Title
Data, Management, and Digital Science

Author
Borgman, Christine L.

Publication Date
2015-06-04

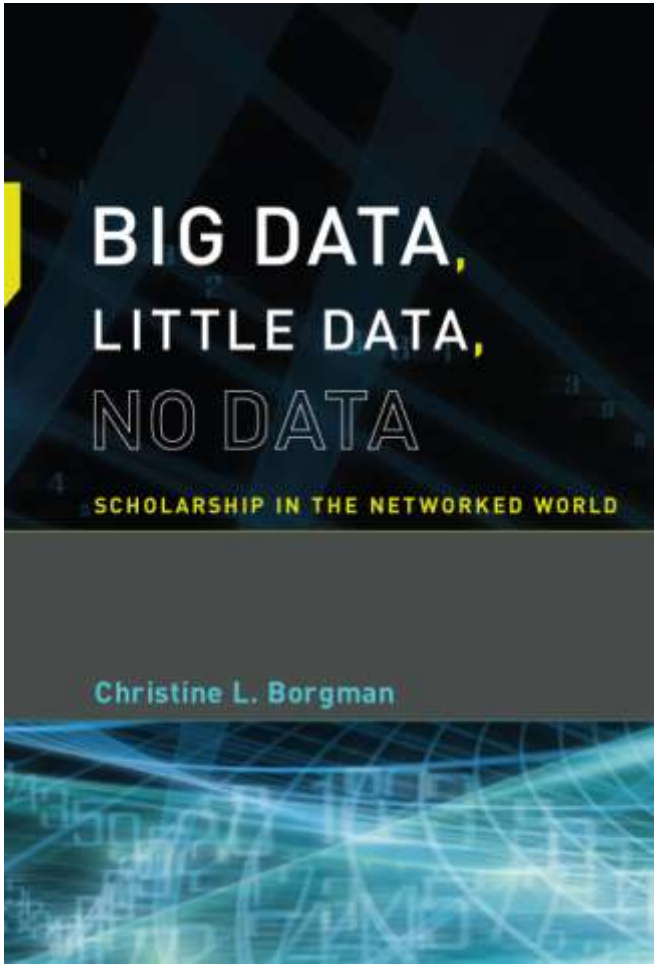
Data, Management, and Digital Science

Christine L. Borgman

Professor and Presidential Chair in Information Studies

University of California, Los Angeles

@scitechprof



Keynote presentation

Digital Science Showcase

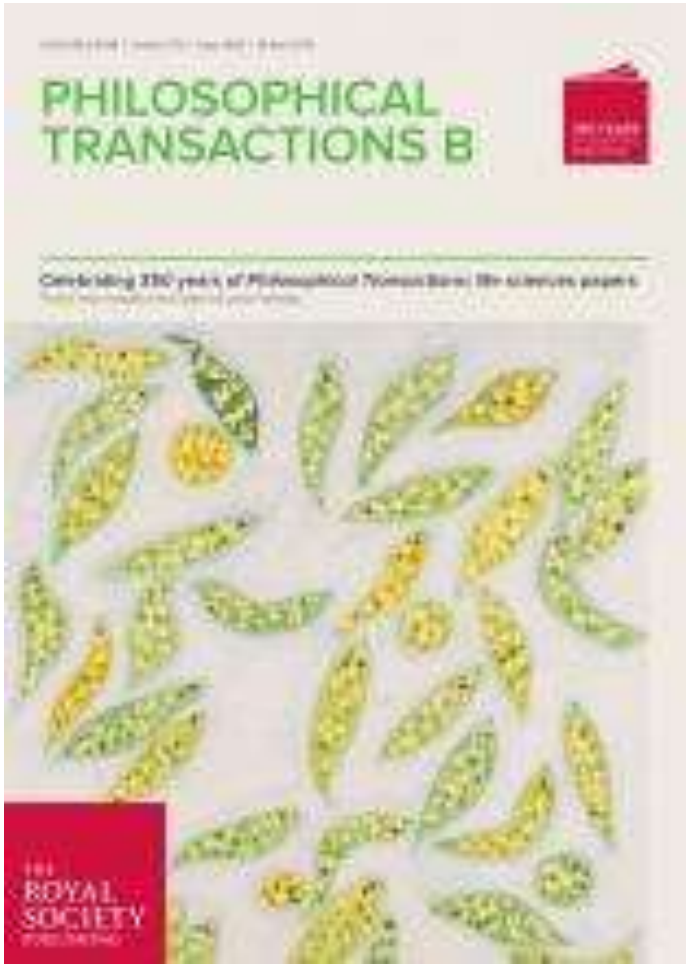
Los Angeles

June 4, 2015

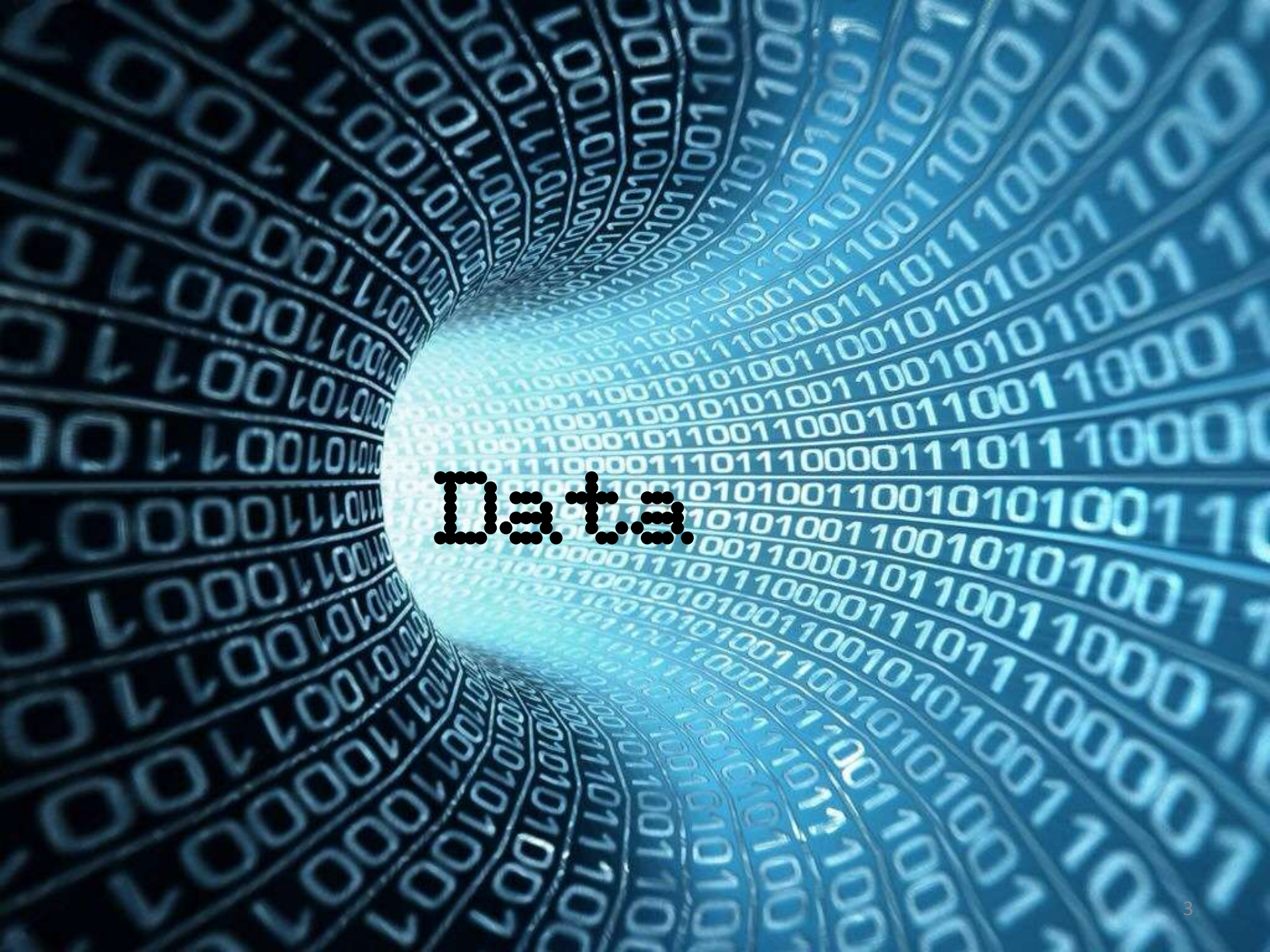
PHILOSOPHICAL
TRANSACTIONS:
GIVING SOME
ACCOMPT
OF THE PRESENT
Undertakings, Studies, and Labours
OF THE
INGENIOUS
IN MANY
CONSIDERABLE PARTS
OF THE
WORLD

Vol I.
For *Anno* 1665, and 1666.

In the *SAVOY*,
Printed by *T. N.* for *John Martyn* at the Bell, a little without *Temple-Bar*, and *James Allestry* in *Duck-Lane*,
Printers to the *Royal Society*.



Theme issue 'Celebrating 350 years of Philosophical Transactions: life sciences papers' compiled and edited by Linda Partridge
19 April 2015; volume 370, issue 1666


Data

Open access policies

- Australian Research Council
  - Code for the Responsible Conduct of Research
  - Data management plans
- National Science Foundation
  - Data sharing requirements
  - Data management plans
- U.S. Federal policy
  - Open access to publications
  - Open access to data
- European Union
  - European Open Data Challenge
  - OpenAIRE
- Research Councils of the UK
  - Open access publishing
  - Provisions for access to data

Australian Government
National Health and Medical Research Council

National Science Foundation
WHERE DISCOVERIES BEGIN

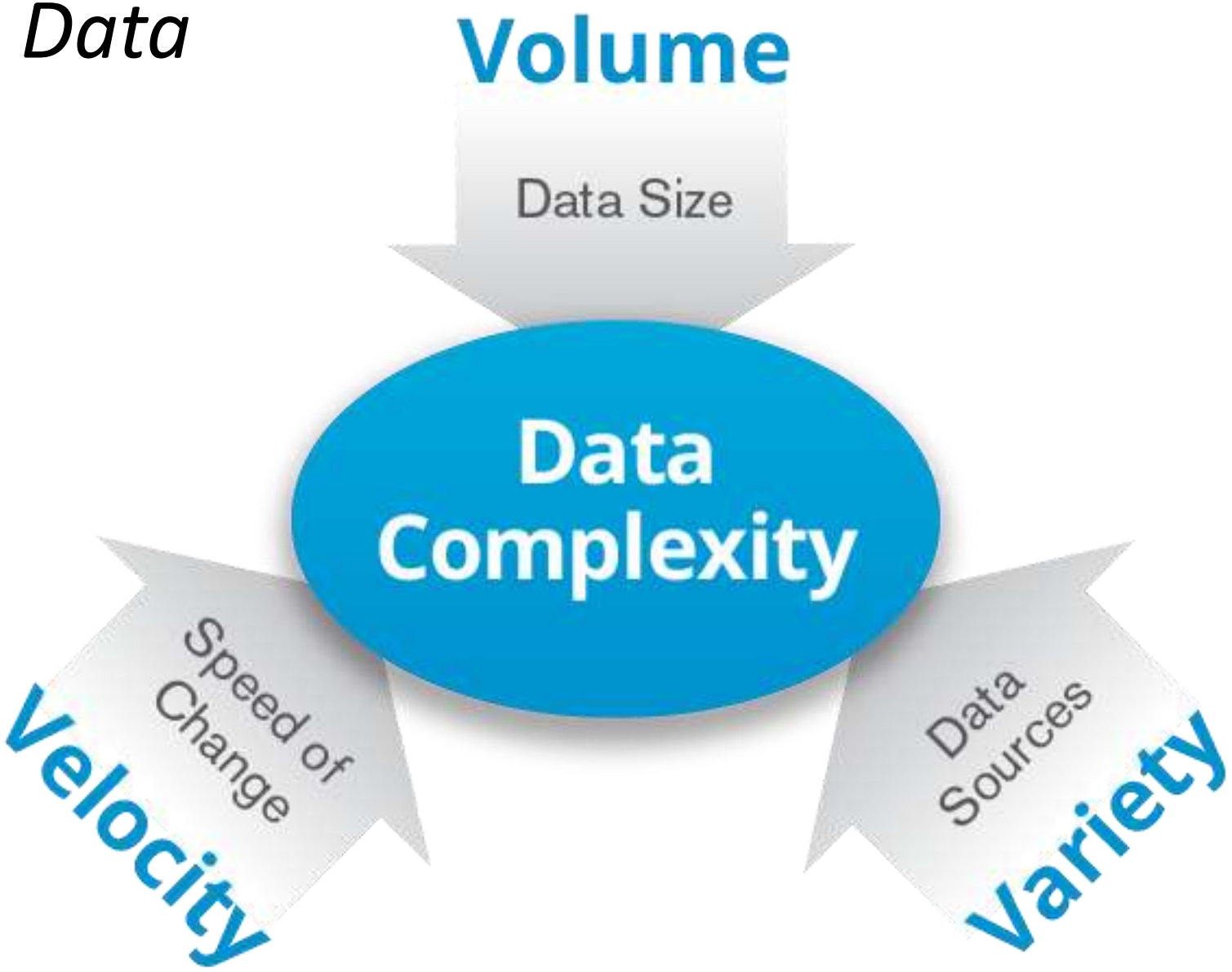
Supported by wellcome trust

Research Data Sharing without barriers

Precondition:

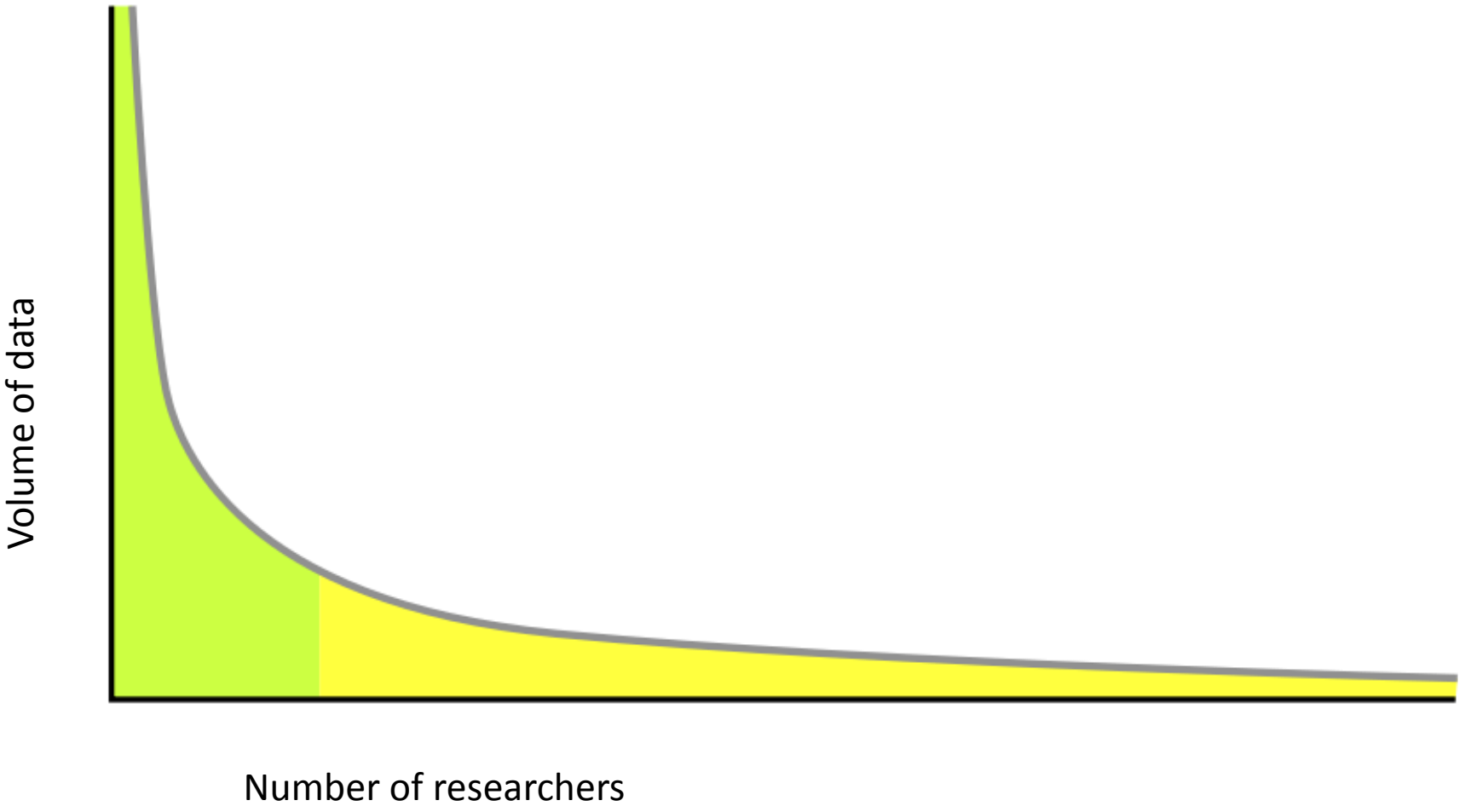
Researchers share data

Big Data



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

Long tail of data



Slide: The Institute for Empowering Long Tail Research

Open Data: Free

- A piece of data or content is open if anyone is free to use, reuse, and redistribute it — subject only, at most, to the requirement to attribute and/or share-alike

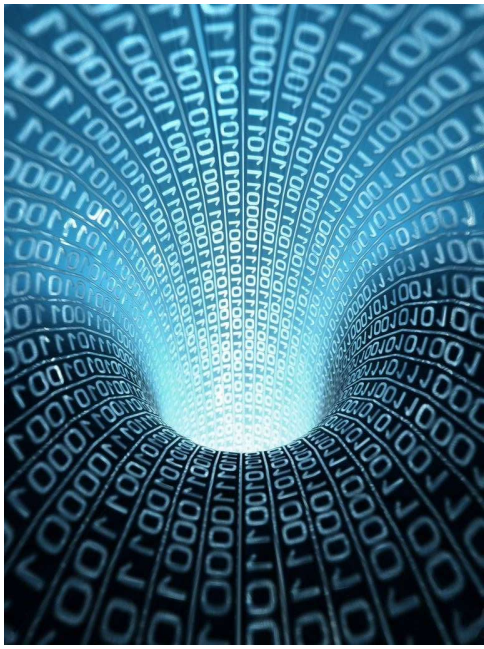
State Library and Archives of Florida, 1922. Flickr commons photo

Open Data Commons. (2013).

```
/FontMatrix matrix def
/FontBBox[2048 -1164 1 index div -628 2 index div 4096 3 index div 2062 5 -1 rol
/sfnts [<
747275650009000000000000637674200000000000000009C000007DA6670676D0000000000000878
019901AC01C101C501C901E101F601F601F60222022202280236023F0243024602670285028502940
01160125011800EA00EA00AE0000003E05BB008A04D70053003FFF8CFFD500150028002200990062
B200402F2B59B002602D2C21B0C051580C6423648BB81555621BB200802F2B59B002602D2C0C6423
B8FF80B30809341CB8FF8040E80809343609352F4A0959045809A7090626062B082C0B280C281328
D4401E091202550C0C047F180118401112025518401D1D02551810251E073C2C04002FED3FEDC42B
1700030209131415002159020071B25090B1E0505262500180C0D0D065518021010065518OCB8FFF8
1F1E4A62A1A4FB432E7900020044FFE40405058E001F002D024BB10202435458402 62F4010100255
2F2B2BDD2B2BC01112392F2BCDD0CD003FED5DC45D5D2B323FED12392F5D5D5DCD31301B4019125F
11242517012B50100110302A2912110608070A101B011BB80152B38F2D012DB802F6B2012A2BB801
025D5972103C103C111217391112393911123901111239391239393130437940 7A4B573A4524351C
2D602D702D04802D902D02B02D01002D102DC02DD02D042D6037A67F182B10F65D5D71723CFD3C10
FFEAB40C0C02550CB8FFE2B40D0D02550CB8FFD6B4101002550CB8FFDEB50F0F02550C05BD03E200
2F2F2F3FDDCD3F3F10ED10ED313001B00D4B5458BF0046FFE80045FFE8002FFFE80030FFE8B51A18
1010065529B8FFF2B70F0F0655292935341112392F2B2B2BDD2B2B2BC01112392F2B2B2BCD2B2B2B
3236353427260200D07E6B76CF7FCF7A677DCC53356B429F82617E694703AF9E87AF7BFC80A58BAD
961A9C1E9621982A9E2BA816A61AAB1CAD2BB916BE2BCD2BDA2BEC2BFB2B2A202D732573288F1397
201A01601A701A021A120B003FC45D5DED5D5D2F3FC45DED5D5D5D1217393F012F2B2BCD2F2BCDD4
363702902126775C4656201F5F92CBBD75546C2115170D21211C9E62455761FEDE2D2D9B7B364D33
393FDD5DCD31301B40350127600D5D362027602770 27B02704340B371F3A20481F4820051A08134F
FFE0400A1339082013391B201339012B2B2B2B002B012B2B5959591321152322061514171313 3635
C0B215391AB8FFF040131539360815392830143929301439110816390 9B8FFE0401B163929401139
0000000900010000000000000000100000721FE4500571000FB74FADF1000000100000000000000
00340363004403630004036 3B2242F1FBA034E006D0800400E1F7F027F037F047F050430440112BF
014A001F000D0126400B1F0DC61F0D571F0D371F0D410D019E0141000D00420141000D001E014100
021E0024455258B90024021E4459594BB8020153205C58B9010F00224544B1222245445958B90C00
2B2B2B2B2B2B2B2B2B2B2B2B2B2B2B0001737500737300456944007373017374 2B2B2B2B2B732B00732B
00>] def
/CharStrings 27 dict dup begin
/.notdef 0 def
            /space 1 def
                        /comma 2 def
                                    /period 3 def
                                                /W 4 def
                                                        /a 5 def
                                                                /b 6 def
                                                                        /c 7
```

Open Data: Useful

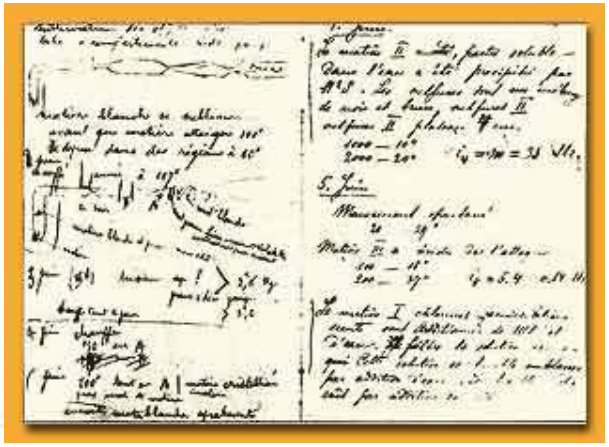
- Openness, flexibility, transparency, legal conformity, protection of intellectual property, formal responsibility, professionalism, interoperability, quality, security, efficiency, accountability, and sustainability.



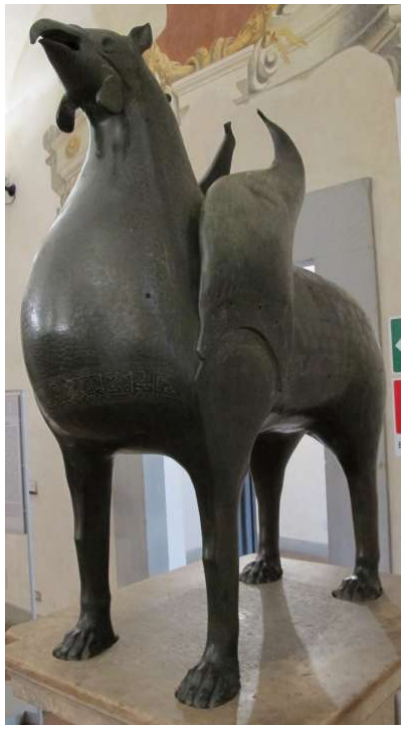
Organization for Economic Cooperation and Development. (2007). *OECD Principles and Guidelines for Access to Research Data from Public Funding*. http://www.oecd.org/dataoecd/9/61/38500813.pdf

What are data?

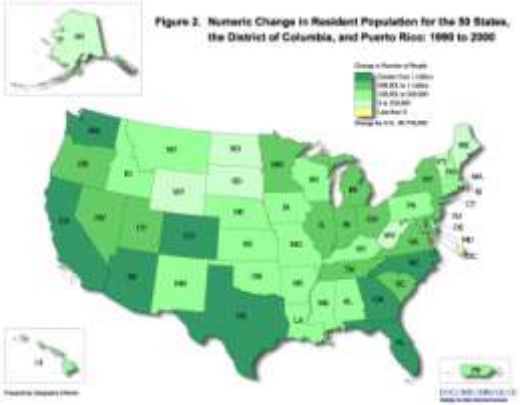
hudsonalpha.org



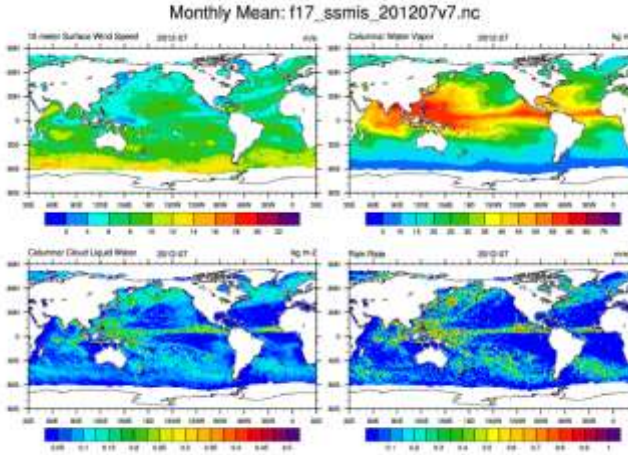
Marie Curie's notebook aip.org



Pisa Griffin



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



ncl.ucar.edu

Date:1/2.07.75 Place:Sakaltutan
Zafor
He will grow old in his present house; new house is for sons - 5 sons. Not sure they want to live in village. He will only build another if they want him to. eS came from Germany and did the plastering. He arranged the carpentry in Kayseri. Çok para gitti. {much money went} Has a tractor.

Date:July1980 Place:Sakaltutan
Zafor:
Household now Zafor and wife; Nazif Unal and wife and youngest son, still a boy. They run two dolmuß; one with a driver from Süleymanli. Goes in and out once a day. He gets 8,000 a month. Zafor then said, keskin de©il. { not sharp - i.e.? not profitable} I said he did very well on 8,000 TL with only two journeys a day. Nazif Unal has "bought" a Durak {dolmuß stop} from Belediye and works all day in Kayseri.

http://onlineqda.hud.ac.uk/Intro_QDA/Examples_of_Qualitative_Data.ph

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

WIRED MAGAZINE: 16.07

SCIENCE : DISCOVERIES

The End of Theory: The Data Deluge Makes the Scientific Method Obsolete

By Chris Anderson 06.23.08

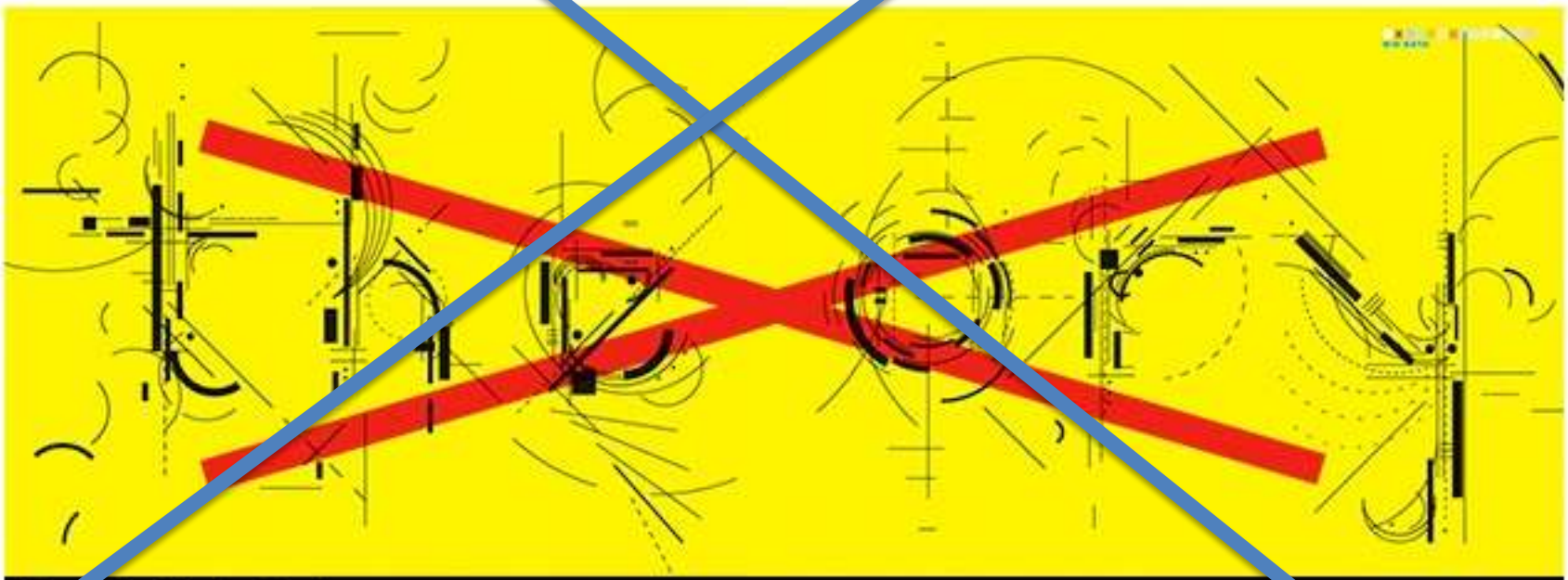
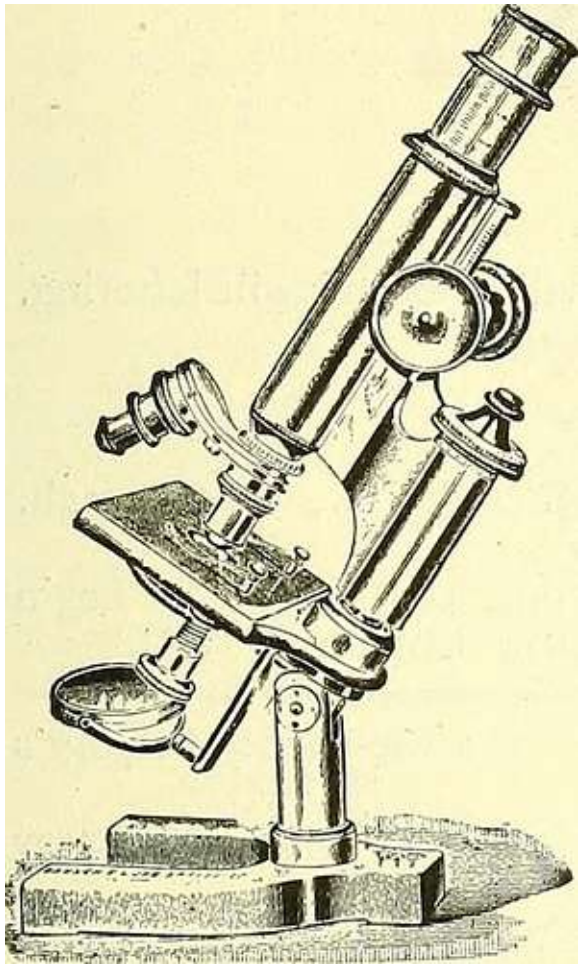


Illustration: Marian Bantjes

Research process

- Models and theories
- Research questions
- Methods
  - Tools
  - Data sources
  - Practices
  - Infrastructure
  - Domain expertise



Commons photo: Science Gossip, 1894

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

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[14], 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.

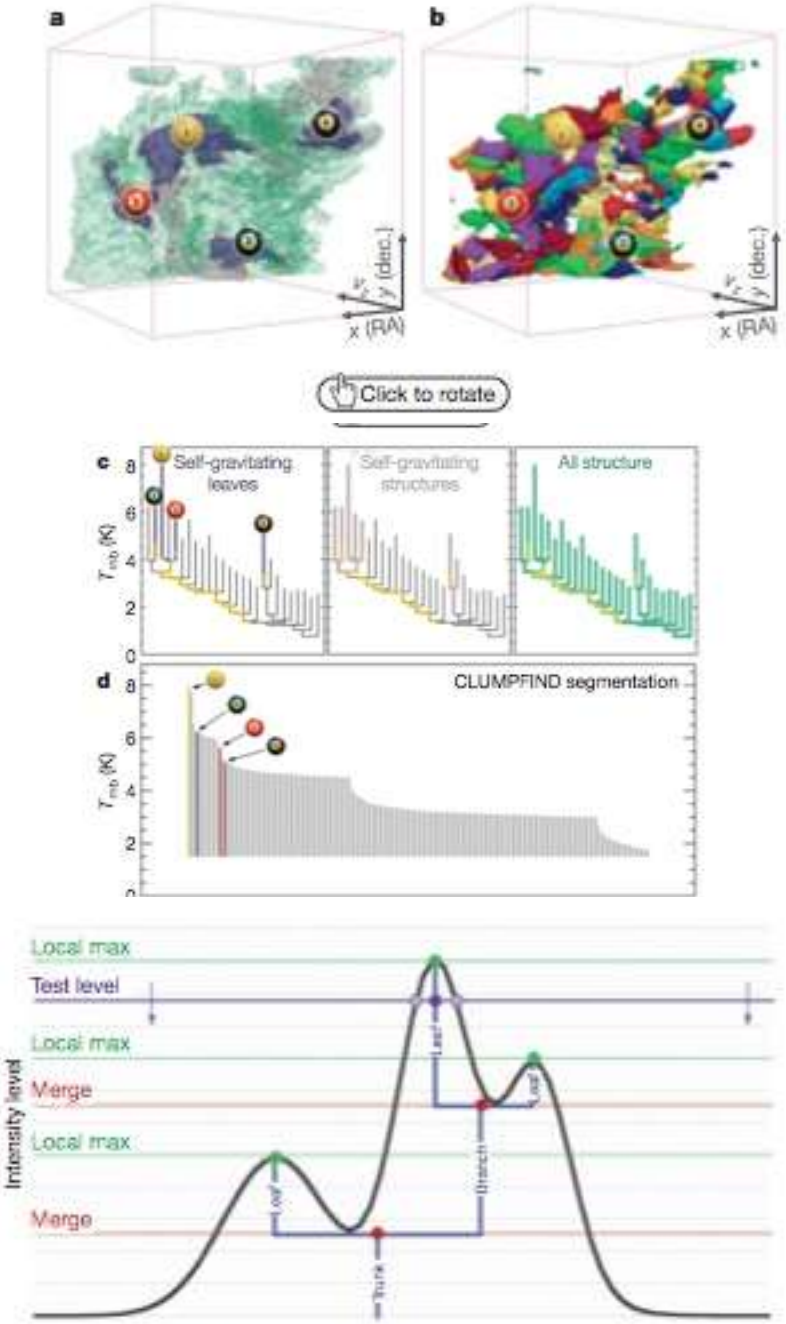
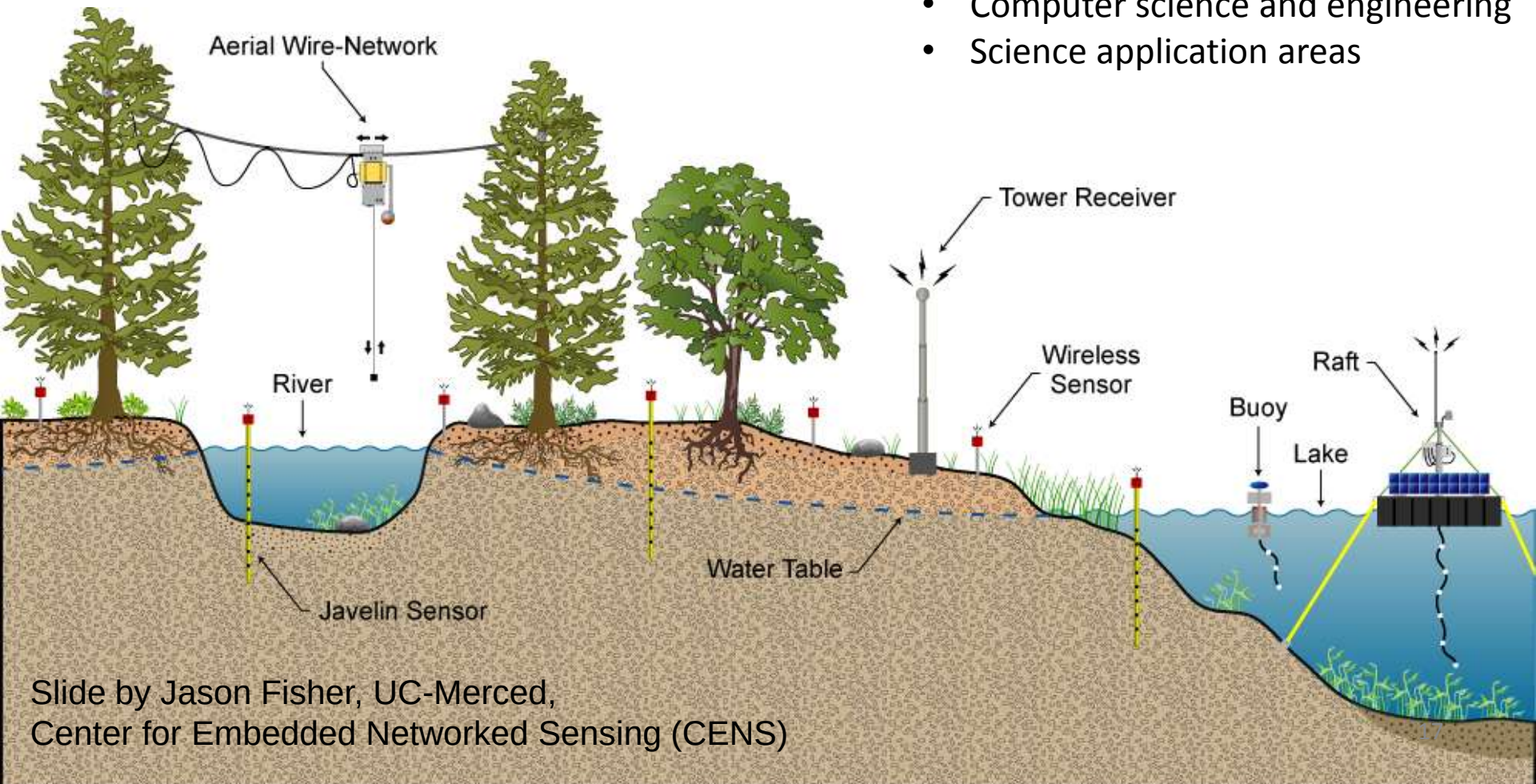


Figure 3 | Schematic illustration of the dendrogram process. Shown is the

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)

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.."*

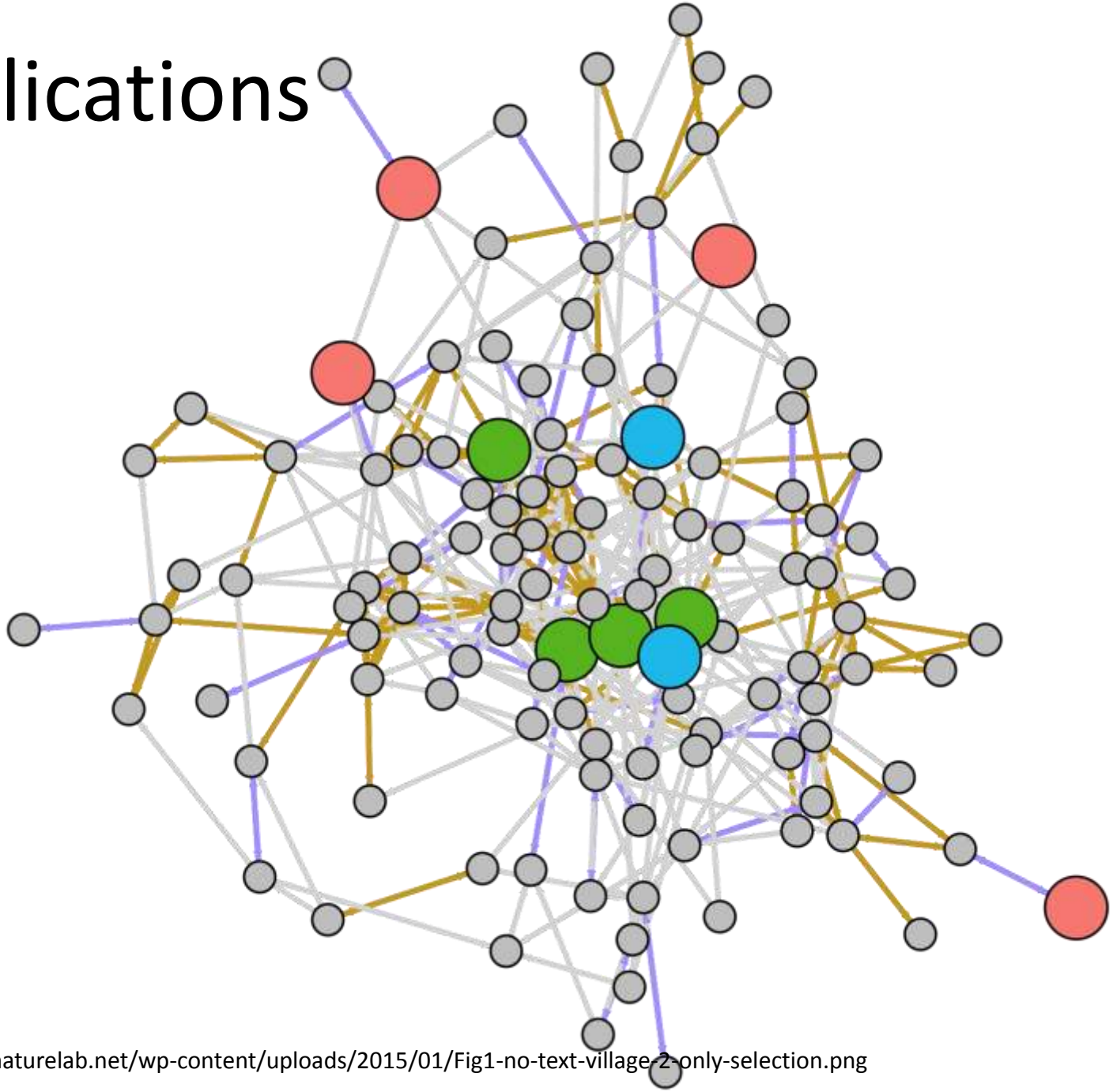
http://vcg.isti.cnr.it/griffin/

Arte islamica, ippogrifo, XI sec 03, own work

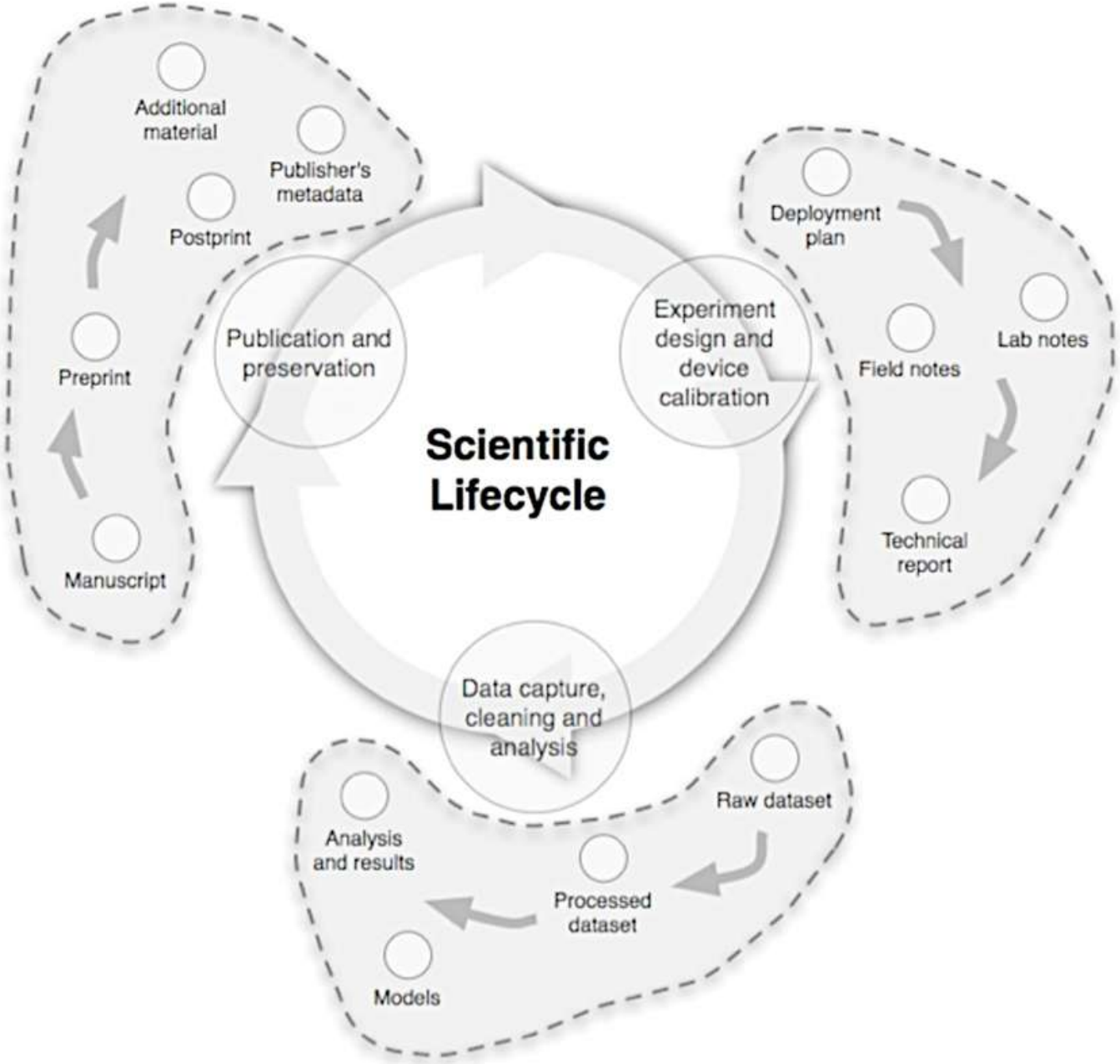
Publications

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



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.

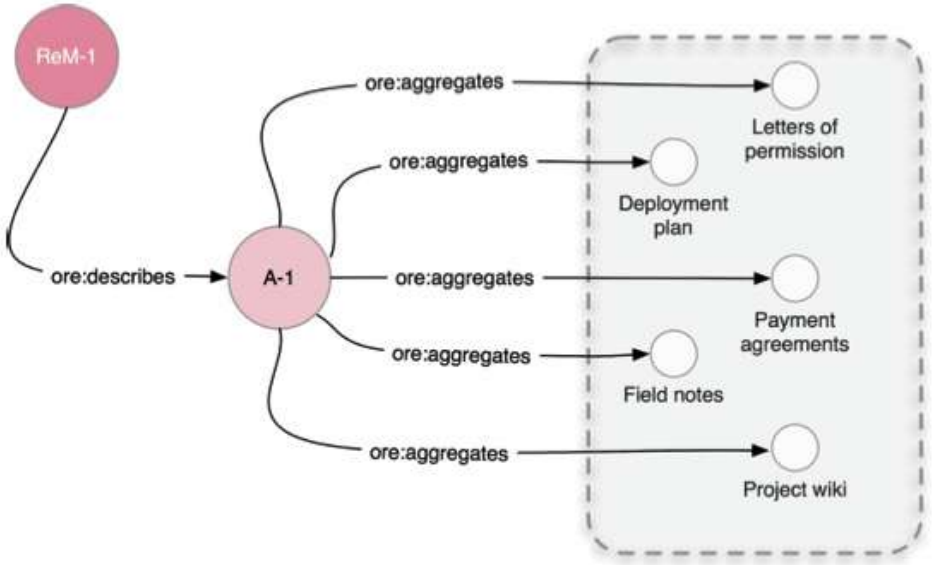


FIG. 4. ORE Aggregation representing the first stage of the scientific life cycle of a sensor network application in seismology (experiment and deployment planning).

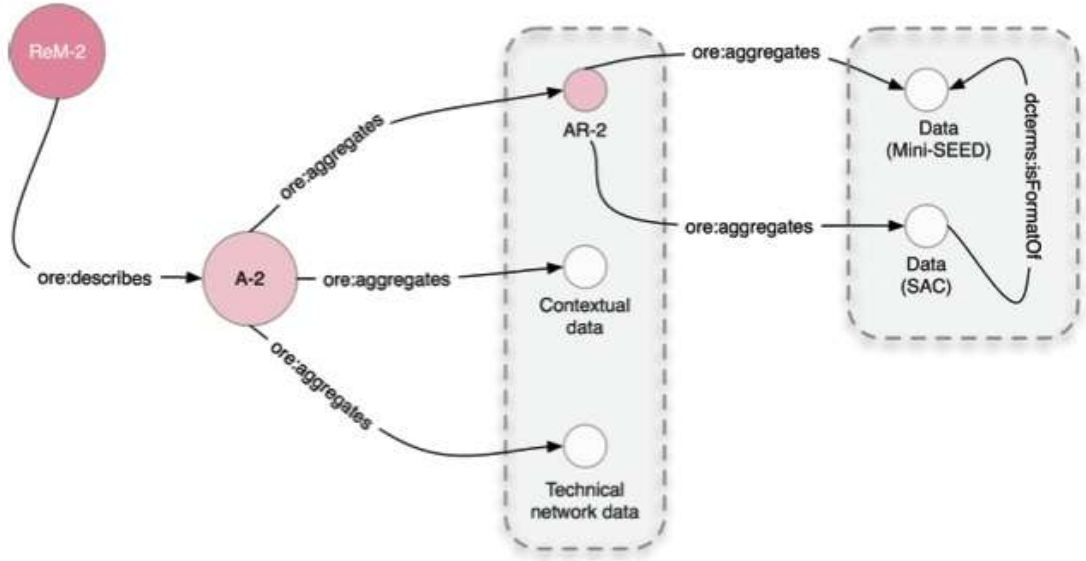
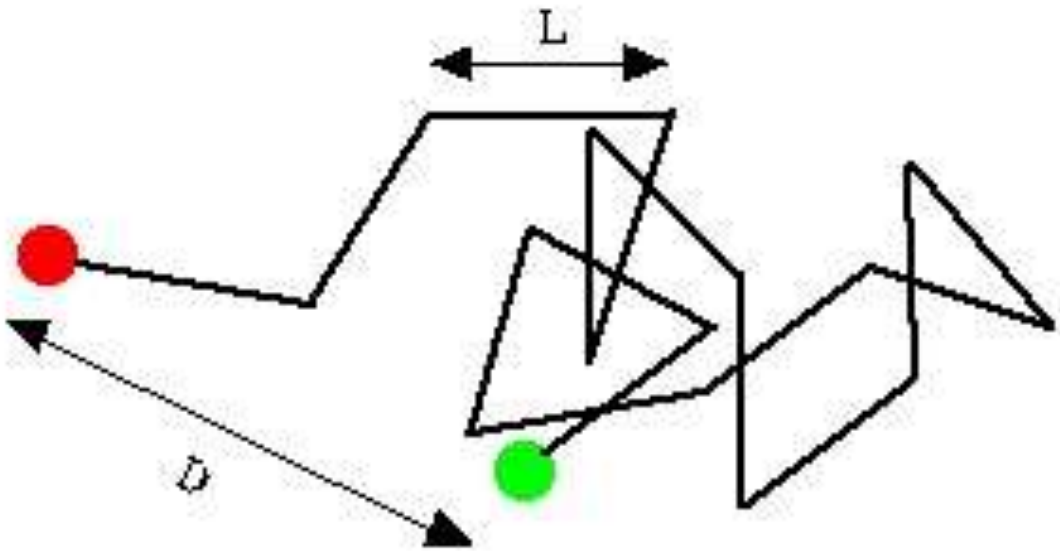


FIG. 5. ORE Aggregation representing the second stage of the scientific life cycle of a sensor network application in seismology (data collection).

Random walk



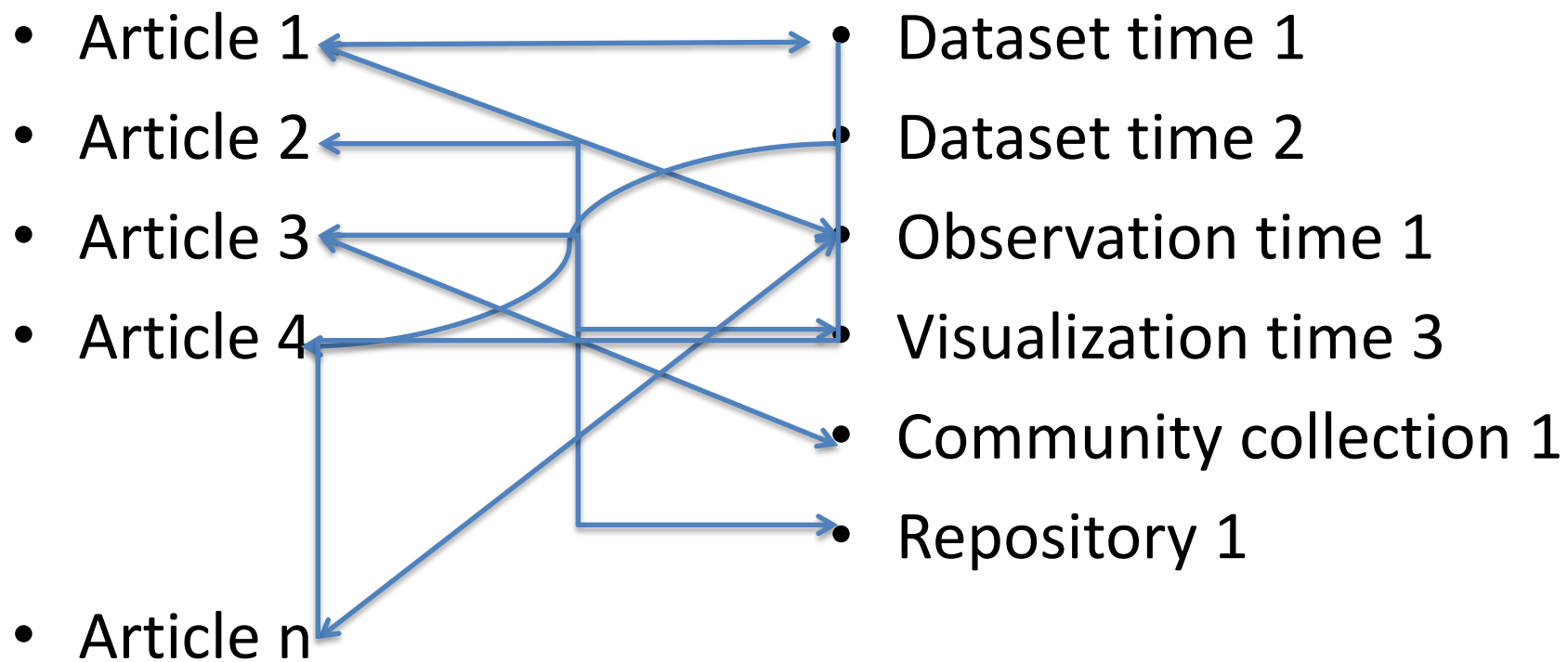
http://www2.ess.ucla.edu/~jewitt/oort2-random.html

Publications <–> Data: Role

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

Publications <–> Data: Mapping



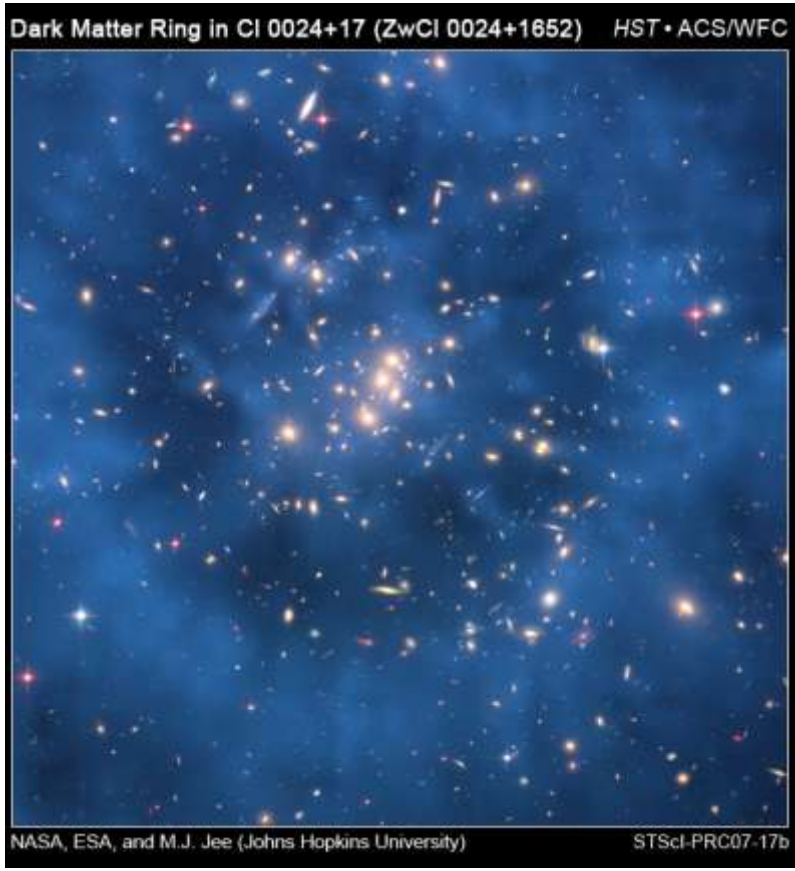
Publications <–> Data: Attribution

- 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

Publications <–> Data: Citations

"If publications are the stars and planets of the scientific universe, data are the 'dark matter' – influential but largely unobserved in our mapping process"*



*Micah Altman, CODATA-ICSTI Task Group on Data Citation Standards and Practices, 2013, p. 54

Data citation and analytics

- Credit
- Attribution
- Discovery

Bibliometrics, Scientometrics, Informetrics, Webometrics…

Broken Promises of Privacy 1709

data—associating stored genes with nonidentifying numbers—to protect privacy.[19] Other guidelines recommend anonymization in contexts such as electronic commerce,[20] internet service provision,[21] data mining,[22] and national security data sharing.[23] Academic researchers rely heavily on anonymization to protect human research subjects, and their research guidelines recommend anonymization generally,[24] and specifically in education,[25] computer network monitoring,[26] and health studies.[27] Professional statisticians are duty-bound to anonymize data as a matter of professional ethics.[28]

Market pressures sometimes compel businesses to anonymize data. For example, companies like mint.com and wesabe.com provide web-based personal finance tracking and planning.[29] One way these companies add value is by aggregating and republishing data to help their customers compare their spending with that of similarly situated people.[30] To make customers comfortable with this type of data sharing, both mint.com and wesabe.com promise to anonymize data before sharing it.[31]

Architecture, defined in Lessig's sense as technological constraints,[32] often forces anonymization, or at least makes anonymization the default choice. As one example, whenever you visit a website, the distant computer with which you communicate—also known as the web server—records some information

19. Roberto Andorno, *Population Genetic Databases: A New Challenge to Human Rights*, in ETHICS AND LAW OF INTELLECTUAL PROPERTY 39 (Christian Lenk, Nils Hoppe & Roberto Andorno eds., 2007).
20. ALEX BERSON & LARRY DUBOV, MASTER DATA MANAGEMENT AND CUSTOMER DATA INTEGRATION FOR A GLOBAL ENTERPRISE 338–39 (2007).
21. *See infra* Part II.A.3.b.
22. G.K. GUPTA, INTRODUCTION TO DATA MINING WITH CASE STUDIES 432 (2006).
23. MARKLE FOUND. TASK FORCE, CREATING A TRUSTED NETWORK FOR HOMELAND SECURITY 144 (2003), *available at* http://www.markle.org/downloadable_assets/nstf_report2_full_report.pdf.
24. *See* THE SAGE ENCYCLOPEDIA OF QUALITATIVE RESEARCH METHODS 196 (Lisa M. Given ed., 2008) (entry for "Data Security").
25. LOUIS COHEN ET AL., RESEARCH METHODS IN EDUCATION 189 (2005).
26. *See* Ruoming Pang et al., *The Devil and Packet Trace Anonymization*, 36 COMP. COMM. REV. 29 (2006).
27. INST. OF MED., PROTECTING DATA PRIVACY IN HEALTH SERVICES RESEARCH 178 (2000).
28. European Union Article 29 Data Protection Working Party, *Opinion 4/2007 on the Concept of Personal Data*, 01248/07/EN WP 136, at 21 (June 20, 2007) [hereinafter 2007 Working Party Opinion], *available at* http://ec.europa.eu/justice_home/fsj/privacy/docs/wpdocs/2007/wp136_en.pdf.
29. *See* Eric Bensdorff, *Spend and Save the Social Way—Personal Technology*, SEATTLE TIMES, Nov. 8, 2008, at A9.
30. *See* Carolyn Y. Johnson, *Online Social Networking Meets Personal Finance*, N.Y. TIMES, Aug. 7, 2007, *available at* http://www.nytimes.com/2007/08/07/technology/07iht-debt.1.7013213.html.
31. *See, e.g.*, Wesabe, Security and Privacy, http://www.wesabe.com/page/security (last visited June 12, 2010); Mint.com, How Mint Personal Finance Management Protects Your Financial Safety, http://www.mint.com/privacy (last visited June 12, 2010).
32. LESSIG, *supra* note 18, at 4.

Ohm, P. (2010). Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization. *UCLA Law Review*, *57*, 1701.

Aad, G., T. Abajyan, B. Abbott, J. Abdallah, S. Abdel Khalek, A. A. Abdelalim, O. Abdinov, et al. 2012. "Observation of a New Particle in the Search for the Standard Model Higgs Boson with the ATLAS Detector at the LHC." *Physics Letters [Part B]* 716 (1):1–29. doi:10.1016/j.physletb.2012.08.020.

Abbate, Janet. 1999. *Inventing the Internet*. Cambridge, MA: MIT Press.

Accomazzi, Alberto. 2010. "Astronomy 3.0 Style." *Astronomical Society of the Pacific Conference Series* 433: 273–281.

Accomazzi, Alberto, and Rahul Dave. 2011. "Semantic Interlinking of Resources in the Virtual Observatory Era." *Astronomical Society of the Pacific Conference Series* 442: 415–424. doi: arXiv:1103.5958.

Acropolis Museum. 2013. "The Frieze." http://www.theacropolismuseum.gr/en/content/frieze-0.

Agosti, Maristella, and Nicola Ferro. 2007. "A Formal Model of Annotations of Digital Content." *ACM Transactions on Information Systems* 26 (1). doi:10.1145/1292591.1292594.

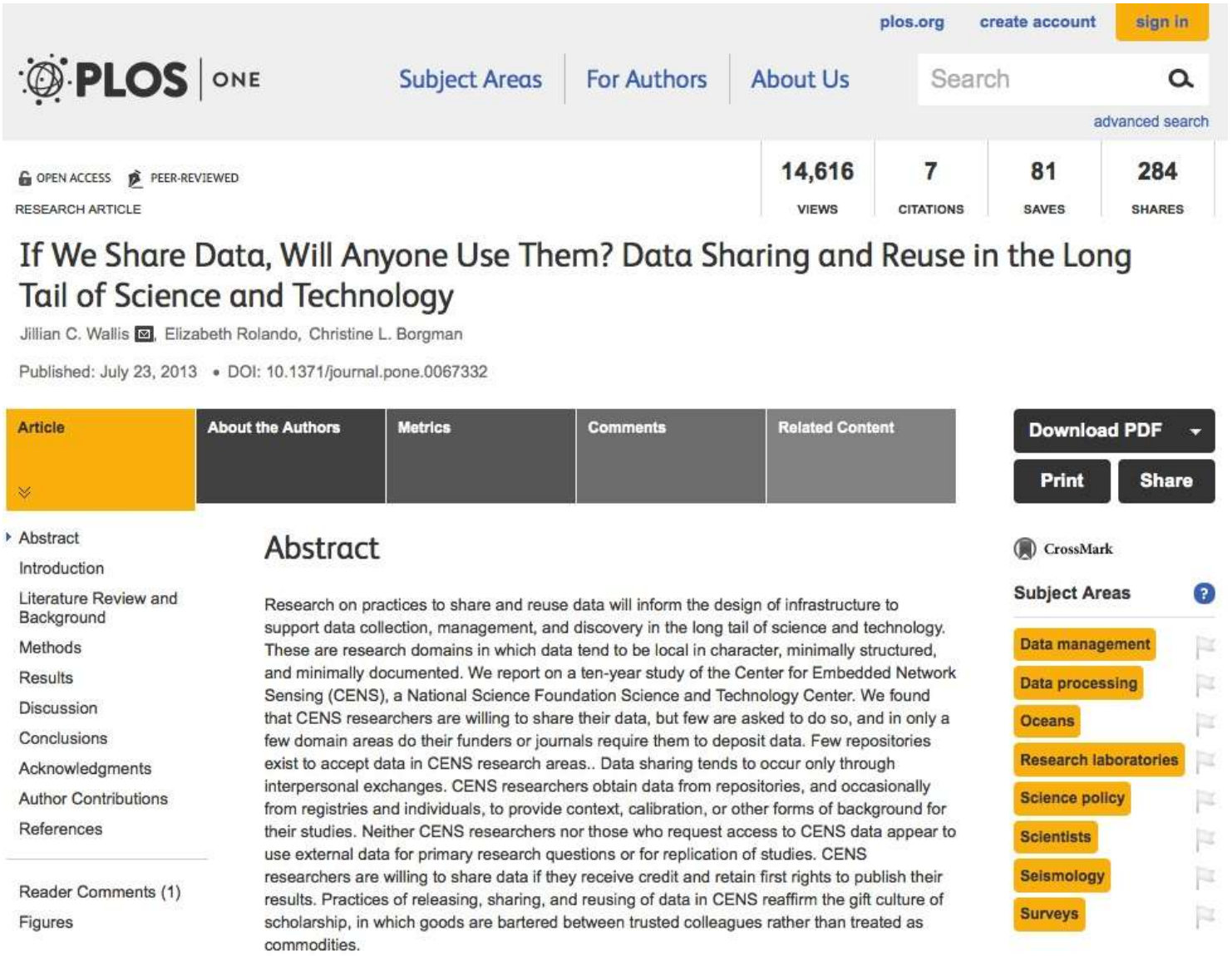
Agre, Philip E. 1994. "From High Tech to Human Tech: Empowerment, Measurement, and Social Studies of Computing." *Computer Supported Cooperative Work* 3 (2):167–195. doi:10.1007/BF00773446.

Ahn, Christopher P., Rachael Alexandroff, Carlos Allende Prieto, Scott F. Anderson, Timothy Anderton, Brett H. Andrews, Éric Aubourg, et al. 2012. "The Ninth Data Release of the Sloan Digital Sky Survey: First Spectroscopic Data from the SDSS-III Baryon Oscillation Spectroscopic Survey." *Astrophysical Journal* 203:21. doi:10.1088/0067-0049/203/2/21.

Akyildiz, I. F., W. Su, Y. Sankarasubramaniam, and E. Cayirci. 2002. "Wireless Sensor Networks: A Survey." *Computer Networks* 38 (4):393–422. doi:10.1016/S1389-1286(01)00302-4.

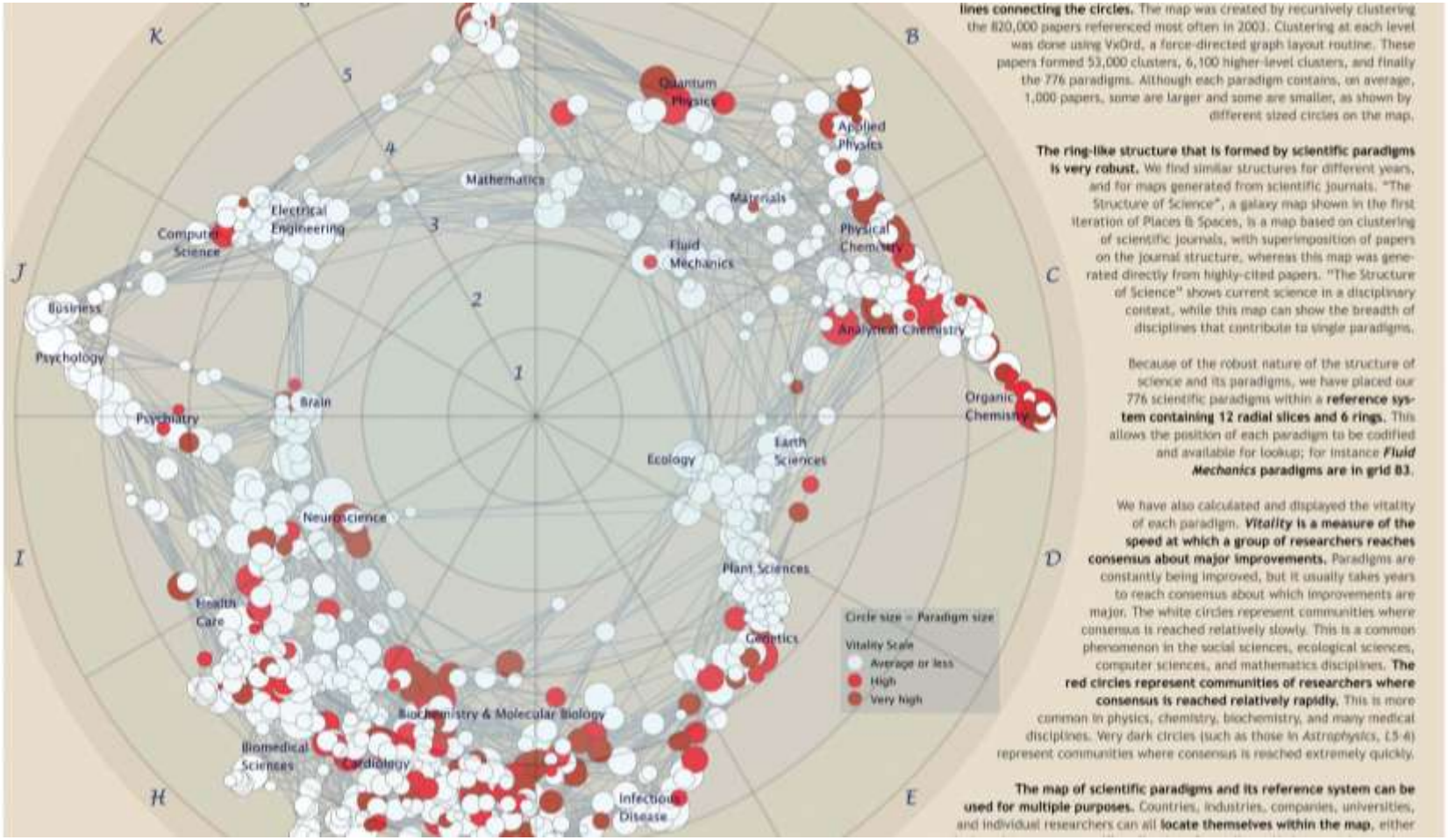
Borgman, C. L. (2015). *Big Data, Little Data, No Data: Scholarship in the Networked World*. Cambridge MA: MIT Press.

Altmetrics



Published July 23, 2013; Altmetrics data as of Nov 5, 2014

Mapping Scholarship



Börner, K. (2010). *Atlas of Science: Visualizing What We Know*. Cambridge, Mass: The MIT Press.

eScholarship
University of California

eScholarship provides Open-Access scholarly publishing services to the University of California and delivers a dynamic research platform to scholars worldwide. Powered by the California Digital Library.

My Account | My Saved Items (0)

Search publications... GO

About eScholarship | Deposit/Publish | Browse Content | Help

UC OPEN ACCESS POLICY

Learn more ▶

DEPOSIT | WAIVER | FAQ

Popular Research This Month

1. Ebola Virus Disease: Essential Public He...
2. An Introduction To Green Marketing
3. How Languages are Learned
4. A Spectral Analysis of World GDP Dynamic...
5. Making Data Count: A Data Metrics Pilot ...

» SEE MORE

Manage Your Publications

Books
Publish online.
Print on demand.

Journals
Peer review and
MS management
made simple.

Working Papers
Share your research
before it gets stale
(or scooped).

Previously Published Works
Open your
publications to
the world.

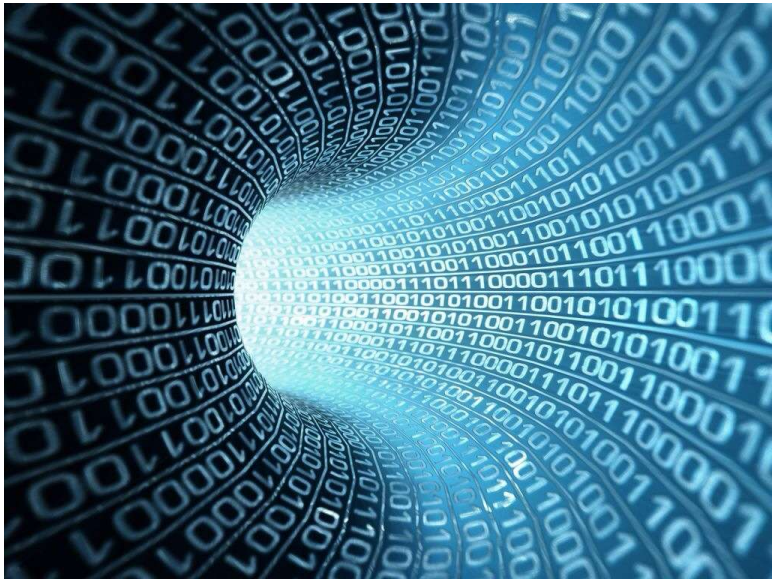
Conferences
Plan your meeting.
Publish the results.

eScholarship by the Numbers

Views Since 2002:	23,216,720
Publications:	75,712
Research Units:	338
Journals:	70

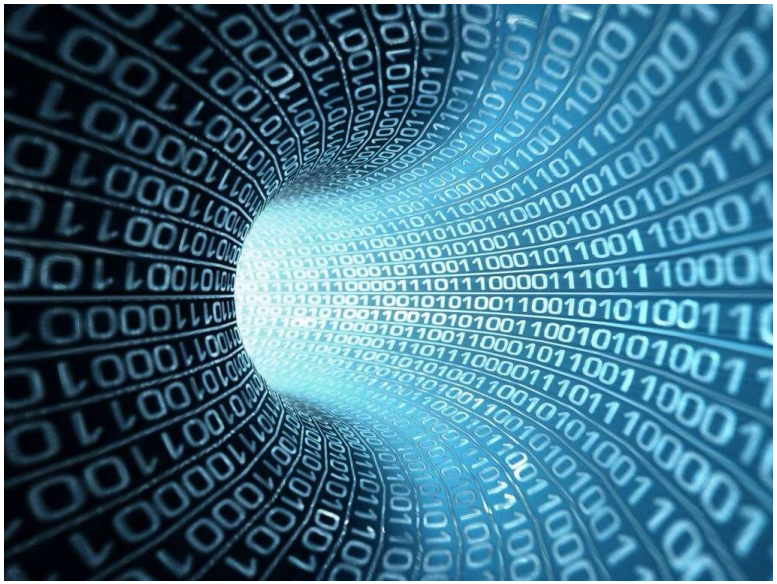
Data issues: Faculty concerns

- Research data
  - Research data management policies
  - Expertise, curation, stewardship, resources
  - Open records laws
  - Human subjects regulations



Data issues: Faculty concerns

- Publication data
  - Library management
  - Academic personnel records
  - Evaluation and credit
  - Public-private partnerships
- Data governance…



Metrics from Open Data

Searches for author: Christine Borgman, Christine L. Borgman, CL Borgman (excluding other C Borgman authors) on July 28, 2014			
Source	Publications	Citations received	H-index
Google Scholar (Google)	380	7766	39
Web of Science (Thomson-Reuters)	145	1629	20
Scopus (Elsevier)	77	1314	14 (after 1995)

idre UCLA Institute for Digital Research and Education

Search this v

ABOUT PEOPLE PROGRAMS RESOURCES RESEARCH NEWS

Home / Featured / Kent Wada and Christine Borgman Lead Data Governance Task Force

KENT WADA AND CHRISTINE BORGMAN LEAD DATA GOVERNANCE TASK FORCE

February 10, 2015 by Stefanie Pietkiewicz

Kent Wada and Christine Borgman

- How should UCLA collect, organize, and use research analytics about our community?
- Who should have access to these data?
  - Within UCLA?
  - In partnership with public and private entities?
- What are the governance principles?
- What are the governance processes?

https://ccle.ucla.edu/course/view/datagov

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…

For Attribution -- Developing Data Attribution and Citation Practices and Standards: Summary of an International Workshop. Washington, D.C.: The National Academies Press. 2012

Discovery and Interpretation

- Identify the form and content
- Identify related objects
- Interpret
- Evaluate
- Open
- Read
- Compute upon
- Reuse
- Combine
- Describe
- Annotate…

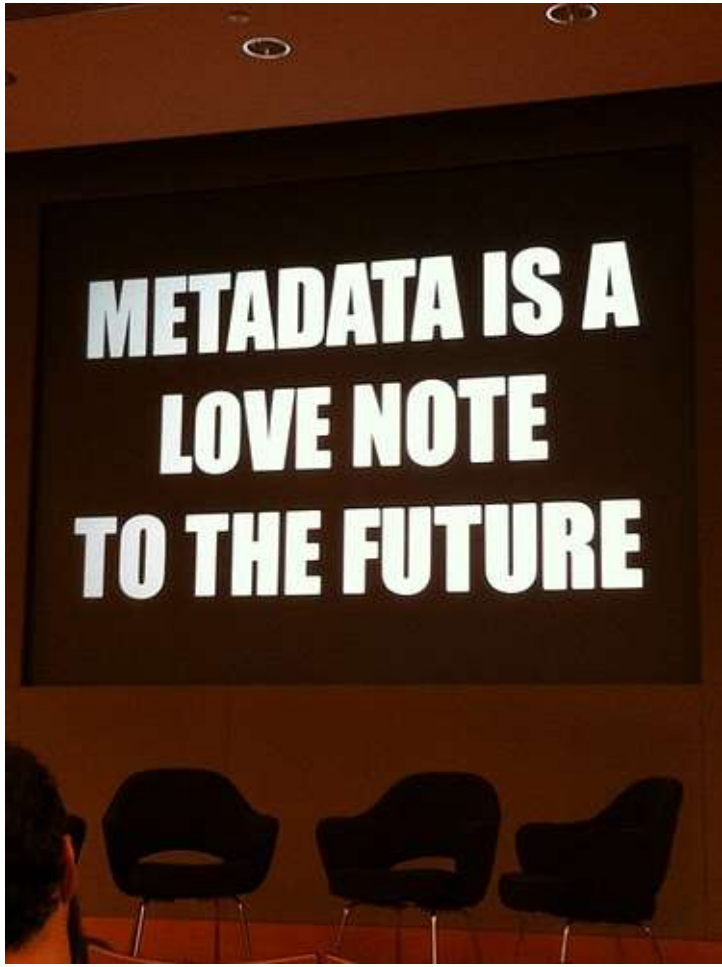
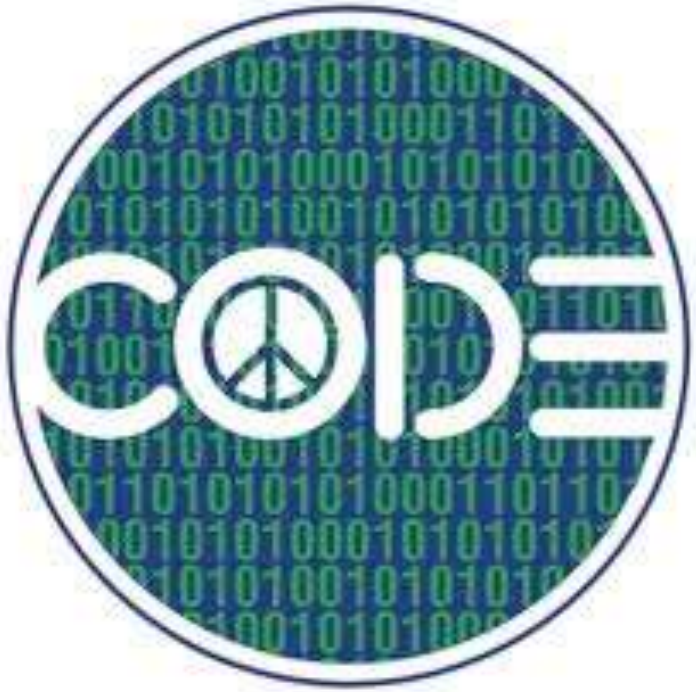


Photo by @kissane; presentation by Jason Scott (@textfiles)

Interpretation and replication

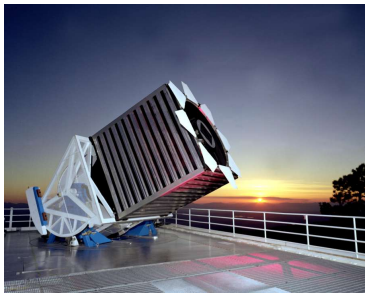
- Datasets
- Methods
  - Collection
  - Cleaning
  - Analysis
  - Codebook
- Publications
- Software and code
- Instrumentation

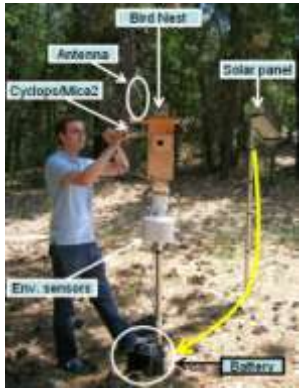


http://peacetour.org/sites/default/files/code4peace-logo2-v3-color-sm.jpg

Some ways to release data

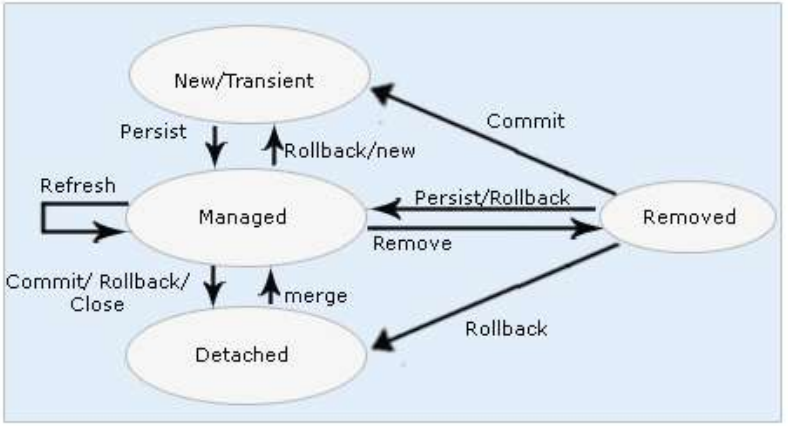
- Centralized data production
  - Top down investments in data
  - Common data archive
- Decentralized data production
  - Bottom up investments in data
  - Pool domain resources later
- Domain-independent aggregators
  - University repositories
  - Figshare, Slideshare, Dataverse…
- Post on lab / personal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





Identity and persistence

- Identity
  - Identifiers
    - DOI, Handles
    - URI, PURL…
  - Naming and namespaces
    - Authors/creators: ORCID, 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

Intellectual property

- What can I do with this object?
- What rights are associated?
  - Reuse
  - Reproduce
  - Attribute
- Who owns the rights?
- How open are data?
  - Open data
  - Open bibliography

http://pzwart.wdka.hro.nl/mdr/research/lliang/mdr/mdr_images/opencontent.jpg/

MODERN DATA SCIENTIST

Data Scientist, the sexiest job of the 21th century, requires a mixture of multidisciplinary skills ranging from an intersection of mathematics, statistics, computer science, communication and business. Finding a data scientist is hard. Finding people who understand who a data scientist is, is equally hard. So here is a little cheat sheet on who the modern data scientist really is.

MATH & STATISTICS

- Machine learning
- Statistical modeling
- Experiment design
- Bayesian inference
- Supervised learning: decision trees, random forests, logistic regression
- Unsupervised learning: clustering, dimensionality reduction
- Optimization: gradient descent and variants

PROGRAMMING & DATABASE

- Computer science fundamentals
- Scripting language e.g. Python
- Statistical computing packages, e.g., R
- Databases: SQL and NoSQL
- Relational algebra
- Parallel databases and parallel query processing
- MapReduce concepts
- Hadoop and Hive/Pig
- Custom reducers
- Experience with xaaS like AWS

DOMAIN KNOWLEDGE & SOFT SKILLS

- Passionate about the business
- Curious about data
- Influence without authority
- Hacker mindset
- Problem solver
- Strategic, proactive, creative, innovative and collaborative

COMMUNICATION & VISUALIZATION

- Able to engage with senior management
- Story telling skills
- Translate data-driven insights into decisions and actions
- Visual art design
- R packages like ggplot or lattice
- Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau

MarketingDistillery.com is a group of practitioners in the area of e-commerce marketing. Our fields of expertise include: marketing strategy and optimization; customer tracking and on-site analytics; predictive analytics and econometrics; data warehousing and big data systems; marketing channel insights in Paid Search, SEO, Social, CRM and brand.

Marketing DISTILLERY

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

Data Curation and Stewardship

- Services and tools
- Data management planning
- Selection and appraisal
- Metadata, provenance
- Migration
- Economics
- Infrastructure

http://www.librarygirl.net/2013/08/putting-your-best-foot-forward-tl.html

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

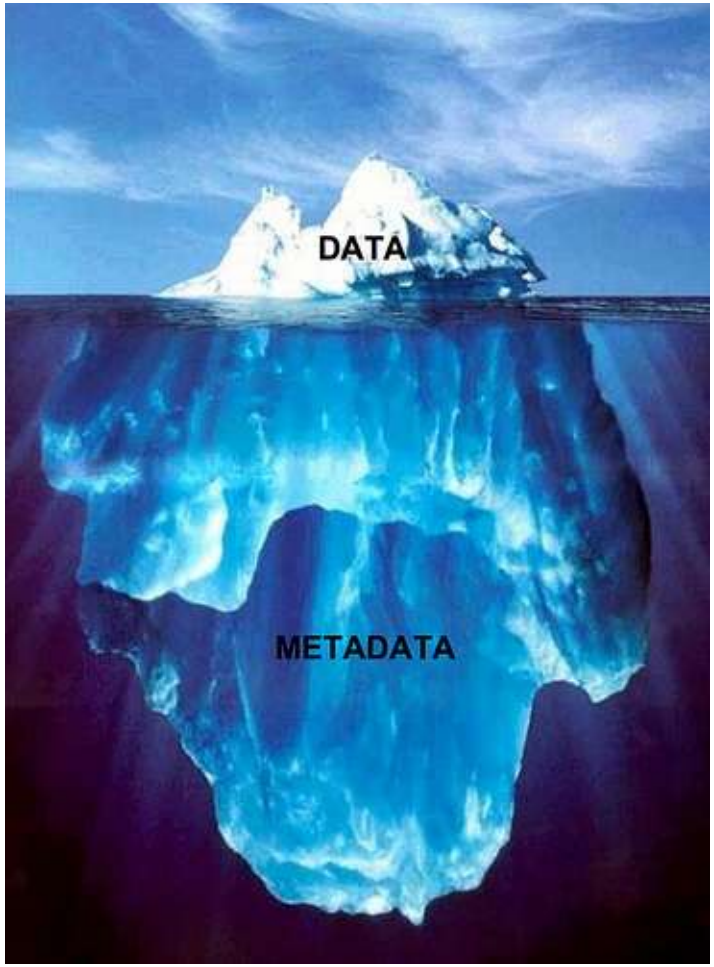
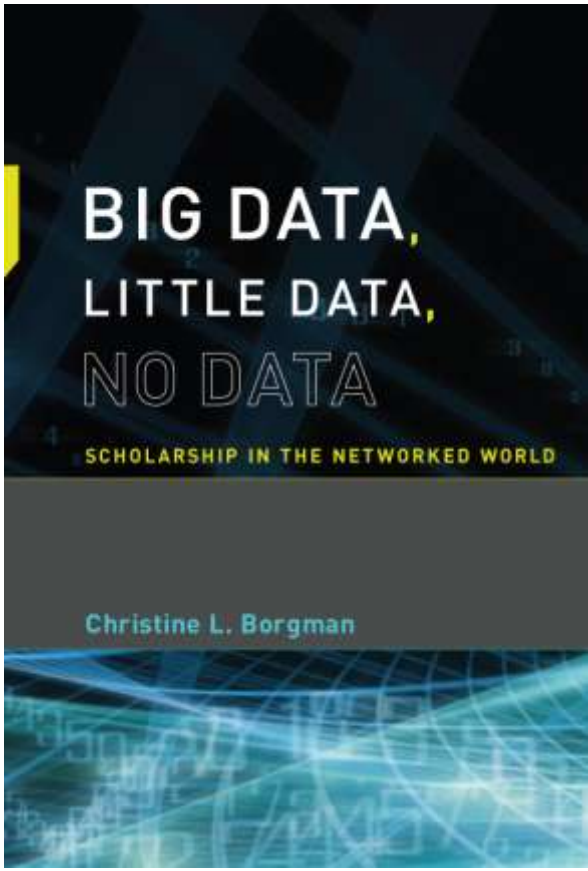


Image from Soumitri Varadarajan blog. Iceberg image © Ralph A. Clevenger. Flickr photo

http://knowledgeinfrastructures.org

Big Data, Little Data, No Data: Scholarship in the Networked World

- Part I: Data and Scholarship
  - Ch 1: Provocations
  - Ch 2: What Are Data?
  - Ch 3: Data Scholarship
  - Ch 4: Data Diversity
- Part II: Case Studies in Data Scholarship
  - Ch 5: Data Scholarship in the Sciences
  - Ch 6: Data Scholarship in the Social Sciences
  - Ch 7: Data Scholarship in the Humanities
- Part III: Data Policy and Practice
  - Ch 8: Releasing, Sharing, and Reusing Data
  - Ch 9: Credit, Attribution, and Discovery
  - Ch 10: What to Keep and Why



Acknowledgements

UCLA Data Practices team

*Peter Darch, Milena Golshan, Irene Pasquetto, Ashley Sands, Sharon Traweek, Camille Mathieu

*Former members: Rebekah Cummings, David Fearon, Ariel Hernandez, Elaine Levia, Jaklyn Nunga, Rachel Mandel, Matthew Mayernik, Alber[illegible], Patricia Smith, [illegible]n Wallis, Laura [illegible] Zhang

*Research funding: National Science Foundation, Alfred P. Sloan Foundation, Microsoft Research, DANS-Netherlands

*University of Oxford: Balliol College, Oliver Smithies