Change Through Data:

A Data Analytics Training Program for Government Employees

Frauke Kreuter[a,b,c], Rayid Ghani[d,e], and Julia Lane[f,g]

[a] Joint Program in Survey Methodology, University of Maryland

[b]School of Social Science, University of Mannheim, Germany

[c]Statistical Methods Unit, Institute of Employment Research, Germany

[e]School of Computer Science, Carnegie Mellon University

[e]Heinz College of Information Systems and Public Policy, Carnegie Mellon University

[f]Wagner Graduate School of Public Service, New York University

[g]Provostial Fellow, New York University

The work described here has been supported by the U.S. Census Bureau, the National Center for Science and Engineering Statistics, NSF SciSIP Awards 1064220 and 1262447; NSF Education and Human Resources DGE Awards 1348691, 1547507, 1348701, 1535399, 1535370, 1633603 Innovative Graduate Education Award Information Infrastructure for Society; NSF NCSES award 1423706; the Laura and John Arnold Foundation; the Overdeck Family Foundation; the Bill and Melinda Gates Foundation; Eric and Wendy Schmidt by recommendation of the Schmidt Futures program; and the Ewing Marion Kaufman and Alfred P. Sloan Foundations.

Abstract

From education to health to criminal justice, governmental decisions pertaining to regulation and policy have important effects on social and individual experience. New data science tools applied to data created by government agencies have the potential to enhance these consequential decisions. However, certain institutional barriers inhibit the realization of this potential. First, we need to provide systematic training of government employees in data analytics. Secondly, we need a careful rethinking of the rules and technical systems that protect data in order to expand access to linked individual-level data across agencies and jurisdictions, while maintaining protections for privacy. Here, we describe a program that has been run for the last three years by the University of Maryland, New York University, and the University of Chicago, with additional partners including Ohio State University, Indiana University Purdue University, Indianapolis, and the University of Missouri. The program—which trains government staff on how to work with confidential individual-level data generated through administrative processes, and extensive project-focused work--provides both online and onsite training components. Training takes place in a secure environment. The aim is to help agencies tackle important policy problems by using modern computational and data analysis methods and tools. We have found that this program accelerates the technical and analytical development of public sector employees. As such, it demonstrates the potential value of working with individual-level data across agency and jurisdictional lines. We plan to build on this initial success by creating a larger community of academic institutions, government agencies, and foundations that can work together to increase the capacity of governments to make more efficient and effective decisions.

Keywords: training programs, evidence-based policy, confidential data, administrative data research facility, government data

Introduction

Today, public policy decisions affect our everyday lives in significant ways, whether it pertains to how we gain and maintain access to health care, or how justice is administered. The massive increase in the availability of data means that it is more important than ever to apply modern data science methods to create, administer, and analyze the impact of public policy interventions in an evidence-based manner while protecting the privacy of individuals. An expanding toolbox of data analysis techniques offers ways to improve understanding of critical questions such as: 'Which individuals graduating from four year colleges are at risk of being long-term unemployed and which education and training programs improve their earnings and employment outcomes?,' 'Which ex-offenders are likely to go back to prison and can proactive outreach to connect them with health and social services reduce their risk of recidivism and improve their outcomes?,' and 'How do regulatory agencies move from reactive, complaint-based, health and safety inspections for workplaces and housing to a more proactive approach that focuses on prevention?'

Our capacity to answer such questions well matters. But in order to answer these questions better, we have to improve our capabilities. More specifically, the provision of better evidence to inform individual and policy decisions relies on analysts being able accurately to combine different existing data sources, perform the right type of analysis, and convey the results to stakeholders in a compelling fashion. However, significant institutional barriers currently prevent this potential from being fully realized, not just in the areas of education, criminal justice, and employment but across the board. Here, we focus on two key gaps.

1. *Lack of workforce capacity*. Too few government employees have the requisite skills in the use of modern computational and data analysis methods, and governments often do not

have the salary flexibility to compete with the private sector to hire data analysts (National Academy of Public Administration, 2017). While professional opportunities for statisticians and data scientists are rapidly expanding, governments struggle to build the capacity of existing staff who have valuable institutional and domain knowledge, but often do not have the necessary new skills. In the event that existing staff are reassigned, and new hires made who possess the skills that their predecessors lacked, agencies run the risk of losing important human capital and institutional knowledge. But retaining and retraining staff who lack knowledge in data analytics is expensive and time-consuming.

2. *Lack of access to confidential micro-data*. Many policy problems require the analysis of individual-level (micro) transactional or administrative data that cross agency lines; in this context, confidentiality rules often limit effective data-sharing (Reamer, Lane, Foster, & Ellwood, 2018). The time to reach agreement can take years—10 years in at least one case (Potok, 2009). In order to obtain the resources necessary to surmount the legal and technical hurdles that prevent cross-agency data collaborations, it is necessary to demonstrate the value of sharing cross-agency data. Such demonstrations are most effectively made by prototyping specific use cases. This situation leads to the current Catch-22: because we cannot demonstrate the value of new data products, agencies cannot get the significant resources necessary to make use of linked data, but lack of resources and data access agreements mean that we cannot demonstrate value.

This situation entails significant monetary and human costs. In monetary terms, the cost of collecting new data instead of using already existing data in different agencies leads to a waste of taxpayer dollars. For example, the cost for the U. S. Census Bureau to count a housing unit has increased from $16 in 1970 to $92 in 2010 and is expected to cost well over $100 by 2020 in

2020 constant dollars (Government Accountability Office, 2017).[1] Much of the data being collected already exists as administrative data within other agencies; it could be repurposed by the Census Bureau, which would allow them to focus on the collection of incremental data.

In human terms, the cost of not combining data in a timely fashion is severe. Dr. Leana Wen, Commissioner of Health, City of Baltimore, has noted that, as "part of Child Fatality Review, department heads in Baltimore City government get together once a month. We review every child death that happened in the city since the previous meeting. We ask what more we might have done to prevent that tragedy. In many cases, each of us has a file on the child or the family at least an inch thick. It's tragic to compare notes after the child has died—what more could we have done when the child was alive?" (Lane, Kendrick, & Ellwood, 2018). In addition to alleviating the risk of catastrophes, the benefits of using the most recent data science innovations include better policymaking, the more efficient use of resources by government, and better data products for businesses and households to use in their decisions and data analytics.

In this article, we describe a program that has been run for the past three years by a consortium of universities. This program enables government agencies to share confidential data, and to train their employees to tackle important policy problems using modern computational and data analysis methods and tools. Our goal in creating this program was (1) to build the technical and analytical capacity of public sector employees and (2) to demonstrate the value of working with data across jurisdictional (state and agency) lines. As an extension of this training program, we also explored (1) the establishment of fellowship programs that could provide the

[1] For the 2020 Census, the Census Bureau plans to use administrative records (data that people have already given to the federal government) to help improve its results and reduce some door-to-door visits. Of course, while only some of the costs include slow innovation in data science areas, and most involve other issues that are largely orthogonal to data science as such, e.g., inefficient implementation of data security procedures and the need for improvement of general management practices, the bureau nevertheless estimated using these data could save $900 million.

basis for continuing the work we have initiated, hardening the results into visible products, and building skills within the public sector, and (2) data infrastructure that can be used after the training programs by government agencies have concluded to develop the initial products and build new ones.

This article describes the training program and its components, our findings based on our administration of the program for the past three years, and our recommendations concerning future paths toward the creation of an infrastructure and workforce capable of tackling policy problems using modern computational and data-driven methods. The interest in the program has been overwhelming: in the last three years the program has drawn more than 450 participants from over 100 federal, state, and local government agencies, resulting in increased data analytics capacity, both in terms of human and technical resources. The program has been increasingly adopted by federal and state agencies. For example, the U.S. Department of Health and Human Services has sponsored an initiative using welfare data (tanfdata.org); the U.S. Department of Agriculture has sponsored a class to inform child nutrition policy, and the states of Illinois, Indiana, Ohio, and Missouri have worked with us to provide multiple classes on education and workforce transitions. Five more programs are planned for 2020.

Existing Approaches

The U.S. government has recognized the importance of data to run effective federal government operations and policy (Commission on Evidence-Based Policymaking, 2017; National Academies of Sciences, Engineering, and Medicine, 2017; Office of Management and Budget (OMB), 2019); the same is true at the local level (Goldsmith & Kleiman, 2017; Mays, 2018). The U.S. Congress has recently passed legislation that puts the apparatus in place to

facilitate the use of data (Hart & Shaw, 2018), and municipalities are also trying to build their own capacity for data science, with groups having been established in multiple cities. Prominent examples include the Mayor's Office of Data Analytics (MODA) and the Center for Innovation through Data Intelligence in New York City, as well as a Mayor's Office of New Urban Mechanics in both Boston, Massachusetts, and Philadelphia, Pennsylvania. Chief information officers and chief data officers in cities as large as Chicago, Illinois, and as small as Asheville, North Carolina, are taking steps to develop data science capacities to tackle a suite of operational policy problems (Pardo, 2014).

In theory, the use of data in the public sector is simple, and its value, self-evident. A government agency administers a program; such a program produces data as by-products or as the product of intentional collection; and agency staff or affiliated researchers analyze the data in order to evaluate outcomes, with a view toward future improvement. In practice, however, the use of data in this context is anything but simple. There are legal issues that must be addressed before data can be accessed and joined, since data are generated by different agencies with different missions. Legal mandates to share information are often lacking.[2] Government agencies, when trying to start a new data-driven project, have to overcome significant challenges: (1) first, they must get access to cross-agency data and link different data sources, then (2) they must get access to (in-house or external) people who can help them with the data science needed to analyze the resultant corpus of data. Below, we highlight some existing work

[2] Exceptions are OMB memorandum M-14-06 (https://www.whitehouse.gov/sites/whitehouse.gov/files/omb/memoranda/2014/m-14-06.pdf) and M-15-15 (https://www.whitehouse.gov/sites/whitehouse.gov/files/omb/memoranda/2015/m-15-15.pdf).

and persistent challenges in the areas of (1) data sharing and linkage across government agencies and (2) training programs in data science for government agencies.

Data Access Infrastructure

We have already mentioned the challenge of getting legal agreements in place for data sharing. In addition, there are technical issues that arise when linking data because the databases are often in different formats, with archaic data management systems and without common identifiers (Reamer & Lane, 2017).

On the technical side, it is now possible to make use of new infrastructures built to provide access to confidential government microdata. These infrastructures can be applied to link data across agency lines, so that agency staff can be trained to use real data to study agency problems. An overview is provided in the report of the Commission on Evidence-Based Policymaking (2017). In particular, the Center for Economic Studies, which was established by Robert McGuckin at the U.S. Census Bureau in the 1980s (McGuckin & Pascoe, 1988), has since evolved into a major source of Census Bureau and other agency administrative and survey data, with access available in 29 federal statistical research data centers at universities across the United States. Similarly, the Longitudinal Employer-Household Dynamics program, which was established in the late 1990s, has also grown into a major national program (Abowd, Haltiwanger, & Lane, 2004; Lane, Theeuwes, & Burgess 1998) that links state and federal data. The NORC/University of Chicago research data center, established in the mid-2000s, provides researchers with access to administrative data (Lane & Shipp, 2007), while Arnold Ventures (previously the Laura and John Arnold Foundation) established policy labs in several key states (Arnold Ventures, 2018). Integrated data systems have been funded by the U.S. Departments of

Education and Labor (Culhane, Fantuzzo, Hill, & Burnett, 2018). Most recently, the Census Bureau commissioned New York University to establish an Administrative Data Research Facility to inform the decision making of the Commission on Evidence-Based Policymaking.

Internal Data Science Capacity

Once data has been linked and made accessible under the right confidentiality protocols, the challenging task of finding people who understand how to make scientific use of the data remains (Barbosa, Pham, Silva, Vieira, & Freire, 2014; Castellani Ribeiro, Vo, Freire, & Silva, 2015; Catlett et al., 2014; Ferreira, Poco, Vo, Freire, & Silva, 2013). Agency employees who have questions to ask of the data often do not have the tools or skills to analyze them. Having capable, in-house data scientists who can demonstrate to their fellow civil servants the value that data has for solving practical problems may be one of the most significant steps any government can take in breaking down the barriers to value creation (Jarmin, Marco, Lane, & Foster, 2014).

In the absence of in-house resources, agencies have resorted to working with outside consultants or academic researchers to build new capacity in this domain. The development of the Census Bureau's Longitudinal Employer-Household Dynamics (LEHD) program is an example of just such an approach (Warsh, 2010). However, reliance on outsiders is not a substitute for internal capacity, particularly as data become more complex. Too often, outsiders only have limited access to data sources, and they often do not know enough about the data-generating process to make appropriate use of the data. In addition, outsiders often do not understand how the analytical results will be used, which makes it difficult properly to scope and design the analysis. Several university groups and programs, such as the Data Science for Social Good (DSSG) program at University of Chicago, have created strong collaborations with local,

state, and federal government agencies, albeit with varying levels of success (Ackerman et al., 2018). Even in the best-case scenario, when outsiders work closely with agency employees, access to and analysis of data rely largely on personal, trusted relationships, rather than a sustainable engagement between data providers and analysts. In the worst-case scenario, consultants generate reports that are not used, and recommend procedures that are rarely implemented (Goerge, 2018).

Training Government Agencies in Data Science

On the skills development side, some notable activities have addressed general workforce issues. For example, a recent National Science and Technology Council Five Year Strategic Plan called for graduate education to be designed that could conceivably provide the existing workforce with options to acquire the skills necessary for success in a broad range of careers (Holdren, Marrett, & Suresh, 2013). The authors recommend strengthening professional development and deepening employer–university engagement in upgrading the skills of the existing workforce.

In the context of data science, these recommendations could be particularly resonant, since the field is inherently applied, and of great value to employers (Davenport & Patil, 2012). Of course, the newness of data science as a field means that there is not a long history of knowledge about how to teach data science. More broadly, the task of educating people how to access and analyze data has been a perpetual challenge (Gould & Cetinkaya-Rundel, 2014), although the community is starting to develop curricula and guidance that address this task (American Statistical Association Undergraduate Guidelines Workgroup, 2014; National Academies of Sciences, Engineering, and Medicine, 2018), ranging from graduate programs at

universities such as the University of Chicago, Carnegie Mellon University, and Georgetown University, all the way down to programs for secondary students (Gould et al., 2016).

A substantial literature exists on how to design programs for new sciences. In an influential series of papers, Handelsman and others argue that new types of science need to adopt active learning techniques (Handelsman et al., 2004).[3] By this, they mean the shifting from a lecture-based format to one that is *inquiry-based and modular* and that *treats students as scientists* who not only develop hypotheses, but also design and conduct experiments or in our case analyze and interpret data, and write about their results. The approach appears to be effective: a recent meta-analysis of 225 studies of the effectiveness of 'learning by telling' vs. 'learning by doing,' albeit in the undergraduate context, suggests that 'learning by doing' increases examination performance, while 'learning by telling' increases failure rates. The positive effects are particularly pronounced for students from disadvantaged backgrounds and for women in male-dominated fields (Freeman et al., 2014).[4]

There is also much to be learned from the experience in other fields in moving from a curriculum based on providing content to one that is interdisciplinary and driven by concepts. In the biological sciences, Gutlerner and Van Vactor (2016) argue forcefully for the development of modular classes—what they call "nanocourses."[5] Rafa Iziarry, creator of a series of nanocourses in Data Science on EdX, also emphasizes the need to have "applications in the forefront rather than a theoretical focus" and to "provide learning experiences that expose students to long-term projects" when teaching Data Science (Iziarry, 2018).

---

[3] Longstanding and prominent examples of bodies of teaching materials were compiled by the Physical Sciences Study Committee and the Biological Sciences Curriculum Study.

[4] Jeff Leek, Professor of Biostatistics and Oncology at of Johns Hopkins Bloomberg School of Public Health, recently published a data science course series (Chromebook Data Science) on Leanpub with the intention to democratize data science education (https://leanpub.com/u/jtleek).

[5] See, for example, https://nanosandothercourses.hms.harvard.edu/node/8.

A substantial complementary literature exists on the value of domain-specific training institutions. One stellar example is that of the agricultural extension program. The 1862 Morrill Land-Grant College Act and the 1887 Hatch Experiment Station Act led to a "longstanding close association and integration of agricultural research with extension and higher education" (Alston & Pardey, 1996, p. 9). The resultant agricultural extension programs built an operational framework, based on Pasteur's Quadrant (Stokes, 2011). Briefly, the program entailed presenting farmers in the field with real operational problems. Correspondingly, agricultural researchers worked to collect data and develop methods to develop solutions, while extension programs created an operational bridge between the two projects. State and local agricultural societies and institutes established farmer's institutes in order to "extend new technologies and improved and best practices from progressive farmers and trained scientists" (Alston & Pardey, 1996, p. 16).

Sustained and meaningful exchanges between practitioners who know how particular problems originate in practice (such as farmers) and those charged with developing solutions, is also key in data science. Thus, one of the key lessons for any aspiring data scientist is to "discover the data generating mechanism" (Peng, 2018).

Science and Technology fellowships established by the American Association for the Advancement of Science (AAAS) have been credited with changing the political landscape in Washington (Morgan & Peha, 2003); they have been notably emulated by the Alfred P. Sloan Foundation, which has recently established similar fellowships through professional associations such as the American Statistical Association, the Association for Computing Machinery, the American Mathematical Society, the Institute for Mathematical Statistics, the Mathematics

Association of America, and the Society for Industrial and Applied Mathematics.[6] Other ways in which governments can learn how to apply new tools include the Intergovernmental Personnel Act Mobility Program (IPA), which provides for the temporary assignment of personnel or the equivalent of presidential management fellowships.

Our work described here builds on the approaches described above and creates a training program that trains existing government employees, using data derived from government agencies, and anchored around problems that they face today. We offer this description of our program in order to stimulate discussion in the community. The next section describes our program further.

Our Approach: Applied Data Analytics Training Program for Governments

Over the past few years, we have developed a training program that demonstrates how a data science program for public policy might be developed at scale to meet the quickly rising demand. Our goals in creating this program were (1) to build the technical and analytical capacity of public sector employees, as well as (2) to demonstrate the value of working with data across jurisdictional (state and agency) lines. The tangible results of the program, as initially envisioned, were to include a trained workforce, new products for agencies to address key problems, and the development of new networks. The overall aim was to provide government agencies with new resources for tackling critical problems, and to create a collaborative community of researchers and practitioners within and beyond this program.

The program we created had three primary components:

[6] https://www.amstat.org/ASA/Your-Career/ASA-Fellowships-and-Grants.aspx

1. Technical (Computational and Data) Infrastructure with access to confidential agency microdata in a secure computing environment;
2. Training Curriculum including lectures, hands-on sessions, and code notebooks using data from multiple jurisdictions;
3. Collaborative Projects based on the needs of government agencies to seed new products.

The participants used a secure cloud-based environment and were introduced to new ways of collecting data and analyzing it using new computational and data analysis methods and tools. The overarching approach included training government staff in how to keep fundamental statistical concepts like population frames, sampling, and valid inference while expanding their skills to include modern computational data analysis tools, such as machine learning. Participants were also trained in the use of new types of data, including administrative data (records generated from the administration of government programs), data captured from websites or through application programming interfaces (APIs), and data with large spatial components, network structures, or text. The program is built on a foundation of social science research principles integrated with current analytic and computer skills, while being rooted in the study of real-world social and economic problems (Foster, Ghani, Jarmin, Kreuter, & Lane, 2016). Figure 1 provides an illustration of the program.

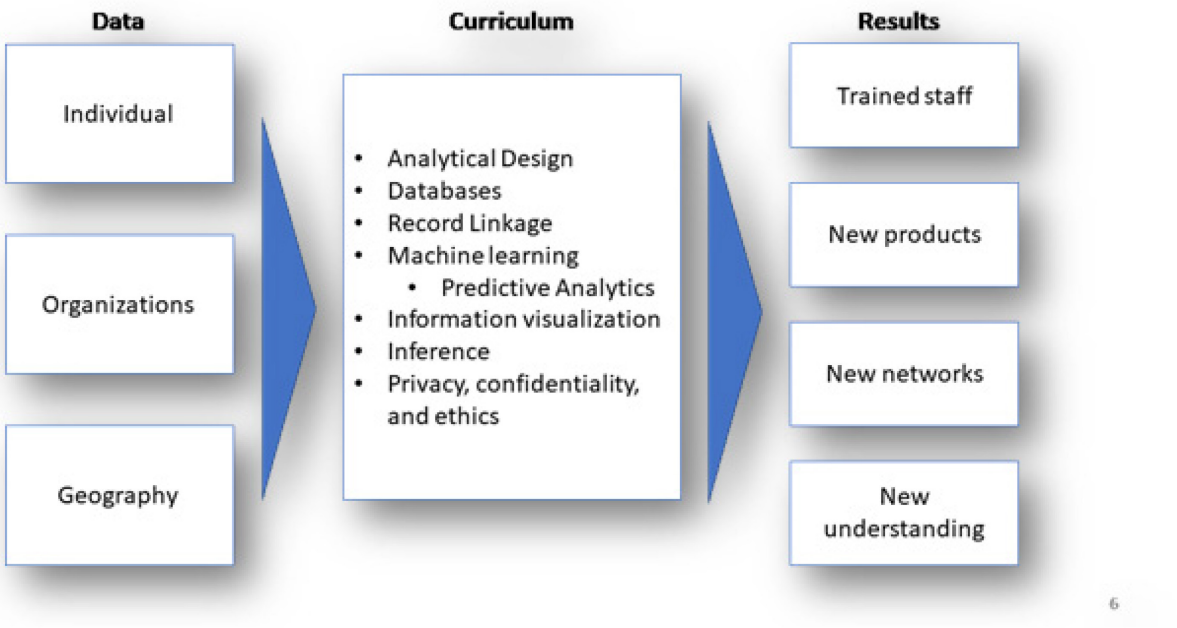


Figure 1. Schematic representation of the training programs' data input, curriculum content, and outputs.

Technical Infrastructure and Access to Confidential Cross-Jurisdictional Data

Because data access is at the core of replicable and reproducible science, we developed a secure collaborative computing environment within a state-of-the-art data facility—the Administrative Data Research Facility (ADRF),[7] which was commissioned by the Census Bureau to inform the decision making of the Commission on Evidence-Based Policymaking. We created a single access-controlled project space that all project members can access from their own computer from any network via a secure client.[8] In order to minimize unauthorized disclosure, access was limited to those with explicit permission. The environment was isolated from the internet, and analytical output was reviewed using standard statistical disclosure

[7] https://coleridgeinitiative.org/computing
[8] https://www.nomachine.com/download

limitation techniques. These shared project spaces allowed users access to shared tools and provided them with the ability to share their code, analysis output, and extracts from the data in a secure and familiar way.[9] The tools used for secure communication of project code and ideas in the ADRF included GitLab, Mattermost, and JupyterHub.

The long-term goal was to build user interfaces that presented rich context to users about the datasets as they work in secure environments, while also incentivizing users to contribute. These interfaces were designed to gather metadata that provided information about who else has used the data, for what purpose, and how others users have accessed and analyzed data in their research work (Yarkoni et al., 2019). Such sharing ideally helps to reduce search and discovery times and code development, but also allows for the replication and reuse of analysis.

Agencies were willing to share their data within the ADRF because it created a secure sandbox environment within which agency staff provided concrete evidence of the value of linking data as part of the course projects; as such, the access and use was consistent with the agency mission.

Training Curriculum to Develop New Skills

Our cross-disciplinary curriculum blends lectures, hands-on sessions, and discussions integrating computer science with statistics and social science. We broadly define data science as the product of this integration. The curriculum was designed to train participants in using modern computational and data analysis methods and tools to solve problems that are critical to their agencies in a scientifically sound manner. We covered the entire spectrum of a project lifecycle: from problem definition and scoping to data collection, linkage, processing, analysis, and

[9] See Cetinkaya-Rundel & Rundel (2018) for a summary of good computing infrastructure for education purposes.

validation as well as how to think about ethical and privacy issues that arise when tackling those issues and communicating the results and impact effectively. We introduced new types of methods from computer science and statistics in the context of their potential to transform the scientific understanding of the dynamics of human behavior. In addition to methods, we included an introduction to new types of data that are now available to solve problems in government agencies as well as a discussion of how to effectively incorporate them in analytical systems.

The core of the program was originally administered in four in-person modules (see Table 1), although the delivery structure can be and has been modified in response to agency needs. Before the class begins, participants filled out a pre-class survey about their skills (self-assessment). This survey was used to group them into teams of four or five. Each team member had a different skill set—one team member might have expertise in policy, another in statistics, another in coding, and another in data. The teams worked together on problems and projects throughout the module series.

The curriculum was delivered through lectures, and hands-on working sessions. Interactive Jupyter notebooks (Perez & Granger, 2015) facilitated the introduction of code and learning of programming skills. These interactive notebooks allowed explanations, code, and output to be all visible in the same place, and code to be executed within the notebooks themselves.

Modules. *Pre-course training.* For those participants unfamiliar with command-line–based languages and new to Python, an online inverted classroom Python and SQL bootcamp was offered prior to the course. We developed the bootcamp for the purpose of the course and made use of Binder,[10] an open-source web application for managing digital repositories. The Binder material was paired with a series of short videos provided every week to be watched at a convenient time for the participants, and a live online video chat with the instructor, to answer questions participants might have regarding the material. We also provided participants with links to a series of online resources for self-paced courses; however, our experience was that, for many employees, it was hard to free up the time without an official course structure.

Module 1. The first module of the in-person meetings covered the fundamentals of problem formulation, inference, and basic programming tools. It tied those fundamentals to a specific topic area (relevant to the agencies the participants come from or to data they otherwise are likely to interact with). The problem formulation section for the first module included such topics as: (1) understanding the science of measurement, (2) identifying research goals, and (3) identifying measurement concepts and data sources associated with the research goals. We conducted a training workshop on project scoping based on the Data Science Project Scoping Guide developed by the Center for Data Science and Public Policy at the University of Chicago.[11] The discussion of the data generation process, quality frameworks, the task of dealing with missing data, and selection issues played an important part. A first graphical inspection of the data was used to familiarize everyone with the various data sources at hand.

[10] https://mybinder.org/
[11] Data Science Project Scoping Guide: https://dsapp.uchicago.edu/home/resources/data-science-project-scoping-guide/

Table 1. Applied Data Analytics Course Modules

Module	**Corresponding Course Learning Objective**	**Corresponding Notebook and Project Work**
Foundations of Data Science	Formulating and scoping research questions and policy projects, matching research questions and data, understanding the social science of measurement, understanding quality frameworks and varying needs, introduction to the data that will be used in this class, case studies, exploring data visually, practice Python, SQL, and Github.	Notebook: -Variables Worksheet: -Project Scoping
Data Management and Curation	Introduction to APIs, building features from administrative record data, understanding data used in the class, introduction to characteristics of large databases, building datasets to be linked, fundamentals of record linkage techniques, create a data science project work flow, data hygiene: curation and documentation, practice SQL.	Notebooks: -Record Linkage -Feature Generation
Data Analysis in Public Policy	What is machine learning, examples of machine learning applications, process and methods, bias and fairness in machine learning, different text analytics paradigms, discovering topics and themes in large quantities of text data, understanding data and networks.	Notebooks: - Machine Learning - Text Analysis - Networks
Presentation, Inference, and Ethics	Using graphics packages for data visualization, error sources specific to found (big) data, examples of big data analysis and erroneous inferences, inference in the big data context, big data and privacy, legal framework, disclosure control techniques, ethical issues, practical approaches.	Notebooks: -Imputing Missing Values for Machine Learning Presentation

Module 2. After completing the first module, teams worked to further develop their research/policy problem before they participate in the second module on data capture and curation. Module 2 focused on the acquisition of new data through web scraping and APIs, data (record) linkage, introduction to the use of relational databases, as well as a brief introduction to dealing with large amounts of data through Hadoop and Mapreduce frameworks. The emphasis here (and in subsequent) modules was on understanding why different data sources are combined, and why and when certain forms of record linkage and database structures are advantageous. Between modules 2 and 3, teams prepared data for their own projects either by creating the appropriate linked table from the data inside the secure environment or by augmenting their data with outside data added during this time to the secure environment.

Module 3. In the third module on modeling and analyses, the focus lay in introducing machine learning techniques, and analyses of networks and text data, depending on the project needs. In each case the focus was on the intuition and assumptions behind the techniques—more specifically, on what these intuitions and assumptions are, and when and why the techniques are applied. A considerable amount of class time was dedicated to the evaluation methodology for machine learning systems, especially as they are used in public policy applications. This included a discussion of methods to define and measure fairness and biases in machine learning models and how to mitigate the risk of decisions made using these models. To the extent it fit their projects, the teams applied those techniques to their data in the month following this module and evaluated them for shortcomings.

Module 4. In the fourth and final module, students learned about the presentation of data, and discussed inferential issues related to their projects, as well as ethical implications. In the visualization segment, the focus was on storytelling and communication over a full survey of all visualization techniques. A large portion of this last module was set aside for teams to work in groups on their projects. Participants provided project-focused peer feedback on their presentations, including comments on the usefulness of their given approach to the relevant agency or organization.

Class Structure

In each module, new topics or techniques were introduced with a lecture and problem sets fully formulated in the interactive notebooks, in order to allow the participants to explore the data with minimal code modifications, before later having to write their own code. Lectures, problem work, and project work were alternated throughout the day in roughly 1.5-hour slots. The problem sets had all the code and documentation developed, and were subsequently made available on GitHub (https://github.com/Coleridge-Initiative) after each class (removing the direct links and outputs from the confidential individual records).[12]

Collaborative Projects Creating Valuable Products

The focus on *projects that produce products* in addition to skills training is the third core element of the program. Rather than simply learning data science in the abstract, a major feature of the curriculum is that the program is structured around teams that work on issues of relevance and interest to them, and draw on the data available in ADRF and tools best suited to solve the problem. Creating teams with mixed skills allows for additional peer-to-peer teaching during the applied project work. We found that this approach has the advantage of (1) pulling together all

[12] All course resources are available for educators interested in adopting the course.

the main skill sets needed to use individual-level cross-agency data, (2) training participants in practical aspects of privacy and confidentiality, and (3) producing analysis that can be operationalized in many other policy contexts.

To determine the projects, agencies, researchers, and students proposed problems (particularly on employment, education, crime, and energy issues). Some of the course projects turned directly into implementation; the machine learning work was repurposed to predict outcomes as varied as sanitation truck breakdowns and recidivism due to technical parole violations. Some results increased our scientific credibility with the agencies, through published work in *Science* (Zolas et al., 2015), the *American Economic Review* (Buffington, Cerf, Jones, & Weinberg, 2016), and *Research Policy* (Chang, Cheng, Lane, & Weinberg, 2019).

Notably, the projects—which are based on real problems agencies are facing and use real data—enabled government employees (often with a social science or legal background) to consider (1) what data are available, and possible errors in that data; (2) what is being missed as data sets are linked; (3) how to draw inferences from new types of sample frames; and (4) how to address ethical issues and to protect privacy and confidentiality.

Three projects, From Prosecuted to Job Recruited: An Exploratory and Machine Learning Approach to Employment After Prison, Addressing Recidivism: Intervening to Reduce Technical Violations and Improve Outcomes for Ex-Offenders, and Mommy Don't Go: Predicting and Preventing Recidivism of Mothers in the Illinois Criminal Justice System, are highlighted on our website (https://coleridgeinitiative.org/training) and materials are available for download as well.

Outcomes

The program we described above was intended to achieve three outcomes:

1. To increase the capacity of public sector employees to use modern computational and data analysis methods and tools;
2. To improve the ability of government agencies to use their own data for evidence-based policymaking;
3. To provide authorized and secure access to administrative records from multiple agencies and states.

In this section, we highlight the results in each of these three key areas as a result of our program.

Increased individual skills

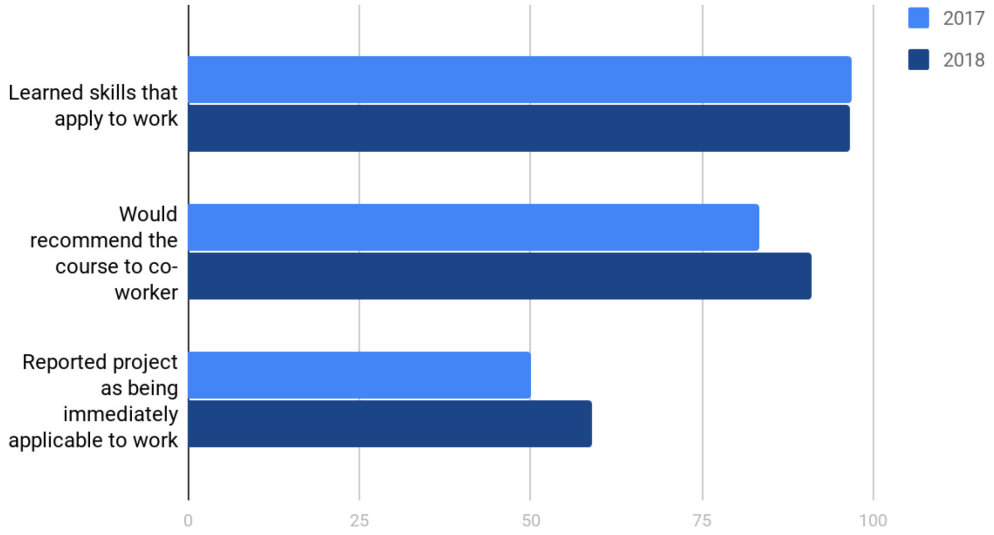
Based on the self-reports in the post-course evaluations, those responding to the evaluation consistently expressed contentment with the course, saw the skills learned as being applicable to their work, and would recommend the course to coworkers (Figure 2, $n = 31$ in 2017; $n = 56$ in 2018). More telling than the individual reports are the visible skills participants acquired.

Working closely with the teams gave us a good sense of the added skills and the deepening of the knowledge necessary to navigate new types of data, as well as new forms of analyzing them within the available infrastructure.

Initially just over half of the participants saw the class projects as being immediately applicable to their agencies. In response, the more recent classes have become more integrated with government agency leadership and the agencies themselves are fully engaged in defining the research questions. We also have found that agencies repeatedly send their employees to the

courses, again indicating satisfaction with the skill enhancement of those who already participated (Lane, 2016).

Fig 4.1 % of 'yes' answers from responding participants



Increased Access to Cross-agency and Jurisdiction Confidential Microdata

A key goal of our program was to motivate government agencies to link data across the agencies and get resources to use that linked data to solve policy problems. During the course of the program, employees from different agencies were exposed to data from multiple agencies that were relevant and useful for them. These employees were able to take advantage of this opportunity and frame projects that made effective use of linked data. They learned the value of linked data as well as the skills necessary to use them. The combination of corrections data with data from housing, employment services, and human services data provided teams with the opportunity to explore the link between neighborhood characteristics and access to jobs, as well as the earnings and employment outcomes of welfare recipients and the formerly incarcerated—

not to mention their subsequent retention of welfare and/or recidivism. The methods and access to new data sets allowed for new insights into approaches for reducing recidivism and dependence on welfare.

In addition, the classes generated new federal and state interest in sharing data across agencies and using it to tackle critical policy problems. The program has been increasingly adopted by federal and state agencies. For example, the U. S. Department of Health and Human Services has sponsored an initiative using welfare data (tanfdata.org); the U.S. Department of Agriculture has sponsored a class to inform child nutrition policy, and the states of Illinois, Indiana, Ohio, and Missouri have worked with us to create multiple classes on education and workforce transitions. Five more programs are planned for 2020.

In the current early stages, the success of the program has been demonstrated by substantial interest on the part of federal, state, and local governments in repeatedly participating in programs devoted to training, sharing data, and establishing dedicated research data facilities.

Increased Ability of Governments to Tackle Policy Problems More Effectively

By improving the skills of the individuals in the program, and prompting the use of linked data, we aim to enhance the short- and long-term capacity of government agencies to tackle policy problems.

To increase the sustainability of the work started in these classes, we awarded four formal fellowships to build upon this work, and to extend it to a product that the agencies would use. Two of the awardees focused on understanding recidivism outcomes for offenders with parole violations. This fellowship was provided to the Illinois Department of Corrections and is beginning to inform criminal justice policy. Another awardee used the course notebooks in order

to develop standardized measures of earnings and employment that can be (and have been) applied to any state's Unemployment Insurance wage records to generate comparable metrics across states. One of the participants in the recidivism project applied one of the course machine learning notebooks on recidivism to predict the probability of individuals returning to prison; another modified the notebooks to estimate the number of Kansas City trucks needing maintenance repairs.

In addition to the participants' subsequent application in their own agencies of what they learned in our program, they were also exposed to a larger network of potential collaborators. One of the participants from the City of Los Angeles City Attorney's Office was able to collaborate with the Civic Analytics Network and the University of Chicago Center for Data Science and Public Policy to use data science to enhance the process by which they identify chronic offenders, conduct background research necessary for individualized interventions, and provide prosecutors with tools to improve outcomes for offenders as well as the residents of Los Angeles.

Future Work

Ultimately, educational programs are measured by results, not just aspirations, however admirable the latter might be. While we have achieved our outcomes on the agency level, as evidenced by the interest that a growing number of government agencies have shown in repeatedly sending their students to us, as well as in sponsoring new programs, we have yet to measure the long-term effects of the program for individual employees, and to see the long-term effects of change through data in the agencies. The short time effects described in section four, however, make us optimistic in this respect.

We also plan to expand our initial assessment to go beyond the self-reports and include a knowledge test covering the class material. This will allow for a more fine-tuned sorting of participants to project groups, and serve as a basis for pre- and post-course evaluations. We are building capacity right now to monitor the access and usage of the asynchronous material, and we can use the resultant data as additional indicators of engagement, and as predictors of learning outcomes.

The upcoming courses will have pre-course surveys with agency heads, class participants, and randomly selected individuals at the agency. After three and after 12 months, all three groups will be surveyed again for a post-course assessment. In parallel agency, heads in states that have been identified as next-round participants will be asked to take the same survey and administer to potential participants.

After the initial start at the University of Maryland, New York University, and the University of Chicago, three more Universities (Ohio State University, University of Missouri and Indiana University/Purdue University Indianapolis) have now joined with us to run the courses with the developed material and the data provided through the ADRF. The next adopter will be University of California, Berkeley. Several foundations and agencies are using the material and infrastructure to run courses within their compounds and with their data, for example, in fall of 2019 at the National Science Foundation, and the U.S. Department of Agriculture.

Our Recommendations

Typical approaches to building the infrastructure we described are predicated on one-off personal relationships that are, all too often, fragile and unsustainable (Goerge, 2018; Lane, 2016; Potok, 2009). We recommend a national initiative that could potentially be supported by a

consortium of both public and private funders. That initiative would combine training programs (either by expanding on our pilot or developing new ones) and enable participants to access confidential data in secure remote access data centers for purposes approved by the relevant agencies. Many universities already boast centers that assist federal, state, and local governments, in the effective use of federal, state, and local data. Building on the existing capacity and interest offers the potential to scale ongoing research as well as develop and test innovative policy programs.

Should the initiative be professionally staffed, it could monitor and oversee the projects, their progress, and the lessons learned. It could also assume responsibility for creating and supporting a data infrastructure standards consortium, the role of which would be to create standards of technology, access, and privacy, and data structures for all projects.

In addition, we propose some recommendations as to how different types of organizations might build upon our work. We focus on recommendations for three types of organizations:

1. Government agencies participating in these programs through employee training and data access may enhance their ability to get access to data across jurisdictions, increase internal capacity to use this data, and develop new products that have value to their clients.
2. Universities participating in these programs will have an excellent way to support federal, state, and local governments, build collaborations with agencies that can support further research and education, and train participants using locally relevant problems and data.
3. Private foundations and federal funding agencies participating in these programs will have a means of connecting with local institutions, academics, and governmental agencies to work on shared local challenges. We believe that this approach helps to

address the technical and human challenges that governments face today, in part by ensuring that the government workforce is equipped with the right skills.

Conclusions

Data science can transform the ways in which governments design policy and improve outcomes for all of its citizens. In order to realize this potential, we need to fill two key gaps: (1) to enable access to data linked across agencies, and (2) to enable government workers to build capacity in using this data to solve critical problems. The applied data analytics program that we describe here begins to fill the infrastructure and skills gap, allowing government agencies to share confidential data and train their employees to tackle policy problems using modern computational and data analysis methods and tools across agency and jurisdictional lines. We have found that this program can accelerate the technical and analytical development of public sector employees. Our hope is to build a larger collaborative community of academic institutions, government agencies, and foundations. It is likely that the curriculum will be continually adapted. The program is designed to allow agencies to choose different focus areas, and technological advances will change how material can and should be taught. We are already working on new tools that allow class participants and the wider community to discover how the data was used by peers, and to build on past experiences and code. Our ultimate aim is to increase the capacity of governments to make more efficient and effective decisions.

Publicly Available Resources

The applied data analytics program described in this article has created a lot of resources that are available for others to use and extend under open-source license. These include:

- Github repository with lectures and interactive Jupyter notebooks available at

  (2019) Ohio State: https://github.com/Coleridge-Initiative/ada-2019-osu

  (2018) UMD: https://github.com/Coleridge-Initiative/ada-2018-umd

  (2018) Kansas City: https://github.com/Coleridge-Initiative/ada-2018-kcmo

  (2018) Chicago: https://github.com/Coleridge-Initiative/ada-2018-uchicago

  (2017) Welfare: https://github.com/Coleridge-Initiative/ada-2017-welfare

  (2017) Justice: https://github.com/Coleridge-Initiative/ada-2017-justice

  (2017) High Needs Pop.: https://github.com/Coleridge-Initiative/ada-2017-high-need
- Curriculum from each class and a sample of projects done as part of the classes are available at https://coleridgeinitiative.org/training
- Infrastructure Information about the ADRF can be found at https://coleridgeinitiative.org/resources

References

Abowd, J. M., Haltiwanger, J., & Lane, J. (2004). Integrated longitudinal employer-employee data for the United States. *American Economic Review*, *94*, 224–229.

Ackermann, K., Walsh, J., De Unanue, A., Naveed, H., Navarrete Rivera, A., Lee, S., Bennett, J., Defoe, M., Cody, C, Haynes, L., & Ghani, R. (2018). Deploying machine learning models for public policy: A framework. *In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '18)* (pp. 15–22). New York, NY: ACM. https://doi.org/10.1145/3219819.3219911

Alston, J., & Pardey, P. (1996). *Making science pay: The economics of agricultural R&D policy*. Washington, DC: American Enterprise Institute Press.

American Statistical Association Undergraduate Guidelines Workgroup. (2014). *Curriculum guidelines for undergraduate programs in statistical science*. Alexandria, VA: American Statistical Association. https://www.amstat.org/asa/education/Curriculum-Guidelines-for-Undergraduate-Programs-in-Statistical-Science.aspx

Arnold Ventures. (2018). *Policy labs*. Washington, DC: Author. Retrieved from https://www.arnoldventures.org/work/policy-labs/

Barbosa, L., Pham, K., Silva, S., Vieira, M., & Freire, J. (2014). Structured open urban data: Understanding the landscape. *Big Data Journal*, *2(3)*, 144–154. https://doi.org/10.1089/big.2014.0020

Buffington, C., Cerf, B., Jones, C., & Weinberg, B. A. (2016). STEM training and early career outcomes of female and male graduate students: Evidence from UMETRICS data linked to the 2010 census. *American Economic Review*, *106*, 333–338.

Castellani Ribeiro, D., Vo, H. T., Freire, J., & Silva, C. T. (2015). An urban data profiler. In *Proceedings of the 24th International Conference on World Wide Web* (pp. 1389–1394). New York, NY: ACM., https://serv.cusp.nyu.edu/~hvo/papers/2015_urban_profiler.pdf

Catlett, C., Malik, T., Goldstein, B., Giuffrida, J., Shao, Y., Panella, A., Eder, D., van Zanten,E., Mitchum, R., Thaler, S., & Foster, I. T. (2014). Plenario: An open data discovery and exploration platform for urban science. *IEEE Data Engineering Bulletin*, *37*(4), 27–42.

Çetinkaya-Rundel, M., & Rundel, C. (2018). Infrastructure and tools for teaching computing throughout the statistical curriculum. *The American Statistician*, *72*(1), 58–65. https://doi.org/10.1080/00031305.2017.1397549

Chang, W. Y., Cheng, W., Lane, J., & Weinberg, B. (2019). Federal funding of doctoral recipients: What can be learned from linked data. *Research Policy, 48,* 1487–1492.

Commission on Evidence-Based Policymaking. (2017). *The promise of evidence-based policymaking: Report of the Commission on Evidence-Based Policymaking.* Washington, DC: Commission on Evidence-Based Policymaking. Retrieved August 21, 2019, from www.cep.gov.

Culhane, D., Fantuzzo, J., Hill, M., & Burnett, T. C. (2018). Maximizing the use of integrated data systems: Understanding the challenges and advancing solutions. *The ANNALS of the American Academy of Political and Social Science*, *675*, 221–239. https://doi.org/10.1177/0002716217743441

Davenport, T. H., & Patil, D. J. (2012). Data scientist. *Harvard Business Review*, *90*(5), 70–76.

Ferreira, N., Poco, J., Vo, H. T., Freire, J., & Silva, C. T. (2013). Visual exploration of big spatio-temporal urban data: A study of New York City taxi trips. *IEEE Transactions on Visualization and Computer Graphics*, *19*, 2149–2158.

Foster, I., Ghani, R., Jarmin, R. S., Kreuter, F., & Lane, J. (2016). *Big data and social science: A practical guide to methods and tools*. London: Chapman and Hall/CRC.

Freeman, S., Eddy, S. L., McDonough, M., Smith, M. K., Okoroafor, N., Jordt, H., & Wenderoth, M. P. (2014). Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the National Academy of Sciences*, *111*, 8410–8415.

Goerge, R. M. (2018). Barriers to accessing state data and approaches to addressing them. *The ANNALS of the American Academy of Political and Social Science*, *675*(1), 122–137.

Goldsmith, S., & Kleiman, N. (2017). *A new city O/S: the power of open, collaborative, and distributed governance*. Washington, DC: Brookings Institution Press.

Gould, R., & Çetinkaya-Rundel, M. (2014). Teaching statistical thinking in the data deluge. In T. Wassong, D. Frischemeier, P. R. Fischer, R. Hochmuth, & P. Bender (Eds.), *Mit Werkzeugen Mathematik und Stochastik lernen–Using Tools for Learning Mathematics and Statistics* (pp. 377–391). Wiesbaden, Germany: Springer Spektrum. https://doi.org/10.1007/978-3-658-03104-6

Gould, R., Machado, S., Ong, C., Johnson, T., Molyneux, J., Nolen, S.,…Zanontian, L. (2016, July). Teaching data science to secondary students: The mobilize introduction to data science curriculum. Iase-Web.org https://iase-web.org/documents/papers/rt2016/Gould.pdf

Government Accountability Office. (2017). *High risk report:2020 Census*. Washington, DC: Author. https://www.gao.gov/highrisk/2020_decennial_census/why_did_study

Gutlerner, J. L., & Van Vactor, D. (2016). Catalyzing curriculum evolution in graduate science education. *Cell*, *153*, 731–736. https://doi.org/10.1016/j.cell.2013.04.027

Handelsman, J., Ebert-May, D., Beichner, R., Bruns, P., Chang A., DeHaan, R.,…Wood, W. B. (2004). Scientific teaching. *Science*, *304*, 521–522.

Hart, N., & Shaw, T. (2018, December 22). Congress provides new foundation for evidence-based policymaking [Blog post]. Washington DC: Bipartisan Policy Center. Retrieved from https://bipartisanpolicy.org/blog/congress-provides-new-foundation-for-evidence-based-policymaking/

Holdren, J. P., Marrett, C., & Suresh, S. (2013). Federal Science, Technology, Engineering, and Mathematics (STEM) education 5-year strategic plan. A report from the Committee on STEM Education National Science and Technology CouncilWashington, DC: Executive Office of the President National Science and Technology Council.

Iziarry, R. (2018, November 1). The role of academia in data science education [Blog post]. In R. Irizarry, R. Peng, & J. Leek (Eds.). Simplystatistics.org Simplystatistics.org. https://simplystatistics.org/2018/11/01/the-role-of-academia-in-data-science-education/

Jarmin, R., Lane, J., Marco, A., & Foster, I. (2014). Using the classroom to bring big data to statistical agencies. *AMSTAT news: The membership magazine of the American Statistical Association*, *449*, 12–13.

Lane, J. I. (2016). Big data for public policy: The quadruple helix. *Journal of Policy Analysis and Management*, *35*, 708–715.

Lane, J. I., Theeuwes, J., & Burgess, S. (1998). The uses of longitudinal matched employer/employee data in labor market analysis. *Proceedings of the American Statistical Association*. Alexandria: American Statistical Association.

Lane, J. I., Kendrick, D., & Ellwood, D. (2018). *A locally based initiative to support people and communities by transformative use of data*. US Partnership on Mobility from Poverty.

https://www.mobilitypartnership.org/locally-based-initiative-support-people-and-communities-transformative-use-data

Lane, J. I., & Shipp, S. (2007). Using a remote access data enclave for data dissemination. *The International Journal of Digital Curation, 1* (2), 128–134.

Mays, J. (2018). Building an infrastructure for evidence-based policymaking: A view from a state administrator. *The ANNALS of the American Academy of Political and Social Science*, *675*(1), 41–43.

McGuckin, R. H., & Pascoe, G. A. (1988). *The Longitudinal Research Database (LRD): Status and research possibilities* (No. 88-2). Washington, DC: U.S. Department of Commerce, Bureau of the Census.

Morgan, M. G., & Peha, J. M. (2003). *Science and technology advice for Congress*. New York: Resources for the Future. Routledge.

National Academies of Sciences, Engineering, and Medicine. (2017). *Innovations in federal statistics: Combining data sources while protecting privacy*. Washington, DC: National Academies Press. https://doi.org/10.17226/24652

National Academies of Sciences, Engineering, and Medicine. (2018). *Data science for undergraduates: Opportunities and options*. Washington, DC: National Academies Press. https://doi.org/10.17226/25104

National Academy of Public Administration. (2017). *No time to wait: Building a public service for the 21st century*. Washington, DC: Author. https://www.napawash.org/studies/academy-studies/no-time-to-wait-building-a-public-service-for-the-21st-century

Office of Management and Budget (OMB). (2019). *Federal data strategy*. Washington, DC: Author. https://strategy.data.gov

Pardo, T. A. (2014, June 16). *Making data more available and usable: A getting started guide for public officials*. Presentation at the Privacy, Big Data and the Public Good Book Launch, New York University, New York, NY.

Peng, R. (2018, December 11). The role of theory in data analysis [Blog post]. In R. Irizarry, R. Peng, & J. Leek (Eds.).Simplystatistics.org, https://simplystatistics.org/2018/12/11/the-role-of-theory-in-data-analysis/

Perez, F., & Granger, B. E. (2015). *Project Jupyter: Computational narratives as the engine of collaborative data science*. http://archive.ipython.org/JupyterGrantNarrative-2015.pdf

Potok, N. F. (2009). *Creating useful integrated data sets to inform public policy* (Doctoral Dissertation). The George Washington University, Washington, DC.

Reamer, A., & Lane, J. (2017). A roadmap to a nationwide data infrastructure for evidence-based policymaking. *The ANNALS of the American Academy of Political and Social Science*, *675*(1), 28–35. https://doi.org/10.1177/0002716217740116

Reamer, A., Lane, J., Foster, I., & Ellwood, D. (Eds.). (2018). Developing the basis for the secure and accessible use of data for high impact program management, policy development, and scholarship. *The ANNALS of the American Academy of Political and Social Science*, *675, 1, 28-35*.

Stokes, D. E. (2011). *Pasteur's quadrant: Basic science and technological innovation*. Washington, DC: Brookings Institution Press.

Warsh, D. (2010, December 13). A few words about the Vladimir Chavrid Award. [Blog Post]. Retrieved from http://www.economicprincipals.com/issues/2010.12.13/1209.html

Yarkoni, T., Eckles, D., Heathers, J., Levenstein, M., Smaldino, P., & Lane, J. I. (2019, January 25). Enhancing and accelerating social science via automation: Challenges and opportunities. https://doi.org/10.31235/osf.io/vncwe

Zolas, N., Goldschlag, N., Jarmin, R., Stephan, P., Owen-Smith, J., Rosen, R. F.,...J. I. (2015). Wrapping it up in a person: Examining employment and earnings outcomes for Ph.D. recipients. *Science*, *350*, 1367–1371.

Appendices

A. Textbook (living document – comments welcome)

B. Notebooks (living repository – individual course folders start with “ada”)

C. Coleridge Initiative (website – overall content collection)