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

UCLA Center for Knowledge Infrastructures

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

THE WHITE HOUSE
WASHINGTON

JST JAPAN SCIENCE AND TECHNOLOGY AGENCY

CHINESE ACADEMY OF SCIENCES

中央研究院
ACADEMIA SINICA 1928

Supported by wellcome trust

Australian Government
National Health and Medical Research Council

NSF National Science Foundation
WHERE DISCOVERIES BEGIN

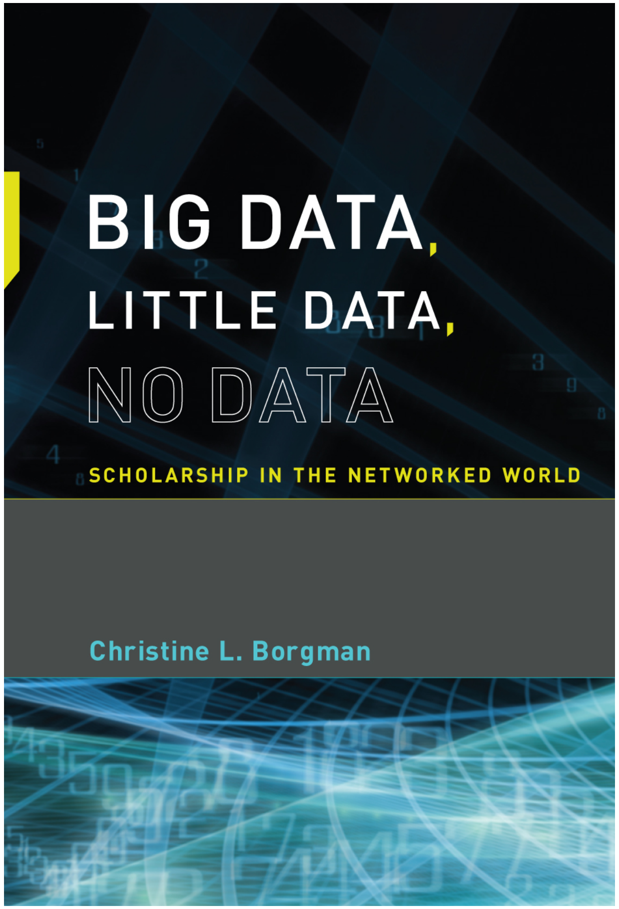
RDA RESEARCH DATA ALLIANCE

NIH National Institutes of Health
Turning Discovery Into Health

E·S·R·C
ECONOMIC & SOCIAL RESEARCH COUNCIL

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



BIG DATA, LITTLE DATA, NO DATA

SCHOLARSHIP IN THE NETWORKED WORLD

Christine L. Borgman

MIT Press, 2015

Lack of incentives to share data

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

http://www.buildingsrus.co.uk/.../ target1.htm

Why Reuse Research Data?

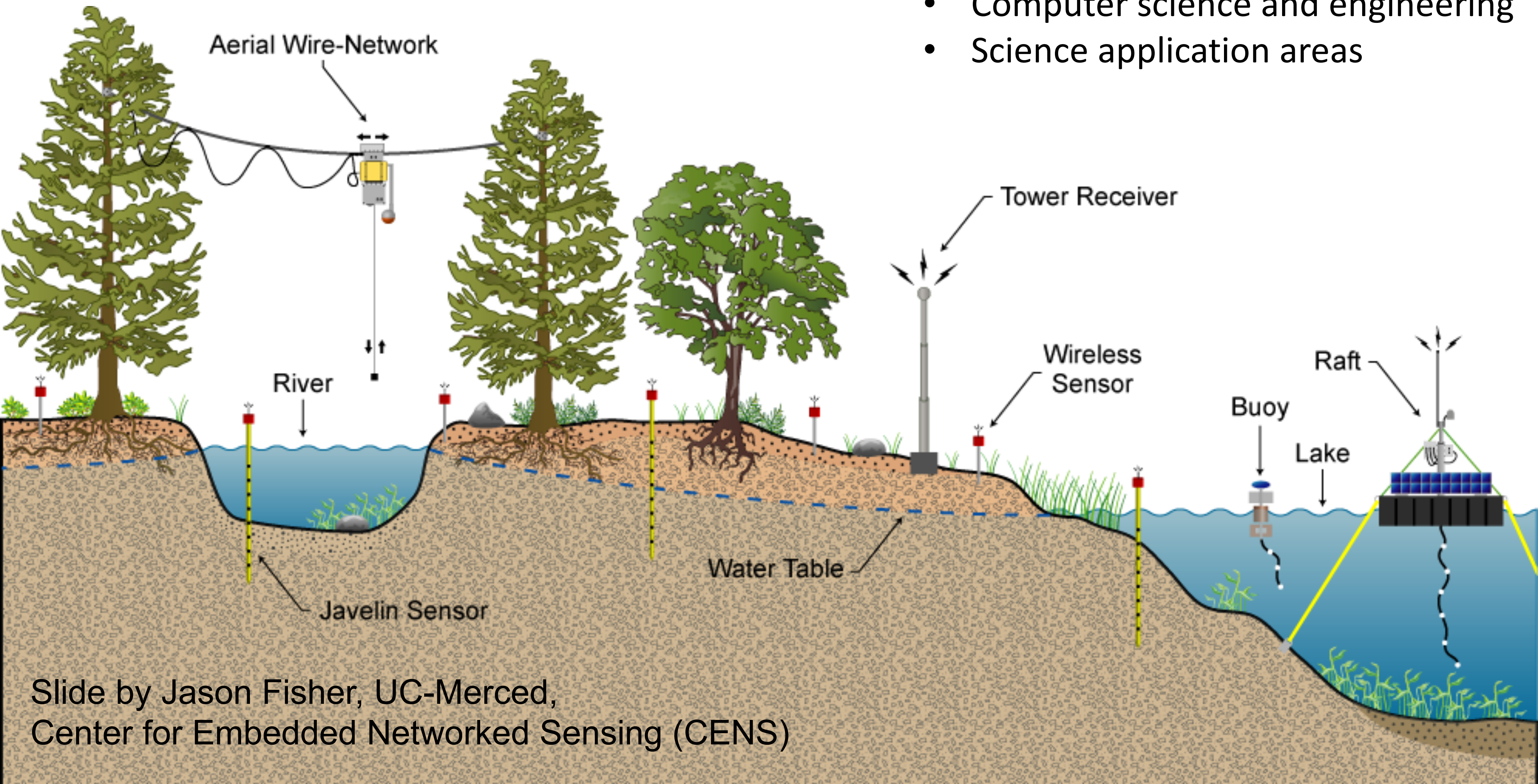
- To reproduce research
- To replicate research
- To verify or validate research
- To integrate with other data

https://www.flickr.com/photos/pagedooley/

Data

Center for Embedded Networked Sensing

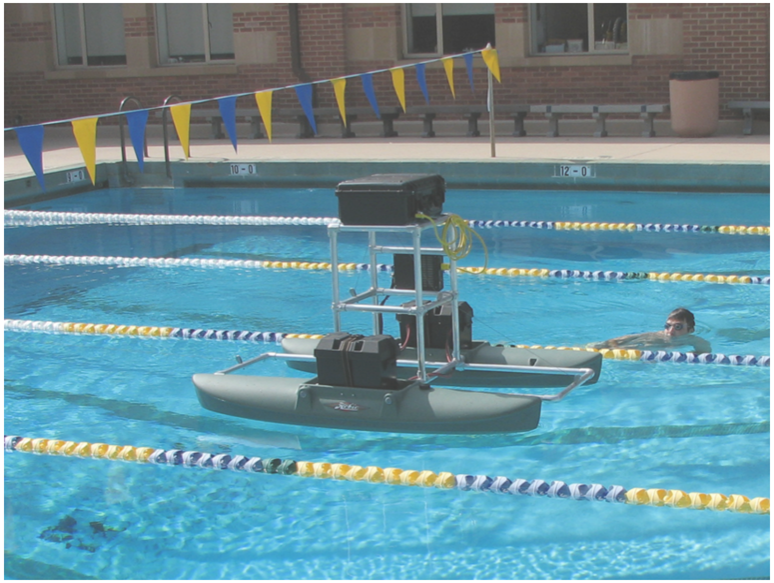
- NSF Science & Tech Ctr, 2002-2012
- 5 universities, plus partners
- 300 members
- Computer science and engineering
- Science application areas



Slide by Jason Fisher, UC-Merced,
Center for Embedded Networked Sensing (CENS)

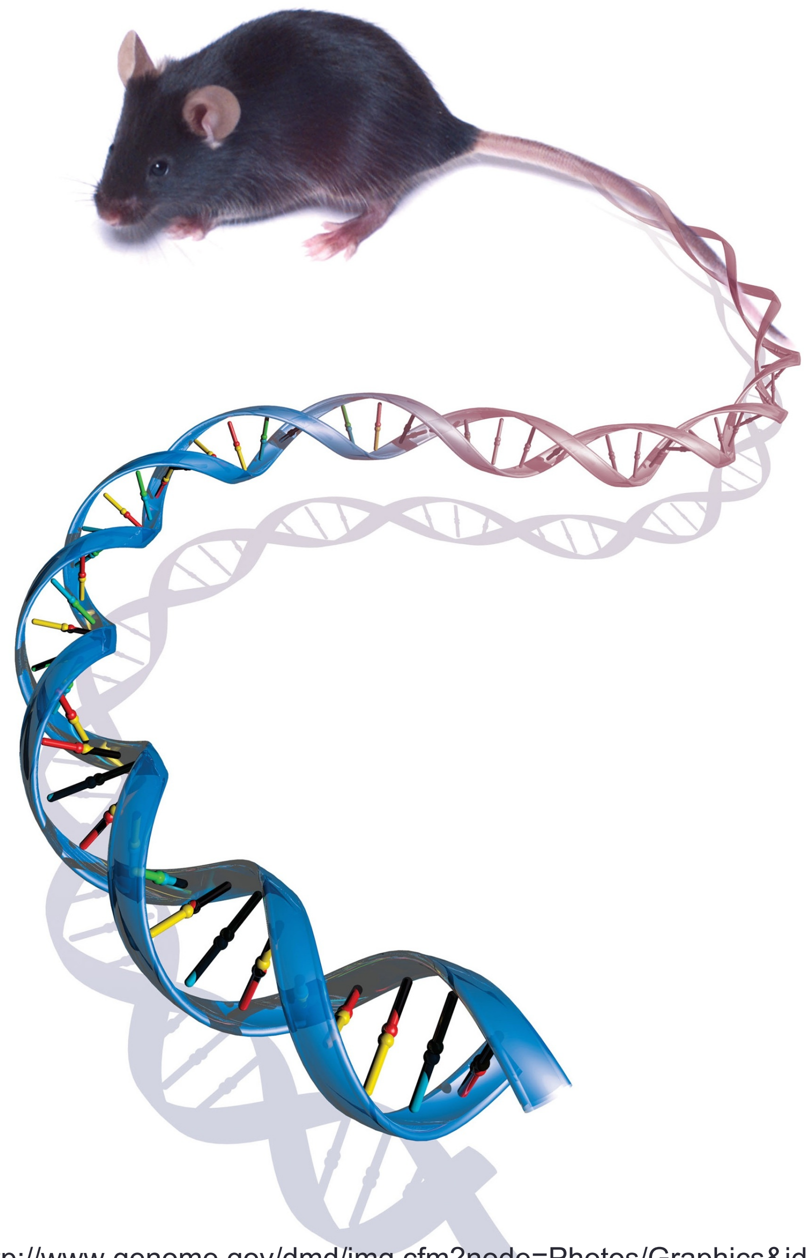
Documenting Data for Interpretation

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.."*



http://www.genome.gov/dmd/img.cfm?node=Photos/Graphics&id=85327

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

UCLA Center for Knowledge Infrastructures

Current Research Sites

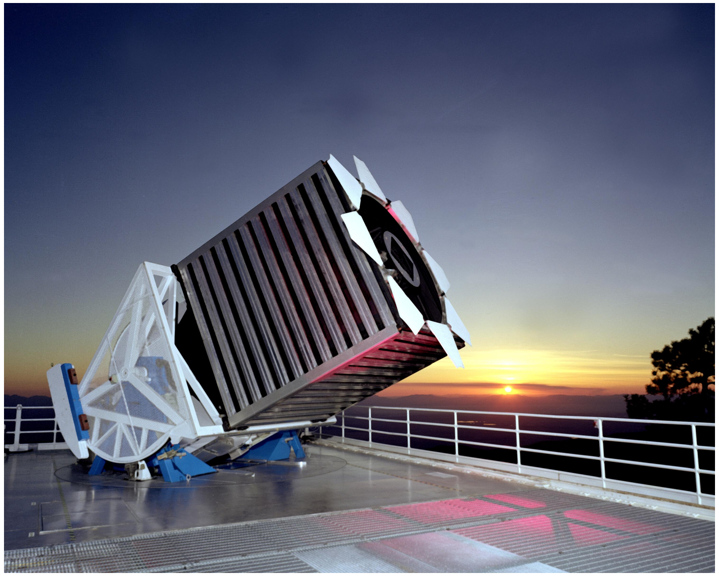
Domain	Focus	Topic
Astronomy sky surveys	Place: sky and universe	Survey of night sky
Deep subseafloor biosphere	Place: under ocean floor	Microbial life and environment
Biomedical collaboration	Problem: data sharing and reuse in an interdisciplinary context	Genomics of four model organisms
Computational 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 subseafloor biosphere
Biomedical research
Computational science
Astronomy phenomena

Sloan Digital Sky Survey (SDSS-I/II)



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

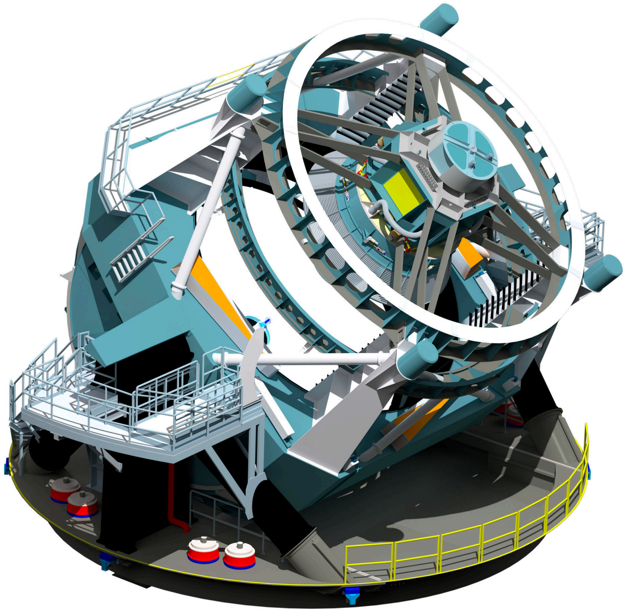
- Survey from 2000-2008
- 160+ TB data total
- Tens of millions of dollars
- Open data
- Proprietary software

Large Synoptic Survey Telescope (LSST)

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

LSST

Large Synoptic Survey Telescope



https://news.slac.stanford.edu/sites/default/files/images/image/lsst_h_0.jpg

LSST telescope, Chile

Mixtures: Astronomy sky surveys

- Domains
  - Astronomy, physics
  - Computer science

SDSS

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

LSST

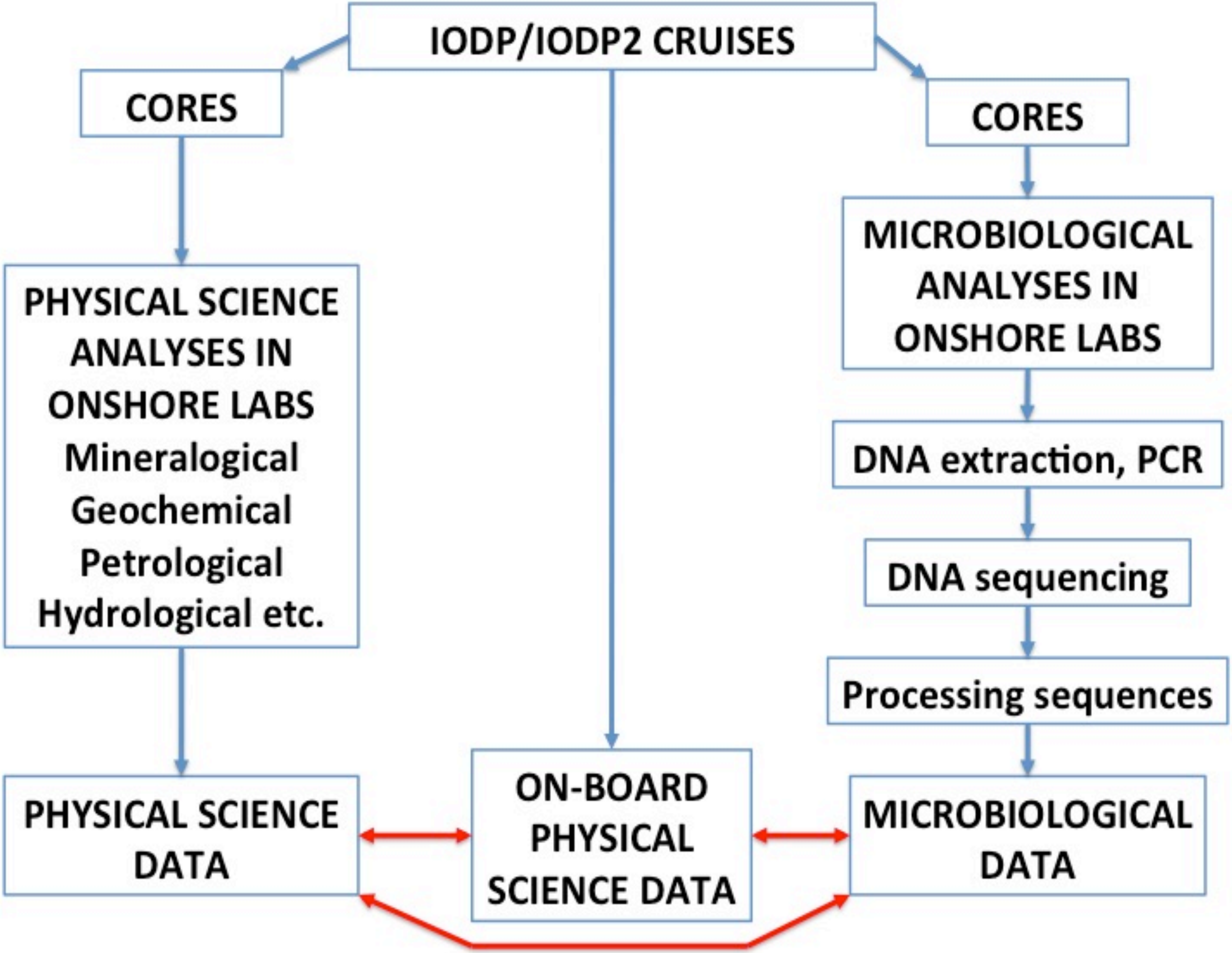
Large Synoptic Survey Telescope

Center for Dark Energy Biosphere Investigations

International Ocean Discovery Program
Iodp.tamu.org

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

Repository for seafloor cores. Photo: Peter Darch



IODP/IODP2 CRUISES

CORES

PHYSICAL SCIENCE ANALYSES IN ONSHORE LABS
Mineralogical
Geochemical
Petrological
Hydrological etc.

PHYSICAL SCIENCE DATA

ON-BOARD PHYSICAL SCIENCE DATA

CORES

MICROBIOLOGICAL ANALYSES IN ONSHORE LABS

DNA extraction, PCR

DNA sequencing

Processing sequences

MICROBIOLOGICAL DATA

Mixtures: Deep subseafloor 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

C-DEBI

Research Question 2

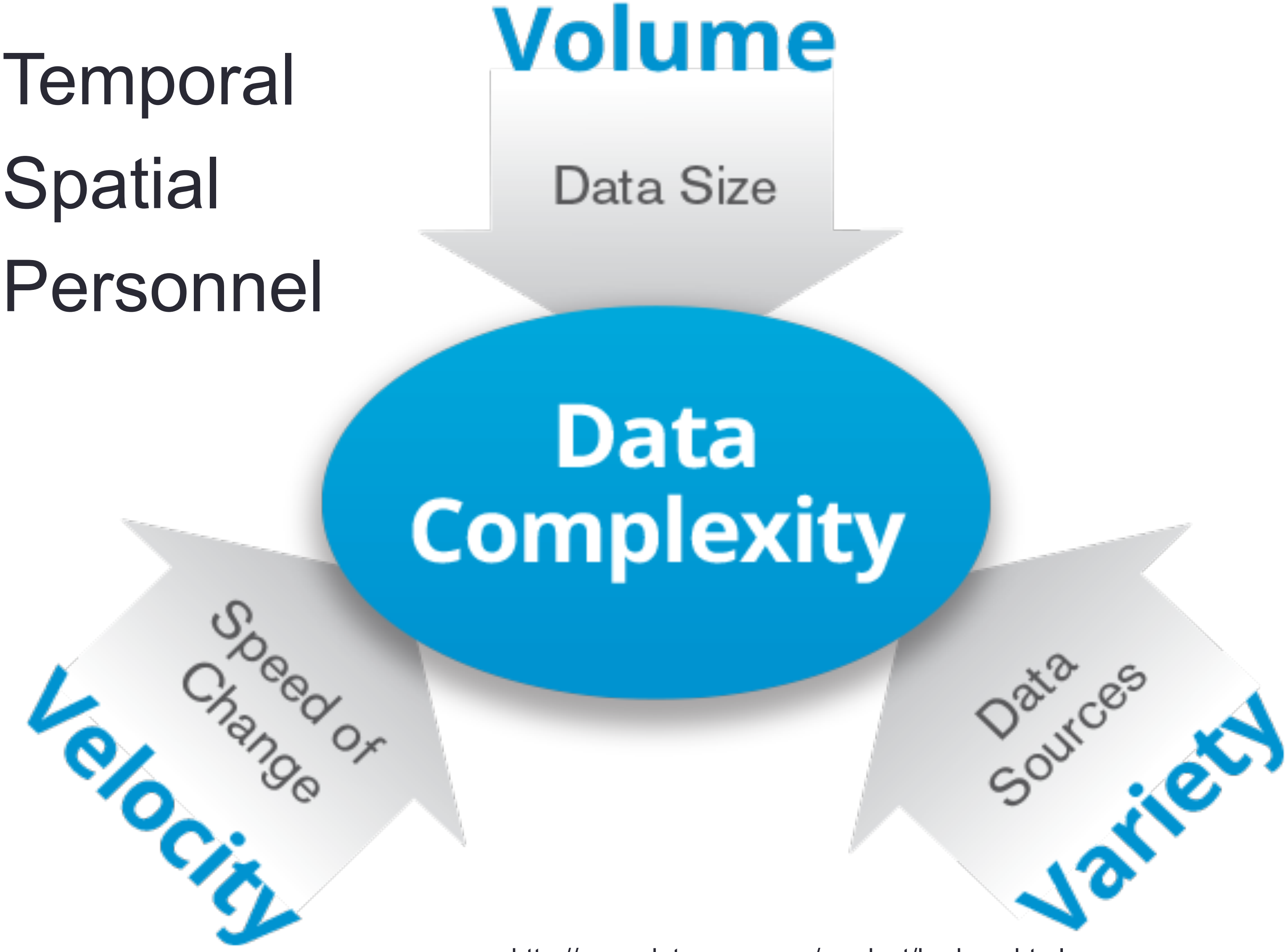
What *factors of scale* influence research practices, and how?

Domain
Astronomy sky surveys
Deep subseafloor biosphere
Biomedical research
Computational science
Astronomy phenomena

UCLA Center for Knowledge Infrastructures

Scale factors

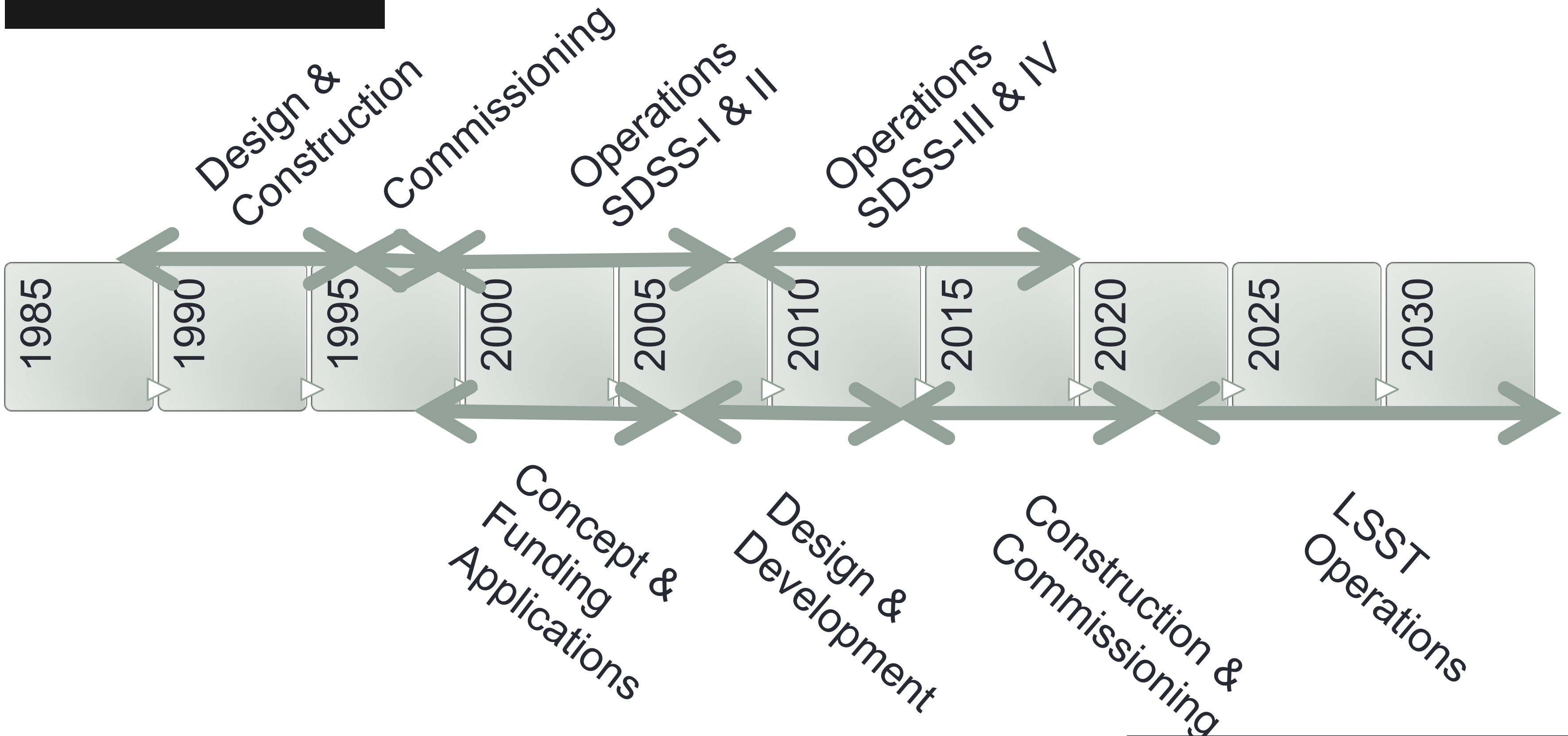
- Temporal
- Spatial
- Personnel



http://www.datameer.com/product/hadoop.html

Project Timelines

SDSS



Design & Construction

Commissioning

Operations SDSS-I & II

Operations SDSS-III & IV

1985

1990

1995

2000

2005

2010

2015

2020

2025

2030

Concept & Funding Applications

Design & Development

Construction & Commissioning

LSST Operations

Large Synoptic Survey Telescope

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 subseafloor biosphere
Biomedical research
Computational science
Astronomy phenomena

UCLA Center for Knowledge Infrastructures

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

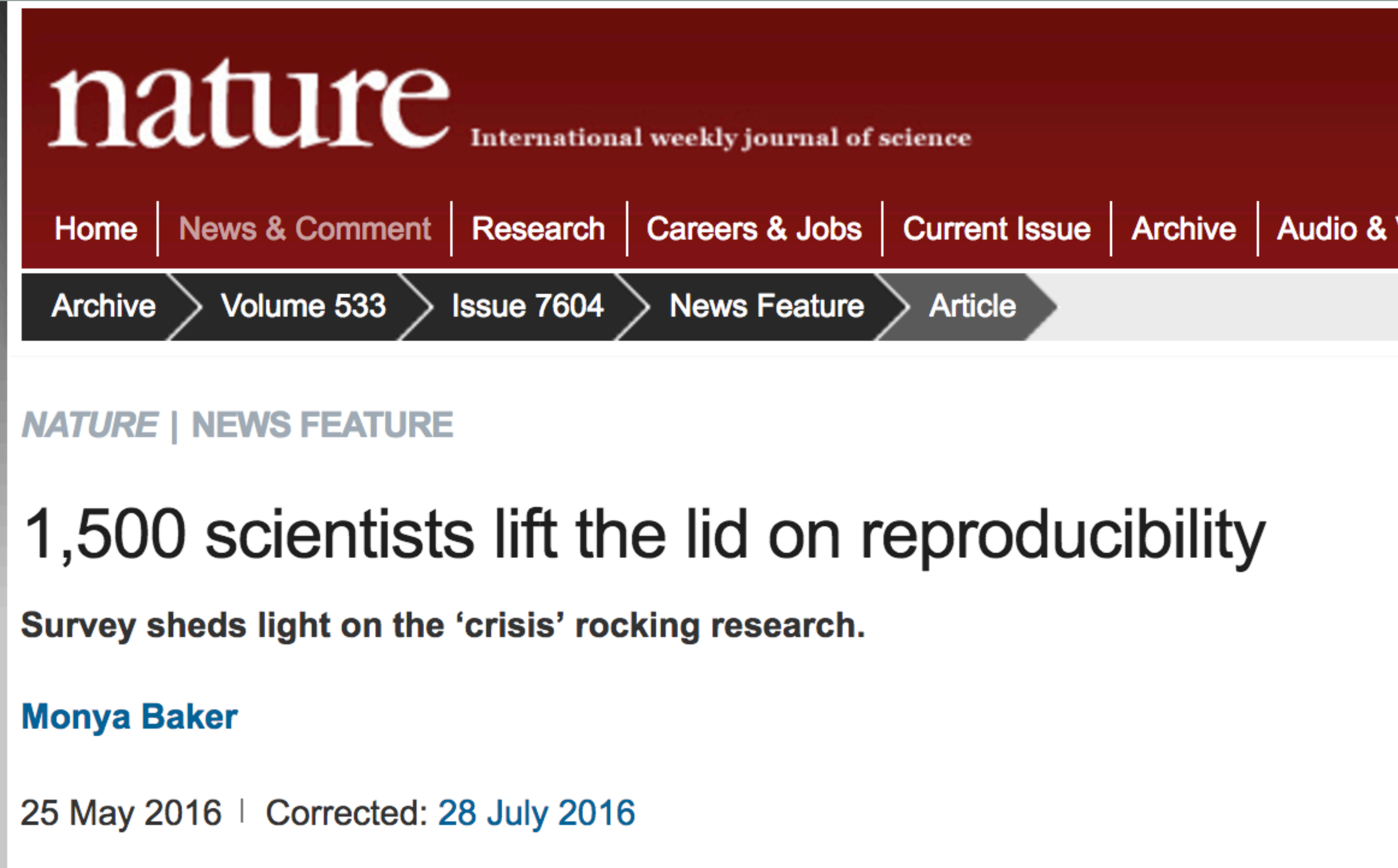
Peter T. Darch, School of Information Sciences, University of Illinois at Urbana-Champaign

DataONE Webinar, May 9, 2017

1867
ILLINOIS
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

Center for Knowledge Infrastructures
UCLA

ALFRED P. SLOAN FOUNDATION
S
1934



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

http://www.statisticsviews.com/common/images/thumbnails/source/14f401c0a35.jpg

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

C-DEBI

Subseafloor Biosphere: Data-Scarce Domain

- “Data scarce” vs. “data abundant”

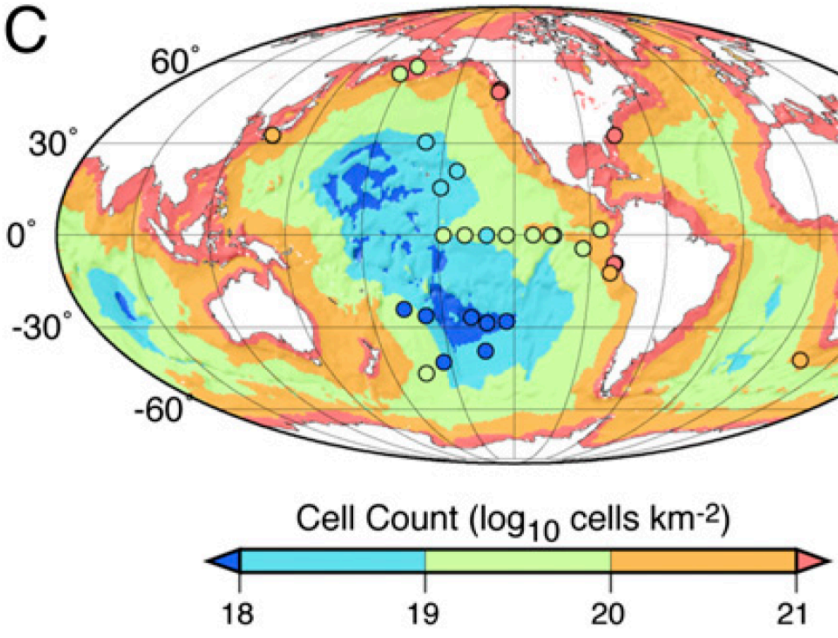
- Objectives of domain scientists
  - Address current scientific debates
  - Transition from “discovery-driven” to “hypothesis-driven” science

- Access to data is limited
  - Domain’s relative newness
  - IODP resources are shared with other domains

Personal photograph

Benefits of Data Reuse

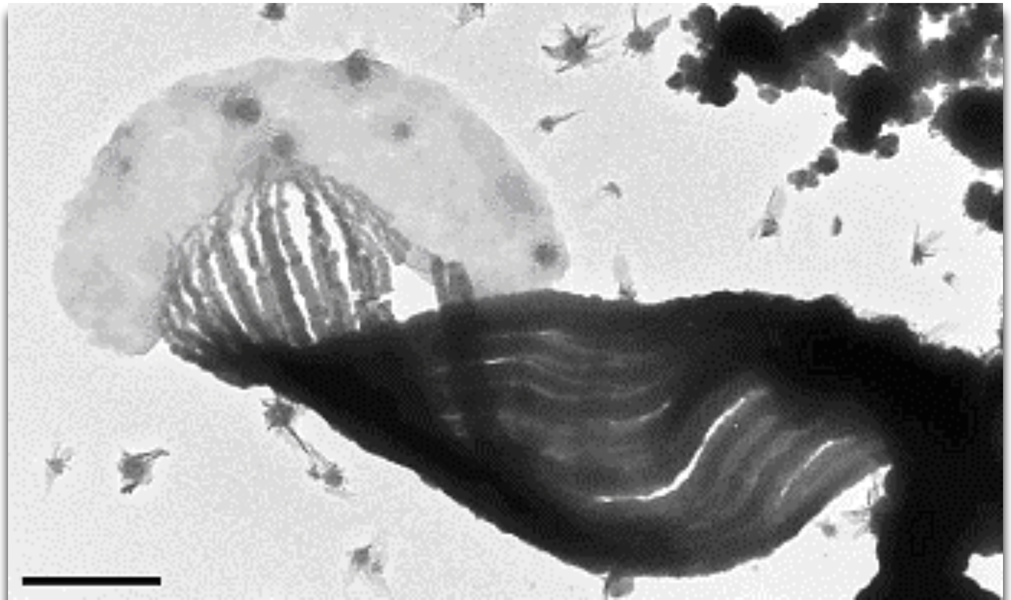
- Improve access to data for researchers
- Build better reference collections for multiple domains
- Answer key questions in microbiology
  - Baas-Becking hypothesis
  - Global distribution of microbes



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.

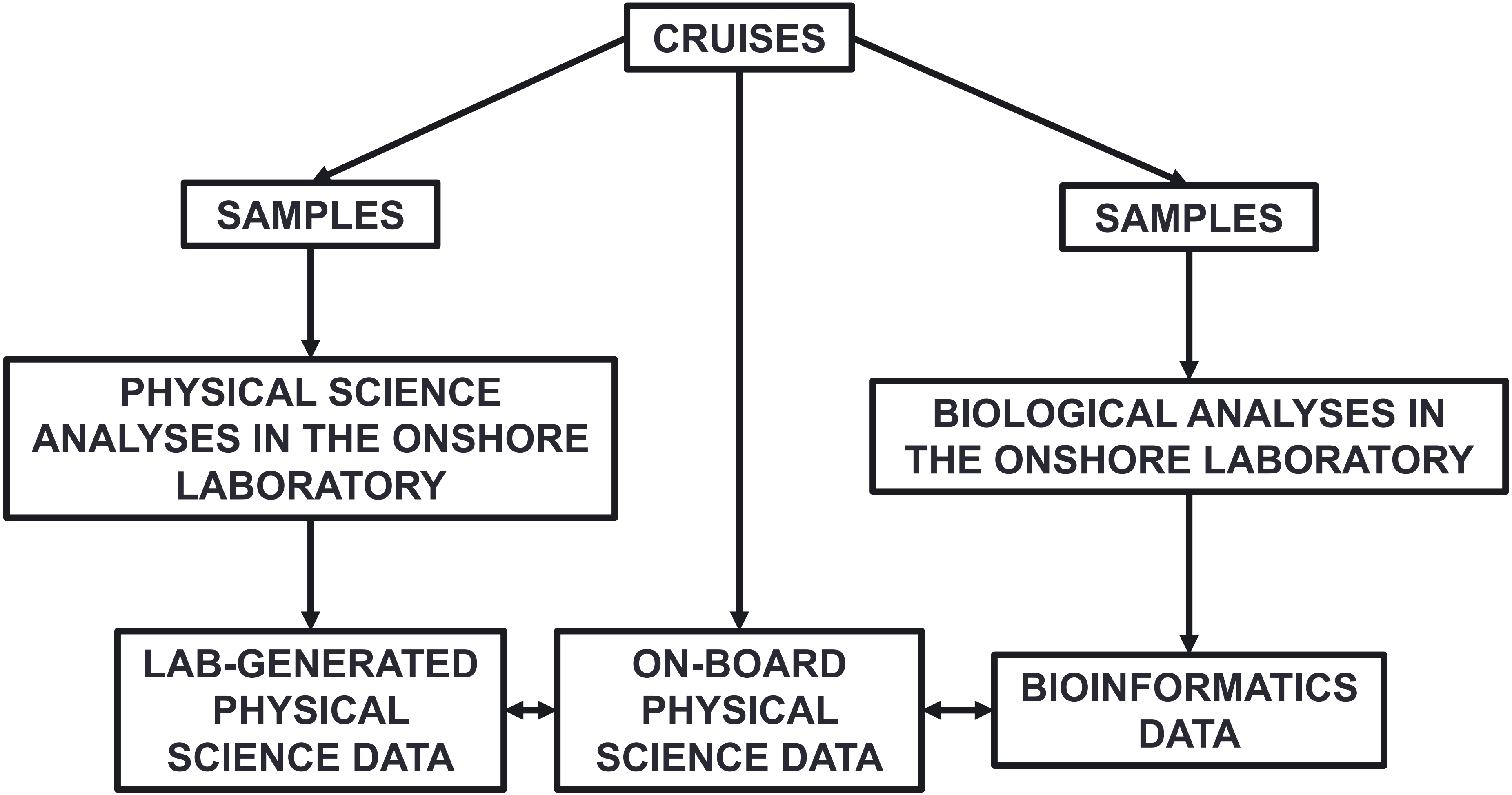
Reproducibility vs. Reuse

- 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



sites.google.com/site/adoptamicrobe/home

Data Diverge During Scientific Work



CRUISES

SAMPLES

SAMPLES

PHYSICAL SCIENCE ANALYSES IN THE ONSHORE LABORATORY

BIOLOGICAL ANALYSES IN THE ONSHORE LABORATORY

LAB-GENERATED PHYSICAL SCIENCE DATA

ON-BOARD PHYSICAL SCIENCE DATA

BIOINFORMATICS DATA

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**

http://iodp.org/expeditions

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