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

Christine L. Borgman

Distinguished Professor and Presidential Chair in Information Studies

University of California, Los Angeles

http://christineborgman.info

https://knowledgeinfrastructures.gseis.ucla.edu

@scitechprof

Christine Borgman

Peter Darch

Ashley Sands

Irene Pasquetto

Bernie Randles

Milena Golshan

Pietro Santachiara

UCLA Center for Knowledge Infrastructures

Closing Keynote

SciDataCon, Denver, CO

September 13, 2016

THE WHITE HOUSE
WASHINGTON

# Data sharing policies

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

JST JAPAN SCIENCE AND TECHNOLOGY AGENCY

中央研究院
ACADEMIA SINICA 1928

CHINESE ACADEMY OF SCIENCES

- 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

Australian Government
National Health and Medical Research Council

NSF National Science Foundation
WHERE DISCOVERIES BEGIN

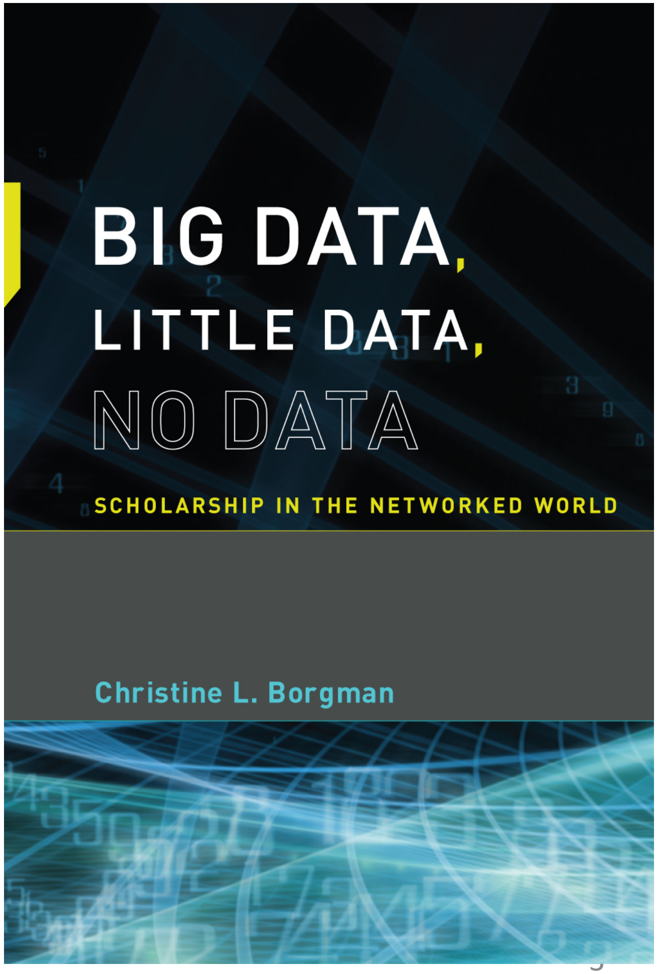
NIH National Institutes of Health
*Turning Discovery Into Health*

Policy RECommendations for Open Access to Research Data in Europe

RECODE

# 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

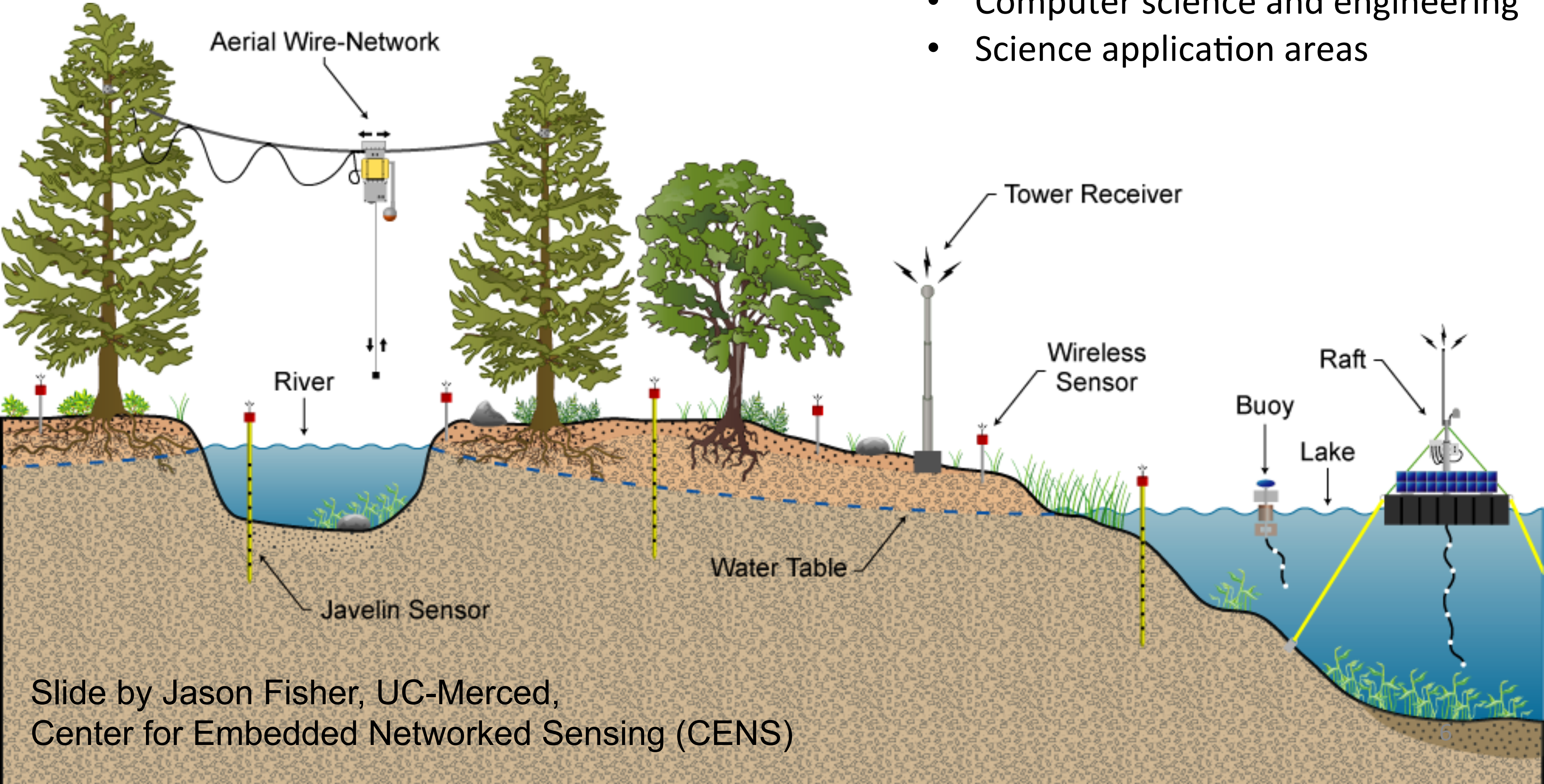
- Rewards for publication
- Effort to document data
- Competition, priority
- Control, ownership

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

# Data

# Center for Embedded Networked Sensing

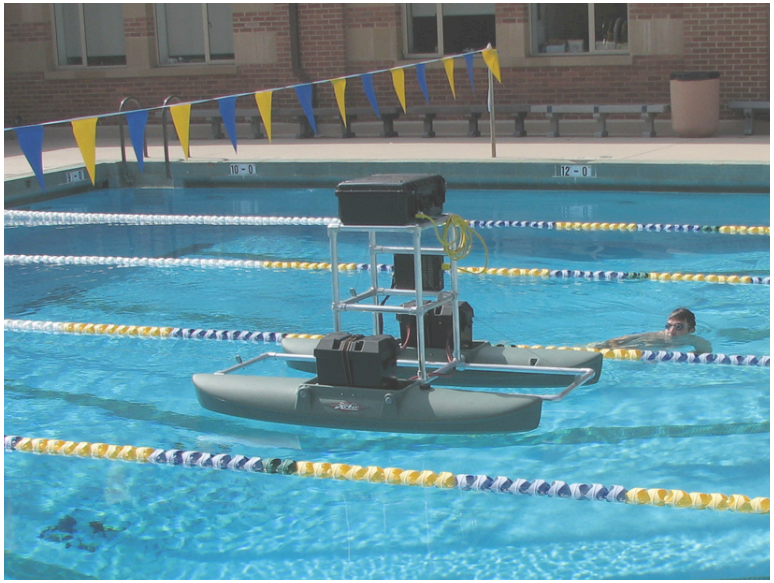
- NSF Science & Tech Ctr, 2002-2012
- 5 universities, plus partners
- 300 members
- Computer science and engineering
- Science application areas



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

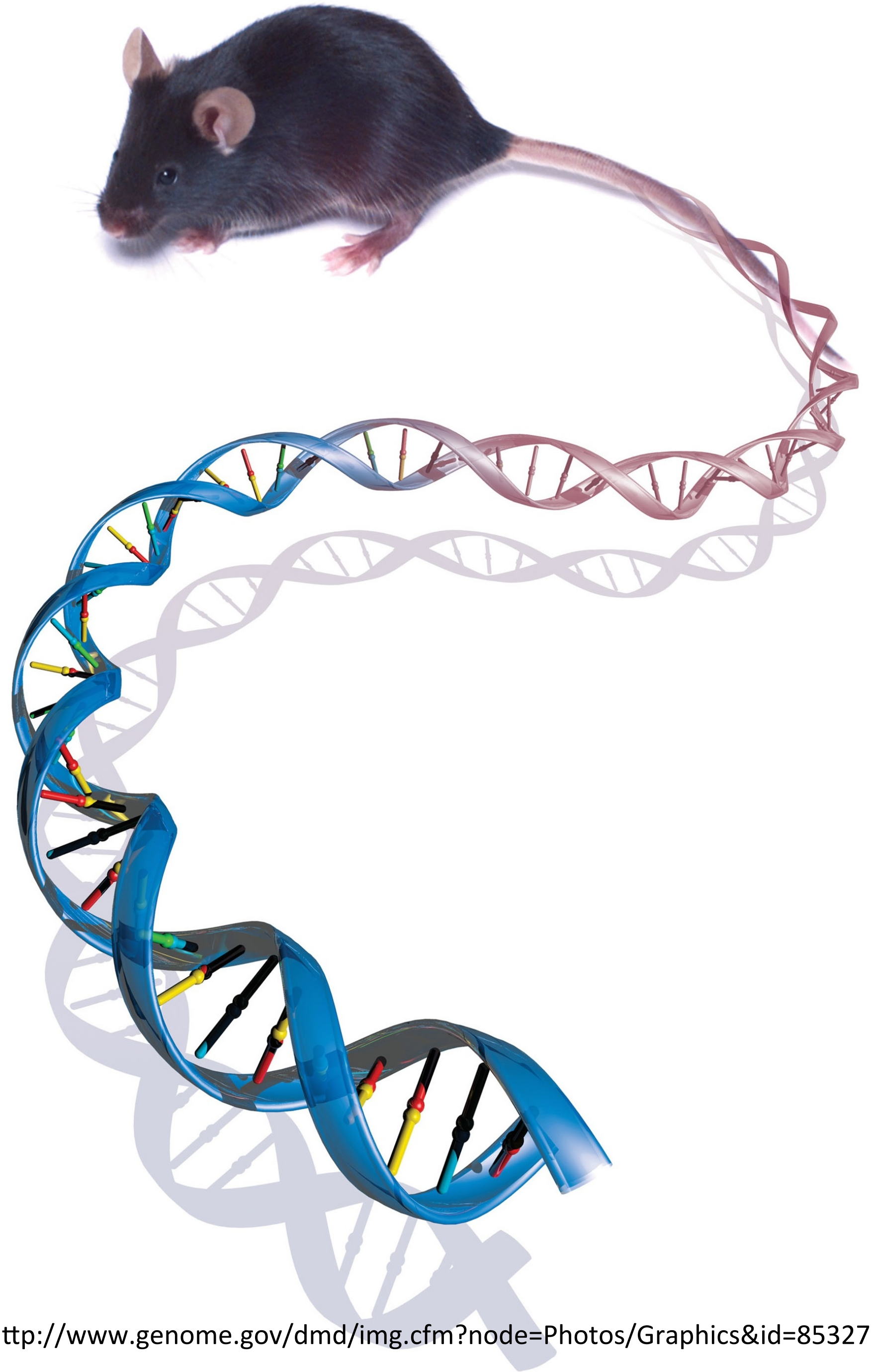
# Documenting Data for Interpretation

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?

- Research Design, 2015-2018
- Methods
- Questions
- Findings
- Conclusions
- DANS study
- Recommendations

Christine Borgman

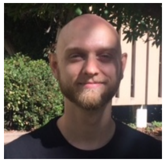
Peter Darch

Ashley Sands

Irene Pasquetto

Bernie Randles

Milena Golshan


Pietro Santachiara

ALFRED P. SLOAN FOUNDATION 1934

UCLA Center for Knowledge Infrastructures

# 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

# 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 birth defects in humans | Genomics of four model organisms |
| Computational science | Problem: Data analysis at scale | Computing platform for astronomy, physics, turbulence, soil science, genomics… |

# Research Question 1

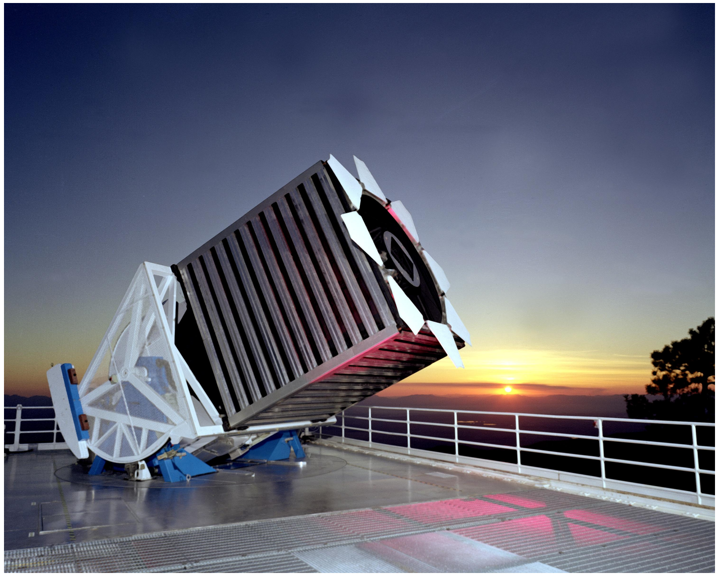
How do the *mixtures of domain expertise* influence the collection, use, and reuse of data – and vice versa?

| Domain |
|---|
| Astronomy sky surveys |
| Deep subseafloor biosphere |
| Craniofacial research |
| Computational science |

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



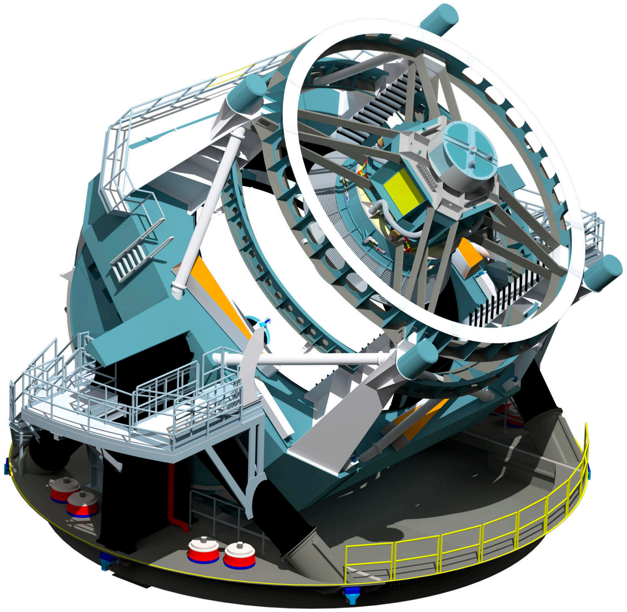
Telescope for the Sloan Digital Sky Survey, Apache Point, New Mexico

# Large Synoptic Survey Telescope (LSST)

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

LSST

Large Synoptic Survey Telescope



LSST telescope, Chile

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

# 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

SDSS

LSST

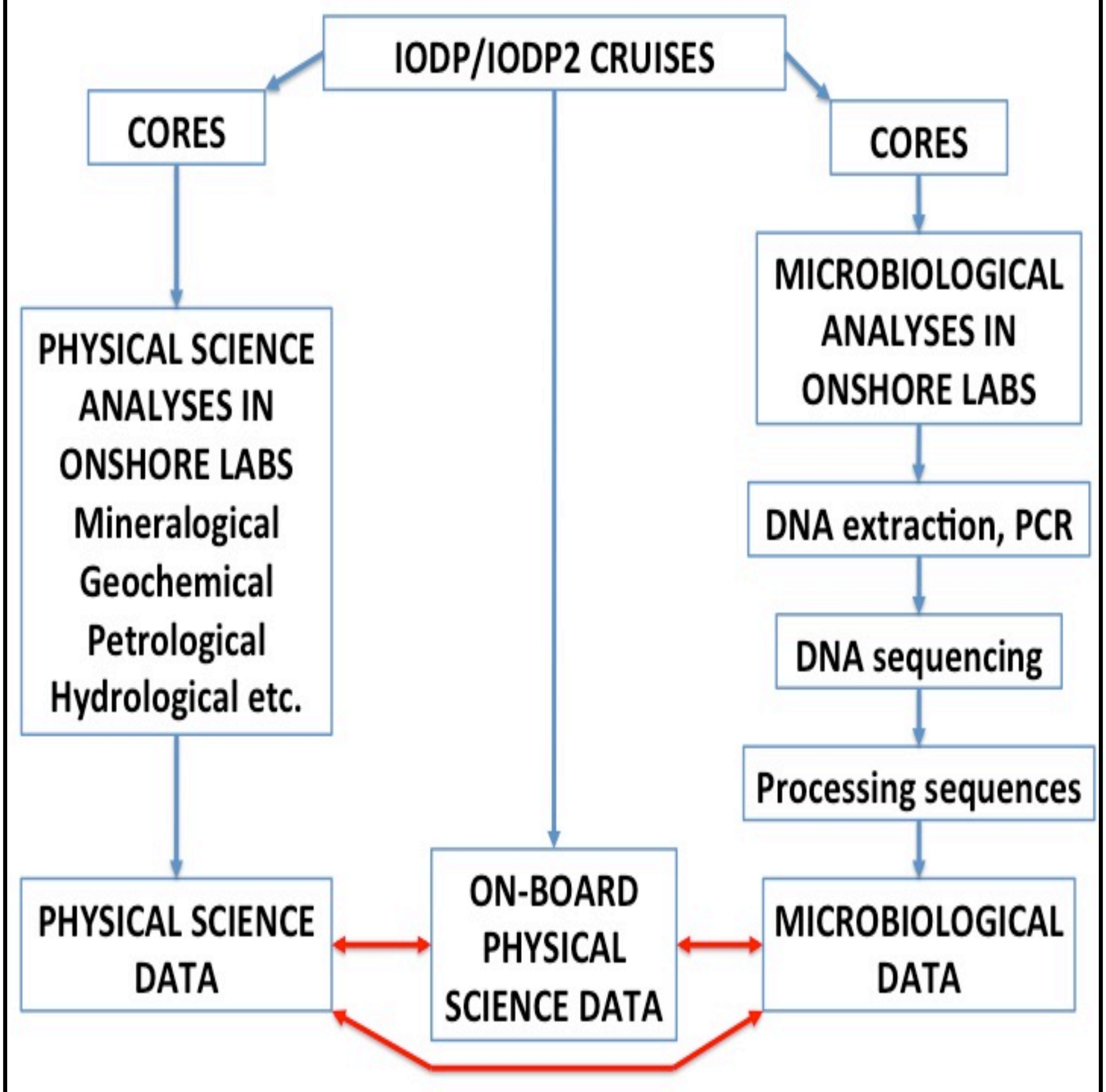
Large Synoptic Survey Telescope

# 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

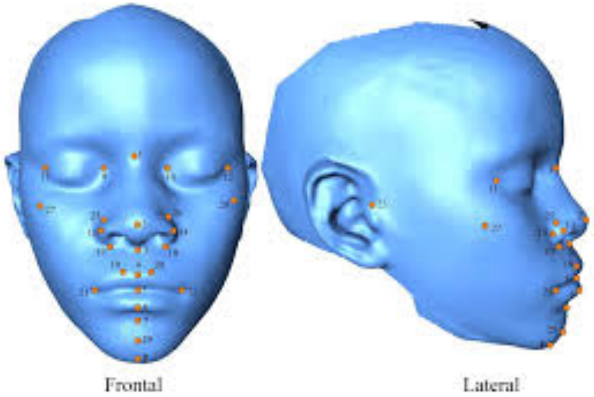
MICROBIOLOGICAL DATA

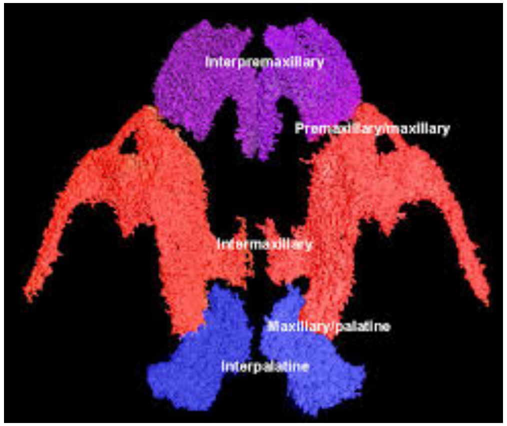
# Mixtures: Deep subseafloor biosphere

- 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

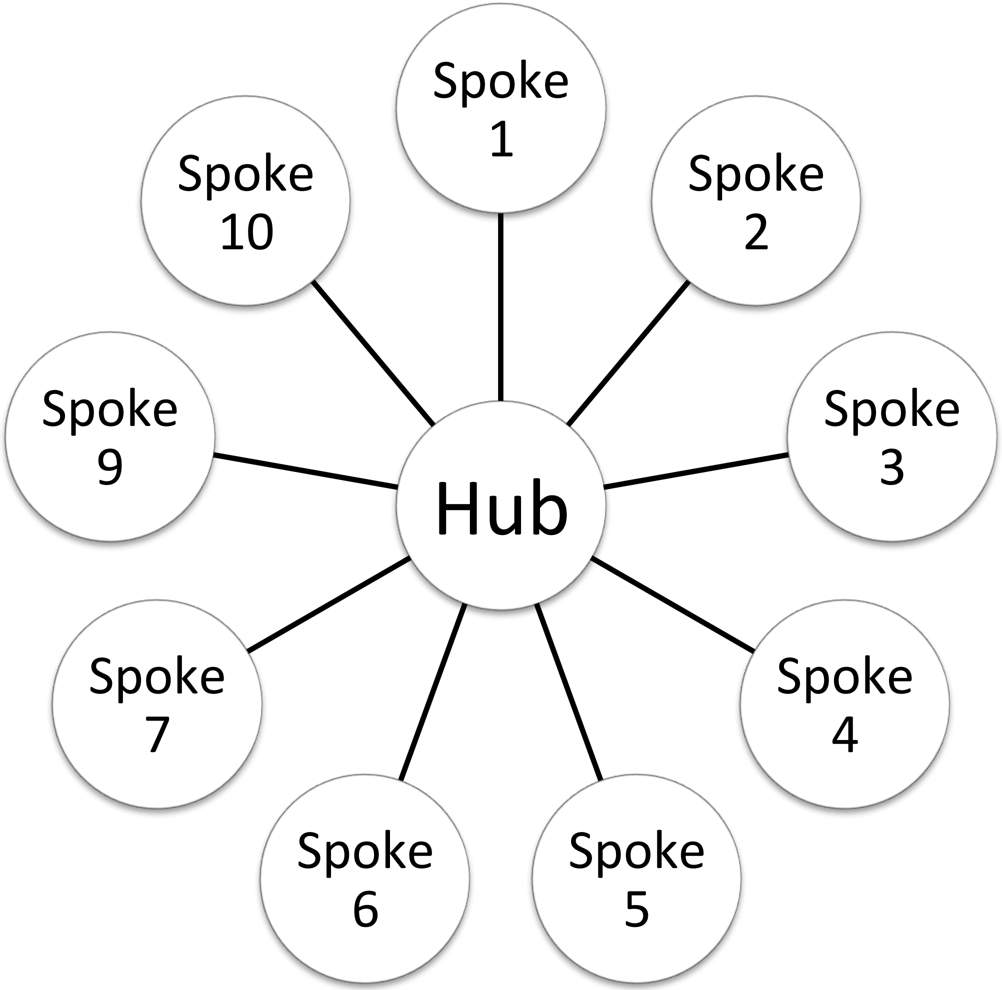
# FaceBase Consortium

- 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



Hub

Spoke 1

Spoke 2

Spoke 3

Spoke 4

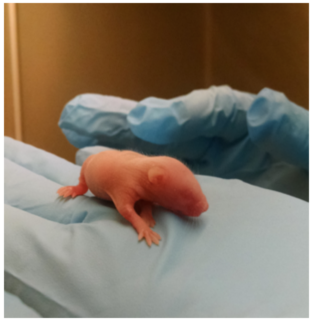
Spoke 5

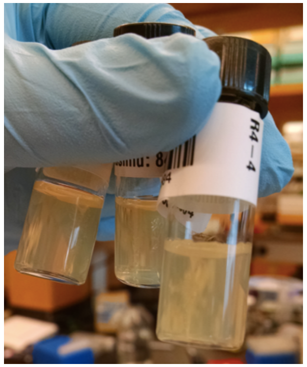
Spoke 6

Spoke 7

Spoke 9

Spoke 10



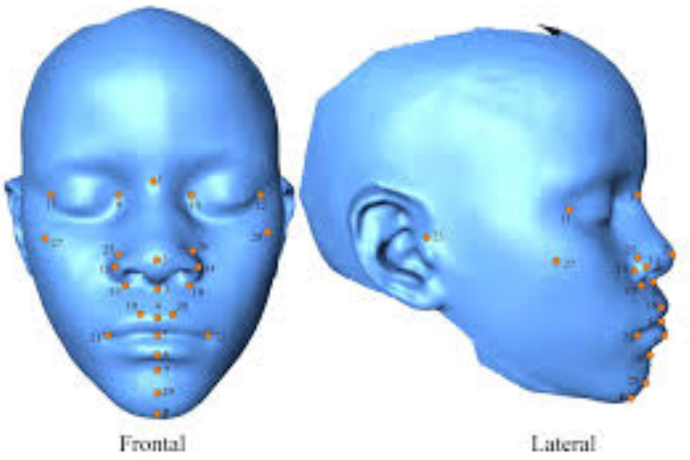


**1 coordinating center**
**10 spokes**

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

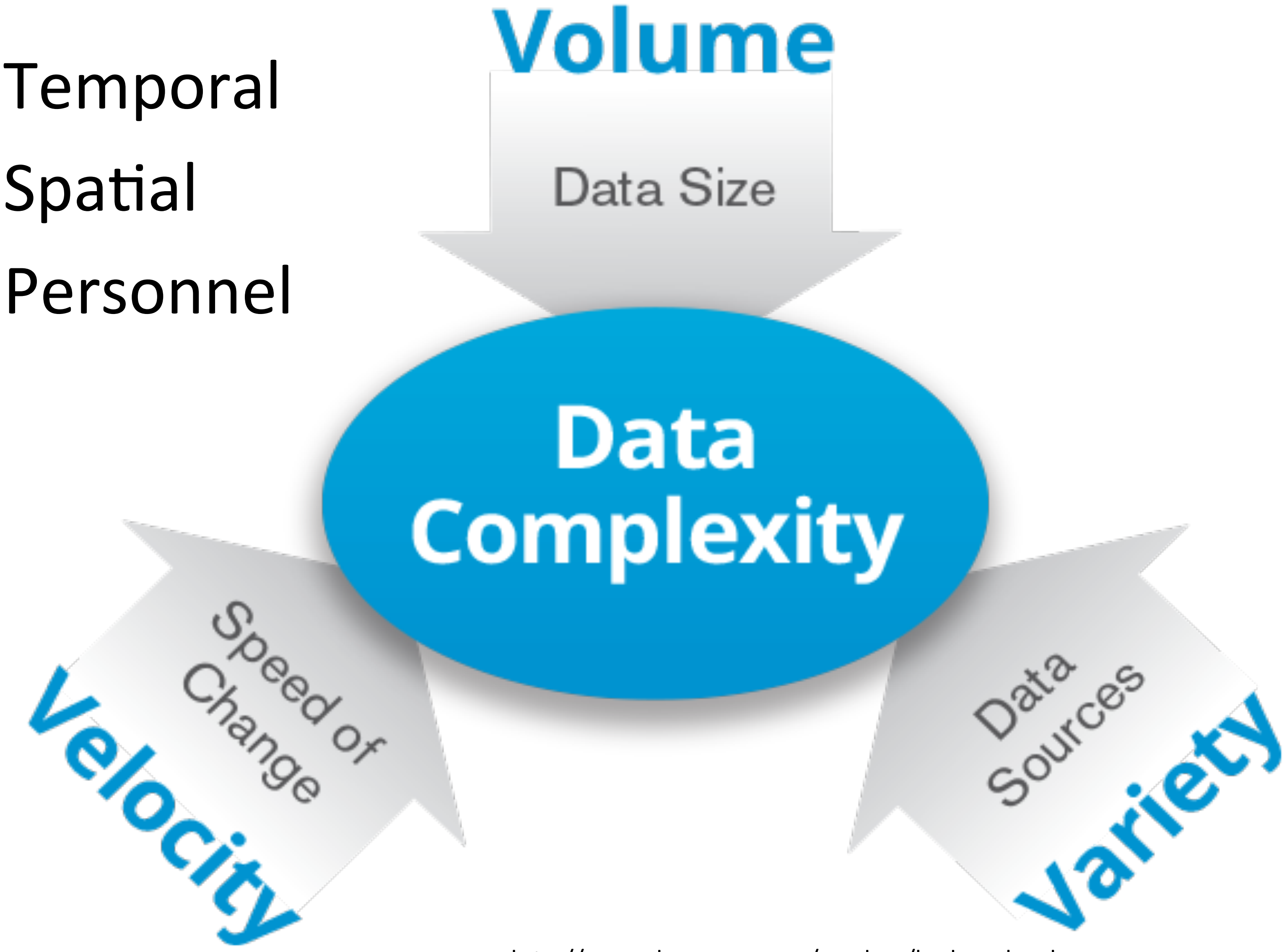
# Research Question 2

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

| Domain |
|---|
| Astronomy sky surveys |
| Deep subseafloor biosphere |
| Craniofacial research |
| Computational science |

# *Scale factors*

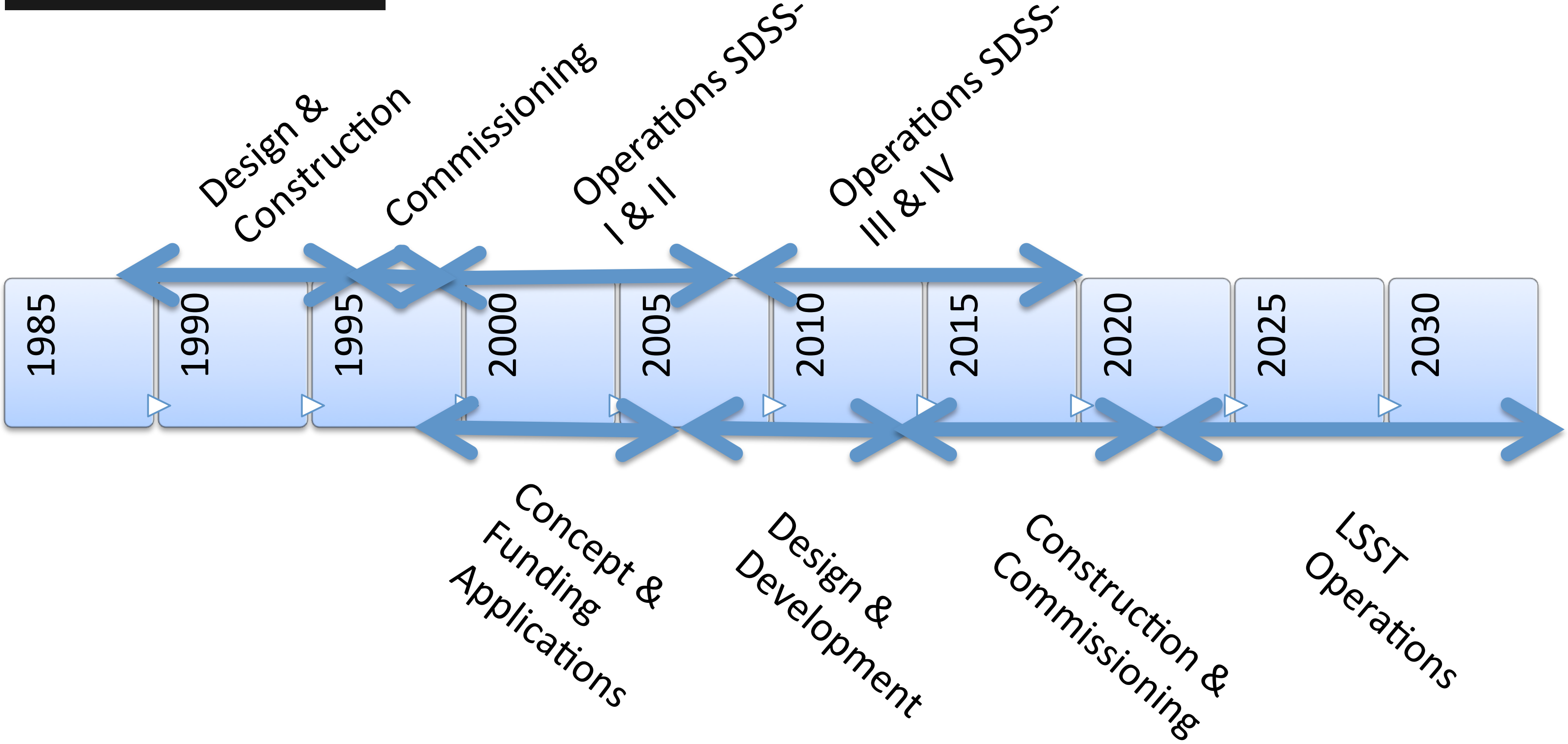
- Temporal
- Spatial
- Personnel



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 |
| 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 |

# 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 |

UCLA Center for Knowledge Infrastructures

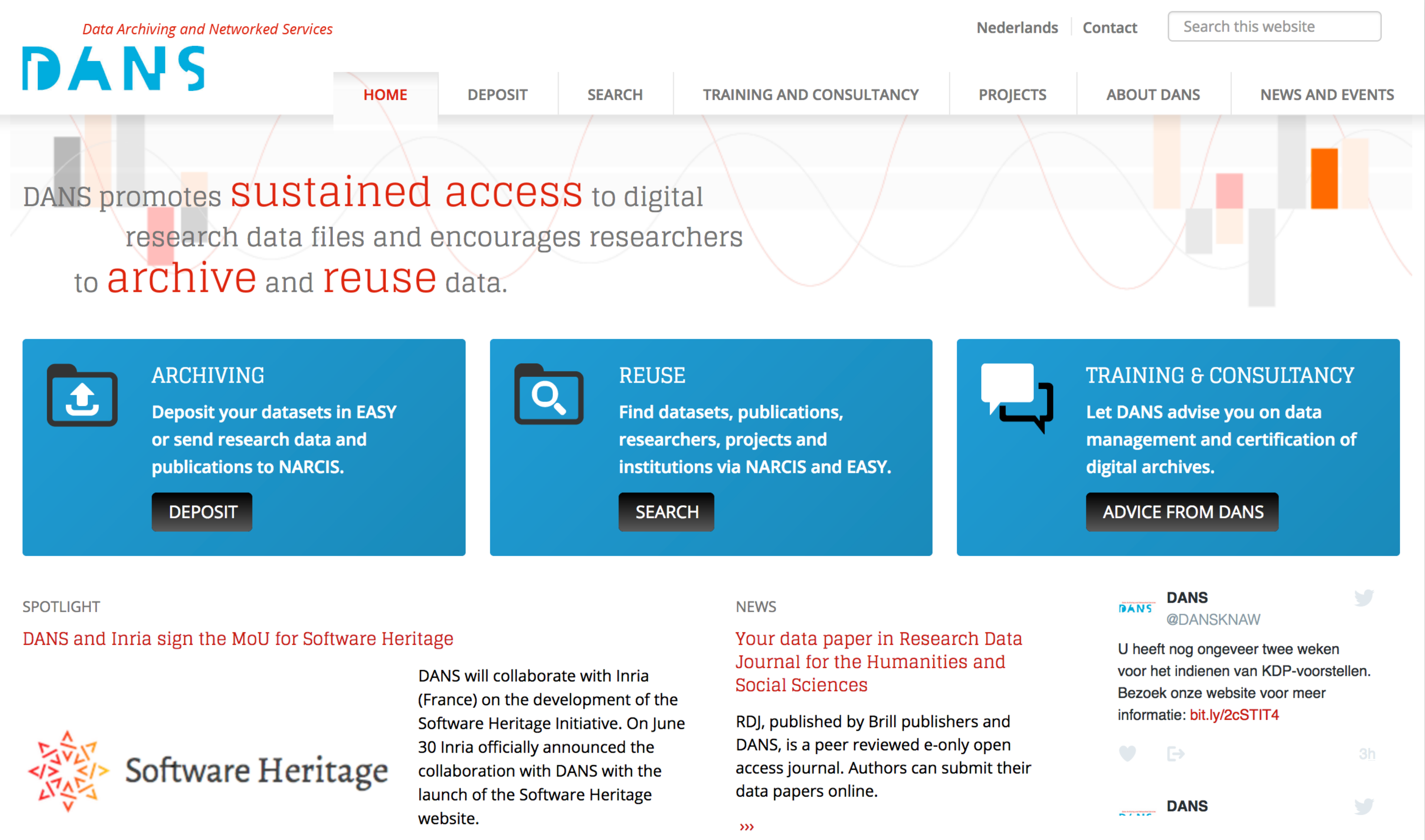
# Centralization factors

| 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 |

# Conclusions so far

- General
  - Data sharing is not one problem, but many
  - 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

# DANS Users and Uses



*Data Archiving and Networked Services*

DANS

Nederlands | Contact | Search this website

HOME | DEPOSIT | SEARCH | TRAINING AND CONSULTANCY | PROJECTS | ABOUT DANS | NEWS AND EVENTS

DANS promotes sustained access to digital research data files and encourages researchers to archive and reuse data.

ARCHIVING

**Deposit your datasets in EASY or send research data and publications to NARCIS.**

DEPOSIT

REUSE

**Find datasets, publications, researchers, projects and institutions via NARCIS and EASY.**

SEARCH

TRAINING & CONSULTANCY

**Let DANS advise you on data management and certification of digital archives.**

ADVICE FROM DANS

SPOTLIGHT

DANS and Inria sign the MoU for Software Heritage

Software Heritage

DANS will collaborate with Inria (France) on the development of the Software Heritage Initiative. On June 30 Inria officially announced the collaboration with DANS with the launch of the Software Heritage website.

NEWS

Your data paper in Research Data Journal for the Humanities and Social Sciences

RDJ, published by Brill publishers and DANS, is a peer reviewed e-only open access journal. Authors can submit their data papers online.

›››

DANS
@DANSKNAW

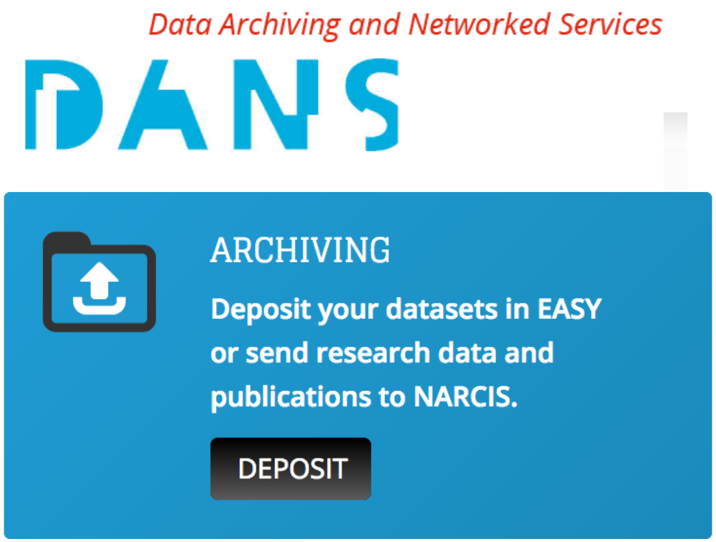
U heeft nog ongeveer twee weken voor het indienen van KDP-voorstellen. Bezoek onze website voor meer informatie: bit.ly/2cSTIT4

3h

DANS

# Why do users put data in DANS?

- Meet legal requirements
- Preserve data for long term
- Get credit for data
- Control access to data
- Use as background service
- Motivate citizen science participants



Data Archiving and Networked Services

DANS

ARCHIVING

Deposit your datasets in EASY or send research data and publications to NARCIS.

Borgman, C. L., Van de Sompel, H., Scharnhorst, A., van den Berg, H., & Treloar, A. (2015). Who Uses the Digital Data Archive? An Exploratory Study of DANS. In *Proceedings of Association for Information Science and Technology*. http://doi.org/10.1002/pra2.2015.145052010096

UCLA Center for Knowledge Infrastructures

# Recommendations so far

- 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

Pietro Santachiara

ALFRED P. SLOAN FOUNDATION

1934

*Data Archiving and Networked Services*

DANS

UCLA Center for Knowledge Infrastructures