data cleaning principles

Karl Broman

Biostatistics & Medical Informatics, UW-Madison

@kwbroman
 kbroman.org
 github.com/kbroman
kbroman.org/Talk_DataCleaning



Tidy data are all alike, but every messy dataset is messy in its own way.

Hadley Wickham

If I clean up [Medicare] data ...
does any of the knowledge I gain ...
apply to the processing of RNA-seq data?

Roger Peng

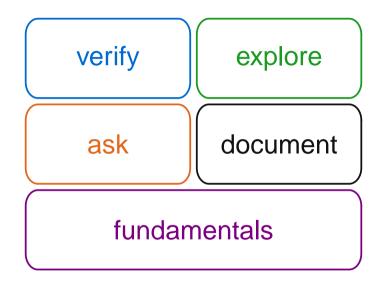
Data Mishaps Night

Join us for the first inaugural Data Mishaps Night! We will feature a lineup of data mistake stories with a focus on the human aspect of data work and lessons learned the hard way.



Data cleaning

- ▶ tedious
- embarrassing
- needs context
- ► doesn't feel like progress
- requires creativity
- ► requires coding prowess
- ► source of most problems



1. Don't clean data when you're tired or hungry.

(paraphrasing Ghazal Gulati)

2. Don't trust anyone (even yourself)

2. Don't trust anyone (even yourself)

"my motto is 'trust no one' ...except maybe @kwbroman?"

Jenny Bryan

3. Think about what might have gone wrong and how it might be revealed

3. Think about what might have gone wrong and how it might be revealed

4. Use care in merging

5. Dates & categories suck

Principle:

A fundamental truth that guides our thinking

5. Dates & categories suck

6. Check that distinct things are distinct

7. Check that matching things match

8. Check calculations

9. Look for other instances of a problem

10. Make lots of plots

11. Look at missing value patterns

12. With massive data, make more plots not fewer

13. Follow up all artifacts

ask

- 14. Ask questions
- 15. Ask for the primary data
- 16. Ask for metadata
- 17. Ask why data are missing

document

- 18. Create checklists & pipelines
- 19. Document not just what but why
- 20. Expect to recheck

Slides: kbroman.org/Talk_DataCleaning



kbroman.org

github.com/kbroman

@kwbroman