Data Culture at the NIH

Story
Slides

Slide 1 Data at the NIH: Some Early Thoughts
Slide 2 Background
Slide 3 Disclaimer
Slide 4 Motivation for Change: PDB Growth in Numbers and Complexity
Slide 5 Motivation for Change: We Are at the Beginning
Slide 6 Motivation for Change: We Are at an Inflection Point for Change
Slide 7 From the Second Machine
Slide 8 Much Useful Groundwork Has Been Done
Slide 9 NIH Data & Informatics Working Group
Slide 10 Big Data To Knowledge (BD2K)
Slide 11 Currently...
Slide 12 This is just the beginning...Some Early Observations
Slide 13 Some Early Observations 1
Slide 14 Consider What Might Be Possible
Slide 15 We Need to Learn from Industries Whose Livelihood Addresses the Question of Use
Slide 16 Some Early Observations 2
Slide 17 We have focused on the why, but not the how
Slide 18 Considering a Data Commons to Address this Need
Slide 19 Some Early Observations 3
Slide 20 Sustainability 1
Slide 21 Sustainability 2
Slide 22 Some Early Observations 4
Slide 23 Training in biomedical data science is spotty
Slide 24 Some Early Observations 5
Slide 25 Press Coverage
Slide 26 I can't reproduce research from my own laboratory?
Slide 27 Characteristics of the Original and Current Experiment
Slide 28 Considered the Ability to Reproduce by Four Classes of User

http://semanticommunity.info/Data_Science/Data_Culture_at_the_NIH
Updated: Fri, 18 Sep 2015 23:53:15 GMT
Powered by mindtouch
Slide 29 A Conceptual Overview of the Method Should be Mandatory
Slide 30 Time to Reproduce the Method
Slide 31 Perhaps it is time we did better?
Slide 32 The Digital Enterprise
Slide 33 Components of The Academic Digital Enterprise
Slide 34 Life in the Academic Digital Enterprise
Slide 35 Life in the NIH Digital Enterprise
Slide 36 Consider the Complete Research Lifecycle
Slide 37 The Research Life Cycle will Persist
Slide 38 Tools and Resources Will Continue To Be Developed
Slide 39 Those Elements of the Research Life Cycle will Become More Interconnected Around a Common Framework
Slide 40 New/Extended Support Structures Will Emerge
Slide 41 We Have a Ways to Go
Slide 42 Next Steps
Slide 43 NIH...Turning Discovery Into Health
Slide 44 Thank you!
Slide 45 Back Pocket Slides for BD2K Programs
Slide 46 Facilitating Broad Use 1
Slide 47 Facilitating Broad Use 2
Slide 48 Facilitating Big Data Analysis 1
Slide 49 Facilitating Big Data Analysis
Slide 50 Enhancing Training
Slide 51 BD2K Training RFAs
Slide 52 BD2K Centers of Excellence

Spotfire Dashboard
- H1N1 Spread
- NIH Data Publication 1

Research Notes
- URLs
- Slide Text

Taking on the Role of Associate Director for Data Science at the NIH – My Original Vision Statement

About BD2K
- Mission Statement
- What is Big Data?
- Major challenges in using biomedical Big Data include:
- Organizational Structure
- Areas of BD2K
Enabling Data Utilization (Click to Watch Video)
Analysis Methods and Software
Enhancing Training
Centers of Excellence
BD2K Executive Committee
Scientific Data Council

H1N1 Flu
Background
The Numbers
CDC Estimates of 2009 H1N1 Cases and Related Hospitalizations and Deaths from April 2009 - March 13, 2010, By Age Group
Discussion
Graphics
Method to Estimate 2009 H1N1 Cases, Hospitalizations and Deaths
Background Emerging Infections Program
Seasonal Influenza-Associated Hospitalizations in the United States
Seasonal Influenza-Associated Deaths
Under-Counting of Flu-Related Deaths
Related Links
Contact Us
History

TB-drugome
Introduction
Download TB-drugome
Summary of Files and Directories
drug_site_info.tab
pdb_structure_info.tab
homology_model_info.tab
SMAP_eHiTs_pdb_structure_info.tab
SMAP_eHiTs_homology_model_info.tab
eHiTs_docking_pdb_structures
eHiTs_docking_homology_models
Contact
Frequently Asked Questions
Reports
General

When I apply for a BD2K grant, do I need to specify a targeted NIH Institute?
Are applications to BD2K FOAs restricted to a particular type of data?
Can I submit an application to this grant that I have previously submitted to another FOA?

KO1

There seems to be a typo in the program announcement, as it states "The NIH will contribute up to per year", without specifying the amount. Could you please clarify?

Is this opportunity targeting investigators whose project(s) are applications of existing methods or development of new technologies?

Eligibility

If I already have a background in computer science, high performance computing, data mining, bioinformatics, or a related field, is my background considered too close to be eligible to apply to this program?

My background is exclusively/preponderantly in one of the the three areas mentioned in the RFA: am I eligible?

If I am already tenured, am I eligible to apply?

If I am a postdoc, am I eligible to apply?

Am I eligible to apply if I already have an R01 (or equivalent)?

If I do not yet have a green card, am I eligible to apply? How will it affect the review of my proposal?

If I intend to study overseas during part of the award period of this grant, am I eligible to be funded during that time?

Mentors

My background is exclusively/preponderantly in one of the three areas mentioned in the FOA: what fields should my mentors be from?

Can my mentor be from outside my campus?

Can a mentor be from a foreign country?

Writing and Submitting an Application

Must my training plan include coursework?

I currently hold a full-time position at an institution, and in the course of this year, I will be taking a position at another institution. Which institution should be the applicant institution?

Can letters of reference come from my proposed mentors?

Can I assume that the reviewers will be familiar with the type of Big Data and the Big Data problem in my project?

Miscellaneous

Can I devote less than 75% effort to this award?

Is there a max percent effort that the K01 will fund? My annual salary is under $167,000, so will the K01 fund 100% of my effort or do I need a minimum amount of institutional support?

Can I delay the award or take a leave of absence?

R25

Can this grant be used to expand an existing program at my institution to accommodate more students?

Would it be responsive to create a course on one of the major MOOC providers, e.g. Coursera, Udacity, or EdX?

Does this grant permit expenditures for computing costs?
Can a training project, which was submitted as a part of a BD2K Centers application, also be submitted as an R25?

Would a proposal for training in the Responsible Conduct of Research (e.g ethics or reproducible research) be responsive to this FOA?

Training Programs
Can an institution submit more than one application for BD2K training grants?

Flu Data
Seasonal Flu
FluView Interactive: Influenza Surveillance Data the Way You Want it!
FluView Surveillance
FluView Hospitalizations
ILI Activity Map
Pediatric Mortality

Google Flu
Flu Trends
Learn more about the research behind Google Flu Trends
Detecting influenza epidemics using search engine query data

Abstract
Introduction
Figure 1: An evaluation of how many top-scoring queries to include in the ILI-related query fraction
Table 1: Topics found in search queries which were found to be most correlated with CDC ILI data
Figure 2: A comparison of model estimates
Figure 3: ILI percentages estimated

Methods
Privacy
Search query database
Model data
Automated query selection process
Computation and pre-filtering
Constructing the ILI-related query fraction
Fitting and validating a final model
State-level model validation

Acknowledgements
Author contributions
Supplementary material
References

1
Quantifying Reproducibility in Computational Biology

Abstract

Introduction

Related Work

Methods and Analysis

Quantifying Reproducibility

Methodology

Conceptual Overview of the Method and Final Workflow

Figure 1. A high-level dataflow diagram of the TB drugome method.

Figure 2. The reproduced TB Drugome workflow with the different subsections highlighted

Reproducibility Analysis

Reproducibility Maps

Figure 3. Reproducibility maps of the three major subsections of the workflow

Productivity and Effort

Table 1. Time to reproduce the method

Table 2. Observations and desiderata for reproducibility

Table 3. Reproducibility Guidelines for Authors

Publishing the Reproduced Workflow

Discussion

Conclusions

Supporting Information

Author Contributions

References
Supplement S1

Comparison of Ligand Binding sites using SMAP

Uncovering the Workflow
  - Figure S1: First attempt to reproduce the SMAP analysis
  - Figure S2: Second attempt to reproduce the SMAP analysis
  - Figure S3: Third attempt to reproduce the SMAP analysis
  - Figure S4: Fourth attempt to reproduce the SMAP analysis

Computational Workflow
  - Figure S5: SMAP workflow

Reproducibility Scores
  - Table S1: Reproducibility scores for the SMAP workflow

Comparison of dissimilar protein structures using FATCAT

Uncovering the Workflow
  - Figure S6: First attempt to reproduce the FATCAT analysis
  - Figure S7: Second attempt to reproduce the FATCAT analysis

Computational Workflow
  - Figure S8: FATCAT workflow

Reproducibility Scores
  - Table S2: Reproducibility scores for the FATCAT workflow
Docking using eHits/AutodockVina

Uncovering the Workflow

Figure S9: First attempt for reproducing the docking analysis

Computational Workflow

Figure S10: Docking workflow

Reproducibility Scores

Table S3: Reproducibility scores for the docking workflow

Original results versus results from the workflow

Table S4: Highly connected drugs, original results
Table S5: Highly connected drugs, results of the workflow.

Figure S11: Visualization of the drug-protein network obtained as a result of the workflow

FIGURES AND TABLES

Figure S1: First attempt to reproduce the SMAP analysis

Figure S2: Second attempt to reproduce the SMAP analysis
Figure S3: Third attempt to reproduce the SMAP analysis
Figure S4: Fourth attempt to reproduce the SMAP analysis
Figure S5: SMAP workflow
Figure S6: First attempt to reproduce the FATCAT analysis
Figure S7: Second attempt to reproduce the FATCAT analysis
Figure S8: FATCAT workflow
Figure S9: First attempt for reproducing the docking analysis
Figure S10: Docking workflow

Figure S11: Visualization of the drug-protein network obtained as a result of the workflow

Table S1: Reproducibility scores for the SMAP workflow
Table S2: Reproducibility scores for the FATCAT workflow
Table S3: Reproducibility scores for the docking workflow
Table S4: Top rated highly connected drugs (original results)
Table S5: Top rated highly connected drugs, results of the workflow

RFI Summary

Executive Summary

RFI Summary: Full Report

Knowledge and Skills Needed to Utilize Biomedical Big Data
Teamwork and Cross-training
Audience for Data and Informatics Training and Education
How Training Should be Achieved

Workshop Report on Enhancing Training

Introduction
Purpose and Goals of the Workshop
Figure Big Data Issues

Background Relevant to the Workshop Discussion

ACD Working Group on Data and Bioinformatics (DIWG)

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

Four Examples of Big Data Challenges and Competencies Needed to Make Full Use of the Data

Electronic Health Records (Daniel Masys)

Imaging (Ron Kikinis)

Genomics (Michael Boehnke)

Integrations of Large and Small Datasets (Mark Musen)

Summary of the Workshop Deliberations and Recommendations

Knowledge and Skills Needed

Long-term Training and Career Award Programs

Short-term Training Programs

Innovative Training Technology

Data Sharing

Curriculum Development

Prioritization

Appendix 1: Rationale for the Workshop

Appendix II: Agenda Big Data to Knowledge (BD2K) Workshop

Monday, 29 July

Tuesday, 30 July

Appendix III: BD2K Workshop on Enhancing Training Invited Guests

Appendix III: BD2K Workshop on Enhancing Training NIH Participants

Appendix IV NIH BD2K Training Working Group Members

Appendix V ACD Working Groups

Story

Philip Bourne: Changing the Data Culture at NIH

Dr. Philip Bourne, Associate Director for Data Sciences, NIH, gave a keynote at the recent Ontology Summit 2014 entitled: Data at the NIH: Some Early Thoughts, which he openly shared on Slide Share.

He said his mission at NIH is to change the data culture. That stuck a resonant chord with me as a former government employee when I went on detail from the US EPA to the USGS and found that the USGS strongly encouraged everyone to have a web server and publish their research. The US EPA never allowed that until I started using state-of-the-art Wiki technology for my Federal CIO Council community building activities for web-services, SOA, and semantics.
So he asked me to describe what I had done. I have worked with Semantic Medline, as suggested by NITRD/NCO Director Dr. George Strawn, and with NLM Open Government Data APIs as suggested by Federal CTO Todd Park.

I told him our data science teams would like to do more collaborative work with NIH data sets and Cray/YarcData is offering to do more free pilots like Semantic Medline with GMU students.

I made the following comments on his excellent presentation as follows:

- Include Dr. Collins TedMed talk last year
- See our recent Joint NSF-NIH Biomedical Big Data Research Meetup (George Strawn, Peter Lister, and Barend Mons)
- Consider my recommendations to Congressional Staff for the Data Act that we are trying to implement in our Meetups:
  - Chief Data Science Officer for the Federal Government and each Agency that directs a team of Data Scientists working on significant data science problems for government business and science that provide answers and web links to the following three basic questions: How was the data collected? Where is it stored? What are the results?
- So you are the Chief Data Science Officer and you have the best Data Science Team that I know of in the government led by Drs. Lindberg, Bodenreider, and Rindflesch that we have helped get Semantic Medline running on the YarcData Graph Appliance.
  - I included this in my recent story and pilot for the Ontology Summit:
- I look forward to getting together to discuss how our Meetup activity might be of further help with NIH Big Data sets.

So I want to help Dr. Bourne lead by example in changing the NIH data culture (and even the entire biomedical data culture) in the following ways: Make his content a data commons that is like a sandbox/dropbox in the cloud, where everything (object) has an identifier, that includes some of his pioneering work (e.g. TB-drugome) and that of others (e.g. H1N1 Flu, CDC FluView, and Google Flu Trends) he would like to emulate. Maybe I should do the same for all the senior NIH Leadership and then drill down into the organization and make more "NIH data publications for data browsers!"

The reproducible results task here is for the Flu Data and I include the "workflow" (see Discussion).

First, the Knowledge Base in MindTouch for the Data Publication of "research objects as digital objects":

http://semanticommunity.info/Data_Science/Data_Culture_at_the_NIH
Updated: Fri, 18 Sep 2015 23:53:15 GMT
Powered by MindTouch
Second, the Knowledge Base in the Spreadsheet for the Data Publication of "research objects as digital objects":

Third, an earlier interactive visualization of HINI data in Spotfire:

Finally, the Knowledge Base and the Flu Data Sets in Spotfire for the Data Publication in the Data Browser:
Cover Page: Google Flu Data

Tab 1: CDC Historic Flu & Current FluView Interactive Data

Tab 2: NIH Data Culture Knowledge Base

Tab 3: Garijo, et al 2013 Paper
It should be noted that I found the Google Flu data easy to work with, I had difficulty downloading the FluView Interactive data, and I could not import the TB-drugome .dat files into Spotfire.

I looked at the RCSB Protein Data Bank and the PLOS Computational Biology that Dr. Bourne cites and it looks like the RCSB PDB Statistics and Journal Archive, respectively would be the place to start with making those data publications like I have done recently for the International Journal of Digital Earth for CODATA.

Finally, Dr. Bourne wants to work like and with the Gordon and Betty Moore Foundation and I am suggesting a "grand data-driven discovery challenge" we have started to work on that would integrate Semantic Medline, JHU geonomic databases, and nutrition databases.

MORE TO FOLLOW

Slides

http://www.slideshare.net/pebourne/ontology-nsf042814

Slide 1 Data at the NIH: Some Early Thoughts

http://www.slideshare.net/pebourne/ontology-nsf042814
Slide 2 Background

Background

- Research in computational biology...
- Co-directed the RCSB Protein Data Bank (1999-2014)
- Co-founded PLOS Computational Biology; First EiC (2005 – 2012)

With Ontologies:
- Extensive work with the Gene Ontology
- Co-developed mmCIF for macromolecular structure

Slide 3 Disclaimer

http://pebourne.wordpress.com/2013/12/
Slide 4 Motivation for Change: PDB Growth in Numbers and Complexity

Slide 5 Motivation for Change: We Are at the Beginning
Slide 6 Motivation for Change: We Are at an Inflection Point for Change

- Evidence:
  - Google car
  - 3D printers
  - Waze
  - Robotics

Slide 7 From the Second Machine

*From the Second Machine Age: Work, Progress, and Popularity in a Time of Bilateral Technologies by Erik Kripal and Andrew M. Hacker*

http://semanticommunity.info/Data_Science/Data_Culture_at_the_NIH
Updated: Fri, 18 Sep 2015 23:53:15 GMT
Powered by mindtouch
Slide 8 Much Useful Groundwork Has Been Done

Slide 9 NIH Data & Informatics Working Group

In response to the growth of large biomedical datasets, the Director of NIH established a special Data and Informatics Working Group (DIWG).

Slide 10 Big Data To Knowledge (BD2K)

http://bd2k.nih.gov
Slide 11 Currently...

1. Facilitating Broad Use
2. Developing and Disseminating Analysis Methods and Software
3. Enhancing Training
4. Establishing Centers of Excellence

http://bd2k.nih.gov

Slide 12 This is just the beginning...Some Early Observations

- Data Discovery Index – under review
- Data Centers – under review
- Training grants – RFA’s issued; under review
- Software index – workshop in May
- Catalog of standards – FOA under development

This is just the beginning...

Some Early Observations
Slide 13 Some Early Observations

1. We don't know enough about how existing data are used.

Slide 14 Consider What Might Be Possible

http://www.cdc.gov/h1n1flu/estimates...l_March_13.htm My Note: See My Illness Tracking Spotfire
Slide 15 We Need to Learn from Industries Whose Livelihood Addresses the Question of Use

Slide 16 Some Early Observations 2

1. We don't know enough about how existing data are used
2. We have focused on the why, but not the how
Slide 17 We have focused on the why, but not the how

- The OSTP directive is the why
- The how is needed for:
  - Any data that does not fit the existing data resource model
  - Data generated by NIH cores
  - Data accompanying publications
  - Data associated with the long tail of science

Slide 18 Considering a Data Commons to Address this Need

http://100plus.com/wp-content/upload...3-1024x825.png

- AKA NIH drive – a dropbox for NIH investigators
- Support for provenance and access control
- Likely in the cloud
- Support for validation of specific data types
- Support for mining of collective intramural and extramural data across IC’s
- Needs to have an associated business model
Slide 19 Some Early Observations

Some Early Observations

1. We don’t know enough about how existing data are used
2. We have focused on the why, but not the how
3. We do not have an NIH-wide sustainability plan for data (not heard of an IC-based plan either)

Slide 20 Sustainability 1

3. Sustainability

- Problems
  - Maintaining a work force – lack of reward
  - Too much data, too few dollars
- Resources
  - In different stages of maturity but treated the same
  - Funded by a few used by many
    - True as measured by IC
    - True as measured by agency
    - True as measured by country
- Reviews can be problematic
3. Sustainability

- Possible Solutions
  - Establish a central fund to support
  - The 50% model
  - New funding models eg open submission and review
  - Split innovation from core support and review separately
  - Policies for uniform metric reporting
  - Discuss with the private sector possible funding models
  - More cooperation, less redundancy across agencies
  - Bring foundations into the discussion
  - Discuss with libraries, repositories their role
  - Educate decision makers as to the changing landscape

Some Early Observations

1. We don't know enough about how existing data are used
2. We have focused on the why, but not the how
3. We do not have an NIH-wide sustainability plan for data (not heard of an IC-based plan either)
4. Training in biomedical data science is spotty
Slide 23 Training in biomedical data science is spotty

4. Training in biomedical data science is spotty

- Problem
  - Coverage of the domain is unclear
  - There may well be redundancies

- Solution
  - Cold Spring Harbor like training facility(s)
  - Training coordinator
  - Rolling hands on courses in key areas
  - Appropriate materials on-line
  - Interagency training initiatives

Slide 24 Some Early Observations

Some Early Observations

1. We don’t know enough about how existing data are used
2. We have focused on the why, but not the how
3. We do not have an NIH-wide sustainability plan for data (not heard of an IC-based plan either)
4. Training in biomedical data science is spotty
5. Reproducibility will need to be embraced
Slide 25 Press Coverage

Slide 26 I can't reproduce research from my own laboratory?

Slide 27 Characteristics of the Original and Current Experiment

http://funsite.sdsc.edu/drugome/TB/
Slide 28 Considered the Ability to Reproduce by Four Classes of User

- REP-AUTHOR – original author of the work
- REP-EXPERT – domain expert – can reproduce even with incomplete methods described
- REP-NOVICE – basic domain (bioinformatics) expertise
- REP-MINIMAL – researcher with no domain expertise

Garpio et al 2013 PLOS ONE 8(11): e60278

Slide 29 A Conceptual Overview of the Method Should be Mandatory

Garpio et al 2013 PLOS ONE 8(11): e60278
Slide 30 Time to Reproduce the Method

![Table 1: Time to reproduce the method.](Ganjo et al 2013 PLOS ONE 8(11): e80279)

Slide 31 Perhaps it is time we did better?

*It's not that we could not reproduce the work, but the effort involved was substantial*

*Any graduate student could tell you this and little has changed in 40 years*

*Perhaps it is time we did better?*
I cast the solutions in a vision … something I call the digital enterprise

Any institution is a candidate to be a digital enterprise, but let’s explore it in the context of the academic medical center

Slide 33 Components of The Academic Digital Enterprise

Components of The Academic Digital Enterprise

- Consists of digital assets
  - E.g. datasets, papers, software, lab notes

- Each asset is uniquely identified and has provenance, including access control
  - E.g. publishing simply involves changing the access control

- Digital assets are interoperable across the enterprise
Slide 34 Life in the Academic Digital Enterprise

- Researchers rely on a variety of technologies and tools to conduct their work. These technologies include software for data analysis, cloud computing services, and collaboration tools. Researchers also use social media platforms to share their findings and collaborate with other researchers.
- The NIH Digital Enterprise is an initiative aimed at improving the efficiency and effectiveness of research by leveraging digital technologies. It includes initiatives such as the NIH Cross-Cutting Initiative and the NIH Vision of the NIH Digital Enterprise.
- The NIH Digital Enterprise also includes efforts to improve data sharing and access, such as the NIH Data Sharing Framework and the NIH Data Commons.


Slide 35 Life in the NIH Digital Enterprise

- Researchers x and y are aware of each other through the NIH Digital Enterprise. They have collaborated in the past and plan to collaborate in the future.
- The NIH Digital Enterprise provides tools and resources to support collaboration, such as the NIH Cross-Cutting Initiative and the NIH Vision of the NIH Digital Enterprise.
- The NIH Digital Enterprise also includes initiatives to improve data sharing and access, such as the NIH Data Sharing Framework and the NIH Data Commons.


http://semanticommunity.info/Data_Science/Data_Culture_at_the_NIH
Updated: Fri, 18 Sep 2015 23:53:15 GMT
Powered by Qmindtouch™
Slide 36 Consider the Complete Research Lifecycle

To get to that end point we have to consider the complete research lifecycle.

Slide 37 The Research Life Cycle will Persist

The Research Life Cycle will Persist

IDEAS -> HYPOTHESES -> EXPERIMENTS -> DATA -> ANALYSIS -> COMPREHENSION -> DISSEMINATION
Slide 38 Tools and Resources Will Continue To Be Developed

Slide 39 Those Elements of the Research Life Cycle will Become More Interconnected Around a Common Framework
Slide 40 New/Extended Support Structures Will Emerge

Slide 41 We Have a Ways to Go
Slide 42 Next Steps

Next Steps

- Support for research objects
  - These objects underpin the various cataloging efforts

- Support for data metrics
  - Such metrics underpin a change in the reward system

Slide 43 NIH...Turning Discovery Into Health

NIH... Turning Discovery Into Health

philip.bourne@nih.gov
Slide 44 Thank you!

Slide 45 Back Pocket Slides for BD2K Programs

Slide 46 Facilitating Broad Use 1

http://bd2k.nih.gov
1. Facilitating Broad Use

- Summary of Data Catalog Workshop and Request for Information: [www.caclab.nih.gov](http://www.caclab.nih.gov)
- Data Discovery Index (DDI)
  - *Will make data findable and citable!*
- RFA-HL-14-031, Data Discovery Index Coordination Consortium (U24) (closed)
  - Will fund one U24 award: community engagement, identification of challenges, and testing of possible solutions.
  - Contacts: Ron Margolis (NIDDK) and Jennie Larkin (NHLBI)

---

Slide 48 Facilitating Big Data Analysis 1

2. Facilitating Big Data Analysis

- **Broad-based, on-going BISTI PARs**
  - BISTI: Biomedical Information Science and Technology Initiative
  - Joint BISTI-B2K effort
  - R01s and S1BRs
  - Contacts: Peter Lyster (NIGMS) and Jennifer Couch (NCI)

- **Planned Workshops:**
  - Software Index (Spring 2014)
    - 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 local/hub computing systems.
    - Contact: Vivien Bonczak (NHLBI)

---

http://semanticommunity.info/Data_Science/Data_Culture_at_the_NIH
Updated: Fri, 18 Sep 2015 23:53:15 GMT
Slide 49 Facilitating Big Data Analysis

2. Facilitating Big Data Analysis

- **RFA for 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)

http://bd2k.nih.gov

Slide 50 Enhancing Training

http://bd2k.nih.gov/faqs_trainingFOA.html

3: Enhancing Training

- **Summary of Training Workshop and Request for Information:**
  - [http://bd2k.nih.gov/faqs_trainingFOA.html](http://bd2k.nih.gov/faqs_trainingFOA.html)
  - Contact: Michelle Dunn (NCI)

- **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.
Spotfire Dashboard

H1N1 Spread

An overview of the spread of the Summer '09 outbreak of the H1N1 virus in the continental United States.

This was done in an earlier version of Spotfire (3.2) and needs to be upgraded and have the Business Process added as follows:

See The Data
Map The Data By State
Map The Data By County
Map The Data With Background Image
Summary, Cross, & Graphical Tables, and Exploratory Data Analysis: Faceted Search & Plots
NIH Data Publication 1
See Story Above for Details

For Internet Explorer Users and Those Wanting Full Screen Display Use: Web Player Get Spotfire for iPad App

Media, iframe, embed and object tags are not supported inside of a PDF.

Research Notes

URLs

http://pebourne.wordpress.com/2013/12/
http://bd2k.nih.gov
http://www.cdc.gov/h1n1flu/estimates...l_March_13.htm
http://100plus.com/wp-content/upload...3-1024x825.png
http://funsite.sdsc.edu/drugome/TB/
http://bd2k.nih.gov/faqs_trainingFOA.html

Slide Text

1. Data at the NIH: Some Early Thoughts Philip E. Bourne Ph.D. Associate Director for Data Science National Institutes of Health

http://www.slideshare.net/pebourne/

2. Background

  ? Research in computational biology…

  ? Co-directed the RCSB Protein Data Bank (1999- 2014)

  ? Co-founded PLOS Computational Biology; First EIC (2005 – 2012)
With Ontologies: – Extensive work with the Gene Ontology – Co-developed mmCIF for macromolecular structure

3. Disclaimer: I only started March 3, 2014 …but I had been thinking about this prior to my appointment

http://pebourne.wordpress.com/2013/12/

4. Number of released entries Year Motivation for Change: PDB Growth in Numbers and Complexity [From the RCSB Protein Data Bank]

5. Motivation for Change: We Are at the Beginning

6. Motivation: We Are at an Inflexion Point for Change


7. From the Second Machine Age From: The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies by Erik Brynjolfsson & Andrew McAfee

8. Much Useful Groundwork Has Been Done

9. NIH Data & Informatics Working NIH Data & Informatics Working Group Group

In response to the growth of large biomedical datasets, the Director of NIH established a special Data and Informatics Working Group (DIWG).

10. Big Data to Knowledge (BD2K) Big Data to Knowledge (BD2K)

1. Facilitating Broad Use

2. Developing and Disseminating Analysis Methods and Software

3. Enhancing Training


11. Currently…

? Data Discovery Index – under review

? Data Centers – under review

? Training grants – RFA’s issued; under review

? Software index – workshop in May

? Catalog of standards – FOA under development
12. This is just the beginning… Some Early Observations

13. Some Early Observations

1. We don’t know enough about how existing data are used


15. We Need to Learn from Industries Whose Livelihood Addresses the Question of Use

16. Some Early Observations

1. We don’t know enough about how existing data are used

2. We have focused on the why, but not the how

17. 2. We have focused on the why, but not the how

? The OSTP directive is the why

? The how is needed for: – Any data that does not fit the existing data resource model

- Data generated by NIH cores
- Data accompanying publications
- Data associated with the long tail of science

18. Considering a Data Commons to Address this Need

? AKA NIH drive – a dropbox for NIH investigators

? Support for provenance and access control

? Likely in the cloud? Support for validation of specific data types

? Support for mining of collective intramural and extramural data across IC's

? Needs to have an associated business model

http://100plus.com/wp-content/upload...3-1024x825.png

19. Some Early Observations

1. We don’t know enough about how existing data are used
2. We have focused on the why, but not the how

3. We do not have an NIH-wide sustainability plan for data (not heard of an IC-based plan either)

20. 3. Sustainability

? Problems – Maintaining a work force – lack of reward – Too much data; too few dollars – Resources

- In different stages of maturity but treated the same
- Funded by a few used by many – True as measured by IC – True as measured by agency – True as measured by country
- Reviews can be problematic

21. 3. Sustainability

? Possible Solutions – Establish a central fund to support – The 50% model – New funding models eg open submission and review – Split innovation from core support and review separately – Policies for uniform metric reporting – Discuss with the private sector possible funding models – More cooperation, less redundancy across agencies – Bring foundations into the discussion – Discuss with libraries, repositories their role – Educate decision makes as to the changing landscape

22. Some Early Observations

1. We don’t know enough about how existing data are used

2. We have focused on the why, but not the how

3. We do not have an NIH-wide sustainability plan for data (not heard of an IC-based plan either)

4. Training in biomedical data science is spotty

23. 4. Training in biomedical data science is spotty

? Problem – Coverage of the domain is unclear – There may well be redundancies ? Solution – Cold Spring Harbor like training facility(s)

- Training coordinator
- Rolling hands on courses in key areas
- Appropriate materials on-line – Interagency training initiatives

24. Some Early Observations

1. We don’t know enough about how existing data are used
2. We have focused on the why, but not the how

3. We do not have an NIH-wide sustainability plan for data (not heard of an IC-based plan either)

4. Training in biomedical data science is spotty

5. Reproducibility will need to be embraced

25. 47/53 “landmark” publications could not be replicated [Begley, Ellis Nature, 483, 2012] [Carole Goble]

26. I can’t reproduce research from my own laboratory?


27. Characteristics of the Original and Current Experiment

? Original and Current: – Purely in silico – Uses a combination of public databases and open source software by us and others


? Current: – Recast in the Wings workflow system


28. Considered the Ability to Reproduce by Four Classes of User

? REP-AUTHOR – original author of the work

? REP-EXPERT – domain expert – can reproduce even with incomplete methods described

? REP-NOVICE – basic domain (bioinformatics) expertise

? REP-MINIMAL – researcher with no domain expertise

Garijo et al 2013 PLOS ONE 8(11): e80278

29. A Conceptual Overview of the Method Should Be Mandatory

Garijo et al 2013 PLOS ONE 8(11): e80278

30. Time to Reproduce the Method

Garijo et al 2013 PLOS ONE 8(11): e80278
31. Its not that we could not reproduce the work, but the effort involved was substantial. Any graduate student could tell you this and little has changed in 40 years. Perhaps it is time we did better?

32. I cast the solutions in a vision ... something I call the digital enterprise. Any institution is a candidate to be a digital enterprise, but let's explore it in the context of the academic medical center.

33. Components of the Academic Digital Enterprise

- Consists of digital assets – E.g. datasets, papers, software, lab notes.
- Each asset is uniquely identified and has provenance, including access control – E.g. publishing simply involves changing the access control.
- Digital assets are interoperable across the enterprise.

34. Life in the Academic Digital Enterprise

- Jane scores extremely well in parts of her graduate on-line neurology class. Neurology professors, whose research profiles are on-line and well described, are automatically notified of Jane's potential based on a computer analysis of her scores against the background interests of the neuroscience professors. Consequently, professor Smith interviews Jane and offers her a research rotation. During the rotation, she enters details of her experiments related to understanding a widespread neurodegenerative disease in an on-line laboratory notebook kept in a shared on-line research space – an institutional resource where stakeholders provide metadata, including access rights and provenance beyond that available in a commercial offering. According to Jane's preferences, the underlying computer system may automatically bring to Jane's attention Jack, a graduate student in the chemistry department whose notebook reveals he is working on using bacteria for purposes of toxic waste cleanup. Why the connection? They reference the same gene a number of times in their notes, which is of interest to two very different disciplines – neurology and environmental sciences. In the analog academic health center, they would never have discovered each other, but thanks to the Digital Enterprise, pooled knowledge can lead to a distinct advantage. The collaboration results in the discovery of a homologous human gene product as a putative target in treating the neurodegenerative disorder. A new chemical entity is developed and patented. Accordingly, by automatically matching details of the innovation with biotech companies worldwide that might have potential interest, a licensee is found. The licensee hires Jack to continue working on the project. Jane joins Joe's laboratory, and he hires another student using the revenue from the license. The research continues and leads to a federal grant award. The students are employed, further research is supported and in time societal benefit arises from the technology.

From What Big Data Means to Me JAMIA 2014 21:194

35. Life in the NIH Digital Enterprise

- Researcher x is made aware of researcher y through commonalities in their data located in the data commons. Researcher x reviews the grants profile of researcher y and publication history and impact from those grants in the past 5 years and decides to contact her. A fruitful collaboration ensues and they generate papers, data sets and software. Metrics automatically pushed to company z for all relevant NIH data and software in a specific.
domain with utilization above a threshold indicate that their data and software are heavily utilized and respected by the community. An open source version remains, but the company adds services on top of the software for the novice user and revenue flows back to the labs of researchers x and y which is used to develop new innovative software for open distribution. Researchers x and y come to the NIH training center periodically to provide hands-on advice in the use of their new version and their course is offered as a MOOC.

36. To get to that end point we have to consider the complete research lifecycle


38. Tools and Resources Will Continue To Be Developed IDEAS – HYPOTHESES – EXPERIMENTS – DATA - ANALYSIS - COMPREHENSION - DISSEMINATION Authoring Tools Lab Notebooks Data Capture Software Analysis Tools Visualization Scholarly Communication


41. We Have a Ways to Go IDEAS – HYPOTHESES – EXPERIMENTS – DATA - ANALYSIS - COMPREHENSION - DISSEMINATION Authoring Tools Lab Notebooks Data Capture Software Analysis Tools Visualization Scholarly Communication Commercial & Public Tools Git-like Resources By Discipline Data Journals Discipline- Based Metadata Standards Community Portals Institutional Repositories New Reward Systems Commercial Repositories Training

42. Next Steps

   ? Support for research objects – These objects underpin the various cataloging efforts

   ? Support for data metrics – Such metrics underpin a change in the reward system

43. NIHNIH…… Turning Discovery Into Health philip.bourne@nih.gov

44. Thank You! Questions? philip.bourne@nih.gov

45. Back Pocket Slides for BD2K Programs

46. 1. Facilitating Broad Use1. Facilitating Broad Use

47. 1. Facilitating Broad Use1. Facilitating Broad Use
? Research use of clinical data

? Workshop held Sept 2013

? Workshop report and plans being finalized

? Contacts: Jerry Sheehan (NLM) and Leslie Derr (OD)

? Community-based data and metadata standards

? Will make data usable

? Workshop held Sept 2013

? Workshop report and plans being finalized

? Contact: Mike Huerta (NLM)

48. 2. Facilitating Big Data Analysis

2. Facilitating Big Data Analysis

? Broad-based, on-going BISTI PARs – BISTI: Biomedical Information Science and Technology Initiative – Joint BISTI-BD2K effort – R01s and SBIRs – Contacts: Peter Lyster (NIGMS) and Jennifer Couch (NCI)

? Planned Workshops: – Software Index (Spring 2014) • 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)

49. 2. Facilitating Big Data Analysis

2. Facilitating Big Data Analysis

? RFA for 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) http://bd2k.nih.gov

50. 3: Enhancing Training

3: Enhancing Training


? 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.
51. 3: BD2K Training RFAs: BD2K Training RFAs Application Receipt Date: April 2, 2014

?K01s for Mentored Career Development Awards, RFA-HG-14-007
?

? Provides salary and research support for 3-5 years for intensive research career development under the
guidance of an experienced mentor in biomedical Big Data Science.
?

? R25s for Courses for Skills Development, RFA-HG- 14-008
?

? Development of creative educational activities with a primary focus on Courses for Skills Development. ?R25
for Open Educational Resources, RFA-HG-14- 009
?

? Development of open educational resources (OER) for use by large numbers of learners at all career levels,
with a primary focus on Curriculum or Methods Development.

52. 4:4: BD2K Centers of Excellence

? Two or more rounds of center awards
?

? 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)

Taking on the Role of Associate Director for Data Science at the NIH –
My Original Vision Statement

Source: http://pebourne.wordpress.com/2013/12/

On March 3, 2014 I will begin the job of Associate Director for Data Science (ADDS) at the National Institute of Health
(NIH). I will report directly to NIH Director, Dr. Francis Collins. When I originally applied for the position in April 2013 I
was asked to prepare a short vision statement. That statement follows here. It does not necessarily reflect what I will
attempt to accomplish in the job, but rather the way I was thinking about data science at the time of my application. In
the spirit of openness which I hope to bring to the position I include it here and invite your comments.

Technology, including information technology, has had a profound impact on health related research at all scales.
Witness the plummeting cost of sequencing and assembling a genome to the emergence of mobile health precipitated
by smart phones. Yet this is just the beginning. I believe the future of research into health and well-being is going to be
tied very much to our ability to sustain, trust, integrate, analyze/discover, disseminate/visualize and comprehend digital
data. The work of the National Library of Medicine and the National Center for Biotechnology Information (NCBI) has
been exemplary among the sciences in getting us this far, but it is just the beginning. Let me address each of these
issues. Two pages preclude a detailed discussion of how to fulfill the vision and I will hope to speak further about the
possibilities. Moreover, it precludes any discussion about the peculiarities, challenges and rewards associated with specific data types. Again, I hope to go beyond this generic discussion on the future of data.

Sustainability is the most critical, yet least addressed aspect of digital health, at least in academia. Sustainability cannot simply mean asking the funding agencies for more money as the data continues to grow at unprecedented rates. We need new business models (academia is a business), including public-private partnerships where private enterprise has been thinking about these problems for a while. We need to recognize that data sustainability is a global not a national problem and finally we need to begin to make informed decisions about what data to discard. Consider examples of the types of discussions that need to be had leading to subsequent policies and procedures, that need to be put in place. Discussions need to be had around business models that provide services atop of free open content that generate revenue to sustain that content. Discussions need to be had that review other global industries, e.g., banking and commerce to consider best (and worst) practices associated with global management of data. Lastly, we need to consider what data we need to sustain. That consideration begins with how we actually use data. To date the study of how data are actually utilized is in its infancy in academia. Funded data providers are required to give global statistics on data use, but this does not speak to how each element of data in a corpus is utilized and why. When we understand this better we can make informed decisions about what to discard with the understanding it could be regenerated later at a cost that makes storing it nonsensical. Data rich private sector companies need to be engaged in this discussion so academia can learn from their best practices.

Sustainability is also an institutional problem. Academic institutions are at this time rarely taking full advantage of their digital assets, including the biomedical data being generated by their faculty and students. The recent Moore and Sloan Foundation initiative (I was involved with this) was a departure in that it rewards institutions for best data science practices rather than individuals. Mechanisms that reward institutions for their careful stewardship and open accessibility of biomedical data should be considered. As should programs that support and promote data scientists in these institutions. Lack of growth paths and perceptions by faculty review committees need to be changed such that the value of institutional data scientists is elevated. Programs can be designed to support this.

Trust in the data has been the biggest factor in the success of the data and knowledge resources (databases and journals) I have been involved with over the years. Trust speaks to the security and quality of the data. Security is temporal and personal. What is secure today may not be secure with the analytical tools of tomorrow. What one person wants to keep secure another wants to make public so as to benefit others. We need to be flexible in our approach to security. Surprisingly quality is not something we pay enough attention to. Current modes of data and knowledge management (database and journals) lack sufficient feedback mechanisms to report on their content. Likewise, there is a data curation-query cycle that is mostly missing in current data management practices. Query of a corpus informs about outliers in that corpus. Such outliers may be discoveries they may be errors that can be corrected or discarded. We need to stimulate more inquiry about the trust in the data we are generating.

Integration of disparate data, often at different biological scales, is a major characteristic of current and future biomedical research discoveries. Optimizing such integration speaks to data representation, metadata, ontologies, provenance and so on. Aspects for which good technical solutions already exist, but where motivation and reward to create well formed datasets from which integration can occur are missing. Facilitating the cataloging and comparison of datasets is one mechanism for creating motivation among researchers, funding mandates another. Data sharing policies are a great step forward, when firmly in place the next step is towards how the data should be presented so that it may be optimally shared need to be put in place.
Analyze/Discover

Discovery informatics is in its infancy. Search engines are grappling with the need for deep search, but it is doubtful they will fulfill the needs of the biomedical research community when it comes to finding and analyzing the appropriate datasets. Let me cast the vision in a use case. As a research group winds down for the day algorithms take over, deciphering from the days on-line raw data, lab notes, grant drafts etc. underlying themes that are being explored by the laboratory (the lab’s digital assets). Those themes are the seeds of deep search to discover what is relevant to the lab that has appeared since a search was last conducted in published papers, public data sets, blogs, open reviews etc. Next morning the results of the deep search are presented to each member as a personalized view for further post processing. We have a long way to go here, but programs that incite groups of computer, domain and social scientists to work on these needs will move us forward.

Disseminate/Visualize/Comprehend

In 2005 I wrote an editorial asking the question, is a biological database really different to a biological journal? The answer then, as now, is no; what is different is the way their value is perceived. What has changed is the emergence of data journals and databases hiring more curators to extract information from the literature and adding it to databases (how cost ineffective is that!). In the world of digital scholarship the paper is a means to execute upon the underlying data and becomes a tool of interactive inquiry. Open access opens the door to these possibilities but there is much to be done.

What I have tried to do here is to simply introduce how I am thinking about the problems of biomedical research data and appreciate there is little in how these problems might be addressed. I hope to have that discussion.

About BD2K

Source: http://bd2k.nih.gov/about_bd2k.html

Mission Statement

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:

• Appropriate access to shareable biomedical data through technologies, approaches, and policies that enable and facilitate widespread data sharing, discoverability, management, curation, and meaningful re-use;
• Development of and access to appropriate algorithms, methods, software, and tools for all aspects of the use of Big Data, including data processing, storage, analysis, integration, and visualization;
• Appropriate protections for privacy and intellectual property;
• Development of a sufficient cadre of researchers skilled in the science of Big Data, in addition to elevating general competencies in data usage and analysis across the behavioral research workforce.

http://semanticommunity.info/Data_Science/Data_Culture_at_the_NIH
Updated: Fri, 18 Sep 2015 23:53:15 GMT
Powered by mindtouch
Overall, the focus of the BD2K initiative is the development of innovative and transforming approaches as well as tools for making Big Data and data science a more prominent component of biomedical research.

What is Big Data?

The term 'Big Data' is meant to capture the opportunities and challenges facing all biomedical researchers in accessing, managing, analyzing, and integrating datasets of diverse data types [e.g., imaging, phenotypic, molecular (including various ‘–omics’), exposure, health, behavioral, and the many other types of biological and biomedical and behavioral data] that are increasingly larger, more diverse, and more complex, and that exceed the abilities of currently used approaches to manage and analyze effectively. Big Data emanate from three sources: (1) a small number of groups that produce very large amounts of data, usually as part of projects specifically funded to produce important resources for use by the entire research community; (2) individual investigators who produce large datasets, often empowered by the use of readily available new technologies; and (3) an even greater number of sources that each produce small datasets (e.g. research data or clinical data in electronic health records) whose value can be amplified by aggregating or integrating them with other data.

Major challenges in using biomedical Big Data include:

- Locating data and software tools.
- Getting access to the data and software tools.
- Standardizing data and metadata.
- Extending policies and practices for data and software sharing.
- Organizing, managing, and processing biomedical Big Data.
- Developing new methods for analyzing & integrating biomedical data.
- Training researchers who can use biomedical Big Data effectively.

Organizational Structure

http://semanticommunity.info/Data_Science/Data_Culture_at_the_NIH

Updated: Fri, 18 Sep 2015 23:53:15 GMT
Powered by mindtouch™
Areas of BD2K

**Enabling Data Utilization (Click to Watch Video)**

- New policies that better encourage data and software sharing
- A catalog of research datasets that will enable researchers to find and cite datasets
- Community-based data and metadata standards

**Analysis Methods and Software**

- Development and hardening of software to meet needs of the biomedical research community
- Access to large-scale computing to enable data analysis on Big Data
- Dynamic community engagement of users and developers

**Enhancing Training**

- Increase number of computationally and quantitatively skilled biomedical trainees
- Strengthen the computational and quantitative skills of all biomedical researchers
- Make training available to NIH staff to enhance NIH review and program oversight

**Centers of Excellence**

- Investigator-initiated centers
- NIH-specified centers

**BD2K Executive Committee**

Vivien Bonazzi (NHGRI)
Lisa Brooks (NHGRI)
Jennifer Couch (NCI)
Allen Dearry (NIEHS)
Leslie Derr (OD)
Michelle Dun (NCI)
Maria Giovanni (NIAID)
Susan Gregurick (NIGMS)
Mark Guyer (NHGRI)
Lynda Hardy (NINR)
Mike Huerta (NLM)
Jennie Larkin (NHLBI)
Peter Lyster (NIGMS)
Scientific Data Council

**Acting Chair:**
Eric Green (Acting ADDS & NHGRI)

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

**Acting Executive Secretary:**
Allison Mandich (NHGRI)

---

**H1N1 Flu**

**Source:** [http://www.cdc.gov/h1n1flu/estimates...l_March_13.htm](http://www.cdc.gov/h1n1flu/estimates...l_March_13.htm)

CDC Estimates of 2009 H1N1 Influenza Cases, Hospitalizations and Deaths in the United States, April 2009 – March 13, 2010

This website is archived for historical purposes and is no longer being maintained or updated. For updated information on the current flu season, see the [CDC Seasonal Flu website](http://www.cdc.gov/flu/).

April 19, 2010, 1:00 PM ET

---

**Background**

Estimating the number of individual flu cases in the United States is very challenging because many people with flu don’t seek medical care and only a small number of those that do seek care are tested. More people who are hospitalized or...
die of flu-related causes are tested and reported, but under-reporting of hospitalizations and deaths occurs as well. For this reason CDC monitors influenza activity levels and trends and virus characteristics through a nationwide surveillance system and uses statistical modeling to estimate the burden of flu illness (including hospitalizations and deaths) in the United States.

When the 2009 H1N1 flu outbreak began in April 2009, CDC began tracking and reporting the number of laboratory-confirmed 2009 H1N1 cases, hospitalizations and deaths as reported by states to CDC. These initial case counts (which were discontinued on July 24, 2009), and subsequent ongoing laboratory-confirmed reports of hospitalizations and deaths, are thought to represent a significant undercount of the actual number of 2009 H1N1 flu cases in the United States. A paper in Emerging Infectious Diseases authored by CDC staff entitled “Estimates of the Prevalence of Pandemic (H1N1) 2009, United States, April–July 2009” reported on a study to estimate the prevalence of 2009 H1N1 based on the number of laboratory-confirmed cases reported to CDC. Correcting for under-ascertainment, the study found that every case of 2009 H1N1 reported from April – July represented an estimated 79 total cases, and every hospitalized case reported may have represented an average of 2.7 total hospitalized people. CDC then began working on a way to estimate, in an ongoing way, the impact of the 2009 H1N1 pandemic on the U.S. in terms of 2009 H1N1 cases, hospitalizations and deaths. CDC developed a method to provide an estimated range of the total number of 2009 H1N1 cases, hospitalizations and deaths in the United States by age group using data on flu associated hospitalizations collected through CDC’s Emerging Infections Program.

The Numbers

(Print table)

On November 12, 2009 CDC provided the first set of estimates on the numbers of 2009 H1N1 cases and related hospitalizations and deaths in the United States between April and October 17, 2009.

Estimates from April – October 17, 2009:

- CDC estimated that between 14 million and 34 million cases of 2009 H1N1 occurred between April and October 17, 2009. The mid-level in this range was about 22 million people infected with 2009 H1N1.
- CDC estimated that between about 63,000 and 153,000 2009 H1N1-related hospitalizations occurred between April and October 17, 2009. The mid-level in this range was about 98,000 H1N1-related hospitalizations.
- CDC estimated that between about 2,500 and 6,000 2009 H1N1-related deaths occurred between April and October 17, 2009. The mid-level in this range was about 3,900 2009 H1N1-related deaths.

Updated Estimates from April – November 14, 2009

Using the same methodology, CDC updated the estimates to include the time period from April through November 14, 2009 on December 10, 2009.

- CDC estimated that between 34 million and 67 million cases of 2009 H1N1 occurred between April and November 14, 2009. The mid-level in this range was about 47 million people infected with 2009 H1N1.
- CDC estimated that between about 154,000 and 303,000 2009 H1N1-related hospitalizations occurred between April and November 14, 2009. The mid-level in this range was about 213,000 H1N1-related hospitalizations.
• CDC estimated that between about 7,070 and 13,930 2009 H1N1-related deaths occurred between April and November 14, 2009. The mid-level in this range was about 9,820 2009 H1N1-related deaths.

Updated Estimates from April – December 12, 2009

Using the same methodology, CDC updated the estimates to include the time period from April through December 12, 2009.

• CDC estimates that between 39 million and 80 million cases of 2009 H1N1 occurred between April and December 12, 2009. The mid-level in this range is about 55 million people infected with 2009 H1N1.
• CDC estimates that between 173,000 and 362,000 2009 H1N1-related hospitalizations occurred between April and December 12, 2009. The mid-level in this range is about 246,000 H1N1-related hospitalizations.
• CDC estimates that between 7,880 and 16,460 2009 H1N1-related deaths occurred between April and December 12, 2009. The mid-level in this range is about 11,160 2009 H1N1-related deaths.

Updated Estimates from April 2009 – January 16, 2010

Using the same methodology, CDC updated the estimates to include the time period from April 2009 through January 16, 2010 on February 12, 2010.

• CDC estimates that between 41 million and 84 million cases of 2009 H1N1 occurred between April 2009 and January 16, 2010. The mid-level in this range is about 57 million people infected with 2009 H1N1.
• CDC estimates that between 183,000 and 378,000 H1N1-related hospitalizations occurred between April 2009 and January 16, 2010. The mid-level in this range is about 257,000 2009 H1N1-related hospitalizations.
• CDC estimates that between about 8,330 and 17,160 2009 H1N1-related deaths occurred between April 2009 and January 16, 2010. The mid-level in this range is about 11,690 2009 H1N1-related deaths.

Updated Estimates from April 2009 – February 13, 2010

Using the same methodology CDC has again updated the estimates to include the time period from April 2009 through February 13, 2010 on March 12, 2010.

• CDC estimates that between 42 million and 86 million cases of 2009 H1N1 occurred between April 2009 and February 13, 2010. The mid-level in this range is about 59 million people infected with 2009 H1N1.
• CDC estimates that between 188,000 and 389,000 H1N1-related hospitalizations occurred between April 2009 and February 13, 2010. The mid-level in this range is about 265,000 2009 H1N1-related hospitalizations.
• CDC estimates that between about 8,520 and 17,620 2009 H1N1-related deaths occurred between April 2009 and February 13, 2010. The mid-level in this range is about 12,000 2009 H1N1-related deaths.

Updated Estimates from April 2009 – March 13, 2010

Using the same methodology CDC has again updated the estimates to include the time period from April 2009 through March 13, 2010 on April 19, 2010.

• CDC estimates that between 43 million and 88 million cases of 2009 H1N1 occurred between April 2009 and March 13, 2010. The mid-level in this range is about 60 million people infected with 2009 H1N1.
• CDC estimates that between about 192,000 and 398,000 H1N1-related hospitalizations occurred between April 2009 and March 13, 2010. The mid-level in this range is about 270,000 2009 H1N1-related hospitalizations.
• CDC estimates that between about 8,720 and 18,050 2009 H1N1-related deaths occurred between April 2009 and March 13, 2010. The mid-level in this range is about 12,270 2009 H1N1-related deaths.
Note: Less than 5% of increases in the estimates from one reporting date to the next are the result of delayed reporting in cases, hospitalizations and deaths.

## CDC Estimates of 2009 H1N1 Cases and Related Hospitalizations and Deaths from April 2009 - March 13, 2010, By Age Group

<table>
<thead>
<tr>
<th>2009 H1N1</th>
<th>Mid-Level Range*</th>
<th>Estimated Range*</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Cases</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0-17 years</td>
<td>~19 million</td>
<td>~14 million to ~28 million</td>
</tr>
<tr>
<td>18-64 years</td>
<td>~35 million</td>
<td>~25 million to ~51 million</td>
</tr>
<tr>
<td>65 years and older</td>
<td>~6 million</td>
<td>~4 million to ~9 million</td>
</tr>
<tr>
<td><strong>Cases Total</strong></td>
<td>~60 million</td>
<td>~43 million to ~88 million</td>
</tr>
<tr>
<td><strong>Hospitalizations</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0-17 years</td>
<td>~86,000</td>
<td>~61,000 to ~127,000</td>
</tr>
<tr>
<td>18-64 years</td>
<td>~158,000</td>
<td>~112,000 to ~232,000</td>
</tr>
<tr>
<td>65 years and older</td>
<td>~26,000</td>
<td>~19,000 to ~39,000</td>
</tr>
<tr>
<td><strong>Hospitalizations Total</strong></td>
<td>~270,000</td>
<td>~192,000 to ~398,000</td>
</tr>
<tr>
<td><strong>Deaths</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0-17 years</td>
<td>~1,270</td>
<td>~900 to ~1,870</td>
</tr>
<tr>
<td>18-64 years</td>
<td>~9,420</td>
<td>~6,700 to ~13,860</td>
</tr>
<tr>
<td>65 years and older</td>
<td>~1,580</td>
<td>~1,120 to ~2,320</td>
</tr>
<tr>
<td><strong>Deaths Total</strong></td>
<td>~12,270</td>
<td>~8,720 to ~18,050</td>
</tr>
</tbody>
</table>

* Deaths have been rounded to the nearest ten. Hospitalizations have been rounded to the nearest thousand and cases have been rounded to the nearest million. [Exact numbers also are available.](http://semanticoncommunity.info/Data_Science/Data_Culture_at_the_NIH)

### Discussion

The latest estimates released on April 19, 2010, incorporate an additional four weeks of flu data (from February 13, 2010 through March 13, 2010) from the previous estimates released on March 12, 2010.
The latest estimates through March 13, 2010 show a relatively small increase in the total number of 2009 H1N1 cases, hospitalizations and deaths since the previous estimates posted on March 12, 2010. The additional four weeks of flu activity data added to derive these updated estimates correlate with a four week period of ongoing but generally low flu activity in the United States.

The United States experienced its first wave of 2009 H1N1 pandemic activity in the spring of 2009, followed by a second wave of 2009 H1N1 activity in the fall. Activity peaked during the second week in October and then declined quickly to below baseline levels in January. The early rise in flu activity in October is in contrast to non-pandemic influenza seasons. Influenza activity usually peaks in January, February or March. (See graph of peak influenza activity by month in the United States from 1976-2009.) Because 2009 H1N1 activity peaked in late October, the greatest increase in the number of estimated 2009 H1N1 cases, hospitalizations and deaths occurred during the period of April through November 14, 2009. The estimates provided for the subsequent four weeks (through December 12, 2009) showed a modest increase in the total number of 2009 H1N1 cases, hospitalizations and deaths and correlated with decreasing but persistent flu activity nationwide. The estimates updated on February 12, 2010 with data from December 13, 2009 through January 16, 2010, correlate with a five week period of generally low flu activity. Although flu activity leveled off and was generally low in the United States from January 17 – February 13, 2010, 2009 H1N1 cases, hospitalizations and deaths continued to occur, though in much smaller increments than during fall 2009. Overall, flu activity remained generally low in the United States from February 14 – March 13, 2010; however, the Southeast United States reported an increase in localized flu activity during this time period. Almost all flu continues to be 2009 H1N1.

Visits to doctors for influenza-like illness (ILI) in general were low nationally between February 14, 2010 and March 13, 2010 (Reporting weeks 7, 8, 9, and 10) and were below the national baseline during all four weeks. ILI is also looked at by regions in the United States. While ILI was consistently low nationwide, during the four-week time period added to the latest estimates, certain regions of the United States – particularly Region 4, which is representative of the Southeast United States – reported ILI activity above their regional baselines between February 14 and March 13, 2010, indicating ongoing localized flu activity, almost all of it thought to be 2009 H1N1. No states reported widespread activity throughout this period; however, a few states continued to report regional flu activity; again an indication of ongoing localized 2009 H1N1 outbreaks. The proportion of deaths attributed to pneumonia and influenza (P&I) based on the 122 Cities Report remained below the epidemic threshold throughout this time period.

Flu activity peaked in October and then declined to below baseline levels in January. At this time, most flu continues to be 2009 H1N1. Flu activity, caused by either 2009 H1N1 or seasonal flu viruses, may rise and fall, but is expected to continue for weeks in the United States.

The data by age provided in the updated estimates continues to confirm that people younger than 65 years of age are more severely affected by this disease relative to people 65 and older compared with seasonal flu. With seasonal influenza, about 60 percent of seasonal flu-related hospitalizations and 90 percent of flu-related deaths occur in people 65 years and older. With 2009 H1N1, approximately 90% of estimated hospitalizations and 87% of estimated deaths from April through March 13, 2009 occurred in people younger than 65 years old based on this method. CDC is continuing to recommend vaccination against 2009 H1N1 at this time for all people 6 months and older, including those people 65 years of age and older because severe illness and deaths have occurred in this age group.

According to the Morbidity and Mortality Weekly Report (MMWR) issued on April 2, 2010, an estimated 72-81 million people have been vaccinated against 2009 H1N1 as of the end of February 2010, with more than one-third of children and nearly one-fifth of U.S. adults having received the vaccine. Counting all doses of vaccine reported, an estimated 81-91 million doses of 2009 H1N1 vaccine were administered. When the numbers of people vaccinated against 2009
H1N1 is combined with the number of people previously infected with 2009 H1N1, a significant number of people in the United States likely have immunity to the 2009 H1N1 virus. However, with a population of more than 300 million in this country, a substantial number of people likely remain susceptible to 2009 H1N1, which continues to circulate at this time. Ongoing vaccination of people with certain health conditions is particularly important because most cases of serious 2009 H1N1 illness (e.g., hospitalizations) have occurred in people with underlying medical conditions. (See "2009 H1N1 Flu: Underlying Health Conditions among Hospitalized Adults and Children.") Health conditions that increase the risk of being hospitalized from 2009 H1N1 include lung disease like asthma or chronic pulmonary disease (COPD), diabetes, heart, or neurologic disease and pregnancy. In addition, minority populations have been harder-hit by the 2009 H1N1 pandemic than non-minority groups. (See "Information on 2009 H1N1 impact by Race and Ethnicity.") There also is growing evidence to support early concerns that people who are morbidly obese are at greater risk of serious 2009 H1N1 complications.

This methodology and the resulting estimates continue to underscore the substantial under-reporting that occurs when laboratory-confirmed outcomes are the sole method used to capture hospitalizations and deaths. CDC has maintained since the beginning of this outbreak that laboratory-confirmed data on hospitalizations and deaths reported to CDC is an underestimation of the true number that have occurred because of incomplete testing, inaccurate test results, or diagnosis that attribute hospitalizations and deaths to other causes, for example, secondary complications to influenza. (Information about surveillance and reporting for 2009 H1N1 is available at Questions and Answers: Monitoring Influenza Activity, Including 2009 H1N1.)

The estimates derived from this methodology provide the public, public health officials and policy makers a sense of the health impact of the 2009 H1N1 pandemic. While these numbers are an estimate, CDC feels that they present a fuller picture of the burden of 2009 H1N1 disease on the United States.

CDC will continue to use weekly data from systems that comprise the National Influenza Surveillance System to monitor geographic, temporal and virologic trends in influenza in the nation.

---

Graphics

(Print graphics)

The graphs below illustrate CDC’s estimates of cumulative 2009 H1N1 cases, hospitalizations and deaths in the United States by age group from April 2009 - March 13, 2010. The vertical black lines represent the range in 2009 H1N1 estimates for each time period. To see the tables associated with the earlier time periods for which CDC provided estimates, please see the April 2009 – February 13, 2010, the April 2009 – January 16, 2010, the April – December 12, 2009 estimates, the April – November 14, 2009 estimates, and the April – October 17, 2009 estimates.

Graph A provides a summary illustration of the various estimates for 2009 H1N1 cases, made over time.
Graph A shows the cumulative estimated 2009 H1N1 cases by age group (0-17 years old, 18-64 years old, and 65 years and older) in the United States for each of the time periods that CDC provided case estimates and illustrates that people in the 18-64 years age group were most heavily impacted by 2009 H1N1 disease followed by people in the 0-17 years age group. People 65 years of age and older were relatively less affected by 2009 H1N1 illness.

Graph B below shows the total cumulative 2009 H1N1 cases (across all age groups) reported for each of the time periods that CDC provided case estimates.

The curved black line in Graph B depicts the increase in 2009 H1N1 cases per the midpoint value of the estimates for each reporting period for which CDC provided 2009 H1N1 case estimates. The curved line in Graph B shows that the greatest increase in 2009 H1N1 cases occurred between October 17, 2009 and November 14, 2009, which correlates with the peak of the fall-winter wave of 2009 H1N1 activity in the United States.

Graphs C and D below display estimates of 2009 H1N1 related hospitalizations in the United States.
Graph C shows cumulative estimated 2009 H1N1 hospitalizations by age group (0-17 years old, 18-64 years old, and 65 years and older) in the United States for each of the time periods that CDC provided estimates of hospitalizations, and illustrates, again, that people 18-64 years of age were most impacted by serious illness (including hospitalizations), followed by people in the 0-17 years old age group. Again, people 65 and older were relatively less affected by 2009 H1N1 hospitalizations than people in other age groups.

Graph D shows the total cumulative 2009 H1N1 hospitalizations reported for each of the time periods that CDC provided estimates of hospitalizations. The curved line in Graph D depicts the increase in 2009 H1N1 hospitalizations per the midpoint value of the estimates of each time period for which CDC provided estimates of hospitalizations. The curved line in Graph D shows that the greatest increase in 2009 H1N1 hospitalizations occurred between October 17, 2009 and November 14, 2009.

Graphs E and F below display estimates related to 2009 H1N1 deaths in the United States.
Graph E shows the cumulative estimated 2009 H1N1 deaths by age group (0-17 years old, 18-64 years old, and 65 years and older) in the United States for each of the time periods that CDC provided estimates of deaths and illustrates, again, that people in the 18-64 years age group were relatively more affected by 2009 H1N1 related deaths than people in other age groups.

Graph F shows the total cumulative 2009 H1N1 deaths reported for each of the time periods that CDC provided estimates of deaths. The curved line in Graph F depicts the increase in 2009 H1N1 deaths per the mid-point value of the estimations for each reporting period for which CDC provided estimates of deaths. The curved line in Graph F shows that the greatest increase in 2009 H1N1 deaths occurred between October 17, 2009 and November 14, 2009.

**Method to Estimate 2009 H1N1 Cases, Hospitalizations and Deaths**

CDC has developed a method to provide an estimated range of the total number of 2009 H1N1 cases, hospitalizations and deaths in the United States since April, 2009, as well as a breakdown of these estimates by age groups. This method uses data on influenza-associated hospitalizations collected through CDC’s Emerging Infections Program (EIP), which conducts surveillance for laboratory-confirmed influenza-related hospitalizations in children and adults in 62 counties covering 13 metropolitan areas of 10 states. To determine an estimated number of 2009 H1N1 hospitalizations...
nationwide, the EIP hospitalization data are extrapolated to the entire U.S. population and then corrected for factors that may result in under-reporting using a multiplier from "Estimates of the Prevalence of Pandemic (H1N1) 2009, United States, April–July 2009."

The lower and upper hospitalization estimates also are calculated using the EIP hospitalization data. The national hospitalization estimates are then used to calculate deaths and cases. Deaths are calculated by using the proportion of laboratory-confirmed deaths to hospitalizations reported through CDC's web-based Aggregate Hospitalization and Death Reporting Activity (AHDRA). Cases are estimated using multipliers derived from "Estimates of the Prevalence of Pandemic (H1N1) 2009, United States, April–July 2009."

The lower and upper end of the ranges for deaths and cases are derived from the lower and upper hospitalization estimates. The methods used to estimate impact may be modified as more information becomes available. More information about this methodology is available.

Throughout the remainder of the 2009 H1N1 pandemic CDC will update the range of estimated 2009 H1N1 cases, hospitalizations and deaths every three or four weeks. While EIP data is reported weekly during influenza season, because the system is based on reviews of patients medical charts there are sometimes delays in reporting and it can take some time for all the data to fill in. CDC will continue to provide weekly reports of influenza activity each Friday in FluView and will update the 2009 H1N1 Situation Update each Friday as well.

The estimated ranges of cases, hospitalizations and deaths generated by this method provide a sense of scale in terms of the burden of disease caused by 2009 H1N1. It may never be possible to validate the accuracy of these figures. The true number of cases, hospitalizations and deaths may lie within the range provided or it’s also possible that it may lie outside the range. An underlying assumption in this method is that the level of influenza activity (based on hospitalization rates) in EIP sites matches the level of influenza like illness (ILI) activity across the states.

This methodology is not a predictive tool and cannot be used to forecast the number of cases, hospitalizations and deaths that will occur going forward over the course of the pandemic because they are based on actual surveillance data.

Background Emerging Infections Program

The Emerging Infections Program (EIP) Influenza Project conducts surveillance for laboratory-confirmed influenza-related hospitalizations in children and adults in 62 counties covering 13 metropolitan areas of 10 states. (This includes San Francisco, CA; Denver, CO; New Haven, CT; Atlanta, GA; Baltimore, MD; Minneapolis/St. Paul, MN; Albuquerque, NM; Santa Fe, NM, Las Cruces, NM; Albany, NY; Rochester, NY; Portland, OR; and Nashville, TN.) Cases are identified by reviewing hospital laboratory and admission databases and infection control logs for children and adults with a documented positive influenza test conducted as a part of routine patient care. EIP estimated hospitalization rates are reported every week during the flu season. More information about the Emerging Infections Program is available.
Seasonal Influenza-Associated Hospitalizations in the United States

An average estimated 200,000 flu-related hospitalizations occur in the United States each year, with about 60 percent of these hospitalizations occurring in people 65 years and older.

Background: A study conducted by CDC and published in the *Journal of American Medical Association (JAMA)* in September 2004 provided information on the number of people in the United States that are hospitalized from seasonal influenza-related complications each year. The study concluded that, on average, more than 200,000 people in the United States are hospitalized each year for respiratory and heart conditions illnesses associated with seasonal influenza virus infections. The study looked at hospital records from 1979 to 2001. In 1979, there were 120,929 flu-related hospitalizations. The number was lower in some years after that, but there was an overall upward trend. During the 1990s, the average number of people hospitalized was more than 200,000 but individual seasons ranged from a low of 157,911 in 1990-91 to a high of 430,960 in 1997-98.

More information about seasonal flu-related hospitalizations is available.

Seasonal Influenza-Associated Deaths

Flu-associated mortality varies by season because flu seasons often fluctuate in length and severity. CDC estimates that about 36,000 people died of flu-related causes each year, on average, during the 1990s in the United States with 90 percent of these deaths occurring in people 65 years and older. This includes people dying from secondary complications of the flu.

Background: This estimate came from a 2003 *Journal of the American Medical Association (JAMA)* study, which looked at the 1990-91 through the 1998-99 flu seasons and is based on the number of people whose underlying cause of death on their death certificate was listed as a respiratory or circulatory disease. During these years, the number of estimated deaths ranged from 17,000 to 52,000. This number was corroborated in 2009, when a CDC-authored study was published in the journal *Influenza and Other Respiratory Viruses*. This study estimated seasonal flu-related deaths comparing different methods, including the methods used in the 2003 JAMA study but using more recent data. Results from this study showed that during this time period, 36,171 flu-related deaths occurred per year, on average. More information about how CDC estimates seasonal flu-related deaths is available.

Under-Counting of Flu-Related Deaths

CDC does not know exactly how many people die from seasonal flu each year. There are several reasons for this:

These are some of the reasons that CDC and other public health agencies in the United States and other countries use statistical models to estimate the annual number of seasonal flu-related deaths. (Flu deaths in children were made a nationally notifiable condition in 2004, and since then, states have reported flu-related child deaths in the United States through the Influenza Associated Pediatric Mortality Surveillance System).
First, states are not required to report individual seasonal flu cases or deaths of people older than 18 years of age to CDC.

Second, seasonal influenza is infrequently listed on death certificates of people who die from flu-related complications.

Third, many seasonal flu-related deaths occur one or two weeks after a person’s initial infection, either because the person may develop a secondary bacterial co-infection (such as a staph infection) or because seasonal influenza can aggravate an existing chronic illness (such as congestive heart failure or chronic obstructive pulmonary disease).

Also, most people who die from seasonal flu-related complications are not tested for flu, or they seek medical care later in their illness when seasonal influenza can no longer be detected from respiratory samples. Influenza tests are most likely to detect influenza if performed soon after onset of illness. In addition, some patients may be tested for influenza using rapid tests that are only moderately sensitive and result in some false-negative results.

Some persons who die of influenza-related complications may not have classic influenza symptoms and thus not be recognized as possibly being influenza-related.

For these reasons, many flu-related deaths may not be recorded on death certificates.
The TB-drugome is a structural proteome-wide drug-target network that includes 274 drugs and 1730 proteins from *Mycobacterium tuberculosis* (*M. tb*). It was constructed by associating the putative ligand binding sites of the *M. tb* proteins with the known binding sites of approved drugs for which structural information was available. The premise being that two entirely unrelated proteins can bind similar ligands if they share similar ligand binding sites, since the ligand binding site can be considered as a negative image of the ligand. In this way, an *M. tb* protein can be connected to a drug through the drug target, irrespective of whether that protein target is from human or another organism. The binding site comparison software SMAP ([http://funsite.sdsc.edu](http://funsite.sdsc.edu)) was used for this purpose in an all-drug-against-all-target manner. For each identified drug-target pair, the atomic details of the interaction were studied using protein-ligand docking. The TB-drugome reveals that approximately one-third of the drugs examined have the potential to be repositioned to treat tuberculosis and that many currently unexploited *M. tb* receptors may be druggable and could serve as novel anti-tubercular targets.
Download TB-drugome

- TB-Drugome.tar.gz (220 M)

Summary of Files and Directories

The compressed file includes the following files and directories:

**drug_site_info.tab**

This file contains information about the 274 approved drugs that were identified in the PDB. For each drug, its name, PDB ligand, isomeric SMILES string and known targets are listed, and the PDB codes of the protein structures with which it has been crystallized are given.

**pdb_structure_info.tab**

This file contains information about the *M.tb* proteins with solved structure(s) in the RCSB PDB that were used in TB-drugome. For each protein, the gene name (if available), gene accession number, protein name and corresponding PDB codes are given.

**homology_model_info.tab**

This file contains information about the reliable homology models of *M.tb* proteins from ModBase that were used in TB-drugome. For each homology model, the ModBase model code is given, as well as the gene accession number, gene name and description of the *M.tb* protein. N.B. Further information about each homology model can be found on the ModBase website.

**SMAP_eHiTs_pdb_structure_info.tab**
This file contains a list of the cross-fold drug-target pairs with a SMAP P-value < 1.0e-5, for solved *M.tb* structures only. For each pair, information about the drug and target structures is given, as well as the corresponding SMAP P-value (indicating the significance of the binding site similarity) and eHiTS energy score (from docking the drug into the predicted binding site in the *M.tb* protein).

**SMAP_eHiTs_homology_model_info.tab**

This file contains a list of the cross-fold drug-target pairs with a SMAP P-value < 1.0e-5, for homology models of *M.tb* proteins only. For each pair, information about the drug and target structures is given, as well as the corresponding SMAP P-value (indicating the significance of the binding site similarity) and eHiTS energy score (from docking the drug into the predicted binding site in the *M.tb* protein).

**eHiTs_docking_pdb_structures**

This directory contains the eHiTS predicted binding poses (in SDF format) for all cross-fold drug-target pairs (solved *M.tb* structures only) with SMAP P-values < 1.0e-5. N.B. A blank file indicates that docking failed for that particular drug-target pair. Each drug binding pose can be viewed within its corresponding *M.tb* protein structure using molecular graphics software. The biological unit PDB files of all solved *M.tb* structures have also been provided for this purpose. In addition, the structures of the proteins with which the drugs were originally crystallized in the PDB have also been provided. These structures have been aligned onto the relevant solved *M.tb* protein structures using SMAP.

**eHiTs_docking_homology_models**

This directory contains the eHiTS predicted binding poses (in SDF format) for all cross-fold drug-target pairs (homology models of *M.tb* proteins only) with SMAP P-values < 1.0e-5. N.B. A blank file indicates that docking failed for that particular drug-target pair. Each drug binding pose can be viewed within its corresponding *M.tb* homology model using molecular graphics software. The homology model PDB files have also been provided for this purpose. In addition, the structures of the proteins with which the drugs were originally crystallized in the PDB have also been provided. These structures have been aligned onto the relevant homology models using SMAP.

**Contact**

Prof. Philip E. Bourne (bourne@sdsc.edu)
Frequently Asked Questions

Source: http://bd2k.nih.gov/faqs_trainingFOA.html

Reports

• BD2K Training Workshop Report
• RFI Summary for BD2K Training

General

The FOAs are at

When I apply for a BD2K grant, do I need to specify a targeted NIH Institute?

No. This is a trans-NIH initiative. For new applications, you do not choose an institute or center to target or apply to. Revision applications (see NOT-HG-14-022 and NOT-HG-14-023) will be assigned to the IC of the parent grant. The other BD2K training applications will be administered as a group and scientifically monitored by a trans-NIH committee.

Are applications to BD2K FOAs restricted to a particular type of data?

These FOAs are open to a wide range of data types (including EHR, imaging, phenotypic, molecular (including -omics), clinical, behavioral, and environmental) as well as a wide range of diseases.
Can I submit an application to this grant that I have previously submitted to another FOA?

Because this is an RFA, an application that was already submitted to NIH can be submitted in response to this announcement as a new application.

However, overlapping applications will result in one being withdrawn. A concurrent submission of an application to both an RFA and a PA is considered an overlapping application.

For more details, please see: [http://public.csr.nih.gov/ApplicantResources/ReceiptReferral/Pages/Evaluation-of-Unallowable-Resubmission-and-Overlapping-Applications.aspx](http://public.csr.nih.gov/ApplicantResources/ReceiptReferral/Pages/Evaluation-of-Unallowable-Resubmission-and-Overlapping-Applications.aspx) Also note that applications must directly address the science described in the BD2K RFA in order to be considered responsive.

KO1

There seems to be a typo in the program announcement, as it states "The NIH will contribute up to per year", without specifying the amount. Could you please clarify?

This section should read: “The NIH will contribute up to $167,000 per year” and has been addressed in NOT-HG-14-021.

Is this opportunity targeting investigators whose project(s) are applications of existing methods or development of new technologies?

The aim of the initiative is to support additional mentored training of scientists who will gain the knowledge and skills necessary to be independent researchers as well as to work in a team environment to develop new Big Data technologies, methods, and tools applicable to basic and clinical research.

Eligibility

If I already have a background in computer science, high performance computing, data mining, bioinformatics, or a related field, is my background considered too close to be eligible to apply to this program?

In general, no background is too close to biomedical data science to apply to this program since “Candidates may enter the program from various backgrounds … [including] biomedical data scientists who already have some background in areas relevant to Big Data Science but who want to gain further expertise.” However, when a candidate is a former PD/PI on a research grant or K award, the candidate must demonstrate a significant shift in research focus to Big Data science.

My background is exclusively/preponderantly in one of the the three areas mentioned in the RFA: am I eligible?

Yes. Since biomedical data science emphasizes: 1) computer science/informatics, 2) statistics/mathematics, and 3) biomedical science, substantial career development (and mentor input) in the areas that complement your background
should be strongly considered. Also keep in mind that the focus of this award is to support independent researchers who will develop technology, methods, and tools.

If I am already tenured, am I eligible to apply?

There is no upper limit on the level of seniority, as long as, when a candidate is a former PD/PI on a research grant or K award, the candidate demonstrates that the proposed work represents a significant shift in research focus.

If I am a postdoc, am I eligible to apply?

Yes, postdocs are eligible to apply as long as 1) their institution allows it and 2) “at the time of award, the candidate must have a ‘full-time’ appointment at the academic institution that is the applicant institution.”

Am I eligible to apply if I already have an R01 (or equivalent)?

Eligibility to apply to this K01 is independent of other research support: “While former PDs/PIs on NIH research project (R01), program project (P01), center grants (P50), sub-projects of program project (P01), sub-projects of center grants (P50), other career development awards (K–awards), or the equivalent NIH or non-NIH grants are eligible to apply, they must demonstrate their commitment to a future career as a full-time biomedical Big Data scientist and a significant shift in research focus to Big Data Science."

Although you can submit an application while you have active R01-equivalent support (including a NCE and including from sources other than NIH), you will need to address both potential overlap and how you will simultaneously satisfy the K01 and R01 requirements. This should be discussed with the program official for your R01 as well as the BD2K training group.

If I do not yet have a green card, am I eligible to apply? How will it affect the review of my proposal?

Yes, you can apply, but you must have received your green card for an award to be made. Green card status is not a review criterion, so it will not have any impact on review.

If I intend to study overseas during part of the award period of this grant, am I eligible to be funded during that time?

Foreign components are not allowed under this K01. Following the link in the FOA, foreign components include

"...he performance of any significant scientific element or segment of a project outside of the United States, either by the grantee or by a researcher employed by a foreign organization, whether or not grant funds are expended."

Mentors
My background is exclusively/preponderantly in one of the three areas mentioned in the FOA: what fields should my mentors be from?

Since biomedical data science emphasizes: 1) computer science/informatics, 2) statistics/mathematics, and 3) biomedical science, substantial mentor input (and career development) in the areas that complement your background should be strongly considered.

Can my mentor be from outside my campus?

Yes, you may choose a mentor from a different institution. However, reviewers might question how you will overcome the difficulties of working across institutions.

Can a mentor be from a foreign country?

The RFA states: “Foreign components, as defined in the NIH Grants Policy Statement, are not allowed.” It is very unlikely that a mentor from a foreign country would not be considered a foreign component, but you should review the definition of “foreign component” within the linked document to determine whether your mentor relationship and proposed research meets the foreign component criterion.

Writing and Submitting an Application

Must my training plan include coursework?

As stated in the RFA, “The [career development] plan should include relevant coursework to provide knowledge and skills in the three areas relevant to Big Data Science.” However, “the career development plan must be tailored to the needs of the individual candidate.” If you can demonstrate that you have sufficient knowledge in all of these three areas beyond the level of available coursework, you may argue that case in your application. We expect the situation that coursework is not needed to be very rare.

I currently hold a full-time position at an institution, and in the course of this year, I will be taking a position at another institution. Which institution should be the applicant institution?

First note that from RFA-HG-14-007, “At the time of award, the candidate must have a “full-time” appointment at the academic institution that is the applicant institution.” Therefore, if you expect to be at the new institution at the time of award, the application should come from there.

In addition, you need to check with the submitting institution because it may have internal policies that do not allow the submission, based on your provisional status.

Can letters of reference come from my proposed mentors?

No. Please see page 136 of this link: http://grants.nih.gov/grants/funding/424/SF424_RR_Guide_General_Adobe_Verb.pdf#page=135. "The letters should be from individuals not directly involved in
the application, but who are familiar with the applicant’s qualifications, training, and interests. The mentor/co-mentor(s) of the application cannot be counted toward the three required references."

**Can I assume that the reviewers will be familiar with the type of Big Data and the Big Data problem in my project?**

Although the reviewers will be experienced in Big Data, your application should include information about both the data and the significance of the problem(s) you propose to work on.

**Miscellaneous**

**Can I devote less than 75% effort to this award?**

This RFA will not accommodate a commitment of less than 75% effort: "The level of effort should be a minimum of 9 person months or 75% of full-time professional effort."

**Is there a max percent effort that the K01 will fund? My annual salary is under $167,000, so will the K01 fund 100% of my effort or do I need a minimum amount of institutional support?**

The BD2K K01 will fund up to 100% of your salary or the salary cap, whichever is smaller.

**Can I delay the award or take a leave of absence?**

Delays, adjustments, and leaves of absence will be decided on a case-by-case basis, once a decision to fund an application has been made, and could include a modification of the requested award length. See the Grants Policy Statement for more details about the following policy: “Leave without award support may not exceed 12 months. Such leave requires prior written approval of the awarding component and will be granted only with justification.”

**R25**

**Can this grant be used to expand an existing program at my institution to accommodate more students?**

The courses R25 can complement existing programs, but it must be distinct from such programs. Also, note that this FOA aims to support broad reaching programs that serve learners from multiple institutions, not just a single institution.

**Would it be responsive to create a course on one of the major MOOC providers, e.g. Coursera, Udacity, or EdX?**

MOOCs are responsive to the R25 FOA and can be particularly useful for dissemination (especially if the MOOC content is available at any time). The R25 FOAs encourage a wide dissemination so that more researchers may benefit.
Does this grant permit expenditures for computing costs?

Computational costs may be considered supplies or “other program-related expenses” and may be included in the proposed budget. These expenses must be justified as specifically required by the proposed program and must not duplicate items generally available at the applicant institution.

Can a training project, which was submitted as a part of a BD2K Centers application, also be submitted as an R25?

In general, the same science cannot be put into two NIH applications under review at the same time. Ordinarily, you have to wait until you get the summary statement from the first application before resubmitting the same science in another application. However, an exception is posted here: http://public.csr.nih.gov/ApplicantResources/ReceiptReferral/Pages/Evaluation-of-Unallowable-Resubmission-and-Overlapping-Applications.aspx, which allows a research application (e.g. R25) identical to a subproject in a Centers application to be submitted concurrently. Please note, though, that if both a BD2K Centers application and an overlapping BD2K R25 application receive potentially fundable scores, the Centers application will take precedence since training is a required component of the BD2K Centers. Also note that the NIH will not pay for the same science twice.

Would a proposal for training in the Responsible Conduct of Research (e.g ethics or reproducible research) be responsive to this FOA?

RCR topics such as ethics research and reproducible research are certainly relevant to the BD2K program and would be responsive to this FOA, as long as the RCR training focuses on the aspects unique to Big Data. In addition, highlighting both the biomedical application(s) and the RCR/ethics considerations, rather than simply focusing on RCR/ethics/reproducible research, may be considered more competitive by the reviewers.

Training Programs

Can an institution submit more than one application for BD2K training grants?

In February 2014, NIH issued 3 notices of intent to publish funding opportunity announcements for training in Biomedical Big Data Science:

• NOT-HG-14-011 for new T32 training programs
• NOT-HG-14-022 for revisions to active T32 training programs
• NOT-HG-14-023 for revisions to active T15 training programs

NOT-HG-14-011 states that only one application may be submitted per institution, i.e. one new application may be submitted per institution. All of the notices state that:

Given the limited amount of funds for these combined initiatives, institutions are encouraged to consider combining the (institutional) expertise in Big Data and submitting only one (single) application, whether for: (a) a new T32 institutional training grant or (b) a revision to a T32 or T15 institutional training grant.
To reiterate, no institution may submit more than one application for a new T32 Biomedical Big Data Science training program. While institutions are permitted to submit more than one revision application, they are strongly urged instead to put together the single best application possible, whether it is for a new training program or a revision of an existing training program.

**Flu Data**

**Seasonal Flu**


Also see: [http://www.cdc.gov/flu/weekly/pastreports.htm](http://www.cdc.gov/flu/weekly/pastreports.htm)

**FluView Interactive: Influenza Surveillance Data the Way You Want it!**

Source: [http://www.cdc.gov/flu/weekly/fluviewinteractive.htm](http://www.cdc.gov/flu/weekly/fluviewinteractive.htm)

The CDC FluView report provides weekly influenza surveillance information in the United States. These applications were developed to enhance the weekly FluView report by better facilitating communication about influenza with the public health community, clinicians, scientists, and the general public. This series of dynamic visualizations allow any Internet user to access influenza information collected by CDC’s monitoring systems.

Influenza surveillance data from the 1997-1998 through current season from the U.S. World Health Organization (WHO) and National Respiratory and Enteric Virus Surveillance System (NREVSS) collaborating laboratories and U.S. Outpatient Influenza-like Illness Surveillance Network (ILINet) can be accessed through the [FluView Interactive website](http://gis.cdc.gov/grasp/fluview/flu...dashboard.html).

The finalized data for previous seasons presented in FluView Interactive may differ from the preliminary data published in the weekly reports due to the receipt of additional data after publication.

**FluView Surveillance**

Source: [http://gis.cdc.gov/grasp/fluview/flu...dashboard.html](http://gis.cdc.gov/grasp/fluview/flu...dashboard.html)

**My Note: See multiple Download Data**
Fluview Hospitalizations

Source: [http://gis.cdc.gov/GRASP/Fluview/FluHospRates.html](http://gis.cdc.gov/GRASP/Fluview/FluHospRates.html)

My Note: See Download Data

ILI Activity Map


My Note: See Download Data
Pediatric Mortality

Source: [http://gis.cdc.gov/GRASP/Fluview/PedFluDeath.html](http://gis.cdc.gov/GRASP/Fluview/PedFluDeath.html)

My Note: See Download Data

---

Google Flu

**Flu Trends**

Source: [http://www.google.org/flutrends/](http://www.google.org/flutrends/)

We’ve found that certain search terms are good indicators of flu activity. Google Flu Trends uses aggregated Google search data to estimate flu activity.

See: [How does this work?](http://www.google.org/flutrends/how-does-it-work) and [FAQ](http://www.google.org/flutrends/faq)
Download world flu activity data  My Note: Got text file - see if can read in Spotfire

Animated flu trends for Google Earth  My Note: Got KMZ file

Compare flu trends across regions in Public Data Explorer  My Note: Got interface for country search

Learn more about the research behind Google Flu Trends

Read the article published by Nature, Detecting influenza epidemics using search engine query data

HTML | PDF (PDF)

Download the Google Flu Trends estimates for the world

Detecting influenza epidemics using search engine query data

Source: http://www.nature.com/nature/journal...ture07634.html (PDF)

Letter

Nature 457, 1012-1014 (19 February 2009) | doi:10.1038/nature07634; Received 14 August 2008; Accepted 13 November 2008; Published online 19 November 2008; Corrected 19 February 2009

Jeremy Ginsberg1, Matthew H. Mohebbi1, Rajan S. Patel1, Lynnette Brammer2, Mark S. Smolinski1 & Larry Brilliant1

Google Inc., 1600 Amphitheatre Parkway, Mountain View, California 94043, USA
Centers for Disease Control and Prevention, 1600 Clifton Road, NE, Atlanta, Georgia 30333, USA
Correspondence to: Matthew H. Mohebbi1 Correspondence and requests for materials should be addressed to J.G. or M.H.M. (Email: flutrends-support@google.com).
Abstract

Seasonal influenza epidemics are a major public health concern, causing tens of millions of respiratory illnesses and 250,000 to 500,000 deaths worldwide each year. In addition to seasonal influenza, a new strain of influenza virus against which no previous immunity exists and that demonstrates human-to-human transmission could result in a pandemic with millions of fatalities. Early detection of disease activity, when followed by a rapid response, can reduce the impact of both seasonal and pandemic influenza. One way to improve early detection is to monitor health-seeking behaviour in the form of queries to online search engines, which are submitted by millions of users around the world each day. Here we present a method of analysing large numbers of Google search queries to track influenza-like illness in a population. Because the relative frequency of certain queries is highly correlated with the percentage of physician visits in which a patient presents with influenza-like symptoms, we can accurately estimate the current level of weekly influenza activity in each region of the United States, with a reporting lag of about one day. This approach may make it possible to use search queries to detect influenza epidemics in areas with a large population of web search users.

Introduction

Traditional surveillance systems, including those employed by the U.S. Centers for Disease Control and Prevention (CDC) and the European Influenza Surveillance Scheme (EISS), rely on both virologic and clinical data, including influenza-like illness (ILI) physician visits. CDC publishes national and regional data from these surveillance systems on a weekly basis, typically with a 1-2 week reporting lag.

In an attempt to provide faster detection, innovative surveillance systems have been created to monitor indirect signals of influenza activity, such as call volume to telephone triage advice lines and over-the-counter drug sales. About 90 million American adults are believed to search online for information about specific diseases or medical problems each year, making web search queries a uniquely valuable source of information about health trends. Previous attempts at using online activity for influenza surveillance have counted search queries submitted to a Swedish medical website, visitors to certain pages on a U.S. health website, and user clicks on a search keyword advertisement in Canada. A set of Yahoo search queries containing the words “flu” or “influenza” were found to correlate with virologic and mortality surveillance data over multiple years.

Our proposed system builds on these earlier works by utilizing an automated method of discovering influenza-related search queries. By processing hundreds of billions of individual searches from five years of Google web search logs, our system generates more comprehensive models for use in influenza surveillance, with regional and state-level estimates of influenza-like illness (ILI) activity in the United States. Widespread global usage of online search engines may enable models to eventually be developed in international settings.

By aggregating historical logs of online web search queries submitted between 2003 and 2008, we computed time series of weekly counts for 50 million of the most common search queries in the United States. Separate aggregate weekly counts were kept for every query in each state. No information about the identity of any user was retained. Each time series was normalized by dividing the count for each query in a particular week by the total number of online search queries submitted in that location during the week, resulting in a query fraction (Supplementary Figure 1).
We sought to develop a simple model which estimates the probability that a random physician visit in a particular region is related to an influenza-like illness (ILI); this is equivalent to the percentage of ILI-related physician visits. A single explanatory variable was used: the probability that a random search query submitted from the same region is ILI-related, as determined by an automated method described below. We fit a linear model using the log-odds of an ILI physician visit and the log-odds of an ILI-related search query:

$$\text{logit}(P) = \beta_0 + \beta_1 \times \text{logit}(Q) + \epsilon$$

where $P$ is the percentage of ILI physician visits, $Q$ is the ILI-related query fraction, $\beta_0$ is the intercept, $\text{logit}(P)$ is the natural log of $P/(1-P)$.

Publicly available historical data from the CDC’s U.S. Influenza Sentinel Provider Surveillance Network \cite{12} was used to help build our models. For each of the nine surveillance regions of the United States, CDC reported the average percentage of all outpatient visits to sentinel providers that were ILI-related on a weekly basis. No data were provided for weeks outside of the annual influenza season, and we excluded such dates from model fitting, though our model was used to generate unvalidated ILI estimates for these weeks.

We designed an automated method of selecting ILI-related search queries, requiring no prior knowledge about influenza. We measured how effectively our model would fit the CDC ILI data in each region if we used only a single query as the explanatory variable $Q$. Each of the 50 million candidate queries in our database was separately tested in this manner, to identify the search queries which could most accurately model the CDC ILI visit percentage in each region. Our approach rewarded queries which exhibited regional variations similar to the regional variations in CDC ILI data: the chance that a random search query can fit the ILI percentage in all nine regions is considerably less than the chance that a random search query can fit a single location (Supplementary Figure 2).

The automated query selection process produced a list of the highest scoring search queries, sorted by mean Z-transformed correlation across the nine regions. To decide which queries would be included in the ILI-related query fraction $Q$, we considered different sets of $N$ top scoring queries. We measured the performance of these models based on the sum of the queries in each set, and picked $N$ such that we obtained the best fit against out-of-sample ILI data across the nine regions (Figure 1).

**Figure 1: An evaluation of how many top-scoring queries to include in the ILI-related query fraction**

Maximal performance at estimating out-of-sample points during cross-validation was obtained by summing the top 45 search queries. A steep drop in model performance occurs after adding query 81, which is “oscar nominations”.

\http://semanticommunity.info/Data_Science/Data_Culture_at_the_NIH
Updated: Fri, 18 Sep 2015 23:53:15 GMT
Powered by Qmindtouch™
Combining the N=45 highest-scoring queries was found to obtain the best fit. These 45 search queries, though selected automatically, appeared to be consistently related to influenza-like illnesses. Other search queries in the top 100, not included in our model, included topics like “high school basketball” which tend to coincide with influenza season in the United States (Table 1).

Table 1: Topics found in search queries which were found to be most correlated with CDC ILI data

The top 45 queries were used in our final model; the next 55 queries are presented for comparison purposes. The number of queries in each topic is indicated, as well as query volume-weighted counts, reflecting the relative frequency of queries in each topic.

<table>
<thead>
<tr>
<th>Search Query Topic</th>
<th>Top 45 Queries</th>
<th>Top 45 Queries</th>
<th>Next 55 Queries</th>
<th>Next 55 Queries</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N</td>
<td>Weighted</td>
<td>N</td>
<td>Weighted</td>
</tr>
<tr>
<td>Influenza Complication</td>
<td>11</td>
<td>18.15</td>
<td>5</td>
<td>3.40</td>
</tr>
<tr>
<td>Cold/Flu Remedy</td>
<td>8</td>
<td>5.05</td>
<td>6</td>
<td>5.03</td>
</tr>
<tr>
<td>General Influenza Symptoms</td>
<td>5</td>
<td>2.60</td>
<td>1</td>
<td>0.07</td>
</tr>
<tr>
<td>Term for Influenza</td>
<td>4</td>
<td>3.74</td>
<td>6</td>
<td>0.30</td>
</tr>
<tr>
<td>Specific Influenza Symptom</td>
<td>4</td>
<td>2.54</td>
<td>6</td>
<td>3.74</td>
</tr>
</tbody>
</table>
Using this ILI-related query fraction as the explanatory variable, we fit a final linear model to weekly ILI percentages between 2003 and 2007 for all nine regions together, thus learning a single, region-independent coefficient. The model was able to obtain a good fit with CDC-reported ILI percentages, with a mean correlation of 0.90 (min=0.80, max=0.96, n=9 regions) (Figure 2).

**Figure 2: A comparison of model estimates**

for the Mid-Atlantic Region (black) against CDC-reported ILI percentages (red), including points over which the model was fit and validated. A correlation of 0.85 was obtained over 128 points from this region to which the model was fit, while a correlation of 0.96 was obtained over 42 validation points. 95% prediction intervals are indicated.
The final model was validated on 42 points per region of previously untested data from 2007-2008, which were excluded from all prior steps. Estimates generated for these 42 points obtained a mean correlation of 0.97 (min=0.92, max=0.99, n=9 regions) with the CDC-observed ILI percentages.

Throughout the 2007-2008 influenza season, we used preliminary versions of our model to generate ILI estimates, and shared our results each week with the Epidemiology and Prevention Branch of Influenza Division at CDC to evaluate timeliness and accuracy. Figure 3 illustrates data available at different points throughout the season. Across the nine regions, we were able to consistently estimate the current ILI percentage 1-2 weeks ahead of the publication of reports by the CDC’s U.S. Influenza Sentinel Provider Surveillance Network.

Because localized influenza surveillance is particularly useful for public health planning, we sought to further validate our model against weekly ILI percentages for individual states. CDC does not make state-level data publicly available, but we validated our model against state-reported ILI percentages provided by the state of Utah, and obtained a correlation of 0.90 across 42 validation points (Supplementary Figure 3).

**Figure 3: ILI percentages estimated**

by our model (black) and provided by CDC (red) in the Mid-Atlantic region, showing data available at four points in the 2007-2008 influenza season. During week 5, we detected a sharply increasing ILI percentage in the Mid-Atlantic region; similarly, on March 3, our model indicated that the peak ILI percentage had been reached during week 8, with sharp declines in weeks 9 and 10. Both results were later confirmed by CDC ILI data.
Google web search queries can be used to accurately estimate influenza-like illness percentages in each of the nine public health regions of the United States. Because search queries can be processed quickly, the resulting ILI estimates were consistently 1-2 weeks ahead of CDC ILI surveillance reports. The early detection provided by this approach may become an important line of defense against future influenza epidemics in the United States, and perhaps eventually in international settings.

Up-to-date influenza estimates may enable public health officials and health professionals to better respond to seasonal epidemics. If a region experiences an early, sharp increase in ILI physician visits, it may be possible to focus additional resources on that region to identify the etiology of the outbreak, providing extra vaccine capacity or raising local media awareness as necessary.

This system is not designed to be a replacement for traditional surveillance networks or supplant the need for laboratory-based diagnoses and surveillance. Notable increases in ILI-related search activity may indicate a need for public health inquiry to identify the pathogen or pathogens involved. Demographic data, often provided by traditional surveillance, cannot be obtained using search queries.

In the event that a pandemic-causing strain of influenza emerges, accurate and early detection of ILI percentages may enable public health officials to mount a more effective early response. Though we cannot be certain how search engine users will behave in such a scenario, affected individuals may submit the same ILI-related search queries used in our model. Alternatively, panic and concern among healthy individuals may cause a surge in the ILI-related query fraction and exaggerated estimates of the ongoing ILI percentage.

The search queries in our model are not, of course, exclusively submitted by users who are experiencing influenza-like symptoms, and the correlations we observe are only meaningful across large populations. Despite strong historical correlations, our system remains susceptible to false alerts caused by a sudden increase in ILI-related queries. An unusual event, such as a drug recall for a popular cold or flu remedy, could cause such a false alert.

Harnessing the collective intelligence of millions of users, Google web search logs can provide one of the most timely, broad reaching influenza monitoring systems available today. While traditional systems require 1-2 weeks to gather and process surveillance data, our estimates are current each day. As with other syndromic surveillance systems, the data are most useful as a means to spur further investigation and collection of direct measures of disease activity.

This system will be used to track the spread of influenza-like illness throughout the 2008-2009 influenza season in the United States. Results are freely available online at http://www.google.org/flutrends.

Methods

Privacy

At Google, we recognize that privacy is important. None of the queries in our project’s database can be associated with a particular individual. Our project’s database retains no information about the identity, IP address, or specific physical location of any user. Furthermore, any original web search logs older than 9 months are being anonymized in accordance with Google’s Privacy Policy (http://www.google.com/privacypolicy.html).
Search query database

For the purposes of our database, a search query is a complete, exact sequence of terms issued by a Google search user; we don’t combine linguistic variations, synonyms, cross-language translations, misspellings, or subsequences, though we hope to explore these options in future work. For example, we tallied the search query “indications of flu” separately from the search queries “flu indications” and “indications of the flu”.

Our database of queries contains 50 million of the most common search queries on all possible topics, without pre-filtering. Billions of queries occurred infrequently and were excluded. Using the internet protocol (IP) address associated with each search query, the general physical location from which the query originated can often be identified, including the nearest major city if within the United States.

Model data

In the query selection process, we fit per-query models using all weeks between September 28, 2003 and March 11, 2007 (inclusive) for which CDC reported a non-zero ILI percentage, yielding 128 training points for each region (each week is one data point). 42 additional weeks of data (March 18, 2007 through May 11, 2008) were reserved for final validation. Search query data before 2003 was not available for this project.

Automated query selection process

Using linear regression with 4-fold cross validation, we fit models to four 96-point subsets of the 128 points in each region. Each per-query model was validated by measuring the correlation between the model’s estimates for the 32 held-out points and CDC’s reported regional ILI percentage at those points. Temporal lags were considered, but ultimately not used in our modeling process.

Each candidate search query was evaluated nine times, once per region, using the search data originating from a particular region to explain the ILI percentage in that region. With four cross-validation folds per region, we obtained 36 different correlations between the candidate model’s estimates and the observed ILI percentages. To combine these into a single measure of the candidate query’s performance, we applied the Fisher Z-transformation to each correlation, and took the mean of the 36 Z-transformed correlations.

Computation and pre-filtering

In total, we fit 450 million different models to test each of the candidate queries. We used a distributed computing framework to efficiently divide the work among hundreds of machines. The amount of computation required could have been reduced by making assumptions about which queries might be correlated with ILI. For example, we could have attempted to eliminate non-influenza-related queries before fitting any models. However, we were concerned that aggressive filtering might accidentally eliminate valuable data. Furthermore, if the highest-scoring queries seemed entirely unrelated to influenza, it would provide evidence that our query selection approach was invalid.
Constructing the ILI-related query fraction

We concluded the query selection process by choosing to keep the search queries whose models obtained the highest mean Z-transformed correlations across regions: these queries were deemed to be “ILI-related”.

To combine the selected search queries into a single aggregate variable, we summed the query fractions on a regional basis, yielding our estimate of the ILI-related query fraction $Q$, in each region. Note that the same set of queries was selected for each region.

Fitting and validating a final model

We fit one final univariate model, used for making estimates in any region or state based on the ILI-related query fraction from that region or state. We regressed over 1152 points, combining all 128 training points used in the query selection process from each of the nine regions. We validated the accuracy of this final model by measuring its performance on 42 additional weeks of previously untested data in each region, from the most recently available time period (March 18, 2007 through May 11, 2008). These 42 points represent approximately 25% of the total data available for the project, the first 75% of which was used for query selection and model fitting.

State-level model validation

To evaluate the accuracy of state-level ILI estimates generated using our final model, we compared our estimates against weekly ILI percentages provided by the state of Utah. Because the model was fit using regional data through March 11, 2007, we validated our Utah ILI estimates using 42 weeks of previously untested data, from the most recently available time period (March 18, 2007 through May 11, 2008).

Acknowledgements

We thank Lyn Finelli at the CDC Influenza Division for her ongoing support and comments on this manuscript. We are grateful to Dr. Robert Rolfs and Lisa Wyman at the Utah Department of Health and Monica Patton at the CDC Influenza Division for providing ILI data. We thank Vikram Sahai for his contributions to data collection and processing, and Craig Nevill-Manning, Alex Roetter, and Kataneh Sarvian from Google for their support and comments on this manuscript.

Author contributions

J.G. and M.H.M. conceived, designed, and implemented the system. J.G., M.H.M., and R.S.P. analysed the results and wrote the paper. L.B. (CDC) contributed data. All authors edited and commented on the paper.

Supplementary material

Figures and other supplementary material is available at http://www.nature.com/nature/journal...ture07634.html

References
1

2

3

4

5

6

7

8

9
Quantifying Reproducibility in Computational Biology

The Case of the Tuberculosis Drugome

Source: http://www.plosone.org/article/fetch...esentation=PDF (PDF)

Daniel Garijo1, Sarah Kinnings2, Li Xie3, Lei Xie4, Yinliang Zhang5, Philip E. Bourne3*, Yolanda Gil6*

1 Ontology Engineering Group, Facultad de Informática, Universidad Politécnica de Madrid, Madrid, Spain, 2 Department of Chemistry and Biochemistry, University of California San Diego, La Jolla, California, United States of America, 3 Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California San Diego, La Jolla, California, United States of America, 4 Department of Computer Science, Hunter College, The City University of New York, New York, New York, United States of America, 5 School of Life Sciences, University of Science and Technology of China, Hefei, Anhui, China, 6 Information Sciences Institute and Department of Computer Science, University of

Abstract

How easy is it to reproduce the results found in a typical computational biology paper? Either through experience or intuition the reader will already know that the answer is with difficulty or not at all. In this paper we attempt to quantify this difficulty by reproducing a previously published paper for different classes of users (ranging from users with little
expertise to domain experts) and suggest ways in which the situation might be improved. Quantification is achieved by estimating the time required to reproduce each of the steps in the method described in the original paper and make them part of an explicit workflow that reproduces the original results. Reproducing the method took several months of effort, and required using new versions and new software that posed challenges to reconstructing and validating the results. The quantification leads to “reproducibility maps” that reveal that novice researchers would only be able to reproduce a few of the steps in the method, and that only expert researchers with advance knowledge of the domain would be able to reproduce the method in its entirety. The workflow itself is published as an online resource together with supporting software and data. The paper concludes with a brief discussion of the complexities of requiring reproducibility in terms of cost versus benefit, and a desiderata with our observations and guidelines for improving reproducibility. This has implications not only in reproducing the work of others from published papers, but reproducing work from one’s own laboratory.


Editor: Christos A. Ouzounis, The Centre for Research and Technology, Hellas, Greece

Received September 18, 2012; Accepted October 10, 2013; Published November 27, 2013

Copyright: 2013 Garijo et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: This research is sponsored by Elsevier Labs, the National Science Foundation with award number -0 , the Air Force Office of Scientific Research with award number FA9550-11-1-0104, internal funds from the University of Southern California’s Information Sciences Institute and from the University of California, San Diego, and by a Formaci study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The research presented here has been sponsored partly by Elsevier Labs. This does not alter the authors’ adherence to all the PLOS ONE policies on sharing data and materials.

* E-mail: pbourne@ucsd.edu (PEB); gil@isi.edu (YG)

---

**Introduction**

Computation is now an integral part of the biological sciences either applied as a technique or as a science in its own right - bioinformatics. As a technique, software becomes an instrument to analyze data and uncover new biological insights. By reading the published article describing these insights, another researcher hopes to understand what computations were carried out, replicate the software apparatus originally used and reproduce the experiment. This is rarely the case without significant effort, and sometimes impossible without asking the original authors. In short, reproducibility in computational biology is aspired to, but rarely achieved. This is unfortunate since the quantitative nature of the science makes reproducibility more obtainable than in cases where< experiments are qualitative and hard to describe explicitly.
An intriguing possibility where potential quantification exists is to extend articles through the inclusion of scientific workflows that represent computations carried out to obtain the published results, thereby capturing data analysis methods explicitly \cite{1}. This would make scientific results more reproducible because articles would have not only a textual description of the computational process described in the article but also a workflow that, as a computational artifact, could be analyzed and re-run automatically. Consequently, workflows can make scientists more productive because they capture complex methods in an easy to use accessible manner \cite{2–3}.

The goal of this article is, by applying a workflow to an existing computational analysis \cite{4}, to describe and quantify the effort involved in reproducing the published computational method and to articulate guidelines for authors that would facilitate reproducibility and reuse. Quantification is achieved by assigning a reproducibility score that exposes the cost of omitting important information from the published paper that then caused problems in creating the workflow. Beyond this no case is made for the value of workflows which is well described elsewhere \cite{3}.

Related Work

As stated, scientific articles describe computational methods informally, as the computational aspects of the method may not be the main focus of the article. We acknowledge that in computer science the method may be described formally and any limitations, it could be argued, reside with the editors and reviewers. However, in the domain of computational biology, which is the focus here, we believe methods to be, for the most part, described informally as formalizations are not typically favored by authors or enforced by reviewers.

Computational methods are often complex and hard to explain in textual form with the given space limitations of many articles. As a result, reproducing methods often requires significant effort from others to reproduce and reuse. Studies have shown that reproducibility is not achievable from the article itself, even when datasets are published \cite{5–7}. The reproducibility process can be so costly that it has been referred to as “forensic” research \cite{8}. Lack of reproducibility also affects the review process and as a result retractions of publications occur more often than is desirable \cite{9}. A recent editorial proposed tracking the “retraction index” of scientific journals to indicate the proportion of published articles that are later found problematic \cite{10}. Publishers themselves are asking the community to end “black box” science that cannot be easily reproduced \cite{11}. Pharmaceutical companies report abandoning efforts to reproduce research that seemed initially promising and worth investigating after substantial investments \cite{12}.

Computational reproducibility is a relatively modern concept. The Stanford Exploration Project led by Jon Claerbout published an electronic book containing a dissertation and other articles from their geosciences lab \cite{13}. Papers are accompanied by zipped files with the software that could be used to reproduce the results, and a methodology was developed to create and manage all these objects that continue today with the Madagascar software \cite{14}. Advocates of reproducibility have sprung up over the years in many disciplines, from signal processing \cite{15} to psychology \cite{16}. Organized community efforts include reproducibility tracks at conferences \cite{17–19}, reproducibility editors in journals \cite{20}, and numerous community workshops and forums (e.g., \cite{21,22}). Active research in this area is addressing a range of topics including copyright \cite{23}, privacy \cite{24}, social \cite{25} and validation issues \cite{26}.

Scientific publications could be extended so that they incorporate computational workflows, as many already include data \cite{1}. Without access to the source codes for the papers, reproducibility has been shown elusive \cite{7}. This would make scientific results more easily reproducible because articles would have not just a textual description of the computational process used but also a workflow that, as a computational artifact, could be inspected and automatically re-executed.
Some systems exist that augment publications with scripts or workflows, such as Weaver for Latex [27–28] and GenePattern for MS Word [29]. Many scientific workflow systems now include the ability to publish provenance records [30–31]. The Open Provenance Model was developed by the scientific workflow community and is extensively used for this purpose [32]. Here we make a contribution to the on-going discussion of reproducibility by attempting to quantify what reproducibility implies.

### Methods and Analysis

#### Quantifying Reproducibility

We focus on an article that describes a method that lends itself to workflow representation, since others can, in principle, use the same exact procedures [4]. The article describes a computational pipeline that, as applied, maps all putative FDA and European drugs to possible protein receptors within a given proteome; Mycobacterium tuberculosis (TB) in the paper under study. Mapping is limited to the accessible structural proteome of experimental structures and high quality homology models. Mapping is performed using a binding site comparison algorithm which compares the binding site of the drug bound to a primary protein receptor to potential binding sites found on every available protein in a given proteome. Docking of the drug to the off-target protein is used to further validate the predicted binding. The study uses data from the RCSB Protein Data Bank (PDB [33]) and Modbase [34]. The resultant “drugome” established multiple receptors to which a given drug can bind and multiple drugs that could bind to a given receptor. As such it is a putative map of possible drug repositioning strategies in treating a given condition caused by a pathogen. Although the article focuses on Mycobacterium tuberculosis (TB), according to the article’s abstract:

…”the methodology may be applied to other pathogens of interest with results improving as more of their structural proteomes are determined through the continued efforts of structural biology/genomics.”

That is, the methodology is likely to be repeated for other organisms and/or repeated in the same organism as more drugs become available and/or more of the structural proteome becomes available. The original work did not use a workflow system; instead the computational steps were run separately and manually. The original work was done over a period of two years, with different authors having different degrees of participation in the design and the programming aspects of the study. There is a TB Drugome project site where many details about the work can be found [35].

The original article was used to challenge participants at the first Beyond the PDF workshop [21]. The workshop attracted participants interested in bettering the communication and comprehension of science. The challenge was to apply the tools they had developed to illustrate their value on a given piece of science to which, as far as possible, all lab notes, raw data, software, drafts of the paper etc. where made available. The work described here is one outcome of these efforts and is aimed at addressing the questions: What can we gain from the process of workflow creation and what does it tell us about reproducibility?

The rest of this paper describes our attempt to answer these questions. Many details of the analysis and how progress was made in reproducing the method are available on the project site [36]. Also Supplement S1 includes a more detailed analysis and the thought processes that occurred.
Methodology

The workflow was reproduced as a joint effort between computer scientists and the original authors of the article. Although some of the authors of the paper had moved to other research groups (notably Kinnings, its first author), they were still available to answer questions and provide software scripts and data as needed.

We present a detailed analysis of the issues that came up in reproducing three major parts of the methods section in the original paper. These three parts were originally fully automated. Other steps of the method, notably the initial steps to obtain the data and the final steps for visualization and presentation, were manually done and not considered as part of the workflow presented here.

We describe how each of the three method subsections was implemented as a workflow. Each computational step corresponds to an execution of an existing tool or a script written by the paper authors. We were able to recreate the workflow in the Wings workflow system [37–39] to make sure it was executable and reproduced the original results reported in the paper. Hence, the workflow explicitly represents the method that the authors meant to convey in the original text, that is, the process by which software and data are used to achieve the published result.

Based on this explicit computational workflow, we present an analysis of the reproducibility of each subsection. We considered reproducibility by researchers of four types:

1. **REP-AUTHOR**, is a researcher who did the original work and who may need to reproduce the method to update or extend the results published. It is assumed that the authors have enough backup materials to answer any questions that arise in reconstructing the method. In practice, some authors may be students that move away from the lab and their materials and notes may or may not be available, confounding reproducibility [40].

2. **REP-EXPERT** is a researcher familiar with the research area. These researchers could reproduce the method even if the methods section of the paper is incomplete and ambiguous. They can use their knowledge of the domain, the software tools and the process to make very complex inferences from the text and reconstruct the method. However, there may be some nontrivial inferences that require significant effort.

3. **REP-NOVICE** is a researcher with basic bioinformatics expertise. They may be asked to use the method with new data, but are only able to make limited inferences based on analyzing the text and software tools. For them reproducibility can be very costly since it may involve a lot of trial and error, or perhaps additional research. In some cases reproducibility may become impossible.

4. **REP-MINIMAL** is a researcher with no expertise in bioinformatics. They need some programming skills to assemble the software necessary to run the different steps of the method. They represent researchers from other areas of science with minimal knowledge about biology, students, and even entrepreneurial citizen scientists (e.g., [41]). Unless the steps of the method are explicitly stated, they would not be able to reproduce the results.

In our work, we did not ask experts to reproduce the method, so we only have three categories of researcher rather than four. We used the following approach:

**N REP-MINIMAL** - The computer scientists in the team read the article and formulated the initial workflows. They have minimal background knowledge in biology.
**N REP-NOVICE** - The computer scientists subsequently consulted the documentation on the software tools mentioned in the article to try to infer how the data were being processed by each of the steps of the method. Based on this, they refined their initial workflows.

**N REP-AUTHOR** - Lastly the computer scientists approached the original paper authors to ask specific questions, resolve execution failures and errors and consult concerning the validity of the results for each step. They created the final workflow based on these conversations with the authors.

We analyzed each of the workflow steps in terms of: whether the existence of the step itself was clear to the reproducers, whether the software that was used to run the step was clear to the reproducers, and whether their inputs and outputs were clear. For example, the existence of a step to compare ligand binding sites is mentioned in the text of the original paper, and the fact that it was carried out using the SMAP software [42] is also explicit in the text, so those would be things that the **REP-MINIMAL** reproducers were able to figure out. The use of a p-value as an input was not mentioned in the text and cannot be easily inferred unless the researcher reproducing the method becomes familiar with the software, so REP-NOVICE reproducers were able to figure out this parameter.

For this analysis, we assigned a reproducibility score to each aspect of the workflow for each of these reproducer categories. A score of 1 in a category means that, in our assessment, a prototypical researcher of that category would be able to figure out the item. A score of 0 means that they would not be likely to figure it out without help from experts.

Based on these scores, we designed a reproducibility map, where the reproducibility of each computational step was highlighted to determine how far each category of researcher could go in reproducing a given workflow fragment.

Finally, we report on the effort involved in creating the workflow, measured as the time spent on various aspects of the work involved in reproducing the method described in the original article.

**Conceptual Overview of the Method and Final Workflow**

An interesting result of our initial discussions of the method was a collaborative diagram that indicated each of the steps in the method and how data were generated and used by each step. This diagram, shown in Figure 1, makes the steps of the method more explicit and adds useful information to the text in the methods section. It also shows where the data in the tables of the article fit into the method.

**Figure 1. A high-level dataflow diagram of the TB drugome method.**

doi:10.1371/journal.pone.0080278.g001
In essence, the bulk of the results in the paper are obtained through three major steps:

1. **Comparison of ligand binding sites**, which compares the putative binding sites of solved protein structures and homology models (obtained from queries to the PDB and other sources) against the binding sites from protein structures where approved drugs are bound. This step used the SMAP software [42].

2. **Comparison of protein structures**, optimizing their alignment as well as reporting on the statistical significance of the structural similarity. This step used the FATCAT software [43] and is in essence a filtering step to remove structures which have overall global similarity and hence likely to be in the same protein family, since we are interested in similar binding sites found in otherwise dissimilar proteins.

3. **Molecular docking**, to predict the binding and affinity of the proteins and drug molecules. This step used the eHits software [44].

**Based on our experience, authors should be encouraged to publish such high-level flow diagrams as a normal part of the materials and methods section of a paper.** The diagrams provide a high level overview of the method, highlights major steps, and offer a roadmap for reproducibility.

The final workflow with the four steps that reproduced the method is shown in Figure 2. We highlight the first three major subsections of the method. In order to validate the new results, we used the same inputs (drug binding sites, solved structures, and homology models) as in the original work. However, these inputs point to external data sources (like the PDB) where the data are stored. These third-party data sources had been updated, and therefore the workflow execution produced slightly different results than the results reported in the original article. A detailed comparison of the original results and the results of the new workflow is provided in Supplement S1.

**Figure 2. The reproduced TB Drugome workflow with the different subsections highlighted**

(1) Comparison of ligand binding sites using SMAP; (2) protein structure comparison using FATCAT; (3) docking using Autodock Vina; and (4) graph network creation (visualization). We focus on the reproducibility of sections 1-3 here.

doi:10.1371/journal.pone.0080278.g002
Reproducibility Analysis

We now analyze each of the subsections of the method as described in the original paper, discussing the difficulties encountered in reproducing the method, highlighting recommendations to improve reproducibility, and show reproducibility scores for each step of the final workflow. An extended analysis of each subsection of the method is available in Supplement S1, detailing the evolution of each sub-workflow in order to achieve the final result.

Comparison of ligand binding sites. The initial workflow design used a single step to compare the three items: the binding sites of experimental structures, the binding sites of the homology models, and the binding sites of the proteins to which drugs were bound. Examining the SMAP software and associated scripts revealed that comparison occurred in two steps: one to compare the experimental binding sites with the drug binding sites, and one to compare the homology model binding sites with the drug binding sites.

To clarify how the outputs of both SMAP invocations were combined, the authors provided the script that invoked the SMAP software. This revealed a new step for sorting the results. In addition, there was an additional step where the results below a given p-value were filtered out.

The SMAP software has several configuration parameters. Without the author's configuration files, default values of the parameters were used not knowing if the workflow would produce questionable results. That is, it is not clear whether without the same parameter settings the original method would be reproduced and similar results would be obtained. For these reasons, the original configuration files were obtained from the authors. This suggests that it would be good practice for authors to publish not just a description of the software used and the data used in the original experiment, but also the configuration files used.

It also became clear that the data published as tables in the original article were not the direct input to the SMAP software, and some transformations would be required in order to use these data in the workflow. We recommend that
when data is published in formats that make it more readable, the actual data that is input for software to run also be made available.

Another issue concerned the constant evolution of the software tools that are used for the method steps. In our case, the SMAP software had evolved since the publication of the original paper. As with many software tools used in biology, SMAP is an active research effort and its functionality continues to improve. When the workflow was reproduced there was a new version of SMAP that had the same basic functionality, but produced slightly different results. Under normal research circumstances, it is not critical that the workflow reproduce the exact execution results, but that the conclusions drawn from those results still hold. An interesting result would be if the workflow was run again with a newer more powerful tool and there were additional findings over and above the original publication. The same can be said for new and more comprehensive sources of input data. The possibility of easily re-running and checking the method periodically with new versions of software tools and/or data that might lead to additional findings may entice researchers to keep their methods more readily reproducible.

Global comparison of protein structures. Inspecting the scripts used by the authors revealed two steps for this subsection not mentioned in the original article. The first step generates a list of significant comparisons, which is used in the second step to remove significantly similar pairs of global structures from the FATCAT output. An expert in the domain would infer the need for these steps from the published article – only one structure from a set with similar global structures is needed to reach the appropriate conclusions. The article mentions the use of a threshold of 0.05, but this value did not appear in any parameter file. The FATCAT documentation mentions that 0.05 is a default value used to filter results, so this threshold did not have to be reflected in the workflow since it was fixed by the software – hard for a novice to know. Thus the workflow for this subsection could not be recreated just from the article alone, but required the scripts from the authors. Authors should be encouraged to publish any software and parameter files that were written by them and that became part of the method, because public domain software tools are only part of the software required to reproduce the method.

An important issue regarding reproducibility came up in this subsection of the workflow. Although the method was reproduced with all of the necessary steps, the execution of the FATCAT step failed. The reason for the failure was that some of the PDB (protein) ids used in the input list had been superseded by newer structures in the PDB. Therefore, an additional component was added to check availability and replace any obsolete protein with its superseded version. This issue will not be unusual in reproducibility. Many experiments rely upon third party data sources that change constantly. Consequently, it is to be expected that these sources may not always be available and that the results that they return for the same given query may not always be the same. In our case, the changes in the PDB were addressed by adding a step that updated the older IDs with the new ones. This suggests that some published results that depend on third party data sources may not always be reproducible exactly, so it would be good practice to publish all intermediate data from the experiment so that the method followed can be examined when re-execution is not possible. An alternative is that data archives provide access to their contents for each version.

Docking. The raw interaction network resulting from the first subsection of the method (comparison of ligand binding sites) was assumed to be the input for docking. It turns out that although the input for docking is data produced by SMAP, it is not the raw interaction network that it outputs. Instead, it is data that SMAP places in an “alignment” folder - only expert users would be aware of this.
The original article refers to adding cofactors to relevant proteins prior to docking, which could be interpreted to be a step prior to docking. As it turns out, there is no explicit step for handling the cofactors since this is handled by manually editing the appropriate PDB file. Again, only expert users would be aware of this.

Examination of the author’s scripts revealed some additional steps: calculating the clip files, which are used for obtaining the ideal ligands before docking. Clip files are mentioned in the article as containing the aligned drug molecules, so it would seem to a non-expert that the aligned molecules would be the output of the initial alignment steps of the overall method.

A major issue with this portion of the workflow is that the docking software used for the original article was no longer used in the laboratory. It is proprietary software, and its license had expired, so alternative software (AutodockVina) with similar functionality has been adopted since the original article was published. Some of the ligands were not recognized by this software, so a transformation step had to be added to the workflow to make Autodock Vina work correctly.

There are reasons why authors use proprietary software, for example, ease of use, support, robustness, visualization and data types supported. However, the authors could replicate the method before publication using open source tools, which would facilitate reproducibility by others. The use of open source software instead of proprietary software facilitates the reproduction of the software steps originally used by the authors, and should be the preferred mode of publication of methods and workflows.

Reproducibility Maps

We present reproducibility maps created as a summary of the reproducibility scores for all the major steps in the workflow. Figure 3 shows the reproducibility maps for each of the subsections, summarizing the reproducibility scores assigned to each step. For each section of the method, we show a progression of steps from left to right, noting on the right hand side the category of reproducer represented (MINIMAL, NOVICE, and AUTHOR). A step is shown in red if it was not reproducible by that category of user, and green if it were.

Figure 3. Reproducibility maps of the three major subsections of the workflow

A step is shown in red if it was not reproducible by that category of user, and green if it were.
doi:10.1371/journal.pone.0080278.g003
Our observation was that a researcher with minimal knowledge of the domain would only be able to reproduce one of the fourteen steps in the workflow. A novice researcher would be able to reproduce seven of the fourteen steps: the six steps to compare ligand binding sites, only one of the four steps to compare the protein structures, and none of the steps for docking. For docking, our conclusion was that only expert researchers with advanced knowledge of the domain would be able to reproduce the steps. The original software was no longer available, and advanced expertise was required to identify equivalent software to replace it, and to write the software necessary to make it work as needed. Expert researchers would be able to reproduce the method, as the original article combined with the data and software published in the site would be sufficient to infer any missing information. A detailed rationale for the scores can be found in the reproducibility scores subsection of Supplement S1.

Regarding the results, we checked that the output of the workflow included all the drugs exposed in the original work (plus new findings). The ranking of drugs in the results of the workflow is almost the same as the original, although the number of connections found for each drug is significantly higher in the results of the workflow. A possible reason is changes in the version of the software tools and updates to the external databases where the structures are stored. A detailed comparison can be seen in the original results versus results from the workflow subsection of Supplement S1.

**Productivity and Effort**

We kept detailed records in a wiki of the effort involved in reproducing the method throughout the project. These records are publicly available from [36].

We estimated the overall time to reproduce the method as 280 hours for a novice with minimal expertise in bioinformatics. The effort included analyzing the paper and the original author’s web site and additional materials (data, scripts, configuration files) to understand the details of the method, locating and preparing the codes, finding appropriate parameter settings, implementing the workflows, asking questions to the authors when necessary, and validating the workflows. It should be noted that the authors of the original experiment were available to answer questions (notably Kinnings, the first author). These questions were related to missing configuration parameters, documentation for the
proper invocation of the tools, and validation of the outcome of the intermediate steps. Table 1 estimates the time
required to reproduce the method and is broken down by major tasks according to our records.

Table 1. Time to reproduce the method

<table>
<thead>
<tr>
<th>Tasks</th>
<th>Time (hours)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Familiarization with workflow and running software</td>
<td>160</td>
</tr>
<tr>
<td>SMAP steps</td>
<td>32</td>
</tr>
<tr>
<td>SMAP result sorter steps</td>
<td>8</td>
</tr>
<tr>
<td>Merger steps</td>
<td>4</td>
</tr>
<tr>
<td>Get significant results</td>
<td>4</td>
</tr>
<tr>
<td>FATCAT URL checker</td>
<td>8</td>
</tr>
<tr>
<td>FATCAT step</td>
<td>4</td>
</tr>
<tr>
<td>Remove significant pairs</td>
<td>4</td>
</tr>
<tr>
<td>Create clip files</td>
<td>8</td>
</tr>
<tr>
<td>Create ideal ligands</td>
<td>8</td>
</tr>
<tr>
<td>Ideal ligand checker</td>
<td>8</td>
</tr>
<tr>
<td>Autodock Vina</td>
<td>16</td>
</tr>
<tr>
<td>Data visualization steps</td>
<td>16</td>
</tr>
<tr>
<td>TOTAL</td>
<td>280 hours</td>
</tr>
</tbody>
</table>

Table 2. Observations and desiderata for reproducibility

<table>
<thead>
<tr>
<th>Observation</th>
</tr>
</thead>
<tbody>
<tr>
<td>• We found that important computational steps were either missing or ambiguous. The paper should make clear all computational steps needed by a novice user.</td>
</tr>
<tr>
<td>• Software is often used with carefully selected parameter settings and configurations. It would be good practice for authors to publish not just a description of the software and data used, but also to publish any parameter settings and configuration files used.</td>
</tr>
</tbody>
</table>
• The possibility of re-running the method periodically with new versions of software tools leading to new findings might help entice researchers to keep their methods readily reproducible.

• Published results that depend on third party data sources may not always be accessible and may make the experiments run by the original authors irreproducible

Where practical, authors should publish all intermediate data from the experiment so that the method they followed can be examined when direct re-execution is not possible.

• To implement some steps of their methods, authors often use proprietary software or software that is not widely available. The use of open source software facilitates the reproduction of the software steps originally used by the authors, and should be the preferred mode of publication for authors of methods and workflows.

• Although many methods are implemented by using public domain software tools, they often contain additional steps that were implemented by the authors. To facilitate reproducibility, authors should publish any software written by them and that became part of the method.

doi:10.1371/journal.pone.0080278.t002

Table 3. Reproducibility Guidelines for Authors

<table>
<thead>
<tr>
<th>Guideline</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Input data: Provide the original datasets used in the experiment reported in the paper</td>
</tr>
<tr>
<td>2. Dataflow Diagram: Provide a diagram that represents a data flow of the computational steps. The nodes in the graph should be computational steps, which include invocations of software tools, scripts and other software that were written, and any additional data manipulations that were carried out manually. The links in the graph specify the data flow, which indicates what the input data for each step are and links to other steps that may have generated the data.</td>
</tr>
<tr>
<td>3. Software: Prefer open software tools that are appropriately documented. Specify the software tools used mentioning versions and download dates. For any scripts or other software that were written, provide the code itself or at least “pseudo-code” (i.e., an informal version of the code that is language-independent)</td>
</tr>
<tr>
<td>4. Configurations: Provide the values of any parameters and configuration files used</td>
</tr>
<tr>
<td>5. Intermediate data: Provide key intermediate data that resulted from important steps and that would help others determine whether they reproduced the method correctly</td>
</tr>
</tbody>
</table>

doi:10.1371/journal.pone.0080278.t003

Publishing the Reproduced Workflow

Now that we had invested significant effort in reproducing the workflow, our goal was to maximize its reusability.
First, the executed workflow was published using the Open Provenance Model [32]. This model is used by many workflow systems, so it increases the workflow reusability because it can be imported into other systems depending on the preference of the particular research group. We also publish the workflow provenance using the PROV ontology [45], a recent standard for provenance from the W3C [46]. This makes the published workflow independent of the workflow system used to create it.

Second, we published an abstract workflow that complements it. The abstract workflow describes the steps in a manner that is independent of the software used to implement them. For this we used an extension of the Open Provenance Model called OPMW [47] that includes new terms to describe abstract steps.

Third, we published the workflow and all of its constituents (including input and output data, software and scripts for the steps) as Linked Data [48], which means that each constituent of the workflow can be accessed by its URI through HTTP, and its properties are described using W3C RDF standards [49]. This means that the published workflow is accessible over the Web, in a way that does not require figuring out how to access institutional catalogs or file systems.

With this maximally open form of publication of the workflow, the effort that we invested in reproducing the workflow does not have to be incurred by others. Each step and its inputs and outputs are explicitly and separately represented as well as linked to the workflow. The software for each step is available as well, as are the intermediate and final results.

The effort involved in creating a workflow is negligible compared with the time to implement the computational method. Implementing the computational method typically takes months, and involves activities such as finding software packages that implement some of the steps, figuring out how to set up the software (e.g., setting up parameters) to suit the data, and writing new code to reformat the data to fit those packages. Once this is all done, creating the workflow can be done in a few hours, and can be as simple as wrapping each step so it can be invoked as a software component and expressing the dataflow among the components. Learning to create simple workflows requires only a few hours, more advanced capabilities clearly require additional time investment (e.g., running workflows in a cluster, depositing results in a catalog, or expressing a complex control flow). Similarly, publishing workflows takes no effort at all since the workflow system takes care of the publication.

Technical details on how the workflow is published can be found in [50]. The OWL ontologies for OPM and PROV that express all the underlying RDF properties can be browsed from [51]. All the materials related to the workflow and its execution results have been published online [36]. Additionally, input and output datasets have been associated to DOIs and uploaded to a persistent data sharing repository [52].

Discussion

Reproducibility is considered a cornerstone of the scientific method and yet rarely is scientific research reproducible without significant effort, if at all [5-7]. Authors submitting papers know this; as do those reading the papers and trying to reproduce the experiment. For computational work like that described here, where data, methods, and control parameters are all explicitly defined there is less of an excuse for not making the work reproducible. Note that making the software available or accessible through a webserver, while commendable, is not the same as making the work reproducible. Workflows, which define the scientific process as well as all the components, provide the tools for improved reproducibility. While workflows are commonly used for highly repetitive tasks, they are less used for earlier
stage research. Whether this is a result of shortcomings in the tools or insufficient emphasis on the need to make work reproducible requires further consideration. This then raises the further issue of whether the emphasis itself is justified. Do we really care if work is exactly reproducible? This generally only becomes important if some variation of the original work cannot be reproduced at all, then the original work is fully scrutinized. This speaks to a need for better quantification of what is really needed to improve productivity in science. When, as is the case here, the experiment is conducted completely in silico, the opportunity to accurately capture what has transpired becomes a relatively straightforward task (i.e., there is a relatively favorable cost:benefit ratio) and raises the question as to whether the community of computational biologists should do better. What does doing better imply?

We believe it is rare that work is purposely made irreproducible; rather the system of peer review speaks to reproducibility but is cursory in demanding it. The scientific reward is in publishing another paper, not making your current paper more reproducible. Tools help, but changes in policy are also needed. It will be a brave publisher indeed that demands that workflows be deposited with the paper. Publishing after all is a business and if one publisher demands workflows, authors are more likely to publish elsewhere than go to the trouble. Journals are beginning to provide guidelines for reproducibility and minimum requirements for method descriptions [53–54]. There is already a concept of “data publication,” where datasets are described and receive a unique identifier and a publication. Similarly, there should be a concept of “workflow publication.” There is no explicit credit for publishing software packages, and many people do it. The credit comes indirectly from acknowledgement by the community that the software is useful. Perhaps publishing end-to-end methods as workflows would bring similar reputation. For this to work, authors must be recognized and credited by other researchers reusing their workflow. We posit that the authors of the original method need not be the ones publishing the workflow.

Third parties interested in reproducing the method could publish the workflow once reproduced, and get credit not for the method but for the workflow as a reusable software instrument. In one sense this is no different than taking other scientists data and developing a database that extends the use of these data to a wider community. It is a value-added service worthy of attention through publication.

Federal mandates similar to those emerging around shared data could also be put in place for reproducibility too. In the end, funding for science ultimately comes from taxes from the public, and we need to be responsible in making science as efficient and productive as possible. Many government agencies already require data to be published and shared with other researchers. Workflows should follow the same path. The recent emphasis on open availability of research products resulting from public funds [55–56] will eventually include the publication of software and the methods (workflows). This will likely be sometime coming as the easier issue of meaningful data provision is not fully understood and solved yet. Notwithstanding, if this remains a difficult issue on a global scale we can make progress in our own laboratories.

A new researcher coming to almost any laboratory and picking up tools used by previous laboratory members can likely testify to what is described in this paper. If we are to accelerate scientific discovery we must surely do better both within a laboratory and beyond. This is particularly important in an era of interdisciplinary science where we often wish to apply methods that we are not experts in. Some would argue that irreproducibility in the laboratory is part of the learning process; we would argue yes, but with so much to learn that is more relevant to discovery we should do better now that we have tools to assist us.
Or should we? Reproducibility aside, is there indeed a favorable cost:benefit ratio in using workflows with respect to productivity? There is a dearth of literature that addresses this question. Rather the value of the workflow is assumed and different workflow systems on different computer architectures are analyzed for their relative performance. At best the question can be addressed by work habits. We must be careful as such work habits could be mandated, in a large company say, rather than by choice, which would be the case in an independent research laboratory. Creating workflows results in overhead for exploratory research, where many paths are discarded. However, once created a workflow can be reused many times. This makes them ideal for repetitive procedures such as might be found in aspects of the pharmaceutical industry. Pharmaceutical companies use workflows for computational experiments [57]. This means there must be a business case for workflows in terms of saving time and effort and/or facilitating quality control. Taking an independent computational biology laboratory, as is the case for this study, it is fair to say that workflows are making inroads into daily work habits. These inroads are still localized to specific subareas of study – Galaxy [58] for high-throughput genomic sequence analysis; KNIME [59] for high-throughput drug screening, and so on, but with that nucleation and with new applications being added by an open source-minded community, adoption is increasing. Adoption would assume a favorable cost:benefit ratio in that use of a workflow system provides increased productivity over not using such a system. This is a cost measured in time rather than money since most academic laboratories in computational biology would use free open source workflow systems. Finally, when articles cannot be easily reproduced the authors are often contacted to clarify or describe additional details. This requires effort that might as well have been invested in writing the article more precisely in the first place.

Workflows can also be seen as an important tool to make the research in a lab more rigorous. Analyses must be captured so they can be inspected by others and errors detected as easily as possible. For example, writing code to transform data makes the transformation inspectable, while using a spreadsheet to do the task makes it much harder to verify that it was done correctly. Ensuring consistency and reproducibility requires more effort without workflows. In our own laboratory we find that the workflow can act as a reference such that new users can more quickly familiarize themselves with the various applications than would be the case without the benefit of the workflow organization, but then choose to go on and run applications outside of the workflow system. As the workflow systems themselves continue to be easier to use and more intuitive we anticipate that more work will be done within the workflow system itself, presumably improving productivity.

For the practitioner, what are the pluses and minuses of workflow use today? An obvious minus is the time required to establish the workflow itself. In some sense this is analogous to documenting a procedure to run a set of software programs. But in most cases once codes are prepared for publication little additional effort is required to include them in a workflow. The advantage of a workflow is that capturing the steps themselves defines the procedure and it can be re-run, in principle, without any further effort. We say “in principle” since as this work has shown workflows decay – the tools available change, the licenses to those tools change, remote data accessibility changes etc. Virtual machines offer the promise of capturing the complete executable environment for future use, however they introduce other issues [26]. For example, virtual machines often act as black boxes that allow repeating the experiment verbatim, but do not allow for any changes to the computational execution pipeline, limiting its reproducibility. Furthermore, virtual machines cannot store external dynamic databases accessed at runtime (like the PDB in our work) due to their size. These databases are commonly used for experiments in computational biology.

All taken together, it may be that we are at this tipping point of broad workflow adoption and it will be interesting to review workflow use by the computational biology community two or more years from now.
Conclusions

We conclude by summarizing the main observations resulting from our work, leading to desiderata for reproducibility shown in Table 2, and a set of guidelines for authors shown in Table 3. We have restrained from making too many absolute conclusions from a single instance of applying a workflow to a scientific method. It would be interesting to carry out similar studies in other domains and compare findings.

Supporting Information

**Supplement S1 A detailed account of the Reproducibility of the TB Drugome Method.**

My Note: See below

Author Contributions

Performed the experiments: DG. Analyzed the data: DG YG. Wrote the paper: DG YG PB. Designed and performed the TB-Drugome experiment: SK Li Xie Lei Xie PB. Tested the automated workflow and reported feedback: YZ. Contributed to the manuscript writing with feedback: SK Li Xie Lei Xie.

References


5

6

7

8

9

10

11

12


http://semanticommunity.info/Data_Science/Data_Culture_at_the_NIH
Updated: Fri, 18 Sep 2015 23:53:15 GMT
Powered by mindtouch
23

24

25

26

27

28

29

30

31
32


33


34


35


36


37


38


39


40

41

42

43

44

45

46

47

48

49


A Detailed Account of the Reproducibility of the TB Drugome Method

We describe the evolution of the reproduction of the method as a series of conceptual diagrams resulting from the interaction between the computer scientist and the authors while gathering more information. Finally, we show the method represented as a workflow of computational steps. We also show reproducibility scores for each of the steps of the resulting workflow.

Comparison of Ligand Binding sites using SMAP

Relevant paragraph of methods section from [Kinnings et al. 2010]:

“Xie et al. recently developed the ligand binding site comparison software SMAP [22], which is based on a sequence order independent profile-profile alignment (SOIPPA) algorithm [24]. Firstly, the protein structure is characterized by a geometric potential; a shape descriptor that is analogous to surface electrostatic potential, but which uses a reduced C-alpha only structural representation of the protein. It has been shown that both the location and the boundary of the ligand binding site can be accurately predicted using the geometric potential [23]. The reduced representation of the protein structure makes the algorithm tolerant to protein flexibility and experimental uncertainty; thus SMAP can be applied to low-resolution structures and homology models. Secondly, two protein structures are aligned, independent of sequence order, using a fast, maximum weighted sub-graph (MWSG) algorithm [80], [81]. The MWSG finds the most similar local structures in the spirit of local sequence alignment. Finally, the aligned surface patches are ranked by a scoring function that combines evolutionary, geometric and physical information. The statistical significance of the binding site similarity is then rapidly computed using a unified statistical model derived from an extreme value distribution [22].

The SMAP software was used to compare the binding sites of the 749 M.tuberculosis protein structures plus 1,446 homology models (a total of 2,195 protein structures) with the 962 binding sites of 274 approved drugs, in an all-against-all manner. While the binding sites of the approved drugs were already defined by the bound ligand, the entire protein surface of each of the 2,195 M.tuberculosis protein structures was scanned in order to identify alternative binding sites. For each pairwise comparison, a P-value representing the significance of the binding site similarity was calculated.”

Uncovering the Workflow

First attempt: The first sentence of the second paragraph was interpreted as using the SMAP software as a single step to compare the three items: the binding sites, the homology models, and the drugs. Therefore the first diagram of this workflow fragment can be seen in Figure S1:
**Figure S1: First attempt to reproduce the SMAP analysis**

**Second attempt:** The SMAP software takes only two source arguments, so it became clear that there were two separate executions of the SMAP software: one to compare the binding sites of experimental structures with the drug-receptor binding sites, and one to compare the homology model binding sites with the drug-receptor binding sites. Note that the two SMAP steps can be run independently, but both invoke the same SMAP software. Another important clarification was that the tables were not the direct input to the SMAP software. Therefore, it was assumed that some processing of the tables occurred to prepare the data for SMAP. A second diagram (shown in Figure S2) was created:

**Figure S2: Second attempt to reproduce the SMAP analysis**

**Third attempt:** Examining the SMAP software, it became clear that the input to the SMAP components were not the tables in the paper but lists of solved structures and homology models obtained from queries to the PDB and other biomedical websites. As it often happens, the tables were derived from the actual data that were analyzed and were not the direct input to the workflow. Figure S3 summarizes the third attempt to reproduce the workflow fragment.

**Figure S3: Third attempt to reproduce the SMAP analysis**

**Fourth attempt:** After becoming familiar with SMAP, the output of the tool became clear. By inspecting the script that invoked the SMAP tool, it became apparent that there was a sorting step. An examination of the code revealed that the results of SMAP were not just sorted, but that all the results below a given p-value were filtered out. The following diagram (shown in Figure S4) was created:

**Figure S4: Fourth attempt to reproduce the SMAP analysis**

**Computational Workflow**

The final workflow that represents all the computational steps has been depicted in Figure S5:

**Figure S5: SMAP workflow**

The left strand of the workflow processes and filters the solved structure comparisons against the drug binding sites, while the right strand of the workflow computes and filters the homology models. A final step, Merger, is used to merge both outputs as a whole new dataset.

There are several things to note in the workflow. First, the p-value parameter is not needed for the SMAP step, only for the sorting step, but it was preserved as input to the SMAP step because both the SMAP and sorting steps were in the same original step, and it was decided to preserve its structure (the p-value was an input to that step). Second, the drug binding sites are input to the sorting steps. This was exposed when the software were examined. The drug binding sites are only used to extract the names of the drugs. Third, the solved structures and homology models are input to the respective sorting steps, being used for extracting the names of the proteins and homology models.

There are also other issues that arise that will be of no surprise to those who have previously attempted to reproduce the work of others based on a research article alone. One issue had to do with the publication of configuration
parameters for tools used in a method. The SMAP software has a configuration file as input. This includes the p-value used, (that could be different than the one used for the threshold of the sorting step), MATCH_SECONDARY_STRUCTURE, LIGAND_CONTACT_DISTANCE_CUTOFF, TEMPLATE_LIGAND_SITE_ONLY, TEMPLATE_LIGAND_ID, ASSOCIATE_GRAPH_NODE_FILTER, TIMES_RANDOM_SHUFFLE, and so on. Without the authors configuration files used, default values of the parameters were used not knowing if the workflow would produce questionable results. That is, it is not clear whether without the same parameter settings the original method would be reproduced and similar results would be obtained. For these reasons, the original configuration files were obtained from the authors. This suggests that it would be good practice for authors to publish not just a description of the software used and the data used in the original experiment, but also the configuration files used.

Another issue concerned the constant evolution of the software tools that are used for the method steps. In our case, the SMAP software had evolved since the publication of the original paper. As with many software tools used in biology, SMAP is an active research effort and its functionality continues to improve. When the workflow was reproduced there was a new version of SMAP that had the same basic functionality but different input requirements. As a result, the intermediate and final data generated by the workflow were not exactly the same as the original results. This is not an unusual situation, as software tools are continuously improved over time. Under normal research circumstances, it is not critical that the workflow reproduce the experiment exactly. What is important is that the original results stand the test of time, and any significant findings are still significant no matter what tools are used. An interesting result would be if the workflow was run again with a newer more powerful tool and there were additional findings over and above the original publication. The same can be said for new and more comprehensive sources of input data. The possibility of easily re-running the method periodically with new versions of software tools and/or data that might lead to additional findings may entice researchers to keep their methods more readily reproducible.

**Reproducibility Scores**

Table S1 summarizes our reproducibility scores of this workflow fragment. The last column of the table is a brief justification of the scores. Some of the justifications simply show a quote from the original article where the step is mentioned, highlighting text that refers to inputs or outputs of the step.

We start with the SMAP1 step. First, we assign a score based on whether the existence of that step can be determined. In this case, it is mentioned in the original article, so the score is a “1” in the category MINIMAL for the existence of the step. Next, we consider whether the software for that step (which we call a software component) can be identified. The article mentions the SMAP software, so the score is a “1”. The inputs are mentioned in the article, so the scores are “1” for MINIMAL. The configuration parameters are not specified in the article, so the score is “0” for MINIMAL. One has to investigate the SMAP software, look at its default configuration, and test whether that works, which requires basic knowledge of the domain. So we assigned a score of “1” for NOVICE. The outputs of the step are mentioned in the article, so the score is “1” for basic.

The SMAP2 step, which deals with the homology models, is not mentioned as a separate step in the original article. So it has a score of “0” for MINIMAL and a score of “1” for NOVICE, since looking at the SMAP software reveals how to use it with different inputs. The rest of the reproducibility scores are analogous to the SMAP1 step.
The SMAPResultSorter1 step is not explicitly mentioned in the article, so it receives a score of "0" for MINIMAL. Anyone with basic knowledge of the domain would infer that sorting was necessary, so it has a score of "1" for NOVICE.

The Mergers step are not mentioned in the article, to the score assigned is "0" for MINIMAL. The need to merge results is obvious once the two SMAP steps are done, so the score is "1" for NOVICE.

**Table S1: Reproducibility scores for the SMAP workflow**

### Comparison of dissimilar protein structures using FATCAT

Relevant paragraph of methods section from [Kinnings et al. 2010]:

> “FATCAT (Flexible structure AlignmenT by Chaining Aligned fragment pairs allowing Twists) [82] is a program for the flexible comparison of protein structures. It optimizes the alignment between two structures, whilst minimizing the number of rigid body movements (twists) around pivot points introduced in the reference structure. In addition to the optimal structural alignment, FATCAT reports the statistical significance of the structural similarity, measured as a P-value. In order to identify pairs of similar binding sites that were from proteins with dissimilar global structures (i.e., cross-fold connections), the first chain of each PDB file was aligned using FATCAT, and those pairs with a significant P-value of less than 0.05 were discarded.”

**Uncovering the Workflow**

**First attempt:** The first interpretation of the paragraph clearly led to two different components: the FATCAT tool plus a filtering step for removing the p-values of less than 0.05. What were not that clear were the inputs to each component: it was interpreted as if FATCAT produced an output which was used together with the SMAP output for the filtering step. Figure S6 shows the first attempt to reproduce FATCAT workflow fragment.

**Figure S6: First attempt to reproduce the FATCAT analysis**

**Second attempt:** After having access to the scripts used by the authors, it became clear that FATCAT uses as input a list of pairs of proteins regardless of whether they were experimentally determined or homology models. So the first part of the workflow (first two ovals and first box in the previous figure), would be replaced by Figure S7:

**Figure S7: Second attempt to reproduce the FATCAT analysis**

**Third attempt:** The list used as input to FATCAT is computed through an additional step, not mentioned in the paper. After having access to the scripts of the authors it was clear that it produces another additional list, called "significant results". None of these inputs and processing steps are mentioned in the paper.

**Computational Workflow**

The WINGS workflow represents the steps as the workflow was finally defined (Figure S8).
Figure S8: FATCAT workflow

Note that checkedList is an input to FATCAT. However, the article refers to “the first chain of each PDB file was aligned using FATCAT”. The article does not state explicitly that the checkedList is in fact from the PDB.

The steps RemoveSigPairsFATCAT and GetSignificantResults are not described in the article. An expert in the domain would infer the need for these steps from the published article, looking at the passage: “identify pairs of similar binding sites.”

In addition, the article mentions a threshold of 0.05 that does not appear anywhere in the workflow: “and those pairs with a significant P-value of less than 0.05 were discarded.” This ended up being a default value in FATCAT, so it does not need to be explicitly set up. An expert would understand that this is the default value by reading the FATCAT documentation.

An important issue for reproducibility came up in this portion of the workflow. Although the method was reproduced with all of the necessary steps, the execution of the FATCAT step failed. The reason for the failure was that some of the protein ids used in the input list had been superseded by other structures in the PDB. Therefore, an additional component was added to check availability and replace any possible protein with its superseded version. This component was named FATCATURLChecker.

This issue will not be unusual in reproducibility. Many experiments rely upon third party data sources that change constantly. Consequently, it is to be expected that these sources may not always be available and that the results that they return for the same given query may not always be the same. In our case, the changes in the PDB were addressed by adding a step that updated the older IDs into the new ones. This suggests that some published results that depend on third party data sources may not always be reproducible exactly as the original authors ran the experiment, so it would be good practice to publish all intermediate data from the experiment so that the method followed can be examined when re-execution is not possible.

**Reproducibility Scores**

Table S2 shows the reproducibility scores for this workflow fragment. The first two steps were not mentioned in the original article. Novices would not be able to determine the need for these steps. Only experts would, so the score is “1” for AUTHOR.

The FATCAT step is mentioned in the original article. However, its inputs are mentioned in an ambiguous manner in the text so some degree of expertise is required. For this reason, we assigned a score of “0” for MINIMAL and “1” for NOVICE.

The RemoveSigPairsFATCAT step is not mentioned in the original article. It requires some basic expertise to determine how to combine the results of the two previous comparisons from SMAP and FATCAT. Therefore we assigned a score of “0” for MINIMAL and “1” for NOVICE.
Table S2: Reproducibility scores for the FATCAT workflow

An important issue that came up with this portion of the workflow is that there were two steps (GetSignificantResults and GetSignificantPairsFATCAT) that were implemented by the authors as scripts that were not initially available. The workflow would have been harder to reproduce without those scripts. Once we had obtained the scripts from the authors, the workflow was easily completed. Note that these scripts are not available in the paper or in the accompanying website. **Authors should be encouraged to publish any software that were written by them and that became part of the method, because public domain software tools do not provide all of the software required to reproduce the method.**

**Docking using eHits/AutodockVina**

Relevant paragraph of methods section from [Kinnings et al. 2010]:

“For those pairs of interest, molecular docking was used to predict the binding pose and affinity of the drug molecule to the M.tb protein. eHiTS Lightning [84] was selected due to its fast speed, relatively high accuracy and ease of automation for large-scale docking studies. Since SMAP had aligned the drug binding site with the M.tb protein binding site, the aligned coordinates of the drug molecule were used to define the search space for docking that drug into the M.tb protein. The aligned drug molecule was used as the clip file with a default search space of 10Å3. As recommended by the manual, the eHiTS accuracy level was set to 6 (default = 3), in order to increase the accuracy of the predicted binding poses. Following all docking, the binding pose with the lowest estimated binding affinity was selected for further investigation. For those proteins with cofactors (e.g., InhA has an NAD cofactor), the cofactor was added as the last residue in the protein structure prior to docking.”

**Uncovering the Workflow**

**First attempt:** The SMAP Output (i.e., the rawInteractionNetwork) was assumed to be the input for this step. However, the inputs used are data produced by SMAP, which places the data in an “alignment” folder. In addition, it seemed that a cofactor step was a component of the workflow: “For those proteins with cofactors, the cofactor was added as the last residue in the protein structure prior to docking.” Figure S9 summarizes the first attempt.

**Figure S9: First attempt for reproducing the docking analysis**

**Second attempt:** As it turns out, there is no explicit step for handling the cofactors. In addition, after email discussions with the authors and having access to the scripts, the workflow turned out to have some additional intermediate steps: calculating the clip files, which were used for obtaining the ideal ligands. Clip files are mentioned in the article (“The aligned drug molecule was used as the clip file with a default search space”…), but they seemed to be part of the SMAP output too.

**Third attempt:** A major issue with this portion of the workflow is that the eHits tool was no longer used in the laboratory. eHits is proprietary, and its license had expired. The new tool being used was AutodockVina, which has similar functionality. The IdealLigandCheckerImpl was added to obtain the docking results. Some of the ligands created by the CreateIdealLigandsImpl were not recognized in the AutodockVina step, so a conversion step had to be created for some
of them. It could be considered a formatting step. If this step had not been included, this portion of the method would not have been reproduced. Figure S10 shows the final workflow developed in this step.

**Computational Workflow**

**Figure S10: Docking workflow**

There are other possible alternatives for this workflow. ClipFiles could be input to AutodockVina, so that the create and check ligand steps would not be needed. Another change is that GetSMAPAlignmentFolder could be eliminated if the data is already extracted from the folders where SMAP outputs the data.

**Reproducibility Scores**

The reproducibility of this subsection of the method was greatly affected by the change to a new software (Autodock Vina instead of eHits). Therefore, the reproducibility scores shown in Table S3 reflect that more advanced expertise would be required to adapt the method to use the new software. Most steps have a score of “0” for MINIMAL and most only reach “1” for the AUTHOR level. Although some steps are mentioned in the article (e.g., docking), advanced expertise is required to implement the software needed to adapt the method to use the software, so for the software component the score is “0” for MINIMAL and NOVICE and “1” for AUTHOR. For example, the docking step is mentioned in the article, and the software as well (eHits), but if the original software is not available then advanced expertise would be required to identify equivalent software (e.g., to know that AutodockVina is an appropriate substitute) and to use it correctly.

An additional step that the new software required was IdealLigandCheckerImpl. This is because Autodock Vina did not recognize some of the ligand identifiers. Recognizing the need for this step, and having the ability to create the configuration scripts requires advanced expertise.

**Table S3: Reproducibility scores for the docking workflow**

The reproducers were not able to reproduce this sub-workflow just from the paper, since it required figuring out how to implement it using the Autodock Vina software. Additionally, there were scripts developed by the authors that eventually were available to the reproducers. By looking at the paper and the documentation of the Autodock Vina tool, the reproducers were not able to figure out any of the steps in this workflow strand. Once the reproducers had access to the scripts originally used by the authors, they were able to set up most of the workflow. The reproducers did need further assistance from the authors in building the checker.

An important issue concerning reproducibility was exposed with this sub-workflow. The software tools used in the original experiments by the authors may not be available to the reproducers. One possible reason, as it was in our case, is that the authors used proprietary software to implement some of the method steps. There are reasons why authors use proprietary software, for example, ease of use, support, robustness, visualization and data types supported. However, the authors could replicate the method before publication using open source tools, which would facilitate reproducibility by others. **The use of open source software facilitates the reproduction of the software steps originally used by the authors, and should be the preferred mode of publication of methods and workflows.**
Original results versus results from the workflow

The summary of the results obtained in the original work can be seen in Table S4, while the results from the workflow are shown in Table S5. Both tables show a ranking of the highly connected drugs (ordered by their number of connections) and the M.tb proteins with solved structures found within those connections. As we can see in the tables, all the drugs belonging to the ranking in the original results have been found in the results of the workflow. The only difference between them is the number of connections (higher in the results of the workflow), and the fact that new highly connected drugs appear in the ranking. After discussing the results from the workflow with the authors of the original article, we can claim that we have successfully reproduced the method. The difference between the results is reasonable, as a different version of the software was used for some of the steps (SMAP) and the external databases (such as the PDB, where the structure of the proteins is stored) are dynamic. Figure S11 shows the drug-protein network obtained as a result from the workflow. The size of each drug node in the figure is proportional to its number of connections.

Table S4: Highly connected drugs, original results

Table S5: Highly connected drugs, results of the workflow.

The entries found in the original results have been highlighted showing the commonalities between both

Figure S11: Visualization of the drug-protein network obtained as a result of the workflow

FIGURES AND TABLES

Figure S1: First attempt to reproduce the SMAP analysis
Figure S2: Second attempt to reproduce the SMAP analysis

Figure S3: Third attempt to reproduce the SMAP analysis

Figure S4: Fourth attempt to reproduce the SMAP analysis
Figure S5: SMAP workflow

Figure S6: First attempt to reproduce the FATCAT analysis
Figure S7: Second attempt to reproduce the FATCAT analysis

Figure S8: FATCAT workflow
Figure S9: First attempt for reproducing the docking analysis

Figure S10: Docking workflow
Figure S11: Visualization of the drug-protein network obtained as a result of the workflow

![Drug-Protein Network](image)

Table S1: Reproducibility scores for the SMAP workflow

<table>
<thead>
<tr>
<th>Workflow Component</th>
<th>MINIMAL</th>
<th>NOVICE</th>
<th>AUTHOR</th>
<th>Justification</th>
</tr>
</thead>
<tbody>
<tr>
<td>SMAP1 step</td>
<td>1</td>
<td>N/A</td>
<td>N/A</td>
<td></td>
</tr>
<tr>
<td>software component</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td></td>
</tr>
<tr>
<td>input1: ligand binding sites</td>
<td>1</td>
<td>N/A</td>
<td>N/A</td>
<td>&quot;The SMAP software was used to compare the binding sites of the 749 protein structures and the 962 binding sites of 274 approved drugs.&quot;</td>
</tr>
<tr>
<td>input2: drug binding site</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td></td>
</tr>
<tr>
<td>input3: configuration parameters</td>
<td>0</td>
<td>1</td>
<td>N/A</td>
<td>The original article does not mention configuration parameters. However, the default values of the parameters in the SMAP software seemed to work.</td>
</tr>
<tr>
<td>output1: ComparisonResults</td>
<td>N/A</td>
<td>N/A</td>
<td>“For each pairwise comparison a <strong>P-value representing the significance of the binding site similarity</strong> was calculated (the result of the component is a p-value)</td>
<td></td>
</tr>
<tr>
<td>--------------------------</td>
<td>-----</td>
<td>-----</td>
<td>------------------------------------------------------------------------------------------------------------------</td>
<td></td>
</tr>
<tr>
<td>output2: AlignmentResults</td>
<td>N/A</td>
<td>N/A</td>
<td>“Secondly, two protein structures are aligned, independent of sequence order…”</td>
<td></td>
</tr>
<tr>
<td><strong>SMAP2 step</strong></td>
<td>0</td>
<td>1</td>
<td><strong>N/A</strong></td>
<td></td>
</tr>
<tr>
<td>software component</td>
<td>N/A</td>
<td>N/A</td>
<td>“The SMAP software was used to compare […] <strong>1446 homology models</strong> with the <strong>962 binding sites</strong> of <strong>274 approved drugs</strong>. However, we have given a 0 in MINIMAL because reading it in the paper it looks like it is a single step.”</td>
<td></td>
</tr>
<tr>
<td>input1: homology models</td>
<td>N/A</td>
<td>N/A</td>
<td><strong>N/A</strong></td>
<td></td>
</tr>
<tr>
<td>input2: drug binding sites</td>
<td>N/A</td>
<td>N/A</td>
<td><strong>N/A</strong></td>
<td></td>
</tr>
<tr>
<td>input3: configurationParameters</td>
<td>N/A</td>
<td>N/A</td>
<td>The original article does not mention configuration parameters. However, the default values seemed to work.</td>
<td></td>
</tr>
<tr>
<td>output1: ComparisonResults1</td>
<td>N/A</td>
<td>N/A</td>
<td>For each pairwise comparison, a <strong>P-value representing the significance of the binding site similarity</strong> was calculated.</td>
<td></td>
</tr>
<tr>
<td>output2: AlignmentResults</td>
<td>N/A</td>
<td>N/A</td>
<td>“Secondly, two protein structures are aligned, independent of sequence order…”</td>
<td></td>
</tr>
<tr>
<td><strong>SMAPResultSorter1 step</strong></td>
<td>N/A</td>
<td></td>
<td><strong>N/A</strong></td>
<td></td>
</tr>
<tr>
<td>software component</td>
<td>1</td>
<td>N/A</td>
<td><strong>N/A</strong></td>
<td></td>
</tr>
<tr>
<td>input1: drugBindingSites1</td>
<td>N/A</td>
<td>N/A</td>
<td><strong>N/A</strong></td>
<td></td>
</tr>
<tr>
<td>input2: ComparisonResults</td>
<td>N/A</td>
<td>N/A</td>
<td><strong>N/A</strong></td>
<td></td>
</tr>
<tr>
<td>input3: pvalue</td>
<td>0</td>
<td>1</td>
<td><strong>N/A</strong></td>
<td></td>
</tr>
<tr>
<td>input4: structureForComparison</td>
<td>N/A</td>
<td>N/A</td>
<td>In the paper there is no reference to a possible sorting of the SMAP results. However, further exploration of the author’s scripts showed that each p-value is filtered and sorted for each structure for comparison (as a list) instead of having each comparison in an individual file.”</td>
<td></td>
</tr>
<tr>
<td>Step</td>
<td>Component</td>
<td>Input 1</td>
<td>Input 2</td>
<td>Output 1</td>
</tr>
<tr>
<td>-----------------------</td>
<td>-----------</td>
<td>--------------------------------</td>
<td>--------------------------------</td>
<td>-------------------------------</td>
</tr>
<tr>
<td>SMAPResultSorter2 step</td>
<td>component 0 1 N/A</td>
<td>input1: drugBindingSites1 N/A</td>
<td>input2: ComparisonResults N/A</td>
<td>output1: RawInteractionNetwork N/A</td>
</tr>
<tr>
<td>Merger step</td>
<td>component 0 1 N/A</td>
<td>input1: RawInteractionNetwork1 N/A</td>
<td>input2: RawInteractionNetwork2 N/A</td>
<td>output1: mergedOutput N/A</td>
</tr>
<tr>
<td>AlignmentResultMerger</td>
<td>component 0 1 N/A</td>
<td>input1: RawInteractionNetwork N/A</td>
<td>output1: mergedOutput N/A</td>
<td></td>
</tr>
</tbody>
</table>

The rationale is the same as the SMAPResultSorter1 step.

The original article does not mention this step, but once the two SMAP steps are apparent then the use of a merger step becomes obvious.

The original article does not mention this step, but once the two SMAP steps are apparent then the use of a merger step for gathering the alignment results becomes obvious.
Table S2: Reproducibility scores for the FATCAT workflow

<table>
<thead>
<tr>
<th>Workflow Fragment/Component</th>
<th>MINIMAL</th>
<th>NOVICE</th>
<th>AUTHOR</th>
<th>JUSTIFICATION</th>
</tr>
</thead>
<tbody>
<tr>
<td>GetSignificantResults step</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>This component is an intermediate step to a) Process the files resultant from the previous step; b) Produce a list for FATCAT in the appropriate format and a list of significant results to compare afterwards. The script provided by the authors showed this step. Experts could infer this step from a note in the article: “identify pairs of similar binding sites”.</td>
</tr>
<tr>
<td>software component</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>input1: MergedOutput</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>input2: TBproteinInfo</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>output1: List</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>output2: SignificantResults</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>FATCATURLchecker step</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>This step was not part of the original method. It was added to address the decay of protein IDs, and check whether the protein IDs are available in the PDB or they had been superseded by new ones. An expert would know to add this step in order to handle the changes in IDs.</td>
</tr>
<tr>
<td>component</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>input1: list</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>output1: CheckedList</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>FATCAT step</td>
<td>1</td>
<td>N/A</td>
<td>N/A</td>
<td>“the first chain of each PDB file was aligned using FATCAT”</td>
</tr>
</tbody>
</table>

http://semanticommunity.info/Data_Science/Data_Culture_at_the_NIH
Updated: Fri, 18 Sep 2015 23:53:15 GMT
Powered by Qmindtouch™ 124
<table>
<thead>
<tr>
<th>component</th>
<th>1</th>
<th>N/A</th>
<th>N/A</th>
</tr>
</thead>
<tbody>
<tr>
<td>input1: checkedList</td>
<td>0</td>
<td>1</td>
<td>N/A</td>
</tr>
<tr>
<td>configuration parameter (not explicit in workflow)</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>output: comparisonOutput</td>
<td>1</td>
<td>N/A</td>
<td>N/A</td>
</tr>
</tbody>
</table>

The original article mentions "the first chain of each PDB file was aligned using FATCAT". However, without relevant expertise it is not clear how these chains of PDB files are obtained from earlier steps.

The original article mentions "those pairs with a significant P-value of less than 0.05 were discarded". It seemed to be a configuration parameter, but it turns out to be part of how FATCAT works. There is no need for a configuration parameter, but only experts would know from the FATCAT publications.

"FATCAT reports the statistical significance of the structural similarity, measured as a P-value."

This step is not explicitly mentioned in the article. However, someone with some degree of expertise would see that the results of the SMAP comparison and the results of the FATCAT comparison need to be combined. That is what this step does.
### Table S3: Reproducibility scores for the docking workflow

<table>
<thead>
<tr>
<th>Workflow Fragment/Component</th>
<th>NOVICE</th>
<th>AUTHOR</th>
<th>QUOTE</th>
</tr>
</thead>
<tbody>
<tr>
<td>CreateClipFiles step</td>
<td>0</td>
<td>1</td>
<td>N/A</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>“The aligned drug molecule was used as the clip file with a default search space of 10Å³”</td>
</tr>
<tr>
<td>software component</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>This step had to be written for the new tool (AutodockVina). Advanced expertise is required to write software for this step.</td>
</tr>
<tr>
<td>input1: SMAPAlignmentFolder</td>
<td>0</td>
<td>1</td>
<td>N/A</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>“Since SMAP had aligned the drug binding site with the M.tb protein binding site, the aligned coordinates of the drug molecule were used to define the search space for docking that drug into the M.tb protein.” The text indicates that the SMAP software gives this result, but it does not return this result explicitly. It takes some investigation into the SMAP software to see that these results are placed in a special folder by SMAP.</td>
</tr>
<tr>
<td>output1: ClipFiles</td>
<td>0</td>
<td>1</td>
<td>N/A</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Advanced expertise is required to write software that generates the output needed.</td>
</tr>
<tr>
<td>CreateIdealLigandsImpl step</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>component</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>input1: clipFiles</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>This step is not in the paper, and by looking at the Autodock Vina software it could not be inferred without some expertise. The authors had written scripts for this step.</td>
</tr>
<tr>
<td>Component</td>
<td>Input1: idealLigandFiles</td>
<td>Input2</td>
<td>Output1: idealLigandsChecked</td>
</tr>
<tr>
<td>-----------</td>
<td>-------------------------</td>
<td>--------</td>
<td>-----------------------------</td>
</tr>
<tr>
<td>IdealLigandCheckerImpl step</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>step</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>component</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>input1: idealLigandFiles</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>output1: idealLigandsChecked</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>AutodockVina</td>
<td>1</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>component</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>input1: SMAPAlignementFolder</td>
<td>1</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>input2: IdealLigandsChecked</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>
These inputs were not specified in the original paper, but the software describes what inputs are needed.

“For those pairs of interest, molecular docking was used to predict the binding pose and affinity of the drug molecule to the M.tb protein”

Table S4: Top rated highly connected drugs (original results)

<table>
<thead>
<tr>
<th>Drug Name</th>
<th>Connection</th>
<th>Connections solved Structures</th>
</tr>
</thead>
<tbody>
<tr>
<td>Methotrexate</td>
<td>48</td>
<td>10 argB, aroF, cmaA2, cyp121, cyp51, Lpd, mmaA4, panC, Rv3676, TB31.7</td>
</tr>
<tr>
<td>Estradiol</td>
<td>38</td>
<td>10 argB, bphD, cyp121, cysM, inhA, mscL, pknB, Wv1264, sigC</td>
</tr>
<tr>
<td>Rifampin</td>
<td>34</td>
<td>6 inA, LpdA, LppX, mscL, prpB, Rv3676</td>
</tr>
<tr>
<td>Amantadine</td>
<td>32</td>
<td>0 homology models only</td>
</tr>
<tr>
<td>Propofol</td>
<td>24</td>
<td>3 clpP, glbN, InhA</td>
</tr>
</tbody>
</table>

Table S4: Top rated highly connected drugs (original results)

<table>
<thead>
<tr>
<th>Drug Name</th>
<th>Connection</th>
<th>Connections solved Structures</th>
</tr>
</thead>
<tbody>
<tr>
<td>Methotrexate</td>
<td>48</td>
<td>10 argB, aroF, cmaA2, cyp121, cyp51, Lpd, mmaA4, panC, Rv3676, TB31.7</td>
</tr>
<tr>
<td>Estradiol</td>
<td>38</td>
<td>10 argB, bphD, cyp121, cysM, inhA, mscL, pknB, Wv1264, sigC</td>
</tr>
<tr>
<td>Rifampin</td>
<td>34</td>
<td>6 inA, LpdA, LppX, mscL, prpB, Rv3676</td>
</tr>
<tr>
<td>Amantadine</td>
<td>32</td>
<td>0 homology models only</td>
</tr>
<tr>
<td>Propofol</td>
<td>24</td>
<td>3 clpP, glbN, InhA</td>
</tr>
</tbody>
</table>
Table S5: Top rated highly connected drugs, results of the workflow

The entries found in the original results have been highlighted in order to show the commonalities between both.

<table>
<thead>
<tr>
<th>Drug Name</th>
<th>Connections</th>
<th>Connections</th>
<th>Solved structures</th>
</tr>
</thead>
<tbody>
<tr>
<td>Methotrexate</td>
<td>156</td>
<td>32</td>
<td>argB, aroF, cyp121, cyp51, lpd, mmaA4, panC, Rv3676, TB31.7, Rv0223c, lipJ, echA6, Rv0948c,</td>
</tr>
<tr>
<td>4-Hydroxytamoxifen</td>
<td>115</td>
<td>25</td>
<td>cysM, inhA, pknB, pknE, Rv1264, Rv1941, Rv3676, cyp130, lppX, gpm1, ligA, folC, bioD,</td>
</tr>
<tr>
<td>Drug</td>
<td>Score</td>
<td>Rank</td>
<td></td>
</tr>
<tr>
<td>--------------------------</td>
<td>-------</td>
<td>------</td>
<td></td>
</tr>
<tr>
<td>Amantadine</td>
<td>79</td>
<td>1</td>
<td>fabG1</td>
</tr>
<tr>
<td>Raloxifene</td>
<td>75</td>
<td>18</td>
<td>inhA,mbtK,pknB,pknE,Rv1264,Rv3676,secA1,sigC,TB31.7,cyp130,aroG,trpD,coaA</td>
</tr>
<tr>
<td>Propofol</td>
<td>54</td>
<td>5</td>
<td>clpP,glbN,inhA,pth,ethR,</td>
</tr>
<tr>
<td>Penicillamine</td>
<td>44</td>
<td>10</td>
<td>groEL,inhA,nusA,Rv1264,Rv3676,lppX,secA1,glmU,glbN, mmmA4</td>
</tr>
<tr>
<td>Daunorubicin</td>
<td>44</td>
<td>12</td>
<td>mmmA4,Rv1264,thyX,lppX,secA1,serA1,Rv3529c,coaA,ptbB,cyp124,ethR,inhA,</td>
</tr>
<tr>
<td>Triclosan</td>
<td>42</td>
<td>5</td>
<td>pepD,Rv1264,thyX,ethR,trxB2</td>
</tr>
<tr>
<td>Darunavir</td>
<td>40</td>
<td>15</td>
<td>cyp125,devB,inhA,lpdA,panC,pknD,pepD,fabH,Rv1941,ppp,ftsZ,hrp1,argR,Rv3676</td>
</tr>
<tr>
<td>Desoxycorticosterone</td>
<td>39</td>
<td>12</td>
<td>mmmA4,Rv0813c,Rv1264,bpoC,thyX,lppX,cysK1,nusA,Rv0760c,sahH, pcaA,embR,</td>
</tr>
<tr>
<td>Diethylstilbestrol</td>
<td>39</td>
<td>7</td>
<td>folP2,Rv1264, pcaA,Rv3676,aroK,mshB,inhA,</td>
</tr>
<tr>
<td>Amprenavir</td>
<td>38</td>
<td>14</td>
<td>pknD,pepD,fabH,Rv3361c,ftsZ,lppX,hrp1,argR,panC,Rv3676,folC,nrdB,trxB2,adk,</td>
</tr>
<tr>
<td>Tadalafil</td>
<td>36</td>
<td>7</td>
<td>nirA,coaA,Rv3676,lppX,secA1,mshB,inhA,</td>
</tr>
<tr>
<td>Lopinavir</td>
<td>35</td>
<td>10</td>
<td>pknD,cyp51,pepD,fabH,ppnK,hrp1,argR,Rv3676, trpB,inhA,</td>
</tr>
<tr>
<td>Saquinavir</td>
<td>34</td>
<td>13</td>
<td>pknD,fabH,Rv1941,cyp130,devB,cyp121,serA1,ephG,hrp1,argR,accD5,adk,inhA,</td>
</tr>
<tr>
<td>Drug</td>
<td>Source 1</td>
<td>Source 2</td>
<td></td>
</tr>
<tr>
<td>--------------</td>
<td>----------</td>
<td>----------</td>
<td></td>
</tr>
<tr>
<td>Indomethacin</td>
<td>cyp51,cmaA2,gpm1,coaA,Rv3676,ethR,glbN, inhA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bepridil</td>
<td>cyp51,Rv1264,bpoC,pknE,secA1,cyp121,mmaA2,Rv3676,ethR,nrdB,ruvA,kasB,sigC,</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bexarotene</td>
<td>Rv1264,mscL,pknE,lppX,glnA1,mshB,uppS,PCA,Rv3676,inhA,</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Imatinib</td>
<td>mmaA4,groEL,lppX,secA1,trpD,coaA,lrpA,Rv3676,cmaA1,ald,dapB,inhA,</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nelfinavir</td>
<td>fabH,pknG,pknD,pepD,hrp1,argR,Rv3676,desA2,inhA,</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**RFI Summary**

Source: [http://bd2k.nih.gov/pdf/bd2k_training_rfi_summary.pdf](http://bd2k.nih.gov/pdf/bd2k_training_rfi_summary.pdf) (PDF)

My Note: I could not copy this PDF file cleanly. It is non-reusable! But I was able to recover it by copying and editing it extensively in NotePad (txt) with Word Wrap Off. I also removed the <br> and </br> markup where needed.

**Executive Summary**

My Note: I think we are doing all 5 specific suggestions for training in the [Federal Big Data Working Group Meetup](http://semanticommunity.info/Data_Science/Data_Culture_at_the_NIH)!

On February 20, 2013, the NIH issued a Request for Information titled “Training Needs In Response to Big Data to Knowledge (BD2K) Initiative.” The response was large, with 103 responses received from individuals, departments, companies, and professional organizations.

A theme that emerged throughout is that meeting the challenge of Big Data requires closely collaborating scientists who are experts in their domain but also cross-trained, so that they can form high-functioning teams.
There is a need to develop both human-focused and technical skills, the latter including skills in three basic categories: (1) statistics and math, (2) computation and informatics, and (3) a domain (e.g., biomedical, behavioral, clinical) science.

RFI respondents confirmed that education and training is needed at all professional levels, from undergraduates all the way through senior faculty and clinicians, and at varying levels of detail, with the assumption that all types of academic institutions need to be included. Respondents gave suggestions about how to train individuals for the new Big Data challenges: a consistent theme was the need for hands-on, immersive experiences with data and a deep understanding of the ideas underlying the methodology in order to learn persistent skills that transcend particular data types or tools. Specific suggestions for training include the following:

1) Modify features of existing training programs (e.g. use dual mentors; start with boot camps to teach basics of less familiar area; offer joint classes and assignments that foster teamwork cross the two areas)

2) Offer short-term experiences and courses on focused topics (workshops, lectures, lab rotations)

3) Develop novel and technology-enabled learning systems that reach a wide audience (MOOCs, online courses, online/physical hybrids, cognitive tutors)

4) Develop and share new curricula (new courses focused on data science principles and practices, new modules to build into existing courses, use cases that can be incorporated into existing classes)

5) Encourage broad dissemination (materials should be publicly available online under open license and shared through web portal of online learning resources)

RFI Summary: Full Report

On February 20, 2013, the NIH issued a Request for Information titled "Training Needs In Response to Big Data to Knowledge (BD2K) Initiative." The response was large, with 103 responses received from individuals, departments, companies, and professional organizations.

Knowledge and Skills Needed to Utilize Biomedical Big Data

Both technical knowledge and skills and human-based competencies are needed to make discoveries with Big Data. Human-based competencies include working well in teams, communicating, and ethics training. At the highest level, technical knowledge and skills could be classified as coming from three major areas: (1) statistics and mathematics, (2) computational sciences, and (3) domain science. A solid foundation in biomedical or behavior domain science, by individuals or within teams, was considered essential.

Also essential for Big Data teams is knowledge or expertise in fundamental “data skills” to manage and process big data, which may require programming in a variety of languages. Platforms such as R or SAS and SQL-based databases are sufficient to a certain scale but eventually break as the data size increases and data sorting and aggregation become more complex. This is the entry point for what can be called Big Data technologies, such as cloud computing, MapReduce, NoSQL, and parallel databases. Scientific and information visualization is critical to make sense of the data. A common theme in the RFI responses was the need for semantic skills with ontologies and metadata, and it was pointed out that academic library science is a good source of this expertise. Efforts to include all possible types of Big

http://semanticommunity.info/Data_Science/Data_Culture_at_the_NIH
Updated: Fri, 18 Sep 2015 23:53:15 GMT
Powered by mindtouch
Data are necessary, particularly from the clinical perspective but also from disciplines outside the typical biomedical sciences.

Much of the data analysis itself involves statistics and machine learning methods. High-level analysis includes modeling, statistical inference, decision theory, and evaluation of properties of estimators. A compelling need for data scientists with the skills for reproducible research was observed. Finally, new ways to present and visualize Big Data are important in fostering greater utilization and understanding of the opportunities.

**Teamwork and Cross-training**

A theme that emerged throughout is that meeting the challenges of Big Data requires teams of closely-collaborating scientists who have been cross-trained in multiple areas. High-functioning teams are critical for success in utilizing Big Data. They are a practical and effective way to find, develop, deploy, and retain the needed combination of skill sets across disciplinary boundaries. Relatively few respondents maintained that a perfect bio-computational hybrid could be trained or was practical. Team members need an “empathy (or at least sympathy) and respect for those in other disciplines.” As one respondent described, “Computation-centric researchers cannot just be viewed as human computers, to whom one delivers Excel spreadsheets and from whom one receives ‘results’. multi-disciplinary teams need to share a common goal and have a vested interest in other team members’ success.”

Another common theme among the responses was the need to cross-train individuals across traditional disciplinary boundaries such as biology and computation. Cross-training is needed so that team members can communicate effectively with each other. Successful cross-training should help individuals to develop and/or strengthen their core expertise in one domain (or more) while learning about, connecting with, and becoming strongly committed to another domain. An effective joint curriculum will strengthen both the data awareness of biomedical researchers and the bioscience awareness of computational researchers; thus, strong teams and cross-training have the same ultimate purpose: to produce successful, expert collaboration.

**Audience for Data and Informatics Training and Education**

RFI respondents confirmed that education and training are needed at all professional levels, from undergraduates all the way through senior faculty and clinicians, and at varying levels of detail for both basic and clinical scientists:

- Practicing scientists and clinicians may need to update their skills to take better advantage of Big Data resources for basic research and the translation of information for health care.
- Trainees (postdocs, doctoral students, Masters-level students, postbacs) should be exposed to data analysis and informatics either in dedicated courses or woven throughout their training.
- Undergraduates should be exposed to more opportunities to develop quantitative and computational skills to improve the baseline skills of entering graduate students.

Respondents also advocated for promoting data science at the high school level to engage interest and cultivate the foundational knowledge of students early in the pipeline. Although NIH rarely enters the K12 domain (the development and distribution of NIH curriculum units is an exception), online courses and modules in introductory subjects would be accessible to K12 students. Another respondent pointed out that managers of the research enterprises in academia and industry need to understand the existing inherent potential of Big Data, as well as the means to achieve functioning collaborative teams.
Some respondents described the difficulty of finding and retaining skilled individuals who often can be better compensated in industry. However, innovative programs and flexible environments could also attract individuals with cutting-edge skills to new intellectual challenges. It is important for NIH to address not only attracting Big Data talent to biomedical research, but also keeping that talent by encouraging the development of rewarding career paths that may accommodate rapidly shifting priorities, responsibilities, and opportunities.

Respondents were supportive of efforts to help increase diversity in the workforce and gave some concrete suggestions to achieve this goal:

- Encourage major public and private institutions to partner with minority serving institutions such as HBCUs and HSIs
- Provide access to educational and technical resources necessary to teach students Big Data skills such as hardware and software or access to cloud computing
- Engage high school students in data, e.g. encourage data management and analysis components in science fair projects or in classroom projects by working with the National Association of Biology Teachers.

Solving the problems of Big Data will a much larger and diverse workforce, so education and training for all interested groups – regardless of professional level, demographic, or educational background – is important.

**How Training Should be Achieved**

Because the target audience is diverse both scientifically and by career stage, a variety of approaches is needed to deliver the knowledge and skills needed to utilize biomedical Big Data. Suggested approaches include development of (1) training programs for long-term training, (2) short-term experiences, and (3) innovative use of technology to reach a broad scale. Also emphasized was the need for curriculum development and sharing, as well as the development of resources for teaching and learning.

1. **Modify the features of existing training programs** to produce cross-trained individuals who work well in Big Data teams while having a deep knowledge in multiple areas. Features include

   - Dual mentors, one in each domain,
   - Boot camps to teach entering graduate students the basics of the less familiar area, and
   - Joint classes and assignments that foster teamwork across the two areas.

A respondent suggested that NIH require that all existing NIH-supported training programs ensure students obtain a set of minimal competencies in computational and statistical methods for biological research and that the availability and success of such efforts be a proposal evaluation criterion.

2. **Develop short-term experiences and courses** on focused topics necessary for Big Data, aimed at all career levels. Examples include the following:

   - Short-term workshops (including lectures, lab rotations, and shadowing) to allow one group of researchers (e.g., biologists, other basic scientists, clinicians, informaticians) to learn more about the field of a different group of researchers and to help break down some of the communication barriers by developing a common vocabulary.
   - Summer intensive programs with topics of general interest (e.g. data mining) along with breakouts that provide specific information (e.g. particular databases or tools)
   - In person workshop attendance could be optimized by having participants participate in some online training sessions before coming together as a group.
Summer programs that aim to recruit undergraduate students, giving intellectually motivated students early exposure to Big Data challenges, similar to the SIBS program in biostatistics.

(3) **Develop novel or technology-enabled learning systems or environments** to reach a wide audience including trainees and researchers. Distance learning has the potential to engage non-traditional sources of talent; in this case, it could be used to expose biomedical scientists to data science and vice versa.

- Examples include MOOCs, online courses, online/physical hybrids, and cognitive tutors.
- Web-based and flexible educational offerings, including certification programs, are being rapidly created and disseminated, particularly in the quantitative and computational sciences.

All three of these approaches to knowledge acquisition require curricula (for new courses, new modules, and use cases). The curricula should have the following content and characteristics:

- Cover the intersection of data science and biomedical science that fuse topics from informatics and the computational and quantitative areas with health, behavioral, and clinical topics, such as health service research.
- Include mentored research projects, which requires access to faculty with appropriate expertise and datasets that can answer important research questions in health and healthcare.
- Include modules (e.g. CME courses) for clinicians (as well as medical students and residents) that focus on skills needed to utilize large datasets, as the importance of Big Data in everyday clinical situations grows.

RFI respondents stressed the need not only to develop but also to share new curricula. Broad dissemination of curricula should be encouraged, including making them publicly available online under open license and shared through a web portal of online learning resources. Sharing is a theme that applies not just to curricula but also to training itself through online lectures, to data sources, and even to virtual machines on the cloud pre-loaded with data and tools so that students and instructors can skip the difficulties of installation and configuration.

The themes that emerged from the RFI responses centered around inclusiveness. Respondents urged NIH to push for sharing of resources such as curricula and courses through web-based portals. They identified knowledge and skills, whether quantitative, computational, or domain, that are needed at all professional levels. Finally, they described the necessity of teams, inclusive of informatics experts, computational scientists, quantitative scientists, as well as the biomedical scientists who are domain experts.

**Workshop Report on Enhancing Training**

for Biomedical Big Data Big Data to Knowledge (BD2K) Initiative

Source: [http://bd2k.nih.gov/pdf/bd2k_trainin...hop_report.pdf](http://bd2k.nih.gov/pdf/bd2k_trainin...hop_report.pdf) (PDF)

July 29-30, 2013

**Introduction**

Big Data to Knowledge (BD2K) is a new NIH initiative that aims to enable scientists to effectively manage and utilize the large, complex data sets (Big Data 1) that are already being generated and whose number and value will only increase in the future. The BD2K initiative is based on a set of recommendations presented on June 12, 2012 by the Data and
Informatics Working Group (DIWG) to the Advisory Committee to the Director, NIH. The DIWG report can be found at http://acd.od.nih.gov/diwg.htm.

"Big Data" is meant to capture the opportunities and address the challenges facing all biomedical researchers in releasing, accessing, managing, analyzing, and integrating datasets of diverse data types.

Such data types may include imaging, phenotypic, molecular (including -omics), clinical, behavioral, environmental, and many other types of biological and biomedical data. They may also include data generated for other purposes. The datasets are increasingly larger and more complex, and they exceed the abilities of currently-used approaches to manage and analyze them. Biomedical Big Data primarily emanate from three sources: 1) a few groups that produce very large amounts of data, usually as part of projects specifically funded to produce important resources for the research community; 2) individual investigators who produce large datasets for their own projects, which might be broadly useful to the research community; and 3) an even greater number of investigators who each produce small datasets whose value can be amplified by aggregating or integrating them with other data.

One of the DIWG recommendations was to “Build Capacity by Training the Workforce in the Relevant Quantitative Sciences such as Bioinformatics, Biomathematics, Biostatistics, and Clinical Informatics.” The NIH organized the “Workshop on Enhancing Training for Biomedical Big Data” as one approach to obtaining input from the biomedical data science community on priorities for training needs and activities. The Workshop was co-chaired by Karen Bandeen-Roche (Professor and Chair of Biostatistics, Bloomberg School of Public Health, Johns Hopkins University) and Isaac Kohane (Professor of Pediatrics and Chair of the Bioinformatics Program, Boston Children’s Hospital and Dana-Farber/Harvard Cancer Center).

In this document, the term “biomedical” will be used in the broadest sense to include biological, biomedical, behavioral, social, environmental, and clinical studies that relate to understanding health and disease.

Purpose and Goals of the Workshop

Michelle Dunn presented the purposes of the workshop as the following: (1) to identify the knowledge, skills, and resources that the biomedical research enterprise needs to organize, process, manage, analyze, and visualize large, complex data sets, and (2) to make recommendations on specific objectives for the NIH in the area of training for the utilization of Big Data and their priorities. The full rationale for the workshop, as distributed in advance to the participants, can be found in Appendix I, the agenda in Appendix II, the roster of participants in Appendix III, and a list of members of the NIH BD2K Training Working Group in Appendix IV. An archived videocast of the workshop can be found at https://videocast.nih.gov/PastEvents.asp.

Dr. Dunn described the issues that need to be addressed in “Big Data” training in four “dimensions”: (1) data that span the NIH mission; (2) applications that span the pipeline from data acquisition and processing to data analysis; (3) scientists, from developers to users; and (4) career stage, from students to professionals.
She noted that the workshop participants embodied many disciplines and scientific areas (including informatics, computational biology, biostatistics, biology, genomics, mathematics, computer science, and education), but that not all relevant disciplines were able to be represented. Therefore, she asked the participants to think broadly, beyond their specific scientific expertise or disease of interest, and to focus in the workshop on the overall needs and priorities for BD2K training. Finally, Dr. Dunn stated that the outcome of the workshop deliberations would be used by NIH staff to develop training initiatives designed to prepare and empower the biomedical research workforce to take full advantage of Big Data for research into the understanding of human biology and improving human health.

Background Relevant to the Workshop Discussion

In June 2012, three reports were presented to the NIH Director by working groups of the NIH Advisory Committee to the Director (ACD), all of which are relevant to training and careers in the area of Big Data. Dr. Sally Rockey presented the recommendations of, and NIH follow up to, two working groups: the ACD Working Group on the Biomedical Workforce and the ACD Working Group on Diversity in the Biomedical Research Workforce. A summary of Dr. Rockey’s presentation can be found in Appendix V. Dr. Mark Guyer presented the recommendations of and follow up to the ACD Working Group on Data and Informatics.

ACD Working Group on Data and Bioinformatics (DIWG)

The DIWG made five recommendations:

1. promote data sharing through central and federated catalogues;

2. support development, implementation, evaluation, maintenance, and dissemination of informatics methods and applications;

3. build capacity by training the workforce in the relevant quantitative sciences;

4. develop an NIH-wide "on-campus" IT strategic plan; and

5. provide a serious, substantial, and sustained funding commitment to enable recommendations 1-4.

The NIH’s initial response to the DIWG report has three components:

- **Appointment of an Associate Director for Data Science (ADDS)** to lead the trans-NIH effort in data science, including the development of a long-term strategic plan. The ADDS will also be the primary NIH focus for...
coordination with data science activities beyond the NIH. Dr. Eric Green, Director, National Human Genome Research Institute, is currently the Acting Director, and a search is underway for a permanent ADDS.

- **Creating a Scientific Data Council** as a high-level internal NIH committee that, working with the ADDS, will provide oversight for trans-NIH data science activities. The Council will be chaired by the ADDS, and its members will be senior leaders from across the NIH.

- **Implementation of the Big Data to Knowledge (BD2K) Initiative** ([http://bd2k.nih.gov](http://bd2k.nih.gov)) as the programmatic arm of the trans-NIH activities in biomedical Big Data. The overarching goal of the BD2K Initiative is to enable, by the end of this decade, a quantum leap in the ability of the biomedical research enterprise to maximize the value of the growing volume and complexity of biomedical data. There is wide-spread support for BD2K across NIH, with at least 24 NIH institutes/centers/offices participating in the initiative.

The DIWG report discussed a number of major problems in the use of Big Data:

- locating the data;
- getting access to the data;
- organizing, managing, and processing the data;
- developing new methods for analyzing data; and
- finding trained researchers who can utilize the data effectively.

The DIWG also noted that cultural changes at NIH are needed, including new approaches to data sharing and recognition that extracting the value of Big Data will require significant resources, i.e. data handling can no longer be considered to be free.

BD2K has identified four programmatic approaches to addressing the DIWG’s recommendations:

I. Facilitating Broad Use of Biomedical Big Data;
II. Developing and Disseminating Analysis Methods/Software for Biomedical Big Data;
III. Enhancing Training for Biomedical Big Data; and
IV. Establishing Centers of Excellence for Biomedical Big Data.

Initial BD2K activities are focused on planning by means of workshops (of which this is the first) and obtaining community input through Requests for Information. The first BD2K Funding Opportunity Announcement, for BD2K Centers of Excellence, was just published ([http://grants.nih.gov/grants/guide/r...HG-13-009.html](http://grants.nih.gov/grants/guide/r...HG-13-009.html)). NIH has set aside $27M in FY2014 for BD2K and has plans to scale up funding up to approximately $100M per year by FY2016.

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

In preparation for the BD2K training workshop, the NIH issued a Request for Information ([http://grants.nih.gov/grants/guide/n...HG-13-003.html](http://grants.nih.gov/grants/guide/n...HG-13-003.html), see Appendix VI) in which the community was invited to comment on 1) the skills and knowledge needed by a BD2K workforce, 2) the characteristics and content of plans for cross-training at all career levels, and 3) how to develop a diverse BD2K workforce. More than 100 responses were received. The NIH BD2K Training Working Group analyzed the responses, and the results were sent to the participants in advance of the workshop. Richard Baird presented a brief summary of the responses.

- **Who to Train:** The BD2K workforce will need both quantitative (statistical and computational) expertise and biomedical domain expertise, taken together as “data science” expertise. Examples of biomedical fields that already incorporate varying amounts and mixtures of quantitative expertise are bioinformatics, computational biology,
biomedical informatics, biostatistics, and quantitative biology. Both basic and clinical researchers at all career levels need to receive training.

• When to Train: Training is needed at all career stages: exposure courses for undergraduates, cross-training for graduate students and postdoctoral fellows, training as needed for researchers at all levels to facilitate their work, refresher courses or certificates in specific competencies for mid-level researchers, and relevant continuing medical education courses for clinical professionals.

• What to Train: Both long- and short-term training is needed, and efforts should be guided by the competency level required for the technical knowledge and skills to be gained. The technical knowledge and skills needed include: (1) computational and informatics skills; (2) mathematics and statistics expertise; and (3) domain science knowledge.

• How to Train: Several ways to cross-train biomedical and quantitative scientists were suggested, including through (1) new or expansion of existing long-term research training programs (which can incorporate activities such as boot camps, joint and team coursework, delayed laboratory rotations, dual or team mentoring, clinical and industrial externships, and team challenges); (2) short-term courses and hands-on immersive experiments (which can span short courses, certificate programs, immersive workshops, summer institutes, clinical immersion and shadowing, and continuing medical education opportunities); (3) curricula for biomedical Big Data; (4) technology-enabled learning systems and environments (e.g., web-based courses and Massive Open Online Courses (MOOCs) to offer training to a much larger audience; and (5) a training laboratory that has tools and resources for self-directed learning and exploration.

Four Examples of Big Data Challenges and Competencies Needed to Make Full Use of the Data

Four types of Big Data research problems were presented as examples of the opportunities offered, the challenges presented, and the competencies needed.

Electronic Health Records (Daniel Masys)

Dr. Masys described Big Data as having two main characteristics: it exceeds the capacity of unaided human cognition for its comprehension, and it strains current technology capacity and is therefore CPU-bound, bandwidth-limited, and/or storage-limited. Electronic Health Records (EHRs) may contain many data types, including quantitative clinical measurements; textual lab reports; narratives; images and signals used to construct images; DNA sequence (and increasingly in the foreseeable future, gene expression, proteome, and metabolome values); complex physiological signal data; as well as billing data, demographics and other coded name-value-pair data, consents and other legal instruments, and e-mail and other forms of patient-provider and provider-provider communications. As a collection of data, important characteristics of the EHR are time sensitivity, the inclusion of both objective and subjective information, an inherent structure understood by users but with the data not always recorded in a structured format, and its confidential nature. Also, the EHR data have primary and secondary uses. Dr. Masys used the Electronic Medical Records and Genomics (eMERGE) Network as an example to describe the challenges of extracting phenotypic data from EHR for use in genotype-phenotype studies. In his view, competencies needed for the use of EHR in this type of correlation research include expertise in human physiology and disease pathophysiology, molecular biology and molecular genetics, clinical documentation rules and business practices, data modeling and database design, natural language processing, use of controlled vocabularies and ontologies, and biostatistics, particularly of methods for association testing using noisy high dimensionality data. Finally, Dr. Masys noted that this type of research typically requires an interdisciplinary team of five to seven people in which each team member has knowledge that spans more than one area.
Imaging (Ron Kikinis)

The amount of imaging data being generated is increasing from gigabytes to terabytes, is becoming more complex, and has more modalities and applications than ever, including both research and clinical. Challenges to using imaging data include the large number of subjects, the length of time (up to years) needed for analysis, and quality assurance. Logistics are challenging in several ways, including standardizing imaging equipment and protocols, getting the patient to the scanner, getting the data to the data center, and post-processing that requires an automated pipeline and large computational resources. As an example, Dr. Kikinis discussed a multi-center COPD genetic epidemiology study that has 21 clinical sites, three image analysis centers, two imaging platforms, involves two contrast mechanisms per visit and two visits per subject, has four processing pipelines, and has collected information on 10,000 subjects. The total analysis takes 320,000 CPU hours per run of the processing pipeline. In his view, competencies needed in imaging research include expertise in medical image computing, medical informatics, image acquisition, and domain science. Cross-training takes three to five years and is apprentice-style. An additional challenge to the use of imaging data is the need for robust, user-friendly tools, the development of which is hindered because tool creation applications generally do not fare well in the NIH peer review system. Dr. Kikinis pointed out that the medical imaging community is a compact one, making it a well-suited as a test bed.

Genomics (Michael Boehnke)

The decreasing cost of sequencing a genome (now less than $10,000 per genome) allows many more genomes to be sequenced, generating much more data to process and analyze. Many current research projects are generating hundreds of terabytes of data. Challenges in using these data include processing raw sequence image files into useable data, aligning sequence reads to the human reference sequence, building error models to allow accurate variant calling, identifying and accounting for DNA sample contamination, imputing dense genotypes from a reference set of sequenced genomes to genomes with less dense genotype data, testing disease-genetic variant association in sequenced and imputed data, and combining data and results across studies. Other challenges include the large amounts of CPU time to analyze the data and memory to store the data; data storage is a major challenge because multiple copies of the data are required (since different software requires different versions of the data) and processed data sets may be as large as the raw data set. In Dr. Boehnke’s opinion, dealing with genomics data requires knowledge in more than one scientific discipline, an aptitude to be actively engaged with the data in order to understand its context and identify problems, the ability to work in teams and communicate with experts in other disciplines, and creativity and flexibility to deal with a rapidly changing landscape. He also emphasized that producing well trained, cross-disciplinary scientists takes longer than training single-discipline scientists and that training needs may differ for developers, creative users, and general users of these methods.

Integrations of Large and Small Datasets (Mark Musen)

Dr. Musen noted that there are a number of challenges to effective data integration. Many databases are not robust, attracting individuals to develop standards is difficult (standards development is not exciting but is essential to data integration), and obtaining support for the development of standards is difficult. He then provided several examples that have developed effective approaches in both biological and clinical arenas. In Dr. Musen’s opinion, trainees need to understand the processes and frameworks needed for data integration. There is a spectrum of data integration, from the very ‘heavy’ integration tools needed for using data warehouses like i2b2 to ‘light’ serendipitous mashups that support discovery of associations. In the latter case, data are integrated on a ‘just in time’ basis, while in the former, data are
integrated on a ‘just in case’ basis. To integrate data, however, people need to find them, use standard metadata descriptions, use standard ontologies to create value sets, and represent the data in frameworks at the right level of granularity (warehouse to mashup). The next generation of investigators will need to understand how to: model biomedical domains to create new ontologies and new metadata specifications, evaluate the appropriateness of an ontology for a given data-integration task, search for data sets using relevant ontologies, and apply semantic technology at different locations on the data-integration spectrum (from data warehousing to mashups).

Summary of the Workshop Deliberations and Recommendations

The discussions ranged broadly over many issues relevant to data science, extracting knowledge from Big Data, and training; many specific ideas and suggestions were offered. There were, however, a number of themes that consistently ran through the different sessions and topics.

- The opportunity for extraction of knowledge from Big Data is often greatest at the intersection of at least two disciplines, and training programs should be designed to develop the ability to work at intersections.
- Multi-disciplinary approaches are critical to taking advantage of Big Data to advance biomedical science and knowledge. While some individuals with skills and expertise in several disciplines will be able to operate on their own as independent investigators, most of the relevant work will be done in well-integrated, multi-disciplinary teams.
- Training programs should be oriented toward providing trainees with the skills to work effectively in Team Science. This will often involve offering the opportunity to develop in-depth expertise in at least two scientific disciplines. To extract knowledge from data, it will be particularly useful if at least one of those disciplines is a quantitative discipline.
- Dual mentoring should be encouraged.
- There is no one right way to implement Big Data training, and it will be important for NIH to allow enough flexibility in its support for this type of training. Flexibility is needed to encourage innovative approaches and to allow training programs at different institutions to take best advantage of the particular talents and expertise available locally. The majority of the workshop attendees did not think it was a good idea to require that all NIH-supported training grants have a required data science component beyond the teaching of the principles of data science.
- The training experience can be enhanced if the trainees have access to large data sets, of multiple types, including -omics, imaging, and clinical data. NIH was encouraged to explore the idea of developing and providing such training sets as a resource.
- Training in quantitative science and experimental design will be increasingly important to clinical researchers and also clinicians. It would be desirable to add such training into medical school curricula, but that will not be easy. It might be easier to add such training to the pre-medical experience. It was also suggested that incorporating questions or problems that require knowledge of quantitative sciences as part of medical board exams would have a strong influence on training curricula.
- The principles of reproducible research should be stressed.
- There are training needs across the full spectrum of scientists, from technology- and tool- developers, to technology- and tool- users, to those who need to be conversant with the challenges and solutions related to big data.
- Realistic goals and limitations must be recognized for short-term training of non-experts, where training should equip the learner to understand enough about the quantitative analysis and tools available to collaborate with expert users or developers in data acquisition and analysis.
- The jobs necessary for Big Data science may not correspond to traditional scientific, particularly academic, jobs. Training individuals to participate across the full spectrum of scientific roles is desirable. In addition, an appropriate career path must be available when those individuals finish their training.
- A diverse workforce should be a major goal of data science training efforts.

http://semanticommunity.info/Data_Science/Data_Culture_at_the_NIH
Updated: Fri, 18 Sep 2015 23:53:15 GMT
Powered by mindtouch™
In addition to these overarching themes, there were many interesting points made in the individual sessions of the workshop, although not all of these were consistent with one another.

Knowledge and Skills Needed

Workshop participants discussed the knowledge and skills that biomedical data science teams and trainees should have as well as strategies for fostering their development.

- It is important for trainees to develop both quantitative (computational and statistical) and domain expertise.
- Working in a scientific team requires learning how to participate in active collaborations, including being respectful of the contributions of those with complementary expertise, being committed to working together, and fostering open communication.
- Faculty trainers must be close collaborators, supportive of the team approach, and have an awareness of and appreciation for all the areas that make up the team.
- Data sharing is critical and is becoming the norm. The concepts of managing and sharing data should be introduced early and continue throughout the training experience.
- Online modules, patterned after the human subjects training modules, could be developed and made widely available to teach principles of data sharing.
- The need to develop multidisciplinary capabilities is likely to lengthen the training period, in contrast to current interest in shortening it.

Long-term Training and Career Award Programs

There are a number of interesting approaches that have worked well in particular training programs, including the following:

- Organizing trainees with complementary expertise to work as a team to solve a specific problem.
- Holding boot camps at the beginning of training programs to introduce trainees to disciplines outside of their experience, e.g., quantitative skills to those with a biology background or biological approaches to those with a quantitative background.
- Pairing advanced graduate students with early-stage PhD students to solve a difficult data analysis problem.
- Having graduate students pursue rotations later in the didactic part of their training so that they have a firmer grasp of the principles of data science as they experience different laboratories.

In terms of environment, it is important to have access to the infrastructure needed to manipulate and analyze large data sets. Partnerships between large institutions and smaller institutions, including community colleges, liberal arts colleges, and minority-serving institutions, can be an effective approach to improving access to training. They are also effective in increasing the flow of students from those institutions into graduate training programs.

Short-term Training Programs

Short-term training opportunities should be made available for both basic and clinical scientists at all career levels to provide ongoing training and career enrichment to both new and established investigators. Short-term training can be used to recruit new people into particular research fields or to allow people to bring new fields into their research programs. Short-term training experiences can take many forms and serve different audiences and purposes, including the following:

http://semanticommunity.info/Data_Science/Data_Culture_at_the_NIH
Updated: Fri, 18 Sep 2015 23:53:15 GMT
Powered by mindtouch
• Workshops
  ◦ To bring trainees from different institutions/programs together
  ◦ For postdoctoral fellows or established scientists to learn new techniques, knowledge, skills, in new or familiar areas of science
  ◦ For experienced investigators to come together several times a year for a couple of days to solve interdisciplinary problems
• Boot camps and summer institutes
• Case study workshops
  ◦ For discussing new solutions to difficult data science or unsolved problems
• Modular training
  ◦ To provide graduate students or postdoctoral fellows an in-depth review of a particular scientific discipline
• Continuing medical education
  ◦ To provide clinicians with information about the complexities of electronic health records
• Team challenges
  ◦ In which groups of students compete to find the best solution for a research design
• Code-a-Thon-like intensive experiences
  ◦ In which individuals with varying expertise come together for a short period to solve a problem that is stated in advance (http://www.health2con.com/devchallenge/code-a-thons/)

Innovative Training Technology

The workshop participants agreed that innovative technology could be used to enhance the experience of face-to-face training and to extend training offerings to a larger audience. Examples of innovative uses of technology include the following:

• Massive Open Online Courses (MOOCs).
  ◦ Advantages include broad reach, scalability, and flexibility.
  ◦ Disadvantages include a lack of physical connection and the inability of teachers to adjust to individual students (although they can adjust based on problems the whole class is having).
  ◦ For which a number of issues must be addressed, such as evaluation of success, the initial expense, and updating/access once the course ends (in a rapidly changing field, updating is particularly important)
• Web-based videos and syllabi.
• Online learning tools.
• Technology-driven personalized content.
• Community-controlled online platforms for information sharing.

Data Sharing

There was a considerable amount of discussion of the need for data sharing in the context of optimal use of Big Data. One approach is to create a data center: large, controlled-access online data sets, together with analytic tools and a (potentially distributed) computing environment, which would provide a sandbox for training and education unique to BD2K. Although there were differing opinions on the value of this approach and not all training will require such online
data sets, a small number of widely accessible resources would enable acquisition of key competencies across a large number of trainees and researchers in a cost-efficient manner. Access control and privacy issues, which can be taught via large online data sets, are as integral to training as analytics.

**Curriculum Development**

Curriculum development was considered in two contexts: as an integral part of an institutional training program and as a standalone curriculum. Broad sharing of curricula across institutions, especially community colleges, small institutions, and minority-serving institutions, was considered to be an important opportunity.

- Sharing outside the group for which the curriculum was designed will require that it be made publicly available and kept up-to-date.
- A tiered curriculum for groups needing different levels of knowledge, e.g. a curriculum to cross-train quantitative and non-quantitative students or members of a research team who bring different expertise to help solve the research problem, should be considered.

**Prioritization**

In the final session of the workshop, each participant offered his or her single highest priority. Many priorities were identified, in which the common themes were interdisciplinary and team-based training, diversity, openly-accessible data, and flexibility in approach to encourage multiple approaches to training.

**Appendix 1: Rationale for the Workshop**

Big Data to Knowledge (BD2K), a new NIH initiative, aims to enable scientists to effectively manage and utilize the large, complex data sets (Big Data) that are already being generated and whose number and value will only increase in the future. The BD2K initiative is based on a set of recommendations on data and informatics from a working group to the Advisory Committee to the Director, NIH (see [http://www.nih.gov/news/health/dec2012/od-07.htm](http://www.nih.gov/news/health/dec2012/od-07.htm)).

The NIH seeks to increase the ability of the scientific workforce to utilize biomedical Big Data. Big Data creates challenges to the data pipeline, from acquisition and processing of the data to analysis and visualization. Utilization and analysis of this data will require new knowledge and skills beyond those traditionally employed in biomedical research. Furthermore, such abilities will be required at all levels, from students through established faculty, in a diverse and sustainable workforce. The workshop will, therefore, consider both a refocus of traditional training programs toward being cross-disciplinary, and the development of focused, short-term training programs that are potentially technology-enabled, web-based, or otherwise widely accessible to investigators at all levels.

The workshop will (a) identify the knowledge and skills needed by individuals and by collaborating teams to work productively with biomedical Big Data, and (b) discuss new resources and programs for educating and training both students and practicing scientists with the necessary knowledge and skills. The workshop will address the long- and short-term training needs of professionals and trainees with the purposes of increasing the number of (1) informaticians and computational and quantitative scientists who wish to apply their skills and knowledge in the biomedical, behavioral, and clinical sciences and (2) biomedical, behavioral, and clinical scientists who have the requisite knowledge and skills to effectively access, organize, analyze, and integrate large and complex data sets.
Appendix II: Agenda Big Data to Knowledge (BD2K) Workshop

on Enhancing Training for Biomedical Big Data
29-30 July 2013
Terrace Level Conference Room
5635 Fishers Lane, Rockville, MD

Workshop Co-chairs: Karen Bandeen-Roche and Isaac Kohane

Workshop Goals:

1) Identify the knowledge, skills, and resources needed by biomedical research to organize, process, manage, and utilize large, complex data sets, and
2) Recommend and prioritize specific objectives for the NIH in training for Big Data.

This information will be used by NIH staff to develop short- and long-term training initiatives that prepare and empower the community to maximize the use of Big Data for research aimed at understanding human biology and improving human health.

My Note: It would be nice to have summary notes/bullets inserted under each of these agenda items.

Monday, 29 July

10:00 Welcome, Introductions, and Mark Guyer

Overview of BD2K Initiative

10:30 Purpose of the Workshop Michelle Dunn

10:45 Summary of Request for Information Responses Richard Baird

11:15 Intersection of BD2K with Director’s Workforce and Sally Rockey

Diversity Initiatives

11:45 Discussion of the Goal and Vision for BD2K Training K. Bandeen-Roche, Z. Kohane

12:15 Lunch – not provided

1:15 Data Challenges and Competencies Needed (10 min presentation + 10 min discussion)
  • Electronic Health Records Dan Masys
  • Imaging Ron Kikinis
  • Genomics Mike Boehnke
  • Integration of Large or Small Datasets Mark Musen

2:45 Discussion of BD2K Knowledge and Skills Participants
• What are the necessary knowledge and skills that a Big Data team must include?
• How do the knowledge and skills needed vary according to the individual’s:
  ◦ Primary relationship to Big Data?
    • needing to be conversant
    • applying routine methods and tools
    • leading novel applications developing new methods and tools
  ◦ Primary training as basic, clinical, or quantitative scientists?
• How do we allow institutions adequate flexibility and still achieve the BD2K goals?

3:15 Break: Refreshments will not be provided

3:30 BD2K Characteristics of Long-term Training and Career Award Programs Participants

Approach
• What type of person should long-term training aim to produce?
• How should individuals be cross-trained?
• How could the curriculum and other program components be modified or developed so that a cross-trained student would not have a longer time from matriculation to graduation?
• The generation of new methods and software are essential for biomedical Big Data. Since computational and quantitative skills are broadly applicable, how should training programs encourage deployment or specialization of these skills in the biomedical field?
• What are the essential elements (e.g. courses, laboratory, clinical, or research rotations in industry, health care organizations, or government labs with big data) of a training program for a cross-trained student?

Environment
• What kind of an environment would be effective for BD2K-supported training?
• What would be a critical mass of students for a viable interdisciplinary program?
• What training program characteristics foster interaction between students trained in different disciplines, so that they learn from one another?

Policy
• Should NIH encourage common core elements in all BD2K-supported training programs?
• Should ALL training programs incorporate some elements of Big Data knowledge and skills into their curriculum?
• What should be the outcome of BD2K training programs and how should they be evaluated?

5:45 Brief Summary of Recommendations Z. Kohane, K. Bandeen-Roche

6:00 Adjourn until Tuesday, 30 July 8:30am

Tuesday, 30 July

8:30 Distillation of Day 1 K. Bandeen-Roche, Z. Kohane
9:00 Characteristics of BD2K Programs for Short-term Training Participants
• Who should the target audience be—undergraduates, faculty at undergraduate institutions, graduate students, postdoctoral fellows, new and experienced investigators, clinicians?
• What can short-term training accomplish? What concepts and skills can be conveyed via in this format? How would the success of such a program be evaluated? What are the metrics of success?

10:00 Characteristics of BD2K Programs for Innovative Training Technology Participants
• What innovative uses of technology could help 1) large numbers of students become familiar with basic core knowledge, or 2) established investigators acquire updated skills or an appreciation of new skills?
• How can online material be made interactive and adaptive to personalize delivery based on the learner’s prior knowledge?
• How can NIH promote the development of training technologies specialized to biomedical Big Data?

11:00 Characteristics of BD2K Programs for Curriculum Participants
• Should NIH support curriculum development to encourage integrated, intersecting curricula?

11:30 Break (Working Lunch) -- Refreshments will not be provided

12:00 General Discussion Participants
• Are there particular challenges to keeping content updated? How can sharing be encouraged?
• How should success of the programs be evaluated? How can this activity be used to increase the number of students in research who are from underrepresented groups or less research-intensive institutions?
• What other training modalities should be considered (e.g. working groups, internships, etc.)?
• Of all the activities discussed, how would you prioritize them?
• Additional advice?

1:00 Summary of Workshop K. Bandeen-Roche, Z. Kohane

1:30 Adjourn

Appendix III: BD2K Workshop on Enhancing Training Invited Guests

My Note: This would be a good start for those that should have Data Publications!

Kristine Alpi, MLS, MPH, AHIP
Director, William Rand Kenan, Jr. Library of Veterinary Medicine,
North Carolina State University
2 Broughton Drive
Raleigh, NC 27695
kmalpi@ncsu.edu
919-513-6219

Karen Bandeen-Roche, PhD
Professor and Chair, Bloomberg School of Public Health,
Johns Hopkins University
615 North Wolfe Street
Baltimore, MD 21205
kbandeen@jhsph.edu
410-955-3067

Mike Boehnke, PhD
Richard G. Cornell Distinguished University Professor of Biostatistics
University of Michigan
1415 Washington Heights
Ann Arbor, Michigan 48109-2029
boehnke@umich.edu
734-936-1001

Alex Bui, PhD
Professor of Radiology and Engineering
University of California, Los Angeles
924 Westwood Boulevard Suite 420
Los Angeles, CA 90095
buia@mii.ucla.edu
310-794-3540

Brian Caffo, PhD
Professor of Biostatistics
Johns Hopkins University
615 North Wolfe Street E3610
Baltimore, MD 21205
bcaffo@jhsph.edu
410-955-3504

Carlos Castillo-Chavez, PhD
Director, Mathematical, Computational and Modeling Sciences Center
Arizona State University
PO Box 871904 Tempe, AZ 85287
ccchavez@asu.edu
480-965-2115

Elissa Chessler, PhD
Associate Professor
Jackson Laboratory
600 Main Street
Bar Harbor, ME 04609
Elissa.chessler@jax.org
207-288-6453
Mark Cohen, PhD  
Professor-in-Residence  
University of California, Los Angeles  
760 Westwood Plaza 17-369  
Los Angeles, CA 90095  
mscohen@ucla.edu  
310-980-7453

Josh Denny, MD, MS  
Associate Professor of Bioinformatics and Medicine  
Vanderbilt University  
2209 Garland Avenue 448  
Nashville, TN 37212  
Josh.denny@vanderbilt.edu  
615-936-1556

Patricia Dombrowski, MA  
Director, Life Science Informatics  
Bellevue College  
3000 Landerholm Circle  
Bellevue, WA 98007  
Patricia.dombrowski@bellevuecollege.edu  
425-564-3164

Ary Goldberger, MD  
Director, Margret & H.A. Rey Institute for Nonlinear Dynamics in Medicine  
Harvard University/Beth Israel Deaconess Medical Center  
330 Brookline Avenue Gz-435  
Boston, MA 02215  
agoldber@caregroup.harvard.edu  
617-667-4267

Betz Halloran, DSc, MD  
Professor of Biostatistics  
University of Washington  
1100 Fairview Avenue North  
PO Box 19024  
Seattle, WA 98109  
betz@u.washington.edu  
206-667-2722

Frank Harrell, PhD  
Chair, Department of Biostatistics  
Vanderbilt University
Larry Hunter, PhD
Director of the Center for Computational Biology and of the Computational Bioscience Program, University of Colorado
12801 East 17th Avenue
MS 8303, RC1-North
Aurora, CO 80045
Larry.hunter@ucdenver.edu
303-724-3574

Robert Kass, PhD
Professor of Statistics,
Carnegie Mellon University
4400 Fifth Avenue Suite 115
Pittsburgh, PA 15213
kass@stat.cmu.edu
412-268-8723

Ron Kikinis, MD
Professor
Brigham and Women's Hospital
1249 Boylston Street 352
Boston, MA 02215
kikinis@bwh.harvard.edu
617-732-7389

Isaac Kohane, MD, PhD
Professor of Pediatrics, Chair of the Informatics Program
Boston’s Children’s Hospital, Dana-Farber/Harvard Cancer Center
300 Longwood Avenue
Boston, MA 02115
Isaac_kohane@harvard.edu
617-919-2184

Andrew Laine, DSc
Professor and Department Chair, Biomedical Engineering
Columbia University
351 Engineering Terrace
1210 Amsterdam Avenue
Mail Code 8904
New York, NY 10027
Elaine Larson, PhD
Professor of Epidemiology, Associate Dean of Research, School of Nursing
Columbia University
Georgian Building
617 West 168th Street 246
New York, NY 10032
Ell23@columbia.edu
212-305-0722

Gary Marchionini, PhD
Professor, School of Information and Library Science
University of North Carolina
Manning Hall 103
Chapel Hill, NC 27599
gary@ils.unc.edu
919-962-8363

Dan Masys, MD
Affiliate Professor, Biomedical and Health Informatics
University of Washington
850 Republican Street, Building C
Seattle, WA 98109-4714
dmasys@uw.edu
360-797-3260

Mark Musen, MD, PhD
Professor, Co-Director, Biomedical Informatics Training Program
Stanford Center for Biomedical Informatics Research
Medical School Office Building
1265 Welch Road
Stanford, CA 94305
musen@stanford.edu
650-725-3390

Mike Newton, PhD
Professor of Statistics, Biostatistics and Medical Informatics
University of Wisconsin
1245A Medical Sciences Center
1300 University Avenue
Madison, WI 53792
Lucila Ohno-Machado, PhD
Professor, Founding Chief, Division of Biomedical Informatics, Associate Dean for Informatics and Technology,
University of California, San Diego
9500 Gilman Drive MC 0505
La Jolla, CA 92093
machado@ucsd.edu
858-822-4931

Sastry Pantula, PhD
Division Director, National Science Foundation
4201 Wilson Boulevard 1025 N
Arlington, VA 22230
spantula@nsf.gov
703-292-9032

Giovanni Parmigiani, PhD
Chair and Professor, Biostatistics and Computational Biology, Dana-Farber/Harvard Cancer Center
450 Brookline Avenue
Boston, MA 02215
gp@jimmy.harvard.edu
617-632-3012

Steve Salzberg, PhD
Professor, Director, Departments of Medicine, Biostatistics, and Computer Science, Center for Computational Biology,
McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University
733 North Broadway
Miller Research Building 459
Baltimore, MD 21205
salzberg@jhu.edu
410-614-6112

Latanya Sweeney, PhD
Professor of Government and Technology in Residence
Harvard University
CGIS 1737 Cambridge Street Knafel 310
Cambridge, MA 02138
latanya@gov.harvard.edu
617-496-3629

Peter Szolovits, PhD
Professor, Computer Science and Engineering
Appendix III: BD2K Workshop on Enhancing Training NIH Participants

My Note: This would be a good start for those that should have Data Publications!

Richard Baird, PhD
Division Director, National Institute of Biomedical Imaging and Bioengineering
bairdri@mail.nih.gov

Vivien Bonazzi, PhD
Program Director, National Human Genome Research Institute
bonazziv@mail.nih.gov

Quan Chen, PhD
Health Scientist Administrator,
National Institute of Allergy and Infectious Diseases
Chenqn2@naiad.nih.gov

Sandra Colombini-Hatch, MD
Medical Officer, National Heart, Lung, and Blood Institute
hatchs@nhlbi.nih.gov

Jennifer Couch, PhD
Branch Chief, National Cancer Institute
Couchj@mail.nih.gov

Leslie Derr, PhD
Health Scientist Administrator,
Office of the Director
derrl@mail.nih.gov
Nancy Desmond, PhD
Office Director and Associate Director,
National Institute of Mental Health
ndesmond@mail.nih.gov

Michelle Dunn, PhD
Program Director, National Cancer Institute
Dunnm3@mail.nih.gov

Valerie Florance, PhD
Division Director, National Library of Medicine
florancev@mail.nih.gov

Nick Gaiano, PhD
Scientific Review Officer, Center for Scientific Review
gaianonr@mail.nih.gov

Jose Galvez, MD
Program Director, National Cancer Institute
galvezij@mail.nih.gov

Maria Giovanni, PhD
Assistant Director for Microbial Genomics, National Institute of Allergy and Infectious Diseases
Mg37u@nih.gov

Bettie Graham, PhD
Division Director, National Human Genome Research Institute
graham@odder.nhgri.nih.gov

Eric Green, MD, PhD
Director, National Human Genome Research Institute
Acting Associate Director for Data Science
green@nhgri.nih.gov

Susan Gregurick, PhD
Division Director, National Institute of General Medical Sciences
susan.gregurick@nih.gov

Mark Guyer, PhD
Deputy Director, National Human Genome Research Institute
guyerm@exchange.nih.gov

Lynda Hardy, PhD, RN
Program Director, National Institute of Nursing Research
hardylr@mail.nih.gov

http://semanticommunity.info/Data_Science/Data_Culture_at_the_NIH
Updated: Fri, 18 Sep 2015 23:53:15 GMT
Powered by mindtouch
Ming Lei, PhD  
Branch Chief, National Cancer Institute  
leim@mail.nih.gov

Peter Lyster, PhD  
Program Director, National Institute of General Medical Sciences  
lysterp@nigms.nih.gov

Ronald Margolis, PhD  
Senior Advisor, National Institute of Diabetes and Digestive and Kidney Diseases  
margolisr@mail.nih.gov

Veerasamy Ravichandran, PhD  
Program Director, National Institute of General Medical Sciences  
ravichanr@nigms.nih.gov

Sally Rockey, PhD  
Deputy Director for Extramural Research, National Institutes of Health  
rockeysa@od.nih.gov

Erica Rosemond, PhD  
Program Officer, National Institute of Mental Health  
rosemonde@mail.nih.gov

Cathrine Sasek, PhD  
Science Education Coordinator, National Institute on Drug Abuse  
csasek@nih.gov

Carol Shreffler, PhD  
Program Administrator, National Institute of Environmental Health Sciences  
Shreffl1@niehs.nih.gov

Heidi Sofia, MPH, PhD  
Program Director, National Human Genome Research Institute  
sofiahj@mail.nih.gov

Erica Spotts, PhD  
Health Scientist Administrator,  
Office of the Director  
spottse@mail.nih.gov

Jennifer Sutton, MS  
Extramural Program Policy and Evaluation Officer, Office of the Director  
suttonj@mail.nih.gov

http://semanticommunity.info/Data_Science/Data_Culture_at_the_NIH  
Updated: Fri, 18 Sep 2015 23:53:15 GMT
Appendix IV NIH BD2K Training Working Group Members

My Note: This would be a good start for those that should have Data Publications!

Richard Baird (NIBIB)
Vivien Bonazzi (NHGRI)
Quan Chen (NIAID)
Sandra Colombini-Hatch (NHLBI)
Leslie Derr (OD)
Michelle Dunn (NCI)
Valerie Florance (NLM)
Nick Gaiano (CSR)
Jose Galvez (NCI)
Bettie Graham (NHGRI)
Mark Guyer (NHGRI)
Linda Hardy (NINR)
Ming Lei (NCI)
Veerasamy Ravichandran (NIGMS)
Erica Rosemond (NIMH)
Catherine Sasek (NIDA)
Carol Shreffler (NIEHS)
Heidi Sofia (NHGRI)
Scott Somers (NIGMS)
Erica Spotts (OD)
Jennifer Sutton (OD)
Appendix V ACD Working Groups

ACD Working Groups on Biomedical Workforce: [http://acd.od.nih.gov/bwf.htm](http://acd.od.nih.gov/bwf.htm)

and ACD Working Group on Diversity in the Biomedical Research Workforce: [http://acd.od.nih.gov/dbr.htm](http://acd.od.nih.gov/dbr.htm)

My Note: There are PDF files at each of these sites that could be re-purposed as subtopics here

Dr. Sally Rockey, NIH Deputy Director for Extramural Research Director, presented an update on the NIH’s responses to the reports from these two ACD working groups. The charge to the Biomedical Workforce Working Group (BMW WG) was to (1) develop a model for a sustainable and diverse U.S. biomedical research workforce that can inform decisions about training of the optimal number of people for the appropriate types of positions that will advance science and promote health and (2) recommend actions that NIH should take to support a future sustainable biomedical infrastructure. Dr. Rockey presented data which showed that (1) the number of PhDs in biomedical sciences is increasing while the number of PhDs in chemistry has remained about the same; (2) most doctoral students are supported as research assistants on research grants; (3) the age when biomedical doctoral students get their first non-postdoc or tenure track job is around 36 years compared to about 33 years for those with doctoral degrees in chemistry; (4) the average age of PIs awarded their first R01 or equivalent is 40 years; (5) early in their careers, biomedical scientists earn less than those with degrees in math, physical and social sciences, and engineering, which results in a significant loss in lifetime earnings; and (6) only 2% of the NIH-trained workforce are unemployed, 43% in academic research, and 55% are employed in other science-related activities. The BMW WG recommended (1) for graduate students--shorten and diversify the training and increase financial support; (2) for postdoctoral fellows--increase financial support and training for more than academic careers; (3) for physician scientists--conduct a focused follow-up study; and (4) for staff scientists--encourage study sections to consider them valuable members of the research team. The BMW WG also recommended that NIH gradually reduce the percentage of funds from NIH grants used for salary support and institute a more vigorous evaluation of programs and encourage stronger coordination amongst programs.

NIH has put several efforts in place to respond to the recommendations:

- The eligibility period for postdoctoral students to apply for the K99/R00 will be shortened from five years to four years, effective February 2014
- NIH is in the process of reviewing applications to institutions that responded to a RFA that called for new approaches to broadening the training experiences of pre-and postdoctoral students to reflect the range of career options of trainees ([http://grants.nih.gov/grants/guide/r...RM-12-022.html](http://grants.nih.gov/grants/guide/r...RM-12-022.html))
- All NIH-supported trainees will be required to have an Individual Development Plan (IDP) in place by October 1, 2014 ([http://grants.nih.gov/grants/guide/n...OD-13-093.html](http://grants.nih.gov/grants/guide/n...OD-13-093.html))
- Postdoctoral stipends will be increased in FY2014.

NIH will also be encouraging institutions to reduce the length of graduate training; mandating that all NIH Institutes and Centers support F30 and F31 fellowships by April 2014; developing a comprehensive survey on benefit policies and NIH support of faculty salaries; developing a comprehensive tracking system for all trainees; and creating a unit at NIH to assess the biomedical workforce.

Dr. Rockey also discussed the recommendations of the Working Group on Diversity in the Biomedical Research Workforce (WGDBRW). She noted that the NIH has been committed to increasing the diversity of the biomedical
workforce, and for over 30 years it has supported programs to achieve this goal through institutional and individual programs. However, a paper in Science in August 2011 highlighted concerns regarding race, ethnicity, and the awarding of research grants. Even when controlled for institutions, African-American scientists had a lower award rate.

On the basis of the recommendations of the WGDBRW, NIH has now developed a comprehensive strategy to redress the problem, which includes the following:

- Establishing a new leadership position, Chief Officer for Scientific Workforce Diversity (the recruitment is underway for a permanent leader).
- Making the effort to increase the pipeline through a new initiative, Building Infrastructure Leading to Diversity (BUILD).
- Developing a National Research Mentoring Network (NRMN).
- Making new efforts to ensure fairness in peer review.

3 Science (v 333), 19 August 2011; pp 1015-1019.