Title
Big Data, Little Data, noData: The Contested Landscape of Data Sharing and Reuse

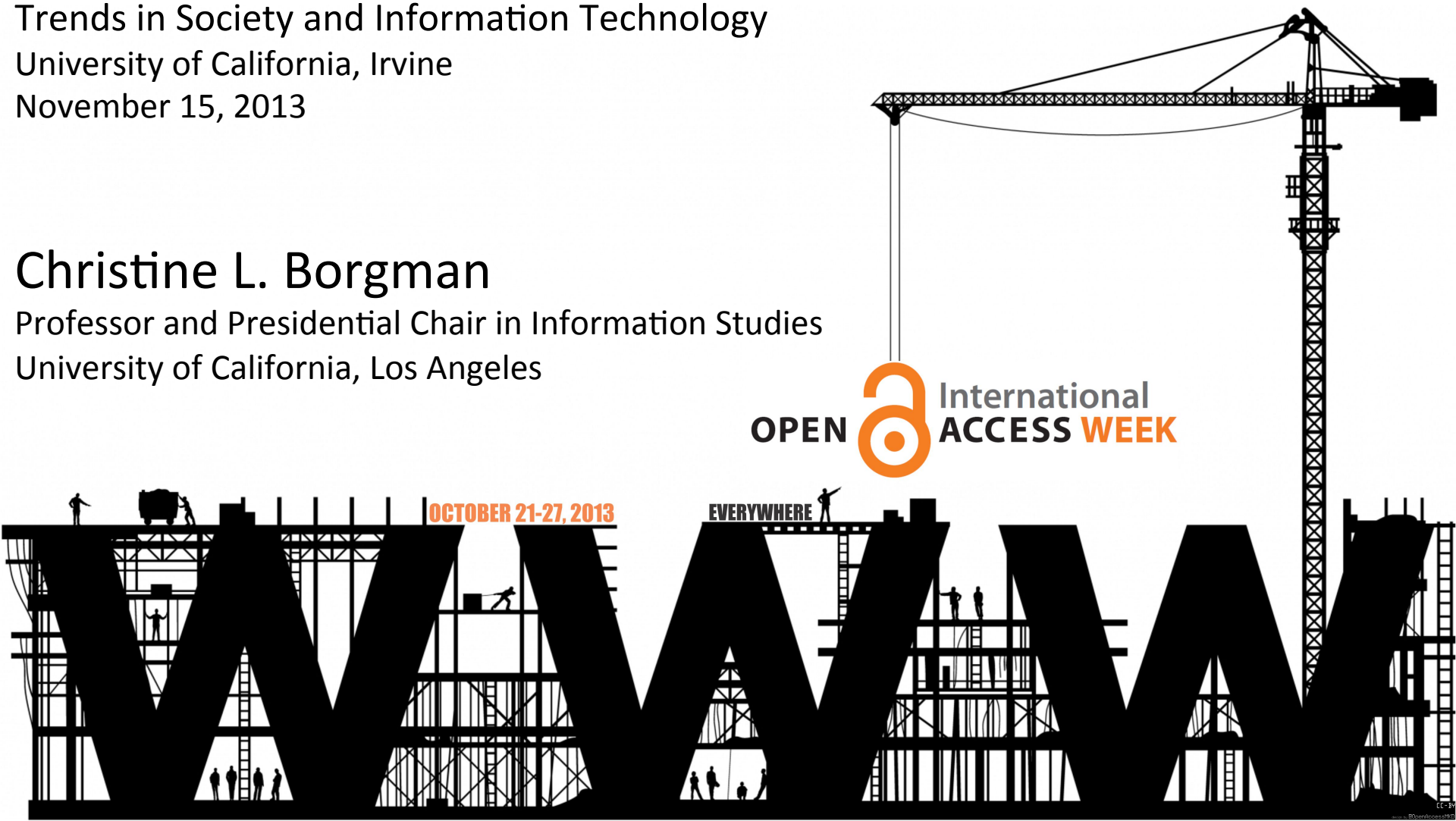
Author
Borgman, Christine L.

Publication Date
2013-11-15

Big Data, Little Data, No Data:
The Contested Landscape of Data Sharing and Reuse

Trends in Society and Information Technology
University of California, Irvine
November 15, 2013

Christine L. Borgman
Professor and Presidential Chair in Information Studies
University of California, Los Angeles




Science
12 September 2008 | $10
AAAS


Science
11 February 2011 | $10
scientific
data
research
information
climate
science
new
analysis
visualization
researchers
access
many
human
knowledge
example
AAAS

Data sharing imperatives

- National Science Foundation
  - Data sharing requirements
  - Data management plans
- U.S. Federal policy-2013
  - Open access to publications
  - Open access to data
- European Union
  - European Open Data Challenge
  - Policy Recommendations for Open Access to Research Data in Europe
  - Riding the wave: How Europe can gain from the rising tide of scientific data
  - OpenAIRE
- Research Councils of the UK
  - Open access publishing requirements
  - Provisions for access to data
- Wellcome Trust
  - Open access publishing
  - Data sharing requirements

Supported by wellcome trust

Overview

- Paradigm shift
- Arguments for sharing data
- Science friction, data friction
- Sharing and reusing data

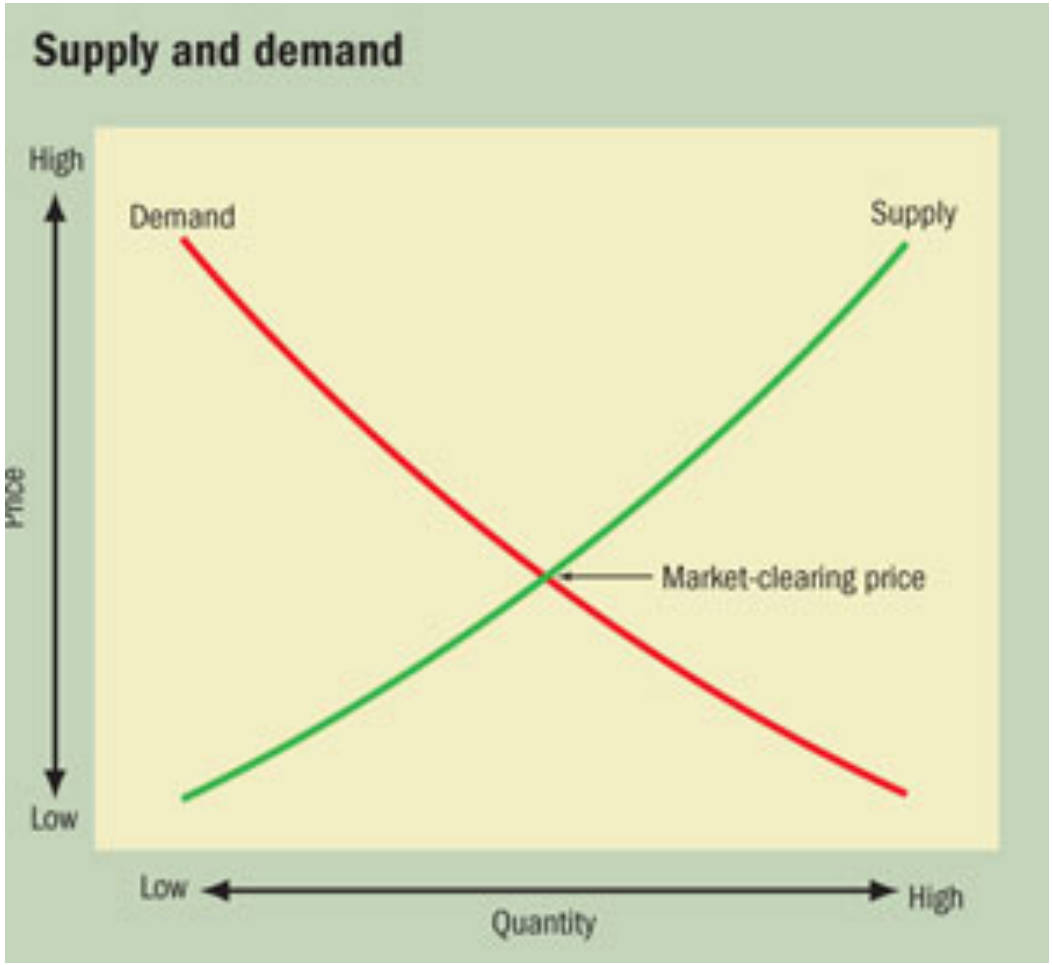
http://inventionmachine.com/the-Invention-Machine-Blog/bid/51703/Three-Key-Challenges-to-Entering-New-Markets

The Conundrum of Sharing Research Data

*If the rewards of the data deluge are to be reaped, then researchers who produce those data must share them, and do so in such a way that the data are interpretable and reusable by others.**

*Borgman, C.L. (2012). The Conundrum of Sharing Research Data. *Journal of the American Society of Information Science and Technology*, 63(6):1059–1078

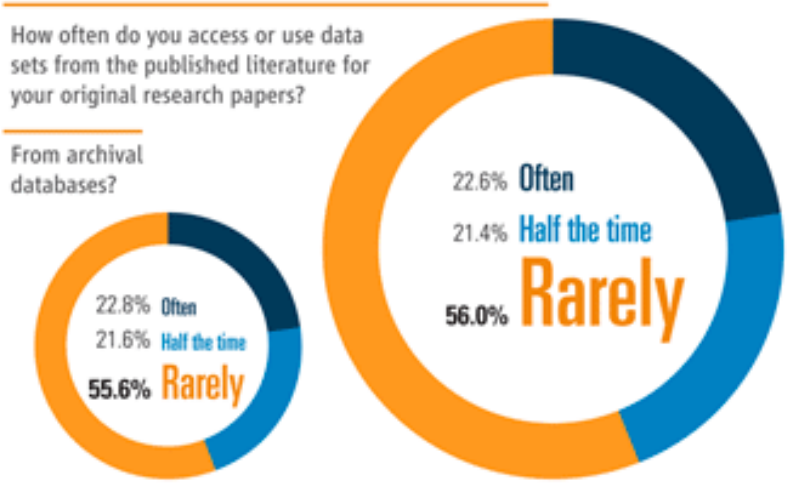
http://www.tzanis.org/tzanisblog/archives/images/push-pull-thumb.jpg



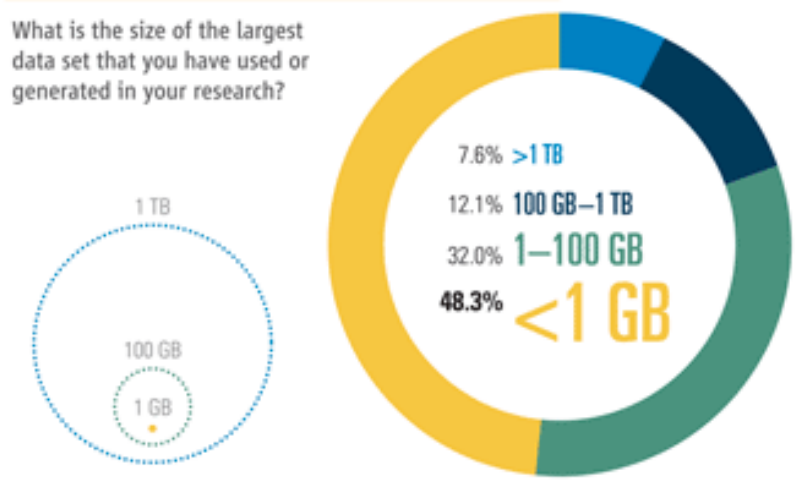
http://www.imf.org/external/pubs/ft/fandd/basics/images/suppdem_cht.jpg

Survey with 1700 respondents from Science (2011) peer reviewers

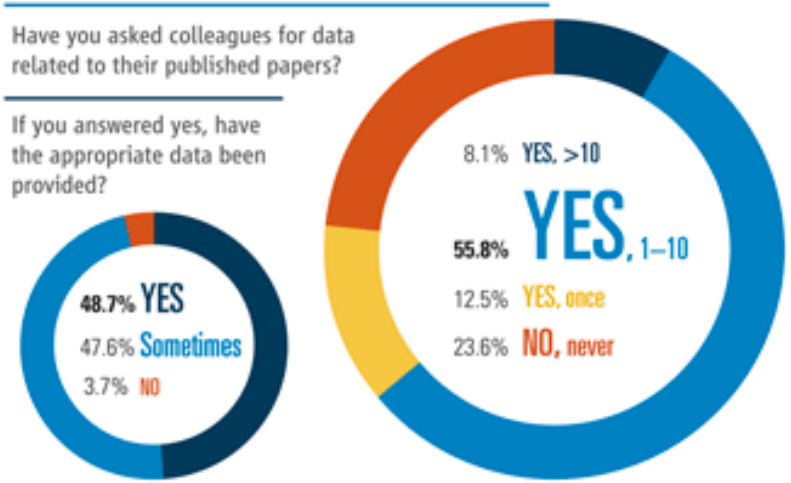
Access data from published work



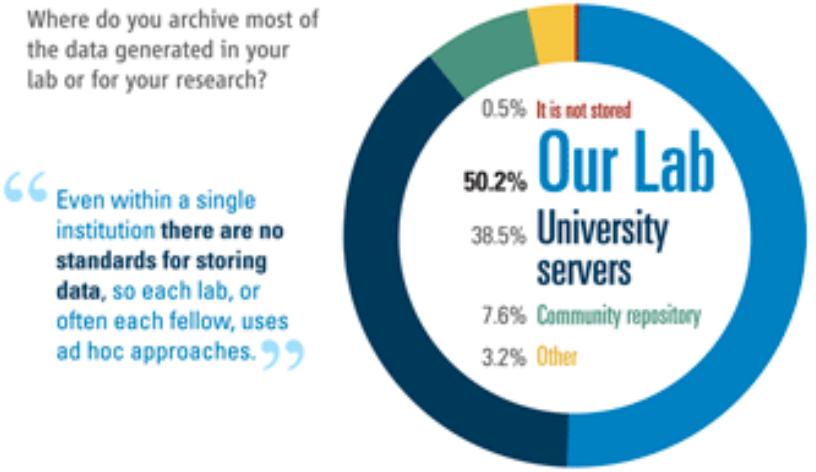
Size of data



Ask colleagues for data



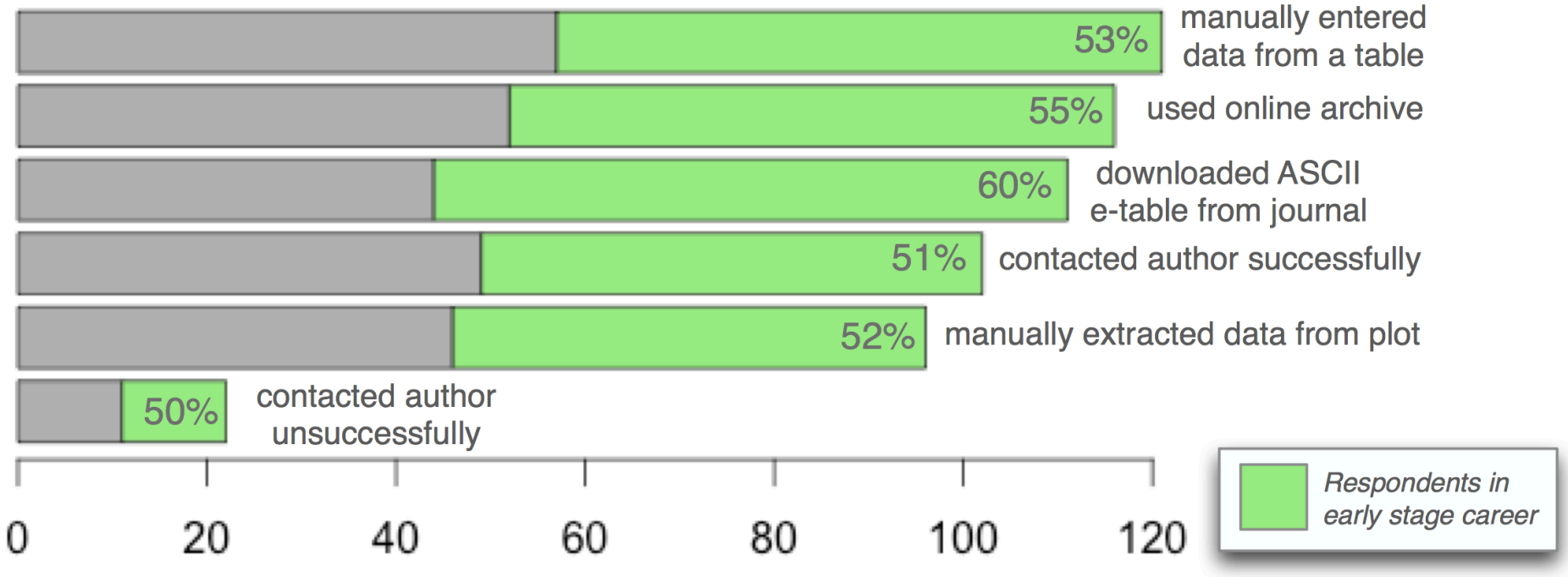
Archival location



Slide courtesy of Merce Crosas, Harvard IQSS

In Astronomy, a field with data standards

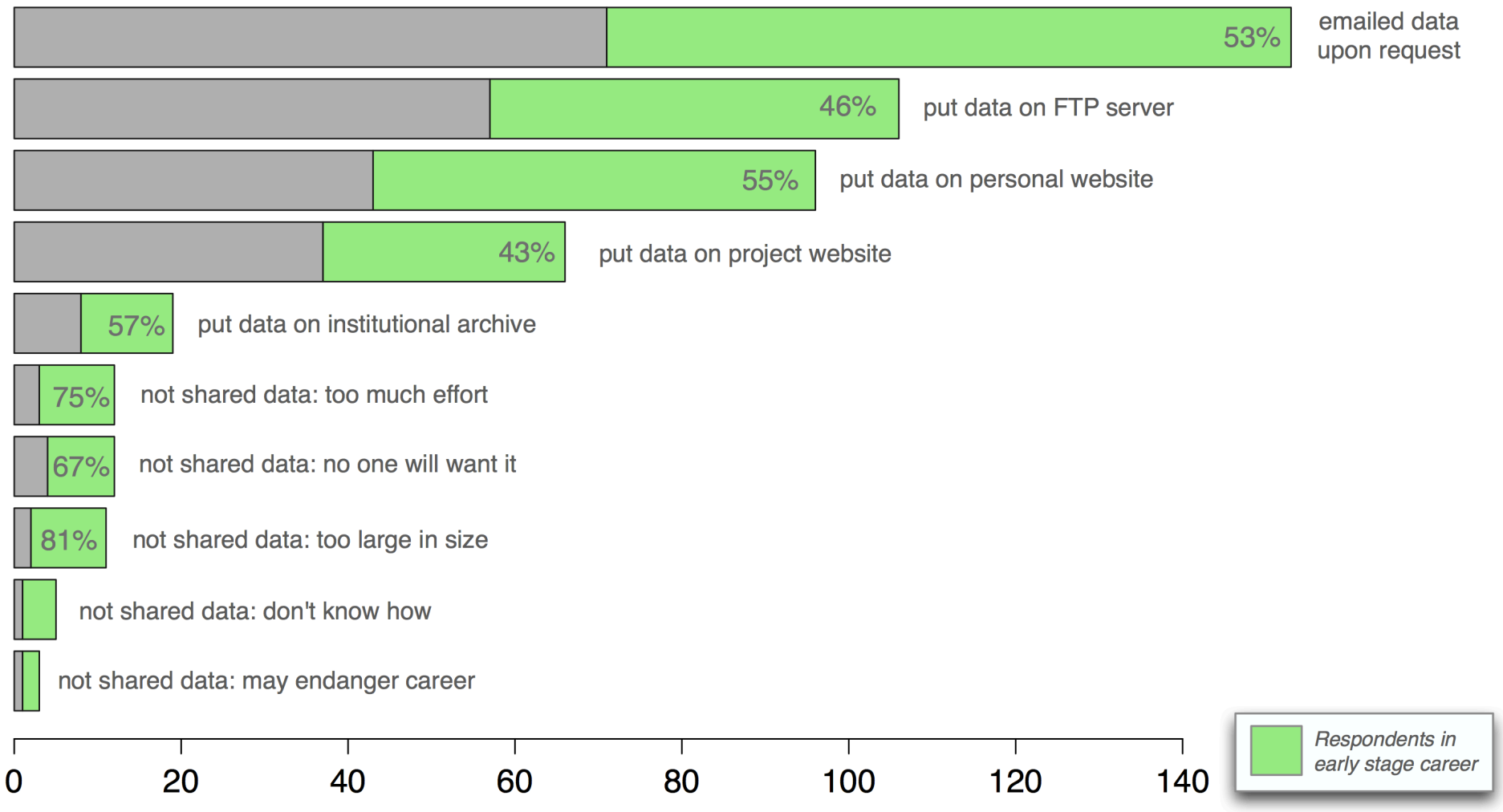
Survey sent to ~ 350 Ph.D. level researchers at the Harvard-Smithsonian Center for Astrophysics; 175 respondents



Have you ever used DATA you learned about from reading a Journal article?
Check ALL that apply

Pepe, Goodman, Muench, Crosas, Erdmann, 2013 "Sharing, Archiving and Citing Data in Astronomy" *Forthcoming*
https://www.authorea.com/users/3/articles/288/_show_article

Slide courtesy of Merce Crosas, Harvard IQSS



When it comes to sharing DATA you've created, collected or curated, you have?

Check ALL that apply.

Pepe, Goodman, Muench, Crosas, Erdmann, 2013 "Sharing, Archiving and Citing Data in Astronomy" *Forthcoming*
https://www.authorea.com/users/3/articles/288/_show_article

Slide courtesy of Merce Crosas, Harvard IQSS

Table 2. Conditions for data sharing.

"I will share my data if...."	Round 1	Round 2	Total
Number of participants in Interview round	*22*	*21*	*43*
Number of participants who mentioned conditions	*16*	*16*	*32*
I have first rights to publish the results from the data	15	5	20
I will receive proper attribution as the data source	5	2	7
The requestor is known to me or my group	2	4	6
My research funder expects me to share	2	4	6
Minimal effort is required to share	1	4	5
Sharing was negotiated in advance of exchange	1	3	4
The data are appropriately sized (not too big or too small)	1	3	4
Research and/or data are developed and stable		3	3
My community expects me to do so		3	3
Minimal effort was required to collect data	2		2
The data will be easily understood by others	1	1	2
The journal requires that the data be shared	1	1	2
Permission was granted by the PI on the project		2	2
Standard methods exist for interoperability	1		1
Shared data are not focus of participant's research	1		1
Data collection is part of my job description	1		1
I do not plan to commercialize the data or technology	1		1
Shared data will be re-shared with others	1		1
Data recipient and I address same research question		1	1
Total Number of Mentions	**36**	**36**	**72**

doi:10.1371/journal.pone.0067332.t002

Wallis JC, Rolando E, Borgman CL (2013) If We Share Data, Will Anyone Use Them? Data Sharing and Reuse in the Long Tail of Science and Technology. PLoS ONE 8(7): e67332. doi:10.1371/journal.pone.0067332
http://www.plosone.org/article/info:doi/10.1371/journal.pone.0067332

Table 3. Methods for sharing data.

Methods for Sharing Data	Round 1	Round 2	Total
Number of participants interviewed	*22*	*21*	*43*
Number of participants mentioning methods to share data	*21*	*15*	*36*
Fulfill personal requests	10	12	22
Post data to a website	15	6	21
Submit data to a repository	2	10	12
Data Publication	2	4	6
Supplement to published journal article	2	1	3
Submit data description to a registry	3	1	4
Total Number of Mentions	**34**	**34**	**68**

doi:10.1371/journal.pone.0067332.t003

Wallis JC, Rolando E, Borgman CL (2013) If We Share Data, Will Anyone Use Them? Data Sharing and Reuse in the Long Tail of Science and Technology. PLoS ONE 8(7): e67332. doi:10.1371/journal.pone.0067332
http://www.plosone.org/article/info:doi/10.1371/journal.pone.0067332

Research practices

- Goal is publications that report the research
- Goal is data that are reusable by others

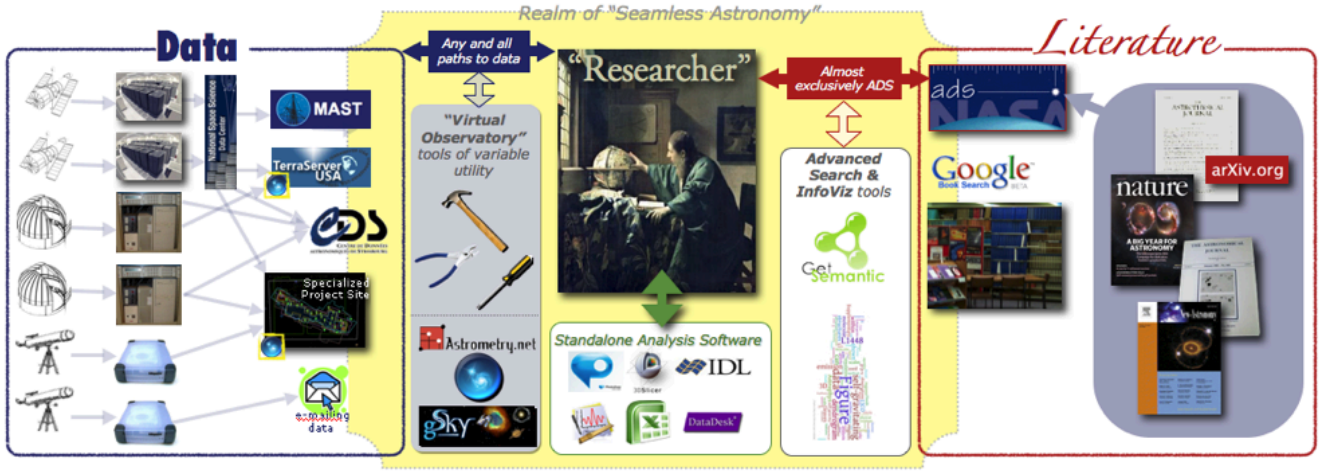
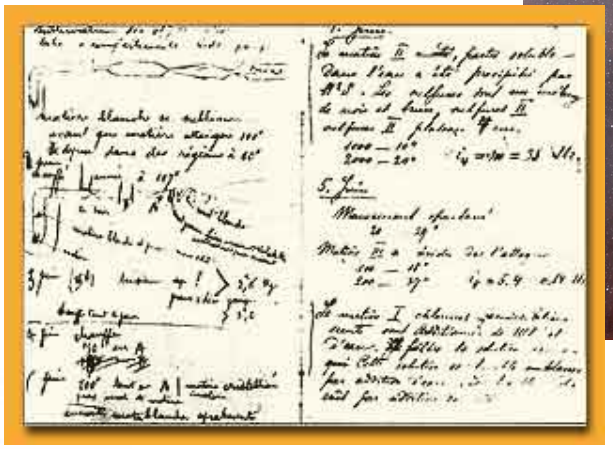


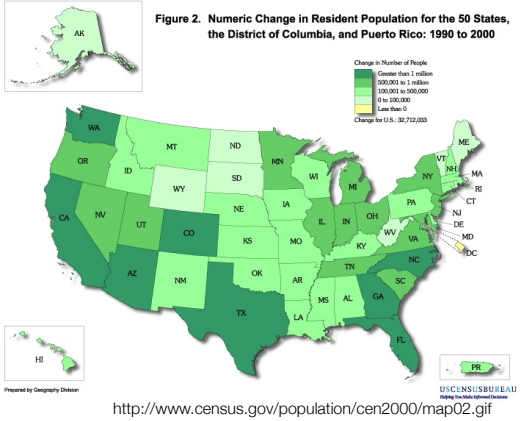
Image: Alyssa Goodman, Harvard Astronomy

What are data?

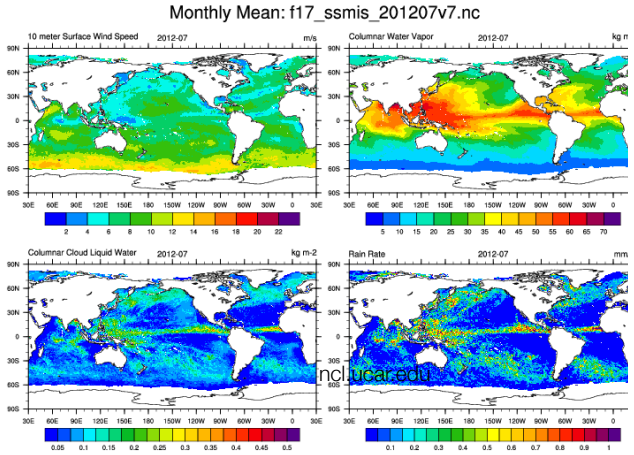
hudsonalpha.org



Marie Curie's notebook aip.org



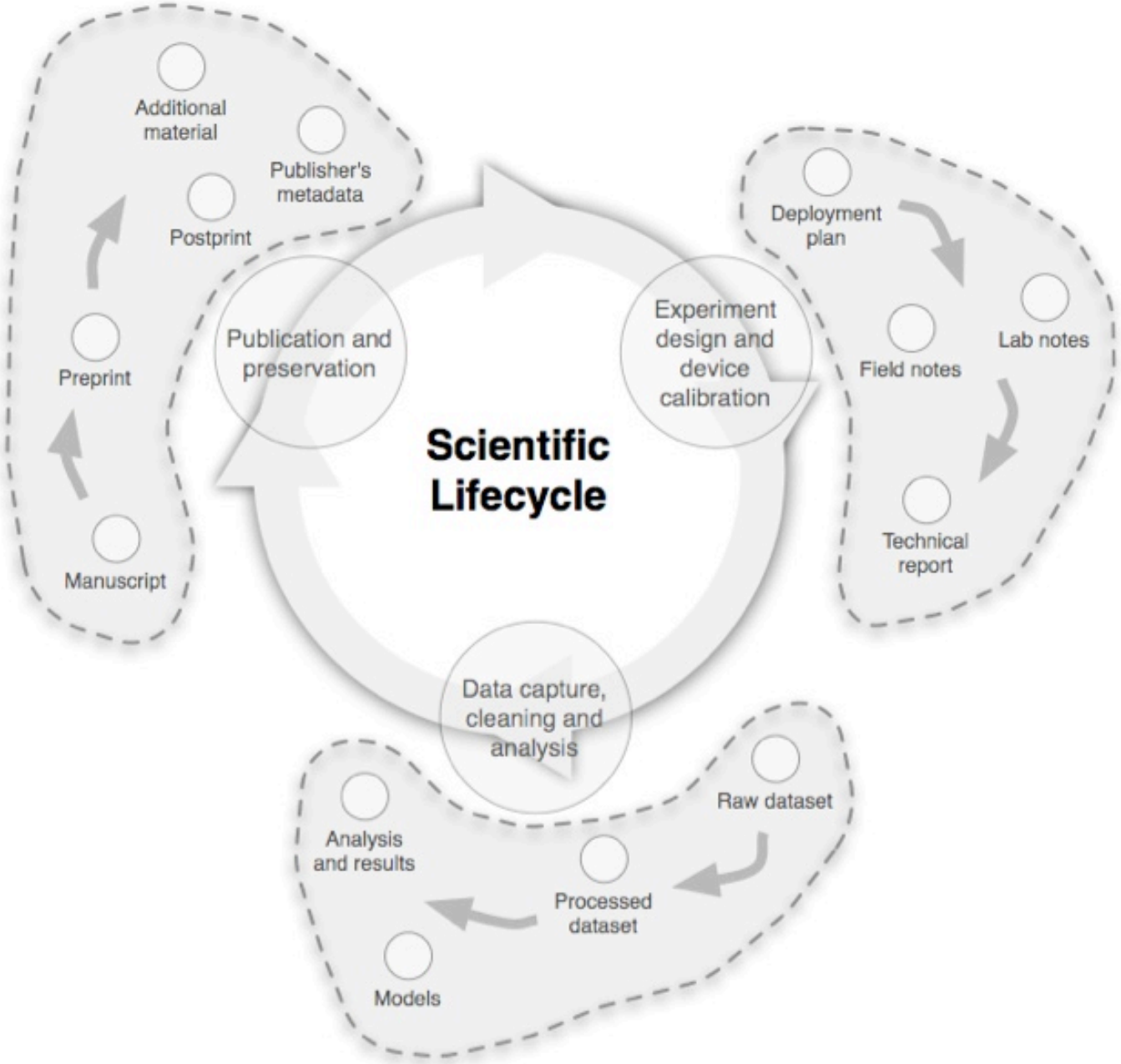
http://www.census.gov/population/cen2000/map02.gif



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



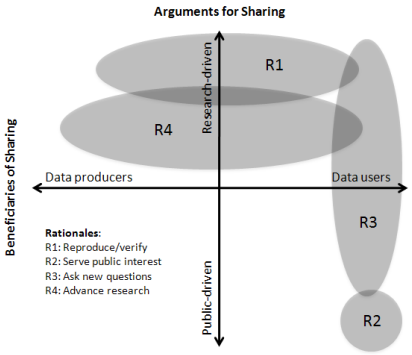
Pepe, A., Mayernik, M. S., Borgman, C. L. & Van de Sompel, H. (2010). From Artifacts to Aggregations: Modeling Scientific Life Cycles on the Semantic Web. Journal of the American Society for Information Science and Technology, 61(3): 567–582.

Overview

- Paradigm shift
- Arguments for sharing data
- Science friction, data friction
- Sharing and reusing data

http://inventionmachine.com/the-Invention-Machine-Blog/bid/51703/Three-Key-Challenges-to-Entering-New-Markets

Why share research data?



Rationales

1. To reproduce research
2. To make public assets available to the public
3. To leverage investments in research data
4. To advance research and innovation

Borgman, C.L. (2012). The Conundrum of Sharing Research Data. JASIST, 63(6):1059–1078

&

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

1. To reproduce research

http://chemistry.curtin.edu.au/research/index.cfm

Benzoic Acid	%yield		IR Peaks (cm^{-1})		Solid (C) or Oil (O) Product	Mp (^0C)
	Gross	Recrystallization	N-H	C=O		
Sodium benzoate		2.58	3327	1638	White C	79-89
Sodium benzoate			3337	1640&1600	O	
Sodium benzoate			3326	1642&1601	O	
Sodium benzoate	37.8		3274	1640	O	
p-nitro	51.84	10.59	3423	1693	Yellow C	152-157
m-nitro	37.38	5.43	3334	1694	Green C	152-157
Benzoic acid		7.44	3293	1642	White C	152-154
m-bromo		47.4	3316	1702	Green paste	
p-bromo		14.53	3344	1638	Pink C	164-166
p-chloro		29.69	3340	1638	Yellow C	
m-chloro		74.53	3410	1637	tan paste	
o-chloro		17.31	3422	1654	Tan C	
3,5-dinitro		44.53	3297	1647	Tan C	139-141
p-hydroxy		3.751	3401	1643	yellow/green C	210
p-amino		8.475	3411	1645	Dark O	
o-methoxy		42.49	3412	1646	Yellow O	

http://serc.carleton.edu/cismi/broadaccess/groupwork.html

Scientific Gold Standard

REPLICATION—THE CONFIRMATION OF RESULTS AND CONCLUSIONS FROM ONE STUDY obtained independently in another—is considered the scientific gold standard.

Jasny, B. R., Chin, G., Chong, L. & Vignieri, S. (2011). Again, and again, and again. Science, 334(6060): 1225.

Victoria Stodden, Columbia

Reproducibility?

- Deductive sciences
  - Check the proof
- Experimental sciences
  - Redo the field work
- Computational sciences
  - Start with the dataset
  - Reconstruct workflow

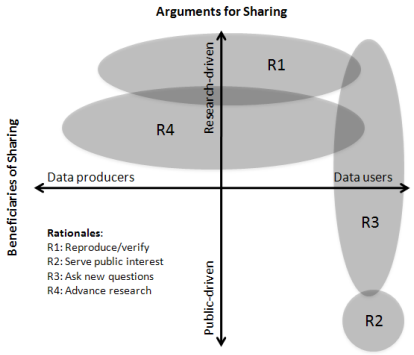
Analytic validity	Do different labs, techniques, and platforms measure the same thing?
Repeatability	Can other scientists access the data and protocols, repeat the analyses, and get the same results?
Replication	Do many different data sets and their combination (meta-analysis) get consistent results?
External validation	Do different data sets by different teams, preferably prospectively and with large-scale evidence, get consistent results?
Clinical validity	Does the discovered information predict clinical outcomes?
Clinical utility	Does the use of the discovered information improve clinical outcomes?

Science AAAS

Published by AAAS

J P A Ioannidis, M J Khoury Science 2011;334:1230-1232

Why share research data?



Rationales

1. To reproduce research
2. To make public assets available to the public
3. To leverage investments in research data
4. To advance research and innovation

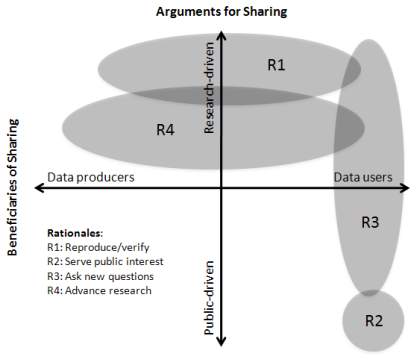
Borgman, C.L. (2012). The Conundrum of Sharing Research Data. JASIST, 63(6):1059–1078

&

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

2. To make public assets available to the public

Why share research data?



Rationales

1. To reproduce research
2. To make public assets available to the public
3. To leverage investments in research data
4. To advance research and innovation

Borgman, C.L. (2012). The Conundrum of Sharing Research Data. JASIST, 63(6):1059–1078

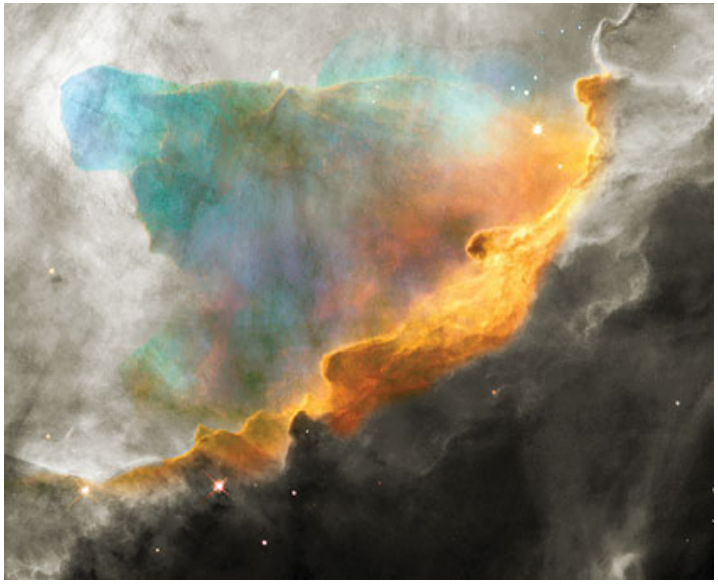
&

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

3. To leverage investments in research data

data

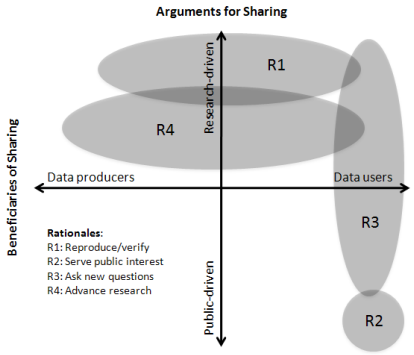
discovery



http://annualreport.ucdavis.edu/2008/images/photos/discovery.jpg

http://digitalassetmanagement.org.uk/2010/02/01/the-winds-of-change-are-blowing-in-the-clouds-favor/

Why share research data?



Rationales

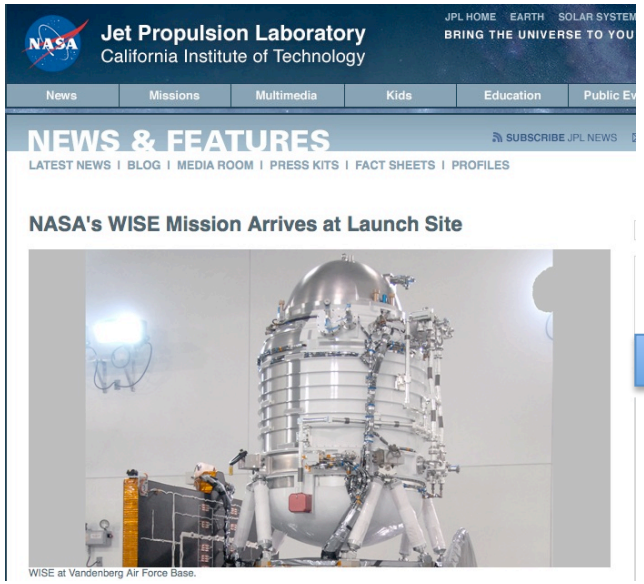
1. To reproduce research
2. To make public assets available to the public
3. To leverage investments in research data
4. To advance research and innovation

Borgman, C.L. (2012). The Conundrum of Sharing Research Data. JASIST, 63(6):1059–1078

&

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

4. To advance research and innovation



International Virtual Observatory Alliance

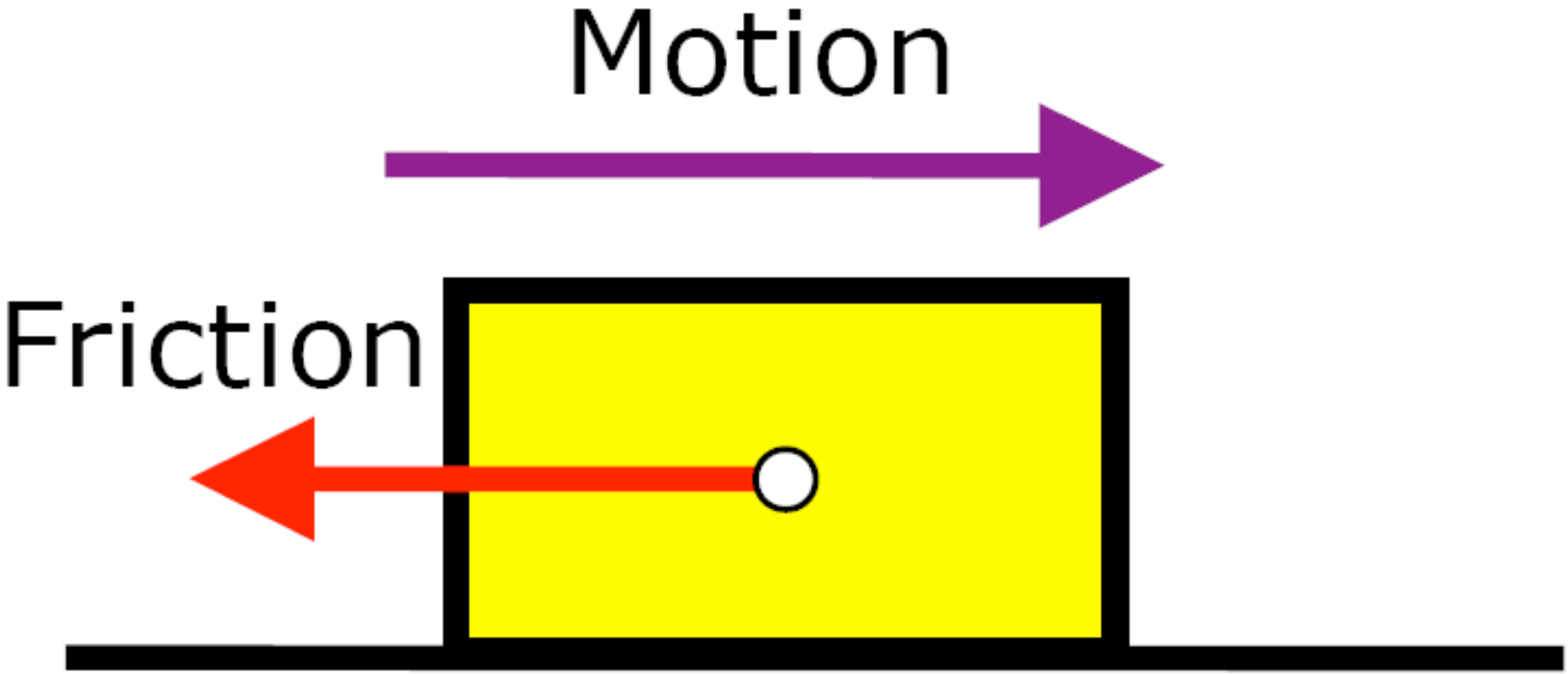
WISE image

Overview

- Paradigm shift
- Arguments for sharing data
- Science friction, data friction
- Sharing and reusing data

http://inventionmachine.com/the-Invention-Machine-Blog/bid/51703/Three-Key-Challenges-to-Entering-New-Markets

Science friction, data friction*



*Edwards, P. N., Mayernik, M. S., Batcheller, A. L., Bowker, G. C., & Borgman, C. L. (2011). Science Friction: Data, Metadata, and Collaboration. *Social Studies of Science*, *41*, 667–690. doi:10.1177/0306312711413314

http://www.stmary.ws/highschool/physics/home/notes/dynamics/friction/imgE2.gif

Lack of incentives to share data

- Rewards for publication
- Effort to document data
- Competition, priority
- Control, ownership
- Legal liability

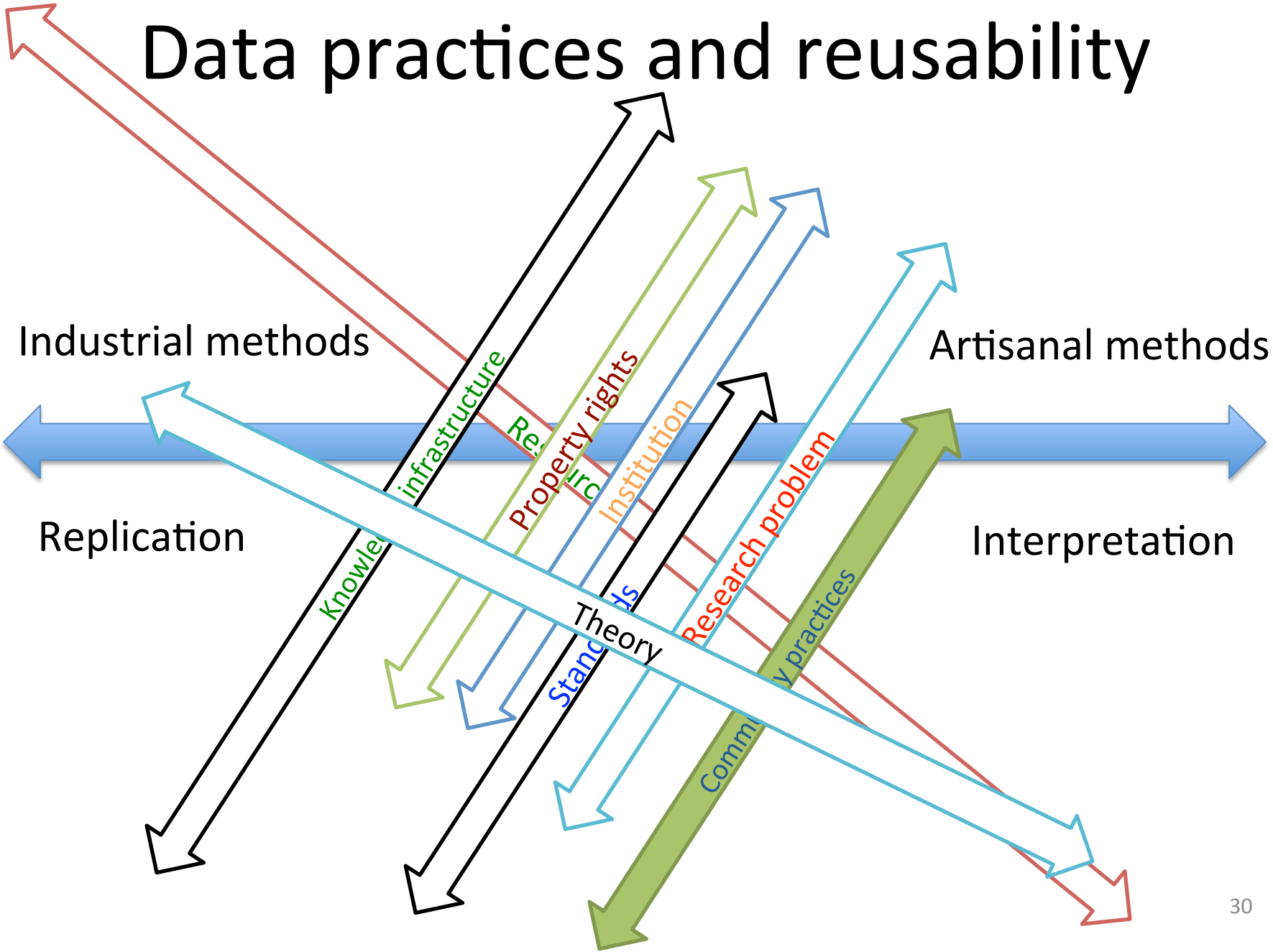
Image source: www.buildingsrus.co.uk/.../ target1.htm

Intractable problems

- Confidentiality
- Anonymization
- Re-identification
- Intellectual property
- Economics

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

Data practices and reusability



Data do not stand alone

- Data are inseparable
  - Code
  - Technical standards
  - Documentation
  - Instrumentation
  - Calibration
  - Provenance
  - Workflows
  - Local practices
  - Physical samples

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

Why openness matters

- Discoverability of related
  - Data
  - Documentation
  - Digital objects
  - Publications
- Interoperability
  - Import and export data
  - Mine and combine
  - Avoid lock-in
- Usability and reusability
  - For research
  - For learning

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

Rewards for publications

Competition, priority

Everyone is overwhelmed with life and email and, in academia, trying to get funding and write papers. Whether something is open or not open is not highest on the priority list. There's still need for making people aware of open science issues and making it easy for them to participate if they want to.

Effort to document data

Control, ownership

Jonathan Eisen, genetics professor at the University of California, Davis

DESPITE BEING GOOD FOR YOU AND FOR SCIENCE, TOO MANY CHALLENGES AND TOO LITTLE TIME

Slide courtesy of Merce Crosas, Harvard IQSS

Overview

- Paradigm shift
- Arguments for sharing data
- Science friction, data friction
- Sharing and reusing data

http://inventionmachine.com/the-Invention-Machine-Blog/bid/51703/Three-Key-Challenges-to-Entering-New-Markets

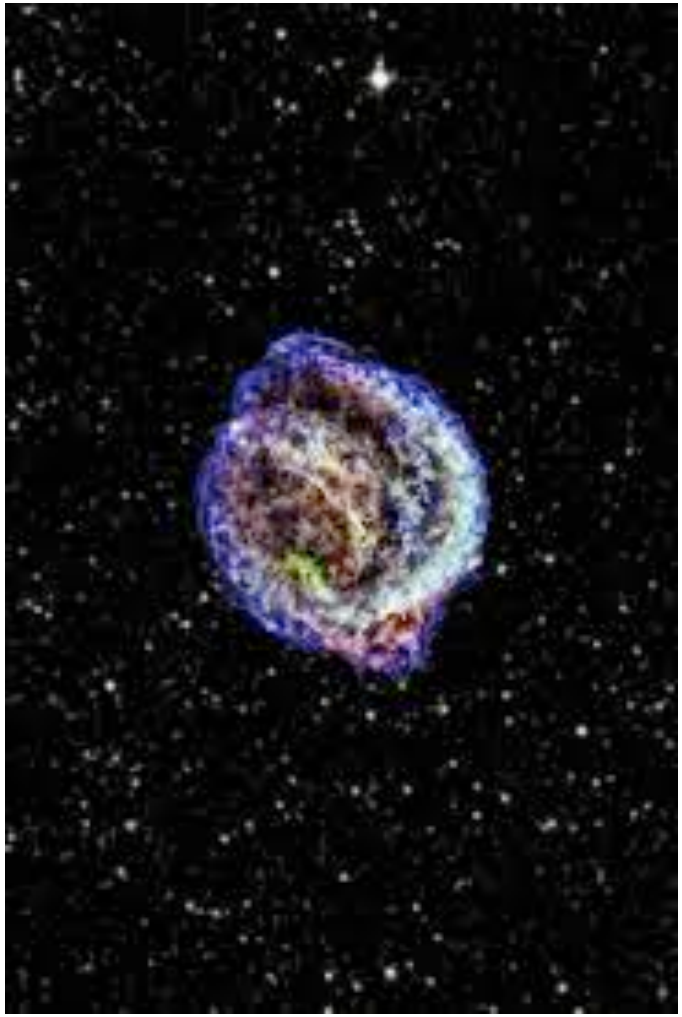
10 Simple Rules for the Care and Feeding of Scientific Data*

1. Love your data, and help others love it too.
2. Share your data online, with a permanent identifier.
3. Conduct science with a particular level of reuse in mind.
4. Publish workflow as context
5. Link your data to your publications as early as possible.
6. Publish your code (even the small bits).
7. Say how you want to get credit for your data (and software).
8. Foster and use data repositories.
9. Reward colleagues who share their data properly.
10. Appendices: Links to useful resources

**Goodman, A.; Pepe, A.; Blocker, A.; Borgman, C.L., et al, (in review), PLOS Computational Biology*
https://www.authorea.com/users/23/articles/1839/_show_article

Distance from origin

- Reuse by investigator
- Reuse by collaborators
- Reuse by colleagues
- Reuse by unaffiliated others
- Reuse at later times
  - Months
  - Years
  - Decades
  - Centuries



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

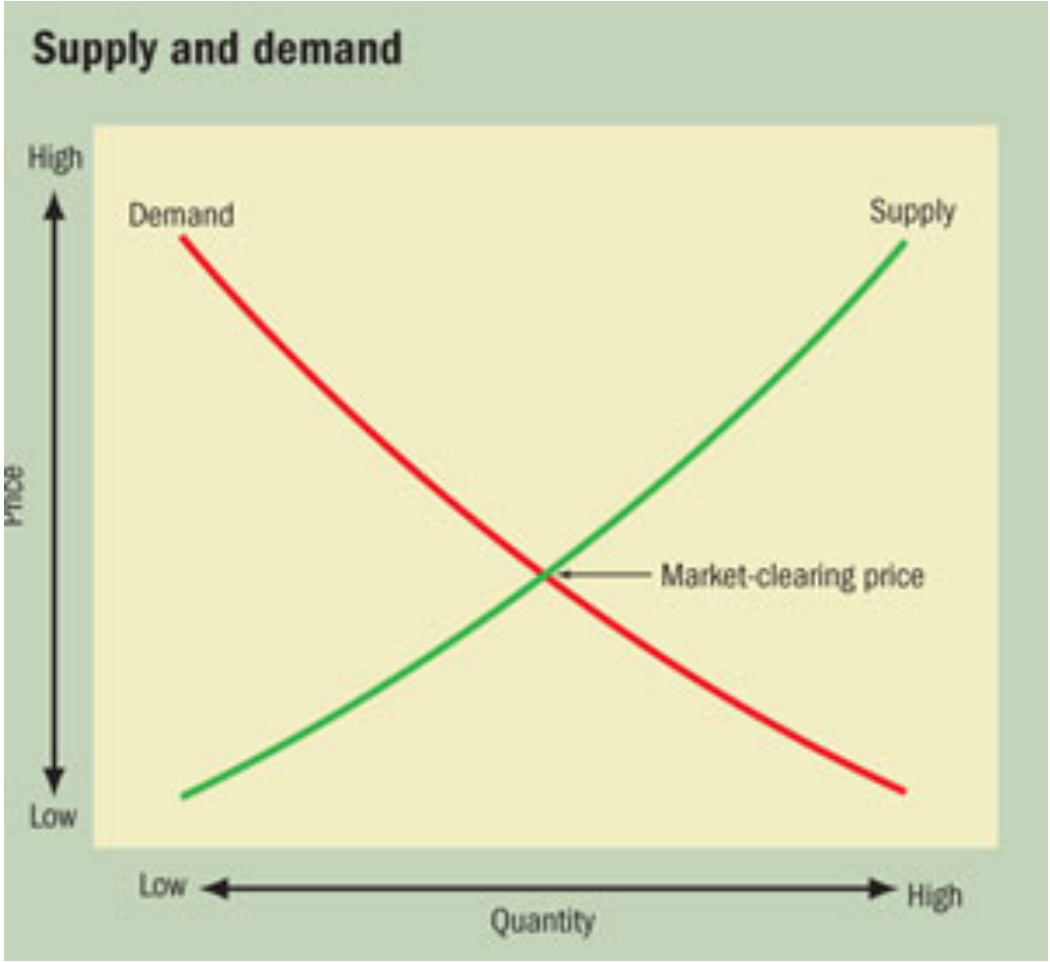
Ways to share data

- Make data publicly available
  - Curated data archive: NASA, UKDA, ICPSR…
  - Author curated data archive
  - University repository
  - Personal website
  - ftp site
- Release upon request*

*Wallis, J. C., Rolando, E., & Borgman, C. L. (2013). If We Share Data, Will Anyone Use Them? Data Sharing and Reuse in the Long Tail of Science and Technology. *PLoS ONE*, *8*(7), e67332. doi:10.1371/journal.pone.0067332

http://www.zippykidstore.com/

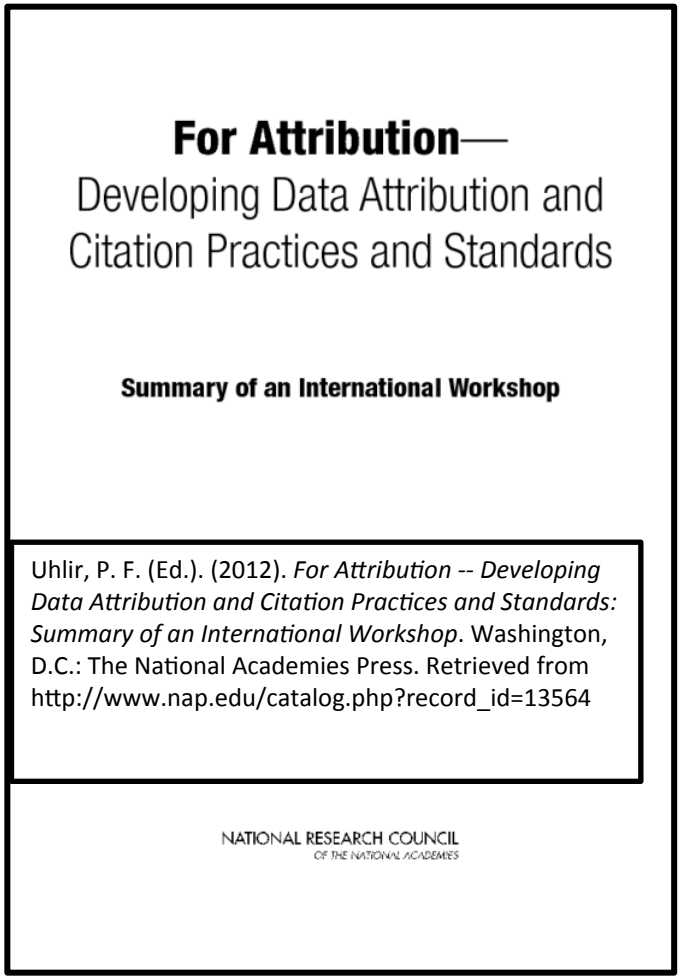
Supply = continuity, trust



Demand = investment, risk

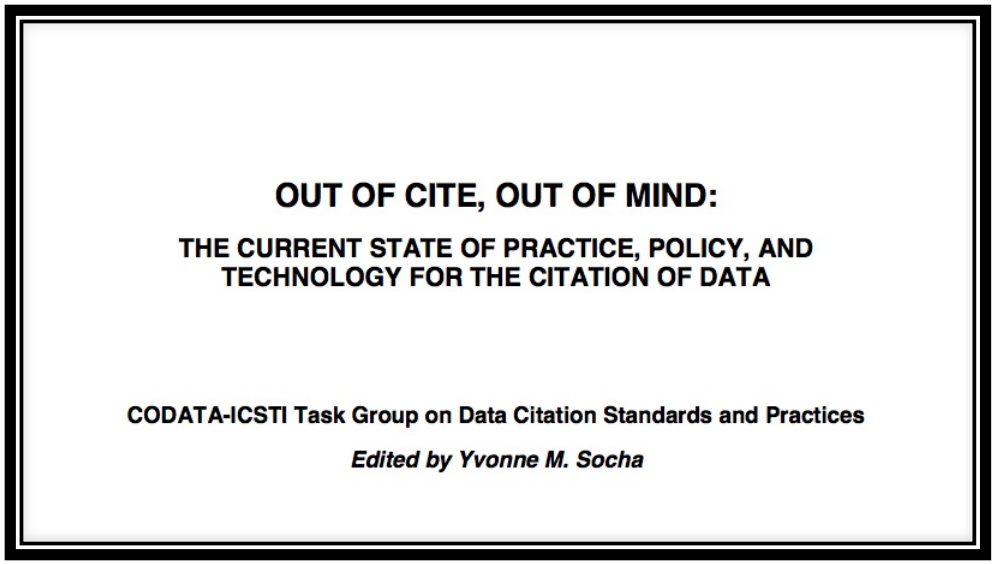
http://www.imf.org/external/pubs/ft/fandd/basics/images/suppdem_cht.jpg

Data Citation and Attribution



Uhlir, P. F. (Ed.). (2012). *For Attribution -- Developing Data Attribution and Citation Practices and Standards: Summary of an International Workshop*. Washington, D.C.: The National Academies Press. Retrieved from http://www.nap.edu/catalog.php?record_id=13564

2012



Data Science Journal, Volume 12, 13 September 2013

Acknowledgements

- UCLA Data Practices team
  - Rebekah Cummings, Peter Darch, David Fearon, Fred Ariel Hernandez, Elaine Levia, Rachel Mandell, Matthew Mayernik, Alberto Pepe, Ashley Sands, Katie Shilton, Sharon Traweek, Jillian Wallis, Laura Wynholds, Kan Zhang
- Research funding
  - National Science Foundation
  - Alfred P. Sloan Foundation
  - Microsoft External Research
- University of Oxford
  - Balliol College
  - Oliver Smithies Fellowship
  - Oxford Internet Institute
  - Oxford eResearch Center
  - Bodleian Library