Reshaping the Federal Sector: Statistical Data and the Role of AI

Dr. Nancy A. Potok

May 2, 2024

AI Day

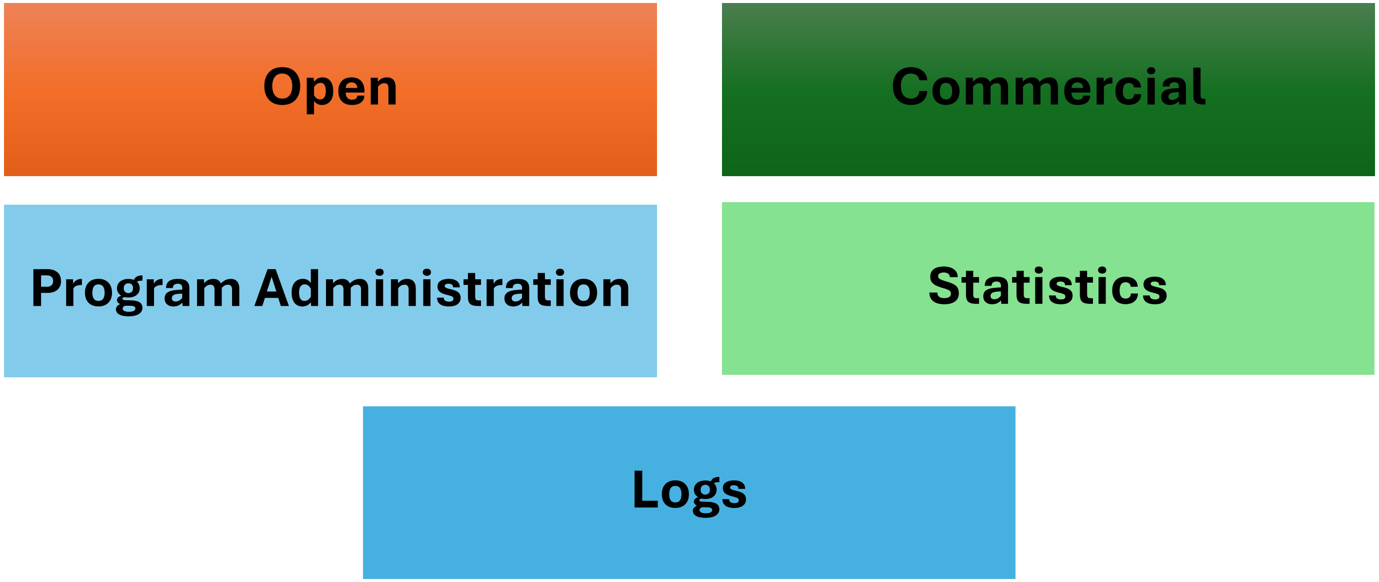
- The current data ecosystem
- The future of AI in federal statistics
- The return on investment for using AI.

VALUE OF DATA DRIVEN POLICY AND OPERATIONS

Trends transforming AI practices:

- Faster timelines and rapid change
- Demand for evidence-based decision-making
- Statutory mandates and federal policies
- Increased data availability
- Data science and technology advances

TYPES OF DATA THAT MAY AVAILABLE:



Open

Commercial

Program Administration

Statistics

Logs

What types of data do you “own”?
Have access to?
Would like to acquire and use?

FEDERAL DATA ECOSYSTEM

Foundations of Evidence Based Policy Act

CHIPS and Science Act

Federal Data Strategy

OMB guidance

GAO/NIST Frameworks

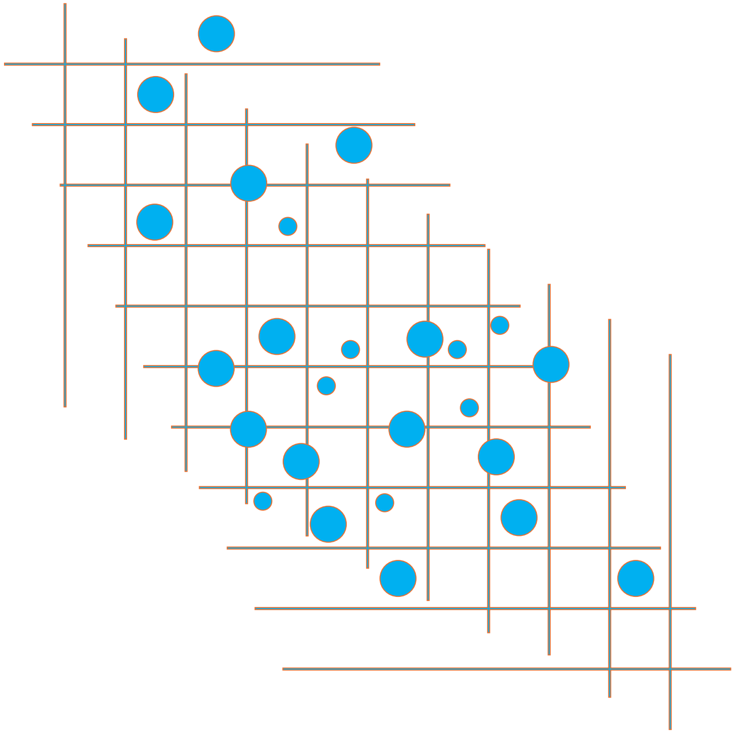
Executive Orders

Agency Data Governance

Data owners

Infrastructure Owners

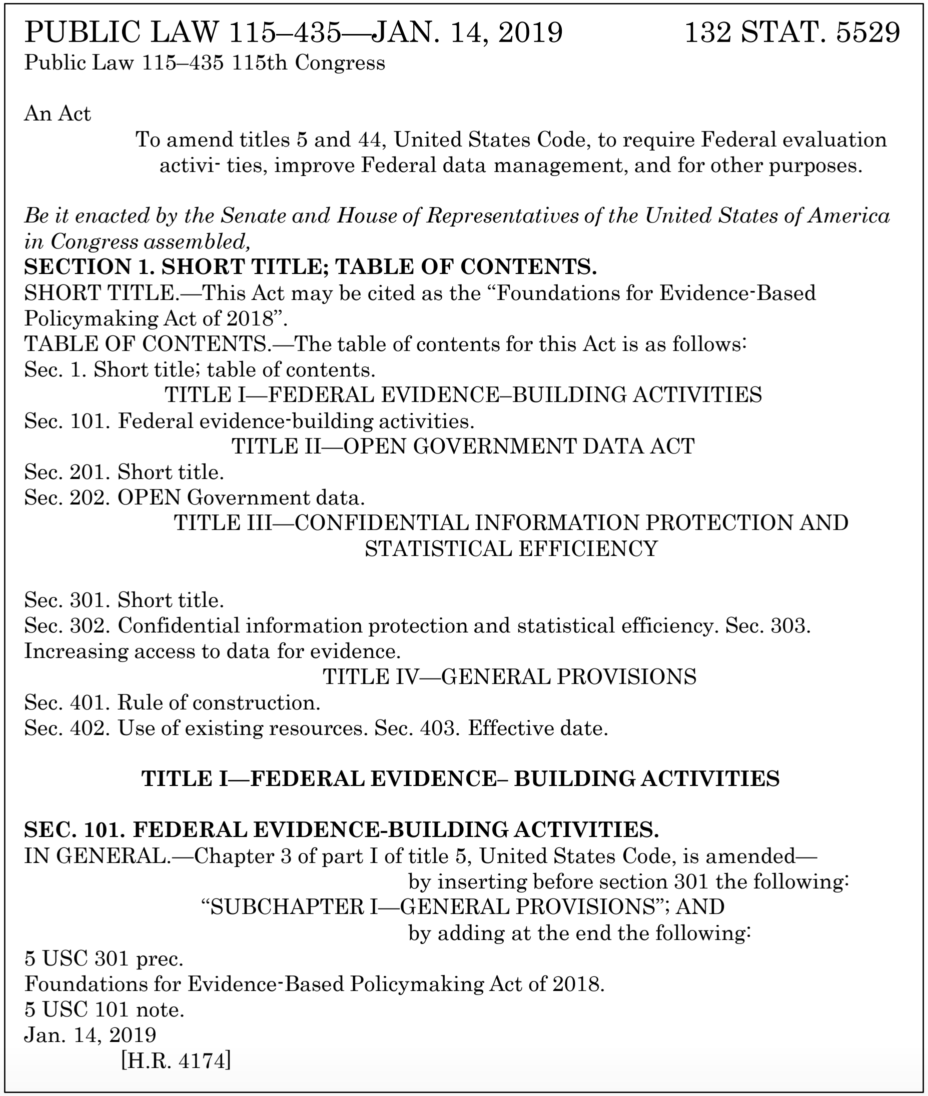
Stakeholders



Center for Excellence
in Public Leadership
THE GEORGE WASHINGTON UNIVERSITY

Foundations of Evidence Based Policy Making Act of 2018

P.L. 115-435



PUBLIC LAW 115–435—JAN. 14, 2019 132 STAT. 5529

Public Law 115–435 115th Congress

An Act

To amend titles 5 and 44, United States Code, to require Federal evaluation activi- ties, improve Federal data management, and for other purposes.

Be it enacted by the Senate and House of Representatives of the United States of America in Congress assembled,

SECTION 1. SHORT TITLE; TABLE OF CONTENTS.

SHORT TITLE.—This Act may be cited as the “Foundations for Evidence-Based Policymaking Act of 2018”.

TABLE OF CONTENTS.—The table of contents for this Act is as follows:

TITLE I—FEDERAL EVIDENCE– BUILDING ACTIVITIES

SEC. 101. FEDERAL EVIDENCE-BUILDING ACTIVITIES.

IN GENERAL.—Chapter 3 of part I of title 5, United States Code, is amended—

by inserting before section 301 the following:

“SUBCHAPTER I—GENERAL PROVISIONS”; AND

by adding at the end the following:

5 USC 301 prec.

Foundations for Evidence-Based Policymaking Act of 2018.

5 USC 101 note.

Jan. 14, 2019

[H.R. 4174]

• EXECUTIVE GUIDANCE ON DATA

EXECUTIVE ORDERS &

OMB GUIDANCE

- M-19-23 and M-21-27 (Evaluation Studies)
- M-19-15 (Information Quality)
- M-14-03 Administrative Data for Statistical Purposes
- M-19-18 Federal Data Strategy
- Circular A-130

OMB M-19-23

- Develop a learning agenda tied to the agency's strategic plan and submit to OMB annually with the budget
- Establish a data governance body chaired by the CDO
- Submit an annual evaluation plan
- Assess capacity annually to assess ability and infrastructure to carry out evidence building activities like foundational fact finding, performance measurement, policy analysis, and program evaluation
- Identify the data needed to answer those questions.

"

M-21-27 LEARNING AGENDAS

“OMB expects agencies to use evidence whenever possible to further both mission and operations, and to commit to build evidence where it is lacking…

Fundamental to this task are effective processes to strategically plan for evidence building, using the Evidence-Building Plans (i.e., Learning Agendas) and Annual Evaluation Plans as tools…

OMB strongly believes that implementing the Evidence Act is not a compliance exercise, and that agencies should develop the required Title I deliverables (i.e., the Learning Agenda, Annual Evaluation Plan, and Capacity Assessment for Statistics, Evaluation, Research and Analysis) in a way that fulfills their purpose as strategic, evidence-building plans…”.

https://www.evaluation.gov/evidence-plans/learning-agenda/

EXECUTIVE ORDERS & OMB GUIDANCE ON AI

Executive Order 13859 of February 11, 2019: Maintaining American Leadership in Artificial Intelligence

Instructed regulatory agencies to publish information about how they plan on regulating AI in compliance with 10 principles laid out by the OSTP and OMB.

Executive Order 13960 of December 3, 2020: Promoting the Use of Trustworthy Artificial Intelligence in the Federal Government

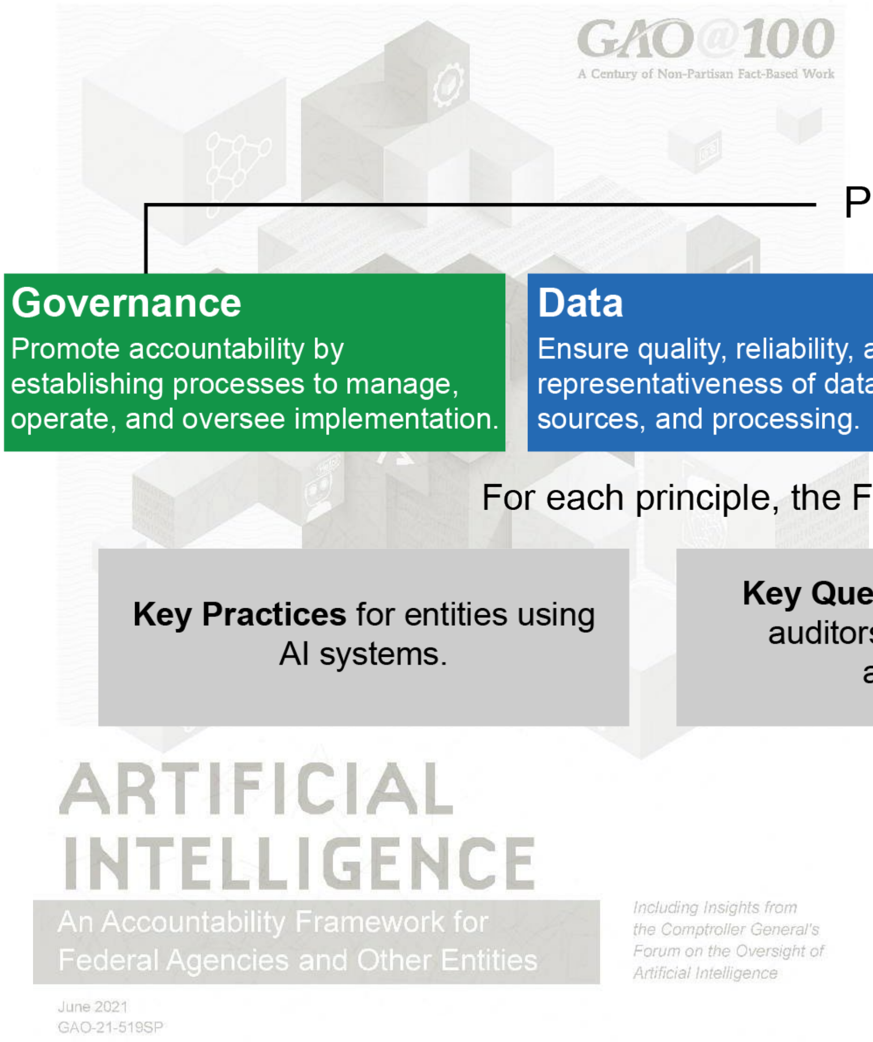
Ordered Federal Chief Information Officers Council to "identify, provide guidance on, and make publicly available the criteria, format, and mechanisms for agency inventories of…use cases of AI by agencies

Executive Order 14110 of October 30,2023: Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence

Establishes a government-wide effort to guide responsible artificial intelligence (AI) development and deployment through federal agency leadership, regulation of industry, and engagement with international partners

M-24-10: Advancing Governance, Innovation, and Risk Management for Agency Use of Artificial Intelligence

Roles and responsibilities of CAIO, AI compliance plan, use case inventories, risk management requirements, advancing innovation, use of AI outputs,



GAO@100

A Century of Non-Partisan Fact-Based Work

ARTIFICIAL INTELLIGENCE

An Accountability Framework for Federal Agencies and Other Entities

Including Insights from the Comptroller General's Forum on the Oversight of Artificial Intelligence

June 2021

GAO-21-519SP

Principles

Governance
Promote accountability by establishing processes to manage, operate, and oversee implementation.

Data
Ensure quality, reliability, and representativeness of data sources, and processing.

Performance
Produce results that are consistent with program objectives.

Monitoring
Ensure reliability and relevance over time.

For each principle, the Framework provides the following:

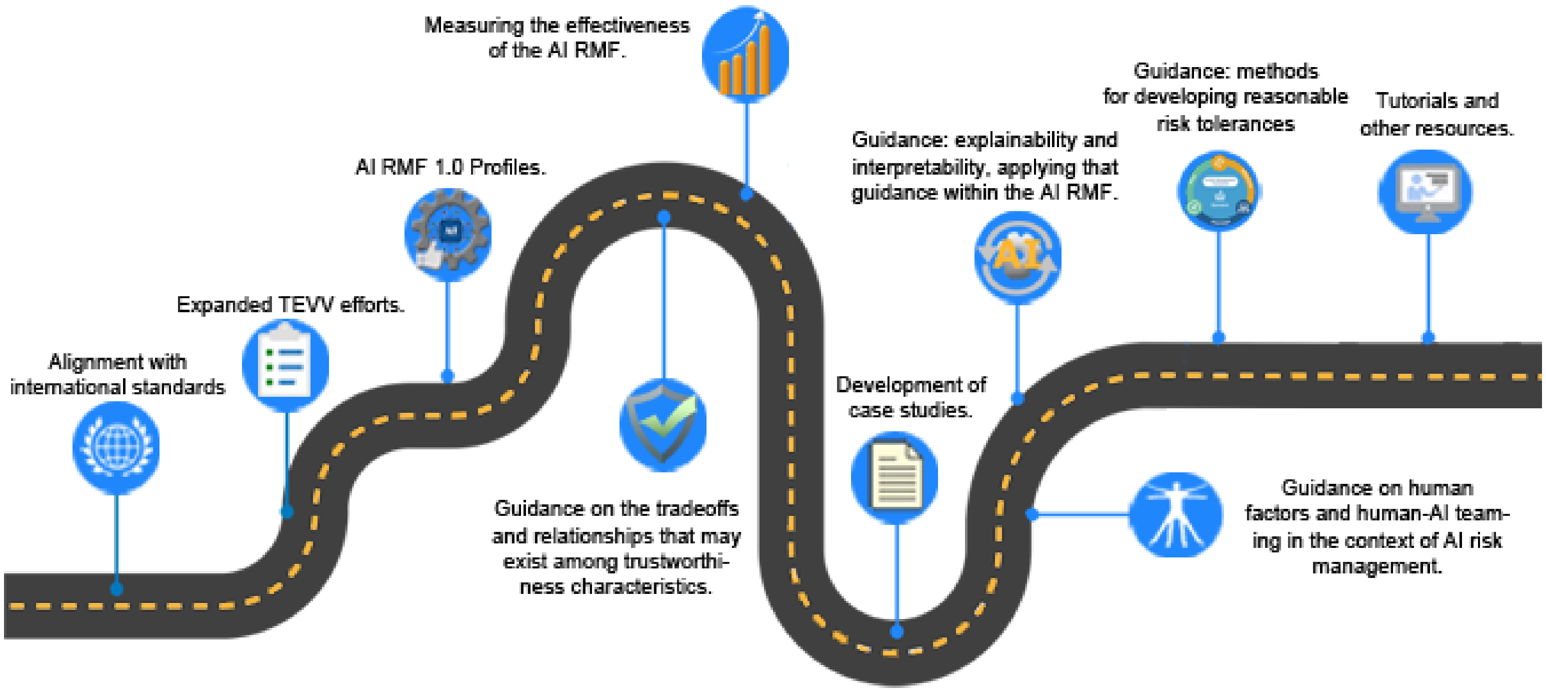
Key Practices for entities using AI systems.

Key Questions for entities, auditors, and third-party assessors.

Audit Procedures with the types of evidence for auditors and third-party assessors to collect.

Alignment with Yellow Book and Green Book standards

NIST AI RISK MANAGEMENT ROADMAP



Alignment with international standards

Expanded TEVV efforts.

AI RMF 1.0 Profiles.

Guidance on the tradeoffs and relationships that may exist among trustworthiness characteristics.

Measuring the effectiveness of the AI RMF.

Development of case studies.

Guidance: explainability and interpretability, applying that guidance within the AI RMF.

Guidance on human factors and human-AI teaming in the context of AI risk management.

Guidance: methods for developing reasonable risk tolerances

Tutorials and other resources.

NIST AI Resource Center: https://airc.nist.gov/Home

NIST AI Risk Management Playbook

Map
Context is recognized and risks related to context are identified

Measure
Identified risks are assessed, analyzed, or tracked

Govern
A culture of risk management is cultivated and present

Manage
Risks are prioritized and acted upon based on a projected impact

AI RMF PLAYBOOK

NIST NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY U.S. DEPARTMENT OF COMMERCE

https://airc.nist.gov/AI_RMF_Knowledge_Base/Playbook

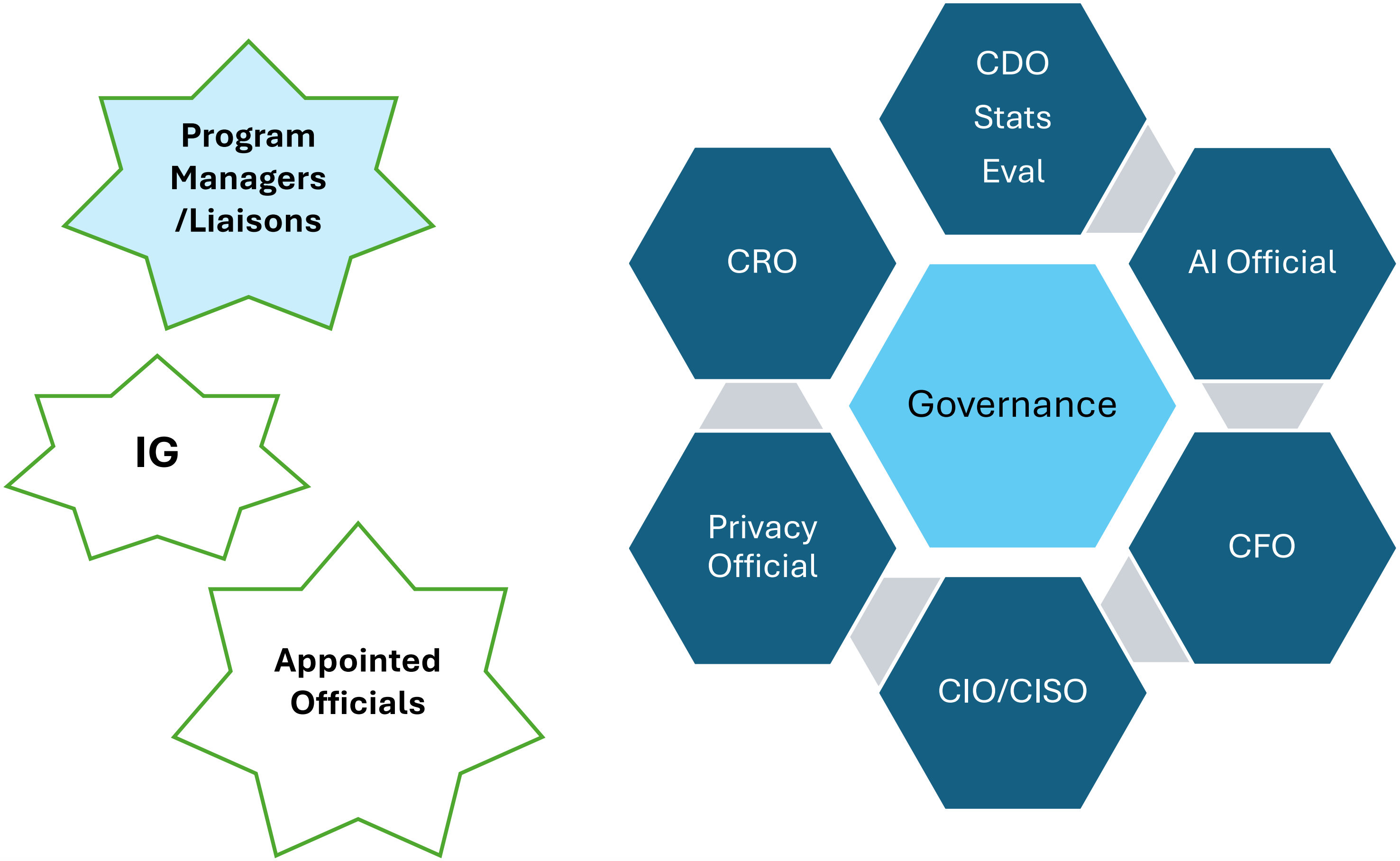
NATIONAL ARTIFICIAL INTELLIGENCE RESEARCH RESOURCE PROJECT (NAIRR)

The U.S. National Science Foundation and collaborating agencies launched the National Artificial Intelligence Research Resource (NAIRR) pilot, a first step towards realizing the vision for a shared research infrastructure that will strengthen and democratize access to critical resources necessary to power responsible AI discovery and innovation.

.

- Partnering with 10 other federal agencies as well as 25 private sector, nonprofit and philanthropic organizations, the NAIRR pilot will provide access to advanced computing, datasets, models, software, training and user support to U.S.-based researchers and educators.
- By connecting researchers and educators with the resources needed to support their work, the NAIRR pilot will power innovative AI research and, as it continues to grow, inform the design of the full NAIRR ecosystem.

Agency Data Roles and Responsibilities



Program Managers /Liaisons

IG

Appointed Officials

CDO

Stats

Eval

CRO

AI Official

Governance

Privacy Official

CFO

CIO/CISO

• PROMINENT AI PROJECT HIGHLIGHTS ACROSS DEPARTMENTS

- Department_of_Energy (197)
- Department_of_Health_and_Human_Services (172)
- Department_of_Commerce (73)
- Department_of_Interior (70)
- Department_of_Homeland_Security (67)
- Department_of_Veterans_Affairs (40)
- Department_of_Agriculture (39)
- National_Aeronautics_and_Space_Administration (33)
- Department_of_State (31)
- Department_of_Labor (17)
- U.S._Agency_for_International_Development (14)
- Social_Security_Administration (14)
- Department_of_Treasury (14)
- Department_of_Transportation (14)

- https://github.com/thoppe/Federal-AI-inventory-analysis-2023/blob/main/results/AI_highlights_by_Department.md

A FEW DOL EXAMPLES…

- “Form Recognizer for Benefits Forms: AI is used to extract data from complex forms and assign data entries to field headers.
- Bureau of Labor Statistics (BLS) ELI Code Labeling: Machine learning is used to label data with Entry Level Item (ELI) codes based on word frequency counts from item descriptions.
- DOL Intranet Website Chatbot Assistant: AI-powered chatbot that answers common procurement questions and provides information about contracts.
- Automatic Data Processing with Form Recognizer: AI technology extracts necessary information from complex forms, streamlining data extraction workflow.
- OEWS Occupation Autocoder: Tool that assigns Standard Occupational Classification (SOC) codes to occupation titles based on state-submitted response files.
- Automatic Document Processing: Technology automatically extracts selection boxes from continuation of benefits forms without manual intervention.
- Document Validation with AI: AI detects discrepancies in addresses and identifies unreadable or distorted text in official documents.”

DEEPER DIVE…NCHS

“Sequential Coverage Algorithm (SCA) and partial Expectation-Maximization (EM) estimation in Record Linkage:
NCHS Data Linkage Program has implemented both supervised and unsupervised machine learning (ML) techniques in their linkage algorithms. The Sequential Coverage Algorithm (SCA), a supervised ML algorithm, is used to develop joining methods (or blocking groups) when working with very large datasets. The unsupervised partial Expectation-Maximization (EM) estimation is used to estimate the proportion of pairs that are matches within each block. Both methods improve linkage accuracy and efficiency.”

“Detecting Stimulant and Opioid Misuse and Illicit Use:
Analyze clinical notes to detect illicit use and miscue of stimulants and opioids”

“Transcribing Cognitive Interviews with Whisper:
Current transcription processes for cognitive interviews are limited. Manual transcription is time-consuming and the current automated solution is low quality. Recently, open-sourced AI models have been released that appear to perform substantially better than previous technologies in automated transcription of video/audio. Of note is the model by OpenAI named Whisper (publication, code, model card) which has been made available for under a fully permissive license. Although Whisper is currently considered state-of-the-art compared to other AI models in standard benchmarks, it has not been tested with cognitive interviews.
We hypothesize Whisper will produce production quality transcriptions for NCHS. “

RESOURCES

Data resources repository materials

https://resources.data.gov

WHAT & HOW?

1

Foster innovation and optimize data assets

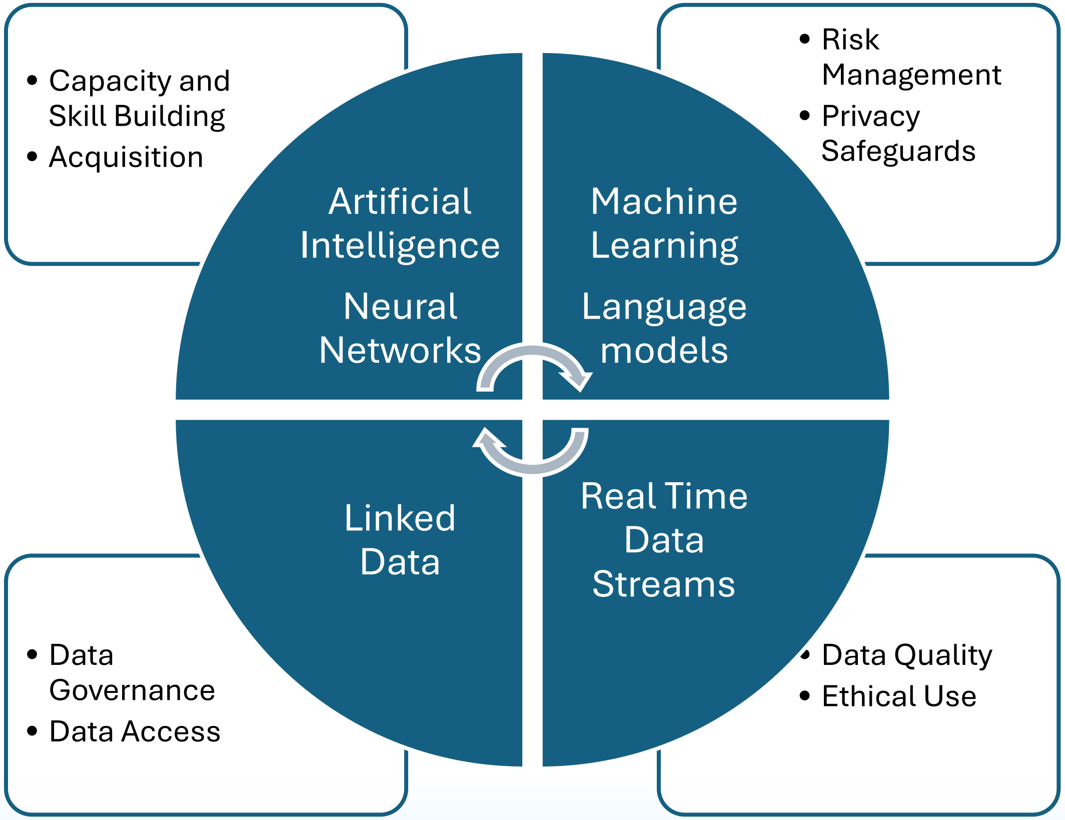
- Identify high value information contributions
- Formulate new ways to use data to provide value
- Build coalitions to implement AI projects

2

Identify opportunities for application of AI tools

- Become familiar with AI trends
- Understand the strengths and weaknesses of AI approaches
- Scope projects for success

Growing use and availability of analytical tools



- Capacity and Skill Building
- Acquisition

- Risk Management
- Privacy Safeguards

Artificial Intelligence

Neural Networks

Machine Learning

Language models

Linked Data

Real Time Data Streams

- Data Governance
- Data Access

- Data Quality
- Ethical Use

CHALLENGES IN THE FEDERAL STATISTICS SPACE

Transparent

Buy vs. Build

Acquisition

Compute Infrastructure

Security

Governance

Operational Sustainment

Workforce

Trustworthy

Ethics

Equity

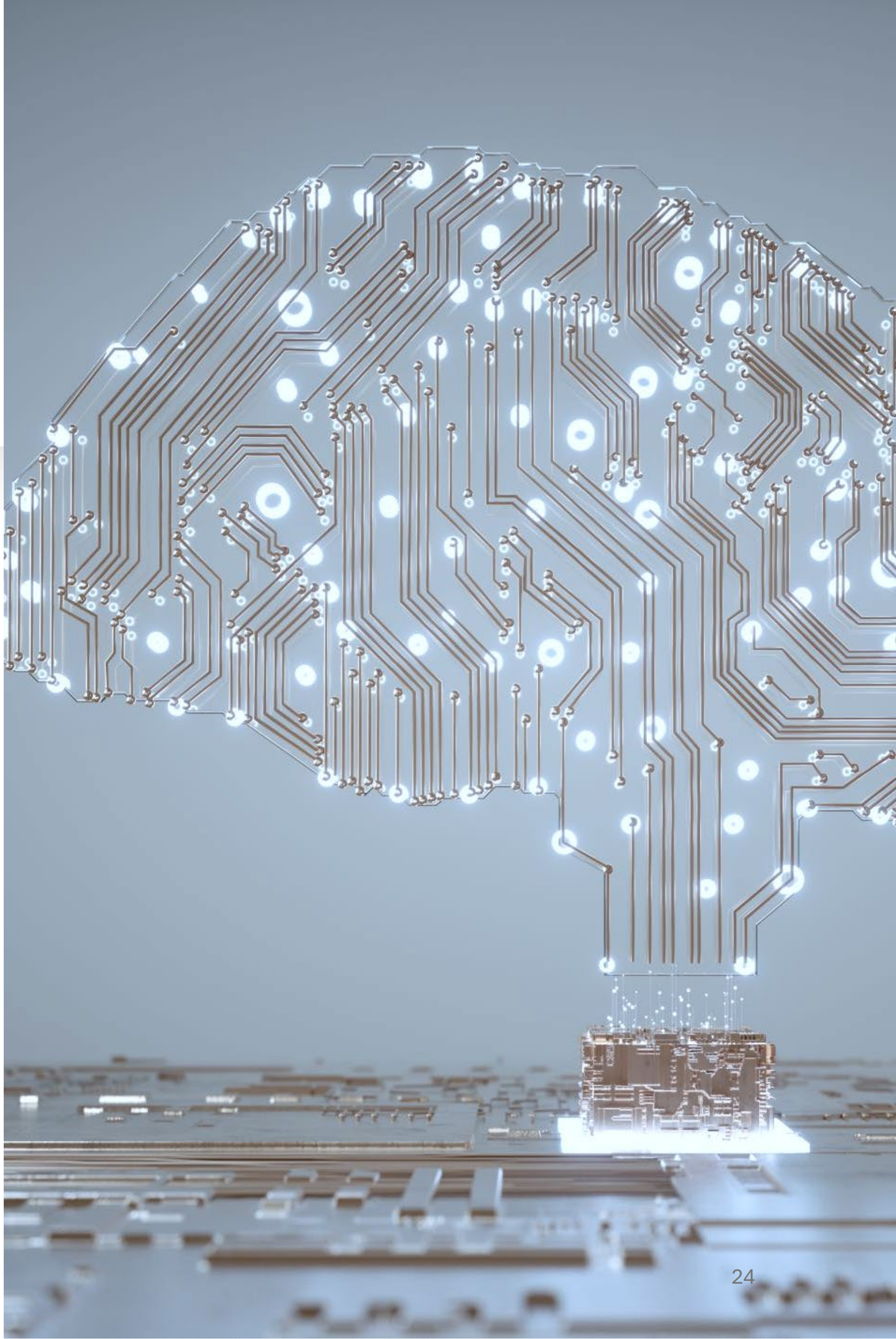
ETHICAL QUESTION

Ask “***SHOULD we?”*** not just “can we” & be prepared to stop if the answer is “no.”

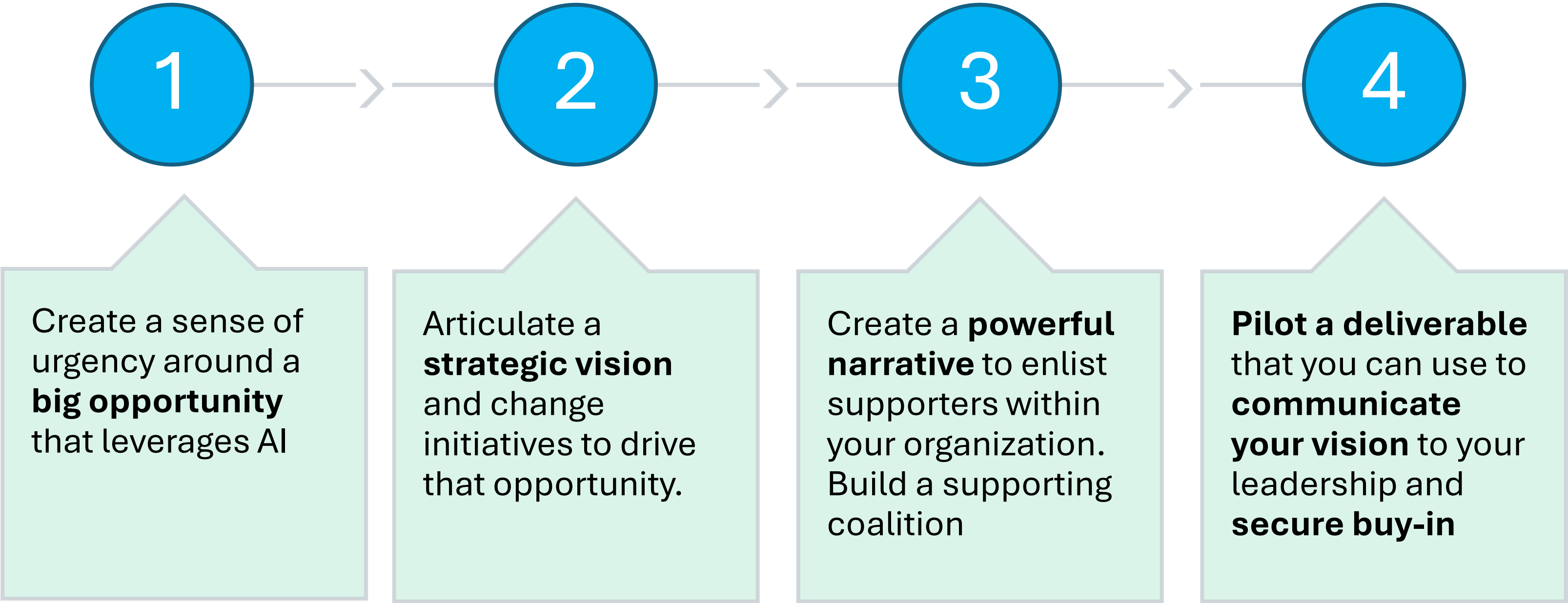
…

Elements of a Successful AI Approach

1. Technical Fundamentals
2. AI Strategy
3. AI-Ready Organizational Culture
4. AI Governance Structure
5. Responsible Leadership/Ethical Framework



CREATING A ROADMAP FOR AI PROJECTS



1

Create a sense of urgency around a **big opportunity** that leverages AI

2

Articulate a **strategic vision** and change initiatives to drive that opportunity.

3

Create a **powerful narrative** to enlist supporters within your organization. Build a supporting coalition

4

Pilot a deliverable that you can use to **communicate your vision** to your leadership and **secure buy-in**

GOVERNANCE: QUESTIONS TO ASK BEFORE STARTING:

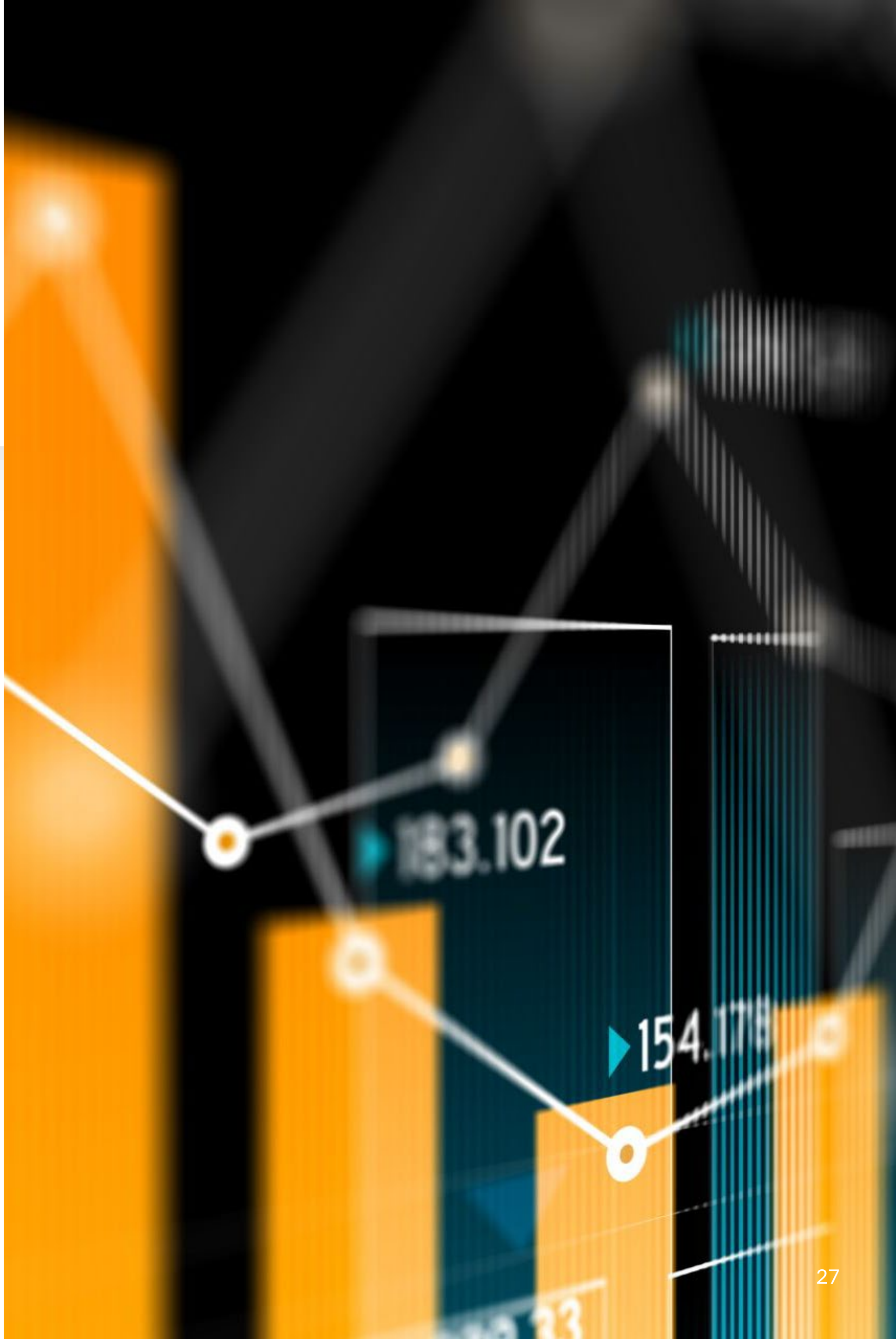
- What AI data governance structures exist in your agency?

- How should you interact with the data governance structure to advance an AI project and manage risk?

- How much time to build in for the governance process?

WHAT DOES THE FUTURE HOLD?

- Data Collection
- Data Processing
- Data Dissemination
- Interacting with Data Users
- Providing Access to Data



DATA COLLECTION

- Target non-respondents in surveys with local information on how data are actually being used (NASS experiment)
- Predictive modeling (NASS – predictive cropland data layer)
- Collecting information from alternative sources (Census satellite pictures for new construction) to improve on small samples
- Improved analysis of nonresponse data to optimize collection times
- Quality control for data linkages from multiple sources
- Self classification for respondents (Economic Census BEACON)

DATA PROCESSING

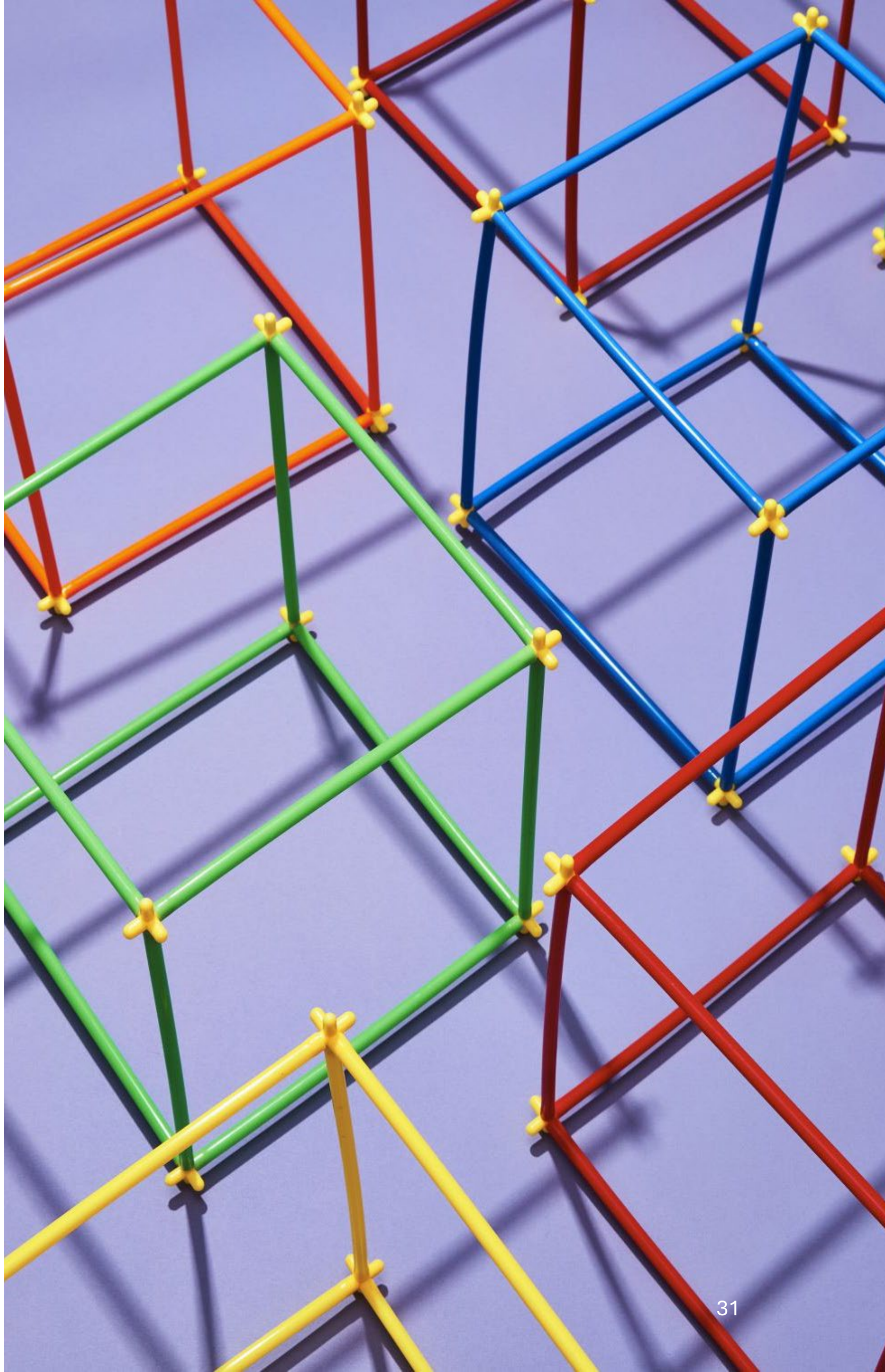
- Coding responses (SOC, etc.)
- Validating addresses and identities
- Writing code for automated processing
- Extracting data from questionnaires
- NLP/ML for reading unstructured text
- Improved estimates for missing data
- Time Series Analysis

DATA DISSEMINATION

- Generative AI to assist with communications
- Chatbots for FAQs
- Improved search and display capabilities using generative AI (data visualizations, etc.)

INTERACTING WITH DATA USERS

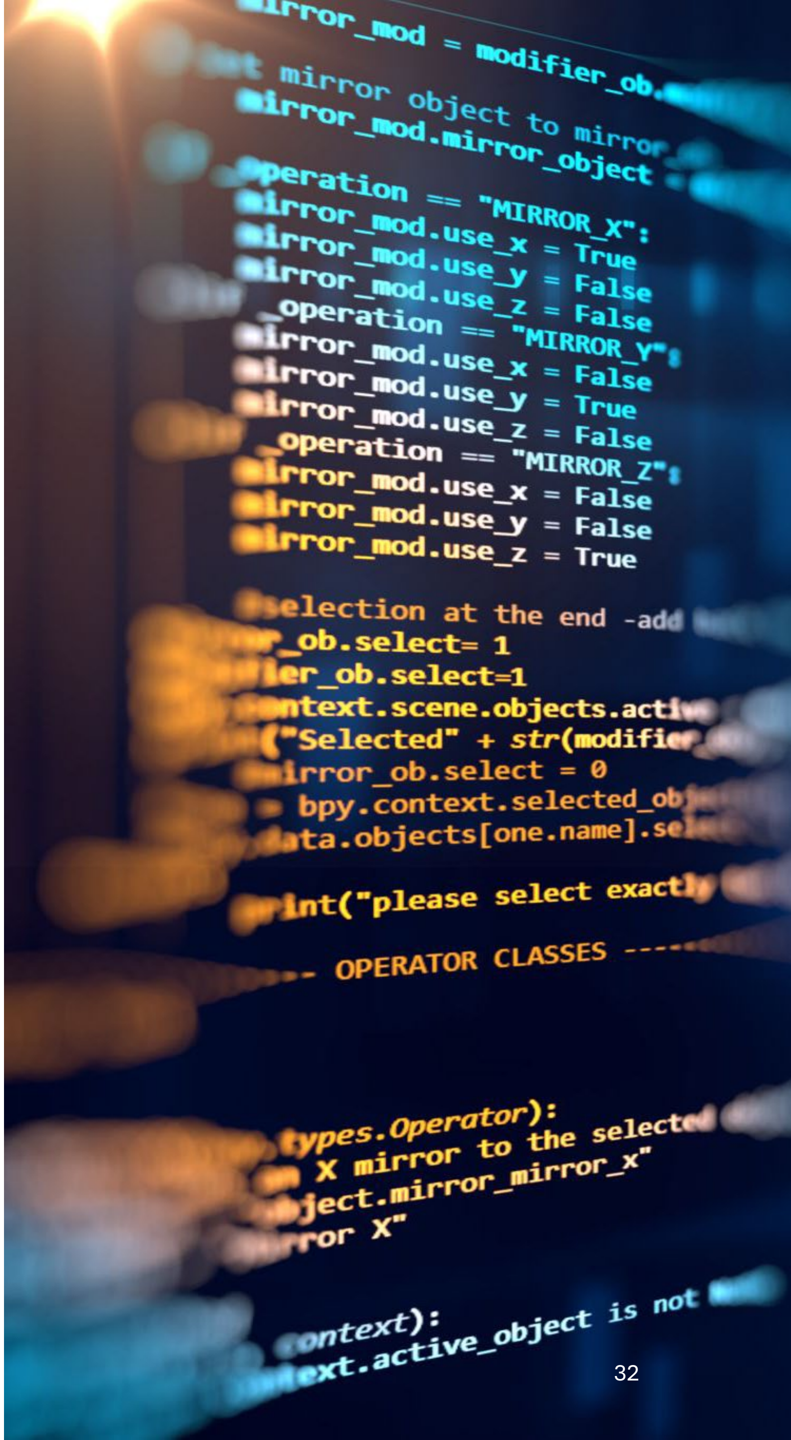
- Identify the users through ML and NLP tools (Search and Discovery Platform)
- Conduct outreach withnew users at smaller institutions and MSIs
- Chatbots to answer FAQs



PROVIDING ACCESS TO DATA

- Guiding Researchers (concierge tools)
  - Search and Discovery Platform
  - Chatbots
  - Easy access to similar research

- Protecting Privacy and Confidentiality
  - Synthetic data
  - Identity validation



```
irror_mod = modifier_ob.
 mirror object to mirror
irror_mod.mirror_object
peration == "MIRROR_X":
irror_mod.use_x = True
irror_mod.use_y = False
irror_mod.use_z = False
_operation == "MIRROR_Y"
irror_mod.use_x = False
irror_mod.use_y = True
irror_mod.use_z = False
_operation == "MIRROR_Z"
irror_mod.use_x = False
irror_mod.use_y = False
irror_mod.use_z = True

election at the end -add
_ob.select= 1
er_ob.select=1
ntext.scene.objects.activ
("Selected" + str(modifie
irror_ob.select = 0
 bpy.context.selected_ob
ata.objects[one.name].se

int("please select exactl

- OPERATOR CLASSES ----

types.Operator):
 X mirror to the selected
ject.mirror_mirror_x"
rror X"

context):
ext.active_object is not
```

Return on Investment

Quantitative:

- Increased efficiency
- Increased productivity
- Demonstrated value of data (usage data)
- Higher response rates on surveys
- More relevant data (faster, more granular, responsive to mission of agency)

Qualitative:

Better relationships with users

Increased public trust

What is the cost of not investing in AI?

https://hdsr.mitpress.mit.edu

HARVARD DATA SCIENCE REVIEW

SPECIAL ISSUE IV | 2024
HDSR. MITPRESS.MIT.EDU

DEMOCRATIZING DATA

Discovering Data Use for Research + Policy

HDSR

HARVARD DATA SCIENCE REVIEW

DEMOCRATIZING DATA

Discovering Data Use + Value for Research + Policy

Join this two-part webinar to hear the authors of the *Harvard Data Science Review* special issue, “Democratizing Data: Discovering Data Use and Value for Research and Policy,” share their knowledge, experience, and ideas on how technology can help build a new, sustainable data ecosystem.

May 21 + May 22, 2024
12:00 PM – 2:00 PM EST
Virtual: Zoom

https://datascience.harvard.edu/calendar_event/democratizing-data-discovering-data-use-and-value-for-research-and-policy/