Title

Keynote: Big Data, Little Data, or No Data? Why Human Interaction with Data is a Hard Problem (slides)

Author

Borgman, Christine L.

Publication Date
2020-03-15

Big Data, Little Data, or No Data?

Why Human Interaction with Data is a Hard Problem

Christine L. Borgman

Distinguished Research Professor

Director, Center for Knowledge Infrastructures

https://knowledgeinfrastructures.gseis.ucla.edu

University of California, Los Angeles

http://christineborgman.info

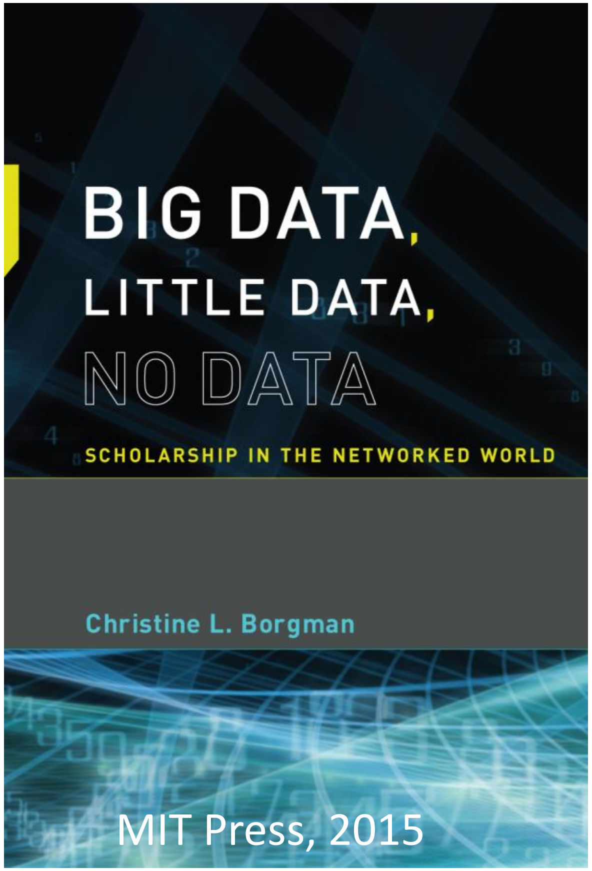
@scitechprof

Keynote Presentation, March 15, 2020 / August 13, 2020

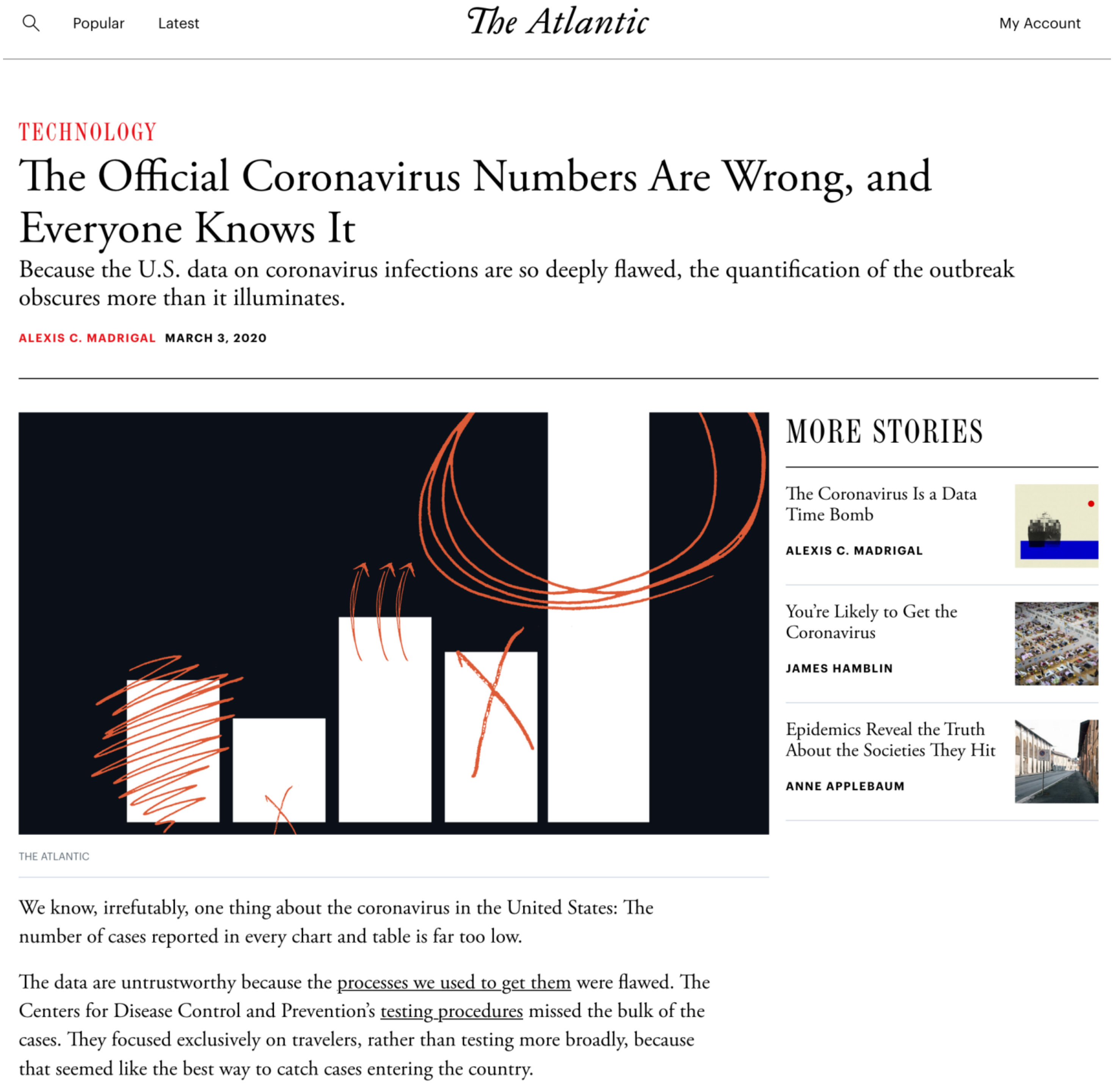
ACM SIGIR Conference on

Human Information Interaction and Retrieval

Vancouver, British Columbia, Canada / Zoom



MIT Press, 2015



Popular Latest **The Atlantic** My Account

TECHNOLOGY

The Official Coronavirus Numbers Are Wrong, and Everyone Knows It

Because the U.S. data on coronavirus infections are so deeply flawed, the quantification of the outbreak obscures more than it illuminates.

ALEXIS C. MADRIGAL MARCH 3, 2020

THE ATLANTIC

MORE STORIES

The Coronavirus Is a Data Time Bomb
ALEXIS C. MADRIGAL

You're Likely to Get the Coronavirus
JAMES HAMBLIN

Epidemics Reveal the Truth About the Societies They Hit
ANNE APPLEBAUM

We know, irrefutably, one thing about the coronavirus in the United States: The number of cases reported in every chart and table is far too low.

The data are untrustworthy because the processes we used to get them were flawed. The Centers for Disease Control and Prevention's testing procedures missed the bulk of the cases. They focused exclusively on travelers, rather than testing more broadly, because that seemed like the best way to catch cases entering the country.

What are data?

THE WHITE HOUSE
WASHINGTON

Data sharing policies

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

JST JAPAN SCIENCE AND TECHNOLOGY AGENCY

中央研究院
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
wellcome trust

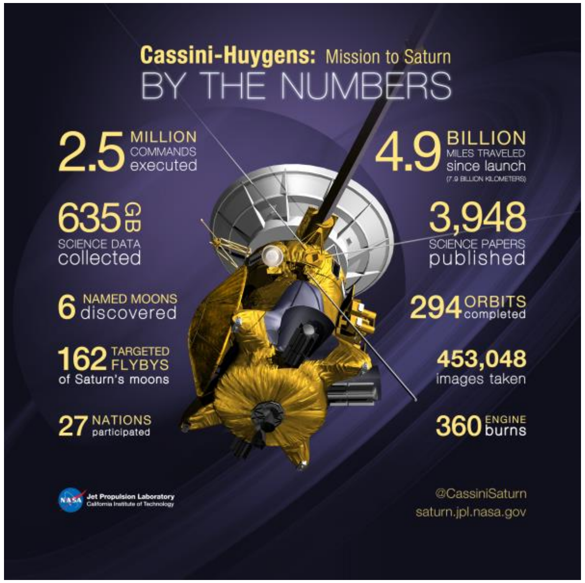
Australian Government
National Health and Medical Research Council

NSF National Science Foundation
WHERE DISCOVERIES BEGIN

Policy RECommendations for Open Access to Research Data in Europe

RECODE

NIH National Institutes of Health
Turning Discovery Into Health


Cassini-Huygens: Mission to Saturn
BY THE NUMBERS
2.5 MILLION COMMANDS executed
4.9 BILLION MILES TRAVELED since launch (7.9 BILLION KILOMETERS)
635 GB SCIENCE DATA collected
3,948 SCIENCE PAPERS published
6 NAMED MOONS discovered
294 ORBITS completed
162 TARGETED FLYBYS of Saturn's moons
453,048 images taken
27 NATIONS participated
360 ENGINE burns
Jet Propulsion Laboratory California Institute of Technology
@CassiniSaturn
saturn.jpl.nasa.gov

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

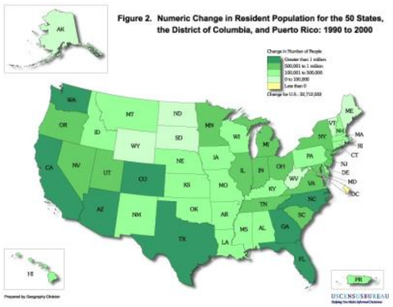
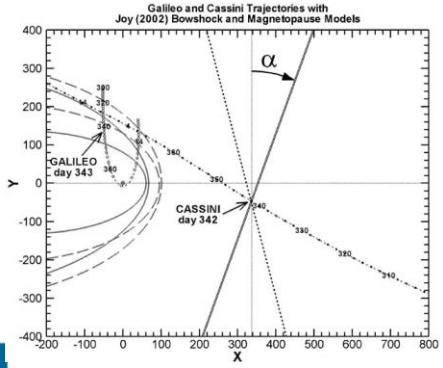
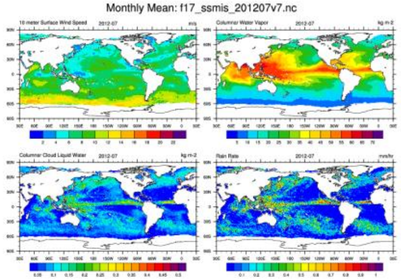


Figure 2. Numeric Change in Resident Population for the 50 States, the District of Columbia, and Puerto Rico: 1990 to 2000

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



Kivelson, M. G., & Southwood, D. J. (2003). First evidence of IMF control of Jovian magnetospheric boundary locations: Cassini and Galileo magnetic field measurements compared. *Planetary and Space Science*, *51*(13), 891–898. https://doi.org/10.1016/S0032-0633(03)00075-8



ncl.ucar.edu

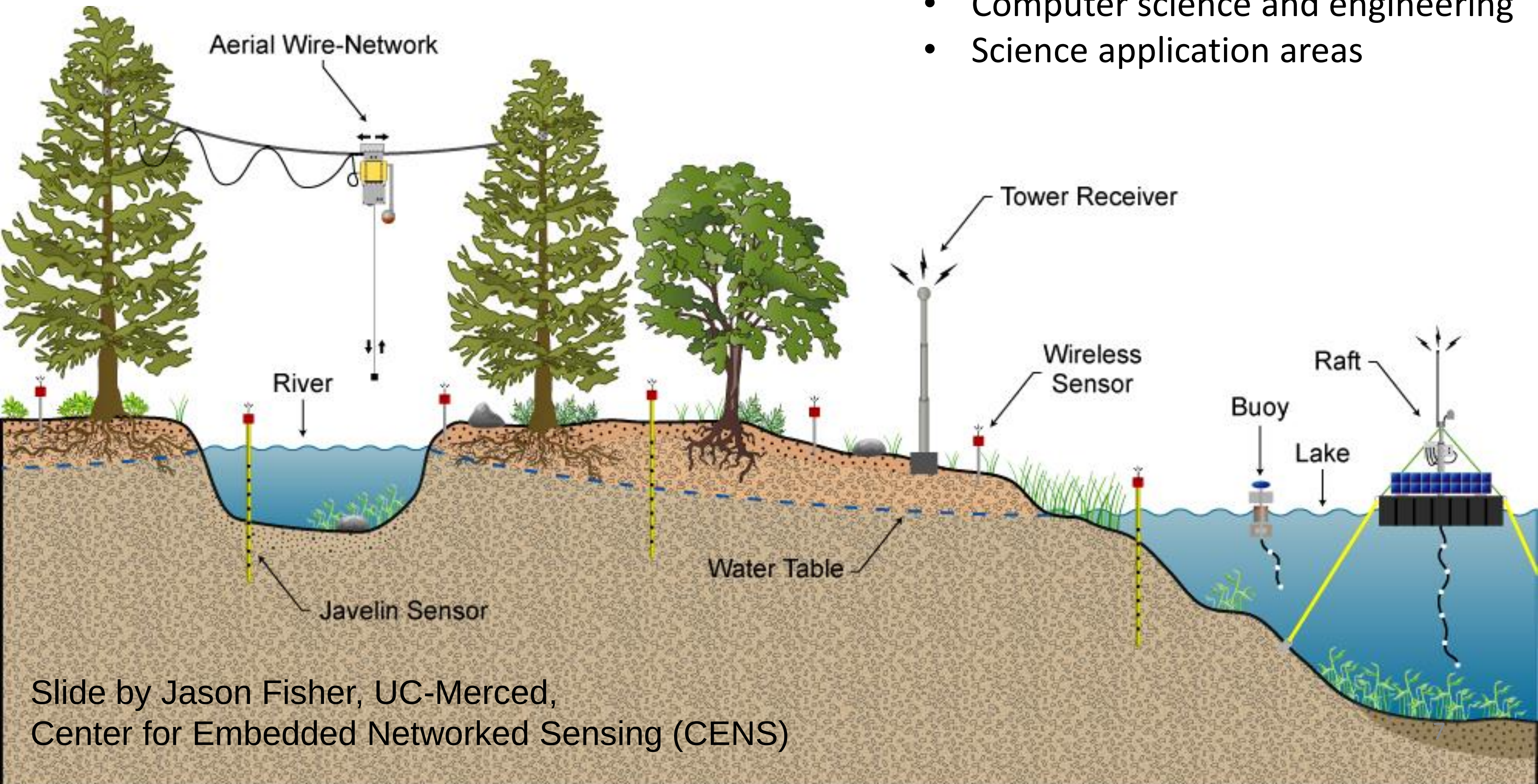
http://www.genome.gov/dmd/img.cfm?node=Photos/Graphics&id=85327

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

How to interpret data?

Center for Embedded Networked Sensing

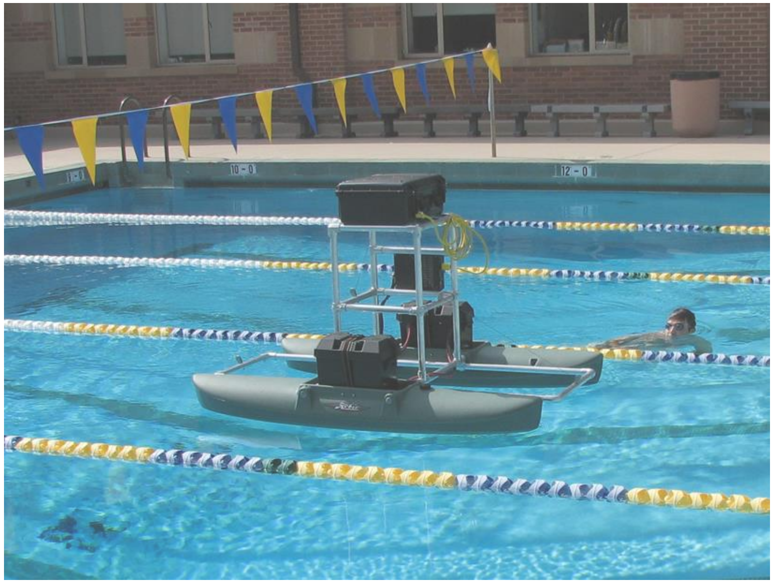
- NSF Science & Tech Ctr, 2002-2012
- 5 universities, plus partners
- 300 members
- Computer science and engineering
- Science application areas



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

Science <–> Data

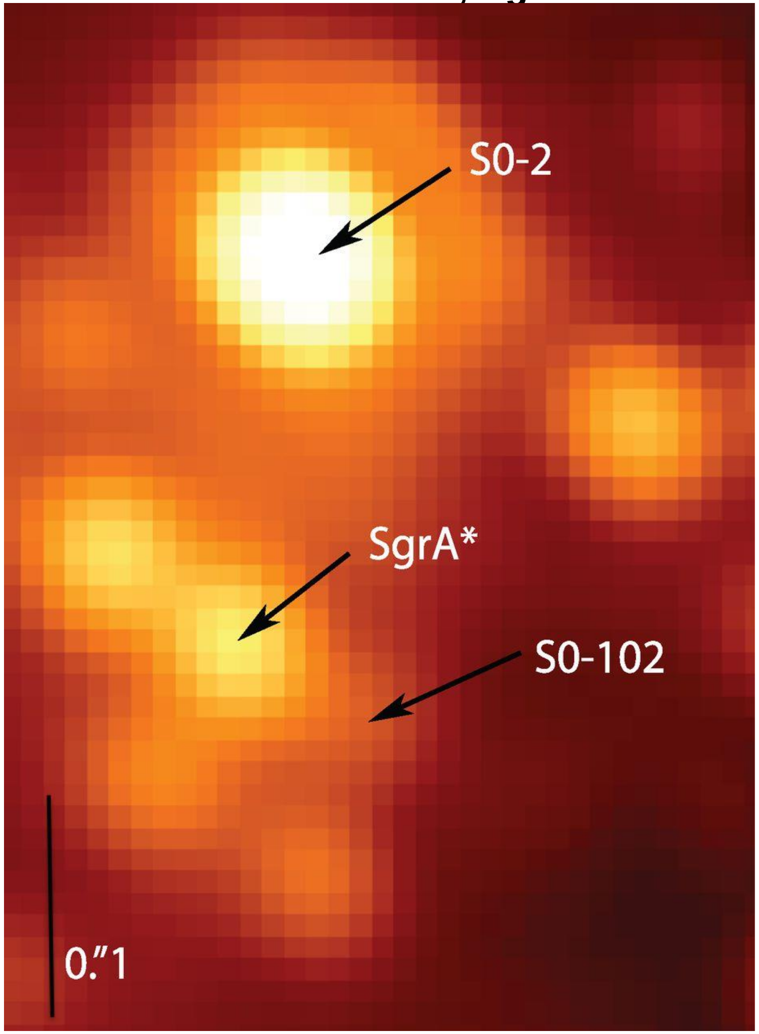
Engineering researcher:
“Temperature is temperature.”



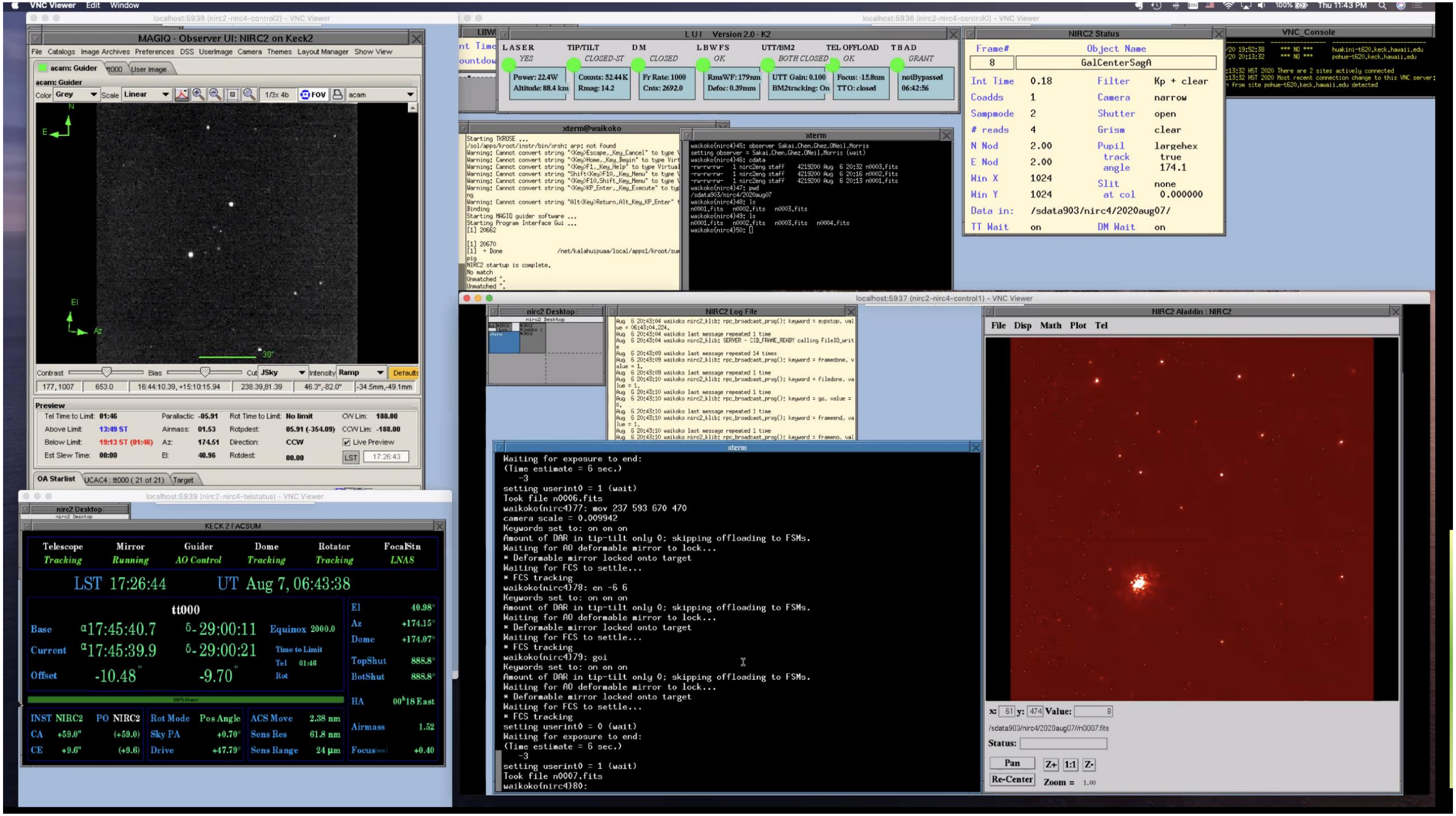
CENS Robotics team

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.."*

Fig. 1 A Keck/NIRC2 AO image from May 2010 showing the short-period star S0-102, which is, besides S0-2, the only star with full orbital phase coverage, and the electromagnetic counterpart of the black hole, Sgr A*.



L. Meyer et al. Science 2012;338:84-87


VNC Viewer Edit Window
100% Thu 11:43 PM
localhost:5938 (nirc2-nirc4-control2) - VNC Viewer
MAGIQ - Observer UI: NIRC2 on Keck2
File Catalogs Image Archives Preferences DSS UserImage Camera Themes Layout Manager Show View
acam: Guider tt000 User Image
acam: Guider
N
E
El
Az
30"
Defaults
OA Starlist UCAC4 : tt000 (21 of 21) Target
localhost:5936 (nirc2-nirc4-control0) - VNC Viewer
LBW
nt Time
ountdow
LUI Version 2.0 - K2
xterm@waikoko
xterm
NIRC2 Status
VNC_Console
localhost:5937 (nirc2-nirc4-control1) - VNC Viewer
nirc2 Desktop
NIRC2 Log File
NIRC2 Aladdin : NIRC2
File Disp Math Plot Tel
localhost:5939 (nirc2-nirc4-telstatus) - VNC Viewer
KECK 2 FACSUM
Pan
Re-Center
Z+ 1:1 Z-

Color: Grey Scale: Linear 1/3x 4b FOV acam

Contrast Bias Cut: JSky Intensity: Ramp

177, 1007 | 653.0 | 16:44:10.39, +15:10:15.94 | 238.39,81.39 | 46.3",-82.0" | -34.5mm,-49.1mm

Preview

Tel Time to Limit:	01:46	Parallactic	-05.91	Rot Time to Limit:	No limit	CW Lim:	188.00
Above Limit:	13:49 ST	Airmass:	01.53	Rotpdest:	05.91 (-354.09)	CCW Lim:	-188.00
Below Limit:	19:13 ST (01:46)	Az:	174.51	Direction:	CCW	☑ Live Preview	
Est Slew Time:	00:00	El:	40.96	Rotdest:	00.00	LST	17:26:43

LASER	TIP/TILT	DM	LBWFS	UTT/BM2	TEL OFFLOAD	TBAD
YES	CLOSED-ST	CLOSED	OK	BOTH CLOSED	OK	GRANT
Power: 22.4W Altitude: 88.4 km	Counts: 52.44K Rmag: 14.2	Fr Rate: 1000 Cnts: 2692.0	RmsWF: 179nm Defoc: 0.39mm	UTT Gain: 0.100 BM2tracking: On	Focus: -15.8um TTO: closed	notBypassed 06:42:56

```
Starting TKROSE ...
/sol/apps/kroot/instr/bin/xrsh: arp: not found
Warning: Cannot convert string "<Key>Escape,_Key_Cancel" to type \
Warning: Cannot convert string "<Key>Home,_Key_Begin" to type Virt
Warning: Cannot convert string "<Key>F1,_Key_Help" to type Virtual
Warning: Cannot convert string "Shift<Key>F10,_Key_Menu" to type \
Warning: Cannot convert string "<Key>F10,Shift_Key_Menu" to type \
Warning: Cannot convert string "<Key>KP_Enter,_Key_Execute" to typ
ng
Warning: Cannot convert string "Alt<Key>Return,Alt_Key_KP_Enter" t
Binding
Starting MAGIQ guider software ...
Starting Program Interface Gui ...
[1] 20662

[1] 20670
[1]  + Done                 /net/kalahuipuaa/local/apps1/kroot/sum
pig
NIRC2 startup is complete.
No match
Unmatched ".
Unmatched ".
```

```
waikoko{nirc4}45: observer Sakai,Chen,Ghez,ONeil,Morris
setting observer = Sakai,Chen,Ghez,ONeil,Morris (wait)
waikoko{nirc4}46: cdata
-rw-rw-rw-   1 nirc2eng staff    4219200 Aug  6 20:32 n0003.fits
-rw-rw-rw-   1 nirc2eng staff    4219200 Aug  6 20:16 n0002.fits
-rw-rw-rw-   1 nirc2eng staff    4219200 Aug  6 20:13 n0001.fits
waikoko{nirc4}47: pwd
/sdata903/nirc4/2020aug07
waikoko{nirc4}48: ls
n0001.fits   n0002.fits   n0003.fits
waikoko{nirc4}49: ls
n0001.fits   n0002.fits   n0003.fits   n0004.fits
waikoko{nirc4}50: []
```

Frame#	8	Object Name	GalCenterSagA
Int Time	0.18	Filter	Kp + clear
Coadds	1	Camera	narrow
Sampmode	2	Shutter	open
# reads	4	Grism	clear
N Nod	2.00	Pupil	largehex
E Nod	2.00	track angle	true 174.1
Win X	1024	Slit	none
Win Y	1024	at col	0.000000
Data in:	/sdata903/nirc4/2020aug07/		
TT Wait	on	DM Wait	on

```
/20 19:52:38     *** NO ***     huakini-t620.keck.hawaii.edu
/20 20:13:32     *** NO ***     pohue-t620.keck.hawaii.edu

:13:32 HST 2020 There are 2 sites actively connected
:13:32 HST 2020 Most recent connection change to this VNC server:
n from site pohue-t620.keck.hawaii.edu detected
```

```
Aug  6 20:43:04 waikoko nirc2_klib: rpc_broadcast_prog(): keyword = expstop, val
ue = 06:43:04.224.
Aug  6 20:43:04 waikoko last message repeated 1 time
Aug  6 20:43:04 waikoko nirc2_klib: SERVER - CID_FRAME_READY calling FileIO_writ
e
Aug  6 20:43:09 waikoko last message repeated 14 times
Aug  6 20:43:09 waikoko nirc2_klib: rpc_broadcast_prog(): keyword = framedone, v
alue = 1.
Aug  6 20:43:09 waikoko last message repeated 1 time
Aug  6 20:43:10 waikoko nirc2_klib: rpc_broadcast_prog(): keyword = filedone, va
lue = 1.
Aug  6 20:43:10 waikoko last message repeated 1 time
Aug  6 20:43:10 waikoko nirc2_klib: rpc_broadcast_prog(): keyword = go, value =
0.
Aug  6 20:43:10 waikoko last message repeated 1 time
Aug  6 20:43:10 waikoko nirc2_klib: rpc_broadcast_prog(): keyword = frameend, va
lue = 1.
Aug  6 20:43:10 waikoko last message repeated 1 time
Aug  6 20:43:10 waikoko nirc2_klib: rpc_broadcast_prog(): keyword = frameno, val
```

```
Waiting for exposure to end:
(Time estimate = 6 sec.)
   -3
setting userint0 = 1 (wait)
Took file n0006.fits
waikoko{nirc4}77: mov 237 593 670 470
camera scale = 0.009942
Keywords set to: on on on
Amount of DAR in tip-tilt only 0; skipping offloading to FSMs.
Waiting for AO deformable mirror to lock...
* Deformable mirror locked onto target
Waiting for FCS to settle...
* FCS tracking
waikoko{nirc4}78: en -6 6
Keywords set to: on on on
Amount of DAR in tip-tilt only 0; skipping offloading to FSMs.
Waiting for AO deformable mirror to lock...
* Deformable mirror locked onto target
Waiting for FCS to settle...
* FCS tracking
waikoko{nirc4}79: goi
Keywords set to: on on on
Amount of DAR in tip-tilt only 0; skipping offloading to FSMs.
Waiting for AO deformable mirror to lock...
* Deformable mirror locked onto target
Waiting for FCS to settle...
* FCS tracking
setting userint0 = 0 (wait)
Waiting for exposure to end:
(Time estimate = 6 sec.)
   -3
setting userint0 = 1 (wait)
Took file n0007.fits
waikoko{nirc4}80:
```

x: 61 y: 474 Value: 8

/sdata903/nirc4/2020aug07//n0007.fits

Status:

Zoom = 1.00

Telescope	Mirror	Guider	Dome	Rotator	FocalStn
Tracking	Running	AO Control	Tracking	Tracking	LNAS

LST 17:26:44 UT Aug 7, 06:43:38

tt000

Base	α17:45:40.7	δ-29:00:11	Equinox 2000.0
Current	α17:45:39.9	δ-29:00:21	Time to Limit
			Tel 01:46
Offset	-10.48"	-9.70"	Rot

El	40.98°
Az	+174.15°
Dome	+174.07°
TopShut	888.8°
BotShut	888.8°
HA	00h18 East
Airmass	1.52
Focus	+0.40

INST	NIRC2	PO	NIRC2	Rot Mode	Pos Angle	ACS Move	2.38 nm
CA	+59.0"		(+59.0)	Sky PA	+0.70°	Sens Res	61.8 nm
CE	+9.6"		(+9.6)	Drive	+47.79°	Sens Range	24 µm

Astronomers' user interface for taking observations at a major ground-based telescope (August 2020)

Publications vs data

Publications

Dégradation ER mRNA...
Estrogen Receptors in Human Nontarget Tissues: Biological and Clinical Implications
DANIEL R. CIOCCA AND LAURA M. VARGAS ROIG
1.3.2.1
Effects of Low Dose Oral Contraceptives on Very Low Density Lipoprotein Metabolism
Abstract
1225
The Estradiol-stimulated Lipoprotein Receptor of Rat Liver

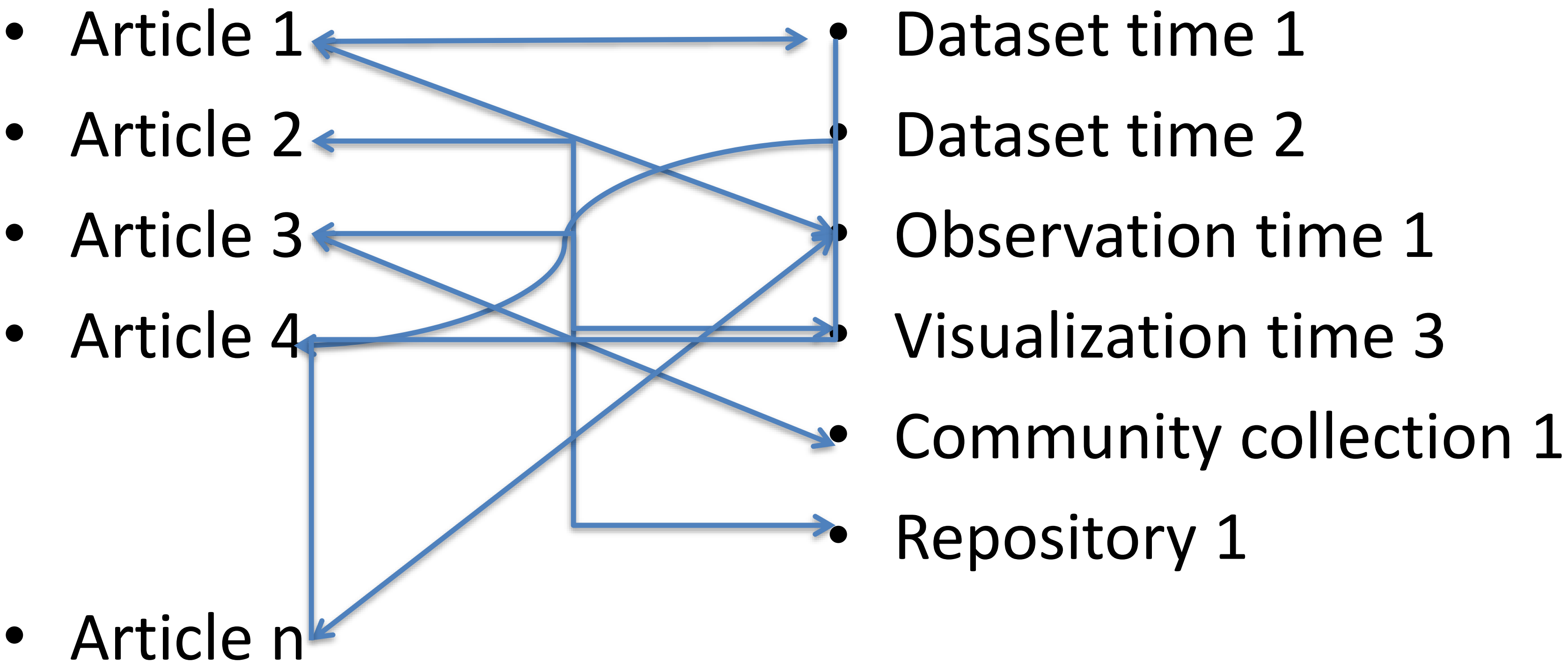
http://www.cse.psu.edu/hpcl/images/publications.jpg

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

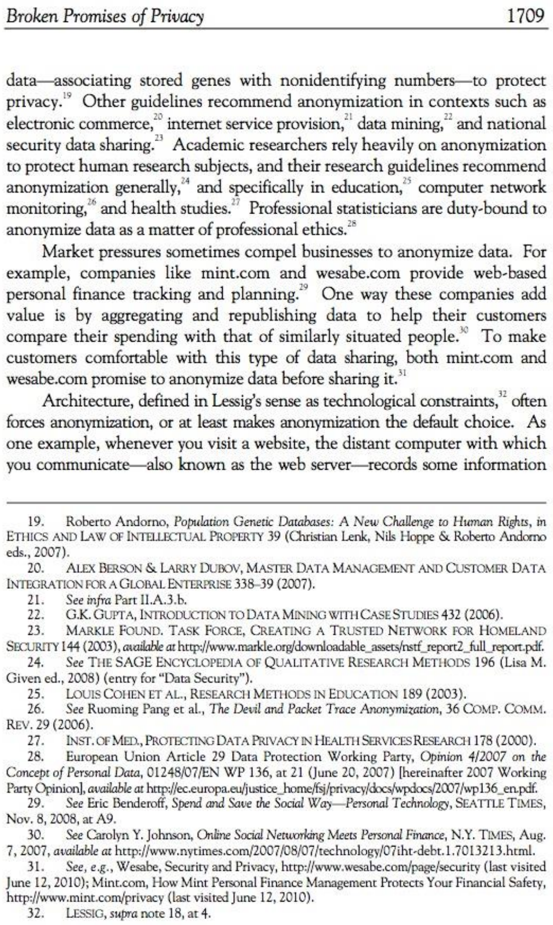
- Dataset time 1
- Dataset time 2
- Observation time 1
- Visualization time 3
- Community collection 1
- Repository 1

Why cite data?

- Credit
- Attribution
- Discovery

Credit

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.

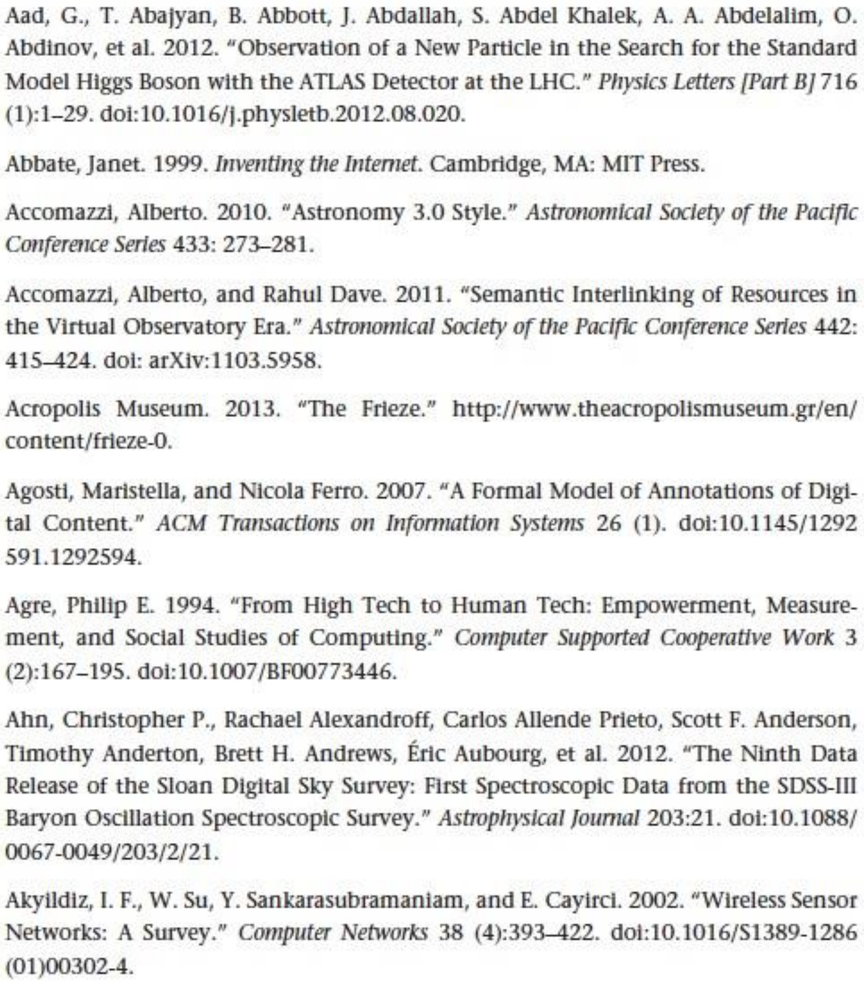
[29] *See* Eric Benderoff, *Spend and Save the Social Way—Personal Technology*, SEATTLE TIMES, Nov. 8, 2008, at A9.

[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.

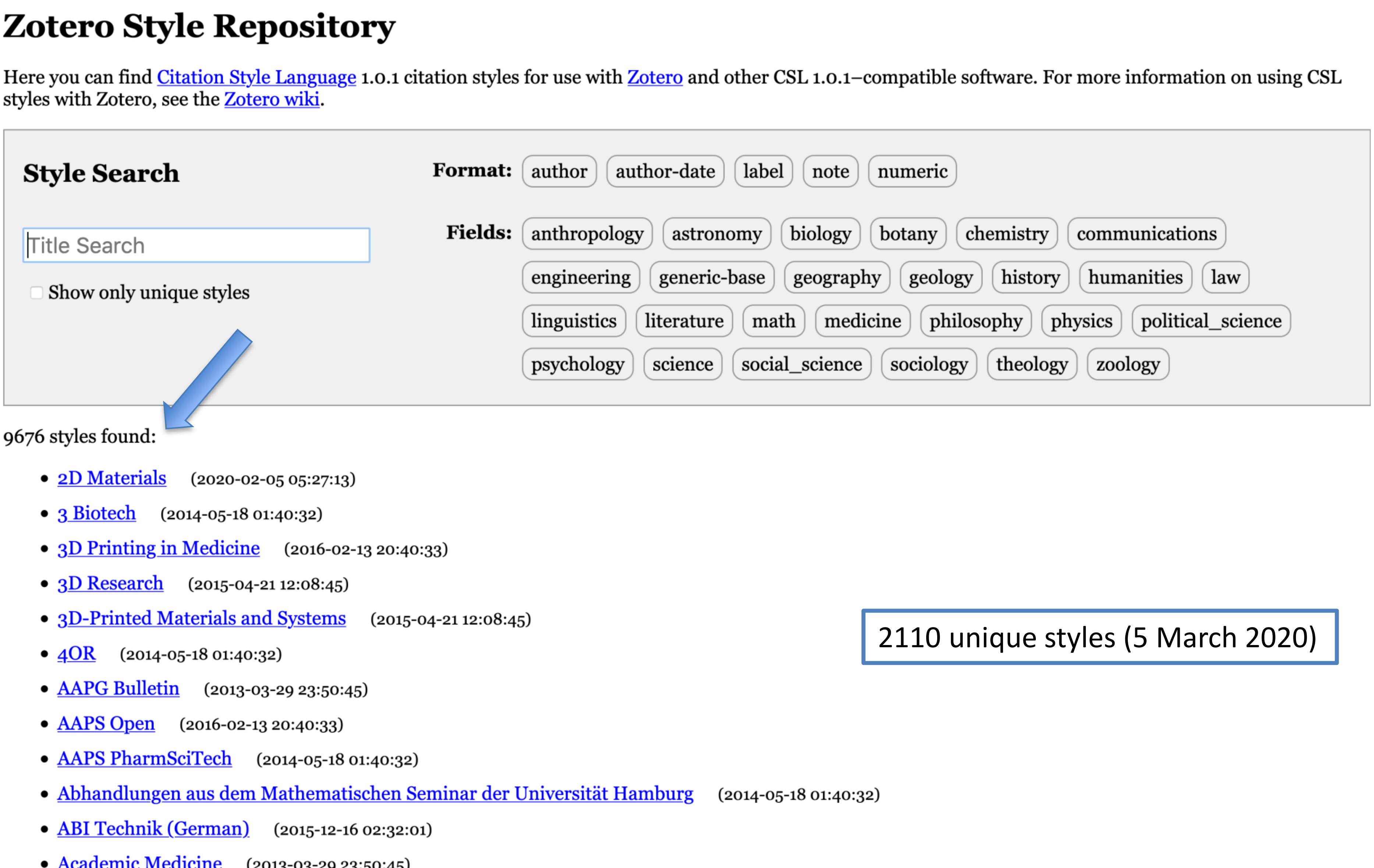
Agre, Philip E. 1994. "From High Tech to Human Tech: Empowerment, Measurement, and Social Studies of Computing." *Computer Supported Cooperative Work* 3 (2):167–195. doi:10.1007/BF00773446.

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

9676 styles found:

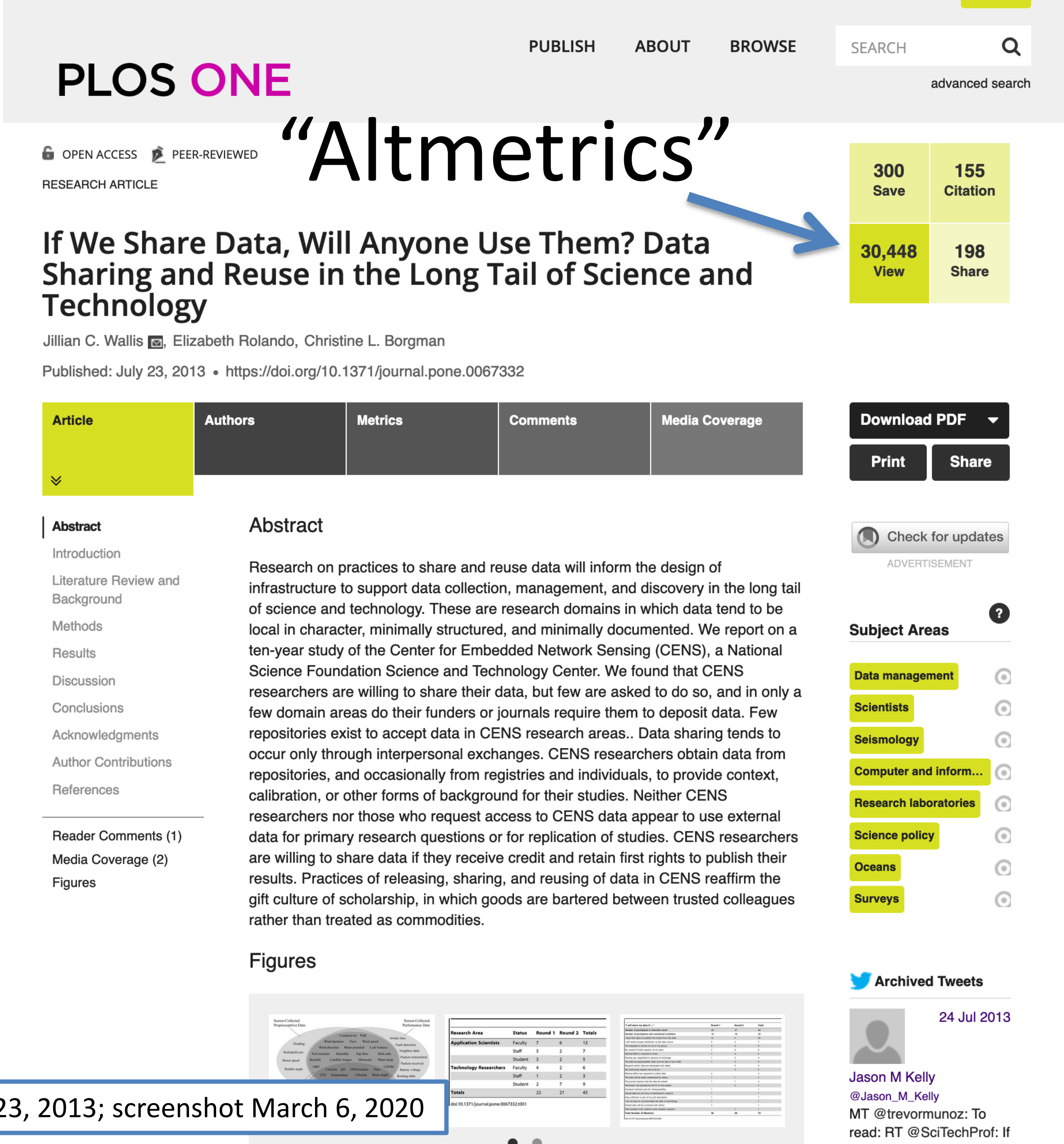
- 2D Materials (2020-02-05 05:27:13)
- 3 Biotech (2014-05-18 01:40:32)
- 3D Printing in Medicine (2016-02-13 20:40:33)
- 3D Research (2015-04-21 12:08:45)
- 3D-Printed Materials and Systems (2015-04-21 12:08:45)
- 4OR (2014-05-18 01:40:32)
- AAPG Bulletin (2013-03-29 23:50:45)
- AAPS Open (2016-02-13 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)
- Academic Medicine (2013-03-29 23:50:45)

2110 unique styles (5 March 2020)

Authorship Credit

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

"Altmetrics"



PLOS ONE

PUBLISH ABOUT BROWSE

SEARCH

advanced search

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

300 Save	155 Citation
30,448 View	198 Share

Article | Authors | Metrics | Comments | Media Coverage

Download PDF | Print | Share

- Abstract
- Introduction
- Literature Review and Background
- Methods
- Results
- Discussion
- Conclusions
- Acknowledgments
- Author Contributions
- References

- Reader Comments (1)
- Media Coverage (2)
- Figures

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 interpersonal exchanges. CENS researchers obtain data from repositories, and occasionally from registries and individuals, to provide context, calibration, or other forms of background for their studies. Neither CENS researchers nor those who request access to CENS data appear to use external data for primary research questions or for replication of studies. CENS researchers are willing to share data if they receive credit and retain first rights to publish their results. Practices of releasing, sharing, and reusing of data in CENS reaffirm the gift culture of scholarship, in which goods are bartered between trusted colleagues rather than treated as commodities.

Figures

Check for updates

ADVERTISEMENT

Subject Areas

- Data management
- Scientists
- Seismology
- Computer and inform...
- Research laboratories
- Science policy
- Oceans
- Surveys

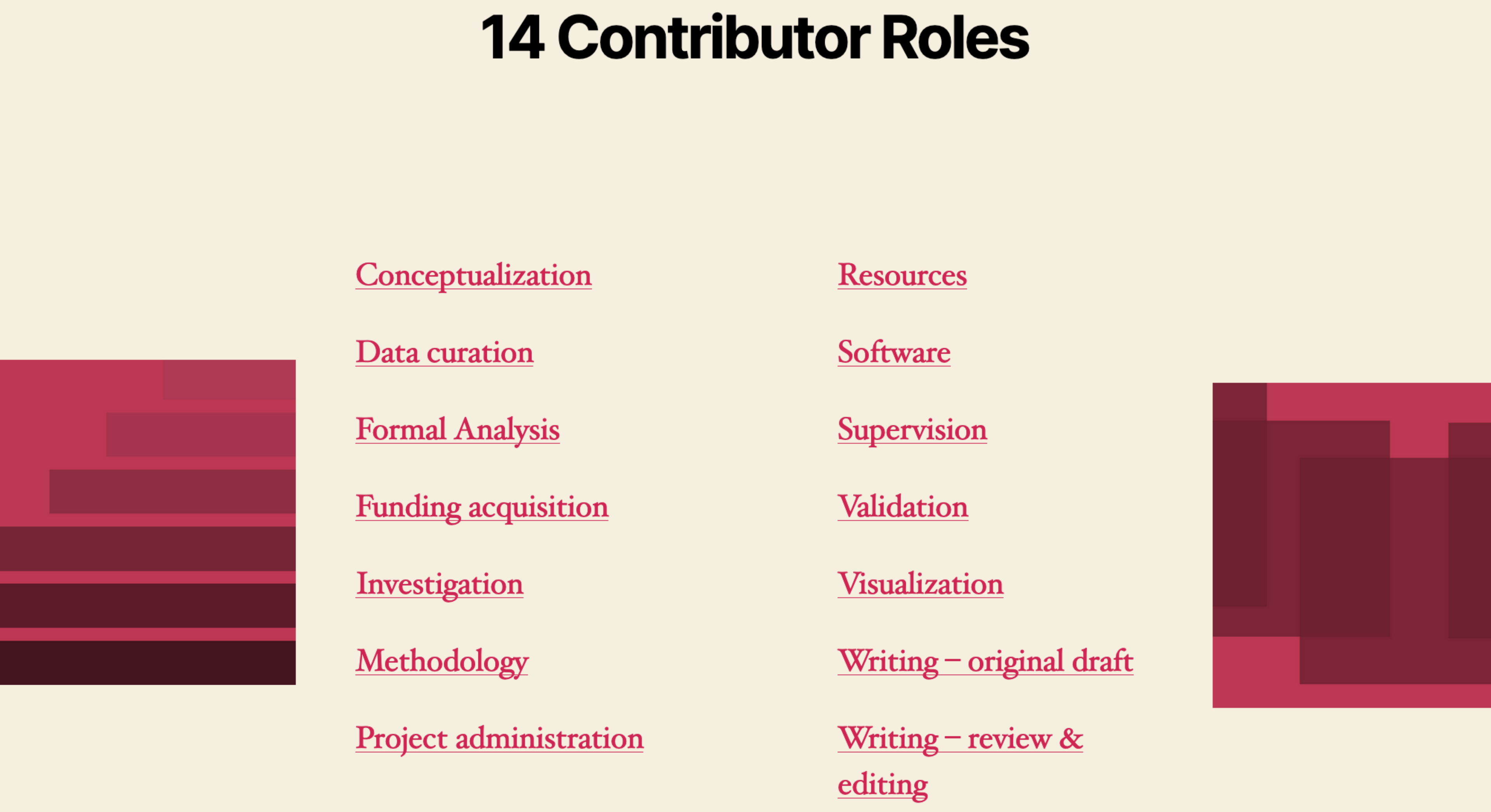
Archived Tweets

24 Jul 2013

Jason M Kelly
@Jason_M_Kelly
MT @trevormunoz: To read: RT @SciTechProf: If

Published July 23, 2013; screenshot March 6, 2020

Attribution



14 Contributor Roles

- Conceptualization
- Data curation
- Formal Analysis
- Funding acquisition
- Investigation
- Methodology
- Project administration
- Resources
- Software
- Supervision
- Validation
- Visualization
- Writing – original draft
- Writing – review & editing

CRediT – Contributor Roles Taxonomy. (2020). http://credit.niso.org/

Publications <–> Data: Attribution

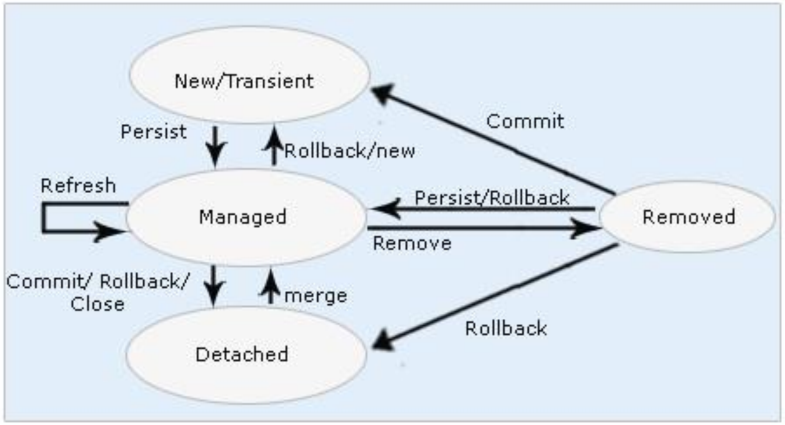
- Publications
  - – Independent units
  - – Authorship is negotiated
- Data
  - – Compound objects
  - – Ownership is rarely clear
  - – Attribution
    - Long term responsibility: Investigators
    - Expertise for interpretation: Data collectors and analysts

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

Discovery

Metadata for Discovery

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



Persistence Content

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

Discovering Useful Data

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

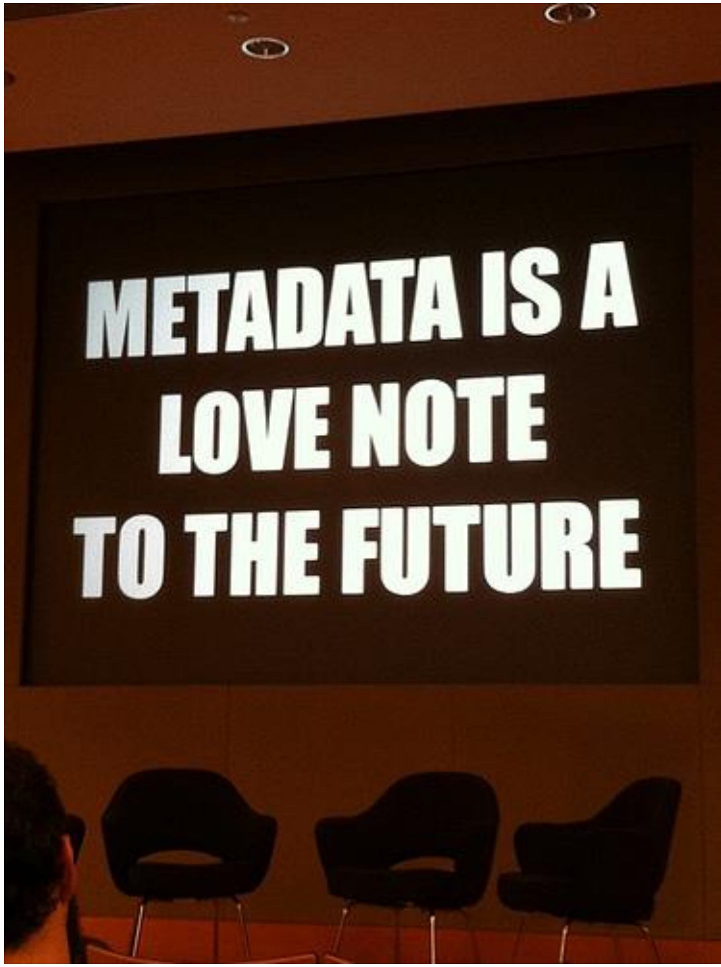
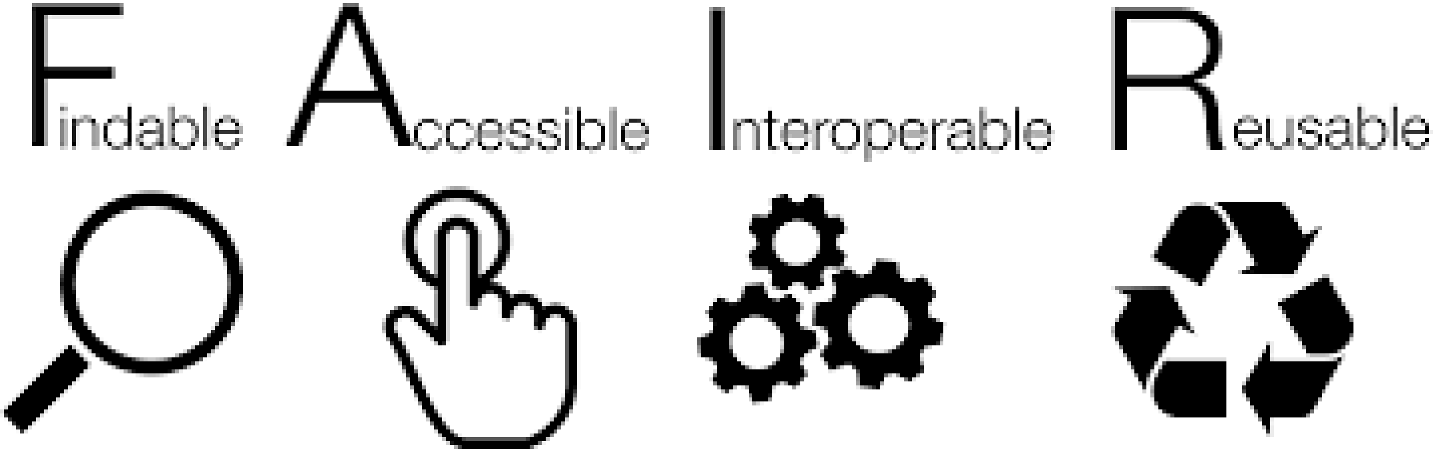


Photo by @kissane; presentation by Jason Scott (@textfiles)

Stewardship, Incentives, and Scientific Practice

Data Stewardship: The Ideal



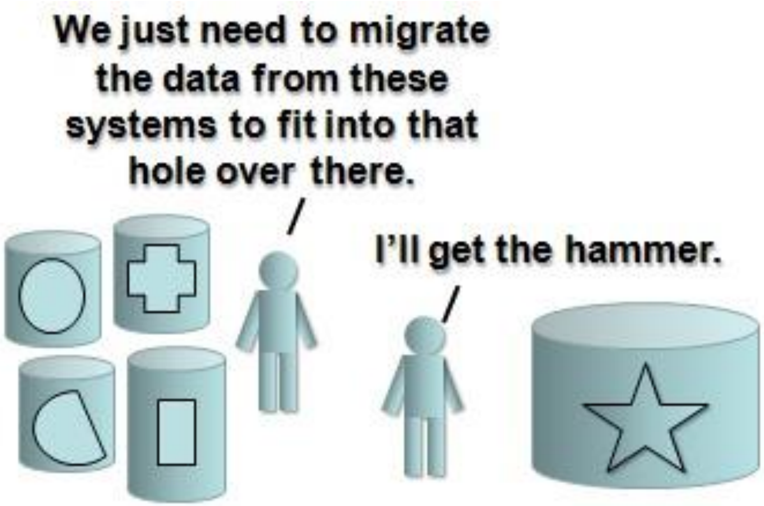
Wilkinson, et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*, http://dx.doi.org/10.1038/sdata.2016.18

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

Lack of incentives to share data

- Labor to document data
- Benefits to unknown others
- Competition
- Control
- Confidentiality
- Lack of expertise and staff
- Lack of sustainability…

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

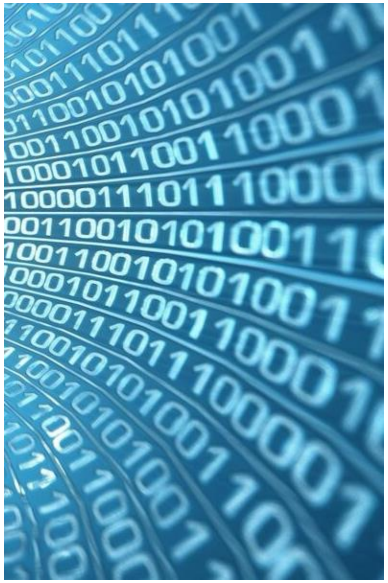
The Data Creators’ Advantage

	Comparative Data Reuse <–> Integrative Data Reuse	
Goal	“Ground truthing:” calibrate, compare, confirm	Analysis: identify patterns, correlations, causal relationships
Example	Instrument calibration, sequence annotation, review summary-level data	Meta-analyses, novel statistical analyses
Frequency	Frequent, routine practice	Rare, emergent practice
Interpretation	Interactional expertise, ‘knowledge that’	Contributory expertise, ‘knowledge how,’ tacit knowledge

Pasquetto, I. V., Borgman, C. L., & Wofford, M. F. (2019). Uses and reuses of scientific data: The data creators’ advantage. *Harvard Data Science Review,* 1:2, https://hdsr.mitpress.mit.edu/. **Winner** of the **2020 ASIS&T SIG SI Social Informatics Best Paper Award**

Why Human Interaction with Data is a Hard Problem

- Data exist in contexts
- Data are complex objects
  - Signals, observations
  - Software, tools, methods, models
  - Digital records, physical objects
- Data management is undervalued
- Data creators have interpretive advantages



UCLA Center for Knowledge Infrastructures

Alberto Pepe, David Fearon, Katie Shilton, Jillian Wallis, Christine Borgman, Matthew Mayernik (2009)

Christine Borgman

Peter Darch

Ashley Sands

Irene Pasquetto

Milena Golshan

Bernie Boscoe

Cheryl Thompson

Morgan Wofford

Michael Scroggins

Sharon Traweek

For a full list of CKI participants, collaborators, and coauthors since ca 2002, see https://knowledgeinfrastructures.gseis.ucla.edu/