Current thoughts at NHLBI regarding a Data Commons

David Goff, MD, PhD

Director
Division of Cardiovascular Sciences
National Heart, Lung, and Blood Institute

March 11, 2017

Charting the Future Together: The NHLBI Strategic Vision

CHARTING THE FUTURE TOGETHER

The NHLBI Strategic Vision

U.S. Department of Health and Human Services
National Institutes of Health
National Heart, Lung, and Blood Institute

Accelerating our journey towards scientific and health advances in the next decade

Understand Human Biology
Develop Workforce and Resources
Reduce Human Disease
Advance Translational Research

The NHLBI Strategic Vision: Objectives

4 mission-oriented goals are at the center of the Strategic Vision

8 objectives provide a framework for moving HLBS science forward

1. Understand normal biological function and resilience

2. Investigate newly discovered pathobiological mechanisms important to the onset and progression of HLBS diseases

3. Investigate factors that account for differences in health among populations

4. Identify factors that account for individual differences in pathobiology and in responses to treatments

5. Develop and optimize novel diagnostic and therapeutic strategies to prevent, treat, and cure HLBS diseases

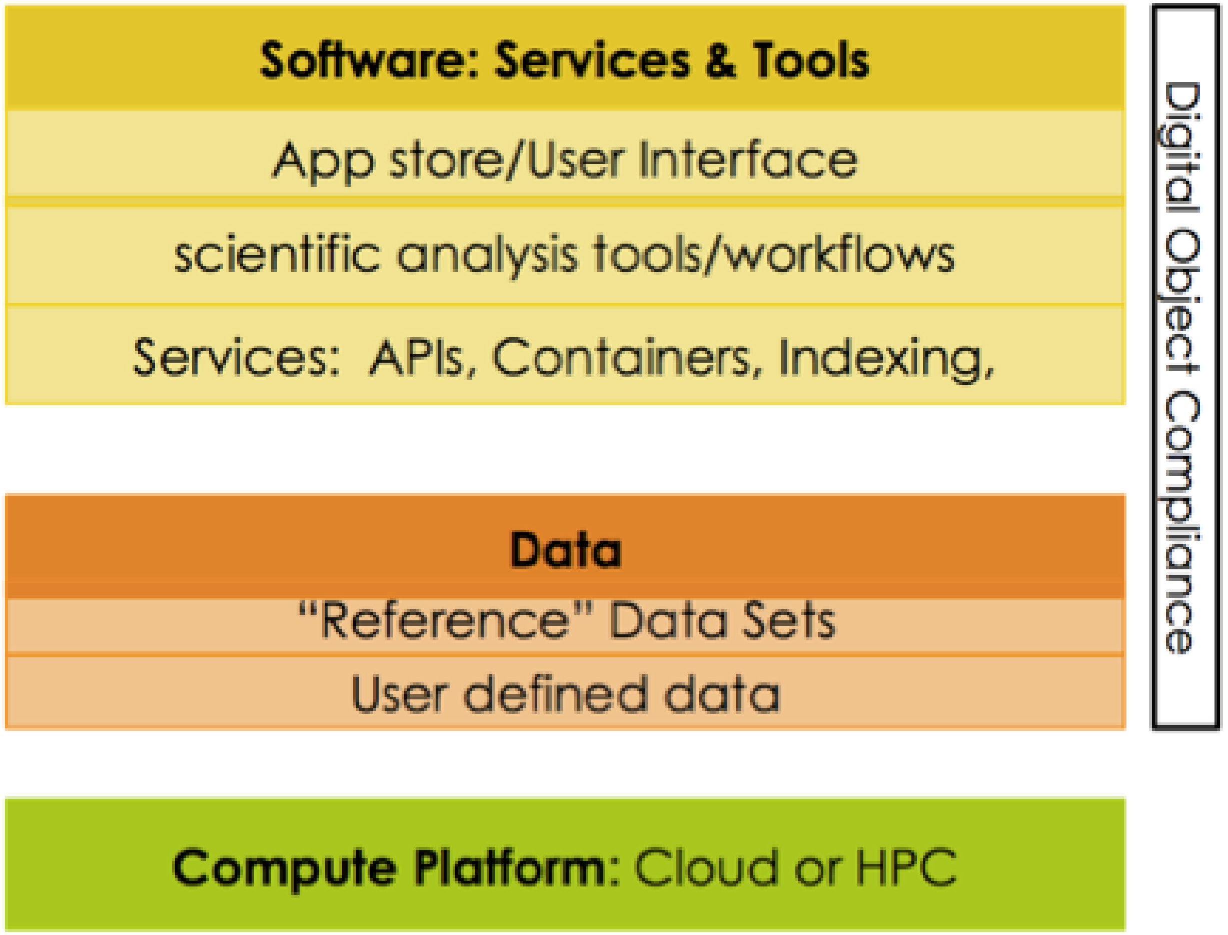
6. Optimize clinical and implementation research to improve health and reduce disease

7. Leverage emerging opportunities in data science to open new frontiers in HLBS research

8. Further develop, diversify, and sustain a scientific workforce capable of accomplishing the NHLBI's mission

NIH Commons Framework

The Commons is a shared virtual space where scientists can work with the digital objects of biomedical research, i.e. it is a system that will allow investigators to find, manage, share, use and reuse data, software, metadata and workflows. It will be a complex ecosystem and thus the realization of the *Commons* will require the use, further development and harmonization of several components.



Software: Services & Tools

App store/User Interface

scientific analysis tools/workflows

Services: APIs, Containers, Indexing,

Data

"Reference" Data Sets

User defined data

Compute Platform: Cloud or HPC

Digital Object Compliance

https://datascience.nih.gov/commons

The Data Lake

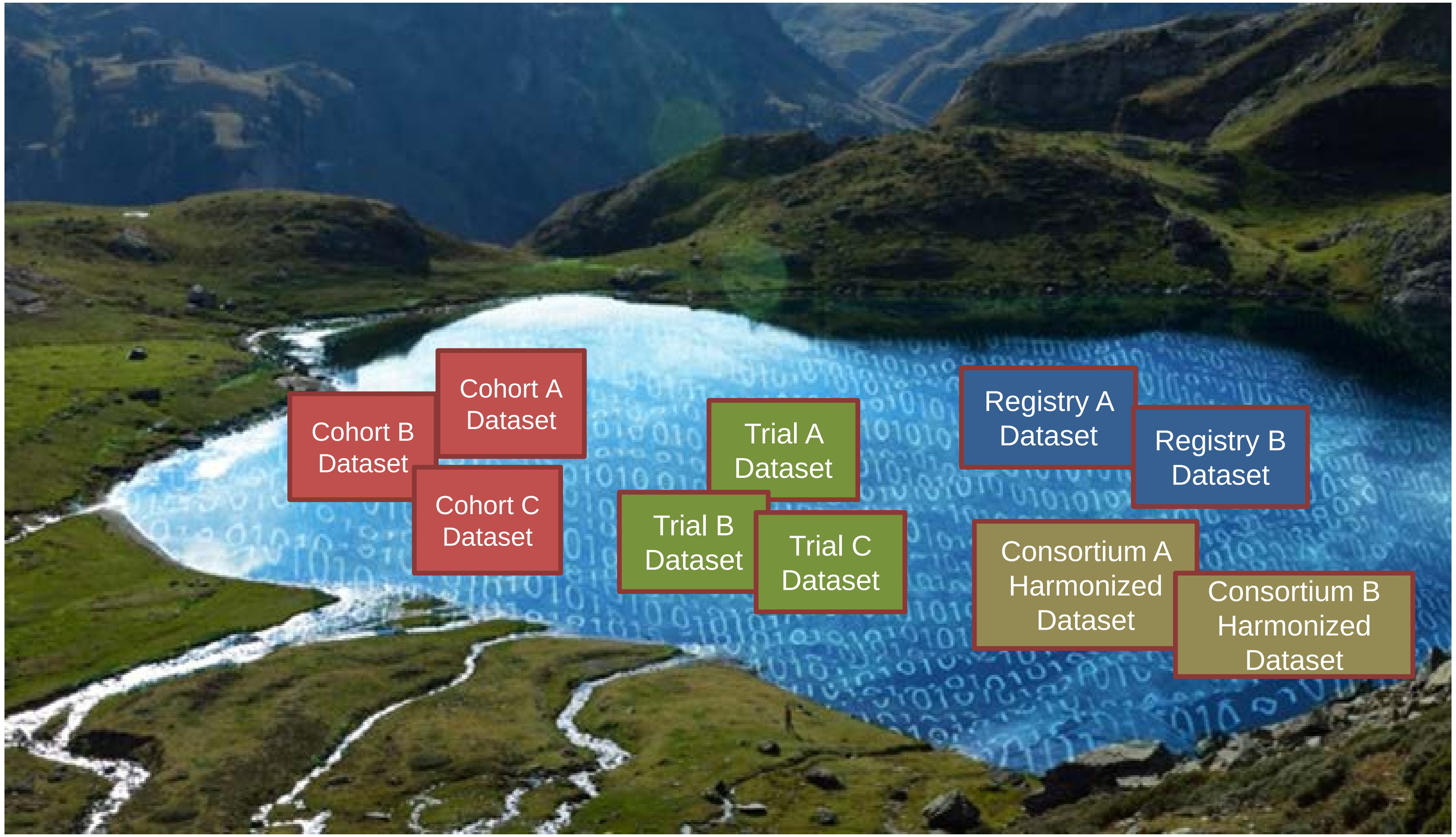
Registries with clinical data and biological samples (GTEX; ENCODE; HMP)

Datasets & Biospecimen from over 100 Clinical & Epi studies

Longitudinal Phenotypic data from diverse populations

Individual participant data from practice-changing clinical trials

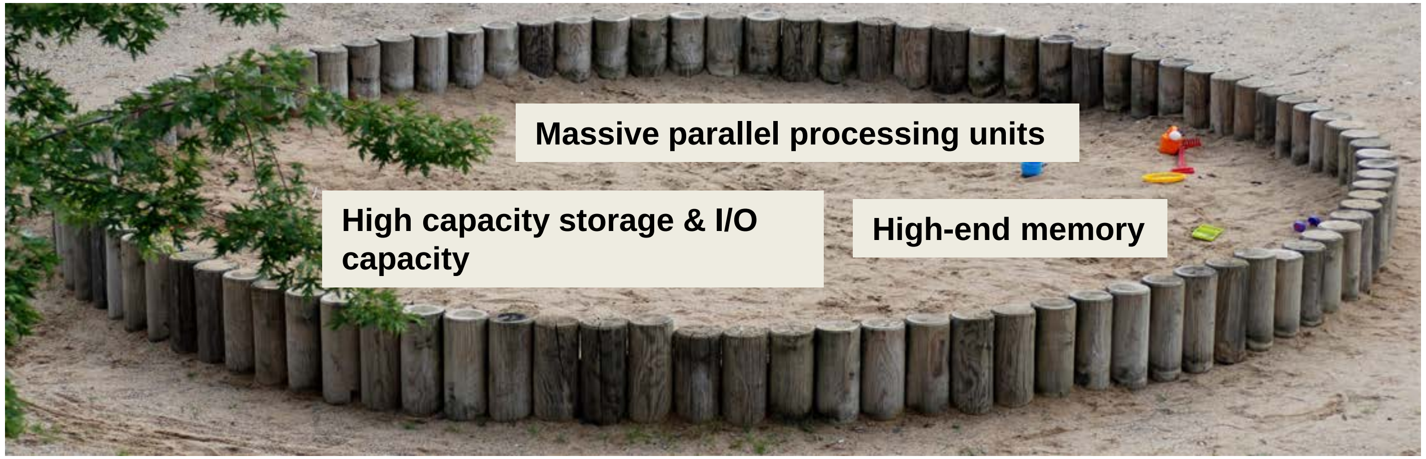
Genomics and Phenomics data from diverse HLBS cohort/clinical studies



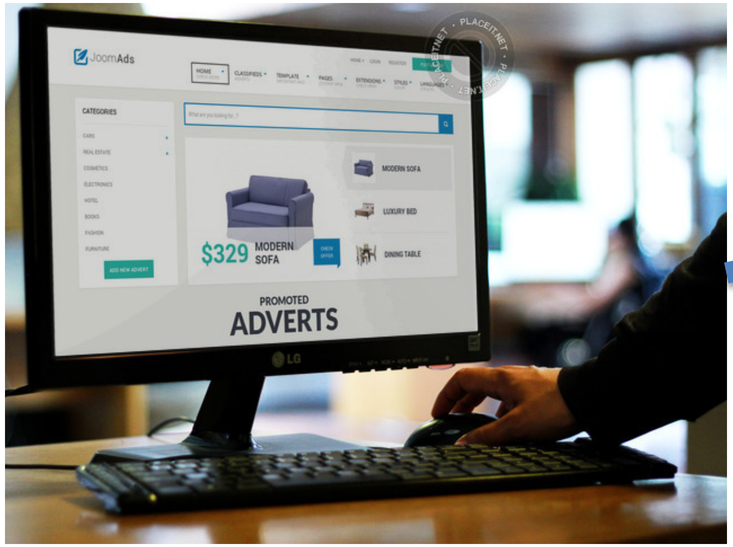
- storage repository that holds a vast amount of raw data in its native format until it is needed.

The Data Commons Sandbox

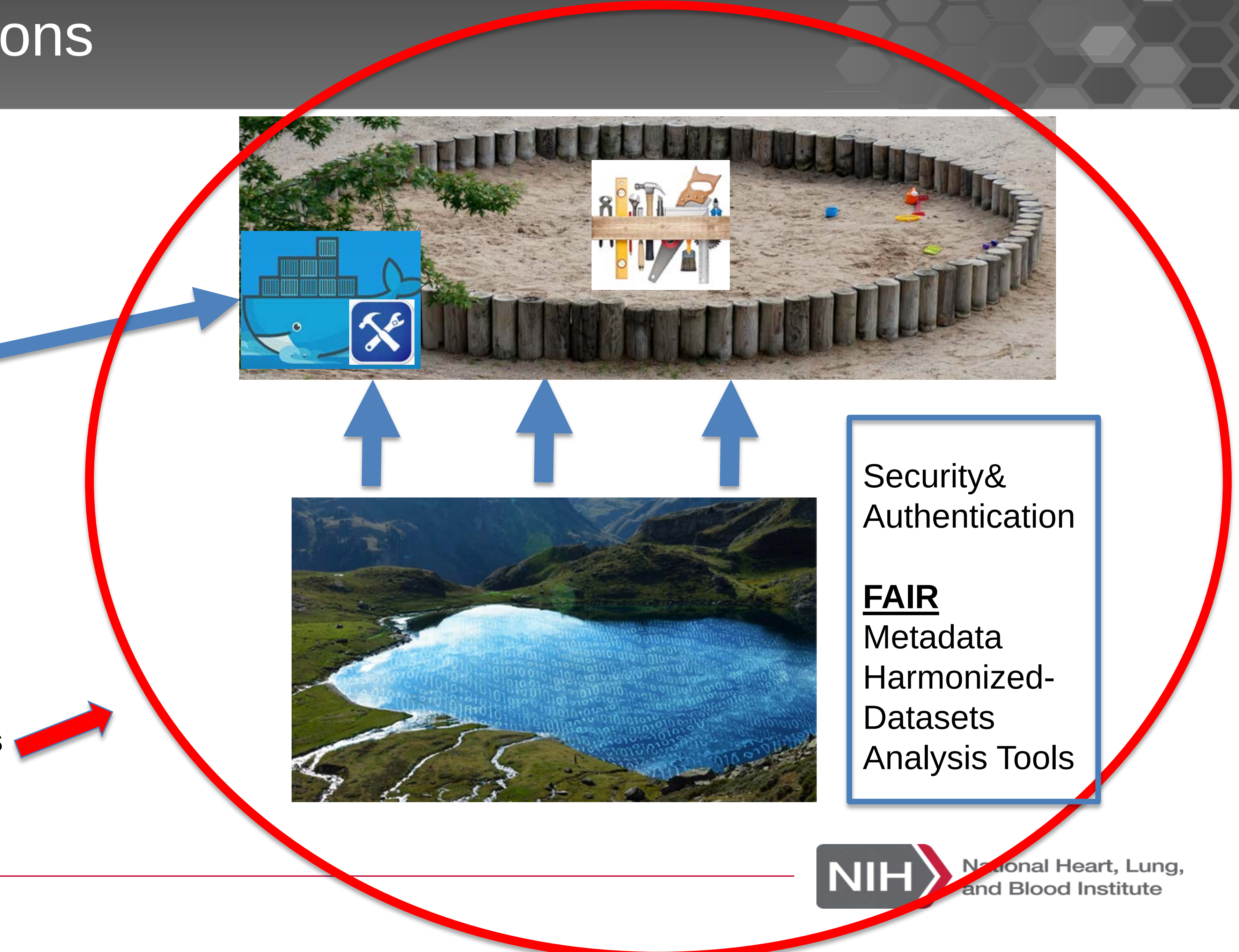
- testing environment that isolates untested code changes and outright experimentation from the production environment or repository


Massive parallel processing units
High capacity storage & I/O capacity
High-end memory

The Data Commons



Approved End User



Security& Authentication

<u>FAIR</u>
Metadata
Harmonized-Datasets
Analysis Tools

The Data Commons

Data Commons Implementation Challenges

- Robust data sharing protocols and standardized documentation

- Large scale data curation and integration among disparate datasets

- High performance computing and network infrastructure
  - Big data and complex processing

FAIR Principles

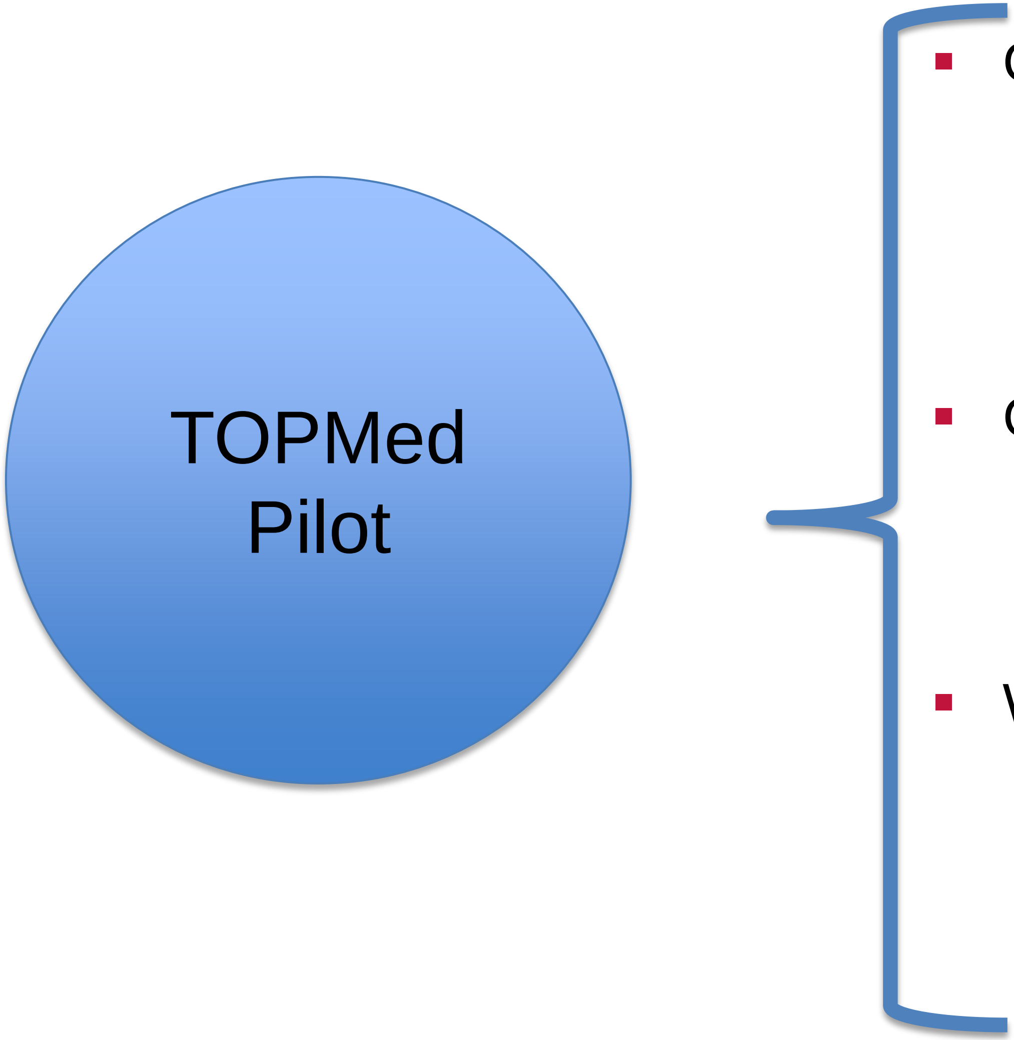
- **F**indable
  - Unique and permanent IDs
  - Metadata for searching

- **A**ccessible
  - Retrievable by ID
  - Authentication/authorization
  (data use limitations)

- **I**nteroperable
  - Broad, applicable language - harmonization
  - Across CSPs, datasets, tools, APIs

- **R**eproducible
  - Historical record of the inputs, entities, systems that influence the data

TOPMed Data Commons Pilot: A Potential Use Case

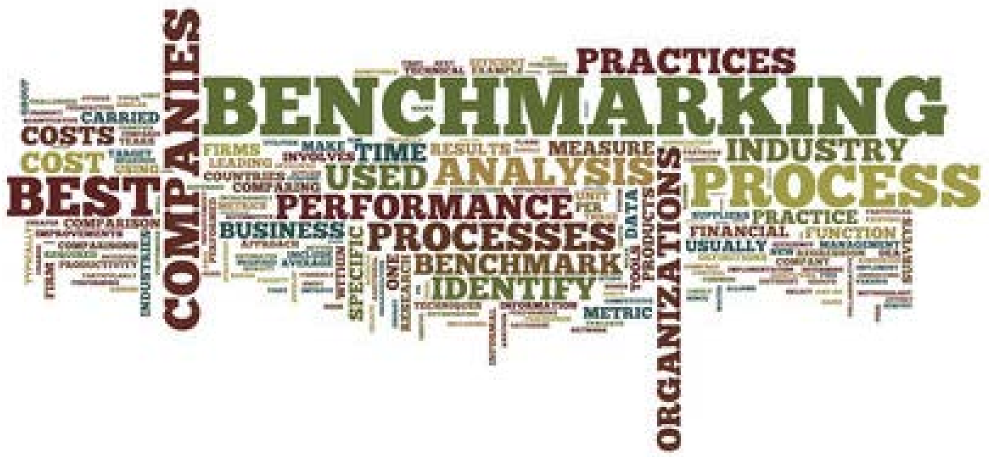


TOPMed Pilot

- Genome-phenome resource
  - Cloud providers (interoperable)
  - Expert users
  - Joint variant calling - across cohorts
  - Analysis
- Graphical User Interface
  - Wider user base
  - Metadata genome/phenome search (findable, reusable)
- Widest access
  - Search and analysis
  - Harmonized data elements
  - Controlled-access/authentication (accessible)
  - Authority to Operate

Where are we now?

- **Benchmarking**
  - NCI cloud pilots, lessons learned
  - Explore potential cloud service providers
  - Big Data to Knowledge (BD2K) collaboration



- **Business plan**
  - Government, Industry, Academia
  - Use what is available – don’t reinvent the wheel

- **NHLBI Data Science Workshop**
  - Planning phase: May, 2017; Chair: Veronique Roger
  - To enable integrated, big-data analysis from NHLBI’s clinical studies
  - Will engage experts in cohorts, trials, computations, data science

Considerations/Questions

- Reality check – value of extensive harmonization efforts, and best approach?
- Retrospectively integrating already-collected disparate datasets (harmonization) vs. prospectively collecting “integratable” datasets (using common data elements)
- How should harmonization be accomplished?
- What data standard(s) should be encouraged for new studies?
- How do we address data sharing limitations imposed by existing consents (e.g. no industry) vs. re-consent of new studies for broader sharing?