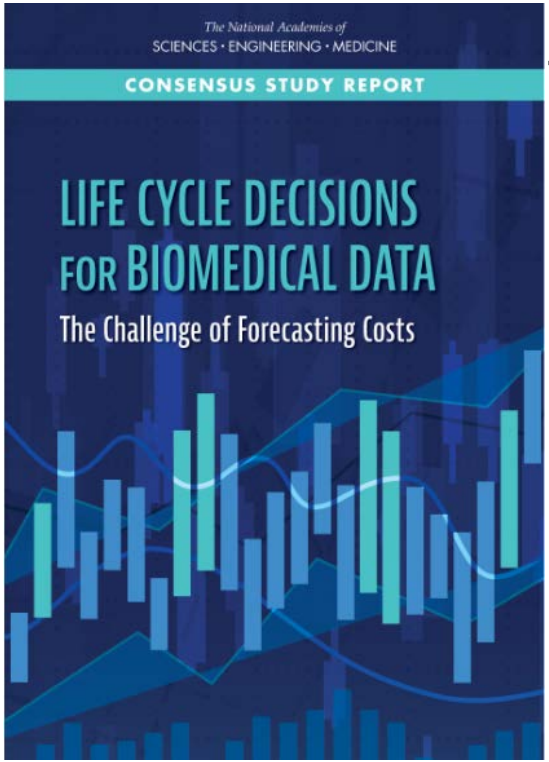
The National Academies of SCIENCES ENGINEERING MEDICINE

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***

Watch webinar videos and learn more about BMSA at https://biomed-data-costs.eventbrite.com/

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

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

https://vimeo.com/showcase/7444639

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

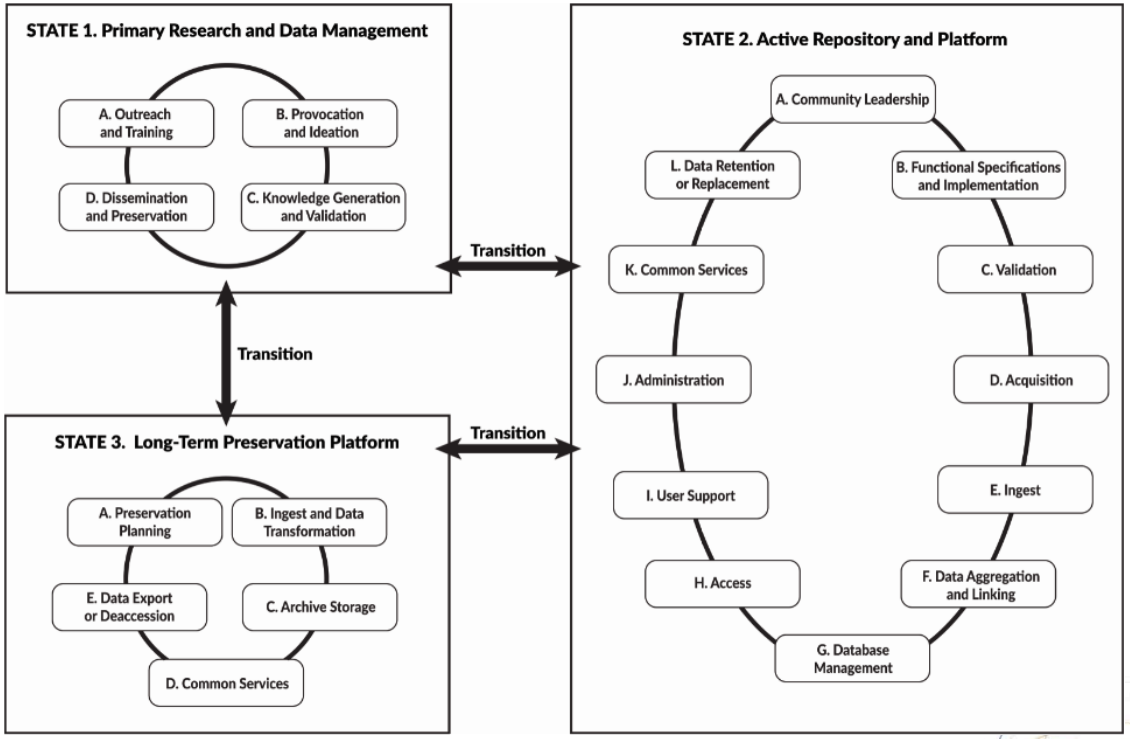
A. Content
B. Capabilities
C. Control
D. External Context
E. Data Life Cycle
F. Contributors and Users
G. Availability
H. Confidentiality
I. Maintenance and Operations
J. Standards, Regulatory, and Governance concerns

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/

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

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

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

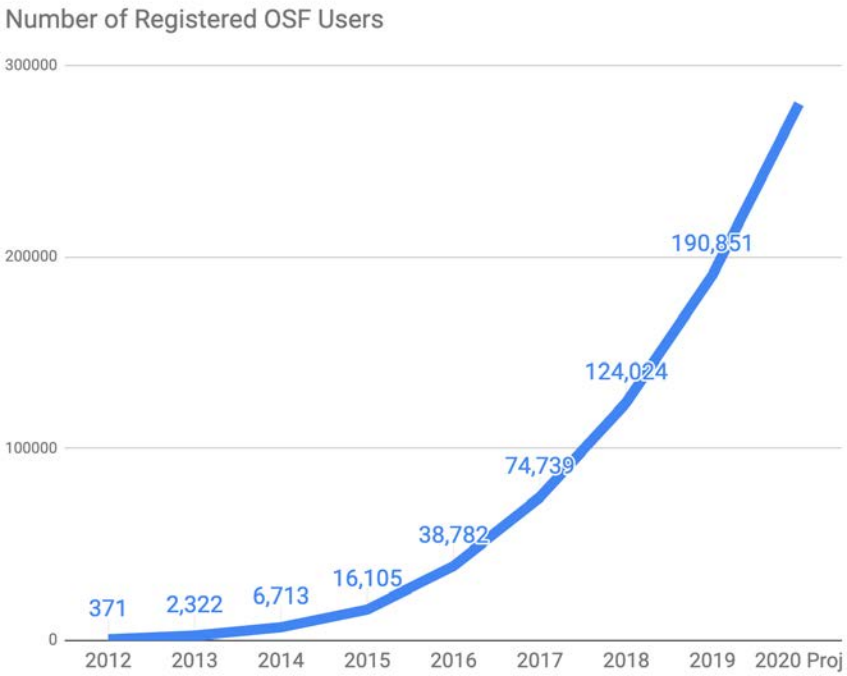
How we can use the cost framework

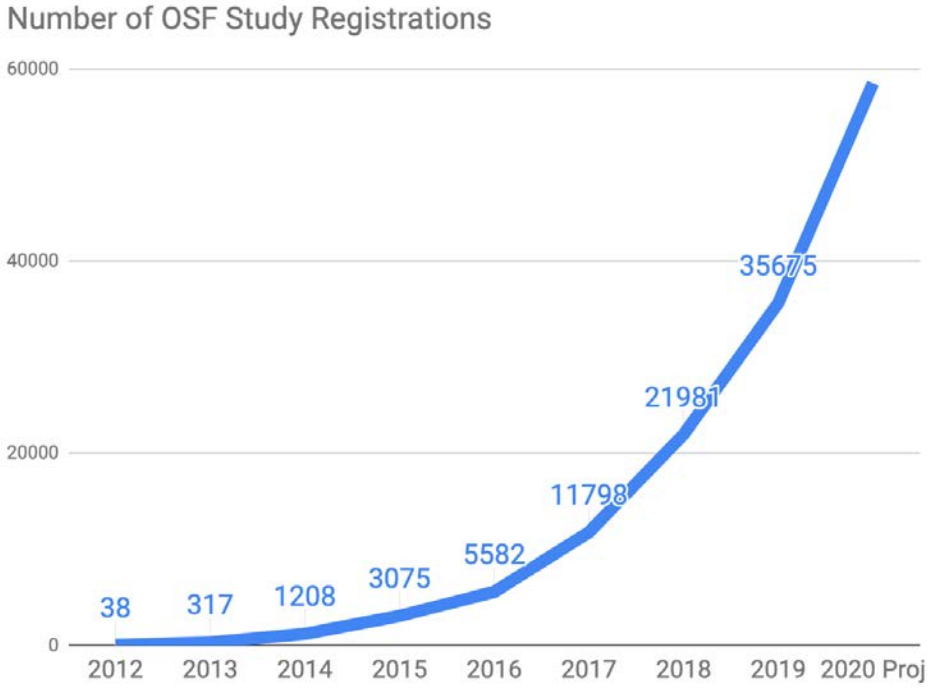
Forecasting

Product Strategy: Sustainability

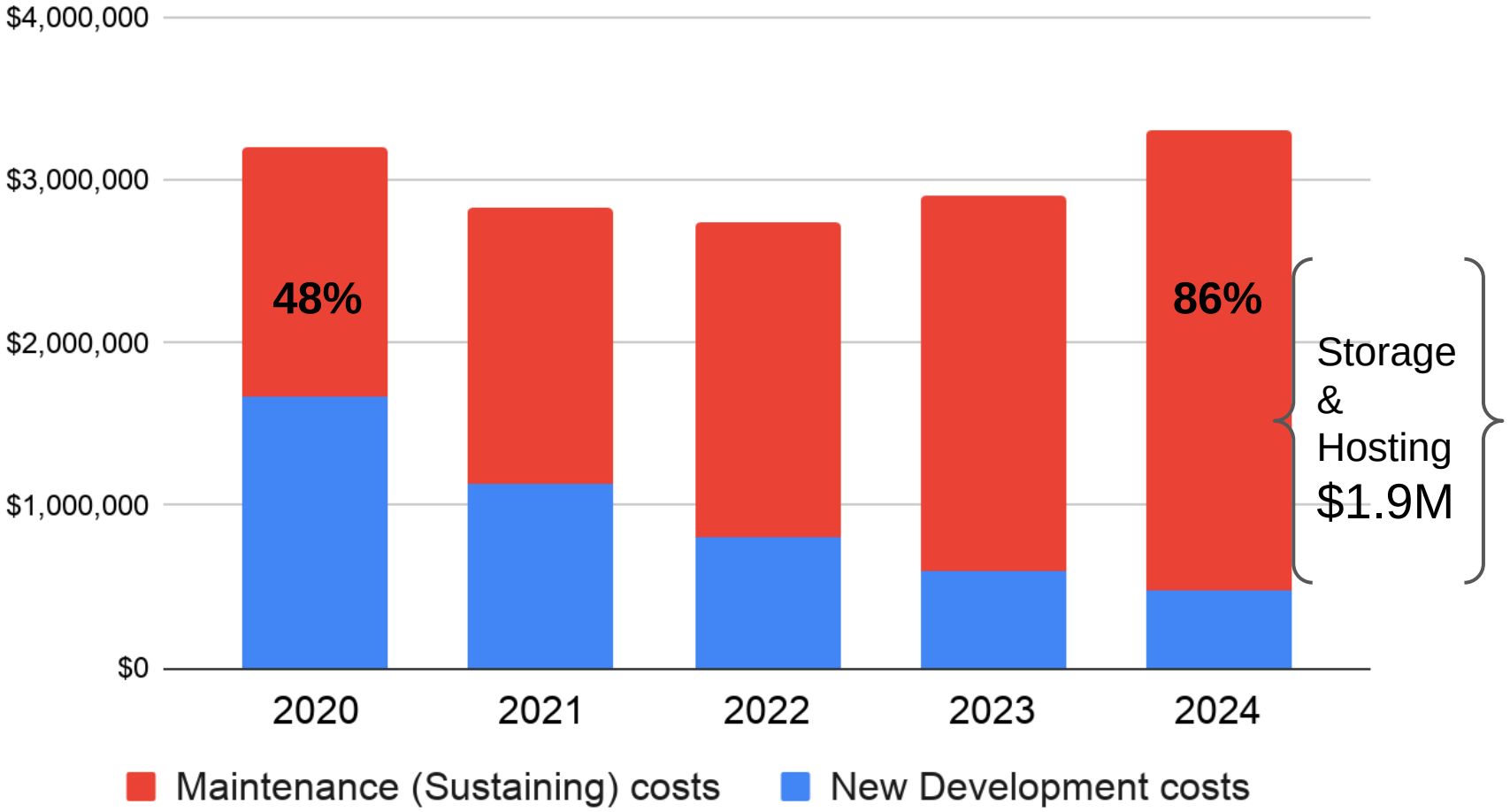
Product Strategy: Design

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





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



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

How we can use the cost framework

Forecasting

Product Strategy: Sustainability

Product Strategy: Design

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

OSF Use Cases

Prospective

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

Retrospective

Report -> Prepare Data -> Archive/Share

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

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

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

Using framework to inform product strategy

Prospective

State 1 State 2

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

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

Using framework to inform product strategy

Prospective

State 1 State 2

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

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

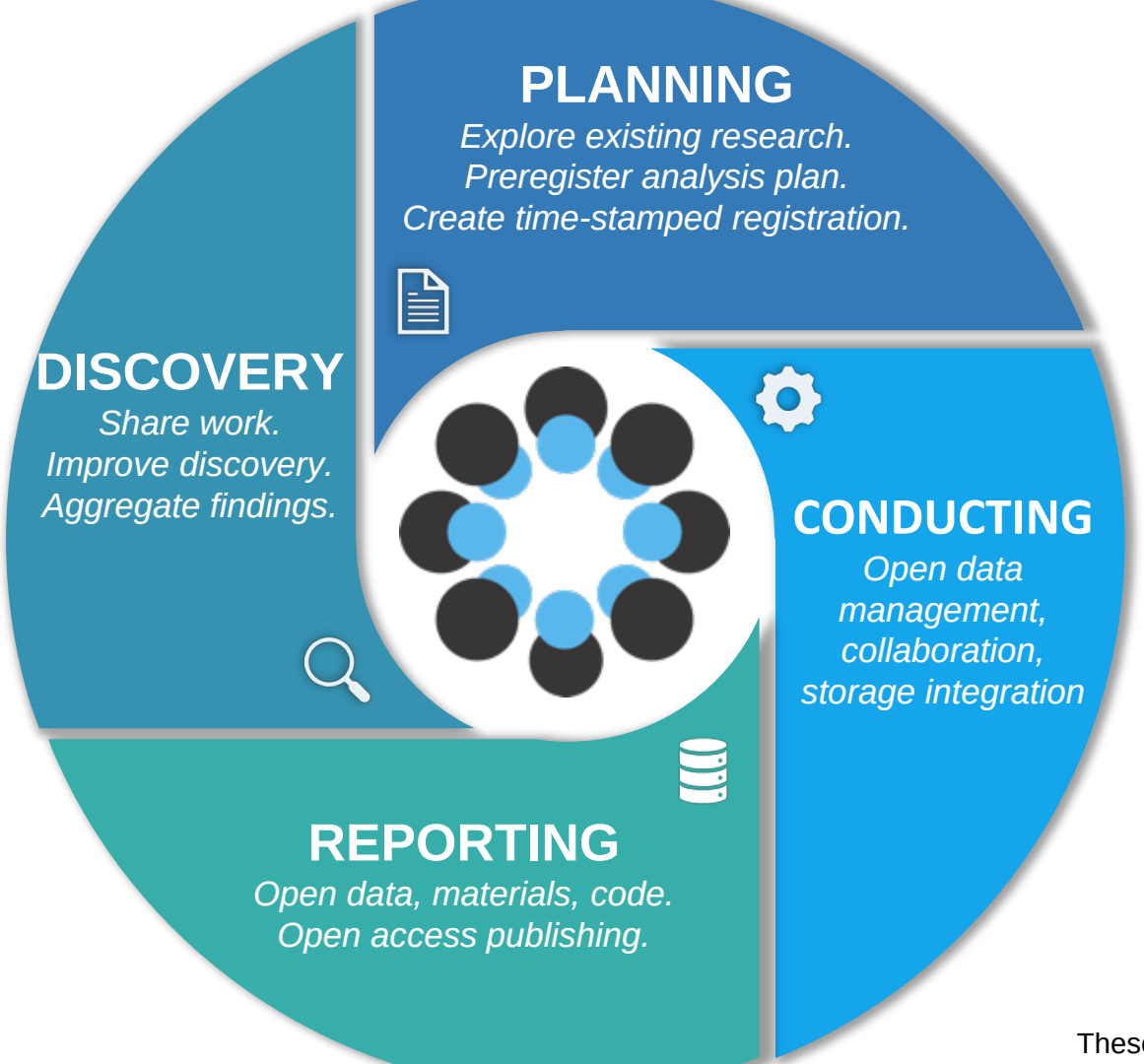
How we can use the cost framework

Forecasting

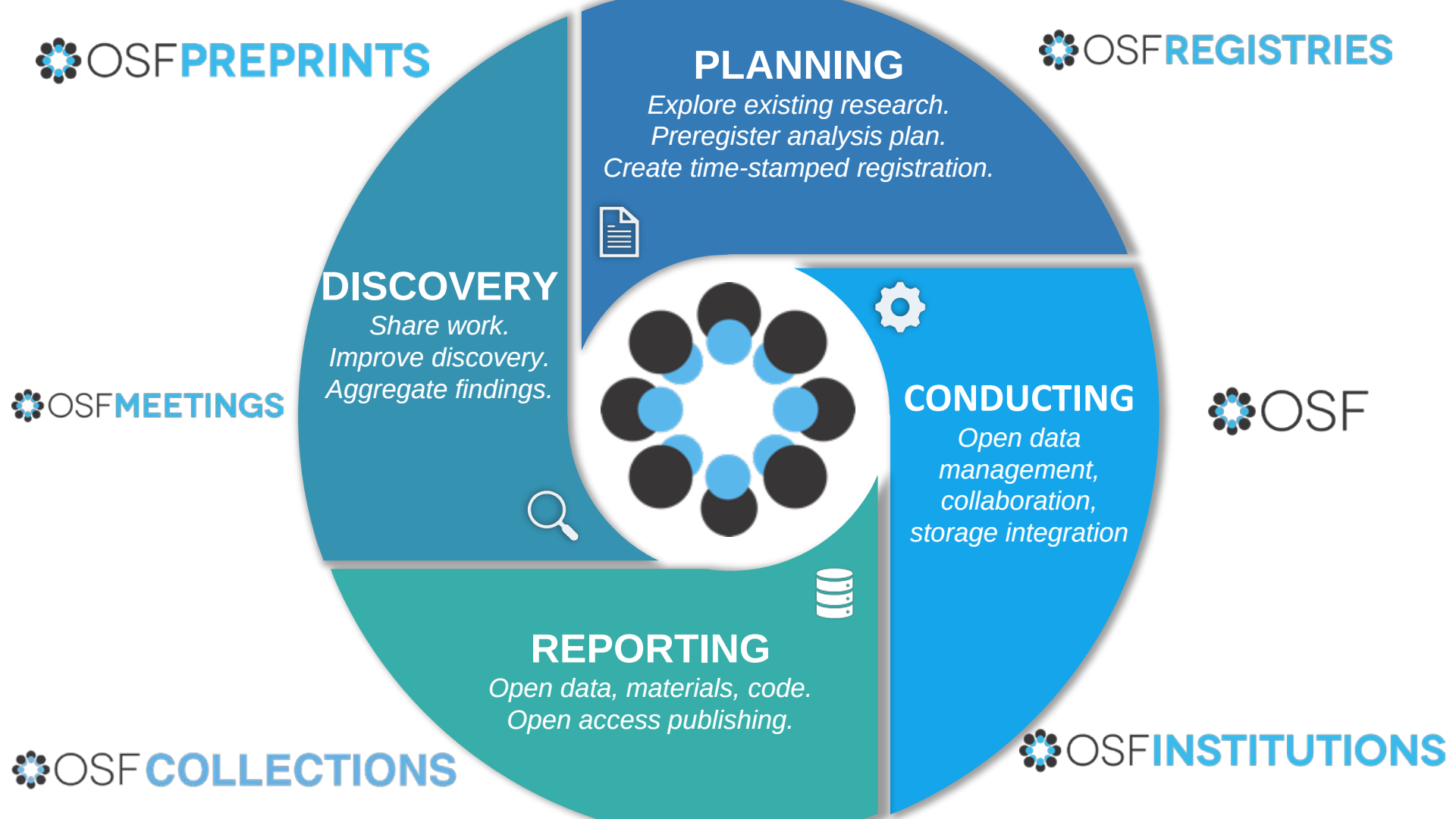
Product Strategy: Sustainability

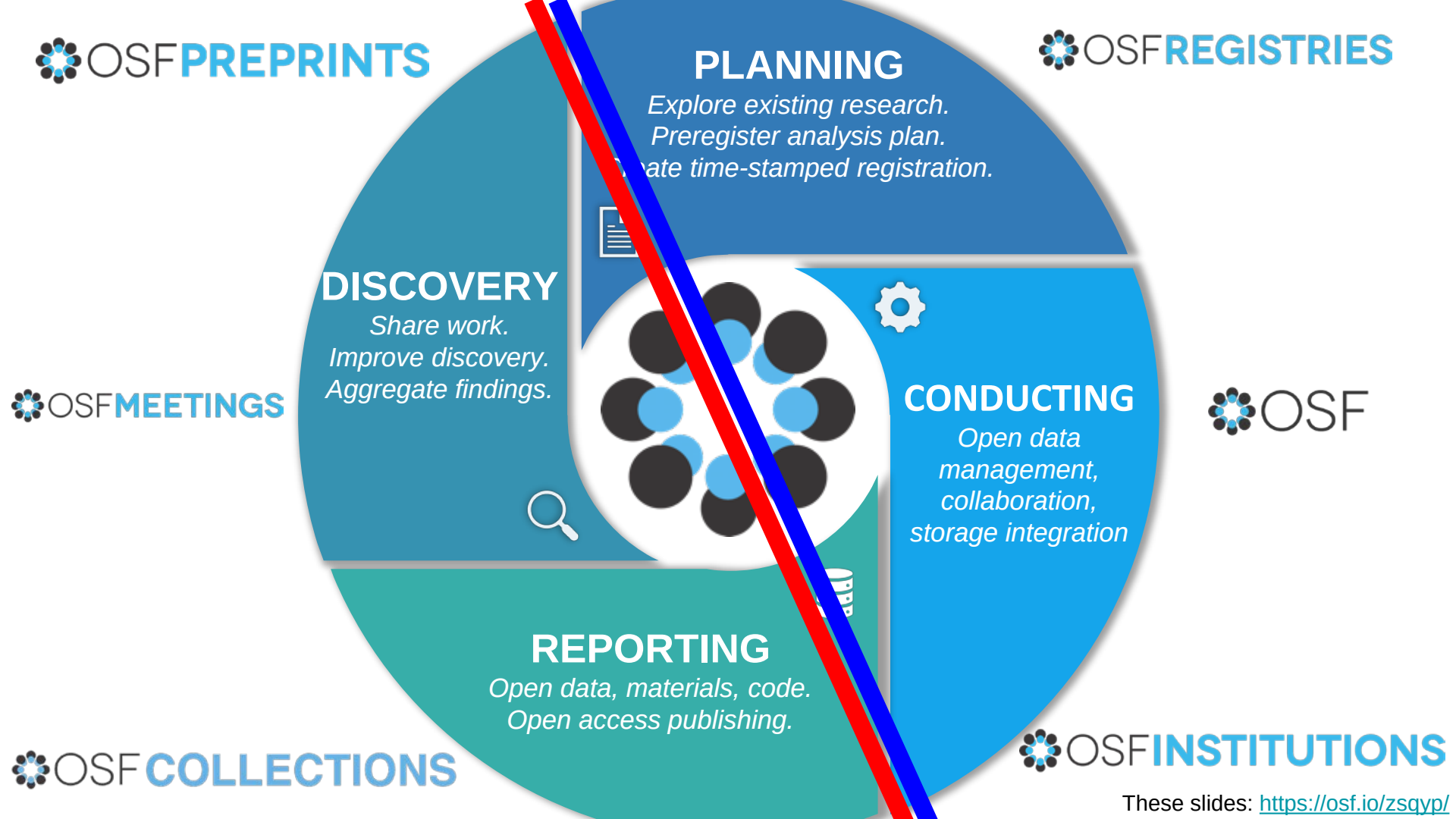
Product Strategy: Design

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



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

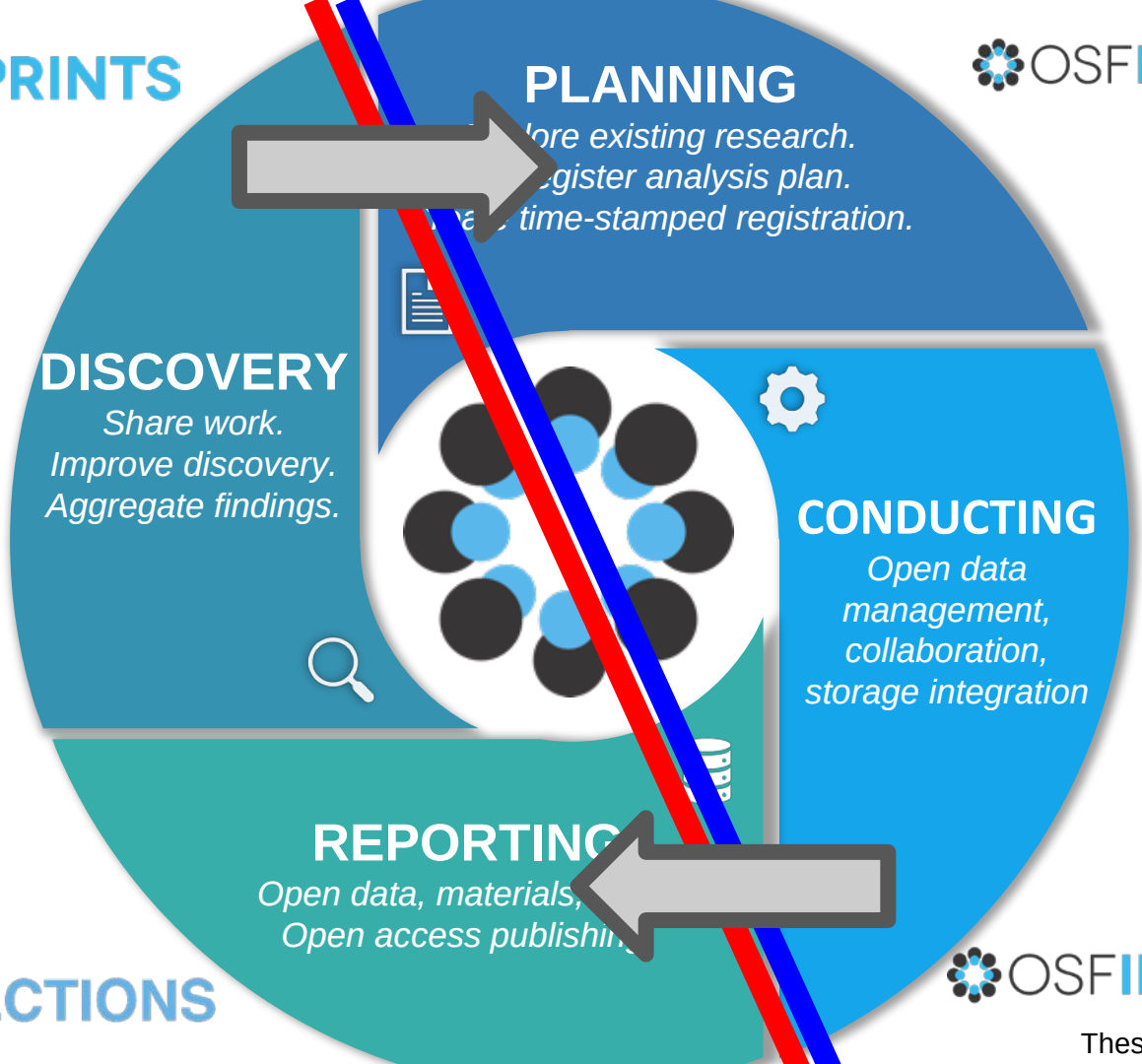

OSF PREPRINTS
OSF REGISTRIES
PLANNING
Explore existing research.
Preregister analysis plan.
Create time-stamped registration.
DISCOVERY
Share work.
Improve discovery.
Aggregate findings.
OSF MEETINGS
CONDUCTING
Open data management, collaboration, storage integration
OSF
REPORTING
Open data, materials, code.
Open access publishing.
OSF COLLECTIONS
OSF INSTITUTIONS



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

OSF PREPRINTS

OSF REGISTRIES



OSF MEETINGS

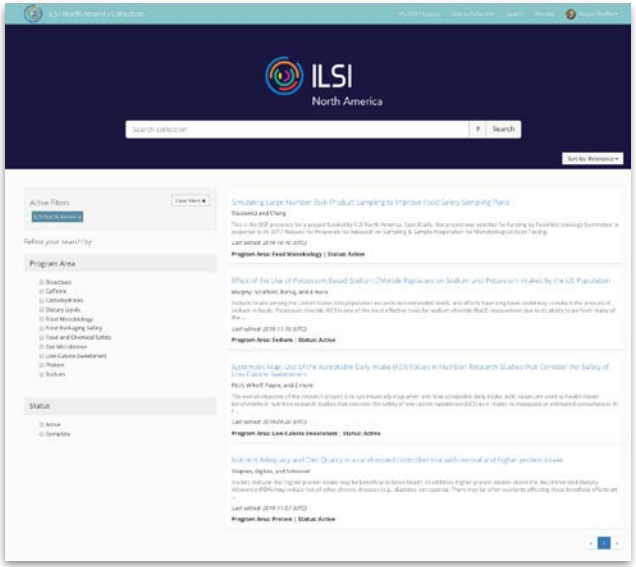
OSF

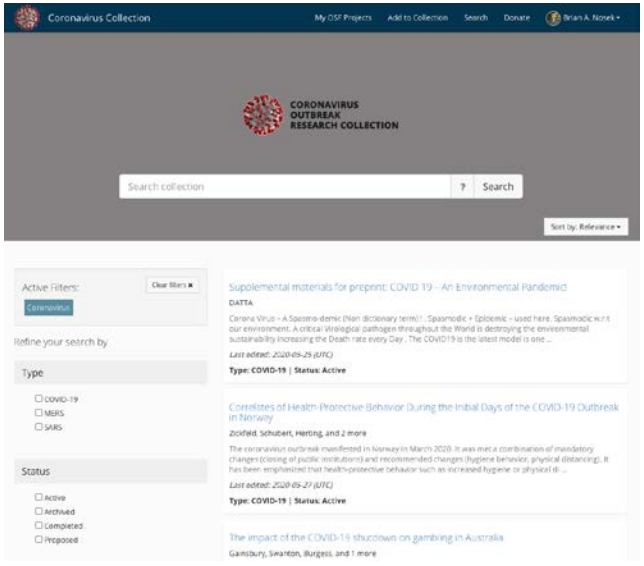
OSF COLLECTIONS

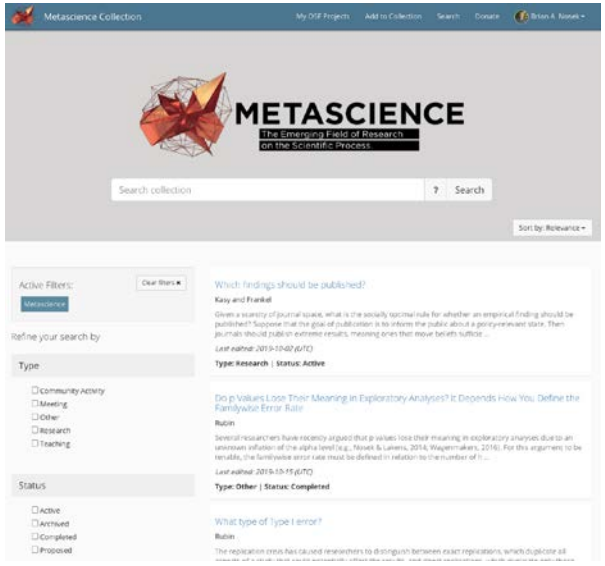
OSF INSTITUTIONS

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

Custom Collections and Repositories on OSF







/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

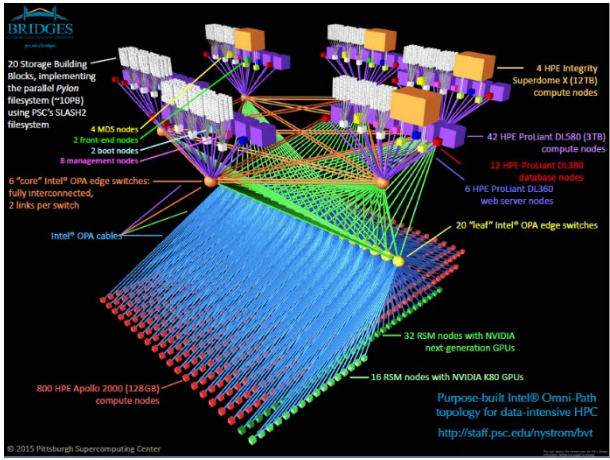
The Brain Image Library: an NIH BRAIN Data Repository

The Brain Image Library

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

Scope:

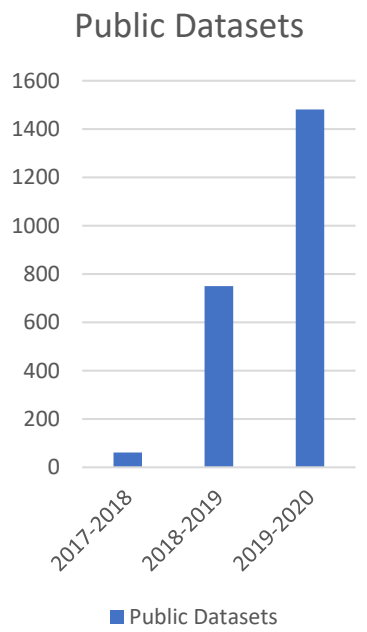
- Permanent repository for high-quality brain microscopy datasets
  - Whole brain images of mouse, rat, other mammals and model organisms
  - Targeted experiments Including connectivity between cells and spatial transcriptomics (*FISH)
  - Historical collections
- Provide HPC computing capability local 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

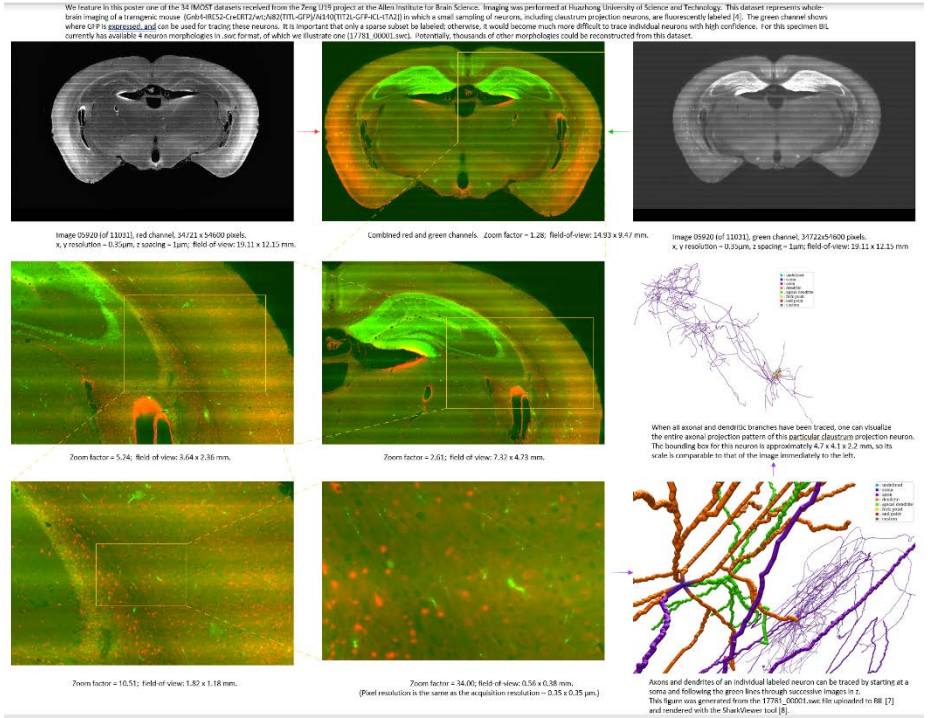


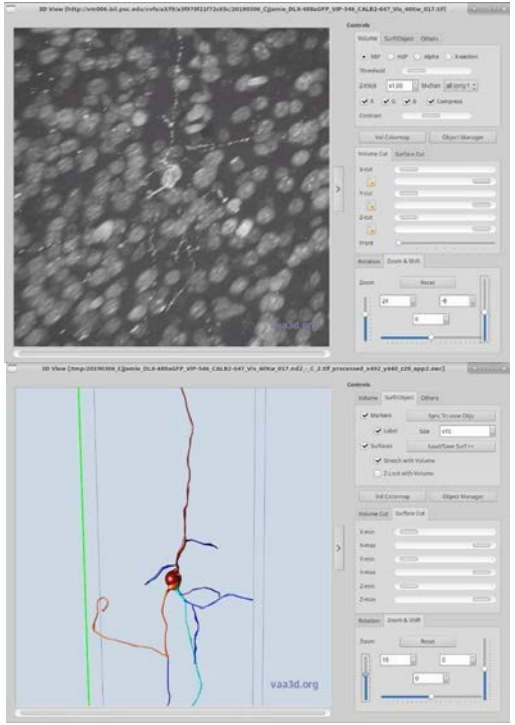
Benninger et. al.2020. Cyberinfrastructure of a Multi-Petabyte Microscopy Resource for Neuroscience Research. (PEARC '20). https://doi.org/10.1145/3311790.3396653

Carnegie Mellon University PITTSBURGH SUPERCOMPUTING CENTER University of Pittsburgh

Data Characteristics







Lifecyle

2017 Helpdesk

2018 Began Receiving Data

Data Submission Portal

Brainball Shippable Disk

First Data Release

2019 Visualization Sandbox

Receive Data Via Tapes

BRAIN Microscopy Metadata Standards

BIL Workshop @SFN

2020 Draft Metadata Standard

New Data Types:
Patch-seq and
Spatial Transcriptomics

DOI Issuance

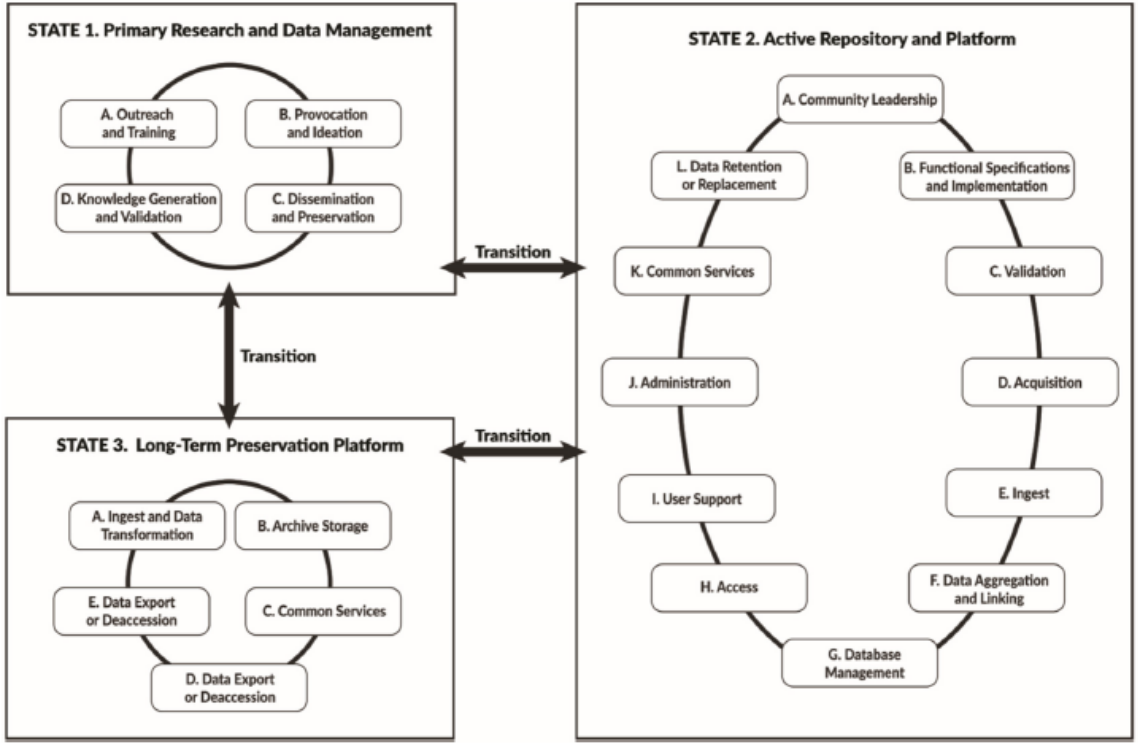


Figure S1: Life Cycle Decisions for Biomedical Data: The Challenge of Forecasting Costs (2020) doi: 10.17226/25639

Carnegie Mellon University PITTSBURGH SUPERCOMPUTING CENTER University of Pittsburgh

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

Carnegie Mellon University | Pittsburgh Supercomputing Center | University of Pittsburgh

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***

Watch webinar videos and learn more about BMSA at https://biomed-data-costs.eventbrite.com/