Title
Data, infrastructure, and stewardship

Author
Borgman, Christine L.

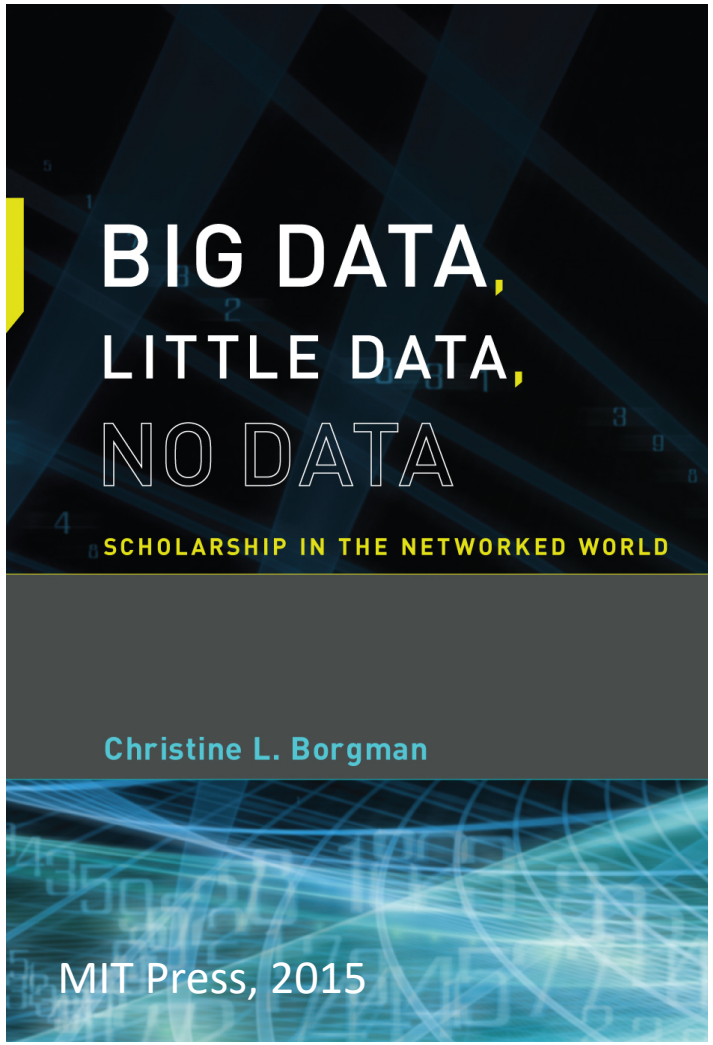
Publication Date
2019-11-07

Data, Infrastructure, and Stewardship

Christine L. Borgman

Distinguished Research Professor
Director, Center for Knowledge Infrastructures
https://knowledgeinfrastructures.gseis.ucla.edu
University of California, Los Angeles
http://christineborgman.info
@scitechprof

Space Studies Board, National Academies of Sciences
Keynote Presentation
Beckman Center, Irvine, 7 November 2019



Data sharing policies

- U.S. Federal research policy
- European Research Council
- 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

National Science Foundation
WHERE DISCOVERIES BEGIN

National Institutes of Health
Turning Discovery Into Health

Policy RECommendations for Open Access to Research Data in Europe

Open Data Practices

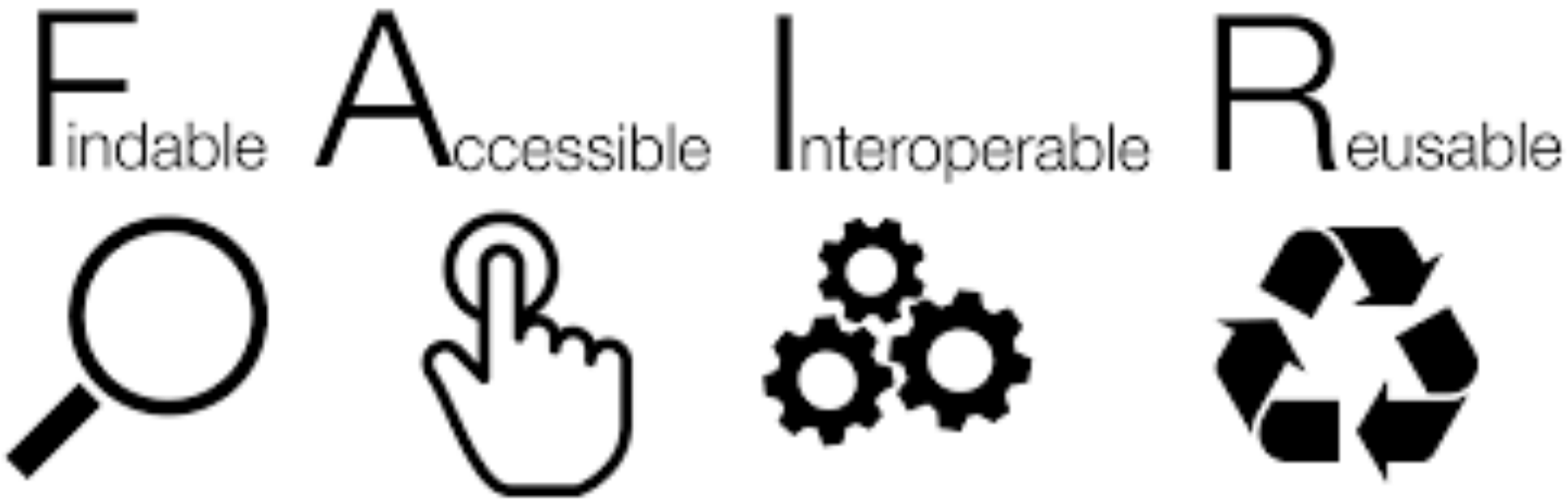
- Deposit datasets in a data archive
- Link datasets to journal article or publication
- Publish data documentation
  - Research protocols
  - Codebooks
  - Software
  - Algorithms
- Cite data and software

PDS: The Planetary Data System

HOME | DATA SEARCH | TOOLS | DATA STANDARDS

Home | About PDS | Data Users | Data Proposers | Data Providers

Data Stewardship: The Ideal



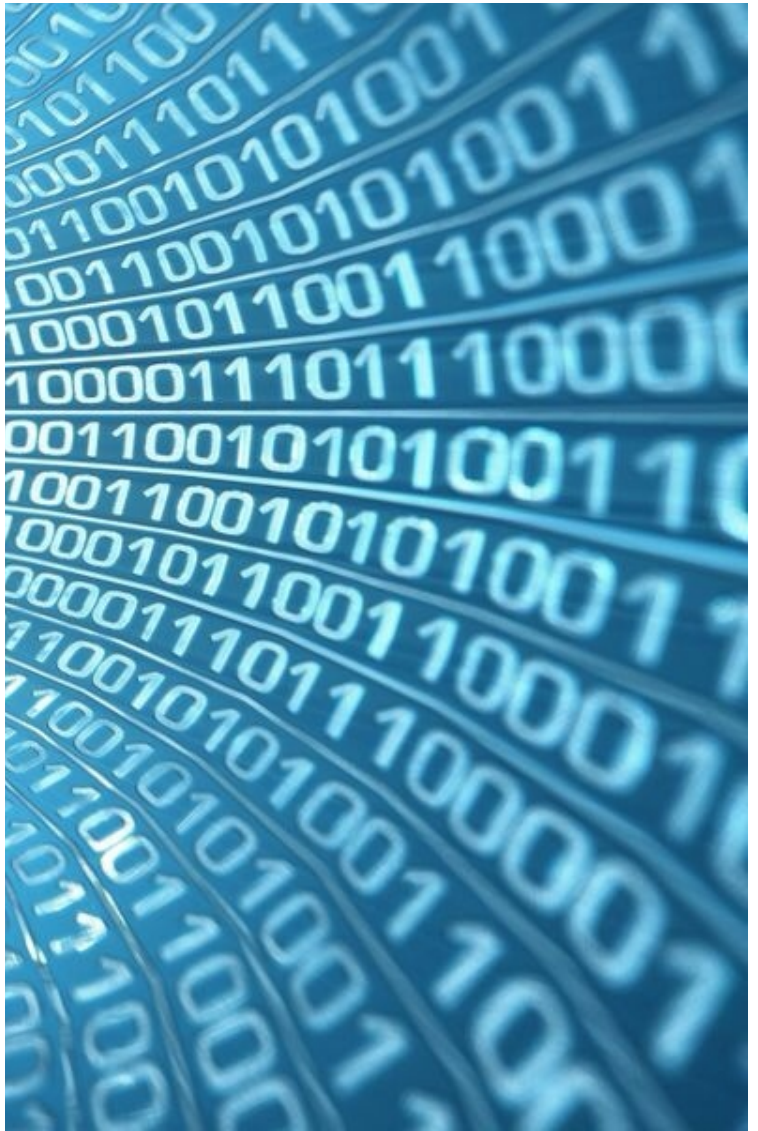
Wilkinson, et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*, http://dx.doi.org/10.1038/sdata.2016.18

4

404 The cosmic object you are looking for has disappeared beyond the event horizon.

Data Challenges in Space Studies

- How to make data useful and reusable?
- How to decide what data are worth keeping?
- How to balance incentives and benefits?
- How to steward data resources?
- Who pays for infrastructure?

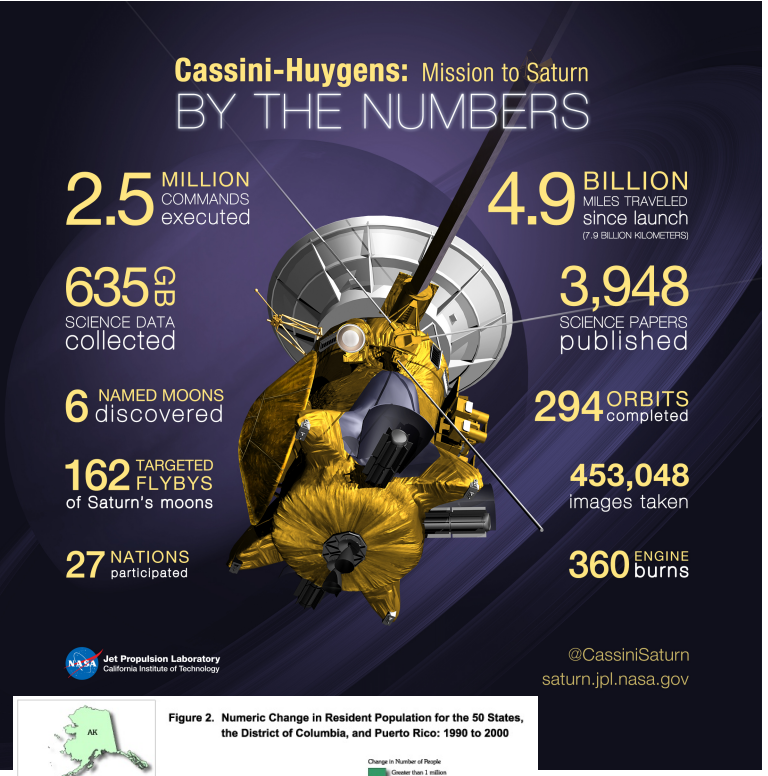


SETI INSTITUTE

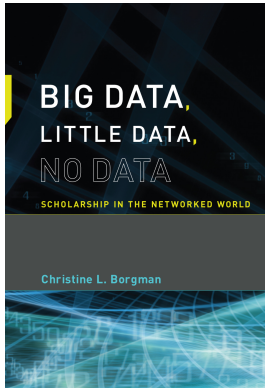
Our Scientists | Core Research | Education & Outreach | About Us | Donate Now

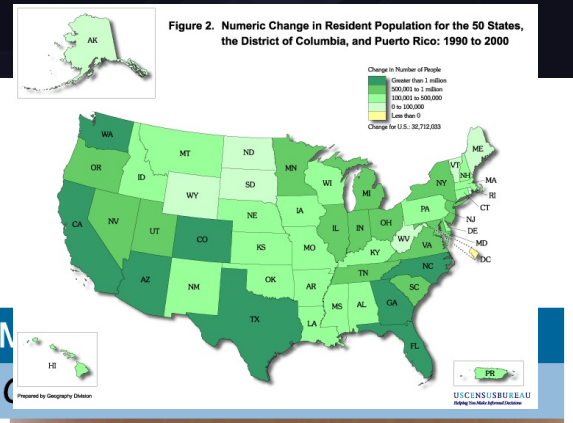
Could We Be Alone in the Cosmos?

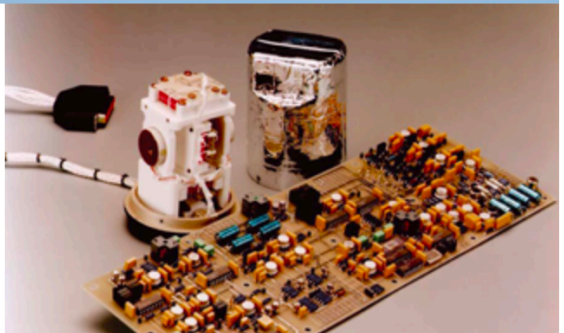
Data

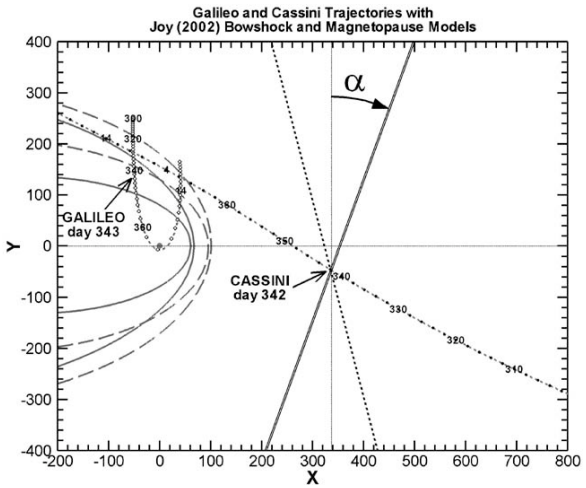


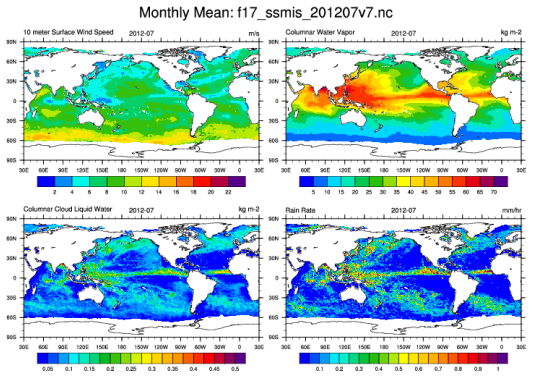
Data are representations of observations, objects, or other entities used as evidence of phenomena for the purposes of research or scholarship.









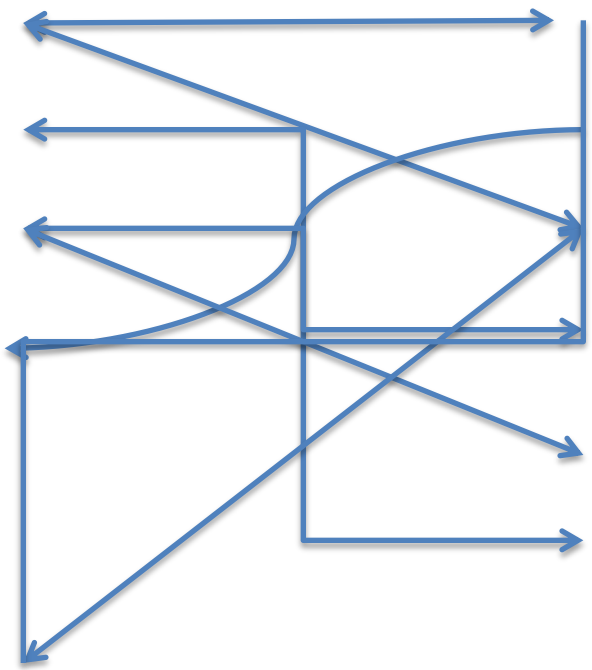


Kivelson, M. G., & Southwood, D. J. (2003). First evidence of IMF control of Jovian magnetospheric boundary locations: Cassini and Galileo magnetic field measurements compared. *Planetary and Space Science*, *51*(13), 891–898. https://doi.org/10.1016/S0032-0633(03)00075-8

Publications <–> Data: Mapping

- Article 1
- Article 2
- Article 3
- Article 4

- Article n

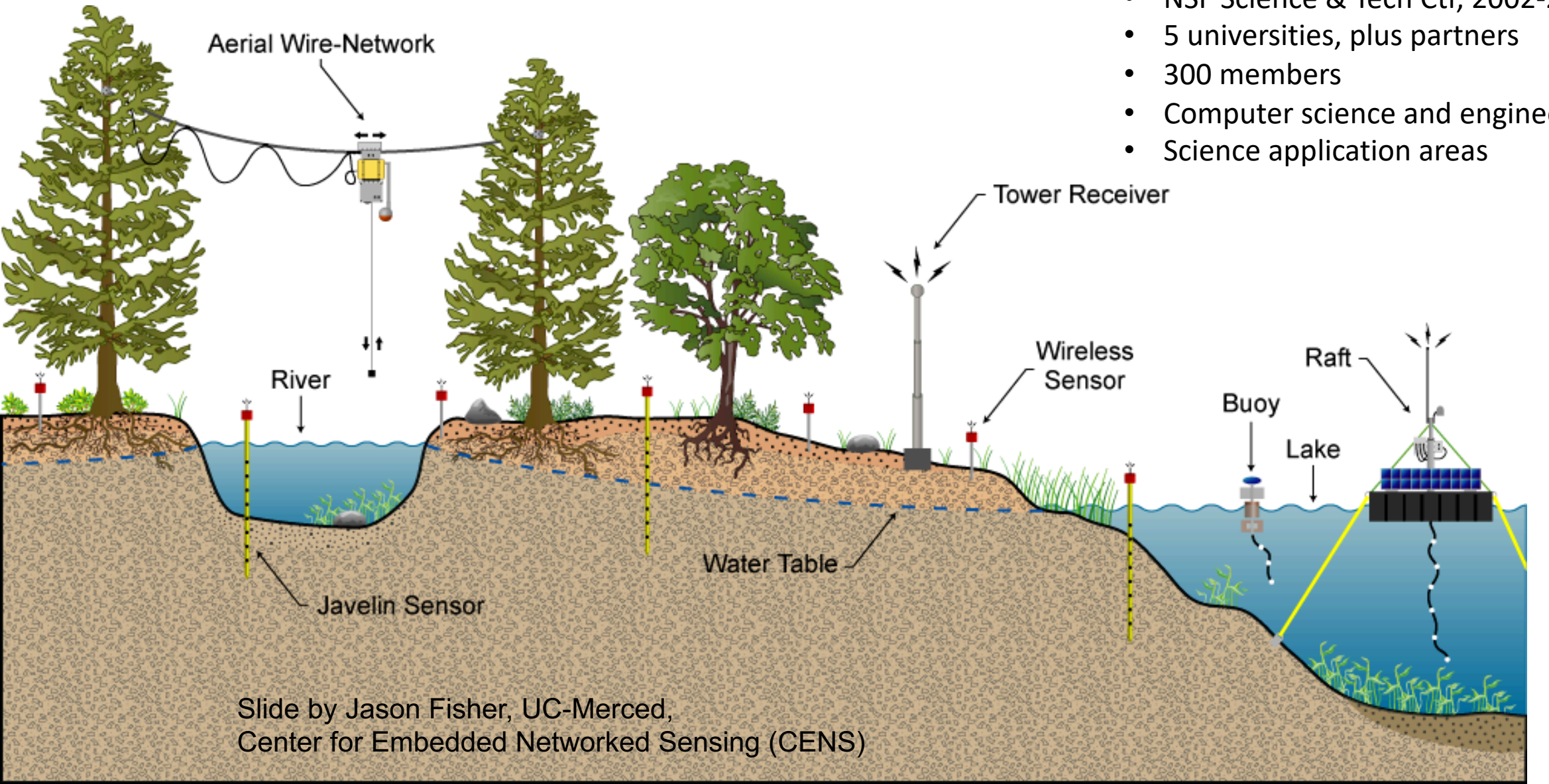


- Dataset time 1
- Dataset time 2
- Observation time 1
- Visualization time 3
- Community collection 1
- Repository 1

Data practices

Center for Embedded Networked Sensing

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

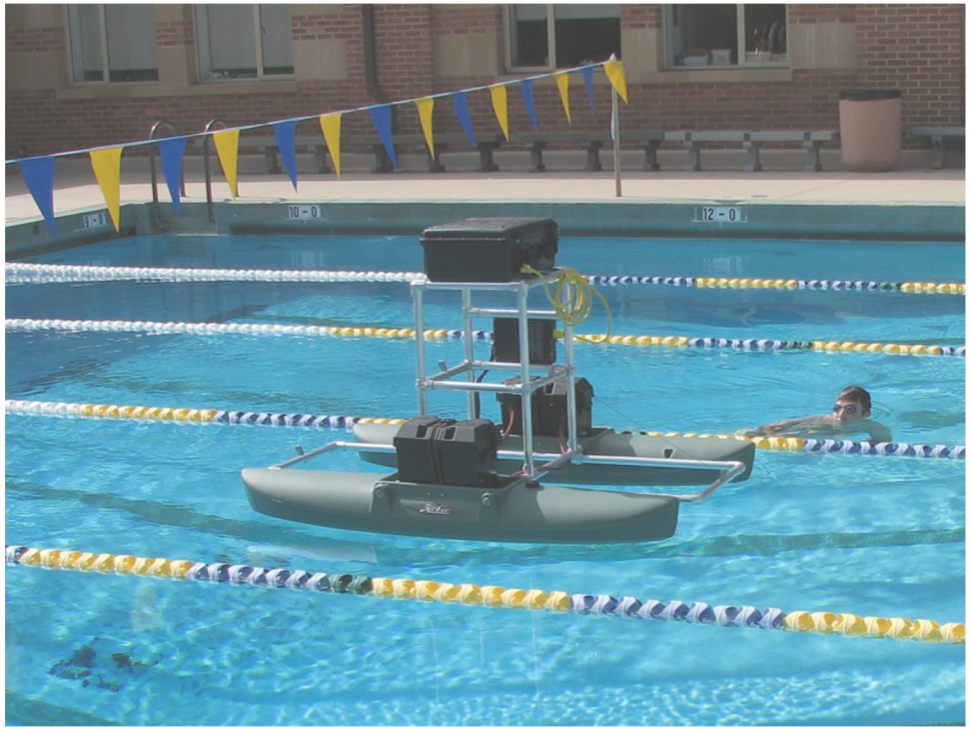


Slide by Jason Fisher, UC-Merced,
Center for Embedded Networked Sensing (CENS)

Science <–> Data

Engineering researcher:

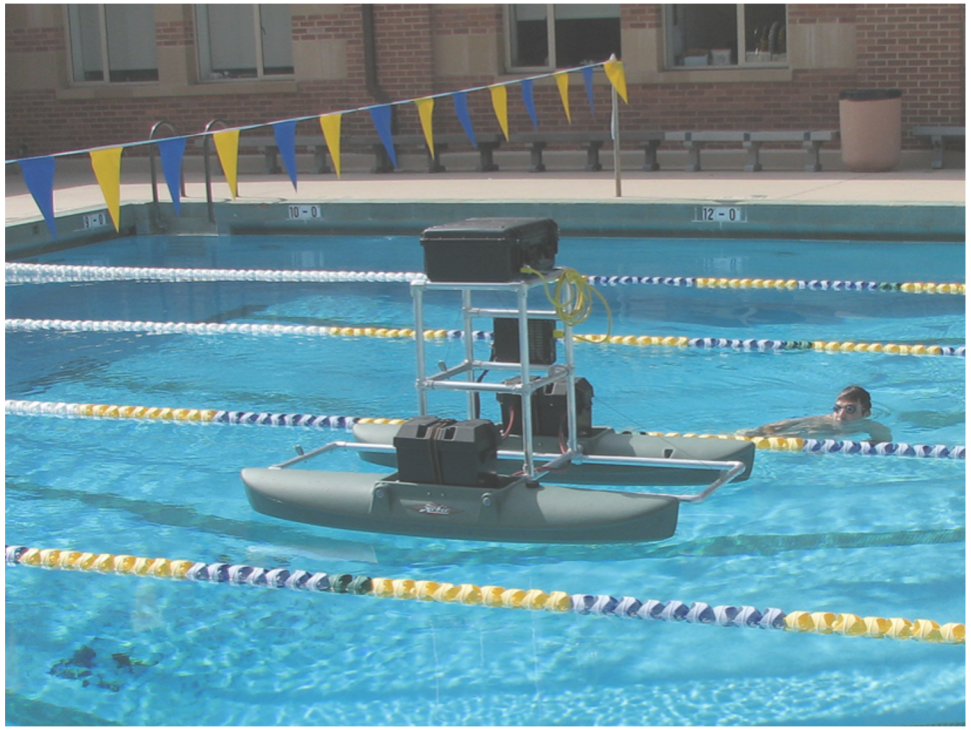
"Temperature is temperature."



CENS Robotics team

Science <–> Data

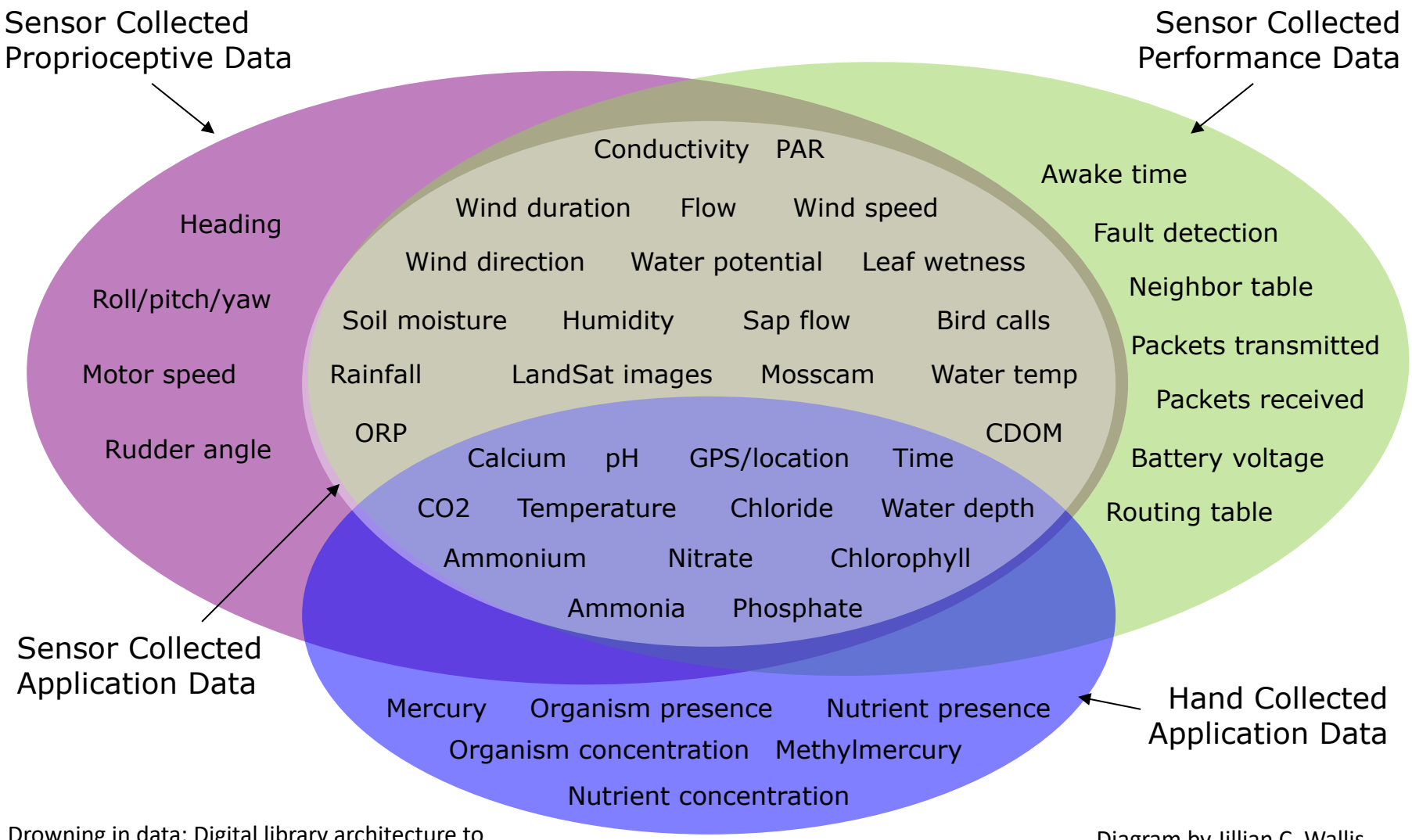
Engineering researcher:
"Temperature is temperature."



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.."

CENS data variation



Borgman, et al. (2007). Drowning in data: Digital library architecture to support scientific use of embedded sensor networks. JCDL

Diagram by Jillian C. Wallis

Deep Subseafloor Biosphere

- Center for Dark Energy Biosphere Investigations (C-DEBI)
- International Ocean Discovery Program (IODP)
- Microbial communities in the seafloor
- Highly multidisciplinary

Darch, P. T., & Borgman, C. L. (2016). Ship space to database: Emerging infrastructures for studies of the deep subseafloor biosphere. *PeerJ Computer Science*, *2*, e97. https://doi.org/10.7717/peerj-cs.97

Center for Dark Energy Biosphere Investigations

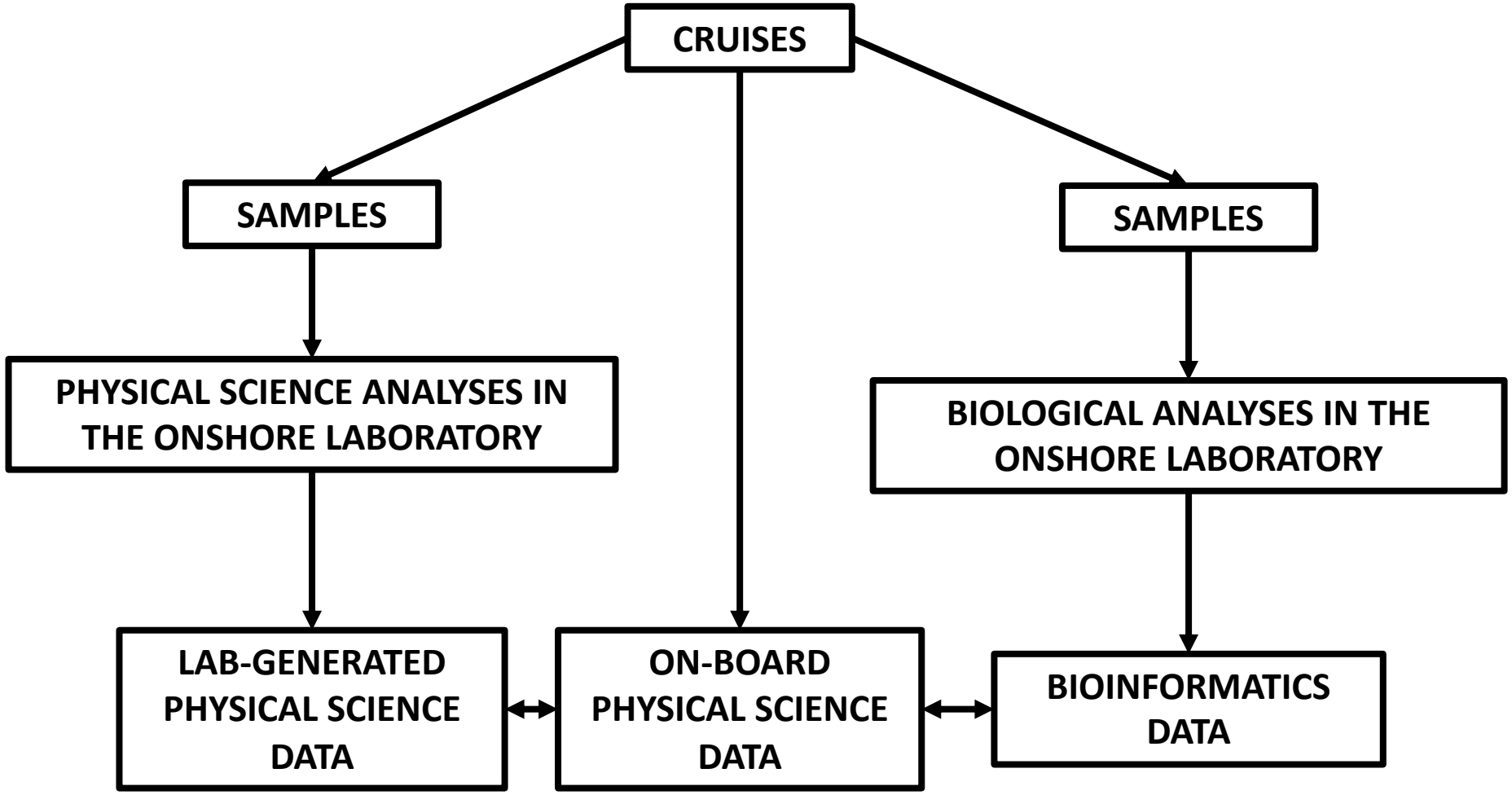
International Ocean Discovery Program
Iodp.tamu.org

Repository for seafloor cores. Photo: Peter Darch

- NSF Science & Tech Ctr, 2010-2020
- 20 universities, plus partners (35 institutions)
- 90 scientists
- Biological sciences
- Physical sciences

Slide by Peter T. Darch, UIUC

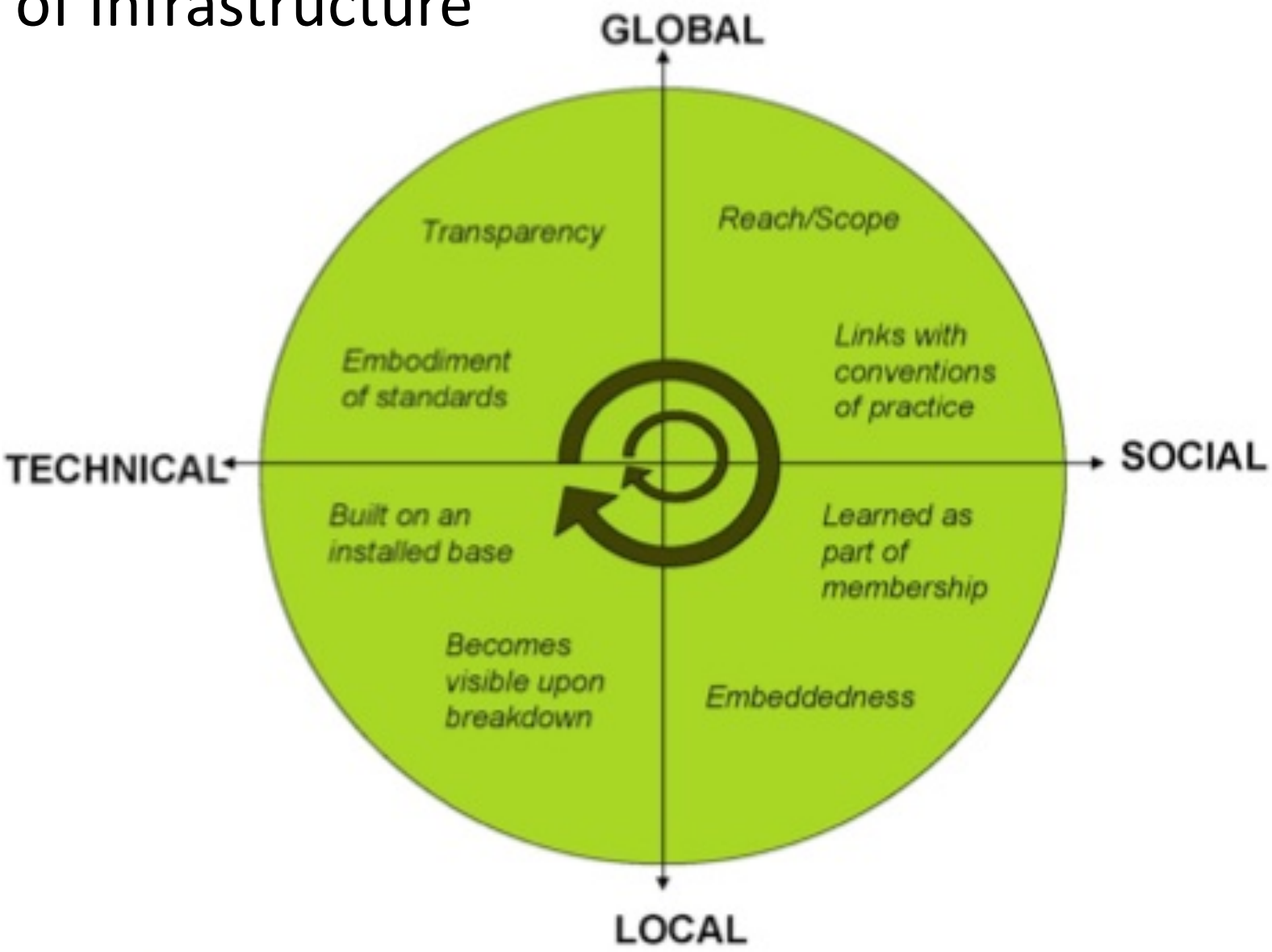
Data Diverge During Scientific Work



Slide by Peter T. Darch, UIUC

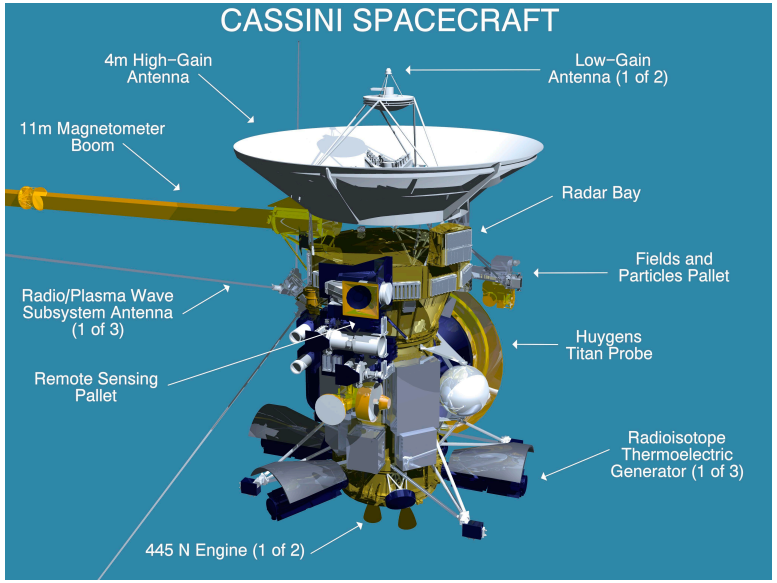
Infrastructure

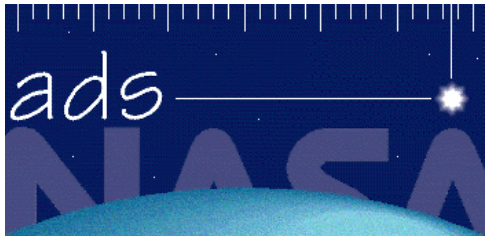
Dimensions of Infrastructure



Star, S. L. & Ruhleder, K. (1996). Steps toward an ecology of infrastructure: Design and access for large information spaces. Information Systems Research, 7(1): 111-134. Figure by Florence Millerand, from: Edwards, P. N., Jackson, S. J., Bowker, G. C. & Knobel, C. P. (2007). Understanding Infrastructure: Dynamics, Tensions, and Design. National Science Foundation: University of Michigan. NSF Grant 0630263.http://hdl.handle.net/2027.42/49353

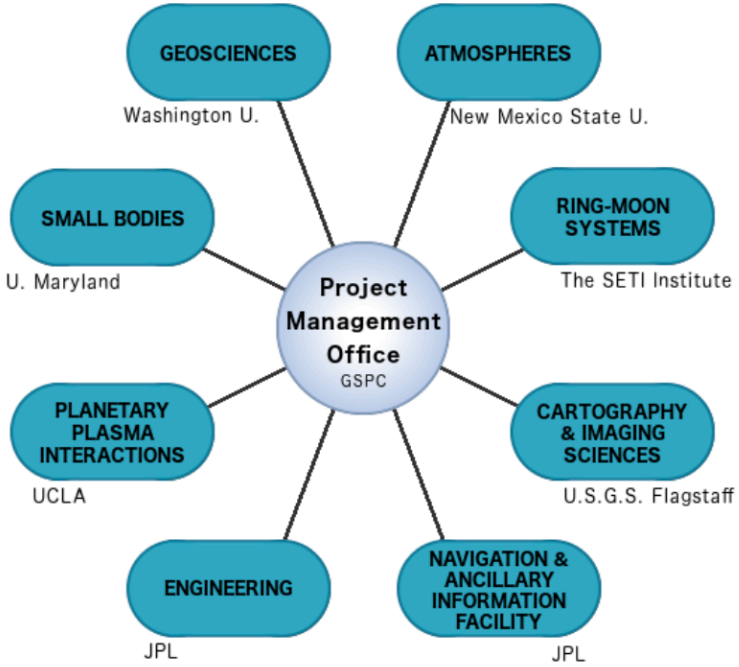
Global and Technical

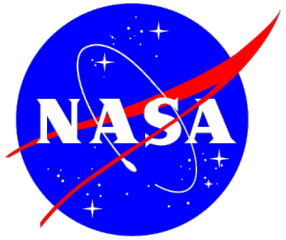


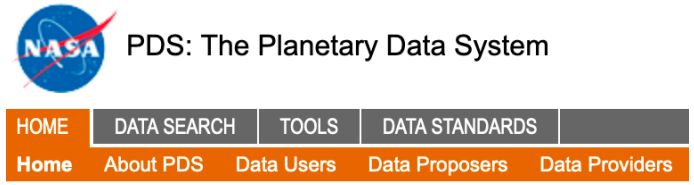


esa

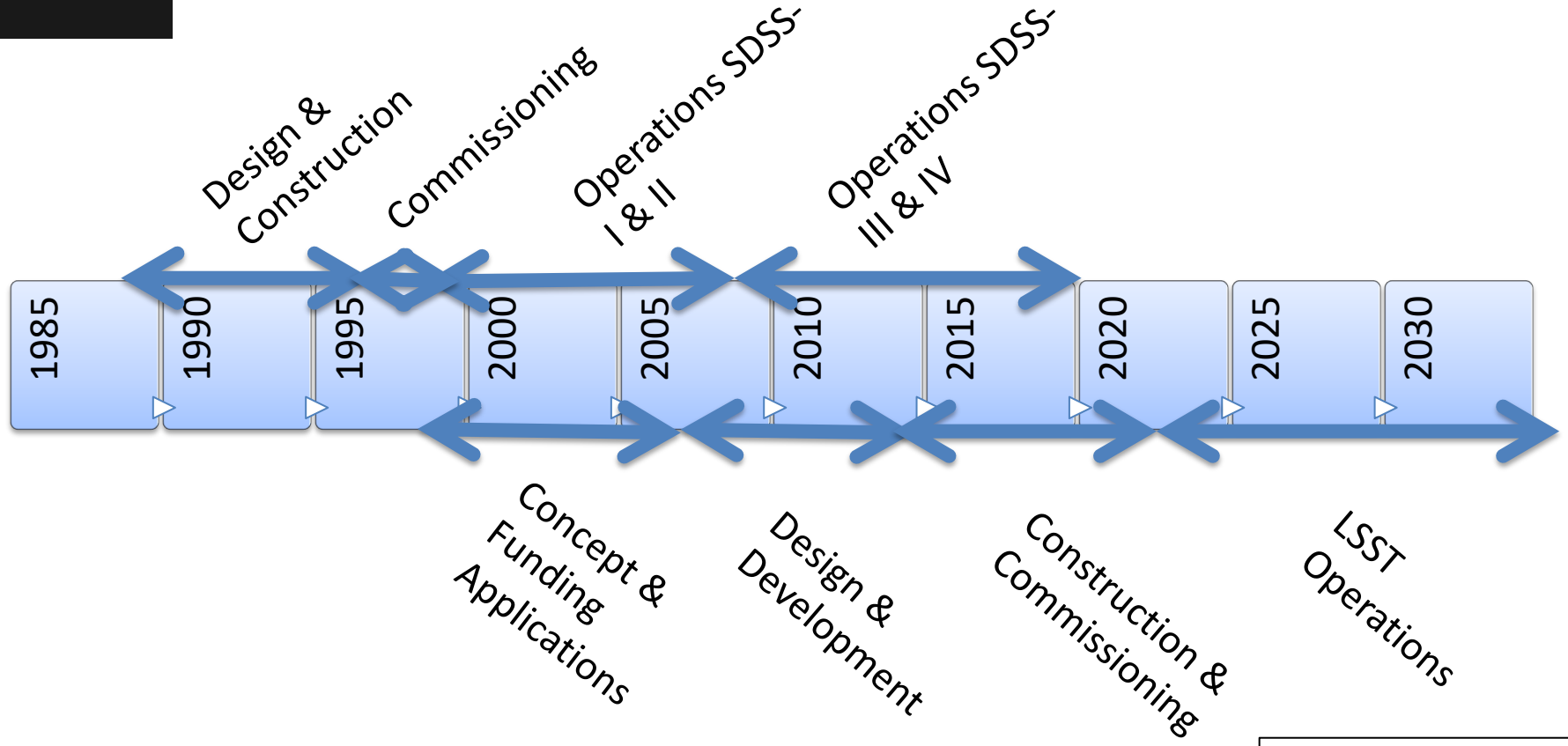
European Space Agency







Project Timelines



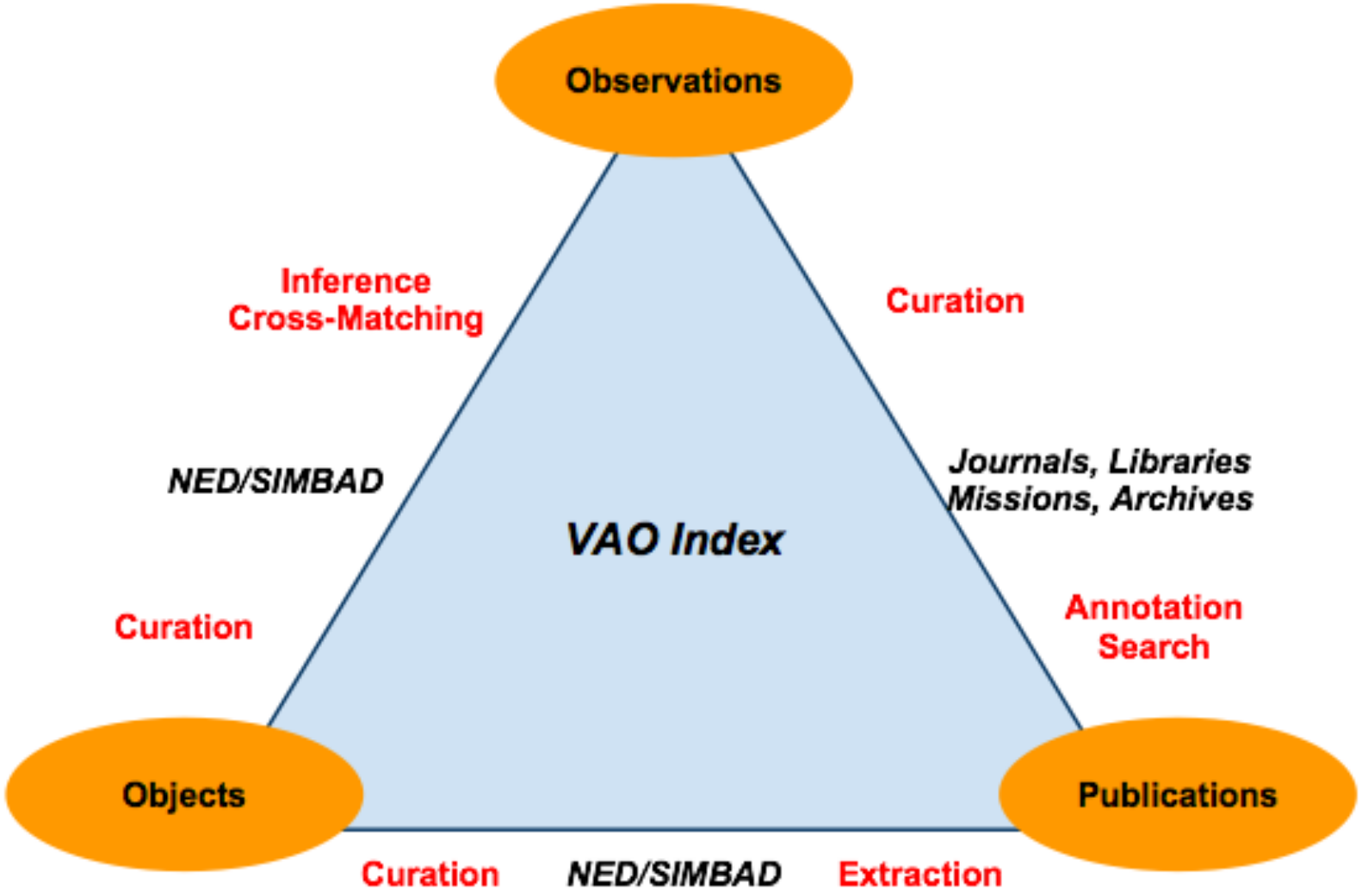
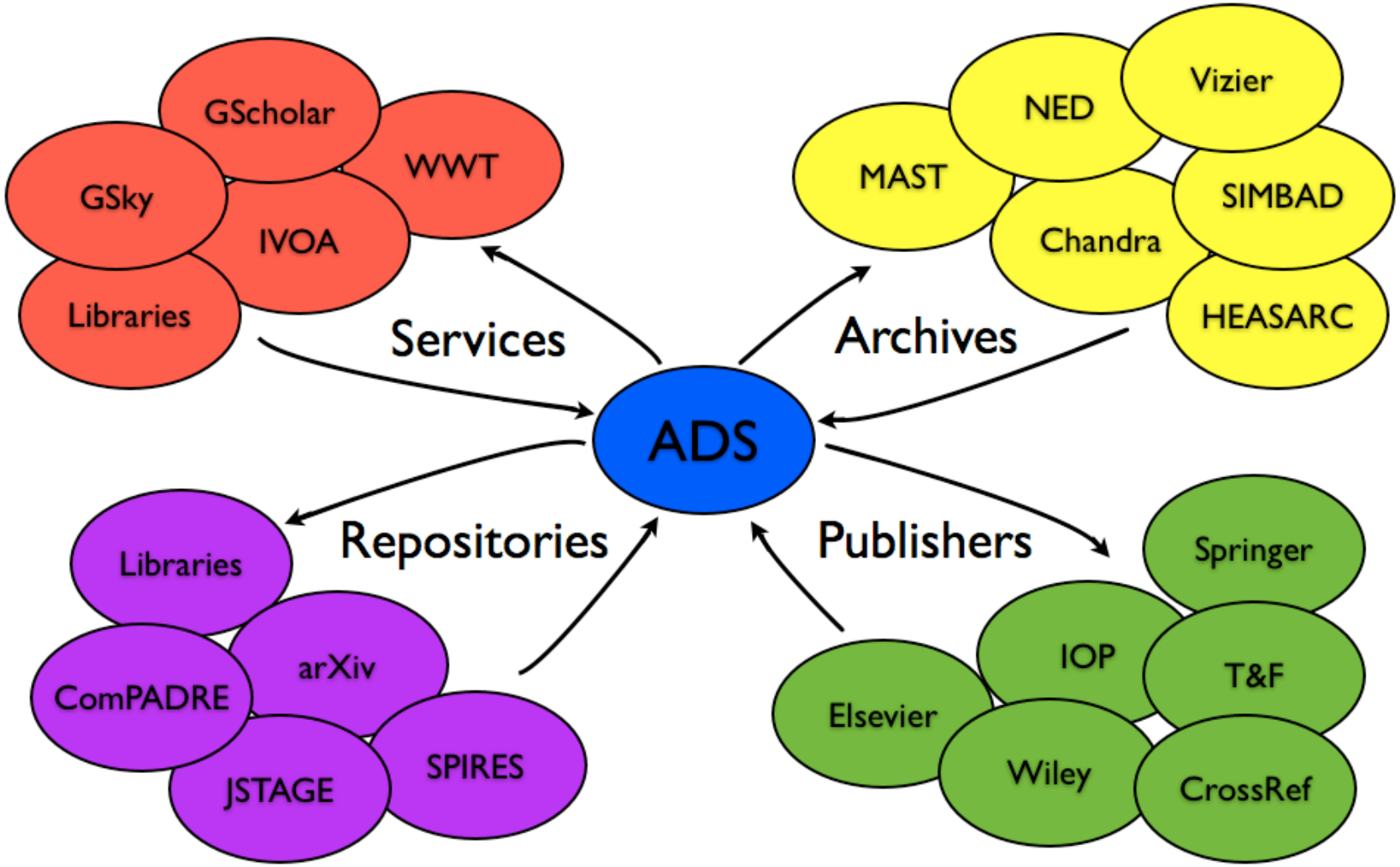


Figure 1. Relationships between Publications, Objects, Observations and the corresponding major actors in the curating process and their activities (in red).

Accomazzi, A., & Dave, R. (2011). Semantic Interlinking of Resources in the Virtual Observatory Era. *arXiv:1103.5958*

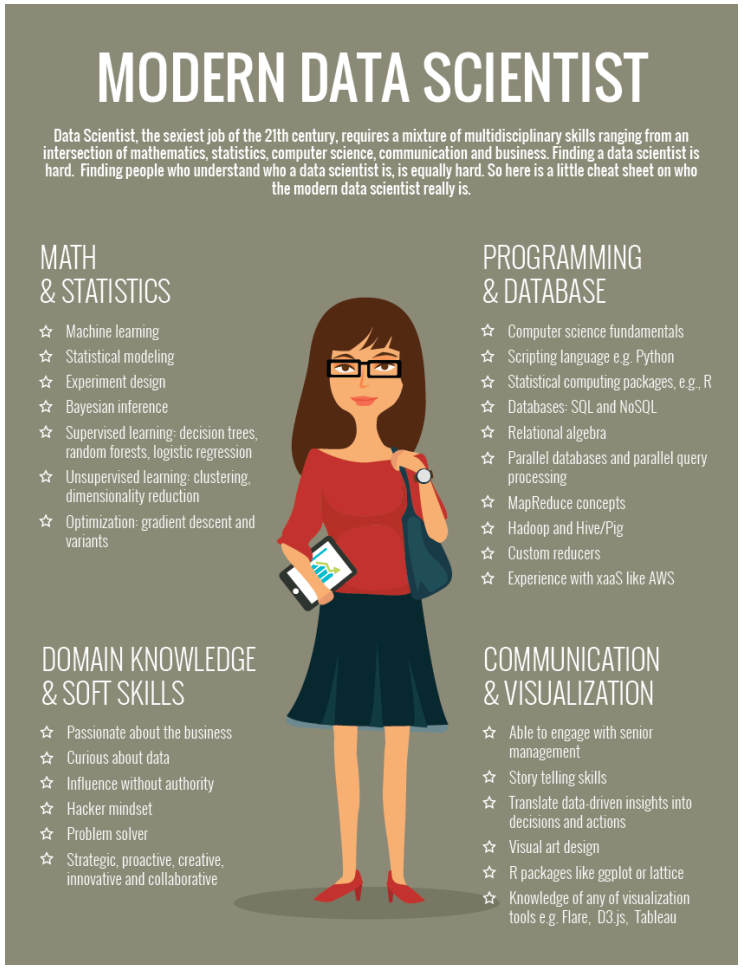
SAO/NASA Astrophysics Data System, 1993-

ADS Collaborators



Slide: Alberto Accomazzi, ADS

Local and Social



https://github.com/okulbilisim/awesome-datascience

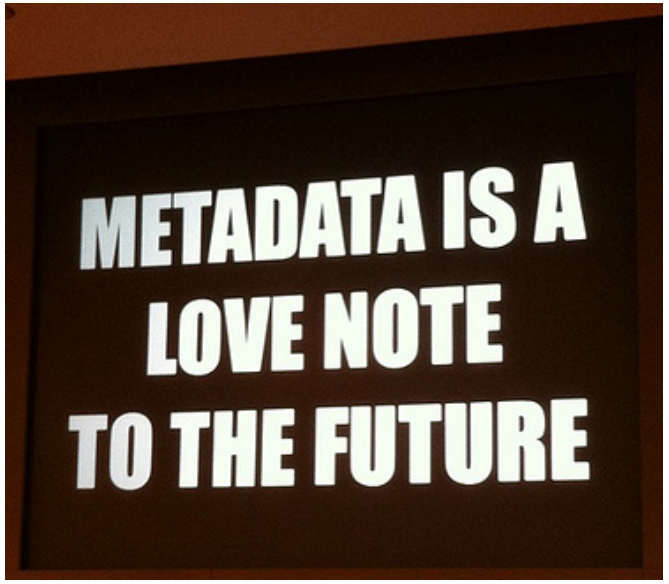
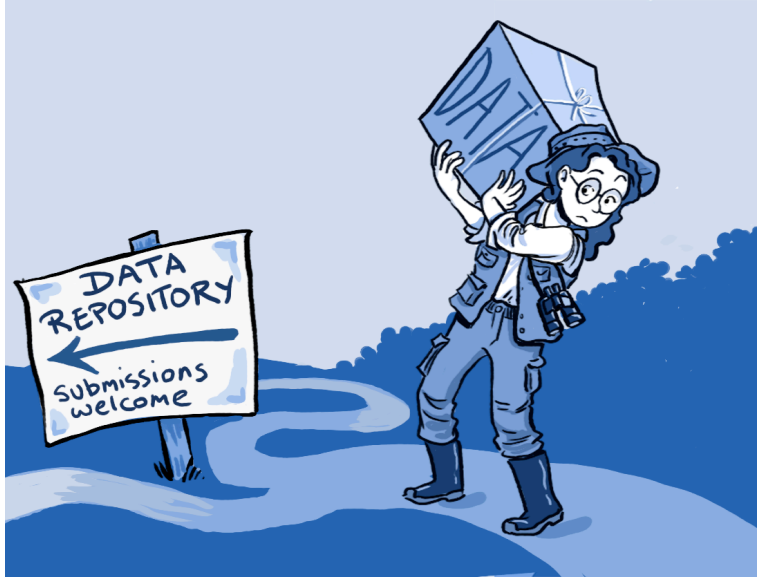


Photo by @kissane; presentation by Jason Scott (@textfiles)

CC Sean MacEntee, Flickr



https://en.wikipedia.org/wiki/Data_sharing

Lack of incentives to share data

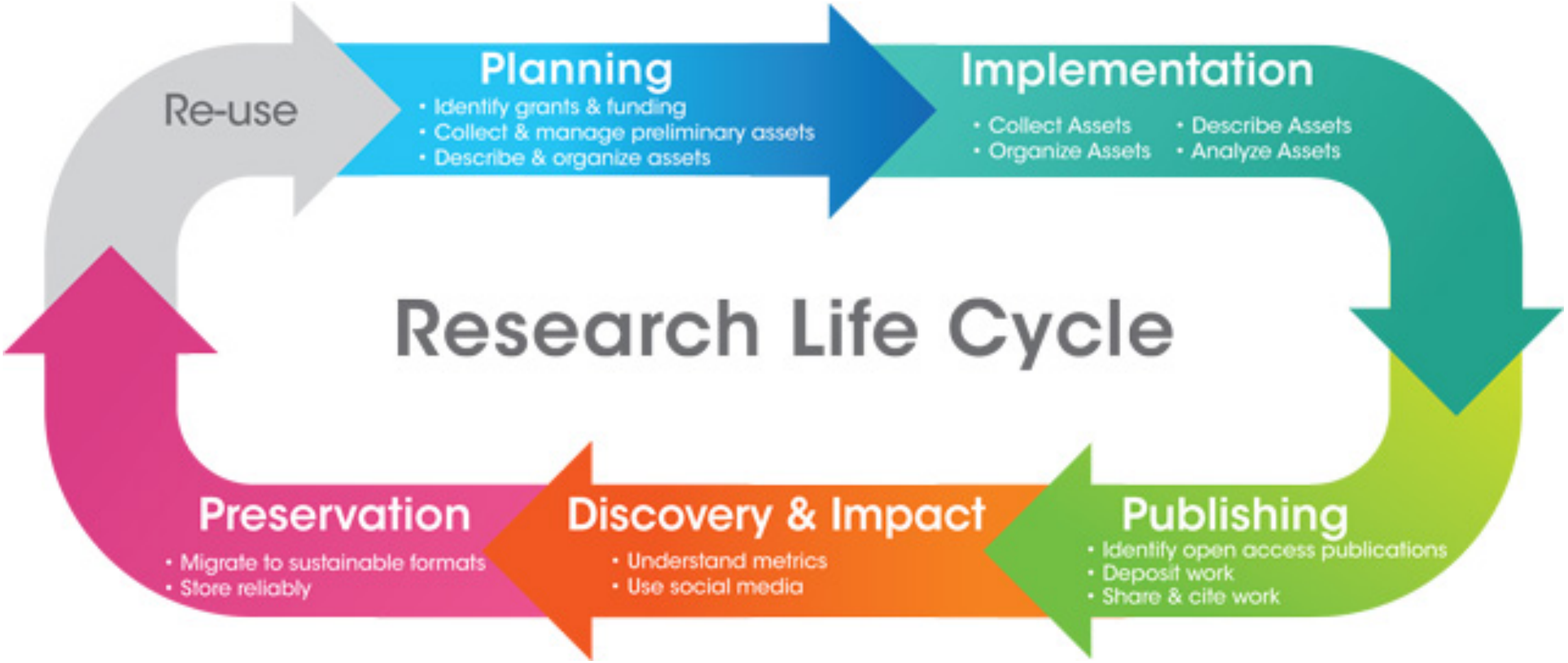
- Labor to document data
- Benefits to unknown others
- Competition
- Control
- Confidentiality
- Lack of expertise and staff
- Lack of sustainability…

Image source: www.buildingsrus.co.uk/.../ target1.htm

	Comparative Data Reuse <–> Integrative Data Reuse	
Goal	"Ground truthing:" calibrate, compare, confirm	Analysis: identify patterns, correlations, causal relationships
Example	Instrument calibration, sequence annotation, review summary-level data	Meta-analyses, novel statistical analyses
Frequency	Frequent, routine practice	Rare, emergent practice
Interpretation	Interactional expertise, 'knowledge that'	Contributory expertise, 'knowledge how,' tacit knowledge

Pasquetto, I. V., Borgman, C. L., & Wofford, M. F. (2019). Uses and reuses of scientific data: The data creators' advantage. *Harvard Data Science Review,* 1:2, https://hdsr.mitpress.mit.edu/.

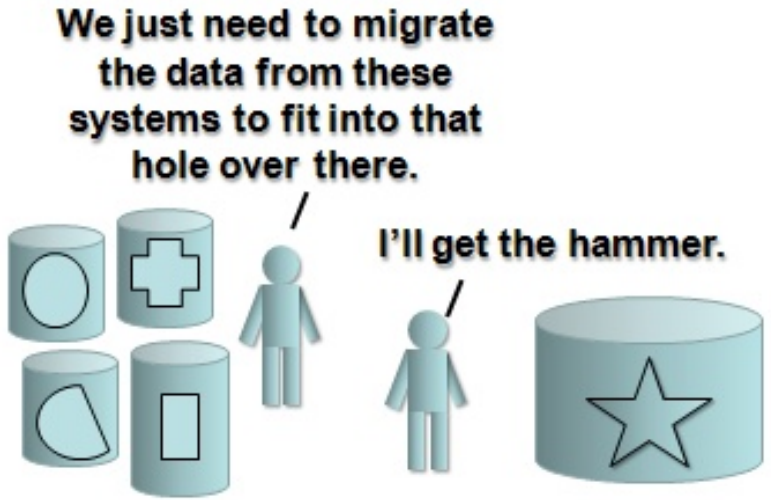
Data creation and reuse: The Ideal



http://www.lib.uci.edu/dss/images/lifecycle.jpg

Data Stewardship: The Reality

http://www.information-age.com/cloud-computing-pharmaceutical-industry-123462676/



http://www.datamartist.com/data-migration-part-1-introduction-to-the-data-migration-delema

Mount Wilson Solar Observatory, 2017

Getty Research Institute

http://gsa.rice.edu/

Graduate students

https://med.nyu.edu/our-community/life-nyu-school-medicine/life-postdoc

Post-doctoral fellows

Infrastructure: Durability

Photos by C.L. Borgman

- Collaboration and openness
- International coordination
- Long-term value of data
- Agreed standards
  - Units of measurement
  - Coordinate systems
  - Data structures
- Shared resources
  - Missions, instruments
  - Data archives
  - Tools and technologies

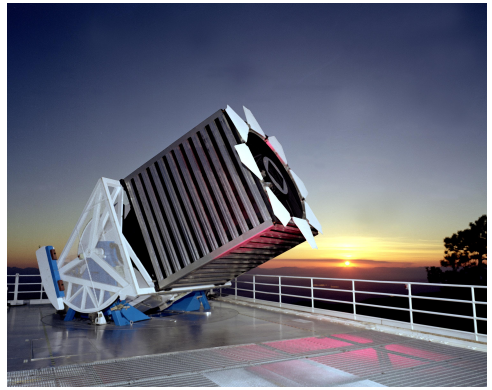
Infrastructure: Fragility

- Investments in data stewardship
  - Mission, instrument
  - Type of research
    - Space-based vs. ground based
    - Large missions vs. observing proposals
    - Shared vs. custom instruments
- Access to data
  - Public archives
  - Local websites
  - Derived data
- Curation investments
  - Open source
  - Proprietary tools
  - Local pipelines, tools, scripts

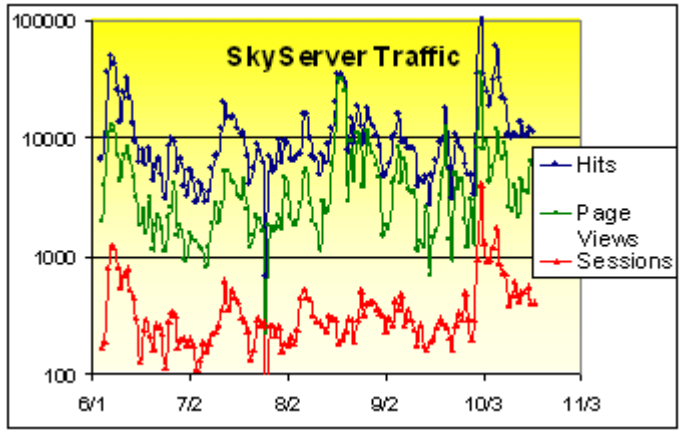
Photos by C.L. Borgman

Discussion

Scientific Data and Infrastructure



- Infrastructures are fragile
- Visible infrastructure
  - Instruments
  - Institutions
- Invisible infrastructure
  - Data, metadata, provenance…
  - Information work
- Interdisciplinary science
  - Global science
  - Local practices



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

LSST All Hands Meeting, August 2014, Arizona State University. Arrow to Peter Darch

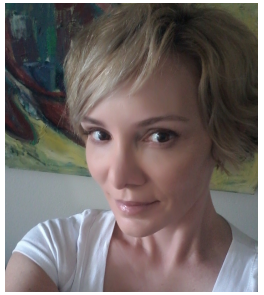
Data, Infrastructure, and Stewardship

- Whose data?
  - Global, comparative, fungible
  - Local, integrative, specific
- Whose infrastructure?
  - Funders, universities, companies
  - Individual investigators
- Whose stewardship?
  - Maintain collections, models, instruments, technology, code…
  - Invest in people, skills, collaborations

Acknowledgements

UCLA Center for Knowledge Infrastructures


Christine Borgman


Bernie Boscoe


Peter Darch


Milena Golshan


Irene Pasquetto


Michael Scroggins


Cheryl Thompson


Morgan Wofford