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

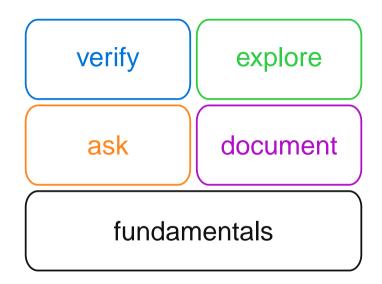
Data Mishaps Night

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



fundamentals

verify

explore

ask

document

Slides: kbroman.org/Talk_DataCleaning



kbroman.org

github.com/kbroman

@kwbroman