

Board on Mathematical Sciences & Analytics

LIFE CYCLE DECISIONS FOR BIOMEDICAL DATA

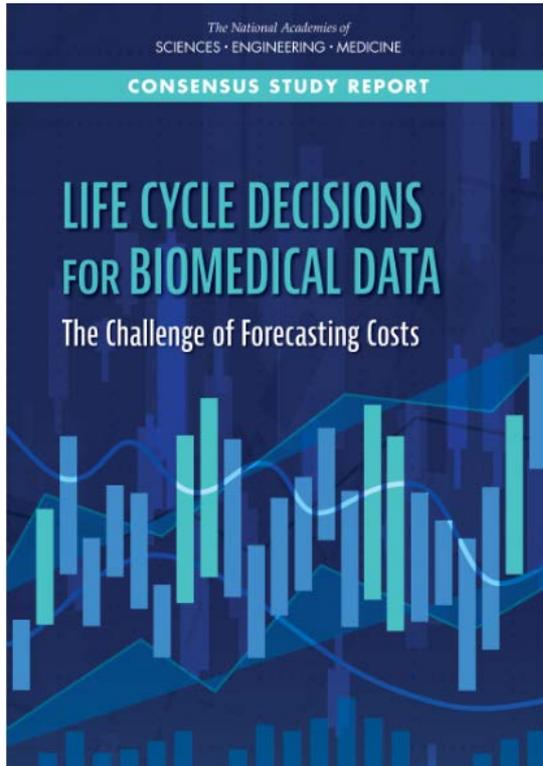
The Challenge of Forecasting Costs





BOARD ON MATHEMATICAL SCIENCES AND ANALYTICS

Forecasting Data Costs for Storage Providers



Life Cycle Decisions for Biomedical Data: The Challenge of Forecasting Costs

Presented to the Public
August 27, 2020

Forecasting Data Costs for Researchers, Funders, and Storage Providers

August 2020 weekly webinar series, 12-1pm ET

Recordings available at <https://vimeo.com/showcase/7444639>

August 13: *Forecasting Data Costs for Researchers*

August 20: *Forecasting Data Costs for Funding Institutions*

August 27: *Forecasting Data Costs for Storage Providers*

This webinar series is sponsored by the
National Library of Medicine of the
National Institutes of Health



U.S. National Library
of Medicine

Forecasting Data Costs *for Storage Providers*



Ilkay Altintas
(Moderator)

University of California San
Diego



Clifford Lynch

Coalition for Networked
Information



Brian Nosek

Center for Open Science



Alex Ropelewski

Pittsburgh Supercomputing
Center

Forecasting Data Costs

for Storage Providers



Clifford Lynch

Executive director, Coalition for Networked Information

Forecasting Data Costs: Highlights for Storage Providers

Summary of my talk

Very brief overview of report & work (see video of earlier committee webinar for a more detailed review)

Closer look at Underlying Framework Foundation of Three Data States
Comments on State Transitions: "Dehydration" and "Hydration"
Comments on Selected Sources of Uncertainty ("Disrupters")
Thoughts on Strategies for State 2 Resources and Terminologies



Context

- Biomedical researchers generate, collect, and store research data in increasing volumes and dimension.
- Sustained data access and preservation generate costs that are difficult to predict and allocate responsibility for.
- The biomedical data landscape is diverse and dynamic, requiring unique and innovative approaches.



Statement of Task

National Library of Medicine of the National Institutes of Health asked for a *framework for forecasting long-term costs* for preserving, archiving, and accessing biomedical data.



Data Value

- The perceived value of data influences decisions regarding their life cycle.
- Data value does not necessarily correlate with the financial investment made to collect those data.
- The value of a data resource compounds if it sparks connections among diverse users.



Cost Forecasting Framework

- *Helps forecaster identify major cost drivers*
- *Basis for a cost forecast (not a one-size-fits-all analysis tool)*
- *Will help forecaster identify decisions that impact short- and long-term costs and data value*
- *The forecaster is encouraged to think beyond the specific data state being developed or managed; about how decisions may affect the costs of data management and access in future data states, the transitions to those states, and the future value of data.*
- Use Case: Estimating costs of a new data repository for the BRAIN Initiative



Cost Components of a Biomedical Information Resource

- *Labor*—direct salaries and benefits
- *IT infrastructure*—computer purchase, upgrade, and replacement; storage servers; networking equipment; software
- *IT services*—installation, operation, and maintenance of IT infrastructure
- *Media*—consumable storage (e.g., tapes, DVDs)
- *Licenses and subscriptions*—periodic payments for access/use of data, software, services
- *Facilities and utilities*—space for people and IT infrastructure, utilities (might be incorporated into institutional overhead)
- *Outside services*—consultants, external auditors, off-site media storage, training
- *Travel*—costs for outreach activities, to convene governing boards, and so on.
- *Institutional overhead*—indirect costs for administrative and other support (might be allowed in a contract or grant)
- *Other “soft” costs* (e.g., time users expend to use the data)

(Box 3.2 in text)

Cost Forecasting Framework: Cost Drivers

Data properties that affect the costs of data access and preservation

- A. Content
- B. Capabilities
- C. Control
- D. External Context
- E. Data Life Cycle
- F. Contributors and Users
- G. Availability
- H. Confidentiality
- I. Maintenance and Operations
- J. Standards, Regulatory, and Governance concerns

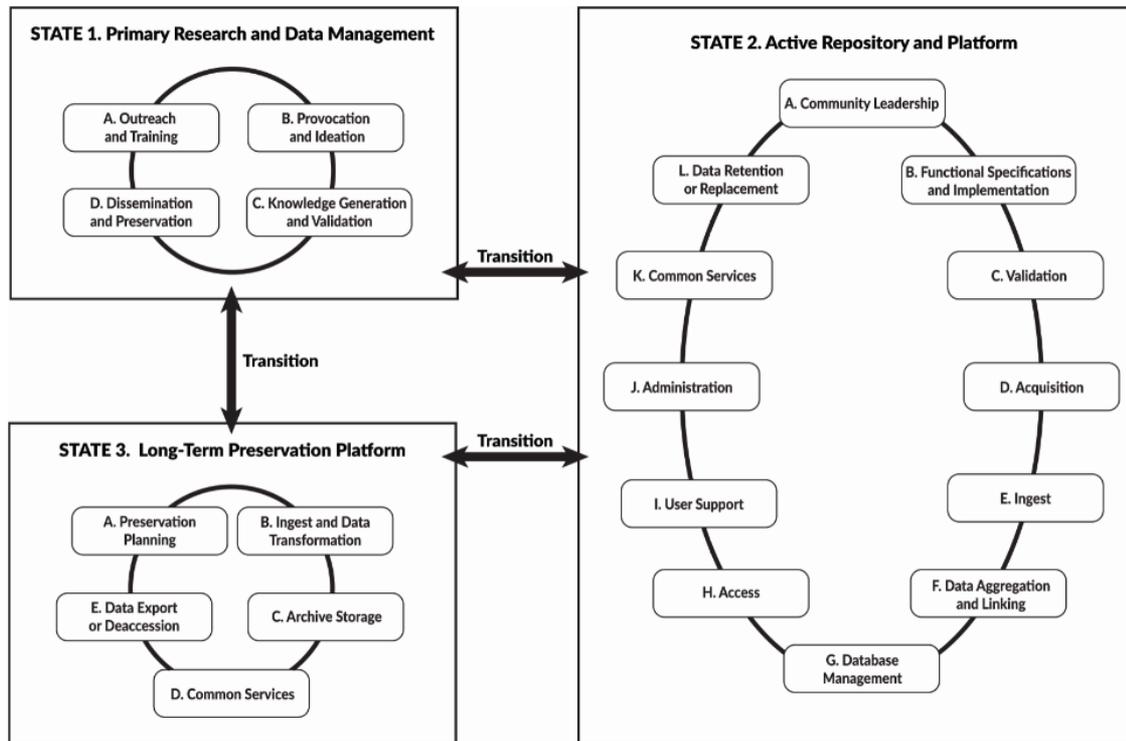


Framework Foundation: Three Data States

State 1: Primary research/data management environment; data are captured and analyzed

State 2: Active repository and platform; data may be acquired, curated, aggregated, accessed, and analyzed

State 3: Long-term preservation platform



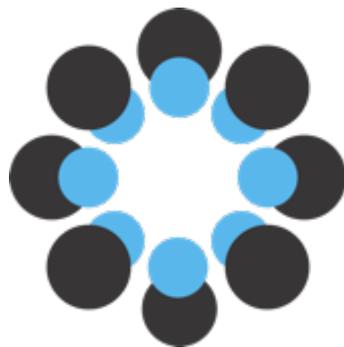
Forecasting Data Costs *for Storage Providers*



Brian Nosek

*co-Founder and Executive Director,
Center for Open Science; Professor of
Psychology, University of Virginia*

Open Science Framework



Open Science Framework

Cost Drivers and Forecasting

<http://osf.io/>

Brian Nosek, Center for Open Science

<http://cos.io/>

OSF: <http://osf.io/>

Launched 2012, free to use (deposit and access), open-source

Full research life-cycle project and data management and archiving

Private, controlled access, and open: Highly configurable

250,000 registered users “producers”; 250 new users/day

>8,000,000 files; 230 TB

>2,500,000 “consumer” users; 16.3M downloads in 2019, pace for 28M in 2020

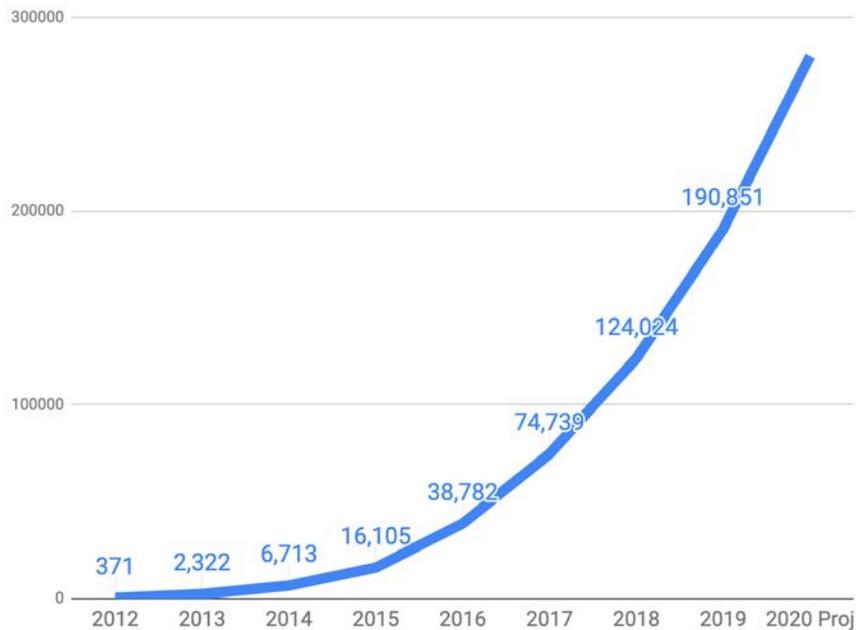
How we can use the cost framework

Forecasting

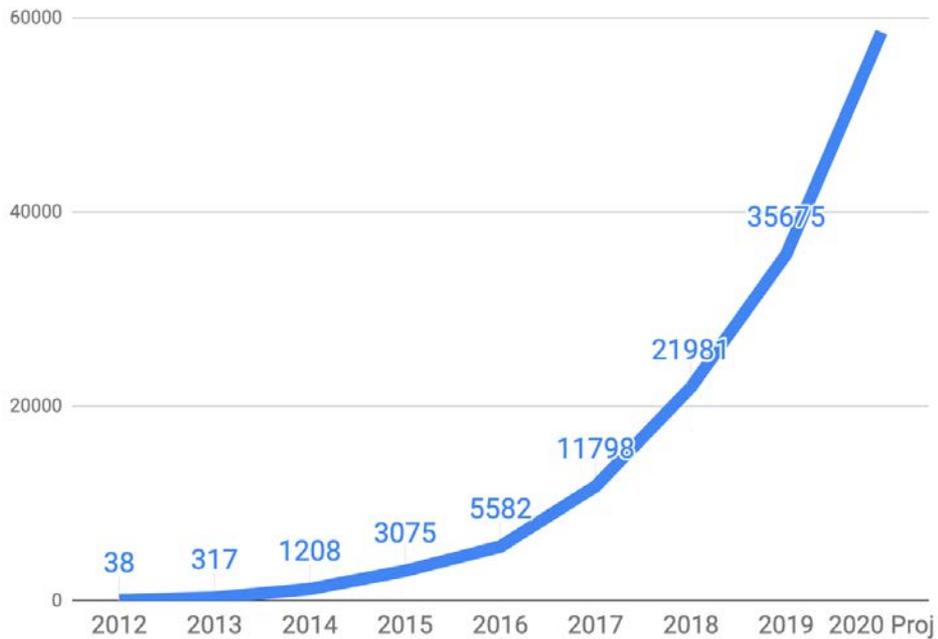
Product Strategy: Sustainability

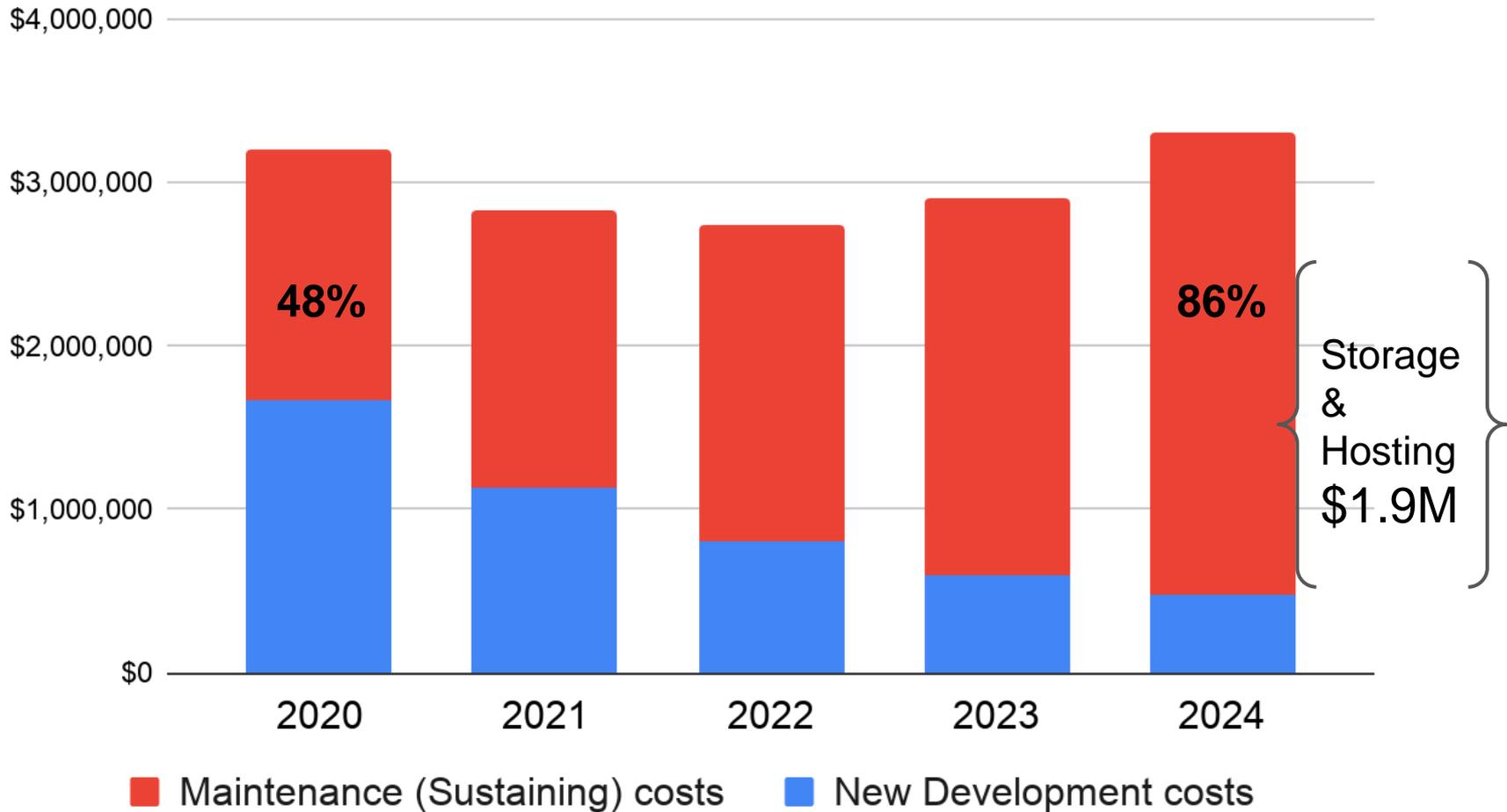
Product Strategy: Design

Number of Registered OSF Users



Number of OSF Study Registrations





How we can use the cost framework

Forecasting

Product Strategy: Sustainability

Product Strategy: Design

OSF Use Cases

Prospective

Plan -> Preregister -> Manage Project/Data -> Archive/Share -> Report

Retrospective

Report -> Prepare Data -> Archive/Share

Using framework to inform product strategy

Prospective

State 1

State 2

Plan -> Preregister -> Manage Project/Data -> Archive/Share -> Report

Retrospective

State 2

Report -> Prepare Data -> Archive/Share

Using framework to inform product strategy

Prospective

State 1

State 2

Plan -> Preregister -> Manage Project/Data -> Archive/Share -> Report



Using framework to inform product strategy

Prospective

State 1

State 2

Plan -> Preregister -> Manage Project/Data -> Archive/Share -> Report



How we can use the cost framework

Forecasting

Product Strategy: Sustainability

Product Strategy: Design

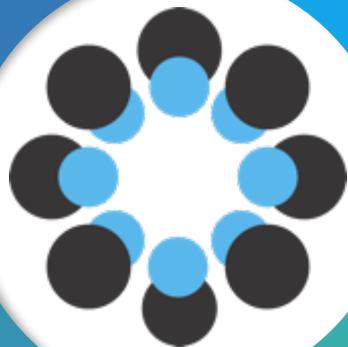
PLANNING

*Explore existing research.
Preregister analysis plan.
Create time-stamped registration.*



DISCOVERY

*Share work.
Improve discovery.
Aggregate findings.*



CONDUCTING

*Open data
management,
collaboration,
storage integration*



REPORTING

*Open data, materials, code.
Open access publishing.*

 OSF **PREPRINTS**

 OSF **REGISTRIES**

PLANNING

*Explore existing research.
Preregister analysis plan.
Create time-stamped registration.*



DISCOVERY

*Share work.
Improve discovery.
Aggregate findings.*



CONDUCTING

*Open data
management,
collaboration,
storage integration*



REPORTING

*Open data, materials, code.
Open access publishing.*

 OSF **MEETINGS**

 OSF

 OSF **COLLECTIONS**

 OSF **INSTITUTIONS**

PLANNING

*Explore existing research.
Preregister analysis plan.
Create time-stamped registration.*

DISCOVERY

*Share work.
Improve discovery.
Aggregate findings.*

CONDUCTING

*Open data
management,
collaboration,
storage integration*

REPORTING

*Open data, materials, code.
Open access publishing.*

PLANNING

Explore existing research.
Register analysis plan.
Create time-stamped registration.



DISCOVERY

Share work.
Improve discovery.
Aggregate findings.

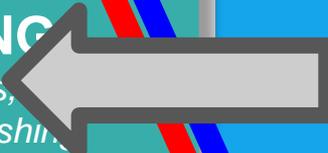


CONDUCTING

Open data
management,
collaboration,
storage integration

REPORTING

Open data, materials,
Open access publishing.



Custom Collections and Repositories on OSF

ILSI North America

Search collection

Active Filters

Program Area

Simulating Large Number Bulk-Product Samplings to Improve Food Safety Sampling Plans
Kawachi and Cheng
This is an OSF preprint for a project funded by ILSI North America. Specifics, this project was selected for funding by the Food Technology Committee in response to its 2017 Request for Proposals for Research on Sampling & Sample Preparation for Manufacturing Food Testing.
Last edited: 2019-01-04 (UTC)

Program Area: Food Microbiology | Status: Active

Effect of the Use of Potassium-Based Sodium Chloride Replacers on Sodium and Potassium Intakes by the US Population
Kang, Verhoff, Wang, and Moore
Sodium intake among the United States population has been recommended to decrease, and efforts have been underway to reduce the amount of sodium in foods. Potassium chloride (KCl) is one of the most effective bulk food sodium chloride (NaCl) replacements due to its ability to perform many of the...

Last edited: 2019-01-09 (UTC)

Program Area: Sodium | Status: Active

Supplemental Materials: Magnitude of the Acceptable Daily Intake (ADI) Values in Nutrition Research Studies that Consider the Safety of Food Additive Residues
Roth, Wilfong, Hines, and Evans
The general objective of the research project is to determine whether when using acceptable daily intake (ADI) values are used to health-based risk estimates, nutrition research studies that consider the safety of one chemical (residue) (EC) are at greater risk to measure or estimate outcomes in...

Last edited: 2019-04-26 (UTC)

Program Area: Food Safety Assessment | Status: Active

Statistical Adequacy and Test Quality in a Retrospective Cohort Study with Normal and Higher Ordinal Scale
Wagner, Rhyne, and Johnson
Statistical methods for higher-ordinal scales may be beneficial to those health-related fields. Higher-ordinal scales should be the preferred format. Although OSF may not host all other chronic diseases, diabetes, osteoporosis. There may be other outcomes affecting those beneficial effects...

Last edited: 2019-01-07 (UTC)

Program Area: Protein | Status: Active

Coronavirus Outbreak Research Collection

My OSF Projects | Add to Collection | Search | Donate | Brian A. Nosek

Search collection

Active Filters

Type

Status

Supplemental materials for preprint: COVID-19 - An Environmental Pandemic
DIATTA
Coronae Virus - A Spasmodic (Non-dictionary term) - Spasmodic + Epidemic - used here, Spasmodic w/r t our environment. A critical viralological pathogen throughout the World is destroying the environmental sustainability. Increasing the Death rate every Day. The COVID-19 is the latest model is one ...
Last edited: 2020-05-29 (UTC)

Type: COVID-19 | Status: Active

Correlates of Health-Protective Behavior During the Initial Days of the COVID-19 Outbreak in Germany
Jozsef, Schubert, Hering, and Jermir
The coronavirus outbreak manifested in Germany in March 2020. It was met a combination of mandatory changes (closing of public institutions) and recommended changes (hygiene behavior, physical distancing). It has been emphasized that health-protective behavior such as increased hygiene or physical di...
Last edited: 2020-05-27 (UTC)

Type: COVID-19 | Status: Active

The impact of the COVID-19 shutdown on gambling in Australia
Gainsbury, Swanton, Burgess, and I more

Metascience Collection

My OSF Projects | Add to Collection | Search | Donate | Brian A. Nosek

Search collection

Active Filters

Type

Status

Which findings should be published?
Key and Frankel
Given a variety of journal topics, what is the socially optimal rule for whether an empirical finding should be published? Suppose that the goal of publication is to inform the public about a given-relevant story. Then journals should publish extreme results, measuring ones their impact for the public.
Last edited: 2019-10-02 (UTC)

Type: Research | Status: Active

Do p-Values Lose Their Meaning in Exploratory Analyses? It Depends How You Define the Familywise Error Rate
Rubin
Several researchers have recently argued that p-values lose their meaning in exploratory analyses due to an uncorrected inflation of the alpha level (e.g., Rosow & others, 2014; Wagenmakers, 2016). For this argument to be valid, the familywise error rate must be defined in relation to the number of t...
Last edited: 2019-10-15 (UTC)

Type: Other | Status: Completed

What type of type I error?
Rubin
The replication crisis has caused researchers to distinguish between exact replications, which duplicate all...

/end

nosek@cos.io

These slides: <https://osf.io/zsqyp/>

OSF is State 1 management and State 2 repository

State 1

Can receive direct input from data acquisition tools (Open Sesame)

Interacts with active analysis pipelines (osfr package; JASP Stats)

Collaborative teams do data management work on OSF (Privately or openly)

Integrations with live data environments (Dropbox, Drive, Box, GitHub, etc.)

Registration of research and data management plans prior to data acquisition

State 2

Archiving and sharing data, protocols, code

Interfaces/Collections for aggregating content

Custom curation/moderation processes

Metadata and FAIR standards

Open and Controlled Access

Integrations with state 2 repositories

Forecasting Data Costs

for Storage Providers



Alex Ropelewski

*Director, Biomedical Applications Group,
Pittsburgh Supercomputing Center; PI and
Operations Director, Brain Image Library*

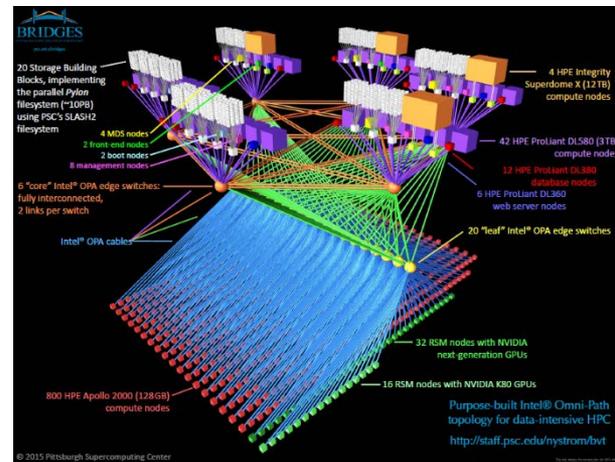
The Brain Image Library: an NIH BRAIN Data Repository

The Brain Image Library

Mission: National public resource enabling researchers to deposit, analyze, mine, share and interact with microscopy datasets of the brain.

Scope:

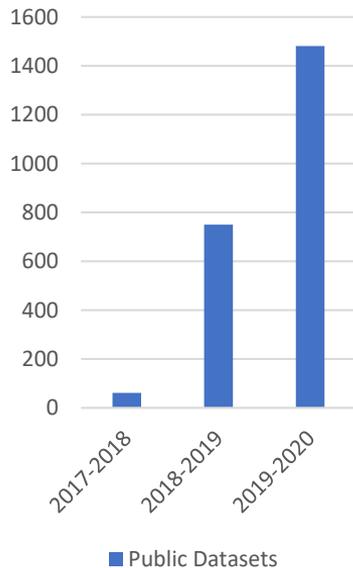
- Permanent repository for high-quality brain microscopy datasets
 - Whole brain images of mouse, rat, other mammals and model organisms
 - Targeted experiments Including connectivity between cells and spatial transcriptomics (*FISH)
 - Historical collections
- Provide HPC computing capability local to the data for pre-submission data processing and post-submission exploration
 - Enclave access to pre-release data
 - Research access to restricted-access, secured data
- Provide user access and support



Benninger et. al.2020.
Cyberinfrastructure of a Multi-Petabyte
Microscopy Resource for Neuroscience
Research. (PEARC '20).
<https://doi.org/10.1145/3311790.3396653>

Data Characteristics

Public Datasets



We feature in this poster one of the 34 IMODAT datasets received from the 2mg UFR project at the Allen Institute for Brain Science. Imaging was performed at Huzhou University of Science and Technology. This dataset represents whole-brain imaging of a transgenic mouse (Gata1-RE32-Cx36RT2-ws1803101-CF79-16100121-CF4-12-17A11) in which a small sampling of neurons, including claustrum projection neurons, are fluorescently labeled [6]. The green channel shows where CFP is expressed, and can be used for tracing these neurons. It is important that only a sparse subset be labeled; otherwise, it would become much more difficult to trace individual neurons with high confidence. For this specimen GL currently has available 4 neuron morphologies in asc format, of which we illustrate one (17781_00001.swc). Intentionally, thousands of other morphologies could be reconstructed from this dataset.

Image 0920 (of 11011), red channel, 34721x49600 pixels, x, y resolution = 0.5µm, z spacing = 1µm; field of view: 19.11 x 12.15 mm.

Combined red and green channels. Zoom factor = 1.28; field-of-view: 14.93 x 9.47 mm.

Image 0920 (of 11011), green channel, 34722x49600 pixels, x, y resolution = 0.5µm, z spacing = 1µm; field of view: 19.11 x 12.15 mm.

Zoom factor = 5.24; field of view: 3.64 x 2.36 mm.

Zoom factor = 2.61; field of view: 7.32 x 4.73 mm.

When all axonal and dendritic branches have been traced, one can visualize the entire axonal projection pattern of this particular claustrum projection neuron. The bounding box for this neuron is approximately 4.7 x 4.1 x 2.2 mm, so its scale is comparable to that of the image immediately to the left.

Zoom factor = 16.51; field of view: 1.82 x 1.18 mm.

Zoom factor = 34.00; field of view: 0.56 x 0.38 mm. (Pixel resolution is the same as the acquisition resolution = 0.35 x 0.35 µm.)

Axons and dendrites of an individual labeled neuron can be traced by starting at a soma and following the green lines through successive images in z. This figure was generated from the 17781_00001.swc file updated to RB [7] and rendered with the StackViewer tool [8].

10D View [http://vsm06.all.psu.edu/vaa3d/395716117345/018030_CGm0e_0L6-486uDF_VIP-546_LALB-647_Vh_405a_011.daf]

10D View [http://vsm06.all.psu.edu/vaa3d/395716117345/018030_CGm0e_0L6-486uDF_VIP-546_LALB-647_Vh_405a_011.daf]

Lifecycle

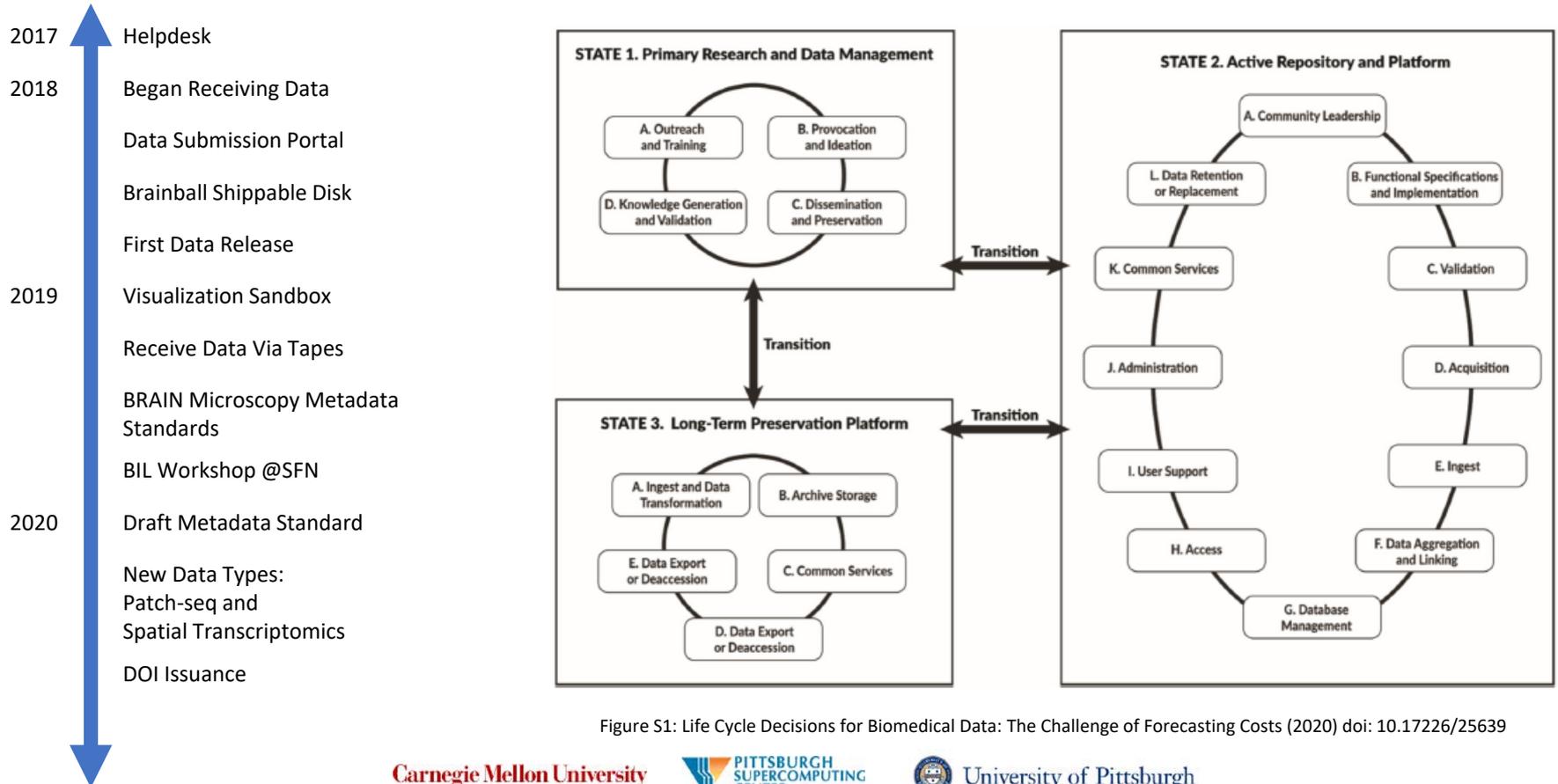


Figure S1: Life Cycle Decisions for Biomedical Data: The Challenge of Forecasting Costs (2020) doi: 10.17226/25639

Cost Drivers

Content	Now	Future
Size	●	●
Complexity/Diversity	●	●
Metadata	●	●
Depth vs Breadth	●	●
Processing/Fidelity	●	●
Replaceability	●	●

Capabilities	Now	Future
User Annotation	●	●
Persistent Identifiers	●	●
Citation	●	●
Search Capabilities	●	●
Data Linking/Merging	●	●
Use Tracking	●	●
Analysis/Visualization	●	●

Control	Now	Future
Content Control	●	●
Quality Control	●	●
Access Control	●	●
Platform Control	●	●

External Content	Now	Future
Resource Replication	●	●
External Information Dependencies	●	●
Distinctiveness	●	●

Content	Now	Future
Size	●	●
Complexity/Diversity	●	●
Metadata	●	●
Depth vs Breadth	●	●
Processing/Fidelity	●	●
Replaceability	●	●

Contributors	Now	Future
Contributor Base	●	●
User Base	●	●
Training/Support	●	●
Outreach	●	●

Availability	Now	Future
Outage Tolerance	●	●
Currency	●	●
Response Time	●	●
Local vs Remote	●	●

Confidentiality	Now	Future
Confidentiality	●	●
Ownership	●	●
Security	●	●

Maintenance	Now	Future
Integrity Check	●	●
Data Transfer	●	●
Risk Management	●	●
System Reporting	●	●
Billing	●	●

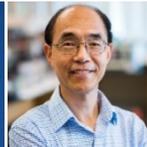
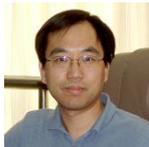
Standards	Now	Future
Applicable Standards	●	●
Regulatory/Legislative Environment	●	●
Governance	●	●
External Consultation	●	●

Relative Cost Potential

- : Low
- : Med
- : High

Modified from Appendix E: Life Cycle Decisions for Biomedical Data: The Challenge of Forecasting Costs (2020) doi: 10.17226/25639

Thank You!



Contact us at: bil-support@psc.edu



Marcel Bruchez (PI)
Greg Fisher (Microscope)



Alexander Ropelewski (Contact PI)
Kathy Benninger (Networking)
Greg Hood (Image Analysis+ HPC)
Derek Simmel (Systems+Data)
Arthur Wetzel (Image Analysis)
Luke Tuite (User Support+Web)

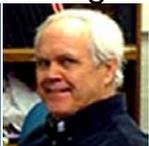


Simon Watkins (PI)
Alan Watson (Microscope)

Marce



Greg



Alex



Greg



Derek



Kathy



Art



Luk



Ivan



Simon



Alan



This project is supported by the National Institute Of Mental Health of the National Institute of Health under Award Number R24MH114683. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Forecasting Data Costs *for Storage Providers*

Please submit questions using the Q&A button in the zoom menu.



Ilkay Altintas
(Moderator)

University of California San
Diego



Clifford Lynch

Coalition for Networked
Information



Brian Nosek

Center for Open Science



Alex Ropelewski

Pittsburgh Supercomputing
Center

Forecasting Data Costs for Researchers, Funders, and Storage Providers

August 2020 weekly webinar series, 12-1pm ET

Recordings available at <https://vimeo.com/showcase/7444639>

August 13: *Forecasting Data Costs for Researchers*

August 20: *Forecasting Data Costs for Funding Institutions*

August 27: *Forecasting Data Costs for Storage Providers*

This webinar series is sponsored by the
National Library of Medicine of the
National Institutes of Health



U.S. National Library
of Medicine