Putting Data in the Spotlight

The case for a Scientific Data Federation

Swen Boehm[1]
Suhas Somnath[2]
Olga Kuchar[2]

May 27, 2022

ORNL is managed by UT-Battelle, LLC for the US Department of Energy

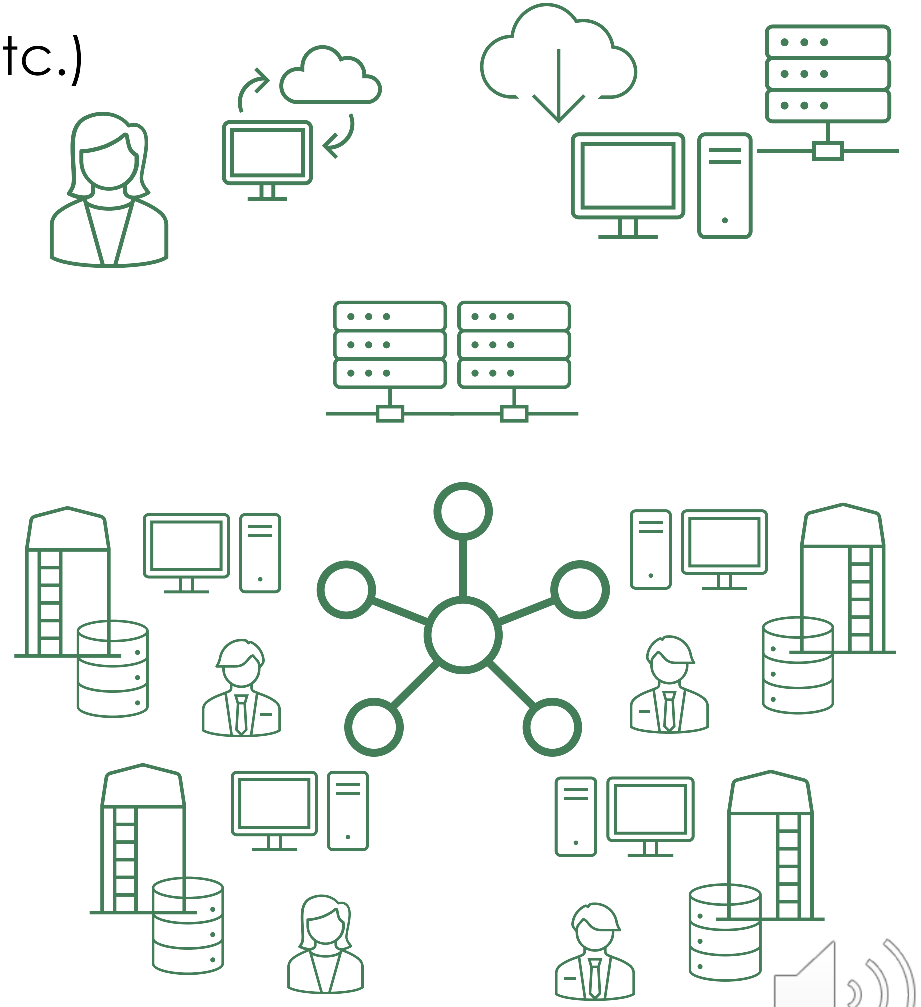
1) Intelligent Systems and Facilities Group,
Computer Science and Mathematics Division, ORNL
2) National Center for Computational Sciences,
Oak Ridge Leadership Computing Facility, ORNL

The Science is in the data

- Growing amount of data produced by sensors, data analytics, and simulations
- Harnessing insights from data requires sophisticated infrastructure and tools
- Metadata augments the data sets with additional information
- Provenance data increases trust in the data
- Lab wide Data Management solution to improve the work with data

Status Quo in Today's Data Infrastructure

- Infrastructure often managed by facilities (SNS, OLCF, etc.)
- Tools typically developed independently by scientists
  - One-off or ad hoc solutions, often reinventing the wheel
- Development limited to scope of projects Manual data movement
- Leads to data silos
- Risk of data loss
- Incompatible or incomplete metadata

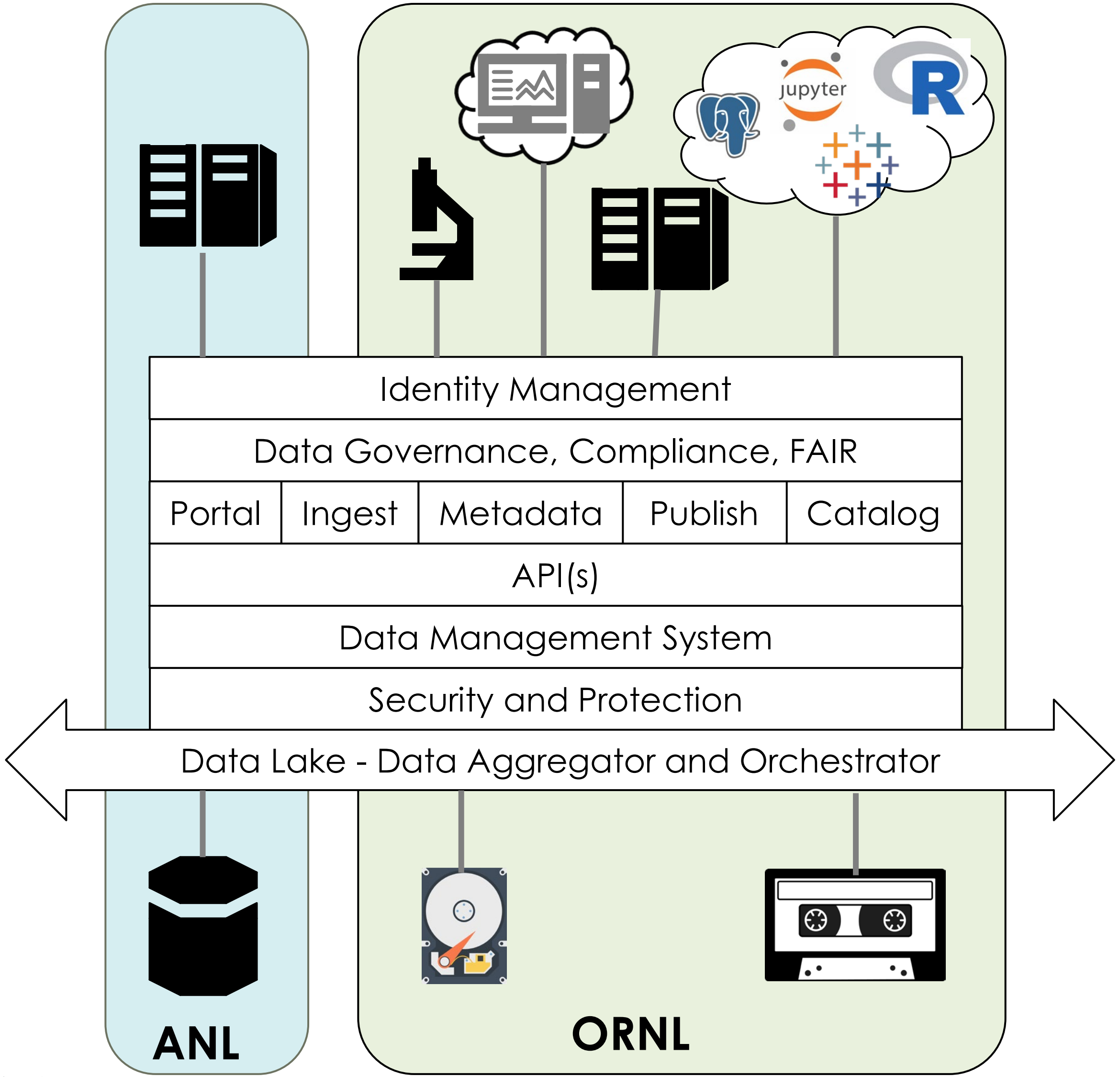


OAK RIDGE
National Laboratory

How to change the status quo

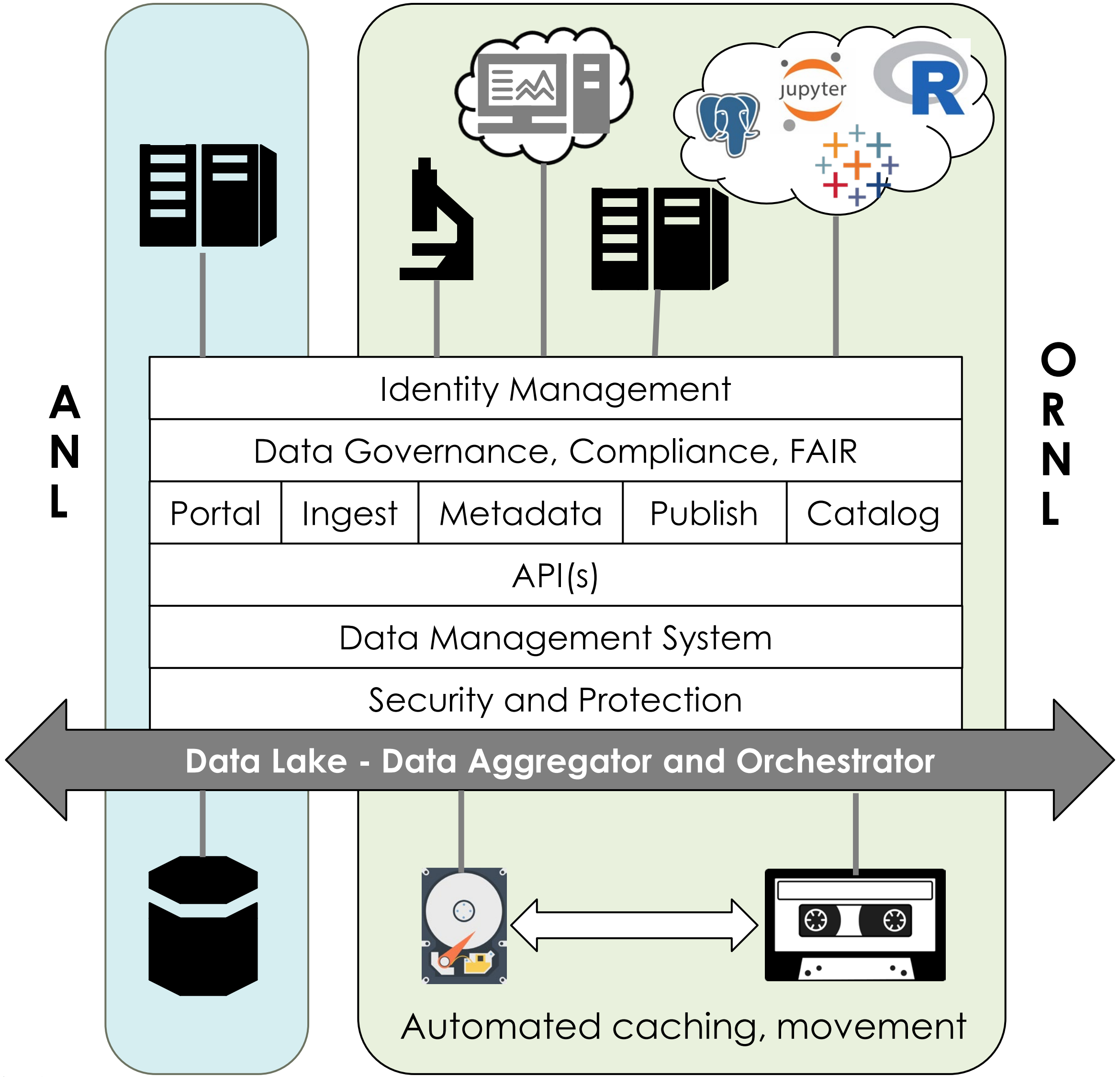
?

Scientific Data Federation



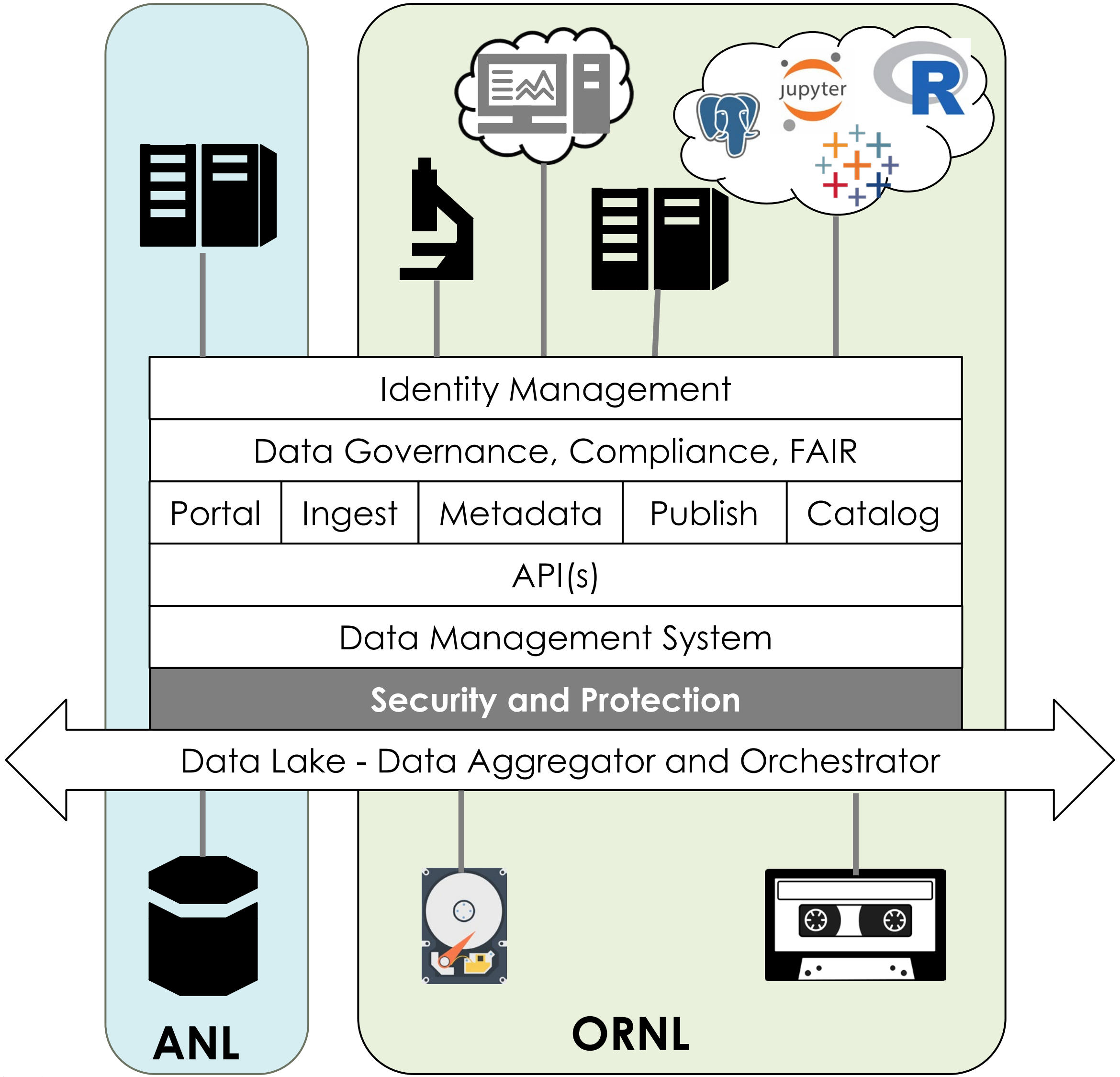
- Tools and a dedicated infrastructure for scientific data
- All facilities at ORNL / DOE complex under a unified umbrella
  - Collaboration opportunities
- Data seamlessly connected to computational resources, analytics platforms, and ancillary services
- Establish common data and metadata standards & data governance

Data Aggregation and Orchestration



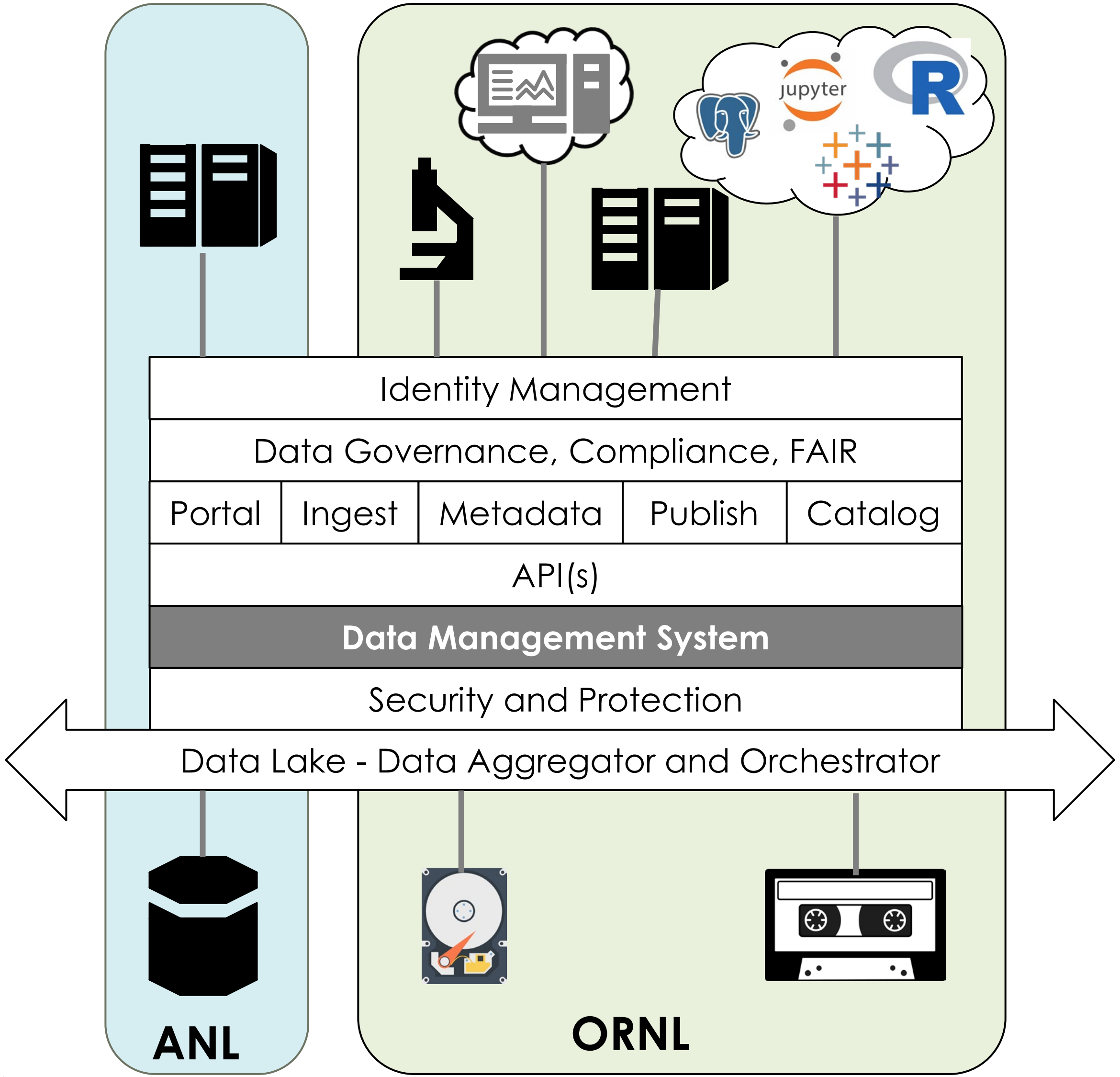
- Mesh all storage solutions into a cohesive data lake
  - File-systems, object stores, etc.
  - Across DOE complex
  - Abstract storage infrastructure related complexities
- Automatic and intelligent data migration
  - Caching frequently used data
  - Archiving stale data
- Efficient data transfer

Security and Protection



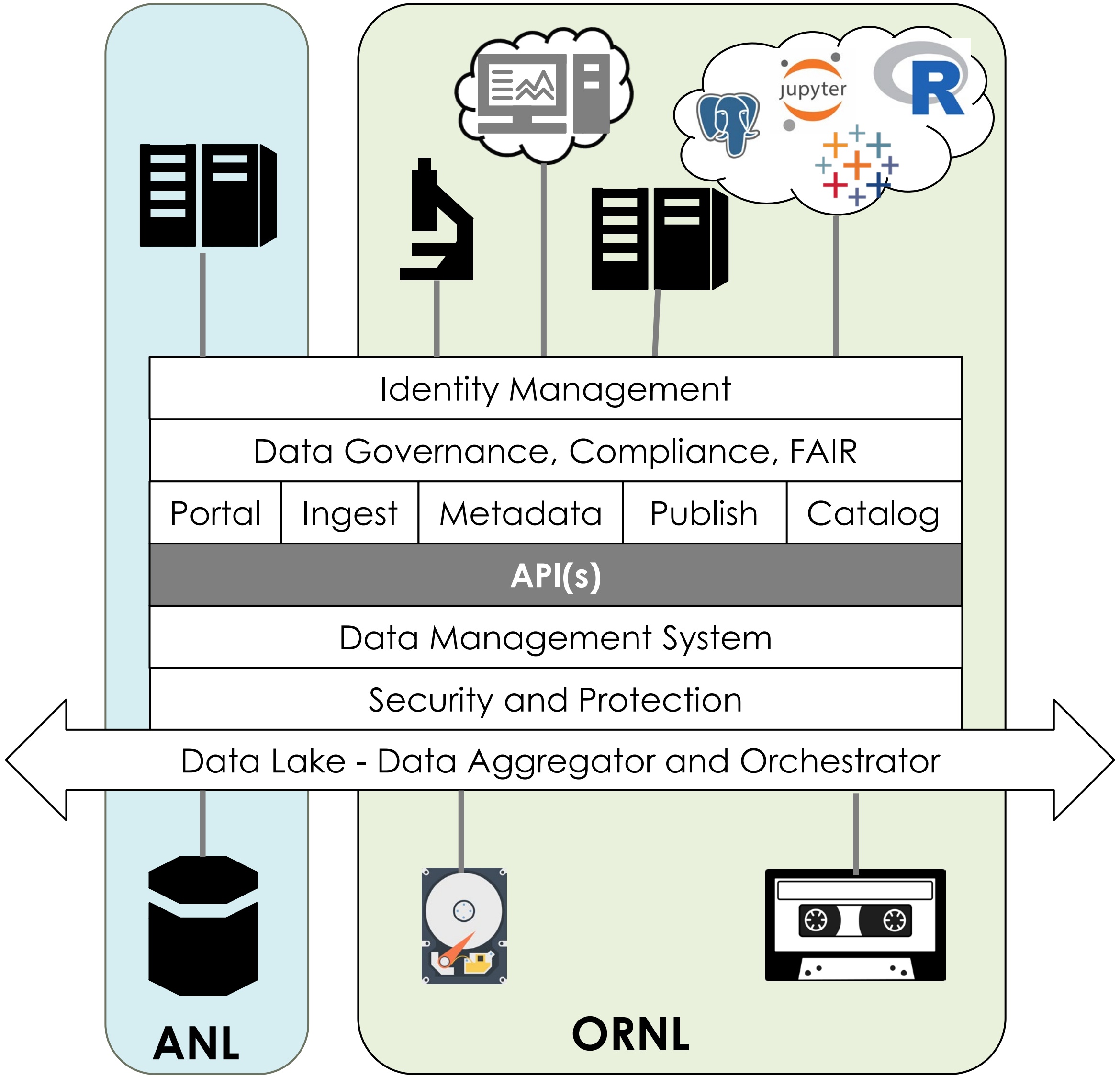
- Reflect and enforce policies (e.g., for sensitive data).
- Enable data to be securely accessible at computing resources (e.g. Citadel at ORNL)
- Track and respect security clearance
- Keep data private by default

Data Management System



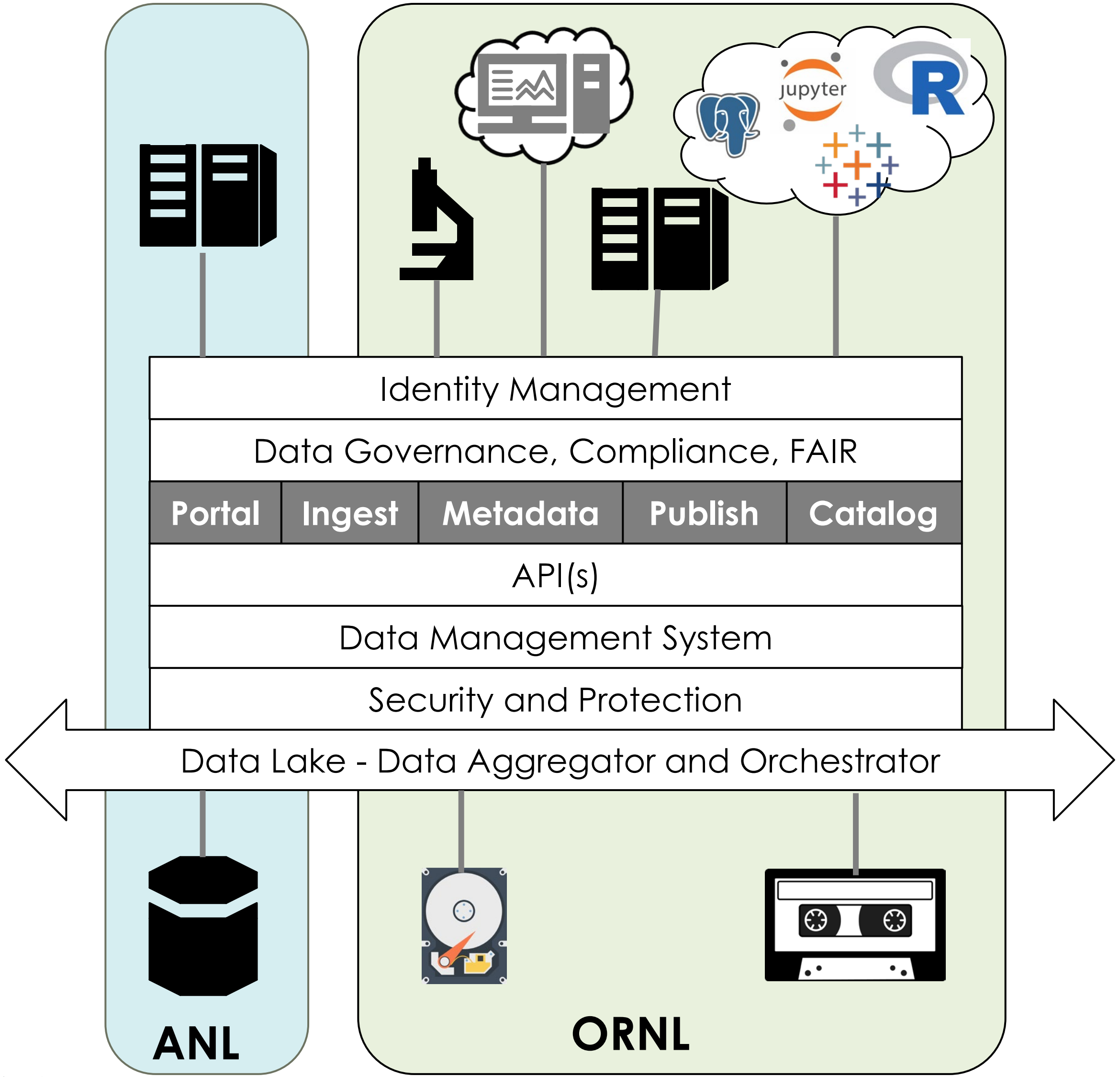
- Unique & persistent identifiers for data
- Abstract file-system-specific complexities
- Capture rich metadata and provenance
- Enable sharing, organization, search and discovery of data across federation
- Support data versioning
- Locate datasets in the federation
- Track physical entities like samples
- Fine-grained access controls

Programming Interfaces



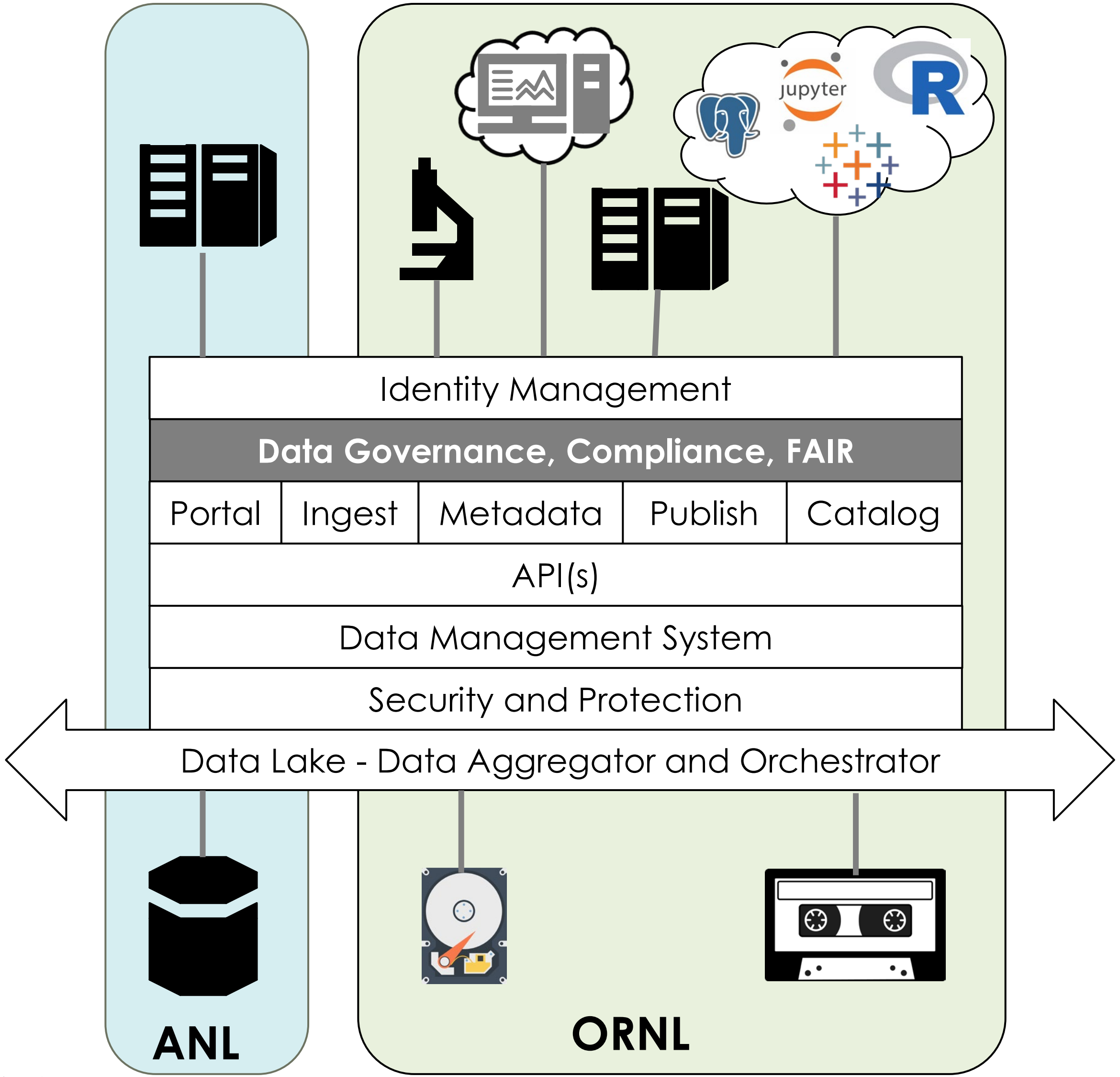
- Intuitive and user-friendly
  - E.g. REST, CLI, Python, etc.
- Accessible everywhere (personal computers, cloud computing, and high-performance computing)
- Consistent interface regardless of
  - underlying storage hardware
  - where data is accessed (e.g. personal computer, cloud…)
- Support cross-facility data pipelines, domain-specific applications

Data Applications



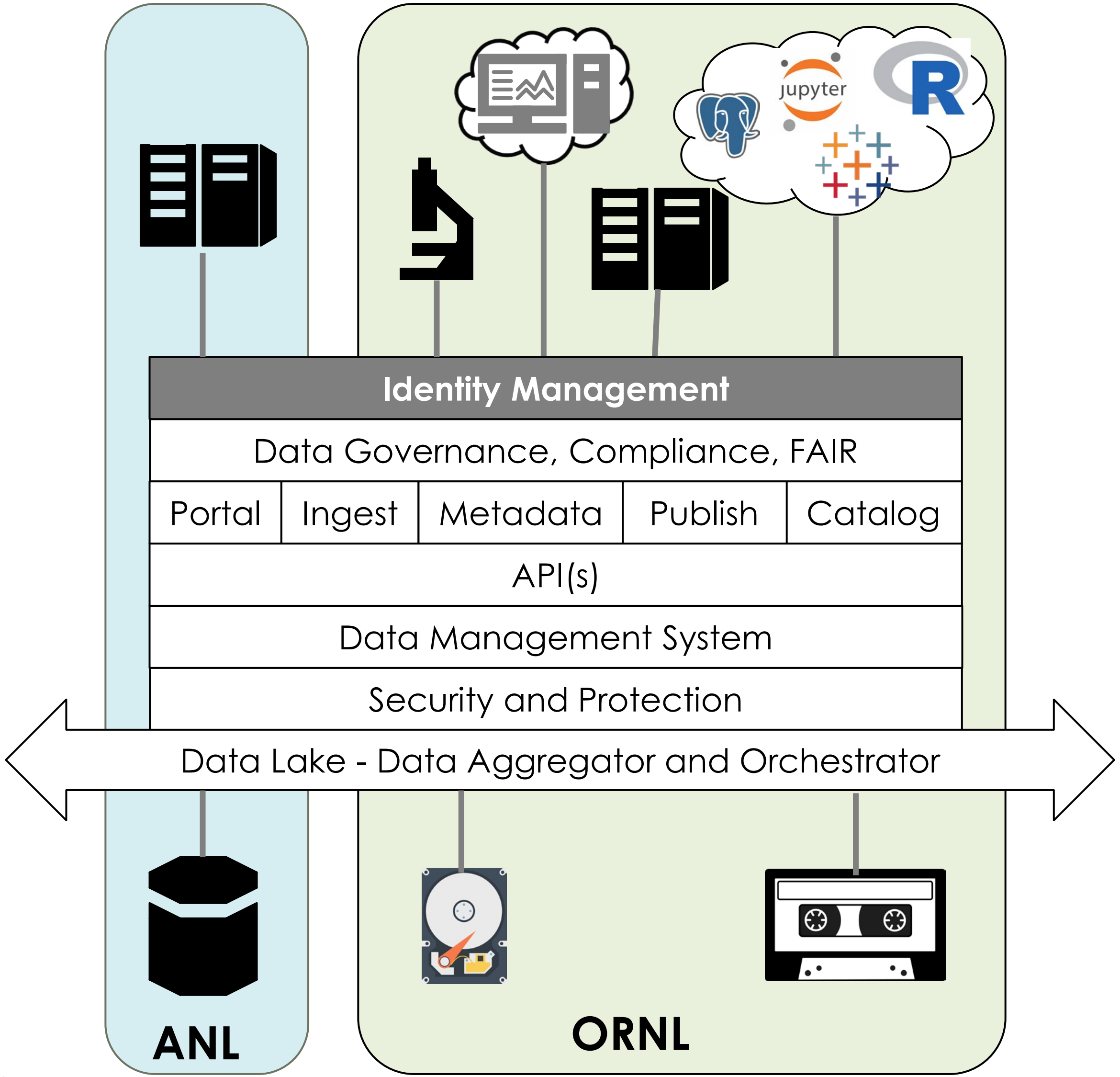
- User-friendly and extensible web-application for data management
- Extract metadata from
  - myriad of different file formats
  - Processes and workflows
- Develop / adopt metadata standards
  - Tools to standardize metadata
- Intuitive data publication interface
- Discover data via catalogs

Governance, Policies and Standards



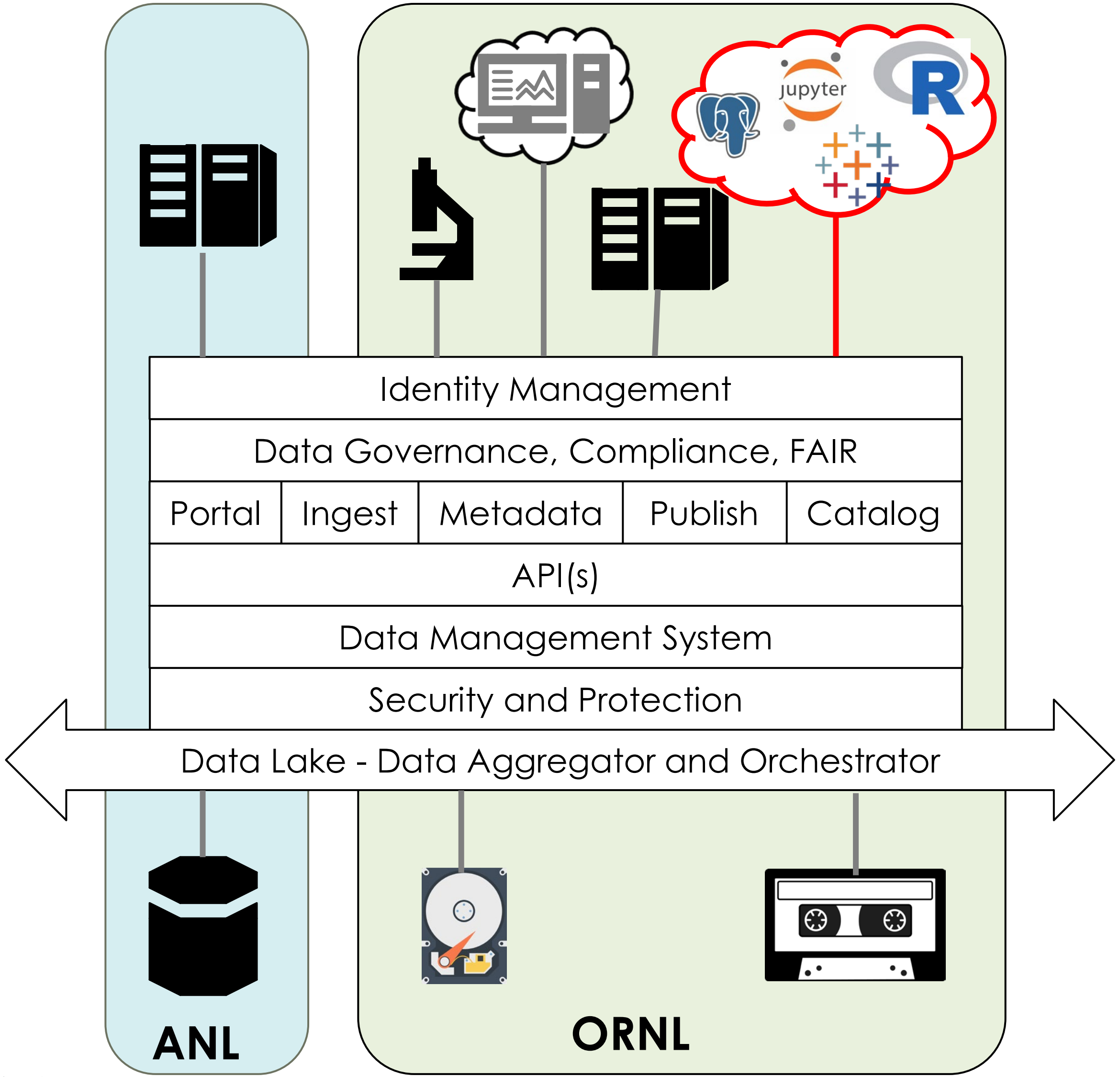
- Provide tools to help comply with FAIR data principles
- Guidance and support to develop and comply with data management plans
- Data governance councils
  - Educate staff about best practices, tools and resources
  - Strategies to extract value out of data
  - Continual improvement of scientific data federation

Trust and Identity Management



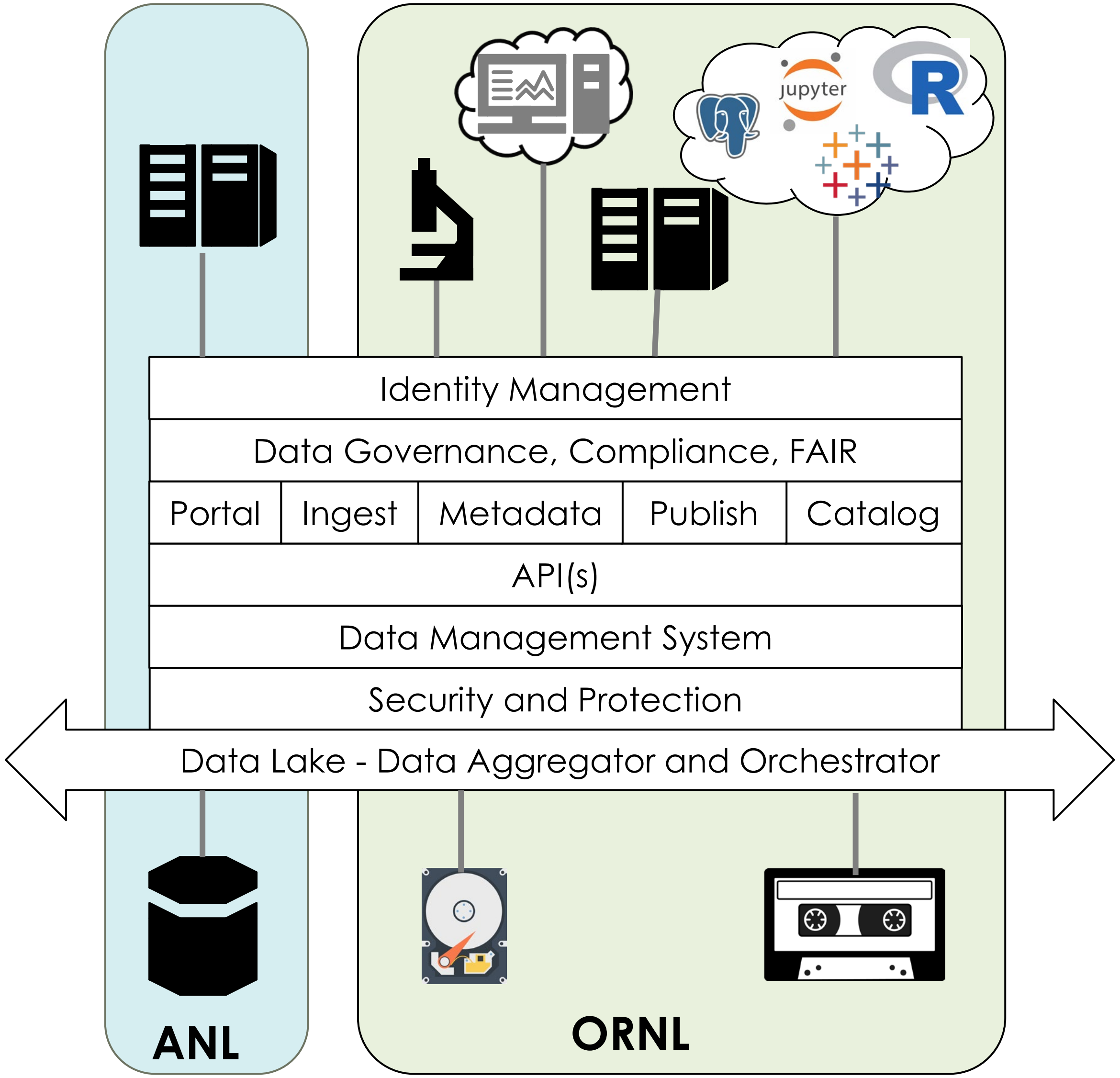
- Adopt Federated Identity Management solutions (e.g., DOE OneID)
- Should work for interactive (e.g. web interface) and scripting / automated use-cases
- Establish or adopt clear standards for authentication and authorization (perhaps as tiers)
  - facilities easily compare / accept each other's security measures
  - compliance checks to access / handle sensitive data from computational resources

Accelerate Data Driven Discovery



- Data analytics platforms (i.e., Jupyter, R Studio, Tableau)
- Platform as a Service for ancillary needs, e.g. - databases, web portals, workflow monitoring
- Software and hardware support for analytic workloads
- Workflows that can be tailored for specific domains

Summary



Identity Management

Data Governance, Compliance, FAIR

Portal	Ingest	Metadata	Publish	Catalog

API(s)

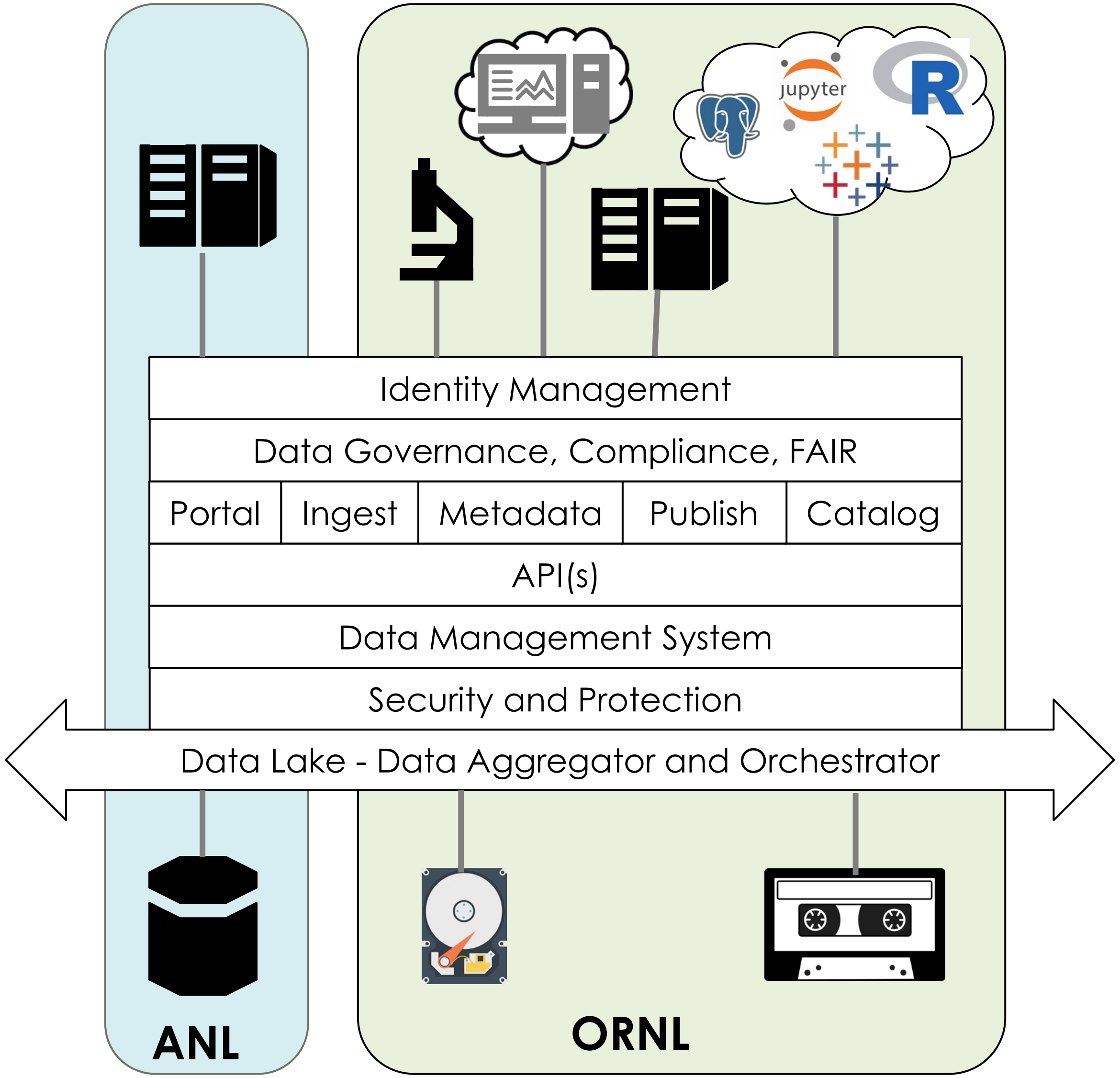
Data Management System

Security and Protection

Data Lake - Data Aggregator and Orchestrator

Acknowledgement

This research used resources of the Oak Ridge Leadership Computing Facility at the Oak Ridge National Laboratory, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC05-00OR22725.



Identity Management

Data Governance, Compliance, FAIR

Portal | Ingest | Metadata | Publish | Catalog

API(s)

Data Management System

Security and Protection

Data Lake - Data Aggregator and Orchestrator