About the AI-Ready Data Matrix

Background and Motivation

In February of 2019, the President signed an Executive Order on "Maintaining American Leadership in Artificial Intelligence". Section 5a of that order directs "...all agencies [to] review their federal data and models to identify opportunities to increase access and use by the greater non-federal AI research community in a manner that benefits that community, while protecting safety, security, privacy, and confidentiality. Specifically, agencies shall improve data and model inventory documentation to enable discovery and usability, and shall prioritize improvements to access and quality of AI data and models based on the AI research community's user feedback."

This Executive Order leads federal stewards of open datasets to ask: What are the properties that make a dataset ready to use for Artificial Intelligence (AI) research and development (R&D)? Beyond the broad categories of quality, access, and documentation… what specific aspects of data improvement would benefit the AI research community? What user feedback has been received from this community? How do we decide if a dataset is AI-ready, and how would we improve its readiness?

The OSTP Subcommittee on Open Science reviewed the available input, conducted a survey of federal agencies, and developed a draft matrix to assess individual datasets for AI-readiness. The matrix is intended to help agencies prioritize data improvements, assess the level of effort that's needed, and support requests for additional resources.

The main goal of the improvements would be to provide federal open data that will help spur AI research and development in universities and the private sector. Federal agencies should also see a benefit to opening their data to a new audience for novel applications. Even more importantly, all of the efforts aimed at meeting the advanced needs of the AI community will also strongly benefit our existing user base with higher-quality data that is easier to use.

Information sources

In developing this matrix, the following sources were examined to find requirements for what makes a dataset most suitable for use:

1. In September 2018, the National Science Foundation published a Request for Information (RFI) to obtain public input to update the 2016 National Artificial Intelligence Research and Development Strategic Plan; 46 responses were received from industry, academia, and the private sector. The RFI was not explicitly focused on open data, but many responses did include some details on their data requirements. (https://www.nitrd.gov/nitrdgroups/index.php?title=AI-RFI-Responses-2018)

2. In July 2019, the Office of Management and Budget (OMB) published an RFI to obtain public input on identifying high-priority improvements for federal data and models for AI; 28 responses were received. These RFI responses contained a wealth of information about data requirements. *(responses not public yet, will include a link)*
3. An internal OMB study was conducted among federal agencies to identify known barriers to use of federal data in AI R&D. The results and initial AI Inventory Guidance were published to code.gov in August 2019.
4. To augment the RFI responses and OMB guidance, the Subcommittee on Open Science conducted a survey within the 16 departments and agencies represented on the Subcommittee. We asked data stewards and Chief Data Officers (CDOs) about their knowledge of what's required for AI R&D and the barriers that exist. We also asked internal federal AI researchers about their data requirements. It is clear from the results that federal data stewards do not feel they understand the AI research community's requirements, and there is a mismatch between organizational data management strategies and what AI researchers actually need. Detailed results are in **Appendix B**.
5. The Department of Energy's Office of Science conducted a "Data for AI" roundtable of AI, data, and IT infrastructure experts from their own labs as well as NIH and NSF (June 2019). The goal was to identify key challenges/opportunities and next steps for the Office of Science. Draft results are available in these slides, full report coming soon: https://science.osti.gov/-/media/ascr/ascac/pdf/meetings/201909/BIVEN_DataForAI_ASCAC_20190923.pdf

About the Readiness Matrix

Factors

From the above sources, we found twelve factors of data that were most important to AI researchers. The numbers in parentheses refer to the information sources listed above to show where each requirement came from. Each factor is a column in the readiness matrix.

- Quality
  - Completeness (**2, 4**): the breadth of a dataset compared to an ideal 100% completion (spatial, temporal, demographic, etc.); important in avoiding bias
  - Consistency (**1, 2, 4**): uniformity within the entire dataset or compared with similar data collections; for example, no changes in units or data types over time; item measured against itself or its counterpart in another dataset or database
  - Lack of Bias (**1, 2**): avoiding a systematic tilt in the dataset, caused for example by instrumentation, incorrect data processing, unrepresentative sampling, or human error; the exact nature of bias and how it is measured will vary depending on the type of data and the research domain
  - Timeliness (**4, 5**): the speed of data release, compared to when an event occurred or measurements were made; requirements will vary depending on the timeframe of the phenomenon (e.g., severe thunderstorms vs. climate change, or disease outbreaks vs. life expectancy trends)

    - Provenance & Integrity (**1, 2, 3**): identification of the data source, how it was processed, who released it, and whether it remains unchanged from the original
- Access
    - Formats (**1, 2, 4**): standards that govern how information is stored in a computer file (e.g., CSV, JSON, GeoTIFF, , etc.); different AI user communities will have different requirements, so best practice is to provide several format options to meet the needs of multiple high priority user communities.
    - Delivery Options (**1, 2, 4, 5**): mechanisms for publishing open data for public use (e.g., direct file download, Application Programming Interface (API), cloud services, etc.); different AI user communities will have different requirements, so best practice is to provide several delivery options to meet the needs of multiple high priority user communities.
    - Usage Rights (**1, 2, 3**): information on who is allowed to use the data and for what purposes, including data sharing agreements, fees, etc.; some federal data needs to have restrictions and some will be fully open, so rights should be documented in detail
    - Security / Privacy (**1, 2, 3, 4**): protection of data that is restricted in some way (privacy, proprietary / business information, national security, etc.)
- Documentation
    - Dataset Metadata (**1, 2, 3, 4, 5**): complete information about the dataset: quality, provenance, location, time period, responsible parties, purpose, etc.
    - Data Dictionary (**1, 2**): (aka Codebook) complete information about the individual variables/measures/parameters within a dataset: type, units, null value, etc.
    - Identifier (**2, 3**): a code or number that uniquely identifies a dataset

Several other dataset factors were identified but are detailed suggestions rather than readiness indicators.

- Metadata must include keywords that tag the dataset as AI-ready, e.g., “usg-artificial-intelligence” or “usg-ai-training-data” (**3**).
- Open data must meet agency Information Quality Act (IQA) guidelines (**3**).
- Data should be submitted to a federal data archive for long-term preservation (**2, 5**).
- Agencies should provide subject-matter experts to help interpret their open data, with contact information provided in the metadata (**4, 5**).
- Merging or aggregating related datasets (e.g., over years or multiple data sources) makes data easier to use and can reduce bias (**1, 2, 5**).
- To improve timeliness and remove analysis bottlenecks, reduce data movement during processing by optimizing storage and memory hierarchies (**5**).

Levels

- Level 0 - Not AI-Ready: The dataset meets basic requirements for Open Data under current federal policy (OPEN Government Data Act, Federal Data Strategy, Public Access to Research Results, etc.), but does not specifically facilitate AI/ML.

- Level 1 – Minimal: The dataset is curated in a standardized and consistent process with manual quality checks. The dataset is well formatted for machine readability and includes robust machine-readable metadata and a data dictionary.
- Level 2 – Intermediate: The dataset is curated in a standardized and consistent process adhering to a documented data standard and subject to in-depth data quality review. The dataset is provided on-demand through one or more options such as an API or downloadable file.
- Level 3 – Optimal: Data collection, curation, and delivery pipeline ensures documentation of versioning, verification of provenance, data integrity against benign or malicious changes, and sanitization of sensitive information. Data has robust machine-readable metadata, license, and data dictionary.

Structural Considerations

The information sources also highlighted several infrastructure factors that are necessary to enable the work of producing and disseminating AI-ready data. These criteria don't apply to individual datasets, but rather to the open data culture and infrastructure within an agency.

- Agencies should work with stakeholders to identify high-impact "benchmark" data, then prioritize those data for investment in curation and preservation (**5**).
- Federal departments and agencies must ensure that they have sufficient staff expertise and funding to support AI-ready data. Hiring and retaining skilled staff is especially challenging in in-demand fields like AI and data engineering (**4, 5**).
- There should be a mechanism for users to submit feedback on datasets, leading to improved future versions (**3**).
- Agencies should also investigate the use of AI techniques (e.g., natural language processing) to automate data cataloging and data curation (**2, 5**).
- Incentives or credit are needed to encourage data sharing-- both for the public sector and the private sector (**2, 5**).
- Federal agencies must have a contract vehicle for use of cloud storage and utilities, which is currently a significant barrier in some agencies. In addition, hiring and retaining staff with cloud expertise is very challenging for federal agencies (**4**).
- Frameworks for tracking relationships between data, models, and tasks are needed to improve reproducibility and streamline progress in AI R&D (**5**).
- The FAIR principles for making data machine-actionable are good guidelines but questions remain about specific needs of AI R&D audience (**5**).
- Federal agencies must mature data repositories, metadata, and data cataloging systems with increased automation and consolidated enterprise data management solutions. These recommendations are beyond initial Open Data requirements set prior to the Foundations for Evidence-Based Policymaking Act of 2018, Title II(a). Communities in the AI/ML domains are among the most advanced group of end users with high-demand requirements from data producers, especially for data quality, machine-readability, and access to large datasets (**4, 5**).

	Quality				
	Completeness	**Consistency**	**Lack of Bias**	**Timeliness**	**Provenance & Integrity**
Level 0 Not AI-Ready	no formal agency effort to ensure dataset completeness before publication	no formal effort to ensure internal consistency before data are published	little to no information on potential bias	long (1yr or more) time lapse between collection/ creation and publication	little to no information on provenance
Level 1 Minimal	manual checks for completeness	manual checks for consistency	metadata contains some info on potential bias, including how data were collected, cleaned, and processed	lag in availability of near-real-time data, long delay for fully quality-controlled data	metadata contains some information on data collection, sources, and processing steps
Level 2 Intermediate	some completeness checks are automated, some documentation of results, and some explanations for missing values	some consistency checks are automated, some documentation of results	some efforts to reduce bias; metadata contains full info on potential bias	minimal lag in availability of near-real-time data, some lag in fully quality-controlled data	metadata contains full information on data collection, sources, and processing steps
Level 3 Optimal	fully-automated completeness checks and reporting, metadata includes explanation for all missing values; some effort to ensure completeness of data in the context of broader community efforts	fully-automated internal consistency checks and reporting; some consideration for external consistency among community datasets	automated bias measurement and correction; metadata contains full info on potential bias	minimal lag in availability of fully quality-controlled data	provenance tracking is automated and included as metadata, plus blockchain/checksum/SHA to ensure data integrity

...

	Access			
	Formats	**Delivery Options**	**Usage Rights**	**Security / Privacy**
Level 0 Not AI-Ready	available (or could be made available) in an open format	open for public use only by request or via an ordering system	meets Federal "open by default" standard, but metadata does not include information on data license	sensitive data not accessible for external use; no aggregated or anonymized version available
Level 1 Minimal	open, machine-readable format is provided as a standard practice	one non-programmatic access option only, such as file download	metadata includes free-text information on data license usage rights	open access to aggregated version at a reduced granularity on request; sensitive data not accessible for external use
Level 2 Intermediate	multiple open formats are provided as a standard practice	multiple delivery options including at least one programmatically accessible method, such as bulk file download plus API or cloud	use of a standard license (e.g. Creative Commons, CDLA, ISO/IEC 19944); metadata includes specific information on data license and usage	open access to anonymized / de-identified data at original granularity as standard agency practice; secure data not accessible for external use
Level 3 Optimal	multiple open formats are provided as a standard practice, including High Performance Computing or cloud-optimized formats	multiple delivery options (download, API, cloud, HPC, data-as-a-service, etc.)	use of a standard license (e.g. Creative Commons, CDLA, ISO/IEC 19944); metadata includes machine-readable information on data license and usage (e.g. Creative Commons Rights Expression Language) or a URL to the full license	open access to desensitized data; secure / encrypted, tiered, permission-based access to sensitive data

...

	Documentation		
	Dataset Metadata	**Data Dictionary**	**Identifier**
Level 0 Not AI-Ready	dataset or collection-level metadata, authored by hand, meets the Federal minimum standards for discoverability	no data dictionary available, or in non machine-readable format (e.g. pdf)	ID internally unique
Level 1 Minimal	dataset or collection-level metadata uses a machine-readable metadata standard, some fields are completed to help with data re-use, and metadata is partially automated to create a more accurate record	data dictionary in machine-readable format (e.g. csv, xml, json)	ID internally unique, plus resolvable or persistent
Level 2 Intermediate	robust metadata meets a defined agency standard for completeness; completion is mostly automated to create an accurate record; may reference controlled vocabulary for keywords	data dictionary uses a machine-readable metadata standard	ID internally unique, resolvable, and persistent
Level 3 Optimal	robust metadata meets a defined agency standard for completeness; completion is automated to record lineage, updates or corrections to data, and enforce standardized keyword vocabulary	data dictionary uses a machine-readable metadata standard; harmonized with other agency datasets, across Federal agencies, or domain standards	ID globally unique, persistent, and resolvable; IDs are structured to handle data versioning

Future Plans

This AI readiness matrix is meant to be a version 1.0 that will improve over time. Over the next year, the Subcommittee on Open Science will socialize the draft matrix and get feedback from external partners who use their open data for AI R&D. Some possible venues include AI conferences or domain-specific conferences with AI sessions. In addition, federal agencies should start applying the matrix to federal agency datasets and give feedback. We envision that agencies could also start to use this assessment as a tool to identify data that is close to being AI-Ready and the level of effort needed to make improvements. It is crucial for the Subcommittee to review this resource often as we refine the requirements and as AI technology rapidly evolves.

Appendix A: Definitions

Artificial Intelligence (AI) is here defined broadly, to encompass a machine-based system designed to make decisions without human intervention. Further explanation can be found, for instance, in the John S. McCain National Defense Authorization Act for Fiscal Year 2019, sec 238(g): "... the term "artificial intelligence" includes the following: (1) Any artificial system that performs tasks under varying and unpredictable circumstances without significant human oversight, or that can learn from experience and improve performance when exposed to data sets. (2) An artificial system developed in computer software, physical hardware, or other context that solves tasks requiring human-like perception, cognition, planning, learning, communication, or physical action. (3) An artificial system designed to think or act like a human, including cognitive architectures and neural networks. (4) A set of techniques, including machine learning, that is designed to approximate a cognitive task. (5) An artificial system designed to act rationally, including an intelligent software agent or embodied robot that achieves goals using perception, planning, reasoning, learning, communicating, decision making, and acting."

Open Data is government data that is available to the public in open formats, free of charge, and without restrictions on its use. Further explanation can be found, for instance, in the Foundations for Evidence-Based Policymaking Act of 2018, Title II(a): "... the term 'open Government data asset' means a public data asset that is (A) machine-readable; (B) available (or could be made available) in an open format; (C) not encumbered by restrictions, other than intellectual property rights ... that would impede the use or reuse of such asset; and (D) based on an underlying open standard that is maintained by a standards organization".

Appendix B: Results from the federal survey

Survey Summary

The purpose of this June/July 2019 survey was to gather information about agency readiness in provisioning data in a form suitable for use in Artificial Intelligence, and use the information to formulate a maturity matrix. Data in a form suitable for AI use we call "analytics-ready data", a concept that is tightly connected to Open Data policies, Cloud First government initiatives, and American leadership in Artificial Intelligence. The survey assessed federal agency progress toward making available high-priority datasets in a form that are “AI-ready” and in the use and creation of new, integrated datasets serving societal benefits. Fifty-three (53) participants responded to the AI survey; the results will not comprehensively represent the entire community, but are sufficient to derive general principles for the first iteration of the maturity matrix. It should also be noted that response rates to each question were between 50 and 75%, this could indicate that AI is still an emerging field in government and many were unable to provide responses. 17 unique agencies/bureaus are represented and another 18 respondents chose not to report their affiliation. The survey results will act as a baseline assessment of the current state of government for informing the content and its granularity for the maturity matrix.

Detailed Survey Results:

A PDF of the raw survey results are available on the USGS Wiki. Note that there were two audiences for this survey: Chief Data Officers (CDOs) and data stewards, and federal scientists engaged in AI research. Based on the role that respondents chose in question 1, they were presented with different questions for the rest of the survey.

Q1. I am responding to this questionnaire as *(select one)*:

- Chief Data Officer or functional equivalent (21%)
- Data Steward (41%)
- Data Scientist (including AI) / Research Scientist (38%)

Audience: All / Answered: 53

The distribution of responses seems representative to the proportion of actual career positions. We would expect fewer CDOs (11) and more Data Stewards (22) and Data Scientists (20).

-- QUESTIONS 2-11 WERE ASKED OF CDOs AND DATA STEWARDS ONLY --

Q2. The groups at my agency for whom open data to support AI is most relevant are *(select all that apply)*:

- Data Scientists (including AI/ML): 84%
- Research Scientists who use AI/ML as a tool: 80%
- IT Operations: 20%
- Business Analysts: 24%

- Data Management Programs: 52%
- Other: 8%

Audience: CDO and Data Steward / Answered: 25 out of 33

The most popular responses by far were Data Scientists and Research Scientists indicating that the focus for most agencies is still in R&D and less so on developing Open Data programs. Conversely, IT and Business Analysts ranked low with Data Management Programs ranking in the middle.

Q3. How well do you understand what's required for your agency's open data to be considered analytics-ready to support AI? *(select one)*

- Not very well. I am not familiar with that user community and the requirements are not clear: 36%
- Somewhat. My agency provides some data for AI users and we have some feedback on their requirements: 52%
- Very well. We provide at least some open data that is specifically analytics-ready for the external AI community and we have a clear understanding of their requirements: 12%

Audience: CDO and Data Steward / Answered: 25 out of 33

The responses to this question support the idea that many who replied to this survey do not have a clear understanding of AI researchers' data needs. Only 3 respondents (12%) had a good understanding of AI requirements.

Q4. Which of these statements best describes your agency's data delivery systems for open data to support AI? "Programmatically accessed" is defined as machine-to-machine interaction (e.g., Application Program Interfaces (APIs) or web/data services). *(select one)*

- Data is stored in siloed systems. Data are frequently copied or downloaded for individual use: 16%
- Data is stored in siloed systems. Some data can be programmatically accessed: 16%
- Some common data systems and standards. Some data can be programmatically accessed: 28%
- Some common data systems and standards. High-value data can be programmatically accessed. Some common tools exist to facilitate use: 24%
- Core common data systems with well-documented standards. High-value data can be programmatically accessed. Common tools are in use across agency and are available externally to facilitate use: 16%

Audience: CDO and Data Steward / Answered: 25 out of 33

Responses represented a fairly evenly distributed bell curve between low maturity (siloed systems) and high maturity (well-developed core data systems with programmatically accessed data). This seems to be representative of the state of federal agencies.

Q5. Which of the following are most needed to improve in your agency for ideal open data delivery to support AI? *(select all that apply)*

- Timeliness: 33%
- Completeness: 42%
- Consistency: 71%
- Accuracy: 38%
- Usefulness: 42%
- Accessibility: 63%

Audience: CDO and Data Steward / Answered: 24 out of 33

Consistency and Accessibility stood out as the highest ranking responses. Consistency is related to delivering the same type of data in the same format repeatedly and the same types of data being in the same format (having the same definition) across data sets, databases, or records. Accessibility is an Open Data concept related to how end users obtain or can interact with data in the context of what is most suitable for their use cases. These two characteristics point out the common pain points of all Data Scientists including AI/ML. A majority of their effort is spent obtaining, formatting, and cleaning data sets to be combined and processed with AI/ML techniques because of a lack of consistency and accessibility from data providers.

Q6. Briefly describe a challenge related to data stewardship (e.g., delivery, access, responsible use, protection of sensitive data) that you feel is a barrier or critical to the success of open data to support AI. *(free text)*

Audience: CDO and Data Steward / Answered: 23 out of 33

This free text question was grouped into topics based on mentions so the total count of mentions is greater than the number of answers. The challenges that elicited the most mentions are two topics: i) data not usable for AI (9 mentions) and ii) data sensitivity issues (7 mentions). Data not usable for AI includes weak metadata, inconsistent formats, and poor quality data. Sensitivity issues cover responsible use, privacy (the General Data Protection Regulation), and lack of guidance for use of sensitive data in AI. The third most cited topic is (raised by 3 respondents) is inadequate cyberinfrastructure where there is a lack of federation mechanisms across agencies, siloed datasets, or insufficient infrastructure to handle the anticipated size of data. Infrequent mentions (1 mention each) were made of missing business need, weak incentives, FAIR metrics, over-regulation in federal space, lack of skilled workforce, and lack of adequate financial support.

Q7. If applicable, how is your agency's enterprise Cloud vendor environment used to support open data for AI? *(select all that apply)*

- Storage: 60%
- High-Performance Computing: 35%
- Data as a Service (i.e., Applications or APIs to support delivery of specific datasets): 55%
- Data Delivery Platform (e.g., data warehouse, data lake, data hub strategies): 50%

- Other: 25%

Audience: CDO and Data Steward / Answered: 20 out of 33

There was a near-even response between Storage, Data/Software as a Service, and Data Delivery Platform, with Storage being the most popular. This seems representative as most agencies exploring cloud start with Storage and mature through Data/ Software as a service into a Data Delivery Platform. However, other free text responses indicated that some agencies are still struggling to move to cloud operations and are only investigating or not using cloud. It should also be noted that High-Performance Computing (HPC, also known as Platform as a Service) ranked relatively low, so there is an opportunity across government to increase use of cloud for HPC, a key component of most data science and AI/ML activities.

Q8. Briefly describe a challenge related to computing infrastructure (e.g., cloud adoption, data storage, computational bandwidth, etc.) that you feel is a barrier or critical to success of open data to support AI. *(free text)*

Audience: CDO and Data Steward / Answered: 18 out of 33

This free text question was grouped into topics based on mentions so the total count of mentions is greater than the number of answers. The challenges around use of cloud resources elicited considerable commonality across the 18 respondents. The most mentions were to high barriers to entry to cloud use. This includes insufficient in-house expertise and regulatory, technical, or funding barriers that slow adoption. Two other topics received the concern of the respondents: limits to local data storage and bandwidth received 5 mentions, while cloud charges received 3 mentions. Concerns about cloud charges include lack of price transparency in cloud pricing especially over time, and an inability to predict usage (such as egress costs) being some of the higher costs in using clouds. This creates an uncertainty that some respondents found problematic. A couple of respondents thought the internal required cloud provider offerings were lacking either because they were behind the technology curve or because they prohibited use of tools that are needed for AI research.

Q9. Which of the following statements best describes your agency's open data culture? We understand that some parts of the agency might be at different stages, but come up with an overall assessment. *(select one)*

- Uncoordinated and ad-hoc. Quality and interoperability issues limit usefulness for providing data that's ready for AI and analytics: 21%
- Data use is by request (i.e., email or file download). Agency-wide data programs are nascent: 29%
- Some data and analytics are routine and have programs supporting key assets: 38%
- High demand for data across agency. Decision-making is driven by data that is ready for AI and analytics: 8%
- High demand for data internally and for external agency partners. AI and/or analytics-ready data is available for all stakeholders to drive decision-making as a community: 4%

Audience: CDO and Data Steward / Answered: 24 out of 33

Culture is developing in the same fashion as data delivery systems (Question 4). This makes sense because the two typically are interrelated, one creating a need for the other. Only 3 respondents felt their agencies had a mature open data culture that supported a high demand for data internally or for external agency partnerships.

Q10. Which of the following skill sets are most needed in your agency to support open data for AI? *(select all that apply)*

- Data Science: 61%
- AI / ML Research and Development: 52%
- Business Intelligence: 22%
- Data Visualization: 43%
- Software Development: 43%
- Data Stewardship: 43%
- Data Engineers / Data Architects: 52%
- High Performance Computing: 39%
- Cloud Computing: 48%
- Software DevOps: 35%
- DataOps / Data as a Service Teams: 30%
- Other: 9%

Audience: CDO and Data Steward / Answered: 23 out of 33

Response rates were consistently high across multiple categories indicating that skills sets across the board are lagging and significant effort is needed to update the workforce. The lowest ranking skill sets were Business Intelligence, High Performance Computing, Software DevOps, and DataOps / Data as a Service Teams which could indicate a lack of understanding about these skill sets or a lack of maturity in Open Data delivery platforms to take advantage of these skill sets. It is appropriate to characterize these skill sets as "emerging".

Q11. Briefly describe a challenge related to organizational culture (e.g., business or research-based practices, skill sets, organizational structure, etc.) that you feel is a barrier or critical to the success of open data to support AI. *(free text)*

Audience: CDO and Data Steward / Answered: 22 out of 33

This free text question was grouped into topics based on mentions so the total count of mentions is greater than the number of answers. The challenges related to organizational culture raised several issues. One is of expertise: inadequate skill sets of existing employees, difficulty attracting and retaining talent (5 mentions). Another is of organizational structural barriers with contract vehicles that are inflexible for R&D prototyping, prior approvals for data access, weakly curated data, and funding aligned with mission systems. The final mentions are around employee culture that prefers holding data rather than making it accessible.

-- QUESTIONS 12-18 WERE ASKED OF DATA SCIENTISTS ONLY --

Q12. Which of the following are the most important data source characteristics to support your AI-related work? *(select all that apply)*

- Timeliness (i.e., near real-time or general speed of delivery): 67%
- Completeness: 75%
- Consistency: 100%
- Accuracy: 83%
- Usefulness: 67%
- Accessibility: 83%
- Other: 25%

Audience: Data Scientist / Answered: 12 out of 20

Again, Consistency and Accessibility were among the highest ranking among Data Scientists. This is in agreement with Question 5 which represented the CDO and Data Steward point of view. In addition, Accuracy and Completeness were highly ranked. Timeliness and Usefulness were the lowest ranked, but still commonly selected. Overall this indicates the key importance of data quality for AI-ready federal data.

Q13. What are the most important characteristics that make data easier to use for your AI-related work? *(free text)*

Audience: Data Scientist / Answered: 11 out of 20

This free text question was grouped into topics based on mentions so the total count of mentions is greater than the number of answers. The most common answer was easy, free data access (5 mentions), but complete documentation was also mentioned often (4 mentions). Three aspects tied for third place (three mentions): labeled training data, delivery modes, and data quality. Two respondents also mentioned file format, with one mention each for timeliness and access to subject matter experts to help users interpret the open data.

Q14. What file formats or standards are needed to support your AI-related work? *(free text)*

Audience: Data Scientist / Answered: 11 out of 20

The answers to this question were extremely diverse. Many respondents emphasized that they are able to use many different standard formats in their work, but the preference is clearly for simple formats (text, XML, JSON, etc.). Specific formats that were mentioned, in order of frequency: csv, JSON/GeoJSON, netCDF, ASCII, and dbf. Each of these formats also received one mention: XML, TFRecord, TIFF, GRIB, HDF, Zarr, and Open Data Metadata Standard v1.1.

Q15. Which data delivery option do you prefer the most to support your AI-related work? *(select one)*

- File download to local environment: 25%

- Data as a Service (i.e., web services / APIs to support data delivery): 33%
- Platform as a Service (i.e., high-performance computing co-located with the data resources): 25%
- Commercial cloud vendors' proprietary platforms (e.g., Google's BigQuery, Amazon EC2/Athena, Microsoft Data Science environment, etc.): 17%
- Other: 0%

Audience: Data Scientist / Answered: 12 out of 20

APIs/Data as a Service was ranked highest, indicating that having data programmatically accessible is a definite preference. File download to local environment and Platform as a Service (HPC) were the next highest ranked, indicating that some data scientists prefer to have the data in an accessible computing environment. An additional 17% of respondents preferred to access the data in commercial cloud platforms. The preferences here were fairly evenly split, indicating that the best practice should be for data providers to offer several different ways to access the data. Free-text answers to other questions support the idea that their preference depends on the size of the dataset, the specific use case, and the tools being used. In comparison to Questions 7 and 10, CDOs and Data Stewards may not recognize the importance of HPC, since HPC was among the lowest ranking options.

Q16. What are the important tools (e.g., software packages, programming languages, Services, etc.) that you use in your AI work? *(free text)*

Audience: Data Scientist / Answered: 12 out of 20

For this audience of federal AI researchers, the list of tools was weighted towards programming languages (as opposed to proprietary software), which may not reflect the broader commercial and academic user community. The specific tools mentioned, in order of frequency: Python, R, TensorFlow, Keras, Jupityr notebooks, and the AWS AI/ML/IoT stack. Each of these tools also received one mention: ArcGIS, Word2Vec, BERT, Excel, SAS, IDL, FORTRAN, GitHub, VIAME, Conda, Spark, Scikit-learn, MatLab, Watson Explorer/Studio, Azure AI/ML/IoT stack, and Google Cloud Platform AI/ML/IoT stack.

Q17. Briefly describe a challenge related to data delivery and use (e.g., data quality, documentation, cloud availability, dataset size, computational bandwidth, etc.) that you feel is a barrier or critical to the success of open data to support AI. *(free text)*

Audience: Data Scientist / Answered: 11 out of 20

This free text question was grouped into topics based on mentions so the total count of mentions is greater than the number of answers. The most common challenges mentioned were data quality and the size of the datasets (3 mentions each). Several respondents also mentioned timeliness, documentation, access to cloud resources, and developing labeled training datasets (2 mentions each).

Q18. Which of the following skill sets are most valuable for your AI-related work? *(select all that apply)*

- Data Science: 83%
- AI / ML Research and Development: 75%
- Business Intelligence: 8%
- Data Visualization: 50%
- Software Development: 50%
- Data Stewardship: 33%
- Data Engineers / Data Architects: 50%
- High-Performance Computing: 58%
- Cloud Computing: 17%
- Software DevOps: 25%
- Data Ops / Data as a Service Teams: 8%
- Other: 17%

Audience: Data Scientist / Answered: 12 out of 20

It is no surprise that Data Science and AI / ML research and development were the highest ranking responses due to their direct relationship of this question for the survey audience. The next highest ranking group was Data Visualization, Software Development, Data Engineers / Data Architects, and High Performance Computing. In comparison to the CDO and Data Steward responses in previous questions where these choices were ranked lowest, these areas could represent gaps in understanding about data science/ AI/ ML at the enterprise/ upper management Open Data strategy level.

Q19. Please tell us the name of your agency, subordinate bureau, or office. *(optional)*

- DHHS / CDC (x2)
- DHHS NIH (x2)
- VA / Veterans Health Administration
- VA / Cooperative Studies Program
- USDA
- USDA / Agriculture Research Service (x3)
- USDA / Economic Research Service
- USDA / Forest Service (x3)
- DOS / USAID Bureau for Management (x2)
- DOI (x2)
- DOI / BOEM (x3)
- DOI USGS (x3)
- DOC / NIST (x2)
- DOC / NOAA (x4)
- DOE Office of Science (x3)
- Smithsonian
- NASA

Major Takeaways from Survey:

1. There are some clear themes in the requirements: several aspects of data quality, good documentation, and offering a variety of data formats and delivery mechanisms
2. A discussion about data sensitivity needs to be included, but will not be included as an element in the matrices
3. CDO/ Data Steward data strategies and visions do not match with AI/ ML Data Scientist user needs
4. While Open Data activities are maturing across the federal government, AI/ ML requirements are largely not understood
5. There are still huge barriers to cloud on-boarding and limited understanding about cloud-optimized strategies
6. Definite across the board lack of employee skill sets and inability to retain highly sought after talent