Data Science at the NIH

Philip E. Bourne Ph.D.
Associate Director for Data Science
National Institutes of Health
Data Science Timeline

6/12

• Findings:
  • Sharing data & software through catalogs
  • Support methods and applications development
  • Need more training
  • Need campus-wide IT strategy
  • Hire CSIO
  • Continued support throughout the lifecycle
Data Science Timeline

- U54 Centers of Excellence - under review
- U54 BD2K-LINCS – under review
- U24 Data Discovery Index – under review
- R01, R41, R42, R43, R44, U01 software and analysis methods grants – on-going
- T32, T15, K01, R25 and R26 training awards – under review
Data Science Timeline

6/12
- U54 Centers of Excellence - under review
- U54 BD2K-LINCS – under review
- U24 Data Discovery Index – under review
- R01, R41, R42, R43, R44, U01 software and analysis methods grants – on-going
- T32, T15, K01, R25 and R26 training awards – under review

2/14

3/14
ADDS Activities Thus Far:
Talked to Stakeholders (Examples)

- 20/27 IC Directors
- Agencies
  - NSF
  - DOE
  - DARPA
  - NIST
- Government
  - OSTP
  - HHS HDI
  - ONC
- Private sector
  - Phrma
  - Google
  - Amazon
- Organizations
  - PCORI
  - CCC
  - CATS
  - FASEB
  - Biophysical Society
  - Sloan Foundation
  - Moore Foundation
ADDS Activities Thus Far: Some Initial Observations

- **Bad News**
  - We do not yet have a data sustainability plan
  - OSTP have defined the *why* but not the *how*
  - We do not know how all the data we currently have are used
  - We can’t estimate future supply and demand
  - Hence we have not projected the resources that will be required to store and analyze data in the future

- **Good News**
  - Genuine willingness to address the problem across IC’s
  - Efficiencies can be achieved
  - BD2K is the beginnings of a plan
  - We are beginning to quantify the issues
  - We have some of the best data scientists in the world to work on the problems
Based on this data gathering we have defined 5 thematic areas to pursue towards a vision...
The Biomedical Research Digital Enterprise

Programmatic Theme

**Sustainability**
- Cloud – Data & Compute
- Search
- Security
- Reproducibility Standards
- App Store

**Education**
- Coordinate
- Hands-on
- Syllabus
- MOOCs

**Innovation**
- Community
- Centers
- Training Grants
- Catalogs
- Standards
- Analysis

**Process**
- Data Resource Support
- Metrics
- Best Practices
- Evaluation
- Portfolio Analysis

**Collaboration**
- IC’s
- Researchers
- Federal Agencies
- International Partners
- Computer Scientists

**Communicate**

Deliverable

- Commons
- Training Center
- BD2K
- Modified Review

* Hires made

Example Features

Scientific Data Council

External Advisory Board
Some Goals of the Digital Enterprise

- Cost savings through sharing of best practices associated with longitudinal clinical studies
- Collaboration through identification of collaborators at the point of data collection not publication
- Improved reproducibility through data and methods sharing
- Integration of data types and data and literature to accelerate discovery
- Availability of clinical data while respecting patient privacy
On Reproducibility Specifically

- Much of the research life cycle is now digital - encourage the reliability, accessibility, findability, usability of data, methods, narrative, publications etc.

- How?
  - Data sharing plans
  - Standards frameworks
  - Data and software catalogs
  - PubMedCentral

- The Commons – PMC for the complete lifecycle
- Machine readable data sharing plans
- Small funding to communities
- Support for training and best practices in eScholarship
The Biomedical Research Digital Enterprise
The Commons (Vivien Bonnazi & George Komatsoulis (NCBI))

- Public/private partnership
- Work with IC’s, NCBI and CIT to identify and run pilots – cloud, HPC centers
- Port DbGAP to the cloud
- ? Experiment with new funding strategies
- Evaluate
Sustainability and Sharing: The Commons

Commons == Extramural NCBI == Research Object Sandbox == Collaborative Environment

**Data**

**The Why:**
Data Sharing Plans

**The How:**
- Government
- Private Sector
- Rest of Academia
- NIH Awardees

**The Commons**

- Software Index
- Standards
- Data Discovery Index
- BD2K Centers
- Sustainable Storage
- Security/Privacy
- Metrics/Standards
- Quality
- Usability
- Knowledge
- Scientific Discovery

**The End Game:**
- NIH Awardees
- Private Sector
- Rest of Academia
- Government

**Cloud, Research Objects, Business Models**
What Does the Commons Enable?

- Dropbox like storage
- The opportunity to apply quality metrics
- Bring compute to the data
- A place to collaborate
- A place to discover

http://100plus.com/wp-content/uploads/Data-Commons-3-1024x825.png
The Biomedical Research Digital Enterprise

**Programmatic Theme**

- **Sustainability***
  - Sustainability

- **Education***
  - Education
  - Training Center
  - BD2K
  - Modified Review

- **Innovation***
  - Innovation
  - Community
  - Centers
  - Training Grants
  - Catalogs
  - Standards
  - Analysis

- **Process***
  - Process
  - Data Resource Support
  - Metrics
  - Best Practices
  - Evaluation
  - Portfolio Analysis

- **Collaboration***
  - Collaboration
  - IC’s Researchers
  - Federal Agencies
  - International Partners
  - Computer Scientists

**Deliverable**

- **Commons**

**Example Features**

- Cloud – Data & Compute
- Search
- Security
- Reproducibility Standards
- App Store
- Coordinate
- Hands-on
- Syllabus
- MOOCs

* Hires made
Training (Michelle Dunn)

- **Training Goals:**
  - Develop a sufficient cadre of researchers skilled in the science of Big Data
  - Elevate general competencies in data usage and analysis across the biomedical research workforce
  - Combat the Google bus

- **How:**
  - Traditional training grants
  - Work with IC’s on a needs assessment
  - Work with institutions on raising awareness
  - Training center(s)?
Programmatic Theme

Sustainability*

Education*

Innovation*

Process

Collaboration

Deliverable

Commons

Training Center

BD2K

Modified Review

Communication

Example Features

- Cloud – Data & Compute
- Search
- Security
- Reproducibility
- Standards
- App Store

- Coordinate
- Hands-on
- Syllabus
- MOOCs

- Community
- Centers
- Training Grants
- Catalogs
- Standards
- Analysis

- Data Resource Support
- Metrics
- Best Practices
- Evaluation
- Portfolio Analysis

- IC’s
- Researchers
- Federal Agencies
- International Partners
- Computer Scientists

The Biomedical Research Digital Enterprise
BD2K Innovation (Jennie Larkin and Mark Guyer)

- **Data Discovery Index** Coordination Consortium (U24) *(under review)*

- **Metadata standards** *(under development)*

- **Targeted Software Development**
  Development of Software and Analysis Methods for Biomedical Big Data in Targeted Areas of High Need (U01)
  - RFA-HG-14-020
  - Application receipt date June 20, 2014
  - Topics: data compression/reduction, visualization, provenance, or wrangling.
  - Contact: Jennifer Couch (NCI) and Dave Miller (NCI)
BD2K Innovation (Jennie Larkin and Mark Guyer)

- **BISTI PARs**
  - BISTI: *Biomedical Information Science and Technology Initiative*
  - Joint BISTI-BD2K effort
  - R01s and SBIRs
  - Contacts: Peter Lyster (NIGMS) and Jennifer Couch (NCI)

- **Workshops:**
  - Software Index (Last week)
    - Need to be able to find and cite software, as well as data, to support reproducible science.
  - Cloud Computing (Summer/Fall 2014)
    - Biomedical big data are becoming too large to be analyzed on traditional localized computing systems.
  - Contact: Vivien Bonazzi (NHGRI)
BD2K Innovation (Jennie Larkin and Mark Guyer)

- FY14
  - Investigator-initiated Centers of Excellence for Big Data Computing in the Biomedical Sciences (U54) RFA-HG-13-009 (closed)
  - BD2K-LINCS-Perturbation Data Coordination and Integration Center (DCIC) (U54) RFA-HG-14-001 (closed)
The Biomedical Research Digital Enterprise

**Programmatic Theme**

- **Sustainability** *
  - Commons
  - **Example Features**
    - Cloud – Data & Compute
    - Search
    - Security
    - Reproducibility Standards
    - App Store

- **Education** *
  - Training Center
  - **Example Features**
    - Coordinate
    - Hands-on
    - Syllabus
    - MOOCs

- **Innovation** *
  - BD2K
  - **Example Features**
    - Community
    - Centers
    - Training Grants
    - Catalogs
    - Standards
    - Analysis

- **Process**
  - Modified Review
  - * Hires made
    - Data Resource Support
    - Metrics
    - Best Practices
    - Evaluation
    - Portfolio Analysis

- **Collaboration**
  - Communication
    - IC’s
    - Researchers
    - Federal Agencies
    - International Partners
    - Computer Scientists
Process (All / OD /CSR)

- **Goals:**
  - Better data sharing e.g., genomic data sharing plan
  - Capture the best investigators

- **How:**
  - Machine readable data sharing plans?
  - Open review?
  - Micro funding?
  - Standing data committees to explore best practices?
  - Crowd sourcing?
Data Science Timeline FY15

- Internal Retreat
- FY14 Awards
- Clinical Hire
- 6/14
- 9/14
- 11/14
- New Metrics
- 11/15
- FY15 Awards
- Training Assessment
- Training Center
- Commons Pilots
- Commons Launch
- External Advisory
- NIH
- Best Practices
Some Acknowledgements

- Eric Green & Mark Guyer (NHGRI)
- Jennie Larkin (NHLBI)
- Leigh Finnegan (NHGRI)
- Vivien Bonazzi (NHGRI)
- Michelle Dunn (NCI)
- Mike Huerta (NLM)
- David Lipman (NLM)
- Jim Ostell (NLM)
- Andrea Norris (CIT)
- Peter Lyster (NIGMS)
- All the over 100 folks on the BD2K team
NIH... philip.bourne@nih.gov

Turning Discovery Into Health