Title
Big Data, Little Data, or No Data? A Social Science Perspective on Data Science [Presentation slides]

Permalink
https://escholarship.org/uc/item/2911049g

Author
Borgman, Christine L.

Publication Date
2021-03-19

Supplemental Material
https://escholarship.org/uc/item/2911049g#supplemental

Big Data, Little Data, or No Data?
A Social Science Perspective on Data Science

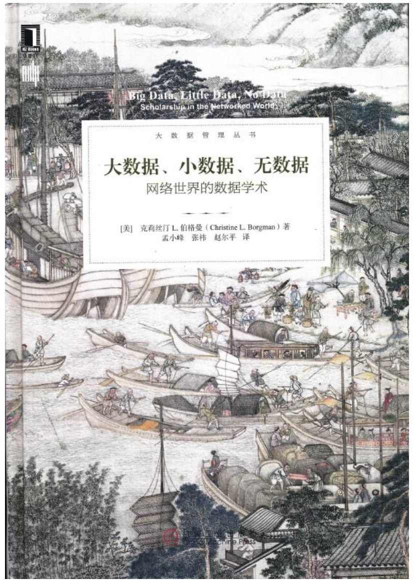
Christine L. Borgman

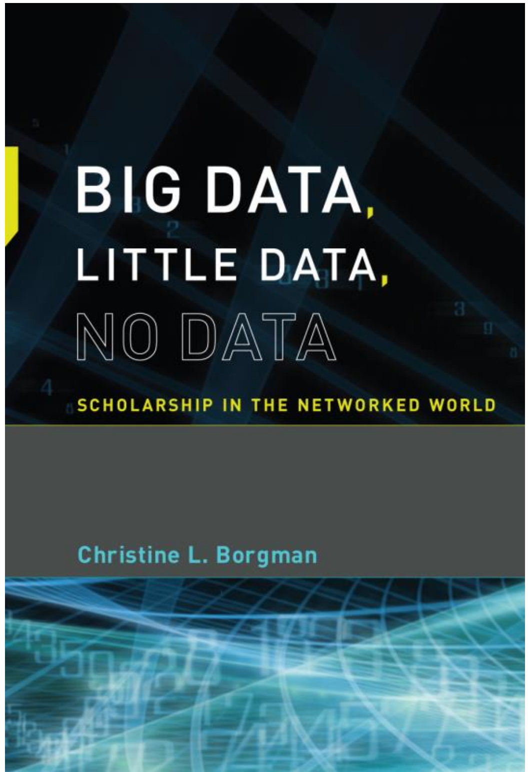
Distinguished Research Professor & Presidential Chair in Information Studies, Emerita

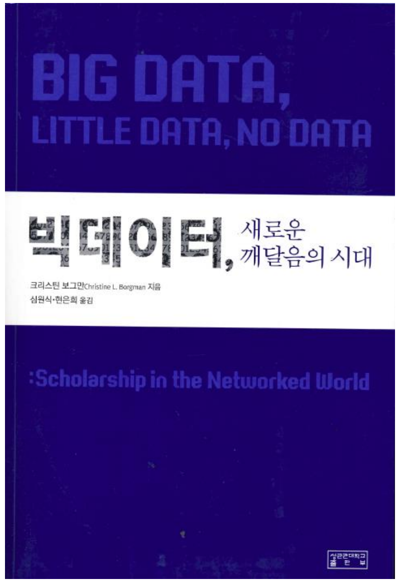
Director, UCLA Center for Knowledge Infrastructures, @scitechprof

Women in Data Science (WiDS), University of Virginia

Keynote Presentation, 19 March 2021

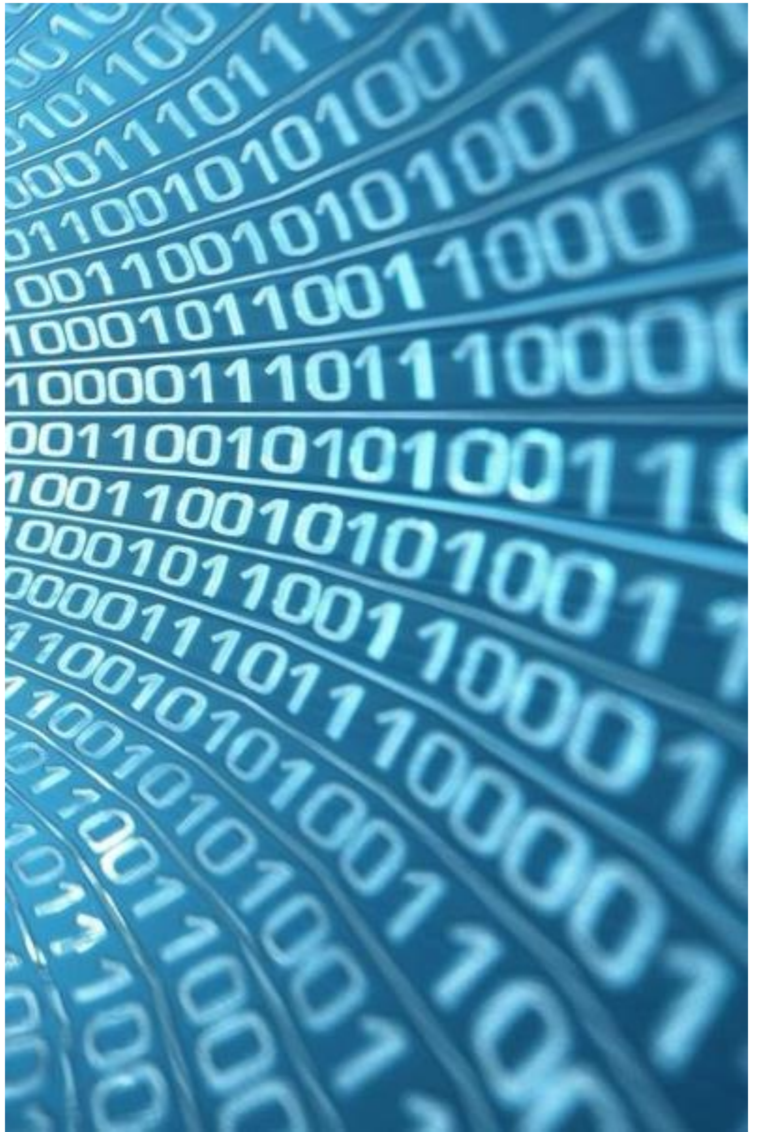






Data Challenges in (Data) Science

- 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?



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

Open Data Practices

dbGaP GENOTYPES and PHENOTYPES

DRYAD

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

NASA PDS: The Planetary Data System

HOME | DATA SEARCH | TOOLS | DATA STANDARDS

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

ICPSR A PARTNER IN SOCIAL SCIENCE RESEARCH

INTER-UNIVERSITY CONSORTIUM FOR POLITICAL AND SOCIAL RESEARCH

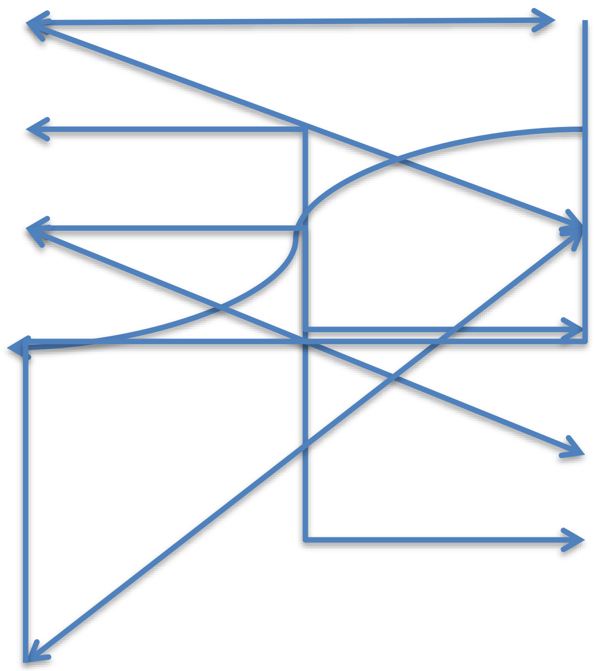
Barbara A. MIKULSKI ARCHIVE FOR SPACE TELESCOPES

UNIVERSITY OF CALIFORNIA dash

HARVARD Dataverse

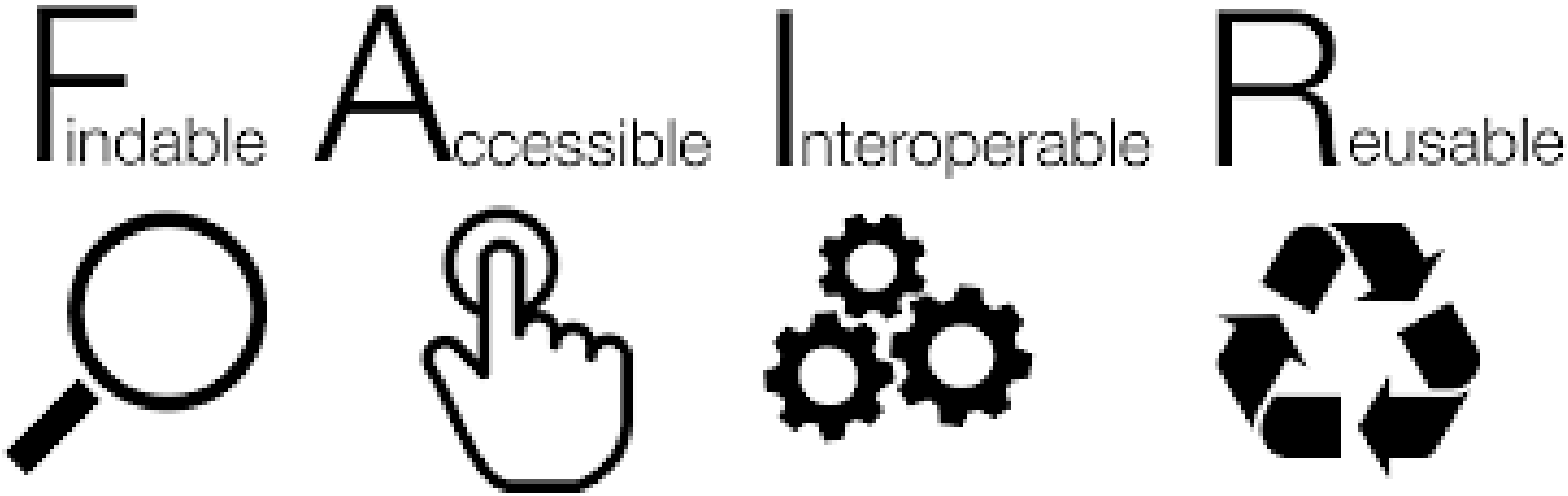
Publications <–> Data: Mapping

- 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 Stewardship: The Ideal

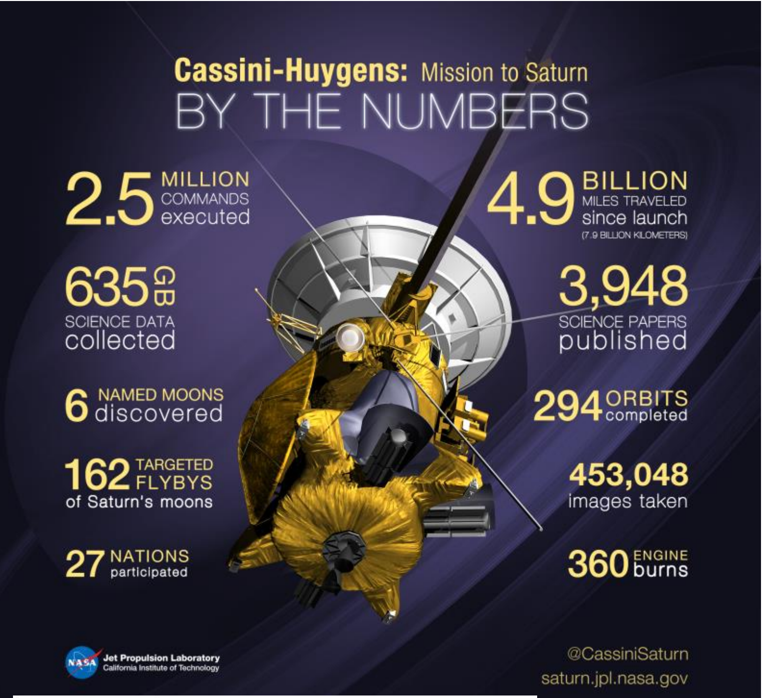


Wilkinson, et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*, http://dx.doi.org/10.1038/sdata.2016.18

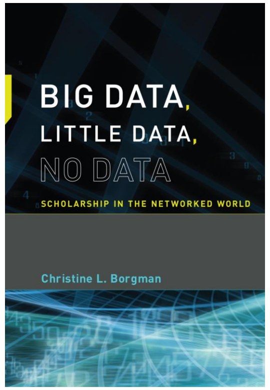
6

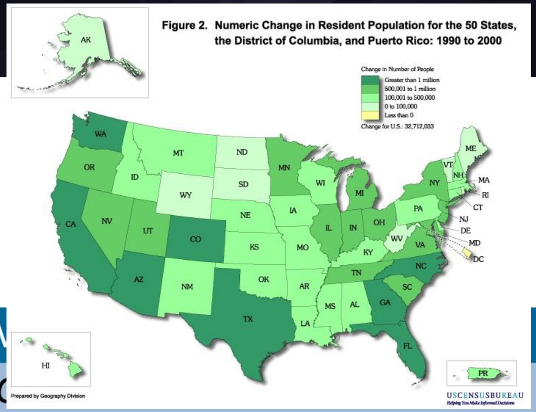

404
The cosmic object you are looking for has disappeared beyond the event horizon.

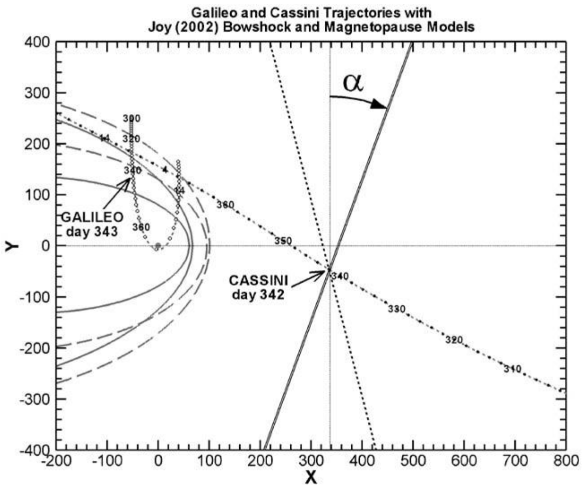
Data

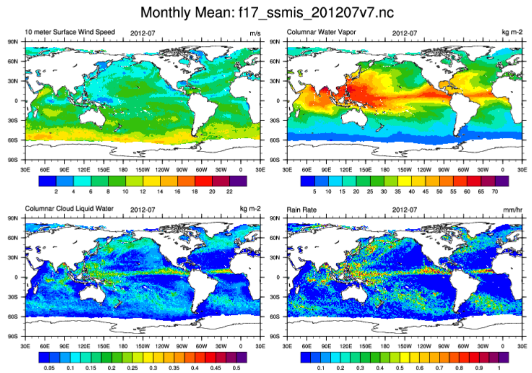


Data are representations of observations, objects, or other entities used as evidence of phenomena for the purposes of research or scholarship.





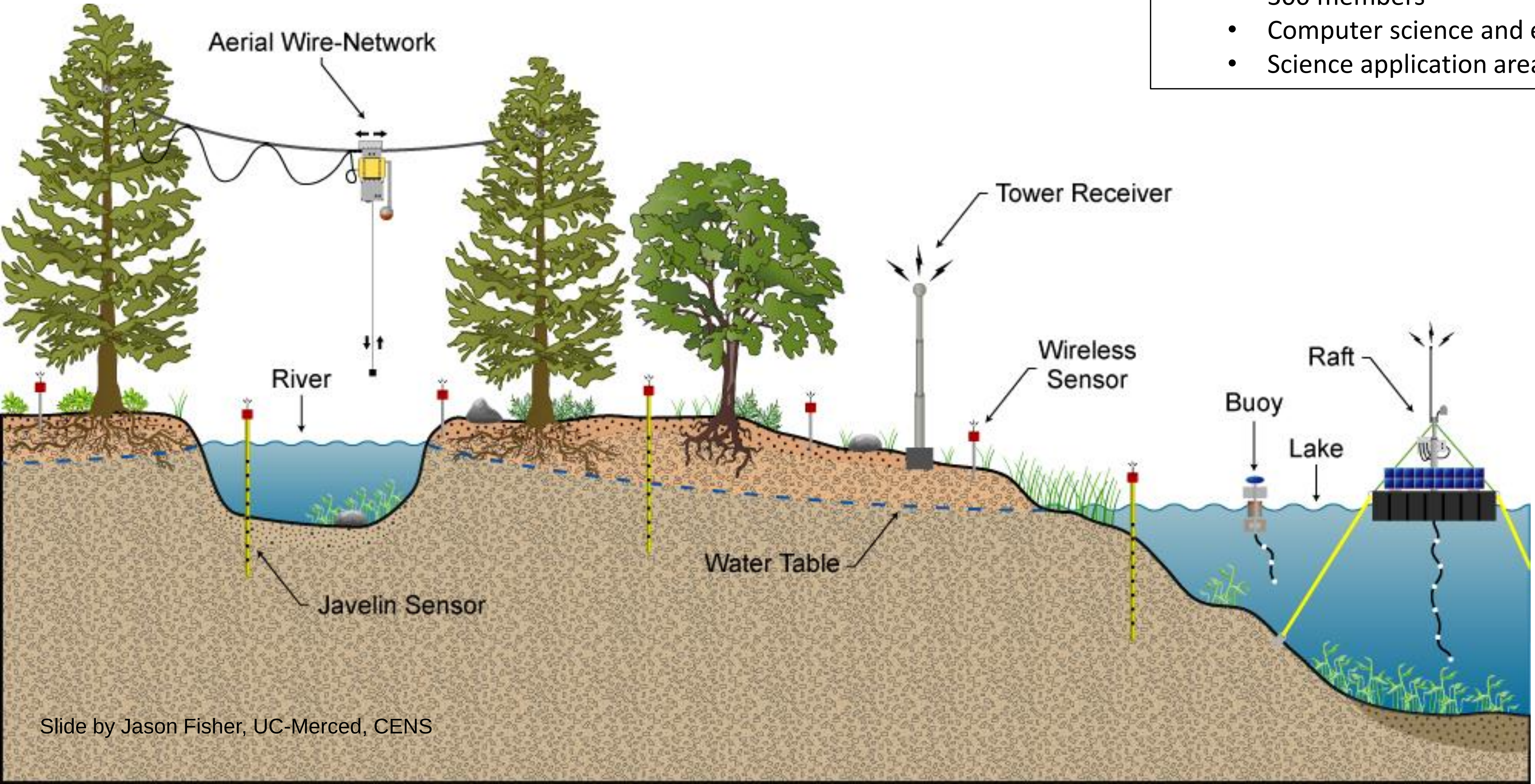




Kivelson, M. G., & Southwood, D. J. (2003). First evidence of IMF control of Jovian magnetospheric boundary locations: Cassini and Galileo magnetic field measurements compared. *Planetary and Space Science*, *51*(13), 891–898. https://doi.org/10.1016/S0032-0633(03)00075-8

Center for Embedded Networked Sensing

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



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.."*

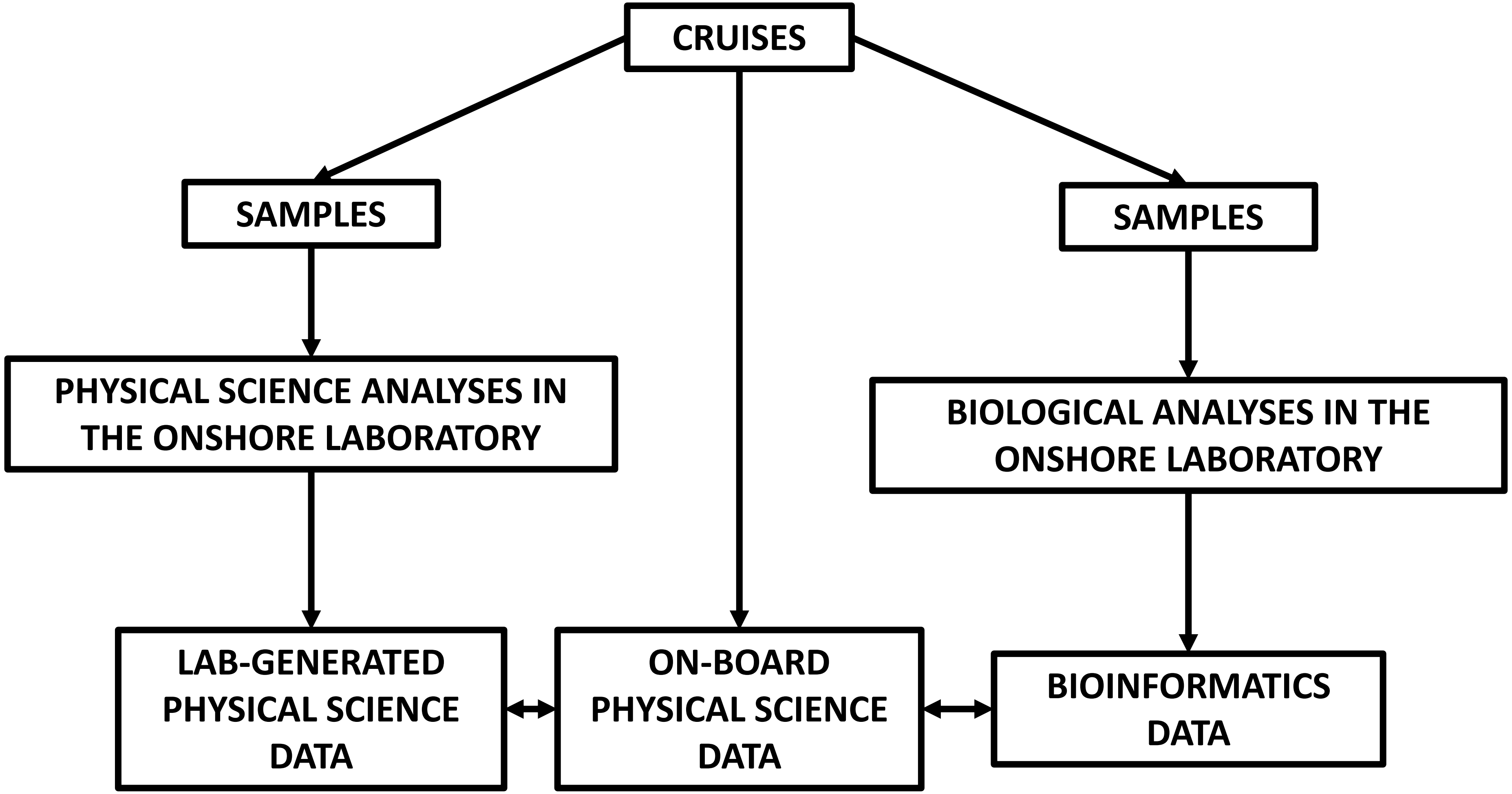
Center for Dark Energy Biosphere Investigations

International Ocean Discovery Program
Iodp.tamu.org

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

Repository for seafloor cores. Photo: Peter Darch

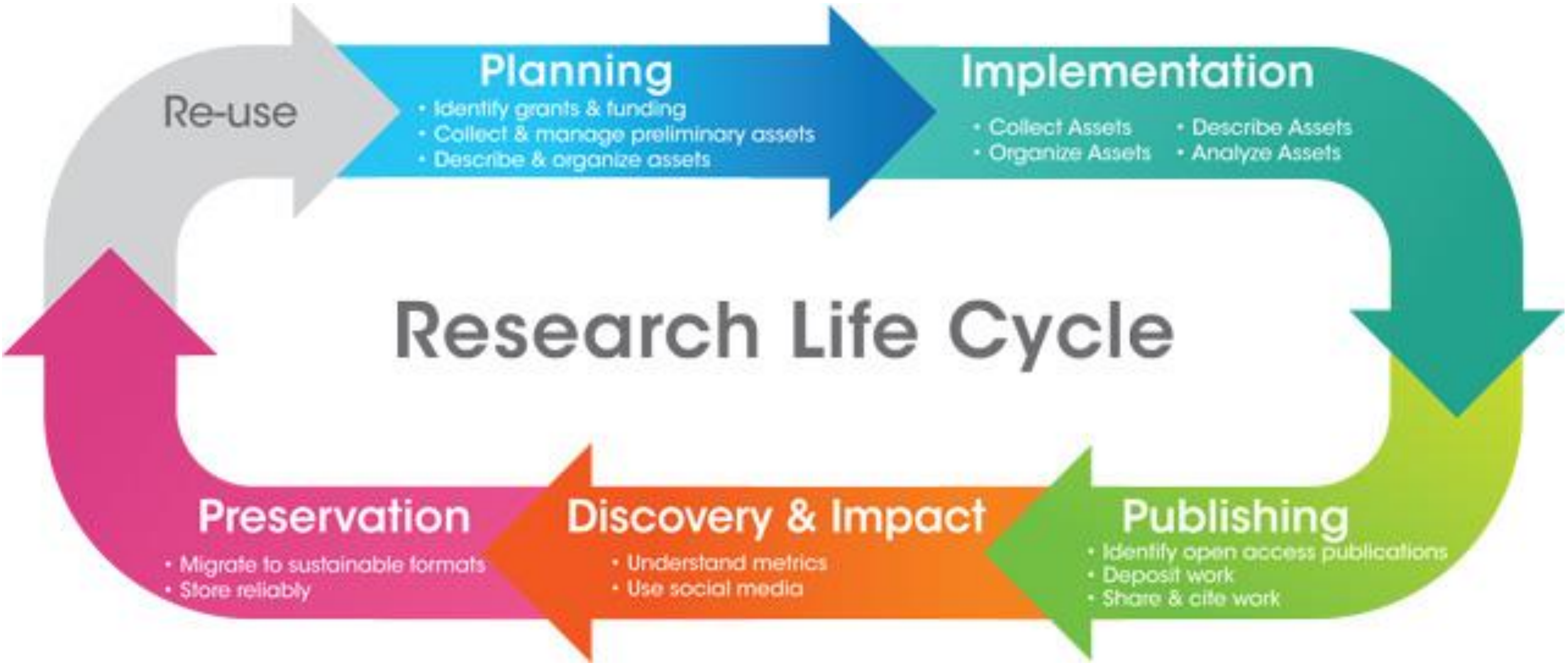
Data Diverge During Scientific Work



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

Data Practices

Data creation and reuse: The Ideal



Borgman, C. L. (2019). The lives and after lives of data. *Harvard Data Science Review*, *1*(1). https://doi.org/10.1162/99608f92.9a36bdb6

Image credit: UC Irvine Libraries

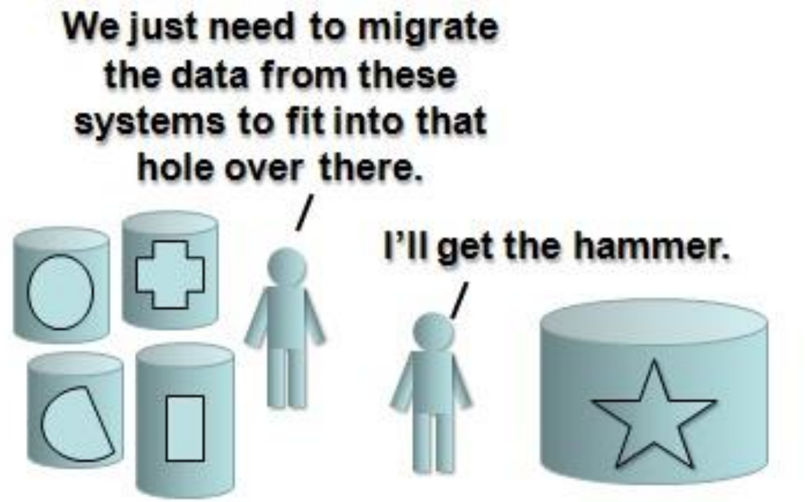
Lack of incentives to share data

- 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

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/

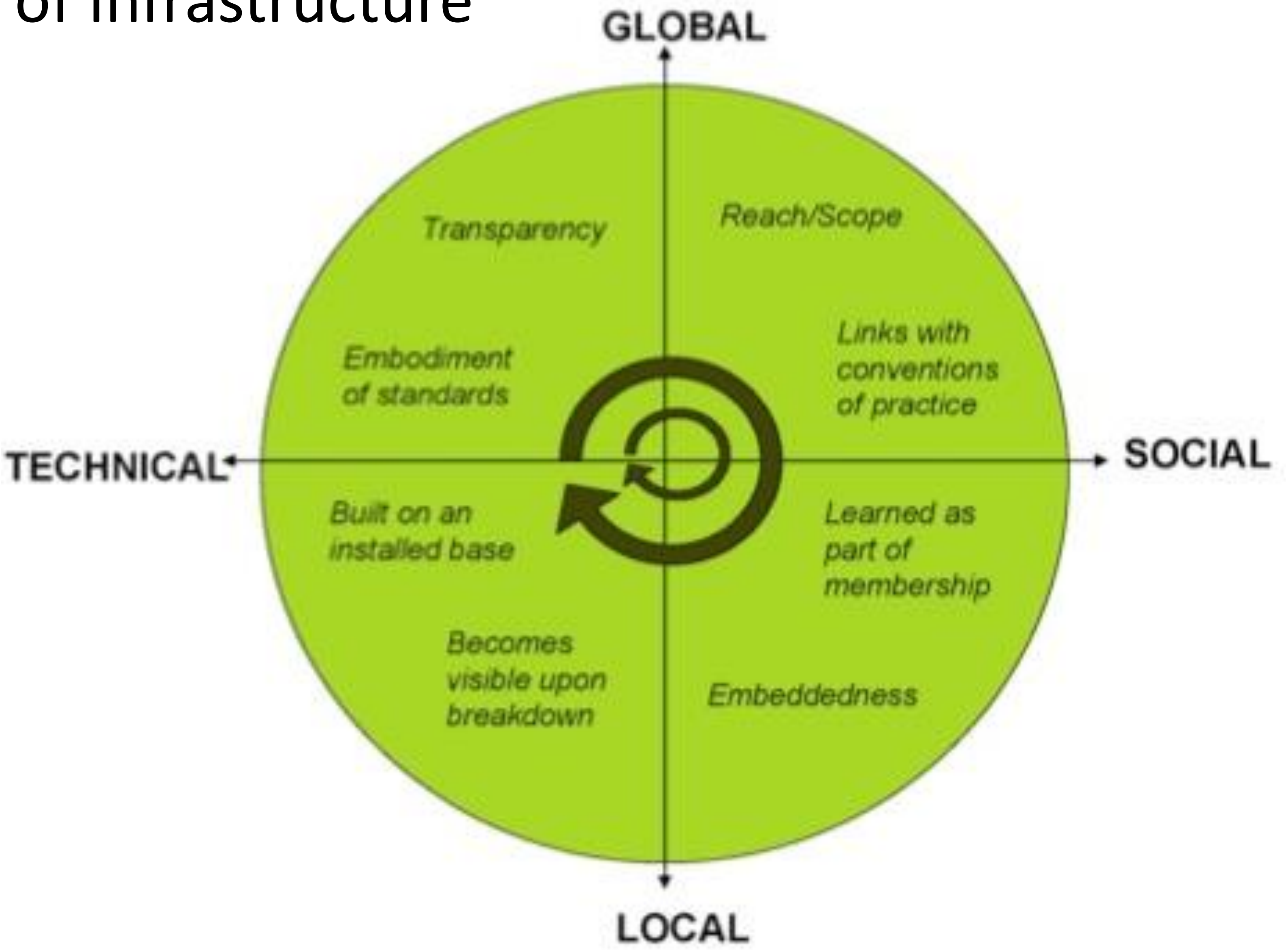
Graduate students

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

Post-doctoral fellows

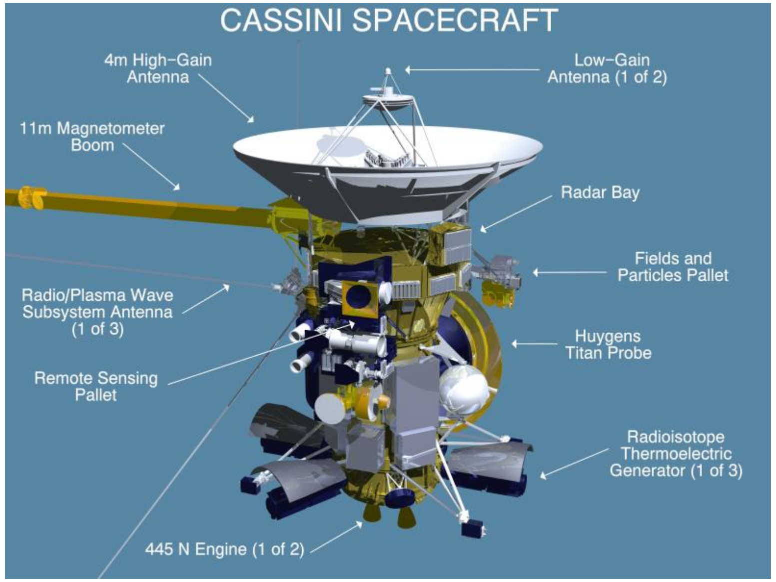
Infrastructure

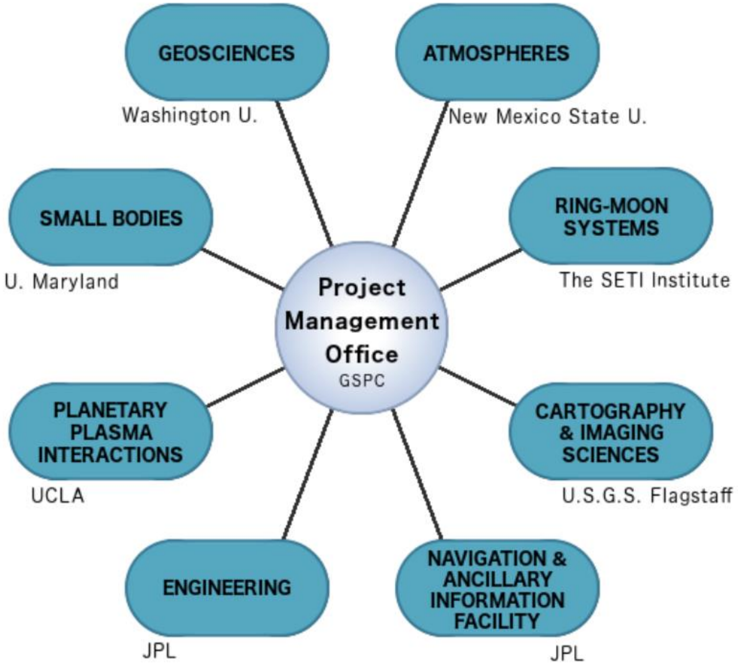
Dimensions of Infrastructure

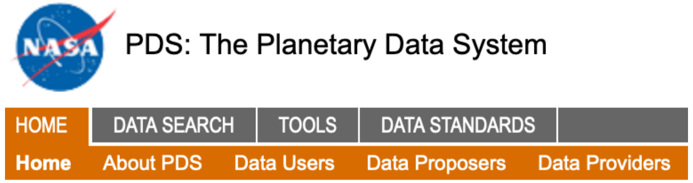


Star, S. L. & Ruhleder, K. (1996). Steps toward an ecology of infrastructure: Design and access for large information spaces. Information Systems Research, 7(1): 111-134. Figure by Florence Millerand, from: Edwards, P. N., Jackson, S. J., Bowker, G. C. & Knobel, C. P. (2007). Understanding Infrastructure: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

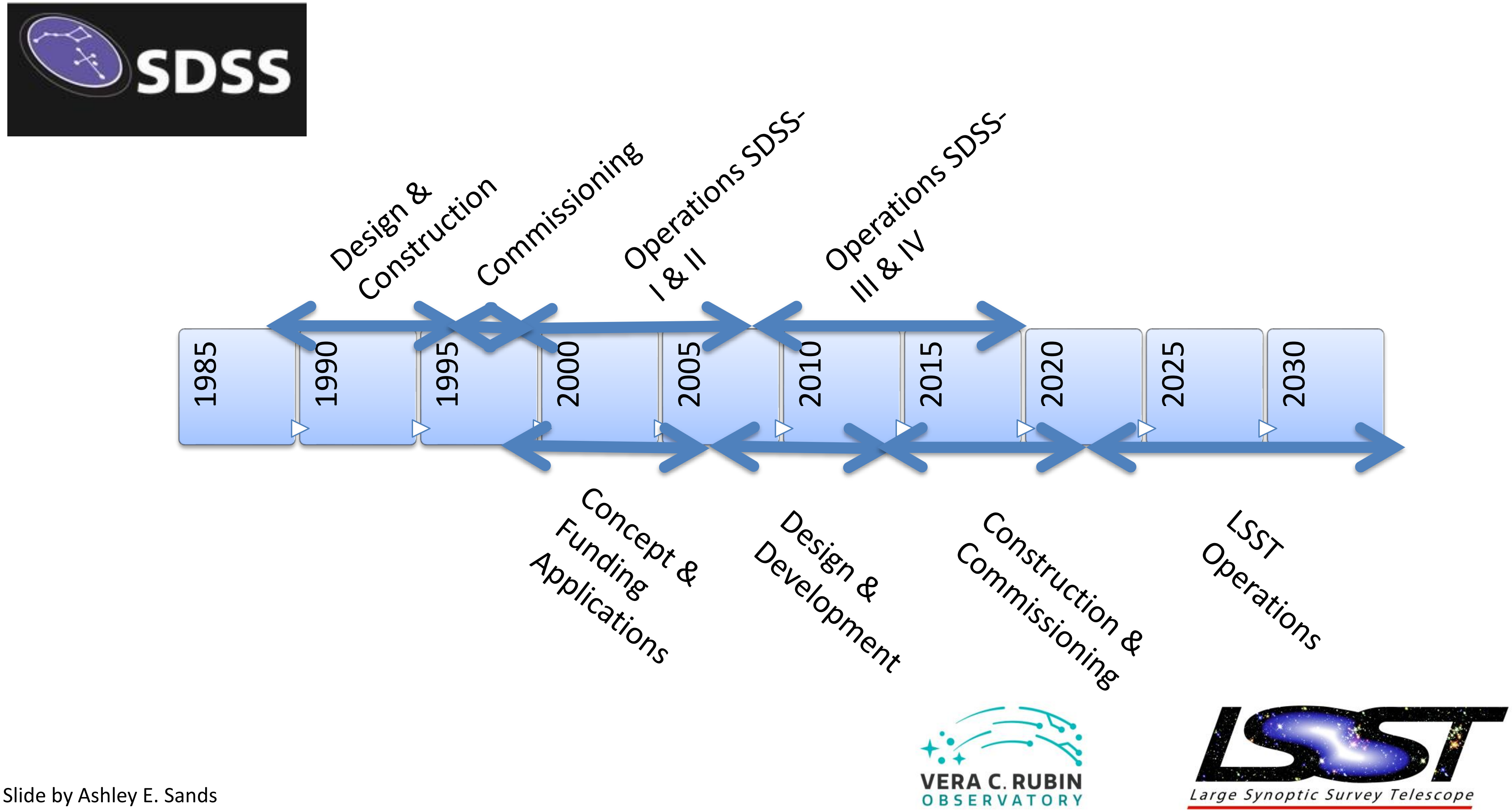
Global and Technical







Project Timelines



Slide by Ashley E. Sands

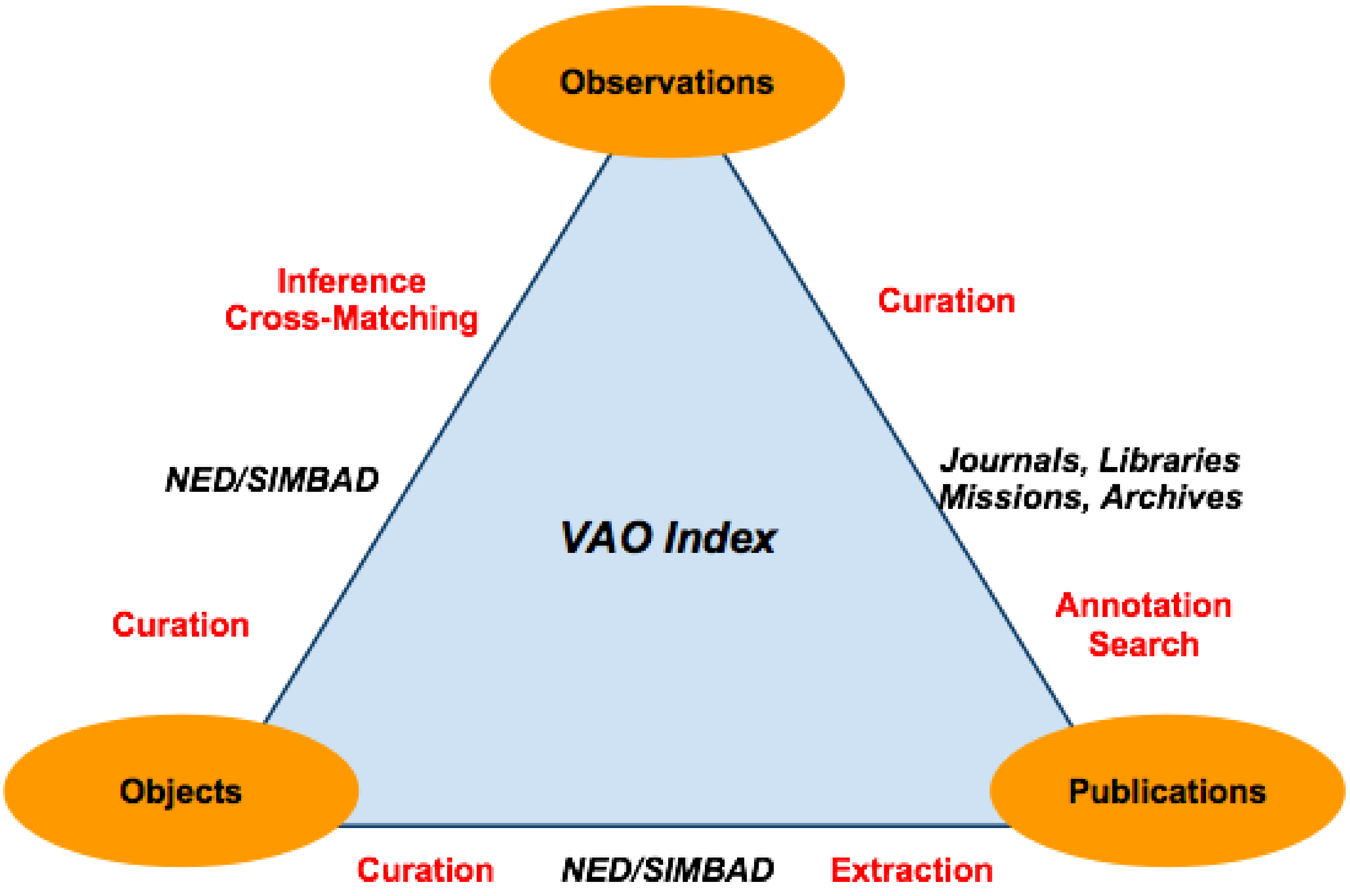
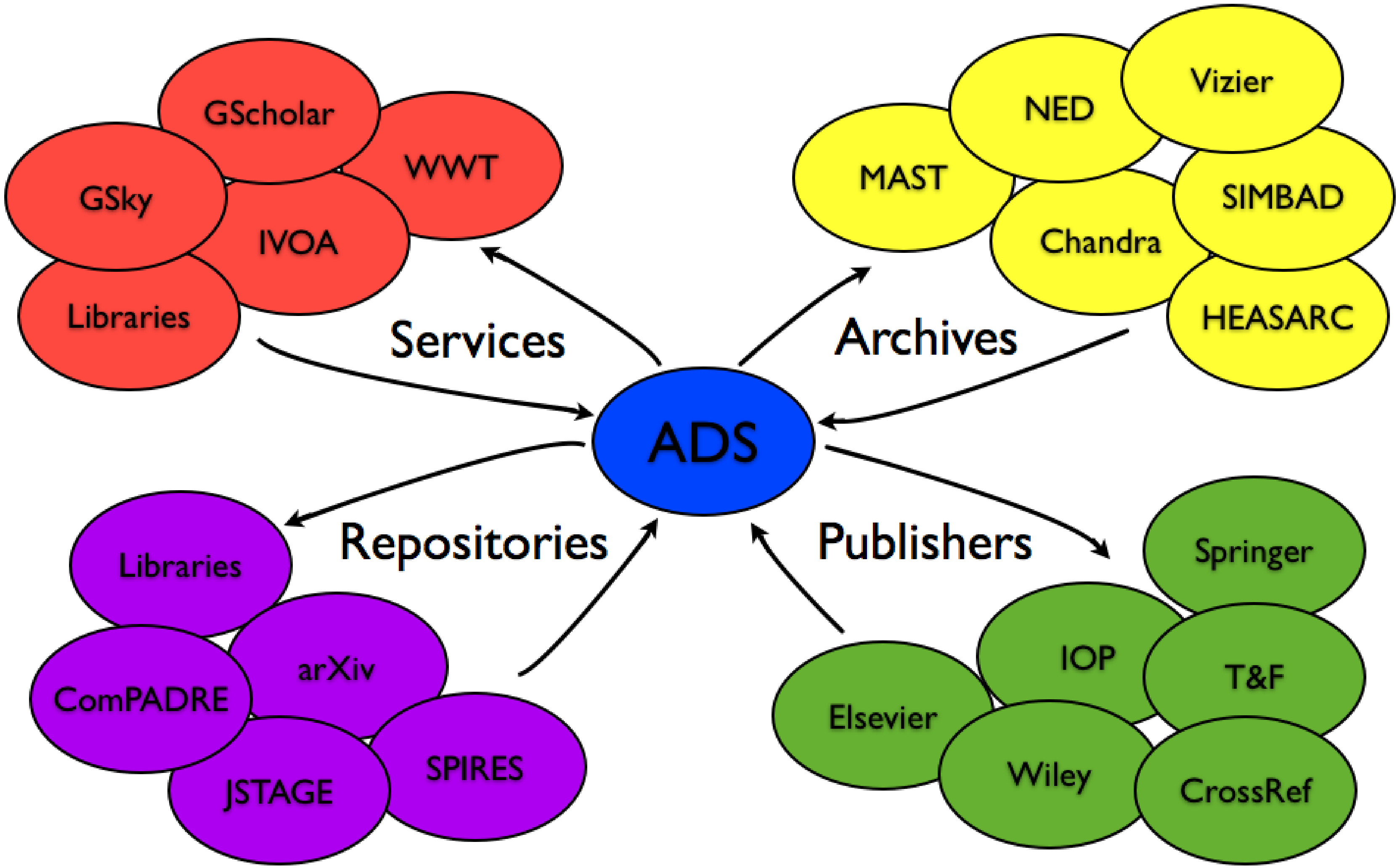


Figure 1. Relationships between Publications, Objects, Observations and the corresponding major actors in the curating process and their activities (in red).

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

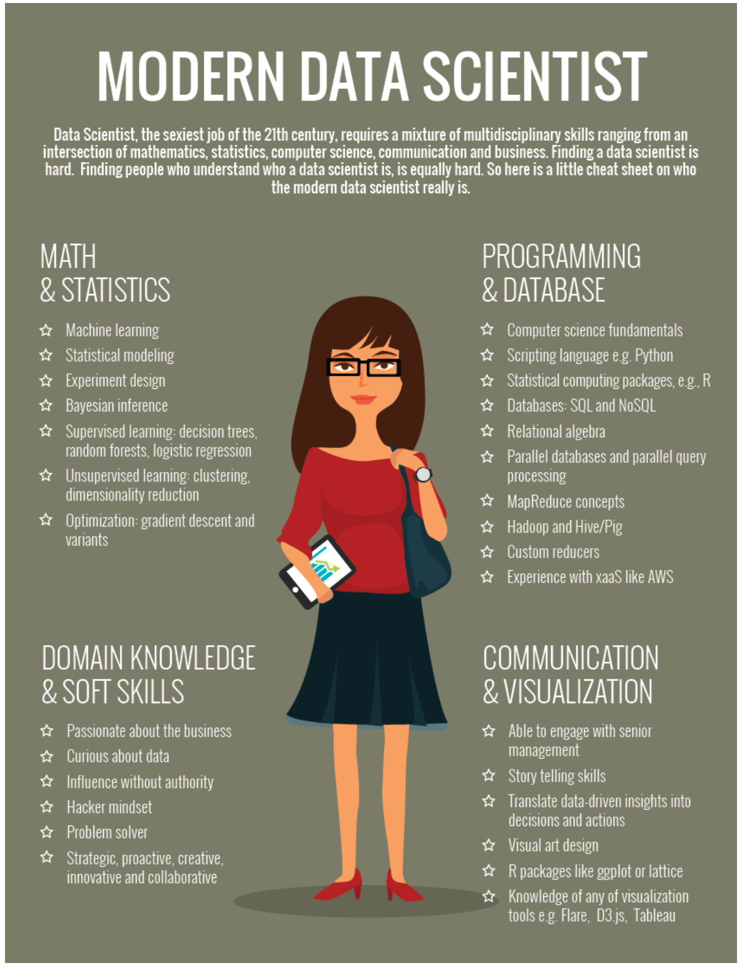
SAO/NASA Astrophysics Data System, 1993-

ADS Collaborators



Slide: Alberto Accomazzi, ADS

Local and Social



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

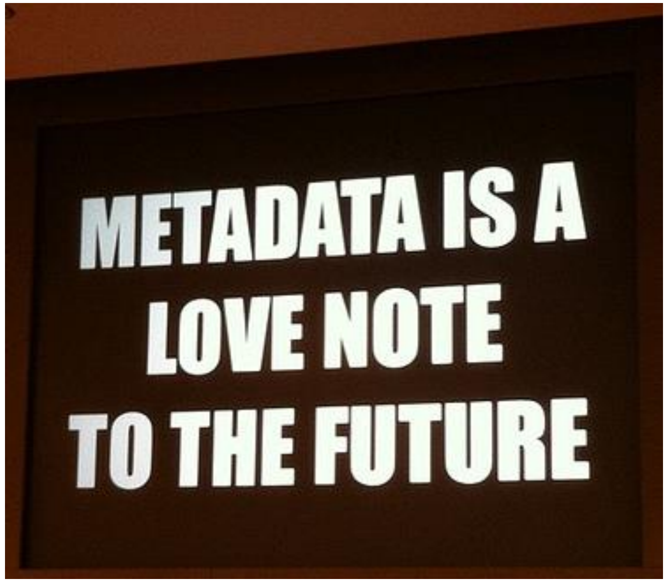


Photo by @kissane; presentation by Jason Scott (@textfiles)

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

CC Sean MacEntee, Flickr

The Data Creators’ Advantage

	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/ or https://doi.org/10.1162/99608f92.fc14bf2d

Infrastructure: Durability

- 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

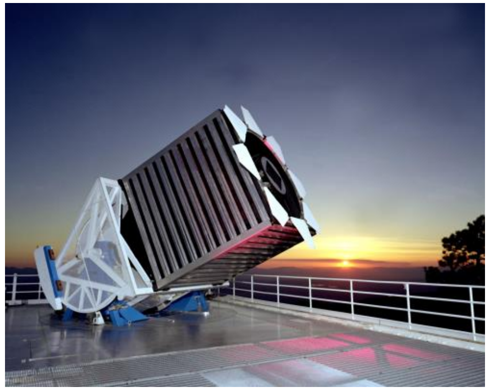
Photos by C.L. Borgman

Infrastructure: Fragility

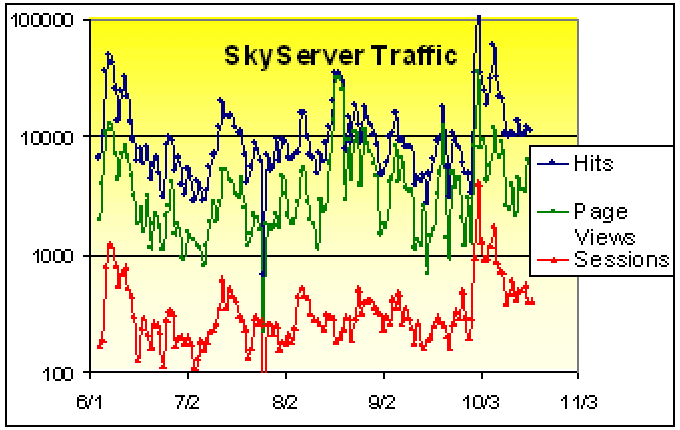
- Investments in data stewardship
  - Mission, instrument
  - Type of research
    - Space-based vs. ground based
    - Large missions vs. observing proposals
    - Shared vs. custom instruments
- Access to data
  - Public archives
  - Local websites
  - Derived data
- Curation investments
  - Open source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

Summary

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

Humboldt Forum, Interdisciplinary Studies, Berlin, May 2019. © SHF / hi.res.cam

UCLA Center for Knowledge Infrastructures

Alberto Pepe, David Fearon, Katie Shilton, Jillian Wallis, Christine Borgman, Matthew Mayernik (2009)

Christine Borgman

Peter Darch

Ashley Sands

Irene Pasquetto

Milena Golshan

Bernie Boscoe

Cheryl Thompson

Morgan Wofford

Michael Scroggins

Sharon Traweek

For a full list of CKI participants, collaborators, and coauthors since ca 2002, see https://knowledgeinfrastructures.gseis.ucla.edu/