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

From "Digital Ethology: Human Behavior in Geospatial Context,"
edited by Tomáš Paus and Hye-Chung Kum. Strüngmann Forum Reports, vol. 33,
Julia R. Lupp, series editor. Cambridge, MA: MIT Press. ISBN 978026254813

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

From "Digital Ethology: Human Behavior in Geospatial Context,"
edited by Tomáš Paus and Hye-Chung Kum. Strüngmann Forum Reports, vol. 33,
Julia R. Lupp, series editor. Cambridge, MA: MIT Press. ISBN 978026254813

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.

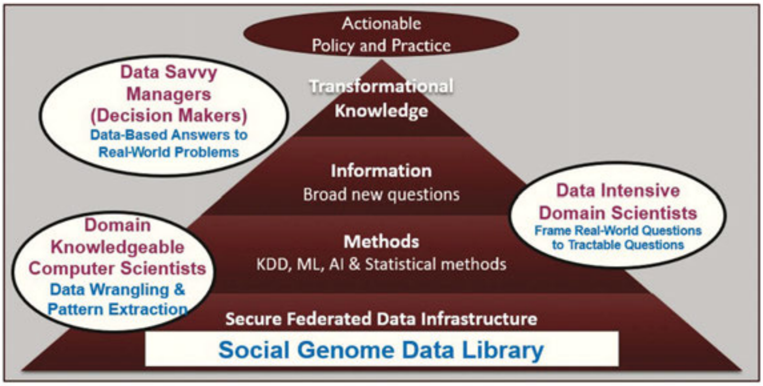

Figure 12.1 Data, information, knowledge, and action (DIKA) pyramid. KDD: Knowledge Discovery and Data Mining; ML: machine learning; AI: artificial intelligence.

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

From "Digital Ethology: Human Behavior in Geospatial Context,"
edited by Tomáš Paus and Hye-Chung Kum. Strüngmann Forum Reports, vol. 33,
Julia R. Lupp, series editor. Cambridge, MA: MIT Press. ISBN 978026254813

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.

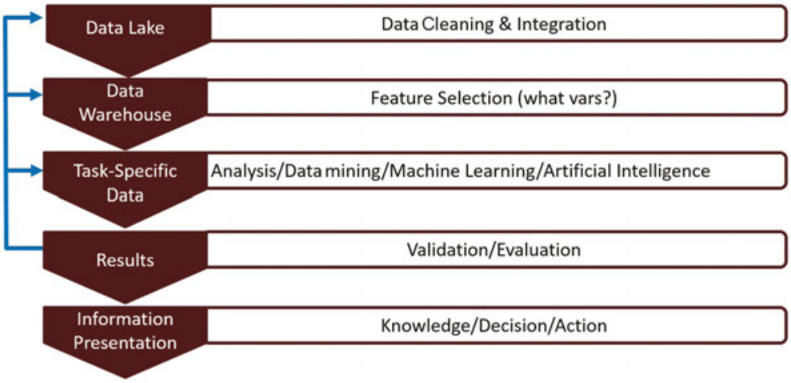
Data Lake
Data Cleaning & Integration
Data Warehouse
Feature Selection (what vars?)
Task-Specific Data
Analysis/Data mining/Machine Learning/Artificial Intelligence
Results
Validation/Evaluation
Information Presentation
Knowledge/Decision/Action

Figure 12.2 Knowledge discovery and data mining process.

From "Digital Ethology: Human Behavior in Geospatial Context,"
edited by Tomáš Paus and Hye-Chung Kum. Strüngmann Forum Reports, vol. 33,
Julia R. Lupp, series editor. Cambridge, MA: MIT Press. ISBN 978026254813

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.

From "Digital Ethology: Human Behavior in Geospatial Context,"
edited by Tomáš Paus and Hye-Chung Kum. Strüngmann Forum Reports, vol. 33,
Julia R. Lupp, series editor. Cambridge, MA: MIT Press. ISBN 978026254813

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

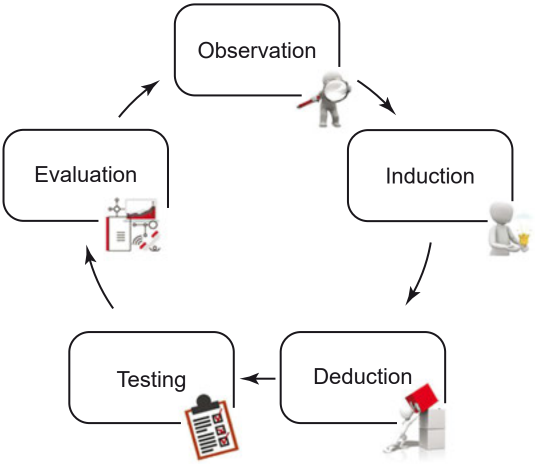

Figure 12.3 The empirical research cycle (De Groot and Spiekerman 1969).

From "Digital Ethology: Human Behavior in Geospatial Context,"
edited by Tomáš Paus and Hye-Chung Kum. Strüngmann Forum Reports, vol. 33,
Julia R. Lupp, series editor. Cambridge, MA: MIT Press. ISBN 978026254813

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

From "Digital Ethology: Human Behavior in Geospatial Context,"
edited by Tomáš Paus and Hye-Chung Kum. Strüngmann Forum Reports, vol. 33,
Julia R. Lupp, series editor. Cambridge, MA: MIT Press. ISBN 978026254813

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).

From "Digital Ethology: Human Behavior in Geospatial Context,"
edited by Tomáš Paus and Hye-Chung Kum. Strüngmann Forum Reports, vol. 33,
Julia R. Lupp, series editor. Cambridge, MA: MIT Press. ISBN 978026254813

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.