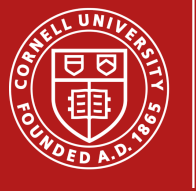


Johnson
Cornell
SC Johnson College of Business

What's going on with your data?

Common Pitfalls and a Few Solutions

Ryan Guggenmos

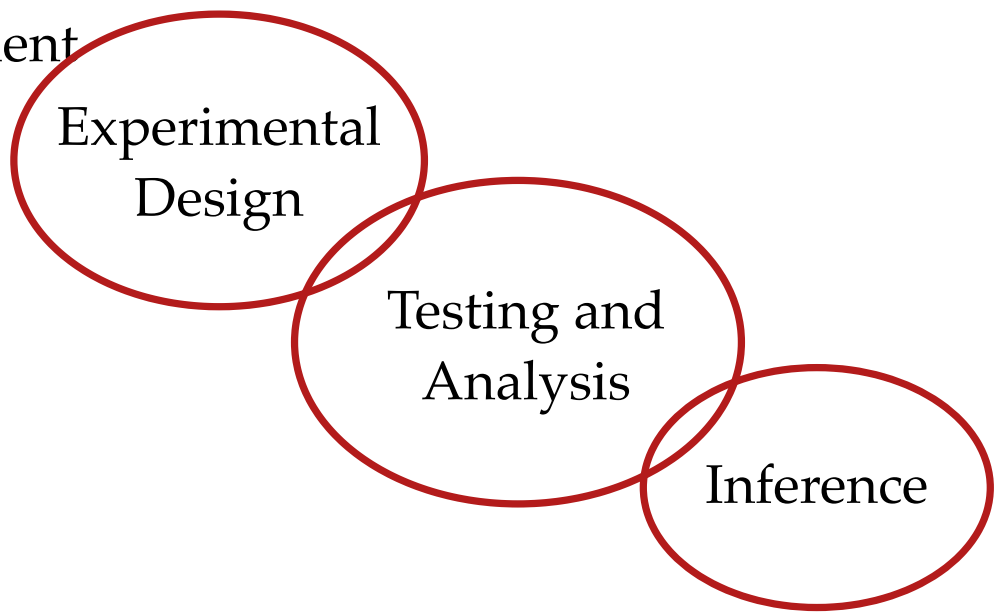
Georgia Tech PhD Seminar - March 31, 2022

Theory-driven Experiments

Theory

Accounting Context

Hypothesis Development

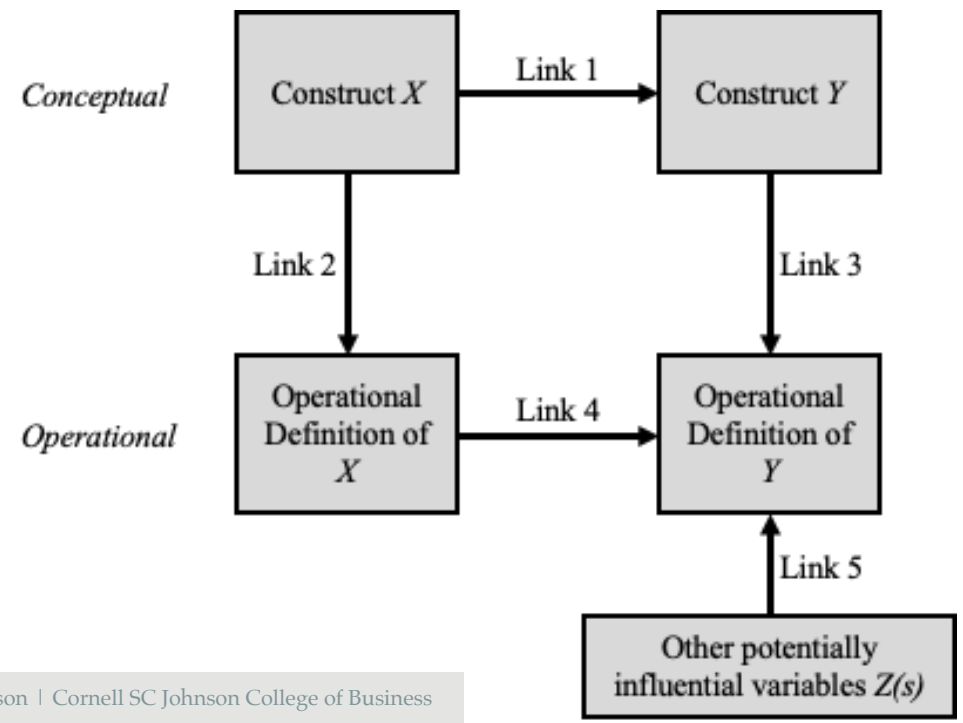


Predictive Validity Framework

(Libby Boxes)

"Link 4 assesses the relations between the operational independent and dependent variables."

Libby, Bloomfield, and Nelson (2002)



But how?

With Statistics!

The missing link...

Link 4 is Statistical Conclusion Validity (SCV).

Statistical Conclusion Validity (SCV) requires that:

- the statistical analysis chosen matches the design employed; and
- the analysis is applied in a way that does not distort the expected probability of Type I or Type II error.

Without adequate SCV, we might not be able to trust our results.

(This is a really big deal.)

Shadish, Cook, and Campbell (2002)

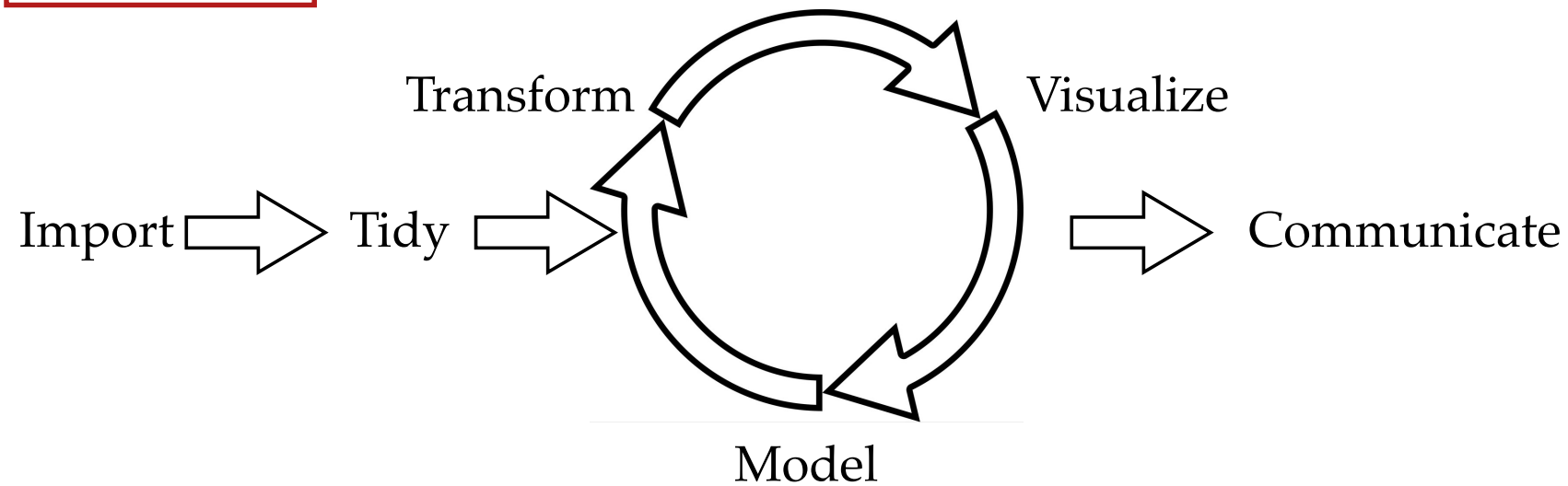
Rethink data analysis.

Make a data analysis plan part of the research ***design*** process,
not something that starts once you have data.

- This means:
  - Starting to think about analysis much before you collect data.
  - Making mindful choices in research design to support SCV.
  - Knowing the implications of your design and analysis choices.
  - Understanding how different statistical methods work (under the hood) and the implications of analysis choices.
  - Documenting your analysis plan so it is reviewable and reproducible.

Data Analysis Workflow

Research Design, first!



Grolemund and Wickham (2019)

Data analysis planning during design

Remember: No amount of stats magic can fix poor design.

Some questions to ask yourself:

- How does my analysis plan map to my hypotheses and theory?
- What, when, and how am I going to measure "stuff"?
  - DV(s), measured IVs, mediators, manipulation checks, items to rule out competing explanations, etc.
  - What do I expect the data collected for each of these measures to look like? Why?
- What purpose does each measure in my instrument serve?
- What are my exclusion criteria? How much data am I collecting?
- What analysis method(s) will I be using? What do they assume? Are these assumptions realistic for my expected data?

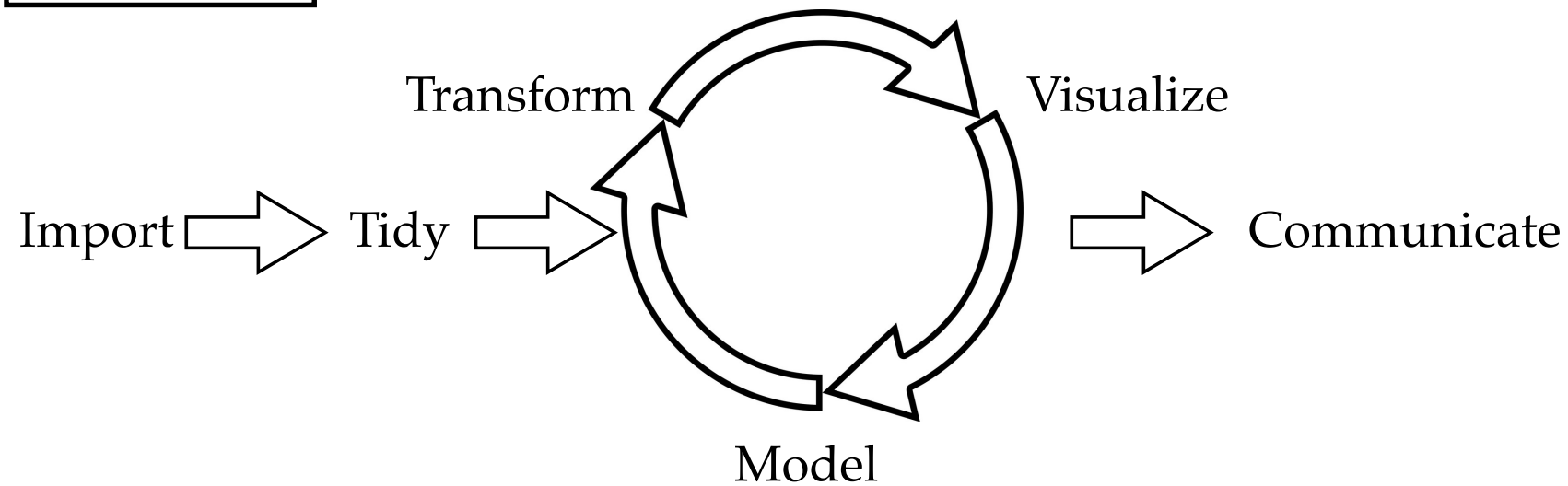
What will these data look like?

In your study, you plan on collecting these measures to test your hypotheses:

- Number of evidence items examined.
- Agreement with the client's position on a fair value estimate:
  - (Scale from 1 to 5, with 1 = "Do not agree with the client" to 5 = "Completely Agree with the client")
  - (Scale from 0 to 100, with 0 = "Do not agree" to 100 = "Completely Agree")
- Reaction time in a psychology-based task used to manipulate the underlying psychological construct.

Data Analysis Workflow

Research Design, first!



Grolemund and Wickham (2019)

The data import process

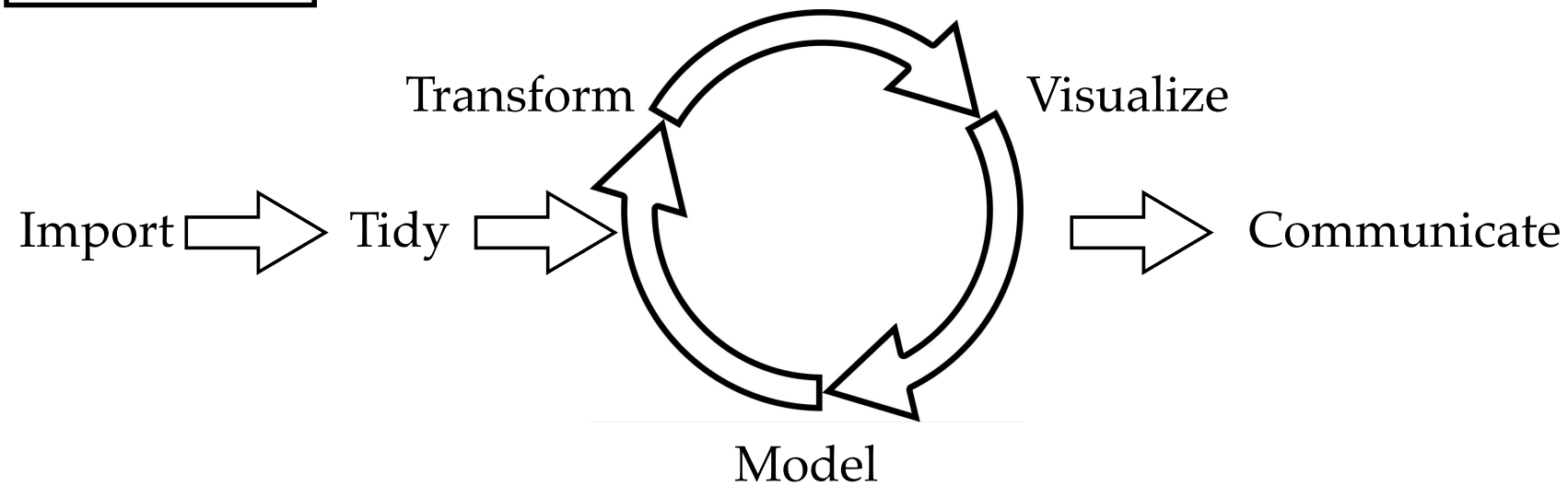
Goal: Enter all the data accurately into the stats software. No more, no less.

Questions to ask yourself:

- Is all the data there? (I paid 153 participants, but I have 150 observations...)
- Do I have "bonus" data? (I paid 150 participants, why do I have 153 observations...)
- Are my cell counts balanced? If not, why?
- How am I going to store this data and how long do I need to store it?

Data Analysis Workflow

Research Design, first!



Grolemund and Wickham (2019)

Tidy

Goal: Get the raw data into a standard format, so you can work with it efficiently and effectively.

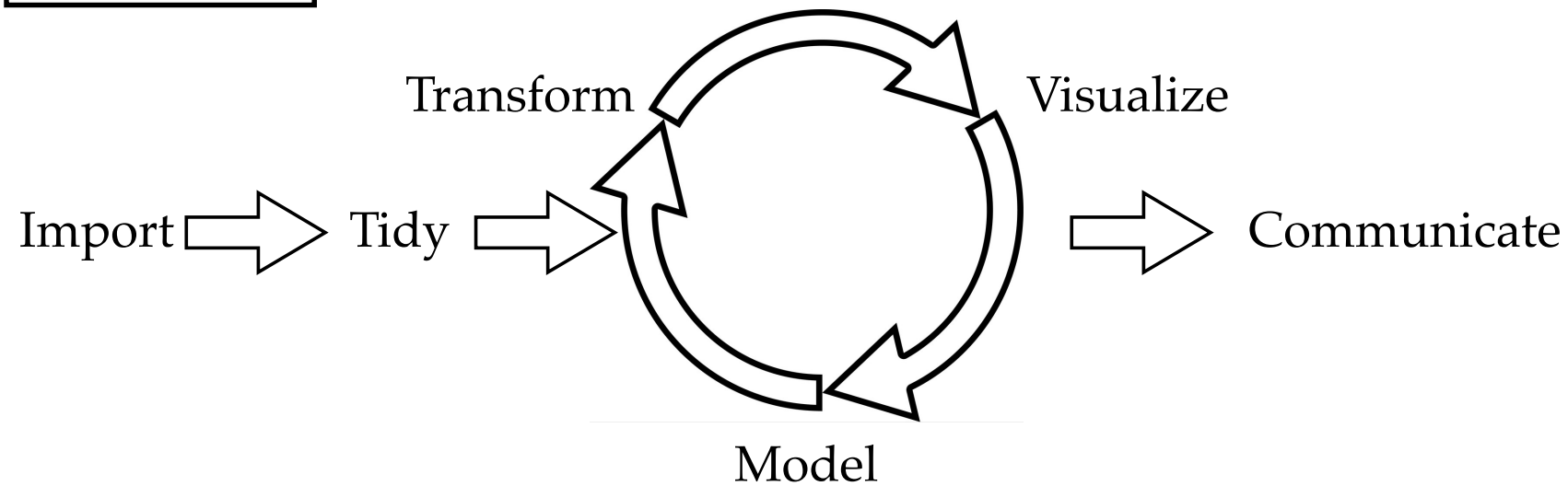
Tidy data - each variable has its own column, each observation has its own row, and each value has its own cell.

Questions to ask yourself:

- What format are each of my variables (numeric, strings / characters, factors, dates)? What format should they be?
- Do I need to combine data sets? (Qualtrics data + coded responses from independent raters)
- Do I have missing observations? If so, how do I want them coded in the data? (NA, -99, .)
- If I come back to this project in a year, have I documented what I've done well enough to reproduce this step?

Data Analysis Workflow

Research Design, first!



Grolemund and Wickham (2019)

Transform

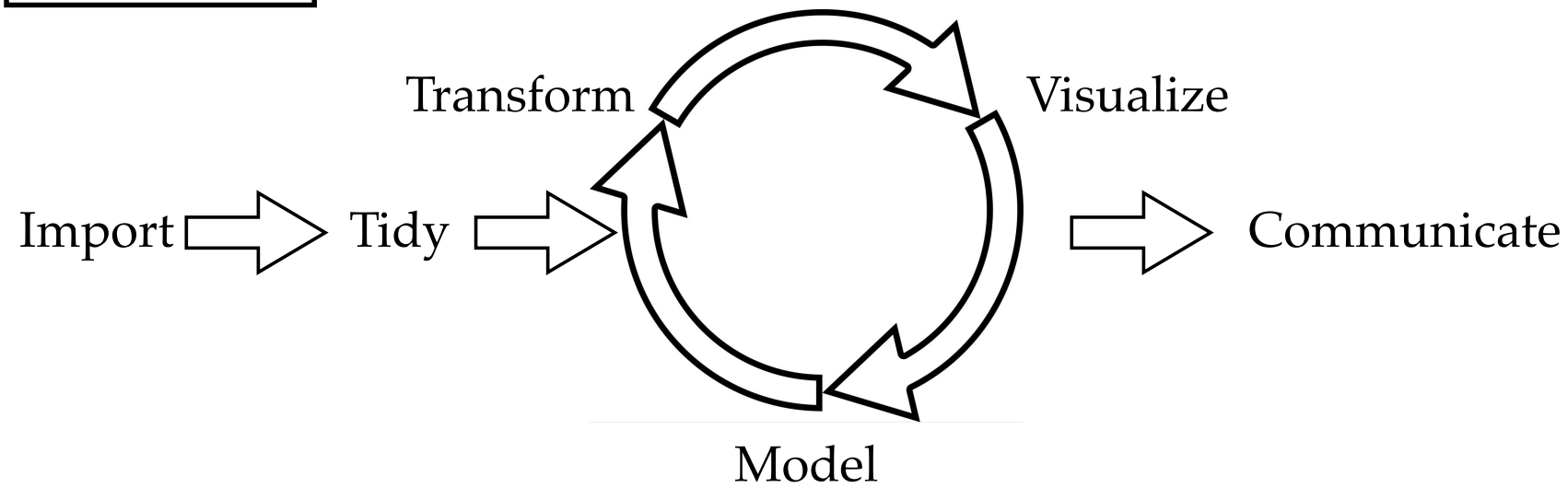
Goal: Take tidy-ed raw data and make it easier to work with and interpret.

Questions to ask yourself:

- Do I need to apply any data transformations? If so, why? and how should I do it?
- Does any of my data need recoded? If so, how should I code it?
- Do new variables need to be created?
- Do any of the cases meet the exclusion criteria?
- Are any of the variable names hard to understand? If so, can I make them more readable?

Data Analysis Workflow

Research Design, first!



Grolemund and Wickham (2019)

Visualize - look before you leap!

Goal: Gain a better understanding of the general properties of your data, in order to guide the rest of the process.

Looking at descriptive statistics is necessary, but not sufficient!

Visualize

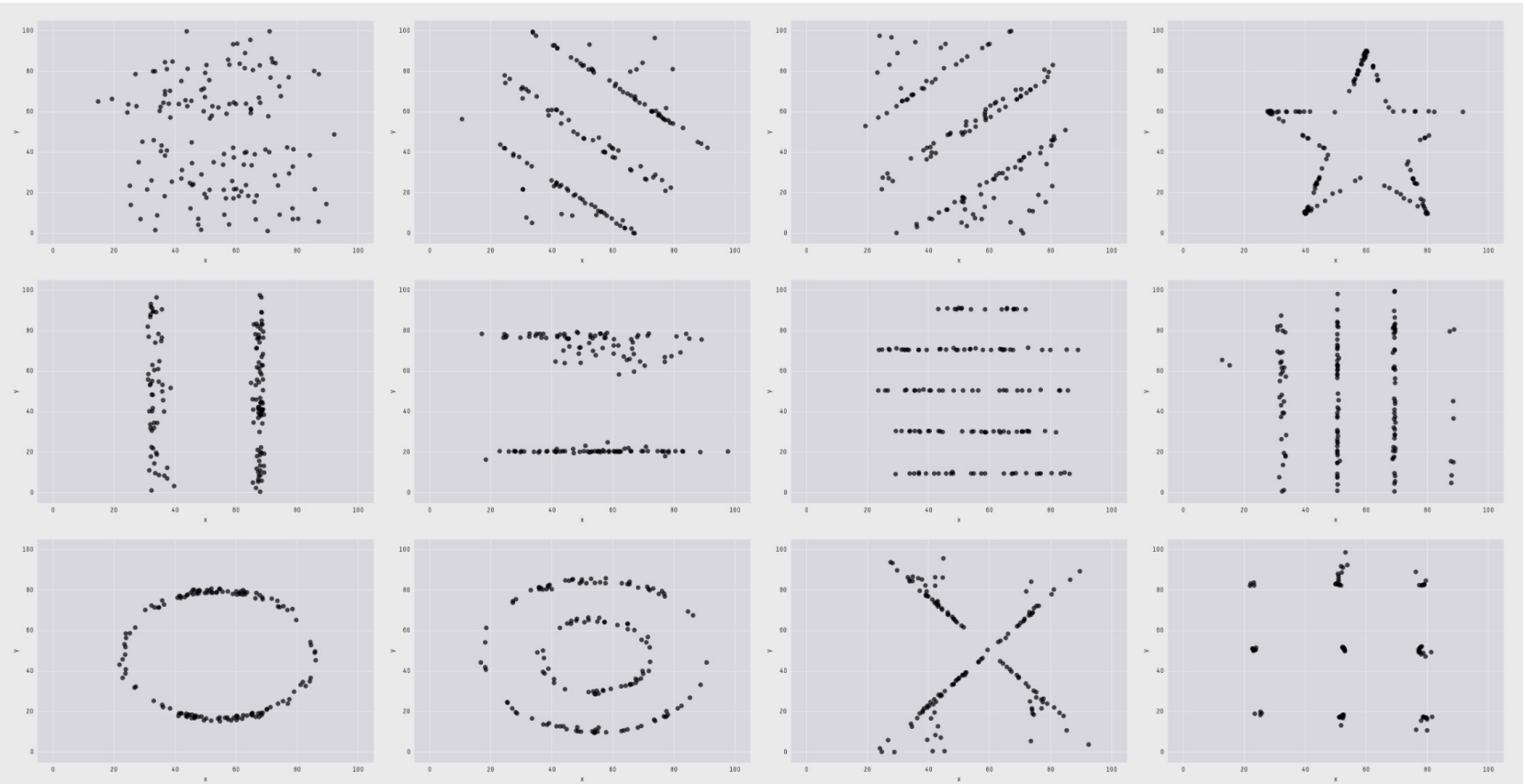
Goal: Gain a better understanding of the general properties of your data, in order to guide the rest of the process.

Questions to ask yourself:

- Before I look at the data, what do I expect to see?
- Which variables do I expect to be correlated (or not)?
- Is my data normal? Should it be?
- Do I have outliers? Are they influential?
- How am I treating missing data?
- Does my data exhibit heteroskedasticity (unequal variance)?
- (after modeling) - what do my residuals look like?

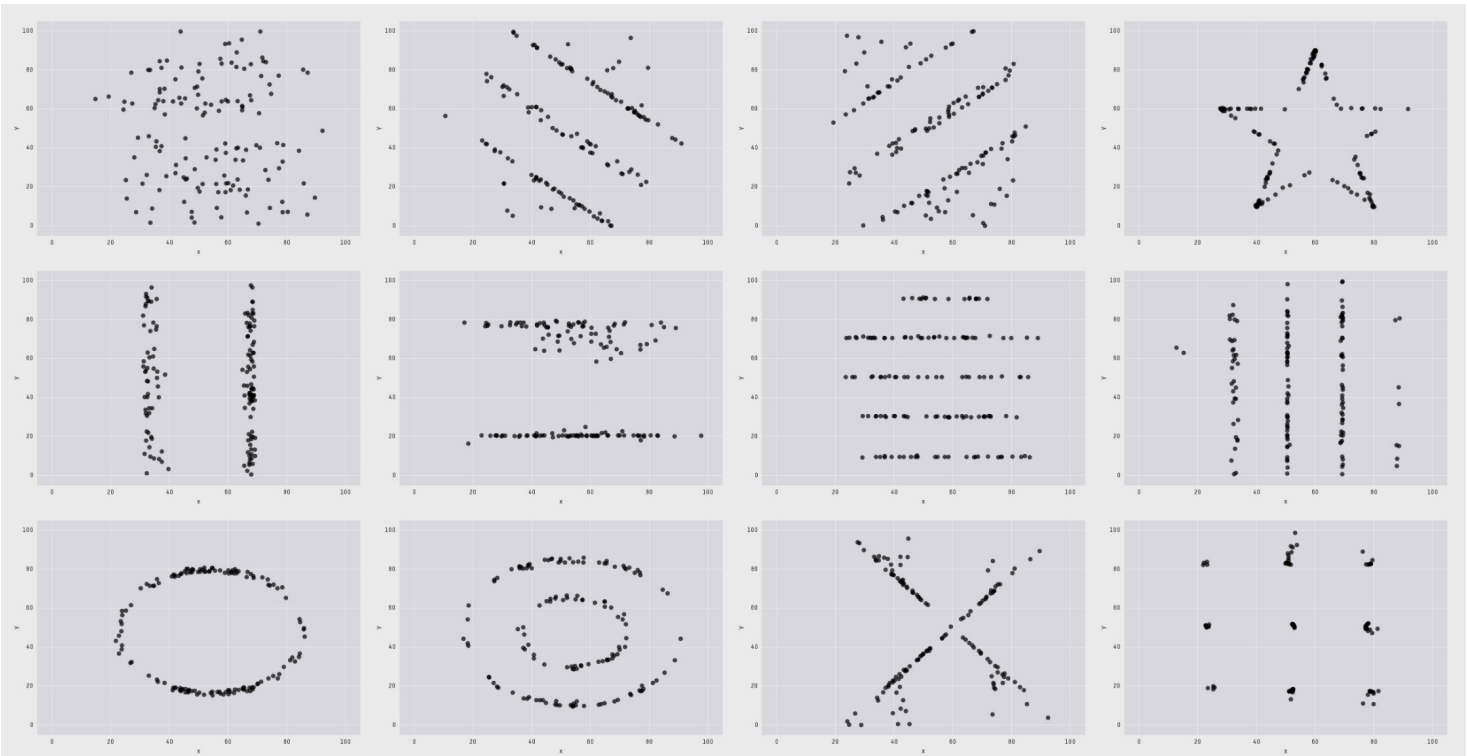
Pick-a-plot!

r = -0.06	X	Y
mean	54.26	47.83
Std. dev.	16.76	26.93



Visualize

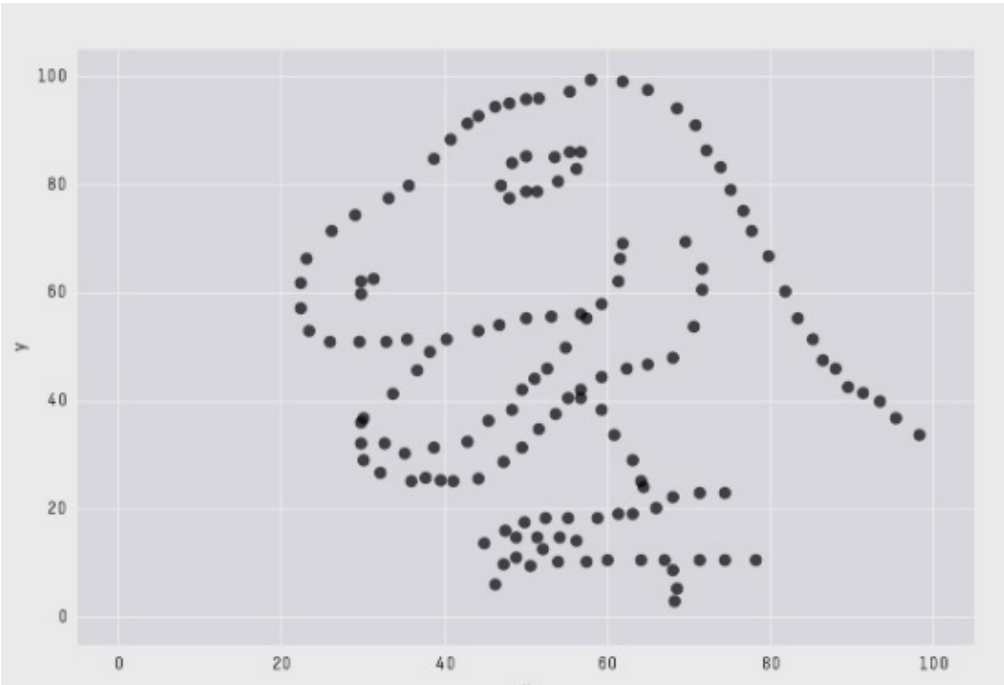
All of them.



Matejka and Fitzmaurice (2017)

Visualize

So does this one.

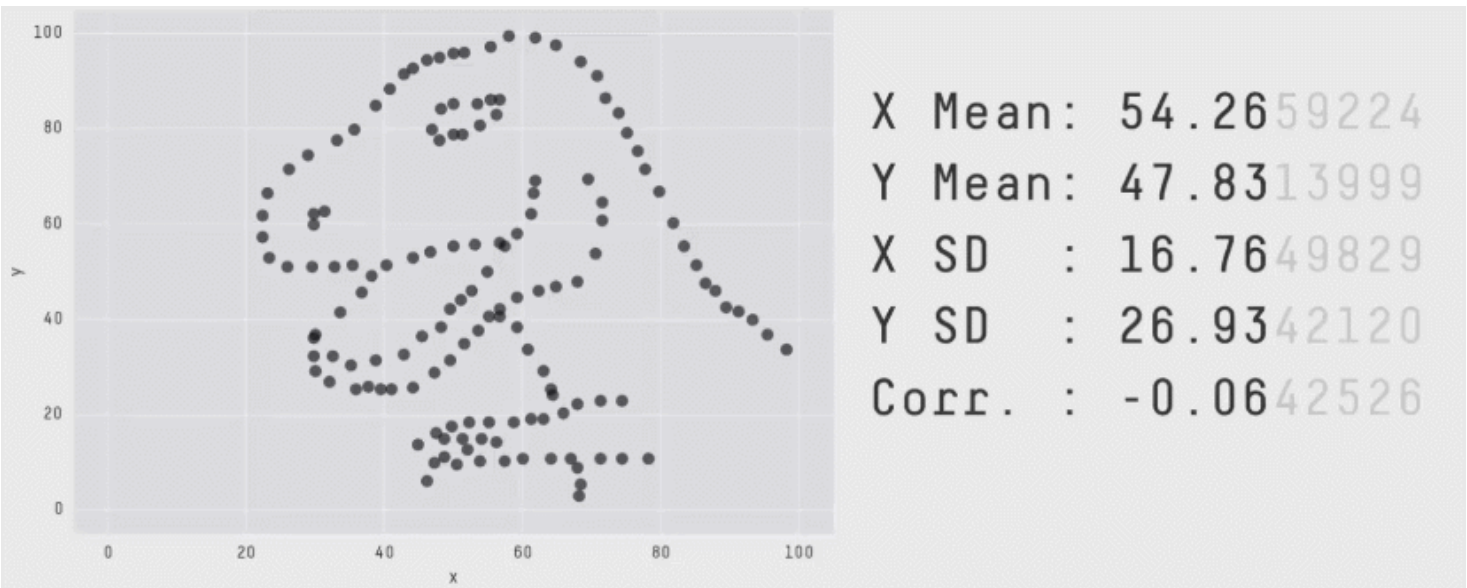


Matejka and Fitzmaurice (2017)

Visualize

"...make both calculations and graphs. Both sorts of output should be studied; each will contribute to understanding."

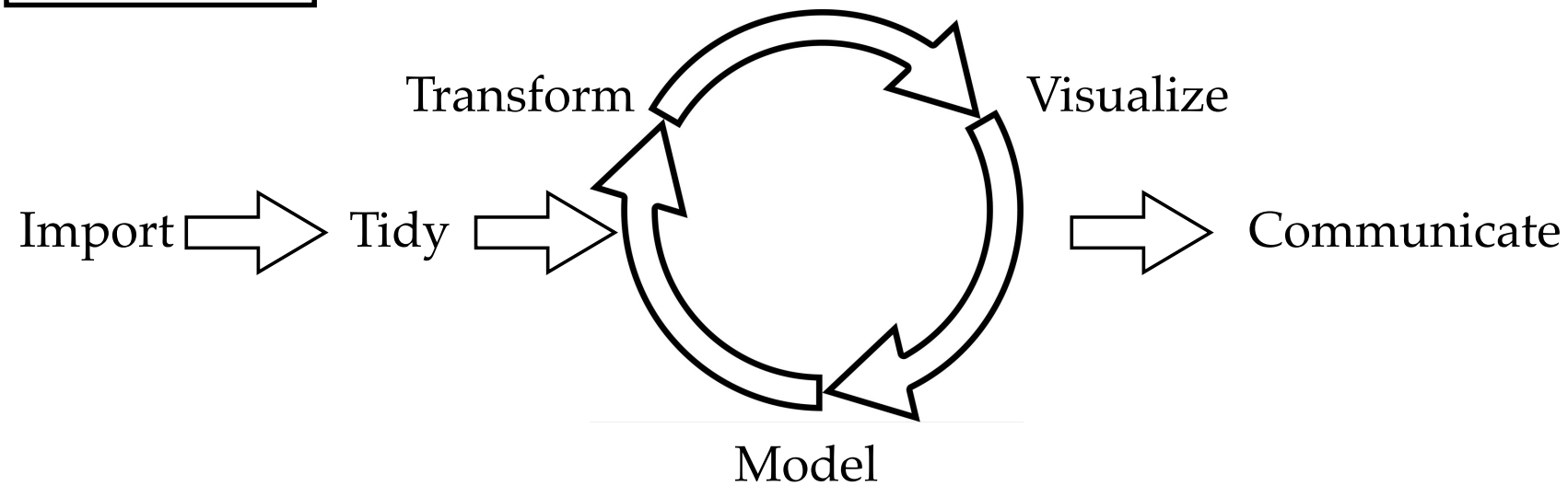
Anscombe (1973)



Matejka and Fitzmaurice (2017)

Data Analysis Workflow

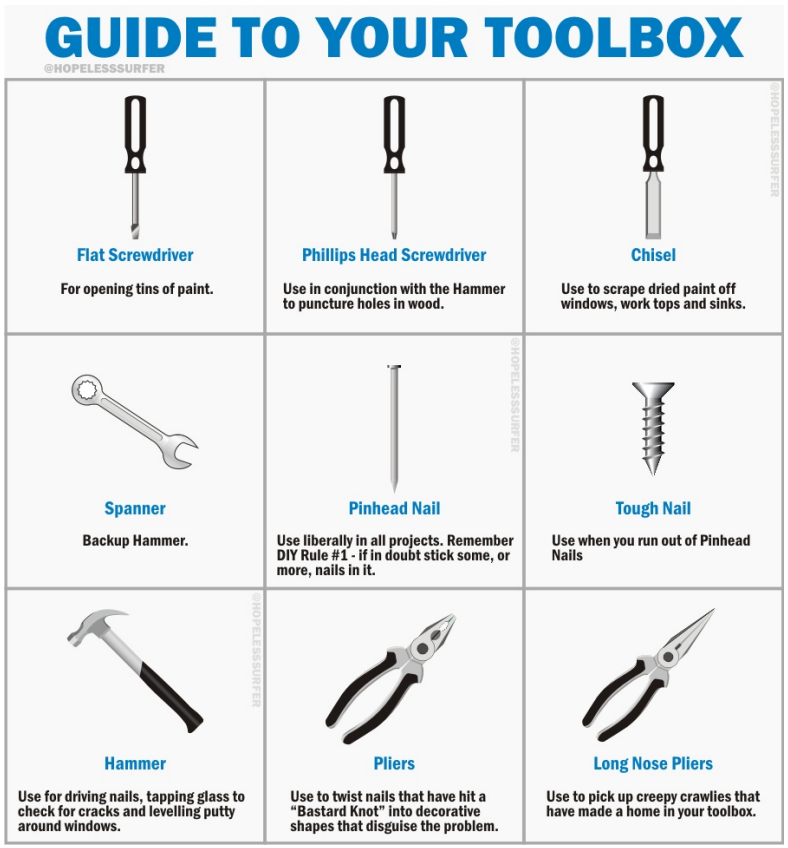
Research Design, first!



Grolemund and Wickham (2019)

Model - Analyze this!

Goal: Partition the data into information and noise (patterns and residuals, effects and error) and draw inference.



Model

Goal: Partition the data into information and noise (patterns and residuals, effects and error) and draw inference.

Questions to ask yourself:

- Why am I using this statistical analysis technique?
- What are the "defaults" for this technique in my software package?
- Are the defaults appropriate?
- What are the assumptions of this method?
- Am I violating these assumptions? What are the consequences?
- How should I interpret my p-value? What if it is a null result?
- Do I need to show robustness of my result? If so, how should I?
- Are my tests over or underpowered? If so, how could that affect my inference?

Example: ANCOVA

You've randomly assigned participants to experimental conditions and realize that one cell has participants that possess greater investment experience. This difference is significant, so you include investment experience as a covariate. The main DV is willingness to invest.

Is this an appropriate use of ANCOVA?

Example: ANCOVA

You've randomly assigned participants to experimental conditions and realize that one cell has participants that possess greater investment experience. This difference is significant, so you include investment experience as a covariate. The main DV is willingness to invest.

Is this an appropriate use of ANCOVA?

You want to evaluate the effect of an intervention that is designed to increase professional skepticism. Because you want to assess the effect of the intervention, controlling for pre-existing levels of skepticism, you administer a skepticism pre-test and use this score as a covariate. The main DV is the post-test score.

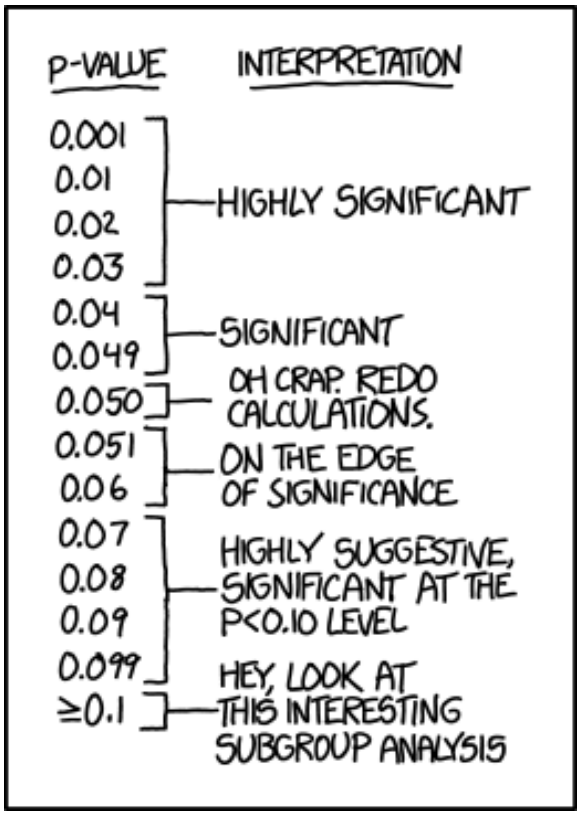
Is this an appropriate use of ANCOVA?

Example: ANCOVA

"[N]o amount of statistical manipulation can tell one what might have been had certain differences been non-existent.. .. The overwhelming weight of logic is on the side of those who warn that neither the analysis of covariance nor any other statistical technique can undo systematic differences which were out of the investigator's control."

Fleiss and Tanur (1973)

Draw inference (in context)



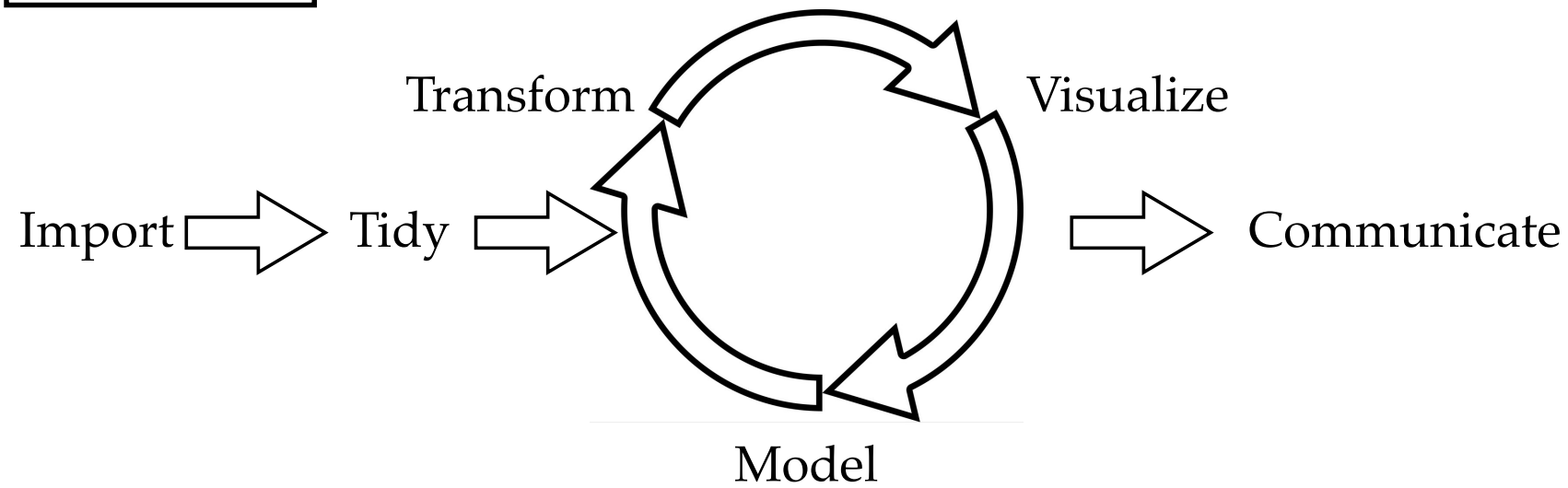
What's a "good" p-value?

The one that is least affected by bias and researcher degrees of freedom.

www.xkcd.com

Data Analysis Workflow

Research Design, first!



Grolemund and Wickham (2019)

Communicate - tell the story!

Communicate

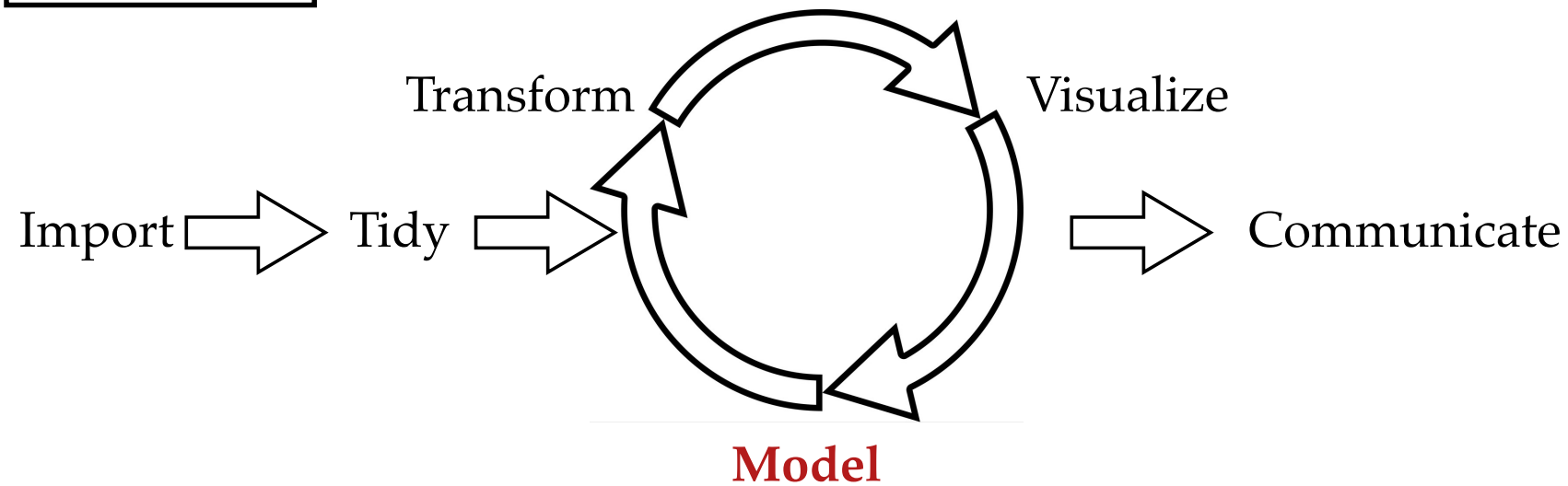
Goal: Report what you did, why you did it, and what it means in a way that tells a story that people can understand.

Questions to ask yourself:

- Is there enough information provided that readers will understand how my analysis provides evidence for my theory or that other plausible explanations are less likely to explain the effect?
- Is my write-up approachable to the average reader?
- Is it clear how my analysis provides evidence to support my theory?
- Will readers understand whether my data are robust to exclusing participants?
- Am I careful not to over or underclaim the evidence that my analysis provides?

Data Analysis Workflow

Research
Design, first!



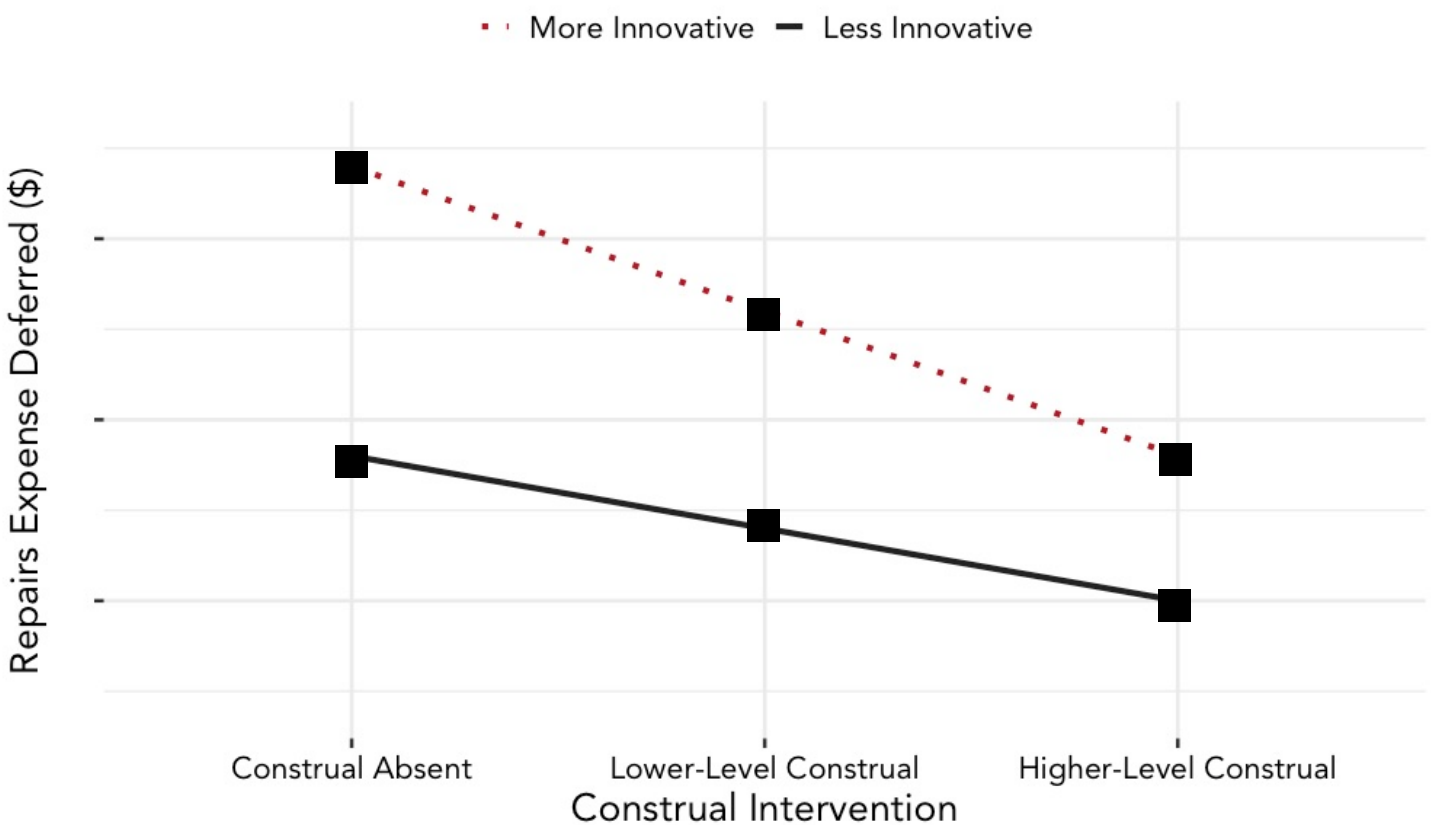
Grolemund and Wickham (2019)

Final thoughts.

- Planning data analysis as part of the research design process can help you anticipate some common analysis pitfalls.
  - If you anticipate these pitfalls, you may be able to steer around them!

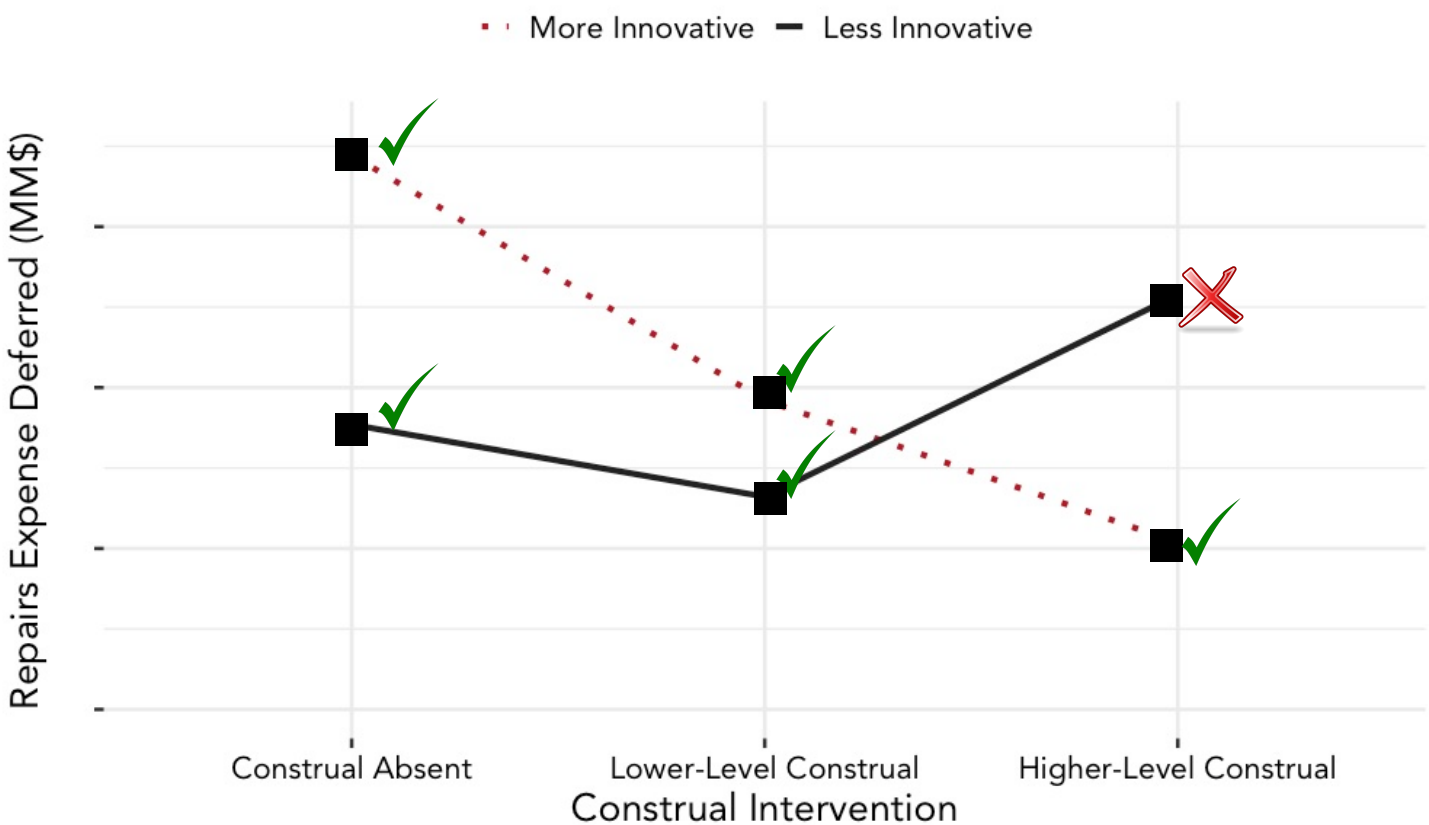
- Data is messy and not everything will work out as planned.
  - That's ok. Data from a well-designed experiment will contain information, even if it's not quite the information you expected.

Predicted Results...



Guggenmos (2019)

Reality...



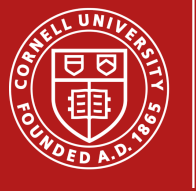
Guggenmos (2019)

Final thoughts.

- Planning data analysis as part of the research design process can help you anticipate some common analysis pitfalls.
  - If you anticipate these pitfalls, you may be able to steer around them!

- Data is messy and not everything will work out as planned.
  - That's ok. Data from a well-designed experiment will contain information, even if it's not quite the information you expected.

- Don't be shady - or anywhere close to shady. Ever. Period.



Johnson
Cornell
SC Johnson College of Business

Thank you.

And good luck with your papers!

Questions?