EDS 240: Discussion 1

Data Wrangling

Week 1 | January 7th, 2025

What do we mean by "data wrangling?"

> *"Data wrangling, sometimes referred to as data munging, is **the process of transforming and mapping data from one "raw" data form into another format** with the intent of making it more appropriate and valuable for a variety of downstream purposes such as analytics."*

-Wikipedia

Wrangling includes (but is not limited to):

- **data cleaning** (e.g. handling missing values, consistent naming of observations, ensuring variables are the correct data types, etc.)
- **restructuring data** (e.g. tidying data (i.e. convert from wide > long format))
- **combining data sets** (e.g. using key values to merge two related data sets)

Wrangling is a critical first step in building any sort of data visualization!

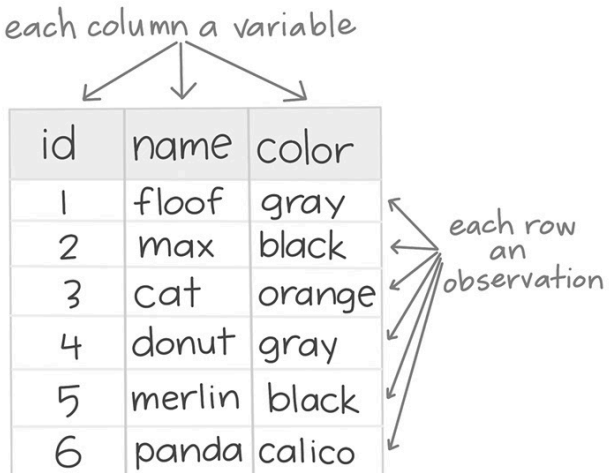
{ggplot2} plays best with *tidy* data

"TIDY DATA is a standard way of mapping the meaning of a dataset to its structure."

–HADLEY WICKHAM

In tidy data:

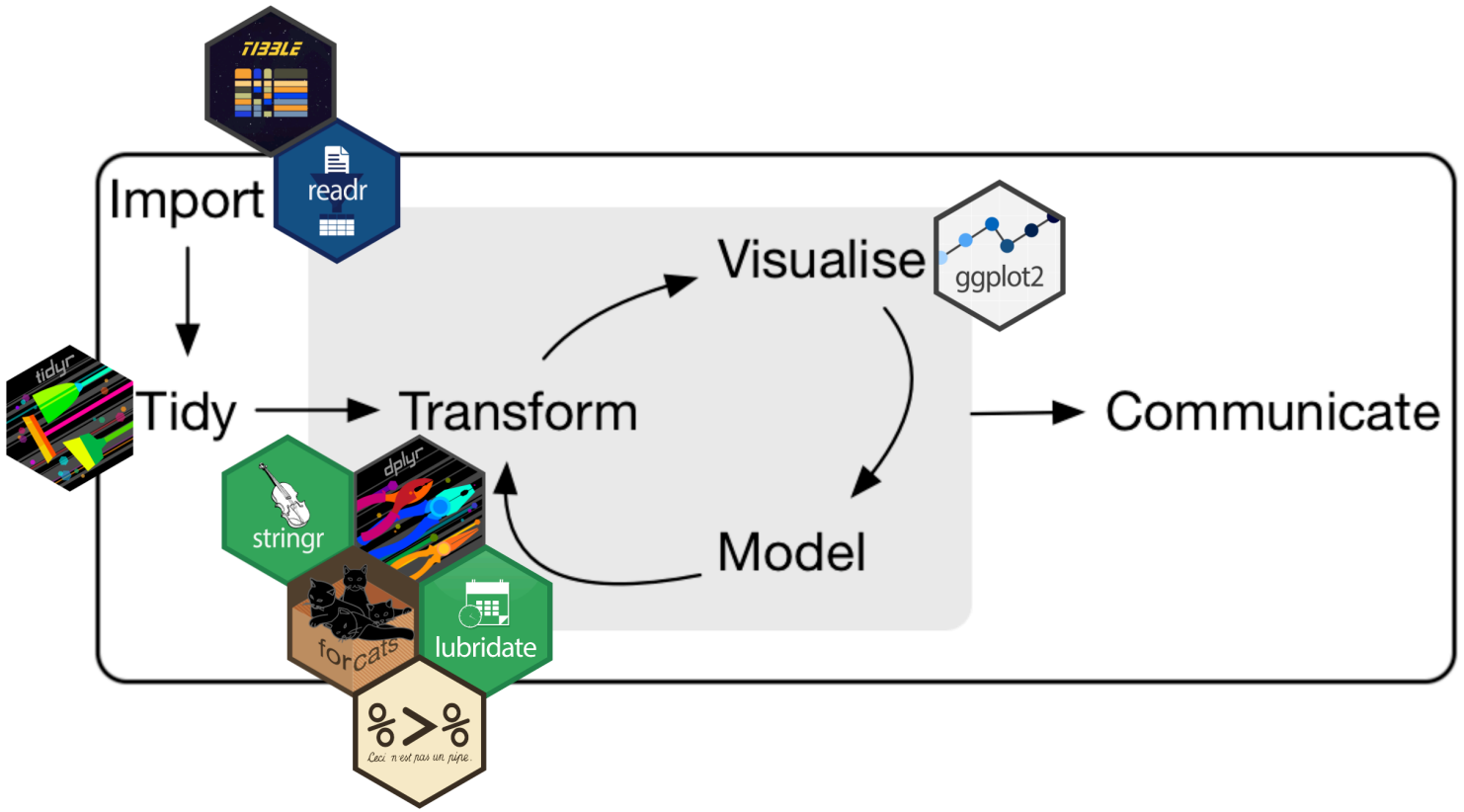
- each variable forms a column
- each observation forms a row
- each cell is a single measurement



Wickham, H. (2014). Tidy Data. Journal of Statistical Software *59* (10). DOI: 10.18637/jss.v059.i10

Artwork by Allison Horst

The `{tidyverse}` provides lots of helpful tools



The data science workflow, as described by Hadley Wickham, Mine Çetinkaya-Rundel and Garrett Grolemund in R for Data Science (2e), with added `{tidyverse}` packages as they fit within this workflow.

Let's wrangle some fracking data

> *Since launching in 2011,* ***FracFocus*** ***has become the largest registry of hydraulic fracturing chemical disclosures in the US****. The database, available to explore online and download in bulk, contains 210,000+ such disclosures from fracking operators; it details the location, timing, and water volume of each fracking job, plus the names and amounts of chemicals used. The project is managed by the Ground Water Protection Council, "a nonprofit 501(c)6 organization whose members consist of state ground water regulatory agencies". As seen in: The latest installment of the New York Times' Uncharted Water series.*

-Jeremy Singer-Vine on Data is Plural (2023.09.27 edition)

Download fracking data from Google Drive!

You should already have downloaded these data from Google Drive

We snagged these data back in November 2023 when they were still *quite* messy. Since then, (it seems that) FracFocus has done a bit more pre-processing of these data – meaning the data you download from their online portal is already a whole lot cleaner. This is great(!), but also defeats the purpose of this exercise ☻.

Open up the Week 1 Discussion: Exercise for instructions / next steps.