



Architecting Discovery

A Capability Driven Approach to Governing Scientific Data at Scale

Asit Sharma, Chief Data Architect, Strategic Deterrence, LLNL

Josh Deotte, Research Engineer, Materials Engineering, LLNL

February 18, 2026



The Opportunity

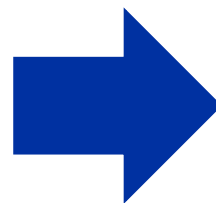
Transitioning from islands of brilliance to a “Brilliant Enterprise”

Project-Specific Tools: Bespoke ingestion and storage solutions

Security & Boundary Friction: Data "bottlenecks" occur whenever research must move across security boundaries or between facilities

Duplicate Efforts: Analysts at different labs spend time reinventing metadata schemas and transformation scripts

Dark Data: Valuable data often remains "dark" and inaccessible for reuse across the wider DOE complex



Standardized Capability Layer: Instead of siloed tools, labs share a "Business Capability Model" that defines common ways to handle acquisition, transformation, and sharing

Technology Enablers: A shared alignment of HPC platforms, analytics frameworks, and metadata services

Interoperable Governance: Architecture governance establishes reference architectures and review processes that work across labs

Mission Alignment: Data flows securely across the "National Security Enterprise" value chain

Capability Modeling: An Overview

What is it?

- An Enterprise Architecture framework that provides a technology-agnostic, organizationally neutral view of the enterprise
- A vehicle to describe “What” an organization does independent of “How” it does it, “Who” does it or “Which” technology enables it
- Used in the industry for years for aligning organization’s investments/spend with business priorities

Structure

- Hierarchical structure, organized into multiple levels:
 1. Level 1: Enterprise Domains of interest (i.e. pillars in Mission-to-Impact canonical value chain for scientific domains: Mission -> Need Definition -> Data Management -> Knowledge -> Impact)
 2. Level 2: Major (stable) Capabilities (i.e. Instrumentation & Source Management)
 3. Level 3: Sub-Capabilities (implementable capabilities where technology & data domains attach)



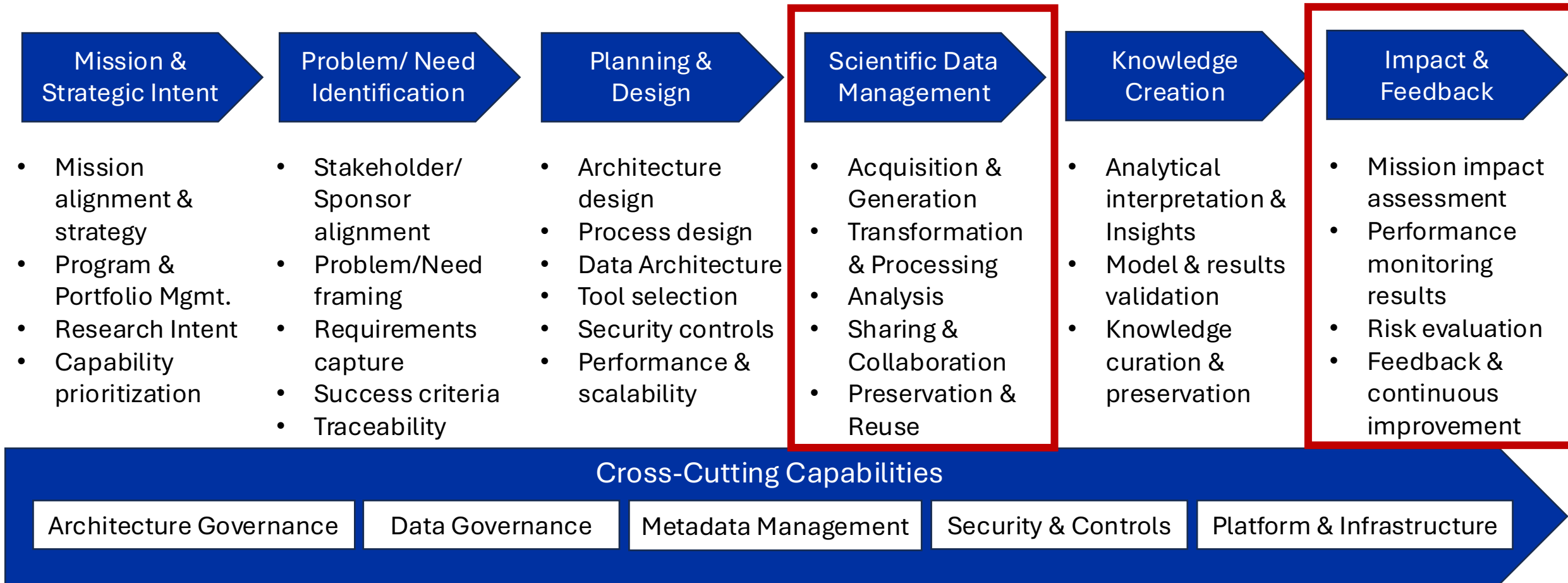
Scientific Data Management Value Chain

- Value Chain serves as a structured abstraction layer to shift focus on “**What**” needs to be accomplished
- Business Operations have used abstractions such as “Lead-to-Cash”, “Procure-to-Pay” etc. to prioritize business capabilities and streamline technology spend
- For scientific data management, we’ll use “Acquisition-to-Impact” (a subset of broader “Mission-to-Impact”) value chain to organize our capabilities
- Using this value chain allows for a shift in conversation from “how many petabytes we store” to “how we drive mission success”
- The architectural reasoning for this approach is briefly summarized below:
 1. Captures the entire research lifecycle
 2. Maps data domains
 3. Enables consistent reasoning across scientific domains (i.e. DIW vs. Climate modeling)
 4. Strengthens mission alignment



Mission-to-Impact: Value Chain Pillars

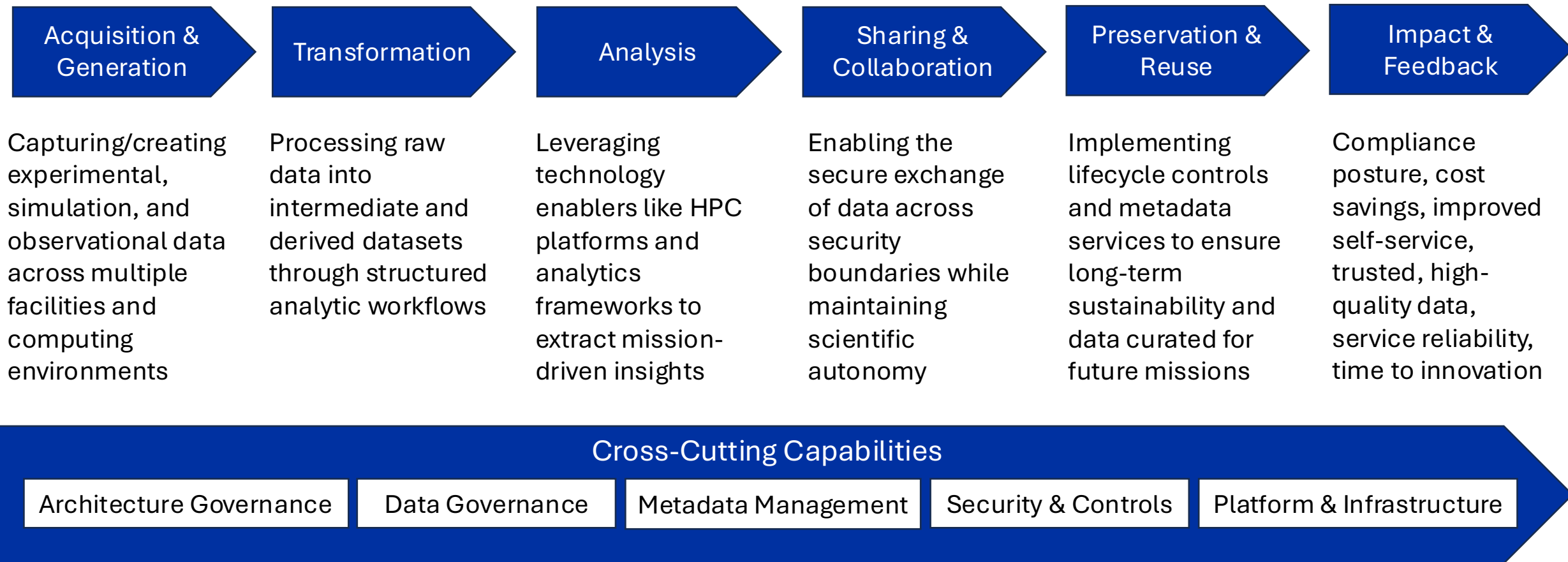
Focusing on Scientific Data Management Capabilities





Acquisition-to-Impact: Scientific Data Management

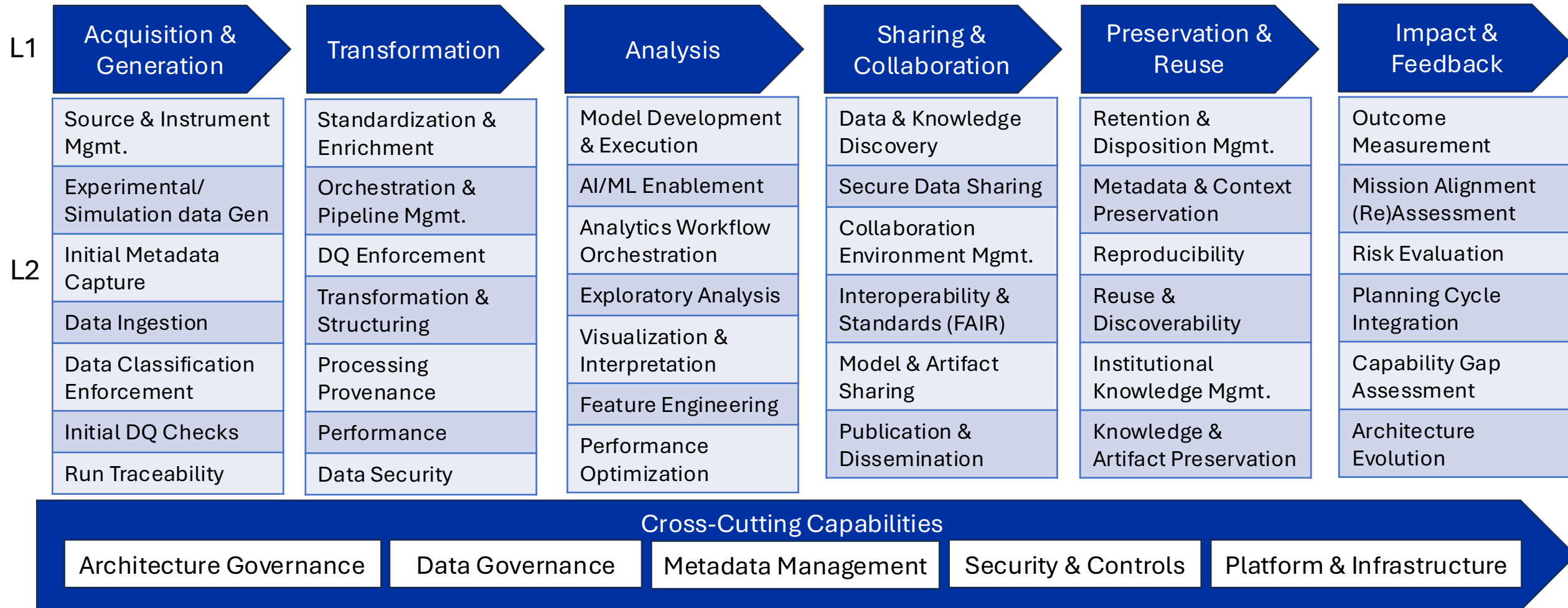
A high-Level definition of value chain pillars





Acquisition-to-Impact: Scientific Data Management

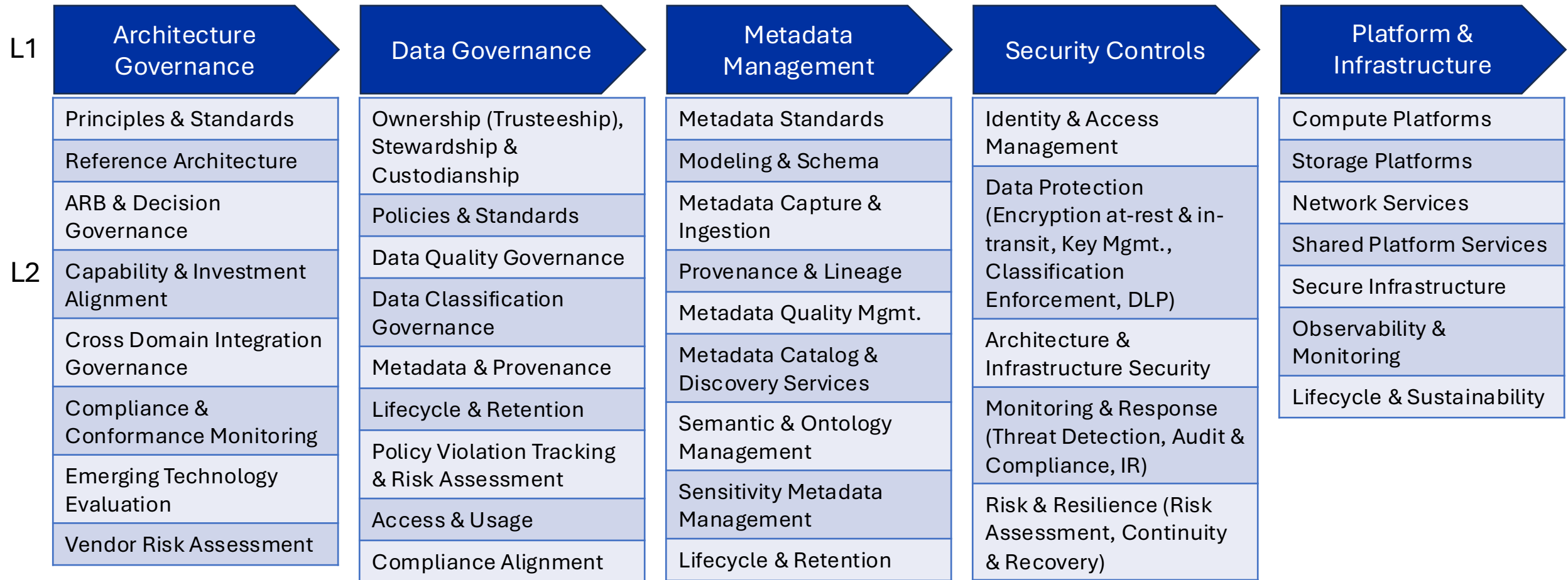
An Illustration of L1/L2 Decomposition of Canonical Value Chain





Cross Cutting Capabilities

An Illustration of L1/L2 Decomposition





Ownership Model

Enterprise vs. Domain Ownership

Enterprise-Owned (Stable, Shared)

- Capability model & reference architectures
- Core data platforms
- Metadata, lineage, identity, security services
- Governance frameworks & standards

Domain-Owned (Federated, Flexible)

- Scientific workflows
- Analysis codes & models
- Domain-specific data schemas
- Research-driven tooling (within guardrails)



Case Study: Direct Ink Writing

Goal: To optimize feedstock rheology and print parameters for high-performance structural components

Feature	“Siloeed” Project Approach	“Capability Driven” EA Approach
Data Capture	Manual logbooks; data stored on local instrument PC hard drives	L1 Capability: “Acquisition & Generation” Automated ingestion from sensors directly to research enclaves
Metadata	Minimal; often missing ambient humidity or batch numbers of the ink	L2 Capability: “Initial Metadata Capture” Mandatory “Metadata Tagging at Source” using enterprise schemas
Processing	One-off Python scripts on a student's laptop	L1 Capability: “Transformation” L2 Capability: “Orchestration & Pipeline Mgmt.” Standardized transformation pipelines
Sharing	Emailed CSVs or USB drives; high risk of version loss	L1 Capability: “Sharing & Collaboration” Secure, governed data exchange across security boundaries
Reuse	Data is “dark” once the paper is published	L1 Capability: “Preservation & Reuse” Curated datasets indexed in a searchable catalog



Case Study: Direct Ink Writing

Example: Using Capability model to map Technology Enablers and Data Domains

L1	L2	L3	Tech Enabler	Data Domain	Lifecycle Stage
Acquisition & Generation	Source & Instrument Mgmt.	Instrument Control	DQ Systems/ Python Scripts	Observational/ Experimental Data	Raw
Transformation	Orchestration & Pipeline Mgmt.	Event Driven Data Movement/ Integration	ETL/Data Processing Platform	Observational/ Experimental Data	Intermediate
Analysis	Model Development & Execution	ML Model Development	ML Platform	Derived Data	Derived
Sharing & Collaboration	Secure Data Sharing	Repository Submission	Darc/Scientific Data Repo	Derived Data	Curated



Highlights from DIW Cases Study

The Reality Gap: In a DIW experiment, the Research Engineer cares about the nozzle pressure, while the Data Architect cares about the provenance of that data. Without EA, those two worlds seldom meet.

From Raw to Curated: Under the old model, the 'Raw' sensor data from the printer stayed on the printer. In our new model, that raw data is systematically mapped to 'Intermediate' and 'Derived' domains, allowing an AI/ML model to later optimize the print path without starting from scratch.

The Governance Win: By establishing a 'Reference Architecture' for additive manufacturing data, we don't have to reinvent the security and storage controls for every new printer we buy. We simply plug the new instrument into the existing capability.

Key Takeaway: By shifting to a capability driven approach, we move the Materials Engineering Division from 'managing files' to 'orchestrating knowledge.' This reduces the time spent on data wrangling and increases the time spent on scientific discovery.



Why Does it Matter? Why Now?

- **"Why":** Just as a business cannot track revenue without a Lead-to-Cash process, a lab cannot track scientific impact without a formal **Data Value Chain**. This model allows us to reason about architectural needs consistently across diverse domains
- **Breaking Silos:** By defining these capabilities, we stop building '**Project X's Database**' and start building an '**Enterprise Metadata Service**' that supports every program at the lab and across sites
- **Governance: This isn't about control; it's about enabling autonomy.** Level 3 capabilities can be implemented locally, but Level 1 and 2 should follow enterprise reference architectures to ensure interoperability

*The **scale, complexity, and diversity** of scientific data is accelerating. The old project-based models weren't designed for the era of AI-driven material science or multi-facility workflows. This shift to an **Architecture-Driven Approach** isn't just a "nice to have"—it is a requirement for long-term sustainability*



Path Forward

Call to Action for NSE Leadership

1. Adopt the Capability-Driven Abstraction

- **Action:** Shift **funding** and **oversight models** from "project-specific silos" to "enterprise-level capabilities"
- **Rationale:** This provides a structured layer between mission-driven objectives and technical implementations, ensuring technology survives organizational shifts

2. Institutionalize Architecture Governance

- **Action:** Empower a **cross-lab architecture board** to establish reference architectures and standards
- **Why:** It balances local scientific autonomy with enterprise-wide consistency, reducing risk and fragmentation

3. Commit to Distributed Data Stewardship

- **Action:** Mandate coordinated data governance that defines ownership and metadata practices across all scientific domains
- **Why:** It ensures data from experimental and simulation workflows remains a searchable, reusable asset across security boundaries

4. Prioritize Systematic Technology Alignment

- **Action:** Use the capability model to guide technology investment decisions in compute, storage, and analytics
- **Why:** This reduces duplication of effort and ensures infrastructure is directly aligned with mission-driven science

Thank You!!

Reference Slides



Scientific Data Domains

Data Categorization for Effective Data Management

Definition:

“Scientific Data Domains” refer to logical groupings of data based on **characteristics, structure and usage**. Scientific data covers a broad spectrum (lab experiments, observations, simulations etc.). Categorizing scientific data into domains will help optimize management and governance of data across functional areas.

Scientific Data Domains

Data Domains	Definition
Observational/Experimental Data	Data collected from observations of natural phenomenon or generated under controlled conditions to test hypotheses
Simulation/Modeling Data	Data generated by computer models simulating real-world processes
Sensor/IoT/Time Series Data	Data from scientific instruments over time
Aggregated Data	Data that has been summarized or combined from multiple sources, raw measurements and/or time intervals
Derived Data	Data that has been computed, contextualized, evaluated or inferred from raw or aggregated data through modeling, statistical analysis, or transformations

Scientific Domains: Master vs. Transactional Data

Decomposing Scientific Data Domains into **Master** and **Transactional** components is critical for the following reasons:

- Allows for enhanced capture and enrichment of metadata
- Design and implementation of robust data cataloging solution
- Improved data management and governance

	Master Data	Transactional Data
Definition	Core, relatively static reference data used to define scientific entities	Dynamic, temporal data from scientific experiments or activities
Volatility	Slowly changing, stable over time	Frequently changing over time
Purpose	Provides context, structure and identity to transactions	Drives analysis, supports hypotheses, records events/processes



Additive Manufacturing

Direct Ink Writing (DIW) Data Domains

AM Functional Domains	Data Domains	Data Sub Domains
Design & Modeling	Observational/Experimental Data	CAD Model library, templates, lattice structures Toolpaths for DIW ink Design Rules Simulation Results
Materials Science	Observational/Experimental Data	Rheological Properties (i.e. ink viscosity, solvent evaporation rate, yield stress, particle loading, curing kinetics, shelf-life) Supplier, lot number, certifications Recycling history (% reuse) Storage and handling specifications Mixing/dispensing records per build
Process	Observational/Experimental Data	Nozzle Geometry (i.e. diameter , nozzle shape, length-to-diameter ratio) Machine Code/G-code (pitch) Machine ID, printer configuration Standard Operating Parameter ranges (extrusion pressure, temperature, layer thickness) Machine alarms/events/alerts Print Job Logs Toolpath Execution Trace Nozzle Performance Outcomes
Quality & Inspection	Observational/Experimental Data	Testing Protocols Acceptance Thresholds (i.e. Porosity %) Inspection Results (dimensions, surface roughness measures) CT images, Defect Maps Monitoring Logs (cameras, sensors)



Direct Ink Writing (DIW) Data Domains

Decomposing Into Master and Transactional Data Elements

AM Functional Domains	Master Data	Transactional Data
Design & Modeling	CAD Model library, templates, lattice structures Design Rules	Simulation Results Toolpaths for DIW ink Build Orientation
Materials Science	Ink Formulation IDs, Codes Rheological Properties (i.e. ink viscosity, solvent evaporation rate, yield stress, particle loading, curing kinetics, shelf-life) Supplier, lot number, certifications Storage and handling specifications	Recycling history (% reuse) Mixing/dispensing records per build Environmental Storage Logs (humidity, temperature)
Process	Nozzle Geometry (i.e. diameter , nozzle shape, length-to-diameter ratio) Machine Code/G-Code (pitch) Machine ID, printer configuration Standard Operating Parameter ranges (extrusion pressure, temperature, layer thickness) Control Software Version	Machine alarms/events/alerts Toolpath Execution Trace Print Job Logs Nozzle Performance Outcomes Sensor Data (flow stability, torque, pressure curves)
Quality & Inspection	Testing Protocols Acceptance Thresholds (i.e. Porosity %)	Inspection Results (dimensions, surface roughness measures) CT images, Defect Maps Monitoring Logs (cameras, sensors) Mechanical Test Data (strength, adhesion)