Why are data sharing and reuse so difficult?

Christine L. Borgman

Professor and Presidential Chair in Information Studies

University of California, Los Angeles

http://christineborgman.info

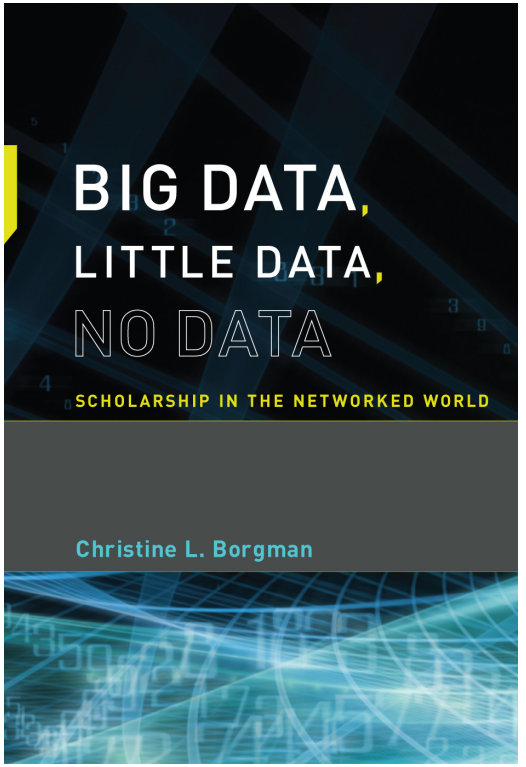
and

UCLA Knowledge Infrastructures Team: Peter Darch, Milena Golshan, Irene Pasquetto, Ashley Sands, Sharon Traweek

FaceBase All Hands Meeting

Information Sciences Institute, Marina del Rey, CA

Thursday, January 8, 2015



knowledge infrastructures
UCLA

@SciTechProf

http://knowledgeinfrastructures.gseis.ucla.edu/

The data deluge has arrived. Data-driven science is accelerating rapidly, but without the necessary social, technical, or policy infrastructure to support the capture, management, curation, use, and reuse of those data. Universities, libraries, funding agencies, and investigators are making critical decisions about what data to keep, in what form, for how long, and at what price. Academic programs are struggling to teach new skills in data management and policy, within the disciplines and within the information professions. All of these efforts are hampered by the lack of robust research that compares sites, disciplines, practices, and policies over a long period of time. The UCLA Knowledge Infrastructures Team studying data, data practices, and data curation brings to this problem several decades of research experience in the social studies of science, digital libraries, and information systems design and development. Related projects by each of the investigators are linked individually.

Knowledge Infrastructures Project
Research Design

	Big Data	Small Data
Ramping up data collection	Large Synoptic Survey Telescope (LSST)	Center for Dark Energy Biosphere Investigations (C-DEBI)
Ramping down data collection	Sloan Digital Sky Survey, Parts I & II (SDSS)	Center for Embedded Network Sensing (CENS)
Knowledge Infrastructures		

Knowledge Infrastructures

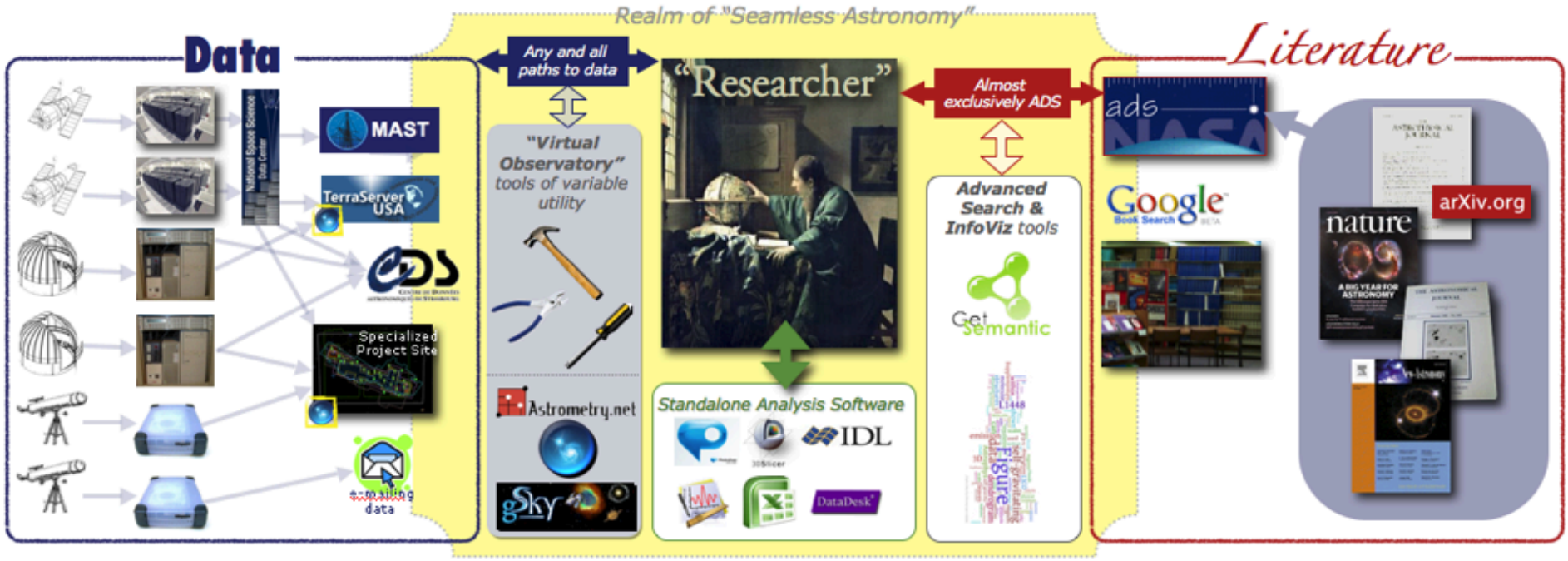


Image: Alyssa Goodman, Seamless Astronomy, Harvard-CfA

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

Precondition:

Researchers share data

Researchers' perspectives on data sharing

- Rewards
- Responsibility
- Data
- Incentives

Persistent URL: photography.si.edu/SearchImage.aspx?id=5799
Repository: Smithsonian Institution Archives

Researchers' perspectives on data sharing

- Rewards
- Responsibility
- Data
- Incentives

Persistent URL: photography.si.edu/SearchImage.aspx?id=5799
Repository: Smithsonian Institution Archives

Rewards may vary…

- Publications
- Grants
- Awards and honors
- Teaching
- Service
- Technologies
- Data
- …

http://blog.startfreshtoday.com/Portals/170402/images/improve-credit-score1.jpg

Researchers' perspectives on data sharing

- Rewards
- Responsibility
- Data
- Incentives

Persistent URL: photography.si.edu/SearchImage.aspx?id=5799
Repository: Smithsonian Institution Archives

Responsibility

Publications are arguments made by authors, and data are the evidence used to support the arguments.

C.L. Borgman (2015). *Big Data, Little Data, No Data: Scholarship in the Networked World*. MIT Press

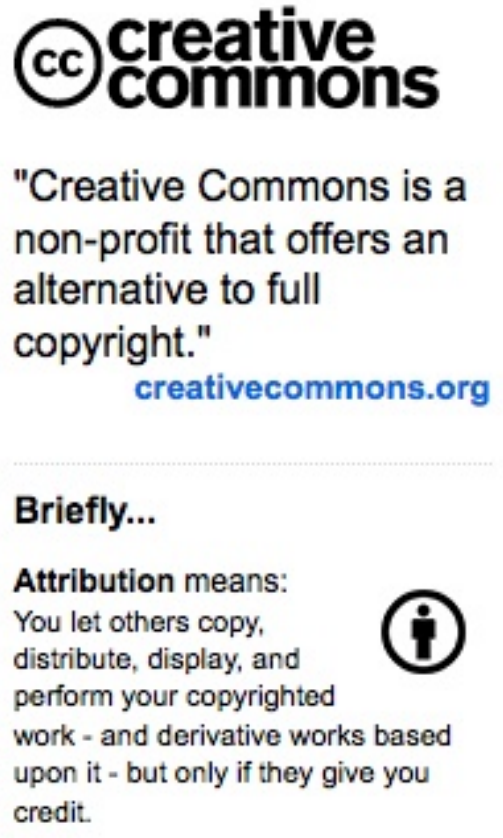
Responsibility

- 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

hudsonalpha.org

Attribution of data

- Legal responsibility
  - Licensed data
  - Specific attribution required
- Scholarly credit: contributorship
  - "Author" of data
  - Contributor of data to this publication
  - Colleague who shared data
  - Software developer
  - Data collector
  - Instrument builder
  - Data curator
  - Data manager
  - Data scientist
  - Field site staff
  - Data calibration
  - Data analysis, visualization
  - Funding source
  - Data repository
  - Lab director
  - Principal investigator
  - University research office
  - Research subjects
  - Research workers, e.g., citizen science…



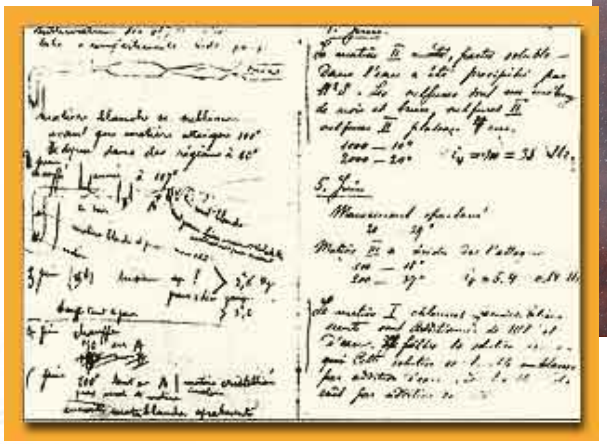
Researchers' perspectives on data sharing

- Rewards
- Responsibility
- Data
- Incentives

Persistent URL: photography.si.edu/SearchImage.aspx?id=5799
Repository: Smithsonian Institution Archives

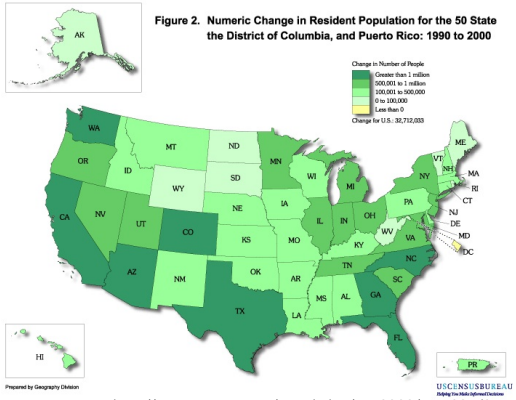
What are data?

NASA Astronomy Picture of the Day

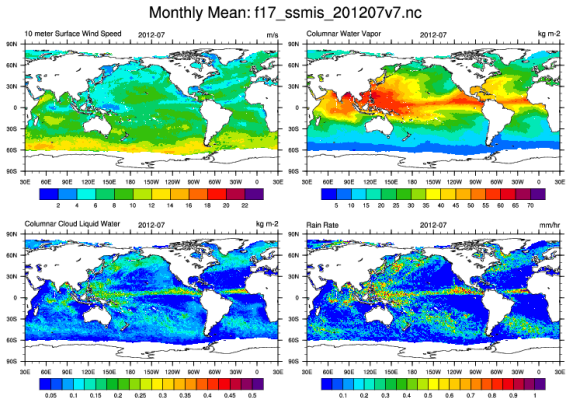


Marie Curie's notebook aip.org

hudsonalpha.org



http://www.census.gov/population/cen2000/map02.gif



ncl.ucar.edu

http://www.bio.umass.edu/biology/karlstrom/FishFacility.html

LETTERS

A role for self-gravity at multiple length scales in the process of star formation

Alyssa A. Goodman[1,2], Erik W. Rosolowsky[2,3], Michelle A. Borkin[1]†, Jonathan B. Foster[2], Michael Halle[1,4], Jens Kauffmann[1,2] & Jaime E. Pineda[2]

Self-gravity plays a decisive role in the final stages of star formation, where dense cores (size ~0.1 parsecs) inside molecular clouds collapse to form star-plus-disk systems[1]. But self-gravity's role at earlier times (and on larger length scales, such as ~1 parsec) is unclear; some molecular cloud simulations that do not include self-gravity suggest that 'turbulent fragmentation' alone is sufficient to create a mass distribution of dense cores that resembles, and sets, the stellar initial mass function[2]. Here we report a 'dendrogram' (hierarchical tree-diagram) analysis that reveals that self-gravity plays a significant role over the full range of possible scales traced by ^{13}CO observations in the L1448 molecular cloud, but not everywhere in the observed region. In particular, more than 90 per cent of the compact 'pre-stellar cores' traced by peaks of dust emission[3] are projected on the sky within one of the dendrogram's self-gravitating 'leaves'. As these peaks mark the locations of already-forming stars, or of those probably about to form, a self-gravitating cocoon seems a critical condition for their existence. Turbulent fragmentation simulations without self-gravity—even of unmagnetized isothermal material—can yield mass and velocity power spectra very similar to what is observed in clouds like L1448. But a dendrogram of such a simulation[4] shows that nearly all the gas in it (much more than in the observations) appears to be self-gravitating. A potentially significant role for gravity in 'non-self-gravitating' simulations suggests inconsistency in simulation assumptions and output, and that it is necessary to include self-gravity in any realistic simulation of the star-formation process on subparsec scales.

Spectral-line mapping shows whole molecular clouds (typically tens to hundreds of parsecs across, and surrounded by atomic gas) to be marginally self-gravitating[5]. When attempts are made to further break down clouds into pieces using 'segmentation' routines, some self-gravitating structures are always found on whatever scale is sampled[6,7]. But no observational study to date has successfully used one spectral-line data cube to study how the role of self-gravity varies as a function of scale and conditions, within an individual region.

Most past structure identification in molecular clouds has been explicitly non-hierarchical, which makes difficult the quantification of physical conditions on multiple scales using a single data set. Consider, for example, the often-used algorithm CLUMPFIND[7]. In three-dimensional (3D) spectral-line data cubes, CLUMPFIND operates as a watershed segmentation algorithm, identifying local maxima in the position–position–velocity (p–p–v) cube and assigning nearby emission to each local maximum. Figure 1 gives a two-dimensional (2D) view of L1448, our sample star-forming region, and Fig. 2 includes a CLUMPFIND decomposition of it based on ^{13}CO observations. As with any algorithm that does not offer hierchically nested or overlapping features as an option, significant emission found between prominent clumps is typically either appended to the nearest clump or turned into a small, usually 'pathological', feature needed to encompass all the emission being modelled. When applied to molecular-line

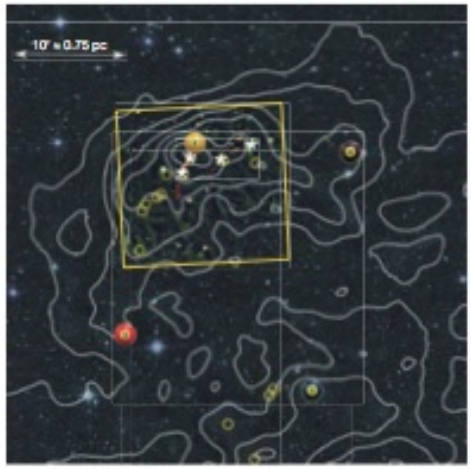


Figure 1 | Near-infrared image of the L1448 star-forming region with contours of molecular emission overlaid. The channels of the colour image correspond to the near-infrared bands J (blue), H (green) and K (red), and the contours of integrated intensity are from ^{13}CO(1–0) emission[8]. Integrated intensity is monotonically, but not quite linearly (see Supplementary Information), related to column density[18], and it gives a view of 'all' of the molecular gas along lines of sight, regardless of distance or velocity. The region within the yellow box immediately surrounding the protostars has been imaged more deeply in the near-infrared (using Calar Alto) than the remainder of the box (2MASS data only), revealing protostars as well as the scattered starlight known as 'Cloudshine'[11] and outflows (which appear orange in this colour scheme). The four billiard-ball labels indicate regions containing self-gravitating dense gas, as identified by the dendrogram analysis, and the leaves they identify are best shown in Fig. 2a. Asterisks show the locations of the four most prominent embedded young stars or compact stellar systems in the region (see Supplementary Table 1), and yellow circles show the millimetre-dust emission peaks identified as star-forming or 'pre-stellar' cores[3].

[1]Initiative in Innovative Computing at Harvard, Cambridge, Massachusetts 02138, USA. [2]Harvard-Smithsonian Center for Astrophysics, Cambridge, Massachusetts 02138, USA. [3]Department of Physics, University of British Columbia, Okanagan, Kelowna, British Columbia V1V 1V7, Canada. [4]Surgical Planning Laboratory and Department of Radiology, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts 02115, USA. †Present address: School of Engineering and Applied Sciences, Harvard University, Cambridge, Massachusetts 02138, USA.

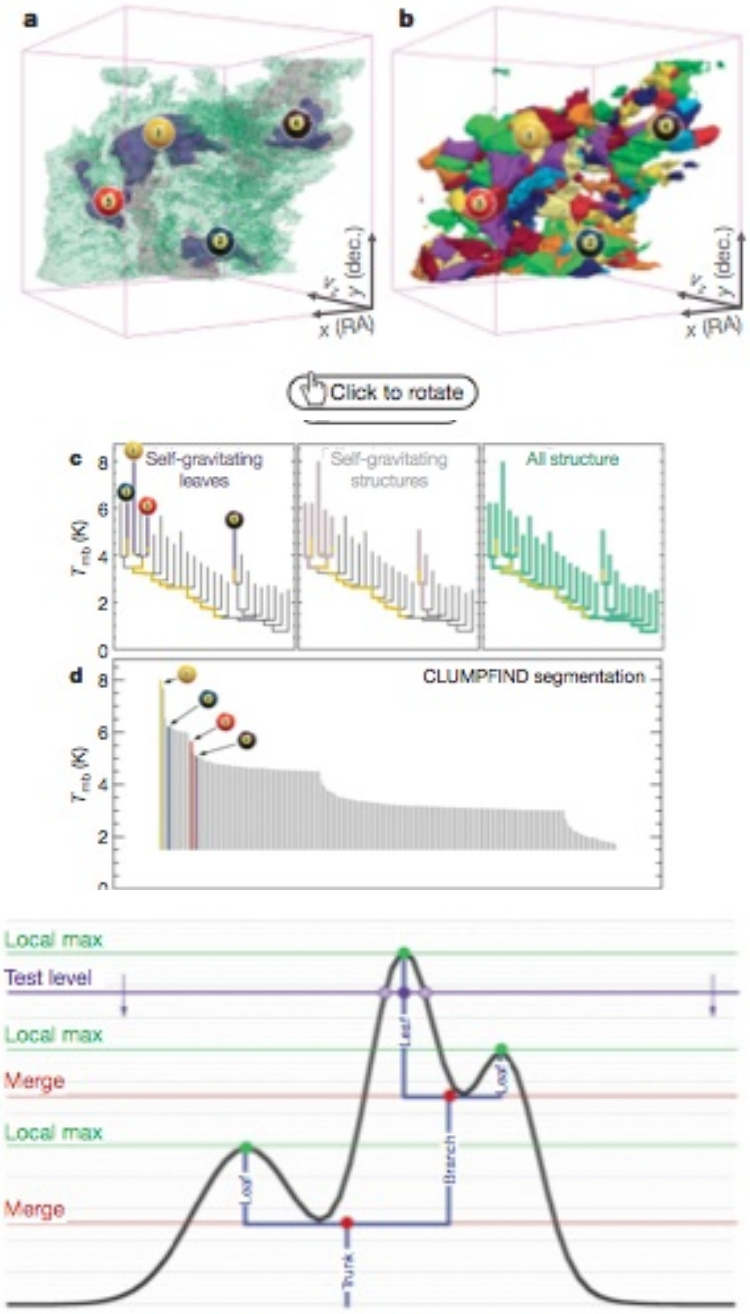
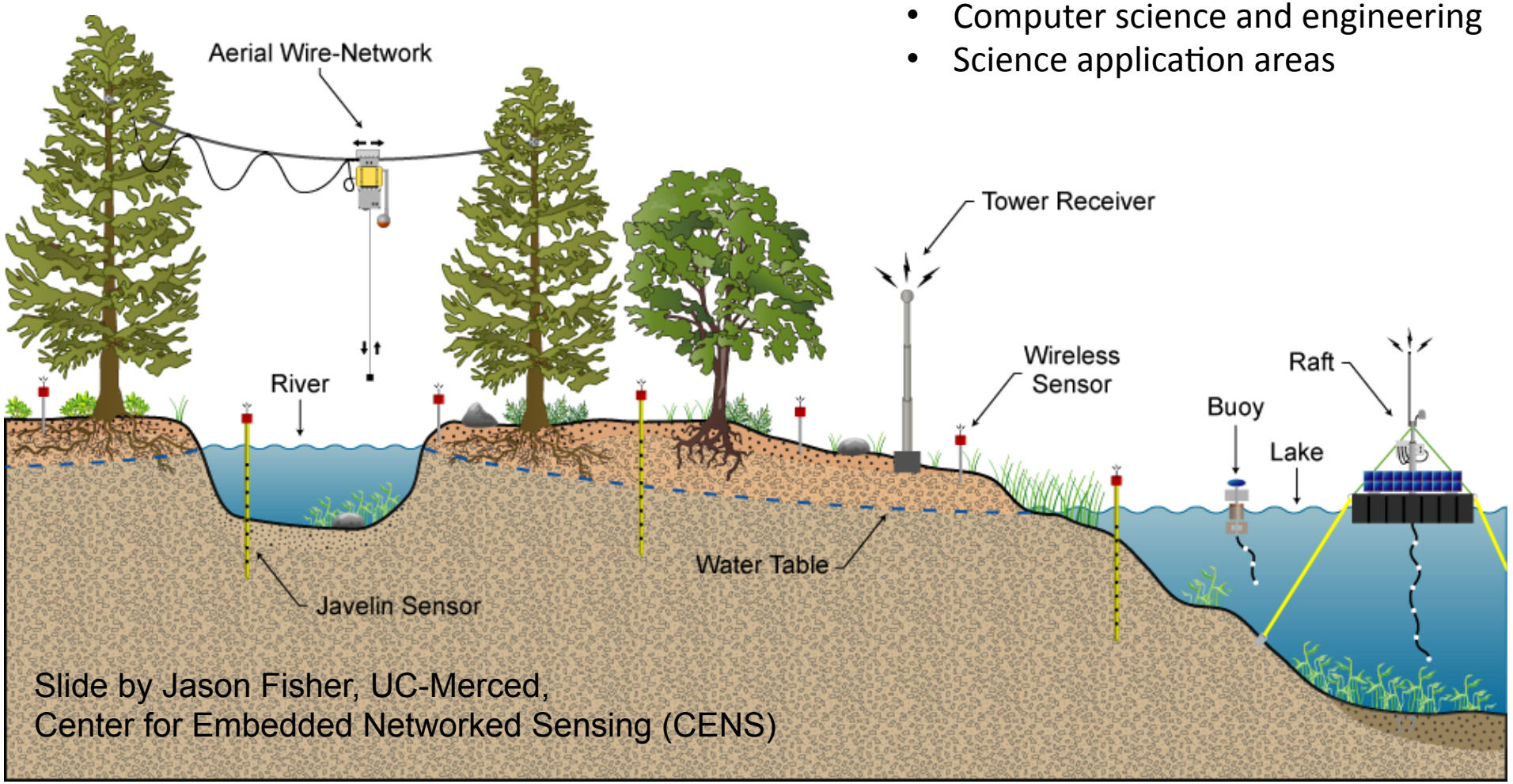


Figure 3 | Schematic illustration of the dendrogram process. Shown is the

Center for Embedded Networked Sensing

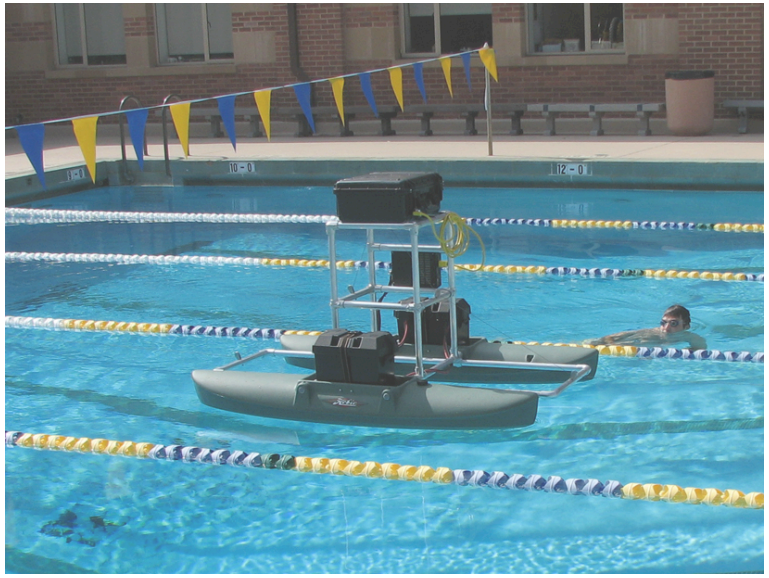
- NSF Science & Tech Ctr, 2002-2012
- 5 universities, plus partners
- 300 members
- Computer science and engineering
- Science application areas



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.."*

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

Repository for seafloor cores. Photo: Peter Darch

Researchers' perspectives on data sharing

- Rewards
- Responsibility
- Data
- Incentives

Persistent URL: photography.si.edu/SearchImage.aspx?id=5799
Repository: Smithsonian Institution Archives

Incentives

- Publications that report the research

Vs.

- Data that are reusable by others

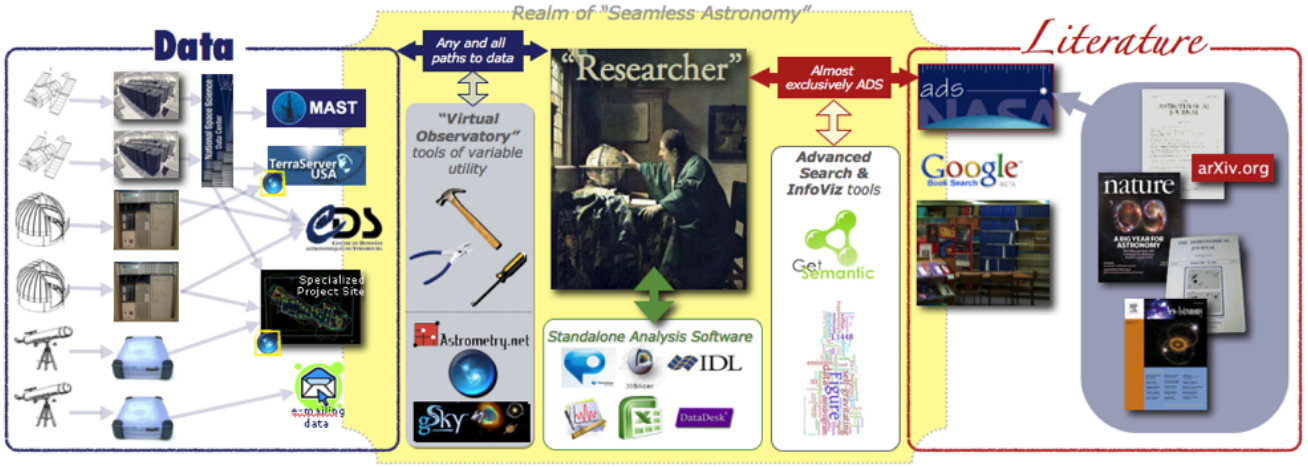
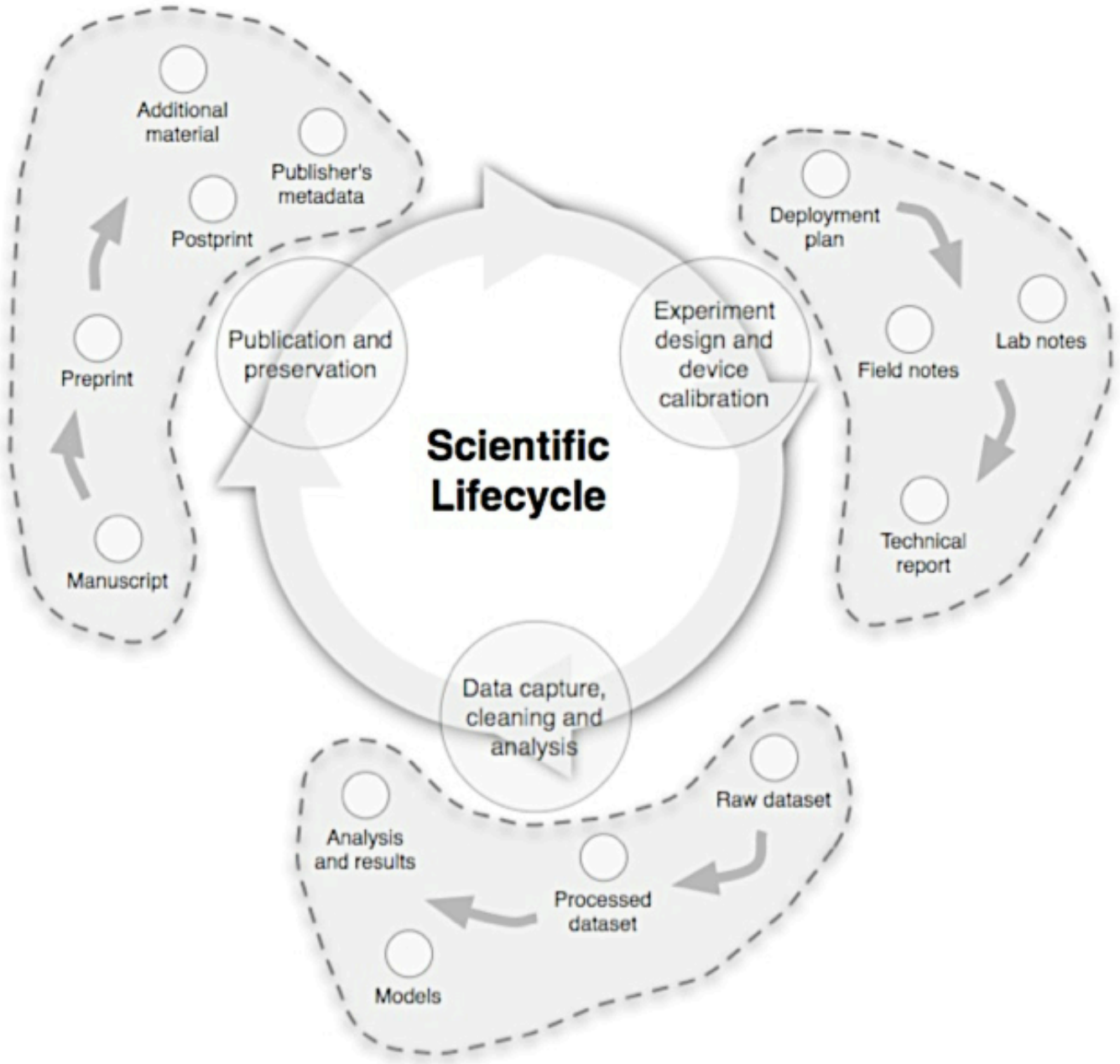


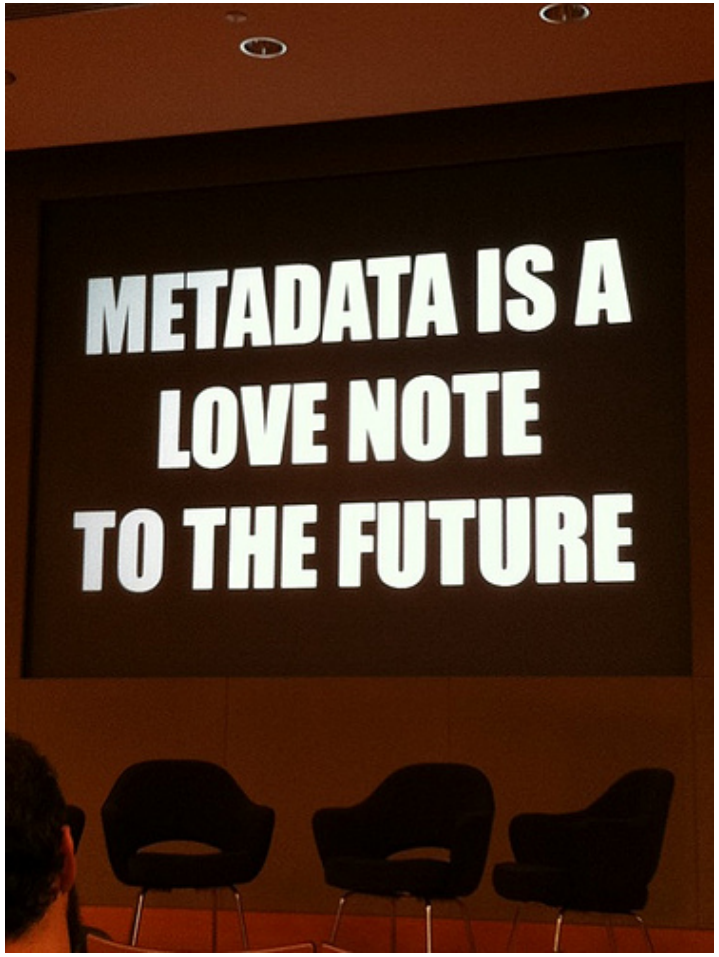
Image: Alyssa Goodman, Harvard Astronomy



Pepe, A., Mayernik, M. S., Borgman, C. L. & Van de Sompel, H. (2010). From Artifacts to Aggregations: Modeling Scientific Life Cycles on the Semantic Web. Journal of the American Society for Information Science and Technology, 61(3): 567–582.

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



*National Information Standards Organization 2004

photo by @kissane

Provenance

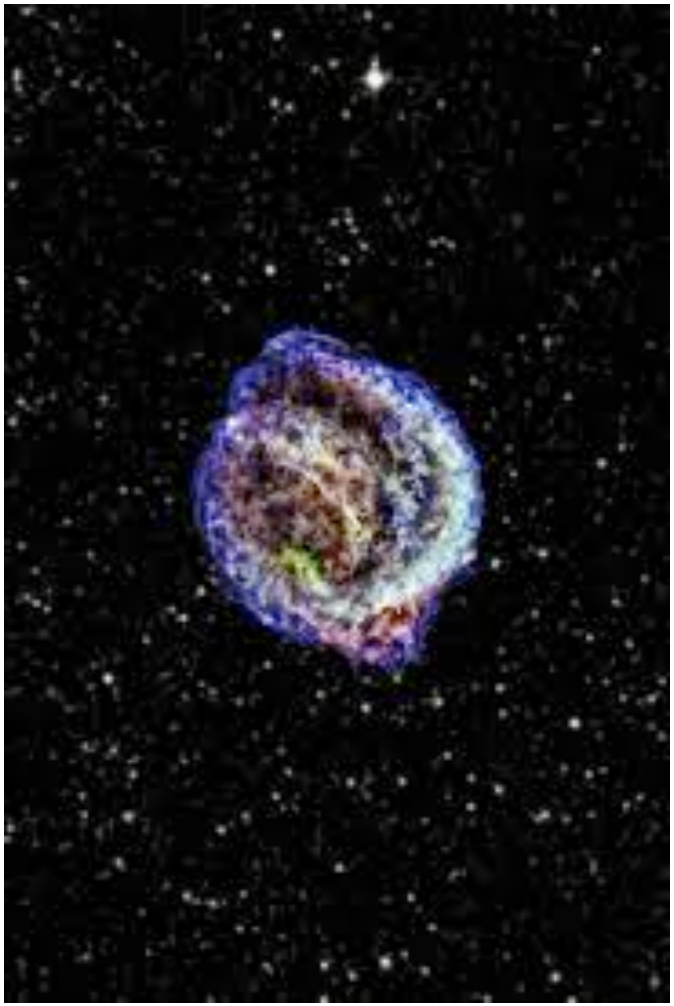
- Libraries: Origin or source
- Museums: Chain of custody
- Internet: Provenance is information about entities, activities, and people involved in producing a piece of data or thing, which can be used to form assessments about its quality, reliability or trustworthiness.*

*World Wide Web Consortium (W3C) Provenance working group

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

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



http://chandra.harvard.edu/photo/2013/kepler/kepler_525.jpg

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.

Q to explore in FaceBase community

- How do you assign credit and responsibility for data creation, curation, use, and reuse?
- How will you balance discipline/species-specific data models and policies with integrative models?
- What data do you expect you to share, with whom, how, and for how long?
- What scientific value do you expect to gain from sharing data via FaceBase?

knowledge infrastructures UCLA

FaceBase

Q to explore in FaceBase community

- Who invest in data curation, and at what stages of sharing and reuse?
- What is the scope of overlap between contributors and users of FaceBase data?
- What scientific value can users obtain from these data, with what kinds of investments?

knowledge infrastructures UCLA

FaceBase

Acknowledgements

UCLA Data Practices team

- Peter Darch, Milena Golshan, Irene Pasquetto, Ashley Sands, Sharon Traweek
- Former members: Rebekah Cummings, David Fearon, Ariel Hernandez, Elaine Levia, Jaklyn Nunga, Matthew Mayernik, Alberto Pepe, Kalpana Shankar, Katie Shilton, Jillian Wallis, Laura Wynholds, Kan Zhang

- Research funding: National Science Foundation, Alfred P. Sloan Foundation, Microsoft Research
- University of Oxford: Balliol College, Oliver Smithies Fellowship, Oxford Internet Institute, Oxford eResearch Center, Bodleian Library