MAKING SOCIAL SCIENCE MORE REPRODUCIBLE

(Starting with Data Curation and Code Review)

Limor Peer
Associate Director for Research, Institution for Social and Policy Studies
Data and Research Specialist, Office of the Provost
Yale University

GESIS Lecture Series

March 13, 2018

INSTITUTION FOR SOCIAL AND POLICY STUDIES

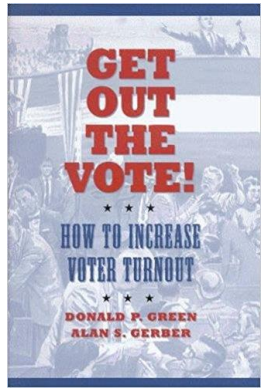
ISPS was founded in 1968 as an interdisciplinary center to support social science and public policy research at Yale University

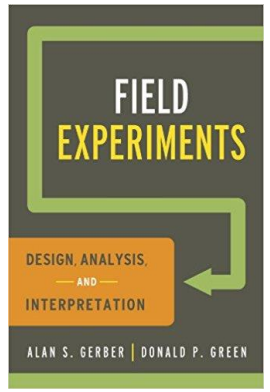
Field Experiments Initiative

Biodiversity Field Experiment (University of Minnesota's Cedar Creek Ecosystem Science Reserve, with permission)

In 2001, the Institution for Social and Policy Studies announced an initiative designed to encourage field experimentation in the social sciences at Yale.

The term 'field experiment' refers to fully-randomized research designs in which observations found in a naturalistic setting – voters, patients, welfare recipients, community organizations, government entities, and the like – are assigned to treatment and control conditions. Recent examples of this kind of research at ISPS include randomized studies of voter mobilization, peer counseling of homeless people, campaign activities in Africa, and the persuasiveness of televised campaign advertisements.





ISPS DATA ARCHIVE

Since 2011

Specialized community

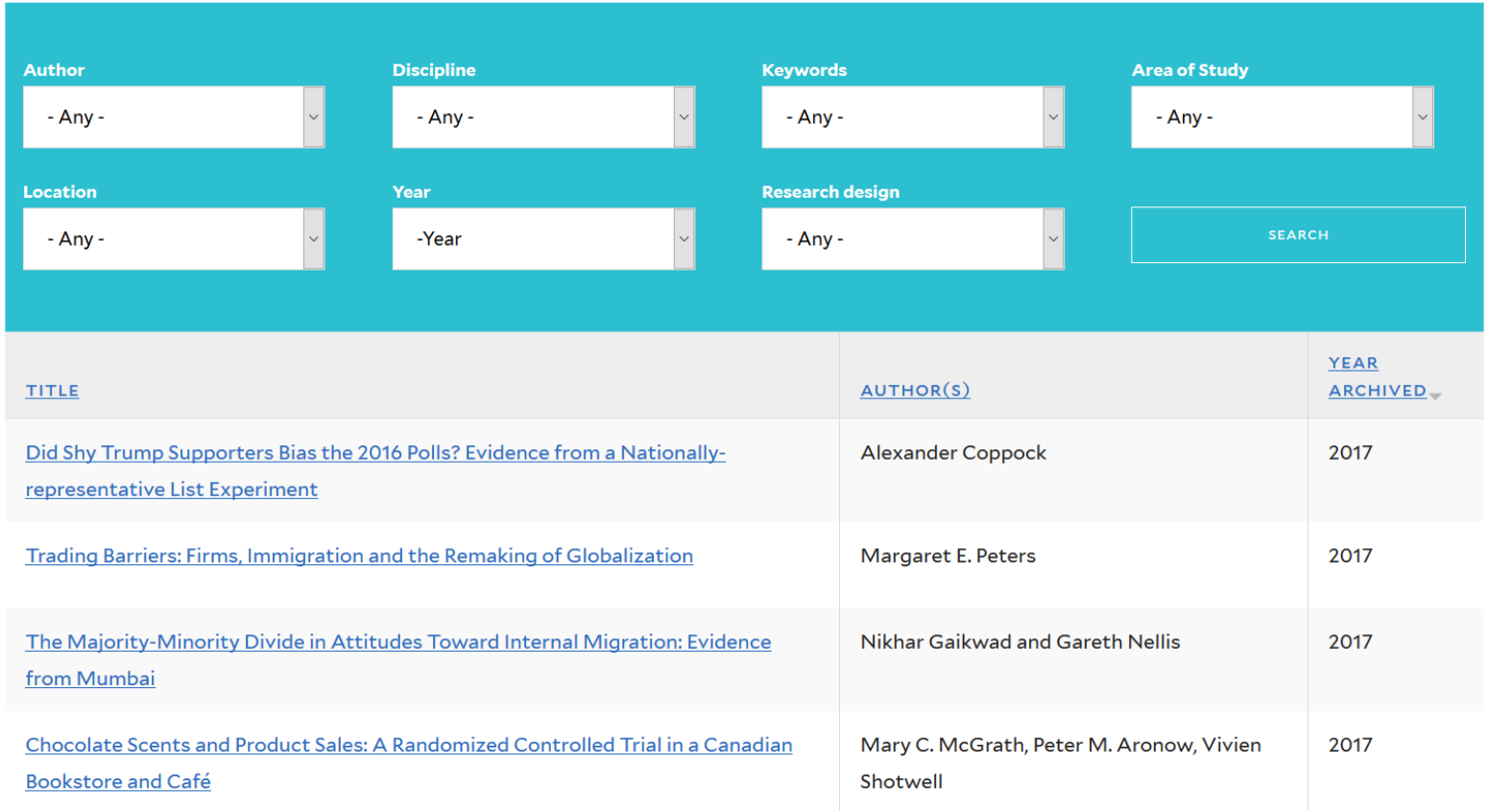
Open access

Website integration

90 studies

1,400 files

15 GB



https://isps.yale.edu/research/data

OUTLINE

Making social science more reproducible: Why, What?

Defining terms

Curating for reproducibility: What? How?

Curation Tool: Yale Application for Research Data (YARD)

OUTLINE

Making social science more reproducible: Why, What?

Defining terms

Curating for reproducibility: What? How?

Curation Tool: Yale Application for Research Data (YARD)

E.G., REINHART & ROGOFF

Free exchange

The 90% question

A seminal analysis of the relationship between debt and growth comes under attack

Print edition | Finance and economics ›

Apr 20th 2013

GOVERNMENT indebtedness matters. Default and financial panic are the stuff of finance-minister nightmares. Government borrowing can crowd out private investment, dragging growth down. Yet economists have struggled to specify when a country needs to worry about its debt load. In a 2010 paper Carmen Reinhart, now a professor at Harvard Kennedy School, and Kenneth Rogoff, an economist at Harvard University, seemed to provide an answer. They argued that GDP growth slows to a snail's pace once government-debt levels exceed 90% of GDP.

https://www.economist.com/news/finance-and-economics/21576362-seminal-analysis-relationship-between-debt-and-growth-comes-under

E.G., REINHART & ROGOFF

In their Excel spreadsheet, Reinhart and Rogoff had not selected the entire row when averaging growth figures: they omitted data from Australia, Austria, Belgium, Canada and Denmark...

University of Massachusetts Amherst researchers uncovered the mistake after they obtained the actual spreadsheet that Reinhart and Rogoff used for their calculations.

[The authors] have acknowledged making a spreadsheet calculation mistake in a 2010 research paper... But the authors stand by their conclusion that higher government debt is associated with slower economic growth.

https://www.bloomberg.com/news/articles/2013-04-18/faq-reinhart-rogoff-and-the-excel-error-that-changed-history

E.G., PIKETTY

The data underpinning Professor Piketty's 577-page tome, which has dominated best-seller lists in recent weeks, contain a series of errors that skew his findings, **according to a Financial Times investigation**.

Piketty findings undercut by errors

Economist and author Thomas Piketty

Chris Giles in London MAY 23, 2014

257

Thomas Piketty's book, 'Capital in the Twenty-First Century', has been the publishing sensation of the year. Its thesis of rising inequality tapped into the zeitgeist and electrified the post-financial crisis public policy debate.

https://www.ft.com/content/e1f343ca-e281-11e3-89fd-00144feabdc0

E.G., PIKETTY - CONTINUED

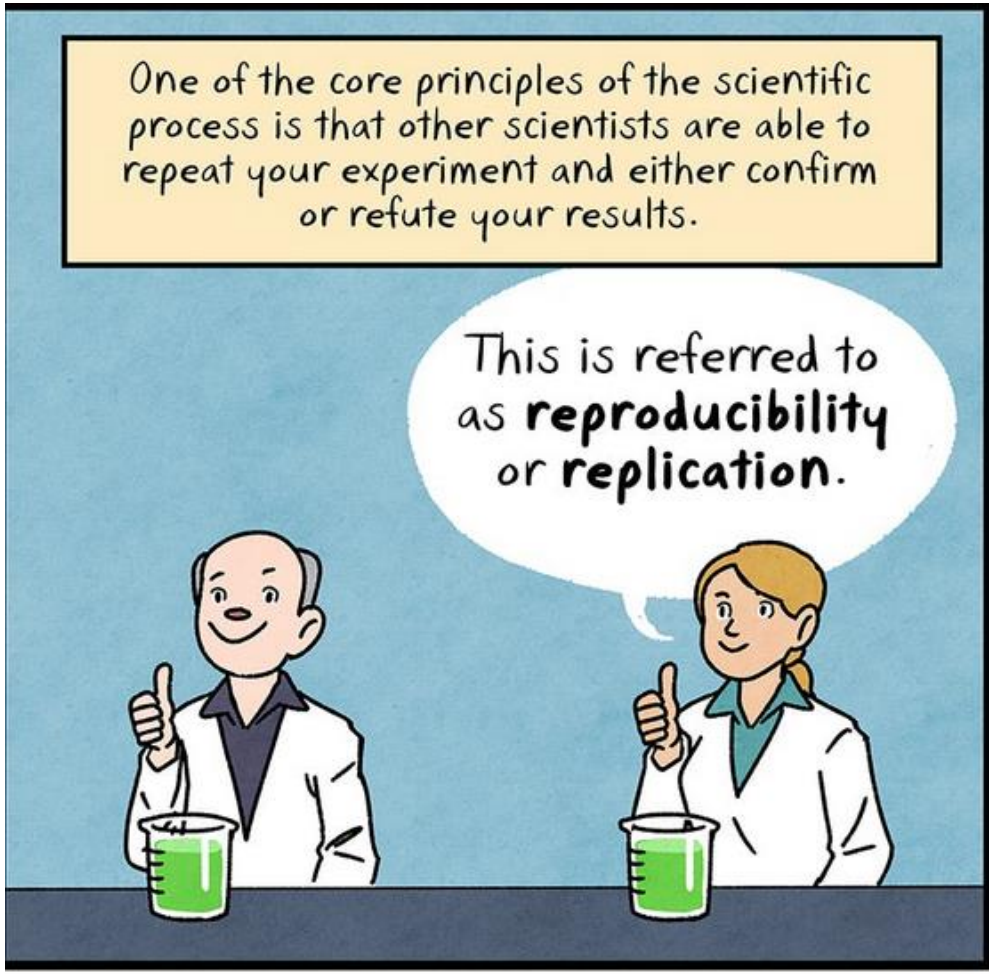
"On the plus side, **he published his code** ... on the negative side, it appears that Piketty's code contains mistakes, fudging and other problems. ... Simply put, spreadsheets are good for quick and dirty work, but they are not designed for serious and reliable work. ... **Spreadsheets make code review difficult**. The code is hidden away in dozens if not hundreds of little cells. **If you are not reviewing your code carefully and if you make it difficult for others to review it, how do expect it to be reliable**?" -- Daniel Lemire

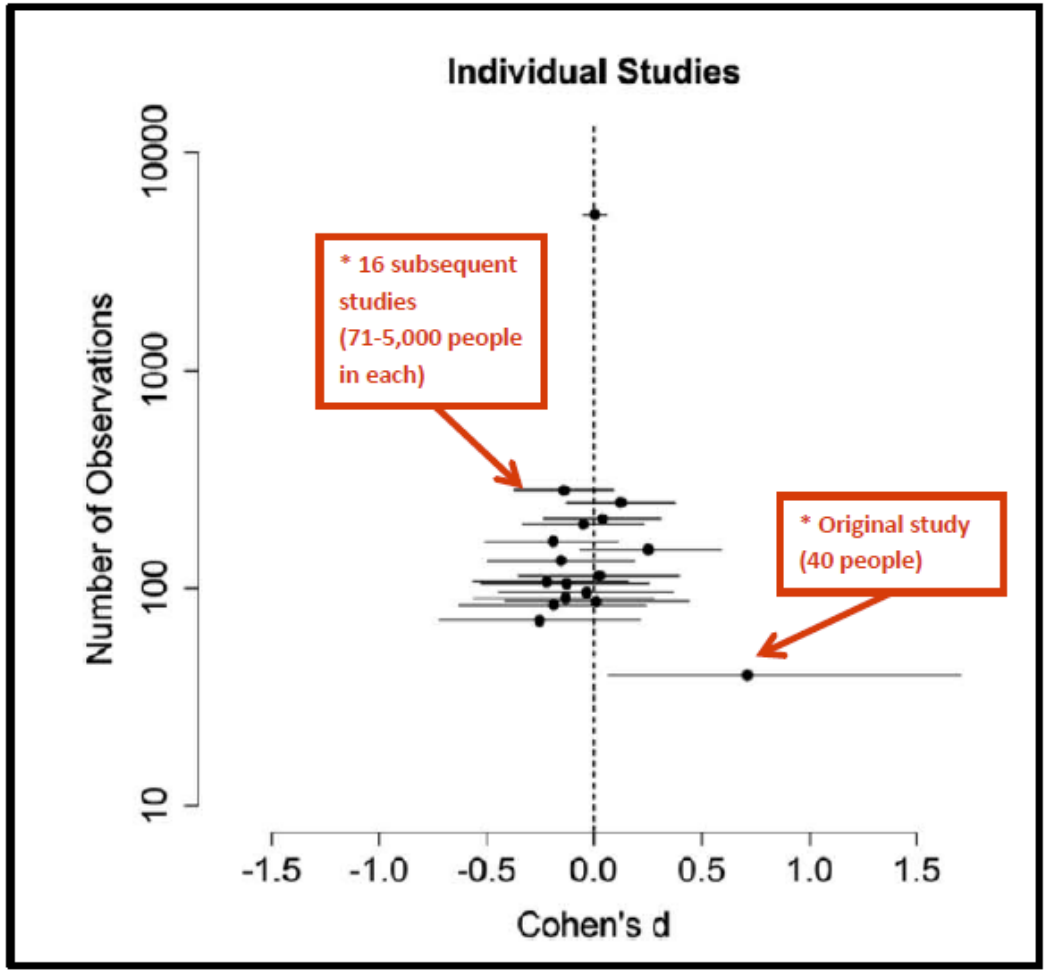
http://lemire.me/blog/archives/2014/05/23/you-shouldnt-use-a-spreadsheet-for-important-work-i-mean-it/

"it's a red herring in the Piketty discussion, except insofar as both [Piketty and Reinhart & Rogoff] are examples that help flesh out **standards and guidelines for data/code release** in economics." -- Victoria Stodden

http://blog.stodden.net/2014/05/27/mistakes-piketty/

SCIENCE ADVANCES
WHEN SCIENTIFIC CLAIMS ARE SUBJECTED TO SCRUTINY





Repeat After Me, Maki Naro https://thenib.com/repeat-after-me

http://andrewgelman.com/2016/06/26/29449/

OUTLINE

Making social science more reproducible: Why, What?

Defining terms

Curating for reproducibility: What? How?

Curation Tool: Yale Application for Research Data (YARD)

REPRODUCIBILITY
DEFINITION

computational reproducibility

empirical reproducibility

VALIDATION

REPEATABILITY

REPLICABILITY

methodological reproducibility

VERIFICATION

REPRODUCIBILITY

statistical reproducibility

REPRODUCIBILITY
DEFINITION

Reproducibility: calculation of quantitative scientific results by independent scientists using the original datasets and methods

Stodden, V. (Ed.), Leisch, F. (Ed.), Peng, R.D. (Ed.). (2014). *Implementing Reproducible Research.* New York: Chapman and Hall/CRC.

Note:

1. cf. **Replication:** implementing experiments and collecting new data for analysis by other researchers to validate the science
2. Different disciplines may interpret these terms differently

SHARING DATA IS ESSENTIAL

- ★ To reproduce or to verify research
- ★ To make the results of publicly funded research available to the public
- ★ To enable others to ask new questions of extant data
- ★ To advance the state of research and innovation

Borgman, C. L. (2012). The conundrum of sharing research data. *Journal of the American Society for Information Science and Technology, 63*(6), 1059-1078. http://doi.org/10.1002/asi.22634

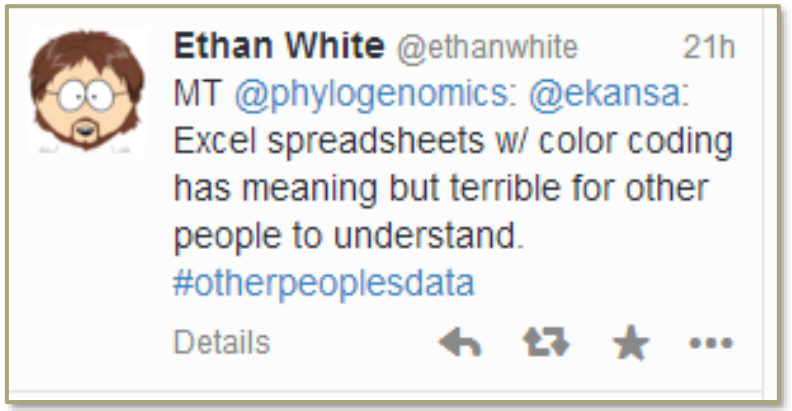
SHARING DATA IS REQUIRED

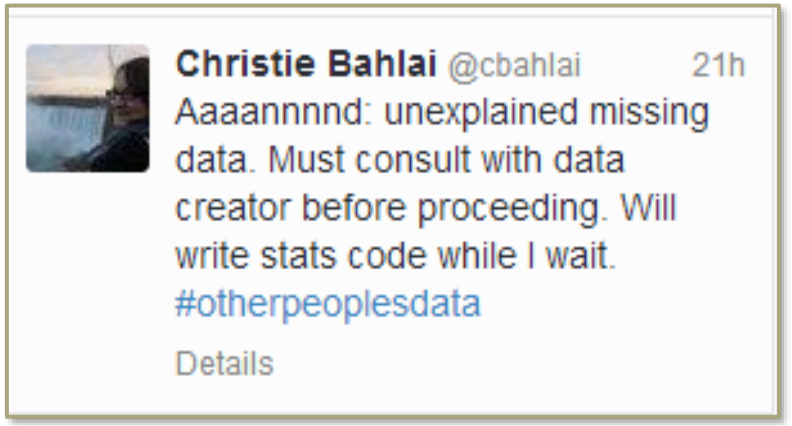
BUT…

SHARING DATA IS NOT SUFFICIENT

The most commonly reported problems associated with [replication] attempts were the lack of… data and code, **followed by insufficient documentation.**

Janz, N., Werfel, S., Wykstra S. (2014). Replication in political science graduate courses: an untapped resource? *Monkey Cage.* https://www.washingtonpost.com/news/monkey-cage





SHARING DATA WITH CARE

Because there are more ways to share data, and because the scholarly landscape supports and encourages that, there is a proliferation of data files on many different types of systems that **do not meet the criterion of quality**…

Peer, L., Green, A., & Stephenson, E. (2014). Committing to data quality review. *International Journal of Digital Curation, 9*(1). http://doi.org/10.2218/ijdc.v9i1.317

REPRODUCIBILITY STANDARD (1)

Reproducibility refers to the ability of a researcher to duplicate the results of a prior study using the same materials as were used by the original investigator. That is, **a second researcher might use the same raw data to build the same analysis files and implement the same statistical analysis in an attempt to yield the same results**. Reproducibility is a minimum necessary condition for a finding to be believable and informative.

K. Bollen, J. T. Cacioppo, R. Kaplan, J. Krosnick, J. L. Olds (2015). *Social, Behavioral, and Economic Sciences Perspectives on Robust and Reliable Science,* National Science Foundation.

REPRODUCIBILITY STANDARD (2)

If a paper makes, or implies, claims that require software, those claims should be backed up.... We are after repeatability, which is simply **the act of checking the claims made in the paper**... Repeatability should become a standard feature of the dissemination of research results.

Krishnamurthi, S., Vitek, J. (2015). *Communications of the ACM*, Vol. 58 No. 3, Pages 34-36, https://doi.org/10.1145/2658987

REPRODUCIBILITY STANDARD (3)

The standard of reproducibility calls for the **data and the computer code** used to analyze the data be made available to others. This standard falls short of full replication because the same data are analyzed again, rather than analyzing independently collected data… However, under this standard, limited exploration of the data and the analysis code is possible and **may be sufficient to verify the quality of the scientific claims**.

Peng, Roger D. (2011). Reproducible research in computational science. *Science*. Vol. 334, Issue 6060, pp. 1226-1227 https://doi.org/10.1126/science.1213847

REPRODUCIBILITY STANDARD (4)

The *replication standard* holds that sufficient information exists with which to understand, evaluate, and build upon a prior work **if a third party could replicate the results without any additional information from the author**.

King, G. (1995). Replication, replication. *PS: Political Science & Politics*, *28*(3), 444–452. http://doi.org/10.2307/420301

REPRODUCIBILITY STANDARD (5)

Could the published computational findings be **reproduced on an independent system by using the data and code** provided?

Stodden, V., McNutt, M., Bailey, D. H., Deelman, E., Gil, Y., Hanson, B., . . . Taufer, M. (2016). Enhancing reproducibility for computational methods. *Science*, *354*(6317), 1240–1241. https://doi.org/10.1126/science.aah6168

REPRODUCIBILITY STANDARD DEMANDS DATA QUALITY

A set of measures that determine if data are **independently understandable for informed reuse**.

Peer, L., Green, A., & Stephenson, E. (2014). Committing to data quality review. *International Journal of Digital Curation, 9*(1). http://doi.org/10.2218/ijdc.v9i1.317

OUTLINE

Making social science more reproducible: Why, What?

Defining terms

Curating for reproducibility: What? How?

Curation Tool: Yale Application for Research Data (YARD)

CURATION
DEFINITION

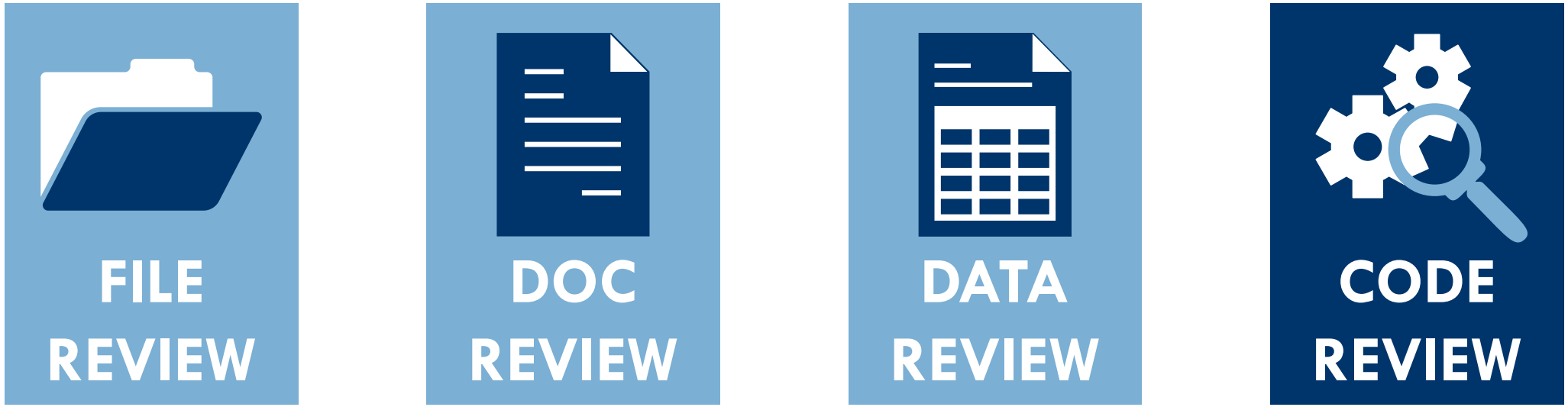
The active and ongoing management of data through its life cycle of interest and usefulness to scholarship, science, and education. Data curation activities enable data discovery and retrieval, maintain its quality, add value, and provide for reuse over time, and this new field includes authentication, archiving, management, preservation, retrieval, and representation.

The University of Illinois' Graduate School of Library and Information Science, https://ischool.illinois.edu/articles/2012/04/palmer-discusses-big-data-u-i-news-bureau

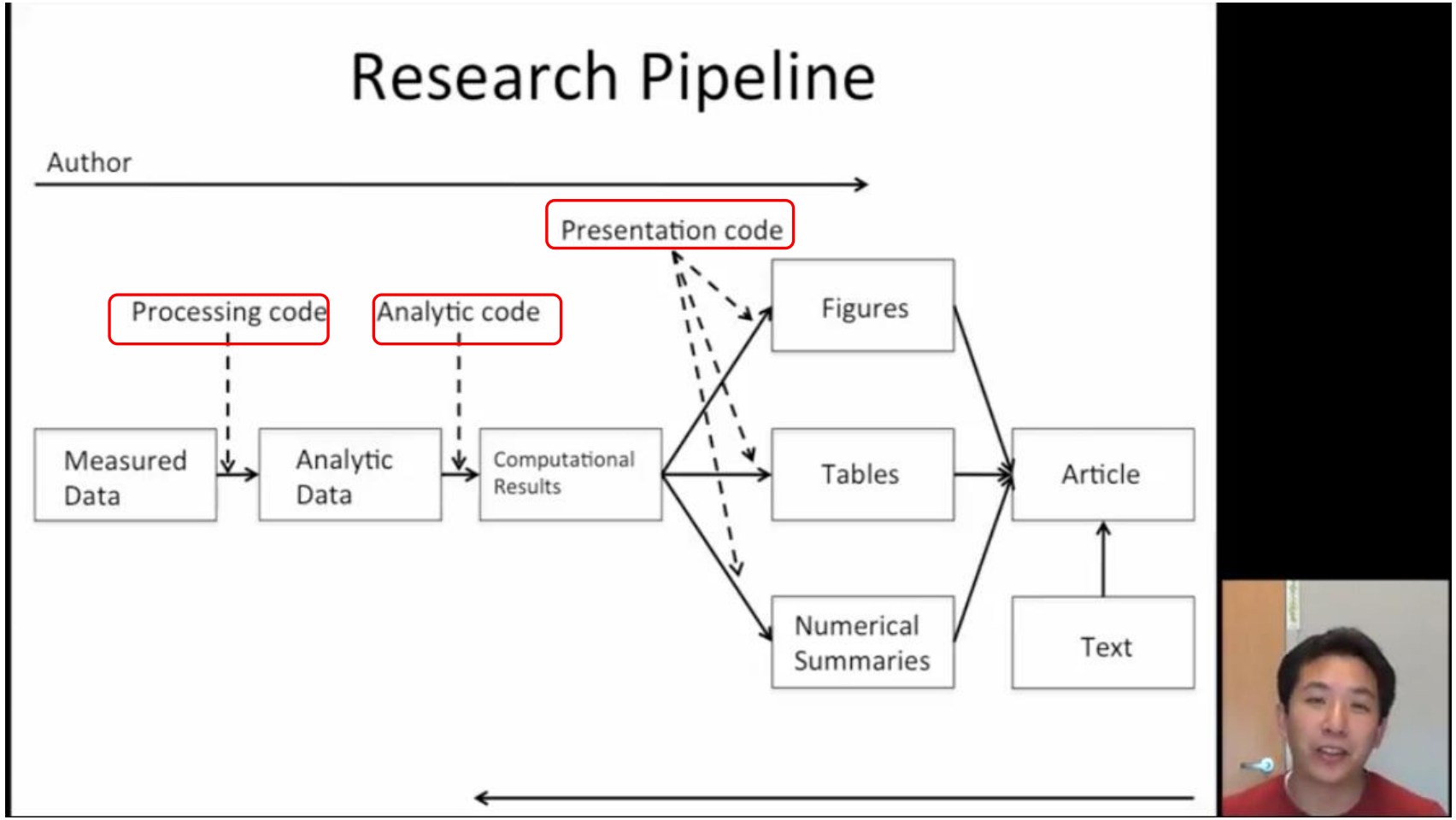
CURATING FOR REPRODUCIBILITY
DATA QUALITY REVIEW

CURATING FOR REPRODUCIBILITY

DATA QUALITY REVIEW



WHAT DO WE MEAN BY CODE?



Roger Peng
(Johns Hopkins University)

http://politicalsciencereplication.wordpress.com/2014/05/08/learning-about-reproducible-research-on-coursera-recap-week-1/

SOFTWARE & CODE

"Code and software are now an integral part of research data generation, usage and analysis"

Nature editorial, 2015

http://www.nature.com/articles/sdata20154

"Deep intellectual contributions now encoded in Software"

Stodden, 2014

https://web.stanford.edu/~vcs/talks/OpenConNov152014-STODDEN.pdf

"Code as first-class research object"

Mozilla Science Lab in collaboration with figshare and github

https://science.mozilla.org/blog/code-as-a-research-object-a-new-project

Science Code Manifesto

Manifesto Discussion Endorse Resources About

Software is a cornerstone of science. Without software, twenty-first century science would be impossible. Without better software, science cannot progress.

But the culture and institutions of science have not yet adjusted to this reality. We need to reform them to address this challenge, by adopting these five principles:

Code	All source code written specifically to process data for a published paper must be available to the reviewers and readers of the paper.
Copyright	The copyright ownership and license of any released source code must be clearly stated.
Citation	Researchers who use or adapt science source code in their research must credit the code's creators in resulting publications.
Credit	Software contributions must be included in systems of scientific assessment, credit, and recognition.
Curation	Source code must remain available, linked to related materials, for the useful lifetime of the publication.

http://sciencecodemanifesto.org/

CODE SHARING REQUIREMENTS

Journals: Stodden et al (2013) find 30% increase in journals requiring code.

Do you want to publish here?

- Nature group
- Science
- PLoS
- AJPS (see APSA DART)
- And the list is growing…

CURATING FOR REPRODUCIBILITY
DATA QUALITY REVIEW

- Assign persistent identifier
- Create study citation and study-level metadata record
- Record file size details
- Check for presence of all files
- Verify content of files matches expected format
- Create non-proprietary versions of files
- Implement migration strategy for file formats

CURATING FOR REPRODUCIBILITY
DATA QUALITY REVIEW

- Confirm presence of comprehensive descriptive information necessary for informed reuse
  - Data definitions
  - Variable construction
  - Methodology
  - Sampling information
  - Original data source citation
  - Analysis software version
- Link to related research products

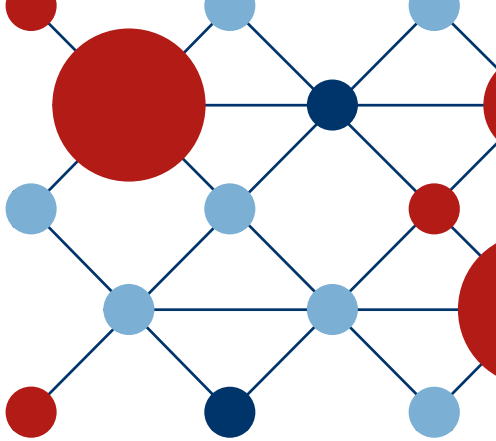
CURATING FOR REPRODUCIBILITY
DATA QUALITY REVIEW

- Check for undocumented variable and value information
- Examine data for inconsistencies and errors
  - Discrepancies in number of observations
  - Out-of-range or wild codes
  - Undefined null values
- Review data for confidentiality issues

CURATING FOR REPRODUCIBILITY
DATA QUALITY REVIEW

- Convert absolute file paths to relative file paths
- Check code for presence of non-executable comments that document analysis processes
- Identify packages required to execute code
- Execute code to ensure code is error-free
- Compare code output to findings presented in article

CURATING FOR REPRODUCIBILITY
THE CURE CONSORTIUM



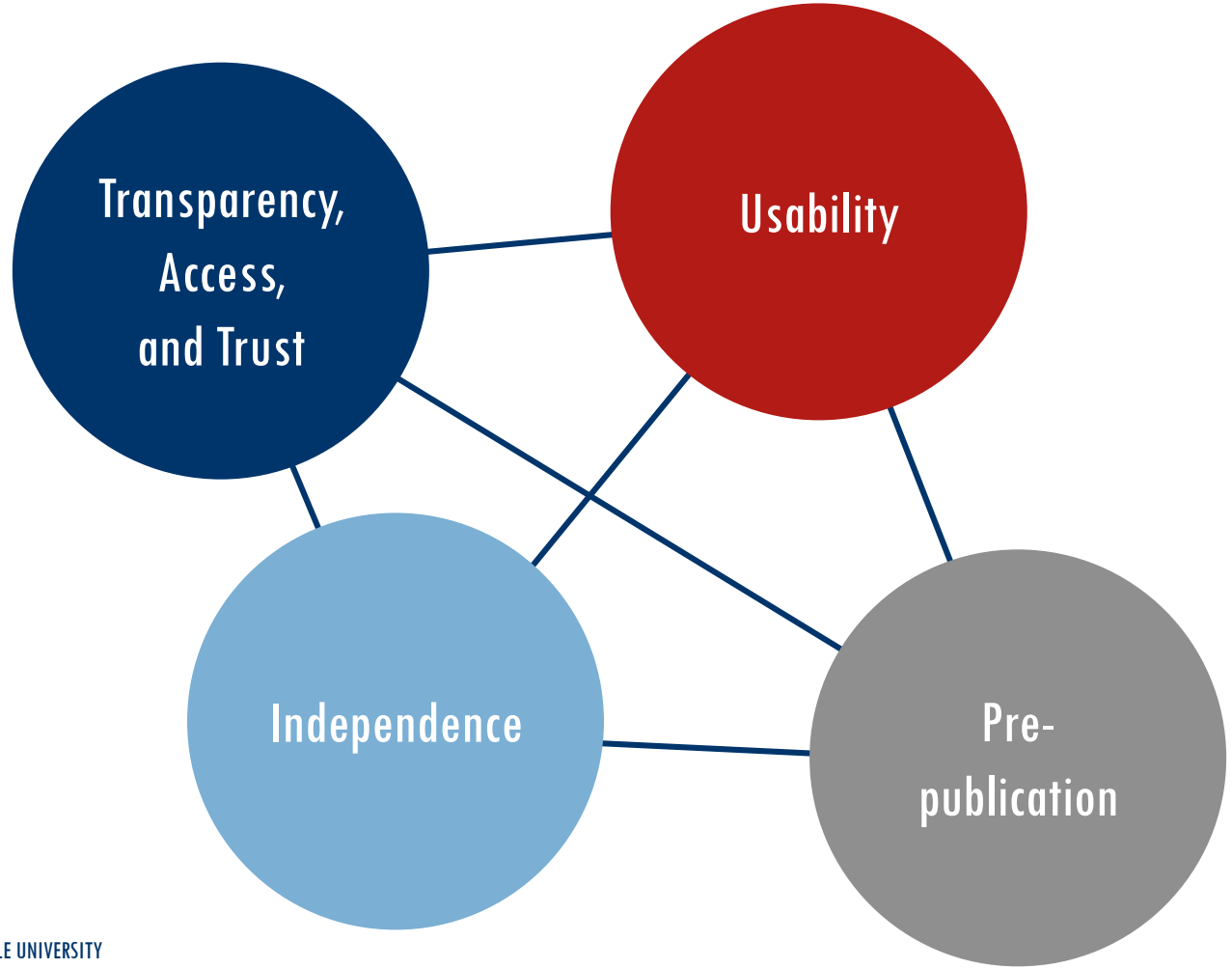
Supporting research data curation and code review for the purpose of facilitating the digital preservation of the evidence base necessary for future understanding, evaluation, and replication of scientific claims.

https://cure.web.unc.edu/

- Establish Standards
- Share Practices
- Promote Data Quality Review

CURATING for REPRODUCIBILITY

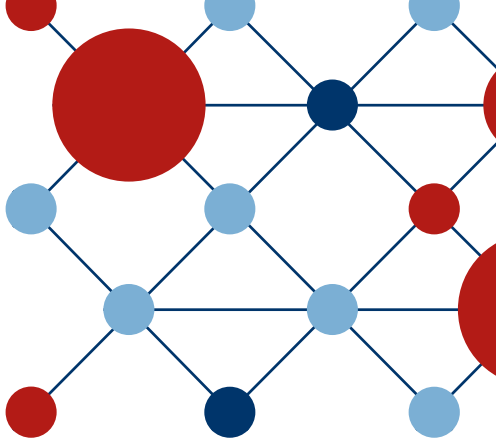
CURATING FOR REPRODUCIBILITY
THE CURE CONSORTIUM



CURATING for REPRODUCIBILITY

https://cure.web.unc.edu/

CURATING FOR REPRODUCIBILITY
MODELS OF PRACTICE



1. **Institution for Social and Policy Studies (ISPS)**
   Aligning Data Curation Workflows with Data Quality Review

2. **Cornell Institute for Social and Economic Research**
   Providing Data Curation and Reproduction of Results (R^2) Services

3. **Odum Institute for Research in Social Science**
   Enforcing Journal Data Replication Policies

OUTLINE

Making social science more reproducible: Why, What?

Defining terms

Curating for reproducibility: How?

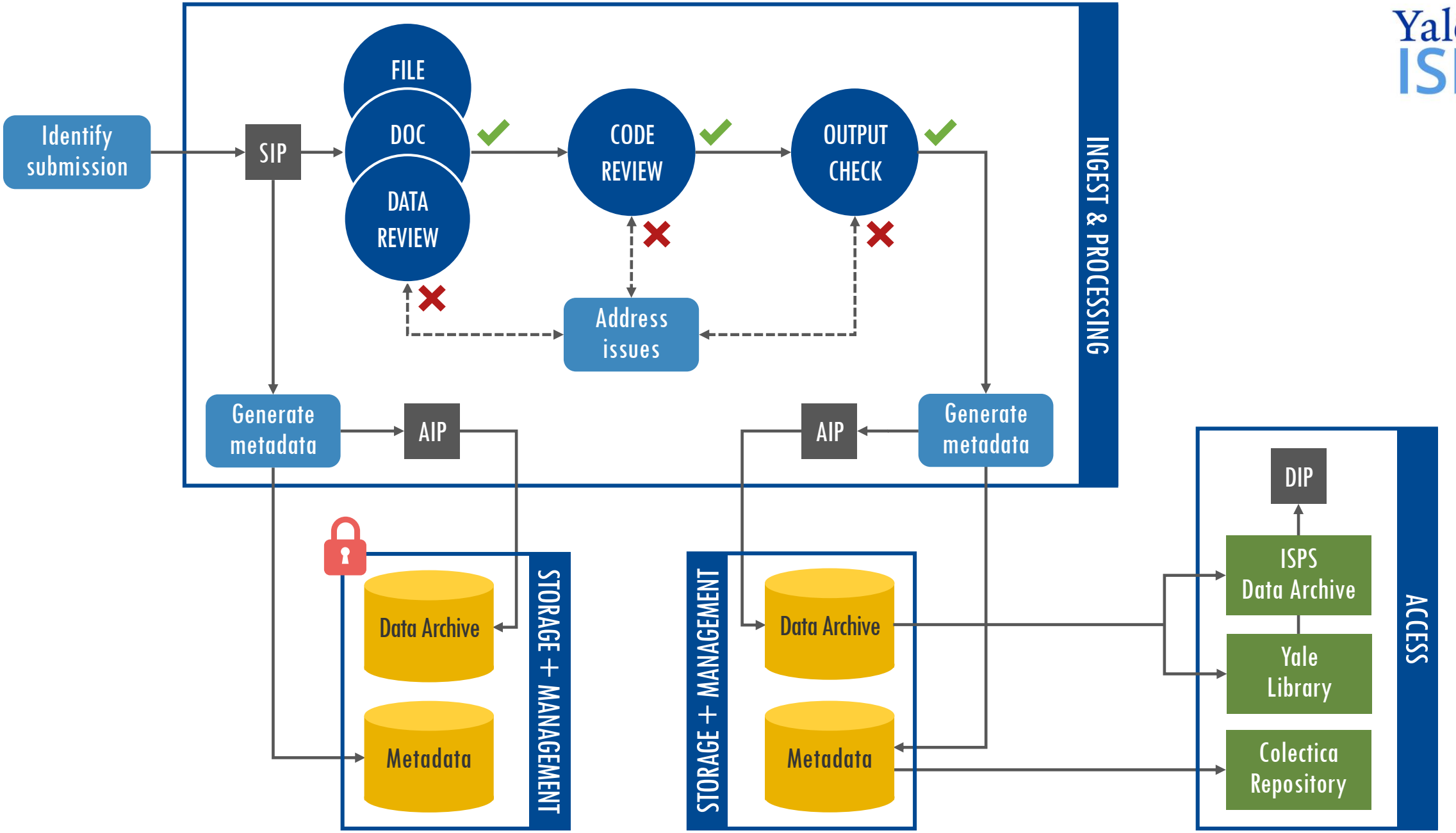
Curation Tool: Yale Application for Research Data (YARD)

CURATING FOR REPRODUCIBILITY

Institution for Social and Policy Studies (ISPS)

Aligning Data Curation Workflows with Data Quality Review

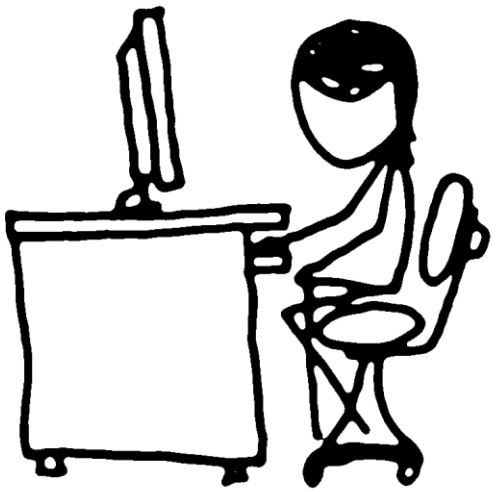
ISPS data archivists developed a data curation workflow that implements the ideals of scientific reproducibility and transparency


Identify submission
SIP
FILE
DOC
DATA REVIEW
CODE REVIEW
OUTPUT CHECK
Address issues
INGEST & PROCESSING
Generate metadata
AIP
AIP
Generate metadata
STORAGE + MANAGEMENT
Data Archive
Metadata
STORAGE + MANAGEMENT
Data Archive
Metadata
DIP
ISPS Data Archive
Yale Library
Colectica Repository
ACCESS

MISSING LABELS

"We are missing labels for the following variables: _n1, _n0, V1 and V0."

Archive staff



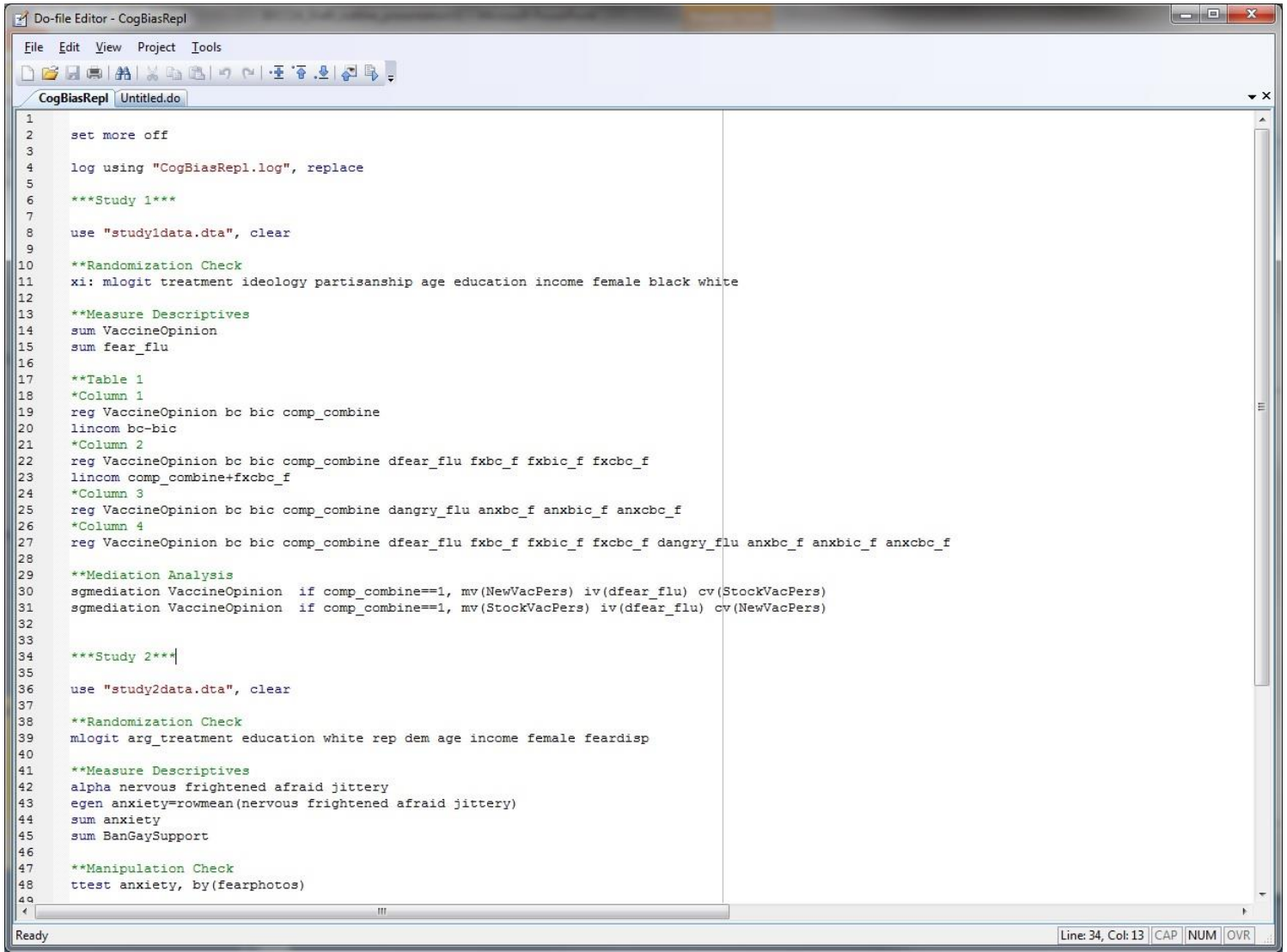
http://xkcd.com/662/ Creative Commons Attribution-Noncommercial

"Here are the labels:
_n1 is the number of observations in the treated strata before matching
_n0 is the number of observations in the comparison strata before matching
v1 = turnout for treated observations
v0 = turnout for comparison observations

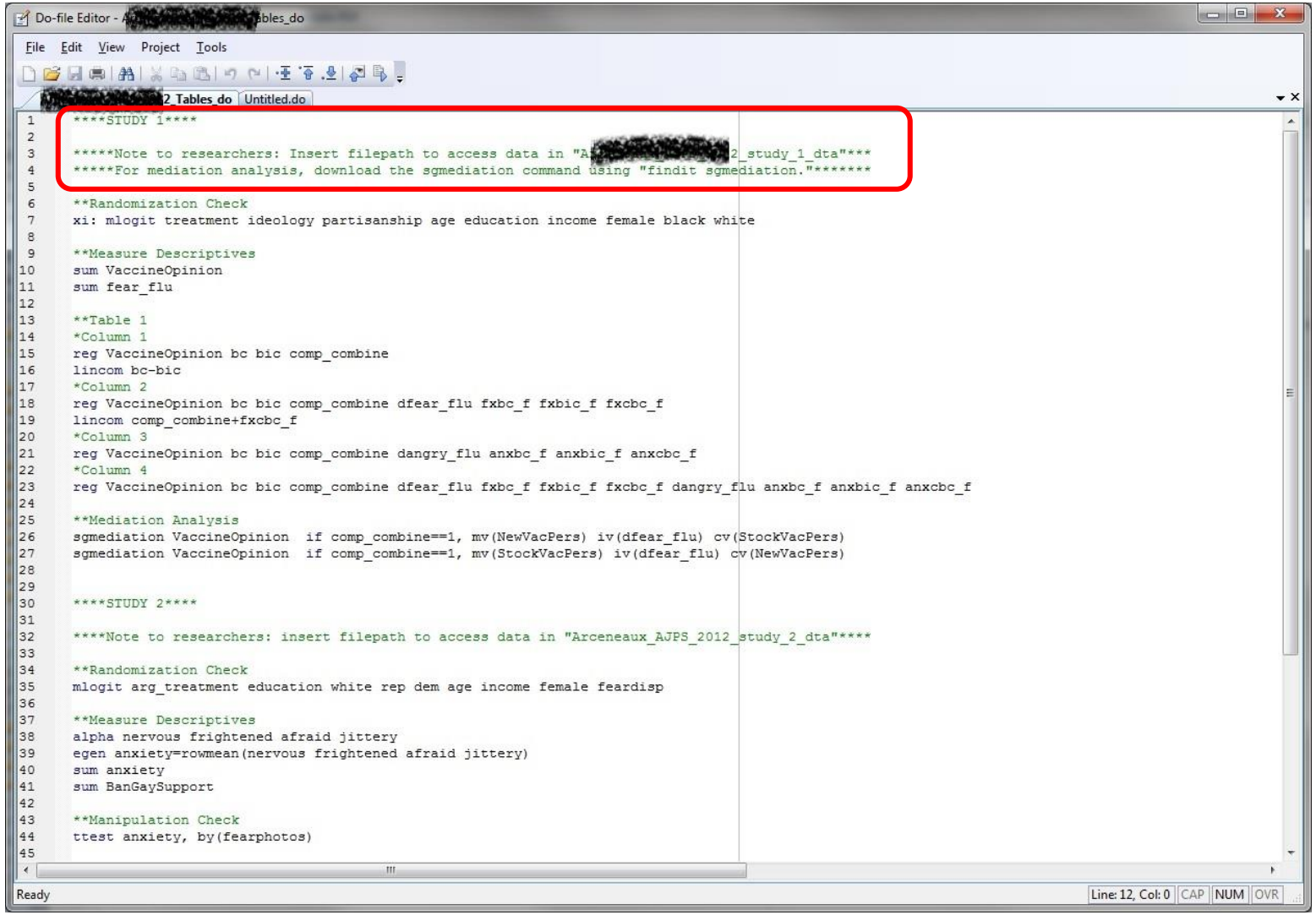
… this reminds me that I needed to include the .ado code in the Matching Code folder. I just did that and updated the readme file. Boy, the things your forget about after not thinking about something for two years!"

Researcher

CODE BEFORE


Do-file Editor - CogBiasRepl
File Edit View Project Tools
CogBiasRepl Untitled.do
set more off
log using "CogBiasRepl.log", replace
Study 1
use "study1data.dta", clear
**Randomization Check
xi: mlogit treatment ideology partisanship age education income female black white
**Measure Descriptives
sum VaccineOpinion
sum fear_flu
**Table 1
*Column 1
reg VaccineOpinion bc bic comp_combine
lincom bc-bic
*Column 2
reg VaccineOpinion bc bic comp_combine dfear_flu fxbc_f fxbic_f fxcbc_f
lincom comp_combine+fxcbc_f
*Column 3
reg VaccineOpinion bc bic comp_combine dangry_flu anxbc_f anxbic_f anxcbc_f
*Column 4
reg VaccineOpinion bc bic comp_combine dfear_flu fxbc_f fxbic_f fxcbc_f dangry_flu anxbc_f anxbic_f anxcbc_f
**Mediation Analysis
sgmediation VaccineOpinion if comp_combine==1, mv(NewVacPers) iv(dfear_flu) cv(StockVacPers)
sgmediation VaccineOpinion if comp_combine==1, mv(StockVacPers) iv(dfear_flu) cv(NewVacPers)
Study 2
use "study2data.dta", clear
**Randomization Check
mlogit arg_treatment education white rep dem age income female feardisp
**Measure Descriptives
alpha nervous frightened afraid jittery
egen anxiety=rowmean(nervous frightened afraid jittery)
sum anxiety
sum BanGaySupport
**Manipulation Check
ttest anxiety, by(fearphotos)
Ready
Line: 34, Col: 13 CAP NUM OVR

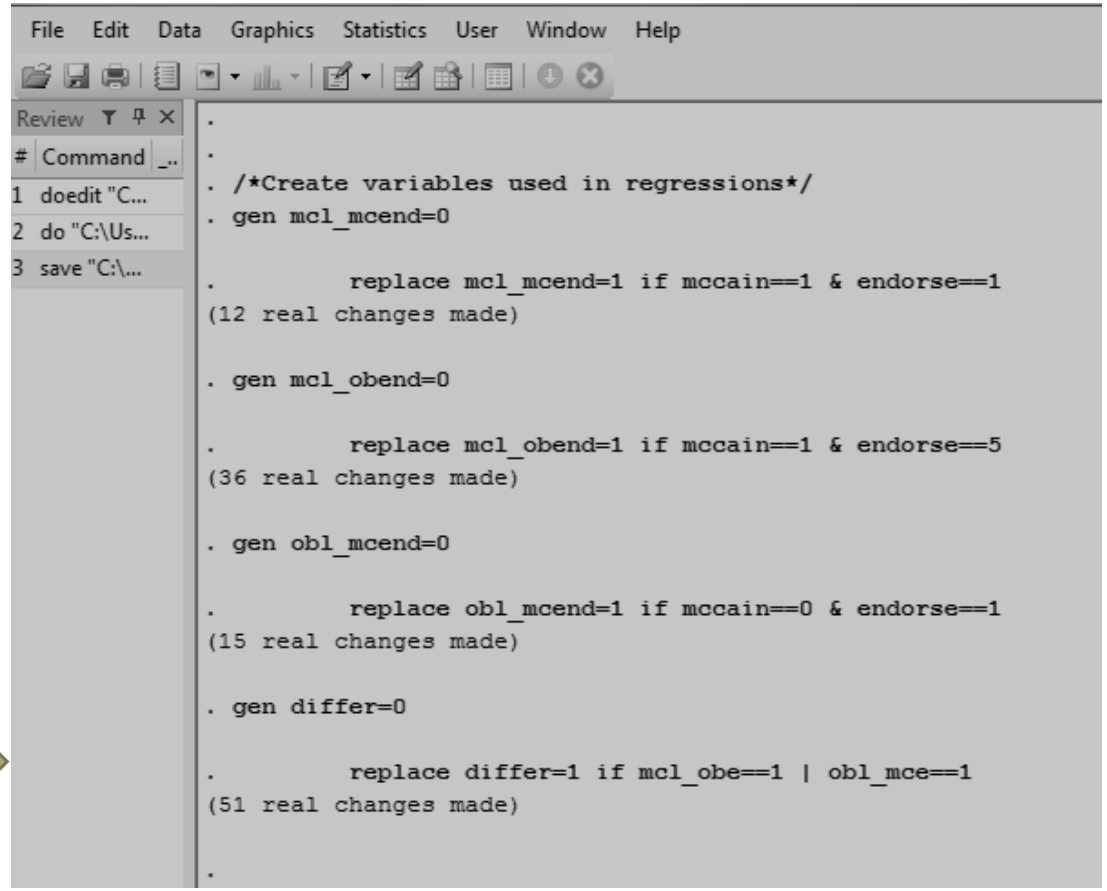
IS THE CODE WELL-ANNOTATED?



DOES THE CODE FULLY EXECUTE?

Example 1:

```
/*Create variables used in regressions*/
gen mcl_mcend=0
    replace mcl_mcend=1 if mccain==1 & endorse==1
gen mcl_obend=0
    replace mcl_obend=1 if mccain==1 & endorse==5
gen obl_mcend=0
    replace obl_mcend=1 if mccain==0 & endorse==1
gen differ=0
    replace differ=1 if mcl_obe==1 | obl_mce==1
```



CHECK THE CODE AGAINST THE PUBLISHED PAPER

Example 2:

```
. /*Table 2*/
.         probit interest_in_letter mccain

Iteration 0:   log likelihood = -57.305692
Iteration 1:   log likelihood =  -56.19505
Iteration 2:   log likelihood = -56.193827
Iteration 3:   log likelihood = -56.193827

Probit regression                                 Number of obs   =        100
                                                  LR chi2(1)      =       2.22
                                                  Prob > chi2     =     0.1359
Log likelihood = -56.193827                       Pseudo R2       =     0.0194

------------------------------------------------------------------------------
interest_in_letter |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------------+----------------------------------------------------------------
           mccain |   .4061989   .2736818     1.48   0.138    -.1302077    .9426054
            _cons |  -.8557124     .20097    -4.26   0.000    -1.249606   -.4618184
------------------------------------------------------------------------------

.               mfx

Marginal effects after probit
      y  = Pr(interest_in_letter) (predict)
         =  .25569496

------------------------------------------------------------------------------
variable |      dy/dx    Std. Err.      z    P>|z|  [    95% C.I.   ]       X
---------+--------------------------------------------------------------------
  mccain*|   .1304522      .08706     1.50   0.134  -.040174  .301079       .49
------------------------------------------------------------------------------
(*) dy/dx is for discrete change of dummy variable from 0 to 1
```

(table)

Table 2. The Effect of Candidate Support on Interest in Letter to the Editor

Dependent Variable = Interested in Publishing Letter Independent Variable	Coefficient (Standard Error) [Change in Probability]	Coefficient (Standard Error) [Change in Probability]
Pro-McCain letter	0.41 (0.27) [13.0%]	0.58** (0.29) [16.3%]
Circulation (in units of 10,000)		−0.028** (0.013) [−16.4%]
Unemployment rate in metro area		−0.009 (0.13) [−0.3%]
Intercept	−0.86** (0.20)	−0.23 (0.75)
N	100	100
Pseudo R^2	.02	.09
Log-likelihood	−56.2	−51.9

Note: The dependent variable is a binary variable that takes the value of 1 if the newspaper either tried to contact the alias for verification purposes or if it published the letter and 0 otherwise. Standard errors are given in parentheses. The estimated predicted probabilities are given in brackets. For the binary variables, the predicted probabilities report the change in the predicted probability when the value of the variable goes from 0 to 1 while holding other variables constant. For the continuous variables, the predicted probabilities report the change in predicted probability when increasing the value of that variable from the mean value to one standard deviation above the mean.
*p < .10. **p < .05.

CURATION TOOL: YARD
YALE APPLICATION FOR RESEARCH DATA

- A Web application that allows Depositors, Curators, and Administrators to submit, review, process, and publish data.

Yale ISPS

ipa INNOVATIONS FOR POVERTY ACTION

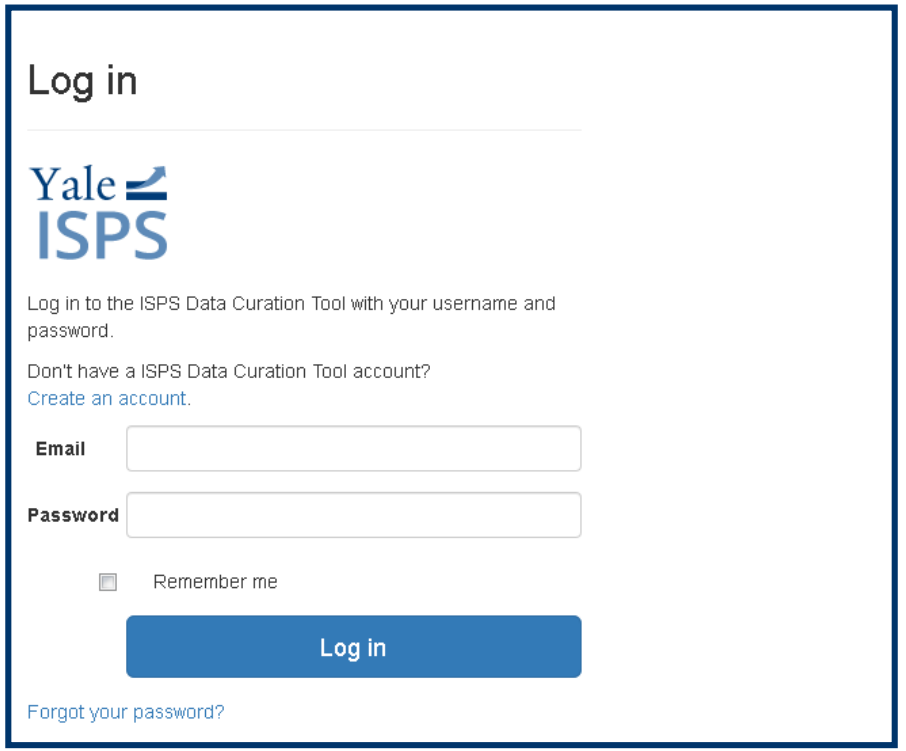
colectica

Production and code release in 2017-2018

CURATION TOOL: REQUIREMENTS

- Curation workflow automation, integration, management, and tracking
- Integrate and capture DDI metadata production with data and code review and cleaning
- Version control for metadata and files
- Preservation metadata and formats
- Secure storage and access
- Preference for open source solutions
- Push out relevant information to pre-determined destinations
  - i.e., a user, the archive administrators, a Web based dissemination system, or preservation systems
- Fit into repository and research workflows

CURATION TOOL: YARD
YALE APPLICATION FOR RESEARCH DATA



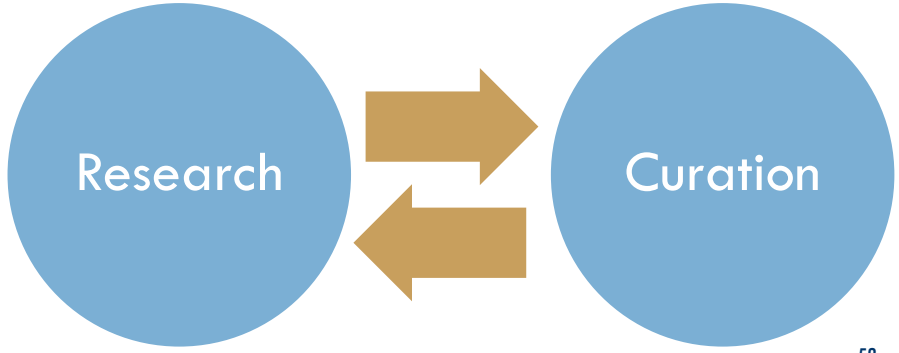
https://docs.colectica.com/curation/technical-documentation/ddi-mapping/

SOME FINAL THOUGHTS...

IF YOU'RE RECEIVING DATA

When do you engage in the research process?

- What type of quality assurance do you offer for the data in your repository?
  - Do you review files? (bits, formats)
  - Do you review data content? (validation, missing info, confidentiality)
  - Do you review documentation, metadata? (comprehensiveness, links)
  - Do you review code / analytic methods? (execute code, replicate results)
- Data repository / archive is the first data re-user

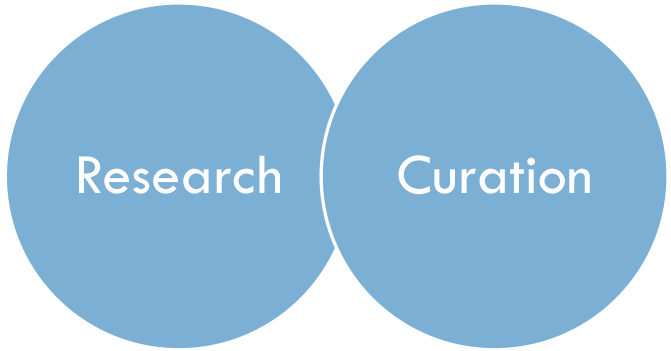


IF YOU'RE RECEIVING DATA

When do you engage in the research process?

- What type of quality assurance do you offer for the data in your repository?
  - Do you review files? (bits, formats)
  - Do you review data content? (validation, missing info, confidentiality)
  - Do you review documentation, metadata? (comprehensiveness, links)
  - Do you review code / analytic methods? (execute code, replicate results)
- Data repository / archive is the first data re-user

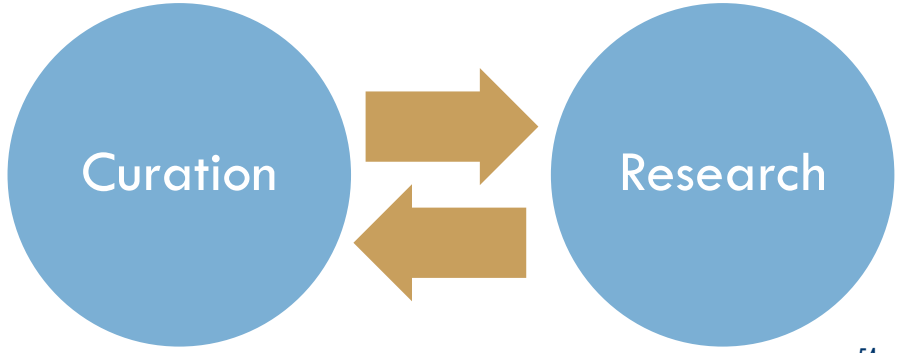
Can curators get involved in the research process earlier?



IF YOU'RE PROVIDING DATA

Do you establish a practice and a culture of transparency?

- Capture your process.
- Check your code, re-check, and check again. Then cross-check.
- Leave little (no?) room for mistakes: Use workflow tools.
- Show (don't tell): Use packaging tools.
- Aim for long-term re-use.

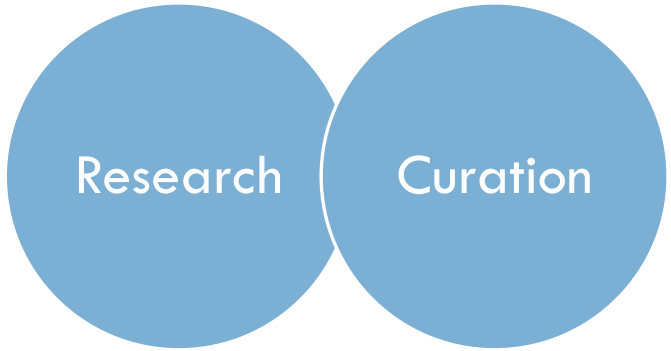


IF YOU'RE PROVIDING DATA

Do you establish a practice and a culture of transparency?

- Capture your process.
- Check your code, re-check, and check again. Then cross-check.
- Leave little (no?) room for mistakes: Use workflow tools.
- Show (don't tell): Use packaging tools.
- Aim for long-term re-use.

Can we embed curation practices in the research workflow?



THANK YOU!

limor.peer@yale.edu

@l_peer

http://isps.yale.edu/