

NIH



# NIH and Biomedical 'Big Data'

---

**Eric Green, M.D., Ph.D.**  
**Director, NHGRI**

**Acting Associate Director for Data Science, NIH**

**TECHNOLOGY FEATURE**

# THE BIG CHALLENGES OF BIG DATA

---

*As they grapple with increasingly large data sets,  
biologists and computer scientists uncork new bottlenecks.*

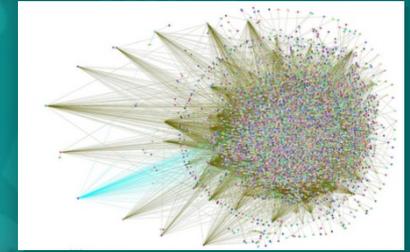
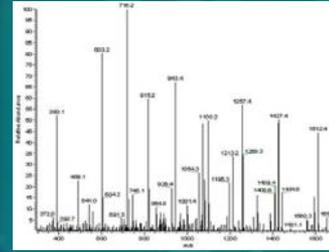
---

***Nature* 2013**

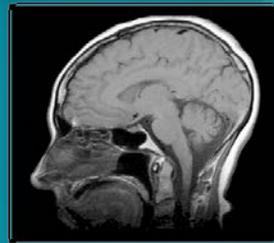
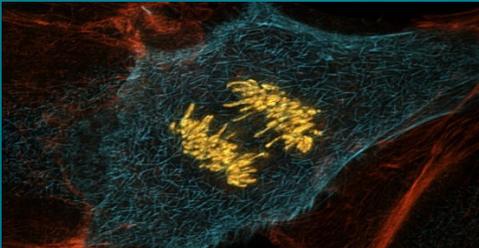
# Myriad Data Types



**Genomic**



**Other 'Omic**



**Imaging**



**Phenotypic**



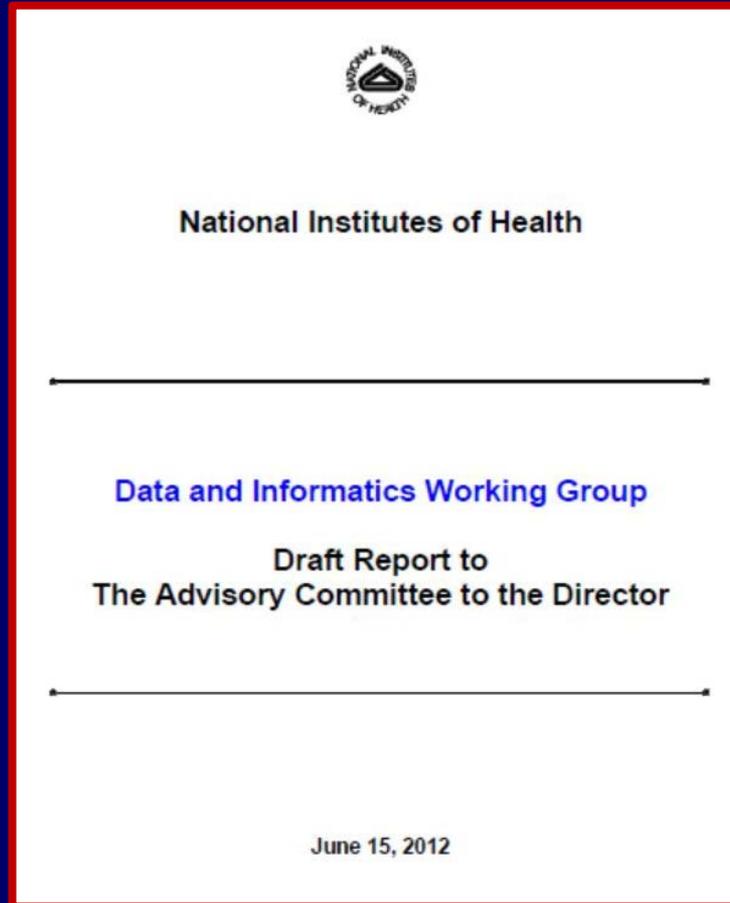
**Exposure**



**Clinical**

# Data and Informatics Working Group

ADVISORY COMMITTEE TO THE DIRECTOR



[acd.od.nih.gov/diwig.htm](http://acd.od.nih.gov/diwig.htm)

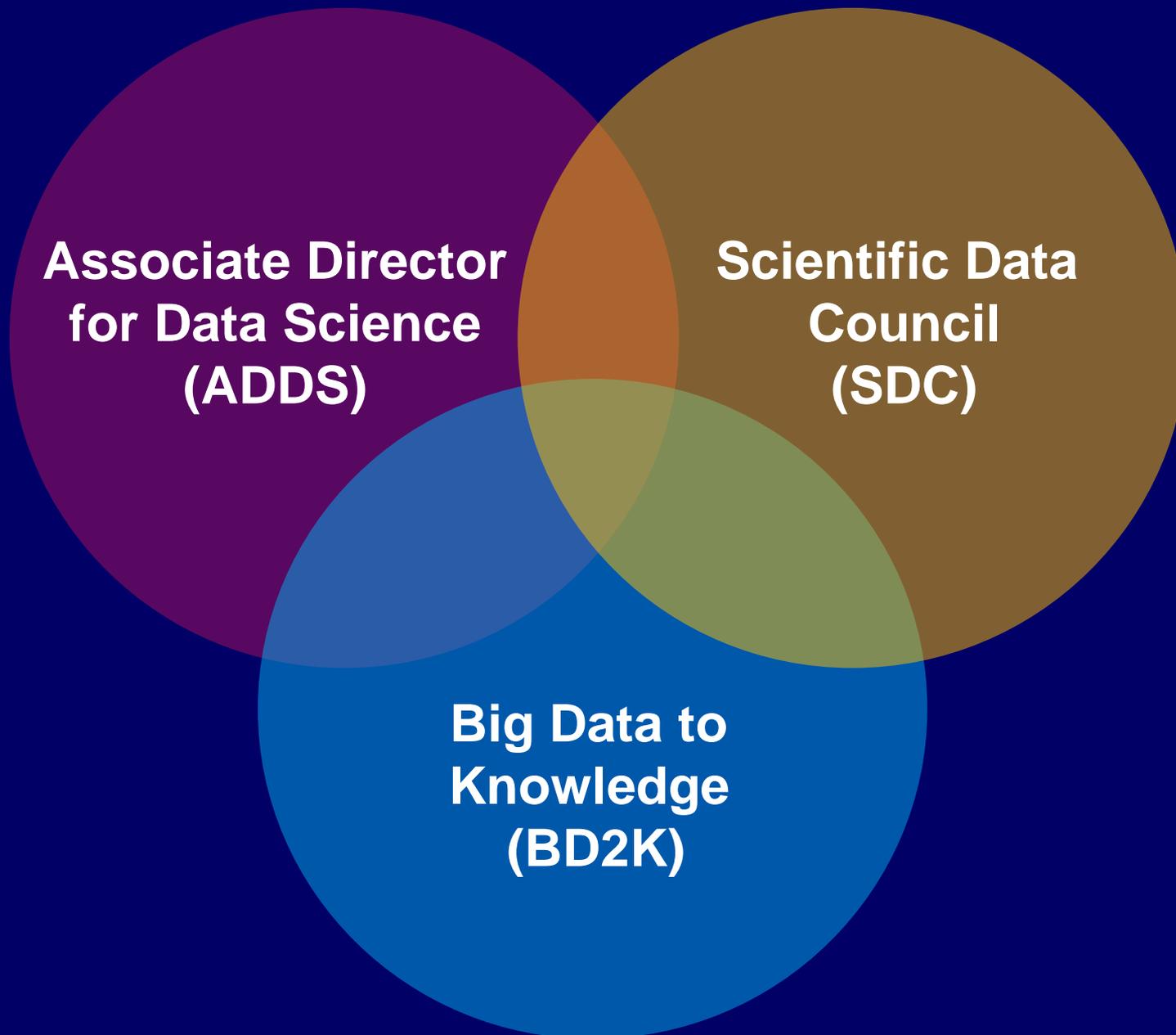
# Overarching Themes

- **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**

# **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



# 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!

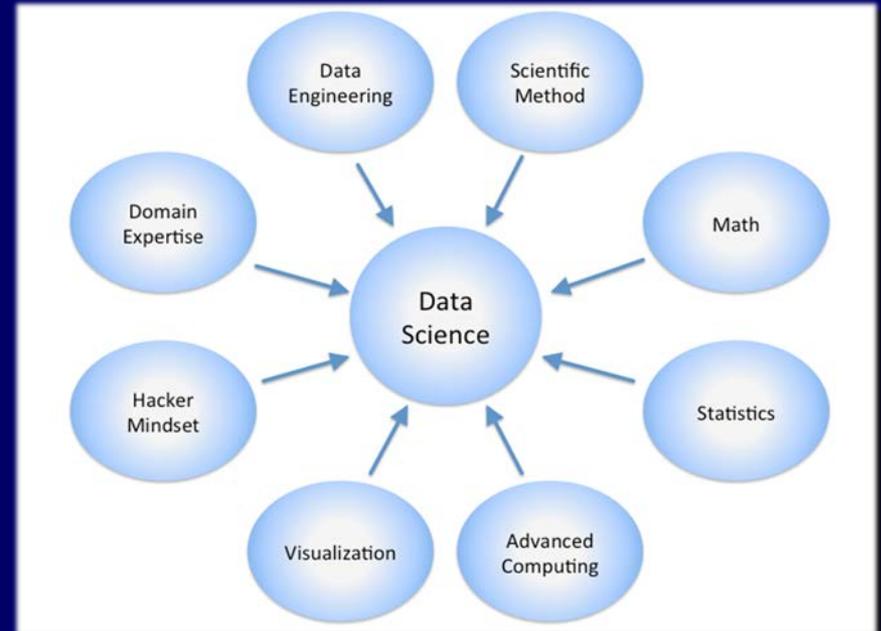
## Data Scientist: *The Sexiest Job of the 21st Century*

Meet the people who can coax treasure out of messy, unstructured data.  
by Thomas H. Davenport and D.J. Patil

**W**hen 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)**

# 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**

# 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

# 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)  
Jon Lorsch (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)

# 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: Other Details



- **Strong support across NIH:**

  - Trans-NIH Working Group with ~125 members

  - 24 Institutes/Centers and several offices involved

- **Revised funding plan:**

	<u>FY14</u>	<u>FY15</u>	<u>FY16</u>
Original:	\$64M	\$96M	\$109M
Revised:	\$27M	\$80M	\$99M

# BD2K: Upcoming Workshops

## Broad Use of Big Data:

Data Catalog (8/21)

Enabling Research Use of Clinical Data (9/13)

Frameworks for Data Standards (9/13)



## Software:

Software Catalog (2/14)

Underserved Areas (TBD)

Platforms for Data Analysis (TBD)



## Training:

Big Data and Training (7/13)



## Centers:

Data Integration (10/13)



# 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 1, 2013

Response Date: May 1, 2013

## Issued by

National Institutes of Health

## Purpose

The National Institutes of Health (NIH) will continue to develop and utilize the large amount of data generated by its research and informatics programs as part of the overall Big Data to Knowledge (BD2K) Initiative. We are soliciting input on the part of the its research programs to increase the effectiveness of its education needs in information and research.

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

**Notice Number:** NOT-HG-13-011

## Key Dates

Release Date: June 11, 2013

Response Date: July 11, 2013

## Issued by

National Human Genome Research Institute

## Purpose

This Request for Information (RFI) is to solicit comments and ideas for the development of analysis methods and software tools, as part of the overall Big Data to Knowledge (BD2K) Initiative. Specifically, this RFI solicits input on needs for software and analysis methods related to data compression/reduction, data visualization, data provenance, and data wrangling.

## 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>).

Request for Information (RFI): Input on Development of Analysis Methods and Software for Big Data

**Notice Number:** NOT-HG-13-014

## Key Dates

Release Date: August 8, 2013

Response Due Date: September 6, 2013

## Issued by

National Human Genome Research Institute ([NHGRI](http://www.nhgri.nih.gov))

## Purpose

This Request for Information (RFI) is to solicit comments and ideas for the development of analysis methods and software tools, as part of the overall Big Data to Knowledge (BD2K) Initiative. Specifically, this RFI solicits input on needs for software and analysis methods related to data compression/reduction, data visualization, data provenance, and data wrangling.

## 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>).

# BD2K: First Funding Opportunity Announcement (FOA) Released

<b>Funding Opportunity Title</b>	<b>Centers of Excellence for Big Data Computing in the Biomedical Sciences (U54)</b>
<b>Activity Code</b>	<a href="#">U54</a> Specialized Center- Cooperative Agreements
<b>Announcement Type</b>	New
<b>Related Notices</b>	None
<b>Funding Opportunity Announcement (FOA) Number</b>	<b>RFA-HG-13-009</b>
<b>Companion Funding Opportunity</b>	None
<b>Number of Applications</b>	See <a href="#">Section III. 3. Additional Information on Eligibility</a> .
<b>Catalog of Federal Domestic Assistance (CFDA) Number(s)</b>	93.172; 93.839; 93.866; 93.273; 93.855; 93.856; 93.846; 93.286; 93.865; 93.279; 93.173; 93.121; 93.847; 93.113; 93.859; 93.242; 93.307; 93.853; 93.361; 93.879; 93.351; 93.350; 93.213; 93.393; 93.394; 93.395; 93.396; 93.397; 93.399; 93.867; 93.233; 93.837; 93.838
<b>Funding Opportunity Purpose</b>	Biomedical research is becoming more data-intensive as researchers are generating and using increasingly large, complex, and diverse data sets. This era of "Big Data" taxes the ability of biomedical researchers to locate, analyze, and interact with these data (and more generally all biomedical data) and associated software due to the lack of tools, accessibility, and training. In response to these new challenges in biomedical research, NIH has developed the Big Data to Knowledge (BD2K) Initiative. Under this FOA, BD2K Centers of Excellence are sought to conduct research to advance the science and utility of Big Data in the context of biomedical and behavioral research, and to create innovative new approaches, methods, software, tools, and related resources. The Centers will advance the ability of the biomedical research enterprise to use Big Data by producing tools and resources from early-stage to mature development that will be broadly useful to the research community.

# Big Data to Knowledge (BD2K): Update



**FOA published for Investigator-Initiated  
Centers of Excellence (U54)**

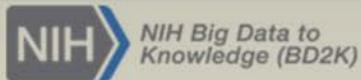
**Applicant Information Webinar: September 12**

**Questions to [BD2KCenterRFA@mail.nih.gov](mailto:BD2KCenterRFA@mail.nih.gov)**

# Capitalizing on Clinical Data



# Capitalizing on Clinical Data



## ENABLING RESEARCH USE OF CLINICAL DATA WORKSHOP

SEPTEMBER 11 - 12, 2013 | FISHERS LANE, ROCKVILLE, MD

[Home](#)

[Agenda](#)

[Biographies](#)

[Relevant Initiatives](#)

[Relevant Publications](#)

[NIH BD2K](#)

# BD2K: Web Site Now Live



NIH Big Data to Knowledge (BD2K)

Advancing Health and Discovery through Big Data



FUNDING OPPORTUNITIES & NOTICES

WORKSHOPS

NEWS

ABOUT BD2K



The NIH Big Data to Knowledge (BD2K) announces funding opportunity for

## CENTERS OF EXCELLENCE FOR BIG DATA COMPUTING IN THE BIOMEDICAL SCIENCES

[LEARN MORE](#)

### WORKSHOPS



**Workshop on Enhancing Training for Biomedical Big Data**  
July 29 - 30, 2013



**NIH Data Catalog**  
August 21 - 22, 2013



**Enabling Research Use of Clinical Data**  
September 11, 2013

[More Workshops >](#)

### NEWS HIGHLIGHT

- **NIH to recruit Associate Director for Data Science**  
January 10, 2013
- **NIH proposes critical initiatives to sustain future of U.S. biomedical research**  
December 7, 2012
- **Following Up On ACD Recommendations, and Paving the Road to Continued, Future Success**  
December 7, 2012

[More News >](#)

The mission of the **NIH Big Data to Knowledge (BD2K)** initiative is to enable biomedical scientists to capitalize more fully on the Big Data being generated by those research communities. With advances in technologies, these investigators are increasingly generating and using large, complex, and diverse datasets. Consequently, the biomedical research enterprise is increasingly becoming data-intensive and data-driven. However, the ability of researchers to locate, analyze, and use Big Data (and more generally all biomedical and behavioral data) is often limited for reasons related to access to relevant software and tools, expertise, and other factors. BD2K aims to develop the new approaches, standards, methods, tools, software, and competencies that will enhance the use of biomedical Big Data by supporting research, implementation, and training in data science and other relevant fields that will lead to: [Read more](#)

[bd2k.nih.gov](http://bd2k.nih.gov)

# 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**

