Journal of Behavioral
Public Administration
Vol 6, pp. 1-31
DOI: 10.30636/jbpa.61.330

Roundtable Article

# Administrative Informatics: A Roundtable on the Conceptual Foundations of a Public Administration-Centered Data Science Subfield

**Michael Overton*, Stephen Kleinschmit,** Mary Feeney***,
Federica Fusi**, Nick Hart†, Spiro Maroulis***,
Kayla Schwoerer♦, Eric Stokan♦♦, Herschel F. Thomas♦♦♦,
and Samuel Workman♦♦♦**

**Keywords:** Data Science, Administrative Informatics, Scientific Knowledge Accumulation, Data Infrastructure, Information Systems

## Defining Administrative Informatics: An Applied Data Science Subfield for the Public Sector —Michael Overton and Stephen Kleinschmit

The increasing prevalence of advanced data applications in the public sector requires the formation of a new subdiscipline that acknowledges, examines, and leverages the transformative nature of data—Administrative Informatics. The goal of informatics fields is to understand applied data applications from a broader scientific discipline, like public health informatics (an offshoot of epidemiology), which focuses on data applications that promote population health. The creation of informatics fields often coincides with emergent data needs that require too much domain-specific expertise for non-experts to address. An examination of various informatics fields suggests that they not only originate around data needs but employ a narrow range of sophisticated applications tailored to extract insights from field-specific data sources.

Urban informatics professionals are not broadly trained data scientists but practitioners with a cultivated data science toolbox for studying urban problems. Administrative informatics must similarly anchor itself in the pursuit of studying data and its applications in the public sector, applying these insights to practice, and developing a data science toolbox specifically for public administration and management. Further, the field must distinguish itself from other manifestations of data science inquiry within the public sector, such as policy analytics, urban informatics, and computational social science.

We believe two core features of administrative informatics will both situate it as an informatics field and distinguish it from other approximate lines of inquiry. First, administrative informatics should have a substantive focus on data and its administrative applications in the public sector. Specifically, this means studying how data and the use of advanced information technologies affects administrative decision-making and the logics of public service delivery—a task well-suited for behavioral public administration scholars. Second, we propose a

*University of Idaho, **University of Illinois Chicago, ***Arizona State University, †The Data Foundation,
♦Vrije Universiteit Amsterdam, ♦♦University of Maryland Baltimore County, ♦♦♦West Virginia University

Address correspondence to Michael Overton at moverton@uidaho.edu.

new approach to scientific knowledge production inspired by the discussions at the 2022 Public Management Research Conference workshop "Developing the Data Science of Public Administration Scholarship," the *Journal of Behavioral Public Administration's* submission standards, informatic journal submission standards, and emergent opportunities arising from the growth of quantitative data. The accumulation of scientific knowledge in administrative informatics needs to be rigorous, defensible, and open to new forms of scientific knowledge.

The purpose of this roundtable is to present, argue, and evaluate the role of administrative informatics in behavioral public administration. This essay is broken up into three sections. The first section introduces the substantive focus of administrative informatics and how it can be studied. The second section introduces the central concepts required to establish a new approach to scientific knowledge production. The final section provides a short overview of the contributions of the roundtable and nests them within the arguments established in this essay.

**Substantive Focus: Data and Administrative Decision-Making**

Administrative informatics broadly focuses on the data processes and tools used to deliver public services. Data create imperfect representations of complex realities and quantify socially defined constructs. The widespread adoption of data-driven technologies is rapidly transforming society, and understanding its effects on the public sector is a natural progression of public administration and management scholarship. Understanding data and its applications will contribute valuable insights into the evolving logics of administrative decision-making as advanced data applications become standard in practice.

For behavioral public administration scholars, a specialized focus on individuals in decision-making is paramount. Current research on data's effects on administrative decision-making is incomplete. Existing research identifies important aspects of behavioral/technological interfaces but is primarily grounded in psychology. Administrative informatics provides a lens to examine the evolving bases of administrative decision-making and its influences on the constituent components of managerial behavior. Prior public technology movements, such as digital era governance and e-government, focused on digitalization and the use of information technology for service delivery and transparency. However, its scholarship is anchored in internet applications, not data. There are open questions concerning how 1) individuals use data in decisions, 2) decision-making is influenced by the capture, curation, analysis, and application of data, and 3) data-driven decision-making affects the sensemaking process and service delivery.

There is much to learn about shifts in administrative behavior resulting from the use of advanced data technologies, from the evolving representations of complex social phenomena to the shifting basis of ethical decision-making. Conversely, many existing theories may need reexamination if data practices render them obsolete, potentially setting off a paradigmatic shift in the field's understanding of administrative behavior.

**Scientific Knowledge Production: Four Principles of Administrative Informatics**

Public administration broadly and behavioral public administration more narrowly would benefit from a thoughtful examination of the current state of scientific knowledge production. Currently, the production of scientific knowledge in public administration is characterized by omnibus articles and a preference for confirmatory studies using quantitative methods. Unfortunately, these practices both unduly tax field researchers and also limit the timely generation of innovative and actionable research.

We propose a more agile, rigorous, and relevant approach to building a body of scientific evidence for public administration. The science of administrative informatics is grounded in four principles: Segmentation of knowledge production, diversity of evidence, the centrality of hard and soft data infrastructure, and the rigorous use of design science to connect scholarship to practice. Together, these principles can advance the production of scientific knowledge and improve the field's understanding of the behavioral foundations of administrative decision-making.

**Segmentation of Knowledge Production**

A limitation surrounding current scientific practices in public administration is an overreliance on omnibus articles, which we characterize as lengthy tomes where all parts of a research cycle are compiled into a single, comprehensive manuscript (Fig.1). Omnibus articles create far-reaching problems for public administration scholars attempting to accumulate knowledge around a guiding research question, particularly for research anchored in emerging technologies. First, articles become incredibly long. Significant contributions throughout such manuscripts get buried in lengthy prose or removed to meet journal word limit requirements. Second, omnibus articles are taxing on reviewers. It is unrealistic to ask reviewers to meaningfully evaluate theory development, measurement approaches, empirical methods, and the nesting of any results within a broader literature *in every reviewed article*. The prevalence of omnibus articles hurts the accumulation of scientific knowledge because important incremental contributions at various stages of the research cycle are underprovided, poorly developed, and inconsistently evaluated in peer review.

The segmentation of research is an essential component of many applied technology fields and provides a framework for timely knowledge development in an era of rapidly accelerating technological change. To remain relevant, public administration scholarship must evolve to value the timeliness of information and value iterative forms of knowledge creation that are common in other applied disciplines.

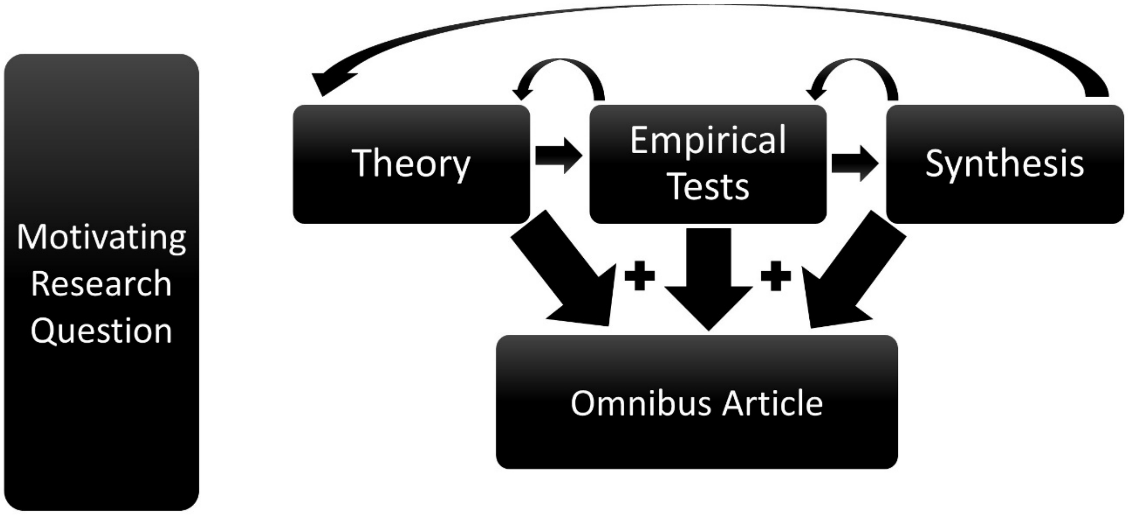

**Figure One: Current State of Knowledge Production**

An agile and adaptive administrative informatics requires the development of a broad body of scientific knowledge and is a necessary foundation for the following three principles in this essay. Segmenting omnibus manuscripts into focused contributions will help reinforce and strengthen every part of the research cycle. Allowing shorter pointed articles on theory, measurement, empirical testing, application, replication, synthesis, etc., enhances the rigorousness and defensibility of a scientific body of knowledge. The emergence of *Perspectives on Public Management and Governance* as an outlet for theoretical and conceptual manuscripts, and the *Journal of Behavioral Public Administration's* range of submission types and short article lengths suggest that the field of public administration is trending in this direction.

### Diversity of Evidence

Not only are there benefits to segmenting research into shorter manuscripts focused on different stages of the research cycle, but the growth of quantitative data and new analytical methods also increases the diversity of beneficial empirical evidence. The proliferation of causal inference methods likely comes at the expense of other empirical approaches. Defensible empirical evidence does not require a causal inference framing to be a valid scientific contribution. Descriptive, exploratory, evaluative, associational, simulation, and prescriptive studies can add meaningful quantitative evidence to the study of behavioral public administration.

Administrative informatics must build a diverse body of empirical studies beyond confirmatory analysis (Fig. 2). The dominance of confirmatory studies in top journals undermines theory development and creates perverse incentives for authors to frame their research as confirmatory when their research would make a valuable descriptive, exploratory, or associational contribution. Administrative informatics strives to be a "big tent" field of inquiry, advancing knowledge accumulation and innovation through rigorous data-driven evidence–confirmatory or otherwise. This practice would align with research trends in other professional fields, which generate high-impact scholarship outside of confirmatory approaches such as the *Academy of Management's* journal *Academy of Management Discoveries.*

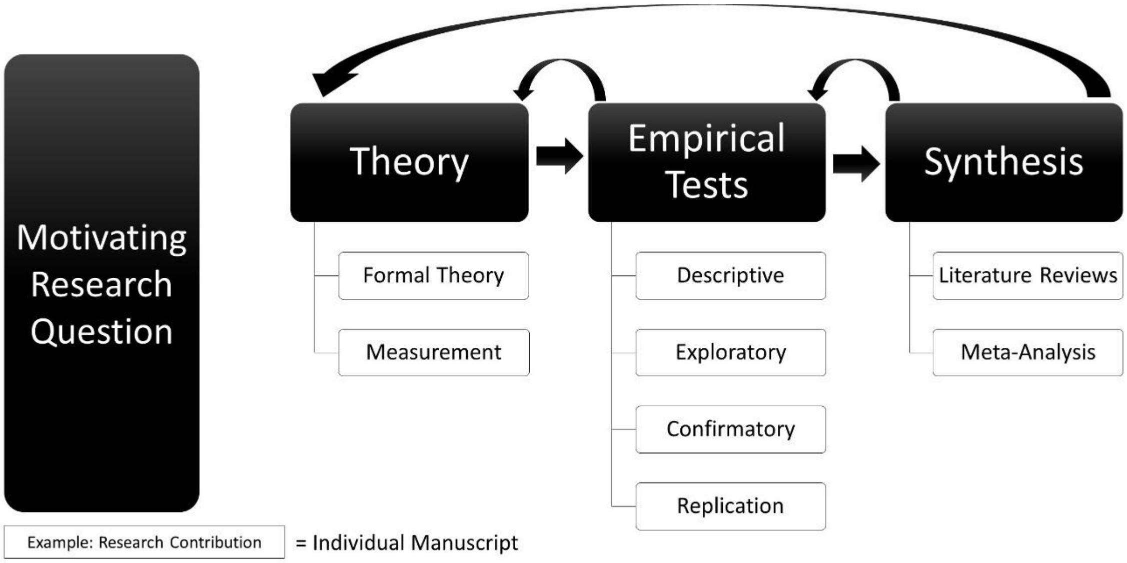

**Figure Two: Ideal State of Knowledge Production**

**Data Infrastructure**

Top informatics journals often have specific submission opportunities and standards for articles that support the development of their field's data infrastructure. Data infrastructure refers to the hard (ex., data repositories) and soft (ex., data science skills) infrastructure that enables reproducible and useful behavioral public administration research. For example, the *BMC Medical Informatics and Decision Making* journal encourages database and software article submissions.

A rigorously evaluated data infrastructure (Fig. 3) will help further the production of scientific knowledge. The quality of empirical studies can be improved through articles profiling data collection procedures, new datasets, repositories, and software applications, in addition to benchmarking commonly used datasets. A well-developed soft data infrastructure opens the doors to possible manuscripts that inform modeling decisions, strengthening the rigor of empirical work. Further, manuscripts demonstrating the value of data applications and technologies can help connect scholarly research to practitioner needs.

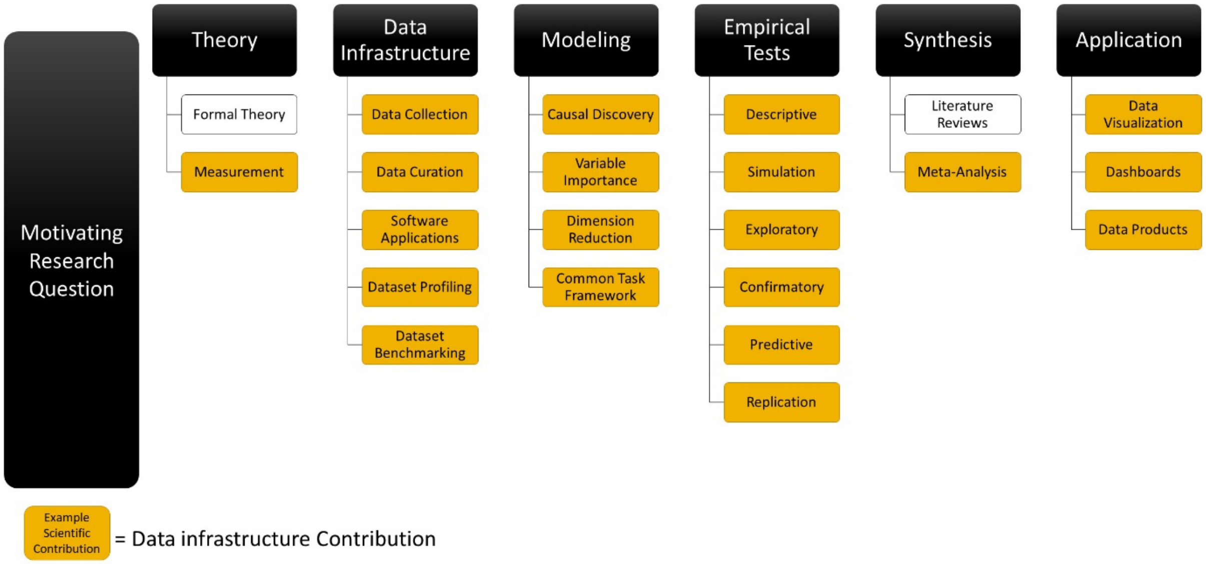

**Figure Three: Data Infrastructure and Data Science Contributions in Knowledge Production**

## Design Science

As a socially constructed field, administrative informatics requires developing new solutions to problems facing practitioners through the acceptance of design science contributions. Design science originated with a familiar name, Herbert Simon, and presents an approach to studying artifacts—constructs created by people, not nature—to address problems in professional fields of study. An embrace of design science will not crowd out theory-driven science. Instead, it develops complementary methods to test and apply theory. Within design science approaches, a problem is identified, and a solution (artifact) is proposed, tested, evaluated, and modified in iterative cycles. An embrace of artifacts and design science would allow scholars to generate new solutions to public problems grounded in scientific knowledge rather than passively evaluating solutions put forth by practitioners.

A design science approach to the development and evaluation of artifacts also provides a rigorous framework to propose, test, and evaluate normative concerns in public administration. Espousing values or public service paradigms in research manuscripts has a rightful place in administrative informatics as it does in public administration scholarship. Design science also provides a framework for studying normative arguments as artifacts. Advocating for public service values is a tradition in public administration scholarship that should not be abandoned in administrative informatics. Instead, scholars should treat these debates as technical artifacts requiring measurement, evaluation, and feedback.

## Roundtable Overview

The essays in this roundtable discuss a range of topics relevant to the development of Administrative Informatics and its impact on behavioral public administration. The roundtable begins with an essay from Drs. Samuel Workman and Herschel Thomas who discuss fundamental tensions that emerge from different types of data infrastructure systems used by governments. Next, Dr. Mary Feeney provides a thoughtful counterargument to the development of another subdiscipline while outlining a path toward accomplishing many of the proposed goals of Administrative Informatics within the current institutional structures of the field.

The remaining six essays overview specific topics within the development of a broad data infrastructure. Dr. Spiro Maroulis outlines when and how domain-specific public administration knowledge can be applied to different computational social science methods. Dr. Stephen Kleinschmit argues that developments in data technology enable public administrators to consider the long-ignored spatial aspects of public problems, driving a methodological shift towards spatial causal inference. Dr. Federica Fusi discusses the need to consider how data affects the populations and communities it measures as data infrastructures are built. Dr. Eric Stokan provides prescriptive advice on developing a protocol for collecting social media data and how such protocols can create a "research commons" for scholarly communities. Dr. Kayla Schwoerer argues that effective communication is a critical bridge required to connect data to decision-making, service delivery, and evidence-based policymaking. Finally, Dr. Nick Hart delivers an overview of Federal legislation guiding the use of data and suggests ways the research community can help support governmental data efforts.

## Data Systems, Information Processing, and Government Learning in Space and Time —Samuel Workman and Herschel F. Thomas

Government information processing is incredibly important for understanding the dynamics of policy change (Workman et al., 2009). Most major public policy theories assume that feedback from the outcomes of policy decisions fuels ongoing efforts to address public problems (Weible & Sabatier, 2017). Feedback comes in the form of systems thinking (Baumgartner & Jones, 2009), policy learning (Weible et al., 2022), citizen response to policy (SoRelle, 2020), or reinforces a particular path of policy reform (Pierson, 2004). Feedback requires information, or data, on policies and outputs or outcomes. The types of data infrastructures that are the linchpin in policy theories receive scant attention.

Meanwhile, in public administration, the study of data infrastructures usually boils down to performance information, necessarily looking inward at organizational routines and procedures and not outward at problems. Often, this means studying outputs rather than outcomes. The triumph of "efficiency" in organizational language means that we have better data on organizations than the problems they are meant to solve.

In part, these limitations derive from our democratic traditions and skepticism of government. The fundamental tension in all democracies between democratic accountability and problem-solving government generates two distinct data infrastructures. Concerns for accountability favor *"look-up"* data infrastructures. While these are useful in fostering representation and citizen accountability, they are not organized in a way to allow policy learning because they hamper data analysis and integration. Data infrastructures built for *analysis* are not as useful for accountability but allow governing systems to integrate information across problems and better solve them. Accountability concerns generate data infrastructures incompatible with learning from policy decisions, severely curtailing the feedback loop that allows policies to adapt to changing problems.

Of course, this state of affairs prevents the development of good social science and hinders the standing of social science in larger policy debates. The data infrastructures built out of concern for accountability encourage policy reform by anecdote. The social sciences collectively devote little resources to applied social science - the lion's share of funding is for basic science (e.g., National Science Foundation). In the disciplines, there is little reward for public-facing, applied, high-quality social science. Under these conditions, social scientists are less apt to adapt "look up" data structures for analytic purposes and much less apt to integrate these with other types of data (e.g., public health). The lack of integration of public policy data with other types is a tremendous impediment to problem-solving, especially at the local governance level.

The implications are beyond academic concerns. Data infrastructure is key to understanding what, if anything, governing systems can learn from their past or peers. Policy learning requires comparison over space or time. We argue that data infrastructures borne of accountability and organizational efficiency concerns are a major impediment to policy learning and are even more acute for complex interdisciplinary problems. The tension between systems built for accountability or analysis has dire consequences for interdisciplinary research on complex problems lying at the juncture of human, physical, and natural systems characteristic of the Anthropocene (Monastersky, 2015). The multidimensional nature of these complex or "wicked" problems (Conklin, 2005) makes data integration necessary for quality applied social science in the public's service.

### Fundamental Tensions of Democracy

The types of data infrastructures one encounters when researching at the juncture of public policy and public affairs have deep roots in fundamental tensions in all democracies. In democracies, citizens and elites want the government to solve their problems. Simultaneously, both desire a government that is hyper-responsive to their priorities and concerns and tend to punish parties that are too out-of-step with prevailing public attitudes (Soroka & Wlezien, 2010). This striking pattern holds across diverse governing systems. Sometimes what citizens want would solve their problems, but most of the time, these two demands represent tradeoffs. We can have a government that listens to us or solves our problems, but many times, not both. The demand for government to be accountable and responsive leads to data collection systems that are inwardly focused and geared toward "look-up." Citizens want to be able to monitor public agencies and know their dollars are spent efficiently. The notion of efficiency here echoes the triumph of economics in discussions of what governments do. It has deep roots in skepticism of government and public solutions to problems. Look-up systems for data collection and organization are the most common by a wide margin in the public sector.

At the same time, citizens demand that governments address their problems. This necessitates data systems that focus outwardly and are amenable to analysis of distributions and trends rather than looking up values. Data systems built for analysis have characteristics that render them difficult for citizens to use as accountability mechanisms but that foster comparison across time and space, allowing for policy learning and applied social science within and across disciplines. These systems are problem-focused rather than organizationally focused. The triumph of "efficiency" as a standard for governing systems and public policy means that data systems built for analysis are rare within the public sector and even more so at local levels of government. Data infrastructures of these sorts are usually compiled by academics, think tanks, businesses, and other entities hoping to learn from policy successes and failures, sometimes from data stored in look-up architectures or created from whole cloth.

### Accountability and "Look-Up" Systems

So, what is a data infrastructure built for "look-up?" As the name implies, these data infrastructures are crafted to allow the easy "look-up" of *values*. One can imagine an actuarial table or a checkbook. The data are collected, organized, and stored so that someone can quickly and easily find particular values for a variable, policy, phenomenon, or behavior of interest. This in itself is not bad on the whole. The public should be able to find accurate figures for all sorts of public endeavors. And certainly, the same values can be used to create a data set more amenable to analysis rather than look-up. However, the types of data organization that make look-up efficient and easy impede reorganizing it for analysis. A great example of this type of organization is multiple headers for columns. Multiple headers on the columns of data allow the user to quickly look up specific data values and recombine them in different ways with ease. However, the same headers give fits when reorganizing and reshaping data for analysis, especially when integrating public sector data with other data sets.

The problem extends to the storage of data. Look-up systems tend to deliver data in forms that are not amenable to integration with other data types and structures. This is because identifying particular values, discrepancies, or comparisons is the end of the exercise. Data is often delivered in PDF form, which is convenient for sharing values. It also fosters sharing without corrupting the base data as it is shared across platforms and readers. Behind the public PDF delivery, data is often stored in inaccessible databases, XML, or other arcane schemas. While modern programming tools, such as R or Python, allow the scraping of data from PDFs or other formats easier, it is worth noting that there are entire cottage industries of analysts devoted to expertise in mining these arcane structures (e.g., Census data or FIPS data). So, beginning from a look-up data system, learning or comparing over time and space is a circuitous process. The analyst must first procure public sector data. Then, the data must be rendered or organized in such a way as to allow analysis and only afterward integrated with other types of data to focus on a particular policy problem.

### Problem-Solving and Analytic Systems

Problem-solving exists alongside accountability as a key goal of data in the public sector. Data infrastructures built for analysis do not allow for the easy look-up of values. They have strict rules for organization that fosters comparison over time and space and deliver the data ready for processing by the analyst. The key to understanding the difference is that the analyst is not typically trying to look up a particular value but rather to understand a distribution of data across time or space. Particular data values are useful for understanding departures from trends or spatial variation but are not useful in generating general lessons from patterns across these two dimensions. The point of look-up systems is the individual trees; the point of analytical systems is the forest.

To generate actionable, systematic analyses and recommendations, the analyst must accurately describe and explain whole distributions or samples of data or trends. The task continues beyond looking up particular values as actionable analyses are only valuable in light of what the data allow us to explain or project, unknowable from specific values. The analyst is discerning *systematic* variation in tables or even data cubes to generate these insights. For instance, these data infrastructures adhere to the one column per variable rule - no nested or recombined headers on the data structures. In addition, they are delivered in formats that feed easily into statistical software and programs (e.g., plain text such as CSV). While looking up particular values is difficult, these data types are easily integrated and combined with other natural, physical, or human indicators to generate analyses on important public problems. For the analyst, the step of compilation is removed. For principles that guide analytic data collection systems, see Workman (2020).

Of course, there is nothing inherent in the process of collecting data about public problems that necessitates one, and only one, delivery system. Here again, the deep-rooted concern for accountability and the inward focus of much public sector data mitigates against analysis. Public agencies are hesitant to provide data in formats that foster analysis because it opens them to the criticism for which so much information on organizational operations is collected - accountability. While this is certainly a democratic good, it impedes learning in the public sector, particularly given the evolving nature of governance and policy problems.

### Complex Systems and Applied Social Research

The limitations of look-up data infrastructures become more severe when viewed from the perspective of modern policy problems. Increasingly, the problems modern governments are called on to solve are complex and multidimensional - characterized by multiple substantive dimensions for decision-making. Actionable analyses on climate change, for instance, must consider decisions about energy, food systems, transportation, and the environment, just to name a few. Look-up systems, focused internally and organized for accountability, are limited in utility for answering cross-sector, cross-domain questions at the confluence of disciplines.

In the age of the Anthropocene, human systems interlock and profoundly shape natural and physical systems. When substantive dimensions of problems are layered on top of different disciplinary expertise, standards, and norms, the problems around systems built solely for look-up compound and amplify in interdisciplinary research aiming to generate actionable, applied social science. Within the academy, land-grant universities share a large part of this burden, imbued with originating missions that mandate working on problems of public importance. Imagine the complexity of linking typical policy or public agency data to public health disparities or environmental justice indicators, and one begins to see the enormity of the problem. It also means some larger scientific debates about things like “nudges” make no sense - like nudging citizens to exercise without a sense of public investment in recreational infrastructure or parks.

### Lessons

The two types of data infrastructures emerge from very different concerns about democracy - that government should be responsive to us on the one hand and solve our problems on the other. The lessons here are simple. Data should be delivered in both ways. While this is difficult from the perspective of legislating data infrastructure, it is certainly possible that scientists - social, physical, and natural alike - adhere to these standards. The more skeptical view, however, is that efficiency is king and generating a functioning data infrastructure that allows the state to learn from itself and its peers is difficult without fundamentally altering conversations around and the framing of what governments do. The lion’s share of this burden will fall on public institutions, especially land-grant institutions, which exist to reckon with important problems in the context of their locations.

## Data Science: Subfield versus Integration—Mary K. Feeney

With the digitization of government, the availability of big data, and the open science movement, it is an exciting time to be conducting research in public administration (PA). We are getting better at drawing from, collaborating with, and contributing to traditional social science disciplines (e.g., sociology, political science). The research questions PA scholars are asking (do policies work, can data and evidence drive government decision making, how does government best achieve outcomes, how do we manage public values) are increasingly relevant to other social sciences. Students are becoming increasingly interested in data science training, scholars are working with larger datasets and machine learning techniques, and governments are grappling with the challenges and opportunities associated with big data. The popularity of data science has public administration scholars looking for more advanced quantitative training while computational mathematicians and information and computer scientists offer their analytical skills to policymakers. What does this mean for the field? Do we need a subfield of data science in public administration? Can public administration scholars train to the level of data scientists in computer and information sciences? Or are the skills and training of public administration scholars better suited to working on interdisciplinary teams with data scientists – bringing expertise about government and policy to big data projects? In this essay, I respond to some of the current challenges facing the integration of data science into public administration scholarship and practice including developing a subfield, crediting dataset creation, advancing interdisciplinary team science, and affecting practice. I begin by describing data science in public affairs. I then outline the challenges for this movement and three strategies for moving forward.

## Data Science and Public Administration

Data science is an interdisciplinary field that uses processes, algorithms, and structured and unstructured data to extract knowledge and make predictions. Advancements in machine learning, data mining and scraping, big data, informatics, computer science, mathematics, and computational science are transforming the way social scientists conduct research. PA scholarship is no exception to this change. PA scholars are learning more advanced methods, and governments around the world are releasing larger, complex datasets enabling researchers to advise them on management, policy, and service delivery. For example, in April 2022, the US White House announced the "Year of Evidence for Action" an initiative to advance evidence-based policy-making that relies heavily on data science. This and initiatives like it demonstrate a commitment to advancing evidence-based policymaking and knowledge production.

As norms around research in PA evolve, so do labels. Scientific fields continually redefine their boundaries with courses, certifications, degree conferral, and peer-review journals. Take, for example, the Behavioral Public Administration (BPA) movement. While scholars in PA and adjacent fields have long studied psychology, human behavior, and behavioral outcomes, a group of researchers – mirroring efforts in behavioral economics - sought to brand this "subfield" of PA research. They built research teams around the term BPA, organized panels and tracks at conferences, added "Behavioral Public Administration" as a keyword to papers, labeled themselves BPA researchers on Google Scholar, and eventually formed a new journal (you're reading it right now!).

In a similar vein, scholars in PA have long-tested research questions using quantitative models, data visualizations, and administrative data. Using data is not new to PA. The difference today is bigger, faster, more granular data, better algorithms, more accurate simulations, and more sophisticated techniques. With this growth in data access and use comes the institutionalization of data science through curriculum, accreditation, certification, degrees, and, maybe someday, peer-review journals. Many PA scholars are working on big data management and manipulation, converting public data into usable formats, and creating data visualizations for public use and education.

Many public affairs programs are offering courses in data science, degrees with 'data analytics' in the title, and specializations in data science or data visualization. Arizona State University offers a degree in Program Evaluation and Data Analytics. The University of South Florida offers a certificate in Data Science for Public Administration. The University of Illinois at Chicago recently renamed its Department of Public Policy, Management, and Analytics, partly to account for its Civic Analytics program. Data science is a regular topic at field conferences and is gaining traction with NASPAA.

But does big data, open administrative data, and a new focus on data construction and manipulation warrant a subfield? We don't typically define academic subfields by, for example, interview methodologies or regression analysis. Does splintering off into subfields (e.g., public management, BPA, data science) strengthen or weaken an already small research field? Does producing data for public use offer more public value than a traditional peer-review research article? How do we evaluate, credit, and reward PA researchers advancing data science? Do PA-data science scholars need a separate journal to reach their audience (e.g., practitioners, data users), or are they better served pairing data work with theory and publishing in established mainstream research journals? Should PA scholars become data scientists, or is the field better served by PA scholars working with data scientists on interdisciplinary teams? Would PA as a field and public administration as a practice be better served by teams of technical data analysts co-producing with social scientists and community members?

**Challenges and Strategies for moving forward**

There is no doubt that data play a critical role in public organizations, from policy design and strategic management to digitized services, algorithmic decision-making, and public service delivery. Scholars need good data for analysis, prediction, and application. Data science has much to offer PA scholarship, but how do we balance the need for more data and advanced analytics with rigorous theory and scholarship? Can academic researchers be experts on all fronts? Do more PA scholars need to embrace data science? Will the data science movement leave qualitative researchers and theorists behind? Does the use of big government data to predict outcomes in policy make a PA scholar a data scientist or simply a quantitative public administration scholar? The challenges being outlined at the intersection of data science and PA often revolve around credit and rewards associated with data work in a traditional academic environment. Specifically, the need to define a data science subfield, create space in peer-review journals (the currency of academics), and credit applied data work in promotion and tenure (P&T) decisions.

We can take individual and collective action to embrace the data science movement in PA and advance knowledge production and practice. In my view, this action does not mean creating a subfield, making new journals, nor further segmenting the field. Rather, I advocate for 1) entwining data science in current systems (e.g., curriculum, peer-review publications); 2) embracing new strategies for allocating credit (e.g., data repositories, pre-prints) and advancing quantitative public administration (e.g., interdisciplinary teams), and 3) reforming academic reward systems (e.g., letters of recommendation, promotion and tenure review letters and evaluations, on committees). Below I describe a course of action.

1. **Entwining Data Science in Current Systems.**

The traditional publication model is broken. We all know the problems associated with for-profit journals, voluntary peer review and editorial service, publications sitting behind paywalls, and so on. Given the current state of affairs in academic publishing, it is always surprising to me that the solution most academics present to the "where can we send this type of work" question is "make a new journal!" We hate this broken system, but we perpetuate it.

Rather than creating a new journal, I suggest working within the system as much as possible, and moving to other venues when not possible. Established peer-review journals can be leveraged to create space for data science work in PA. A good start would be getting on an editorial team and pitching a new "data science" section or feature for data releases. Special issues are another mechanism for creating space for research not normally published in a journal. Researchers can submit to journals that have research notes or data release sections. Another strategy is to write a descriptive research paper from the data and publish it in a mainstream research journal. A descriptive manuscript provides an overview of the dataset and sets the stage for the paper to be the core citation for data use.

2. **New Strategies. Alternative venues and interdisciplinary projects**

This era of big data, administrative data, and open science requires a new way of thinking about research and research outputs – specifically leveraging spaces outside the for-profit, traditional publication model that puts peer-reviewed research behind paywalls. Moving to other venues is not only an option, it's increasingly the popular option because other venues are often more open and accessible. The first and most obvious venue for data science outputs is open data portals and repositories. Posting datasets puts one's stamp on the dataset, enables more use of the data (e.g., advancement of science and knowledge), and enables others to cite and credit the researcher who collected, built, or cleaned the dataset. Second, researchers can post their data work (e.g., codebooks, descriptive papers, analytical code, practitioner reports, pre-prints) to research sites such as SSRN, ResearchGate, arXiv, and their own websites. They can also announce data releases through association newsletters and magazines. Including a cover page with a clear citation ensures the research is credited for the work, and the citations can be tracked and reported in performance and evaluation reports.

Data science, broadly, refers to an interdisciplinary field centered on mathematics and computational methods that aims to understand, describe, and predict patterns in data. These predictions are often made without a theoretical understanding of or focus on the social contexts that produce the data. The possibilities for data science applications in social science abound. As public administration scholars shift toward using larger datasets, administrative data, and more advanced predictive models, there is an opportunity to collaborate with data scientists in cross-disciplinary teams. As a program officer at the National Science Foundations, I have seen many amazing research proposals, applying unbelievably complicated data collection and analysis approaches to understand government, public funding, or public outcomes. The most quantitatively sophisticated proposals often come from researchers trained in computer science, informatics, computational mathematics, and engineering. Those same proposals, however, often fall short on theory, context, and the many social and political explanations behind the patterns found in their data. Technical teams are often looking for social science collaborators. PA scholars have a huge opportunity to bring expertise on how and why governments work (or don't) to computational data science proposals. As governments around the world open their administrative data systems for input from researchers, it will be vital for public administration scholars to team up with data scientists to help answer the big questions facing governments and society.

### 3. Reform Systems through Action

One reason academic researchers resist using data repositories is that their professional evaluations are typically centered on peer-review publications. Providing data to the community has its rewards, but ideally, these rewards would include professional recognition and advancement. We all know the limitations and problems with the current academic publication model, so rather than attach one more research output (a data set) to that model, let's evaluate intellectual and practical contributions in more than one way.

I would suggest that rather than taking the approach of "Academic publications matter most for tenure, so let's publish data production in academic journals," it would make more sense to say, "Let's evaluate academic productivity and contribution on something more than peer-reviewed articles." Just as patents are included in the promotion and tenure evaluation of engineers, producing publicly available datasets that enhance the research and practice community should be included in the promotion and tenure evaluation of PA scholars. These datasets can potentially advance practice, research, training, and academic productivity. A person can easily track the use and citation of a dataset and report it as an academic and practical output of their work. For this shift in evaluation and credit to work, PA scholars centering their work on data science need to articulate and demonstrate the value of their contribution, letter writers need to recognize and highlight that value, and P&T Committees, department heads, and deans need to recognize and reward the value produced.

A shift in how we evaluate the contribution of datasets and data science to the production of knowledge in our field does not require a formal change in P&T policies, but a change in the mindset of those of us producing research and using and evaluating the work and contributions of others. We should (1) track and report the contributions that data science makes to the field, (2) include evaluation of these contributions in review letters, (3) recognize and reward the hard work behind data production and sharing in the field, (4) clearly cite and attribute credit to those who produce and share valuable data, and (5) push leadership in our institutions, associations, and departments to expand their views on how we measure and value research outputs and outcomes. Essentially, we need to *make, track, report, and reward it.*

## Conclusions

We have more and more data to test research questions, advance theory, and analyze evidence. But the promise of knowledge advancement with more data is only as good as our theory, research designs, critical thinking, acknowledgment of what is missing, and acceptance of the limitations of any social science seeking to understand, describe, and predict human behavior. Bigger data with more observations and complexity cannot solve the problems of poorly trained researchers, unethical scholars, and fickle human beings. And while large-scale data and advanced computational methods can enable the analysis of more observations, it generally cannot answer questions related to nuance, perceptions, and human experience. Qualitative research remains vital to understanding and assessing data production, quality, sources, and procedures; analyzing trust and decision-making among researchers and research participants; and interpreting results and producing theory (Grigoropoulou & Small, 2022).

The data science movement is here to stay. The challenge remains to create space to support and value this work within public administration and policy – integrating data science as a research approach rather than a disconnected subfield. We should leverage data science and advanced methods through interdisciplinary teams, while maintaining our focus on theory and knowledge about public organizations and public service delivery. Big data and advanced modeling alone cannot make a subfield, rather, we should integrate these data advancements and techniques to further strengthen our field. This integration would be best achieved through collaborative teams, drawing on the expertise of data scientists, computational mathematicians, engineers, and information and computer scientists to answer key public administration questions and exposing students to the development of data skills and techniques within the broader context of research in public administration and policy. The first steps are working across disciplines; recognizing the theoretical limitations of data science as a subfield; acknowledging and valuing our field for its strengths and interdisciplinary knowledge of institutions, actors, and governance; and embracing the opportunities data science can bring to advancing public administration knowledge, theory, and practice.

## Thinking Beyond "Big Data" Applications of Computational Methods in Public Administration—Spiro Maroulis

The explosion of data availability and computational power in the last two decades has led to increasing use of computational approaches for modeling and analyzing social problems and phenomena, an area of study often referred to as "computational social science." These computational approaches include machine learning algorithms for fitting models to "big data," and modeling tools for simulating the behavior of social systems over time. One reaction to the emergence of such techniques is to argue, as the Introduction to this Roundtable does, that importing technical tools is not enough, and perhaps even folly, without domain-specific knowledge. A specialized subfield focused on the data and data processing tools used in public administration is required to effectively apply the new ideas and tools to public organizations.

The advancement of a specialized Administrative Informatics subfield presumes that there is high value in fostering the integration of technical skills and public administration knowledge. Undoubtedly, there are many applications of modern computational methods for which this is the case. However, the degree to which technical and domain-specific knowledge must be intertwined depends highly on the specific end-use. Consequently, in this essay, I examine the types of knowledge needed for various applications of computational methods in public administration, ranging from data classification tasks requiring little domain-specific expertise to theory development using computational models of causal explanation. I argue that regardless of whether a specialized subfield materializes, to fully realize the potential of computational social science in our field, we must avoid the temptation of taking a too data-centric view of how modern computational methods can contribute to public administration. We can make more unique contributions by focusing on hybrid end-uses of computational methods, such as developing simulation-based learning environments to understand and contextualize the output of predictive models, and using computational models as platforms for behavioral experiments.

### Low-Stakes Classification and High-Stakes Prediction

There are a surprising number of cases in public administration where very little domain-specific knowledge is required to reap the benefits of computational methods. This is true for end-uses that share two characteristics: i) they hinge on classification or prediction, and ii) the consequences of misclassification are either relatively small or easily detected. For example, the Alberta Environment and Parks system enabled more rapid implementation of operational improvements by using text analytics to process the high volume of visitor feedback received from various sources (Robinson, 2017). This example meets both criteria: The organizational benefit relies on having a system that quickly classifies visitor comments. Additionally, errors in the classification, like if the model incorrectly labeled a tweet as pertaining to a restroom issue when it was really about a rest area, can be readily detected and most likely do not drown out the primary signal. In low-stakes classification cases, as well as applications such as topic modeling or anomaly detection where the model groups or characterizes the data based on data similarities or linguistic rules, an atheoretical algorithm will do just fine.

The situation becomes more fraught for end-uses that, while still depending on prediction, are accompanied by a larger price when the prediction is wrong. When computational models are used to provide risk assessments for high-stakes decisions, tight integration of technical and domain-specific knowledge is required to develop and implement the decision-support tool in a beneficial and equitable manner. A good example is the use of predictive models in the criminal justice system to inform bail, sentencing, or parole decisions. Technical knowledge alone is not enough to create an accurate and fair model predicting the likelihood of a defendant's future misbehavior. In particular, domain-specific knowledge is extremely important in selecting outcome variables that minimize the risk of reproducing historical bias in the model's training data. Conversely, after model development, domain-specific knowledge alone is not enough to create decision-making guidelines that help a judge interpret output from an opaque process. Contextual knowledge must be tightly coupled with a solid understanding of the algorithm.

## Causal Description and Causal Explanation

While the advent of modern machine learning techniques has brought welcome attention to prediction issues, many questions in public administration and policy instead ask about the impact particular actions have on organizational or societal outcomes. Here it is important to distinguish between "causal" models that estimate the effect of one particular factor or action on an outcome, and models that capture the mechanisms that underlie the effect. Shadish et al. (2002, p. 9) labeled the former "causal description" and the latter "causal explanation."

A growing area of research involves using machine learning techniques for causal description. Most promising for public administration and policy applications are methods to identify, estimate, and draw statistical inference about heterogeneous treatment effects of a policy or program. For example, in a multi-site trial of a "growth mindset" intervention helping students view their intelligence as malleable rather than predetermined, a novel machine learning method for causal description confirmed that the intervention impact was greater at schools with supporting peer norms, and lower at very high achieving schools (Yeager et al., 2019). Despite the potential of these models to provide information that can better target programs, their use in public administration, and indeed most areas of social science, has yet to take hold.

Computational models focused on causal explanation of public sector phenomena are more common but still limited. Examples include using computational models to refine theory about the connection between individual motivation and collaborative governance outcomes (Choi & Robertson, 2019), micro-level processes underlying public organization adaptation to extreme weather risks (Zhang & Maroulis, 2021), and the interaction between social and technical factors in the implementation of frontline innovation (Maroulis & Wilensky, 2015). Computational models capturing the structure of complex social systems are also occasionally used as the focal points of intentionally designed learning environments. An excellent example of this approach is the Rethink Health model of a regional health ecosystem used in community workshops and classrooms to illustrate how different types of local healthcare investments translate into varying levels of care, access, and cost (Homer et al., 2020).

Models of causal explanation require deeper domain expertise than models of causal description. For example, consider whether a governance innovation, such as providing charters to non-public organizations to run schools, improves academic outcomes for students. Developing a computational model of causal description requires having enough domain-specific knowledge to incorporate variables that account for salient preexisting differences between students who attend charter schools and those that do not. Developing a computational model of causal explanation also requires articulating how student differences interact with the strategies, practices, and resources used by the schools, how the teachers and students interact with each other, and how those processes unfold over time (Maroulis et al., 2010; Maroulis, 2016).

## Moving Forward: Simulation-Based Behavioral Experiments and Learning Environments

The initial impulse with new technology is to use it to help us do what we are already doing better and more efficiently. For example, looking for patterns and imposing some structure on loosely structured text data -- such as data from organizational reports or public websites – is a research activity familiar to many public administration scholars. Access to machine learning tools has made improving and scaling such efforts much easier, making text analysis a popular early application of machine learning in public administration. However, to realize the full potential of computational social science tools and ideas, we must also think creatively about opportunities computational advances provide for new, and not just existing, applications.

The most promising application areas lie at the intersection of causal explanation, causal description, and prediction. One specific opportunity is creating simulation-based learning environments to help students and decision-makers better understand and contextualize high-stakes predictive models used in the public sector. Judges, public health officials, and other public servants using these models are called upon to combine probabilistic assessments of an event occurring, or risk scores, with outside information when making a decision. To use risk scores effectively and fairly, decision-makers must be provided with information to understand their origin. Simulation-based learning environments, where participants can control and observe the data-generating process and then build models that predict related outcomes, are particularly well-suited to the challenge of preparing students to manage the implementation of predictive decision-support tools. They can also prepare students to conduct research on the interpretability of models used in public organizations or develop new models. Versions of such learning environments tailored to real-world decision-makers would have an even greater and more immediate impact.

A second opportunity is to use computational models focused on causal explanation as platforms for conducting behavioral experiments. Experimental methods are becoming a core component of the public administration scholar’s toolkit for causal description. Computational models can enhance experimental research by enabling researchers to combine controlled experimentation with hard-to-capture process-oriented data, accelerating the integration of causal description and causal explanation. In cultural evolution, for example, simulations are used in behavioral experiments investigating the mechanisms related to the intergenerational selection and adaptation of norms and knowledge (Thompson et al., 2020). In a more management-related context, my colleagues and I conducted simulation-based behavioral experiments examining the performance of networked groups attempting to solve a problem that required aggregating information from diverse sources (Maroulis et al., 2020). The computational model allowed us to capture the actions of each individual and the flow of all communication during each time step, in addition to observing the collective success of 20-person teams under randomly assigned experimental conditions. Consequently, we were able to estimate the effect of communication network characteristics on collective problem-solving (causal description), and also understand how the behaviors and interactions of the individuals over time contributed to the end result (causal explanation).

Developing a computational model for use with behavioral experiments also provides benefits beyond process-oriented, researcher-controlled data generation. In the network experiment above, running the model with hypothetical agent behaviors allowed us to make refined predictions about group outcomes before executing the experiment. After all the data had been collected, calibrating the model with the observed behavior of the participants and running additional computational experiments facilitated a theoretically generative exploratory analysis.

Regardless of whether future efforts to capitalize on the affordances of modern computational methods take the form of an Administrative Informatics subfield, those efforts will benefit from not taking a too data-centric view of the contribution computational methods provide. Data need models; models are chosen depending on one’s end-use; and end-uses vary in the mix of technical and domain-specific knowledge required. End-uses that integrate causal explanation with prediction or causal description provide the strongest case for a specialized subfield, as well as the greatest promise for realizing the full potential of computational social science tools and ideas in public administration.

## Understanding Problems of Place Through Data: Spatial Causal Inference in Public Administration Research—Stephen Kleinschmit

Historically siloed within the disciplinary confines of geography, spatial analysis has received little focus within public administration research, despite widespread adoption in adjacent academic fields and professional practice. Consequently, it is little surprise that the field's growing emphasis on causal inference omits an important emerging development within this area: Spatial causal inference. To date, the field's tepid adoption of spatial methods limits its ability to conceptualize and model complex realities. By favoring the use of context-agnostic social constructions that fail to acknowledge problems of place, public administration has a systemic problem with understanding how the physical world influences behavior. This essay argues that for more realistic assessments of public problems, public administration's methodological evolution will be strengthened by incorporating spatial considerations into its understanding of causal inference.

Nearly every public problem has a deterministic spatial component, yet such considerations are curiously absent in public administration research, as "few scholars of public administration have examined spatial interdependence; most of the discipline has yet to explicitly test theories of spatial relationships" (Cook et al., 2018, p. 594). Aside from policy diffusion studies, the field's limited consideration has yielded few articles detailing these relationships or using spatial methods. Consequently, the field also largely ignores an existing ecosystem of spatial data, tools, and research methodologies. Without addressing these structural limitations, public administration will continue to provide suboptimal explanations of public behavior and suboptimal solutions to their problems. Broadly, wider adoption of spatial methods would drive important shifts in how practitioners understand and address public problems through data. As part of a robust administrative informatics framework, spatial causal inference could prove to be a transformational concept not only for data modeling and visualization but also to the field's conceptual understanding of social interfaces with the physical world.

## Public Administration's Capacity for Addressing Spatial Questions

Public administration research is awash with long-established theories anchored in underacknowledged (often unacknowledged) spatial relationships. Our understanding of public goods, institutional design, and collective choice are rooted in spatial determinants. Program evaluators study the efficiency, effectiveness, and equitability of public service delivery, yet rarely discern the effects of distance, proximity, or scale underlying their magnitude. The field has yet to embrace an understanding of real-world phenomena beyond aggregation into defined regions, a structural limitation to research validity. Public problems, like environmental features, rarely conform to abstract political boundaries, creating imprecision in estimating their effects. As the computational and data availability limitations that once precluded more granular units of analysis have receded, the imperative grows for a shift in the field's methods to meet the new capacities for precision afforded by modern data structures.

One can speculate about the reasons for the lack of spatial emphasis within field research. Public services, delivered under a market-failures framework, lack traditional price mechanisms that drive spatial attention within private enterprises. The clustering and proximity considerations underlying the efficiency of business operations are commonplace but substantially less prevalent in public-sector research. An orientation towards physical geography might have driven social science researchers to consider geographic information systems (GIS) a poor fit for social research. Perhaps the hegemony of inferential statistics hampered public administration's methodological evolution, with few formally trained in spatial methods to contribute to field development. With the small number of academic programs offering courses in GIS, let alone the far smaller number that incorporate them into their core curricula, the field's future capacity for spatial research remains dismal given prevailing pedagogical trends.

Public health and urban affairs have long understood the benefits of spatial analysis; urbanization drove a need for an improved understanding of the spatial dependencies underlying public problems. Affordable housing, environmental quality, sanitation, traffic congestion, crime, economic development, residential segregation, and poverty are just a few of the problems whose solutions require an understanding of the physical world. The clustering effects of these phenomena have direct consequences for the public, yet are rarely considered in a systematic way within public administration research. To date, we have not seen a movement to provide explanatory frames underpinned by social-physical interfaces, though "spatial analysis could also be used to grapple theoretically with what space means in an administrative or policy context" through the examination of "geographic proximity, structural similarity, contextual similarity, or many other possibilities" (Zhu et al., 2019). As a field predicated on policy implementation, public administration should show a much greater interest in understanding spatial relationships – aspects that heavily underpin the effective and efficient delivery of public services.

### Spatial Causal Inference

With public data science and causal inference growing as areas of inquiry, a promising movement is emerging that incorporates elements of both domains. Spatial causal inference has arisen as a new area of emphasis within ecology, economics, and most prominently within the field of epidemiology (Akbari et al., 2021). Consequently, adopting spatial causal inference could offer revolutionary advances in knowledge of public problems and administrative behavior, and could also help define a distinct methodological framework that differentiates administrative informatics from the generalized principles of computational social science.

A precondition to adopting spatial causal inference within the field of behavioral public administration is the development of spatial reasoning capacity - the ability to conceptualize public problems in multi-dimensional space. Field research often considers the concepts of clustering and dispersion of observations within its statistical models but not within our conceptual framing of behavior occurring in physical space. In a historical sense, spatial relationships are one of the most deterministic dimensions underlying human behavior; the expenditure of time and resources drive decisions about where we live, how we travel through physical space, and countless other fundamental facets of human behavior. It is difficult to understate the importance of spatial relationships as a predictor of human behavior, making their absence from public administration research much more concerning.

Spatial causal inference addresses the limitations of prevailing causal inference research. Traditional experimental design is often impossible due to the clustering and non-random occurrence of its phenomena. Akabari et al. (2021) note several methodological challenges for traditional causal inference in spatial processes, including spatial spillovers, spatial heterogeneity, modifiable area unit problems (ecological fallacies), selection biases, confounding biases, omitted variables, and reverse causal relationships. They further note, "Current causal inference approaches attempt to port methods from nonspatial processes to the spatial domain, and do not systematically manage spatial effects" (p.24). Consequently, administrative informatics must be purposive about the construction of its methods and avoid the flawed imposition of spatially agnostic research methodologies onto questions that evidence deterministic spatial dependencies.

### The Limitations of Spatially Agnostic Research Frames

As a sensemaking framework, social sciences are primarily concerned with logics underlying human behavior. As a social science, public administration research has primarily situated itself to examine questions of administrative behavior and the implementation of public policy. What it has failed to do thus far is systematically address the limitations of its spatially agnostic research frames, particularly when human behavior is heavily rooted in spatial dependencies. As Waldo Tobler (1970) noted in what later became known as the First Law of Geography, "Everything is related to everything else, but near things are more related than distant things." This simple yet profound idea led to several important conceptual developments in geography and beyond, from challenging basic assumptions underlying human behavior to criticizing prevailing research methodologies for failing to address the impossibility of homogeneous regions existing in the real world. Public administration theory, thus far, has not endured such a reckoning.

Future research is based on conceptual framing anchored in the path dependency of existing research and methods, creating a form of institutional myopia that artificially bounds its inquiry. Reliance on spatially agnostic social constructs as explanatory mechanisms neglect spatially confounding environmental variables that may prove more explanatory than the social constructs selected by the researcher. Can disparities in K-12 student educational performance best be explained by levels of fiscal appropriation, exposure to violent crime, the socioeconomic status of parents, or cognitive development issues resulting from the presence of lead water service lines in older housing stock? These elements are critical and may be the primary determinants in some contexts but not others. Often spatial variables are completely omitted from consideration due to the bounded conceptual frame of the researcher, likely a continuation of the bounded conceptual frames and historical narratives introduced through their education and practice. Another important consideration is that missing confounding variables may have prominent spatial patterns, and methods must mitigate the bias caused by their omission (Reich et al., 2020). Additionally, counterfactual spatial distributions, modifiable areal unit problems, spatiotemporal events, and hierarchical causal relationships further the complexity of modeling real-world effects. Without an understanding of spatially confounding variables, it is possible that a substantial amount of research in the field is imposing suboptimal causal explanations onto public problems, logically hampering the design and implementation of policy interventions.

### Information Technology Provides a Catalyst for Spatial Research

Administrative informatics, as an emerging field of inquiry, provides the conceptual home for operationalizing spatial causal inference through the principles of data science. The rise of a public data paradigm invites consideration of a future of practice awash in spatial data. The collection of spatial data through IT platforms and sensory technologies presents new opportunities for advanced analytical procedures not previously possible - spatial big data. Mobile spatial data provides new capacities for societal impact assessment, while spatial-temporal analysis extends analysis by adding an extra dimension, allowing an improved understanding of recurrent problems in physical space. For example, adopting methodologies utilized in predictive policing could enable a range of new possibilities for effective public service delivery in other domains.

While public administration's interest in network analysis is largely grounded in the analysis of organizational behavior, network applications grounded in spatial topology could prove to be important for generating new theories on route strategy and displacement effects. Sensor-driven data streams provide opportunities for "certainty through saturation," modeling becomes less probabilistic and more deterministic as confidence intervals become negligible and confidence levels approach 100% (Overton et al., 2022). Such capacities might be relatively new within the confines of social science research, but in the physical world, it is not difficult to identify a well-worn path. In the future, spatially linked administrative data will allow us to reliably "ground truth" field research with similar precision.

### Conclusion

Spatial causal inference, like administrative informatics, is a logical next step in the evolution of public administration research. As a hybrid field informed by intellectual traditions of many disciplines, its intellectual locus is constantly evolving. Existing scholarship has identified field deficiencies in public administration's theoretical and methodological toolboxes, evidencing its discomfort for spatial considerations; this essay extends this conversation into the emergent domain of spatial causal inference. Public administration inquiry is bounded by a body of existing research that largely reduces the dimensionality of its real-world phenomena into sets of behavioral relationships defined by often arbitrary political regions, rather than offering comprehensive causal models of the social-physical relationships underpinning reality. From a technical perspective, computational capacity and data availability are no longer limitations for public administration. Whether the field can overcome the inertia of its existing institutional and theoretical limitations remains to be seen.

## Building the administrative informatics field: Suggestions for a data infrastructure framework—Federica Fusi

Across all disciplines, we observe the emergence of new sub-fields, such as urban informatics and bioinformatics, which leverage data and digital technologies to support decision-making and innovatively solve public problems. Administrative informatics brings this approach into public administration (PA) scholarship by moving forward two new streams of research. First, it invites PA scholars to examine how public organizations and managers use data and how their use and analysis affect decision-making and government outcomes. Second, it draws from informatics principles, such as agile development and design science to rethink knowledge production. This essay speaks to this second opportunity, particularly the role of data infrastructures, to reflect on how PA scholars approach data and which data practices we want to design, promote, and maintain.

Data infrastructures refer to a vast range of digital technologies to store, share, analyze, and visualize data. Some data infrastructures offer a shared online space to store and share code scripts and data (e.g., Dataverse). Others facilitate code reproducibility (e.g., GitHub) or dashboard development (e.g., Shiny). Some others focus on data sharing and visualization (e.g., open data portals, GIS mapping tools). Overall, data infrastructures promote research replicability and reproducibility and foster new knowledge production by facilitating access to pools of data and capacity to analyze them. They are fundamental to reorganizing knowledge production towards greater task specialization and segmentation. By building infrastructures that allow researchers to share data and related outcomes, such as code scripts, maps, and dashboards, we can attribute scholarly value to independent activities such as data collection, code development, data analysis and visualization, and software design.

However, other informatics fields highlight the importance of building a framework to consciously guide the design, development, and implementation of data infrastructures. Data infrastructures are not neutral artifacts, and research and society outcomes largely depend on how they are designed and used. Informatics fields often prioritize technical issues and researchers' needs, failing to recognize and engage other data constituencies – e.g., individuals and communities, practitioners, and vulnerable populations. This aspect is critical for the PA research community, which aims to integrate equity and social responsibility with traditional academic values. Without a data infrastructure framework, we risk falling short in our goal to both build a solid academic scholarship and produce knowledge that advances communities' priorities. At the core of an administrative informatics field, we need a framework that simultaneously: (1) accounts for the socio-technical nature of data infrastructures, and (2) does not exclude or ignore but rather recognizes and advocates for the multiplicity of interests and rights surrounding data.

The first issue acknowledges the social practices, norms, and rules regulating data access, use, and sharing. In academia, data represent a "labor of love" which implicitly confers rights to researchers to decide with whom, where, and how to share them. Studies consistently show that researchers are reluctant to change their data practices towards more open and community-oriented approaches (Fusi et al, 2018). Reasons vary, from concerns of being scooped or exposed for mistakes, to practical matters, such as forgetting the dataset's location, missing or malfunctioning coding scripts, and lack of time and resources (Devriendt et al., 2021).

Data infrastructures aim to lower barriers to sharing and accessing data and, therefore, encourage greater data use and re-use thanks to innovative technical solutions. However, they are often developed as standalone technological artifacts removed from the communities that are supposed to use them. There is a recursive relationship between contribution to and usage of data infrastructures, where use encourages additional contributions and increased contributions further use (Kane & Ransbotham, 2016). Data infrastructures fail if they do not initiate and sustain this relationship by attracting a community of both contributors and users (Fusi et al., 2018). This failure is often due to a misalignment between data infrastructures and the social norms and incentives of the community. Devriendt and colleagues (2021) find that researchers are less willing to contribute data if data platforms do not offer collaboration opportunities as "only sharing data is not considered to be intellectually stimulating" (p. 7).

Accounting for the socio-technical nature of data infrastructures is an important exercise to ensure that sufficient data inputs are provided for research and practice while avoiding large "data dumps" that do not produce meaningful outcomes. It means acknowledging that data users and contributors will actively design and negotiate social norms around data infrastructures. Hence, data infrastructure development cannot only focus on technical solutions, but it requires ongoing discussions around data practices and goals of the interested community. Technical features of data infrastructures should be flexible enough to accommodate diverse uses, practices, and goals. Results from Devriendt and colleague's study, for instance, suggest that to attract a research community, data infrastructures should focus on features that promote collaboration along with data sharing[1].

The second issue concerns the multiplicity of interests and rights surrounding data and whose values and social norms should guide the design of data infrastructures. PA scholars strive to engage public managers, policymakers, and communities as part of our research. For administrative informatics to do the same, we need to discuss what data count and what data mean for them.

Accessing data is a way for communities, groups, and organizations to build evidence to support their position, break the monopolies that historically characterize certain policy domains, and influence policymaking. Similarly, providing data is a way for underserved populations to have their voices heard. Community data offer local knowledge and insights on policy issues and elevate the concerns of local residents beyond the mainstream narrative of government officials and researchers (Ottinger, 2013). Yet, practices surrounding community data

[1] For instance, by designing infrastructures that support small group interactions or by limiting the use of data infrastructures to specialized communities and data.

use have undergone great scrutiny because of the low level of trust between communities and other data users, including researchers. Communities reluctantly share data with researchers, who have historically taken from them without producing valuable knowledge in return (Vera et al., 2019). Researchers are also distrustful of data collected and produced by communities (Ottinger, 2013; Vera et al., 2019), perceiving them as less reliable than other types of data. Government data are often preferred despite community warnings on the incomplete and biased nature of such data, especially when self-reported by interested parties (e.g., industries). New technologies, such as social media and remote sensors, allow researchers to extract data from a distance, further removing the need to connect with the community and potentially exacerbating scholar-community gaps (Vera et al., 2019).

Some initiatives are emerging to change data practices and build new trust among data stakeholders. The CARE principles (Collective benefit, Authority to control, Responsibility, and Ethics) exemplify a different approach that primarily highlights the power differentials and social context in which data are embedded. Designed to protect and recognize Indigenous rights to self-determination, these principles encourage "open and other data movements to consider both people and purpose in their advocacy and pursuits" (*CARE Principles of Indigenous Data Governance*, 2022). Similarly, the Environmental Data & Governance Initiative (EDGI) is a distributed, consensus-based organization that seeks to create new data practices around environmental data to promote fairness and justice (Vera et al., 2019). The EDGI recognizes the need to quantify and measure environmental damage but it also challenges the supremacy of government data over communities' voices and narratives.

These initiatives question the paradigm that data producers – e.g., researchers or government agencies – are the only ones holding data rights and reject the idea that data can be collected "for free, without relations or responsibilities" (Vera et al., 2019, p. 7). Current data practices should be re-negotiated to account for the historical power dynamics that underlie data. This process might lead to new data practices, such as ongoing consensus procedures with research participants to share or even delete data (Ottinger, 2013) or community-led projects to collect and analyze data. In addition, the design and development of data infrastructures should reflect the perspectives of a diverse group of stakeholders. Dillon and colleagues (2017) suggest "networked open-source data infrastructure that can be modified, adapted, and supported by local communities" (p. 186) to concretely enact the notions of shared data ownership and community participation.

These two overarching issues raise several questions that PA scholars should carefully examine moving forward: Whose data and data practices are central to administrative informatics? If we aim to foster social responsibility and equity, how do we create data infrastructures that attract active contributors and users from the PA community at large (e.g., scholars but also local communities)? How do we identify the uses, practices, and goals of these different communities to design data infrastructures? How do we integrate new data practices into our research, and how do our research practices fit with current data infrastructures? Who holds data rights, and what counts as data ownership? How do we design governance models for data infrastructures in line with the values of multiple communities? As briefly shown in this essay, some questions are not new; research and advocacy work on data and data infrastructures in other fields offer a solid starting point for further discussion and investigation. Similarly, public administration scholarship on social equity, open government data, behavioral public administration, among others, offers theoretical frameworks to investigate and contextualize these emerging issues. These two research streams have rarely been integrated, and PA has yet to investigate how macro-level trends related to data sharing and use influence behaviors and cognitions of the involved stakeholders, including researchers, public managers, and local communities.

This call offers excellent opportunities to theoretically address these questions and empirically apply some of these concepts. For instance, articles benchmarking or profiling datasets should develop new evaluation criteria that account for researchers' bias towards community-produced data. Interviews and focus groups with local communities could offer the opportunity to develop new metrics and indicators as well as design participatory processes to involve community members in data infrastructure development. Commentaries to data curation should offer insights into best practices for data documentation and user agreements, which are fundamental to designing social norms that foster a positive relationship between data contribution and use. Scholars could also design experiments to understand how communities, researchers, and public officials use data platforms and how different social norms affect the use and contribution of data. It is critical, however, to integrate these individual contributions into a collective data infrastructure framework (e.g., principles) to guide the shift towards a more data-intensive PA field in line with our community's goals.

## Curating Novel Data for the Research Commons: A Data Science Approach—Eric Stokan

The growth of social media and digital trace data have opened new avenues for scientific inquiry. In the fields of public administration and public policy scholarship increasingly relies on data sources and emergent analytic techniques to answer novel questions. Despite the benefits to scholarly activity, there are "hidden steps" and decision points in these data that are glossed over, relegated to a footnote, or nonexistent. Each new project using data drawn from a social media platform requires assumptions and necessitates key decisions with little insight from the field as to best practices. I highlight many key decision points and processing steps required for using large-scale social media data (e.g., Twitter). I draw from my experience in an ongoing collaborative research effort with Dr. Ian Anson and Dr. Nate Jensen. In this project, we utilize Twitter data to understand attitudes regarding the second headquarters of Amazon (Amazon HQ2.0). As I identify these hidden steps, I make recommendations for points of collaboration in the hopes of developing a data infrastructure commons for the scholarly community. The costs of redundant processes for our field are steep, as are the learning costs of the tools and technologies we utilize. Thus, I advocate for collective engagement among scholars in the capture, curation, processing, and dissemination phases of projects relying on social media data.

### Twitter Data for Public Sentiment

My colleagues and I are using Twitter data to understand the perceptions of different actors in the policymaking process on the Amazon HQ2.0 decision. With few surveys measuring public perception, we know little about perceptions regarding Amazon HQ2.0. We expect that these perceptions vary across place, over time, and by policymaking process status (e.g., politician, media, development actor, public, etc.). We aim to understand whether public perceptions impacted the decision to make incentive packages transparent and affected the amount offered.

Many have used Twitter and other social media sources to discern public sentiment on topics ranging from perceptions of the Affordable Care Act to understanding the spread of Zoonotic diseases. However, an important note on the use of Twitter data is warranted. During the development of this article, the fate of Twitter has become less certain. Elon Musk has become CEO, and the company has shed employees, it currently faces an uncertain financial future, and concerns over fake accounts are growing. Those with developer accounts should consider using Twitter data before this period (October 27, 2022) and may wish to collect data quickly if the company ceases to offer these data going forward for researchers. Nonetheless, the general ideas advanced in this article extend to other social media data sources. Each platform maintains its own standards for making data available.

### Pulling Twitter Data

For Twitter, an easy-to-use API is currently available with a developer account. This account provides information on the user (e.g., user description, # of followers, # of follows, # of Tweets of the user) and the Tweet itself (e.g., text of the Tweet, hashtags in the Tweet, URLs in the Tweet). These two data sources require separate calls to the API- meaning each requires distinct programming code (e.g., in R or Python) and different rate limits (how many elements you can collect in a given time without violating Twitter terms). The *Rtweets* package in the R programming language can help ease the collection process. The two sources can be merged via the user ID for a fuller picture of which users are discussing what topics.

**Recommendation: The scholarly community should share Python and R code for various social media platforms to collect these data. Care should be taken to remove personal account information- developer keys.**

### Filtering Users

Twitter has not identified the extent of the users that are bots as opposed to humans, which is an increasing concern for researchers. Thus, to ensure useable content, the first step often involves removing users that are bots or outside the population of interest (e.g., users from outside the geographic area of focus). For small data sets, this may necessitate simply removing a few users. As projects scale into more than a thousand observations, the concerns increase, and the demands on one's time and resources become a challenge. Fortunately, programs have been written in the R programming language that estimate the probability that users are bots (aptly named *BotOrNot2* by Matthew Kearney). Algorithms, such as BotOrNot, will produce both false positives and false negatives. The consideration left to the researcher is at what probability the researcher should exclude the user or inspect the user more closely. As a discipline, we might consider standards for determining how best to handle these cases.

### Classifying Users

Classifying the type of user is an essential task when the unit of analysis is the individual. Researchers may be interested in understanding the user's demographic characteristics (e.g., race, ethnicity, gender, age), employment, partisanship, socio-economic status, position in the policymaking process, and more. Some classifications are easier to identify than others.

As it relates to automating the process with machine learning algorithms, care must be taken in managing these data, and critical assumptions should be documented. With Twitter data, the "user description" field often provides information about the *user's self-identity, workplace, and professional positions*. My account, for example, states my professional position and workplace. I list my topical research areas, and I also identify as a runner, coffee fan (thought people should know), urbanist, and chess enthusiast. My profile is consistent with many others in the field, however, not all will choose to identify employment, profession, or any other aspects of themselves. Some will use satirical references in their self-descriptions or in other data fields. This can present a challenge for understanding by humans and machines without the relevant context.

Researchers have been able to discern *ideology and partisanship* through information in the user descriptions, relevant Tweets, and even through their networks by tracking the user's connection to other users. Computer visioning can even assist researchers in determining the ideology of legislators, based upon symbols they display in their social media pictures.

Through these same computer vision techniques, one's profile image may provide useful, though far from perfect, information about the *user's gender and race*. The predicted probabilities and confidence in the results might also be improved using the user's name field and an R package like *predictrace*.

For these classifications, each strategy may introduce some bias; however, machine learning algorithms can be combined through ensembling methods, which improves the overall accuracy of prediction tasks. Researchers should make clear the assumptions built in at each stage when discerning identity and determine the best ordering and methods to increase this certitude.

**Recommendations 1: The scholarly community should share best practices and benchmark different strategies against each other to improve the process of identifying and classifying users for research in public administration and policy.**

**Recommendation 2: Some users would be of particular interest to researchers (e.g., all state or local legislators, governors or mayors, various news outlets and commentators, etc.). Sharing lists of known usernames can reduce redundancy in the classification step and improve productivity and scholarship in the discipline.**

### Coding Sentiment/Stance

Often, researchers seek to understand citizens' perceptions of key political actors, their governments, and the adoption and implementation of policies and programs within their communities. As a result, discerning the sentiment of social media text may be critical in understanding their perspectives. Existing algorithms, such as BERT-based models trained with Wikipedia data and VADER (capable of classifying emojis and images as positive/negative), offer easy-to-use algorithms to make such assessments. These algorithms are pre-trained on large datasets using machine learning models allowing researchers to predict whether a Tweet is positive, negative, or neutral. However, topics of study using unique jargon may require the user to build their own models. In such cases, researchers should develop and communicate clear codification schemes with conventional standards of interrater reliability. For Twitter data, the researcher may need to codify tens or hundreds of thousands of Tweets- a task that even large research teams will find unmanageable. Suggestions for scaling up to this level are found below.

Stance is a closely related topic to sentiment. The distinction is often important with social media data. Understanding public perceptions of the Amazon HQ2.0 deal is illustrative. A Tweet might state, "Representative Alexandria Ocasio-Cortez is wrong in her thinking that Amazon HQ2 is bad for NYC, because it's a job creator." In this case, most algorithms would score this as negative due to terms such as "wrong" and "bad." The stance toward the HQ2.0 opportunity, however, is positive. Thus, researchers should discern between these components when inferring how citizens feel about political actors, governmental agencies, policies, or other governmental decisions or effectiveness. At present, this also requires training one's own model.

**Recommendation 1: The scholarly community should consider building training models, like VADER, uniquely focused and trained on public administration, public policy, and governance topics.**

**Recommendation 2: Building and sharing, models aimed at discerning stance toward policies and government programs could improve scholarship in the field.**

### Topic Modelling

For some projects, researchers are interested in understanding broader themes of social media comments. Suppose I want to know how people are describing COVID-19. The topic COVID-19 might be focused on specific components like hospitalization, vaccination, mask-wearing perceptions, workforce concerns, or supply chain issues. The desire to understand topics from the data is an increasing focus of legal and planning scholarship. Researchers can adapt many distinct approaches to discern meaning from raw text. This includes analyzing term frequency, using bag of words approaches, clustering through Principal Components Analysis, analyzing with Latent Dirichlet Allocation, or even mapping the geometric distance between documents based on word choice. Deciding which approach is most reasonable is best driven by the research question.

### Identifying Geolocation

Surprisingly, about 70-80% of Twitter users provide information that can easily be geolocated to a city, state, and/or country. Those that fail to clearly define one geography do so for several reasons. First, many users list multiple places (e.g., DC/NYC/Paris). A user may provide a familiar reference that indicates their location, often requiring additional knowledge (e.g., city of brotherly love in the US context means Philadelphia; home of the Tar-Heels means the Chapel Hill area). For others, they simply fail to offer information or make nonsensical references. In each case, this leads to decision points on how best to handle these situations. Researchers may gain useful contextual information from other Tweets or the occasional geotag of a user's location.

**Recommendations: The scholarly community should consider building lists of known short-hand references to identify locations. Exceptions to these lists might be added as a caution field. These lists could be connected with respective FIPS codes to allow easy connection with governmental and other third-party data sources.**

### Scaling Up

One of the key strengths of social media and digital trace data is the scale of user information that can be collected. While it is an advantage for the analysis in a large-N study, it is also a potential constraint that researchers face in classifying, coding, and cleaning data.

Researchers might consider relying on services like Amazon Mechanical Turk or similar crowdsourcing marketplaces to help assist in the research process. Researchers will need to conduct careful quality control checks on data supplied by each Turker. This action may include pre-testing with a known sample of users or social media posts and then building mechanisms for quality control throughout the data cleaning and classification processes. While these processes can largely be automated, they are not trivial tasks.

**Recommendations: The scholarly community should develop collective best practices for using crowdsourcing platforms to classify and clean social media data with careful quality control with ethical standards for compensating crowd-sourced assistance (e.g., Turkers).**

**Next Steps**

There are many other strategies for utilizing new data sources that go beyond this short essay's scope. However, we should collectively work towards a greater understanding of how we use these new data sources, collaborate around cleaning and processing tasks, develop useful code in conventional programming languages, and identify best practices in the sequential processing, cleaning, classifying, and properly scaling up these efforts. These strategies and novel data sources need not replace traditional approaches but should complement and supplement them.

I have outlined an example of the data science curation process using social media data, but there are many more examples of other novel large-N data sources that require big data approaches and open new research lines of research. For a similar example of a common task of interest, such as classifying political ideology of local government officials, see Neumann, Linder, and Desmarais (2022). They use a web-based crawling algorithm and text-based analysis to discern ideology in even small governments that are challenging to reach by survey.

The learning curve for these practices is steep, as are the number of decision points made in these projects, but collective efforts could lead to an improved understanding of core public administration and policy questions. Through GitHub and other web-based platforms, we can share and collaborate on code, processes, ideas, and data. While the steps, and corresponding recommendations for collaboration, were outlined sequentially and in isolation, collaboration across phases could vastly improve our ability to answer fundamental questions in the discipline.

## How to DATAFY Public Administration: An Introductory Guide—Kayla Schwoerer

The rise of big data and data science applications in public administration (PA) offers novel opportunities to explore both the "new" and "old" questions of our field. At the same time, they present a number of new and yet unknown challenges for us as researchers as well as the public organizations, public servants, and citizens our work is intended to help.

But what is "big data" anyway, and what is meant by data science applications? Generally speaking, big data is used as an umbrella term to describe the process, analysis, and application of large amounts of data to generate data-driven insights. In the public sector, big data consists of combinations of administrative data and data collected from sensors, satellites, mobile apps, social networks, and other data sharing arrangements. Still, collecting and processing large, diverse, and complex datasets is just one piece of the puzzle. As others have provocatively argued, big data is not really about the data at all. What matters most is what sorts of actionable insights come out of the data. This is where data science comes in.

Put simply, data science, in the broadest sense, guides the process of selecting the appropriate analyses to identify patterns in and give meaning to raw data. While data science has a long and storied history, for now, it is only necessary to note that data science applications in PA are integral to making sense of available data. Additionally, there are a number of existing, but underutilized, data science tools that can prove especially useful for PA in communicating the insights that come out of big data. These tools help individuals make sense of enormous amounts of data, identify patterns, translate those patterns into actionable insights, and, most importantly, communicate those insights effectively to generate stakeholder buy-in, influence behavior, and inform decisions. In other words, to fully realize the data's potential.

Therefore, this essay aims to demonstrate that we cannot discuss the application of either big data or data science in PA without also discussing the role of communication. More specifically, the role that effective communication practices play in implementing decisions and policies made based on data-driven insights and applications. Ultimately, I argue that communication is integral to linking data not only to decision-making, policy formation, and service delivery but to *effective* decisions and policies and the achievement of public value, broadly defined.

In this way, *Administrative Informatics* must acknowledge the critical role of communication in realizing the potential of data and its various applications in the public sector.

I introduce a framework to help guide how we think about the role of effective communication in an increasingly datafied world. I argue that effective communication about data and its applications to administrative decision-making demands 1) engaging data visualizations, 2) audience-driven strategies, 3) storytelling, 4) attention to accessibility, 5) sufficient recognition of context, and 6) a "why." These six principles form the foundation for DATAFY.

### A Quick Note on Terminology

It is important to acknowledge upfront that these principles are in no way comprehensive nor are they absolute. There are a vast number of tools available to the many stakeholders tasked with collecting, curating, making sense of, and using data across the wide range of contexts in which data-driven insights can be especially useful. These are just a few of them.

Nonetheless, I believe that these are some of the most relevant and actionable tools for PA, especially as it pertains to how we use data to understand and influence behavior. Therefore, this essay directly addresses both researchers and practitioners and concerns how we (as PA researchers and practitioners) communicate in an increasingly datafied field and a rapidly datafied world.

The activities in which we are required to communicate with and about data are varied including cataloging how data was collected or has been curated, our analyses, the presentation of findings, and how findings informed specific decisions in practice. These are the types of activities I refer to in this essay as they are essential activities in the production and dissemination of knowledge. Additionally, they also demand that we practice effective communication to successfully apply that knowledge. Neglecting to do so can render efforts to leverage "the power of big data" unproductive; or, worse, pointless.

### How to DATAFY Public Administration

Below, I briefly introduce the six principles and the value I believe they offer to PA vis á vis an *Administrative Informatics* lens. There is much more to say about these principles than space will allow here so I conclude with a brief discussion of how we might explore them and their impact on our understanding of behavioral elements in PA in future research.

### Data Visualization

Visualizations that make appropriate use of colors, shapes, sizes, and other visual elements such as typography effectively capture attention and help emphasize the most important information. The research shows that effective data visualizations not only increase individuals' understanding of the information presented but also help individuals recall that information at a later date.

Data science applications offer a host of data visualization tools that can help communicate findings in compelling, engaging, and easy-to-understand ways. With the recent rise in low-code (e.g., data visualization packages in R and Python) and code-free tools (e.g., Tableau, Power BI, etc.), there is no longer the expectation that one must have advanced coding skills to make useful visualizations with their data.

The form of visualization should be appropriate for the question being asked and the type of data presented therefore, data visualizations will change depending on the context in which the data are being used. Nonetheless, effective visualizations are those that make it easy for audiences to quickly and efficiently understand the information being presented.

### Audience-Driven

Effective strategies for communicating with data lean into humans' natural tendencies to search for and identify patterns. However, we all possess our own mental models, cognitive biases, and personal experiences, which inform the types of information we seek out and what information resonates with us most. Therefore, it is important that communications are tailored to the intended audience.

When communicating with data, it is especially important to consider the audience's level of prior understanding of the topic or their general level of data literacy. What makes sense to a colleague with particular domain expertise may confuse a lay citizen. Similarly, what resonates with citizens may fail to engage policy makers. Therefore, an important rule of thumb is that communication about and with data should always be informed by and, thus, appropriate for the intended audience.

### Tell a Story

In data science, data visualization is often used as a form of storytelling as it benefits both the audience and the storyteller. Storytelling helps scientists make sense of data by tying them to a logic model that helps to explain their findings. Relaying that logic model using contextually relevant, audience-informed stories are one of the most effective strategies for communicating with data as it helps the audience connect that story to existing mental models, which enhance information processing, including attention, understanding, and memory.

### Accessibility

There are several important factors to consider to ensure accessibility when communicating to diverse audiences using data. A great place to start in designing communications of any type for accessibility is the Universal Design for Learning (UDL) Guidelines.

There have been widespread efforts in recent years to make design more inclusive for those with disabilities. As a result, the data science and open-source software communities have developed a number of free guides that are helpful for designing more accessible ways of communicating with and about data. While there is still a long way to go, there have been many advancements in designing visualizations for those with color blindness (see https://jfly.uni-koeln.de/color/), creating text that is readable by assistive devices (see Microsoft's "Everything you need to know about writing effective alt-text" ), and designing government information for accessibility (see U.S. Consumer Financial Protection Bureau's "Design System").

### Frame of Reference

The collection and use of big data in the public sector is context specific but there have still been widespread concerns about potential ethical, privacy and security implications of data applications in the public sector. This reality impacts how administrative data are collected, coded, and released to the public. As discussed, big data often involves combining data sets from many different sources. While a sufficient discussion of the implications of this for measurement, analysis, and decision-making is beyond the scope of this short article, the importance of providing audiences with a frame of reference by accurately communicating where data come from, limitations, biases, and other relevant contextual information remain the same. Furthermore, it is important not to overstate claims which may be mitigated by clearly communicating the role of context in data collection and analysis and acknowledging how context shapes the findings and their application in the public sector.

### Why?

The last principle is simple but increasingly important in our data-driven world. The amount of data collected and processed daily is growing exponentially which suggests that access to data is rarely the issue. Again, it is what we do with it that matters most. Therefore, it is crucial that we not only answer for ourselves, but clearly convey to others what we are trying to understand, improve, and advance in using these tools.

Data science has no shortage of "shiny tools," but the ability to leverage those tools for the tools that they, in fact, are is increasingly important. In the end, what is the value that new forms of data, advanced analytics, and even compelling data visualizations provide in our understanding of administrative behavior at the individual, organizational, and institutional levels? Once we have those answers, how can we best leverage the data science tools available to us to effectively communicate the importance of these questions and the implications of the answers we uncover?

### Now What?

Big data and applications of data science in PA indeed offer new ways for public managers and organizations to use data in administrative decision-making. Further, they offer PA researchers the opportunity to empirically test new and existing theories using novel data and analytical tools. Together, they may also provide opportunities for PA researchers and public organizations to work together to analyze unprecedented amounts of data to generate useful insights for both theory and practice. Nonetheless, this essay argued that for big data and data science to achieve such potential in PA, greater attention must be paid to the range of communication tools that come with them.

In a world where the amount of data we produce, collect, and process is growing exponentially, the value lies less in the data and more in our ability to effectively, efficiently, and equitably communicate about - and increasingly to - our datafied world. Therefore, our understanding of administrative applications of data through an *Administrative Informatics* lens must also include further attempts to understand what comes after specific decisions or policies have been made on the basis of that data.

Future research can examine the validity of these claims by empirically testing the effect of data visualizations, storytelling strategies, and inclusive design principles on individual-level behavior, including information processing and potential links to behavioral intentions and outcomes. At the same time, it is important to examine how these practices are shaped by organizational contexts and existing institutional factors that may or may not be shifting with these technological advancements. Doing so can help us to truly understand what an increasingly datafied world means for the public sector and DATAFY public administration in ways that positively impact theory and practice.

## Using Data in Government: Advancing Policy Frameworks to Support Operational Decision-Making—Nick Hart

In the United States (U.S.), the use of data to inform government decision-making started at the country's inception with the incorporation of demographic counts in the Constitution. The use of information collected through the decennial census now embodies core political processes, determines the allocation of funds, and is foundational for research activities – in both the public and private sector.

With data critical to the national ethos of the U.S., a series of federal laws have emerged to promote and encourage the use of data in decision-making, though applying these laws has proven challenging in practice. Early legislation focused on broad policymaking leading to a compliance-oriented mindset, and ultimately led to mixed results in their implementation. Recent changes to these laws provide new promise. This essay presents a brief context about the national legal authorities that facilitate the use of data for public sector employees, outlines the challenges facing public administration for bridging the ethos and practice, and offers suggestions for the research community in the years ahead.

### National Frameworks Encouraging Use of Data for Government Decision-Making

While the legacies of data management and evaluation have roots scattered across history, the key roots of the U.S. government-wide activities began in the early 1990s. First, the Chief Financial Officers Act of 1990 outlined expectations for aligning financial information and program performance plans. These reporting requirements were expanded substantially with the Government Performance and Results Act of 1993 (GPRA), which required most agencies to develop strategic plans and initiate methodical performance reporting. GPRA led to a proliferation of performance information and, in 2010, was reauthorized as the GPRA Modernization Act to focus on higher-level priority goals. It intended to provide more summary information, cutting through the vast trove of performance indicators now available but largely unused. The hope was that priority goals would prove more useful for policymakers. Congress even included mechanisms to punish agencies for not fulfilling targets on priority goals. Unfortunately, these mechanisms established perverse incentives for agencies as they set performance targets and selected annual performance indicators.

GPRA and GPRA Modernization provided opportunities to collect, manage, and use data at an operational level for decision-making, though that was not typically how the information was managed because performance management and measurement systems were designed for high-level budget reporting. The George W. Bush Administration rightfully observed that budget decisions were not based on performance information and launched the Program Assessment Rating Tool as a mechanism to consistently incorporate the information into budget decisions (Newcomer & Hart, 2022). Reported performance measures were presented to support budgetary decisions rather than broader operational activities, regulatory actions, and other day-to-day uses. Nonetheless, the performance information was published and made available through a central clearinghouse, performance.gov, in addition to static budget documents known as congressional justifications published annually by federal agencies.

In 2002, Congress passed the Confidential Information Protection and Statistical Efficiency Act (CIPSEA). While not an authority well-known outside the federal statistical system, it facilitates the protection of highly restricted and confidential information. It encourages greater use and sharing of restricted data across national statistical agencies. CIPSEA, in its original form, had a narrow focus on administrative and operational analytics because of the additional privacy safeguards under the framework. However, its focus was on "statistical activities," meaning group-level analysis of any kind of data. The authority and system could have been, but were not, oriented to demographic, economic, or operational questions alike.

In parallel with these activities, all agencies were directed in 2013 by the then-Barack Obama Administration to initiate a series of open data activities under existing administrative authority (Obama, 2013). By emphasizing the need for data to be open and machine-readable by default, government data would become more accessible to everyone, including program administrators and the American public. However, the effort led to the publication of largely existing datasets available through other mechanisms on data.gov.

Despite the open data efforts and the ongoing publication of assets from the federal statistical system, many gaps remained in performance, statistical, and administrative data systems. Recognizing these gaps and hoping to use more administrative data for research and operational activities, two members of Congress advocated for the establishment of an expert panel. In 2018, on the heels of the U.S. Commission on Evidence-Based Policymaking's unanimous recommendations for improving the government's data infrastructure, Congress passed the Foundations for Evidence-Based Policymaking Act (Evidence Act), which included the OPEN Government Data Act and expanded CIPSEA (Abraham et al., 2017).

The Evidence Act's authorities are expansive for encouraging government use of data for research, evidence-building, and operational analytics. Key features of the law include:

- *Data Governance Requirements*. Most fundamentally, the law created chief data officers, established data governance protocols, required data catalogs and inventories, and reinforced the 2013 Executive Order for open data by default.

- *Evaluation Capacity and Planning*. The law requires large agencies to designate evaluation officers, produce annual evaluation plans, and expects the establishment of a national evaluation workforce.

- *Statistical Data Sharing*. As part of the CIPSEA reauthorization, the law expanded statistical data sharing capabilities with a new provision referred to as the "presumption of accessibility" allowing federal statistical agencies to access administrative records unless there is an explicit prohibition to access and use for statistical activities.

- *Multi-Year Learning Agendas*. Addressing one of the longstanding challenges of the chasm between the supply and demand for research and evidence, the law requires agencies to produce periodic plans that articulate major questions and data needs aligned with quadrennial strategic plans.

Taken together, these provisions of the Evidence Act enable U.S. national agencies to collect, manage, analyze, and share data in new ways – even outlining an explicit expectation to do so. The Evidence Act provides a blended focus on top-level policy and front-line operational questions.

**Navigating the Challenges of Using Data for Public Sector Operational Analytics**

Encompassing the Evidence Act and the other frameworks are other major laws like the Privacy Act of 1974, the Paperwork Reduction Act of 1995, the Freedom of Information Act, and others – all of which have positive and negative effects on the ability for public sector officials to use administrative records for decision-making. In addition, sector and topic-specific authorities impose additional limits and incentives. These laws include the Financial Data Transparency Act, the Digital Accountability and Transparency Act (DATA Act), and others.

This matrix of authorities demonstrates the complexity that public sector officials navigate when using data. Even for seemingly simple tasks, officials must navigate a legal and bureaucratic puzzle to comply with relevant rules. To make matters, federal agencies must also consider over 3,000 sub-national privacy laws in the U.S.

For those seeking to conduct operational analytics, doing so responsibly within the existing legal matrix can be challenging. The Evidence Act framework provides useful pathways because of the alignment between the statistical, evaluation, and data governance authorities. Key challenges addressed within this framework include (Hart & Carmody, 2018):

- Chief data officers (CDO) provide a new capability to support the implementation of data standards, improve data quality, and champion data access in agencies. For public sector employees who need administrative records or information that may fall beyond the immediate scope of control but is relevant for conducting operational analytics, a properly resourced CDO can ensure successful data governance and access.

- Privacy has long been used as a shield among government attorneys at the expense of data use because privacy was imagined as an absolute. The Evidence Act in the U.S. establishes a protocol for data use that conceptualizes and manages privacy risk along a spectrum. This reimagined and realistic approach ensures that analysis is prioritized to extract value from data while also protecting confidential and sensitive information.

- The formal recognition of the field of program evaluation in the Evidence Act is not insignificant for those pursuing operational analytics. Evaluation supports a range of analytic activities, but is especially well-positioned to ask questions about performance, personnel, and operational tasks, including – critically – whether those activities are the right ones, should be undertaken in the first place, and are appropriately deployed. The marriage of evaluation to data in the new legal framework is an opportunity, but requires data analysts to potentially receive uncomfortable feedback about whether they are even asking the right questions.

For those working in or alongside public sector agencies to conduct operational analytics, the Evidence Act creates new prospects to analyze government data. But implementing this law continues to be perplexing due to the inertia from cultural, institutional, and capacity challenges in agencies, which are all realistic challenges to be overcome, and provide opportunities for the research community.

**Opportunities for Researchers and the Evidence-Building Community**

The research community and the broader evidence-building community have multiple opportunities to support implementation in the years ahead as the Evidence Act evolves. A few possibilities include:

- *Apply the Agency Learning Agendas to Research Opportunities.* In 2022, federal agencies published multi-year learning agendas, many for the first time. The research community can review those plans, participate in filling research needs, and support the provision of new data assets. Scholars can also identify gaps in the agendas that might include a lack of stakeholder engagement, misalignments for key priorities around operational analytics, or other areas that should be addressed or prioritized. The plans are available at evaluation.gov.

- *Request New Data Assets – And Use Them.* Historically public sector agencies are reluctant to publish open data assets where quality is questionable. The scholarly community can help improve data quality over time by encouraging agencies to publish their data catalogs and inventories, enable access to open data assets, and provide appropriate access to restricted data. Some data are available on data.gov, and restricted data from the federal statistical system can be applied at researchdatagov.org.

- *Advocate for System-wide Improvements.* The belief that researchers cannot advocate is misguided, especially when the advocacy can improve data access and quality. The Evidence Act is an opportunity for the research community to call on the government to ensure data portals, standards, access infrastructure, privacy reviews, and other data-related activities are adequately resourced to permit timely reviews and approvals. Data will only be available from government agencies with adequate, sustained resources and capacity; to achieve this, agencies and Congress must view data as a priority. Someone needs to ask and advocate for it to be a constant priority.

We need an effective administrative data ecosystem that includes data for operational analytics, statistical data, and more. Fundamental to this goal is coherent coordination between CDOs evaluation officers, statistical officials, and other disciplines. The legal framework is complex in the U.S., but has improved in recent years because of the overarching and evolving capabilities provided by the Evidence Act. The Evidence Act provides opportunities for public sector employees in the U.S. to support more operational analytics, overcoming longstanding challenges, but only if implemented well. The research community can also support effective implementation by participating in the process, making priorities known, and, when necessary, advocating for resources and capacity.

## References

Abraham, K. G., Haskins, R., Glied, S., Groves, R. M., Hahn, R., Hoynes, H., & Wallin, K. R. (2018). The Promise of evidence-based policymaking: Report of the commission on evidence-based policymaking. *Washington, DC: Commission on Evidence-Based Policymaking.*

Akbari, K., Winter, S., & Tomko, M. (2021). Spatial Causality: A systematic review on spatial causal inference. *Geographical Analysis* (0), 1-34. https://doi.org/10.1111/gean.12312

Baumgartner, F. R., & Jones, B. D. (2009). *Agendas and instability in American politics* (2nd ed). The University of Chicago Press.

*CARE Principles of Indigenous Data Governance.* (n.d.). Global Indigenous Data Alliance. Retrieved 29 October 2022, from https://www.gida-global.org/care

Choi, T., & Robertson, P. J. (2019). Contributors and free-riders in collaborative governance: A computational exploration of social motivation and its effects. *Journal of Public Administration Research and Theory*, *29*(3), 394-413. https://doi.org/10.1093/jopart/muy068

Conklin, J. (2005). *Dialogue Mapping: Building Shared Understanding of Wicked Problems.* John Wiley & Sons, Inc.

Cook, S. J., An, S. H, & Favero, N. (2018). Beyond policy diffusion: spatial econometric models of public administration. *Journal of Public Administration Research and Theory, 29*(4), 591–608. https://doi.org/10.1093/jopart/muy050

Devriendt, T., Borry, P., & Shabani, M. (2021). Factors that influence data sharing through data sharing platforms: A qualitative study on the views and experiences of cohort holders and platform developers. *PLOS ONE*, *16*(7), e0254202. https://doi.org/10.1371/journal.pone.0254202

Dillon, L., Walker, D., Shapiro, N., Underhill, V., Martenyi, M., Wylie, S., Lave, R., Murphy, M., & Brown, P. (2017). Environmental Data Justice and the Trump Administration: Reflections from the Environmental Data and Governance Initiative. *Environmental Justice*, *10*(6), 186–192. https://doi.org/10.1089/env.2017.0020

Fusi, F., Manzella, D., Louafi, S., & Welch, E. (2018). Building Global Genomics Initiatives and Enabling

Data Sharing: Insights from Multiple Case Studies. *OMICS: A Journal of Integrative Biology*, *22*(4), 237-247. https://doi.org/10.1089/omi.2017.0214

Grigoropoulou, N., & Small, M. L. (2022). The data revolution in social science needs qualitative research. *Nature Human Behaviour*, 1-3. https://doi.org/10.1038/s41562-022-01333-7

Hart, N. and K. Carmody. (2018). *Barriers to Using Government Data: Extended Analysis of the U.S. Commission on Evidence-Based Policymaking's Survey of Federal Agencies and Offices*. Washington, D.C.: Bipartisan Policy Center. Available at: (Hart & Carmody, 2018)

Homer, J., Milstein, B., and Hirsch, B. (2020). System Dynamics Modeling to Rethink Health System Reform in Yorghos Apostolopoulos, Michael K. Lemke, and Kristen Hassmiller Lich (Eds.) *Complex Systems and Population Health*, Oxford University Press.

Kane, G. C., & Ransbotham, S. (2016). Content as Community Regulator: The Recursive Relationship Between Consumption and Contribution in Open Collaboration Communities. *Organization Science*, *27*(5), 1258–1274. https://doi.org/10.1287/orsc.2016.1075

Maroulis, S., Diermeier, D., & Nisar, M. A. (2020). Discovery, dissemination, and information diversity in networked groups. *Social Networks*, *61*, 67-77. https://doi.org/10.1016/j.socnet.2019.08.007

Maroulis, S. (2016). Interpreting school choice treatment effects: Results and implications from computational experiments. *Journal of Artificial Societies and Social Simulation*, *19*(1), 7. https://doi.org/10.18564/jasss.3002

Maroulis, S., & Wilensky, U. (2015). Social and task interdependencies in the street-level implementation of innovation. *Journal of Public Administration Research and Theory*, *25*(3), 721-750. https://doi.org/10.1093/jopart/mut084

Maroulis, S., Guimera, R., Petry, H., Stringer, M. J., Gomez, L. M., Amaral, L. A. N., & Wilensky, U. (2010). Complex systems view of educational policy research. *Science*, *330*(6000), 38-39. https://doi.org/10.1126/science.1195153

Monastersky, R. (2015). Anthropocene: The human age. *Nature*, *519*(7542), 144–147. https://doi.org/10.1038/519144a

Neumann, M., Linder, F., & Desmarais, B. (2022). Government websites as data: a methodological pipeline with application to the websites of municipalities in the United States. *Journal of Information Technology & Politics, 19*(4), 411-422. https://doi.org/10.1080/19331681.2021.1999880

Newcomer, K. and N. Hart. (2022). *Evidence-Building and Evaluation in Government*. Sage.

Obama, B. (2013). Executive order--making open and machine readable the new default for government information. *The White House*.

Ottinger, G. (2013). Changing Knowledge, Local Knowledge, and Knowledge Gaps: STS Insights into Procedural Justice. *Science, Technology, & Human Values*, *38*(2), 250–270. https://doi.org/10.1177/0162243912469669

Overton, M., Larson, S., Carlson, L., & Kleinschmit, S. (2022). Public data primacy: The changing landscape of public service delivery as big data gets bigger. *Global Public Policy and Governance*, *2*(3). https://doi.org/10.1007/s43508-022-00052-z

Pierson, P. (2004). *Politics in time: History, institutions, and social analysis*. Princeton University Press.

Robinson, J. (2017). Brilliant analytics for smart cities in Marie Lowman (Ed.) *A practical guide to analytics for governments: using big data for good*. John Wiley & Sons.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton, Mifflin and Company.

SoRelle, M. E. (2020). *Democracy declined: The failed politics of consumer financial protection*. The University of Chicago Press.

Soroka, S. N., & Wlezien, C. (2010). *Degrees of democracy: Politics, public opinion, and policy*. Cambridge University Press.

Thompson, B., van Opheusden, B., Sumers, T., & Griffiths, T. L. (2022). Complex cognitive algorithms preserved by selective social learning in experimental populations. *Science*, *376*(6588), 95-98. https://doi.org/10.1126/science.abn0915

Tobler, W. (1970). A computer movie simulating urban growth in the Detroit region. *Economic Geography*, *46*(2), 234–240. https://doi.org/10.2307/143141

Vera, L. A., Walker, D., Murphy, M., Mansfield, B., Siad, L. M., & Ogden, J. (2019). When data justice and environmental justice meet: Formulating a response to extractive logic through environmental data justice. *Information, Communication & Society*, *22*(7), 1012–1028. https://doi.org/10.1080/1369118X.2019.1596293

Weible, C. M., Olofsson, K. L., & Heikkila, T. (2022). Advocacy coalitions, beliefs, and learning: An analysis of stability, change, and reinforcement. *Policy Studies Journal*, psj.12458. https://doi.org/10.1111/psj.12458

Weible, C. M., & Sabatier, P. A. (Eds.). (2017). *Theories of the policy process* (Fourth edition). Westview Press.

Workman, S. (2020). Four Principles of Data Collection. *Towards Data Science*. https://towardsdatascience.com/four-principles-of-data-collection-82ad07938ac1

Workman, S., Jones, B. D., & Jochim, A. E. (2009). Information Processing and Policy Dynamics. *Policy Studies Journal*, *37*(1), 75–92. https://doi.org/10.1111/j.1541-0072.2008.00296.x

Yeager, D. S., Hanselman, P., Walton, G. M., Murray, J. S., Crosnoe, R., Muller, C., ... & Dweck, C. S.

(2019). A national experiment reveals where a growth mindset improves achievement. *Nature*, *573*(7774), 364-369. https://doi.org/10.1038/s41586-019-1466-y

Zhang, F., & Maroulis, S. (2021). Experience is not enough: A dynamic explanation of the limited adaptation to extreme weather events in public organizations. *Global Environmental Change*, *70*, 102358. https://doi.org/10.1016/j.gloenvcha.2021.102358

Zhu, L., Witko, C., and Meier, K. (2019). The Public Administration manifesto II: Matching methods to theory and substance. *Journal of Public Administration Research and Theory*, *29*(2), 287–298. https://doi.org/10.1093/jopart/muy079Addi-Raccah, A., & Ainhoren, R. (2009). School governance and teachers' attitudes to parents' involvement in schools. *Teaching and Teacher Education, 25*(6), 805–813.