Digital Ethology

Human Behavior in Geospatial Context

edited by Tomáš Paus and Hye-Chung Kum

ERNST STRÜNGMANN FORUM

Digital Ethology

Human Behavior in Geospatial Context

Strüngmann Forum Reports

Julia R. Lupp, series editor

The Ernst Strüngmann Forum is made possible through
the generous support of the Ernst Strüngmann Foundation,
inaugurated by Dr. Andreas and Dr. Thomas Strüngmann.

This Forum was supported by the
Deutsche Forschungsgemeinschaft

Digital Ethology

Human Behavior in Geospatial Context

Edited by

Tomáš Paus and Hye-Chung Kum

Program Advisory Committee

Kimmo Kaski, Hye-Chung Kum, Julia R. Lupp,
Maria Melchior, and Tomáš Paus

The MIT Press

Cambridge, Massachusetts
London, England

Series Editor: J. R. Lupp
Editorial Assistance: A. Gessner, C. Stephen
Lektorat: BerlinScienceWorks

The book was set in TimesNewRoman and Arial.

Library of Congress Cataloging-in-Publication Data is available.

ISBN: 978-0-262-54813-7

10 9 8 7 6 5 4 3 2 1

Contents

Human Behavior in Geospatial Context

Mapping Place-Based Context

Human Behavior: Real and Digital

Context and Health

List of Contributors

Balsa-Barreiro, José CITIES, Division of Engineering, New York University Abu Dhabi, Saadiyat Island, Abu Dhabi, United Arab Emirates; MIT Media Lab, Massachusetts Institute of Technology, Cambridge, MA 02139, U.S.A.

Bard, Kim A. Dept. of Psychology, University of Portsmouth, Portsmouth PO1 2DY, U.K.

Bedrick, Steven Medical Informatics and Clinical Epidemiology, Oregon Health & Science University, Portland, OR 97239, U.S.A.

Brauer, Michael School of Population and Public Health, University of British Columbia, Vancouver, BC V6T 1Z3, Canada; Institute for Health Metrics and Evaluation, University of Washington, Seattle, WA 98105, U.S.A.

Brinkhoff, Thomas Dept. of Geoinformation, Jade University of Applied Sciences Oldenburg, and Institute for Applied Photogrammetry and Geoinformatics, 26121 Oldenburg, Germany

Chawla, Nitesh V. Dept. of Computer Science and Engineering, University of Notre Dame, Notre Dame, IN 46556, U.S.A.

Dávid-Barrett, Tamas Trinity College, University of Oxford, Oxford, OX1 3BH, U.K.

Doerr, Megan Applied Ethical, Legal, and Social Implications (ELSI) Research, Sage Bionetworks, Seattle, WA 98121, U.S.A.

Dumas, Guillaume Dept. of Psychiatry and Addiction, University of Montreal, CHU Sainte Justine Research Center,and Mila – Quebec AI Institute, Montreal, QC H3T 1C5, Canada

Ejbye-Ernst, Peter Netherlands Institute for the Study of Crime and Law Enforcement (NSCR), 1081 HV Amsterdam, The Netherlands

Frangou, Sophia Dept. of Psychiatry, University of British Columbia, Vancouver, BC V6T 1Z3, Canada and Icahn School of Medicine, Mt. Sinai, New York, NY 10029, U.S.A.

Friis, Camilla Bank Department of Sociology, University of Copenhagen, Copenhagen K, Denmark

Gilliland, Jason Dept. of Geography and Environment, Western University, London, ON N6A 3K7, Canada

Kaski, Kimmo Dept. of Computer Science, Aalto University, 00076 Aalto, Finland

Keller, Heidi Human Sciences, Osnabrück University, 29074 Osnabrück, Germany

Kon, Fabio Dept. of Computer Science, University of São Paulo, São Paulo, SP 05508-090, Brazil

Kum, Hye-Chung Population Informatics Lab, School of Public Health; Computer Science and Engineering; Industrial Systems and Engineering, Texas A&M University, College Station, TX 77843, U.S.A.

Liebst, Lasse Suonperä Department of Sociology, University of Copenhagen, Copenhagen K, Denmark

Lindegaard, Marie Rosenkrantz NSCR and Department of Sociology, University of Amsterdam, 1018 WV Amsterdam, The Netherlands

Lovasi, Gina S. Urban Health Collaborative, Dornsife School of Public Health, Drexel University, Philadelphia, PA 19104, U.S.A.

Lupp, Daniel P. Division for Methods, Statistics Norway, 0177 Oslo, Norway

Medeiros, Claudia Bauzer Institute of Computing, University of Campinas, Campinas, SP 13083-852, Brazil

Melchior, Maria Sorbonne Université, INSERM, Institut Pierre Louis d'Epidémiologie et de Santé Publique, IPLESP, ERES (Department of Social Epidemiology), 75012 Paris, France

Menendez, Mónica CITIES, Division of Engineering, New York University Abu Dhabi, Saadiyat Island, Abu Dhabi, United Arab Emirates

Pallante, Virginia Netherlands Institute for the Study of Crime and Law Enforcement, De Boelelaan 1077a, 1081 HV Amsterdam, The Netherlands

Paus, Tomáš Dept. of Psychiatry and Neuroscience and Centre Hospitalier Universitaire Sainte-Justin, University of Montreal, Montreal, QC H3T 1C5, Canada

Ritz, Beate Dept. of Epidemiology, Fielding School of Public Health, UCLA, Los Angeles, CA 90095, U.S.A.

Sandin, Sven Dept. of Psychiatry, Icahn School of Medicine at Mount Sinai, New York, NY 10029, U.S.A.; Dept. of Medical Epidemiology and Biostatistics, Karolinska Institutet, 17177 Stockholm, Sweden

Sarker, Abeed Dept. of Biomedical Informatics, School of Medicine, Emory University, Atlanta, GA 30322, U.S.A.

Schmit, Cason D. Health Policy and Management, School of Public Health, Texas A&M University, College Station, TX 77843, U.S.A.

Smith, Lindsey Dept. of Geography and Planning, University of Toronto, Toronto, ON M5S 3G3, Canada

Thompson, Kimberly M. Kid Risk, Inc., Orlando, FL 32819, U.S.A.

Tiemeier, Henning Dept. of Social and Behavioral Sciences, Harvard University, Boston, MA 02115, U.S.A.

Weigle, Michele C. Dept. of Computer Science, Old Dominion University, Norfolk, VA 23529, U.S.A.

Preface

Science is a highly specialized enterprise—one that enables areas of enquiry to be minutely pursued, establishes working paradigms and normative standards, and supports rigor in experimental research. All too often, however, "problems" are encountered that fall outside the scope of any single discipline, and to progress, new perspectives are needed to expand conceptualization, increase understanding, and define trajectories for research to pursue.

The Ernst Strüngmann Forum was established in 2006 to address such topics. Founded on the tenets of scientific independence and the inquisitive nature of the human mind, we provide a platform for experts to scrutinize topics that require input from multiple areas of expertise. Our gatherings, or Forums, take the form of intellectual retreats: disciplinary idiosyncrasies are put aside, existing perspectives are questioned. Importantly, consensus is not necessarily the goal. Instead, participants work to expose gaps in current knowledge and ways to fill these gaps are collectively sought. To ensure access to emerging insights, the results of the entire process are disseminated through the Strüngmann Forum Report series.

This volume reports on the discussions surrounding the topic of "digital ethology" (i.e., the study of human behavior revealed through multifaceted digital footprints). Tomáš Paus (Professor of Psychiatry and Neuroscience, University of Montreal) brought this topic to our attention in 2019. Having participated in two earlier forums, Paus was keen to explore how digital ethology might be used as a conceptual framework and tool to quantify the social environment, and what novel insights into the social dynamics of populations might emerge to generate new knowledge about human behavior across various communities. He invited Hye-Chung Kum (Professor of Health Policy and Management, and Computer Science & Engineering, at Texas A&M University) to join him in preparing a proposal. After review and approval by our scientific advisory board, the Program Advisory Committee was formed to transform the proposal into a framework that would support an extended, multidisciplinary discussion. Joining us on the committee were Kimmo Kaski (Dept. of Computer Science, Aalto University) and Maria Melchior (Sorbonne Université, INSERM, Institut Pierre Louis d'Epidémiologie et de Santé Publique). Together, the committee identified participants and formulated the following overarching goals to guide the discussion:

- To expand understanding of how the environment shapes human development across the life span
- To examine ways through which digital data can broaden research into human behavior and support future comparative behavioral studies across species

- To construct a conceptual and methodological framework for integrating various data sources

Further, the committee established four primary areas around which work would focus and invited "background papers" key topics to initiate the discussion. Originally scheduled to take place from September 20–25, 2020, the Forum experienced delays due to travel restrictions associated with COVID. Ultimately, people traveled to Frankfurt from July 24–29, 2022, for the Forum and a lively discussion ensued between experts from geospatial and data science, behavioral and brain science, epidemiology and public health, ethics, and law, as well as urban planning. This volume synthesizes the ideas and perspectives that emerged.

An endeavor of this kind, especially one developed during COVID lockdowns, creates unique group dynamics and puts demands on everyone. I wish to thank each person who participated in the Forum for their time, efforts, and positive attitudes. A special word of thanks goes to the members of the Program Advisory Committee as well as to the authors and reviewers of the background papers. Importantly, the work of the discussion groups' moderators—Kim A. Bard, Beate Ritz, Jason Gilliland, and Kimmo Kaski—and rapporteurs—Guillaume Dumas, Gina S. Lovasi, Michele C. Weigle, and Claudia Bauzer Medeiros—deserves special recognition: To support lively debate and transform this into a coherent, multiauthor report is never a simple matter. Finally, I extend my sincere appreciation to the scientific chairs, Tomáš Paus and Hye-Chung Kum. Their expertise and leadership accompanied the entire project and contributed greatly to its outcome.

The Ernst Strüngmann Forum is able to conduct its work in the service of science and society due to the generous backing of the Ernst Strüngmann Foundation, established by Dr. Andreas and Dr. Thomas Strüngmann in honor of their father. I also wish to acknowledge the support received from our Scientific Advisory Board as well as the Deutsche Forschungsgemeinschaft, which provided supplemental financial support.

In the attempt to extend the boundaries of knowledge, it is never easy to relinquish long-held views or ideas. Yet once such limitations are recognized, the act of formulating strategies to get past this point becomes a most invigorating activity. On behalf of everyone involved, I hope this volume is able to transfer some of this excitement and be used to create a greater understanding of the relationships between human behavior and the environment through their digital footprints.

Julia R. Lupp, Director, Ernst Strüngmann Forum
Frankfurt Institute for Advanced Studies
Ruth-Moufang-Str. 1, 60438 Frankfurt am Main, Germany
https://esforum.de/

Human Behavior in Geospatial Context

1

Human Brain and Behavior in Geospatial Context

Why and How

Tomáš Paus

General Background

From conception onward, the individual is developing, maturing, working, playing, and aging in their[1] context. As illustrated in Figure 1.1, multiple layers of environment (context) surround an individual across space and time: from the uteroplacental circulation connecting the fetus and their mother before birth, to the influence of their caregivers, extended family, and peers during childhood, adolescence, and adulthood. This "proximal" context (light gray) is embedded in larger geospatial units, such as specific neighborhoods, cities, or countries (dark gray). All environmental influences unfold in time throughout the individual's lifespan. Needless to say, the different layers interact, in a bidirectional manner, with each other. Thus, for instance, a pregnant person responds to signals generated by the fetus, and vice versa (Fowden et al. 2022; Kolle et al. 2020; Menon 2019), the pregnant person interacts with their partner, and vice versa (Khaled et al. 2021; Saxbe et al. 2018), and the caregiver interacts with the child, and vice versa (Carollo et al. 2023; Paquette and St. George 2023). At the same time, the individual and those in their proximal context (e.g., caregivers and peers) act as both recipients and co-creators of their area-level environment along all its dimensions, including physical environment (e.g., air quality), built environment (e.g., parks and transportation network), and social environment (e.g., social cohesion). Different aspects of the environment change over time in an interdependent fashion (e.g., air quality, vehicular traffic, lack of green space, demographic characteristics), often

[1] Throughout this chapter, "they" (and its derivations) is used as a gender-neutral third-person pronoun.

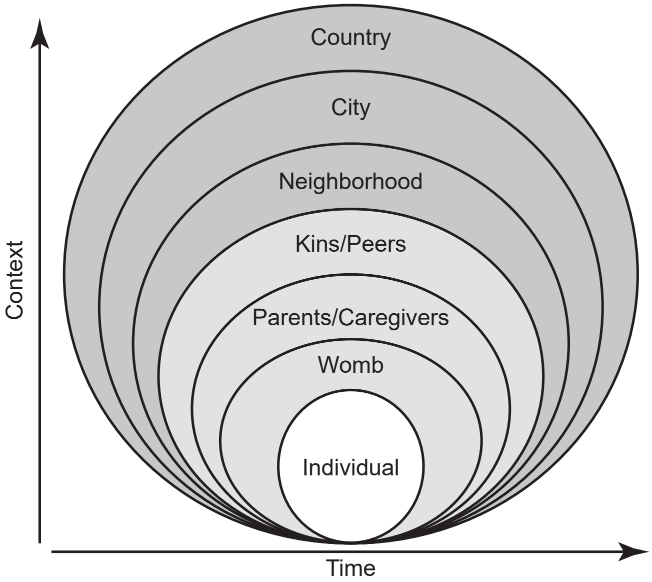


Figure 1.1 Conceptualization of the multiple layers that comprise the contextual environment of an individual across space and time.

reflecting the resources and policies in place at different levels of geospatial granularity (e.g., country, city, neighborhood). Both within and across countries, the lack of environmental justice is reflected in disproportional exposures of marginalized communities to various combinations of adverse environments and, in turn, their *combined* health effects (Van Horne et al. 2023).

For those of us interested in understanding the forces that shape the human brain and behavior, from conception onward, the complexity of this multilayered "exposome" (Munzel et al. 2023; Wild 2005) is staggering. The field of population neuroscience emerged to face this challenge; it brings together epidemiology, genetics, and neuroscience to gain insights into factors underpinning the interindividual variability in the structure and function of the human brain (Paus 2010, 2013, 2016). Owing to the ease of characterizing the individual's genome and the advances in our understanding of related biological processes, initial studies focused on the genetic side of the equation. Working mostly in the context of international consortia, such as ENIGMA (Thompson et al. 2014) and CHARGE (Psaty et al. 2009), we have learned a great deal about the molecular architecture of various quantitative traits derived from magnetic resonance images of the human brain (Grasby et al. 2020; Satizabal et al. 2019; Shin et al. 2020), but efforts on the environment front lags behind. This is understandable given the difficulty of characterizing an individual's environment. Published studies in this area have addressed a handful of factors—one at the time—from the different context layers illustrated in Figure 1.1, such as intrauterine environment (e.g., exposure to maternal cigarette smoking during pregnancy; Muller et al. 2013; Toro et al. 2008), family environment (e.g., family socioeconomic status; Noble et al. 2015), population density (Xu et al. 2022a), as well as variations in the physical (e.g., air pollution; Sukumaran et al. 2023), built (e.g., green space; Kardan et al. 2015) and social (e.g., income inequality; Parker et al. 2017) environments across neighborhoods, cities, and/or countries. Although encouraging, major gaps remain. The Ernst Strüngmann Forum on Digital Ethology, convened in Frankfurt am Main, Germany, in July 2022, brought together scholars and experts to address a number of conceptual and practical gaps in this area.

As pointed out above, the scarcity of multidimensional data that can be used to characterize an individual's environment in an integrated fashion represents the key challenge for studying relationships between the multilayered, multi-domain environment and individual-level outcomes, such as brain development and aging. The Forum addressed this challenge in two ways. *Conceptually*, it called for adopting an ethological approach whereby human behavior is observed, or inferred, in the "wild"; that is, without influencing the observed individual (e.g., by asking them questions). *Practically*, it called for focusing on data sources that either exist or can be readily harnessed at an aggregate level, with different area-level (spatial) granularity (e.g., neighborhood, city, country). The ethological framework presented in Dumas et al. (Chapter 2) underpins the name to this Forum. By "digital ethology," we mean the observation of human behavior through its digital manifestations, such as the use of a search engine, a payment card, or through posting on social media. This behavior leaves "digital footprints" that are particularly relevant for characterizing the social environment of a given area-level unit, as discussed by Weigle et al. (Chapter 4). Human behavior is also reflected and constrained by the surrounding physical and built environments, as outlined by Lovasi et al. (Chapter 3). Finally, variations in individual-level outcomes as a function of the multidimensional area-level environment can best be studied using large datasets; the practicalities as well as legal and ethical considerations are addressed by Medeiros et al. (Chapter 5). Finally, Chapters 6 through 12 provide primers to many of the concepts and strategies that underpin digital ethology.

A Case Study: Inequalities in Area-Level Environment and Brain Health[2]

Social, economic, and political conditions produce *health inequalities* within and across countries (Metzl and Hansen 2018; Scambler 2012; Stuart and Soulsby 2011). In high-income countries, for instance, individuals are more likely to experience poor mental health if they grow up in households with low

[2] This section is a modified version of an article published in *Frontiers in Neuroimaging* (Paus et al. 2022).

income (Bjorkenstam et al. 2017) or affluence (Elgar et al. 2015; Rajmil et al. 2014), live in areas with high deprivation (Kivimaki et al. 2020), or experience inequalities in income distribution (Mangalore et al. 2007). Certain communities are disadvantaged more than others (Waldron 2018). This is especially true for Indigenous (Ogilvie et al. 2021) and racialized (Castro-Ramirez et al. 2021) communities, which are at higher risk for mental-health problems and simultaneously experience a lower likelihood of receiving evidence-based treatment (Castro-Ramirez et al. 2021). At the area level, our physical, built, and social environments combine to create ecosystems in which we live and work. Together, these ecosystems, as well as the structures and systems that produce them, contribute to what has been termed "social and structural determinants of health" (Diderichsen et al. 2001; Vandenbroucke 1990).

As described elsewhere (Paus 2016), there are countless permutations of the physical, built, and social environments that surround us in space and time. We both "receive" and "create" our environments (Kendler et al. 2003), thus co-determining what air we breathe, how many steps we take, how hot or cold we are, as well as what and who we see, hear, and interact with during our commutes. Together with our genes, these "external exposures" contribute to "internal" environments that exist in our body: on body surfaces (e.g., microbes on our skin and in the gut), in the lungs (e.g., particulate matter), circulating blood (e.g., toxins, micronutrients, inflammatory molecules), and the brain (e.g., stress- and reward-related neurotransmitters, cumulative engagement of specific neural circuits).

As pointed out above, the use of aggregate-level (spatial) data, produced from multiple locations and time points, is one strategy for characterizing physical, built, and social environments surrounding the individual. In turn, linking such aggregate-level data with individual-level information about a person's health in general, and brain health in particular, provides the first step toward understanding these relationships. Below, the basic steps in this process are reviewed, which are covered in depth in Chapters 6–12.

Geospatial Mapping of Area-Level Environments

Geospatial science and related tools enable spatial analysis and visualization of the external environments in which we spend considerable amount of our lives (e.g., our residence, place of work, school, recreation or a commute path) and an evaluation of their impact on our health. Datasets can be created at different levels of spatial granularity matching the goals of a given study and availability of relevant data. In Canada, for example, geographic units include six-digit postal codes, Canadian Census geographic units such as dissemination areas (400 to 700 persons), and census tracts (2,500 to 8,000 persons), or larger areas such as city districts. The spatial unit used to link geospatial datasets to health data varies; depending on the study and actions necessary to protect confidentiality of study participants, this can be as precise as the exact street address or a postal code (half of a city block in dense urban areas), or as coarse as a city district, a county, a province/state or a country. The temporal dimension depends on the type of data; it may range from data sampled monthly (e.g., air quality), annually (e.g., public transportation), or up to every five years (e.g., the Canadian Census).

Spatiotemporal datasets can be created using existing tools and databases provided by large GIS-based (geographic information systems) organization and companies, such as ESRI, DMTI Spatial, Google Earth Engine, as well as open sources (e.g., Open Street Map), government sources (e.g., Statistics Canada), and academic organizations. In Canada, we have acquired, curated, and disseminated geospatially coded information about the physical and built environments through the Canadian Urban Environmental Health Research Consortium, CANUE (Brook et al. 2018). Metrics derived from different sources can be combined to ask, for example, questions about the relationship between socioeconomic indicators (e.g., household income) and the built environment (e.g., access to parks), and thereby used to assess inequity in the spatial distribution of environmental good or hazards. Figure 1.2 illustrates inequality in the access to parks and recreation (derived from Open Street Map data) across areas with a high level (top 20%) of material deprivation (derived from Canadian Census data; Pampalon et al. 2012).

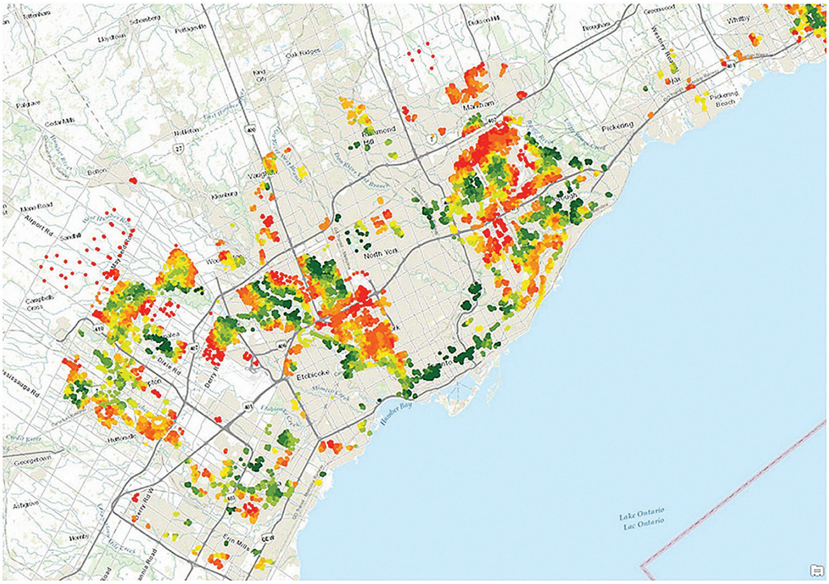


Figure 1.2 Material deprivation and access to parks and recreation in the Greater Toronto Area. All colored areas represent postal codes characterized by high (top 20%) material deprivation (Pampalon et al. 2012). Green indicates postal codes in the highest 10% density of park and recreational amenity within 1 km; red indicates postal codes in the lowest 10% (Source: Open Street Map).

In addition to sourcing and creating data about physical and built environments from existing databases (see Table 1 in Paus 2016), one can also derive relevant measures from new data streams such as high-resolution satellite and street-level imagery combined with machine-learning techniques (Weichenthal et al. 2019). For example, Google Street View allows investigators to assess different features of the built environment using panoramic street-level images taken mostly by camera-equipped cars, while recent satellite technology provides daily coverage of most inhabited areas on Earth at a resolution of only a few meters. These geocoded images can be rated for various features, such as signs of physical disorder (e.g., litter, graffiti), physical decay (e.g., poor conditions of sidewalks), type of stores, traffic, or street walkability (Less et al. 2015; Odgers et al. 2012); this approach does have, however, some limitations (Curtis et al. 2013). In turn, computer vision and machine-learning algorithms can exploit these image data to generate indirect indices of the social environment (e.g., psychosocial stress) and physical environment (e.g., air or noise pollution) in a manner similar to that used by others to derive measures characterizing living environment, health, and crime (Suel et al. 2019).

As summarized in Table 1.1 (social environment), a wealth of data speak to basic (often self-reported) measures of socioeconomic factors (e.g., education, employment, immigration, household spending habits, volunteering, and giving) collected by governmental agencies (e.g., census) and national surveys. One can, however, also use data from digital streams (e.g., search engines, social media) to generate new measures of the social environment that are relevant for attitudes vis-à-vis health and health interventions (e.g., vaccination), as well as social cohesion, social support and role models and, most recently, for the emerging issues related to environmental anxiety (Hickman et al. 2021; Soutar and Wand 2022; To et al. 2021; Usher 2022).

Once properly curated, all aggregate-level data (e.g., see Table 1.1) should be described using comprehensive metadata and coded to different geographic units (e.g., postal codes, dissemination areas, and other census geographies), as has been previously done by CANUE.

Linkage with Individual-Level Data

Ultimately, what we are interested in doing is to link aggregate-level "exposures" described above to individual-level "outcomes." In this section, two examples illustrate how this can be achieved using administrative health databases and data acquired in research cohorts.

Administrative Data

In the recent past, we have all seen the power of mapping administrative data related to COVID-19 (across countries, provinces/states, or cities) and communicating these numbers to the public. In Canada, administrative health data

Table 1.1 Examples of measures, with the corresponding sources of raw geospatially coded data and examples of the new types of data to be derived.

Physical and Built Environment	Social Environment
Air quality (NO_2, O_3, SO_2, PM2.5)[1]	Demographic (b)[6]
Greenness (greenest pixel, tree canopy)[1]	Households (c)[6]
Nighttime light[1]	Socioeconomic (d)[6]
Noise[2]	Water quality concerns[7]
Public transportation[3]	Composting and recycling behavior[7]
Proximity to roads[4]	Involvement in outdoor activities[7]
Proximity to retail outlets and sales of alcohol, tobacco, cannabis, gambling[4]	Caregiving and care receiving[8]
Green roads[5]	Social identity[8]
Facility index[5]	Giving, volunteering, and participating[8]
Cumulative opportunities (a)[5]	Victimization[8]
	Social media and search engine use by youth: frequency and time of day[9]
	Social media and search engine use by youth: content[9]
	Built environment predictors of psychosocial stress[10]
	Built environment predictors of social cohesion[10]
(a) Travel times (walking, public transport) to jobs, leisure, and shopping, as well as health, medical, and social services (b) population (total and densities), proportions (by age, sex, ethnicity, marital status, mobility/migration status, religion, mother tongue)	(c) household size, total housing units, proportion rented, type of dwelling (d) household income, unemployment rate, proportion below poverty line, proportion (by age/sex) in labor force

Sources:

[1] Landsat
[2] CANUE
[3] OpenStreetMap (OSM)
[4] DMTI Spatial
[5] OSM and CANUE
[6] census
[7] Household and the Environment Survey (Canada)
[8] The General Social Survey (Canada)
[9] newly derived measures from raw data streams (e.g., Twitter/X, Google search engines),
[10] newly derived measures from raw data streams (satellite and street view imagery)

(i.e., data captured during the course of providing services or running programs) are made available for research use by provincial governments and other agencies, often in close partnership with academic organizations (Lucyk et al. 2015). In all provinces, these data are longitudinal and population-based, covering all residents who have received health care and social services (e.g., education), from birth onward. This creates comprehensive and important data for the population of interest, such as youth.

In the province of Ontario, for example, administrative health data are curated and made available for research by the Institute for Clinical Evaluative Sciences (ICES), a not-for-profit research institute made up of a community of research, data, and clinical experts that provide a secure and accessible inventory of Ontario's health-related data. Behind a firewall, ICES provides access to coded and linkable databases containing, for example, the Ontario Mental Health Reporting System. Just in the City of Toronto, these data are available for about 270,000 adolescents and youth (12–22 years of age). In addition to health data, many of the provincial custodians of administrative data provide access to other linked datasets, such as education, workplace or justice data (e.g., Population Data BC). When linking administrative data with geospatial datasets containing area-level characteristics of the physical, built, and social environment, one would typically use the residential six-digit postal codes (Canada) and relevant geographies (e.g., dissemination blocks) reported in the administrative data for each individual. Postal code-indexed geospatial datasets are linked in the secure environments controlled by the custodian of the individual-level health data (Boyd et al. 2013; Kum and Ahalt 2013; Pencarrick Hertzman et al. 2013). Here, ethical and legal guidance is necessary to provide assurance to data stewards that this form of data linkage and access can be done in a privacy-preserving and transparent manner that respects all applicable legal, regulatory, and ethical requirements. Ongoing efforts address issues relevant for ensuring public trust, such as transparency of the current practices and systems of governance, and understanding public opinion regarding the use of "big data" in the service of population health (Aitken et al. 2016; O'Brien et al. 2019; Schmit et al. 2021).

Cohort Studies

One of the key advantages of administrative health data is their population-wide coverage. By definition, these data show only the tip of the "health iceberg"; namely, individuals with health issues significant enough to enter the health-care system. This is where community-based cohort studies come in as a complementary source of information, with longitudinal birth cohorts being most valuable. For example, birth cohorts are well suited for investigating relationships between brain health (individual-level data) and context (aggregate-level characteristics of the environment) for several reasons:

1. Many birth cohorts, such as ALSPAC (Boyd et al. 2013), Generation R (Tiemeier et al. 2012) and Northern Finland Birth Cohorts (Rantakallio 1988), ascertained their participants (pregnant women) in a relatively small geographic area.
2. Each cohort includes a relatively large sample size of individuals (~10,000).

3. Brain (e.g., mental) health of cohort members is assessed using a number of instruments, often on a continuous scale.

The combination of the first two features makes it likely that a reasonable number of participants live in each geospatial unit, hence providing sufficient statistical power to investigate these relationships. The third feature (assessment) permits the capture of “subclinical” mental-health problems. Finally, additional deep-phenotyping of cohort members through, for example, cognitive assessment, neuroimaging, blood-based biomarkers (e.g., inflammation), genotyping and epigenotyping provides rich information suitable for detailed modeling of exposure–outcome relationships and their mediators and moderators (Paus 2013).

Social Inequality and Mental Health

To close this section, let us consider a hypothetical example illustrating how one can use aggregate-level information about the physical, built, and social environments to unpack the relationship between poverty and mental health. As pointed out by Diderichsen et al. (2001), and represented in Figure 1.3, social stratification—with poverty being but one example of social, economic, and political inequalities—generates a vicious circle: Disadvantaged persons

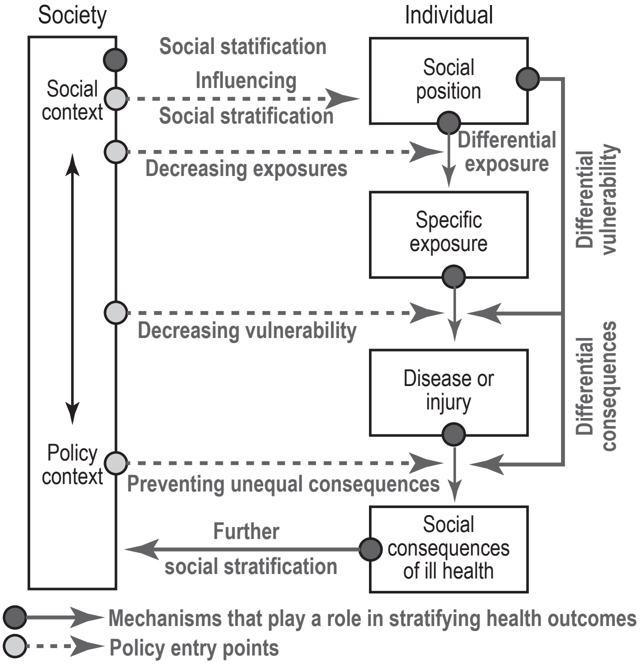


Figure 1.3 From structural inequalities to ill health (Diderichsen et al. 2001).

are more likely to be exposed to harmful or deprived physical (e.g., air pollution), built (e.g., access to food stores), and social (e.g., lack of social support) environments as well as to population-level challenges (e.g., heat waves, SARS-CoV-2). These exposures lead to an increased vulnerability to other exposures (e.g., victimization), and both the exposures and vulnerabilities combined precipitate (mental) illness. This vicious circle is closed by the illness leading to further social stratification (e.g., lost educational and employment opportunities). Having extensive multi-domain area-level datasets that can be used to characterize the physical, built, and social environments would enable us to test a variety of possible pathways (and their combination) leading from social stratification to brain health; decomposition analysis is but one method that can be used to quantify contributions of various factors to the observed outcomes (O'Donnell et al. 2008).

Looking Forward

As discussed by Lovasi et al. (Chapter 3) and outlined in the environmental justice framework for exposure science (Van Horne et al. 2023), complexities of the multilayered relationships between the individual and their environment require not only top-quality data and conceptual and analytical approaches but also meaningful engagement with communities and their policy makers, as well as development and implementation of adequate strategies by funders, academic institutions, and journal editors working in this research field. Innovative methods should be employed to address one of the main limitations of observational strategies, namely the difficulty of making causal inferences (see Dumas et al., Chapter 2). For example, dense time-series of multiple exposures and outcomes offer an opportunity for estimating Granger causality (Imran et al. 2023). Pseudo-experimental design can explore causality and directionality in cases of discrete events that affect local environment; note that events such as forest fires may impact not only physical (air quality; Khraishah et al. 2022) but also built (loss of infrastructure) and social (evacuation) environments. To begin to address this issue, at least for certain environments, Mendelian Randomization (Smith and Ebrahim 2003) could be used. For example, using genetic variants associated with biomarkers of low-grade inflammation (Liu et al. 2019; Xu et al. 2022b), one can test a mechanistic path by which air pollution affects brain-related outcomes (Fani et al. 2021). Finally, as reflected in the diversity of the Forum participants, this enterprise requires experts from wide-ranging domains, including geospatial and data science, behavioral and brain science, epidemiology and public health, ethics and law, as well as urban planning. Finding a common language and purpose will allow us to work together toward the understanding of how humans transform their environments and how environments shape human brain and behavior.

2

How Can Concepts of Ethology Be Applied to Large-Scale Digital Data?

Guillaume Dumas, Sophia Frangou, Heidi Keller, Daniel P. Lupp, Virginia Pallante, Tomáš Paus, and Kim A. Bard

Abstract

The ethological approach is used to study naturally occurring behavior. In the modern world, many such behaviors are connected to, and recorded by, a wide array of digital services (e.g., social networking, information search, closed-circuit television). How can ethological concepts be applied to help us characterize the environment in which humans live? What aspects of the ethological approach can guide us to obtain measures captured directly from digital data generated by our everyday activities? What kinds of models do we need to understand how human behaviors/activities can be inferred from the physical and built environment? This chapter explores the bidirectional nature of these relationships; namely, how individuals create their environment, and how the environment shapes the individual. It discusses how to proceed from observation and data sampling to knowledge extraction and causal inference. The complementary nature of common and specific are addressed as well as the challenge of integrating niches at both physical and social levels. Finally, all these concepts and associated methods are illustrated through a hypothetical study.

Reflecting on Observation

> What we observe is not nature itself, but nature exposed to our method of questioning. —Werner Heisenberg (1958)

The early ethologists Karl von Frisch, Konrad Lorenz, and Nikolaas Tinbergen, relied on observations as their core method of inquiry. After careful observation, they would fully describe the behaviors of interest. Then, as a second step, they would contemplate the function of these behaviors, assessed through a process of classification of possible behaviors, field experiments, and comparison of behaviors within and across contexts or species. Tinbergen (1963) argued that behavior could be explained on four levels: ontogenetic, phylogenetic, proximate (immediate cause), and ultimate (evolutionary reasons). By systematizing their observations and focusing on the more fixed behaviors, they could easily replicate their findings. While watching, they wondered about the parameters of the behavior and conducted field experiments from which they could determine proximal causes. For example, one of Tinbergen's experiments involved understanding how digger wasps locate their home burrow after flying away in search of food (Tinbergen 1972). He conducted a series of experiments that involved placing a pinecone at the burrow entrance as the wasps were leaving and then moving it to a nearby location while the wasps were away. Upon their return, the wasps flew to the relocated pinecone rather than to their burrow entrance. In this way, Tinbergen discovered that digger wasps use landmarks to identify their burrow.

Another important study by Tinbergen focused on courtship in stickleback fish (1952), where he identified the specific behaviors of the male to which its prospective mate responded. While this may seem anecdotal, this led to the creation of one of the first ethograms (i.e., a comprehensive list, inventory, or description of the behavior of an organism) (Figure 2.1a), the key tool of ethology. Eibl-Eibesfeldt adopted these same principles of detailed observation coupled with experimental causal inference when beginning the field of human ethology, focusing on the structure of human behavior from recorded observations of people living in diverse settings around the world (Eibl-Eibesfeldt 1989). Since then, the field has grown considerably to encompass a wide range of activities, highlighted by the International Society for Human

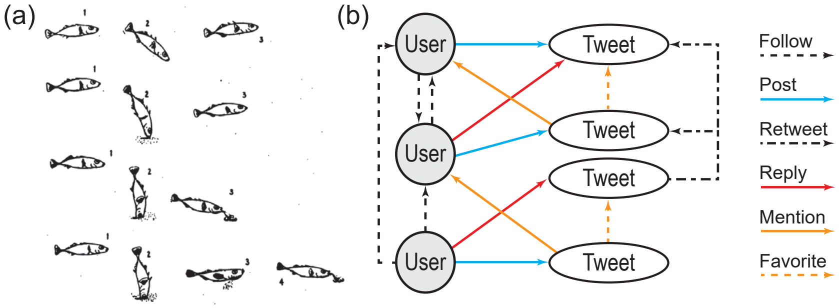


Figure 2.1 Two types of ethograms: (a) A traditional one showing four levels of intensity during the excavation of sand by the three-spined stickleback (*Gasterosteus aculeatus*) (Tinbergen 1951). Intensity ranges from minimal (top) to maximal (bottom); numbers indicate the sequence of behaviors: (1) swimming, (2) digging sand for a nest pit, (3) losing sand through the gills, and (4) spitting out the sand. (b) An example of a digital ethogram based on a user–tweet interaction model (after Belkaroui et al. 2015), showing the six canonical behaviors (arrows) on the social media network Twitter/X.

Ethology. Recently, the interdisciplinary field of computational ethology has emerged at the crossroad of physical computer and life sciences. Here, the idea is to leverage recent progress in machine learning to create ethograms through automatic detection and analysis of behavior while still using individual-level behaviors as the main unit of observation (Anderson and Perona 2014). In this chapter, we focus on digital ethology, which relies on large-scale digital data coupled with geocoding of physical and social environments. In this respect, individual-level behaviors are aggregated at the level of geospatial units, thus guaranteeing better safeguarding of the privacy of individuals while allowing researchers to examine human–environment bidirectional relationships.

Ethologists observe behavior and ponder about its function. So, when Tinbergen watched the digger wasps, he saw them fly in a pattern over the burrow as they emerged, as well as before they reentered the burrow. After extensive observation, he was able to draw the pattern in which they flew. He wondered why they flew like that, under those circumstances. Through field experiments, he determined the proximate cause: the digger wasps encoded features of the landscape that identified their own burrow. Behavior is what is observed, and the construct is either an explanatory or a functional mechanism. The documentation of behaviors allows the ethologist to determine ethograms. Current ethologists develop ethograms of select behaviors to answer specific questions. The constellation of behaviors found to be related to a specific outcome measure might therefore constitute a construct, such as environmental variables related to an increased risk for depression. Interestingly, ethology can also gather information from the constraints of those behaviors (see Figure 2.1a).

Digital ethology poses unique challenges that must be managed if we are to generate an ethogram. Figure 2.1b provides an illustration of how a digital ethogram could be generated using the “tweets” with “influence” as the construct of interest. Like a typical ethogram, Figure 2.1b shows the observed activity of Twitter/X users and the behavioral patterns that arise from their interactions. In this example, constraints are built into the Twitter/X platform; similar to the epigenetic landscape of Waddington (1957), constraints may also be embodied in the physical environment through space, resources, and risks. Pallante et al. (this volume) demonstrate how such constraints influence the probability of engaging in a specific behavior, such as resolution or reconciliation. Digital ethology could thus gather information about area-level constraints on certain behaviors or activities coming from other domains, such as the built environment (e.g., accessing “resources” such as food when stores are not nearby). A digital ethogram would contain selected digital data thought to represent area-level aggregates of behaviors, naturally occurring variations in these behaviors, and assumptions about their functional significance and underlying mechanisms.

The creation of an ethogram implies a process of reduction and simplification aimed at controlling and describing the observations. The boundaries that delimit a behavioral pattern underpin the quantitative approach in ethological studies: the ethogram is meant to be the coding scheme that quantifies the observations, which necessarily leads to a reduction of the variability observed. For the early ethologists, ethograms refer to the complete set of behaviors. In typical modern-day studies, especially in humans, it is nearly impossible to construct an ethogram of all behaviors. Thus, ethograms must necessarily be comprised of select behaviors of interest. Notwithstanding, ethograms are descriptions of (parts of) the observed behavioral repertoire, aimed at capturing and explaining the variability of this repertoire in time and space.

Early in the development of an ethogram, *ad libitum* observations of behaviors are required (Altmann 1974). This means that an inductive approach must be taken, usually by nonsystematically recording the behavioral patterns observed in a group of animals. This helps the researcher to become familiar with and gain insights into the behavioral repertoire of the species. The *ad libitum* nature of such a sampling technique relies on the ability to collect as many observations as possible when behaviors, individuals, and time sessions are chosen without restrictions. The behaviors classified into the ethogram are those that can be clearly described, follow a specific pattern, are limited and repeated over time, and are usually performed by several individuals in the colony. The recording of behaviors is conducted by naming and describing the specific behavioral patterns observed. These notes will turn into items of the ethogram.

Interobserver reliability is a major concern for modern-day ethologists, especially since the behaviors under observation are not usually fixed (i.e., with fixed releasers and fixed forms). Thus, after the development phase of an ethogram and before all observations are coded, a second independent observer is trained to apply the ethogram to evaluate if the definitions of the behaviors included are clear enough to allow for their coding. This may lead to a modification of the original ethogram. Once agreement is reached at an acceptably high level (e.g., Bakeman 2023), the ethogram can be applied for the data collection. Modern-day ethologists use observational methodologies of observing systematically with well-defined ethograms, coding schemes specifying how observations are collected, as well as inter- and intra-observer reliability, etc. (For a review of observational methodology, see Bakeman and Quera 2011.)

Despite technical methods, one may still question how "reliability" differs from "subjectivity," especially when qualitative and quantitative approaches are compared (for an overview of terms, see Appendix 2.1). This discussion is about the quality of assessment procedures, including observations. Reliability and validity are the main characteristics of quality in quantitative approaches (Mays and Pope 2000). While quantitative methodology tries to objectify subjectivity, qualitative methodology tries to represent the subjective meaning systems of the research participants. Because of this subjective component, qualitative methods need alternative validation methods. Just as in quantitative methods, different methods exist, but the rigor of a qualitative study is mostly represented by its trustworthiness, defined by the confidence in the data.

From Data Sampling to Knowledge Extraction

You cannot see things till you know roughly what they are. —C. S. Lewis (1943)

Data collection—the first practical step in the journey—is open to discussions about validity and reliability. Classic ethology relies on prolonged observations of activity in humans or nonhuman species from which the observer can begin to identify distinct patterns (i.e., behavior) to build an ethogram. Even in purely observational studies, the influence of the observer on the activity being observed, and hence the validity of observations, is often a matter of debate. This issue becomes even more complex when one attempts to access internal states (e.g., motives, salience attributions) that can be reported or inferred in human studies but only conjectured in nonhuman species. In all instances, whether research is considered "qualitative" or "quantitative," the observer remains part of the observed in terms of the behaviors selected for observation, the instruments used to record behavior, and the attribution of function and cause. The reliability (i.e., consistent reproducibility) of behavioral measures is also beset with difficulties as simply having a quantitative index of behavior is not sufficient.

The availability of technological platforms (e.g., video cameras, sensors) can minimize the perceived presence of the observer and aid in measurement reliability as they can provide quantitative estimates of behavior that minimize inter- and intra-rater variability. The role of the observer, however, remains integral to the process. For example, Twitter/X users are aware that their behavior is being observed and exploit this to make their behavior accessible to a large number of observers. Similarly, the use of video or CCTV data to study human behavior, such as conflict resolution, requires significant observer input even though the observed are usually not aware that they are being recorded at the time. Often, members of a lab sit together in front of videotapes and discuss jointly what they observe and what the observed behavior could mean. Other methods involve triangulation and respondent validation. In data science, this triangulation weights the validity of a given approach by comparing it to other data, other technical approaches, or both (Oppermann 2000). In epidemiology, a similar triangulation is used (Lawlor et al. 2016), but validity can also be inferred from counterfactual reasoning (Höfler 2005). In both qualitative and quantitative research, it is essential that the research process, including the subjective perception of the researcher, is made transparent and conscious.

Digital ethology needs to adapt those different approaches to data collection and evaluation that address the unprecedented scale of the data and their heterogeneity. Similar to other forms of ethology, the goal of digital ethology is constant: to observe and ponder what is meaningful at different functional levels for the problem under consideration. Whether it is precision medicine, understanding a mechanism, or identifying external stressors, all the previous scientific knowledge will partially constrain the search space of constructs and measures. In this sense, before any data have been collected, some choices are already made, explicitly or even unconsciously. As there are no definitive answers to these issues, awareness and transparency are essential.

The road from data to knowledge passes through information before reaching it and, hopefully, heads off in the direction of wisdom. This road rarely follows a straight line, however, and to avoid getting lost en route, a map is useful. Data science is an informed exploratory process (Huber 1996). Intelligent data exploration requires a clear hierarchy of analysis plans, from the definition of which data structure (e.g., variable, relationship, model) should be considered to the choice of plan (or plan type, in the case of higher-level decisions) that is appropriate. Thereafter, one is faced with the problem of deciding between different plausible but not always consistent results produced by different analytical procedures (e.g., least squares, resistant line, number of data partitions). In optimization, the "units" or "domains" of analysis may vary by the predefined outcome. If a massive dataset is to be analyzed without the direct supervision of a human user, then a representation of the process conducted becomes an even more necessary component of the result. This is one of the key distinctions between planning and other forms of search: the aim is to generate a sequence (or more complex combination) of operations, not simply a result.

It all starts with data and, in the case of digital ethology, with *big data*. Big data is an umbrella term used to describe datasets whose size and structure are so large and complex that conventional computational tasks become unfeasible. The term is commonly associated with vast amounts of data, although it should not be constrained by such a narrow definition. Doug Laney (2001) defines big data with respect to the three Vs:

1. Volume, which refers to the size of the dataset in multiple dimensions (i.e., in the number of records or the number of recorded variables).
2. Velocity or the speed at which data is gathered and processed.
3. Variety, which describes the heterogeneity in the structure of data gathered.

Laney's definition serves as a basis for many alternative definitions that often add additional Vs (e.g., veracity). Still, consensus is lacking on one specific definition. Independently of the definition to which one subscribes, a key point to keep in mind is that big data does not solely refer to size but entails various aspects of complexity within the data and data collection. Thus, for instance, it is entirely possible to have a big dataset with comparatively low volume, but high velocity and variety.

How then do we move from (big) data to knowledge? Data are usually considered as raw measures that have not yet been contextualized. Although the choice of recording one measure rather than another is already contextual, the switch from data to information is usually when those measures are contextualized at the moment of analysis. In this sense, raw data are "dry" and one needs to "rehydrate the data" (Claudia Bauzer Medeiros, pers. comm.) to be able to interpret it. Information emerges by moving from “raw” data to “minimally processed” (e.g., satellite images cleaned from artifacts) or “pre-processed” data (e.g., engineered features extracted from satellite images such as roads, sidewalks, and building types). The move from information to knowledge is then linked to the interpretation of the information and the generation of meaningful claims. In a sense, knowledge generation cannot happen solely based on analyses of information. It needs an outcome; that is, explanatory theories must be generated about a specific phenomenon. Big data might be considered as a catchall ethogram for all possible behaviors, but a more explicit ethogram is needed to answer specific questions and extract laws that govern the associated phenomenon.

With big data come big analyses, and the technological progress in computing has enabled the fast-paced development of artificial intelligence methods such as deep learning. A key issue in machine learning is the generalizability of the results to another context. In this respect, metadata are critical to understanding the link “who was collecting, how, why, and where?” (For further discussion, see Lovasi et al., this volume.) It is important for the sampling to be as representative as possible of the population of interest. Still, it is hard to guarantee the representativeness of the sample regarding the whole population, as the whole population is not a *representation*; it is a *description* (see Medeiros et al., this volume). Knowledge extraction is thus not a monolithic activity; it can come from imposing or discovering structure.

What happens when no structure at all is imposed? Here, bioinformatics offers some clues. Indeed, technological developments in genetics have driven a move from the traditional single-gene approach to a polygenic and even “omnigenic” perspective (Boyle et al. 2017). This new way of viewing the genome emphasizes the interdependence of genes and leverages new digital tools to measure holistic effects at the molecular level. In digital ethology, progress in artificial intelligence may induce a similar shift in ethogram construction, from considering a few discrete canonical behaviors to embracing all observable behavioral patterns. The challenge becomes one of interpretability, since algorithms may detect and use behaviors that are imperceptible to the eye of a human observer. The result could outsmart traditional human ethology in prediction, but at the price of having a clear inference of underlying mechanisms. Thus, beyond prediction, a clear challenge for digital ethology in knowledge generation remains the inference of causality.

Inferring Causality

> You are smarter than your data. Data do not understand causes and effects; humans do. —Judea Pearl and Dana Mackenzie (2018)

Humans and, in particular, scientists often like causal explanations. The concept of causality has been debated for centuries by diverse disciplines that have emphasized different aspects of causality. For instance, sociology, anthropology, and psychology place significant emphasis on the *context* in which causal relationships are examined. Causality thus comes in different flavors and requires different approaches to assess it. In ethology, there are at least four different types: ontogenetic (caused by the development), phylogenetic (caused by the evolution of species), proximate (caused by immediate physiological or environmental factors), and ultimate (associated with goals and function from an evolutionary point of view; Tinbergen 1963). In a broader scientific context, three general classes of causality are usually specified: direct, structural, and logical (Craver 2007). Direct and structural causality require, most of the time, an experiment (an empirical intervention with planned perturbation of the system) or at least a quasi-experiment (the use of events that occur independently from the research planning but could nevertheless be exploited to infer how those events causally impact the system). While direct causality is associated with physical events and mechanisms in time (e.g., an earthquake destroyed a city), structural causality is associated with physical objects and mechanisms in space (e.g., the transportation system constrains the growth of the city). Finally, logical causality is independent of time and space since it relies on abstract propositions, reasoning, and implication.

In the case of big data, when trying to determine a causal link between variables A and B (i.e., "the presence of property A causes the likely presence of property B"), one invariably stumbles across known problems. For instance, when analyzing a given dataset D for causal links, one wishes to determine whether property A is necessary and sufficient for the (likely) presence of property B. In this context, both "necessary" and "sufficient" are needed: without "necessary" we cannot deduce that A is the cause of B, and without "sufficient" we cannot deduce that B is the result of A. Though this is the method to determine causal links, it would be incorrect to infer from this that a causal link is a necessary and sufficient condition. Indeed, mathematically speaking, necessity and sufficiency is a characterization of equivalence, not implication or causation. So how can it be that we determine causation (itself a form of implication) by checking equivalence? Here, it is important to note that necessity and sufficiency are both determined with respect to the given dataset D. In other words, property A is necessary and sufficient for property B within the dataset D. This does not, however, mean that A is necessary and sufficient for B in all other datasets. In practical terms, this is the same as saying that there might be other settings where something other than A causes B as well.

This apparent discrepancy is precisely the gap between the *closed-world assumption* (CWA) and *open-world assumption* (OWA) (Reiter 1978). Under the CWA, inference is made with respect to a given dataset D: a statement is considered true if and only if it is true in D. For example, for a dataset containing three records of people—Anna, Bob, and Catherine—the statement "David is not a person" would be considered true. Under the OWA, a statement is considered valid[1] or true if and only if it can be proven to be true; that is, it is true for all possible datasets. In the previous example, "David is not a person" cannot be proven so it can be both true and false depending on context.

What does this have to do with the actual analysis of causality? Unless the dataset being analyzed provides an accurate description of the entire universe of discourse, there will always be a limitation to the relationships we can determine as causal via necessity and sufficiency (inference of probabilistic equivalence using the CWA) and "actual" causation (inference of probabilistic implication using the OWA), since we will never manage to prove that nothing other than A can cause B. In a different context, it might well be the case that C causes B. To summarize, necessity and sufficiency within a dataset only provide a tool for revealing possible causal links; they do not define causality. Though we will likely never be able to close completely the gap between the CWA and OWA, analysis of multiple datasets and access to big data describing a more complete view of the relevant data can narrow the gap.

In biomedical research, causality is usually considered when a change in one parameter (cause) within a system is associated with a change in another parameter or the wider state of a system (effect).[2] We can consider two subtypes of causality. First, idiographic causality concerns itself with causal relationships for specific units or events such as a group (considered as singular entities, regardless of how they are defined), an individual, or specific event (Molenaar 2004). For example, the idiographic approach to depression would focus on the cause of depression in a single individual without requiring or even being concerned as to whether the same causal factors may or may not apply to other people. Second, nomothetic causality is concerned with factors that are generalizable to other contexts (i.e., individuals, groups, events), such as the causes of depression whenever and wherever it occurs.

Another aspect of causality refers to its *nature*, which is conventionally considered in terms of deterministic or probabilistic and necessary or sufficient (Khemlani et al. 2014). Deterministic causal relationships require that a change in a specific parameter (parameter A) is *always* followed by a change in another specific parameter (parameter B). In probabilistic causality, "always" is replaced by "frequently"; in other words, a change in the parameter A increases the probability that a change in parameter B will occur. Such "causes" are referred to as "risk factors" or "enabling conditions" with the latter avoiding assumptions about the desirability of the effect. A causal relationship is considered necessary when a change in parameter B can never happen unless there is a change in parameter A. A causal relationship is considered sufficient when a change in parameter A can cause a change in parameter B, although changes

[1] Note: one refers to validity under the OWA as opposed to truth.

[2] "Cause" and "effect" may be named differently in other fields; for example, "exposure" and "outcome" in epidemiology or "independent variable" and "dependent variable" in psychology.

in parameter B can occur through changes in other parameters (Rothman and Greenland 2005).

A further important dimension of causality is the criteria that need to be met to consider an association between changes in parameters A and B as cause and effect. The most influential set of considerations in the context of medicine was established by the British statistician and epidemiologist, Austin Bradford Hill (1965), and are

- strength of association,
- consistency of association,
- specificity of association,
- temporality, biological gradient (dose-response relationship),
- biological plausibility,
- coherence with previous knowledge,
- experimental evidence (e.g., clinical trials, intervention studies including natural experiments), and
- analogy (i.e., testing that there are analogous causal mechanisms in certain animal models and humans).

Modern medicine and epidemiology tend to rely increasingly on counterfactual reasoning and related approaches to infer causality (Höfler 2005).

In the case of digital ethology (and many other fields), the problem of causality becomes more complicated because causality may be bidirectional, where changes in parameters A and B can reiteratively influence each other. For instance, our social and built environments form ecosystems that contribute to what has been termed "social and structural determinants of health." Thus, we both "receive" and "create" our environments, while codetermining what air we breathe, how many steps we take, how hot or cold we are, and what and who we see, hear, and interact with during our commutes (Paus et al. 2022). Ultimately, causality also moves across levels of organization, from the emergence of collective dynamics to the downward causation when individuals tune their behavior in response to estimates of collectively computed macroscopic properties (e.g., social inequality; Flack 2017).

Common and Specific

> Brian: You're all individuals!
> Followers: Yes, we're all individuals!
> Brian: You're all different!
> Followers: Yes, we are all different!
> Dennis: I'm not.
> —Monty Python, from the *Life of Brian* (1979)

The availability of large-scale digital data has the potential to enable the interrogation of behaviors across diverse cultural, ethno-racial, and socioeconomic human groups. Also of interest is how human behaviors may relate to behaviors observed in other species, such as nonhuman primates. Behavioral patterns (either human or nonhuman) are typically assigned to different constructs that are theoretically defined (e.g., attachment).

Differences between groups may arise at the theoretical meaning of a construct. If constructs are theoretically deemed to be similar, differences may arise in the behaviors assigned to the construct in distinct groups (or species) or the measures developed to assess the construct in these distinct groups (or species). Further, even though the construct and context remain constant, there may still be different measures for assessing this construct representing preferences or conventional practices among researchers. Establishing equivalence of constructs and measures is a prerequisite for comparative studies and a complex task in itself because there is no universally agreed definition of what constitutes equivalence and how it can be established.

We advocate for the scheme provided by Hui and Triandis (1985), who consider equivalence between constructs at the conceptual, functional, item, and scalar levels. Conceptual equivalence requires that a construct has the same meaning across groups (or species). Functional equivalence requires that constructs have similar nomological properties across groups (i.e., same predictors, consequences, and correlates). Conceptual equivalence is conventionally established through a process of building a theoretical consensus, whereas establishing functional equivalence involves statistical strategies that aim to identify common patterns of associations between constructs and their nomological properties across groups (or species). Item equivalence and scalar equivalence can only be considered for constructs that are conceptually and functionally equivalent. Item equivalence refers to the instruments used to assess a construct and their goodness of fit for that construct. Finally, scalar equivalence requires that the same instrument yields similar results when used in different groups. Item and scalar equivalence can be assessed by a variety of methods including reliability coefficients, examination of the internal structure of an instrument, measurement invariance across groups, or using tools from item response theory.

For example, attachment is a construct that refers to a child's relationships to their social partners and their embeddedness in their social world (e.g., Keller and Chaudhary 2017). It is crucial for a child's development of trust—both in themselves as well as in others—and sense of self. Historically, research on attachment has focused on Western middle-class families, often described as WEIRD (western, educated, industrialized, rich, democratic; Henrich et al. 2010). In these contexts, attachment typically unfolds within the framework of a nuclear family, where there is usually one primary caregiver, often the mother, engaging in exclusively dyadic interactions with the child. These interactions, characterized by distal communication such as face-to-face interaction, language, and play with toys, are structured to foster psychological autonomy and self-consciousness in the child from an early age (Keller 2021).

This WEIRD perspective does not represent, however, the diverse nature of attachment across different sociocultural contexts (Henrich et al. 2010; Keller and Bard 2017). In many non-WEIRD societies, including traditional farming, hunter-gatherer, and fishing communities, childcare involves a more extensive network of caregivers, which may include up to 20 people, both related and unrelated (Keller 2021). The mother, while often a central figure, may be one among many caregivers or even play a marginal role. In these settings, children's interactions are mainly proximal, involving bodily based communication processes emphasizing rhythm and synchrony. These societies prioritize the development of a communal self, teaching children to be integral and responsible members of their community, and often have hierarchical social structures that influence communication and interaction rules. This contrasts sharply with the WEIRD model of fostering individual autonomy and self-reliance (Keller and Chaudhary 2017; Morelli et al. 2017).

Digital ethology, with its potential for analyzing large-scale digital data capturing a wide array of behaviors, offers a unique opportunity to examine how the construct of attachment is expressed and understood differently across cultures. By exploring behavior patterns in digital communication, digital ethology can reveal how attachment and socialization strategies are expressed across various cultures. This approach can also be relevant in examining the formation and expression of multiple cultural identities, especially in a globalized world where migration plays pivotal roles (Garcia Coll and Marks 2011). Nevertheless, it is essential to be aware of the potential for an even more narrow bias toward the "digital WEIRD" subpopulation in digital ethology (i.e., the part of the WEIRD population that is accustomed to digital technologies). This means ensuring that digital ethology does not simply reinforce the attachment models based on research conducted on Western societies, but instead captures the rich diversity of attachment expressions globally. To obtain a more representative and comprehensive understanding of global behaviors, it is imperative to analyze digital interactions not only through the lens of advanced technologies prevalent in Western societies (e.g., expensive smartphones) but also through technologies and platforms used widely in non-WEIRD contexts. This includes focusing on more popular tools in developing countries (e.g., affordable mobile models) and exploring messaging apps and social platforms that are available as globally as possible (e.g., apps that are avoiding censorship). Thoughtfully applying the framework proposed by Hui and Triandis (1985) becomes particularly relevant in this context. This framework emphasizes ensuring that the construct of attachment is inclusively and consistently defined across cultures (conceptual equivalence), as its meaning can vary significantly. It is crucial for researchers to also verify that the role and significance of attachment behaviors are comparable across different groups (functional equivalence). This includes adapting measurement tools, like questionnaires or digital analysis algorithms (item equivalence), to suit each cultural context and ensuring these tools yield consistent results (scalar equivalence) across various cultures. This approach, especially challenging in digital ethology due to the diversity of online platforms and communication styles, demands careful construction, adaptation, and validation of research methods. This process allows researchers to draw more reliable conclusions, recognizing the richness of cultural variations while maintaining scientific rigor and comparability of data.

Conclusion

Digital ethology is grounded in the established core methods of observation and knowledge extraction of traditional ethology yet it faces burgeoning challenges associated with large-scale digital data. Major challenges are associated with causality, especially when humans are bidirectionally coupled to their environment: "…enough people participating in an individual activity can result in structural change and vice versa" (Lovasi et al., this volume, p. 33). This becomes obvious when certain behaviors have no meaning at the individual level (e.g., Gini index or synchronization phenomena). Thus, at the methodological level, we need to develop "collective" ethograms and mathematical tools to account properly for these niche constructions at ecological and social levels (Krakauer et al. 2020). In addition, at the legal and ethical level, we should keep in mind that data ownership can go beyond individuals, for instance, in the case of Indigenous communities where communal structures override individual claims.

Acknowledgments

We are thankful to the organizers of the Ernst Strüngmann Forum and all our colleagues for the pleasant and stimulating "mind jazz" over the week. We especially thank the Forum participants who took part in our group's discussions: Beate Ritz, Claudia Bauzer Medeiros, Fabio Kon, Gina S. Lovasi, Henning Tiemeier, Hye-Chung Kum, José Balsa-Barreiro, Michael Brauer, Michele C. Weigle, Nitesh V. Chawla, Steven Bedrick, and Sven Sandin. GD wants to thank Alexandra Glavaschi for providing a high-definition scan of the original Tinbergen's drawing in Figure 2.1.

Appendix 2.1: Glossary

Ascertainment bias, differential recording of outcomes or imbalance screening for outcomes among exposed individuals compared to unexposed individuals.

Convergent validity, often measured by applying different tests and observational methods that intend to measure the same construct with the same individual or groups of individuals and test the consistency or interrelationship.

Discriminant validity assesses how much tests/other methods that are not intended to measure the construct in question, deviates/differs from assessments intended to measure the construct. Reliability and validity belong together.

Dissemination area, the smallest standard geographic area for which all census data are disseminated, usually a small area composed of one or more neighboring dissemination blocks (400–800 inhabitants).

Internal consistency means that individuals/groups respond consistently across items measuring the same construct. If you have a questionnaire measuring one construct, you can, for example, split the items and correlate the two sets. Challenges are, for example, the quality of formulation and preciseness of the items and the extent to which they measure the construct.

Inter-rater reliability in which two trained raters observe the same situation, or the same videotape. Their agreement is statistically assessed, most simply in percentage, more usually with a Cohen's Kappa coefficient.

Outcome (variable) is an event or metric that captures a construct or a predicted behavior. It is measured as categorical (nonparametric statistics), ordinal (nonparametric statistics), or continuous (parametric statistics) values.

Reflexivity means sensitivity to the ways in which the researcher and the research process have shaped the collected data, including the role of prior assumptions and experience, which can influence even the most avowedly inductive inquiries. Personal and intellectual biases need to be made plain at the outset of any research reports to enhance the credibility of the findings.

Reliability refers to the consistency of a measurement. Three types of consistency are usually considered: over time (test–retest reliability), across items (internal consistency), and across different observers/coders (inter-rater reliability).

Respondent validation, or "member checking," includes techniques in which the investigator's account is compared with those of the research subjects to establish the level of correspondence between the two sets. Participants' reactions to the analyses are then incorporated into the study findings.

Sampling Frame:

Test–retest reliability means measuring the same construct/variable at two different points in time on the same individual or group of individuals and testing the correlation of the two measurements. One challenge in this method is the potential for learning effects; for example, if the same items are used, participants might remember their previous responses, which can influence the consistency of the construct over time.

Triangulation compares the results from either two or more different methods of data collection (e.g., interviews and observation) or, more simply, two or more data sources (e.g., interviews with members of different interest groups). The researcher looks for patterns of convergence to develop or corroborate an overall interpretation.

Validity refers to the extent to which a measure represents the variable or construct intended to measure. There are also different kinds and different ways to define validity, most often it is convergent and discriminant validity.

3

Paths to Public Benefit

Constructing Meaning from Our Physical and Built Environments through Digital Observation

Gina S. Lovasi, Steven Bedrick, Michael Brauer,
Megan Doerr, Fabio Kon, Lindsey Smith, and Beate Ritz

Abstract

Digital data can be used to observe human behavior as well as aspects of the physical, built, and natural environment that provide context for such behaviors. Data extracted from communities through surveillance have rightfully been the subject of concern, yet such data hold great potential for benefits, including knowledge generation and dissemination to advance human health and equity. Benefits will depend on what is measured and who sets the agenda. Here, ways to organize available and future physical, built, and natural environment measures are discussed, and approaches are proposed to guide the use of such data to generate knowledge while keeping in mind varied value judgments and goals. Metadata are identified as a key tool to deter misrepresentation and misuse of data. To serve this purpose, metadata could be expanded in several ways, including historical context and intent of data collection as well as limitations and permissions to be aware of while planning use and interpreting findings. As data are used, subsequent versions of metadata could record information to inform future use, including a statement of social license updated as the individuals and communities affected by use of the data reflect on harms and benefits. The process of seeking social license for use of geographically referenced data itself has potential to add to our understanding of human agency and to inform ethical inquiry about the structural determinants and individual choices that play out in communities. Opportunities to fill gaps and meet future challenges are identified. Further, attention must be given to incentives across the funding, publishing, and institutional landscape so that envisioned change can be realized and sustained.

Introduction: Physical, Natural, and Built Environment Measurement for Digital Ethology

We are living amidst a revolution of geospatial data generation and use. Such data have the power to be transformative by improving our understanding of the physical, natural, and built environment and benefiting the public and individuals through valued outcomes such as health. Here, we consider the potential of place-based data within the emerging interdisciplinary field of digital ethology, which brings a multimodal perspective to the potential for accumulating data to describe and explain the bidirectional relationship between human behavior and its geospatial context (see Paus, this volume).

Implications of Accumulating Place-Based Digital Data

As we live our lives, data accumulate in records that are increasingly in a digital format. When we access our phones, we generate vast amounts of behavioral data, much of which can be anchored to our location at the moment and our recurring travel patterns. Further, we may add sensors to our homes to detect water leaks or other disturbances, and municipalities and governments digitally monitor and report on air quality and temperature. Individually, we benefit when we use location, satellite imagery, and real-time traffic congestion data to navigate to a restaurant or clinic. To attain these benefits efficiently, we may agree to monitoring of our mobility as a part of traffic density surveillance, which is then made available for broader use that extends far beyond our own planning. Stored imagery or video footage of public spaces, such as that recorded for security-related purposes, could additionally be used by researchers to study human behavior in daily life, as highlighted by Pallante et al. (this volume). Thus, knowledge generation[1] goals may be among uses that extend beyond those originally envisioned in planning or permitting digital data collection. Such research applications could use geospatial data resulting from digital surveillance for the common good; there is also a need, however, to manage and mitigate potential harm.

Potential Harms

While digital surveillance is an increasing and nearly ubiquitous reality,[2] digital surveillance has negative connotations due to known, suspected, and

[1] In line with Kum et al. (this volume), we view knowledge as being created from data and usable to inform action. Many steps and much potential for missteps lie along the path from data to knowledge to action.

[2] Some applications of geolocated data are designed for public health surveillance purposes, such as to monitor infectious disease outbreaks, as highlighted by Sarker (this volume). Here, we include not only passive methods that capture information about people, but also those that capture spatiotemporal variation in the spaces inhabited, traversed, or otherwise used by people.

feared uses and abuses. Many possible uses are not anticipated or may not be welcomed by the individuals whose data are assembled. For example, commercially marketed cell phone–location data have been used by police during criminal investigations (Burke and Dearen 2022) and can reveal presence at sites where sensitive medical services are provided (e.g., reproductive, mental, or behavioral health clinics) (Fair 2022). These present uses add to historic precedent to substantiate concerns that place-based digital data can reify and exacerbate systemic inequities and power imbalances. Harmful use of digital place-based data can include restriction of individual and collective rights.[3] Other costs to individuals and communities may result from commercialization of these data in ways that are misaligned with or undermine advances toward equity. Documented harms from use and abuse of digital data, even if well-intended, provoke questions about the legitimacy and acceptance of digital observation and resultant data use.

To mitigate harms that can arise from digital surveillance, strategies that increase transparency and limit abuse potential are required. In some instances, the risk of harm or lack of consent may be most appropriately addressed by not accumulating data. Where digital data can be ethically collected, however, their use should benefit the individuals and communities whose surroundings and activities are represented, such as through remediation of environmental harms that undermine health. Here, we frame this as *using geospatial data for good*, while recognizing that notions of "good" are highly subjective. Proactively thinking in these terms frames our obligation to produce public benefits while averting harm. Further, it highlights the need to include and amplify the voices of the communities who contribute to the data from the outset. Finally, investment in dissemination and translation is needed so that observation and knowledge generation can contribute to communities' data-informed advocacy and action.

This chapter distills our multifaceted discussions from the Ernst Strüngmann Forum in July 2022. During this week-long immersive event, we put forward a vision to advance scientific and societal benefits made possible by assembling digital data on the physical, natural, and built environment. To do so we identified types of data to be included, implications of sharing access to and power over such data, and strategies for creating and disseminating knowledge with attention to challenges specific to spatial data and to the values and needs of communities represented in this data. Before concluding the chapter, we highlight opportunities for team formation, cross-disciplinary training, and ways to shape our funding allocation, publication, and institutional incentives to support sustained progress toward our vision.

[3] Here we are referring to rights, such as the right to life and liberty, but also note that the right to privacy is closely connected to concerns raised about digital surveillance. For further reading, see Chapters 10, 11, and 12 (this volume); for an overview of how human rights could inform ethical work with big data, see Mantelero (2018).

Types of Digital Data on Physical, Built, and Natural Environments

As noted by Smith (this volume), multiple existing data sources capture aspects of what is present in the environment (e.g., land cover such as pavement), how it is used (e.g., parking, playground), and quantitative characteristics that vary spatially (e.g., surface temperature, air pollutant concentration, annual precipitation, daily average sound levels). The lens of digital ethology suggests making human habitual behavior central to our typology of environmental measurement.

Human-Centric Quest for Measurement

For the purposes of this chapter, we chose a human-centric approach to classifying measures of the environment. Our emphasis is on public benefits and harms, where the humans who make up this public have lived experience expertise and value perspectives that need to be considered. This should not be interpreted as the only lens through which one can view potential global benefits of digital geospatial data; alternatives may emphasize aspects of the biosphere affecting multiple species. Here, we identify that a human-centric approach can bring attention to the following questions:

- Fitness for fulfilling human needs: How fit is the environment for fulfillment of human needs? In what ways does the environment create opportunities from the most fundamental (e.g., breathing clean air) to the most aspirational (e.g., artistic expression, co-creation of knowledge)?
- Suitability to how a place is actually being used: How fit is the environment for the currently enacted or desired use[4] by the community? What features may enhance uses for which a place was designed? What features contribute to unintended side effects, including those that arise from the mismatch between originally intended and current de facto uses? What constructs relate to fitness for the current de facto or emergent proposed use, such as livability, walkability, or accessibility?
- Design and redesign to encourage intended uses: What immediate- and long-term uses were deliberately accommodated or discouraged as the environment was built and rebuilt over time? Do we have direct accounts of the intentions[5] (e.g., oral history, transcribed discussion at

[4] Uses of the surrounding built environment range widely, including the acquisition of food and other goods and services, mobility and physical activity, and social interactions from casual greetings to building collective identity and action (see Weigle et al., this volume, on social environment).

[5] Intentions might, for example, be revealed by noting exclusive attention to private vehicle use in a planning document for gridded streets.

planning meetings or public hearings, archived documents, legislation) or can intentions be inferred based on the specific features present (e.g., hostile architecture to deter homeless encampments, loitering, or skateboarding (Petty 2016)?

These questions help to organize existing measures and can also lead us toward what additional data are required as new uses of environments are initiated or proposed, or as needs are newly articulated by communities.

Notably, the intentions of those who design and the needs of communities who use the environment are brought into the foreground. These intentions can be mutually informed, as emphasized by architects such as the Brazilian landscape designer Roberto Burle Marx who revisited design decisions after actual use has been observed (Montero and Marx 2001).[6] For example, Burle Marx maintained that the paths in newly opened public gardens should be formalized only a year after the space becomes available, reflecting the footpaths created by frequent community use (i.e., those routes through the space that have been demonstrated to be convenient and useful). This attention to emerging use can be applicable even in cities with a long history of human habitation and built environment change. There is also the possibility for observation across domains, measurement scales and time periods, and across emerging frontiers of measurement to inspire entirely new questions as we wonder about ways in which humans respond to the built environment "in the wild."[7]

Domains: What to Measure That Is Relevant to Human Needs and Uses of the Environment

As we explore digital data related to the lived environment, we find ourselves encountering a wide variety of domains of data, situated at varied levels of resolution and abstraction.

Beginning with impediments to foundational needs such as breathing clean air and sustaining thermal comfort, we may first consider data describing atmospheric properties of the physical environment (e.g., particulate concentration, humidity). Topological characteristics and type of land cover may affect these properties, along with how suited the landscape is for providing nourishment and shelter, what resources can be accessed, and what uses the spaces may support. Beyond describing the places used for housing, work, and leisure,

[6] Other practices applied by Burle Marx in his work in Brazil may have relevance to natural features integrated into the built environment (e.g., specifying that gardens should prioritize native species and taking into consideration the preexisting natural and physical landscape). While on the surface these may not seem crucial to a human-centric approach, the perspectives of Indigenous peoples may bring further attention to these and other aspects of how we build.

[7] The phrase "in the wild" is used here to convey that these are not settings artificially contrived to manipulate human behavior for research purposes, as might be seen in a laboratory setting. Humans in their current habitat largely means humans surrounded by structures and urban spaces built by and for humans.

geographic data can incorporate notions of secure tenure (ownership), safety, and private or restricted use spaces.

Going beyond physical attributes of terrain (i.e., topology, geology) to consider fitness for intended uses requires distinct measurement approaches even when situated at a similar geographic scale. Remote sensing and stationary sensors are especially valuable for visible environment measurement, including the presence of buildings and transportation-related structures. In contrast, administrative and participatory digital data collection approaches are often needed to capture aspects of the built environment[8] that relate to intended and actual use over time (e.g., availability and accessibility of health-care delivery or food establishments). Notably, quantities such as auditory noise may be operationalized via relatively objective measurements of physical properties at a particular point in space and time (i.e., ambient decibel level), yet whether a given decibel level is perceived as *unwelcome* noise can depend on the source, the listener, and the surrounding context. Many measures that relate to fitness for use are quite complex and inherently subjective, such as a "walkability score" (Wang and Yang 2019), which may be computed in any number of different ways and often relies on combining multiple sources of data. Developing and agreeing on methods of measurement is critical for deriving value from geospatial data in terms of how these data relate to human–environment interactions. Clarity about what to measure is a prerequisite to selection of relevant data sources,[9] and also to noting limitations specific to the task at hand. Going beyond methodological limitations, it is also important to explicitly examine and document sources of bias within one's data and choice of measures.

Further enhancing our understanding of the environment, we may consider data representing (and possibly directly generated by) discrete and ongoing human activities, including "sensor data."[10] This could include readings from traffic counters with relevance to mobility and vehicle emissions, as well as data sources providing insight into how people feel or act in a given space, such as geotagged social media posts. These data types may illuminate barriers to realizing benefits of intended land use. For example, two otherwise similar parks may be quite different with respect to physical and mental health benefits due to differences in surrounding vehicle traffic and associated noise, air pollution, and injury hazards.

[8] We define the built environment as including human-built or modified structures, transportation systems, and features such as buildings, roads, plazas, and parks as well as fixed features such as fire hydrants and light posts.

[9] The proposed use will determine whether available data are sufficiently relevant, and correspondence between what we aspire to measure and what we have represented in our data is never perfect.

[10] This is intended broadly to include not only stationary sensors deployed for purposes of measurement but also device-based data such as accelerometry and geolocation data generated as people carry cell phones throughout their activity space.

There is a feedback relationship between the design and use of the environment; enough people participating in an individual activity can result in structural change, and vice versa. Further, observational data generation and subsequent knowledge generation can make ongoing use of a space more evident, and awareness of how a space is used can itself change use (e.g., people changing their behavior or chosen route in response to the presence of a cycle-counting monitor) or can bolster the case for sustained investment to facilitate use (e.g., monitoring the number of cyclists following the development of protected bicycle lanes can be used to make the case to maintain and scale up such protections).

The same physical feature may simultaneously span multiple domains categorized based on type of human use. For example, a bus stop could be both relevant to current community use for mobility as well as providing for rest or shelter because of the presence of a bench. Likewise, a mixed-use building including ground-floor retail and apartments may play a role in both the food environment and walkability at the neighborhood scale.[11]

Variable Scale and Timing Require Attention to Human-Drawn (and Redrawn) Boundaries

Data representing features of the physical and built environment range in scale in terms of spatial and temporal resolution, density, and precision.

Advantages of digital data include its volume and frequency—for example, measures that capture seasonal and even hourly fluctuations in air quality. Some digital data can be archived and later processed and transformed to limit the uncertainty due to temporal gaps in an observation series.

At a given time point, geographic space is divided into units of observation in ways that may align with how they are designed or used, ranging from a simple grid to human-drawn administrative units, including parcels, zoning areas, and plots of variable shape and size.

A challenge in digitally derived environment data is posed by human-drawn boundaries and features (rather than those that are naturally occurring and enduring). Human-drawn boundaries have different social, economic, and political functions, and are commonly used in research relying on geographic information systems (GIS), as described by Smith (this volume). Our organization and depiction of such boundaries benefit from a notion of spatial hierarchy, yet the spatial nesting of smaller areas within larger ones may be imperfect. In some scenarios, these hierarchies may be complex, and involve plural,

[11] Neighborhoods have been variously defined to include activity spaces frequently visited, areas important to resident identity, or the postal and other administrative units that provide a convenient but imperfect operationalization of neighborhoods (Lovasi et al. 2012). Together, neighborhoods contribute to larger geographic contexts such as the city-level patterns that connect physical and social environments; for further discussion, see Balsa-Barreiro and Menendez (this volume).

non-overlapping, and highly irregular attributes. Boundaries may be closely tied to elements of physical geography or existing infrastructure, such as bodies of water or utility and sewage networks. Historic processes shaping delineation may themselves be harmful, as in the case of gerrymandering (Sánchez 2018) or municipal fragmentation (André Hutson et al. 2012). Understanding the origin of these boundaries may have implications for contemporary use, including efforts to explain how various land uses arose and changed over time. For example, analyses concerned with equity and resource distribution benefit from use of historical information about boundaries such as those associated with *redlining* in the United States, which determined unequal access to loans and housing by race (Rothstein 2017).

Human-drawn boundaries may be driven by power or bias and are subject to challenge or overthrow. The built spaces marked by these changing boundaries may respond incrementally or suddenly.[12] An example of the latter can be noted in the city in which our Forum discussions to conceptualize this chapter took place: Frankfurt am Main, Germany. Frankfurt's old town underwent substantial physical rebuilding and administrative changes after bombing during World War II had destroyed most of its physical infrastructure (Lehné et al. 2013). Changing boundaries may occur in response to population growth or migration, which requires attention when working with longitudinal population characteristics based on census boundaries (Logan et al. 2014). These shifts can pose a challenge to systems of data management, knowledge representation, and statistical analysis.

Contextualizing Imposed Labels and Current Practices: A Case for Increasingly Inclusive Teams

Digital data that we have access to or envision to create arise from a legacy of geospatial work. Further, those engaged in generating and using environmental data to explain human behavior and health are influenced by our training to think of the world as compartmental and to formulate questions according to our specific professional lens as well as other aspects of identity. This can impede the match between community needs and what is measured about the environment. For example, within mobility research, amenities and services may not be equally matched to the needs of all demographic groups, and in particular, accessible toilets and benches that are critical mobility determinants for seniors have not been routinely captured in walkability measures. Thus, collaborations inclusive of perspectives across demographic categories such as age may yield new insights even for commonly addressed topics.

Provenance and identification of those who should set the agenda for measurement of and changing use of a space (e.g., public vs. private control) may

[12] In addition, the location and nature of boundaries can, of course, be disputed between groups of people or organizations, adding an additional layer of complexity.

not be easily established. The concepts and definitions discussed here and by Smith (this volume) are illustrative of measures commonly encountered in urban spaces in western, educated, industrialized, rich, democratic (WEIRD) countries where the systematic collection of data from the physical and built environment started several decades ago. Yet, the vast majority of humanity does not live in areas that have data availability typical of WEIRD countries. As we consider a truly global research agenda for digital data about the environment, collaboration with a broader cross section of researchers, communities, and policy makers from around the world will be essential.

Future research teams may include perspectives we are missing and may accordingly judge some ways of categorizing or labeling domains of environment measurement to be inappropriate. It is particularly important to recognize the need for bridging to work on topics of importance to equity, such as housing instability, with research in understudied parts of our human habitat. For example, Weinstein pointed out the narrow view of North American scholars on the topic of evictions and wrote an article on reconceptualizing housing insecurity by looking at the work carried out by scholars in India and South Africa on urban "slum" evictions (Weinstein 2021). The field will benefit from critically appraising current practice, assessing ways in which our categorization of data is or is not appropriate to other research scenarios, and articulating additional concepts that need to be developed.

Expanding Frontiers of Environment Measurement

Digital observation may open the door to research efforts, collaborations, and exchanges that cross national boundaries. Some types of data are already collected globally (e.g., Landsat, which collects satellite imagery from the entire Earth). Such data can now also be used with tools such as artificial intelligence algorithms in innovative ways (e.g., remote sensing images from the Amazon Rainforest to detect deforestation areas with the help of machine learning and citizen science) (Dallaqua et al. 2021).

Some types of data offer flexibility in generating data categories and constructs, and we note that imagery is one such type of data. Imagery can be used to capture pre-determined features and to enable future uses not envisioned at the start of data collection. For example, at present digital data to characterize quality and use of indoor environments is limited, even though these are the environments where most people around the world spend most of their time. Potential indoor environment data sources include indirect information derived from exterior imagery (e.g., building structure and details visible through remote sensing or façade features from street-level imagery) as well as imagery that more directly shows the indoor environment, but which may not represent the typical condition of that environment over time (e.g., from online real estate resources which include indoor images). Importantly, despite the flexibility of working with imagery, challenges arise due to measurement that is inferred, available only for a biased sample of places or times, or unreliable such as due to variation in weather, lighting conditions, obstructions, or other temporal aspects that may affect observation (e.g., image capture of a street before or after trash collection). Further, when using human raters of imagery to capture information such as perceived safety, there is a risk of embedding into resultant metrics any salient human biases, such as an implicit association of racial composition of a neighborhood with perceived safety. Preferences and perceptions differ, which makes an inclusive research team composition and practices like community consultation valuable in understanding what is being observed in digital data. Relevant to safety perceptions and equity, for example, experiences of over-policing may result in divergent responses by race to police presence.

Alongside digital datasets about the built and physical environment, spatially referenced human reactions to events can be captured, particularly through data sources like social media, as discussed by Sarker (this volume).[13] Social media can capture conscious reactions to physical features or associated construction efforts, possibly leading to behavior change or public demands. In contrast, users of a space may not be able to sense air pollution or notice resultant cumulative health effects, and therefore reactions to unseen or gradually harmful exposures are unlikely to be captured in social media posts. Novel insights and innovations may be facilitated by the increasing use of social media as a source of data, including insights into the perspectives of geographically delimited communities and other social or professional groups. Representativeness of such data must be considered, however, as different social media platforms may have greater affinity from particular user groups while other parts of society may be entirely excluded.

Beyond what data are presently recorded or monitored describing our built and physical environment, it is important to be aware of what *is not* being measured. Even where certain aspects of the physical or built environment are currently challenging to measure, determining that something is worth measuring or sensing digitally has the potential to drive down costs of data acquisition, as has been the case with the cost of remote sensing imagery.

As a metaphor, we find it helpful to think of the digital measures that are currently in common use for understanding the environment as those found "under the lamp post." As data needs are articulated and the range of domains covered by available geospatial data broadens, we will expand and spread the light of the lamp post and increasingly be able to see what has until now been hidden. In full awareness of our current imperfect vision of what is possible, we endeavor to provide ideas and questions about how emerging frontiers of

[13] It must be remembered that social media data introduces issues of sampling bias; for example, a dataset comprised of geolocated Twitter/X posts will underrepresent voices of older users or of users without smartphones. We discuss this issue in detail later in this chapter.

data generation could fit with previously used data sources. In doing so we aspire to catalyze continued conversation and elaboration by others.

Who Has Input and Access to Environmental Data and Metadata from Digital Surveillance?

In this section, we consider how we can improve access to digital data toward an overall goal of "data for public good." In doing so, we consider that with improved technology there may be opportunities to measure previously understudied aspects of the environment.

In considering who has access to data, there are a number of existing constraints. Not all data can be shared without relevant security or legal clearance. For example, some imagery is classified (collected for military purposes) or when released obscures specific features. Data may come at a financial cost or require payment for transformations needed to make it ready for use. Storage systems could create barriers to access or pose additional costs. Some data may only be available for a limited period of time, either after an embargo period or before it must be deleted. Of course, beyond data access, appropriate and informed use of data requires understanding the underlying methodology and purpose, making metadata invaluable.

Metadata Wishlist

As noted by Miller (2022), metadata is data about data, taking the form of structured statements that inform efforts to organize, describe, locate, index, structure, navigate, and manage data resources. Metadata creation and contribution of metadata to repositories are important ways to increase responsible use of digital data (Leipzig et al. 2021). Both those sharing and accessing data and data repositories (e.g., Dataverse; King 2007) will benefit from the skills of data governance experts and data librarians (Lagoze et al. 2006).

Some novel aspects of metadata that we propose below go beyond fixed technical specifications and may need updates subsequent to initial data dissemination. This means that a system that handles versioning is needed, perhaps building on practices developed for GitHub (Crystal-Ornelas et al. 2021).

We note that the data versus metadata distinction can seem arbitrary and, in fact, the same observations may both be represented as data in one database and be summarized in the metadata for a different but spatially overlapping database. The use as data or metadata will depend on the specifics of any given analytical or data management scenario.

Metadata about Original Purpose for Data Collection

Some recontextualization of data can be achieved when including information about the original data collection purpose within metadata. For example, Google imagery and maps have become useful tools for the characterization of the built environment for health research (Rzotkiewicz et al. 2018). Google Street View (Gallo and Kettani 2020) had a primary purpose of improving the spatial and temporal accuracy of Google Maps, for purposes which included identifying commercial locations and increasing advertising revenue. As a consequence, derived data based on these private sector efforts are expected to represent retail settings more accurately than other aspects of the environment such as bike routes. Commensurate with its primary purpose, the image availability and recency vary systematically with socioeconomic conditions (Fry et al. 2020). Researchers may, however, use Google Maps/Google Street View for efficient characterization of the environment at a scale that would not be feasible using field audits.[14]

Other examples in environment measurement likewise benefit from understanding the original purpose and potential for blind spots and bias in the data. Differing susceptibility to bias based on origin can be articulated even among data sources in a similar domain, such as traffic counts computed by a city's bureau of transportation as compared with user-contributed data for smartphone-derived traffic apps such as TomTom or Waze. Whereas a transportation bureau may collect data for meeting reporting requirements or informing intersection changes to improve safety, traffic apps are likely seeking to increase user engagement and associated revenue. Data users could be more cognizant of the data origins and differences in sampling density if these are routinely contained in metadata.

Privatization of data generation intensifies the need for metadata to highlight the reasons for data collection and the related implications for their secondary use. Potential biases, blind spots, or inconsistency may arise related to the original commercial purpose motivating data generation.

Public Open Data projects (e.g., Open Street Map) where "the community" can upload and update data are an alternative that is commonly used in research, especially in locations where government or private sector data may not exist, are not trusted, or lack granularity. In working with such community-generated data, users should be aware of ongoing updates and gaps based on data provider capacity or interest in specific locations (e.g., locations with higher proportions of populations with technical GIS proficiency may have more detailed information; points of interest to specific groups, such as caregivers, may be underrepresented).

Metadata Relevant to Generalizability: Incorporating Structured Information about Communities

Metadata illuminate how data are viewed from multiple perspectives (Lagoze 2001), including attention to the communities represented or omitted.

[14] For example, de Macedo Oliveira and Hirata Jr. developed a system that analyzes thousands of Google Street View images with machine learning to investigate the greenery in a megalopolis like São Paulo (de Macedo Oliveira and Hirata Jr. 2021).

Omission can be the result of structural racism, marginalization, and related social processes that the data creator may not acknowledge or endorse. Thus, a structured requirement for attention to representativeness within the metadata itself is useful, especially if accompanied by an inclusive process. Multiple perspectives can allow a team to draft more robustly and update metadata, documenting a range of cautions to consider when generalizing to a larger set of individuals or geographic areas.

When a community is not represented in data on the environment, there will be missed opportunities to inform decision making (e.g., due to insufficient attention to hazards in the environment or incorrect attribution of harmful effects to the wrong cause), potentially resulting in a community missing out on beneficial place-based or policy changes as a result. As an example where errors in attribution could result in missed benefits, consider how a focus on physical signs of disinvestment could be interpreted as supporting different action strategies. One response might involve attending to the visible signs (e.g., by fixing broken windows or planting trees in deprived neighborhoods); however, even if appreciated by residents, this may fall short of enduring change if the underlying cause is not identified. Alternatively, the underlying disinvestment could be addressed more directly, such as through investment in education or job creation in the same neighborhoods to foster social mobility. Thus, attributing any observed harm to what is proximal and visible risks superficial action; that is not only ineffective, it also diverts attention from alternative actions responsive to the underlying cause and with greater potential for enduring benefits.

Metadata that incorporate a structured ontology (Norris et al. 2019) for social context could help researchers delimit their findings by identifying populations that were entirely or disproportionately excluded. This would facilitate systematic efforts to describe and fill gaps in the availability of actionable knowledge that result from historical and present inequity.

Metadata about Data Sharing and Social License

Data access and sharing practices can also be highlighted in metadata, for example, as described by the Data Use Ontology standard (Lawson et al. 2021). Crucially, this can include who can access data and potentially also how the data have been used over time. Through data reuse ("secondary use") of large-scale data, community data may become divorced from their source context. Structured approaches are needed to reconnect datasets to their originating and dynamic social context. We propose that this can be achieved by bringing community voice alongside application of established guidelines, such as Maelstrom for data harmonization (Eva et al. 2022).

How a dataset is used is subject to change over time, requiring updates to information about how it has been or could be used. Such information includes who has used the data and how data transformations and linkages have been made or could be made. Such metadata would bring users' attention to any distinction between the data in circulation and aggregated or enriched data available upon request. For example, food environment data from the Canadian Urban Environmental Health Research Consortium (CANUE) repository is being released to general users at a higher level of aggregation than the version released to approved research teams (Doiron et al. 2018).

A consideration relevant to stewardship of data is social license, defined as the acceptance granted by a given community or public to a company or organization for a particular activity. Social license could both be described in metadata and seen as a prerequisite to using data about the physical, natural, and built environment. One example of a deliberative process leading to documented social license is the vast network of CCTV cameras in the United Kingdom. These cameras generate data about the environment and human interactions with and within those environments. The use of CCTV for surveillance is considered an extension of the principles of "policing by consent" established in 1829 (GOV.UK 2012a). Permitted use of the resulting data is formalized through the Protection of Freedoms Act (GOV.UK 2012b) which includes the Surveillance Camera Code of Practice. This legislation established the Surveillance Camera Commissioner (updated in 2022 to the Biometrics and Surveillance Camera Commissioner) as an authority to guide the use of this technology for one of the most visually surveilled countries in the world. Systematic attention to social license as metadata is created and shared could promote communication among users and with the original community or its descendants, and also capture efforts over time to reconfirm or revise the agreed terms.

Metadata about Other Data Limitations

Metadata should be designed to include aspects relevant to understanding and communicating data limitations, such as coarseness of the data that potentially masks important variation. Thus, metadata should note quantifiable sources of error and uncertainty. A critical component of data is the characterization of measurement uncertainty (as distinct from true observable variability). Uncertainty may be due to the quality of the measurement itself and may also arise due to sampling error (e.g., gaps in spatial and temporal sampling). No measurement is exact, but measurements may be compared against some practical benchmark or reference value.

Attention to error and uncertainty can aid in not only articulation of limitations but also harmonization and triangulation with other sources. For example, a current measurement with an improved spatial resolution (measurement A) could be combined with a historical measurement with more coarse resolution (measurement B) using comparative analyses (e.g., by linear regression) which itself has some uncertainty (MacEachren et al. 2012). This uncertainty should be propagated together with the uncertainty of the initial measurements. This permits all available data to be used in a way that reflects the reduced certainty of estimated values as compared with measured values. Likewise, interpolation (e.g., filling in missing data that is within the spatial or temporal bounds of the measured data) is another process where uncertainty should be propagated based on both the original measurements and their modeled relationships.

Using Data

Once data are assembled and access is being provided alongside metadata, further steps allow data to be used to generate knowledge and benefits. Importantly, even before turning to strategies for dissemination, we consider how to make GIS-informed knowledge replicable and reproducible (Peng and Hicks 2021). Key steps include integration, analysis, and interpretation refined through multiple perspectives.

Integration

Across disciplines, good practices are needed for data stewardship (Wilson et al. 2017a), including planning for data storage (Hart et al. 2016). Errors are caught and transparency of algorithms improved through practices such as code review (Vable et al. 2021) and code sharing (Peng and Hicks 2021). We note, however, that code that only works on a transformed dataset is not sufficient to allow for external verification, even if the raw data are publicly available; sharing the code (or at least a narrative) that details steps involved in the data transformation and integration can limit redundant work or use of unnecessarily flawed data.

Two major approaches can be adopted when dealing with large amounts of data from different sources, differentiated by whether transformations are done up front or later as needed.

First, a *data warehouse* (Vaisman and Zimányi 2014) is a very large, highly structured database built by extracting, transforming, and loading data from its original sources. The warehouse can be updated periodically through an automated process. Metadata are included. Transformation may include spatial, temporal, and semantic alignment. Data warehouses are designed to enable their users to perform analytical queries (e.g., summarizing data, computing aggregate measures). As such, designers consider the specific analytical needs that the warehouse will support, embedding aspects of their knowledge and intention into the resulting data warehouse design.

In contrast, a *data lake* (Gorelik 2019) is a repository of heterogeneous data collected from multiple sources and stored in its original, raw format. A data lake typically holds a huge amount of data, in a similarly huge variety of different formats. A key advantage of a data lake is that new types of data can be added quickly and with minimal effort. The trade-off, of course, is that the end user of a data lake will need more work to harmonize, format, or link data before starting analyses than is typical for users of a data warehouse. The greater control that the user has over decisions on how to transform or link data may have advantages, however, especially if the analytical needs differ greatly between data users.

Incentive structures that encourage or impede data integration are themselves considered by Balsa-Barreiro and Menendez (this volume), including factors that influence the perceived opportunity costs and benefits (both intrinsic and extrinsic).

Analysis

Emulating study designs that can provide a strong basis for causal inference and pre-specification of analysis plans are among practices whose benefits have been articulated elsewhere (see Dumas et al. and Medeiros et al., this volume). As such, we acknowledge these but focus mainly on challenges related to interdisciplinary collaborations and place-based analyses typically encountered in work with geographically referenced data relevant to environmental constraints on human well-being and behavior.

Expectations for rigor and transparency (such as use of code sharing platforms like GitHub) vary across fields, and collaboration with computer science researchers from an early planning phase can help to ensure adequate resources and capacities. Care is needed for analyses of geospatial data. Current practice ranges from regression approaches to neural network techniques. Widely used programming environments bundled as libraries (such as in R, Python; ESRI, QGIS) may reduce user error and encourage code checking. Investigators focused on causal hypothesis testing may benefit from applying approaches such as directed acyclic graphs to identify confounders or colliders (Pearl and Mackenzie 2018); in other phases of research, undirected exploratory analyses may be more useful.

Even when considering analyses of a single environmental measure, independence assumptions may be violated because spatially closer or neighboring units are similar. Data reduction and modeling techniques can help to quantify or account for this, such as through hotspot analyses or geo-aware clustering algorithms. Highly correlated spatial characteristics are also commonly encountered in datasets derived from geospatial sources, requiring analytical methods to take this correlated nature of spatial measures into consideration.[15]

[15] For example, Dias et al. (2023) were able to find a causal connection between the use of glyphosate in genetically modified soybean crops and infant mortality by taking into consideration the geographical dispersion of the pesticide via Brazilian rivers. Aleixo et al. (2022) were able to develop a machine learning model capable of predicting dengue fever outbreaks in individual neighborhoods of Rio de Janeiro by carefully analyzing the geographical distribution of tens of variables.

Geographic location or other characteristics derived from geographic data may have a relationship to the dependent variable outcome that is nonlinear and, sometimes, completely unpredictable. For example, the distribution of bicycle-based mobility flows within a city is influenced by geography but also by city points of interest, residential, work, and leisure areas, as well as the existing transit infrastructure (Kon et al. 2022). A robust understanding of how multiple characteristics of the environment contributed to observed spatial patterns can be promoted by considering multiple measures, study designs, and statistical analysis methodologies.

Interpretation

Initial interpretation of analysis output by researchers should be informed by known or likely data limitations, including those documented in the metadata, as well as other questions and considerations shown in Figure 3.1. Importantly, this should be a starting point to participatory input from others, allowing potential harms or overlooked aspects to be considered. Involvement of broader communities to inform interpretation as conclusions are reached will be more effective if it is based on a prior foundation of working with communities as true partners across the entire research life cycle. Models for such engagement include citizen or community-based science practices and participatory research methods, including community-based participatory research; place-based work on human use of environments may be an especially good fit for such approaches.

Interactive visualizations are a promising way to allow audiences to select options aligned with their interests and needs, increasing the opportunities for engagement in ways that inform interpretation. There exist specialized information visualization platforms for communicating narratives with a geospatial component, such as ESRI's Story Map platform (Alemy et al. 2017).[16] Interactive geographic dashboards (e.g., InterSCity; see Batista et al. 2016) provide another powerful visual tool capable of giving insights and evidence for stakeholders including health professionals and urban planners.

Audience Engagement to Refine and Disseminate Knowledge

For the knowledge generated from geospatial studies to result in public good, efforts to disseminate knowledge must be tailored to multiple audiences. This requires methods and skills for effective dissemination as well as an investment

[16] For example, see the story map produced by the Confederated Tribes of the Grande Ronde (available at https://arcg.is/0v1TO0), which illustrates the geographic history of the various original treaties with the United States and includes a series of interactive maps, narrative text, and other multimedia elements. This story map was one of the winning entries in the 2019 ESRI "Tribal Story Map Challenge."

Enhancing the Use of Data for Public Good: Key Considerations

What do data users need to be aware of?
What is the optimal description of the data?
Who has access and under what conditions?

Collection

What data we collect ("under the lamp post")

- Spatial, temporal coverage and resolution
- May be agnostic to (e.g., Landsat) or aligned to human-drawn boundaries (e.g., census tracts)
- Resolution/quality varies over space and time

What data we don't collect

- Deep historical context (e.g., changes in use and boundaries)
- Relationships between layers
- Data settings that have presented logistical challenges (e.g., informal communities, indoor environments)

Access / Who owns the data

Open / Government / Research / Private sector

- Made publically accessible
- Licensing/cost
- Maintenance / storage control (e.g., servers)

Context of data collection

- Intended uses
- What was not collected/processed
- What was collected but not accessible

Description (what "rides along")

Resolution, coverage, quality (and how this varies over time)

Collection instruments, methods, and context

Post-processing (availability of raw/more granular versions)

Access and use restrictions

Social license: how data were collected, are being used

Limitations

What do data not depict and what is incomplete/missing

What may be lost in raw to processed conversion

What are known problems with representativeness based on incomplete coverage of target areas/population

Errors and uncertainty in measured or estimated values

Figure 3.1 To enhance the use of data for public good, the following must be taken into consideration: What do data users need to be aware of? How can data be optimally described? Who has access, and under which conditions, to the data?

of time and money. Below, we summarize the prosocial motivations by group and offer guidance to aid optimal dissemination to:

1. The *general public*: to encourage critical appraisal, build acceptance and support of data-driven activities, and increase the potential for collective action potential, and enable citizens to hold policy makers accountable.
2. *Target populations*: to benefit and empower specific populations or communities and to support relationships between researchers and these communities.

3. *Study populations*: to benefit directly/indirectly and empower those who have contributed to the greater understanding of their environment, to invest in reciprocity, transparency, and accountability of the research process, and to return value through capacity building.

To reach these first three groups, direct outreach efforts will be invaluable, as will working with journalists, creating community-driven data browsers and visualization tools, utilizing popular platforms (e.g., social media, television, podcasts, online courses, museum talks), and leveraging features of social media platforms that promote dissemination (e.g., consider "bots for scientific good"). Additional groups are important to reach:

4. *Practitioners*: to align practice with evidence, to drive practice change and innovation, and to inform interventions and planning. This can best be achieved through existing structures for training/skill development (e.g., continuing education, professional associations).
5. *Policy makers*: to promote evidence-driven policy making, to support critical appraisal and adjustment of existing policies, and to prevent distortion of scientific results. Here, the preparation of contextually framed policy briefs is imperative as well as tapping into existing mechanisms for public comment as well as funder and advocacy organizations' lobbying networks.
6. S*cientific community*: to ensure accountability, to foster collaboration, to generate new ideas, and to receive feedback. Efforts should focus on refereed journals, scientific meetings, etc., and through cooperation with the various professional societies.
7. *Private sector*: to increase knowledge that supports harm avoidance, thereby increasing public good; to increase legal culpability; and to encourage doing good while doing well. Efforts will need to communicate contextually framed information on the potential for harm and interventions for future harm avoidance.

Across all audiences, the following pitfalls need to be considered as strategies are developed:

- Reification of bias
- Limited expertise, experience, and resources
- Competing priorities and time commitments of research team
- Competing demands for audience attention
- Difficulty of contextualization
- Temptation to oversimplify or overhype
- Insufficient accessibility of language or terminology
- The elongated timescale of science generally or of a given research project specifically as compared with the "media cycle"

To engage people who could be affected by physical and built environment characteristics,[17] it is important to distinguish between the individuals directly engaged in a study (participants), other residents/users of the studied settings, and broader populations spanning other settings to whom the research conclusions may be generalized.[18] Investigators need to adopt specific strategies to reach groups affected by their research, address any risks for group harm, and ensure transparency through communication. Working with journalists, and science or data journalists in particular, can be a strategy to reach multiple audiences and foster trust and recognition for scientific endeavors and advances in our fields.

Yet not all dissemination efforts will be equally effective. Brevity is valued. Timing of dissemination can determine receptivity to research findings. The optimal time to speak to the media or to publish op-eds based upon research findings may not be when the research has been completed/published, but rather when a relevant issue arises within the public discourse or policy agenda. For example, emerging attention to wildfires may present an unanticipated window of opportunity to disseminate research on health effects from air pollution or land management decisions. Complementary engagement may be considered across multiple formats (e.g., reaching policy makers through both producing a two-page policy brief and contributing comments on a public notice with obligatory response to relevant comments). Beyond the efforts of individual investigators, there may be a role for professional societies to scan for relevant actions (such as public comment periods for certain subjects) so that membership can be made aware of relevant rulemaking.

Nuances and caveats typically reserved for communication to a specialist audience may be important to translate to a broader audience in a concise and accessible way that conveys which findings are fragile versus robust; oversimplifying the message may be expedient but could later backfire as subsequent and seemingly conflicting findings impede understanding or undermine trust. Without appropriate training, dissemination opportunities can be mishandled, communities offended or harmed, misinformation reinforced, and disinformation propagated. Inaccessible language can impede effective communication that meets audience needs, including due to unintended connotations of some commonly used scientific terms (Somerville 2012).

How research should be disseminated will also depend on characteristics of the researcher and audience. In some instances, rules established by funder

[17] Often, those affected by the environment are discussed as "stakeholders." We note, however, that there was not agreement within our group about the utility and appropriateness of that term.

[18] The known limitations of such generalization are important to articulate, though generalization beyond the observations included requires strong assumptions. A key assumption for generalizing claims about an environmental effect on human health, for example, would be that there is no effect modification (even if unmodeled) that results in a different strength or direction of association, or that any effect modifier is not differently distributed between the measured and target population or settings.

organizations or governing the organization where the research is conducted may shape what is allowed to be communicated to elected officials. Other potential obstacles may be institutional fear of upsetting funders or board members. In order to reach the private sector where the aim includes establishments of culpability, contextualization of the information may be warranted such that not only possibility for harm is demonstrated but interventions for future harm avoidance are proposed.

Ultimately, a dissemination process that reaches a broader audience is critical to professional advancement of researchers individually and stands to benefit the field through influence on funders and policy maker priorities (Dudo and Besley 2016). Dissemination should be wide ranging, encourage multidirectional discourse, and may frequently extend beyond the conclusions of a particular project (e.g., massively open online courses [MOOCs] or other formats for continuing education, TED talks, science museum events, podcasts, engagement on social media platforms, and other ongoing public outreach roles). Specific cultural contexts may also create other avenues to move science messaging to be more resonant with popular culture forms. For example, a Canadian public service announcement designed to reach a Punjabi-speaking population used video featuring bhangra dance and a well-known Indo-Canadian actress to enliven delivery of the message about pesticide safety and laundry instructions (Murphy and Nicol 2010).

Critical consumption of knowledge warrants attention across all sectors of society and all career stages as the methods used continue to advance (Few 2019). Especially with complex topics, a craving for simplistic solutions and a rush to attribution may lead to trendy, overhyped, and misleading science. To avoid dissemination of simplistic conclusions which can undermine public trust, attention is needed to who is engaged in research, how we train investigators, and system-wide incentive structures.

Future Directions

For digital environmental measurement and research to advance in ways envisioned as beneficial to humanity, attention will be needed on inclusive team formation, cross-disciplinary training, and leverage points for sustaining change.

Team Formation: Include Diverse Perspectives Early

The need to include diverse perspectives is a unifying theme throughout our vision for determining what to measure about the environment, documenting how and why, and bringing intentionality to the generation and dissemination of knowledge. A team science approach offers the potential to combine the strengths of multiple fields. For place-based digital ethology, teams should consider including computer research scientists, geographers, and others skilled in working with geographically referenced data (as highlighted by Brinkhoff, this volume); those with expertise in the building of the environment such as urban planners and architects; and domain scientists with expertise in how environments affect the people who live or spend time in them (e.g., environmental health, environmental psychology, urban health). Bringing together these multiple perspectives early recognizes the depth of expertise and limits the risk of redundant effort rediscovering what is already established in another field. Further, teams and partnerships that bring together a diversity of disciplinary, identity-based, and lived experience may be especially crucial to scrutinizing unsupported assumptions and addressing shortcomings of conventional practices.

Interdisciplinary researcher teams are poised for effectiveness by encompassing knowledge about standard and emerging practices for measuring and investigating the environment, including promising practices from across disciplines in data stewardship, responsible use of data, and supporting audiences in the critical consumption of knowledge produced with data. Skills and roles that allow at least some team members to work with nonacademic partners (e.g., inviting input and exploring social license for use of data directly with communities, cultivating audience ties including through interactions with media) are most beneficial if incorporated from the earliest stages of collaboration planning.

Training Needs to Leverage Place-Based Digital Ethology for Good

Cross-disciplinary training was identified as a priority to support working across silos, as well as building networks that span disciplines, allowing for rapid dissemination of promising approaches and ideas. Cross-disciplinary training relevant to digital ethology should provide a foundation for future collaboration among those with knowledge of geographic settings (e.g., urban planners, architects), physical and mental health in humans (e.g., medicine, public health, psychiatry, neuroscience), social processes (e.g., sociology, anthropology), technology to collect and use digital data (e.g., computer science research), and outreach to partners and audiences (e.g., community-engaged research, communication and implementation science).

As no one person can reasonably cover this full range of skills and expertise, disciplinary silos impede collaboration. To foster awareness and appreciation of other disciplines among disciplinary specialists, it will be necessary to develop models for cross-training at different career stages. This may take the form of courses offering basic skill building or a primer for topics outside of one's own discipline. There can also be value in dual degree or exchange programs, for instance embedding journalism trainees within science teams, which promotes two-way learning.

Even among fully trained professionals who are active in the field, ongoing learning opportunities are needed. Workshops and discussions across disciplines (such as the Forum which resulted in this volume) may create the opportunity to discuss and mitigate potential harms of siloed research (e.g., awareness of legal, ethical considerations that are relevant to avoiding harmful consequences of technically possible research).

Setting the Stage for Sustainable Change

Given the challenge of competing priorities, attention is needed on upstream structures and incentives that can sustain change in project planning and implementation, dissemination, and other scientific professional practice. Here we present some preliminary ideas concerning strategies for organizations that provide research funding, journal editors, and academic institutions.

Funders

Organizations that fund research teams and projects have leverage to incentivize sustained change. Some funders already incorporate planning grants, dissemination requirements, or other approaches designed to foster inclusive team formation and knowledge generation that reaches those who can take action to reduce harms or create benefits. Since harvesting metadata on a large scale can be challenging (Lagoze et al. 2006), standards and citation rules for meta-analyses should be supported and incentivized by funders. Funders can require and financially support the generation of user-ready GIS data repositories, such as efforts supported currently by the Lacuna Fund. This may require special initiatives to generate such repositories similar to one designed to address environmental influences on child health outcomes (Smith et al. 2018), or special funding that would be provided for the work necessary to prepare and add data to repositories at the end of studies. Complementary to traditional seed funding to incubate a nascent project, funders may consider "harvest funding" to amplify the ability of teams to articulate, document, and disseminate both insights and data from a project as it concludes.

Journal Editors

As peer-reviewed articles continue to be an important signal of research reputation used in making funding and promotion decisions, journal editors can set the stage for improved practice through required reporting and flexible formats.

Journals increasingly ask authors to include statements about data sharing or the involvements of affected populations (e.g., the British Medical Journal required reporting on patient and public involvement) (Boivin et al. 2018). The acknowledgment of acceptable social license could, in the future, be considered a requirement for publication with digital data of certain types, in a way similar to the ethics assessment carried out by Institutional Review Boards.

Beyond current reporting requirements for articles, journal editors can create spaces for published datasets and metadata with shareable digital object identifiers and promote citing these as a way to connect related work and ensure credit. This could take the form of an article type or follow the model of a dedicated journal as illustrated by the *Nature* journal *Scientific Data*, which includes examples such as a description of global emissions mapping data (Weng et al. 2020).

Academic Institutions

Academic institutions can update promotion and tenure guidelines and processes such that they reward the investment of time in activities described above, from building of database structures and repositories to developing trusting partnerships with communities and other knowledge dissemination audiences. Institutions may decide to set up or further invest in structures to facilitate cross-disciplinary collaboration in the form of Institutes or Centers. These can foster the necessary policy-oriented communication capacities and practical collaborative infrastructure that allows for feasible incorporation of skills that any given research project may not have sufficient funding to sustain (e.g., web designers to provide audiences with access to knowledge synthesis and interactive mapping).

Conclusion

With the considerations described in this chapter, we envision a world in which digital data on the physical, natural, and built environment are useful and used for public good. Data volume, scope, depth, and quality are likely to increase in the future. We are already seeing multiple benefits, although missteps may be an inevitable part of our path forward as the field evolves. Boundless potential gives us optimism for appropriate use, while recognizing that attention is needed to amplify responsible use of digital data for knowledge generation, equity, and other public benefits.

Acknowledgments

The authors would like to acknowledge funding from Dr. Ernst Strüngmann Foundation; inspiration and feedback from those who participated in the Digital Ethology Forum, and assistance from Julia Lupp, Marta Gruszka, and all at the Ernst Strüngmann Forum who set the stage for and supported our work.

4

Characterizing Social Environments in the Physical and Virtual Worlds Using Digital Data

Michele C. Weigle, José Balsa-Barreiro,
Nitesh V. Chawla, Tamas Dávid-Barrett, Maria Melchior,
Virginia Pallante, Abeed Sarker, and Jason Gilliland

Abstract

The study of social environments has typically revolved around interactions in the physical world. Here, a contemporary perspective of social environments weaves together multidisciplinary viewpoints and considers both physical and virtual spaces that offer opportunities for interaction. In the intersection where virtual and physical spaces collide, how does the structure of the social environment in the physical world affect that in the virtual world, and vice versa? How can abundant area-level digital data, produced at multiple locations and points in time, be used to study these social environments? This chapter examines the role that digital data plays in the study of human interactions, with considerations for context, in terms of physical proximity, history, and culture, as well as the advantages and challenges presented in using social media data for this type of study. The long-term goal is to examine how the social environment extends from the physical verse to the metaverse. This provides an unprecedented opportunity to characterize not only social environments using digital data but also to juxtapose them with the influence of physical environments.

Introduction

An important aspect of digital ethology is the role that digital data[1] can play in characterizing the *social environments* of an individual. The accelerating use of digital information technologies has allowed researchers to access and

[1] Digital data are data stored in digital form. In our context, digital data may refer to digital behavior (e.g., social media posts, number of followers/likes/shares on a social media platform, online search queries) or nondigital behavior (e.g., geolocation, census data, emergency room records).

analyze digital records about our behavior and interactions in the physical world (e.g., movement patterns, purchase history, mobile phone interactions). Traditionally, social environments were constrained to physical spaces such as neighborhoods or workplaces—the context in which people have the opportunity to interact by reflecting socioeconomic characteristics, including social networks[2] and levels of social support. These aspects of individuals' social environments determine their quality of life and collective behavior within their communities. While these aspects of the traditional social environment still hold true, the margins of the environment themselves have shifted through an increased impact by the virtual world.[3]

In addition to physical places, people interact online, and when they do, they leave many digital traces that encapsulate and define their collective behavior. Our understanding of the concept of social environment has been sharply modified by the advent of social media[4] and social media platforms,[5] which have been used to facilitate online communication and interactions. The social environments in which individuals live today consist of a combination of their physical and virtual environments. The emergence of this *new digital social environment* provides us with new opportunities and challenges to understand individuals' behavior and their interactions within their social environments. In this chapter, we attempt to characterize this new digital social environment in the context of the ongoing digital revolution. An individual's activities and their social interactions in the virtual world can influence and supplement their physical social environment, even covering some gaps or deficiencies in it. The online social networks[6] that an individual builds can become key factors in their social environments. In some cases, such as in augmented reality[7] and the metaverse,[8] the virtual and physical social environments intersect, thus blurring the boundaries between both social environments.

[2] A social network consists of the connections and relationships made between individuals in the physical world. Social networks can consist of strong ties (i.e., close friendships) and weak ties (i.e., acquaintances, work colleagues).

[3] The virtual world represents the online world and consists of interactions and connections made using online platforms. We use the term virtual world, or environment, as a contrast to the physical world.

[4] Social media are communications between users in text, image, or video form shared over the Internet on a third-party platform (as opposed to direct communication using mobile phones).

[5] A social media platform is a third-party software platform used to facilitate communications and connections between individuals and groups over the Internet.

[6] Online social networks are social networks developed via social media platforms. In some literature, the terms online social networks and social media platforms are used interchangeably, but we make the distinction here.

[7] Augmented reality presents an overlay of a virtual world or virtual objects onto a view of the physical world.

[8] The metaverse is an emerging virtual world that combines simulated virtual reality (facilitated by 3D headsets) and elements of social interaction and connection to create an emotional and believable immersive experience.

This chapter reflects our multifaceted discussions during the Forum, aimed at developing a framework to characterize social environments, both in the physical and virtual worlds, using digital data. We start by describing social environments from different viewpoints, including those based on ethology, social norms, geography and place, social epidemiology, and social networks. We then develop a view of social environments based on digital ethology, bringing in components of social environments from both the physical and virtual worlds. To be able to use digital data to study social environments, we must first consider how aspects related to the context influence social behavior. We emphasize factors related to physical proximity and human emotions as key aspects. We refer to the impact of context on the behavior of particular collectives (e.g., people who immigrate) and discuss how social bridges and social mobility can affect context. We then dive into how an individual's social environments can be influenced by the virtual world. This includes the effects of building social capital in the virtual world and considering how interactions in the virtual world can affect behavior in the physical world. In an attempt to understand our behavioral evolution in the near future, we discuss what ethology means in the metaverse and what implications the metaverse might have in our social environments and how they are perceived. We then turn to the various types of digital data that can capture traces of human behavior and provide a detailed discussion of social media data, which can be a rich source of information about the interactions in the virtual world. We discuss the advantages of using data extracted from social media (e.g., size, speed, capturing emergent knowledge) in comparison to more traditional sources (e.g., paper-based surveys) and the challenges of deriving knowledge from these data sources (e.g., making individual vs. group-level inferences, use of colloquial language, generalizability). We focus, in particular, on the various types of bias that might be present in social media data and discuss several studies that have used social media data to examine aspects of human behavior. Finally, we conclude with open questions to be addressed in the near future.

Describing Social Environments

There is not a singular definition of what constitutes a social environment nor a consensus across different disciplines, but there are overlapping elements. In addition, there is not one single social environment, rather, an individual can have multiple social environments that reflect the different aspects, spaces, functions, or interactions present in their life. Here we present several different, but related, views on social environments and conclude with a unifying view of social environments for digital ethology.

Ethology View

We first consider social environments at the ethological level by considering the social behavior of nonhuman primates. In the wild, social environments vary according to species' characteristics and communicative system. They are not restricted to a particular defined physical environment and might change in scale, but they are related to the *opportunity that an individual has to interact socially with other individuals*. The social interactions do not necessarily have to take place, but the boundaries defining the social environment are delimited by the potential for the interaction. This potential varies among species, depending on their social structure and communicative system.

The social system of distinct animal species shape their social environments because it includes individuals with different characteristics (e.g., the prevalence of one sex over the other), and these individuals might use the physical environment in different ways and interact with different conspecifics (Mitani et al. 2012). For instance, social systems of nonhuman primates might vary from extended families organized around matrilineal hierarchies, where social interactions are more or less biased toward kin (Sueur et al. 2011), to troops that are organized within families. Families, in turn, are organized hierarchically across the troop, where friendships with immigrant males might be common (Smuts 2017). A species' social system shapes, therefore, the identity of the individuals with whom the interaction takes place in the social environment.

Different social systems can build different social environments according to the use that individuals make of it. For example, different ecological selective pressures can generate different social organizations, as in the case of species living in fission-fusion societies, which are characterized by a temporal and fluidly dynamic separation of the individuals in subgroups (Symington 1990). These dynamics challenge the opportunity to interact socially and moreover they create distinctive selective pressures affecting, in turn, the communicative system of the species (Aureli et al. 2008). The potential for communication between individuals shapes the physical space that allows the social interaction between the individuals of a community, concurring in defining a flexible social environment.

Therefore, the influence of the social structure and the communicative system of a species creates a specific social environment at many levels (e.g., groups, species, orders), which is reflected in the potential for the individuals to interact. Examples range from cultural differences in the responsiveness during joint attention interactions (Bard et al. 2021) to the variability in the behaviors used to communicate in different but phylogenetically close species in relation to the evolutionary history (bonobos vs. chimpanzees, Gruber and Clay 2016; different macaques' species, Maestripieri 2005). Both human infants and young chimpanzees show a significant cultural variability when they interact socially with an adult female to share attention toward an object (joint attention); this highlights that different forms of engagement need to be contextualized to be fully understood in their expression (Bard et al. 2021).

Whatever the selective pressure, the social environment and its complexity influence the evolution of the social traits of a species in a feedback loop that redefines the social environment itself and the social interactions that characterize it.

Norm-Based View

Humans, like all other social animals, spend most of their lives in proximity to and in interaction with conspecifics. We largely live in "nests" built by others, eat food harvested and prepared by others, engage in conversations about knowledge created by others, and develop new ideas in cooperation with others, in a full behavioral synchrony. These interactions build a basis for a social environment that is both influenced by and influences social norms. Social norms guide our interactions, technologies (whether engineering or social), rules, laws, and how we perceive the universe around us. This broader societal context is a key component of our social environments and includes prevailing cultural norms, religious beliefs, structural racism, legal frameworks, political institutions, and other factors that may shape human attitudes, behaviors, and opportunities. We live in complex social networks, whether we are hunter-gatherers (Apicella et al. 2012) or live in modern societies (Dunbar and Spoors 1995). There are often consequences for individuals who violate social norms, including reduced opportunity for interaction or even being ostracized from the community (Kam and Bond 2009; van Kleef et al. 2015; van Leeuwen et al. 2012). In fact, perhaps the most important factor in the human brain, being of such an exceptionally large size, is the need to manage the cognitive demands of interacting with others inside complex social environments (Dávid-Barrett and Dunbar 2013), whether it is the task of computing strategic social action (Dunbar and Shultz 2007) or coordinating to achieve behavioral synchrony (Dávid-Barrett and Dunbar 2012).

Place-Based View

Social environments are the social settings or contexts in which people live and potentially interact with others. Interactions can occur in both physical settings (i.e., occupying physical space with a geographic location) and virtual settings (e.g., an online community). A place-based view focuses, however, on physical settings: the places people inhabit and live their lives. In human geography, "place" is traditionally defined as a location that has been constructed by human experiences; it is distinguished by the sociocultural or subjective meanings through which it is created and differentiated (Relph 1976; Tuan 1977). In this place-based view, an individual's social environment begins at home (i.e., where one sleeps at night). Here, an individual may interact regularly with other members in the household (family or unrelated roommates) and is both influenced by, and helps shape, the social norms of others who occupy the same home. People are also influenced by the social environments of their respective workplaces and/or educational centers, where they interact with their friends, colleagues/classmates, and supervisors/teachers. In these places, there are rules (explicit), social norms (implicit), and power relations (both explicit and implicit) that influence social relations and an individual's behavior within these social environments (Cresswell 2004). These social environments are typically experienced several times per week. In addition, other influential places make up one's social environments, such as places of worship, commerce, and recreation. These places, typically located within one's immediate neighborhood or are at least geographically accessible within one's settlement or population center (e.g., city, town, village), are usually accessed less frequently than home and work/school, but vary according to personal, cultural, and geographic factors. Balsa-Barreiro and Menendez (this volume) describe several ways in which geography and population density in urban versus rural settings impact the opportunities for and types of social interactions. Furthermore, an individual's social environment extends to their neighborhood, city, state, and country of residence. These administrative/governmental entities influence human behavior in that they exert power over society through, for example, laws and norms.

Social Epidemiology View

Following seminal works from Durkheim (1897) and Villermé (2008), social environment refers to social interactions and relations among people at different levels of analysis determined by the household, the family, the school, the workplace, the neighborhood, the society in which one lives (Berkman et al. 2014), as well as more recently the digital environment where people evolve in, willingly or not. The key idea is that these social relations, organized in networks, are essential for individuals' well-being and behavior, as well as for other outcomes. While we generally think of social environments as being resources, they can also be sources of negative interactions and exposures, such as conflict, violence, and incentives to engage in unhealthy or dangerous behaviors (Villalonga-Olives and Kawachi 2017). One of the questions that arises with recent changes in social interactions, increased by the dissemination of digital media, is the extent to which virtual social interactions replace, compensate, or augment face-to-face interactions. Moreover, the social environment also refers to the hierarchy of social relations in a society, which conditions access to socioeconomic resources related to education, employment, occupation, housing, and place of residence, which determine individuals' status in society.

Social Network View

If an individual's social environments are based on the opportunity for interaction, then they depend necessarily on the individual's social network (i.e., the network of personal connections that the individual has with others). Not all personal connections are equal in weight. These connections, or ties, have been generally classified as strong ties (e.g., close friendships with frequent meaningful interactions) or weak ties (e.g., acquaintances with fewer meaningful interactions). The importance of social ties in well-being has been recognized for a long time, both in terms of number (Dunbar and Spoors 1995; Hill and Dunbar 2003; Shultz and Dunbar 2010) and quality of ties (Granovetter 1973; Seyfarth and Cheney 2012). It was previously assumed that the well-being effect comes from the intensity of the relationships, which is usually determined by the frequency of meaningful interactions (Pollet et al. 2013; Roberts and Dunbar 2011). There is, however, a further effect that stems from the level of interconnectedness of the social network itself (Brondino et al. 2017; Dávid-Barrett 2022a; Dunbar 1998). For us to feel safe, we need to perceive a highly integrated social network around us, despite the fact that some studies on complex systems have demonstrated how networks with many interdependences tend to be more unstable (Balsa-Barreiro et al. 2020b). This integration (or lack thereof) can highly shape how we view our social environments, either as a benefit or a drawback. An organizing principle of social networks is also the notion of structural diversity, which suggests the number of connected components and their influence in forming the network connection (Dong et al. 2017b). For example, just being from the same larger physical space (zip code) might imply a more diverse common neighborhood between two connected nodes in a network, as each of those nodes may have their own friends or workplace connections; nonetheless, two close college friends may have more similarities in their connections, thus creating a less diverse common social neighborhood. Such diverse or common neighborhoods create the spectrum of social resources available to an individual.

Through much of human history, the primary organizing principle of all human communities was kinship. The fall in family size, especially when combined with urbanization, has led to the rise of friendship as the dominant form of social relationships (Dávid-Barrett 2019). Friendship is fundamentally different in its nature to kinship, in that the latter is mostly preset (and in a network sense, prewired), but the former is flexible. Such flexibility poses, however, a network organization problem, as friendship groups, if organized randomly, have a much lower level of integration (lower clustering coefficient, in network science terms) than kinship groups (Dávid-Barrett 2022a). One possible solution to this problem is the use of trait similarity (homophily) in friendship choice (Dávid-Barrett 2020). This mechanism explains the importance of homophily in friendship choice, a well-established phenomenon (Kossinets and Watts 2009; Laakasuo et al. 2020; McPherson et al. 2001). The presence of homophily in social networks reflects the interplay of selection, where an individual may choose to form ties with others who have similar characteristics or interests, and social influence, where an individual's existing ties contribute to the development of new interests (Easley and Kleinberg 2010).

Digital Ethology View

Going forward, we take into consideration the various views on social environments that have been presented thus far. In doing so, we formulate a unifying definition of social environments that can encompass both the physical and virtual worlds and the various factors that impact an individual's social environments. When we include the virtual world, one's social environment also includes the interactions that can occur within personal online communities. For example, one can belong to, interact with, and be influenced by (and help create) the content within various social media platforms (e.g., Reddit, Twitter/X, Facebook). While these online social environments do not occupy a precise physical space, for those who spend a large amount of time on these platforms, they may exert a powerful influence over their real-life behavior.

In our view, and as related to digital ethology, *social environments are spaces where the opportunity for interaction occurs, whether physical or virtual, personal or societal*.

Context Can Affect Social Behavior

Societal context, in terms of physical proximity, history, and/or culture, is an important component of social environment as it can affect our social environments even when there are no direct interactions, such as the influence of proximity. For example, to understand the social and economic behavior of Mexican citizens, their proximity to the United States must be considered, even for those who do not travel to the United States or interact directly with Americans. Something similar happens in many Eastern European countries, where collective behavior is sharply influenced by recent history. For instance, even though more than 30 years have transpired since the Reunification of East and West Germany (Andor 2019), difficulties associated with converging the two populations are evident in terms of health disparities (Grigoriev and Pechholdová 2017), educational opportunities (Klein et al. 2018), and political attitudes (Weisskircher 2020). In this way, past history can lead to human emotions related to fear, mistrust, or guilt being mutually shared by whole communities.

When using data to study social environments, it is important to frame the data in the appropriate context and consider the source of the data. In some cases, what appears to be the same data points can lead to different conclusions (Balsa-Barreiro et al. 2022). This can happen even with indicators that can be quantified from an objective perspective. For example, how is poverty defined? The concept of poverty is contextual, varying according to different situations. An individual could be deemed impoverished within an affluent community while being relatively wealthy in a deprived neighborhood, even with an identical income in both settings.

Throughout history, humans have developed a series of survival strategies based on the simplification of information. An evident instance is the establishment of straightforward stereotypes about individuals from diverse countries, commonly held by many. Past relationships through history, popular traditions, books, and broadcast media contribute to spread and perpetuate these stereotypes. This societal survival strategy once made sense in terms of biological machinery for generations, yet such simple and binary thinking has become a problem in a society where the number of interactions and information available has grown exponentially over the last few decades (Dutton 2021). Therefore, properly contextualizing datasets is crucial to prevent biased outcomes and potentially misleading conclusions, which can result in weak and inadequate decisions. Incorporating context and a comprehensive grasp of spatial scales is vital, particularly given the extensive use of data-driven tools in decision-making processes.

Below, we discuss several examples that demonstrate how societal context can affect behavior in sometimes nonintuitive ways. These examples highlight the need to include context when drawing conclusions about group-level phenomena observed in data.

Impact of Context on the Behavior of People Who Immigrate

People who immigrate move from one context into another, often vastly different, context. Here, we consider, at a group level, how this change of location can affect behavior and the way in which these behaviors evolve over time. This provides important insights into the roles of different types of environments.

Research conducted as far back as the 1970s by Len Syme's group (Robertson et al. 1977) showed that over time, the behaviors of people who immigrate come to resemble those of the host population. For example, they found that men who immigrated to the United States from Japan had higher levels of cardiovascular disease than those living in Japan; further, levels of risk for cardiovascular disease varied depending on whether they resided in California or Hawaii. More recent research on this topic has shown that among persons who migrate from Ghana to Europe, dietary patterns change and cardiovascular risk is higher than if they had stayed in rural or urban Ghana, as well as across the destination cities (Galbete et al. 2017). Most notably, consumption of sugar, principally through soda drinks, varies greatly for a Ghanaian residing in London, Berlin, and Amsterdam, regardless of any other characteristics. Galbete et al. (2017) also showed that over time, Ghanaians who immigrated to Europe have an elevated risk of hypertension. The factors associated with that increased risk actually vary, however, across places of residence, further highlighting the role of context (van der Linden et al. 2022).

Moreover, the behavior of those who immigrate changes over time as well as across generations; differences are also possible across communities and contexts. For instance, research conducted in the United States shows that descendants of immigrants from Asia or South America follow similar diets as the host population, whereas South Asians appear to have distinct dietary patterns that resemble those of first-generation immigrants from South Asia (Rodriguez et al. 2020).

Sociological research has also shown that those who immigrate tend to converge with the majority population over time. Patrick Simon's group has studied the way in which individuals name their children in a nationally representative study of people who immigrate and their descendants living in France. Data show that while traditional French names are not common among descendants of immigrants, these children are also not given names that are most common in their parents' country or culture of origin either. Rather, they are given names that lie somewhere in-between the standards of the culture of origin and the French setting (Coulmont and Simon 2019). Consistent data have shown that persons who have a foreign, and particularly a Muslim-sounding, name are at high risk of experiencing discrimination with regard to education, employment, housing, and possibly other domains of life. Thus, giving a particular name to a child may shape the social environment and experiences of the child later in life (Simon 2017).

Impact of Social Bridges

Social bridges are individuals whose social networks serve to connect multiple communities and facilitate information exchange between the communities. Dong et al. (2017a) studied the impact of social bridges on purchase behavior between different communities when their social bridges worked at locations near each other. Their main assumption was that because they worked in proximity to each other, these social bridges had the opportunity to foster information exchange, which could then be transferred back to their home communities. The authors analyzed millions of credit card transactions and found more similarity in purchase behavior between communities that had higher numbers of social bridges linking them. This similarity was even present for nonbridge individuals in the communities. Further, the number of social bridges between two communities was a stronger indicator of purchase similarity than other factors, such as income, gender, or age.

Impact of Social Mobility

Studies examining the causal link between socioeconomic status and health/behaviors often evaluate this through investigations into the effects of social mobility, by considering both upward and downward shifts. Dohrenwend et al. (1992) conducted a key study by comparing the risk of psychiatric disorder to the level of educational attainment across two different ethno-racial groups in Israel. They showed that among young people who do not belong to a socioeconomically disadvantaged group, a low level of education was associated with a higher risk for a psychotic disorder. This effect was not observed among young people with similar level of education who came from a socioeconomically disadvantaged group. This suggests that downward social mobility can be related to poor health; that is, individuals are "selected" into a social group because of impaired mental health. In contrast, individuals who came from a socially disadvantaged group but achieved higher levels of education were at low risk for psychotic disorders, indicating that upward social mobility could be protective. Similarly, intergenerational upward mobility has been found to predict health habits (Mok et al. 2018) and mental health levels (Melchior et al. 2018) generally comparable to those of individuals who always experienced favorable socioeconomic conditions.

In the 1990s, vivid debates played out in the scientific literature between Michael Marmot's and George Davey-Smith's groups, regarding why social hierarchy and one's place within it influences behavior, with opposing views on the role of material versus psychosocial pathways. There is now evidence that both these mechanisms contribute to socioeconomic inequalities in behaviors and health (Fleitas Alfonzo et al. 2022). Moreover, extensive research has documented that social, economic, and the physical characteristics of places where individuals reside and spend most of their time contribute as well to condition certain behaviors (Daniels et al. 2021). Importantly, data from Ana Diez-Roux's group show that if one lives in a deprived neighborhood, the proximity to a wealthy area is also relevant (Auchincloss et al. 2006), indicating that the concentration of poverty is detrimental to health behaviors possibly because of reduced access to resources as well as higher stress resulting from spatial segregation.

Following the hypothesis proposed by Putnam (2000) and translated to epidemiology by Kawachi and Berkman (2014), social cohesion and social capital within communities and neighborhoods have been proposed to be protective in terms of health and health behaviors. This is based on the idea that tight social ties in a community provide a setting for individuals' supportive social networks and a source of social control, which can help taper unwanted behaviors. While much research has shown that a cohesive social environment can be positive, some evidence indicates that it is not, particularly when a person is excluded. Consistent research, primarily based in the United Kingdom and the Netherlands, show that members of ethno-racial minority groups who reside in neighborhoods populated primarily by members of nonminority groups have elevated rates of psychosis (Baker et al. 2021). The relative heterogeneity in findings across settings suggests that different social contexts exert varying effects. The main mechanisms that have been proposed to explain this counterintuitive finding relate to individuals' experiences of racism and discrimination in neighborhoods where members of ethno-racial minority groups are few, leading overall to experiences of social exclusion and elevated acute as well as chronic stress levels (Henssler et al. 2020).

Social Environments and the Virtual World

The social environment in a virtual world offers a large capacity to develop interactions that span neighborhoods, regions, and countries, which are things limited by the physical boundaries and constraints imposed by the physical world. In a virtual world, "travel documents" are not needed to communicate with someone across national borders, as there would be in the physical world. This creates fertile ground for unique social environments that may be particular to an individual and the development of communities and may even be of assistance to the individual. Nonetheless, the individual may also be exposed to risks that stem from the wide mix of social interactions that emerge in the virtual world. Virtual world and physical world interactions also intersect, however. Do virtual world interactions create strong ties, or are strong ties preordained as kinship or ties that originate in the physical world?

Social environment can also be a relative concept. If an individual is devoid of an accessible physical neighborhood for living or work, do virtual interactions create a complementary set of opportunities? If so, how do we develop a union or intersection of interactions between the virtual and physical environments that collectively build the social environment? What if an individual does not have good access to technology to enable a virtual social environment? Do these factors contribute further to inequities? Will an individual who is already facing disparities in the physical environment be further disadvantaged in the virtual world because of a lack of access to technology? As we imagine the construction of social environments, it becomes important to consider these questions and be able to develop a utility metric to characterize or measure the level of quality of social environments.

Social Capital and the Economy of Attention

One major element of the virtual world are online social networks that are facilitated by social media platforms, which take advantage of the Internet to allow interactions between individuals and groups in text, voice, image, or video format (Sarker, this volume). The emerging business model of social media platforms is to sell advertisements targeted toward specific groups of users based on behaviors tracked by the social media platforms. The longer a user stays engaged on a platform, the more advertisements they will see, so social media platforms have the incentive to hold users' attention for as long as possible. In essence, the use of social media platforms may be free for users, but it comes at the expense of providing extensive behavioral information to advertisers.

From the human behavioral perspective, online social networks are framed within the economy of attention. Franck (2019) outlines the shape of this new, quaternary sector of the economy, characterized by dematerialization and virtualization, where our attention is the main asset. Online social networks are presented as means or tools that allow us to connect with the world and communicate with our friends. They can be, however, somewhat more complex. For Wei (2019), online social networks are basically tools to extract and show status or social capital. This status is calculated as the sum of all the elements of prestige existing in our social life. In the past, this social capital was highly fragmented and difficult to estimate, at least until the advent of digital social media. Online social networks generate a new market where it is possible to quantify our social capital based on our communication and interactions on them, by checking the images we see/like/share, our comments, and our connections, among others. Competition in online social networks intensifies as more users seek increasing attention, but our limited attentional capacity means favoring some neglects others. Consequently, a paradox emerges: over time, all our online friends become competitors or adversaries in the medium to long term.

Interactions in Digital Spaces Can Affect Behavior

Some studies have explored how digital spaces can impact human behavior in real life. For example, in 2016, the first mobile phone-based augmented reality game, Pokémon Go, was released and became popular worldwide. Pokémon Go superimposed a virtual world based on augmented reality on top of the physical world; imaginary creatures called Pokémons could be seen and captured as part of the game. The game required players to walk around and explore the physical world in search of the virtual creatures. Althoff et al. (2016) showed that over a period of 30 days, engaged game players increased their average step count by 1,473 steps per day, approximately 25% more than usual. They estimated that within the brief time span of the study, the game resulted in a total of 144 billion additional steps to the overall U.S. physical activity. This was the first study that reported on the impact of augmented reality on the real-life physical activity of humans. Similar follow-up studies around the world (Laato et al. 2021; Ma et al. 2018) showed that connecting virtual spaces with physical world objects had the benefits of increasing physical activity and supporting social meetings.

Interactions on social media platforms can also change real-world behavior. A study of young girls who use the photo-based social media platform Instagram found more negative levels of body image than those who did not use the platform, likely due to social comparison. (Pedalino and Camerini 2022). Others found that cosmetic surgery consultations related to interventions similar to the filters used in social media increased (Maes and de Lenne 2022) and have demonstrated how social media can manifest a distorted view of physical reality (Hong et al. 2020; Perrotta 2020). In recent years, the idea has spread that online social networks function as echo chambers (Bail et al. 2018; Cinelli et al. 2021), where users interact exclusively with others (users and/or media) with similar ideologies and are no longer exposed to information that differs or contradicts their own ideas. In this way, it is hypothesized that online social networks are closed systems where one's own ideas would seem true due to amplification and continuous repetition of them.

In recent years, virtual reality technology has improved to the point where consumer devices are both affordable and provide a believably immersive experience. Combining virtual reality environments with elements of social interaction creates the opportunity for a new virtual space, the *metaverse*. An early example of such a community was the virtual world Second Life, released initially in 2003. Because the technology that powers these environments provides a more realistic experience, we can start to ask questions about how interactions in the metaverse could be different from physical interactions or interactions that take place on traditional online social media platforms. These platforms could allow individuals to break out from the social environments they experience in the physical world, which are influenced by culture, history, and social norms centered on a place. This is a new and exciting avenue for research. How will the metaverse impact an individual's social environments? What would digital ethology look like in the metaverse? What is behavior in the metaverse? Who is the actor, or who engages in the behavior? Who is the observer, or ethologist, in the metaverse? How might the ability to interact in ways that are impossible in the physical world, or to set up social norms and conditions that would take years to develop in the physical world, allow experimentation and incubation of ideas that could later be manifested in the physical world?

Using Digital Data to Learn about Social Environments

Many of the interactions that make up our social environments can be characterized using digital data, yet there are distinctions to be made: digital data may reflect *digital* behavior (e.g., content of social media posts, number of social media followers) or *nondigital* behavior (e.g., census reports, hospital data), which in turn may reflect the consequences of human behavior or activities (e.g., traffic-related air pollution). Whatever the target, digital data offer a great potential for the study of human ethology. Some sources are, however, underused due to lack of knowledge about what is available, methodological complexity for using, and restricted access due to issues of user privacy and industry ownership. For a sampling of digital data sources that can be used to study human behavior, see Balsa-Barreiro and Menendez (this volume), with ethical considerations discussed by Lovasi et al. (this volume) and Medeiros et al. (this volume). We begin by considering two examples:

First, digital data can be used to study social networks in the physical world. For a relatively brief period of time, mobile phone networks provided an exceptionally useful source of social network data. Between the mid-2000s and the mid-2010s, when mobile phone usage became prevalent in the general population in most societies in the world, patterns extracted from mobile phone data created a rich and comprehensive image of human social networks. As the use of social networking apps for communication began to spread, however, much of the dyadic and polyadic digital communication shifted to platforms such as WhatsApp, Telegram, and Signal. Since the communication pattern on these apps tends to be opaque, the parallel use of several of these have made social network detection nearly impossible. Several mobile phone call studies allowed the recognition of a large number of social behaviors ranging from gender differences in social behavior (Bhattacharya et al. 2016; Palchykov et al. 2012; Yang et al. 2019), structural properties of social networks (Jo et al. 2014; Onnela et al. 2007), inference of demographics from communication patterns (Dong et al. 2014), and life course dependent social behaviors (Dávid-Barrett et al. 2016b). Although today it is far more difficult to acquire common behavioral patterns from mobile phone data, these can still be useful for extracting mobility patterns for some particular communities based on demographics and socioeconomic factors by using different aggregation levels of data (Pullano et al. 2020; Valdano et al. 2021).

Second, digital data can be used to unravel society's response to a pandemic. COVID-19 presented wide-ranging challenges (e.g., scientific, policy, economic, and behavioral), and there was variance in society's response to COVID-19 restrictions and expectations. As the scientific community raced to develop vaccines and therapeutics in record-breaking time, policy makers grappled with how to communicate and influence sociopolitical-economic decisions that could require individuals to take uncomfortable decisions. To inform the ethology of a society's response to a pandemic, it became important to leverage digital data. Krieg et al. (2020) leveraged several streams of digital data, including COVID-19 case data, demographic data, longitudinal news and web search trends, media bias data, and mobility reports to inform an understanding of society's response, norms, attitudes, and beliefs.

Social Media Data

We recognize the importance of digital data in general and their usefulness in learning about an individual's social environment. Here, however, we focus on social media data, which is a subset of the larger digital data, because of the relatively novel complexities involved in using such data to study human behavior. We consider social media data to be the traces of interactions between individuals and groups in text, voice, image, or video format taking place over the Internet (Sarker, this volume). These social media data consist of the posts as well as the metadata about posts and their authors obtained via application programming interfaces (APIs) offered by the social media platforms or via web scraping from the platforms' user interfaces.

There is a wide diversity in social media platforms in terms of the types of interactions that are enabled, the types of media that can be shared, and researcher access to that data and metadata. For instance, Twitter/X has a limit of 280 characters per post whereas Facebook, LinkedIn, Reddit have no character limits. While other platforms' main post type is text, Instagram is image based. Instagram users can include text captions, but they must accompany an image or video. Between platforms, there are also differences in how users can interact with each other. On Facebook, connections are largely bidirectional; if you "friend" another user, not only can you see their posts, but they also become your friend and can see your posts. On Twitter/X, however, relationships are unidirectional: you can "follow" another user, but they may not follow you back. In terms of access, Twitter/X had served as a favored platform for academic research due to its widespread accessibility as most posts are public. Additionally, Twitter/X offered a powerful API for accessing the posts and author and post metadata, including geolocation, though free use of this API on Twitter/X has been restricted. By contrast, Instagram has a much larger user base (1.5 billion vs. 425 million) (Statista 2022a) but offers only a limited API (via CrowdTangle) for researchers to access posts or metadata. Dong et al. (2017b) used some of the structural differences among social media platforms to uncover three main superfamilies of platforms, based on how users develop connections with each other. For instance, this explained how social networks developed via Facebook are different from those developed via LinkedIn.

Estimates suggest that globally over 4.26 billion people (around 58.4% of the global population) currently use social media (Statista 2022b). Consequently, the digital footprint of collective human behavior on social media is enormous, leading to a plethora of information on many topics of interest. The utility of such data was realized by the social media companies and the advertising industry as it provides insights about user-level and group-level interests and can be used to conduct targeted advertising. More recently, the utility of social media data has been realized by researchers with noncommercial interests. Data from social media sources have been used in different fields of knowledge, such as public health and social sciences. In public health, for example, social media chatter has been leveraged to study and detect infectious disease outbreaks (Hossain et al. 2016; Ting et al. 2020; Tsao et al. 2021) and adverse drug reaction patterns (Bulcock et al. 2021; Sarker et al. 2015).

Social media can present rich individualized or aggregated data about individuals, communities, and society at large. New data-harnessing technologies allow us to capture individual behavior and activities across a variety of social media platforms, and to link or integrate those with aggregated data generated from public record platforms (e.g., census records) or to combine the signals derived from social media with traditional survey instruments. This provides a unique opportunity to explore human behavior, attitudes, beliefs, and how they cascade, but it also opens possible pathways of risk (e.g., privacy invasion, bullying, or stalking). There are technical considerations involved in such linking and integration, as well as important ethical issues; for further discussion, see Lovasi et al. (this volume) and Medeiros et al. (this volume).

As researchers, we must ponder about how and when to use social media data, for what purposes, and what reliable methods and results could be derived. The normative question becomes: What is the focus of our study? Should we study the humans who create the content, and how attitudes, beliefs, and opinions develop or cascade as a result? Should we study the object of conversations (e.g., social media chatter on drugs) and the side effects that emerge? Or should we study some social network phenomena on how links emerge or how information flows on a social media platform?

Framing the normative question that guides data collection and research process is essential to determine whether the use or data sample derived from social media is sufficient for the research method and the conclusions that emerge. While social media holds the promise of large sample sizes (large *N*), it also presents the challenge of not knowing who (or what) it represents. We romanticize the idea of data availability at scale, but just because data are available and potentially accessible, it does not mean that data are sufficient to address the question being considered, and we lack a formal definition of sufficiency. We do not attempt to define this here, but rather aim to highlight that it is important to raise such a definition to inform the use of data. For a discussion of the use of such large-scale datasets, see Kum et al. (this volume).

Advantages to Using Social Media Data

From the perspective of research, social media data present several advantages compared with traditional data sources, as evidenced by the following examples:

- *Reach*: Social media potentially offers greater reach compared to other platforms or data sources. Social media adoption is globally at an all-time high. Many hard-to-reach populations (e.g., refugees, people without health insurance, victims of violence, people with disabilities who are unable to leave home) can make their voices heard through social media. Social media-based studies can include data generated from such populations, who may not be accessible through any other channels.
- *Size*: Social media data are massive. Thus, it is possible to generate reliable population-level insights for the population of social media users studied, though there are limitations to this, as will be discussed in the next section.

- *Speed*: Specific APIs make social media data available in real time or close to real time. These insights can be crucial for many studies, particularly in the space of public health, where it can be used to detect the outbreak of infectious disease faster than other sources. Social media is a compelling source for use cases that require scale and speed, which traditional surveys might not be able to provide.
- *Capturing emergent knowledge*: Social media data are constantly being updated, so emergent knowledge can quickly be captured. For example, if we are collecting streaming social media data and identifying the topics of discussion, we may suddenly notice a new topical construct that emerges. This can be a quick indicator of a change, possibly generated by an exogenous event of concern and can serve to be a leading indicator of a phenomenon. Kryvasheyeu et al. (2016) evaluated how online social media contributes to rapid assessment of disaster damage by improving situational awareness, facilitates dissemination of emergency information, enables early warning systems, and helps coordinate relief efforts. Similarly, Sarker et al. (2020) demonstrated the utility of social media in characterizing acute COVID-19 before widespread knowledge about its symptom spectrum was available.
- *Anonymity*: Social media often allows people to share information anonymously. Hence, discussions about sensitive topics (e.g., substance use, intimate partner violence) are frequently available on social media but often not available from other sources. Anonymous online data, such as Google search queries, can be more reliable indicators than answers to survey questions. For example, Google search queries were used to characterize the racial animus in the years leading up to the election of Barack Obama as president of the United States in 2008 (Stephens-Davidowitz 2013). Because of the sensitive nature of the behavior under study, it could be difficult to obtain truthful answers on a survey.
- *Cost*: Collecting data over social media is typically much cheaper than traditional methodologies (e.g., surveys). This is particularly true at the national or international level. Conducting national surveys, for example, can be very expensive, whereas social media data can be collected at little cost.
- *Breadth*: Traditional instruments, such as surveys, only collect information about the questions that are asked. Because the information shared over social media is not constrained by such questions, the breadth of the information can be much larger and may enable deep, longitudinal studies on the evolution of culture, behavior, opinions, and beliefs.
- *Discovery of knowledge using natural language processing*: Advances in the broader field of data science, particularly natural language processing and machine learning, have created new opportunities in social media-based research. Natural language processing may allow inference of knowledge that is not explicitly encoded in the metadata. For

example, even when geolocation information is not explicitly present, mentions of locations by a specific social media user can be identified and/or extracted using named entity recognition methods (Batbaatar and Ryu 2019; Chen et al. 2018). Meanings of expressions, including nonstandard or colloquial expressions, can be inferred by advanced natural language processing and machine-learning methods.

- *Collective information*: Social media can help identify common issues faced by groups or communities. For example, by understanding the challenges faced by individuals, we might be able to study substance use, depression, or drug side effects: what interventions work and how supportive communities form. In public health research related to substance use, insights derived from social media data in the United States have been validated against traditional sources of information, such as overdose deaths from the CDC Wonder database, the National Survey on Drug Use and Health (NSDUH), and the Nationwide Emergency Department Sample (Sarker et al. 2019; Yang et al. 2021). Compared with some traditional survey-based instruments, social media-based insights may be better representative of population-level behaviors because they integrate marginalized groups who may not complete surveys. For example, Yang et al. (2021) showed that gender distributions for opioid use, estimated from Twitter/X geodata in the United States, had better agreement with emergency department visits for opioid use related injuries compared with the NSDUH estimates. This ability to infer population-level insights for a specific geolocation has been shown to hold even for anonymous social media channels, such as Reddit (Harrigian 2018). At a country-level scale, Nigam et al. (2017) leveraged social media data to determine the outcome of the Colombian peace process and infer the underlying challenges or pain points of the population (Madan et al. 2010).

Challenges to Using Social Media Data

While the advantages described above make the use of social media data appealing, there are numerous challenges associated with the use of such data. We present a non-exhaustive sample below:

- *Presence of bots*: Digital data from social media can be used for prediction and analyses, but bots or fake posts can influence such tasks. At the individual level, particularly, bots can improperly influence analyses or predictions by contaminating the data collected. Relying on group-level data (e.g., posts from many users) can mitigate this problem. Some recent studies have also proposed methods for detecting bots automatically (Davis et al. 2016; Davoudi et al. 2020;

Sayyadiharikandeh et al. 2020), which may allow for their impact to be removed prior to analysis.

- *Making individual-level inferences*: Individual-level inferences should not be made from the data because data are incomplete and may even be false. For example, an individual post from a certain geolocation may be fake or posted by an automated account (i.e., a bot). At a group level, or with aggregated analyses, it is possible to mitigate some of these problems or risks. For example, if 10,000 posts from the geolocation are analyzed, it is likely that the number of fake posts, and consequently their influence on the overall inference, could be mitigated. Similarly, while missing data at the individual level can largely constrain our understanding of an individual, aggregation of large data can fill the gaps left by missing data at the individual level and help obtain more reliable population-level insights.
- *Natural language processing*: Most data available from social media are in free text format. The language of social media is often colloquial and contains nonstandard expressions and misspellings. While advances in natural language processing and machine learning have made it easier to derive knowledge from social media posts, the methods are not perfect, and in most cases, not even near perfect, especially with the nuance often present in communication using social media. As a result, knowledge is often not accurately detected or extracted from social media data.
- *Generalizability*: Conclusions derived from social media are typically not generalizable to the entire population of a given location. People on digital social media generally skew younger. Often, they are more tech savvy compared with the general population. The demographic representation also varies between social media platforms. For example, Facebook has a larger representation of older people, whereas TikTok is more popular among younger people (Auxier and Anderson 2021). These limitations must be established in any study and boundaries provided for the use of any insight or finding that emerges from the study. Further, social media data does not offer the opportunity of a deep understanding that might emerge from longitudinal ethnographic studies that stem from immersion into a community.
- *Representativeness*: Related to generalizability, representativeness refers to whether the characteristics of the sample population captured in the data are considered to reflect accurately the characteristics of a larger population from which it is drawn. Determination of sample representativeness is hampered by the fact that key demographic (e.g., age, sex, gender, ethnicity) and socioeconomic (e.g., income, education, employment status) information is often missing on the subpopulation captured in social media sources. In addition, to determine representativeness in social media data, one must carefully consider what

is the largest population that the data are meant to represent (e.g., a particular community or the society as a whole). Certain groups are excluded from access to social media, and thus no social media platform should be used to represent a population as a whole (Blank and Lutz 2017). In particular, individuals with a higher socioeconomic status and Internet use skills are typically overrepresented in social media (Hargittai 2020). Assessing the representativeness of social media is a moving target: social networks evolve continuously as do the populations that use them. While social networks were mostly popular among younger people, larger numbers of older people are gradually adopting them. Hence, a specific social media platform, such as Facebook, does not necessarily represent the same population now as it did five years ago nor will it represent a similar population five years from now. Unanswered questions about representativeness do not necessarily diminish the utility of social media in digital ethology research, but researchers need to be mindful of this when leveraging social media data.

- *Unknown denominator*: While population-level behaviors can be studied using social media data, a major obstacle to conducting epidemiological studies using data from social networking platforms is that the denominator is typically not known. For example, while nonmedical use of prescription opioids can be detected from social media data and the relative volume of nonmedical use can be assessed, the total number of people who report using opioids for medical purposes remains unknown. Adding to the complexity, the proportion of people who consume opioids and report this on social media is also unknown. To date, we have no specific strategy to overcome this challenge and need to be mindful of this characteristic.
- *Ill-defined control groups*: Many studies require an intervention/experimental group and a comparison/control group. Currently, however, there is no well-defined mechanism for generating control or comparison groups from social media data. While observational studies of virtual cohorts can reveal group characteristics, there are no meaningful ways of comparing these characteristics with other groups. For example, while it is possible to create a virtual cohort of people who use opioids and study group-level patterns from the data posted by the cohort, the patterns may not be meaningfully compared with a control group. While a virtual cohort of people who never report using opioids can be created relatively easily, there is no guarantee that the members of the comparison group actually never used opioids nonmedically in real life. Rather this group would represent those that do not report on nonmedical use of opioids on social media. Nonetheless, it is also important to note that issues of gaps in reality and reporting are prevalent in most population studies (e.g., surveys only represent what participants choose to reveal, and emergency department data analysis

only represent those who went to the emergency service being studied) and measuring real truth is a fundamental issue in all research. Noting the limitations, learning what one can, and being thoughtful about the interpretation and inferences being made is most important.

- *Missing data*: Social media-based studies, behavioral or otherwise, can only incorporate information that is reported by individuals voluntarily. It is impossible to determine, particularly at the individual level, what information is and what is not reported. Additionally, many social media platforms allow users to edit or delete posts or may ban users, removing their posts from public view. This may not be an issue for studying group-level behavior, but many prominent public figures, including politicians, have deleted embarrassing or incriminating posts or have been banned from social media platforms. Especially for government figures, deletion of posts or account bans can impact the digital preservation of government public records (Kriesberg and Acker 2022). Some original social media posts may be found in web archives, but due to the prevalent use of JavaScript, many social media posts are difficult to archive (Bragg et al. 2023; Brunelle et al. 2016; Garg et al. 2021, 2023).
- *Self-editing*: Researchers should be cautious about taking social media data at "face value." In a personal profile, people may project their lives by posting what they want others to see, typically the most positive aspects of their lives (e.g., their most attractive photos on Instagram). Self-editing also means that people will share different pieces of data on different platforms, such as professional details on LinkedIn, but nothing about family (Hollenbaugh and Ferris 2015). Self-editing is not only restricted to limiting the type and amount of information that is shared; it also includes dishonesty. For example, photo filters can be applied to make one look more attractive, and people lie about various aspects of their lives (e.g., height, number of sex partners) to show that they are happier than they are in reality.
- *Legality/privacy*: The sociopolitical-legal structure informs the use of the social media platform. Different countries or cultures have different permissible uses or activities that can be done on social media platforms; this directly limits the replicability or reproducibility of the work. There are also risks involved when linking social media data that may have been considered by the author of a post to be anonymous with other sources of data that might personally identify the author. There is, thus, a particular need to study potential risks in parallel to any study utilizing social media data. It must also be noted that while academic researchers continuously regulate themselves from the perspective of ethics (e.g., through institutional review board reviews), the ultimate power lies with the companies that host the social networks and the ultimate risks perhaps lie with the commercial interests

of these companies. Little is known about how these companies use the data they host themselves. Perhaps there should be a greater push for transparency. For further discussion, see Medeiros et al. (this volume).

Issues of Bias in Social Media Data

Because there are many different types of bias that can be present in social media data, we have separated the challenge of bias from the above list. Even though data from social media might be large *N*, it might still be difficult to define the statistical power and mixing of potential biases. We provide an outline of different categories of bias below, though these categories are not exhaustive:

- *Selection bias*: Social media users do not typically represent the general population. As pointed out above, subscribers of social media platforms tend to be younger and tech savvy, and older populations are often underrepresented. Access to digital devices such as smartphones, digital literacy, and local policies (physical environment) also influence selection bias.
- *Behavioral bias*: Olteanu et al. (2019) described systematic distortions in how user behaviors are represented across different social media platforms and contexts. The same individual may express different behavioral traits based on the particular social media platform being used. Thus, data from one platform may contain quite different digital footprints compared with another network even though the underlying user base is similar.
- *Reporting bias*: The rate of reporting certain events on social media may deviate from their real-world frequencies. For example, social media posts may excessively amplify topics that receive coverage on traditional news media, while some topics can be underrepresented. Certain behavioral traits may also be overrepresented over social media, as people want to broadcast those behaviors to their networks (e.g., travel, exercise, dining), while others may be underrepresented (e.g., substance use). As another example, people using dating sites tend to represent themselves strategically and to behave strategically (e.g., women report lower age and lower weight than the reality, while men tend to report being taller and earning more than the reality) (Drouin et al. 2016). The distortion is so large, that despite the presence of exceptionally large datasets, the use of these for scientific understanding of human dating choice behavior is limited, apart from the fact, of course, that such dishonesty exists. Data from social media also often overrepresent extreme views on topics while underrepresenting non-extreme ones. Not all social media subscribers are equally active. Those who are most vocal are represented better by the data (Baeza-Yates 2020).

- *Group attribution bias*: This bias is more associated with the interpretation of behavioral data from social media rather than as a bias in the data itself. Often behaviors observed in individuals or groups of individuals are overgeneralized to a broader cohort to which they belong. There is also a tendency to stereotype individuals to groups in which they do not belong. Since insights from social media data are typically derived from aggregated cohorts, unique individual characteristics may be lost in favor of group characteristics.
- *Platform-imposed bias*: A significant limitation of social media data for research relates to the platform. For example, the sampling rate and algorithm that a platform provides can lead to a biased or uncertain sample, which directly impacts the method being considered and the result that emerges from the triangulation of data and methods. Thus, there is an accessibility versus representativeness dilemma.
- *Temporal bias*: Even on the same social network platform, data from different time periods can exhibit biases based on the user base of the platform, its usability, and constraints/rules imposed by it. For example, Twitter/X had a character limit of 140 per post at the beginning, which was increased to 280 characters later. The data generated on the platform, consequently, could change substantially over time. The evolution of social networking platforms, such as Twitter to X, lead to evolving biases. Thus, when using data to study human behavior, findings from one time period may not hold over time (Liu et al. 2014); it may only offer a snapshot from that specific time period.
- *Data processing bias*: Biases may also be introduced to the data when processing it to study human behavior. Over recent years, many studies have attempted to derive knowledge from user-generated social media data using machine learning and other data-centric methods. Machine learning algorithms themselves add biases when interpreting the data. Machine-learning models are vulnerable to, for example, algorithm bias (i.e., the algorithm favors specific data or is biased toward amplifying specific phenomena) as well as measurement bias (machine-learning algorithms are biased toward specific criteria).

Examples

Having discussed several advantages and challenges to using social media data to study human behavior, the following examples illustrate how social media data could be used in research.

To Test a Particular Social Behavior

In a Facebook profile picture study, social networking data collected in 2011 were used, for the first time, to evaluate whether a particular social behavior constituted a universal human behavior (Dávid-Barrett et al. 2015). Here, the aim was to assess the hypothesis that women have a larger number of close friends than men do. The study coded approximately 112,000 Facebook profile pictures for the number of people and the gender composition in profile pictures, which was used to determine close friendships. The assumption behind this methodology was that if the same behavior was detected in all populations, then it is likely to be universal, and thus it is valid to ask whether it is also genetically inherited. Finding universal human behaviors had been extremely difficult in the past, because for this to be the case, not only the same behavior should be observed in all human cultures, but manifestation of the behavior should also be within the same social context. Using Facebook allowed observation of behavior from an exceptionally large number of people in different cultures within the same platform, and thus solved both problems. A significant gender difference was found, in particular in the formation of close friendships. The pattern was the same on all continents, in line with the hypothesis that there might be an at least partial genetic underpinning behind the behavior. The dataset yielded results beyond the initial question, suggesting that life course drives social behavior on social networking sites (Dávid-Barrett et al. 2016a). The initial social media study was followed by a real-life observation of 1.2 million people in 46 countries across the world, which supported the original study's findings (Dávid-Barrett 2022b).

To Study Problems for Which Data Are Not Available from Other Sources

Social media serves as a valuable tool to study issues lacking data from conventional sources, thereby providing a voice to marginalized communities typically excluded from such data sources. A recent study focusing on opioid use disorder as discussed on Reddit (Spadaro et al. 2022) revealed insights about the concerns of patients receiving or looking to receive treatment through medications for opioid use disorder (e.g., buprenorphine). Specifically, the study revealed that people with opioid use disorder on Reddit discussed experiences and fear of precipitated withdrawal when initiating buprenorphine treatment. The study further showed that the Reddit subscribers had collectively discovered potential reasons for precipitated withdrawal, and the community discussed successful self-management strategies that worked better (according to their shared experiences) than the protocols followed in clinical settings. This study illustrated the utility of social media data for leveraging insights that addresses the true concerns of targeted communities.

Combining Social Media with Additional Data Sources

In the Tesserae project, Mattingly et al. (2019) studied how a suite of sensors could measure workplace performance, psychological traits, and physical characteristics over a one-year period. The study enrolled more than 750 information workers across the United States, who participated using sensors (e.g., smartwatch, beacons, phone agent). Shared data included measures such as heart rate, physical activity, activity patterns, and social context. Participants also shared access to their social media data (Facebook). The variety of such (unobtrusive) sensing streams for a diverse user group allowed a detailed understanding of patterns of life and activities in these people's natural environments (Robles-Granda et al. 2021). Based on naturalistic observation, this methodology was implemented to infer driving behavior, showing advantages such as the limited intervention of the researcher in the experiment (Balsa-Barreiro et al. 2019b, 2020a). The social media in this case presents an opportunity for verbal and social sensing, in addition to physical and environmental sensing which the smartwatches, beacons, and smart phones may provide. While social media sensing might be driven by an individual's self-selection bias on participating and sharing, the physical sensing could capture complementary contextual attributes that could explain or model the propensity to participate on social media or individual-/group-level outcomes (Saha et al. 2019).

To Measure Social Fragmentation

Social fragmentation refers to the breakdown in connectedness in a community. Dong et al. (2020) analyzed how income segregation determines social interactions both in the physical and virtual world. They checked preferred discussion topics in the online space according to income in some Western cities. Discussions in wealthy neighborhoods typically included lifestyle topics (e.g., travel, leisure activities), whereas in poor neighborhoods discussions were focused primarily on sports and TV shows. Balsa-Barreiro et al. (2022) investigated global communication patterns through data sourced from Twitter/X. They constructed a global network where edges linked locations when users mentioned others in different places, with edge weights indicating communication intensity between locations. Using the Louvain algorithm, they identified 14 major communities initially, expanding to 86 minor communities as the analysis scaled up, analyzing 70 million tweets by 4 million users worldwide between August and September 2019. Their study highlighted the intricate multiscale nature of social spaces based on human communication patterns. Bakker et al. (2019) implemented different measures extracted from mobile phone metadata for checking the level of integration of Syrian refugees in Turkey. Their integration was estimated based on three dimensions: social, spatial, and economic integration. This study found striking differences both in the distributions of these dimensions, but also in the relationships between them.

Open Questions

Based on our discussion of social environments and the challenges of using digital and social media data to study social environments, we list several open questions that should be considered in the future.

- Does the online social environment shape behavior as much as the physical social environment? How will the emergence of the metaverse change this?
- If an individual is devoid of an accessible physical neighborhood for living or work, do virtual interactions create a complementary set of opportunities? If so, how do we develop a union or intersection of interactions between the virtual and physical environments that collectively build the social environment? What if the individual does not have good access to technology to enable a virtual social environment?
- How does the particular online social media platform used relate to strength of relationship tie? For instance, being friends on Facebook may be more related to some physical interaction and may produce stronger ties, but connections on LinkedIn, Twitter/X, or Reddit may never meet physically, so those ties may be weaker. What factors are more relevant: time spent on the online social platforms, or number of online interactions with people that have physically met?
- How will the metaverse impact an individual's social environments? What would digital ethology look like in the metaverse? What is behavior in the metaverse? Who is the actor, or who is the behavior by? Who is the observer, or ethologist, in the metaverse?
- Could one calculate online inequality, similar to how income inequality is characterized with the Gini coefficient? What would be a meaningful metric for this inequality? Number of followers? Of likes? The scenes that someone is projecting on his/her social networks?
- Could one trace the variation in similarity (related to social fragmentation) across regions? In large regions where we collect abundant social media posts, greater diversity and heterogeneity of hashtags are expected. Yet, do these patterns unfold similarly in areas where people have varying income levels?

Conclusion

In this chapter, we have discussed the concept of social environments from various viewpoints, starting with a basic ethological definition and moving to more complex notions of social environments that humans may encounter in both the physical and virtual worlds. We considered how context, in terms of physical location, which then brings in that location's culture and history, can affect an individual's social environments. We also discussed how the virtual world can affect social environments through its impact on social capital and the ability of interactions in the virtual world to affect individuals' behavior in the physical world. This exploration culminated in a discussion of how the emerging metaverse could further affect individuals' behaviors and interactions, even more than interactions in more simple virtual worlds. With these considerations of social environments in hand, we then discussed how the vast amounts of digital data generated can be used to learn about social environments. In particular, we focused on social media data and various considerations for their use. Data scientists and others should be aware of the many challenges and potential pitfalls to using social media data to study social environments. The relative ease of data collection and volume of social media data make it an easy target for study, but researchers should be careful in making broad generalizations based on what could be individual-level data points. In closing, we hope that future studies will pursue the open questions that we identified to provide greater understanding in how digital data can be used to study social environments.

Acknowledgments

We are grateful to the organizers and staff of the Ernst Strüngmann Forum for facilitating such an enriching experience. We thank all of the Forum participants, and name those in particular who contributed during our group's discussion sessions: Kim Bard, Steven Bedrick, Megan Doerr, Guillaume Dumas, Sophia Frangou, Kimmo Kaski, Fabio Kon, Hye-Chung Kum, Claudia Bauzer Medeiros, Tomás Paus, Beate Ritz, Sven Sandin, Cason D. Schmit, Kimberly M. Thompson, and Henning Tiemeier.

5

Integrating Knowledge from Individual Data to Population-Level Data

Claudia Bauzer Medeiros, Hye-Chung Kum,
Sven Sandin, Cason D. Schmit,
Kimberly M. Thompson, Henning Tiemeier,
and Kimmo Kaski

Abstract

Knowledge integration permeates all scientific endeavors, which increasingly depend on interdisciplinary collaboration as well as on combining data from multiple sources and knowledge domains. Advances in digital ethology progressively rely on knowledge integration, which is enhanced, but also hampered, by the large volumes of heterogeneous data that need to be considered, the multiple aggregation levels to be considered, and the human expertise involved in answering research questions. Though considerable research efforts have focused on leveraging knowledge creation through data integration, many challenges remain. This chapter identifies and investigates some of these challenges, pointing out strategies toward the generation of knowledge while bearing incentives and barriers in mind. To investigate human behavior in the built, social, and/or natural environments, for example, what kinds of considerations exist when integrating individual and population data? Are big data an asset or a hindrance to such integration? Why should (or should not) researchers go through the effort of curating, documenting, and integrating multiscale data?

First and foremost, despite all the technological advances, human judgment remains a key factor in the selection of datasets to be integrated, in monitoring and validating the integration process, as well as in interpreting the results to extract knowledge. Moreover, quality factors, such as reproducibility or robustness, must be considered at all stages: data selection, design and implementation of the integration process, and result analysis. Appropriate documentation of data and processes must be ensured for fairness and reproducibility, and metadata quality is essential for sharing of data and processes. In conclusion, ethical and legal considerations interact in many complex ways, but there exist paths to move forward and overcome the barriers posed.

Introduction

Incentives behind Multiscale Knowledge Integration

Opportunities to integrate individual- and population-level data[1] to approach innovative research questions continue to expand as researchers recognize the benefits of interdisciplinary scholarship to better understand human behavior in the context of built, social, and natural environments. Although similar to the value of research partnerships and collaboration within domains of expertise, the need to combine individual data, often aggregated at multiple levels, to address some types of research questions, usually expands the number and types of disciplines and experts required to engage cooperatively in the process. Such efforts thus promote cross-fertilization of ideas and improve interdisciplinary understanding in the process of reaching shared insights.

Research projects that integrate data can also uncover new hypotheses as well as novel lines of inquiry, provide better insights about existing hypotheses and theories, and refine our understanding of observed phenomena, driving us to dig deeper to explain any differences in outcomes observed. By investigating some questions using integrated datasets, analysts can increase the ecological validity of findings and the generalizability of results. Recognizing the connectivity of different domains can provide further understanding of the structure and mechanisms operating in the complex human systems in which we observe patterns of behavior. Moreover, by linking individual and population data, the insights that exist primarily in one domain may take on broader relevance and importance.

The development of technological resources, the appearance of new platforms, and increased availability and access to digital data, including "data repositories," "data lakes," and federations thereof, contribute to this expansion of opportunities. Code repositories (e.g., GitHub), data repositories (e.g., NASA's satellite image repository), and registries of repositories (e.g., re3data.org) facilitate the identification of datasets and analysis code in different domains, which can then be reused or repurposed to answer new research questions. In addition, the development of ontologies[2]—such as LOINC for health-related measurements (McDonald et al. 2003), the human phenotype ontology (Robinson et al. 2008), and gene ontology for bioinformatics (Gene Ontology Consortium 2018)—help to support the translation and linking of data across datasets (Kamdar et al. 2019) as well as the characterization of geo-related scenarios (Huang et al. 2019).

Researchers who integrate individual and population data can benefit from using existing data (e.g., acquiring data much more quickly than collecting new data, saving time and money) and their reuse of the data may increase the value

[1] Population-level data is the result of aggregating individual data into groups that abstract some of the individual-level properties to run an analysis.

[2] We define an ontology as a data structure that organizes some field(s) of knowledge, by connecting terms to their meanings, usages, and relationships.

of existing data. In addition, reuse of previously collected data may be the only means of acquiring historical information. The recognition of researchers engaged in such efforts may lead to their identification as "connectors" and "translators" across disciplines who "think big," and increase the visibility of their domain-specific contributions in other domains. The development of broad expertise (i.e., across more than one domain) and increased professional visibility may further provide rewards in the form of increased funding opportunities, attraction of students and/or other collaborators, influence, and greater dissemination of results. The process of working on interesting and challenging research questions that require the integration of datasets and collaboration across domains can provide fun and intellectually challenging opportunities for learning with others about interdependent and multiple factors affecting outcomes that otherwise cannot be observed.

The Data Deluge and Research Questions

Integration of knowledge is always prompted by research questions, some of which can only now be answered thanks to the so-called data deluge (which, at the same time, poses new challenges to eliciting these answers). Technological advances in data collecting and processing devices have allowed massive availability of data on human behavior and activity at individual, group, community, and population levels, in different forms and storage organizations (e.g., databases, repositories, data lakes, and others). An estimate published by The Economist (2017) claimed that, by 2025, data generated per year will have reached an order of 170 zettabytes (zettabyte = 10^{21} bytes), and that it would take some 450 million years to transfer this amount of data from one place to another using the current data transfer technology. According to the same source, some 80% of these data are privately held or in hard-to-access forms; only 20% are found in various kinds of records (e.g., social or health data in registries) that are more accessible and regulated. Indeed, a wide variety of open big data sources provide select information on individuals and populations, summarized at different geographic or administrative levels (e.g., municipal, district, state, city, and country level) and by specific characteristics (e.g., age, education level, behaviors), as well as a myriad of files on conditions associated with the built, social, and natural environments (e.g., transportation, social networks, weather). The handling of such heterogeneous sources of information poses a number of conceptual and technical challenges.

Indeed, integration of data is arguably an important step in the attempt to develop new knowledge on human behavior and its constraints. While some platforms function primarily as repositories for data access, others support data collection, curation, security, anonymization/pseudonymization, as well as software tools, methodologies, and algorithms.

The key construct of science, namely the *formulation of a research question*, involves a long path to extract new knowledge through integration of various sources of data. The nature of the question will determine the choice of research method. For example, is the research question guided by an existing theory or *hypothesis* to be tested, or is the research question related to empirical *exploration* without a hypothesis? In the exploratory study approach, the focus of the research question is on the properties of the system concerning its structure, how it functions and responds to different external conditions, in that order, using the methods of data analysis, computational modeling and simulation, respectively (Kumpula et al. 2007). This exploratory process to generate knowledge from data can be viewed as a continuum such that the data analysis primarily leads to insights about structural properties or correlations between entities. After this, additional studies may provide and would be required to obtain further insight into the functions or processes of the system. The progression from poorly structured mental models to mechanistic models that capture causal and dynamic relationships in physical and/or social systems may then support simulations to answer "what-if" questions and/or predictions of likely outcomes of future experiments or interventions (Saramäki and Kaski 2005). Learning and knowledge generation is not a linear process; rather, the knowledge obtained at each step may require going back to any of the previous steps, for example, to acquire more data or to change the modeling approach. In addition, individuals and their interactions with social, built, and natural environments (including the technology) continue to change over time, which means that our understanding of human behavior and our world also continues to evolve.

Research methods can take advantage of a number of well-established statistical analysis tools that are readily available for drawing inference from large amounts of data. Computational tools may use a phenomenological approach, a statistical approach, or a holistic approach that combines both. The phenomenological approach uses methods associated with, for example, network science and modeling to analyze links between entities, functions, or processes, in search of plausible mechanisms to understand the formation of human social networks and dynamics of human behavior in them. Statistical approaches are an integral part of data science, and cover statistical analysis or modeling, in which various machine-learning methods may play an increasing role for regression, clustering, and inference. Integration of data from various sources points, however, to the need to develop novel computational approaches, methods, and algorithms to get more detailed insight into human social behavior and population-level phenomena; regardless of the approach, human expertise is generally required (discussed further below).

As West (2017) pointed out in his data-driven studies of human social systems: "The underlying laws of complex social systems are not known, yet, but they show regularities so there must be governing principles." This, in turn, signals the relevance of integrating knowledge from data in multiple scales, collected at the individual, group, community, and/or population level. Our discussion begins with an overview of the main steps and approaches for this kind of integration, and briefly delves into identifying questions while stressing the importance of human intervention. We then discuss some kinds of studies that lead to claims that may be made as a result of integration and analyze the soundness of data to support these claims. We address quality issues through the integration process and look at some of the factors that may hinder integration activities. Finally, we analyze ethical and legal questions that arise during integration and suggest future research directions for consideration.

Knowledge Acquisition through Data Integration: From the Individual to Populations

The integration of data to acquire knowledge can be seen as an iterative process that comprises four interrelated steps:

1. Defining and acquiring the data to be integrated.
2. Curating and preprocessing as needed.
3. Performing the integration through a number of strategies.
4. Performing computational analyses on the results of the integration.

This process may require backtracking to re-execute any activity, with potentially new data or strategies that may, for example, indicate the need for alternative or new data sources, or additional curation, or alternative integration methods, in which case one or more of the activities will be repeated until the researcher is satisfied with the result. In the context of place-based digital ethology, integration combines individual-level data (e.g., tabular records from administrative health databases; see Sandin, this volume) with area-level data about physical, built, and social environment (see also chapters by Smith, Lovasi et al., and Weigle et al., this volume).

Given a specific research question and datasets, results may be different and lead to distinct (and even contradictory) conclusions and claims depending on the choices made during steps 2, 3, and 4 and their interactions. This points to the need for separating the *concept* of integration from the *algorithmic strategies* used, as well as from the kind of *underlying physical storage* mechanisms (e.g., are the data in warehouses, repositories, or data lakes; are they provided through a single site or via a federation of sites or institutions). Here we will concentrate on concepts and high-level strategies and ignore computational implementation issues.

Many Names, One Goal: Acquiring Knowledge through Multiscale Data Integration

The integration of individual- and aggregate-level (in our context, most often area-level) data to derive new knowledge has been discussed in different disciplines and research domains under a variety of names and contexts. It is sometimes called "multiscale data integration," in which the scale may be associated with the geographic space (Cui et al. 2022), but may also refer to different scales in human biology for health studies (Phan et al. 2012). Multiscale integration may interweave the data with the models that were used to produce data at different levels of complexity (Peng et al. 2021). Other names include "multilevel analysis" (Snijders 2011), "combination" of individual and aggregate data (Haneuse and Bartell 2011; Mezzetti et al. 2020; Raghunathan et al. 2003), "linking" (Paus et al. 2022), or "merging" (Gaubatz 2015; Hernández and Stolfo 1998) datasets.

Regardless of the name used, the ultimate goal is to acquire knowledge and get new insights about relationships among the real-world entities being represented by the data so that we can answer research questions. Through integration, new relationships emerge (Jo et al. 2014; Monsivais et al. 2017). Relationships may be explicit, such as those between "attributes"[3] (data properties) associated with a particular geolocation in multiple domains (e.g., rural, urban, demographics, records of social or medical services). Geolocation can be further enhanced by related information, such as temperature and length of daylight (Kovanen et al. 2013), obtained from open national meteorological and geophysical registries. Nonexplicit relationships (e.g., behavior patterns in a social network) can be obtained algorithmically by using, for instance, machine-learning techniques (see section on the Importance of Human Judgment in Data Integration).

Though ideally the research question at hand should decide which data source to use (step 1 of the iterative process), other considerations, like convenience and data availability, might also influence the selection of data. Regardless, the data sources chosen will have consequences on all analyses performed, statistical and scientific inferences, as well as which claims and conclusions we are able to draw. It is crucial for researchers to be explicit and clear about what they are proposing to measure and combine, and to ensure that the data they use are relevant to the task at hand. They must also understand and acknowledge the limitations in the data, analysis methods, and strategies (see Lovasi et al., and Kum et al., this volume).

Metadata

In parallel, researchers are often concerned about issues such as data access (how can I get the data I need; how do I know whether it exists, and where), data provenance (where did the data come from, how were the files produced,

[3] An attribute refers to a field in a file record and is sometimes called a property or feature, depending on the research domain. The term usually refers to textual or tabular files but may extend to nontextual files. Examples are the name of a person in a table, the coordinates of a region covered in a satellite image, or the amplitude of a sound wave.

and by whom), and responsible data management[4] as a whole. When looking for data that may be used in a research effort, metadata[5] are a valuable resource, since they describe a file and give information on authorship, provenance, quality, access rights, as well as other fields that may help in understanding the context in which sharing and reuse are allowed. (For a discussion on metadata and its value, see Lovasi et al., this volume.) Data registries, repositories, and federations thereof always contain catalogs of metadata—albeit of varying quality—that help to find the datasets of interest therein, in line with FAIR principles (Wilkinson et al. 2016). Given all these roles played by metadata records, metadata quality is a serious issue, often ignored by researchers. Issues related to quality are discussed below; see also Lovasi et al. and Weigle et al. (this volume) for further discussion on provenance and associated quality issues for data derived from interactions of humans with and within the built and social environments.

Integration Strategies

Integration strategies (step 3) are high-level procedures that can be applied to combine data from individual to multiple aggregate levels (e.g., area levels with different spatial granularities). Each strategy is refined depending on the data being integrated as well as quality and provenance issues. The actual computational implementation requires taking additional factors into account, such as performance, data volume, data placement, and even privacy concerns. The main groups of strategies relevant to the discussion in this chapter include:

- *Fusion* combines datasets into a single one by joining them along common attributes; this is often applied to tabular data (Bleiholder and Naumann 2009; Gagolewski 2015). Overlay is an example of a fusion technique in which the data to integrate are images whose contents, in digital ethology, are combined based on geolocation (Tsou 2004). In this case, the result is a compound image, in which each pixel corresponds to a value that represents a combination of the values of pixels of the overlaid images at that location. Individual- and population-level data can be fused, based on geolocation, when each individual is connected to a place; aggregate-level data refers to a polygon that contains

[4] For a comprehensive set of resources and standards on research data management and governance, see Research Data Alliance (https://www.rd-alliance.org/).

[5] Metadata are data that describe the contents of a file to help find and characterize it at a high level so as to preclude having to open the file to see what is inside. Metadata are always textual records. Metadata standards are domain- and research-group dependent and define which are the attributes of these records. A metadata record describing a satellite image includes attributes such as information on the sensors that captured the image, the date it was taken, and coordinates covered. A metadata record on a questionnaire applied in qualitative research may contain information on how interviewers were trained, or even a pointer to a particular term of consent.

the place, for example, as described by spatial join integration techniques (Brinkhoff et al. 1994).

- *Linkage* typically does not fuse datasets; rather, they may be kept apart but linked together (e.g., using tables) to form clusters of information about a given entity. An example is record linkage, also called entity resolution, which corresponds to recognizing different manifestations of the same entity in different files, and connecting their records based on an identifier, such as geolocation. Each integrated entity becomes a cluster of records, each of which addresses a specific kind of information, from individual to multilevel aggregates (e.g., income tax, criminal record, employment history, hospitalization history, census sectors). Linkage when the identifier is not unique or does not exist is a research problem. Herzog et al. (2007) treat a different aspect of this problem, and Kum et al. (this volume) discuss some approaches to dealing with privacy in record linkage when using individual-level data.
- *Semantic integration* connects separate files via ontology links (Noy 2004) by examining the semantics of their contents. Individual- and population-level data are connected together by the concepts they have in common and, in our case, considering geographic characteristics (Huang et al. 2019). Semantic integration often results in large graphs with millions of elements. Social networks are often processed using semantic integration mechanisms, in which clusters arise due to, for example, common behavior, expressed beliefs, or discussion topics (see Weigle et al., this volume). For a discussion of behavior patterns in digital ethology and associated data, see Dumas et al. (this volume).

Since integration starts by trying to identify commonalities across the files to be integrated, it is important to assess whether all files are minimally compatible. In particular, a combination of the above strategies may need to be applied, depending on the kinds of data types to be integrated (e.g., textual data, images, data streams, graphs of social networks, surveillance videos). The following is a succinct set of questions that need to be asked to identify commonalities among two or more datasets to facilitate integration:

- Is there any common set of features/fields/attributes/properties[6] that will allow, for instance, spatial or temporal units to be integrated, or the associated entity or characteristic to be represented, such as spatial extent, geographical characterization, measured variables, or category (e.g., land-use or socioeconomic factors)?
- At what granularity were attributes collected (e.g., meters, census units, years), and how were they expressed (e.g., frequency, intensity, time it takes to do something)? Are they qualitative or quantitative? Is there any kind of conversion between qualitative and quantitative that will

[6] Distinct research domains use these names to mean the same thing.

allow meaningful comparison? What does "near" mean in a location-based system, or "frequent" in medical reports? For a discussion on spatiotemporal granularity, see Lovasi et al. (this volume).

- Are these common sets of attributes compatible: Do they cover the same or overlapping spatial regions? Do they refer to the same or overlapping temporal windows?
- Did the datasets to be integrated already exist, or were they collected for the research effort? Are they raw, or derived, or synthetic? If derived or synthetic, what code was used to generate them? Note that synthetic data are common in situations where raw (real) data are hard to get, such as to protect individual privacy (Arora and Arora 2022).

The answers to these questions may indicate the need for data curation (a step toward increasing quality) or preprocessing (e.g., to fill in blanks or missing values, or to perform conversions). Examples of preprocessing involve converting temporal or measurement units, or aggregating/disaggregating records (e.g., transforming schools into school districts). Preprocessing may also involve additional methods, such as transforming images, sound, or videos into arrays that encode them in a more compact way (also called "descriptors" in image or sound processing).

An example of the need for such questions when integrating individual and population-level data is the so-called "modifiable areal unit problem" (MAUP) (Manley 2019). In a MAUP, the level of aggregation (e.g., administrative or census units) and the shape of the units will affect integration and subsequent analysis. Indeed, there is often an underlying assumption of population homogeneity within each aggregation unit, which is not always the case. Here, it is not enough to perform linking, or fusion, or semantic integration, without understanding the fitness for use of the individual and the population data.

Importance of Human Judgment

The consequences of different approaches to integrate individual and population data depend on how and when the integration occurs. The process affects the variability and clustering of the data ultimately used in the analysis (e.g., as in the MAUP situation just described) as well as the transparency in the judgment of the investigators involved in the process. Whatever approaches are used, substantial human judgment is involved, and domain expertise is essential (see the discussion on the importance of "human in the loop" by Kum et al., this volume).

Prior to the development of big data algorithms, analysts traditionally combined data through a process that involved the identification of data for potential aggregation and undertook a careful process of data curation with domain expertise to combine only what was needed. More is not better in these situations; rather, integration typically required pulling together only the relevant parts of the data based on domain expertise.

In comparison, in the newer machine-learning approaches, the data selection process can include a wide range of data associated with the research question, but some of the integration relationships may not be known or established, or the association with the research question can be challenged. Here, more may be better. This is because the first step of data integration for these approaches is to get as much (potentially) relevant data together as possible, and then follow this step by data reduction and correlational analyses that identify relationships. In this process, the researchers face challenges to explain the data, and this process can lead to deeper investigation to identify causal relationships and sources of variability. Here, modern statistical and computational methods and techniques (e.g., machine learning) can elicit relationships that would not be identifiable in the more traditional knowledge integration processes.

In either scenario, the role of the domain experts is important for pulling together as much data as possible, for data reduction (Mattingly et al. 2019), or in understanding and validating the emerging relationships and results.

Some Typical Study Types and Associated Claims

Claims can be contextualized by the kind of studies with which they are associated, such as:

- *Descriptive studies.* These studies typically do not involve any elaborate claims and may be free from more formal statistical analyses and rely more on basic statistical methods (e.g., mean, median, distributions, confidence interval) but, ideally, include data representative of the target population. The goal is to describe what is being observed.
- *Estimating studies.* Can be seen as a deeper and more focused descriptive study, usually including formal statistical methods and inference quantifying an estimate of interest. The claim would relate to estimated effect size and magnitude. Their goal is to go beyond a simple description to look at relationships.
- *Hypothesis testing studies.* A study testing a prespecified scientific and statistical hypothesis, alternatively supporting equivalence of some kind. This study would include statistical methods; inference, estimation, and description would be included as supporting information. The claim would be very specific—for example, declaring presence of a difference.
- *Causality and mechanistic studies*. While hypothesis testing studies can be based on group-level data using population averages and correlations, this would be less likely for a study concluding causality where we would require a high degree of support from the data in order to claim that an association is a measure of a truly causal effect and not

driven by confounding factors or biases from mediating or unbalanced moderating factors. Patterns supporting a mechanism and causal effect include patterns across time or age, or dose response.

- *Normative studies.* These are studies that seek to make claims about whether observations are consistent with available prior observations. For example, normative claims using individual physiological data are familiar to most people (e.g., blood pressure is normal, or low or high relative to the reference range). Although there are no universally standardized reference ranges for human behavior, in human development some reference ranges exist (e.g., growth curves or developmental milestones, some of which vary by country). Similarly, psychologists and psychiatrists categorize some types of mental disorders and cognitive heuristics that impact behavior. In economics, there is a focus on observing human behavior by understanding choices/decisions, often in the context of preferences revealed by participation (or not) in markets.
- *Methodological studies.* These are studies that focus on demonstrating the functionality of algorithms and tools that make claims about the utility of the algorithm/process. These studies often start from defining an important computational problem that is useful to addressing human behavior if the problem can be solved with some algorithm. Here, results about human behavior may not be novel, but it is still important to demonstrate usefulness of the proposed methods through real case studies.

Evaluating the Use of Data to Support a Claim

There are at least two complementary approaches to evaluate how integrated data are used to support a claim. Both approaches address bias in research: one relies on statistical methods to check whether bias in integrated data produces biased claims; the other concentrates on methodological aspects in data collection and integration that may lead to misinterpretation of results.

In the first case, a major statistical approach is the analysis of confounding variables; namely, those in which external factors of no interest may influence integration outcomes, and thus the claims. Consider, for example, the use of environmental data to qualify the claim that "people work less than normal when it is hot." For this research question, and associated claim, consideration of the potential role of the omitted variables may explain the phenomenon (e.g., school vacations occur during the summer). Thus, temperature may not be the primary driver of this behavior, but rather the fact that children are out of school, which encourages families to take vacation and work less at the same time. In addition, high temperatures may spur government regulation when schools are open. In a more general sense, when integrating data to support claims, the analysis needs to include a process of not only validating the accuracy and reliability of the data, but also the role that different variables may have on the research question at hand, and understanding the context in which the question is posed. A sequence of models and analyses may be required to test and evaluate the outcome under different assumptions related to the role of the variables as relevant with respect to discussion of a direct effect on the outcome or for the indirect effect on the outcome (i.e., as a valid indicator of something for which we do not have sufficient data). As always, analysts must remember "garbage in garbage out," and that quality issues must be considered at all research stages. The increasing use of machine learning as part of the analysis process has spurred the development of a wide range of statistical methods to check data and analysis bias on integration results (Ntoutsi et al. 2020).

The methodological approach (Brazhnik and Jones 2007), instead, is guided by questions on steps 1–4 of the integration process presented above. These questions can only be answered when the datasets and the steps were appropriately documented, in particular using metadata. The first set of questions concerns step 1: data selection. Was the choice of datasets to be integrated appropriately justified? Did these datasets already exist, or were they created for that research effort? If they existed, why were they chosen, and how were they found? Were they included just because they exist and are big (a self-justified choice)? Are they representative of the phenomena they purportedly describe?

Additional questions refer to how these datasets and their integration were documented. Are they appropriately described as to the spatiotemporal context in which they were created/collected? Are all units that characterize them stated? Are there standards against which we can analyze the suitability of the integration strategies adopted? What were the integration strategies performed, what kinds of preprocessing was conducted (e.g., curation, unit conversion)? Are they overly described (too many variables), requiring integration via multicriteria decision analysis? Or are they under-described, which would result in a poor analysis process and unsupported claim?

We now proceed to a discussion on quality, which is directly associated with all aspects previously discussed in this chapter.

Quality Considerations

Quality considerations permeate the knowledge acquisition process, from stating the research question to the final claims. During integration, quality checks apply to the four steps previously mentioned: data collection, curation and preprocessing, data integration, and computational analysis (and the selection of analysis methods and datasets). Such checks apply to the data (e.g., appropriateness of choice) as well as to the processes involved in integration and analysis. Which quality factors should be applied, and how should they be evaluated? Here, one must remember that data quality is also defined as "fitness for use" (de Bruin et al. 2001) or "fitness for purpose," so that quality factors and their evaluation have to be specified relative to the research framework and acceptability of the results within that framework.

These factors, often called "quality dimensions" (Fox et al. 1994), include robustness, trustworthiness, generalizability, and reproducibility. When talking about big data, the term "veracity" is sometimes related to quality; namely, to which degree results or processes represent what they are supposed to. Weigle et al. (this volume) present many examples of quality dimensions associated with social media data, such as cohesion or coverage.

Here, we discuss how the integration of individual and population (area-level) data impacts the robustness, reproducibility, and generalizability of results. In particular, we offer recommendations on how to improve the trustworthiness and generalizability of results.

Robustness of associations and relationships found in integrated datasets will depend on modeling practice, measurement error, and sampling uncertainty. In all population studies, the robustness of associations is influenced by factors such as the basic model choice (e.g., structural equation vs. regression models), the degree to which model assumptions are met (such as the normal distribution of the outcome in linear models), and modeling choices, such as the number of knots in a spline regression (Klau et al. 2021).

In large-scale social media studies or large registries, some aspects of modeling are less impactful if the sample size increases. For example, a certain deviation from the normal distribution is more likely to influence results if less than a few hundred individuals are studied; very large datasets are often more robust to these assumptions (Schmidt and Finan 2018). The impact of many other model assumptions is independent of sample size. For example, any aggregated data will have to be analyzed accounting for the clustering of individuals in the study. Although standard practice, this is sometimes overlooked, in particular if the exposure of interest is based on individual-level data or if only confounders were obtained from aggregating data.

Measurement error is often only superficially discussed in datasets resulting from integration of population and individual-level data. It can occur in exposure, confounder, and outcome measures, but has been shown to impact results even if only occurring in one variable and even if datasets are large. Although measurement error is often nondifferential (i.e., associations would be weakened), it can also lead to overestimation of results. If adjustment variables are measured poorly, effect inflation is common. Aggregate-level data are often imprecise; for example, neighborhood data or measurements to model environmental data may have poor spatial resolutions. Hence, some scientists advocate careful analyses of measurement error, such that the possible degree of error is reviewed, modeled, and tested. Sensitivity analyses can be used to show the degree of measurement error that would make results disappear (Bennett et al. 2017). Good practice in the analyses of combined data with some reasonable doubt about measurement error should incorporate such analyses to quantify robustness to measurement error. The practice is, however, uncommon.

Reproducibility: In this discussion, we will follow the report by the National Academy of Sciences (NASEM 2019) and distinguish reproducibility from replicability. Reproducibility is defined as obtaining the same results with the same protocol (measurements and model) in the same population. Open science advocates have called for codes or syntax and, ideally, the data to be made available publicly or at least on request. Likewise, analytical protocols and preregistration of analyses are suggested. These protocols should be specific and uploaded to registries and are ideally presented and discussed prior to any analyses. This increasingly common practice is important and useful even if no specific hypothesis is tested. That said, a formal evaluation of the progress in reproducibility achieved by the open science initiatives is lacking. Some guidelines have been suggested (e.g., transparency and openness guidelines; Nosek et al. 2015), but it is important that guidelines do not stifle innovation.

Replicability is the capacity to obtain consistent results across studies aimed at answering the same scientific question (NASEM 2019), "each of which has obtained its own data." The so-called replication crisis (Schooler 2014) has been discussed for over a decade. Several scientists have attempted to estimate the lack of replication in observational research and some state that many research findings are "false" (Ioannidis 2005). Although such claims cannot be quantified easily, combining data not initially collected for a certain research question or using large-scale social media data raise similar concerns. Replicability of results is important to guide policy and other implementation efforts. As Lash (2022) pointed out, however, replicability should not be judged by whether two results are both significant (or not). Rather, it is the slow accumulation of knowledge that mostly guides policy; and replication endeavors are an important part of this accumulation process. Limited sample size, chance findings, reliance on statistical testing, different forms of bias, selective reporting, and publication bias severely impact the ability of researchers to replicate results. Good practice in analyzing integrated data is not different from any other form of science. Some advocate analytical protocols and preregistration of analyses, but there are reasons to assume that this might improve reproducibility but not replicability (Hicks 2021). Others advocate for the use of careful multiple testing controls to reduce chance findings and data dredging. This, however, addresses only one problem of replicability and can increase the type II error (i.e., false negatives): the most significant associations are not necessarily the reproducible ones. Replication efforts using other samples to examine an association or other findings can be part of the original investigation. The current practice in some fields, like machine learning, is to reproduce a statistical model obtained in one sample by applying it to another independent sample, typically split off from the same dataset prior to analysis, and formally examine if the same result is obtained. In this framework, an algorithm is fitted on the training data and the model performance is tested on such independent, unseen, test data. Well-powered studies could be redefined as allowing such replication. Yet, this is not a typical practice in population studies with aggregate and individual-level data, often because sample size does not permit such splitting of data and effects are commonly small.

A result is replicable if the design and findings of the original study and replication attempts are *qualitatively similar*. Because similarity of study includes the design, measurements, sampling frame, and analyses, and these assumptions are often implicit and involve judgment, replicability is almost always ambiguous if not put into context and can be highly controversial (Feest 2016). Another important facet of replicability endeavors is that they can unravel why associations differ, how variability in measurement or exposure distribution impact results. Large-scale social media or geocoded data may offer scientists the possibility to study the seeming lack of consistency and poor replicability of results, which point to cultural specificity, or the impact of measurement or design.

Generalizability of results is a major determinant of the usefulness of data. If findings cannot be generalized or extrapolated to specific, even if limited, populations or population subgroups, there is a limit to generalizable knowledge that can be obtained. Insights may still arise from studies where generalizable knowledge is not the goal (e.g., case studies), but care is needed not to over interpret the insights, especially when implementing interventions or policies. Generalizability is inherently subjective. It is conditional on a careful description of the research, not only the study population (characteristics, ascertainment, inclusion/exclusion criteria), but the exact study question, methodology adopted, and also outcome and exposure definition and assessment. Importantly, generalizability (external validity) is conditional on the (internal) validity of results. A biased finding may be reproducible even in different settings but generalizing it to the larger or any other population makes no sense. Hence a careful evaluation of possible measurement error, selection bias, and confounding is key.

Representativeness on key characteristics such as race/ethnicity or urbanicity is often used as an indicator of the generalizability of results to different populations. Such representativeness can also indicate lack of selection bias (internal validity) and that results apply widely to the general population. The degree to which a population is representative of a larger population is an indicator of sample generalizability. Without a clear sampling frame, however, representativeness may create the illusion of generalizability, for example, if the minorities included differ from minorities not sampled. Despite the appeal of representativeness, we encourage researchers to consider sampling nonrepresentative populations for certain questions. This may make sense for many reasons beyond practical ones. Opting for nonrepresentative populations can minimize bias (certain groups may be more reliable reporters), it can increase variability of the exposure, and it can help include or focus on subgroups (e.g., Indigenous or LGBTQ+ populations), which are often poorly represented in large population-based samples (Richiardi et al. 2013).

To increase generalizability, we recommend out-of-study reproducibility efforts as outlined above. Such efforts can truly help judge the extent of generalizability and further the process of evidence accumulation. Although replication efforts can best begin with populations and designs that are as similar as possible, often sample characteristics, settings, or measurements will differ to some degree. While some researchers recommend that analytical strategies and modeling practices should be kept the same in replication efforts, we argue research design can and, if possible, should be improved according to current insights. No single reproducibility study will show or refute generalizability, but out-of-study, rather than just an out-of-sample reproducibility, is needed to evaluate whether results hold in a different context and thus are generalizable. Even hypothesis-generating analyses of integrated data should aim to implement out-of-study reproducibility. In sum, replicability and generalizability are not tested, but depend on the quality of research that is carefully evaluated in a complex and often slow process.

Barriers to Multiscale Integration of Individual and Population Data

While most of this chapter focuses on the many benefits of multiscale integration, we must also consider some of the barriers. Table 5.1 summarizes some of the main incentives that influence multiscale/multilevel data integration. While the benefits that appear in the left column were emphasized in the introduction and assumed as given throughout the chapter, here we discuss potential disincentives listed in the right column.

While data reuse may come with savings in time and resources needed to collect new data, the process of understanding the study design and data selection processes that led to the reused datasets and obtaining these datasets may also require substantial investments of time. The failure to understand sufficiently the domain expertise that led to some of the data runs the risk of producing invalid results (see Lovasi et al., this volume). This implies further risk of the research becoming an example of "bad science" (Ritchie 2020) with potential criticism from domain experts and affected communities who may assert that the researchers did not sufficiently recognize the domain context and the need for relevant expertise, which can present a reputational risk to individuals, the group of collaborators, and any institutions with which they affiliate. Recognizing this possibility at the beginning of a project may lead to the need for an expanded research team with additional expertise, which implies the need for up-front resource investments to create new or negotiate expansion of research partnerships (e.g., to enable intentional and purposeful stakeholder involvement using value-sensitive designs; Friedman and Hendry

Table 5.1 Incentives that promote or hinder the integration of data.

Promote	Hinder
• Better insights, knowledge • Technological resources, tools, ontologies, statistical methods, GitHub, code sharing, open data, repositories, data lakes • Opportunities to innovate • Partnership, collaboration • Fun • Ecological validity, connectivity, generalizability, relevance • Quality, deeper, refined understanding • Collaboration • Refined understanding • Funding, reward, efficiency, value in influencing activities • Internal collaboration, visibility of research for expansion • Interest in interdisciplinary scholarship, publication, broader recognition • Domain expertise expansion	• Technological barriers • Legal agreements, divergence/complexity • Lack of expertise • Time pressure, time taken • Bureaucracy • Discrimination: communication, publication • Promotion of interdisciplinary work (independent vs. collaborative work) • Data quality • Collaboration • Risk of invalid results, domain context, expertise • Possession of data, fear of discovery • Stakeholders expanded • Fear of trying • Head in the sand, research suppression • Lack of recognition • Restrictions due to nature of data • Funding, variability across domains • People, training • Peer group • Demand for technical support, inertia associated with sharing data • Lack of training, opportunity costs • Misperception of costs, risks, benefits

2019; McIntyre 2008; Viswanathan et al. 2004). In addition, the nature of the datasets to be integrated may expand the size and number of stakeholders, specifically affected communities interested in engaging in the research process in some capacity, and may come with some restrictions that include ethical, legal, and institutional review. For example, if the research involves using data that are subject to a data use agreement, then the process of reusing the data may require negotiation with those involved in the specific data use agreement, and navigation of complicated and potentially divergent interests. In addition, if the data use agreement precludes sharing the data outside of the collaboration, then this may restrict options for publication of the results to journals that do not require deposition of the data into a repository.

Ethical and legal requirements can preclude the sharing of data at the same level of granularity (e.g., across country borders), leading to both bureaucratic and methodological challenges when collaborating researchers have access to different levels of detail. Interdisciplinary collaboration may introduce another complicating factor when distinct disciplines adopt noncompatible data sharing and reuse policies.

The identification of datasets for potential integration (step 1) does not mean that the researcher will gain access to the datasets in a usable format (or at all). Specifically, not all researchers (or, for that matter, institutions) share data. This may reflect their compliance with agreements they made to collect or assemble the data, interests in protecting data that they are actively analyzing or expect to analyze once the data collection ends (e.g., for a longitudinal study), or simply because not sharing data maintains control of further discovery, evaluation, and communications of the data and prevents misuse or uses that might harm the reputations of those who possess the data (e.g., discovery of errors in the data). Similarly, restricting access to data to prevent discoveries by others may reflect the preferences of some data owners to maintain the uncertainty and ambiguity that comes from lack of analysis, because providing data to independent researchers might lead to real or perceived risks of negative attention. For example, analyses of integrated datasets may result in identification of previously unidentified issues that some stakeholders may prefer not to become aware of (an attitude summarized as "head in the sand" in Table 5.1), lead to claims that require further resource investments, or create new risks for the data owners or stakeholders. In this regard, research that integrates data that may directly affect the activities of one stakeholder may encounter active research suppression efforts by others. Resistance for sharing data may also be due to fear of data misuse—the so-called dual use issues (Bezuidenhout 2013).

In spite of the development of tools, platforms, and advances in technology, research efforts that integrate individual and population data may encounter technological hurdles related to the nature of the datasets, issues with data quality, challenges with incompatibility between platforms and software used for processing data, inappropriate and/or insufficient ontologies required for coherent understanding of the concepts behind the data, insufficient data documentation, and computational demands that necessitate the engagement of technological or computational expertise in addition to any subject matter expertise. For example, while open data repositories mean that researchers may access datasets simply by downloading them, lack of documentation on these data, such as poor metadata, may render them unusable.

Along these lines, researchers who are willing to share data can upload the data with different levels of processing (e.g., raw, curated, derived) and their responsibility for data sharing ends with depositing the data into an adequate repository. Nevertheless, good data management practices, together with FAIR properties, require that datasets be documented by use of appropriate metadata records. Adequate documentation is itself a time-consuming activity that goes largely unrewarded, yet another barrier to good practices in data sharing.

There is, moreover, an expectation from researchers who want to reuse the data that the depositor of the data is responsible for answering questions, producing details about the data, essentially providing free technical support to potential data users. Since this kind of stewardship is seldom available, this means that those seeking to use data from a repository may need to at least attempt to establish a collaboration with the data collector or generator and/or engage others to ensure appropriate interpretation of the data during integration. Alternatively, if data repositories come with expectations of perpetual stewardship of the data and responsibility for spending time helping any and all potential users, then this may create a disincentive for depositing data for reuse, or deposit only "self-explanatory" data, when research projects would potentially benefit from a more complete dataset.

At an individual level, engaging in research that integrates individual and population (e.g., area-level) data as part of a collaboration will likely mean sharing credit for the work. This may have substantial career implications for new and less-established researchers whose scholarship and promotion are judged by their independent contributions, and who may not receive sufficient recognition for their contributions as part of the team. In addition, the dissemination of the results may come with challenges related to communication of added complexities associated with the multiscale data integration, and difficulties finding an appropriate journal and/or opportunities to publish in high impact journals that may view the work as not a good disciplinary (or domain) fit. Similarly, the people who developed the original idea and intellectual property are rarely acknowledged, even though they managed to obtain funding, and performed data collection, cleaning, and storage to a level that would allow other researchers to use and access the data later are rarely acknowledged. This lack of acknowledgment may hamper data collection and sharing more broadly. The group dynamics of collaborative activities can provide a substantial disincentive and discourage even attempts to engage due to real or perceived pressures that researchers face to meet productivity targets ("publish or perish"), secure funding for research outside of established domain-specific funding streams or in domains with variable or little funding opportunities, and opportunity costs associated with investing in additional training and acquisition of staff with less-familiar skills and expertise.

The results that may come with the innovation of research that integrates multiscale data may also face challenges due to the absence of peer groups, or to experts in related domains who may perceive the research as a threat. All research projects come with some risk of failure (e.g., not resulting in outcomes worthy of publication or further pursuit), but some unique pathways of failure come from combining individual and population data. For example, the effort may fail after substantial investments in the up-front activities that lead to the integration process if the collaborators determine that the quality and fitness of the data when integrated do not support the analysis required to answer the research question. Those who perceive this and other risks as potentially very substantial may fear even trying to engage in this type of research. As with any research activity, individual researchers may misperceive the risks, costs, and benefits of participating in activities that integrate multiscale data, particularly in the context of evaluating the opportunity costs. With time, as more efforts either succeed or fail, the risks may become more easily understood.

Ethical and Legal Considerations

Ethics and law are essential tools when making decisions about data use, but they are different constructs that provide different types of answers (Hulkower et al. 2020). If the question is, "Can I use these data?" ethics will help distinguish whether the answer is "right" or "wrong" or "should" or "should not." In contrast, the law helps distinguish between "yes," "no," or "maybe" and answers of "must" versus "may." It is also essential to recognize that ethical activities might not be legal (Hulkower et al. 2020) and that legal activities might not be ethical. Legal and ethical issues on data integration and use must be important considerations in determining whether a research project can or should proceed. Here, we focus on two critical considerations of integrating knowledge from individual and aggregate-level data: *group harms* and *legal uncertainty*.

Group Harms

In the ethical review process, the overwhelming focus is on the mitigation of individual-level risks. These risks are well documented, and research ethics committees are accustomed to weighing these risks against the perceived value of a proposed research project. To this end, the principal strategies include seeking an individual's consent, where practicable, and de-identification (for definitions on distinct forms of data privacy, see Table 1 in Kushida et al. 2012). Informed consent rests on the idea that the individual is best situated to evaluate the risks and benefits of participating in a research project. De-identification rests on the assumption that rendering individuals more difficult to identify will reduce the risks faced by those individuals ("data subjects"). Both strategies can, however, be legitimately criticized in big data contexts. When integrating large datasets involving individual- and aggregate-level data, the objective is often to gain insights about groups of people with similar characteristics (e.g., their geospatial location at a particular level of spatial granularity). These insights—well-meaning or not—can lead to substantial harm to these groups and the individuals within them. Thus, research that uses big data, especially when it involves integration, implies a different type of risk that is largely ignored by research ethics committees: group harms (Ienca et al. 2018).

Group harms are those harms that adversely affect the collective interests of individuals sharing common characteristics (Xafis et al. 2019). Some of these groups might have legal protections (Wachter 2022); for example, in the United States, various antidiscrimination laws protect racial groups legally. Other groups might have substantial predictive importance but lack any protections under the law. For example, owning a dog is an important grouping characteristic used by many data brokers, yet "dog owners" is not a legally protected class under U.S. antidiscrimination laws (Federal Trade Commission 2014). Still other groupings, such as those derived through artificial intelligence, are entirely incomprehensible to humans (Wachter 2022). These incomprehensible groupings might include, for example, individuals with specific mouse movement patterns, or specific web-browsing behaviors (Wachter 2022).

Using de-identification or aggregation may protect the individual data subjects, but it shifts the focus of analysis, and the risks that come with it, to an identified and identifiable group. As a consequence, the grouping might aggravate risks for group members. For example, data aggregated using racial grouping criteria could facilitate erroneous stereotypes of that group and discrimination. Behavioral insights about a group like "dog owners," mentioned above, could enable harmful and potentially legal discrimination against individuals within the group. Also, de-identification may be meaningless as a privacy protection mechanism to individuals whose identity is strongly linked to the group they belong to, as is the case of many Indigenous groups, for which specific data governance principles exist (Carroll et al. 2020).

Similarly, consent is an imperfect tool to manage group harms. An individual who provides consent to research could face minimal individual risks, but the group the individual belongs to could face substantial group harms. For example, genetic data can be collected with minimal risk to an individual, but the use of the genetic data can have far-reaching impacts on the individual's family, community, and even culture, as was made painfully clear when genetic information from the Havasupai Native American tribe was used for research that caused significant cultural harm, stigma, and embarrassment. Genetic data collection is also a good example of another kind of group harm: by being "aggregated" into a group, the individual may not only incur harms–other members of that group, and sometimes even outside the group, may be harmed as well (e.g., allowing discovery of new knowledge through use of bioinformatics).

Moreover, most individuals cannot fully know or appreciate the implications of their "consent." For example, most Meta (Facebook) users might not appreciate that the broad consent they provided to Meta permitted widespread emotional experimentation on vulnerable social media users (Reilly 2017). An individual's ability to protect against group harms through withholding consent depends substantially on the individual's awareness of the group(s) they belong to.

Importantly, aggregation and grouping decisions during integration steps 1–4, described earlier in this chapter, can affect the distribution of group harms. Individuals and the communities they belong to have a right to be counted (Fairchild 2015). This right derives from the fact that information can empower individuals and communities to act. For example, the discovery that an industry is harming a community empowers the individuals within that community to act to seek new regulations for the industry; that action would not, however, occur but for the knowledge of the harm. Similarly, the act of counting informs crucial resource allocation decisions. Consequently, inequitable counting begets inequitable resource distributions. In the extreme, inequitable counting can lead to so-called data genocide, whereby the undercounting of a particular group contributes to systemic exclusion of a group (and eventual extermination) (Urban Indian Health Institute 2021). For example, a 2021 report by the Urban Indian Health Institute alleged that inadequate reporting and sharing of COVID-19 surveillance data with tribal communities and governments contributed to ongoing data genocide of American Indian and Alaskan Native populations. For these and other reasons, great care should be taken to ensure that aggregation and grouping decisions do not contribute to systematic and inequitable disenfranchisement of vulnerable groups.

Critically, the group harms can extend beyond the specific subject matter of the data being aggregated or integrated. For example, consider a research project on school performance, where researchers report only aggregated student performance data at the school level to protect individual students. Although the reported data concern only specific schools, there might be group harms that extend beyond the study's focus. Neighborhoods surrounding poorly performing schools might see falling property values and increasing community stigma. Since the neighborhood residents were not the focus of the study, they might not have had an appropriate opportunity to raise their concerns with the researchers. In this way, researchers and research ethics committees should consider what groups, internal and external to the research focus, could face group harm from the research activity and weigh the risks and benefits to both individuals and groups accordingly.

Seeking a "social license" from relevant communities or groups is one approach to address potential group harms. Social license refers to the informal permission given by a community to a public or private entity to engage in a specific activity (Shaw et al. 2020; see also Weigle et al. this volume). In the context of digital ethology and other big data activities, a social license provides legitimacy to collect, use, or share data that is tied to the data subjects' communities. Additionally, the social license helps establish credibility and builds trust between the parties (Jijelava and Vanclay 2017). Careful and appropriate community consultation and engagement (Dickert and Sugarman 2005) can help develop a social license (Corscadden et al. 2012). For example, in the context of public health surveillance, the World Health Organization (WHO 2017) cites community consultation and involvement as one way to support ethical surveillance activities.

Legal Uncertainty

There are multiple dimensions of *legal uncertainty* in digital ethology and big data generally. First, the technology to easily share digital data across great distances has existed for decades, but laws often make collecting, accessing, sharing, and using data exceptionally difficult in practice (Schmit et al. 2019). Laws vary across jurisdictional lines, and organizations interpret and operationalize laws into their internal policies in a variety of ways. Moreover, laws can regulate different types of data (e.g., health, census) or certain data activities (e.g., research, public health) differently (Schmit et al. 2022). These differences in laws must be carefully navigated when data that are regulated by different laws are integrated. This complexity creates both real and perceived legal barriers to data use. Consequently, the first and most challenging aspect of legal uncertainty in a data project is often understanding what legal rules apply (Public Health Informatics Institute 2021).

In addition to the legal complexity, technological innovation in big data analytics far outpaces the ability of regulators to manage new and emerging social risks. Bowman describes this problem using the parable of the race between the tortoise and the hare (Bowman 2013). In this analogy, technology is the hare—progressing at a rapid pace—and law is the tortoise—progressing at a much slower pace. When the gap between the two is too great, technological progress is impeded (i.e., the hare sleeps). This can happen when an out-of-date law is used to regulate a technological practice it was never intended to regulate, or when the uncertainty and legal risk of operating under out-of-date laws is too great. For example, the relative failure of the United States to keep pace with other countries' regulation of data protection led to the invalidation of the international data sharing agreement, the EU–U.S. Privacy Shield Framework by the Court of Justice of the European Union (Kerry 2021). This decision led to the cessation of many data sharing activities between European and U.S. researchers, and even questions concerning data transfer across the Atlantic (Hallinan et al. 2021). In this way, the failure of regulators to keep pace can interrupt scientific progress.

Rapid innovation also challenges regulators by making it difficult to define the subject of proposed regulation. Laws work by attaching legal prohibitions or permissions to words. Consequently, legal definitions of these words are incredibly important. Innovation-laden terms like "big data" or "artificial intelligence" have been difficult to define, and thus difficult to regulate. Rapid innovations in how the technologies are used make it difficult to balance precautionary risk-mitigation with appropriate room for technological progress.

Legal definitions can also lead to tremendous confusion because they can be counterfactual. A law might provide a definition for a de-identified dataset, but an individual can in fact be identified within a dataset by using an appropriate reidentification method. Here, the purpose of the legal definition is not to describe what is true, but rather to describe the thing that is subject to the law. Nevertheless, the discrepancy between legal definitions and technical terminology can lead to considerable confusion between parties—such as researchers, data custodians, research ethics committees, and data subjects. In these situations, it is important to clarify the intent of the terminology being used when describing a data activity. For example, if data must be legally de-identified to comply with the law, then the legal definition is important. If, however, data identifiability is part of the data governance approach to manage an ethical concern like risk of harm, then the legal definition is less relevant and might either overmanage or undermanage the ethical issue.

Complexities and uncertainties in law and ethics can lead to both real and perceived barriers to data use. Well-intentioned individuals can reach reasonable (and seemingly intractable) disagreements on whether a data use is legal or ethical. Successfully navigating legal and ethical issues in digital ethology requires identifying these real and perceived barriers to data use. This will in turn require subsequent negotiation among all actors involved (legal, administrative, researchers) to achieve an agreement at some level ("getting to yes").

In these challenges, lawyers have a duty to advise their clients of the legal and ethical risks of a proposed activity. Ultimately, however, clients have the decision about whether to proceed with an activity in the face of the legal and ethical risks. For research institutions, there is unlikely to be a risk-free course of action in the face of these and other legal and ethical challenges. Unfortunately, often data sharing agreement negotiations can be bogged down by organizations (or their attorneys) aggressively pursuing a zero-risk agreement, resulting in protracted delays or restrictions that are neither legally nor ethically required. Some tolerance of risks—known and unknown—is necessary to ensure that socially beneficial research continues and the key to progress may require different ways for balanced risk management (e.g., Table 1 in Kum et al. 2014) rather than risk avoidance.

Conclusions and Additional Directions

This chapter analyzed some of the factors involved in generating knowledge from multiscale data integration, ranging from the individual to the population level. As seen throughout the text, in digital ethology such integration requires interdisciplinary cooperation. Indeed, one must never forget that data integration is not "just" integration of data, but also of the knowledge of domain experts.

The role of human expertise must be emphasized all through the integration process, since there will always be limits to what technology can provide. Humans intervene in selecting and curating the data, choosing the integration strategies, analyzing and interpreting results, documenting data and metadata, and checking quality at all integration stages. Quality assessment and monitoring throughout integration planning and execution are essential. Indeed, quality questions must be embedded into integration efforts. This might even be called a "quality by design" approach, in the sense that quality must be planned for, and designed into the integration of knowledge. The need for appropriate documentation, including metadata, is a requirement for checking quality and also supporting FAIR principles. Multiscale data integration also requires navigation of ethical and legal paths and pitfalls to access, integrate, and analyze the integrated results. The associated risks must be acknowledged and considered by all actors involved in knowledge generation and governance, so that the barriers these risks pose can be overcome through cooperation.

While research collaborations are traditionally implemented through *direct* interactions among groups of researchers, the worldwide movement toward open science has introduced a new kind of interdisciplinarity in which groups collaborate through making the digital resources produced by their research (data, software, code) publicly available for reuse. This second type of collaboration, an *indirect* one, has been enabled thanks to progress in digital technologies. Here, the digital resources that are made available through, for example, repositories, data lakes, or federations, become de facto "collaboration mediators." Researchers who painstakingly prepare data to become available for sharing are assisting groups they may never meet; they are helping to solve yet-to-be-formulated research questions and, as such, are, indeed, collaborating with the future.

In this sense, open access to data and code are to be encouraged, and acknowledged, as a means of fostering scientific progress and new kinds of knowledge creation. Encouragement and acknowledgment also apply to institutions that provide resources to support appropriate data management and archival, thereby helping researchers to extend their cooperation networks. Digital ethologists whose research involves integration of multiscale data typically rely on datasets made available by others. Providing broad access and transparency can moreover foster reproducible research as well as scientific innovation in the methodologies developed, in the algorithms, in the code, and in the results themselves.

While the emphasis was on population-level data as a powerful kind of data aggregation that can help advance research in this field, other kinds of aggregation may also be considered, to which many of the issues raised in this chapter apply. This is the case, for instance, of satellite images, in which each pixel is a spatiotemporal aggregate of remotely sensed data that indicates human activity (or lack thereof). Spatialized pixels can be integrated with data on individuals and communities that inhabit that region or vicinity through use of coordinates and geo-statistics. Satellite images are aggregators of human activity, as in land-use maps, or as reflecting change in patterns of human behavior due to changes in the built or natural environment. For instance, forest fires or riverine pollution or erosion reflected in such images can be correlated with displacement of Indigenous populations, individual reports of respiratory diseases, or patterns in the spread of zoonotic diseases (Mishra et al. 2021). These are examples of aggregations that are not specifically computed as such; rather, they emerge from direct observation via the instruments used to collect such data.

Acknowledgments

The authors gratefully acknowledge the support of the Ernst Strüngmann Forum and the organizers of the Forum on Digital Ethology. Many thanks to all Forum colleagues for suggestions, comments, and stimulating discussions throughout and beyond the event.

Mapping Place-Based Context

6

Geospatial Information Technology Systems for Digital Ethology

Thomas Brinkhoff

Abstract

Today, large amounts of digital data about human activities are generated and stored in databases. These data are often geospatial (i.e., locations on Earth are directly or indirectly referenced). To analyze the digital footprint of human activities in their environment, geospatial information is essential because spatial (and temporal) proximity to events may indicate meaningful relationships. The processing, analysis, and presentation of such information require a deliberate handling of geospatial data as well as the use of suitable software tools and frameworks. This chapter provides a short review of the geospatial information technology (IT) systems that can be used for digital ethology. It introduces the main concepts of geospatial information, presents several types of IT systems for handling geospatial data, and discusses their suitability for digital ethology. Special attention is given to the handling of very large geospatial datasets, to the use of geospatial analysis and aggregation methods, as well as to the application of comprehensible visualization techniques. Besides the usage of out-of-the-box functions, more complex geospatial analyses may need to use application programming interfaces for specific solutions.

Introduction

As introduced by Paus (this volume), the objective of digital ethology is to study human behavior—as well as its constraints and consequences vis-à-vis the built environment—in the natural environment by analyzing its digital footprint. In many cases, behavior and information about the environment relate to some geographical place(s) on Earth. This statement is obvious when one considers that human activities directly influence the built environment, such as when people cover land areas with buildings or a street artist covers the wall of a building with a mural. Many digital datasets also contain direct or indirect spatial references, such as place names, postal codes, and coordinates. Thus, a suitable management and analysis of geospatial data can foster digital ethology. This mainly results from the first law of geography by Waldo Tobler (1970), who stated that "everything is related to everything else, but near things are more related than distant things." The claim "from individuals to communities and back" requires (among others) the handling of very large geospatial datasets, the use of suitable geospatial analysis and aggregation methods, as well as the provision of comprehensible visualization techniques.

In this chapter, the main concepts of geospatial information, georeferencing and geospatial data models are introduced. Discussion then follows on IT systems that are typically used for handling geospatial data, including their key properties as well as their assets and drawbacks for digital ethology.

Geospatial Information

The main characteristic of geospatial information and data is their reference to locations relative to Earth. For example, geospatial information describes the surface of Earth, refers to real-world objects like buildings and bridges, allows the planning of cities or other areas, defines abstract entities like municipal or postcode areas, or describes spatiotemporal events like traffic congestions and floods (Bartelme 2022). Objects with geospatial information are called geospatial features. The digital impacts of human behavior are often geospatial. In many cases, spatial and temporal extensions are important and can occur at different levels of granularity, with exact or fuzzy boundaries. For a digital representation of geospatial information, we need suitable data models. These models are encoded for storing, processing, and exchanging geospatial data.

Georeferencing

The spatial reference of a geospatial feature can be established in different ways (Longley et al. 2015):

1. *Names and codes*: A location is described by place names, address data, code numbers, or similar information. Common codes are postal codes as well as codes for administrative or statistical areas, such as the ISO 3166 (International Organization for Standardization) and NUTS (Nomenclature of Territorial Units for Statistics) used in the European Union. Code schemas subdivide areas and typically have a hierarchical structure. For instance, a NUTS-1 unit consists of one or several disjoint NUTS-2 units. A typical drawback encountered when place names are used from web pages, tweets, or similar sources as data is that the place names are often ambiguous or have vague boundaries (Markowetz et al. 2005).

2. *Symbolic reference*: Spatial reference is created through information that reflects a situation in a way that is comprehensible to humans and that refers to other objects. Driving instructions from navigation systems (e.g., "turn right at the next intersection") are an example of symbolic references. Further examples are often contained in social media messages (e.g., tweets, Facebook posts).
3. *Coordinate reference systems* (CRS) provide a framework that allows locations to be described as coordinates. This framework consists, among others, of a mathematical figure approximating the surface of Earth and a horizontal (geodetic) datum for assigning coordinates to points on this surface. For a geographic CRS, a position of the Earth's surface is defined by angular measures related to the equator and prime meridian of the ellipsoid (see Figure 6.1). For example, WGS84 coordinates are geographic coordinates. To display geoinformation on a flat surface (e.g., on paper or on a screen), a mathematical mapping of positions of the Earth's surface onto the plane is required and provides simpler and faster computations than geographic coordinates. Depending on the projected CRS chosen for this purpose, smaller or larger distortions may occur in terms of the location and the size of the area. National institutions and web applications often use projected coordinates.
4. *Linear referencing systems* provide another form of georeferencing that describes positions on a line feature (e.g., a road or pipeline) by a distance measure from a defined point of reference. These distances are typically stored by measures or m-coordinates.

Typically, IT systems require two- or three-dimensional coordinates to represent, exchange, and analyze geospatial data. In addition, operations on linear coordinates are often supported. Missing spatial references are a common

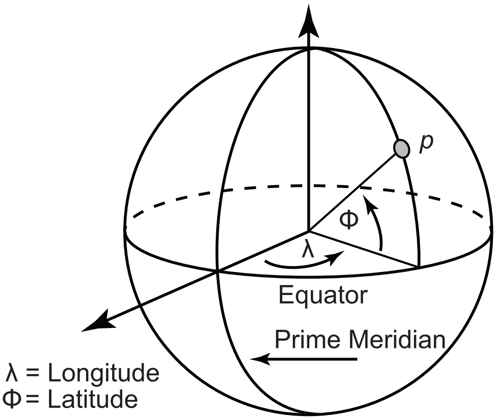


Figure 6.1 Illustration of geographic coordinates.

problem. Many photos exist, for example, that show some place on Earth but without a geotag to describe its position.

Modeling Geospatial Information

For an IT system to process real-world information, a suitable data model is required (Herring et al. 2022). To represent geospatial information, its essential properties must be considered. In addition to geometry and topology, nonspatial (thematic) and temporal properties may exist, yet only the first two properties are specific for geospatial data. The combination of space and time is of special importance because it can be used to describe the dynamics of a feature (e.g., the expansion of an urban area).

Similar geospatial features are typically grouped in layers (e.g., buildings, roads, rivers), as illustrated in Figure 6.2. Thematically related layers (e.g., all traffic layers) can form a grouped layer.

Geometry Models

Geometric properties of geospatial data are used to describe the location and extent of a place in space. As illustrated in Figure 6.3, two basic approaches are used: a vector and a raster model (Gröger and George 2022; Herring et al. 2022):

In a *vector model*, points are the base element that generates lines, surfaces, and (3D) solids. Coordinates describe the position of a point, and a sequence of two or more points creates a line. A surface is bounded by one or more closed lines, and it may have one or more holes. Figure 6.3 illustrates the vector model on the left side. A geospatial feature stores a vector geometry by a corresponding attribute.

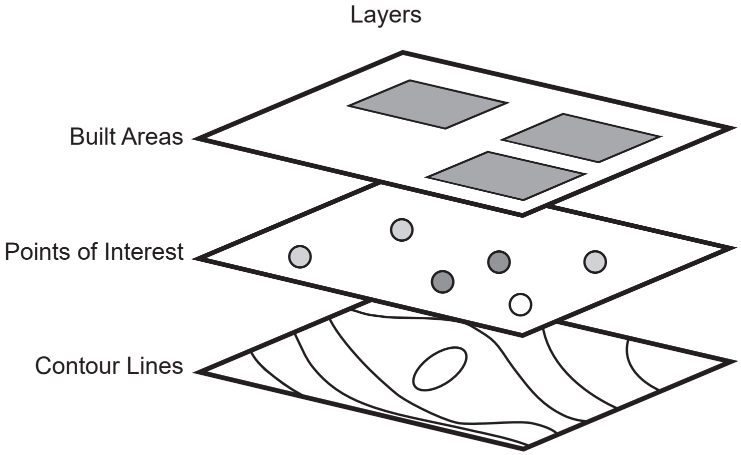


Figure 6.2 Layering of geospatial features.

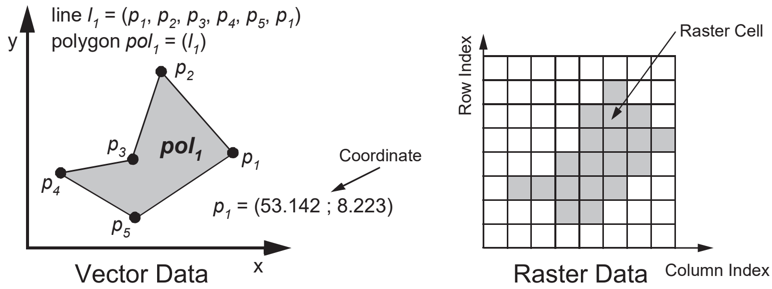


Figure 6.3 Vector model (left) and raster model (right).

In a raster model, coverages are functions from positions in space to values of some type. The most common implementation of a coverage is a raster. It decomposes the data space into similar raster cells (also called pixels), which are usually squares or rectangles and are identified by a column and row index. Each cell stores a single or composed data value. In the case of raster images, this value corresponds to color or brightness. In general, any type of value can be stored in the cells ("raster data"). The spatial reference must be established by georeferencing; for instance, by specifying the coordinates of the corner points of a raster. A georeferenced raster image is called raster map.

The properties of these two geometry models differ significantly. The vector model permits greater accuracy and better resolution scaling. In the vector model, a feature bundles an identifier, its geometry, and other properties. This connection can be used for further analyses. The raster model harmonizes well with important acquisition methods (e.g., aerial or satellite images) and output devices (screen).

For digital ethology, both models are useful. As described by Smith (this volume), vector-based administrative data and raster-based remote sensing data are important digital data sources. Balsa-Barreiro and Menendez (this volume) also list vector data (e.g., mobility patterns, census data, locations from personal wearables, point clouds from laser scanning) as well as raster data.

Topology

Topological properties describe the relative spatial relationships between geospatial features. Typical questions they address include: Which areas touch another area? Which lines intersect an area? Which lines are connected with another line?

Topological properties can either be derived from geometric properties or explicitly represented by a data model. The former will typically be used to address questions related to digital ethology (e.g., which tweets correspond to an area of interest), since typically they do not need to be answered in a precise and consistent form (Kwan 2012). Routing is, however, an important exception. To compute the shortest path or to follow a road network for some distance requires a topological node-edge model. The nodes can represent the points in space and the edges the direct connections between two nodes with their essential properties like distance or travel time. In public health, for instance, this approach can be used to delineate hospital service areas (Wang 2020). In urban analysis, network analyses are used to determine the accessibility of particular areas such as parks for neighborhoods or specific population groups (Unal et al. 2016).

Maps

Maps visualize geospatial information and allow its contents to be communicated (Kraak and Ormeling 2021). For the presentation of geospatial data in maps on screen or in printed form, styling must be defined. Vector data can be visualized using graphical surrogates (symbols). Since points have no extension, they must be made presentable by special point symbols (icons). Specific symbols may be used for illustrating the semantics of line and polygon features.

Appropriate design rules must be defined for the thematic properties of geospatial objects or of raster cells (Buckley et al. 2022). Qualitative properties, which can be represented by (finite) enumerations (e.g., place category), can be visualized by a graduated color scheme or by symbols. Quantitative properties, which originate from a (in principle infinite) number range, can be represented by a color gradient. Alternatively, intervals can be formed, such places with less than 1,000 inhabitants or places with 1,000 to 4,999 inhabitants. Nominal properties such as names and codes as well as quantitative or qualitative properties that are difficult to represent by icons or colors can be added to a map by using labels. In addition to the definition of properties like font and text decoration, the application of label placement rules is important for comprehensible maps (Been et al. 2006). Appropriate design rules can be defined for different scale ranges with respect to a layer.

Generalization is an important concept for the design of maps (Brassel and Weibel 1988) and comprises

- the selection of important information (e.g., only cities with more than 100,000 inhabitants are displayed),
- the simplification, aggregation, and/or classification of data depending on the current scale (e.g., the presentation of individuals vs. the visualization of groups of a minimum size),
- the emphasis of important information (e.g., by using a special symbol or color), and
- the displacement of features so that they do not overlap with other objects (e.g., schematic plans of transport networks abstracted from exact position and emphasize topological connections).

Well-designed maps allow a broad audience to visualize the results of an analysis: from a wide array of experts to common citizens. They help lead the viewer to draw proper conclusions and identify the next steps of analysis.

Standardization and Data Formats

As discussed by Kum et al. (this volume), data access and cleaning are important steps for data analysis. To enable a smooth data exchange or "interoperability," the provided data must be accessible through standardized models and formats (Sondheim et al. 1999). In the field of geoinformation, two organizations play an important role for the standardization of data at the international level (Kresse et al. 2022). The Open Geospatial Consortium (OGC) has established a large number of specifications and other recommendations, many of which are reviewed and ultimately published by the ISO Technical Committee 211 Geographic information/Geomatics (ISO/TC 211) as standards of the 19100 series. Important standards include the following:

- ISO 19107 Spatial Schema is a conceptual data model that describes the spatial properties of geospatial features. ISO 19136 GML (Geography Markup Language) implements this model for interoperable data exchange using XML (Extensible Markup Language). GML is often integrated into an application-specific data model. For example, CityGML (Kolbe 2009), which is a well-known OGC specification for digital city models, follows this approach.
- ISO 19125 specifies a subset of ISO 19107, especially for use in spatial databases and geospatial web services. This simple feature model defines also WKT (Well-Known Text) and WKB (Well-Known Binary) as open encodings for data exchange.
- ISO 19115 Metadata is the accepted metadata model for geospatial data (Brodeur et al. 2019). In addition to obtaining basic information like content, representation, and geometric extent, metadata is useful for accessing quality (Dassonville et al. 2002) and for determining provenance (Beilschmidt et al. 2017).

The data formats GeoPackage and KML (Keyhole Markup Language) are two important OGC standards for data exchange. In addition to the ISO and OGC standards, other encodings are used for geospatial data. For vector data, shapefiles and the text-based GeoJSON (JavaScript Object Notation), formats are most important. For georeferenced raster data, GeoTIFF is often used for representation. The identification of coordinate reference systems is mostly done by EPSG codes.

Geospatial IT Systems

Location is a central aspect of human activities. Thus, digital ethology requires IT systems to analyze, process, and visualize geospatial data. Here, a suitable selection of systems is presented and discussed with respect to their applicability for digital ethology.

Many IT systems are available for processing geospatial data. In a broad sense, each may be referred to as a geographic information system (GIS). This term, however, refers to more specific types of systems. To distinguish them from other geospatial IT systems, the term is used only for traditional GIS in the following sense.

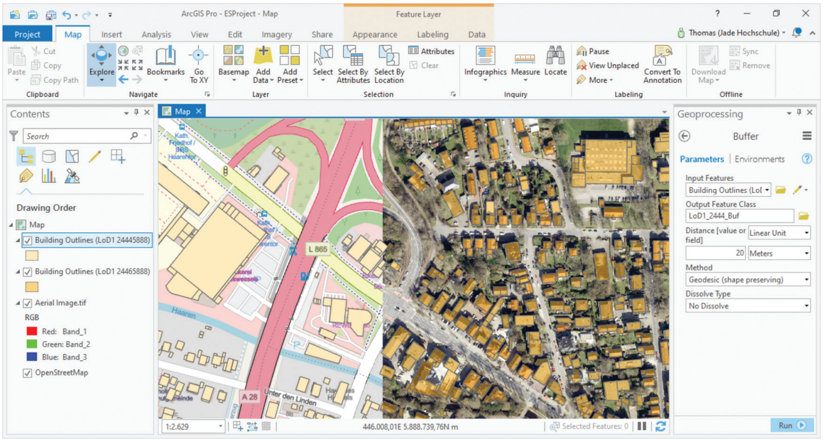
Geographic Information Systems

A GIS is a computer-based system designed to collect, manage, analyze, and present geospatial information (Bartelme 2022). The acquisition of geoinformation comprises not only the data input, by using the GIS, but also the import of geospatial data encoded by different data formats. Update functionality is provided as well. Management includes the appropriate description, structuring, storage, and retrieval of geospatial data. GIS supports both the vector and raster models. Data are typically organized by corresponding layers. In addition to the proper datasets, metadata describing the geospatial information (e.g., area coverage, date of origin, data format, acquisition type) should also be provided.

The analysis functionality of a GIS serves to gain information and knowledge from existing geospatial data. It allows measurements, coordinates transformations and geometry analysis functions such as buffering, overlay and nearest-neighbor search, topology analysis functions (e.g., routing), interpolations, approximations, and simulations. Results can be new geospatial datasets, alphanumerical data, statistical evaluations, reports, and other forms of data. A parameterized execution of combined processing steps allows an automation. Users can define workflows in GIS by visual programming languages like the ModelBuilder in ArcGIS.

A GIS provides a graphical user interface. The main component is a map that displays one or several layers according to user-defined styling rules. Temporal developments can be depicted by animated maps.

Well-known GISs include ArcGIS Pro from ESRI, GeoMedia from Hexagon, and the open-source system QGIS. In Figure 6.4, the user interface of ArcGIS Pro is shown as an example. The map is visualized in the center. The layer list on the left contains four layers: two layers are vector layers and represent outlines of buildings. An aerial image is depicted as raster layer in the right part of the map. Open Street Map (OSM) is used as background layer in the left part. The ribbon contains basic GIS tools. The dialog on the right asks for the input parameters of a buffer computation.



- Map
  - ☑ Building Outlines (LoD1 24445888)
  - ☑ Building Outlines (LoD1 24465888)
  - ☑ Aerial Image.tif
    - RGB
    - 🟥 Red: Band_1
    - 🟩 Green: Band_2
    - 🟦 Blue: Band_3
  - ☑ OpenStreetMap

1:2.629
446.008,01E 5.888.739,76N m
Selected Features: 0

Input Features: Building Outlines (LoI
Output Feature Class: LoD1_2444_Buf
Distance [value or field]: Linear Unit
20 Meters
Method: Geodesic (shape preserving)
Dissolve Type: No Dissolve

Figure 6.4 The user interface of ArcGIS Pro. Attribution for the building outlines and aerial image: Extract from geodata of the Landesamt für Geoinformation und Landesvermessung Niedersachsen ©2023, used in accordance with the data license (https://www.govdata.de/dl-de/by-2-0). OSM used per the Open Data Commons Open Database License by the OpenStreetMap Foundation (https://www.openstreetmap.org/copyright/en).

In digital ethology, a GIS can be used to select, prepare, and combine different types of source data, to analyze and visualize geospatial data, and for data export. For this purpose, a large set of analyses are possible. The map overlay is the traditional analysis operation for combining geospatial information, and has been used since the 1960s to detect areas of urban expansion (Steinitz 2016). To extract geospatial data from individuals, neighborhood and proximity functions are essential. A typical example are studies that investigate whether minority and low-income populations are disproportionately exposed to industrial pollution (Sheppard et al. 1999). In another example, Yin and Shaw (2015) examined the relationships between physical movements and social closeness evolution. Aggregation and cluster algorithms are important to form groups and detect spatial patterns. In a study by Leong and Sung (2015), clustering was used for crime analysis. Charreire et al. (2012) identified built environmental patterns by using a GIS cluster analysis and investigated relationships between walking and cycling facilities and body mass index. Westerholt (2019) estimated hot spots from geospatial social media datasets and found, in contrast to other city districts, that the Asian quarter in San Francisco was a hot spot of messages during Chinese New Year celebrations.

It is important to note, however, that a GIS has limitations for use in digital ethology. For instance, GIS is not designed to handle very large sets of single features. A certain grade of aggregation should thus be done beforehand. Analyses that are very time consuming and may need a distributed processing of data should not be done within a GIS. Furthermore, a user must be aware that a GIS is a proprietary software.

Virtual Globes

Virtual globes are programs that enable the representation of Earth based on a three-dimensional model. In addition to satellite and aerial images, a virtual globe contains further geospatial data (e.g., streets, railroad lines, place names) and georeferenced objects (e.g., photos, Wikipedia articles, 3D building models). Google Earth is a well-known example of a virtual globe.

Virtual globes also allow the creation of mashups by integrating user data. In Google Earth, the KML data format can be used for defining point, line, and area geometries as well as corresponding visualization rules. It is also possible to integrate raster maps and map services as well as user-defined 3D building models.

For digital ethology, virtual globes are helpful for checking hypotheses, validating the quality of given geospatial datasets, for finding explanations for outliers or unexpected results, and for verifying the results of analyses. Figure 6.5 shows the results of an algorithm that computes center points for localities (Brinkhoff 2020). Corresponding locality boundaries and Google Earth's imagery enable a user to check the results of the algorithm.

Spatial Database Systems

Database systems allow the permanent and secure storage of large amounts of information in databases and support efficient retrieval of data. The structure of data in a database follows the specifications of a database model to ensure uniform and consistent storage and update. In particular, the management software supports simultaneous multiuser operations. The relational database model is most commonly used. It allows the retrieval of data by SQL (Structured Query Language). Sometimes the relational model is extended by

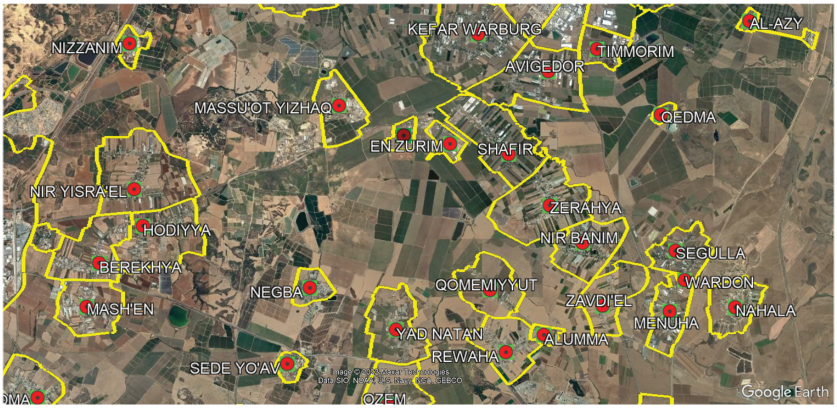


Figure 6.5 The use of Google Earth to compare the computed centers of Israeli localities with the locality boundaries and the globe's imagery. Attribution: Image © 2023 Maxar Technologies Data SIO, NOAA, U.S. Navy, NGA, GEBCO.

object-oriented functionality ("object-relational model"). For supporting large-scale application, NoSQL databases are gaining importance.

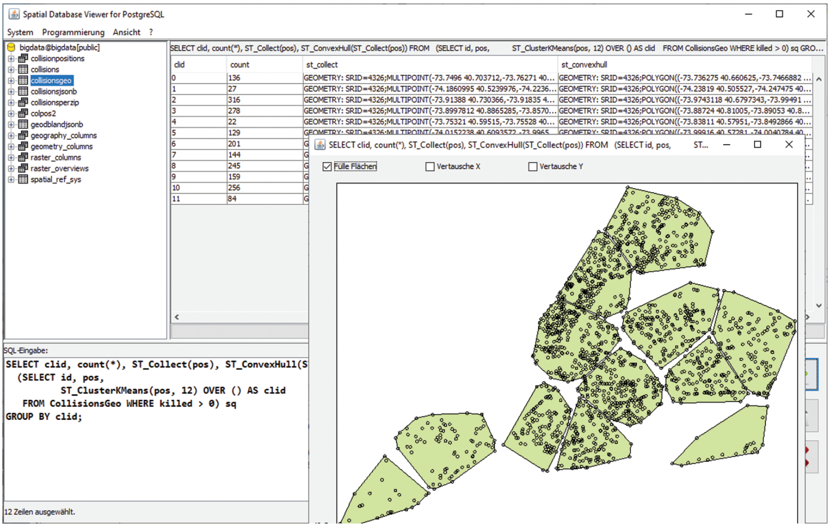
Spatial database systems allow the integrated storage and spatial retrieval of geospatial data (Brinkhoff 2022; Rigaux et al. 2011). The structure and semantics of geospatial data and the functionality of the spatial database system follow internationally accepted standards to ensure interoperability (ISO 19125 and ISO/IEC 13249-3 SQL/MM Spatial). A spatial database system can be used to manage the data of a GIS. It can also be run independently of a GIS to provide data to geospatial services or applications. Object-relational spatial database systems with extensive functionality are Oracle Spatial and PostgreSQL with the PostGIS extension. First NoSQL databases like MongoDB and Neo4j also support geospatial data (Guo and Onstein 2020). Their functionality is, however, rather limited compared with the aforementioned systems.

A primary task of spatial database systems is to support spatial queries. An example is the point query, which determines all geospatial features whose geometry contains a given query point. Other examples are window and region queries that compute all features intersected by a given query rectangle and polygon, respectively. A distance query finds all features whose geometry is located within a given distance to a query geometry, and a *k*-nearest-neighbor query (*k*-NNQ) determines the *k* nearest features for a query geometry. A spatial join allows combining two datasets and provides all pairs of features that fulfill a topological or distance condition. For an efficient processing of spatial queries, spatial database systems use spatial indexes like linear quadtrees (Samet 2006) or R-trees (Guttman 1984).

High-level spatial database systems support topological data models, 3D data, raster data, and spatial data mining techniques. The analysis of spatio-temporal data (e.g., for moving objects) is currently not supported by common spatial database systems.

For digital ethology and other fields, spatial database systems are extremely helpful in organizing large and exceptionally large sets of geospatial data, enabling cooperation between different users and software systems, and for applying aggregation and analysis operations on geospatial datasets. In Figure 6.6, the SQL query selects from 1.7 million accidents in New York City those collisions that killed persons and clusters them by the k-means algorithm into 12 spatial clusters. Their convex hulls are depicted.

For a visualization of data, GIS or other tools are required. Although some spatial database systems provide a large set of geospatial functions, their capabilities are limited in comparison to GIS. Despite SQL standardization, the handling of data by spatial database systems involves the use of proprietary solutions. In very large datasets (e.g., geotagged information from social networks), big data analytics can be performed by special frameworks and NoSQL databases (Bordogna et al. 2017; Hoel 2022).



SELECT clid, count(*), ST_Collect(pos), ST_ConvexHull(ST_Collect(pos)) FROM (SELECT id, pos,   ST_ClusterKMeans(pos, 12) OVER () AS clid   FROM CollisionsGeo WHERE killed > 0) sq GRO...

clid	count	st_collect	st_convexhull
0	136	GEOMETRY: SRID=4326;MULTIPOINT(-73.7496 40.703712,-73.76271 40...	GEOMETRY: SRID=4326;POLYGON((-73.736275 40.660625,-73.7466882 ...
1	27	GEOMETRY: SRID=4326;MULTIPOINT(-74.1860995 40.5239976,-74.2236...	GEOMETRY: SRID=4326;POLYGON((-74.23819 40.505527,-74.247475 40...
2	316	GEOMETRY: SRID=4326;MULTIPOINT(-73.91388 40.730366,-73.91835 4...	GEOMETRY: SRID=4326;POLYGON((-73.9743118 40.6797343,-73.99491 ...
3	278	GEOMETRY: SRID=4326;MULTIPOINT(-73.8997812 40.8865285,-73.8570...	GEOMETRY: SRID=4326;POLYGON((-73.88724 40.81005,-73.89053 40.8...
4	22	GEOMETRY: SRID=4326;MULTIPOINT(-73.75321 40.59515,-73.75528 40...	GEOMETRY: SRID=4326;POLYGON((-73.83811 40.57951,-73.8492866 40...
5	129	GEOMETRY: SRID=4326:MULTIPOINT(-74.0152238 40.6093572,-73.9965	GEOMETRY: SRID=4326:POLYGON((-73.99916 40.57281,-74.0040784 40.
6	201	G	
7	144	G	
8	245	G	
9	159	G	
10	256	G	
11	84	G	

SQL-Eingabe:

```
SELECT clid, count(*), ST_Collect(pos), ST_ConvexHull(S
  (SELECT id, pos,
          ST_ClusterKMeans(pos, 12) OVER () AS clid
   FROM CollisionsGeo WHERE killed > 0) sq
GROUP BY clid;
```

12 Zeilen ausgewählt.

Figure 6.6 Performing a spatial clustering of collisions in New York City by using the PostGIS extension of PostgreSQL. Attribution: NYC Open Data (https://data.cityofnewyork.us/Public-Safety/Motor-Vehicle-Collisions-Crashes/h9gi-nx95).

Geospatial Services

Geospatial web services provide an important service by making geospatial data and maps available to a broad audience. Geospatial data or processing functionalities are available on the Internet (or Intranet) through common web protocols. Spatial database systems usually serve as the data source.

Both closed and open geospatial services are available. Closed services provide geospatial data exclusively for specific applications or libraries. Their protocol is not open and cannot be used by other systems. An example is the aerial and satellite imagery retrieved by Google Earth. The map services used by Google Maps, Microsoft Bing Maps, or Apple Maps also fall into this category.

Open geospatial services are usually based on general geospatial standards. In many cases, they are made available by public institutions, such as surveying agencies or statistical offices (Kresse and Danko 2022):

- The Web Map Service (WMS) (ISO 19128) is a portrayal service that computes user-specified map sections and provides them using common raster and vector map formats.
- The Web Map Tile Service (WMTS) produces raster tiles in predefined bounds and resolutions. This restriction can significantly increase server performance.

- The Web Feature Service (WFS) (ISO 19142) provides vector-based geospatial features that fulfill a given spatial and nonspatial query condition.
- The Web Coverage Service (WCS) provides raster-based coverages according to ISO 19123. A WCS implementation typically supports clipping, scaling, interpolation, and CRS transformation.
- The Web Processing Service (WPS) is a general framework for server-side geospatial computations.

Geospatial services can be integrated into GIS and web mapping applications. For digital ethology, portrayal services can be used to support a visual analysis of other geospatial datasets. The data and processing service are more relevant for data access. The main advantage of open geospatial services is their high grade of standardization and interoperability. Proprietary web services also exist (e.g., the feature service by ArcGIS Server). A WPS may be used to provide specific functionality to a broad range of users.

Geospatial Sensor Data

Human activities influence the physical environment. As described by Smith (this volume), data collected from sensors are a valuable source for deriving measures of the physical and built environment. In the case of *in situ* sensors, the location of the sensor and the observed area are (almost) the same. For remote sensing, these two locations differ. In both cases, however, the location of the observed area is of high importance.

For geospatial sensor data, the OGC has developed an architecture for a geosensor web (Botts et al. 2013). It comprises several data models and geospatial services. First, the Sensor Model Language (SensorML) provides a metadata model to describe sensors and includes information about sensor identification, observable properties (phenomena), and the location of measurement. Second, the Observations and Measurements (O&M) data model (ISO 19156) defines an encoding of observations, including phenomenon and measurements. Finally, the Sensor Observation Service (SOS) allows querying sensor descriptions and observations by spatial, temporal, and further filter conditions.

The SensorThings API (application programming interface) represents another current OGC approach to process and provide geospatial sensor data. It takes concepts from the Internet of Things (IoT) into account (Işıkdağ 2020) and specifies a data model. In this model, a data stream groups observations that refer to the same phenomenon and are measured by the same sensor. The location can be given by the location of the sensor or by a feature of interest that describes the location being observed by the sensor. This may be the same as the sensor's location but it may also differ. The SensorThings API provides a web-based interface for requests and operations and is based on the REST paradigm (Representational State Transfer) and JSON (JavaScript Object Notation). It supports spatial, temporal, and alphanumerical query conditions as well as insert, update, and delete operations. The SensorThings API extends MQTT (Message Queuing Telemetry Transport), which is an important IoT protocol, and enables the transmission of measurements between devices, even if the bandwidth is low or delays occur.

Services for geospatial sensor data allow for the simple integration of large sets of sensor data into other applications. An important field of application are smart cities (Al Nuaimi et al. 2015; Meier and Portmann 2016). Smart cities utilize multiple technologies to improve health, transportation, energy, education, and other services important for their residents. Sensor measurements are a central ingredient for controlling these services as well as the focus of many studies aimed, for instance, at understanding how cities influence social behavior (see Balsa-Barreiro and Menendez, this volume). Privacy issues are of paramount importance and must be observed.

Geospatial sensor data are often not available for all locations in a study area. If the measured phenomenon is continuous, samples can be extrapolated using geostatistical methods provided by GIS, in particular by kriging[1] (Lorkowski 2021).

Geospatial Application Programming Interfaces

Because data processing in the context of digital ethology often requires complex and time-consuming processing steps and algorithms, one solution is to include these into statistical software, data mining software, big data frameworks, or similar packages. Still, the capabilities for processing and visualizing geospatial data vary. If a problem requires more geospatial operations, the integration into a GIS, discussed above, might offer a solution.

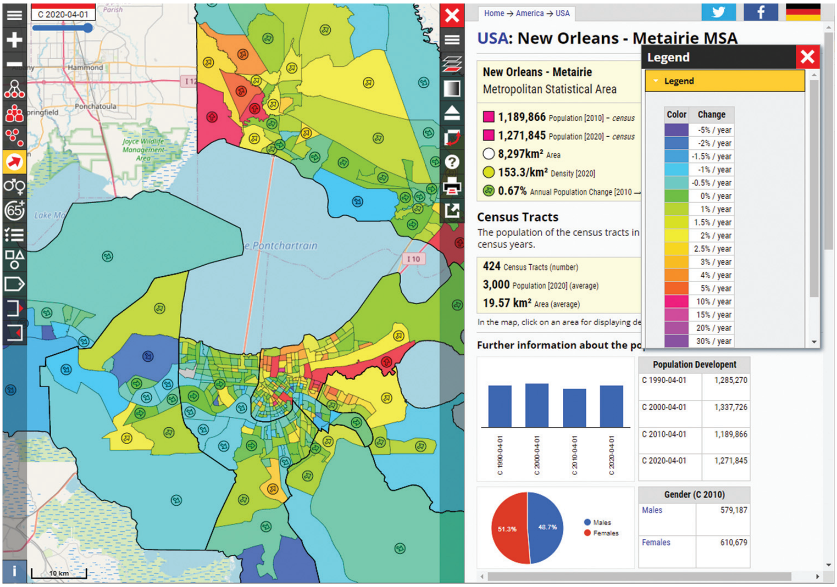
Another solution is to program a stand-alone software that uses a geospatial programming library. Such APIs are provided by GIS vendors (e.g., ArcGIS Maps SDK) or exist as independent solutions. For the Java programming platform, the open-source library JTS (Java Topology Suite) is often used as implementation of the simple feature model defined by ISO 19125. Ports into other programming languages are available.

A more comprehensive solution is the open-source GIS toolkit GeoTools. This Java API allows the representation of geospatial features and coverages. They can be uniformly queried from databases and web services. Further processing capabilities, such as coordinate transformations, raster-vector conversions, and rendering, are provided.

[1] From a limited set of sampled data points, kriging estimates the value of a variable over a continuous spatial field: (a) the spatial covariance structure of the sampled points is determined by fitting a variogram; (b) weights derived from this covariance structure are used to interpolate values for unsampled points or blocks across the spatial field.

APIs can also be used to present geospatial data and maps via web applications. In the case of web mapping, the map is embedded into a web page. Navigation and information functionality are provided. Geospatial data are usually obtained via services and spatial database systems. The creation of web mapping applications is supported by various JavaScript-based geospatial APIs. For example, most GIS vendors offer specific software packages that can be used to convert a GIS project into a web application. Corresponding APIs also exist for proprietary geospatial services. Another approach is to use independent software libraries. A prominent representative is the free open-source software OpenLayers, which allows the straight integration of various data formats (including GeoJSON, GML, KML, OSM) and open geospatial services (including WMS, WMTS, OSM Tile Service). A popular alternative is the JavaScript library Leaflet, which offers less functionality but is easier to apply.

Figure 6.7 depicts a web page that visualizes the population development of census tracts for the New Orleans–Metairie Metropolitan Statistical Area between the 1990, 2000, 2010, and 2020 census. The original census data were aggregated to small census blocks and larger census tracts by the U.S. Census Bureau for reasons of manageability and privacy. The census tract geometries are also provided by the same agency. To produce the depicted map, several further steps, outlined below, are necessary that use some of the presented geospatial tools.



USA: New Orleans - Metairie MSA

New Orleans - Metairie
Metropolitan Statistical Area

■ 1,189,866 Population [2010] - *census*
■ 1,271,845 Population [2020] - *census*
○ 8,297km² Area
● 153.3/km² Density [2020]
● 0.67% Annual Population Change [2010 →

Census Tracts

The population of the census tracts in census years.

424 Census Tracts (number)
3,000 Population [2020] (average)
19.57 km² Area (average)

In the map, click on an area for displaying de

Further information about the po

Color	Change
■	-5% / year
■	-2% / year
■	-1.5% / year
■	-1% / year
■	-0.5% / year
■	0% / year
■	1% / year
■	1.5% / year
■	2% / year
■	2.5% / year
■	3% / year
■	4% / year
■	5% / year
■	10% / year
■	15% / year
■	20% / year
■	30% / year

Population Developent	
C 1990-04-01	1,285,270
C 2000-04-01	1,337,726
C 2010-04-01	1,189,866
C 2020-04-01	1,271,845

Gender (C 2010)	
Males	579,187
Females	610,679

Figure 6.7 Population development of census tracts in the New Orleans–Metairie Metropolitan Statistical Area. Attribution: Thomas Brinkhoff, City Population, https://www.citypopulation.de/en/usa/metroneworleans/.

Census tracts of different years differ in their boundaries and are not immediately comparable. For computing adapted population figures, previous census blocks need to be assigned to 2020 census tracts. This task can be solved by overlaying the census block polygons with census tracts polygons. Another (simpler and more robust) approach is to determine a representative point for a census block and use it for a unique assignment to a census tract. The U.S. Census Bureau provides a centroid for a census block. Since centroids may lay outside of the original geometry, they are not really suitable for this task. Instead, it is better to determine (exactly or approximatively) the "visual center" of a polygon. In Brinkhoff (2020), the method of Garcia-Castellanos and Lombardo (2007) is favored because it can be programmed using an API like JTS or a script within a GIS.

The original census tract polygons are too bulky for a web application and need to be generalized for this purpose. An individual generalization of polygons would, however, produce gaps and slivers between the polygons. Thus, a topological data model has to be constructed before performing the generalization. Comprising parish polygons can be neatly computed by merging related census tract polygons.

The map is rendered by using the OpenLayer API, which requests the polygons from a geospatial service and retrieves them from a spatial database system. The original CRS is WGS84. The integration of other geospatial services works best, however, with the Pseudo-Mercator projection. Thus, the web mapping API transforms the coordinates. The background OSM is requested as raster tiles. Other background maps can be integrated by a user by specifying a WMS or a WMTS service. For the visualization, a suitable color gradient is defined. Arrow icons depict in the map the population increase or decrease. They are placed on the visual center of the corresponding polygon. Only icons that fit into the corresponding polygon are shown.

Conclusions

This chapter has highlighted geospatial IT systems that can be used for digital ethology. For such analyses, vector as well as raster data are often required. To achieve a high grade of interoperability, geospatial standards for data models and data formats should be observed. This requirement concerns not only the access to input data but also the provisioning of research results and can be fulfilled by using standardized geospatial web services.

GIS is the basic tool for geospatial analyses. It supports the acquisition, management, analysis, and visualization of geospatial data. For digital ethology, the combination of various databases as well as neighborhood and proximity functions are essential. The latter group of functions is often accompanied by network analyses. Aggregation and cluster algorithms provided by GIS are important for forming groups as well as for detecting spatial patterns and hot spots. Results can be new geospatial datasets, alphanumerical data, statistical evaluations, reports, and other types of data. In order to check hypotheses or validate data quality as well as to find explanations for outliers or unexpected results and to verify the results of analyses, we need a suitable visualization of geospatial data by maps in GIS or in virtual globes.

For large geospatial datasets, the use of spatial database systems is advisable as they support spatial queries and provide a basic (or in some cases a rather extensive) set of geospatial analysis functions. Spatial database systems also serve as a data source of geospatial web services. Sensor data are an important source for data about the environment and human activities. Geospatial standards and frameworks facilitate the access and the analysis of such sensor data.

For repeated data access or complex analyses, an automated execution of combined processing steps is necessary. Visual programming languages in GIS support the definition of such workflows. For the specification of more complex geospatial analyses as well as for geospatial web applications, several application programming interfaces exist.

7

What Types of Physical and Built Environment Can We Find in Digital Data?

Lindsey Smith

Abstract

Processes such as urbanization demonstrate how human activity influences the physical environment and the subsequent implications for Earth's natural systems. Correspondingly, changes to different environments, and in particular built environments, are linked with human behavior and health. Understanding these relationships requires the definition and measurement of environments. Considering advancements in the collection and processing of high-volume and high-velocity geospatial data, this chapter seeks to outline features of physical and built environments that may be identified from digital data. Attention is given to open data with varying spatial and temporal resolutions. Administrative data, remote sensing imagery, and data from stationary sensors provide contextual information such as the rate of urban expansion and changes in air quality. Mobile and social sensing enable the collection of high-resolution data that contribute to the identification of smaller-scale features. Developments in classification techniques, such as deep learning, provide the opportunity to explore human–environment interactions in real time. Although challenges exist related to data integration and categorization and must be resolved by future research, the combination of data from multiple sources adds value and holds promise for improving our understanding of the patterns that rapidly change landscapes, and the role of environments in shaping human behavior.

Introduction

Urbanization has modified Earth's surface and contributed to considerable changes in land cover and land use. Understanding how humans use and alter the landscape has long been important for disaster risk planning and sustainable resource management purposes (Meyer and Turner 1992; Turner et al. 2007). While human activity has influenced the form and quality of both the physical and built environments through processes such as urbanization, deforestation, and agriculture, a bidirectional association exists whereby the environment simultaneously shapes human behaviors and health.

Socioecological models recognize multiple drivers of behavior (Sallis et al. 2006). Alongside social and interpersonal factors, spatial and temporal factors related to the environments with which people interact are now embraced as determinants of behavior and health (Rainham et al. 2010). Conditions and opportunities vary by city and neighborhoods, producing social and health inequalities (Marmot 2005; Santana 2017). As a potentially modifiable target for intervention, the built environment has been increasingly addressed by research and policy. For example, a wealth of urban health literature has explored air pollution, water contamination, food environments, green spaces, and active travel infrastructure as well as their relationships with related behavioral and health outcomes (Brunekreef and Holgate 2002; Houlden et al. 2018; Lytle and Sokol 2017; Van Holle et al. 2012). Health, well-being, responsible production, as well as sustainable and resilient cities feature in the United Nations Sustainable Development Goals as strategic approaches for tackling inequality.

Understanding different environments and their relationships with human activity is therefore important. These relationships are, however, complex in nature and operate through multiple mediators and operators (Dahlgren and Whitehead 1991; Rutter et al. 2017). To date, much research has been limited by a focus on single attributes of the environment, narrow or simplistic conceptualizations of space, and either measures of risk exposure or behavioral change without consideration for how these interact (Frank et al. 2019). Advances in geospatial and computational technologies have contributed to the emergence of big spatial data. For example, remote sensing, geographic information systems (GIS), and global positioning systems (GPS) enable the collection of data with high spatial and temporal resolution. This, in turn, creates new opportunities to characterize environments and enhance understanding of human–environment relationships.

This chapter outlines features of the physical and built environments that have important implications for social science research. Digital data sources that enable the identification of such features are subsequently discussed. Focus is given to big, open datasets that are accessible to researchers and may be used to create comparable and scalable measures. Approaches developed for processing data, however, may also hold relevance for individual-level information, such as the collection of imagery from wearable cameras in cohort studies. For each data type, an overview and examples will be provided, along with a discussion of strengths and limitations. In closing, opportunities and key challenges that pertain to all data sources are highlighted to guide discussions on how measures and frameworks may be developed to advance future research.

Physical and Built Environments

The physical environment refers to physical surroundings such as air, geological and climate conditions, water, vegetation, and the built environment. The built environment, more specifically, encompasses spaces and places that have been created or modified by humans to support human needs and activities. This ranges in scale from cities to neighborhoods to infrastructure and features of urban form such as buildings and urban parks.

Table 7.1 provides a sampling of physical and built environments. The concepts discussed in this chapter are contingent on the author's research in urban spaces in high-income countries such as the United Kingdom and Canada (see also Appendix 7.1). A different research scenario may require additional or alternative concepts not listed.

Physical environments that occur on Earth's surface may be described by land cover type such as vegetation (e.g., forest, grassland, cropland), water types (e.g., wetland, open water), urban area, ice, bare soil, and rock. These environments may be further categorized by land use (i.e., the purpose for which land is utilized by people). Areas with the same land cover type can have different land uses, which may be influenced by geographical factors including the availability of resources, existing infrastructure, and proximity to urban populations. A range of land use categories have been identified and studied across a number of disciplines. Common land use types studied within an urban context include residential, commercial, transportation, recreational, and institutional. These uses, and how they change over time, provide information for planning and may influence the types and spatial configuration of built environment features that are developed. The built environment may be subject to administrative boundaries and notions of access or ownership. Features of the built environment include transport infrastructure and services such as roads, footpaths, and health-care facilities. Measures of features (e.g., the density of intersections, retail outlets, and residential units) may also be used to derive information about the value of spaces in relation to human–environment interactions, such as a walkability score.

Environments may be considered at a range of scales that correspondingly affect the type and frequency of human behavior associated with them. Macroscale environments, such as heat and rainfall, may contribute to hazardous events and affect displacement, migration, and food production. Microscale environments, such as food retailers and facilities designed for physical activity in the built environment, may encourage specific and more regular behaviors, such as types of food purchased and modes of travel used. Environments, particularly at the microscale, are moderated by quality. For instance, a park close to a busy road may experience higher rates of air and noise pollution, affecting pathways to use and associated health and well-being outcomes. While quality of environments may have an objective measure in the data, factors

Table 7.1 Examples of physical and built environments and corresponding open digital data sources. Both spatial scale and population impact decrease from top to bottom of the table.

Environments	Digital Data Type and Example Source
Climate and weather: season, temperature, precipitation, natural hazard event	*Remote sensing*: global coverage of surface temperature *Stationary sensory*: *in situ* weather recordings *Social media*: citizen response to flood event
Land cover: vegetation, water, soil, urban	*Remote sensing*: Classification of land cover from multispectral satellite sensors
Land use: agricultural, conservation, residential, commercial, industrial, institutional	*Administrative data*: food and agriculture, business registry, census population data *Social media*: spatiotemporal clustering of users
Boundaries: protected areas, municipalities, plots, buildings	*Administrative data*: census boundary files, land information
Features: *Greenspace*: parks, gardens, trees *Bluespace*: harbors, lakes, rivers, water features *Transport infrastructure*: roads, footpaths, cycle lanes, bus stops *Services*: health care, education, leisure, housing, retail *Utilities*: wastewater system, power station and lines	*Administrative data*: digitized land survey data *Participatory sensing*: volunteered points of interest plotted through open-source platform *Mobile sensing*: street view imagery of store fronts
Quality: traffic, air quality, noisescape, lighting, litter, human perceptions	*Stationary sensory*: estimated surface models of air pollution from sensor network *Mobile sensing*: street view imagery of building damage *Participatory sensing*: mobile crowdsensing of noise *Social media:* semantic analysis of geotagged tweets

such as safety, cleanliness, and noisescape can be perceived or experienced uniquely by different groups and individuals.

Digital Data Sources

In social science research, environmental data may be quantified by a geographical unit (e.g., point location, administrative boundary, address buffer, activity space) to measure exposure, access, change, or use of space. Resultant metrics may be subsequently linked to social data (e.g., based on home or work address) for analysis.

Administrative Data

Spatial data representing the physical and built environments may be obtained from existing datasets. Administrative data, such as land survey data or census data, are typically derived from organizations and institutions. Although administrative data has largely been neglected from the discussions associated with big data, such data can provide reliable information at national scales (Connelly et al. 2016). In contrast to raw sensing data, administrative data are usually cleaned (e.g., inaccuracies and inconsistencies in the data are rectified) and organized by data specialists into a format available for download and use within a GIS.

The ESRI open data hub provides access to over 210,000 open GIS datasets that have been collected from organizations around the world. Searches of key features and areas return related web maps, live dashboards, and datasets. Data can be downloaded in a variety of formats, including vector formats such as Shapefile and GeoJSON, which are commonly used for representing geographic features as points, lines, and polygons, and raster formats such as GeoTIFF that stores geospatial information as grids of pixels (see also Brinkhoff, this volume). Each dataset includes metadata describing the data source and the date when data were most recently updated. Often, local governments provide access to regional vector files representing infrastructure, land use, and municipal facilities through an open data portal. The features available as well as the temporal and spatial range of these data may, however, be limited. Separate files for individual characteristics, such as parks, schools, and transport infrastructure, can make it difficult to map cities and spaces fully. More systematic examples of administrative data collected at the national level include DMTI Spatial (Digital Mapping Technologies Inc.) data in Canada and Ordnance Survey data in the United Kingdom; both are available to education institutions in their respective countries. These repositories enable consistent matching of environmental characteristics to national cohorts with geographical heterogeneity, such as the U.K. Biobank dataset (Sarkar et al. 2015), and novel analytical approaches that incorporate a combination of environmental characteristics at scale (Smith et al. 2019b).

In addition to spatial data files of vector features representing specific environmental characteristics, aggregated information can also describe environments by geographic units. Statistics Canada, for instance, provides annual and biannual information on greenness, parks, and trees on properties at national, provincial, and metropolitan levels. As recorded in the Canadian census, which is updated every five years, population density also provides a proxy for residential density by dissemination area, census tract,[1] or larger regions.

Ultimately, administrative data can provide independent measures of the environment that range in type and scale, as well as information on groups or areas that may not be represented in social sensing data (see below). Further, administrative data are less likely to contain processing and measurement errors compared with those collected from human input or sensing (Groen 2012). While time and cost associated with data production may be reduced, the spatial coverage and availability of data may be uneven, both within and between countries, and the temporal frequency of data updates is often slow and may be inconsistent across datasets of the same area.

Sensor Data

In addition to administrative data, data collected from sensors provide a valuable source for deriving measures of the physical and built environment. Below, four primary types of sensing data are outlined: remote sensing, stationary sensing, mobile sensing, and social sensing.

Remote Sensing

Remote sensing provides information about the Earth's surface based on reflected or emitted radiation. Information is recorded by instruments at a distance, typically aboard a satellite or aircraft, and can be processed subsequently to identify features in the physical and built environment (Read and Torrado 2009). Key advantages of satellite data include its global and relatively long temporal coverage, which allows patterns and impacts of the global landscape to be systematically captured (Wulder et al. 2019). Correspondingly, satellite data have long been used to monitor land surface temperature, meteorological and climate conditions, and greenness, as well as to map and detect change in large-scale land cover such as water bodies, vegetation, bare soil, and urban infrastructure.

Reflective of the range of applications, a multitude of satellite-based remote sensing instruments exist (Horning 2019). Depending on age and purpose, sensors vary in spectral, spatial, and temporal resolution, which can affect image quality and accuracy of object detection. A characteristic example of satellites used to monitor land surface is the Landsat program, which was first launched in 1972. Since 2008, Landsat data have been available for download at no cost due to a change in data policy; this has contributed to an expansion of scientific studies (Hemati et al. 2021; Zhu et al. 2019). The most recent Landsat satellite,

[1] In the Canadian census, Statistics Canada uses geographic units such as dissemination areas (i.e., the smallest standard area available with a population of 400 to 700 persons) and census tracts (i.e., areas in large urban centers with a population of 2,500 to 8,000).

Landsat 9, carries the Operational Land Instrument (OLI) and the Thermal Infrared Sensor (TIRS), which permits multispectral images to be produced at a spatial resolution of 30 m. Compared with the coarse resolution of early land cover maps (300 m to 1 km), resultant data have improved land cover classification accuracy (Chen et al. 2015; Gong et al. 2013) and have been utilized at local, national, and global scales.

Studies utilizing Landsat data have explored common drivers of land cover change, including deforestation, urbanization, human activities, and abrupt events such as wildfires (Hemati et al. 2021). For example, a loss of 1.5 million km^2 in global forest cover was recorded between 2000 and 2012 (Hansen et al. 2013), and urban land cover in China was reported to have doubled between 1990 and 2010, replacing existing cropland (Wang et al. 2012). Such findings are substantiated by alternative remote sensing sources such as MODIS data (collected from NASA's Moderate Resolution Imaging Spectroradiometer sensors) (Schneider et al. 2010) and nighttime light observations used to identify urban clusters (collected from The Defense Meteorological Program Operational Line-Scan System) (Liu et al. 2012; Zhou et al. 2018).

Processing raw remote sensing data and, particularly, identifying land cover types requires the application of machine-learning algorithms and specialist knowledge of spectral classification. Techniques including support vector machine, random forest, decision tree and artificial neural networks have been increasingly used to classify land cover (Talukdar et al. 2020). The development and availability of global land cover products, such as FROM-GLC10 (Gong et al. 2019), allow users to access classified data without having to perform any data processing (Li et al. 2020). These products are, however, often limited in terms of their temporal coverage. As an intermediary approach, advances in cloud computing platforms such as Google Earth Engine have reduced the need for storage requirements and provide access to myriad large-scale datasets and algorithms for image processing (Gorelick et al. 2017).

While high-resolution remote sending data can aid the process of monitoring climate conditions and mapping land cover and changes, inherent constraints of the data limit classification accuracy (Talukdar et al. 2020). For example, difficulties arise in distinguishing subtle variations in vegetation types with similar spectral reflectance and small-scale features such as parking lots or small residential structures can be challenging to classify due to limited spatial resolution of Landsat imagery. Furthermore, information about land use (e.g., use of forestry for conservation or use of urban spaces for residential purposes) cannot be inferred, particularly in urban environments where single spaces may be used in multiple ways. Integrating remote sensing data with complementary sources such as administrative or social sensing data may therefore be important for improving accuracy, understanding how people interact with environments, and identifying finer-resolution built environments relevant to social science research (Yin et al. 2021).

Stationary Sensors

In contrast to the large-scale coverage of data collected by remote sensing, stationary sensors (e.g., environmental sensors and cameras) collect high-frequency information from single locations. They are suited for detecting daily changes in features and quality of smaller-scale environments.

Monitoring sites in cities are commonplace for observing traffic flows, travel modes, monitoring weather conditions, and detecting urban air quality and noise. Data from dynamic sensor streams may be broadcast for the purposes of real-time visualization (e.g., the Toronto ESRI live stream dashboard, which reports on transit, traffic, weather, and air quality for the metropolitan area). Alternatively, historic data collected at each sensor may be downloaded for analysis. Example applications include studies in the United Kingdom that utilize data collected hourly from fixed weather stations (Meteorological Office Integrated Data Archive System) to explore urban heat effects for climate change resilience (Emmanuel and Krüger 2012; Heaviside et al. 2015) and weather effects on mobility (Brum-Bastos et al. 2018). Researchers acknowledge, however, the limitations of sparse spatial coverage of stationary sensors.

To estimate exposure between networks of monitoring sites, surfaces such as Weather Research and Forecasting, dispersion, or land use regression models may be developed within a GIS. As part of the European Study of Cohorts for Air Pollution Effects project (ESCAPE), a land use regression model was generated using data from up to 80 passive samplers at 36 sites across Europe (Beelen et al. 2013). Additional predictors of land use, traffic, and geographic characteristics from administrative datasets were input into the model to derive average annual concentrations of particulate matter with aerodynamic diameter ≤ 2.5 μm ($PM_{2.5}$), nitrogen dioxide (NO_2), and nitrogen oxides (NO_X) as continuous variables, enabling the attribution of value to the home addresses of participants in multiple cohort studies across Europe.

Despite the ability to estimate exposure from surface models, low density networks of monitoring points limit the capacity of models to capture accurately the spatial variability in concentrations being measured (Marshall et al. 2008). To address this and improve spatial and temporal variability, studies increasingly incorporate mobile (Deville Cavellin et al. 2016) and crowdsourced social sensor data, such as citizen weather station networks (Brousse et al. 2022) and microphone-enabled smartphone apps, to measure ambient noise levels (Marjanovic et al. 2017).

Mobile Sensors

Mobile and portable sensors have contributed to increased spatial coverage of sensor networks. Street view datasets, such as Google Street View (GSV), Bing Streetside, and Tencent (specific to China where these is no official coverage of GSV) are examples of mobile sensor data that capture small-scale features in the built environment. GSV, the largest of these datasets, has full or partial coverage in 102 countries. Data are collected as 360° panoramic images from vehicles and updated (at most) annually, depending on location and urbanicity. Images are available for download via the GSV application programming interface (API) which provides access to the most recent imagery. The GSV "Time Machine" function and open-source packages (e.g., the module for downloading photos from GSV) further enable users to view and access historic data to assess retrospectively environmental change (Cândido et al. 2018; Cohen et al. 2020).

The emergence of street view imagery has proven useful for identifying features such as retail outlets and validating existing GIS datasets. In addition, fine image resolution has facilitated the identification of visual factors that affect the quality of the built environment, such as greenness, broken windows, potholes, property damage, litter, and the estimation of urban canyons based on sky openness and building height. Street imagery also makes it possible to identify "nudge factors," such as the presence of billboards advertising junk food (Egli et al. 2019; Huang et al. 2020), and relative perceptions of environmental quality across space. For example, comparison of GSV imagery with hand-drawn maps revealed Latin American schoolchildren were more aware of litter in natural compared with urban environments (De Veer et al. 2022). Lastly, street view imagery makes it possible to explore human interactions with the environment, such as the number of street users and their modes of travel (Goel et al. 2018; Ibrahim et al. 2021).

Identifying features at scale requires the application of deep learning techniques whereby models are first trained on a large sample of images. Applied examples include the use of semantic segmentation and convolutional neural networks to predict human perceptions of images (Zhang et al. 2018) as well as to identify streetscape green and blue spaces and to examine relationships with behavior and health outcomes, such as depression in the elderly in Beijing, China (Helbich et al. 2019), or walking in Hong Kong (Lu 2018). These complex approaches rely on pixel-level classification to recognize and understand the subtle differences of features within an urban scene.

Mobile sensor data are more cost- and time-effective than field audits for identifying visible environmental features. Previous studies report accurate and consistent agreements between field audits and the use of street view imagery, highlighting its potential for filling in missing information from stationary sensor and administrative datasets. Critiques include irregular spatial coverage and variable collection frequency. In the case of street view data, only a snapshot of locations is provided, and this does not capture dynamics and flows of urban spaces or account for differences by time of day, day of the week, or season. In addition, data coverage may be biased toward more commercial streets, given the focus on businesses within Google Maps. As with remote sensing and administrative datasets, value is added to mobile sensor data when integrated with complementary information such as social sensing data.

Social Sensing

Social sensing involves the collection, processing, and analysis of crowd-sourced data from humans using devices (Pandharipande 2021). Given the proliferation of smartphone usage with GPS and camera capabilities as well as the unprecedented use of social media platforms for broadcasting information, social sensing has received attention as a means to acquire data about cities and human–environment interactions at scale (Aggarwal and Abdelzaher 2013; Wang et al. 2015). Consequently, social sensing offers potential for applications in urban planning, transport, health, and crime prevention.

Social sensing can be categorized into (a) participatory social sensing, where participants are recruited or voluntarily contribute data about a geographical area, (b) the use of social media data, where the user is not purposively generating data to map environment features. Although not discussed here, population-level GPS data from smartphones also provides useful information for traffic flows and crowding.

Participatory Social Data. In general, the process of participatory sensing involves citizens voluntarily and intentionally uploading local information through a platform or application. This instance of user-generated content, coined volunteered geographic information (VGI) by Goodchild (2007), has enabled collaborative mapping of environmental features based on local knowledge of participants worldwide. Citizens have become empowered to collect and map features that may not traditionally be included in administrative datasets, such as cycling facilities and wheelchair routes.

One prominent example of VGI is OpenStreetMap (OSM), which has around 37,000 active contributors per month. OSM is an openly accessible and editable map of the world; contributors typically input points of interest (e.g., retail outlets, education and health facilities, transit stops) and linear features (e.g., rivers, roads, bus routes) that can be downloaded via various repositories. Key strengths of OSM include its community input and global coverage, yet concerns have been raised over biases in mapped features and the validity of data. As a result, researchers have sought to demonstrate reasonable comparisons with administrative datasets in specific regions of the world (Dorn et al. 2015; Haklay 2010), and tools such as TagInfo have been developed to guide users and encourage consistency when tagging features. Attempts to standardize OSM tags also enables them to be mapped to standard codes, such as the North American Industry Classification System, allowing for linkage and comparison with administrative data.

In addition to existing platforms for logging volunteered information, geotagging campaigns may utilize mobile crowdsensing to build a denser network and more reliable measures of environmental conditions. For example, in Brazil, *Guiaderodas* (a technology company that promotes accessibility in built environments) relied on crowdsourcing to evaluate the accessibility of over 250,000 establishment locations for wheelchair users in over 115 countries. Data are subsequently used to provide information for app users to plan accessible routes. Measures of noise have also been recorded for studies using microphones on personal smartphones. Capturing more complex measures of weather and air pollution, for example, may require the use of specialist devices, which can limit the number of contributors (Brousse et al. 2022; Marjanovic et al. 2017).

Social Media Data. Social media platforms such as Twitter/X, Facebook, Instagram, Flickr, YouTube, online blogs, and review ratings sites allow billions of users to generate and share data in the form of text, image, or video. The use of social media on smartphones can also provide detailed contextual information, such as location and time, based on GPS.

Geotagged social media data may be downloaded through an API. Although not initially intended to provide environmental information, spatial and temporal clustering of check-in activities and geotagged tweets have been used to infer land use (Soliman et al. 2017; Zhan et al. 2014) and quality of parks by incorporating semantic content analysis (Kovacs-Györi et al. 2018). The potential of using a framework to integrate image, text, and maps has also been demonstrated in the context of a localized event: the release of water from flood control reservoirs in Houston during Hurricane Harvey in 2017 (Fan et al. 2020). A graph-based approach was first used to detect critical tweets, then an image-ranking algorithm for selecting relevant images, and lastly a kernel density estimation of geotagged locations was used to map the geographic coverage of disruptions. The combination of social media data types and approaches may therefore contribute to enhanced real-time situational awareness of rapid environmental changes, such as wildfires and flooding, as well as slower changes, such as evolving perceptions and definitions of land use (e.g., use of residential spaces for employment which accelerated during the COVID pandemic).

While social media data provide new opportunities for understanding human–environment interactions at scale, key limitations lie in its reliability and representativeness. It is difficult to infer the validity of information in text, and data remain biased toward social media users, specifically those who enable geo-location services. Only 1–2% of tweets are geotagged, calling for geocoding and geoparsing methods to extract additional locations (Middleton et al. 2018). In addition, concerns around geoprivacy due to the disclosure of individuals' sensitive locations have been raised. As a result, spatial data may need to be masked or aggregated to protect individuals from being identified through their location records.

Key Considerations

Below, areas that require further discussion are highlighted to guide the application of big data in environment–behavior research.

Data Integration

Each data source is associated with unique strengths and capabilities for identifying environmental features. For example, remote sensing imagery provides global coverage but cannot capture small-scale features or land use. Street view imagery provides greater resolution but often at lower temporal frequency, whereas social sensing can capture high-frequency information but data quality is limited. Selecting a single data source may be appropriate for identifying a single environmental feature; there will likely be trade-offs, however, in spatial and temporal coverage, and data quality. Furthermore, human behavior is embedded in a complex system of places, times, and environments. Much of the literature exploring relationships between the environment, behavior, and health has focused on single features. While useful for identifying associations with specific outcomes (e.g., walkability and walking), environmental characteristics coexist and interrelate. Reflecting on the growing recognition of the broader determinants of behavior and health, a more holistic and integrative approach to measuring environments is required if we are to gain a better understanding of these complex interactions.

Combining digital data sources enables multiple environments and outcomes of varying scales to be explored: from broad city-level influences on population health to feature-level influences on more personal behaviors. Curated data libraries provide the first step in bringing spatial data about the physical and built environment from diverse sources into a single location. Subsequent consideration must therefore be given to how data are integrated, particularly given different formats, time frames for collecting data, and disparate scales and coverage. Here, deep learning may provide an opportunity to bridge gaps between different data types (Zhang et al. 2019).

Data Categorization

Linking environmental data with information related to social, behavioral, and health outcomes creates possibilities for analyzing associations and exploring inequalities by place. The unit at which data are aggregated and categorized, however, may have implications for causality.

Sociodemographic, social, or health data, such as that acquired from the census or social media, may be aggregated to an administrative boundary such as a census tract. Matching data with environmental features quantified within the same unit enables broad population-level patterns to be observed. Such analyses are limited by the modifiable areal unit problem, whereby the chosen spatial unit differentially impacts results. For example, the same data aggregated by census tract, postal code area, or an individual's neighborhood may yield different effects. Ecological fallacy (i.e., inferences made about individuals from group data) may also rise when investigating features of the built environment (Houston 2014). Exposure is considered to be the same for all who live in the same administrative unit, irrespective of mobility patterns and transport opportunities. Linking individual-level data from cohort studies helps to overcome this. Still, much work in this area has focused solely on the home address by quantifying features within a residential buffer. In doing so, researchers are at risk of the "uncertain geographic context problem" as relevant environments beyond the home neighborhood where behaviors occur are missed (Kwan 2012). Increasingly, studies use GPS data to capture more relevant spaces. Features are quantified within individuals' activity spaces, based on locations they have visited over time. Delineation of activity spaces has been inconsistent and studies have conflated measures of access with use of space (Smith et al. 2019a).

Consideration needs to be given, therefore, to the quantification and categorization of data to ensure its conceptual relevance for meeting study aims and enabling comparisons. Here, metadata can help ensure that data are not only findable but described transparently in terms of collection processes and applicable for previous use cases. As high-volume location data become increasingly available, consideration must be given to how spatial and temporal sequence patterning may be incorporated into measures, beyond simple delineations of activity spaces (Fuller and Stanley 2019). As measures begin to reflect environments of importance better, researchers must not lose sight of causal thinking and strive to develop stronger evidence on the pathways that act to influence use of spaces and changes in behavior.

Reproducibility

Given the volume of digital data and application of machine-learning methods to process data, it is important that methods are reported transparently. Code sharing sites such as GitHub allow researchers to test, collaborate, and build upon existing approaches. This has implications for replication, scalability, and comparison in future studies.

Conclusions

Taken together, the wealth and quality of openly available digital data enables the identification of a range of physical and built environments at different spatial and temporal scales. Remote sensing imagery, administrative data, and stationary sensors provide contextual information such as the rate of urban expansion, changes in air pollution, and coverage of green spaces. Meanwhile, methodological advances in mobile and social sensing enable the collection and analysis of highly granular longitudinal data on small-scale features through VGI. Developments in classification techniques, such as deep learning algorithms, also permit real-time behaviors to be explored in place through social media and mobile devices. Compared with traditional field audits or the collection of information from study participants, acquiring and processing digital data can be much less resource intensive. This holds promise for improving our understanding of the patterns that are rapidly changing global and local landscapes, and the role of environments in shaping human behavior and health over time.

While the volume and velocity of openly available data may not yet match that of commercial or privatized data, the availability and variety of open data for characterizing physical and built environments is continually increasing. As computational capacity and data expand, users must provide key considerations as to (a) the representativeness and relevancy of data and (b) to integration, categorization, and reproducibility. Value is added when variable data from multiple sources are combined to explore spatial patterns or develop immediate strategies, such as the direction of humanitarian aid following a natural hazard event. Approaches to data preparation and analysis also have implications for causality, findings, and potential inferences. Transparency in communicating methods and findings, with efforts toward reproducibility, is therefore key to ensuring integrity and reliability in research.

Appendix 7.1: Explanation of Useful Terms

Big data: High-volume and velocity data which may be too large to be processed with traditional software applications. May be analyzed to reveal patterns, trends, and associations, especially relating to human behavior.

Open data: Freely available data that may be downloaded and modified.

Crowdsourced data: Contribution of information from a large number of people.

Geographic information system (GIS): Computer system for creating, storing, and analyzing spatial data.

Application programming interface (API): Intermediary software that enables the transmission of data between two applications.

Machine learning: A type of artificial intelligence (AI) that finds and learns from patterns in big data.

Deep learning: A type of machine learning that uses multiple layers of processing to find smaller patterns in big data.

Semantic image segmentation: Computer vision task in which each pixel of an image is labeled with a corresponding class of what is being represented.

Convolutional neural networks: A form of deep learning which uses multiple layers to process arrays of data such as those in images, and extract features.

Semantic analysis: Process of finding meanings in text.

8

How Cities Influence Social Behavior

José Balsa-Barreiro and Monica Menendez

Abstract

Over the past century, urbanization has witnessed a significant rise, with the global population in urban areas surpassing 55% today and expected to reach nearly 70% by 2050. While cities contribute to productivity and innovation, dense urban living can bring challenges such as increased living costs, social segregation, traffic congestion, and rising levels of air pollution. The COVID-19 pandemic, coupled with technological advancements and social shifts, has reshaped urban landscapes. Since the majority of the world's population resides in urban areas, addressing societal and environmental challenges necessitates a focus on cities. This chapter explores the intricate relationship between urban form and social behavior, drawing insights from an extensive review of literature across various themes: human cooperation, mobility, social interactions, integration, quality of life, health, and safety perception. These findings provide a comprehensive framework to understand the complexities of social dynamics in urban environments.

Cities as Complex Systems

The world is experiencing an unprecedented, substantial trend toward urbanization. As reported by the United Nations (UN-Habitat 2022), more than 55% of the global population currently lives in urban areas, and projections indicate this proportion will reach approximately 70% by 2050. This progression will lead to intensified concentration of people, goods, means of production, and services within increasingly confined spaces.

The driving force behind urban growth lies in the advantages that are linked to *economies of scale* (Gill and Goh 2010; Wheaton and Shishido 1981). Urban environments serve as hubs that concentrate a diverse array of job opportunities by facilitating the convergence of key agents, including people and workplaces. This concentration optimizes essential resources, reduces infrastructure investments, and encourages the development of collective transportation systems (Pentland 2014). Presently, urban economies form the backbone of the most high-income countries (Frick and Rodríguez-Pose 2018), leading to the continual growth of cities in terms of population and economic prosperity over time (Thisse 2018). In the last decade, Dobbs and Remes (2013) estimated that the 2,600 largest global cities accommodated 38% of the global population while contributing to 72% of the global gross domestic product (GDP). Recent projections from the World Bank (2023) suggest that the contribution percentage might have already surpassed 80%.

Some particular discrepancies contradict previously mentioned arguments, one of which lies in the nonuniform correlation between global urbanization and wealth expansion. Balsa-Barreiro et al. (2019a) analyzed the sustainability of global economic growth from the 1960s, using factors such as wealth generation (in terms of GDP), environmental impact (CO_2), and population indicators, particularly urbanization. By estimating the average location of the planet's activity for each indicator annually, they illustrated the trajectory of these indicators over time. The findings revealed diverging trends: while global wealth gravitates toward the East, population growth and urbanization trend toward the South.

The progression of urbanization brings forth a multitude of challenges, particularly in environmental and social contexts. Notably, the upsurge in mobility and resulting traffic congestion poses significant costs, potentially impeding urban competitiveness (Sweet 2011). Urban residents face the risks of exposure to the environmental impacts stemming from cities, which currently contribute to two-thirds of global energy consumption and over 70% of greenhouse gas emissions (World Bank 2023). Moreover, rising social tensions, including urban segregation and gentrification, arise from imbalances in supply and demand within a fiercely competitive global context. These complex issues underscore the concept of *urban diseconomies*, a notion highlighted by scholars to portray the compounding challenges associated with agglomeration economies (Richardson 1995).

Regional disparities in urbanization rates highlight distinct patterns. In low- and middle-income countries, rapid urbanization stems primarily from limitations in rural areas rather than urban opportunities. This phenomenon has resulted in pseudo-urbanization processes (Balsa-Barreiro et al. 2021; Hashimov et al. 2013), posing risks of environmental unsustainability and social exclusion. This includes the rise of poverty pockets, which can increase crime rates, and the expansion of informal settlements with inadequate services, heightening vulnerability to potential hazards for their dwellers (Williams et al. 2019; Zerbo et al. 2020).

Four additional aspects are pertinent to comprehend the magnitude of the global urbanization process. The first involves evaluating the accuracy of estimated projections for the mid-century within the current intricate context. Factors such as the impact of the COVID-19 pandemic and the technological transformation derived from artificial intelligence (AI) have contributed to deepen the economic deglobalization initiated in the mid-2010s, potentially leading to short- to medium-term structural changes in the labor market (Balsa-Barreiro et al. 2020b; Rossi and Balsa-Barreiro 2020). Some authors have initiated discussions on the future impacts of these factors over cities (Williams 2023) and population distribution, debating the potential for urban resilience (Foster 2020) versus the urbanization crisis (Kotkin 2020) over the next decades. The second aspect concerns the granularity and scaling of the urbanization process (Bettencourt 2013). Although urbanization processes are commonly linked to large metropolises, they manifest at various scales and levels. Balsa-Barreiro et al. (2021) illustrated how the urbanization process operates across multiple scales, not solely confined to large cities, displaying fractal patterns characterized by repetitive geometry across scales (Batty and Longley 1994; Mandelbrot 1982). Consequently, smaller cities may experience rapid population growth, leading to heightened traffic congestion and pollution beyond their capacity (Borck and Tabuchi 2019). The third aspect highlights the growing role of cities as primary economic centers for entire regions, emphasizing the need for an urban-focused approach to tackle global social and environmental challenges (UN 2016). Finally, the fourth aspect refers to spatial disparities among cities based on wealth levels. High-income countries demonstrate steady urbanization rates and low demographic growth, featuring built and well-established cities. Conversely, low- and middle-income countries experience rapid urbanization, often marked by unregulated and informal construction in many cities.

The intricate nature of urban complexities underscores the critical need for a more profound comprehension of urban dynamics to address proactively forthcoming social and environmental challenges. Achieving this requires a deeper understanding of the driving mechanisms that shape city performance, encompassing both physical and social dimensions. In this chapter, we investigate the correlation between the physical structure of cities and the social behavior of their residents. To achieve this, we conduct an extensive literature review of prominent studies that link these factors. Our goal is to establish a robust framework for future research, shedding light on how physical and human factors interact in urban environments.

We will begin with an introduction to the factors influencing urban form. We then define the concept of urban morphology and its treatment in current literature. Next, we explore the reciprocal relationship between cities and human behavior through a literature review across seven major social themes: human cooperation, mobility, social interactions, integration, quality of life, health, and safety perception. We then outline potential data sources for gathering information related to both social behavior and urban morphology. We conclude by summarizing key insights to consider in planning sustainable, efficient cities for the future.

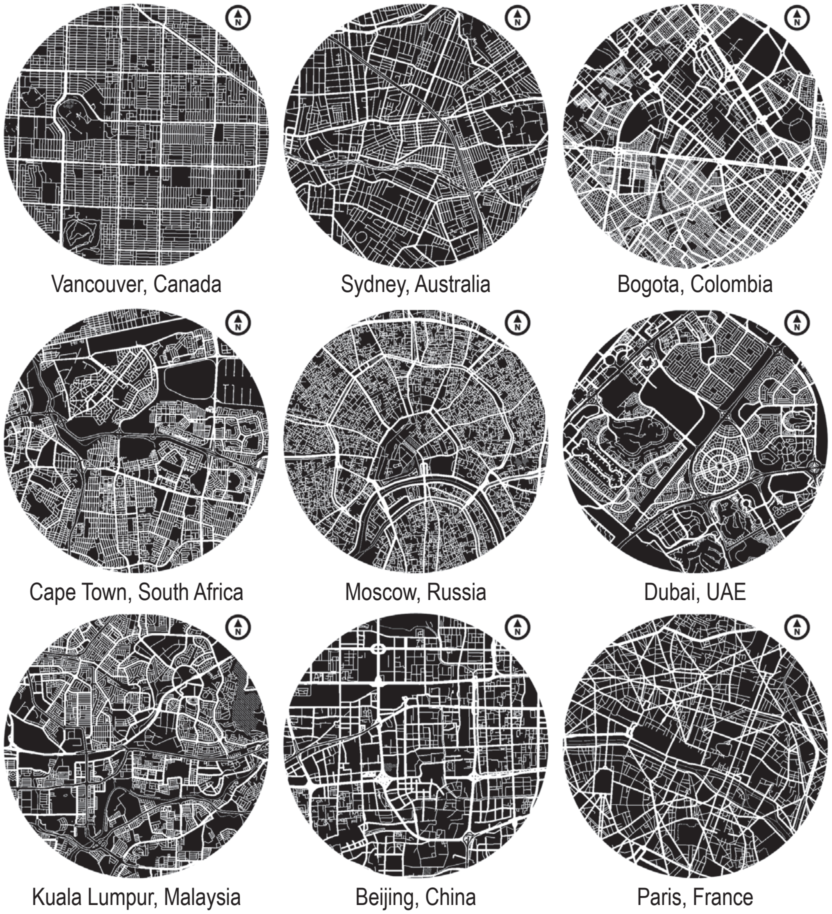
Urban Form across Scales

There are two primary factors explaining the urban configuration of cities. First, the physical context surrounding cities constitutes a primary determinant of their layout. Proximity to natural features such as rivers, coastlines, or valleys strongly influences and constrains the directional expansion of a city. Many cities are strategically planned to capitalize on their natural potential. Urban designs in tropical regions prioritize maximizing cooling breezes, while cities in desert areas often feature narrower road networks to mitigate sun exposure (Hang et al. 2009; Masoud et al. 2020). Second, the socioeconomic aspect, reflected in the diverse urban forms and their evolution, emerges from the interplay between physical and human factors, particularly regarding the economic utilization of natural resources. Throughout history, a notable portion of the world's population has settled in coastal zones, utilizing water resources for various industrial purposes and enabling ease of navigation and coastal fisheries. Approximately 40% of the global population resides within 100 kilometers of coastal regions (Moser 2014), although this ratio significantly rises when accounting for riverbanks, lakes, and other water bodies. This underscores the critical role of these natural elements in human development. Medieval cities were historically located in strategically favorable and well-connected sites, serving as pivotal hubs for the development of markets catering to vast rural regions (Fujita et al. 2001). The expansion of these exchange centers and their transformation into substantial urban centers stemmed from the significance of their potential market.

Technological advancements, particularly in transportation, played a pivotal role in both the expansion and morphology of cities across scales (Balsa-Barreiro and Menendez 2021, 2022) Globally, the emergence of large cities centered on major commercial ports is attributed to low costs associated with maritime freight traffic, fostering extensive trade (Fujita and Mori 1996). Likewise, the extensive growth of suburbanization processes, known as *urban sprawl*, in American cities is primarily linked to the widespread use of private vehicles, allowing point-to-point mobility. Within cities, this influence explains the prevalence of regular city grids, characterized by broad streets designed primarily for vehicular traffic flows.

The historical arrangement of elements reveals a diverse array of shapes and sizes that trace the complexity of the *urban tissue* (Marshall 2004), highlighting human influence on constructing the built environment over time. Actual urban forms define its present usage, as reflected in its varying degrees of physical accessibility, social integration, and economic functionality (Martino et al. 2021). This underscores the reciprocal connection between urban form and the socioeconomic factors molding cities.

Drawing upon these factors, urban forms can be examined across two distinct scales. First, the macro-scale focus on the *city as a whole* provides the most comprehensive perspective. At this scale, urban sprawl encompasses building blocks and configurations of street patterns. The distribution, configuration, hierarchy, orientation, and connectivity of the street network are crucial elements that define the city layout, as shown in Figure 8.1. Second, the meso- and micro-scales are mainly characterized by neighborhoods, districts, or any aggregated units, such as visible building blocks observed at a more detailed level. This scale allows the evaluation of socioeconomic disparities in urban forms, showcasing differences in land uses, economic prosperity, and social segregation based on income and racial factors. The integration of metrics related to urban form and social indicators enables the assessment of accessibility to infrastructures, open spaces, and basic services across different regions of the city.



Vancouver, Canada

Sydney, Australia

Bogota, Colombia

Cape Town, South Africa

Moscow, Russia

Dubai, UAE

Kuala Lumpur, Malaysia

Beijing, China

Paris, France

Figure 8.1 Configurations of street networks in densely populated cities on different continents. Different traces result from the interaction between physical and human factors. Each figure illustrates a distinct region, displaying diverse spatial scales for each city. The thickness of the lines represents the hierarchy within the street network.

Urban Morphology

The study of urban environments involves various perspectives including visual, perceptual, and social aspects. Krier (1979) defines urban spaces as "all sorts of space between buildings," emphasizing physical construction and referring to spaces where interactions between people and places occur. The form of the urban core is made up of essential physical elements, including building blocks, plots, and streets (Moudon 1997). Subsequently, other elements like land use (Levy 1999), natural environments, and green spaces (Kropf 2009) were integrated later. Building blocks, which delineate the smallest enclosed spaces within an urban grid, and streets, which comprise the public network for movement across the urban landscape, are widely recognized as key indicators by most authors.

Urban design encompasses primary dimensions, including form. *Urban morphology*, as a distinct discipline, investigates physical structures, spatial layout, and changes of cities over time (Kropf 2017). This discipline, traditionally qualitative and visual, has been transformed due to the abundance of data and enhanced computational capabilities, resulting in the emergence of quantitative methods known as *urban morphometrics* (Dibble et al. 2017). This advancement contributes significantly to measuring and categorizing urban form, particularly enhancing *typo-morphology* studies (Samuels 2008) and *space syntax* theories for the analysis of spatial configurations of urban networks (Elek et al. 2020; Hillier 1996).

Urban morphologists have developed indicators to estimate various morphological relations (orientations, areas and dimensions in 2D, volumes in 3D) between discrete elements, describing the morphology, geometry, and typology of the built environment. Vertical indicators aid in studying building façades, horizontal indicators cover building distribution factors (density, distances), while volumetric indicators define compactness. Street indicators refer to road network configuration describing urban grid and axial lines, but also street composition, which include width, position, length, area, and orientation of roads. Additionally, land use, particularly the presence of green spaces, is a significant factor in these studies. In this case, we must consider aspects related to total area as well as spatial distribution and fragmentation of green spaces throughout the city. Hence, urban morphology involves physical characteristics such as shape, size, and density, yet its complexity lies in assessing spatial relationships among its elements. A simplified proposal for the classification of urban indicators is shown in Figure 8.2.

The estimation of these attributes involves the development of specific methodologies to derive a set of metrics or indicators. Methods and outcomes might differ depending on factors such as the basic spatial unit, the spatial scale, and level of data aggregation, among others. For instance, studies conducted by Hermosilla et al. (2014) and Boeing (2019) estimated indicators at the street level, whereas Biljecki and Chow (2022) conducted their analysis

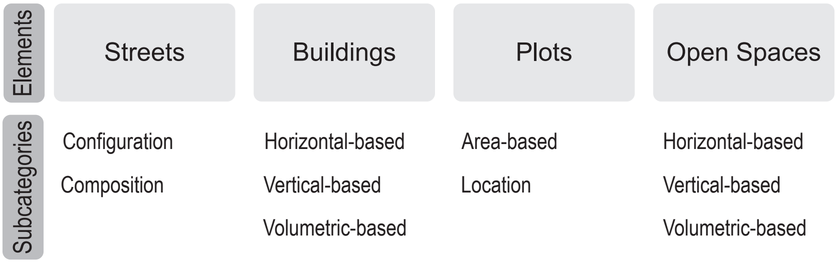


Figure 8.2 Classification of urban morphology indicators, aligned with Elzeni et al. (2022).

at the building level. Unlike many studies that propose a limited number of indicators, the latter is one of the most comprehensive studies, presenting a list of 43 morphology indicators. Noteworthy contributions also come from Bourdic et al. (2012) and the enhanced version by Fleischmann et al. (2020). Their comprehensive review encompasses 72 quantitative studies, identifying a list of 465 measurable indicators of urban form. Due to terminological inconsistencies and vague methodological descriptions in some studies considered, they refined the list to 361 valid indicators. These indicators were classified across six categories (dimension, shape, spatial distribution, intensity, connectivity, and diversity) and three conceptual scales (small, medium, and large). A brief summary of this proposal is given in Table 8.1.

Cities and Human Behavior

Some environmental factors contribute significantly to our perception of a place. In urban settings, these encompass physical and built environments shaping and defining the existing urban morphology. Their interaction highlights the complexity of urban spaces and their impact on our perception, extending beyond purely aesthetic or subjective comfort criteria. Cities located in naturally favorable environments with pleasant climates may possess poor urban planning, leading to varying perceptions among individuals. To address this ambivalence, numerous studies have employed diverse approaches to understand the impact of built environments on experience and perception. Multiple approaches span different disciplines, including the ones coming from subjective geography (Hiss 1991; Lynch 1960), psychogeography (Self 2007), and environmental psychology (Kopec 2012).

Geographical and sociological approaches offer significant insights into the impact of the environment on social behavior. *Geographic determinism* emphasizes environmental factors as primary influencers on human behavior, cultural development, and societal progress. In contrast, *possibilism* highlights societies' capacity to overcome these natural constraints (de Quadros 2020). The *Chicago School* (ecological school) made substantial sociological

Table 8.1 Condensed version of the morphological indicator's list developed by Fleischmann et al. (2020), including categories with their respective definitions and a list of relevant indicators.

Category	Definition	Indicators
Dimension	Geometric properties of individual objects	• Length • Height • Number of floors • Mesh size • Area
Shape	Geometric dimensions' mathematical properties	• Height-to-width ratio • Compactness index • Form factor • Fractal dimension • Rectangularity index • Complexity index
Distribution	Spatial distribution of objects in space	• Built front ratio • Distance • Continuity • Concentration index
Intensity	Density of elements by unit area	• Covered area ratio • Floor area ratio • Number of plots • Weighted number of intersections
Connectivity	Spatial interconnection of street networks	• Closeness centrality • Clustering coefficient • Node/edge connectivity • Node connectivity
Diversity	The diversity and complexity of the elements	• Power law distribution of areas • Plot area heterogeneity • Plot area diversity • Intersection type proportion

contributions. *Symbolic interactionism* theories proposed that both the built environment and social structures shape human behavior (Bulmer 1984). In the early 20th century, these ideas were tested in Chicago through compelling experiments. Thomas and Znaniecki (1918) argued that immigrants' transition from controlled European societies to the more competitive urban environments fueled Chicago's dynamic growth. The school explored a wide array of social behavior in urban settings, analyzing specific behaviors such as alcoholism, homicide, suicide, psychosis, and poverty. Their findings suggested that the urban lifestyle weakened primary social relationships, leading to social disorganization with significant impacts on human behavior. Recent studies reinforce these conclusions, demonstrating that insufficient integration and higher mobility rates are associated with increased crime rates (Caminha et al. 2017).

Aligned with the theories of *possibilism*, a thorough understanding of the intricate interplay between urban form and human behavior necessitates recognizing their reciprocal influence. At the outset, an individual's socioeconomic status shapes their urban preferences. Initially, individuals may identify with their place of birth, but social class increasingly shapes housing choices within cities. The gradual process of urban development significantly reflects social hierarchies, with income playing a pivotal role in shaping urban landscapes. Traditionally, higher-income households tend to favor exclusive areas with superior amenities, opting for spacious homes in less congested suburbs, often surrounded by extensive green spaces.

In recent decades, Western countries have transformed their urban landscapes. Criticism of the urban sprawl model, reliant on automobiles and incurring associated costs, has sparked renewed interest in revitalizing city centers. This revitalization has been driven by factors like gentrification, prompting affluent individuals to relocate to urban cores. It has favored interaction between affluent and lower-income groups, particularly during transitional periods with both groups sharing spaces (Lees et al. 2007). The changing dynamics are urging a reconfiguration of urban spaces within central neighborhoods, emphasizing policies that target their revitalization by limiting motor vehicle access and expanding green spaces. This indicates a partial transformation of urban metrics in established areas, demonstrating the continuous interplay between human behavior and urban design, guiding adjustments according to socioeconomic conditions.

Several notable studies have analyzed the substantial impact of urban design on the collective lives of residents. Jacobs (1961) criticized the mid-20th-century American urban planning, advocating for diverse neighborhoods and community-driven city growth. She emphasized vibrant streets and highlighted the significance of intricate urban environments in fostering community interaction and creativity. Her urban planning model challenges conventional approaches by emphasizing the significance of neighborhoods and local communities in the development of cities. More recently, Hern (2017) also contested prevailing narratives regarding urban development, questioning preferences that prioritize economic growth and urban rejuvenation at the expense of social equity and community welfare.

At present, urban policies adopted by many Western cities prioritize human-centric approaches. They focus on environmentally sustainable models, such as eliminating industrial pollution, improving urban green spaces, reducing reliance on private vehicles, and expanding pedestrian-friendly zones. This transformation advocates for city models that emphasize human livability through compact, interconnected, and economically diverse forms (Burgess 2000; Kain et al. 2022). The urban transition presents three key criticisms. First, certain once-praised urban models, like *sprawling suburbs*, now face criticism due to extended commuting times, environmental impacts, and their tendency to foster social isolation despite their perceived design benefits. The city models perceived as being optimized today may lose their efficiency in the face of ongoing socioeconomic or technological changes. The current focus on compact, pedestrian-friendly, human-centered cities might yield considerable adverse effects, potentially resulting in less resilient urban economies, marked by job decentralization and fewer industries due to limited accessibility or an overreliance on the services sector. Second, the influence of digital economies and remote work may diminish the attractiveness of cities as business hubs, potentially causing residential dispersion. Alternatively, increased labor flexibility and mobility levels could ultimately reshape many cities into tourist hubs, intensifying gentrification and the displacement of their residents (Moskowitz 2017). Third, a comprehensive evaluation of city-specific forms, exemplified in Figure 8.3, necessitates contextual assessments within their respective geographical contexts (Balsa-Barreiro et al. 2022). Hence, certain urban designs and policies may be suitable for specific cities but not universally applicable across all.

The growing accessibility of building-level and individual-level data has amplified research exploring the intricate correlation between urban form and social behavior. To grasp this relationship comprehensively, we conducted an extensive literature review centered on key social aspects influenced by city configurations. Our review comprises seven primary subsections: human cooperation and altruism, human mobility, social interactions, social integration, quality of life and livability, health, and crime and safety perception.

Human Cooperation and Altruism

The urban environment influences (negatively) individuals' tendencies toward prosocial behavior and helpfulness. Various studies indicate that residents in urban settings exhibit lower inclinations to engage in activities such as responding to postal surveys (Couper and Groves 1996), assisting a distressed stranger (Levine et al. 1976), rectifying accidental overpayments in stores (Korte and Kerr 1975), or contributing to charitable causes (Chen and Mace 2019).

Korte and Ayvalioglu (1981) conducted a Turkish field study to compare urban settings' impact on individuals' willingness to help. They assessed four indicators: giving change, cooperating in an interview, responding to an accident, and reacting to a lost postcard. Their findings revealed lower helpfulness among urban residents compared with those living in small towns. Moreover, they noted behavioral variations among urban dwellers based on specific urban districts. In a recent U.K. study, Zwirner and Raihani (2020) conducted a similar experiment across 37 neighborhoods in 12 cities (200,000 to 1,000,000 residents) and 12 towns (fewer than 20,000 residents). Analyzing actions like posting a lost letter or assisting pedestrians, their results diverged from Korte and Ayvalioglu's study and showed no link between urban residency and willingness to help strangers. Their findings highlighted, however, that the neighborhood's deprivation level was a significant factor influencing helping

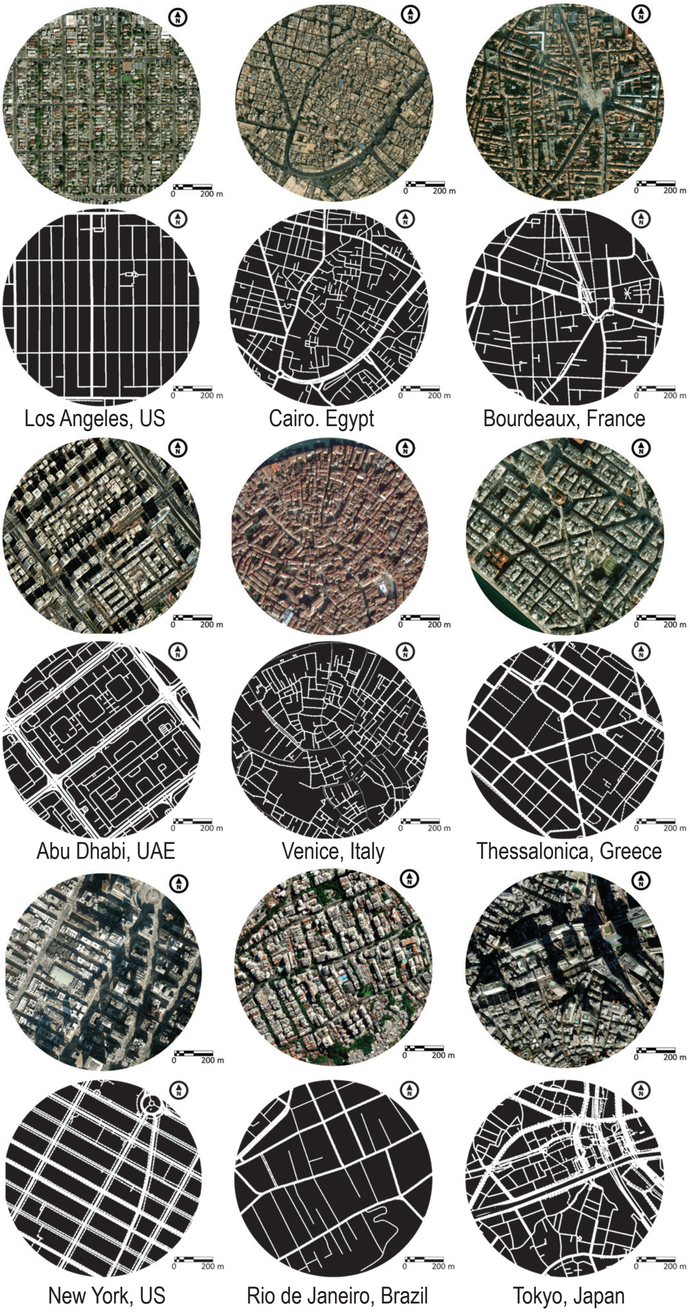


Figure 8.3 Configurations of street network and urban typologies in densely populated cities on different continents. Each figure represents a specific layout in each particular city, all standardized to the same scale. Aerial imagery collected from Google Earth (2022).

behavior. This underscores that prosocial tendencies depend more on the income factors than population size.

The ambiguity in the outcomes of the aforementioned studies was already evidenced more than four decades ago by Amato (1983), who scrutinized numerous studies examining urban–rural differences in helping behavior. In his assessment of six helping measures across 55 cities and towns stratified by population size and geographical isolation, he found a negative correlation between population size and helping behavior in four of the measures examined, with many studies exhibiting contradictory results.

Human Mobility

The surge of big data over the last decade has enhanced our comprehension of human mobility at a profound level of detail. Some prominent studies (e.g., Lu et al. 2013; Song et al. 2008) exemplify the high predictability and consistency observed in our mobility patterns, notably accentuated within urban settings. This illustrates how our commuting and leisure patterns can be remarkably similar, contingent upon our socioeconomic and demographic conditions. An alternative perspective, albeit yielding closely aligned outcomes, emerges from transportation studies. Ambuhl et al. (2021) conducted an extensive analysis of traffic behaviors across various cities spanning a year, utilizing data collected from loop detectors placed at diverse points within the urban network. Their findings revealed a remarkable consistency in the aggregated patterns exhibited by the majority of cities over time.

The urban form influences our mobility patterns. Leck (2006) illustrated how land use mixing in built environments strongly predicts our travel behavior. In general, city models characterized by extensive urban sprawl lead to widespread mobility challenges (Batty et al. 2003), resulting in larger commuting distances and exacerbated traffic congestion (Travisi et al. 2010). Prolonged congestion times not only increase commuting durations but also pose a potential surge in road fatalities (Yeo et al. 2015). In 2022, one-way commuting time in the top 50 U.S. metropolitan areas averaged 28 minutes, reflecting a 20% increase from 2019 (Candiloro 2023), mainly due to the resurgence of urban sprawl driven by COVID-19 (Peiser and Hugel 2022).

Advocates of compact city models emphasize their advantages in fostering shorter commutes and encouraging preferences for active transportation (Mouratidis et al. 2019). The opposite scenario may, however, occur, leading to higher traffic density and consequent congestion in urban centers (ADB 2019). Consequently, the debate regarding urban sprawl versus traffic externalities remains ambiguous (Wang 2023).

Confronting this, numerous cities are adopting comprehensive policies aimed at discouraging the use of private vehicles, limiting road capacities, and promoting alternative transportation modes. Many European cities, for instance, are implementing urban designs based on *shared spaces* that encourage drivers to adopt more pedestrian-friendly behavior. The efficacy of this strategy is, however, a subject of debate, particularly concerning its capacity to establish secure mobility models (Methorst et al. 2007). The impact of these policies evidence that individuals residing in urban areas drive significantly less frequently compared with residents in other regions. In cities like Tokyo, the average car ownership value stands at merely 0.32 cars per household, which is three times lower than the national average (Japan) of 1.06 cars per household (Knowles 2023).

From an intraurban perspective, Wang and Debbage (2021) underscored the substantial influence of urban form on traffic congestion. Their research indicated that cities characterized by intensified urban land use or with multiple centers (polycentric configurations) are more susceptible to traffic congestion. Examining the impact of the size of a city block on urban mobility, Zhang and Menendez (2020) revealed that opening superblocks to certain traffic flows notably improved traffic conditions. Loder et al. (2019) demonstrated how congestion hinges on urban network topology, observing that certain indicators (e.g., network density and the number of road intersections) contribute to congestion by amplifying conflict zones. Likewise, Choi and Ewing (2021) explored additional topological indicators in the Wasatch Front metropolitan area in Utah, United States. Their findings indicated that urban networks with higher density and connectivity typically experience lower levels of traffic congestion.

The uncertainty in assessing the impact of specific topological indicators on urban traffic congestion may be attributed to factors related to the location of each city and substantial variations in the spatial orientation of urban networks across the globe (Boeing 2019). Nevertheless, despite some ambiguity and conflicting results, the paradox lies in the feedback loop between these factors, where traffic congestion can induce urban sprawl, leading cities to become more extensive and less densely populated (Legey et al. 1973). Once again, this raises the question of whether this urban model represents the problem or the solution.

Social Interactions

City structure influences resident interactions. Public spaces, such as streets, squares, and parks, play a pivotal role in fostering social integration and communal life. Talen (1999) refers to the concept of "sense" that pertains to the capacity of built environments to foster a feeling of community belonging among urban residents. Many cities emphasize the need to expand public spaces (Mahmoud et al. 2013), considering aspects like spatial distribution and fragmentation as relevant metrics. The type and frequency of social relationships within public spaces depend on a wide range of factors including urban design, pollution levels, and collective safety, among others. In a study conducted in San Francisco, Appleyard et al. (1981) investigated the influence of urban design and traffic on residents. His analysis of three streets with different traffic levels revealed that dwellers in high-traffic areas had fewer social connections and a diminished sense of community compared with those in low-traffic zones.

Cities offer significant advantages by facilitating social interactions among diverse individuals, leading to competitive benefits in terms of innovation (Pentland 2014). In the physical realm, Schläpfer et al. (2014) demonstrated a close relationship between the total number of contacts, communication activity, and population size according to well-defined scaling relations. Sato and Zenou (2015) analyzed interaction types and revealed that while individuals in densely populated regions interact with more people, these interactions are more random due to weaker social ties compared with residents in rural regions. Examining factors like distance and population density, Büchel and von Ehrlich (2020) discovered a positive correlation between cell phone usage and population, especially in close proximity, suggesting a complementary relationship between face-to-face and mobile interactions. Their findings validate the operation of economies of scale facilitated by cell phones. Moreover, Dong et al. (2017a) investigated how urban dwellers' social interactions influence their purchasing behavior: individuals working in nearby locations, despite living in different communities, often act as "social bridges" between their communities, leading to similar purchase behaviors within those communities.

Social Integration

Some studies have focused on the social integration of individuals and communities within urban areas. In the United States, Baum-Snow (2007) conducted a study investigating the impact of the interstate highway system, authorized in the Federal-Aid Highway Act of 1956, on some major metropolises. The construction of high-capacity roads to new suburban areas through existing Afro-American communities contributed to the decline and spatial isolation of these communities within cities as a result of *redlining* policies.[1] At the same time, it facilitated the migration of White middle-class populations to suburban areas. Dmowska and Stepinksi (2018, 2019) evaluated the long-term consequences of these policies by analyzing the spatial patterns of residential racial segregation in 41 American cities from 1990 to 2010. Interestingly, urban segregation extends beyond physical spaces. Morales et al. (2019) analyzed interactions on Twitter/X among urban residents across diverse European and American cities. Their findings demonstrated that the physical segregation of

[1] *Redlining* is a discriminatory practice of withholding services, particularly financial, in neighborhoods labeled "risky" due to high minority and low-income populations. This practice began in the United States with housing programs from the 1930s New Deal and initially targeted areas where Black residents lived.

some communities extended into the virtual space, visible through their online interactions and the diverse topics discussed.

Koramaz (2014) also investigated spatial aspects of urban segregation in Istanbul. She observed that social groups with lower levels of structural integration, particularly in the job market and education system, tend to reside in informally developed residential areas with poor environmental quality. Conversely, groups with higher levels of structural integration live in formally developed areas with optimal public services and environmental conditions. Beyond residential segregation, Legeby (2010) confirmed that the structure and layout of public spaces in Swedish cities also play a role. Bakker et al. (2019) analyzed large volumes of cell phone data to examine the social integration of Syrian refugees in Turkey. They found that refugees in Istanbul lived in more integrated neighborhoods compared with those living in less populated regions. Moreover, regions like southeastern Anatolia showed a higher positive correlation between refugee employment and their interaction with locals, indicating a potential relationship between job opportunities and social integration.

Quality of Life and Livability

Quality of life represents a dimension that can be complex to estimate as it depends on various factors. Dubois and Ludwinek (2014) compared quality of life in both urban and rural Europe by examining a spectrum of factors, encompassing subjective elements like life satisfaction and more objective metrics such as living conditions, material deprivation, trust in institutions, and social exclusion. Their research highlighted significant disparities in the perception of various indicators based on the place of residence. Residents in urban areas within some of Europe's wealthiest countries (e.g., France, the United Kingdom, or Germany) showed higher rates of social exclusion and dissatisfaction with their living conditions and accommodation compared with their rural counterparts. Conversely, in other Northern and Eastern European countries (e.g., Denmark, Finland, and Romania), opposite findings were observed.

A significant dimension in quality of life pertains to individual perceptions of happiness. Burger et al. (2020) discovered that, on average, urban populations tend to be happier than rural ones. They attributed this perception to factors such as higher living standards, higher access to diverse activities and services, and better economic prospects, particularly for individuals with higher educational attainment. Similarly, Leyden et al. (2011) highlighted that key factors contributing to this perception include physical accessibility, affordability, and a wider array of cultural and recreational amenities.

The correlation between urban form and quality of life has been a subject of examination in various recent studies. Residents in suburban areas of sprawling cities experience longer commuting times and often display lower subjective well-being (Clark et al. 2020; Stutzer and Frey 2008). Sapena et al. (2021) conducted an analysis of the spatial structures of 31 cities in North Rhine-Westphalia, Germany, revealing a significant correlation between the spatial structure (e.g., compactness, spatial distribution, and fragmentation of built areas) and quality of life indicators. Venerandi et al. (2018) found that the most deprived neighborhoods in the six major UK conurbations commonly exhibited higher population densities, larger areas of undeveloped land, an increased prevalence of dead-end roads, and more uniform street patterns. This observation aligns with a recent report from the Economist Intelligence Unit (EIU 2022) that ranked the most livable cities globally, where notably, none of these cities showcased a highly regular urban network. Mumford (1961) offered a compelling perspective, suggesting that American gridiron plans, designed for efficient car traffic, lacked differentiation between main arteries and residential streets. This oversight potentially prioritizes car traffic over sustainable transportation modes, potentially impeding social interactions among urban residents.

Health

According to the U.S. County Health Rankings, rural residents are more likely to have higher rates of obesity, sedentary behavior, and smoking habits, along with higher risks of various health issues such as diabetes, heart attacks, and high blood pressure (University of Wisconsin Population Health Institute 2022). Conversely, urban dwellers face greater exposure to air pollution, exhibit higher rates of sexually transmitted diseases, and are more prone to excessive alcohol consumption. Additional studies indicate higher likelihoods among urban dwellers to experience mental illnesses and depression (Fauzie 2015).

Effective urban design can affect the physical and psychological health of urban residents (Mehta 2014), especially benefiting active older adults. The presence of green spaces in cities correlates with lower morbidity by promoting physical activity, aiding psychological relaxation, and stress reduction (Braubach et al. 2017). The integration of more green spaces into urban streetscapes has been associated with better mental health and higher social cohesion among city residents (de Vries et al. 2013). Moreover, the structure of tree canopies contributes to mitigating traffic pollution, noise, and heat-related stress (Fisher et al. 2022; McDonald et al. 2020).

Urban form influences residents' health, though with some ambiguity. Compact cities may initially appear dense, potentially leading to traffic congestion and higher pollution levels. Many authors argue, however, that this model enhances efficiency concentrating and mixing land uses, fostering economic diversity, reducing work–home commutes, and encouraging sustainable transportation modes like public transit, cycling, and walking. This approach diminishes car usage and pollution levels substantially (Mansfield et al. 2015).

Crime and Safety Perception

Studies like Parkinson et al. (2006) commonly uphold the belief that crime rates tend to be higher in cities, predominantly concentrated in the most deprived neighborhoods. Labbrook (1988) conducted a study in Japan, suggesting that higher urban crime rates may be attributed to various demographic factors. These factors encompass quantitative aspects, such as higher population densities and growth rates, as well as qualitative factors, such as younger populations and higher immigration. Interestingly, although crime rates are generally higher in cities, they do not necessarily escalate in direct correlation with city size (Oliveira et al. 2017).

Understanding criminal patterns relies on urban design factors. Kimpton et al. (2016) showcased a negative correlation between crime rates and green spaces. Their study highlighted that the existence of green spaces, at both micro and macro levels, is linked to lower crime rates. This trend is observable on a global scale, even in areas known for high crime rates, such as South Africa. Venter et al. (2022) observed that for every 1% rise in overall green space within urban settings, there was a corresponding decrease of 1.2% in the rate of violent crime.

Various studies have juxtaposed real crime rates with perceptions of safety, investigating the impact of built environments. For example, Zhang et al. (2021) scrutinized streets in Houston by comparing officially reported crime rates to safety perceptions via Google Street View imagery. Their results revealed intriguing paradoxes: places with elevated daytime activity seemed safer than perceived, whereas those with increased nighttime activity were perceived as more hazardous.

Mapping Context

Throughout this chapter, we have explored the intricate relationship between urban form and social behavior by synthesizing insights from various papers covering diverse social behavior topics. Some studies base conclusions on limited datasets or confined areas, whereas others speculate on the impact that urban forms have on social behavior. Notably, most emphasis focused on examining how urban morphology influences social behavior, thus revealing a research gap in investigating the reciprocal relationship and prompting the need for further exploration in future studies.

The proliferation of big data has ushered in a wealth of building-level and individual-level information, providing a robust framework for understanding the bidirectional relationship between urban form and social behavior (Balsa-Barreiro et al. 2018). Individual-level data facilitate the reconstruction of mobility patterns, purchase behavior, and social interactions in both physical and virtual spaces, gathered at high frequencies across extensive populations. To derive meaningful insights into human behavior, managing aggregated data becomes crucial, ensuring the confidentiality and privacy of information (Carballada and Balsa-Barreiro 2021; Hardjono et al. 2019). Concurrently, the extraction of building-level data offers a spectrum of variables and indicators pertinent to urban forms. Table 8.2 outlines various data sources and technologies utilized for data collection at both individual and building levels.

Conclusions

Cities are expanding rapidly and evolving into the predominant dwelling for the global populace. Future projections underscore a heightened inclination toward urban lifestyles in the forthcoming decades. As we confront substantial global challenges, cities will bear a considerable impact, underscoring the critical necessity to delve deeper into the underlying factors influencing urban

Table 8.2 Sources and information technologies for data collection related to social behavior at the individual level (I-L) and urban forms at the building level (B-L). P: physical space; V: virtual space.

Level	Data Source	Information	Data Description	Scope
I-L	Mobile phones	Call detail records Apps (profile, type)	Social/communication/mobility patterns Social/purchase behavior	P/V P/V
	Social networks	Interactions	Social patterns	P/V
	Personal wearables	Various	Social/health patterns	P/V
	Crowdsourcing	Volunteer data	Social/communication/mobility patterns	P/V
	Banking	Credit card transactions	Purchase behavior	P
	Mobility services	GPS traces	Social/mobility patterns	P
	Surveys	Experimental	Social/communication/mobility patterns	P
B-L	Aerial imagery	Imagery	Urban form/greenery	P
	Remote sensing imagery	Imagery	Urban form/greenery	P
	Laser scanner	Point cloud	Urban form/digital elevation models	P
	Cadastral plans	Thematic data	Urban form/building heights	P
	Official reports	Thematic data	Urban form/household data	P
	Census	Socioeconomic data	Household data	P
	Historical maps	Thematic data	Urban form	P
	Photogrammetry	Imagery/point cloud	Urban form/digital elevation models	P

functionality and their interconnectedness with human behavior. Nonetheless, the intricate nature of cities, entangled within a web of multifaceted elements, leaves numerous inquiries and uncertainties unaddressed.

This chapter explored the intricate dynamics between urban morphology and social behavior. To achieve this, we analyzed pivotal aspects of both fields and conducted an in-depth literature review focusing on social aspects influenced by city configurations. Our exploration spanned human cooperation and altruism, human mobility, social interactions, social integration, quality of life and livability, health, and crime and safety perception.

The primary goal was to establish a comprehensive framework that facilitates a holistic understanding of the reciprocal relationship between urban form and social behavior. This study caters to a broad spectrum of interests across multiple disciplines, from urban planning to social sciences. The implications of our findings hold substantial significance for experts and policy makers, offering insights crucial for the development of future cities that prioritize sustainability and efficiency.

Acknowledgments

José Balsa-Barreiro expresses gratitude to the organizers and staff of the Ernst Strüngmann Forum for inviting him to the event. Both authors acknowledge the support of the New York University Abu Dhabi (NYUAD) Center for Interacting Urban Networks (CITIES), funded by Tamkeen under the NYUAD Research Institute Award CG001.

Human Behavior: Real and Digital

9

Leveraging Video Footage for Ethological Observation of Human Behavior

Virginia Pallante, Lasse Suonperä Liebst, Peter Ejbye-Ernst, Camilla Bank Friis, and Marie Rosenkrantz Lindegaard

Abstract

Originating in biology, the ethological approach to studying human behavior has increasingly spread across various disciplines, including the social sciences. In addition to offering biologically proximate and evolutionary explanations, ethology provides a methodological framework for systematically observing and analyzing human behavior in natural face-to-face settings. This chapter discusses the relevance of using the ethological approach for the study of human behavior, particularly by leveraging video recordings of public behavior for ethological observation. This prospect is demonstrated through an outline of recent video-observational research on violent and bystander helping behaviors. Further avenues are discussed to advance video-based human ethology.

A Video-Based Human Ethology

The use of digital data has the potential to reshape how social science is fundamentally conducted, as the digital footprint left on digital and social media platforms provides unique insight into human behavior (Blok and Pedersen 2014; Zhang et al. 2020). Digital data takes many forms and shapes. In this chapter, we argue that the use of visual digital data, especially video recordings of behavior in public places, offers a unique but as yet underutilized potential to examine human behavior. While people's online presence and digital footprint bear witness to many aspects of human social life, a great deal of human behavior remains nondigital in nature and leaves none or only a shallow digital footprint and may thus be better observed *in situ* (Molotch and Boden 1993). Here, video data offers great potential as it opens a window into the nondigital social world, where people behave bodily, are co-present, and often interact face to face (Nassauer and Legewie 2022).

By capturing our daily routines and rare encounters, public cameras provide a versatile tool for conducting detailed and unobtrusive field observational research on human behavior. Within social science, however, systematic naturalistic observation—whether video-based or conducted on-site—has been surprisingly underutilized compared to self-reports in qualitative interviews and surveys (Reiss Jr 1992). This indirect approach to studying human behavior allows understanding of people's motivations for their actions, but it provides only a coarse-grained picture of how people actually behave. As such, for decades there has been a call for wider use of naturalistic observation techniques within the social sciences (Baumeister et al. 2007). Reflecting this, Erving Goffman (1971), an extremely influential and early pioneer of the study of interpersonal behavior, suggested that the subfield of micro-sociology should be practiced as "interaction ethology," a kind of human ethology with a particular focus on the *interactional* aspect of social behavior. His realization was that ethologists had developed the most detailed methodological skill set and procedures for systematically observing human interaction *in situ*, and that this should be taken as a methodological model for how micro-sociology should be conducted. Recently, Goffman's vision has begun to show its methodological potential in the social sciences. While Goffman needed to rely on on-site observations of human behavior decades ago, high-quality video recordings (captured, e.g., by surveillance cameras and smartphones) are now available to scholars (Gerrard and Thompson 2011). When such video data is utilized for ethological observation, it opens groundbreaking possibilities for the study of human behavior (Nassauer and Legewie 2022; Philpot et al. 2019).

First, the sampling of human behavior through video recordings may dramatically increase the sample size of rare events, which an on-site observer may never (e.g., terrorist attacks) or only rarely witness (e.g., street fights) (Lindegaard 2022). Second, observation and quantification of human behavior through video recordings have higher reliability and precision than on-site observations. This is because the event and subsequent behavior can (and often needs to) be observed many times, in slow motion, to cross-validate records between observers. As such, evidence suggests that the dynamic or interactional part of social encounters cannot be reliably captured with on-site observations (Morrison et al. 2016). For example, while ethnographic participant observation has excellent ecological validity, its reliability in capturing micro-interactional details is low.

Finally, video recordings are a highly unobtrusive data source. In many countries, recording devices, such as surveillance systems installed by police and municipalities in public settings, are an accepted part of the natural environment. What the videos reveal, therefore, is people's unstructured behavior, unaffected by the observer. The unobtrusiveness of video data further has the benefit that even dangerous human behavior may be observed without exposing the observer to direct threat, as could happen in on-site observations (Lindegaard et al. 2020).

In sum, there is little doubt that if Goffman had lived to experience the digital era, he would have embraced video data, just as several of his students did (e.g., Collins 2008). This current chapter follows in Goffman's footsteps and highlights the value of a video-based human ethology to study face-to-face interaction. With this ambition, it must be acknowledged that other attempts have been made to develop video-based approaches for the micro-sociological study of social interaction, such as the Video Data Analysis approach (Nassauer and Legewie 2018). This latter approach does not, however, take its point of departure in ethology and therefore has no separate interest in biological considerations of proximate and evolutionary explanations or cross-species comparisons. In addition, the current approach puts a stronger emphasis on intercoder reliability tests and has an explicit ambition to test the generalizability of behavioral hypotheses and, as such, in quantitative and large-*N* applications.

The exciting prospect of video-based human ethology is that many of the assumptions regarding human interpersonal behavior embedded within the social sciences can be checked against systematic observational evidence (Mortensen and Cialdini 2010). Often, such reality testing is not a priority within the social sciences, reflecting a weak interest in replicable testing (Makel and Plucker 2014) and the fact that the available methods offer low reliability and validity for examining human behavior. This has put the social sciences in a puzzling position where, as sociologist John Levi Martin (2017:118) summarizes, "probably more is known about interactions between chimpanzees than interactions between humans." We believe that a video-based human ethology is one way to address these issues.

Video Observation as a Method

Video-based human ethology can be applied to the study of various human behaviors, and here we will focus on one area where its methodological value has been clearly demonstrated: the study of interpersonal violence. Traditionally, this field of study relied almost exclusively on self-reported data and laboratory experiments—despite the limitations of these methods for examining violent behavior (for a review, see Philpot et al. 2019): Self-reports of violence and other crimes are subject to social desirability and recollection bias, likely exacerbated by the distress of these events. Laboratory experiments are limited by the practical and ethical circumstances that actual violence cannot be realistically simulated.

The growing availability of video data offers a way to overcome this methodological impasse in studying actual, unstructured violence. In analyzing these data, we largely follow a procedure developed and applied within human ethology (Eibl-Eibesfeldt 1989; Jones et al. 2018). This involves a strong emphasis on inductive observation of the behavior under study to inform the construction of an ethogram or behavioral inventory with detailed behavioral definitions. In ethology, ethograms are taken as the point of departure to study the behavioral repertoire of a species (Lehner 1998). The development of an ethogram is the product of nonsystematic *ad libitum* observations to select which behaviors to include, particularly those that are more discernible, delimited, and repeated over time (Altmann 1974). These behaviors can range from social interactions to individual activities, postures, or movements. This phase includes testing and revising the inter-reliability of the ethogram by comparing the ratings of two or more independent observers. Once high agreement is reached, the ethograms are applied systematically to observe and code the behavior of interest.

Ethograms are often refined and validated in subsequent studies; the aim is to develop a standardized ethogram of a given category of behavior. To illustrate, consider bystander behavior at violent public events. Initially developed by Levine et al. (2011), the ethogram of bystander behavior has been applied and validated in a number of studies (Ejbye-Ernst 2022; Liebst et al. 2019; Philpot et al. 2020). The resulting standardized ethogram includes bystander behaviors such as "pacifying gesturing," "calming touches," "blocking contact," "holding, pushing, or pulling an aggressor away from the conflict," and "consoling a victim of aggression." An example of an ethogram for nonviolence includes face-touching behaviors (Liebst et al. 2022) and was developed during the COVID-19 pandemic to examine the potentially adverse (self-inoculation) effects of mask-wearing. This ethogram describes fine-grained distinctions between whether a person touched a mucosa area (e.g., the T-zone of eyes, nose, and mouth), which is the main entry point for viral infection.

Although video records allow for the application of the ethological method to humans, the recorded social contexts can differ compared to animal studies. Ethological research frequently focuses on closed animal communities, where repeated interactions among the same individuals are possible, and where kin and social relationships are known or can be determined through repeated observations of the same subjects. By contrast, public security cameras record public spaces where people are present for only a limited amount of time, and typically no repeated observations of the same person are possible. Even though video records may thus fail to document some of the dynamics that occur between affiliated individuals during recurrent interactions, they provide a realistic insight into what is at the core of urbanized human ecology: a social structure organized between interacting strangers (Christakis 2019).

Bystander Helping in the Wild

In our violence research, we have specifically utilized video observation to examine the role of bystanders in violent incidents. For decades, the leading theory of bystander behavior within the social sciences has been the so-called “bystander effect” hypothesis (Darley and Latane 1968). This theory posits that people lose their moral compass when present in crowds, and thus remain passive and apathetic when witnessing someone in need of help. In other words, in crowds, the responsibility for taking action is diluted among those present, which, in turn, inhibits the helping likelihood. The bystander effect hypothesis was initially developed to explain the case of Kitty Genovese, who was raped and murdered in public in New York in 1964, while 38 bystanders allegedly remained passive. To study bystander passivity, field experiments were conducted: researchers staged emergencies in public places and then documented how the likelihood of intervention decreased when co-present with additional bystanders.

Despite being initiated by real-life violent events, research soon became uncoupled from the reality it set out to explain, due to the experimental approach used. The staged emergencies were often very trivial in nature (e.g., people dropping coins in an elevator), thus questioning the generalizability of these results to actual violent events, such as the Kitty Genovese case. Stressing this concern, a meta-analysis showed that the bystander effect was attenuated in experiments that simulated emergencies with some level of danger, albeit none simulating direct violence (Fischer et al. 2011). For the most dangerous situations included, the analysis indicated that additional bystanders offered welcome support, making the intervention more—not less—likely. The problem remained, however, that no meta-analysis is better than the studies included. Without analyzing any violent studies, it could not provide ecologically valid insights into how bystanders act in actual violent events.

The field of bystander studies encapsulates the concern of Tinbergen (1963:411) that researchers “skipped the preliminary descriptive stage that other natural sciences had gone through, and so was soon losing touch with the natural phenomena.” A reality check is needed, based on detailed naturalistic observations of real-life bystander behavior (Lindegaard 2022). The first video-based study of this kind was conducted by Levine et al. (2011) who, in direct contradiction to the bystander effect narrative, showed that bystanders play an active and effective role in regulating violent events. Building on this insight, Philpot et al. (2020) conducted a video-based study to investigate whether bystanders intervened in 219 street violence assaults captured on video in the Netherlands, the United Kingdom, and South Africa. They found that in nine out of ten situations, at least one, and typically four, bystanders did something to help the victim (Philpot et al. 2020). Furthermore, it was found that the likelihood of victims receiving help increased with the number of bystanders present. In other words, intervention is the norm, and there is safety in numbers. This is the reality of real-life bystander behavior outside the artificiality of the experimental setting. Characteristically, this was also the case in the Kitty Genovese case: historical analysis has documented that bystanders actually tried to intervene, although unsuccessfully (Manning et al. 2007).

These insights represent only the beginning of an ecologically valid and detailed understanding of bystander behavior. With the use of video-based ethological methods, a range of additional insights into bystander behavior has been revealed:

1. Regarding the causes of individual bystander intervention, the video data offer mixed results concerning the number of bystanders present, in contrast to the uniformly robust result that social relationship ties between bystanders and conflict victims dramatically increase the likelihood of individual intervention. Friends help friends (Liebst et al. 2019; Lindegaard et al. 2017). This conclusion is consistent with social psychological and evolutionary theory, which stresses that individuals have stronger empathic feelings toward in-group members with whom they have interdependent social ties (de Waal and Preston 2017; Stürmer et al. 2006).
2. Further, the level of violent danger is a very influential predictor of intervention. This suggests that people act when it really matters, especially when events become explicitly aggressive and dangerous (Lindegaard et al. 2021).
3. Bystander intervention is not a single act, as often portrayed in experimental settings. Instead, it is an intervention trajectory involving various actions that follow a specific behavioral pattern. Bystanders who intervene tune into the aggression level of the conflict, and stopping the fight requires consistent insistence and preparedness to scale up the intervention intensity (Ejbye-Ernst et al. 2021).
4. Relatedly, bystander intervention is not merely performed by an individual but is typically carried out in collaboration with others. This is because the violent conflict may require the actions of several individuals acting in concert to be stopped (Bloch et al. 2018; Levine et al. 2011; Weenink et al. 2022).
5. Bystander interventions may take place during all phases of the conflict, including in its aftermath where bystanders may provide consolation to victims of aggression (Bloch et al. 2018; Lindegaard et al. 2017). This behavior is similar to what has been documented among human children (Verbeek 2008) and nonhuman primates (de Waal and van Roosmalen 1979).
6. Given the high bystander intervention rate, there might be a concern that intervening bystanders may be victimized themselves when helping others. In general, however, the likelihood of bystander victimization is low (around 5%), and if victimization occurs, it is often relatively non-severe (Liebst et al. 2020).
7. Bystander intervention is actually effective in terminating violence, especially when performed as forceful interventions rather than as mere expressions of disapproval (Ejbye-Ernst 2022).

Prospects and Challenges

Video-based human ethology holds great promise, but the journey ahead is replete with a plethora of possibilities and unaddressed issues. Numerous questions emerge from the fact that although humans are a great ape and should be studied as such (Turner and Maryanski 2018), we are also different—not in kind but in degree (Darwin 1871)—from other animals and primates. This difference has important methodological and theoretical implications. *Homo sapiens*, like every other species, has unique characteristics that must be considered. Critical specificities for humans include evolved cognitive skills, which enable advanced capacities for collaboration, symbolic communication, and cultural learning (Tomasello and Herrmann 2010; cf. Bard et al. 2021). Specialized methods and theories have been developed to grasp these human social qualities, which cannot be fully captured through ethological observation of nonverbal behavioral displays (Geertz 1973).

The limitations of human ethological observation may be further magnified by the technology of public security cameras, which typically do not capture sound and thus do not permit content analysis of verbal communication. Considering that ethology is the study of behavior, and that humans often use speech when they interact (Austin 1975), video data is not optimal for examining verbal human behavior. With respect to the study of violence, this is crucial because during the initial phase of conflicts or in low-intensity disputes, verbal exchanges often unfold prior to the use of physical force (Friis et al. 2020). Thus, the inherently dispute-related nature of many violent crimes, involving mutual verbal insults and retaliations (Felson 1982), cannot be fully grasped with public security cameras.

One way to overcome this limitation is to analyze how verbal behavior is often expressed in conjunction with nonverbal cues (Eibl-Eibesfeldt 1989) and to use this to make some rough inferences from observations of nonverbal behavior to their verbal counterparts. Alternatively, scholars are using video data recorded by devices that capture sound, such as mobile phones and body cameras (Friis et al. 2020; Sytsma et al. 2021). Finally, the lack of sound may be compensated for by triangulating with other verbal data sources (e.g., interview data in combination with video-observational data). This could provide insights into the cultural, motivational, and meaningful content of social life, which aids in understanding why people do what they do (Friis 2022; Small and Cook 2021).

Furthermore, video data combined with additional information on the locations of public security cameras may offer a fruitful basis for explaining the behavioral data captured on camera. For instance, Sampson and Raudenbush (2004) combined systematic observations with census data, police records, and surveys to examine whether racial stigma, the economic context, and the actual observation of social disorder shape how people perceive social disorder. Similarly, in the analysis of criminal events, participant observation and interviews with people in specific locations may help to contextualize and explain observed interactions. These data sources provide information on the characteristics of the neighborhood as well as the people living in the area, including the subjective motivations that underpin the observed behaviors (Lindegaard and Bernasco 2018).

When adopting ethological insights to study human behavior, it is necessary to consider the extent to which its methodological aspects are distinct from its theoretical aspects. For Goffman, inspiration from ethology should be methodological, not theoretical. In his view, the ethological application of a "Darwinian frame" leads to "very unsophisticated statements," but "if we politely disattend this feature of ethology, its value for us as a model stands clear." (Goffman 1971:xvii). While Goffman seems to be referring to a reductionist evolutionary perspective that was prevalent at the time, contemporary ethology and evolutionary theories today are interested in questions central to sociological reasoning (Meloni 2014): prosociality, empathy, and how social relationships create group structures and influence conflict management strategies (de Waal 2000; de Waal and Preston 2017). As such, recent micro-sociology, inspired by Goffman, is engaging in fruitful dialogues with evolutionary and biological schools of thought, which support rather than erode the importance of sociological mechanisms (Heinskou and Liebst 2016; Lindegaard et al. 2017; Turner and Maryanski 2018).

A precondition for this type of interdisciplinary exchange is cross-species comparisons of behaviors between humans and nonhuman primates (Turner and Maryanski 2018), and in this area, we lack human *adult* ethological data. The limited evidence available is biased toward human children (Verbeek 2008), leading to the constrained conclusion that "other primates are *mentally* like human children" (de Waal 1989:249). Using a video-based human ethological approach, Lindegaard et al. (2017) conducted the first study on human adult post-conflict consolation behavior, comparing the observed patterns with those of chimpanzees (Lindegaard et al. 2017). We strongly recommend that future research examine other human adult behaviors with a view toward cross-species comparisons.

Video-based interaction ethology offers a way to compare human behavior in different conflict phases, conflict types, and cultural contexts. Behavioral variations are found in different steps of conflicts—for example, affiliative touching is more frequent in the aftermath than before or during robberies (Lindegaard et al. 2017; Philpot et al. 2022)—and intervention behavior is more physical at the end of the conflict than at the beginning (Ejbye-Ernst et al. 2021). Further, cultural comparisons allow us to theorize about the mechanisms of the observed behavior and may help us understand the extent to which human nature is universal (Brown 1991). Cross-cultural comparisons between South Africa, the Netherlands, and the United Kingdom, for instance, reveal similar bystander intervention frequencies in street violence events (Philpot et al. 2020).

Bystander intervention in street fights might be, however, conflict-type specific and thus not generalizable to other kinds of conflict (e.g., robberies, partner violence, war atrocities). For example, in an analysis of bystander intervention during armed robberies in the Netherlands, we found that bystanders only intervened in a minority of robberies (unpublished data), in contrast to street violence (Philpot et al. 2020). Further, when bystanders do intervene in robberies, the risk of victimization is much higher during armed robberies than in street fights (Liebst et al. 2020). Such examples highlight variations in bystander behavior across different types of conflict, potentially related to differences in conflict dynamics, cultural settings, and causal mechanisms, and underscore the need for further research in this area.

The primary strength of the ethological approach is its focus on detailed and naturalistic description (Lorenz 1973) yet integrating this with a focus on explaining causal mechanisms remains challenging. For Goffman, this was less of a concern in his vision of interaction ethology, as he deliberately refrained from moving beyond description to test causal hypotheses (Verhoeven 1993). Recent Goffman-inspired research using video data, however, argues that such a step should and could be taken, given the strength of video observation is how it allows one to “study if there is causality at the microlevel” (Nassauer and Legewie 2018:163). Considering the studies mentioned above, the issue is that most rely on cross-sectional, between-subject (or between-situation) designs—a weak approach for testing causality. One solution could be to employ field-experimental methods more extensively, as it is commonly done in ethology (Cuthill 1991), despite the obvious ethical limitations with respect to how violence or danger may be simulated in field experiments.

An alternative to testing causality in aggression and bystander behaviors is to match subjects with themselves under different study conditions, a powerful method for controlling both observed and unobserved confounders (Dawkins 2007). Typically, this involves observing the same subjects under different situations, which is often not feasible with public video data. Recently, however, a few studies have shown that subjects can be measured several times *within* unfolding situations, allowing for a fixed-effect panel regression approach that is considered a robust approximation to causality (Listl et al. 2016). This was done, for example, in the cited video-based study that examined danger levels as a predictor of intervention (Lindegaard et al. 2021): By following the same individuals throughout the unfolding situation, we established which level of danger caused the bystanders to intervene. The success of video-based human ethology hinges on how its descriptive and causal-explanatory potentials are united. This should be a priority in future work.

While one of the strengths of video-based human ethology is its high ecological validity and reliability, a potential weakness is its generalizability, often due to working with nonrepresentative samples of low statistical power (Taborsky 2010). Manually coding behavior second by second is very labor-intensive. To increase sample sizes, computer programs could be used for automatic behavior annotation, creating observational datasets much larger than would otherwise be possible; this approach may also be considered digital ethology (Anderson and Perona 2014). Specifically, computer vision scientists have worked for decades on training algorithms to detect automatically different kinds of behavior in video clips (Jain et al. 2015). Instead of relying on costly and potentially biased human observers, such as municipal employees or law enforcement agents, computer vision tools could identify relevant study situations from large pools of video clips. Although using computer vision tools to identify conflict situations might yield numerous false positives, filtering out these erroneous clips would still significantly reduce the time required for sampling relevant situations compared to human observers. An example of integrating computer vision tools in video-based interaction ethology is our research project investigating social distancing behavior during the COVID-19 pandemic. We were able to measure automatically when people failed to keep the recommended distance from each other on the street. Using computer vision enabled us to analyze the behavior of over half a million individuals across thousands of hours of footage (Bernasco et al. 2022), a task that would have been impossible with human coders.

Closing Remarks

While video-based human ethology shows significant potential for future research on human behavior, its primary development has been within the domain of interpersonal violence. In our view, many other fields could benefit from utilizing video observations. Broadly speaking, the use of this approach may be one means to make the social sciences a more high-consensus and rapid-discovery science, similar to what has been achieved within the natural and medical sciences. Compared to these disciplines, a limitation of the social sciences is that they are to a lesser degree propelled forward by innovations in research hardware and technologies (Collins 1994). For example, Galileo's brilliance was not only his novel ideas but how he made use of research hardware and technologies: lenses assembled into microscopes and telescopes that led to a series of groundbreaking discoveries. The social sciences have embraced such research hardware to a lesser extent, but this is likely to change with the advent of a more hardware-driven and computational science that harnesses the potential of digital data, simulations, and artificial intelligence (Sallach 2003). For the micro-sociological study of interpersonal behavior, video technology is specifically suggested to hold potential for scientific advancement, given its possibilty to map the micro-world of human behavior (Collins 1994).

While this application remains to be fully embraced within academia, the groundbreaking potential of video data has already proven its worth outside academia. In a certain sense, living in contemporary society implies being a video-trained human ethologist, given our massive exposure to video-recorded content. Video data allow us to see behavioral realities that cannot and should not be unseen, whether as scholars or citizens (Goold 2006). Poignant examples includes the murder of George Floyd in 2020, where security and witness footage drove the global outrage over the atrocity we all observed, or the video documentation of war crimes in Syria and Ukraine. Video technology, ever-present in contemporary society, is already revolutionizing our perception of the world.

Acknowledgments

This study was funded by the Dutch Research Council (NWO VI.Vidi.195.083) by means of a grant awarded to Marie Rosenkrantz Lindegaard. This project has also received funding from the European Union's Horizon 2020 research and innovation program under the Marie Sklodowska-Curie grant (101029161) awarded to Virginia Pallante.

10

Geolocation-Centric Monitoring and Characterization of Social Media Chatter for Public Health

Abeed Sarker

Abstract

The adoption of social media is currently at an all-time high. More than half of the world has access to social media. The large-scale adoption and growth of social media have demonstrated the benefits and drawbacks of human activities over such platforms. As the digital footprint of human behavior via social media platforms continues to evolve, it is essential to identify strategies and execute actions that can utilize the data generated for the benefit of humankind. Since most of the human footprint on social media is in the form of free text, the field of natural language processing holds substantial promise in converting such data into valuable and actionable knowledge. Geolocation-related metadata available with or inferred from social media posts enable knowledge to be aggregated at various spatiotemporal granularities. Fine-grained area-level insights about human behavior can, for instance, be obtained through social media-based surveillance in close to real time. Geolocation-specific statistics derived from social media data may also be combined with other area-level data from more traditional sources to obtain comprehensive knowledge on chosen topics. Following a brief introduction to social media and natural language processing, the utility of social media data, particularly when combined with geolocation-based information, is discussed. Two examples—COVID-19 and substance use—are used as case studies.

Introduction

Social media refer to Internet-based platforms over which communications involving text, voice, video, and/or images take place. Growth in the use of social media has been primarily driven by social networking websites, which enable people to connect with others and share information. The adoption of social media is currently at an all-time high, and it is estimated that over 4.5 billion people in the world use social media (Statista 2022c). Despite the staggering number of existing social media users, the adoption of such platforms continues to grow. Globally, the most commonly used social network is Facebook. Other popular social networks include but are not limited to Instagram (primarily used for image sharing), Twitter/X (supports microblogging), YouTube (video sharing), and Reddit (topic-specific forums that allow subscribers to remain anonymous if they desire). While social media are still disproportionately popular among younger people, adoption is currently happening at a faster rate among older people according to the Pew Research Center (2021). As demographics shift, it is only a matter of time before the global social media user base becomes quite accurately reflective of the world population. In fact, there is perhaps no other platform currently available that has a better reach than social media.

The widespread use of social media has resulted in the continuous generation of massive data. Such data encapsulate knowledge on essentially any topic. Connected networks also enable the rapid dissemination of information to many people, typically without any geolocation-based limitation. Both the volume of knowledge and the rapidity with which it can spread have the potential to be leveraged to determine and influence population-level behaviors. Consequently, over the last decade, social media platforms have been utilized for a variety of purposes, including (but not limited to) politics, health, and finance. The role of social media in the presidential elections of the United States, for example, has been extensively studied (Bossetta 2018). In the broad field of finance, the power of social media-based communication and behavioral influence was demonstrated in 2021 when a group of subscribers coalesced on a Reddit forum to invest collectively in stocks of GameStop—a company in the United States that was on the verge of bankruptcy according to many institutional investors (Anand and Pathak 2022). It was reported that the collective trading of small investors on Reddit in January 2021 surpassed the previous trading volume record set in 2008 in the New York Stock Exchange by a factor of six (from approximately four billion shares to 24 billion). This collective behavior, which was specific in the United States from the perspective of geolocation, led to a steep, unprecedented rise in the market valuation of the company, by over 1000% in two weeks, baffling institutional and seasoned investors. These events demonstrated the utility of social media and the influence that social media-based human activities can have within specific spatial and temporal windows. The utility of social media-based data for health-related tasks, particularly the possibility of deriving geolocation-specific insights for public health, has been realized over recent years, and substantial research efforts are currently ongoing to utilize data effectively from this ever-growing resource. The primary focus of this chapter is to outline some of the opportunities associated with social media data in the realm of public health, with particular emphasis on the geospatial aspects, and the research challenges that such data present. Two case studies—COVID-19 and substance use—are used to illustrate the use of social media data in real life.

Social Media and Health

A considerable portion of chatter on social media is concerned with health-related topics. People often share their health problems, discuss treatment options and efficacies, ask questions, describe personal experiences, and provide suggestions, including self-management strategies for myriad health conditions. These discussions capture important information about health topics in an unstructured form. Such data are often referred to as patient-generated big data and have been shown to contain information not available through other, more traditional, sources such as electronic health records and published literature. The information is typically enriched with metadata[1] including geolocations, which may enable spatial aggregation. Even when geolocation information is not explicitly present in the metadata, researchers have developed tools that can estimate geolocation based on other profile-level data (Dredze et al. 2013). In theory, patient-generated social media data can be categorized, aggregated, and analyzed to obtain population- and area-level insights in close to real time and at low cost. Importantly, the data are collected in an unobtrusive manner, which may mitigate biases that typically arise in synthetic experimental settings (Fan et al. 2018). The value of patient-generated data from social media for public health has been realized over recent years, and it is being used increasingly for health-related tasks, such as pandemic surveillance (Chen et al. 2020), pharmacovigilance (Sarker et al. 2015), mental health-related topics (Chancellor et al. 2021), and substance use surveillance (Sarker et al. 2019), to name but a few.

Challenges and Limitations of Social Media Data Processing

While the knowledge contained in social media big data holds considerable promise, the extraction and utilization of such knowledge have been limited for years by our capabilities, or lack thereof, in big data and natural language processing (NLP). NLP is the field of computer science that broadly addresses the problem of automatic understanding of human language in text or verbal form. The flow of natural language, by nature, is nondeterministic; thus, traditional, rule-based computational models are not capable of effectively processing such data. Automatic processing of health-related natural language data from social media is particularly difficult due to the presence of colloquial expressions, misspellings, noise, and context-ambiguous statements. The conversion of health-related social media big data into valuable and actionable knowledge has required the development of advanced NLP and machine-learning (ML) methods—research areas in which enormous advances have been made in recent years. We are therefore at an important point in time in our understanding

[1] Data that summarizes or provides additional information about other data.

of how best to leverage social media chatter for improving public health, including in the context of geolocation-centric surveillance.

Currently, social media text mining systems employ pipelines of NLP and ML modules that gradually filter out the noise and convert unstructured chatter into aggregated knowledge. NLP methods do not have to rely on explicitly coded rules; rules are learned automatically from the chatter itself via ML methods, which have in some fields reached human-level performances (Montejo-Ráez and Jiménez-Zafra 2022). Within NLP, the most exciting advances have perhaps been brought about by innovative strategies in text representations. Early NLP methods simply used rules such as character patterns (often referred to as regular expressions) on the text-based representations. The incorporation of ML into NLP approaches necessitated the use of vector-based representations of texts, resulting in the creation of sparse vector models such as the bag-of-words[2] and n-gram[3] models. The next leap was in the generation of dense vector representations of words or phrases that required large, unlabeled datasets—of which there is an abundance on the Internet and social media—and the representations were capable of capturing the semantics of the texts such that similar words/phrases would appear close together in vector space (e.g., word2vec models; Mikolov et al. 2013). One shortcoming of such word- or phrase-level models was that they were unable to capture contextual differences; for instance, homonyms[4] would have the same vector representations. These challenges were overcome very recently with the creation of contextual vector models that better captured meanings with large sequences of texts, as in the bidirectional encoder representation from transformers or BERT (Devlin et al. 2019). In addition to these advances in text representation, the capabilities of computing large volumes of data and optimizing complex ML models have also made large strides. While many challenges still exist in the automatic processing of health-related natural language data (e.g., in cases when the relevant concepts are sparse or rare), advances have enabled the utilization of social media chatter for many targeted tasks. Parallel advances in geolocation inference strategies when metadata are not available (Harrigian 2018; Mahajan and Mansotra 2021) have improved our capabilities to conduct geolocation-specific studies.

In addition to the technical challenges associated with mining knowledge from social media, there are limitations inherent to this resource that may not be solvable through technological advances. At the area-level, a major limitation concerns the issue of representativeness. Social media cohorts at specific

[2] A text representation commonly used for sentences or documents. Each word is represented as a number, in a list or vector, that specifies its presence/absence or count. Word order is not preserved.

[3] A text representation approach that uses contiguous sequences of *n* words. Unlike bag-of-words models, n-gram models preserve information about word sequences.

[4] Two or more words with the same meaning or pronunciation but different meanings.

geolocations are not necessarily representative of the entire population. It is well known that social media data generally underrepresent older age groups while overrepresenting younger ones. Representations may also vary based on the problem being studied. Given a specific health problem, certain segments of the population may be more likely to self-report personal information than others. In the case of substance use, discussed in more detail below, studies have shown that college students are more likely to report nonmedical use of stimulants (Sarker et al. 2016), whereas opioid use may be underreported due to stigma and other factors (Chenworth et al. 2021; Graves et al. 2022). In areas where substance use is criminalized, people may also underreport compared with those living in areas where the issue is treated as a public health issue. The extent to which such under- and overreporting happens among particular cohorts is not fully understood. Absent this knowledge, the best strategy to validate findings from social media is perhaps to compare them with information from traditional sources, such as surveys. Some recent studies have attempted to calibrate problem-specific demographic distribution statistics by developing automatic methods to detect self-reported demographic information (e.g., gender, age-group, and race) and then adjusting the distributions against the distribution detected using the same methods from generic social media data (Yang et al. 2023). Such methods are promising, but the limitations associated with representativeness, and other limitations of social media data, remain important open problems.

Types of Social Networks and Their Contents

While this discussion has mostly projected social media as a sphere of homogeneous data, in reality, that is not the case. Data generated over each social network are unique, as are the utilities associated with the data. Facebook, Twitter/X, Instagram, and Reddit, mentioned above, can be broadly classified as generic social networks. On such networks, subscribers can essentially post on any topic they desire. Consequently, much of the content can simply be considered to be noise, and NLP pipelines processing the chatter must first filter out such noise. The structures of the posts can also be significantly different. Facebook and Reddit, for example, allow long posts. In contrast, Twitter/X posts are length-limited, and so posts are short and often lack context. In addition to these generic social networks, others are dedicated specifically to health-related topics (e.g., MedHelp, PatientsLikeMe), and are generally rich in information but lack metadata, such as geolocation, and attract lower numbers of daily active users. The distinct structures and contents of these social networks have naturally led to distinct digital footprints of their subscriber cohorts. Since subscriber behaviors evolve over time based on the characteristics of the social networks, these differing behaviors provide exciting data for digital ethology.

Geolocation-Centric Data Analysis and Application Programming Interfaces

As mentioned above, the knowledge encapsulated in social media data goes beyond just natural language chatter or images. Most social networks allow subscribers to make geolocation information visible in their posts. Posts by subscribers on Twitter/X, for example, often contain geolocation information in the form of exact coordinates or information obtained at the city or state level. Such information that complements the contents of social media posts are called metadata. Metadata, such as geolocation and timestamps, are automatically encoded in the posts. Therefore, geolocation-based metadata, when available, can be used to study geolocation-centric digital behavioral patterns among the subscribers. Data posted at specific geolocations by many subscribers at defined time periods can be aggregated and analyzed to study subpopulation-level behavior digitally. Studying aggregated data from many subscribers, as opposed to data from a single subscriber, is invariably more valuable from social media sources. Individual subscribers may not post all information relevant for behavioral or other analyses, but when posts from large numbers of subscribers are aggregated, the most important topics relevant to that group of subscribers tend to become visible as they surface above the rest. Aggregating by geolocations may reveal important distinctions in topics relevant to people from different locations.

Due to the growing utility of social media data, many platforms have made them available through application programming interfaces (APIs), which allow computer programs to connect to the data streams on networks and collect data based on the relevant protocols. Twitter/X, for example, recently released an academic API to support noncommercial research. This API allows researchers to collect the contents of the posts as well as the metadata associated with such posts. Two key metadata elements that have been utilized heavily in research are timestamps and geolocation. Specifically, these meta contents are used to aggregate posts on the platform, given a specific time and topic, and to analyze them over specific geolocations. As mentioned earlier, recent studies have also proposed methods for inferring geolocation from social media posts when explicit geocoding is not available (Dredze et al. 2013; Mahajan and Mansotra 2021). These inference methods have substantially increased the proportion of posts that can be aggregated by geolocation to derive insights. Researchers use geolocation-specific data for tasks such as infectious disease outbreak surveillance, and a number of recent studies have utilized such data to study the COVID-19 pandemic. Below, we look at two case studies that have utilized metadata from Twitter/X for geolocation-centric analyses.

COVID-19

The pandemic caused by the novel coronavirus provides a current example of a recent global health crisis and has received considerable research attention since the outbreak of the virus in late 2019. This research led to the development of effective mRNA and other vaccines in record time. Ongoing research includes, but is not limited to, studies that focus on the long-term impacts of COVID-19 infection (typically referred to as long COVID), identification and analysis of new mutated variants of the virus, and methods for detecting potential future outbreaks in a timely manner. There is now general understanding and acceptance that future infectious disease outbreaks like COVID-19 may happen. It is also generally accepted that no one mechanism of infectious disease surveillance is by itself sufficient to provide timely alerts; a combination of approaches is required. Localized infectious disease outbreaks, including future variants of COVID-19, can exert tremendous strain on health systems, causing large numbers of deaths (Carinci 2020), as was observed in some countries (e.g., Italy and Spain) as well as in big metropolitan cities (e.g., New York City) during the early waves of the pandemic. Traditional surveillance methods struggled to keep up with the pace of the outbreaks due to the time and effort required to compile data (González-Padilla and Tortolero-Blanco 2020; Gupta and Katarya 2020; Lakamana et al. 2022; Sabouret et al. 2020), which typically comes from sources such as hospitals. The need to develop novel surveillance strategies with the potential to forecast upcoming outbreaks was realized during the COVID-19 outbreak. Infodemiology-oriented data-centric methods for surveillance (Eysenbach 2009), such as those that rely on social media posts, have the potential to detect patterns in chatter associated with geolocation-specific outbreaks and provide timely alerts to relevant health agencies.

Social media proved to be of high utility during the early COVID-19 outbreaks, as it became the primary mode of communication for many, particularly after "lockdowns" and/or "social distancing" measures went into effect. Research during the early months of COVID-19 revealed that social media chatter was rich in first-person reports of COVID-19 positive test results (Guo et al. 2021; Myrick and Willoughby 2022). Many people shared the symptoms they were experiencing, often with day-to-day updates. Research also showed that these self-reports of positive test results and expressions of symptoms can be detected and extracted automatically using NLP methods. In fact, early research showed that about one-third of the people discussed symptoms up to two weeks before they tested positive for COVID-19, and some relevant symptoms were reported before their associations with COVID-19 were common knowledge. For example, the first report of anosmia was observed on Twitter/X in the first week of March, while Google Trends showed that search queries for the symptom peaked after March 20, 2020 (Sarker et al. 2020). This suggests that information specific to COVID-19, including self-reported symptoms, may be available and detectable from social media. Self-reported symptoms represented only a subset of all COVID-19 topics covered in social media chatter, and the knowledge generated often preceded those from other, more traditional, sources. Findings from numerous studies conducted during the COVID-19 pandemic suggest that strategic mining of knowledge from social media chatter can provide valuable and timely insights about infectious diseases, including insights on upcoming outbreaks and potential communities at risk, enabling relevant experts to plan appropriate responses (Callard and Perego 2021; Chen and Wang 2021; Golinelli et al. 2020; Li and Zhang 2021; Matharaarachchi et al. 2022; Tsao et al. 2021; Turiel et al. 2021).

Geolocation-Centric Surveillance

In response to the COVID-19 pandemic, Twitter/X released a customized API for the collection of data in real time. Combining metadata, particularly about geolocation, with NLP methods showed excellent potential for conducting localized surveillance of outbreaks. Since this was the first global pandemic in the era of social media, the data collected since its beginning served as testbeds for innovative approaches. In summary, research showed that social media chatter provided both challenges and opportunities for geolocation-centric monitoring. An outline of these is provided below.

Prior to COVID-19, studies using social media to detect infectious disease outbreak relied primarily on volumes of data that emerged from targeted geolocations. Keyword-based methods were used to detect relevant social media posts (e.g., about *flu* or *flu-like* symptoms) to estimate the incidence of new infections (Broniatowski et al. 2013). Such methods proved too simplistic in the case of COVID-19, as the behaviors of people over digital platforms evolved substantially during the COVID-19 pandemic compared with past infectious outbreaks. First, the volume of chatter associated with COVID-19 was dissimilar to any previous infectious disease outbreak. Once awareness about it increased throughout the population, posts on the topic surged. The vast majority of posts were not first-person experiences; rather, they were mostly focused on sharing information. Second, the large-scale sharing of information also led to the spread of misinformation, which included but not limited to, conspiracies about the pandemic and vaccines, promotion of fraudulent products for the treatment and diagnosis of COVID-19, and the sharing of unverified news. Consequently, an effective strategy for conducting geolocation-centric surveillance proved to be a multistep process.

The first step in utilizing social media chatter for geolocation-centric monitoring of infectious disease outbreaks is to *collect* the right data. For COVID-19, the specialized API provided by Twitter/X served this purpose. The metadata that accompanies the posts often includes geolocation information. Even if a small fraction of the post-level metadata contains geolocation information, having a large volume of data generated on a daily basis enables the collection of very representative geolocation-specific behavioral digital footprints on the chosen topic. For a topic such as COVID-19, however, most of the digital footprint may be noise or misinformation; thus, a crucial step in processing is to *characterize* the data so that irrelevant or unwanted content (e.g., misinformation) can be separated from relevant or useful content (e.g., firsthand reports of positive tests). This characterization problem is also perhaps the hardest to automate. Here, the latest developments in NLP research can help. To solve this characterization step, recent studies have proposed modeling it as a supervised classification[5] problem. Supervised classification is an ML approach where models are trained based on manually annotated data. In this case, efforts were made to annotate data manually to identify misinformation, firsthand reports of symptoms, and informative contents (Gerts et al. 2021). Next, state-of-the-art supervised classification models, such as transformer-based ones (Li and Zhang 2021; Nguyen et al. 2020), were trained on the annotated data and deployed to characterize streaming data automatically. Posts deemed to be relevant are mapped onto their origin location. In the United States, significant correlation (Spearman $r=0.88$, $p=0.000$) was found between the distribution of automatically detected posts at the state level and real COVID-19 case counts. Figure 10.1 shows the population-adjusted distribution of automatically characterized Twitter/X posts from early 2021 at the state level.

Social media-based surveillance is not limited to the United States or high-income countries. Due to its widespread global adoption, social media-based surveillance can be implemented almost anywhere in the world. This may be particularly useful for geographical areas where testing services are limited or slow and traditional surveillance of outbreaks is ineffective. To test the utility of social media-based geolocation-centric monitoring outside of the United States, one study focused on India—a large country with a population of over one billion where surveillance at the national level is extremely challenging (Lakamana et al. 2022). In the study, between February and June 2021, over 500,000 tweets about COVID-19 were geolocated to be from India. The chatter about COVID-19 increased almost at the same time as the number of confirmed cases in India, with a high correlation (Spearman $r=0.944$; $p=0.001$). The top tweeting states were Maharashtra, Karnataka, Tamil Nadu, and Uttar Pradesh—states that also recorded some of the highest numbers of COVID-19 cases. There was also a significant correlation between the state-level case numbers and the number of tweets emerging from those states (Spearman $r=0.84$, $p=0.0003$). Fatigue, dyspnea, and cough were the top reported symptoms emerging from India, and emotion analysis showed a surge in negative emotions in 2021 compared with the previous year. Anxiety levels and concerns about black fungus (mucormycosis) also surged—the latter was known as a problem during the outbreak there. The strong correlations between actual

[5] A machine-learning approach in which labeled examples are used to train a model, which is then used to classify unlabeled samples.

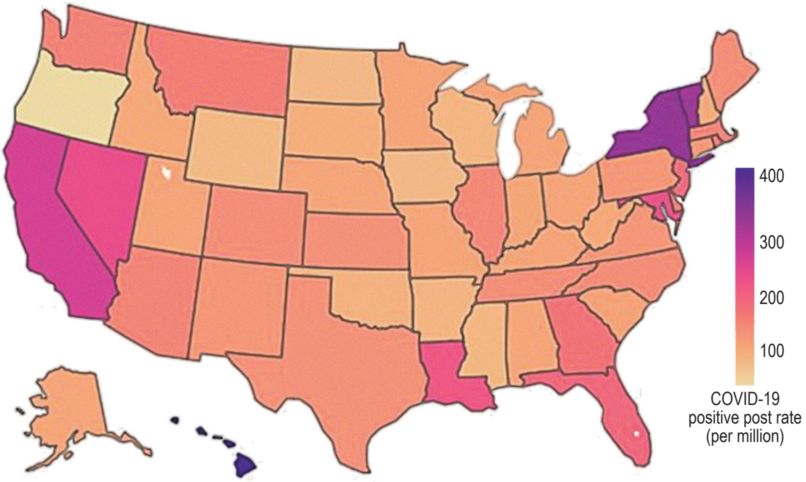


Figure 10.1 Population-adjusted state-level distribution of firsthand reports of positive COVID-19 tests on Twitter/X.

COVID-19 cases and the numbers reported on social media, as discussed above, illustrate the potential of social media-based geolocation-centric pandemic monitoring. With the growing adoption of social media globally, it has the potential to serve as a future platform for geolocation-centric surveillance on a global level. In fact, social media-based surveillance has the potential to reach populations that are hard to reach via traditional surveillance mechanisms—in close to real time and at low cost.

Substance Use

Social media platforms have emerged as potential sources of knowledge for studying topics about which information is either not available or scarce to obtain from traditional sources. One such topic is substance use and substance use disorder. Substance use and its impact constitute a major public health problem globally, and in some countries, like the United States, it is currently considered to be a national crisis. In 2020, over 90,000 Americans died from drug overdoses according to the National Center for Health Statistics (2021), and more than 100,000 overdose deaths occurred in the 12 months leading up to December 2021, an average of over 270 deaths per day. Whereas nonmedical use of prescription medications has historically contributed significantly to the drug overdose epidemic, recent years have seen notable increases in the use of synthetic opioids and psychostimulants. The current epidemic of substance use-related deaths and substance use disorder, including opioid use disorder, is the continuation of decades of constantly evolving trends (Jalal et al. 2018). Within the United States, inequitable access to treatment and enforcement of drug use laws have led to racial disparities in substance use, addiction, treatments, and outcomes (Sanmartin et al. 2020). Over recent years, many studies have highlighted disparities in the treatment of people who use substances that can be traced to socioeconomic status, race/ethnicity, gender identity, community, criminal history, and health-care coverage (Burlew et al. 2021; Lagisetty et al. 2019). The COVID-19 pandemic exacerbated the substance use epidemic, disproportionally affecting communities of color and minority populations (Volkow and Blanco 2021). It has been realized that the implementation of public health approaches to fight the substance use crisis across the globe needs to be multifaceted, focusing on evidence-based programs (Becker et al. 2021), actively addressing barriers to treatment, such as treatment access and stigma (Volkow 2020), and improved surveillance of emerging substance use trends (Kolodny and Frieden 2017; Strickland et al. 2019). Surveillance must be timely to detect emerging "waves" of the epidemic, which is currently believed to be in the early phases of a "fourth wave" in the United States, involving polysubstance use, illicit fentanyl analogs and stimulants (Ciccarone 2021), and responses to these evolving trends need to be tailored to the underlying needs of the affected populations.

A necessary aspect of curbing the epidemic of substance use is to obtain insights about its trends in a timely manner so that responses can be executed accordingly. Traditional surveillance systems consist of surveys (e.g., those conducted by the National Survey on Drug Use and Health, NSDUH), poison control centers, hospital data about treatment admissions and discharge, overdose-related emergency department visits, overdose death records, and others. Such traditional surveillance systems have considerable lags associated with them (Flores and Young 2021). For mortality data, for example, there is a lag time of 12 to 18 months (Anwar et al. 2020). Due to these major delays, emerging trends can only be detected and understood retrospectively. Here, social media can potentially provide an excellent source of real-time information. Indeed, the utility of social media for conducting substance use surveillance (toxicovigilance) has been realized in recent years, resulting in a fast increase in the number of studies exploring social media for substance use-related topics. Social media sources hold substantial promise for toxicovigilance research and signals comparable to NSDUH surveys and NEDS[6] can be discovered from social media via automatic characterizing and mapping of data (Chary et al. 2017; Sarker et al. 2019). Social media are also well-suited for studying aggregated behaviors from targeted cohorts since the social media user base is fairly diverse. Social media data can potentially be used to understand

[6] NEDS (Nationwide Emergency Department Sample) is part of a family of databases and software tools developed for the Healthcare Cost and Utilization Project.

substance use and substance use disorder among subpopulations such as those who have rising rates of overdose deaths and worse treatment outcomes (e.g., Black and Hispanic populations), lower chances of seeking treatment (e.g., women), or are often excluded (e.g., uninsured).

Social Media-Based Monitoring Strategies

Strategies for conducting monitoring of social media chatter effectively for substance use are similar, in principle, to those for COVID-19 described above. Unlike COVID-19, however, substance use-related chatter is sparse, so the data collection process requires additional innovations. In the past, studies have used automatic, data-centric methods for generating street names and misspellings of substances for data collection (Sarker and Gonzalez-Hernandez 2018). Following data collection, supervised ML needs to be applied to filter out most of the posts that mention substances but are not reports of personal use. Once self-reported substance use posts are identified, they can be mapped to geolocations to obtain an understanding of how substance use is distributed spatially at specific time periods. Figure 10.2 shows the distribution of county-level substance use-related chatter in the United States in 2019, estimated purely via automatic characterization of Twitter/X data.

Research that attempted to establish social media as a potential source for geolocation-centric monitoring had to first validate whether signals detected from these resources were meaningful. Since it is not possible to ascertain if individual social media subscribers at specific geolocations are self-reporting accurate information, this validation focused on comparing aggregated social media data on substance use with established traditional sources. In the case of substance use, these established sources include, for example, overdose deaths from the CDC WONDER database (Spencer et al. 2022) and national surveys such as the NSDUH (SAMSHA 2017, 2020). A study conducted using this strategy of geolocation-centric analysis showed that for the state of Pennsylvania, estimates derived from Twitter/X about opioid use were correlated with opioid overdose-related deaths (Spearman $r=0.331$, $P=.004$) at the county level (Sarker et al. 2019). At the substate level, tweet-level estimates were also found to be correlated with prescription opioid use (Spearman $r = 0.346$), illicit drug use (Spearman $r=0.341$), illicit drug dependence (Spearman $r=0.495$), and illicit drug dependence or abuse (Spearman $r=0.401$). This study demonstrated the utility of analyzing geolocation-specific patterns of Twitter/X chatter on substance use, as it can be applied to understand behavior at a large scale accurately and in close to real time. Social media-based monitoring thus offers the possibility of detecting patterns faster than any other traditional form of surveillance.

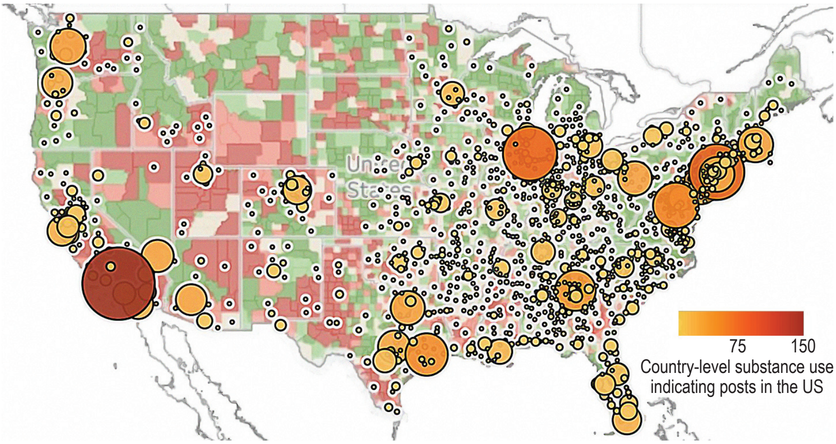


Figure 10.2 Self-reported substance use rates in the United States at the county level on Twitter/X.

Ethical Considerations of Utilizing Social Media Data

The rapid growth of social media and its utility in digital ethology raises important questions regarding the ethical considerations that need to be made when using such data (see also Medeiros et al., this volume). Because of the evolving nature of this research area, there are currently no standardized and universally accepted guidelines for the usage of social media data in health-related or other tasks. The protocols for data use are primarily driven by the organizations behind the social networks and their end-user agreements. For research within the broader medical domain, the protocols for the inclusion of social media data in research are largely guided and approved by institutional review boards. By and large, these boards attempt to ensure that the inclusion of data from social media does not pose any additional risks to the people whose data are being used. Generally speaking, the use of data is considered to be acceptable as long as the data are publicly available. Over recent years, researchers in this space have made efforts to reach consensus regarding the acceptable use of data. Many research groups have also outlined efforts to promote safe use of the data that go beyond what is required by the data use agreements specified by the social networking companies. These efforts include, for example, the removal of user data from studies if subscribers themselves delete their data or make their data private.

Due to the evolving nature of the data and research in this sector, ensuring standards for the ethical use of such data for digital ethology is a moving target. This will perhaps continue to be the case in the near future, much like the field of artificial intelligence itself. This fact is being increasingly recognized by researchers in this space, and efforts are in place to reach consensus collaboratively and/or raise awareness about concerns.

Conclusions

The evolution of social media, its large-scale adoption, and the rapid advances in data science have opened up unprecedented opportunities for digital ethology. Here, I have focused specifically on the utility of geolocation-centric social media chatter analysis for public health tasks and have outlined two case studies. As the digital footprint of human civilization on social media continues to grow, it is reasonable to expect new opportunities and challenges will arise in the future. The currently known limitations of this data source will also likely evolve over time. The evolving nature of research in this domain means that ethical guidelines will evolve. Consequently, it is imperative that experts from different domains and stakeholders with diverse intentions collaborate to establish protocols that will ensure the responsible use of such data, leveraging it for the common good and minimizing potential harm.

Acknowledgment

The research on social media-based substance use surveillance was supported in part by the National Institute on Drug Abuse (NIDA) of the National Institutes of Health (NIH) in the United States (DA057599, DA046619).

Context and Health

11

Integrating Knowledge from Individual- and Aggregate-Level Data

Sven Sandin

Abstract

Modern technologies and societal changes have generated vast amounts of data, personal and individual or aggregated in clusters or geographic regions. Even though this development has stimulated a wealth of research aimed at understanding disease etiologies and promoting lifestyle changes, opportunities remain, and the integration of data is underutilized.

This chapter describes how geographic and aggregate-level data, with information about environmental and social exposures, can be combined with individual-level health data to increase our understanding of disease etiologies. With an emphasis on data primarily available in Nordic countries, it provides a summary of data sources, references for further reading, approaches and methods for analyses, legal aspects, and limitations.

Compared with data at the individual level, analysis of data at the aggregate level has many advantages in terms of access and privacy. Nonetheless, because the availability of individual-level data is the main strength of data from the Nordic countries, the summary starts with a description of these data and ends with aggregate and geographical (area-level) data. Note that in the Nordic countries, all register-based individual-level data can be linked to geographic regions (e.g., hospital, city, county) associated, for example, with place of birth or current residence. The information provided here should be helpful for anyone interested in disease-specific research and public health work to understand better underlying risks and causal paths.

Introduction

In 1943, the national Danish cancer register was created as a national research register, and in the 1950s, the other Nordic countries followed suit; reporting of malignant cancers became mandatory by law. National population registers and registers for vital statistics combined with unique personal numbers opened the door for population-based epidemiology (Pukkala et al. 2018). In the wake of cancer research and cancer epidemiology, a multitude of different registers and data sources have since been developed and become available for research purposes in the Nordic countries (Laugesen et al. 2021). Today, the Nordic countries (Denmark, Finland, Iceland, Norway, and Sweden) comprise a total population of approximately 27 million. The countries provide unique opportunities for joint health register-based research in large populations with long and complete follow-up, facilitated by shared features such as tax-funded public health-care systems, similar population-based registers, and the personal identity number as a unique identifier of all citizens (Laugesen et al. 2021). Notwithstanding these similarities, joint Nordic data resources remain underutilized in health research, and it should be possible to combine a wider array of data sources and apply modern methods to address research questions with better precision and accuracy. Examples of such data sources include weather data (temperature, rain, humidity, sun hours), pollution and air quality, road traffic and population density in different geographic regions as well as socially informative data (education, income, occupation, work). Furthermore, multigeneration and twin registers can provide information about inherited risks, opening the door for statistical analyses strengthening causal interpretation of results. In all Nordic countries, repositories of official statistics act as hubs linking different data sources through the personal identification number, which is in turn linked to tax records that provide information about geographic location.

The national population data sources available in the Nordic registers are not universally available to any citizen. For behavioral and lifestyle data or phenotype information not provided by national registers, cohort studies can be linked.

Whereas national registers and population-based cohorts are unique in their ability to generate unbiased estimates thanks to the complete (or almost complete) subject selection, other data sources offer methodological challenges, such as case-control studies, case cohorts, and self-selected samples. Consequently, the landscape of data sources has grown exponentially and includes a large variety of different designs, as well as data collected with no a priori design, or a lack of design. And whereas in the past national registers, cohort studies, and special case-control studies have provided undisputed information and knowledge useful for development of health measures, an efficient mapping and utilization of new data sources is required to keep up the pace of discovery. The development of statistical and computational methods, such as artificial intelligence, machine learning, and modern computer processing capabilities, provide useful tools.

With the goal of facilitating the creation of data informative for human health, the purpose of this chapter is to provide an overview of the different data sources available (see also Appendix 11.1), to demonstrate how to find, combine, and share the data, and to identify analytical challenges associated with their use.

Nordic National Registers

In each Nordic country, every citizen has a unique national identification number provided at birth or at immigration. The authorities use this number in all correspondence or registrations to ensure that citizens can be uniquely identified. Tax offices in the Nordic countries keep records on date of birth, emigration, or immigration. In addition, each country has a range of nationwide registers on health-related and other topics relevant for the authorities to monitor. Some of these were established decades ago, whereas others are more recent. As detailed below, all Nordic countries administer medical birth registers where information related to all births, preceding pregnancies, and maternal and perinatal conditions are recorded; patient registers record diagnoses by clinical specialists and vital statistics registers provide information about date of birth, death, immigration, and emigration.

Reporting to many registers (e.g., patient register) is mandatory by law and with few exceptions does not require consent (e.g., smoking during pregnancy in the Medical Birth Registry of Norway). Outside of the national cancer registers, the main purpose is not research but administration, monitoring, and quality assurance. Since personal ID numbers are used in all registrations, information from one register can be linked to information from others. This is permitted for research purposes under special circumstances (see below for description for each country). There are also requirements as to how data can be stored, used, and shared. When those circumstances are met, the researchers can apply for data from the register-keeping authorities. Health registers, for instance, are usually administered by different institutions than registers containing social information. When applications are approved, researchers receive data files containing copies of data they requested. These data usually require a lot of reorganization and cleaning before they can be used for statistical analyses. In addition, when combining data from two or more countries, extensive harmonization work is needed before analyses can be conducted in a similar (or as similar as possible) manner.

Medical Birth Registers

All Nordic countries have nationwide birth registers with complete coverage of live and stillbirths (Table 11.1). This register contains information on infant and maternal characteristics as well as on the pregnancy and delivery. The Swedish and Norwegian registers also collect information on fertility treatments, their indications, and procedures. The midwife or physician overseeing the delivery collects the following data at the hospital or home in the case of

Table 11.1 Overview of Nordic registers, showing the starting year that social and health-care data began to be collected in Finland, Denmark, Norway, and Sweden.

Type of data	Finland	Denmark	Norway	Sweden
Unique personal identifier of all residents	1968	1968	1967	1961
Medical birth register	1987	1973	1967	1973
Cause of death	1971	1970	1951	1961
Inpatient specialist care diagnoses	1969	1977	2008	1987/1973[1]
Outpatient specialist care diagnoses	1998	1995	2008	2001
Primary care diagnoses	2011	—	2006	—
Detailed neonatal specialty care	2005[2]	—	2009[3]	2001[3]
Cancer	1953	1943	1953	1958
Prescribed medicine/drugs	1964	1995	2004	2005
Medical pension and sickness leave (date, diagnosis)	1962/1999[4]	1976	1992	1990
Unemployment and social welfare	1970	1976	1992	1990
Taxable income	1970	1970	1993	1990
Educational attainment	1970	1973	1974	1970s
Occupation	1970	1981	—	1960s
Military draft cognition tests[5]	1982	1957	1970	1951–2010[6]

[1] Nationally, all psychiatric diagnoses from 1973 and somatic diseases from 1987
[2] Birth weight under 1500 gram or born before 32 weeks of gestation
[3] All children admitted to neonatal care
[4] Sickness leave from 1994
[5] Finnish data include personality; Finnish/Norwegian data include physical fitness
[6] Also from 2017 but with very low number summoned and tested

planned home deliveries: maternal height and weight, smoking status, parity and complications during pregnancy or delivery, infant gestational age, weight, length, head circumference, live/dead-born, and malformations and complications at birth.

Patient Registers

The national patient registers (NPR) are similar in the Nordic countries (Table 11.1). Since each Nordic country has a publicly financed health system with equal access, this ensures complete coverage of the population. NPRs include information about a patient's geographic location; the hospital, department, and clinical specialty needed; admission and discharge date; whether the visit was acute, planned, in- or outpatient; the type of diagnosis according to the International Classification of Diseases (ICD) diagnosis as well as surgical and medical procedure codes. Currently, ICD-10 diagnostic codes are used.

NPRs have evolved over the years. In Sweden, for example, its NPR was founded in 1964 but national coverage began only 1987 (except for psychiatric diagnoses when national coverage began in 1973 for inpatient specialist care). From the beginning, only inpatient visits with diagnoses from specialist care were included; diagnoses from outpatient specialist care were added sequentially, county-by-county, between 1999 and 2005. Extensive validation efforts had been made for different diseases with good results. Coverage and reliability vary, however, depending on the type of condition. Acute conditions requiring inpatient care (e.g., myocardial infarction) have better coverage than conditions such as obesity, type 2 diabetes, or subclinical depression and mood disorders which are typically treated by general practitioners.

Drug Prescription Registers

All Nordic countries have nationwide prescription registers that contain information about prescribed and collected drugs coded using the Anatomical Therapeutic Chemical (ATC) classification system. ATC has five levels: The first level indicates anatomical main group and contains 14 codes (e.g., N = nervous system and C = cardiovascular). The second level indicates the therapeutic subgroup. Levels three to five indicate finer details that describe chemical and pharmacological subgroups. The last level contains 5,067 codes. Even though there is information about drug dosage, the dose information is entered as free text and is therefore difficult to use. Limitations include the lack of information about drugs dispensed in hospitals and over-the-counter drugs. One practical limitation is the lack of data on why the drug was dispensed, which may provide information that helps to avoid biases due to confounding by indication (Catalog of Bias 2018; Greenland and Neutra 1980).

Sweden's Multigeneration Register

The multigeneration register (Ekbom 2011) is a register administered by Statistics Sweden (SCB) and is comprised of persons who have been registered in Sweden after 1961 as well as those born in 1932 or later. These people are referred to as index persons. The register contains connections between index persons and their biological parents. In 2016, there were about ten million index persons in the register. Information is also collected for certain index persons from older national registration material. For index persons who were adopted, there is also information on their adoptive parents. Currently, there are about 150,000 index persons with information on adoptive mother or adoptive father. Thus, pedigree information on a child, mother, father, maternal, and paternal grandparents is available, and information about siblings (full, maternal and paternal half siblings), cousins (of different types), and aunts and uncles can be derived. This information on pedigrees has allowed family studies separating inherited risk from the environment without the need for genetic data. It has also the additional strength of capturing the entire inherited genetic information, whereas genome-wide association studies (GWAS) capture only a fraction (Bai et al. 2019, 2020). Another important use of this data source is analyses adjusting for family confounding; that is, factors related to the family cluster (including genes) and not the individual per se (also unobserved factors). For example, one study estimated the relative risk for individuals in the lowest Swedish income quintile of being convicted of violent criminality, compared with the highest quintile, to be a sevenfold increased risk. When adjusted for (unmeasured) family risk factors, the risk difference disappeared (Sariaslan et al. 2014). In another study, offspring exposed to higher levels of smoking during pregnancy had greater rates of severe mental illness rates than did unexposed offspring. This study failed, however, to find support for a causal effect of smoking when adjusting for (unobserved) family risk factors (Quinn et al. 2017).

In the other Nordic countries, the mother–child information from medical birth registers and information about the father can be used to derive similar information (Bai et al. 2019).

Cause of Death Register

All Nordic countries have cause of death registers, which include information about date and place of death, cause of death, and whether the death was natural, an accident, or suicide (Brooke et al. 2017; Helweg-Larsen 2011; Norwegian Institute of Public Health 2022; Statistics Finland 2021; Tolonen et al. 2007). All registers were founded before 1970.

Registers Informative for Social Exposures

All Nordic countries administer national registers for education, work and unemployment, occupation, income and taxation, housing, and other social factors. One register example is LISA (Longitudinal Integrated Database for Health Insurance and Labour Market Studies) in Sweden, with similar databases available in the other countries.

Created by SCB (Ludvigsson et al. 2019), LISA integrates existing data from the labor, education, and social sectors with the goal of enabling analysis and evaluation in the field of health/illness. LISA currently comprises 28 vintages and covers the period from 1990 to 2017. The database is expanded with a new vintage every year, with a delay of about 15 months, and is longitudinal: data for the same person can be linked for all years the person is in the population. Between 1965 and 1990, an extensive survey was sent out every five years to all inhabitants of Sweden, and this information is also linked to LISA. This detailed questionnaire, completed by all citizens, provided information about work and type of occupation as well as information on the conditions and standards of living. LISA includes data on yearly income and taxation, the highest level of education attained, occupation, number of days unemployed, income due to unemployment, early retirement, marital status, disposable income, number of children of different ages in a household, and the European Socioeconomic index created from the International Standard Classification of Occupations (ISCO).

Country-Specific Procedures for Data Access

Norway

In Norway, the use of register data for medical research is regulated by the General Data Protection Regulation (GDPR), the Health Research Act, the Health Registry Act, and the Statistics Act. In addition, most health registers have their own specific regulations.

In general, the use of health-related information for research purposes requires informed consent from the participant, yet information reported to the national health registries is confidential and reported without consent requirements. Therefore, the use of individual-level health-related information for research requires the approval and exemption from confidentiality from a Regional Committee for Medical and Health Research Ethics. Application to the ethics committee must include a project description that specifies the project aims and justifies the need for new knowledge, along with details on the planned data linkages and reasons why this information is needed to conduct the project. The application must also describe who will have access to data and how data will be stored. After acceptance, if someone not mentioned in the original application needs to have access to data, an amendment must be submitted.

Anonymized data (i.e., data which cannot be traced back to an individual living person) from the health registers (even linked between registers) can be used freely without applying for ethical approval. In such cases, the registry-keeping authorities are responsible for ensuring that the data provided to the researcher are "truly anonymized" (i.e., the data are indeed impossible to trace back to an individual) as judged by the responsible Norwegian authorities.

Statistics Norway administers data on education, income, social, and work-related information. The Statistics Act forbids any individual-level data from Statistics Norway to be stored in countries other than Norway. This severely constrains the use of Norwegian data in international research.

In practice, analyses involving such data must be carried out in Norway, and only the results can be shared. Researchers at an approved research institution or body within the EU/EEA may, as an exception, be granted access to indirectly identifiable data (pseudo anonymized) from the health registers. In its assessment, Statistics Norway places importance on measures to address the increased risk of data processed outside Norway's jurisdiction. In such cases, requirements are generally set for a specially adapted agreement with the foreign research institution/authority to ensure that Norwegian rules of law are applied and that a Norwegian legal venue is established (Statistics Norway 2022).

In practice, data may be shared in a common repository if there is no possibility to extract raw data on individuals and there is strict control of access to data. This “human restriction” security level includes contracts with register-keeping authorities and usually involves very few analysts (ideally only one for each study). This person is known and selected by the data processor who also ensures the competency level for the data processing. Together with the technical solution (SSH tunnel and time-limited certificates), this guarantees data protection.

Finland

In Finland, register data can be used for secondary purposes, including medical research, according to the Act on the Secondary Use of Health and Social Data (552/2019), the Personal Data Act, and the Act on the Openness of Government Activities. Other associated laws include the Statistics Act, the Act on National Personal Records Kept under the Health Care System, and the Medical Research Act. The Data Protection Ombudsman guides and controls the processing of personal data and provides related consultation.

The general principle regarding medical research is that whenever possible, non-individual-level data is preferred by the authorities (as stated in the Personal Data Act). If individual-level information is needed for research, informed consent is requested from the participants whenever possible. If getting consent is not possible, for example, due to a high number of individuals in the dataset (as is often the case in register studies) or because historical data is needed, a permission for research can be requested from the Health and Social Data Permit Authority (Findata) or, in some cases, directly from the authority keeping the register. Consent is always needed if register data are linked, for example, with survey data. If there is a need to combine data from the registers of multiple owners or obtain data from private social welfare and health-care service providers, the permits are issued by the Findata authority. If data are needed from a single register owner, the authority that oversees that register takes final responsibility for all research use of their data.

In principle, if a study uses only register-based information, an approval of an ethics committee is not required by law. In practice, however, research institutions where the study is conducted can require ethics committee approval for all studies conducted by that institution. Medical studies using register data usually apply for a statement from the regional ethics committee in the hospital district. In Finland, as in Norway, application to the ethics committee must include a project description/research plan specifying its aims and detailing planned data linkages, as well as an explanation as to why this information is needed to carry out the project.

As with the application to the ethics committee, the application for a data permission must include a data utilization plan, a list of individuals who will process the data, and a description of the requested data. If someone not mentioned in the original application needs to have access to data, an amendment must be submitted. Data from health registers can be shared with research collaborators in other countries if data security is sufficiently high. This applies primarily to collaborators in Europe (EU and EEA countries). Data sharing outside Europe is much more strictly regulated.

In most cases, remote access to pseudonymized data is granted. Identifiable data can be delivered to researchers in some restricted cases, if data security is sufficiently high; for instance, if the researcher already has the identification numbers (e.g., own cohort) or if the researcher will link additional data to the dataset (e.g., medical records from the hospitals). Permission and processing of the register data for research purposes is liable to charges.

Sweden

In Sweden, research using the Swedish registers requires affiliation with a university and approval from the Swedish Ethical Review Authority (2023). The registers are primarily administered by three government bodies: SCB, The National Board of Health and Welfare (*Socialstyrelsen*, or SOS), and the Swedish tax agency. As of 1947, all Swedish citizens are assigned a unique personal identification number at birth, which makes it technically possible to link all governmental registers. In research, the personal ID number is always replaced by a random identifier by the register holder for privacy reasons. To request data for research purposes from a national register, an ethics permission is needed. Approval is not, however, sufficient to enable access to the register data; each authority alone decides on what information can be provided to the applicant. After approval from the national ethics board, a lawyer at each register reviews and approves the use of data through a process that does not need to take research aims into consideration. Their goal is solely to protect the privacy of individual Swedish citizens, based on regulations to which the respective authorities are subject. When ordering data for research purposes, a main responsible person is usually assigned at either Statistics Sweden or the National Board of Health and Welfare to coordinate the activities linking the different registers and selecting the appropriate records. This work will be charged to the researcher ordering the data.

The National Board of Health and Welfare (SOS) is a government agency under the Ministry of Health and Social Affairs. The primary register for medical research is the NPR, which contains records of all visits to a clinical specialist; nationwide inpatient care since 1987 (1973 for psychiatric diagnoses). Outpatient specialist diagnoses are available in the patient register between 1999 and 2005 for different counties. All diagnoses are recorded using ICD 7, 8, 9, and 10. SOS is also responsible for the cause of death and the cancer registers. It is not the policy of SOS to provide individual-level data to researchers outside Sweden and the EU/EES. Instead, they advise researchers from other countries to cooperate with colleagues affiliated with a Swedish university, to whom SOS can provide data according to standard legal provisions and procedures. Over the last few years, the Swedish government has invested in health registers, which has resulted in the creation of the National Quality Registries. The National Quality Registries have been built up by dedicated health-care professionals with the aim of monitoring the outcome of specific health conditions (e.g., breast cancer, psychiatry, heart disease). The objective has been to generate valuable knowledge to improve health care and support research.

SCB is the Swedish government agency responsible for producing official statistics in Sweden. It is the holder for registers of vital statistics (date of birth, death, immigration and emigration), for education, as well as social measures. SCB collects, supports, and coordinates official statistics. It produces statistics from many subject areas with different kinds of geographic divisions, such as county, municipality, partial areas, and postal code areas. The products are developed by Statistics Sweden as commissioned work. In Sweden, data for individual respondents (microdata) are protected by the Secrecy Act. It is, however, possible for researchers to apply for access to microdata for use in specified research projects. The system for researchers' access to microdata stored at Statistics Sweden is called Microdata Online Access (MONA). Data are described through Statistics Sweden's standard system for documentation of microdata. Information about MONA and the documentation is published on the website in Swedish. The SCB longitudinal database LISA contains individual data on sickness, parental, and unemployment insurance.

Denmark

In Denmark, there are two main owners of data from national registers: Statistics Denmark and the Danish Health Data Authority. As public authorities and data processors, both are subject to Danish laws for treatment of personal data, including the Act on Processing of Personal Data and the Danish Act of Health.

Statistics Denmark manages data registers on the total population, including information on various demographic factors and social conditions. To obtain access to data from Statistics Denmark, a research project must be associated with a Danish public research unit. Furthermore, the Danish Data Protection Agency must approve the research project if data are linked to data from other authorities or registers. If data from Statistics Denmark are linked with data from the Danish Health Data Authority, approval from the Danish Health Data Authority is also required. Subsequently, Statistics Denmark extracts data from the registers and places all the data on a server at Statistics Denmark (EIT Health Scandinavia 2022).

The Danish Health Data Authority is the supreme authority of health care in Denmark and is part of the Ministry of Health. The Danish Health Data Authority is responsible for all health registers, including the medical birth register, the cause of death register, and NPR. To access data from the Danish Health Data Authority, the Danish Data Protection Agency must approve the research project; if the research project includes direct contact with humans or human biological material, approval must also be obtained from the National Committee on Health Research Ethics. Over the Scientific Service of the Danish Health Data Authority, researchers can obtain access to these data in a safe IT environment, known as "the Research Machine" (*Forskermaskinen*) (Sundhedsdatastyrelsen 2022). The Research Machine allows remote access to most health registers in a secure environment; it requires a personal user ID and two-factor login and no remote access. The user is allowed to use email to send out results from the Research Machine but may send individual-level data.

Aggregate-Level Data

While individual-level data provide the most precise information on individuals, aggregate-level data offers valuable insight. For instance, different *occupations* are often associated with different environmental exposures (e.g., the exposure of workers in sawmills and lumberyards to wood fiber dust). This information can be exploited after individuals are linked to occupational registers. Although individual exposures may vary depending on the exact job task and length of work, such classification can provide important information (Knight et al. 2010) and relate occupation to health outcomes. It is important, however, to adjust for confounding since occupation is strongly linked to education and other socioeconomic factors which are also generally associated with health.

Urbanicity, another type of information defined on an aggregate level, has often been proposed to influence psychiatric outcome and mental illness (e.g., schizophrenia) and is available from national registers and polls as well as from cohorts. For example, SCB offers information from demographic areas (DeSO), using unique codes to indicate nine positions. The first four positions indicate the county and municipality to which an area belongs, as it consists of the county and municipality code. The fifth position is a letter: A, B, or C, which groups the DeSO into three different categories: A is located primarily outside major population concentrations or urban areas; B is mostly located in a population concentration or urban area, but not in the central city of the municipality; C is located in the central part of the municipality (Figure 11.1). In each area, information about age, education, and living conditions is available.

Geographic variations in disease frequency, or exposure (e.g., air pollution), can be used in the search for underlying risk factors. Geographic variations in

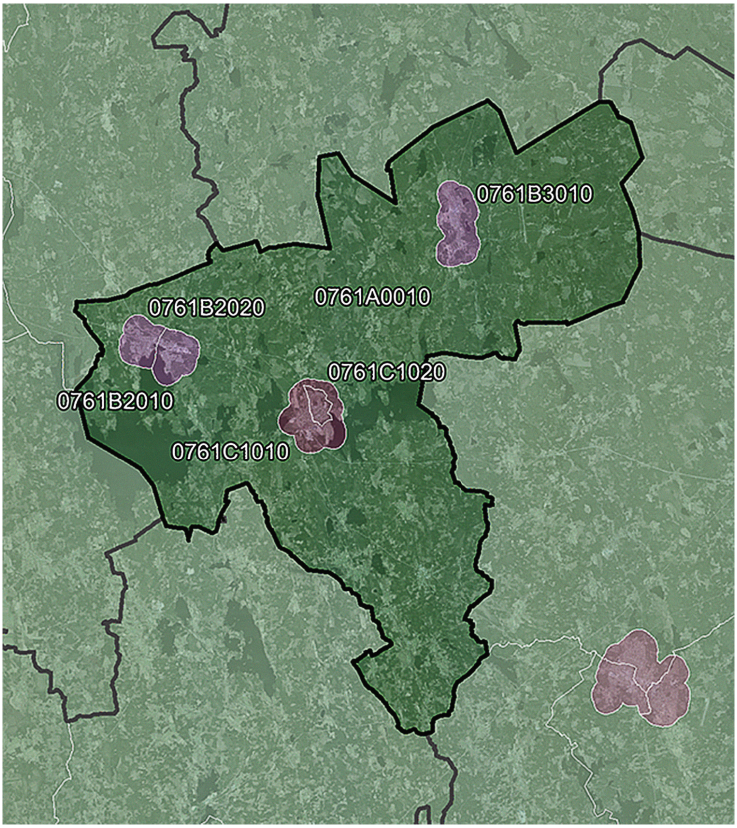


Figure 11.1 Demographic areas coding. Image source: Processing © SCB, other geodata © SCB, Lantmäteriet.

physical environment (e.g., temperature, humidity, wind, and sun exposure) may give insight in health and wellness and may be increasingly important with future changes in the climate (Beauté et al. 2016; Bhopal 1993). Such data is generally available on a geographic regional level from national meteorological institutes. These measures are available across the European Union using the *Nomenclature des Unités Territoriales Statistiques* (NUTS), the hierarchical geographical region classification system (Publications Office of the European Union 2003). The aim was to obtain comparable areas in terms of, for example, surface area and population size in the various EU member states. Introduced in 1988 by EUROSTAT, it is also used by the Nordic countries for area classification, which can be linked to the individual data in the national registers. NUTS can then, in a next step, be linked to national geographic areas such as postal codes (zip codes). Using different units or different definitions of geographic units can result in increased variations in health outcomes; see the study of Legionnaires disease (Beauté et al. 2016).

Combining Heterogeneous Data Sources

For a number of health outcomes, the immense gain in statistical power achieved by pooling research studies has allowed a detailed examination of various relationships, such as attempts to quit smoking, the duration of hormonal replacement therapy after menopause, and combined effects of maternal and paternal age in autism (Collaborative Group on Hormonal Factors in Breast Cancer 1997; Doll et al. 2004; Sandin et al. 2015; Sundström et al. 2019). Naturally, the pooling of studies is constrained by the available exposure data collected in different sites/countries, sometimes several decades after the original study was created. Perhaps more important, pooling studies often requires a reduction of the exposure level to a lowest common denominator. Aligning data measured in many different ways and for different purposes is a challenge and can result in severe oversimplification. When combining several, more or less heterogeneous, data sources, the following issues must be considered.

Selection of Data Sources

Finding relevant data can be a challenge. It is clear that documenting various data collections and data samples in public access databases would facilitate such a task. For population-based studies, the Maelstrom project can serve as a good example (Bergeron et al. 2018). The Maelstrom project administers a database where studies can be registered, the study variables can be mapped to existing variables that facilitate cross-study comparisons, longitudinal measurements are displayed, and the contact information for study principal investigators is accessible.

The lack of generally accepted and utilized variable standard(s) hamper pooling efforts. Thus, harmonization work is repeated for each pooling project, which is a waste of sparse resources. Again, Maelstrom may offer a solution or a start.

To evaluate validity and reliability of collected data, local experts (e.g., clinicians) are usually needed, thus allowing human knowledge to be embedded. For example, a pooling study including longitudinal clinical diagnoses of type 1 diabetes and a second data sample using self-reports can make the overall results impossible to interpret. The clinical diagnoses may change over time as well as the coding system, and changes in the health system may affect ascertainment.

Study Design

Integrating knowledge from different data sources is influenced by the underlying study design. A cohort created as a random sample from a well-defined population has the advantage of allowing many different research questions to be addressed, but other designs may offer different advantages. While cohort studies are often considered easiest to combine, data from different designs should be considered complementary, not competing. Even though much of the criticism of case-control studies is valid, as in biased case selections or lack of a relevant control group, it is not a feature of the design per se. In the Nordic countries, there exists an infrastructure for designing and creating case-control studies with at least the same quality as a prospective cohort study (e.g., where the full population can be enumerated). Strategies for generating new knowledge should be open for inclusion of data from different designs, and data from most designs may be combined using statistical techniques.

Sharing Data

International collaboration as well as data pooling and sharing is key to modern research. Not all collaborators and data sources are positioned within the same legal system. Thus, ways of sharing and combining data must be considered. The most common and, from an analyst's perspective, best way to share data is by *sending the original data*. Encryption in combination with data transfer using secure protocols (e.g., https/TLS) ensure sharing of data with minimal risk of data theft during data transport. Combining all data onto one site optimizes the analytical choices. While it is now difficult to share data between the European Union and the United States (Hallinan et al. 2021), it is possible to share data within each region. In the European Union, data is shared by applying standard agreements for data transfer agreements.

When this is not always possible, more advanced and restricted ways of data sharing must be considered. If the safety concern is related to sharing of individual-level personal and sensitive data, sharing aggregated data may offer an alternative. This, however, comes with restrictions on the analytical tools that are available to analyze the data and will therefore not always fit. A simple example of *aggregated data* is the study of mortality between males and females. For a country of Germany's size, a table of 80 million rows and two columns would be needed yet to calculate the difference in proportion, a 2×2 table containing the number of rows where males and females die and survive will suffice. These information lossless measures are called *sufficient statistics* (e.g., for estimating the difference in proportions of dead males and females). Only statistical analyses where sufficient statistics can be derived from aggregate-level data can be performed without losing any information (Hallinan et al. 2021; Persson et al. 2020; Sandin et al. 2006). When the aggregated data is too crude (too high loss of information) for the intended analysis to be executed, simulation approaches may be used. Applying statistical simulation methods made possible by the power of modern computers allows us to "simulate" or *generate a synthetic database* with the same numeric properties as the original data, but where all links to original (individual-level) data have been eliminated (Nowok et al. 2016). Once the synthetic database has been shared and analyzed, the computer code can be sent back to the original data owner and applied to the original (real) data.

For data sharing in larger collaborations across several sites, data federation techniques offer a viable solution to this problem by permitting controlled access to datasets located and managed in disparate locations without the need for permanent storage at a single location (Haas et al. 2002). Under this scenario, each study site retains control of their own data in separate databases at their respective site (Figure 11.2). The GenomEUTwin project stored epidemiological data for around 600,000 twins from across Europe and Australia (Muilu et al. 2007). In the iCARE project—a collaboration of national registers for autism research between Sweden, Denmark, Norway, Finland, Israel, and West Australia—software was developed to share data as well as to analyze data in a central node using data aggregated at each site (Figure 11.2) (Carter et al. 2016).

Depending on legal requirements, an even more privacy protective approach may be applied, such as by using technologies offered by Datashield (Wilson et al. 2017b; Wolfson et al. 2010). Datashield implements a database federation but in combination with statistical computational techniques similar to the aggregated data (above). Here, only minimal statistics are shared to the

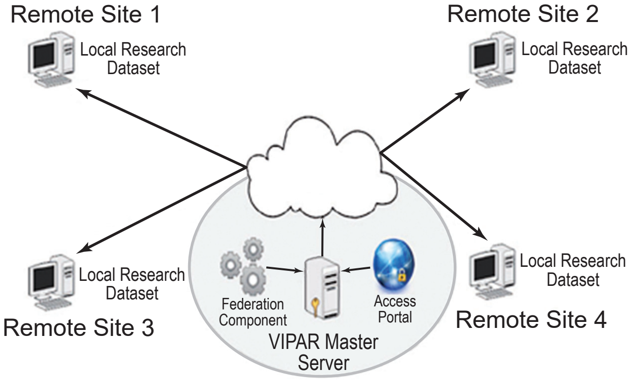


Figure 11.2 Topology of ViPAR (from Carter et al. 2016; https://creativecommons.org/licenses/by/4.0/deed.en). This database application is built around a master server, linked to remote sites. Each site maintains their own data. Analysts access the web-based portal where they run analyses. During analysis, the federation component retrieves data from the sites into computer RAM on the master server where they are analyzed and removed without ever permanently being stored.

central analytical server; individual-level data never leave the local site. As an example, for calculating a linear regression line, only measures (n, $\sum x$, $\sum x^2$) are needed from each site ("sufficient statistics").

Generating Knowledge

Internal and External Validity

While many associations and treatment contrasts can be reliably estimated within single studies, external validity may be less dependable. For instance, even if the relative risk of a health outcome is estimated close to the underlying truth internally in the study, the absolute measures may be biased. This needs to be considered and may be addressed by weighting (Wang et al. 2020).

Confounding needs to be considered as an important topic, both when designing new studies or gathering data from different data sources. Healthy worker effect is such an example. Originally observed in occupational cohort studies, healthy worker effect refers to a situation where people available for and willing to participate in a study tend to be healthier than the target population. This specific form of selection bias usually results in an underestimation of risks, such as for mortality caused by occupational exposures (Naimi et al. 2013).

Replication

One single study, no matter how well designed or implemented, is unable to provide irrefutable evidence regarding the correctness of an association. By using study replication design with independent data samples, the generalizability of results can be addressed as well as the increasing concern of bias and non-reproducibility of results from research studies (Ioannidis 2005; Moonesinghe et al. 2007). This is a current priority of the NIH (National Institute of Dental and Craniofacial Research 2018), which also calls for large population-based studies with contemporary and accurate clinical diagnoses and for studies that can adjust for individual and familial confounding as well as temporal trends.

Knowledge Embedding

Failing to embed properly human knowledge, experience, and empirical knowledge is wasteful. Immediate examples of this include the integration of clinical knowledge about case ascertainment and clinical exposure or known features of health system(s). An analytical example is when applying known genetic correlations in equations (Bai et al. 2019; Svensson et al. 2009) instead of estimating the correlations from the data itself. On the other hand, embedding human knowledge in the wrong way or embedding less solid knowledge could increase both bias and measurement errors.

Challenges

Combining data sources, especially large data with high statistical power, but sometimes limited validity, can lead to false alarms (e.g., warning for spurious association between diet or other environmental exposures and health outcomes). False alarms undermine the credibility of science, move the focus from more important and causally true associations, and increase the anxiety of consumers of the research literature. The reasons for false alarms include badly designed studies, nontransparent (or entirely lacking) analysis plans (often with extensive and ad hoc subgroup analyses), lack of adjustment for multiplicity of statistical tests, and findings uncritically promoted by the investigators. Media attention often worsens the problem when a potentially large proportion of the population may be concerned about a particular exposure. Examples of this include the fear that one might contract brain cancer from using a cell phone use (IARC Working Group on the Evaluation of Carcinigenic Risks to Humans 2013) or that vaccines increase the risk for autism. It took over a decade to ease public concern over media alarms regarding cellular phones (IARC Working Group on the Evaluation of Carcinigenic Risks to Humans 2013). The false claim of autism risk following vaccination has yet to be dismantled in the minds of a large proportion of society and has affected other health outcomes (Madsen et al. 2002). An important yet often neglected consequence of false alarms is that they can undermine efforts to promote healthy lifestyles based on well-established evidence. False alarms increase the risk that the general public will deny all evidence and leave them with a sense that nothing matters.

Given the many rare outcomes and sparse exposures, big data approaches are needed. Geographic disparities as well as temporal trends in disease risk and health markers may indicate the presence of environmental factors. Still, it is a challenge to bring in such innovation, which must be paired with funding, into these and related fields.

Summary and Notes about Future Needs

There are many key issues that need to be addressed in the future:

1. High-quality data do not occur automatically. As researchers, or users of data, we all have a responsibility to *generate new data*. Current research models give too little credit to these issues. After years of planning, generating funding, and collecting quality assurance data, analysts often expect to take first and last place in the list of authors

by arguing they made the scientific contribution. We need a new model to reward the creation and management of new data. New data in a new area should be designed, not as isolated islands, but with the aim and target to combine with other data sources upfront. All new studies want to perform new measures (e.g., new risk scores, new measures of physical activity, rating scales). Each new study, however, should not overlook the importance of reusing existing measures, which would allow the field to connect multiple related studies to facilitate pooling and replication.

2. Study documentation: Projects like Maelstrom should be supported.
3. Legal concern: To solve issues around data access and sharing, more conscious, *brave, and scientifically engaged lawyers are needed as collaborators* to move the field forward. Too often lawyers act in the role of guardians of a company, university, or database. As such, denial of data access is often the first level of defense. For the community, this may result in unethical procedures where accrual and generation of new knowledge is hampered—often contrary to the wish of patients or study participants (Dufva et al. 2021).
4. Privacy and data sharing: In a globalized world where collaborations are key for fast and efficient development, we need to *develop community-based agreements on how to use personal data*. Whereas the European Union has taken one standpoint in strengthening the rights of individual citizens to own and control their personal data, other countries do not agree and have instead adopted laws where governments have the right to all data. This situation seriously hampers collaborations.
5. Methods and competence: There is an urgent need for *advanced statistical methods*, analysts, and software tools to apply these methods to optimize the use of data, targeted for the research question at hand.
6. Data, method, and software: Publications in health research, and other work, should include not only a written description of the analytical approach. In Open Science publications, and for science funded by the NIH, there is often a requirement that data should be made available after publication. While this is a step forward, it is not sufficient. The analytical method should be documented through *publication of the software code* used, and comments on the different steps taken to reach the final conclusions.
7. Replication: We need models and approaches that *encourage replication* and verification of research results. Not only should "new" hypotheses be rewarded; more credit should be given to replication studies. This will allow studies that cannot be replicated to be downplayed and studies which are replicated, but where results cannot be replicated and verified, to be shamed.

Acknowledgments

Part of the country-specific information on accessing register data was taken from work done within the Research on European Children and Adults Born Preterm project (RECAP preterm 2022). The publication by former colleagues at the Karolinska Institutet, Hans-Olov Adami and Nyrén, provided inspiration for future requirements and the need to generate new data (Adami and Nyrén 2016).

Appendix 11.1: Useful Links

Europe

- Infrastructure for spatial information in Europe (INSPIRE): https://inspire-geoportal.ec.europa.eu/
- European Union official statistics (EUROSTAT): https://ec.europa.eu/eurostat
- Data at the World Health Organization: https://www.who.int/data

Sweden

- The National Board of Health and Welfare web page: http://www.socialstyrelsen.se/english
- The Swedish Medical Birth Register: http://www.socialstyrelsen.se/register/halsodataregister/medicinskafodelseregistret/inenglish
- National Patient Register: http://www.socialstyrelsen.se/register/halsodataregister/patientregistret/inenglish
- The Swedish Cancer Registry: http://www.socialstyrelsen.se/register/halsodataregister/cancerregistret/inenglish
- Swedish National Quality Registries, a unique research base: http://kvalitetsregister.se/englishpages/useregistrydatainyourresearch.2251.html
- Ethical aspects of registry-based research in the Nordic countries: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4664438/
- Statistics Sweden (SCB), government of Sweden’s bureau for official statistics: https://www.scb.se
- Regional statistical products: https://www.scb.se/en_/Services/Regional-statistical-products/
- Guidance for researchers and universities: https://www.scb.se/en_/Services/Guidance-for-researchers-and-universities/
- Longitudinal integration database for health insurance and labor market studies (LISA): https://www.scb.se/en_/Services/Guidance-for-researchers-and-universities/SCB-Data/Longitudinal-integration-database-for-health-insurance-and-labour-market-studies-LISA-by-Swedish-acronym/
- MONA – *leveranssystemet för microdata*: https://www.scb.se/sv_/Vara-tjanster/Bestalla-mikrodata/MONA/

Denmark

- Statistics Denmark, government of Denmark's bureau for official statistics: https://www.dst.dk/en
- The Danish Health Data Authority: https://www.sst.dk/da
- Overview of Danish health data: https://www.danishhealthdata.com
- Publications for several Danish registers: https://journals.sagepub.com/toc/sjp/39/7_suppl

Finland

- The Act on the Secondary Use of Health and Social Data: https://stm.fi/en/secondary-use-of-health-and-social-data
- Personal Data Act (unofficial translation): http://www.finlex.fi/fi/laki/kaannokset/1999/en19990523.pdf
- Act on the Openness of Government Activities: http://www.finlex.fi/fi/laki/kaannokset/1999/en19990621
- Statistics Act: http://tilastokeskus.fi/meta/lait/tilastolaki_en.html & http://tilastokeskus.fi/meta/lait/2013_tilastolaki_en.pdf
- Medical Research Act: http://www.finlex.fi/fi/laki/kaannokset/1999/en19990488.pdf
- Data Protection Ombudsman: http://www.tietosuoja.fi/en/index.html
- Findata, Health and Social Data Permit Authority: https://www.findata.fi/en/
- Finnish Information Centre for Register Research: https://rekisteritutkimusen.wordpress.com/
- Institute for Health and Welfare (THL): https://www.thl.fi/en/web/thlfi-en (e.g., Medical Birth Register, Hospital Discharge Register, Care Register for Health Care, Register of Primary Health Care visits)
- Statistics Finland, government of Finland's bureau for official statistics: http://www.stat.fi/index_en.html
- Population Register Centre: http://vrk.fi/en/frontpage (data e.g., on address, nationality, mother tongue, and family relations)
- Social Insurance Institution of Finland: http://www.kela.fi/web/en (data e.g., on reimbursed prescription medication purchases and welfare benefits)
- Finnish Cancer Registry: http://www.cancer.fi/syoparekisteri/en/
- Finnish Centre for Pensions: http://www.etk.fi/en/ (data on all old-age and disability pensions)

Norway

- The Regional Committees for Medical and Health Research Ethics: https://helseforskning.etikkom.no/?_ikbLanguageCode=us
- A translated (unofficial) version of The Health Research Act: https://app.uio.no/ub/ujur/oversatte-lover/data/lov-20080620-044-eng.pdf
- Statistics Norway, government of Norway's bureau for official statistics: http://www.ssb.no/en/
- The Norwegian Institute of Public Health (NIPH), which administers the Medical Birth Registry of Norway, the Norwegian Cause of Death Registry,

the Norwegian Neonatal Network (a quality registry for neonatal medicine) and the Norwegian Prescription Database: https://www.fhi.no/en/

- The Norwegian Patient Registry administered by the Norwegian Directorate of Health: https://helsedirektoratet.no/english/norwegian-patient-registry
- The Cancer Registry of Norway: https://www.kreftregisteret.no/en/

12

Challenges in Data Science in the Use of Large-Scale Population Datasets for Scientific Inquiry

Hye-Chung Kum, Steven Bedrick, and Michele C. Weigle

Abstract

In today's digital world, traces of almost all human activity are logged in various databases, which some have termed the *social genome data*. When appropriate methods are applied to this real-world data, the potential for new insights is endless. The social genome data may transform many fields of science, just as the human genome data has transformed biology. Yet, obtaining, accessing, integrating, cleaning, and using the social genome data to realize its full potential has many computational, statistical, and ethical challenges. The general methodological approach adopted to study human behavior found in the social genome data is *data science*. The application of data science in an iterative spiral process can result in the transformation of data to information to knowledge to action by iterating between inductive and deductive reasoning. Data science applies methods from both computer science and statistics, and also seeks to synthesize them and develop new methods to address the context and needs of a particular disciplinary field. In this paper, the importance of incorporating human judgment and expert domain knowledge into the data science activities at all steps and the numerous design decisions required to obtain valid results and ultimately useful insights is emphasized. Challenges and open questions in applying data science to the emerging field of digital ethology for scientific inquiry follow. In sum, data science teams must have a wide view to see the context, understand ethical considerations of the data, and be able to communicate both the insights and the limitations inherent in the data.

Introduction

Over the last few decades, most of the processes in our society have been digitized, leading to a new digital world where almost all traces of human activities, from birth to death, are captured in various databases. In our previous work, we have referred to the digital footprints left by humans as the *social genome data* (Kum et al. 2014*)*; that is, large-scale datasets of records collected from a large proportion of individuals in a population that report on people's interactions with governments, businesses, and other individuals—collected and linked from many data sources (e.g., the health, education, financial, Census, location, shopping, employment, or social networking records). This encompasses all aspects of human activity including exposure and outcome data. Social genome data are the basis of *population informatics* (Kum et al. 2014), also called population data science (McGrail and Jones 2018), which leverages these large, complex, diverse, integrated individual-level real-world data to address population scale research questions and gain insights by observing human behavior in the digital traces (Kum et al. 2014).

Just as human genome data has transformed, for example, biology in many ways, the potential for new insights when appropriate methods are applied to social genome data is endless and may transform many fields of science. Yet, obtaining, accessing, integrating, cleaning, and using the social genome data to realize its full potential has many computational, statistical, and ethical challenges (Blei and Smyth 2017; Cesare et al. 2018; Haneef et al. 2022). We adopt the view that social genome data are *big data* as characterized by some aspects of the scale, complexity, heterogeneity, and uncertainty of the data sometimes referred to as the five Vs of big data: volume, velocity, variety, veracity, and value. This requires a new way of synthesizing insight from the raw real-world data beyond the traditional methods, regardless of the size of data (Borgman et al. 2015; Ekbia et al. 2015).

In this paper, we first present a brief overview of *data science* as we define the phrase, the general methodological approach we adopt to study human behavior in the social genome data. This includes a description of how data science results in the transformation of data to information to knowledge to action. Then we present challenges and open questions in applying data science to the emerging field of digital ethology.

To ground and motivate our discussion, we introduce a case study involving a hypothetical (but in many ways realistic) analysis into the impacts of a wildfire smoke event on the population of a city. Wildfire smoke is rapidly becoming a significant public health and climate justice issue (Black et al. 2017; Liu et al. 2015; Reid and Maestas 2019). Smoke events affect many aspects of behavior and activity, and, as such, our hypothetical analysts must work with data regarding many aspects of the life and structure of the city including, for instance, data about emergency department (ED) visits, meteorological conditions, and traffic patterns. This includes both data about individuals and also data about the environment—both physical and social—around those individuals, which are all part of the social genome data (see chapters by Smith, Pallante et al., and Sandine, this volume). Analyzing this diverse collection of data to produce actionable policy that could improve residents' well-being will require a data science team that includes expertise in domain science,[1] statistics/math, and computer science/IT (Cao 2017). We will use this scenario to illustrate different aspects of the data science analysis process.

Data, Information, Knowledge, and Action (DIKA)

The main methodological approach that is needed to extract information and knowledge from the social genome data to obtain new insights about human behavior is *data science*. We adopt the view that data science applies methods from both computer science and statistics but also seeks to "blend them, refocus them, and develop new methods to address the context" and needs of a particular disciplinary field (Blei and Smyth 2017). In addition, we emphasize the importance of incorporating human judgments and expert domain knowledge into the data science activities in all steps to obtain valid results and ultimately useful insights. Data science requires sensemaking techniques borrowed from cognitive science (Grolemund and Wickham 2014) that allow the data scientists to apply their work to a larger framework. Further, the methods and techniques required for this may vary by domain, and even by research question. Data scientists need to be able to have a wide view to see the context of the question at hand, understand ethical considerations of the data, and be able to communicate both the insights and the limitations inherent in the data (Blei and Smyth 2017). Data science overlaps in many ways with the field of Knowledge Discovery and Data Mining (KDD), traditionally defined as "the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data" (Fayyad et al. 1996). We postulate that data science as a methodology goes beyond KDD in that it explicitly includes the timely and effective communication of the patterns to the relevant stakeholders to support knowledge, decisions, and actions.

Figure 12.1 depicts our framework for leveraging the digital traces in the social genome data to support evidence-based action. This "data to action hierarchy" is adapted from the standard DIKW (data–information–knowledge–wisdom) pyramid (Ackoff 1989) with an added focus on data science and its application to social genome data. At the foundation (level 1), we find our social genome data library with an appropriate infrastructure for its secure and compliant access. One of the critical steps in data science is to define a research question that will inform the domain but will also be feasible to answer with the data on hand. We need domain scientists who are able to map domain-level questions and inquiries into tractable, or even abstract, tasks that provide

[1] In this scenario, the specific kinds of "domain science" needed will depend on the ultimate analytical and policy goals of the study, but might include, for example, public and environmental health, forestry, botany, meteorology, and urban planning. Furthermore, domain expertise in history, sociology, demography, and political science, with an emphasis on the local community, may be essential.

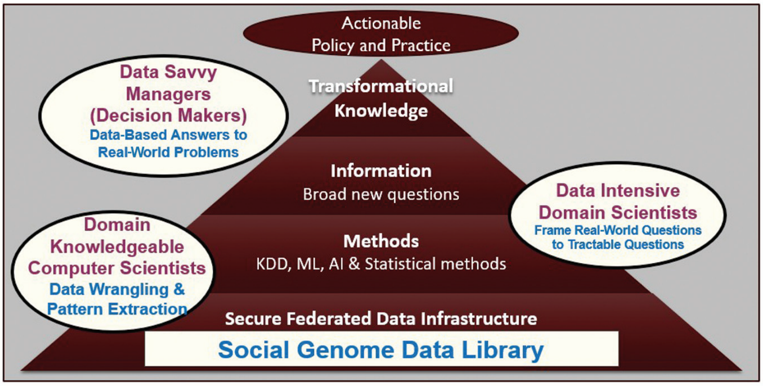


Figure 12.1 Data, information, knowledge, and action (DIKA) pyramid. KDD: Knowledge Discovery and Data Mining; ML: machine learning; AI: artificial intelligence.

a target for the investigation. They will need a good understanding of what data are available in the social genome data library. These questions can then be used to help drive the analysis, using various methods (level 2), including KDD, machine learning (ML), artificial intelligence (AI), and other statistical methods. The expertise of domain knowledgeable computer scientists is essential here to be able to extract relevant data and determine appropriate analysis techniques to effectively address the given questions. The outcomes of these analyses are the answers to the tractable data questions that were posed. Ideally the initial results, this new information (level 3), can be used to generate new questions that can then lead to more insights. Though this is depicted as a pyramid, it is really an iterative process where questions are asked about data, which are analyzed using methods that produce information, leading to new questions, potentially requiring additional sources of data for analysis (and thus, perhaps, new methods). This process continues until the information that is produced results in new knowledge (level 4). This happens when multiple pieces of information can be combined by a human with their domain knowledge, expertise, and experience. The new knowledge in the domain expert can then lead to actions and decisions based on data. Unlike highly automated analysis tasks, this process often requires a team of data and domain scientists with a wide range of expertise iterating through many deliberations, judgments, and analyses. In the following sections, we will expand upon the issues and challenges faced on this path from data to action.

Level 1: Data Infrastructure

The first step for data to support action is to build a compliant data infrastructure (level 1) where social genome data can be ingested, often in the form of a data lake: a "repository storing raw data in their native format," without a pre-defined purpose or specific intended use (Ravat and Zhao 2019). Along with the raw data, some type of metadata about the raw data must be managed so that use can be supported dynamically as the need arises. In addition to the data and metadata, we posit that the underlying software code used throughout the full data pipeline, (involving, e.g., ingest, cleaning, transformation) is also an integral part of the data infrastructure, in order to facilitate good science through replication and reuse (Goodman et al. 2014). Typical users of the data lake are skilled data scientists trained in both computer science and statistics with technical skills in data wrangling and pattern extraction.

By "compliant data infrastructure," we mean the combination of secure computer systems along with the associated policy and procedure layers for data governance that facilitate compliance with legal and ethical obligations (Kum and Ahalt 2013) for use of the data. As data move through the different processes described in the DIKA pyramid, access requirements will change, and an effective infrastructure will have more than one level of access (e.g., restricted, controlled, monitored, and open access). Access controls often relate to granularity of data; for instance, as in a scenario in which some users in some contexts are only able to access data that has been aggregated to a certain degree. Beyond purely technical controls, institutions generally require policy controls when analyzing social genome data, in the form of various types of security, privacy, and human subject research approvals. The details of the data governance and ethical issues are beyond the scope of this paper, but the myriad of laws that apply to the different data sources and purpose of use, and different institutional policies on how to manage the risks involved, is not trivial and is often one of the major barriers to this type of research becoming mainstream. Managing these kinds of data governance complexities is one of the core methodological components of population informatics as a field.

Case Study

Using the example of our wildfire scenario, our data lake will contain several different datasets from a variety of sources and take a variety of forms:

- Admission and discharge data reported by area hospitals and EDs to the county public health authority (structured, de-identified, both individual level as well as aggregated into geospatial units)
- Data about transit and automotive traffic patterns from the city's Department of Transportation (geospatial, time series)
- Demographic information from city, state, and federal records (structured, possibly aggregated to varying levels ranging from county to neighborhood or Census tract, geospatial)

- Geospatial/geographic features of the region (e.g., terrain height map, locations of bodies of water) and readings from air quality monitors (structured, dense time series, geospatial)
- Meteorological data (structured, dense time series, geospatial)
- Emergency management and wildfire reporting data (structured, geospatial)
- Corpora of news articles, social media posts, shared photos produced before, during, and after the event (individual-level, unstructured text and images)

Simply assembling such a dataset represents a substantial technical challenge; each component will bring its own difficulties in terms of collection, storage, maintenance, errors, uncertainty, and documentation. Some may be obtainable from published sources, while generating others may require close collaboration with data providers. The scale and volume of each component data source is likely to be quite different, and each will use fundamentally different file formats, data models, and sampling frames. Furthermore, from a governance standpoint, different parts of this dataset will require different levels of care and oversight when being collected and used. Some of the information is generally publicly available (e.g., air quality readings), while other subsets of the dataset are of a more clearly sensitive nature (e.g., ED admissions), and may come with rules around who is able to access the data and in what ways or may require auditable records of when the data were accessed. Additionally, consider the corpus of social media posts; in this particular scenario, such data are best understood as being public but still sensitive (Martin and Shilton 2016; Nissenbaum 2011; Olteanu et al. 2019; Zimmer 2018) and, as such, must be treated with care (see Weigle et al., this volume). Note that the process of building a data lake, just like that of the KDD process as a whole (Figure 12.2), is typically iterative: new data sources will likely be added as they become available, and, as our scope of analysis changes over time, different sources may suddenly become relevant. It is also important to remember that our different sources of data may play different roles over the course of the project; one kind of data may be considered as an exposure of possible interest in one analysis, and then in another analysis that same data element may be considered as an outcome (dependent) variable, or as a moderator for some other effect. One advantage of the data lake model (as compared to a model relying on a more formally structured data repository) is that it preserves the maximum flexibility in how its constituent datasets may be used.

Levels 2 and 3: Application of Methods for Information

After assembling our data lake, the next step is to define research questions that may be answered using the diverse data available in the social genome data library to extract new information in the field (level 3). In this step, the main task is to frame real-world questions into tractable data questions that can be addressed with the data available. This step is often led by domain scientists who are trained and have more experience in the newer data intensive methods in their field. They often work closely with a strong data science team to determine the most appropriate methods to apply to the raw data to address the question. Methods here are used very broadly to include the full KDD process (Figure 12.2) such as experiential design, measurement definitions, feature and sample selection, as well as modeling and validation.

Once the research question, general methods, and data have been determined, the heavy lifting data science implementation begins (level 2). This is an iterative process that is often referred to as a spiral process, where each iteration will improve on the limitations of the previous spiral until the final results meet the goals of the project. More computationally trained data scientists may adopt the philosophy of agile development (Wells 2009), more often used for software development. This starts from the minimum viable product (MVP), by setting up the data pipeline from beginning to end to check all the basics and test feasibility in the first spiral, then specifies more details in different parts of the data pipeline over the different spirals. This way of implementing the data science project will allow for more reproducible and tractable results, ultimately leading to more valid results. It also easily allows for engaging the domain scientist with different levels of skills at the end of each spiral to do quick checks for staying on track to address the main research question in the domain. These meetings are critical to having results that are relevant to the domain and not getting pulled into the data too far from reality. It is important that the design of the study is well thought out ahead of time since it will be expensive in terms of time and effort to redo things if the setup is wrong. Testing out all aspects using the MVP in the first spiral is one way to check on feasibility before the project gets too far into the weeds.

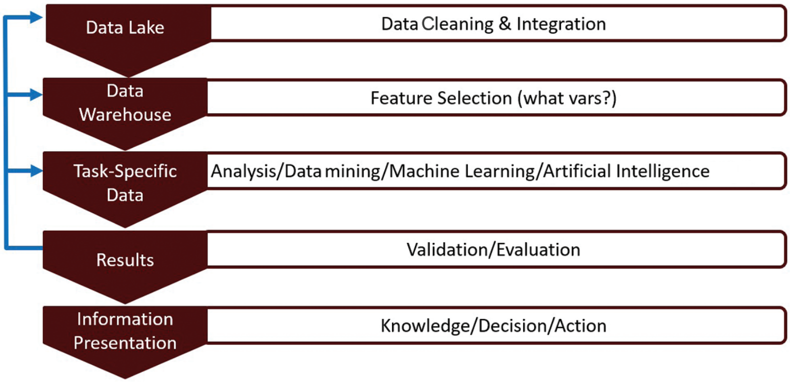

Data Lake
Data Cleaning & Integration
Data Warehouse
Feature Selection (what vars?)
Task-Specific Data
Analysis/Data mining/Machine Learning/Artificial Intelligence
Results
Validation/Evaluation
Information Presentation
Knowledge/Decision/Action

Figure 12.2 Knowledge discovery and data mining process.

Case Study

In the context of our wildfire scenario, imagine an epidemiologist interested in health disparities associated with this particular exposure across racial and ethnic groups as part of a larger effort around determining how best to allocate resources from emergency preparedness funds. Wildfire smoke is known to have heterogeneous impacts across the population (Davies et al. 2018; Liu et al. 2017; Masri et al. 2021), and environmental justice requires that this be taken into account when planning interventions (Brulle and Pellow 2006; D'Evelyn et al. 2022). Our epidemiologist's "big picture" research question might be something along the lines of: Are there differences in how wildfire smoke is affecting the respiratory health of Latino and White residents of the city? There are many possible ways to address this question, depending on how one operationalizes various elements. Which path one takes will depend heavily on what specific data are available. The data scientist, then, will work closely with the epidemiologist to make the question more concrete, and to determine what aspects are feasible (and, just as importantly, what aspects are not). Beginning with the question of how to measure respiratory health impact, we may decide to focus on ED visits with certain groups of diagnosis codes; deciding which codes to include will require a certain amount of domain expertise in working with medical data.

Next, we turn to how to address the question of ethnicity. In our scenario, let us assume that the dataset of ED visits does not turn out to contain reliable information about the ethnicity of patients, which means we must rely on a less indirect statistical approach. We may not have direct ethnicity data, but we do have approximate mailing addresses from the billing records (approximate because they have been blurred/fuzzed as part of a de-identification effort by the original data provider), and, in combination with Census data, it may or may not be possible to use geography as a proxy to get at questions of racial and ethnic disparities; determining this aspect of the analysis will require not only statistical and computational expertise but also domain knowledge in racial disparities, and it may prove necessary to obtain additional or different sources of data.

Finally, to quantify the amount of exposure to smoke, the data scientist will work with the epidemiologist to review available data from the air quality monitoring network in the city; they may also need to involve additional domain experts, for example, specialists in environmental monitoring and sensing, or people with location-specific knowledge about the city's air quality monitoring infrastructure. Together, they will make determinations about (a) the adequacy of coverage and quality, (b) modeling considerations around granularity (in terms of temporal and spatial resolution), and (c) possible issues integrating data from multiple sensor networks. At each of these steps, the original research question will be refined, and new questions may be generated.

Levels 4 and 5: Knowledge and Action

After valid results are obtained to the tractable data questions, the fourth step is to translate the data answer back to the real-world answer to the original real-world question. It will be important at this phase to be transparent, describing exactly what population was used, how features were defined, what, if any, algorithmic black boxes were used, and the limitations of the study including the degree of generalizability of the results. The devil is in the details in any data intensive study, and the details matter in how to interpret the results in the appropriate context. Research involving social genome data typically involves numerous datasets from a variety of sources, meaning that these details have a way of multiplying in their complexity and subtlety. If the error and uncertainty is not well managed by data science experts, then the results will be meaningless.

Another common task at this step is to design and conduct sensitivity analysis that can more clearly delineate the scope of the information obtained. When the full details of the study are effectively presented to data savvy decision makers, we posit that they will synthesize the data details and results into transformational knowledge that can support evidence-based decisions and actions. We believe that information becomes knowledge in a person once the information is understood well enough to apply to decision-making processes and actions. These data savvy decision makers in the domain are the third type of data scientists that have expertise in the domain as well as an intuition for what data can and cannot do, and good judgment on how best to use evidence from data. Many of them are not trained at the PhD level and are key to having real-world impact from the new information and knowledge obtained from data-intensive scientific inquiry.

Case Study

Recall that the underlying motivation behind our analysis of wildfire smoke impact was to help inform decision making about how to allocate emergency preparedness funds, with the goal of maximizing their impact on the community's health. Suppose that our analysts have now computed per-neighborhood estimates of air quality impact, and by linking them with Census data have found what appear to be disparities across ethnicities in terms of that impact. Intriguingly, however, they have also noticed some "outlier" neighborhoods (i.e., neighborhoods more heavily impacted by wildfire smoke than the model would have predicted based on their demographic and geographic properties). In looking more closely at our data, we have spotted that these outlier neighborhoods are ones that have a larger-than-baseline number of living facilities for the elderly, and that the bulk of the larger-than-baseline number of ED visits from those neighborhoods are indeed from older members of the community.

We have now produced *knowledge* and must translate it into *action*. This becomes an entirely different matter, requiring a different set of skills. Earlier in the analysis, our research questions and analytical plan were shaped by the data that were available to us. Now, we must be shaped by two variables that lay somewhat outside the realm of what is usually thought of as "data science": our community's values and the space of possible actions that may be possible.

In terms of our community's values, recall that our goal is to "maximize impact" on the community's health. It is of course crucial to ask what form this is to take; answering this question must necessarily include some consideration of our community's values. Are we primarily concerned with equality? If so, we may set our goals as being to provide a (possibly smaller) benefit to the largest number of people possible. Alternatively, we may wish to prioritize equity, and focus on providing assistance to those who are more vulnerable or more heavily impacted, even if it means helping a smaller number of people overall. We may wish to prioritize justice and thus take historical patterns of inequality and oppression into account as we decide which parts of our community to focus on. Of course, these are not necessarily mutually exclusive ways of thinking, but the important thing to note here is that this is not a question that we are able to answer using ML methods.

In terms of the action space available, we are similarly at the end of our road in terms of computational and statistical tools. We can offer suggestions informed by our analysis (some of the grant funding could go to cover air filter maintenance and upgrades to living facilities for the elderly, or to public outreach materials in specific languages) but fundamentally this may well be beyond our control.

We are *not*, however, at the limit of what we as data scientists can (and must) contribute from a methodological standpoint. Resolving questions of values and choices of action will involve disseminating the results of our analyses to the community as a whole and to policy makers and will involve a great deal of communication. A key part of this will involve helping the consumers of our results to understand the provenance of our findings, as well as what our level of uncertainty might be around individual conclusions. This may take, for example, the form of written reports, data visualizations (interactive or static), and simulations (answering "what if" or "for instance" questions), all of which are core parts of the data science process.

Open Questions and Challenges

Human in the Loop

One of the difficulties in being able to leverage fully the social genome data is that for a given research question, more data are not always better. In fact, as the following sections will demonstrate, often the plethora of what, at first glance, may seem like relevant data often will not turn out to be useful after more careful investigation. There are several reasons for this:

1. The sampling frame is unknown.
2. The variables are not measured in the right unit.
3. There are not sufficient variables available in the data to address the questions.
4. The different available datasets cannot be integrated to address the question.
5. Similar constructs are measured on different perspectives that do not align well.

"Garbage in, garbage out" is a principle all data scientists must heed. It is too easy to drown in data and lose sight of your research objectives. Thus, we posit that data science is a human-intensive intellectual activity that requires much thoughtful deliberation over many parts of the research, including research question development based on an understanding of current theories in the field, the feasibility of the study using available data, a thoughtful research plan based an appreciation for experimental design, inferential statistics principles, and measurement. Data science is more art than science due to the countless human judgments that are required. Data science is ultimately about sensemaking from raw data and trying to put the puzzle together to see the big picture. But for a particular puzzle, even though there may be a lot of pieces from lots of different puzzles, there may not be enough relevant pieces to complete the puzzle of interest. The data science team will usually have to fill in the blanks with good human judgment based on prior theories in the field, good empirical research, and understanding the limitations of big data. There are many barriers (e.g., aligning funding and authorship conventions with different disciplinary expectations and incentives) to working in interdisciplinary team science that will be important to navigating the field. For in-depth discussion, see Medeiros et al. (this volume).

Good Science, Bad Science, and Data Science

We should not confuse scientific inquiry with just running statistics on data. All statistical methods require subjective choices, and there is no objective decision machine for automated scientific inference. Thus, inference from the sample to a larger population must be scientific rather than statistical, even if we use inferential statistics. It must be scientists who make the inference, and "claims about a larger population will always be uncertain" (Amrhein et al. 2019; Gelman and Hennig 2017). We must remember that the acceptable level of uncertainty for scientific inquiry and public policy decision making is different from when recommending products online, and it requires a higher level of rigor and precision. In sum, good science naturally requires much thinking, judgment, dealing with uncertainties, hypothesis generation, hypothesis testing, and making correct interpretations after properly applying inferential statistics.

The full empirical scientific research cycle, as illustrated in Figure 12.3, involves observation–induction–deduction–testing–evaluation (De Groot and Spiekerman 1969). The first phase of observation and induction is the exploratory data analysis phase where broad general inquiries are being made to generate good research questions and hypotheses using inductive reasoning based on observed patterns and past theories in the field. The second phase of deduction, testing, and evaluation is the confirmatory data analysis phase where worthy hypotheses are carefully selected and tested through good experimental design, data collection, and analysis contributing to the knowledge base in the field including both the positive and negative results. What we learn from the confirmatory analysis should inform the next iteration of exploratory analysis, providing direction for what next questions should be investigated. Note that "finding the question is often more important than finding the answer" (Tukey 1980). Good empirical science has always been an iterative spiral process of exploratory analysis and confirmatory analysis, one careful analysis at a time giving insight, leading to a body of literature that together produces knowledge through many costly and time-consuming iterations between inductive and deductive reasoning.

What has changed with big data and data science is that now it allows for the full empirical scientific research cycle in one study. There is the potential in some studies to have enough data for even multiple iterations in the data lake, allowing for a much faster process of iterating between hypothesis generation and testing than ever before. In many ways data science is iterating between (a) traditional qualitative research and quantitative exploratory analysis, where the goal is to listen to the data and all of its constituent details as much

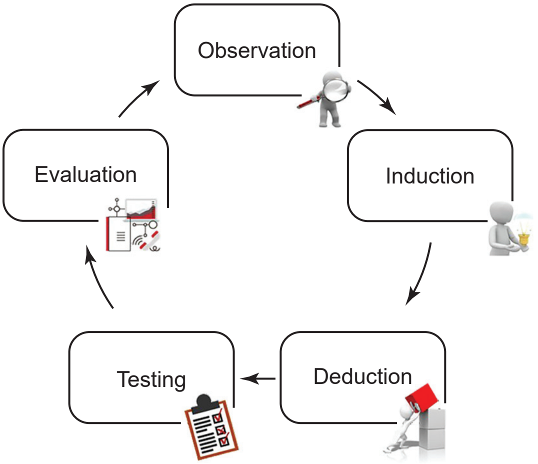


Figure 12.3 The empirical research cycle (De Groot and Spiekerman 1969).

as possible in an attempt to find the common patterns and generate good hypothesis through inductive reasoning, and (b) traditional quantitative research, where we conduct strict confirmatory analysis to test the hypothesis through deductive reasoning. In any particular study in science, however, these two phases are not always so clearly separated, and it is easy to blur the lines and lose track of what analysis is being done. This can lead to bad science, where we forget that hypothesis testing cannot be conducted on the very data used to suggest the hypothesis (Wagenmakers et al. 2012). There is a risk in the spiral iterating process to confuse hypothesis testing and generation, leading to a fishing expedition and over interpretation of the findings. To guard against this danger, we must remember two important statistical principles that are key to good science: proper sampling from a well-defined sampling frame and clearly planning out your research question and approach before touching the data, regardless of whether it is exploratory or confirmatory analysis.

First, in traditional sciences, one of the most important steps to get right is the sampling method. We must ensure that we use a representative random sample of the study population, including using stratified sampling to ensure smaller subgroups are properly represented. In studies where the target population is difficult to control, it is very important to clearly state the study population and note the acceptable scope of generalizability. For example, in our wildfire scenario, if only English-language social media posts were analyzed, noting it as a limitation of the study sample is very important to the interpretation of the results. Whenever possible, obtaining access to the full representative set of social media posts regardless of language and reporting out the percentage of the English-language posts in relation to the full universe will provide much better context for interpretation, even if limited time and resources only allow for analyzing English posts. In data science, because data collection often happens "out of band" as a separate activity as opposed to being part of the planned research itself, this key principle of sampling frame can get lost. We must remember, however, that no matter how much data we have, if there is not a proper understanding and description of the sampling frame, the results may be misleading or useless because it is not possible to interpret the results appropriately. A good example of this is the fact that even now, many years since the pandemic started, without a well-designed, nationally representative random sample for tracking infections and outcomes, we still do not know the incidence of COVID-19 in the United States, even with the many sources of data online about COVID-19 cases and deaths (Dean 2022). The rates estimated in the Stanford COVID-19 antibody study (Bendavid et al. 2021) were quickly challenged by statisticians (Gelman 2020; Gelman and Carpenter 2020).

Second, in confirmatory analysis, two major issues with applying inferential statistics on big data are that p-values are directly related to sample size, and that there are no good solutions to multiple statistical tests being performed on one dataset (Tukey 1980). Some consider the most conservative approach with Bonferroni Correction in these situations, but many believe it can create more problems than it solves (Perneger 1998). Others will pay more attention to the effect size rather than the p-values. Some have argued that there is no real alternative, and in most truly confirmatory studies, one must have "a single main question in which a question is [pre]specified by ALL of design, collection, monitoring, AND ANALYSIS" (Tukey 1980; see also Wagenmakers et al. 2012 and Miguel et al. 2014). The pre-specification of the study plan for confirmatory hypothesis testing analysis is very important but also often difficult to follow because there will likely be something that does not go as planned in real research; furthermore, in many scenarios involving secondary use of data, pre-specification is difficult because it is not always clear what data will be available and in what form. Further, the distinctions between exploratory analysis for hypothesis generation and confirmatory analysis are too often not understood, and results of exploratory analysis are reported and interpreted as confirmatory analysis, leading to bad science (Wagenmakers et al. 2012). Even in exploratory analysis where there are no hypotheses, it will be important to have thought through the main research question and be aware of the relevant literature in the field to guide the descriptive study (Miguel et al. 2014; Tukey 1980).

Data Integration, Aggregation, and Measurement

The data that form the basis for this type of research come from a variety of sources and are linked together to overcome the limitations of data collected for operations from one source, because alone they often do not contain sufficient information for a study. On one hand, the integration of the different data sources can augment the primary source and improve the completeness and comprehensiveness of information and potentially provide the important context for the data. On the other hand, errors, which exist in all real-world data, may get amplified when more datasets are linked together, making it more complex to track and bound error in the results (Baldi et al. 2010; Bollier 2010; Harron et al. 2017). Dealing with uncertainty and error is fundamental to working with any real-world data, but if it cannot be bounded in some way, it renders the results mostly useless and can often be misleading. Managing and bounding errors throughout the full data science process for proper interpretation of results is an open area of research.

Integrating data from disparate sources is rife with methodological as well as technical challenges. The method of linking individual or organizational level data is often referred to as record linkage (RL), or entity resolution (Dusetzina et al. 2014; Getoor and Machanavajjhala 2012; Gilbert et al. 2017; Karim et al. 2021). In the wildfire case study, ED data from different hospitals are likely to require RL to obtain unique people counts because different hospital systems will not have a common ID system. The absence of a common, error-free, unique identifier makes exact matching solutions inadequate, leading to approximate methods (probabilistic or deterministic) that require cleaning and standardizing data as well as manual resolution of ambiguous matches. It is an open area of research that is further complicated with issues of privacy and confidentiality due to the need to use identifiable information.

One line of research is the privacy preserving RL methods based on hashing. These methods are computationally set up to solve the private RL problem, which focuses on linking data securely given a predetermined linkage mapping function. These algorithms assume a machine-only system that limits human interaction, making it very difficult to determine the linkage function, clean and standardize data, as well as check on the validity of the results, which is critical in real applications (Hall and Fienberg 2010; Vatsalan et al. 2017). Another issue with machine-only RL systems is selection bias as a result of preferentially selecting patients with complete information on required identifiers. This can underrepresent particular groups, including the socioeconomically disadvantaged and racial/ethnic minorities (Bronstein et al. 2009; Harron et al. 2014). Thus, balancing the accuracy of RL with privacy is an active research area without a known technical solution (Hall and Fienberg 2010; Kum et al. 2013; Vatsalan et al. 2017). Recently, a more human-centered AI RL system has been proposed that allows researchers to integrate directly, but securely, individual-level data (Kum et al. 2013). MiNDFIRL (Minimum Necessary Disclosure For Interactive Record Linkage) uses ML for the automated components (Antonie et al. 2014; Ramezani et al. 2021) and interactive on-demand incremental information disclosure for privacy-aware manual review components (Kum et al. 2019; Ragan et al. 2018) that allow for optimizing both utility and privacy. It further facilitates data governance through template documents for privacy statement, DUA, and IRB application that communicate the complex parts of the technology used in the appropriate language for each community (Giannouchos et al. 2021; Kum et al. 2022; Schmit et al. 2020, 2024).

Besides technical issues, there are deeper and more fundamental problems that come from integrating data in this way. Datasets do not arise *ex nihilo*: they are designed, collected, postprocessed, and distributed by humans, in response to specific needs, values, and constraints. Along the way, those same humans make numerous conscious and unconscious choices that shape the final form of a dataset. Examples of such choices might include which underlying phenomena to capture and what abstraction and modeling compromises to make in order to represent the phenomena of interest; where and how to collect observations; which observations to include (and which to exclude); and what unit to aggregate to. Those humans are themselves operating within a variety of structural constraints that affect everything from the fundamental questions they are asking to the mechanics of how their data are collected. As such, datasets are in no way neutral (i.e., value-free) artifacts (Boyd and Crawford 2012).

It is important to note that this is not a critique; it is, rather, a reminder, and a simple observation about the nature of real-world data. It is crucial, then, to consider carefully the "story" behind any given dataset. This is particularly true in a secondary use scenario, in which, for instance, the values, constraints, and priorities that shaped one dataset may differ from another and may furthermore be quite different from those shaping your analysis. In practice, what does this look like? In the case of our wildfire scenario, one example might be a dataset of air quality monitoring records in which, due to logistics of how sensors are placed, there is an uneven spatial coverage across a city. In such a situation, some parts of the city may have been thoroughly covered, whereas the coverage in others may be sparse due to variability in budgeting and departmental priorities over time at the local branch office of the local Department of Environmental Quality. For its original scenarios of use, this irregular placement may not have posed issues. Our current analysis, however, needs to model conditions across the entire city; without taking this underlying issue into account, we could easily end up with estimates of our outcome of interest that varied in their accuracy according to geography.

When choosing whether and how to use a given dataset, one must ensure that the assumptions made by its originators are compatible with our present study. Even given that level of compatibility, though, we may encounter practical difficulties in directly integrating data points from disparate datasets if the underlying numbers are measuring qualitatively different phenomena. For example, to continue our wildfire analogy, let us imagine that the city and the state both have air quality monitoring programs, neither of which has complete geographic coverage of the metro area on their own, but which taken together have good coverage. May we combine the datasets?

To guide us in thinking through these kinds of challenges, and successfully integrating data in this way, we turn to measurement theory and its notions of *constructs* and *measurement models*. By *construct* we refer to a theoretical abstraction of the underlying phenomenon that a dataset is attempting to describe (e.g., air quality). Generally, such phenomena are unobservable and abstract, and must instead be explored using observable properties of the world. The process of doing so is referred to as *operationalizing* our construct via a measurement model. For example, consider the (unobservable) construct of "air quality": in the context of a wildfire smoke event, we would expect that a resident of our city might experience a decrease in their air quality; further, we might expect the amount of decrease to vary according to a number of different factors (e.g., wind, geography, the HVAC configuration of their home). Because "air quality" may mean many different things (e.g., concentration of a specific pollutant, or the presence or absence of some set of chemical pollutants), it may be operationalized (i.e., estimated via one or more observable phenomena) in a number of ways, depending on the specific needs of a given project. For instance, one study might operationalize air quality via a quantitative estimate of the concentration of particulate matter of a certain size (e.g., PM2.5), while another might focus on carbon monoxide concentrations. A third study might not have access to appropriate sensor data from a given geographic area, and thus might measure something more indirect, such as the number of ED visits with respiratory complaints. The degree to which a measure meaningfully models and reflects its underlying construct is referred to as its *construct validity*; often specific methodological and engineering choices are made around how to record an observable phenomenon in order to capture a particular construct adequately. The same observable phenomenon may furthermore be recorded in a very different manner (e.g., at a different timescale) depending on the underlying construct of interest.

For purposes of data integration, the first prerequisite, then, is that the data elements that we wish to integrate are attempting to represent the same construct. From there, many things become at least theoretically possible; assuming that our two measures (PM2.5 and ED visits) are indeed valid, it may be possible to combine them in some useful way, perhaps by calibrating them to one another and then computing a proxy variable of some kind, under the close guidance of a statistician accustomed to such methods.

Moving beyond integration of continuous data, similar issues can also arise with categorical data. A particularly common area of difficulty in data integration involves sociodemographic data (e.g., race and ethnicity categories). This is an extremely complex and challenging issue (Bowker and Star 1999) and there are no “good” answers, only more or less imperfect ones.

Matching Comparison Group in Observational Studies

One of the key characteristics of data science is that it relies on existing data sources. In scientific terms, it relies on observational data that were collected for another primary purpose (e.g., operating a hospital) outside of research. Thus, conducting research with these data is a secondary purpose. This means that researchers have no control over the data collection process and methodology and are limited to existing data. Thus, these studies are often called observational studies, secondary database studies, or retrospective studies. One of the main challenges when working with large existing databases is extracting meaningful measures and adjusting for the sampling that can address the research question, taking into account the limitations of how the data were collected, which often does not align well with the research question. This is very different from controlled experimental studies where data collection is carefully designed to manipulate the variables so that their effect upon other variables can be directly observed while other conditions are kept constant (Shadish et al. 2001).

Unfortunately, there are many experimental studies in sciences that are not possible for a variety of reasons, and the next best alternative may be observational studies using treatment and comparison groups that are carefully designed to adjust for covariates to the extent possible, either through multivariable modeling or matching. In our case study, investigating the differential impact of the wildfire across racial groups may benefit from gathering similar data from a matching comparison group from a city with similar characteristics but no wildfire to provide a baseline.

There are numerous variations for matching, using propensity scores, that can lead to many decisions:

- What are the appropriate covariates to match on?
- How many comparison samples should be matched to one treatment sample?
- What minimum caliper should be used?
- How exact does the match need to be?
- Should sampling be done with or without replacement?

Thus, it is important to think through "the design and compare several matched designs for an observational study just as one compares experimental designs before picking a satisfactory design" (Rosenbaum 2020). It is crucial that matching is conducted without access to any outcome data, thereby assuring the objectivity of the design. It is important to note that outcome data are specific to a given project and must occur after the event of interest (e.g., wildfire), and it should be distinguished from exposure data for the project, which occurs before the event of interest and may look similar to outcome data. For example, in the wildfire example, ED admission data from after the wildfire are outcome data, but ED admission data from before the wildfire may be covariates that measure the baseline condition of the community that should be accounted for in the analysis. This may be done in different ways such as baseline level of ED visits by zipcode before the wildfire. Thus, ED admission data may be used for matching, as long as it is a measurement that occurred before the event of interest. In addition, matching does not preclude additionally adjusting an estimate through multivariable modeling using the matched sample when appropriate (Rubin 1979). A good review of matching can be found in Rosenbaum (2020), who notably describes a methodological approach that follows very closely with the general data science approach, in that it involves exploring many different implementations iteratively for best insight and produces its final conclusion by synthesizing all results using human judgment. Another relevant approach is to use inverse probability of treatment weighting (IPTW), to weight the subjects to obtain unbiased estimates of average treatment effects in observational studies (Austin and Stuart 2015). Nonetheless, adjusting for observable differences in these ways does not fully address concerns that the treated and comparison groups may still differ in terms of unobserved covariates. This is a limitation of all observational studies, and it may be further exacerbated through matching if not carefully designed, because the matching process exacerbates the imbalance in the unobservable across groups (Brooks and Ohsfeldt 2013).

Other Considerations

We focused this paper on the role of human judgment in data science, which limited our discussion of other important topics. In this section, we briefly mention other open challenges to consider. First, randomly splitting the data into training, validation, and testing datasets is common practice in ML projects to avoid overfitting the data, and this technique is critical to having valid results in data science. This process facilitates finding the most generalizable model to keep the balance between bias and variance. On the one hand, this strict rule has parallels to exploratory analysis (training/validation phase) and confirmatory analysis (testing) phase in traditional science. On the other hand, there are sufficient differences between ML models and regression models, and better understanding of the commonality and distinctions would be helpful. One key distinction lies in the fact that ML is based on inductive reasoning while hypothesis testing using regression models are based on deductive reasoning, which gives rise to differences in interpretation. Recently, there have been advances in the bias-variance trade-off that may be of interest to those using ML (Belkin et al. 2019), but this is beyond the scope of this paper. Second, we have scoped this paper on challenges to analyzing existing secondary data sources, precluding discussion on simulations and electronic data collection (e.g., app, social media based), which may also be relevant to digital ethology. We refer interested readers in simulations to San Miguel et al. (2012) for a discussion on challenges in complex system science. In addition, some key topics were not included in this paper because they are discussed elsewhere in this volume. These include limited discussions on challenges to using social media data specifically, covered by Weigle et al. (this volume), as well as important discussions on ethics and data governance covered by Medeiros et al. (this volume).

Conclusion

We have outlined the challenges in using data science approaches to study large-scale population datasets, which we refer to as social genome data because the term has been used in other related fields to refer to the digital footprints left by humans (Kum et al. 2014; McGrail and Jones 2018). The library of social genome data can be used as a basis for inquiry, allowing analysts to answer complex high-level domain questions. We describe this process based on the DIKA pyramid, which provides a framework for approaching such problems. The ultimate goal is to allow data scientists, domain experts, and decision makers together to use the social genome data to produce actionable policy through the generation of new knowledge. We highlight many of the challenges in this process, including several difficult aspects of working with heterogeneous, error prone, real-world data, and we emphasize the essential role of data scientists in producing quality science in this area. A successful scientific inquiry using data science methods requires an expert toolsmith who can navigate the data lake with many computational and statistical tools to meet the domain goals, bringing in domain experts in the many decisions as appropriate (i.e., to help generate meaningful and feasible questions, decide on the right experimental design, operationalize measures, correctly interpret the findings, disseminate to appropriate audiences) to build a well-documented, transparent, and reusable process. The data scientist must pay attention to experimental details, remembering the key principles of statistical inference, such as sampling frames, uncertainty management, and the difference between exploratory and confirmatory analysis. This requires sensemaking by iteratively zooming in and out as appropriate. There is no one formula or method for how to analyze such data, and there are many pitfalls that can be encountered. Applying data science for rigorous scientific inquiry depends upon the judgment, expertise, and experience of the entire study team.

Acknowledgments

We thank Mark Fossett and Ernesto Amaral for their in-depth discussion on good science practices, and Robert Ohsfeldt for his review and suggestions for the section on matching comparison group.

Bibliography

Note: Numbers in square brackets denote the chapter in which an entry is cited.

Ackoff, R. L. 1989. From Data to Wisdom: Presidential Address to ISGSR, June 1988. *J. Appl. Syst. Anal.* **16**:3–9. [12]

Adami, H. O., and O. Nyrén. 2016. Enigmas, Priorities and Opportunities in Cancer Epidemiology. *Eur. J. Epidemiol.* **31**:1161–1171. [11]

ADB. 2019. Asian Development Outlook 2019. Fostering Growth and Inclusion in Asian´S Cities. Manila, Philippines: Asian Development Bank. [8]

Aggarwal, C. C., and T. Abdelzaher. 2013. Social Sensing. In: Managing and Mining Sensor Data, ed. C. C. Aggarwal, pp. 237–297. Boston: Kluwer. [7]

Aitken, M., J. de St Jorre, C. Pagliari, R. Jepson, and S. Cunningham-Burley. 2016. Public Responses to the Sharing and Linkage of Health Data for Research Purposes: A Systematic Review and Thematic Synthesis of Qualitative Studies. *BMC Med Ethics* **17**:73. [1]

Al Nuaimi, E., H. Al Neyadi, N. Mohamed, and J. Al-Jaroodi. 2015. Applications of Big Data to Smart Cities. *J. Internet Serv. Applicat.* **6**:25. [6]

Aleixo, R., F. Kon, R. Rocha, M. Santos Camargo, and R. Y. De Camargo. 2022. Predicting Dengue Outbreaks with Explainable Machine Learning. In: 22nd Intl. Symposium on Cluster Computing and the Grid (CCGRID), pp. 940–947. Taormina, Italy: IEEE. [3]

Alemy, A., S. Hudzik, and C. N. Matthews. 2017. Creating a User-Friendly Interactive Interpretive Resource with ESRI's ArcGIS Story Map Program. *Hist. Archaeol.* **51**:288–297. [3]

Althoff, T., R. W. White, and E. Horvitz. 2016. Influence of Pokémon Go on Physical Activity: Study and Implications. *J. Med. Internet Res.* **18**:e315. [4]

Altmann, J. 1974. Observational Study of Behavior: Sampling Methods. *Behaviour* **49**:227–266. [2, 9]

Amato, P. R. 1983. Helping Behavior in Urban and Rural Environments: Field Studies Based on a Taxonomic Organization of Helping Episodes. *J. Pers. Soc. Psychol.* **45**:571–586. [8]

Ambuhl, L., A. Loder, L. Leclercq, and M. Menendez. 2021. Disentangling the City Traffic Rhythms: A Longitudinal Analysis of MFD Patterns over a Year. *Transport. Res. C Emerg. Technol.* **126**:103065. [8]

Amrhein, V., D. Trafimow, and S. Greenland. 2019. Inferential Statistics as Descriptive Statistics: There Is No Replication Crisis If We Don't Expect Replication. *Am. Stat.* **73**:262–270. [12]

Anand, A., and J. Pathak. 2022. The Role of reddit in the Gamestop Short Squeeze. *Econ. Lett.* **211**:110249. [10]

Anderson, D. J., and P. Perona. 2014. Toward a Science of Computational Ethology. *Neuron* **84**:18–31. [2, 9]

Andor, L. 2019. Fifteen Years of Convergence: East-West Imbalance and What the EU Should Do About It. *Intereconomics* **54**:18–23. [4]

André Hutson, M., G. A. Kaplan, N. Ranjit, and M. S. Mujahid. 2012. Metropolitan Fragmentation and Health Disparities: Is There a Link? *Milbank Q.* **90**:187–207. [3]

Antonie, L., K. Inwood, D. J. Lizotte, and J. Andrew Ross. 2014. Tracking People over Time in 19th Century Canada for Longitudinal Analysis. *Mach. Learn.* **95**:129–146. [12]

Anwar, M., D. Khoury, A. P. Aldridge, S. J. Parker, and K. P. Conway. 2020. Using Twitter to Surveil the Opioid Epidemic in North Carolina: An Exploratory Study. *JMIR Public Health Surveill.* **7**:e17574. [10]

Apicella, C. L., F. W. Marlowe, J. H. Fowler, and N. A. Christakis. 2012. Social Networks and Cooperation in Hunter-Gatherers. *Nature* **481**:497–501. [4]

Appleyard, D., M. S. Gerson, and M. Lintell. 1981. Livable Streets. Oakland: Univ. California Press. [8]

Arora, A., and A. Arora. 2022. Synthetic Patient Data in Health Care: A Widening Legal Loophole. *Lancet* **399**:1601–1602. [5]

Auchincloss, A. H., A. V. Diez Roux, D. G. Brown, E. S. O'Meara, and T. E. Raghunathan. 2006. Association of Insulin Resistance with Distance to Wealthy Areas: The Multi-Ethnic Study of Atherosclerosis. *Am. J. Epidemiol.* **165**:389–397. [4]

Aureli, F., C. M. Schaffner, C. Boesch, et al. 2008. Fission-Fusion Dynamics: New Research Frameworks. *Curr. Anthropol.* **49**:627–654. [4]

Austin, J. L. 1975. How to Do Things with Words. Oxford: Oxford Univ. Press. [9]

Austin, P. C., and E. A. Stuart. 2015. Moving Towards Best Practice When Using Inverse Probability of Treatment Weighting (IPTW) Using the Propensity Score to Estimate Causal Treatment Effects in Observational Studies. *Stat. Med.* **34**:3661–3679. [12]

Auxier, B., and M. Anderson. 2021. Social Media Use in 2021. Washington, D.C.: Pew Research Center. [4]

Baeza-Yates, R. 2020. Biases on Social Media Data: (Keynote Extended Abstract). In: WWW '20: The Web Conference 2020, pp. 782–783. New York: ACM. [4]

Bai, D., N. Marrus, B. H. K. Yip, et al. 2020. Inherited Risk for Autism through Maternal and Paternal Lineage. *Biol. Psychiatry* **88**:480–487. [11]

Bai, D., B. H. K. Yip, G. C. Windham, et al. 2019. Association of Genetic and Environmental Factors with Autism in a 5-Country Cohort. *JAMA Psychiatry* **76**:1035–1043. [11]

Bail, C. A., L. P. Argyle, T. W. Brown, et al. 2018. Exposure to Opposing Views on Social Media Can Increase Political Polarization. *PNAS* **115**:9216–9221. [4]

Bakeman, R. 2023. Kappaacc: A Program for Assessing the Adequacy of Kappa. *Behav. Res. Meth.* **55**:633–638. [2]

Bakeman, R., and V. Quera. 2011. Sequential Analysis and Observational Methods for the Behavioral Sciences. Cambridge: Cambridge Univ. Press. [2]

Baker, S. J., M. Jackson, H. Jongsma, and C. W. N. Saville. 2021. The Ethnic Density Effect in Psychosis: A Systematic Review and Multilevel Meta-Analysis. *Br. J. Psychiatry* **219**:632–643. [4]

Bakker, M. A., D. A. Piracha, P. J. Lu, et al. 2019. Measuring Fine-Grained Multidimensional Integration Using Mobile Phone Metadata: The Case of Syrian Refugees in Turkey. In: Guide to Mobile Data Analytics in Refugee Scenarios, pp. 123–140, A. Salah, Pentland, A., Lepri, B., Letouzé, E., series ed. Cham: Springer. [4, 8]

Baldi, I., A. Ponti, R. Zanetti, et al. 2010. The Impact of Record-Linkage Bias in the Cox Model. *J. Eval. Clin. Pract.* **16**:92–96. [12]

Balsa-Barreiro, J., Y. Li, A. J. Morales, and A. Pentland. 2019a. Globalization and the Shifting Centers of Gravity of World's Human Dynamics: Implications for Sustainability. *J. Clean. Prod.* **239**:117923. [8]

Balsa-Barreiro, J., and M. Menendez. 2021. Cómo Son y Cómo Se Mueven Las Redes Urbanas. Los Angeles Times. https://www.latimes.com/espanol/opinion/articulo/2021-10-27/opinion-como-son-y-como-se-mueven-las-redes-urbanas (accessed Jan. 23, 2024). [8]

———. 2022. Fisionomia y Flujos de Trafico ¿Como Entender la Movilidad en Las Ciudades? Foreign Affairs Latinoamerica. https://revistafal.com/fisionomia-urbana-y-flujos-de-trafico/ (accessed Jan. 23, 2024). [8]

Balsa-Barreiro, J., M. Menendez, and A. J. Morales. 2022. Scale, Context, and Heterogeneity: The Complexity of the Social Space. *Sci. Rep.* **12**:9037. [4, 8]

Balsa-Barreiro, J., A. J. Morales, and R. C. Lois-González. 2021. Mapping Population Dynamics at Local Scales Using Spatial Networks. *Complexity* **2021**:8632086. [8]

Balsa-Barreiro, J., A. M. Morales, and E. Castelló. 2018. Datos, Inteligencia Artificial y Complejidad. Una Visión de la Sociedad del Futuro. Instituto de Ingeniería de España. Madrid: Instituto de Ingeniería de España. [8]

Balsa-Barreiro, J., P. M. Valero-Mora, J. L. Berné-Valero, and F. Varela-García. 2019b. GIS Mapping of Driving Behavior Based on Naturalistic Driving Data. *ISPRS Intl. Journal of Geo-Information* **8**:226. [4]

Balsa-Barreiro, J., P. M. Valero-Mora, M. Menéndez, and R. Mehmood. 2020a. Extraction of Naturalistic Driving Patterns with Geographic Information Systems. *Mob. Netw. Appl.* **28**: 619–635. [4]

Balsa-Barreiro, J., A. Vié, A. J. Morales, and M. Cebrian. 2020b. Deglobalization in a Hyper-Connected World. *Palgrave Commun.* **6**:28. [4, 8]

Bard, K. A., H. Keller, K. M. Ross, et al. 2021. Joint Attention in Human and Chimpanzee Infants in Varied Socio-Ecological Contexts. *Monogr. Soc. Res. Child Devel.* **86**:7–217. [4, 9]

Bartelme, N. 2022. Geographic Information Systems. In: Springer Handbook of Geographic Information, ed. W. Kresse and D. Danko, pp. 121–149, Springer Handbooks. Cham: Springer. [6]

Batbaatar, E., and K. H. Ryu. 2019. Ontology-Based Healthcare Named Entity Recognition from Twitter Messages Using a Recurrent Neural Network Approach. *Int. J. Environ. Res. Public Health* **16**:3628. [4]

Batista, D. M., A. Goldman, R. Hirata, et al. 2016. Interscity: Addressing Future Internet Research Challenges for Smart Cities. In: 7th Intl. Conf. on the Network of the Future, p. 10.1109/NOF.2016.7810114. loBuzios, Brazil: IEEE. [3]

Batty, M., E. Besussi, and N. Chin. 2003. Traffic, Urban Growth and Suburban Sprawl
London: Bartlett Centre for Advanced Spatial Analysis. [8]

Batty, M., and P. Longley. 1994. Fractal Cities: A Geometry of Form and Function. Cambridge, MA: Academic Press. [8]

Baum-Snow, N. 2007. Did Highways Cause Suburbanization? *Q. J. Econ.* **122**:775–805. [8]

Baumeister, R. F., K. D. Vohs, and D. C. Funder. 2007. Psychology as the Science of Self-Reports and Finger Movements: Whatever Happened to Actual Behavior? *Perspect. Psychol. Sci.* **2**:396–403. [9]

Beauté, J., S. Sandin, S. A. Uldum, et al. 2016. Short-Term Effects of Atmospheric Pressure, Temperature, and Rainfall on Notification Rate of Community-Acquired Legionnaires' Disease in Four European Countries. *Epidemiol. Infect.* **144**:3483–3493. [11]

Becker, S. J., B. R. Garner, and B. J. Hartzler. 2021. Is Necessity Also the Mother of Implementation? COVID-19 and the Implementation of Evidence-Based Treatments for Opioid Use Disorders. *J. Subst. Abuse Treat.* **122**:108210. [10]

Beelen, R., G. Hoek, D. Vienneau, et al. 2013. Development of NO2 and NOx Land Use Regression Models for Estimating Air Pollution Exposure in 36 Study Areas in Europe: The Escape Project. *Atmos. Environ.* **72**:10–23. [7]

Been, K., E. Daiches, and C. Yap. 2006. Dynamic Map Labeling. *IEEE Trans. Vis. Comput. Graph.* **12**:773–780. [6]

Beilschmidt, C., J. Drönner, M. Mattig, et al. 2017. VAT: A Scientific Toolbox for Interactive Geodata Exploration. *Datenbank-Spektrum* **17**:233–243. [6]

Belkaroui, R., R. Faiz, and P. Kuntz. 2015. User-Tweet Interaction Model and Social Users Interactions for Tweet Contextualization. In: Computational Collective Intelligence, ed. M. Núñez et al., pp. 144–157. Cham: Springer. [2]

Belkin, M., D. Hsu, S. Ma, and S. Mandal. 2019. Reconciling Modern Machine-Learning Practice and the Classical Bias–Variance Trade-Off. *PNAS* **116**:15849–15854. [12]

Bendavid, E., B. Mulaney, N. Sood, et al. 2021. COVID-19 Antibody Seroprevalence in Santa Clara County, California. *Int. J. Epidemiol.* **50**:410–419. [12]

Bennett, D. A., D. Landry, J. Little, and C. Minelli. 2017. Systematic Review of Statistical Approaches to Quantify, or Correct for, Measurement Error in a Continuous Exposure in Nutritional Epidemiology. *BMC Med. Res. Methodol.* **17**:146. [5]

Bergeron, J., D. Doiron, Y. Marcon, V. Ferretti, and I. Fortier. 2018. Fostering Population-Based Cohort Data Discovery: The Maelstrom Research Cataloguing Toolkit. *PLOS ONE* **13**:e0200926. [11]

Berkman, L. F., I. Kawachi, and M. M. Glymour, eds. 2014. Social Epidemiology. New York: Oxford Univ. Press. [4]

Bernasco, W., E. Hoeben, D. Koelma, et al. 2022. Promise into Practice: Application of Computer Vision in Empirical Research on Social Distancing. *Sociol. Methods Res.* **May 9**:1239–1287. [9]

Bettencourt, L. M. A. 2013. The Origins of Scaling in Cities. *Science* **340**:1438–1441. [8]

Bezuidenhout, L. 2013. Data Sharing and Dual-Use Issues. *Sci. Eng. Ethics* **19**:83–92. [5]

Bhattacharya, K., A. Ghosh, D. Monsivais, R. I. M. Dunbar, and K. Kaski. 2016. Sex Differences in Social Focus across the Life Cycle in Humans. *R. Soc. Open Sci.* **3**:160097. [4]

Bhopal, R. S. 1993. Geographical Variation of Legionnaires' Disease: A Critique and Guide to Future Research. *Int. J. Epidemiol.* **22**:1127–1136. [11]

Biljecki, F., and Y. S. Chow. 2022. Global Building Morphology Indicators. *Comput. Environ. Urban Syst.* **95**:101809. [8]

Bjorkenstam, E., S. Cheng, B. Burstrom, et al. 2017. Association between Income Trajectories in Childhood and Psychiatric Disorder: A Swedish Population-Based Study. *J. Epidemiol. Commun. Health* **71**:648–654. [1]

Black, C., Y. Tesfaigzi, J. A. Bassein, and L. A. Miller. 2017. Wildfire Smoke Exposure and Human Health: Significant Gaps in Research for a Growing Public Health Issue. *Environ. Toxicol. Pharmacol.* **55**:186–195. [12]

Blank, G., and C. Lutz. 2017. Representativeness of Social Media in Great Britain: Investigating Facebook, Linkedin, Twitter, Pinterest, Google+, and Instagram. *Am. Behav. Sci.* **61**:741–756. [4]

Blei, D. M., and P. Smyth. 2017. Science and Data Science. *PNAS* **114**:8689–8692. [12]

Bleiholder, J., and F. Naumann. 2009. Data Fusion. *ACM Comput. Surv.* **41**:Article 1. [5]

Bloch, C., L. S. Liebst, P. Poder, J. M. Christiansen, and M. B. Heinskou. 2018. Caring Collectives and Other Forms of Bystander Helping Behavior in Violent Situations. *Curr. Sociol.* **66**:1049–1069. [9]

Blok, A., and M. A. Pedersen. 2014. Complementary Social Science? Quali-Quantitative Experiments in a Big Data World. *Big Data Soc.* **1**:10.1177/2053951714543908. [9]

Boeing, G. 2019. Urban Spatial Order: Street Network Orientation, Configuration, and Entropy. *Appl. Netw. Sci.* **4**:67. [8]

Boivin, A., T. Richards, L. Forsythe, et al. 2018. Evaluating Patient and Public Involvement in Research. *Br. Med. J.* **363**:k5147. [3]

Bollier, D. 2010. The Promise and Peril of Big Data. Washington, D.C.: Aspen Institute. [12]

Borck, R., and T. Tabuchi. 2019. Pollution and City Size: Can Cities Be Too Small? *Journal of Economic Geography* **19**:995–1020. [8]

Bordogna, G., S. Capelli, and G. Psaila. 2017. A Big Geo Data Query Framework to Correlate Open Data with Social Network Geotagged Posts. In: Societal Geo-Innovation, ed. A. Bregt et al., pp. 185–203, Lecture Notes in Geoinformation and Cartography. Cham: Springer. [6]

Borgman, C. L., P. T. Darch, A. E. Sands, et al. 2015. Knowledge Infrastructures in Science: Data, Diversity, and Digital Libraries. *Int. J. Digit. Libr.* **16**:207–227. [12]

Bossetta, M. 2018. The Digital Architectures of Social Media: Comparing Political Campaigning on Facebook, Twitter, Instagram, and Snapchat in the 2016 U.S. Election:. *JMCQ* **95**:471–496. [10]

Botts, M., G. Percivall, C. Reed, and J. Davidson. 2013. OGC Sensor Web Enablement: Overview and High Level Architecture. Open Geospatial Consortium. https://docs.ogc.org/wp/07-165r1/ (accessed Jan. 19, 2024). [6]

Bourdic, L., S. Salat, and C. Nowacki. 2012. Assessing Cities: A New System of Cross-Scale Spatial Indicators. *Building Research & Information* **40**:592–605. [8]

Bowker, G. C., and S. L. Star. 1999. Sorting Things Out: Classification and Its Consequences. Cambridge, MA: MIT Press. [12]

Bowman, D. M. 2013. The Hare and the Tortoise: An Australian Perspective on Regulating New Technologies and Their Products and Processes. In: Innovative Governance Models for Emerging Technologies, pp. 155–175. Cheltenham: Edward Elgar Publ. [5]

Boyd, A., J. Golding, J. Macleod, et al. 2013. Cohort Profile: The "Children of the 90s"—the Index Offspring of the Avon Longitudinal Study of Parents and Children. *Int. J. Epidemiol.* **42**:111–127. [1]

Boyd, D., and K. Crawford. 2012. Critical Questions for Big Data. *Inform. Commun. Soc.* **15**:662–679. [12]

Boyle, E. A., Y. I. Li, and J. K. Pritchard. 2017. An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell* **169**:1177–1186. [2]

Bragg, H., H. R. Jayanetti, M. L. Nelson, and M. C. Weigle. 2023. Less Than 4% of Archived Instagram Account Pages for the Disinformation Dozen Are Replayable. In: Proc. of the ACM/IEEE Joint Conference on Digital Libraries (JCDL), pp. 102–106. New York: ACM. [4]

Brassel, K. E., and R. Weibel. 1988. A Review and Conceptual Framework of Automated Map Generalization. *Int. J. Geographic. Inf. Syst.* **2**:229–244. [6]

Braubach, M., A. Egorov, P. Mudu, et al. 2017. Effects of Urban Green Space on Environmental Health, Equity and Resilience. In: Theory and Practice of Urban Sustainability Transitions, ed. N. Kabisch et al., pp. 187–205. Cham: Springer. [8]

Brazhnik, O., and J. F. Jones. 2007. Anatomy of Data Integration. *J. Biomed. Inform.* **40**:252–269. [5]

Brinkhoff, T. 2020. Determining Point Locations of Populated Places by Using Area Datasets. In: Geospatial Technologies for Local and Regional Development, ed. P. Kyriakidis et al., paper 74, Springer Ebooks Earth and Environmental Science. Cham: Springer. [6]

———. 2022. Geodatenbanksysteme in Theorie und Praxis: Einführung unter besonderer Berücksichtigung von PostGIS und Oracle. Berlin: Wichmann. [6]

Brinkhoff, T., H.-P. Kriegel, R. Schneider, and B. Seeger. 1994. Multi-Step Processing of Spatial Joins. *SIGMOD Rec.* **23**:197–208. [5]

Brodeur, J., S. Coetzee, D. Danko, S. Garcia, and J. Hjelmager. 2019. Geographic Information Metadata: An Outlook from the International Standardization Perspective. *ISPRS Int. J. Geo-Inf.* **8**:280. [6]

Brondino, N., L. Fusar-Poli, and P. Politi. 2017. Something to Talk about: Gossip Increases Oxytocin Levels in a near Real-Life Situation. *Psychoneuroendocrinol.* **77**:218–224. [4]

Broniatowski, D. A., M. J. Paul, and M. Dredze. 2013. National and Local Influenza Surveillance through Twitter: An Analysis of the 2012-2013 Influenza Epidemic. *PLOS ONE* **8**:e83672. [10]

Bronstein, J. M., C. T. Lomatsch, D. Fletcher, et al. 2009. Issues and Biases in Matching Medicaid Pregnancy Episodes to Vital Records Data: The Arkansas Experience. *Matern. Child Health J.* **13**:250–259. [12]

Brook, J. R., E. M. Setton, E. Seed, et al. 2018. The Canadian Urban Environmental Health Research Consortium: A Protocol for Building a National Environmental Exposure Data Platform for Integrated Analyses of Urban Form and Health. *BMC Public Health* **18**:114. [1]

Brooke, H. L., M. Talbäck, J. Hörnblad, et al. 2017. The Swedish Cause of Death Register. *Eur. J. Epidemiol.* **32**:765–773. [11]

Brooks, J. M., and R. L. Ohsfeldt. 2013. Squeezing the Balloon: Propensity Scores and Unmeasured Covariate Balance. *Health Serv. Res.* **48**:1487–1507. [12]

Brousse, O., C. Simpson, N. Walker, et al. 2022. Evidence of Horizontal Urban Heat Advection in London Using Six Years of Data from a Citizen Weather Station Network. *Environ. Res. Lett.* **17**:044041. [7]

Brown, D. 1991. Human Universals. Philadelphia: Temple Univ. Press. [9]

Brulle, R. J., and D. N. Pellow. 2006. Environmental Justice: Human Health and Environmental Inequalities. *Annu. Rev. Public Health* **27**:103–124. [12]

Brum-Bastos, V. S., J. A. Long, and U. Demšar. 2018. Weather Effects on Human Mobility: A Study Using Multi-Channel Sequence Analysis. *Comput. Environ. Urban Syst.* **71**:131–152. [7]

Brunekreef, B., and S. T. Holgate. 2002. Air Pollution and Health. *Lancet* **360**:1233–1242. [7]

Brunelle, J. F., M. Kelly, M. C. Weigle, and M. L. Nelson. 2016. The Impact of Javascript on Archivability. *Int. J. Digit. Libr.* **17**:95–117. [4]

Büchel, K., and M. von Ehrlich. 2020. Cities and the Structure of Social Interactions: Evidence from Mobile Phone Data. *J. Urban Econ.* **119**:103276. [8]

Buckley, A., P. Hardy, and K. Field. 2022. Cartography. In: Springer Handbook of Geographic Information, ed. W. Kresse and D. Danko, pp. 315–352. Springer Handbooks. Cham: Springer. [6]

Bulcock, A., L. Hassan, S. Giles, et al. 2021. Public Perspectives of Using Social Media Data to Improve Adverse Drug Reaction Reporting: A Mixed-Methods Study. *Drug Saf.* **44**:553–564. [4]

Bulmer, M. 1984. The Chicago School of Sociology: Institutionalization, Diversity, and the Rise of Sociological Research. Chicago: Univ. Chicago Press. [8]

Burger, M. J., P. S. Morrison, M. Hendriks, and M. M. Hoogerbrugge. 2020. Urban-Rural Happiness Differentials across the World. https://worldhappiness.report/ed/2020/urban-rural-happiness-differentials-across-the-world/ (accessed Jan. 23, 2024). [8]

Burgess, R. 2000. The Compact City Debate: A Global Perspective. In: Compact Cities. Sustainable Urban Forms for Developing Countries, ed. R. Burgess and M. Jenks, pp. 21–36. London: Routledge. [8]

Burke, G., and J. Dearen. 2022. Tech Tool Offers Police "Mass Surveillance on a Budget". Associated Press. https://apnews.com/article/technology-police-government-surveillance-d395409ef5a8c6c3f6cdab5b1d0e27ef (accessed Nov. 1, 2022). [3]

Burlew, K., C. McCuistian, and J. Szapocznik. 2021. Racial/Ethnic Equity in Substance Use Treatment Research: The Way Forward. *Addict. Sci. Clin. Pract.* **16**:1–6. [10]

Callard, F., and E. Perego. 2021. How and Why Patients Made Long Covid. *Soc. Sci. Med.* **268**:113426. [10]

Caminha, C., V. Furtado, T. H. C. Pequeno, et al. 2017. Human Mobility in Large Cities as a Proxy for Crime. *PLOS ONE* **12**:e0171609. [8]

Cândido, R. L., M. Steinmetz-Wood, P. Morency, and Y. Kestens. 2018. Reassessing Urban Health Interventions: Back to the Future with Google Street View Time Machine. *Am. J. Prev. Med.* **55**:662–669. [7]

Candiloro, T. 2023. The Best and Worst Cities for Commuters in 2022. https://listwithclever.com/research/best-and-worst-cities-for-commmuters-2022/ (accessed Jan. 23, 2024). [8]

Cao, L. 2017. Data Science: A Comprehensive Overview. *ACM Comput. Surv.* **50**:1–42. [12]

Carballada, A., and J. Balsa-Barreiro. 2021. Geospatial Analysis and Mapping Strategies for Fine-Grained and Detailed COVID-19 Data with GIS. *ISPRS Int. J. Geo-Inf.* **10**:602. [8]

Carinci, F. 2020. Covid-19: Preparedness, Decentralisation, and the Hunt for Patient Zero. *Br. Med. J.* **368**:bmj.m799. [10]

Carollo, A., P. Montefalcone, M. H. Bornstein, and G. Esposito. 2023. A Scientometric Review of Infant Cry and Caregiver Responsiveness: Literature Trends and Research Gaps over 60 Years of Developmental Study. *Children (Basel)* **10**:1042. [1]

Carroll, S. R., I. Garba, O. L. Figueroa-Rodríguez, et al. 2020. The CARE Principles for Indigenous Data Governance. *Data Sci. J.* **19**:43. [5]

Carter, K. W., R. W. Francis, K. Carter, et al. 2016. ViPAR: A Software Platform for the Virtual Pooling and Analysis of Research Data. *Int. J. Epidemiol.* **45**:408–416. [11]

Castro-Ramirez, F., M. Al-Suwaidi, P. Garcia, et al. 2021. Racism and Poverty Are Barriers to the Treatment of Youth Mental Health Concerns. *J. Clin. Child Adolesc. Psychol.* **50**:534–546. [1]

Catalog of Bias. 2018. Confounding by Indication. https://catalogofbias.org/biases/confounding-by-indication/ (accessed Nov. 9, 2022). [11]

Centers for Disease Control and Prevention. 2020. Wide-Ranging Online Data for Epidemiologic Research (WONDER). Centers for Disease Control and Prevention. https://wonder.cdc.gov/ (accessed Dec. 5, 2023). [10]

Cesare, N., H. Lee, T. McCormick, E. Spiro, and E. Zagheni. 2018. Promises and Pitfalls of Using Digital Traces for Demographic Research. *Demography* **55**:1979–1999. [12]

Chancellor, S., S. A. Sumner, C. David-Ferdon, T. Ahmad, and M. de Choudhury. 2021. Suicide Risk and Protective Factors in Online Support Forum Posts: Annotation Scheme Development and Validation Study. *JMIR Ment. Health* **8**:e24471. [10]

Charreire, H., C. Weber, B. Chaix, et al. 2012. Identifying Built Environmental Patterns Using Cluster Analysis and GIS: Relationships with Walking, Cycling and Body Mass Index in French Adults. *Int. J. Behav. Nutr. Phys. Act.* **9**:59. [6]

Chary, M., N. Genes, C. Giraud-Carrier, et al. 2017. Epidemiology from Tweets: Estimating Misuse of Prescription Opioids in the USA from Social Media. *J. Med. Toxicol.* **13**:278–286. [10]

Chen, E., K. Lerman, and E. Ferrara. 2020. Tracking Social Media Discourse About the COVID-19 Pandemic: Development of a Public Coronavirus Twitter Data Set. *JMIR Public Health Surveill.* **6**:e19273. [10]

Chen, J., J. Chen, A. Liao, et al. 2015. Global Land Cover Mapping at 30 M Resolution: A POK-Based Operational Approach. *ISPRS J. Photogramm. Remote Sens.* **103**:7–27. [7]

Chen, J., and Y. Wang. 2021. Social Media Use for Health Purposes: Systematic Review. *J. Med. Internet Res.* **23**:e17917. [10]

Chen, W., and R. Mace. 2019. Large-Scale Cooperation Driven by Reputation, Not Fear of Divine Punishment. *R. Soc. Open Sci.* **6**:190991. [8]

Chen, X., C. Faviez, S. Schuck, et al. 2018. Mining Patients' Narratives in Social Media for Pharmacovigilance: Adverse Effects and Misuse of Methylphenidate. *Front. Pharmacol.* **9**:541. [4]

Chenworth, M., J. Perrone, J. S. Love, et al. 2021. Methadone and Suboxone® Mentions on Twitter: Thematic and Sentiment Analysis. *Clin. Toxicol.* **59**:982–991. [10]

Choi, D.-a., and R. Ewing. 2021. Effect of Street Network Design on Traffic Congestion and Traffic Safety. *Journal of Transport Geography* **96**:103200. [8]

Christakis, N. A. 2019. Blueprint: The Evolutionary Origins of a Good Society. New York: Little, Brown Spark. [9]

Ciccarone, D. 2021. The Rise of Illicit Fentanyls, Stimulants and the Fourth Wave of the Opioid Overdose Crisis. *Curr. Opin. Psychiatry* **34**:344–350. [10]

Cinelli, M., G. De Francisci Morales, A. Galeazzi, W. Quattrociocchi, and M. Starnini. 2021. The Echo Chamber Effect on Social Media. *PNAS* **118**:e2023301118. [4]

Clark, B., K. Chatterjee, A. Martin, and A. Davis. 2020. How Commuting Affects Subjective Wellbeing. *Transportation* **47**:2777–2805. [8]

Cohen, N., M. Chrobok, and O. Caruso. 2020. Google-Truthing to Assess Hot Spots of Food Retail Change: A Repeat Cross-Sectional Street View of Food Environments in the Bronx, New York. *Health Place* **62**:102291. [7]

Collaborative Group on Hormonal Factors in Breast Cancer. 1997. Breast Canc7er and Hormone Replacement Therapy: Collaborative Reanalysis of Data from 51 Epidemiological Studies of 52 705 Women with Breast Cancer and 108 411 Women without Breast Cancer. *Lancet* **350**:1047–1059. [11]

Collins, R. 1994. Why the Social Sciences Won't Become High-Consensus, Rapid-Discovery Science. *Sociol. Forum* **9**:155–177. [9]

———. 2008. Violence: A Micro-Sociological Theory. Princeton: Princeton Univ. Press. [9]

Connelly, R., C. J. Playford, V. Gayle, and C. Dibben. 2016. The Role of Administrative Data in the Big Data Revolution in Social Science Research. *Soc. Sci. Res.* **59**:1–12. [7]

Corscadden, K., A. Wile, and E. Yiridoe. 2012. Social License and Consultation Criteria for Community Wind Projects. *Renew. Energy* **44**:392–397. [5]

Coulmont, B., and P. Simon. 2019. Quels Prénoms Les Immigrés Donnent-Ils À Leurs Enfants en France? *Popul. Soc.* **565**:1–4. [4]

Couper, M. P., and R. M. Groves. 1996. Social Environmental Impacts on Survey Cooperation. *Qual. Quant.* **30**:173–188. [8]

Craver, C. F. 2007. Explaining the Brain: Mechanisms and the Mosaic Unity of Neuroscience. Oxford: Clarendon Press. [2]

Cresswell, T. 2004. Place: A Short Introduction. Hoboken: Wiley-Blackwell. [4]

Crystal-Ornelas, R., C. Varadharajan, B. Bond-Lamberty, et al. 2021. A Guide to Using GitHub for Developing and Versioning Data Standards and Reporting Formats. *Earth Space Sci.* **8**:e2021EA001797. [3]

Cui, Y., K. M. Eccles, R. K. Kwok, et al. 2022. Integrating Multiscale Geospatial Environmental Data into Large Population Health Studies: Challenges and Opportunities. *Toxics* **10**:403. [5]

Curtis, J. W., A. Curtis, J. Mapes, A. B. Szell, and A. Cinderich. 2013. Using Google Street View for Systematic Observation of the Built Environment: Analysis of Spatio-Temporal Instability of Imagery Dates. *Int. J. Health Geogr.* **12**:53. [1]

Cuthill, I. 1991. Field Experiments in Animal Behaviour: Methods and Ethics. *Anim. Behav.* **42**:1007–1014. [9]

D'Evelyn, S. M., J. Jung, E. Alvarado, et al. 2022. Wildfire, Smoke Exposure, Human Health, and Environmental Justice Need to Be Integrated into Forest Restoration and Management. *Curr. Environ. Health Rep.* **3**:366–385. [12]

Dahlgren, G., and M. Whitehead. 1991. Policies and Strategies to Promote Social Equity in Health: Background Document to WHO Strategy Paper for Europe. Stockholm: Institute for Future Studies. [7]

Dallaqua, F. B. J. R., Á. L. Fazenda, and F. A. Faria. 2021. Foresteyes Project: Conception, Enhancements, and Challenges. *Future Gener. Comput. Syst.* **124**:422–435. [3]

Daniels, K. M., L. H. Schinasi, A. H. Auchincloss, C. B. Forrest, and A. V. Diez Roux. 2021. The Built and Social Neighborhood Environment and Child Obesity: A Systematic Review of Longitudinal Studies. *Prev. Med.* **153**:106790. [4]

Darley, J. M., and B. Latane. 1968. Bystander Intervention in Emergencies: Diffusion of Responsibility. *J. Pers. Soc. Psychol.* **8**:377–383. [9]

Darwin, C. 1871. The Descent of Man, and Selection in Relation to Sex. London: John Murray. [9]

Dassonville, L., F. Vauglin, A. Jakobsson, and C. Luzet. 2002. Quality Management, Data Quality and Users, Metadata for Geographical Information. In: Spatial Data Quality, ed. W. Shi et al., pp. 214–227. London: CRC Press. [6]

Dávid-Barrett, T. 2019. Network Effects of Demographic Transition. *Sci. Rep.* **9**:2361. [4]

———. 2020. Herding Friends in Similarity-Based Architecture of Social Networks. *Sci. Rep.* **10**:4859. [4]

———. 2022a. Kinship Is a Network Tracking Social Technology, Not an Evolutionary Phenomenon. *arXiv* **Mar. 2022**:2204.02336v02331. [4]

———. 2022b. World-Wide Evidence for Gender Difference in Sociality. *arXiv* **Mar. 2022**:2203.02964. [4]

Dávid-Barrett, T., I. Behncke Izquierdo, J. Carney, et al. 2016a. Life Course Similarities on Social Networking Sites. *Adv. Life Course Res.* **30**:84–89. [4]

Dávid-Barrett, T., and R. I. M. Dunbar. 2012. Cooperation, Behavioural Synchrony and Status in Social Networks. *Journal of Theoretical Biology* **308**:88–95. [4]

———. 2013. Processing Power Limits Social Group Size: Computational Evidence for the Cognitive Costs of Sociality. *Proc. R. Soc. B* **280**:20131151. [4]

Dávid-Barrett, T., J. Kertesz, A. Rotkirch, et al. 2016b. Communication with Family and Friends across the Life Course. *PLOS ONE* **11**:e0165687. [4]

Dávid-Barrett, T., A. Rotkirch, J. Carney, et al. 2015. Women Favour Dyadic Relationships, but Men Prefer Clubs: Cross-Cultural Evidence from Social Networking. *PLOS ONE* **10**:e0118329. [4]

Davies, I. P., R. D. Haugo, J. C. Robertson, and P. S. Levin. 2018. The Unequal Vulnerability of Communities of Color to Wildfire. *PLOS ONE* **13**:e0205825. [12]

Davis, C. A., O. Varol, E. Ferrara, A. Flammini, and F. Menczer. 2016. BotOrNot: A System to Evaluate Social Bots. In: Proceedings of the 25th International Conference Companion on World Wide Web, pp. 273–274. Montréal: Intl. World Wide Web Conferences Steering Committee. [4]

Davoudi, A., A. Z. Klein, A. Sarker, and G. Gonzalez-Hernandez. 2020. Towards Automatic Bot Detection in Twitter for Health-Related Tasks. In: AMIA Summits on Translational Science Proc., pp. 136–141. Rockville, MD: AMIA. [4]

Dawkins, M. S. 2007. Observing Animal Behaviour: Design and Analysis of Quantitative Data. New York: Oxford Univ. Press. [9]

de Bruin, S., A. Bregt, and M. van de Ven. 2001. Assessing Fitness for Use: The Expected Value of Spatial Data Sets. *Int. J. Geographic. Inf. Sci.* **15**:457–471. [5]

De Groot, A. D., and J. A. A. Spiekerman. 1969. Methodology: Foundations of Inference and Research in the Behavioral Sciences. Berlin: De Gruyter Mouton. [12]

de Macedo Oliveira, A. A. A., and R. Hirata Jr. 2021. INACITY: INvestigate and Analyze a CITY. *SoftwareX* **15**:100777. [3]

de Quadros, J. A. 2020. Determinism and Possibilism: A Critical Epistemological Analysis: Independently Published. [8]

De Veer, D., A. Drouin, J. Fischer, et al. 2022. How Do Schoolchildren Perceive Litter? Overlooked in Urban but Not in Natural Environments. *J. Environ. Psychol.* **81**:101781. [7]

de Vries, S., S. van Dillen, P. Groenewegen, and P. Spreeuwenberg. 2013. Streetscape Greenery and Health: Stress, Social Cohesion and Physical Activity as Mediators. *Soc. Sci. Med.* **94**:26–33. [8]

de Waal, F. B. M. 1989. Peacemaking among Primates. Cambridge, MA: Harvard Univ. Press. [9]

———. 2000. Primates: A Natural Heritage of Conflict Resolution. *Science* **289**:586–590. [9]

de Waal, F. B. M., and S. D. Preston. 2017. Mammalian Empathy: Behavioural Manifestations and Neural Basis. *Nat. Rev. Neurosci.* **18**:498–509. [9]

de Waal, F. B. M., and A. van Roosmalen. 1979. Reconciliation and Consolation among Chimpanzees. *Behav. Ecol. Sociobiol.* **5**:55–66. [9]

Dean, N. 2022. Tracking COVID-19 Infections: Time for Change. *Nature* **602**:185. [12]

Deville Cavellin, L., S. Weichenthal, R. Tack, et al. 2016. Investigating the Use of Portable Air Pollution Sensors to Capture the Spatial Variability of Traffic-Related Air Pollution. *Environ. Sci. Technol.* **50**:313–320. [7]

Devlin, J., M.-W. Chang, K. Lee, K. Toutanova, and Google Language A. I. 2019. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. In: Proc. of NAACL-HLT, pp. 4171–4186. Stroudsburg, PA: ACL. [10]

Dias, M., R. Rocha, and R. R. Soares. 2023. Down the River: Glyphosate Use in Agriculture and Birth Outcomes of Surrounding Populations. *Rev. Econ. Stud.* **90**:2943–2981. [3]

Dibble, J., A. Prelorendjos, O. Romice, et al. 2017. On the Origin of Spaces: Morphometric Foundations of Urban Form Evolution. *Environ. Plann. B Urban Anal. City Sci.* **46**:707–730. [8]

Dickert, N., and J. Sugarman. 2005. Ethical Goals of Community Consultation in Research. *Am. J. Public Health* **95**:1123–1127. [5]

Diderichsen, F., T. Evans, and M. Whitehead. 2001. The Social Basis of Disparities in Health. In: Challenging Inequities in Health: From Ethics to Action, ed. M. Whitehead et al., pp. 12–23. Oxford: Oxford Univ. Press. [1]

Dmowska, A., and T. F. Stepinski. 2018. Spatial Approach to Analyzing Dynamics of Racial Diversity in Large US Cities: 1990–2000–2010. *Comput. Environ. Urban Syst.* **68**:89–96. [8]

———. 2019. Imperfect Melting Pot: Analysis of Changes in Diversity and Segregation of US Urban Census Tracts in the Period of 1990–2010. *Comput. Environ. Urban Syst.* **76**:101–109. [8]

Dobbs, R., and J. Remes. 2013. Trends: The Shifting Urban Economic Landscape: What Does It Mean for Cities? World Bank's Sixth Urban Research and Knowledge Symposium. Washington, D.C.: World Bank. [8]

Dohrenwend, B. P., I. Levav, P. E. Shrout, et al. 1992. Socioeconomic Status and Psychiatric Disorders: The Causation-Selection Issue. *Science* **255**:946–952. [4]

Doiron, D., E. Setton, E. Seed, M. Shooshtari, and J. Brook. 2018. The Canadian Urban Environmental Health Research Consortium (CANUE): A National Data Linkage Initiative. *Int. J. Popul. Data Sci.* **3**:114. [3]

Doll, R., R. Peto, J. Boreham, and I. Sutherland. 2004. Mortality in Relation to Smoking: 50 Years' Observations on Male British Doctors. *Br. Med. J.* **328**:1519. [11]

Dong, X., A. J. Morales, E. Jahani, et al. 2020. Segregated Interactions in Urban and Online Space. *EPJ Data Sci.* **9**:20. [4]

Dong, X., Y. Suhara, B. Bozkaya, et al. 2017a. Social Bridges in Urban Purchase Behavior. *ACM Trans. Intell. Syst. Technol.* **9**:1–29. [4, 8]

Dong, Y., R. A. Johnson, J. Xu, and N. V. Chawla. 2017b. Structural Diversity and Homophily: A Study across More Than One Hundred Big Networks. In: KDD '17: The 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 807–816. New York: ACM. [4]

Dong, Y., Y. Yang, J. Tang, Y. Yang, and N. V. Chawla. 2014. Inferring User Demographics and Social Strategies in Mobile Social Networks. In: KDD '14: Proc. of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 15–24. New York: ACM. [4]

Dorn, H., T. Törnros, and A. Zipf. 2015. Quality Evaluation of VGI Using Authoritative Data: A Comparison with Land Use Data in Southern Germany. *ISPRS Int. J. Geo-Inf.* **4**:1657–1671. [7]

Dredze, M., M. J. Paul, S. Bergsma, and H. Tran. 2013. Carmen: A Twitter Geolocation System with Applications to Public Health. In: Expanding the Boundaries of Health Informatics Using Artificial Intelligence (Papers from the 2013 AAAI Workshop), pp. 20–24. Palo Alto: AAAI Press. [10]

Drouin, M., D. Miller, S. M. J. Wehle, and E. Hernandez. 2016. Why Do People Lie Online? "Because Everyone Lies on the Internet". *Comput. Hum. Behav.* **64**:134–142. [4]

Dubois, H., and A. Ludwinek. 2014. Quality of Life in Urban and Rural Europe. Luxembourg: Publications Office of the EU. [8]

Dudo, A., and J. C. Besley. 2016. Scientists' Prioritization of Communication Objectives for Public Engagement. *PLOS ONE* **11**:e0148867. [3]

Dufva, Y. E., H. Westman, U. Khilbom, P. F. Sullivan, and V. Johansson. 2021. Swedish Large-Scale Schizophrenia Study: Why Do Patients and Healthy Controls Participate? *Schizophr. Res.* **228**:360–366. [11]

Dunbar, R. I. M. 1998. Grooming, Gossip, and the Evolution of Language. Cambridge, MA: Harvard Univ. Press. [4]

Dunbar, R. I. M., and S. Shultz. 2007. Evolution in the Social Brain. *Science* **317**:1344–1347. [4]

Dunbar, R. I. M., and M. Spoors. 1995. Social Networks, Support Cliques, and Kinship. *Hum. Nat.* **6**:273–290. [4]

Durkheim, E. 1897. Le Suicide: Étude Sociologique. Paris: Presses Universitaires de France. [4]

Dusetzina, S. B., S. Tyree, A.-M. Meyer, et al. 2014. Linking Data for Health Services Research: A Framework and Instructional Guide. AHRQ Methods for Effective Health Care. Rockville, MD: Agency for Healthcare Research and Quality. [12]

Dutton, K. 2021. Black-and-White Thinking: The Burden of a Binary Brain in a Complex World. London: Bantam Press. [4]

Easley, D., and J. Kleinberg, eds. 2010. Networks, Crowds, and Markets: Reasoning About a Highly Connected World. Cambridge: Cambridge Univ. Press. [4]

Economist. 2017. The World's Most Valuable Resource Is No Longer Oil, but Data. *The Economist* May 6, 2017. [5]

Egli, V., C. Zinn, L. Mackay, et al. 2019. Viewing Obesogenic Advertising in Children's Neighbourhoods Using Google Street View. *Geographic. Res.* **57**:84–97. [7]

Eibl-Eibesfeldt, I. 1989. Human Ethology. Piscataway: Transaction Publishers. [2, 9]

EIT Health Scandinavia. 2022. Access to Registers in Denmark [Internet]. https://www.eithealth-scandinavia.eu/biobanksregisters/access/registers-denmark/ (accessed Nov. 14, 2022). [11]

EIU. 2022. The Global Liveability Index. https://www.eiu.com/n/campaigns/global-liveability-index-2022 (accessed Jan. 23, 2024). [8]

Ejbye-Ernst, P. 2022. Does Third-Party Intervention Matter? A Video-Based Analysis of the Effect of Third-Party Intervention on the Continuation of Interpersonal Conflict Behaviour. *Br. J. Criminol.* **azab121**:78–96. [9]

Ejbye-Ernst, P., M. R. Lindegaard, and W. Bernasco. 2021. How to Stop a Fight: A Qualitative Video Analysis of How Third-Parties De-Escalate Real-Life Interpersonal Conflicts in Public. *Psychol. Violence* **12**:84–94. [9]

Ekbia, H., M. Mattioli, I. Kouper, et al. 2015. Big Data, Bigger Dilemmas: A Critical Review. *J. Assoc. Inform. Sci. Technol.* **66**:1523–1545. [12]

Ekbom, A. 2011. The Swedish Multi-Generation Register. *Methods Mol. Biol.* **675**:215–220. [11]

Elek, P., D. Kim, and D. P. A. A. Gideon. 2020. Limits of Space Syntax for Urban Design: Axiality, Scale and Sinuosity. *Environment and Planning B* **47**:508–522. [8]

Elgar, F. J., T. K. Pfortner, I. Moor, et al. 2015. Socioeconomic Inequalities in Adolescent Health 2002-2010: A Time-Series Analysis of 34 Countries Participating in the Health Behaviour in School-Aged Children Study. *Lancet* **385**:2088–2095. [1]

Elzeni, M., A. Elmokadem, and N. Badawy. 2022. Classification of Urban Morphology Indicators Towards Urban Generation. *Port-Said Engineering Research Journal* **26**:43–56. [8]

Emmanuel, R., and E. Krüger. 2012. Urban Heat Island and Its Impact on Climate Change Resilience in a Shrinking City: The Case of Glasgow, UK. *Build. Environ.* **53**:137–149. [7]

Eva, G., G. Liese, B. Stephanie, et al. 2022. Position Paper on Management of Personal Data in Environment and Health Research in Europe. *Environ. Int.* **165**:107334. [3]

Eysenbach, G. 2009. Infodemiology and Infoveillance: Framework for an Emerging Set of Public Health Informatics Methods to Analyze Search, Communication and Publication Behavior on the Internet. *J. Med. Internet Res.* **11**:e11. [10]

Fair, L. 2022. FTC Says Data Broker Sold Consumers' Precise Geolocation, Including Presence at Sensitive Healthcare Facilities. FTC Business Blog. https://www.ftc.gov/business-guidance/blog/2022/08/ftc-says-data-broker-sold-consumers-precise-geolocation-including-presence-sensitive-healthcare (accessed Nov. 1, 2022). [3]

Fairchild, A. L. 2015. The Right to Know, the Right to Be Counted, the Right to Resist: Cancer, AIDS, and the Politics of Privacy and Surveillance in Post-War America. *J. Med. Law Ethics* **3**:45–64. [5]

Fan, C., Y. Jiang, and A. Mostafavi. 2020. Social Sensing in Disaster City Digital Twin: Integrated Textual–Visual–Geo Framework for Situational Awareness during Built Environment Disruptions. *J. Manage. Eng.* **36**:me.1943–5479.0000745. [7]

Fan, R., O. Varol, A. Varamesh, et al. 2018. The Minute-Scale Dynamics of Online Emotions Reveal the Effects of Affect Labeling. *Nat. Hum. Behav.* **3**:92–100. [10]

Fani, L., M. K. Georgakis, M. A. Ikram, et al. 2021. Circulating Biomarkers of Immunity and Inflammation, Risk of Alzheimer's Disease, and Hippocampal Volume: A Mendelian Randomization Study. *Transl. Psychiatry* **11**:291. [1]

Fauzie, A. K. 2015. Impacts of Urbanization on Mental Health and Problem Behaviour. *Int. J. Multidiscip. Res. Dev.* **2**:87–89. [8]

Fayyad, U., G. Piatetsky-Shapiro, and P. Smyth. 1996. From Data Mining to Knowledge Discovery in Databases. *AI Mag.* **17**:37. [12]

Federal Trade Commission. 2014. Data Brokers: A Call for Transparency and Accountability. https://www.ftc.gov/system/files/documents/reports/data-brokers-call-transparency-accountability-report-federal-trade-commission-may-2014/140527databrokerreport.pdf (accessed Oct. 7, 2022). [5]

Feest, U. 2016. The Experimenters' Regress Reconsidered: Replication, Tacit Knowledge, and the Dynamics of Knowledge Generation. *Stud. Hist. Phil. Sci. A* **58**:34–45. [5]

Felson, R. B. 1982. Impression Management and the Escalation of Aggression and Violence. *Soc. Psychol. Q.* **45**:245–254. [9]

Few, S. 2019. The Data Loom: Weaving Understanding by Thinking Critically and Scientifically with Data. El Dorado Hills, CA: Analytics Press. [3]

Findata. The Finnish Health and Social Data Permit Authority. https://findata.fi/en/what-is-findata/ (accessed Nov. 9, 2022). [11]

Fischer, P., J. I. Krueger, T. Greitemeyer, et al. 2011. The Bystander-Effect: A Meta-Analytic Review on Bystander Intervention in Dangerous and Non-Dangerous Emergencies. *Psychol. Bull.* **137**:517–537. [9]

Fisher, J., E. Rankin, K. Irvine, and M. Goddard. 2022. Can Biodiverse Streetscapes Mitigate the Effects of Noise and Air Pollution on Human Wellbeing? *Environ. Res.* **212**:113154. [8]

Flack, J. C. 2017. Coarse-Graining as a Downward Causation Mechanism. *Philos. Trans. A Math. Phys. Eng. Sci.* **375**:20160338. [2]

Fleischmann, M., O. Romice, and S. Porta. 2020. Measuring Urban Form: Overcoming Terminological Inconsistencies for a Quantitative and Comprehensive Morphologic Analysis of Cities. *Environ. Plann. B Urban Anal. City Sci.* **48**:2133–2150. [8]

Fleitas Alfonzo, L., T. King, E. You, et al. 2022. Theoretical Explanations for Socioeconomic Inequalities in Multimorbidity: A Scoping Review. *BMJ Open* **12**:e055264. [4]

Flores, L., and S. D. Young. 2021. Regional Variation in Discussion of Opioids on Social Media. *J. Addict. Dis.* **39**:316–321. [10]

Foster, N. 2020. The Pandemic Will Accelerate the Evolution of Our Cities. https://www.theguardian.com/commentisfree/2020/sep/24/pandemic-accelerate-evolution-cities-covid-19-norman-foster (accessed Jan. 23, 2024). [8]

Fowden, A. L., O. R. Vaughan, A. J. Murray, and A. J. Forhead. 2022. Metabolic Consequences of Glucocorticoid Exposure before Birth. *Nutrients* **14**:2072–6643. [1]

Fox, C., A. Levitin, and T. Redman. 1994. The Notion of Data and Its Quality Dimensions. *Inf. Process. Manag.* **30**:9–19. [5]

Franck, G. 2019. The Economy of Attention. *J. Sociol.* **55**:8–19. [4]

Frank, L. D., N. Iroz-Elardo, K. E. MacLeod, and A. Hong. 2019. Pathways from Built Environment to Health: A Conceptual Framework Linking Behavior and Exposure-Based Impacts. *J. Transp. Health* **12**:319–335. [7]

Frick, S. A., and A. Rodríguez-Pose. 2018. Big or Small Cities? On City Size and Economic Growth. *Growth Change* **49**:4–32. [8]

Friedman, B., and D. Hendry. 2019. Value Sensitive Design: Shaping Technology with Moral Imagination. Cambridge, MA: MIT Press. [5]

Friis, C. B. 2022. Ticket Inspection in Action: Managing Impressions, Status, and Emotions in Contested Everyday Encounters. PhD dissertation, Univ. of Copenhagen, Copenhagen. [9]

Friis, C. B., L. S. Liebst, R. Philpot, and M. R. Lindegaard. 2020. Ticket Inspectors in Action: Body-Worn Camera Analysis of Aggressive and Nonaggressive Passenger Encounters. *Psychol. Violence* **10**:483–492. [9]

Fry, D., S. J. Mooney, D. A. Rodríguez, W. T. Caiaffa, and G. S. Lovasi. 2020. Assessing Google Street View Image Availability in Latin American Cities. *J. Urban Health* **97**:1–9. [3]

Fujita, M., P. R. Krugman, and A. Venables. 2001. The Spatial Economy: Cities, Regions, and International Trade. Cambridge, MA: MIT Press. [8]

Fujita, M., and T. Mori. 1996. The Role of Ports in the Making of Major Cities: Self-Agglomeration and Hub-Effect. *Journal of Development Economics* **49**:93–120. [8]

Fuller, D., and K. G. Stanley. 2019. The Future of Activity Space and Health Research. *Health Place* **58**:102131. [7]

Gagolewski, M. 2015. Data Fusion: Theory, Methods, and Applications. Institute of Computer Science. Warsaw: Polish Academy of Sciences. [5]

Galbete, C., M. Nicolaou, K. A. Meeks, et al. 2017. Food Consumption, Nutrient Intake, and Dietary Patterns in Ghanaian Migrants in Europe and Their Compatriots in Ghana. *Food Nutr. Res.* **61**:1341809. [4]

Gallo, P., and H. Kettani. 2020. On Privacy Issues with Google Street View. *South Dakota Law Rev.* **65**:608. [3]

Garcia-Castellanos, D., and U. Lombardo. 2007. Poles of Inaccessibility: A Calculation Algorithm for the Remotest Places on Earth. *Scott. Geographic. J.* **123**:227–233. [6]

Garcia Coll, C., and A. K. Marks. 2011. The Immigrant Paradox in Children and Adolescents: Is Becoming American a Risk Factor? Washington, D.C.: American Psychological Association. [2]

Garg, K., H. R. Jayanetti, S. Alam, M. C. Weigle, and M. L. Nelson. 2021. Replaying Archived Twitter: When Your Bird Is Broken, Will It Bring You Down? In: 2021 ACM/IEEE Joint Conference on Digital Libraries (JCDL), pp. 160–169. New York: ACM. [4]

———. 2023. Challenges in Replaying Archived Twitter Pages. *Int. J. Digit. Libr.* **2023**:10.1007/s00799-00023-00379-w. [4]

Gaubatz, K. T. 2015. A Survivor's Guide to R: An Introduction for the Uninitiated and the Unnerved. https://methods.sagepub.com/book/a-survivors-guide-to-r. (accessed Nov. 7, 2022). [5]

Geertz, C. 1973. The Interpretation of Cultures: Selected Essays. New York: Basic Books. [9]

Gelman, A. 2020. Assessing Evidence vs. Truth in the Coronavirus Pandemic. *Chance* **33**:58–60. [12]

Gelman, A., and B. Carpenter. 2020. Bayesian Analysis of Tests with Unknown Specificity and Sensitivity. *J. R. Stat. Soc. Ser. C* **69**:1269–1283. [12]

Gelman, A., and C. Hennig. 2017. Beyond Subjective and Objective in Statistics. *J. R. Stat. Soc. Ser. A* **180**:967–1033. [12]

Gene Ontology Consortium. 2018. The Gene Ontology Resource: 20 Years and Still Going Strong. *Nucleic Acids Res.* **47**:D330–D338. [5]

Gerrard, G., and R. Thompson. 2011. Two Million Cameras in the UK. *CCTV Image* **42**:9–12. [9]

Gerts, D., C. D. Shelley, N. Parikh, et al. 2021. "Thought I'd Share First" and Other Conspiracy Theory Tweets from the COVID-19 Infodemic: Exploratory Study. *JMIR Public Health Surveill.* **7**:e26527. [10]

Getoor, L., and A. Machanavajjhala. 2012. Entity Resolution: Theory, Practice and Open Challenges. *Proc. VLDB Endow.* **5**:2018–2019. [12]

Giannouchos, T. V., A. O. Ferdinand, G. Ilangovan, et al. 2021. Identifying and Prioritizing Benefits and Risks of Using Privacy-Enhancing Software through Participatory Design: A Nominal Group Technique Study with Patients Living with Chronic Conditions. *J. Am. Med. Inform. Assoc.* **28**:1746–1745. [12]

Gilbert, R., R. Lafferty, G. Hagger-Johnson, et al. 2017. Guild: Guidance for Information about Linking Data Sets. *J. Public Health* **40**:191–198. [12]

Gill, I. S., and C.-C. Goh. 2010. Scale Economies and Cities. *World Bank Res. Observ.* **25**:235–262. [8]

Goel, R., L. M. T. Garcia, A. Goodman, et al. 2018. Estimating City-Level Travel Patterns Using Street Imagery: A Case Study of Using Google Street View in Britain. *PLOS ONE* **13**:e0196521. [7]

Goffman, E. 1971. Relations in Public: Microstudies of the Public Order. New York: Basic Books. [9]

Golinelli, D., E. Boetto, G. Carullo, et al. 2020. Adoption of Digital Technologies in Health Care During the COVID-19 Pandemic: Systematic Review of Early Scientific Literature. *J. Med. Internet Res.* **22**:e22280. [10]

Gong, P., H. Liu, M. Zhang, et al. 2019. Stable Classification with Limited Sample: Transferring a 30-M Resolution Sample Set Collected in 2015 to Mapping 10-M Resolution Global Land Cover in 2017. *Sci. Bull.* **64**:370–373. [7]

Gong, P., J. Wang, L. Yu, et al. 2013. Finer Resolution Observation and Monitoring of Global Land Cover: First Mapping Results with Landsat TM and ETM+ Data. *Int. J. Remote Sens.* **34**:2607–2654. [7]

González-Padilla, D. A., and L. Tortolero-Blanco. 2020. Social Media Influence in the COVID-19 Pandemic. *Int. Braz. J. Urol.* **46**:120–124. [10]

Goodchild, M. F. 2007. Citizens as Sensors: The World of Volunteered Geography. *GeoJournal* **69**:211–221. [7]

Goodman, A., A. Pepe, A. W. Blocker, et al. 2014. Ten Simple Rules for the Care and Feeding of Scientific Data. *PLoS Comput. Biol.* **10**:1–5. [12]

Goold, B. J. 2006. Open to All? Regulating Open Street CCTV and the Case for "Symmetrical Surveillance". *Crim. Justice Ethics* **25**:3–17. [9]

Gorelick, N., M. Hancher, M. Dixon, et al. 2017. Google Earth Engine: Planetary-Scale Geospatial Analysis for Everyone. *Remote Sens. Environ.* **202**:18–27. [7]

Gorelik, A. 2019. The Enterprise Big Data Lake: Delivering the Promise of Big Data and Data Science. Sebastopol, CA: O'Reilly Media. [3]

GOV.UK. 2012a. Definition of Policing by Consent. https://www.gov.uk/government/publications/policing-by-consent/definition-of-policing-by-consent (accessed Dec. 30, 2023). [3]

———. 2012b. Protection of Freedoms Act 2012. https://www.legislation.gov.uk/ukpga/2012/9/contents/enacted (accessed Dec. 30, 2023). [3]

Granovetter, M. S. 1973. The Strength of Weak Ties. *Am. J. Sociol.* **78**:1360–1380. [4]

Grasby, K. L., N. Jahanshad, J. N. Painter, et al. 2020. The Genetic Architecture of the Human Cerebral Cortex. *Science* **367**:eaay6690. [1]

Graves, R. L., J. Perrone, M. A. Al-Garadi, et al. 2022. Thematic Analysis of reddit Content About Buprenorphine-Naloxone Using Manual Annotation and Natural Language Processing Techniques. *J. Addict. Med.* **16**:454–460. [10]

Greenland, S., and R. Neutra. 1980. Control of Confounding in the Assessment of Medical Technology. *Int. J. Epidemiol.* **9**:361–367. [11]

Grigoriev, P., and M. Pechholdová. 2017. Health Convergence between East and West Germany as Reflected in Long-Term Cause-Specific Mortality Trends: To What Extent Was It Due to Reunification? *Eur. J. Popul.* **33**:701–731. [4]

Groen, J. A. 2012. Sources of Error in Survey and Administrative Data: The Importance of Reporting Procedures. *J. Off. Stat.* **28**:173–198. [7]

Gröger, G., and B. George. 2022. Geometry and Topology. In: Springer Handbook of Geographic Information, ed. W. Kresse and D. Danko, pp. 297–313, Springer Handbooks. Cham: Springer. [6]

Grolemund, G., and H. Wickham. 2014. A Cognitive Interpretation of Data Analysis. *Int. Stat. Rev.* **82**:184–204. [12]

Gruber, T., and Z. Clay. 2016. A Comparison between Bonobos and Chimpanzees: A Review and Update. *Evol. Anthropol.* **25**:239–252. [4]

Guo, D., and E. Onstein. 2020. State-of-the-Art Geospatial Information Processing in NoSQL Databases. *ISPRS Int. J. Geo-Inf.* **9**:331. [6]

Guo, J.-W., S. M. Sisler, C.-Y. Wang, and A. S. Wallace. 2021. Exploring Experiences of COVID-19-Positive Individuals from Social Media Posts. *Int. J. Nurs. Pract.* **27**:e12986. [10]

Gupta, A., and R. Katarya. 2020. Social Media Based Surveillance Systems for Healthcare Using Machine Learning: A Systematic Review. *J. Biomed. Inform.* **108**:103500. [10]

Guttman, A. 1984. R-Trees. *ACM SIGMOD Record* **14**:47–57. [6]

Haas, L. M., E. T. Lin, and M. A. Roth. 2002. Data Integration through Database Federation. *IBM Syst. J.* **41**:578–596. [11]

Haklay, M. 2010. How Good Is Volunteered Geographical Information? A Comparative Study of Openstreetmap and Ordnance Survey Datasets. *Environ. Plann. B Plann. Des.* **37**:682–703. [7]

Hall, R., and S. E. Fienberg. 2010. Privacy-Preserving Record Linkage. In: Privacy in Statistical Databases, ed. J. Domingo-Ferrer and E. Magkos, pp. 269–283. Heidelberg: Springer. [12]

Hallinan, D., A. Bernier, A. Cambon-Thomsen, et al. 2021. International Transfers of Personal Data for Health Research Following Schrems II: A Problem in Need of a Solution. *Eur. J. Hum. Genet.* **29**:1502–1509. [5, 11]

Haneef, R., M. Tijhuis, R. Thiébaut, et al. 2022. Methodological Guidelines to Estimate Population-Based Health Indicators Using Linked Data and/or Machine Learning Techniques. *Arch. Public Health* **80**:9. [12]

Haneuse, S., and S. Bartell. 2011. Designs for the Combination of Group- and Individual-Level Data. *Epidemiology* **22**:382–389. [5]

Hang, J., M. Sandberg, and Y. Li. 2009. Effect of Urban Morphology on Wind Condition in Idealized City Models. *Atmos. Environ.* **43**:869–878. [8]

Hansen, M. C., P. V. Potapov, R. Moore, et al. 2013. High-Resolution Global Maps of 21st-Century Forest Cover Change. *Science* **342**:850–853. [7]

Hardjono, T., D. L. Shrier, and A. S. Pentland. 2019. Trusted Data, revised and Expanded edition. A New Framework for Identity and Data Sharing. Cambridge, MA: MIT Press. [8]

Hargittai, E. 2020. Potential Biases in Big Data: Omitted Voices on Social Media. *Soc. Sci. Comput. Rev.* **38**:10–24. [4]

Harrigian, K. 2018. Geocoding without Geotags: A Text-Based Approach for reddit. In: Proc. of the 2018 EMNLP Workshop W-Nut: The 4th Workshop on Noisy User-Generated Text, pp. 17–27. Stroudsburg, PA: ACL. [4, 10]

Harron, K., A. Wade, R. Gilbert, B. Muller-Pebody, and H. Goldstein. 2014. Evaluating Bias Due to Data Linkage Error in Electronic Healthcare Records. *BMC Med. Res. Methodol.* **14**:36. [12]

Harron, K. L., J. C. Doidge, H. E. Knight, et al. 2017. A Guide to Evaluating Linkage Quality for the Analysis of Linked Data. *Int. J. Epidemiol.* **46**:1699–1710. [12]

Hart, E. M., P. Barmby, D. LeBauer, et al. 2016. Ten Simple Rules for Digital Data Storage. *PLoS Comput. Biol.* **12**:e1005097. [3]

Hashimov, A., I. Pathiddinov, M. Makhmutova, et al. 2013. Urbanization in Central Asia: Challenges, Issues and Prospects. Tashkent, Uzbekistan: Center for Economic Research. [8]

Heaviside, C., X. M. Cai, and S. Vardoulakis. 2015. The Effects of Horizontal Advection on the Urban Heat Island in Birmingham and the West Midlands, United Kingdom during a Heatwave. *Q. J. R. Meteorol. Soc.* **141**:qj.2452. [7]

Heinskou, M. B., and L. S. Liebst. 2016. On the Elementary Neural Forms of Micro-Interactional Rituals: Integrating Autonomic Nervous System Functioning into Interaction Ritual Theory. *Sociol. Forum* **31**:354–376. [9]

Heisenberg, W. 1958. The Revolution in Modern Science. Amherst, N.Y.: Prometheus Books. [2]

Helbich, M., Y. Yao, Y. Liu, et al. 2019. Using Deep Learning to Examine Street View Green and Blue Spaces and Their Associations with Geriatric Depression in Beijing, China. *Environ. Int.* **126**:107–117. [7]

Helweg-Larsen, K. 2011. The Danish Register of Causes of Death. *Scand. J. Public Health* **39**:26–29. [11]

Hemati, M., M. Hasanlou, M. Mahdianpari, and F. Mohammadimanesh. 2021. A Systematic Review of Landsat Data for Change Detection Applications: 50 Years of Monitoring the Earth. *Remote Sens.* **13**:2869. [7]

Henrich, J., S. J. Heine, and A. Norenzayan. 2010. The WEIRDest People in the World. *Brain Behav. Sci.* **33**:61–83. [2]

Henssler, J., L. Brandt, M. Müller, et al. 2020. Migration and Schizophrenia: Meta-Analysis and Explanatory Framework. *Eur. Arch. Psychiatry Clin. Neurosci.* **270**:325–335. [4]

Hermosilla, T., J. Palomar, A. Balaguer Beser, J. Balsa Barreiro, and L. A. Ruiz. 2014. Using Street Based Metrics to Characterize Urban Typologies. *Comput. Environ. Urban Syst.* **44**:68–79. [8]

Hern, M. 2017. What a City Is For: Remaking the Politics of Displacement. Cambridge, MA: MIT Press. [8]

Hernández, M. A., and S. J. Stolfo. 1998. Real-World Data Is Dirty: Data Cleansing and the Merge/Purge Problem. *Data Min. Knowl. Discov.* **2**:9–37. [5]

Herring, J., C. Roswell, and D. Danko. 2022. Modeling of Geographic Information. In: Springer Handbook of Geographic Information, ed. W. Kresse and D. Danko, pp. 3–19, Springer Handbooks. Cham: Springer. [6]

Herzog, T. N., F. J. Scheuren, and W. E. Winkler. 2007. Data Quality and Record Linkage Techniques. New York: Springer. [5]

Hickman, C., E. Marks, P. Pihkala, et al. 2021. Climate Anxiety in Children and Young People and Their Beliefs about Government Responses to Climate Change: A Global Survey. *Lancet Planet Health* **5**:e863–e873. [1]

Hicks, D. J. 2021. Open Science, the Replication Crisis, and Environmental Public Health. *Account. Res.*1–29. [5]

Hill, A. B. 1965. The Environment and Disease: Association or Causation? *Proc. R. Soc. Med.* **58**:295–300. [2]

Hill, R. A., and R. I. M. Dunbar. 2003. Social Network Size in Humans. *Hum. Nat.* **14**:53–72. [4]

Hillier, B. 1996. Space Is the Machine. Cambridge: Cambridge Univ. Press. [8]

Hiss, T. 1991. The Experience of Place: A New Way of Looking at and Dealing with Our Radically Changing Cities and Countryside. New York: Knopf Doubleday. [8]

Hoel, E. 2022. Big Data Analytics. In: Springer Handbook of Geographic Information, ed. W. Kresse and D. Danko, pp. 107–118, Springer Handbooks. Cham: Springer. [6]

Höfler, M. 2005. Causal Inference Based on Counterfactuals. *BMC Med. Res. Methodol.* **5**:1–12. [2]

Hollenbaugh, E. E., and A. L. Ferris. 2015. Predictors of Honesty, Intent, and Valence of Facebook Self-Disclosure. *Comput. Hum. Behav.* **50**:456–464. [4]

Hong, S., M. R. Jahng, N. Lee, and K. R. Wise. 2020. Do You Filter Who You Are?: Excessive Self-Presentation, Social Cues, and User Evaluations of Instagram Selfies. *Comput. Hum. Behav.* **104**:106159. [4]

Horning, N. 2019. Remote Sensing. In: Encyclopedia of Ecology, ed. B. B. T. Fath, pp. 404–413. Oxford: Elsevier. [7]

Hossain, L., D. Kam, F. Kong, R. T. Wigand, and T. Bossomaier. 2016. Social Media in Ebola Outbreak. *Epidemiol. Infect.* **144**:2136–2143. [4]

Houlden, V., S. Weich, J. P. de Albuquerque, S. Jarvis, and K. Rees. 2018. The Relationship between Greenspace and the Mental Wellbeing of Adults: A Systematic Review. *PLOS ONE* **13**:e0203000. [7]

Houston, D. 2014. Implications of the Modifiable Areal Unit Problem for Assessing Built Environment Correlates of Moderate and Vigorous Physical Activity. *Appl. Geogr.* **50**:40–47. [7]

Huang, D., A. Brien, L. Omari, et al. 2020. Bus Stops near Schools Advertising Junk Food and Sugary Drinks. *Nutrients* **12**:1192. [7]

Huang, Y., M. Yuan, Y. Sheng, X. Min, and Y. Cao. 2019. Using Geographic Ontologies and Geo-Characterization to Represent Geographic Scenarios. *ISPRS Int. J. Geo-Inf.* **8**:566. [5]

Huber, P. J. 1996. Massive Data Sets Workshop: The Morning After. Massive Data Sets: Proceedings of a Workshop. Washington, D.C.: National Academies Press. [2]

Hui, C. H., and H. C. Triandis. 1985. Measurement in Cross-Cultural Psychology: A Review and Comparison of Strategies. *J. Cross Cult. Psychol.* **16**:131–152. [2]

Hulkower, R., M. Penn, and C. Schmit. 2020. Privacy and Confidentiality of Public Health Information. In: Public Health Informatics and Information Systems, ed. J. A. Magnuson and B. E. Dixon, pp. 147–166. Cham: Springer. [5]

IARC Working Group on the Evaluation of Carcinigenic Risks to Humans. 2013. Non-Ionizing Radiation, Part 2: Radiofrequency Electromagnetic Fields. Lyon: Intl. Agency for Research on Cancer. [11]

Ibrahim, M. R., J. Haworth, and T. Cheng. 2021. URBAN-i: From Urban Scenes to Mapping Slums, Transport Modes, and Pedestrians in Cities Using Deep Learning and Computer Vision. *Environ. Plann. B Urban Anal. City Sci.* **48**:76–93. [7]

Ienca, M., A. Ferretti, S. Hurst, et al. 2018. Considerations for Ethics Review of Big Data Health Research: A Scoping Review. *PLOS ONE* **13**:e0204937. [5]

Imran, M., S. Khan, A. A. Nassani, et al. 2023. Access to Sustainable Healthcare Infrastructure: A Review of Industrial Emissions, Coal Fires, and Particulate Matter. *Environ. Sci. Pollut. Res. Int.* **30**:69080–69095. [1]

Ioannidis, J. P. A. 2005. Why Most Published Research Findings Are False. *PLOS Med.* **2**:e124. [5, 11]

Işıkdağ, Ü. 2020. An IoT Architecture for Facilitating Integration of Geoinformation. *Int. J. Engineer. Geosci.* **5**:15–25. [6]

Jacobs, J. 1961. The Death and Life of Great American Cities. New York: Random House. [8]

Jain, M., J. C. Van Gemert, and C. G. Snoek. 2015. What Do 15,000 Object Categories Tell Us about Classifying and Localizing Actions? In: Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 46–55. Piscataway: IEEE. [9]

Jalal, H., J. M. Buchanich, M. S. Roberts, et al. 2018. Changing Dynamics of the Drug Overdose Epidemic in the United States from 1979 through 2016. *Science* **361**:eaau1184. [10]

Jijelava, D., and F. Vanclay. 2017. Legitimacy, Credibility and Trust as the Key Components of a Social Licence to Operate: An Analysis of BP's Projects in Georgia. *J. Clean. Prod.* **140**:1077–1086. [5]

Jo, H.-H., J. Saramäki, R. I. M. Dunbar, and K. Kaski. 2014. Spatial Patterns of Close Relationships across the Lifespan. *Sci. Rep.* **4**:6988. [4, 5]

Jones, L. K., B. M. Jennings, M. K. Higgins, and F. B. M. De Waal. 2018. Ethological Observations of Social Behavior in the Operating Room. *PNAS* **115**:7575–7580. [9]

Kain, J.-H., M. Adelfio, J. Stenberg, and L. Thuvander. 2022. Towards a Systemic Understanding of Compact City Qualities. *J. Urban Des.* **27**:130–147. [8]

Kam, C. C.-S., and M. H. Bond. 2009. Emotional Reactions of Anger and Shame to the Norm Violation Characterizing Episodes of Interpersonal Harm. *Br. J. Soc. Psychol.* **48**:203–219. [4]

Kamdar, M. R., J. D. Fernández, A. Polleres, T. Tudorache, and M. A. Musen. 2019. Enabling Web-Scale Data Integration in Biomedicine through Linked Open Data. *npj Digit. Med.* **2**:90. [5]

Kardan, O., P. Gozdyra, B. Misic, et al. 2015. Neighborhood Greenspace and Health in a Large Urban Center. *Sci. Rep.* **5**:11610. [1]

Karim, M., M. Ramezani, T. Sunbury, R. L. Ohsfeldt, and H.-C. Kum. 2021. VIEW: A Framework for Organization Level Interactive Record Linkage to Support Reproducible Data Science. *arXiv* **2102**:08273. [12]

Kawachi, I., and L. F. Berkman. 2014. Social Capital, Social Cohesion, and Health. In: Social Epidemiology, ed. L. F. Berkman et al., pp. 290–319. New York: Oxford Univ. Press. [4]

Keller, H. 2021. The Myth of Attachment Theory. New York: Routledge. [2]

Keller, H., and K. A. Bard, eds. 2017. Contextualizing Attachment: The Cultural Nature of Attachment. Strüngmann Forum Reports, vol. 22. J. R. Lupp, series ed. Cambridge, MA: MIT Press. [2]

Keller, H., and N. Chaudhary. 2017. Is the Mother Essential for Attachment? Models of Care in Different Cultures. In: Contextualizing Attachment: The Cultural Nature of Attachment, ed. H. Keller and K. H. Bard, pp. 109–138, Strüngmann Forum Reports, vol. 22, J. R. Lupp, series ed. Cambridge, MA: MIT Press. [2]

Kendler, K. S., C. O. Gardner, and C. A. Prescott. 2003. Personality and the Experience of Environmental Adversity. *Psychol. Med.* **33**:1193–1202. [1]

Kerry, C. 2021. The Oracle at Luxembourg: The EU Court of Justice Judges the World on Surveillance and Privacy. Washington, D.C.: Brookings. [5]

Khaled, M., G. W. Corner, A. Morris, et al. 2021. Physiological Linkage in Pregnancy: Couples' Cortisol, Negative Conflict Behavior, and Postpartum Depression. *Biol. Psychol.* **161**:108075. [1]

Khemlani, S. S., A. K. Barbey, and P. N. Johnson-Laird. 2014. Causal Reasoning with Mental Models. *Front. Hum. Neurosci.* **8**:849. [2]

Khraishah, H., B. Alahmad, R. L. Ostergard, Jr., et al. 2022. Climate Change and Cardiovascular Disease: Implications for Global Health. *Nat. Rev. Cardiol.* **19**:798–812. [1]

Kimpton, A., J. Corcoran, and R. Wickes. 2016. Greenspace and Crime: An Analysis of Greenspace Types, Neighboring Composition, and the Temporal Dimensions of Crime. *J. Res. Crime Delinq.* **54**:303–337. [8]

King, G. 2007. An Introduction to the Dataverse Network as an Infrastructure for Data Sharing. *Sociol. Methods Res.* **36**:173–199. [3]

Kısar Koramaz, E. 2014. The Spatial Context of Social Integration. *Social Indicators Research* **119**:49–71. [8]

Kivimaki, M., G. D. Batty, J. Pentti, et al. 2020. Association between Socioeconomic Status and the Development of Mental and Physical Health Conditions in Adulthood: A Multi-Cohort Study. *Lancet Public Health* **5**:e140–e149. [1]

Klau, S., S. Hoffmann, C. J. Patel, J. P. Ioannidis, and A. L. Boulesteix. 2021. Examining the Robustness of Observational Associations to Model, Measurement and Sampling Uncertainty with the Vibration of Effects Framework. *Int. J. Epidemiol.* **50**:266–278. [5]

Klein, M., K. Barg, and M. Kühhirt. 2018. Inequality of Educational Opportunity in East and West Germany: Convergence or Continued Differences. *Sociol. Sci.* **6**:1–26. [4]

Knight, A., S. Sandin, and J. Askling. 2010. Occupational Risk Factors for Wegener's Granulomatosis: A Case-Control Study. *Ann. Rheum. Dis.* **69**:737–740. [11]

Knowles, D. 2023. How Tokyo Became an Anti-Car Paradise. https://heatmap.news/economy/tokyo-anti-car-pedestrian-paradise (accessed Jan. 23, 2024). [8]

Kolbe, T. H. 2009. Representing and Exchanging 3D City Models with CityGML. In: 3D Geo-Information Sciences, ed. J. Lee and S. Zlatanova, pp. 15–31, Lecture Notes in Geoinformation and Cartography. Berlin: Springer. [6]

Kolle, S., B. Hughes, and H. Steele. 2020. Early Embryo-Maternal Communication in the Oviduct: A Review. *Mol. Reprod. Dev.* **87**:650–662. [1]

Kolodny, A., and T. R. Frieden. 2017. Ten Steps the Federal Government Should Take Now to Reverse the Opioid Addiction Epidemic. *JAMA* **318**:1537. [10]

Kon, F., É. C. Ferreira, H. A. de Souza, et al. 2022. Abstracting Mobility Flows from Bike-Sharing Systems. *Public Trans.* **14**:545–581. [3]

Kopec, D. 2012. Environmental Psychology for Design. New York: Fairchild Books. [8]

Korte, C., and N. Ayvalioglu. 1981. Helpfulness in Turkey: Cities, Towns, and Urban Villages. *J. Cross Cult. Psychol.* **12**:123–141. [8]

Korte, C., and N. Kerr. 1975. Response to Altruistic Opportunities in Urban and Nonurban Settings. *J. Soc. Psychol.* **95**:183–184. [8]

Kossinets, G., and D. J. Watts. 2009. Origins of Homophily in an Evolving Social Network. *Am. J. Sociol.* **115**:405–450. [4]

Kotkin, J. 2020. The Coming Age of Dispersion. https://quillette.com/2020/03/25/the-coming-age-of-dispersion/ (accessed Jan. 23, 2024). [8]

Kovacs-Györi, A., A. Ristea, R. Kolcsar, et al. 2018. Beyond Spatial Proximity-Classifying Parks and Their Visitors in London Based on Spatiotemporal and Sentiment Analysis of Twitter Data. *ISPRS Int. J. Geo-Inf.* **7**:378. [7]

Kovanen, L., K. Kaski, J. Kertész, and J. Saramäki. 2013. Temporal Motifs Reveal Homophily, Gender-Specific Patterns, and Group Talk in Call Sequences. *PNAS* **110**:18070–18075. [5]

Kraak, M.-J., and F. Ormeling. 2021. Cartography: Visualization of Geospatial Data. Boca Raton: CRC Press. [6]

Krakauer, D., N. Bertschinger, E. Olbrich, J. C. Flack, and N. Ay. 2020. The Information Theory of Individuality. *Theory Biosci.* **139**:209–223. [2]

Kresse, W. H., and D. Danko, eds. 2022. Springer Handbook of Geographic Information. Springer Handbooks, vol. Cham: Springer. [6]

Kresse, W. H., D. Danko, and K. Fadaie. 2022. Standardization. In: Springer Handbook of Geographic Information, ed. W. Kresse and D. Danko, pp. 383–492, Springer Handbooks. Cham: Springer. [6]

Krieg, S. J., J. J. Schnur, J. D. Marshall, M. M. Schoenbauer, and N. V. Chawla. 2020. Pandemic Pulse: Unraveling and Modeling Social Signals During the COVID-19 Pandemic. *Digit. Gov. Res. Pract.* **2**:1–9. [4]

Krier, R. 1979. Urban Space. New York: Rizzoli Intl. Publ. [8]

Kriesberg, A., and A. Acker. 2022. The Second US Presidential Social Media Transition: How Private Platforms Impact the Digital Preservation of Public Records. *J. Assoc. Inform. Sci. Technol.* **73**:1529–1542. [4]

Kropf, K. 2009. Aspects of Urban Form. *Urban Morphology* **13**:105–120. [8]

———. 2017. The Handbook of Urban Morphology. Chichester: John Wiley & Sons Ltd. [8]

Kryvasheyeu, Y., H. Chen, N. Obradovich, et al. 2016. Rapid Assessment of Disaster Damage Using Social Media Activity. *Sci. Adv.* **2**:e1500779. [4]

Kum, H.-C., and S. Ahalt. 2013. Privacy-by-Design: Understanding Data Access Models for Secondary Data. *AMIA Jt. Summits Transl. Sci. Proc.* **2013**:126–130. [1, 12]

Kum, H.-C., A. Krishnamurthy, A. Machanavajjhala, and S. C. Ahalt. 2014. Social Genome: Putting Big Data to Work for Population Informatics. *Computer* **47**:56–63. [5, 12]

Kum, H.-C., A. Krishnamurthy, A. Machanavajjhala, M. K. Reiter, and S. Ahalt. 2013. Privacy Preserving Interactive Record Linkage (PPIRL). *J. Am. Med. Inform. Assoc.* **21**:212–220. [12]

Kum, H.-C., E. Ragan, A. O. Ferdinand, and C. D. Schmit. 2022. Developing Methods for Record Linkage That Protect Patient Privacy. PCORI Final Research Report. https://www.pcori.org/sites/default/files/Kum404-Final-Research-Report.pdf (accessed Sept. 5, 2023). [12]

Kum, H.-C., E. D. Ragan, G. Ilangovan, et al. 2019. Enhancing Privacy through an Interactive On-Demand Incremental Information Disclosure Interface: Applying Privacy-by-Design to Record Linkage. In: 15th Symposium on Usable Privacy and Security (SOUPS 2019), pp. 175–189. Santa Clara: USENIX Association. [12]

Kumpula, J. M., J.-P. Onnela, J. Saramäki, K. Kaski, and J. Kertész. 2007. Emergence of Communities in Weighted Networks. *Phys. Rev. Lett.* **99**:228701. [5]

Kushida, C. A., D. A. Nichols, R. Jadrnicek, et al. 2012. Strategies for de-Identification and Anonymization of Electronic Health Record Data for Use in Multicenter Research Studies. *Med. Care* **50(Suppl)**:82–101. [5]

Kwan, M. P. 2012. How GIS Can Help Address the Uncertain Geographic Context Problem in Social Science Research. *Ann. GIS* **18**:245–255. [6, 7]

Laakasuo, M., A. Rotkirch, M. van Duijn, et al. 2020. Homophily in Personality Enhances Group Success among Real-Life Friends. *Front. Psychol.* **11**:710. [4]

Laato, S., N. Inaba, and J. Hamari. 2021. Convergence between the Real and the Augmented: Experiences and Perceptions in Location-Based Games. *Telemat. Informat.* **65**:101716. [4]

Labbrook, D. A. 1988. Why Are Crime Rates Higher in Urban Than in Rural Areas? Evidence from Japan. *Aust. N. Z. J. Criminol.* **21**:81–103. [8]

Lagisetty, P. A., R. Ross, A. Bohnert, M. Clay, and D. T. Maust. 2019. Buprenorphine Treatment Divide by Race/Ethnicity and Payment. *JAMA Psychiatry* **76**:979–981. [10]

Lagoze, C. 2001. Keeping Dublin Core Simple: Cross-Domain Discovery or Resource Description? *D-Lib Mag.* **7**:Jan. 2001. [3]

Lagoze, C., D. Krafft, T. Cornwell, et al. 2006. Metadata Aggregation and Automated Digital Libraries: A Retrospective on the NSDL Experience. In: Proc. of the 6th ACM/IEEE-Cs Joint Conference on Digital Libraries, pp. 230–239. New York: ACM. [3]

Lakamana, S., Y.-C. Yang, M. A. Al-Garadi, and A. Sarker. 2022. Tracking the COVID-19 Outbreak in India through Twitter: Opportunities for Social Media Based Global Pandemic Surveillance. *AMIA Annu. Symp. Proc.* **May 23**:313–322. [10]

Laney, D. 2001. 3D Data Management: Controlling Data Volume, Velocity and Variety. *META Group Res. Note* **6**:1. [2]

Lash, T. L. 2022. Getting over TOP. *Epidemiology* **33**:1–6. [5]

Laugesen, K., J. F. Ludvigsson, M. Schmidt, et al. 2021. Nordic Health Registry-Based Research: A Review of Health Care Systems and Key Registries. *Clin. Epidemiol.* **13**:533–554. [11]

Lawlor, D. A., K. Tilling, and G. Davey Smith. 2016. Triangulation in Aetiological Epidemiology. *Int. J. Epidemiol.* **45**:1866–1886. [2]

Lawson, J., M. N. Cabili, G. Kerry, et al. 2021. The Data Use Ontology to Streamline Responsible Access to Human Biomedical Datasets. *Cell Genom.* **1**:100028. [3]

Leck, E. 2006. The Impact of Urban Form on Travel Behavior: A Meta-Analysis. *Berkeley Plan. J.* **19**:37–58. [8]

Lees, L., T. Slater, and E. Wyly. 2007. Gentrification. New York: Routledge. [8]

Legeby, A. 2010. Urban Segregation and Urban Form: From Residential Segregation to Segregation in Public Space. PhD dissertation, KTH Royal Institute of Technology, Stockholm. [8]

Legey, L., M. Ripper, and P. Varaiya. 1973. Effects of Congestion on the Shape of a City. *Journal of Economic Theory* **6**:162–179. [8]

Lehné, R. J., C. Hoselmann, H. Heggemann, H. Budde, and A. Hoppe. 2013. Geological 3D Modelling in the Densely Populated Metropolitan Area Frankfurt/Rhine-Main. *Zeitschrift der Dtsch. Gesellschaft fur Geowissenschaften* **164**:591–603. [3]

Lehner, P. N. 1998. Handbook of Ethological Methods. Cambridge: Cambridge Univ. Press. [9]

Leipzig, J., D. Nüst, C. T. Hoyt, K. Ram, and J. Greenberg. 2021. The Role of Metadata in Reproducible Computational Research. *Patterns* **2**:100322. [3]

Leong, K., and A. Sung. 2015. A Review of Spatio-Temporal Pattern Analysis Approaches on Crime Analysis. *Int. J. Crim. Sci.* **9**:1–33. [6]

Less, E. L., P. McKee, T. Toomey, et al. 2015. Matching Study Areas Using Google Street View: A New Application for an Emerging Technology. *Eval. Program Plann.* **53**:72–79. [1]

Levine, M., P. J. Taylor, and R. Best. 2011. Third Parties, Violence, and Conflict Resolution: The Role of Group Size and Collective Action in the Microregulation of Violence. *Psychol. Sci.* **22**:406–412. [9]

Levine, M. E., J. Vilena, D. Altman, and M. Nadien. 1976. Trust of the Stranger: An Urban/Small Town Comparison. *J. Psychol.* **92**:113–116. [8]

Levy, A. 1999. Urban Morphology and the Problem of the Modern Urban Fabric: Some Questions for Research. *Urban Morphology* **3**:79–85. [8]

Lewis, C. S. 1943. Out of the Silent Planet. London: Bodley Head. [2]

Leyden, K. M., A. Goldberg, and P. Michelbach. 2011. Understanding the Pursuit of Happiness in Ten Major Cities. *Urban Affairs Rev.* **47**:861–888. [8]

Li, W., R. Dong, H. Fu, et al. 2020. Integrating Google Earth Imagery with Landsat Data to Improve 30-M Resolution Land Cover Mapping. *Remote Sens. Environ.* **237**:111563. [7]

Li, Y., and K. Zhang. 2021. Using Social Media for Telemedicine during the COVID-19 Epidemic. *Am. J. Emerg. Med.* **46**:667–668. [10]

Liebst, L. S., P. Ejbye-Ernst, M. de Bruin, J. Thomas, and M. R. Lindegaard. 2022. Face-Touching Behaviour as a Possible Correlate of Mask-Wearing: A Video Observational Study of Public Place Incidents during the COVID-19 Pandemic. *Transbound. Emerg. Dis.* **3**:1319–1325. [9]

Liebst, L. S., R. Philpot, W. Bernasco, et al. 2019. Social Relations and Presence of Others Predict Bystander Intervention: Evidence from Violent Incidents Captured on CCTV. *Aggress. Behav.* **45**:598–609. [9]

Liebst, L. S., R. Philpot, M. Levine, and M. R. Lindegaard. 2020. Cross-National CCTV Footage Shows Low Victimization Risk for Bystander Interveners in Public Conflicts. *Psychol. Violence* **11**:11–18. [9]

Lindegaard, M. R. 2022. Violence in Action: What We Know and What We See. In: Inaugural address, Faculty of Social and Behavioural Sciences. Amsterdam Institute for Social Science Research: Univ. of Amsterdam. [9]

Lindegaard, M. R., and W. Bernasco. 2018. Lessons Learned from Crime Caught on Camera. *J. Res. Crime Delinq.* **55**:155–186. [9]

Lindegaard, M. R., M. Boeri, and R. K. Shukla. 2020. Going Native with Evil. In: Inside Ethnography: Researchers Reflect on the Challenges of Reaching Hidden Populations, pp. 27–48. Oakland: Univ. California Press. [9]

Lindegaard, M. R., L. S. Liebst, W. Bernasco, et al. 2017. Consolation in the Aftermath of Robberies Resembles Post-Aggression Consolation in Chimpanzees. *PLOS ONE* **12**:e0177725. [9]

Lindegaard, M. R., L. S. Liebst, R. Philpot, M. Levine, and W. Bernasco. 2021. Does Danger Level Affect Bystander Intervention in Real-Life Conflicts? Evidence from CCTV Footage. *Soc. Psychol. Person. Sci.* **13**:795–802. [9]

Listl, S., H. Jürges, and R. G. Watt. 2016. Causal Inference from Observational Data. *Community Dent. Oral Epidemiol.* **44**:409–415. [9]

Liu, J. C., G. Pereira, S. A. Uhl, M. A. Bravo, and M. L. Bell. 2015. A Systematic Review of the Physical Health Impacts from Non-Occupational Exposure to Wildfire Smoke. *Environ. Res.* **136**:120–132. [12]

Liu, J. C., A. Wilson, L. J. Mickley, et al. 2017. Who among the Elderly Is Most Vulnerable to Exposure to and Health Risks of Fine Particulate Matter from Wildfire Smoke? *Am. J. Epidemiol.* **186**:730–735. [12]

Liu, Q., X. Gu, F. Deng, et al. 2019. Ambient Particulate Air Pollution and Circulating C-Reactive Protein Level: A Systematic Review and Meta-Analysis. *Int. J. Hyg. Environ. Health* **222**:756–764. [1]

Liu, Y., C. Kliman-Silver, and A. Mislove. 2014. The Tweets They Are a-Changin': Evolution of Twitter Users and Behavior. *Proc. Int. AAAI Conf. Web Soc. Media* **8**:305–314. [4]

Liu, Z., C. He, Q. Zhang, Q. Huang, and Y. Yang. 2012. Extracting the Dynamics of Urban Expansion in China Using DMSP-OLS Nighttime Light Data from 1992 to 2008. *Landsc. Urban Plann.* **106**:62–72. [7]

Loder, A., L. Ambühl, M. Menendez, and K. W. Axhausen. 2019. Understanding Traffic Capacity of Urban Networks. *Sci. Rep.* **9**:16283. [8]

Logan, J. R., Z. Xu, and B. J. Stults. 2014. Interpolating U.S. Decennial Census Tract Data from as Early as 1970 to 2010: A Longitudinal Tract Database. *Prof. Geogr.* **66**:412–420. [3]

Longley, P. A., M. F. Goodchild, D. J. Maguire, and D. W. Rhind. 2015. Geographic Information Science and Systems. Hoboken: Wiley. [6]

Lorenz, K. Z. 1973. The Fashionable Fallacy of Dispensing with Description. *Naturwissenschaften* **60**:1–9. [9]

Lorkowski, P. 2021. Monitoring Continuous Phenomena: Background, Methods and Solutions. Boca Raton: CRC Press. [6]

Lovasi, G. S., S. Grady, and A. Rundle. 2012. Steps Forward: Review and Recommendations for Research on Walkability, Physical Activity and Cardiovascular Health. *Public Health Rev.* **33**:484–506. [3]

Lu, X., E. Wetter, N. Bharti, A. J. Tatem, and L. Bengtsson. 2013. Approaching the Limit of Predictability in Human Mobility. *Sci. Rep.* **3**:2923. [8]

Lu, Y. 2018. The Association of Urban Greenness and Walking Behavior: Using Google Street View and Deep Learning Techniques to Estimate Residents' Exposure to Urban Greenness. *Int. J. Environ. Res. Public Health* **15**:1576. [7]

Lucyk, K., M. Lu, T. Sajobi, and H. Quan. 2015. Administrative Health Data in Canada: Lessons from History. *BMC Med. Inform. Decis. Mak.* **15**:69. [1]

Ludvigsson, J. F., P. Svedberg, O. Olén, G. Bruze, and M. Neovius. 2019. The Longitudinal Integrated Database for Health Insurance and Labour Market Studies (LISA) and Its Use in Medical Research. *Eur. J. Epidemiol.* **34**:423–437. [11]

Lynch, K. 1960. The Image of the City. Cambridge, MA: MIT Press. [8]

Lytle, L. A., and R. L. Sokol. 2017. Measures of the Food Environment: A Systematic Review of the Field, 2007–2015. *Health Place* **44**:18–34. [7]

Ma, B. D., S. L. Ng, T. Schwanen, et al. 2018. Pokémon Go and Physical Activity in Asia: Multilevel Study. *J. Med. Internet Res.* **20**:e217. [4]

MacEachren, A. M., R. E. Roth, J. O'Brien, et al. 2012. Visual Semiotics & Uncertainty Visualization: An Empirical Study. *IEEE Trans. Vis. Comput. Graph.* **18**:2496–2505. [3]

Madan, A., M. Cebrian, D. Lazer, and A. Pentland. 2010. Social Sensing for Epidemiological Behavior Change. In: UBICOMP '10: The 2010 ACM Conference on Ubiquitous Computing, pp. 291–300. New York: ACM. [4]

Madsen, K. M., A. Hviid, M. Vestergaard, et al. 2002. A Population-Based Study of Measles, Mumps, and Rubella Vaccination and Autism. *N. Eng. J. Med.* **347**:1477–1482. [11]

Maes, C., and O. de Lenne. 2022. Filters and Fillers: Belgian Adolescents' Filter Use on Social Media and the Acceptance of Cosmetic Surgery. *J. Child. Media* **May 25**:587–605. [4]

Maestripieri, D. 2005. Gestural Communication in Three Species of Macaques (*Macaca mulatta* , *M. Nemestrina* , *M. Arctoides*): Use of Signals in Relation to Dominance and Social Context. *Gesture* **5**:57–73. [4]

Mahajan, R., and V. Mansotra. 2021. Predicting Geolocation of Tweets: Using Combination of CNN and BiLSTM. *Data Sci. Eng.* **6**:402. [10]

Mahmoud, H., I. M. El Araby, K. Al Hagla, and S. El Sayary. 2013. Human Social Behavior in Public Urban Spaces: Towards Higher Quality Cities. *Spaces Flows* **3**:23–35. [8]

Makel, M. C., and J. A. Plucker. 2014. Facts Are More Important Than Novelty: Replication in the Education Sciences. *Educ. Res.* **43**:304–316. [9]

Mandelbrot, B. B. 1982. The Fractal Geometry of Nature. New York: Freeman Press. [8]

Mangalore, R., M. Knapp, and R. Jenkins. 2007. Income-Related Inequality in Mental Health in Britain: The Concentration Index Approach. *Psychol. Med.* **37**:1037–1045. [1]

Manley, D. 2019. Scale, Aggregation, and the Modifiable Areal Unit Problem. In: Handbook of Regional Science, ed. M. M. Fischer and P. Nijkamp, pp. 1–15. Heidelberg: Springer. [5]

Manning, R., M. Levine, and A. Collins. 2007. The Kitty Genovese Murder and the Social Psychology of Helping: The Parable of the 38 Witnesses. *Am. Psychol.* **62**:555. [9]

Mansfield, T. J., D. A. Rodriguez, J. Huegy, and J. MacDonald Gibson. 2015. The Effects of Urban Form on Ambient Air Pollution and Public Health Risk: A Case Study in Raleigh, North Carolina. *Risk Anal.* **35**:901–918. [8]

Mantelero, A. 2018. AI and Big Data: A Blueprint for a Human Rights, Social and Ethical Impact Assessment. *Comput. L. Sec. Rev.* **34**:754–772. [3]

Marjanovic, M., S. Grubeša, and I. P. Žarko. 2017. Air and Noise Pollution Monitoring in the City of Zagreb by Using Mobile Crowdsensing. In: 25th Int. Conf. on Software, Telecommunications and Computer Networks, pp. 1–5. Piscataway: IEEE. [7]

Markowetz, A., T. Brinkhoff, and B. Seeger. 2005. Geographic Information Retrieval. In: Next Generation Geospatial Information, ed. P. Agouris, pp. 5–14, ISPRS Book Series/International Society for Photogrammetry and Remote Sensing. Leiden: Balkema. [6]

Marmot, M. 2005. Social Determinants of Health Inequalities. *Lancet* **365**:1099–1104. [7]

Marshall, J. D., E. Nethery, and M. Brauer. 2008. Within-Urban Variability in Ambient Air Pollution: Comparison of Estimation Methods. *Atmos. Environ.* **42**:1359–1369. [7]

Marshall, S. 2004. Street and Patterns: The Structure of Urban Geometry. London: Routledge. [8]

Martin, J. L. 2017. Thinking through Methods: A Social Science Primer. Chicago: Univ. Chicago Press. [9]

Martin, K., and K. Shilton. 2016. Putting Mobile Application Privacy in Context: An Empirical Study of User Privacy Expectations for Mobile Devices. *Inform. Soc.* **32**:200–216. [12]

Martino, N., C. Girling, and Y. Lu. 2021. Urban Form and Livability: Socioeconomic and Built Environment Indicators. *Buildings and Cities* **2**:220–243. [8]

Masoud, B., H. Coch, and B. Beckers. 2020. The Correlation between Urban Morphology Parameters and Incident Solar Radiation Performance to Enhance Pedestrian Comfort, Case Study Jeddah, Saudi Arabia. In: Sustainability in Energy and Buildings, ed. J. H. Littlewood, Robert J. et al., pp. 543–554. Singapore: Springer. [8]

Masri, S., E. Scaduto, Y. Jin, and J. Wu. 2021. Disproportionate Impacts of Wildfires among Elderly and Low-Income Communities in California from 20002020. *Int. J. Environ. Res. Public Health* **18**:3921. [12]

Matharaarachchi, S., M. Domaratzki, A. Katz, and S. Muthukumarana. 2022. Discovering Long COVID Symptom Patterns: Association Rule Mining and Sentiment Analysis in Social Media Tweets. *JMIR Form. Res.* **6**:e37984. [10]

Mattingly, S. M., J. M. Gregg, P. Audia, et al. 2019. The Tesserae Project: Large-Scale, Longitudinal, *in Situ,* Multimodal Sensing of Information Workers. In: CHI '19: Conf. on Human Factors in Computing Systems, pp. 1–8. New York: ACM. [4, 5]

Mays, N., and C. Pope. 2000. Assessing Quality in Qualitative Research. *Br. Med. J.* **320**:50–52. [2]

McDonald, C. J., S. M. Huff, J. G. Suico, et al. 2003. LOINC, a Universal Standard for Identifying Laboratory Observations: A 5-Year Update. *Clin. Chem.* **49**:624–633. [5]

McDonald, R. I., T. Kroeger, P. Zhang, and P. Hamel. 2020. The Value of US Urban Tree Cover for Reducing Heat-Related Health Impacts and Electricity Consumption. *Ecosystems* **23**:137–150. [8]

McGrail, K., and K. Jones. 2018. Population Data Science: The Science of Data about People. *Int. J. Popul. Data Sci.* **3**:ijpds.v3i4.918. [12]

McIntyre, A. 2008. Participatory Action Research. Thousand Oaks: Sage. [5]

McPherson, M., L. Smith-Lovin, and J. M. Cook. 2001. Birds of a Feather: Homophily in Social Networks. *Annu. Rev. Sociol.* **27**:415–444. [4]

Mehta, V. 2014. Evaluating Public Space. *J. Urban Des.* **19**:53–88. [8]

Meier, A., and E. Portmann, eds. 2016. Smart City: Strategie, Governance und Projekte. Edition HMD, vol. XXXII. Wiesbaden: Springer Vieweg. [6]

Melchior, M., A. Ziad, E. Courtin, et al. 2018. Intergenerational Socioeconomic Mobility and Adult Depression. *Am. J. Epidemiol.* **187**:260–269. [4]

Meloni, M. 2014. How Biology Became Social, and What It Means for Social Theory. *Sociol. Rev.* **62**:593–614. [9]

Menon, R. 2019. Initiation of Human Parturition: Signaling from Senescent Fetal Tissues via Extracellular Vesicle Mediated Paracrine Mechanism. *Obstet. Gynecol. Sci.* **62**:199–211. [1]

Methorst, R., J. Gerlach, D. Boenke, and J. Leven. 2007. Shared Space: Safe or Dangerous? A Contribution to Objectification of a Popular Design Philosophy. Walk21 Conference. Toronto: Walk21. [8]

Metzl, J. M., and H. Hansen. 2018. Structural Competency and Psychiatry. *JAMA Psychiatry* **75**:115–116. [1]

Meyer, W. B., and B. L. Turner. 1992. Human Population Growth and Global Land-Use/Cover Change. *Annu. Rev. Ecol. Syst.* **23**:39–61. [7]

Mezzetti, M., D. Palli, and F. Dominici. 2020. Combining Individual and Aggregated Data to Investigate the Role of Socioeconomic Disparities on Cancer Burden in Italy. *Stat. Med.* **39**:26–44. [5]

Middleton, S. E., G. Kordopatis-Zilos, S. Papadopoulos, and Y. Kompatsiaris. 2018. Location Extraction from Social Media: Geoparsing, Location Disambiguation, and Geotagging. *ACM Trans. Inf. Syst.* **36**:1–27. [7]

Miguel, E., C. Camerer, K. Casey, et al. 2014. Promoting Transparency in Social Science Research. *Science* **343**:30–31. [12]

Mikolov, T., K. Chen, G. Corrado, and J. Dean. 2013. Distributed Representations of Words and Phrases and Their Compositionality. *arXiv* **1310**:4546. [10]

Miller, S. J. 2022. Metadata Resources: Selected Reference Documents, Web Sites, Books, Articles, and Other Resources. Univ. Wisconsin-Milwaukee. https://sites.uwm.edu/mll/metadata-resources/ (accessed Nov. 1, 2022). [3]

Mishra, J., P. Mishra, and N. K. Arora. 2021. Linkages between Environmental Issues and Zoonotic Diseases: With Reference to COVID-19 Pandemic. *Environ. Sustain.* **4**:455–467. [5]

Mitani, J. C., J. Call, P. M. Kappeler, R. A. Palombit, and J. B. Silk. 2012. The Evolution of Primate Societies. Chicago: Univ. Chicago Press. [4]

Mok, P. L. H., S. Antonsen, C. B. Pedersen, et al. 2018. Family Income Inequalities and Trajectories through Childhood and Self-Harm and Violence in Young Adults: A Population-Based, Nested Case-Control Study. *Lancet Public Health* **3**:e498–e507. [4]

Molenaar, P. C. M. 2004. A Manifesto on Psychology as Idiographic Science: Bringing the Person Back into Scientific Psychology: This Time Forever. *Measurement* **2**:201–218. [2]

Molotch, H., and D. Boden. 1993. The Compulsion of Proximity. In: Now/Here: Time, Space and Social Theory, ed. R. Friedland and D. Boden, pp. 257–286. Oakland: Univ. California Press. [9]

Monsivais, D., K. Bhattacharya, A. Ghosh, R. I. M. Dunbar, and K. Kaski. 2017. Seasonal and Geographical Impact on Human Resting Periods. *Sci. Rep.* **7**:10717. [5]

Montejo-Ráez, A., and S. M. Jiménez-Zafra. 2022. Current Approaches and Applications in Natural Language Processing. *Appl. Sci.* **12**:4859. [10]

Montero, M. I., and R. B. Marx. 2001. Roberto Burle Marx: The Lyrical Landscape. Berkeley: Univ. California Press. [3]

Moonesinghe, R., M. J. Khoury, and A. C. J. W. Janssens. 2007. Most Published Research Findings Are False: But a Little Replication Goes a Long Way. *PLOS Med.* **4**:e28. [11]

Morales, A. J., X. Dong, Y. Bar-Yam, and A. Sandy Pentland. 2019. Segregation and Polarization in Urban Areas. *R. Soc. Open Sci.* **6**:190573. [8]

Morelli, G., N. Chaudhary, A. Gottlieb, et al. 2017. Taking Culture Seriously: A Pluralistic Approach to Attachment. In: Contextualizing Attachment: The Cultural Nature of Attachment, ed. K. H. Bard and H. Keller, pp. 139–169, Strüngmann Forum Reports, vol. 22, J. R. Lupp, series ed. Cambridge, MA: MIT Press. [2]

Morrison, C., J. P. Lee, P. J. Gruenewald, and C. Mair. 2016. The Reliability of Naturalistic Observations of Social, Physical and Economic Environments of Bars. *Addict. Res. Theory* **24**:330–340. [9]

Mortensen, C. R., and R. B. Cialdini. 2010. Full-Cycle Social Psychology for Theory and Application. *Soc. Person. Psychol. Compass* **4**:53–63. [9]

Moser, S. C. 2014. Raising the Seas, Rising to Greatness? Meeting the Challenge of Coastal Climate Change. In: Applied Studies in Climate Adaptation, pp. 175–180. [8]

Moskowitz, P. E. 2017. How to Kill a City: Gentrification, Inequality, and the Fight for the Neighborhood. New York: Bold Type Books. [8]

Moudon, A. V. 1997. Urban Morphology as an Emerging Interdisciplinary Field. *Urban Morphology* **1**:3–10. [8]

Mouratidis, K., D. Ettema, and P. Næss. 2019. Urban Form, Travel Behavior, and Travel Satisfaction. *Transportation Research Part A: Policy and Practice* **129**:306–320. [8]

Muilu, J., L. Peltonen, and J. E. Litton. 2007. The Federated Database: A Basis for Biobank-Based Post-Genome Studies, Integrating Phenome and Genome Data from 600,000 Twin Pairs in Europe. *Eur. J. Hum. Genet.* **15**:718–723. [11]

Muller, K. U., E. Mennigen, S. Ripke, et al. 2013. Altered Reward Processing in Adolescents with Prenatal Exposure to Maternal Cigarette Smoking. *JAMA Psychiatry* **70**:847–856. [1]

Mumford, L. 1961. The City in History. San Diego: Harcourt, Brace & World. [8]

Munzel, T., M. Sorensen, O. Hahad, M. Nieuwenhuijsen, and A. Daiber. 2023. The Contribution of the Exposome to the Burden of Cardiovascular Disease. *Nat. Rev. Cardiol.*651–669. [1]

Murphy, D., and A. Nicol. 2010. Wash with Care: Public Service Announcement. http://hdl.handle.net/2429/33872 (accessed Nov. 1, 2022). [3]

Myrick, J. G., and J. F. Willoughby. 2022. A Mixed Methods Inquiry into the Role of Tom Hanks' COVID-19 Social Media Disclosure in Shaping Willingness to Engage in Prevention Behaviors. *Health Commun.* **37**:824–832. [10]

Naimi, A. I., D. B. Richardson, and S. R. Cole. 2013. Causal Inference in Occupational Epidemiology: Accounting for the Healthy Worker Effect by Using Structural Nested Models. *Am. J. Epidemiol.* **178**:1681–1686. [11]

NASEM. 2019. Reproducibility and Replicability in Science. Washington, D.C.: National Academies Press. [5]

Nassauer, A., and N. M. Legewie. 2018. Video Data Analysis: A Methodological Frame for a Novel Research Trend. *Sociol. Methods Res.* **50**:135–174. [9]

———. 2022. Video Data Analysis: How to Use 21st Century Video in the Social Sciences. London: Sage. [9]

National Center for Health Statistics. 2021. Provisional Drug Overdose Death Counts. https://nchstats.com/category/opioid/ (accessed Nov. 7, 2022). [10]

National Institute of Dental and Craniofacial Research. 2018. NIDCR Strategic Plan 2014-2019. https://www.nidcr.nih.gov/about-us/strategic-plan (accessed Nov. 14, 2022). [11]

Nguyen, D. Q., T. Vu, and A. T. Nguyen. 2020. BERTweet: A Pre-Trained Language Model for English Tweets. In: Proc. of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pp. 9–14. Stroudsburg, PA: ACL. [10]

Nigam, A., H. K. Dambanemuya, M. Joshi, and N. V. Chawla. 2017. Harvesting Social Signals to Inform Peace Processes Implementation and Monitoring. *Big Data* **5**:337–355. [4]

Nissenbaum, H. 2011. A Contextual Approach to Privacy Online. *Daedalus* **140**:32–48. [12]

Noble, K. G., S. M. Houston, N. H. Brito, et al. 2015. Family Income, Parental Education and Brain Structure in Children and Adolescents. *Nat. Neurosci.* **18**:773–778. [1]

Norris, E., A. N. Finnerty, J. Hastings, G. Stokes, and S. Michie. 2019. A Scoping Review of Ontologies Related to Human Behaviour Change. *Nat. Hum. Behav.* **3**:164–172. [3]

Norwegian Institute of Public Health. 2022. Norwegian Cause of Death Registry. https://www.fhi.no/en/hn/health-registries/cause-of-death-registry/ (accessed Nov. 14, 2022). [11]

Nosek, B. A., G. Alter, G. C. Banks, et al. 2015. Promoting an Open Research Culture. *Science* **348**:1422–1425. [5]

Nowok, B., G. M. Raab, and C. Dibben. 2016. Synthpop: Bespoke Creation of Synthetic Data in R. *J. Stat. Softw.* **74**:1–26. [11]

Noy, N. F. 2004. Semantic Integration: A Survey of Ontology-Based Approaches. *SIGMOD Rec.* **33**:65–70. [5]

Ntoutsi, E., P. Fafalios, U. Gadiraju, et al. 2020. Bias in Data-Driven Artificial Intelligence Systems: An Introductory Survey. *Data Min. Knowl. Disc.* **10**:e1356. [5]

O'Brien, E. C., A. M. Rodriguez, H. C. Kum, et al. 2019. Patient Perspectives on the Linkage of Health Data for Research: Insights from an Online Patient Community Questionnaire. *Int J Med Inform* **127**:9–17. [1]

O'Donnell, O., E. van Doorslaer, A. Wagstaff, and M. Lindelow. 2008. Analyzing Health Equity Using Household Survey Data. Washington, D.C.: World Bank. [1]

Odgers, C. L., A. Caspi, C. J. Bates, R. J. Sampson, and T. E. Moffitt. 2012. Systematic Social Observation of Children's Neighborhoods Using Google Street View: A Reliable and Cost-Effective Method. *J. Child Psychol. Psychiatry* **53**:1009–1017. [1]

Ogilvie, J. M., S. Tzoumakis, T. Allard, et al. 2021. Prevalence of Psychiatric Disorders for Indigenous Australians: A Population-Based Birth Cohort Study. *Epidemiol. Psychiatr. Sci.* **30**:e21. [1]

Oliveira, M., C. Bastos-Filho, and R. Menezes. 2017. The Scaling of Crime Concentration in Cities. *PLOS ONE* **12**:e0183110. [8]

Olteanu, A., C. Castillo, F. Diaz, and E. Kıcıman. 2019. Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries. *Front. Big Data* **2**:fdata.2019.00013. [4, 12]

Onnela, J.-P., J. Saramäki, J. Hyvönen, et al. 2007. Structure and Tie Strengths in Mobile Communication Networks. *PNAS* **104**:7332–7336. [4]

Oppermann, M. 2000. Triangulation: A Methodological Discussion. *Int. J.Tour. Res.* **2**:141–145. [2]

Palchykov, V., K. Kaski, J. Kertész, A.-L. Barabási, and R. I. M. Dunbar. 2012. Sex Differences in Intimate Relationships. *Sci. Rep.* **2**:370. [4]

Pampalon, R., D. Hamel, P. Gamache, et al. 2012. An Area-Based Material and Social Deprivation Index for Public Health in Quebec and Canada. *Can. J. Public Health* **103**:17–22. [1]

Pandharipande, A. 2021. Social Sensing in IoT Applications: A Review. *IEEE Sens. J.* **21**:12523–12530. [7]

Paquette, D., and J. M. St. George. 2023. Proximate and Ultimate Mechanisms of Human Father-Child Rough-and-Tumble Play. *Neurosci. Biobehav. Rev.* **149**:105151. [1]

Parker, N., A. P. Wong, G. Leonard, et al. 2017. Income Inequality, Gene Expression, and Brain Maturation during Adolescence. *Sci. Rep.* **7**:7397. [1]

Parkinson, M., T. Champion, and J. Simmle, Turok, I. 2006. State of the English Cities. A Research Study, vol. 1. London: Office of the Deputy Prime Minister. [8]

Paus, T. 2010. Population Neuroscience: Why and How. *Hum. Brain Mapp.* **31**:891–903. [1]

———. 2013. Population Neuroscience. Heidelberg: Springer. [1]

———. 2016. Population Neuroscience. *Handb. Clin. Neurol.* **138**:17–37. [1]

Paus, T., J. Brook, and D. Doiron. 2022. Mapping Inequalities in the Physical, Built and Social Environment in Population-Based Studies of Brain Health. *Front. Neuroimaging* **1**:884191. [1, 2, 5]

Pearl, J., and D. Mackenzie. 2018. The Book of Why: The New Science of Cause and Effect. New York: Basic Books. [2, 3]

Pedalino, F., and A.-L. Camerini. 2022. Instagram Use and Body Dissatisfaction: The Mediating Role of Upward Social Comparison with Peers and Influencers among Young Females. *Int. J. Environ. Res. Public Health* **19**:1543. [4]

Peiser, R. B., and M. Hugel. 2022. Is the Pandemic Causing a Return to Urban Sprawl? *Journal of Comparative Urban Law and Policy* **5**:26–41. [8]

Pencarrick Hertzman, C., N. Meagher, and K. M. McGrail. 2013. Privacy by Design at Population Data BC: A Case Study Describing the Technical, Administrative, and Physical Controls for Privacy-Sensitive Secondary Use of Personal Information for Research in the Public Interest. *J Am Med Inform Assoc* **20**:25–28. [1]

Peng, G. C. Y., M. Alber, A. Buganza Tepole, et al. 2021. Multiscale Modeling Meets Machine Learning: What Can We Learn? *Arch. Comput. Meth. Eng.* **28**:1017–1037. [5]

Peng, R. D., and S. C. Hicks. 2021. Reproducible Research: A Retrospective. *Annu. Rev. Public Health* **42**:79–93. [3]

Pentland, A. 2014. Social Physics: How Social Networks Can Make Us Smarter. New York: Penguin Books. [8]

Perneger, T. V. 1998. What's Wrong with Bonferroni Adjustments. *Br. Med. J.* **316**:1236–1238. [12]

Perrotta, G. 2020. The Concept of Altered Perception in "Body Dysmorphic Disorder": The Subtle Border between the Abuse of Selfies in Social Networks and Cosmetic Surgery, between Socially Accepted Dysfunctionality and the Pathological Condition. *J. Neurol. Neurologic. Sci. Disord.* **6**:1–7. [4]

Persson, M., S. Opdahl, K. Risnes, et al. 2020. Gestational Age and the Risk of Autism Spectrum Disorder in Sweden, Finland, and Norway: A Cohort Study. *PLOS Med.* **17**:e1003207. [11]

Petty, J. 2016. The London Spikes Controversy: Homelessness, Urban Securitisation and the Question of "Hostile Architecture". *Int. J. Crime, Justice Soc. Democr.* **5**:67. [3]

Pew Research Center. 2021. Social Media Fact Sheet. https://www.pewresearch.org/internet/fact-sheet/social-media/ (accessed Jan. 5, 2023). [10]

Phan, J. H., C. F. Quo, C. Cheng, and M. D. Wang. 2012. Multiscale Integration of -Omic, Imaging, and Clinical Data in Biomedical Informatics. *IEEE Rev. Biomed. Eng.* **5**:74–87. [5]

Philpot, R., L. S. Liebst, M. Levine, W. Bernasco, and M. R. Lindegaard. 2020. Would I Be Helped? Cross-National CCTV Footage Shows That Intervention Is the Norm in Public Conflicts. *Am. Psychol.* **75**:66–75. [9]

Philpot, R., L. S. Liebst, M. R. Lindegaard, P. Verbeek, and M. Levine. 2022. Reconciliation in Human Adults: A Video-Assisted Naturalistic Observational Study of Post Conflict Conciliatory Behaviour in Interpersonal Aggression. *Behaviour* **159**:1225–1261. [9]

Philpot, R., L. S. Liebst, K. K. Møller, M. R. Lindegaard, and M. Levine. 2019. Capturing Violence in the Night-Time Economy: A Review of Established and Emerging Methodologies. *Aggress. Violent Behav.* **46**:56–65. [9]

Pollet, T. V., S. G. B. Roberts, and R. I. M. Dunbar. 2013. Going That Extra Mile: Individuals Travel Further to Maintain Face-to-Face Contact with Highly Related Kin Than with Less Related Kin. *PLOS ONE* **8**:e53929. [4]

Psaty, B. M., C. J. O'Donnell, V. Gudnason, et al. 2009. Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium: Design of Prospective Meta-Analyses of Genome-Wide Association Studies from 5 Cohorts. *Circ. Cardiovasc. Genet.* **2**:73–80. [1]

Public Health Informatics Institute. 2021. Toolkit for Planning an EHR-Based Surveillance Program. https://phii.org/course/toolkit-for-planning-an-ehr-based-surveillance-program/ (accessed Oct. 7, 2022). [5]

Publications Office of the European Union. 2003. Union Po of the E. Regulation (EC) No 1059/2003 of the European Parliament and of the Council of 26 May 2003 on the Establishment of a Common Classification of Territorial Units for Statistics (NUTS), CELEX1 http://op.europa.eu/en/publication-detail/-/publication/dff690b3-c02d-11e9-9d01-01aa75ed71a1/language-en/format-PDF (accessed Nov. 14, 2022). [11]

Pukkala, E., G. Engholm, L. K. Højsgaard Schmidt, et al. 2018. Nordic Cancer Registries: An Overview of Their Procedures and Data Comparability. *Acta Oncol.* **57**:440–455. [11]

Pullano, G., E. Valdano, N. Scarpa, S. Rubrichi, and V. Colizza. 2020. Evaluating the Effect of Demographic Factors, Socioeconomic Factors, and Risk Aversion on Mobility during the COVID-19 Epidemic in France under Lockdown: A Population-Based Study. *Lancet Digit. Health* **2**:e638–e649. [4]

Putnam, R. D. 2000. Bowling Alone: The Collapse and Revival of American Community. New York: Simon & Schuster. [4]

Quinn, P. D., M. E. Rickert, C. E. Weibull, et al. 2017. Association between Maternal Smoking During Pregnancy and Severe Mental Illness in Offspring. *JAMA Psychiatry* **74**:589–596. [11]

Ragan, E. D., H.-C. Kum, G. Ilangovan, and H. Wang. 2018. Balancing Privacy and Information Disclosure in Interactive Record Linkage with Visual Masking. In: Proc. of the 2018 CHI Conference on Human Factors in Computing Systems, pp. 1–12. New York: ACM. [12]

Raghunathan, T. E., P. K. Diehr, and A. D. Cheadle. 2003. Combining Aggregate and Individual Level Data to Estimate an Individual Level Correlation Coefficient. *J. Educ. Behav. Stat.* **28**:1–19. [5]

Rainham, D., I. McDowell, D. Krewski, and M. Sawada. 2010. Conceptualizing the Healthscape: Contributions of Time Geography, Location Technologies and Spatial Ecology to Place and Health Research. *Soc. Sci. Med.* **70**:668–676. [7]

Rajmil, L., M. Herdman, U. Ravens-Sieberer, et al. 2014. Socioeconomic Inequalities in Mental Health and Health-Related Quality of Life (HRQOL) in Children and Adolescents from 11 European Countries. *Int. J. Public Health* **59**:95–105. [1]

Ramezani, M., G. Ilangovan, and H.-C. Kum. 2021. Evaluation of Machine Learning Algorithms in a Human-Computer Hybrid Record Linkage System. In: Proc. of the AAAI 2021 Spring Symposium on Combining Machine Learning and Knowledge Engineering (AAAI-Make 2021), ed. A. Martin et al., CEUR Workshop Proc., vol. 2846, paper 25. Palo Alto: AAAI Press. [12]

Rantakallio, P. 1988. The Longitudinal Study of the Northern Finland Birth Cohort of 1966. *Paediatr. Perinat. Epidemiol.* **2**:59–88. [1]

Ravat, F., and Y. Zhao. 2019. Data Lakes: Trends and Perspectives. In: Database and Expert Systems Applications, ed. S. Hartmann et al., pp. 304–313. Cham: Springer. [12]

Read, J. M., and M. Torrado. 2009. Remote Sensing. In: International Encyclopedia of Human Geography, ed. R. Kitchin and N. Thrift, pp. 335–346. London: Elsevier. [7]

RECAP preterm. 2022. Research on European Children and Adults Born Preterm (RECAP). https://recap-preterm.eu/ (accessed Nov. 14, 2022). [11]

Reid, C. E., and M. M. Maestas. 2019. Wildfire Smoke Exposure under Climate Change: Impact on Respiratory Health of Affected Communities. *Curr. Opin. Pulm. Med.* **25**:179–187. [12]

Reilly, M. 2017. Is Facebook Targeting Ads at Sad Teens? MIT Technology Review. https://www.technologyreview.com/2017/05/01/105987/is-facebook-targeting-ads-at-sad-teens/ (accessed Oct. 7, 2022). [5]

Reiss Jr, A. J. 1992. The Trained Incapacities of Sociologists. In: Sociology and Its Publics, ed. T. Halliday and M. Janowitz, pp. 297–315. Chicago: Univ. Chicago Press. [9]

Reiter, R. 1978. On Closed World Data Bases. In: Logic and Data Bases, ed. H. Gallaire and J. Minker, pp. 55–76. Boston: Springer. [2]

Relph, E. C. 1976. Place and Placelessness. London: Pion. [4]

Richardson, H. W. 1995. Economies and Diseconomies of Agglomeration. In: Urban Agglomeration and Economic Growth, ed. H. Giersch, pp. 123–155. Heidelberg: Springer. [8]

Richiardi, L., C. Pizzi, and N. Pearce. 2013. Commentary: Representativeness Is Usually Not Necessary and Often Should Be Avoided. *Int. J. Epidemiol.* **42**:1018–1022. [5]

Rigaux, P., M. Scholl, and A. Voisard. 2011. Spatial Databases: With Applications to GIS. The Morgan Kaufmann Series in Data Management Systems. San Francisco: Morgan Kaufmann. [6]

Ritchie, S. J. 2020. Science Fictions: How Fraud, Bias, Negligence, and Hype Undermine the Search for Truth. New York: MacMillan. [5]

Roberts, S. G. B., and R. I. M. Dunbar. 2011. The Costs of Family and Friends: An 18-Month Longitudinal Study of Relationship Maintenance and Decay. *Evol. Hum. Behav.* **32**:186–197. [4]

Robertson, T. L., H. Kato, G. Rhoads, et al. 1977. Epidemiologic Studies of Coronary Heart Disease and Stroke in Japanese Men Living in Japan, Hawaii and California. *Am. J. Cardiol.* **39**:239–243. [4]

Robinson, P. N., S. Köhler, S. Bauer, et al. 2008. The Human Phenotype Ontology: A Tool for Annotating and Analyzing Human Hereditary Disease. *Am. J. Hum. Genet.* **83**:610–615. [5]

Robles-Granda, P., S. Lin, X. Wu, et al. 2021. Jointly Predicting Job Performance, Personality, Cognitive Ability, Affect, and Well-Being. *IEEE Comput. Intell. Mag.* **16**:46–61. [4]

Rodriguez, L. A., Y. Jin, S. A. Talegawkar, et al. 2020. Differences in Diet Quality among Multiple US Racial/Ethnic Groups from the Mediators of Atherosclerosis in South Asians Living in America (MASALA) Study and the Multi-Ethnic Study of Atherosclerosis (MESA). *J. Nutr.* **150**:1509–1515. [4]

Rosenbaum, P. R. 2020. Modern Algorithms for Matching in Observational Studies. *Annu. Rev. Stat. Appl.* **7**:143–176. [12]

Rossi, E., and J. Balsa-Barreiro. 2020. The Future of Work in the Post-COVID-19 World. *Economic and Political Weekly* **55**:23–26. [8]

Rothman, K. J., and S. Greenland. 2005. Causation and Causal Inference in Epidemiology. *Am. J. Public Health* **95**:144–150. [2]

Rothstein, R. 2017. The Color of Law: A Forgotten History of How Our Government Segregated America. New York: Liveright Publishing. [3]

Rubin, D. B. 1979. Using Multivariate Matched Sampling and Regression Adjustment to Control Bias in Observational Studies. *J. Am. Stat. Assoc.* **74**:318–328. [12]

Rutter, H., N. Savona, K. Glonti, et al. 2017. The Need for a Complex Systems Model of Evidence for Public Health. *Lancet* **390**:2602–2604. [7]

Rzotkiewicz, A., A. L. Pearson, B. V. Dougherty, A. Shortridge, and N. Wilson. 2018. Systematic Review of the Use of Google Street View in Health Research: Major Themes, Strengths, Weaknesses and Possibilities for Future Research. *Health Place* **52**:240–246. [3]

Sabouret, P., P. P. Bocchino, and G. Biondi-Zoccai. 2020. Positive and Negative Impact of Social Media in the COVID-19 Era. *Rev. Cardiovasc. Med.* **21**:489–492. [10]

Saha, K., A. E. Bayraktaroglu, A. T. Campbell, et al. 2019. Social Media as a Passive Sensor in Longitudinal Studies of Human Behavior and Wellbeing. In: CHI '19: CHI Conference on Human Factors in Computing Systems, pp. 1–8. New York: ACM. [4]

Sallach, D. L. 2003. Social Theory and Agent Architectures: Prospective Issues in Rapid-Discovery Social Science. *Soc. Sci. Comput. Rev.* **21**:179–195. [9]

Sallis, J. F., R. B. Cervero, W. Ascher, et al. 2006. An Ecological Approach to Creating Active Living Communities. *Annu. Rev. Public Health* **27**:297–322. [7]

Samet, H. 2006. Foundations of Multidimensional and Metric Data Structures. The Morgan Kaufmann Series in Data Management Systems. San Francisco: Morgan Kaufmann. [6]

Sampson, R. J., and S. W. Raudenbush. 2004. Seeing Disorder: Neighborhood Stigma and the Social Construction of "Broken Windows". *Soc. Psychol. Q.* **67**:319–342. [9]

SAMSHA. 2017. Key Substance Use and Mental Health Indicators in the United States: Results from the 2016 National Survey on Drug Use and Health (HHS Publication No. SMA 17-5044, NSDUH Series H-52). Rockville, MD: Center for Behavioral Health Statistics and Quality, Substance Abuse and Mental Health Services Administration. [10]

———. 2020. Key Substance Use and Mental Health Indicators in the United States: Results from the 2019 National Survey on Drug Use and Health (HHS Publication No. Pep20-07-01-001, NSDUH Series H-55). Rockville, MD: Center for Behavioral Health Statistics and Quality, Substance Abuse and Mental Health Services Administration. [10]

Samuels, I. 2008. Typomorphology and Urban Design Practice I. Samuels. *Urban Morphology* **12**:58–62. [8]

San Miguel, M., J. H. Johnson, J. Kertesz, et al. 2012. Challenges in Complex Systems Science. *Eur. Phys. J. Spec. Top.* **214**:245–271. [12]

Sánchez, F. 2018. Racial Gerrymandering and Geographic Information Systems: Subverting the 2011 Texas District Map with Election Technologies. *Technical Commun.* **65**:354–370. [3]

Sandin, S., H. Hjalgrim, B. Glimelius, et al. 2006. Incidence of Non-Hodgkin's Lymphoma in Sweden, Denmark, and Finland from 1960 through 2003: An Epidemic That Was. *Cancer Epidemiol. Biomarkers Prev.* **15**:1295–1300. [11]

Sandin, S., D. Schendel, P. Magnusson, et al. 2015. Autism Risk Associated with Parental Age and with Increasing Difference in Age between the Parents. *Mol. Psychiatry* **21**:693–700. [11]

Sanmartin, M. X., R. M. McKenna, M. M. Ali, and J. D. Krebs. 2020. Racial Disparities in Payment Source of Opioid Use Disorder Treatment among Non-Incarcerated Justice-Involved Adults in the United States. *J. Mental Health Pol. Econ.* **23**:19–25. [10]

Santana, P., ed. 2017. Atlas of Population Health in European Union Regions. Coimbra: Coimbra Univ. Press. [7]

Sapena, M., M. Wurm, H. Taubenböck, D. Tuia, and L. A. Ruiz. 2021. Estimating Quality of Life Dimensions from Urban Spatial Pattern Metrics. *Comput. Environ. Urban Syst.* **85**:101549. [8]

Saramäki, J., and K. Kaski. 2005. Modelling Development of Epidemics with Dynamic Small-World Networks. *J. Theor. Biol.* **234**:413–421. [5]

Sariaslan, A., H. Larsson, B. D'Onofrio, N. Långström, and P. Lichtenstein. 2014. Childhood Family Income, Adolescent Violent Criminality and Substance Misuse: Quasi-Experimental Total Population Study. *Br. J. Psychiatry* **205**:286–290. [11]

Sarkar, C., C. Webster, and J. Gallacher. 2015. UK Biobank Urban Morphometric Platform (UKBUMP): A Nationwide Resource for Evidence-Based Healthy City Planning and Public Health Interventions. *Ann. GIS* **21**:135–148. [7]

Sarker, A., R. Ginn, A. Nikfarjam, et al. 2015. Utilizing Social Media Data for Pharmacovigilance: A Review. *J. Biomed. Inform.* **54**:202–212. [4, 10]

Sarker, A., and G. Gonzalez-Hernandez. 2018. An Unsupervised and Customizable Misspelling Generator for Mining Noisy Health-Related Text Sources. *J. Biomed. Inform.* **88**:98–107. [10]

Sarker, A., G. Gonzalez-Hernandez, Y. Ruan, and J. Perrone. 2019. Machine Learning and Natural Language Processing for Geolocation-Centric Monitoring and Characterization of Opioid-Related Social Media Chatter. *JAMA Netw. Open* **2**:e1914672. [4, 10]

Sarker, A., S. Lakamana, W. Hogg-Bremer, et al. 2020. Self-Reported COVID-19 Symptoms on Twitter: An Analysis and a Research Resource. *J. Am. Med. Inform. Assoc.* **27**:1310–1315. [4, 10]

Sarker, A., K. O'Connor, R. Ginn, et al. 2016. Social Media Mining for Toxicovigilance: Automatic Monitoring of Prescription Medication Abuse from Twitter. *Drug Saf.* **39**:231–240. [10]

Satizabal, C. L., H. H. H. Adams, D. P. Hibar, et al. 2019. Genetic Architecture of Subcortical Brain Structures in 38,851 Individuals. *Nat. Genet.* **51**:1624–1636. [1]

Sato, Y., and Y. Zenou. 2015. How Urbanization Affects Employment and Social Interaction. *Eur. Econ. Rev.* **75**:131–155. [8]

Saxbe, D., G. W. Corner, M. Khaled, et al. 2018. The Weight of Fatherhood: Identifying Mechanisms to Explain Paternal Perinatal Weight Gain. *Health Psychol. Rev.* **12**:294–311. [1]

Sayyadiharikandeh, M., O. Varol, K.-C. Yang, A. Flammini, and F. Menczer. 2020. Detection of Novel Social Bots by Ensembles of Specialized Classifiers. In: Proc. of the 29th ACM International Conference on Information & Knowledge Management (CIKM), pp. 2725–2732. New York: ACM. [4]

Scambler, G. 2012. Health Inequalities. *Sociol. Health Illn.* **34**:130–146. [1]

Schläpfer, M., B. Luis, S. Grauwin, et al. 2014. The Scaling of Human Interactions with City Size. *J. R. Soc. Interface* **11**:20130789. [8]

Schmidt, A. F., and C. Finan. 2018. Linear Regression and the Normality Assumption. *J. Clin. Epidemiol.* **98**:146–151. [5]

Schmit, C., K. V. Ajayi, A. O. Ferdinand, et al. 2020. Communicating with Patients about Software for Enhancing Privacy in Secondary Database Research Involving Record Linkage: Delphi Study. *J. Med. Internet Res.* **22**:e20783. [12]

Schmit, C., A. O. Ferdinand, T. V. Giannouchos, and H.-C. Kum. 2024. Case Study on Communicating with Research Ethics Committees about Minimizing Risk through Software: An Application for Record Linkage in Secondary Data Analysis. *Am. Med. Inform. Assoc. Open*, in press. [12]

Schmit, C., T. Giannouchos, M. Ramezani, et al. 2021. US Privacy Laws Go against Public Preferences and Impede Public Health and Research: Survey Study. *J. Med. Internet Res.* **23**:e25266. [1]

Schmit, C., K. Kelly, and J. Bernstein. 2019. Cross Sector Data Sharing: Necessity, Challenge, and Hope. *J. Law. Med. Ethics* **47**:83–86. [5]

Schmit, C., B. Larson, and H.-C. Kum. 2022. Data Privacy in the Time of Plague. *Yale J. Health Pol. Law Ethics* **21**:152–227. [5]

Schneider, A., M. A. Friedl, and D. Potere. 2010. Mapping Global Urban Areas Using MODIS 500-M Data: New Methods and Datasets Based on "Urban Ecoregions". *Remote Sens. Environ.* **114**:1733–1746. [7]

Schooler, J. W. 2014. Metascience Could Rescue the "Replication Crisis". *Nature* **515**:9. [5]

Self, W. 2007. Psychogeography: Disentangling the Modern Conundrum of Psyche and Place. New York: Bloomsbury. [8]

Seyfarth, R. M., and D. L. Cheney. 2012. The Evolutionary Origins of Friendship. *Annu. Rev. Psychol.* **63**:153–177. [4]

Shadish, W. R., T. D. Cook, and D. T. Campbell. 2001. Experimental and Quasi-Experimental Designs for Generalized Causal Inference. Boston: Houghton Mifflin. [12]

Shaw, J. A., N. Sethi, and C. K. Cassel. 2020. Social License for the Use of Big Data in the COVID-19 Era. *npj Digit. Med.* **3**:128. [5]

Sheppard, E., H. Leitner, R. B. McMaster, and H. Tian. 1999. GIS-Based Measures of Environmental Equity: Exploring Their Sensitivity and Significance. *J. Expo. Anal. Environ. Epidemiol.* **9**:18–28. [6]

Shin, J., S. Ma, E. Hofer, et al. 2020. Global and Regional Development of the Human Cerebral Cortex: Molecular Architecture and Occupational Aptitudes. *Cereb. Cortex* **30**:4121–4139. [1]

Shultz, S., and R. I. M. Dunbar. 2010. Bondedness and Sociality. *Behaviour* **147**:775–803. [4]

Simon, P. 2017. Les Descendants D'immigrés Et la Question de L'intégration. *Reg. crois. éconon.* **20**:81–92. [4]

Small, M. L., and J. M. Cook. 2021. Using Interviews to Understand Why: Challenges and Strategies in the Study of Motivated Action. *Sociol. Methods Res.* **Mar 17**:1591–1631. [9]

Smith, G. D., and S. Ebrahim. 2003. 'Mendelian Randomization': Can Genetic Epidemiology Contribute to Understanding Environmental Determinants of Disease? *Int. J. Epidemiol.* **32**:1–22. [1]

Smith, L., L. Foley, and J. Panter. 2019a. Activity Spaces in Studies of the Environment and Physical Activity: A Review and Synthesis of Implications for Causality. *Health Place* **58**:102113–102113. [7]

Smith, L., J. Panter, and D. Ogilvie. 2019b. Characteristics of the Environment and Physical Activity in Midlife: Findings from UK Biobank. *Prev. Med.* **118**:150–158. [7]

Smith, P. B., S. Knox, and D. K. Benjamin Jr. 2018. Coordination of the Environmental influences on Child Health Outcomes (ECHO) Program: So the Whole Is Greater Than the Sum of Its Parts. *Curr. Opin. Pediatr.* **30**:263. [3]

Smuts, B. B. 2017. Sex and Friendship in Baboons. London: Routledge. [4]

Snijders, T. A. B. 2011. Multilevel Analysis. In: International Encyclopedia of Statistical Science, ed. M. Lovric, pp. 879–882. Heidelberg: Springer. [5]

Soliman, A., K. Soltani, J. Yin, A. Padmanabhan, and S. Wang. 2017. Social Sensing of Urban Land Use Based on Analysis of Twitter Users' Mobility Patterns. *PLOS ONE* **12**:e0181657. [7]

Somerville, R. C. J. 2012. Communicating the Science of Climate Change. *Phys. Today* **64**:48. [3]

Sondheim, M., K. Gardels, and B. K. 1999. GIS Interoperability. In: Geographical Information Systems, ed. P. A. Longley et al., pp. 347–358. New York: Wiley. [6]

Song, C., Z. Qu, N. Blumm, and A. Barabasi. 2008. Limits of Predictability in Human Mobility. *Science* **327**:1018–1021. [8]

Soutar, C., and A. P. F. Wand. 2022. Understanding the Spectrum of Anxiety Responses to Climate Change: A Systematic Review of the Qualitative Literature. *Int. J. Environ. Res. Public Health* **19**:990. [1]

Spadaro, A., A. Sarker, W. Hogg-Bremer, et al. 2022. Reddit Discussions about Buprenorphine Associated Precipitated Withdrawal in the Era of Fentanyl. *Clin. Toxicol.* **60**:694–701. [4]

Spencer, M. R., A. M. Miniño, and M. Warner. 2022. Drug Overdose Deaths in the United States, 2001–2021. *NCHS data brief* **457**:1–8. [10]

Statista. 2022a. Most Popular Social Networks Worldwide as of January 2022, Ranked by Number of Monthly Active Users. Statista. https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/ (accessed Nov. 2, 2022). [4]

———. 2022b. Number of Global Social Network Users 2018-2027. Statista. https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/ (accessed Nov. 2, 2022). [4]

———. 2022c. Number of Worldwide Social Network Users 2027. https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/ (accessed Oct. 19, 2022). [10]

Statistics Finland. 2021. Causes of Death. https://www.stat.fi/til/ksyyt/index_en.html (accessed Nov. 9, 2022). [11]

Statistics Norway. 2022. Access to Norwegian Government Microdata. https://www.ssb.no/en/data-til-forskning/utlan-av-data-til-forskere (accessed Nov. 9, 2022). [11]

Steinitz, C. 2016. Beginnings of Geodesign: A Personal Historical Perspective. *Res. Urban Ser.* **4**:9–24. [6]

Stephens-Davidowitz, S. I. 2013. The Cost of Racial Animus on a Black Presidential Candidate: Using Google Search Data to Find What Surveys Miss. https://ssrn.com/abstract=2238851 (accessed Nov. 7, 2022). [4]

Strickland, J. C., J. R. Havens, and W. W. Stoops. 2019. A Nationally Representative Analysis of "Twin Epidemics": Rising Rates of Methamphetamine Use among Persons Who Use Opioids. *Drug Alcohol Depend.* **204**:107592. [10]

Stuart, K., and E. J. Soulsby. 2011. Reducing Global Health Inequalities. Part 1. *J. R. Soc. Med.* **104**:321–326. [1]

Stürmer, S., M. Snyder, A. Kropp, and B. Siem. 2006. Empathy-Motivated Helping: The Moderating Role of Group Membership. *Pers. Soc. Psychol. Bull.* **32**:943–956. [9]

Stutzer, A., and B. S. Frey. 2008. Stress That Doesn't Pay: The Commuting Paradox. *The Scandinavian Journal of Economics* **110**:339–366. [8]

Suel, E., J. W. Polak, J. E. Bennett, and M. Ezzati. 2019. Measuring Social, Environmental and Health Inequalities Using Deep Learning and Street Imagery. *Sci. Rep.* **9**:6229. [1]

Sueur, C., O. Petit, A. De Marco, et al. 2011. A Comparative Network Analysis of Social Style in Macaques. *Anim. Behav.* **82**:845–852. [4]

Sukumaran, K., C. Cardenas-Iniguez, E. Burnor, et al. 2023. Ambient Fine Particulate Exposure and Subcortical Gray Matter Microarchitecture in 9- and 10-Year-Old Children across the United States. *iScience* **26**:106087. [1]

Sundhedsdatastyrelsen. 2022. The Danish Secure Research Platform. https://sundhedsdatastyrelsen.dk/da/english/health_data_and_registers/research_services/secure_research_platform (accessed Nov. 14, 2022). [11]

Sundström, J., M. Söderholm, S. Söderberg, et al. 2019. Risk Factors for Subarachnoid Haemorrhage: A Nationwide Cohort of 950 000 Adults. *Int. J. Epidemiol.* **48**:2018–2025. [11]

Svensson, A. C., S. Sandin, S. Cnattingius, et al. 2009. Maternal Effects for Preterm Birth: A Genetic Epidemiologic Study of 630,000 Families. *Am. J. Epidemiol.* **170**:1365–1372. [11]

Swedish Ethical Review Authority. 2023. Etikprövningsmyndigheten. https://etikprovningsmyndigheten.se/ (accessed Nov. 9, 2023). [11]

Sweet, M. 2011. Does Traffic Congestion Slow the Economy? *Journal of Planning Literature* **26**:391–404. [8]

Symington, M. M. 1990. Fission-Fusion Social Organization in*Ateles* and*Pan*. *Int. J. Primatol.* **11**:47–61. [4]

Sytsma, V. A., V. F. Chillar, and E. L. Piza. 2021. Scripting Police Escalation of Use of Force through Conjunctive Analysis of Body-Worn Camera Footage: A Systematic Social Observational Pilot Study. *J. Crim. Justice* **74**:101776. [9]

Taborsky, M. 2010. Sample Size in the Study of Behaviour. *Ethology* **116**:185–202. [9]

Talen, E. 1999. Sense of Community and Neighbourhood Form: An Assessment of the Social Doctrine of New Urbanism. *Urban Stud.* **36**:1361–1379. [8]

Talukdar, S., P. Singha, S. Mahato, et al. 2020. Land-Use Land-Cover Classification by Machine Learning Classifiers for Satellite Observations: A Review. *Remote Sens.* **12**:1135. [7]

Thisse, J.-F. 2018. Human Capital and Agglomeration Economies in Urban Development. *The Developing Economies* **56**:117–139. [8]

Thomas, W. I., and F. Znaniecki. 1918. The Polish Peasant in Europe and America. Boston: Gorham Press. [8]

Thompson, P. M., J. L. Stein, S. E. Medland, et al. 2014. The ENIGMA Consortium: Large-Scale Collaborative Analyses of Neuroimaging and Genetic Data. *Brain Imaging Behav.* **8**:153–182. [1]

Tiemeier, H., F. P. Velders, E. Szekely, et al. 2012. The Generation R Study: A Review of Design, Findings to Date, and a Study of the 5-HTTLPR by Environmental Interaction from Fetal Life Onward. *J. Am. Acad. Child Adolesc. Psychiatry* **51**:1119–1135 e1117. [1]

Tinbergen, N. 1951. The Study of Instinct. Oxford: Clarendon Press. [2]

———. 1952. The Curious Behavior of the Stickleback. *Sci. Am.* **187**:22–27. [2]

———. 1963. On Aims and Methods of Ethology. *Z. Tierpsychol.* **20**:410–433. [2, 9]

———. 1972. The Animal in Its World: Explorations of an Ethologist, 1932-1972, vol. 84. Cambridge, MA: Harvard Univ. Press. [2]

Ting, D. S. W., L. Carin, V. Dzau, and T. Y. Wong. 2020. Digital Technology and COVID-19. *Nat. Med.* **26**:459–461. [4]

To, P., E. Eboreime, and V. I. O. Agyapong. 2021. The Impact of Wildfires on Mental Health: A Scoping Review. *Behav. Sci. (Basel)* **11**:126. [1]

Tobler, W. R. 1970. A Computer Movie Simulating Urban Growth in the Detroit Region. *Econ. Geogr.* **46**:234. [6]

Tolonen, H., V. Salomaa, J. Torppa, et al. 2007. The Validation of the Finnish Hospital Discharge Register and Causes of Death Register Data on Stroke Diagnoses. *Eur. J. Cardiovasc. Prev. Rehabil.* **14**:380–385. [11]

Tomasello, M., and E. Herrmann. 2010. Ape and Human Cognition: What's the Difference? *Curr. Dir. Psychol. Sci.* **19**:3–8. [9]

Toro, R., G. Leonard, J. V. Lerner, et al. 2008. Prenatal Exposure to Maternal Cigarette Smoking and the Adolescent Cerebral Cortex. *Neuropsychopharmacol.* **33**:1019–1027. [1]

Travisi, C. M., R. Camagni, and P. Nijkamp. 2010. Impacts of Urban Sprawl and Commuting: A Modelling Study for Italy. *Journal of Transport Geography* **18**:382–392. [8]

Tsao, S.-F., H. Chen, T. Tisseverasinghe, et al. 2021. What Social Media Told Us in the Time of COVID-19: A Scoping Review. *Lancet Digit. Health* **3**:e175–e194. [4, 10]

Tsou, M.-H. 2004. Integrating Web-Based GIS and Image Processing Tools for Environmental Monitoring and Natural Resource Management. *J. Geogr. Syst.* **6**:155–174. [5]

Tuan, Y.-F. 1977. Space and Place: The Perspective of Experience. Minneapolis: Univ. Minnesota Press. [4]

Tukey, J. W. 1980. We Need Both Exploratory and Confirmatory. *Am. Stat.* **34**:23–25. [12]

Turiel, J., D. Fernandez-Reyes, and T. Aste. 2021. Wisdom of Crowds Detects COVID-19 Severity Ahead of Officially Available Data. *Sci. Rep.* **11**:13678. [10]

Turner, B. L., E. F. Lambin, and A. Reenberg. 2007. The Emergence of Land Change Science for Global Environmental Change and Sustainability. *PNAS* **104**:20666–20671. [7]

Turner, J. H., and A. Maryanski. 2018. Discovering Human Nature through Cross-Species Analysis. In: Handbook of Evolution, Biology, and Society, ed. R. L. Hopcroft, pp. 89–112. New York: Oxford Univ. Press. [9]

UN-Habitat. 2022. World Cities Report 2022 . Envisaging the Future of Cities. https://unhabitat.org/world-cities-report-2022-envisaging-the-future-of-cities (accessed Jan. 23, 2024). [8]

UN. 2016. The New Urban Agenda: UN Conference on Housing and Sustainable Urban Development (Habitat III). https://www.habitat3.org/the-new-urban-agenda (accessed Jan. 23, 2024). [8]

Unal, M., C. Uslu, and A. Cilek. 2016. GIS-Based Accessibility Analysis for Neighbourhood Parks: The Case of Cukurova District. *J. Digit. Landsc. Architect.* **1**:46–56. [6]

University of Wisconsin Population Health Institute. 2022. County Health Rankings National Findings 2022. https://www.countyhealthrankings.org/reports/2022-county-health-rankings-national-findings-report (accessed Jan. 23, 2024). [8]

Urban Indian Health Institute. 2021. Data Genocide of American Indians and Alaska Natives in COVID-19 Data. https://www.uihi.org/projects/data-genocide-of-american-indians-and-alaska-natives-in-covid-19-data/ (accessed Oct. 7, 2022). [5]

Usher, C. 2022. Eco-Anxiety. *J. Am. Acad. Child Adolesc. Psychiatry* **61**:341–342. [1]

Vable, A. M., S. F. Diehl, and M. M. Glymour. 2021. Code Review as a Simple Trick to Enhance Reproducibility, Accelerate Learning, and Improve the Quality of Your Team's Research. *Am. J. Epidemiol.* **190**:2172–2177. [3]

Vaisman, A., and E. Zimányi. 2014. Data Warehouse Systems: Design and Implementation. Data-Centric Systems and Applications, M. J. Carey and S. Ceri, series ed. Heidelberg: Springer. [3]

Valdano, E., J. T. Okano, V. Colizza, H. K. Mitonga, and S. Blower. 2021. Using Mobile Phone Data to Reveal Risk Flow Networks Underlying the HIV Epidemic in Namibia. *Nat. Commun.* **12**:2837. [4]

van der Linden, E. L., K. A. C. Meeks, K. Klipstein-Grobusch, et al. 2022. Hypertension Determinants among Ghanaians Differ According to Location of Residence: Rodam Study. *J. Hypertens.* **40**:1010–1018. [4]

Van Holle, V., B. Deforche, J. Van Cauwenberg, et al. 2012. Relationship between the Physical Environment and Different Domains of Physical Activity in European Adults: A Systematic Review. *BMC Public Health* **12**:807–807. [7]

Van Horne, Y. O., C. S. Alcala, R. E. Peltier, et al. 2023. An Applied Environmental Justice Framework for Exposure Science. *J. Expo. Sci. Environ. Epidemiol.* **33**:1–11. [1]

van Kleef, G. A., F. Wanders, E. Stamkou, and A. C. Homan. 2015. The Social Dynamics of Breaking the Rules: Antecedents and Consequences of Norm-Violating Behavior. *Curr. Opin. Psychol.* **6**:25–31. [4]

van Leeuwen, F., J. H. Park, and I. S. Penton-Voak. 2012. Another Fundamental Social Category? Spontaneous Categorization of People Who Uphold or Violate Moral Norms. *J. Exp. Soc. Psychol.* **48**:1385–1388. [4]

Vandenbroucke, J. P. 1990. Epidemiology in Transition: A Historical Hypothesis. *Epidemiology* **1**:164–167. [1]

Vatsalan, D., Z. Sehili, P. Christen, and E. Rahm. 2017. Privacy-Preserving Record Linkage for Big Data: Current Approaches and Research Challenges. In: Handbook of Big Data Technologies, ed. A. Y. Zomaya and S. Sakr, pp. 851–895. Cham: Springer. [12]

Venerandi, A., G. Quattrone, and L. Capra. 2018. A Scalable Method to Quantify the Relationship between Urban Form and Socio-Economic Indexes. *EPJ Data Sci.* **7**:4. [8]

Venter, Z. S., C. Shackleton, A. Faull, et al. 2022. Is Green Space Associated with Reduced Crime? A National-Scale Study from the Global South. *Sci.Total Environ.* **825**:154005. [8]

Verbeek, P. 2008. Peace Ethology. *Behaviour* **145**:1497–1524. [9]

Verhoeven, J. C. 1993. An Interview with Erving Goffman, 1980. *Res. Lang. Soc. Interact.* **26**:317–348. [9]

Villalonga-Olives, E., and I. Kawachi. 2017. The Dark Side of Social Capital: A Systematic Review of the Negative Health Effects of Social Capital. *Soc. Sci. Med.* **194**:105–127. [4]

Villermé, L.-R. 2008. De la Mortalité Dans Les Divers Quartiers de la Ville de Paris (1830). Paris: La Fabrique. [4]

Viswanathan, M., A. Ammerman, E. Eng, et al. 2004. Community-Based Participatory Research: Assessing the Evidence. *Evid Rep Technol Assess* **99**:1–8. [5]

Volkow, N. D. 2020. Stigma and the Toll of Addiction. *N. Eng. J. Med.* **382**:1289–1290. [10]

Volkow, N. D., and C. Blanco. 2021. Research on Substance Use Disorders during the COVID-19 Pandemic. *J. Subst. Abuse Treat.* **129**:108385. [10]

Wachter, S. 2022. The Theory of Artificial Immutability: Protecting Algorithmic Groups under Anti-Discrimination Law. *Tulane Law Rev.* **97**:1–49. [5]

Waddington, C. H. 1957. The Strategy of the Genes: A Discussion of Some Aspects of Theoretical Biology. London: Routledge. [2]

Wagenmakers, E.-J., R. Wetzels, D. Borsboom, H. L. J. van der Maas, and R. A. Kievit. 2012. An Agenda for Purely Confirmatory Research. *Perspect. Psychol. Sci.* **7**:632–638. [12]

Waldron, I. 2018. There's Something in the Water: Environmental Racism in Indigenous & Black Communities. Winnipeg: Fernwood Publ. [1]

Wang, D., T. Abdelzaher, and L. Kaplan. 2015. Social Sensing: Building Reliable Systems on Unreliable Data. Burlington, MA: Morgan Kaufmann. [7]

Wang, F. 2020. Why Public Health Needs GIS: A Methodological Overview. *Ann. GIS* **26**:1–12. [6]

Wang, H., and Y. Yang. 2019. Neighbourhood Walkability: A Review and Bibliometric Analysis. *Cities* **93**:43–61. [3]

Wang, K. 2023. Planning for Livability? State-Built New Towns and Urban Traffic Externalities in China. *Asian Economic Papers* **22**:96–117. [8]

Wang, L., B. I. Graubard, H. A. Katki, and Y. Li. 2020. Improving External Validity of Epidemiologic Cohort Analyses: A Kernel Weighting Approach. *J. R. Stat. Soc. Ser. A Stat. Soc.* **183**:1293–1311. [11]

Wang, L., C. C. Li, Q. Ying, et al. 2012. China's Urban Expansion from 1990 to 2010 Determined with Satellite Remote Sensing. *Chinese Sci. Bull.* **57**:2802–2812 [7]

Wang, M., and N. Debbage. 2021. Urban Morphology and Traffic Congestion: Longitudinal Evidence from US Cities. *Comput. Environ. Urban Syst.* **89**:101676. [8]

Weenink, D., R. Dhattiwala, and D. van der Duin. 2022. Circles of Peace: A Video Analysis of Situational Group Formation and Collective Third-Party Intervention in Violent Incidents. *Br. J. Criminol.* **62**:18–36. [9]

Wei, E. 2019. Status as a Service (StaaS). https://www.eugenewei.com/blog/2019/2/19/status-as-a-service (accessed Nov. 2, 2022). [4]

Weichenthal, S., M. Hatzopoulou, and M. Brauer. 2019. A Picture Tells a Thousand... Exposures: Opportunities and Challenges of Deep Learning Image Analyses in Exposure Science and Environmental Epidemiology. *Environ. Int.* **122**:3–10. [1]

Weinstein, L. 2021. Evictions: Reconceptualizing Housing Insecurity from the Global South. *City Commun.* **20**:13–23. [3]

Weisskircher, M. 2020. The Strength of Far-Right AfD in Eastern Germany: The East-West Divide and the Multiple Causes Behind "Populism". *Politic. Q.* **91**:614–622. [4]

Wells, D. 2009. Agile Software Development: A Gentle Introduction. http://www.agile-process.org/ (accessed Nov. 9, 2022). [12]

Weng, H., J. Lin, R. Martin, et al. 2020. Global High-Resolution Emissions of Soil NOx, Sea Salt Aerosols, and Biogenic Volatile Organic Compounds. *Sci. Data* **7**:1–15. [3]

West, G. 2017. Scale: The Universal Laws of Growth, Innovation, Sustainability, and the Pace of Life in Organisms, Cities, Economies, and Companies. New York: Penguin Press. [5]

Westerholt, R. 2019. Methodological Aspects of the Spatial Analysis of Geosocial Media Feeds: From Locations Towards Places. *GIS.Science* **2**:65–76. [6]

Wheaton, W. C., and H. Shishido. 1981. Urban Concentration, Agglomeration Economies, and the Level of Economic Development. *Economic Development and Cultural Change* **30**:17–30. [8]

WHO. 2017. WHO Guidelines on Ethical Issues in Public Health Surveillance. Geneva: World Health Organization. [5]

Wild, C. P. 2005. Complementing the Genome with an "Exposome": The Outstanding Challenge of Environmental Exposure Measurement in Molecular Epidemiology. *Cancer Epidemiol. Biomarkers Prev.* **14**:1847–1850. [1]

Wilkinson, M. D., M. Dumontier, I. J. Aalbersberg, et al. 2016. The FAIR Guiding Principles for Scientific Data Management and Stewardship. *Sci. Data* **3**:160018. [5]

Williams, B. 2023. Global Cities Unveiled: Understanding the Transformation of Urban Spaces in a Connected World: Amazon. [8]

Williams, D. S., M. Máñez Costa, C. Sutherland, L. Celliers, and J. Scheffran. 2019. Vulnerability of Informal Settlements in the Context of Rapid Urbanization and Climate Change. *Environment and Urbanization* **31**:157–176. [8]

Wilson, G., J. Bryan, K. Cranston, et al. 2017a. Good Enough Practices in Scientific Computing. *PLoS Comput. Biol.* **13**:e1005510. [3]

Wilson, R., O. Butters, D. Avraam, et al. 2017b. Datashield: New Directions and Dimensions. http://datascience.codata.org/articles/10.5334/dsj-2017-021/ (accessed Nov. 14, 2022). [11]

Wolfson, M., S. E. Wallace, N. Masca, et al. 2010. Datashield: Resolving a Conflict in Contemporary Bioscience: Performing a Pooled Analysis of Individual-Level Data without Sharing the Data. *Int. J. Epidemiol.* **39**:1372–1382. [11]

WorldBank. 2023. Urban Development. World Bank. https://www.worldbank.org/en/topic/urbandevelopment/overview (accessed Jan. 23, 2024). [8]

Wulder, M. A., T. R. Loveland, D. P. Roy, et al. 2019. Current Status of Landsat Program, Science, and Applications. *Remote Sens. Environ.* **225**:127–147. [7]

Xafis, V., G. O. Schaefer, M. K. Labude, et al. 2019. An Ethics Framework for Big Data in Health and Research. *Asian Bioeth. Rev.* **11**:227–254. [5]

Xu, J., X. Liu, Q. Li, et al. 2022a. Global Urbanicity Is Associated with Brain and Behaviour in Young People. *Nat. Hum. Behav.* **6**:279–293. [1]

Xu, Z., W. Wang, Q. Liu, et al. 2022b. Association between Gaseous Air Pollutants and Biomarkers of Systemic Inflammation: A Systematic Review and Meta-Analysis. *Environ. Pollut.* **292**:118336. [1]

Yang, Y.-C., M. A. Al-Garadi, J. S. Love, et al. 2023. Can Accurate Demographic Information about People Who Use Prescription Medications Non-Medically Be Derived from Twitter? *PNAS* **120**:e2207391120. [10]

Yang, Y.-C., M. A. Al-Garadi, J. S. Love, J. Perrone, and A. Sarker. 2021. Automatic Gender Detection in Twitter Profiles for Health-Related Cohort Studies. *JAMIA Open* **4**:ooab042. [4]

Yang, Y., N. V. Chawla, and B. Uzzi. 2019. A Network's Gender Composition and Communication Pattern Predict Women's Leadership Success. *PNAS* **116**:2033–2038. [4]

Yeo, J., S. Park, and K. Jang. 2015. Effects of Urban Sprawl and Vehicle Miles Traveled on Traffic Fatalities. *Traffic Injury Prevention* **16**:397–403. [8]

Yin, J., J. Dong, N. A. S. Hamm, et al. 2021. Integrating Remote Sensing and Geospatial Big Data for Urban Land Use Mapping: A Review. *Int. J. Appl. Earth Obs. Geoinf.* **103**:102514. [7]

Yin, L., and S.-L. Shaw. 2015. Exploring Space–Time Paths in Physical and Social Closeness Spaces: A Space–Time GIS Approach. *Int. J. Geographic. Inf. Sci.* **29**:742–761. [6]

Zerbo, A., R. C. Delgado, and P. A. González. 2020. Vulnerability and Everyday Health Risks of Urban Informal Settlements in Sub-Saharan Africa. *Global Health Journal* **4**:46–50. [8]

Zhan, X., S. V. Ukkusuri, and F. Zhu. 2014. Inferring Urban Land Use Using Large-Scale Social Media Check-in Data. *Netw. Spat. Econ.* **14**:647–667. [7]

Zhang, C., D. Song, C. Huang, A. Swami, and N. V. Chawla. 2019. Heterogeneous Graph Neural Network. In: Proc. of the ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, pp. 793–803. New York: ACM. [7]

Zhang, F., Z. Fan, Y. Kang, Y. Hu, and C. Ratti. 2021. "Perception Bias": Deciphering a Mismatch between Urban Crime and Perception of Safety. *Landsc. Urban Plann.* **207**:104003. [8]

Zhang, F., B. Zhou, L. Liu, et al. 2018. Measuring Human Perceptions of a Large-Scale Urban Region Using Machine Learning. *Landsc. Urban Plann.* **180**:148–160. [7]

Zhang, J., W. Wang, F. Xia, Y.-R. Lin, and H. Tong. 2020. Data-Driven Computational Social Science: A Survey. *Big Data Res.* **21**:100145. [9]

Zhang, L., and M. Menendez. 2020. Modeling and Evaluating the Impact of City Block Size on Traffic Performance. *J. Urban Plann. Dev.* **146**:04020021. [8]

Zhou, Y., X. Y. Li, G. R. Asrar, S. J. Smith, and M. Imhoff. 2018. A Global Record of Annual Urban Dynamics (1992–2013) from Nighttime Lights. *Remote Sens. Environ.* **219**:206–220. [7]

Zhu, Z., M. A. Wulder, D. P. Roy, et al. 2019. Benefits of the Free and Open Landsat Data Policy. *Remote Sens. Environ.* **224**:382–385. [7]

Zimmer, M. 2018. Addressing Conceptual Gaps in Big Data Research Ethics: An Application of Contextual Integrity. *Social Media Society* **4**:10.1177/2056305118768300. [12]

Zwirner, E., and N. Raihani. 2020. Neighbourhood Wealth, Not Urbanicity, Predicts Prosociality Towards Strangers. *Proc. R. Soc. B* **287**:1936. [8]

Subject Index

Strüngmann Forum Report Series*

Migration Stigma: Understanding Prejudice, Discrimination, and Exclusion
Edited by Lawrence H. Yang, Maureen A. Eger and Bruce G. Link
ISBN: 9780262548120

Exploring and Exploiting Genetic Risk for Psychiatric Disorders
Edited by Joshua A. Gordon and Elisabeth Binder
DOI: https://doi.org/10.7551/mitpress/15380.001.0001
ISBN electronic: 9780262377423

Intrusive Thinking: From Molecules to Free Will
Edited by Peter W. Kalivas and Martin P. Paulus
ISBN: 9780262542371

Deliberate Ignorance: Choosing Not to Know
edited by Ralph Hertwig and Christoph Engel
ISBN 9780262045599

Youth Mental Health: A Paradigm for Prevention and Early Intervention
Edited by Peter J. Uhlhaas and Stephen J. Wood
ISBN: 9780262043977

The Neocortex
Edited by Wolf Singer, Terrence J. Sejnowski and Pasko Rakic
ISBN: 9780262043243

Interactive Task Learning: Humans, Robots, and Agents Acquiring New Tasks through Natural Interactions
Edited by Kevin A. Gluck and John E. Laird
ISBN: 9780262038829

Agrobiodiversity: Integrating Knowledge for a Sustainable Future
Edited by Karl S. Zimmerer and Stef de Haan
ISBN: 9780262038683

Rethinking Environmentalism: Linking Justice, Sustainability, and Diversity
Edited by Sharachchandra Lele, Eduardo S. Brondizio, John Byrne,
Georgina M. Mace and Joan Martinez-Alier
ISBN: 9780262038966

Emergent Brain Dynamics: Prebirth to Adolescence
Edited by April A. Benasich and Urs Ribary
ISBN: 9780262038638

The Cultural Nature of Attachment: Contextualizing Relationships and Development
Edited by Heidi Keller and Kim A. Bard
ISBN (Hardcover): 9780262036900 ISBN (ebook): 9780262342865
Winner of the Ursula Gielen Global Psychology Book Award

* Available at https://mitpress.mit.edu/books/series/strungmann-forum-reports

Investors and Exploiters in Ecology and Economics: Principles and Applications
edited by Luc-Alain Giraldeau, Philipp Heeb and Michael Kosfeld
ISBN (Hardcover): 9780262036122 ISBN (eBook): 9780262339797

Computational Psychiatry: New Perspectives on Mental Illness
edited by A. David Redish and Joshua A. Gordon
ISBN: 9780262035422

Complexity and Evolution: Toward a New Synthesis for Economics
edited by David S. Wilson and Alan Kirman
ISBN: 9780262035385

The Pragmatic Turn: Toward Action-Oriented Views in Cognitive Science
edited by Andreas K. Engel, Karl J. Friston and Danica Kragic
ISBN: 9780262034326

Translational Neuroscience: Toward New Therapies
edited by Karoly Nikolich and Steven E. Hyman
ISBN: 9780262029865

Trace Metals and Infectious Diseases
edited by Jerome O. Nriagu and Eric P. Skaar
ISBN 9780262029193

Pathways to Peace: The Transformative Power of Children and Families
edited by James F. Leckman, Catherine Panter-Brick and Rima Salah,
ISBN 9780262027984

Rethinking Global Land Use in an Urban Era
edited by Karen C. Seto and Anette Reenberg
ISBN 9780262026901

Schizophrenia: Evolution and Synthesis
edited by Steven M. Silverstein, Bita Moghaddam and Til Wykes,
ISBN 9780262019620

Cultural Evolution: Society, Technology, Language, and Religion
edited by Peter J. Richerson and Morten H. Christiansen,
ISBN 9780262019750

Language, Music, and the Brain: A Mysterious Relationship
edited by Michael A. Arbib
ISBN 9780262019620

Evolution and the Mechanisms of Decision Making
edited by Peter Hammerstein and Jeffrey R. Stevens
ISBN 9780262018081

Cognitive Search: Evolution, Algorithms, and the Brain
edited by Peter M. Todd, Thomas T. Hills and Trevor W. Robbins,
ISBN 9780262018098

Animal Thinking: Contemporary Issues in Comparative Cognition
edited by Randolf Menzel and Julia Fischer
ISBN 9780262016636

Disease Eradication in the 21st Century: Implications for Global Health
edited by Stephen L. Cochi and Walter R. Dowdle
ISBN 9780262016735

Better Doctors, Better Patients, Better Decisions: Envisioning Health Care 2020
edited by Gerd Gigerenzer and J. A. Muir Gray
ISBN 9780262016032

Dynamic Coordination in the Brain: From Neurons to Mind
edited by Christoph von der Malsburg, William A. Phillips and Wolf Singer,
ISBN 9780262014717

Linkages of Sustainability
edited by Thomas E. Graedel and Ester van der Voet
ISBN 9780262013581

Biological Foundations and Origin of Syntax
edited by Derek Bickerton and Eörs Szathmáry
ISBN 9780262013567

Clouds in the Perturbed Climate System: Their Relationship to Energy Balance, Atmospheric Dynamics, and Precipitation
edited by Jost Heintzenberg and Robert J. Charlson
ISBN 9780262012874
Winner of the Atmospheric Science Librarians International Choice Award

Better Than Conscious? Decision Making, the Human Mind, and Implications For Institutions
edited by Christoph Engel and Wolf Singer
ISBN 978-0-262-19580-5