Data Science Applications in Politics and Policy
Lecture 3

Melissa Sands
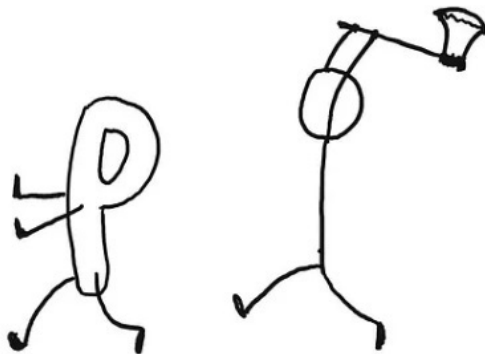
London School of Economics

GV4L3

Overview

P-HACKING

p-hacking; image source: Twitter KC

At the Researcher-level...

Science has always relied on transparency:

- Philosophy
- Mathematical/theoretical proofs
- Engineering (which required the creation of the patent system)
- Observational data
- Experimental data

As science has become bigger (more scientists doing more science) and increasingly empirical (greater emphasis on data and analysis), the transparency problem has also become bigger.

What about solutions?

- Norms are ever-changing

- Tools and "standards" to be applied throughout the research lifecycle.

  - Berkeley Initiative for Transparency in the Social Sciences (BITSS): https://www.bitss.org/

  - Open Science Framework (OSF): https://osf.io/

  - Dataverse: https://dataverse.org/

  - Evidence in Governance and Politics (EGAP): https://egap.org/

  - Institute for Replication (I4R): https://i4replication.org/

- Main emphasis: Ways to improve transparency in why, what, and how researchers do science.

The BITSS perspective looks like this:
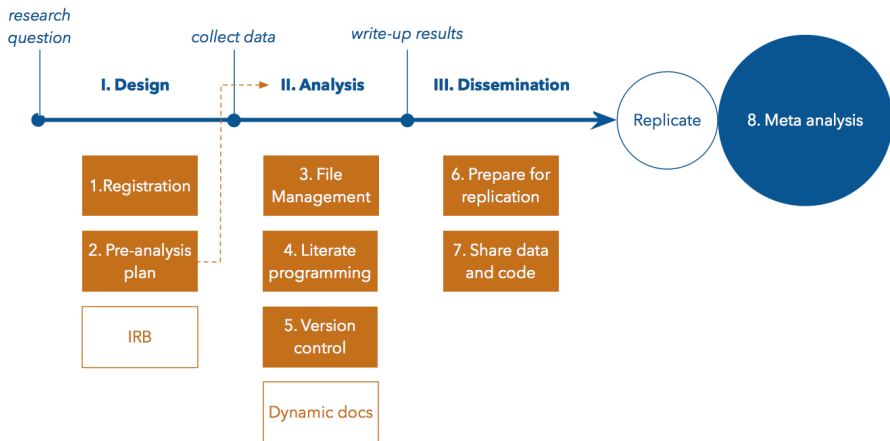
Overview

Registration

- **What:** Enter your study into a "registry"
    - Now required for experimental studies submitted to *Journal of Politics*
    - In many other review contexts you will be penalised for not having one

- **Why:** Combat the file-drawer problem and publication bias.

- **Where:**
    - Open Science Framework: http://osf.io
    - American Economics Association (AEA): http://socialscienceregistry.org
    - As Predicted (Wharton): http://aspredicted.org

Pre-Analysis Plans (PAPs)

- **What:** Detailed description of research design and data analysis plans, submitted to a registry *before* looking at the data.
  - Rule of thumb: PAP should be specific enough such that an expert from the same subfield could conduct and replicate all of the analyses on their own if they had 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
  - Transparent methods make it easier for others to build on your work

Registration typically includes a pre-analysis plan.

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

One step further: Registered reports

A registered report is a slightly newer idea, and is like a paper combined with a PAP.

- You literally write the paper, write your code, and then submit the paper without having collected or analysed any data.
- Editors and reviewers then judge the paper based purely on the motivation for writing it, not the results you get.
- *APSR* and *JOP* (among many other journals) both have dedicated registered report streams.

This takes a particular philosophical perspective to its (somewhat natural) conclusion: Science should be judged without knowing the results.

Discuss: What are the up-sides and down-sides of this perspective?

What to include in a PAP (or RR)

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, etc.), hetereogenous treatment effects (subgroups, interactions), covariates, standard errors, corrections for multiple hypothesis testing, missing values, outliers
Team	members, affiliations, conflicts of interest

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.
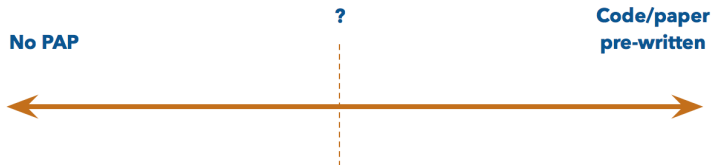
→ **requires a lot of forethought!**

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"

- Ofosu & Posner (2023) weigh the arguments for and against

- More debate on utility for observational work but can be done (see Neumark 2001)

Research Ethics

The New York Times

Syphilis Victims in U.S. Study Went Untreated for 40 Years

By JEAN HELLER
The Associated Press

WASHINGTON, July 25—For 40 years the United States Public Health Service has conducted a study in which human beings with syphilis, who were induced to serve as guinea pigs, have gone without medical treatment for the disease and a few have died of its late effects, even though an effective therapy was eventually discovered.

The study was conducted to determine from autopsies what the disease does to the human body.

Officials of the health service who initiated the experiment have long since retired. Current officials, who say they have serious doubts about the morality of the study, also say that it is too late to treat the syphilis in any surviving participants.

Doctors in the service say they are now rendering whatever other medical services they can give to the survivors while the study of the disease's effects continues.

Dr. Merlin K. DuVal, Assistant Secretary of Health, Education and Welfare for Health and Scientific Affairs, expressed shock on learning of the study. He said that he was making an immediate investigation.

The experiment, called the Tuskegee Study, began in 1932 with about 600 black men,

The Tuskegee Syphilis Study (1932–1972) was a U.S. Public Health Service study in which Black men with syphilis were deliberately left untreated without their informed consent, even after effective treatment became available. The study is now regarded as a profound ethical failure and a key catalyst for modern research ethics regulations.

Human subjects research protection

IRB (Institutional Review Board) requirements to protect human subjects $\longrightarrow$ necessary for ethical research, though not always sufficient

Key regulatory bodies in the UK:

- UK Research and Innovation (UKRI): Sets ethical standards for funded projects
- Health Research Authority (HRA): Oversees the UK Policy Framework for Health and Social Care Research.
- **University research ethics committees**, e.g., https://info.lse.ac.uk/staff/divisions/research-and-innovation/research/committees-and-working-groups/Research-ethics-committee

For more on ethics in experiments, see EGAP Committee Memo https://egap.org/wp-content/uploads/2020/05/egap-committee-memo-on-the-report-of-the-apsa-ad-hoc-committee-on-human-subjects-research.pdf

Overview

1. Design

2. **Analysis**

3. Dissemination

4. Institutional/Discipline Level Solutions

The big picture

Reproducibility $\iff$ transparency

A reproducible workflow needs to account for at least the following steps:

1. Ingestion: Acquiring and reading raw data
2. Transformation: Turning raw data into usable data
3. Analysis: Extracting meaning from the usable data
4. Output: Formatting and producing results (e.g. tables or figures)
5. Reporting: Putting the results into a shareable document (e.g. slides or papers)

Worth knowing: In the data science/analytics engineering world, the first two parts are called an ELT (extract, load, transform) or ETL pipeline.

The big picture: Ultimate goal

Ultimately, a reproducible workflow should be:

1. Code-based, structured
2. Documented, literate, version-controlled
3. Functional, modular, tested, DRY
4. Dependency-proof
5. Automated

We will cover tiers 1 and 2 now (and do some of 3-5 in Seminar). If you are doing 1 and 2, you are doing better than most social scientists!

The big picture: The BITSS Approach

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

File management
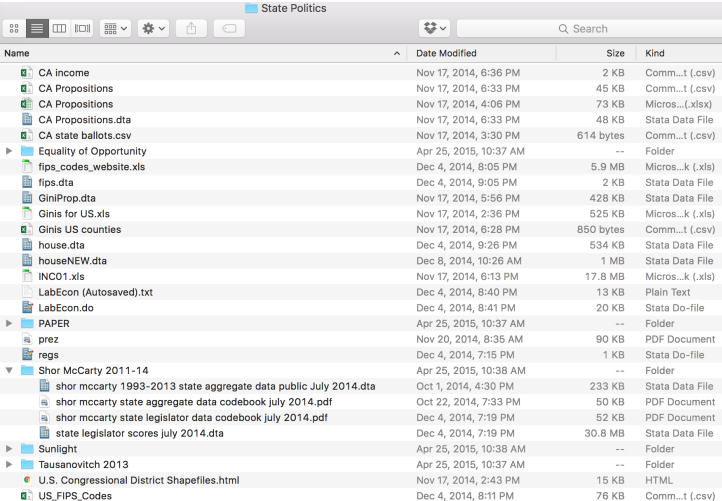
- **What:** Organizing, manipulating, and managing files hygienically, transparently, and intuitively

- **Why:**
  - Preserve original data
  - Improve error detection
  - Streamline workflows
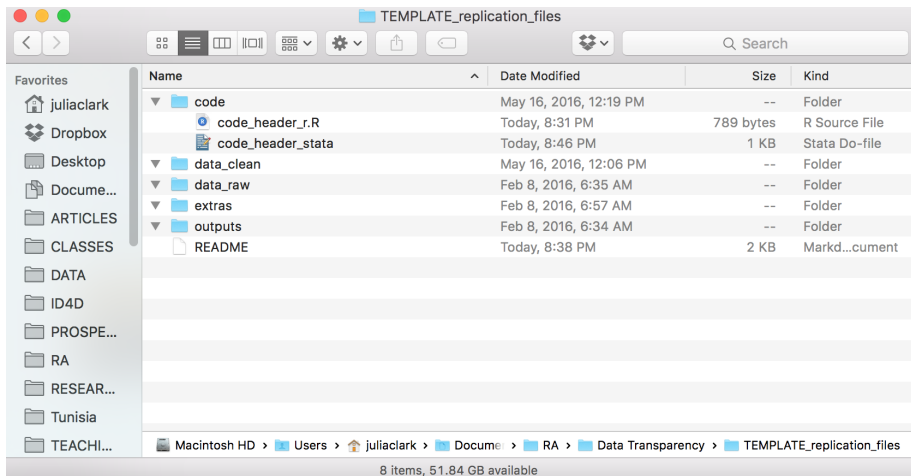 - Reduce overhead

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

Instead, use (something like) PDEL template

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

Literate programming

- **What:** Writing code that it's legible to *humans*.

- **Why:**
  - Document what you are doing and why
  - Allow others to reproduce (and ultimately replicate) your work
  - Avoid mistakes

(The most) basic principles

- Structure and name files, output, variables logically and intuitively
  - Number scripts sequentially in the order they should be run (e.g., `1_main_analysis.R`, `2_robust_checks.R`)
  - Label output with descriptive names that aren't likely to change (e.g., `figure_hte.png` is much better than `figure_1.png`))
  - Label variables intuitively (e.g., `edu_percent` rather than `v76`)

- Thoroughly comment code as you go, or use something like R Markdown or a Jupyter Notebook
  - Add headers to script (see PDEL template)
  - Use this to document every decision you make.
  - Should cover the what, the why, and how.
- Make the contents of scripts and files easy to navigate
  - Create separate scripts for processing and analysing data, with a runner-script for executing it all
- Streamline (DRY) code to minimize repetition

Header example

```r
1 - ######## INFO ########
2
3   # PROJECT
4     # Paper:|
5     # Authors:
6
7   # R Script
8     # Purpose:
9     # Created: <date> by <author> # you don't need this if using Git!
10    # Updated: <date> # you don't need this if using Git!
11    # Inputs: <files required>
12    # Outputs: <tables and figures>
13
14 - ######## SETUP ########
15
16    rm(list = ls()) # clear workspace
17    setwd("~/Documents/replication_files")
18
19 - ######## PACKAGES ########
20
21   # Check system and installs packages user doesn't have, load needed packages
22
23    need <- c("dplyr", "foreign", "ggplot2", "stargazer") # list packages needed
24    have <- need %in% rownames(installed.packages()) # checks packages you have
25    if(any(!have)) install.packages(need[!have]) # install missing packages
26    invisible(lapply(need, library, character.only=T)) # load needed packages
27
28 - ######## ANALYSIS ########
29
```

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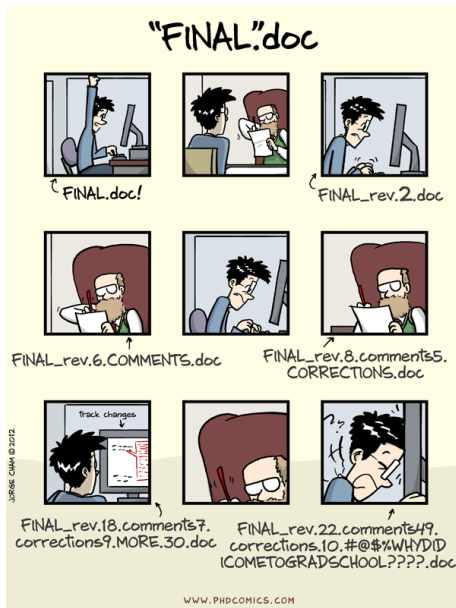
- **What:** A system for managing iterative versions of files (code, data, manuscripts) over time and across collaborators

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

Principles of version control

Principles of version control:

- Vault original, raw data files (do not save over raw, original data)

- Changes to files should be documented and reversible

- Keep "main" versions of files in working order; create copies before experimenting

- Reconcile independent changes by different users

While you can do this manually, in week 1 you learned about Git (and GitHub), so do that instead!

Overview
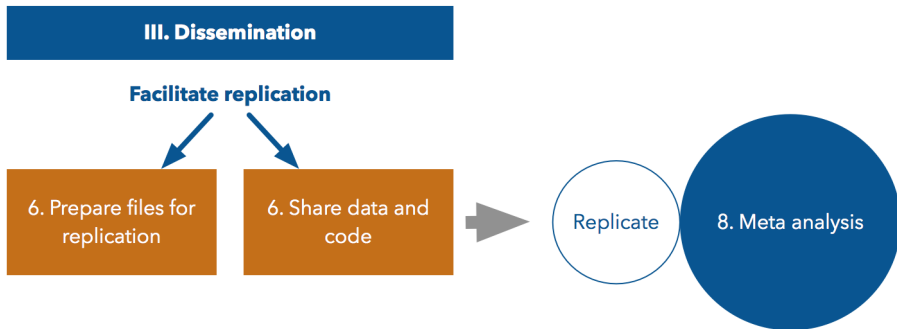
How to prepare files to facilitate replication/reproducibility

"Replication" files should ...

- Be complete but parsimonious
- Run and reproduce results, ideally from the raw data, w/ 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.

Five 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. If you have been disciplined about the earlier parts of your process, this should not be hard.

Purpose:

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

1. Set-up

Create the following:

1. **A new, empty replication folder** *within* your project directory (e.g., "`replication_files/`")

2. **Subfolders:**
   - `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

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.*

Data identifiability is not binary. Data lies on a spectrum with multiple shades of identifiability. See this useful infographic:
https://fpf.org/wp-content/uploads/2016/04/FPF_
Visual-Guide-to-Practical-Data-DeID.pdf

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

- Birthdate (month, day and year of birth), gender, and 5-digit postal code (ZIP) uniquely identifies most people (87%) in the United States. (Sweeney 2000)

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.
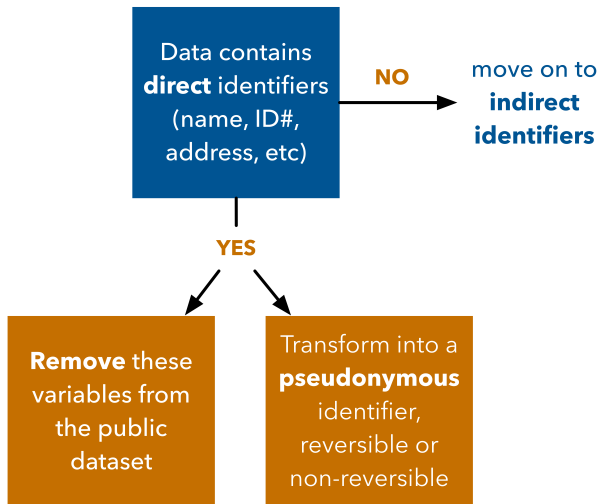
Dealing with direct Iidentifiers

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

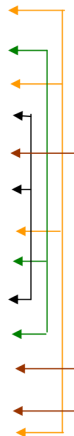

Data contains **direct** identifiers (name, ID#, address, etc)

NO → move on to **indirect identifiers**

YES

Remove these variables from the public dataset

Transform into a **pseudonymous** identifier, reversible or non-reversible

What is sufficient de-identification for indirect identifiers?
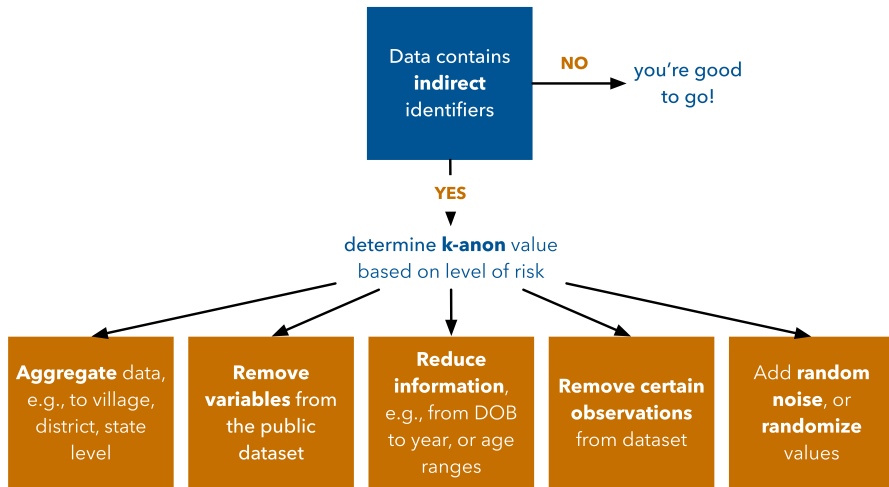
1. **Determine Risk:** Pr(being identified) $\times$ sensitivity of data
2. **Set "k-anonymous" level:** each record cannot be distinguished from at least $k-1$ other individuals who also appear in the data set
3. **Select appropriate method(s) of de-identification:** aggregating data, removing certain 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

Solutions for indirect identifiers

Data contains **indirect** identifiers

NO → you're good to go!

YES

determine **k-anon** value based on level of risk

Aggregate data, e.g., to village, district, state level

Remove variables from the public dataset

Reduce information, e.g., from DOB to year, or age ranges

Remove certain observations from dataset

Add **random noise**, or **randomize** values

- **Aggregating**: lose ability to replicate any individual-level analysis

- **Removing variables**: may not be able to replicate specific models

- **Remove observations**: adds bias if non-random

- **Reducing information in variables**: adds noise to models

- **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 all code even if it manipulates/analyzes identified data, as long as it doesn't compromise anonymity, e.g., censor code that sets the seed for a random draw to generate pseudonymous ID numbers

- If identifiers are not used for analysis, de-identify early in merging/cleaning process

- Store original data with personally identifiable information securely (at LSE you're required to complete a Data Management Plan detailing this)

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 R/python/etc. memory

- Rerun the entire process – merging, cleaning and analysis – to make sure your edits didn't break anything

- Testing on a friend's computer can also be a final check

- Once any discrepancies are addressed, the files are ready for sharing

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 Dataverse

- Open Science Framework

- OpenICPSR

- And, of course, on your own GitHub page!

- **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"

    - One study = one data point

- **Why:** Because any estimate in an individual study may be biased or contain random error

Same Data, Different Conclusions

Twenty-nine research teams were given the same set of soccer data and asked to determine if referees are more likely to give red cards to dark-skinned players. Each team used a different statistical method, and each found a different relationship between skin color and red cards.
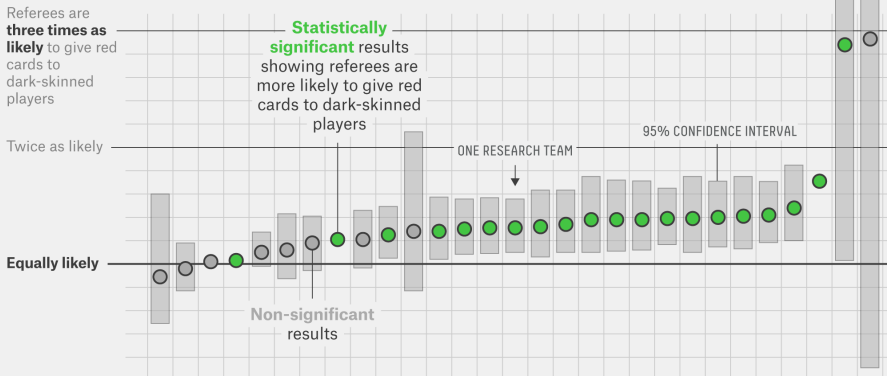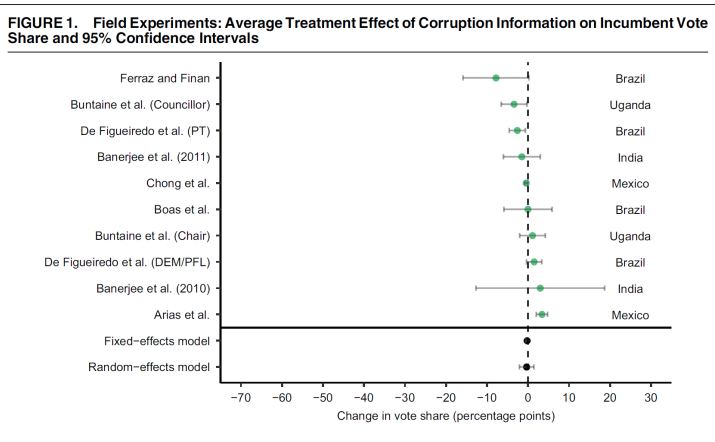
Source: Graph = fivethirtyeight.com, see `https://osf.io/j5v8f/` for study materials

Meta-analysis: effect of corruption information

FIGURE 1. Field Experiments: Average Treatment Effect of Corruption Information on Incumbent Vote Share and 95% Confidence Intervals

Incerti, T. (2020). Corruption information and vote share: A meta-analysis and lessons for experimental design. *American Political Science Review*, 114(3), 761-774.

Overview

Solutions at the Institutional/Discipline Level

- **Design-based publication:** "registered reports" moves peer review before data analysis
- **Incentives for transparency, replication, meta-analysis:** See BITSS prizes and awards, OSF pre-registration challenge, I4R, etc.
- **Change norms:** e.g., journal/disciplinary standards for data sharing
- **Training:** BITSS, Center for Open Science, etc.
- **Tenure:** "Adherence to the replication standard should be part of [tenure] judgment" (King 1995)

Selected Reading & Citations

- **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), Gelman & 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)