

UCLA

Presentations

Title

Whose Science? Whose Data? Whose Evidence?

Permalink

<https://escholarship.org/uc/item/9qc8c2qh>

Author

Borgman, Christine L.

Publication Date

2019-07-25

Copyright Information

This work is made available under the terms of a Creative Commons Attribution-NonCommercial-NoDerivatives License, available at <https://creativecommons.org/licenses/by-nc-nd/4.0/>

Whose Science? Whose Data? Whose Evidence?

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

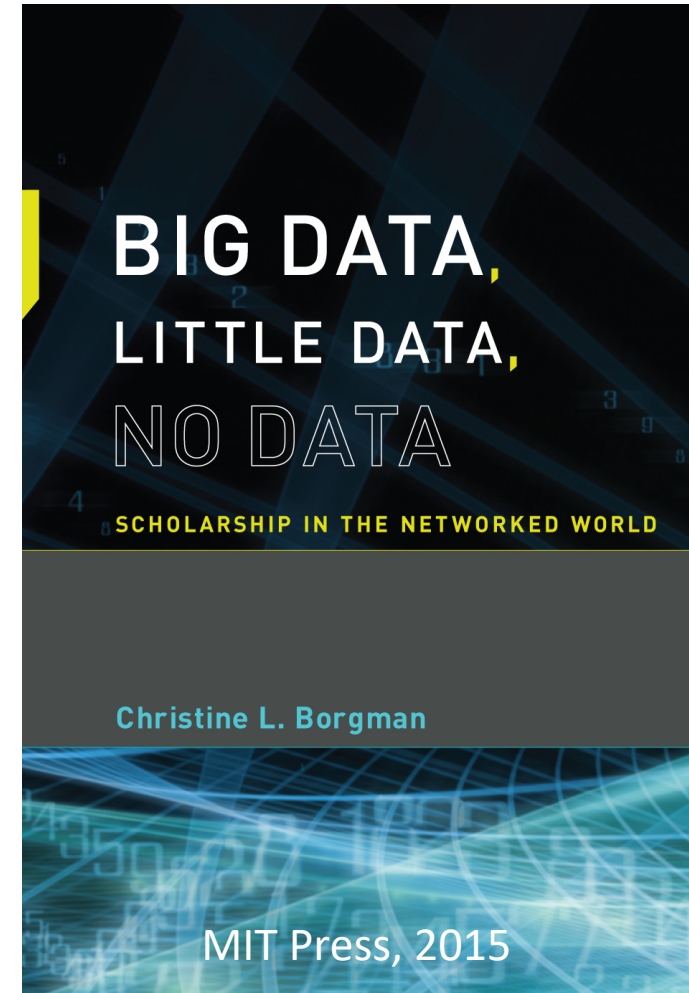
Keynote presentation

Open Science and the Role of Common Evidence

Sage Bionetworks 10th Anniversary Assembly

Seattle, WA, July 25, 2019

<https://sagebionetworks.org/events/2019-assembly/>





Data sharing policies



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



Supported by
wellcometrust



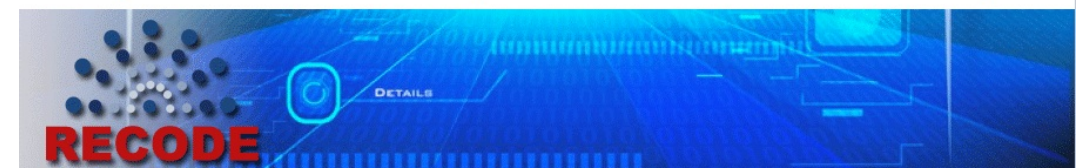
Australian Government
National Health and Medical Research Council



National Science Foundation
WHERE DISCOVERIES BEGIN



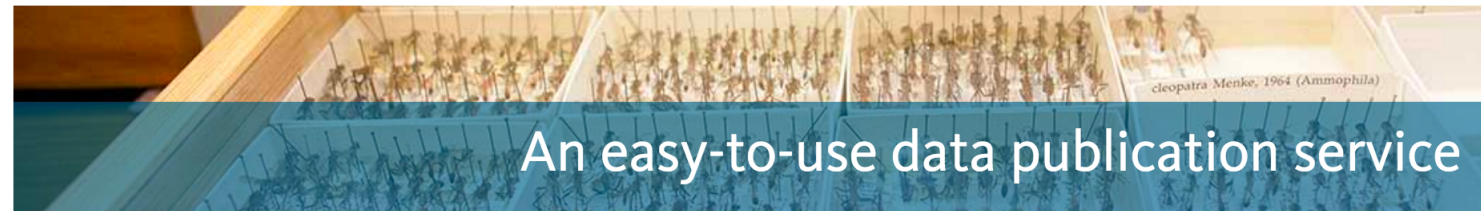
Policy RECommendations for Open Access to Research Data in Europe





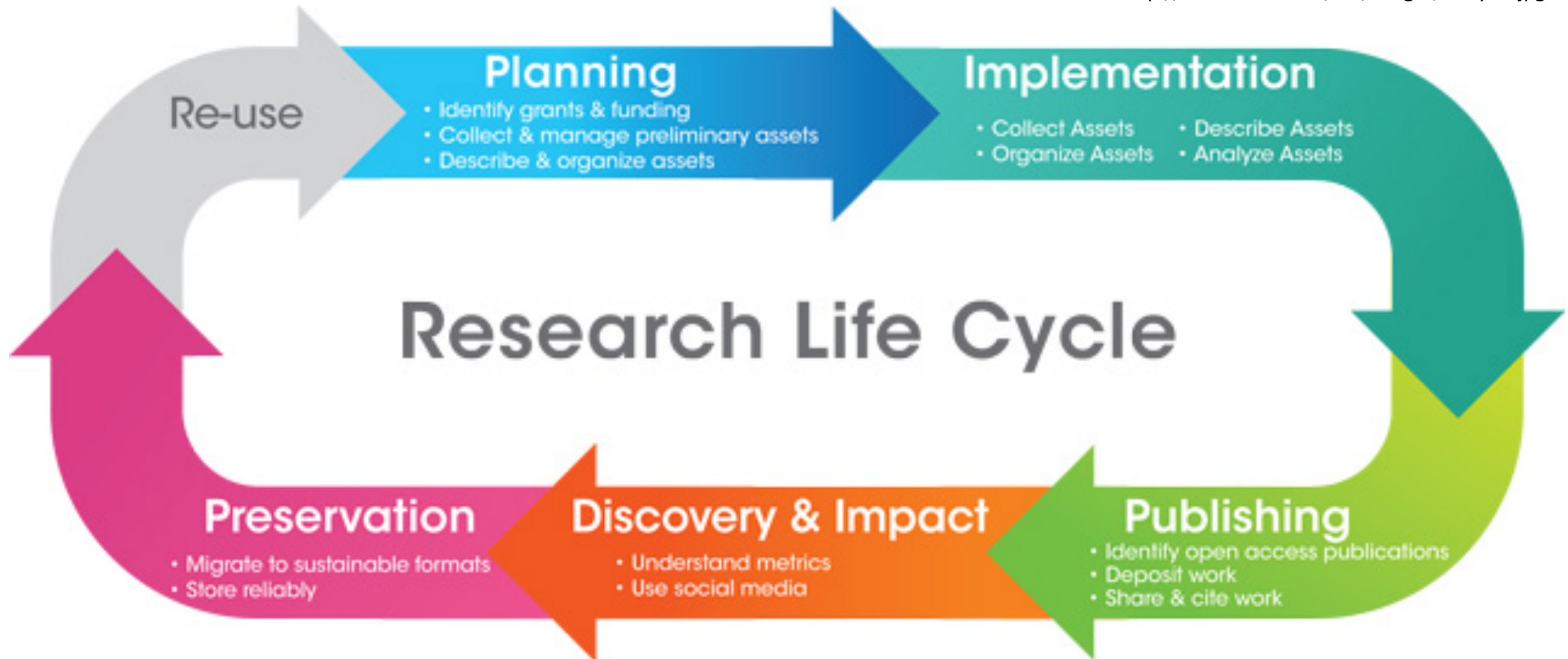
Open Data Practices

- Deposit datasets in a data archive
- Link datasets to journal article or publication
- Publish data documentation
 - Research protocols
 - Codebooks
 - Software
 - Algorithms
- Cite data and software



Data creation and reuse: The Ideal

<http://www.lib.uci.edu/dss/images/lifecycle.jpg>



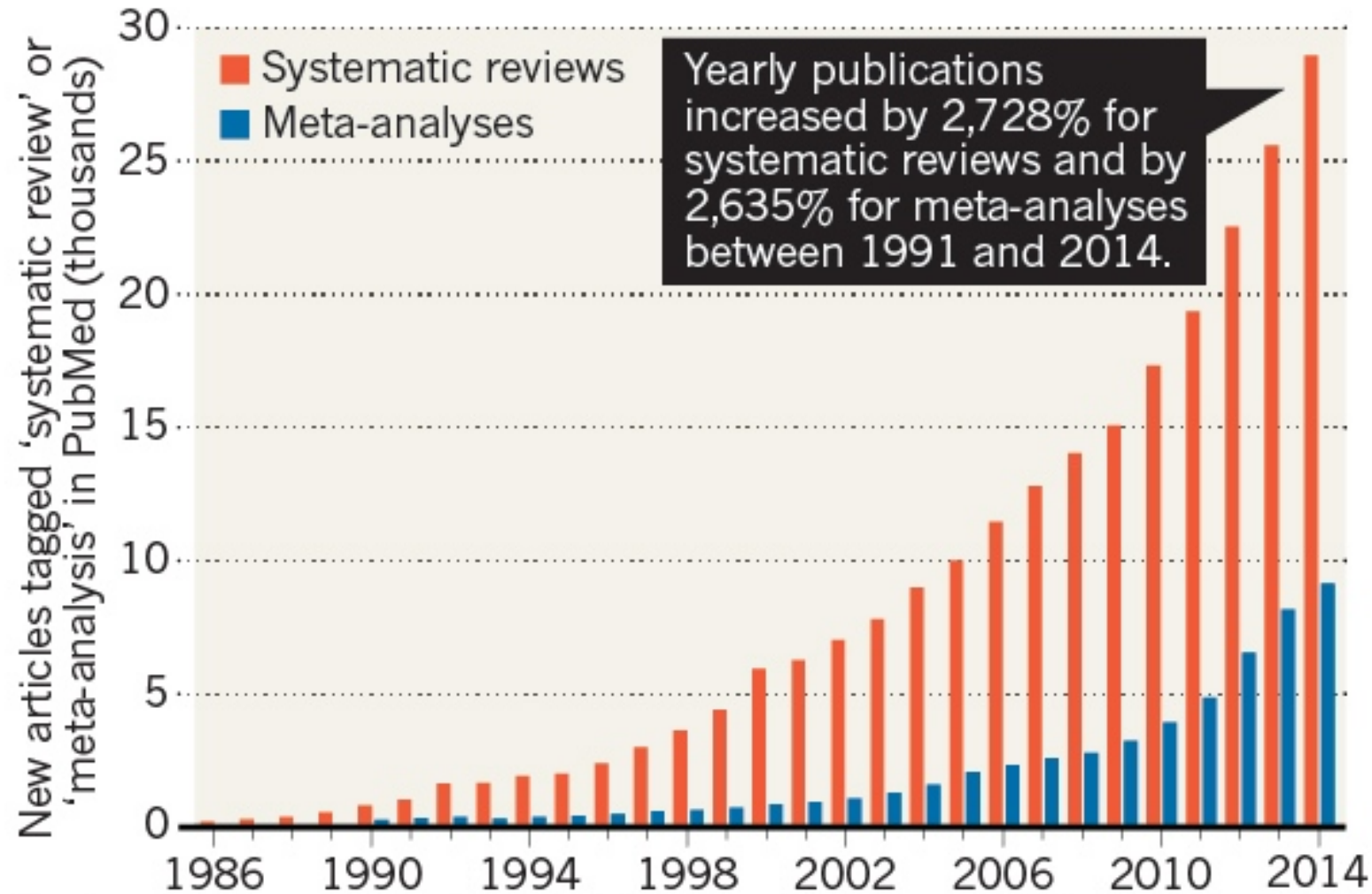
Borgman, C. L. (2019). The Lives and After Lives of Data. *Harvard Data Science Review*, 1(1). <https://doi.org/10.1162/99608f92.9a36bdb6>

Pasquetto, I. V., Randles, B. M., & Borgman, C. L. (2017). **On the Reuse of Scientific Data**. *Data Science Journal*, 16. <https://doi.org/10.5334/dsj-2017-008>


Evidence-based medicine

META MASS PRODUCTION

The number of systematic reviews and meta-analyses published each year has proliferated since 1986.



A systematic review analyses and compiles all papers, and sometimes unpublished work, on a topic. A meta-analysis is a systematic review that combines data from multiple papers.

Original Investigation |  Full Access |

The Mass Production of Redundant, Misleading, and Conflicted Systematic Reviews and Meta-analyses

JOHN P.A. IOANNIDIS 

First published: 13 September 2016 | <https://doi.org/10.1111/1468-0009.12210> | Cited by: 212

Publications \leftrightarrow Data

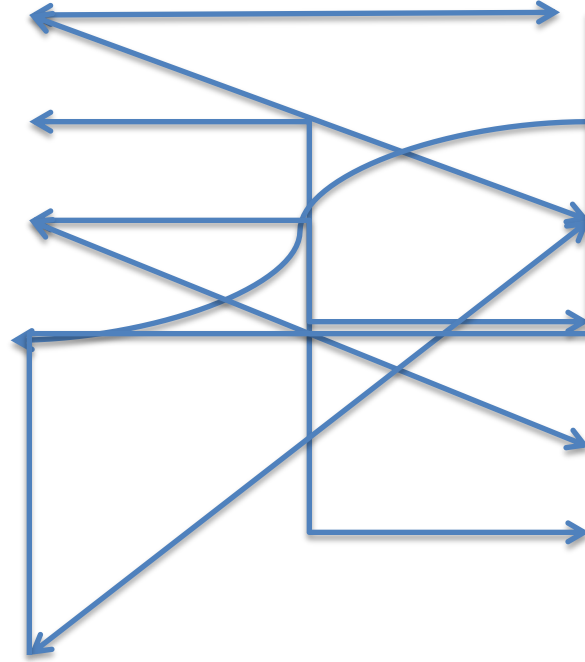
Publications are arguments made by authors, and data are the evidence used to support the arguments.

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



Publications \leftrightarrow 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

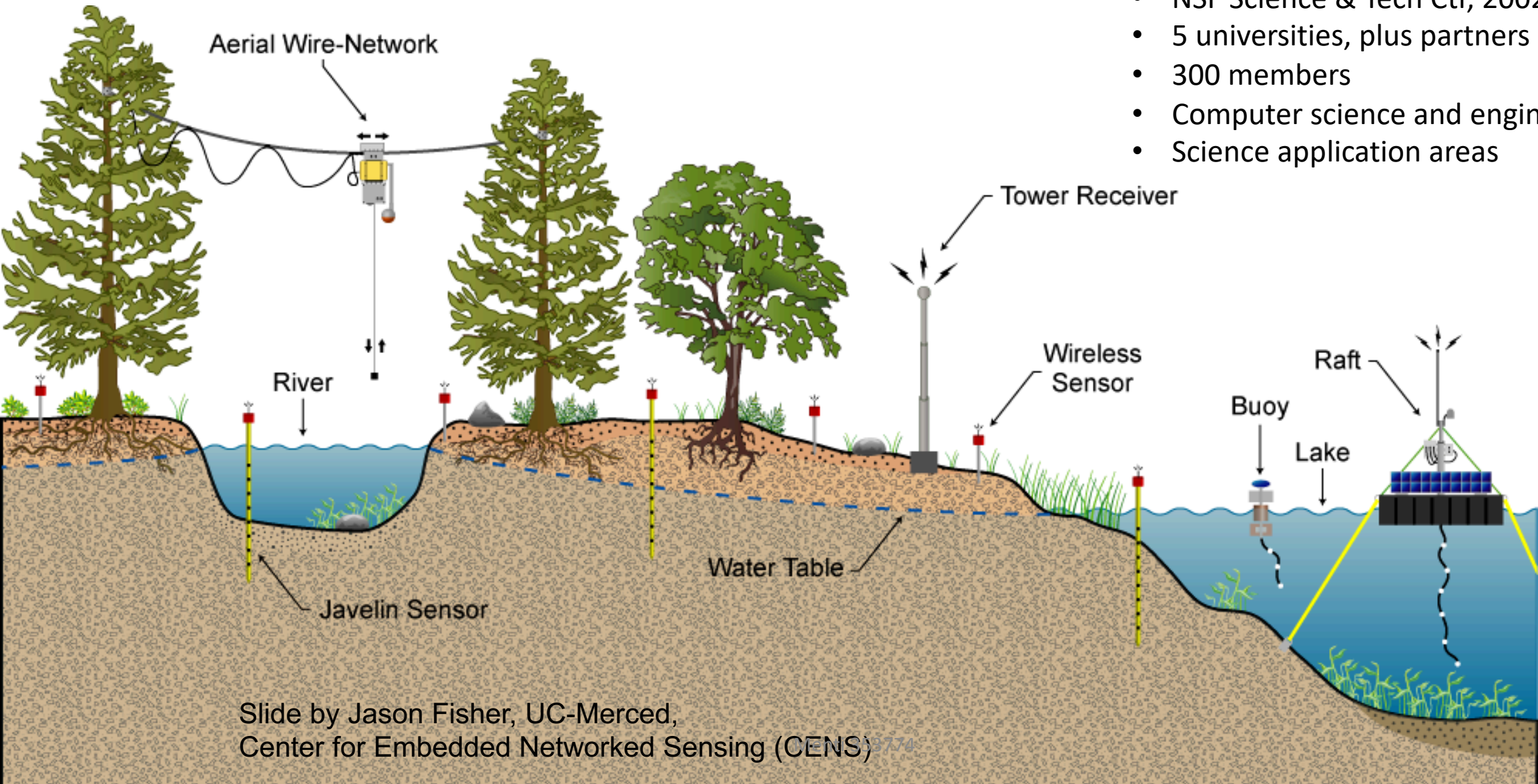


Data are representations of observations, objects, or other entities used as evidence of phenomena for the purposes of research or scholarship.

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

Center for Embedded Networked Sensing

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

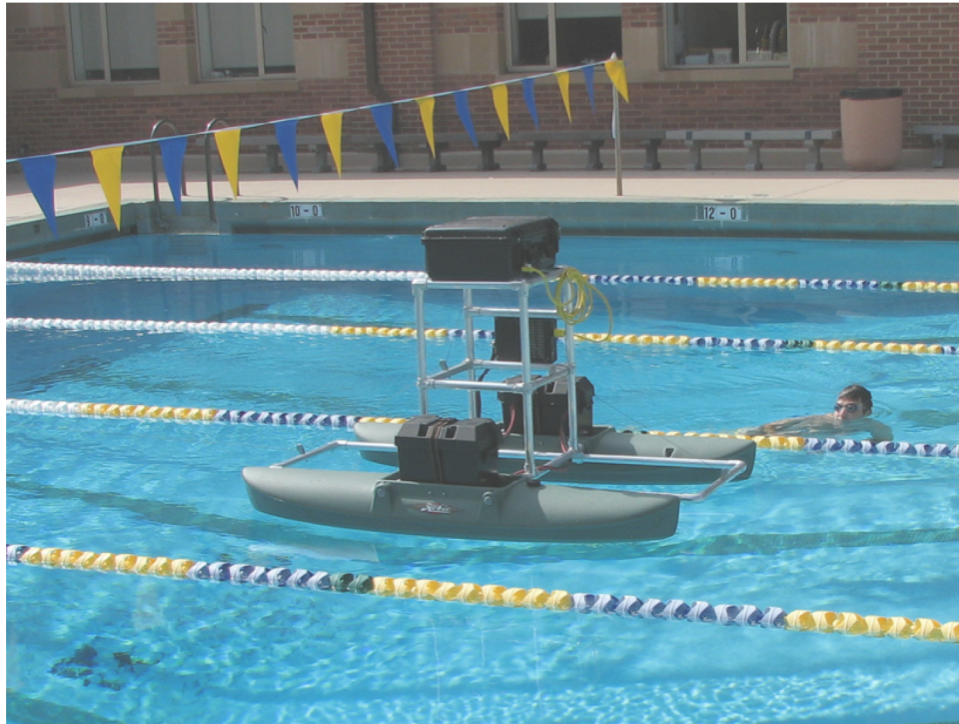


Slide by Jason Fisher, UC-Merced,
Center for Embedded Networked Sensing (CENS)

Science \leftrightarrow Data

Engineering researcher:

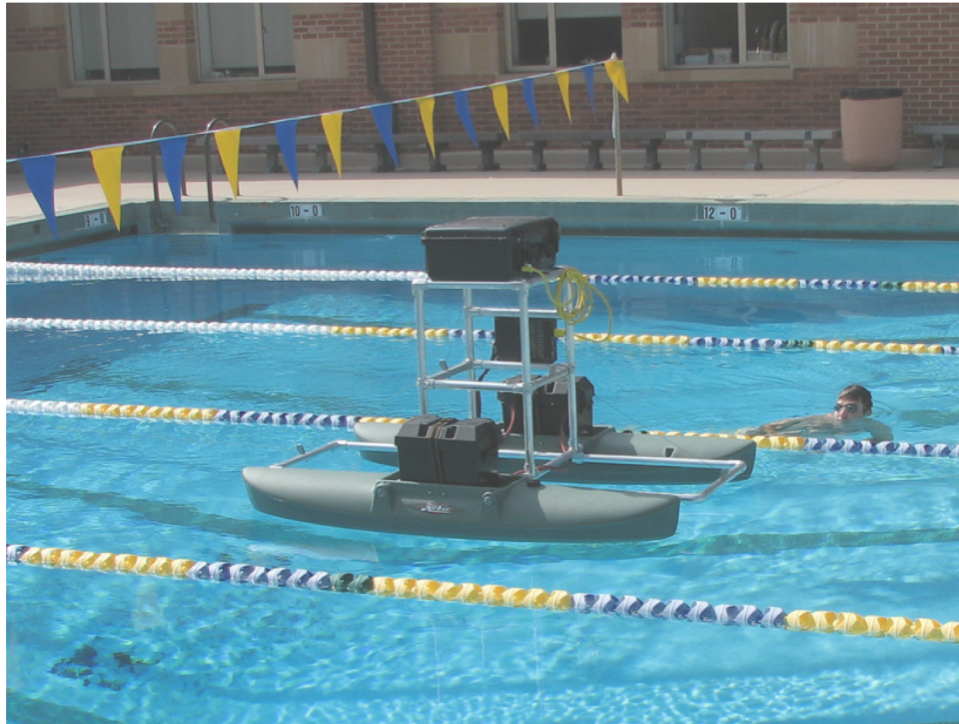
“Temperature is temperature.”



CENS Robotics team

Science \leftrightarrow Data

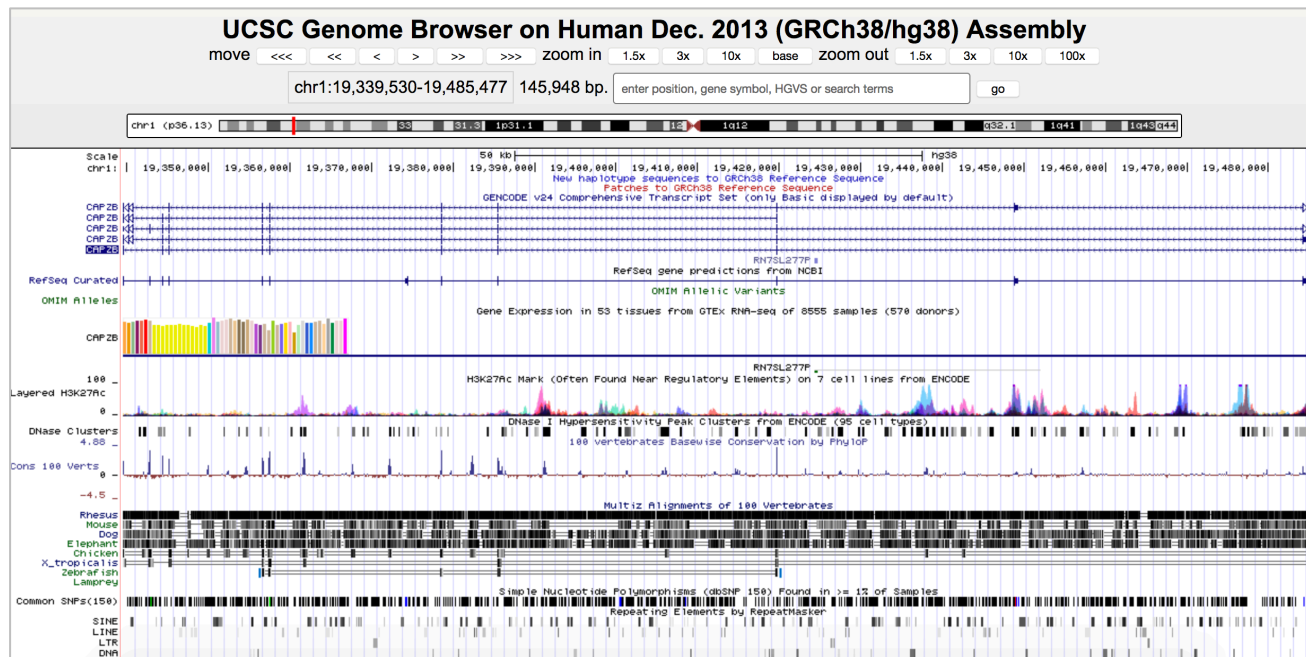
Engineering researcher:
“Temperature is temperature.”



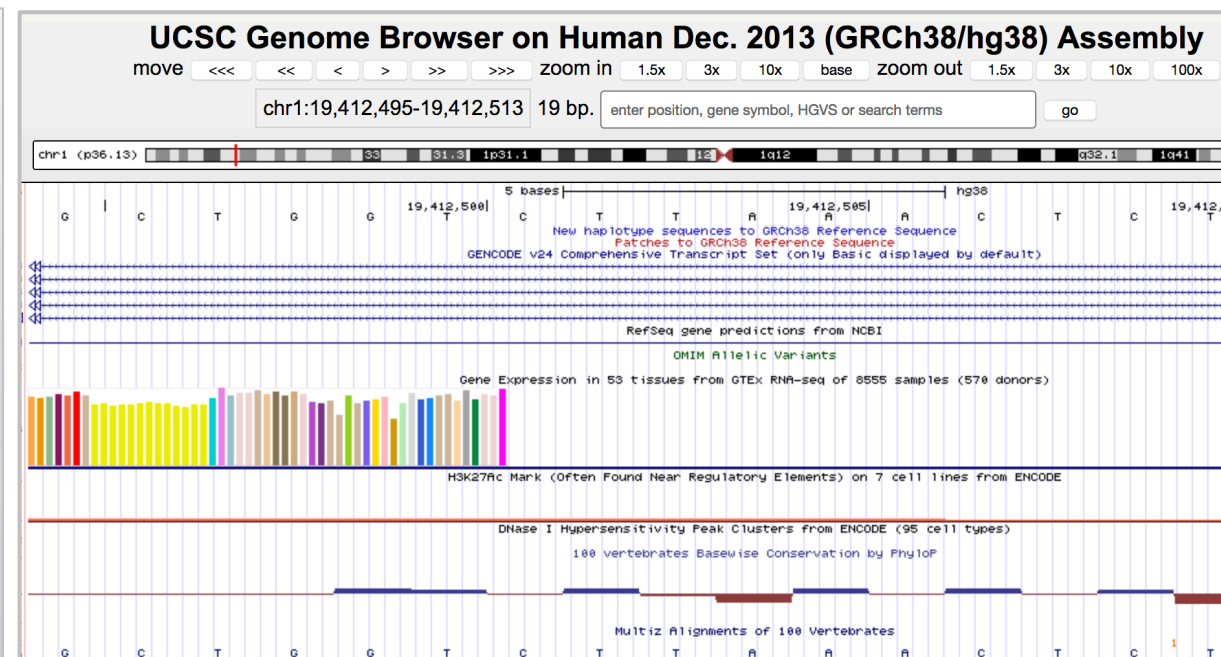
CENS Robotics team

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..”

Comparative data reuse: calibrate, control, “ground truth”

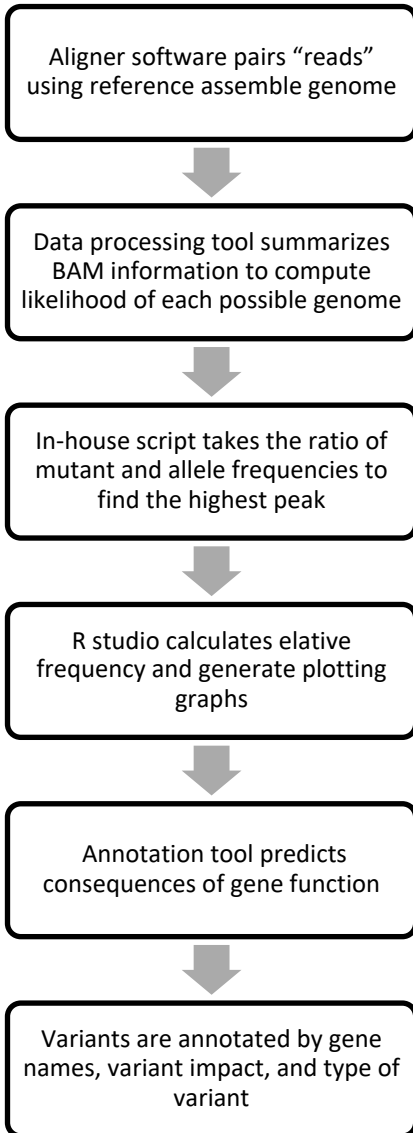


UCSC Genome Browser – Search example
(CAPZB gene)



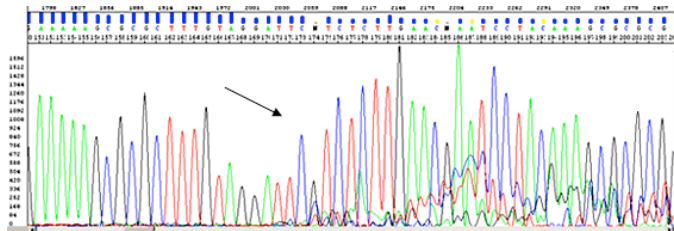
UCSC Genome Browser – Zoom IN

Integrative Data Reuse: Hypothesis Testing and Statistical Analysis

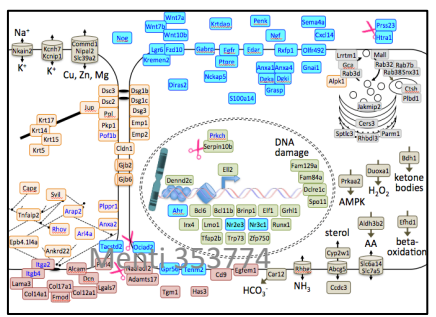


```
@HD VN:1.0 S0:coordinate
@SQ SN:chr20 LN:64444167
@PG ID:tophat VN:2.0.14 CL:/srv/dna_tools/tophat/tophat -N 3 --read-edit-dist 5 --read-read-
lign-edit-dist 2 -i 50 -I 5000 --max-coverage-intron 5000 -M -o out /data/user446/mapping_tophat/index/chr
20 /data/user446/mapping_tophat/L6.18.GTAAA.L007.R1.001.fastq
HWI-ST1145:74:C181DACXX:7:1102:4284:73714 16 chr20 199930 3 100M * 0 0 0
CCGTGTTAAAGTGGATGGGCTACCTCCAGCTAGGCTTAGGATTCCTAGTTGGCTAGGAAATCCAGTAGTCCTGCTCTCAGTCCCCCTC
C BBDDCCDDDDDDDDDDDDDDCCDCBC?DDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDD
AS:i:-15 XM:i:3 XO:i:0 XG:i:0 MD:Z:55C20C13A9 NM:i:3 NH:i:2 CC:Z:= CP:1:55352714 HI:1:0
HWI-ST1145:74:C181DACXX:7:1114:2759:41961 16 chr20 193953 50 100M * 0 0 0
TGCTGGATCATGTTAGTGGCTTCTGACTCAGAGACCTCTGCCCCCTGGGCGAGTGGACCTTCAGTGATTCCCTGCATAAGGGGCGATGGACGA
G DCDDDEDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDD
AS:i:-16 XM:i:3 XO:i:0 XG:i:0 MD:Z:60G16T18T3 NM:i:3 NH:i:1
HWI-ST1145:74:C181DACXX:7:1204:14760:4030 16 chr20 270877 50 100M * 0 0 0
GGCTTTATGGTAAAAAAGGAATGAGAGATTAAATCAGAAATCCACCTGGCCAGCAGACCAACGAGAAGAAGGAAGAACAGCAAGAAAAACA
C DDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDD
AS:i:-11 XM:i:2 XO:i:0 XG:i:0 MD:Z:0A85G13 NM:i:2 NH:i:1
HWI-ST1145:74:C181DACXX:7:1210:11167:8699 0 chr20 271218 50 50M4700NS0M * 0 0 0
0 GTGGCTCTCCACAGGAATGTTGAGGATGACATCCATGTCTGGGGTGACTTGGCTCCGGAAGCAGAACATCCTCAATATGACCTCTCG
accepted_hits.sam
```

“RAW” DATA



Pipeline



RESULTS

	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

Data Stewardship: The Reality



We just need to migrate the data from these systems to fit into that hole over there.



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

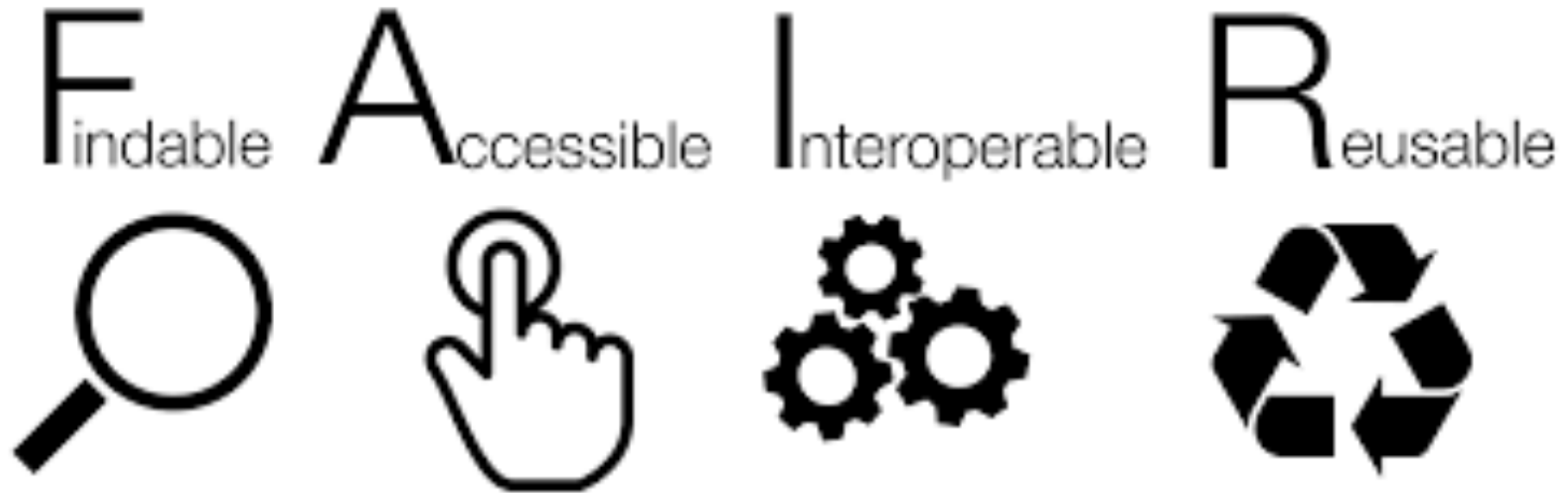


Graduate students



Post-doctoral fellows ¹⁶

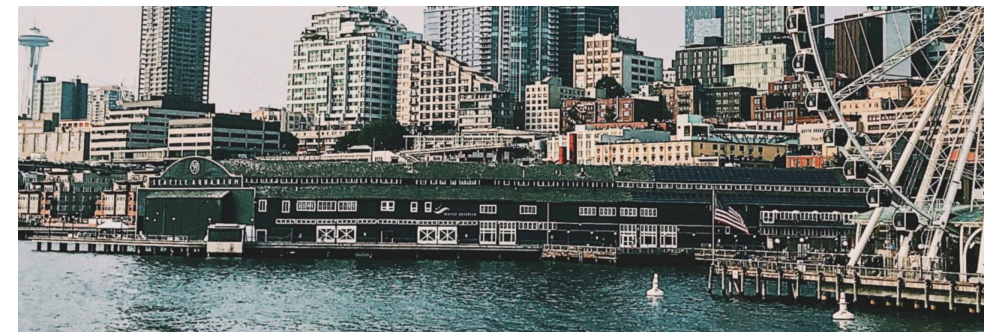
Data Stewardship: The Ideal



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>

Open Science and the Role of Common Evidence

- Whose science?
 - Community practices
 - Investigators' hypotheses, theories, models, methods, code...
- Whose data?
 - Global, comparative, fungible
 - Local, integrative, specific
- Whose evidence?
 - Common, cumulative, collaborative
 - Data creators' advantage



Acknowledgements



Christine Borgman



Bernie Boscoe



Peter Darch



Milena Golshan



Irene Pasquetto



Michael Scroggins



Cheryl Thompson



Morgan Wofford