AI and Official Statistics: Responsibly Leveraging Large Language Models in Support of Open Data

August 8, 2024

The views expressed in this presentation are those of the presenters and not the Census Bureau.

Sallie Ann Keller, Ph.D., Chief Scientist and Associate Director

Kenneth Haase, Ph.D., Senior Computer Scientist for Artificial Intelligence Applications

Michael B. Hawes, Senior Survey Statistician for Scientific Communication

Research and Methodology Directorate
U.S. Census Bureau

United States® Census Bureau

Commerce Content and Generative AI—Anticipating Potential while Mitigating Risks

How can we promote responsible AI in society?

How do we leverage Generative AI responsibly within government?

How can government leverage AI to democratize access to data?

United States Census Bureau

History is likely to repeat itself

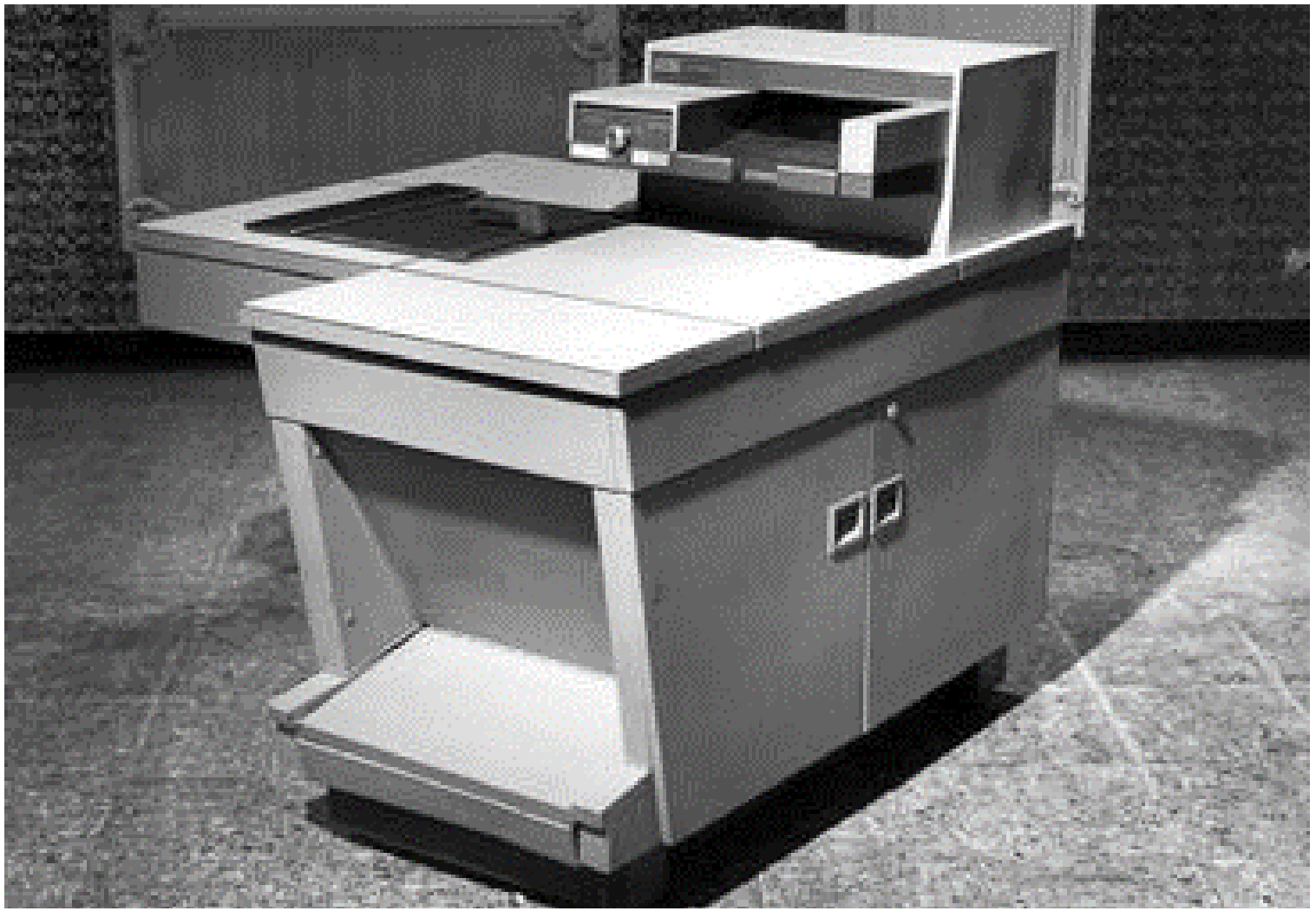
Technology emerges

Public expectations change

Policy follows

United States® Census Bureau

Xerox Photocopier Led to FOIA Laws



APIs Led to Open Data Executive Order and Data.gov

United States® Census Bureau

Open Data in the Pandemic

Unprecedented appetite for data, catalyzed during pandemic

The shift in the public data landscape

20th Century	21st Century
Federal government was dominant user	**Many diverse users**
Statistical system was a near monopoly	**Many more organizations that produce similar statistical products**
Output was mostly cross-tabs	**Output varied and complex featuring visualizations and analysis tools**
Published in books and deposited in libraries, then electronically largely in book formats	**Data accessed electronically online or in secure enclaves**
Source data acquisition was difficult and costly	**Source data more abundant, structured and unstructured formats, available and less costly**
Privacy and confidentiality risks were small	**Privacy and confidentiality risks are much greater**
Computation was expensive and limited	**Computation vastly improved**

The Expectations for Open Data

United States® Census Bureau

How can Commerce leverage AI to democratize access to data?

Democratizing access to public data through generative AI

Emergence of AI technologies has the potential to provide improved information and data access to users, from novice to expert.

Generative AI applications digest disparate sources of text, images, audio, video, and other types of information to produce new content and interpretations for users.

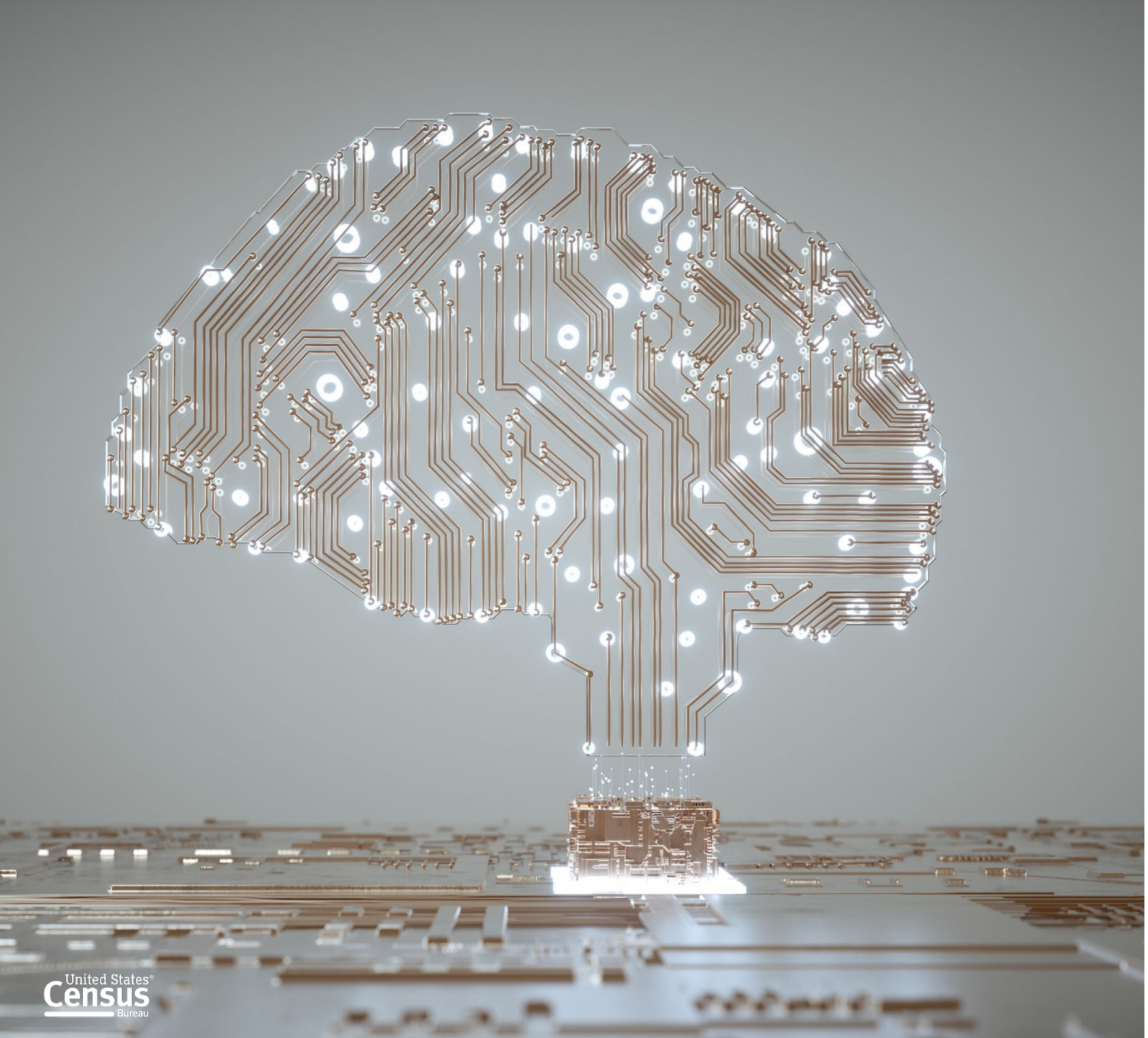
Generative AI and other AI technologies present opportunities and challenges for data providers and data users—including government entities, industry, academia, and the public at large.

United States Census Bureau

We are not AI-ready until our data are AI-ready

The public will expect a natural language interface to accomplish many (most) tasks—especially data analysis.

United States Census Bureau



We are here

Generative AI systems—powered by very large numeric models—are fluent in answering questions, generating explanations, and performing a range of creative tasks.

BUT they struggle with hallucinations, biases, factual errors, and fabrications.

Which is not surprising because the training of LLMs is about fluency and flexibility, not *facts or precision*.

Adapting to Today’s AI Systems

United States® Census Bureau

AI-readiness requires a shift in how government thinks about data publishing standards

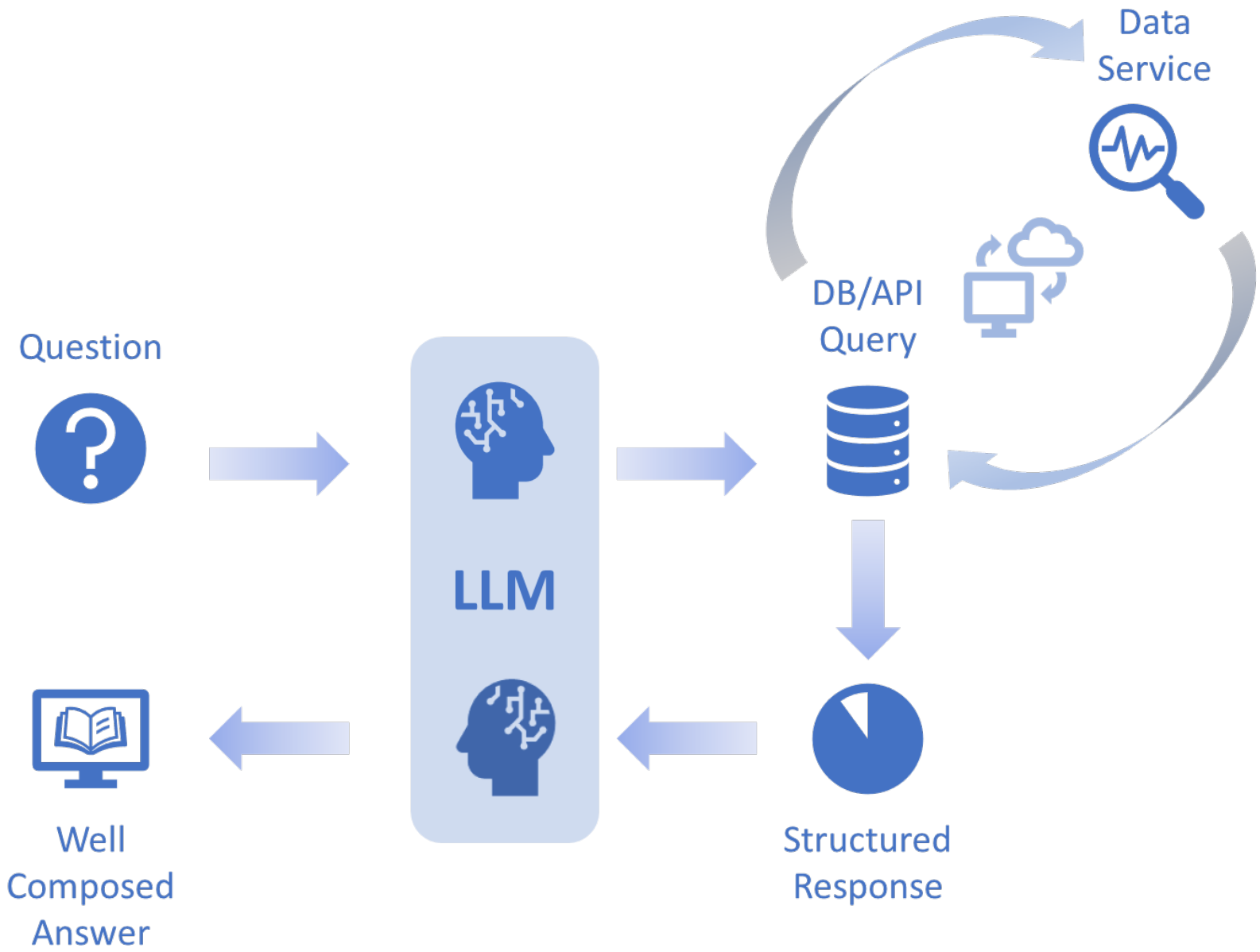
From machine readable

- Can be automatically processed by a computer
- Common formats (.csv, JSON, HTML)

To machine understandable

- Structured data
- Standardized schemas with enriched metadata
- Semantics expressed in a knowledge graph

Hybrid Solutions



United States Census Bureau

Why LLMs and APIs?

LLM-powered Generative AI works by regurgitating and composing variations of the content on which they were trained

This creative process is very different from the responsible use of data in generating abstractions, predictions, or models

Responsible use of published data must articulate both context (like sampling, geography) and relations of the data (like correlation)

Generative systems must be able to use "semantic" metadata to both license and constrain inference, abstraction, or (potentially) explanation

A promising solution is to embrace data formats which capture that rich metadata and combine it with APIs

Generative AI can use these APIs by actively constructing queries and explaining results

Our Assumptions for Achieving Machine Understandable Data

- Create knowledge graphs for variable level metadata, allowing systems to better link human terms to data elements
- Embrace standardized open and extensible ontologies such as schema.org or NIEM, to jumpstart annotation and leverage common knowledge
- Harmonize and link internal ontologies and vocabularies using knowledge graphs grounded in standardized ontologies

United States® Census Bureau

Our Assumptions for Achieving Machine Understandable Data

- To perform these transformations in both directions, databases or APIs must expose the data's meaning and structure
- The categories and connections of variables and results must be able to match the detail and complexity of human languages
and purposes
- This kind of machine-understandable data is needed to guide the use, combination, and explanation of data in “responsible” ways

United States® Census Bureau

Our Assumptions for Achieving Machine Understandable Data

- Use open standards for APIs with the ability to link into knowledge graphs
- Improve guidance and metadata around appropriate data usage, permissions and requirements for purposes such as research analytics, text-and-data mining, and AI system ingestion

United States Census Bureau

First Steps Toward Machine Understandable Data

- Gathering internal and external written documentation of existing data products and:
  - Mining them for terminology to use in metadata harmonization and linking; or
  - Releasing them in raw formats for the training of AI models
- Adopting data formats which allow for rich metadata as well as generating metadata “sidecars” for more traditional formats such as CSV or SAS

United States® Census Bureau

AI and Open Government Data Assets

United States® Census Bureau

Data Dissemination Standards

Data dissemination standards that support human-readable and machine-understandable public data

Formats, metadata, and documentation prioritized to facilitate AI applications

Metadata standards that distinguish between raw data (like data from sensor networks) and derived data (like statistical data from the Census Bureau)

Data Accessibility and Retrieval

Consider users when disseminating AI-ready data—including atypical users

Take measures to ensure user-friendly interfaces, such as clear labeling and readable formats, for Commerce's online data resources

Understand both the needs of data users and the ROI in making our data more AI-ready

Partnership and Engagement

Industry and academic stakeholders collaborating with government to shape the design and dissemination of AI-ready open data

Potential areas of industry/academia partnership and contribution include enhancing data quality, integrity, and usefulness for AI purposes

Encourage open-source connections, collaborations, and platforms

Data Integrity and Quality

- Collectively address challenges related to authenticity bias, privacy, data quality, equity, and ethical use while maintaining transparency and accountability
- Develop security protocols to mitigate risks of unauthorized data access and manipulation
- Promote transparency in data sourcing and processing methods to enhance trust and reliability
- Set expectations for reporting data quality and ensuring that information will be carried through to the end user

Data Ethics

Establish clear legal and ethical guidelines for AI data usage that …

- Ensure privacy rights
- Preserve property rights
- Focus on equitable outcomes

THANK YOU

Michael Hawes
Senior Statistician for Scientific Communication
Research and Methodology Directorate
U.S. Census Bureau

michael.b.hawes@census.gov

United States® Census Bureau