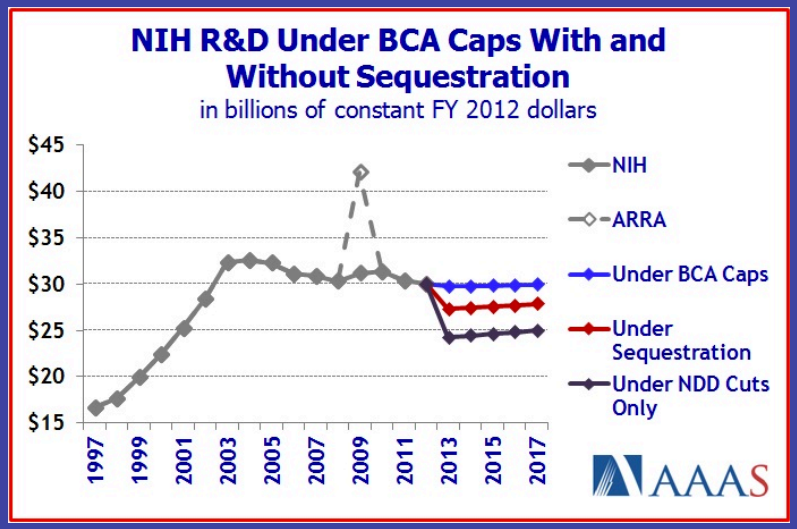
Biomedical Research as an Open Digital Enterprise

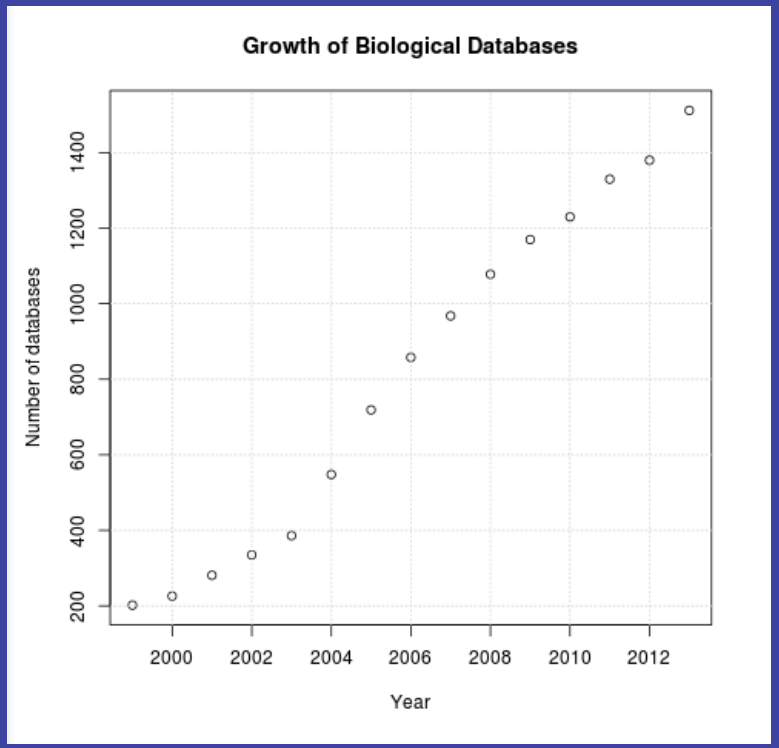
Philip E. Bourne Ph.D.
Associate Director for Data Science
National Institutes of Health
http://www.slideshare.net/pebourne

A View from the Funding Agencies

“It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness, it was the epoch of belief, it was the epoch of incredulity, it was the season of Light, it was the season of Darkness, it was the spring of hope, it was the winter of despair …”

A Tale of Two Numbers





Source Michael Bell http://homepages.cs.ncl.ac.uk/m.j.bell1/blog/?p=830

We (the NIH) Are Working On, But As Yet Do Not Have Good Answers To:

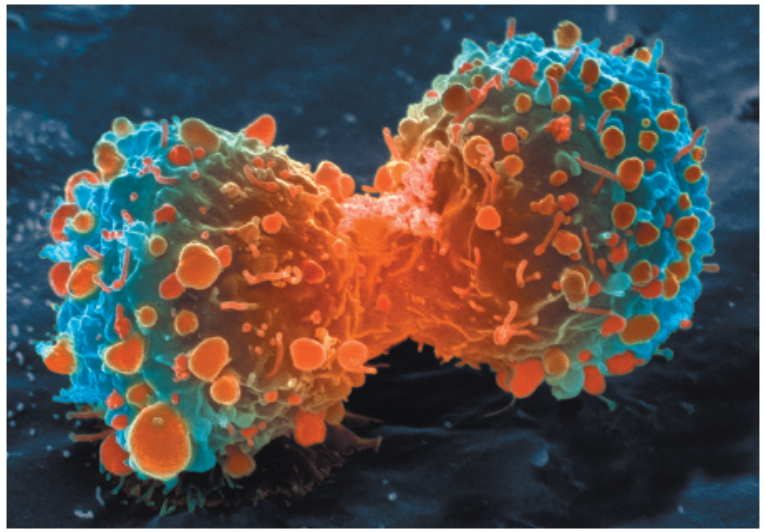
1. Today, how much are we actually spending on data and software related activities?

2. How much should we be spending to achieve the maximum benefit to biomedical science relative to what we spend in other areas?

There are other drivers of change out there besides economics and an increasing emphasis on data and analytics

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]

Reproducibility

- Most of the 27 Institutes and Centers of the NIH are currently reviewing the ability to reproduce research they are funding

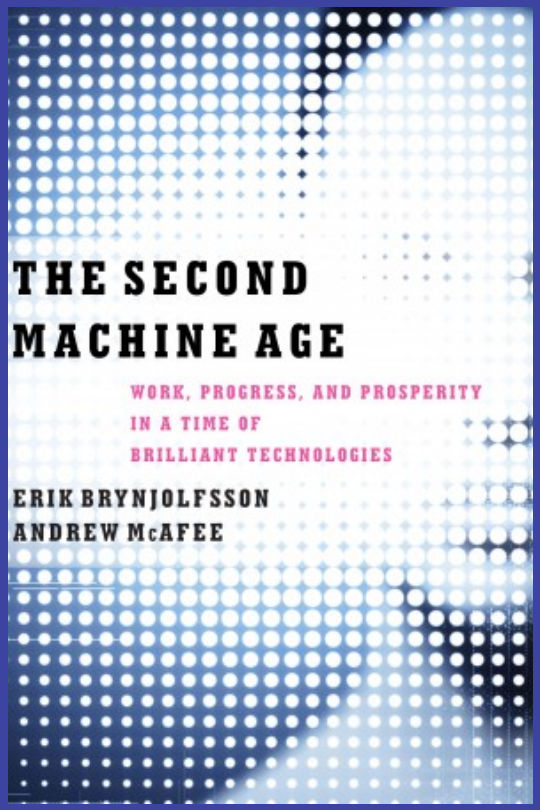
- The NIH recently convened a meeting with publishers to discuss the issue – a set of guiding principles arose

Reproducibility – More is in the Works

- Much of the research life cycle is now digital - encourage the reliability, accessibility, findability, usability of data, methods, narrative, publications etc.
- How?
  - ✓ Data sharing plans
  - ✓ Standards frameworks
  - ✓ Data and software catalogs
  - ✓ PubMedCentral

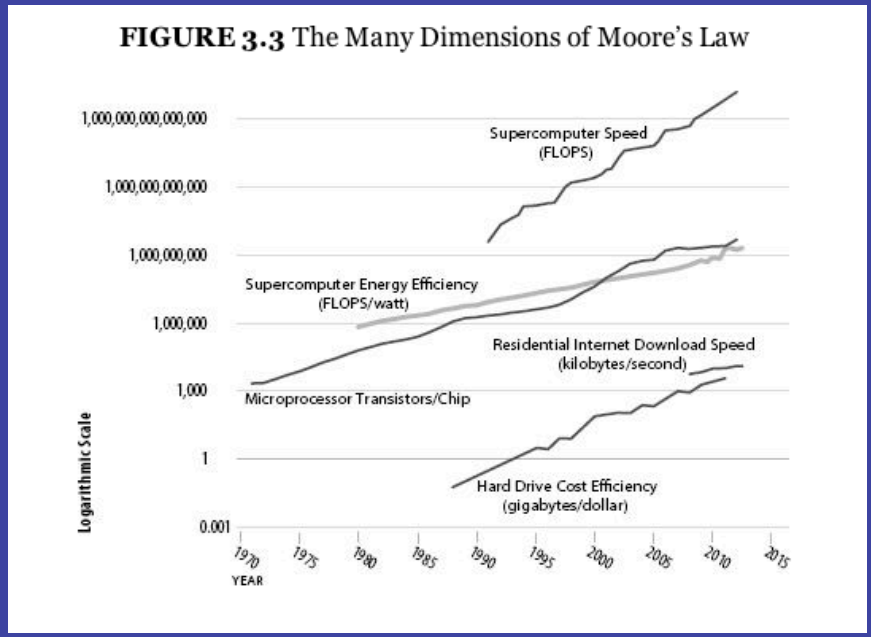
  - ? The Commons – PMC for the complete lifecycle
  - ? Machine readable data sharing plans
  - ? Small funding to communities
  - ? Support for training and best practices in eScholarship

Growth as Another Driver



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

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



To Summarize Thus Far …

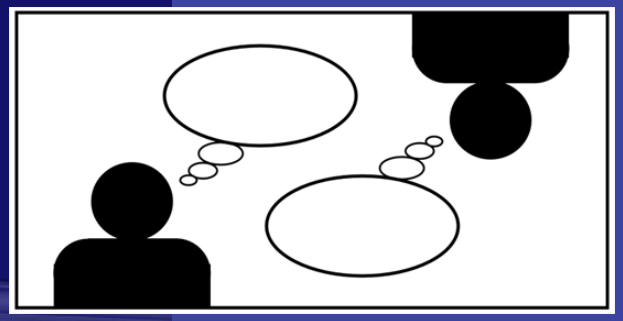
A time of great (unprecedented?) scientific development but limited funding

A time of upheaval in the way we do science

From a funders perspective…

A time to squeeze every cent/penny to maximize the amount of research that can be done

A time when top down approaches meet bottom up approaches



Top Down vs Bottom Up

- Top Down
  - Regulations e.g. US: Common Rule, FISMA, HIPPA
  - Data sharing policies
    - OSTP
    - GWAS
    - Genome data
    - Clinical trials
  - Digital enablement
  - Moves towards reproducibility
- Bottom Up
  - Communities emerge and crowd source
    - Collaboration
    - Data shared
    - Open source software
    - Common principles
    - Standards

And Considering This Audience…

It was the age when software developers are in the greatest demand for science..

It was the age when the rewards outside academia are greater than the rewards inside


The New York Times

Technology

WORLD | U.S. | N.Y. / REGION | BUSINESS | TECHNOLOGY | SCIENCE | HEALTH | SPORTS | OPINION

CAMCORDERS | CAMERAS | CELLPHONES | COMPUTERS | HANDHELDS | HOME VIDEO | MUSIC | PERIPHERALS

Google's Buses Help Its Workers Beat the Rush

Randi Lynn Beach for The New York Times

Google offers shuttle bus service to and from its main office in Mountain View, Calif. About 1,200 employees use the service.

By MIGUEL HELFT

Published: March 10, 2007

TWITTER

Optimistically This is a Time of Opportunity

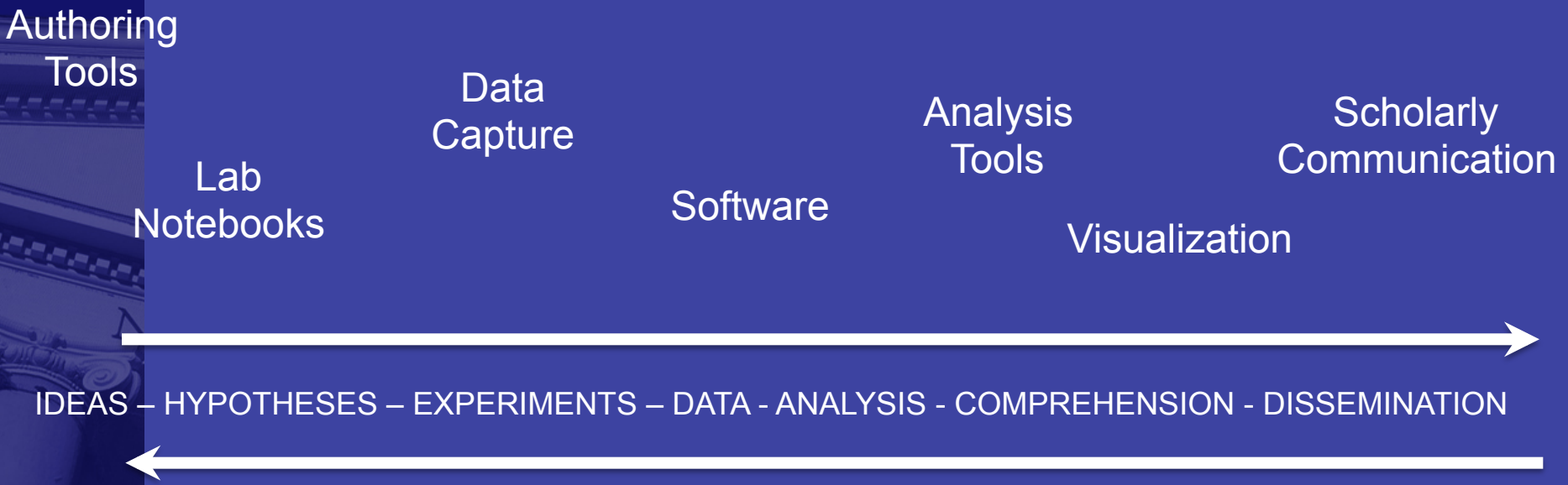
- The time for software developers is here
- The time to derive new business models is here
- The time to foster best software practices is here
- ….

Okay so what are we doing about it?

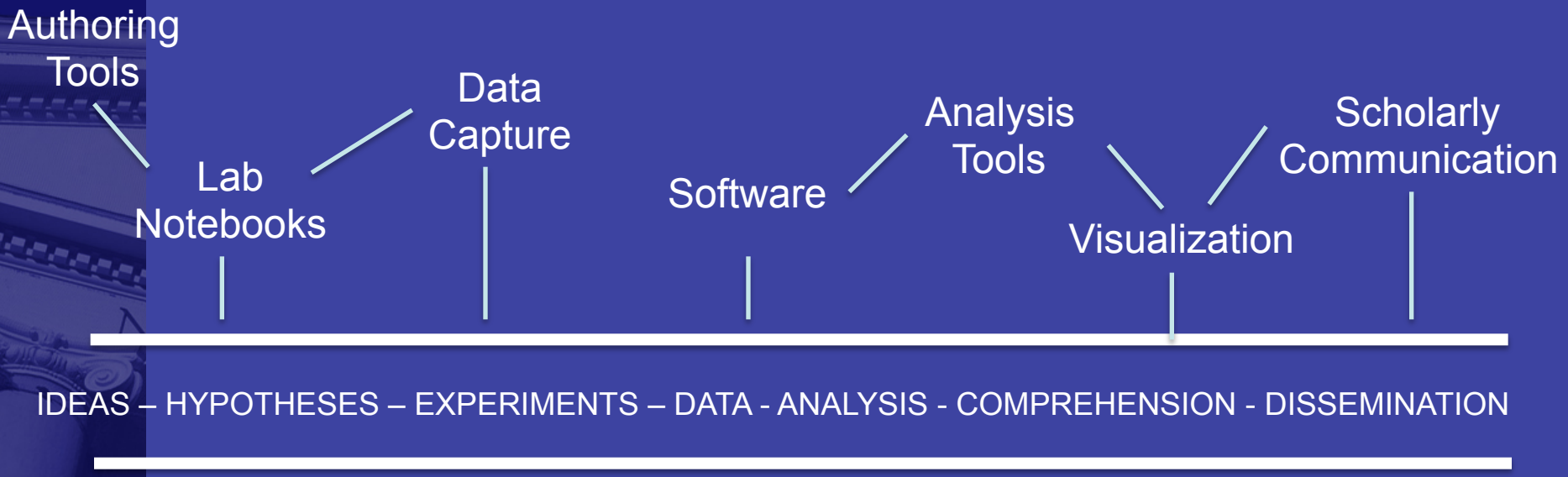
To start with we are thinking about the complete research lifecycle

The Research Life Cycle

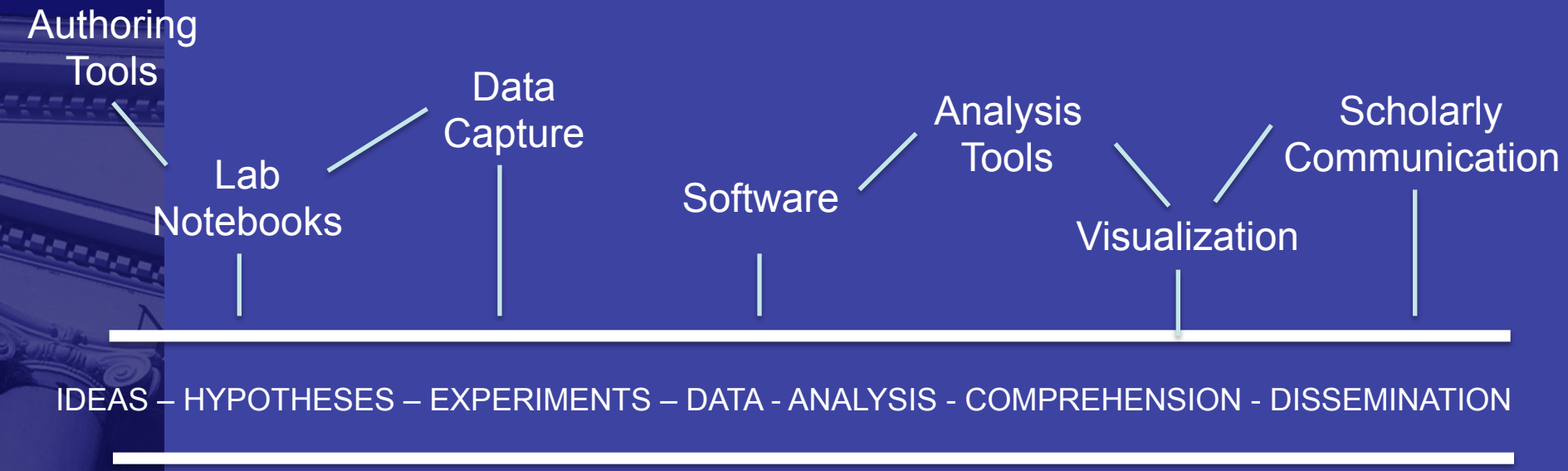
Tools and Resources Will Continue To Be Developed



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



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



Discipline-Based Metadata Standards

Git-like Resources By Discipline

Community Portals

Data Journals

Commercial & Public Tools

New Reward Systems

Training

Institutional Repositories

Commercial Repositories

NIH

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

Authoring Tools

Lab Notebooks

Data Capture

Software

Analysis Tools

Visualization

Scholarly Communication

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

Data Journals

Git-like Resources By Discipline

Community Portals

Commercial & Public Tools

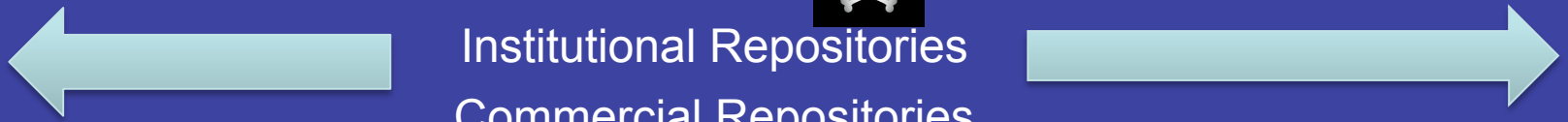
Discipline-Based Metadata Standards

New Reward Systems

Training

Institutional Repositories

Commercial Repositories



What are we proposing as that common framework?

The Commons Is …

- A public/private partnership
- An agile development starting with the evaluation of a few pilots
- An example: porting DbGAP to the cloud
- An experiment with new funding strategies

What The Commons *Is* and *Is Not*

- Is Not:
  - A database
  - Confined to one physical location
  - A new large infrastructure
  - Owned by any one group
- Is:
  - A conceptual framework
  - Analogous to the Internet
  - A collaboratory
  - A few shared rules
    - All research objects have unique identifiers
    - All research objects have limited provenance

Sustainability and Sharing: The Commons

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

Data

The Long Tail

Core Facilities/HS Centers

Clinical /Patient

The Why:
Data Sharing Plans

The How:

NIH Awardees

Government

Software Index

Standards

The Commons

Data Discovery Index

BD2K Centers

Rest of Academia

Private Sector

Cloud, Research Objects, Business Models

The End Game:

Scientific Discovery

Knowledge

Usability

Quality

Security/ Privacy

Metrics/ Standards

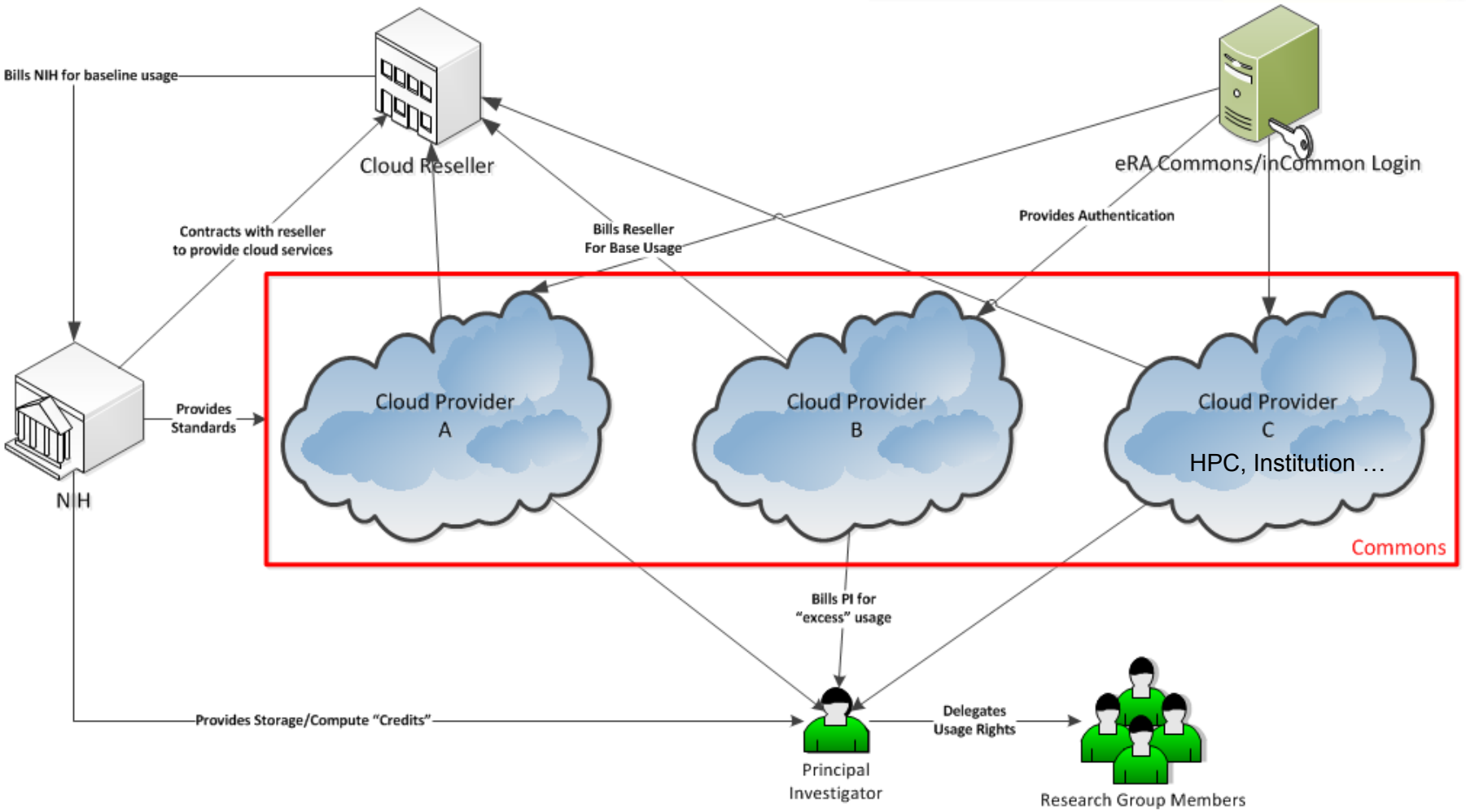
Sustainable Storage

What Does the Commons Enable?

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

http://100plus.com/wp-content/uploads/Data-Commons-3-1024x825.png

One Possible Commons Business Model



[Adapted from George Komatsoulis]

Commons Pilots

- Define a set of use cases emphasizing:
  - Openness of the system
  - Support for basic statistical analysis
  - Embedding of existing applications
  - API support into existing resources
- Evaluate against the use cases
- Review results & business model with NIH leadership
- Design a pilot phase with various groups
- Conduct pilot for 6-12 months
- Evaluate outcomes and determine whether a wider deployment makes sense
- Report to NIH leadership summer 2015

What Will Software Development Look Like in the Commons?

- Software identifiers make software:
  - Easy to find
  - Easy of use
  - Easy to cite
- Which means:
  - Need a standard citation scheme
  - Publishers must be encouraged to use it
  - The software index should facilitate the above AND
    - Provide metrics for use
    - Ability to provide commentary

Minimal Software Specification

- Title
- Version
- License
- Links to source
- Human readable synopsis
- Author names, affiliations
- Ontological terms describing software
- Dependencies
- Acknowledgements
- Publications

Examples of Folks We Want to Engage

- Other funding agencies – national and international
- Open Science Framework https://osf.io/
- Evernote https://evernote.com/
- Simtk https://simtk.org/xml/index.xml
- MyExperiment http://www.myexperiment.org/
- Galaxy http://galaxyproject.org/
- Lab notebook systems
- Other systems used already by NIH

Putting it all together in a coherent strategy….

Associate Director for Data Science

Scientific Data Council

External Advisory Board

Programmatic Theme

Sustainability*	Education*	Innovation*	Process	Collaboration

* Hires made

Deliverable

Commons	Training Center	BD2K	Modified Review	Communication

Example Features

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

- Coordinate
- Hands-on
- Syllabus
- MOOCs

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

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

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

NIH

The Biomedical Research Digital Enterprise

BD2K – Commons Users

- Centers of Excellence in Data Science (Awards 9/14)
- Data Discovery Index Consortium (Award 9/14)
- Training grants awarded (Awards 9/14)

- Software development (Awards 15)

- Standards framework (Awards 15)
- Software index consortium (Award 15)

- Awards next year ~$100M

Mission Statement

To foster an ecosystem that enables biomedical research to be conducted as a digital enterprise that *enhances health, lengthens life and reduces illness and disability*

Some Acknowledgements

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