Data Science at NIH: Opportunities and Challenges

Philip E. Bourne, PhD, FACMI

Associate Director for Data Science

National Institutes of Health

Council of Councils

September 1, 2015

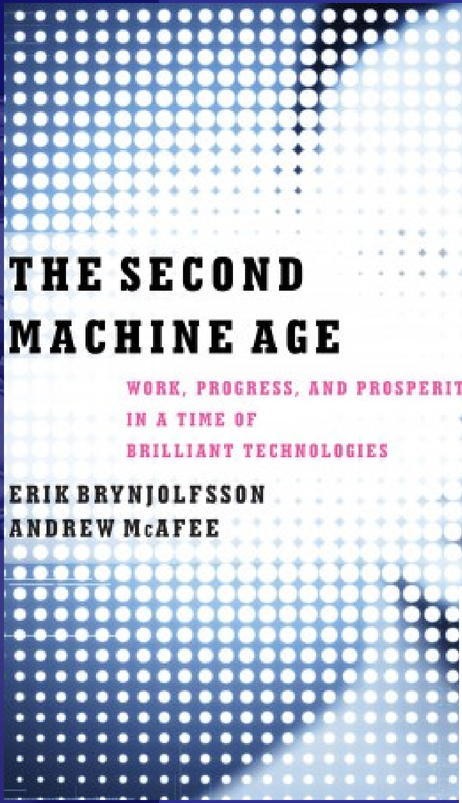
Agenda

- Drivers of change
- Scientific motivators
- The NIH response
  - The Office of Data Science
  - The Big Data to Knowledge (BD2K) initiative
- The 3-legged stool strategy
  - Infrastructure
  - Policies
  - Communities
- Diversity
- Your thoughts here?

Agenda

- Drivers of change
- Scientific motivators
- The NIH response
  - The Office of Data Science
  - The Big Data to Knowledge (BD2K) initiative
- The 3-legged stool strategy
  - Infrastructure
  - Policies
  - Communities
- Diversity
- Your thoughts here?

We are at a Point of Deception …


THE SECOND MACHINE AGE
WORK, PROGRESS, AND PROSPERITY IN A TIME OF BRILLIANT TECHNOLOGIES
ERIK BRYNJOLFSSON
ANDREW McAFEE

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

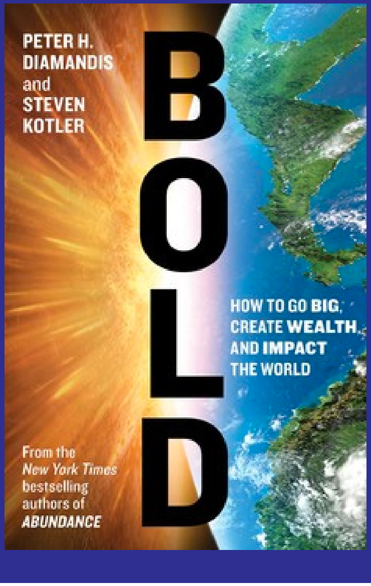
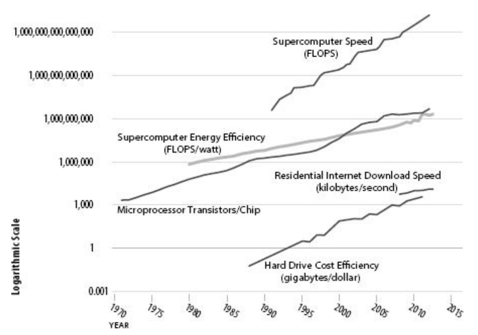
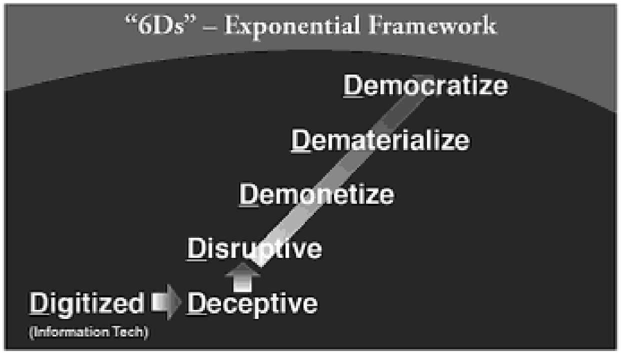

PETER H. DIAMANDIS and STEVEN KOTLER
BOLD
HOW TO GO BIG, CREATE WEALTH, AND IMPACT THE WORLD
From the New York Times bestselling authors of ABUNDANCE

FIGURE 3.3 The Many Dimensions of Moore's Law



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



"6Ds" – Exponential Framework

Democratize

Dematerialize

Demonetize

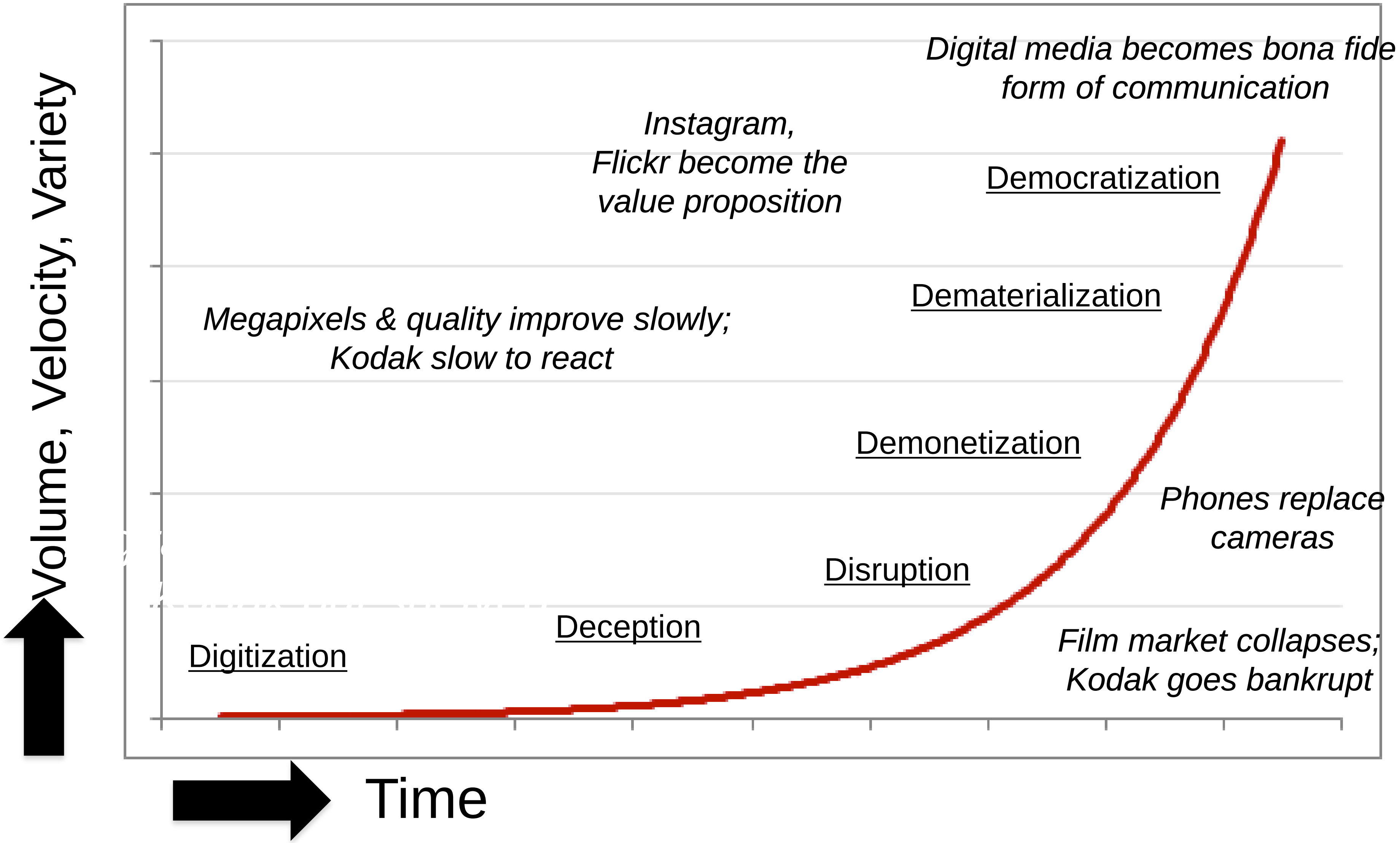
Disruptive

Digitized (Information Tech) Deceptive

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

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

Example - Photography



Digital media becomes bona fide form of communication

Instagram, Flickr become the value proposition

Democratization

Dematerialization

Megapixels & quality improve slowly; Kodak slow to react

Demonetization

Phones replace cameras

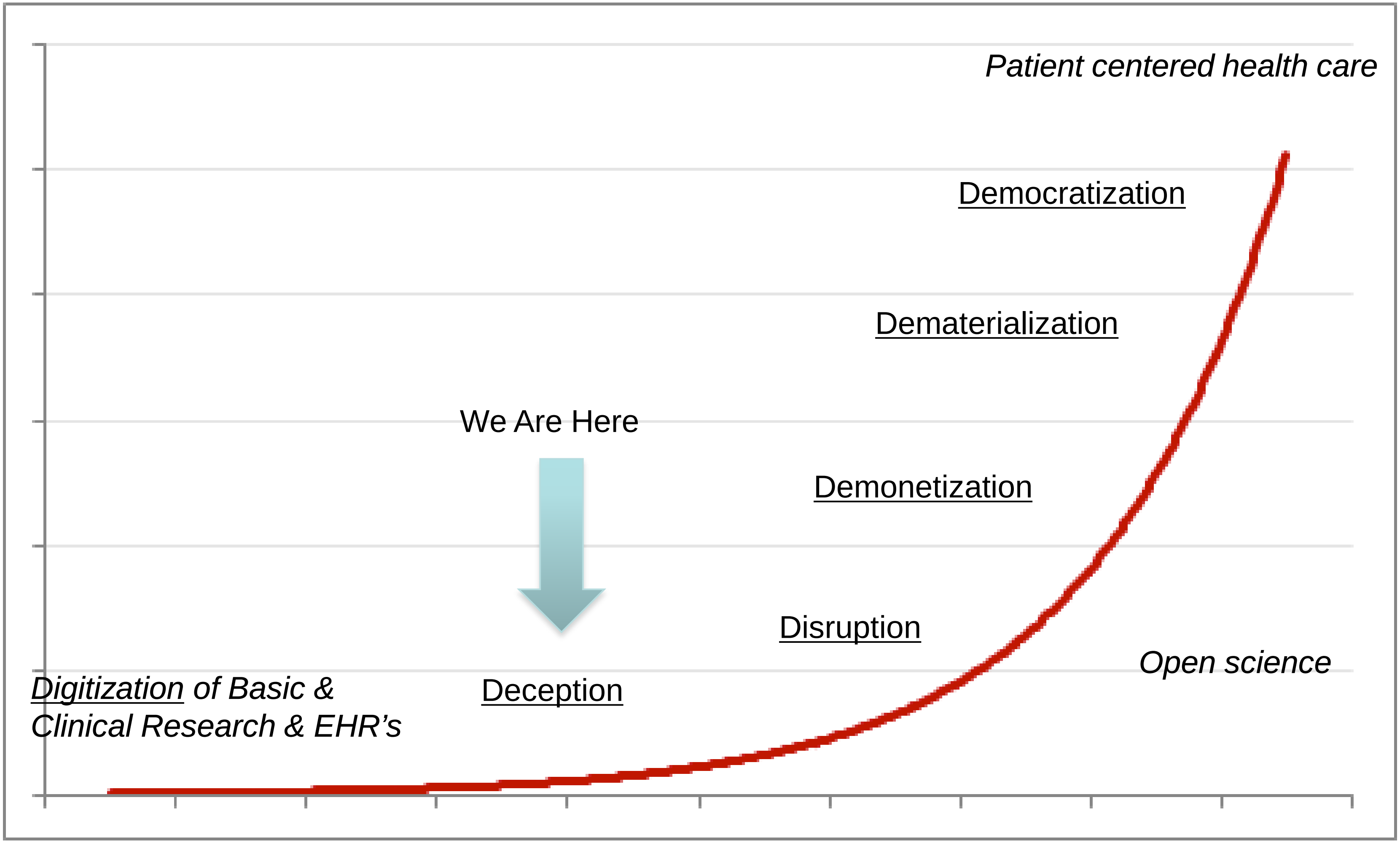
Disruption

Deception

Digitization

Film market collapses; Kodak goes bankrupt

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



Patient centered health care

Democratization

Dematerialization

We Are Here

Demonetization

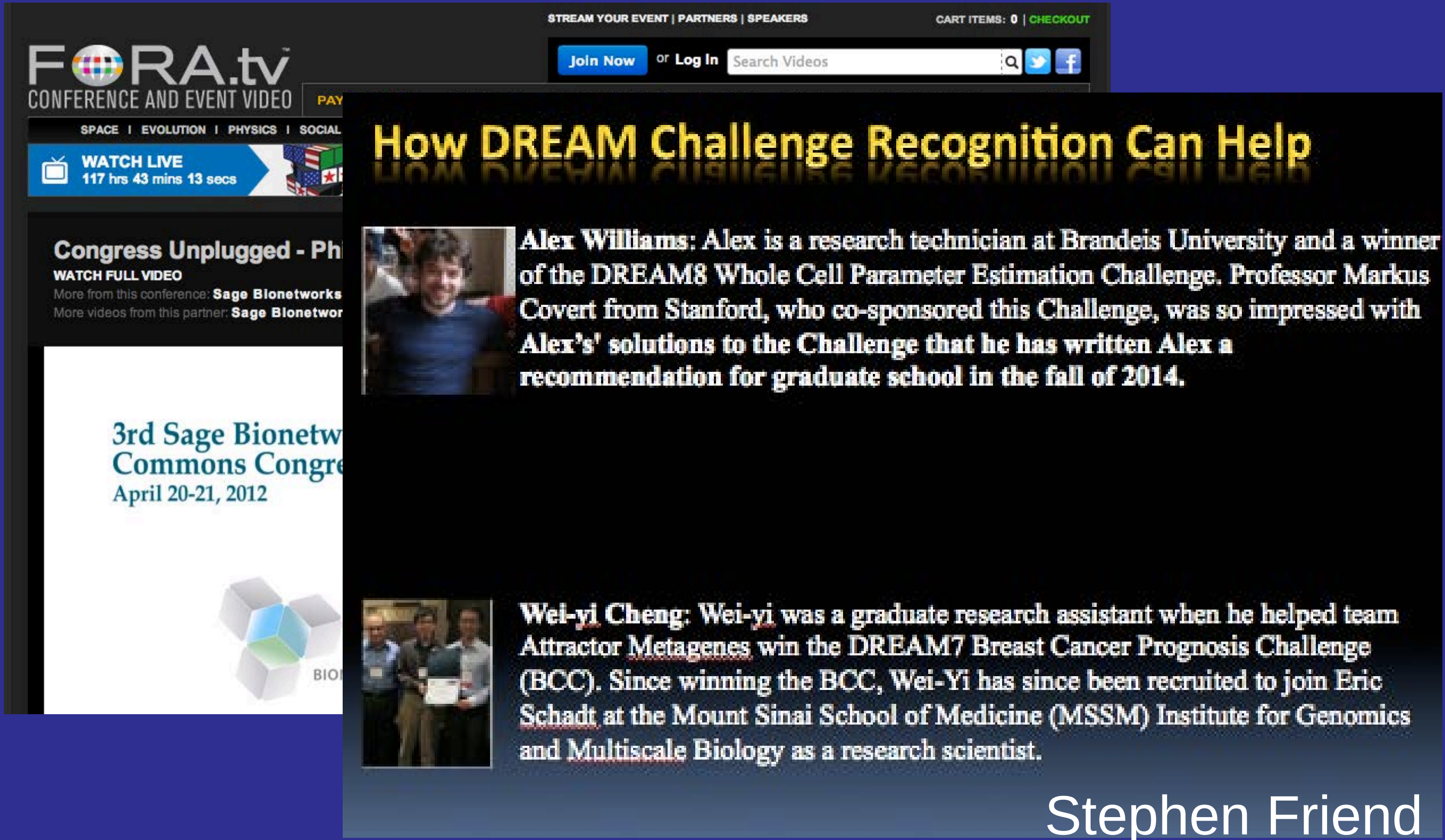
Disruption

Open science

Digitization of Basic & Clinical Research & EHR's

Deception

Another Driver of Change



STREAM YOUR EVENT | PARTNERS | SPEAKERS
CART ITEMS: 0 | CHECKOUT
FORA.tv
CONFERENCE AND EVENT VIDEO
PAY
Join Now
or Log In
Search Videos
SPACE | EVOLUTION | PHYSICS | SOCIAL
WATCH LIVE
117 hrs 43 mins 13 secs

Congress Unplugged - Phi

WATCH FULL VIDEO

More from this conference: Sage Bionetworks

More videos from this partner: Sage Bionetwor

3rd Sage Bionetw
Commons Congre
April 20-21, 2012

How DREAM Challenge Recognition Can Help

Alex Williams: Alex is a research technician at Brandeis University and a winner of the DREAM8 Whole Cell Parameter Estimation Challenge. Professor Markus Covert from Stanford, who co-sponsored this Challenge, was so impressed with Alex's' solutions to the Challenge that he has written Alex a recommendation for graduate school in the fall of 2014.

Wei-yi Cheng: Wei-yi was a graduate research assistant when he helped team Attractor Metagenes win the DREAM7 Breast Cancer Prognosis Challenge (BCC). Since winning the BCC, Wei-Yi has since been recruited to join Eric Schadt at the Mount Sinai School of Medicine (MSSM) Institute for Genomics and Multiscale Biology as a research scientist.

Stephen Friend

http://fora.tv/2012/04/20 Congress_Unplugged_Phil_Bourne

Let's Make Gender Diversity in Data Science a Priority Right from the Start
2015 Berman & Bourne *PLOS Biology* 13(7): e1002206

Agenda

- Drivers of change
- Scientific motivators
- The NIH response
  - The Office of Data Science
  - The Big Data to Knowledge (BD2K) initiative
- The 3-legged stool strategy
  - Infrastructure
  - Policies
  - Communities
- Some future activities
- Your thoughts here?

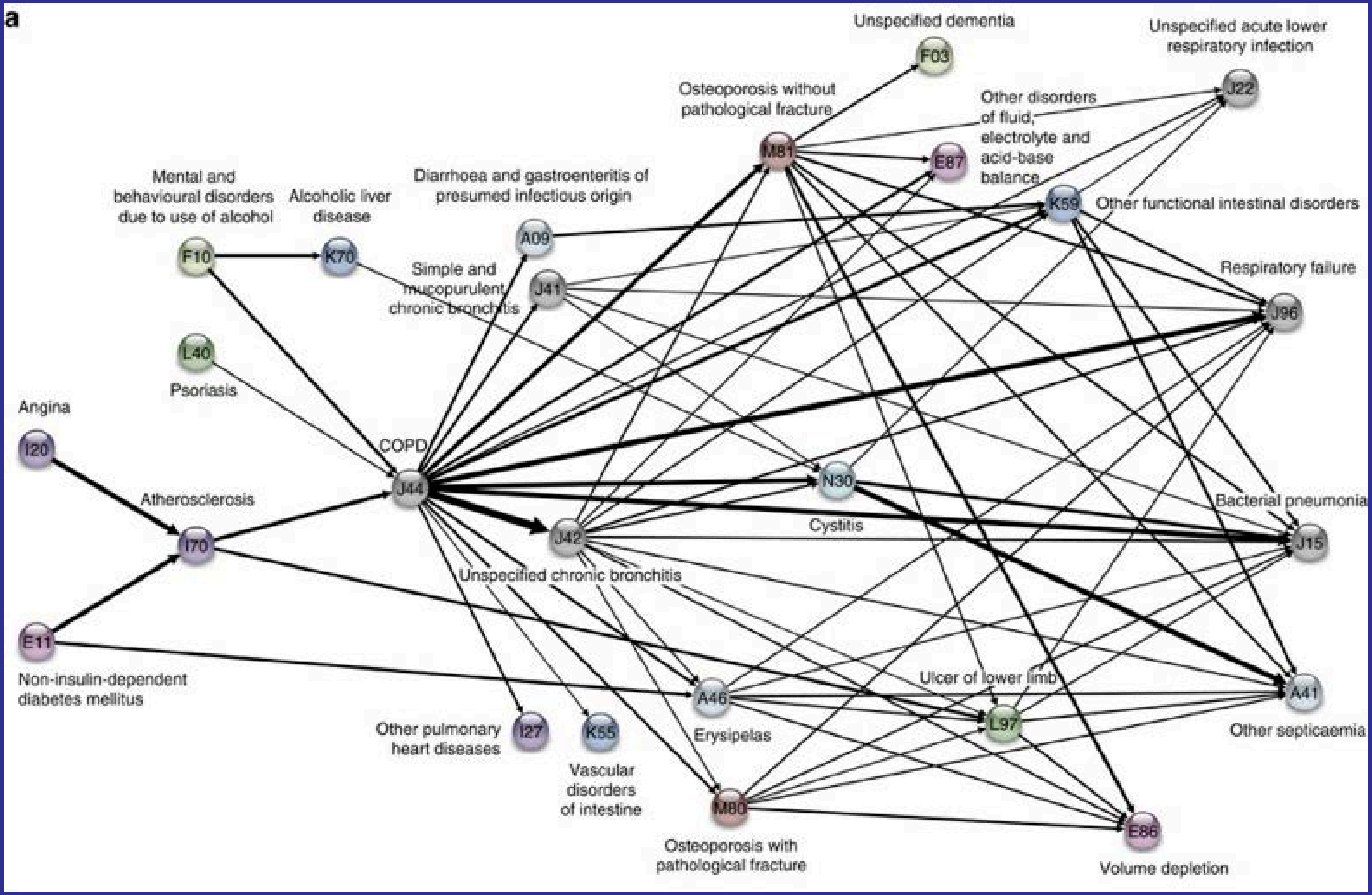
“And that’s why we’re here today. Because something called precision medicine … gives us one of the greatest opportunities for new medical breakthroughs that we have ever seen.”

President Barack Obama
January 30, 2015

Precision Medicine Initiative

- 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


a
Unspecified dementia
F03
Unspecified acute lower respiratory infection
J22
Osteoporosis without pathological fracture
M81
Other disorders of fluid, electrolyte and acid-base balance
E87
K59
Other functional intestinal disorders
Mental and behavioural disorders due to use of alcohol
F10
Alcoholic liver disease
K70
Diarrhoea and gastroenteritis of presumed infectious origin
A09
Simple and mucopurulent chronic bronchitis
J41
Respiratory failure
J96
L40
Psoriasis
Angina
I20
COPD
J44
N30
Cystitis
Bacterial pneumonia
J15
Atherosclerosis
I70
J42
Unspecified chronic bronchitis
E11
Non-insulin-dependent diabetes mellitus
A46
Erysipelas
Ulcer of lower limb
L97
A41
Other septicaemia
Other pulmonary heart diseases
I27
K55
Vascular disorders of intestine
M80
Osteoporosis with pathological fracture
E86
Volume depletion

Jensen et al 2014 Nat Comm 5:4022

Agenda

- Drivers of change
- Scientific motivators
- The NIH response
  - The Office of Data Science
  - The Big Data to Knowledge (BD2K) initiative
- The 3-legged stool strategy
  - Infrastructure
  - Policies
  - Communities
- Some future activities
- Your thoughts here?

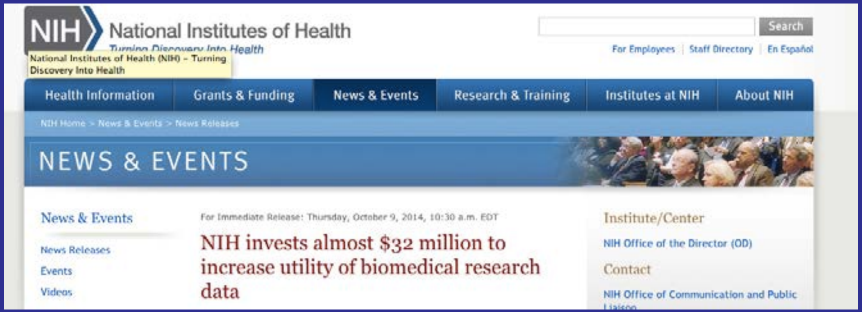
NIH Office of Data Science Mission Statement

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

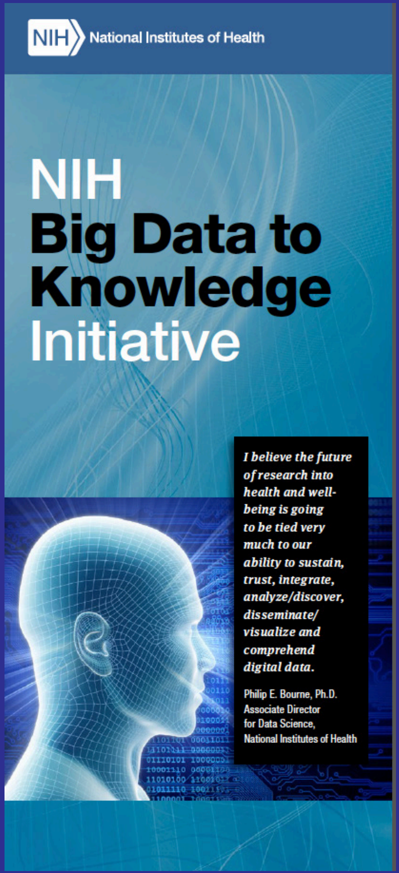
Goals expanded from recommendations in the June 2012 DIWG and BRWWG reports.

The BD2K Program is Central to the Mission


NIH National Institutes of Health
Turning Discovery Into Health
National Institutes of Health (NIH) – Turning Discovery Into Health
Search
For Employees | Staff Directory | En Español
Health Information
Grants & Funding
News & Events
Research & Training
Institutes at NIH
About NIH
NIH Home > News & Events > News Releases
NEWS & EVENTS
News & Events
News Releases
Events
Videos
Institute/Center
NIH Office of the Director (OD)
Contact
NIH Office of Communication and Public Liaison

For Immediate Release: Thursday, October 9, 2014, 10:30 a.m. EDT

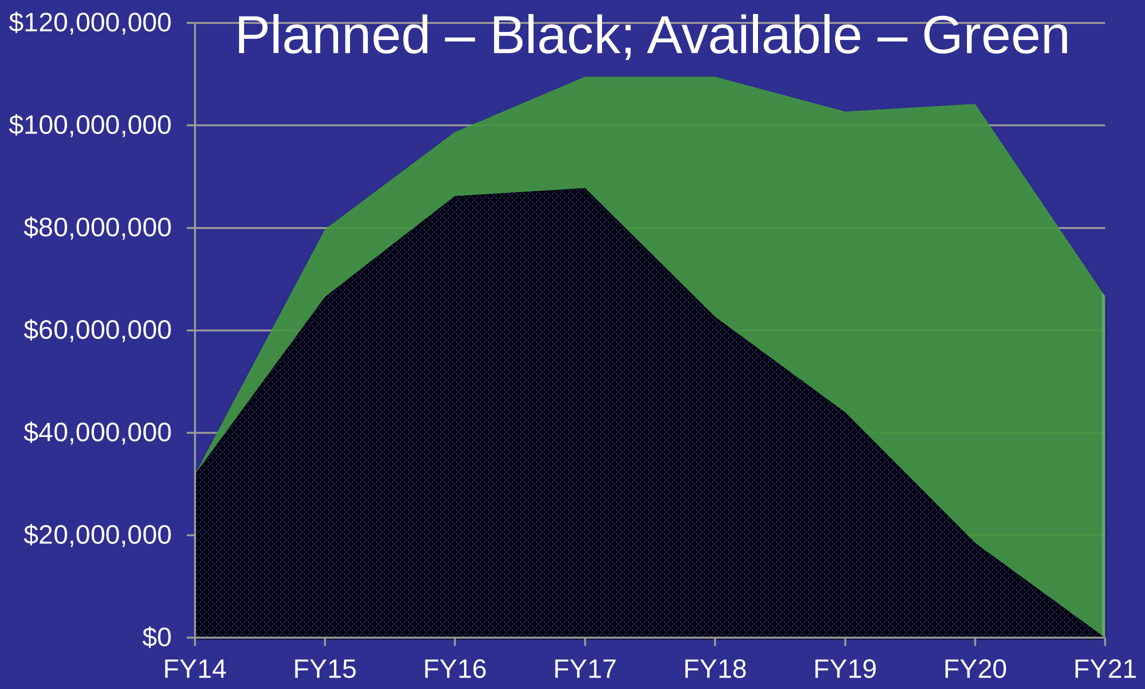
NIH invests almost $32 million to increase utility of biomedical research data



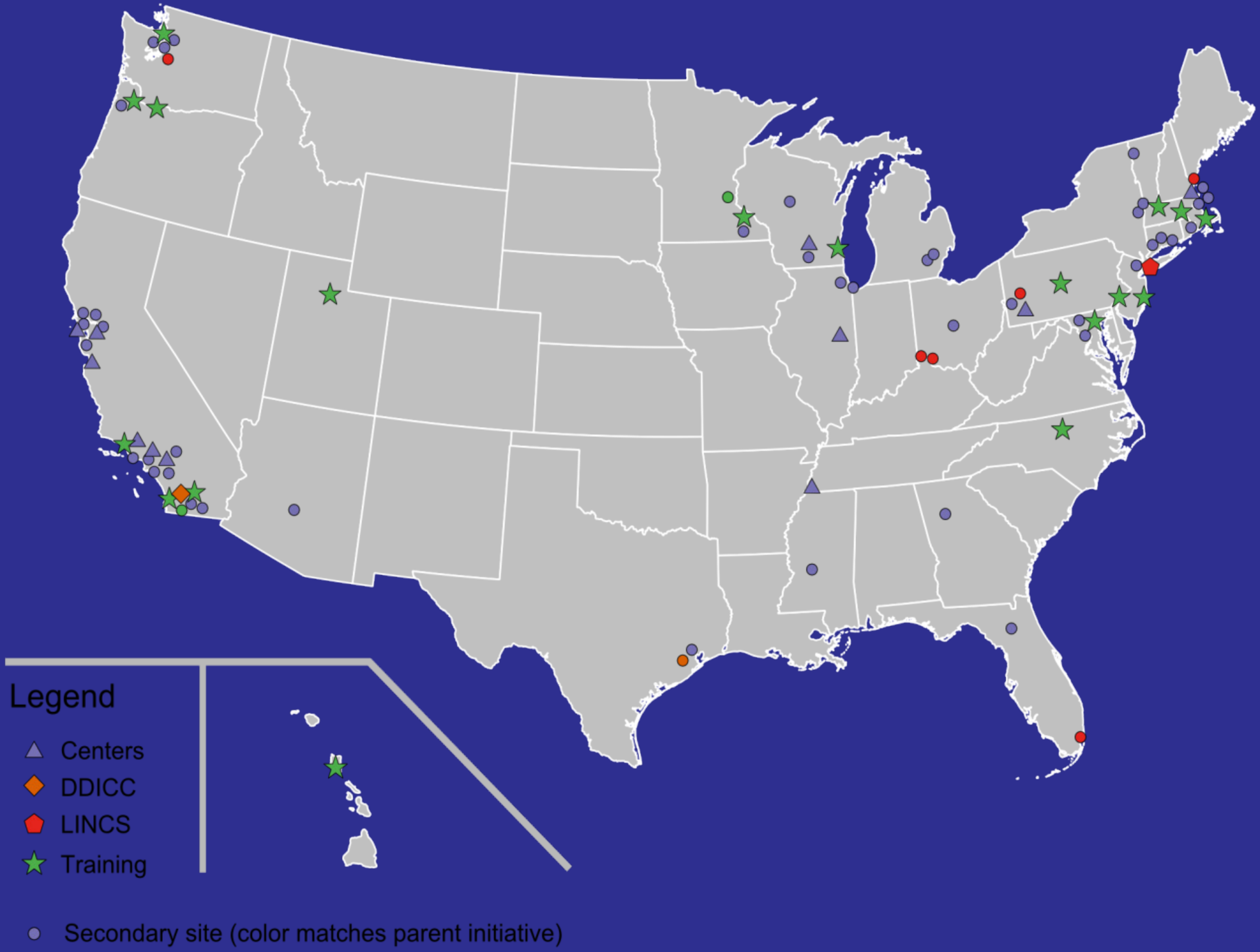
NIH
Big Data to Knowledge Initiative

I believe the future of research into health and well-being is going to be tied very much to our ability to sustain, trust, integrate, analyze/discover, disseminate/visualize and comprehend digital data.

Philip E. Bourne, Ph.D.
Associate Director for Data Science,
National Institutes of Health



BD2K Awards


Legend
▲ Centers
◆ DDICC
⬟ LINCS
★ Training
● Secondary site (color matches parent initiative)

Example: BD2K Center

Working Across Strategic Areas

Strategic Areas

Sustainability

Workforce Development & Diversity

Discovery & Innovation

Policy & Process

Leadership

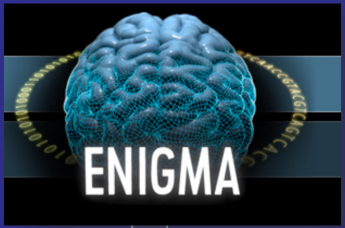
Research Objects in the Commons

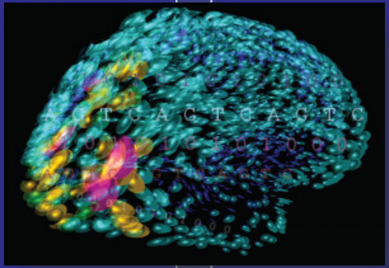
Over 100 Public Lectures
Collaboration with a Minority Institution

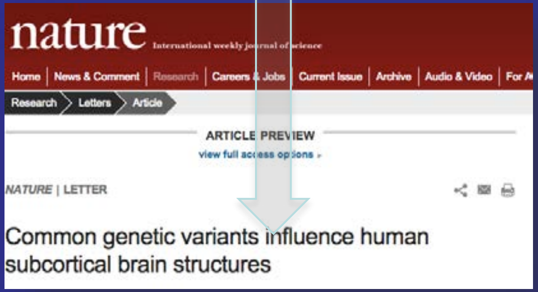
Voxel Wide Genome Scanning
MRI standardization

Genomic Data Sharing
Policy

185 Institutions Involved






nature International weekly journal of science
Home | News & Comment | Research | Careers & Jobs | Current Issue | Archive | Audio & Video | For A
Research > Letters > Article

ARTICLE PREVIEW

view full access options ›

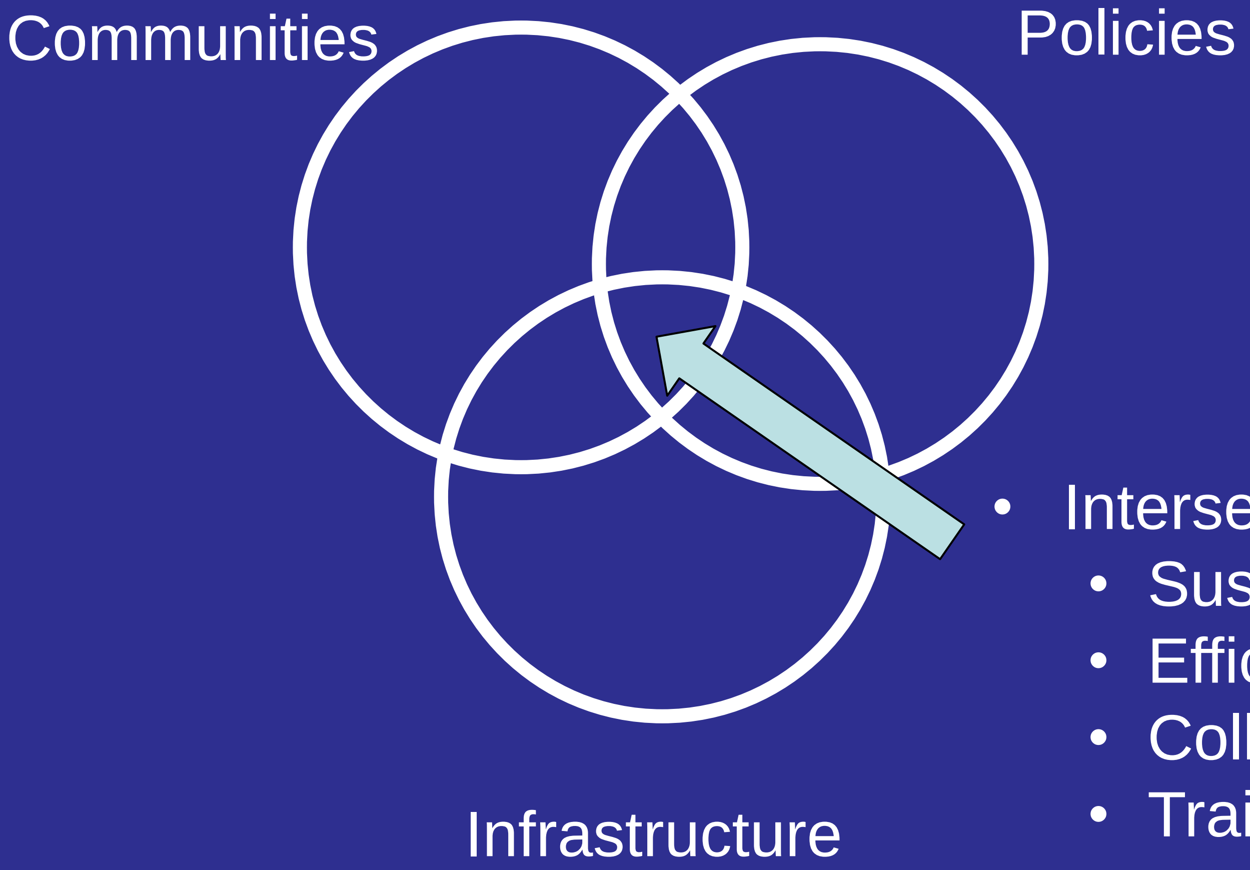
NATURE | LETTER

Common genetic variants influence human subcortical brain structures

Agenda

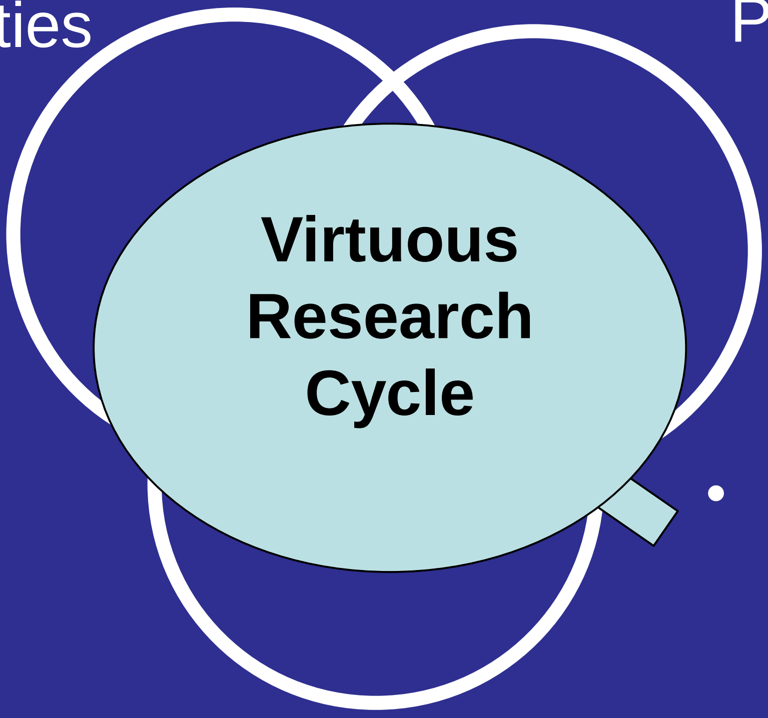
- Drivers of change
- Scientific motivators
- The NIH response
  - The Office of Data Science
  - The Big Data to Knowledge (BD2K) initiative
- The 3-legged stool strategy
  - Infrastructure
  - Policies
  - Communities
- Some future activities
- Your thoughts here?

Elements of Our Strategy



- Intersection:
  - Sustainability
  - Efficiency
  - Collaboration
  - Training

Elements of Our Strategy



Communities

Policies

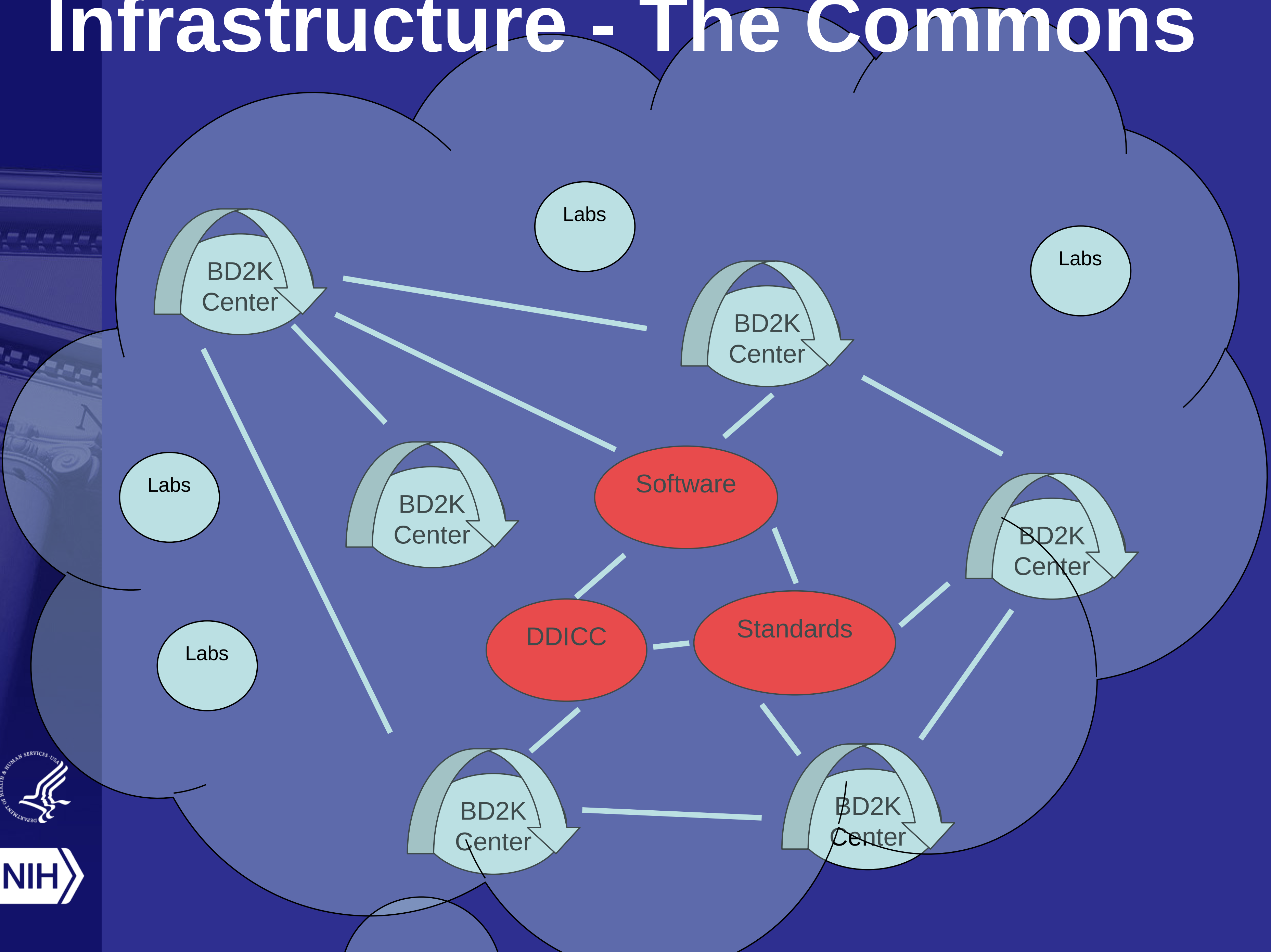
Virtuous Research Cycle

Infrastructure

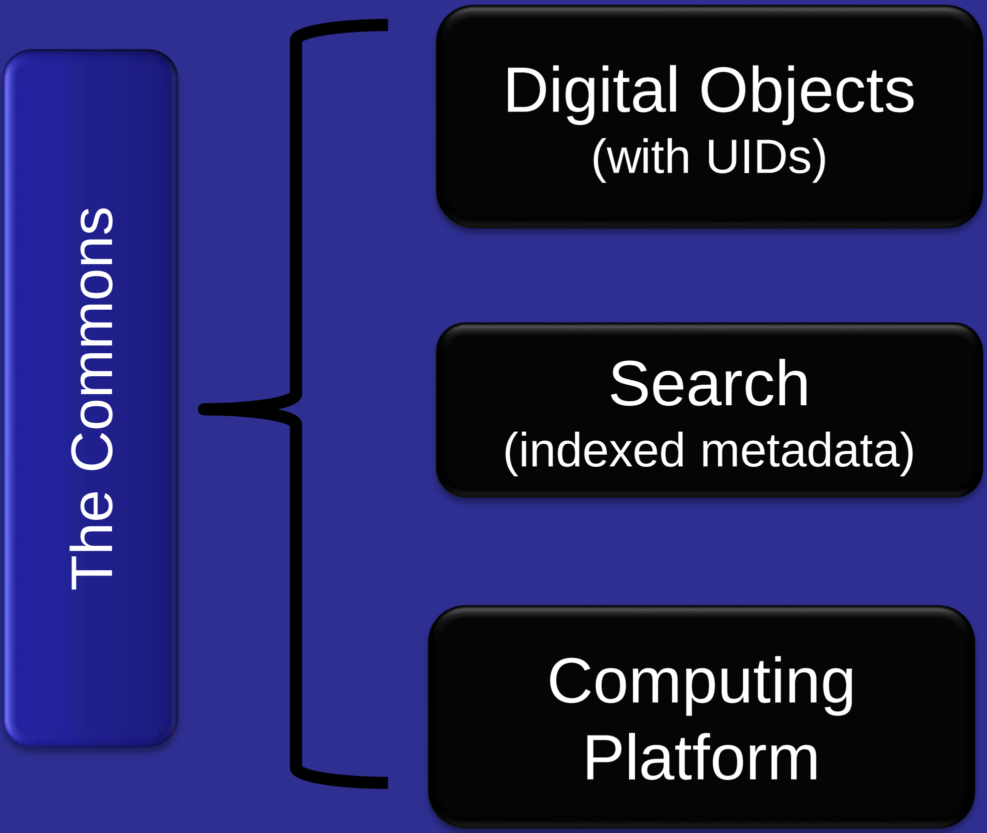
- Intersection:
  - Sustainability
  - Efficiency
  - Collaboration
  - Training

Consider Each Component Starting with the Infrastructure…

Infrastructure - The Commons

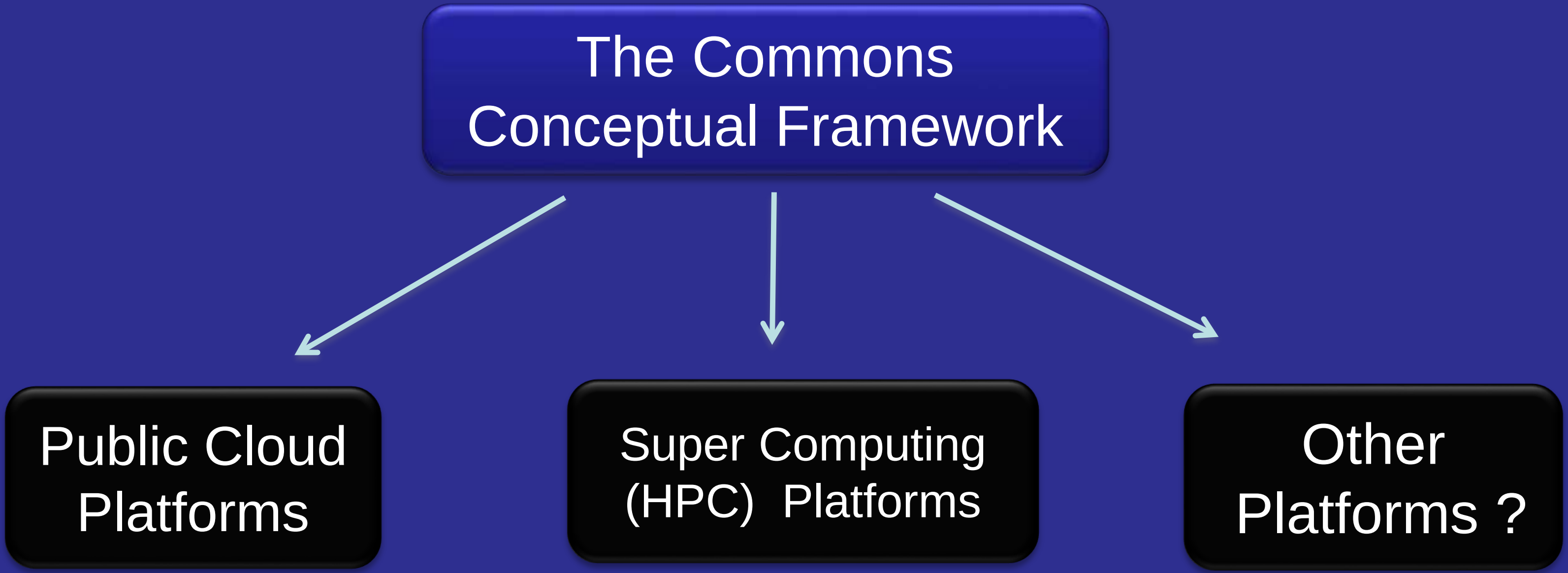


The Commons



Vivien Bonazzi
George Komatsoulis

The Commons: Compute Platforms



The Commons Conceptual Framework

Public Cloud Platforms

Super Computing (HPC) Platforms

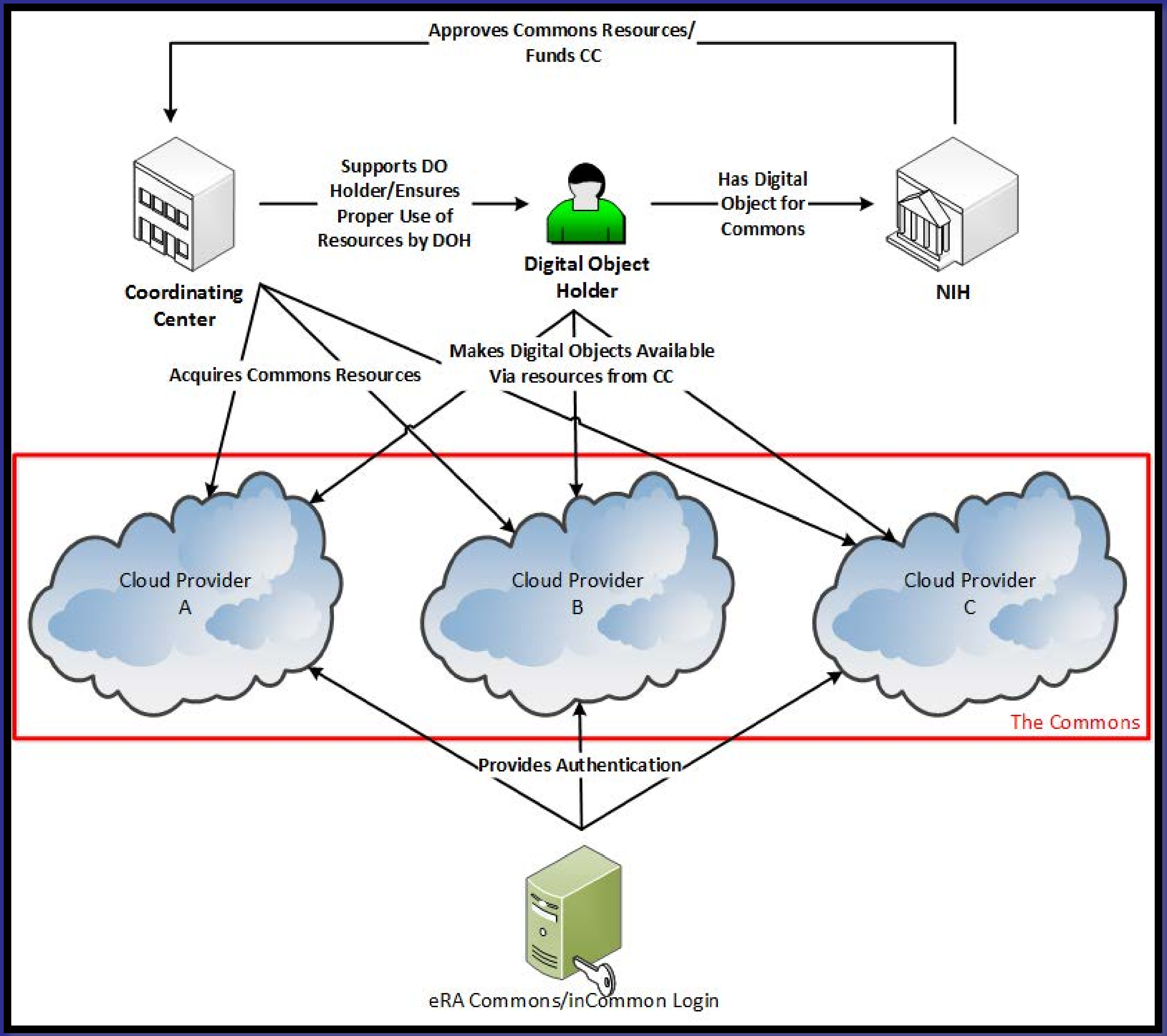
Other Platforms ?

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

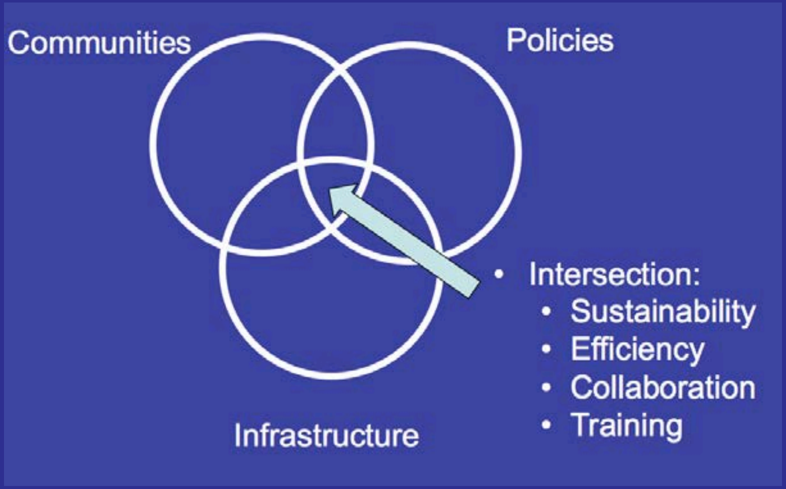
- Traditionally low access by NIH

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

The Commons:
Business Model



[George Komatsoulis]

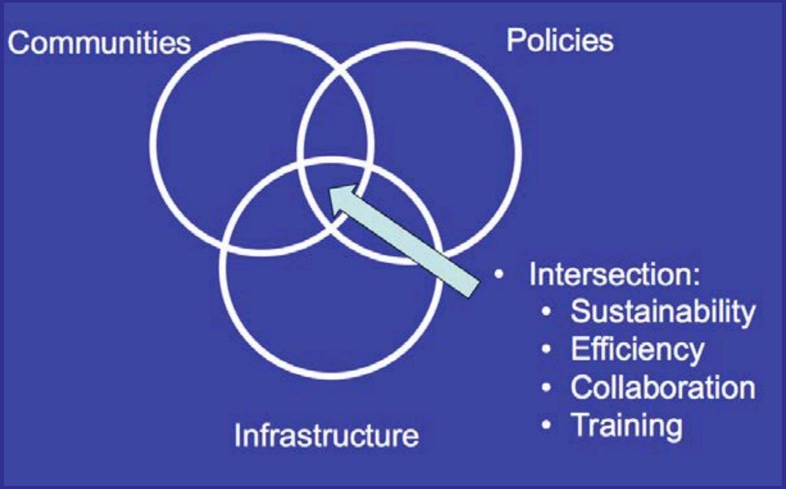


- Intersection:
  - Sustainability
  - Efficiency
  - Collaboration
  - Training

And Now Communities …

Communities: Example Activities

- 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



- Intersection:
  - Sustainability
  - Efficiency
  - Collaboration
  - Training

And Lastly Policies …

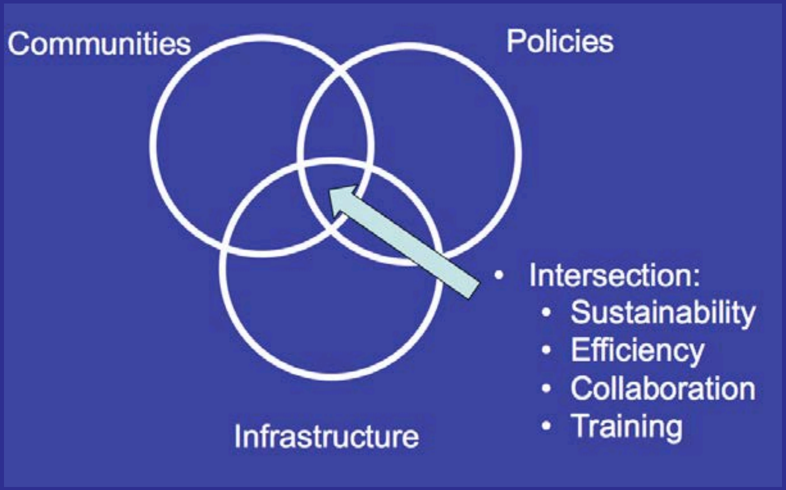
Policies: Now & Forthcoming

- 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!)



Communities

Policies

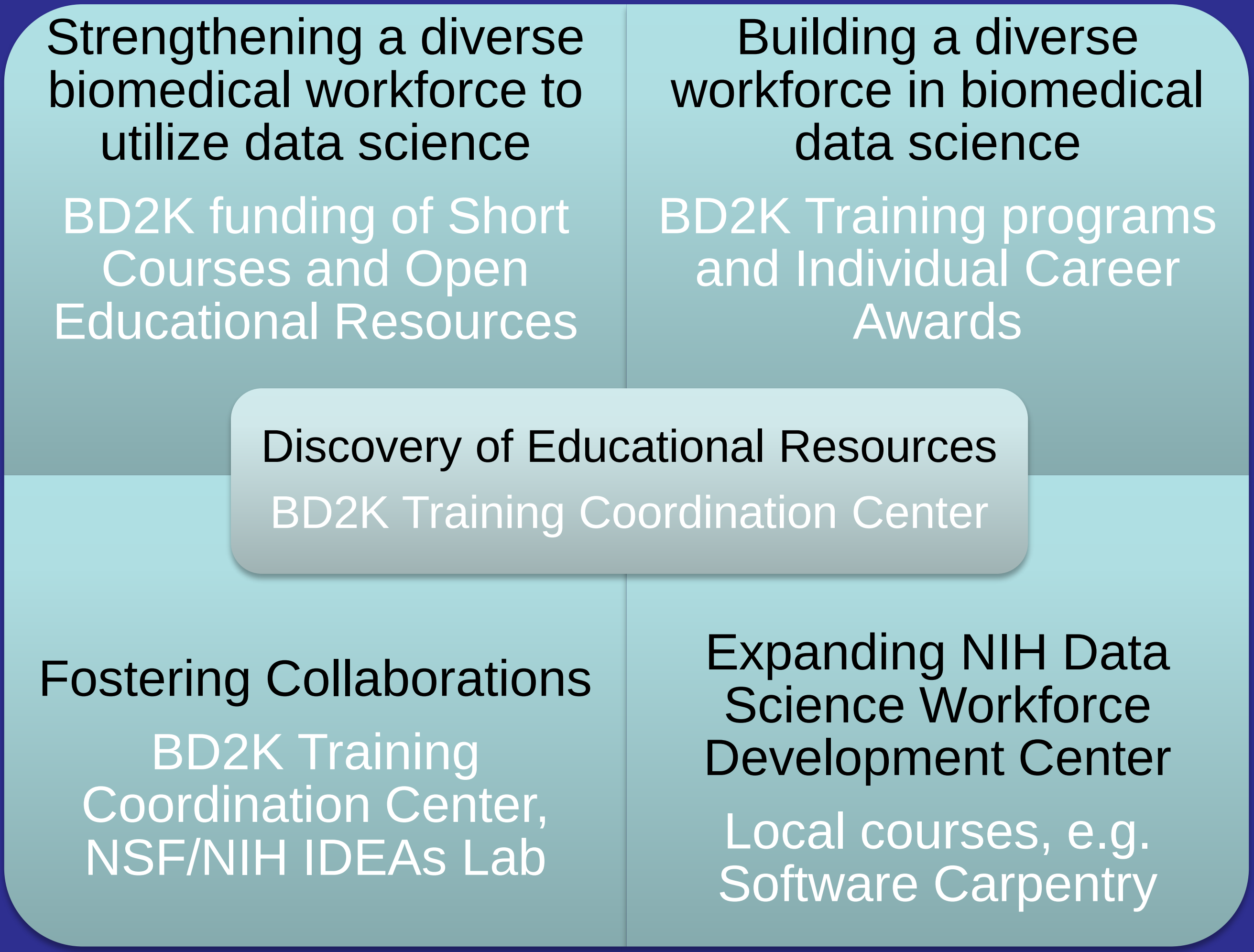
Infrastructure

- Intersection:
  - Sustainability
  - Efficiency
  - Collaboration
  - Training

Issues At The Intersection…

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

Agenda

- Drivers of change
- Scientific motivators
- The NIH response
  - The Office of Data Science
  - The Big Data to Knowledge (BD2K) initiative
- The 3-legged stool strategy
  - Infrastructure
  - Policies
  - Communities
- Some future activities
- Your thoughts here?

New BD2K Workshops

- NIH Common Data Elements (CDE) Workshop
  - in conjunction with NCI, NLM, and BMIC
- Academic deans and data science career paths workshop
  - in conjunction with NSF
- Data Science of Citizen Science
  - Possibly in conjunction with the Heart of Data Science (Ping) Center
- National Academies (CATS) workshop on big data inference
  - in conjunction with NSF/CISE
- National Academies (CSTB) workshop on data science curriculum development
  - in conjunction with NSF/CISE

New BD2K Activities

- **Reference Datasets**
  - Will move important, FAIR, digital resources into the cloud to support increased access and utility.
  - Will release an RFI, to inform details of an FOA.
  - Will follow up and formalize activities started in the Commons Administrative supplements
- **Sustainability of Data Repositories**
  - BD2K sustainability group is doing a financial and portfolio analysis of digital data repositories across NIH, and has been studying current sustainability approaches of such repositories.
  - BD2K is working in conjunction with NIGMS, NHGRI, and BISTI to develop FOA on this topic.
  - Will follow up on initial activities of the interoperability supplements.
- **Software Hardening Resource**
  - BD2K is standing up a group to develop a proposal to ensure useful software can develop from useful academic grade prototypes to more robust, commons-compliant tools.

FAIR = Findable, Accessible, Interoperable, Reusable

Agenda

- Drivers of change
- Scientific motivators
- The NIH response
  - The Office of Data Science
  - The Big Data to Knowledge (BD2K) initiative
- The 3-legged stool strategy
  - Infrastructure
  - Policies
  - Communities
- Diversity
- Your thoughts here?

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

– Woodrow Wilson

NIH…

philip.bourne@nih.gov

Turning Discovery Into Health