Building Data Science & Analytics Capabilities within Government

Every day, state, county, and municipal governments improve citizen-facing services through the power of data science. Although it might sound complicated, "data science" simply means leveraging technology to find patterns in data. Identifying these patterns helps organizations better understand their business, communities, or the environment in which they operate.

While the term "data science" is most commonly associated with advanced techniques like machine learning and predictive analytics, even its simplest techniques can provide significant insights. Data scientists refer to these techniques as "descriptive" analytics—meaning they join together or visualize exsisting data. This illuminates patterns or connections that are hard to see otherwise.

For example, many cities are connecting data from health departments, child protective services, and law enforcement to develop a more comprehensive profile of families at risk. Similarly, neighborhood-level census data in combination with 311 call report data has been used to better identify neighborhood-specific city service needs.

Moving from "description" to "prediction" can unlock even greater benefits — public sector organizations are using data science tools to more efficiently target resources to where they are needed most.

For example...

- **State agencies** use predictive analytics to find households that are unenrolled but likely eligible for benefit programs, or to find those at risk of being victimized by financial scams.
- **Localities** are using predictive analytics to focus enforcement resources on restaurants likely to have health code violations, or homes unlikely to have smoke detectors.
- **School districts, large employers, and universities** are developing early warning tools to predict student dropouts and employees likely to quit.

These tools prioritize scarce resources, helping organizations save money, and improve the quality of citizen-facing services.

In this white paper, we provide practical advice for decision makers who want to develop or strengthen data science capabilities within their department, agency, city or county as a whole. We focus on implementing organizational structures and the people, tools, data and security that will allow your data science competency to thrive.

Finding Success in Data Science

Local government organizations are just beginning to use data science, and it can sometimes seem overwhelming. But developing a data science competency is just like developing any other competency in an organization — you don't need a huge team; you need resources deployed and organized in the right way.

Organizational Structure: The structure of your data science organization is important and should be considered carefully. Most municipal governments choose to adopt a citywide data science capability that is housed within one department and consults across departments. States often implement data science capabilities at the agency level.

People: People, as with so many other things, are the most important piece of the puzzle. Organizations need to have a team, ideally in-house, with the skill sets required to use data for decision-making. You probably already have these skills within your organization, they just need to be deployed the right way.

Data: The ability to access data in an organization, which is nearly always siloed across departments and systems, is crucial. To become a data-driven organization, data should be accessible across departments and offices. You don't have to get your data all in one place, but you do need an environment where you can merge and explore data from different departments.

Tools: Proper analysis requires access to a set of software tools, most of which are available through "open source" technologies that you do not have to buy. Proper analysis also requires computational and data storage resources.

Security: Data security and privacy protection are key concerns for any organization implementing data science. Following best practices for data security, privacy protection, and compliance is crucial.

In the following sections, we go through each component in detail.

Organizational Structure

It's crucial for organizations to figure out where, organizationally speaking, their data science teams will live. Data science teams can be embedded in one department, serve as a cross-departmental resource, or even exist as a separate entity.

A best practice is for data science teams (often simply called "analytics teams") to work as internal, "multipurpose" consulting organizations which serve various offices and departments as their clients, and work across topic areas, and domains. Another approach is to create a consolidated, cross-agency data science capability around a specific issue, or set of issues. These focused efforts, which can be centered around issues like disaster response or childhood health, can optimize resources from across organizational jurisdictions. This approach also has the benefit of providing more focus, which can lead to quicker organizational wins that illustrate the power of data science.

Data science teams are housed in different agencies or offices based on budgetary and organizational factors that are often specific to particular localities. Boston, for instance, has a citywide analytics team integrated into their IT department, while Washington, D.C. integrates their data science team, referred to as "The Lab @ DC," into the mayor's office.

One great resource with lots of information about how to implement a data analytics capability at a municipal level is the Data-Smart City Solutions Initiative, within the Ash Center at Harvard. They publish a great set of resources and case studies.

People

Successful data science teams require individuals to hold several key roles, with different skill sets which can be found within a government organization. Note that in a small team, one person can fill multiple roles.

Key Members of an Effective Data Science Team

Role	Responsibilities	Key Skills	Candidates within Government Organizations
Project Owner	• Team leader and chief marketer • Main point of contact with external stakeholders • Final decider for task prioritization	• Strong communication skills, strategic vision for the organization	• Program management staff, Innovation Officer
Project Manager	• Own day-to-day management of the project (prioritize tasks, estimate timelines, communicate with stakeholders) • Solicit user feedback • Maintain project status bookkeeping	• Strong project management skills, a good understanding of the user perspective for tools resulting from data science projects	• Project managers from throughout the organization who have managed cross-departmental projects
Data Engineer	• Build and maintain data pipelines in and out of system • Clean, format and consolidate data	• Data management and cleaning, strong programming skills (often in SQL) and a clear understanding of database design and data modeling	• Data analysts or IT staff that work with large databases
Data Scientist	• Scientific lead: build and maintain statistical and machine learning models • Understand business context, encode in modeling software • Work with data engineer on feature engineering	• Strong understanding of statistics, and/or machine learning techniques. Programming skills in a language like R or python	• Social science researchers, more sophisticated data analysts with a statistics background
Software Engineer	• Build and maintain general purpose software (dashboard, monitoring systems, deployment code, etc.) • Architect computing resources • Guide data scientists (and potentially engineers) in software engineering best practices	• Excellent working knowledge of hardware and software. Often programs in multiple languages (often JavaScript), understanding of how to put a data science tool into production in the existing IT environment	• Existing IT staff

Building a Team

When building organizational data science capabilities, it is important to start developing a team structure, even if you need to fill roles with staff who also serve other non-data science related functions in your organization. A data science capability cannot be built around a single employee or contractor, no matter how talented.

One successful strategy could be to hire a single data science specialist with experience in data engineering and data science, and then to fill the other team roles with existing staff who already have key skill sets.

When starting out, a simple team might look like this:

- Project Manager/Owner (existing management)
- Data Scientist/Data Engineer (one new hire, and one existing staff member)
- Software Engineer (existing IT staff member)

A more complex team at a larger organization may look like the following:

- Owner
- Project Manager
- Two Data Scientists
- Three Data Engineers
- Two Software Engineers

When structuring a potential team, focus on these principles:

- **Data engineering comes first:** Before hiring a speciality data scientist, make sure you have people who can clean, merge, and move data. Without that critical first step, even the most advanced algorithms are useless.
- **Use the buddy system:** It can be difficult for small organizations to find data scientists and it can be tempting to hire a data scientist as a team of one. However, data scientists cannot deliver results on their own. They need to interface with others throughout a project lifecycle to be successful.
- **Don't create a team entirely composed of new team members:** Make sure that at least one current staff member has a role on the new team, even if it is just as an advisor. Having current staff members on your team will prevent misunderstandings around data resources and provide a greater appreciation for the operational, regulatory or cultural barriers that public sector organizations face.

> Tip: If you are just getting started in data science, it may be some time before you need a true data scientist for predictive modeling or machine learning — focus on hiring a data engineer first.

Finding the Right People

The most important piece when introducing data science into an organization is locating the right people. Typically, there are three options:

1. Finding data-literate staff and training them in data science
2. Hiring data science expertise; or
3. Contracting with an outside organization to get started.

We provide best practices for those three options below.

Train Your Own Staff

We encourage local government organizations to find and upskill potential data scientists within their organizations. Because data science is a critical skill for many new roles in government, building skills internally can pay massive dividends for your organization.

The ideal candidates for a data science team should possess the following:

- A background in data analysis within the applicable domain (e.g., public health, law enforcement)
- The ability to use data to tell a story
- Have experience using tools like SQL, R, Python, Stata, or ArcGIS

These individuals do not necessarily have to hold a "data scientist" or "data engineer" title. They may be the go-to person when someone needs to get new data pulled from a website, or someone who makes tables or graphs from data that tell a story really well. Odds are good that there are already people on your team with the ability to interpret data who may be interested in building on those skills.

Although educational backgrounds in computer science or statistics can be helpful for building a data science team, you do not necessarily need someone with a particular degree to be a data scientist. Most data scientists and data engineers do not have a computer science background and many come from liberal arts or social science backgrounds. Because of the explosion of interest in data science, there are academies, "bootcamps," and online courses your data-literate staff can take to prepare for a role as a data engineer or data scientist without going back to school full-time. Some of these resources are:

- Coursera: The Data Scientist's Toolbox (https://www.coursera.org/learn/data-scientists-tools)
- DataSF Data Academy (https://datasf.org/academy/)
- edX: Data Science: Inference & Modeling (https://www.edx.org/course/data-science-inference)
  Lynda.com (https://www.lynda.com)
- Udacity: Intro to Data Science (https://www.udacity.com/course/intro-to-data-science--ud359)

> Looking for **nascent data scientists?** Start with those who work with GIS data in the Planning, Community Development or Public Works department.

Hiring a Team

In many cases, scaling up internal data science capabilities means hiring new employees with the right background and skill set. Here are some tips:

- **Look to your local community.** Talk to local businesses that already use data science, and see if you can leverage a staff member who can help vet skills. You may be barred from directly involving someone not in government for direct vetting, but staff from local organizations can help you develop simple assessment tools, or coach you on the right technical interview questions based on your specific needs.
- **Find nearby data science groups or hackathons on Meetup.com, Facebook, or LinkedIn,** especially those with a focus on "data science for the public

good." These events often attract attendees who are learning data science and are looking for a job in the field. Code for America Brigade meetings are often full of people who understand coding and data science, and are excited about working for the public good.

Working With Outside Organizations

There are organizations, universities, and even individuals who can work with local governments and nonprofit organizations to develop data science solutions to problems on a pro-bono basis. These external resources can be a good first step to exploring the possibilities of data science in your organization. Local universities, particularly analytics programs, may be looking for real-world experience for their students through hands-on projects.

If working with external consultants, make sure that you have an infrastructure to maintain what others build. They should help train someone on your team to update any analysis or predictive model that they build, and ensure your team has infrastructure in place to update any analysis conducted. Make sure your consultants are willing to support you, not just with a one-off project, but with skill-building for data science internally.

Here are some questions to ask outside consultants:

1. What languages and libraries will you use? Are they open source? Will I need access to a particular software to view and re-run your code?

2. Can you share the source code you use to do your analysis with me?

3. What computing power do I need to perform similar types of projects?

4. Can you train one of my staff to update this analysis or model for future projects?

Make sure that your consultants use **open source libraries,** not proprietary tools, and that your organization has access to raw data so analyses can be updated. Otherwise, organizations can be "locked out" of their own data and tools, which makes completing projects more expensive and time-consuming.

Our Recommendations

- Look around your organization for data-literate team members who can train to become data scientists. Build skills from within if you can.

- Hire and train for data engineering skills first, before looking for data scientist skill sets. Understanding how to clean, merge, and move data are much more important than machine learning skills when you are just getting started.

- Build a team. A single data scientist working alone will add limited value to your organization.

- If hiring outside help or consultants, always build a pathway for bringing data science work in-house.

Data

There are three primary types of data used for data science in the public sector, each of which involves unique challenges and opportunities:

Administrative Data

In most government agencies, most of the data is operational in nature. This includes, for instance, data from incoming citizen calls, applications for building permits, or social services visits. Data scientists call this "administrative data," which means that it was collected for administrative purposes. These types of data present two challenges: One technical, and one organizational.

Technical: Administrative data is often not formatted in ways that are immediately useful. This type of data is not originally intended for analysis, and may have to undergo a great deal of formatting and cleaning to be useful. This is where a data engineer is a helpful resource.

Organizational: Groups may find implementing the permissions and structures for sharing data across different parts of an organization can be difficult. Departments or agencies are hesitant to share data with other governmental organizations. This happens both because of the high level of scrutiny government agencies receive, and because of fear that the data will be used to misrepresent the organization's work.

Here are some strategies for empowering internal data sharing:

- **Provide a service:** Has the office or agency always wanted to document or clean their data, but lacked the time? Offer to develop a data dictionary, or code for cleaning and merging data, and give it back to the agency for their own use.
- **Ensure data privacy for agencies or departments:** Make it clear that you are not going to create any open data without their consent, that data will only be shared with others on a need-to-know basis, and that analysis will be shared with the department first.
- **Set up internal MOUs:** Set up MOUs (memorandums of understanding) to memorialize the above, or even include other data-sharing initiatives.

What is "ETL"? One term that you will hear data scientists use is "ETL", which means Extract, Transform, Load. ETL means preparing data for analysis, and formatting it so organizations can use it in data science projects. In less-complex situations, this simply means extracting data so it can be analyzed later. For larger data ecosystems, this means setting up an analytic datamart (a structure for analyzing data) which pulls data from many different systems for analysis.

Public Data

Governments and local organizations have access to amazing public data sources for their communities. The Census Bureau provides a wealth of data on demographic, economic, and housing patterns in the U.S. In the past few years, the federal government has taken significant steps toward consolidating public-facing data and making it more discoverable by Federal, State, and local government entities. Here are some excellent starting points that often contain community-specific public data:

- Data.gov (https://www.data.gov)
- IRS Tax Statistics (https://www.irs.gov/statistics)
- HealthData.gov (https://www.healthdata.gov)

There are various other public data sources which can be used to augment and better understand your community.

Commercial Data Sources

Private companies license data from brokers who can provide data on the local business environment, real estate market, or consumer data on businesses and households in your community. Oftentimes, this data is based on information from public sources like the Census Bureau or business licenses. It is then cleaned, merged, and collated to create a more versatile product. Commercial data sources can also give you more insight into your community than public data sources. For example, you can learn more about where renters live, or which households are likely to be pet owners.

Buying commercial data can be expensive, but never assume that you can't afford it. Local government agencies and nonprofits often qualify for preferential pricing. Alternately, many data brokers offer discounts if you are willing to talk publicly about using their services.

New Data Collection

When you want to analyze a particular issue of interest, first look around at what data you have available from administrative, public and commercial sources before going to the expense of collecting new data. However, there will be situations in which you will need data on public opinions or attitudes that you will have to collect directly from citizens through surveys.

Some organizations treat survey data as a one-off data resource, to be used for answering one specific question and then discarded. However, survey data should always be understood as a perennial and valuable resource for benchmarking public opinion over time. Keeping survey data in one place can add to the full picture of your community and how it changes over time.

> **Not everything can be "Open Data."**
> The open data movement has huge benefits to transparency and empowering citizen decisionmaking. However, data does not have to be open to be useful. Some data can also never become public because of confidentiality and privacy concerns. When talking with different offices about data access, make it clear to all parties when data will be kept confidential and used for internal analysis only, and what will ideally be open data.

Our Recommendations

- Before collecting new data, make sure your organization inventories data already available from public sources, from other offices in your organization, or data that can be purchased or licensed from third parties.
- When requesting data from other offices in your organization or from other organizations, offer to provide a service for the data owner. Make it clear that you will protect privacy and set up an MOU on data usage to memorialize the agreement.

Tools

Once you have data to work with and people to analyze it, you need resources to both host that data and enable data scientists to answer questions, find patterns, and develop models that predict the future.

Programming Languages & Libraries

A programming language is similar to a written language — it is a set of syntax and grammar which allows programmers to communicate with computers. A programming library is a set of reusable programs within a particular language that data scientists can use when performing analysis.

For local governments, if possible, encourage your team to work with open-source programming languages. Two of the most commonly used open-source languages in data science are R and Python. Both of these languages have specific "libraries" already developed just for data science tasks, many of which are open-source, and free to use.

Hosting and Computational Resources

Many small data science teams working on the types of applications and data scale available at the local government level start by having a single employee use data extracts downloaded from a production system. They then locally run code on the extracts using a laptop or desktop. On a quality computer, this can work as a basic (although slow) analysis infrastructure for performing routine tasks on a fairly small dataset.

Although this process may be low-cost and convenient for a single data scientist working on their own (which, again, we don't recommend), this is not an appropriate infrastructure for supporting a data science team, even if it consists of a single employee. This type of infrastructure traps your analysis with one person and a single computer.

If that employee happens to leave, or the computer is lost, this endangers organization-wide data science initiatives. There are also additional day-to-day concerns.

An "open source" programming language is developed by a community, and is free to use, share, and modify. Programmers add libraries to the language over time, so that it continues to evolve. This is why many of the most cutting-edge data science libraries are written in open source languages.

Performing quality control or allowing other employees to run analysis is harder when data is stored on a single computer with a single gatekeeper.

Building an Analytics Environment for Data Science

Empowering a data science program of any size requires building a dedicated analytics environment. An analytics environment is commonly called a "sandbox" both because it's a place to play, and a place for performing exploratory analysis without impacting production-level systems. This means creating a non-production

Don't fell into the trap of "I'll just do it on my laptop." Even if your data science team says it's faster and easier for them to run the type of analysis they do on a laptop, this is not a sustainable practice. It leads to data hoarding and sharing of confidential datasets by email and USB drive, which constitutes a serious security risk.

environment for analysts to test out analysis and build models without requiring access to source system data, or being able to "break" anything that is in production.

There are two key decisions to make when establishing an analytics environment:

1. Cloud infrastructure versus on-premises infrastructure

2. Developing your own environment versus buying an existing analytics platform

Cloud vs On-premises

On-premises solutions, usually referred to as "on-prem" solutions, depend on a set of physical IT resources that you own and are physically resident on your organization's premises. On-prem solutions can store millions of records.

Cloud-based solutions are resources managed by a third party such as Amazon Web Services, Microsoft Azure, and Google Cloud, which can be quickly provisioned over the Internet. Although they are not physically based on your organization's premises, they offer fast and reliable service that is often cheaper than an on-prem option. If your organization is working with large datasets, then your organization will likely need to adopt a cloud solution.

Because cloud storage is incredibly cheap compared to on-prem solutions, it makes more budgetary sense for many organizations. In addition, cloud solutions provide automated backups and remove the need for internal IT support and maintenance. This makes them an attractive option to data science teams who may not always have dedicated IT resources.

Because of the more attractive cost, scalability, reliability, and maintenance options, we always suggest that organizations look to the cloud first when deploying analytics environments, even if all other data stores are on-prem.

> Don't think that if you have some data that need to stay on-prem, and that you can't leverage cloud-based solutions for analysis. **Cloud-based solutions can usually be configured to "talk to" on-prem data stores.**

Building or Buying

Another choice set is whether to build your own analytics environment, or leverage software products that provide a ready-made environment. When considering buy vs. build, however, make sure not to underestimate how difficult the build process can be.

Data scientists can generate a set of analytics resources for your organization on the cloud, but a separate set of skills is needed for maintaining or administering clouds and on-prem environments over time. Maintaining cloud or on-prem environments over time requires IT support, additional software development, and operations assistance.

If building a custom analytics environment, make sure your organization has dedicated software engineering developmental operations (commonly called DevOps) and both the budget and capacity for creating an internal system. In addition, your organization needs to ensure the custom environment adheres to expected security standards.

If your team does not have access to those types of resources, it can be helpful to use one of the SaaS products already on the market.

For example, the Civis Platform provides a set of collaboration tools and resources that can help data science teams work together using open source languages and libraries, in a highly secure environment. Because Civis supports the underlying architecture of the platform, data science teams can run analysis at scale without draining scarce information technology resources.

Our Recommendations

- Open-source data science libraries offer the most flexible and cutting-edge solutions for most governmental organizations and nonprofits.
- Buy or build a dedicated data science development environment for your data science team, even if it only consists of one individual.
- Cloud-based resources usually make more financial, logistical, and organizational sense for your data science environment.
- If building your own data science environment, make sure you have significant software engineering resources available, and the ability to manage to common security standards.

Security

In an age of constant hacker attacks and data leaks, security is critical to public sector operations. Every day, citizens trust the government with their precious personal data. Leaking or breaching this data is unacceptable for any organization, and especially for a public sector organization working hard to serve citizens.

Although the rules vary from state to state, most states have data security regulations which cover most government use. Two common private sector security standards often leveraged by the state and local government are SOC (System and Organization Controls) and ISO (International Organization for Standardization). A SOC-certified or ISO-certified organization or software product has been independently audited to ensure that all safeguards and procedures in place.

> **Many of the best security standards are simply common sense.** For example, unless they are strictly necessary for data analysis (and they very rarely are), strip names and social security numbers from all data sets. Share data only on a need-to-know basis, and set access controls to data sets so only the right people see data.

In addition, the type of data your organization works with may require particular standards, which then need to be adopted throughout the whole data program. For example, agencies which come into contact with personal health data may be required to be HIPAA (Health Insurance Portability and Accountability Act) compliant. Organizations working with crime data may need to be compliant with CJIS (Criminal Justice Information Systems) standards.

Talk to your organization's head of information security about what needs to be put in place first, so your organization can choose the best infrastructure and toolset for its needs. Even if your organization does not have any required data security standards, adopting tools based around known security standards is highly recommended to maintain public trust and avoid embarrassing leaks.

Besides adhering to compliance regimes, make sure your team has robust data security training, and that best practice data security measures are put into place. Most security breaches aren't the result of system breaches caused by malicious hackers or disaffected employees. Instead, the bulk of breaches are the result of simple human error and the careless handling of data.

In addition, make sure your data science staff is trained on phishing attacks and physical security, and understands not to store sensitive data on laptops. It is crucial to train, remind, and put policies in place to make sure data scientists understand these concepts. Security is an area where a little bit of advance work can pay off in major dividends over the long term.

Finally, make sure that your data science team is the standard bearer for smart information security standards organization-wide, above and beyond what is required by IT. This instills confidence throughout the organization.

Our Recommendations

- Make security the central focus of your data science team. Plan ahead for the types of data you will analyze now and in the future when considering what security regime to adopt.
- Always ask about the security standards and certifications of the products you are considering.
- Implement common-sense security standards, such as stripping data sets of names and social security numbers, share sensitive data on a need-to-know basis, and avoid storing data on laptops.

Conclusion

All public sector organizations already have access to data that can improve their operations. Launching a data science capability can sound daunting, but you probably already have many analytical skills within your current staff — they just need to be deployed the right way. By formally organizing a data science capability, you can turn those skills into a scalable resource that can help you better deploy scarce resources, and improve the quality of citizen-facing services. By deciding on an organizational structure, and empowering a team with the people, tools, data and security, you can develop a data science capability that can lead to better citizen-facing services.

Let us Help!

Civis has helped public sector organizations across the country build data science competency within their organizations. Find out how Civis can improve your organization's data science efforts and generate top results.

For more information, visit **civisanalytics.com** or email Amy Deora at **adeora@civisanalytics.com**

civisanaytics.com