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/ontology-nsf042814

Background

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

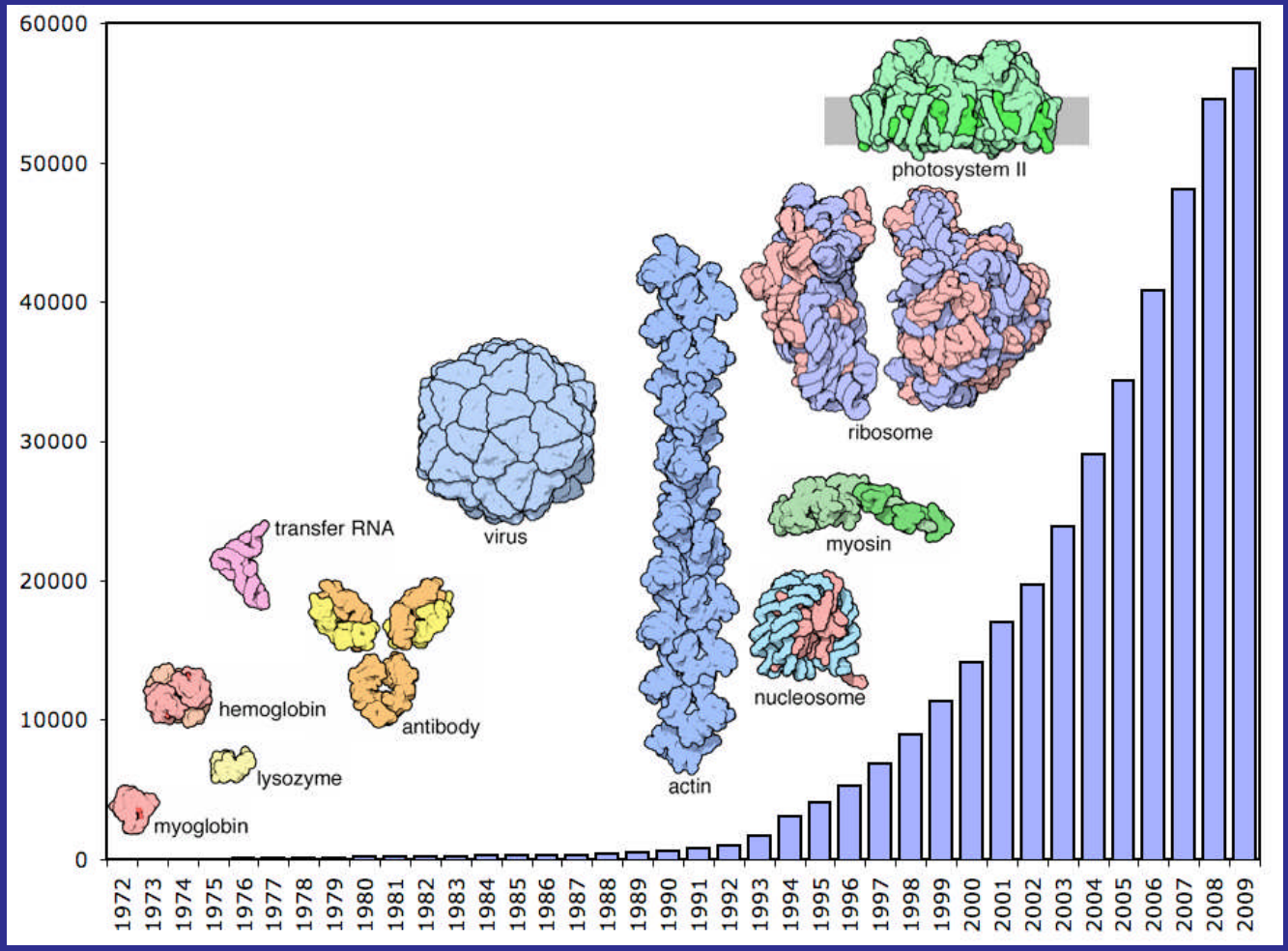
- With Ontologies:
  - Extensive work with the Gene Ontology
  - Co-developed mmCIF for macromolecular structure

Disclaimer: I only started March 3, 2014

…but I had been thinking about this prior to my appointment

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

Motivation for Change:
PDB Growth in Numbers and Complexity



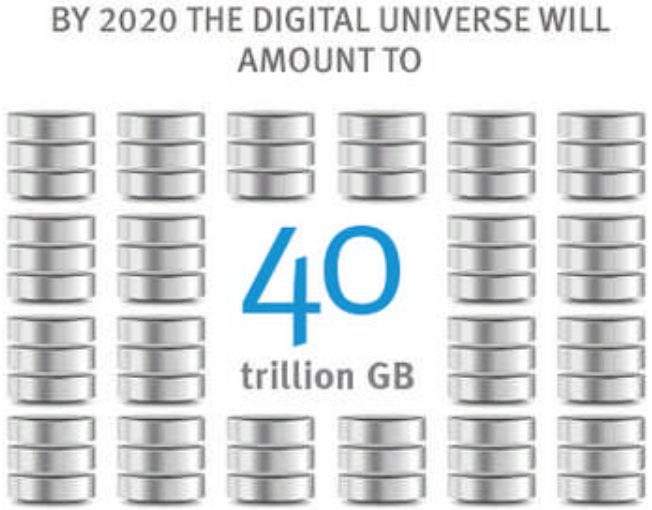
[From the RCSB Protein Data Bank]

Motivation for Change: We Are at the Beginning

DATA KEEPS GROWING

The volume of digital data worldwide is growing rapidly, as the annual IDC Digital Universe study reveals. From 2005 to 2020, the digital universe will grow by a factor of 300, from 130 exabytes to 40,000 exabytes, or 40 trillion gigabytes (more than 5,200 gigabytes for every man, woman, and child in 2020). From now until 2020, the digital universe will about double every two years.

The majority of information in the digital universe, 68% in 2012, is created and consumed by consumers watching digital TV, interacting with social media, sending camera phone images and videos between devices and around the Internet, and so on.

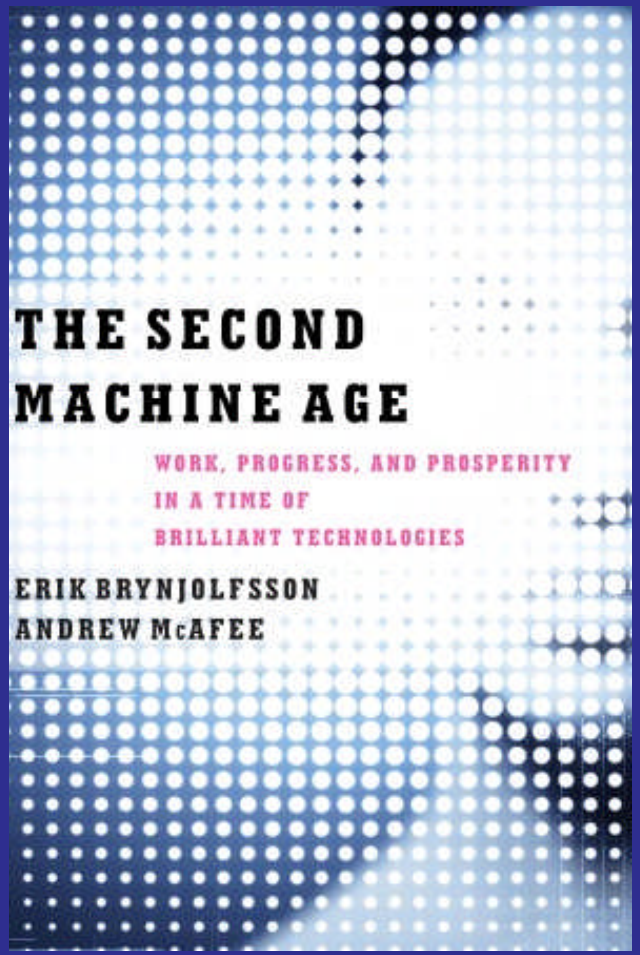


INTERACTIVE: IDC Report: The Digital Universe in 2020.

Explore the IDC report, watch videos, and more.

NIH

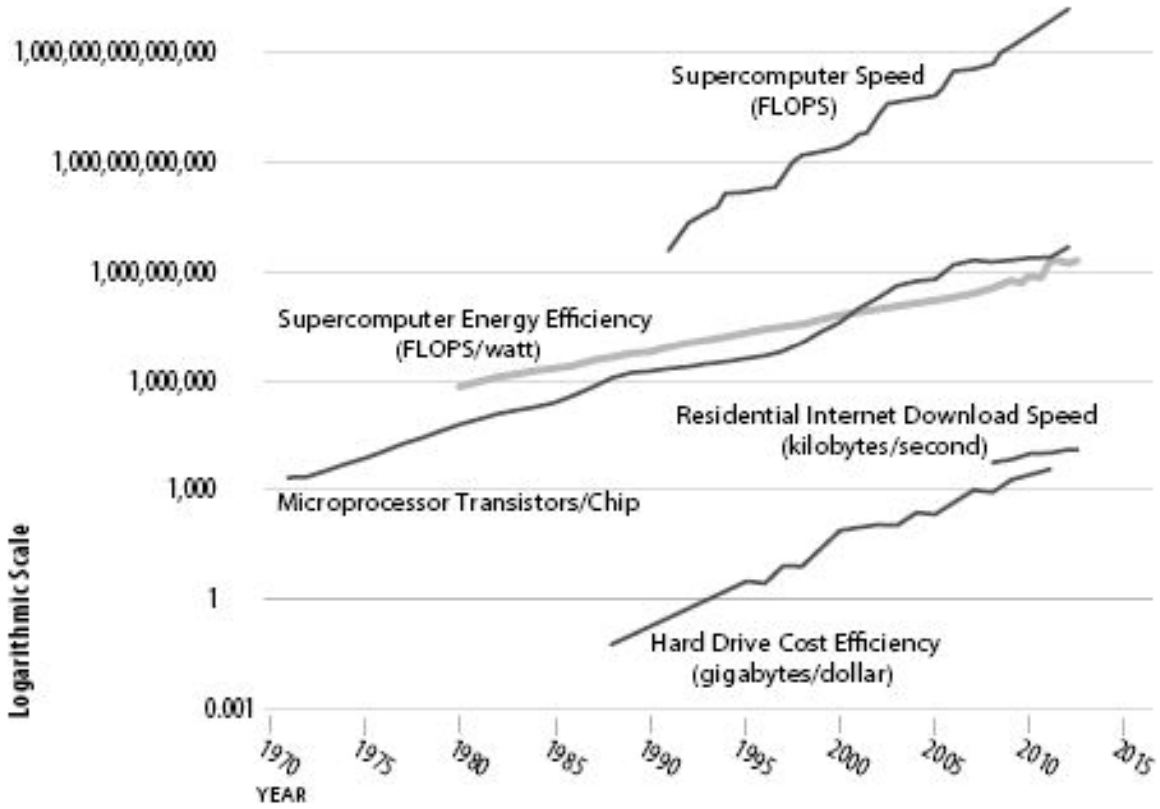
Motivation:
We Are at an Inflexion Point for Change



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

From the Second Machine Age

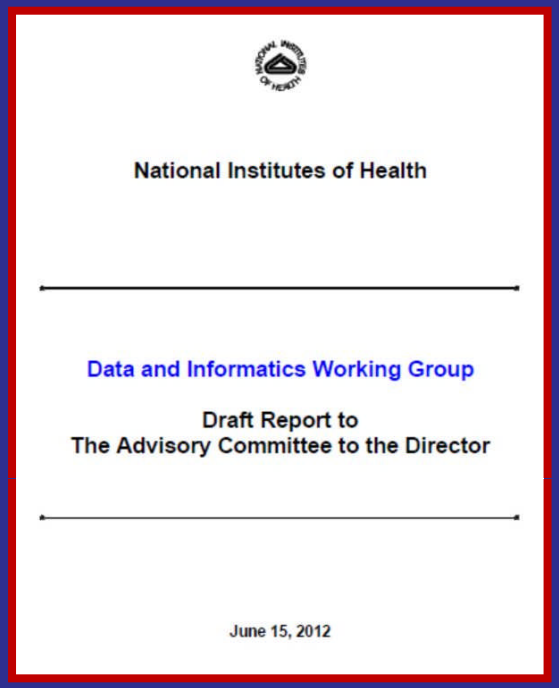
FIGURE 3.3 The Many Dimensions of Moore's Law



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

Much Useful Groundwork Has Been Done

NIH Data & Informatics Working Group



National Institutes of Health

Data and Informatics Working Group

Draft Report to
The Advisory Committee to the Director

June 15, 2012

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

Big Data to Knowledge (BD2K)

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

http://bd2k.nih.gov

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

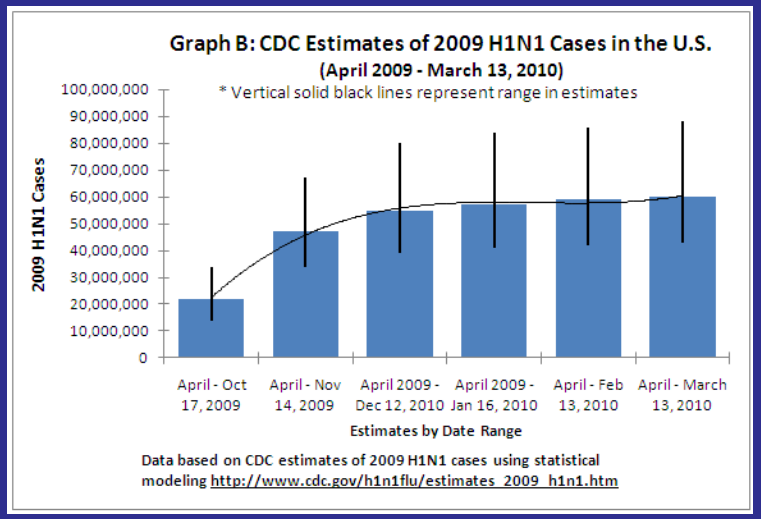
This is just the beginning…

Some Early Observations

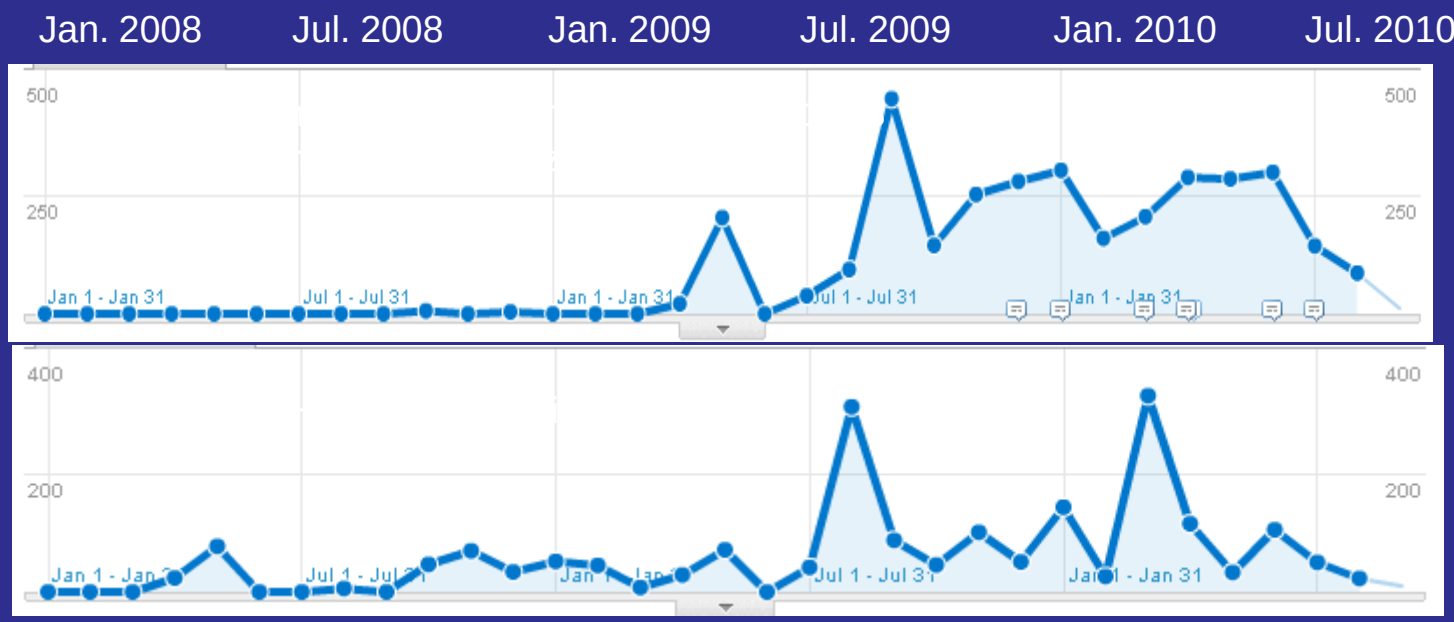
Some Early Observations

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

Consider What Might Be Possible



* http://www.cdc.gov/h1n1flu/estimates/April_March_13.htm



Structure Summary page activity for H1N1 Influenza related structures

[Andreas Prlic]

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

Google

amazon.com

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

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

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/uploads/Data-Commons-3-1024x825.png

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)

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 makes 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

4. Training in biomedical data science is spotty

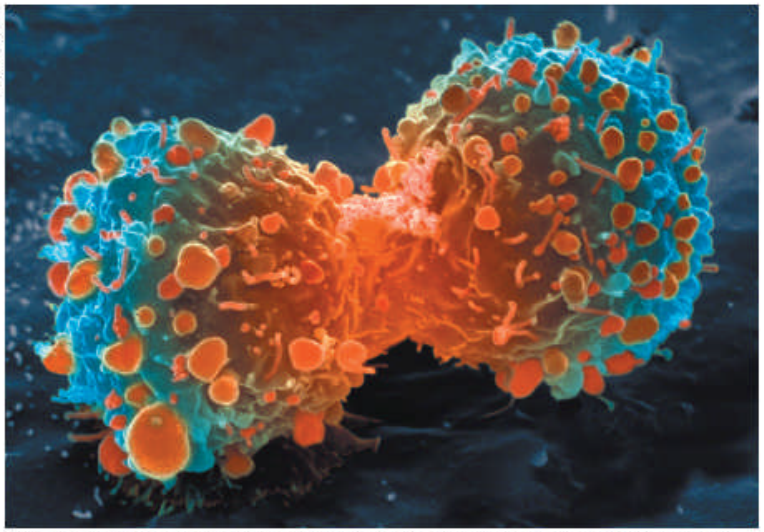
- Problem
  - Coverage of the domain is unclear
  - There may well be redundancies
- Solution
  - Cold Spring Harbor like training facility(s)
    - Training coordinator
    - Rolling hands on courses in key areas
    - Appropriate materials on-line
  - Interagency training initiatives

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

COMMENT

AVIAN INFLUENZA Shift expertise to track mutations where they emerge p.534 | EARTH SYSTEMS Past climates give valuable clues to future warming p.537 | HISTORY OF SCIENCE Descartes' lost letter tracked using Google p.540 | OBITUARY Wylie Vale and an elusive stress hormone p.542



Many landmark findings in preclinical oncology research are not reproducible, in part because of inadequate cell lines and animal models.

Raise standards for preclinical cancer research

C. Glenn Begley and Lee M. Ellis propose how methods, publications and incentives must change if patients are to benefit.

Efforts over the past decade to characterize the genetic alterations in human cancers have led to a better understanding of molecular drivers of this complex set of diseases. Although we in the

trials in oncology have the highest failure rate compared with other therapeutic areas. Given the high unmet need in oncology, it is understandable that barriers to clinical development may be lower than for other

investigators must reassess their approach translating discovery research into gre clinical success and impact.

Many factors are responsible for the h failure rate, notwithstanding the inh

47/53 “landmark” publications could not be replicated

[Begley, Ellis Nature, 483, 2012]

Must try harder

Too many sloppy mistakes are creeping into scientific papers. Lab heads at the data — and at themselves.

Error prone

Biologists must realize the pitfalls of wor massive amounts of data.

If a job is worth doing, it is worth doing twice

Researchers and funding agencies need to put a premium on ensuring that results are reproducible, argues **Jonathan F. Russell**.

The case for open computer programs

Six red flags for suspect work

C. Glenn Begley explains how to recognize the preclinical papers in which the data won’t stand up.

Know when your numbers are significant

[Carole Goble]

I can’t reproduce research from my own laboratory?

Daniel Garijo et al. 2013 Quantifying Reproducibility in Computational Biology: The Case of the Tuberculosis Drugome *PLOS ONE* 8(11) e80278 .

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
- Original:
  - http://funsite.sdsc.edu/drugome/TB/
- Current:
  - Recast in the Wings workflow system

Daniel Garijo et al. 2013 Quantifying Reproducibility in Computational Biology: The Case of the Tuberculosis Drugome *PLOS ONE* 8(11) e80278 .

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

A Conceptual Overview of the Method Should Be Mandatory

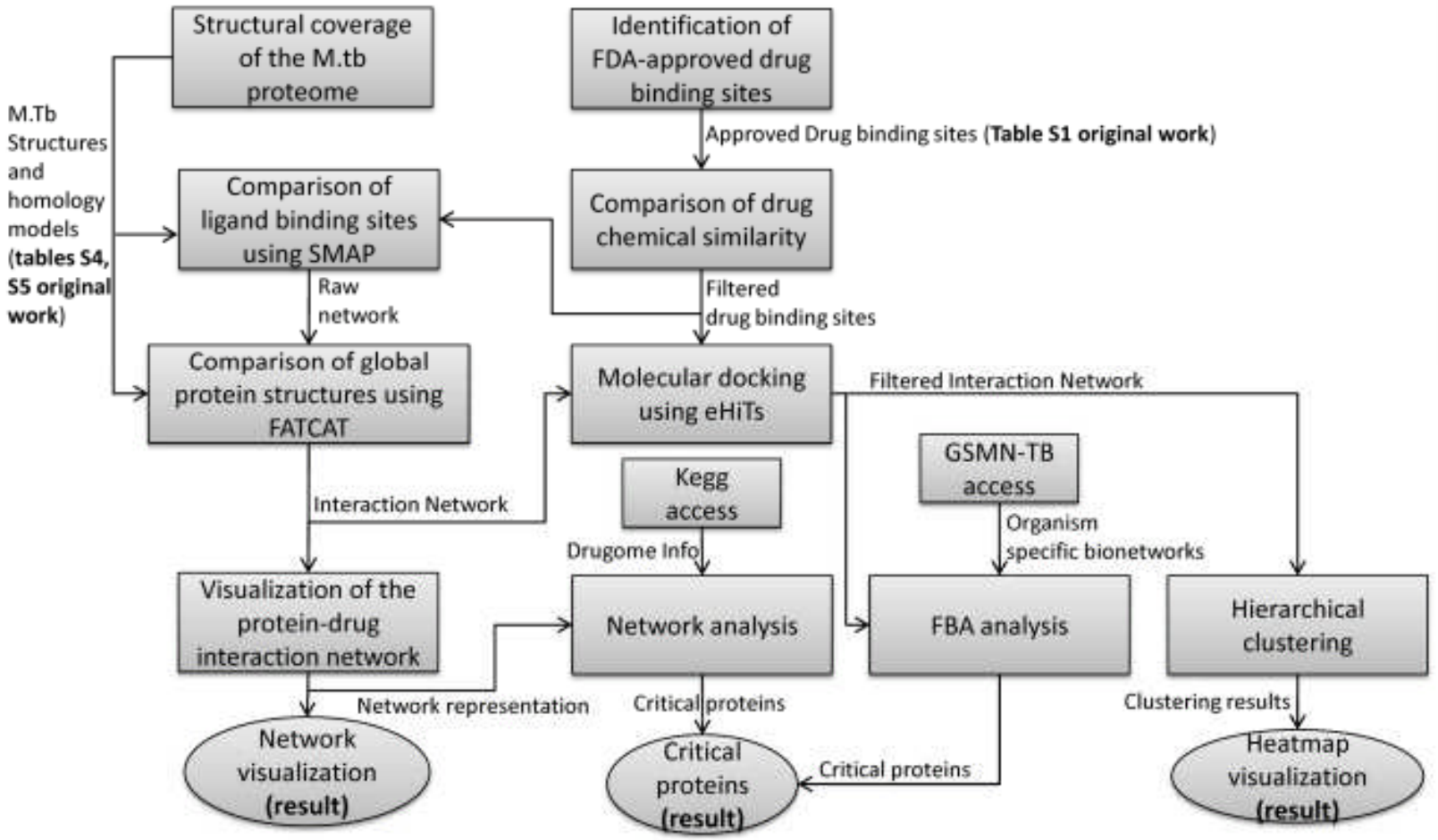


Figure 1. A high-level dataflow diagram of the TB drugome method.
doi:10.1371/journal.pone.0080278.g001

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

Time to Reproduce the Method

Table 1. Time to reproduce the method.

Tasks	Time (hours)
Familiarization with workflow and running software	160
SMAP steps	32
SMAP result sorter steps	8
Merger steps	4
Get significant results	4
FATCAT URL checker	8
FATCAT step	4
Remove significant pairs	4
Create dip files	8
Create ideal ligands	8
Ideal ligand checker	8
Autodock Vina	16
Data visualization steps	16
TOTAL	280 hours

doi:10.1371/journal.pone.0080278.t001

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

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?

I cast the solutions in a vision …

something I call the digital enterprise

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

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*

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

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.

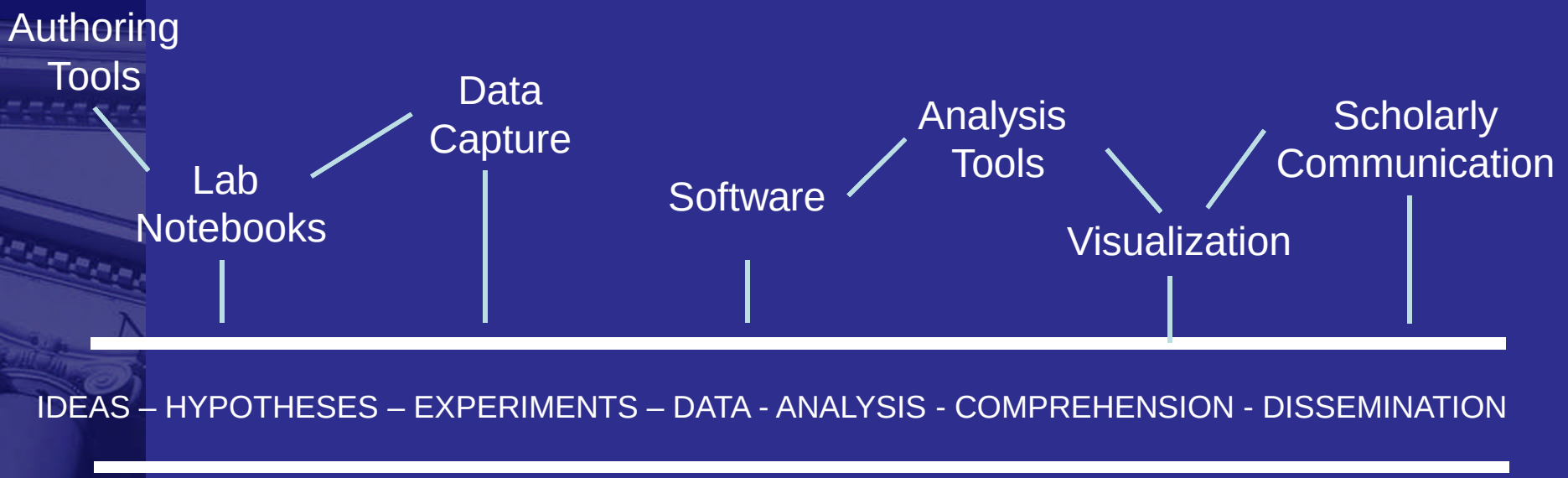
To get to that end point we have to consider the complete research lifecycle

The Research Life Cycle will Persist

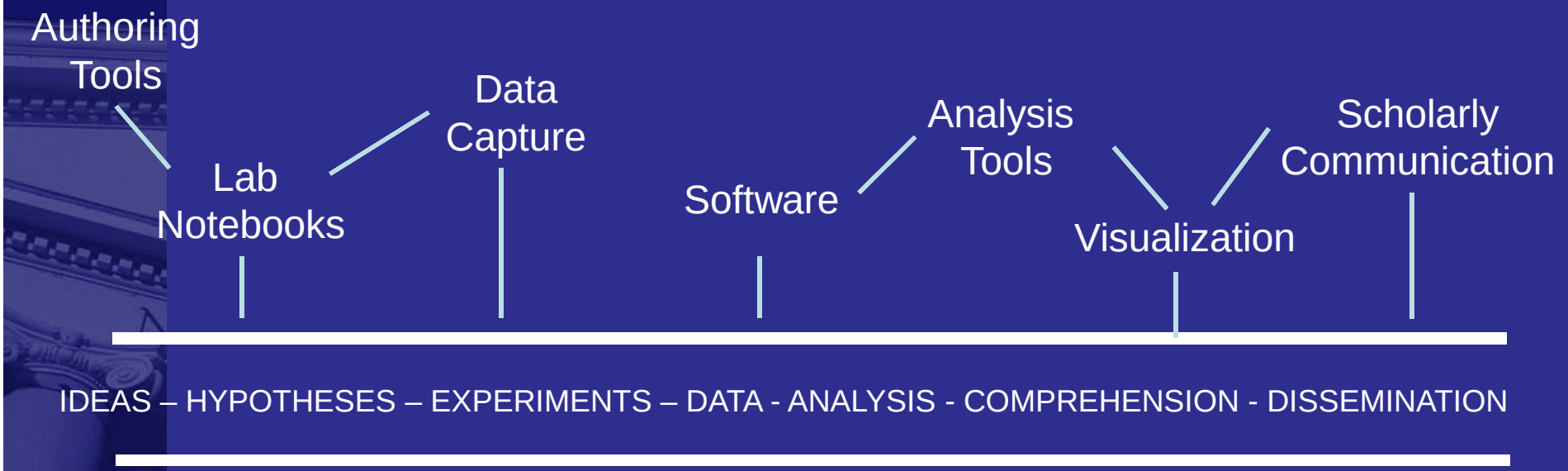
IDEAS – HYPOTHESES – EXPERIMENTS – DATA - ANALYSIS - COMPREHENSION - DISSEMINATION

Tools and Resources Will Continue To Be Developed

Those Elements of the Research Life Cycle will Become More Interconnected Around a Common Framework



New/Extended Support Structures Will Emerge



Discipline-Based Metadata Standards

Git-like Resources By Discipline

Community Portals

Data Journals

Commercial & Public Tools

New Reward Systems

Training

Institutional Repositories

Commercial Repositories

We Have a Ways to Go

Authoring Tools

Lab Notebooks

Data Capture

Software

Analysis Tools

Visualization

Scholarly Communication

IDEAS – HYPOTHESES – EXPERIMENTS – DATA - ANALYSIS - COMPREHENSION - DISSEMINATION

Commercial & Public Tools

Discipline-Based Metadata Standards

Git-like Resources By Discipline

Community Portals

Data Journals

New Reward Systems

Training

Institutional Repositories

Commercial Repositories

Next Steps

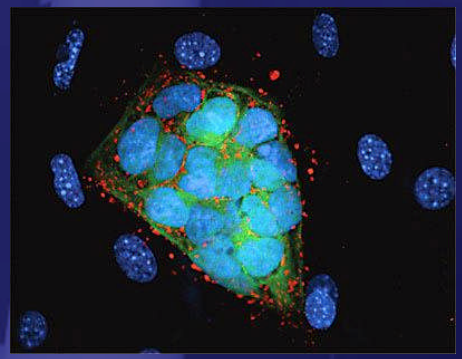
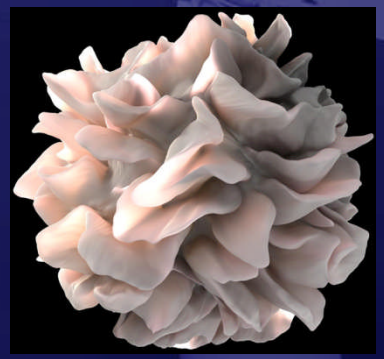
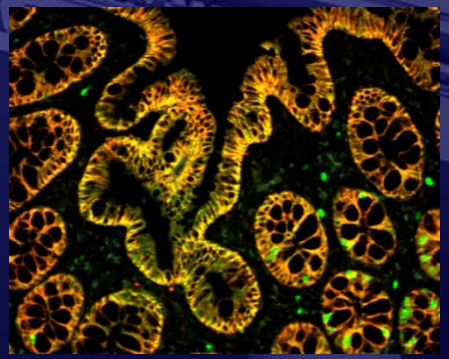
- Support for research objects
  - These objects underpin the various cataloging efforts
- Support for data metrics
  - Such metrics underpin a change in the reward system




NIH...

philip.bourne@nih.gov

Turning Discovery Into Health

philip.bourne@nih.gov

Thank You!
Questions?

Back Pocket Slides for BD2K Programs

1. Facilitating Broad Use

- Summary of Data Catalog Workshop and Request for Information: www.bd2k.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)

1. 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)

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)

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

3: Enhancing Training

- Summary of Training Workshop and Request for Information:
  - 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.

3: 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**.

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*)