

UCLA

Presentations

Title

The Challenges of Reproducibility in Data-Scarce Fields

Permalink

<https://escholarship.org/uc/item/529611vj>

Authors

Borgman, Christine L.
Darch, Peter T.

Publication Date

2017-05-09

Copyright Information

This work is made available under the terms of a Creative Commons Attribution-NonCommercial-NoDerivatives License, available at <https://creativecommons.org/licenses/by-nc-nd/4.0/>

The Challenges of Reproducibility in Data-Scarce Fields

Christine L. Borgman

Distinguished Professor and Presidential Chair in Information Studies

University of California, Los Angeles

<http://christineborgman.info>

@scitechprof

Peter T. Darch

Assistant Professor

School of Information Sciences

University of Illinois at Urbana-Champaign

DataONE Webinar, May 9, 2017

<https://www.dataone.org/webinars/challenges-reproducibility-data-scarce-fields>

<https://knowledgeinfrastructures.gseis.ucla.edu>



Christine Borgman



Peter Darch



Ashley Sands



Irene Pasquetto



Bernie Randles



Milena Golshan



ILLINOIS
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

Data sharing policies

- European Union
- U.S. Federal research policy
- Research Councils of the UK
- Australian Research Council
- Individual countries, funding agencies, journals, universities



Supported by
wellcometrust



Australian Government
National Health and Medical Research Council



National Science Foundation
WHERE DISCOVERIES BEGIN

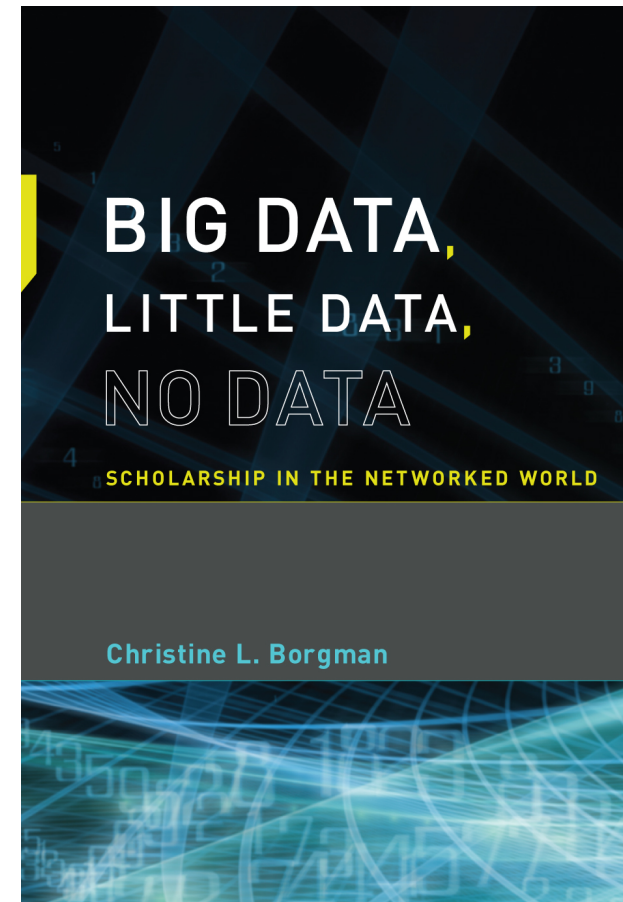


National Institutes of Health
Turning Discovery Into Health



Why Share Research Data?

- To reproduce research
- To make public assets available to the public
- To leverage investments in research
- To advance research and innovation



MIT Press, 2015

Lack of incentives to share data



- Rewards for publication
- Effort to document data
- Competition, priority
- Control, ownership

Why Reuse Research Data?

- To reproduce research
- To replicate research
- To verify or validate research
- To integrate with other data

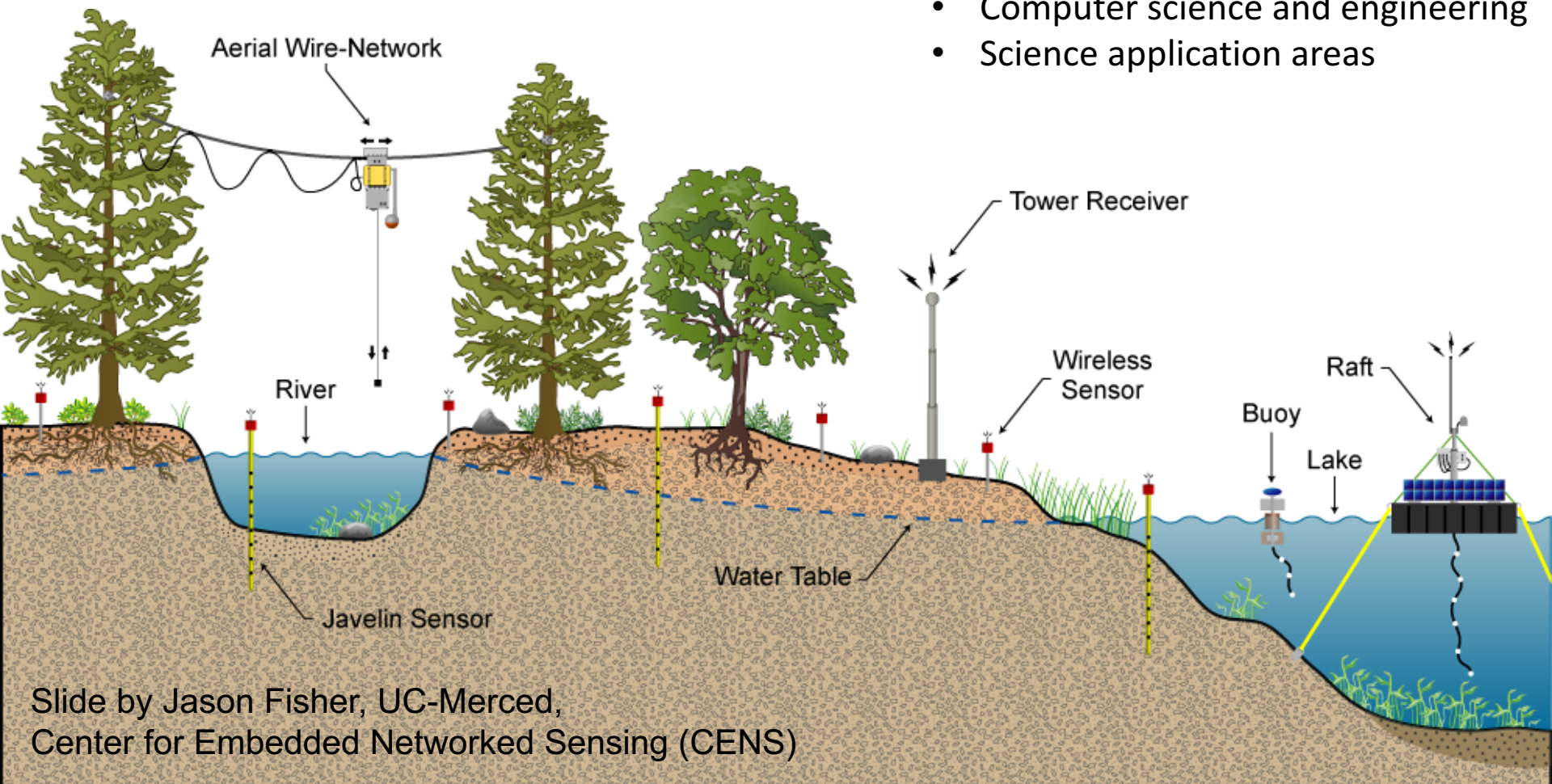




Data

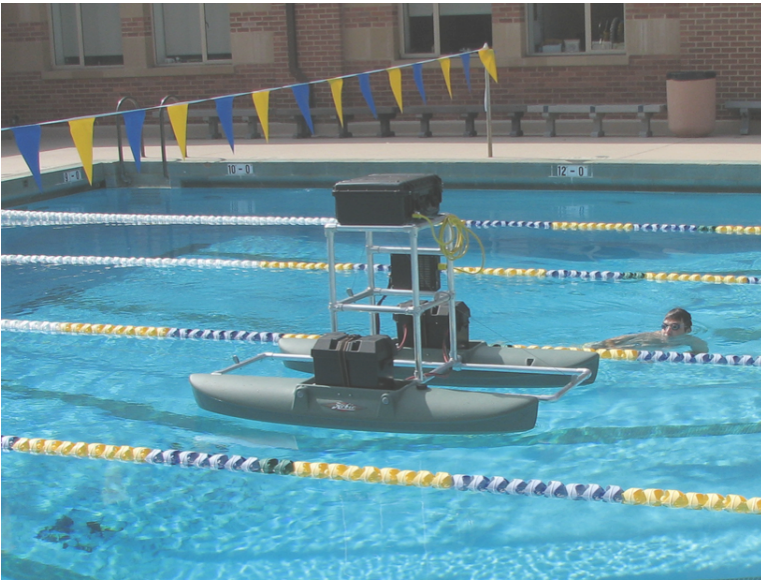
Center for Embedded Networked Sensing

- NSF Science & Tech Ctr, 2002-2012
- 5 universities, plus partners
- 300 members
- Computer science and engineering
- Science application areas



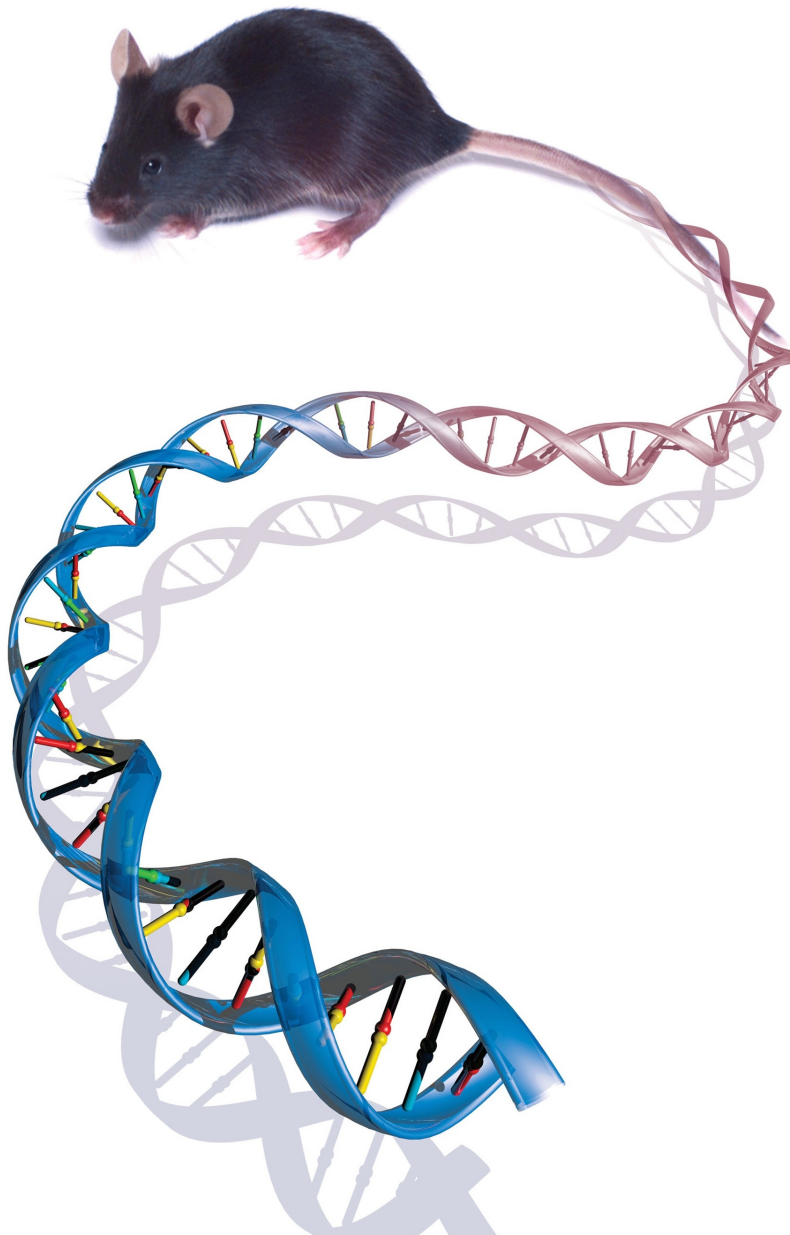
Documenting Data for Interpretation

Engineering researcher:
“Temperature is temperature.”



CENS Robotics team

Biologist: ***“There are hundreds of ways to measure temperature.*** ‘The temperature is 98’ is low-value compared to, ‘the temperature of the surface, measured by the infrared thermopile, model number XYZ, is 98.’ That means it is measuring a proxy for a temperature, rather than being in contact with a probe, and it is measuring from a distance. The accuracy is plus or minus .05 of a degree. I [also] want to know that it was taken outside versus inside a controlled environment, how long it had been in place, and the last time it was calibrated, which might tell me whether it has drifted..”



Data are representations of observations, objects, or other entities used as evidence of phenomena for the purposes of research or scholarship.

C.L. Borgman (2015). *Big Data, Little Data, No Data: Scholarship in the Networked World*. MIT Press

If Data Sharing Is the Answer, What Is the Question?

- Goals

- Explicate data, sharing, reuse, openness, infrastructure across scientific domains
- Identify new models of scientific practice

- Dimensions

- Mixtures of domain expertise
- Factors of scale
- Centralization of data collection and analysis



Qualitative Methods

- Document analysis
 - Public and private documents and artifacts
 - Official and unofficial versions of scientific practice
- Ethnography
 - Observing activities on site and online
 - Embedded for days or months at a time
- Interviews
 - Questions based on our research themes
 - Compare multiple sites over time



Current Research Sites

Domain	Focus	Topic
Astronomy sky surveys	Place: sky and universe	Survey of night sky
Deep subseafloor biosphere	Place: under ocean floor	Microbial life and environment
Biomedical collaboration	Problem: data sharing and reuse in an interdisciplinary context	Genomics of four model organisms
Computational science	Problem: Data analysis at scale	Computing in physical and life sciences
Astronomy phenomena	Place: sky and universe	Orbits, black holes, gravity

Research Question 1

How do the *mixtures of domain expertise* influence the collection, use, and reuse of data – and vice versa?

Domain

Astronomy sky surveys

Deep seafloor biosphere

Biomedical research

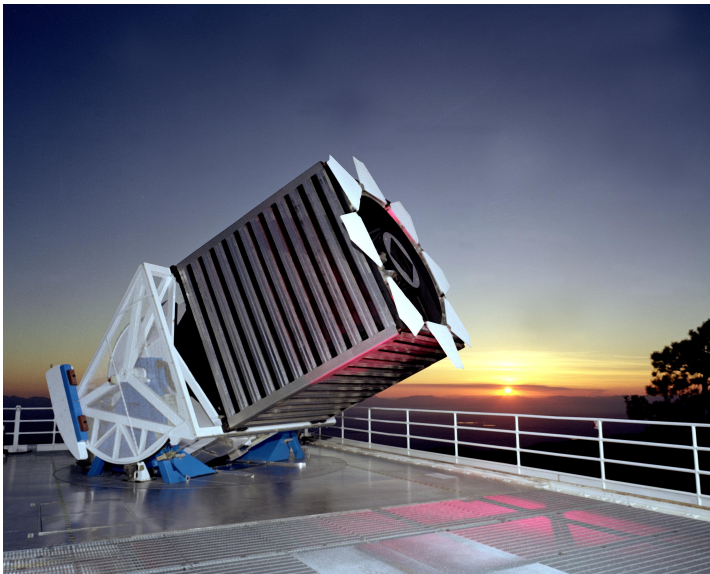
Computational science

Astronomy phenomena

Sloan Digital Sky Survey (SDSS-I/II)



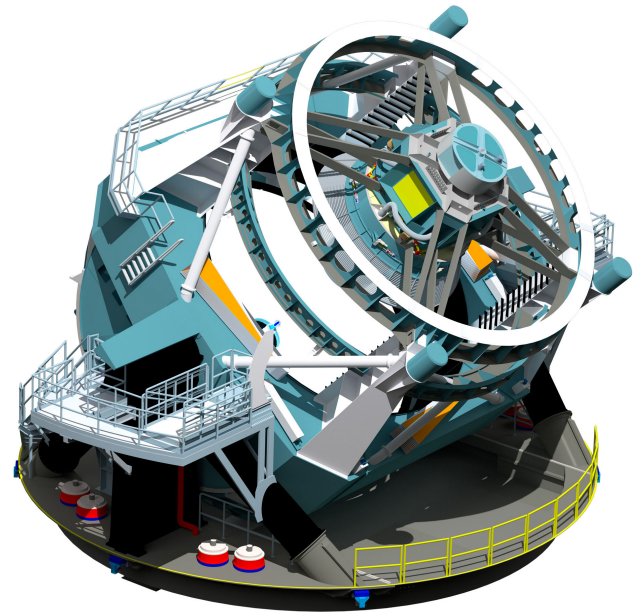
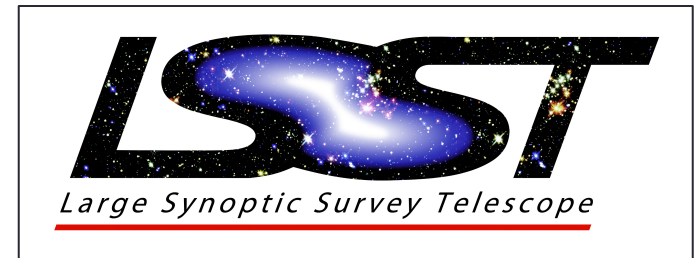
- Survey from 2000-2008
- 160+ TB data total
- Tens of millions of dollars
- Open data
- Proprietary software



Telescope for the Sloan Digital Sky Survey, Apache Point, New Mexico

Large Synoptic Survey Telescope (LSST)

- Survey from 2022-2032
- 15 TB data per night
- 1+ Billion dollars
- Data open to partners
- Open source software



Mixtures: Astronomy sky surveys

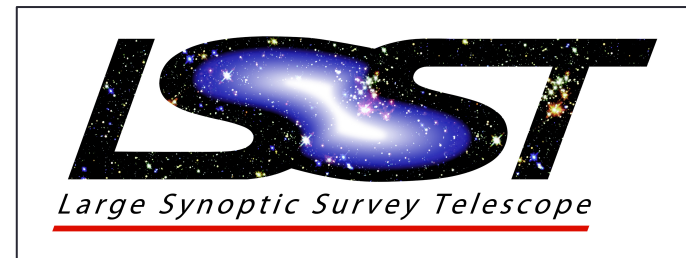
- Domains

- Astronomy, physics
- Computer science



- Project characteristics

- Mature discipline
- Abundant data
- Trusted archives
- Shared tools, methods
- Established infrastructure for data access and use



Center for Dark Energy Biosphere Investigations



Repository for seafloor cores. Photo: Peter Darch

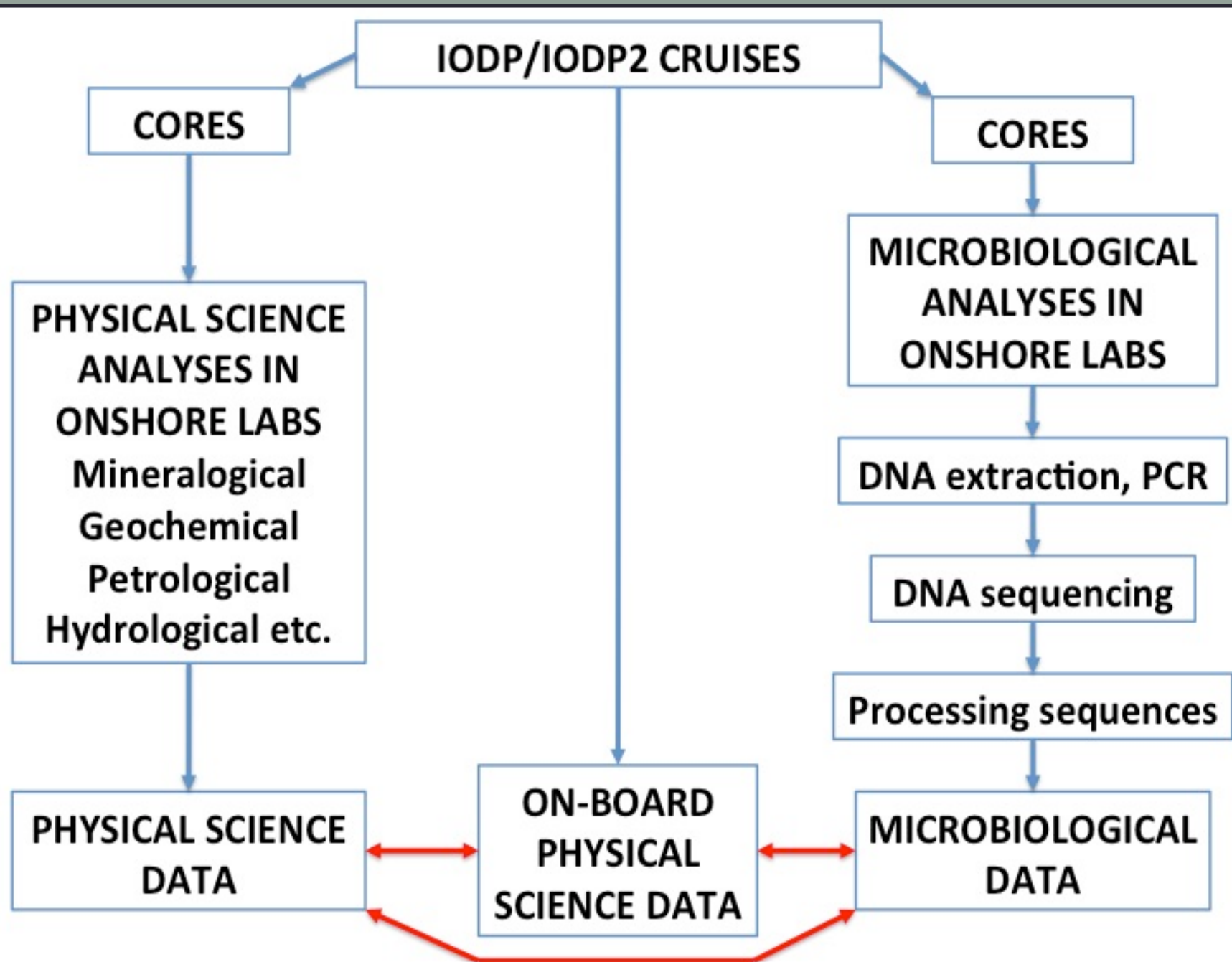


International Ocean Discovery Program

lodp.tamu.org

- NSF Science & Tech Ctr, 2010-2020
- 35 institutions
- 90 scientists
- Biological sciences
- Physical sciences





Mixtures: Deep seafloor biosphere

- Domains
 - Biological sciences
 - Physical sciences
 - 50+ self-identified specialties
- Project characteristics
 - Emergent scientific problem area
 - Scarce data
 - Disparate, exploratory methods
 - Building capacity for data collection
 - Sharing established infrastructures



Research Question 2

What *factors of scale* influence research practices, and how?

Domain

Astronomy sky surveys

Deep seafloor biosphere

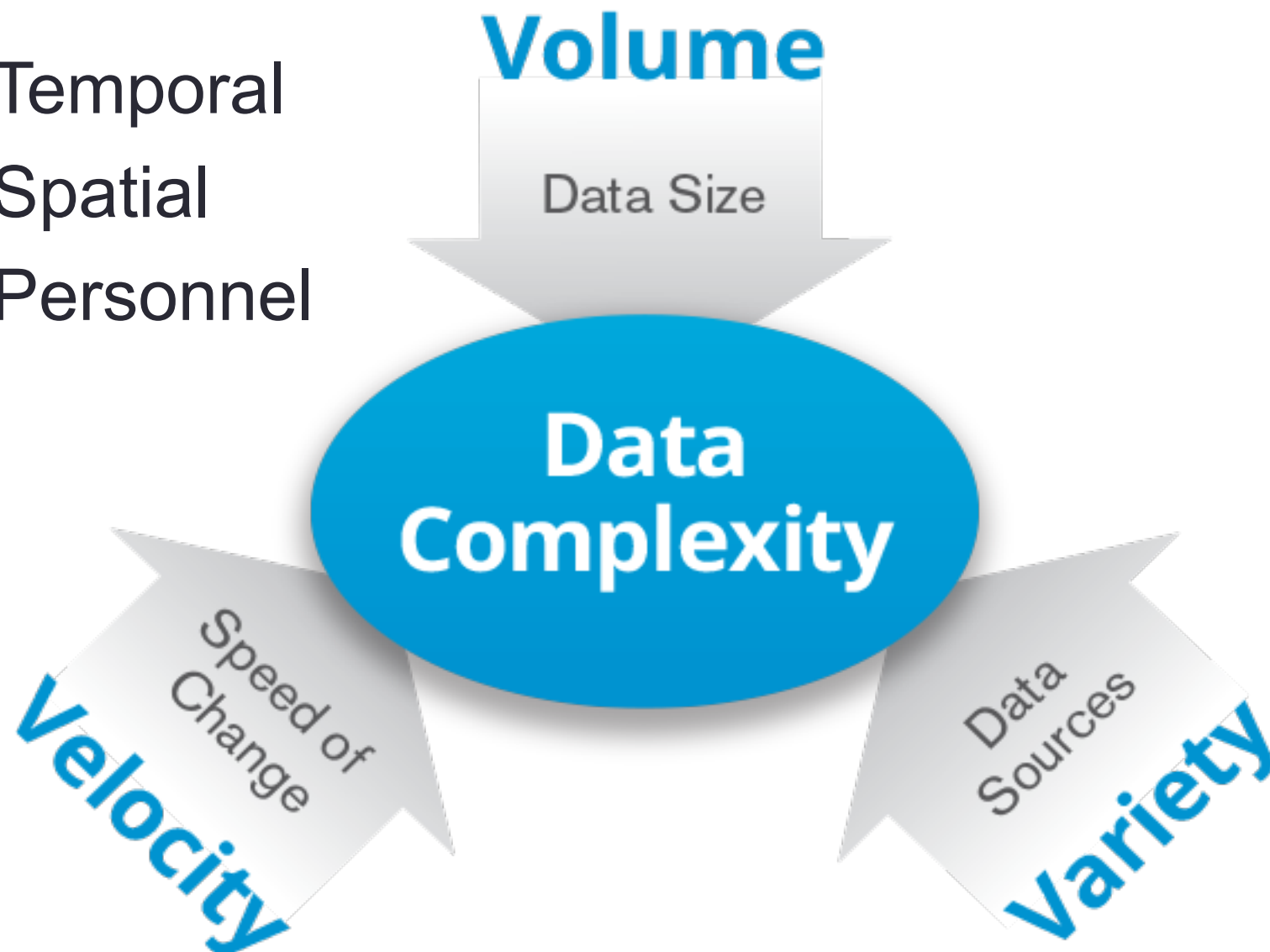
Biomedical research

Computational science

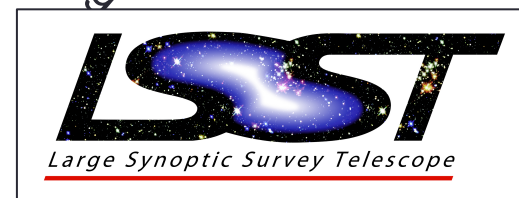
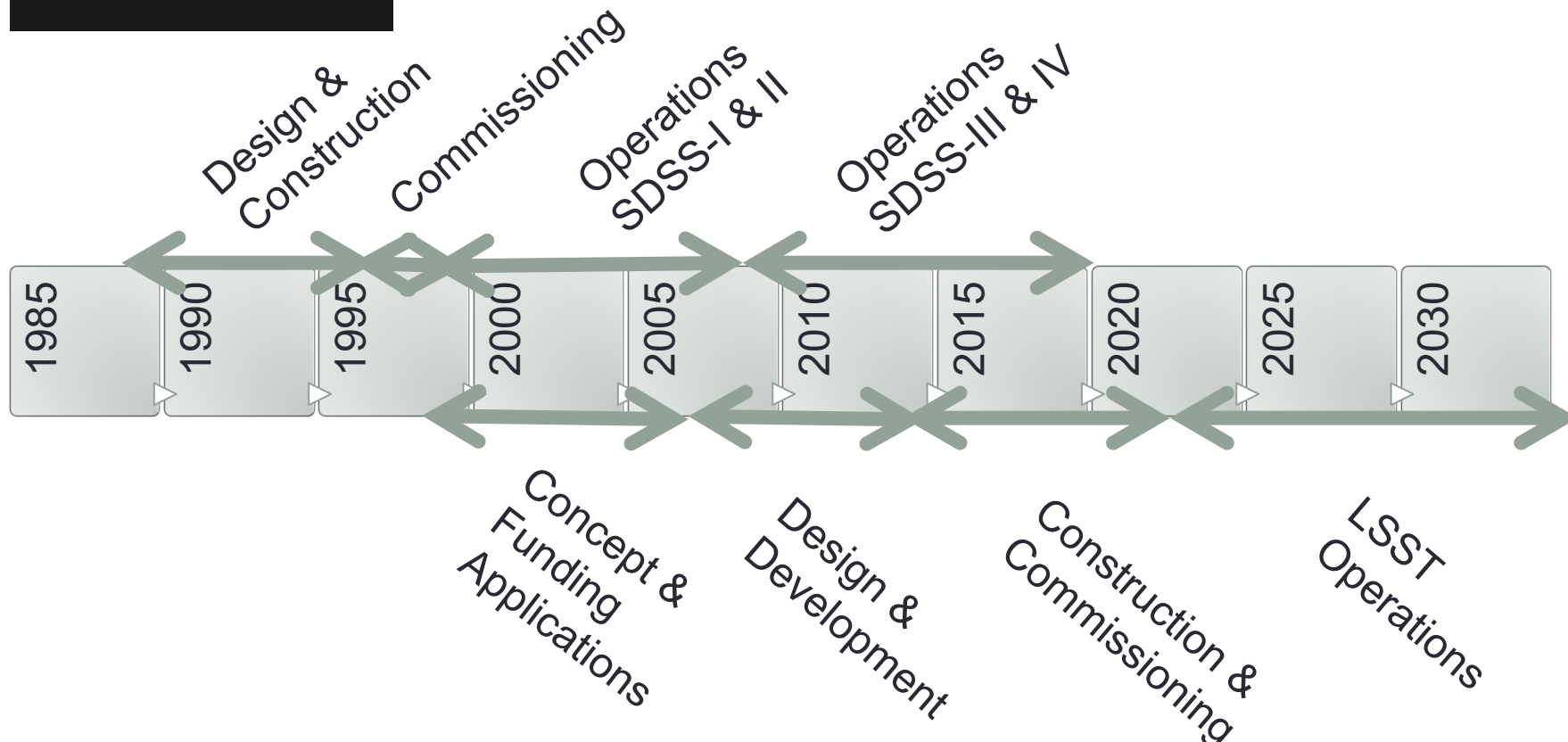
Astronomy phenomena

Scale factors

- Temporal
- Spatial
- Personnel



Project Timelines



Scale factors

Research site	Scale factors
Astronomy sky surveys	Uncertainty due to long temporal frame; paradigm shifts
Deep subseafloor biosphere	Scarce data are sparse data; high variety; difficult to standardize
Biomedical research	High variety in genomes studied, models, methods, duration of analysis; difficult to standardize
Computational sciences	High variety in data, methods, tool expertise; difficult to standardize

Research Question 3

How does the degree of *centralization of data collection and analysis* influence use, reuse, curation, and project strategy?

Domain

Astronomy sky surveys

Deep seafloor biosphere

Biomedical research

Computational science

Astronomy phenomena

Centralization factors

Research Site	Centralization factors
Astronomy sky surveys	Centralized data collection and initial processing; decentralized use and analysis
Deep subseafloor biosphere	Common data source, shared repositories of cores; decentralized analysis
Biomedical research	Decentralized data collection; efforts to integrate data for centralized analysis reveal lack of commonalities
Computational sciences	Decentralized data collection; efforts to integrate data for centralized analysis reveal lack of commonalities

REPRODUCIBILITY IN THE DEEP SUBSEAFLOOR BIOSPHERE

Peter T. Darch, School of Information Sciences, University of
Illinois at Urbana-Champaign

DataONE Webinar, May 9, 2017



ILLINOIS
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

Center for
Knowledge Infrastructures
UCLA



nature

International weekly journal of science

[Home](#)[News & Comment](#)[Research](#)[Careers & Jobs](#)[Current Issue](#)[Archive](#)[Audio & Video](#)[Archive](#)[Volume 533](#)[Issue 7604](#)[News Feature](#)[Article](#)

NATURE | NEWS FEATURE

1,500 scientists lift the lid on reproducibility

Survey sheds light on the 'crisis' rocking research.

Monya Baker

25 May 2016 | Corrected: [28 July 2016](#)

<http://www.nature.com/news/1-500-scientists-lift-the-lid-on-reproducibility-1.19970>

Reproducibility

- Reproducing an analysis requires access to:
 - Data
 - Methods
 - Source code
 - Workflows
- Access means:
 - Availability
 - Usability
 - Interpretability



Deep Subseafloor Biosphere

- Microbial communities in the seafloor
- Highly-multidisciplinary
- Center for Dark Energy Biosphere Investigations (C-DEBI)
 - 10-year NSF Science and Technology Center
- International Ocean Discovery Program (IODP)



<http://iodp.org/expeditions>



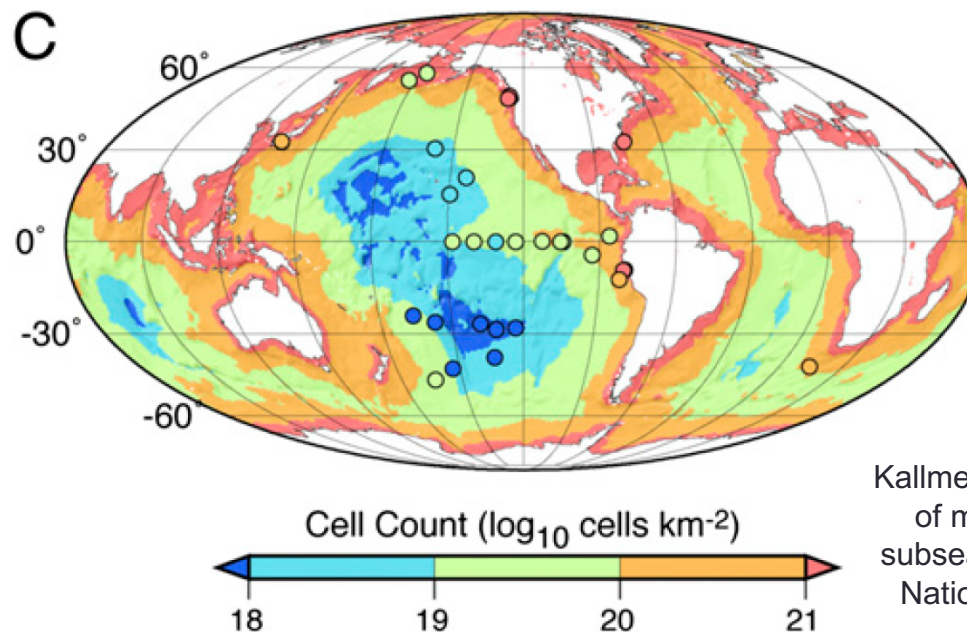
Subseafloor Biosphere: Data-Scarce Domain

- “Data scarce” vs. “data abundant”
- Objectives of domain scientists
 - Address current scientific debates
 - Transition from “discovery-driven” to “hypothesis-driven” science
- Access to data is limited
 - Domain’s relative newness
 - IODP resources are shared with other domains



Benefits of Data Reuse

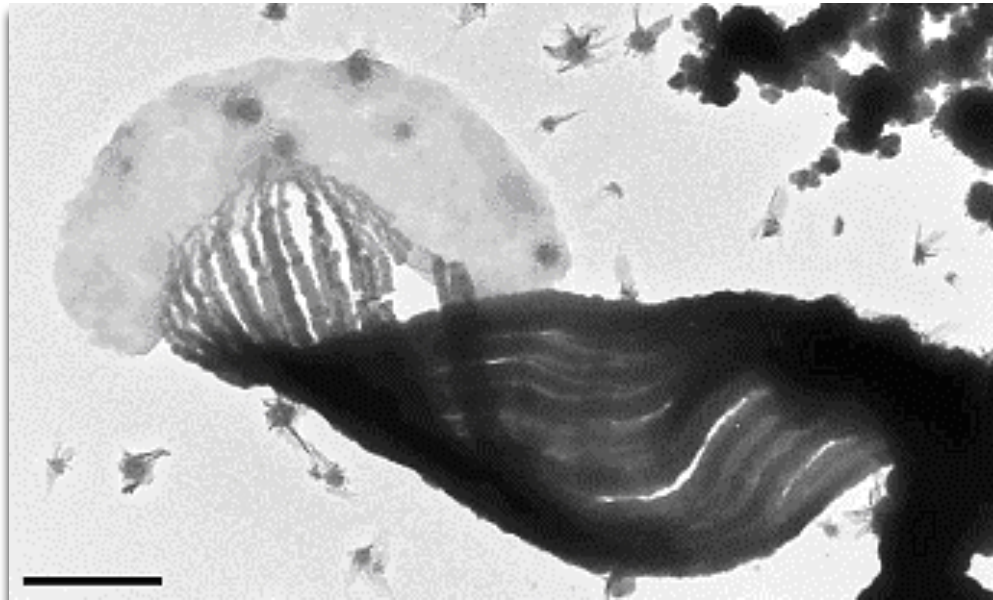
- Improve access to data for researchers
- Build better reference collections for multiple domains
- Answer key questions in microbiology
 - Baas-Becking hypothesis
 - Global distribution of microbes



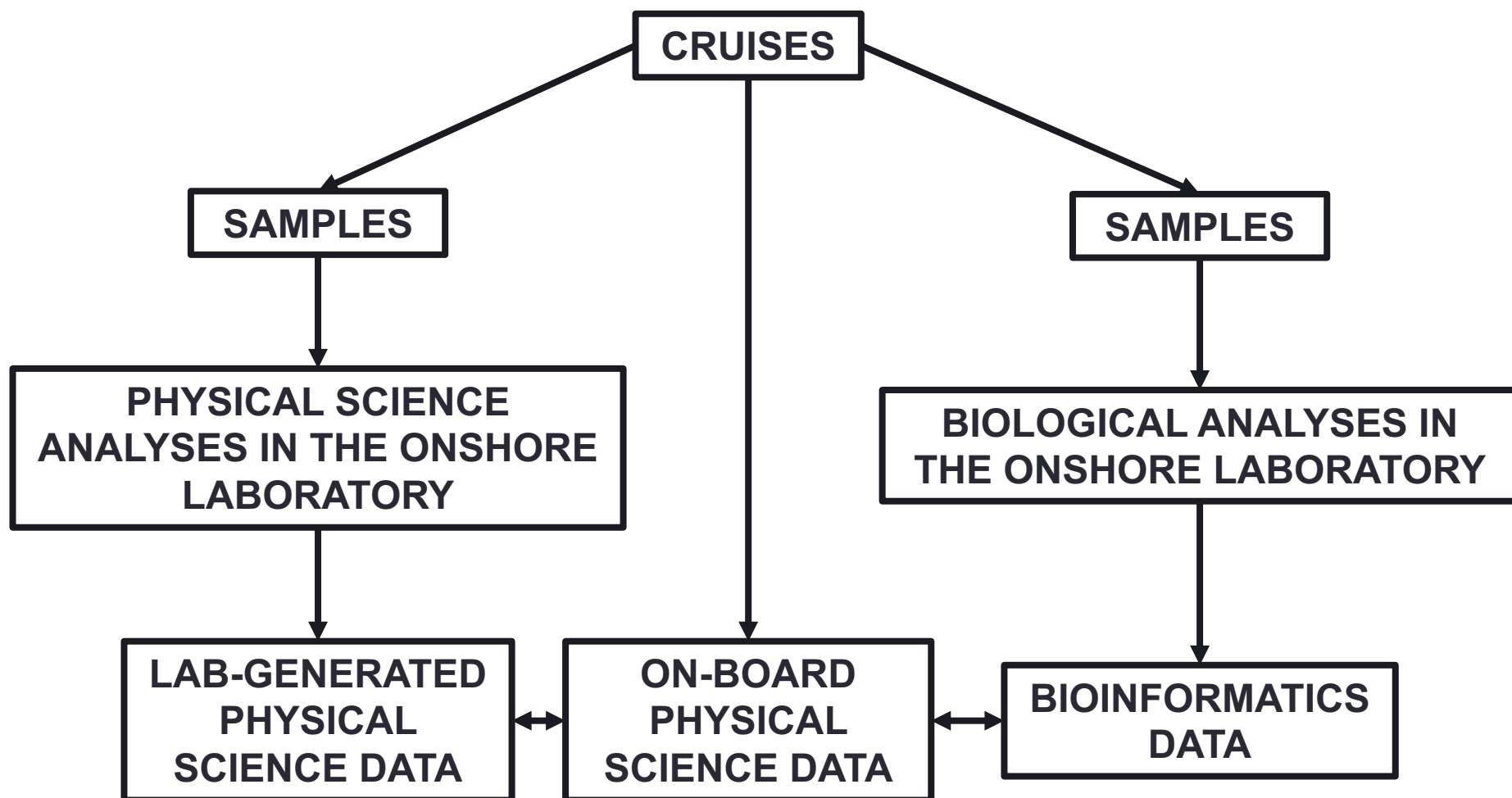
Kallmeyer et al. (2012). Global distribution of microbial abundance and biomass in subseafloor sediment. *Proceedings of the National Academy of Sciences*, 109(40), 16213–16216.

Reproducibility vs. Reuse

- Reuse is more effective strategy in data-scarce domains
- Heterogeneous data types complicate reproducibility
- Sharing for reuse affects researchers' relationships in a different way to sharing for reproducibility



Data Diverge During Scientific Work



Reproducibility when Data Diverge

- Reproducibility requires access to:
 - Bioinformatics data
 - Physical science data



PANGAEA.

Data Publisher for Earth & Environmental Science

- Different types of data can be:
 - Subject to different policies for curation
 - Deposited in different databases



GREENGENES
The 16S rRNA Gene Database and Tools

- These differences inhibit access and integration of data

Goals for Sharing and Reusing Data

- Nurturing personal relationships is critical for the domain
 - Domain is in the early stages of establishing itself
 - Domain is relatively small
 - Domain is highly-distributed
- Exchanges of data and software can affect relationships

Sharing for reuse	Sharing for reproducibility
Links researchers together	Links researchers together
Allows researchers to display good faith in each other	Can imply mistrust in competence or good faith of other researchers
Reinforces positive collaborative relationships	Uncertain effect on collaborative relationships

Implications for Data Scarce Environments

- Data-scarce domains experience good pay-offs from data reuse
- Barriers to reproducibility emerge early in the scientific process
- A focus on reproducibility may obscure data reuse opportunities
- **Reproducibility goals may inhibit scientific progress in data-scarce domains**



Acknowledgements

- C-DEBI and IODP personnel who participated in our research
- Sloan Foundation (Awards #20113194, #201514001)
- Other members of the Center for Knowledge Infrastructures at UCLA: Milena Golshan, Irene Pasquetto, Bernie Randles; past members: Ashley Sands, Sharon Traweek

<http://knowledgeinfrastructures.gseis.ucla.edu>

And thank you for listening