Barriers in the Research Data Lifecycle: From Collection to Reuse

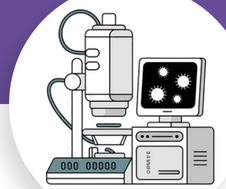
Roberto dos Reis

The Scientific Data Lifecycle

01

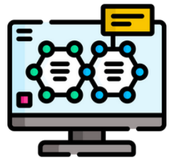
PLANNING & DESIGN

02



ACQUISITION

03



PROCESSING & ANALYSIS

04

CURATION & PRESERVATION

05

SHARING & REUSE

- Barriers Compound: Gaps at each stage amplify downstream - e.g., poor planning hinders FAIR sharing
- Legacy Systems: Many workflows still rely on disconnected, manual, vendor-specific tools
- At-Risk Data: Without persistent IDs or metadata, up to 80% of scientific datasets become unusable in 5 years

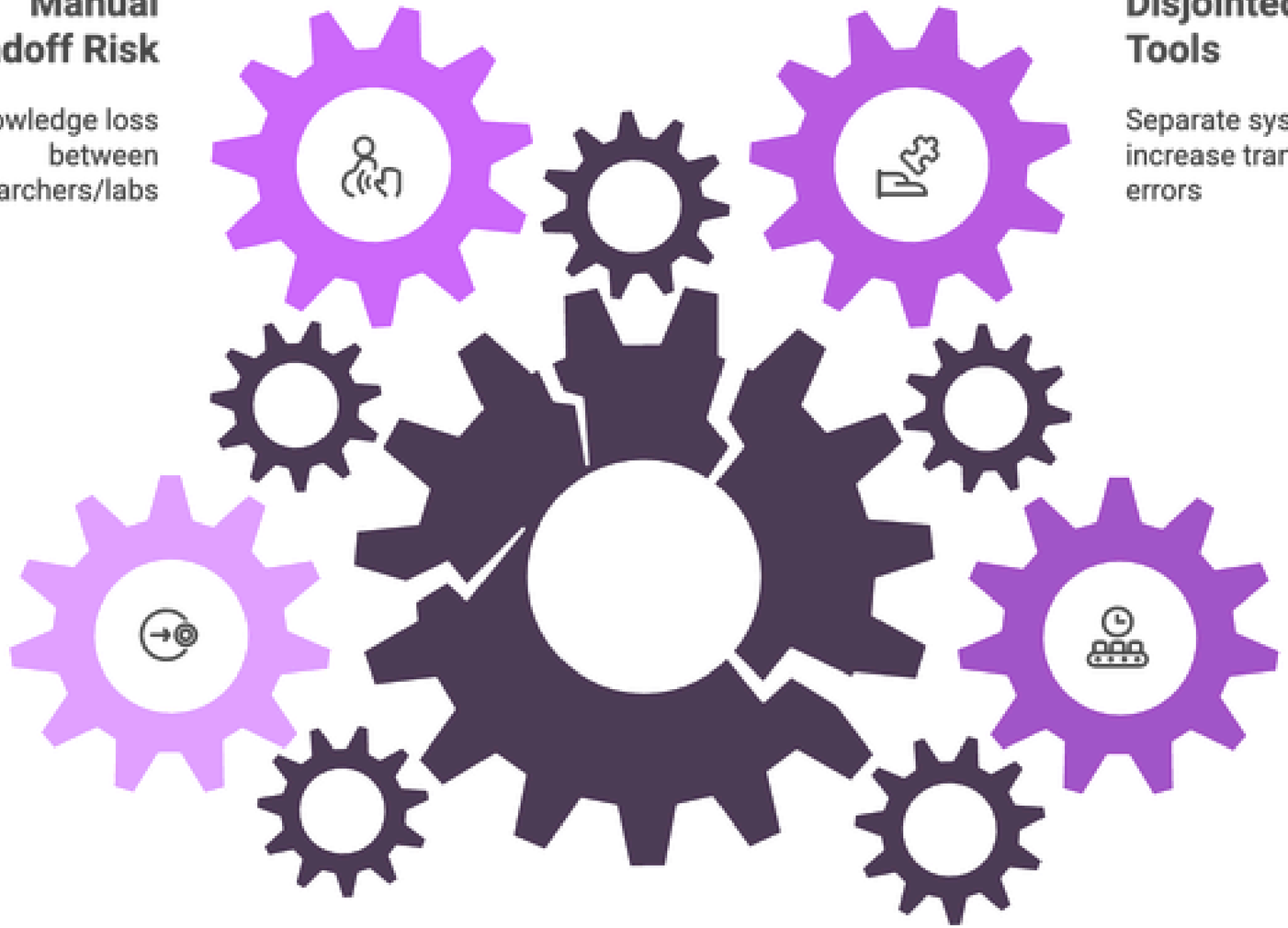
Fragmentation Across Lifecycle Stages

Manual Handoff Risk

Knowledge loss between researchers/labs

Disjointed Tools

Separate systems increase transition errors



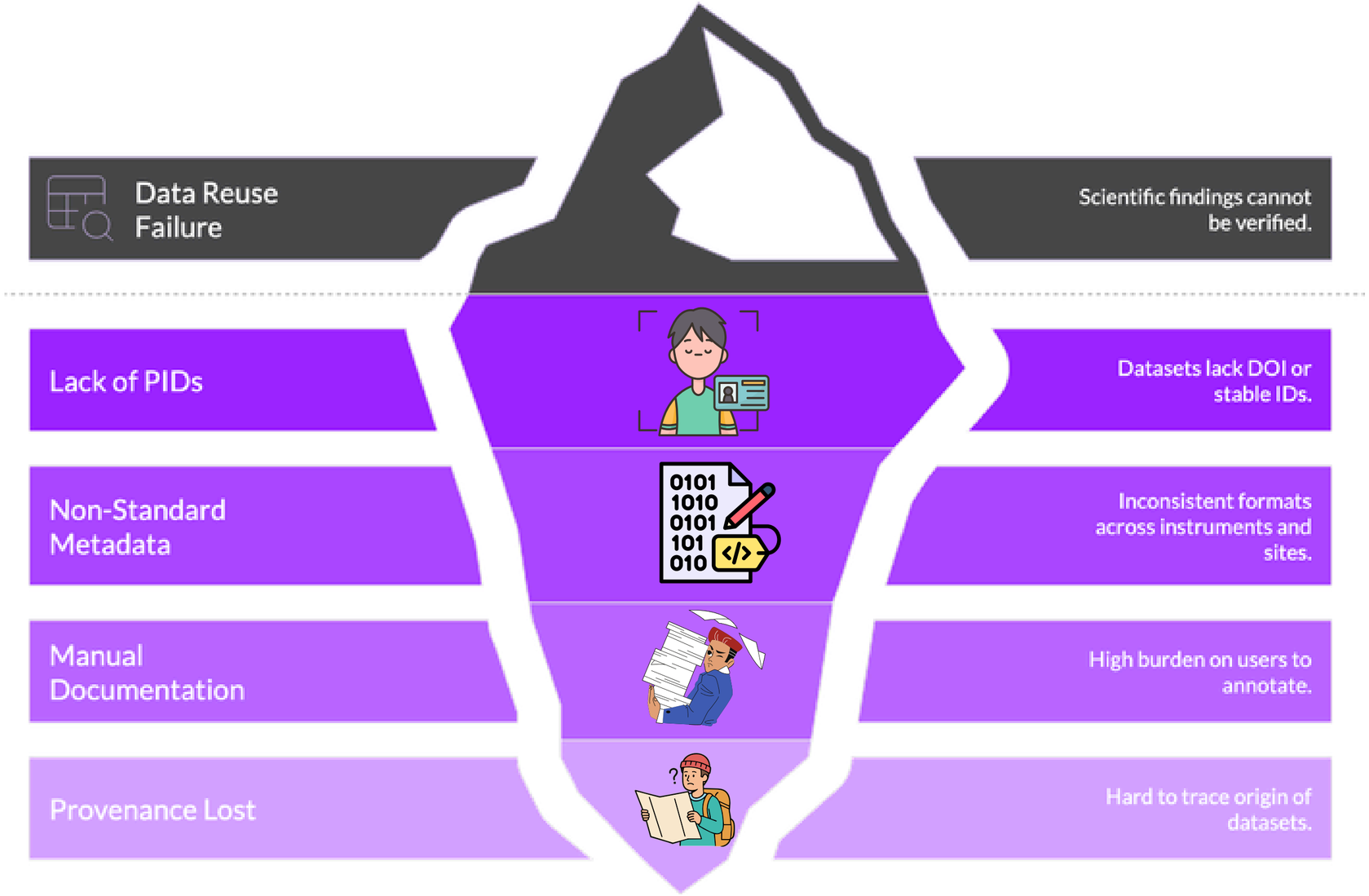
No Closed Loop

Analysis insights don't inform planning

Sequential Processing

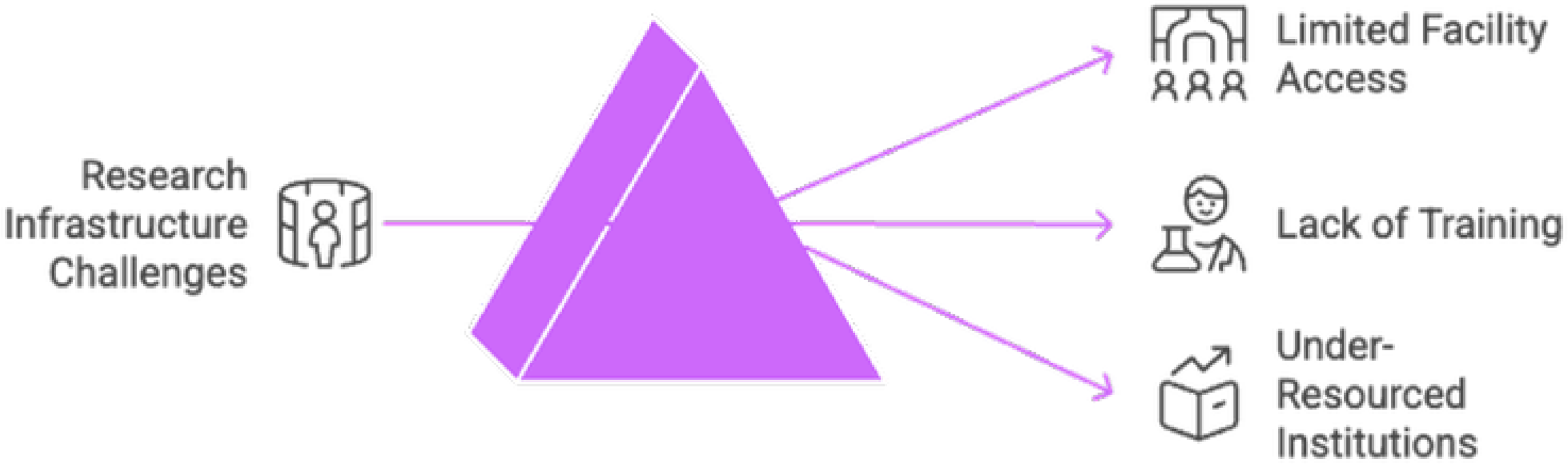
Linear data flow causes delays

Metadata Inconsistency & Provenance Gaps



Silos and Institutional Inequity

Access Barriers Limit Research Participation



Manual Curation & FAIR Compliance Burden

Unsustainable Data Stewardship Models

Expertise Required

Users need FAIR training

Low Adoption

Outputs lack FAIR compliance

Petabyte Problem

Data scale overwhelms processes

Manual Metadata

Time-consuming, error-prone process

Reuse & Discoverability Deficiencies

Data Exists – But Can't Be Found

Limited Search Capabilities

AI-Optimized Search

No Semantic Discovery

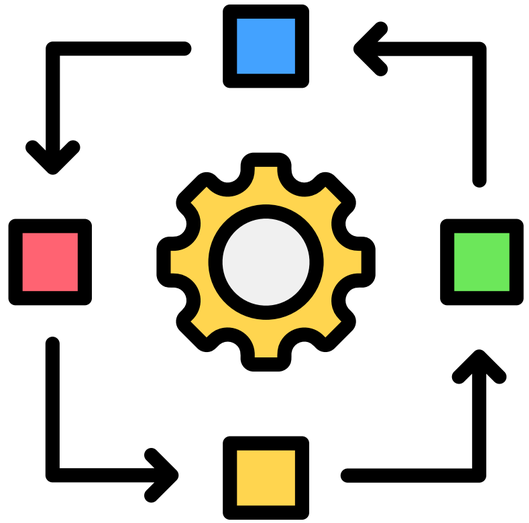
Semantic Data Discovery

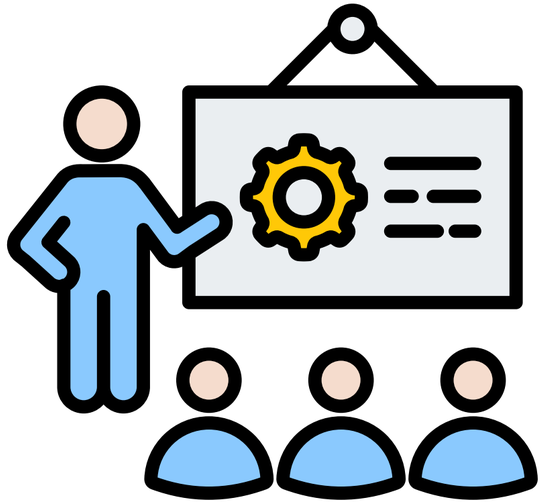
No AI Optimization

AI Optimization

Incompatible Instrumentation & Protocols

- **Vendor Lock-In:** Every platform (JEOL, Thermo, Hitachi, etc.) uses different control protocols
- **Non-Portable Workflows:** Protocols and settings are not transferrable
- **Fragmented Training:** Users must retrain for each platform





Implications for NUANCE

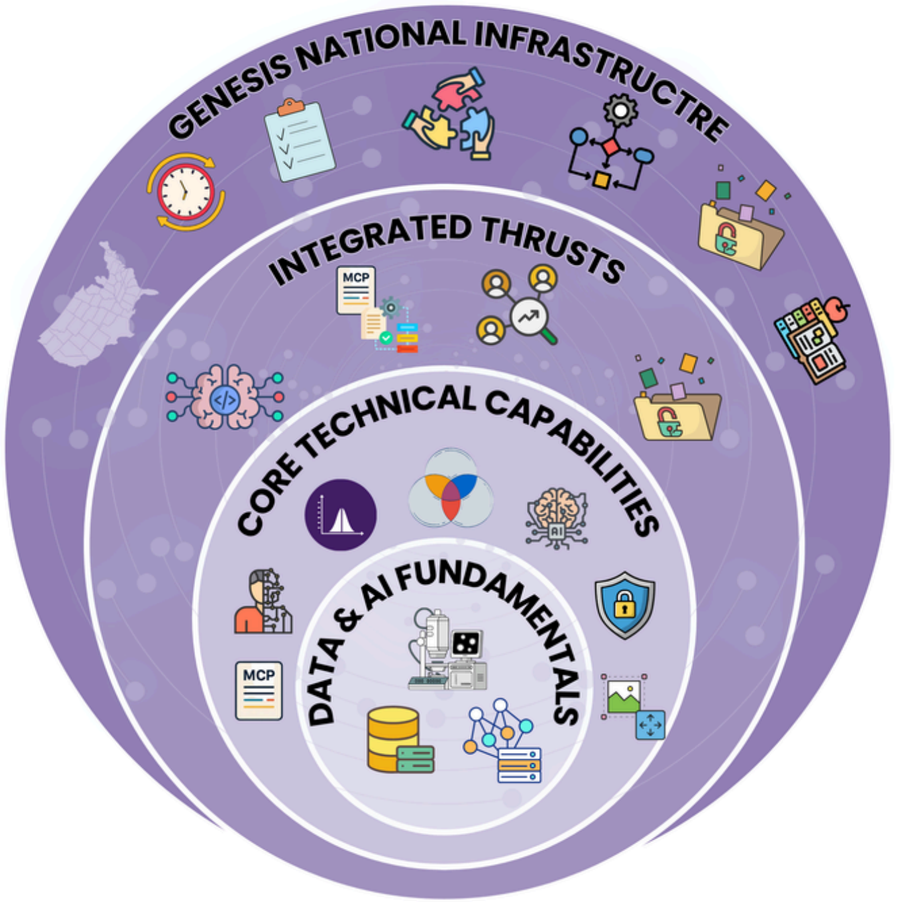
What's at Stake

- **Slowed Discovery:** Long time-to-data → long time-to-insight
- **Wasted Infrastructure:** National investment not fully utilized
- **Equity Gaps Widen:** Under-resourced institutions fall further behind
- **Compliance Risks:** Failing to meet FAIR/open mandates can threaten funding

Lost Datasets = Lost Science

Solutions to Lifecycle Barriers

Systemic Innovation Across the Stack



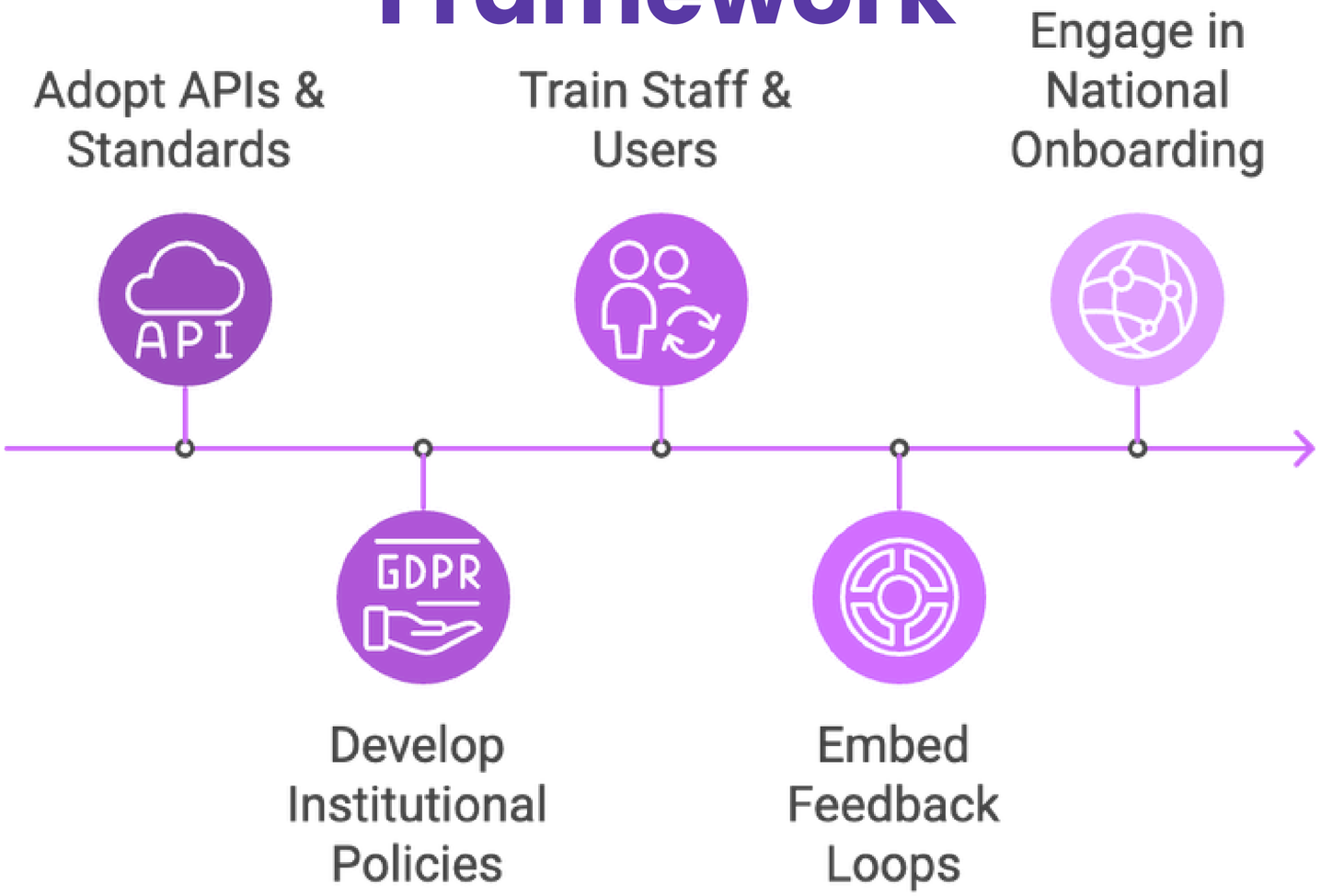
AI-Guided Design: Reduces trial-and-error and planning burden

Universal Instrument Control: Vendor-agnostic, secure, and remote

Federated, Compressed Analysis: Petabyte-scale computation and similarity search

Automated FAIR Compliance: Real-time metadata, DOIs, provenance, auditability

Pathways to a Robust Lifecycle Framework



Next Steps

01

AI-Guided Design

Automated protocol optimization, multi-modal technique coordination, real-time safety validation.

02

Universal Control

Different instrument types supported, automated quality assurance, zero equipment damage (AI safety).

04

Automated FAIR

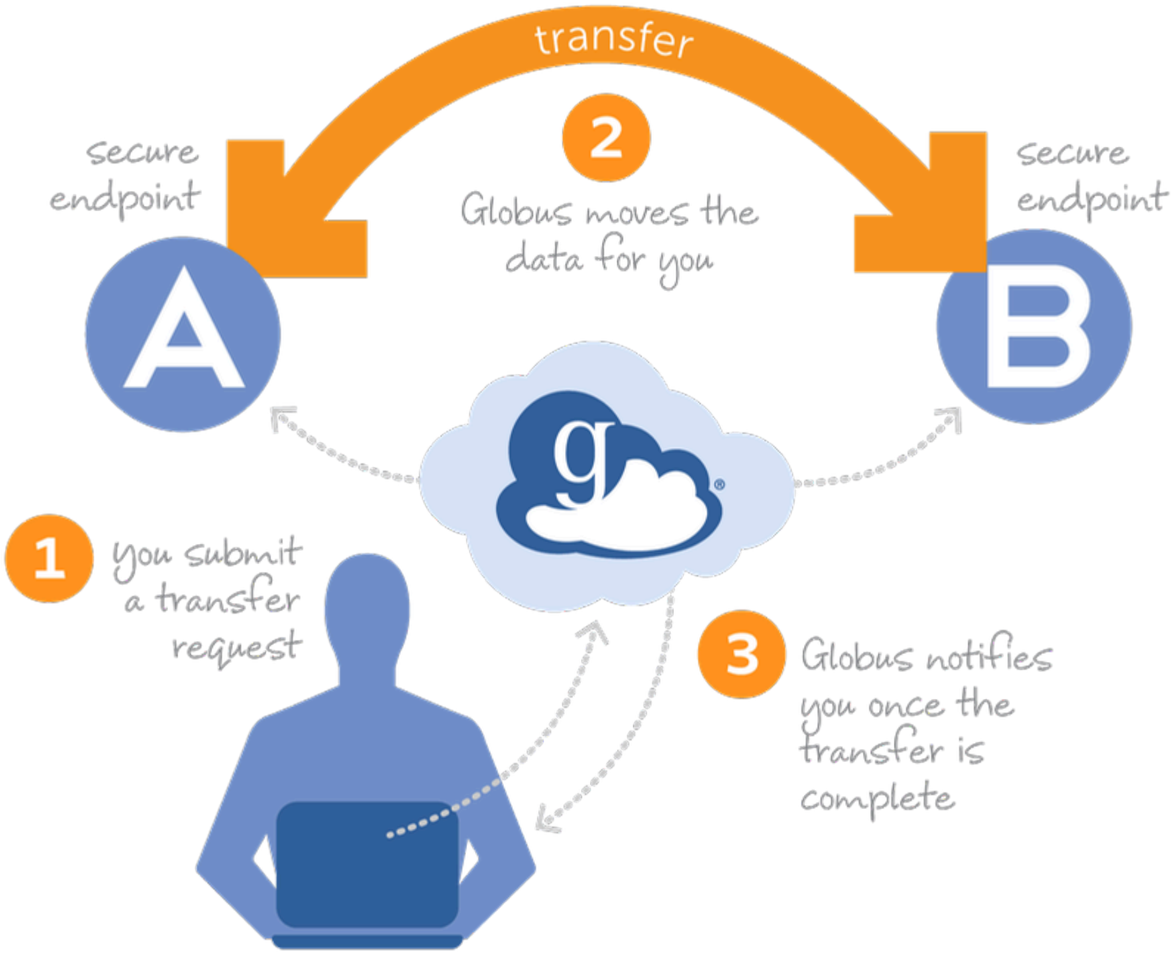
Standardized vocabularies, cryptographic provenance, and cross-institutional discovery.

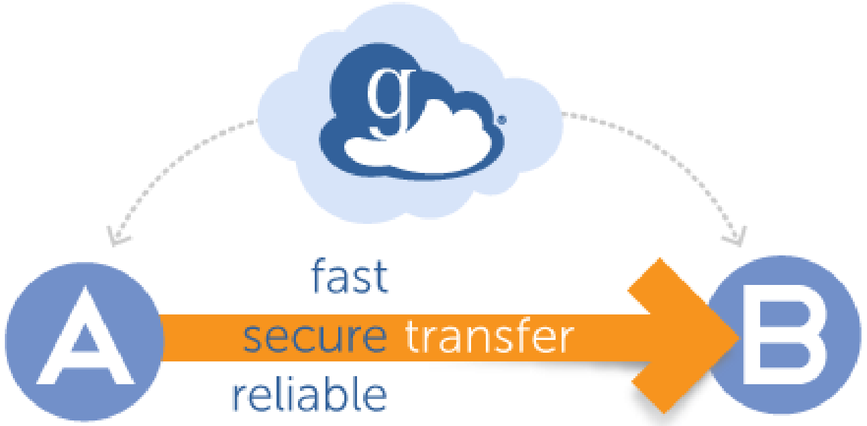
03

Federated Analysis

100-1000x semantic compression, multi-modal data fusion, and continuous AI model improvement.

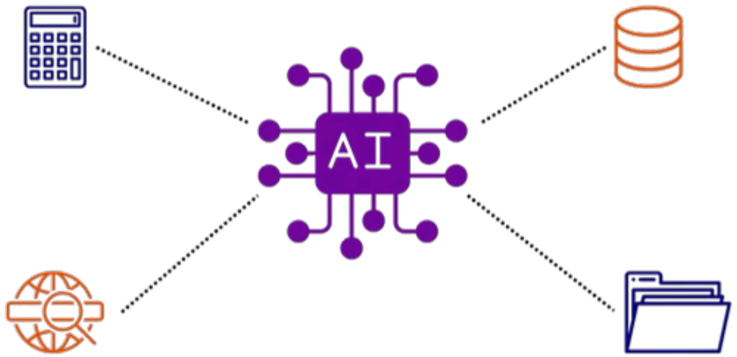
Next Steps


transfer
2
Globus moves the data for you
secure endpoint
secure endpoint
A
B
1
you submit a transfer request
3
Globus notifies you once the transfer is complete

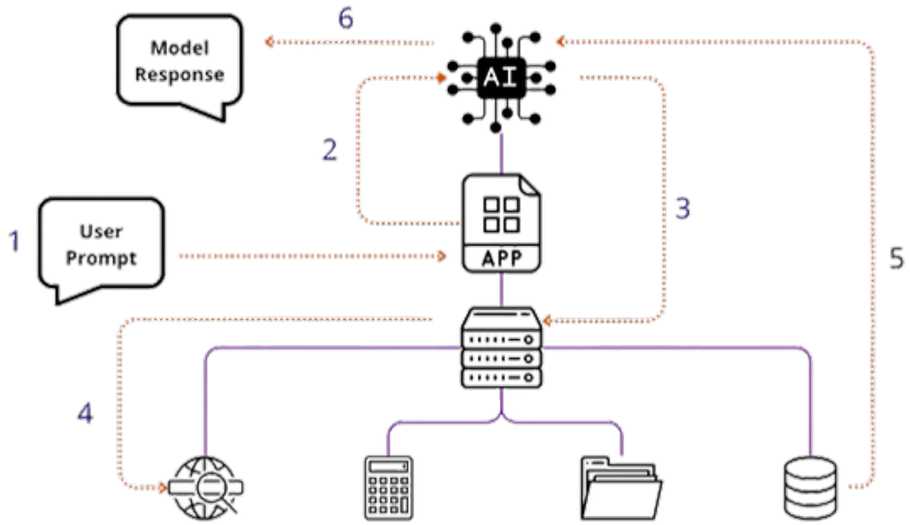

fast
secure transfer
reliable
A
B

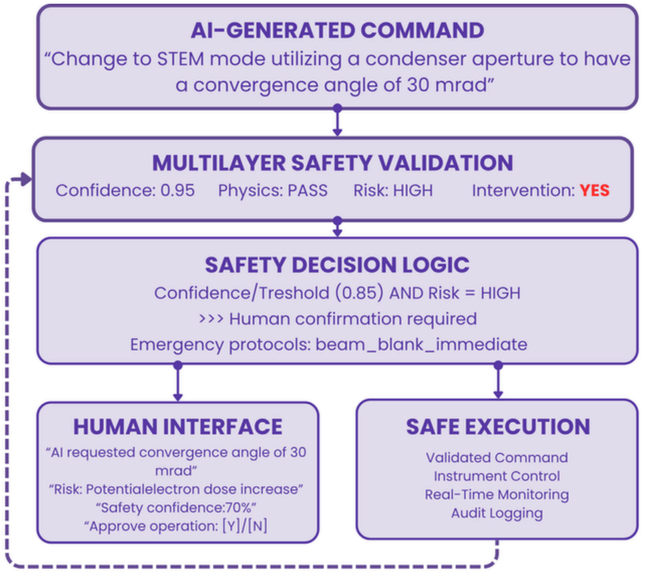
Next Steps

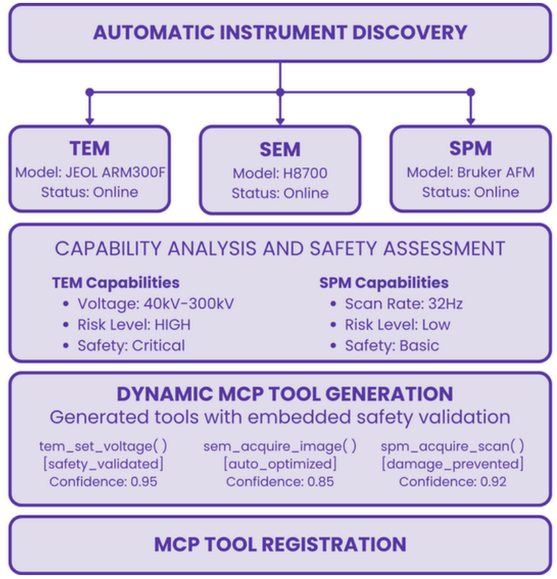
An open protocol for connecting AI models to tools, APIs, and external context.



MCP acts as a bridge between models and the outside world.







Next Steps

01

Local Acquisition & Watermarking

02

Transfer to Institutional RDSS

03

Cloud Backup & Processing

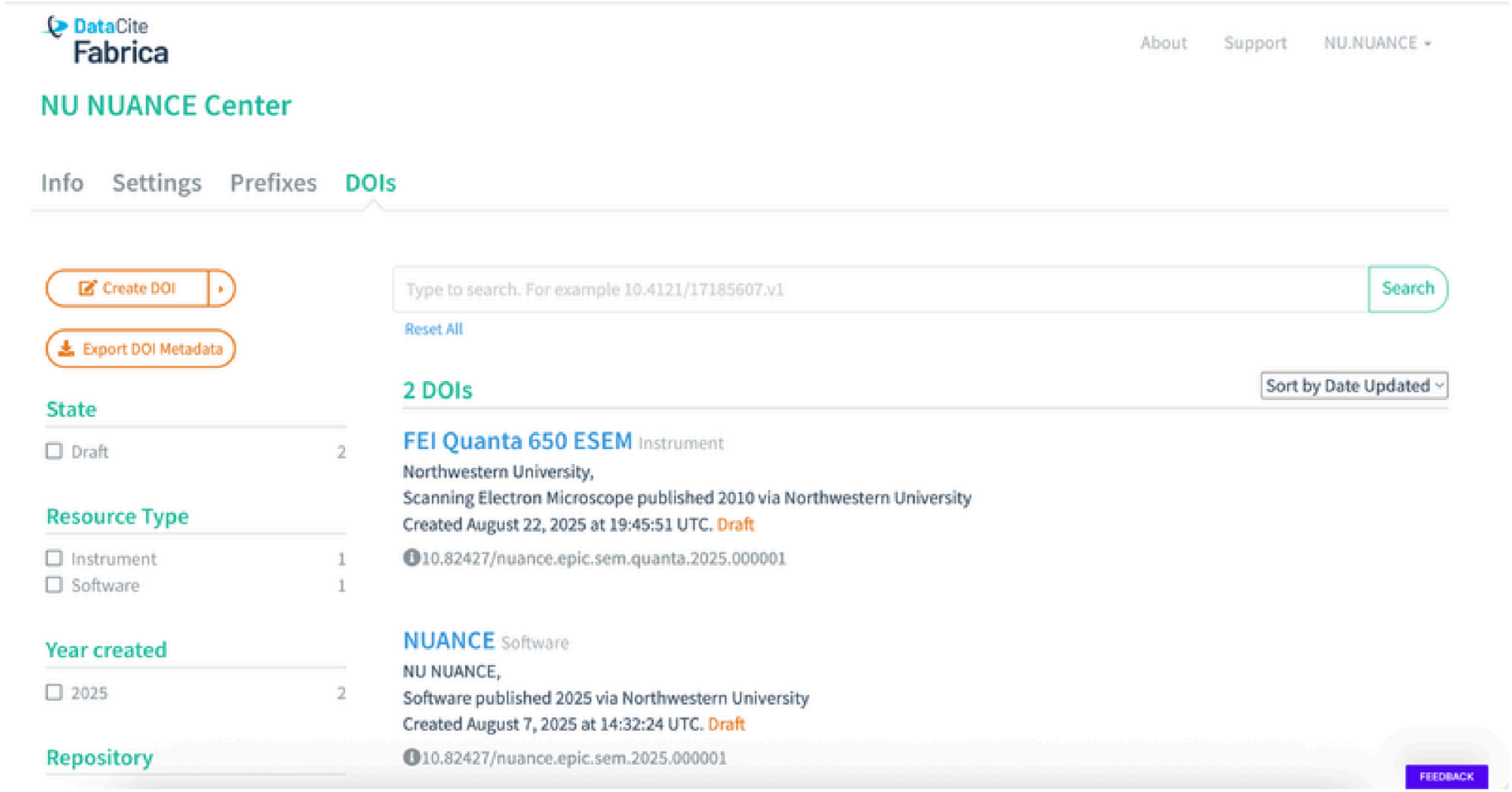
04

Provenance Registration

05

Verification Protocol

Next Steps


DataCite
Fabrica
About
Support
NU.NUANCE
NU NUANCE Center
Info
Settings
Prefixes
DOIs
Create DOI
Export DOI Metadata
Type to search. For example 10.4121/17185607.v1
Search
Reset All
2 DOIs
Sort by Date Updated
State
Draft 2
Resource Type
Instrument 1
Software 1
Year created
2025 2
Repository
FEI Quanta 650 ESEM Instrument
Northwestern University,
Scanning Electron Microscope published 2010 via Northwestern University
Created August 22, 2025 at 19:45:51 UTC. Draft
10.82427/nuance.epic.sem.quanta.2025.000001
NUANCE Software
NU NUANCE,
Software published 2025 via Northwestern University
Created August 7, 2025 at 14:32:24 UTC. Draft
10.82427/nuance.epic.sem.2025.000001
FEEDBACK

Next Steps

Data and Metadata Standards

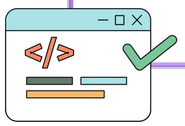
Ensures data consistency and interoperability

Storage, Preservation, and Access

Provides secure and accessible data storage solutions

Compliance, Timeline, and Sustainability

Aligns with NSF guidelines and ensures long-term viability



Data Management and Sharing Plan

Acquisition, Processing, and Curation Workflows

Streamlines data handling from collection to archiving

Federated Analysis, Privacy, and Security

Enables collaborative analysis while protecting data integrity

Conclusion & Next Steps

Barriers Are Real – But Solvable:
With the right architecture, culture, and automation

NU*ANCE* at the Vanguard:
Already pioneering solutions, now scaling them

Your Role:
Help shape an equitable, AI-ready, interoperable future of data-intensive science

HAPPY HOLIDAYS

roberto.reis@northwestern.edu