Big Data: An NIH Perspective

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

National Institutes of Health

October 16, 2015

Big Data Challenges in the Life Sciences …

This speaks to something more fundamental that more data …

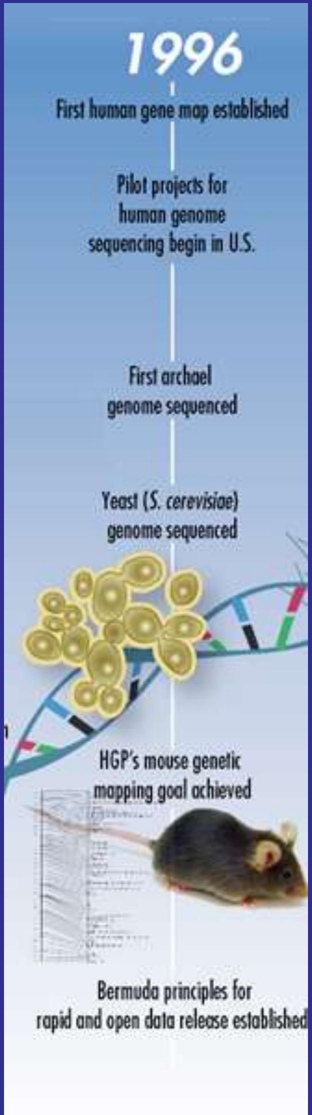
It speaks to new methodologies, new skills, new emphasis, new cultures, new modes of discovery …

New types of funding

Conversation Cards

- A brief historical perspective
- What could happen in the future
- The implications for this future
- NIH initiatives in this landscape
  - – Big Data to Knowledge (BD2K)
  - – Precision Medicine

Laying the Foundation for Open Access: HGP, Bermuda, 1996



1996

First human gene map established

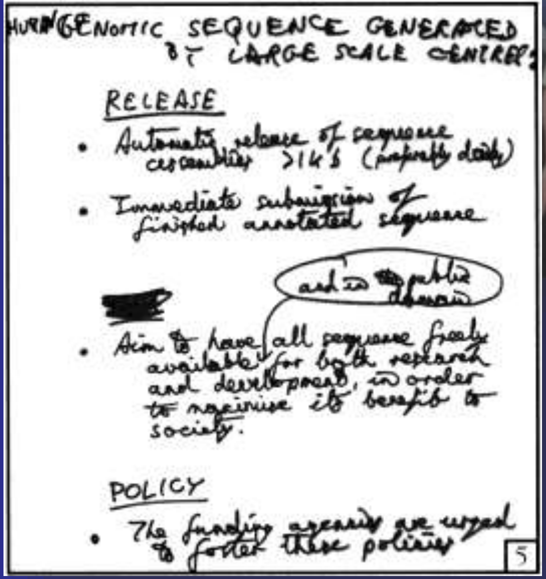
Pilot projects for human genome sequencing begin in U.S.

First archael genome sequenced

Yeast (*S. cerevisiae*) genome sequenced

HGP's mouse genetic mapping goal achieved

Bermuda principles for rapid and open data release established



HUMAN GENOMIC SEQUENCE GENERATED BY LARGE SCALE CENTRES:

RELEASE

- Automatic release of sequence assemblies >1k's (preferably daily)
- Immediate submission of finished annotated sequence

[redacted]

- Aim to have all sequence freely available and in the public domain for both research and development, in order to maximise its benefit to society.

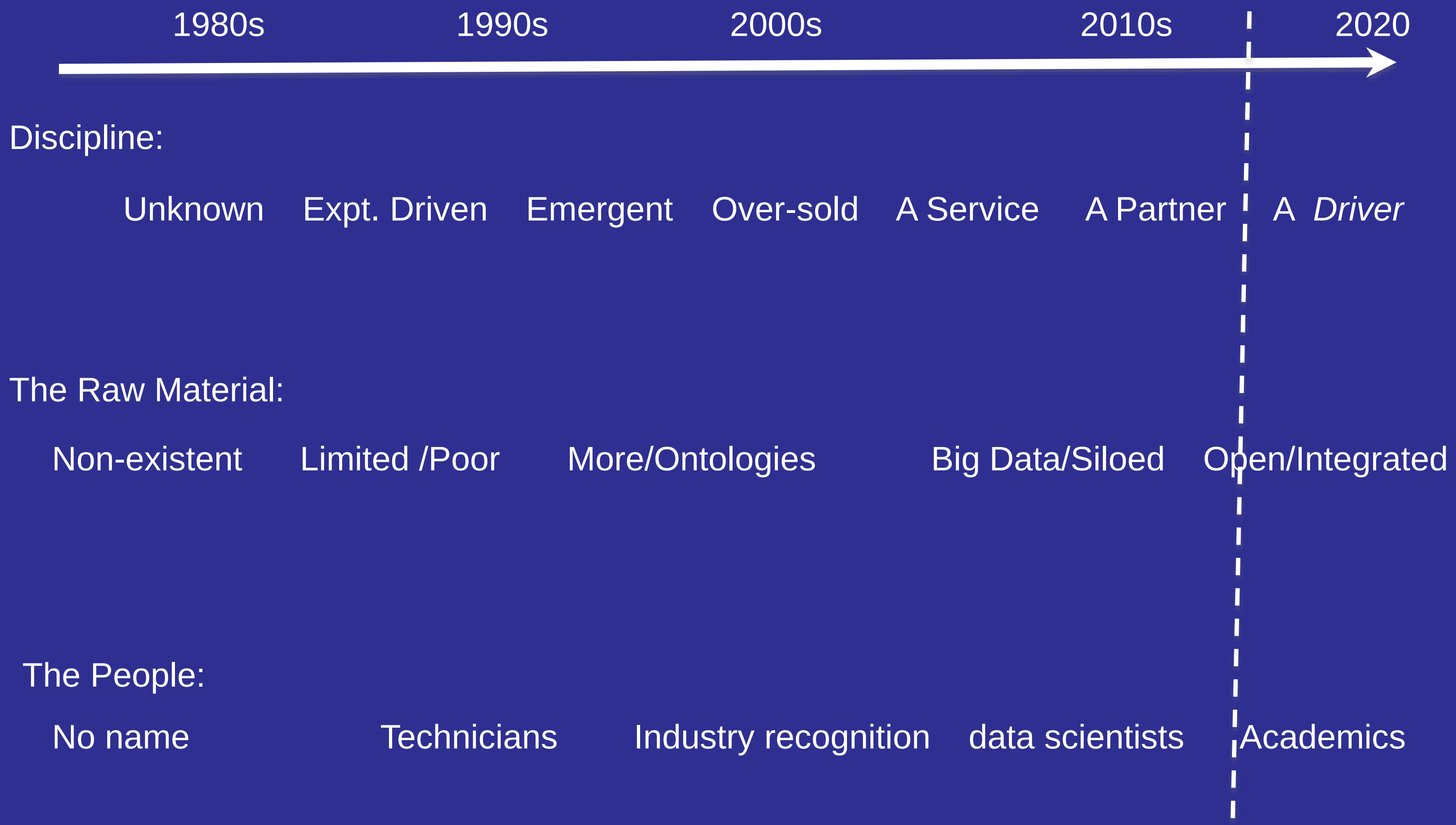
POLICY

- The funding agencies are urged to foster these policies

5

The History of Computational Biomedicine According to Bourne

Searls (ed) The Roots in Bioinformatics Series *PLOS Comp Biol*



1980s 1990s 2000s 2010s 2020

Discipline:

Unknown Expt. Driven Emergent Over-sold A Service A Partner A *Driver*

The Raw Material:

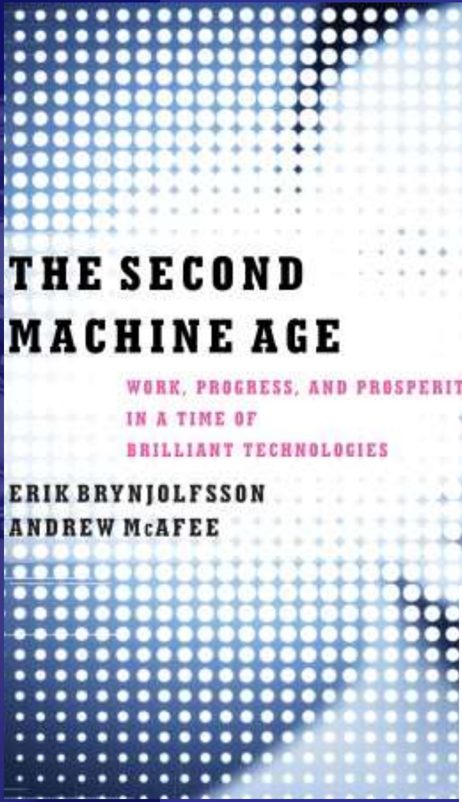
Non-existent Limited /Poor More/Ontologies Big Data/Siloed Open/Integrated

The People:

No name Technicians Industry recognition data scientists Academics

What could happen in the future?

We are at a Point of Deception …

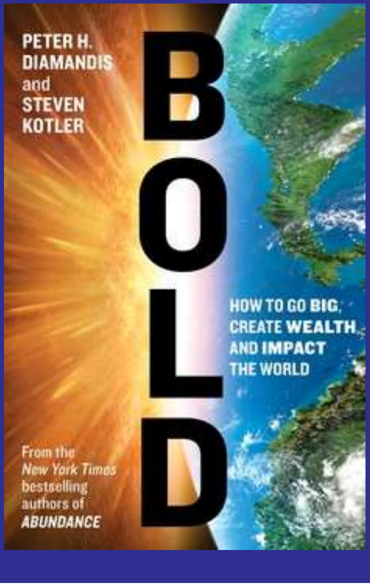


THE SECOND MACHINE AGE

WORK, PROGRESS, AND PROSPERITY IN A TIME OF BRILLIANT TECHNOLOGIES

ERIK BRYNJOLFSSON
ANDREW McAFEE

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



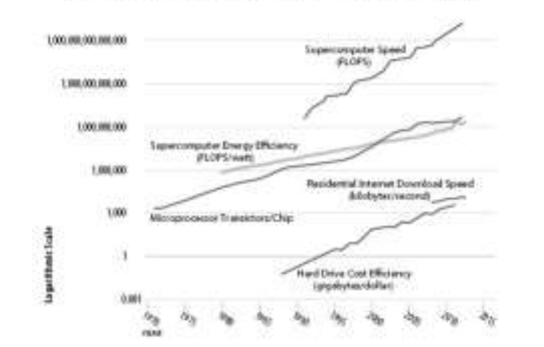
PETER H. DIAMANDIS and STEVEN KOTLER

BOLD

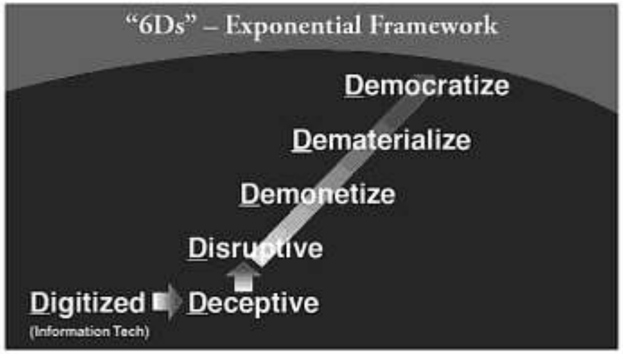
HOW TO GO BIG, CREATE WEALTH, AND IMPACT THE WORLD

From the New York Times bestselling authors of ABUNDANCE

FIGURE 3.3 The Many Dimensions of Moore's Law



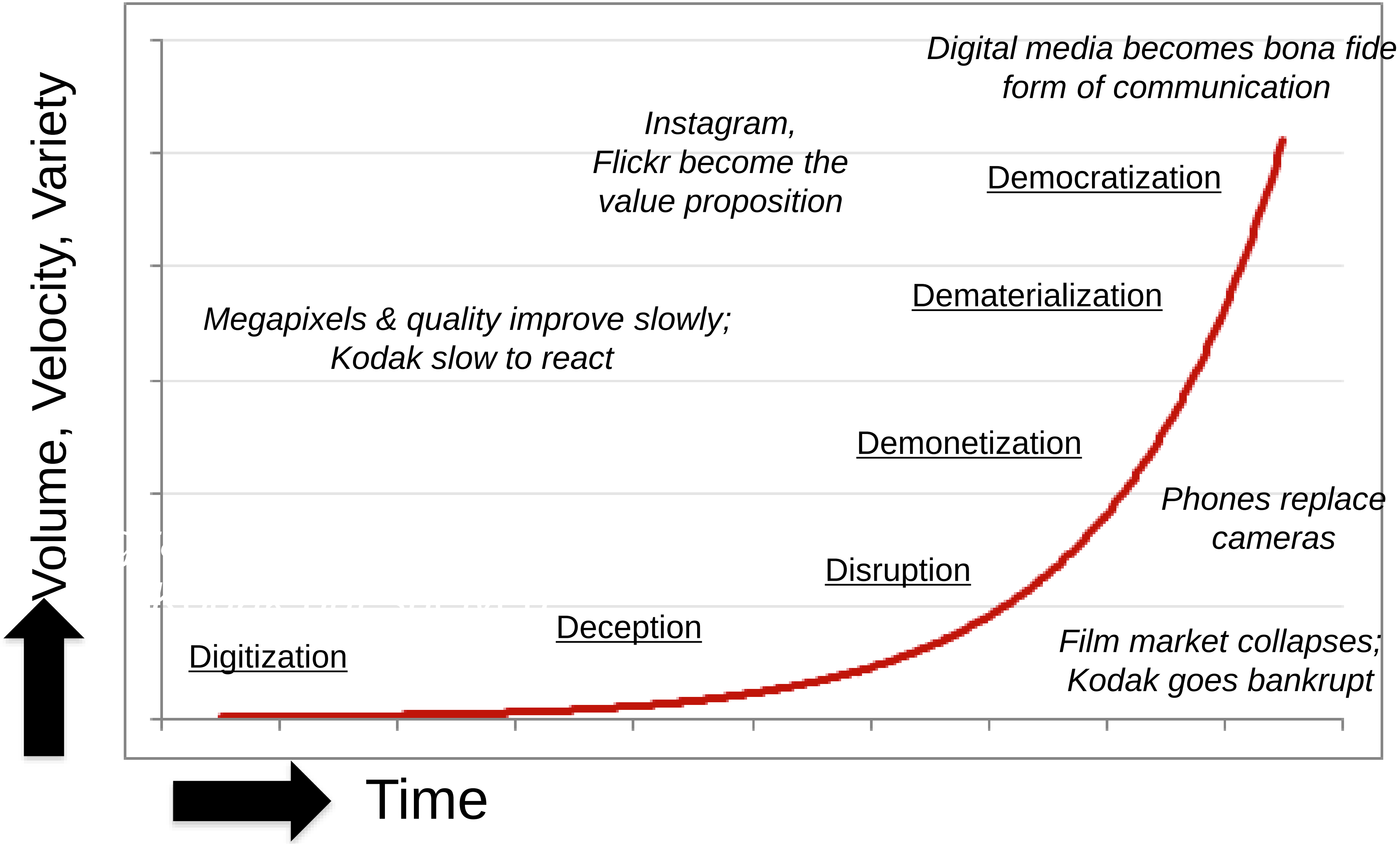
From: The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies by Erik Brynjolfsson & Andrew McAfee



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

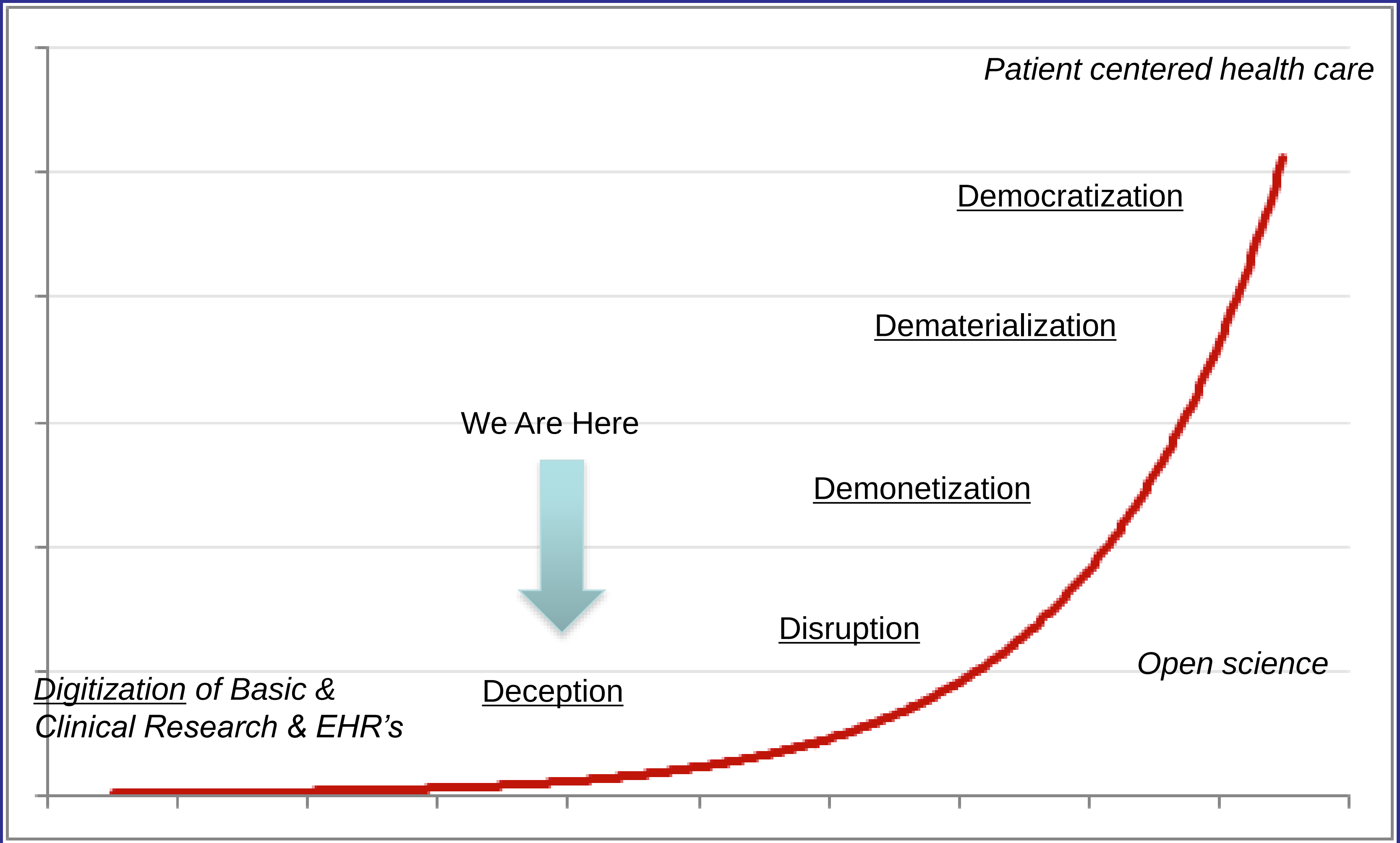
Disruption

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

What Are Some Implications of Such a Future?

- Open collaborative science becomes of increasing importance
- The value of data and associated analytics becomes of increasing value to scholarship
- Opportunities exist to improve the efficiency of the research enterprise and hence fund more research
- Cooperation between funders will be needed to sustain the emergent digital enterprise
- Current training content and modalities will not match supply to demand
- Precision medicine is indeed a reality

What Are Some Implications of Such a Future?

- Open collaborative science becomes of increasing importance
- The value of data and associated analytics becomes of increasing value to scholarship
- Opportunities exist to improve the efficiency of the research enterprise and hence fund more research
- Cooperation between funders will be needed to sustain the emergent digital enterprise
- Current training content and modalities will not match supply to demand
- Precision medicine is indeed a reality

“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.”

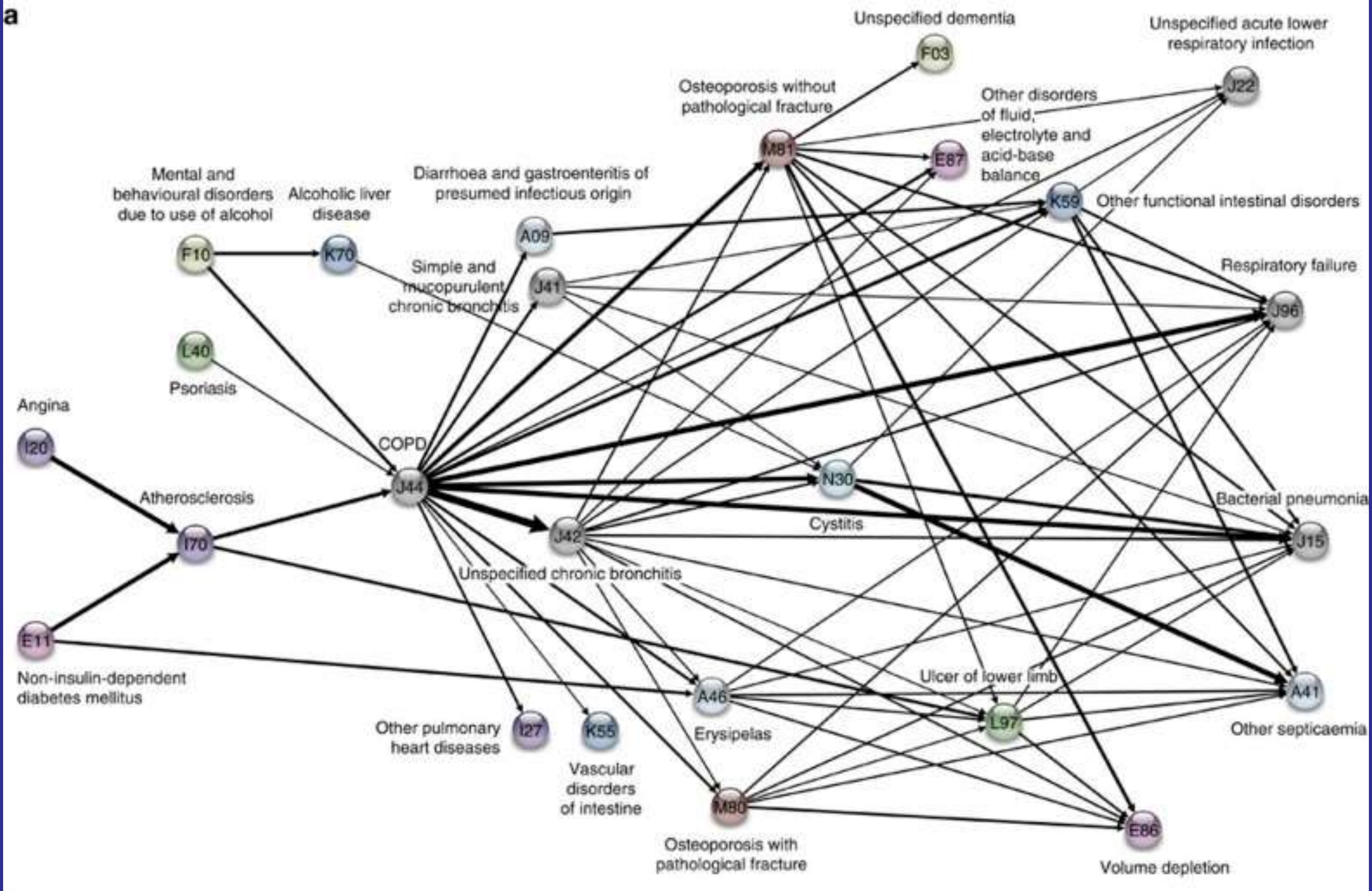
President Barack Obama

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

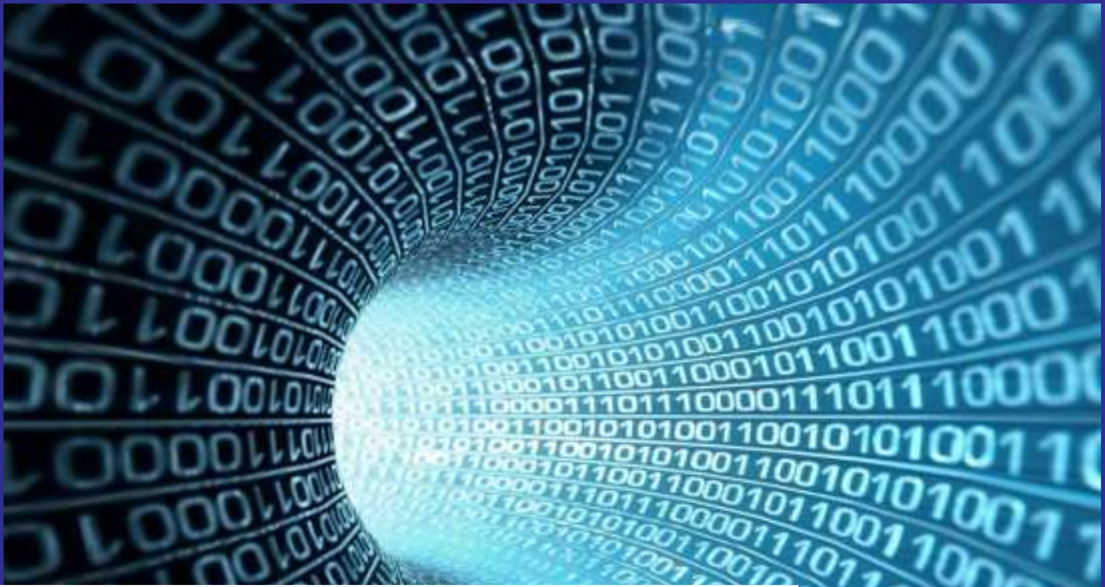
An Example of That Promise: Comorbidity Network for 6.2M Danes Over 14.9 Years

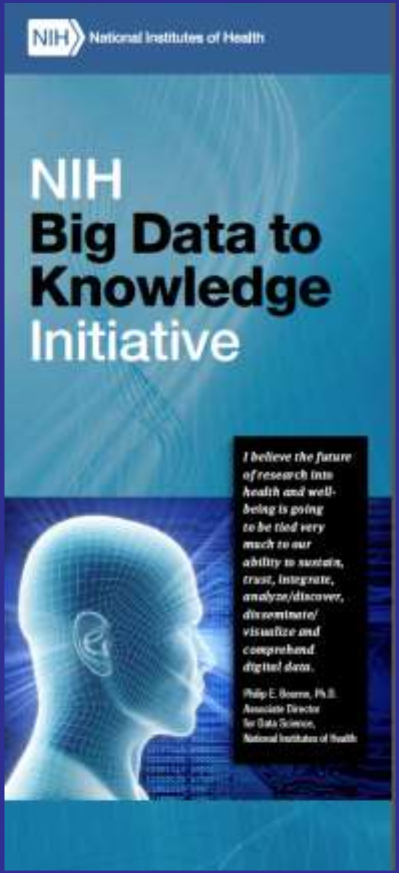

a
Unspecified dementia
F03
Unspecified acute lower respiratory infection
J22
Osteoporosis without pathological fracture
M81
Other disorders of fluid, electrolyte and acid-base balance
E87
K59
Other functional intestinal disorders
Mental and behavioural disorders due to use of alcohol
Alcoholic liver disease
Diarrhoea and gastroenteritis of presumed infectious origin
A09
F10
K70
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

Conversation Cards

- A brief historical perspective
- What could happen in the future
- The implications for this future
- NIH initiatives in this landscape
  - – Big Data to Knowledge (BD2K)
  - – Precision Medicine



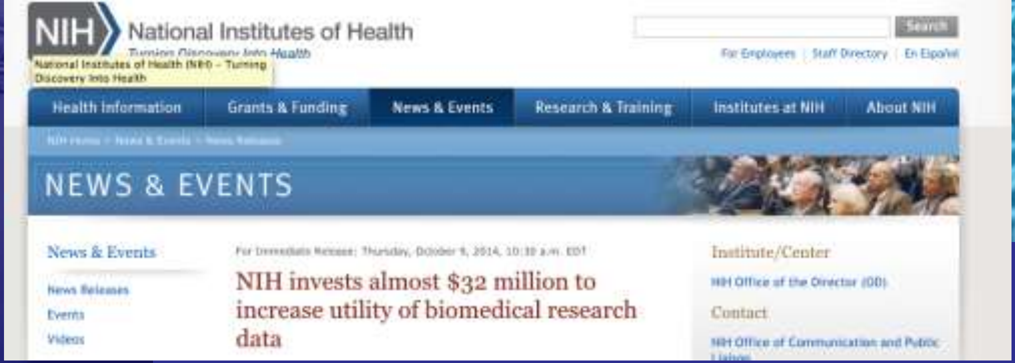


NIH National Institutes of Health

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



NIH National Institutes of Health
Turning Discovery Into Health

National Institutes of Health (NIH) - Turning Discovery Into Health

NEWS & EVENTS

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

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

Institute/Center

NIH Office of the Director (OD)

Contact

NIH Office of Communication and Public Liaison

What is the NIH Doing to Fulfill That Promise?

ADDS Mission Statement

To foster an open ecosystem that enables biomedical* research to be conducted as a digital enterprise that *enhances health, lengthens life and reduces illness and disability*

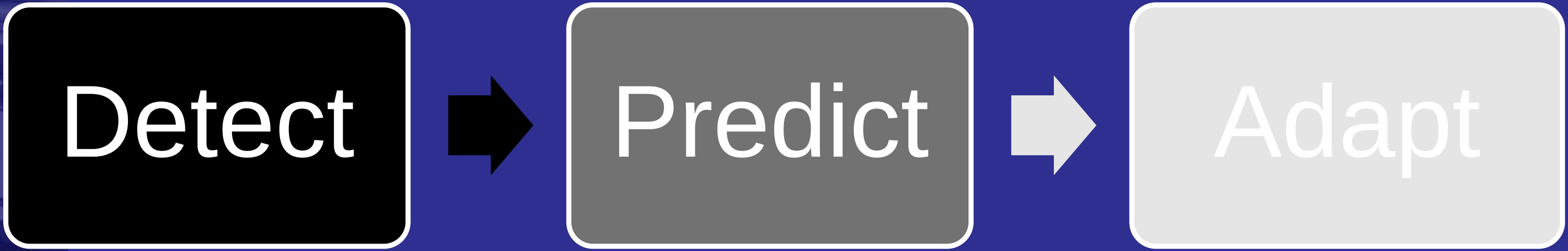
** Includes biological, biomedical, behavioral, social, environmental, and clinical studies that relate to understanding health and disease.*

Center of Excellence for Mobile Sensor Data-to-Knowledge (MD2K)

Santosh Kumar, Ph.D.
Director, MD2K Center of Excellence
Professor & Moss Chair of Excellence in Computer Science
University of Memphis

https://datascience.nih.gov/bd2k/funded-programs/centers

MD2K Applications – CHF and Smoking


Detect

Predict

Adapt

Example: BD2K Center

Working Across Strategic Areas

Strategic Areas

Sustainability

Workforce Development & Diversity

Discovery & Innovation

Policy & Process

Leadership

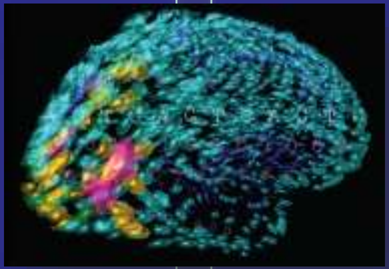
Research Objects in the Commons

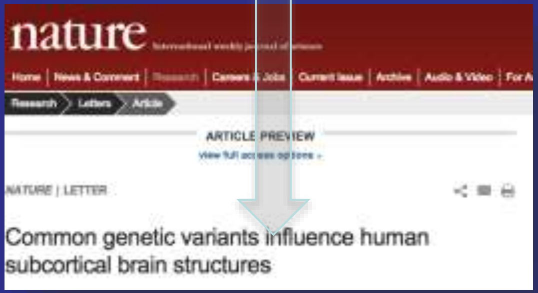
Over 100 Public Lectures
Collaboration with a Minority Institution

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

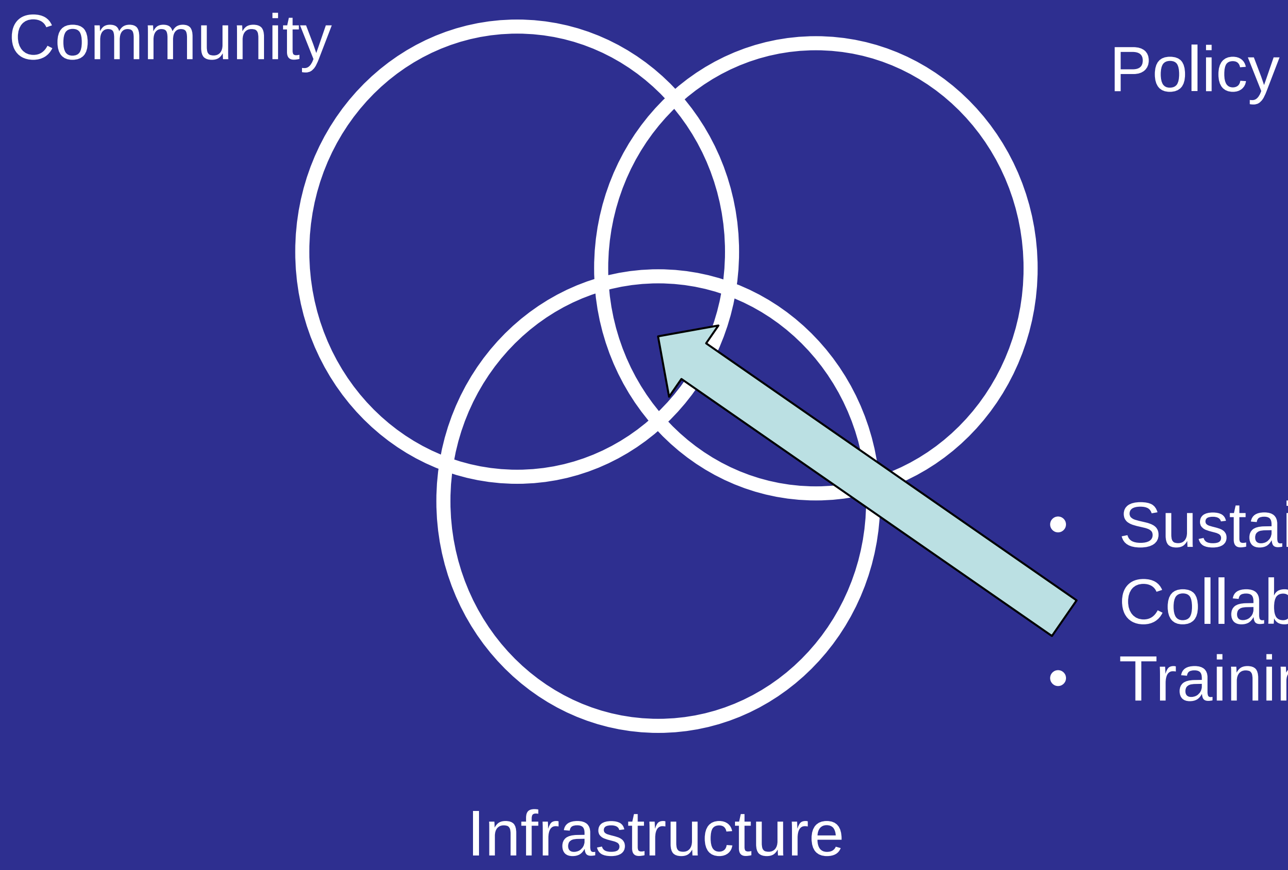
ARTICLE PREVIEW

View full access options

NATURE | LETTER

Common genetic variants influence human subcortical brain structures

Elements of The Ecosystem



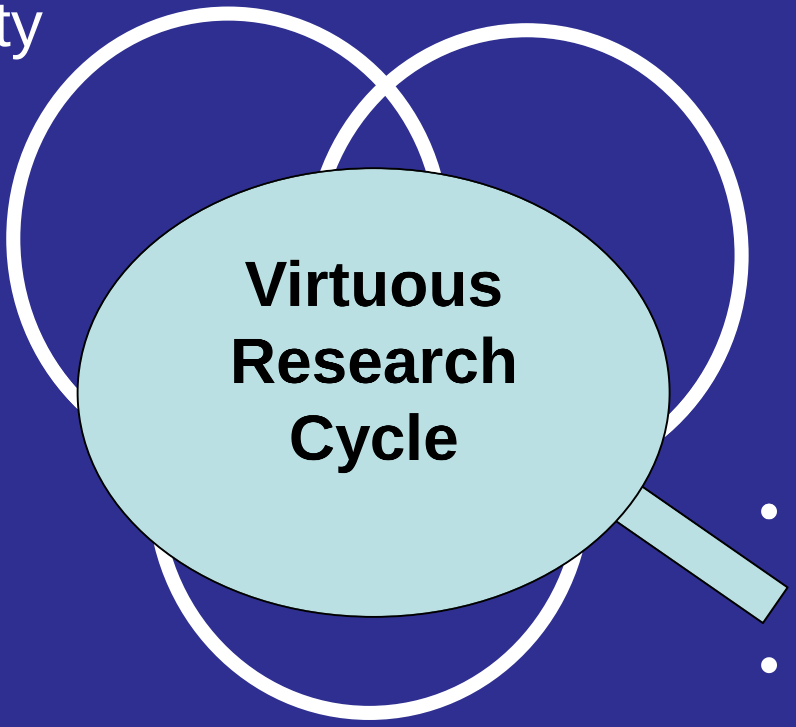
Community

Policy

Infrastructure

- Sustainability
  Collaboration
- Training

Elements of The Ecosystem



Community

Policy

Virtuous Research Cycle

Infrastructure

- Sustainability
  Collaboration
- Training

Policies – Now & Forthcoming

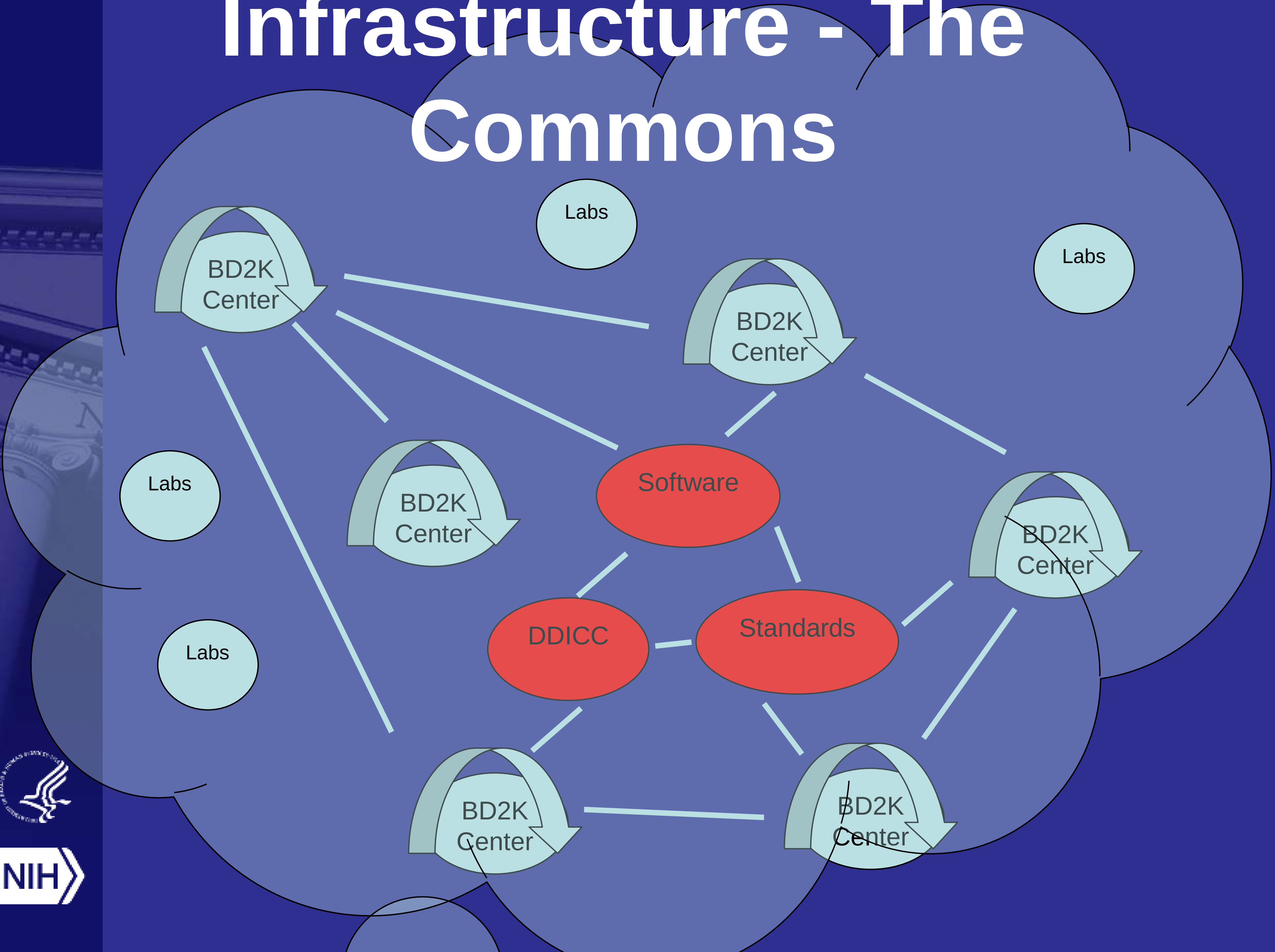
- Data Sharing
  - Genomic data sharing announced
  - Data sharing plans on all research awards
  - Data sharing plan enforcement
    - Machine readable plan
    - Repository requirements to include grant numbers

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

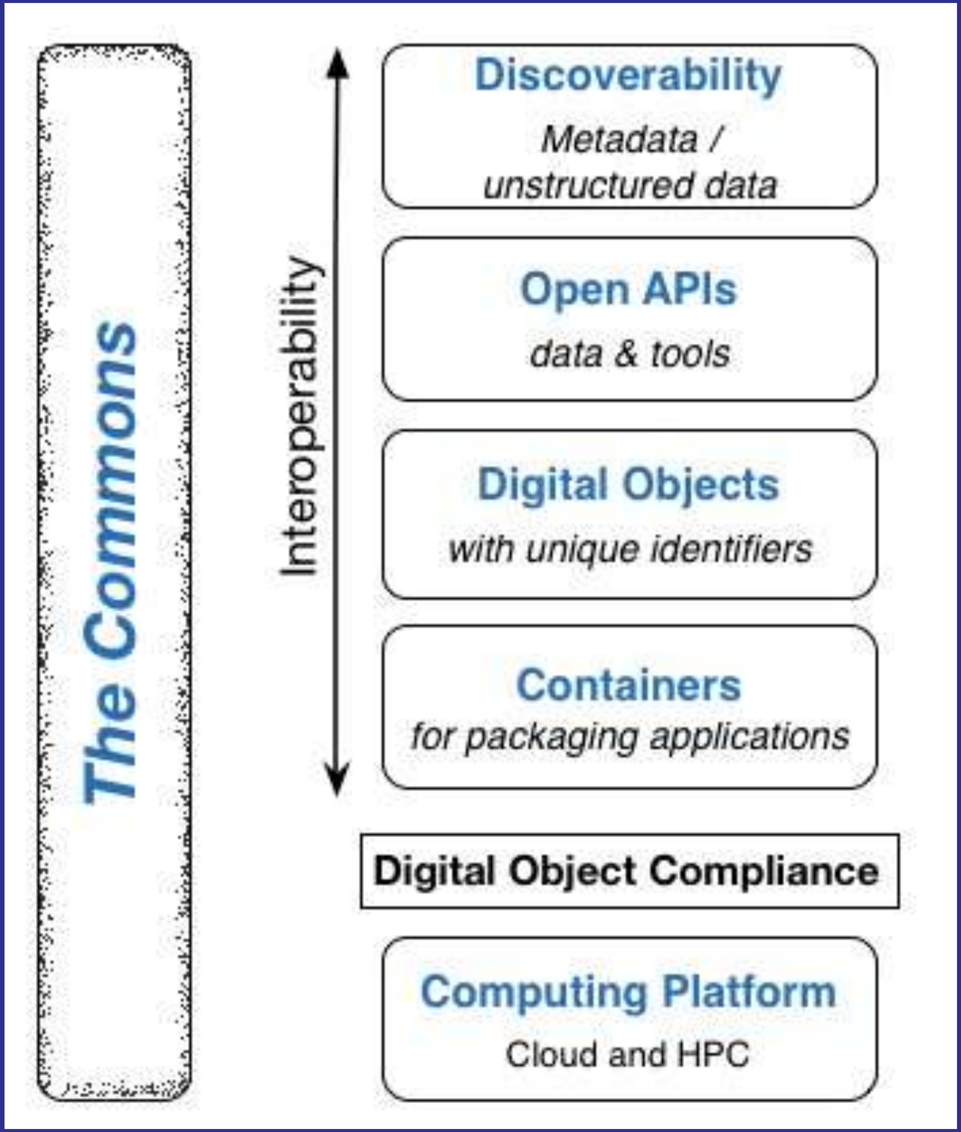
Policies - Forthcoming

- 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

Infrastructure - The Commons



The Commons: Components



The Commons

Digital Object Compliance: FAIR

- **Attributes of digital objects in the Commons**
- **Initial Phase**
    - Unique digital object identifiers of some type
    - A minimal set of searchable metadata
    - Physically available in a cloud based *Commons* provider
    - Clear access rules (especially important for human subjects data)
    - An entry (with metadata) in one or more indices

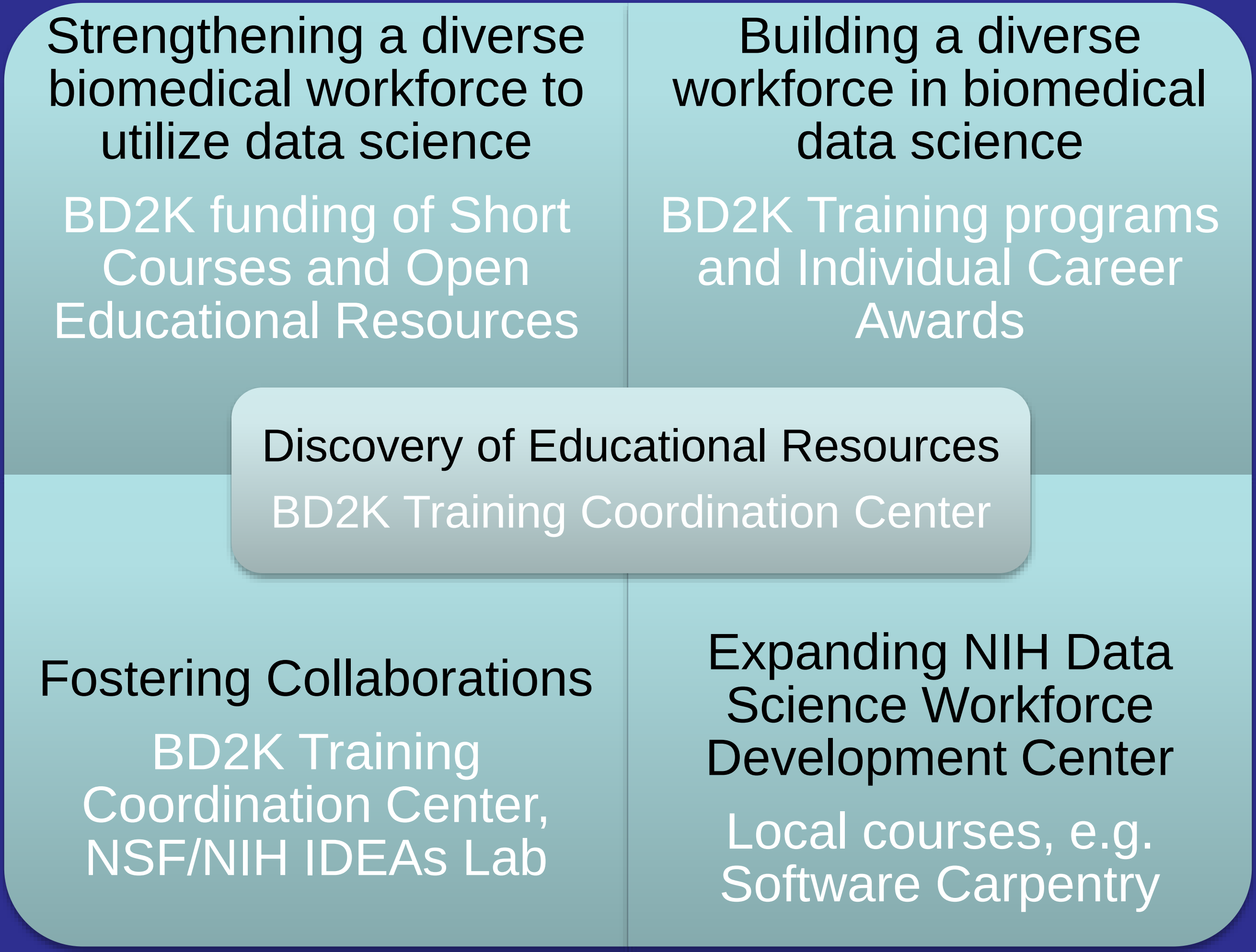
  - **Future Phases**
    - Standard, community based unique digital object identifiers
    - Conform to community approved standard metadata for enhanced searching
    - Digital objects accessible via open standard APIs
    - Are physically and logical available to the commons

The Commons:
Evaluation Pilots Underway

Evaluation Criteria	Pilot
Implementation	BD2K Centers
Interoperability	Model organism databases
Computation on Big Data	HMP data and tools in the cloud
Multi-cloud accessibility	NCI cloud pilots & genomic data commons
Business model	Supply and demand via credits

A Quick Word on 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**

The Team

philip.bourne@nih.gov

https://datascience.nih.gov/

http://www.ncbi.nlm.nih.gov/research/staff/bourne/

NIH…

Turning Discovery Into Health