NIH Big Data to Knowledge (BD2K)

Data Science & BD2K Update

Philip Bourne, PhD, FACMI
Associate Director for Data Science

Advisory Committee to the NIH Director
June 10, 2016

http://datascience.nih.gov

Slides: http://www.slideshare.net/pebourne

Data Science Agenda

- What problems are we trying to solve?
- What are the solutions we are exploring?
- How does BD2K facilitate those solutions?

What Problems Are We Trying to Solve?

- Data are extensive, complex and growing
- Data are in silos while science transcends those silos
- Data are expensive to maintain and share while demands for sharing are increasing
- There is an insufficient workforce with the needed data analytical skills
- A collective (trans NIH) solution

NIH Big Data to Knowledge (BD2K)

Quantifying the Problem

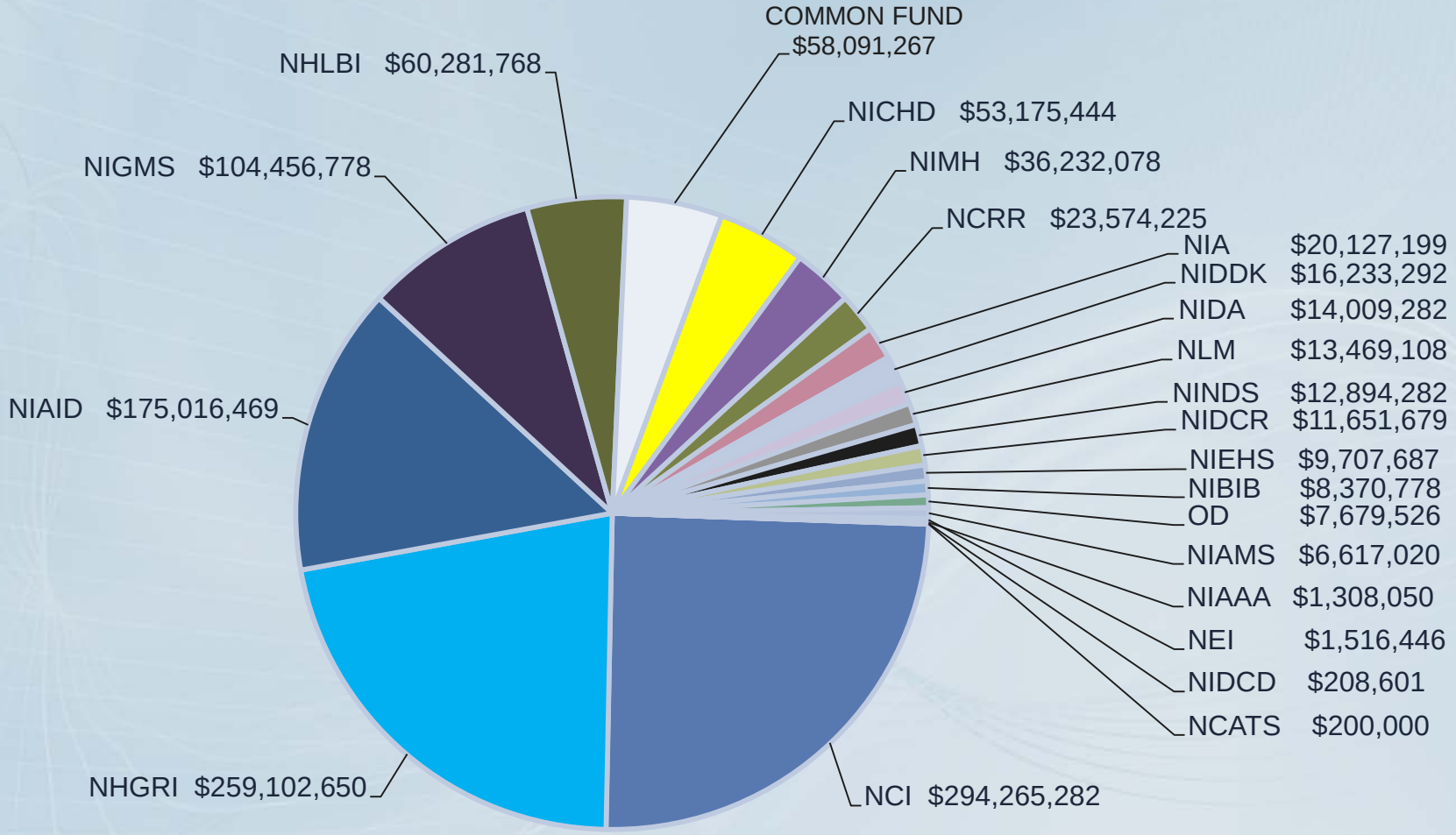
- Big Data
  - Total data from NIH-funded research currently estimated at 650 PB*
  - 20 PB of that is in NCBI/NLM (3%) and it is expected to grow by 10 PB this year
- Dark Data
  - Only 12% of data described in published papers is in recognized archives – 88% is dark data^
- Cost
  - 2007-2014: NIH spent ~$1.2Bn extramurally on maintaining data archives

* In 2012 Library of Congress was 3 PB

^ http://www.ncbi.nlm.nih.gov/pubmed/26207759

NIH Big Data to Knowledge (BD2K)

NIH Extramural Data Repository Funding, 2007-2014



Number of Extramural Data Repositories Sampled * 124
Grand Total of Awards 777
Estimated Gross Total Spent $1,188,188,911

NIH Big Data to Knowledge (BD2K)

**Note: Award data confirmed as of 03/2016. Some repositories funded by hybrid mechanisms (eg. grants-contracts, IAA-contracts, etc.)*

Biomedical Digital Data Repository Survey by Institute and Center (IC)

- Leadership meeting late in 2015 requested a survey of IC approaches and plans for data repositories
- Responses received from 18 IC's
- Clear challenges were identified

The Major Challenge Encountered When Considering Repository Funding

* Some IC's identified multiple challenges equally

NIH Big Data to Knowledge (BD2K)

What Solutions Are We Exploring?

The *Commons* is one solution that leverages the experiences in cloud-based computing and is being enabled by BD2K research

Examples of Cloud Based Initiatives

Program Snapshot

40TB AWS

The Common Fund's **Human Microbiome Project (HMP)** is developing research resources to enable the study of the microbial communities that live in and on our bodies and the roles they [illegible] human health and disease.

5 PB

The NCI Genomic Data Commons

The NCI Genomic Data Commons (GDC) is a unified knowledge base that promotes sharing of genomic and clinical data between researchers and facilitates precision medicine in oncology.

Cancer is fundamentally a disease of the genome, caused by mutations and other harmful genomic changes that alter its function and contribute to the malignant behavior of cancer cells. Genomic aberrations can influence the aggressiveness of tumors and the response of tumors to particular drugs.

The NCI Genomic Data Commons is housed at the University of Chicago Kenwood Data Center
Credit: University of Chicago

The Commons – The Internet of Data

The Commons offers a path forward to integrate these discreet cloud-based initiatives using BD2K developments to make data FAIR*

- Findable
- Accessible
- Interoperable
- Reusable

The internet started as discreet networks that merged - the same could happen with data

* http://www.ncbi.nlm.nih.gov/pubmed/26978244

NIH Big Data to Knowledge (BD2K)

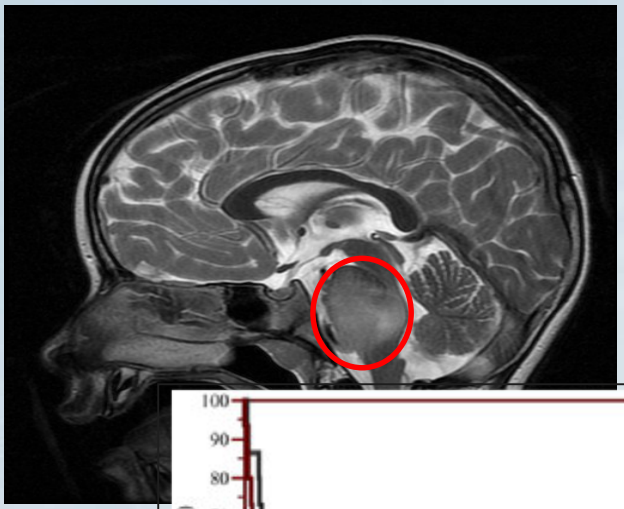
Use Case:

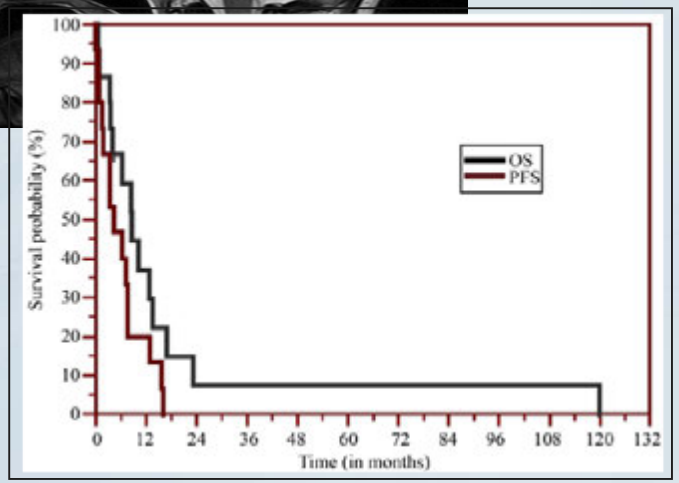
Aggregate integrated data offers the potential for new insights into rare diseases …

As we get more precise every disease becomes a rare disease

Diffuse Intrinsic Pontine Gliomas (DIPG): In need of a new data-driven approach

- Occur 1:100,000 individuals.
- Peak incidence 6-8 years of age.
- Median survival 9-12 months.
- Surgery is not an option
- Chemotherapy ineffective and radiotherapy only transitive.





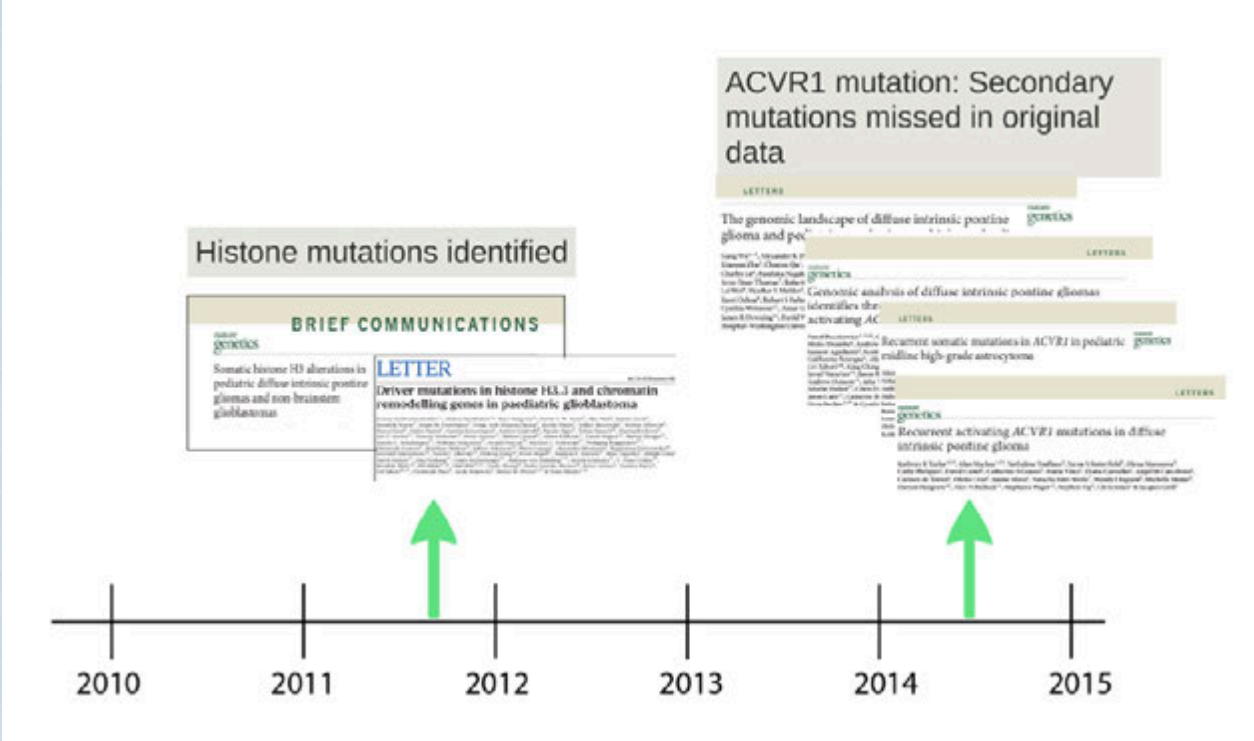
From Adam Resnick

NIH Big Data to Knowledge (BD2K)

Timeline of Genomic Studies in DIPG

- Landmark studies identify histone mutations as recurrent driver mutations in DIPG ~2012

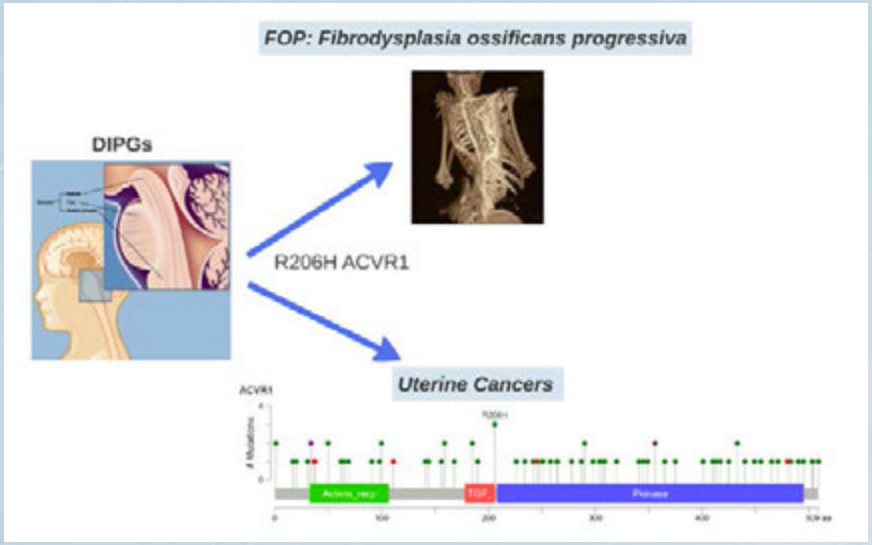
- Almost 3 years later, in largely the same datasets, but partially expanded, the same two groups and 2 others identify ACVR1 mutations as a secondary, co-ocurring mutation.



From Adam Resnick

NIH Big Data to Knowledge (BD2K)

Hypothesis: The Commons would have revealed ACVR1



- ACVR1 is a targetable kinase
- Inhibition of ACVR1 inhibited tumor progression in vitro
- ~300 DIPG patients a year
- ~60 are predicted to have ACVR1
- If large scale data sets were only integrated with TCGA and/or rare disease data in 2012, ACVR1 mutations would have been identified
- 60 patients/year X 3 years = 180 children's lives (who likely succumbed to the disease during that time) could have been impacted if only data were FAIR

From Adam Resnick

NIH Big Data to Knowledge (BD2K)

A 3 Year BD2K Sponsored Commons Pilot is Under Way

– Questions to be addressed:

- Does the ability to compute across very large datasets lead to new discoveries?
- Are data and analytics more easily located and shared and does this improve productivity?
- Is there an advantage to have the results of those large calculations also available?
- Is research more reproducible?
- Is this environment more cost-effective than what we do now?

NIH Big Data to Knowledge (BD2K)

Another use case…

Let’s review the Commons pilot using the Model Organism Databases (MODs) as an example …

Example of the Problem:

The Model Organism Databases (MODS)

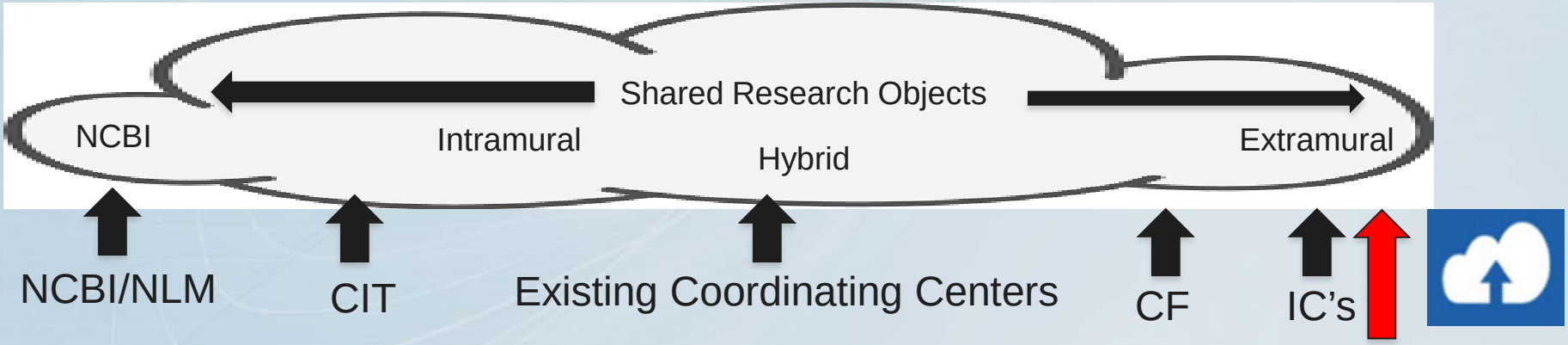
NHGRI & NHLBI

- Highly curated and valuable data
- Siloed / Not interoperable
- Cumbersome to compute over all the data
- Costly to maintain as individual resources

NIH Big Data to Knowledge (BD2K)

Step 1: Data & Analytics Moved to the Commons

- Moved as *Commons compliant* shared research objects, including:
  - Identifiers
  - Minimal metadata standards



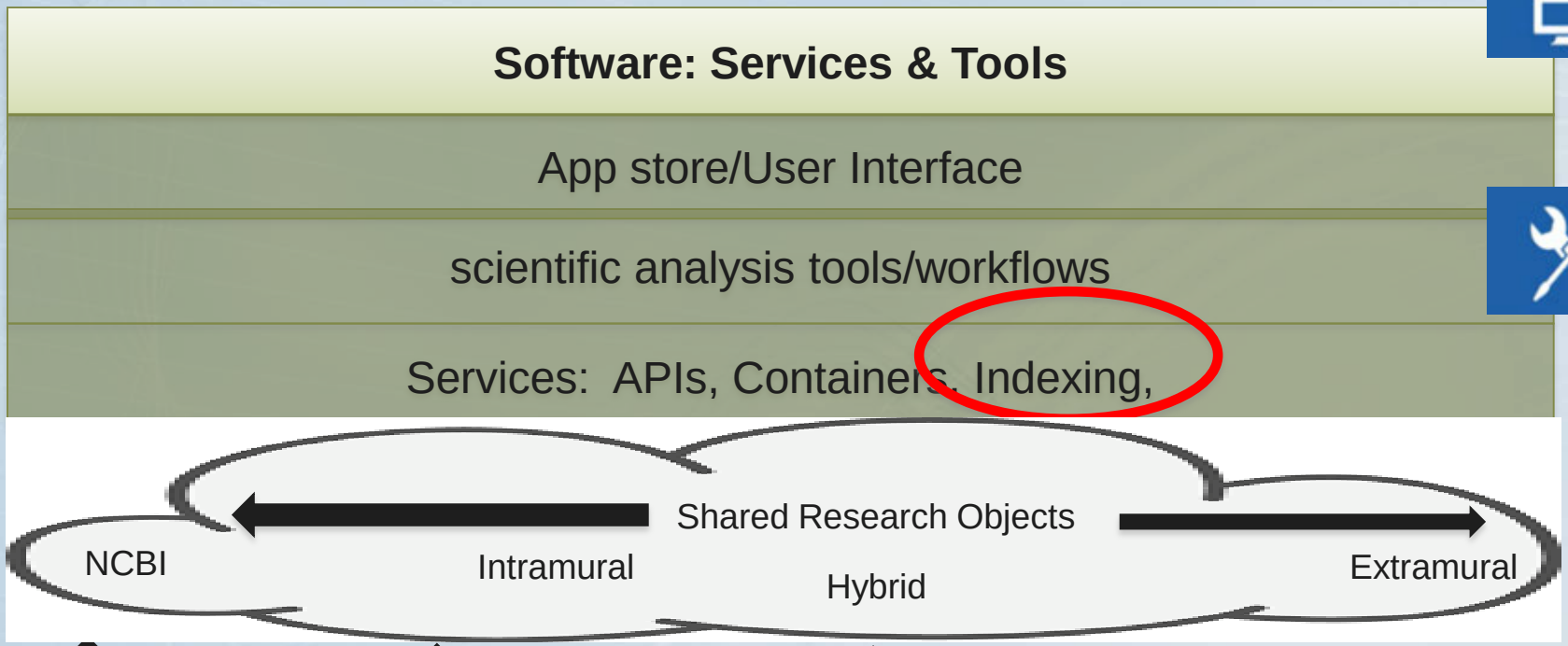
NIH Big Data to Knowledge (BD2K)

BD2K is Providing Those Metadata Standards

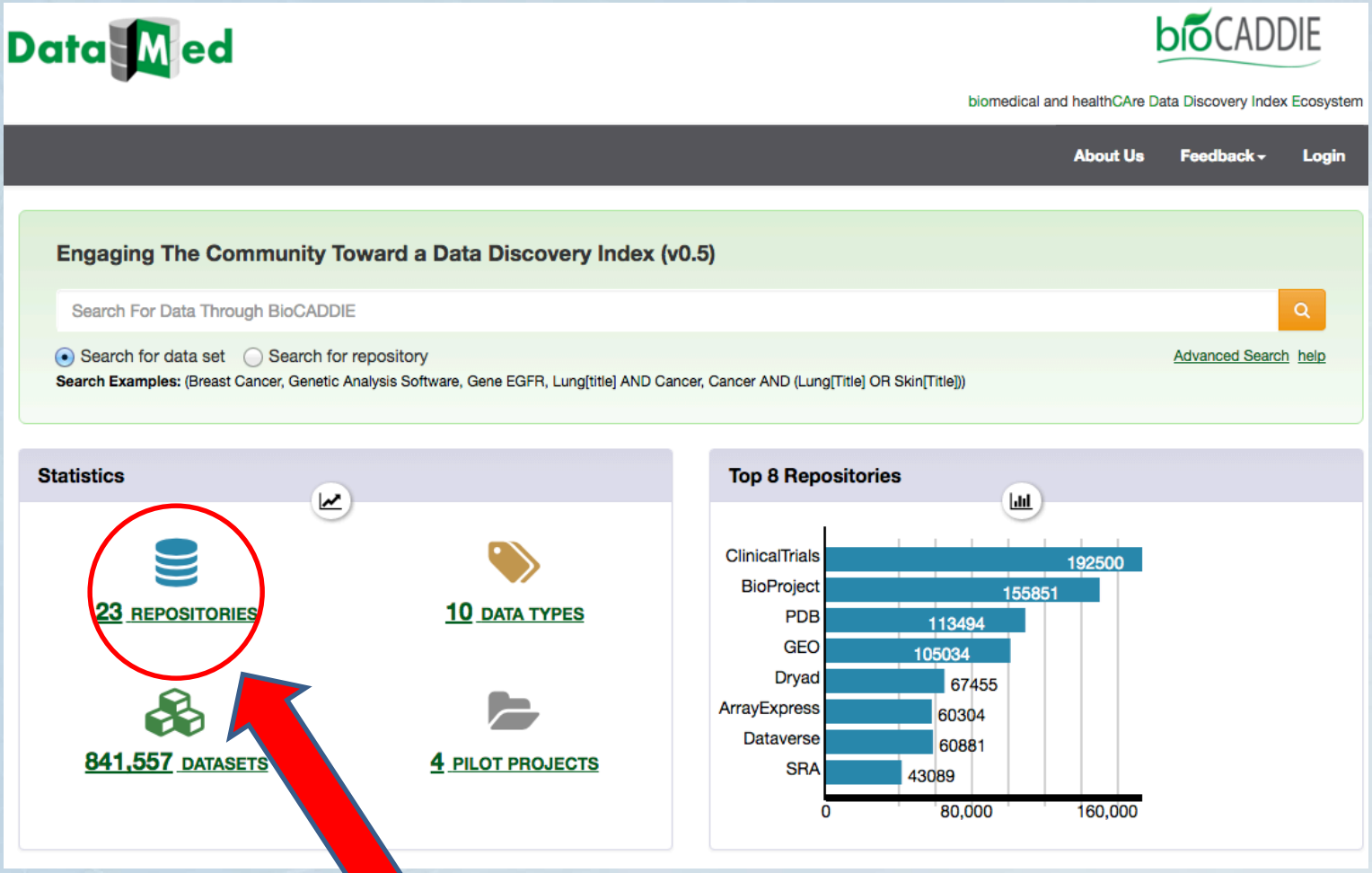
BD2K
BD2K
Standards Coordinating Center

NIH Big Data to Knowledge (BD2K)

Step 2: Layers of Software & Services Added



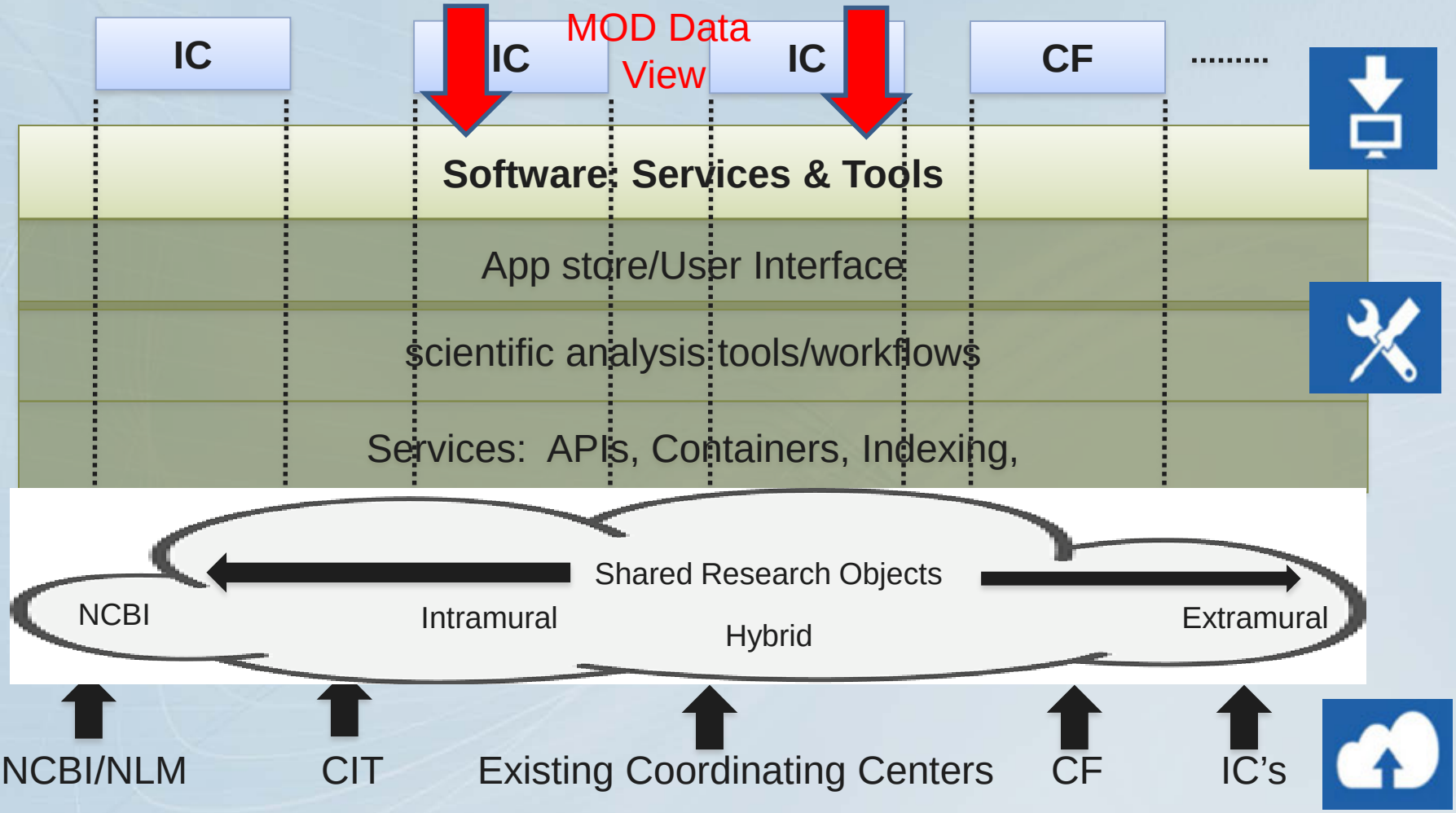
DataMed is a “Find” Service Developed by BD2K



MOD Data indexed

NIH Big Data to Knowledge (BD2K)

Step 3: Commons Content Shared While Maintaining Autonomous Views



NIH Big Data to Knowledge (BD2K)

BD2K Commons Pilot Timeline

Step 0: Initiation

- Finalize conformance requirements
- Arrange initial providers

Step 1: ~5 Initial Projects including Common Fund ← **We Are Here**

Step 2: ~50 projects

Step 3: Evaluation & Next Steps

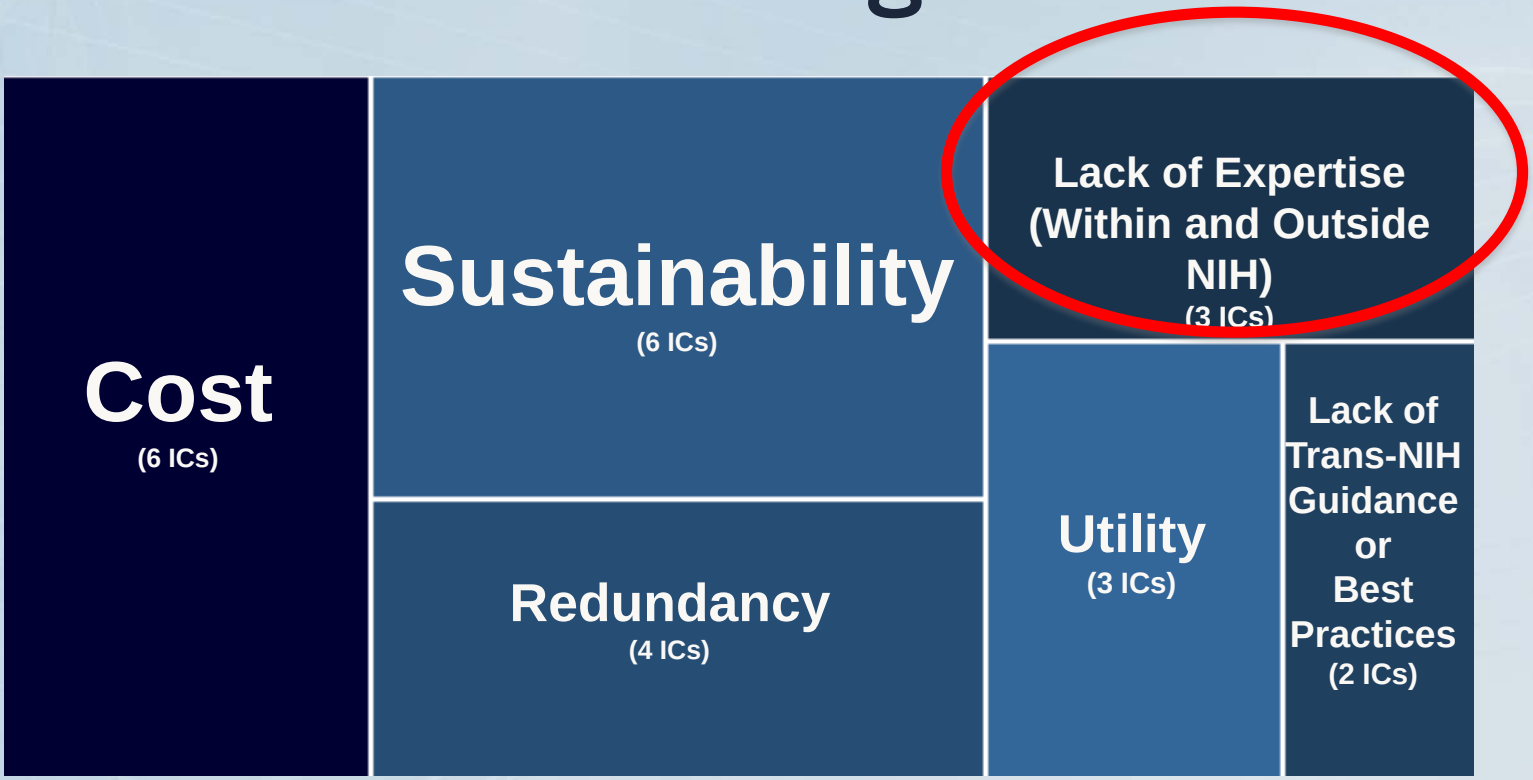
Project Year 1
FY 2015
Oct 2015 – Sep 2016

Project Year 2
FY 2016
Oct 2016 – Sep 2017

Project Year 3
FY 2017
Oct 2017 – Sep 2018

NIH Big Data to Knowledge (BD2K)

The Major Challenge Encountered When Considering Repository Funding



* Some IC's identified multiple challenges equally

NIH Big Data to Knowledge (BD2K)


16 T32/T15
Predoctoral
Training
Programs
21
Postdoctoral and
Faculty Career
Awards

Enhancing Diversity

- Focus on low-resourced institutions
  - Supports curriculum and faculty development
  - Supports research experiences for undergraduates
- Builds partnerships with BD2K Centers

Improving Data Science Skills Among all Biomedical Scientists

BD2K supports the development of individual educational resources

BD2K supports an Educational Resource Discovery Index that enables scientists to find relevant materials

NIH Big Data to Knowledge (BD2K)

The Role of BD2K

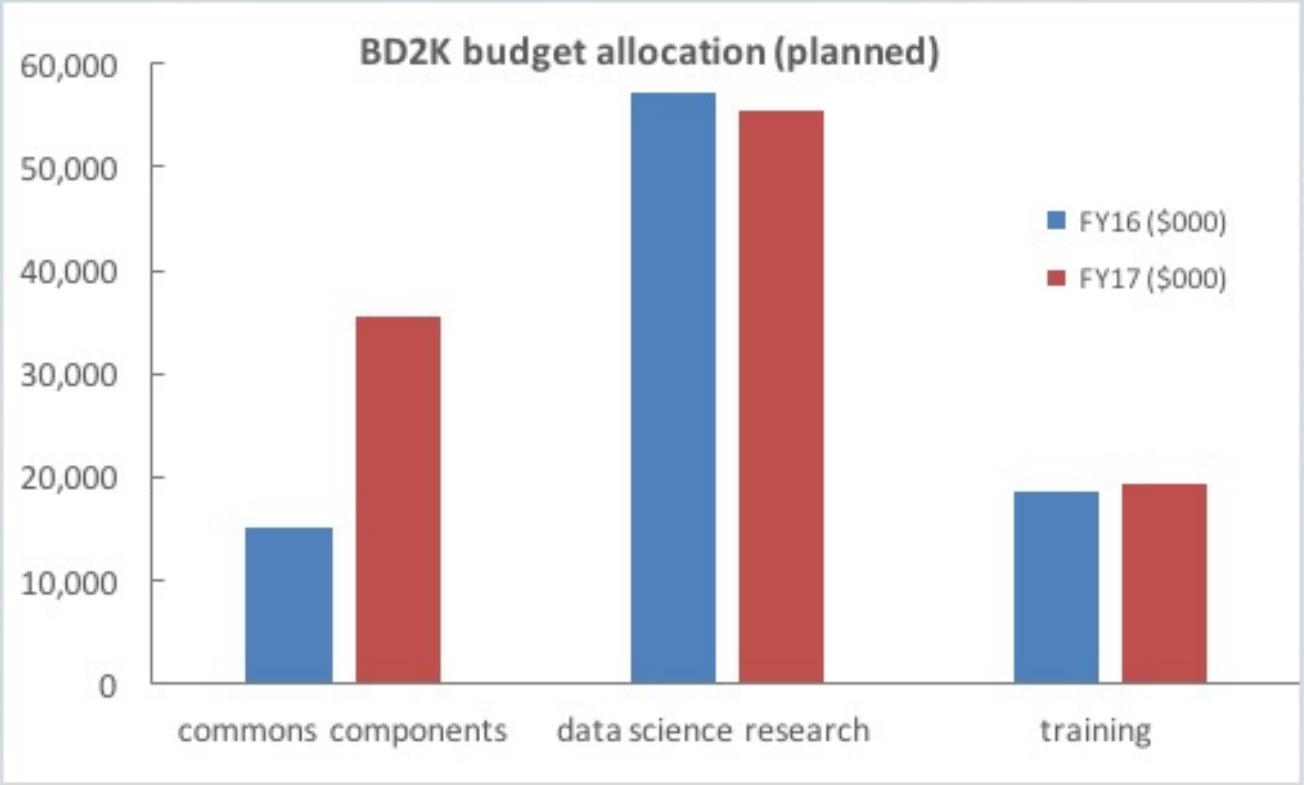
1. Commons

- Resource Indexing
- Standards
- Cloud & HPC
- Sustainability

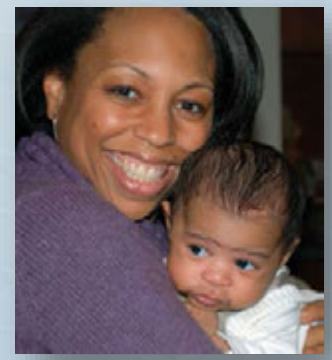
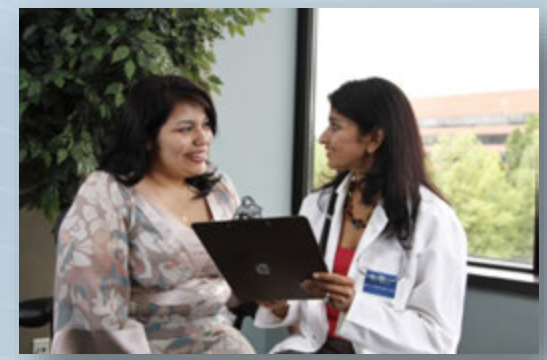
2. Data Science Research

- Centers
- Software Analysis & Methods

3. Training & Workforce Development



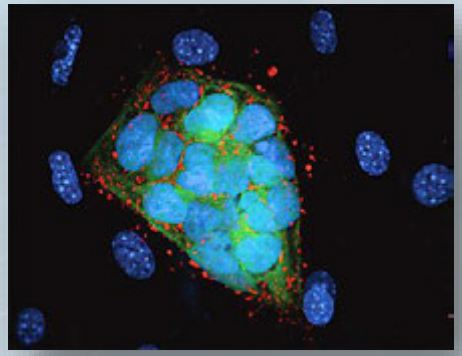
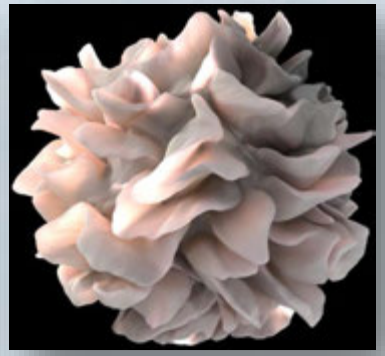
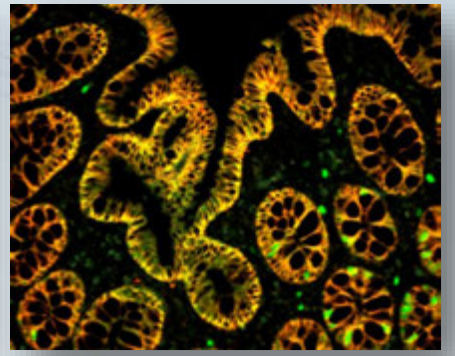
NIH Big Data to Knowledge (BD2K)

philip.bourne@nih.gov
https://datascience.nih.gov/

NIH...

Turning Discovery Into Health

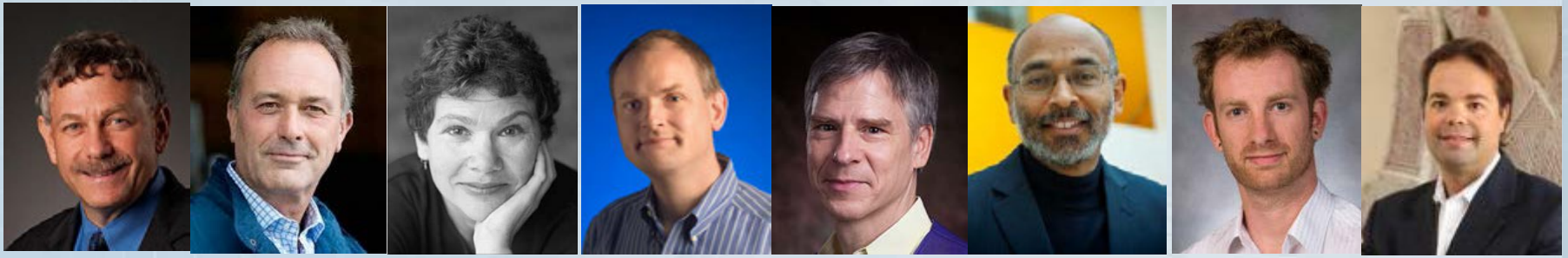





NIH Big Data to Knowledge (BD2K)

Data Science Events at NIH

- Pi Day
  - 2016 Lecture by Carlos Bustamante
  - Poster Session with Pies
  - PiCo Lightning Talks
  - Pi Day Scholars: outreach to high schools
  - Workshop: Reproducible Research

- Lecture Series: Distinguished and Frontiers in Data Science



- Data Science Courses
  - Machine learning
  - Hackathons

software carpentry

DATA CARPENTRY

NIH Big Data to Knowledge (BD2K)