NIH and Biomedical ‘Big Data’

Eric Green, M.D., Ph.D.
Director, NHGRI
Acting Associate Director for Data Science, NIH
Myriad Data Types

Genomic

Imaging

Exposure

Other ‘Omic

Phenotypic

Clinical
Data and Informatics Working Group

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
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
“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

The shortage of data scientists is becoming a serious constraint in some sectors.
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)
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](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 launching the BD2K Initiative to capitalize on the large amounts of biomedical data being generated and to utilize the large amounts of biomedical data being generated and to utilize the large amounts of biomedical data being generated and to utilize the large amounts of biomedical data being generated. The BD2K Initiative is focused on training and educating the workforce to meet the needs of the BD2K Initiative. This training and education will be critical for the success of the BD2K Initiative.

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:

<table>
<thead>
<tr>
<th></th>
<th>FY14</th>
<th>FY15</th>
<th>FY16</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original:</td>
<td>$64M</td>
<td>$96M</td>
<td>$109M</td>
</tr>
<tr>
<td>Revised:</td>
<td>$27M</td>
<td>$80M</td>
<td>$99M</td>
</tr>
</tbody>
</table>
ADDS, SDC, BD2K: Governance

NIH Director

NIH Director’s Steering Committee

Scientific Data Council (SDC)

Associate Director for Data Science (ADDS)

BD2K
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 advocacy organizations from across the world have joined together to form an international alliance dedicated to enabling secure sharing of genomic and clinical data. Over 500 million one-millionfold, and more and more people are choosing to make available for research, clinical, and personal use. However, the evidence base for biomedicine that is larger than any other is the highest standards of ethics and privacy. These are best served if we work together to develop and share data that makes it possible to share and interpret both effective and responsible.

Alliance will build data-sharing future

World’s health researchers join together to share and use “big data”

More than 50 leading health care, research, and disease advocacy organizations from across the world have joined together to form an international alliance dedicated to enabling secure sharing of genomic and clinical data. Each of these organizations has signed a ‘Letter of Intent’, pledging to work together to create a not-for-profit, inclusive, public-private, international, non-governmental organization (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 standardized biobank 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 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."

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

June 3, 2013
Accord Aims to Create Trove of Genetic Data

By GINA KOLATA

June 5, 2013

Geneticists push for global data-sharing

Erika Check Hayden
05 June 2013

It is a paradox that bedevils agreement that doctors are there has been scant mov...
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

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
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
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.
<table>
<thead>
<tr>
<th>Signatories to the Letter of Intent</th>
<th>73 institutions active in 40 countries</th>
</tr>
</thead>
<tbody>
<tr>
<td>American Association for Cancer Research (US)</td>
<td>Massachusetts General Hospital (US)</td>
</tr>
<tr>
<td>American Association of Clinical Oncology (US)</td>
<td>Memorial Sloan Kettering Cancer Center (US)</td>
</tr>
<tr>
<td>American Society of Human Genetics (US)</td>
<td>McGill University (Canada)</td>
</tr>
<tr>
<td>BGI –Shenzhen (China)</td>
<td>Mount Sinai Hospital (Canada)</td>
</tr>
<tr>
<td>Boston Children’s Hospital (US)</td>
<td>National Cancer Institute (US)</td>
</tr>
<tr>
<td>Brigham and Women’s Hospital (US)</td>
<td>National Cancer Center (Japan)</td>
</tr>
<tr>
<td>Broad Institute of MIT and Harvard (US)</td>
<td>National Human Genome Research Institute (US)</td>
</tr>
<tr>
<td>Canadian Cancer Society (Canada)</td>
<td>National Institute for Health Research (UK)</td>
</tr>
<tr>
<td>Cancer Research UK (UK)</td>
<td>National Institutes of Health (US)</td>
</tr>
<tr>
<td>Centre for Genomic Regulation (Spain)</td>
<td>New York Genome Center (US)</td>
</tr>
<tr>
<td>Chinese Academy of Sciences (China)</td>
<td>Ontario Institute for Cancer Research (Canada)</td>
</tr>
<tr>
<td>Dana Farber Cancer Institute (US)</td>
<td>Partners HealthCare (US)</td>
</tr>
<tr>
<td>European Bioinformatics Institute (UK)</td>
<td>P3G (International based in Canada)</td>
</tr>
<tr>
<td>European Molecular Biology Laboratory (Germany)</td>
<td>Queens University, Belfast (U.K.)</td>
</tr>
<tr>
<td>European Society of Human Genetics (Europe)</td>
<td>Sage Bionetworks (US)</td>
</tr>
<tr>
<td>Genetic Alliance (US and also UK)</td>
<td>Simons Foundation (US)</td>
</tr>
<tr>
<td>Genome Canada (Canada)</td>
<td>Stanford University (US)</td>
</tr>
<tr>
<td>Global Genes/RARE Project (US)</td>
<td>St. Jude Children’s Research Hospital (US)</td>
</tr>
<tr>
<td>Howard Hughes Medical Institute (US)</td>
<td>The Hospital for Sick Children (Canada)</td>
</tr>
<tr>
<td>Human Variome Project (Australia)</td>
<td>University Health Network (Canada)</td>
</tr>
<tr>
<td>H3BioNet Africa (Africa)</td>
<td>University of California at San Francisco (US)</td>
</tr>
<tr>
<td>Institut National du Cancer (France)</td>
<td>University of California at Santa Cruz (US)</td>
</tr>
<tr>
<td>Institute of Health and Welfare (Finland)</td>
<td>University of Cape Town (South Africa)</td>
</tr>
<tr>
<td>International Cancer Genome Consortium</td>
<td>University of Chicago (US)</td>
</tr>
<tr>
<td>Johns Hopkins University School of Medicine (US)</td>
<td>University of Oxford (UK)</td>
</tr>
<tr>
<td>Lund University (Sweden)</td>
<td>Wellcome Trust Sanger Institute (UK)</td>
</tr>
<tr>
<td>MD Anderson Cancer Center (US)</td>
<td>Wellcome Trust (UK)</td>
</tr>
<tr>
<td>Signatories to the Letter of Intent: funders</td>
<td></td>
</tr>
<tr>
<td>-------------------------------------------</td>
<td></td>
</tr>
<tr>
<td><strong>American Association for Cancer Research (US)</strong></td>
<td></td>
</tr>
<tr>
<td><strong>American Association of Clinical Oncology (US)</strong></td>
<td></td>
</tr>
<tr>
<td><strong>American Society of Human Genetics (US)</strong></td>
<td></td>
</tr>
<tr>
<td><strong>BGI –Shenzhen (China)</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Boston Children’s Hospital (US)</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Brigham and Women's Hospital (US)</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Broad Institute of MIT and Harvard (US)</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Canadian Cancer Society (Canada)</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Cancer Research UK (UK)</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Centre for Genomic Regulation (Spain)</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Chinese Academy of Sciences (China)</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Dana Farber Cancer Institute (US)</strong></td>
<td></td>
</tr>
<tr>
<td><strong>European Bioinformatics Institute (UK)</strong></td>
<td></td>
</tr>
<tr>
<td><strong>European Molecular Biology Laboratory (Germany)</strong></td>
<td></td>
</tr>
<tr>
<td><strong>European Society of Human Genetics (Europe)</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Genetic Alliance (US and also UK)</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Genome Canada (Canada)</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Global Genes/RARE Project (US)</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Howard Hughes Medical Institute (US)</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Human Variome Project (Australia)</strong></td>
<td></td>
</tr>
<tr>
<td><strong>H3BioNet Africa (Africa)</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Institut National du Cancer (France)</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Institute of Health and Welfare (Finland)</strong></td>
<td></td>
</tr>
<tr>
<td><strong>International Cancer Genome Consortium</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Johns Hopkins University School of Medicine (US)</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Lund University (Sweden)</strong></td>
<td></td>
</tr>
<tr>
<td><strong>MD Anderson Cancer Center (US)</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Massachusetts General Hospital (US)</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Memorial Sloan Kettering Cancer Center (US)</strong></td>
<td></td>
</tr>
<tr>
<td><strong>McGill University (Canada)</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Mount Sinai Hospital (Canada)</strong></td>
<td></td>
</tr>
<tr>
<td><strong>National Cancer Institute (US)</strong></td>
<td></td>
</tr>
<tr>
<td><strong>National Cancer Center (Japan)</strong></td>
<td></td>
</tr>
<tr>
<td><strong>National Human Genome Research Institute (US)</strong></td>
<td></td>
</tr>
<tr>
<td><strong>National Institute for Health Research (UK)</strong></td>
<td></td>
</tr>
<tr>
<td><strong>National Institutes of Health (US)</strong></td>
<td></td>
</tr>
<tr>
<td><strong>New York Genome Center (US)</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Ontario Institute for Cancer Research (Canada)</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Partners HealthCare (US)</strong></td>
<td></td>
</tr>
<tr>
<td><strong>P3G (International based in Canada)</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Queens University, Belfast (U.K.)</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Sage Bionetworks (US)</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Simons Foundation (US)</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Stanford University (US)</strong></td>
<td></td>
</tr>
<tr>
<td><strong>St. Jude Children’s Research Hospital (US)</strong></td>
<td></td>
</tr>
<tr>
<td><strong>The Hospital for Sick Children (Canada)</strong></td>
<td></td>
</tr>
<tr>
<td><strong>University Health Network (Canada)</strong></td>
<td></td>
</tr>
<tr>
<td><strong>University of California at San Francisco (US)</strong></td>
<td></td>
</tr>
<tr>
<td><strong>University of California at Santa Cruz (US)</strong></td>
<td></td>
</tr>
<tr>
<td><strong>University of Cape Town (South Africa)</strong></td>
<td></td>
</tr>
<tr>
<td><strong>University of Chicago (US)</strong></td>
<td></td>
</tr>
<tr>
<td><strong>University of Oxford (UK)</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Wellcome Trust Sanger Institute (UK)</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Wellcome Trust (UK)</strong></td>
<td></td>
</tr>
</tbody>
</table>
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?