

---

# From Data Gravity to Discovery

## Secure-by-Design and AI-Ready Data Pipelines

**Dr. Seth Berl**

**Former Deputy Chief Data Officer, US Department of Energy**

**DOE Data Days  
March 3, 2026**

# Dr. Seth Berl

## Former DOE Deputy Chief Data Officer | US Department of Energy



- Co-Founder, **OpSpark** – Mission-Driven AI-Ready Data
- Member of the **Board of Directors** for **NANO Nuclear Energy**
- Former **Deputy Chief Data Officer** for the US Department of Energy where our team delivered the Enterprise Data Strategy with a **unified strategy and vision to accelerate mission goals**.
- Former **Chief Technology Officer** for GovSmart, Inc. focusing on Federal Defense & Civilian agencies to solve complex operational challenges.
- Principally architected and implemented projects including **high performance compute, data science, AI/ML, enterprise networks, cybersecurity** & more.
- Integrating my academic experience, including a **Ph.D. & M.A. in atomic physics from UVA** with industry experience to **support the success of each organization's diverse missions**.

# The AI Era: Transformational Changes

AI is already embedded in nearly all facets of our lives



- **AI and Gen AI**
  - Operational Efficiency Through Automation
  - Augmenting decision-making
  - Accelerated content and asset generation
  - Hyper-personalized experiences
  - Workforce transformation & skills evolution
  - Embedding digital assistants in critical workflows
- **AI is no longer experimental**
  - AI is operational in government, industry, and academia

Public-impacting decisions are increasingly AI-assisted.

# Data Bookends AI: The Consequences of Poor Data Quality



In 2025 only 10% of organization are “completely ready” to adopt AI. (Harvard Business Review)

July 30, 2024

**Gartner Predicts 30% of Generative AI Projects Will Be Abandoned After Proof of Concept By End of 2025**

SYDNEY, July 30, 2024 — At least 30% of generative AI projects will be abandoned by the end of 2025, due to poor data quality, inadequate data governance, and other factors, according to Gartner Inc.

## Data Quality Across the Digital Landscape

ArcGIS Data Reviewer

Summer 2024

According to consulting firm Gartner, bad data costs organizations an average of \$12.9 million per year. Other reports show similarly staggering figures. The McKinsey Global Institute, for example, found that poor-quality data can lead to a 20 percent decrease in productivity and a 30 percent increase in costs.

**By 2028, 80% of GenAI Business Apps Will Be Built On Existing Data Management Platforms**

By Admin — ON JUL 19, 2025

According to leading research firm Gartner Inc., by 2028, a staggering 80% of Generative AI (GenAI) business applications will be developed on existing data management platforms, dramatically cutting down delivery timelines and complexity by up to 50%.

INTERNATIONAL NEWS

acceldata

## The Compounding Cost of Delay

The financial toll compounds as bad data travels through your systems:

- 20-30% of enterprise revenue is lost due to data inefficiencies (Gartner)
- Data teams spend 50% of their time on remediation (Ataccama)
- By the time a quality issue hits a boardroom dashboard, fixing it can cost 100x more than catching it at ingestion (1x10x100 rule)

## The Hidden Costs of Poor Data Governance



Exelegant  
7,307 followers

July 16, 2025

In today's data-driven economy, information is one of the most valuable assets a business can possess. Yet for many enterprise organizations, data remains a liability rather than an advantage.

Why? Because of one persistent issue: poor data governance.

The cost of poor data governance goes far beyond misfiled records or compliance gaps—it silently drains millions from company resources each year. In this article, we'll explore how the absence of a robust governance framework leads to real financial, operational, and strategic damage—especially in companies with over 200 employees.

**Poor Data Quality has substantial direct, indirect, hidden, and cultural costs.**

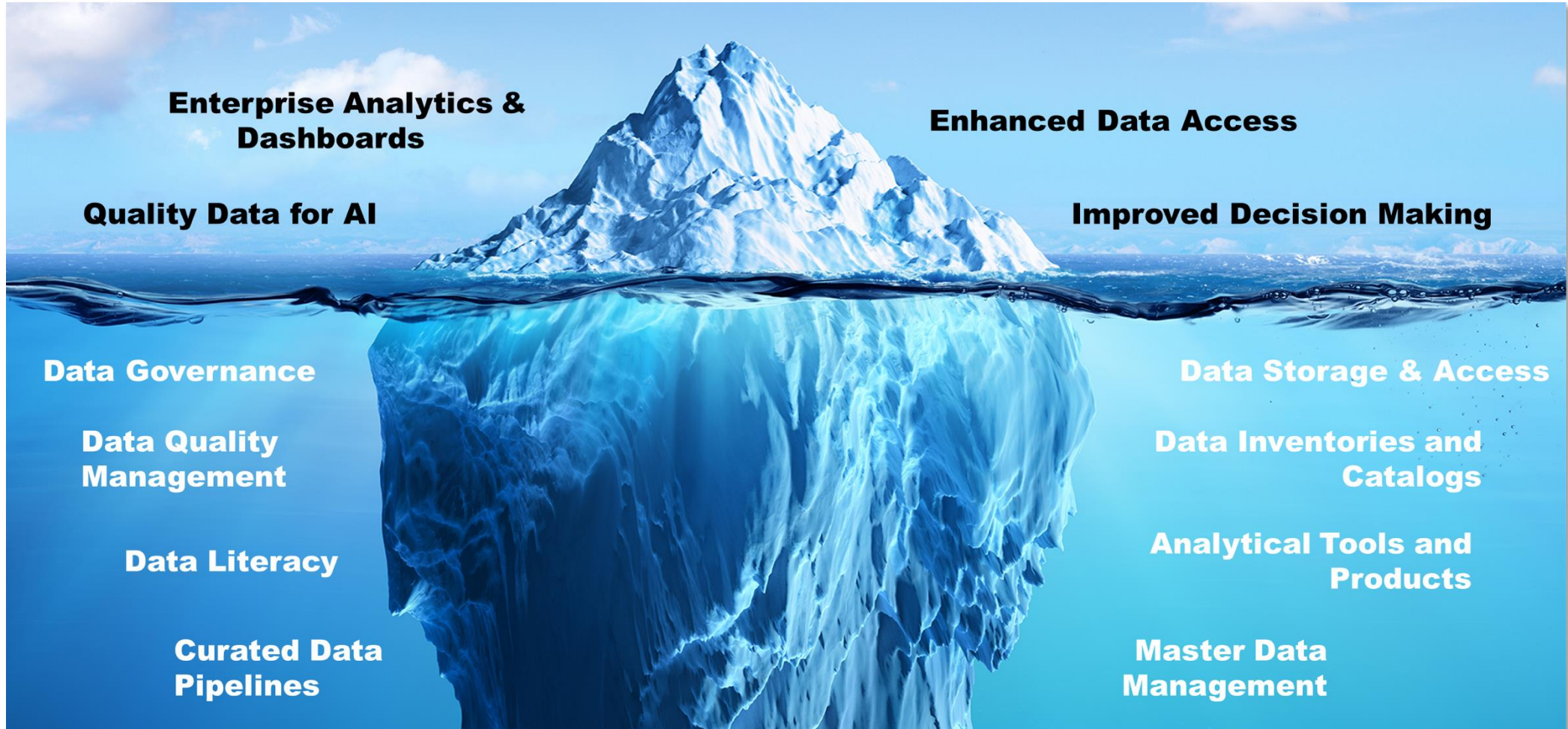
# Data Gravity

- As data accumulates, applications and services are increasingly pulled toward the data
- For DOE, data gravity is magnified by:
  - Geographical and mission **federation**
  - **Broad domain** purview (mission, operational, and types)
  - **Systems heterogeneity** (HPC, on-prem enterprise IT, cloud, scientific instruments, edge sensors, etc.),
  - Growing **AI workloads** that demand low-latency access to governed data rather than ad hoc copies
- Yet...
  - Mission Leaders, Data Consumers, **and** Scientists want **access** to **quality**, timely, **authoritative data** and decision support tools
  - Delivering these capabilities is much more complex than generally understood

Image: [NASA](#)

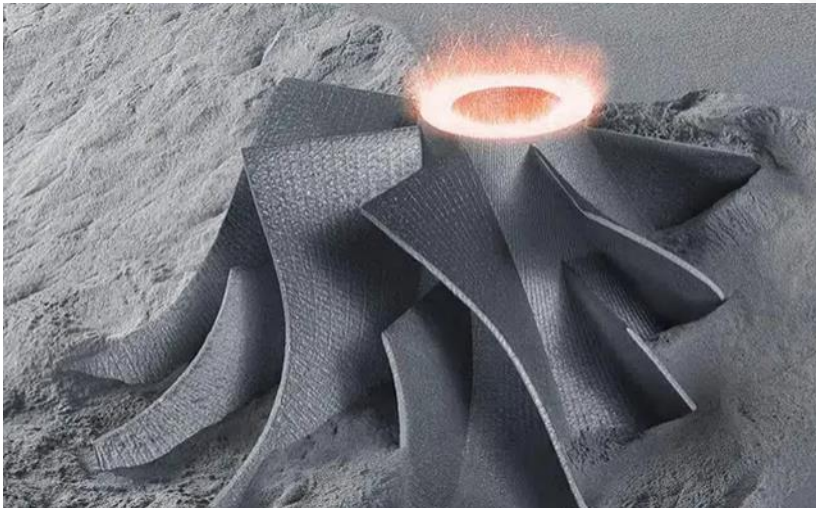
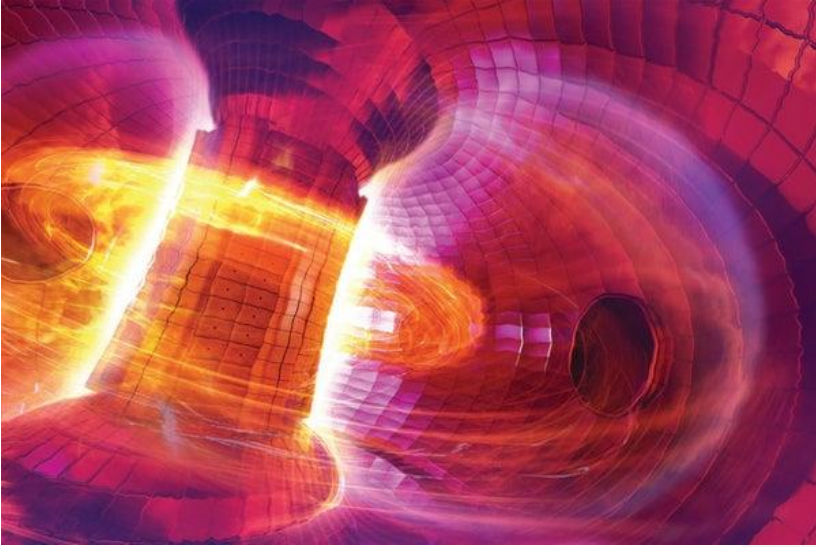
# Scaling data management for AI

The same foundational data management is required, but AI has introduced new management approaches.



# The Catalyst

---



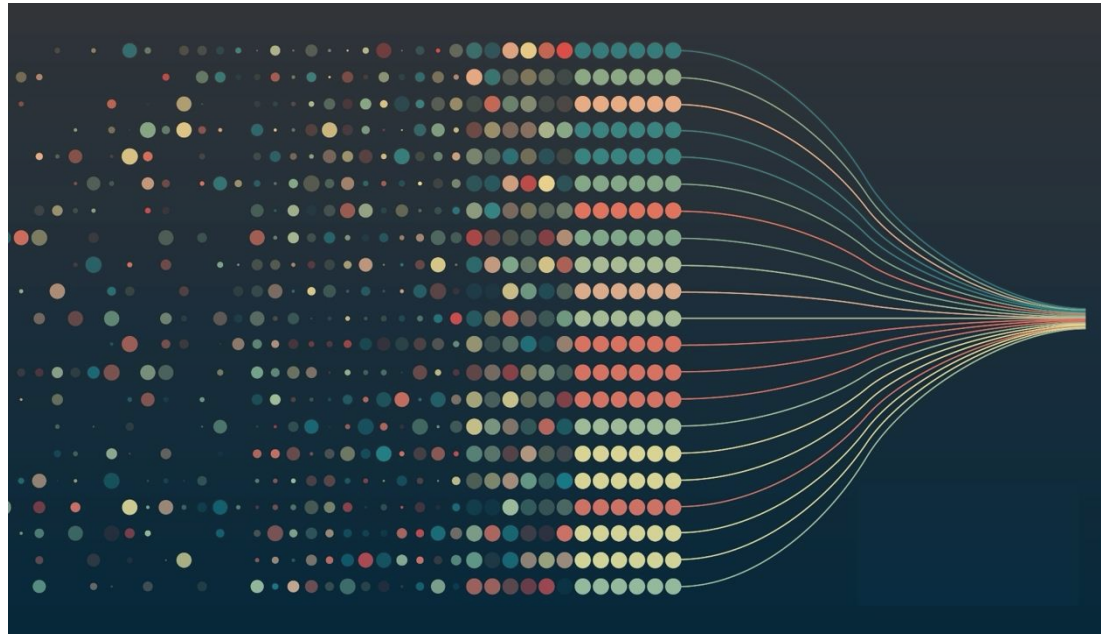
- **Genesis Mission**<sup>1,2</sup> is a call to action
  - National effort to build the American Science and Security Platform, a closed-loop AI experimentation platform that links federal scientific datasets with HPC to train scientific foundation models and AI agents.
  - Genesis targets areas where the U.S. government controls the most sensitive datasets and where private labs cannot train equivalent models due to classification or export restrictions.
- 26 High-Impact AI Projects Identified<sup>3</sup>
- Building an integrated AI platform<sup>1</sup>

1. <https://www.whitehouse.gov/presidential-actions/2025/11/launching-the-genesis-mission/>

2. <https://genesis.energy.gov/>

3. <https://www.energy.gov/documents/genesis-mission-science-and-technology-challenges>

# Genesis Mission Data Standards



- + ○ • “Secure access to appropriate datasets, including proprietary, federally curated, and open scientific datasets, in addition to synthetic data generated through DOE computing resources.”

Data-Access  
Policies

Data-  
Management  
Processes

Cybersecurity  
Standards

Classification,  
Privacy, & Export  
Compliance

Intellectual  
Property &  
Licensing  
Procedures

# Data Provenance, Metadata, and Traceability Framework

- **Identify “initial data and model assets ... including digitization, standardization, metadata, and provenance tracking.”**
- **Crucial for:**
  - Reproducibility
  - Auditability
  - Scientific integrity

What is the data?

How was it collected?

When was it collected?

By whom?

What is the origin and history?

Were any transformations made and when?

# FY25-28

## ENTERPRISE DATA STRATEGY

U.S. Department of Energy

# US Department of Energy Enterprise Data Strategy

### Guiding Principles

-  *Treat Data as a Strategic Asset*
-  *Align to Community Best Practices*
-  *Measure & Solve for What Matters*
-  *Embrace Incremental Progress*
-  *Foster Continuous Learning & Data Fluency*
-  *Design for Scale & Collaboration*
-  *Embed Data Ethics, Equity & Justice*

### Goals & Objectives

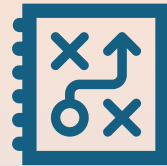


- Maximizing the value of data for AI
- Scaling impact across energy, science & innovation, safety, security, and operations

# Collective Progress



Into early 2025, **DOE Data Governance Board** grew to 500+ DOE participants.



**Enterprise Data Strategy** formally approved and published (initiated Oct. '24 and approved Jan '25)



Formally **chartered governance board** with **80% representation from DOE labs, plants, and sites**



Collectively **created products for full DOE complex use** to increase data fluency & workforce development



Executed the **Data Catalog, Lexicon, & Strategy Integrated Working Groups** to deliver enterprise products



Collaboratively developed an **FY25 implementation plan** with prioritized activities

# Dimensions of Successful Data Programs

Dimension	Meaning	Maturity
<b>Infrastructure &amp; Architecture</b>	Collection, storage, integration	Catalogs, data lake, real-time pipelines, software suites
<b>Analytics &amp; AI</b>	Ability to derive insights, automate decision-making	Predictive analytics, AI-driven service delivery, performance dashboards
<b>Domain Applications</b>	Categorization of related mission data	Consistent, trusted, well-governed, and interoperable data.
<b>Strategy &amp; Governance</b>	Alignment between data initiatives and mission, governance structure	Data strategy document, CDO/CAIO role, cross-agency governance body
<b>Culture &amp; Processes</b>	Embedding data in everyday operations, decision making	Data literacy programs, democratized access, continuous improvement loops
<b>Performance &amp; Value</b>	Demonstrable outcomes (efficiency, economic value, innovation)	Metrics showing cost-savings, service improvement, business outcomes

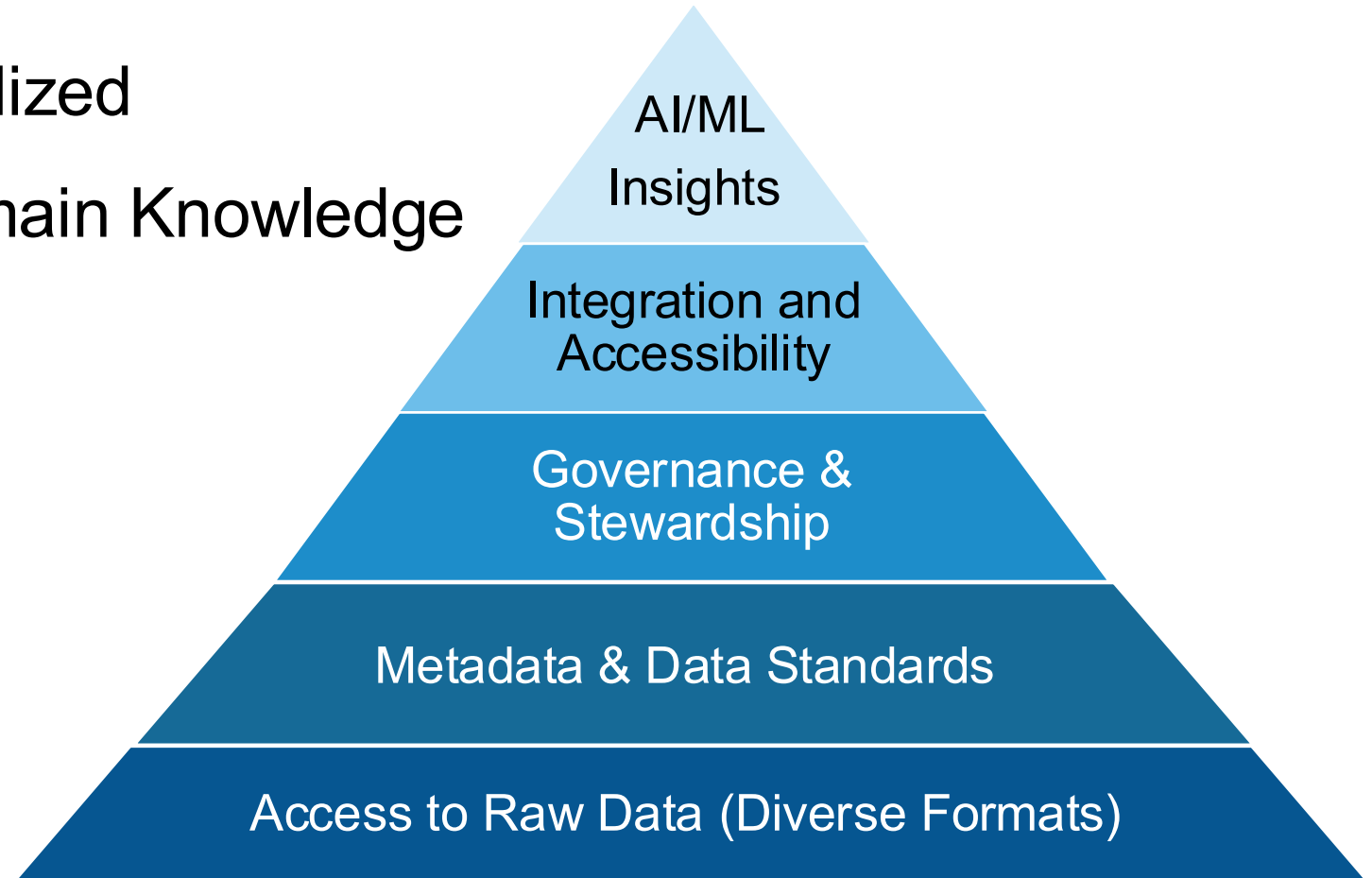
# Tailoring Data Governance to DOE's Primary Missions

DG Goals	Science & Research (S&R) Mission	Weapons Design & Manufacturing
<b>Access Control</b>	Tiered access (internal, public, restricted)	Clearance-based access (e.g., Confidential/Secret/TS, RD/FRD)
<b>AI Readiness</b>	Enabling reproducible science and accelerating discovery	Validated, explainable AI for simulation, design, and manufacturing assurance
<b>Compliance Standards</b>	Open Science, Open Data, FAIR Principles	DOE Orders (e.g., 471.1B, 205.1C), NNSA-specific directives, DoD nuclear security standards
<b>Data Sharing</b>	Open access where possible; OSTI submission, public repositories	No public release; distribution limited under DOE/NNSA classification; cross lab/site data sharing is a key priority
<b>Incentives</b>	Academic and peer review, worldwide collaboration, science acceleration	Mission assurance, accelerate weapons production, risk avoidance, compliance, program certification
<b>Metadata</b>	FAIR-aligned and scientific discipline-specific domains (e.g., chemistry, materials, physics)	Security-tagged metadata with classification markings, provenance tracking, chain-of-custody
<b>Retention</b>	Always accessible, reusable, records management policies, and funding mandates	Driven by mission continuity and records management policies; retention can be decades to indefinite
<b>Stewardship Model</b>	Collateral-duty stewards may be feasible in research groups, but fulltime institutional should be considered.	Full-time, cleared stewards should be required for compliance and security

# Characteristics of AI-Ready Data

---

- Structured and Standardized
- Contextualized with Domain Knowledge
- Accessible and Secure
- Provenance-Tracked
- Validated and Curated
- Efficiently Managed



# Data Governance Practices to Enable AI-Ready Data

---



## Data Governance Board:

- Key leaders from across organization and must be decision-oriented.
- Identify and tie key organizational priorities.
- Provide sponsorship, resources and accountability.



Require a **standard Data Management Plan (DMP)** for every IT investment



## Develop **Executive and Workforce Data and AI Literacy**

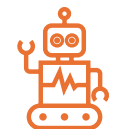
- Leaders must establish a data-driven culture and drive organization change
- Workforce must understand how to interact with data and utilize AI



Deploy, dedicate and incentive **Data Stewards to embed best practices**



**Adopt FAIR principles** (Findable, Accessible, Interoperable, Reusable) for Metadata Standards



**Automate metadata capture and cataloging**, especially in HPC environments, while maintaining “human in the loop” verification via Data Stewards.

# Data Governance Practices to Enable AI-Ready Data

---

## Responsible AI:



- Invest in validation of data pipelines for “fit for purpose” contextualization for AI use cases,
- Maintain dataset provenance for all AI model training,
- Utilize and pilot mini-Rag AI systems and Vector databases with SME-curated and validated datasets



Develop mechanisms to **value and measure data and its ROI** through tracking dataset reuse, reproducibility, cost savings.

- Potential Industry’s demand for datasets (potential monetization)



For data access and security, **establish a “need to know” data policy** for access controls and apply **tiered access controls**



To enable Infrastructure Governance, adopt compute-to-data plans to **minimize network bottlenecks**. **Use tiered storage and edge processing** to manage large datasets.

# Federated Data Pipeline Architecture

## Governance



Meta-data Management

Reference Data Management

Data Quality

Data Catalog

Data Policy Management

Master Data Management

Workflow Management

## Sources



Structured Data Sources

Unstructured Data Sources

Semi-structured Data Sources

External Data Sources

Streaming Data Sources

Event Data Sources

Enterprise Integration Hub

## Ingestion



Batch Processing

API Integrations

Streaming Integrations

## Processing



ETL Services

## Knowledge & Discovery



Enterprise Search

Classification Services

Semantic Layer

Indexing & Profiling

## Storage



Structured Data Lake

Data Warehouse

Graph Database

Unstructured Data Lake

AI/ Gen AI Databases

Sandbox Workspaces

Enterprise Integration Hub

## Consumption



### Analytics & BI

Scalable Compute

Data Science Notebooks

Reports & Dashboards

Geospatial Analytics



### Advanced AI/ML

AI Services

MLOps

AI Governance

Gen AI Services

## Infrastructure & Security



Identity & Access Management

Audit & Monitoring

Data Protection

Environment Management

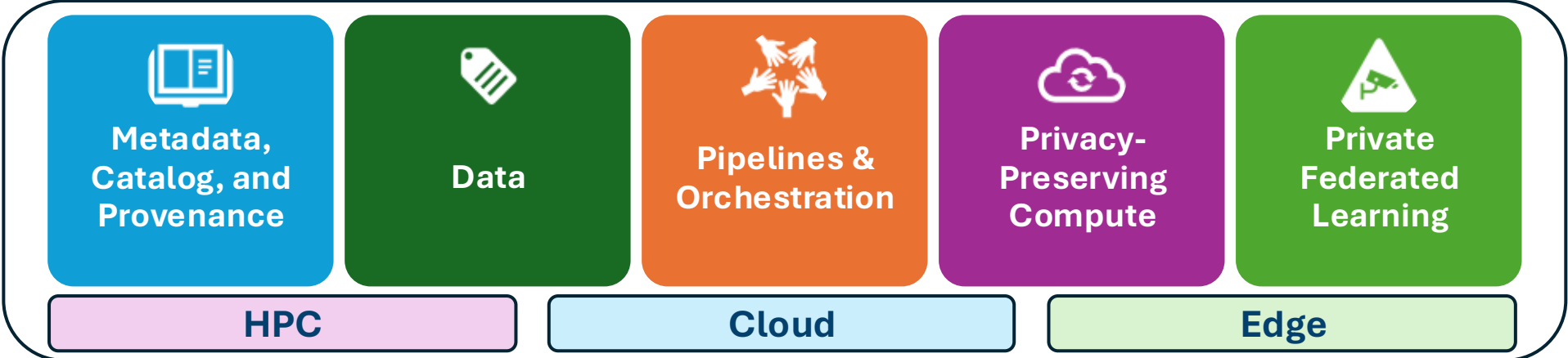
Networking

Infrastructure Provisioning

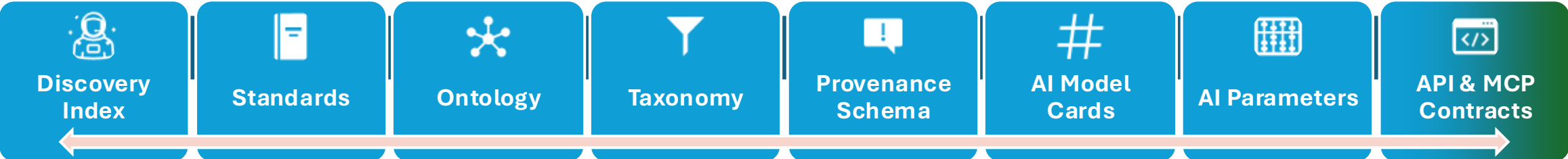
DevOps Pipelines

# Privacy-Preserving Federated Learning Architecture

## Site Boundary



## DOE Complex-Wide Interoperability



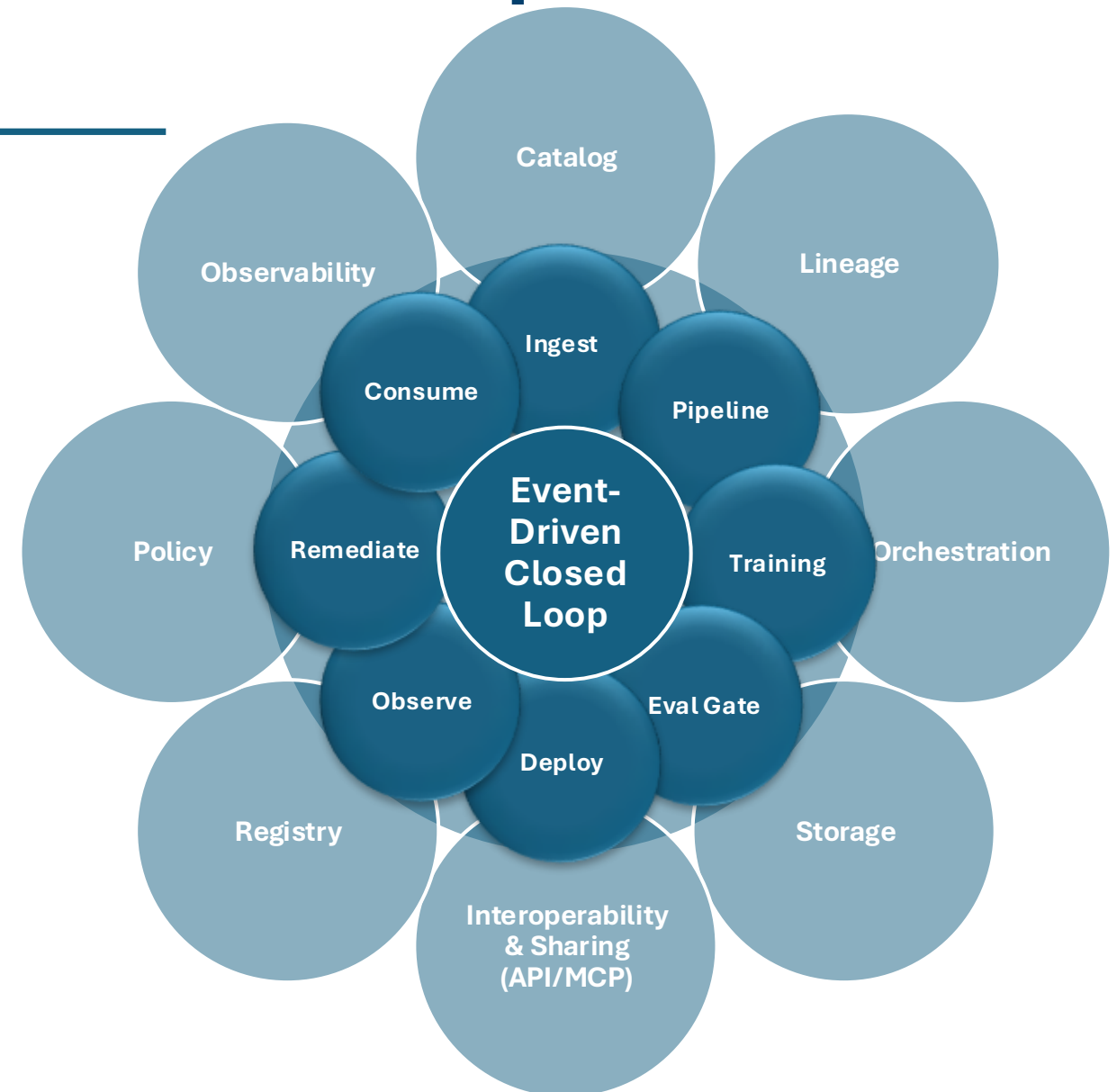
## Permissioned Network, Identity, Trust, and Policy Control Plane



# Example: Complex-wide closed-loop feedback systems

---

- **Goal:** Continuous improvement loops for AI-directed experiments and operational decisioning
- **Deliverables**
  - Standard “model release pipeline”
  - Drift/poisoning resilience playbooks integrated into platform
  - Governance operating model: domain autonomy + complex-wide interoperability
- **Exit criteria**
  - Reduced experiment-to-insight cycle time is demonstrated (baseline vs after)
  - Model degradation is detected + remediated within agreed SLO
  - Multiple domains onboarded using the same repeatable reference architecture



# Accelerating Innovation & Mission Outcomes

## Challenges

**Data gravity** and **Inaccessible Data**

**Federated, heterogeneous platforms**

Data accuracy issues with **the lack of a single source of truth**

Need to **report** on key information internally and publicly

Difficulty **locating** data or **securing authorizations** for sharing data with organization components

Inability to monitor **data poisoning / integrity threats**

**Deploying** AI and RPA in production

**Mission assurance**

## Federated Data & AI Platform

 *Infrastructure & Security*

 *Data Governance*

 *Data Storage*

 *Data Integration*

 *Visualization, Analytics & BI*

 *Knowledge & Discovery*

 *Advanced AI/ML*

## Outcomes

Cost savings through consumption-based models and data deduplication

Built-in security controls & compliance

Connects data sources with AI and analytics

Improved documenting & reporting with semantic search & smart data classification tools

Accelerated time to market for data & AI capabilities

Optimized advanced reporting & AI governance

Increased reusability & trustworthiness of data

# Key Enablers to Mission Success



Clear **strategy** and **executive sponsorship**



Strong **Data & AI Governance**



Modern **infrastructure**



**Interoperability & standards**



**Share first** with controls



**Capacity building** – data scientists, analytics, stewards



Demonstrating **quick wins** and **iterative value**



**Security Controls & Compliance**



**Technology is NOT a silver bullet**

# Dr. Seth Berl

Former DOE Deputy Chief Data Officer | US Department of Energy



Seth Berl, PhD

CTO | CDO | Director | Advisor | Speaker



# Thank you!

# Questions?

[LinkedIn.com/in/SethBerl](https://www.linkedin.com/in/SethBerl)

[seth@opspark.ai](mailto:seth@opspark.ai)



# Appendix

# Is DOE ready for vibe coding?



Image: MIT Technology Review

# Core Principles for High-Impact Program Strategies

Principles adopted and embraced by DOE

1



**It takes a village**

*We seek widespread buy-in and forge partnerships*

2



**Do no harm**

*We use data responsibly*

3



**The whole is greater than the sum of its parts**

*We realize a collective impact*

4



**Measure what matters**

*We focus on meaningful mission impacts*

5



**Don't reinvent the wheel**

*We leverage & optimize what we have within our organizations*

6



**All boats rise with the tide**

*We foster fluency & collaboration for collective success*

# Mission-Driven Technology

---



- Data is increasingly being described as a “core asset”, like: people, technology, and capital.
- A strong data foundation enables:
  - Faster, more accurate policy and operational decisions
  - Improved services that citizens expect
  - Public-private partnerships
  - Improved disaster response
- Governments are using data to drive efficiency, effectiveness, innovation, responsiveness



# Genesis Mission Science & Technology

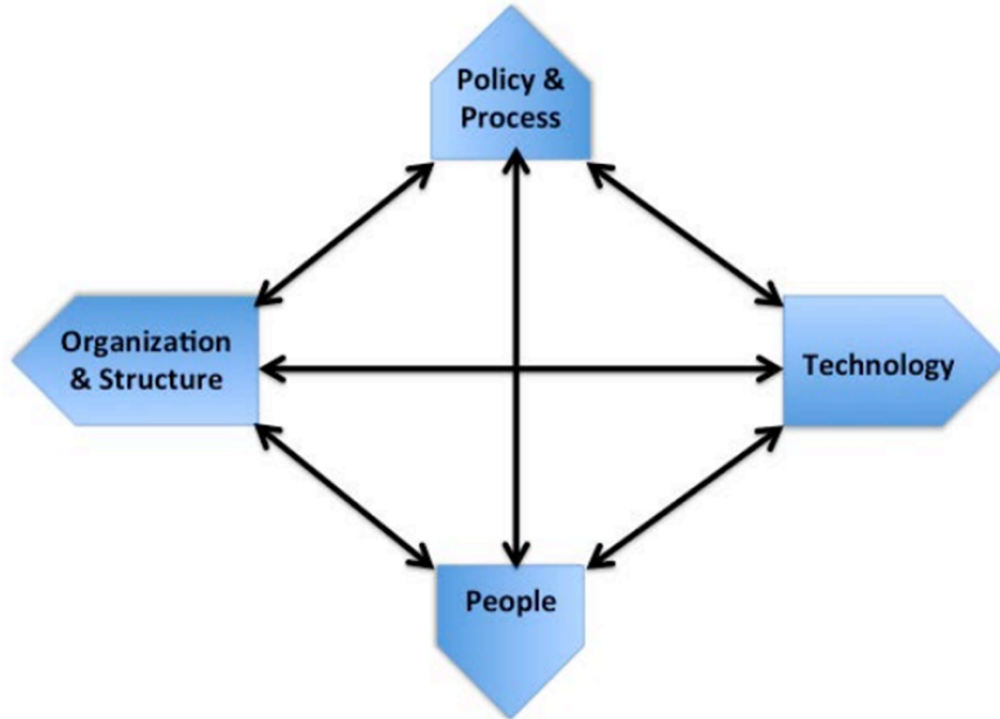
## Challenges to Accelerate AI-Enabled Leadership

---

- **Scaling the Grid to Power the American Economy:** Using AI to improve power grid planning, interconnection, operations, and security — enabling decisions up to 20–100 times faster and improving electricity cost and reliability by up to 10 percent.
- **Harnessing America’s Historic Nuclear Data:** Digitizing eight decades of nuclear research to create a secure, searchable database to inform future energy and security decisions.
- **Enhancing Particle Accelerators for Discovery:** Deploying AI to make accelerators adaptive and autonomous, accelerating breakthroughs in medicine, materials, and energy.
- **Designing Materials with Predictable Functionality:** Using AI to design materials based on performance goals, shrinking development timelines from decades to months.
- **Unleashing Subsurface Strategic Energy Assets:** Applying AI to model underground environments for responsible, cost-effective energy development.
- **Achieving AI-Driven Autonomous Laboratories:** Automating experiments to speed discovery of new drugs, advanced materials, and next-generation energy technologies.
- **Reenvisioning Advanced Manufacturing and Industrial Productivity:** Bridging research and production with AI-driven systems that strengthen supply chains, improve manufacturing productivity and capability, speed the design to production loop, and create American jobs.
- **Discovering Quantum Algorithms with AI:** Accelerating quantum algorithm development to unlock breakthroughs in energy, chemistry, and logistics.
- **Recentering Microelectronics in America:** Advancing next-generation microelectronics to secure U.S. technological leadership, economic prosperity, and national security.

# Organizational Change – Starts with Culture

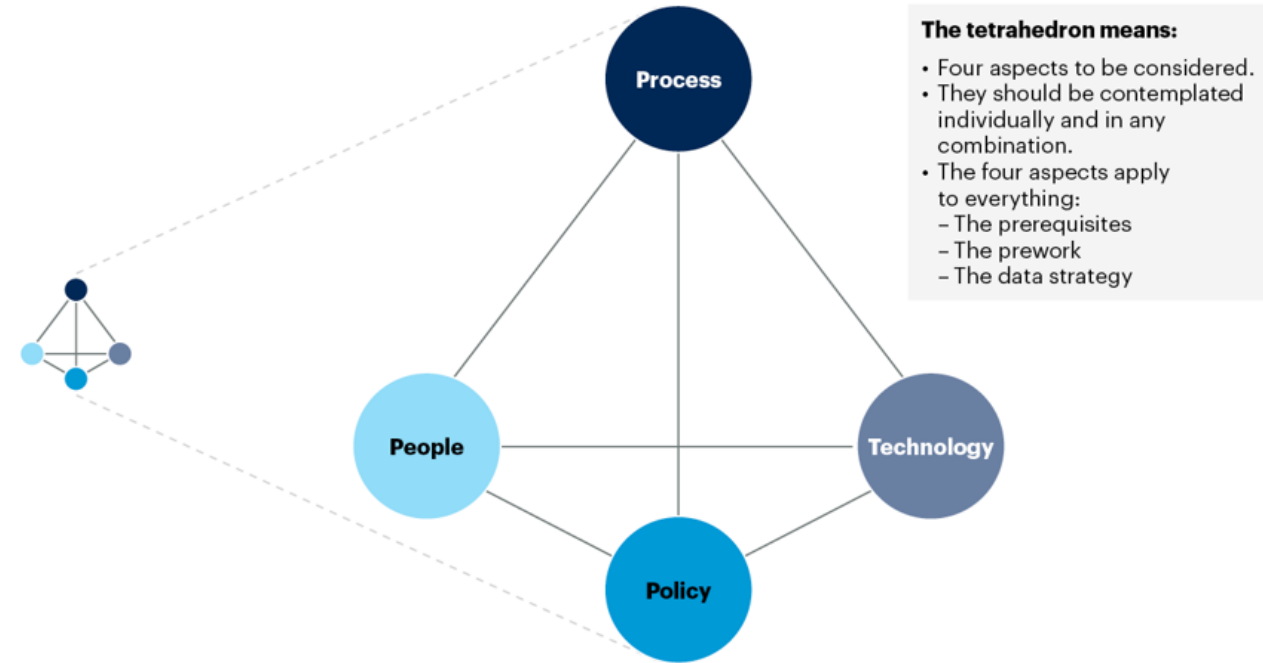
People, Process, Policy, and Technology



[Leavitt's Diamond: Change Framework](#)

(published in 1965)

## Multidimensional Data Deliverables



[Gartner Tetrahedron: How Technical Professionals Help With Effective Data and Analytics Strategies](#)

(published in 2025)

**Data can no longer be seen as a backend function  
nor AI as some futuristic buzzword**