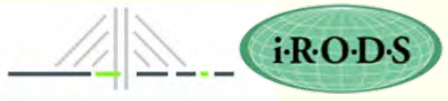
Apr 9th, 12:00 AM

The Librarian & the Big Data: Bridging the Gap

Arcot Rajasekar
University of North Carolina at Chapel Hill

The Librarian & the Big Data: Bridging the Gap

Arcot Rajasekar
rajasekar@unc.edu
The University of North Carolina
at Chapel Hill



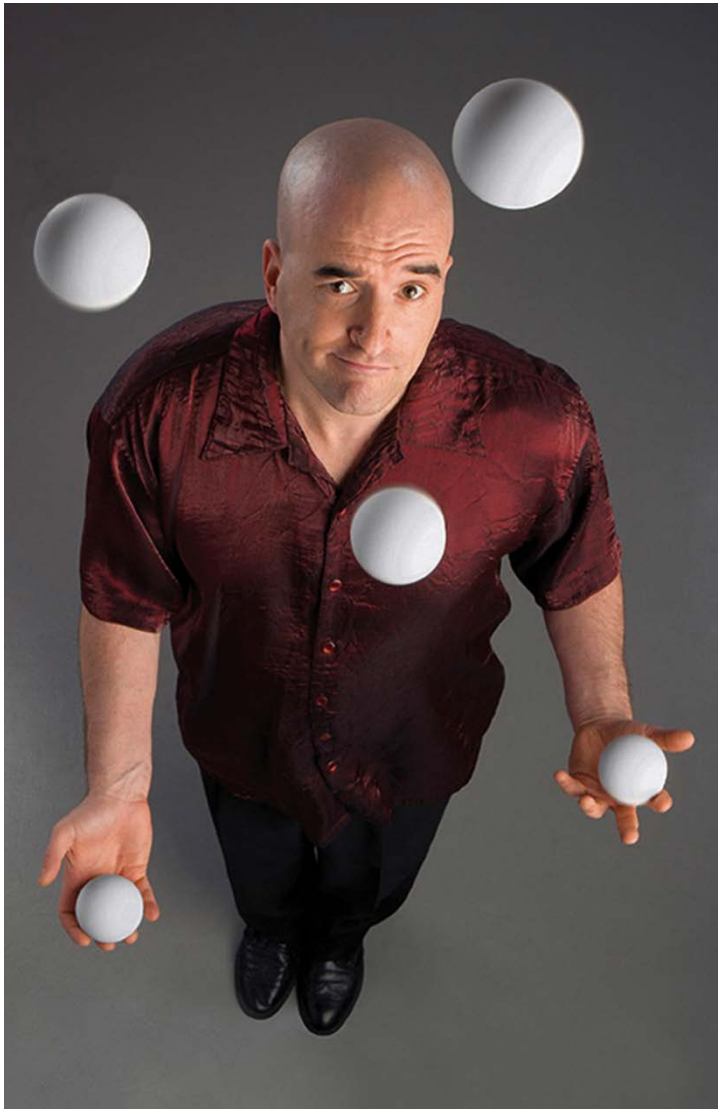
Outline

- Challenges in Big Data
  - Scientific Data Explosion & Role of Librarians
- Projects at UNC
  - Gearing to Meet the Challenges
- Looking Towards the Future
  - Integration of Data, Computing & Networks

Big Data Challenges for Information & Library Science

Lets Start with an Analogy

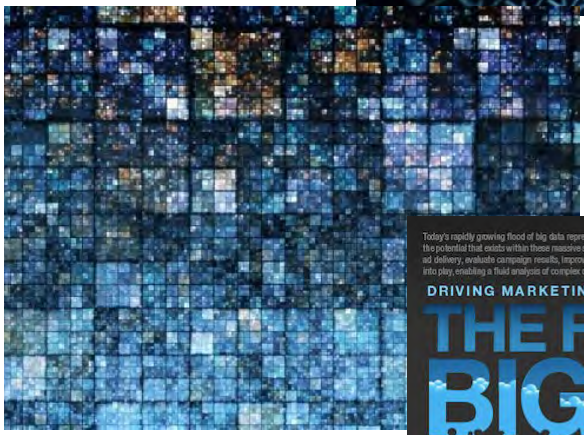
ILS - Today

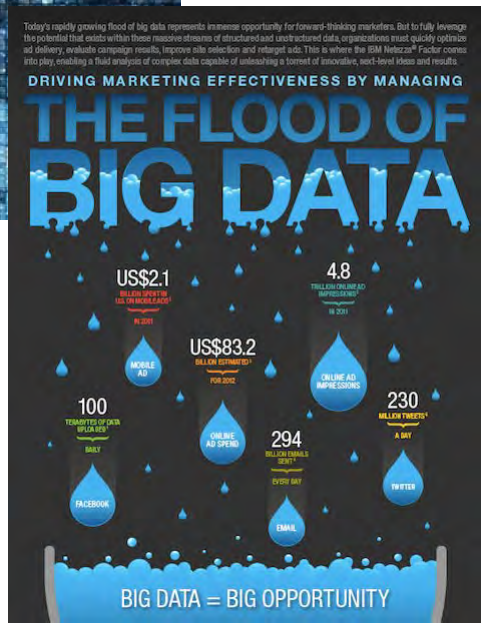


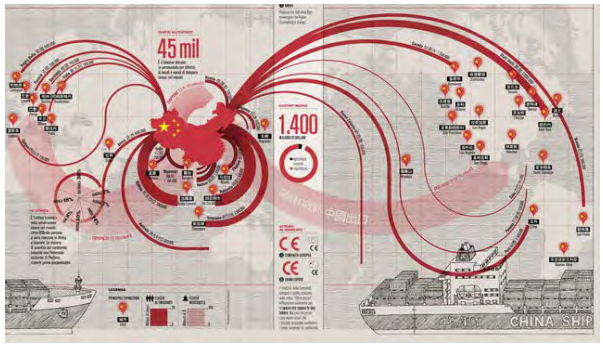
ILS - Tomorrow

Big Data EveryWhere!

- Lot of data collected and analyzed
  - Web data, e-commerce
  - Scientific projects
  - Commercial/Financial transactions
  - Social Network data
  - Medical & Health Information

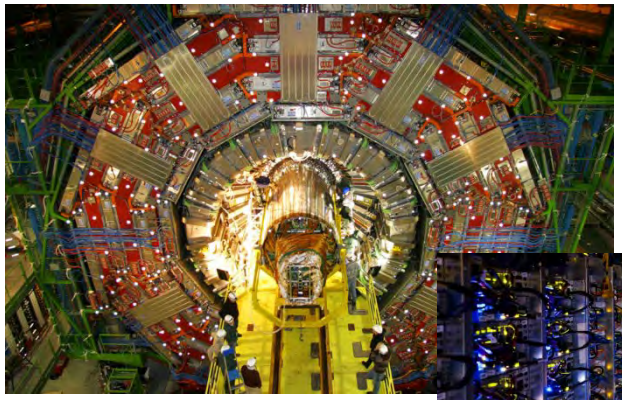


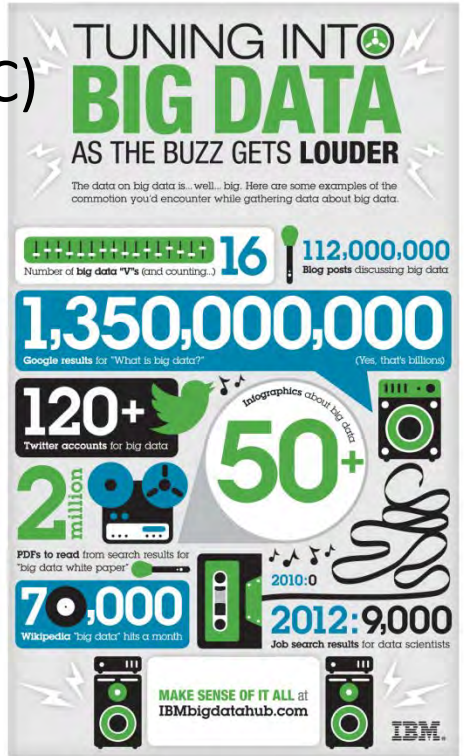




How much data?

- Google processes 20 PB a day (2008)
- Wayback Machine has 3 PB (3/2009)
  - Growing at 100 TB/month
- Facebook has 2.5 PB of user data (4/2009)
  - Growing at 30 TB/day
- eBay has 6.5 PB of user data(5/2009)
  - Growing at 50 TB/day
- CERN's Large Hydron Collider (LHC) generates 15 PB/year





Characteristics of Big Data

- Five V_s -
  - Volume – Exponential Increase in Size & Count
  - Velocity – Speed at which Data is Created, Processed or Used
  - Variety – Multi-dimensionality, arrangement, format, etc.
  - Veracity – Integrity & Fidelity
  - Value – Worth

Paradigm Shift

- Compute Intensive to
  Data Intensive
- Large Actions on Small Amounts of Data to
  Small Actions on Large Numbers of Data
- Move Data to Processing Site (Supercomputer Model)
  Move Process to Data Site (Map-Reduce Model)
- Function Chaining to
  Service Chaining
- Model-based Science to
  Data-based Science (Data Mining, Knowledge Discovery)

Information Science in the Future

- Organization
- Classification
- Ontologies
- Metadata
- Retrieval
- Management
- Collection building
- Analysis
- Information seeking
- Knowledge Representation
- Human Computer Interaction
- Social Skills
- Information Behavior
- Ethics
- Privacy
- Security
- Information Technology
- Transformation
- Interpretation
- Dissemination
- Application
- Reference Collections
- Information Processing
- Data Mining
- Information Visualization
- Communication Network
- Policy

The List is the Same for Tomorrow – No Different than Today
But Changes are needed to meet the 5Vs
New Methodology, New Way of Thinking, New Processing Paradigms, New Interactions

THE FUTURE IS BRIGHT

Building a Big Data Platform

Integrated Rule Oriented Data Systems (iRODS)

- Life Time Library – Personal Digital Library
- Carolina Digital Repository – Institutional Repository
- DataNet Federation Consortium – National-scale Cross Disciplinary Collaboration
- Data Bridge – Long-tail of Science “Data Communities”

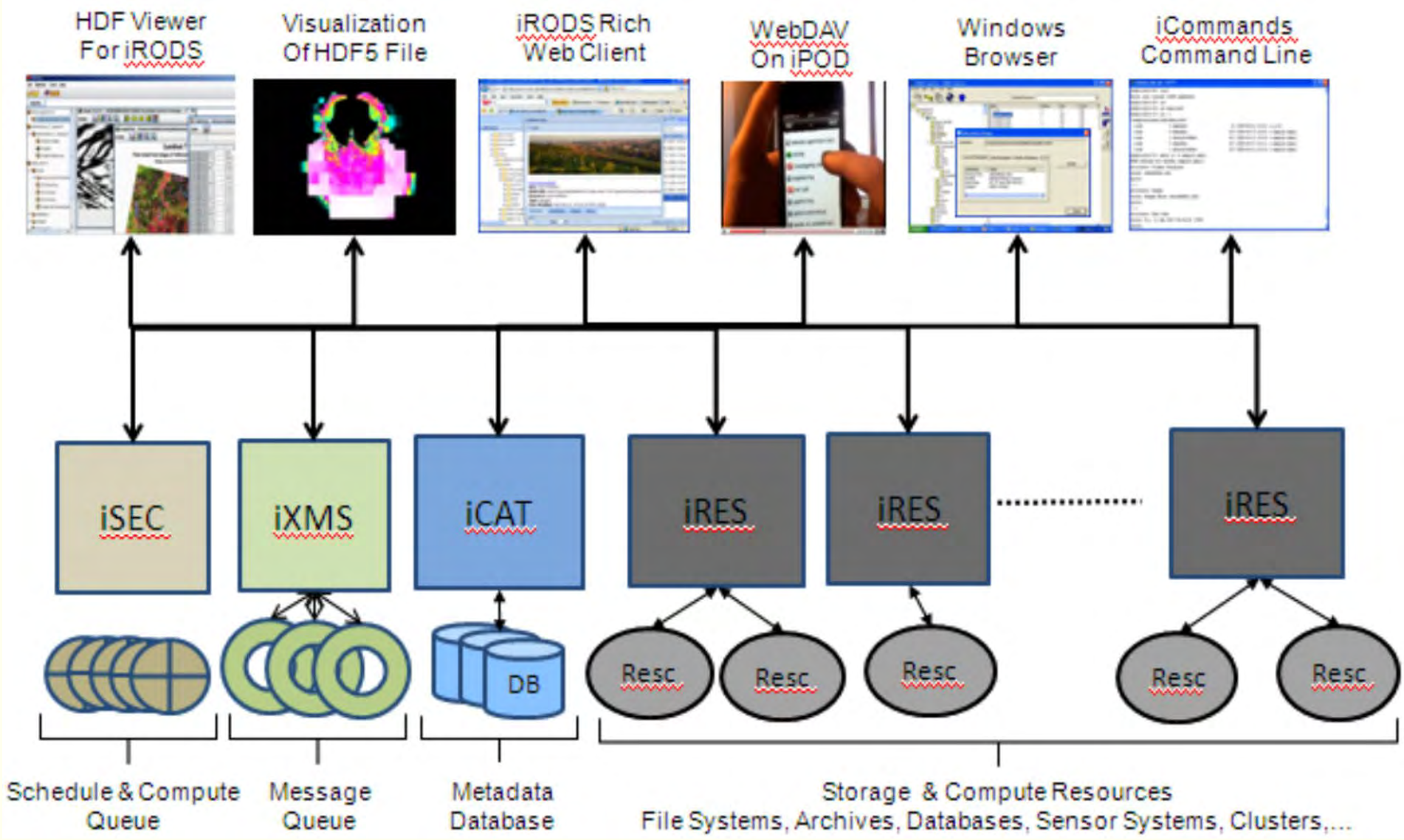
Motivations & Ingredients

- Two Early Projects: 1996 -1999
  - Massive Data Analysis System (MDAS) - DARPA
  - Distributed Object Computation Testbed (DOCT) – DARPA, USPTO, NARA
  - Motivation: Perform data-intensive computation across distributed resources administered by multiple organizations
  - Ingredients:
    - Hide the Data Distribution
    - Virtualize Resources
    - Uniform Access Mechanisms
  - Platform: Storage Resource Broker (SRB) Data Grid
- Multiple Follow-on Projects: 1999 - 2006
  - Transcontinental Persistent Archives Prototype (TPAP) – NARA, NSF
  - National Partnership for Advanced Computing Infrastructure – NSF
  - Others (DOE, DOD, NASA, NIH)
  - Motivation: Federation of multiple data grids for scientific collaboration
  - Ingredients:
    - Associate Metadata
    - Virtualize the User
    - Empower Sharing and Collaboration
  - Platform: Storage Resource Broker (SRB) Federated Data Grid

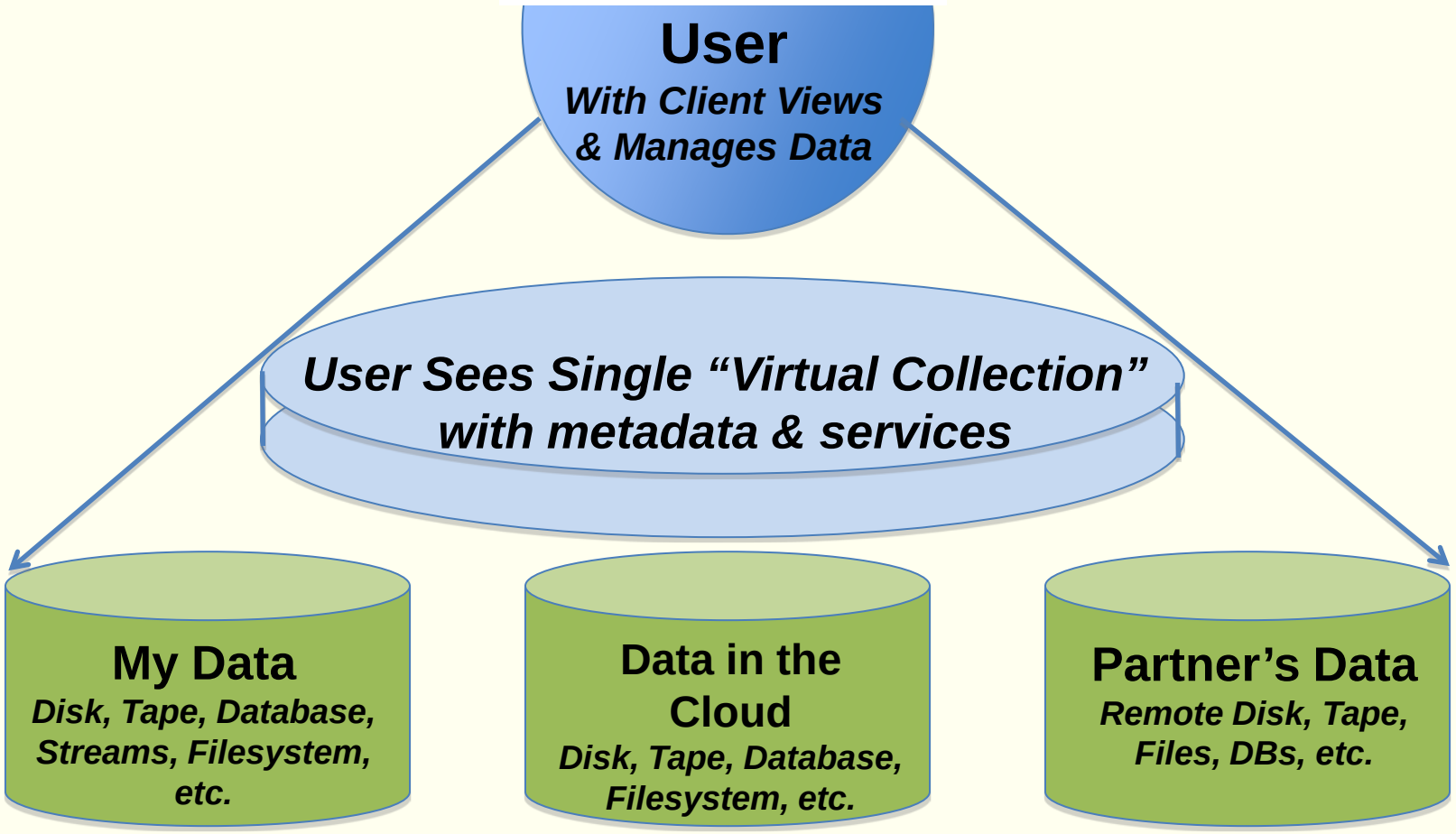
Motivations & Ingredients(Contd.)

- Paradigm Shift: 2005 - 2013
  - Policy-based Data Management- NARA, NSF
  - Motivation: Improve Customizability and Integrate Distributed Processing
  - Ingredients:
    - Automate Administration
    - Manage with Policy
    - Enable server-side processing
    - Empower User-centric Customization
    - Support Workflows and Computational Services
  - Platform: DICE integrated Rule Oriented Data Systems (iRODS)
- Future : 2012 -
  - 1000s of projects all over the world and growing
  - DataNet Federation Consortium at UNC-Chapel Hill (NSF)
  - Motivation: Sustainability of the iRODS software and extensions
  - Ingredients:
    - Encapsulate domain knowledge in procedures
    - Enable reproducible data driven research
    - Sustianability through Formation of an iRODS Consortium
  - Platform: Consortium iRODS - released early April 2014

iRODS Distributed Data Management

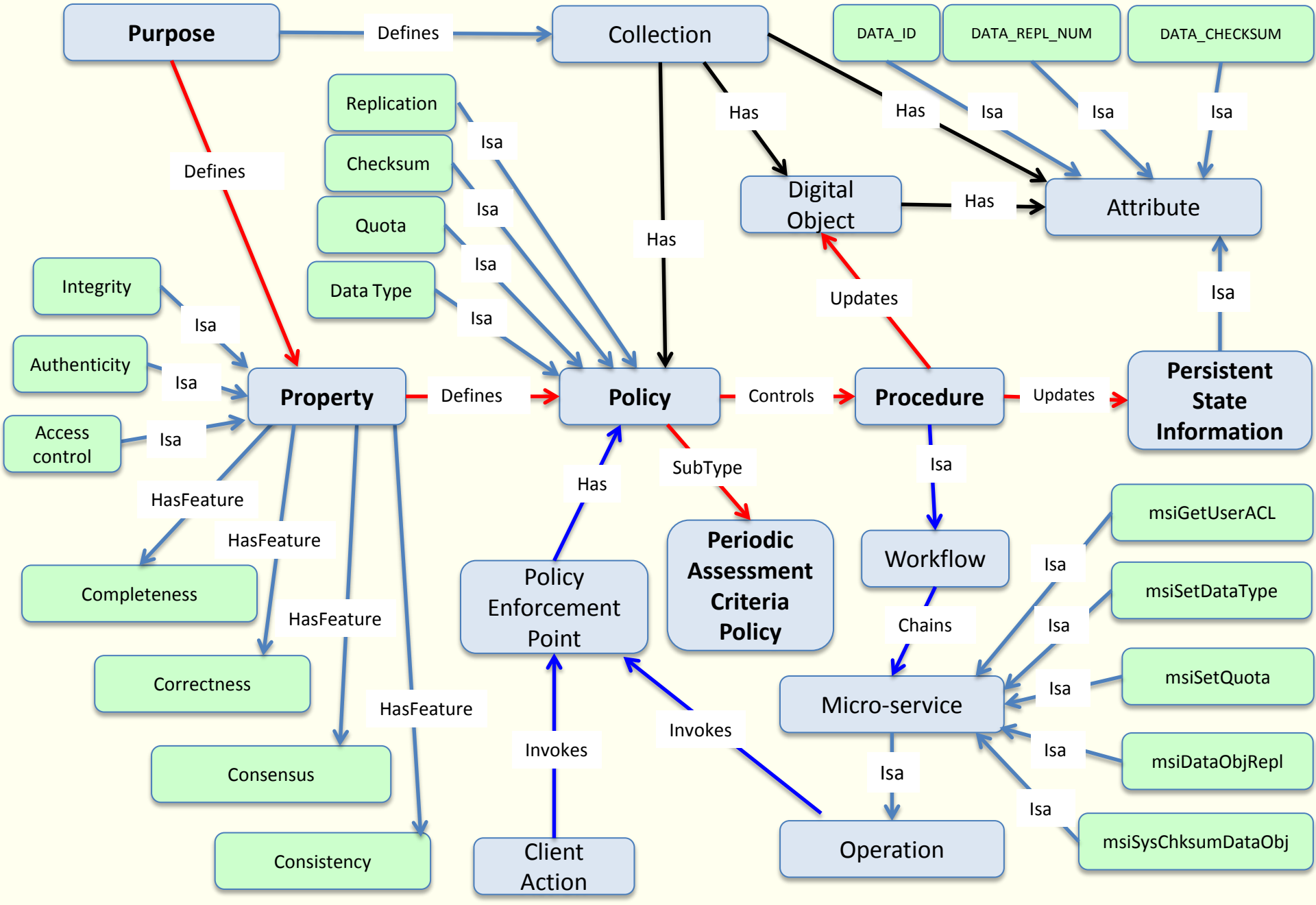


iRODS Shows Unified "Virtual Collection"



The iRODS Data System can install in a "layer" over existing or new data, letting you view, manage, and share part or all of diverse data in a unified Collection.

Policy-based Collection Management



iRODS: Types of Scalability

- Largest data grid
  - > 10 Petabytes (French National Institute for Nuclear Physics and Plasma Physics)
  - > 500+ storage resources (Australian Research Collaboration Service)
  - > 15,000 users (the iPlant Collaborative)
  - > 300 million attributes (NASA Center for Climate Simulations)
  - > Inter-continental sharing (Cinegrid – Americas, Asia, Europe)

Data Management Applications

- International projects
  - BaBar, International Neuroinformatics Coordinating Facility, Cyber Square Kilometer Array (radio astronomy), Cinegrid (movies)
- National data grids
  - Australia-ARCS, New Zealand, Portugal, UK, France-IN2P3
- Federal agency archives
  - NOAA National Climatic Data Center, NASA Center for Climate Simulation, National Optical Astronomy Observatories, NSF XSEDE
- Grand challenge research projects
  - iPlant Collaborative, Ocean Observatories Initiative
- Institutional repositories
  - Carolina Digital Repository, UNC-SILS LifeTime Library, Texas Digital Library, French National Library, Broad Institute genomics data grid, Sanger Institute genomics data grid
- Projects in 39 countries, 62 academic institutions in the US

Using the Big Data Platform

Integrated Rule Oriented Data Systems (iRODS)

- Life Time Library – Personal Digital Library
- Carolina Digital Repository - Institutional Repository
- DataNet Federation Consortium – National-scale Cross Disciplinary Collaboration
- Data Bridge – Long-tail of Science “Data Communities”

SILS LifeTime Library Vision

- Inculcate digital collection assembly habits
- Teach insights into policies – hands on
- Provision with micro-services for extraction of metadata
- Provision with indexing mechanisms
- Assignments in class
- Keep beyond the UNC degree program
- Student digital libraries
  - Enable students to build collections of
    - ✓ Photographs
    - ✓ MP3 audio files
    - ✓ Class documents – presentations, projects, homeworks, reading material
    - ✓ Video
    - ✓ Web site archive
    - ✓ Track social media

SILS LifeTime Library

- Resources provided by School of Information and Library Science at UNC-CH
- Replication at RENCI – two copies
- Policies are student driven
- Collections are student assembled:
  - Student collections range from 2 GBytes to 150 Gbytes
  - Number of files from 2000 to 12,000 per student
  - Policies: 5 policies on the average

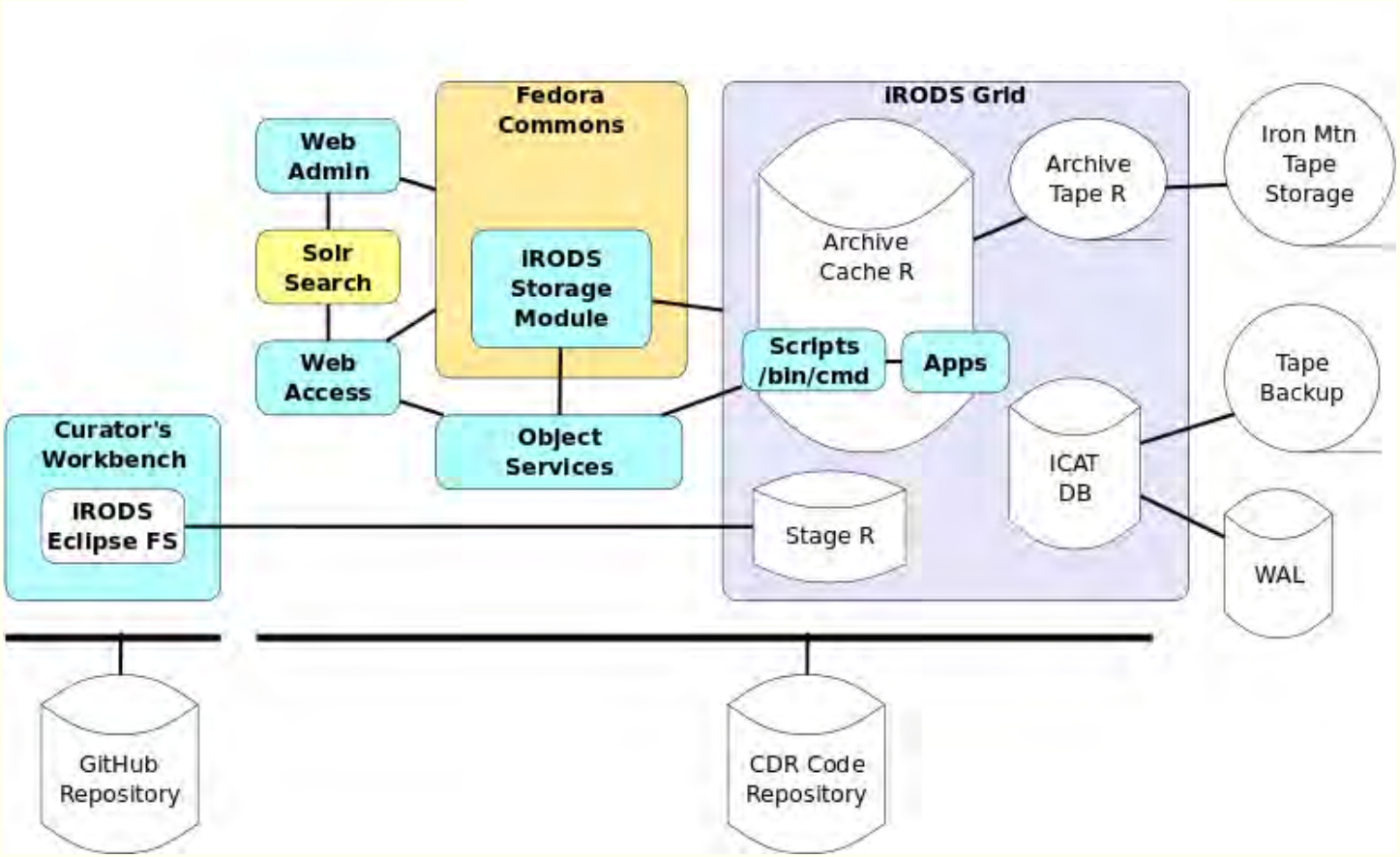
SILS LifeTime Library Policies

- Library management policies
  - Replication
  - Checksums
  - Versioning
  - Strict access controls
  - Quotas
  - Metadata catalog replication
  - Installation environment archiving
- Ingestion
  - Automated synchronization of student directory with LifeTime Library
  - Automated loading or extraction of metadata

Sample Student Collections (2012)

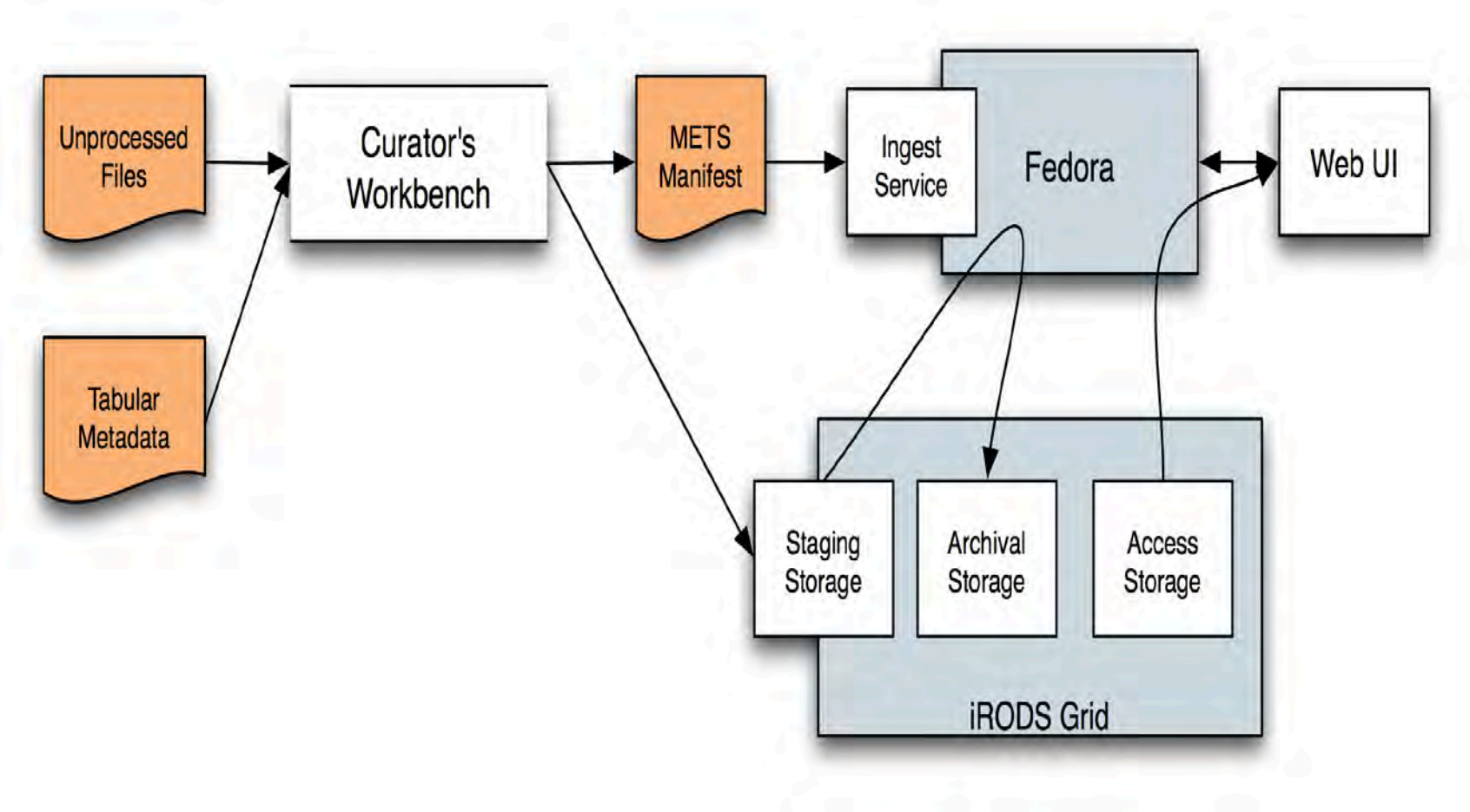
# files	# metadata	Size	#metadata/file	Collection	Metadata type	Metadata load
2111	8684	16.0 GB	4.1	iTunes	AVUs	XML load
2734	4500	4.3 GB	1.6	Photo	Tags	Hand
1109	8174	1.2 GB	7.4	Photo, Music	Tags, AVUs	Hand
5697	15472	47.0 GB	2.7	iTunes	AVUs	ASCi load
1692	8098	0.1 GB	4.8	Photo	AVUs	Hand
125	1100	0.8 GB	8.8	iTunes	AVUs	XML load

Carolina Digital Repository



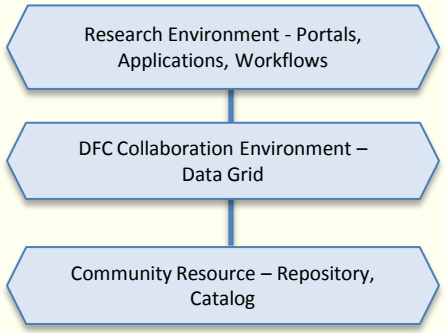
Policy-Driven Repository Infrastructure project
funded by the Institute for Museum and Library Services

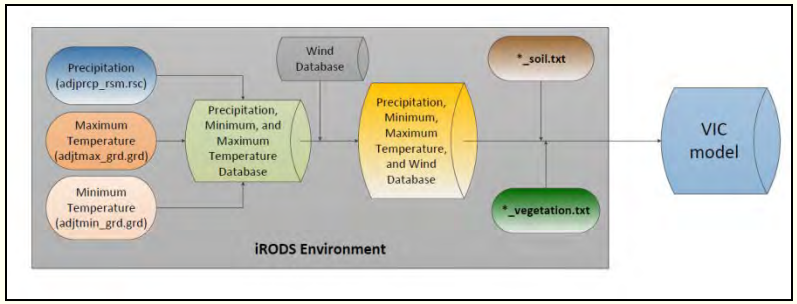
Carolina Digital Repository
Ingest Workflow



DataNet Federation Consortium Vision

- Enable collaborative research
  - Sharing of data, information, and knowledge
- Build national data cyberinfrastructure
  - Federation of existing data management systems
- Support reproducible data-driven research
  - Encapsulate knowledge in shared workflows
- Enable student participation in research
  - Policy-controlled access to “live” data

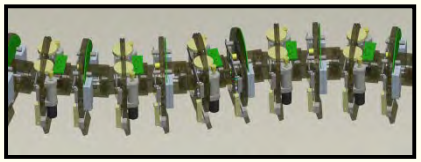


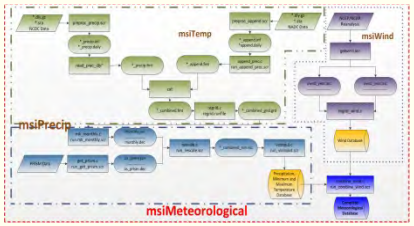


Data Driven Science and Engineering

- Collaboration Environments
  - Oceanography – Ocean Observatory Initiative
    - Archiving of climatic data records from real-time sensor data streams, replay of sensor data
  - Engineering – CIBER-U
    - Engineering Digital Library: curation of civil engineering data, student training materials
  - Hydrology - CUAHSI, …
    - Automation of hydrology research workflows (reproduce, reuse and repurpose)
  - Plant Biology – iPlant Collaoratory
    - Project data I sharing and integration, virtualized metadata services
  - Social Science – Odum Institute
    - Survey data and Statistical data processing
  - Cognitive Science – Temporal Dynamics Learning
    - Inter-team collaboration policies, human data

Engineering Representation

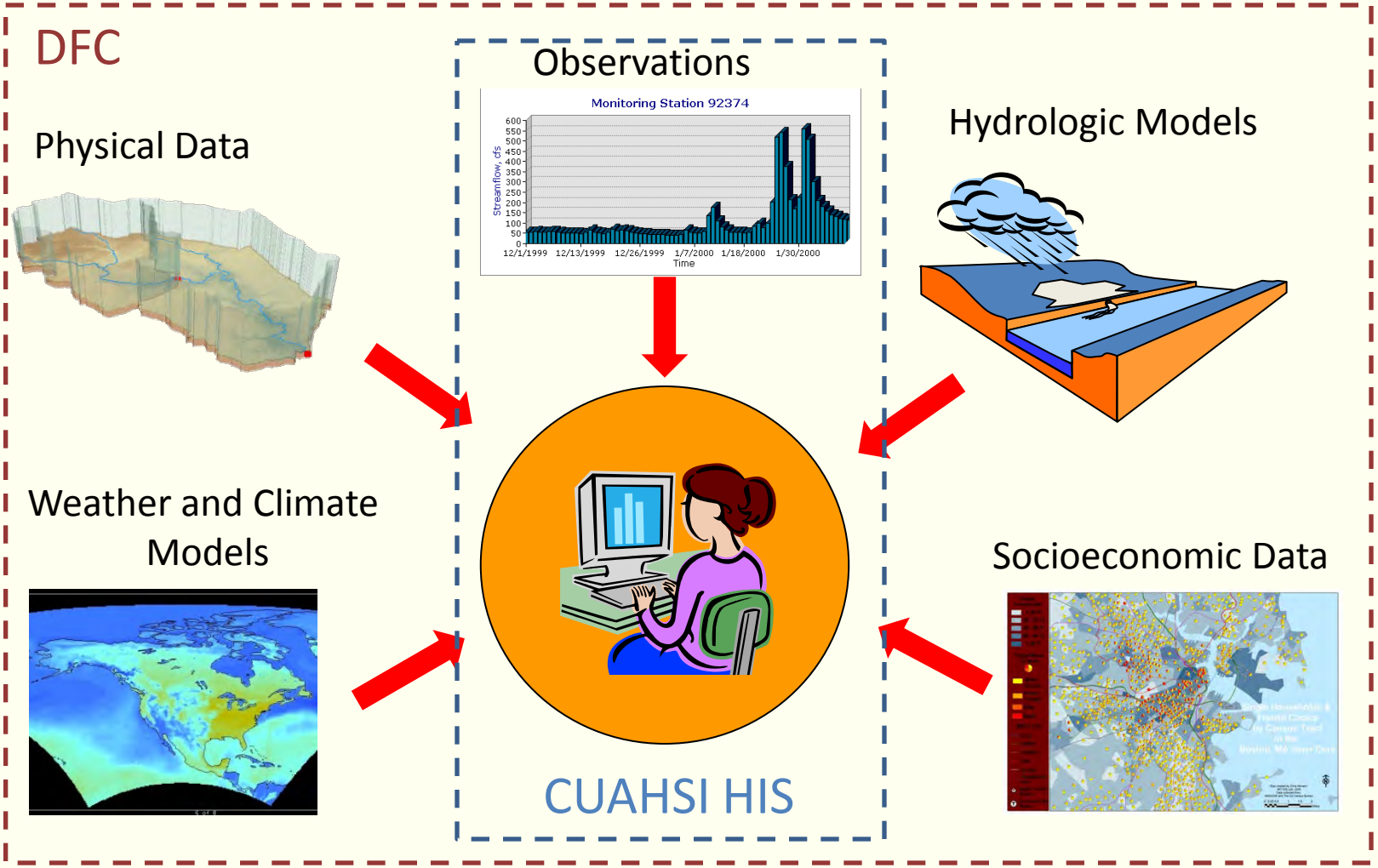




Knowledge Management

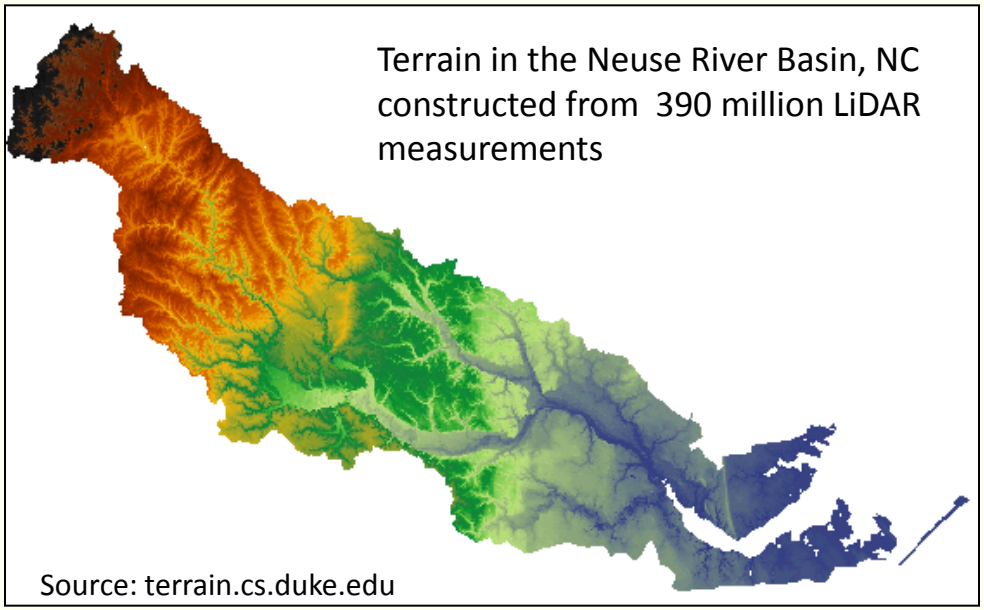
- DFC is exploring three types of knowledge management
  - Encapsulate knowledge needed to enforce a community consensus on shared collections
    - Policies and procedures
  - Encapsulate knowledge needed to automate a research analysis
    - Research workflows
  - Encapsulate knowledge needed for interoperability between data cyberinfrastructure components
    - Micro-services

Data and Model Integration Needed to Support Hydrologic Science



A Use Case: National Water Model

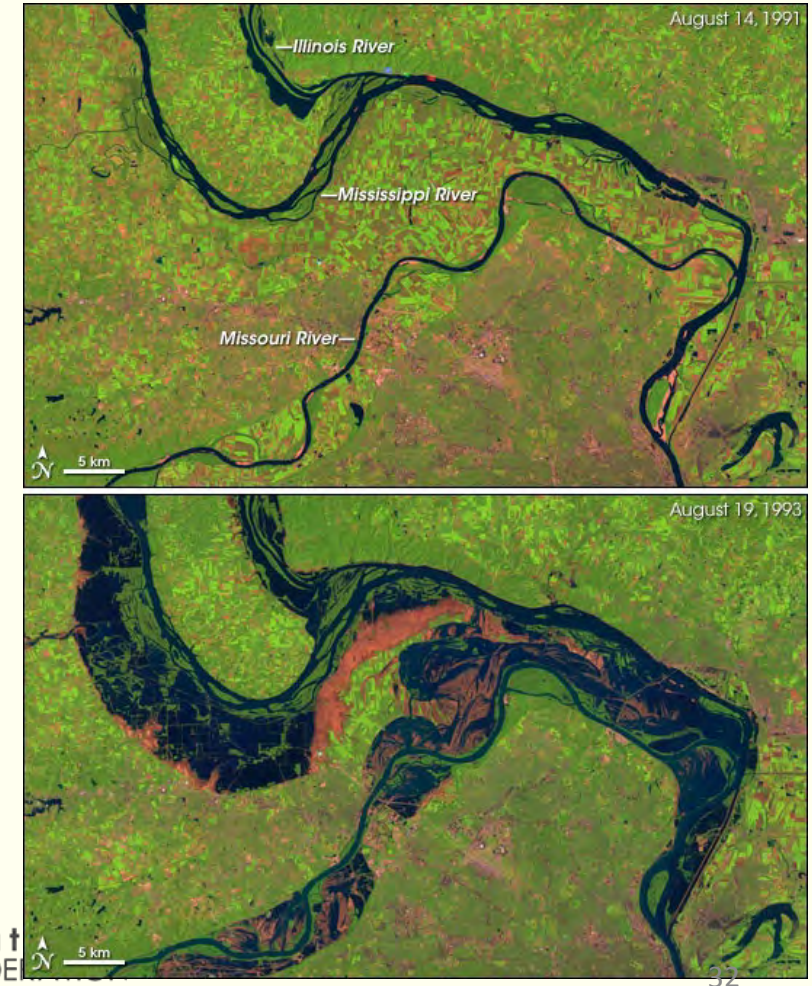
Hydrologic scientist have expressed a "grand research challenge" of building a National Water Model for flood and drought applications.



Achieving this goal will require a system like DFC to handle the massive data requirements.

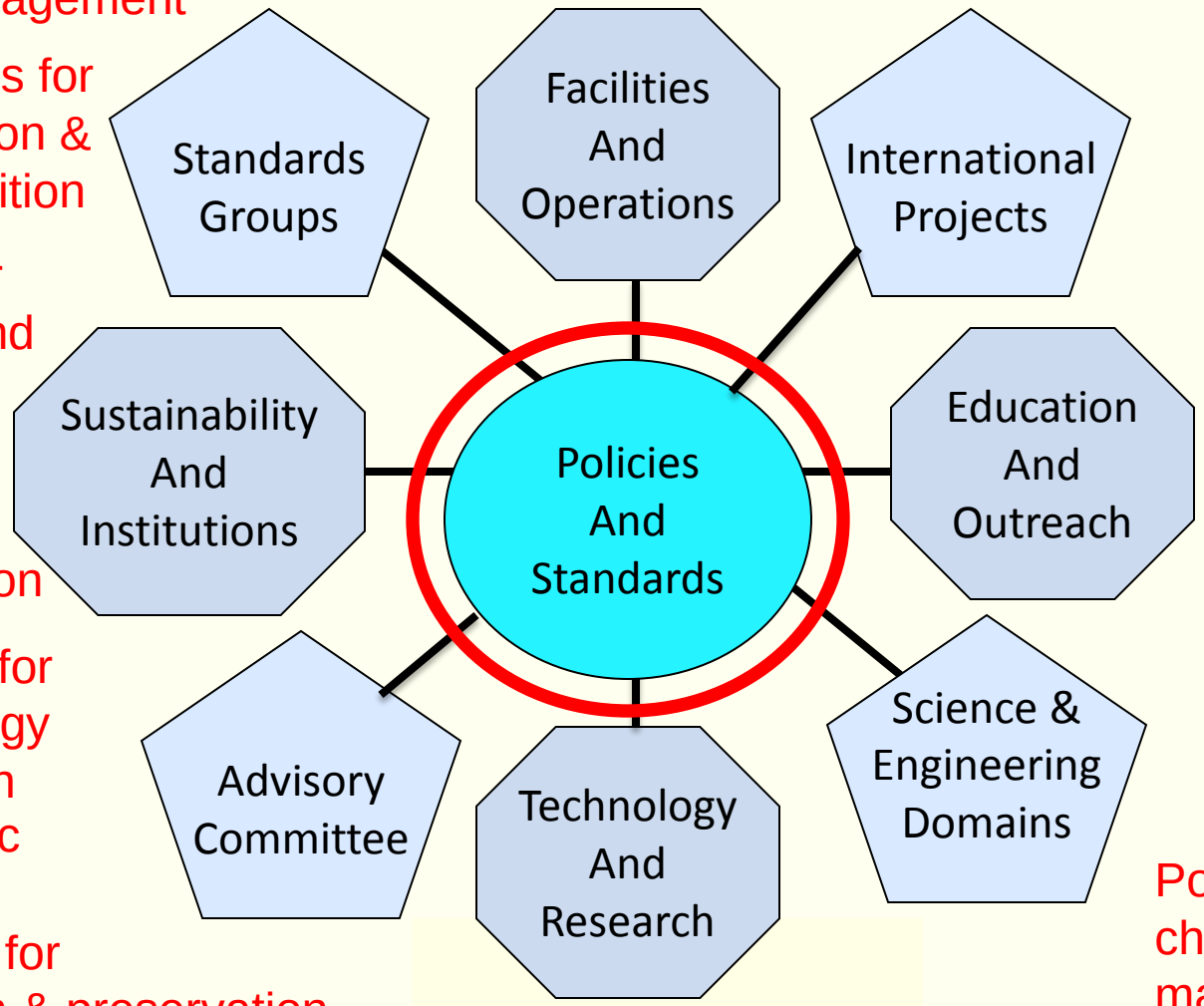
Can it be done in real-time with streaming data?

Flooding in the Mississippi River Basin, August 1993 observed from satellite imagery



Source: nasa.gov

Policies at the Center of the DFC



Policies for automating data management

Policies for metadata extraction

Policies for Deletion & redaction

Policies for collaboration and reuse

Policies for trust

Policies for authentication & authorization

Policies for retention & disposition

Policies for archiving, staging & caching

Policies for analysis and workflow

Policies for IPR & citations

Policies for replication & synchronization

Policies for publication & federation

Policies for technology migration

Policies for provenance & sustainability

Domain-centric policies

Policies for change management

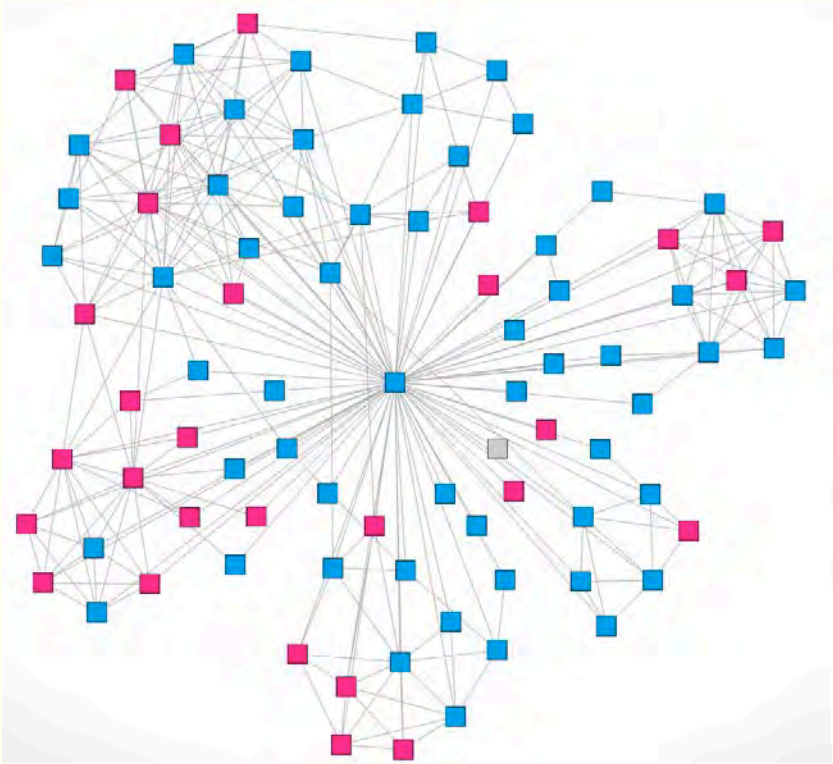
Polices for curation & preservation

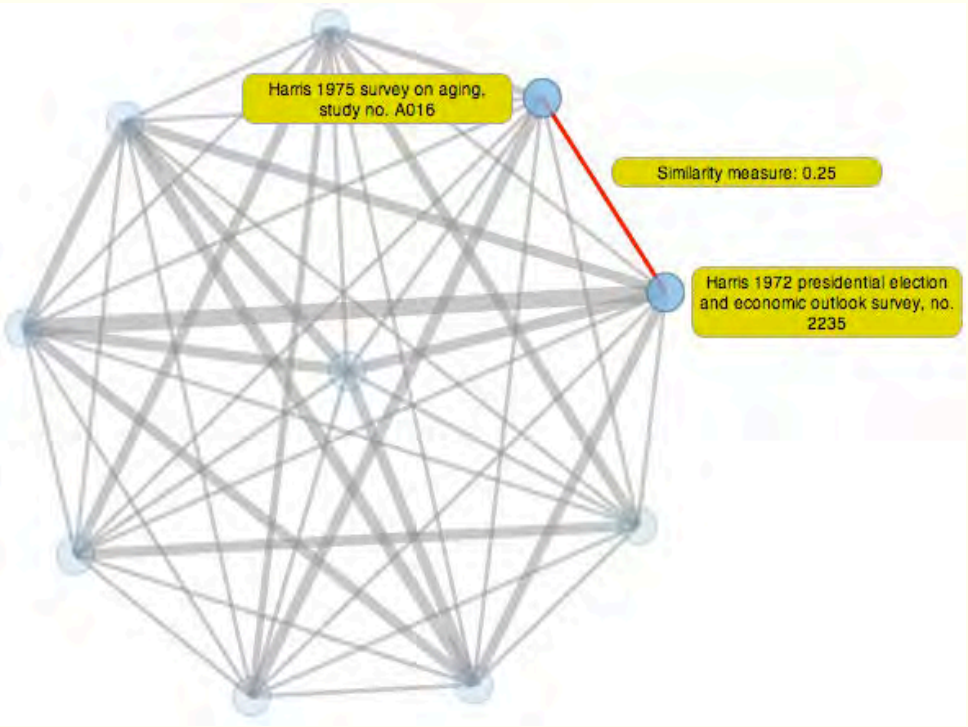
The DataBridge Vison

Build a Social Network for Scientific Data – Data as a Citizen

- Long-tail of Science Data
  - Enormous amounts of atomistic, distributed and unconnected dark data!
  - Non homogenous formatting and metadata
  - Gathered by one scientist or a small group
  - Not easily sharable or discoverable.
- Community detection through multi-dimensional sociometric analyses.
- Three Tasks and Challenges:
  - Evaluate the similarity/relevancy of data sets
  - Perform community detection on the resulting set of similarities
  - Provide ingestion, query and access to this multi-dimensional network

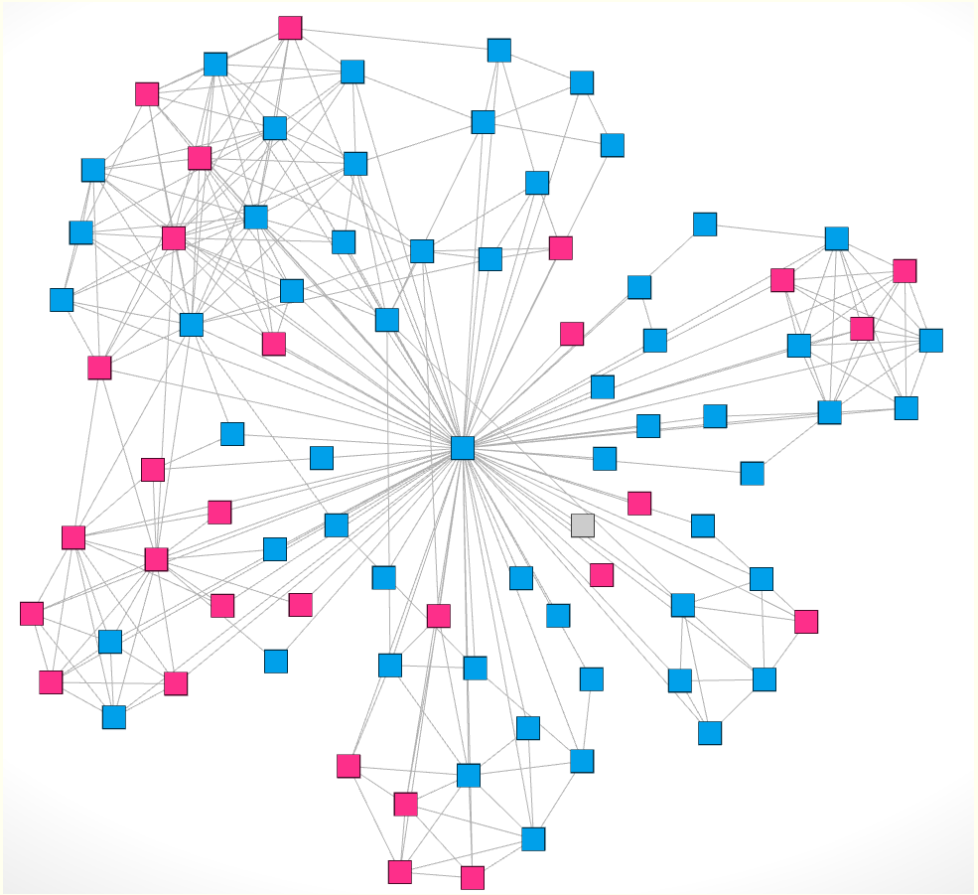
The DataBridge: A Social Network for Data





The DataBridge : Simple Social Network Example

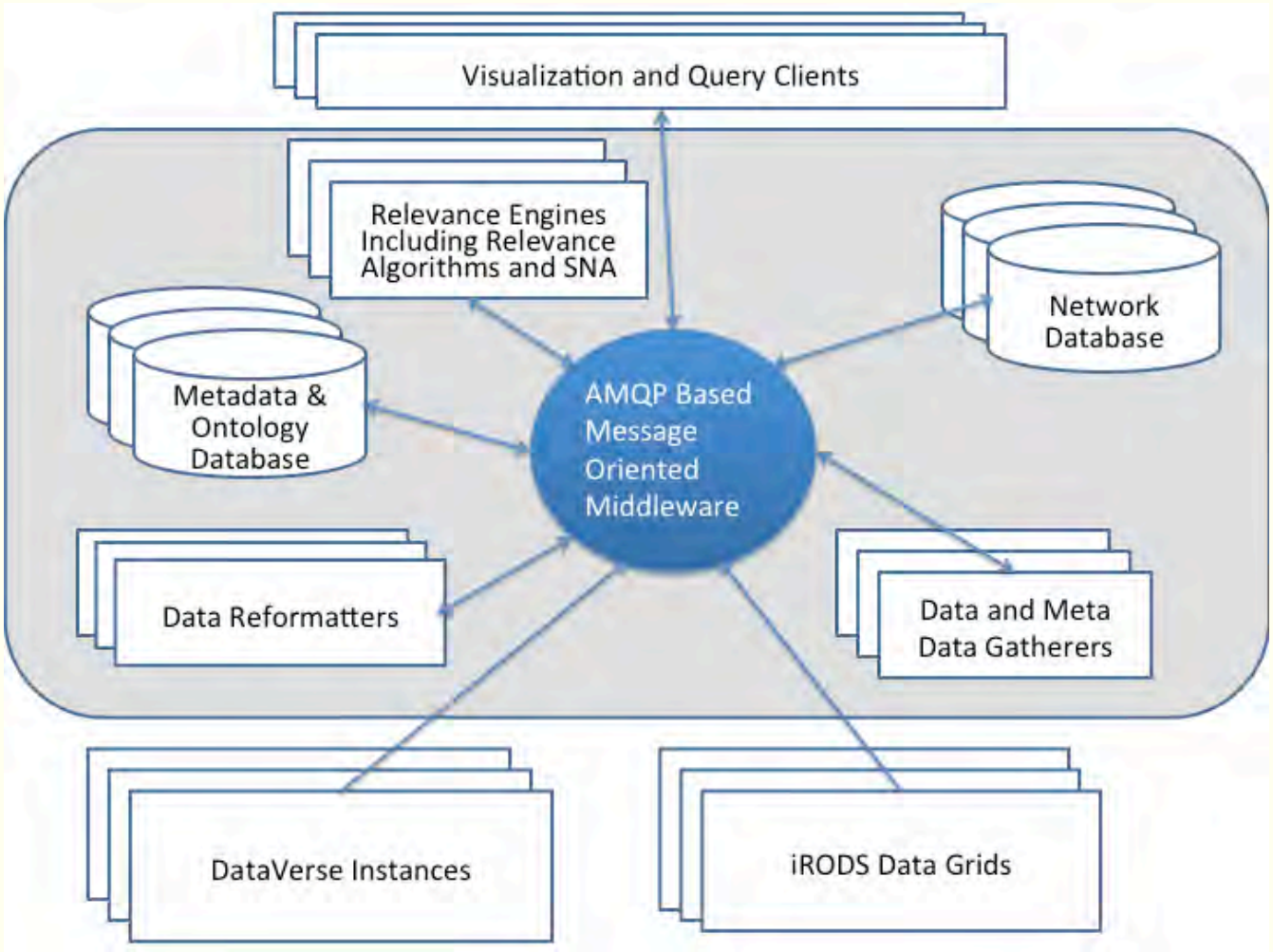
- Basic similarity: people who are Facebook friends with me
- Not a lot of additional information



The DataBridge Strategy: Building a Social Network for Scientific Data

- Investigate similarity measures:
  - Data to Data Connections: metadata and derived data about the data set
  - User to Data Connections: metadata about the usage and users of the data set
  - Method to Data Connections: metadata about the analyses of the data set
- Multi-dimensional similarities Examples
  - Eigen values and vector spaces
  - Cosine Similarities
  - Term Frequency and Inverse Doc Frequency
  - Principle Component Analysis
  - Latent Semantic Analysis
  - Digital Signature Analysis

DataBridge Implementation



Future Big Data Platforms

Building Informatics for Next-Generation Genomics

TAGCTAGGATCGAAGCA
…

> Genomic data currently for ~3000 patients

Next Gen Sequencing

Computational Workflows, High Performance Computing, Distributed Data, Security

RENCI

Sequencing Informatics

Informatics & Cyberinfrastructure R&D

Clinical Practice

Clinical Research

Basic Research

Identifying genomic variants relevant to clinical care

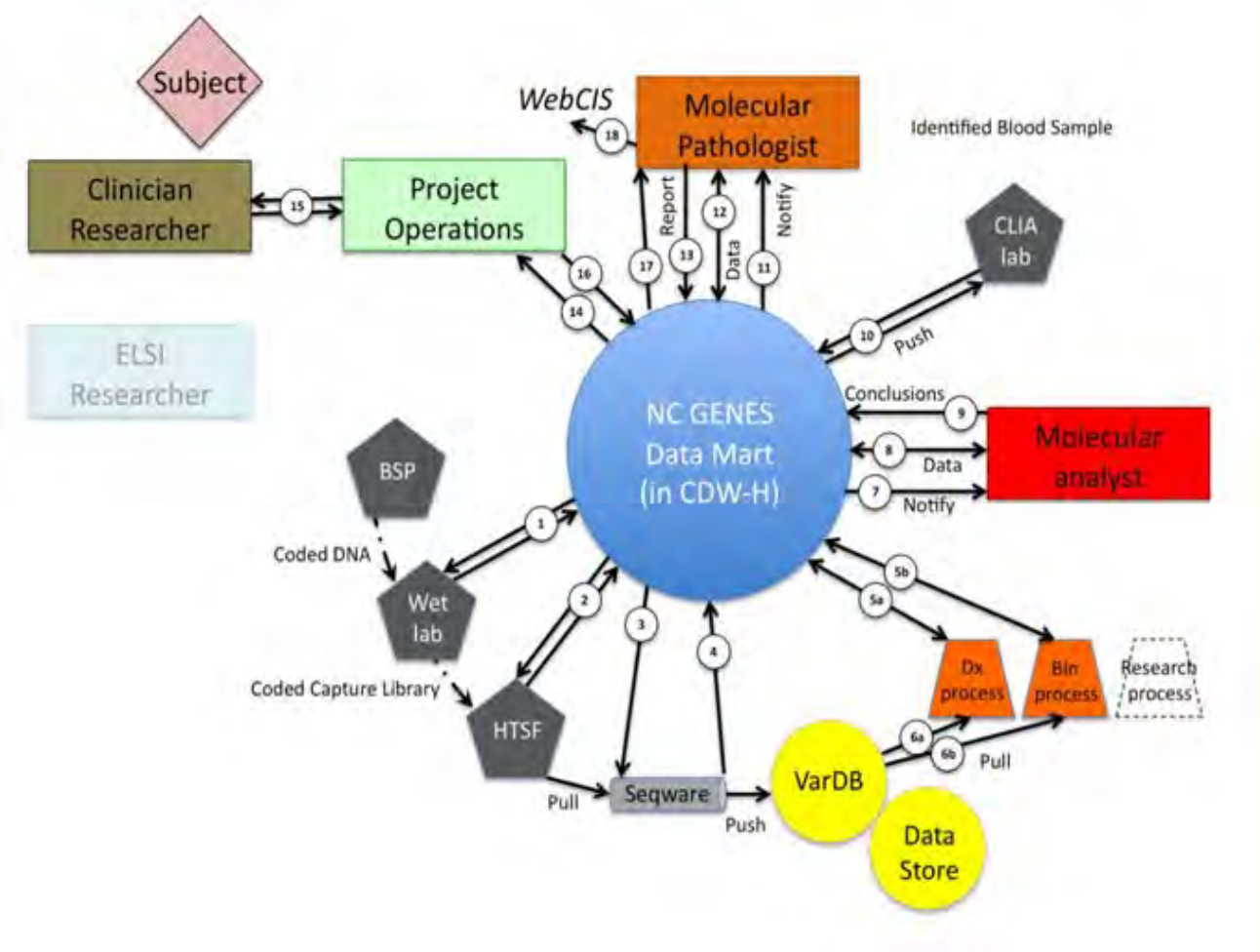
Exploring ethical/legal issues around reporting genomic findings

Determining relationships between genomic variants and disease

Finding new ways to understand the relationship between genes and disease/behavior

In collaboration with UNC Research Computing, UNC Dept of Medical Genetics, Lineberger Comprehensive Cancer Center, Institute for Pharmacogenetics and Personalized Treatment, UNC High Throughput Sequencing Core, UNC Center for Bioinformatics, DICE Center

End-to-End Clinical Genomics Informatics



- Blood draw to clinical relevant variants
- High performance analysis pipelines
- Large-scale data storage systems
- System-level workflow management
- Laboratory information management systems
- Orchestration around multiple storage and computer systems
- Closed loop system with independent validation paths (CLIA lab and exom chips)

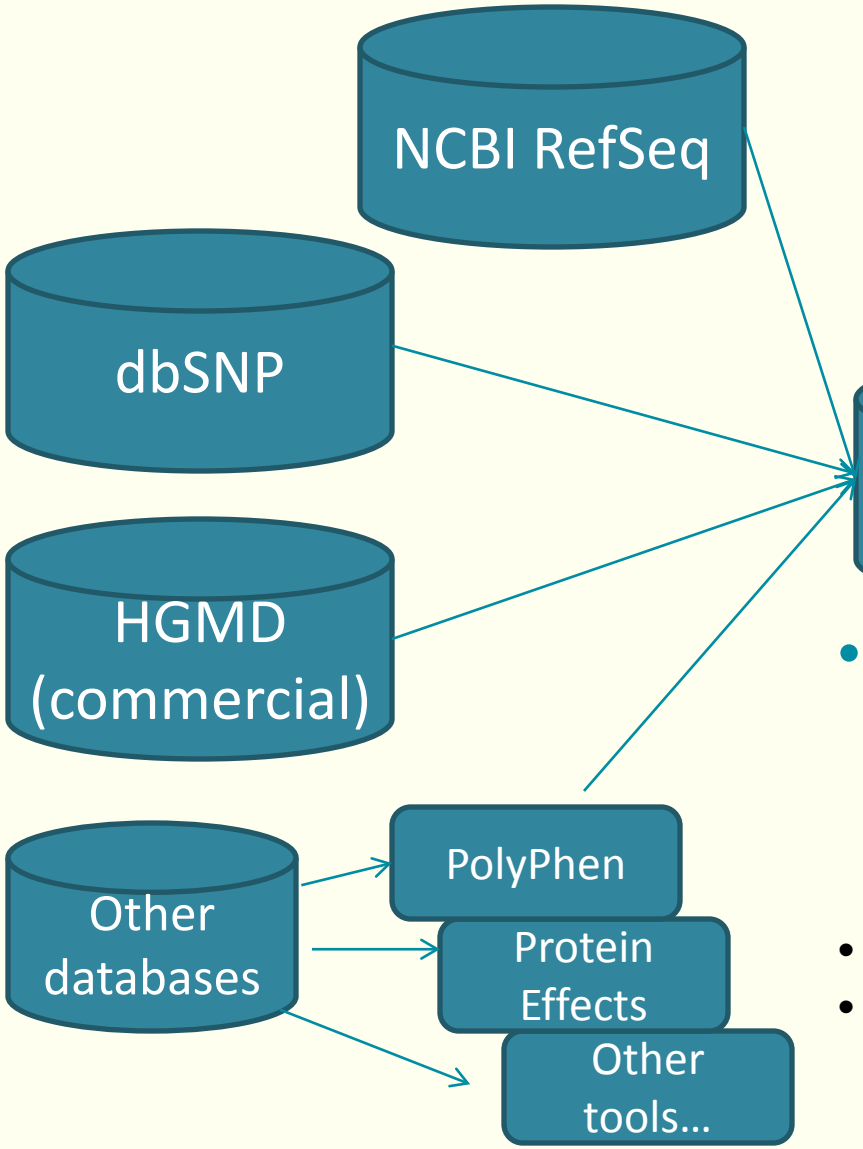
The challenge of storing the genome

- $15 to $75 billion dollars is the cost for disk space alone

- High costs push for limiting sequencing and data deletion
  - Exomic: only store around the genes
  - Genome chip: only store around known variants
  - Delete unaligned reads

- Yet, science is pushing for even more human genomic data
  - Epigenomic data
  - Cancer-specific data from sequencing
  - Micro-organism sequences (e.g., HIV virus) which are captured in human sequence data
  - All these vary over tissue and time

Aggregating Knowledge



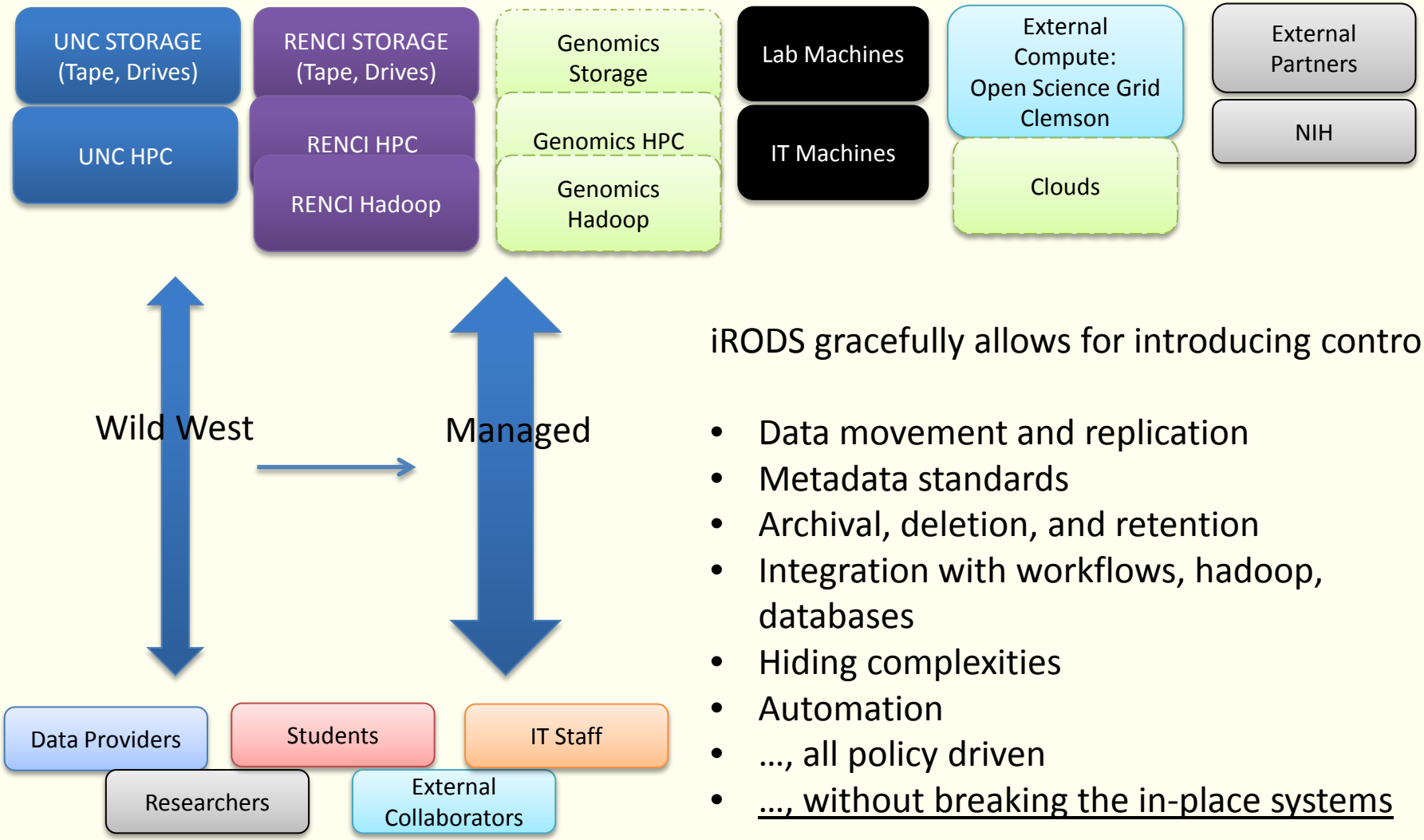
Approach:

-Traditional ETL and SQL queries

-Leverages RDBMS strengths

- **VarDB: ~1.5 TB**
  - Reference Genomes
  - Canonical Variants
  - Annotations
  - Indexes
- Dell PowerEdge 2950 (8 core) 32 Gb RAM
- MD 1000 storage array 750Gb SAS, ~ 4Tb

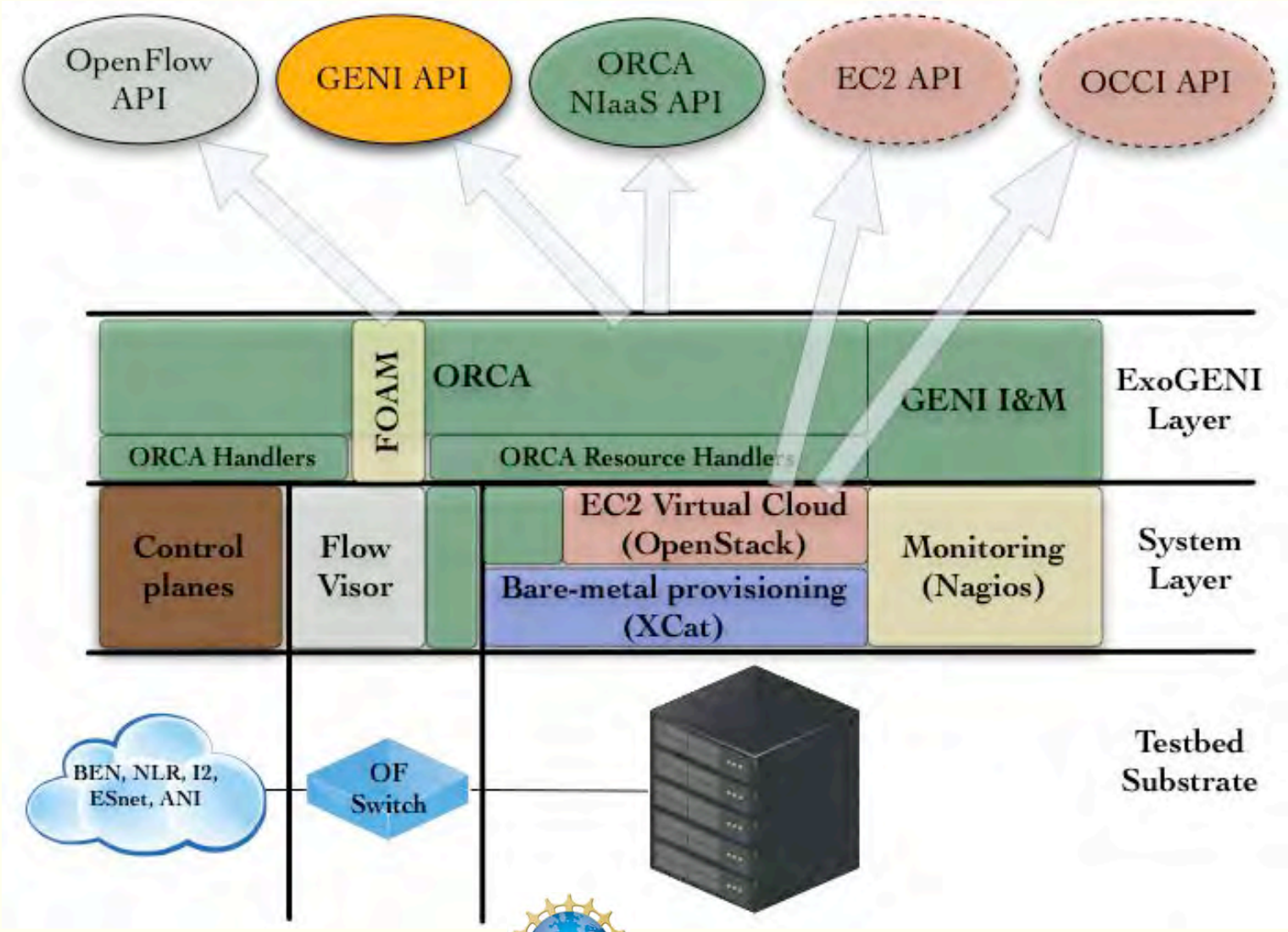
Managing Data on the Research Side



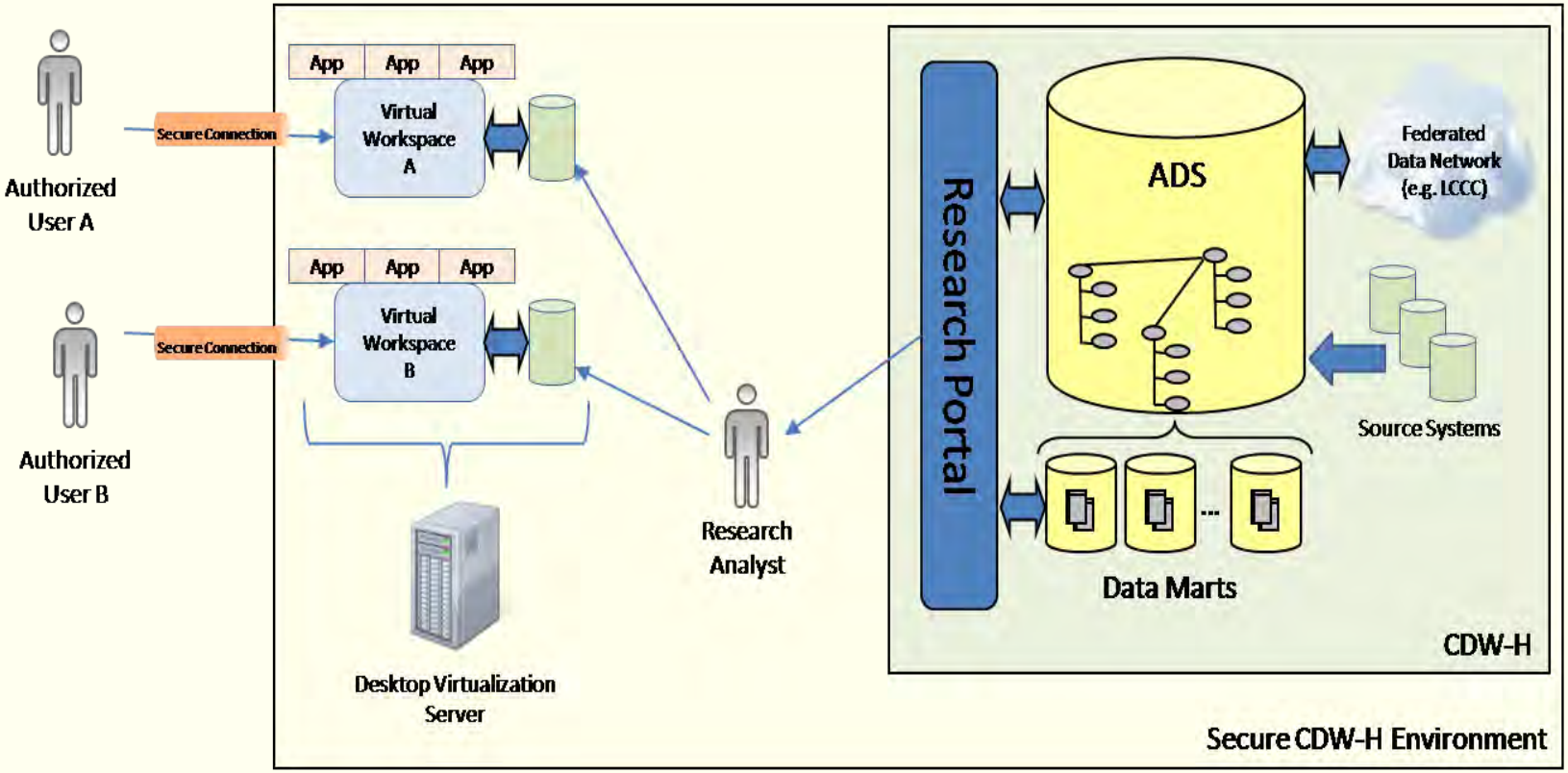
iRODS gracefully allows for introducing control:

- Data movement and replication
- Metadata standards
- Archival, deletion, and retention
- Integration with workflows, hadoop, databases
- Hiding complexities
- Automation
- …, all policy driven
- …, without breaking the in-place systems

ExoGENI Rack Software Stack



Secure Virtual Research Workspace Architectural Overview



Conclusion

- Challenges in Big Data
  - Scientific Data Explosion & Role of Librarians
- Projects at UNC
  - Gearing to Meet the Challenges
- Looking Towards the Future
  - Integration of Data, Computing & Networks