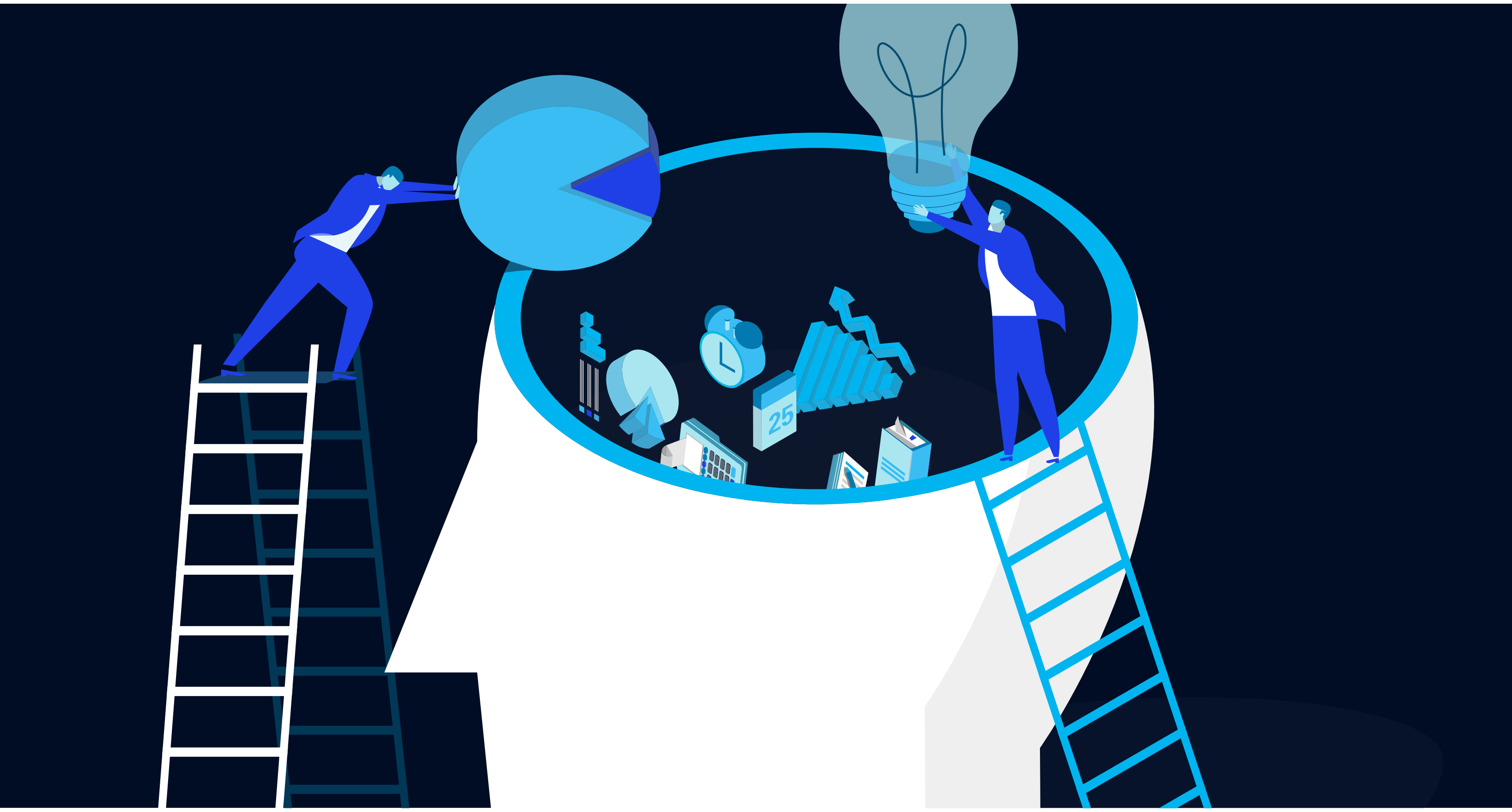
McKinsey
& Company

Public & Social Sector Practice

Accelerating AI impact by taming the data beast

Government agencies seeking to deploy artificial intelligence face hurdles in data awareness, availability, and quality. A five-step, mission-based data strategy can help sidestep these challenges.

by Anusha Dhasarathy, Ankur Ghia, Sian Griffiths, and Rob Wavra



© Z_Wei/Getty Images

March 2020

Artificial intelligence (AI) has the power to dramatically enhance the way public-sector agencies serve their constituents, tackle their most vexing issues, and get the most out of their budgets. Several converging factors are pressuring governments to embrace AI's potential. As citizens become more familiar with the power of AI through digital banking, virtual assistants, and smart e-commerce, they are demanding better outcomes from their governments. Similarly, public servants are pushing for private sector–like solutions to boost on-the-job effectiveness. At the same time, AI technology is maturing rapidly and being incorporated into many offerings, making it increasingly accessible to all organizations.

Most government agencies around the world do not yet have all of the building blocks of successful AI programs—clear vision and strategy, budget, high-quality available data, and talent—in place. Even as AI strategy is formulated, budget secured, and talent attracted, data remains a significant stumbling block. For governments, getting all of an organization's data "AI ready" is difficult, expensive, and time-consuming (see sidebar, "AI-ready data defined"), limiting the impact of AI to pilots and projects within existing silos.

How can governments get past pilots and proofs-of-concept to achieve broader results? To raise the return on AI spending, leading organizations are prioritizing use cases and narrowing their aperture to focus only on improving the data necessary to create an impact with AI. A five-step, mission-driven process can ensure data meets all AI requirements and that every dollar invested generates tangible improvements.

Navigating the data labyrinth

As governments seek to harness the power of AI, one of the first questions that AI programs may need to answer concerns analytical adequacy: Is there data, and is it of sufficient quality to address the specific business need?[1] On the whole, the public sector has more data than private-sector organizations, but it's often in unusable, inconsistent formats. On average, only 3 percent of an organization's data meet the quality standards needed for analytics.[2] And unlike tools, infrastructure, or talent, a complete set of AI-ready data cannot typically be purchased because an agency's unique use cases and mission demand bespoke data inputs.

The most powerful AI solutions often require a cocktail of internal data about constituents, programs, and services as well as external data from other agencies and third parties for enrichment. The core—existing internal agency data—is often in a format and a quality that make it incompatible with AI approaches. A Socrata survey highlighted these challenges:[3]

- Only 45 percent of developers agreed that government data was clean and accurate; the same percent agreed that it was in a usable format for their work
- Less than 35 percent thought it was well documented

In addition, sharing data between agencies often requires an intergovernmental agreement (IGA)—which can take years to secure, even with the most willing counterparties. Within a single national agency, policy restrictions require signed data-sharing agreements and adherence to multiple security standards. State agencies face similar problems with inconsistent confidentiality, privacy requirements, and legal frameworks for sharing data. The result is a hodgepodge of conflicting memorandums of understanding and IGAs.

Locating data and determining ownership can also pose challenges. In many organizations, data have accumulated uncontrollably for years. It's not uncommon for agencies to be unaware of where the data reside, who owns them, and where they came from. As a result, little AI-relevant data is accessible to any given office or "problem owner" in

[1] Oliver Fleming, Tim Fountaine, Nicolaus Henke, and Tamim Saleh, "Ten red flags signaling your analytics program will fail," May 2018, McKinsey.com.
[2] Tadhg Nagle et al., "Only 3% of companies' data meets basic quality standards," *Harvard Business Review,* September 11, 2017, hbr.org.
[3] Developers Rate the Current State of Gov Data Accessibility, Socrata, updated August 23, 2016, benchmarkstudy.socrata.com.

AI-ready data defined

Data that can support artificial intelligence (AI) solutions must meet five criteria:

1. **Known**
   The agency is aware of its available enterprise and local data sources.

2. **Understood**
   Users and leaders are aware of what's in the data set (and what isn't), where it came from (its provenance and lineage), as well as its format, size, and potential to link to other data sets.

3. **Available**
   Data must "live" somewhere that makes it available to users and analysts doing AI work.

4. **Fit for purpose**
   The data are right for the AI goal and of sufficient quality, variety, and scale.

5. **Secure**
   The data are being handled appropriately and are compliant with information security guidelines, confidentiality, relevant civil rights and civil liberties rules, and data privacy regimes (for example, the General Data Protection Regulation).

the organization. According to a McKinsey Global Survey about AI capabilities, only 8 percent of respondents across industries said their AI-relevant data are accessible by systems across the organization.[4] Data-quality issues are compounded by the fact that governments have a multitude of different systems, some of which are obsolete, so aggregating data can be exceedingly difficult. Both state and federal agencies grapple with aging infrastructure: in some instances, the whole stack of hardware, data storage, and applications is still in use—decades after reaching end of life. And annual budget cycles make it difficult to implement long-term fixes.

The scale of the challenge can lead government officials to take a slower, more comprehensive approach to data management. Realizing the importance of data to AI, agencies often focus their initial efforts on integrating and cleaning data, with the goal of creating an AI-ready data pool over hundreds or even thousands of legacy systems. A more effective approach focuses on improving data quality and underlying systems through surgical fixes.

All of these factors make getting data AI ready expensive and time-consuming; the undertaking also demands talent that is not always available in the public sector. It also puts years of IT projects and data cleansing between current citizen needs and the impact of AI-enabled solutions. The number of records needed for effective analytics can range from hundreds to millions (see box, "Number of records needed for effective analytics").

Five steps to AI-ready data

The best way for public-sector agencies to start their AI journey is by defining a mission-based data strategy that focuses resources on feasible use cases with the highest impact, naturally narrowing the number of data sets that need to be made AI ready. In other words, governments can often accelerate their AI efforts by emphasizing impact over perfection.

[4] Michael Chui and Sankalp Malhotra, "AI adoption advances, but foundational barriers remain," November 2018, McKinsey.com.

Box

Number of records needed for effective analytics

Deep learning on images: Millions

Artificial neural networks: Hundreds of thousands to millions

Classifiers: Hundreds to thousands

In addition, while prioritizing use cases, governments should ensure that data sources are available and that the organization is building familiarity and expertise with the most important sources over time.

Proper planning can allow bundling of related use cases—that is, exploiting similar tools and data sets, reducing the time required to implement use cases. By expending resources only on use cases prioritized by mission impact and feasibility, governments can ensure investments are closely tied to direct, tangible mission results and outcomes. These early wins can build support and excitement within agencies for further AI efforts.

Governments can select the appropriate data sets and ensure they meet the AI-ready criteria by following five steps.

1. Build a use case–specific data catalog
The chief data officer, chief information officer, or data domain owner should work with business leaders to identify existing data sets that are related to prioritized use cases, who owns them, in which systems they live, and how one gains access. Data-discovery approaches must be tailored to specific agency realities and architectures. Many successful efforts to build data catalogs for AI include direct collaboration with line- and supervisor-level system users, interviews with technical experts and tenured business staff, and the use of smart or automated data discovery tools to quickly map and categorize agency data.

One federal agency, for example, led a digital assessment of its enterprise data to highlight the most important factors for achieving enhanced operational effectiveness and cost savings. It built a data catalog that allowed data practitioners throughout the agency to find and access available data sets.

2. Evaluate the quality and completeness of data sets
Since the prioritized use cases will require a limited number of data sets, agencies should assess the state of these sources to determine whether they meet a baseline for quality and completeness. At a national customs agency, business leaders and analytics specialists selected priority use cases and then audited the relevant data sets. Moving forward on the first tranche of use cases tapped less than 10 percent of the estimated available data.

In many instances, agencies have a significant opportunity to tailor AI efforts to create impact with available data and then refine this approach over time. A state-level government agency was able to use data that already existed and predictive analytics to generate a performance improvement of 1.5 to 1.8 times. They then used that momentum to pursue cross-agency IGAs, focusing their investments on the data with the highest impact.

3. Aggregate prioritized data sources

Agencies should then consolidate the selected data sources into an existing data lake or a microdata lake (a "puddle")—either on existing infrastructure or a new cloud-based platform put together for this purpose. The data lake should be available to the business, client, analytics staff, and contractors. One large civil engineering organization quickly collected and centralized relevant procurement data from 23 enterprise resource planning systems on a single cloud instance available to all relevant stakeholders.

4. Gauge the data's fit

Next, government agencies must perform a use case–specific assessment about the quantity, content, quality, and joinability of available data. Since such assessments depend on a specific use case's context or problem to be solved, data can't objectively be fit for purpose. For example, data that are highly aggregated or missing certain observations may be insufficiently granular or low quality to inform person-level decision support. However, they may be perfectly suited for community-level predictions. To assess fit, analytics teams must do the following:

- Select available data related to prioritized use cases.
- Develop a reusable data model for the analytic, identifying specific fields and tables needed to inform the model. Notably, approaches that depend on working directly with raw data, exploiting materialized views, or developing custom queries for each feature often do not scale and may result in data inconsistency.
- Systematically assess the quality and completeness of prioritized data (such as error rate and missing fields) to understand gaps and potential opportunities for improvement.
- Bring the best of agile approaches to data development, iteratively enriching the reusable data model and its contents. Where quality is lacking, analytics teams can engineer new features or parameters, incorporating third-party data sets or collecting new data in critical domains.

A state agency decided to build a machine-learning model to help inform the care decisions of a vulnerable population. The model required a wide range of inputs—from demographics to health. Much of this data was of poor quality and in a suboptimal format. The agency conducted a systematic assessment of the required data by digesting paper-based data and made targeted investments to improve data quality and enrich existing data sets. It also generated the analytics model to improve outcomes.

5. Govern and execute

The last step is for agencies to establish a governance framework covering stewardship, security, quality, and metadata. This need not immediately be an exhaustive list of rules, controls, and aspirations for data maturation. Still, it is crucial to define how data sets in different environments will be stewarded by business owners, how their quality will be increased, and how they will be made accessible and usable by other agencies.

Many security governance issues may already be met by keeping data in a compliant environment or accredited container, but agencies still need to pinpoint any rules that remain unaddressed. Finally, they should determine required controls based on standard frameworks—for example, the National Institute of Standards and Technology—and best practices from leading security organizations. One large governmental agency was struggling with the security and sharing requirements for its more than 150 data sources and specialized applications. It did not have agency-level security procedures tailored for such a complex, role-based environment where dozens of combinations of roles and restrictions could exist. To resolve this issue, leaders developed a comprehensive enterprise data strategy with use case–level security requirements, dramatically simplifying the target architecture and application stack. The agency is currently executing a multiyear implementation road map.

These important governance and security responsibilities must be paired with a strong bias toward impact. Most public-sector agencies have found that legacy waterfall-development life cycles and certification and accreditation processes are incompatible with AI projects. Agile approaches to development—from scrum-based methods of leading development efforts to fully mature DevSecOps approaches to continuous delivery—are central to ensuring that process and culture are also AI ready. While this change is often slow, decelerated by risk-averse cultures and long-established policies, it is a critical element in AI success stories.

By adopting a mission-based data strategy, governments can avoid many common roadblocks and immediately focus their technical talent, knowledge, and limited budgets on the subset of data needed for prioritized use cases. This strategy avoids creating data and tool capabilities without a plan. The iterative process—translating mission priorities into requirements and data engineering tasks, generating AI-ready data, and translating data into insights—keeps investments focused and maximizes their impact.

Anusha Dhasarathy is a partner in McKinsey's Chicago office and **Ankur Ghia** is a senior partner in the Washington, DC, office, where **Sian Griffiths** and **Rob Wavra** are associate partners.