Introduction: A Roadmap to a Nationwide Data Infrastructure for Evidence-based Policymaking

Andrew Reamer

Julia Ingrid Lane

A Data Infrastructure for Evidence-based Policymaking

Throughout the United States, there is broad interest in expanding the nation's capacity to design and implement public policy based on solid evidence. That interest has been stimulated by the new types of data are available that can transform the way in which policy is designed and implemented. Yet progress in making use of sensitive data has been hindered by the legal, technical, and operational obstacles to access for research and evaluation. Progress has also been hindered by an almost exclusive focus on the interest and needs of the data users, rather than the interest and needs of the data providers. In addition, data stewardship is largely artisanal in nature.

There are very real consequences that result from lack of action. State and local governments are often hampered in their capacity to effectively mount and learn from innovative efforts. Although jurisdictions often have treasure troves of data from existing programs, the data are stove-piped, underused, and poorly maintained. The experience reported by one large city public health commissioner is too common: "We commissioners meet periodically to discuss specific childhood deaths in the city. In most cases, we each have a thick file on the child or family. But the only time we compare notes is after the child is dead." In reality, most localities lack the technical, analytical, staffing, and legal capacity to make effective use of existing and emerging resources.

It is our sense that fundamental changes are necessary and a new approach must be taken to building data infrastructures.   In particular,

1. Privacy and confidentiality issues must be addressed at the beginning—not added as an afterthought.
2. Data providers must be involved as key stakeholders throughout the design process
3. Workforce capacity must be developed at all levels.
4. The scholarly community must be engaged to identify the value to research and policy

There is substantial interest in building a data infrastructure that effects such changes.   In a very high profile example, Congress established the federal Commission on Evidence-Based Policymaking (CEP) through the bipartisan Evidence-Based Policymaking Commission Act of 2016 (P.L. 114-140), jointly sponsored by Speaker Paul Ryan (R-WI) and Senator Patty Murray (D-WA), and signed by President Barack Obama on March 30, 2016. On September 7, 2017, the CEP released its final report, "The Promise of Evidence-based Policymaking," laying out a series of recommendations regarding approaches to building the capacity of the federal government to generate and make effective use of evidence for public policy purposes while fully protecting privacy and confidentiality.

There is also an emerging opportunity to build such an infrastructure at scale. Recent years have seen dramatic growth in the technological capabilities to organize, link, integrate, and analyze enormous volumes of data from multiple, disparate sources[1, 2]. The new data make it possible to gather more information from a larger population (of people, households, or businesses) at far lower cost.

In order to develop a roadmap for the creation of such an infrastructure, the Bill and Melinda Gates Foundation, together with the Laura and John Arnold Foudation hosted a day-long workshop of more than sixty experts to discuss the findings of 12 commissioned papers and their implications for action. This issue of The ANNALS showcases those twelve articles. The workshop papers were grouped into three thematic areas: privacy and confidentiality, the views of data producers, and comprehensive strategies that have been used to build data infrastructures in other contexts. The authors and the attendees included computer scientists, social scientists, practitioners and data producers.

This introductory chapter places the research in both a historical and a current context. It also provides a framework for understanding the contribution of the 12 papers.

The Historical Context of Evidence-based Policymaking

While the term "evidence-based policymaking" is of recent origin, the notion of designing public policies based on data analysis has been practiced in the English-speaking world since the 1600s (see Box 1).

BOX 1

A Brief History of Statistical Capacity for Evidence-based Policymaking

- 1662 — John Graunt publishes Observations on the Bills of Mortality, the first time that British health statistics are collected and used for public health policy design.
- 1690 — William Petty and Charles Davenant invent "political arithmetic" to calculate optimal British tariffs and taxation.
- 1751 — Benjamin Franklin writes and privately circulates "Observations Concerning the Increase of Mankind, Peopling of Countries, Etc," laying the groundwork for demographic analysis for the purpose of public policy. The next year Franklin creates the first successful fire insurance company in the colonies. [?/ I think it's a fire company, not a fire insurance company] [He did both – the fire company in 1736 and the fire insurance company in 1752. See http://www.pbs.org/benfranklin/l3_citizen_insurance.html]

- 1790 — Congress approves Representative James Madison's proposal to add questions to the First Census so that Congress could "adapt the public measures to the particular circumstances of the community." Madison's idea lives today as the Census Bureau's American Community Survey.
- 1791 — Sir John Sinclair publishes the first comprehensive survey-based statistical study in the English-speaking world. Sinclair adopted new German "statistical" methods (the "science of dealing with data about the condition of a state") to measure the "quantum of happiness" that existed in Scotland and identify ways of improving this.
- 1884 — Carroll D. Wright proposes in an address to the American Social Science Association that tariffs be set "scientifically," based not on politics, but on data. Congress takes up the cause, expanding the federal capacity to collect data, and in so doing created the foundations of today's federal economic statistical agencies.
- 1946 — The Employment Act of 1946 establishes the infrastructure for data-driven macroeconomic (fiscal and monetary) policy.
- 1966 — As part of efforts to measure the impact and improve Lyndon Johnson's Great Society programs, the White House proposes a National Data Center, essentially a federal data clearinghouse, but Congress does not approve the proposal because of broadly voiced concerns about the potential of the center to invade personal privacy.

A perennial problem—overcoming decision-makers' penchants for ignoring or ridiculing evidence—must therefore be addressed, but that is beyond the scope of this volume. What we propose here is a path toward capitalizing on advances in IT that make it possible for policy to be more evidence-based than ever before.

Current Context

Recently, multiple disparate efforts have been initiated to promote data infrastructure development for evidence-based policymaking. These efforts are housed in governments at all levels, in nonprofits, and in universities. Many have been supported by philanthropic foundations (see Box 2).

BOX 2

Examples of Current Data Centers and Networks

- Centers
  - Federal
    - Federal Statistical Data Research Center Program – partnerships between federal statistical agencies and leading research institutions. (This program has one database and multiple access points for qualified researchers.)
    - Census Bureau's Administrative Records Clearinghouse for the Evaluation of Federal and Federally Sponsored Programs (funded by Congress; $10 million annually)
    - Commission on Evidence-Based Policymaking – created by Congress to ascertain desirability and design of federal data clearinghouse
  - Nonprofit
    - NORC Data Enclave

        - Private Capital Research Institute
        - Health Care Cost Institute

    - University
        - Institute for Research on Innovation and Science (IRIS), University of Michigan
        - Inter-university Consortium for Political and Social Research, University of Michigan
        - Institute for Research in the Social Sciences, Stanford University
        - Dataverse, Harvard University
        - Massive Data Institute, Georgetown University
        - Minnesota Population Center, University of Minnesota
- Networks (associations of centers)
    - Actionable Intelligence for Social Policy – organized by University of Pennsylvania
        - The network consists of thirteen sites, providing data on five states (Florida, Michigan, South Carolina, Washington, Wisconsin) and eight counties or cities (Allegheny County [Pittsburgh], Cook County [Chicago], Cuyahoga County [Cleveland], Los Angeles County, Mecklenburg County [Charlotte], New York City, and Philadelphia)
        - Made possible by a grant from the MacArthur Foundation
    - Administrative Records Data Network – initiative of the Sloan Foundation
    - National Neighborhood Indicators Partnership, The Urban Institute

- Civics Analytics Network – national peer network of urban Chief Data Officers (CDOs) funded by the Arnold Foundation and managed by Harvard University

Operators of data repositories and systems may include nonprofit organizations, universities, state governments, local governments, federal governments, think tanks, and industry associations. Data repositories and systems may be accessed by researchers and analysts employed by universities, governments, think tanks, other nonprofits, and for-profits. They may receive technical assistance—in business operations and data management and methods—from network organizations (via network staff or peers), universities and other nonprofits, and for-profit consultants

Data can have either a geographic or topical focus.   They can be held by (i) *individual government programs*, particularly administrative records (e.g., Social Security); (ii) *corporate and nonprofit* providers of own data (e.g., Uber or Mastercard); (iii) *data repositories* that bring together data from multiple sources without attempting to integrate them (e.g., data.gov); (iv) *enterprise data systems* that bring together, and make compatible, data from multiple programs (e.g., the Census Bureau's enterprise data system); (v) *integrated data systems* that bring together, and make compatible, data from multiple enterprises, e.g., state- or county-wide (e.g., Data Share in Milwaukee); and (vi) *nationwide networks* of programs, data repositories, enterprise data systems, or integrated data systems (such as Chapin Hall at the University of Chicago), which are networks of organizations and typically do not (as yet) involve data-sharing among members.

Building Robust National Data Infrastructure

The articles commissioned for this project make clear while there are considerable barriers to the development of a robust data infrastructure ecosystem, it is possible to address them. The papers point to six areas in which progress can be made: legitimacy, public data providers, law, transaction costs, data protection, and ecosystem self-organization.

Legitimacy

It is important to treat data subjects as a stakeholder group and offer an opportunity for input on data infrastructure uses and operations. Rare disease research consortia provide a model for doing this. Notions that may prove useful include "collective consent" and "social license."

The legitimacy of data infrastructure organizations could also be called into question if stakeholders have concerns about other dimensions of trustworthiness, such as capabilities and practices. To address this issue, some project participants suggested that a process be created by which organizations are certified by a national authorized body as having the desired capabilities and practices.

Also, it is important for data infrastructure proponents to have the capacity to disentangle concerns expressed about protecting subjects' privacy from political concerns, say about the proper role of government programs in a market economy. Sometimes, project participants said, their experience suggests that political concerns are raised in the guise of a desire to protect privacy, for example, in the case of the Family Education Rights and Privacy Act (FERPA).

Particularly important to legitimacy is identifying means to institutionalize the use of evidence in policymaking in the executive and legislative branches. In this regard, it is helpful, but certainly not sufficient, that the Evidence-Based Policymaking Commission Act of 2016 (H.R. 1831) passed both Houses of Congress with strong bipartisan support.

Public data providers

Data infrastructure development and evidence-based policymaking can be successful only if public data providers, particularly public program managers, are invested in making their data available and usable and acting on the results of analysis. Project participants identified a number of challenges that need to be addressed in any design, including individual agencies' willingness to be evaluated and give up control of their data. In addition, public data providers suffer from lack of staff trained to document, extract, and transmit data and lack of capacity to establish data standards and IT capacity. Projects that are inordinately long term in nature are unlikely to succeed, since providers do not have the budget or the mandate to take on long-term data-driven performance improvement projects. Project participants found that rather than bemoaning these problems, it is essential to identify projects that provide short term wins while moving toward a long-term goal.

Law

Project participants found that laws regarding data access and privacy serve as a barrier to data infrastructure development in multiple ways. Most obviously, laws are complex and siloed. At the federal level, data access is governed by numerous laws, including the Health Insurance Portability and Accountability Act (HIPAA), Family Educational Rights and Privacy Act (FERPA), Video Privacy Protection Act (VPPA), Children's Online Privacy Protection Act (COPPA), and regulations concerning personally identifiable information (PII). Similar complexity exists at the state level. Some laws unnecessarily restrict data access and some laws are difficult to clearly interpret. Many states restrict federal re-use of their data. Many lawyers are risk averse, choosing to serve as gatekeepers rather than problem-solvers. Many government agency lawyers do not have proper and adequate training in privacy and confidentiality law.

To address these issues, project participants proposed a number of steps. These include conducting a systematic review and analysis of existing laws to facilitate greater consistency, clarity, and legitimate data access; creating a roster of qualified legal experts; and creating training curricula in data privacy and confidentiality, which include legal toolkits in data privacy and confidentiality, such as model legislative and legal agreement templates. The political approaches would include passing federal and state laws that require use of data for analysis.

As an example of efforts in the field, the State Data Sharing Initiative (SDS), funded by the Arnold Foundation, works in a number of states to reduce the extent to which law and regulation are barriers to data access.

Transaction costs

Successful expansion of the nation's data infrastructure for evidence-based policymaking requires a substantial reduction in the transaction costs now borne by data providers, repositories, systems, and networks.

At present, due to the artisanal nature of most efforts, organizations too often need to recreate what others have already developed and put much time and effort into crafting standards, templates, methods, practices, and laws that enable legitimacy and good organizational, project, and data management. Project participants proposed the following approaches for reducing transaction costs: (i) carry out ecosystem-wide efforts to develop, adopt, and disseminate widely affirmed and "evidence-tested" standards, templates, methods, practices, and laws; (ii) establish repositories of technical assistance and coaching talent that would be available to eligible organizations; (iii) facilitate information sharing through peer learning and executive education; (iv) create laboratories for testing new methods; and (v) learn lessons from existing federal-state cooperative data programs.

Organizations also face significant transaction costs in data management. To address this, they suggested that the new capacity in computer science could be used to develop (i) processes for automated data stewardship and (ii) cloud-based data enclaves for shared data, methods, results, and expertise.

Data protection

Trustworthy, effective data protection is essential for successful infrastructure expansion. While data protection challenges are numerous and ever-growing, good practices have been identified and can be built upon. These include (i) statistical disclosure prevention techniques, such as deidentification, fuzzy data, synthetic data, simulation, differential privacy algorithms; and (ii) access management mechanisms, such as interactive query systems, two-factor authentication, unforgeable query logs, and secure data enclaves. Concerted efforts should be made to enhance, test, standardize, and disseminate good data protection practices and technologies.

Conclusion

Examples are now frequently emerging of “artisanal” efforts to build data infrastructures. Too often, however, one individual must work tirelessly to overcome the legal and bureaucratic barriers to linking and using existing administrative or survey data on a one-off basis. Often the system is built on the trust of a few key actors, and unique rules and work-arounds are developed. Enough has been learned to build a new approach. It is clear that to move from “artisanal” integrated data systems to professional, routinized and sustainable systems requires starting with the value propositions to data owners and other stakeholders.

Achieving long term useable and sustainable integrated systems will require focusing on creating standardized systems for creating access while ensuring privacy, secure data centers, and useable standards for key elements such as identifiers, core data structures, and metadata. It will be important to make investments that demonstrate the impact of innovations on a large scale and provide blueprints and frameworks that make it possible for other localities to create capacity far more effectively.

Any approach should focus on building sustainable, coordinated and scalable data infrastructures built around state-of-the-art technology with effective and safe privacy protections. In so doing, this effort will be a valuable partner to many existing efforts, starting with highly motivated state and local programs with compelling project ideas, and inspire other jurisdictions. It is possible to build national framework for sharing and learning both about effective policies and programs and highly useable data infrastructure.

References

Commission on Evidence-based Policymaking, "The Promise of Evidence-Based Policymaking: Report of the Commission on Evidence-based Policymaking," Washington, DC: September 7, 2017.

Deloitte Center for the Edge. 2014. Shaping strategies: Executive summary. New York, NY: Deloitte Center for the Edge.

Hagell III, John, John Seely Brown, and Lang Davison. 2010. *The power of pull: How small moves, smartly made, can set big things in motion*. New York, NY: Basic Books.

Reamer, Andrew, “Congressional Attitudes to Evidence-based Policymaking: An Historical Review,” presentation to the Legislative Branch Capacity Working Group, Washington, DC, July 17, 2017.

Reamer, Andrew, “Before the U.S. Tariff Commission: Congressional Efforts to Obtain Statistics and Analysis for Tariff-setting, 1789–1916,” prepared for the U.S. International Trade Commission, March 2017.

U.S. Census Bureau, “Statistics of U.S. Businesses – Noise Infusion,” at https://www.census.gov/programs-surveys/susb/technical-documentation/methodology.html, accessed September 25, 2017.

Williams, Anthony. 19 March 2017. What data-driven mayors don’t get. *The Atlantic CityLab*. Available from https://www.citylab.com/politics/2017/03/what-data-driven-mayors-dont-get/520092/.