Harvard Data Science Review • Issue 2.1, Winter 2020

Doing Data Science: A Framework and Case Study

Sallie Ann Keller[1] Stephanie S. Shipp[1] Aaron D. Schroeder[1] Gizem Korkmaz[1]

[1]Social and Decision Analytics Division, Biocomplexity Institute and Initiative, University of Virginia, Charlottesville, Virginia, United States of America

Published on: Feb 21, 2020

DOI: https://doi.org/10.1162/99608f92.2d83f7f5

ABSTRACT

Today's data revolution is not just about big data, it is about data of all sizes and types. While the issues of volume and velocity presented by the ingestion of massive amounts of data remain prevalent, it is the rapidly developing challenges being presented by the third *v*, variety, that necessitates more attention. The need for a comprehensive approach to discover, access, repurpose, and statistically integrate *all* the varieties of data is what has led us to the development of a *data science framework* that forms our foundation of *doing data science*. Unique features in this framework include problem identification, data discovery, data governance and ingestion, and ethics. A case study is used to illustrate the framework in action. We close with a discussion of the important role for data acumen.

Keywords: data science framework, data discovery, ethics, data acumen, workforce

Media Summary

In the words of Thomas Jefferson, 'knowledge is power,' an adage that data scientists understand too well given that data science is quickly becoming the new currency of sound decision making and policy development. But not all data are created equal.

Today's data revolution is not just about big data, but the emergence of all sizes and types of data. Advances in information technology, computation, and statistics now make it possible to access, integrate, and analyze massive amounts of data over time and space. Further, massive repurposing (using data for purposes other than those for which it was gathered) is becoming an increasingly common practice and concern. These data are often incomplete, flawed, challenging to access, and nonrepresentative.

That predicament is driving a significant need for a data-literate population to move from simple data analytics to actually 'doing data science.' To bring this to bear, researchers from the University of Virginia's (UVA) Biocomplexity Institute and Initiative have developed a research model and data science framework to help mature data science. The data science framework and associated research processes are fundamentally tied to practical problem solving, highlight data discovery as an essential but often overlooked step in most data science frameworks, and, incorporate ethical considerations as a critical feature to the research. Finally, as data are becoming the new currency across our economy, the UVA research team emphasizes the obligation of data scientists to enlighten decision makers on data acumen (literacy). The need to help consumers of research to understand the data and the role it plays in problem solving and policy development is important, as is building a data-savvy workforce interested in public good applications, such as the Data Science for the Public Good young scholars program led by the Biocomplexity Institute team.

Today's data revolution is more about *how* we are 'doing data science' than just 'big data analytics,' a buzzword with little value to policymakers or communities trying to solve complex social issues. With a proper research model and framework, it is possible to bring the *all* data revolution to *all* organizations, from local, state, and federal governments to industry and nonprofit organizations, expanding its reach, application, understanding, and impact.

---

1. Introduction

Data science is the quintessential translational research field that starts at the point of translation—the real problem to be solved. It involves many stakeholders and fields of practice and lends itself to team science. Data science has evolved into a powerful transdisciplinary endeavor. This article shares our development of a framework to build an understanding of what it means to *just do data science*.

We have learned how to do data science in a rather unique research environment within the University of Virginia's Biocomplexity Institute, one that is an intentional collection of statisticians and social and behavioral scientists with a common interest in channeling data science to improve the impact of decision making for the public good. Our data science approach to research is based on addressing real, very applied public policy problems. It is a research model that starts with translation by working directly with the communities or stakeholders and focusing on their problems. This results in a 'research pull' versus a 'research push' to lay the research foundation for data science. Research push is the traditional research paradigm. For example, research in biology and life sciences moves from basic bench science to bedside practice. For data science, it is through working several problems in multiple domains that the synergies and overarching research needs emerge, hence a research pull.

Through our execution of multiple and diverse policy-focused case studies, synergies and research needs across the problem domains have surfaced. A data science framework has emerged and is presented in the remainder of this article along with a case study to illustrate the steps. This data science framework warrants refining scientific practices around data ethics and data acumen (literacy). A short discussion of these topics concludes the article.

2. Data Science Framework

Conceptual models are being proposed for capturing the life cycle of data science, for example, Berkeley School of Information (2019) and Berman et al. (2018). A simple Google search of 'data science' brings forward pages and pages of images. These figures have overlapping features and are able to nicely summarize several components of the data science process. We find it critical to go beyond the conceptual framing and have created a framework that can be operationalized for the actual practice of data science.

Our data science framework (see Figure 1) provides a comprehensive approach to data science problem solving and forms the foundation of our research (Keller, Korkmaz, Robbins, & Shipp, 2018; Keller, Lancaster, & Shipp, 2017). The process is rigorous, flexible, and iterative in that learning at each stage informs prior and subsequent stages. There are four features of our framework that deviate from other frameworks and will be described in some detail. First, we specify the problem to be addressed and keep it ever-present in the framework, hence grounding the data science research in a problem to be solved. Second, we undertake data discovery, the search for existing data sources, as a primary activity and not an afterthought. Third, governance and data ingestion play a critical role in building trust and establishing data-sharing protocols. Fourth, we actively connect data science ethics to all components of the framework.

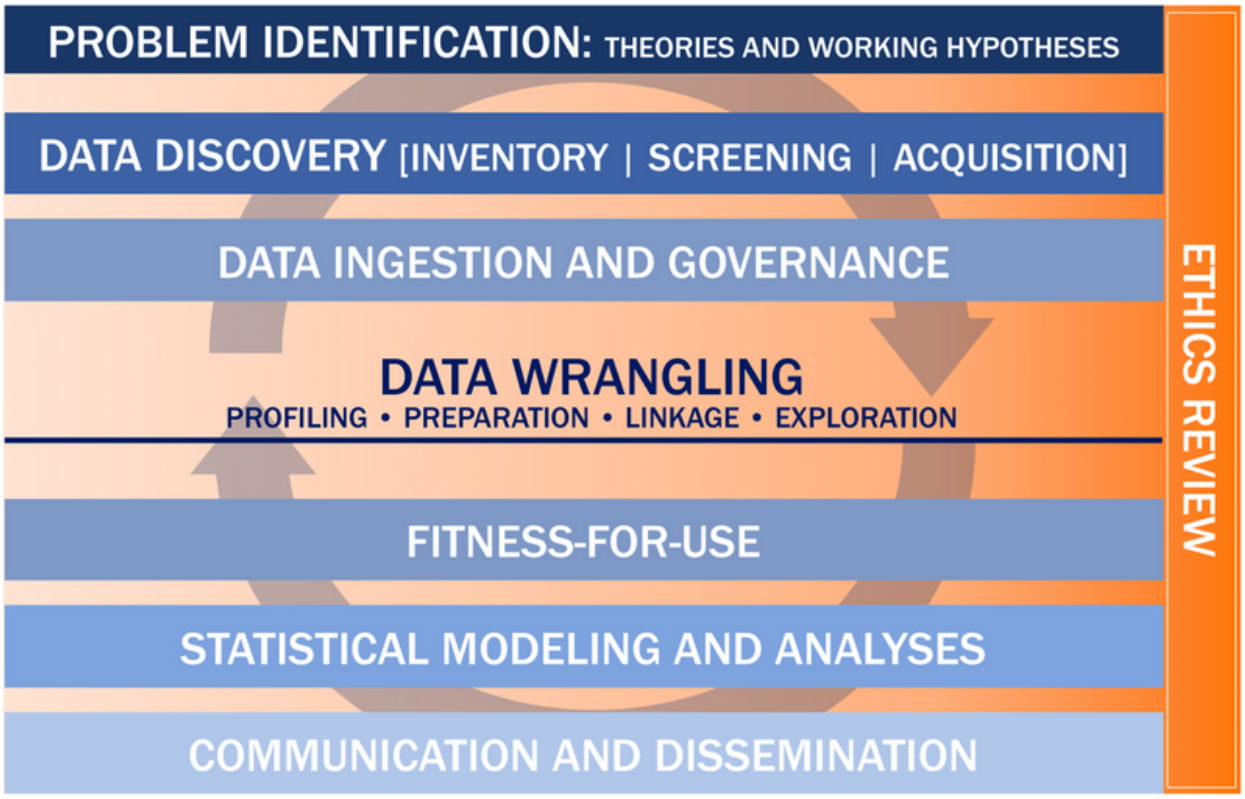

Figure 1. Data science framework. The data science framework starts with the research question, or problem identification, and continues through the following steps: *data discovery*—inventory, screening, and acquisition; *data ingestion and governance; data wrangling*—data profiling, data preparation and linkage, and data exploration; *fitness-for-use* assessment; *statistical modeling and analyses*; *communication and dissemination* of results; and *ethics review*.

In the following, we describe the components of the data science framework. Although the framework is described in a linear fashion, it is far from a linear process as represented by a circular arrow that integrates the process. We also provide a case study example for youth obesity and physical activity in Fairfax County, Virginia, that walks through the components of the framework to demonstrate how a disciplined implementation of the steps taken to do data science ensures transparency and reproducibility of the research.

2.1. Problem Identification

Data science brings together disciplines and communities to conduct transdisciplinary research that provides new insights into current and future societal challenges (Berman et al., 2018). Data becomes a common language for communication across disciplines (Keller, 2007; Keller et al., 2017). The data science process starts with the identification of the problem. Using relevant theories and framing hypotheses is achieved through traditional literature reviews, including the review of the grey literature (e.g., government, industry, and nonprofit organization reports) to find best practices. Subject matter (domain) expertise also plays a role in translating the information acquired into understanding the underlying phenomena in the data (Box, Hunter, & Hunter, 1978). Domain knowledge provides the context to define, evaluate, and interpret the findings at each stage of the research (Leonelli, 2019; Snee, DeVeaux, & Hoerl, 2014).

Domain knowledge is critical to bringing data to bear on real problems. It can take many forms, from understanding the theory, the modeling, or the underlying changes observed in data. For example, when we repurpose local administrative data for analyses, community leaders can explain underlying factors and trends in the data that may not be apparent without contextual knowledge.

> **Case Study Application—Problem Identification**
>
> The Health and Human Services (HHS) of Fairfax County, Virginia, is interested in developing capacity for data-driven approaches to gain insights on current issues, such as youth obesity, by characterizing social and economic factors at the county and subcounty level and creating statistical models to inform policy options. Fairfax County is a large county (406 square miles) with 1.1 million people across all income groups and ethnicities.
>
> The obesity rate in the United States has steadily increased since the 1970s due to growing availability of food and declining physical activity that occurs as people get older. The project aims are to identify trends and activities related to obesity across geographies of interest for local policy and program development. The HHS sponsors provided insight and context in identifying geographic regions of interest for Fairfax County decision makers. Instead of using traditional census tracts to analyze subcounty trends, they requested that the analyses be based on Fairfax County high school attendance areas and political districts. As described in the following, this led to innovations in our research through the creation of synthetic information technology to align data by these geographic dimensions.

2.2. Data Discovery (Data Inventory, Screening, and Acquisition)

Data discovery is the identification of potential data sources that could be related to the specific topic of interest. Data pipelines and associated tools typically start at the point of acquisition or ingestion of the data (Weber, 2018). A unique feature of our data science framework is to start the data pipeline with data discovery.

The goal of the data discovery process is to think broadly and imaginatively about all data, capturing the full potential variety of data (the third *v* of the data revolution) that could be useful for the problem at hand and literally assemble a list of these data sources.

An important component of doing data science is to first focus on *massive repurposing of existing data* in the conceptual development work. Data science methods provide opportunities to wrangle these data and bring them to bear on the research questions. In contrast to traditional research approaches, data science research allows researchers to explore all existing data sources before considering the design of new data collection. The advantage of this approach is that data collection can be directly targeted at current gaps in knowledge and information.

Khan, Uddin, and Gupta (2014) address the importance of variety in data science sources. Even within the same type of data, for example, administrative data, the problem (research question) drives its use and applicability of the information content to the issue being addressed. This level of variety drives what domain discoveries can be made ("Data Diversity," 2019). Borgman (2019) notes that data are human constructs. Researchers and subject matter experts decide "what are data for a given purpose, how those data are to be interpreted, and what constitutes appropriate evidence." A similar perspective is that data are "relational," and their meaning relies on their history (how the data are born and evolve), their characteristics, and the interpretation of the data when analyzed (Leonelli, 2019).

Integrating data from disparate sources involves creating methods based on statistical principles that assess the usability of the data (United Nations Economic Commission for Europe, 2014, 2015). These integrated data sources provide the opportunity to observe the social condition and to answer questions that have been challenging to solve in the past. This highlights that the usefulness and applicability of the data vary depending on its use and domain. There are barriers to using repurposed data, which are often incomplete, challenging to access, not clean, and nonrepresentative. There may also exist restrictions on data access, data linkage, and redistribution that stem from the necessity of governance across multiple agencies and organizations. Finally, repurposed data may pose methodological issues in terms of inference or creating knowledge from data, often in the form of statistical, computational, and theoretical models (Japec et al., 2015; Keller, Shipp, & Schroeder, 2016).

When confronted over and over with data discovery and repurposing tasks, it becomes imperative to understand how data are born. To do this, we have found it useful to define data in four categories, designed, administrative, opportunity, and procedural. These definitions are given in Table 1 (Keller et al., 2017, Keller et al., 2018). The expected benefits of data discovery and repurposing are the use of timely and frequently low-cost (existing) data, large samples, and geographic granularity. The outcomes are a richer source of data to support the problem solving and better inform the research plan. A caveat is the need to also weigh the costs of repurposing existing data compared to new data collection, questioning whether new experiments would provide faster results and more unbiased results than finding and repurposing data. In our experience, the

benefits of repurposing existing data sources often outweigh these costs and, more importantly, provides guidance on data gaps for cost effective development of new data collection.

Table 1. Data types.

Designed Data involve statistically designed data collections, such as surveys and experiments, and intentional data collections such as astronomical observations, remote sensing, and health registries.
Administrative Data are collected for the administration of an organization or program by entities, such as government agencies as they provide services, companies to track orders, and universities to record registered students.
Opportunity Data are derived from Internet-based information, such as websites and social media and captured through application programming interfaces (APIs) and Web scraping.
Procedural Data focus on processes and policies, such as a change in health care coverage or a data repository policy that outlines procedures and metadata required to store data.

Note. Adapted from Keller et al. (2018).

The typology of data (designed, administrative, opportunity, and procedural) provides a systematic way to think about possible data sources and a foundation for the data discovery steps. *Data inventory* is the process by which the data sources are first identified through brainstorming, searching, and snowballing processes (see Figure 2).

A short set of data inventory questions is conducted to assess the usefulness of the data sources to support the research objectives for a specific problem. The process is iterative, starting with the data inventory questions to assess whether the data source meets the basic criteria for the project with respect to the type of data, recurring nature of the data, data availability for the time period needed, geographic granularity, and unit of analysis required. If the data meet the basic criteria, then they undergo additional *screening* to document the provenance, purpose, frequency, gaps, how used in research, and other uses of the data. We employ a ‘data map’ to help drive our data discovery process (see Figure 3). Throughout the course of the project, as new ideas and data sources are discovered, they are inventoried and screened for consideration.

The *acquisition* process for existing data sources depends on the type and source of the data being accessed and includes downloading data, scraping the Web, acquiring it directly from a sponsor, or purchasing data from aggregators, or other sources. It also includes the development and initiation of data sharing agreements, as necessary.

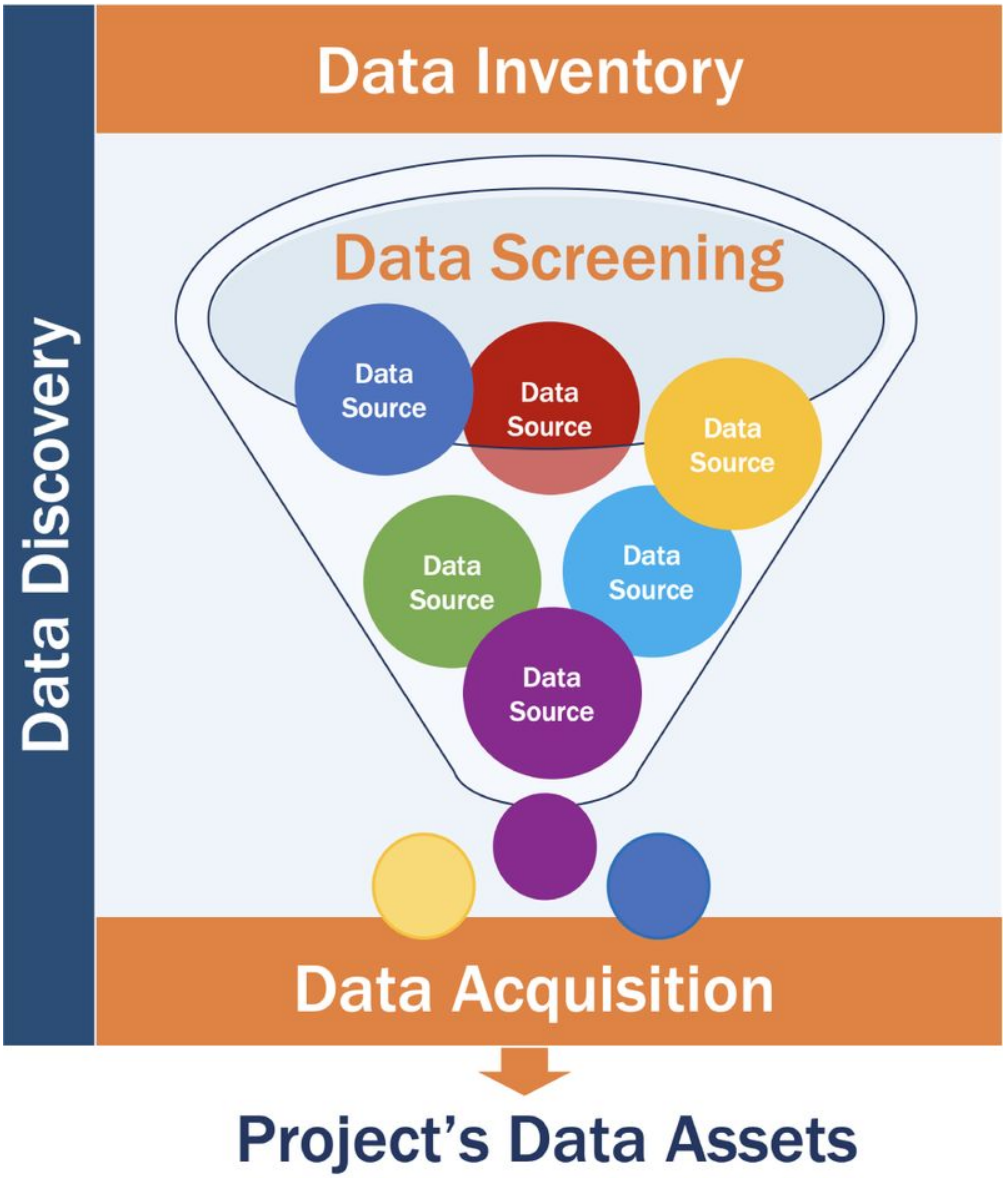

Figure 2. Data discovery filter. Data discovery is the open-ended and continuous process whereby candidate data sources are identified. Data inventory refers to the broadest, most far-reaching 'wish list' of information pertaining to the research questions. **Data screening** is an evaluative process by which eligible data sets are sifted from the larger pool of candidate data sets. ***Data acquisition is the process of acquiring the data from a sponsor, purchasing it, downloading it using an application programming interface (API), or scraping the web.***

Case Study Application—Data Discovery

The creation of a data map highlights the types of data we want to 'discover' for this project (see Figure 3). This is guided by a literature review and Fairfax County subject matter experts that are part of the

Community Learning Data Driven Discovery (CLD3) team for this project (Keller, Lancaster, & Shipp, 2017). This data map immediately captures the multiple units of analysis that will need to be integrated in the analysis. The data map helps the team identify potential implicit biases and ethical considerations.

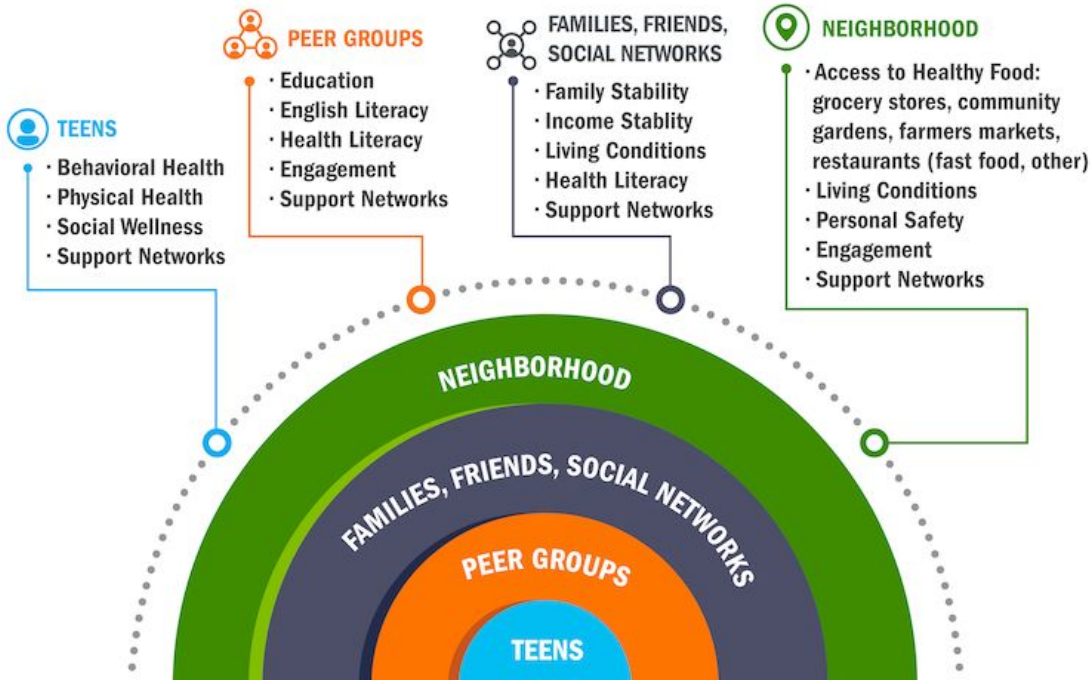

Figure 3. Data Map. The data map highlights the types of data desired for the study and is used as a guide for data discovery. The lists are social determinants and physical infrastructure that could affect teen behaviors. The map highlights the various units of analysis that will need to be captured and linked in the analyses. These include individuals, groups and networks of individuals, and geographic areas.

Data Inventory, Screening, and Acquisition.

The data map then guides our approach to identify, inventory, and screen the data. We screened each one to assess its relevance to this project, as follows.

For surveys and administrative data:

- Are the data at the county or subcounty level? (Note: This question screened out several national sources of data that are not available at the geographic granularity needed for the study.)
- What years are the data available, i.e., are they for the same years as the American Community Survey (ACS) and Fairfax Youth Survey?
- Can we acquire and use the data in the timeframe of the project, e.g., March to September?

For place-based data:

- Is an address provided?
- Can the type of establishment be identified?
- Can we acquire and use the data in the timeframe of the project?

Following the data discovery step, we identified and acquired survey, administrative, and place-based (opportunity) data to be used in this study. These are summarized in Table 2. The baseline data are the ACS, which provides demographic and economic data at the census block and census tract levels. We characterize the housing and rental stock in Fairfax County through the use of property tax assessment administrative records. Geo-coded place-based data are scraped from the Web and include location of grocery stores, convenience stores, restaurants (full-service and fast food), recreation centers, and other opportunities for physical activity. We also acquired Fairfax County Youth Survey aggregates (at the high school boundary level) and Fairfax Park Authority administrative data.

Table 2. Selected data sources.

Data Source	Geography
American Community Survey data (Census), 2012–2016*	Census tracts and block groups
American Time Use Survey (BLS), 2017	National
CDC Youth Risk Behavior Surveillance System, 2015	State
County Health Rankings, 2017	County
Fairfax County Built Environment, e.g., grocery stores, Supplemental Nutrition Assistance Program (SNAP) retailers, recreation centers, community gardens*	Address level
Fairfax County rental unit data from Dept. of Management and Budget*	Address-level rental unit data
Fairfax County real estate tax assessment data from CoreLogic	Address-level owned-home unit data
Fairfax County real estate tax assessment data from Dept. of Management and Budget*	Address-level owned-home unit data
Fairfax County Open data: Zoning, Environment, Water, Parks, Roads*	Shapefiles
Fairfax County Youth Survey, 2016; 8th, 10th, 12th graders*	High school attendance area

Fairfax County Park Authority Data, 2017*	Address level
National Center for Education Statistics, 2014-2015	High school
Virginia Department of Education, 2017	High school
* indicates data used in the final analysis	

2.3. Data Governance and Ingestion

Data governance is the establishment of and adherence to rules and procedures regarding data access, dissemination, and destruction. In our data science framework, access to and management of data sources is defined in consultation with the stakeholders and the university's institutional review board (IRB). *Data ingestion* is the process of bringing data into the data management platform(s).

Combining disparate data sources can raise issues around privacy and confidentiality, frequently from conflicting interests among researchers and sponsors working together. For clarity, *privacy* refers to the amount of personal information individuals allow others to access about themselves and *confidentiality* is the process that data producers and researchers follow to keep individuals' data private (National Research Council, 2007).

For some, it becomes intoxicating to think about the massive amounts of individual data records that can be linked and integrated, leading to ideas about following behavioral patterns of specific individuals, such as what a social worker might want to do. This has led us to a data science guideline distinguishing between ensuring confidentiality of the data for research and policy analyses versus real-time activities such as casework (Keller et al., 2016). Casework requires identification of individuals and families for the data to be useful, policy analysis does not. For casework, information systems must be set up to ensure that only social workers have access to these private data and approvals granted for access. Our focus is policy analysis.

Data governance requires tools to identify, manage, interpret, and disseminate data (Leonelli, 2019). These tools are needed to facilitate decision making about different ways to handle and value data and to articulate conflicts among the data sources, shifting research priorities to consider not only publications but also data infrastructures and curation of data. Our best practices around data governance and ingestion are included as part of the training of all research team members and also captured in formal data management plans.

Resulting modified read-write data, or code that can generate the modified data, produced from the original data sources are stored back to a secure server and only accessible via secured remote access. For projects involving protected information, unless special authorization is given, researchers do not have direct access to data files. For those projects, data access is mediated by the use of different data analysis tools hosted on our

own secure servers that connect to the data server via authenticated protocols (Keller, Shipp, & Schroeder, 2016).

Case Study Application—Data Governance and Ingestion

Selected variables from data sources in Table 2 were profiled and cleaned (indicated by the asterisks). Two unique sets of data requiring careful governance were discovered and included in the study. First is the Fairfax County Youth Survey, administered to 8th, 10th, and 12th graders every year. Access to these data requires adhering to specific governance requirements that resulted in aggregate data being provided for each school. These data include information about time spent on activities, e.g., homework, physical activity, screen time; varieties of food eaten each week; family structure and other support; and information about risky behaviors, such as use of alcohol and drugs. Second, the Fairfax County Park Authority data include usage data at their nine recreation centers, including classes taken, services used, and location of recreation center.

2.4. Data Wrangling

These next phases of executing the data science framework activities of data profiling to assess quality, preparation, linkage, and exploration can easily consume the majority of the project's time and resources and contribute to assessing the quality of the data (Dasu & Johnson, 2003). Details of data wrangling are now readily available from many authors and are not repeated here (e.g., DeVeaux, Hoerl, & Snee, 2016; Wickham, 2014; Wing, 2019). Assessing the quality and representativeness of the data is an iterative and important part of data wrangling (Keller, Shipp, & Schroeder, 2016).

2.5. Fitness-for-Use Assessment

Fitness-for-use of data was introduced in the 1990s from a management and industry perspective (Wang & Stone, 1996) and then expanded to official statistics by Brackstone (1999). Fitness-for-use starts with assessing the constraints imposed on the data by the particular statistical methods that will be used and if inferences are to be made whether or not the data are representative of the population to which the inferences extend. This assessment extends from straightforward descriptive tabulations and visualizations to complex analyses. Finally, fitness-for-use should characterize the information content in the results.

Case Study Application—Fitness-for-Use

After linking and exploring the data sources, a subset of data was selected for the fitness-for-use analyses to benchmark the data. We were unable to gain access to individual student-level data and also important health information (even in aggregate) such as body mass indices (BMI is a combination of height and weight data). An implicit bias discussion across the team ensued and given these limitations the decisions

on which data would be carried forward into the analyses were guided by a refocusing of the project to characterize the social, economic, and behavioral features of the individual high schools, their attendance areas, and county political districts. These characterizations could be used to target new programing and policy development.

2.6. Statistical Modeling and Analyses

Statistics and statistical modeling are key for drawing robust conclusions using incomplete information (Adhikari & DeNero, 2019). Statistics provide consistent and clear-cut words and definitions for describing the relationship between observations and conclusions. The appropriate statistical analysis is a function of the research question, the intended use of the data to support the research hypothesis, and the assumptions required for a particular statistical method (Leek & Peng, 2015). Ethical dimensions include ensuring accountability, transparency, and lack of algorithmic bias.

Case Study Application—Statistical Modeling and Analyses

We used the place-based data to calculate and map distances between home and locations of interest by political districts and high school attendance areas. The data include the availability of physical activity opportunities and access to healthy and unhealthy food. Figure 4 gives an example of the distances from home to locations of fast food versus farmers markets within each political district.

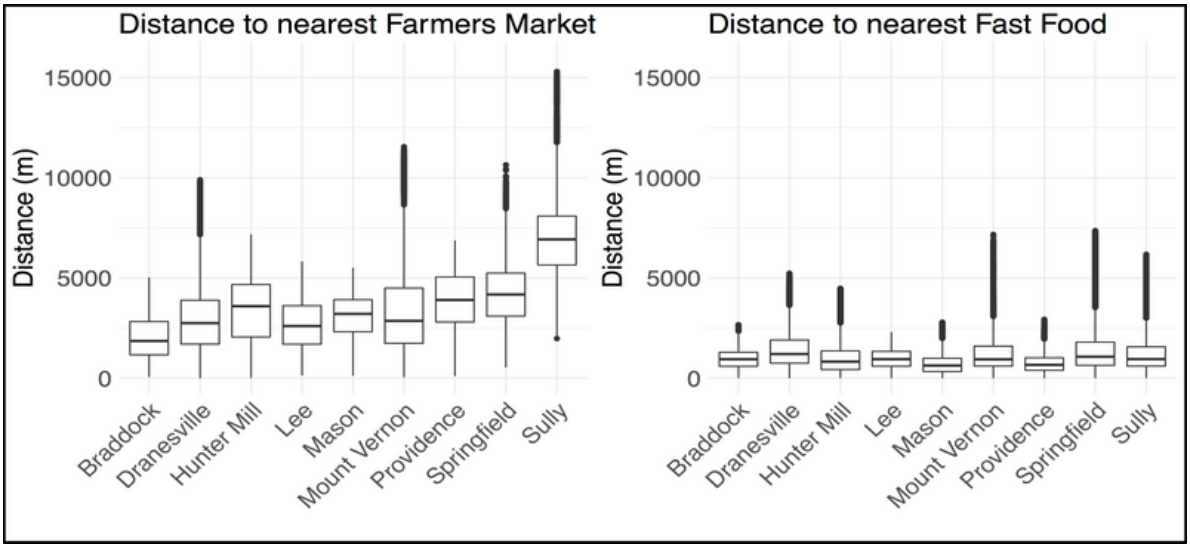

Figure 4. Exploratory analysis— direct aggregation of place-based data based on location of housing units. The box plots show the distance from each housing unit to farmers market or fast food by each of the 9 Fairfax County political districts. The take-away is that people live closer to fast food restaurants than to farmers markets.

Synthetic information methods. Unlike the place-based data, the survey data do not directly align with geographies of interest, e.g., 9 Supervisor Districts and 24 School Attendance Areas. To realign the data

and the subsequent composite indicators to the relevant geographies, we used synthetic information technology to impute social and economic characteristics and attach these to housing and rental units across the county. Multiple sets of representative synthetic information about the Fairfax population based on iterative proportional fitting were constructed allowing for estimation of margins of errors (Beckman, Baggerly, & McKay, 1996). Some of the features of the synthetic data are an exact match to the ACS marginal tabulations, while others are generated statistically using survey data collected at varying levels of aggregation. Synthetic estimates across these multiple data sources can then be used to make inferences at resolutions not available in any single data source alone.

Creation of composite indicators. Composite indicators are useful for combining data to create a proxy for a concept of interest, such as the relative economic position of vulnerable populations across the county (Atkinson, Cantillon, Marlier, & Nolan, 2002). Two composite indicators were created, the first to represent economically vulnerable populations and the second to represent schools that have a larger percent of vulnerable students (see Figure 5). We defined the indicators as follows:

- **Economic vulnerability** is the statistical combination of four factors: the percent of households with housing burden greater than 50% of household income, with no vehicle, receiving Supplemental Nutrition Assistance Program (SNAP) benefits, and in poverty.
- **High school vulnerability indicators** are developed as a statistical combination of percentage of students enrolled in Limited English Proficiency programs, receiving free and reduced meals, on Medicaid, receiving Temporary Assistance for Needy Families, and migrant or homelessness experiences.

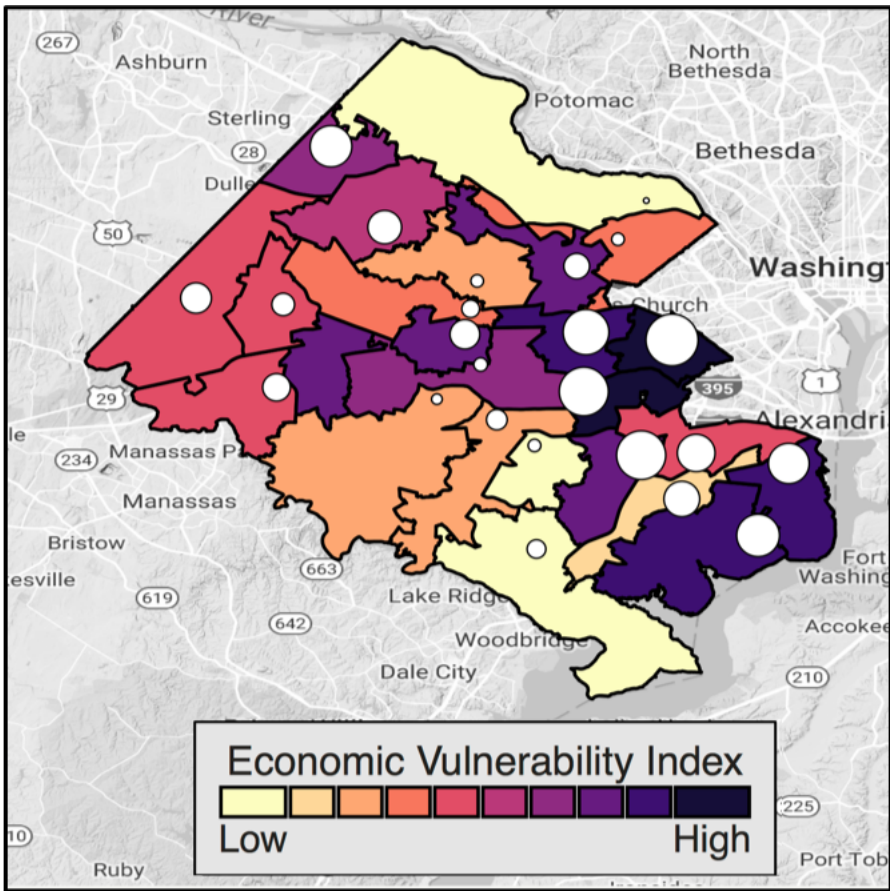

Figure 5. School and economic vulnerability indicators for Fairfax County, Virginia. Economic vulnerability indicators are mapped by the 24 high school attendance areas and by color; the darker the color, the more vulnerable is an area. The overlaid circles are high school vulnerability indicators geolocated at the high school locations. The larger the circle, the higher the vulnerability of the high school population.

Figure 6 presents correlations between factors that may affect obesity.

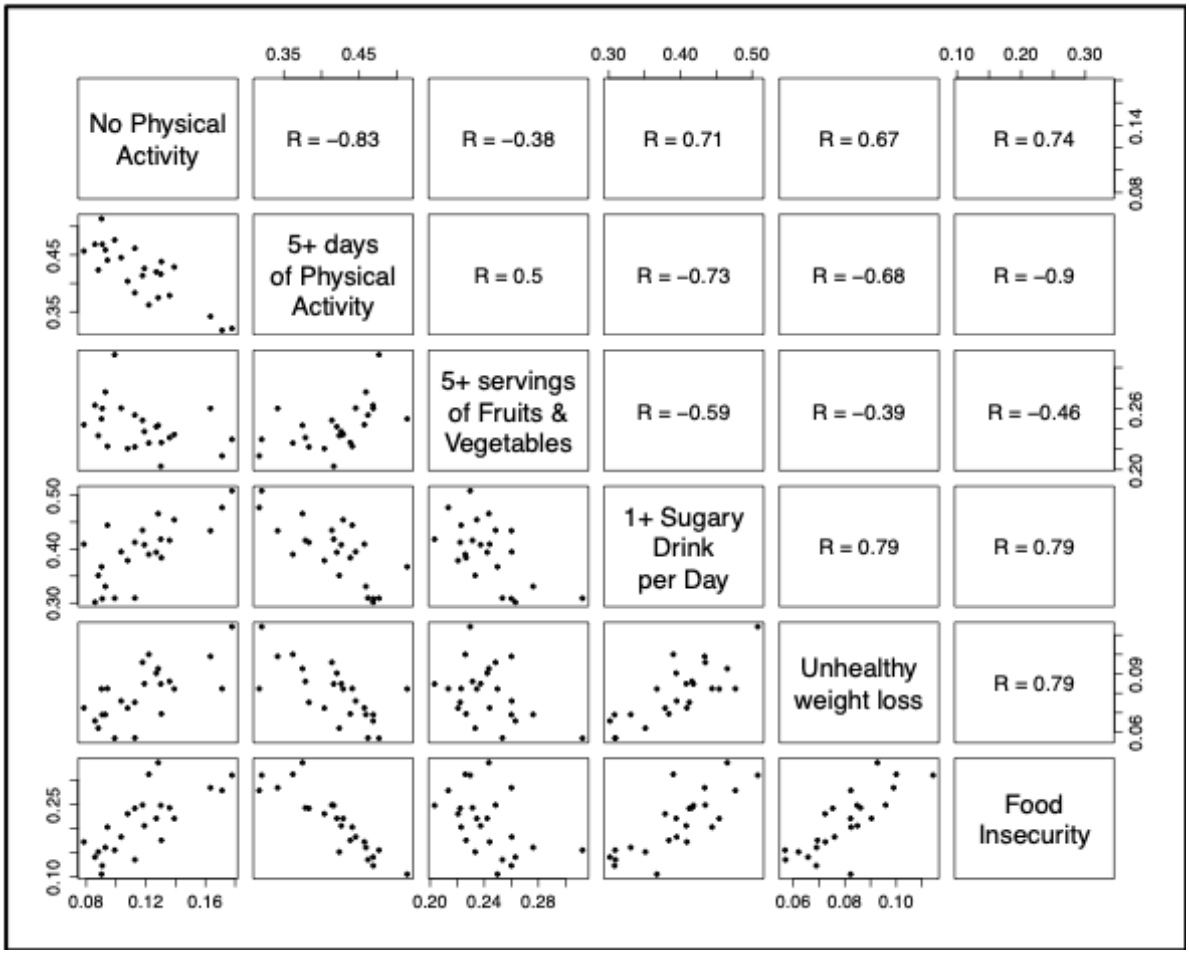

Figure 6. Correlations of factors that may affect obesity. The factors are levels of physical activity (none or 5+ days per week), food and drink consumed during past week, unhealthy weight loss, and food insecurity. As an example, the bottom left-hand corner shows a positive correlation between no physical activity and food insecurity.

The next phase of the analyses was to build statistical models that would give insights into the relationships between physical activity and healthy eating based on information from the Youth Surveys. Based on the full suite of data, several machine learning models were used.

Fitness-for-use assessment revisited.

While we were asked to examine youth obesity, we did not have access to obesity data at the subcounty level or student level. Yet, we decided to move from descriptive analysis to more complex statistical modeling to assess if existing data could still provide useful results. First, we used Random Forest, a supervised machine learning method that builds multiple decision trees and merges them together to get a more accurate and robust prediction. Our Random Forest results did not predict any reasonable or statistically significant results. Next, we used LASSO (least absolute shrinkage and selection operator), a regression analysis method that performs both variable selection and regularization (the process of adding

information) to enhance the prediction accuracy and interpretability of the statistical model it produces. However, the LASSO method consistently selected the model with zero predictors, suggesting none are useful. A partial least squares regression had the best performance when no components were used, mirroring LASSO. Instead of using the original data, partial least squares regression reduces the predictors to a smaller set of uncorrelated components and performs least squares regression on these components.

Our conclusion is that more complex statistical modeling does not provide additional information beyond the (still clearly useful) descriptive analysis. As noted below, BMI data and stakeholder input to identify the relative importance of composite indicator components are needed to extend the modeling.

2.7. Communication and Dissemination

Communication involves sharing data, well-documented code, working papers, and dissemination through conference presentations, publications, and social media. These steps are critical to ensure processes and findings are transparent, replicable, and reproducible (Berman et al., 2018). An important facet of this step is to tell the story of the analysis by conveying the context, purpose, and implications of the research and findings (Berinato, 2019; Wing, 2019). Visuals, case studies, and other supporting evidence reinforce the findings.

Communication and dissemination are also important for building and maintaining a community of practice. It can include dissemination through portals, databases, and repositories, workshops, and conferences, and the creation of new journals (e.g., *Harvard Data Science Review*). Underlying communication and dissemination is preserving the privacy and ethical dimensions of the research.

Case Study Application—Communication and Dissemination

We summarized and presented our findings at each stage of the data science lifecycle, starting with the problem asked, through data discovery, profiling, exploratory analysis, fitness-for-use, and the statistical analysis. We provided new information to county officials about potential policy options and are continuing to explore how we might obtain data-sharing agreements to obtain sensitive data, such as BMI.

The data used in this study are valuable for descriptive analyses, but the fitness-for-use assessment demonstrated the statistical models required finer level of resolution of student-level data to obtain better predictive measure, for example, body mass index (BMI) or height and weight data. The exploratory analysis described earlier provided many useful insights for Fairfax County Health and Human Services about proximity to physical activity and healthy food options for each political district and high school attendance area. We encourage Fairfax County Health and Human Services to develop new data governance policies that allow researchers to access sensitive data, while ensuring that the privacy and confidentiality of the data are maintained.

> Until we can access BMI or height and weight data, we propose to seek stakeholder input to develop composite indicators, such as the economic vulnerability indicator described in this example. These composite indicators would inform stakeholders and decision makers about where at-risk populations live, and changes over time in how those populations are faring from various perspectives such as economic self-sufficiency, health, access to healthy food, and access to opportunities for physical activity.

2.8. Ethics Review

The ethics review provides a set of guiding principles to ensure dialogue on this topic throughout the lifecycle of the project. Because data science involves interdisciplinary teams, conversations around ethics can be challenging. Each discipline has its own set of research integrity norms and practices. To harmonize across these fields, data science ethics touches every component and step in the practice of data science as shown in Figure 1. This is illustrated throughout the case study.

When acquiring and integrating data sources, ethical issues include considerations of mass surveillance, privacy, data sovereignty, and other potential consequences. Research integrity includes improving day-to-day research practices and ongoing training of all scientists to achieve "better record keeping, vetting experimental designs, techniques to reduce bias, rewards for rigorous research, and incentives for sharing data, code, and protocols—rather than narrow efforts to find and punish a few bad actors" ("Editorial: Nature Research Integrity," 2019, p. 5). Research integrity is advanced by implementing these practices into research throughout the entire research process, not just through the IRB process.

Salganik (2017) proposes a principles-based approach to ethics to include standards and norms around the uses of data, analysis, and interpretation, similar to the steps associated with implementing a data science framework. Similarly, the "Community Principles on Ethical Data Sharing," formulated at a Bloomberg conference in 2017, is based on four principles—fairness, benefit, openness, and reliability (Data for Democracy, 2018). A systematic approach to implementing these principles is ensuring scientific data are **FAIR**:

- '**Findable**' using common search tools;
- '**Accessible**' so that the data and metadata can be explored;
- '**Interoperable**' to compare, integrate, and analyze; and
- '**Reusable**' by other researchers or the public through the availability of metadata, code, and usage licenses (Stall et al., 2019).

Underlying the FAIR principles is to also give credit for curating and sharing data and to count this as important as journal publication citations (Pierce, Dev, Statham, & Bierer, 2019). The FAIR movement has taken hold in some scientific disciplines where issues surrounding confidentiality or privacy are not as prevalent. Social sciences, on the other hand, face challenges in that data access is often restricted for these

reasons. However, the aim should be to develop FAIR principles across all disciplines and adapt as necessary. This requires creating repositories, infrastructures, and tools that make the FAIR practices the norm rather than the exception at both national and international levels (Stall et al., 2019).

Building on these principles, we have developed a ***Data Science Project Ethics Checklist*** *(see the Appendix for an example)*. We find two things useful to do to instantiate ethics in every step of 'doing data science.' First, we require our researchers to take IRB) and the Responsible Conduct of Research training classes. Second, for each project, we develop a checklist to implement an ethical review at each stage of research to address the following criteria:

- Balance simplicity and sufficient criteria to ensure ethical behavior and decisions.
- Make ethical considerations and discussion of implicit biases an active and continuous part of the project at each stage of the research.
- Seek expert help when ethical questions cannot be satisfactorily answered by the research team.
- Ensure documentation, transparency, ongoing discussion, questioning, and constructive criticism throughout the project.
- Incorporate ethical guidelines from relevant professional societies (for examples, see ACM Committee on Professional Ethics. (2018), American Physical Society (2019), Committee on Professional Ethics of the American Statistical Association (2018),

Creating the checklist is the first step for researchers to agree on a set of principles and serves as a reminder to have conversations throughout the project. This helps address the challenge of working with researchers from different disciplines and allow them to approach ethics through a variety of lenses. The Data Science Ethics Checklist given in the Appendix can be adapted to specific data science projects, with a focus on social science research. Responsible data science involves using a set of guiding principles and addressing the consequences across the data lifecycle.

Case Study Application—Ethics

Aspects of the ethics review, a continuous process, have been touched on in the earlier steps of the case study, specifically, the ethics examination of the methods used, including the choice of variables, the creation of synthetic populations, and the models used. In addition, our findings were scrutinized, vetted, and refined based on internal discussions with the team, with our sponsors, Fairfax County officials, and external experts. The primary question asked throughout was whether we were introducing implicit bias into our research. We concurred that some of the findings had the potential to appear biased, such as the finding about level of physical activity by race and ethnicity.

However, in this case, these findings would be important to school officials and political representatives.

3. Data Acumen

In the process of doing data science, we have learned that many of the consumers of this research do not have sufficient data acumen and thus can be overwhelmed with how to make use of data-driven insights. It is unrealistic to think that the majority of decision makers are data scientists. Even with domain knowledge, some literacy in data science domains is useful, including the underpinnings of probability and statistics to inform decision making under uncertainty (Kleinberg, Ludwig, Mullainathan, & Obermeyer, 2015).

Data acumen, traditionally referred to as data literacy, appears to be first introduced in the 2000s as social sciences began to embrace and use publicly open data (Prado & Marzal, 2013). We define data acumen as the ability to make good judgements about the use of data to support problem solutions. It is not only the basis of statistical and quantitative analysis; it is a critical mechanism to improve society and a necessary first step to statistical understanding. The need for policy and other decision makers with data acumen is growing in parallel with the massive repurposing of all types of data sources (Bughin, Seong, Manyika, Chui, & Joshi, 2018).

We have found it useful to conceptualize data acumen across three levels or roles (Garber, 2019). The first are the data scientists, trained in statistics, computer science, quantitative social sciences, or related fields. The second are researchers trained in a specific field, such as public health or political science, who also have a range of training in data science, obtained through a master's degree, certificate programs, or hands-on programs such as the University of Virginia's Data Science for the Public Good program (UVA, 2019). This second group plays a bridging role by bringing together multidisciplinary teams. The third group are the consumers of data science applications. The first and second groups may overlap with respect to skills, expertise, and application. The third group requires a basic understanding of data science, that is, they must be data literate (Garber, 2019).

Data acumen is both a baseline and overarching concept. A data literate person should conceptually understand the basics of data science, (e.g., the data science framework described in Figure 1 is a good guide), and be able to articulate questions that require data to provide evidence:

- What is the problem?
- What are the research questions to support the problem?
- What data sources might inform the questions? Why?
- How are these data born? What are the biases and ethical considerations?
- What are the findings? Do they make sense? Do I trust them? How can I use them?

A data literate person understands the entire process, even if they do not have the skills to undertake the statistical research. Data acumen requires an understanding of how data are born, and why that matters for evaluating the quality of the data for the research question being addressed. As many types of data are discovered and repurposed to address analytical questions, this aspect of data literacy is increasingly important.

Being data literate is important to know why our intuition may not often be right (Kahneman, 2011). We believe that building data capacity and acumen of decision makers is an important facet of data science.

4. Conclusion

Without applications (problems), doing data science would not exist. Our data science framework and research processes are fundamentally tied to practical problem solving and can be used in diverse settings. We provide a case study of using local data to address questions raised by county officials. Some contrasting examples that make formal use of the data science framework are the application to industry supply chain synchronization and the application to measuring the value and impact of open source software (Keller et al., 2018; Pires et al., 2017).

We have highlighted data discovery as a critical but often overlooked step in most data science frameworks. Without data discovery, we would fall back on data sources that are convenient. Data discovery expands the power of data science by considering many new data sources, not only designed sources. We are also developing new behaviors by adopting a principles-based approach to ethical considerations as a critical underlying feature throughout the data science lifecycle. Each step of the data science framework involves documentation of decisions made, methods used, and findings, allowing opportunity for data repurposing and reuse, sharing, and reproducibility.

Our data science framework provides a rigorous and repeatable, yet flexible, foundation for doing data science. The framework can serve as a continually evolving roadmap for the field of data science as we work together to embrace the ever-changing data environment. It also highlights the need for supporting the development of data acumen among stakeholders, subject matter experts, and decision makers.

Acknowledgments

We would like to acknowledge our colleagues who contributed to the research projects described in this paper: Dr. Vicki Lancaster and Dr. Joshua Goldstein, both with the Social & Decision Analytics Division, Biocomplexity Institute & Initiative (BII), University of Virginia, Dr. Ian Crandell, Virginia Tech, and Dr. Emily Molfino, U.S. Census Bureau. We would also like to thank Dr. Cathie Woteki, Distinguished Institute Professor, Biocomplexity Institute & Initiative, University of Virginia and Professor of Food Science and Human Nutrition at Iowa State University, who provided subject matter expertise and review of the research. Our sponsors, Michelle Gregory and Sophia Dutton, Office of Strategy Management, Fairfax County Health and Human Services, supported the research and provided context for many of the findings.

Disclosure Statement

This research was partially supported by US Census Bureau under a contract with the MITRE Corporation; National Science Foundation’s National Center for Science and Engineering Statistics under a cooperative

agreement with the US Department of Agriculture, National Agriculture Statistical; U.S. Army Research Institute for Social and Behavioral Sciences; Fairfax County, Virginia.

References

ACM Committee on Professional Ethics. (2018). Association for Computing Machinery (ACM) code of ethics and professional conduct. Retrieved December 1, 2019, from https://www.acm.org/binaries/content/assets/about/acm-code-of-ethics-and-professional-conduct.pdf

Adhikari, A., & DeNero, J. (2019). The foundations of data science. Retrieved December 1, 2019, from https://www.inferentialthinking.com/chapters/intro#The-Foundations-of-Data-Science

American Physical Society. (2019). Ethics and values. Retrieved from https://www.aps.org/policy/statements/index.cfm

Atkinson, T., Cantillon, B., Marlier, E., & Nolan, B. (2002*). Social indicators: The EU and social inclusion.* Oxford, UK: Oxford University Press.

Beckman, R. J., Baggerly, K. A., & McKay, M. D. (1996). Creating synthetic baseline populations. *Transportation Research Part A: Policy and Practice*, *30*(6), 415–429. https://doi.org/10.1016/0965-8564(96)00004-3

Berinato, S. (2019). Data science and the art of persuasion: Organizations struggle to communicate the insights in all the information they've amassed. Here's why, and how to fix it. *Harvard Business Review, 97*(1). Retrieved from https://hbr.org/2019/01/data-science-and-the-art-of-persuasion

Borgman, C. L. (2019). The lives and after lives of data. *Harvard Data Science Review*, *1*(1). https://doi.org/10.1162/99608f92.9a36bdb6

Box, G. E. P., Hunter, W. G., & Hunter J. S. (1978). *Statistics for experimenters*. Hoboken, NJ: Wiley. pp.563-571

Bughin, J., Seong, J., Manyika, J., Chui, M., & Joshi, R. (2018). *Notes from the AI frontier: Modeling the impact of AI on the world economy.* Stamford, CT: McKinsey Global Institute.

Berkeley School of Information. (2019). What is data science? Retrieved December 1, 2019, from https://datascience.berkeley.edu/about/what-is-data-science/

Berman, F., Rutenbar, R., Hailpern, B., Christensen, H., Davidson, S., Estrin, D.,…Szalay, A. (2018). Realizing the potential of data science. *Communications of the ACM*, *61*(4), 67–72. https://doi.org/10.1145/3188721

Brackstone, G. (1999). Managing data quality in a statistical agency. *Survey Methodology, 25*(2), 139–150. https://repositorio.cepal.org//handle/11362/16457

Committee on Professional Ethics of the American Statistical Association. (2018). Ethical guidelines for statistical practice. Retrieved from https://www.amstat.org/asa/files/pdfs/EthicalGuidelines.pdf

Dasu, T., & Johnson, T. (2003). *Exploratory data mining and data cleaning*. Hoboken, NJ: Wiley.

Data diversity. (2019, January 11). *Nature Human Behaviour, 3*, 1–2. https://doi.org/10.1038/s41562-018-0525-y

Data for Democracy. (2018). A community-engineered ethical framework for practical application in your data work. Global Data Ethics Project. Retrieved December 1, 2019, from https://www.datafordemocracy.org/documents/GDEP-Ethics-Framework-Principles-one-sheet.pdf

De Veaux, R., Hoerl, R., & Snee, R., (2016). Big data and the missing links. *Statistical Analysis and Data Mining, 9*(6), 411–416. https://doi.org/10.1002/sam.11303

Editorial: Nature research integrity is much more than misconduct [Editorial]. (2019, June 6). *Nature* 570, 5. https://doi.org/10.1038/d41586-019-01727-0

Garber, Allan. (2019). Data science: What the educated citizen needs to know. *Harvard Data Science Review, 1*(1). https://doi.org/10.1162/99608f92.88ba42cb

Japec, L., Kreuter, F., Berg, M., Biemer, P., Decker, P., Lampe, C., . . .Usher, A. (2015). Big data in survey research: AAPOR task force report. *Public Opinion Quarterly, 79*, 839–880. https://doi.org/10.1093/poq/nfv039

Kahneman, D. (2011). *Thinking, fast and slow*. New York, NY: Farrar, Straus and Giroux.

Keller-McNulty, S. (2007). From data to policy: Scientific excellence is our future. *Journal of the American Statistical Association, 102*(478), 395–399. https://doi.org/10.1198/016214507000000275

Keller, S. A., Shipp, S., & Schroeder, A. (2016). Does big data change the privacy landscape? A review of the issues. *Annual Review of Statistics and Its Application, 3*, 161–180. https://doi.org/10.1146/annurev-statistics-041715-033453

Keller, S., Korkmaz, G., Orr, M., Schroeder, A., & Shipp, S. (2017). The evolution of data quality: Understanding the transdisciplinary origins of data quality concepts and approaches. *Annual Review of Statistics and Its Application, 4*, 85–108. https://doi.org/10.1146/annurev-statistics-060116-054114

Keller, S., Korkmaz, G., Robbins, C., Shipp, S. (2018) Opportunities to observe and measure intangible inputs to innovation: Definitions, operationalization, and examples. *Proceedings of the National Academy of Sciences*

(PNAS), *115*(50), 12638–12645. https://doi.org/10.1073/pnas.1800467115

Keller, S., Lancaster, V., & Shipp, S. (2017). Building capacity for data driven governance: Creating a new foundation for democracy. *Statistics and Public Policy*, *4*(1), 1–11. https://doi.org/10.1080/2330443X.2017.1374897

Khan, M. A. Uddin, M. F., & Gupta, N. (2014, April). Seven V's of Big Data understanding Big Data to extract value. In *Proceedings of the 2014 Zone 1 Conference of the American Society.* https://doi.org/10.1109/ASEEZone1.2014.6820689

Kleinberg, J., Ludwig, J., Mullainathan, S., & Obermeyer, Z. (2015). Prediction policy problems. *American Economic Review, 105*(5), 491–495. https://doi.org/10.1257/aer.p20151023

Leek, J. T., & Peng, R. D. (2015). What is the question? *Science*, *347*(6228), 1314–1315. https://doi.org/10.1126/science.aaa6146

Leonelli, S. (2019). Data governance is key to interpretation: Reconceptualizing data in data science. *Harvard Data Science Review, 1*(1). https://doi.org/10.1162/99608f92.17405bb6

National Research Council. (2007). *Engaging privacy and information technology in a digital age*. Washington, DC: National Academies Press.

Office for Human Research Protections. (2016). Institutional Review Board (IRB) Written Procedures: Guidance for Institutions and IRBs (draft guidance issued in August 2016). Department of Health and Human Services. Washington DC. Retrieved from: https://www.hhs.gov/ohrp/regulations-and-policy/requests-for-comments/guidance-for-institutions-and-irbs/index.html

Prado, J. C., & Marzal, M. Á. (2013). Incorporating data literacy into information literacy programs: Core competencies and contents. *Libri*, *63*(2), 123–134. https://doi.org/10.1515/libri-2013-0010

Pierce, H. H., Dev, A., Statham, E., & Bierer, B. E. (2019, June 6). Credit data generators for data reuse. *Nature*, *570*(7759), 30–32. https://doi.org/10.1038/d41586-019-01715-4

Pires, B., Goldstein, J. Higdon, D., Sabin, P., Korkmaz, G., Shipp, S., ... Reese, S. (2017). A Bayesian simulation approach for supply chain synchronization. In the *2017 Winter Simulation Conference* (pp. 1571–1582). New York, NY: IEEE. https://doi.org/10.1109/WSC.2017.8247898

Salganik, M. J. (2017). *Bit by bit: Social research in the digital age*. Princeton, NJ: Princeton University Press.

Snee, R. D., DeVeaux, R. D., & Hoerl, R. W. (2014). Follow the fundamentals. *Quality Progress*, *47*(1), 24–28. https://search-proquest-com.proxy01.its.virginia.edu/docview/1491963574?accountid=14678

Stall, S., Yarmey, L., Cutcher-Gershenfeld, J., Hanson, B., Lehnert, K., Nosek, B., ... & Wyborn, L. (2019, June 6). Make all scientific data FAIR. *Nature*, *570*(7759), 27–29. https://doi.org/10.1038/d41586-019-01720-7

United Nations Economic Commission for Europe (UNECE). (2014). A suggested framework for the quality of big data. Retrieved December 1, 2019, from https://statswiki.unece.org/download/attachments/108102944/Big%20Data%20Quality%20Framework%20-%20final-%20Jan08-2015.pdf?version=1&modificationDate=1420725063663&api=v2

United Nations Economic Commission for Europe (UNECE). (2015). Using administrative and secondary sources for official statistics: A handbook of principles and practices. Retrieved December 1, 2019, from https://unstats.un.org/unsd/EconStatKB/KnowledgebaseArticle10349.aspx

University of Virginia (UVA). (2019). Data Science for the Public Good Young Scholars Program. Retrieved December 1, 2019, from https://biocomplexity.virginia.edu/social-decision-analytics/dspg-program

Wang, R. Y., & Strong, D. M. (1996). Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems*, *12*(4), 5–33. https://doi.org/10.1080/07421222.1996.11518099

Weber, B. (2018, May 17). Data science for startups: Data pipelines (Part 3). *Towards Data Science*. Retrieved from https://towardsdatascience.com/data-science-for-startups-data-pipelines-786f6746a59a

Wickham, H. (2014). Tidy data. *Journal of Statistical Software*, *59*(10), 1–23. https://doi.org/10.18637/jss.v059.i10

Wing, J. M. (2019). The data life cycle. *Harvard Data Science Review*, *1*(1). https://doi.org/10.1162/99608f92.e26845b4

Appendix

Data Science Project Ethics Checklist

Box 1. Data Science Project Ethics Checklist
Project Initiation. *Recognize and affirm that all project plans incorporate regular checks, discussion, and documentation to ensure adherence to the ethical principles of research.*

Problem Identification. *Establish the ethical basis for undertaking the project as well as the project requirements for both the protection of research participants and the equitable allocation of all potential project benefits and risks.*

1. What are the expected benefits of the project to the 'public good,' and do they outweigh potential risks to certain populations?
2. Are there implicit assumptions and biases in the framing of the project regarding the studied communities and how will they be addressed?
3. What type of institutional review board approval process is needed? Has the team reviewed the protocol?

Data Discovery, Inventory, Screening, & Acquisition. Consider potential biases that may be introduced through the choice of datasets and variables.

4. Do the data include disproportionate coverage for different communities under study?
5. Do data have adequate geographic coverage?
6. Have checks and balances been established to identify and address implicit biases in the data?

Data Ingestion and Governance. Put in place data platforms and processes to ensure data transfer, storage, and database development adheres to data governance agreements and best practices for data quality assurance.

7. Have team members reviewed standard operating procedures (SOPs) and data management plans?
8. Do additional procedures need to be defined for this project?

Data Wrangling[1] **–** *Cleaning, transforming, linking, and exploratory analysis are critical steps in understanding data quality, how representative the data are, and potential biases in the data.*

1. What is the quality of the data?
2. How representative are the data? What populations are covered, not covered?
3. Are your assumptions correct?

Fitness-for-Use Assessment. Critically assess the overall utility of the results in achieving the predicted benefits of the study, to be transparent about potential limitations of the study, and to ensure that unintended biases haven't been introduced as a result of data choice and model refinement.

9. What are the limitations of the results? Are the results useful given the purpose of the study?

10. Do the statistical results support the potential benefits of the study previously stated?

11. Do the statistical results support the mitigation of the potential risks of the study previously stated?

12. Do any of the data require revisiting the question of potential biases being introduced through the choice of datasets and variables?

Statistical Modeling & Analysis. Establish transparency in methods, results and limitations. 13. Have project methods and outputs been made as transparent as possible? 14. Are the potential limitations of the research clearly presented? 15. Should the research be used as the basis for policy action? Have the predicted benefits and social costs to all potentially affected communities been considered?
Communication and Dissemination. Summarize ethics-related questions and actions taken, to reinforce the process of ethical consideration in continuing and future projects. Refine protocols for replication and expansion of the research findings, and information dissemination. 16. Did key ethical questions arise during the research and, if so, how were they addressed? How could they be addressed differently in future projects? 17. Are research protocols, methods and data available to other researchers? If so, in what way, and, if not, what factors are limiting the ability to do so?

Addendum

6/8/20: The authors added description and questions about Data Wrangling to the Ethics Checklist in the Appendix.

Footnotes

1. Revision, 6/8/20: Added description and questions about Data Wrangling to the Ethics Checklist. ↩