The Big Data to Knowledge (BD2K) Guide to the Fundamentals of Data Science

WHY DATA SHARING AND REUSE ARE HARD TO DO

CHRISTINE BORGMAN
Distinguished Professor

IRENE PASQUETTO
Ph.D. Candidate

UCLA

APRIL 21, 2017

TCC BD2K Training Coordinating Center

NIH Data Science at NIH

BD2K CCC

CHRISTINE BORGMAN

- Distinguished Professor & Presidential Chair in Information Studies at UCLA
- Author of more than 250 publications in information studies, computer science, and communication.
- Directs the Center for Knowledge Infrastructures with research grants from the Alfred P. Sloan Foundation, the National Science Foundation, and other sources.

IRENE PASQUETTO

- A research assistant at the Center for Knowledge Infrastructures.
- A Ph.D. Candidate in the Department of Information Studies at UCLA.
- Working on her dissertation on data sharing and reuse practices in molecular biology and genomics.
- Research interests include open science frameworks, science governance models, and, more in general, the ethics and policies of data and code practices.

Why Data Sharing and Reuse are Hard to Do

Christine L. Borgman

Distinguished Professor and Presidential Chair in Information Studies

University of California, Los Angeles

http://christineborgman.info

@scitechprof

Irene V. Pasquetto

Doctoral Candidate in Information Studies

University of California, Los Angeles

@IrenePasquetto

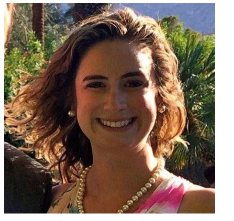
NIH BD2K Data Science Webinar

April 21, 2017

https://datascience.nih.gov/FundamentalsSeries

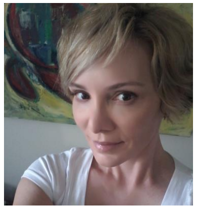
Christine Borgman

Peter Darch



Ashley Sands

Irene Pasquetto



Bernie Randles

Milena Golshan

UCLA Center for Knowledge Infrastructures

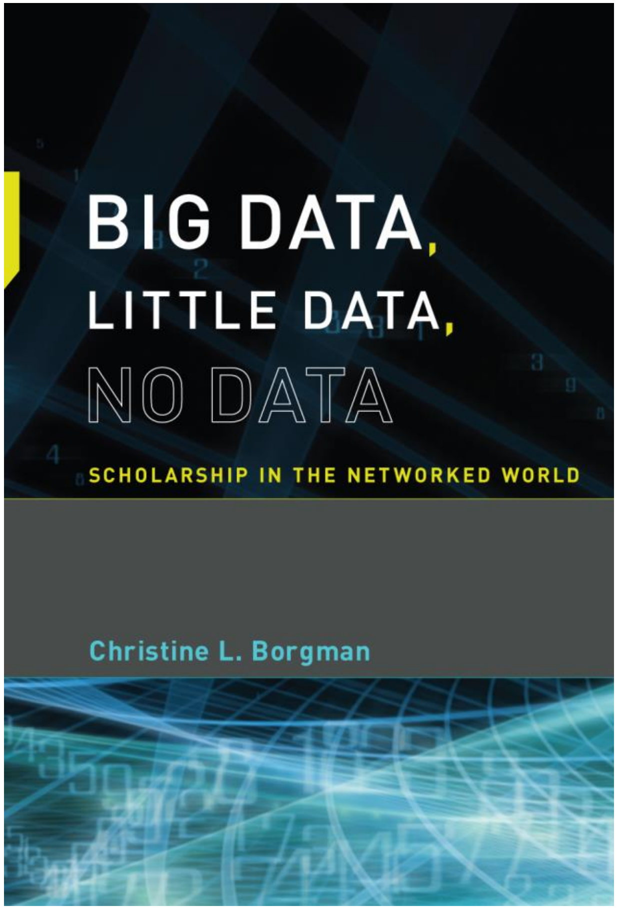
https://knowledgeinfrastructures.gseis.ucla.edu

Data sharing policies

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

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

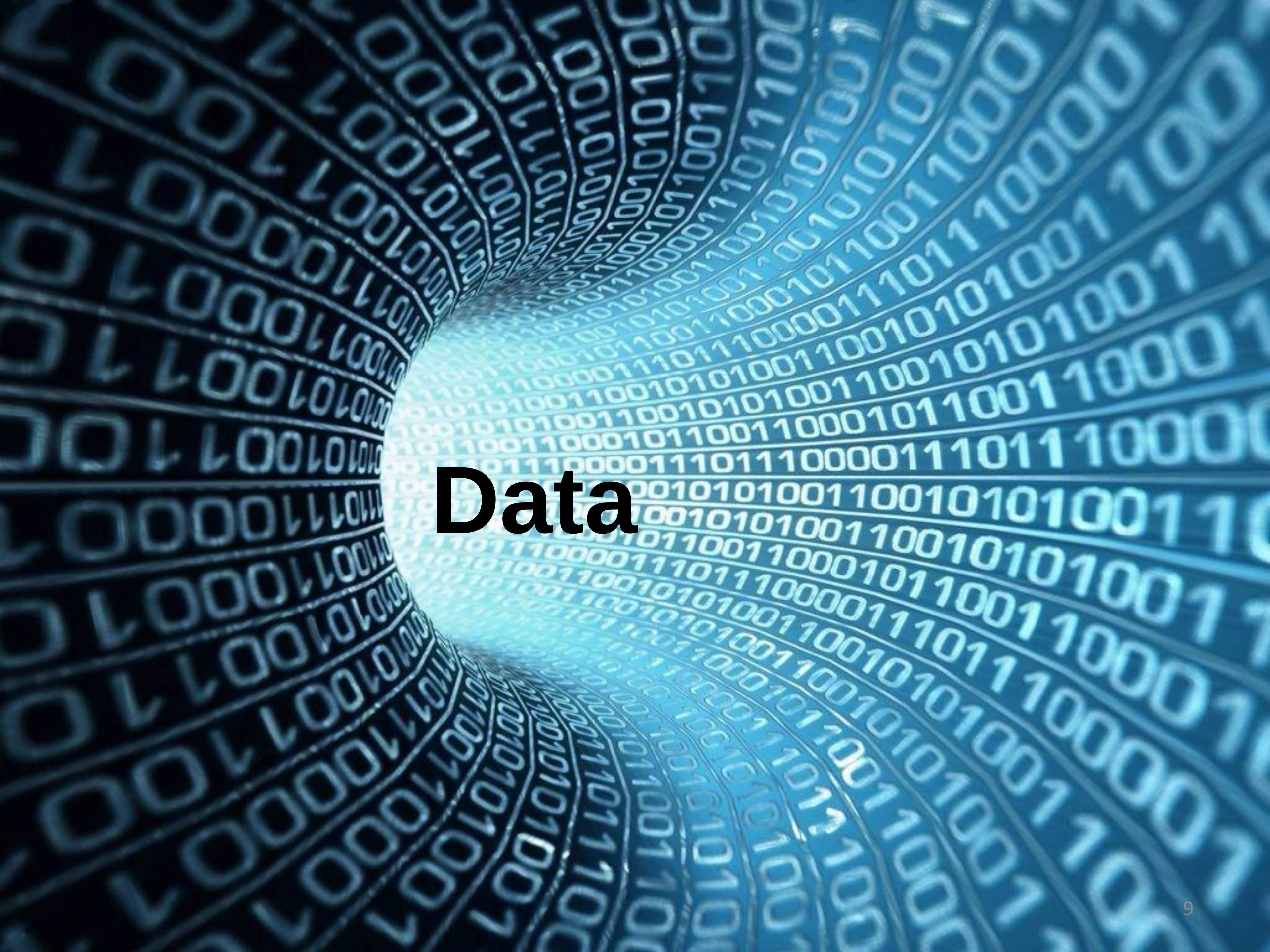
http://www.buildingsrus.co.uk/.../ target1.htm

Why Reuse Research Data?

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

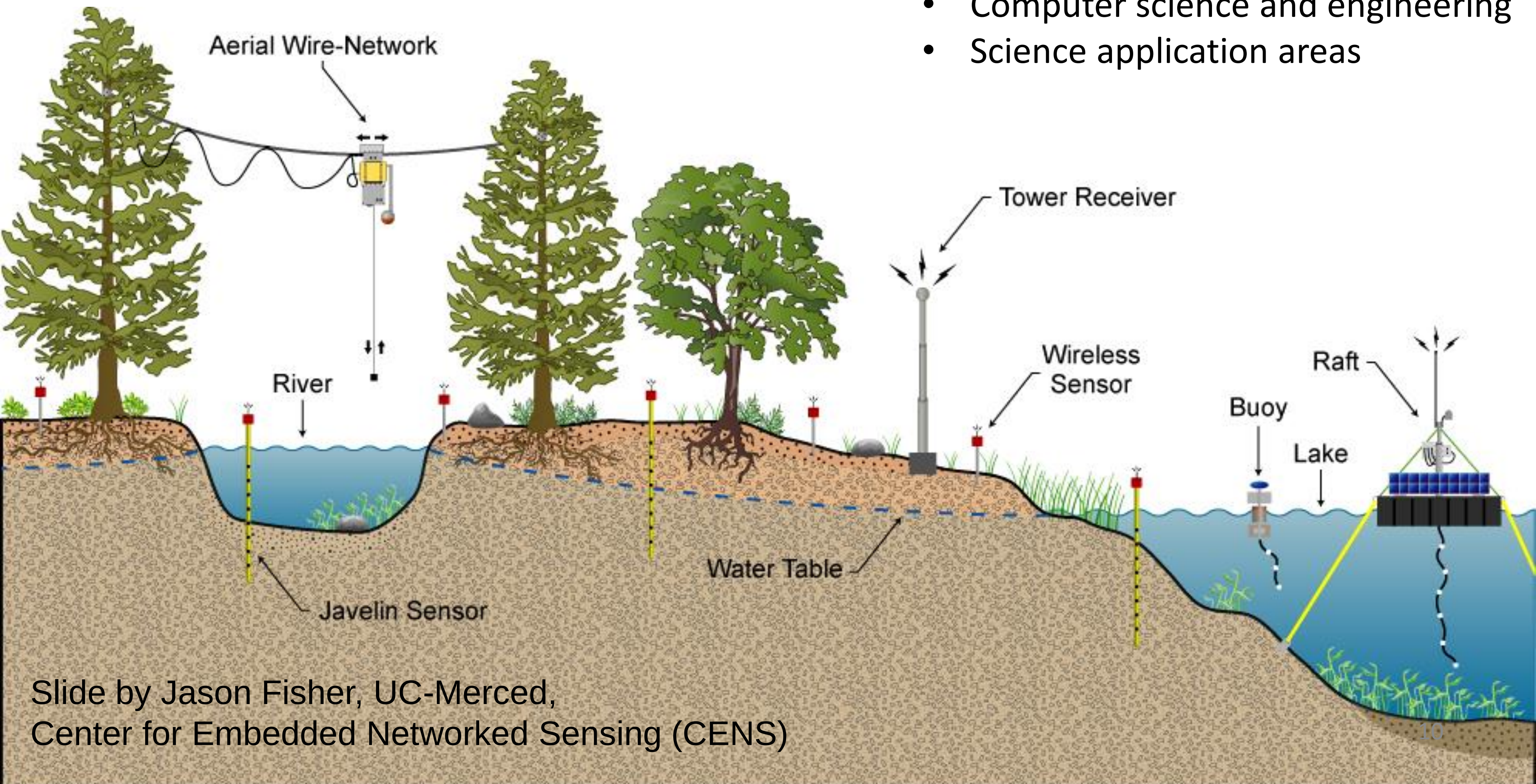
https://www.flickr.com/photos/pagedooley/

Data



Center for Embedded Networked Sensing

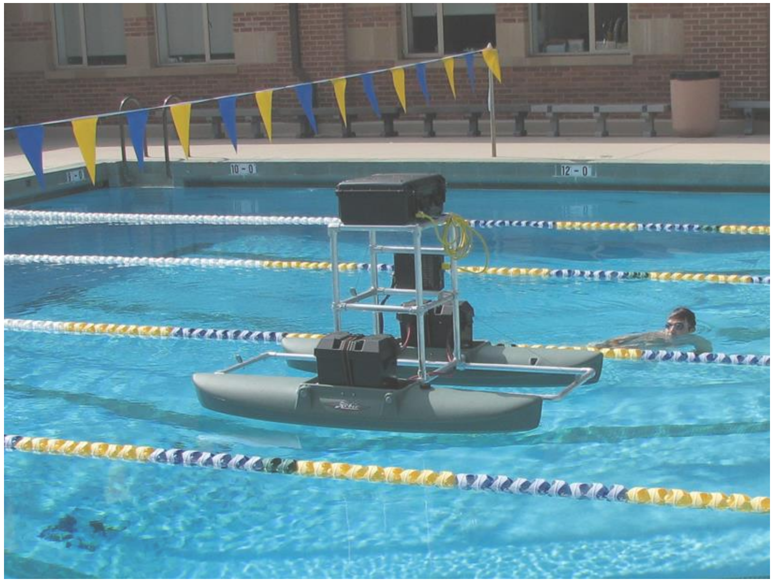
- NSF Science & Tech Ctr, 2002-2012
- 5 universities, plus partners
- 300 members
- Computer science and engineering
- Science application areas



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.."*

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

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

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
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

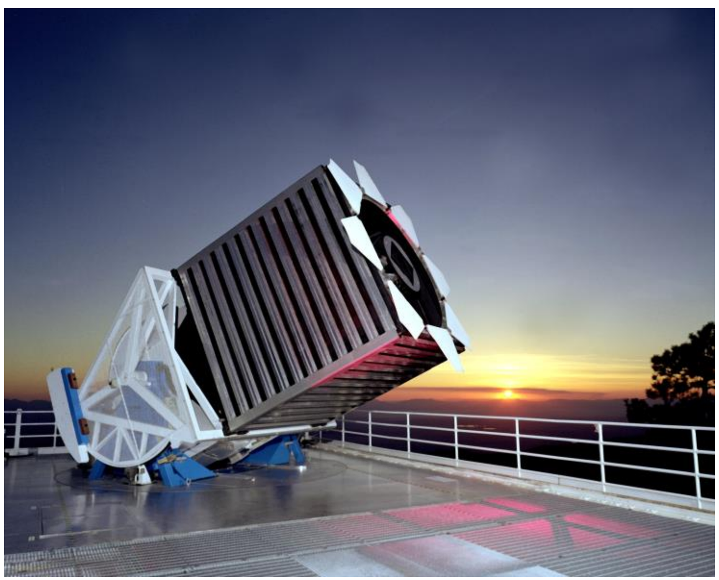
How do the *mixtures of domain expertise* influence the collection, use, and reuse of data – and vice versa?

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

UCLA Center for Knowledge Infrastructures

Sloan Digital Sky Survey (SDSS-I/II)

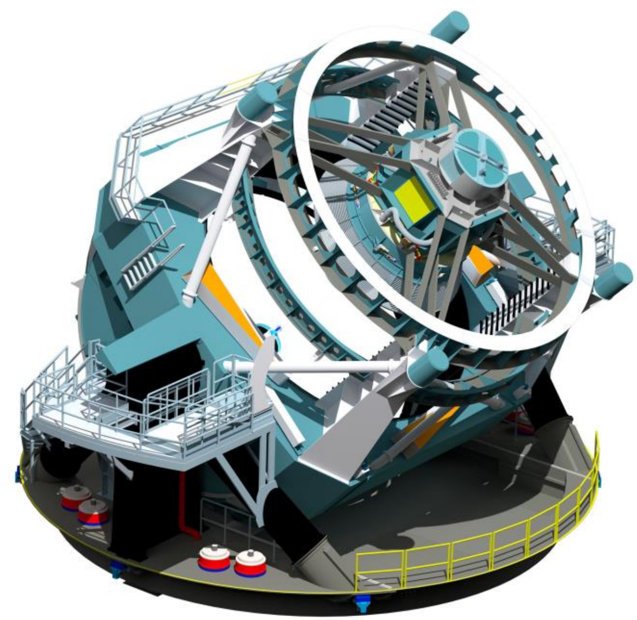
- Survey from 2000-2008
- 160+ TB data total
- Tens of millions of dollars
- Open data
- Proprietary software



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

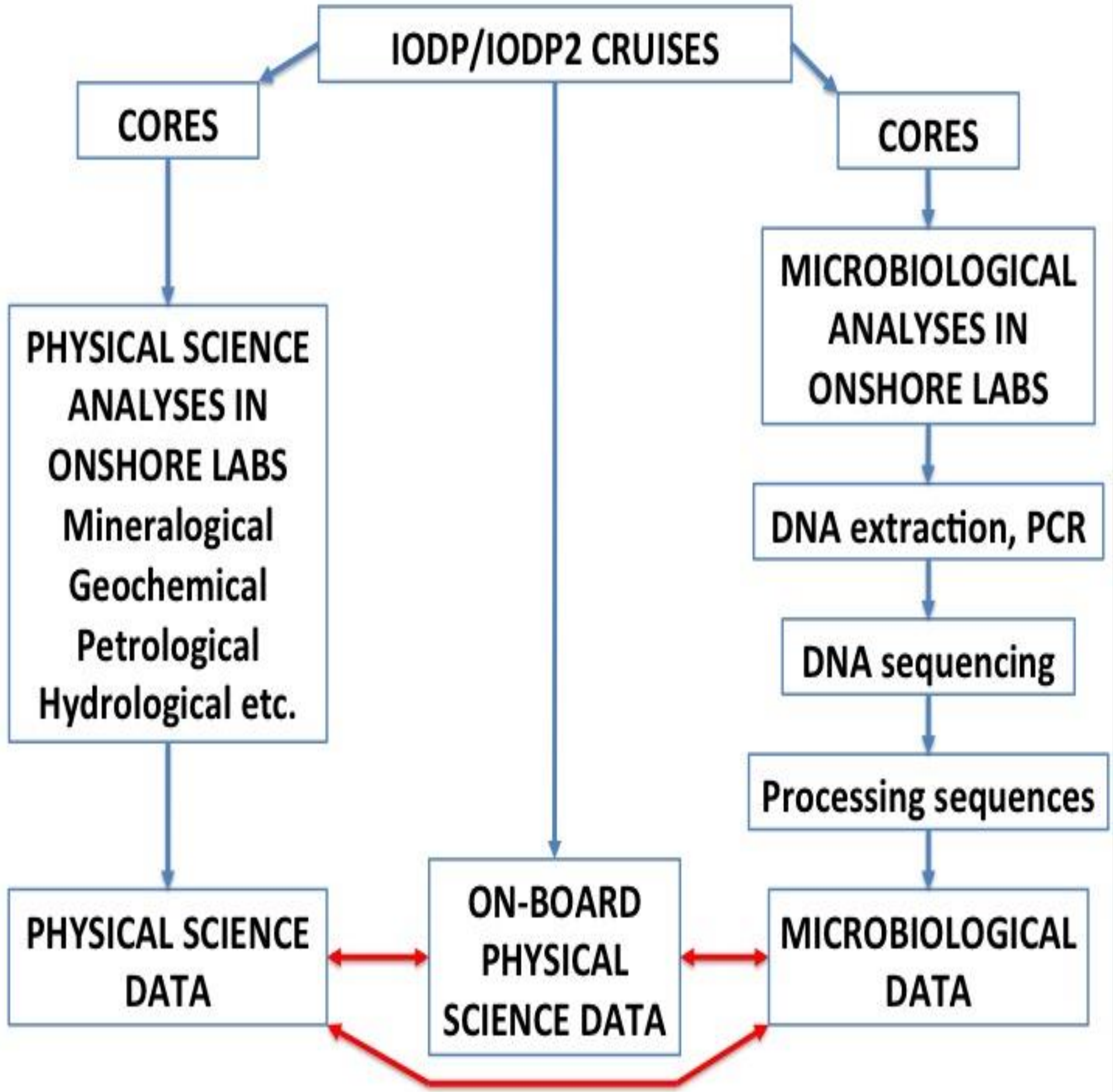
- Domains
  - Astronomy, physics
  - Computer science
- Project characteristics
  - Mature discipline
  - Abundant data
  - Trusted archives
  - Shared tools, methods
  - Established infrastructure for data access and use

Center for Dark Energy Biosphere Investigations

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

Research Question 2

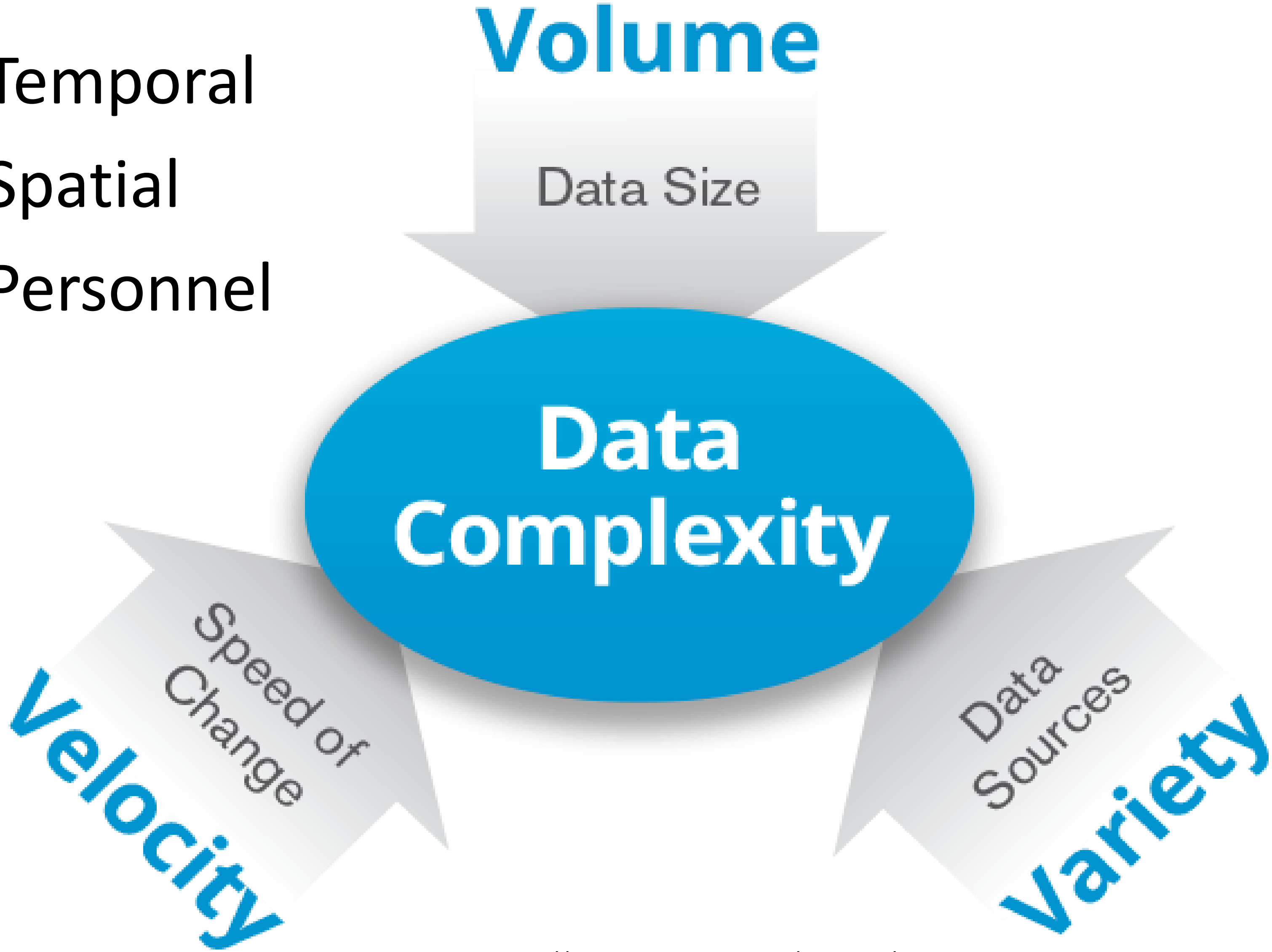
What *factors of scale* influence research practices, and how?

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

UCLA Center for Knowledge Infrastructures

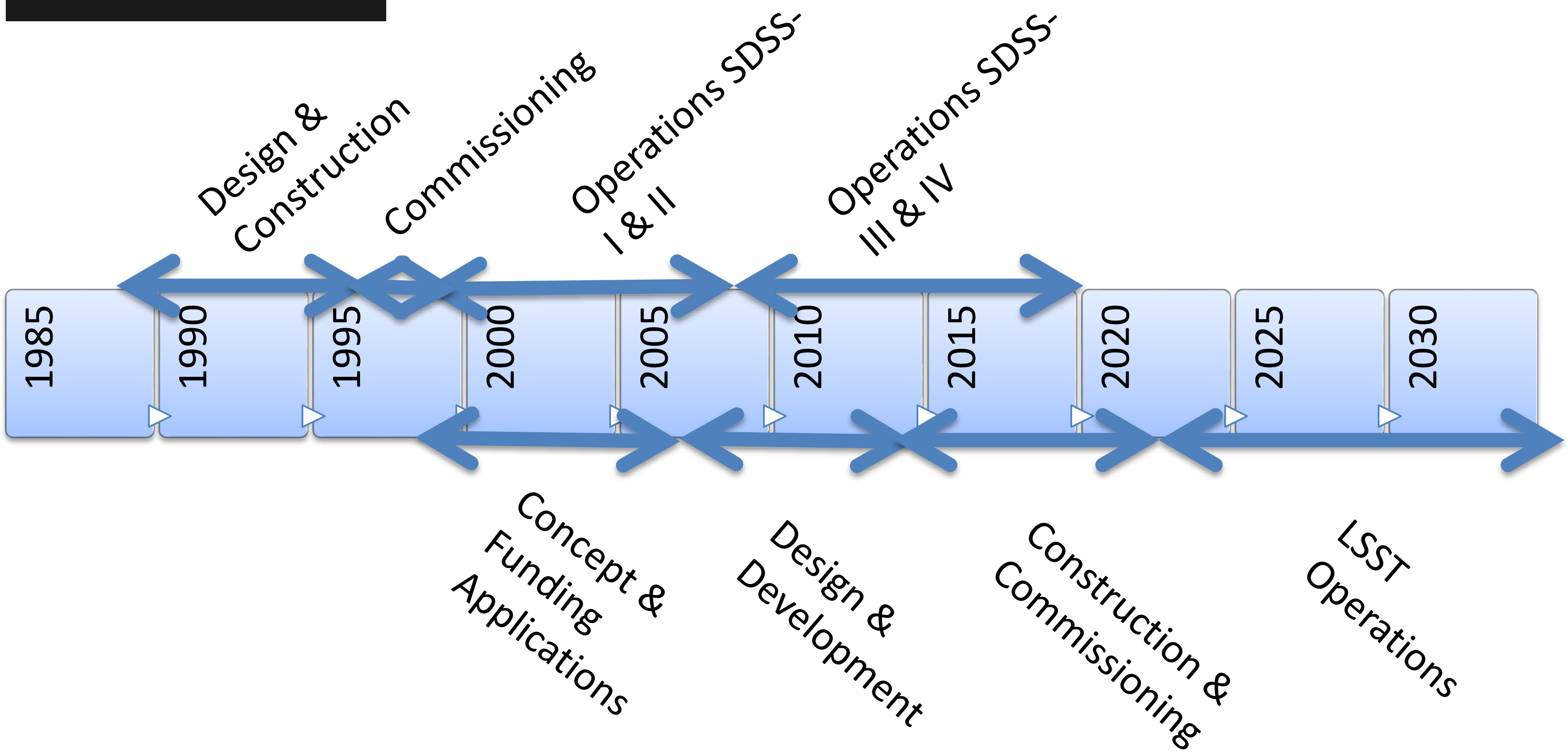
Scale factors

- Temporal
- Spatial
- Personnel



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

Project Timelines



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

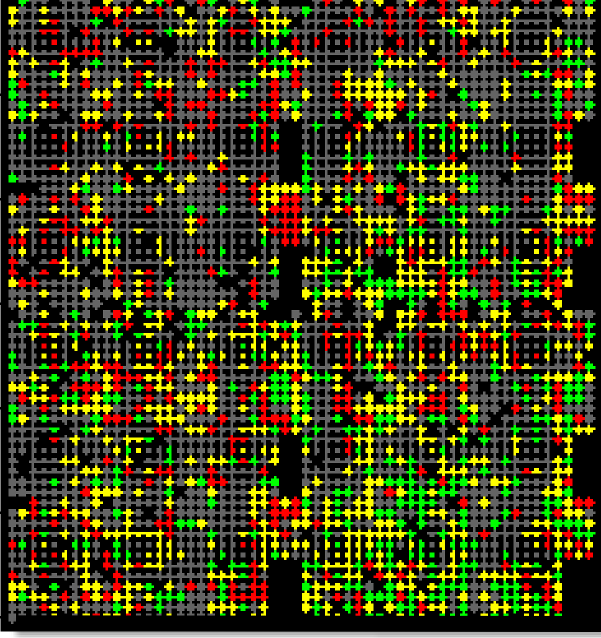
Research Site	Centralization factors
Astronomy sky surveys	Centralized data collection and initial processing; decentralized use and analysis
Deep subseafloor biosphere	Common data source, shared repositories of cores; decentralized analysis
Biomedical research	Decentralized data collection; efforts 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

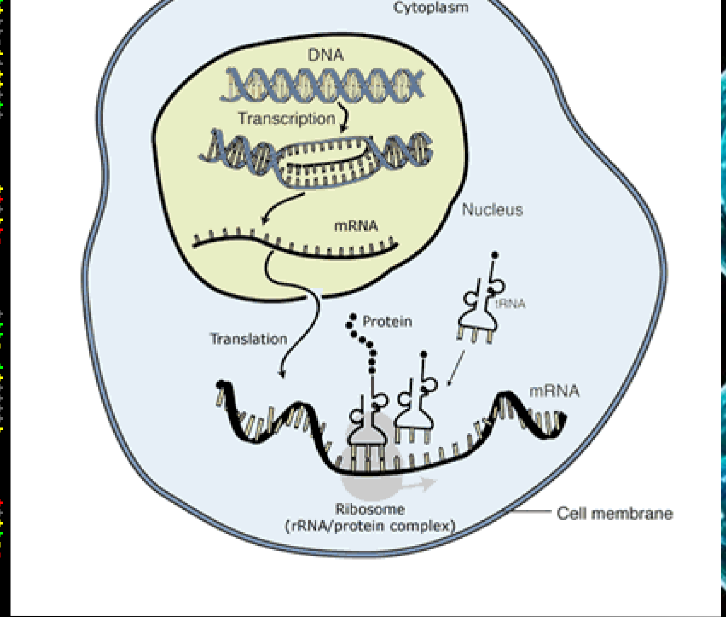
Biomedical Case Study in Data Sharing and Reuse

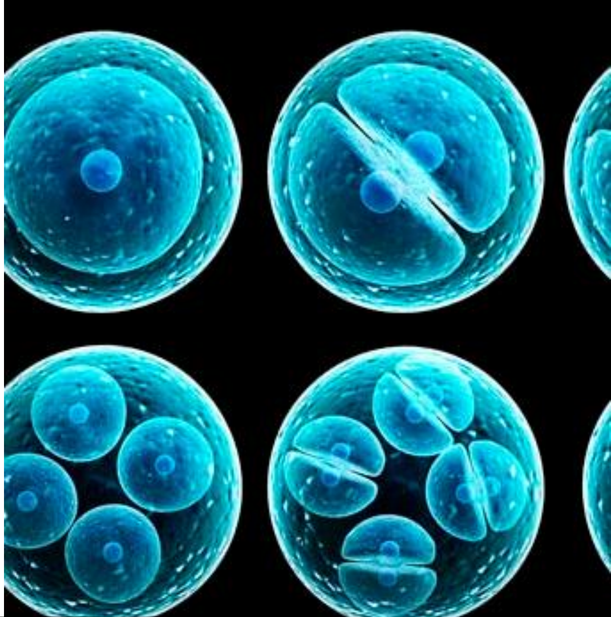
Irene V. Pasquetto

PhD Candidate

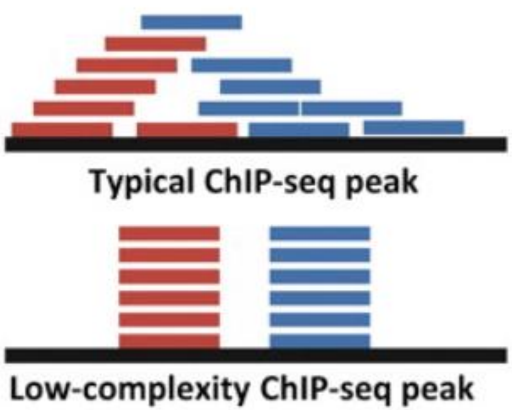
Information Studies, UCLA

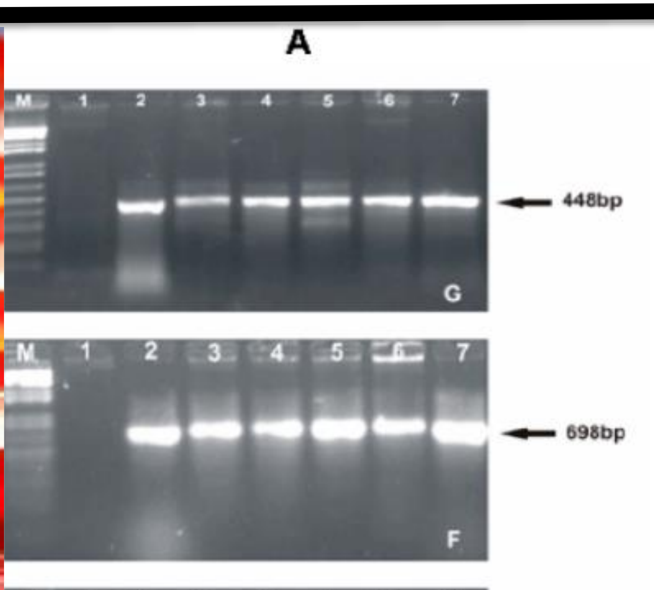






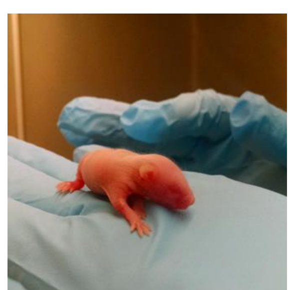
THE DATA ARE SHARED.
WHO IS REUSING THEM? WHY? HOW?

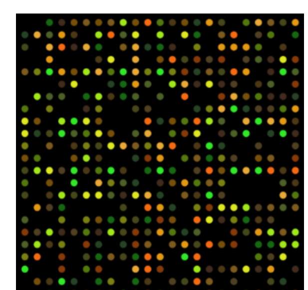




A CONSORTIUM FOR DATA SHARING

- **Data collected**: images, anatomy norms, sequences, gene expression drawing
- **11 interdisciplinary projects**: clinical, biology, bioinformatics
- **4 model organisms**: human, primates, mice, zebrafish
- **Types of data:** patients' images, metrics for anatomy norms, phenotypic images, sequences from genome wide association studies, results from genes/proteins function validation studies
- **Goals**: data integration, systemic approaches to knowledge discovery





DATA REUSE AND SHARING PRACTICES

- **DATA SHARING PRACTICES.** Most scientists are willing to release data prior to publication
- **DATA REUSE PRACTICES.** Reusing data is more challenging than sharing. Data reuse practices vary by skills, expertise, disciplinary focus, and type of data
- Concerns about data reuse:
  1. **Common issues**
  2. **Disciplinary-specific challenges**

1. COMMON ISSUES with DATA REUSE

1.1 Reusing “unpublished data.” Most scientists expressed concern about reusing data if not associated with peer-reviewed publications.

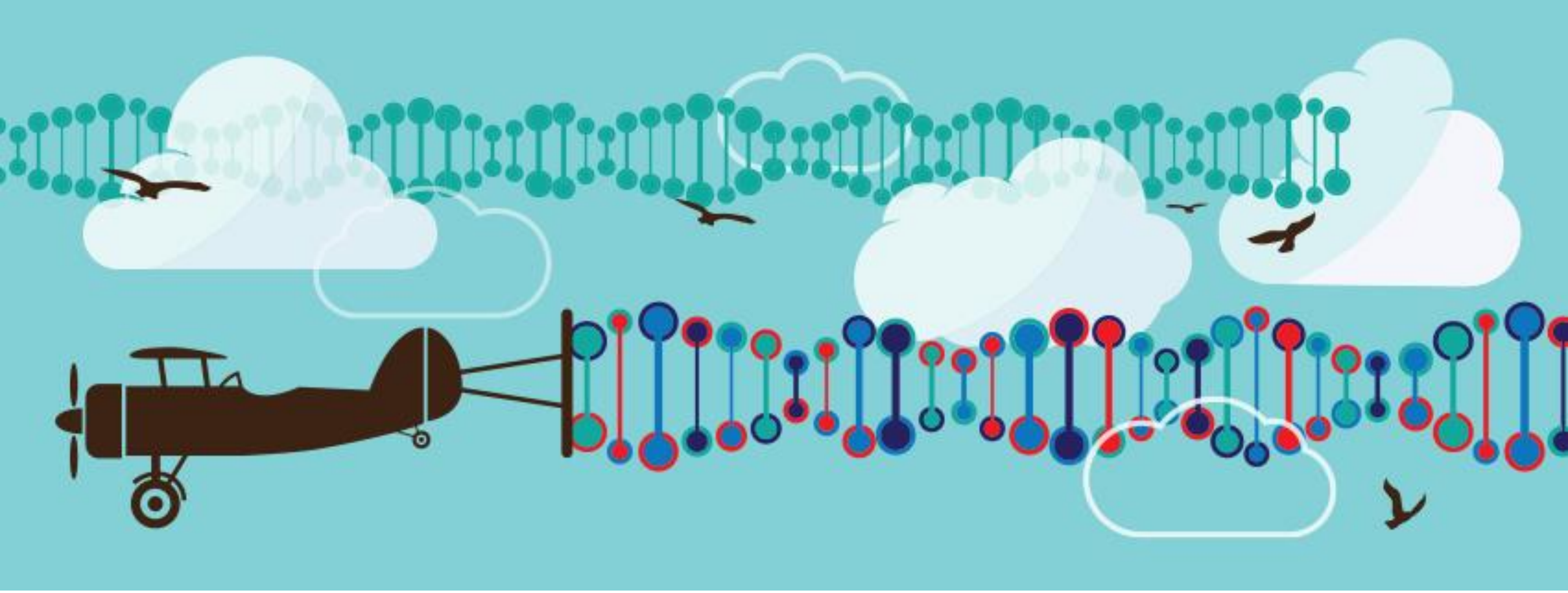
1.2 Verifying data quality. Establishing trust in others’ data requires time and effort. Most study participants are concerned about trusting others’ data and about the quality of data.

1.3 Reusing data knowledge production vs. quality control. Data may be reused to check quality of newly collected data, rather than to produce new knowledge.

2. DISCIPLINARY-SPECIFIC CONCERNS

2.1 Expertise and skills for reusing data. Membership in a domain or specific skills influences how data are accessed and reused by the community.

2.2 Concerns about data reuse and approximation of results. When data are scarce, reusing available data can reduce accuracy and validity. Some scientists are concerned that reusing “close enough” data may produce misleading or approximate results.



- Data reuse is a set of heterogeneous practices
- Data reuse is a process, rather than a single act
- Data are validated at multiple points in the research process

Summary of Research Themes

- Domains consist of subdomains with fluid boundaries
- Volume of data may be least important scale factor
- Centralized data collections become decentralized in analysis
- Decentralized data collections are hardest to integrate for analysis

UCLA Center for Knowledge Infrastructures

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

Conclusions

- Data can be shared in many ways
- Data sharing is not an end in itself
- Data reuse requires
  - Knowledge about the data
  - Validation at multiple stages
  - Stewardship and sustainability
  - Trust

UCLA Center for Knowledge Infrastructures

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

Recommendations for practice

- Identify practices of subdomains and interactions
- Seek right level of abstraction for data sharing, integration, curation, reuse
- Invest in data curation early in project design
- Promote infrastructure solutions
  - Shared tools and services
  - Data discovery mechanisms
  - Iterative stewardship

UCLA Center for Knowledge Infrastructures

Acknowledgements


Christine Borgman


Peter Darch


Ashley Sands


Irene Pasquetto


Bernie Randles


Milena Golshan

UCLA Center for Knowledge Infrastructures