

HONOR *the* FUTURE

THE CAMPAIGN FOR THE UNIVERSITY OF VIRGINIA



UNIVERSITY
of VIRGINIA

Advancement

THINKING REPRODUCIBLY

James Rogol

APRA OverDRIVE/

23 March 2020



Advancement



AGENDA

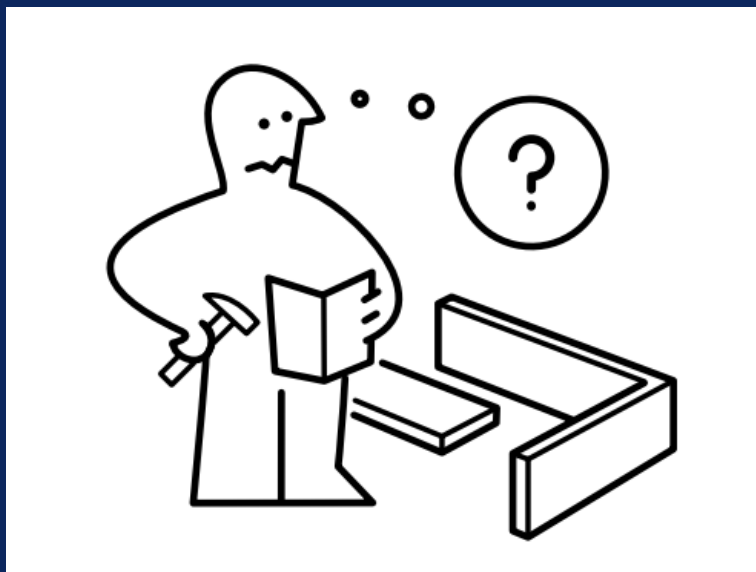
- What is reproducibility?
- Why does this matter?

- How to think reproducibly
 - The Data
 - What to do with it?
 - How to work with it?

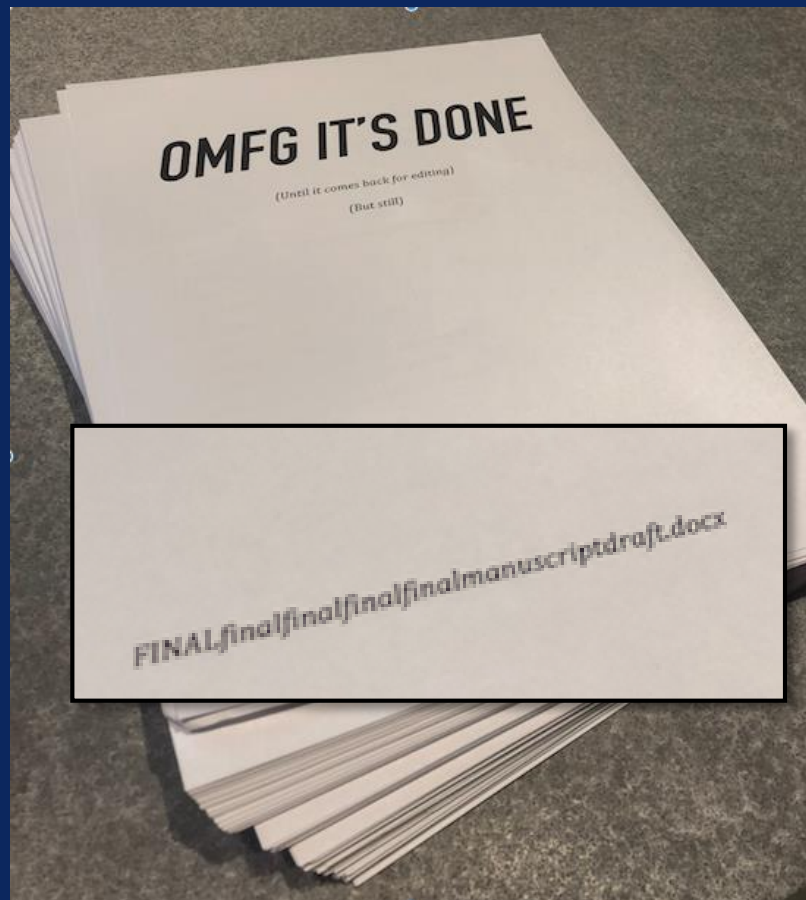
re·pro·duc·i·bil·i·ty, *noun*

- Repeatable
- Identical
- Shareable





XLII

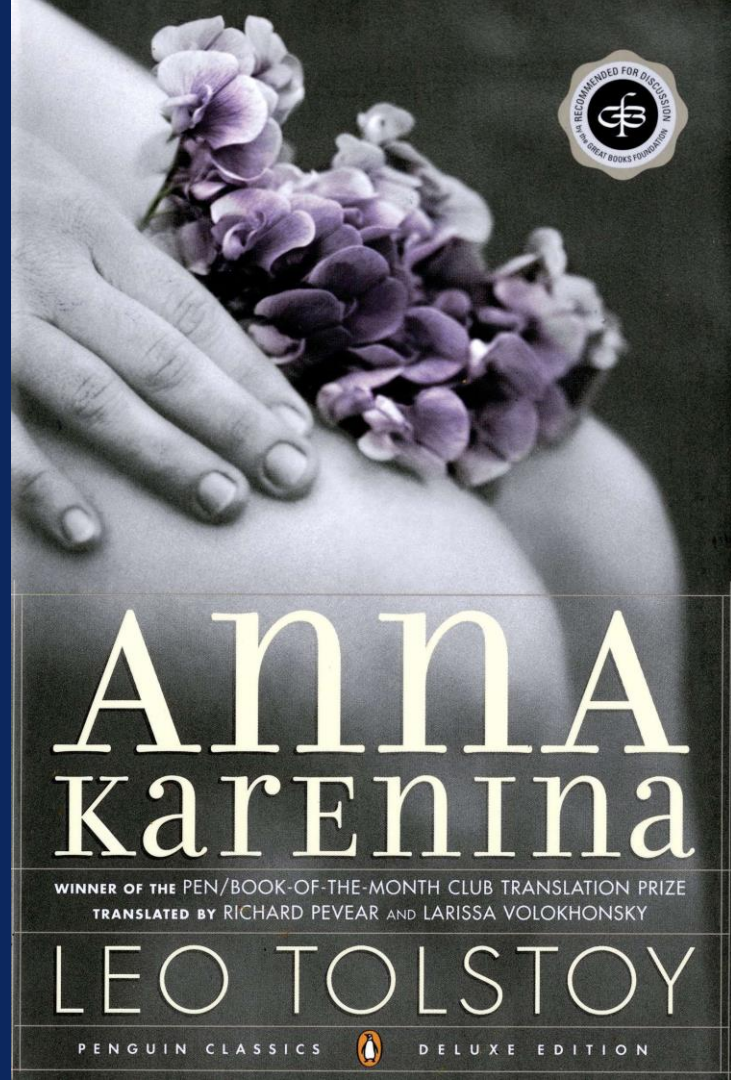


.....

**“All happy families are alike;
each unhappy family is
unhappy in its own way.”**

- Leo Tolstoy

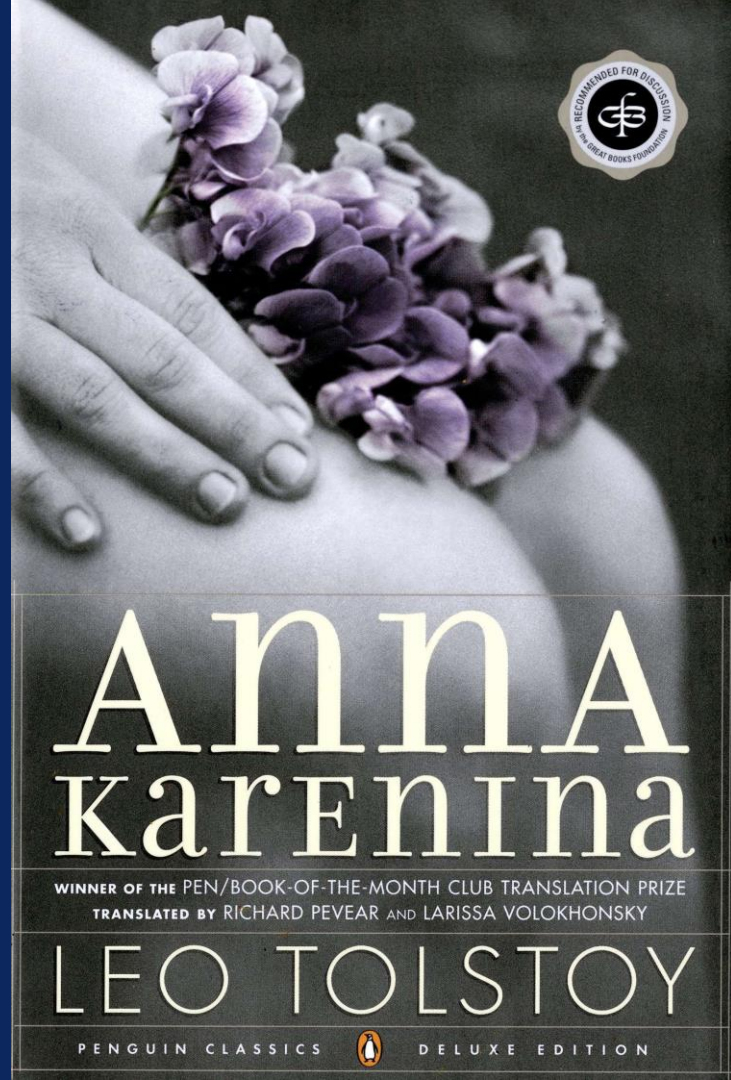
.....



.....

All ***reproducible projects***
are alike;
Each ***non-reproducible***
one ***is not*** in its own way.

.....





192
experiments



27%
success rate



“Seriously, it doesn’t have to be about sharing your code (although that is an added benefit)
It is about saving yourself time.”

Hilary Parker
Data Scientist, StitchFix



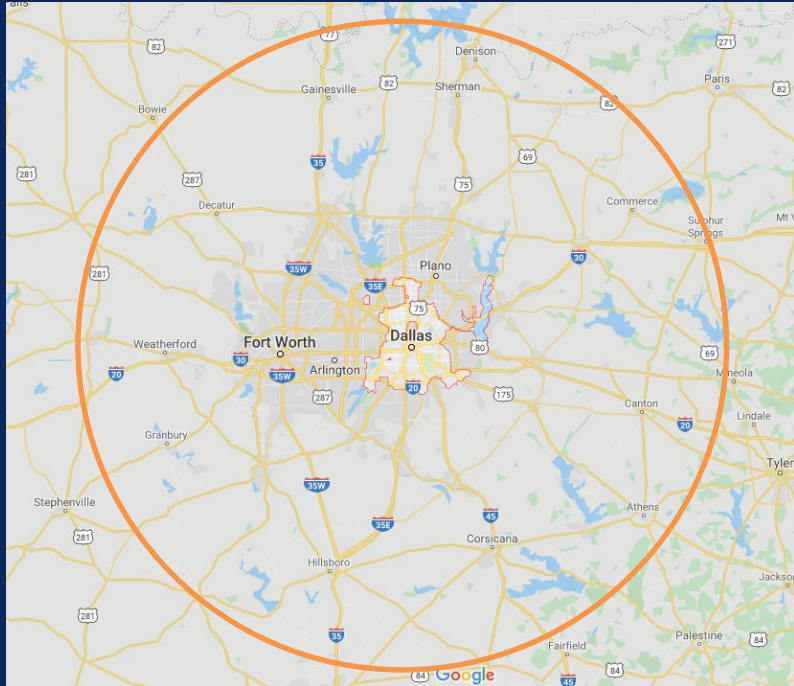
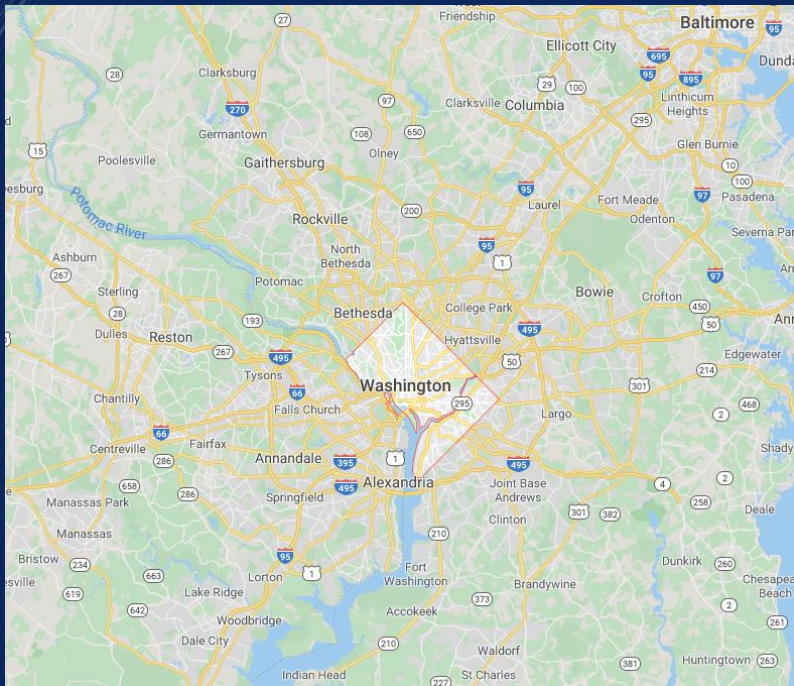


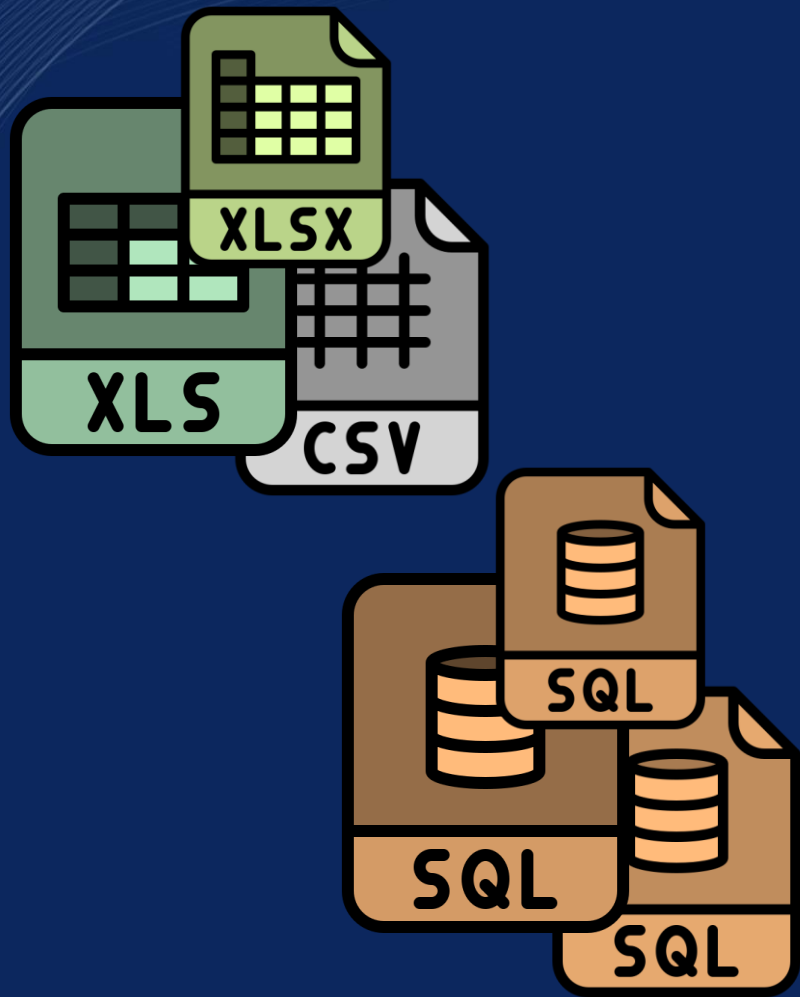
HOW?



SLOW DOWN

- Ask Better Questions



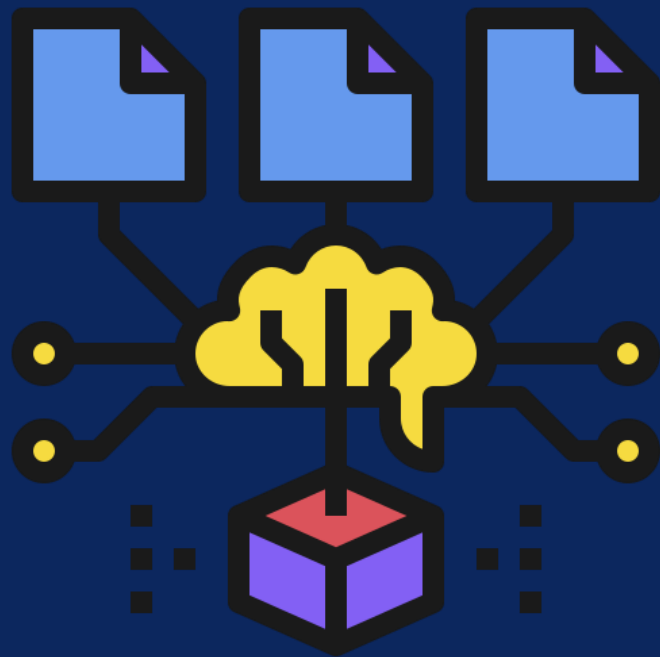


SLOW DOWN

- Ask Better Questions
- Inputs & Outputs
- Know Your Data

KNOW YOUR DATA

- What does the data look like?
- Who has access?
- How does it connect?



KNOW YOUR DATA

- What does the data look like?
- Who has access?
- How does it connect?
- Inputs

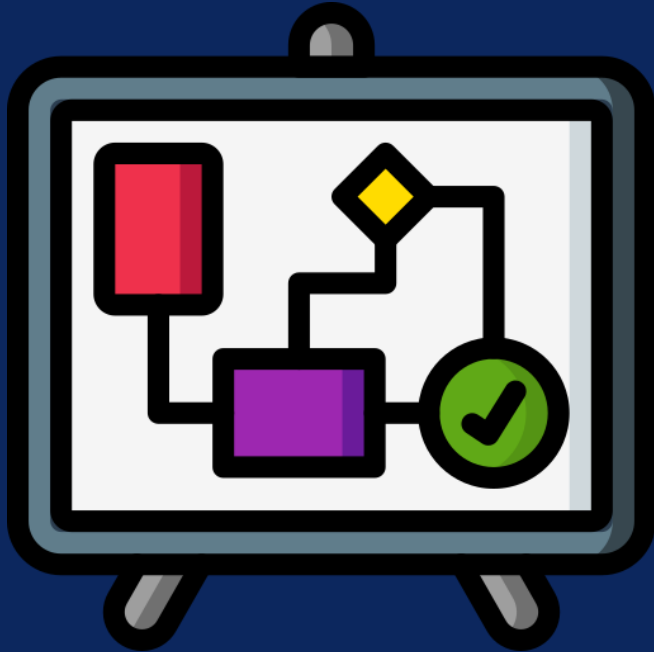


KNOW YOUR DATA

- What does the data look like?
- Who has access?
- How does it connect?
- Inputs and Outputs

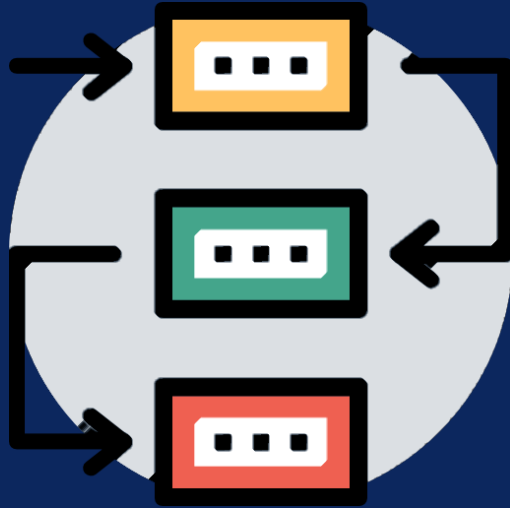


DESIGN



- Start with a question
- Provide context
- Specific, actionable tasks
- Diagram

DESIGN



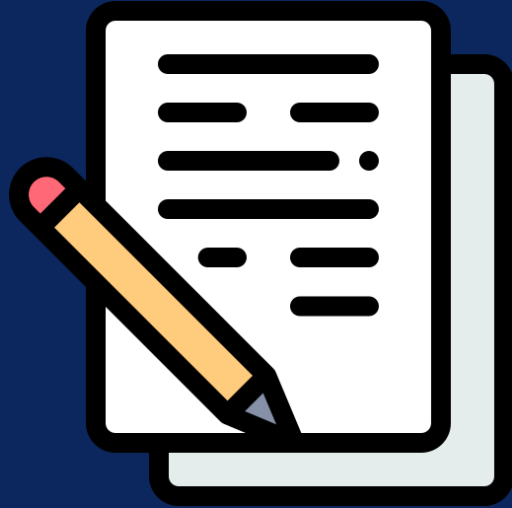
- How to access the data?
- What does the data *need* to look like?
- Preprocessing steps
 - Accounting for randomness

ABSTRACT

- Generalize Tasks
 - Points of Variation
- Simple, monotonic
 - Functions, Files, Scripts



DOCUMENT



- Code can't speak
- The Why
- Leave Breadcrumbs
- External Dependencies

THINKING REPRODUCIBLY

- Slow Down
- Know Your Data
- Design
- Abstract
- Document



QUESTIONS?

CONTACT



rogol@virginia.edu



[/in/jrogol](https://www.linkedin.com/in/jrogol)



[/jrogol](https://twitter.com/jrogol)



[/jrogol](https://github.com/jrogol)

REFERENCES

Brent Waller. "Benny's cousins are in town."

Ikea instructions excerpt from Ikea Koppang Dresser Instructions.

Karthik Ram. "A Guide to Making Your Data Analysis More Reproducible." *rstudio::conf(2019)*. 15 January 2019.

Manuscript Image from Jenny Lawson.

Vincent Warmerdam. "Roman Reasoning." 6 December 2019.

Leo Tolstoy. *Anna Karenina*. Translated by Richard Pevear and Larissa Volokhonsky. 2004.

Jack Grove. "Trouble Replicating Cancer Studies a 'Wake-up Call' for Science." *Times Higher Education*. 23 January 2020.

Center for Open Science Logo from the Center for Open Science.

REFERENCES

Icons made by [Freepik](#), [Darius Dan](#), [Good Ware](#), [iconixar](#), [Becris](#), [Smashicons](#), [ProSymbols](#) [Gregor Cresnar](#) and [Pixel Perfect](#) from www.flaticon.com, licensed [CC 3.0 BY](#).

Caitlin Lukacs. [“Data Storytelling.”](#) *Currents*. CASE. 1 November 2019.

Hilary Parker. [“Writing an R Package from Scratch.”](#) 29 April 2014.

Payday Loans image from [Consumer Financial Protection Bureau](#).

Gordon Shotwell. [“Technical Debt for Data Scientists.”](#) 19 April 2019.

DMV Sloth via [giphy](#).

Washington and Dallas maps from [Google Maps](#).

Microsoft Excel logo from Microsoft.

Emily Robinson and Jacqueline Nolis. [Build a Career in Data Science](#). April 2020.

REFERENCES

Hadley Wickham. R for Data Science. Chapter 15: Functions.

Garrett Grolemond. “Reproducibility in Production (Webinar).” 4 September 2019. Webinar.

Jenny Bryan. “Code Smells and Feels.” *useR!2018*. 13 July 2018.

Vincent Warmerdam. “Untitled12.ipynb.” *PyData Eindhoven 2019*. 30 November 2019.

ROpenSci. “Reproducibility Guide.”

Victoria Stodden. “Reproducibility.” 2014.

Victoria Stodden. “Setting the Default to Reproducible.” *Reproducibility in Computational and Experimental Mathematics*. 16 February 2013.

Andrew Bray, Mine Çetinkaya-Rundel, and Dalene Stangl. “Five Concrete Reasons Your Students Should Be Learning to Analyze Data in the Reproducible Paradigm.” CHANCE. 27 June 2014.