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

Christine L. Borgman
Distinguished Professor & Presidential Chair in Information Studies
Director, Center for Knowledge Infrastructures
https://knowledgeinfrastructures.gseis.ucla.edu
University of California, Los Angeles
http://christineborgman.info
@scitechprof


Christine Borgman


Peter Darch


Irene Pasquetto


Bernie Boscoe


Michael Scroggins


Milena Golshan

Information Access Seminar
UC Berkeley iSchool
17 November 2017

UCLA Center for Knowledge Infrastructures

Overview

- Data sharing policy drivers
- Project Design, 2015-2019
- Methods
- Questions
- Findings
- Comparisons, late 2016
- New themes, late 2017

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 wellcome trust

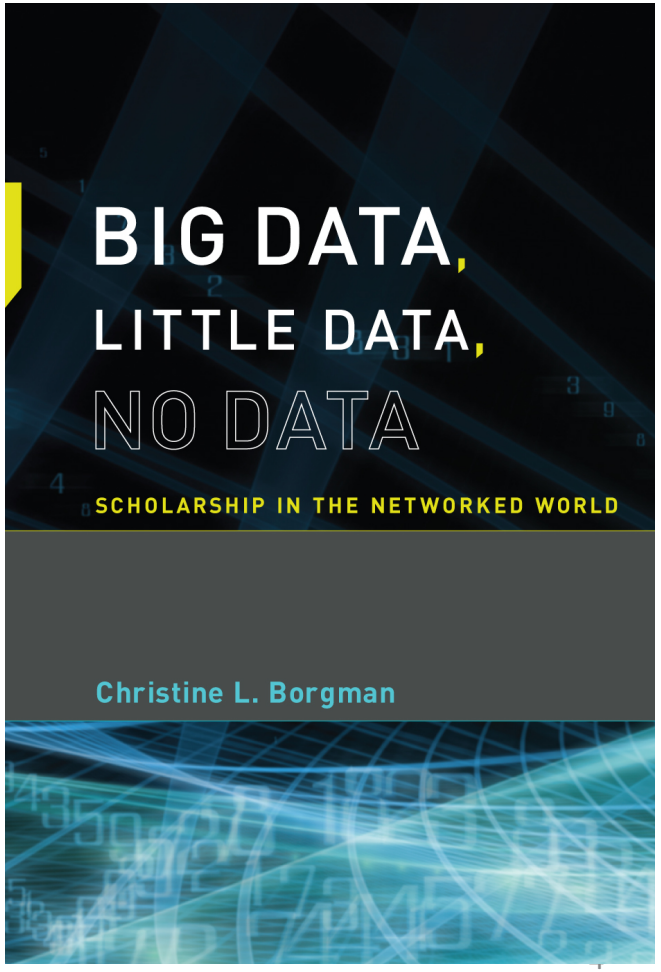
National Science Foundation
WHERE DISCOVERIES BEGIN

National Institutes of Health
Turning Discovery Into Health

Policy RECommendations for Open Access to Research Data in Europe

Why Share Research Data?

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



MIT Press, 2015

Lack of incentives to share data

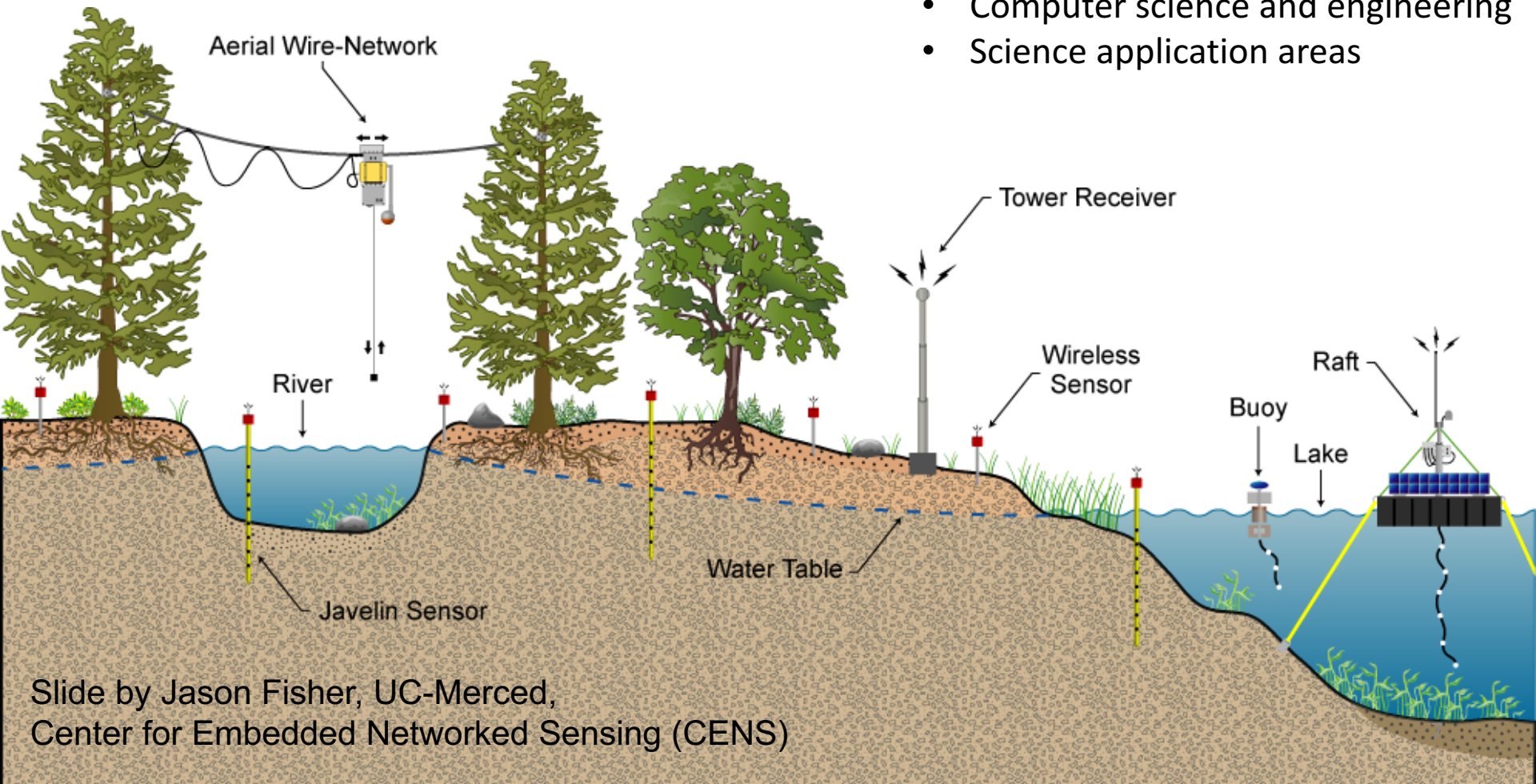
- Rewards for publication
- Effort to document data
- Competition, priority
- Control, ownership

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


Data

Center for Embedded Networked Sensing

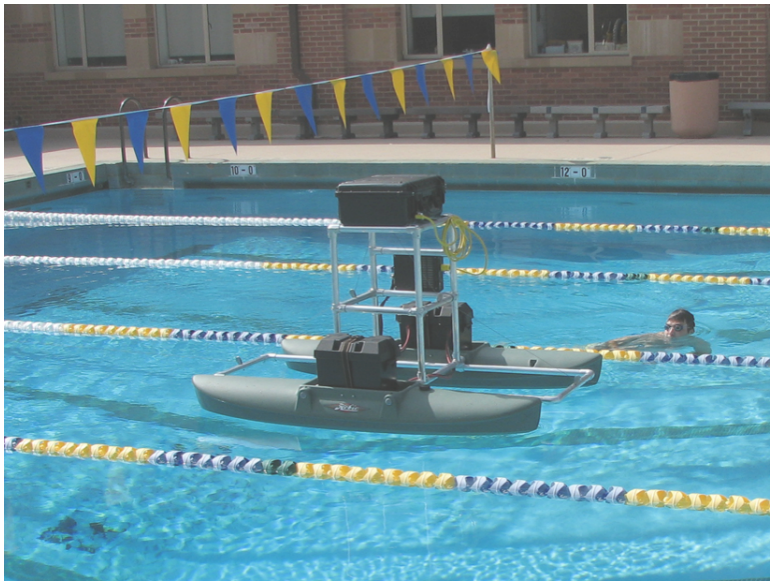
- NSF Science & Tech Ctr, 2002-2012
- 5 universities, plus partners
- 300 members
- Computer science and engineering
- Science application areas



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.."*

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

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

Research Design

- 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

UCLA Center for Knowledge Infrastructures

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
Craniofacial research	Problem: Craniofacial syndromes	Genomics of four model organisms
Computational science	Problem: Data analysis at scale	Computing platform for sciences
Astrophysics phenomena	Problem: Behavior of an object over time	Super massive black hole

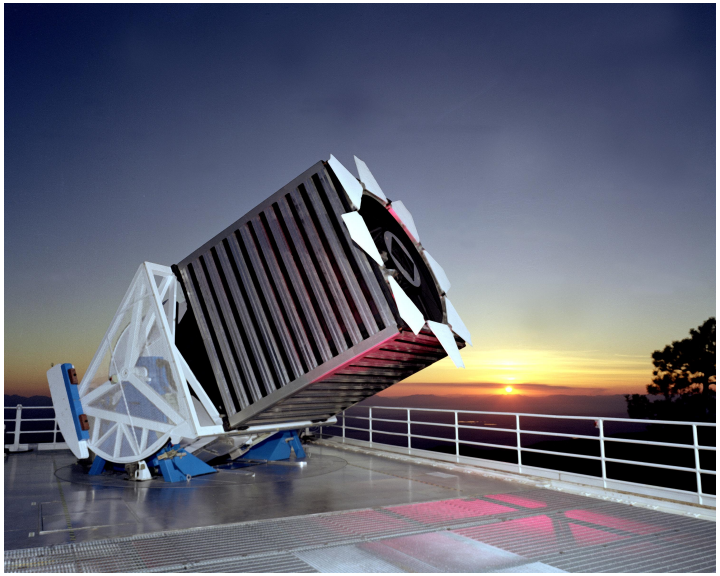
Research Question 1

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
Craniofacial research
Computational science
Astrophysics phenomena

UCLA Center for Knowledge Infrastructures

Sloan Digital Sky Survey (SDSS-I/II)

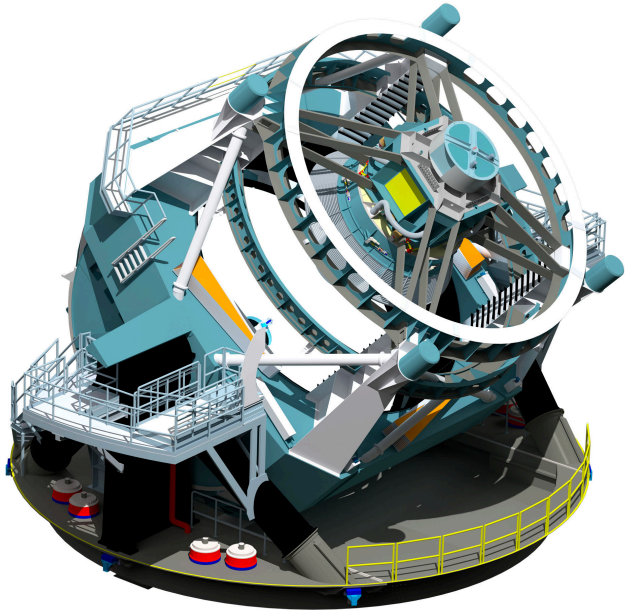


- Survey from 2000-2008
- 160+ TB data total
- Tens of millions of dollars
- Open data
- Proprietary software

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

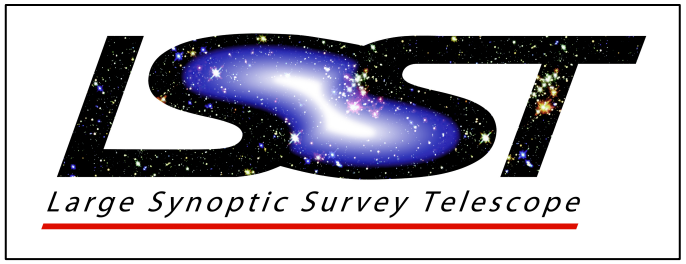


https://news.slac.stanford.edu/sites/default/files/images/image/lsst_h_0.jpg

LSST telescope, Chile

Mixtures: Astronomy sky surveys

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

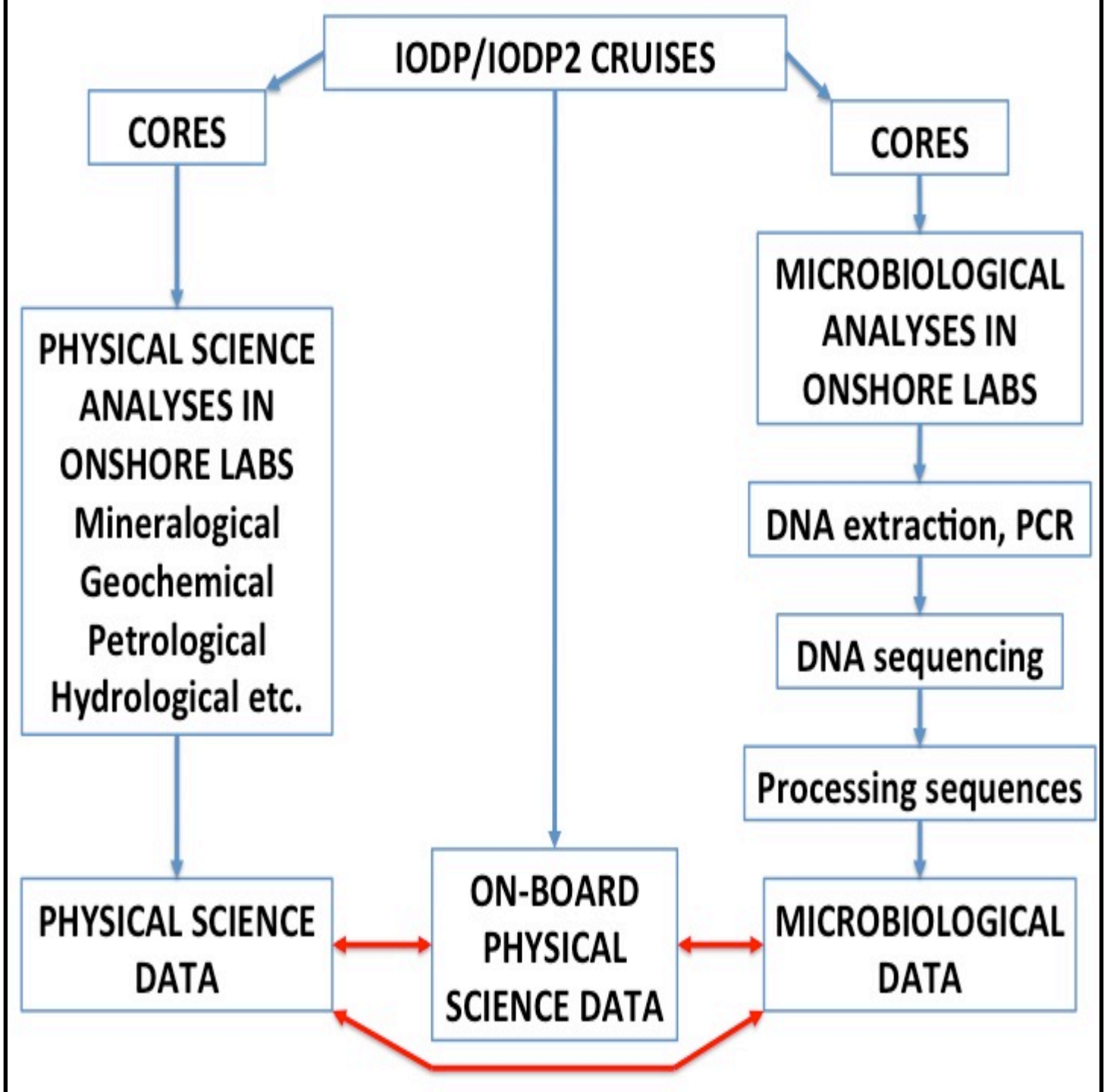


Center for Dark Energy Biosphere Investigations

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

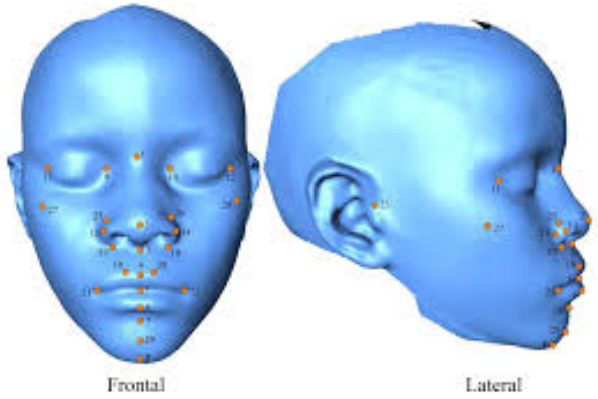

IODP/IODP2 CRUISES
CORES
CORES
PHYSICAL SCIENCE ANALYSES IN ONSHORE LABS Mineralogical Geochemical Petrological Hydrological etc.
MICROBIOLOGICAL ANALYSES IN ONSHORE LABS
DNA extraction, PCR
DNA sequencing
Processing sequences
PHYSICAL SCIENCE DATA
ON-BOARD PHYSICAL SCIENCE DATA
MICROBIOLOGICAL DATA

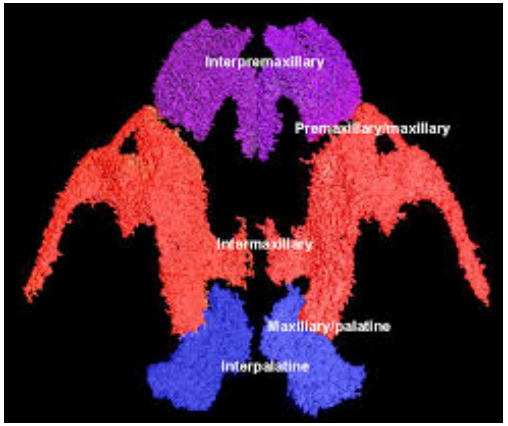
Mixtures: Deep subseafloor biosphere

- 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

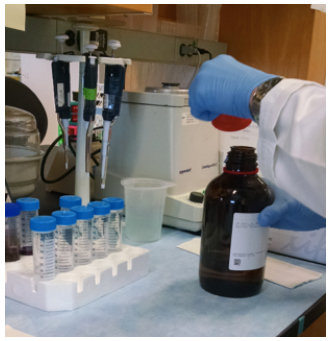
FaceBase Consortium

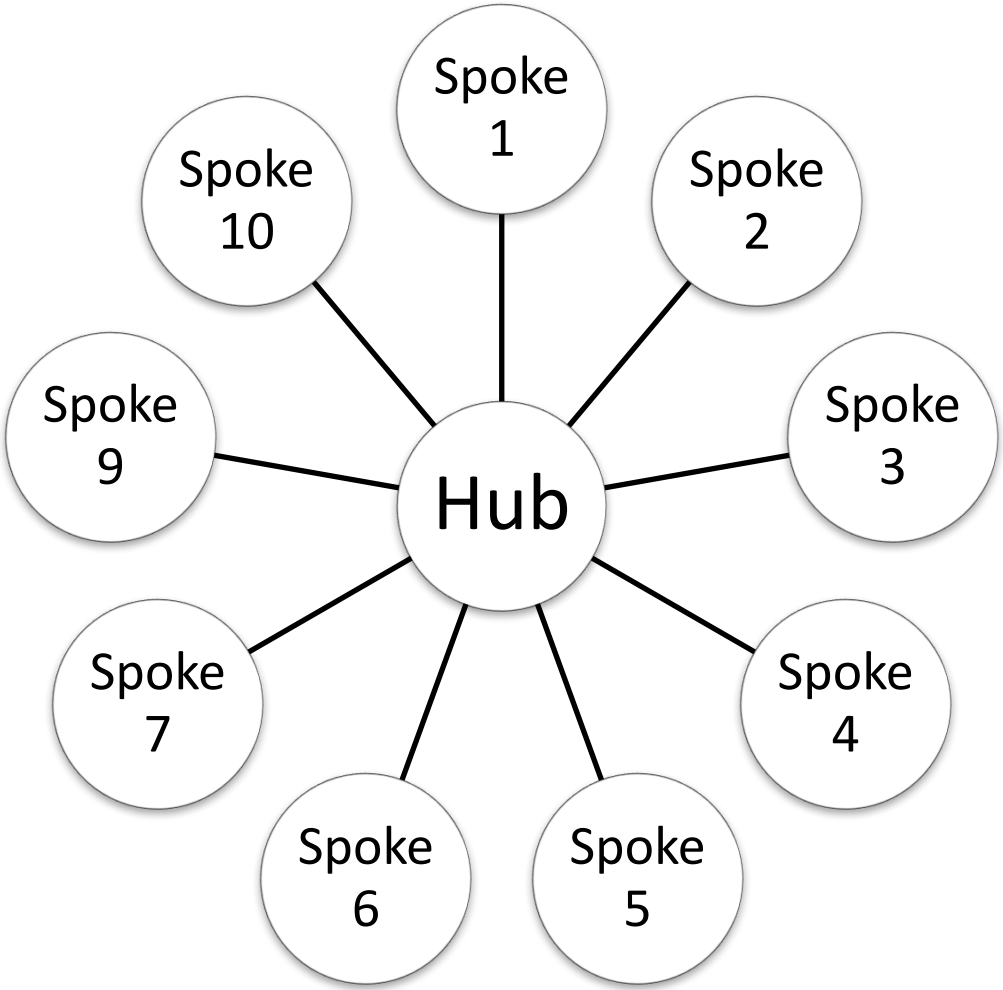
- National Institute for Dental and Craniofacial Research
- Genetics, imaging data: craniofacial development
- 11 projects: clinical, biology, bioinformatics
- 4 model organisms: human, primates, mice, zebrafish
- Make data available on hub **www.facebase.org**

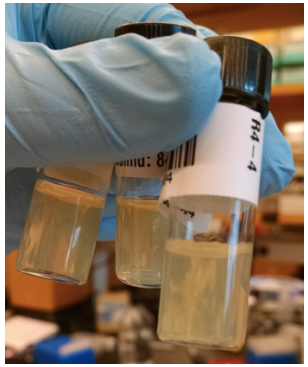




FaceBase Spokes and Hub





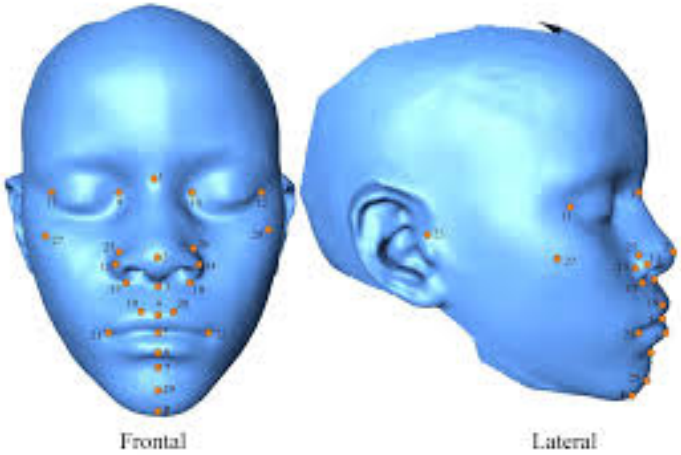


1 coordinating center
10 spokes

Photos: Irene Pasquetto

Mixtures: Craniofacial deformities

- Domains
  - Genomics, bioinformatics
  - Molecular, developmental biology
  - Dentistry, plastic surgery
- Project characteristics
  - Urgent medical problem
  - Species-specific data
    - Humans
    - Primates
    - Mice
    - Zebrafish
  - Competing tools, methods
  - Multiple established infrastructures



FaceBase

Research Question 2

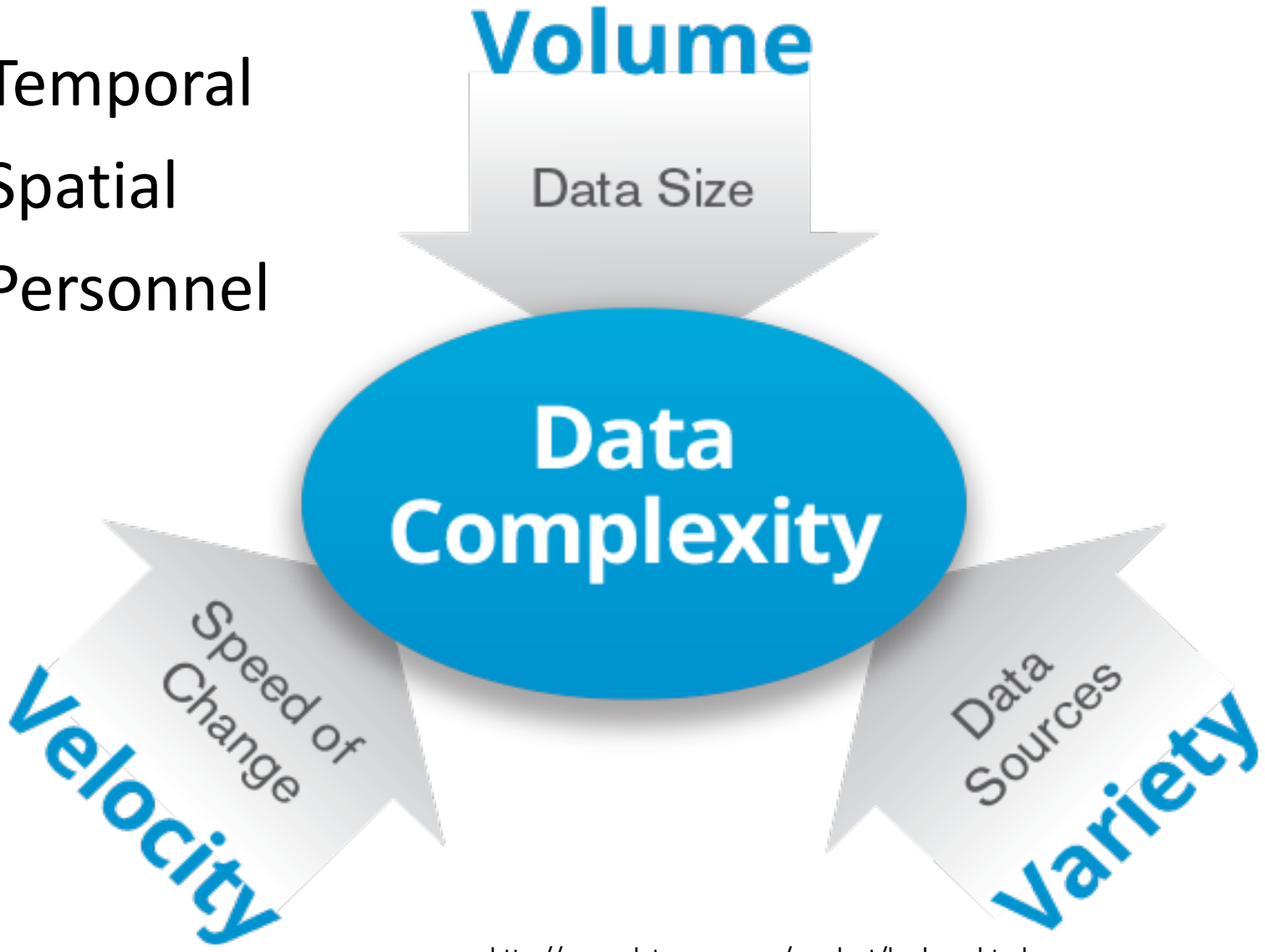
What *factors of scale* influence research practices, and how?

Domain
Astronomy sky surveys
Deep subseafloor biosphere
Craniofacial research
Computational science
Astrophysics phenomena

UCLA Center for Knowledge Infrastructures

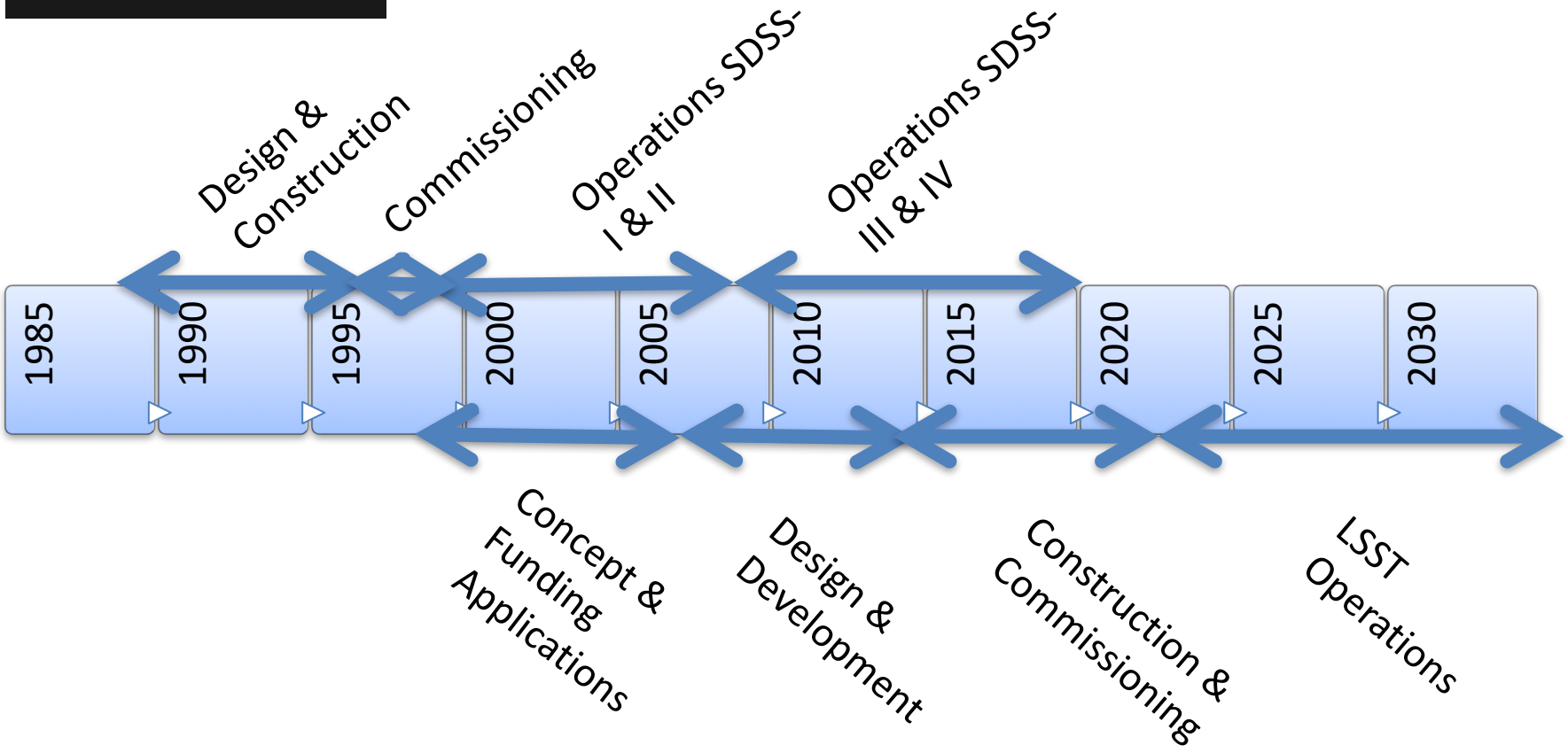
Scale factors

- Temporal
- Spatial
- Personnel



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

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
Craniofacial 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
Astrophysics phenomena	Long time frame of data collection, continuous integration

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
Craniofacial research
Computational science
Astrophysics 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
Craniofacial 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
Astrophysics phenomena	Centralized data collection; ad hoc curation over two decades

Conclusions so far (2016)

- General
  - Data reuse and sharing are distinct yet varied
  - Factors interact: domain mixtures, scale, centrality
- Research themes
  - Domains consist of subdomains with fluid boundaries
  - Volume might be least important scale factor
  - Centrality contradictions
    - Centralized data collections become decentralized in analysis
    - Decentralized data collections are hardest to integrate for analysis

UCLA Center for Knowledge Infrastructures

Emerging Threads (2017): People and Infrastructure in the Context of Data Sharing and Reuse

1. Invisible Work : Expertise; Repair and Maintenance; Technicians; Relationships
2. The Politics of Infrastructure: The Afterlives of Projects; Reproducibility; Open Science
3. Digital Science as Scientific Labor: Reproducibility; Data Abundance; Discovery vs Hypothesis; Open Science and Career Formation
4. Machine Learning: Reproducibility; Algorithms; Expertise

UCLA Center for Knowledge Infrastructures

Paper Ideas:
The Politics of Infrastructure

Setting the Cadence (1)

An examination of governance in setting the cadence of the LSST telescope. This is also setting the agenda for research and the distribution of resources to astronomic subfields. It also examines the role of the LSST book in setting the cadence for potential funders - and against potential rivals.

Federated or Formalized: Political Organization and Knowledge Infrastructures

C-DEBI and CENS as federations of associated disciplines compared to the more formalized structure of the astronomy projects.

UCLA Center for Knowledge Infrastructures

Paper Ideas:
New Expertise and New Norms

More Data, New Problems: Data Reuse and Perceptions of Researcher Misconduct

Traditional norms of scientific practice are changing in fields touched by digital science. Big data generation requires the efforts of multiple researchers and technicians from fields with (oft) differing assumptions about both data sharing, reuse and norms of researcher (mis)conduct.

New Knowledge Infrastructures, Emerging Forms of Expertise (2)

A critical revisit of Collins and Evans taxonomy of expertise - particularly a rethinking of what constitutes "constitutive expertise" in the realm in digital science. This paper uses SDSS data and maybe CENS to examine the making and unmaking of careers in light of emergent forms of expertise.

UCLA Center for Knowledge Infrastructures

Paper Ideas:
Knowledge Infrastructures and Invisible Work

The Technicians of Digital Science (3)

This paper examines the technicians/staff who care for logistics, archiving, and tend to personal relationships on distributed projects. Examples are drawn from Biocurious, CENS, and Facebase.

Repairing and Maintaining Knowledge Infrastructures

Everyone wants to build new infrastructure, nobody wants to do maintenance. What happens when a project ends or infrastructure needs mending? This might compare CENS and C-DEBI or draw some examples from astronomy.

UCLA Center for Knowledge Infrastructures

Paper Ideas: Reproducibility

Irreproducible Science : Some Muddles in the Model of Digital Science

This paper takes a critical look at digital work flows – from cleaning data, to developing pipelines, to attempts at establishing portable environments (virtual machines, Jupyter notebooks, Smalltalk environments) in the production of irreproducible science. It takes irreproducible science as a two-part problem of malleability (of digital tools) and expertise.

Reproducibility and Disciplinary Imperialism (3)

Two disciplines, chemistry and physics, are not suffering from a crisis of reproducibility.

Acknowledgements

Christine Borgman

Peter Darch

Michael Scroggins

Irene Pasquetto

Bernie Boscoe

Milena Golshan

Ashley Sands

Data Archiving and Networked Services

DANS

UCLA Center for Knowledge Infrastructures