Toward Building State and Local Data Analytic Capacity through the Nationwide Data Initiative

Amy O'Hara
Stanford University

Introduction

In the United States, over 19,000 municipal governments and 3,000 counties generate data. How many of them are aware of their opportunities for deeper data use? Data can shape their programs and policies. This paper focuses on state and local data, including cities and counties. These are the places people live and work, interact with government and consume services. Some needs for data analytics are common across local areas, and some cities and counties have unique needs to support their populations or locations.

What can prompt the development of data-driven cultures? We need a campaign that is wide-ranging, with combined forces, and keen strategies. I propose three areas to prompt action:

- Build interest
- Build human capital
- Build technology

Building interest involves increasing awareness in the areas producing the data. This involves sharing what has worked – and what hasn't worked – in other data infrastructure projects and building trust with the officials who will agree to data access and evaluation.

Building human capital includes human capital development across many layers of government, but especially analytics capacity in program and policy offices.

Building technology ranges from tools to systems, and everything in between. The local government's information technology (IT) life cycle, budget, and security needs affect technology choices to build data capacity. Opportunities to obtain managed services for data hosting and analysis may allow governments to ease into new platforms with less disruption to current operational systems.

This paper expands on these three goals, then considers what needs to be learned to spread data-driven cultures throughout local governments.

Build interest

Some cities, counties, and states now use their administrative data to extract insights and utility. A few jurisdictions have been doing so for decades through state- or university-led efforts.[1] Their data use has improved program administration, community health and safety, and policymaking. Early data infrastructure adopters have realized cost savings through more efficient operations and have quicker responses to emergencies.

Thousands of other jurisdictions may lack awareness, resources, or enthusiasm to use their data. These state and local agencies are busy providing services on existing budgets for education, health and human services, and public safety. While some are connected through associations (e.g., National Conference of State Legislatures, National Governors Association, National Association of Counties, National League of Cities, and National Conference of Mayors) or networks (e.g., Smart Cities, MetroLab, and National Network Indicators Project), broad buy-in for data driven decision-making and governing is absent. They need to embrace the opportunity to improve and innovate.

Describing the payoff

Better communication among capacity-seeking and experienced cities, counties and states can spread innovation stories, spurring both insight and action. Interested governments need to voice their concerns, so that more experienced areas can explain how data can make a difference. Perhaps they seek ways to avoid waste, fraud, and abuse, or improved outreach and program participation. Perhaps they simply want to improve their planning and program administration. These governments can access tools to upgrade data management and documentation, producing better data to feed dashboards and scorecards. By linking across programs and over time, they can engage in robust research and evaluations.[2]

The dialogue should include multiple layers of government, from front-line workers to policymakers. The data generators need to understand that data can be a substantial asset, not only a liability or risk to be managed. Evaluators and policy analysts can explain how data can generate forecasts and inform interventions. Researchers can share examples of projects using longitudinal data to observe outcomes and targeting long-standing problems invisible on any given quarter's dashboard.

Trust and enthusiasm

Interested jurisdictions need to develop trust in data use, trust in systems that can protect the data, and trust in the people using the data. Their enthusiasm will grow once they see how data addresses their gaps, and they will discover solutions that can address their needs. They can learn about support available from foundations and federal grants, and technical assistance to use open source tools. They can also get moral support to push through problems.

There are many risks facing jurisdictions considering a capacity-building project. These include loss of data control, potential embarrassment, potential data misuse, and higher operations and maintenance costs. Methods that mitigate these risks must be shared to build trust.[3]

Build human capital

A variety of training initiatives provide data analytics skills in cities, counties, and states. Some efforts train existing state and local employees, others train people who will provide contract services, or train the next generation of data analytics leads and policy analysts. To reach state and local employees across thousands of agencies, training should be offered in-

person, online, and onsite.

Training analysts and leaders

Numerous universities and non-profits offer data analytics training. In some cases, groups signing up for the training agree to bring their data or use a vendor's platform such as the Administrative Data Research Facility (ADRF). Many classroom and online data science bootcamps exist. Some, like the Coleridge Initiative, specifically cater to government employees. Once individuals are trained, they need access to the right tools and systems to use their new skills. For example, the government units may seek platform-as-a-service solutions to keep their newly trained staff engaged.

A growing number of data science programs offer training to university students, ranging from undergrad summer programs to graduate degrees. These include Data Science and Public Policy/Data Science for Social Good (University of Chicago), Data Science for Public Policy (Georgetown University), and Data Science for the Public Good (Virginia Tech). How are these similar or different? Where are graduates being placed, are any in state or local government? How can public service jobs get matched to these students? What will it take for this space to mature into a "social informatics" curriculum that is recognized across domains and employers, like health informatics?

Beyond number-crunching for analysts, training is also available for agency or program leaders to cultivate data-driven cultures or learn about IT innovations. The leaders need to be ready to see new reports and insights – and act on the evidence. Where is the "now-what layer"? After a newly trained analyst produces a report and visualization showing inefficiencies, who designs the next step? Partnering with an intermediary, like Actionable Intelligence for Social Policy (AISP) or the Government Performance Lab (GPL), can help.

Leaders also need training on IT options and risk mitigation. Most jurisdictions lack Chief Data Officers or analytics teams, many reject the notion of copying their data into a new place – fearing procurement issues, reliability of the vendor, concern over control, concern over cloud storage, and interoperability issues. Training on security, privacy, and transparency can help leaders understand their options.

Embedding support

Some in the community of data-intensive problem solvers, such as the GPL, California Policy Lab (CPL), and Center for Government Excellence (CGE), work directly with government offices to create sustainable systems improving performance, contracting, evaluations, and transparency.[4] Other groups embed expertise to reform state Integrated Data Systems (IDS). These include the long-standing technical assistance from AISP and, more recently, Code for America. This is far from a complete list, and groups that embed assistance are not alike. CGE and GPL work closely with existing staff, bringing in a new way of thinking and doing. They deploy systems (including roles and governance) before they leave. CPL sends experts into government to improve and analyze data to address pressing policy questions.

We should understand these models more: It is important to understand the motives for embedding, particularly whether it is driven by an inability or unwillingness to move data.
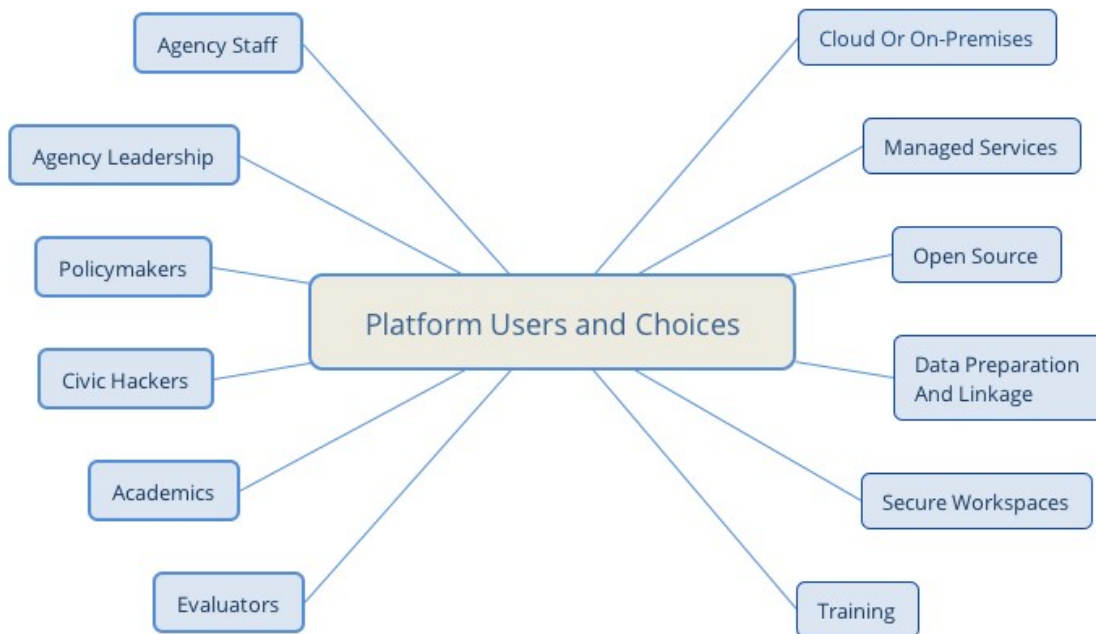
Some places may be unable to move data due to laws or regulations. We can assemble a list of places that will not use a platform, even a free one, and embedding may be the only way to analyze restricted data. Other area may face a cultural barrier, not a legal one. They may fear the change and need to be educated on security practices and risk mitigation.

Build the technology

Platform development initiatives vary across cities, counties, and states, some to aid operations and others to support research and evaluation. The platforms offer an array of choices for cities, counties, and states depending on their tastes for the cloud, control over their data, and the need for support to clean and harmonize across systems. Some platforms have been developed with a focus on a specific geography (e.g., Rhode Island Innovative Policy Lab), topic (e.g., ReEmployUSA), or cohort (e.g., Children's Data Network). Some IT approaches redesign the underlying systems that a city, state, or county is using, replacing old systems or making the existing interoperable. Other approaches intake and host the data (e.g., Palantir, ADRF, American Institutes for Research, Sailfish, NORC Data Enclave). The platforms vary in their management of personally identifiable information and protocols and methods of linking data.

Other initiatives have proposed a different approach, using a standardized, reformatted set of data elements that feed an analytics system. This approach, to identify and clean a small number of variables, works well for compliance or dashboard needs. While the appropriate limited set of elements can answer many evaluation and policy questions, it cannot support deep dives into a program or observe effects from nuanced policy or population changes. The standardized applications will be useful to areas getting support to build or improve their IDS (e.g., AISP, Code for America).

The ideal platform for each jurisdiction will depend on who will use it, for what purposes, and their IT requirements. The technology space is rapidly changing, with many platforms emerging in the past few years that serve certain users. The diagram below shows a wide range of users with varying levels of technical expertise. Platforms must have tools and interfaces that match diverse users' needs. Different users will require different interfaces and different tools. Some need individual-level data and others need aggregates, some need longitudinal data, and others need real-time data. Platform characteristics and choices will be determined by location, maintenance, and service requirements. In addition, platforms must be safe places, contain safe data, enable safe projects, permit only safe people, and support processes that disseminate safe outputs.

Platform Users and Choices

- Agency Staff
- Agency Leadership
- Policymakers
- Civic Hackers
- Academics
- Evaluators
- Cloud Or On-Premises
- Managed Services
- Open Source
- Data Preparation And Linkage
- Secure Workspaces
- Training

How to proceed

Pursuing these three goals – building interest, human capital, and technology – requires information collection and sharing among many factions. This campaign for data capacity has many fronts, requiring operatives to:

- Gather data on potential platform users and needs.
- Define the range of services and products available to cities, counties and states.
- Ask potential users what they need and classify the responses.
- Create a spectrum of involvement for the data owners.

Regarding the last point, the lightest touch is where data management duties are outsourced and a service provider is responsible for warehousing and access.[5] At the other end of the spectrum are places with data analytics offices and chief data officers that use their data intensively. According to Government Technology, eight states have Chief Data Officers (CDO), two counties have CDOs, and twelve cities have CDOs.[6] Only four of these CDOs have been in place since 2014, the rest were placed more recently. Most jurisdictions fall in the middle of the spectrum, managing their own data and wanting to get more utility from it.

Moving them forward needs a great deal of planning. Much information must be gathered and synthesized to understand the best next steps. To start that process, I offer twenty questions to expose role definitions, incentives, precedents, and scope.

Twenty questions

1) Who will engage representatives from associations of mayors, counties, governors, public health, and other policy stakeholder groups?

2) Who will involve stakeholders in domains including health, education, labor, and

justice?

3) Who will engage intermediaries who know current projects (e.g., in the operational space, in the Randomized Controlled Trial (RCT) and Pay for Success space, in the evidence-based budgeting and procurement space)?

4) Who will connect with community and advocacy group representatives to explain how and why data are being used?

5) Who will engage the public, and how?

6) Who will determine how these state and local initiatives relate to federal efforts?

7) Who will align these initiatives with university-led efforts (e.g., data hosting and analytics, RCTs)?

8) Who will consider data use and interoperability guidelines?

9) What places have applied for technical assistance or infrastructure development through other initiatives?

10) What do jurisdictions want out of capacity building?

11) What happens to data infrastructure or analytics labs when political leadership changes? What does history show?

12) Can regional analytics hubs support the dissemination and technical assistance needed nationwide?

13) What are foundations already funding and when will those projects be completed?

14) What can be learned from states that have data hubs? Where did their initial investment come from, and how long ago?

15) What are procurement strategies for contracting IT services, hardware, or training? Are Chief Information Officers on board?

16) What can be learned from the open data movement, did early vendors/platforms survive?

17) What are incentives for a city, county, or state to improve their data quality and documentation?

18) What is the incentive for existing topic-specific infrastructures (e.g., child-focused or education-focused) to branch into other areas or open their data to other users?

19) Who can develop code and tools to produce, reports, dashboards, graphics, predictions, and regressions?

20) What is a realistic goal for success – some capacity at any level in every state? Capacity in some percentage of the largest cities? Of all counties? What would ambitious goals look like?

Conclusion

We must develop incentives to instill data-driven cultures in state and local governments, encourage data owners to update their IT infrastructure and analyze their data on a research platform. We also need to develop incentives to support cross-agency and cross-

domain analyses. Such incentives will need to be sold to multiple layers across government agencies. Successes will be hard-won, and likely possible by following the work of experts who have succeeded in past conflicts. This campaign will advance as we increase awareness of deeper data use and will be ultimately successful if we gain acceptance, implementation, and policy action.

---

[1] Long running examples include Washington State Institute for Public Policy, Chapin Hall at the University of Chicago, AISP, and Institute for Research on Poverty at the University of Wisconsin-Madison.

[2] See case studies in "Building the Smarter State: The Role of Data Labs" by Anirudh Dinesh, @TheGovLab https://medium.com/data-labs/building-the-smarter-state-the-role-of-data-labs-5b5428920f0f

[3] Projects led by Virginia Tech described in http://theconversation.com/how-to-put-data-to-work-in-your-neighborhood-85350.

[4] Note that CGE is embedding to help cities generate open data, not restricted data, and not joins across program silos. Their success demonstrates that the right incentives (including funding and herd behavior), get mayors to embrace capacity building and culture change.

[5] Some jurisdictions view their data as held hostage by the service providers – they lack ready access to their information and are constrained to reports that the vendor supplies.

[6] From "Chief Data Officers: Mapping Which State and Local Governments Have a CDO," http://www.govtech.com/people/Chief-Data-Officers-Mapping-Which-State-and-Local-Governments-Have-a-CDO.html.