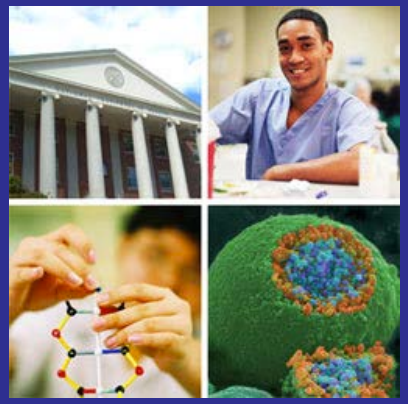
Biomedicine As A Data Driven Science

Philip E. Bourne, PhD, FACMI
Associate Director for Data Science

National Institutes of Health

National Data Integrity Conference
Colorado State University
May 7, 2015

Office of Biomedical Data Science Mission Statement



To use data science to foster an

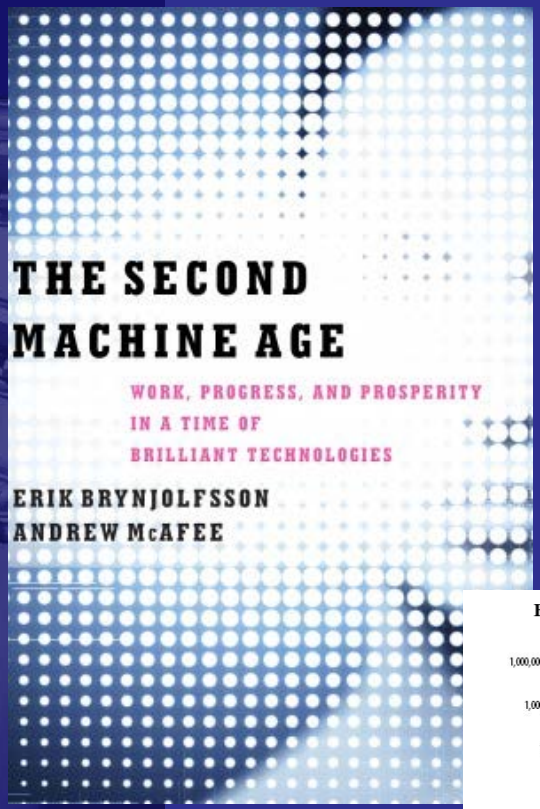
open *digital ecosystem* that will accelerate **efficient, cost-effective** biomedical research

to enhance health, lengthen life, and reduce illness and disability

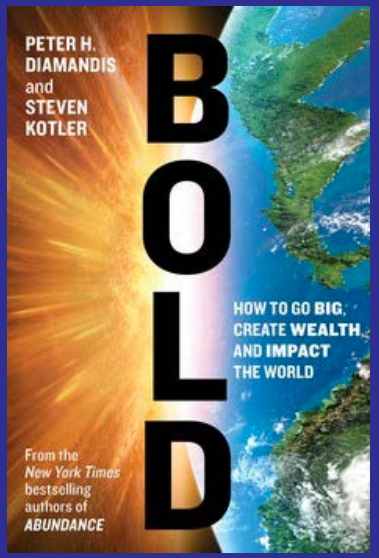
Goals expanded from recommendations in the June 2012 DIWG and BRWWG reports.

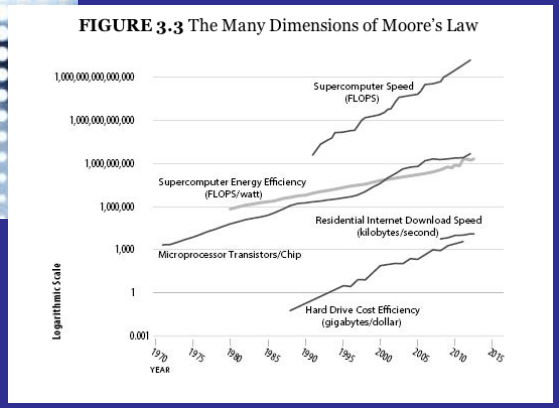
Let Me Give You 4 Examples of What Drives Us …

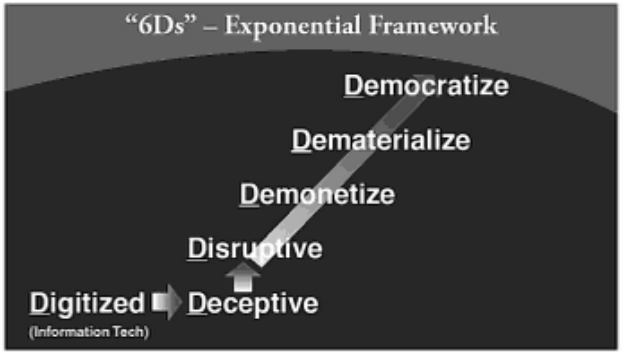
1. We are at a Point of Deception …



- Evidence:
  - Google car
  - 3D printers
  - Waze
  - Robotics
  - Sensors





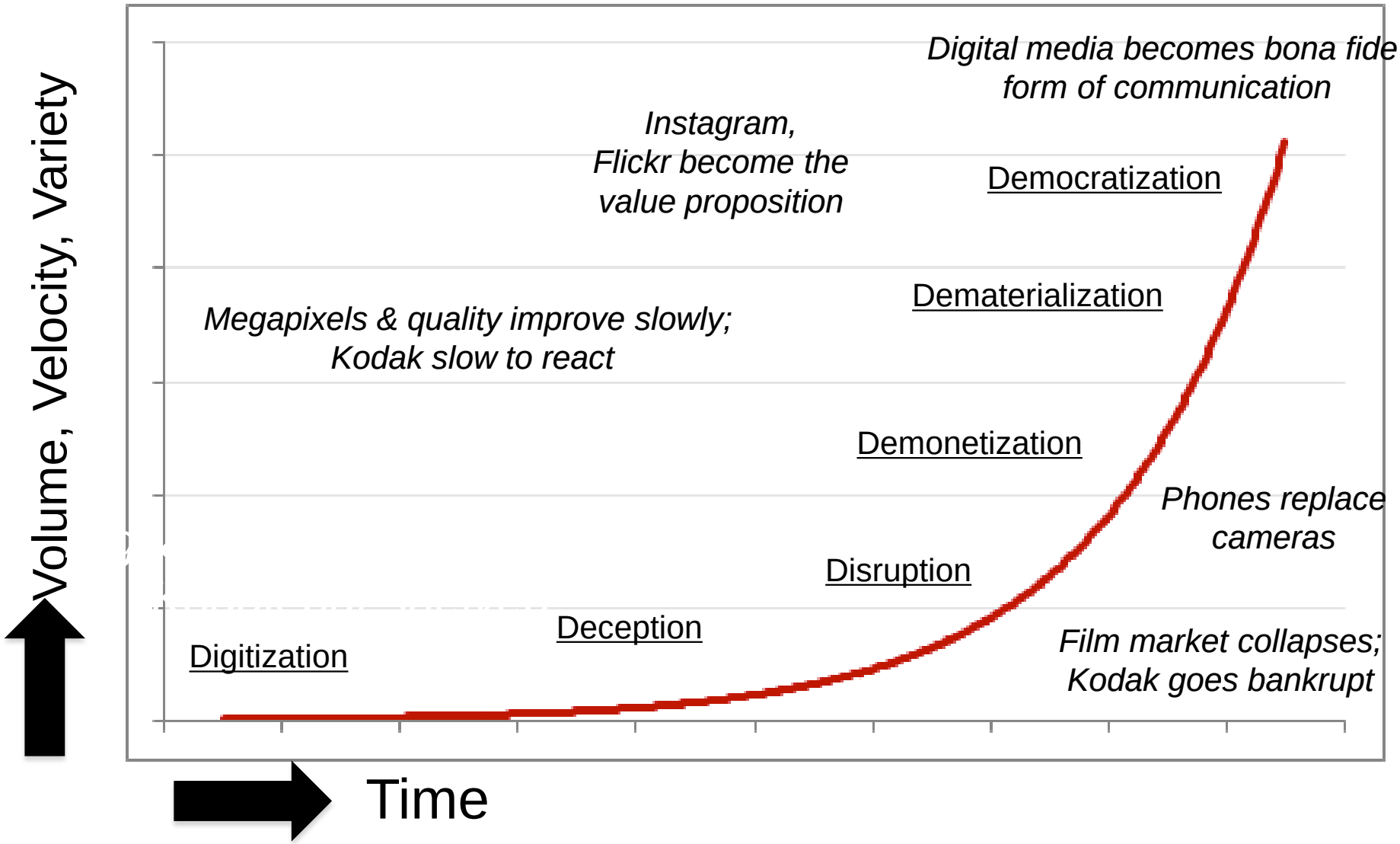


The 6 Ds of Exponentials: Digitalization, Deception, Disruption, Demonetization, Dematerialization, and Democratization

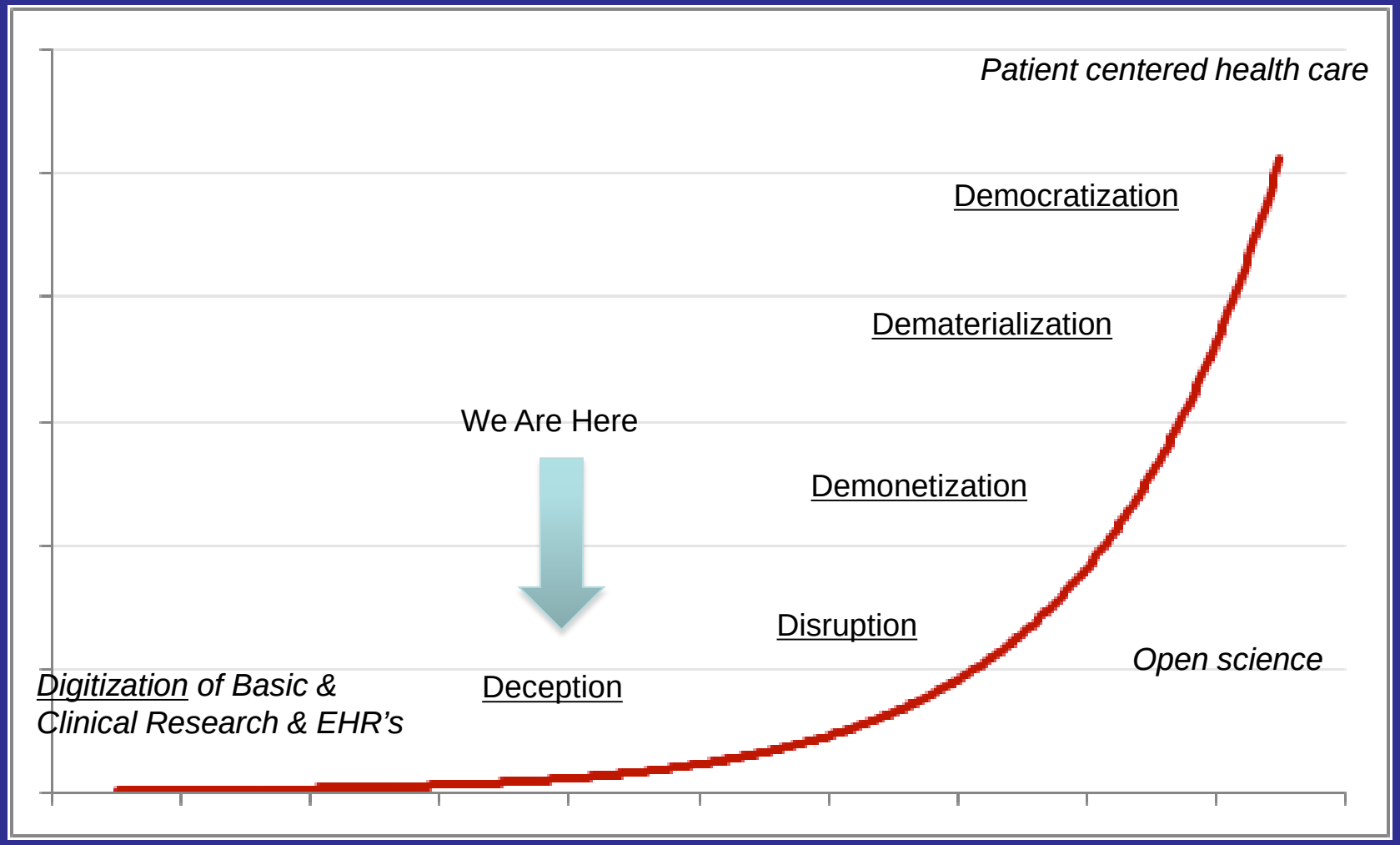
Source: Peter H. Diamandis, www.abundancehub.com

From: The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies by Erik Brynjolfsson & Andrew McAfee

Example - Photography



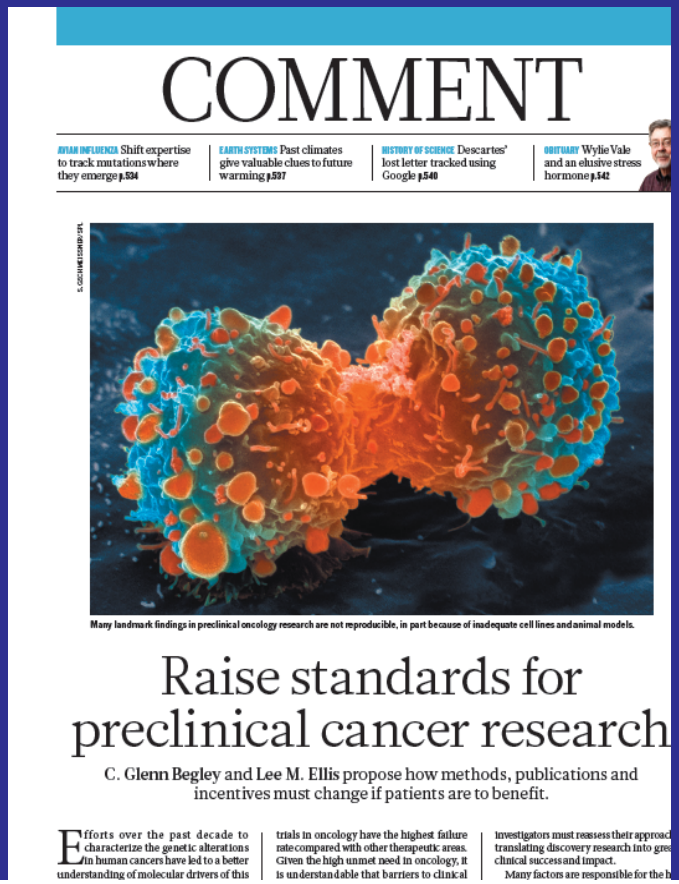
1. We Are At a Point of Deception
The 6D Exponential Framework



2. Democratization Will Follow
The Story of Meredith

http://fora
Phil_Bou

3. Disruption Can Occur



COMMENT

AVIAN INFLUENZA Shift expertise to track mutations where they emerge p.534 | EARTH SYSTEMS Past climates give valuable clues to future warming p.537 | HISTORY OF SCIENCE Descartes' lost letter tracked using Google p.540 | OBITUARY Wylie Vale and an elusive stress hormone p.542

S. GSCHMEISSNER/SPL

Many landmark findings in preclinical oncology research are not reproducible, in part because of inadequate cell lines and animal models.

Raise standards for preclinical cancer research

C. Glenn Begley and Lee M. Ellis propose how methods, publications and incentives must change if patients are to benefit.

Efforts over the past decade to characterize the genetic alterations in human cancers have led to a better understanding of molecular drivers of this

trials in oncology have the highest failure rate compared with other therapeutic areas. Given the high unmet need in oncology, it is understandable that barriers to clinical

investigators must reassess their approach translating discovery research into gre clinical success and impact.

Many factors are responsible for the h

47/53 “landmark” publications could not be replicated

[Begley, Ellis Nature, 483, 2012]

Must try harder

Too many sloppy mistakes are creeping into scientific papers. Lab heads must l at the data — and at themselves.

Error prone

Biologists must realize the pitfalls of work or massive amounts of data.

If a job is worth doing, it is worth doing twice

Researchers and funding agencies need to put a premium on ensuring that results are reproducible, argues **Jonathan F. Russell**.

The case for open computer programs

Six red flags for suspect work

C. **Glenn Begley** explains how to recognize the preclinical papers in which the data won’t stand up.

Know when your numbers are significant

[Carole Goble]

4. Demonetization, Democratization?

"And that's why we're here today. Because something called precision medicine … gives us one of the greatest opportunities for new medical breakthroughs that we have ever seen."

President Barack Obama

January 30, 2015

Precision Medicine Initiative

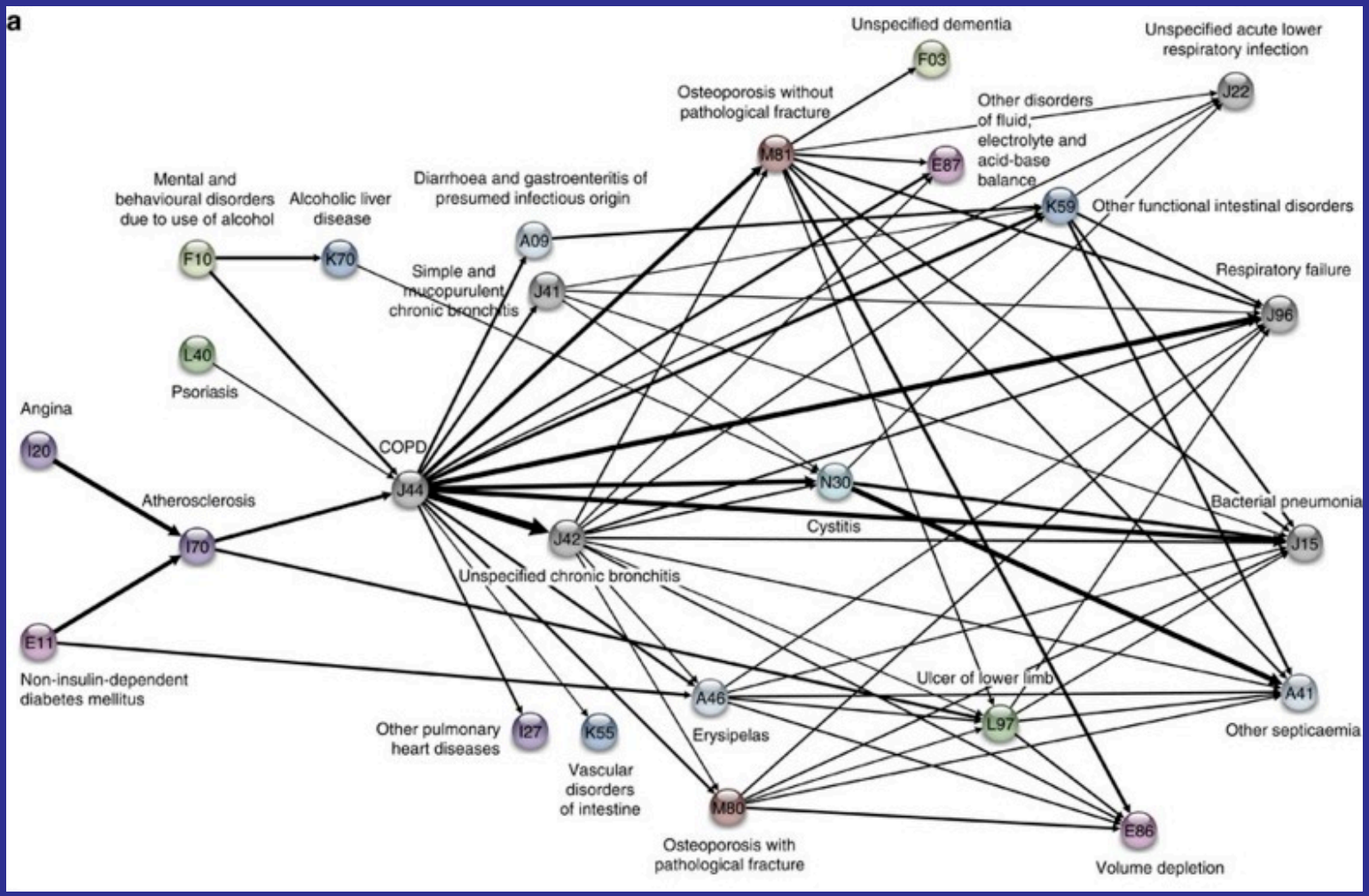
Vision: Build a broad research program to encourage creative approaches to precision medicine, test them rigorously, and, ultimately, use them to build the evidence base needed to guide clinical practice.

- Near Term: apply the tenets of precision medicine to a major health threat – cancer
- Longer Term: generate the knowledge base necessary to move precision medicine into virtually all areas of health and disease

Precision Medicine Initiative

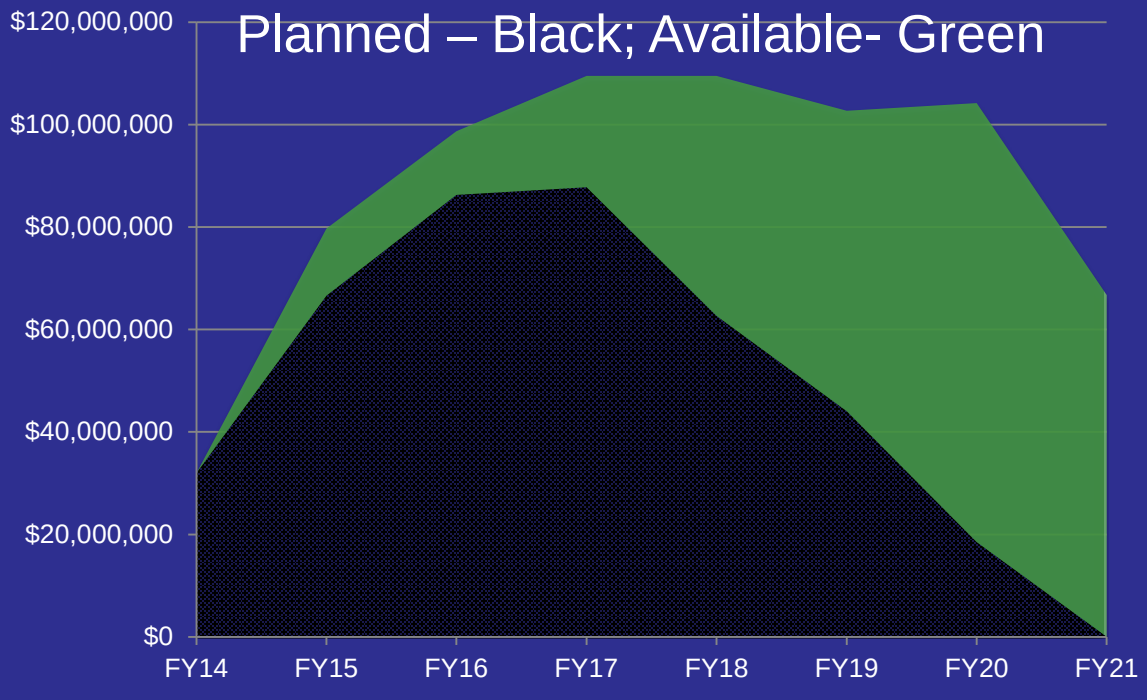
- National Research Cohort
  - >1 million U.S. volunteers
  - Numerous existing cohorts (many funded by NIH)
  - New volunteers
- Participants will be centrally involved in design and implementation of the cohort
- They will be able to share genomic data, lifestyle information, biological samples – all linked to their electronic health records

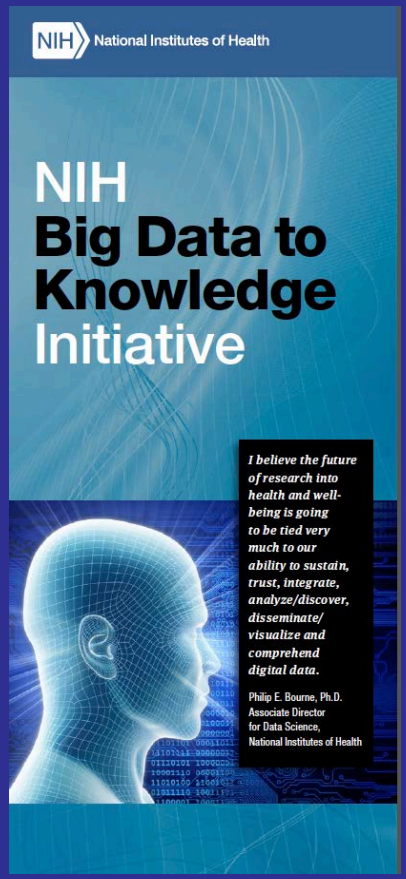
An Example of That Promise: Comorbidity Network for 6.2M Danes Over 14.9 Years



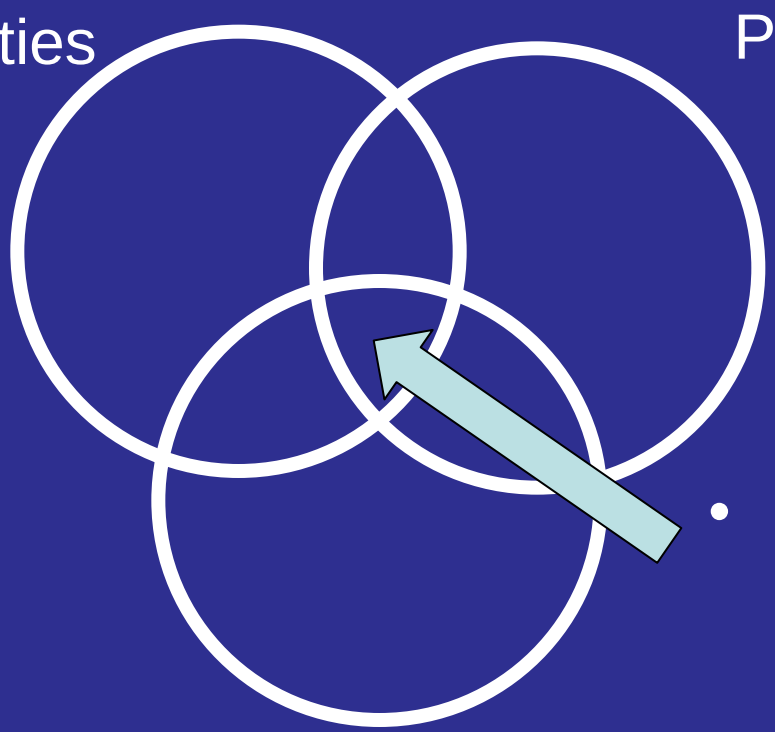
Jensen et al 2014 Nat Comm 5:4022

The BD2K Program is Central to the Mission



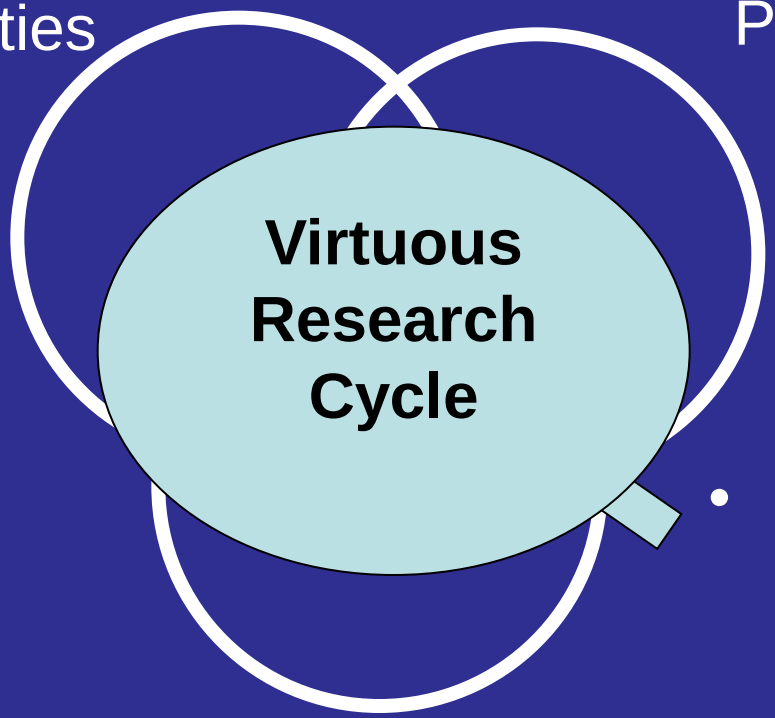


Elements of The Digital Enterprise



- Intersection:
  - Sustainability
  - Efficiency
  - Collaboration
  - Training

Elements of The Digital Enterprise

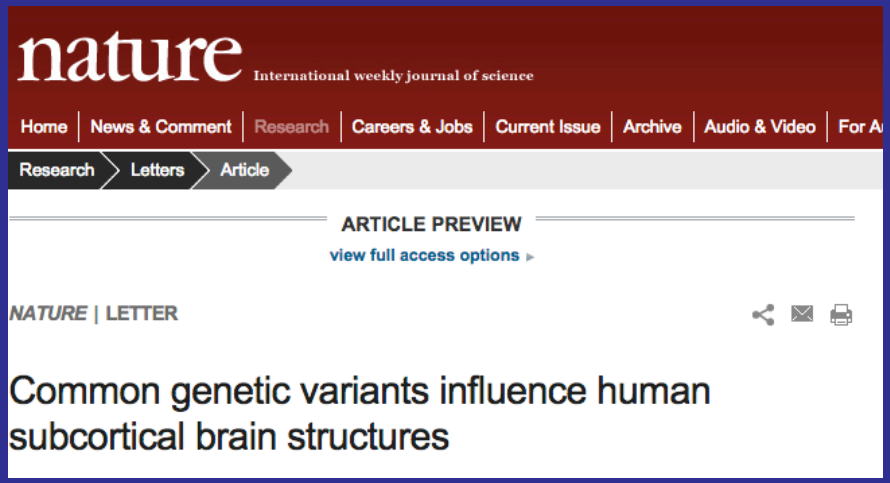


- Intersection:
  - Sustainability
  - Efficiency
  - Collaboration
  - Training

Consider an example…

NIH NIH Big Data to Knowledge (BD2K)

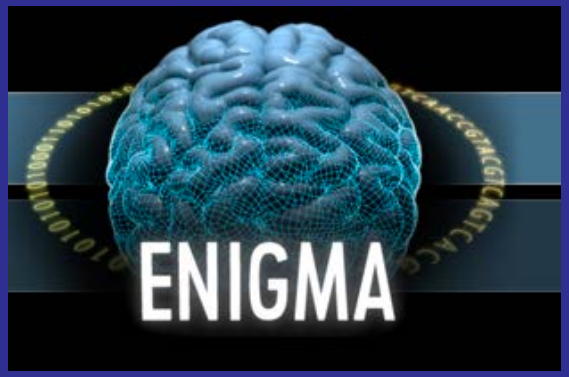
ENIGMA



- Big Data: The study involved MRI images & GWAS data from over 30,000 people
- Collaboration: Data came from many different sights affiliated with the ENIGMA consortium
- Methods: To homogenize data from different sites, the group designed standardized protocols for image analysis, quality assessment, genetic imputation, and association

- Found five novel genetic variants
- Results provided insight into the variability of brain development, and may be applied to study of neuropsychiatric dysfunction

NIH



- Community – Enigma, BD2K
- Policy
  - Improved consent methods
  - Cloud accessibility for human subjects data
  - Trusted partners
  - Data sharing
- Infrastructure
  - Standards, compute resources, software

Communities: Thus Far

- Visioning workshop convened 9/3/14
- Launched BD2K ($32M)
  - 12 Centers of data excellence
  - Data Discovery Index Coordination Consortium (DDICC)
  - Training awards
- First successful consortia meeting 11/3-4
- Workshops to inform future funding
  - Software indexing and discoverability
  - Gaming

Communities: 2015 Activities

- New FOAs with outreach to new communities – math, stats, comp science etc.
- Work with e.g GA4GH, RDA, FORCE11, NDS ….
- IDEAS lab with NSF
- Competition with international funders
- Software carpentry, hackathons, Pi Day

Communities: Questions?

- Societies of the modern age?
- How to enable these groups?
- How to marry the funding of individuals with the funding of communities?

Policies: Now & Forthcoming

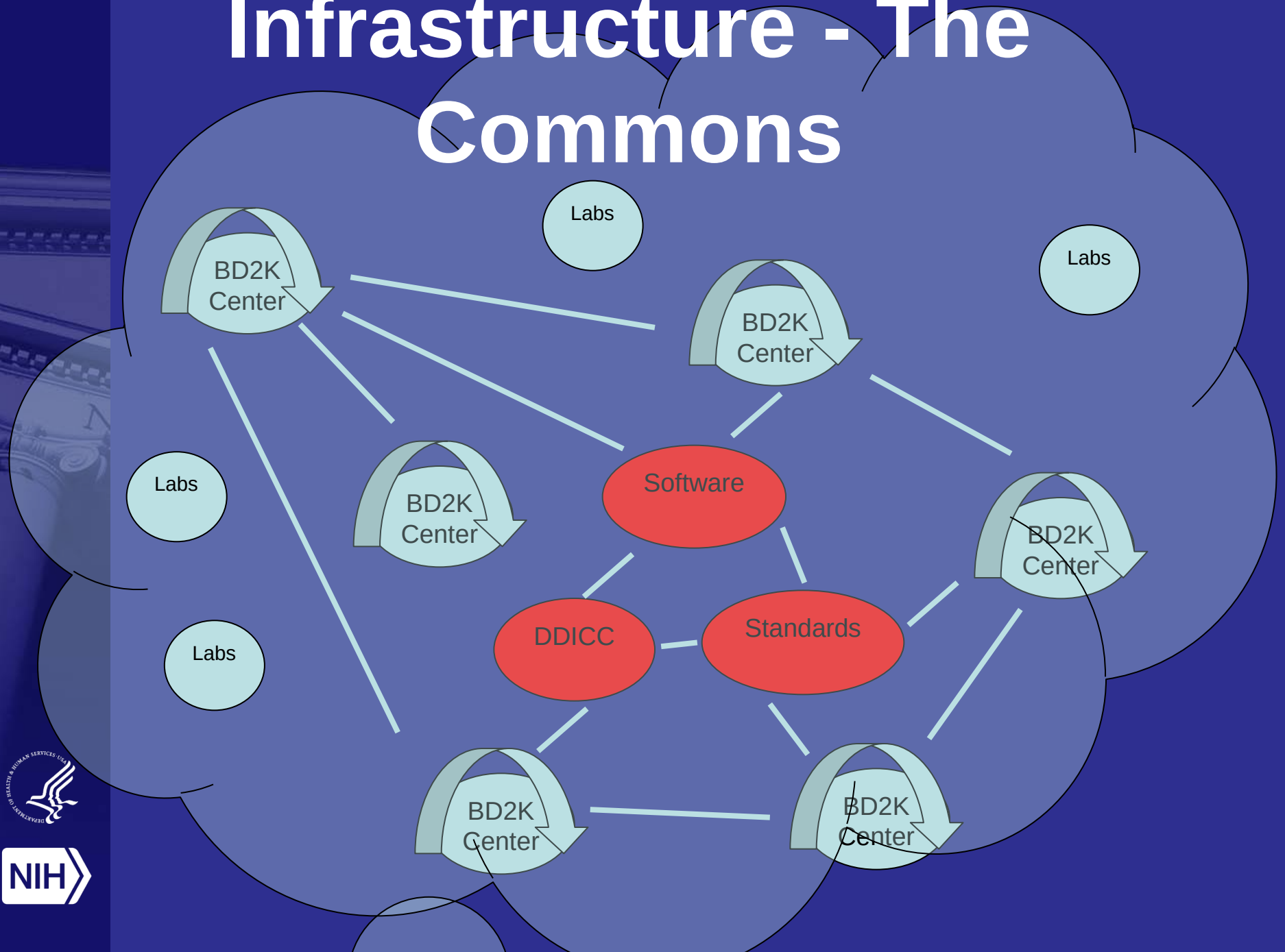
- Data Sharing
  - Genomic data sharing announced
  - Data sharing plans on all research awards
  - Data sharing plan enforcement
    - Machine readable plan
    - Repository requirements to include grant numbers

http://www.nih.gov/news/health/aug2014/od-27.htm

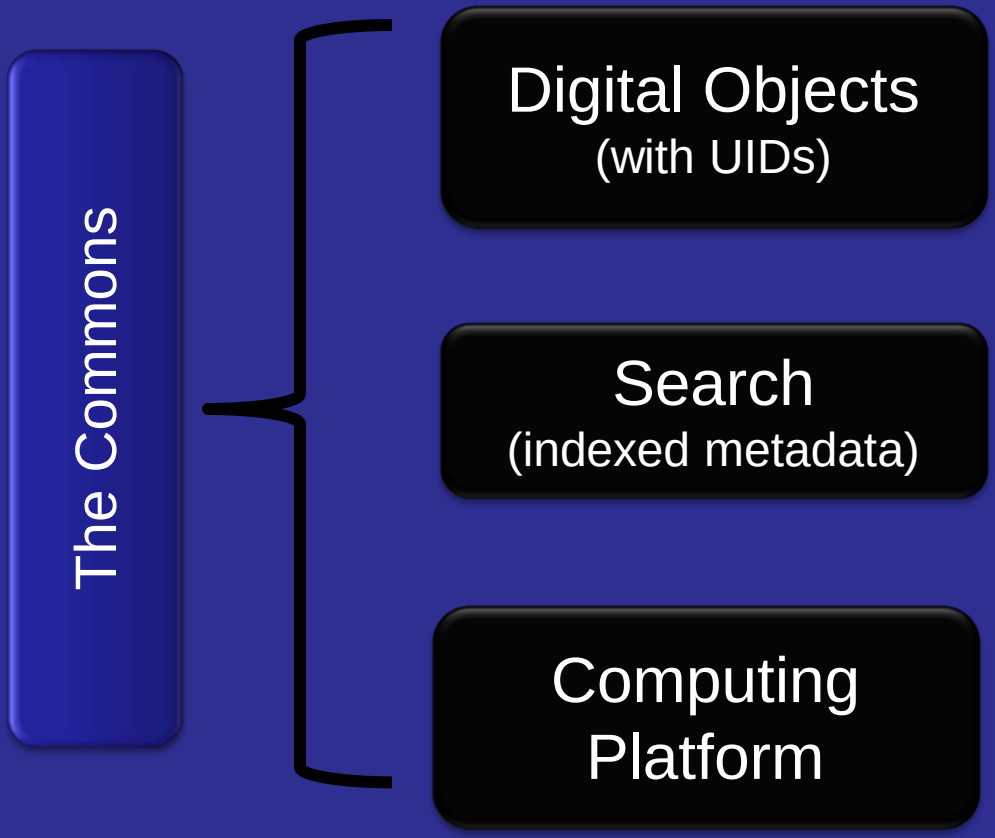
Policies - Forthcoming

- Data Citation
  - Goal: legitimize data as a form of scholarship
  - Process:
    - Machine readable standard for data citation (done)
    - Endorsement of data citation for inclusion in NIH bib sketch, grants, reports, etc.
    - Example formats for human readable data citations
    - Slowly work into NLM/NCBI workflow
- dbGaP in the cloud (done!)

Infrastructure - The Commons

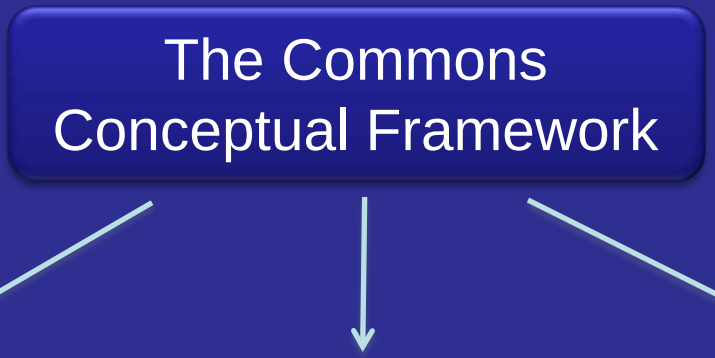


The Commons



Vivien Bonazzi
George Komatsoulis

The Commons: Compute Platforms



Public Cloud Platforms

- Google, AWS (Amazon)
- Microsoft (Azure), IBM, other?

Super Computing (HPC) Platforms

- Traditionally low access by NIH

Other Platforms ?

- In house compute solutions
- Private clouds, HPC
  - Pharma
  - The Broad
  - Bionimbus

Commons – Simple Implementation Stack

APIs

App Store

Biomedical Data Software

Biomedical DATA

Big Data Software

Scalable Hardware

Direct access to data, software

User friendly Interface

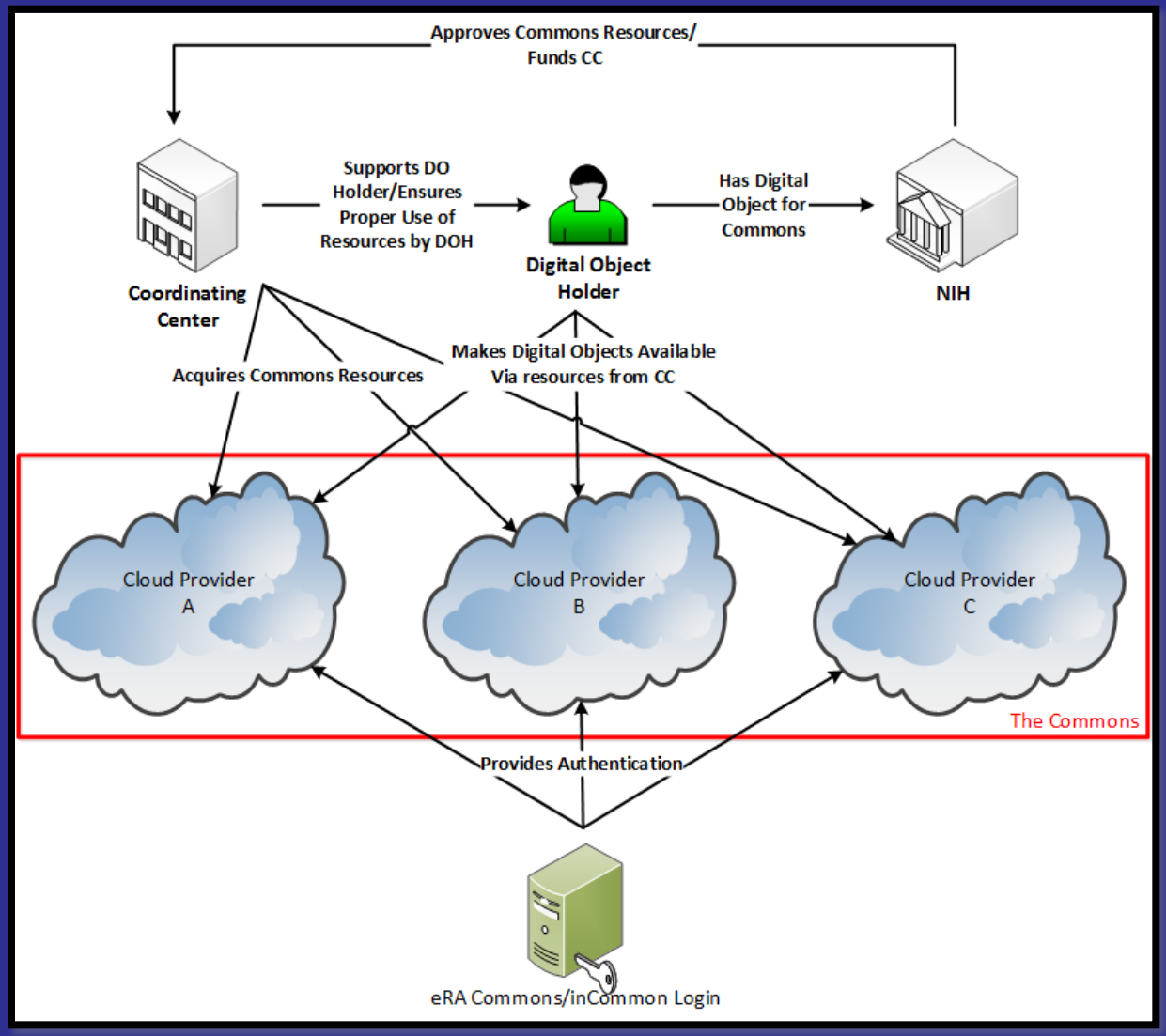
Genome Assembly

1000 genomes
HMP
SRA

Hadoop
search, clustering, indexing, NLP

Compute cores

The Commons:
Business Model

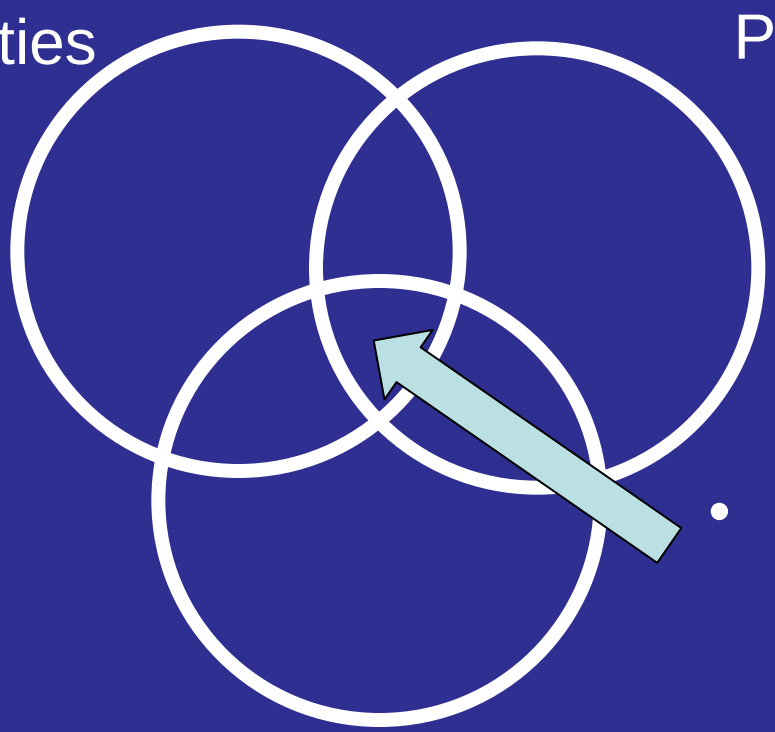


[George Komatsoulis]

Infrastructure: Standards

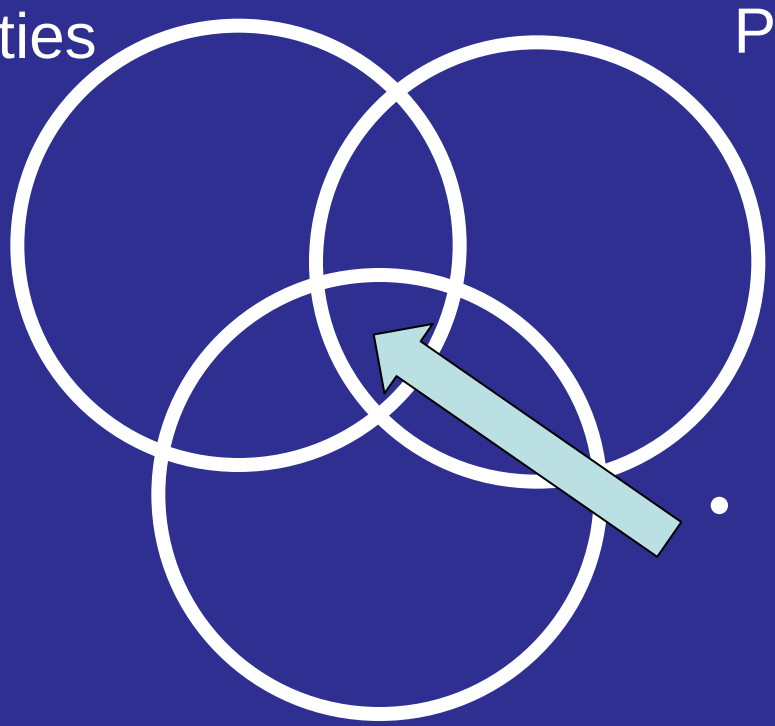
- 2013 Workshop on Frameworks for Community-Based Standards
- August 2014 Input on Information Resources for Data-Related Standards Widely Used in Biomedical Science – 30 responses
- Feb 2015 Workshop Community-based Data and Metadata Standards
- Internal CDE Registry project

Elements of The Digital Enterprise



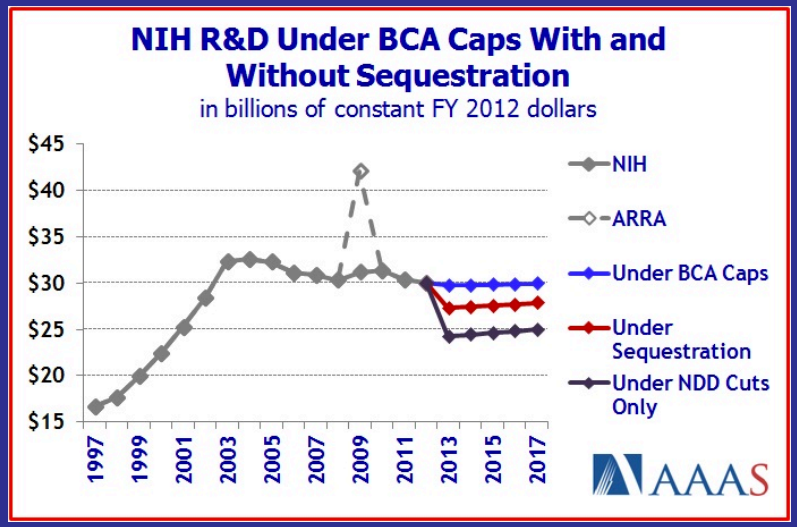
- Intersection:
  - Sustainability
  - Efficiency
  - Collaboration
  - Training

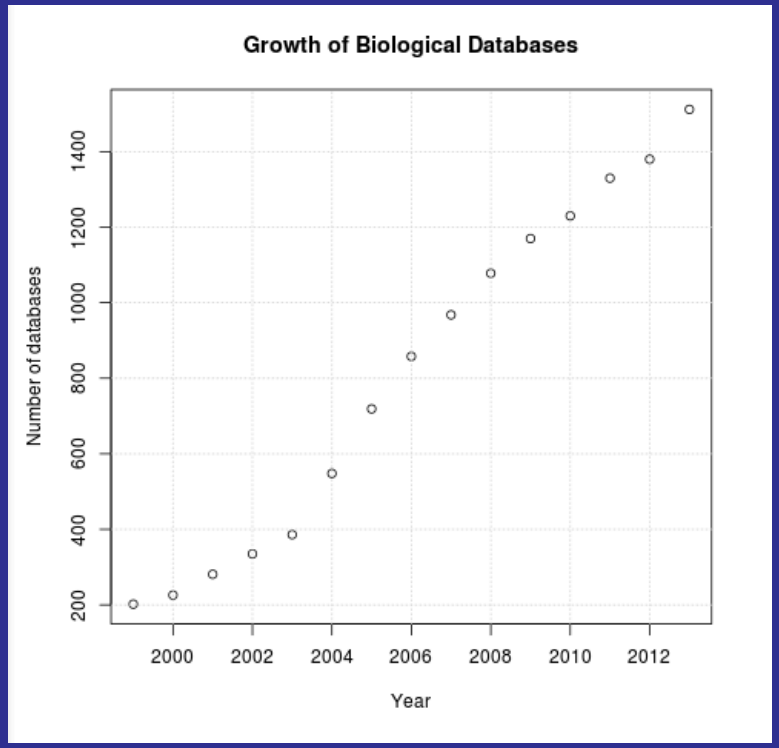
Elements of The Digital Enterprise



- Intersection:
  - Sustainability
  - Efficiency
  - Collaboration
  - Training

Sustainability 101





Source Michael Bell http://homepages.cs.ncl.ac.uk/m.j.bell1/blog/?p=830

Workforce Training

Goal: To strengthen the ability of a diverse biomedical workforce to develop and benefit from data science

Strengthening a diverse biomedical workforce to utilize data science

BD2K funding of Short Courses and Open Educational Resources

Building a diverse workforce in biomedical data science

BD2K Training programs and Individual Career Awards

Discovery of Educational Resources

BD2K Training Coordination Center

Fostering Collaborations

BD2K Training Coordination Center, NSF/NIH IDEAs Lab

Expanding NIH Data Science Workforce Development Center

Local courses, e.g. Software Carpentry

I not only use all the brains I have, but all I can borrow.

– Woodrow Wilson

Associate Di

Data Science

Scientific Data Council

External Advisory Board

Programmatic Theme

Ed on

Co ion

Deliverable

Exa

The Biomedical Research Digital Enterprise




NIH…

philip.bourne@nih.gov

Turning Discovery Into Health