Open and Reproducible Research: Goals, Obstacles, and Solutions

Matthew J. Salganik
Department of Sociology and Office of Population Research
Princeton University

January 6, 2016
OPR Workshop

For other OPR workshops:
http://opr.princeton.edu/workshops

Imagine a world in which every single person on the planet is given free access to the sum of all human knowledge. – Jimmy Wales

http://commons.wikimedia.org/wiki/File:The_Earth_seen_from_Apollo_17.jpg

Imagine a world where you have the job of your dreams.

http://commons.wikimedia.org/wiki/File:Nassau_Hall_Princeton.JPG

Open and reproducible research

Background

- caveats
- introductions

Background

Chatham House Rule: "When a meeting, or part thereof, is held under the Chatham House Rule, participants are free to use the information received, but neither the identity nor the affiliation of the speaker(s), nor that of any other participant, may be revealed."

More information:
https://www.chathamhouse.org/about/chatham-house-rule

A proposed working standard for open and reproducible research

For each published paper:

- code is available
- data is available
- paper is available

A proposed working standard for open and reproducible research

For each published paper:

- code is available
- data is available
- paper is available

There are complications and qualifications, and we'll address those later.

Introduction

HARVARD gazette

SCIENCE & HEALTH > HEALTH & MEDICINE

Ebola genomes sequenced

Speedy analysis reveals mutations, insights into outbreak, along with clues to origin, spread

August 28, 2014 |

Introduction

"By making the data immediately available to the community, we hope to accelerate response efforts," said co-senior author Pardis Sabeti, a senior associate member at the

Introduction

The New York Times | http://nyti.ms/11TfPht

The Opinion Pages | OP-ED CONTRIBUTOR

The Problem With Prostate Screening

By RICHARD J. ABLIN NOV. 25, 2014

Introduction

TUCSON — SCIENTIFIC data from clinical trials provides the foundation of medical decision making, from a doctor's prescription pad to sweeping public health policies. Public trust that the data is accurate and unbiased is the glue that binds our $3 trillion health care system. I worry that this trust, particularly when it comes to American men and their physicians and screening programs for prostate cancer, is now at risk.

Introduction

In 1970 I discovered the prostate-specific antigen, or PSA, which is now the most widely used tool in prostate screenings. But there has been a growing concern about whether the use of the PSA test has led to overdiagnosis and overtreatment, with millions of unnecessary surgeries, complications and deaths.

Nevertheless, the medical community has roundly embraced the results of a recent study finding that PSA screening reduced prostate cancer deaths by 20 percent. The study, the European Randomized Study of Screening for Prostate Cancer, joined another survey, the so-called Swedish Goteborg study (the results of which provided a basis for the European Randomized Study), which found an astounding 44 percent reduction.

Introduction

But there's a big problem with both of these studies: In March the Goteborg study's authors announced in the British Medical Journal that their data "are not available to outside investigators."

That the researchers would block access to government- and charity-supported research is bad enough. Even worse, it calls into question why, if the data was strong, the researchers wouldn't open it up to independent scrutiny.

Introduction

Not just in the biomedical sciences . . . .

Introduction

What Can We Learn about Neighborhood Effects from the Moving to Opportunity Experiment?[1]

Jens Ludwig
University of Chicago

Jeffrey B. Liebman
Harvard University

Jeffrey R. Kling
Brookings Institution

Greg J. Duncan
University of California, Irvine

Lawrence F. Katz
Harvard University

Ronald C. Kessler
Harvard Medical School

Lisa Sanbonmatsu
National Bureau of Economic Research

Neighborhood Effects on Economic Self-Sufficiency: A Reconsideration of the Moving to Opportunity Experiment[1]

Susan Clampet-Lundquist
Saint Joseph's University

Douglas S. Massey
Princeton University

Moving to Inequality: Neighborhood Effects and Experiments Meet Social Structure[1]

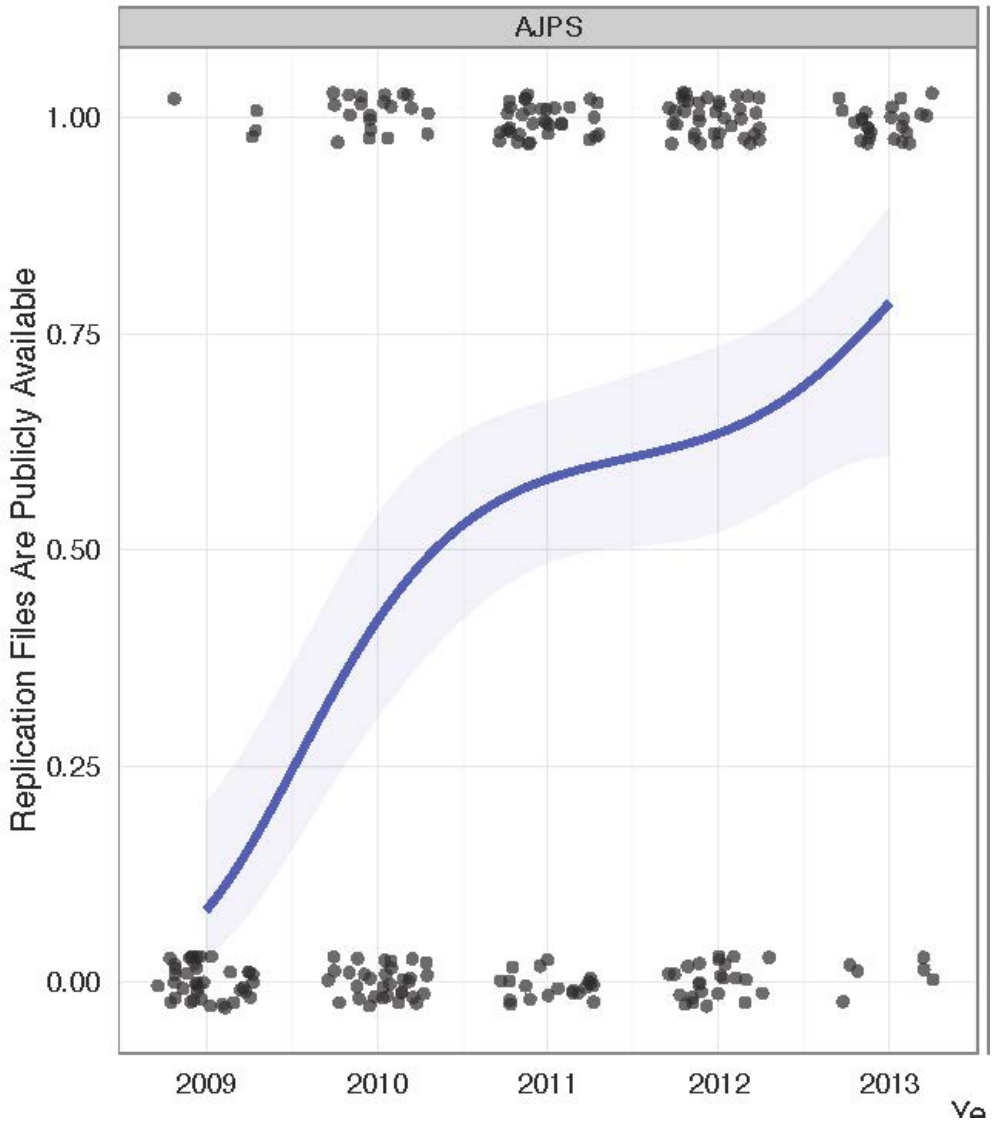
Robert J. Sampson
Harvard University

Made possible with open data

Current system is historical artifact.

http://commons.wikimedia.org/wiki/File:IBM_card_storage.NARA.jpg

But, change is coming. . . .



http://isps.yale.edu/news/blog/2013/09/the-imperative-to-share-complete-replication-files

- open and reproducible research is about making us better scientists
- open and reproducible research is not about advancing your career by bringing others down

Questions? Comments?

A proposed working standard for open and reproducible research

For each published paper:

- code is available
- data is available
- paper is available

Code is available

Understandable code that reproduces all the numbers, tables, and figures in your paper

- someone like me could understand the code in one afternoon
- does not need to include every piece of code you wrote for the project
- does not need to be beautiful; coding is about trade-offs
- code should turn rawest data into final results

Code is available

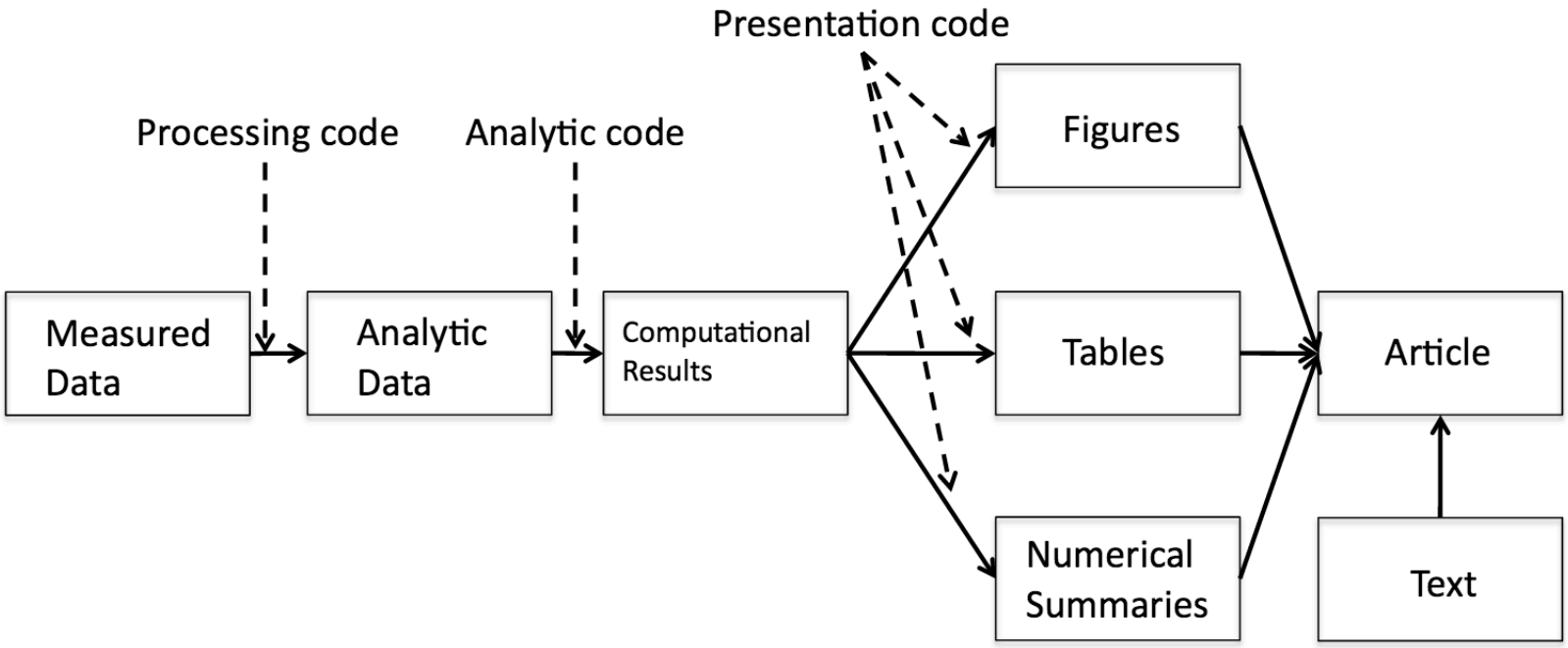


Image from presentation by Roger Peng

Code is available

Further reading:

- Publish your computer code: it is good enough by Barnes, *Nature*, 2010.
- A Decade of Replications: Lessons from the *Quarterly Journal of Political Science* by Eubank, Blog post, 2014.
- Reproducible Research: A View from the Social Sciences by Marwick, Presentation, 2014.

Code is available

Questions about making your code available . . .

A proposed working standard for open and reproducible research

For each published paper:

- code is available
- data is available
- paper is available

Data is available

Documented files that enables others to reproduce all figures, tables, and numbers

- someone like me could start using the data in one afternoon
- avoid propriety formats
- bonus points for releasing extra variables that are not need to reproduce specific analysis

Data is available

Documented files that enables others to reproduce all figures, tables, and numbers

- someone like me could start using the data in one afternoon
- avoid propriety formats
- bonus points for releasing extra variables that are not need to reproduce specific analysis

But . . .

- potentially creates ethical issues: it is very difficult to de-anonymize data

Data is available

Risks come from combining data sources

$$\underbrace{\text{Baking soda}}_{\text{Safe}} + \underbrace{\text{Vinegar}}_{\text{Safe}} =$$

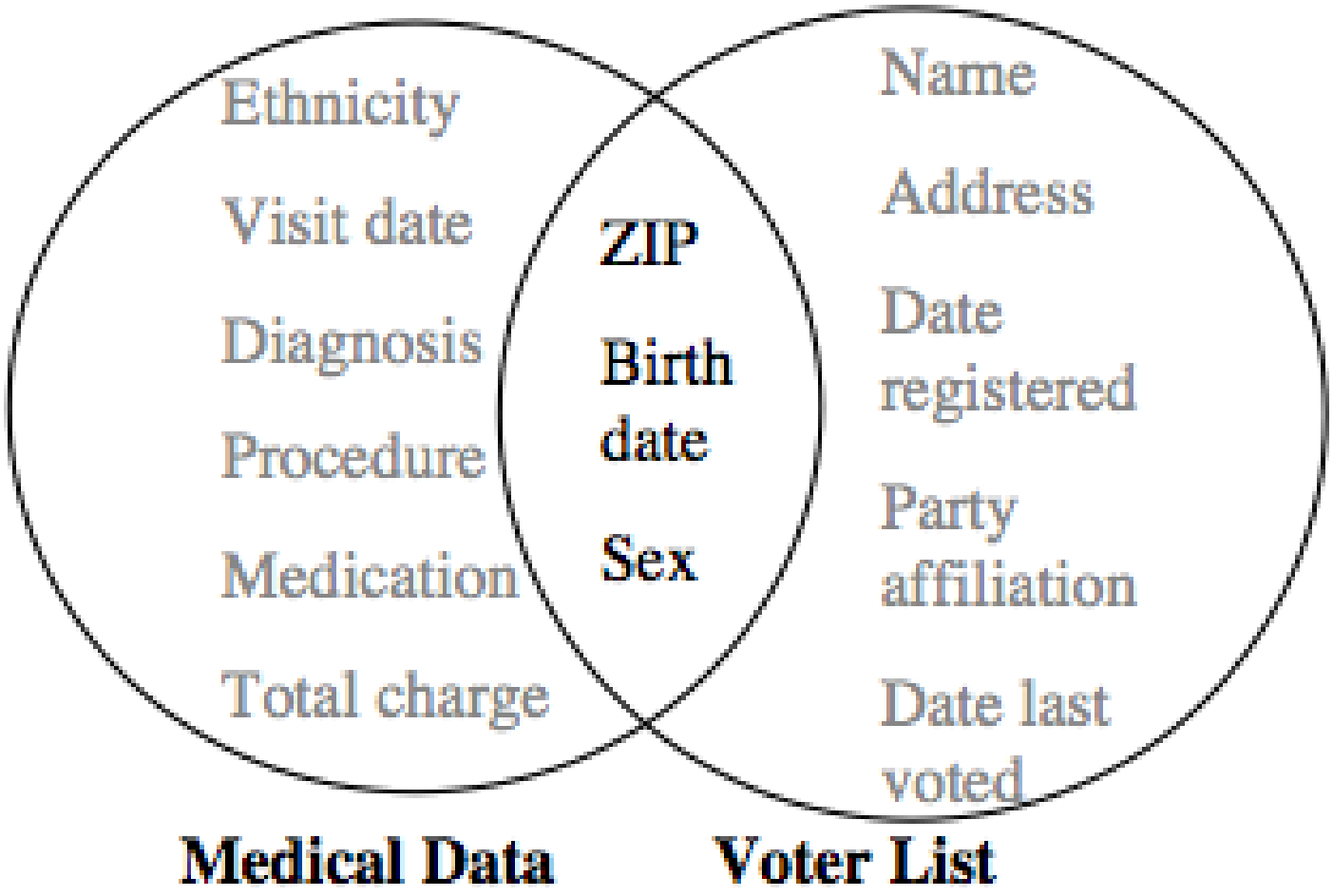
Data is available

Risks come from combining data sources

$$\underbrace{\text{Baking soda}}_{\text{Safe}} + \underbrace{\text{Vinegar}}_{\text{Safe}} =$$

https://www.flickr.com/photos/edenpictures/15962352215/

Data is available



Source: Sweeney (2002)

Data is available

To learn more:

- Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization by Ohm, *UCLA Law Review*, 2010.
- Privacy and Data-Based Research by Heffetz and Ligett, *Journal of Economic Perspectives*, 2014.
- No silver bullet: De-identification still doesn't work by Narayanan and Felten, *Working paper*, 2014.
- How to de-identify your data by Angiuli, Blitzstein, and Waldo, *Communications of the ACM*, 2015.

Data is available

Ways to manage the ethical dilemma

- consider data release from the beginning (consent form, IRB application, etc)
- learn about data anonymization (e.g., coarsening and hashing)
- submit your plan to the IRB

Data is available

Ways to manage the ethical dilemma

- consider data release from the beginning (consent form, IRB application, etc)
- learn about data anonymization (e.g., coarsening and hashing)
- submit your plan to the IRB

Research ethics involves both minimizing risks *and* maximizing benefit

Data is available

The Dataverse Network™ Project

A Web Application for Publishing, Citing, Analyzing and Preserving Research Data

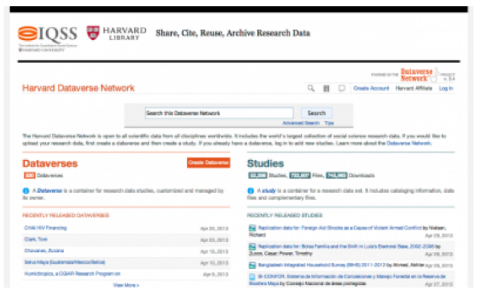
IQSS
The Institute for Quantitative Social Science
HARVARD UNIVERSITY

ABOUT | GETTING STARTED | BEST PRACTICES | SOFTWARE | APPS | GUIDES

The Dataverse Network project develops software, protocols, and community connections for creating research data repositories that automate professional archival practices, guarantee long term preservation, and enable researchers to share, retain control of, and receive web visibility and formal academic citations for their data contributions.

Share & Find Data

The **Harvard Dataverse Network** is free* and open to all researchers worldwide to share, cite, reuse and archive research data.



Institutions or organizations may also choose to download the open source software for their own use. Here are some other **Dataverse Networks around the world**.

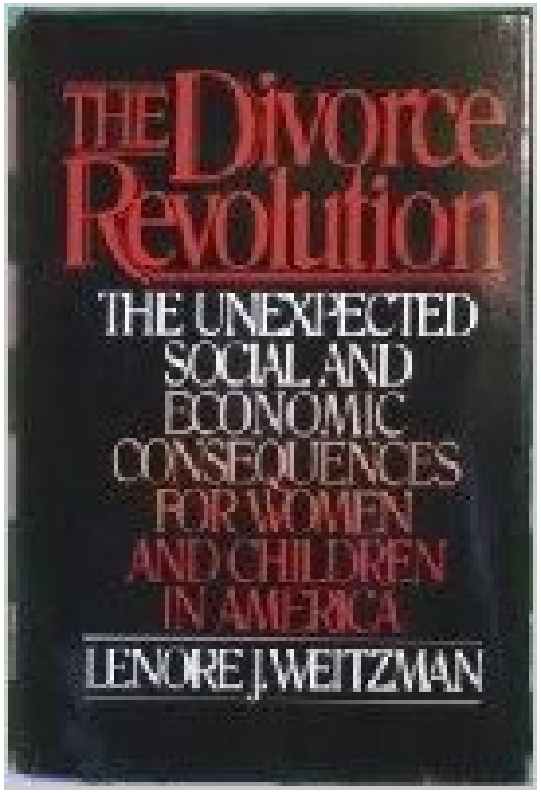
* *If you plan to upload >1TB please contact us.*

Data is available

Questions about making your data available . . .

Not having everything in order can come back to haunt you . . .

https://www.youtube.com/watch?v=66oNv_DJuPc



main empirical finding about changes in living standard after divorce

- for women declines 73%
- for men increases 42%

- American Sociological Association Book Award in 1986
- Between 1986 and 1993, cited in 348 social science articles and 250 law review articles
- Between 1986 and 1993, cited in 24 legal cases and *by* the Supreme Court
- Led to changes in divorce law in California

A RE-EVALUATION OF THE ECONOMIC CONSEQUENCES OF DIVORCE*

Richard R. Peterson
Social Science Research Council

Over the last 20 years, researchers have focused considerable attention on the economic consequences of divorce. One book, Weitzman's The Divorce Revolution *(1985), reports a 73 percent decline in women's standard of living after divorce and a 42 percent increase in men's standard of living. These percentages, based on data from a 1977–1978 Los Angeles sample, are substantially larger than those from other studies. I replicate* The Divorce Revolution*'s analysis and demonstrate that the estimates reported in the book are inaccurate. This reanalysis, which uses the same sample and measures of economic well-being as* The Divorce Revolution, *produces estimates of a 27 percent decline in women's standard of living and a 10 percent increase in men's standard of living after divorce. I discuss the implications of these results for debates about divorce law reform.*

"First, let me begin with Peterson's implied question: Was this responsible research and did I meet professional standards in analyzing these data?"

Weitzman (1996)

" . . . .Changes to the original raw data file resulting from this data cleaning process were made by a series of programming statements on a master SPSS system file. *The raw data file that is stored at the Murray Center is the original "dirty data" file and does not include these cleaning changes*. . . ."

Weitzman (1996), emphasis in original

"Unfortunately, the original cleaned master SPSS system file no longer exists. I assumed it was being copied and reformatted as I moved for job changes and fellowships from the project's original offices in Berkeley to Stanford (in 1979), then to Princeton (in 1983), back to Stanford (in 1984) and then to Harvard (in 1986). With each move, new programmers worked on the files to accommodate different computer systems."

Weitzman (1996), emphasis in original

"Before I left Stanford I instructed my programmers to prepare all my data files for archiving. I know now (but did not know then) that the original master SPSS system file that I used for my book had been lost or damaged at some point and was not included among these files. The SPSS system file that I thought was the master SPSS system file was the result of the merging of many smaller subfiles that had been created for specific analyses. It later became apparent that a programming error had been made, and the subfiles were not "keyed" correctly: Not all of the data from each individual respondent were matched on the appropriate case ID number, and data from different respondents were merged under the same case ID. At present it is not possible to disentangle exactly what mismatch occurred for any specific respondent."

Weitzman (1996)

"When I could not replicate the analyses in my book with what I had mistakenly assumed was the archived master SPSS system file, I hired an independent consultant, Professor Angela Aidala from Columbia University, to help me untangle what had happened. She reviewed all of the project files, documentation, and codebooks, as well as the available data and programming files to determine a possible computational error in the standard of living statistic. But she could not do this without an accurate data file to work with. We then went back to the original questionnaires and recoded a random sample of about 25 percent of the cases. There were so many discrepancies between the questionnaires and the "dirty data" raw data file, and between the questionnaires and the mismatched SPSS system file, that we finally abandoned the effort and left a warning to all future researchers *that both files at the Murray Center were so seriously flawed that they could not be used.* It was a very sad, time consuming, and frustrating experience. . . "

Weitzman (1996)

Lessons:

Lessons:

- Great that Weitzman released the data into an archive

Lessons:

- Great that Weitzman released the data into an archive
- You need to be able to reproduce results from *start* to *finish*

Lessons:

- Great that Weitzman released the data into an archive
- You need to be able to reproduce results from *start* to *finish*
- You need to be able to reproduce your results 11 years after they are published

Lessons:

- Great that Weitzman released the data into an archive
- You need to be able to reproduce results from *start* to *finish*
- You need to be able to reproduce your results 11 years after they are published
- This was harder in the past

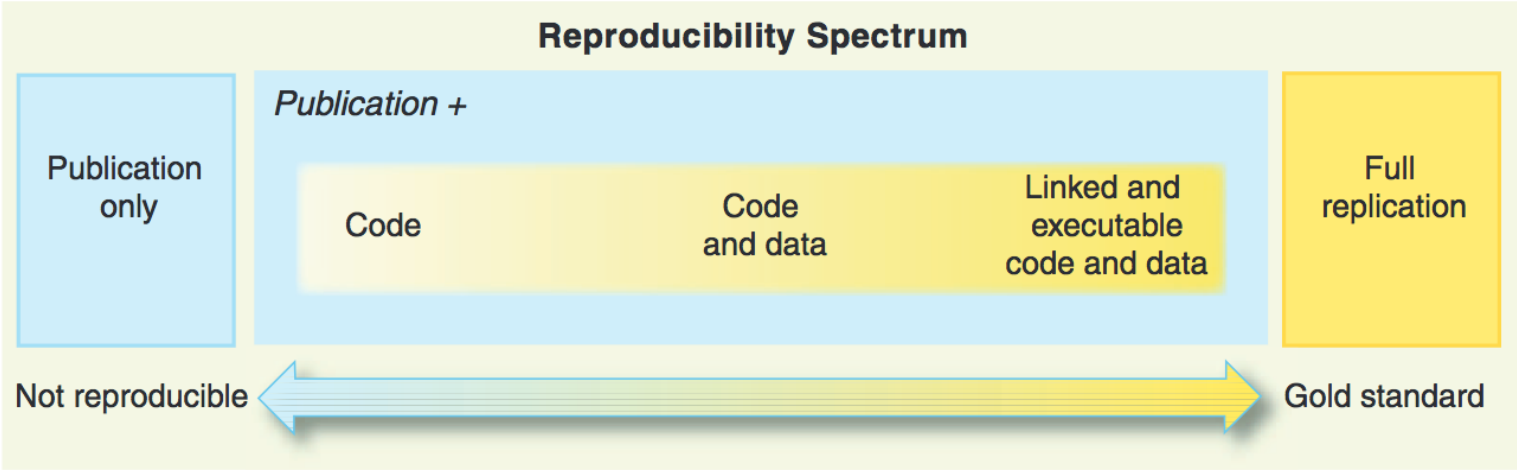
Lessons:

- Great that Weitzman released the data into an archive
- You need to be able to reproduce results from *start* to *finish*
- You need to be able to reproduce your results 11 years after they are published
- This was harder in the past
- You, not your RAs, are the one who is responsible

My personal experiences:

- releasing data and code
- using data and code from others

I've been everywhere on this spectrum:



Source: Peng (2011)

Releasing your code and data will force you to be better

http://commons.wikimedia.org/wiki/File:Odysseus_Sirens_BM_E440_n2.jpg

Wonderful advice about how to manage your data and code:

- Gentzkow and Shapiro (2014) Code and Data for the Social Sciences: A Practitioner's Guide, *Working paper*

Questions? Comments?

A proposed working standard for open and reproducible research

For each published paper:

- code is available
- data is available
- paper is available

Paper is available

Paper can be downloaded for free by anyone with an internet connection

Paper is available

Paper is available

THE LANCET

Volume 377, Issue 9778, 14–20 May 2011, Pages 1633–1635

Comment

Science as a public enterprise: the case for open data

Geoffrey Boulton[a], Michael Rawlins[b], Patrick Vallance[c], Mark Walport[d]

[a] Grant Institute, Edinburgh University, Edinburgh EH9 3JW, UK

[b] National Institute for Health and Clinical Excellence, London, UK

[c] GlaxoSmithKline, London, UK

[d] Wellcome Trust, London, UK

http://dx.doi.org/10.1016/S0140-6736(11)60647-8, How to Cite or Link Using DOI

https://www.flickr.com/photos/dullhunk/8653757940/

Paper is available

Do you want your work to be available to
everyone in the world or just to academics at rich universities?

Paper is available

Sarah K. Cowan
@SarahKCowan

Follow

My favorite citation yet! The #DNC cited my work in their amicus brief to #SCOTUS on #evenwel. scotusblog.com/wp-content/upl …

Her paper was published open access

Paper is available

Examples of people who are harmed by closed-access publication system

- public health researchers in developing countries
- public interest lawyers
- people with rare diseases fighting to get medical treatment

Paper is available

Why would a publisher – dedicated to spreading knowledge – try to hoard information?

Paper is available

Why would a publisher – dedicated to spreading knowledge – try to hoard information?

http://commons.wikimedia.org/wiki/File:Money_Cash.jpg

Paper is available

Options:

- gold open access (open access journals)
  - Fees: *Sociological Science*, *Socius*, *PLOS One*
  - No Fees: *Journal of Statistical Software*

Paper is available

Options:

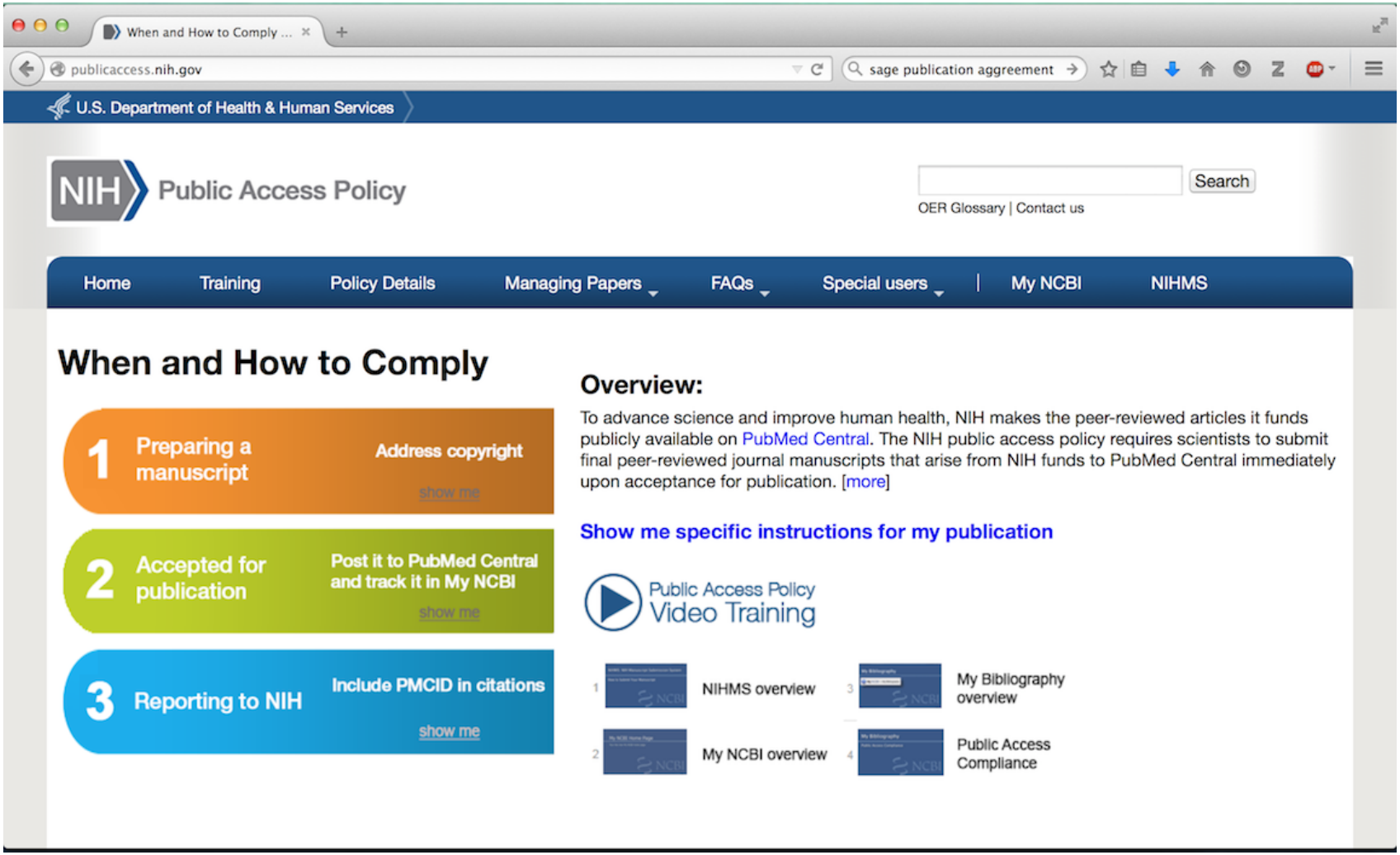
- gold open access (open access journals)
  - Fees: *Sociological Science*, *Socius*, *PLOS One*
  - No Fees: *Journal of Statistical Software*
- green open access (self archiving)
  - PubMed, arXiv.org, SSRN

Paper is available

Options:

- gold open access (open access journals)
  - Fees: *Sociological Science*, *Socius*, *PLOS One*
  - No Fees: *Journal of Statistical Software*
- green open access (self archiving)
  - PubMed, arXiv.org, SSRN
- open access per article

Paper is available



Paper is available

Overview:

To advance science and improve human health, NIH makes the peer-reviewed articles it funds publicly available on PubMed Central. The NIH public access policy requires scientists to submit final peer-reviewed journal manuscripts that arise from NIH funds to PubMed Central immediately upon acceptance for publication. [more]

Paper is available

To learn more:

- Talking about Open Access: SMASH and Subtler Tactics by Cirasella, *Presentation for Open Access Week*, 2014.
- Peer Review as a Service: It's not about the journal by Lintott et al., *Blog post*, 2014.
- Princeton Scholarly Communication Office
- "How Open Is It?" Open Access Spectrum (OAS)

Paper is available

Questions about making your paper available . . .

A proposed working standard for open and reproducible research

For each published paper:

- code is available
- data is available
- paper is available

Concerns

If this is so great, why isn't everyone doing it already?

Concerns

If this is so great, why isn't everyone doing it already?
I don't know. Here are some guesses:

- inertia (remember this was not easy 10 years ago)
- mis-estimation of costs and benefits

Concerns

Sociological Methods & Research
Volume 36 Number 2
November 2007 153-172

10.1177/0049124107306659
http://smr.sagepub.com
hosted at
http://online.sagepub.com

Replication Standards for Quantitative Social Science

Why Not Sociology?

Jeremy Freese
Northwestern University, Evanston, Illinois

The credibility of quantitative social science benefits from policies that increase confidence that results reported by one researcher can be verified by others. Concerns about replicability have increased as the scale and sophistication of analyses increase the possible dependence of results on subtle analytic decisions and decrease the extent to which published articles contain full descriptions of methods. The author argues that sociology should adopt standards regarding replication that minimize its conceptualization as an ethical and individualistic matter and advocates for a policy in which authors use independent online archives to deposit the maximum possible information for replicating published results at the time of publication and are explicit about the conditions of availability for any necessary materials that are not provided. The author responds to several objections that might be raised to increasing the transparency of quantitative sociology in this way and offers a candidate replication policy for sociology.

Concerns

Possible objections included in Freese (2007)

- Won't this mean more work for researchers?

Concerns

Possible objections included in Freese (2007)

- Won't this mean more work for researchers?
- Won't this mean more work for editors?

Concerns

Possible objections included in Freese (2007)

- Won't this mean more work for researchers?
- Won't this mean more work for editors?
- There are good reasons for researchers not to make data publicly available

Concerns

Possible objections included in Freese (2007)

- Won't this mean more work for researchers?
- Won't this mean more work for editors?
- There are good reasons for researchers not to make data publicly available
- There are good reasons for researchers not to make code publicly available

Concerns

Possible objections included in Freese (2007)

- Won't this mean more work for researchers?
- Won't this mean more work for editors?
- There are good reasons for researchers not to make data publicly available
- There are good reasons for researchers not to make code publicly available
- What about qualitative research?

Concerns

Possible objections included in Freese (2007)

- Won't this mean more work for researchers?
- Won't this mean more work for editors?
- There are good reasons for researchers not to make data publicly available
- There are good reasons for researchers not to make code publicly available
- What about qualitative research?
- Not enough interest exists in reproducing results to justify changes in existing policy

Concerns

When Firebaugh proposed replication standards for *American Sociological Review*:

- The freeloading problem: Why should I go to the effort to obtain grants and collect my own data if I am then required to share my data with others?
- The "I-might-be-scooped" problem: Not only will there be freeloaders, but they might become famous at my expense by publishing key results before I am able to.
- The question of qualitative research: Should qualitative research be held to the same standards? If so, how? If not, why not?
- Too much work: The extra work for authors and editors would be onerous.

Concerns

- Freese (2007) Replication Standards for Quantitative Social Science: Why Not Sociology, *Sociological Methods & Research*.
- King (2007) An Introduction to the Dataverse Network as an Infrastructure for Data Sharing, *Sociological Methods & Research*.
- Firebaugh (2007) Replication Data Sets and Favored-Hypothesis Bias: Comment on Jeremy Freese (2007) and Gary King (2007), *Sociological Methods & Research*.
- Abbott (2007) Notes on Replication, *Sociological Methods & Research*.
- Freese (2007) Overcoming Objections to Open-Source Social Science, *Sociological Methods & Research*.

Top six lessons from my own struggles with these issues:

Top six lessons from my own struggles with these issues:

- This will be hard at first, but then it will become easier

Top six lessons from my own struggles with these issues:

- This will be hard at first, but then it will become easier
- This is probably work that you should be doing anyway

Top six lessons from my own struggles with these issues:

- This will be hard at first, but then it will become easier
- This is probably work that you should be doing anyway
- Whatever you do, it will not be perfect; don't let that stop you

Top six lessons from my own struggles with these issues:

- This will be hard at first, but then it will become easier
- This is probably work that you should be doing anyway
- Whatever you do, it will not be perfect; don't let that stop you
- There is no single right way; there are many reasonable ways

Top six lessons from my own struggles with these issues:

- This will be hard at first, but then it will become easier
- This is probably work that you should be doing anyway
- Whatever you do, it will not be perfect; don't let that stop you
- There is no single right way; there are many reasonable ways
- You must make this decision at the beginning of your project not the end

Top six lessons from my own struggles with these issues:

- This will be hard at first, but then it will become easier
- This is probably work that you should be doing anyway
- Whatever you do, it will not be perfect; don't let that stop you
- There is no single right way; there are many reasonable ways
- You must make this decision at the beginning of your project not the end
- Once you start, you will never go back

There are no insurmountable obstacles
preventing you from doing open and reproducible research

You can choose what kind of scholar you want to be.