data cleaning principles

Karl Broman

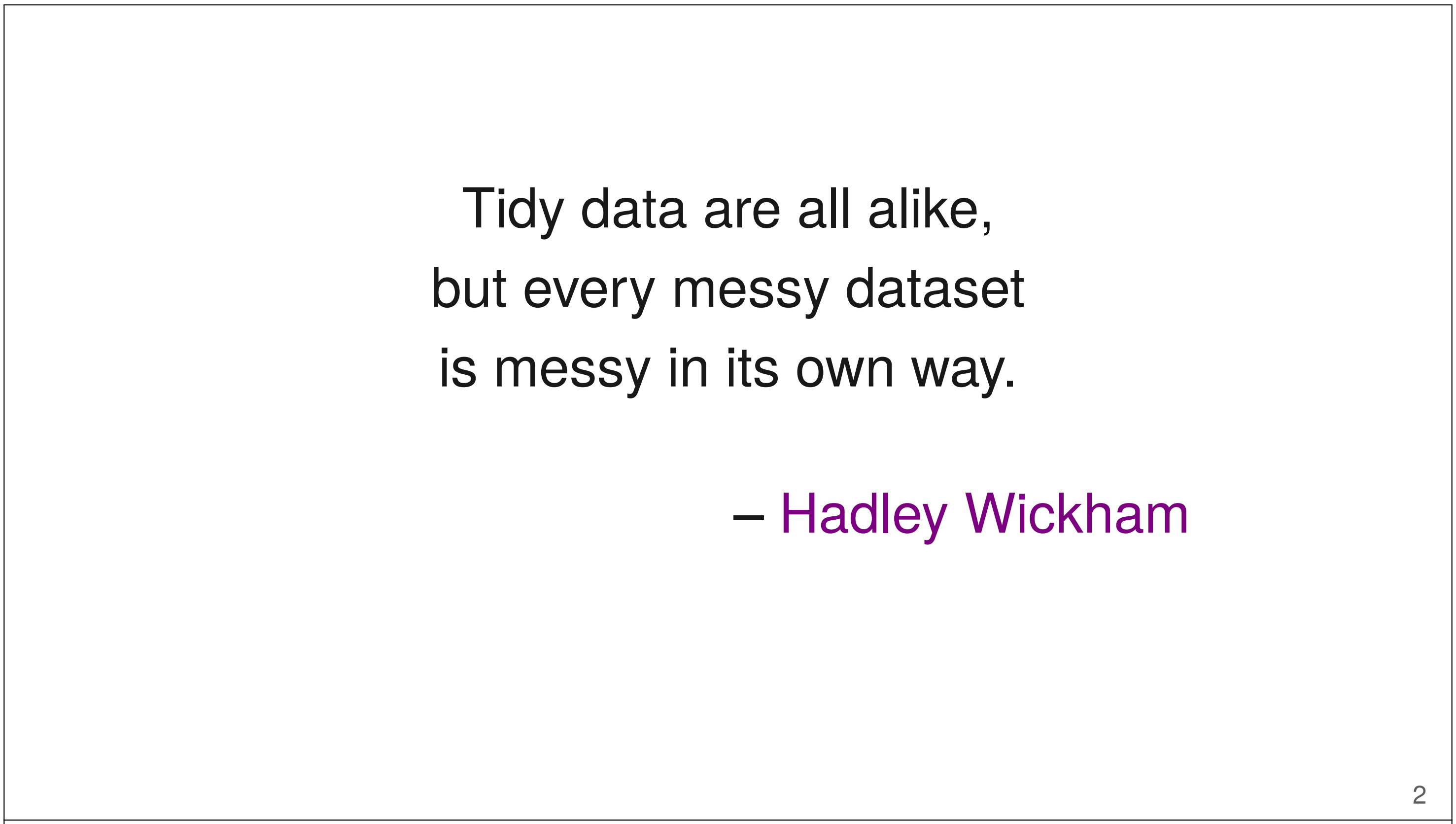
Biostatistics & Medical Informatics, UW–Madison

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

These are slides for a talk for the csv,conf,v6 (https://csvconf.com/) on May 4-5, 2021.

Data analysts spend a lot of time organizing and cleaning data, but few of us have been trained to do so. Why is that?

Some say that data cleaning is difficult to generalize. But I think there are some general principles. Moreover, I think we have an important shared experience in data cleaning that we can commiserate about, and through which we can learn from each other.



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

Hadley's talking more about data organization than data cleanliness. And his point is that if you make data tidy, it simplifies all the downstream analyses.

But is every messy dataset uniquely messy?

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

In his discussion of David Donoho's paper about data science, Roger Peng spoke about how data cleaning is frustratingly difficult to generalize.

But my answer to his question is absolutely!

A person with experience cleaning one dataset has important experience to draw upon when moving to another dataset even if it's of a totally different nature.

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.

dataMishapsNight.com

In February, 2021, Caitlin Hudon and Laura Ellis organized an Friday evening conference where 16 people gave short presentations on data mishaps.

Many of the stories concerned mistakes in data cleaning, and while these weren't necessarily the most amusing stories, they did seem to bring out a strong sense of shared experience. We have suffered and struggled through very similar data problems.

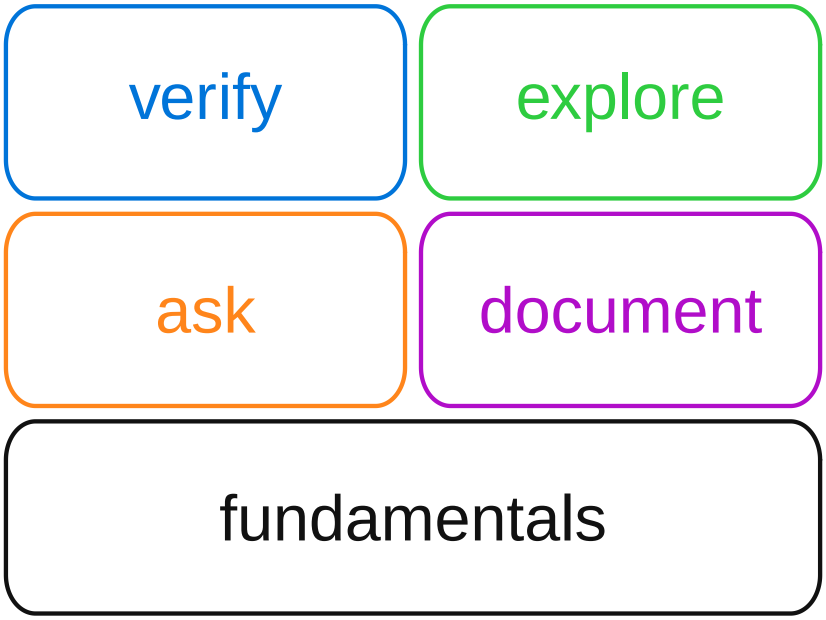
Data cleaning

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

Really, I think we don’t usually teach data cleaning because it’s something we prefer to keep private.

We’re shy about it.

And data cleaning code is our ugliest code.



verify

explore

ask

document

fundamentals

fundamentals

verify

explore

ask

document

Slides: kbroman.org/Talk_DataCleaning

kbroman.org

github.com/kbroman

@kwbroman