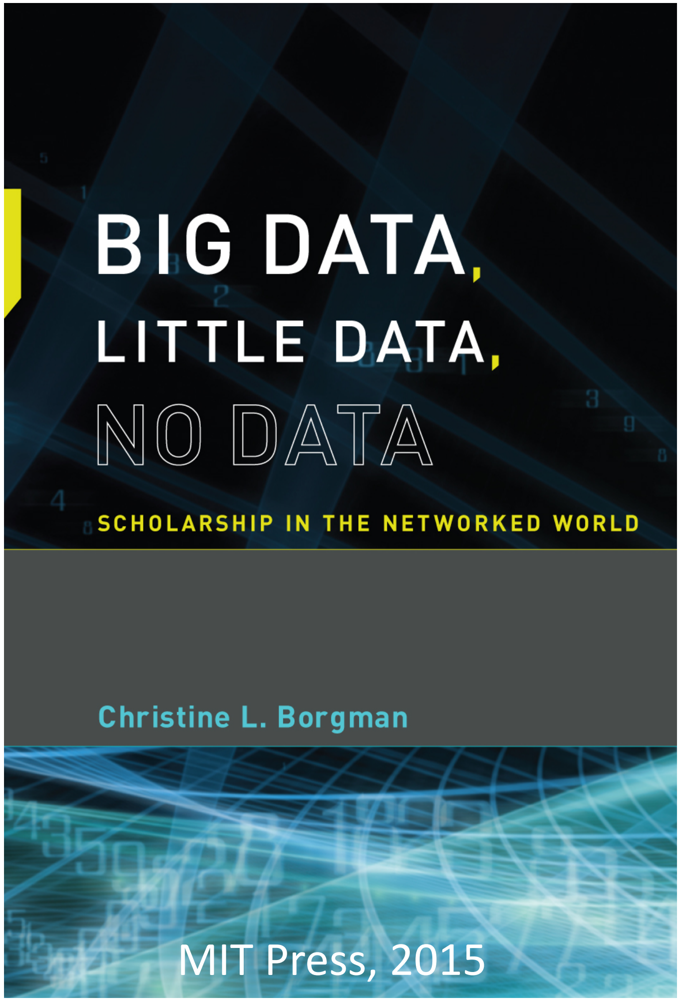
Big Data, Little Data, or No Data?
Scholarship and Stewardship to Build the UC Digital Library

Christine L. Borgman
Distinguished Professor &
Presidential Chair in Information Studies
Director, Center for Knowledge Infrastructures
https://knowledgeinfrastructures.gseis.ucla.edu
University of California, Los Angeles
http://christineborgman.info
@scitechprof

Digital Library Federation X Conference
University of California, Riverside
February 27, 2018

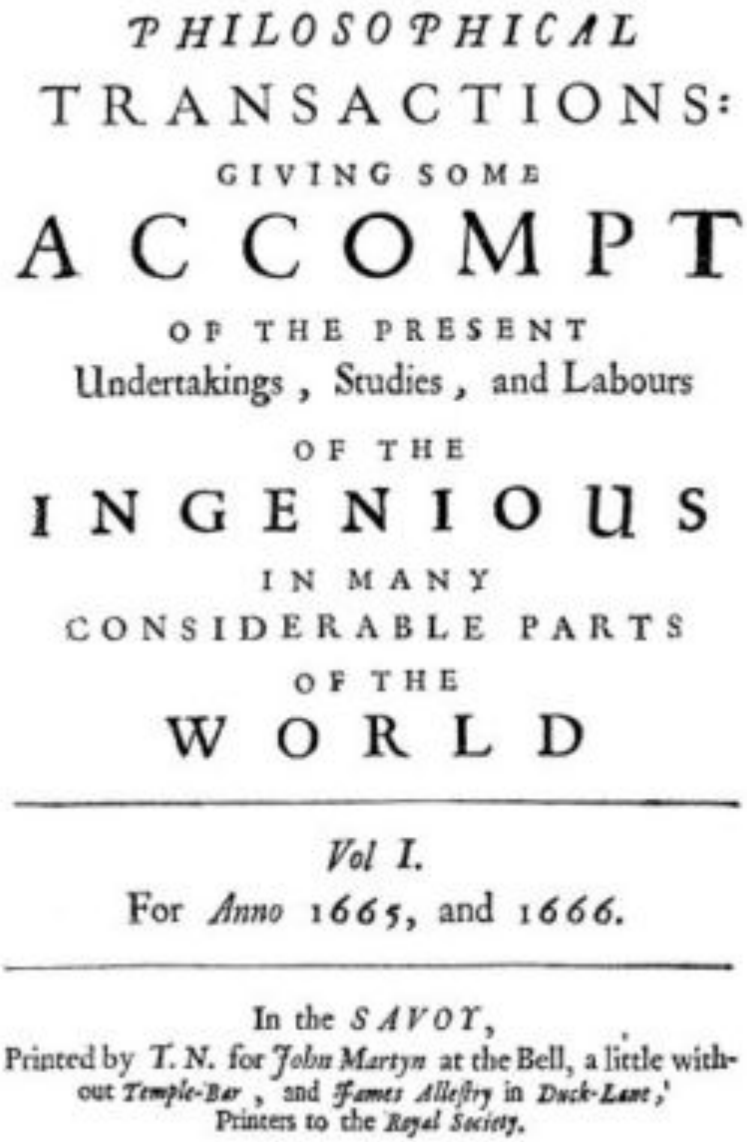


BIG DATA, LITTLE DATA, NO DATA

SCHOLARSHIP IN THE NETWORKED WORLD

Christine L. Borgman

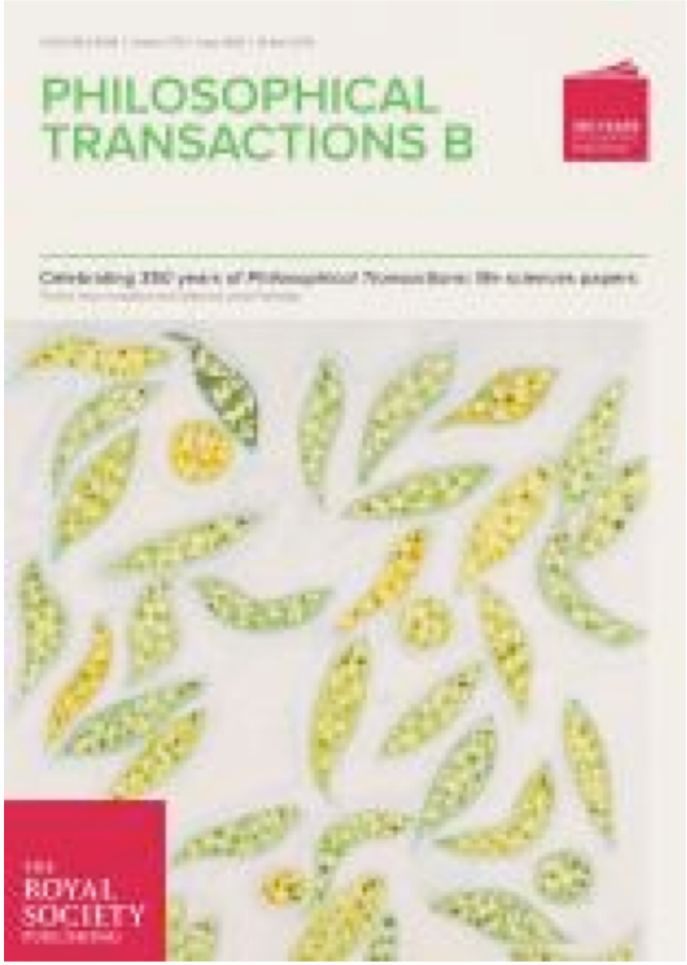
MIT Press, 2015



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*.



PHILOSOPHICAL
TRANSACTIONS B

Celebrating 350 years of Philosophical Transactions: life sciences papers

THE ROYAL SOCIETY PUBLISHING

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

THE WHITE HOUSE
WASHINGTON

Data sharing policies

E·S·R·C
ECONOMIC & SOCIAL RESEARCH COUNCIL

JAPAN SCIENCE AND TECHNOLOGY AGENCY

中央研究院
ACADEMIA SINICA 1928

CHINESE ACADEMY OF SCIENCES

- European Union
- U.S. Federal research policy
- Research Councils of the UK
- Australian Research Council
- Individual countries, funding agencies, journals, universities

Supported by wellcometrust

Australian Government
National Health and Medical Research Council

National Science Foundation
WHERE DISCOVERIES BEGIN

Policy RECommendations for Open Access to Research Data in Europe

RECODE

NIH National Institutes of Health
Turning Discovery Into Health

RDA

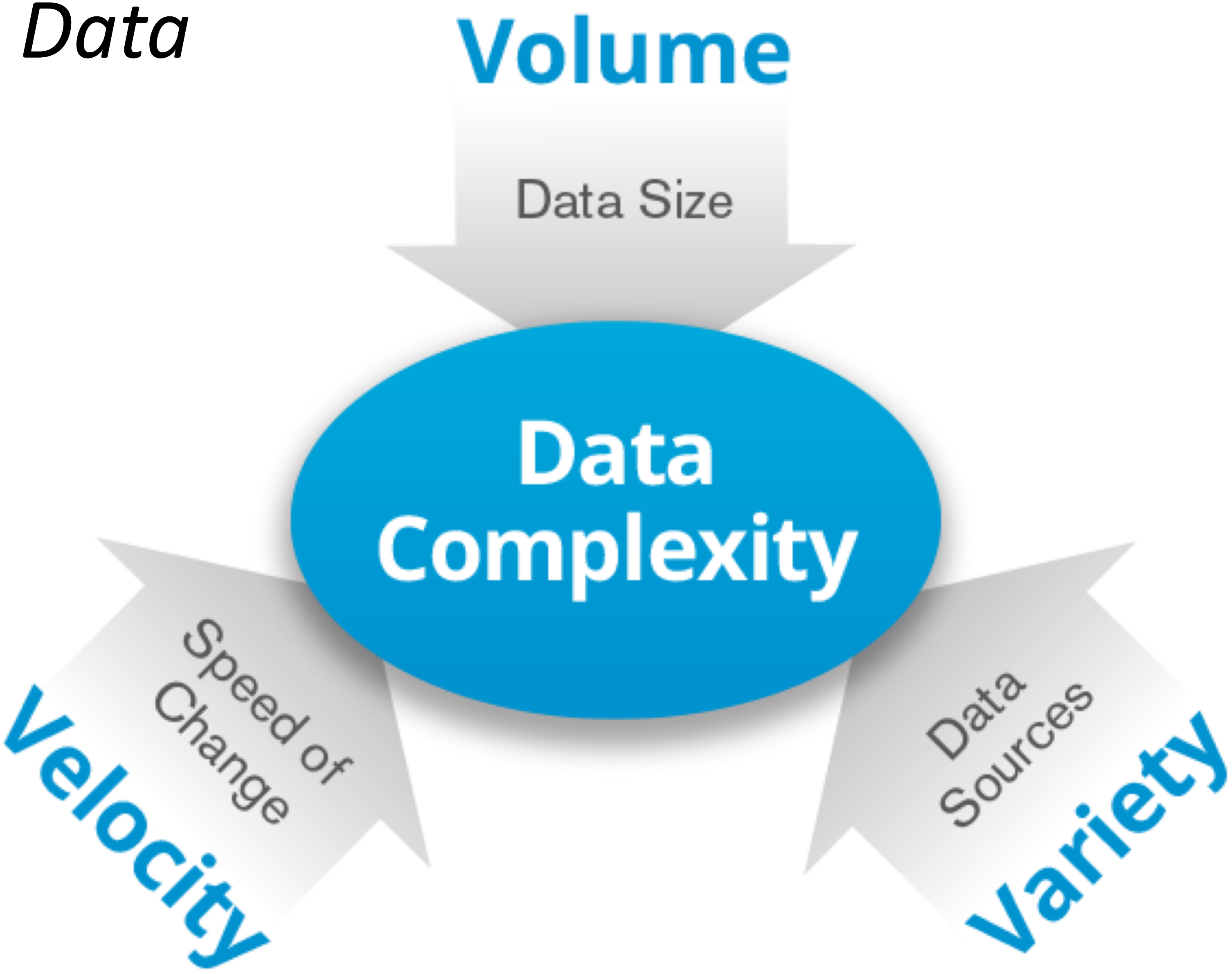
RESEARCH DATA ALLIANCE

Research Data Sharing
without barriers

Precondition:

Researchers share data

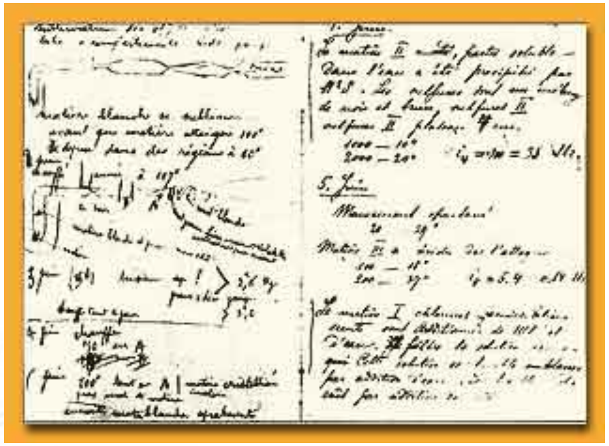
Big Data



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

What are data?

hudsonalpha.org



Marie Curie’s notebook aip.org

Pisa Griffin

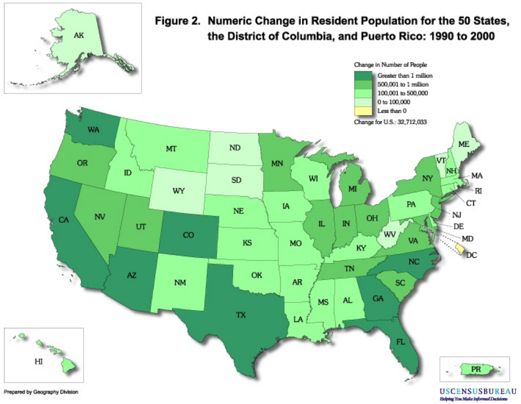
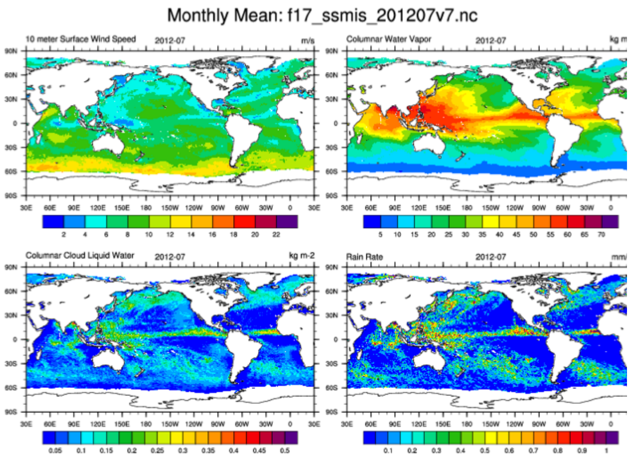
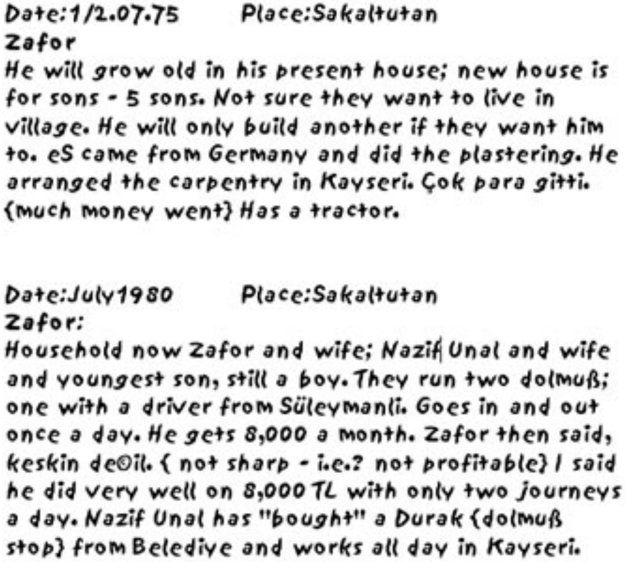


Figure 2. Numeric Change in Resident Population for the 50 States, the District of Columbia, and Puerto Rico: 1990 to 2000

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

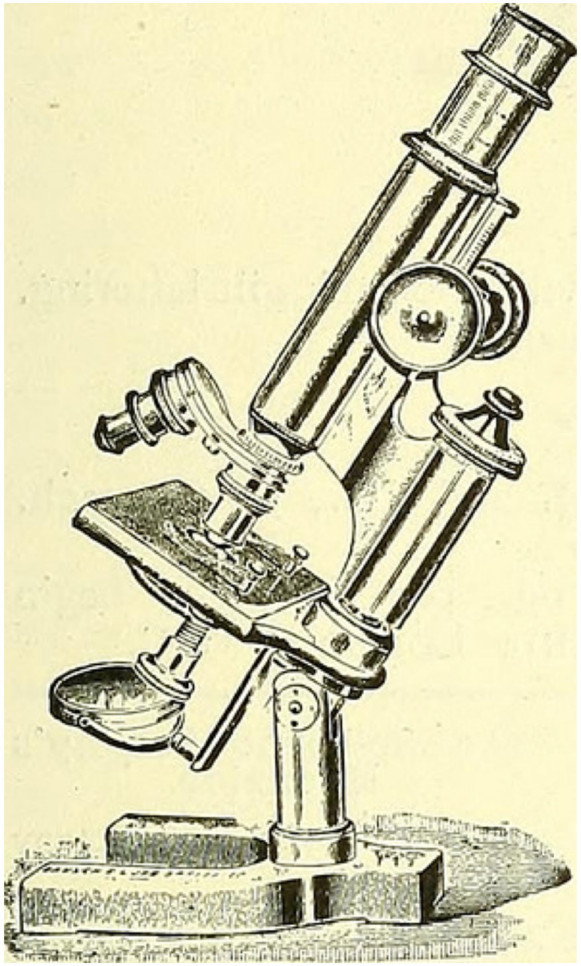
http://www.genome.gov/dmd/img.cfm?node=Photos/Graphics&id=85327

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

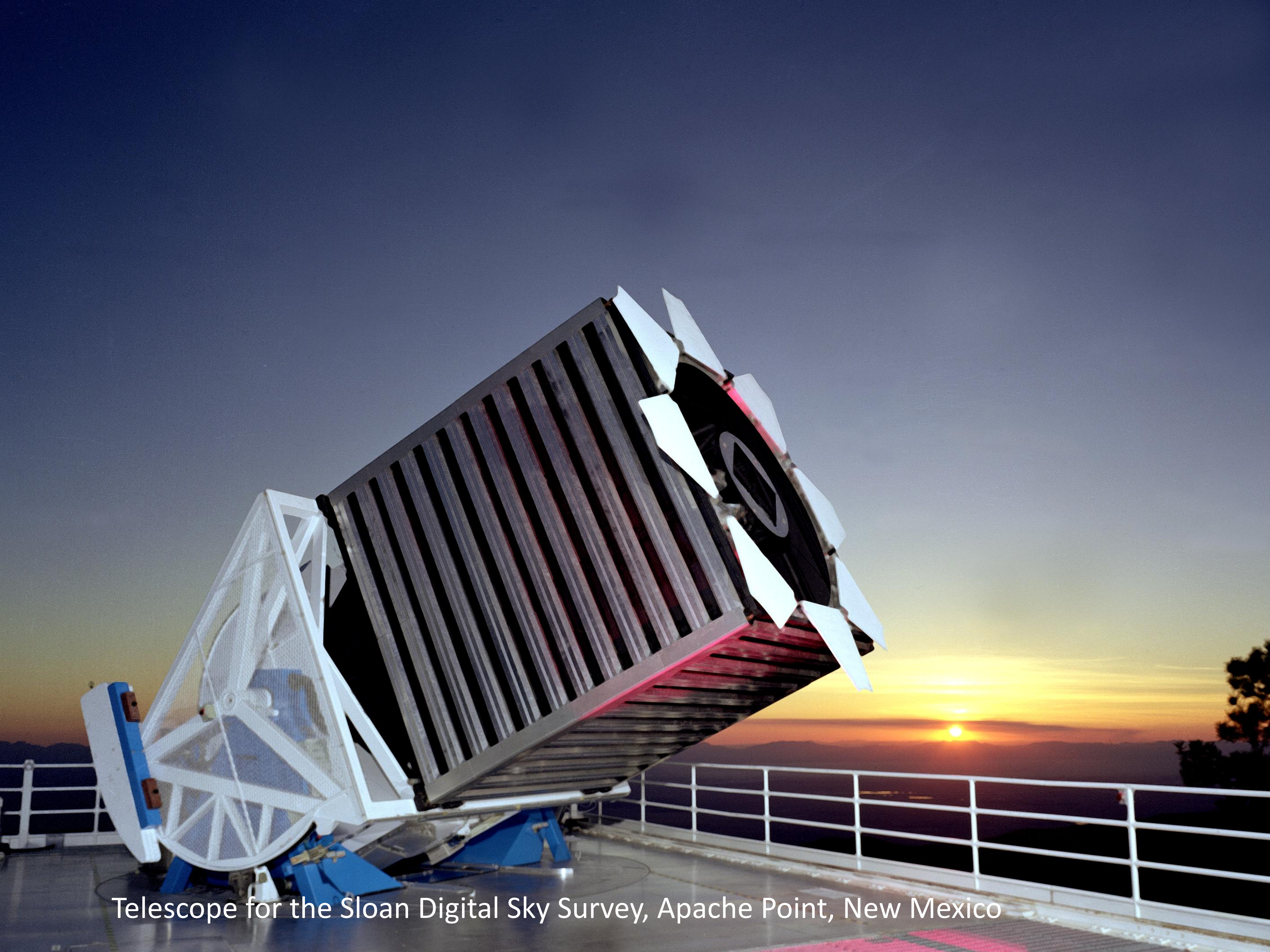
C.L. Borgman (2015). *Big Data, Little Data, No Data: Scholarship in the Networked World*. MIT Press

Research process

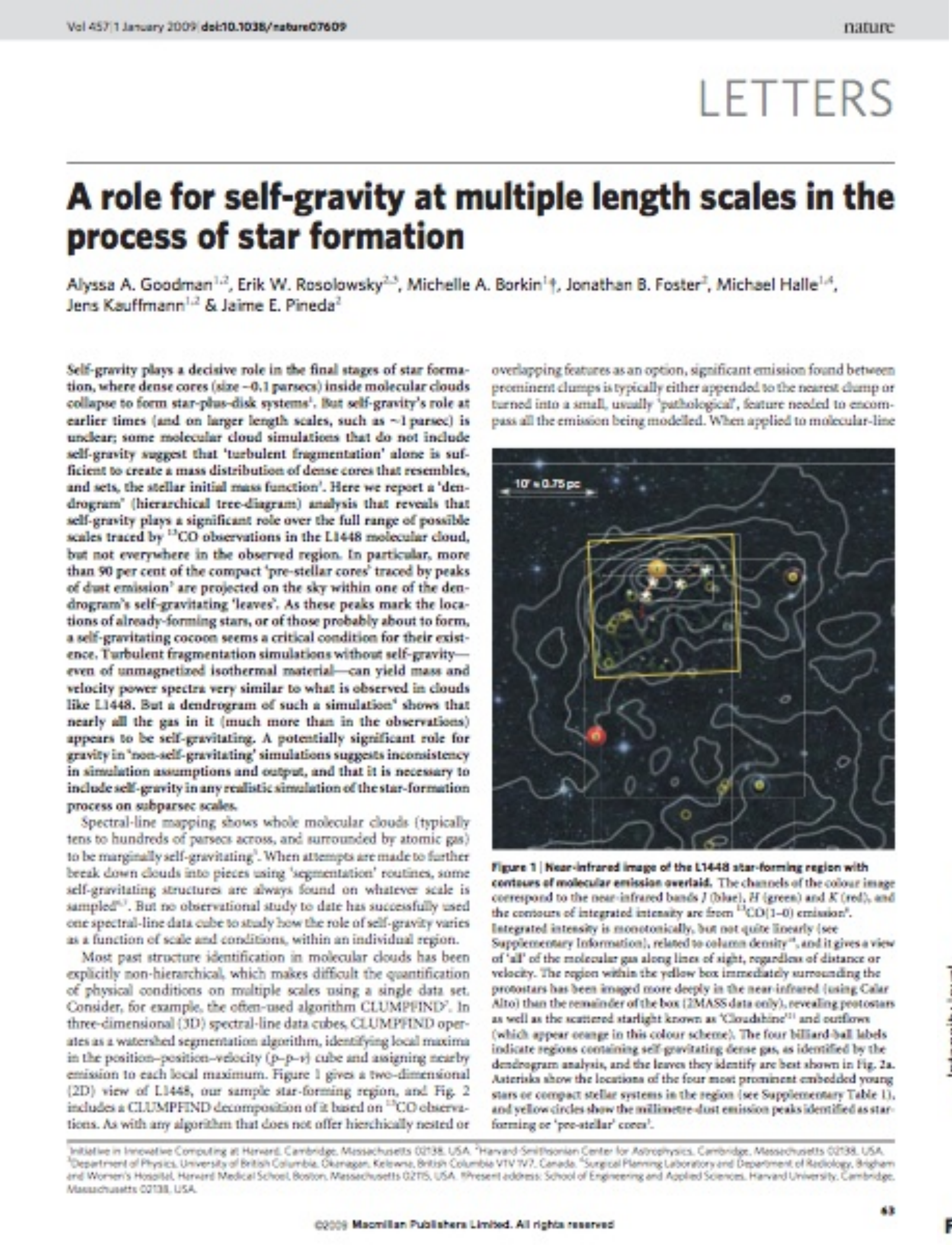
- Models and theories
- Research questions
- Methods
  - – Domain expertise
  - – Practices, protocols
  - – Data sources
  - – Instruments, software
  - – Infrastructure



Commons photo: Science Gossip, 1894



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


Vol 457 | 1 January 2009 | doi:10.1038/nature07609

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]

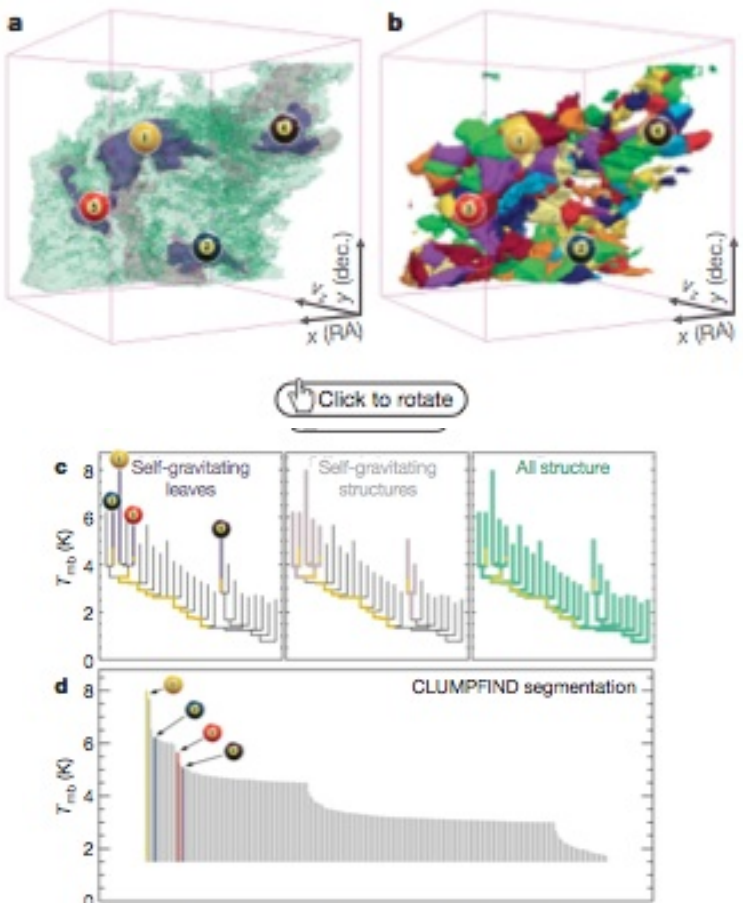
Self-gravity plays a decisive role in the final stages of star formation, where dense cores (size ~0.1 parsecs) inside molecular clouds collapse to form star-plus-disk systems[1]. But self-gravity's role at earlier times (and on larger length scales, such as ~1 parsec) is unclear; some molecular cloud 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[10], 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.



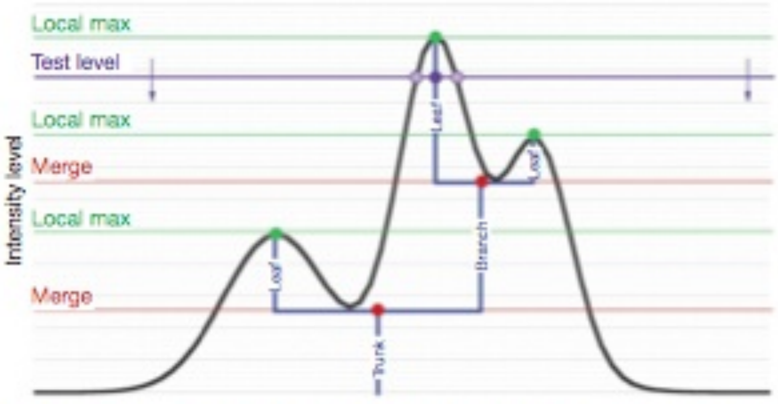
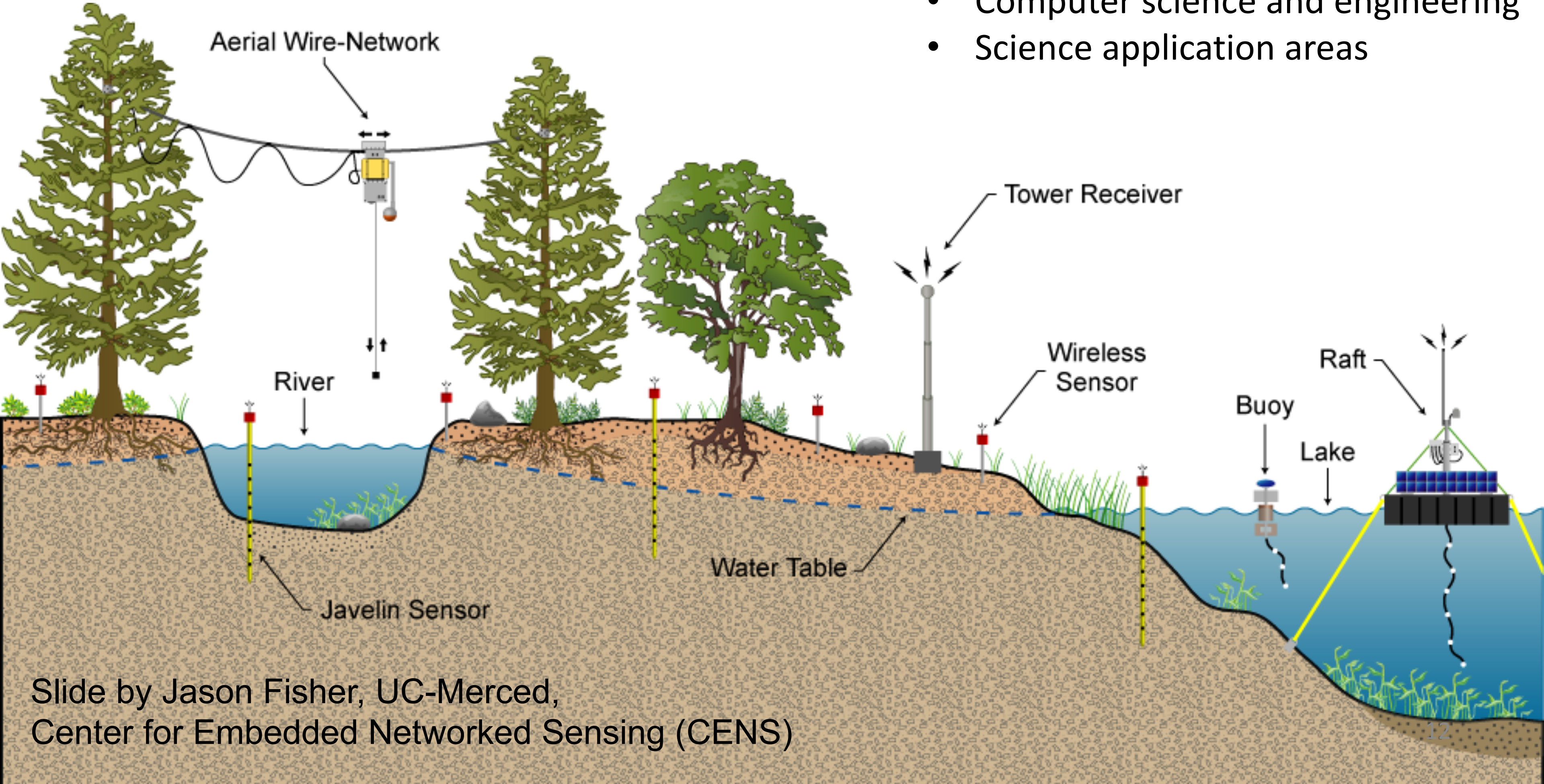


Figure 3 | Schematic illustration of the dendrogram process. Shown is the

Center for Embedded Networked Sensing

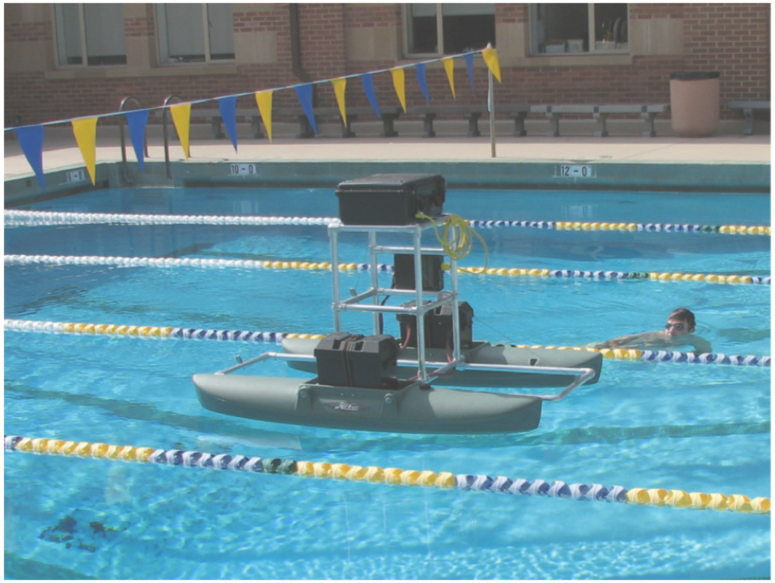
- NSF Science & Tech Ctr, 2002-2012
- 5 universities, plus partners
- 300 members
- Computer science and engineering
- Science application areas



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

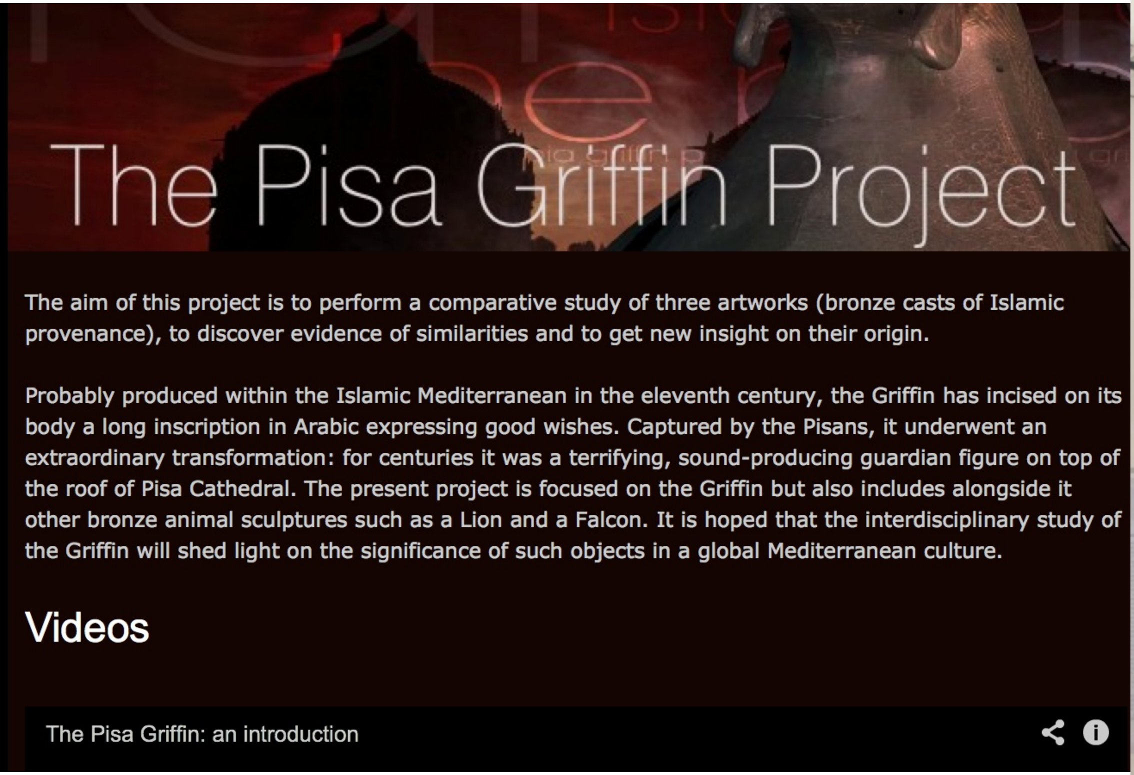
Science <–> Data

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.."*



The Pisa Griffin Project

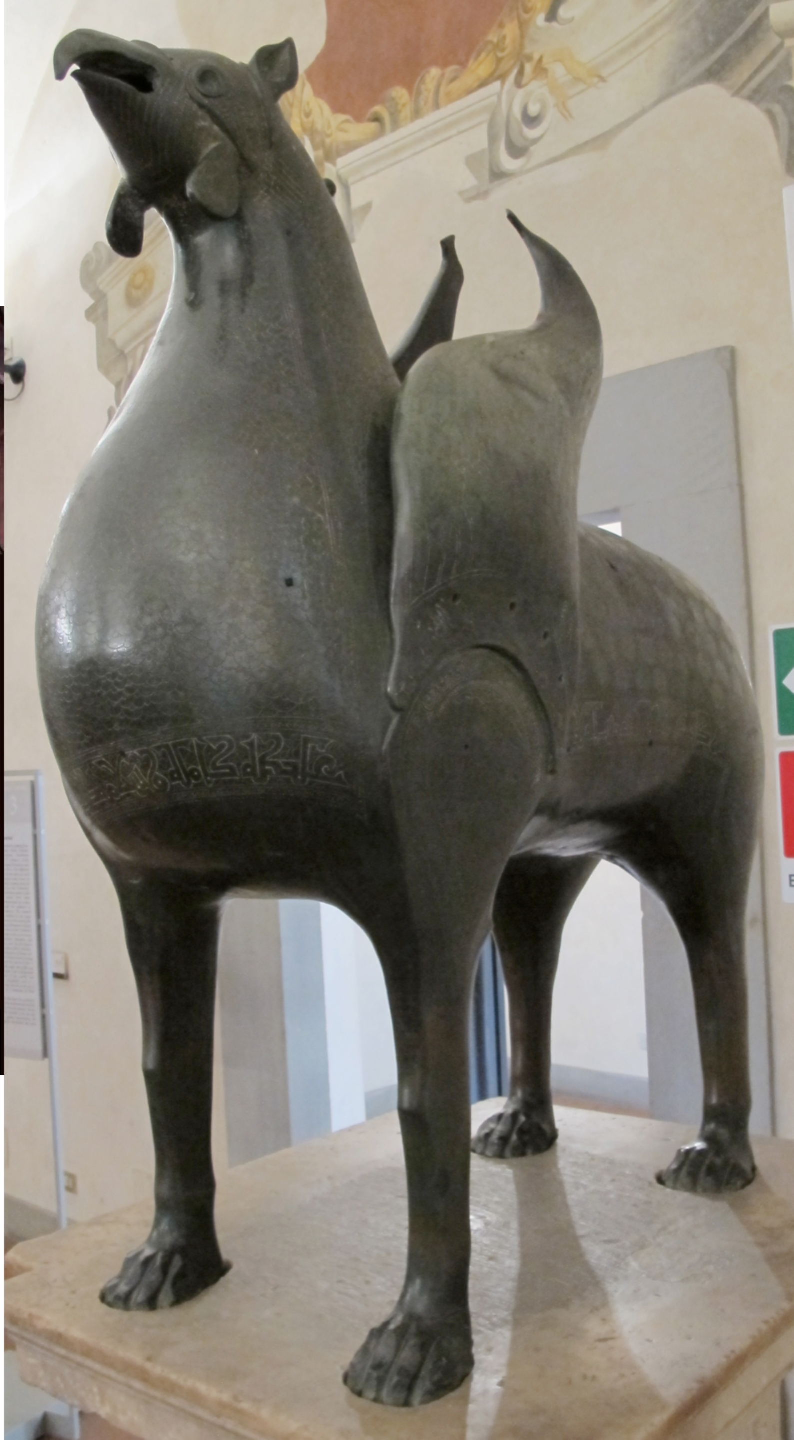
The aim of this project is to perform a comparative study of three artworks (bronze casts of Islamic provenance), to discover evidence of similarities and to get new insight on their origin.

Probably produced within the Islamic Mediterranean in the eleventh century, the Griffin has incised on its body a long inscription in Arabic expressing good wishes. Captured by the Pisans, it underwent an extraordinary transformation: for centuries it was a terrifying, sound-producing guardian figure on top of the roof of Pisa Cathedral. The present project is focused on the Griffin but also includes alongside it other bronze animal sculptures such as a Lion and a Falcon. It is hoped that the interdisciplinary study of the Griffin will shed light on the significance of such objects in a global Mediterranean culture.

Videos

The Pisa Griffin: an introduction

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



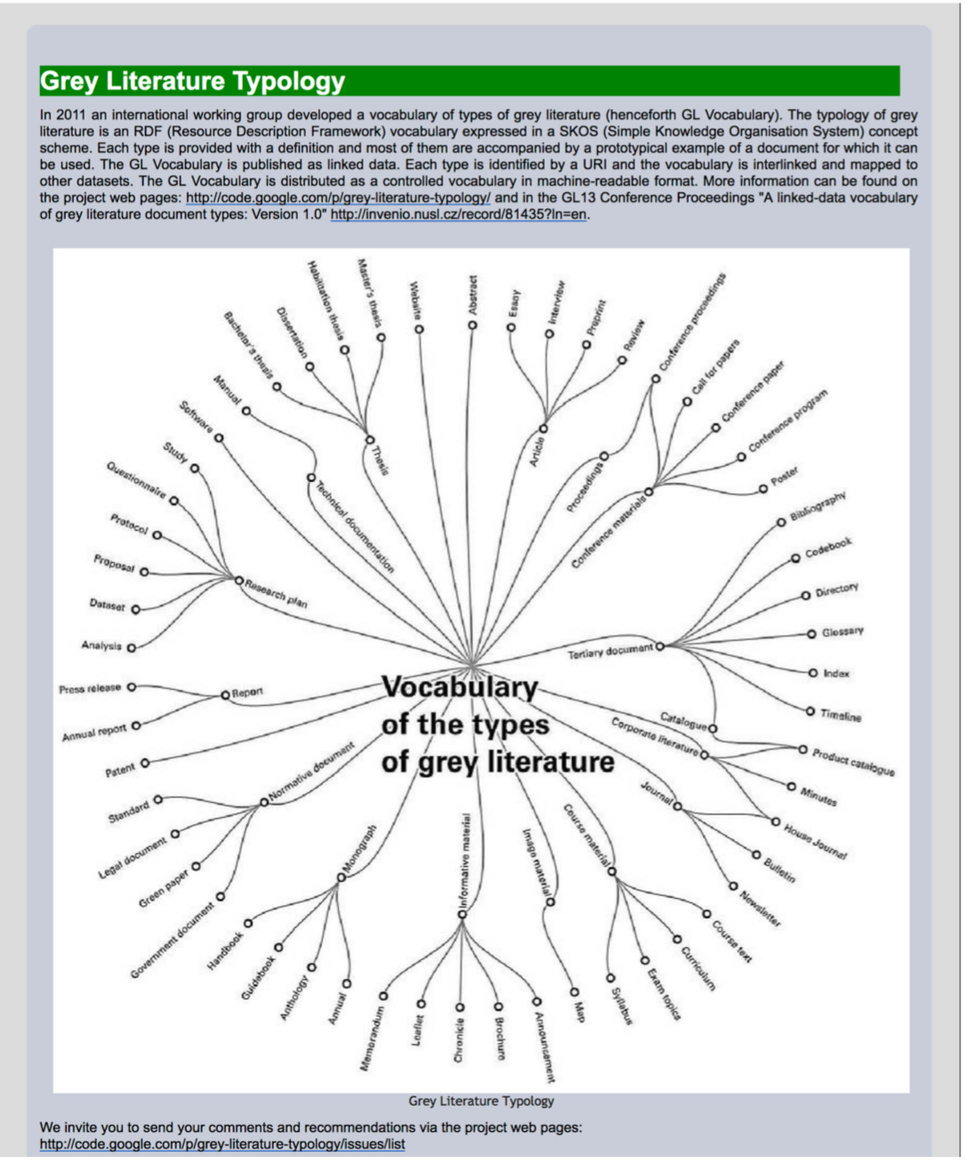
Arte islamica, ippogrifo, XI sec 03, own work

Publications

Dégradation ER mRNA...
Estrogen Receptors in Human Nontarget Tissues: Biological and Clinical Implications*
DANIEL R. CIOCCA AND LAURA M. VARGAS ROIG
Endocr Rev 1995
1.3.2.1
Effects of Low Dose Oral Contraceptives on Very Low Density Lipoprotein Metabolism
Abstract
1.2.2.5
The Estradiol-stimulated Lipoprotein Receptor of Rat Liver

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

Grey Literature



Grey Literature Typology

In 2011 an international working group developed a vocabulary of types of grey literature (henceforth GL Vocabulary). The typology of grey literature is an RDF (Resource Description Framework) vocabulary expressed in a SKOS (Simple Knowledge Organisation System) concept scheme. Each type is provided with a definition and most of them are accompanied by a prototypical example of a document for which it can be used. The GL Vocabulary is published as linked data. Each type is identified by a URI and the vocabulary is interlinked and mapped to other datasets. The GL Vocabulary is distributed as a controlled vocabulary in machine-readable format. More information can be found on the project web pages: http://code.google.com/p/grey-literature-typology/ and in the GL13 Conference Proceedings "A linked-data vocabulary of grey literature document types: Version 1.0" http://invenio.nusl.cz/record/81435?ln=en.

Grey Literature Typology

We invite you to send your comments and recommendations via the project web pages: http://code.google.com/p/grey-literature-typology/issues/list

- Reports
- Working papers
- Conference papers
- Preprints
- Patents
- Datasets
- Audio
- Video
- Slides
- Posters
- Codebooks
- Course syllabi
- Proposals
- Memos

http://www.greynet.org/

Grey Data

Stuart Miles: FreeDigitalPhotos.net

HIPAA

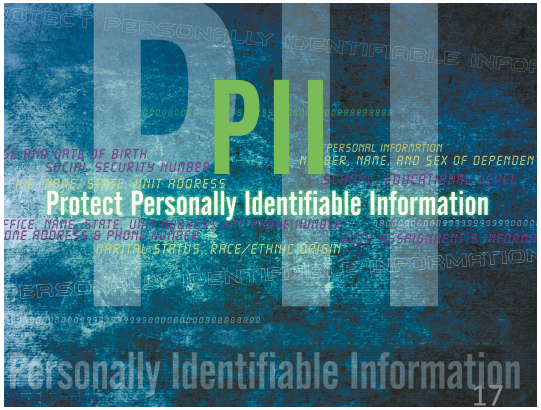
Health Insurance Portability and Accountability Act

https://www.linkedin.com/pulse/hipaa-privacy-rule-compliance-understanding-new-rules-syed-najaf

- Student applications
- Registrar records
- Learning management systems
- University ID cards: library, health, recreation, dorms, food service, transportation…
- Academic personnel dossiers
- Regulation and compliance data
- Staff surveys
- Sensor networks
- Security cameras
- Network traffic
- Street traffic…

RESPECT FOR PERSONS • JUSTICE • BENEFICENCE •
UNIVERSITY OF WASHINGTON
IRB
HUMAN SUBJECTS DIVISION

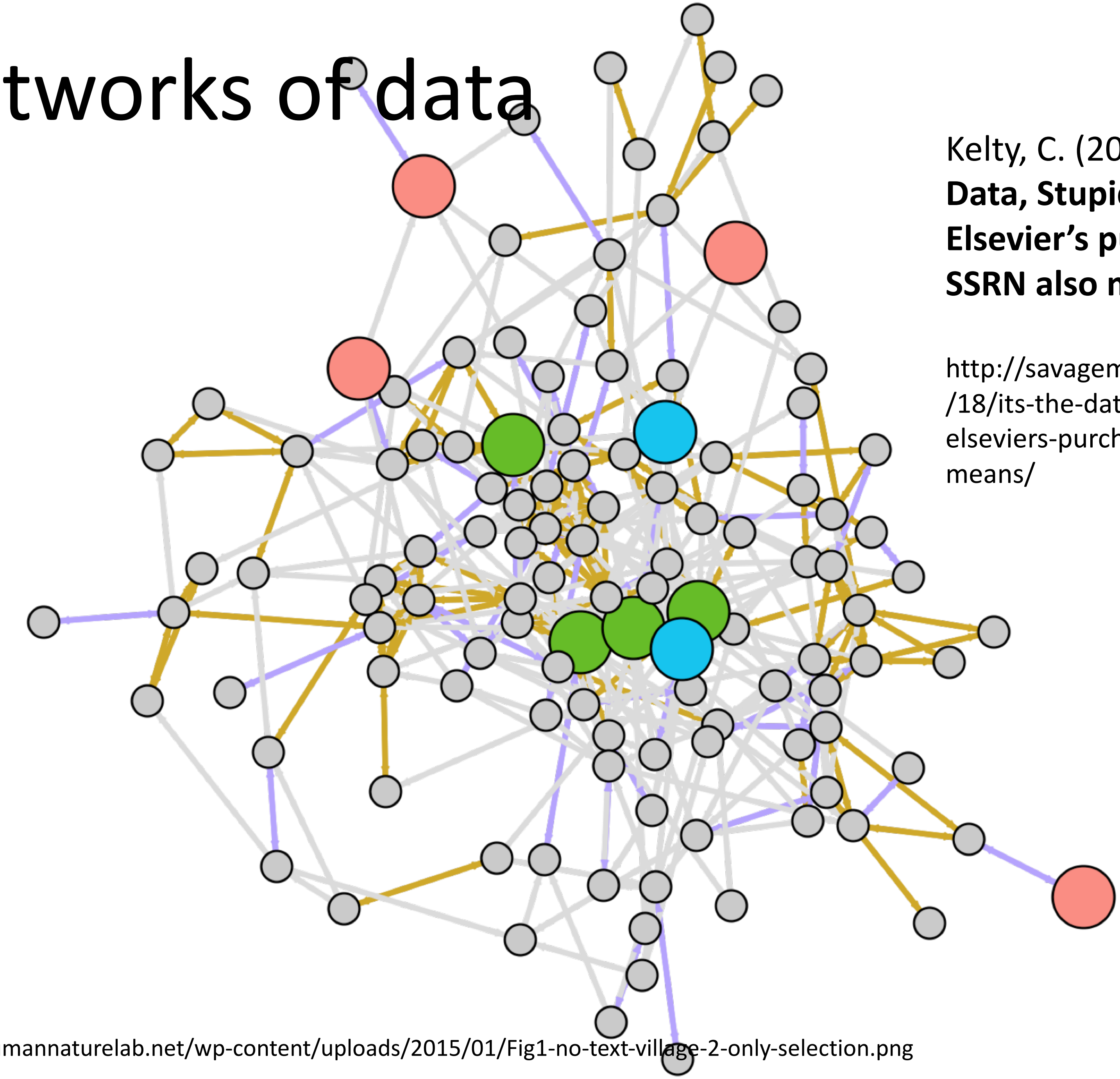
DEPARTMENT OF EDUCATION • UNITED STATES OF AMERICA
FERPA
Family Educational Rights and Privacy Act


PII
Protect Personally Identifiable Information
Personally Identifiable Information

Borgman, C. L. (2018). Open Data, Grey Data, and Stewardship: Universities at the Privacy Frontier. *Berkeley Technology Law Journal*, *33*(2). https://arxiv.org/abs/1802.02953

http://www.aetc.af.mil/News/Article-Display/Article/559551/think-before-sending-protecting-pii/

Networks of data



Kelty, C. (2016). **It's the Data, Stupid: What Elsevier's purchase of SSRN also means.**

http://savageminds.org/2016/05/18/its-the-data-stupid-what-elseviers-purchase-of-ssrn-also-means/

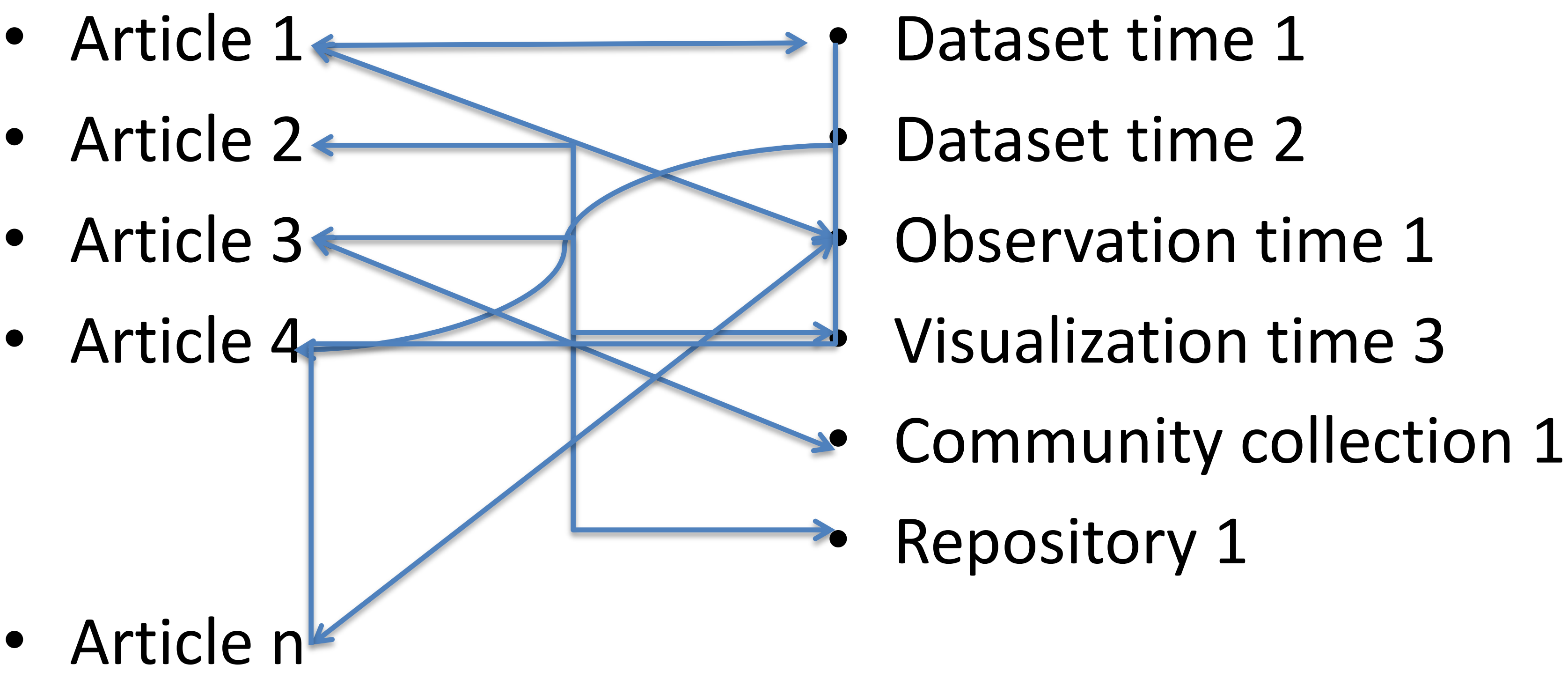
http://humannaturelab.net/wp-content/uploads/2015/01/Fig1-no-text-village-2-only-selection.png

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



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

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

Publications <–> Data: Attribution

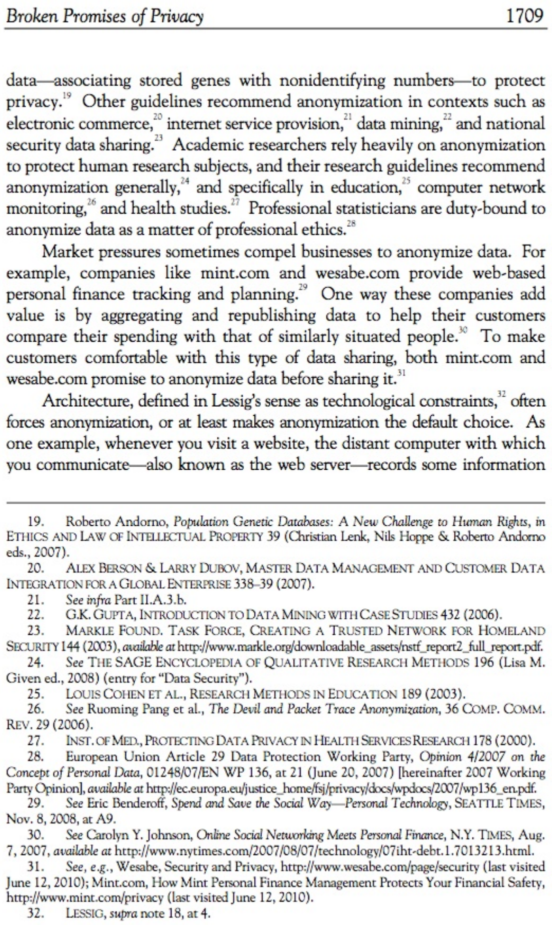
- Publications
  - Independent units
  - Authorship is negotiated
- Data
  - Compound objects
  - Ownership is rarely clear
  - Attribution
    - Long term responsibility: Investigators
    - Expertise for interpretation: Data collectors and analysts

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

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 (2003).

[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.

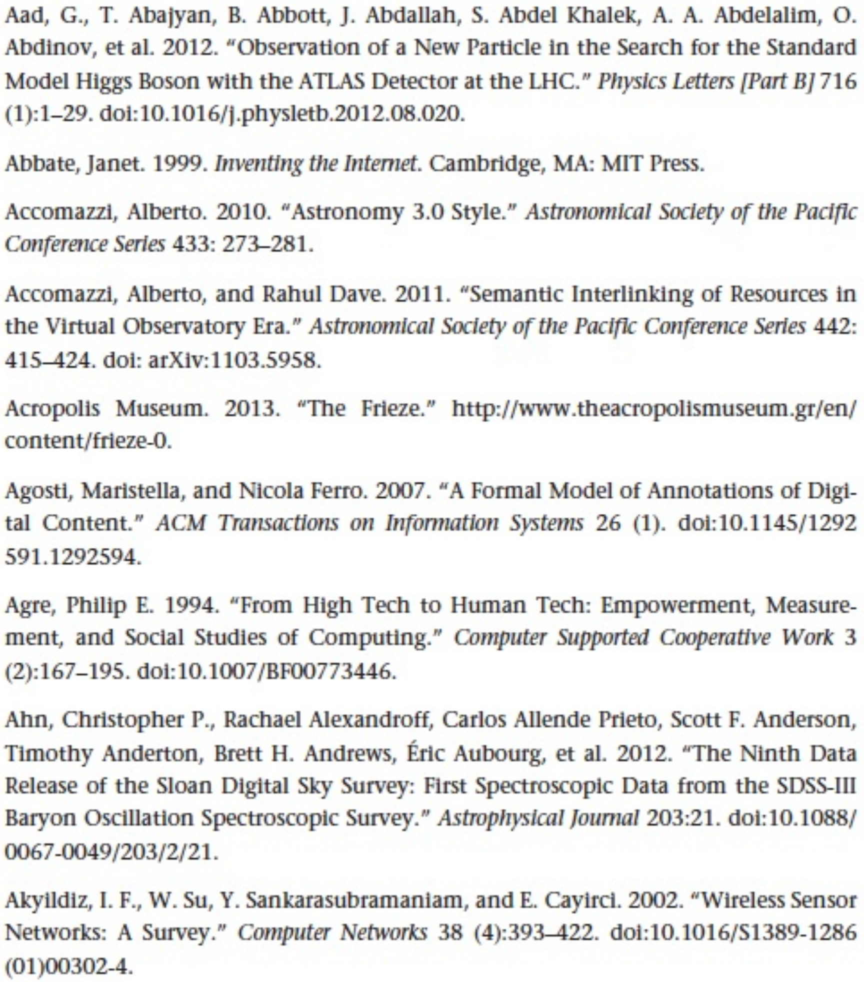
[29] *See* Eric Benderoff, *Spend and Save the Social Way—Personal Technology*, SEATTLE TIMES, Nov. 8, 2008, at A9.

[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.

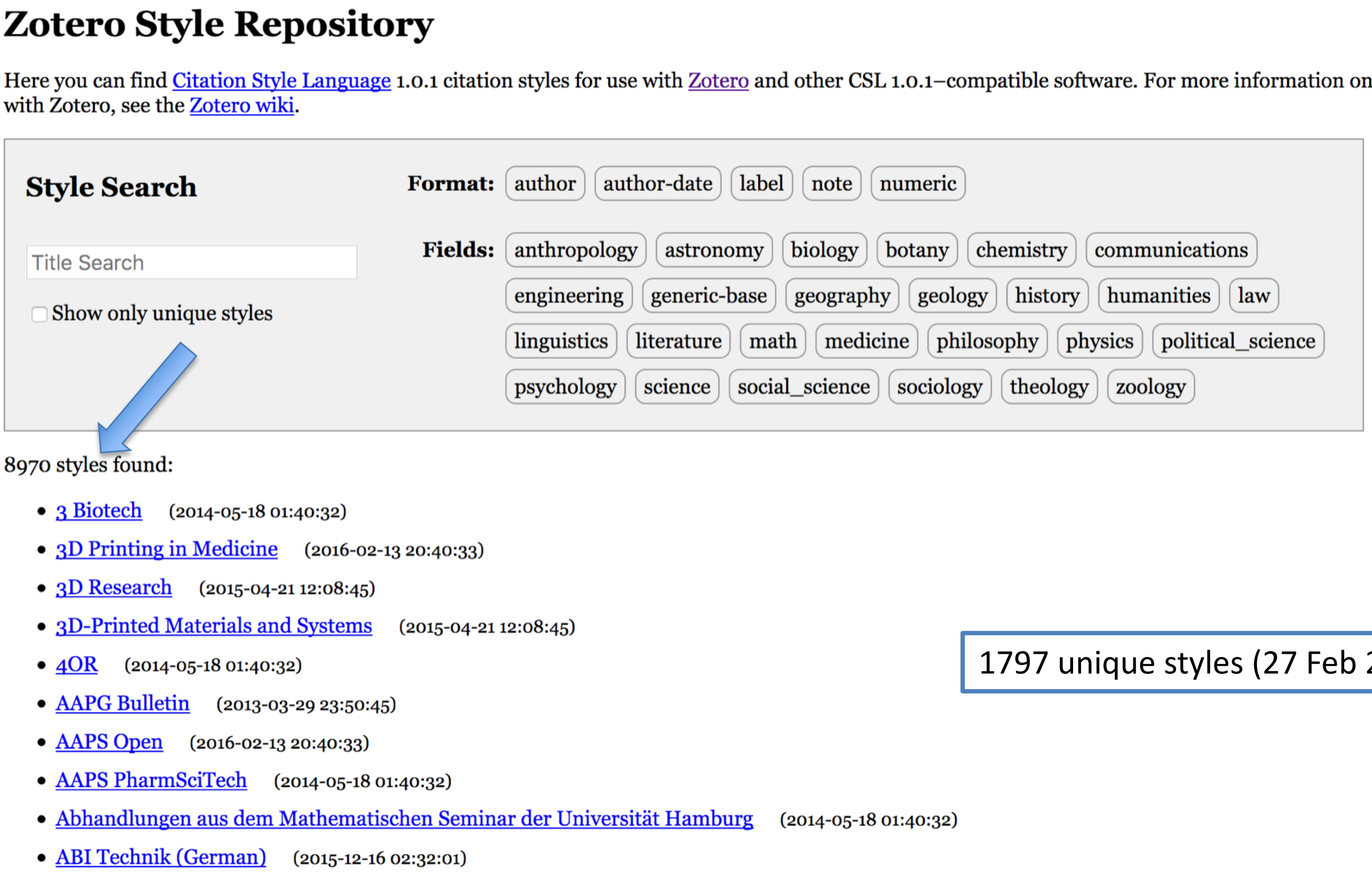
Agre, Philip E. 1994. "From High Tech to Human Tech: Empowerment, Measurement, and Social Studies of Computing." *Computer Supported Cooperative Work* 3 (2):167–195. doi:10.1007/BF00773446.

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.

Bibliographic styles



Zotero Style Repository

Here you can find Citation Style Language 1.0.1 citation styles for use with Zotero and other CSL 1.0.1–compatible software. For more information on using CSL styles with Zotero, see the Zotero wiki.

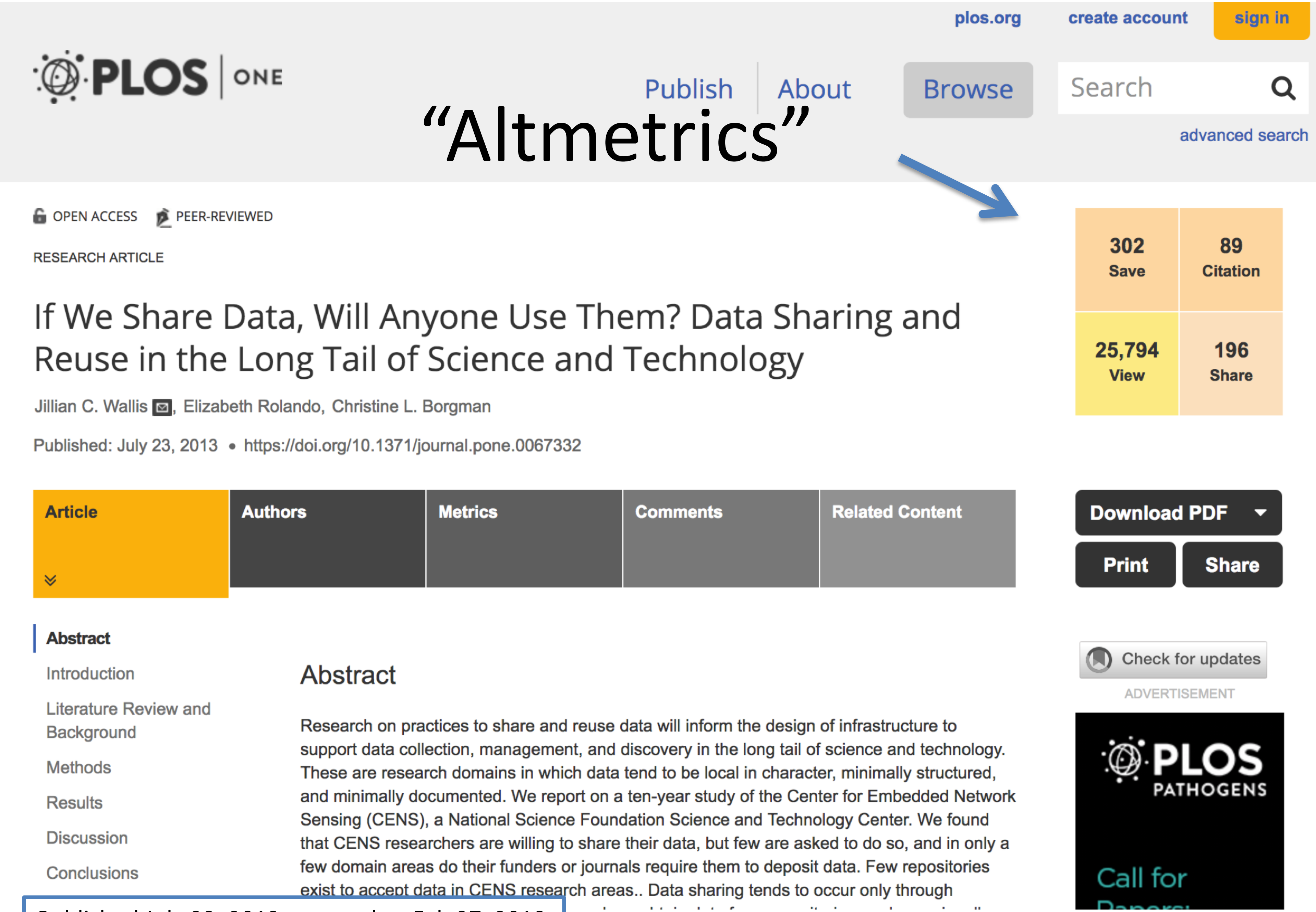
Style Search
Title Search
□ Show only unique styles
Format: author author-date label note numeric
Fields: anthropology astronomy biology botany chemistry communications engineering generic-base geography geology history humanities law linguistics literature math medicine philosophy physics political_science psychology science social_science sociology theology zoology

8970 styles found:

- 3 Biotech (2014-05-18 01:40:32)
- 3D Printing in Medicine (2016-02-13 20:40:33)
- 3D Research (2015-04-21 12:08:45)
- 3D-Printed Materials and Systems (2015-04-21 12:08:45)
- 4OR (2014-05-18 01:40:32)
- AAPG Bulletin (2013-03-29 23:50:45)
- AAPS Open (2016-02-13 20:40:33)
- AAPS PharmSciTech (2014-05-18 01:40:32)
- Abhandlungen aus dem Mathematischen Seminar der Universität Hamburg (2014-05-18 01:40:32)
- ABI Technik (German) (2015-12-16 02:32:01)

1797 unique styles (27 Feb 2018)

"Altmetrics"



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 • https://doi.org/10.1371/journal.pone.0067332

302 Save	89 Citation
25,794 View	196 Share

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. These are research domains in which data tend to be local in character, minimally structured, and minimally documented. We report on a ten-year study of the Center for Embedded Network Sensing (CENS), a National Science Foundation Science and Technology Center. We found that CENS researchers are willing to share their data, but few are asked to do so, and in only a few domain areas do their funders or journals require them to deposit data. Few repositories exist to accept data in CENS research areas.. Data sharing tends to occur only through

Published July 23, 2013; screenshot Feb 27, 2018

Bibliometrics by Source

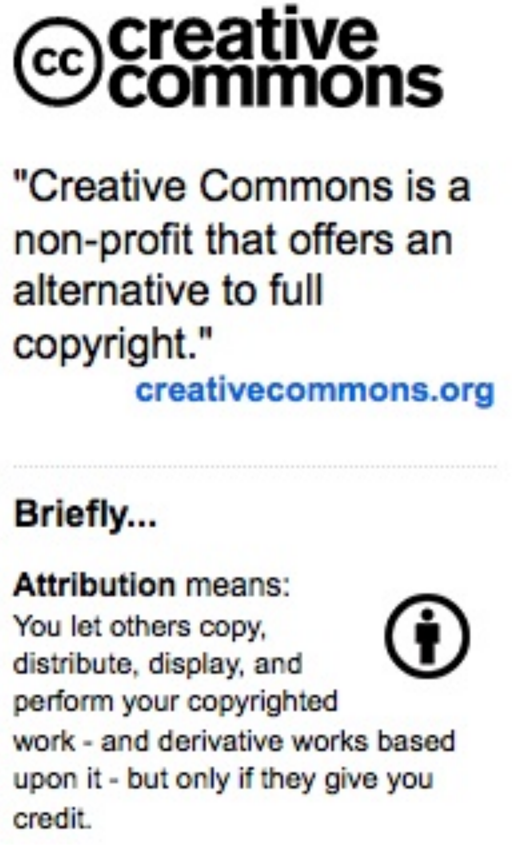
Searches for author: Christine Borgman, Christine L. Borgman, CL Borgman (excluding other C Borgman authors) on July 28, 2014 and February 25, 2016 for Google Scholar, Web of Science, Scopus

UCLA cancelled Scopus subscription by 2016

Source	Publications		Citations received		H-index	
	2014	2016	2014	2016	2014	2016
Google Scholar (Google)	380	443	7766	9701	39	43
Web of Science (Thomson-Reuters)	145	150	1629	1967	20	23
Scopus – July 2014 (Elsevier)	*77*		*1314*		*14 (after 1995)*	

Attributing responsibility

- 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…



creative commons

"Creative Commons is a non-profit that offers an alternative to full copyright."

creativecommons.org

Briefly...

Attribution means:
You let others copy, distribute, display, and perform your copyrighted work - and derivative works based upon it - but only if they give you credit.

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…

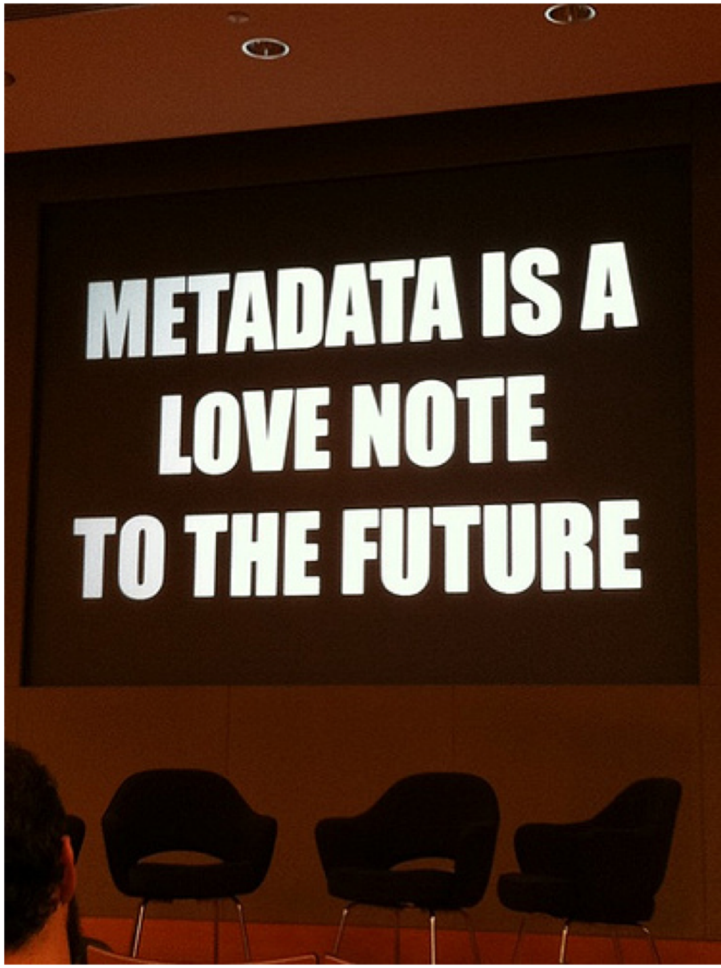
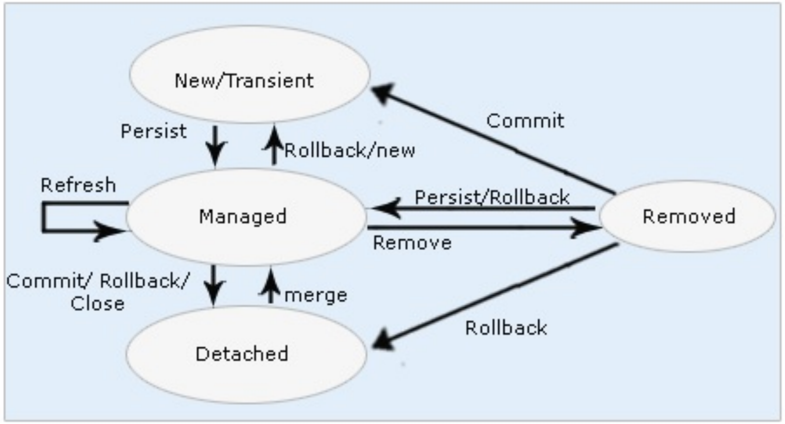


Photo by @kissane; presentation by Jason Scott (@textfiles)

Identity and persistence

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



Persistence Content

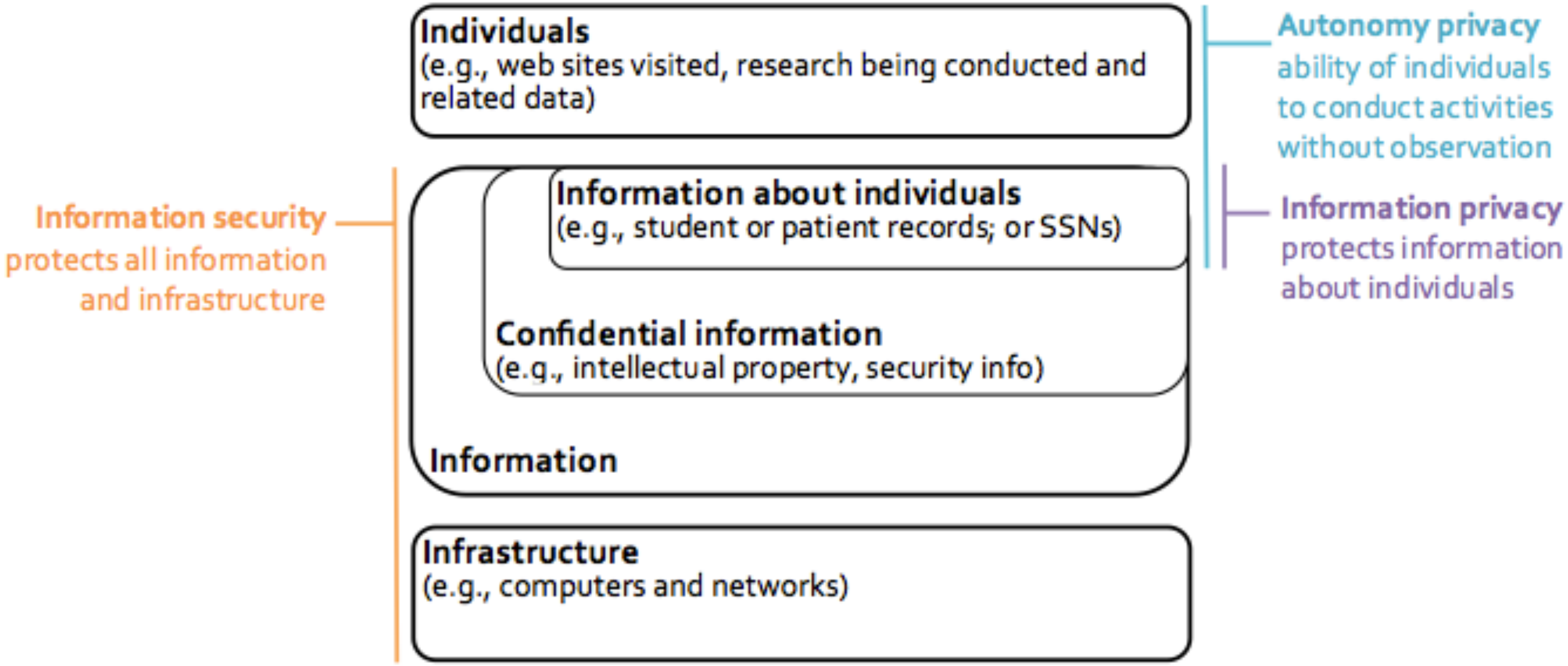
http://web-interview-questions.blogspot.com/2010_06_21_archive.html

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/

Information and Autonomy Privacy



Individuals
(e.g., web sites visited, research being conducted and related data)

Information about individuals
(e.g., student or patient records; or SSNs)

Confidential information
(e.g., intellectual property, security info)

Information

Infrastructure
(e.g., computers and networks)

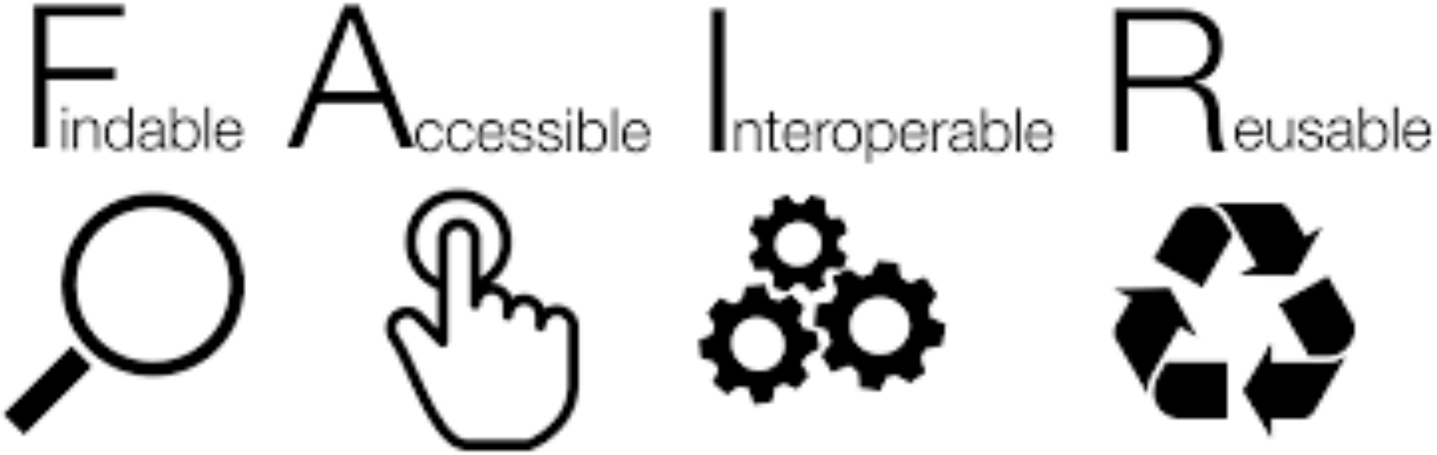
Information security
protects all information and infrastructure

Autonomy privacy
ability of individuals to conduct activities without observation

Information privacy
protects information about individuals

UCOP Privacy and Information Security Initiative. (2013). http://ucop.edu/privacy-initiative/

Data Stewardship: The Ideal



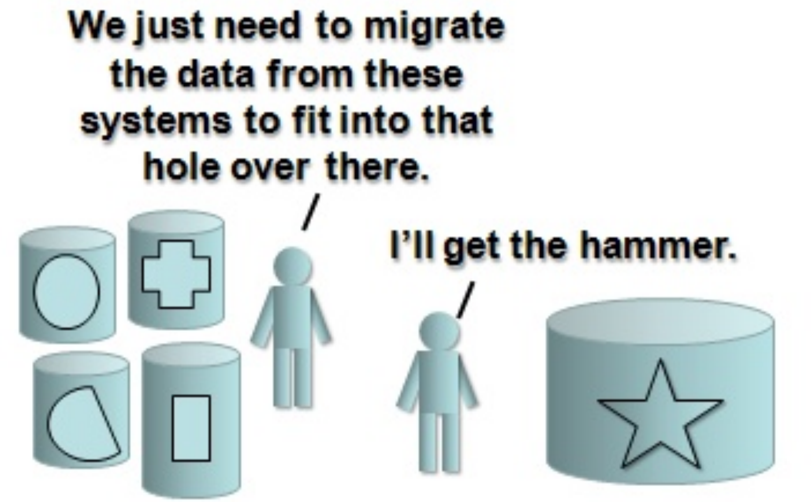
Wilkinson, et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*, http://dx.doi.org/10.1038/sdata.2016.18

Data Stewardship: the Reality

http://www.information-age.com/cloud-computing-pharmaceutical-industry-123462676/

Getty Research Institute

Mount Wilson Solar Observatory, 2017



We just need to migrate the data from these systems to fit into that hole over there.
I'll get the hammer.

http://www.datamartist.com/data-migration-part-1-introduction-to-the-data-migration-delema

http://gsa.rice.edu/

Graduate students

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

Post-doctoral fellows

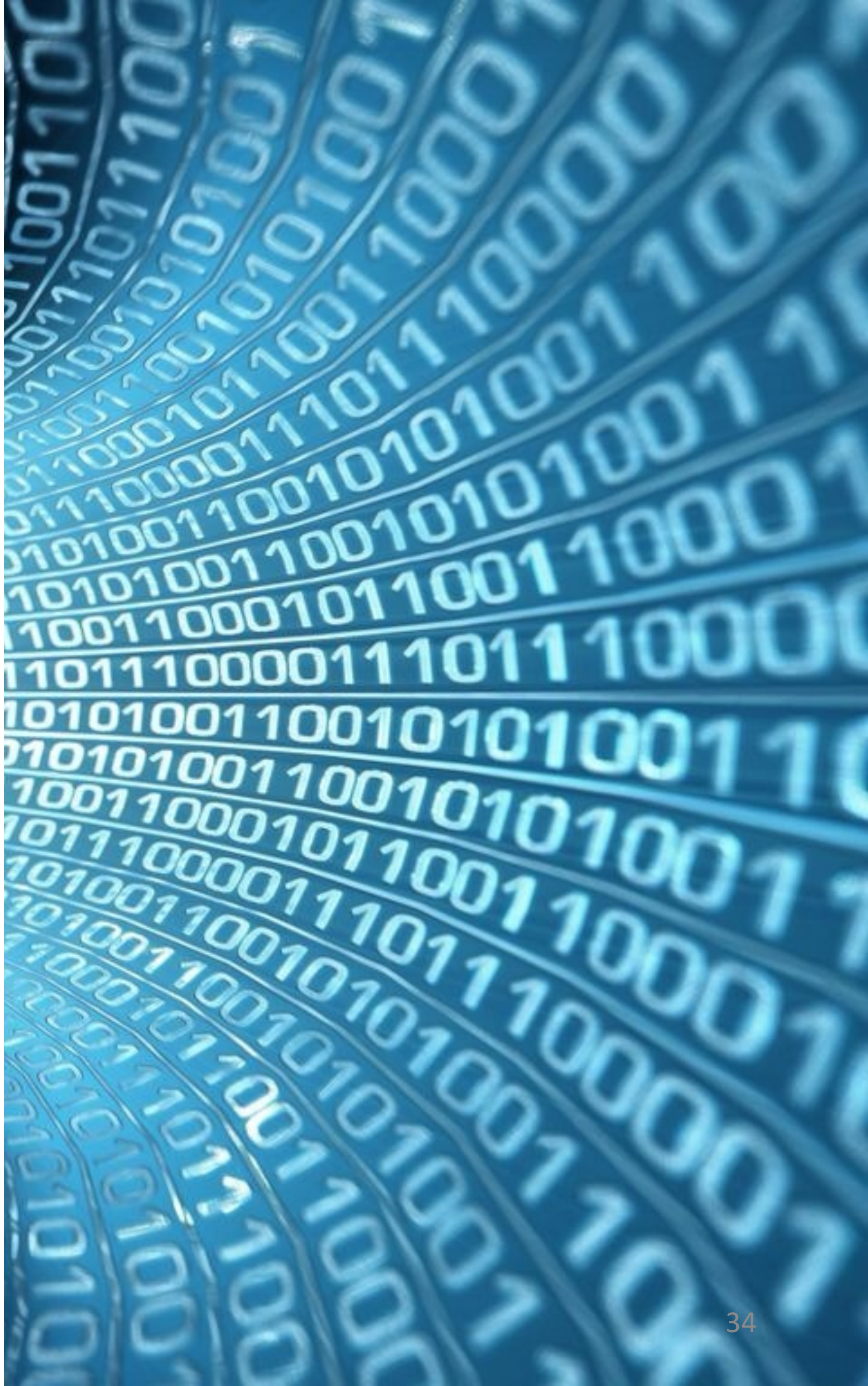
Data

If you can’t protect it, don’t collect it.

(privacy and security aphorism)

Therefore:

If you collect it, you must protect it.



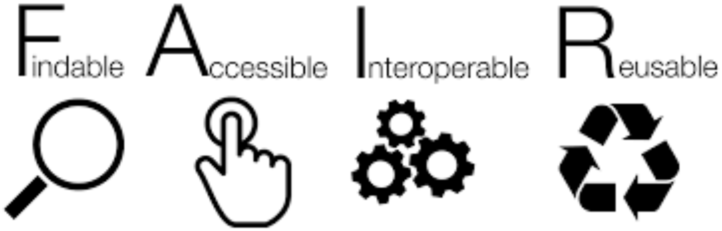
Protect Data and Privacy

open by design

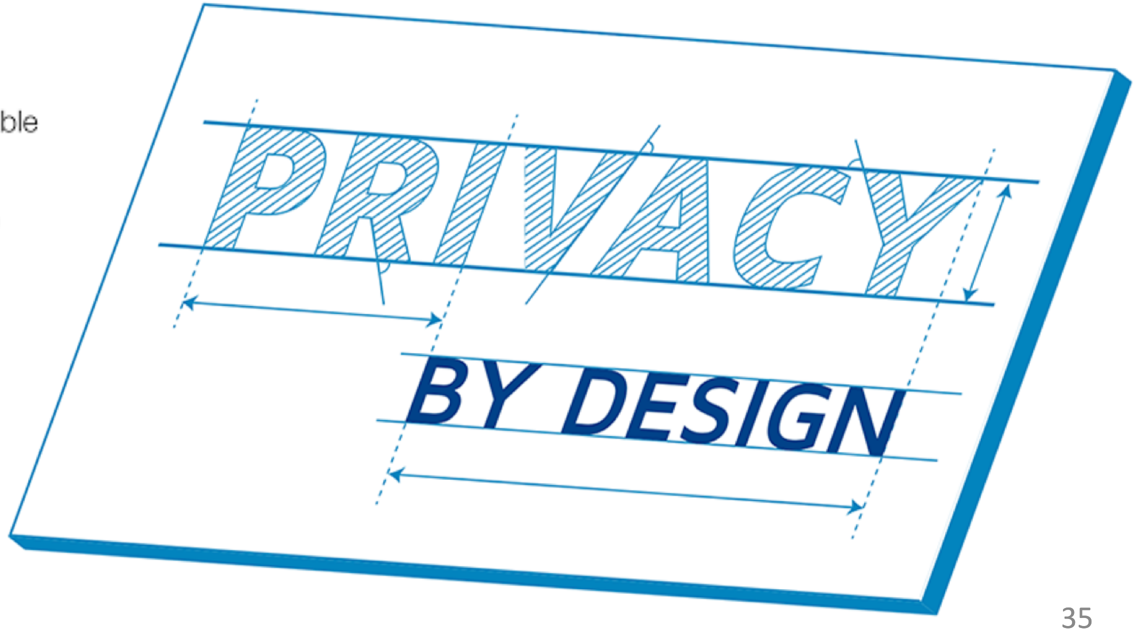
http://democracyos.eu/blog/open-by-design

OPEN DATA

https://wwwdb.inf.tu-dresden.de/opendatasurvey/

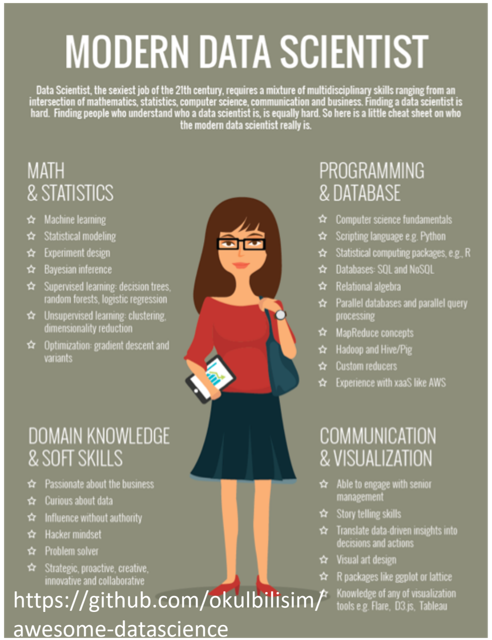


Wilkinson, et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*, http://dx.doi.org/10.1038/sdata.2016.18



https://privacybydesign.foundation/en/

Protect Data and Privacy



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

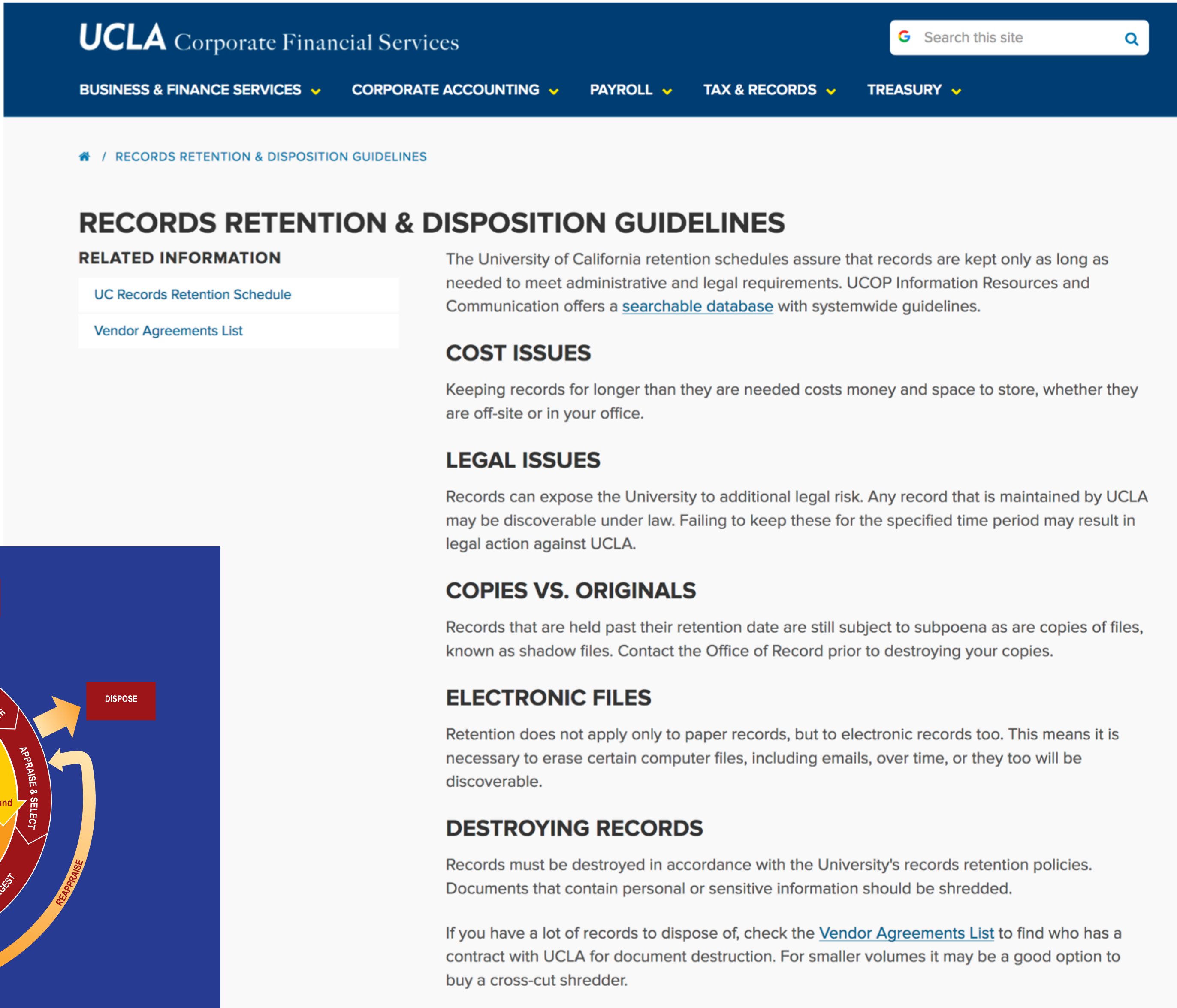
DOMAIN KNOWLEDGE
& SOFT SKILLS

- ☆ 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

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



UCLA Corporate Financial Services
Search this site
BUSINESS & FINANCE SERVICES
CORPORATE ACCOUNTING
PAYROLL
TAX & RECORDS
TREASURY
/ RECORDS RETENTION & DISPOSITION GUIDELINES

RECORDS RETENTION & DISPOSITION GUIDELINES

RELATED INFORMATION

- UC Records Retention Schedule
- Vendor Agreements List

The University of California retention schedules assure that records are kept only as long as needed to meet administrative and legal requirements. UCOP Information Resources and Communication offers a searchable database with systemwide guidelines.

COST ISSUES

Keeping records for longer than they are needed costs money and space to store, whether they are off-site or in your office.

LEGAL ISSUES

Records can expose the University to additional legal risk. Any record that is maintained by UCLA may be discoverable under law. Failing to keep these for the specified time period may result in legal action against UCLA.

COPIES VS. ORIGINALS

Records that are held past their retention date are still subject to subpoena as are copies of files, known as shadow files. Contact the Office of Record prior to destroying your copies.

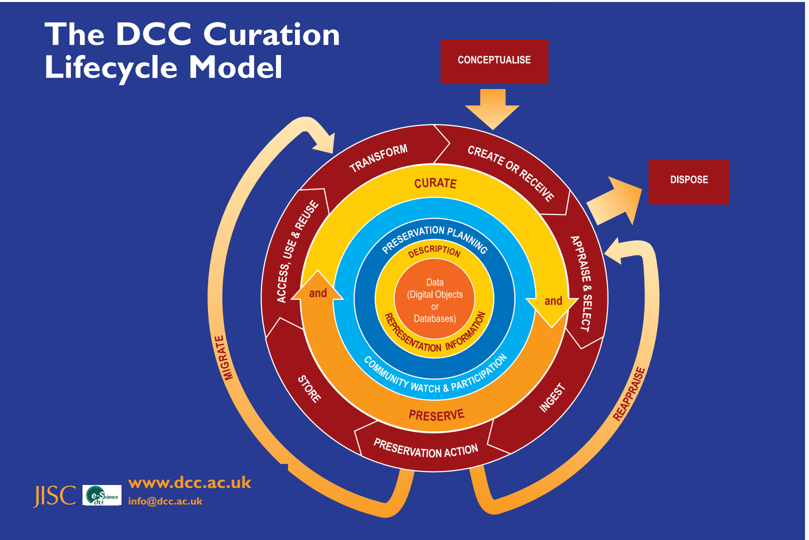
ELECTRONIC FILES

Retention does not apply only to paper records, but to electronic records too. This means it is necessary to erase certain computer files, including emails, over time, or they too will be discoverable.

DESTROYING RECORDS

Records must be destroyed in accordance with the University's records retention policies. Documents that contain personal or sensitive information should be shredded.

If you have a lot of records to dispose of, check the Vendor Agreements List to find who has a contract with UCLA for document destruction. For smaller volumes it may be a good option to buy a cross-cut shredder.



The DCC Curation Lifecycle Model

Promote Responsible Data Practices

- Respect information and autonomy privacy
  - Open data: release and reuse
  - Data collection and use
  - Data management
  - Collaborations
  - Publications
- Community
  - Faculty
  - Librarians
  - Staff
  - Students
  - External partners
- Joint governance process

https://www.universityofcalifornia.edu/subject/term/technology-engineering

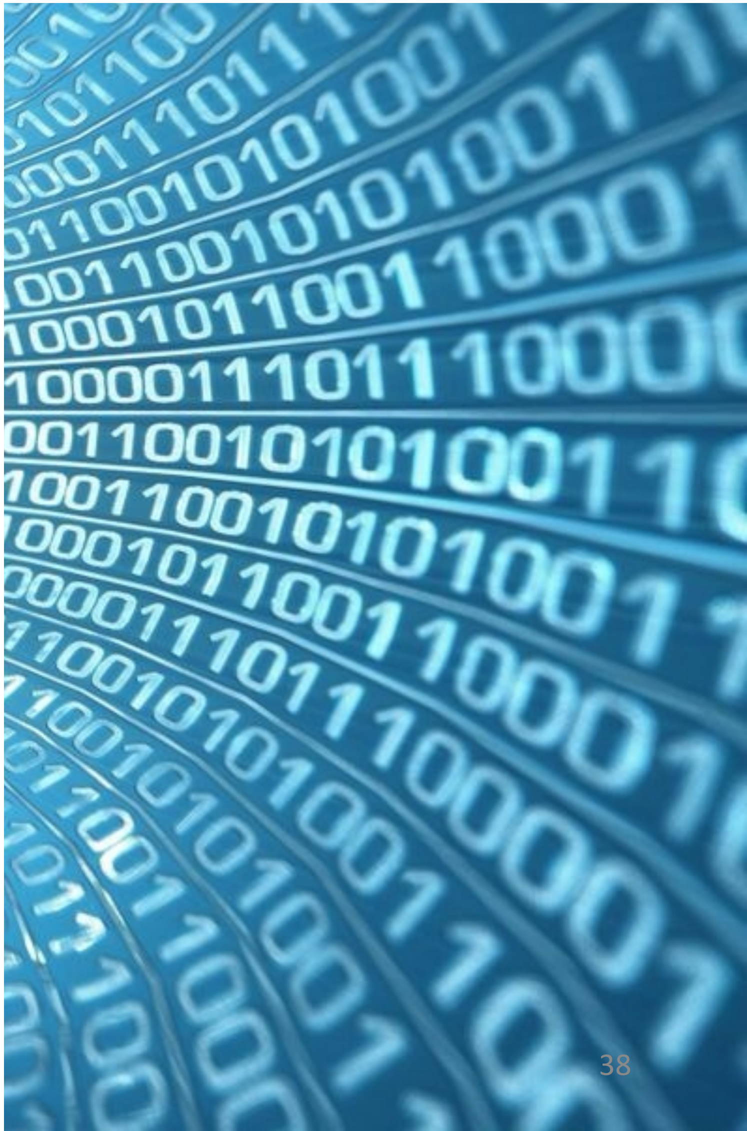
http://www.berkeley.edu/utility/jobs

http://gsa.rice.edu/

https://www.commondreams.org/views/2014/09/20/corporations-your-diet

Scholarship and Stewardship to Build the UC Digital Library

- Mission-drive stewardship
  - Research
  - Teaching
  - Services
- Steward the scholarly record
  - Integrated workflows
  - Version of record
  - Record of versions (Van de Sompel)
- Support discovery at scale
  - Human readable
  - Machine readable
  - Lawyer readable
- Sustain trust of community
  - Privacy: information, autonomy
  - Academic freedom
  - Stewardship and governance



UC Leadership in Data Policy

- We must maximally enable the **mission** of the University by supporting the values of **academic and intellectual freedom**.
- We must be **good stewards** of the **information entrusted** to the University.
- We must ensure that the University has **access to information** resources for **legitimate business purposes.**
- We must have a University community with **clear expectations of privacy**—both **privileges and obligations** of individuals and of the institution.
- We must make decisions within an **institutional context**.
- We must acknowledge the **distributed nature** of information stewardship at UC, where **responsibility for privacy and information security** resides at every level.

UNIVERSITY OF CALIFORNIA Academic Senate

UCOP Privacy and Information Security Initiative. (2013). http://ucop.edu/privacy-initiative/

Acknowledgements

UCLA Center for Knowledge Infrastructures

Christine Borgman

Peter Darch

Irene Pasquetto

Bernie Boscoe

Michael Scroggins

Milena Golshan

UCLA Center for Knowledge Infrastructures

ALFRED P. SLOAN FOUNDATION 1934