NASPAA Data Science Curriculum for Public Service

Summary of Proposed Approach at Cornell University's Institute for Public Affairs

Elizabeth Day, Maria Fitzpatrick, Thomas O'Toole

"What are the elements of successful curricula or pedagogical models that both develop capacity to solve public policy problems and leave explicit space for local experimentation and modification?"

Recent decades have seen exponential growth in the collection and amassing of data. Although many people, particularly social scientists, tend to think in terms of traditional survey sources when thinking about data, the current explosion of data creation and availability includes data from an ever-growing variety of sources: administrative data, internet, video, audio, sensor, and data from businesses. These new data provide new opportunities for making better informed decisions and provide better information to citizens, yet relatively few policymakers and public servants have the appropriate training to do so.

Many institutions of higher education, including a handful in the public policy space, have focused their data science training efforts on creating "teched-up" data scientists. In this proposal, we argue that the lack of basic understanding of data science by most public affairs graduates, and the associated inability to manage high-quality data use by governments and business, necessitates ***basic data science training for all students*** instead of, or in addition to, specialized data science tracts.

If understanding and use of data are limited to just those with specialized skills, "data" will remain a siloed and foreign concept in the public sector and we will never realize its true potential for improving citizens' lives. ***In order to truly revolutionize the way the public sector and business in related industries use data, everyone in public affairs must know what data is***

and how to use it. If management cannot develop support for data use and build consensus about proper data use, the data will go unused or it will get used improperly. If policy analysts cannot work with data scientists to create systems adequately designed to address policy needs, those systems will never be created. To create and implement policy effectively will require *all* public servants and people in public affairs-related occupations to truly understand data' promise and pitfalls, and how to use it effectively.

Basic training in data science for all students must include deep understanding of 1) privacy, confidentiality and the ethics of data use; 2) the basics of data (sources, format, linkages); 3) question formulation and data analysis; 4) data visualization; and 5) dissemination. Our proposed curriculum provides students with this understanding through introductory coursework, follow-on modules, and integration throughout other coursework. Hands-on, lab-based, and experiential learning are crucial features built into the curriculum so that students can learn directly from real-world experiences.

In what follows, we first provide more information on the challenge at hand, including the wide array of data available to policymakers and the difficulties using the data effectively. We then turn to i) an overview of the current state of training in data science for students and professionals in public service and public affairs, including discussion of the specialized training now offered at a handful of top public administration programs across the nation, ii) an overview of the current needs and gaps of training government staff, and iii) examples provided by discussions with our alumni. We then outline our own curriculum proposal. We conclude with a discussion of how public policy schools and NASPAA can encourage collaboration and facilitate sharing of ideas so that high-quality data science training for public service-related professionals can become more widespread.

Considerations for Data Science and Public Policy

Data is now available from a vast range of sources, including surveys, administrative records, sensors, online chat forums, internet usage, phone and video recording, the Internet of things, and more. These data have the potential to be used in a multitude of ways to improve public decision-making and ultimately better citizens' lives. The Partnership for Public Service, notes that advanced analytics can make policies and programs more effective and efficient by 1) ensuring that data is usable, 2) improving the transparency and accessibility of information for stakeholders, 3) enhancing the ability of agencies to forecast, and 4) allowing agencies to better identify and prevent fraud, waste, and abuse (Partnership for Public Service 2019).

However, the new data are not without pitfalls and policymakers and public servants must take care to use data appropriately. For example, each of these data sources involves questions of privacy, confidentiality, and ethics. How can the government ensure that health records created through administration of health insurance are kept private while still allowing for synthesis of the records? Can sensor data tracking citizen movements within a city be useful while simultaneously limiting the ability to track individual citizens? Can responses to online commentary and phone conversations also be synthesized without revealing citizens?

Additional questions arise around whether or how to connect data from different sources in order to form a more complete portrait. For example, would it be ethical, or even possible, to connect administrative data and censor data in order to more efficiently provide services to program users? What are the challenges in combining various sources of data in order to target services most efficiently? When is it ethical to use predictive risk modeling techniques?

Still more concerns arise when considering how to use these new forms of data. Are the data from online forums and other outlets representative of the general population? Who may be

left out of such forums? How can we appropriately translate and catalog data from forums, such as audio or video, or any other text form for that matter?

Unfortunately, the public sector is currently ill-equipped to handle these questions and to make use of the vast amount of data available. For example, in a survey of 283 federal employees, while 78 percent of respondents conveyed that data is integral to their roles, more than 60 percent reported that their agencies were either not effective or only somewhat effective at leveraging data. 57 percent strongly agreed that their agency needs to invest more in training personnel in data analytics (SAS/GovLoop 2014). Recognizing this shortcoming at the federal level, one can only speculate the extent of the shortcoming in state and local governments, as well as outside the United States. As Gamage (2016) notes, similar concerns have been expressed regarding public sector organizations in the United Kingdom, Australia, and India. It is very likely that gap is therefore a global phenomenon, and one that is largely attributable to a lack of university and professional education in this space (Yerak 2013). Improving our training of existing and the future generations of public servants in data science is crucial.

Current Training in Data Science for Public Servants

Most training in data sciences occurs either within schools of information science (and possibly computer science) or in narrowly defined spaces within other disciplines. In recent years, there have been calls from various social science disciplines and beyond for greater attention to data science and the potential that new sources of data provide (Athey 2018, Brady 2019, Mergel 2016). These summary papers highlight many examples of how new sources of data have been useful at providing new information on society that can inform policy design and

implementation. They also provide many examples of how new data can be used to evaluate the effectiveness of policies.

However, translating this increased interest among academics to increased knowledge for public policy-related training has been slow to occur. For this proposal, we surveyed the website of the top 25 public policy and public affairs programs in the country as ranked by U.S. News and World Report in 2019.[1] Among these MPA and MPP programs, we found ***none that required all students to take even basic coursework*** in data science related areas. Instead, there were only a handful that offered specialization tracts in data science and data analytics. Others indicated that they allowed students to take coursework in the area as part of their required coursework more broadly. And, while a few of these programs offered coursework in data science and analytics designed for public policy students, most relied on coursework offered elsewhere on campus (e.g., in a School of Information Studies or Sciences).

This is also true at Cornell, where our MPA students have access to courses throughout the University, including the renowned Computing and Information Sciences. However, our own experience is that many, if not most, students come to the MPA program without enough existing technical and quantitative skills to take coursework from these pother programs. And indeed, since many of them do not want to be Chief Technology Officers or even data analysts (see below), it makes little sense for them to spend their professional training time on a deep dive into the specialty. Instead, it is most important for them to understand the data and tools available, their uses and limitations, so that they can use them to improve their policy- and decision-making.

[1] https://www.usnews.com/best-graduate-schools/top-public-affairs-schools/public-affairs-rankings (August 13, 2019)

Current Needs in Government: Overview and Case Studies

In support of our proposal to supplement the current expansion of specialty training in data scientists in public policy and public affairs programs with proficiency training for all students, our survey of the public service landscape revealed that two types of training will be necessary to ensure the public sector and related industries become 'fully functioning' with data. First, there is the already recognized need to train data scientists with an understanding of public policy and the needs of the public sector. Second, there is a need for even the least technical public servants to become proficient and up to date in their understanding of the potential benefits and the complexities of data. The need for training across both of these dimensions is supported by a study by the McKinsey Global Institute that projects that "the United States alone faces a shortage of 140,000 to 190,000 people with deep analytical skills, as well as 1.5 million managers and analysts to analyze big data and make decisions based on their findings" (Manyika et. al. 2011, p. 3). This shortage is compounded in government, which has historically lagged several generations behind the private sector in ensuring technical skills proficiency for a modern workforce.

Understanding that data were going under-utilized by the federal government, Congress passing the Foundations for Evidence-Based Policymaking Act (FEBPA) in 2018. This legislation, which followed years of lagging behind the private sector in strategic planning, business intelligence, and data analytics, seeks to establish data as a critical public asset, as well as *promote more strategic thinking around data in government*. To further the goals of FEBPA, in July 2019, the Office of Personnel Management formally announced the creation of a "data scientist" job classification as a means of both encouraging the addition of "Chief Data Officers" to agency c-suites and attracting support staff in this strategic area (Wagner 2019).

Nonetheless, federal, state, and local government still faces considerable challenges in luring those with data analysis proficiency away from the private sector. There are several reasons for this shortcoming, including overall lack of qualified applicants for these roles, lack of human resource expertise for evaluating applicants for technical positions, and a cumbersome application process relative to the private sector (Anastasoff et. al. 2018). Though not the focus of this proposal, creating high-quality specialized training in data science for public servants is a necessary first step to adequately populating government with data scientists with the "deep analytical skills" noted in the McKinsey report.

Promoting more strategic thinking around data in government will require more than just an army of data scientists. Specifically, the SAS/GovLoop survey of federal agency employees mentioned above found that, beyond data skills and knowledge of technical solutions, the most prominent deficiency respondents identified at their agency was the lack of both "hard and soft" management skills required to bring data to bear. Soft skills identified as critical included the ability to generate consensus around data among agency leadership, utilizing critical thinking and problem-solving approaches to data, strategizing around how to align data to agency mission, fostering inter-agency collaboration, assessing workforce needs around data, and developing performance indicators (Florenza 2014). Thus, traditional graduate and professional programs training technologists, like those in computing or information sciences departments, are unlikely to produce employees capable of addressing the data skills gap. Instead, professional programs aimed at public service professionals must work to provide training in both technical skills and management and leadership for a world with increasing technical and data-related needs.

To get a better understanding of how these shortfalls in training may be playing out in the public sector and in the public policy world more generally, we turned to our alumni for examples of how data science has been useful in the public sector, or in private sector firms that engage with the public sector, and for information on what tools and training are most useful. Below, we describe some of the most salient parts of those discussions. Going forward, our goal will be to flesh out specific examples from alumni into case studies for use in training students.

Several of our alumni working in government agencies, with those agencies as consultants, or in private sector firms affected by policy described similar experiences of how data skills at multiple levels were required to create new datasets to address shortfalls of existing data. Some of these datasets were built to track transient populations that had not been captured by existing efforts, thereby enabling the agency to provide better services to these marginal populations. Other datasets were built by linking data across agencies (and sometimes across jurisdictions) so that fraud could be more easily detected. Still others were data built in order to enable companies to be sure they were in compliance with existing and new regulations. Employees with the involved organizations had to identify the drawbacks of the existing system, convince colleagues and collaborators that a new system was needed, and work with data scientists to design a new system. This required contribution from both data scientists and others, all of whom had to be capable enough at understanding data, and is potential, to build support for the projects and work together to make them successful. The involved alumni specifically discussed that, although data scientists are required to build and merge systems and conduct analyses, they oftentimes lack the policy knowledge to thinking through how systems should be architected. As such, they recommended that training in both basic data science skills

(e.g. what is data, how is it used) and data leadership and policy thinking skills will be essential for training public affairs students.

Our alumni at the management level in the public sector and in consulting firms that work with the public sector echoed how basic data science training is useful for all types of MPA graduates. Data are increasingly used to assess and benchmark market trends, to form the basis for arguments for how to move plans forward, and to assess whether initiatives have worked. However, most management consultants have not had much training in data science. Without proper training in understanding data, and the methods used to collect and analyze it, the data may be used improperly and interpreted incorrectly. Therefore, basic skills are needed even for those in management who do not expect to use data themselves.

These conversations with alumni have helped us shape our proposed path forward in data sciences training. As detailed below, going forward, we will work with our alumni and community partners to be sure our curriculum continues to provide the skills useful to public service.

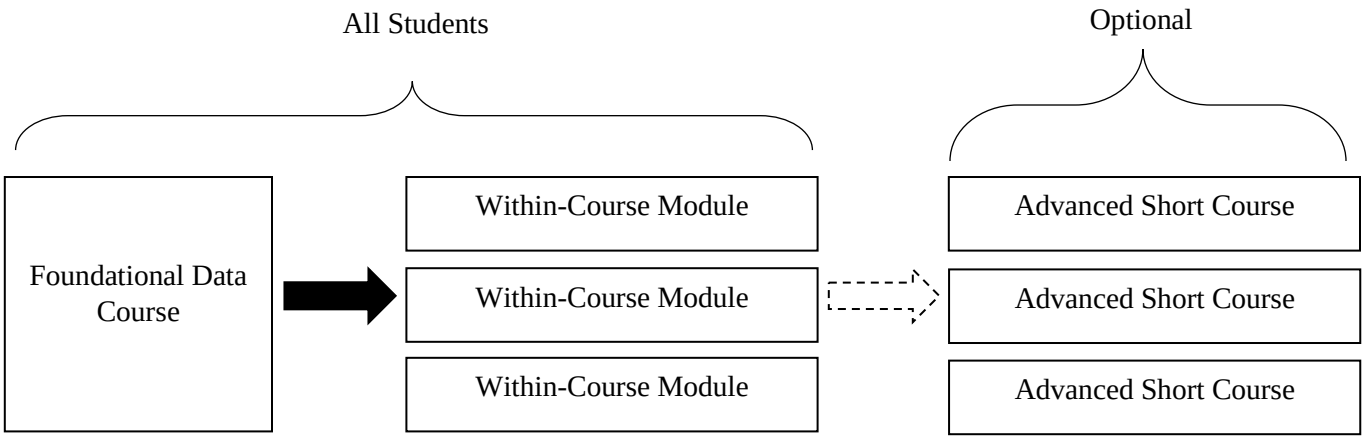
The Cornell Data Science and Public Policy Curriculum Proposal

The Cornell Data Science and Public Policy Curriculum seeks to meet these pressing needs for public administrators with foundational data science skills. There are five key elements that shape our proposed curriculum: scaffolded instruction, integration into existing curricula, hands-on learning, connection to local needs and policymaking, and design thinking.

Core Elements of Proposed Basic Training in Data Science

Scaffolded instruction: As is true in many programs, our students arrive with various levels of prior knowledge of data analysis. We recognize the importance of scaffolding instruction to meet students at their level and plan to provide gradual progression of skills throughout their coursework. We envision this happening through a full introductory course on data to provide foundational skills, followed by short courses and mini-modules students take over the remainder of their program, based on their areas of interest (see Figure 1).

Figure 1. Curricula Model: Foundational Course, Modules, and Short Courses



Integration into existing curricula: For students to truly build data science skills, it will be useful for them to employ those skills repeatedly. As such, we are redesigning our coursework in all areas to include some components of data science. In one example we will be asking students to create data visualizations for local and state legislators for our course called Bridging Research and Policy. Similarly, students will work on doing data analysis and visualizations in our Microeconomics, Statistics, and Public Finance Coursework. Case studies aimed at teaching leadership and management skills for situations involving big data will be woven into public

administration, ethics, leadership, and other management courses. We will also be asking students to incorporate their data science skills as part of their experiential learning in courses like Consulting and Capstone. With this repeated use and integration across courses and throughout their training, students will become even more skilled at using data to answer questions and understand policy problems.

Hands-on learning: We foresee coursework that provides a broad, but hands-on introduction to a variety of analytic techniques, including those described below. These tools and techniques would be introduced using open-source software (R, Python) to allow for transferability to more advanced courses on relevant techniques, and to align with new industrial and public sector standards. In order to truly understand the relevant concepts, it is paramount that students have practice in using them. Therefore, organizationally, we anticipate coursework structured to involve both lecture- and lab-work in a lab-based setting. In this way, students will be able to 'get their hands dirty' in a way that will allow them to conceptualize the relevant topics using detailed examples.

Connection to local need and policymaking: Along with hands-on lab-based learning, we also plan to incorporate data projects that meet the needs of local stakeholders and policymakers. Students will have the chance to design a culminating project, separate from a capstone, to answer real-world questions. We envision these projects leveraging our existing collaborations with local organizations, the Cooperative Extension system across the state, and both local and state policymakers, through which we have access to datasets and relationships with stakeholders with unanswered questions.

Design thinking: Students will also learn and implement design thinking throughout their coursework. The design thinking process, which emphasizes understanding audience needs,

defining key questions with the end goal in mind, and brainstorming and creating solutions to meet client needs (Brown, 2008), maps well onto current curriculum in Consulting and Capstone, and will be integrated throughout hands-on learning experiences.

Regular review and updating of curriculum: Our design of this curricular proposal involved a close review of the existing demand for data-related skills in the public sector, in part through consultation with our alumni. Going forward, we regularize a process for consultation with alumni and community partners that will inform us of industry standards and ongoing and developing needs, as well as help to provide up to date and relevant case studies for coursework.

Core Competencies Covered in the Proposed Basic Training in Data Science

There are several technical and statistical skills essential for a basic data science public policy curriculum. We also have grouped these into five main areas: ethics and privacy, data basics, data analysis, data visualization, and dissemination.

Ethics and privacy: What are key considerations for data ethics and privacy? What issues arise when thinking about using existing administrative data – ownership, confidentiality, networks and partnerships, communication, etc.?

Data basics: What are data? What forms can data take? What are possible sources of data? How do we think about linkages and connections between data from different sources? This includes direct and fuzzy matching, as well as thinking about combining data at different levels (e.g. information about children with their school and neighborhood characteristics).

Data analysis: What questions can we answer with which types of data? How do we go from knowing what datasets exist to using it to say something useful for decision-making, as well as how to go from a question stakeholders want answered to an informative answer using

existing data? Both of these require both an understanding of statistical analysis (means, t tests, etc.) and an understanding of the assumptions underlying various uses of data and different forms of causal inference.

Data visualization: What approaches exist for data visualization (e.g., charts, tables, figures, maps)? When is it useful to make an interactive visualization?

Dissemination: How can data be used in a way that lay audiences will understand? How would use of data vary across different formats (e.g., policy briefs, internal reports, presentations, webpages, email)?

Working Together Going Forward

As our curriculum proposal highlights, any resolution of the critical need for data scientists at all levels of government, and in policymaking more generally, must represent a collaborative effort between academia and the public, private, and nonprofit sectors. Several concrete steps that public policy schools and NASPAA could take include the following:

Coordinating Government Relations Efforts Around New Recruitment Programs: Similar to NASPAA's successful efforts to encourage transparency and reform around the Presidential Management Fellowship (PMF), NASPAA could coordinate a government relations initiative with public policy schools to both raise awareness of the need for data scientists in government, as well as encourage government to invest in recruitment programs that address this critical need. This might involve, for instance, the development of special recruitment programs modeled after PMF that expose professionals to challenges in data science on a rotational basis across agencies.

As competition for talent with the private sector will likely be a challenge that government will face in this area, public policy schools and NASPAA should advocate for a

streamlined recruitment process that will allow government to effectively compete with the private sector for top candidates. One replicable model to consider is the recruitment and selection process used by United States Digital Service (USDS). Created in 2014 as a means of closing personnel gaps which led to the failure of healthcare.gov, USDS adopted a private sector recruitment model emphasizing 1) oversight of the recruitment process by a diverse group of specialized recruiters (as opposed to human resource generalists), 2) candidate evaluation by subject matter experts, and 3) constant monitoring and evaluation of the recruitment process itself to ensure parity with private sector best practices (Anastasoff et. al. 2018). *In summary, NASPAA and policy schools should work together to advocate for a government recruitment and selection process that is effective at attracting those with the skills sets required to close the data science gap.*

Encourage Dual Degree Programs: While NASPAA has traditionally focused on accreditation for MPA and MPP degree programs, this paper demonstrates that both strong general management and technical skills are required to effectively address the data science gap in government. With this in mind, NASPAA and policy schools should work together to encourage dual degree programs for students interested in data science careers. This might include dual MPA/MPP-Master of Data Science programs, or MPA/MPP-Master of Engineering Programs. *Ideally, this would entail some incorporation of data science teaching or professional development in the NASPAA accreditation process itself, or perhaps even a separate accreditation process for schools seeking to develop specializations in data science.*

Serve as a Clearinghouse for Program Curricula/Syllabi: Since data science is a "new frontier" for many public policy schools and NASPAA, closing the skills gap will require considerable guidance on program curricula and syllabi to ensure schools are teaching the

cutting-edge skills required of data science professionals. NASPAA and public policy schools at the forefront of this initiative could, for example, develop a digital clearinghouse of program curricula and syllabi (for both courses dedicated to data science, as well as general management instructors looking to integrate data science content into their syllabi). Such a clearinghouse might also include cases illustrating real world pitfalls and best practices in this area, as well as resources for non-specialists looking to integrate data science content into their courses. NASPAA might also sponsor Annual Conference events around teaching/professional development best practices in this area.

Sponsor Data Science Simulations/Hackathons: Similar to the successful NASPAA-Batten Simulation, NASPAA and policy schools could jointly sponsor data science simulations or hackathons. These events would allow students to test their data skills in a time-sensitive environment, as well as collaborate with peers across MPA/MPP programs.

Committee/Section Development: Due to the technical and rapidly changing nature of the data science space, many of these recommended initiatives presume that a body of fully invested professionals will be assembled to continuously advise public policy schools and NASPAA on emerging trends and best practices. Such a committee would require more than faculty—data science professionals from across sectors should be invited to participate to ensure policy schools are not myopic in addressing this complex need. These committee/section members could support advocacy efforts, help design accreditation standards for public policy schools in this area, write cases that can be used in the classroom, and evaluate simulation/hackathon participants. Regular committee/section meetings should be coordinated by NASPAA and policy schools to ensure ongoing feedback, in addition to a dedicated committee/section meeting at the Annual Conference. Public policy schools could also leverage participation as a means of

engaging alumni, which would, in turn, generate internship, capstone, and job opportunities in the data science space.

References

Anastasoff J, Smith J, Stier M. (2018). Mobilizing tech talent: hiring technologists to power better government. Retrieved from www.ourpublicservice.org.

Athey, S. (2018). The impact of machine learning on economics. Draft chapter, National Bureau of Economic Research, Cambridge MA.

Brady, (2019). The challenge of big data and big science. American Review of Political Science. 22:297-323.

Brown, T. (2008). Design thinking. *Harvard Business Review, 86*(6), 84-92.

Florenza, P.. (2014). Closing the data and analytics skills gap. *Government Workforce in Focus*. Retrieved from www.govloop.com.

Gamage, P. (2016). New development: leveraging 'big data' analytics in the public sector. *Public Money and Management*, *36(6)*, 385-390.

Manyika J, Chui M, Brown B. et. al. (2011). Big data: the next frontier for innovation, competition, and productivity. Retrieved from www.mckinsey.com.

Mergel, I. (2016). Big data in public affairs education. Journal of Public Affairs Education. 22(2):231-48.

Partnership for Public Service (2019). Seize the data: using evidence to transform how federal agencies do business. Retrieved from www.ourpublicservice.com.

Wagner, E. (2019, July 1). OPM announces new 'data scientist' job title. *Government Executive*. Retrieved from www.govexec.com.

Yerak, B. (2013, August 25). In growing field of big data, jobs go unfilled. *Chicago Tribune*. Retrieved from www.chicagotribune.com.