Data Acumen in Action

Sallie Ann Keller and Stephanie Shipp

1. The Need for Data Acumen

Advances in information technology, computation, applied mathematics, and statistics make it possible to access, integrate, and analyze massive amounts of data to support a wide range of applications. Open source software platforms, tools, and libraries have democratized access to many different types of data making the repurposing of data from multiple sources common practice. While these advances help solve numerous complex problems, they also greatly increase the need for data acumen and sensitivity to ethical challenges in using and integrating data.

This article discusses the necessity for data acumen and data science ethics and describes the fundamentals necessary to apply these skills to real-world problems. Whereas the examples and case study used to demonstrate the concepts are public-good problems at the intersection of data science and the social and behavioral sciences, they are generalizable to other disciplines, including engineering and physical and life sciences.

The desire to apply data to solving societal problems is not new, though modern technology has brought new challenges. As data have become more available and prevalent in the general public, many terms have evolved to describe a broad understanding of the role for data, analyses, and statistics in a modern democratic society. One popular term, data literacy, moved from a focus on the ability to collect, store, retrieve, and view data to a concern about the ability to comprehend and apply data in the required context. Additional terms have come into use including quantitative literacy, statistical literacy, statistical thinking, and numeracy [Sch03, PR11].

Historically, these literacy terms have focused on the ability to do and understand statistics, with an emphasis on basic quantitative reasoning. In the 1990s to early 2000s, data literacy (and related terms) meant knowledge and understanding sufficient to lead, use, and inform decision-making. These early insights foreshadow data acumen, noting that quantitative literacy "empowers people by giving them tools to think for themselves, to ask intelligent questions of experts, and to confront authority confidently" [Ste99, p. 2] and that "statistical thinking is essential for anyone who wants to be an informed citizen, intelligent consumer, or skilled worker" [Sch03, p. 146]. We subsume all of these terms under the rubric of data science. As data science took hold over the next decades,

Sallie Ann Keller is an endowed Distinguished Professor in Biocomplexity, director of the Social and Decision Analytics Division within the Biocomplexity Institute at the University of Virginia and professor of Public Health Sciences. Her email address is sak9tr@virginia.edu.

Stephanie Shipp is the deputy director and professor at the Social and Decision Analytics Division within the Biocomplexity Institute at the University of Virginia. Her email address is sss5sc@virginia.edu.

Communicated by Notices *Associate Editor Richard Levine.*

For permission to reprint this article, please contact: reprint-permission@ams.org.

DOI: https://doi.org/10.1090/noti2353

the definition and terminology has matured to mean the collection of abilities to do data science and the central role of data science ethics.

The need for additional thinking around data literacy stems for the emergence of data science. Borrowing from statistics, operations research, and computer science [Ros16], we describe ***data science as learning from data in the context of problems***. Bringing data to bear on real problems is the genius of this new transdisciplinary field—without the presence of a problem, there is no need for the application of data, hence no data science. Problem solutions require carefully framed questions, and a rigorous, replicable, and repeatable process to scientifically address the problems. This must include clear dissemination of results, the choices made in the development of the data science methods applied, ethical reviews, and an assessment of how well the solution works in real life. This is the data science life-cycle process [Win19, KSSK20].

This rich-problem focus of data science lends itself to a modern expression of literacy called data acumen [oS18]. **Data acumen** is the ***"ability to make good judgements about the use of data to support problem solutions"*** [KSSK20]. Data acumen focuses on the need for understanding the data, tools, and problem solutions to make good decisions using data. Data acumen ranges from basic data literacy to acquiring quantitative and computational skills for doing data science. Three types of data acumen are subsumed under the concept, although data acumen is, in fact, a continuum of understanding ranging from sufficient knowledge to ask the right questions and understand the answers to deep knowledge and experience addressing data science problems [Gar19]. In fact, a person might be a subject matter expert, a data scientist, or at times a consumer. Different levels of data acumen span the type of expertise and use of data science, which may vary depending on the question being asked. These roles include:

- Subject matter or domain experts trained in a specific field, e.g., public health, engineering, or political science, who also have a range of training in data analytics capabilities. These individuals are able to advise on the subject area nuances of the problems, placing them in the proper context.
- Data scientists trained in quantitative sciences to include statistics, computer science, quantitative social sciences, operations research, and applied mathematics. These individuals have acquired the collection of skills giving them the abilities to do data science, often each with their own specialties.
- Consumers, stakeholders, or the recipients of the data science applications to support decisions, include people such as policymakers, community leaders, company managers, organizational administrators, and ordinary citizens. These people need to translate the results into actions.

Across all levels, applying data acumen involves asking a series of questions about the situation and process. These questions include:

- What is the **problem**? What are the **questions**?
- What **kinds of decisions** need to be made?
- What **data are available or may need to be collected**? Why should the **data be trusted**?
- What **types of analyses** are needed to inform the decisions?
- What are the **ethical considerations**?
- What part of the **data science life-cycle process** is being discussed?

Rarely will one person be able to provide the best answers to each of these questions. This is consistent with the notion that data science is a team activity. Most situations of significant impact require some cooperation and collaboration among individuals with differing types of expertise and levels of data acumen. That cooperation is likely to work most efficiently and effectively if all parties have some fundamental knowledge of data acumen—some understanding of what data are, the kinds of decisions necessary to be made, and the analytical frameworks to reach some set of reasonable solutions. Examples and a case study below help to explain these ideas.

1.1. Fundamentals of data acumen.

To accomplish the goals of data acumen some fundamental elements must be present. Data acumen requires an understanding of (1) data, (2) the kinds of decisions to be made, and (3) the analytical methods to inform the different kinds of decisions.

1.1.1. Data. To apply data acumen, one must have a clear understanding of what constitutes data and what data are relevant. ***"Data are any facts assumed to be a matter of direct observation"*** [Dat95]. Thus, data derive from three components—observations, problems and questions, and measurement. Humans have been making observations since they came into existence. When we desire to use observations to support problem solutions, we observe with intent. To use these observations in a meaningful way requires that the problems are translated into concrete research questions and observations sought that relate to the questions. Measurement is needed to create facts from observation and thus create data [KSSK20].

Over time, the tools to ask questions, observe, measure, and record to create data have become increasingly sophisticated.

With all the information that can be observed today and translated into data, the question of relevance becomes critical. More data makes the solution more effective only if the data are relevant. The primary determinant of relevance is whether the data address the decision situation and decision requirements. Determination of relevance can be achieved by answering the data acumen questions

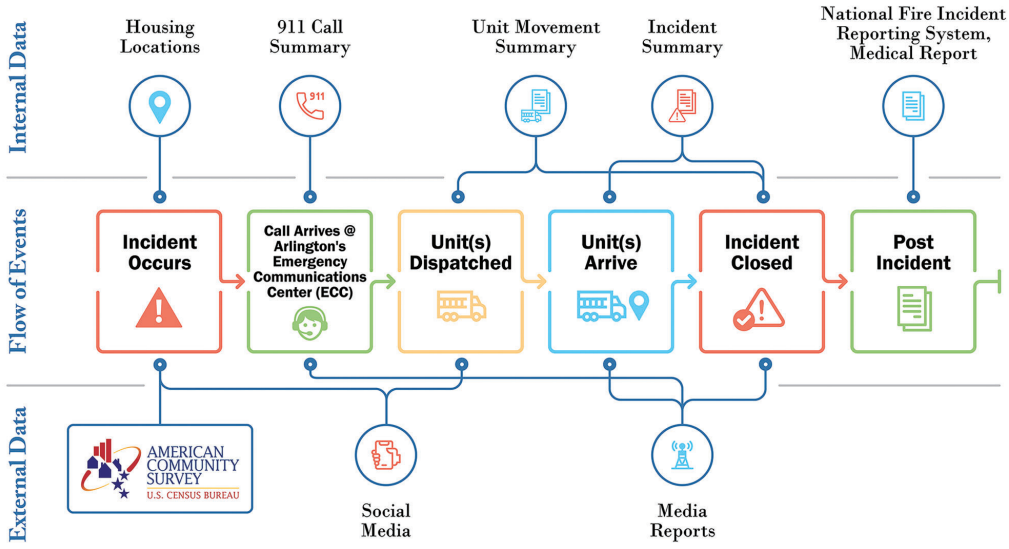

Figure 1. Information flow for Arlington County, Virginia, Fire Department.

above. Relevance is a key part of our data science lifecycle process (described below). Our data discovery process identifies and screens for relevance all possible data sources (i.e., survey, designed experiments, opportunity (e.g., social media), and procedural data) that might be used to address the problem being addressed [KSSK20]. Only after such screening can data gaps be identified and new data collection developed, if needed.

Example (Creating data for Arlington County, Virginia, Fire Department). The Fire Chief in Arlington, Virginia, wanted to gain better situational awareness on how well the Fire/Emergency Management Service (EMS) was serving the residents, particularly the vulnerable populations. The county had silos of fire/EMS operational systems where information flowed during the course of an incident to manage the deployment of resources and respond to state and national reporting as shown in Figure 1. The systems were not designed with unique identifiers to connect the information across the systems for the incidents.

To turn this information into data required statistically linking across these systems and recreating the incidents over time and geographic locations. To make these data useful for addressing issues associated with vulnerable populations, more information was needed. We linked the data by matching on time, date, and location of incident and integrated American Community Survey socioeconomic data to describe Arlington neighborhoods and social media data to provide context (see Figure 1) [KLS17, KSK+18]. By linking these data, the Fire Chief had a corpus of data to answer these questions as well as others about types of incidents by season, when special events occurred, and by neighborhood.

1.1.2. Decision requirements. The next fundamental of data acumen is an understanding of the decision(s) that must be made. To develop good and relevant data, we must understand the context of the situation and the kinds of decisions. Decision-making involves three kinds of decisions [oNI19]:

- **Strategic** decision-making informs and enriches understanding of problems and issues to plan for future operations. It involves developing an understanding of the context, knowledge, and intelligence to support planning. The steps to develop a strategic plan include identifying future capabilities, activities, and expectations for change and evaluating it against the current capabilities to identity gaps.
- **Anticipatory** decision-making is used to detect, identify, and plan for emerging issues and discontinuities. It involves collecting and analyzing information to identify emerging trends, changing conditions, and unexpected consequences of these trends and changes. The process is to question assumptions and advance new perspectives from many disciplines, as well as identify new opportunities and risks.
- **Tactical** decision-making supports ongoing actions and current operations.

Underlying these kinds of decision-making is cultivating the acumen to ask questions to gain clarity and at the same time accept uncertainty as the norm.

Example (Decision-making for Arlington County, Virginia, Fire Department). Continuing with the above example, the Fire Chief in Arlington, Virginia, also wanted to

use Fire/Emergency Management Service (EMS) data for both strategic and anticipatory decision-making concerning resource utilization. As shown in Figure 1, and collaborating with the Fire Chief and his staff, we identified a corpus of information that could be used to *strategically* support future force and resource development and at the same time better *anticipate* to plan the weekly allocation of resources across the county. Once linked, the data provided valuable insights for unit utilization as shown in Figure 2. The Fire Chief's focus was to ensure that the limited number of medical units were in the right place at the right time [KLS17].

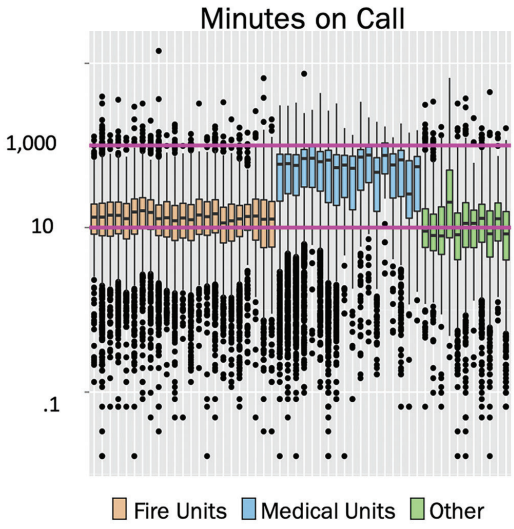

Figure 2. Time on call. The chart illustrates the linked data from Figure 1 for the distributions of minutes on call (in log base 10 scale) for fire (gold), medical (blue), and other (green). Medical units are out longer on call and require increased anticipatory planning to ensure they are located where needed.

1.1.3. Taxonomy of analyses. Data acumen requires understanding the types of analyses being applied to support the different kinds of decisions described above. We find it useful to think about five types of analysis—descriptive; explanatory; predictive; modeling, simulation, and analysis (MS&A); and prescriptive analysis. These definitions are described below and the relationship among each of them is presented in Figure 3 [oS17].

Descriptive analytics refers to the various ways to explore data. It is a set of processes and technologies that summarize data to infer what is happening now or what has happened in the past. This type of analytics is used to identify patterns, outliers, and other factors. Querying, frequencies, distributions, tabulations, visualizations, and geospatial maps are areas of descriptive analytics. **Explanatory** analytics move beyond describing the data or underlying population and develop statistical models for **inference** and testing **causal hypotheses** such as the likelihood of events or outcomes occurring.

In contrast to descriptive analytics that investigate the present or past, **predictive analytics** provide a forward-looking perspective. Predictive analytics rely on statistical and mathematical models to predict new or future outcomes. **Modeling and Simulation** is a popular form of prediction useful for exploring what-if questions to study current and future scenarios. It is one type of predictive analysis that can be used to develop and analyze well parameterized plans, options, or possible actions. The approach relies on descriptive analysis of the population to understand the current situation being studied and the specific future scenario being proposed. Modeling and Simulation methods do not automatically identify the best outcomes from the perspective of optimization and do not mathematically evaluate the quality or implications of future actions.

Prescriptive analytics are less mature than the other forms of analytics, but extremely exciting [LBAM20]. They combine prediction and scenario options to identify the best, mathematically optimal, solutions. Sometimes called optimization under uncertainty, prescriptive analytics methods use stochastic modeling and mathematical optimization to understand why a particular future could happen and what actions should be taken.

Example (Analytics support for Arlington County, Virginia, Fire Department). Building on the example in Figures 1 and 2, the Arlington County, Virginia, Fire Chief wanted to identify potential differences in response time by fire stations and unit dispatches. National policy standards state that the first fire engine should reach an incident within four minutes and other apparatus units (e.g., medical units such as basic or advanced life support transport units and rescue units) within eight minutes.

A descriptive analysis did not find statistically significant differences across fire stations. This is because when the data were pooled across types of incidents the variability in the data masked potential differences due to the apparatus types that were deployed. An explanatory analysis was developed by fitting a spatial regression model patterned after [GH18] to control for apparatus type and station [AGK+21]. It was this statistical controlling for apparatus type and station that surfaced significant differences in response times across several stations (see Figure 4). The Fire Chief used this information to inform his strategic plans for locations of new fire stations and anticipatory planning about changes in procedures to decrease time to incidents [AGK+21].

1.2. Data acumen in the data science life-cycle process.

To facilitate the rigorous application of data science to

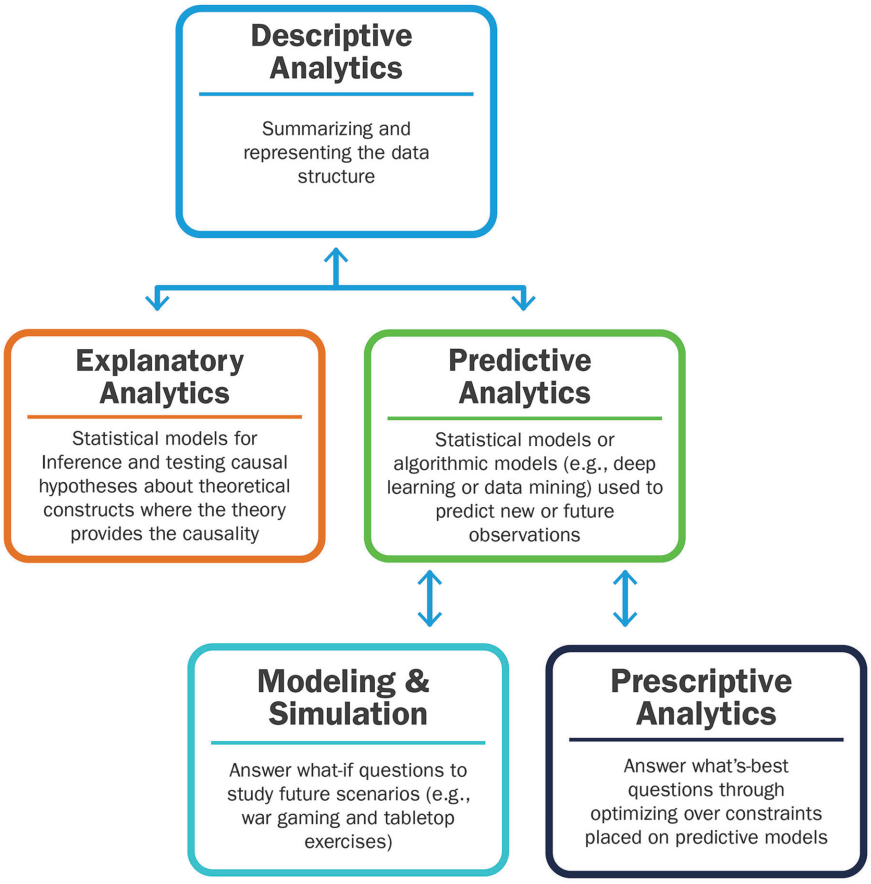

Figure 3. Five types of data analytics and their relationships across the types of analyses.

societal problems, we have developed a data science framework to guide the process (see Figure 5). Our data science framework provides a comprehensive, rigorous, and disciplined foundation for doing data science utilizing a scaffold for transparent communication and ethical considerations at every step of the data science process [KSSK20]. Learning at each step informs other prior and subsequent steps in the process. Here we briefly describe the data science framework and then follow that with a more detailed discussion of data science ethics.

1.2.1. The data science framework. As depicted in Figure 5, the data science framework starts with the research question, or problem identification, and continues through the project life cycle: data discovery (inventory, screening, and acquisition); data ingestion and governance; data wrangling (data profiling, data preparation and linkage, and data exploration); fitness-for-use assessment; statistical modeling and analyses; communication and dissemination of results; and ethics reviews throughout the entire process.

There are three main components of the data science framework: communications on the left, analytics in the middle, and data science ethics on the right. All three are essential to achieving data acumen and are intertwined. The analytics component emphasizes working in collaboration with the consumers to identify the problem(s) and develop a shared understanding of the context. Defining clear, focused unbiased research questions is the first step for the research to progress. This is achieved through conversations with stakeholders, experts, and a review of academic and gray literature. The same is true for implementing methods and developing findings and results.

Throughout the process, communication across the team (e.g., subject matter experts, data scientists, and consumers) and external dissemination provides opportunities for feedback and vetting about the analytical steps. In parallel, discussions about the ethical dimensions of the research are ongoing underlying the thinking at each stage of the data science process and through the ongoing communications and dissemination.

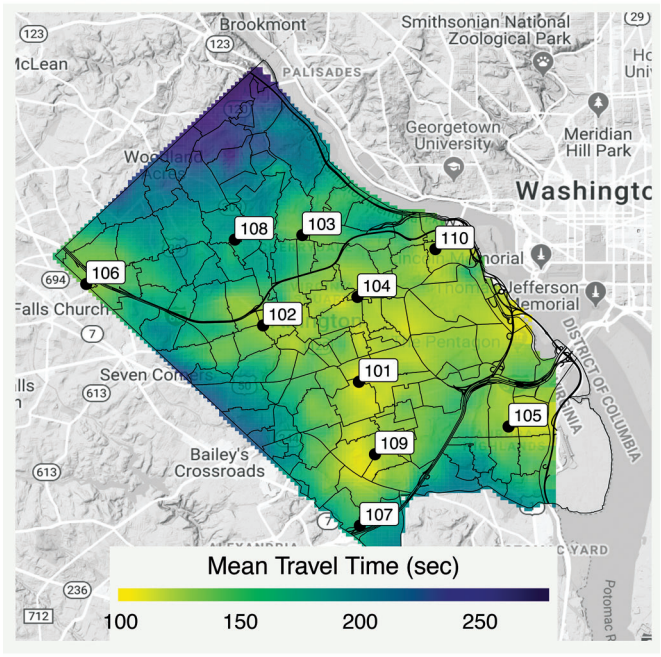

Figure 4. Travel time. The fitted mean surface for travel times of first fire engines to structure fires in Arlington. The national policy recommendation for travel time is 240 seconds. Reprinted with permission from [AGK+21].

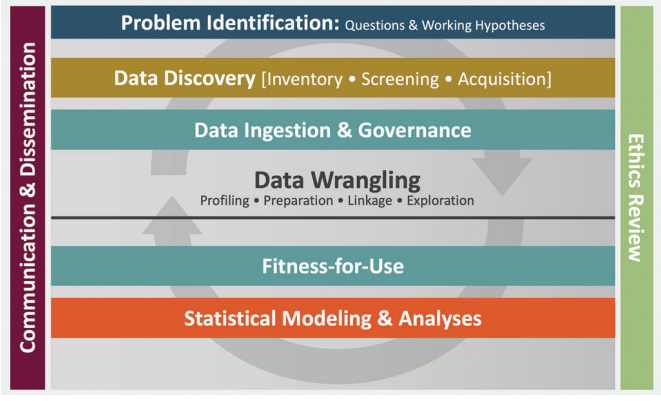

Figure 5. The data science framework provides a repeatable and iterative process for addressing a data science issue, from identifying the research question through to the statistical modeling and analysis. Communication and addressing the ethical dimensions is critical at each step of the life cycle. Adapted from [KSSK20].

1.2.2. The criticality of data science ethics. The problems that are addressed by data science are often some of the most fundamental and critically important, affecting broad swaths of our society, especially vulnerable populations. To address these problems, data science teams are assembled that bring together researchers and stakeholders across many areas of expertise, each with their own set of research integrity norms and data acumen. Consequently, ethics must be woven into every aspect of doing data science to ensure effective, equitable, and efficient solutions. The data science framework reinforces this as data science ethics touches every component and step in the data science project life cycle.

Data acumen requires an understanding of data science ethics and its application to real problems. This is critically important if we want the public to trust the results and to convey research and findings in clear and transparent ways. It is not simply about using a particular statistical or AI method but rather understanding where it fits. There are a variety of challenges to address, for example:

- Are data being used to manipulate people against their best interest?
- Are there implicit biases in the research question, the data chosen, methods used, and the analysis and findings?

Data science ethics often focus on examples of bias in algorithms that might have been avoided had there been dialogue throughout the research. For example, in 2016, Amazon algorithmically determined which neighborhoods would get Prime Same Day delivery using customer data. They did not consider race, yet the data mapped almost perfectly by race. Areas designated for same day delivery were primarily white neighborhoods and those not designated were primarily minority neighborhoods [IS16].

In another example, hospitals and insurance companies used algorithms to assign risk scores to their patients. The scores were derived from bills and insurance payouts that track illnesses, hospitalizations, and other variables. Independent health researchers found that black patients are assigned low risk scores, even when they have poor health. By tweaking the algorithm to predict the number of chronic illnesses that a patient will likely experience each year, rather than the cost of treating those illnesses, the researchers were able to eliminate most of the disparities [Pri19].

These examples demonstrate the need for active conversations about ethics and vetting results throughout the entire process. These mistakes might have been avoided if there had been earlier conversations within the team and with experts. Changes in ethical guidelines are occurring as data science becomes more pervasive in our lives, requiring creativity and data acumen by all involved in the research.

1.2.3. Evolving ethical guidelines to support the data revolution. Ethical principles continue to evolve as research is changing to not only include research directly conducted on "human subjects" to observing subjects through the massive repurposing of existing data without consent or awareness by those providing the data [Sal19]. These principles are often established in response to ethical failures in research. For example, the Nuremberg Code was created in 1947 following the notorious World War II experiments.

This written document established ten ethical principles for protecting human subjects. Principles 1, 2, and 10 are presented below to demonstrate this clarity

- Voluntary consent is essential.
- The results of any experiment must be for the greater good of society
- ...
- The scientist in charge must be prepared to terminate the experiment when injury, disability, or death is likely to occur.

These clearly written and thought-out principles lay the foundation for the Belmont principles, although it is unclear why it took 30 years.

As a result of medical studies on vulnerable populations, the Belmont Commission formed in 1979. They issued three principles for the conduct of ethical research. These principles are

- Respect for persons—(1) treating people as autonomous and honoring their wishes, and (2) protecting people with diminished autonomy, e.g., prisoners.
- Beneficence—understanding the risks and benefits of the study and weighing the balance between (1) do no harm and (2) maximize possible benefits and minimize possible harms.
- Justice—ensuring that the risks and benefits of research are distributed fairly. There are many ways to define "fair"—equal shares, individual need, individual effort, societal contribution, or merit. The last four are subjective criteria.

The principles were codified into US federal law and are now referred to as the Common Rule that governs all federally funded research. The Belmont principles provide the foundation for the Institutional Review Board's guidelines and focus on "research involving human subjects."

In 2012, the Department of Homeland Security created the Menlo Commission to bring the Belmont principles to the Information Technology (IT) world. Their rationale for doing this was to acknowledge the scale of data available and the speed and interconnectedness of IT systems and data via networks. They were also concerned about mitigating actual harm because the data are decentralized, widely distributed, and are opaque in that data users are not privy to the inner workings of applications, devices, and networks.

The Menlo Report expanded the Belmont principles in two important ways. They added a fourth principle—Respect for Law and Public Interest—that extends the principle of beneficence to include all relevant stakeholders and they expanded the focus to include "research with human-harming potential," not just a focus on "research on human subjects." This change focuses on society in a digital age where the use of technologies and repurposing of data can expose people to risks [Sal19]. Although foundational to data science ethics, oddly this report has gained little traction. This is surprising given its applicability to data science methods and integration of many types of data.

Today, federally-funded research is officially guided by the Belmont principles. However, the Menlo Report provides broader criteria for the conduct of ethical data science. Including the Menlo criteria in the project's ethical review allows researchers to make decisions and communicate their decision process for cases where official rules do not yet exist [Sal19]. The Belmont and Menlo principles together provide the foundation for ethical practice of data science research. There are other laws, definitions, and best practices that must also be followed such as IRB processes [KKO+17, KSS16].

2. Data Acumen in Practice

This section provides an end-to-end example of posing and answering the data acumen questions as the data science process unfolds.

What is the *problem*? What are the *questions*? This example is a community-based research partnership with our research division and the Roanoke Valley-Alleghany Regional Commission (RVARC) in Virginia. RVARC is one of 21 Planning District Commissions in Virginia composed of five counties and three cities in western Virginia, much of it rural. Despite being located near many universities, few graduates remain in the region. The Commission would like to build a strong manufacturing industry in the area and therefore wants to attract experienced workers from other areas and retain graduating students and current workers. We worked with the Commission to identify attributes that could attract workers and then created social, economic, health, and well-being indicators at the subcity and subcounty (neighborhood) levels to define potential regional attractiveness attributes. Discussion about implicit biases in the indicators ensued.

The following research questions were formed by the team through discussions with RVARC and assembling and disseminating early exploratory socioeconomic descriptive analyses about the region:

- What are the factors that make the Roanoke Valley-Alleghany Regional Commission attractive to workers, singles and families, and industry?
- What challenges does the region face?

What *kinds of decisions* need to be made? The Commission was interested in improving their understanding of the strengths (attractiveness) and challenges (unattractiveness) of the region to inform and understand problems and issues. They planned to use this to develop a strategic plan. This plan was to include indicators that could be

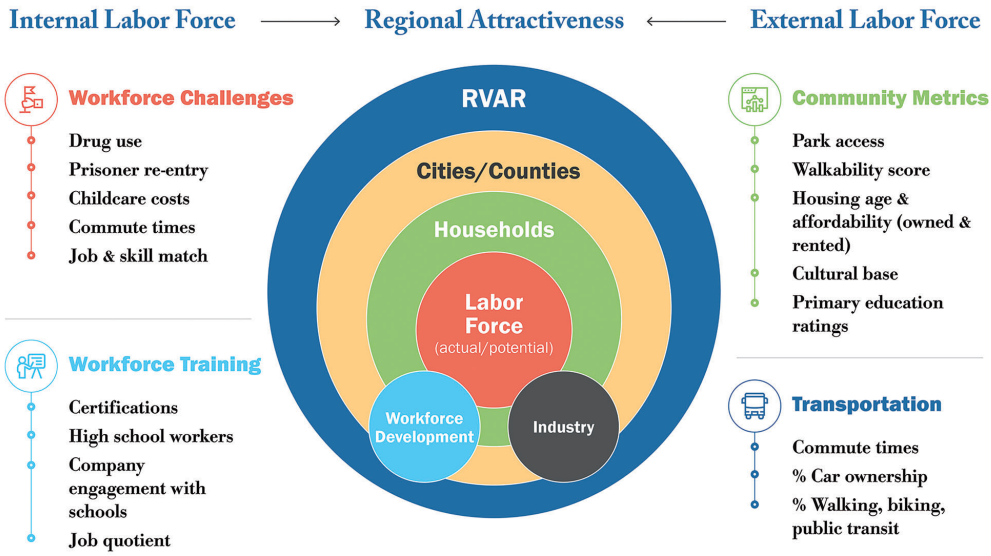

Figure 6. The data map identifies the data sources to characterize attractiveness and the units of analysis, i.e., labor force, households, cities and counties, and the region. The data map sets the stage for data discovery.

monitored to measure yearly progress and anticipate productive changes to the plan.

What *data are available*? Why should the *data be trusted*? Working with the Commission, factors were identified they deemed important to make an area attractive or not and the challenges the region faces. Through this process, the team created a data map to identify the kind of data we would ideally like to find to characterize attractiveness (see Figure 6). The map was developed based on local knowledge of the region and review of related literature.

The data map includes four categories—workforce challenges, workforce training, community metrics, and transportation. For example, to support a growing workforce requires availability of affordable childcare, reasonable commute times, and programs that support drug rehabilitation and prisoner re-entry into the workforce and community. An attractive region has a strong educational and training infrastructure. These include support to encourage high school completion, ability to participate in additional training and certifications, availability of community college, and on-the-job training, as well as company engagement with schools at all levels. Community metrics include factors that make a place desirable to live, including access to parks, affordable and up-to-date housing, cultural events, and good schools. In a largely rural area, transportation options and alternatives are needed. The Commission was interested in learning about attractiveness at the neighborhood level in the region. They proposed using the voting district level of geography to approximate neighborhoods. Finally, the data map clearly highlighted the different units of analysis across the kinds of data needed to support this problem. This integration required statistical and mathematical methods to build relevant indicators.

The identification of these attractiveness factors set the stage for data discovery—seeking out all possible sources of data to use in the analysis through a structured data discovery process [KSSK20]. We discovered, screened, and acquired a mix of data sources:

- Designed data—American Community Survey;
- Administrative data—local property records;
- Opportunity data—place-based data to identify childcare, education and training locations, and other services, voting district shapefiles); and
- Procedural data—regional strategic plans and annual reports.

At each stage, we discussed whether our research questions, data map, data discovery, and choice of data sources had implicit biases or might produce biased results. These are not easy questions to answer but important to discuss to keep ethical thinking integrated into the data science process.

What *types of analyses* are needed to inform the decision? To support the development of the data map and subsequent data discovery, we characterized the region across a collection of socioeconomic variables using descriptive analyses. For example, geospatial descriptive analysis was used to map the location of childcare facilities, public transportation, drug rehabilitation, and workforce development training facilities to identify access (or lack of access) to services.

To support the research questions and commission's strategic planning, we developed attractiveness indicators. These needed to combine relevant data represented in the data map. The analyses would also need to account for the varying units of analyses and to provide indicators at the subcounty and subcity levels to compare the attractiveness across the 48 voting districts to approximate neighborhoods.

Modeling and simulation was used to develop synthetic data at the level of housing units across the region. This was done using the American Community Survey (ACS) and local property records to impute microdata for each housing unit controlling for ACS age, income, race, and poverty. We used a Bayesian model to capture the multivariate distributions of these variables. Synthetic data over multiple sets of imputations were used to create summaries and estimate margins of error for the variables of interest over the 48 voting districts.

Descriptive analyses were applied to examine the individual synthetic variables by voting districts. Composite indicators were created by statistically combining and then ranking the aggregate synthetic data across each voting district for transportation, housing, and community variables and for two groups—singles and families. These indicators were then displayed geographically to visually present the attractiveness indexes across the region (see Figure 7). On the left-hand map, the urban areas are more attractive to singles (the lighter colors), and on the right-hand map, the more rural areas are more attractive to families.

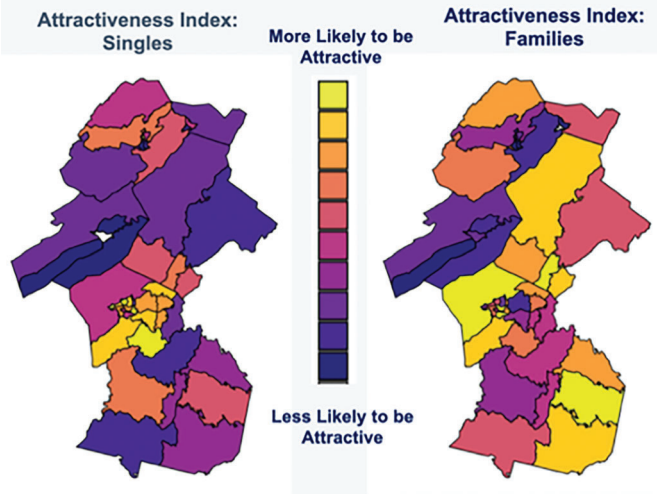

Figure 7. Attractiveness composite indicators are presented for singles and families using synthetic data created from American Community Survey and local property data. The indicator has three categories of data (1) Transportation (commute time, vehicle ownership, and access to public transportation), (2) Housing (size, type, and age of home); and (3) Community (renter or owner, singles or families, diversity (based on race), percent employed or in school).

What are the *ethical considerations*? At each stage of the research, we addressed whether our research questions, variable and data sources, methods, and findings had implicit biases. We had conversations with the Commission members, county and city planners, those providing services, and discussions with people in different parts of the region. There was sensitivity around issues related to drug rehabilitation and prisoner re-entry into the community and workforce. After discussion, the Commission felt that these issues were too important to ignore and that policies are needed to support these populations, with the motivation that they can be part of the growing workforce to make the region attractive. The Commission is now using this information in their planning. They are identifying transportation alternatives to transport workers to training and to jobs, providing incentives for more childcare, and planning new amenities to attract and retain workers in the region. This collaborative process increased the data acumen of all involved.

3. Conclusions

Data science is increasingly becoming a foundational part of our economy and lives. This requires that everyone obtain a level of data acumen that allows them to make good judgements about the use of data to solve problems. Understanding the three components of the data acumen (data, kinds of decisions, and types of analyses) along with the data science process will equip everyone involved in data science problem solving to take advantage of the data revolution and build capacity to embrace data-driven decision-making.

ACKNOWLEDGMENTS. We would like to thank the following for their contributions to the Roanoke Valley-Alleghany Regional Commission (RVARC) case study: Wayne G. Strickland, Executive Director, RVARC, Matt Miller, Director of Information Services, RVARC, Joshua Goldstein, Assistant Research Professor, Social and Decision Analytics Division, Biocomplexity Institute, University of Virginia, and Davon Woodard, PhD Candidate in Planning, Governance and Globalization at Virginia Tech.

References

[AGK+21] Madison Arnsbarger, Joshua Goldstein, Claire Kelling, Gizem Korkmaz, and Sallie Keller, *Modeling response time to structure fires*, Amer. Statist. **75** (2021), no. 1, 92–100, DOI 10.1080/00031305.2019.1695664. MR4203485

[Dat95] Data.com, `https://www.dictionary.com/browse/datum`, 1995. [Online; accessed 06-March-2021].

[Gar19] Alan M. Garber, *Data science: What the educated citizen needs to know*, Harvard Data Science Review **1** (2019), no. 1.
[GH18] Yawen Guan and Murali Haran, *A computationally efficient projection-based approach for spatial generalized linear mixed models*, J. Comput. Graph. Statist. **27** (2018), no. 4, 701–714, DOI 10.1080/10618600.2018.1425625. MR3890863
[IS16] David Ingold and Spencer Soper, *Amazon doesn't consider the race of its customers. Should it?*, Bloomberg (2016).
[KKO+17] Sallie Keller, Gizem Korkmaz, Mark Orr, Aaron Schroeder, and Stephanie Shipp, *The evolution of data quality: Understanding the transdisciplinary origins of data quality concepts and approaches*, Annual Review of Statistics and Its Application **4** (2017), 85–108.
[KLS17] Sallie Keller, Vicki Lancaster, and Stephanie Shipp, *Building capacity for data-driven governance: Creating a new foundation for democracy*, Statistics and Public Policy **4** (2017), no. 1, 1–11.
[KSK+18] Sallie Keller, Stephanie Shipp, Gizem Korkmaz, Emily Molfino, Joshua Goldstein, Vicki Lancaster, Bianica Pires, David Higdon, Daniel Chen, and Aaron Schroeder, *Harnessing the power of data to support community-based research*, Wiley Interdiscip. Rev. Comput. Stat. **10** (2018), no. 3, e1426, 10, DOI 10.1002/wics.1426. MR3799917
[KSS16] Sallie Ann Keller, Stephanie Shipp, and Aaron Schroeder, *Does big data change the privacy landscape? A review of the issues*, Annual Review of Statistics and Its Application **3** (2016), 161–180.
[KSSK20] Sallie Ann Keller, Stephanie S. Shipp, Aaron D. Schroeder, and Gizem Korkmaz, *Doing data science: A framework and case study*, Harvard Data Science Review **2** (2020), no. 1.
[LBAM20] Katerina Lepenioti, Alexandros Bousdekis, Dimitris Apostolou, and Gregoris Mentzas, *Prescriptive analytics: Literature review and research challenges*, International Journal of Information Management **50** (2020), 57–70.
[oNI19] Director of National Intelligence, *National intelligence strategy of the United States* (2019).
[oS17] National Academies of Sciences, *Strengthening data science methods for Department of Defense personnel and readiness missions*, National Academies Press, 2017.
[oS18] National Academies of Sciences, *Data science for undergraduates: Opportunities and options*, National Academies Press (US), 2018.
[PR11] Jean-François Plante and Nancy Reid, *Statistics in the news*, Amer. Statist. **65** (2011), no. 2, 80–88, DOI 10.1198/tast.2011.11018. MR2829050
[Pri19] Michael Price, *Hospital risk scores prioritize white patients*, Science (2019).
[Ros16] Robert Rose, *Defining analytics: A conceptual framework*, ORMS Today **43** (2016), no. 3.
[Sal19] Matthew J. Salganik, *Bit by bit: Social research in the digital age*, Princeton University Press, 2019.
[Sch03] Richard L. Scheaffer, *Statistics and quantitative literacy*, Quantitative literacy: Why numeracy matters for schools and colleges, The National Council on Education and the Disciplines (2003), 145–152.
[Ste99] Lynn Arthur Steen, *Numeracy: The new literacy for a data-drenched society*, Educational Leadership **57** (1999), 8–13.
[Win19] Jeannette M. Wing, *The data life cycle*, Harvard Data Science Review **1** (2019), no. 1.


Sallie Ann Keller


Stephanie Shipp

Credits

Opening image is courtesy of malerapaso via Getty.
Figures 1–3, 6, and 7 are courtesy of Sallie Ann Keller and Stephanie Shipp.
Figure 4 is reprinted by permission of the American Statistical Association.
Figure 5 is adapted from [KSSK20]. Licensed under CC-BY 4.0.
Photo of Sallie Ann Keller is courtesy of Johnny Shryock Photography.
Photo of Stephanie Shipp is courtesy of Jack Looney Photography.