The NIH Data Commons

Council of Councils

Vivien Bonazzi, PhD

May 26, 2017

The NIH Data Commons – Update on Progress

- Provide a brief background
- Explain evolution of thought
- Propose an approach
- Obtain your perspectives on the proposed approach and plan
- And, what are the deliverables?

Background

A conceptual framework for data management

- June 25, 2015 – Data Science Update from Phil Bourne to IC Directors
  - Introduced the Commons as a way to sustain, track and analyze data
  - Stressed FAIR principles: Findable, Accessible, Interoperable, Reusable
  - Described projects that were piloting various aspects of a Commons with various datasets

- November, 2016 – NIH Director convened a Task Force for Data Science to develop recommendations for launching an NIH Data Commons: how to bring it all together
  - Recommended piloting an NIH Commons with 3 high-value but disparate datasets
    - NHLBI's TOPMed – data from multiple cohorts
    - NHGRI's Model Organism Databases – most-used MOs
    - Common Fund's GTEx

Need to establish and test a physical substantiation of the Commons concepts

- January, 2017 – Phil Bourne resigned as ADDS; DPCPSI charged with moving forward with the Commons Pilot

Clarification

- The NIH Data Commons will include:

The NIH Data Commons … Sharable, More Cost-Effective Compute and Storage Platforms

The NIH Data Commons … A Set of *Community-Endorsed* Unifying Principles and Standards

TOPMed

GTEx

The NIH Data Commons … a Portal to Find and Access Data Sets and Tools

Clarification

- Some pilots are underway to test the Commons Concept in the ICOs.
  - The NIH Data Commons Pilot will provide a framework for integrating across

- BD2K funds will provide resources to:
  - *Support people, process, and tools to find data, access it, and work with it*
    - **PLATFORMS**
    - **PORTAL**
    - **FUNCTIONALITIES**
    - **GUIDELINES**
    - **SUSTAINABLE ENVIRONMENT**
  - *Evaluate/analyze how data and tools are used in the cloud*
  - *Evaluate costs/business practices of Commons management*
- *ICOs will need to commit additional resources to enable their datasets to adopt FAIR guidelines and become integrated into the NIH Data Commons*

- Costs to sustain – and locus of responsibility - are currently unknown

Data Science Challenges → Deliverables

- Establish metadata standards for data types (where needed)
- Crosswalks between equivalent terms/ontologies
- Robust, shared approaches to data access/authentication for access to controlled-access data
- A platform for data interoperability
- Encourage transition of tools in data sets to tools that operate across data sets
- Determine value of cloud storage and processing in the cloud versus other environments
- Develop policies from the lessons learned and incorporate them within future terms of award.
- Best practices that will enable existing data to become FAIR and will guide generation of future datasets

Management Challenges → Deliverables

- Rapidly evolving field makes approaches/tools/etc subject to change – approaches need to be adaptable
- Effort is required to adapt data to community standards and move data to the cloud
  - How much does that cost and how long does it take?
- Cumbersome processes for contracting with cloud providers and Trusted Partner organizations
- Lack of interoperability between cloud providers
- Making data FAIR comes with a cost
  - How much does it actually cost?
  - How can we minimize the cost?
  - How do we determine whether any one set of data warrants the expense?
- What is the value added to the data by making it FAIR?
  - What new science can be achieved?
  - How can new derived data or new computational approaches be incorporated into the Commons?
  - What are the limitations of FAIR'ness from dataset to dataset?

Proposed Approach: 5 Components

Developing and Testing Compute/Storage Marketplace

Establishing Community-Endorsed Unifying Principles and Standards

Portal to Find and Access Data Sets and Tools

Learning by doing

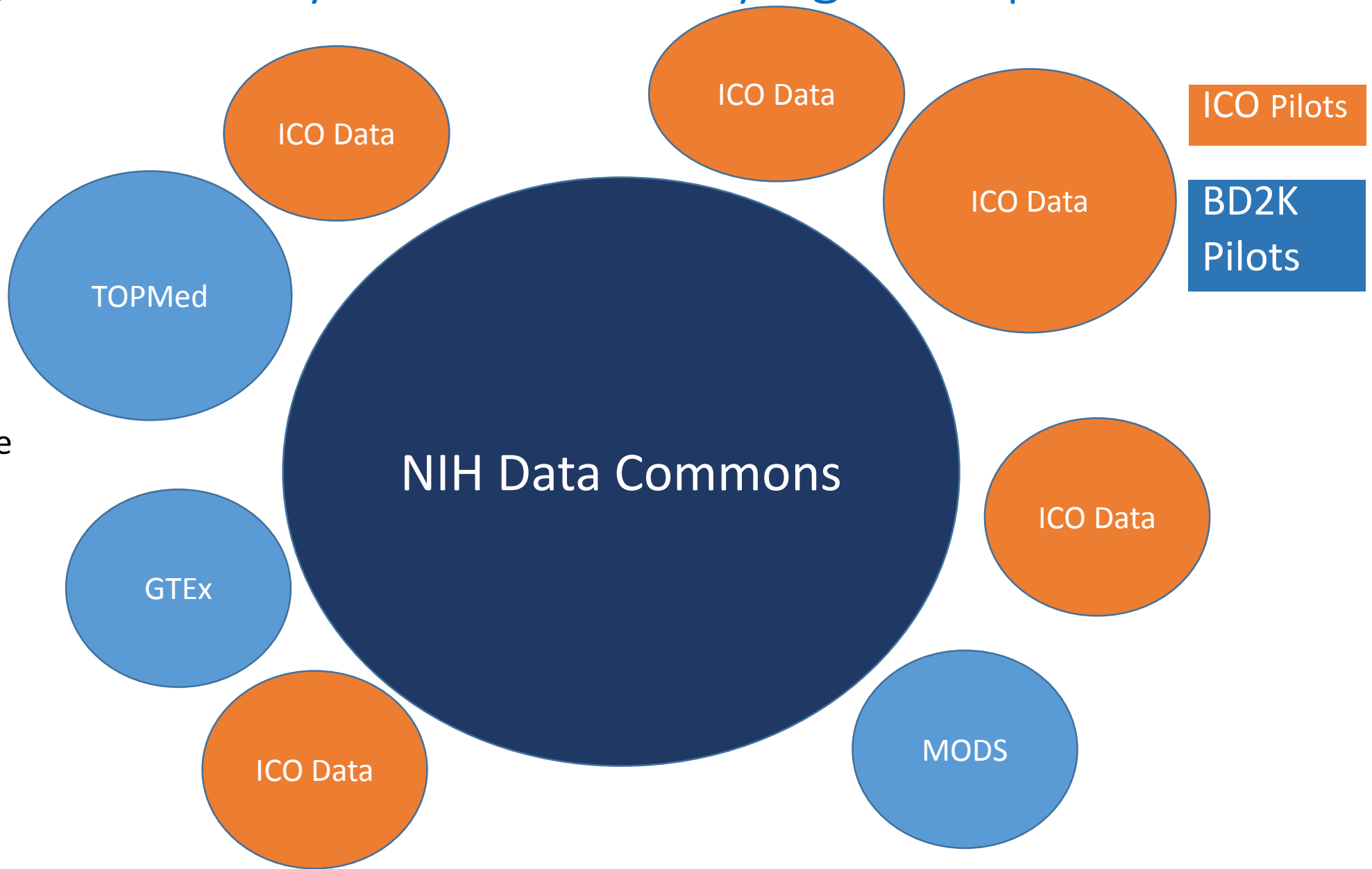
Kicking the Tires

Component 1:
Developing and Testing Compute/Storage Marketplace

- ***Contract:*** Will leverage relationships established as part of BD2K, and develop a “marketplace” for ICOs to easily engage with cloud providers, resellers, and trusted partners
- ***CIT:*** Will work with contractor and transition the marketplace to CIT for a sustainable service to all components at NIH
- ***ICO’s:*** may bring their own funds to the marketplace for individual datasets in a process to be developed via the contract

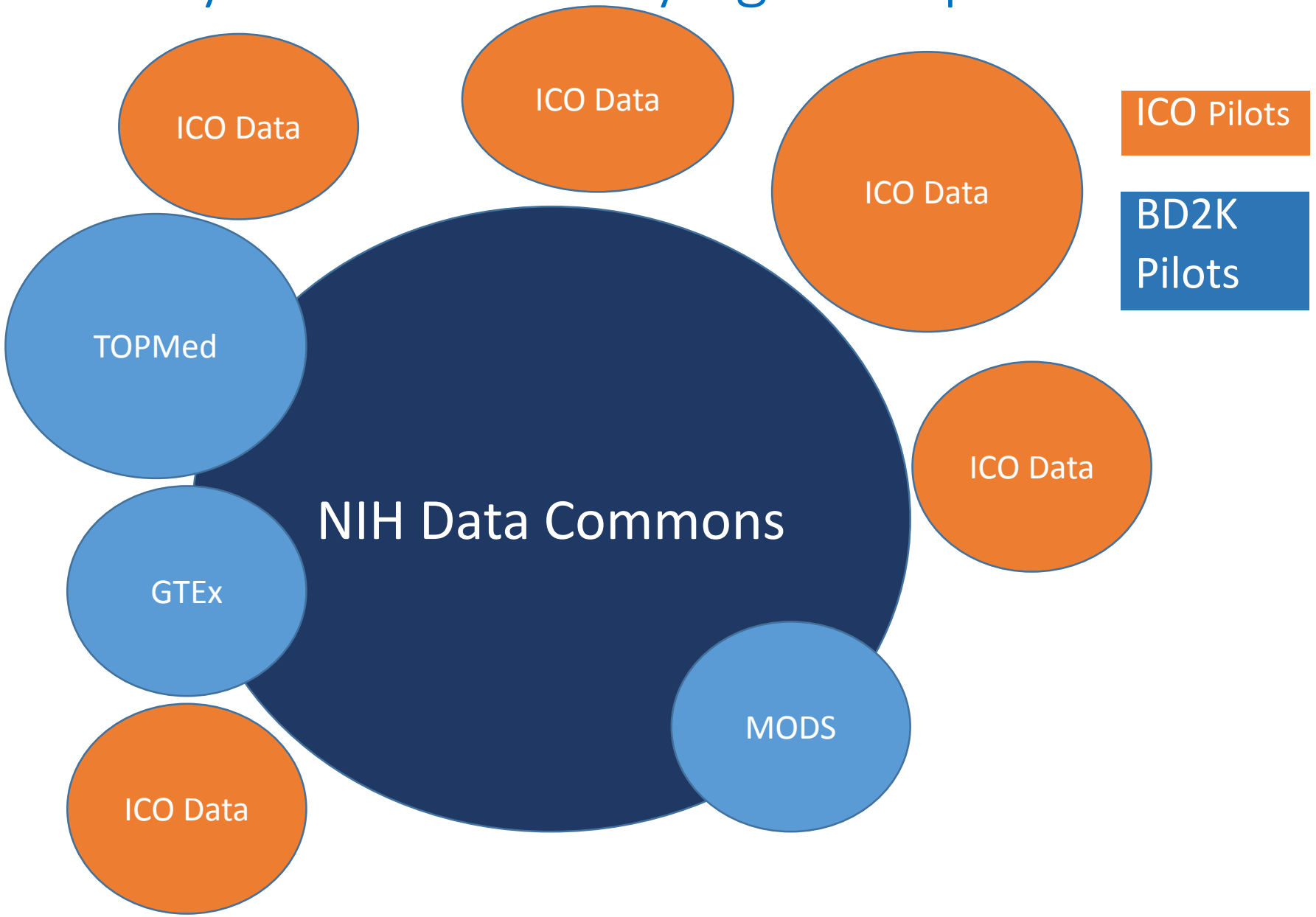
Component 2: Establishing Community-Endorsed Unifying Principles and Standards

- Current state of the art for FAIR Guidelines
- FAIR Metrics
- Current technical limitations to guide technology/tool enhancements



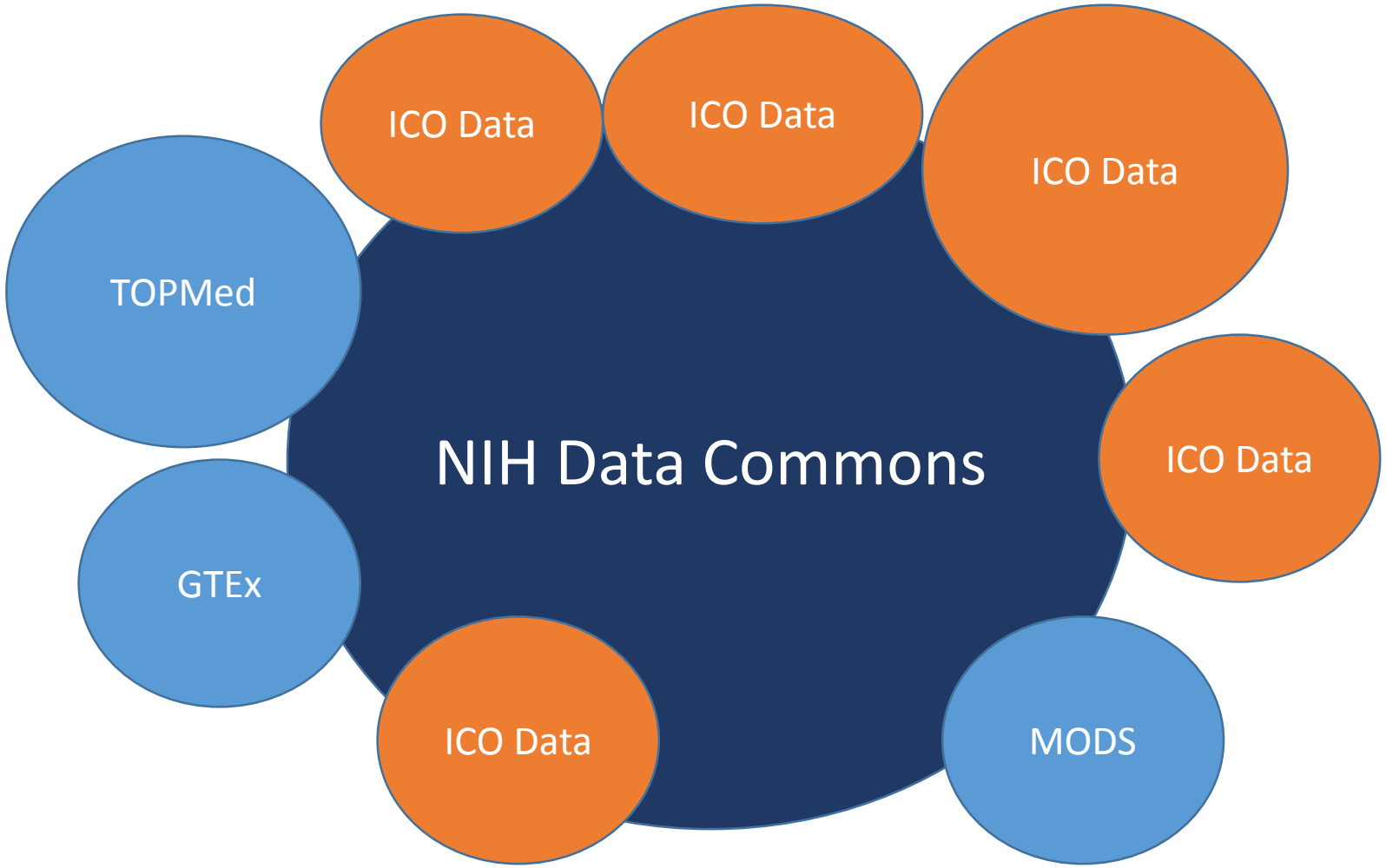
Component 2: Establishing Community-Endorsed Unifying Principles and Standards

- Current state of the art for FAIR Guidelines
- FAIR Metrics
- Current technical limitations to guide technology/tool enhancements



Component 2:
Establishing Community-Endorsed Unifying Principles and Standards

- Current state of the art for FAIR Guidelines
- FAIR Metrics
- Current technical limitations to guide technology/tool enhancements



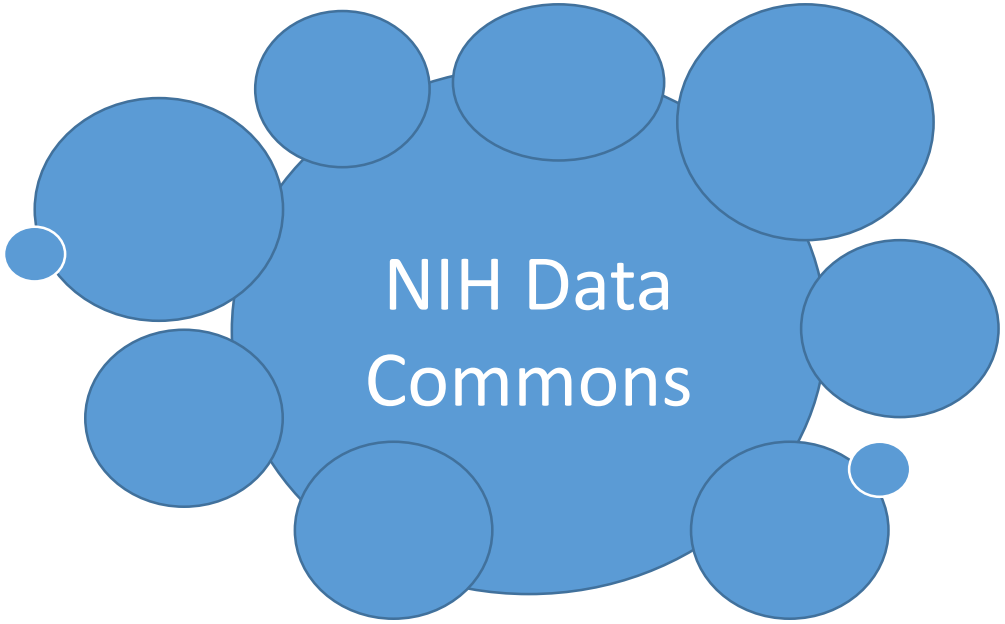
Component 3: Portal to Find and Access Data Sets and Tools

- Develop a portal through which users of all levels of expertise can access data
- Incorporate new authentication/authorization protocols
- Develop an “app store” model for access to tools; allows new tools to be added
- Establish collaborative workspaces for expert and general researchers
- Enable researchers to easily search data and analyze data across different data types and sets
- Provide an environment to test and evaluate existing/emerging open data, workflow tools, and technology standards

The goal is to expand to other datasets as an NIH-wide resource

Component 4:
Learning by doing

- Establish/reuse data architecture to enable FAIR'ness of TOPMed, GTEx, and MODs – *extensible to ICO datatypes and datasets*
- Support these datasets to adopt community-endorsed principles and standards
- Migrate data to cloud environment
- Establish new use cases to drive evolution
- Encourage use cases that involve ICO Commons Pilots
- Support development of tools/methods to test the use cases
- Assess scientific value of Commons



Component 5: Kicking the Tires

Contract: Analyze and evaluate financial and process components:

- What are the real costs?
- What business models and practices are most effective…
  - To pay for interactions with cloud providers/Trusted Partners?
  - To charge users for use of the data?
  - To add new data and tools to datasets in the Commons?
  - To add new datasets?
- Who uses the data? How often?

Open calls to the community will:

- develop new use cases and test the ability of the Commons structure to allow those to be tested
- develop new metrics of FAIRness so that the system can improve
- test usability of the portal

Budget and Deliverables

FY17	FY18	FY19	FY20
$10.5 M	$15 M	$15 M	$15 M

Costs are estimates!

- Deliverables
  - Portal through which users can find, access, and analyze data; add tools
  - Robust authentication/authorization process for controlled-access data
  - Extensible data architecture, tools, processes, guidelines for data management
  - Cost-effective strategies for engaging cloud providers
  - Assessment of costs for enabling FAIR data – costs for storage, access, use, and growth
  - Assessment of scientific value-added for FAIR data management
  - Business model to manage costs, charging for some uses to users

Commons Pilot Working Group:

- Jim Anderson (DPCPSI)
- Vivien Bonazzi (ADDS/DPCPSI)
- Patti Brennan (NLM/iADDS)
- Valentina Di Francesco (NHGRI)
- Gary Gibbons (NHLBI)
- Maria Giovanni (NIAID)
- Eric Green (NHGRI)
- Warren Kibbe (NCI)
- JJ McGowan (NIAID)
- Marie Nierras (DPCPSI)
- Andrea Norris (CIT)
- Mary Perry (DPCPSI)
- Ajay Pillai (NHGRI)
- Alastair Thomson (NHLBI)
- Simona Volpi (NHGRI)
- Betsy Wilder (DPCPSI)
- Ken Wiley (NHGRI)