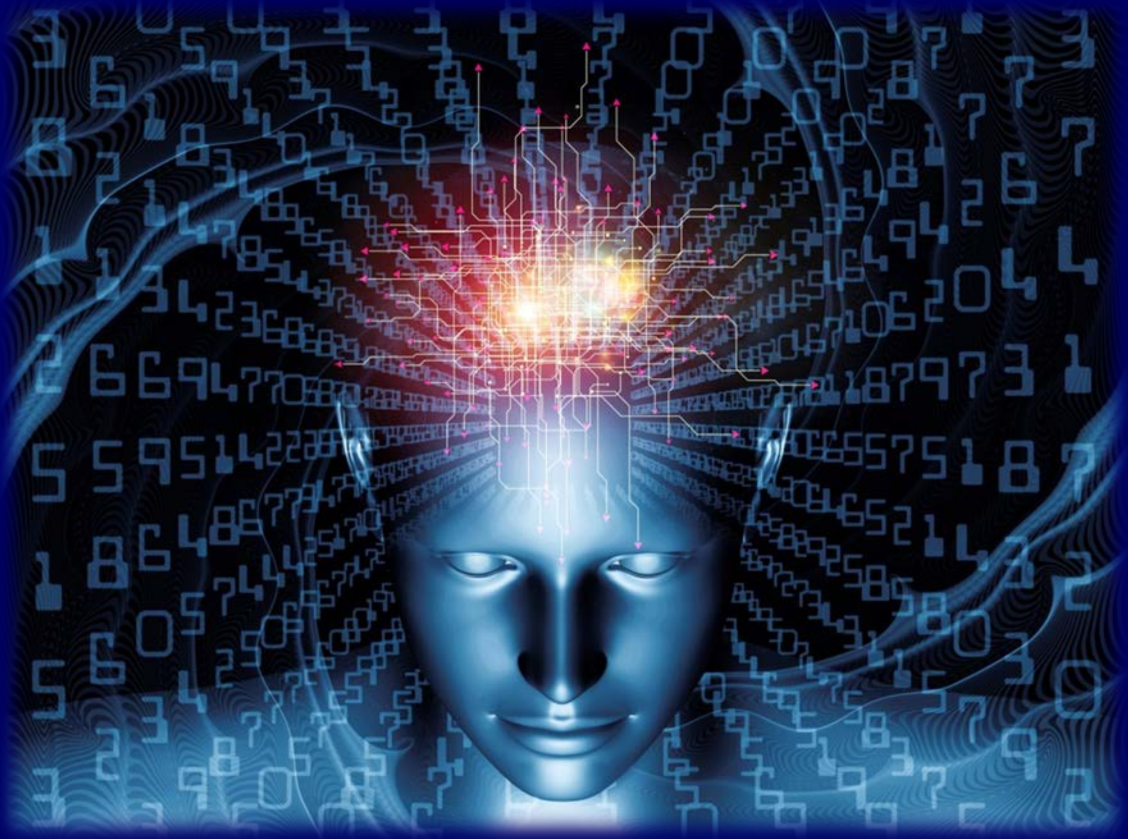
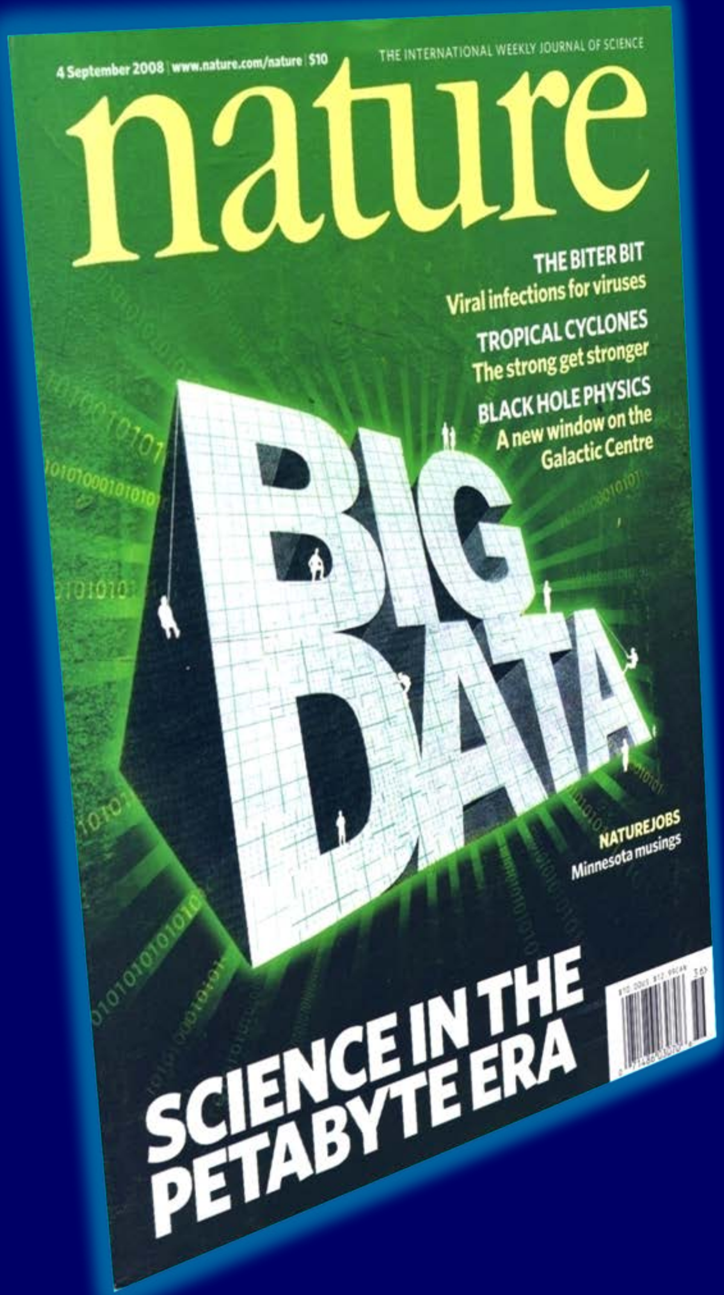


NIH and Biomedical ‘Big Data’

Eric Green, M.D., Ph.D.

Director, NHGRI

Acting Associate Director for Data Science, NIH



4 September 2008 | www.nature.com/nature | $10

THE INTERNATIONAL WEEKLY JOURNAL OF SCIENCE

nature

THE BITER BIT
Viral infections for viruses

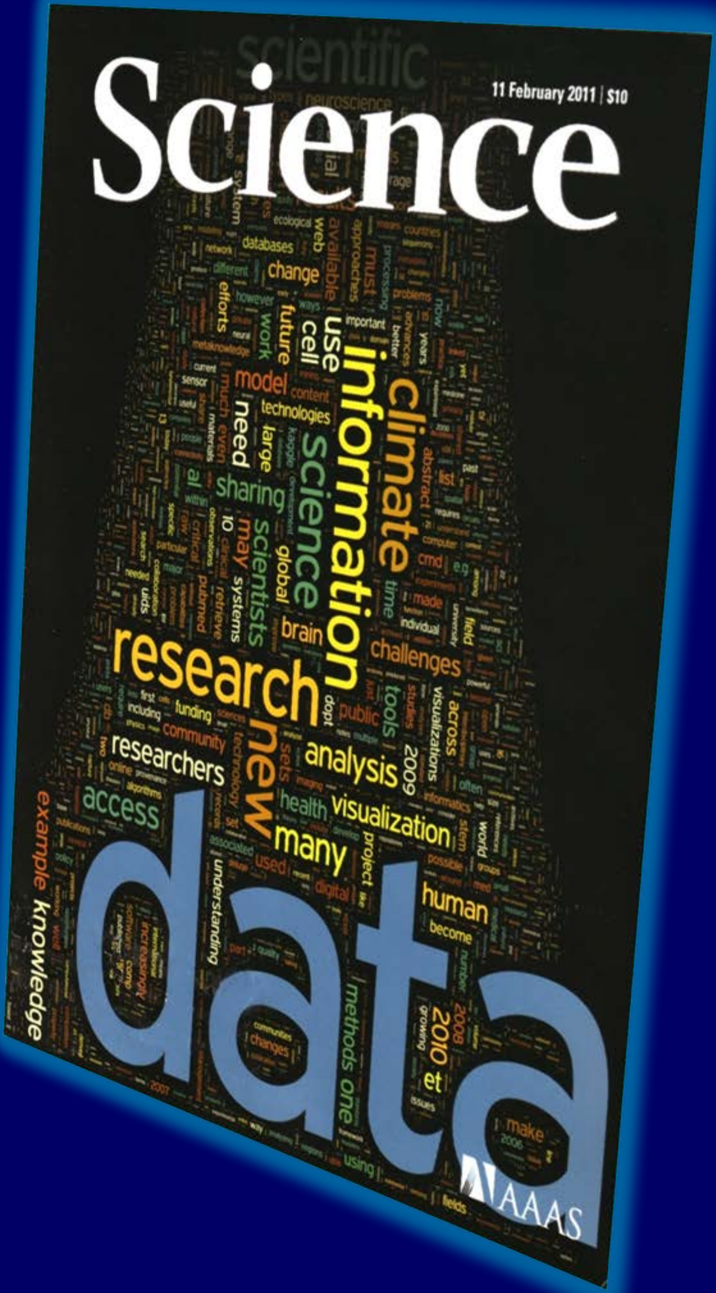
TROPICAL CYCLONES
The strong get stronger

BLACK HOLE PHYSICS
A new window on the Galactic Centre

BIG DATA

NATUREJOBS
Minnesota musings

SCIENCE IN THE PETABYTE ERA

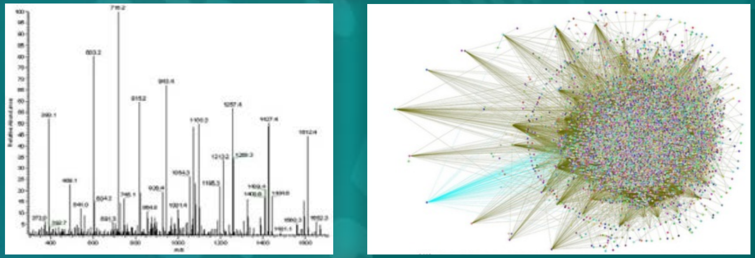


Science

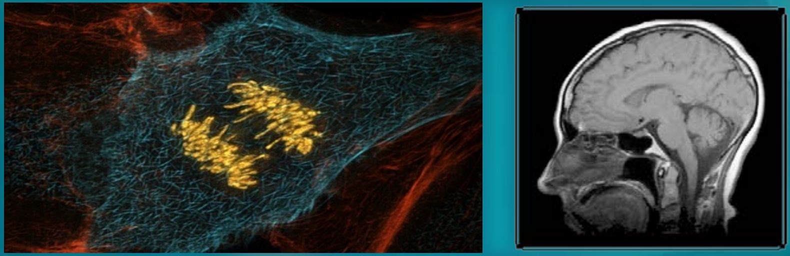
11 February 2011 | $10

Myriad Data Types

Genomic



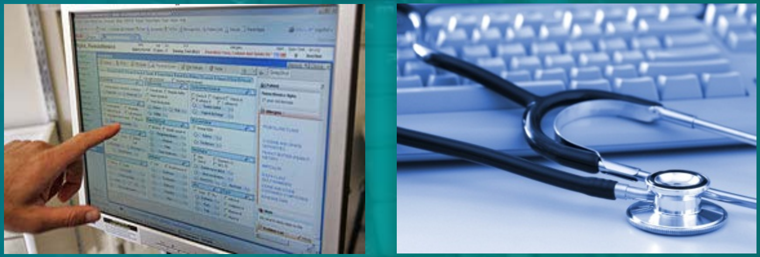
Other 'Omic



Imaging

Phenotypic

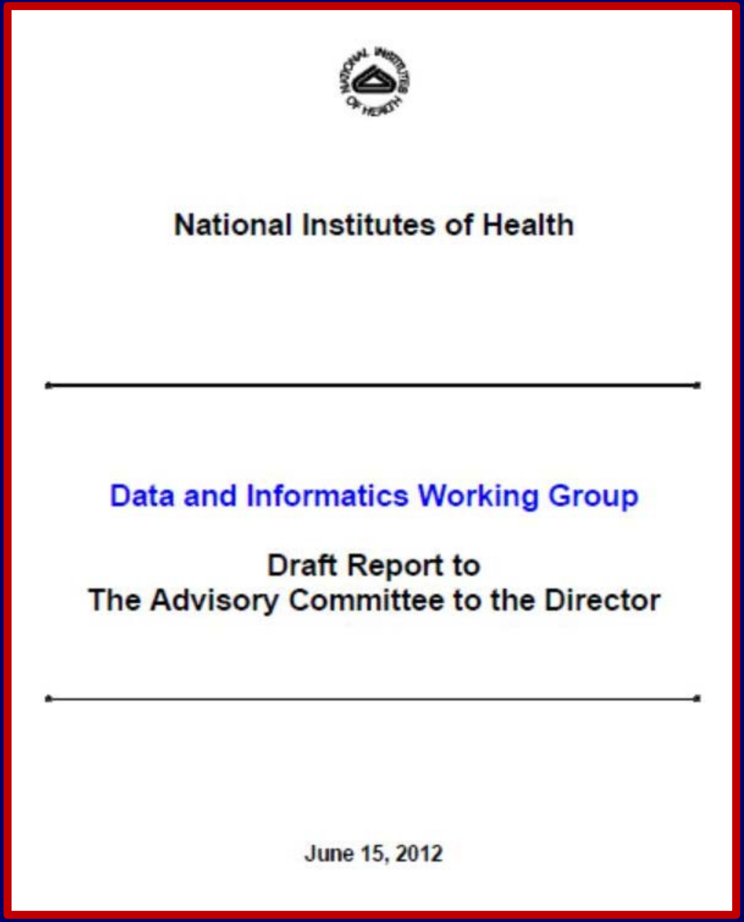
Exposure



Clinical

Data and Informatics Working Group

ADVISORY COMMITTEE TO THE DIRECTOR

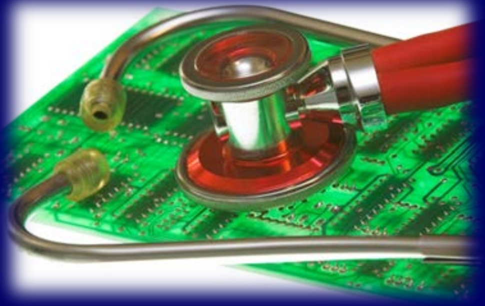


National Institutes of Health

Data and Informatics Working Group

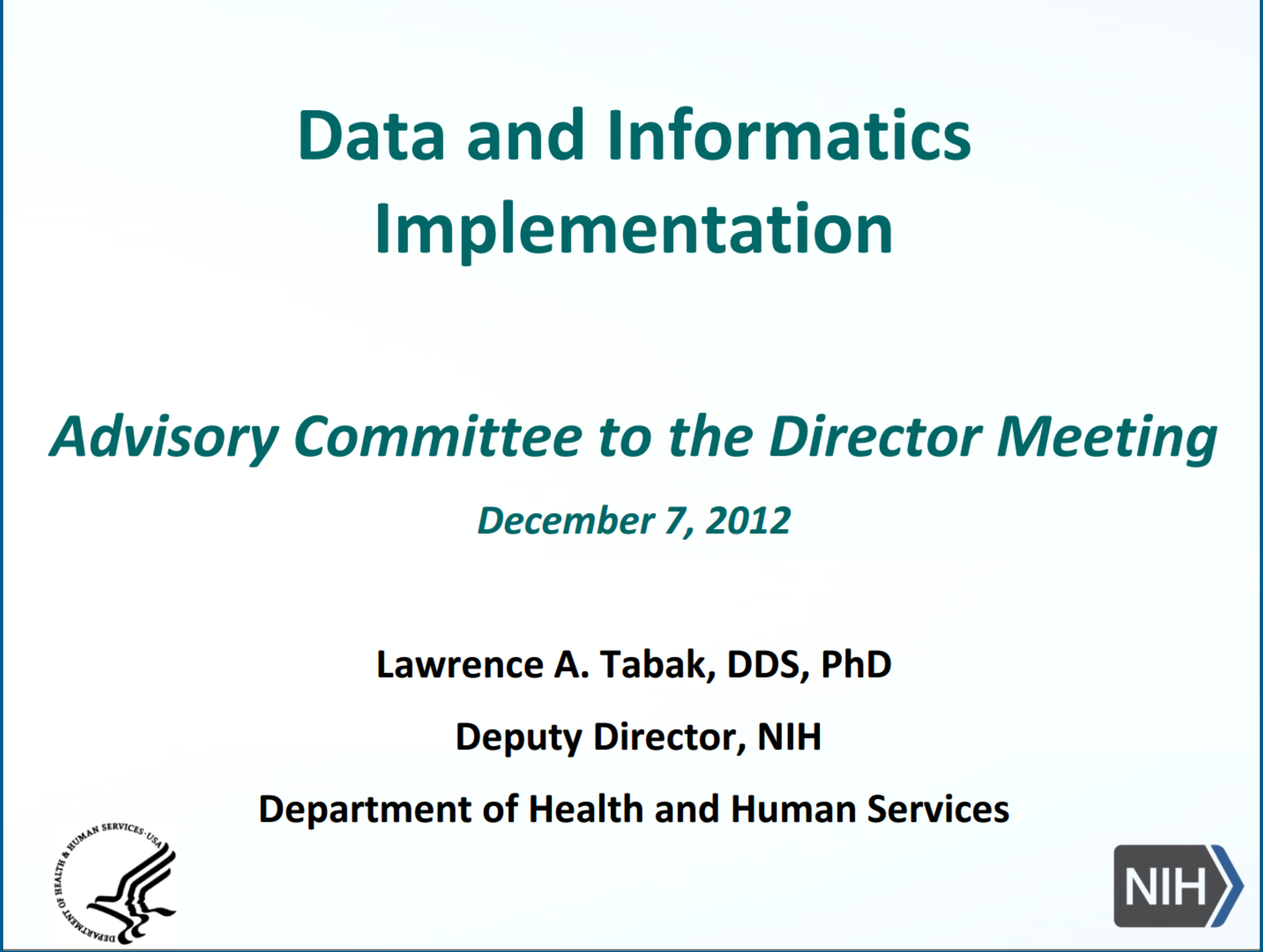
Draft Report to
The Advisory Committee to the Director

June 15, 2012



acd.od.nih.gov/diwg.htm

Implementation Plan: December 2012



Data and Informatics Implementation

Advisory Committee to the Director Meeting

December 7, 2012

Lawrence A. Tabak, DDS, PhD

Deputy Director, NIH

Department of Health and Human Services

Overarching Themes

- **At a pivotal point:**
  - **Risk failing to capitalize on technology advances**
  - **Bordering on “institutional malpractice”**
- **Cultural changes at NIH are essential**
- **Aim to develop new opportunities for:**
  - **Data sharing**
  - **Data analysis**
  - **Data integration**
- **Long-term NIH commitment is required**

Relevant Quote to Set the Stage…

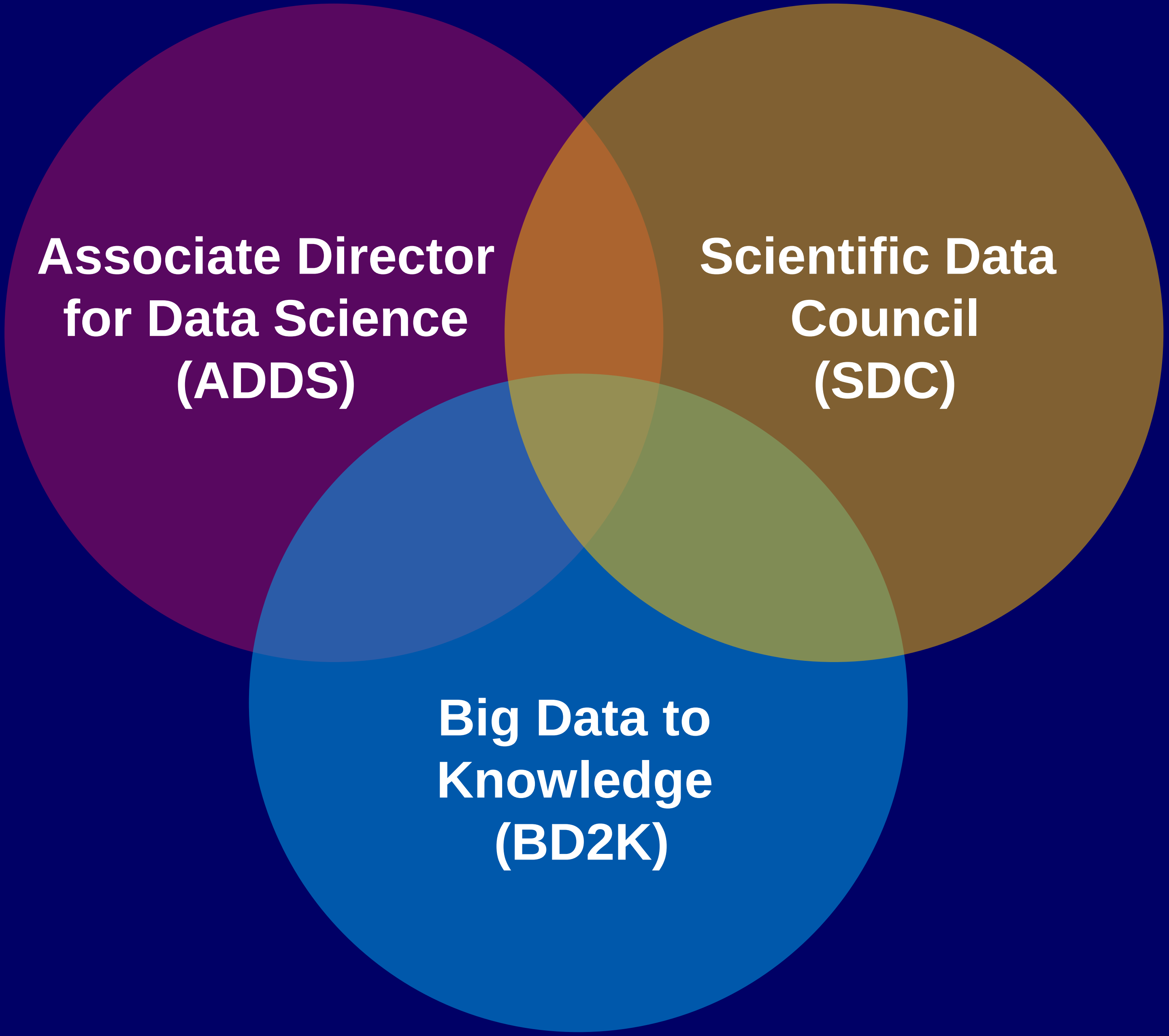
“A final key strategic challenge is to ensure that [the] NIH culture changes [are] commensurate with recognition of the key role of informatics and computation for every IC’s mission. Informatics and computation should not be championed by just a few ICs, based on the personal vision of particular leaders. Instead, NIH leadership must accept a distributed commitment to the use of advanced computation and informatics toward supporting the research portfolio of every IC.”

Data and Informatics Working Group
(June 2012 Report, p. 25)

Among the Major Problems to Solve…

1. Locating the data
2. Getting access to the data
3. Extending policies and practices for data sharing
4. Organizing, managing, and processing biomedical Big Data
5. Developing new methods for analyzing biomedical Big Data
6. Training researchers who can use biomedical Big Data effectively

NIH is Tackling the ‘Big Data’ Problem



Associate Director for Data Science (ADDS)

Scientific Data Council (SDC)

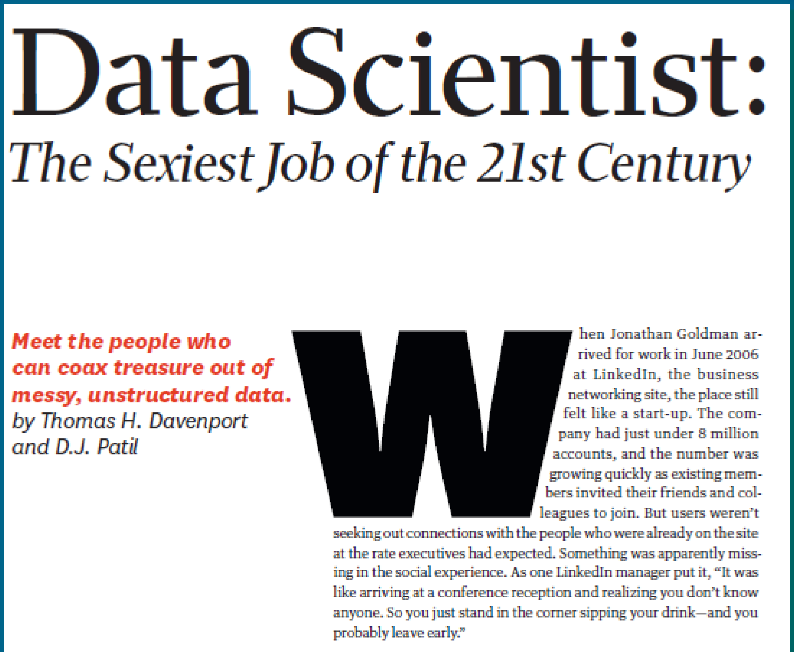
Big Data to Knowledge (BD2K)

What's in a Name?

- Big Data
- Bioinformatics
- Computational Biology
- Biomedical Informatics
- Information Science
- Biostatistics
- Quantitative Biology

Data Science

When in Doubt… Go with Sexy!



Data Scientist:

The Sexiest Job of the 21st Century

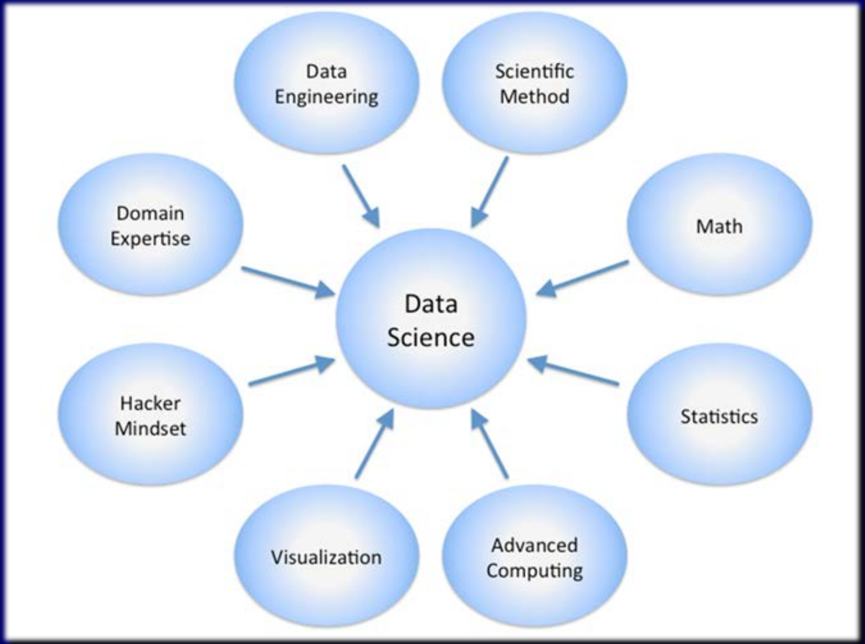
Meet the people who can coax treasure out of messy, unstructured data.

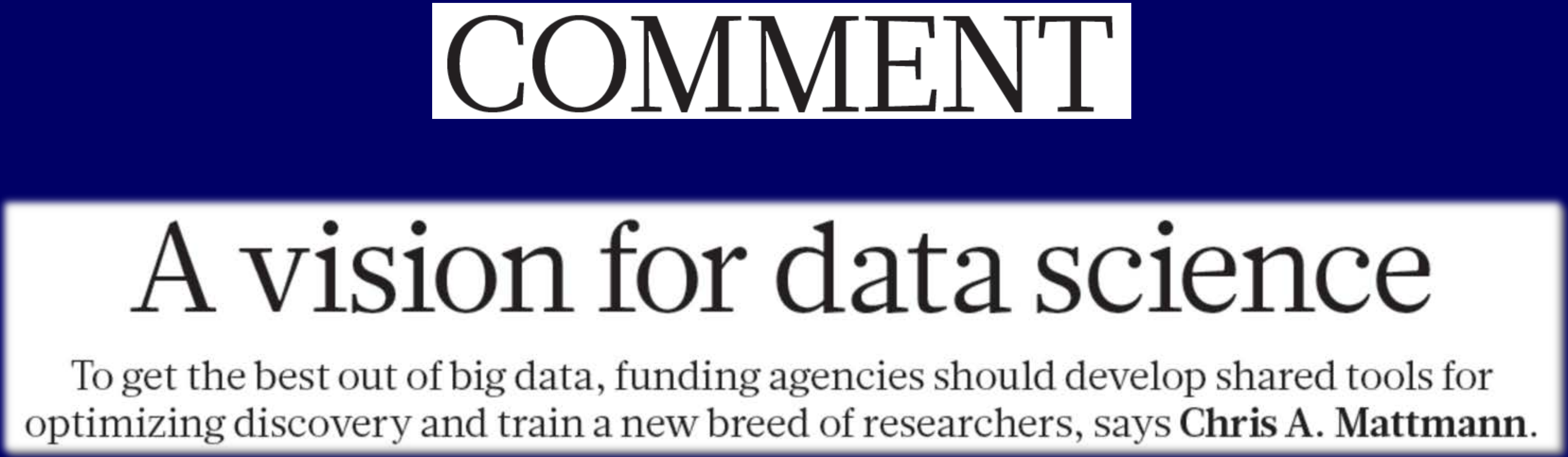
by Thomas H. Davenport and D.J. Patil

When Jonathan Goldman arrived for work in June 2006 at LinkedIn, the business networking site, the place still felt like a start-up. The company had just under 8 million accounts, and the number was growing quickly as existing members invited their friends and colleagues to join. But users weren't seeking out connections with the people who were already on the site at the rate executives had expected. Something was apparently missing in the social experience. As one LinkedIn manager put it, "It was like arriving at a conference reception and realizing you don't know anyone. So you just stand in the corner sipping your drink—and you probably leave early."

***Harvard Business Review* (2012)**

The shortage of data scientists is becoming a serious constraint in some sectors.

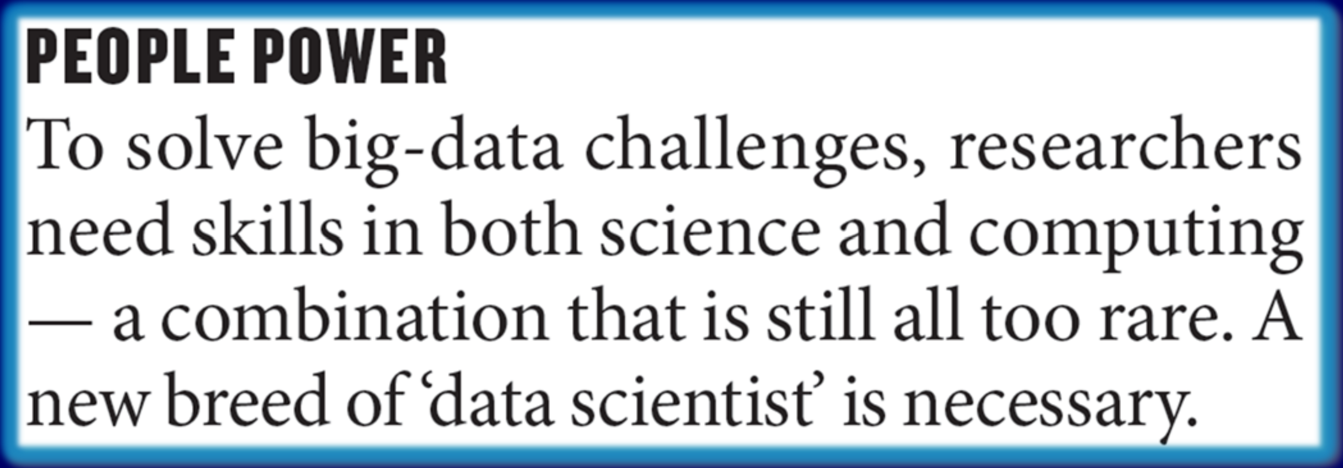




COMMENT

A vision for data science

To get the best out of big data, funding agencies should develop shared tools for optimizing discovery and train a new breed of researchers, says **Chris A. Mattmann**.



PEOPLE POWER

To solve big-data challenges, researchers need skills in both science and computing — a combination that is still all too rare. A new breed of 'data scientist' is necessary.

Nature **2013**

Associate Director for Data Science: Overview

- NIH Data Science ‘Programmatic Czar’ (aka, Point Person, Strategic Leader, etc.)
- Reports to NIH Director
- Eric Green, Acting
- Search underway (Eric Green & Jim Anderson, Co-Chairs of Search Committee)

Associate Director for Data Science
Office of the Director, National Institutes of Health, Department of Health and Human Services

The NIH is the center of medical and behavioral research for the Nation
----making essential medical discoveries that improve health and save lives.

Are you a top-level Scientific Researcher or Scientific Administrator seeking a career at the one of the preeminent biomedical research institutions in the Nation and the world? Are you at that point in your career where you're ready to "give back?" The position of Associate Director for Data Science (ADDS), Office of the Director (OD), National Institutes of Health (NIH), offers a unique and exciting opportunity to provide critical leadership for basic and translational research. The era of "Big Data" has arrived for the biomedical sciences. There is an urgent need and, with it, spectacular opportunities for NIH to enhance its programs in data science, such as those involving data emanating from different sources (e.g., genomics, imaging, and phenotypic information from electronic health records). The ADDS provides a vision for the utilization and extraction of knowledge from the data generated by, and relevant to, NIH research, and advises experts throughout the agency on a variety of complex, unique, and/or sensitive situations and issues in data science to ensure continual achievement of NIH's dynamic biomedical research mission.

We are looking for applicants with senior-level experience who have a commitment to excellence and the energy, enthusiasm, and innovative thinking necessary to lead a dynamic and diverse organization.

The successful candidate for this position will be appointed at a salary commensurate with his/her qualifications. Full Federal benefits will be provided including leave, health and life insurance, long-term care insurance, retirement, and savings plan (401k equivalent).

If you are ready for an exciting leadership opportunity, please see the detailed vacancy announcement at http://www.jobs.nih.gov (under Executive Careers). **Applications will be reviewed starting May 13, 2013, and will be accepted until the position is filled.**

THE NATIONAL INSTITUTES OF HEALTH AND THE DEPARTMENT OF HEALTH AND HUMAN SERVICES ARE EQUAL OPPORTUNITY EMPLOYERS

Scientific Data Council: Overview

- High-level internal NIH group providing programmatic leadership and coordination of data science activities
- Chaired by Associate Director for Data Science
- Trans-NIH representation

Scientific Data Council: Membership

Acting Chair: Eric Green (Acting ADDS & NHGRI)

Members:
James Anderson (DPCPSI)
Sally Rockey (OER)
Michael Gottesman (OIR)
Kathy Hudson (OD)
Amy Patterson (OSP)
Andrea Norris (CIT)
Judith Greenberg (NIGMS)
Betsy Humphreys (NLM)
Douglas Lowy (NCI)
John J. McGowan (NIAID)
Alan Koretsky (NINDS)
Michael Lauer (NHLBI)
Belinda Seto (NIBIB)

Acting Executive Secretary: Allison Mandich (NHGRI)

ADDS + SDC: Joint Responsibilities

- Oversight of Big Data to Knowledge (BD2K) initiative
- Trans-NIH intellectual and programmatic ‘hub’ for data science (coordination and convening functions)
- Coordination with data science activities beyond NIH (e.g., other government agencies, other funding agencies, and private sector)
- Long-term NIH strategic planning in data science
- Key role in data sharing policy development & oversight
- Coordination with ‘parallel’ administrative data efforts

Big Data to Knowledge (BD2K): Overview

- Major trans-NIH initiative addressing an NIH imperative and key roadblock
- Aims to be catalytic and synergistic
- Overarching goal:

By the end of this decade, enable a quantum leap in the ability of the biomedical research enterprise to maximize the value of the growing volume and complexity of biomedical data

BD2K: Four Programmatic Areas

I. Facilitating Broad Use of Biomedical Big Data

II. Developing and Disseminating Analysis Methods and Software for Biomedical Big Data

III. Enhancing Training for Biomedical Big Data

IV. Establishing Centers of Excellence for Biomedical Big Data

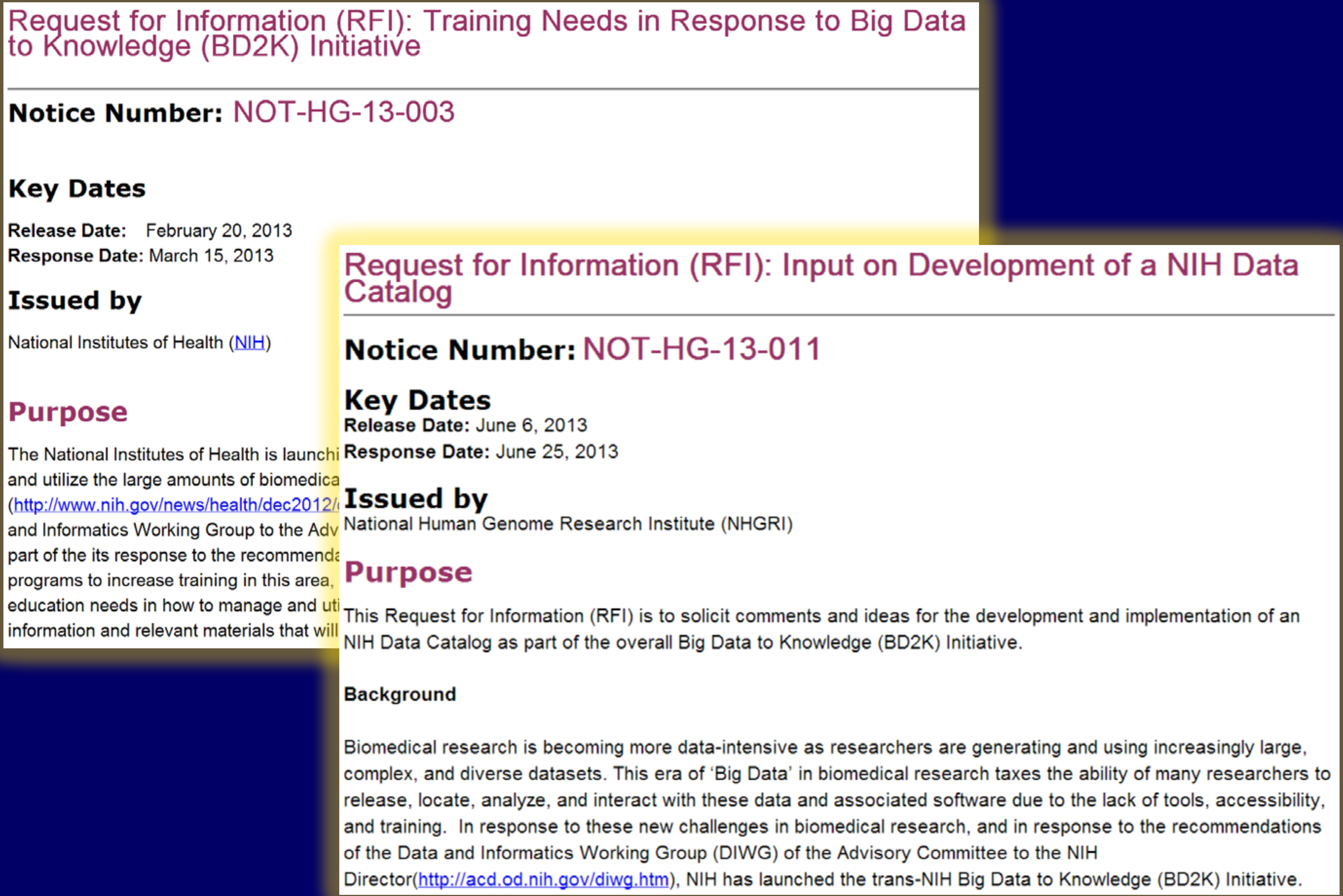
BD2K: Funding Plan

- Initial 7-year funding plan (thru FY2020)
- Begins in FY2014
- Ramps to slightly over $100M by FY2017
- Novel funding model:

  1. Early front-loading contributions by Common Fund
  2. Increasing Institutes/Centers' contributions

- Complete budgetary 'adoption' by Institutes/Centers by FY2020 to ensure sustainability

BD2K: Requests for Information (RFIs)



Request for Information (RFI): Training Needs in Response to Big Data to Knowledge (BD2K) Initiative

Notice Number: NOT-HG-13-003

Key Dates

Release Date: February 20, 2013
Response Date: March 15, 2013

Issued by

National Institutes of Health (NIH)

Purpose

The National Institutes of Health is launchi
and utilize the large amounts of biomedica
(http://www.nih.gov/news/health/dec2012/
and Informatics Working Group to the Adv
part of the its response to the recommenda
programs to increase training in this area,
education needs in how to manage and uti
information and relevant materials that will

Request for Information (RFI): Input on Development of a NIH Data Catalog

Notice Number: NOT-HG-13-011

Key Dates

Release Date: June 6, 2013
Response Date: June 25, 2013

Issued by

National Human Genome Research Institute (NHGRI)

Purpose

This Request for Information (RFI) is to solicit comments and ideas for the development and implementation of an NIH Data Catalog as part of the overall Big Data to Knowledge (BD2K) Initiative.

Background

Biomedical research is becoming more data-intensive as researchers are generating and using increasingly large, complex, and diverse datasets. This era of ‘Big Data’ in biomedical research taxes the ability of many researchers to release, locate, analyze, and interact with these data and associated software due to the lack of tools, accessibility, and training. In response to these new challenges in biomedical research, and in response to the recommendations of the Data and Informatics Working Group (DIWG) of the Advisory Committee to the NIH Director(http://acd.od.nih.gov/diwg.htm), NIH has launched the trans-NIH Big Data to Knowledge (BD2K) Initiative.

BD2K: Upcoming Workshops

Broad Use of Big Data:

Enabling Research Use of Clinical Data (9/13)

Frameworks for Data Standards (9/13)

Data Catalog (8/13)

Software:

Software Catalog (10/13)

Underserved Areas (TBD)

Platforms for Data Analysis (TBD)

Training:

Big Data and Training (7/13)

Centers:

Data Integration (10/13)

BD2K: Other Details

- **Strong support across NIH:**

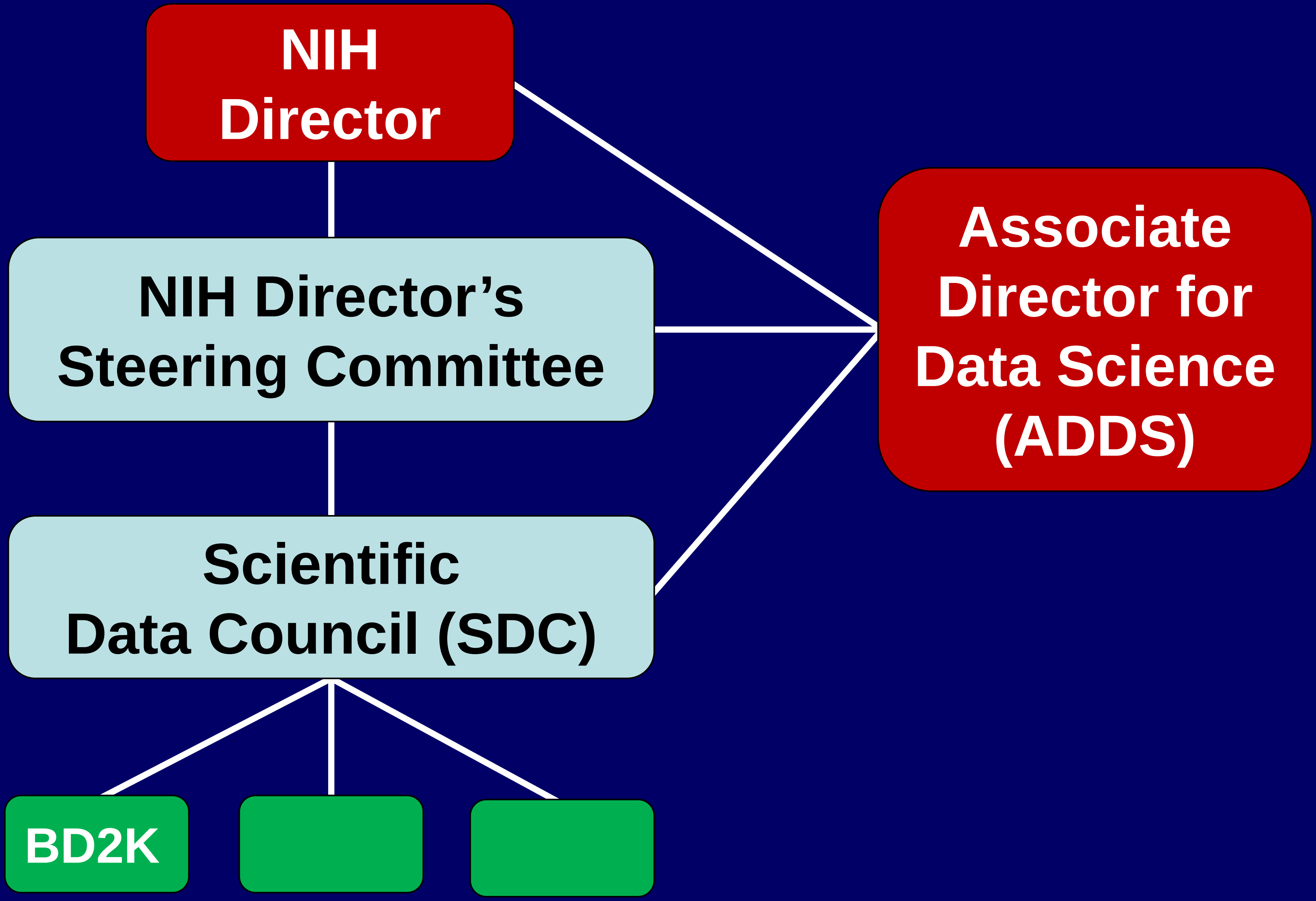
  Trans-NIH Working Group with ~125 members

  24 Institutes/Centers and several offices involved

- **Revised funding plan:**

	FY14	FY15	FY16
Original:	**$64M**	**$96M**	**$109M**
Revised:	**$27M**	**$80M**	**$99M**

ADDS, SDC, BD2K: Governance



NIH Director

NIH Director's Steering Committee

Scientific Data Council (SDC)

Associate Director for Data Science (ADDS)

BD2K

Closing Thoughts

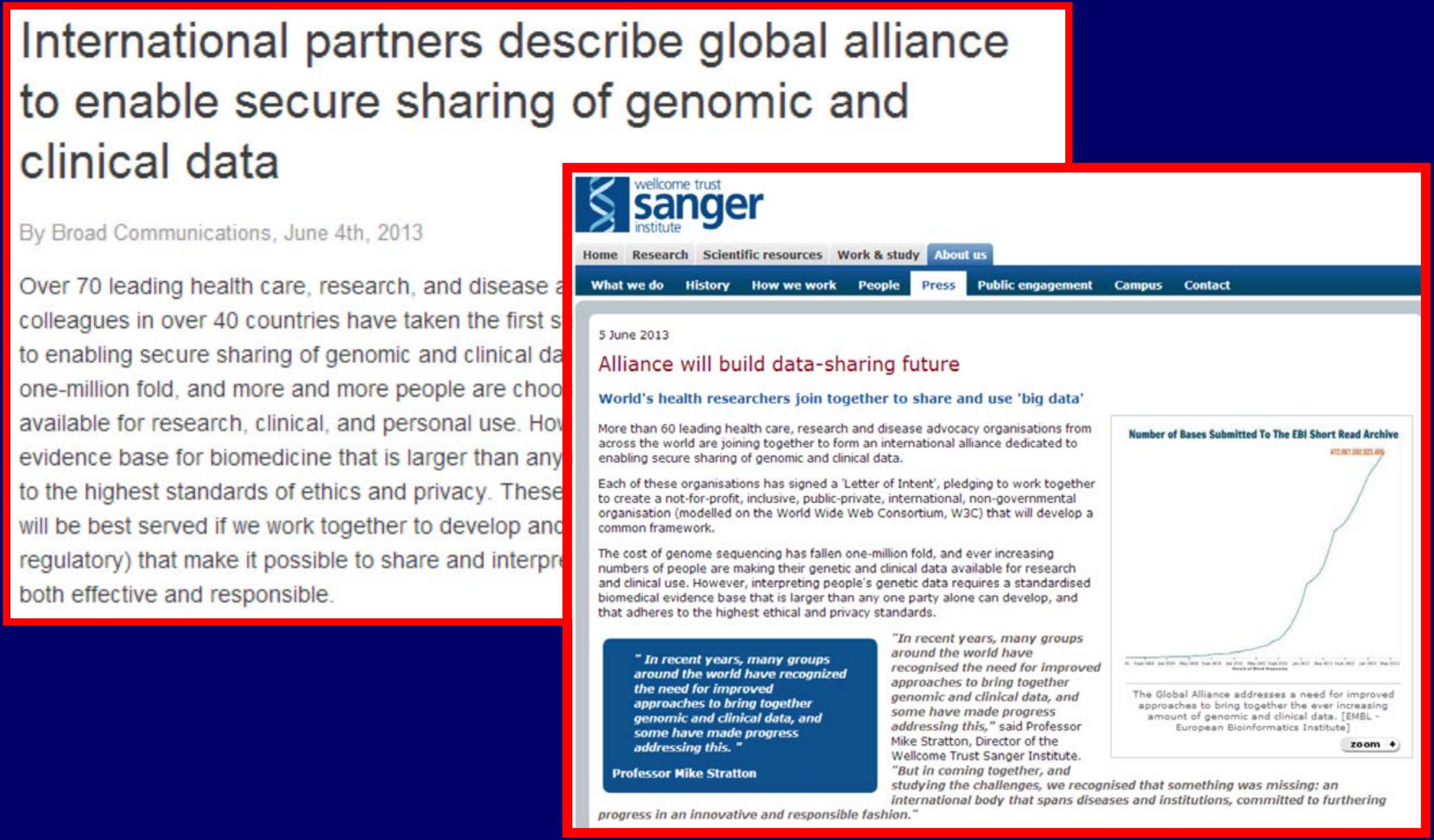
- The biomedical research enterprise is undergoing a major ‘phase change’ with respect to Big Data and data science
- Trans-NIH problem needing trans-NIH solutions
- Solutions include multifaceted cultural changes
- New NIH plans are:

Mission critical
Transformational
Transitional-- en route to longer-term commitment

Breaking
News

‘Global Alliance’ to Enable Responsible Sharing of Genomic and Clinical Data

Announced June 4, 2013



International partners describe global alliance to enable secure sharing of genomic and clinical data

By Broad Communications, June 4th, 2013

Over 70 leading health care, research, and disease a colleagues in over 40 countries have taken the first s to enabling secure sharing of genomic and clinical da one-million fold, and more and more people are choo available for research, clinical, and personal use. Ho evidence base for biomedicine that is larger than any to the highest standards of ethics and privacy. These will be best served if we work together to develop and regulatory) that make it possible to share and interpre both effective and responsible.

5 June 2013

Alliance will build data-sharing future

World's health researchers join together to share and use 'big data'

More than 60 leading health care, research and disease advocacy organisations from across the world are joining together to form an international alliance dedicated to enabling secure sharing of genomic and clinical data.

Each of these organisations has signed a 'Letter of Intent', pledging to work together to create a not-for-profit, inclusive, public-private, international, non-governmental organisation (modelled on the World Wide Web Consortium, W3C) that will develop a common framework.

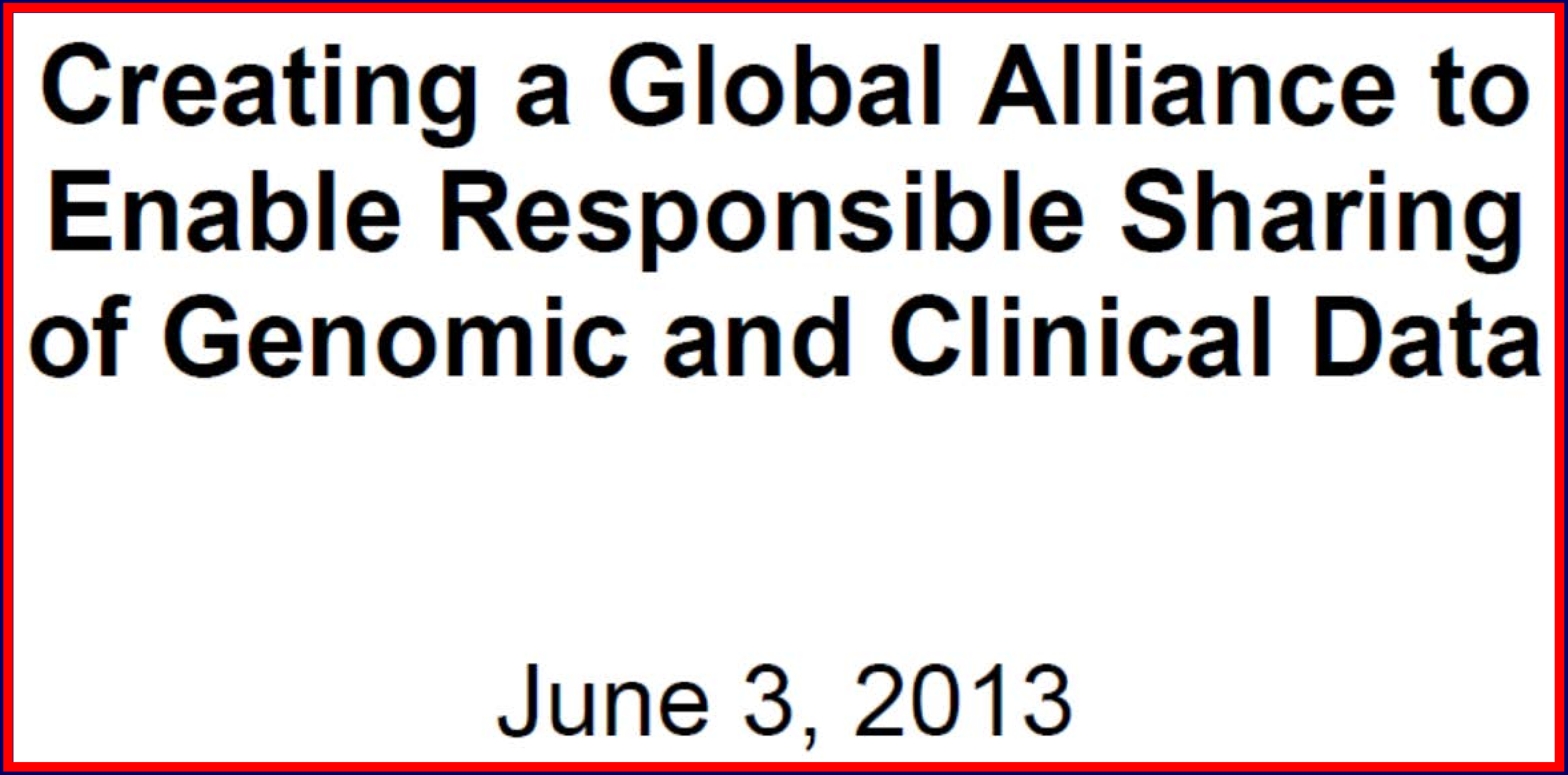
The cost of genome sequencing has fallen one-million fold, and ever increasing numbers of people are making their genetic and clinical data available for research and clinical use. However, interpreting people's genetic data requires a standardised biomedical evidence base that is larger than any one party alone can develop, and that adheres to the highest ethical and privacy standards.

> *" In recent years, many groups around the world have recognized the need for improved approaches to bring together genomic and clinical data, and some have made progress addressing this. "*
>
> **Professor Mike Stratton**

"In recent years, many groups around the world have recognised the need for improved approaches to bring together genomic and clinical data, and some have made progress addressing this," said Professor Mike Stratton, Director of the Wellcome Trust Sanger Institute. *"But in coming together, and studying the challenges, we recognised that something was missing: an international body that spans diseases and institutions, committed to furthering progress in an innovative and responsible fashion."*

The Global Alliance addresses a need for improved approaches to bring together the ever increasing amount of genomic and clinical data. [EMBL - European Bioinformatics Institute]

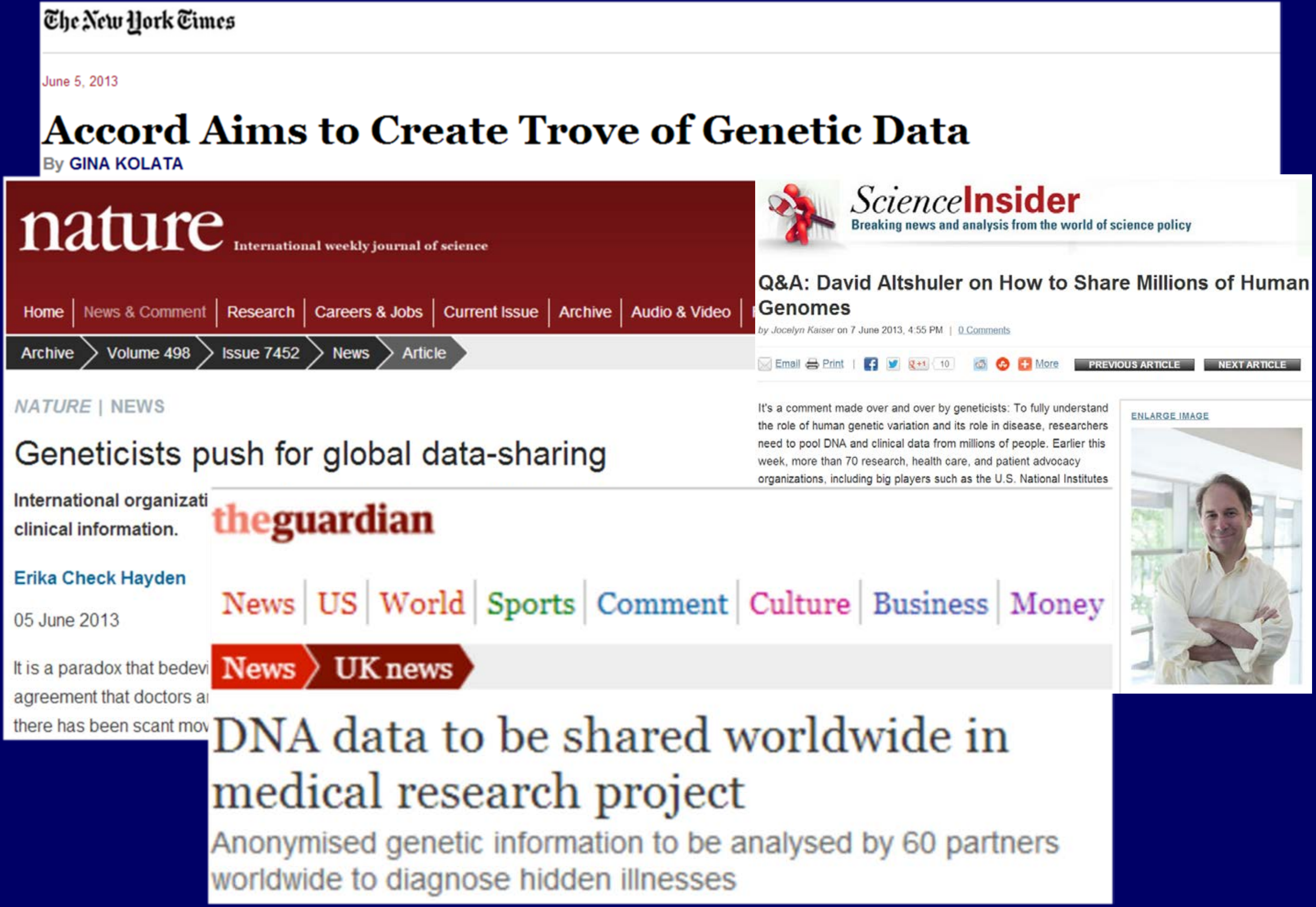
‘Global Alliance’: White Paper



Creating a Global Alliance to Enable Responsible Sharing of Genomic and Clinical Data

June 3, 2013

‘Global Alliance’: Press Coverage



The New York Times

June 5, 2013

Accord Aims to Create Trove of Genetic Data

By GINA KOLATA

nature International weekly journal of science

Home | News & Comment | Research | Careers & Jobs | Current Issue | Archive | Audio & Video | I

Archive > Volume 498 > Issue 7452 > News > Article

NATURE | NEWS

Geneticists push for global data-sharing

International organizati
clinical information.

Erika Check Hayden

05 June 2013

It is a paradox that bedevi
agreement that doctors ai
there has been scant mov

ScienceInsider Breaking news and analysis from the world of science policy

Q&A: David Altshuler on How to Share Millions of Human Genomes

by Jocelyn Kaiser on 7 June 2013, 4:55 PM | 0 Comments

Email Print | 10 More PREVIOUS ARTICLE NEXT ARTICLE

ENLARGE IMAGE

It's a comment made over and over by geneticists: To fully understand the role of human genetic variation and its role in disease, researchers need to pool DNA and clinical data from millions of people. Earlier this week, more than 70 research, health care, and patient advocacy organizations, including big players such as the U.S. National Institutes

theguardian

News | US | World | Sports | Comment | Culture | Business | Money

News > UK news

DNA data to be shared worldwide in medical research project

Anonymised genetic information to be analysed by 60 partners worldwide to diagnose hidden illnesses

The scientific opportunity

An explosion of information about the genome sequences of individuals with associated clinical characteristics and outcomes

Learning from extensive data on genome sequence with clinical annotation, it should be possible to accelerate progress in:

- Cancer outcomes and targeted therapy
- Inherited pediatric diseases
- Common diseases and drug responses
- Infectious diseases

Moreover, clinical interpretation of individual genome sequences will require a robust evidence base

Slides courtesy of David Altshuler

The challenge

Very large comparator data sets (millions) needed

Stakeholders not organized to seize the opportunity:

- Data in silos: by disease, institution, platform, method
- Regulation and consent: didn't anticipate need to share
- Informatics capabilities: non-standardized, few at scale

If we don't act: a hodge-podge of Balkanized systems
If we don't act: great uncertainty in privacy and ethics

Slides courtesy of David Altshuler

The process

Over the past several years, many groups have identified this set of issues, and organized meetings on related topics

- Some meetings focused on one disease (e.g., cancer)
- Some meetings limited to a single country (e.g., US)

On January 28, 2013 a meeting focusing on these topics was held in NYC, bringing together 50 participants from 8 countries, spanning disease areas, disciplines, and countries

The group wrote a White Paper and then invited organizations to sign a (non-binding) letter of intent to create a new alliance

Slides courtesy of David Altshuler

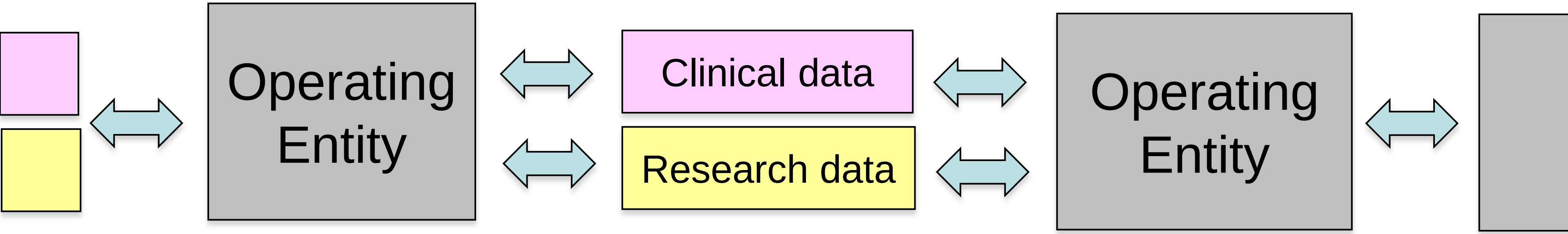
A vision for the ecosystem

Clinical risk assessment

Disease-specific portals

Innovative Apps for analysis

Transformative research projects



Operating Entity

Clinical data

Research data

Operating Entity

Global Alliance: technology standards, harmonization of ethics

Slides courtesy of David Altshuler

Core principles: global alliance

Respect – data sharing and privacy preferences of participants

Transparency – of governance and operations

Accountability – best practices in technology, ethics, and outreach

Inclusivity – partnering and building trust among stakeholders

Collaboration – sharing information to advance human health

Innovation – developing an ecosystem that accelerates progress

Agility – acting swiftly to benefit those suffering with disease

Slides courtesy of David Altshuler

Shared and open technical standards

To spark innovation in information platforms that are embody these core principles and will be interoperable and we need open technology standards for data sharing and analysis:

- Agnostic to (inclusive of) the specific platforms for sequence data generation, cloud providers, etc.
- Open so many parties can innovate, shared so that these innovations can speak to one another

An inspiration is the World Wide Web Consortium (**W3C**), which spurred innumerable and unanticipated applications

Slides courtesy of David Altshuler

Harmonization of ethics, privacy, consent

In forming an international partnership that brings together ethics, privacy, medicine, research, and technology under one tent, we aim to develop harmonized solutions that are both responsible and that can be implemented.

We reject a “one size fits all” approach, and rather look to a menu of options so that different parties have choice

The White Paper and Letter of Intent commit to a **founding principle** of respect for the data sharing choices of participants, including sharing broadly, or narrowly, or not at all.

The alliance won’t have any authority over stakeholders, but rather aims to lead by example and advocate for shared solutions

Slides courtesy of David Altshuler

Signatories to the Letter of Intent
73 institutions active in 40 countries

American Association for Cancer Research (US)
American Association of Clinical Oncology (US)
American Society of Human Genetics (US)
BGI –Shenzhen (China)
Boston Children's Hospital (US)
Brigham and Women's Hospital (US)
Broad Institute of MIT and Harvard (US)
Canadian Cancer Society (Canada)
Cancer Research UK (UK)
Centre for Genomic Regulation (Spain)
Chinese Academy of Sciences (China)
Dana Farber Cancer Institute (US)
European Bioinformatics Institute (UK)
European Molecular Biology Laboratory (Germany)
European Society of Human Genetics (Europe)
Genetic Alliance (US and also UK)
Genome Canada (Canada)
Global Genes/RARE Project (US)
Howard Hughes Medical Institute (US)
Human Variome Project (Australia)
H3BioNet Africa (Africa)
Institut National du Cancer (France)
Institute of Health and Welfare (Finland)
International Cancer Genome Consortium
Johns Hopkins University School of Medicine (US)
Lund University (Sweden)
MD Anderson Cancer Center (US)

Massachusetts General Hospital (US)
Memorial Sloan Kettering Cancer Center (US)
McGill University (Canada)
Mount Sinai Hospital (Canada)
National Cancer Institute (US)
National Cancer Center (Japan)
National Human Genome Research Institute (US)
National Institute for Health Research (UK)
National Institutes of Health (US)
New York Genome Center (US)
Ontario Institute for Cancer Research (Canada)
Partners HeathCare (US)
P3G (International based in Canada)
Queens University, Belfast (U.K.)
Sage Bionetworks (US)
Simons Foundation (US)
Stanford University (US)
St. Jude Children's Research Hospital (US)
The Hospital for Sick Children (Canada)
University Health Network (Canada)
University of California at San Francisco (US)
University of California at Santa Cruz (US)
University of Cape Town (South Africa)
University of Chicago (US)
University of Oxford (UK)
Wellcome Trust Sanger Institute (UK)
Wellcome Trust (UK)

Signatories to the Letter of Intent: funders

American Association for Cancer Research (US)
American Association of Clinical Oncology (US)
American Society of Human Genetics (US)
BGI –Shenzhen (China)
Boston Children’s Hospital (US)
Brigham and Women's Hospital (US)
Broad Institute of MIT and Harvard (US)
Canadian Cancer Society (Canada)
Cancer Research UK (UK)
Centre for Genomic Regulation (Spain)
Chinese Academy of Sciences (China)
Dana Farber Cancer Institute (US)
European Bioinformatics Institute (UK)
European Molecular Biology Laboratory (Germany)
European Society of Human Genetics (Europe)
Genetic Alliance (US and also UK)
Genome Canada (Canada)
Global Genes/RARE Project (US)
Howard Hughes Medical Institute (US)
Human Variome Project (Australia)
H3BioNet Africa (Africa)
Institut National du Cancer (France)
Institute of Health and Welfare (Finland)
International Cancer Genome Consortium
Johns Hopkins University School of Medicine (US)
Lund University (Sweden)
MD Anderson Cancer Center (US)

Massachusetts General Hospital (US)
Memorial Sloan Kettering Cancer Center (US)
McGill University (Canada)
Mount Sinai Hospital (Canada)
National Cancer Institute (US)
National Cancer Center (Japan)
National Human Genome Research Institute (US)
National Institute for Health Research (UK)
National Institutes of Health (US)
New York Genome Center (US)
Ontario Institute for Cancer Research (Canada)
Partners HeathCare (US)
P3G (International based in Canada)
Queens University, Belfast (U.K.)
Sage Bionetworks (US)
Simons Foundation (US)
Stanford University (US)
St. Jude Children's Research Hospital (US)
The Hospital for Sick Children (Canada)
University Health Network (Canada)
University of California at San Francisco (US)
University of California at Santa Cruz (US)
University of Cape Town (South Africa)
University of Chicago (US)
University of Oxford (UK)
Wellcome Trust Sanger Institute (UK)
Wellcome Trust (UK)

There is a tremendous amount to do

Establish the global alliance as an organization

- governance
- funding
- structure
- membership (nonprofit and for-profit)

Working groups

- technical (genomic data, security, interoperability)
- ethics (consent, privacy, patient centric initiatives)
- clinical data
- outreach and communication

Establish operating entities and start pilot projects

Slides courtesy of David Altshuler

‘Global Alliance’: Summary

- International alliance that will enable secure sharing of genomic and clinical data by:
  - Establishing inter-operable standards for genomic and clinical data (initially)
  - Develop framework for harmonizing data-sharing practices to address issues related to ethics, privacy, and consent
- Signatories of Letter of Intent include 73 institutions in 40 countries (13 funding agencies)
- Just getting off the ground, with much to be done
- Aims to tackle several of the major problems that NIH identified and that are components of BD2K

Questions?

Ben Chams – Fotolia