Change through Data:

A Public Extension Program for Government Employees[1]

Frauke Kreuter
Joint Program in Survey Methodology, University of Maryland
School of Social Science, University of Mannheim, Germany
Statistical Methods Unit, Institute of Employment Research, Germany

Rayid Ghani
Center for Data Science and Public Policy
Department of Computer Science, University of Chicago
Harris School of Public Policy, University of Chicago

Julia Lane
Wagner Graduate School of Public Service, New York University
Provostial Fellow, New York University

[1] 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

New data types and access have fundamentally transformed the private sector. A similar level of change has yet to happen in the public sector, although exciting developments are on the horizon. However, two significant challenges need to be addressed before government agencies can succeed: (1) workforce capacity must be increased to effectively combine, curate and analyze data; (2) data must flow more easily across agency lines. This paper describes an approach that has successfully done both. The approach is inspired by the success of agricultural extension programs that use a problem-focused approach to create projects with value proposition attractive enough to surmount the many legal and technical hurdles that have historically prevented cross-agency data collaborations. We discuss the possibility to scale this effort nationally based on examples we have piloted with federal, state, and local employees.

Key phrases

Training programs; Evidence-based policy; Confidential data; Administrative Data Research Facility; government data

1. Introduction

The use of data at scale has profoundly changed business. The largest companies in the United States are now Apple, Amazon, Alphabet (Google), Microsoft, and Facebook —no longer General Electric or General Motors— because they have used data to transform the way value is created in the private sector (Galloway, 2017).

Although exciting developments are underway in the private sector, there have not yet been similar changes in the U.S. public sector. Past efforts to improve data use across the federal enterprise to improve the quality of evidence for policy making have been mixed (Hidalgo, 2016; Lee, Almirall, & Wareham, 2015). Recently, the U.S. government launched a Federal Data Strategy (Office of Management and Budget, 2019) informed by many stakeholders[2], and signed the Foundations for Evidence-Based Policymaking Act of 2018 into law.[3]

There are two significant barriers to transforming the government's use of data. The first is a lack of workforce capacity. Too few government employees have the requisite skills, and governments often do not have the salary flexibility to compete with the private sector to hire and retain enough in-house data analysts (National Academy of Public Administration, 2017). The second is that many policy problems require analyzing data that cross agency lines, and confidentiality rules prohibit data-sharing (Reamer, Lane, Foster, & Ellwood, 2018). Absent a clearly sufficient value proposition, it is difficult to obtain the resources necessary to surmount the many legal and technical hurdles that prevent cross-agency data collaborations. The time to reach agreement can take years—10 years in at least one case! (Potok, 2009). These combined challenges have led to the current Catch-22: because they 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 mean that they cannot demonstrate value.

The costs of this situation are apparent. In monetary terms, the cost to the tax payer (in current dollars) for the 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).[4] In human terms, the cost of not combining data (in a timely fashion) is also evident. Dr. Leana Wen, Commissioner of Health, City of Baltimore has noted, 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)

---

[2] Feedback was solicited by through the recent Federal Register notice and public-domain OMB memo M-19-01.

[3] https://www.congress.gov/bill/115th-congress/house-bill/4174 signed into law on January 14, 2019.

[4] https://www.gao.gov/products/GAO-17-664 For the 2020 Census, the Census Bureau plans to use administrative records (data that people have already voluntarily given to the federal government) to help improve its results and reduce some door-to-door visits. The Bureau estimated that using these data could save $900 million.

We argue it is possible to create a virtuous cycle of change within agencies whereby employees can share data, share knowledge, and apply modern technology to transform the way problems are solved across agency and jurisdictional lines. New environments have been built in which confidential data can be shared across agency lines. New educational programs are being designed for graduate and undergraduate students (National Academies of Sciences and Medicine, 2018b; National Research Council, 2015). To accelerate the development of the current workforce new executive education style courses for public sector employees are also needed, ideally in a way that not only provide technical trainings, but also demonstrate the potential value of working with a wide variety of data. We are inspired by the agricultural extension programs which have been a major source of the productivity growth in U.S. agriculture and suggest that a good way to start such programs in the data science space is to use a problem-focused approach (Peng & Parker, 2018) that generates results that both demonstrate clear impact and are highly visible. We also suggest the establishment of fellowship programs that can provide the basis for continuing the work seeded by these trainings, hardening the results into visible products, and building skills within the public sector. We describe a possible path forward with examples derived from a successful approach that we have piloted with federal, state, and local employees.

2. Challenges

The U.S. federal government has clearly recognized the importance of data to run government operations and policy (Commission on Evidence based Policy, 2017; National Academies of Sciences, Engineering, and Medicine, 2017; Office of Management and Budget, 2019); the same is true at the local level (Goldsmith & Kleiman, 2017; Mays, 2017). And Congress has recently passed legislation that puts the apparatus in place to do so (Hart & Shaw, 2018). Municipalities are also trying to build their own capacity for data science, with groups 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 and a Mayor's Office of New Urban Mechanics in both Boston, MA and Philadelphia, PA, while Chief Information Officers in cities as large as Chicago, IL and as small as Asheville, NC are taking steps to develop data science capacities (Pardo, 2014).

On the surface, making use of data in the public sector seems straightforward. Programs are administered, data are produced as by-products or collected intentionally, and outcomes can be analyzed and evaluated. Unfortunately, the reality is often different. 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 and no legal mandate to share information. Work by the information science, management information systems, and e-government research communities has documented barriers to value creation from open data platforms—including problems of diverse user needs and capabilities, the limitations of internally-oriented data management techniques, untested assumptions about information content and accuracy, and issues associated with information quality and fitness for use (Dawes & Helbig, 2010). There are additional technical issues before data can be analyzed, because the databases are often in different formats, with archaic data management systems and without common identifiers (Reamer & Lane, 2017).

When open data platforms are created, finding people who understand how to make scientific use of the data is often challenging (Barbosa, Pham, Silva, Vieira, & Freire, 2014; Castellani Ribeiro, Vo, Freire, & Silva, 2015; Catletta et al., 2014; Ferreira, Poco, Vo, Freire, & Silva, 2013). Agency employees with questions to ask of the data do not have the skills to analyze them. This gap is a major issue, since having capable, in-house data scientists who can demonstrate to their fellow civil servants the value 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). However, scaling the capacity building is a major challenge.

Agencies have, as a result, resorted to working with outside consultants or academic researchers to build new capacity. One of the authors of this paper, Julia Lane, started a new statistical program - the Census Bureau's Longitudinal Employer-Household Dynamics (LEHD) program - as a result of just such an approach (Warsh, 2010), with the subsequent collaboration of John Abowd, currently the Associate Director for Research and Methodology and Chief Scientist of the U.S. Census Bureau. However, reliance on outsiders is not a substitute for internal capacity, particularly as data become more complex. Outsiders too often 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 to properly scope and design the analysis. Lane was able to make progress because she was awarded an American Statistical Association Fellowship that enabled her to work inside the Census Bureau and connect with internal experts to understand what data were available and what was possible. However, 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 never implemented (Goerge, 2017).

In sum, governments lack both the human and technical infrastructure to securely store, combine, analyze and disseminate a variety of different administrative and other process data on human subjects, and make them accessible for replicability. The workforce issues are arguably the most serious, since it is easier to replace physical than human capital, given government human resource management constraints (National Academy of Public Administration, 2017). There are, however, several ways in which the data science community could be organized to address the challenges.

3. Data Science in the Public Service

On the technical side, it is now possible to make use of new infrastructures built to provide access to confidential government microdata. These infrastructures provide a technical solution that can be applied to link data across agency lines, so that agency staff can be trained to use real data to study actual agency problems. An overview is provided in the report of the Commission on Evidence based Policy (Commission on Evidence based Policy, 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 US. The Longitudinal Employer-Household Dynamics program which was established in the late 1990s has grown into a major national program (Abowd, Haltiwanger, & Lane, 2004; Burgess, Lane, & Theeuwes, 1998) that links state and federal data. The NORC/University of Chicago research data center, established in the mid 2000s provides researcher access to administrative data (Lane & Shipp, 2008) and the Laura and John Arnold Foundation has established policy labs in several key states (Laura and John Arnold Foundation, 2018). Integrated data systems have been funded by the U.S. Departments of Education and Labor (Culhane, Fantuzzo, Hill, & Burnett, 2017). Most recently, the U.S. Census Bureau commissioned New York University to establish an Administrative Data Research Facility to enable secure remote access for government data (Government Computer News Staff, 2018) in order to demonstrate to the Commission on Evidence Based Policy that such an infrastructure could be put in place.

On the skills development side, some notable activities have addressed workforce issues more generally. A recent National Science and Technology Council Five Year Strategic Plan called for graduate education to be designed to 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 who pay a substantial premium for data scientists (Patil & Davenport, 2012). Of course, the newness of data science as a field means that there is not a long history of knowledge of how to teach data science, and educating people how to access and analyze data has been a challenge (Gould & Cetinkaya-Rundel, 2013), although the community is starting to develop curricula and guidance (American Statistical Association, 2014; National Academies of Sciences and Medicine, 2018a), all the way down to secondary students (Gould, 2016).

Fortunately, a substantial literature exists on how to design programs for new sciences. In an influential series of papers, Handelsman and others argue that the new types of science need to adopt active learning techniques (Handelsman, Ebert-May, Beichner, & Bruns, 2004). By this they mean changing teaching from a lecture-based format to one which is *inquiry-based and modular* and which *treats students as scientists* who "develop hypotheses, design and conduct experiments, collect 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).[5]

[5] 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).

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 both interdisciplinary and driven by concepts. In the biological sciences, Gutlerner and Van Vactor argue forcefully for the development of modular classes – what they call "nanocourses"[6] (Gutlerner & Van Vactor, 2016). Rafa Iziarry (Harvard University), 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).

There is also a substantial complementary literature 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, which led to a "longstanding close association and integration of agricultural research with extension and higher education"(p9) (Alston & Pardey, 1996). The resultant agricultural extension programs built an operational framework, based on Pasteur's Quadrant (Stokes, 1997). Briefly, farmers in the field were challenged by real operational problems, agricultural researchers worked to collect data and develop methods to develop solutions, and the extension programs created the operational bridge between the two. State and local agricultural societies and institutes established farmer's institutes to "extend new technologies and improved and best practices from progressive farmers and trained scientists" (p16) (Alston & Pardey, 1996).

This element of exchange between those knowing how the problem originates (such as the farmer) and those charged with developing solutions, is also key in the Data Science work space. One of the key lessons for any aspiring data scientist is to "discover the data generating mechanism" (Peng, 2018). Training programs are a way to work with subject matter experts to determine mechanisms that work in each domain..

Robert Oppenheimer noted that "the best way to transmit knowledge is to wrap it up in a human being" (N. Zolas et al., 2015), and that principle has been well understood in both federal and state government (Peters & Savoie, 2000). 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, 2016); 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, Applied and Computational Mathematics, the American Mathematical Society, the Institute for Mathematical Statistics, Mathematics Association of America, and the Society for Industrial and Applied Mathematics.[7] 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. Prominent examples for successful IPAs in the Federal Statistical Agency space leading to new tools and products are Katharine Abraham (Bureau of Labor Statistics—American Time Use

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

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

Survey), Julia Lane (Census Bureau – LEHD), Rod Little (Census Bureau – Bayesian Methods), John Abowd (Census Bureau – Differential Privacy), to name just a few.

4. From predicting recidivism to truck failures: A training program

Over the last five years we have developed a training program which demonstrates how a data science program for public policy might be developed at scale to meet the quickly rising demand. The interest and investment into the program has been overwhelming: in the last two years the program drew over 250 participants from over 100 federal, state, and local government agencies.

The key elements of the program include (1) access to confidential agency microdata in a secure computing environment, (2) developing new skills, and (3) creating products that have value for government agencies. Participants work in a secure cloud-based environment and are introduced to new ways of collecting data and making use of new methods and tools. The overarching approach includes training government staff in how to keep fundamental statistical concepts like population frames, sampling, and valid inference but expand their skillset to include modern computational data analysis tools as well as using new types of data[8]. The program is built on a foundation of social science research principles integrated with current analytic and computer skills, though rooted in the study of real-world social and economic problems (Foster, Ghani, Jarmin, Kreuter, & Lane, 2016). Figure 1 provides an illustration of an extension program targeted at governments—a public extension program.

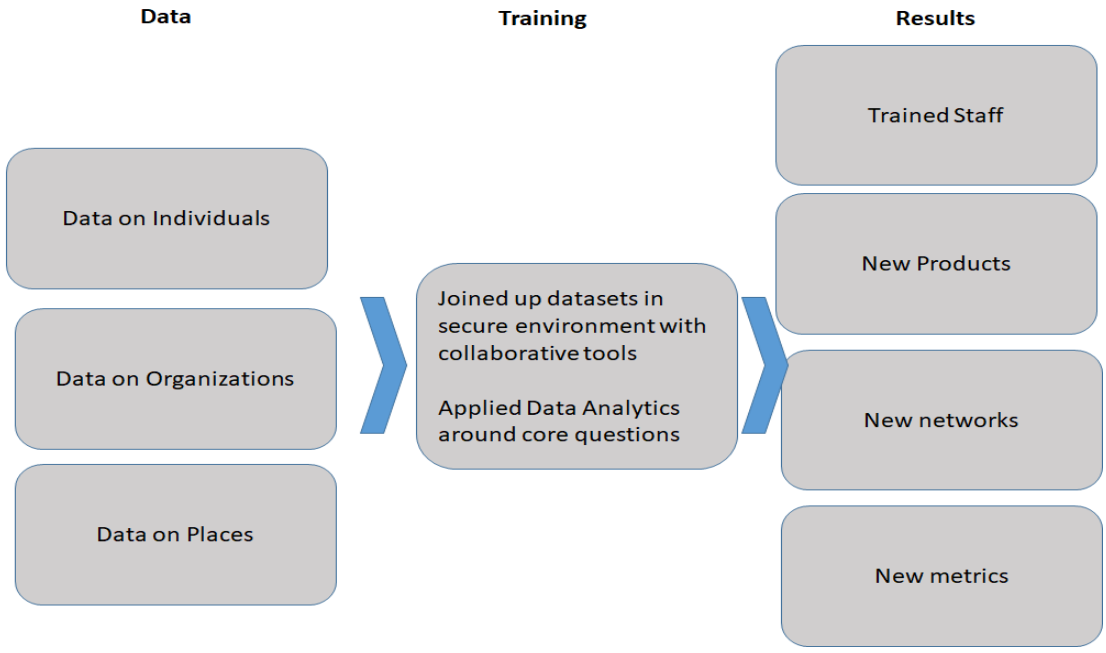

Figure 1: The "public extension" program visualized

[8] The new types of data include records from administrative processes, data captured from websites or via direct data exchanges through application programming interfaces (APIs), data with large spatial components or network structures.

4.1 Access to Confidential Agency Microdata

Because data access is at the core of replicable and reproducible science, we developed a team-centric work environment within a state-of-the-art data facility—the Administrative Data Research Facility (ADRF)[9] which was commissioned by the Census Bureau to inform the decision making of the Commission on Evidence Based Policy. ADRF staff created a single access-controlled project space that all project members can access from their own computer from any network via a secure client.[10] In order to eliminate unauthorized disclosure, access is limited to those with explicit permission. These shared project spaces allow users access to the same tools and provides them with the ability share their code, analysis output, and extracts from the data in a safe and intuitive way[11]. The tools are used for secure communication of project code and ideas in the ADRF.

The long-term goal is to build user interfaces that present rich context to users about the datasets as they work in secure environments, while also incentivizing users to contribute . These interfaces will be designed to gather metadata that provide 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 are willing to share their data within the ADRF because it creates a sandbox environment within which agency staff provide concrete evidence of the value of linking data as part of the course projects; as such, the access and use is consistent with the agency mission.

4.2 Developing New Skills

The cross-disciplinary curriculum integrates computer science with statistics and social science which is the essence of Data Science. The curriculum is designed to train participants in the creative and scientific use of available data. It covers all phases of addressing social issues including problem formulation, data collection, pre-processing, and analysis, as well as how to think about ethical and privacy issues that arise when tackling those issues. At the same time, new types of methods and data generated by computer scientists are introduced for their potential to transform the scientific understanding of the dynamics of human behavior.

The core of the program is administered in four in-person modules (see Table 1). Before the class begins, participants fill out a pre-class survey about their skills. This survey is used to group them into teams of four or fiveEach team member has a different skill set – one team member might have expertise in policy, another in statistics, another in coding, and another in data. The teams work together on problems and projects throughout the module series.

Interactive notebooks facilitate the introduction of code and learning of programming skills. At their core these interactive notebooks allow explanation, code, and output to be all visible in the

[9] https://coleridgeinitiative.org/computing
[10] https://www.nomachine.com/download
[11] See Cetinkaya-Rundel & Rundel (2017) for a summary of good computing infrastructure for education purposes.

same place, and code to be executed within the notebooks themselves. Here specifically, Jupyter notebooks are introduced as learning devices (Granger, 2015).[12] For those participants unfamiliar with command-line based languages and new to Python, an online inverted classroom Python and SQL bootcamp is offered prior to the course. We make use of Binder, an open-source web application for managing digital repositories. The Binder material is 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 though our experience is that for many it is hard to free up the time without an official course structure.

The first module of the in-person meetings covers the fundamentals of problem formulation, inference, and basic programming tools. It ties 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 includes 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. The discussion of the data generation process, quality frameworks, dealing with missing data, and selection issues plays and important part. A first graphical inspection of the data is used to familiarize everyone with the various data sources at hand.

Table 1 Applied Data Analytics Course Modules

Module	**Corresponding Course Learning Objective**	**Corresponding Notebook and Project Work**
Foundations of Data Science	Formulating research questions, matching research questions and data, "Big Data" - definitions, 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 and Github.	Notebook: Variables Worksheet: Project Scoping
Data Management and Curation	Introduction to APIs, building features from administrative record data, understanding earnings data, introduction to characteristics of large databases, building datasets to be linked, linkage in the context of big data, fundamentals of record linkage techniques, create a big data work flow, data hygiene: curation and documentation, practice SQL.	Notebooks: Record Linkage I+II 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

After completing the first module, teams work to further develop their research problem before they participate in the second module on data capture and curation, which includes in varying

[12] Juypter currently supports the programming languages Julia, Python, and R. Popular alternatives are for example R Markdown notebooks (https://rmarkdown.rstudio.com/) or Texdoc for Stata (Jann 2016).

detail the acquisition of data through web scraping and APIs, data linkage, database basics, as well as a brief introduction to programming with big data. The emphasis here (and in subsequent) modules is on the understanding why the different sources are combined, and why and when certain forms of record linkage and data base structures are advantageous. Between module two and three teams prepare data for their own research projects either by creating the appropriate linked table from the data inside the secure environment or by augmenting those with outside data added during this time to the secure environment.

In the third module on modeling and analyses, the focus lies in discussing machine learning techniques, analyses of networks, and sometimes text analysis, depending on the project needs. In each case the focus is on when and why the techniques are applied rather than on their theoretical underpinnings. A considerable amount of class time is dedicated to the evaluation of machine learning models, and public sector applications of such models as well as their risks. This includes the discussion of indicators to assess fairness and biases in machine learning models. To the extent it fits their projects, the teams apply those techniques to their data in the month following this module and evaluate them for their shortcomings.

In the fourth and final module, students learn about presentation of data, discuss inferential issues related to their projects, and ethical implications. In the visualization segment the focus is on story-telling and communication over a full survey of all visualization techniques. A large portion of this last module is set aside for teams to work in groups on their projects. Participants provide project focused peer feedback to teach team on their presentations including on the usefulness of the approach to their agency or organization.

In each module new topics or techniques are introduced with a lecture and a problem sets fully formulated in the interactive notebooks, 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 are alternated throughout the day in roughly 1.5-hour slots (see Appendix D). Each module has at least on lunch talk showcasing applications of the respective methodologies.

The problem sets have all the code and documentation developed, and are subsequently made available on github (https://github.com/Coleridge-Initiative) after each class (removing the direct links and outputs from the confidential microdata).[13] The results have been exceptionally positive, as the following personal communication from a class participant demonstrates.[14]

> "I am writing to let you know about a panel I was on at this year's RECS conference[15] which focused almost exclusively on content produced in your excellent data science courses! One team (from the New York City mayor's office) discussed a project they did to cluster welfare recipient job placements based on various measures of job quality available in administrative data. This project has sparked widespread diffusion of these methods within New York City government. […] Illinois Department of Human Services also presented on an interesting predictive modeling project which focused on

[13] All course resources are available for educators interested in adopting the course.
[14] The original email contained acronyms, replaced here for readability.
[15] http://recsconference.net/2018agenda_detail.htm

> Supplemental Nutrition Assistance Program recipients. Illinois is so enthused about the training and the Administrative Research Data Facility (ADRF) platform that they are building their own state ADRF […]. Finally, I spoke about some of MDRC data science initiatives which focus on building nonprofit capacity to do this work. Our approach was also influenced by the excellent material [..] learned in the classes. […] it is great to see how much of a ripple effect the classes are having across the human services community. What most excites me is that this is not just researchers. A lot of state personnel are using these techniques on the job. That will have a huge impact. These are the right courses at the right time. By democratizing data science the sky is the limit for data driven policy." (Rick Hendra, personal communication, June 3, 2018)

4.3 Creating Valuable Products

The focus on projects that produce products rather than mere skill 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 greatest interest to them, and draw on the data, models and tools best suited to solve the problem. Participants have some experience with statistics and applied data work coming into the program. 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 the main skill sets needed to use big data, (2) training participants in privacy and confidentiality, and (3) producing insights that can be used in many other policy contexts.

To determine the projects, agencies, researchers, and students nominated 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) and the American Economic Review (Buffington, Cerf, Ikudo, Lane, & Weinberg, 2017).

The focus on projects is a key difference from most courses in computer science departments. The projects—which are built on real problems with real data—enable government employees (often with a social science or legal background) to consider (1) what data are available and possible error; (2) what is being missed as data sets are linked; (3) how to draw inference from new types of sample frames; and (4) how to address ethical issues and protect privacy and confidentiality.

Three projects are highlighted on our website (https://coleridgeinitiative.org/training) and are available for download. The title of each of them show the nature of the work: "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." One of the participants in the recidivism project

actually then applied one of the course notebooks on recidivism and the machine learning approach on predicting the probability of individuals returning to prison to calculating the Kansas City trucks needing maintenance repairs.

Four formal fellowships were awarded by the program. Two of the awardees focused on the analysis of technical evaluations of parole and the reports on technical violations of parole. This attracted attention by the Illinois Department of Corrections and is beginning to inform policy. Another awardee used the course notebooks 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.

In addition, the classes generated new federal and state interest.. At the federal level, the TANF (Temporary Assistance for Needy Families)Data Innovations project, was initiated by the Department of Health and Human Services, and began to institutionalize our approach by engaging state agencies to support the innovation and efficiency of state-level TANF programs by enhancing the use of data from TANF and related human services programs programs. quite complicated in structure; T he project, which is joint with the nonprofit evaluation organization, MDRC, Chapin Hall at the University of Chicago and the University of Pennsylvania has three main objectives: (i) to advance the creation and understanding of and access to integrated data in the service of helping states to address high priority questions they have identified, (ii) to facilitate innovative analytic approaches to these integrated data; and (iii) to institutionalize the capacity built by cultivating analytic mindsets through training programs and by installing governance structures that will outlast the project. At the state level, the state of Illinois has requested its own Illinois Administrative Data Research Facility so that the Departmentsof Corrections, Human Services, and Employment Security, among others, can share data across agency lines.

Evaluating the effectiveness of the approach using standard methods like randomized controlled trials is not possible, because of the verythe nature of the program However, the success of the program is demonstrated by the interest on the part of federal, state and local governments to participate in the training, share data, and establish their own research data facility. In addition, we logged robust usage statistics within the research data facility and got enthusiastic responses from pre- and post-class surveys about the usage of the skills learned back on the job. We also have found that agencies repeatedly send their employees to the courses, again indicating satisfaction with the skill enhancement of those that already participated (Lane, 2016). Going forward, several universities are interested adopting the program or integrating it into their educational offerings, notably Ohio State University and the University of Indiana, with two others also indicating interest in learning from our approach.

5. A forward-looking agenda

Both private foundations and federal funding agencies have indicated interest in the community extension approach. It should be possible for data scientists to develop a successful two-pronged public extension approach based on addressing both the technical and human challenges so that

the interests of a disparate group of data providers are addressed and that the government workforce is equipped with the right skills.

An integral part of this infrastructure the institutionalization of public ecosystem; the current approach which is predicated on one-off personal relationships is fragile and unsustainable (Goerge, 2017; Lane, 2016; Potok, 2009). In the United States, for example, a national initiative could potentially be supported by a consortium of both public and private funders. That initiative would combine training programs that expand on our successful pilot andenable participants to access confidential data in secure remote access data centers for purposes approved by the relevant agencies. .. Many universities already boast centers that are deeply engaged with federal, state and local governments, assisting them 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.

At the national level, the initiative should be professionally staffed, and led by an executive director. The director would be charged with establishing a process for identifying state and local projects eligible for large-scale investments, with local champions, that have high promise for key policy and research issues of interest.

The initiative would monitor and oversee the projects, their progress, and the lessons learned. It would also be responsible for creating and supporting a data infrastructure standards consortium whose role would be to create standards of technology, access and privacy, and data structures which would be adopted by all projects.

Building a longer-term sustainable basis for public sector work presents an enormous opportunity for the field of data science. Put simply, curricula could be developed that fill the gap between the skills needed to work with the new types of data and the skills embodied in the extant public sector workforce. There is very real potential for data science to be as transformational in changing the way in which governments do business as it has been in the private sector—but potentially with higher rewards. Enormous sums of tax-payer money will not need to be expended to count the population, and fewer children need to be at risk for lack of data analysis.

References

Abowd, J. M., Haltiwanger, J., & Lane, J. (2004). Integrated Longitudinal Employer-Employee Data for the United States. *American Economic Review*, *94*(2), 224–229.

Alston, J., & Pardey, P. (1996). *Making science pay*. American Enterprise Institute.

Association Statistical Association (2014). Curriculum guidelines for undergraduate programs in statistical science. *Retreived from Http://Www. Amstat. Org/Education/Curriculumguidelines. Cfm*.

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.

Buffington, C., Cerf, B., Ikudo, A., Lane, J., & Weinberg, B. (2017). *Research Funding and the Foreign Born*. New York.

Burgess, S., Lane, J., & Theeuwes, J. (1998). The Uses of Longitudinal Matched Employer/Employee Data in Labor Market Analysis. In *Proceedings of the American Statistical Association*.

Castellani Ribeiro, D., Vo, H. T., Freire, J., & Silva, C. T. (2015). An Urban Data Profiler. In *WWW2015 Workshop on Web Data Science and Smart Cities* (p. to appear).

Catletta, C., Malika, T., Goldsteina, B., Alessandro, J. G. Y. S., van Zanten, P. D. E. E., & Fosterc, R. M. S. T. I. (2014). Plenario: An Open Data Discovery and Exploration Platform for Urban Science.

Cetinkaya-Rundel M., Rundel C. (2017). Infrastructure and tools for teaching computing throughout the statistical curriculum. https://peerj.com/preprints/3181/

Commission on Evidence based Policy. (2017). *The Promise of Evidence-Based Policymaking*. Washington DC. Retrieved from www.cep.gov

Culhane, D., Fantuzzo, J., Hill, M., & Burnett, T. C. (2017). 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*(1), 221–239. https://doi.org/10.1177/0002716217743441

Dawes, S., & Helbig, N. (2010). Information Strategies for Open Government: Challenges and Prospects for Deriving Public Value from Government Transparency. In M. Wimmer (Ed.), *Electronic Government: Lecture Notes in Computer Science,*. EGOV 2010, LNCS 6228.

Ferreira, N., Poco, J., Vo, H., Freire, J., & Silva, C. (2013). Visual Exploration of Big Spatio-Temporal Urban Data: A Study of New York City Taxi Trips. *IEEE Transactions on Visualization and Computer Graphics*, 2149–2158.

Foster, I., Ghani, R., Jarmin, R. S., Kreuter, F., & Lane, J. I. (2016). *Big Data and Social Science: A Practical Guide to Methods and Tools*. Taylor & Francis Group. Available on https://coleridge-initiative.github.io/big-data-and-social-science/

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*(23), 8410–8415.

Galloway, S. (2017). *The Four: The Hidden DNA of Amazon, Apple, Facebook, and Google*. New York, NY, USA: Portfolio/Penguin.

Goerge, R. M. (2017). 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. https://doi.org/10.1177/0002716217741257

Goldsmith, S., & Kleiman, N. (2017). *A New City O/S: The Power of Open, Collaborative, and Distributed Governance*. Brookings Institution Press.

Gould, R., Machado, S., Ong, C., Johnson, T., Molyneux, J., Nolen, S., Tangmunarunkit, H., Trusela, L, Zanontian, L. (2016). Teaching Data Science to Secondary Students: The Mobilize Introduction to Data Science Curriculum. Proceedings of the International Association for Statistical Education 2016 Roundtable on Promoting Understanding of Statistics About Society. Berlin, Germany, July 2016. (http://iaseweb. org/Conference_Proceedings.php?p=Promoting_Understanding_of_Statistics_about_Society_201)

Gould, R and Cetinkaya-Rundel, M. (2013). Teaching Statistical Thinking in the Data Deluge. In *Mit Werkzeugen Mathematik und Stochastik lernen-Using Tools for Learning Mathematics and Statistics*, Wassong, T., Frischemeier, D., Fischer, P., Hochmuth, R., Bender, P (eds). Springer International Publishing.

Government Accountability Office. (2017). *High Risk Report:2020 Census*. Washington DC.

Government Computer News Staff. (2018). Data mashups at government scale: The Census Bureau ADRF. *GCN Magazine*.

Granger, B. E. (2015). Jupyter and JupyterHub. Retrieved from https://jupyter.org/

Gutlerner, J. L., & Van Vactor, D. (2016). Catalyzing Curriculum Evolution in Graduate Science Education. *Cell*, *153*(4), 731–736. https://doi.org/10.1016/j.cell.2013.04.027

Handelsman, J., Ebert-May, D., Beichner, R., & Bruns, P. (2004). Scientific teaching. *Science*, *304*(5670), 521.

Hart, N., & Shaw, T. (2018). Congress Provides New Foundation for Evidence-Based Policymaking. Washington DC: Bipartisan Policy Center. Retrieved from https://bipartisanpolicy.org/blog/congress-provides-new-foundation-for-evidence-based-policymaking/

Hidalgo, C. A. (2016). What's Wrong with Open-Data Sites--and How We Can Fix Them. *Scientific American*. Retrieved from https://blogs.scientificamerican.com/guest-blog/what-s-wrong-with-open-data-sites-and-how-we-can-fix-them/

Holdren, J. P., Marrett, C., & Suresh, S. (2013). Federal Science, Technology, Engineering, and Mathematics (STEM) Education 5-Year Strategic Plan. *National Science and Technology Council: Committee on STEM Education*.

Iziarry, R. (2018). The rol of Academia in Data Science Education. Irizarry, R., Peng, R., & Leek, J. (Eds.) Simplystatistics.org. 2018/11/01.

Jann, B. (2016). Creating LaTeX Documents from within Stata using Texdoc. *The Stata Journal*, *16*(2), 245–263. https://doi.org/10.1177/1536867X1601600201

Jarmin, R., Marco, A., Lane, J., & Foster, I. (2014). Using the Classroom to Bring Big Data to Statistical Agencies. Amstat News.

Lane, J. (2016). Big data for public policy: The quadruple helix. *Journal of Policy Analysis & Management*, *Summer*.

Lane, J., Kendrick, D., & Ellwood, D. (2018). *A Locally Based Initiative to Support People and Communities by Transformative Use of Data*. Washington DC. Retrieved from https://www.mobilitypartnership.org/locally-based-initiative-support-people-and-communities-transformative-use-data

Lane, J., & Shipp, S. (2008). Using a remote access data enclave for data dissemination. *International Journal of Digital Curation*, *2*(1).

Laura and John Arnold Foundation. (2018). Policy Labs. Washington DC. Retrieved from http://www.arnoldfoundation.org/initiative/evidence-based-policy-innovation/policy-labs/

Lee, M., Almirall, E., & Wareham, J. (2015). Open data and civic apps: first-generation failures, second-generation improvements. *Communications of the ACM*, *59*(1), 82–89.

Mays, J. (2017). 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. https://doi.org/10.1177/0002716217739275

McGuckin, R. H., & Pascoe, G. A. (1988). *The longitudinal research database (LRD): Status and research possibilities*. US Department of Commerce, Bureau of the Census.

Morgan, G., & Peha, J. (2016). *Science and Technology Advice for Congress*. Routledge.

National Academies of Sciences, Engineering, and Medicine (2017). *Innovations in Federal Statistics: Combining Data Sources While Protecting Privacy*. Washington, DC: The National Academies Press. https://doi.org/10.17226/24652

National Academies of Sciences and Medicine, E. (2018a). *Data science for undergraduates: opportunities and options*. National Academies Press.

National Academies of Sciences and Medicine, E. (2018b). *Envisioning the Data Science Discipline: The Undergraduate Perspective: Interim Report*. Washington, DC: The National Academies Press. https://doi.org/10.17226/24886

National Academy of Public Administration. (2017). *No Time to Wait: Building a Public Service for the 21st Century*. Washington DC.

National Research Council. (2015). *Training Students to Extract Value from Big Data: Summary of a Workshop*. (M. Mellody, Ed.). Washington, DC: The National Academies Press. https://doi.org/10.17226/18981

Office of Management and Budget. (2019). Federal Data Strategy. Washington DC. Retrieved from https://strategy.data.gov

Pardo, T. (2014). Making Data More Available and Usable: A Getting Started Guide for Public Officials. Retrieved from http://cusp.nyu.edu/data-privacy-book/

Patil, T., & Davenport, T. (2012). Data Scientist: The Sexiest Job of the 21st Century. *Harvard Business Review*.

Peng, R., & Parker, H. (2018). Not so Standard Deviations (Episodes 63-70). Retrieved from http://nssdeviations.com/

Peng, R. (2018). The Role of Theory in Data Analysis. Irizarry, R., Peng, R., & Leek, J. (Eds) *Simplystatistics.org*, 2018/12/11.

Peters, G., & Savoie, D. J. (2000). *Governance in the twenty-first century: revitalizing the public service*. McGill-Queen's Press-MQUP.

Potok, N. F. (2009). *Creating useful integrated data sets to inform public policy*. THE GEORGE WASHINGTON UNIVERSITY.

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. (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*.

Stokes, D. E. (1997). *Pasteur's quadrant : basic science and technological innovation*. Washington, D.C.: Brookings Institution Press.

Warsh, D. (2010). A Few Words about the Vladimir Chavrid Award.

Yarkoni, T., Eckles, D., Heathers, J., Levenstein, M., Smaldino, P., & Lane, J. I. (2019). *Enhancing and accelerating social science via automation: Challenges and Opportunities*.

Zolas, N., Goldschlag, N., Jarmin, R., Stephan, P., Owen-Smith, J., Rosen, R. F., … Lane, J. I. (2015). Wrapping it up in a person: Examining employment and earnings outcomes for Ph.D. recipients. *Science*, *350*(6266). https://doi.org/10.1126/science.aac5949

Zolas, N., Goldschlag, N., Jarmin, R., Stephan, P., Smith, J. O.-, Rosen, R. F., … Lane, J. I. (2015). Wrapping it up in a person: Examining employment and earnings outcomes for Ph.D. recipients. *Science*, *350*(6266), 1367–1371. https://doi.org/10.1126/science.aac5949

SECTION IV: APPENDICES

A. Textbook (living document – comments welcome)
B. Notebooks
C. Coleridge Initiative
D. Generic course outline