BD2K Update

Philip Bourne, PhD, FACMI
Associate Director for Data Science

▶ BD2K Multi-Council Working Group ▶ January 11, 2016

National Institutes of Health
Office of the Director
Data Science at NIH

This is a public summary of the major highlights arising from the BD2K initiative in 2015 and reported to the BD2K Multi-Council Working Group (MCWG)

National Institutes of Health
Office of the Director
Data Science at NIH

BD2K is Implementing the ACD Data & Informatics Recommendations*

DIWG Recommendations

1. Sharing data & software through indexes
2. Advance big methods, tools & applications
3. Expand data science training
4. Continued support throughout the data & software lifecycle

BD2K Implementation

1. Implement the Commons (indices, standards, etc.)
2. Data science research programs (Centers, U01s, etc.)
3. Training and workforce development programs
4. Addressing sustainability of science, technology, and funding mechanisms

* http://acd.od.nih.gov/diwg.htm

The All-Hands Meeting Provides a Yardstick for Progress

- ❑ 439 participants
- ❑ 167 remote viewers
- ❑ Breakout sessions
- ❑ 133 Posters
- ❑ 16 Demos
- ❑ 3 BOFs

http://www.scgcorp.com/bd2k2015/Default

National Institutes of Health
Office of the Director
Data Science at NIH

Some Trends

- **Large datasets -** 46M Aetna EHRs
- **Data integration -** Mobile health + Yelp
- **Analysis -** Machine learning to predict phenotype from EHRs
- **Diverse data types -** Genomics, proteomics, imaging, clinical trials, EHRs
- **Collaboration -** Joint API development, use and requests for metadata templates, data sharing

National Institutes of Health
Office of the Director
Data Science at NIH

Lets look at this progress in terms of the original ACD recommendations

See our strategic plan for 2016-17

National Institutes of Health
Office of the Director
Data Science at NIH

1. Sharing Data & Software Through Indexes

- Protect privacy, proprietary interests, and preserve the balance between the benefits of access/preservation and the costs
- Ensure that all NIH-funded researchers prepare data management and sharing plans
- Ensure that plans are reviewed during peer review
- Encourage use of established repositories and community-based standards
- Develop approaches to ensure discoverability of data
- Implement the Commons

What is The Commons?

- A shared virtual space that:
- Contains digital research objects (data, software, methods, papers, etc.)
- Conforms to **FAIR** principles:
  - **F**indable
  - **A**ccessible (*and usable)*
  - **I**nteroperable
  - **R**eusable

The Commons Framework

Software: Services & Tools

App store/User Interface

Scientific analysis tools/workflows

Services: APIs, Containers, Indexing

Data

"Reference" Data Sets

User defined data

Compute Platform: Cloud or HPC

Digital Object Compliance

National Institutes of Health
Office of the Director
Data Science at NIH

Mapping BD2K Activities to the Commons

Cloud Credits →

HMP
MODS
GDC →

Indexing →

NIH +
Community
defined data sets →

Cloud Credits →

Software: Services & Tools

App store/User Interface

Scientific analysis tools/workflows

Services: APIs, Containers, Indexing

Data

"Reference" Data Sets

User defined data

Compute Platform: Cloud or HPC

Digital Object Compliance

National Institutes of Health
Office of the Director
Data Science at NIH

- Need to find and share data/metadata standards?
  - **Standards Coordinating Center**
- Need to find tools to make annotation and curating easier?
  - **Center for Expanded Data Annotation and Retrieval**
- Need to find resources related to Data Science training and education?
  - **Training Coordination Center**
- Need to tools and resources arising from the BD2K Centers?
  - **Centers Coordinating Center**

National Institutes of Health
Office of the Director
Data Science at NIH

2. Advance big methods, tools, and applications

Examples…

National Institutes of Health
Office of the Director
Data Science at NIH

Supports innovative analytical methods and software tools that address critical current and emerging needs of the biomedical research

- 2015 Topics (18 awards, U01s)
  - Data Compression
  - Data Provenance
  - Data Visualization
  - Data Wrangling
- 2016 Topics (U01s, under review)
  - Data Privacy
  - Data Repurposing
  - Applying Metadata
- 2016 Crowdsourcing and Interactive Digital Media (UH2)

3. Expand data science training

Biomedical Science Specialists

Courses (R25) [11 awards]

Diversity (R25) [4]

Open Educational Resource (R25s) [8 awards]

General Public and K12 → Undergrad → Graduate → Postdoc → Junior faculty → Senior faculty

Museum [1]

Training Programs (T32/T15)[6 awards]

Career Development (K01) [20 awards]

Data Science Specialists

National Institutes of Health
Office of the Director
Data Science at NIH

- 2016 Lecture by Carlos Bustamante, Ph.D.
- Posters
- PiCo Lightening Talks
- Event for High School Students
- Workshop on Reproducible Research
- Pies

- Distinguished Lecture Series
- Frontiers in Data Science Lecture Series
- Software Carpentry
- Hackathons

National Institutes of Health
Office of the Director
Data Science at NIH

The Open Science Prize

NIH National Institutes of Health
Turning Discovery Into Health

wellcome trust

hhmi | Howard Hughes Medical Institute

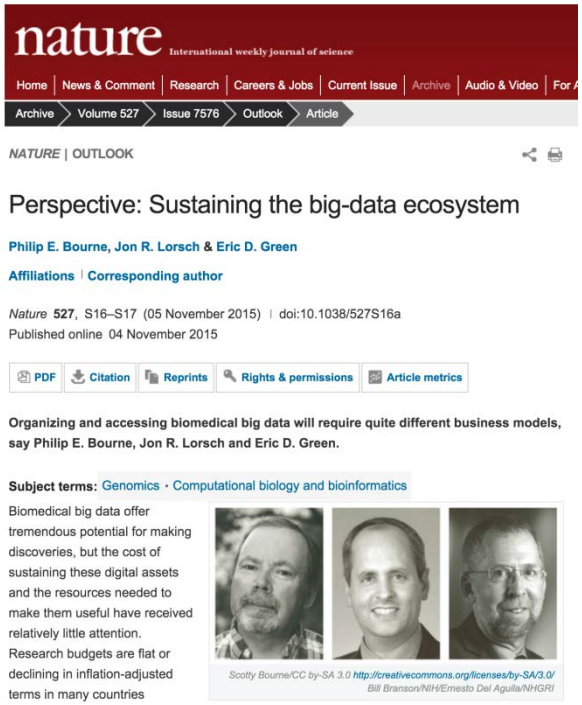
DEPARTMENT OF HEALTH & HUMAN SERVICES · USA

NIH National Institutes of Health
Office of the Director
Data Science at NIH

4. Continued support throughout the data & software lifecycle

National Institutes of Health
Office of the Director
Data Science at NIH

Sustaining the Big Data Ecosystem



nature International weekly journal of science

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

Archive > Volume 527 > Issue 7576 > Outlook > Article

NATURE | OUTLOOK

Perspective: Sustaining the big-data ecosystem

Philip E. Bourne, Jon R. Lorsch & Eric D. Green

Affiliations | Corresponding author

Nature **527**, S16–S17 (05 November 2015) | doi:10.1038/527S16a
Published online 04 November 2015

PDF | Citation | Reprints | Rights & permissions | Article metrics

Organizing and accessing biomedical big data will require quite different business models, say Philip E. Bourne, Jon R. Lorsch and Eric D. Green.

Subject terms: Genomics · Computational biology and bioinformatics

Biomedical big data offer tremendous potential for making discoveries, but the cost of sustaining these digital assets and the resources needed to make them useful have received relatively little attention. Research budgets are flat or declining in inflation-adjusted terms in many countries

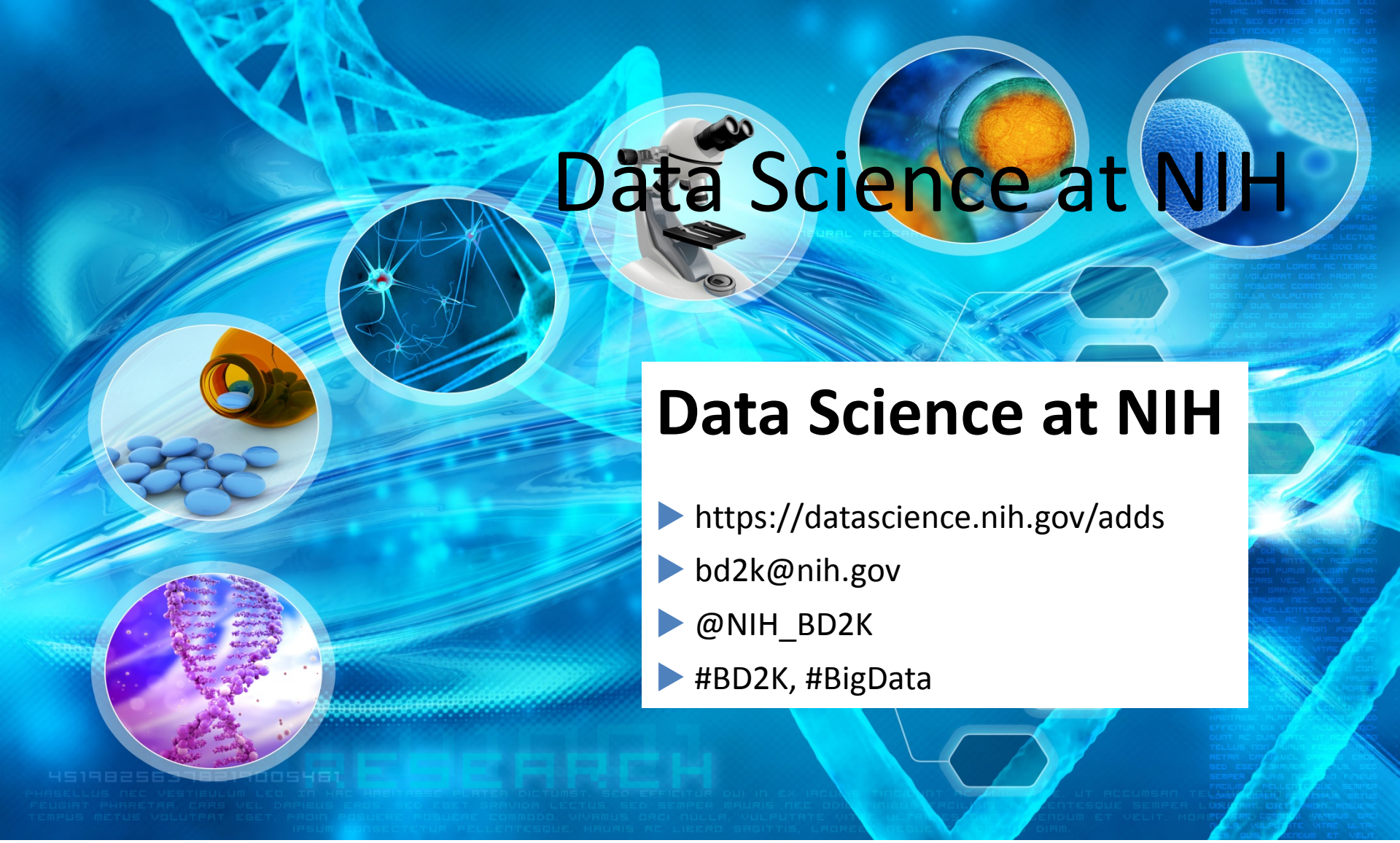
Scotty Bourne/CC by-SA 3.0 http://creativecommons.org/licenses/by-SA/3.0/
Bill Branson/NIH/Ernesto Del Aguila/NHGRI

- Revised governance structure
- Inventory of NIH data repositories and costs
- The Commons
- Interoperability pilots
- Sustainability FOAs
- Policy recommendations

DEPARTMENT OF HEALTH & HUMAN SERVICES · USA

NIH National Institutes of Health
Office of the Director
Data Science at NIH

Data Science at NIH



Data Science at NIH

- https://datascience.nih.gov/adds
- bd2k@nih.gov
- @NIH_BD2K
- #BD2K, #BigData

National Institutes of Health
Office of the Director
Data Science at NIH