Bringing the Scientific Process Into the Social Sciences

A practical guide from start to finish

Craig McIntosh & Julia Clark

Designing Field Experiments

30 May 2017

Motivation

"Academia is a social enterprise that is usually most successful when individual researchers compete and collaborate in contributing toward common goals.

In contrast, when we work in isolation on unrelated problems, ignoring work that has come before, we lose the benefits of evaluating each other's work, analyzing the same problem from different perspectives, improving measurement techniques and methods, and, most important, building on existing work rather than repeatedly reinventing the wheel."

– Gary King 1995, p. 445

(Social) Science!

In order to accumulate knowledge, we need to be able to replicate and extend the work that came before. This means we need to know:

- The universe of previous studies
- Their methods and specifications
- Their results (from ALL studies, not just the ones that got published!)

$\rightarrow$ Requires research transparency

Without transparency, we have issues ...

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

$\rightarrow$ fragmented and biased body of knowledge

“Replication Crisis”

- **Problem:** In the past decade, many studies have failed to replicate across the social, behavioral, and medical sciences
- Ideally, replications help determine if original results are robust to alternative specifications or 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

Dewald et al. (1986)

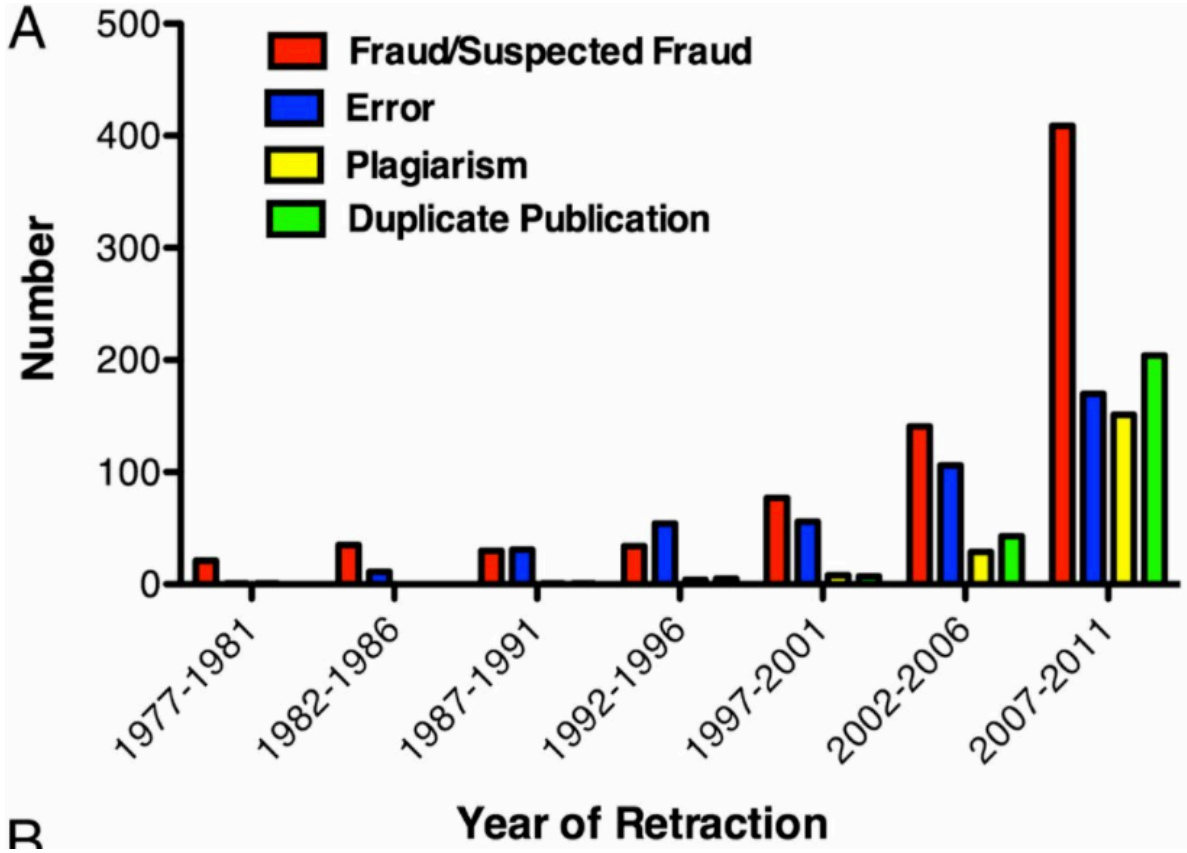
Attempted to replicate 62 papers submitted to ***Journal of Money, Credit and Banking***, got data/code for only 22:

TABLE 2—PROBLEMS IN SUBMITTED DATA SETS

	Published before Data Requested	Accepted before Data Requested	Under Review when Data Requested
No Problems	1	3	4
Problems Identified:			
Incomplete Submission	6	3	5
Sources Cited Incorrectly	0	4	4
Sources Cited Imprecisely	11	7	10
Data Transformations Described Incompletely	3	4	1
Data Element Not Clearly Defined	2	3	2
Other	0	3	1
Problems	22	24	23
Data Sets Examined	19	14	21

Fang et al. (2012

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



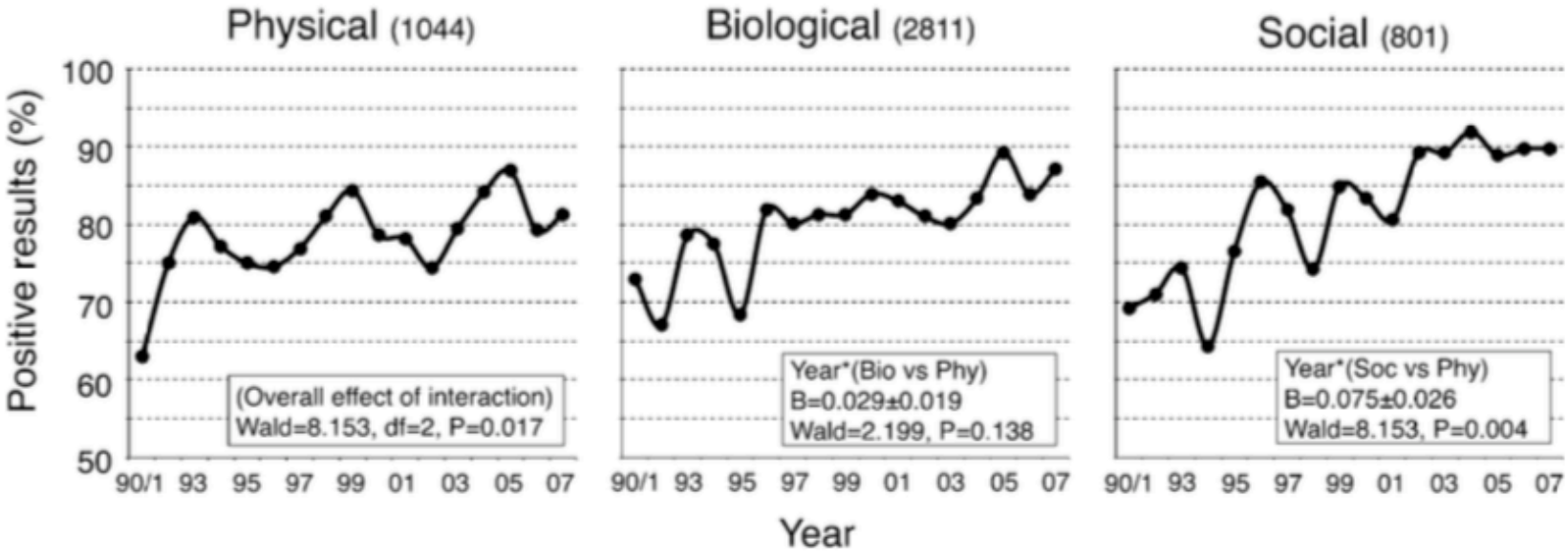
Publication Bias

- **"The File Drawer Problem":** Studies more likely to be published when findings are significant $\rightarrow$ studies with null (or negative) findings are hidden

- **Result:** Missing full universe of studies and results; what we believe to be a solid body of evidence could be due to random chance (e.g., if we expect 5% of results of all studies to be significant)

Publication Bias

Increase in % of papers with positive results over time, across scientific disciplines (Fanelli 2010, 2011):



Bias in the Social Sciences

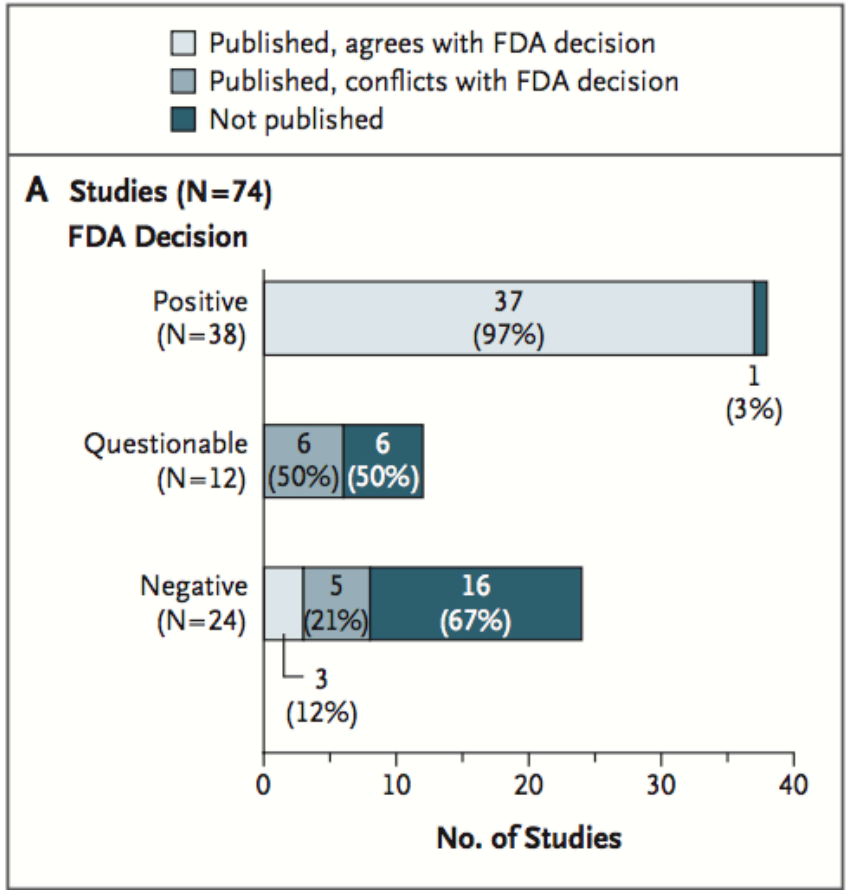
Strong results 60pp more likely to be written up than null results, 40pp more likely to be published (Franco, Malhotra, Simonovits 2014):

Table 3. Cross-tabulation between statistical results of TESS studies and their publication status (column percentages reported). Pearson χ^2 test of independence: χ^2 (6) = 80.3, $P < 0.001$.

	Null (%)	Mixed (%)	Strong (%)
Not written	64.6	12.2	4.4
Written but not published	14.6	39.0	34.1
Published (non–top-tier)	10.4	37.8	38.4
Published (top-tier)	10.4	11.0	23.1
Total	100.0	100.0	100.0

This has consequences!

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

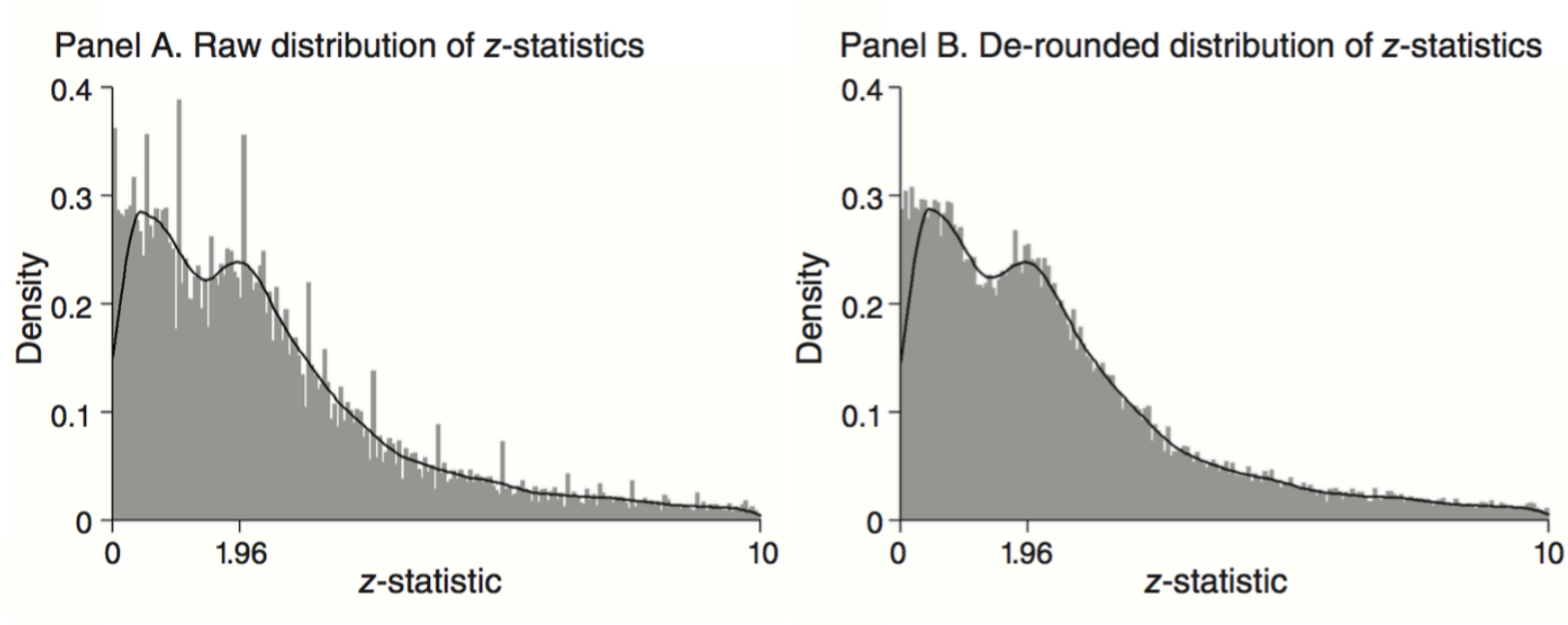


About P-Hacking (AKA fishing, data mining, specification searching, etc.)

- **Motive:** Researchers have incentives (from journals, tenure requirements, etc.) to find significance
- **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)
- **Result:** Biased evidence base (also contributes to replication crisis)

Evidence of P-Hacking

Brodeur et al. (2016):



Researcher Degrees of Freedom

Wicherts et al. 2016 identify 34 key RDFs (see article for full list):

Table 1
Checklist for different types of degrees of freedom in the planning, executing, analyzing, and reporting of psychological studies

Code	Related	Type of Degrees of Freedom
		Hypothesizing
T1	R6	Conducting explorative research without any hypothesis
T2		Studying a vague hypothesis that fails to specify the direction of the effect
		Design
D1	A8	Creating multiple manipulated independent variables and conditions
D2	A10	Measuring additional variables that can later be selected as covariates, independent variables, mediators, or moderators
D3	A5	Measuring the same dependent variable in several alternative ways
D4	A7	Measuring additional constructs that could potentially act as primary outcomes
D5	A12	Measuring additional variables that enable later exclusion of participants from the analyses (e.g., awareness or manipulation checks)
D6		Failing to conduct a well-founded power analysis
D7	C4	Failing to specify the sampling plan and allowing for running (multiple) small studies

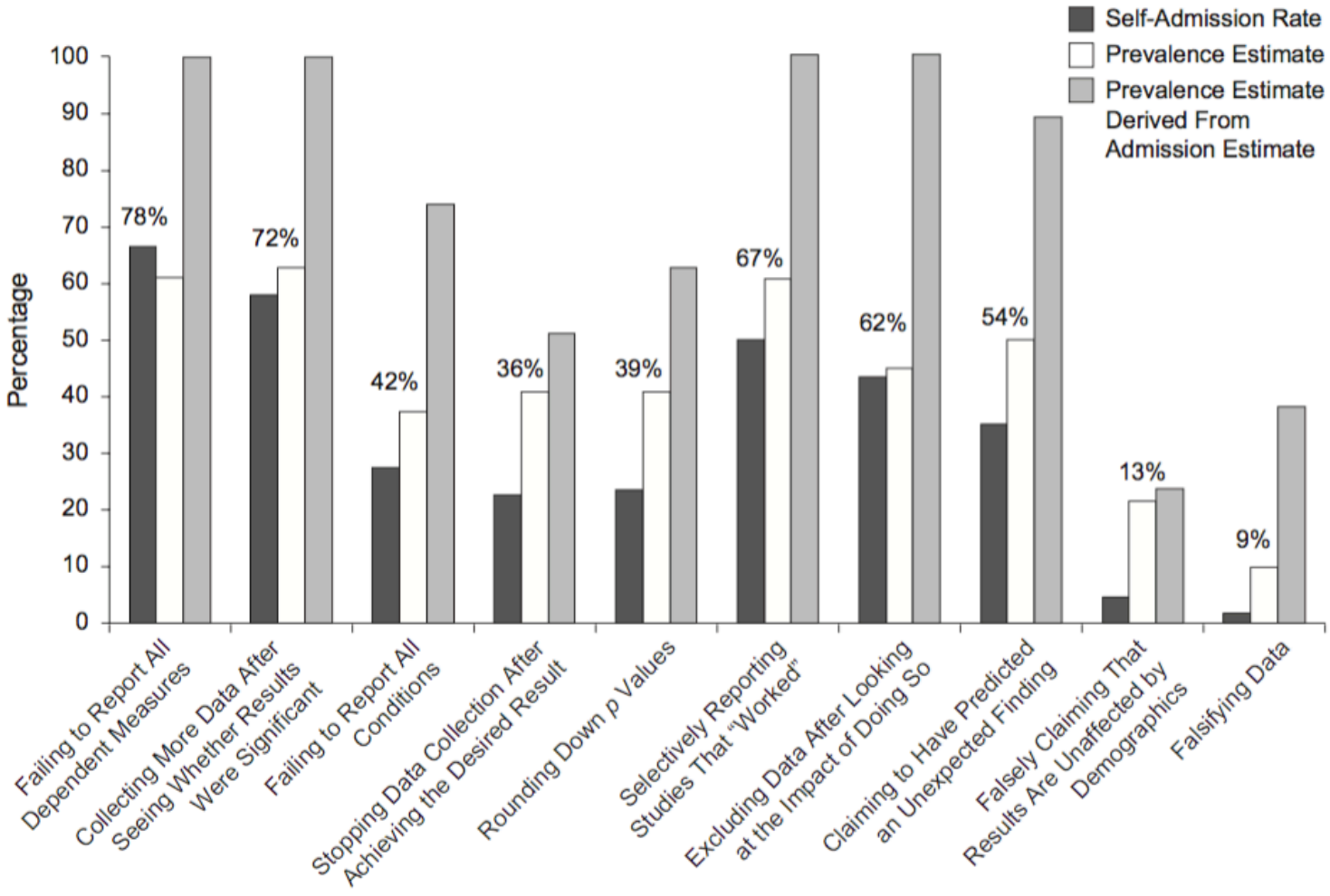
...

Misconduct & Fraud

- **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"

Questionable Research Practices

Survey of 2000 psychologists (John et al. 2012):

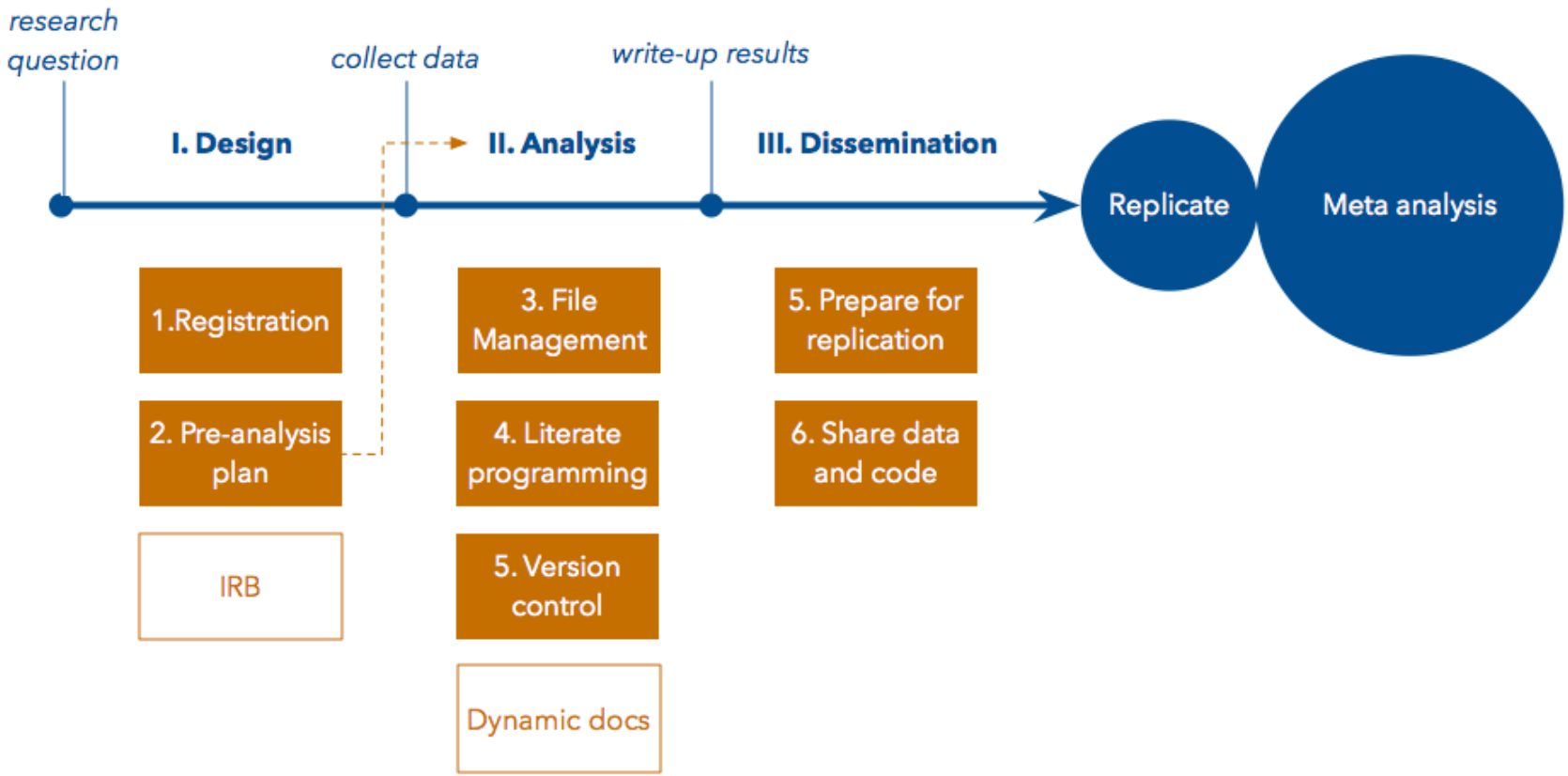


Good news, everyone!

Norms are changing. Smart people are working on these issues in the social sciences and developing standards and tools to help throughout the research lifecycle.

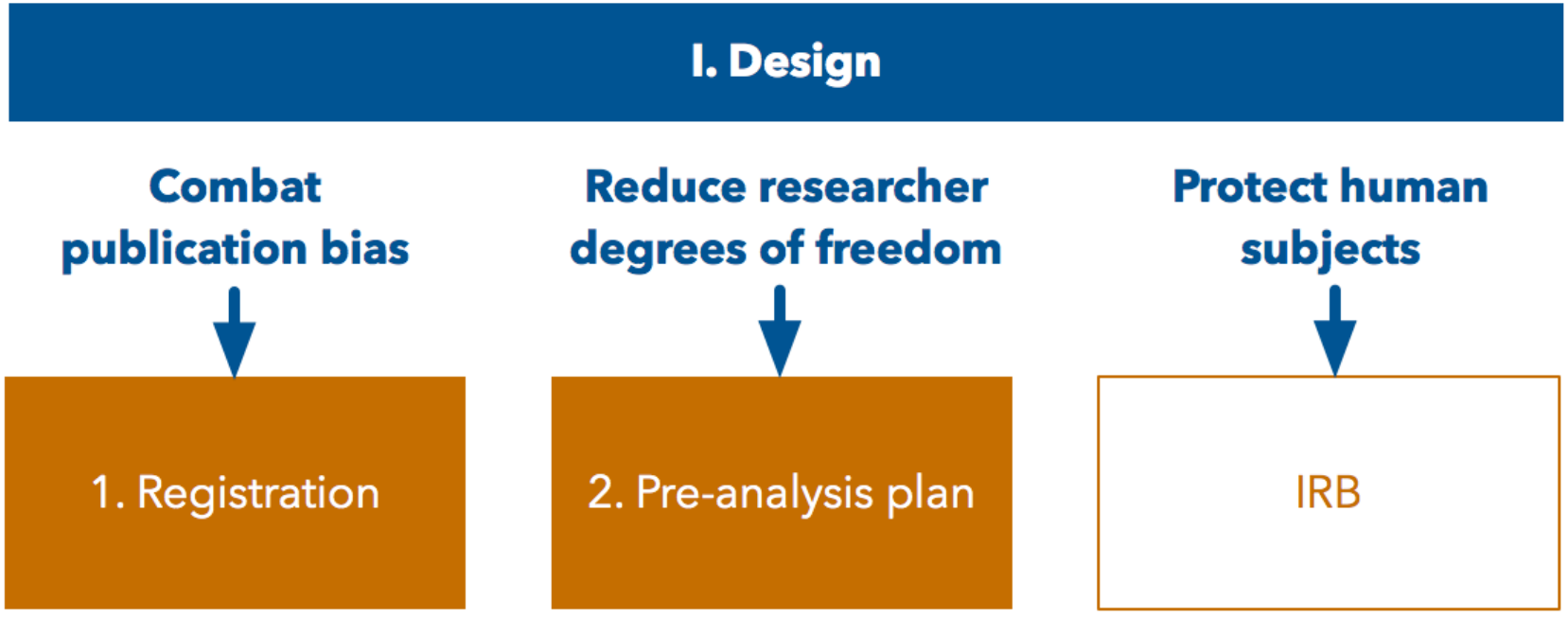
- PDEL, BITSS, OSF, DART, Dataverse, EGAP, etc. etc.

Research Lifecycle



Phase 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 [visibility]; 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

AEA

To register an experimental study with AEA ...

1. Create an account at https://www.socialscienceregistry.org

2. Click on “register a trial” and enter basic information—including title, country, status, keyword, abstract, start and end dates, outcomes, experimental design, whether treatment clustered, planned number of clusters and observations, IRB information

EGAP

To register an experimental (or non-experimental) study with EGAP ...

1. If you're not already in the EGAP author database, go to http://egap.org/node/add/people to add your name and basic information

2. Go to http://egap.org/node/add/registration and complete the registration form—including faculty affiliation, prospective vs. retrospective, whether experimental, start date, background on study, hypotheses to be tested, basic research design, sample size, whether power analysis, IRB information, and keywords

2. Pre-Analysis Plan

About Pre-Analysis Plans (PAP)

- **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, multiple hypothesis testing, 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 adresses p-hacking**—separate confirmatory vs. exploratory analysis

Where to Submit a PAP

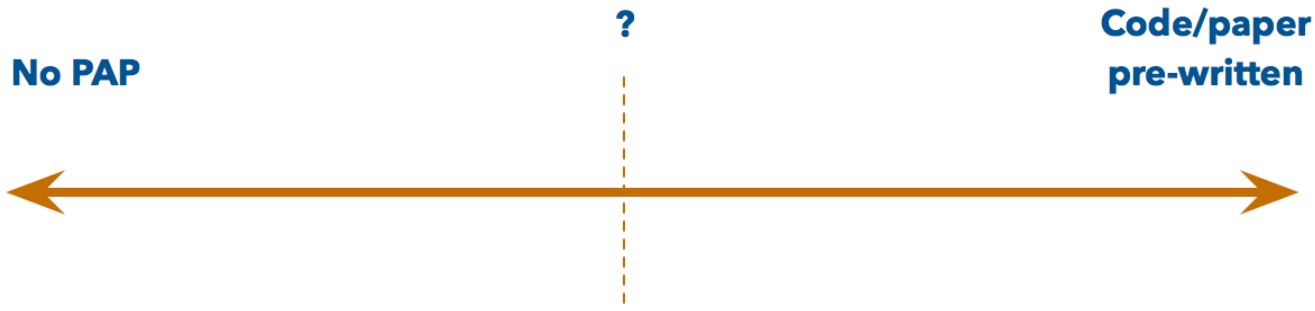
Generally, upload as ***part*** of registration process ...

- 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

No universal standard, recommendations include ...

1. **Background:** Abstract, motivation, questions
2. **Design:** Treatment, sampling & randomization, attrition, spillover, survey instruments, power calculations, plan for data collection, processing & management
3. **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
4. **Research team:** Members, affiliations, conflicts of interest
5. **Logistics:** Fieldwork, timeline, budget

Tie your hands in the right places



$\rightarrow$ PAP requires a lot of thought!

Olken's PAP Checklist (2013)

Item	*Brief description*
Primary outcome variable	The key variable of interest for the study. If multiple variables are to be examined, one should know how the multiple hypothesis testing will be done.
Secondary outcome variable(s)	Additional variables of interest to be examined.
Variable definitions	Precise variable definitions that specify how the raw data will be transformed into the actual variables to be used for analysis.
Inclusion/Exclusion rules	Rules for including or excluding observations, and procedures for dealing with missing data.
Statistical model specification	Specification of the precise statistical model to be used, hypothesis tests to be run.
Covariates	List of any covariates to be included in analysis.
Subgroup analysis	Description of any heterogeneity analysis to be performed on the data.
Other issues	Other issues include data monitoring plans, stopping rules, and interim looks at the data.

Ongoing Debate

- Olken (2013) on "Promises and Perils of Pre-analysis Plans"
- Coffman & Niederle (2015) argue that "Pre-analysis Plans Have Limited Upside, Especially Where Replications Are Feasible"
- More debate on utility for observational work but can be done (see Neumark 2001)

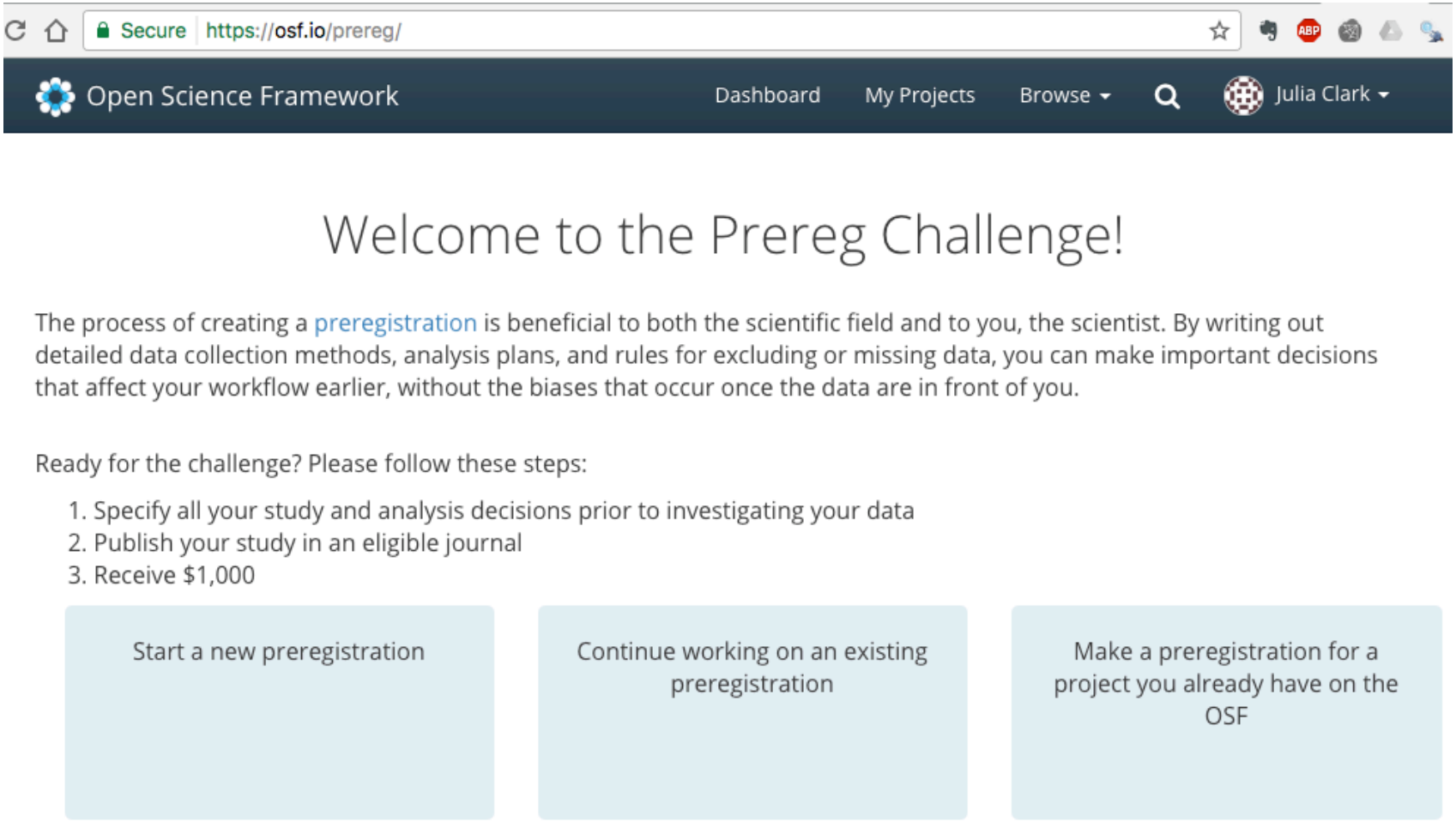
Spotlight on OSF

“A scholarly commons to connect the entire research cycle”

(Trying to be) a one-stop hub for managing projects, collaboration, materials, with integrated registration, pre-analysis plans, file management (Dropbox), version control (GitHub), etc., across scientific disciplines.

... Make an account and explore at https://osf.io/

Win $1000 with Preregistration Challenge



[IRB]

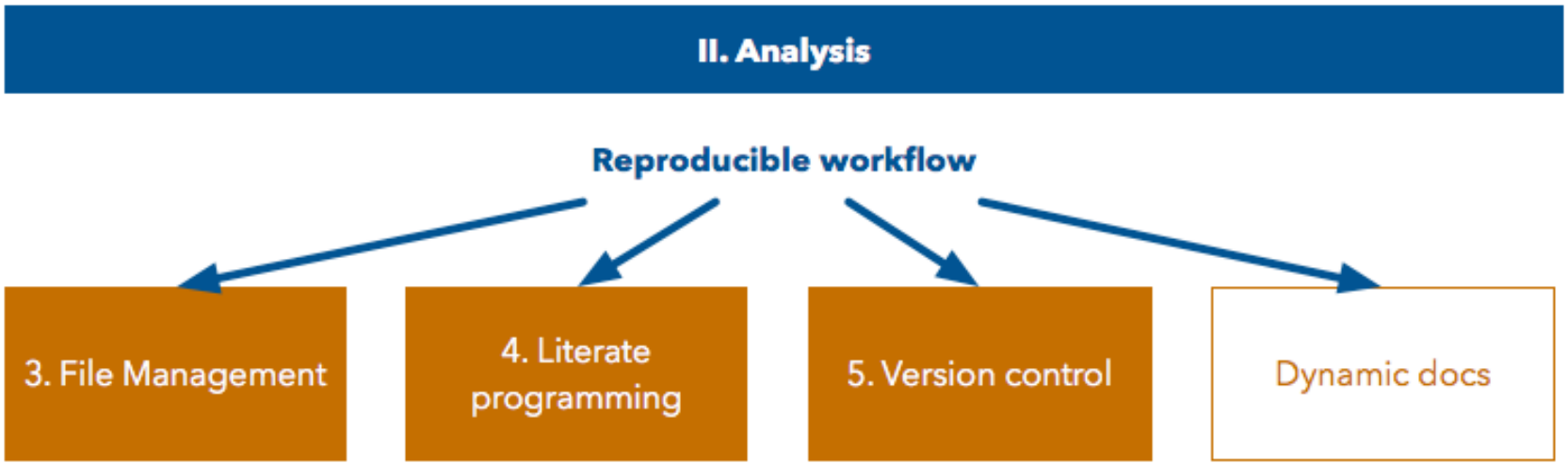
Don’t forget about IRB requirements to protect human subjects!

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

Phase II: Analysis

Steps

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



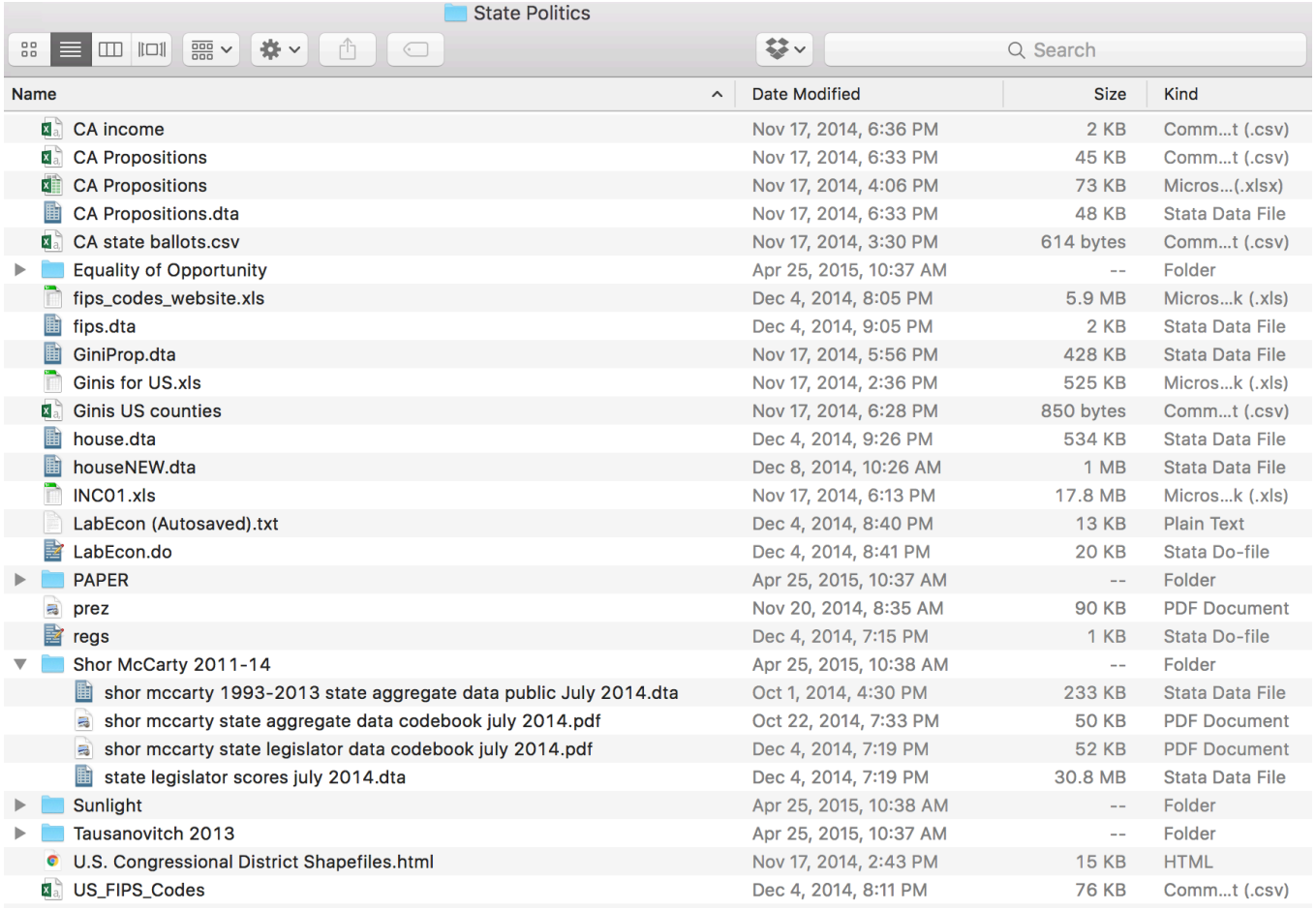
3. File Management

About File Management

- **What:** Organizing and managing files cleanly and intuitively

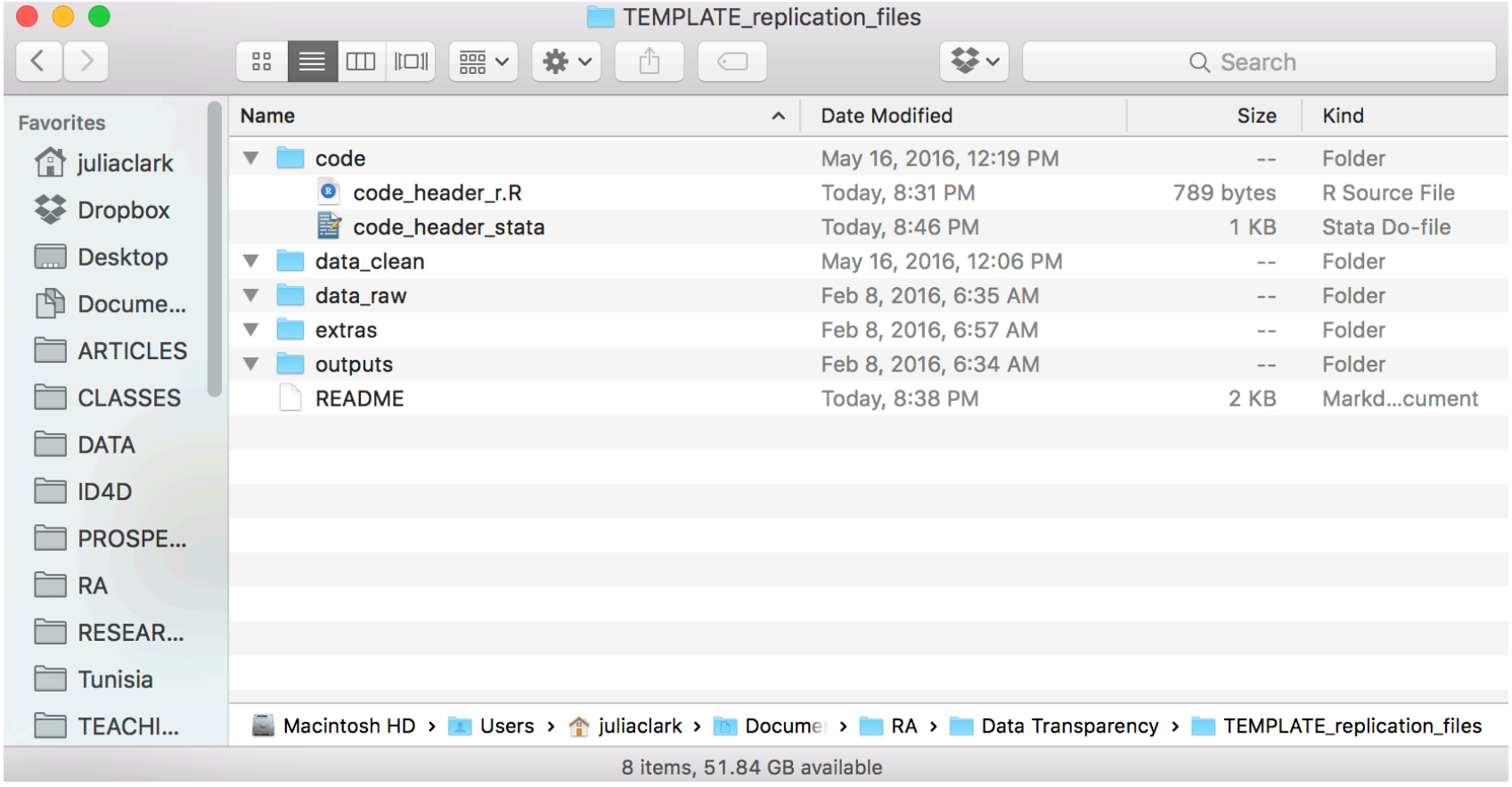
- **Why:** To preserve original data, streamline workflow, and reduce clean-up time when sharing files

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



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
- Difficult to comment/annotate too much
- White-space (e.g., spaces, tabs, returns) is your friend

Structure and Name Files

- Create separate scripts for merging/cleaning and data analysis, with a master-script for running it all
- Give code, data files, and output logical names where possible
    - e.g., Number folders/files sequentially in the order they should be run

Streamline & Annotate Code

- Add headers (see PDEL template)
- Format scripts so they're easily readable—e.g., indent code, use ample line breaks and spaces, standardize comment syntax
- Add comments to improve reader understanding; remove unhelpful/embarrassing comments
- Clearly label code sections, main analyses, outputs
- Give variables intuitive names like `edu_percent` rather than `v76`
- Label variables and values in Stata

Use working directories

R: `setwd("~/Documents/replication_files")`
Stata: `capture cd "~/Documents/replication_files"`

- Saves you time, since you (or someone replicating your study) only have to change the path once if the files move AND your code will be shorter
- Particularly helpful if co-authors alternate between Mac (“/”) and Windows (“\”) file extensions

5. Version Control

About Version Control

- **What:** A system for managing iterative versions of files (code, data, manuscripts) over time and across collabrators

- **Why:** Keep original files, protect work, collaborate efficiently, streamline workflow, etc., etc.

"FINAL".doc

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

- 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 text file log to remind yourself of the location/content of major changes

Or let version control software do this for you!

GitHub

Version control software > Git > GitHub

- **Version control software:** helps manage versions and edits to files (e.g., Microsoft Word's "track changes", or Google Doc's "suggestion" feature)—many options!
- **Git:** Open-source, "distributed model" of version control developed by creator of Linux
- **GitHub:** Free, web-based service that hosts Git "repositories" and offers a variety of features for collaboration

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

How do I use GitHub?

- **GitHub website**—necessary for collaboration, but limitations
- **GitHub Desktop**—free desktop client for Windows/Mac, more user friendly than website
- **Command line (shell)**—optimal for advanced users

How to think about Git: A Metaphor

Tell Big Brother Git to watch a set of files (a “repository”) and it tracks every change within them, line-by-line.

What GitHub is NOT ...

1. **Dropbox.** Git tracks changes in files and GitHub let's you sync/track/view/edit those changes online and with collaborators. Dropbox is a hard drive in the cloud.

2. **File Manager/GUI.** GitHub desktop may *look* sort of like Finder (Mac) or Explorer (PC), but its function is not to let you find, click on, or open files; it's unit of operation is ***changes within those files***.

Also, most useful for text files (.txt, LaTeX, Markdown, Stata, R code, etc.); not useful for PDF, Word, Excel.

GitHub Prep

If you haven't already ...

- **Make sure you have a good text editor.** Notepad (PC) and TextEdit (Mac) will work (if you set TextEdit default file format to `.txt` and not `.rtf`). Or get a more powerful editor like Atom.
- **Create an account at GitHub.** This gives free *public* repositories, but go to the education section and click "request a discount" to get free *private* repositories.
- **Download and install GitHub Desktop.** Then open and log in using your GitHub account.

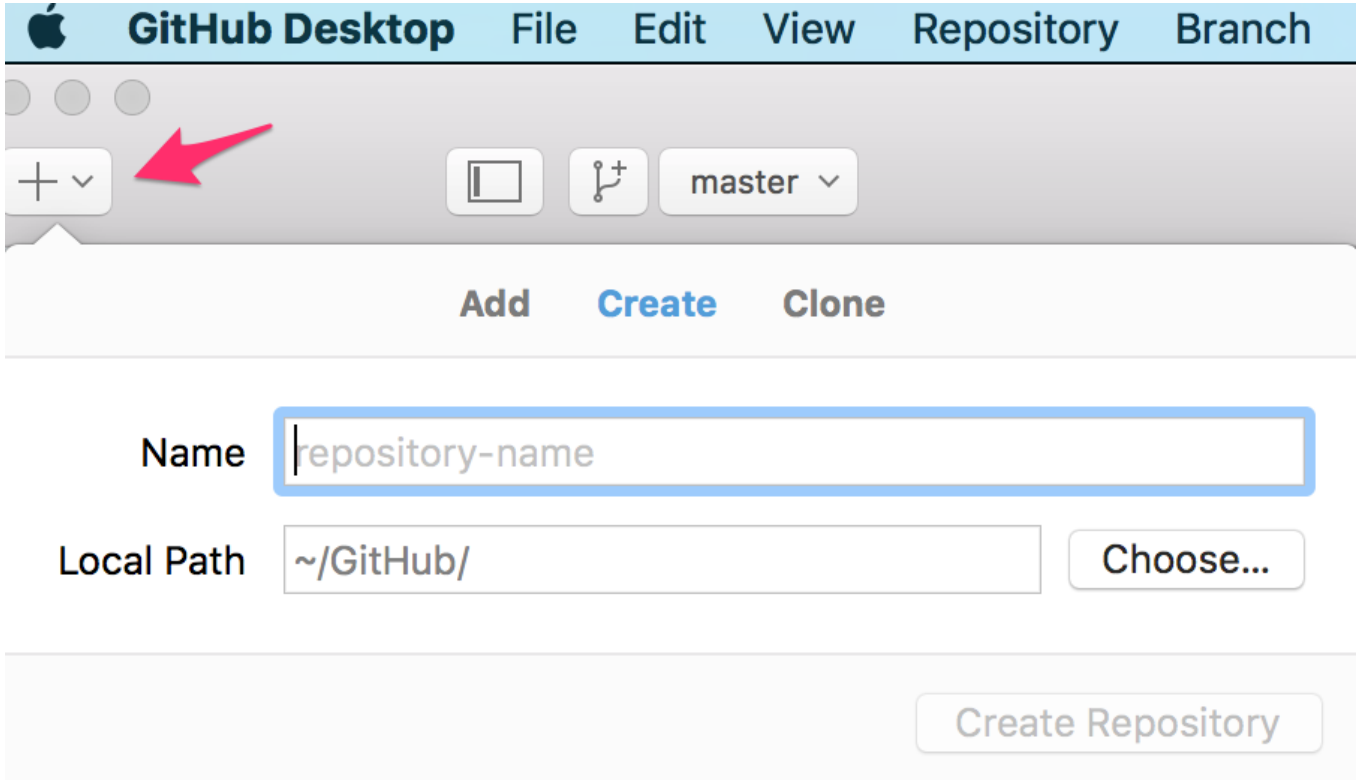
Basic vocabulary

- **Repository:** A set of files (in a folder) that you have told Git to track, along with its associate .git files. A local repository is the copy on your computer; a remote repository is the copy synced online.

- **Commit:** A labeled change or series of changes to files. Git tracks every change you make, and then you group these changes as desired into a “commit” that can be commented on, reverted, etc.

1. Create a NEW repository

Within **GitHub Desktop**, click on "+" and then "create" to make a new repository with a name and location of your choice. This creates a new folder that will be empty except for some hidden files (e.g., a .git directory).



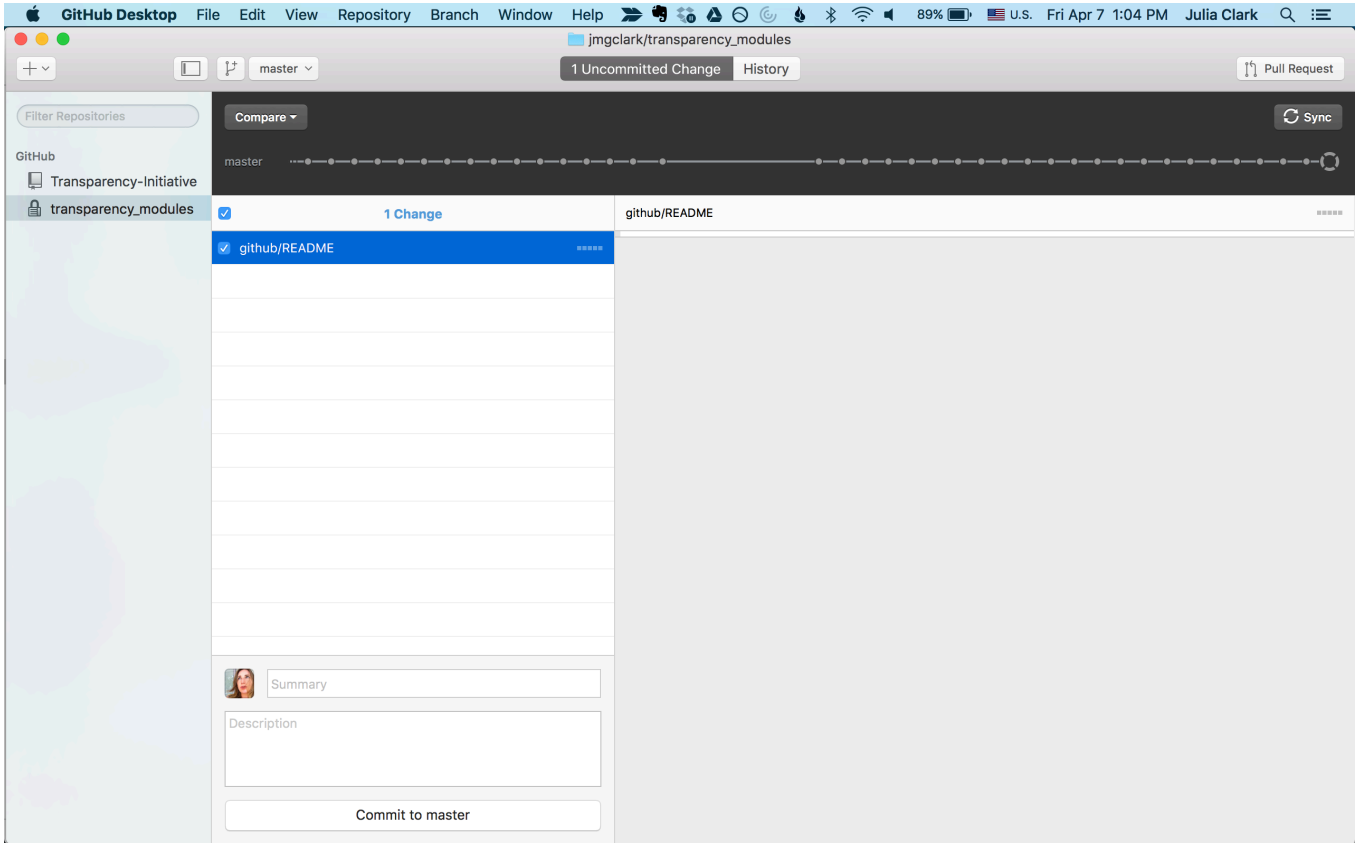
2. Add a text file to your repository

- Create a new text file on your computer called “README” as you normally would, and ***save it in your repository location***.

- This should be a plain text file (.txt) or Markdown file (.md), not a rich text format file (.rtf).

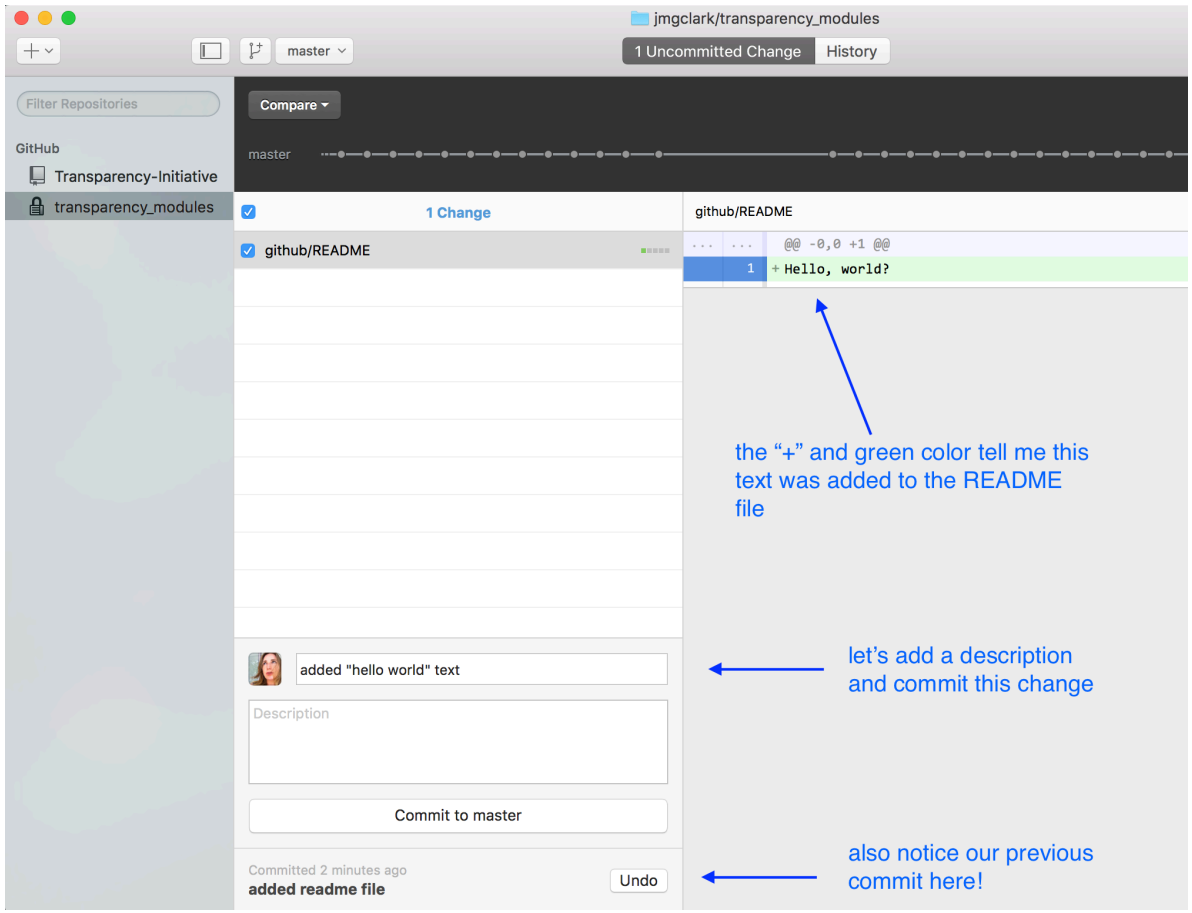
3. Commit this change in GitHub Desktop

If you open up the desktop client, it will look something like this:



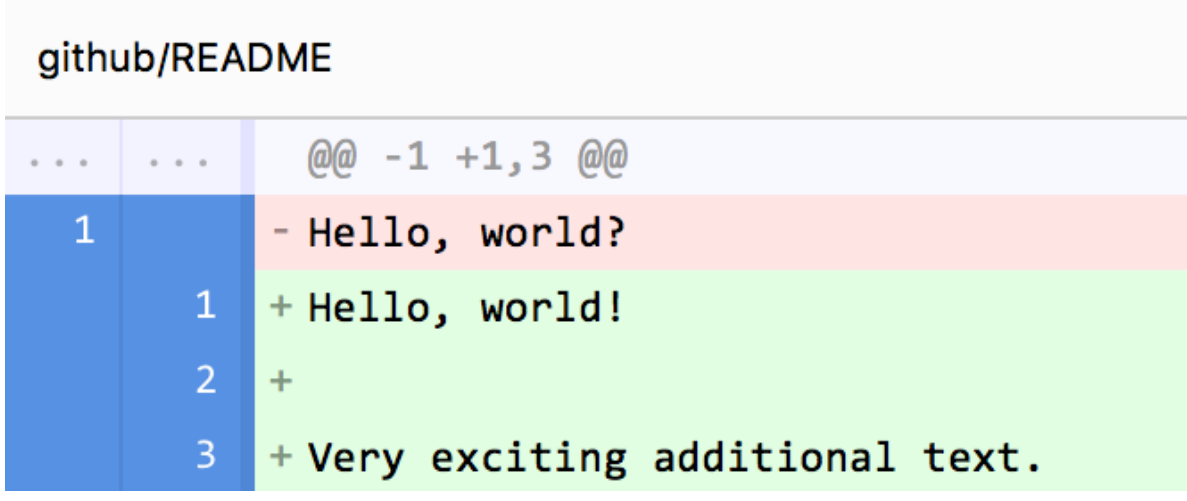
4. Add text to README and commit changes

Add some text to your file and save. If you go back to the Desktop client, you will now see something like this:



5. Edit README text and commit changes

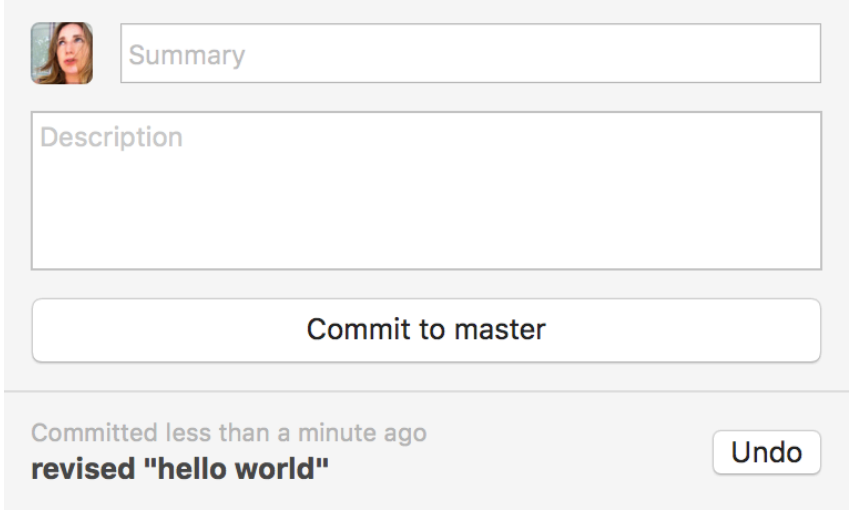
Make and save changes to your text, then go back to GitHub Desktop. In the right-hand pane, additions will appear in green and deletions will appear in red:



Note that the unit of change is the ***paragraph***, so changing “?” to “!” involved deleting/adding the whole phrase.

6. Undo the last commit

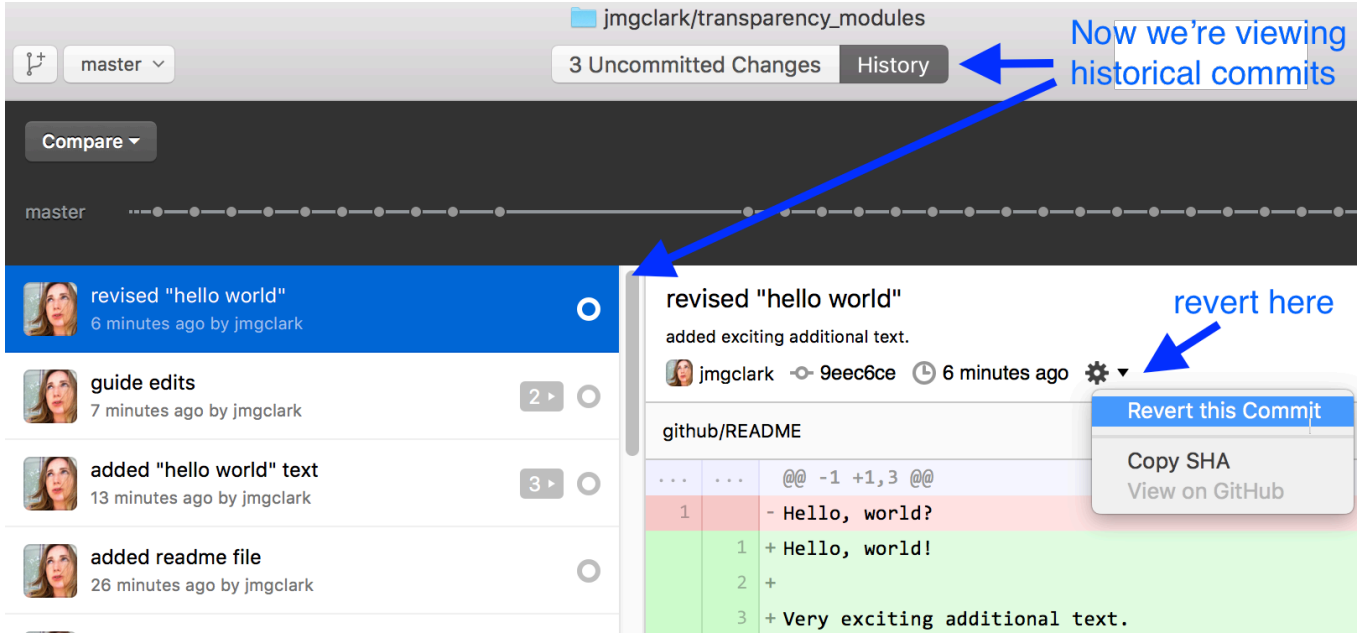
If you're unhappy with your LAST commit (i.e., you disliked how it was grouped or labeled), you can **click the "Undo"** button at the bottom of the list of uncommitted changes:



Now, these changes will appear again in the "uncommitted changes" window for you to regroup or relabel.

6. Revert a previous commit

If you're unhappy with the CHANGES in a commit themselves, you can undo them by **reverting** a commit. To do this, **switch to the "History" tab** at the top of the window and view all your previous commits. Select the commit, navigate to the dropdown menu, and click **"Revert"**:



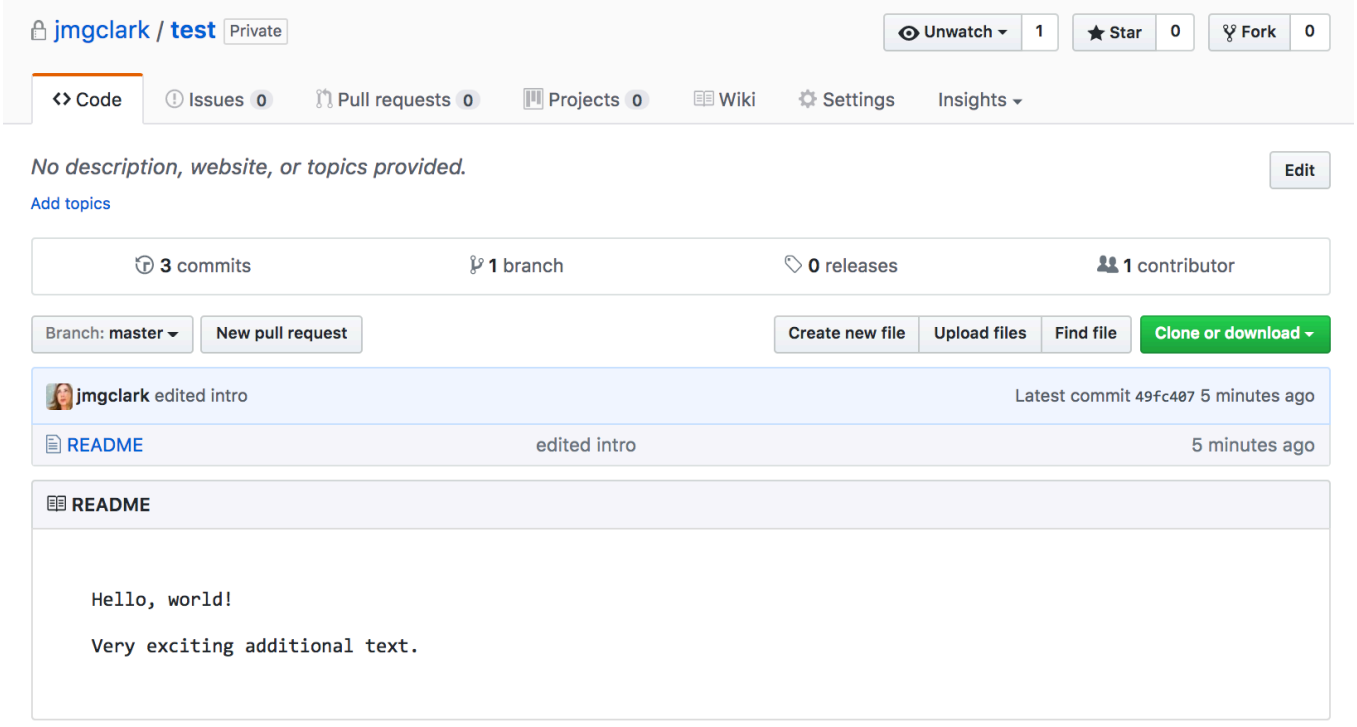
7. Publish repository to your online account

So far, we've been working in a **local repository**–one that that you created on your computer.

This may be useful to you, but to collaborate on projects you'll need to publish the repository to the web (i.e., make a **remote repository**). → Do this by **clicking "publish"**:

8. View your repository & changes online

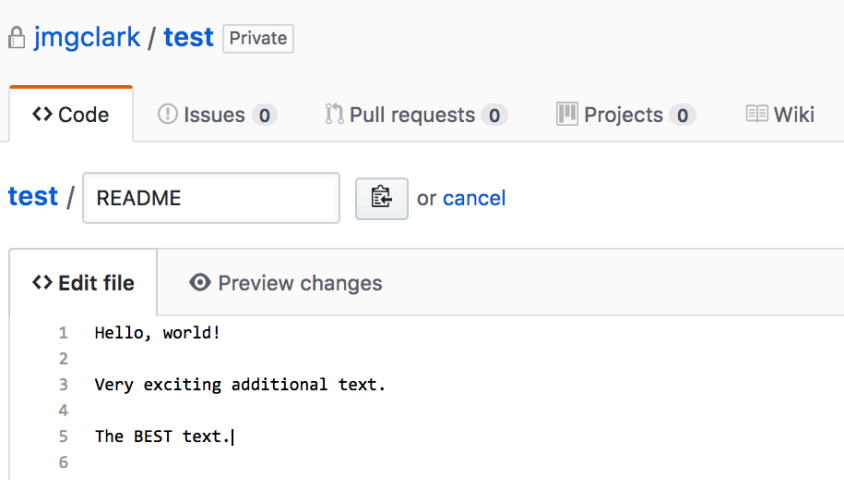
When you login to GitHub online, you'll see the new repository and file you've added.



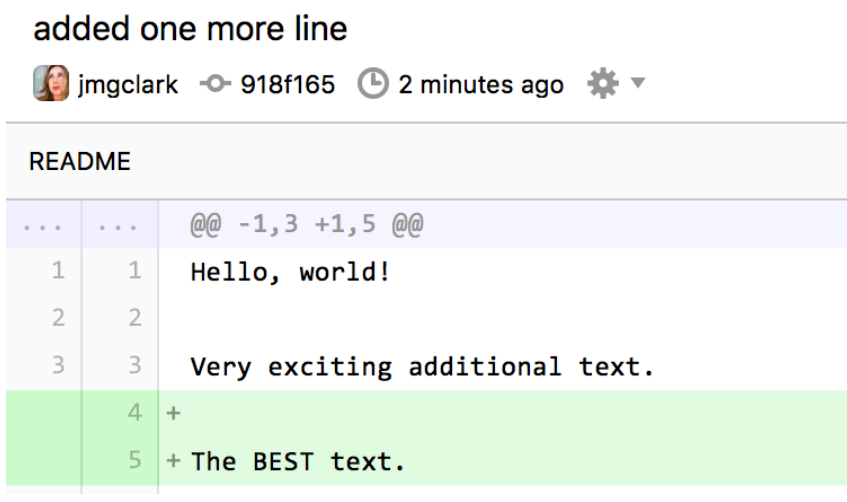
9. Edit the file online & sync with local repository

Click on the README file and then click the edit button (the pen). (A) **Make some changes and then commit**. Then go back to the Desktop client and **click "Sync"**. (B) Your new commit will appear in the **history tab**.

(A)



(B)



What's next?

That was very very basic. To really use Git, explore these great features with weird names ...

- Cloning **online repositories**—copies an online repository onto your local hardrive so you can work offline
- Forking **online repositories**—duplicates ***someone else's*** shared repository and lets you use/change/build on it without affecting their original work
- Branching **a repository**—lets you (and collaborators) experiment with changes that can be merged into the "master" version of the files
- **Initiating a** pull request—submits your commits to be merged into a forked/branched repository (accepted/rejected by collaborators)

Git Resources

Too many to name, but some good places to start:

- Gentle intro to version control
- GitHub and collaborative writing in academia
- Forks and pull requests
- Non-programmer's intro to Git using command line
- Fork-branch workflow using command line (but useful to read for Desktop as well)

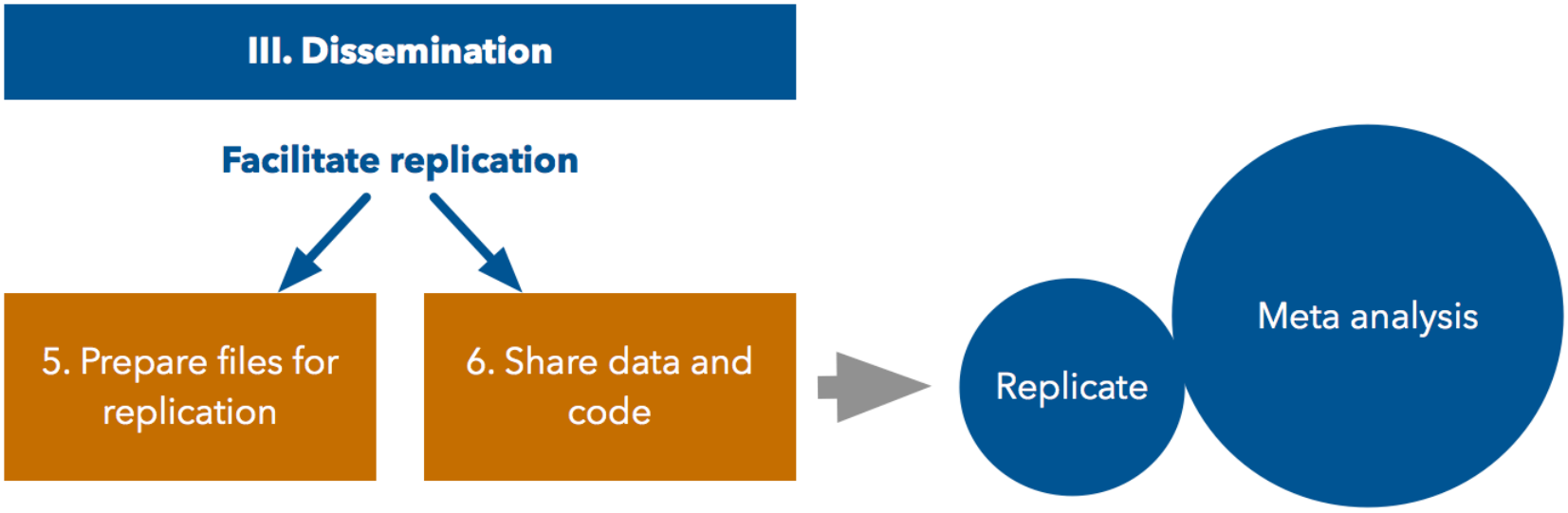
[Dynamic Docs]

You can take reproducible research a step further by integrating code ***into your manuscript***. No need to copy-and-paste results from statistical software (annoying and error-prone); just click the button and everything reproduces.

- RMarkdown
- Stata Markdoc or Stata texdoc

Phase III: Dissemination

Steps


III. Dissemination
Facilitate replication
5. Prepare files for replication
6. Share data and code
Replicate
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!) over 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

Check Analysis

Easier to start with final analysis and work backwards to data cleaning/merging.

1. Copy original analysis script(s) into `replication_files/code`
2. Copy cleaned dataset(s) used for analysis into `replication_files/data_clean`
3. Run code without changes (except for wd)
4. Fix any bugs in the code, address discrepancies with previous results

Check Data Clean/Merge

1. If separate from analysis, copy original merge/cleaning script(s) into `replication_files/code`
2. Copy original dataset(s) to `replication_files/data`
3. Run merge/clean code without changes (except for wd)
4. Rerun the analysis code from above on the newly cleaned/merged data file
5. If you get different results than step #1, there is a discrepancy with merging/cleaning code—fix it!

3. De-Identifying Individual-Level Data

Now you know the code works and replicates, congratulations! The next step is to ensure that any shared files ***do not contain*** data that could be used to identify individuals.

Purpose:

- Ensure you are protecting individuals' identity and private information—this is an ethical issue for researchers, and a 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 that is explicitly linked to the subject—*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*, *classes 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

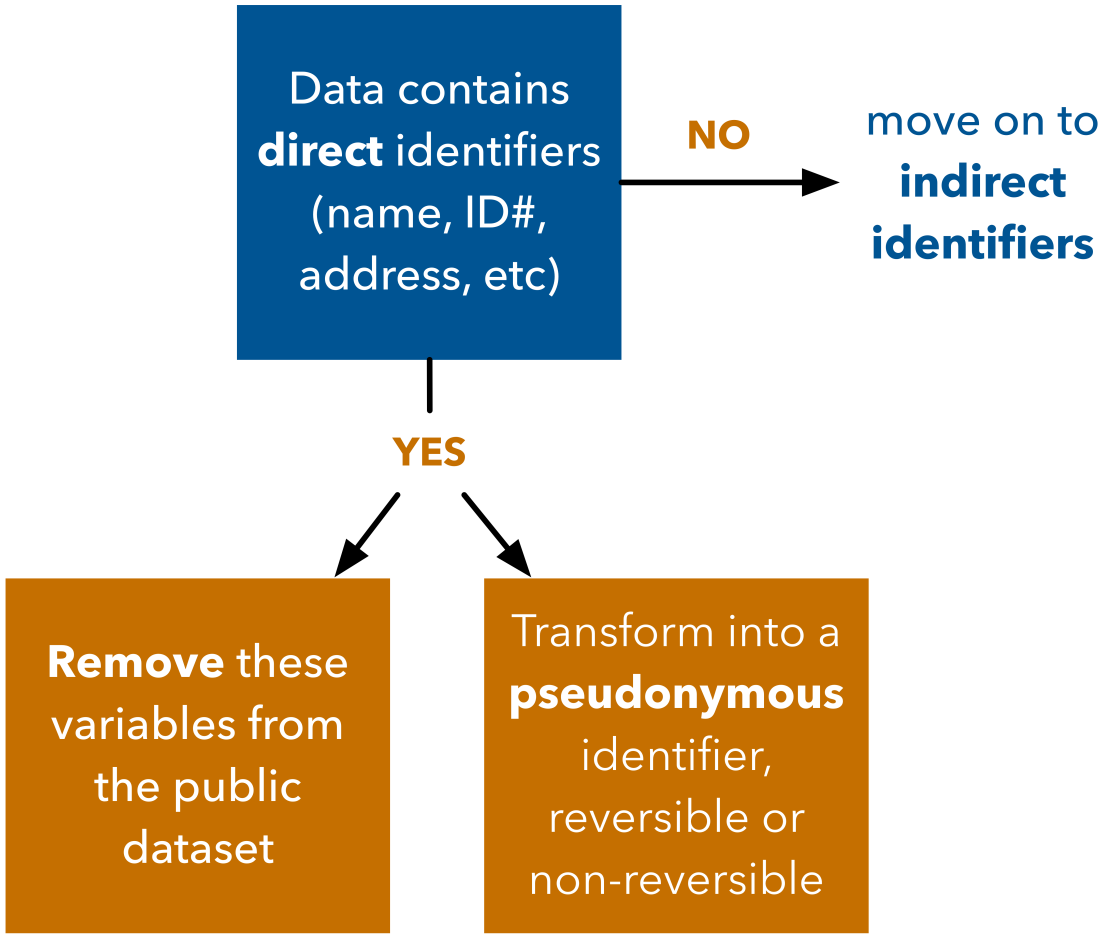
Dealing with Direct Identifiers

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: replace identifiers with "pseudonyms" that may be reversible or non-reversible—e.g., give people random names or ID numbers—goal is to be able to link datasets

Solutions for Direct Identifiers



What is Sufficient De-Identification for Indirect Identifiers?

1. **Determine Risk** = Pr(de-identifying) $\times$ 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

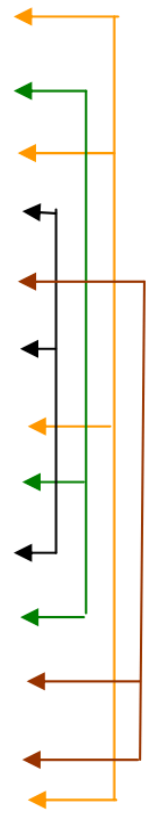
The Problem

ID	Study	Pub Year [§]	Health data included?	Profession of adversary	Number of individuals re-identified	Country of adversary	Proper de-identification of attacked data ?	Re-identification verified ?
A	[70]	2001	No	Researchers	29 of 273	Germany	"Factually anonymous"	Yes (records containing insurance numbers only)
B	[71]	2001	No	Researchers	75% of 11,000	USA	Direct identifiers removed	No
C	[67]	2002	Yes	Researcher	1 of 135,000	USA	Removal of names and addresses	Yes
	[56]	2003	No	Researchers	219 unique matches, 112 with 2 possibilities, 8 confirmed	UK	Yes	Verified matches, but not identities
D	[22]	2006	No	Journalist	1 of 657,000	USA	No	Yes (with individual)
E	[72]	2006	Yes	Researchers	79% of 550	USA	No	Verified (with original data set)
	[73]	2006	No	Researchers	Of 133 users, 60% of those who mention at least 8 movies	USA	Direct identifiers removed	No
F	[52]	2006	Yes	Expert Witness	18 of 20	USA	Only type of cancer, zip code and date of diagnosis included in request	Yes (verified by the Department of Health)
G	[74]	2007	No	Researchers	2,400 of 4.4 million	USA	Identifying information removed	Verified using original data
	[53]	2007	Yes	Broadcaster	1	Canada	Direct Identifiers removed & possibly other unknown de-id methods used	Yes
H	[23]	2008	No	Researchers	2 of 50	USA	Direct identifiers removed+maybe perturbation	No
I	[75]	2009	Yes	Researcher	1 of 3,510	Canada	Direct identifiers removed	Yes
J	[76]	2009	No	Researchers	30.8% of 150 pairs of nodes	USA	Identifying information removed	Verified using ground-truth mapping of the 2 networks
K	[57,58][???]	2010	Yes	Researchers	2 of 15,000	USA	Yes - HIPAA Safe Harbor	Yes

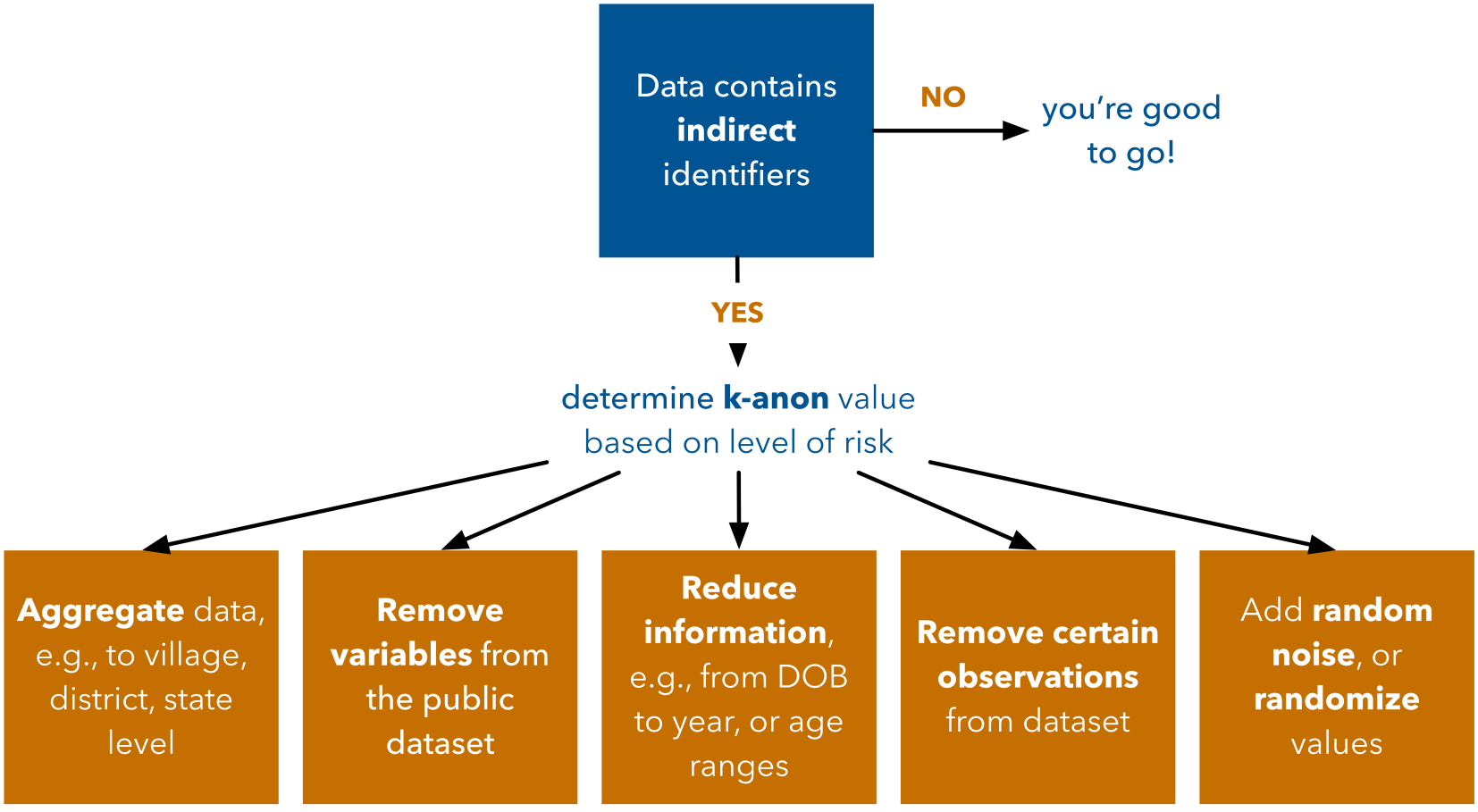
Source: El Emam et al. 2015. "A Systematic Review of Re-Identification Attacks on Health Data." PLOS One.

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



Solutions for Indirect Identifiers



Trade-off: Usefulness $\Longleftrightarrow$ Anonymity

- **Aggregating**—lose ability to replicate any individual-level analysis
- **Removing variables**—may not be able to replicate specific models
- **Reducing information**—adds noise to models
- **Remove observations**—adds bias if non-random
- **Adding random noise/values**—adds noise, obviously

See here and here for more discussion of appropriate thresholds, methods, and tools for de-identification.

Good Practices

- Include code for de-identified data for transparency (as long as the code itself doesn't compromise anonymity)
  - e.g., censor code that sets the seed for a random draw to generate new ID numbers and could be used to re-identify individuals
- If identifiers ***aren't*** used for analysis, de-identify early in merging/cleaning process
- Store original data with PII securely—if you're using Dropbox, see PDEL GitHub wiki for tips on sharing with RAs in a way that protects PII data

4. Edit and Organize Files for Clarity

Now we have working files that are de-identified; the next step is to clean and annotate them 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
- Track and document packages, software versions
  - R: `sessionInfo()`
  - Stata: `version`

5. Final Replication

Now that you have cleaned/reorganized script files …

- Shutdown or clear your Stata/R memory
- Rerun the entire process—including data merging, cleaning and analysis—to make sure the editing process 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 study materials to an online repository
- **Why:** lasts longer than personal website, more searchable, future proof
- **Concerns:**
  - Can usually be embargoed, or you can provide only what is necessary for replication (e.g., leave out other 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

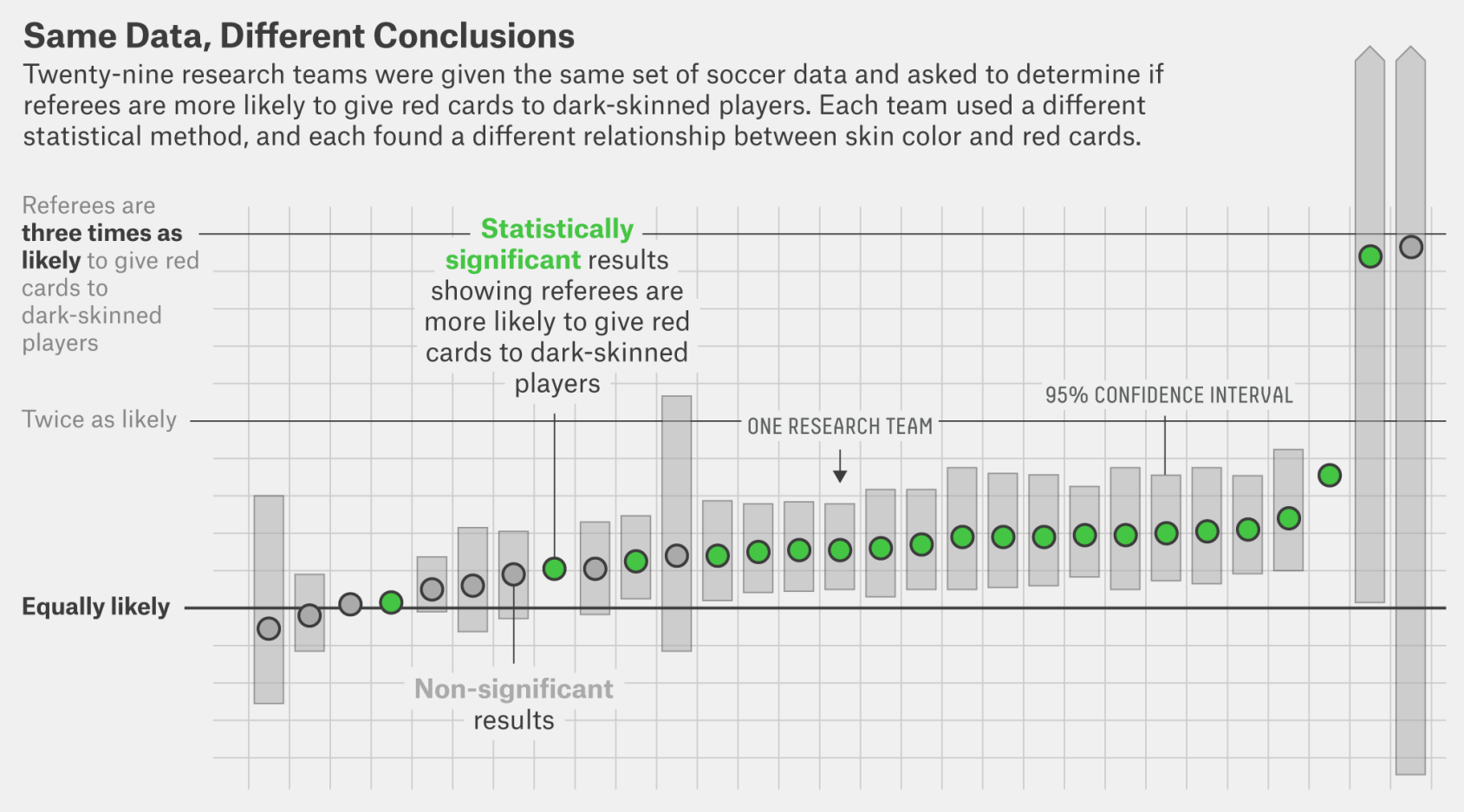
About Meta-Analysis

- **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 individual study estimate may be biased or contain random error

One Study = One Data Point

That experiment you just ran with 3,685 participants? It's one data point among many other studies. ***Even with the same data***, results may vary ...



Who Does Meta-Analysis?

- Campbell Collaboration (policy)
- 3ie (development)
- Cochrane Collaboration (medicine)
- What Works Clearinghouse (US Gov't, Education)
- CLEAR (US Gov't, Labor)
- MAER-NET (Economics)
- You!

Basic Steps

Using a PAP or “protocol” (and assuming NO publication bias!) ...

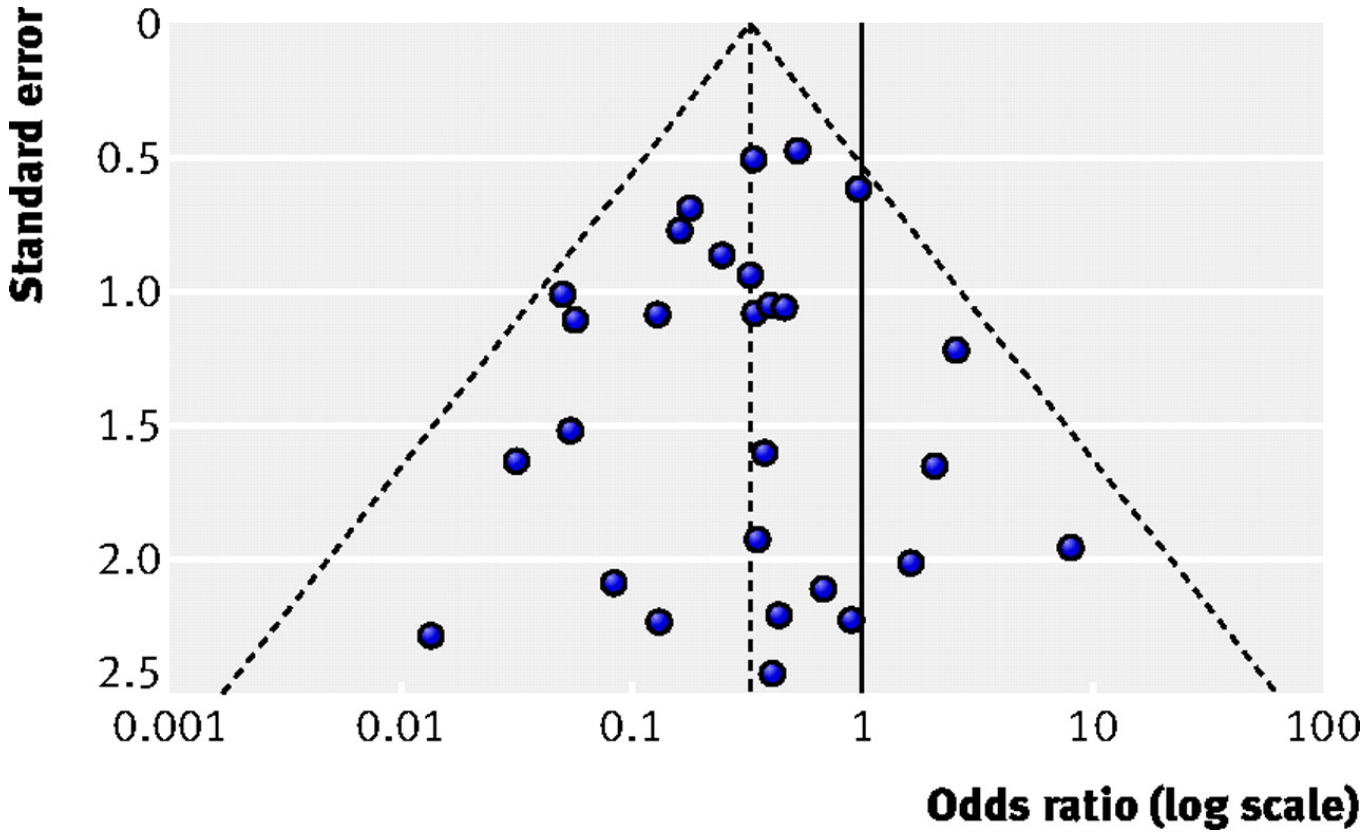
1. Determine which studies to include

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. study prevision (e.g., SE of study treatment effect)



Source: BMJ 2011

Etc.

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)