Big Data, Little Data, or No Data?

Knowledge Infrastructures for the Earth Sciences

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

Keynote Presentation
All Hands Meeting, Seattle, June 7, 2017

EarthCube

Christine Borgman

Peter Darch

Ashley Sands

Irene Pasquetto

Bernie Randles

Milena Golshan

UCLA Center for Knowledge Infrastructures

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 wellcometrust

Australian Government
National Health and Medical Research Council

NSF National Science Foundation
WHERE DISCOVERIES BEGIN

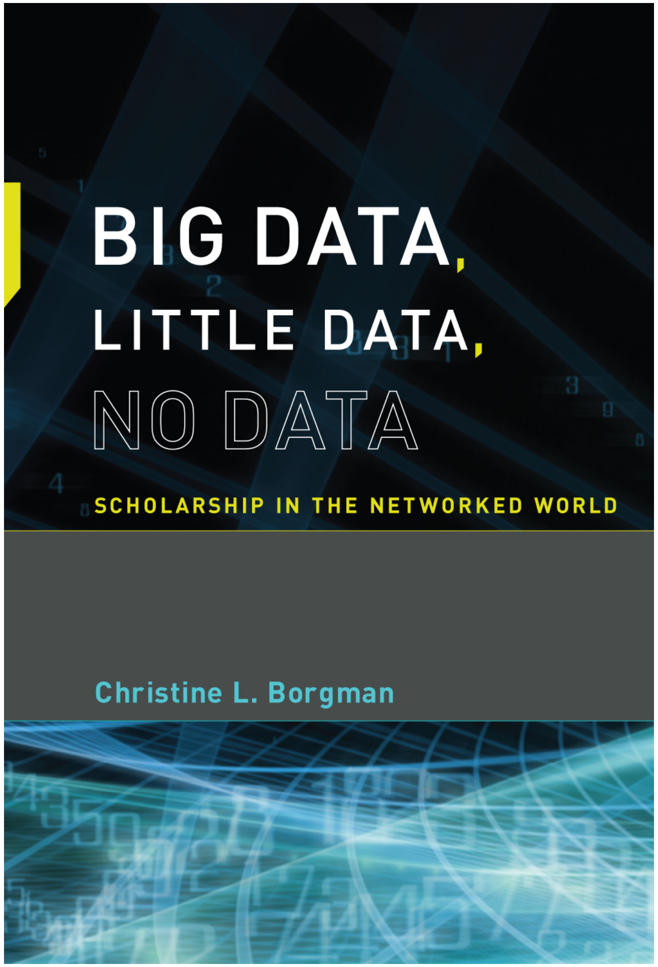
Policy RECommendations for Open Access to Research Data in Europe

RECODE

NIH National Institutes of Health
Turning Discovery Into Health

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



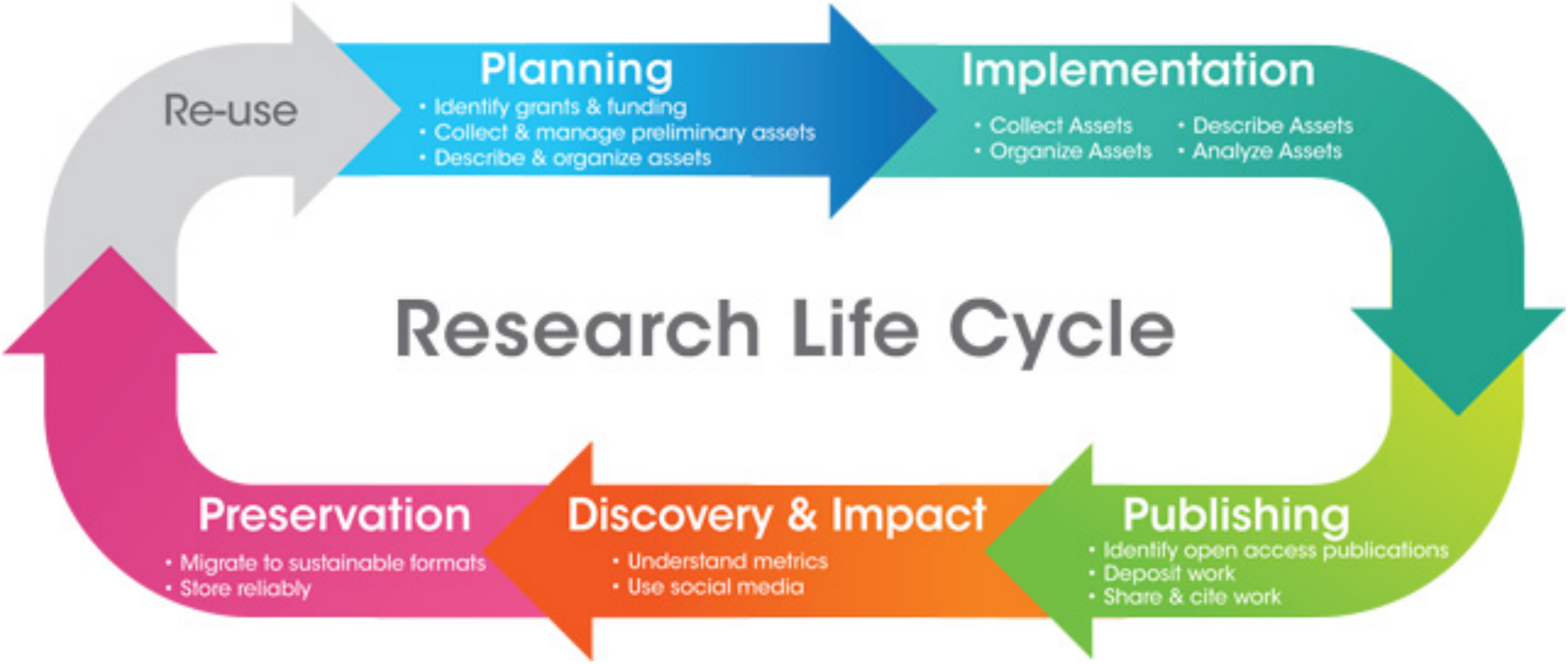
MIT Press, 2015

Lack of incentives to share data

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

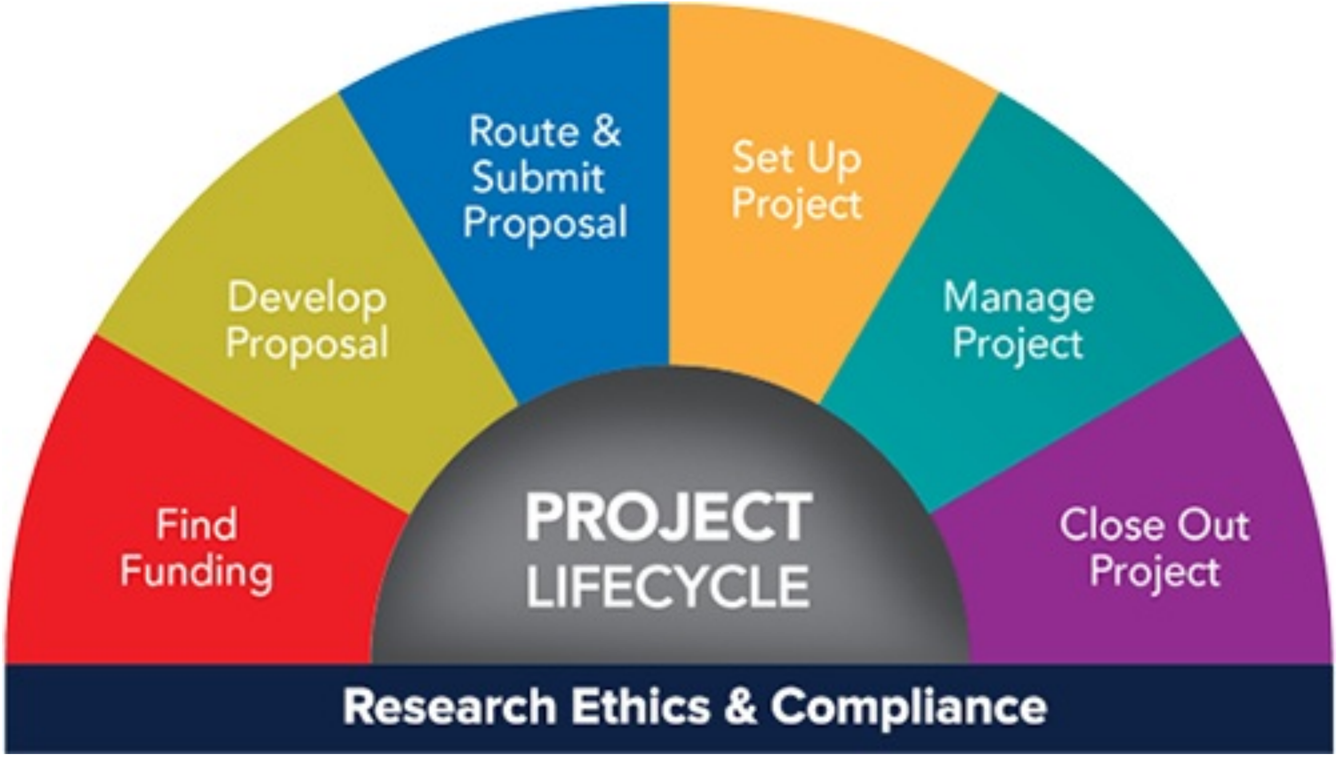
http://www.buildingsrus.co.uk/.../ target1.htm

When to invest in data?



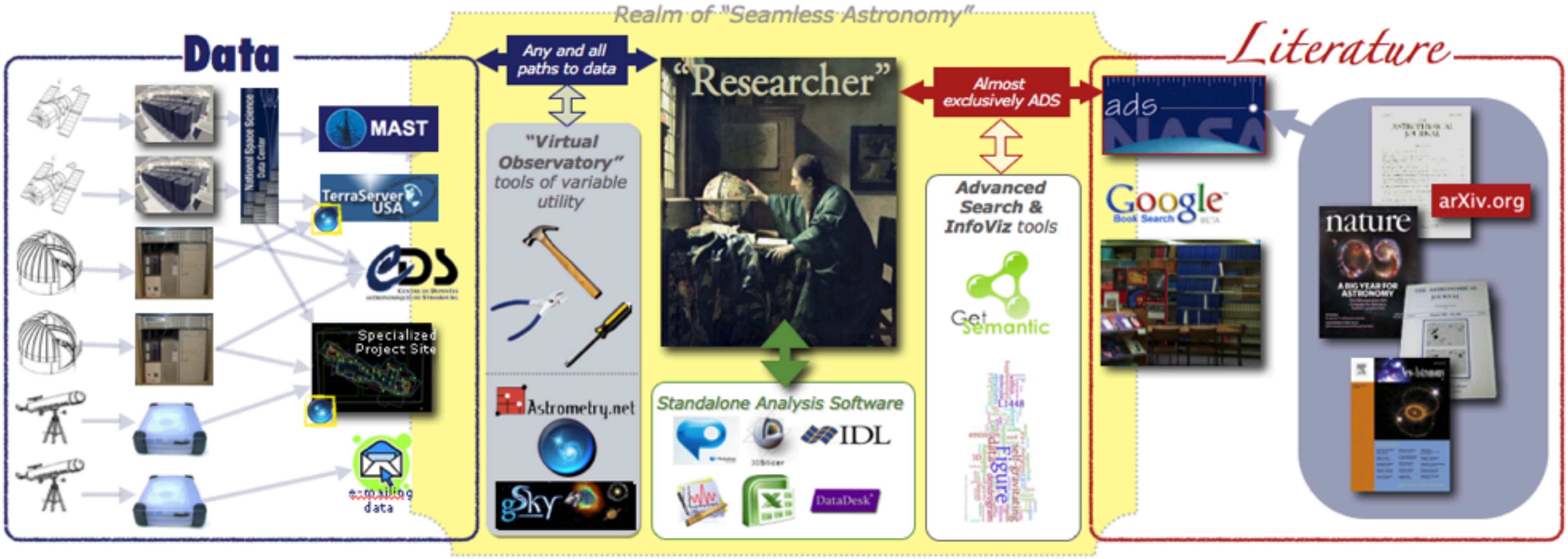
http://www.lib.uci.edu/dss/images/lifecycle.jpg

When to invest in data?



http://www.finance.umich.edu/programs

Knowledge Infrastructures



Slide: Alyssa Goodman, Harvard Astronomy

Knowledge Infrastructures:
Intellectual Frameworks and Research Challenges

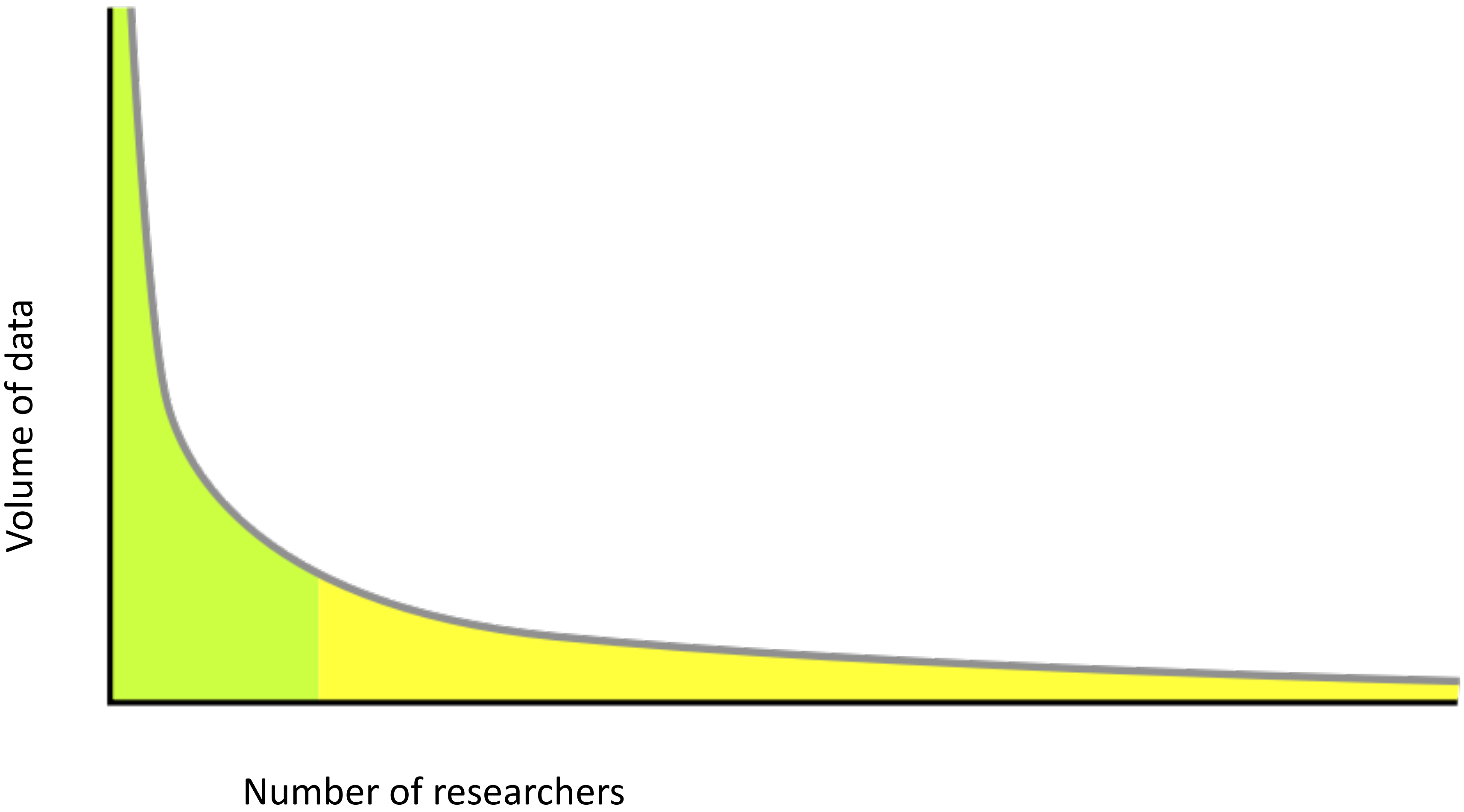
Report of a workshop sponsored by the National Science Foundation and the Sloan Foundation

University of Michigan School of Information, 25-28 May 2012

http://knowledgeinfrastructures.org

Data

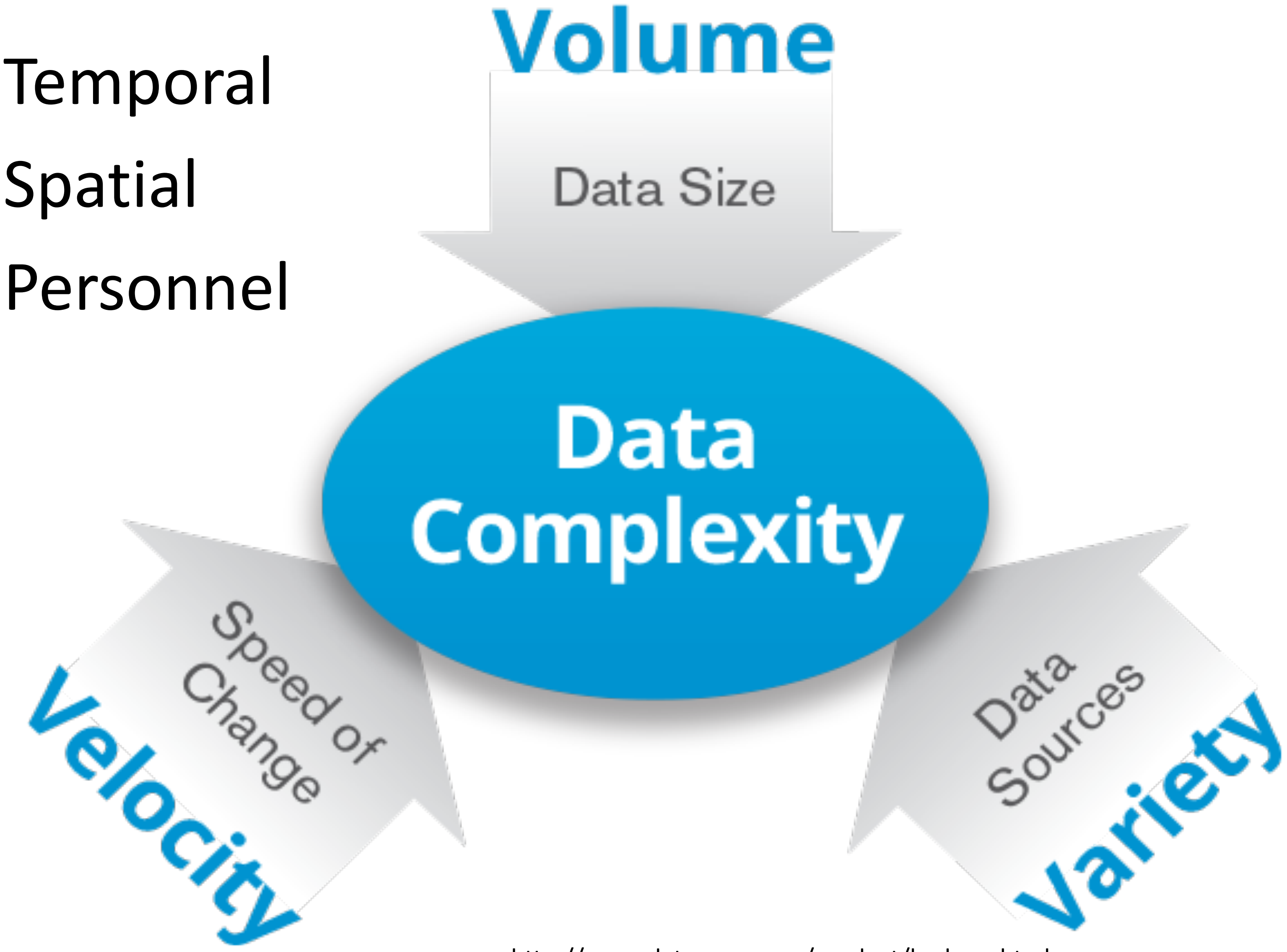
Long tail of data



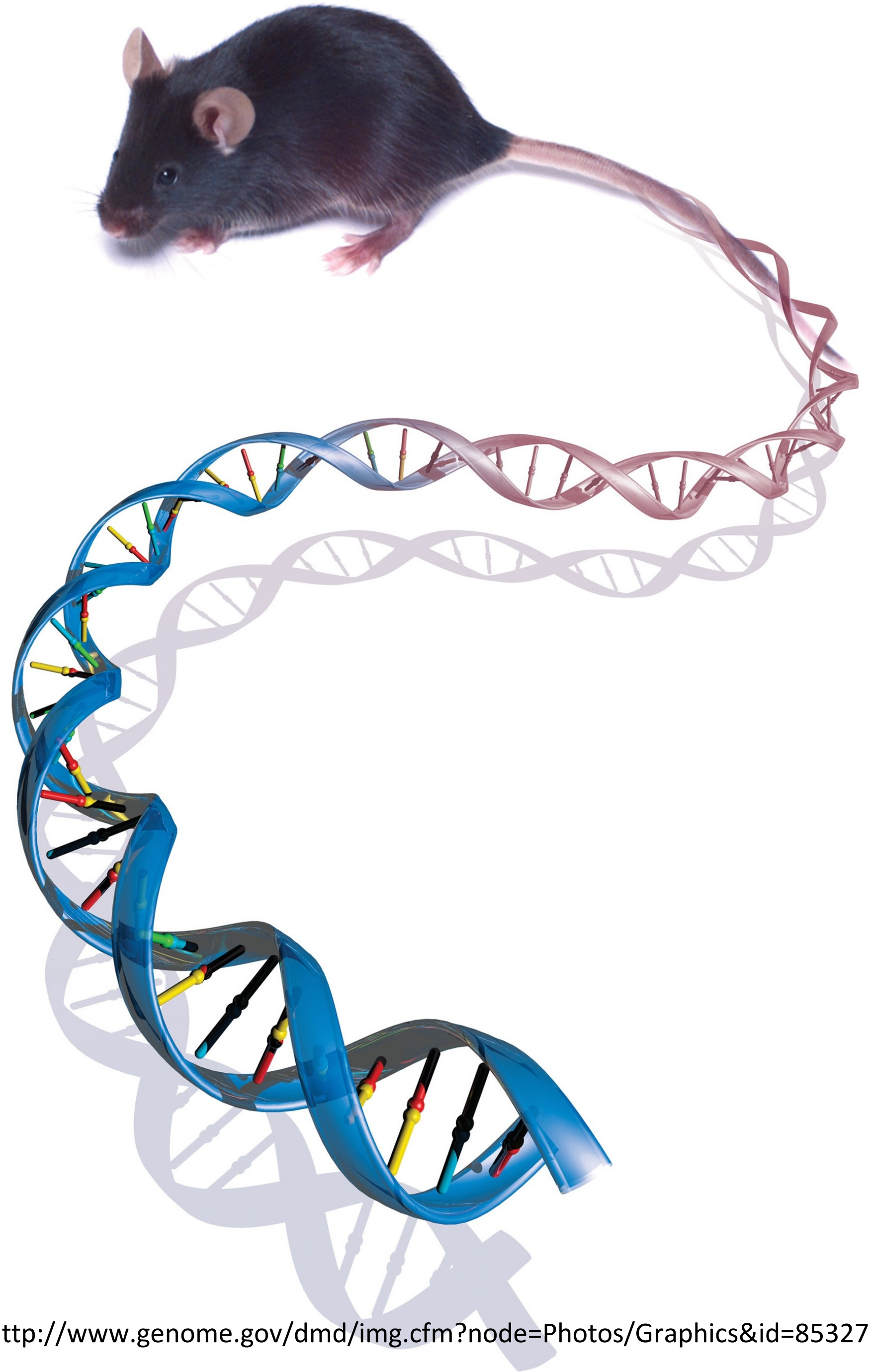
Slide: The Institute for Empowering Long Tail Research

Scale factors

- Temporal
- Spatial
- Personnel



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



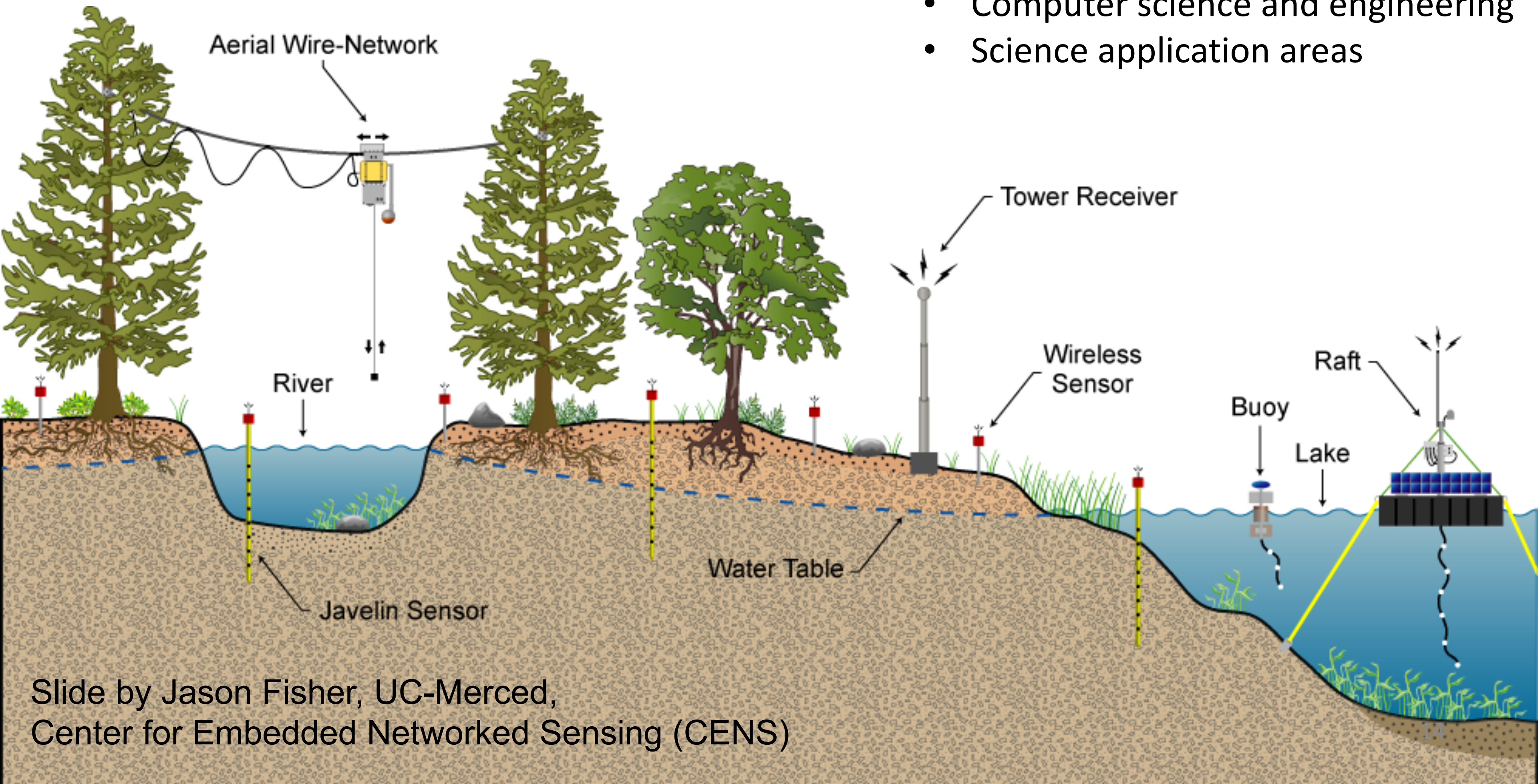
http://www.genome.gov/dmd/img.cfm?node=Photos/Graphics&id=85327

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

Center for Embedded Networked Sensing

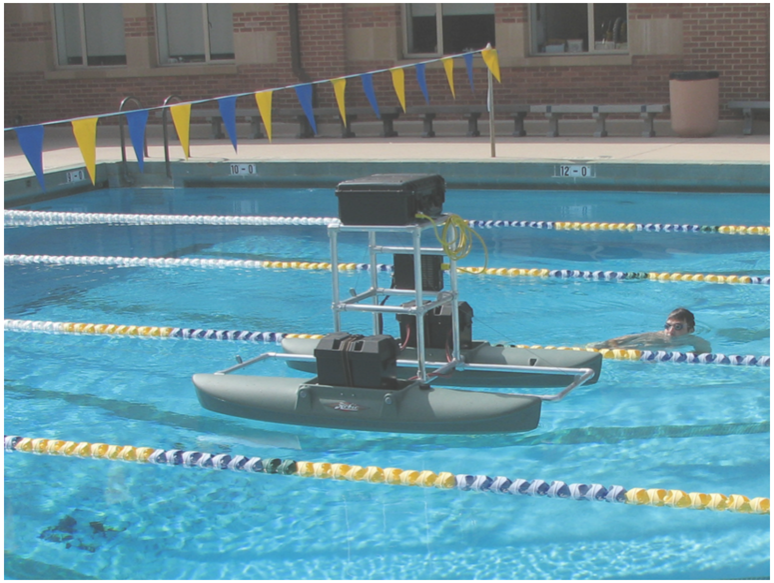
- NSF Science & Tech Ctr, 2002-2012
- 5 universities, plus partners
- 300 members
- Computer science and engineering
- Science application areas



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

Science <–> Data

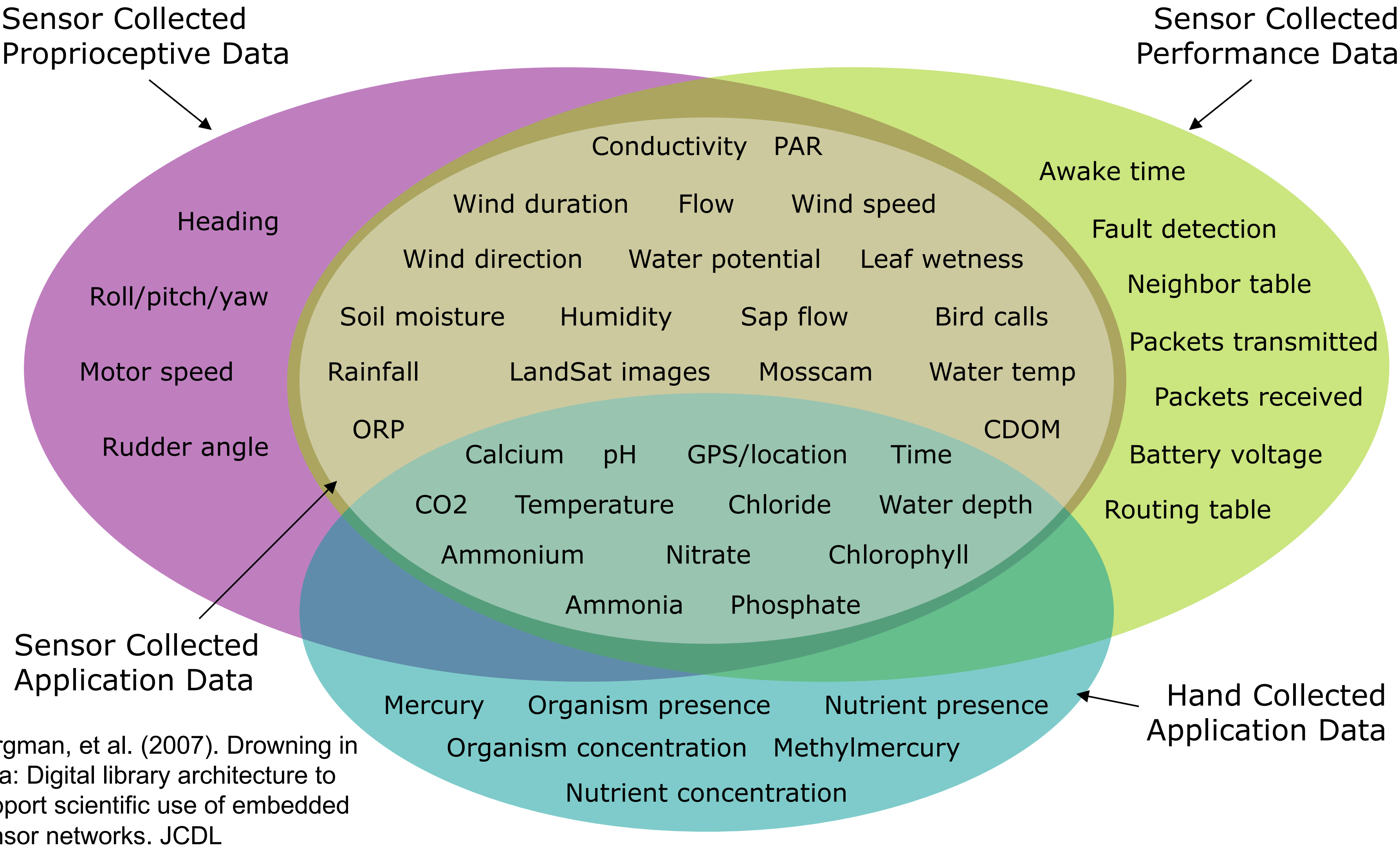
Engineering researcher: ***"Temperature is temperature."***



CENS Robotics team

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



Sensor Collected Proprioceptive Data

Heading

Roll/pitch/yaw

Motor speed

Rudder angle

Sensor Collected Performance Data

Awake time

Fault detection

Neighbor table

Packets transmitted

Packets received

Battery voltage

Routing table

Conductivity PAR

Wind duration Flow Wind speed

Wind direction Water potential Leaf wetness

Soil moisture Humidity Sap flow Bird calls

Rainfall LandSat images Mosscam Water temp

ORP CDOM

Calcium pH GPS/location Time

CO2 Temperature Chloride Water depth

Ammonium Nitrate Chlorophyll

Ammonia Phosphate

Sensor Collected Application Data

Mercury Organism presence Nutrient presence

Organism concentration Methylmercury

Nutrient concentration

Hand Collected Application Data

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

Deep Subseafloor Biosphere

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

http://iodp.org/expeditions

Slide by Peter T. Darch, UIUC

Center for Dark Energy Biosphere Investigations

International Ocean Discovery Program
Iodp.tamu.org

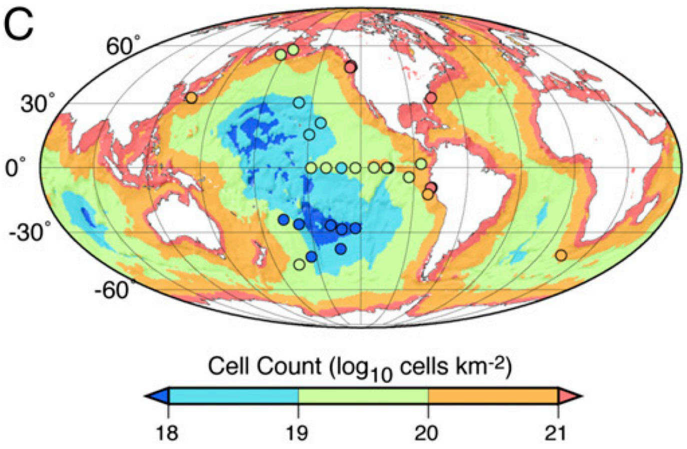
- NSF Science & Tech Ctr, 2010-2020
- 20 universities, plus partners (35 institutions)
- 90 scientists
- Biological sciences
- Physical sciences

Slide by Peter T. Darch, UIUC

Repository for seafloor cores. Photo: Peter Darch

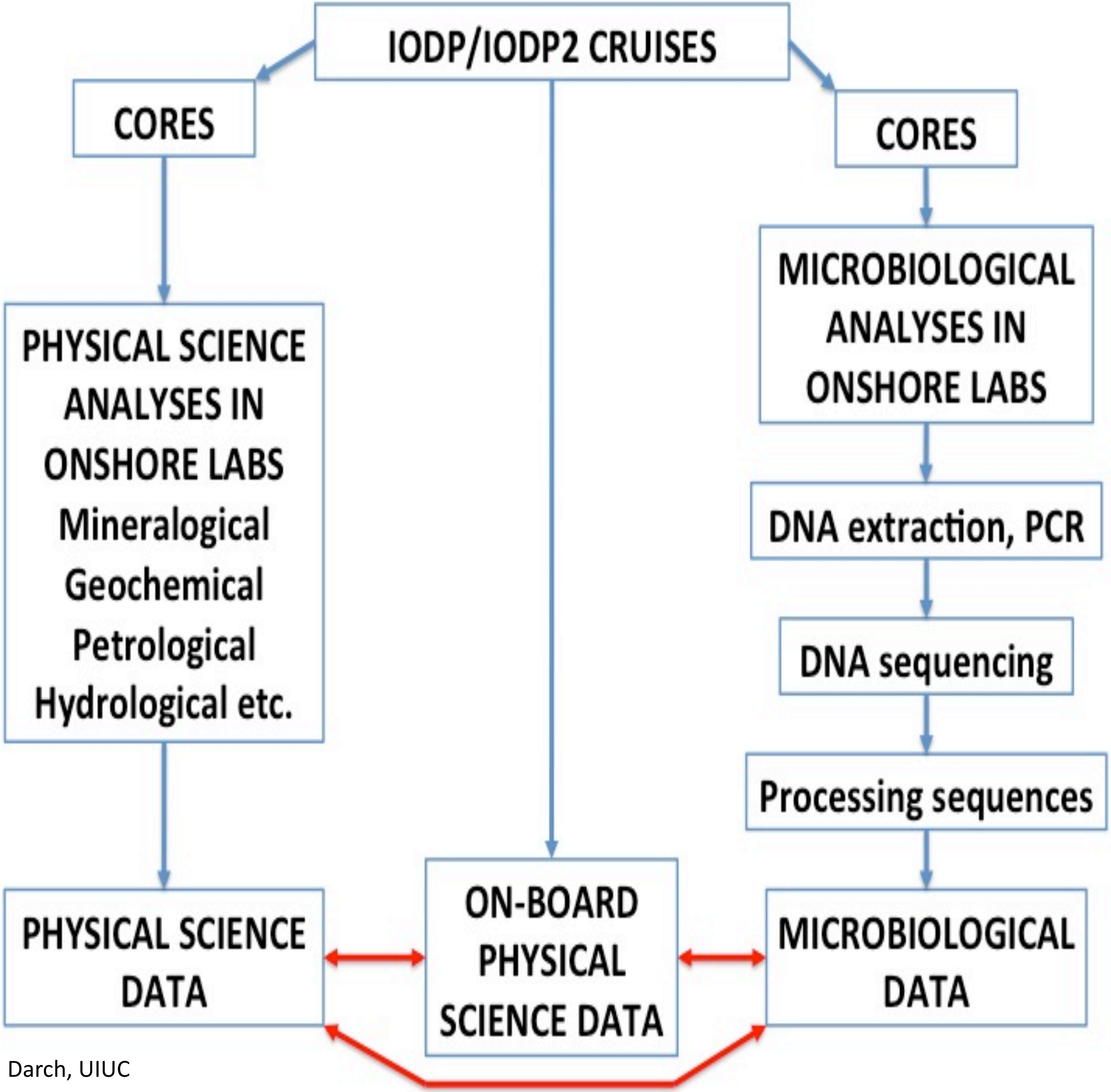
Benefits of Data Reuse

- Increase access to data
- Address complex questions
- Build shared reference collections



Kallmeyer et al. (2012). Global distribution of microbial abundance and biomass in subseafloor sediment. Proceedings of the National Academy of Sciences, 109(40), 16213–16216.

Slide by Peter T. Darch, UIUC



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

Slide by Peter T. Darch, UIUC

Availability of Earth Science Data

- Abundant data vs. Scarce data
- Scientific objectives
  - – Discovery-driven
  - – Hypothesis-driven
- Scientific constraints
  - – Emergent domain
  - – Shared IODP resources

Photograph by Peter T. Darch

Reuse vs. Reproducibility

- Data reuse can be productive in data-scarce domains
- Reproducibility requires standards
  - Maturity varies by domain
  - Standards may be non-existent, inappropriate, or premature
- Reproducibility goals may
  - Inhibit scientific progress
  - Obscure data reuse opportunities

http://iodp.org/expeditions

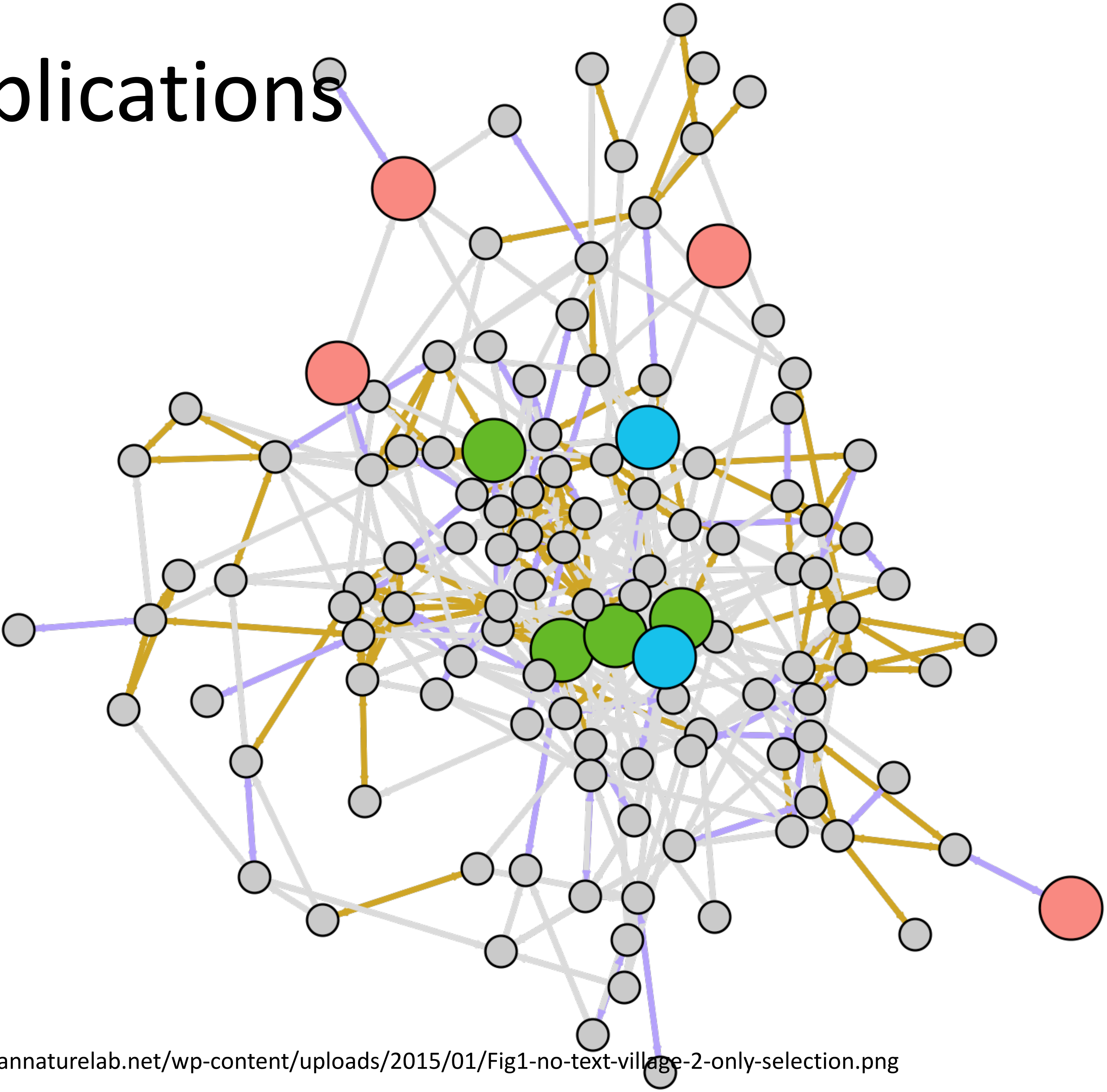
Darch, P. T., & Borgman, C. L. (2016). Ship space to database: emerging infrastructures for studies of the deep subseafloor biosphere. *PeerJ Computer Science*, *2*, e97. https://doi.org/10.7717/peerj-cs.97

Publications

Estrogen Receptors in Human Nontarget Tissues: Biological and Clinical Implications
Effects of Low Dose Oral Contraceptives on Very Low Density Lipoprotein Metabolism
The Estradiol-stimulated Lipoprotein Receptor of Rat Liver
Dégradation ER mRNA...
1.3.2.1
1225

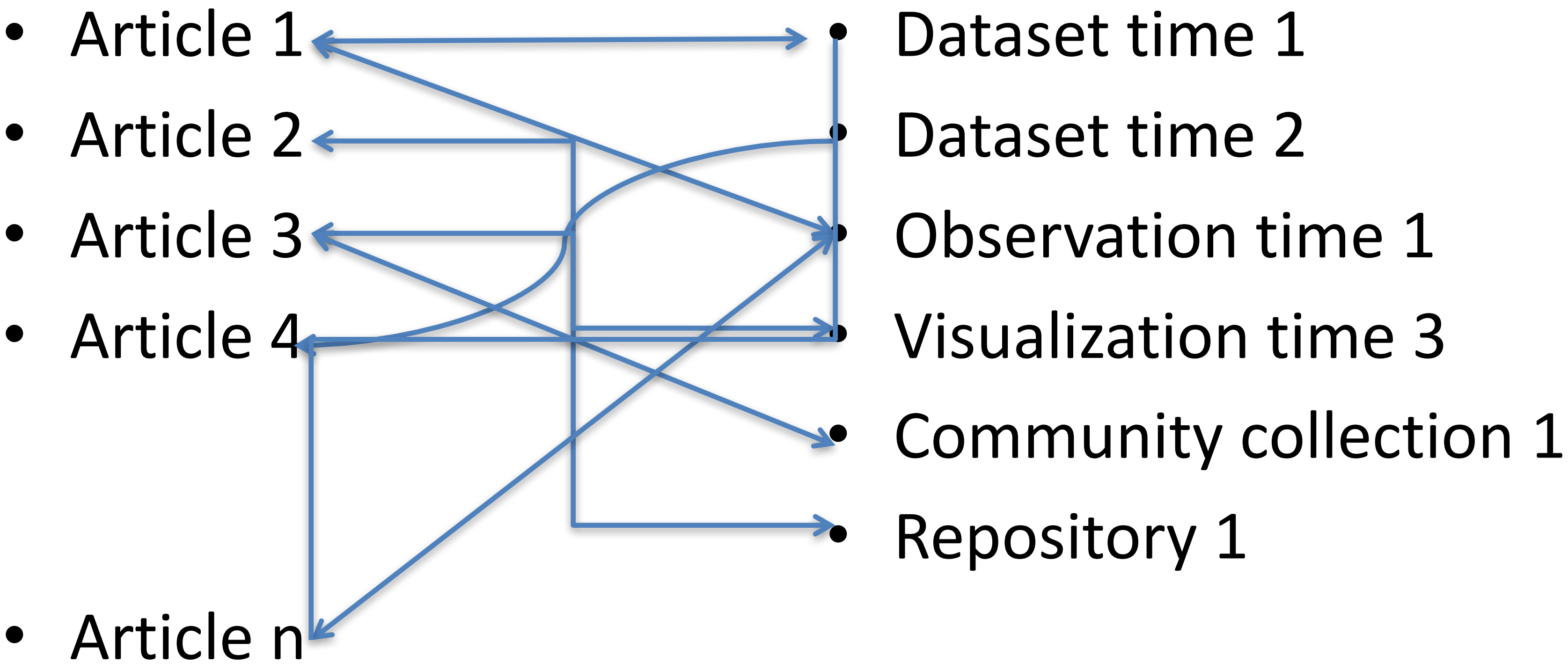
http://www.cse.psu.edu/hpcl/images/publications.jpg

Publications



http://humannaturelab.net/wp-content/uploads/2015/01/Fig1-no-text-village-2-only-selection.png

Publications <–> Data: Mapping



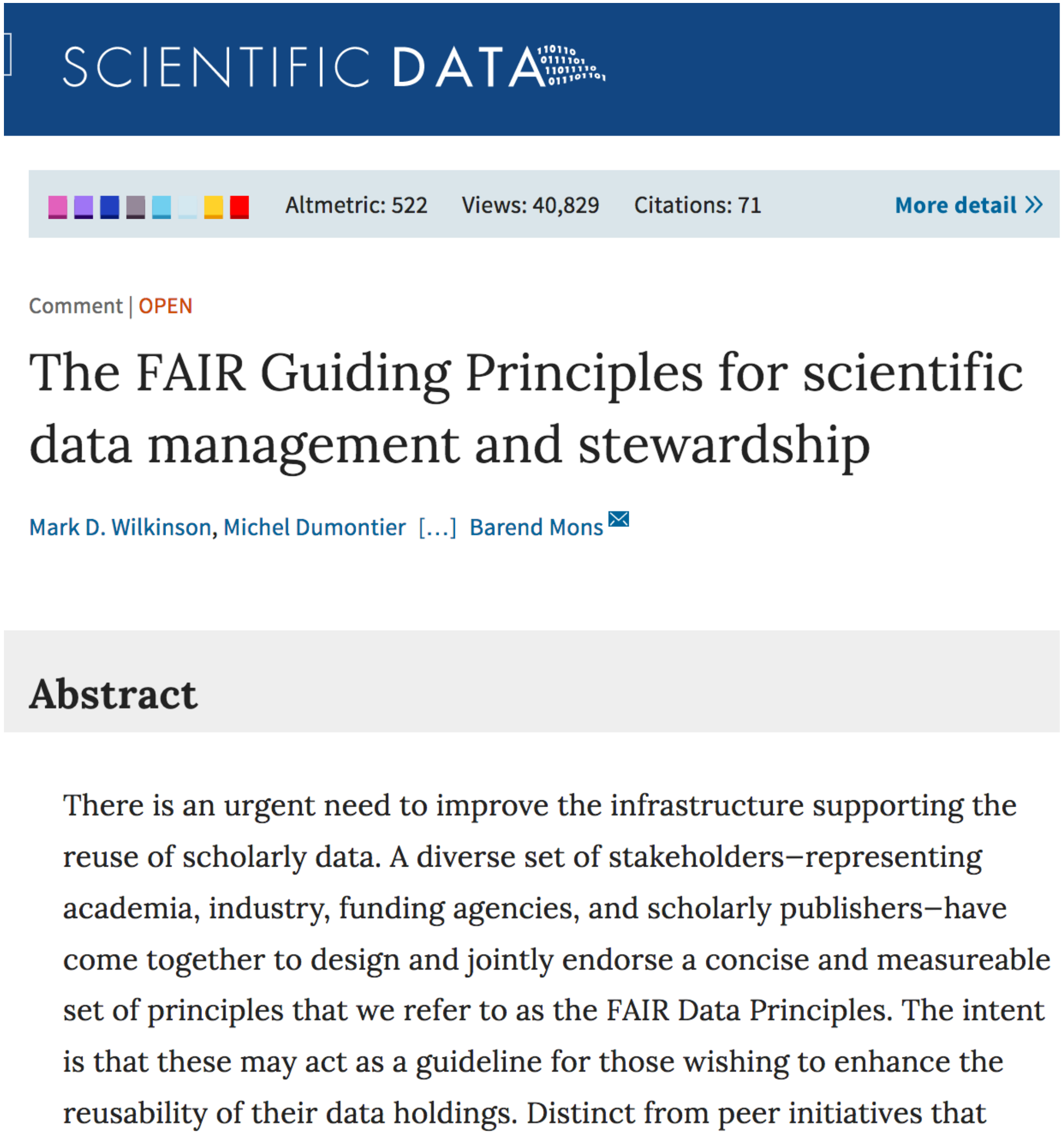
- Article 1
- Article 2
- Article 3
- Article 4
- Article n

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

Publications <–> Data: Attribution

- Publications
  - Independent units
  - Authorship is negotiated
- Data
  - Compound objects
  - Ownership is rarely clear
  - Attribution
    - Long term responsibility: Investigators
    - Expertise for interpretation: Data collectors and analysts

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



SCIENTIFIC DATA

Altmetric: 522 Views: 40,829 Citations: 71 More detail »

Comment | OPEN

The FAIR Guiding Principles for scientific data management and stewardship

Mark D. Wilkinson, Michel Dumontier […] Barend Mons

Abstract

There is an urgent need to improve the infrastructure supporting the reuse of scholarly data. A diverse set of stakeholders—representing academia, industry, funding agencies, and scholarly publishers—have come together to design and jointly endorse a concise and measureable set of principles that we refer to as the FAIR Data Principles. The intent is that these may act as a guideline for those wishing to enhance the reusability of their data holdings. Distinct from peer initiatives that

- Findable
- Accessible
- Interoperable
- Reusable

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data, 3, 160018. Retrieved from http://dx.doi.org/10.1038/sdata.2016.18

Metadata

- Metadata is structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage an information resource.*
  - – descriptive
  - – structural
  - – administrative

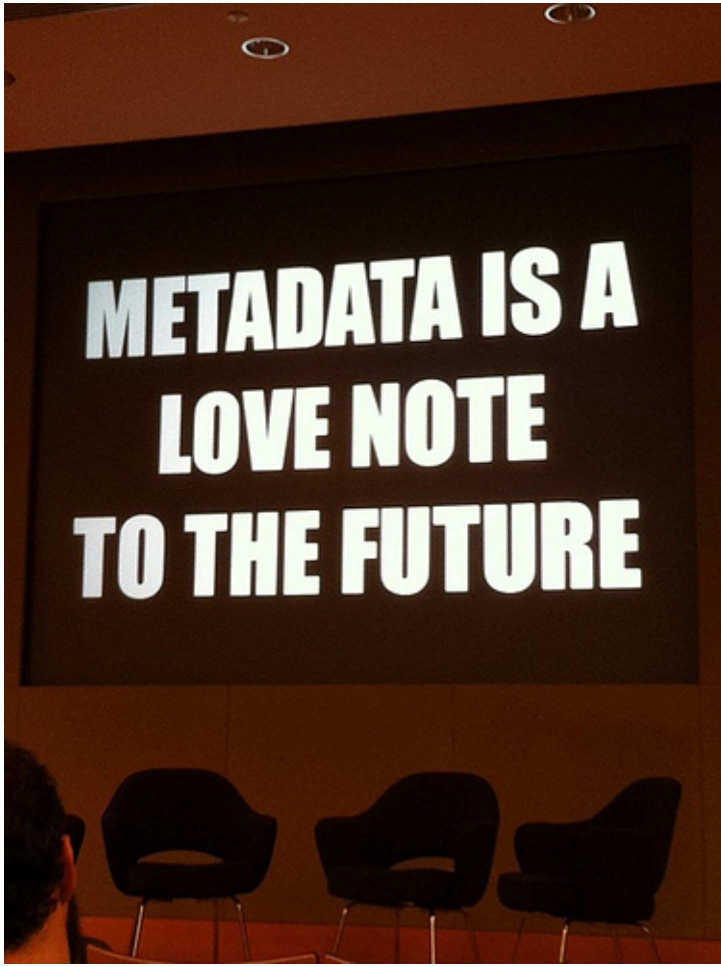
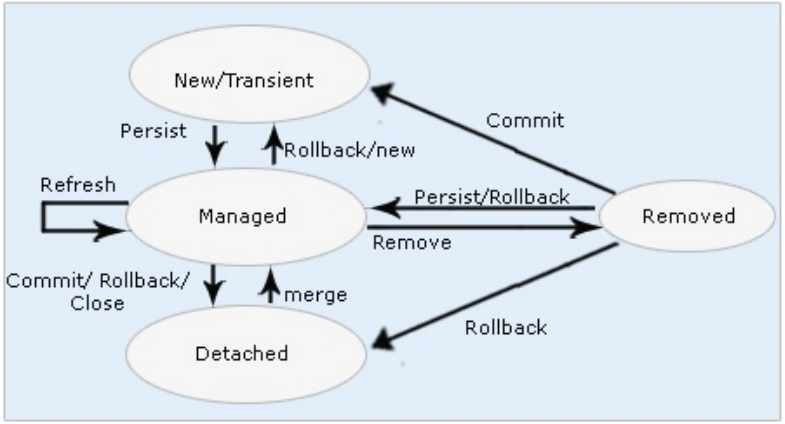


Photo by @kissane; presentation by Jason Scott (@textfiles)

*National Information Standards Organization 2004

Identity and persistence

- Identity
  - Identifiers
    - DOI, Handles
    - URI, PURL…
  - Naming and namespaces
    - Authors/creators: ORCID, ISNI, VIAF…
    - Generic/specific: registry number…
  - Description
    - Self-describing
    - Metadata augmentation
- Persistence
  - Perishable
  - Long-lived
  - Permanent



Persistence Content

http://web-interview-questions.blogspot.com/2010_06_21_archive.html

Provenance

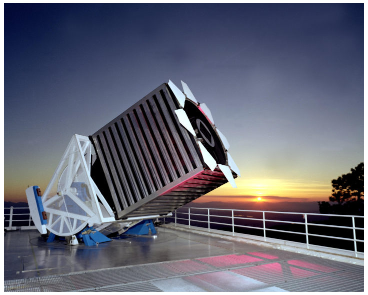
- Libraries: Origin or source
- Museums: Chain of custody
- Internet: Provenance is information about entities, activities, and people involved in producing a piece of data or thing, which can be used to form assessments about its quality, reliability or trustworthiness.*

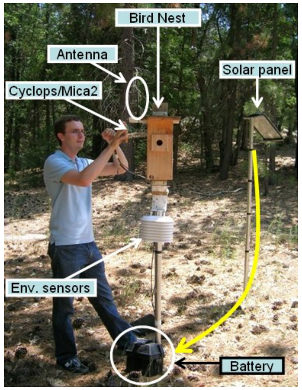
*World Wide Web Consortium (W3C) Provenance working group

British Library, provenance record:
Bestiary - caption: 'Owl mobbed by smaller birds'

Data sharing and access

- Centralized data production
  - Top down investments in data
  - Common data archive
- Decentralized data production
  - Bottom up investments in data
  - Pool domain resources later
- Domain-independent aggregators
  - University repositories
  - Dataverse, Figshare, Slideshare, …
- Post on lab / personal websites
- Share privately upon request





Wallis, J. C., Rolando, E., & Borgman, C. L. (2013). If we share data, will anyone use them? Data sharing and reuse in the long tail of science and technology. *PLoS ONE*, *8*(7), e67332. https://doi.org/10.1371/journal.pone.0067332

Reuse across place and time

- Reuse by investigator
- Reuse by collaborators
- Reuse by colleagues
- Reuse by unaffiliated others
- Reuse at later times
  - Months
  - Years
  - Decades
  - Centuries

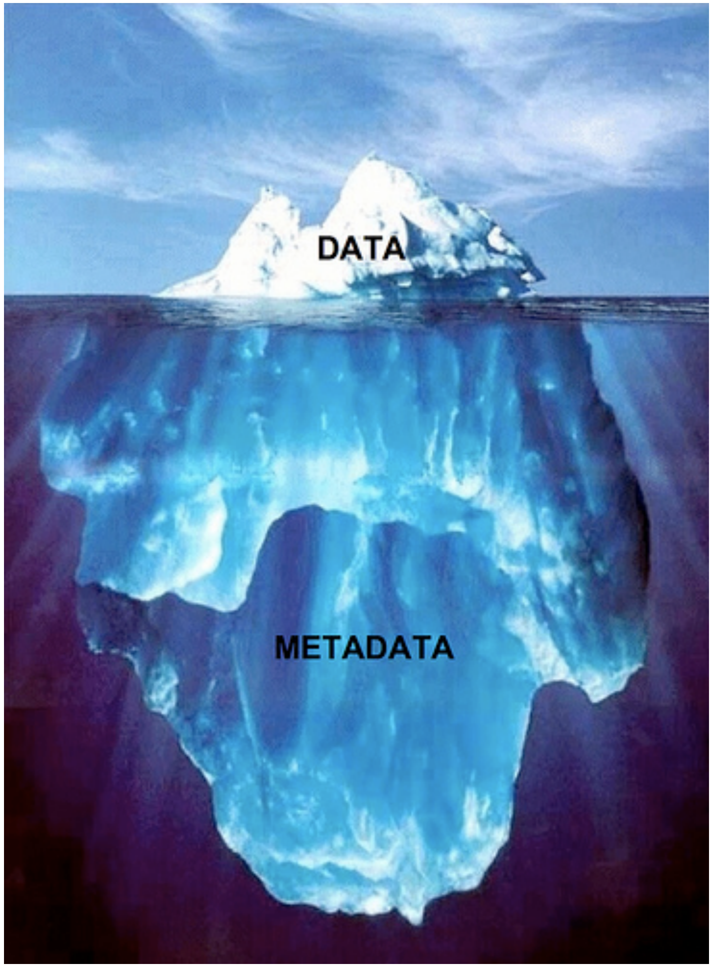


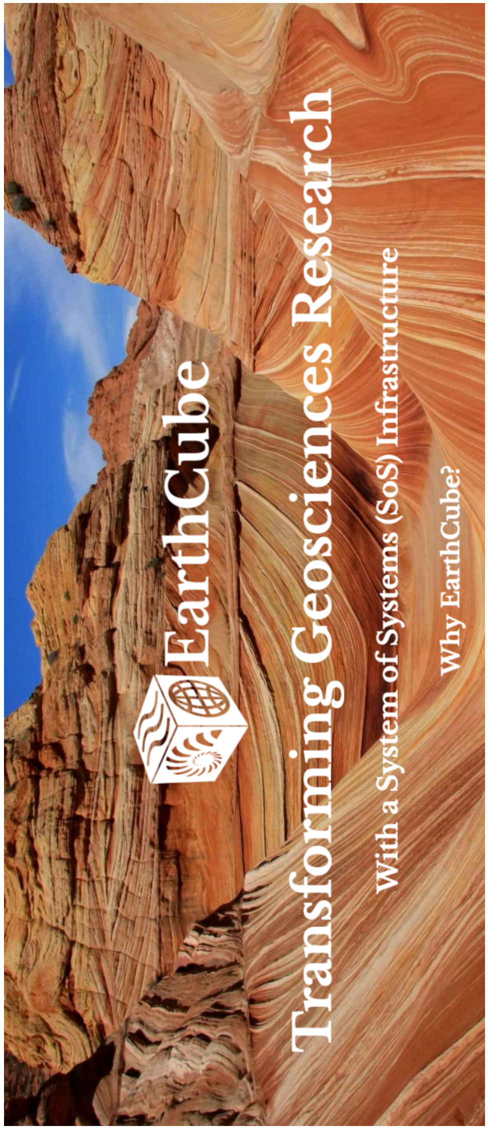
Image from Soumitri Varadarajan blog. Iceberg image © Ralph A. Clevenger. Flickr photo

Economics of the Knowledge Commons

	Subtractability / Rivalry		
		Low	High
Exclusion	Difficult	**Public Goods** General knowledge Public domain data	**Common-pool resources** Libraries Data archives
	Easy	**Toll or Club Goods** Subscription journals Subscription data	**Private Goods** Printed books Raw or competitive data

Adapted from C. Hess & E. Ostrom (Eds.), *Understanding knowledge as a commons: From theory to practice*. MIT Press.

Suggestions for EarthCube



EarthCube

Transforming Geosciences Research

With a System of Systems (SoS) Infrastructure

Why EarthCube?

- Follow the FAIR principles
- Invest in data early and often
- Sustain access to observational data
- Invest in domain repositories
- Invest in data documentation
  - Data, metadata, provenance
  - Research questions
  - Protocols, instrumentation
  - Software

There is no plan B, because there is no PLANET B!

- UN Secretary-General Ban Ki-moon

Acknowledgements

Christine Borgman

Peter Darch

Ashley Sands

Irene Pasquetto

Bernie Randles

Milena Golshan

UCLA Center for Knowledge Infrastructures

NSF

ALFRED P. SLOAN FOUNDATION 1934