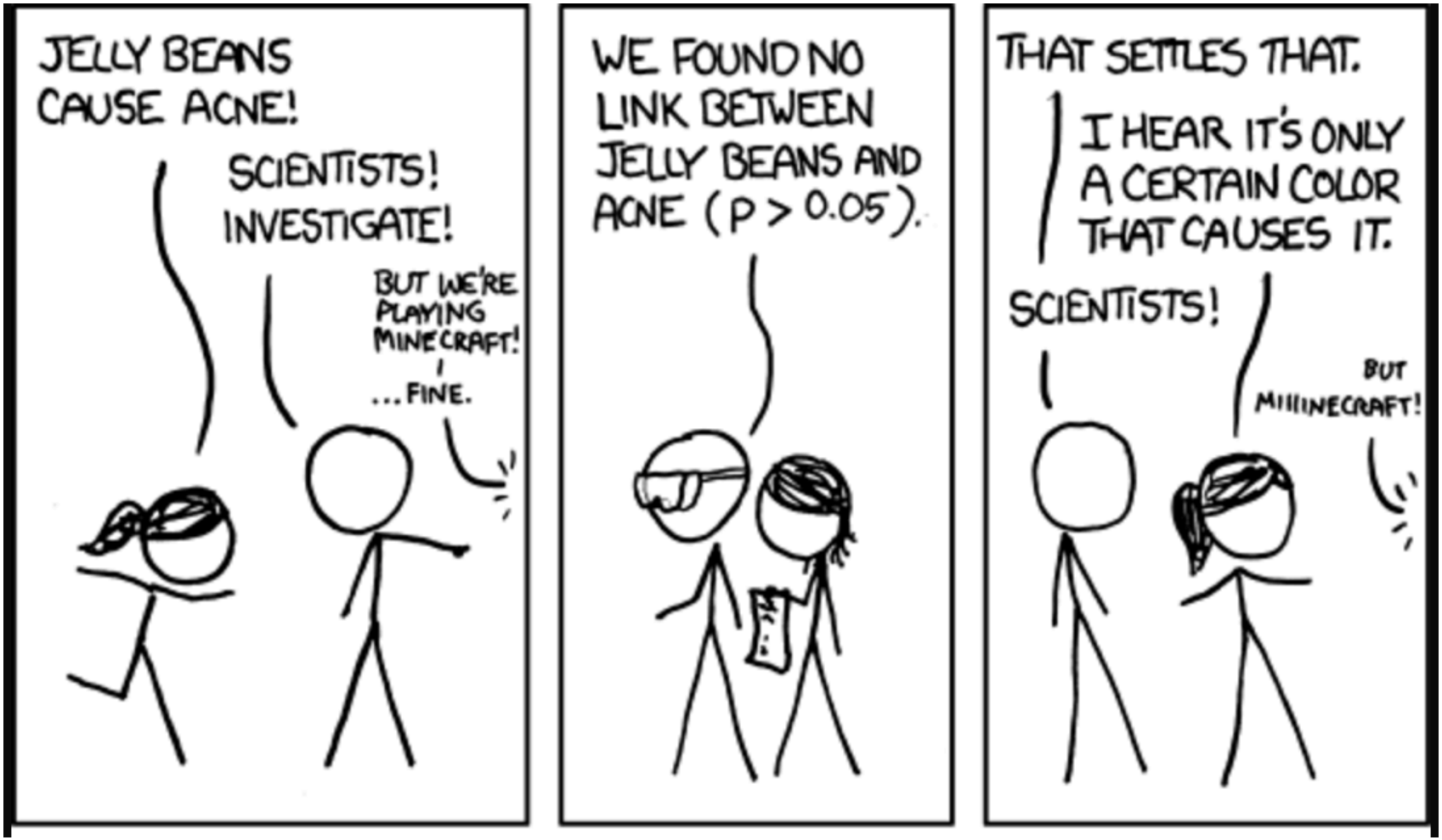
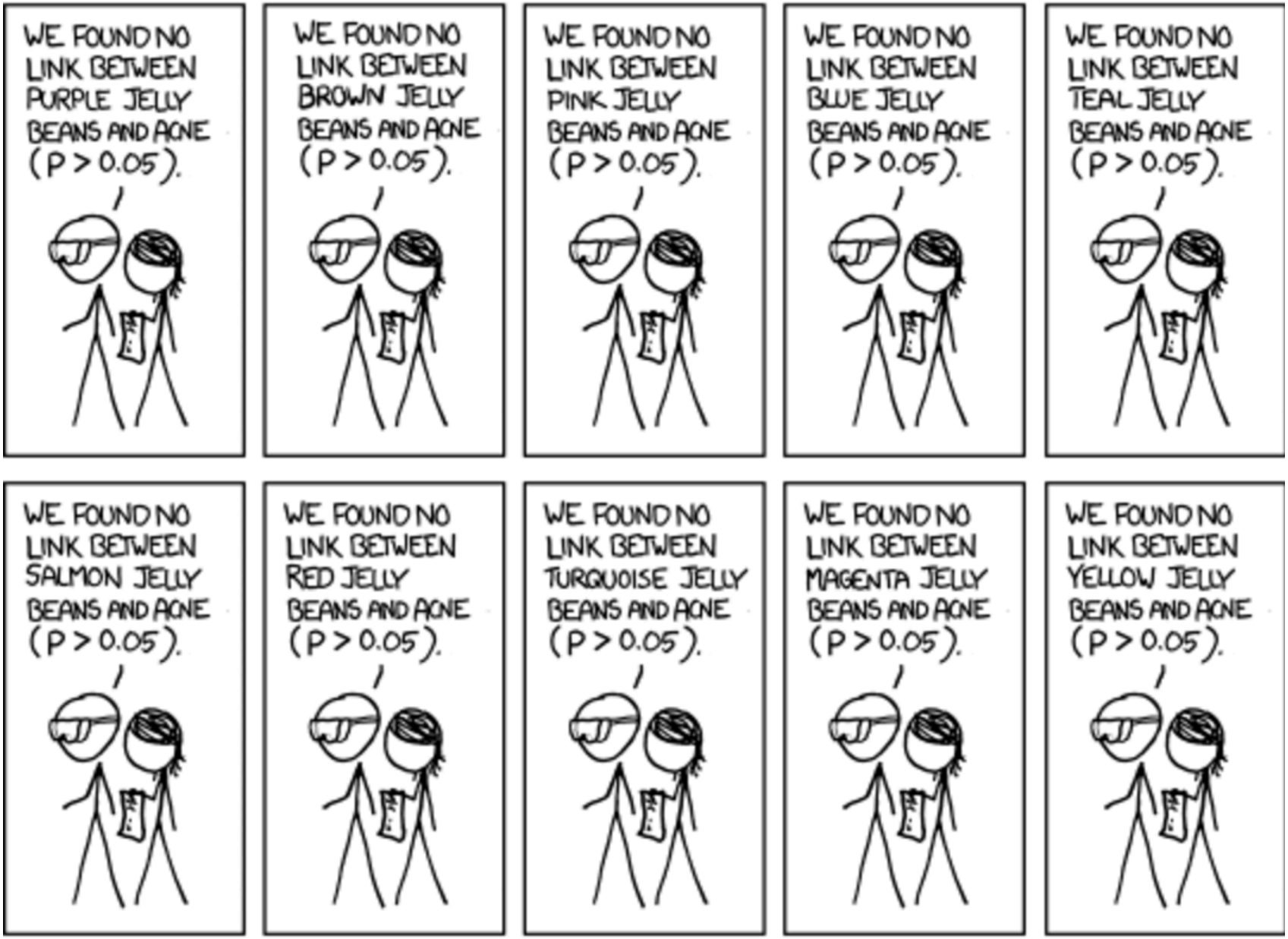
How to improve the credibility of (your) social science

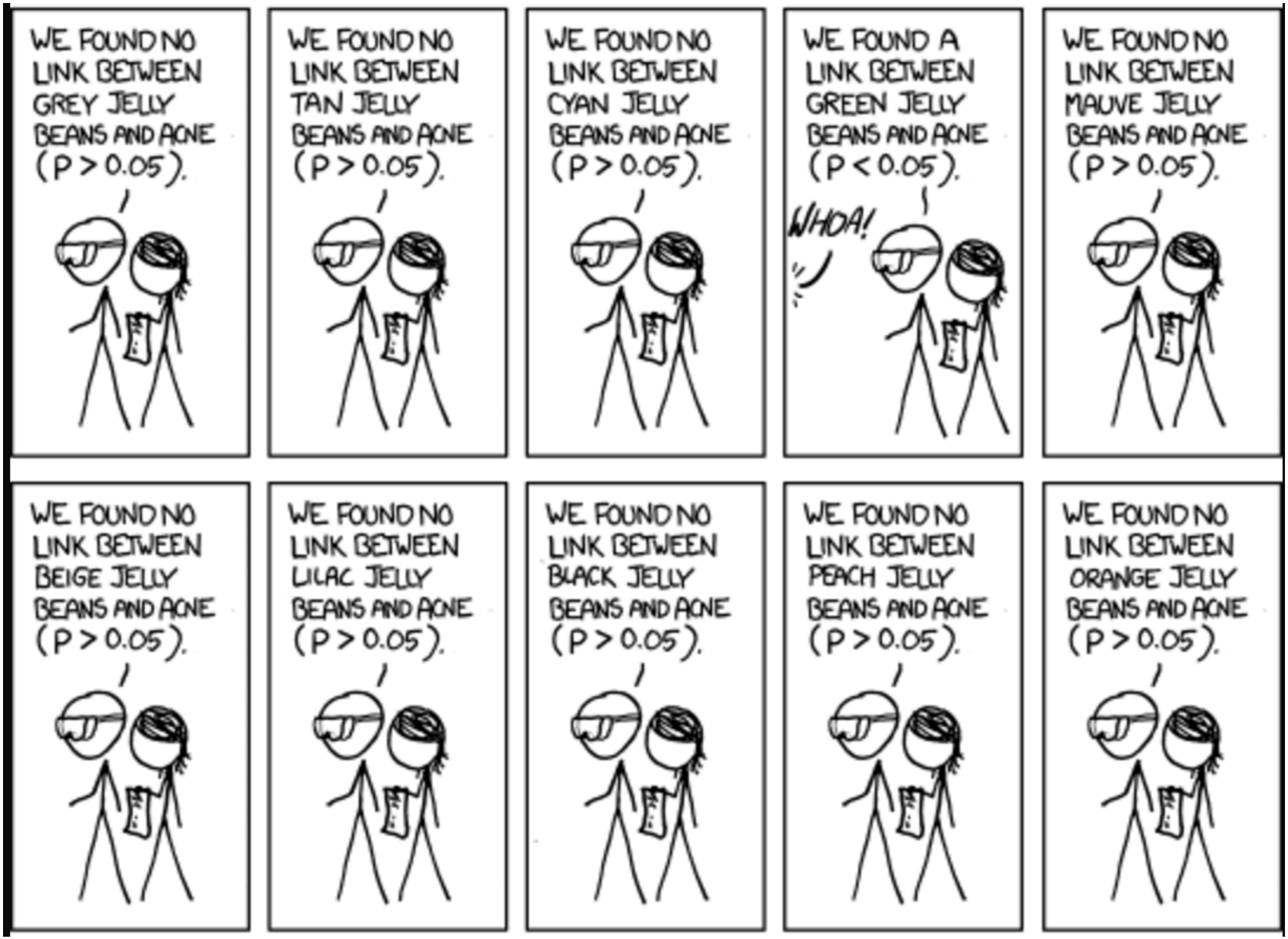
A practical guide for researchers

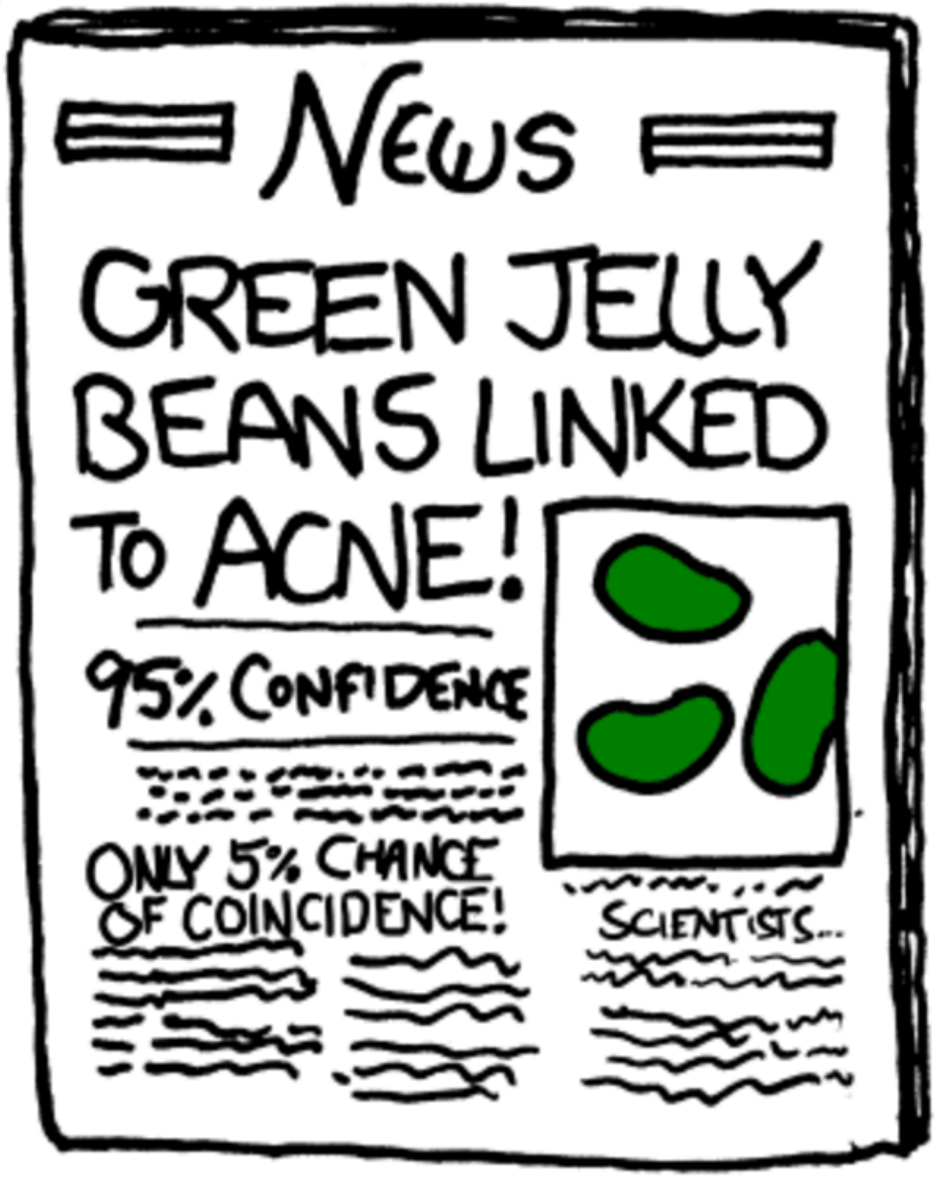
October 2, 2017

The Problem


JELLY BEANS CAUSE ACNE!
SCIENTISTS! INVESTIGATE!
BUT WE'RE PLAYING MINECRAFT!
... FINE.
WE FOUND NO LINK BETWEEN JELLY BEANS AND ACNE ($p > 0.05$).
THAT SETTLES THAT.
I HEAR IT'S ONLY A CERTAIN COLOR THAT CAUSES IT.
SCIENTISTS!
BUT MIIIINECRAFT!


WE FOUND NO LINK BETWEEN PURPLE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BROWN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PINK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLUE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TEAL JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN SALMON JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN RED JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TURQUOISE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN MAGENTA JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN YELLOW JELLY BEANS AND ACNE (P > 0.05).


WE FOUND NO LINK BETWEEN GREY JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN CYAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND A LINK BETWEEN GREEN JELLY BEANS AND ACNE (P < 0.05).
WHOA!
WE FOUND NO LINK BETWEEN MAUVE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BEIGE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN LILAC JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLACK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PEACH JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN ORANGE JELLY BEANS AND ACNE (P > 0.05).


NEWS
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE
ONLY 5% CHANCE OF COINCIDENCE!
SCIENTISTS...

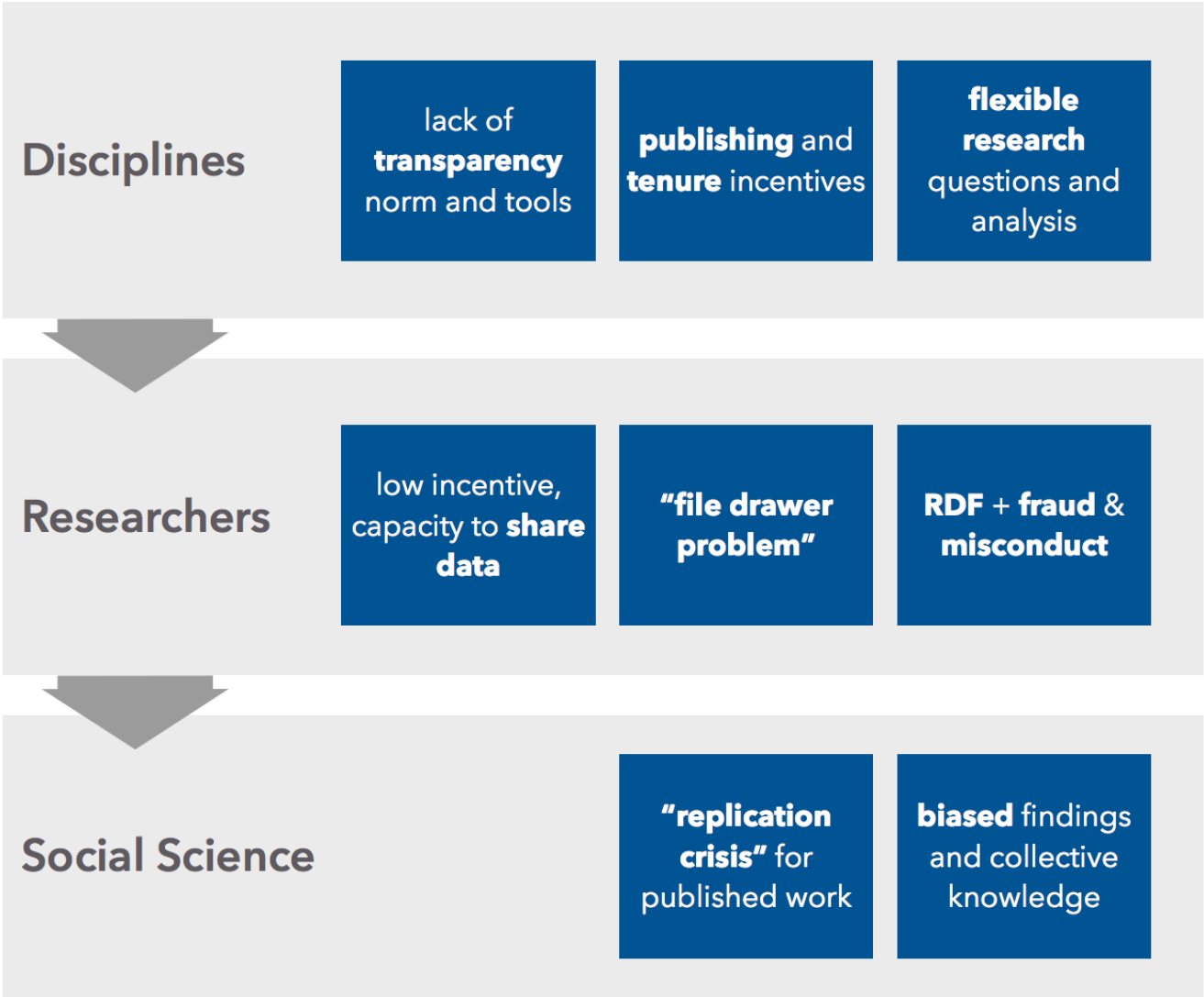
Source: XKCD

We often hear about ...

1. **"Replication crisis"**—studies fail to replicate (psych, econ, polisci, medicine, etc.)
2. **Publication bias**—published studies only represent fraction of results, biased toward significant positive findings
3. **P-hacking/researcher degrees of freedom**—published studies use only a fraction of possible specifications, biased toward significance
4. **Misconduct/fraud**—relatively easy to get away with!

$\rightarrow$ adds up to biased body of knowledge

Why do we have this credibility crisis?



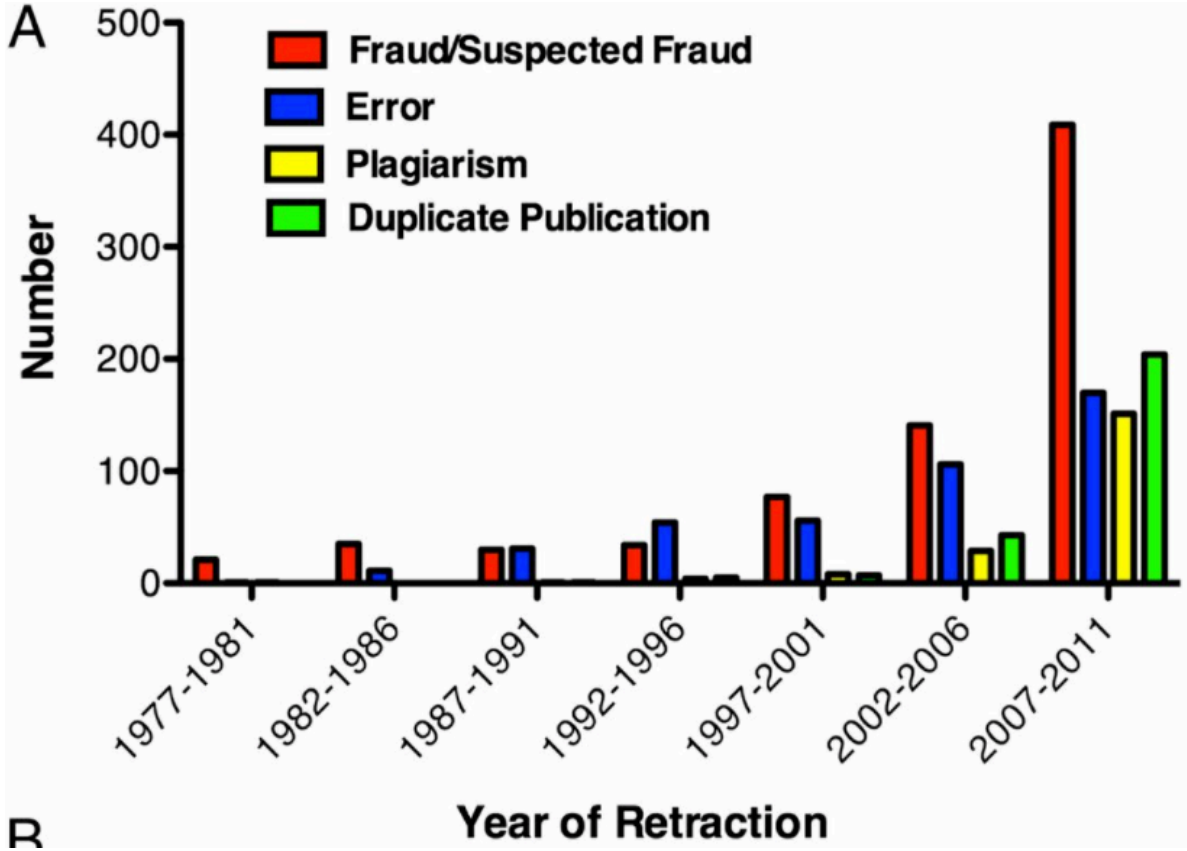
1. “Replication Crisis”

Social, behavioral, and medical studies often don't replicate

- Ideally, replications determine if original results are robust to alternative specifications or samp if they were due to *random chance*.
- In reality, failure to replicate often a result of ...
  - Lack of transparency in sharing data/code
  - Errors in data/code
  - Misconduct or fraud

Fang et al. (2012)

Review of 2,047 retracted biomedical and life-science articles on PubMed:



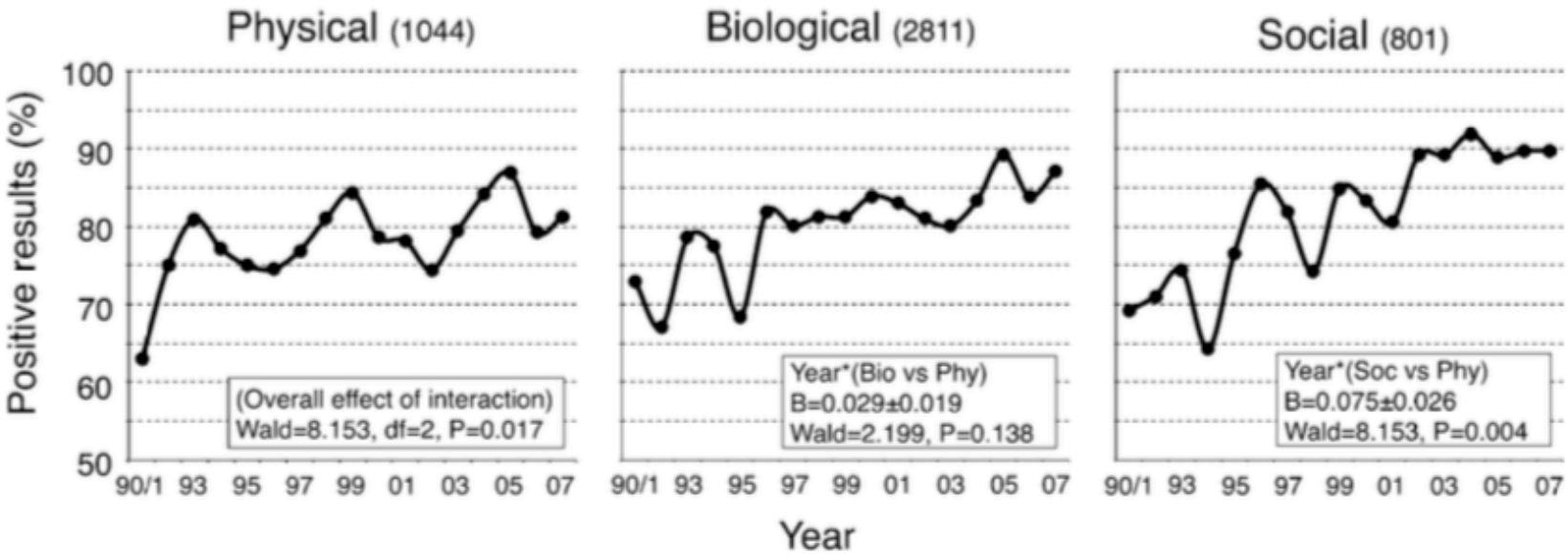
2. Publication Bias

AKA the “file drawer problem”

- **Problem:** Studies more likely to be submitted/published when findings are significant $\rightarrow$ studies with null (or negative) findings are hidden
- **Result:** Bias evidence base—we’re missing full universe of studies and results; what gets published could be due to random chance (e.g., if we expect 5% of results of all studies to be significant)

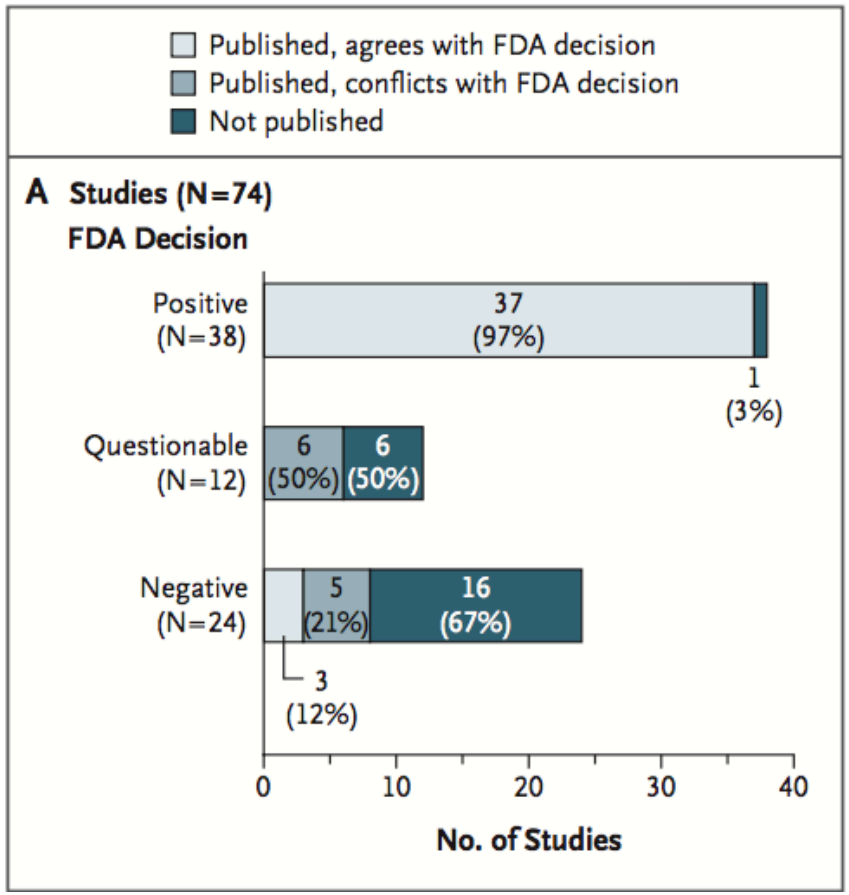
Fanelli (2010 & 2011)

Increase in % of papers with positive results over time, across scientific disciplines:



This has consequences!

E.g., studies that agree with FDA decisions more likely to be published (Turner et al. 2008):



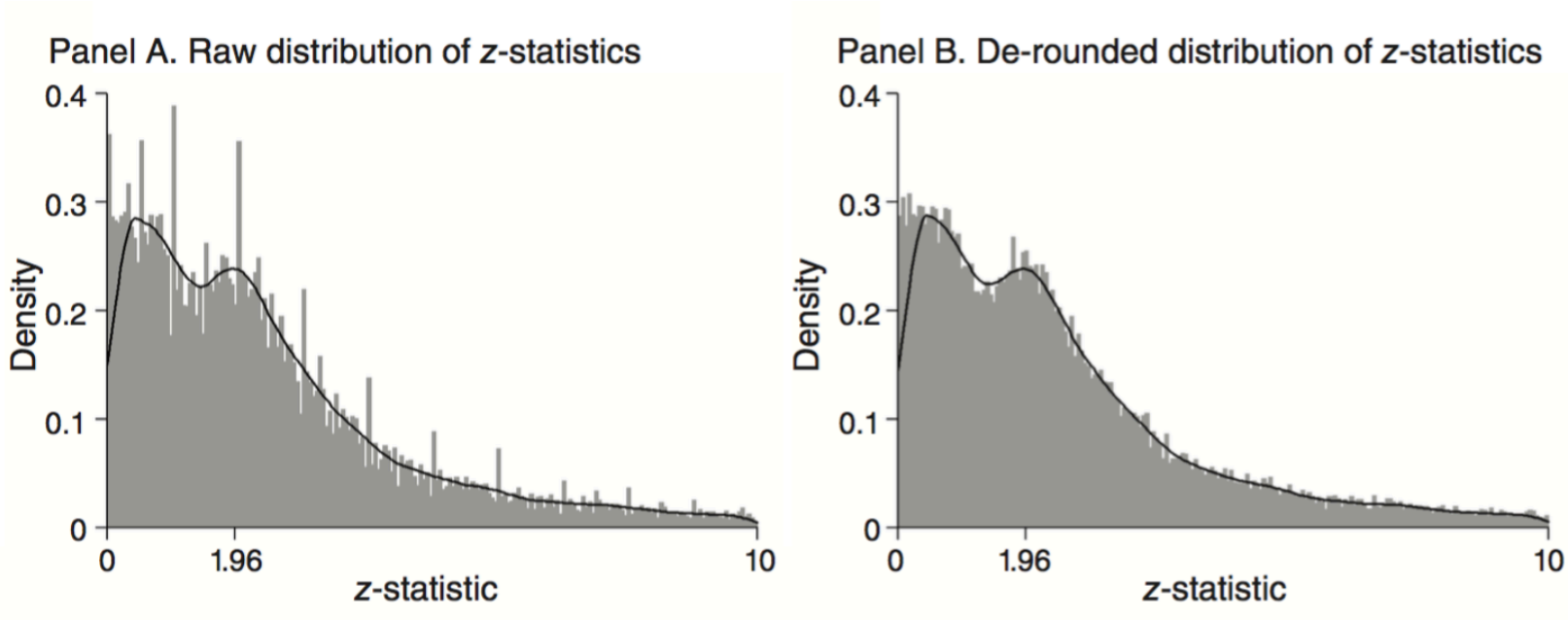
3. P-hacking—AKA fishing, data mining, specification searching, etc.

"Torture the data until it tells you what you want to hear"

- **Opportunity:** Researchers also have many "degrees of freedom" (RDF) in the design and analysis of a study $\rightarrow$ p-hacking (may not always be intentional, see Gelman & Loken 2013)
- **Motive:** Researchers have incentives (from journals, tenure requirements, etc.) to find significance
- **Result:** Biased evidence base (also contributes to replication crisis)

Brodeur et al. (2016)

Evidence of P-Hacking:



4. Misconduct & Fraud

Rare(?) but serious

- **Includes**: Falsifying some or all data and/or results, as well as plagiarism and other forms of misconduct
- **Result:** False or biased evidence base, (also contributes to replication crisis)
- **Note**: Fabrication of data (e.g., LaCour, Fujii, Foster, Staple) less common than other "questionable research practices"

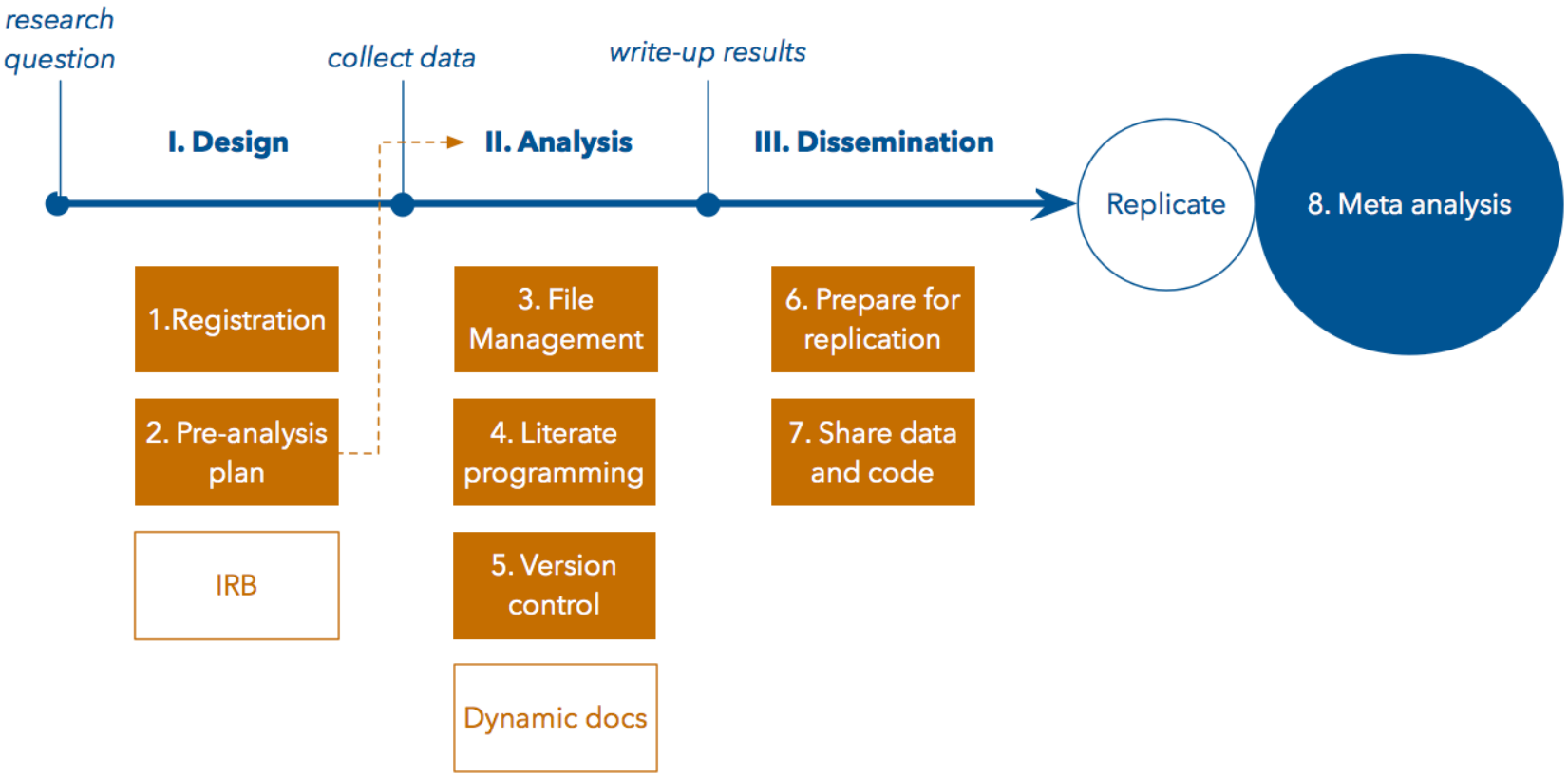
But all hope is not lost ...

Norms are changing

Smart people are working on these issues and developing standards and tools to help throughout the research lifecycle.

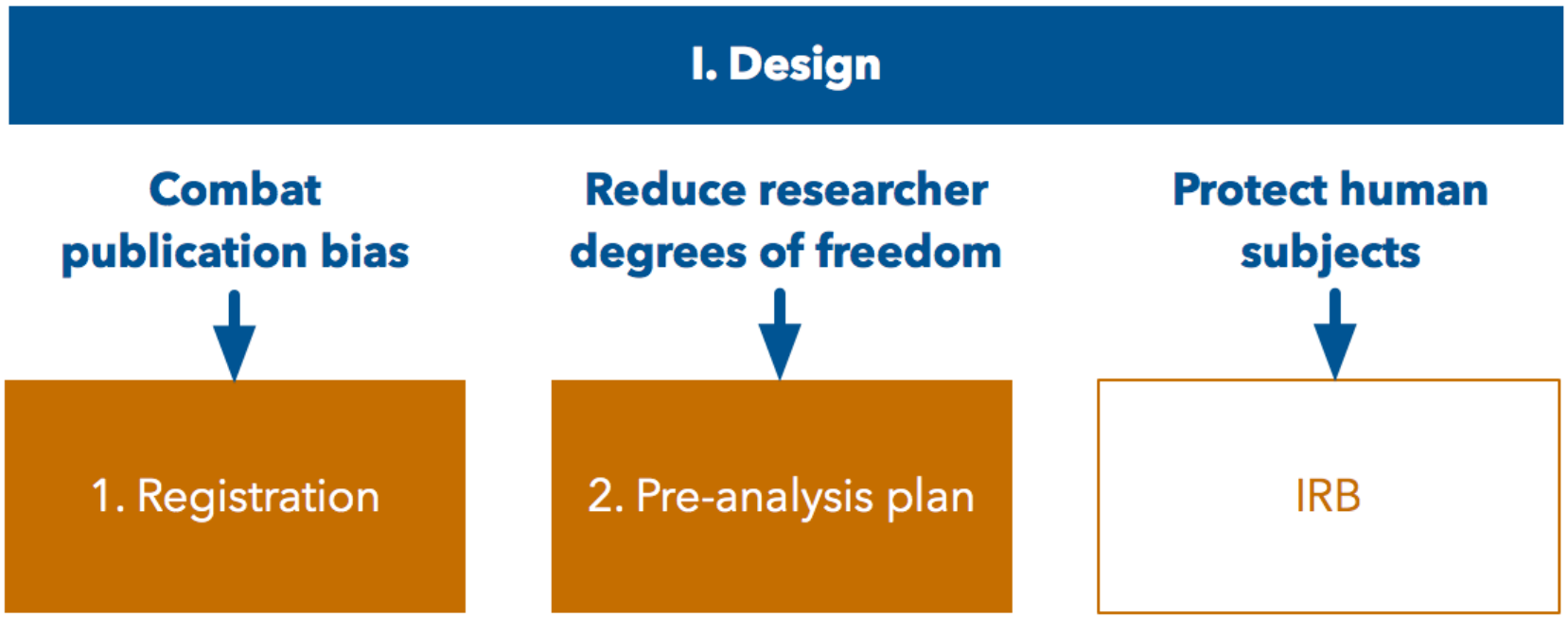
- PDEL, BITSS, OSF, DART, Dataverse, EGAP, etc. etc.

Research lifecycle: Individual-level solutions



Solutions I: Design

Steps


I. Design
Combat publication bias
Reduce researcher degrees of freedom
Protect human subjects
1. Registration
2. Pre-analysis plan
IRB

1. Registration

About Registration

- **What:** Enter your study into the appropriate disciplinary "registry"—basically a requirement for experiments (especially in medicine)

- **Why:** To combat the file-drawer problem, publication bias— also, stake out intellectual claim!

Where to Register

- American Economics Association (AEA): http://socialscienceregistry.org
- Experiments in Governance and Politics (EGAP): http://egap.org/design-registration
- Registry for International Development Impact Evaluations (3ie): http://ridie.3ieimpact.org
- Open Science Framework: http://osf.io—OSF is integrated with other formats, soon with AEA!
- http://aspredicted.org

2. Pre-Analysis Plan

About Pre-Analysis Plans (PAPs)

- **What:** Detailed description of research design and data analysis plans, submitted to a registry BEFORE looking at the data.
- **Why:**
  - Tie your hands for data analysis (address researcher degrees of freedom, etc.)
  - Distinguish between *confirmatory* and *exploratory* analysis
  - Boost credibility of research (get a badge from OSF!)
  - Transparent methods make it easier for others to build on your work

PAP vs. Registration

Registration often—but not always—includes a pre-analysis plan. BUT, purpose is different ...

- **Registration addresses publication bias**—study enters the universe, no matter the outcome
- **PAP addresses p-hacking**—limiting degrees of freedom

No universal standard, can include ...

Background	abstract, motivation, questions
Design	treatment, sampling & randomization, attrition, spillover, survey instruments, power calculations, plan for data collection, processing & management
Analysis	hypotheses (main, auxiliary), outcome measures (primary, secondary), variable operationalization, balance checks, estimation of treatment effects (ATE, ITT, TOT, etc.), HTEs (subgroups, interactions), covariates, standard errors, corrections for multiple hypothesis testing, missing values, outliers
Team	members, affiliations, conflicts of interest
Logistics	fieldwork, timeline, budget

[IRB]

Not covered here, but ...

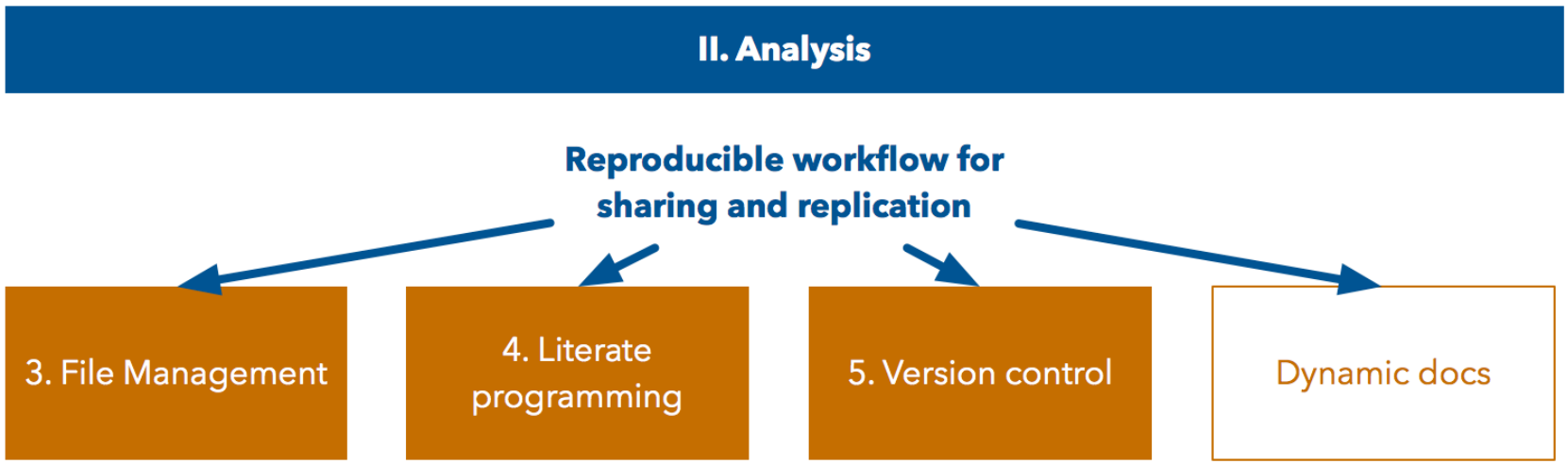
Don't forget IRB requirements to protect human subjects!

Necessary for ethical research, though not sufficient (see http://desposato.org/ethicsfieldexperiments.pdf for more on ethics in experiments).

Solutions II: Analysis

Steps

"**Reproducibility** is just collaboration with people you don't know, including yourself next week" — Philip Stark, UC Berkeley

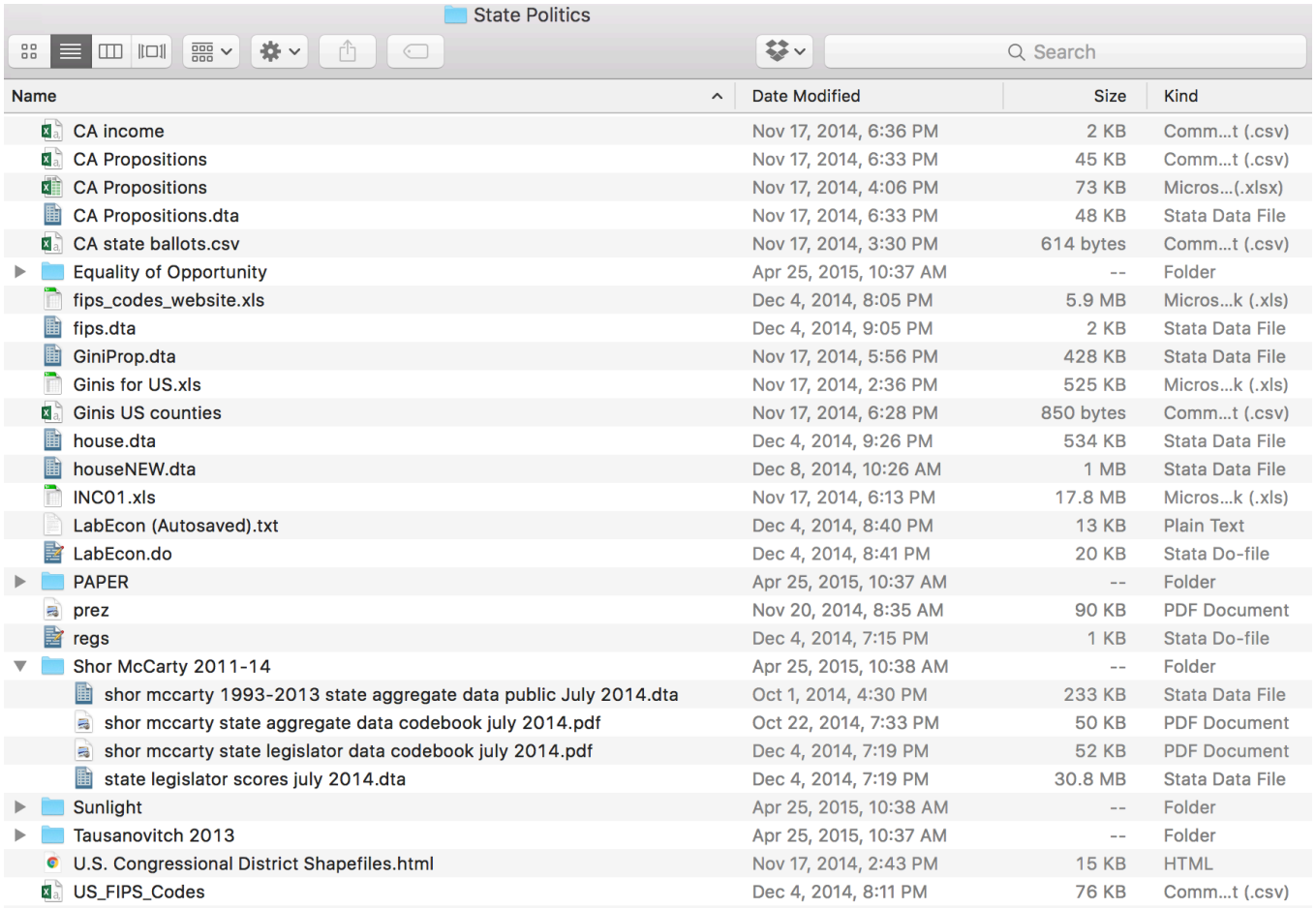


3. File Management

About File Management

- **What:** Organizing and managing files cleanly and intuitively
- **Why:** To preserve original data, streamline workflow, and reduce prep time when sharing files

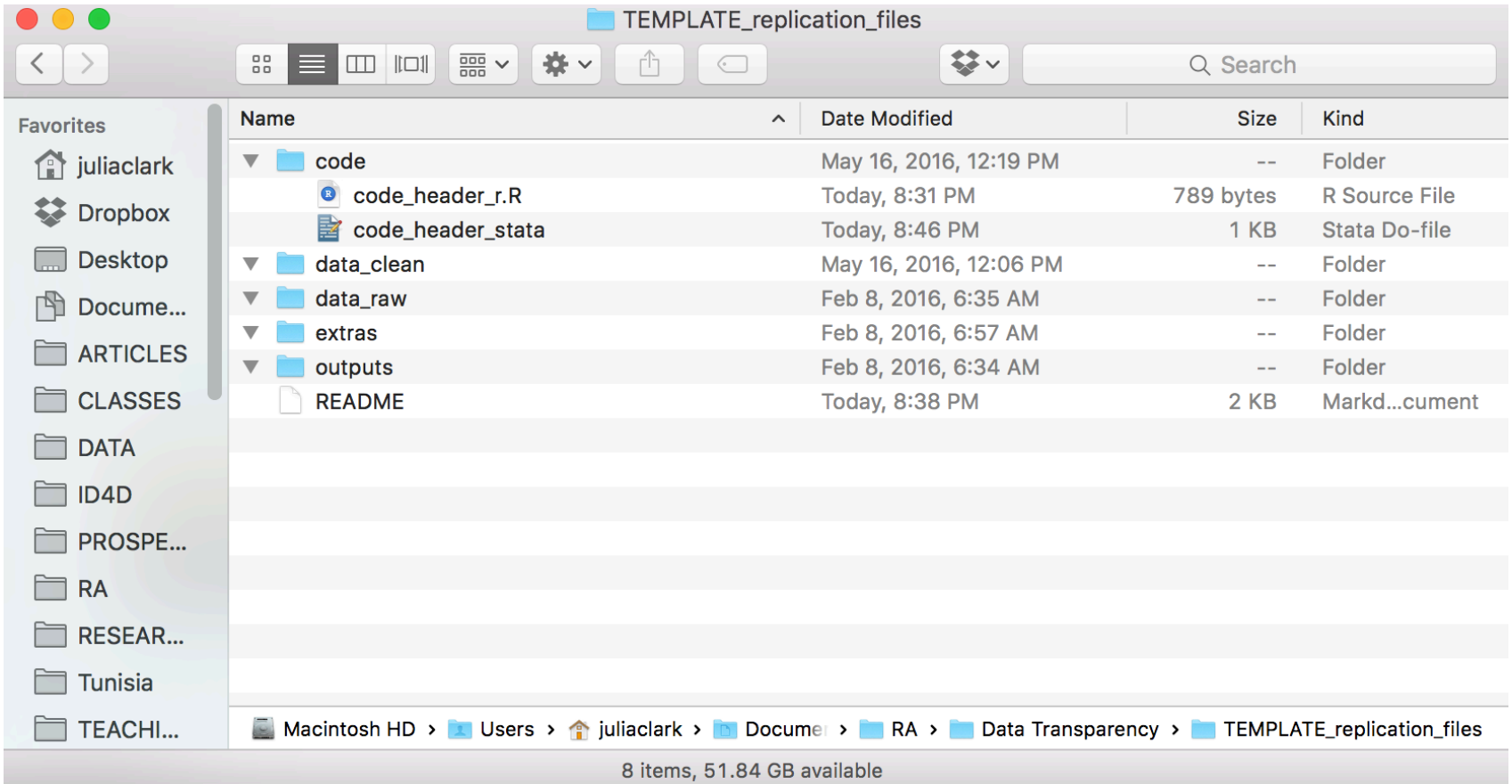
Don’t let your files look like this ...



Name	Date Modified	Size	Kind
CA income	Nov 17, 2014, 6:36 PM	2 KB	Comm…t (.csv)
CA Propositions	Nov 17, 2014, 6:33 PM	45 KB	Comm…t (.csv)
CA Propositions	Nov 17, 2014, 4:06 PM	73 KB	Micros…(.xlsx)
CA Propositions.dta	Nov 17, 2014, 6:33 PM	48 KB	Stata Data File
CA state ballots.csv	Nov 17, 2014, 3:30 PM	614 bytes	Comm…t (.csv)
Equality of Opportunity	Apr 25, 2015, 10:37 AM	--	Folder
fips_codes_website.xls	Dec 4, 2014, 8:05 PM	5.9 MB	Micros…k (.xls)
fips.dta	Dec 4, 2014, 9:05 PM	2 KB	Stata Data File
GiniProp.dta	Nov 17, 2014, 5:56 PM	428 KB	Stata Data File
Ginis for US.xls	Nov 17, 2014, 2:36 PM	525 KB	Micros…k (.xls)
Ginis US counties	Nov 17, 2014, 6:28 PM	850 bytes	Comm…t (.csv)
house.dta	Dec 4, 2014, 9:26 PM	534 KB	Stata Data File
houseNEW.dta	Dec 8, 2014, 10:26 AM	1 MB	Stata Data File
INC01.xls	Nov 17, 2014, 6:13 PM	17.8 MB	Micros…k (.xls)
LabEcon (Autosaved).txt	Dec 4, 2014, 8:40 PM	13 KB	Plain Text
LabEcon.do	Dec 4, 2014, 8:41 PM	20 KB	Stata Do-file
PAPER	Apr 25, 2015, 10:37 AM	--	Folder
prez	Nov 20, 2014, 8:35 AM	90 KB	PDF Document
regs	Dec 4, 2014, 7:15 PM	1 KB	Stata Do-file
Shor McCarty 2011-14	Apr 25, 2015, 10:38 AM	--	Folder
shor mccarty 1993-2013 state aggregate data public July 2014.dta	Oct 1, 2014, 4:30 PM	233 KB	Stata Data File
shor mccarty state aggregate data codebook july 2014.pdf	Oct 22, 2014, 7:33 PM	50 KB	PDF Document
shor mccarty state legislator data codebook july 2014.pdf	Dec 4, 2014, 7:19 PM	52 KB	PDF Document
state legislator scores july 2014.dta	Dec 4, 2014, 7:19 PM	30.8 MB	Stata Data File
Sunlight	Apr 25, 2015, 10:38 AM	--	Folder
Tausanovitch 2013	Apr 25, 2015, 10:37 AM	--	Folder
U.S. Congressional District Shapefiles.html	Nov 17, 2014, 2:43 PM	15 KB	HTML
US_FIPS_Codes	Dec 4, 2014, 8:11 PM	76 KB	Comm…t (.csv)

Instead, use PDEL template (or similar)

Download at https://github.com/PolicyDesignEvaluationLab/Transparency-Initiative



4. Literate Programming

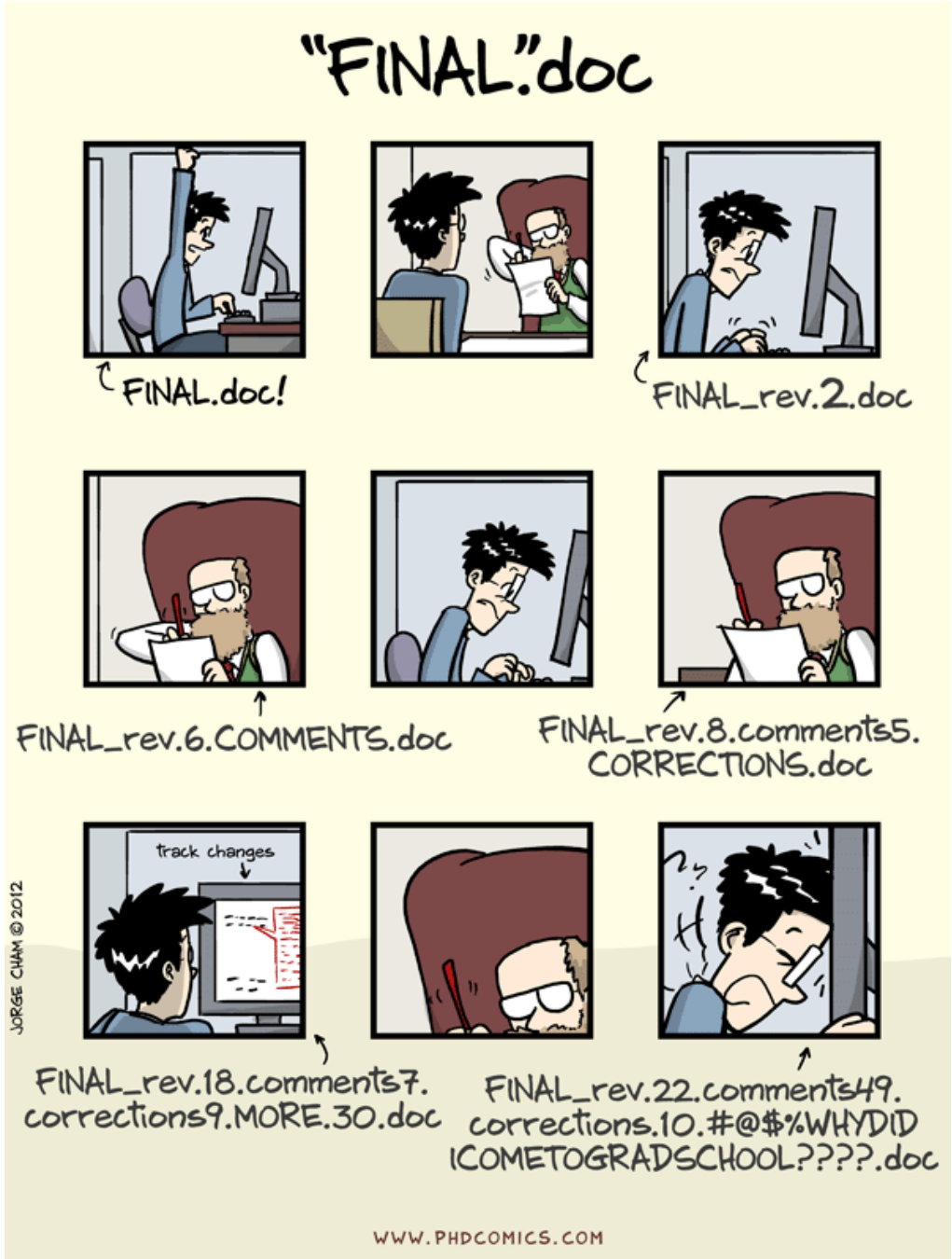
About Literate Programming

- **What:** Writing code that it's legible to ***humans***
- **Why:** So you and others can better replicate your work (and to help you avoid mistakes!)

(The Most) Basic Principles

- Structure and name files intuitively
- Make the contents of files easy to navigate
- Streamline code to avoid repetition

5. Version Control


"FINAL".doc
FINAL.doc!
FINAL_rev.2.doc
FINAL_rev.6.COMMENTS.doc
FINAL_rev.8.comments5.
CORRECTIONS.doc
track changes
FINAL_rev.18.comments7.
corrections9.MORE.30.doc
FINAL_rev.22.comments49.
corrections.10.#@$%WHYDID
ICOMETOGRADSCHOOL????.doc
JORGE CHAM © 2012
WWW.PHDCOMICS.COM

About Version Control

- **What:** A system for managing iterative versions of files (code, data, manuscripts) over time and across collaborators
- **Why:** Keep original files, protect work, collaborate efficiently, streamline workflow, etc., etc.

Principles of Version Control

- Vault original, raw data files—do not save over!
- Changes to files should be documented and reversible
- Keep “master” versions of files in working order; create copies before experimenting
- Reconcile independent changes by different users

Manual Solutions (not ideal, but better than nothing)

- Create dated versions of files (save-as) for each substantive change
- With each modification, re-run ALL code to make sure nothing is broken—helps if you have a master file to run all scripts!
- Check-in with coauthors to ensure multiple people aren't working on the same files at the same time
- Keep a simple log to remind yourself of the location/content of major changes

Or let version control software do this for you!

GitHub

Common problems that GitHub helps solve

- **Tracking changes in code/text files**—who, what, where, when, preserved forever
- **Selectively reverting changes**—better than `ctrl + Z`
- **Experimenting**—easier than "my_code_v2_new.R"
- **Collaborating**—sharing/vetting/reconciling changes

[Dynamic Docs]

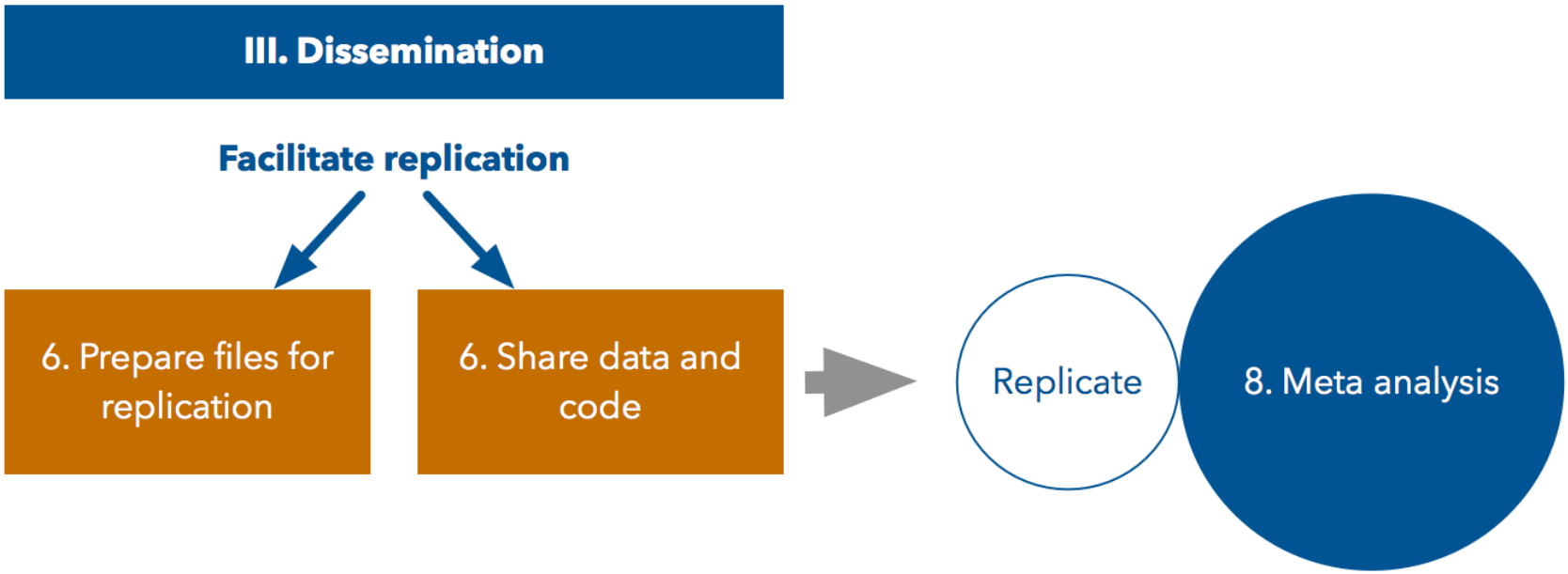
Not covered here, but ...

You can take reproducible research a step further by integrating code *into your manuscript*.

- RMarkdown
- Stata Markdoc or Stata texdoc

Solutions III: Dissemination

Steps


III. Dissemination
Facilitate replication
6. Prepare files for replication
6. Share data and code
Replicate
8. Meta analysis

6. Prepare for Replication

Why do we care if our code is reproducible?

- **Unselfish reasons**—part of the scientific process and a public good
- **Selfish reasons**—make code more usable for yourself, catch potentially embarrassing errors before they become public, boost your transparency credibility

Replication files should ...

- Be complete but parsimonious
- Run and reproduce results with one click
- Be readable and interpretable by humans
- Protect personal information

Caveat: There is no single, perfect way to organize or prepare files for replication. Do what works for you (as long as it meets the above criteria)!

5 Steps for Prepping Files

1. Set-up
2. Initial replication
3. De-identify
4. Edit
5. Final replication

1. Set Up

Create a ***new***, clearly organized folder structure for replication that you add to selectively.

- Purpose:
  - Ensure files are complete/parsimonious, legible
  - Protect original files

Create

1. A new, empty replication folder *within* your project directory (e.g., "`replication_files/`")
2. Subfolders: *Same as File Management tips!*
   - `code/` — scripts
   - `data_clean/` — manipulated data
   - `data_raw/` — original data
   - `output/` — generated tables, graphs, etc.
   - `extra/` — misc. extras (e.g., code book)
3. A "README.txt" file to document contents, sources, software/system versions, other info necessary for replication/comprehension.

2. Initial Replication

Copy (don't move!) data and code files into the replication folder and try to replicate your results.

Purpose:

- Make sure your code actually runs and reproduces before you tinker with structure and formatting
- Build up your replication folder with complete and parsimonious data/code files

3. De-Identifying Individual-Level Data

If you haven't already, make sure replication files ***do not contain*** data that could be used to identify individuals.

Purpose:

- Protect individuals' identity and private information—ethical issue for researchers, potential safety issue for participants
- Comply with legal, research board or funder requirements (e.g., HIPAA and IRB in the US)

What does “de-identifying” mean?

Two types of identifiers:

1. Direct: Variables explicitly linked to subjects—***e.g., name, email, address, ID number, phone number, etc.***

2. Indirect: Variables that, in combination, could be used to identify individuals—***e.g., gender, dates (birth, program admission, etc.), geographic location (village, GPS), unusual occupations or education, etc.***

See this useful infographic.

Example of Indirect Identifiers

- You survey teachers and collect information on *gender*, *grade-level taught*, and *age*.
- If there is only one *female*, *third-grade* teacher *aged 40-49* at a particular school, she is not anonymous in your data

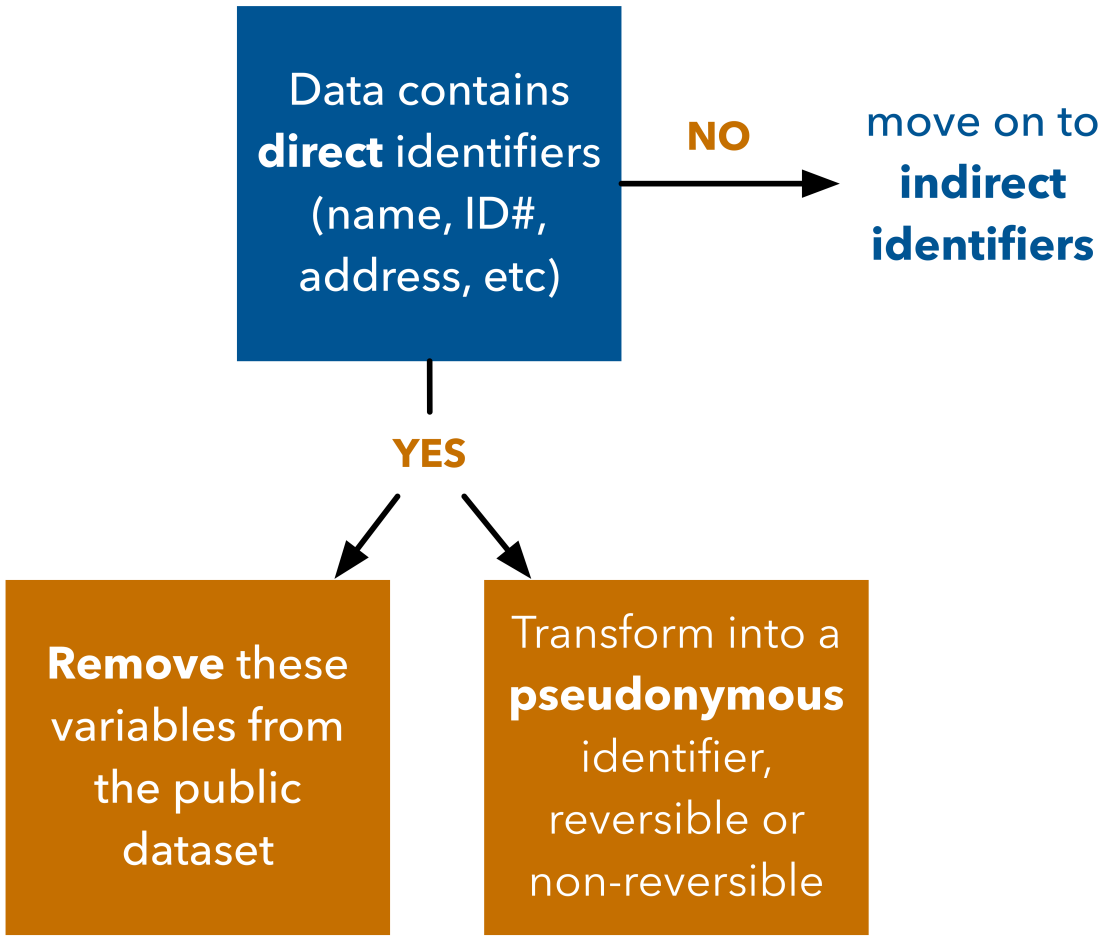
Dealing with Direct Identifiers

In general, direct identifiers—e.g., name, address, mobile number, ID number—should ***never*** be made public.

Options:

- Remove variables from shared dataset
- Pseudonymize data in order to be able to link datasets: replace identifiers with “pseudonyms” that may be reversible or non-reversible, e.g., give people random names or ID numbers

Solutions for Direct Identifiers

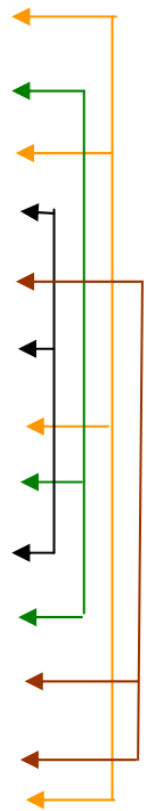


What is sufficient de-identification for indirect identifiers?

1. **Determine Risk:** Pr(being identified) $\times$ sensitivity of data
2. **Set "k-anonymous" level:** each record cannot be distinguished from at least $k-1$ other individuals who also appear in the data set
3. **Select appropriate method(s) of de-identification:** aggregating data, removing certain variables or observations, reducing information/detail, adding random noise or values

Example of K-anon where k=3

Pseudo ID	Age	Gender	ICD-10 Code
Patient 1	0 to 10 yrs	M	F106
Patient 2	20 to 35 yrs	F	F106
Patient 3	0 to 10 yrs	M	F106
Patient 4	51 to 65 yrs	F	F106
Patient 5	20 to 35 yrs	M	F106
Patient 6	51 to 65 yrs	F	F106
Patient 7	0 to 10 yrs	M	F106
Patient 8	20 to 35 yrs	F	F106
Patient 9	51 to 65 yrs	F	F106
Patient 10	20 to 35 yrs	F	F106
Patient 11	20 to 35 yrs	M	F106
Patient 12	20 to 35 yrs	M	F106
Patient 13	0 to 10 yrs	M	F106



4. Edit and Organize Files for Clarity

Next step is to clean and annotate data, code, and other files to improve usability.

Purpose:

- Ensure files are legible in terms of structure and content

Basic steps

- Structure and name files*
- Streamline and annotate code*
- Document file and folder contents

*Already done if you follow the literate programming tips in Phase II!

Document File and Folder Content

- Update the README file to describe contents of replication folders
- If necessary, include codebook in "`extra/`" folder
- Document packages & software versions used
  - R: `sessionInfo()`
  - Stata: `version`

5. Final Replication

- Shutdown or clear your Stata/R/etc. memory
- Rerun the entire process—merging, cleaning and analysis—to make sure your edits didn't break anything
- Testing on a friend (or RA's) computer can also be a final check
- Once discrepancies are addressed, the files are ready for sharing!

7. Share Data and Code

About Sharing Data and Code

- **What:** add replication files to an online repository
- **Why:** lasts longer than personal website, more searchable, future proof
- **Concerns:**
  - Can usually be embargoed, or provide only what is necessary for replication (e.g., unused survey Qs)
  - Biggest risk isn't having your data/ideas stolen, it's having your research ignored! (King 1995)
  - Difficult if proprietary

Where to Share

Depends on discipline: find appropriate registry at http://www.re3data.org/, or check out ...

- **Harvard's Dataverse**
- Open Science Framework
- OpenICPSR
- figshare
- Data Dryad
- University library (e.g., http://library.ucsd.edu/dc/rdcp/collections)

8. Meta-Analysis

About Meta-Analysis

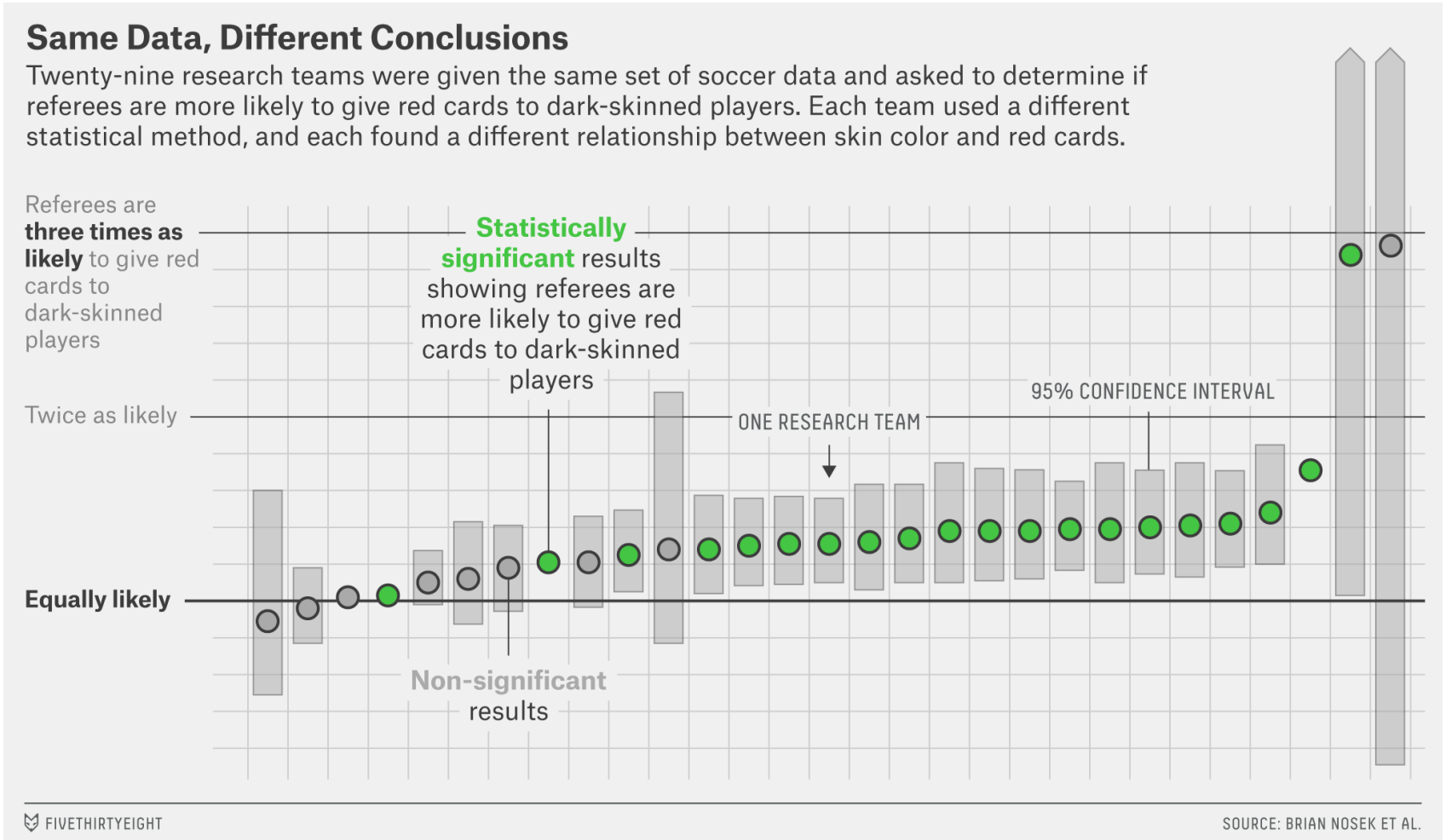
- **What:** Statistical analysis of a group of studies to derive a pooled estimate of the effect of a treatment; may be part of a "systematic review"
- **Why:** Because any estimate in an individual study may be biased or contain random error (note: assumes NO publication bias!)

One Study = One Data Point

That experiment you just ran with 3,685 participants? It's one data point among many other potential studies.

- What if the results are due to random chance?
- What if there was bias in your sample?
- What if someone else had analyzed your data?

Even with the same data, results may vary ...



Source: Graph = fivethirtyeight.com, see https://osf.io/j5v8f/ for study materials

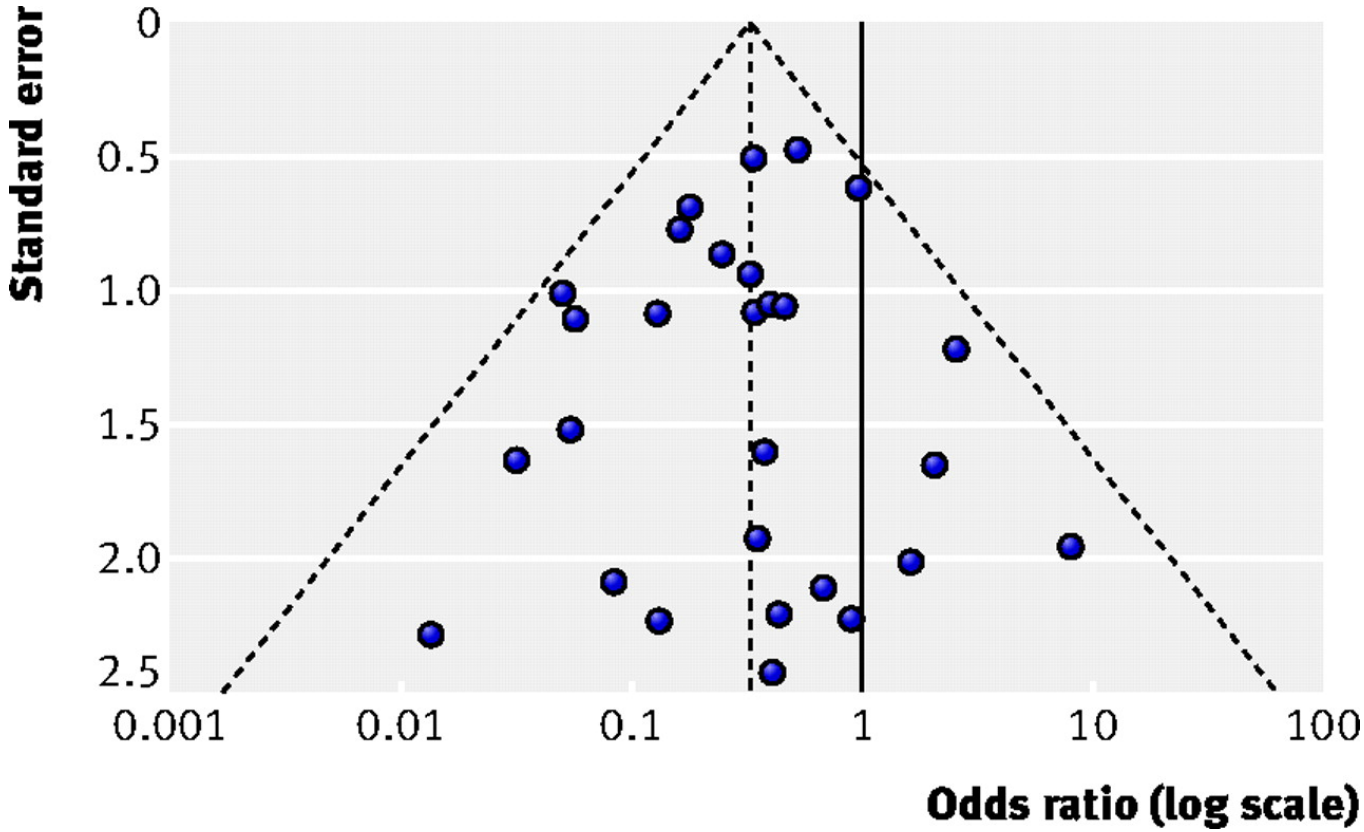
Basic Steps

Using a PAP or “protocol” ...

1. Determine which studies to include
2. Determine which outcomes to measure (e.g., discrete, continuous)
3. Select model for “meta-regression” (e.g., RE, FE, etc.)

Funnel Plots

Scatter plot of study effect sizes vs. precision (e.g., SE of treatment effect)



Source: BMJ 2011

Extra

Solutions at the Institutional/Discipline Level

- **Design-based publication:** AKA "registered reports," moves peer review before data analysis (example)
- **Incentives for transparency, replication, meta-analysis:** See BITSS prizes and awards, OSF pre-registration challenge, etc.
- **Change norms:** e.g., journal/disciplinary standards for data sharing
- **Training:** Like this! More at BITSS, Center for Open Science, etc.
- **Tenure:** "Adherence to the replication standard should be part of [tenure] judgment" (King 1995)

Selected Reading

- Transparency: BITSS Best Practices Manual
- Replication: Dewald et al. (1986), King (1995), Fang et al. (2012), FiveThirtyEight (2015), Clemens (2015)
- Publication bias: Turner et al. (2008), Gerber & Malhotra (2008) Fanelli (2010), Fanelli (2011), Franco et al. (2014)
- P-hacking, fishing, researcher degrees of freedom, fraud: Simons, Nelson, Simonsohn (2011), Gelmen & Loken (2013), Brodeur et al. (2016), John et al. (2012)
- PAPs: Olken 2013, Coffman & Niederle (2015), Neumark 2001
- De-identifying data: Tools for De-Identification, El Emam (2010)
- Literate programming: Long (2008), Gandrud (2013), Gentzkow & Shapiro (2014)
- Meta-analysis: Card & Krueger (1995), Stanlet & Doucouliagos (2012), BMJ (2011)

Thank you!

About this Presentation

This presentation was developed by Julia Clark, Scott Desposato, and Craig McIntosh of UCSD's Policy Design and Evaluation Lab (PDEL) as part of an effort to integrate good research transparency practices into methods training at UCSD.

Funding for this project was generously provided by the Berkeley Initiative for Transparency in the Social Sciences (BITSS) through a Catalyst grant.

This presentation and associated materials are available online at GitHub and are licensed under CC BY-NC 4.0 

. You are free to share and adapt them for any non-commercial purpose with proper attribution. Please cite as "Clark, J., Desposato, S., and McIntosh, C. 2017. 'How to improve the credibility of (your) social science: A practical guide for researchers'. Policy Design and Evaluation Lab (PDEL). University of California, San Diego."