Title
Why data are not publications: Potential potholes for STM publishers

Author
Borgman, Christine L.

Publication Date
2015-04-23

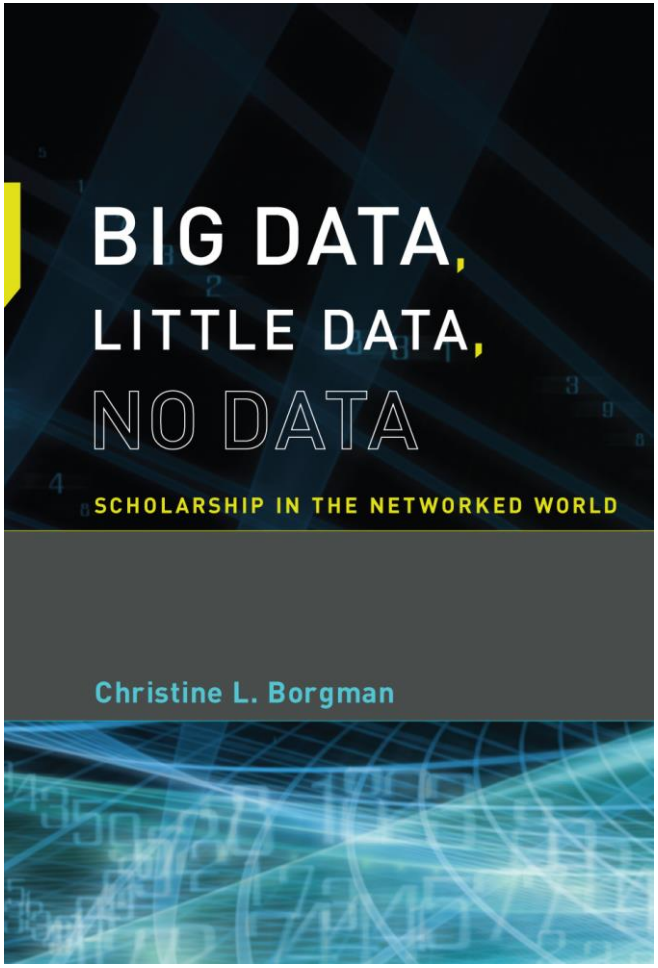
Why data are not publications: Potential potholes for STM publishers

Christine L. Borgman

Professor and Presidential Chair in Information Studies

University of California, Los Angeles

@scitechprof



Keynote presentation

STM Publishers U.S. Conference

Washington, D.C.

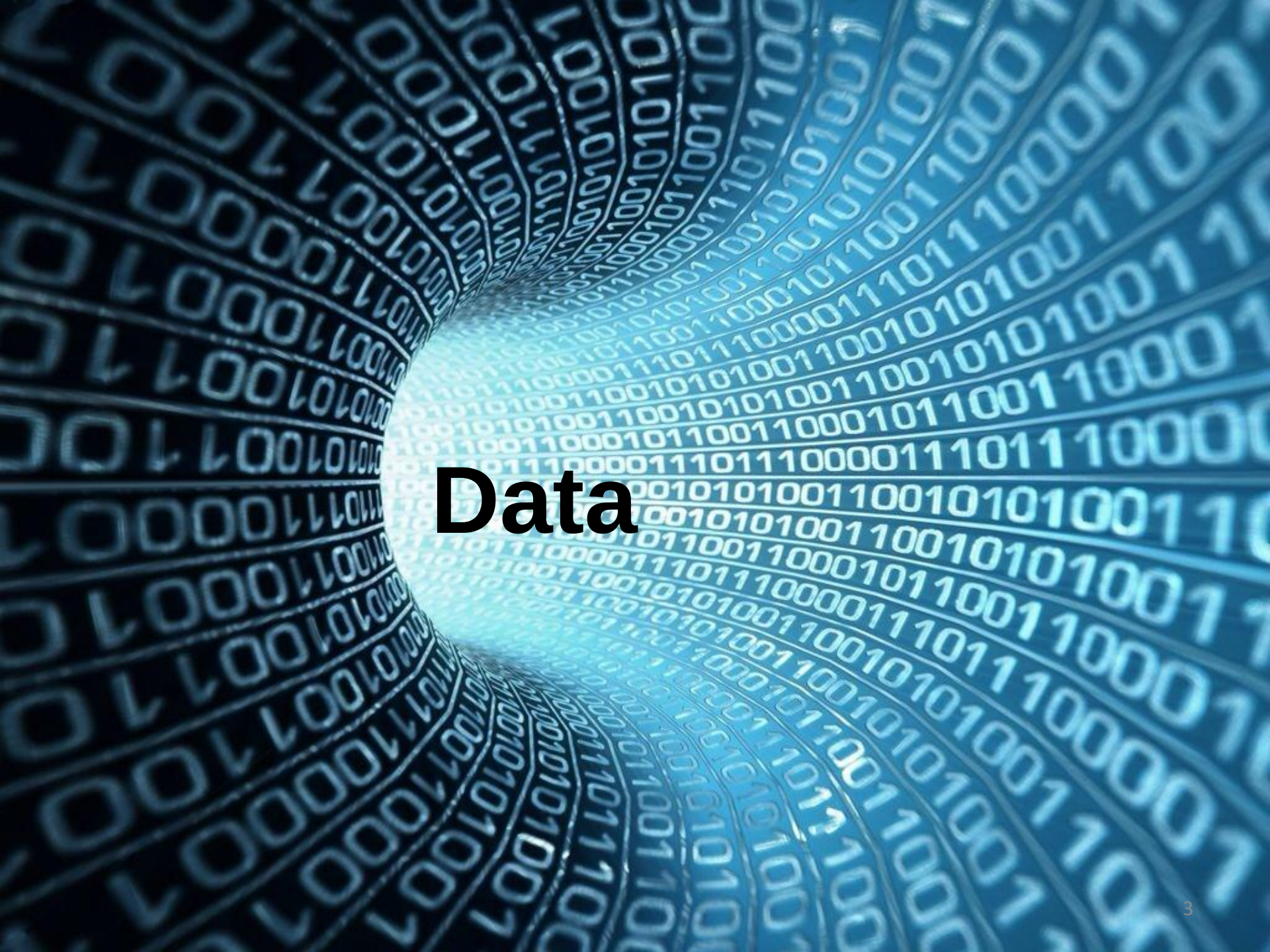
April 22, 2015

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

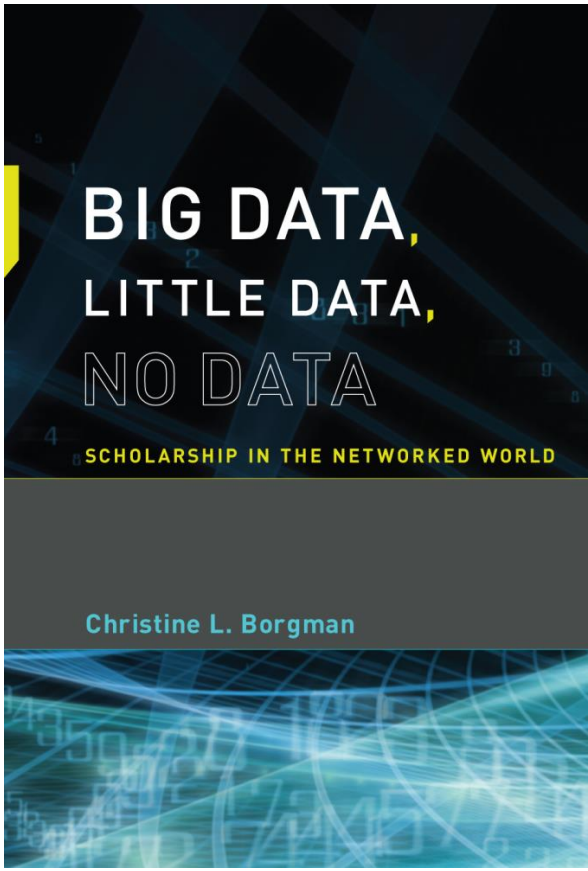
Australian Government
National Health and Medical Research Council

National Science Foundation
WHERE DISCOVERIES BEGIN

Supported by wellcome trust

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

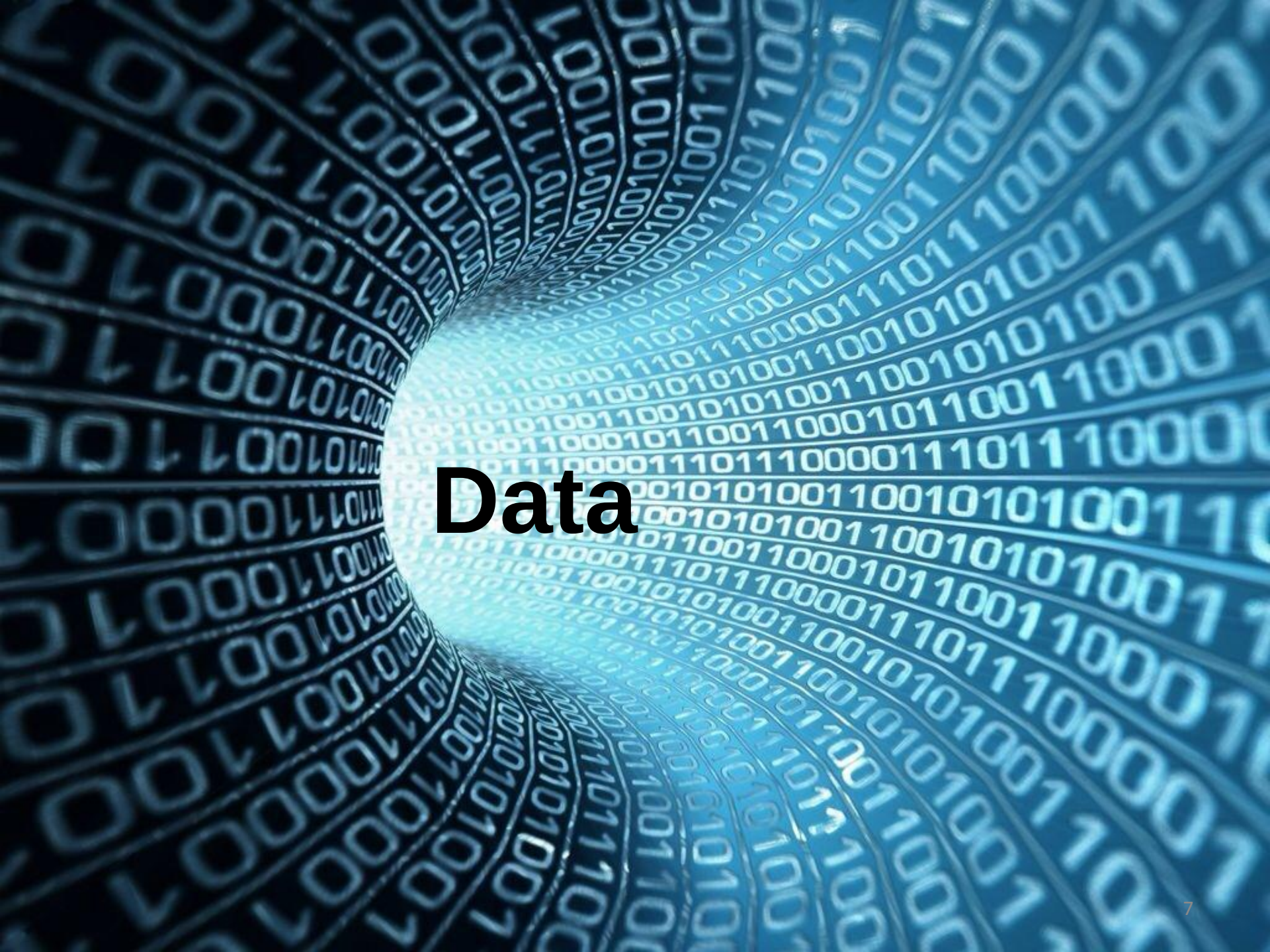
- Part I: Data and Scholarship
  - Ch 1: Provocations
  - Ch 2: What Are Data?
  - Ch 3: Data Scholarship
  - Ch 4: Data Diversity
- Part II: Case Studies in Data Scholarship
  - Ch 5: Data Scholarship in the Sciences
  - Ch 6: Data Scholarship in the Social Sciences
  - Ch 7: Data Scholarship in the 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



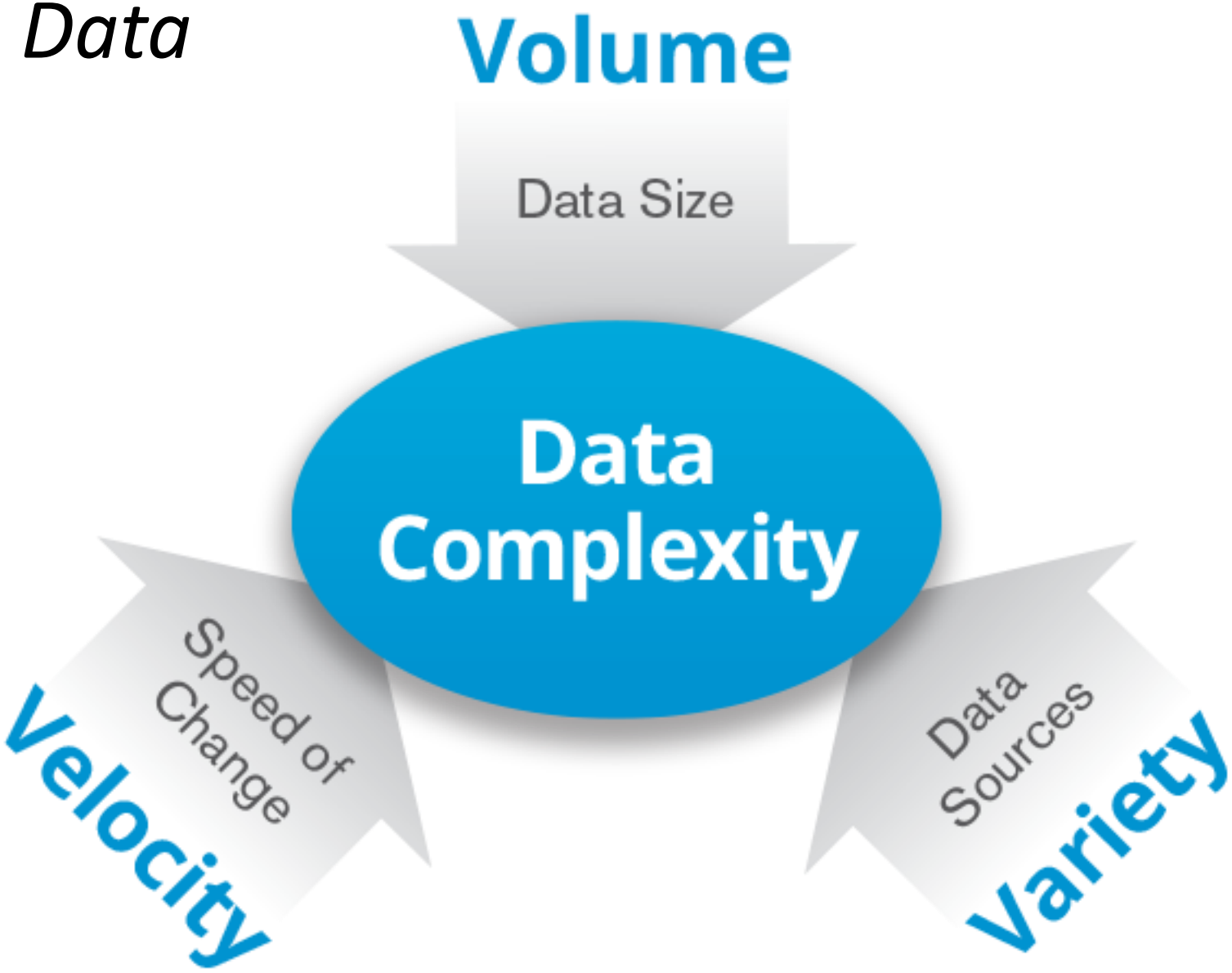
Data, Publications, and Scholarship

- Data
- Publications
- Publications vs. Data
- Releasing data
- Keeping data useful

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

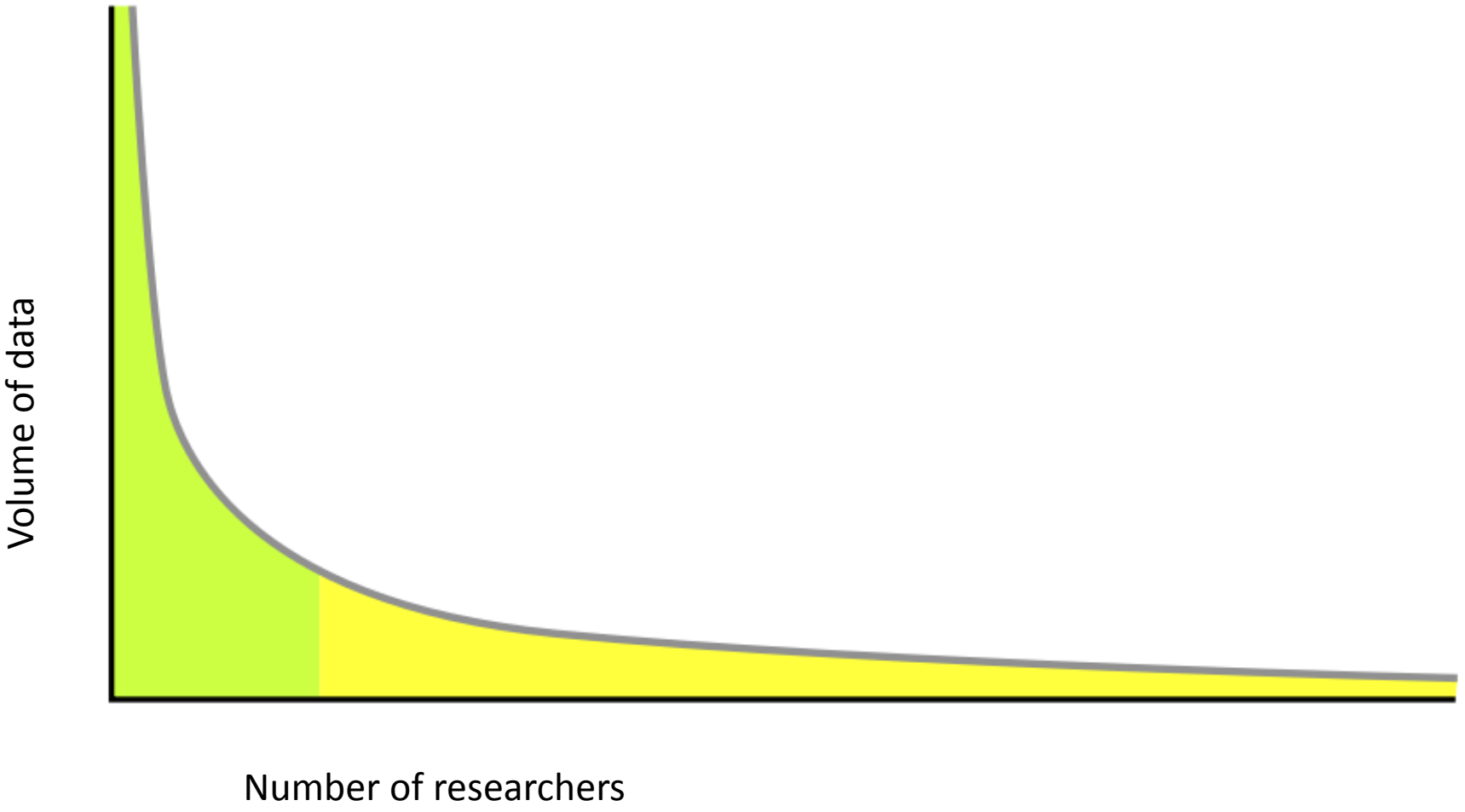

Data

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

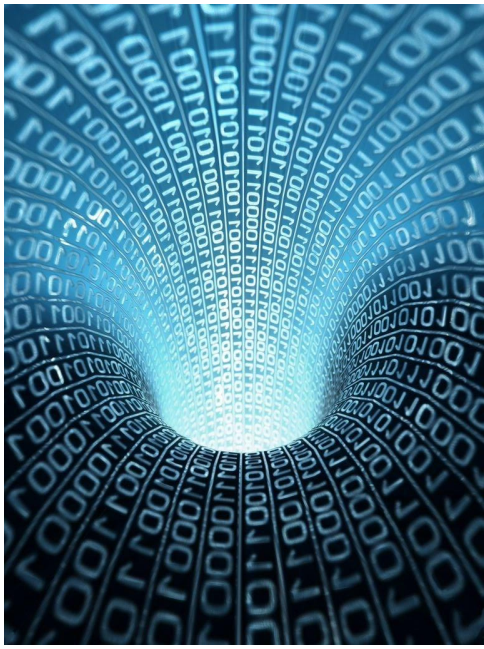
State Library and Archives of Florida, 1922. Flickr commons photo

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
019901AC01C101C501C901E101F601F601F60222022202280236023F024302460267028502850294
01160125011800EA00EA00AE0000003E05BB008A04D70053003FFF8CFFD500150028002200990062
B200402F2B59B002602D2C21B0C051580C6423648BB81555621BB200802F2B59B002602D2C0C6423
B8FF80B30809341CB8FF8040E80809343609352F4A0959045809A7090626062B082C0B280C281328
D4401E091202550C0C047F180118401112025518401D1D02551810251E073C2C04002FED3FEDC42B
1700030209131415002159020 71B25090B1E050526250018 0C0D0D065518021010065518 0CB8FFF8
1F1E4A62A1A4FB432E7900020044FFE40405058E001F002D024BB1020243545840262F4010100255
2F2B2BDD2B2BC01112392F2BCDD0CD003FED5DC45D5D2B323FED12392F5D5D5DCD31301B4019125F
11242517012B50100110302A2912110608070A101B011BB80152B38F2D012DB802F6B2012A2BB801
025D5972103C103C11121739111239391112390111123939123939313043794 07A4B573A4524351C
2D602D702D04802D902D02B02D01002D102DC02DD02D042D6037A67F182B10F65D5D71723CFD3C10
FFEAB40C0C02550CB8FFE2B40D0D02550CB8FFD6B4101002550CB8FFDEB50F0F02550C05BD03E200
2F2F2F3FDDCD3F3F10ED10ED313001B00D4B5458BF0046FFE80045FFE8002FFFE80030FFE8B51A18
1010065529B8FFF2B70F0F0655292935341112392F2B2B2BDD2B2B2BC01112392F2B2B2BCD2B2B2B
3236353427260200D07E6B76CF7FCF7A677DCC53356B429F82617E694703AF9E87AF7BFC80A58BAD
961A9C1E9621982A9E2BA816A61AAB1CAD2BB916BE2BCD2BDA2BEC2BFB2B2A202D732573288F1397
201A01601A701A021A120B003FC45D5DED5D5D2F3FC45DED5D5D5D1217393F012F2B2BCD2F2BCDD4
363702902126775C4656201F5F92CBBD75546C2115170D21211C9E62455761FEDE2D2D9B7B364D33
393FDD5DCD31301B40350127600D5D362027602770 27B02704340B371F3A20481F4820051A08134F
FFE0400A1339082013391B201339012B2B2B2B002B012B2B5959591321152322061514171313 3635
C0B215391AB8FFF040131539360815392830143929301439110816390 9B8FFE0401B163929401139
0000000900010000000000000000100000721FE4500571000FB74FADF1000000100000000000000
00340363004403630004036 3B2242F1FBA034E006D0800400E1F7F027F037F047F050430440112BF
014A001F000D0126400B1F0DC61F0D571F0D371F0D410D019E0141000D00420141000D001E014100
021E0024455258B90024021E4459594BB8020153205C58B9010F00224544B1222245445958B90C00
2B2B2B2B2B2B2B2B2B2B2B2B2B2B2B2B0001737500737300456944007373017374 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

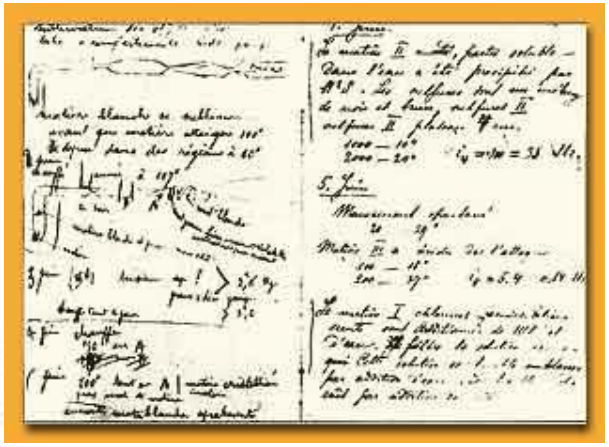
- Openness, flexibility, transparency, legal conformity, protection of intellectual property, formal responsibility, professionalism, interoperability, quality, security, efficiency, accountability, and sustainability.



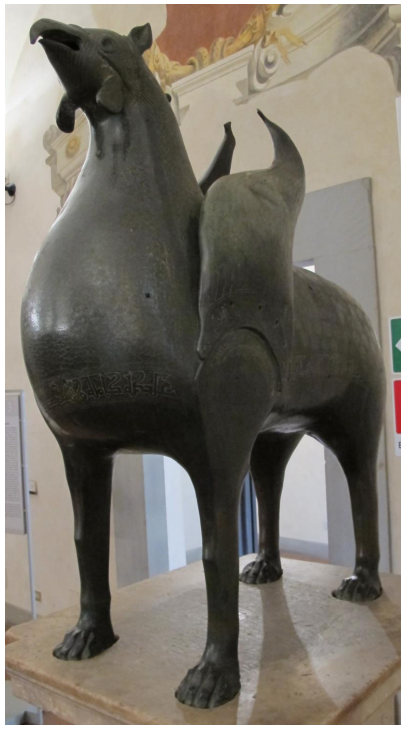
Organization for Economic Cooperation and Development. (2007). *OECD Principles and Guidelines for Access to Research Data from Public Funding*. http://www.oecd.org/dataoecd/9/61/38500813.pdf

What are data?

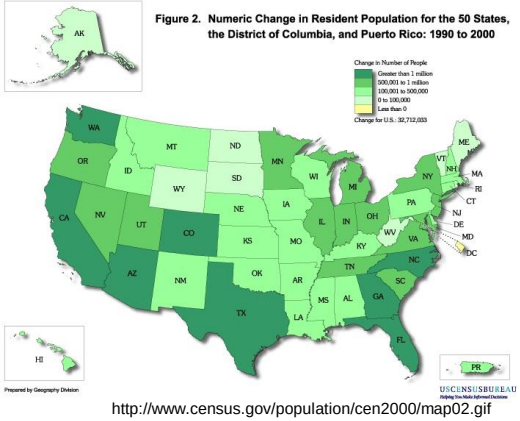
hudsonalpha.org



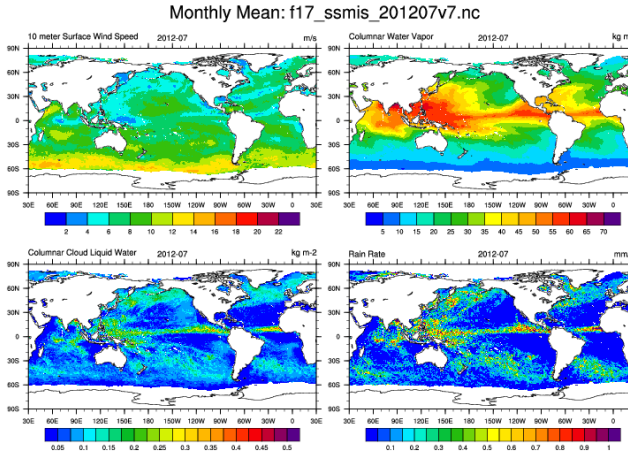
Marie Curie's notebook aip.org



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.

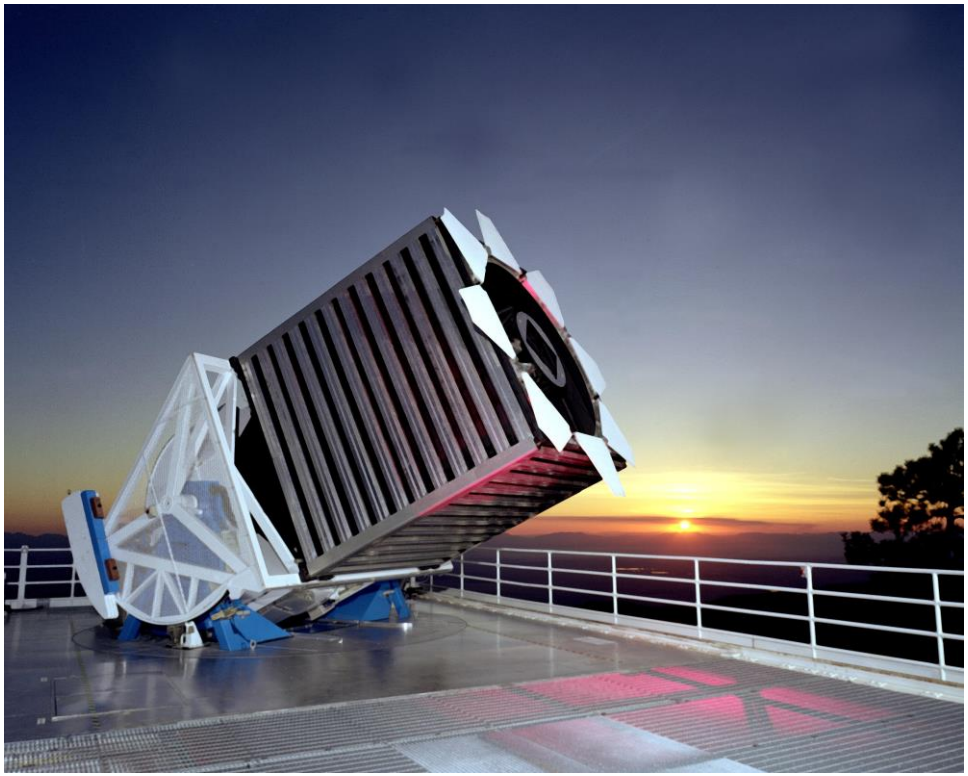
http://onlineqda.hud.ac.uk/Intro_QDA/Examples_of_Qualitative_Data.php

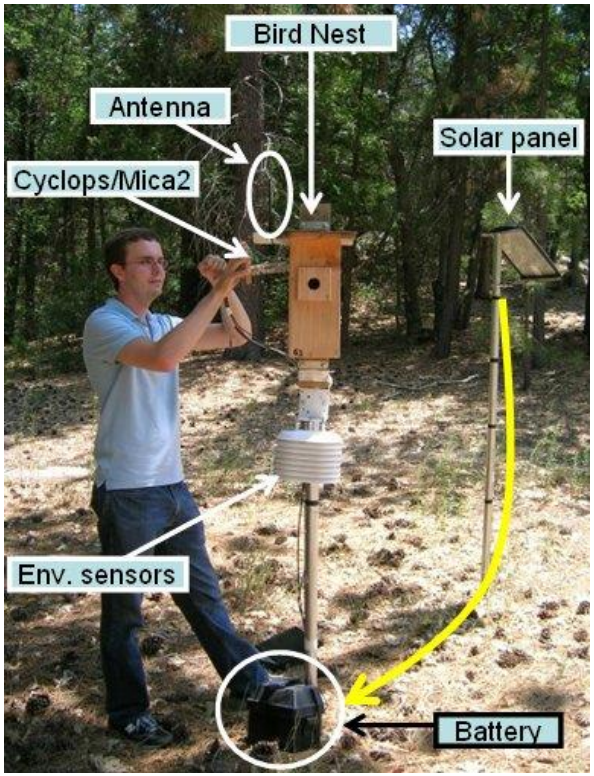
http://www.genome.gov/dmd/img.cfm?node=Photos/Graphics&id=85327

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

Making useful data





Sloan Digital Sky Survey Telescope, Apache Point, New Mexico

Sensor networks

http://astro.uchicago.edu/~frieman/SDSS-telescope-photos/

http://enl.usc.edu/~jpaek/data/cyclops/bird_nest_2008/figures/nestbox2.jpg

WIRED MAGAZINE: 16.07

SCIENCE : DISCOVERIES

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

By Chris Anderson 06.23.08

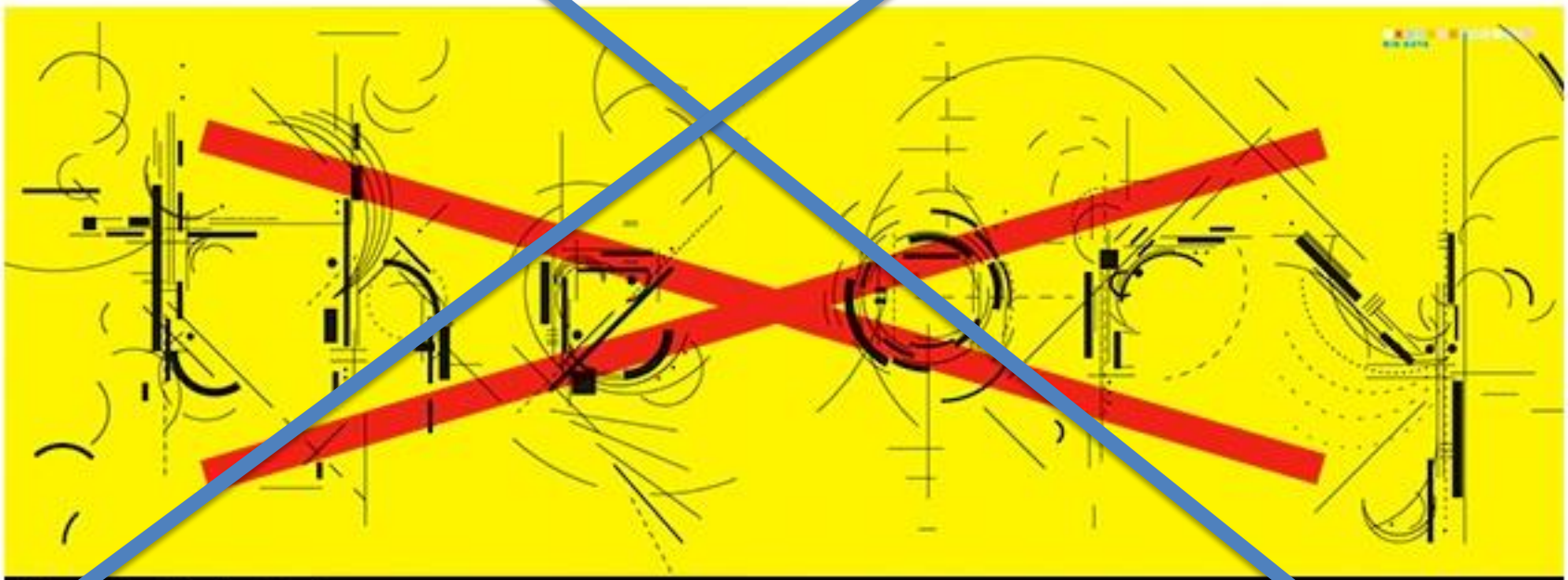
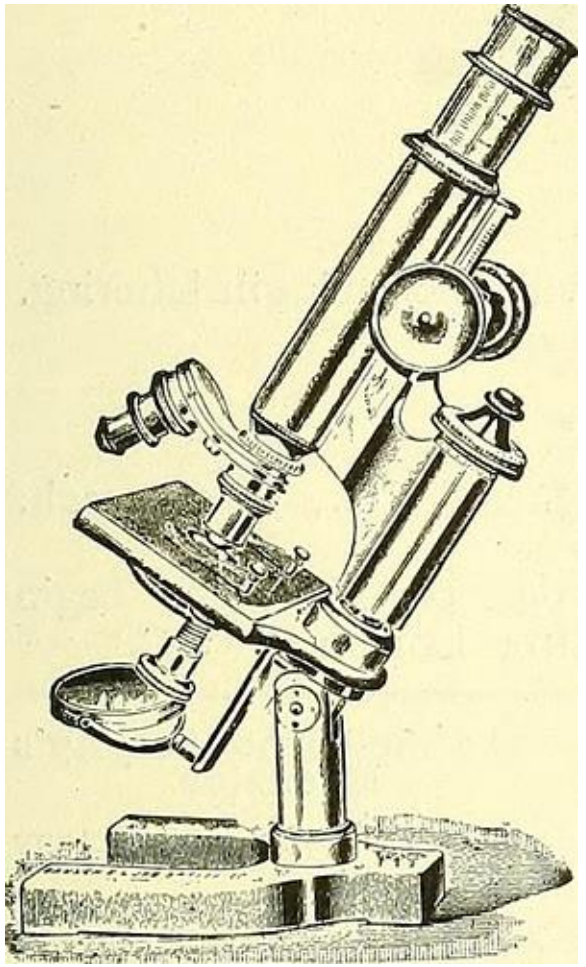


Illustration: Marian Bantjes

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 02215, USA. †Present address: School of Engineering and Applied Sciences, Harvard University, Cambridge, Massachusetts 02138, USA.

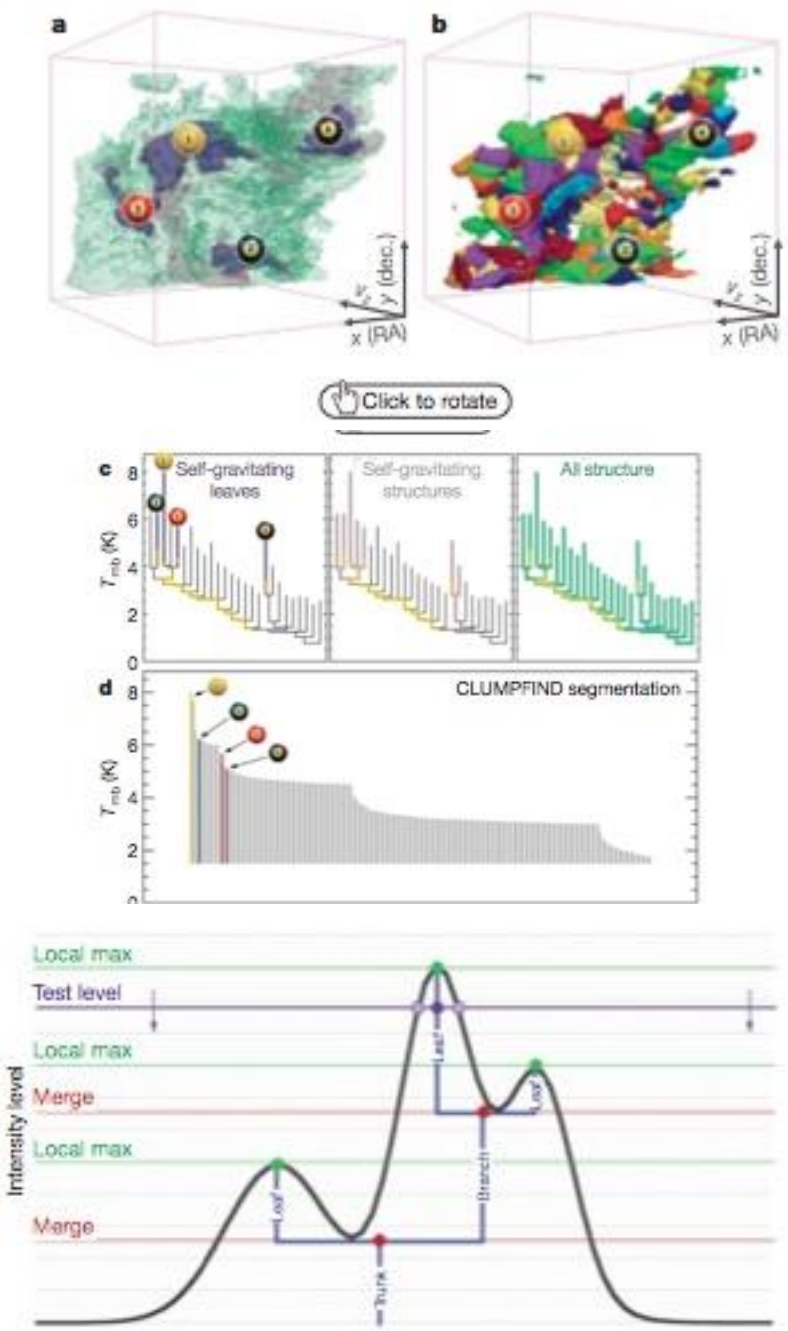
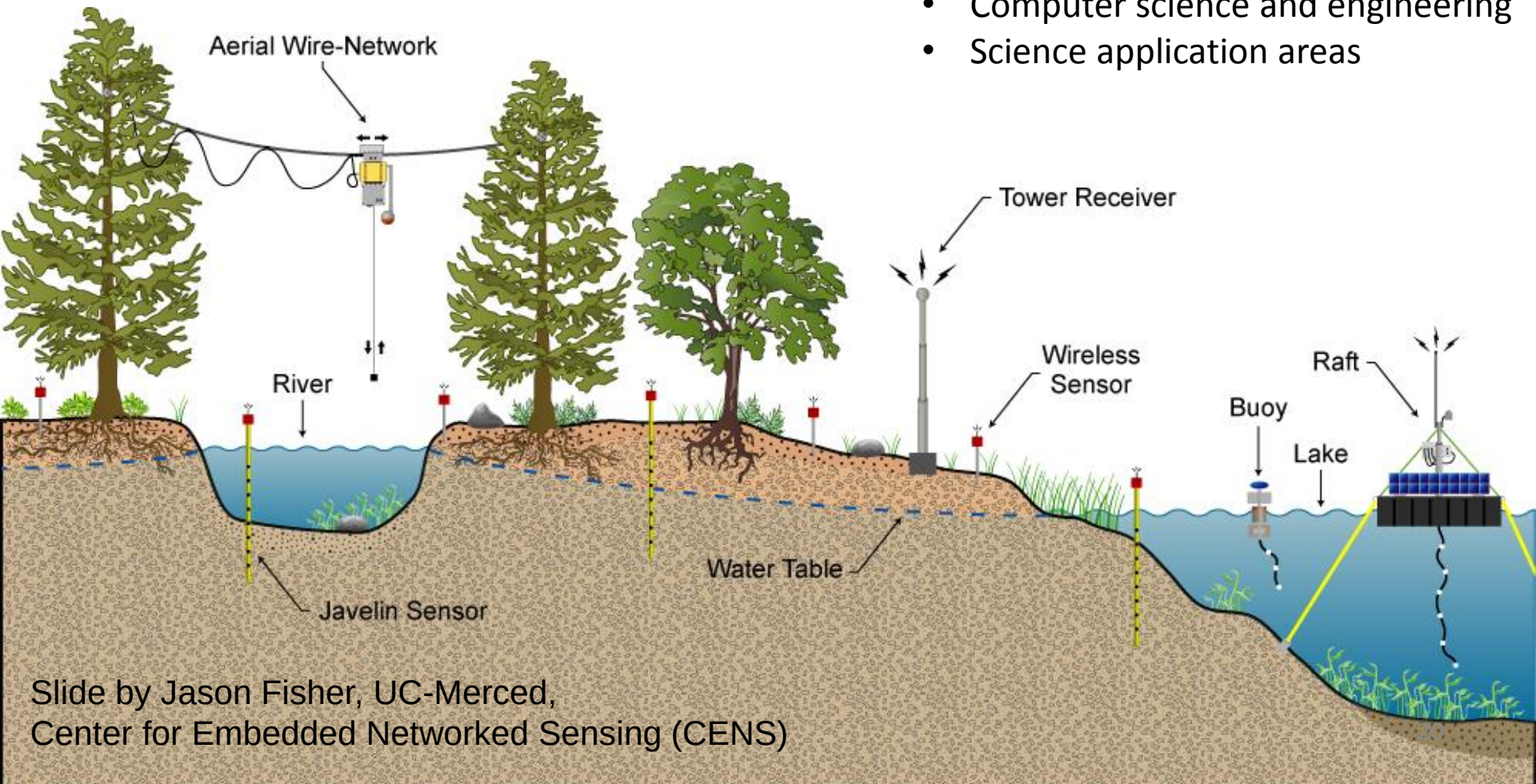


Figure 3 | Schematic illustration of the dendrogram process. Shown is the

Center for Embedded Networked Sensing

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



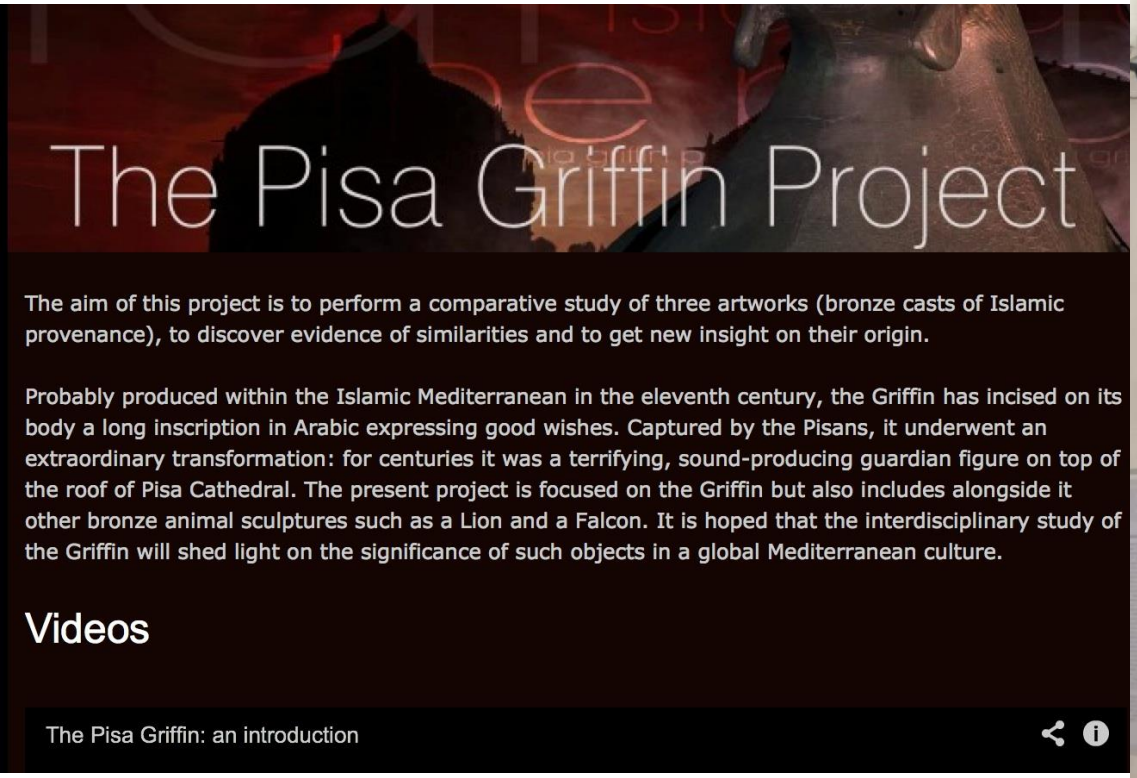
Slide by Jason Fisher, UC-Merced,
Center for Embedded Networked Sensing (CENS)

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.."*



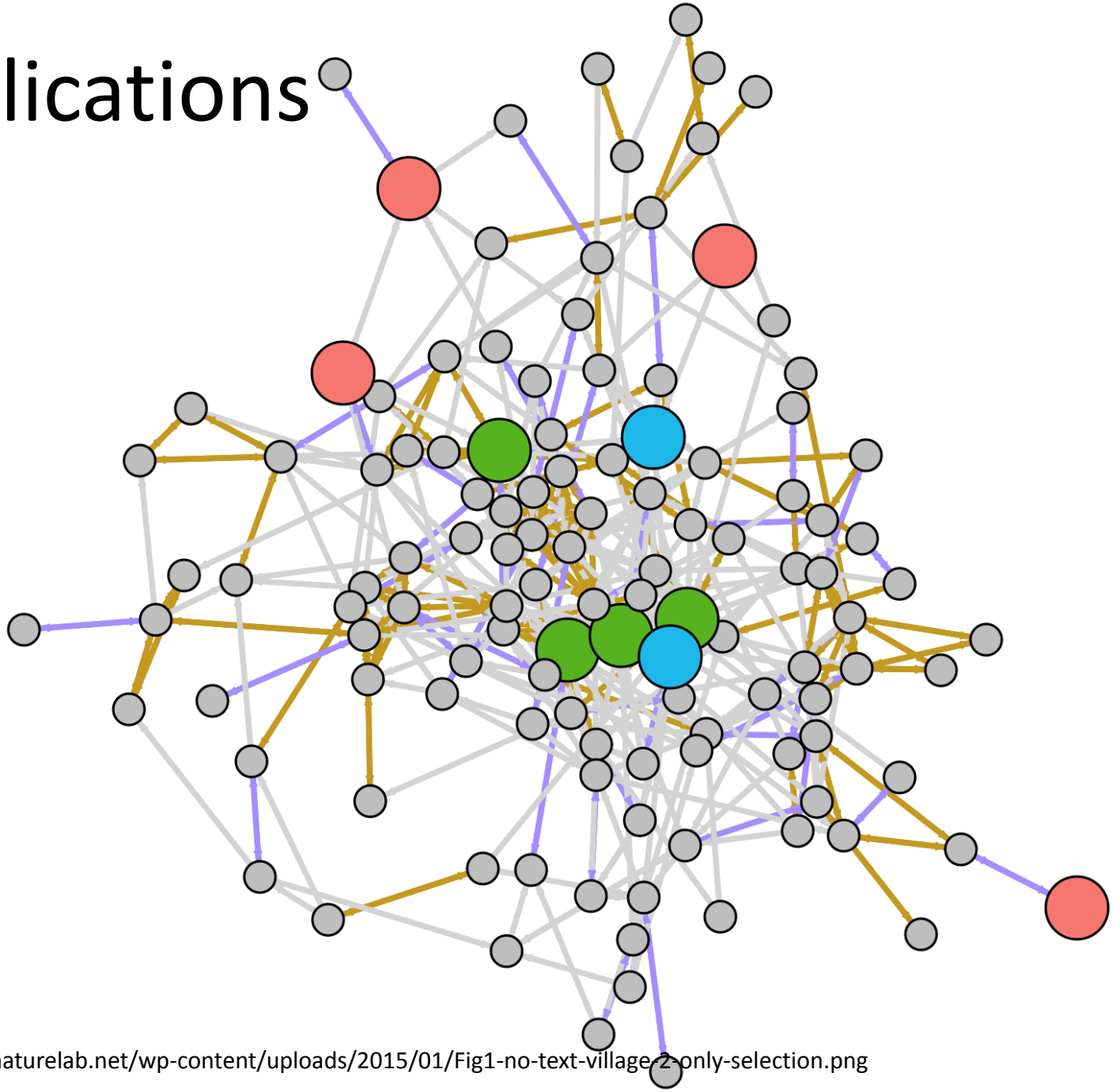
http://vcg.isti.cnr.it/griffin/

Arte islamica, ippogrifo, XI sec 03, own work

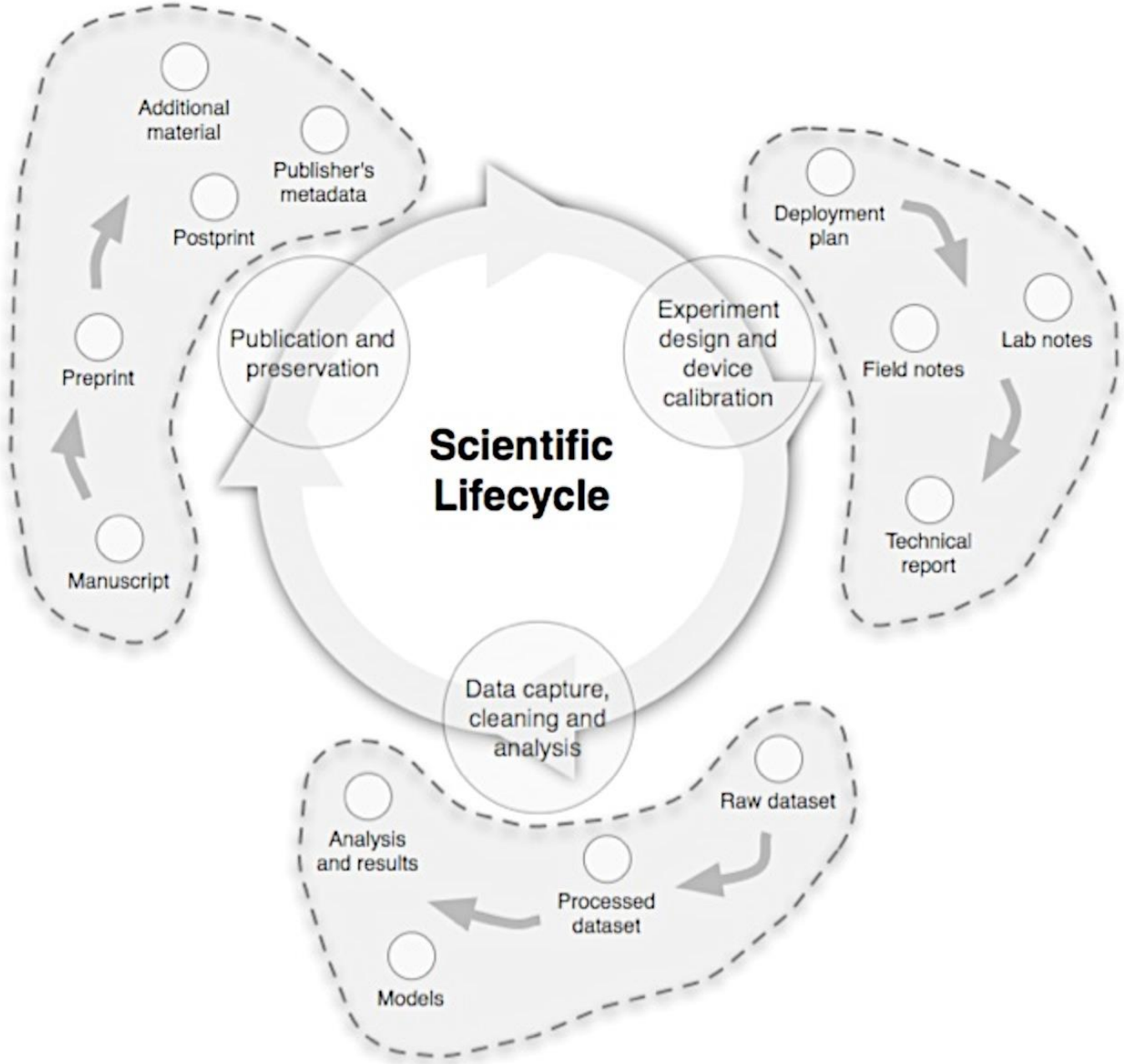
Publications

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.

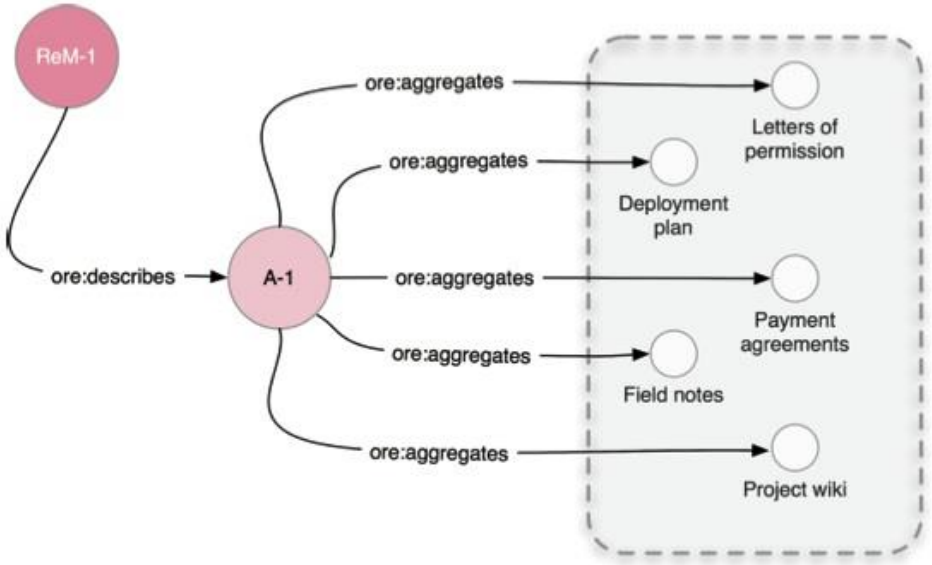


FIG. 4. ORE Aggregation representing the first stage of the scientific life cycle of a sensor network application in seismology (experiment and deployment planning).

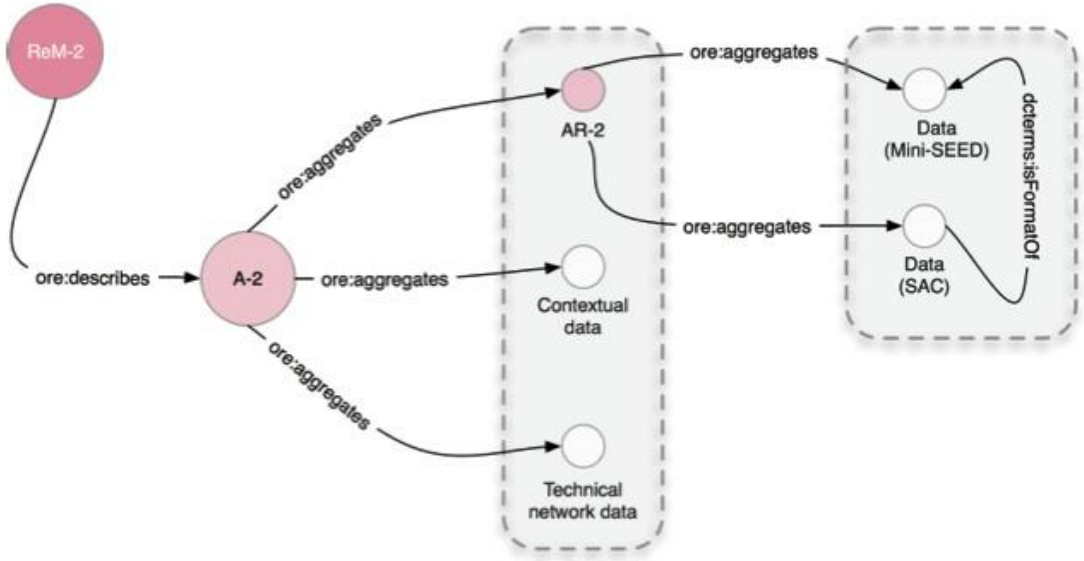
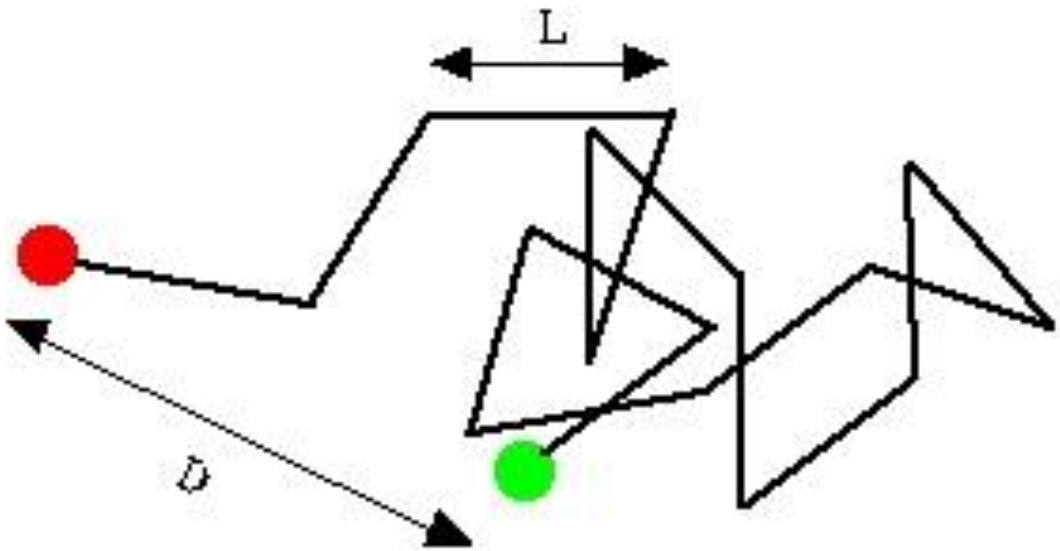


FIG. 5. ORE Aggregation representing the second stage of the scientific life cycle of a sensor network application in seismology (data collection).

Random walk



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

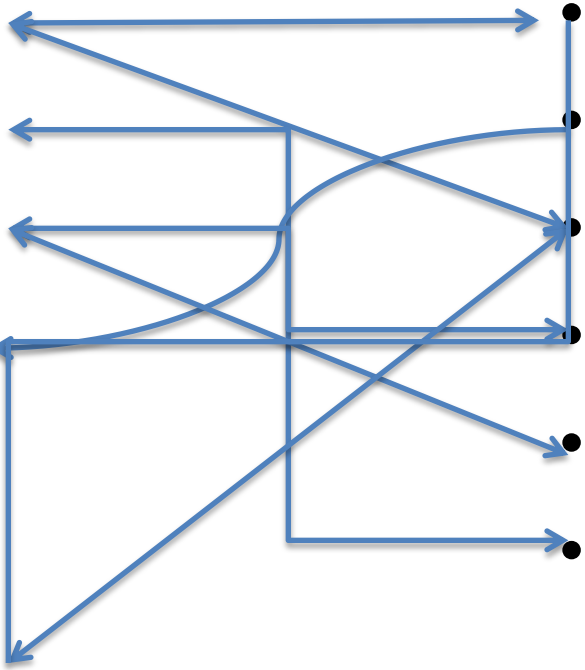
Publications <–> Data

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

- 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

- 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

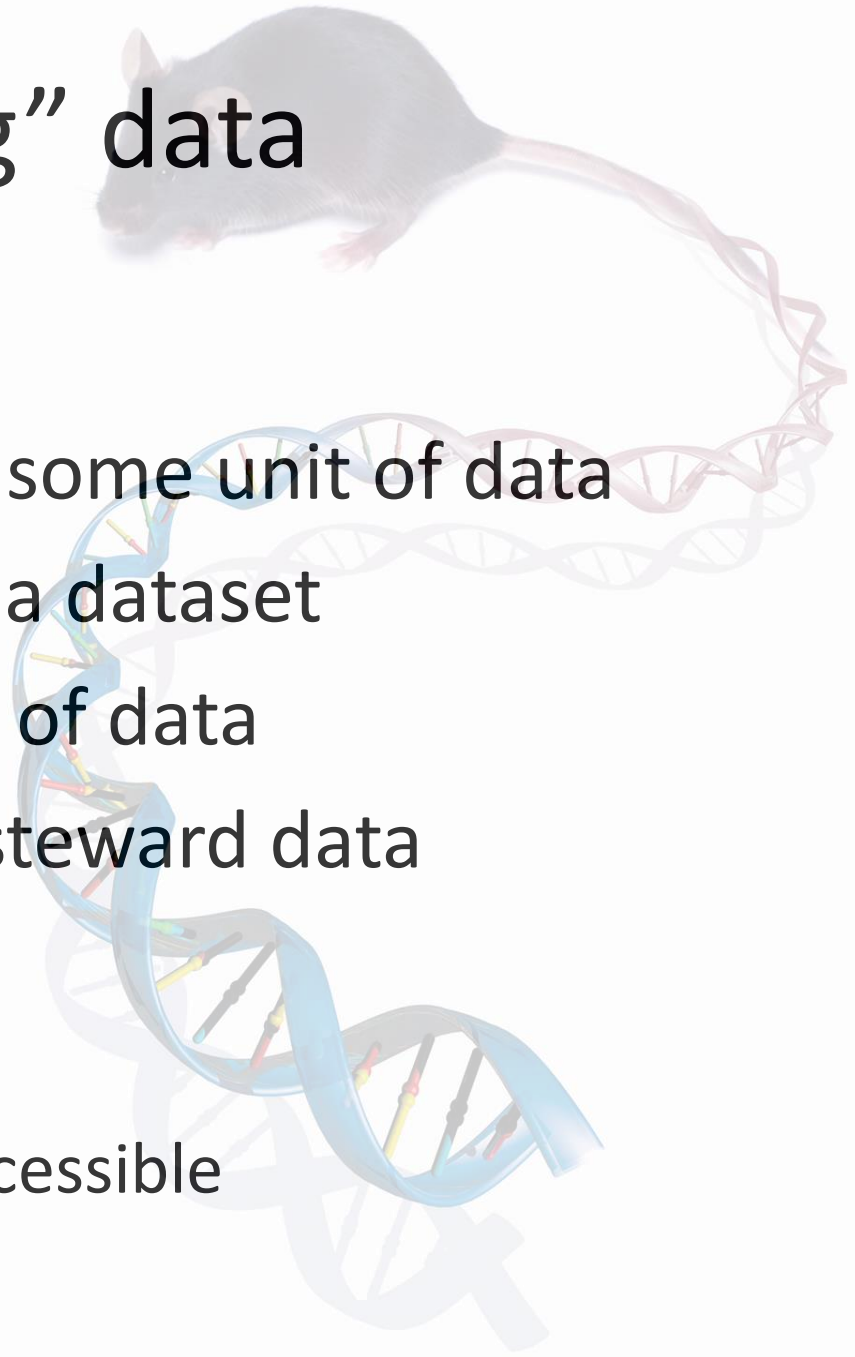
Attribution of data

- 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

“Publishing” data

- To make public
- To release, post, or share some unit of data
- To link a publication with a dataset
- To peer review some unit of data
- To curate, preserve, and steward data
  - Build coherent collections
  - Sustain scientific value
  - Make discoverable and accessible



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

ROSC Community Group Charter

Jun Zhao | Posted on: April 17, 2013

Research Object for Scholarly Communication (ROSC) Community Group Charter

Goal

The primary goal of the Community Group is to provide a platform for scholars, librarians, publishers, archivists and policy makers to exchange requirements and expectations for supporting a new form of scholarly communication, i.e. making the actual research assets available as first-class objects to enable better reuse and reproduce of research results and knowledge. These research assets, including data used and generated in an investigation, methods used for producing the data, as well as people and organisations involved in the study, are what we call **Research Objects**.

It is within the scope of the Community Group to articulate these requirements, gather use cases, survey related work, and analyse and propose various best practices and guidance towards establishing community-wide practice of facilitating the publishing and sharing of the so-called research objects. This Community Group does NOT intend to produce a new specification to achieve the above goals, but may collaborate on forming requirements for other related specifications and vocabularies produced within or outside W3C groups.

Deliverables

Untitled — Edited

PeerJ, Inc. (US) https://peerj.com/preprints/697v4/ research objects manchester

PeerJ | PeerJ Computer Science | ARTICLES | PREPRINTS | More | FOR AUTHORS | SUBMIT ARTICLE | Login | Search

Download as

Citations in Google Scholar
Feedback (2)
Questions
Links

Versions

v1 (14 December, 2014)
v2 (16 December, 2014)
v3 (9 January, 2015)
v4 (11 February, 2015)

554 Visitors 1,199 Views 836 Downloads

View all metrics + mentions on the Web

PREPRINT - *PeerJ PrePrints* is not a peer-reviewed venue.

Achieving human and machine accessibility of cited data in scholarly publications

Joan Starr[1], **Eleni Castro**[2], **Mercè Crosas**[2], **Michel Dumontier**[3], **Robert R. Downs**[4], **Ruth Duerr**[5], **Laurel Haak**[6], **Melissa Haendel**[7], **Ivan Herman**[8], **Simon Hodson**[9], **Joe Hourclé**[10], **John Ernest Kratz**[1], **Jennifer Lin**[11], **Lars Holm Nielsen**[12], **Amy Nurnberger**[13], **Stefan Pröll**[14], **Andreas Rauber**[15], **Simone Sacchi**[13], **Arthur P. Smith**[16], **Michael Taylor**[17], **Tim Clark**[18]

[1] California Digital Library, University of California, Office of the President, Oakland, CA, USA
[2] Institute of Quantitative Social Science, Harvard University, Cambridge, MA, USA
[3] Stanford Center for Biomedical Informatics Research, School of Medicine, Stanford University, Palo Alto, CA, USA
[4] Center for International Earth Science Information Network (CIESIN), Columbia University, Palisades, NY, USA
[5] University of Colorado at Boulder, National Snow and Ice Data Center, Boulder, CO, USA
[6] ORCID, Inc., Bethesda, MD, USA
[7] Oregon Health and Science University, Portland, OR, USA
[8] W3C/CWI, Amsterdam, The Netherlands
[9] CODATA (ICSU Committee on Data for Science and Technology), Paris, FR
[10] Solar Data Analysis Center, NASA Goddard Space Flight Center, Greenbelt, MD, USA
[11] Public Library of Science, San Francisco, CA, USA
[12] European Organization for Nuclear Research (CERN), Geneva, Switzerland
[13] Columbia University Libraries/Information Services, New York, NY, USA
[14] SBA Research, Vienna, AT
[15] Institute of Software Technology and Interactive Systems, Vienna University of Technology / TU Wien,

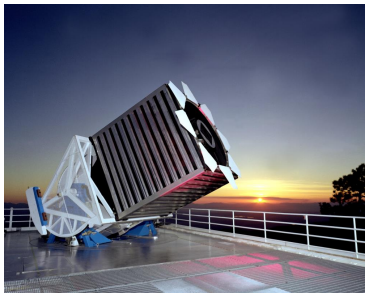
Research Data Sharing without barriers

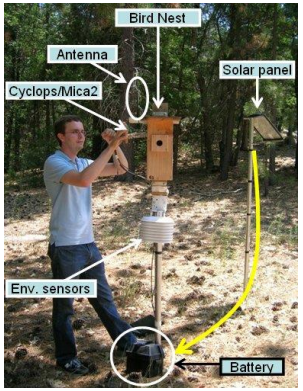
Precondition:

Researchers share data

Ways to release data

- Centralized data production
  - Top down investments in data
  - Pooled data resources for the community
- Decentralized data production
  - Bottom up investments in data
  - Local data resources pooled later
- Post on websites
- Share privately upon request…





Sloan Digital Sky Survey

SDSS

Data Surveys Instruments Collaboration Results Education Contact

Search

The Sloan Digital Sky Survey has created the most detailed three-dimensional maps of the Universe ever made, with deep multi-color images of one third of the sky, and spectra for more than three million astronomical objects. Learn and explore all phases and surveys—past, present, and future—of the SDSS.

EXPLORE OUR DATA

Go to Data Access

Current data: Data Release 12

News

SDSS Press Releases

The Sloan Digital Sky Survey Opens a New Public View of the Sky
January 6, 2015

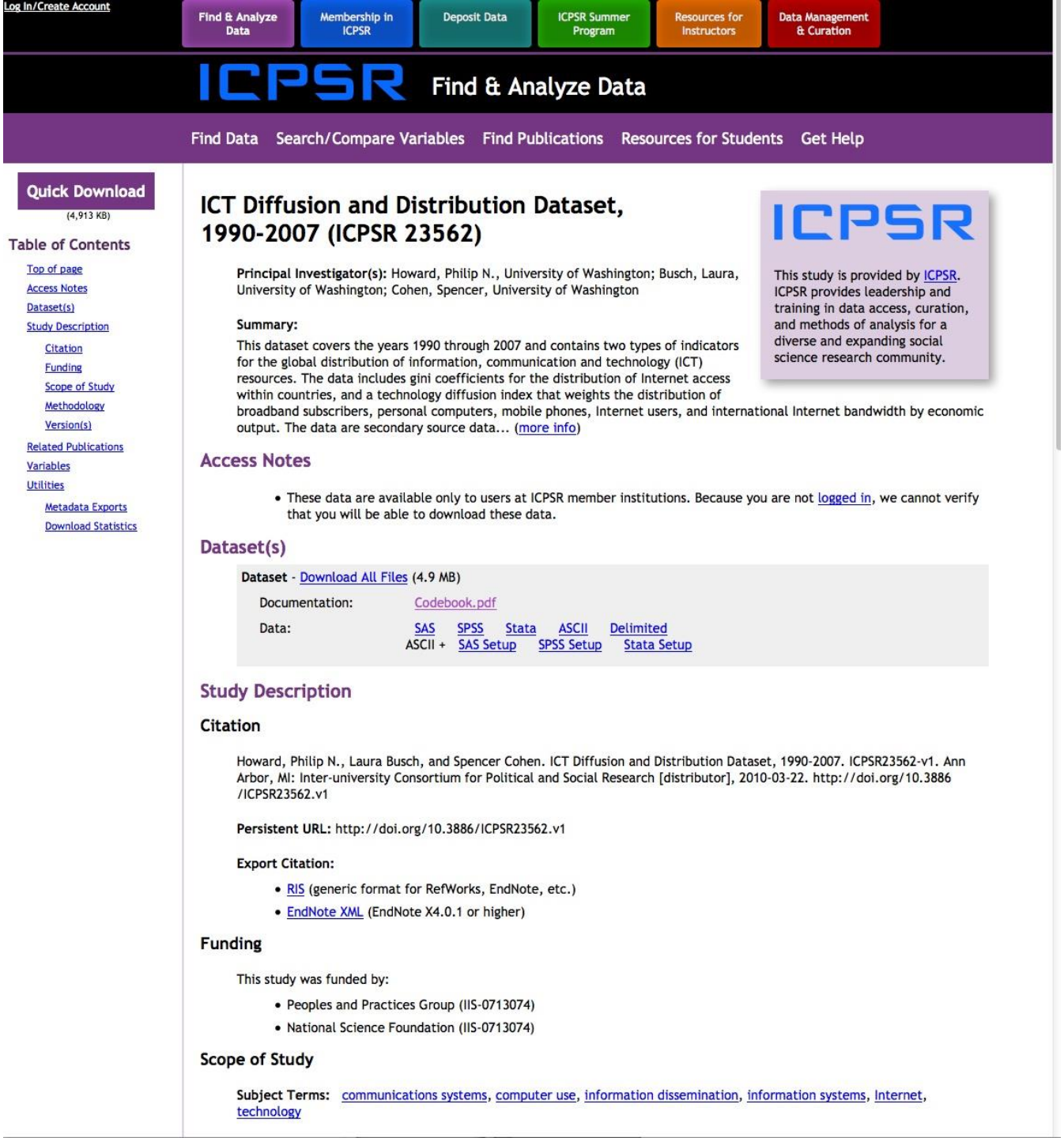
SDSS Science Blog

How SDSS Uses Light to Measure the Distances to Galaxies
February 27, 2015

SDSS Data Lead to Discovery of 12 Billion Solar Mass Black Hole in Young Universe
February 25, 2015

Spotlight on APOGEE: Duy Nguyen and Binary Stars
February 14, 2015

Social Science Surveys



Log In/Create Account

Find & Analyze Data | Membership in ICPSR | Deposit Data | ICPSR Summer Program | Resources for Instructors | Data Management & Curation

ICPSR Find & Analyze Data

Find Data | Search/Compare Variables | Find Publications | Resources for Students | Get Help

Quick Download
(4,913 KB)

Table of Contents

- Top of page
- Access Notes
- Dataset(s)
- Study Description
  - Citation
  - Funding
  - Scope of Study
  - Methodology
  - Version(s)
- Related Publications
- Variables
- Utilities
  - Metadata Exports
  - Download Statistics

ICT Diffusion and Distribution Dataset, 1990-2007 (ICPSR 23562)

ICPSR

This study is provided by ICPSR. ICPSR provides leadership and training in data access, curation, and methods of analysis for a diverse and expanding social science research community.

Principal Investigator(s): Howard, Philip N., University of Washington; Busch, Laura, University of Washington; Cohen, Spencer, University of Washington

Summary:

This dataset covers the years 1990 through 2007 and contains two types of indicators for the global distribution of information, communication and technology (ICT) resources. The data includes gini coefficients for the distribution of Internet access within countries, and a technology diffusion index that weights the distribution of broadband subscribers, personal computers, mobile phones, Internet users, and international Internet bandwidth by economic output. The data are secondary source data... (more info)

Access Notes

- These data are available only to users at ICPSR member institutions. Because you are not logged in, we cannot verify that you will be able to download these data.

Dataset(s)

Dataset - Download All Files (4.9 MB)

Documentation: Codebook.pdf

Data: SAS SPSS Stata ASCII Delimited
ASCII + SAS Setup SPSS Setup Stata Setup

Study Description

Citation

Howard, Philip N., Laura Busch, and Spencer Cohen. ICT Diffusion and Distribution Dataset, 1990-2007. ICPSR23562-v1. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2010-03-22. http://doi.org/10.3886/ICPSR23562.v1

Persistent URL: http://doi.org/10.3886/ICPSR23562.v1

Export Citation:

- RIS (generic format for RefWorks, EndNote, etc.)
- EndNote XML (EndNote X4.0.1 or higher)

Funding

This study was funded by:

- Peoples and Practices Group (IIS-0713074)
- National Science Foundation (IIS-0713074)

Scope of Study

Subject Terms: communications systems, computer use, information dissemination, information systems, Internet, technology

OPEN ACCESS PEER-REVIEWED

RESEARCH ARTICLE

If We Share Data, Will Anyone Use Them? Data Sharing and Reuse in the Long Tail of Science and Technology

Jillian C. Wallis, Elizabeth Rolando, Christine L. Borgman

Published: July 23, 2013 • DOI: 10.1371/journal.pone.0067332

Abstract

Research on practices to share and reuse data will inform the design of infrastructure to support data collection, management, and discovery in the long tail of science and technology.

Subject Areas

Data management

Lack of incentives to release data

- Labor to document data
- Benefits to unknown others
- Competition
- Control
- Confidentiality…

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

Lack of incentives to reuse data

- Identify useful data
  - Documentation
  - Interpretation
  - Software
- Cleaning
- Trust
- Credit
- Licensing…

http://fyi.uiowa.edu/wp-content/uploads/2011/10/utopia_in_four_movements_filmstill5_utopiasign.jpg

Keeping Data Useful

Flickr Commons Photo: Women working in the Pinion Department at Bulova Watch, Southern Methodist University Libraries; Creator: Richie, Robert Yarnall (1908-1984), 1937

Discovery and Interpretation

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

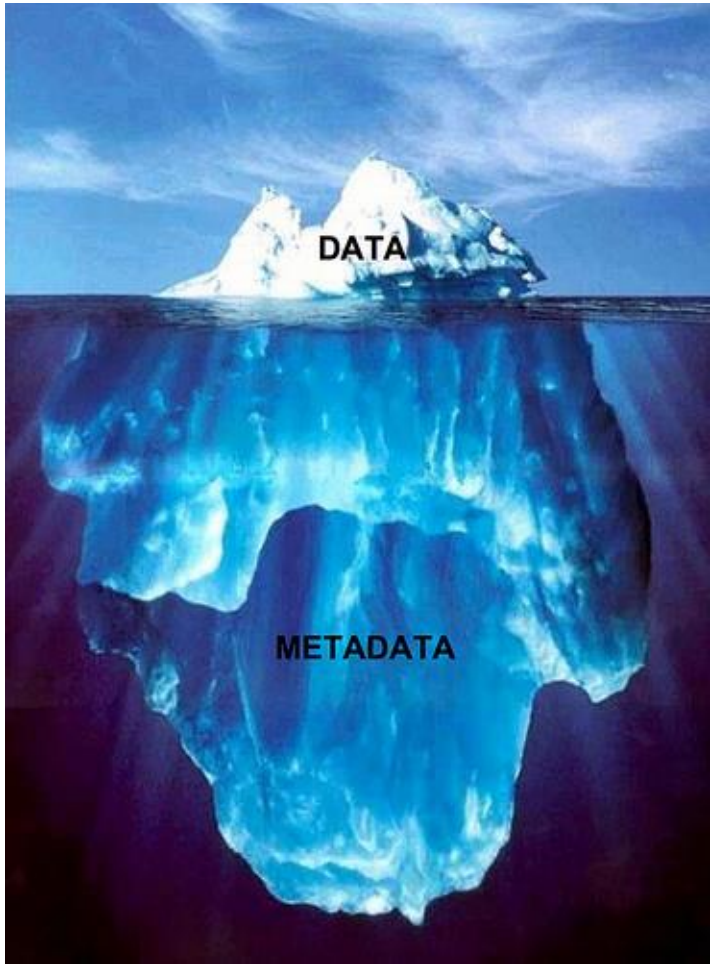
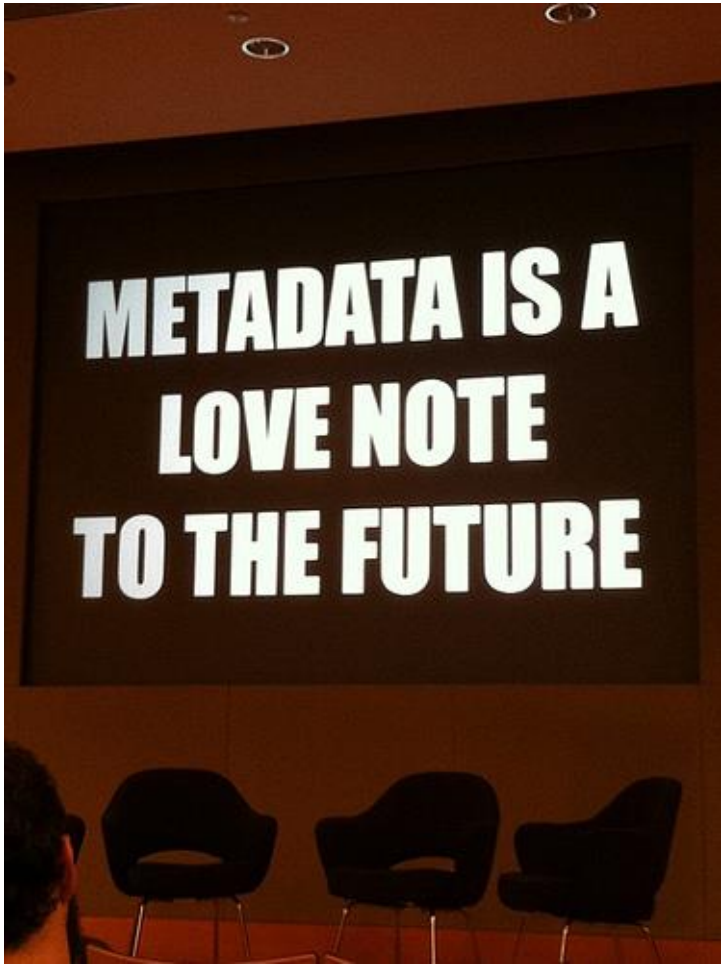


Image from Soumitri Varadarajan blog. Iceberg image © Ralph A. Clevenger. Flickr photo

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

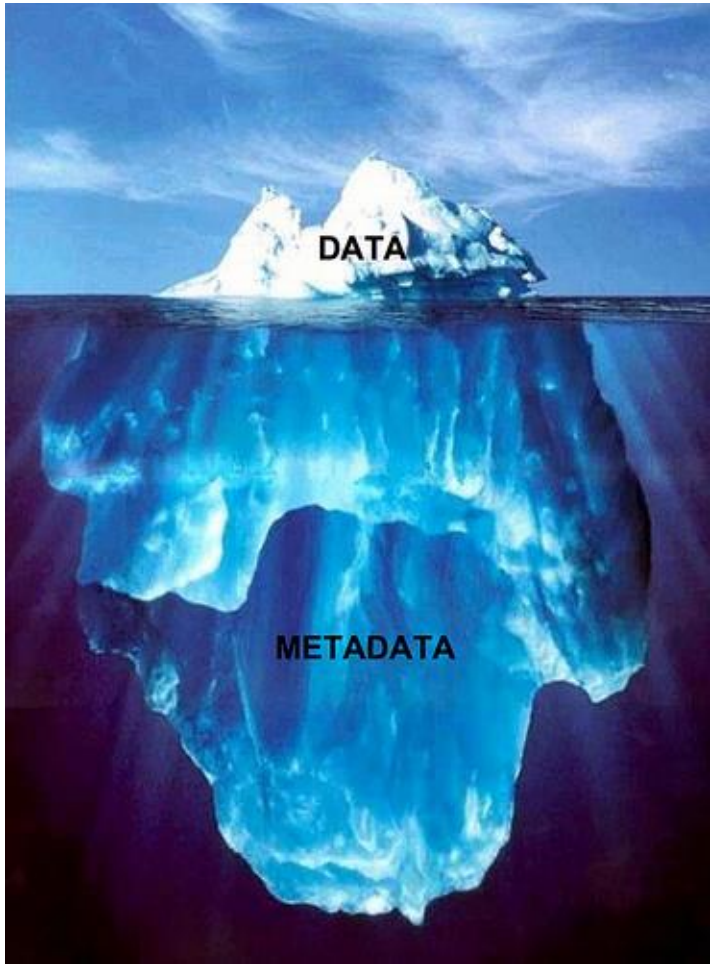
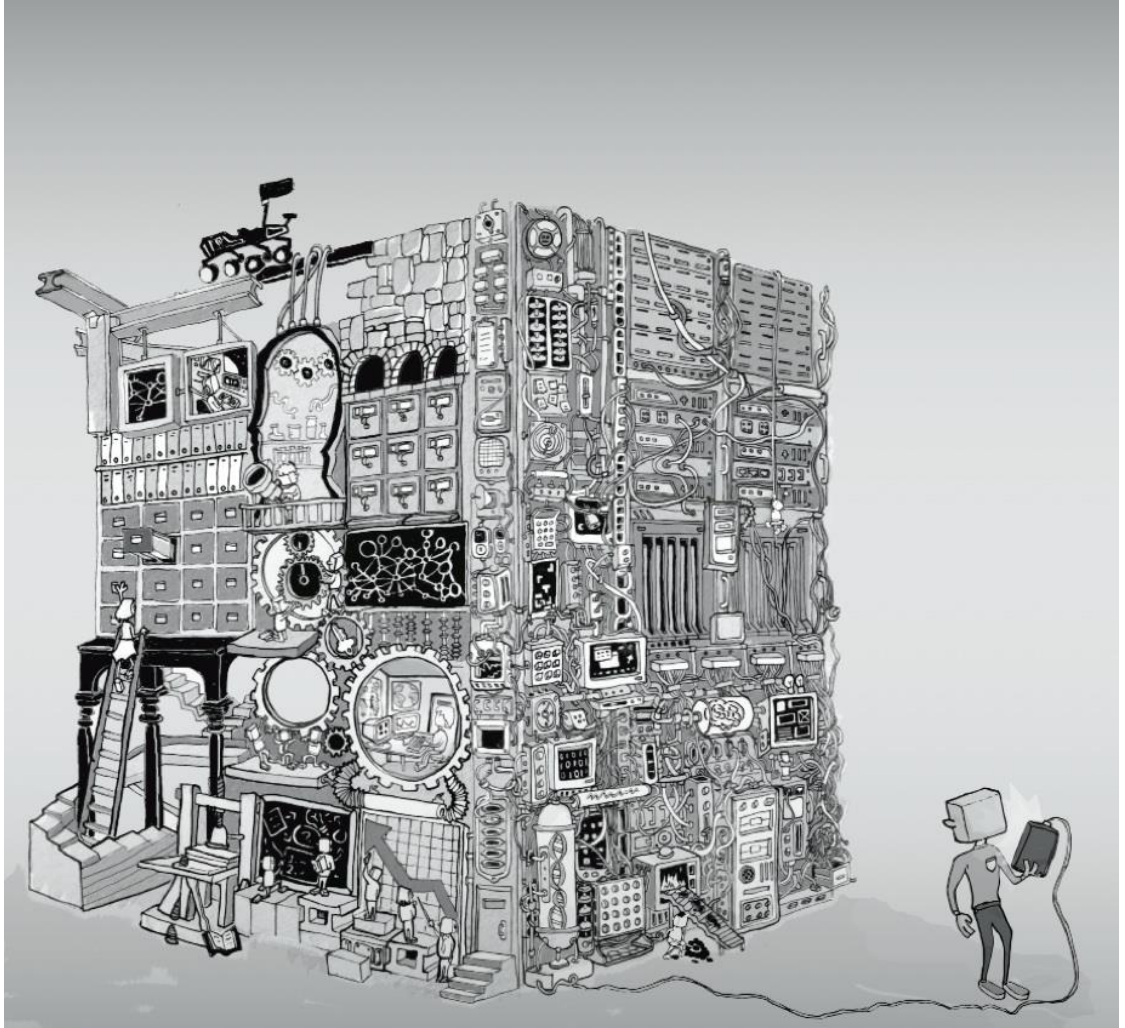


Image from Soumitri Varadarajan blog. Iceberg image © Ralph A. Clevenger. Flickr photo

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



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

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.

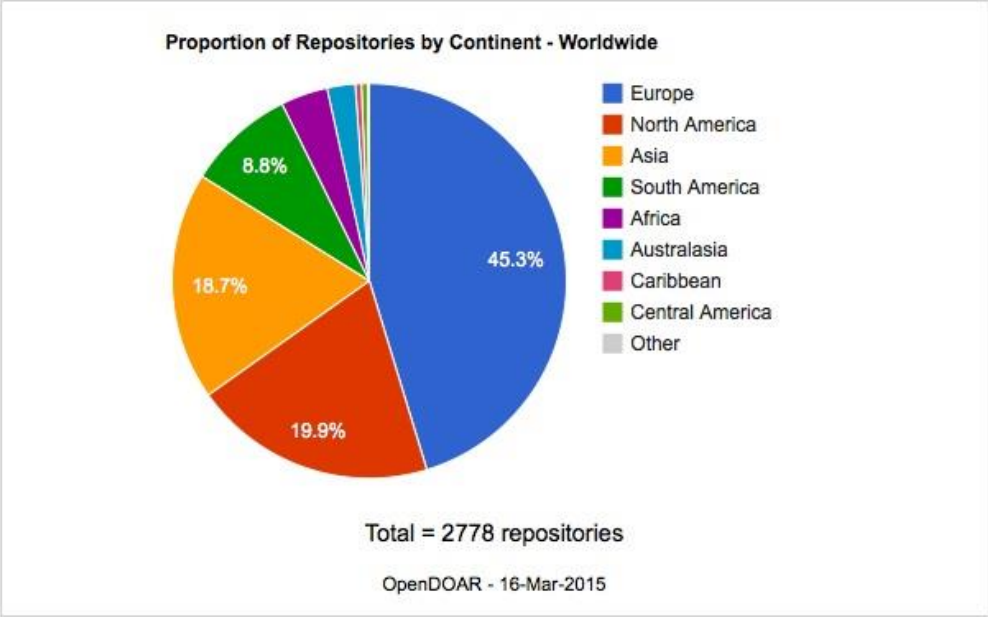
Data Repositories

OpenDOAR

Directory of Open Access Repositories

Home | Find | Suggest | Tools | FAQ | About | Contact Us

Proportion of Repositories by Continent - Worldwide



This chart is based on the number of repositories in each Continent. However, some organisations have two or more repositories - over 20 in some cases - and this arguably skews the results.

For a different viewpoint, please see the equivalent chart for Repository Organisations, in which each organisation only counts once, regardless of how many repositories it hosts.

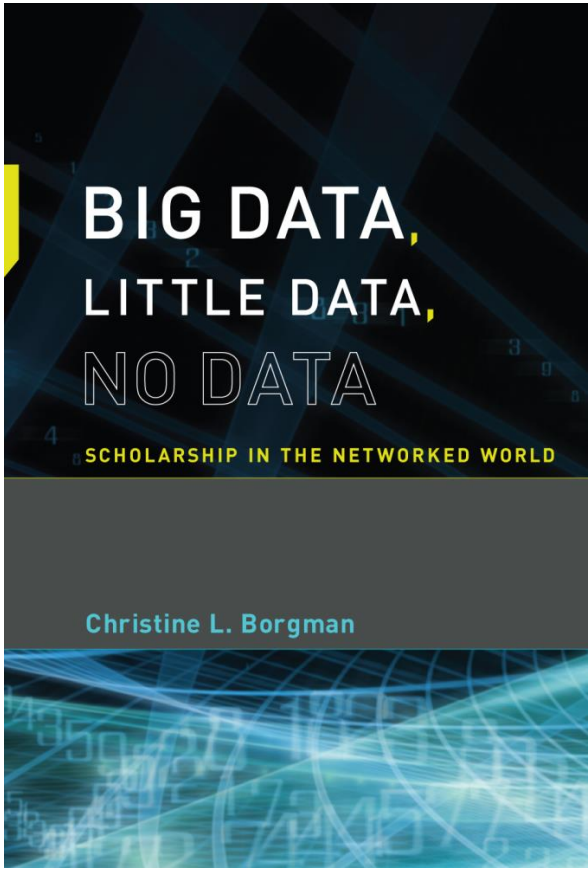
For further data, please see the corresponding table of repositories sorted by country.

Show embedding code

Show legacy chart and embedding code

Conclusions

- Data
  - Representations used as evidence
  - One person's signal is another's noise
- Publications
  - Peer-reviewed documents
  - Networks of research objects
- Publications vs. data
  - Authorship vs. attribution, credit
  - Relationships are many to many
- Releasing data
  - "Publishing data" is an oxymoron
  - Lack of incentives to release or reuse
- Keeping data useful
  - Invest in repositories, infrastructure, expertise
  - Value propositions vary



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, Alberto Pepe, Kalpana Shankar, Katie Shilton, Jillian Wallis, Laura Wynholds, Kan Zhang

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