UCLA
Presentations

Title
Big Data, Little Data, or No Data? Systematic Reviews in an Age of Open Data

Permalink
https://escholarship.org/uc/item/9wh1n7jn

Authors
Borgman, Christine L.
Pasquetto, Irene V.

Publication Date
2018-09-15

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

eScholarship.org Powered by the California Digital Library
University of California

Big Data, Little Data, or No Data?
Systematic Reviews in an Age of Open Data

Christine L. Borgman, PhD

Distinguished Research Professor
Center for Knowledge Infrastructures
University of California, Los Angeles
http://christineborgman.info
https://knowledgeinfrastructures.gseis.ucla.edu
@scitechprof

Irene V. Pasquetto, PhD

UCLA Center for Knowledge Infrastructures


Christine Borgman


Bernie Boscoe


Peter Darch


Milena Golshan


Irene Pasquetto


Michael Scroggins


Cheryl Thompson


Morgan Wofford

Keynote Presentation, Menti 353774
Cochrane Colloquium
Edinburgh, 15 September 2018

Cochrane
Trusted evidence.
Informed decisions.
Better health.

UCLA Center for Knowledge Infrastructures

Data sharing policies

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

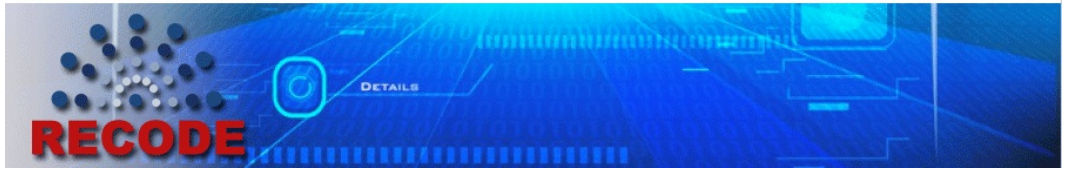
Supported by wellcome trust

Australian Government
National Health and Medical Research Council

National Science Foundation
WHERE DISCOVERIES BEGIN

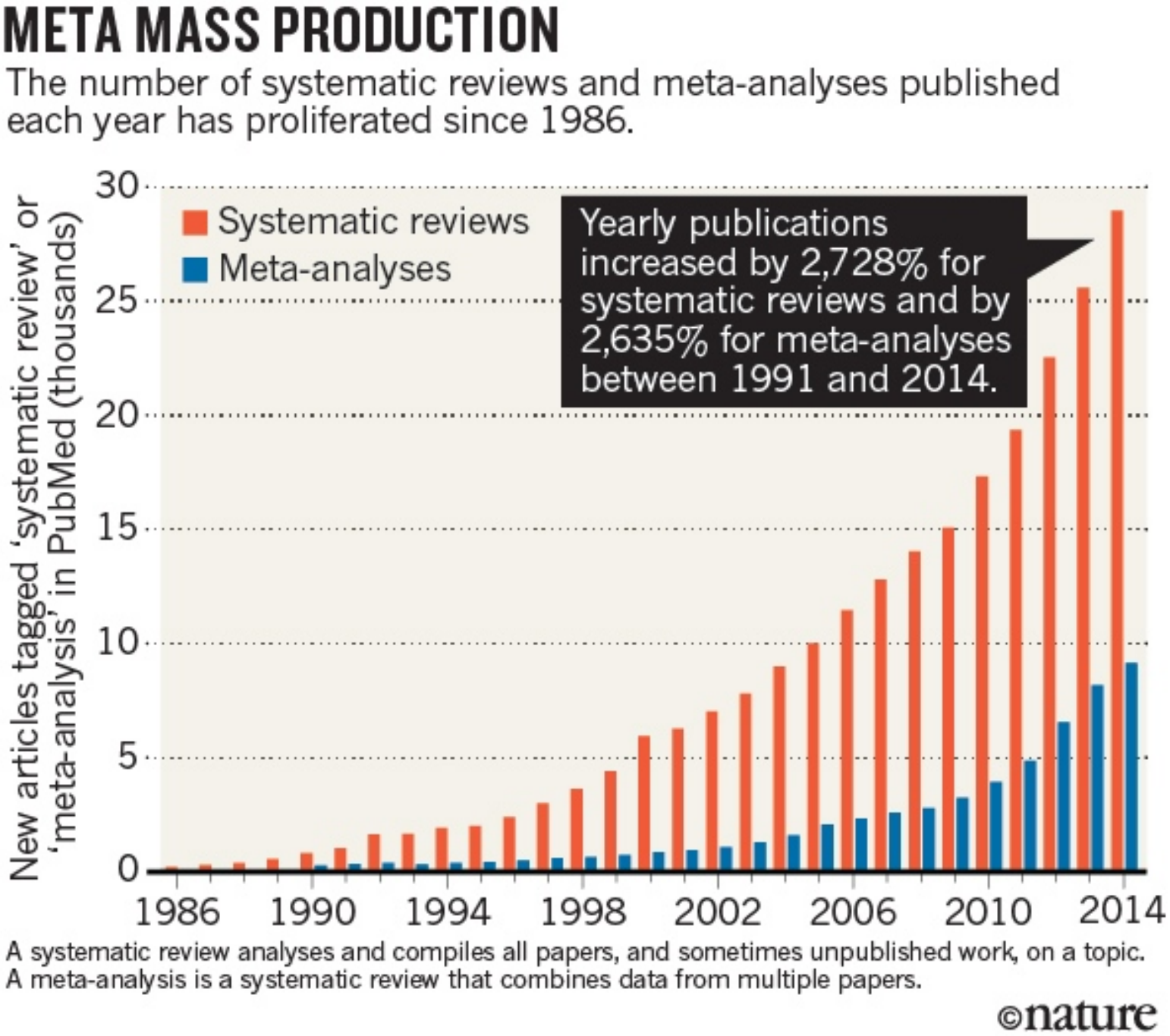
National Institutes of Health
Turning Discovery Into Health

Policy RECommendations for Open Access to Research Data in Europe



Menti 353774

Evidence-based medicine



Baker, M. (2016). Mass production of review articles is cause for concern. Nature News. https://doi.org/10.1038/nature.2016.20617

THE MILBANK QUARTERLY
A MULTIDISCIPLINARY JOURNAL OF POPULATION HEALTH AND HEALTH POLICY

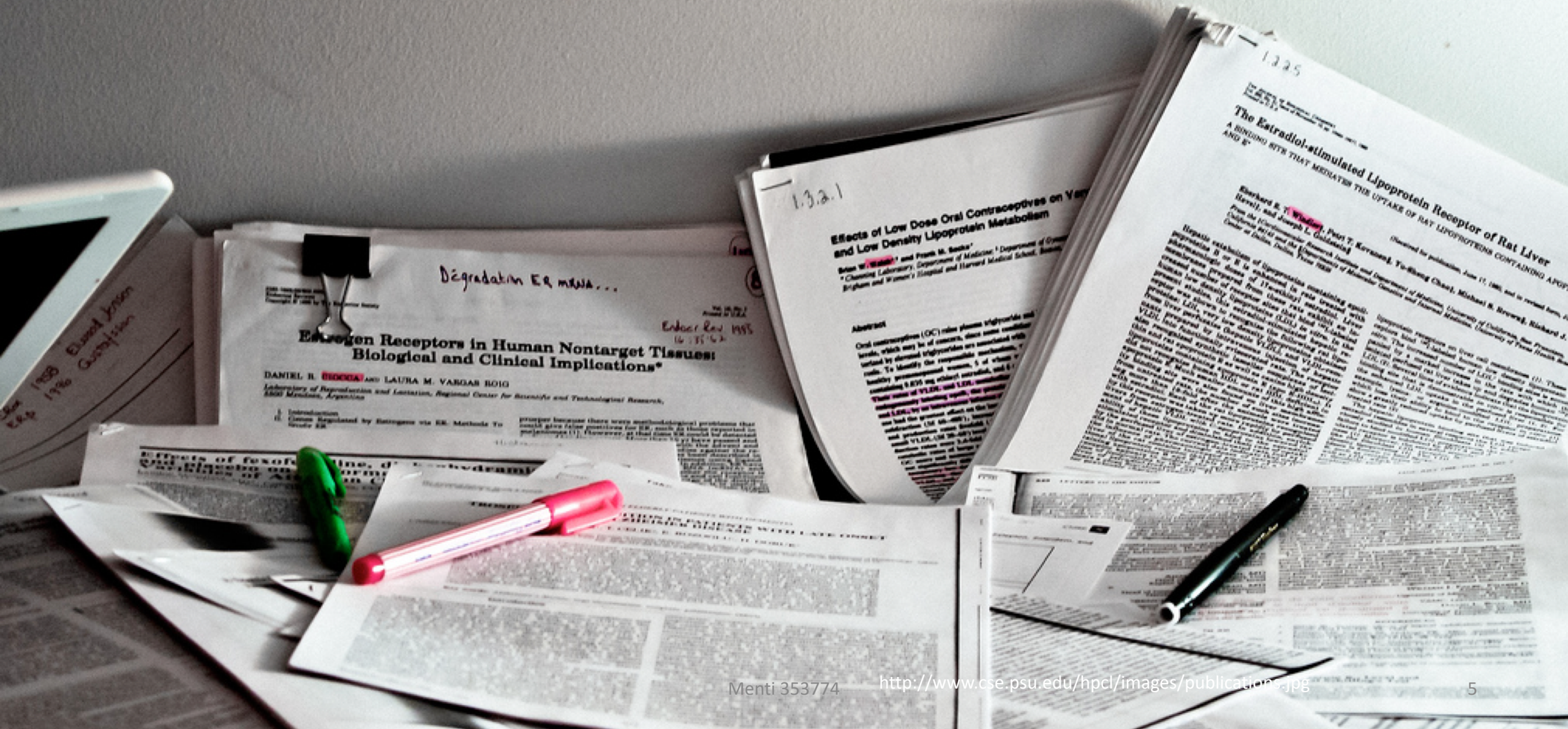
Original Investigation

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: 80

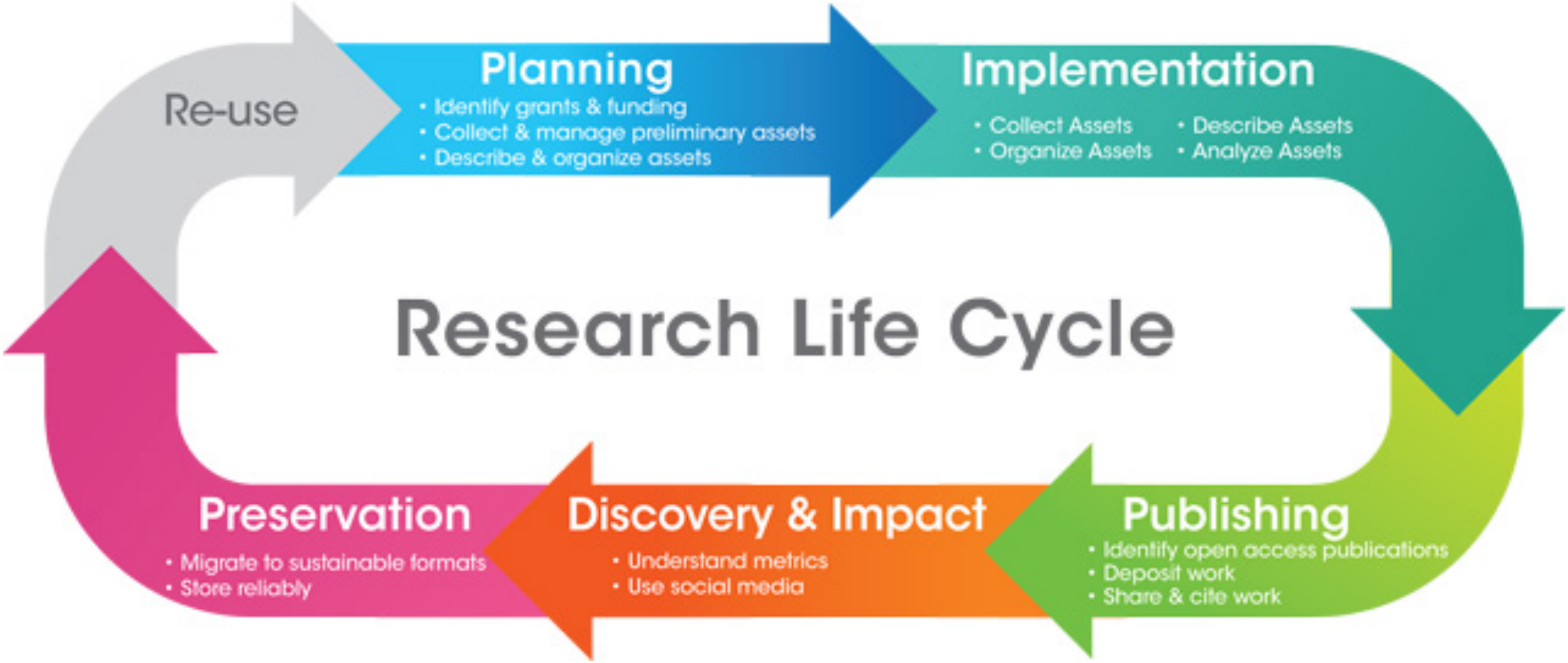
Publications



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


Data

Data creation and reuse: The Ideal



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

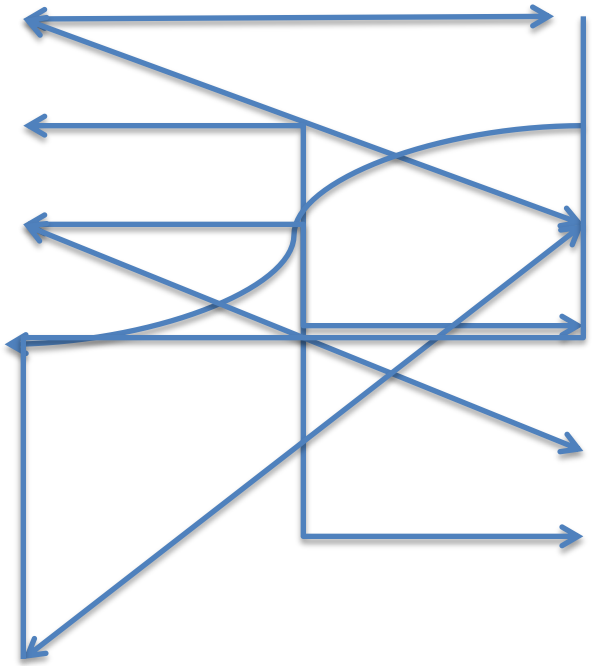
Publications <–> Data: Role

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

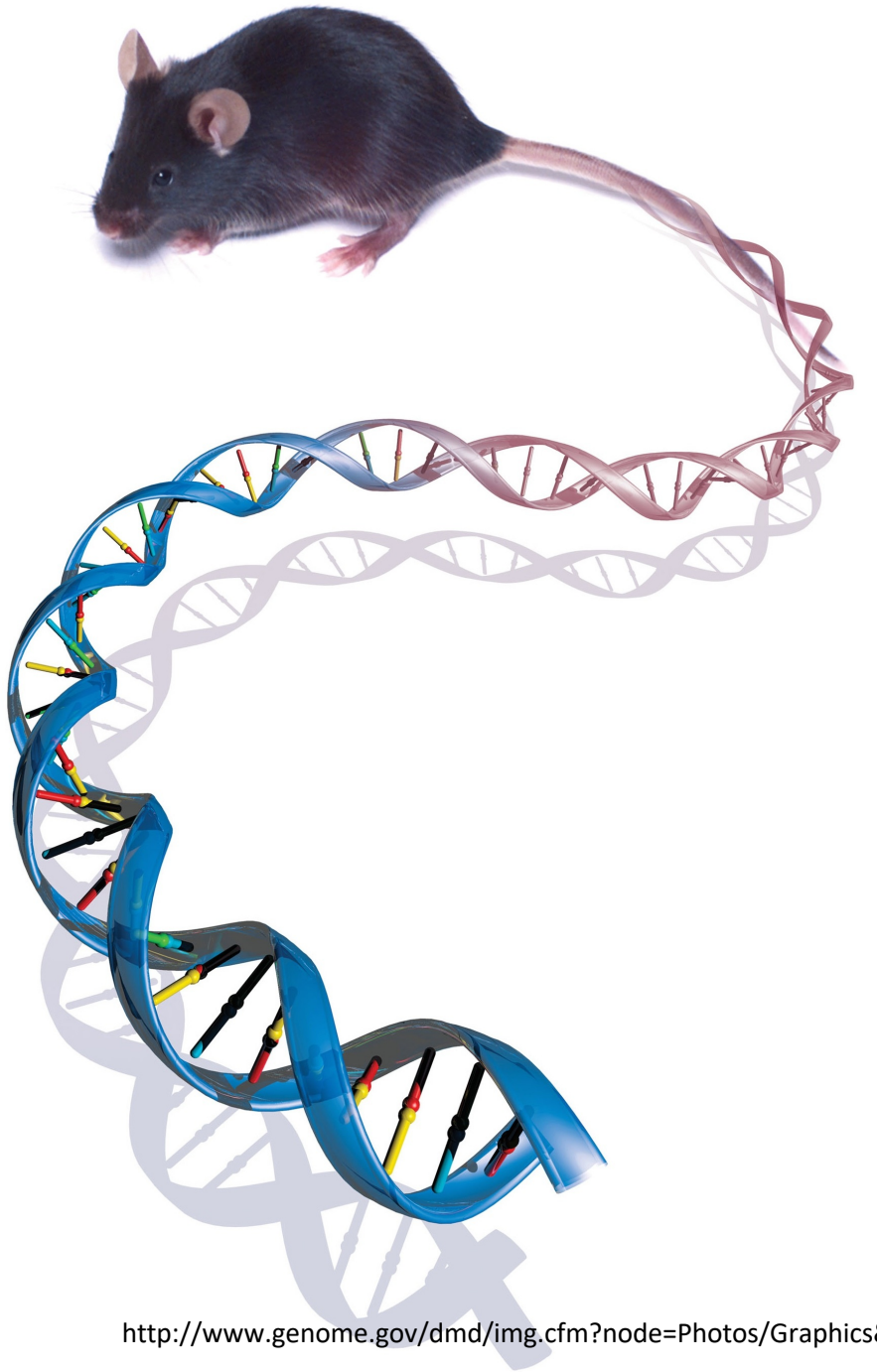
Publications <–> Data: Mapping

- 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



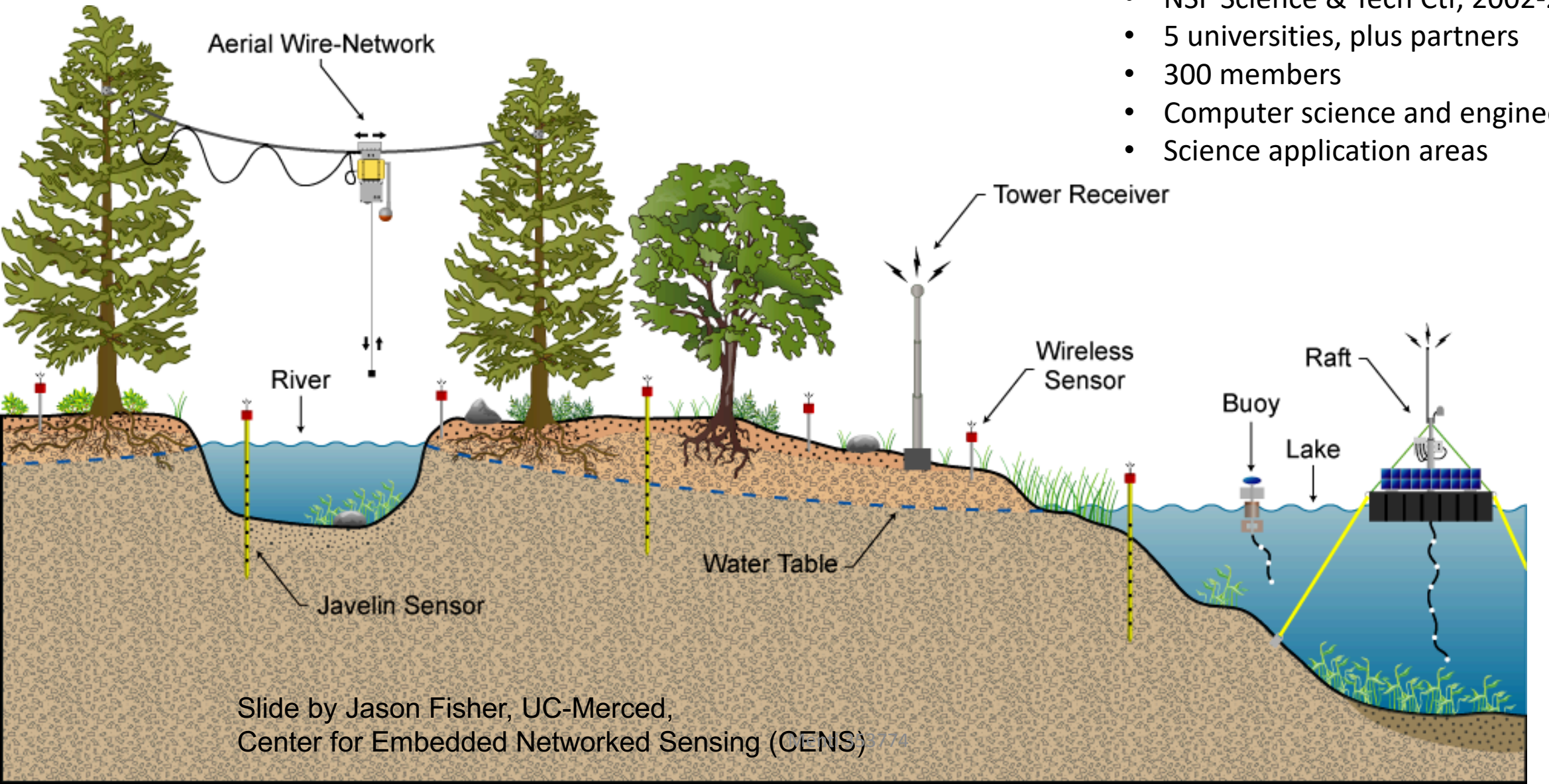
Data are representations of observations, objects, or other entities used as evidence of phenomena for the purposes of research or scholarship.

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

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

Center for Embedded Networked Sensing

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

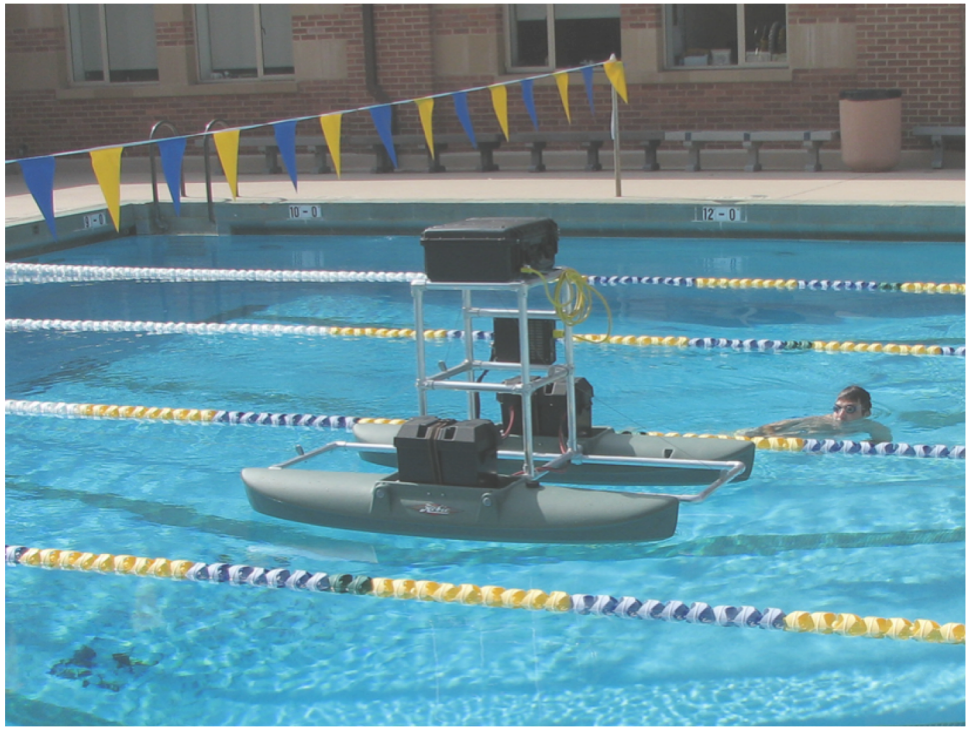


Slide by Jason Fisher, UC-Merced,
Center for Embedded Networked Sensing (CENS)

Science <–> Data

Engineering researcher:

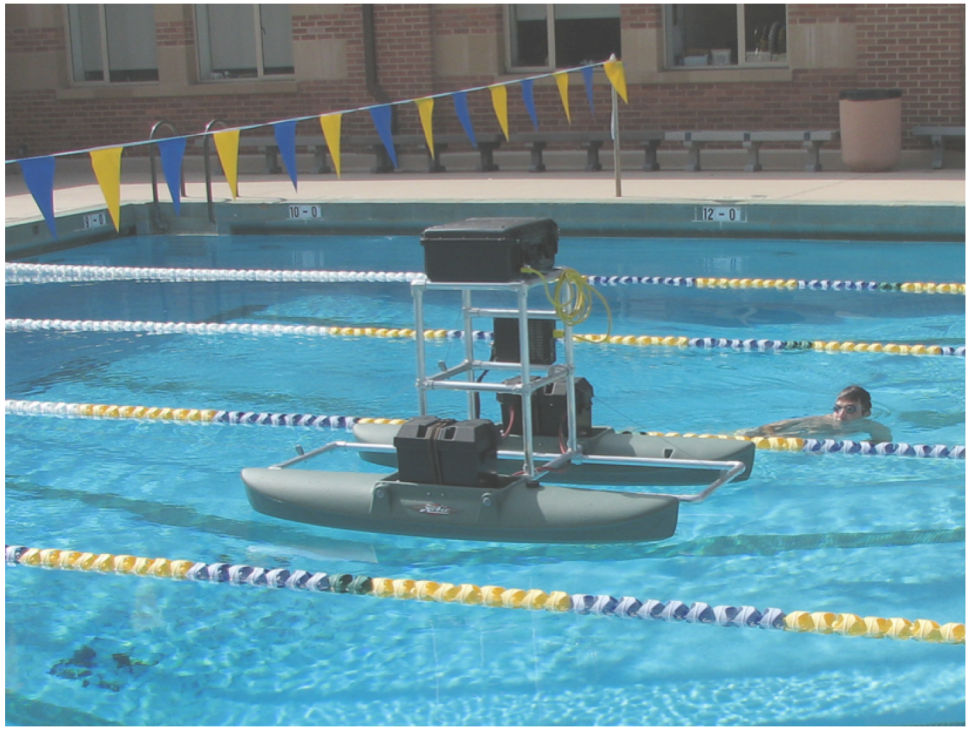
“Temperature is temperature.”



CENS Robotics team

Science <–> Data

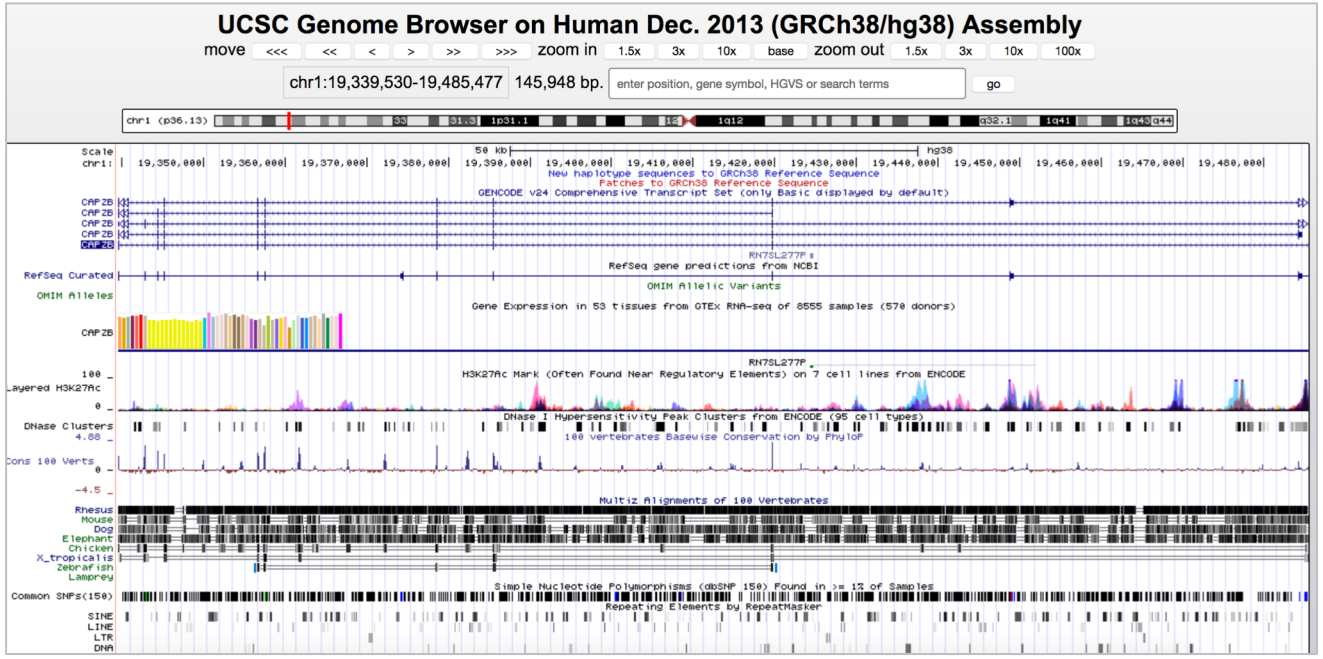
Engineering researcher:
"Temperature is temperature."



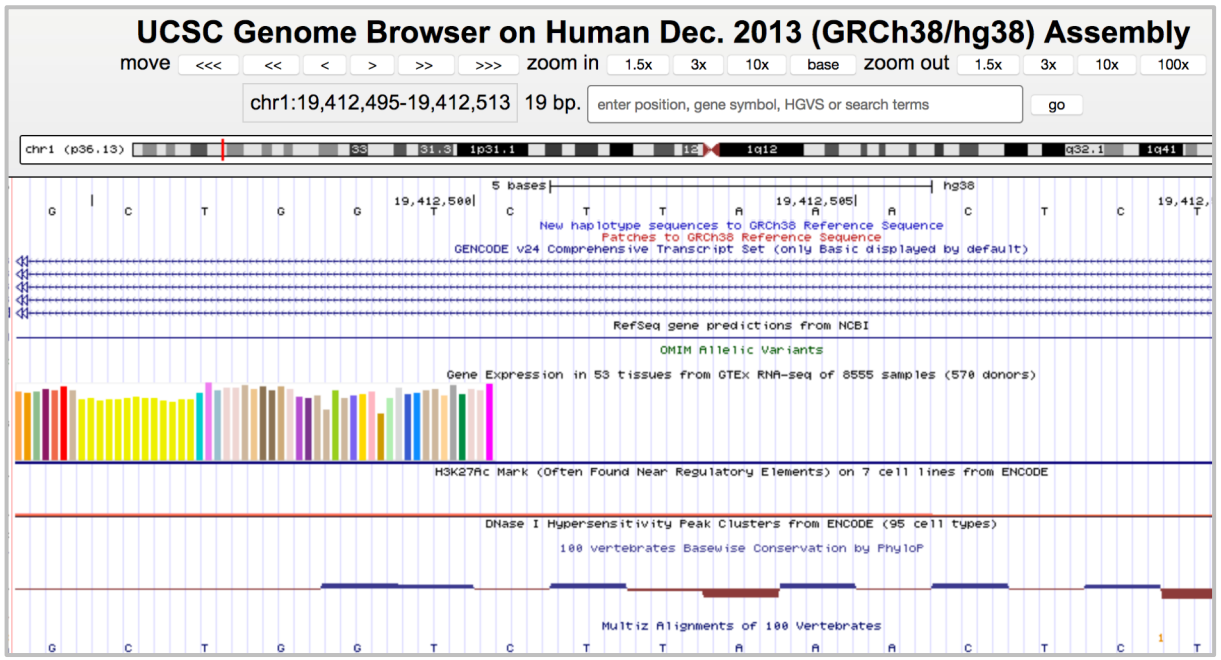
CENS Robotics team

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.."

Background Reuse in Biomedicine: Comparison, control, verification



UCSC Genome Browser – Search example (CAPZB gene)



UCSC Genome Browser – Zoom IN

Menti 353774

Foreground Reuse:
Hypothesis Testing and Statistical Analysis

Aligner software pairs “reads” using reference assemble genome

Data processing tool summarizes BAM information to compute likelihood of each possible genome

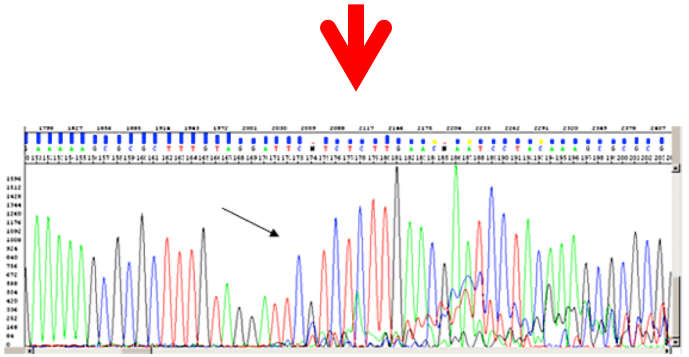
In-house script takes the ratio of mutant and allele frequencies to find the highest peak

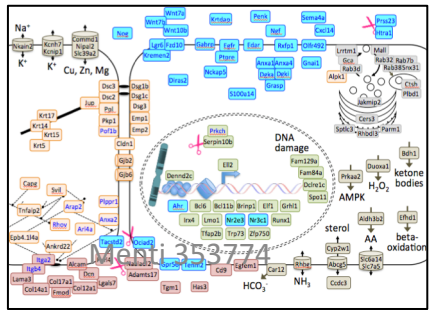
R studio calculates elative frequency and generate plotting graphs

Annotation tool predicts consequences of gene function

Variants are annotated by gene names, variant impact, and type of variant

```
@HD     VN:1.0  SO: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-rea
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_GTGAAA_L007_R1_001.fastq
HWI-ST1145:74:C101DACXX:7:1102:4284:73714        16      chr20   190930  3       100M    *       0       0
       CCGTGTTTAAAGGTGGATGCGGTCACCTTCCCAGCTAGGCTTAGGGATTCTTAGTTGGCCTAGGAAATCCAGCTAGTCCTGTCTCTCAGTCCCCCCTCT
C    BBDCCDDCCDDDDCDDDDDDCDCCCDBC?DDDDDDDDDDDDDDDDCCDCDDDDDDDDDDCCCCEDDDC?DDDDDDDDDDDDDDDDDDDDDBDHFFFFDC@@
    AS:i:-15        XM:i:3  XO:i:0  XG:i:0  MD:Z:55C20C13A9 NM:i:3  NH:i:2  CC:Z:=  CP:i:55352714   HI:i:0
HWI-ST1145:74:C101DACXX:7:1114:2759:41961        16      chr20   193953  50      100M    *       0       0
       TGCTGGATCATCTGGTTAGTGGCTTCTGACTCAGAGGACCTTCGTCCCCTGGGGCAGTGGACCTTCCAGTGATTCCCCTGACATAAGGGGCATGGACGA
G    DCDDDDEDDDDDDDCDDDDDDDCCCDDDCDDDDDEEC>DFFFEJJJJJIGJJJJIHGBHHGJIJJJJJJGJJJIJJJJJJIHJJJJJJHHHHHFFFFFCCC
    AS:i:-16        XM:i:3  XO:i:0  XG:i:0  MD:Z:60G16T18T3 NM:i:3  NH:i:1
HWI-ST1145:74:C101DACXX:7:1204:14760:4030        16      chr20   270877  50      100M    *       0       0
       GGCTTTATTGGTAAAAAAGGAATAGCAGATTTAATCAGAAATTCCCACCTGGCCCAGCAGCACCAACCAGAAAGAAGGGAAGAAGACAGGAAAAAACCA
C    DDDDDDDDDCCDDDDDDDDDDEEEEEEEEFFFEFFEGHHHHFGDJJIHJJIJIJJJIIIIGGFJJIHIIIIIJJJJJJIGHHFAHGFHJHFGGHFFFDD@BB
    AS:i:-11        XM:i:2  XO:i:0  XG:i:0  MD:Z:0A85G13    NM:i:2  NH:i:1
HWI-ST1145:74:C101DACXX:7:1210:11167:8699        0       chr20   271218  50      50M4700N50M     *       0
       0       GTGGCTCTTCCACAGGAATGTTGAGGATGACATCCATGTCTGGGGTGCACTTGGGTCTCCGAAGCAGAACATCCTCAAATATGACCTCTCG
accepted_hits.sam
```





“RAW” DATA

Pipeline

RESULTS

	BACKGROUND Reuse of Data	**FOREGROUND Reuse of Data**
Goal of reuse	"Ground truthing:" calibrate, compare, confirm	Analysis: identify patterns, correlations, causal relationships
Example of reuse	Instrument calibration, sequence annotation, review summary-level data	Meta-analyses, novel statistical analyses
Frequency of reuse	Frequent, routine practice	Rare, emergent practice

Pasquetto, I. V., Borgman, C. L., & Wofford, M. F. (in review). The Who, What, When, and Why of Reusing Data in Scientific Practice. *Harvard Data Science Review*.

	BACKGROUND Reuse of Data	FOREGROUND Reuse of Data
Goal of reuse	"Ground truthing:" calibrate, compa… confirm	Analyses: identi… patterns, co… …ns, causal r…
Example of reuse	Instrume… sequ… …on, r… …mary-level da…	Met… …s.
Frequency of reuse	Freq…nt - routine practice	Rare - emergent practice

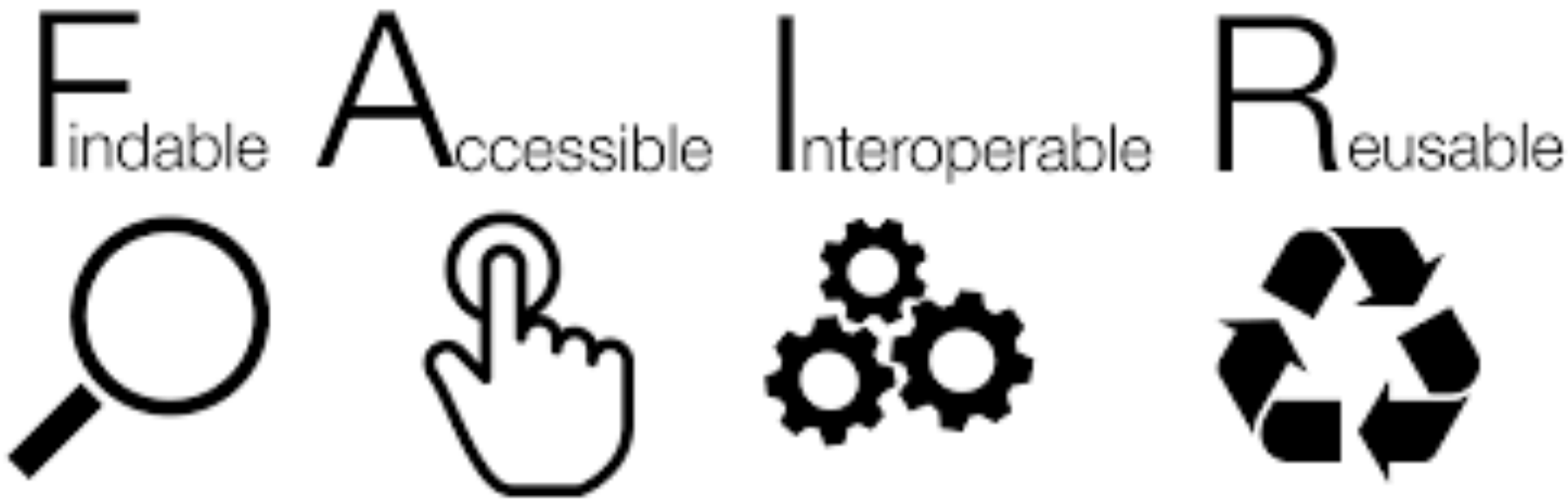
INDEPENDENT REUSE OF DATA

COLLABORATIVE REUSE WITH DATA CREATORS

Pasquetto, I. V., Borgman, C. L., & Wofford, M. F. (in review). The Who, What, When, and Why of Reusing Data in Scientific Practice. *Harvard Data Science Review*.

Menti 25 8774

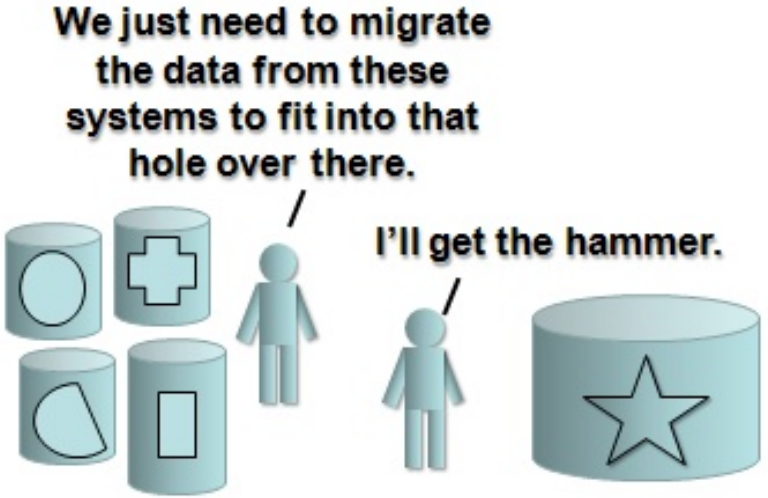
Data Stewardship: The Ideal



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/



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

Mount Wilson Solar Observatory, 2017

Getty Research Institute

http://gsa.rice.edu/

Graduate students

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

Post-doctoral fellows

Implications for Cochrane

- What are the opportunities in open data reviews?
  - Background reuse for broader surveys
  - Foreground reuse for new knowledge production
- What are the threats in data integration?
  - Interpretation, provenance, data cleaning, statistical error, …
  - Investment in skills and resources
  - Data stewardship commitments
- How can Cochrane collaborate with data creators to improve systematic reviews and metanalyses?

Acknowledgements

UCLA Center for Knowledge Infrastructures

ALFRED P. SLOAN FOUNDATION 1934


Christine Borgman


Bernie Boscoe


Peter Darch


Milena Golshan


Irene Pasquetto


Michael Scroggins


Cheryl Thompson


Morgan Wofford