Online CDS 101 *Introduction to Computational and Data Sciences*

Data transformation

What is data wrangling?

The word "wrangle"

The word "wrangle"

wrangle

> *verb*
>
> *to tend or round up (cattle, horses, or other livestock).*
>
> — dictionary.com

The word "wrangle"

wrangle

> *verb*
>
> *to tend or round up (cattle, horses, or other livestock).*
>
> — dictionary.com

- So, by analogy, "wrangling data" means to collect, clean, and organize digital information (tend and round up)

The word "wrangle"

wrangle

> *verb*
>
> *to tend or round up (cattle, horses, or other livestock).*
>
> — dictionary.com

- So, by analogy, "wrangling data" means to collect, clean, and organize digital information (tend and round up)
- Informal word, but data scientists will understand what you mean if you use it

The word "wrangle"

wrangle

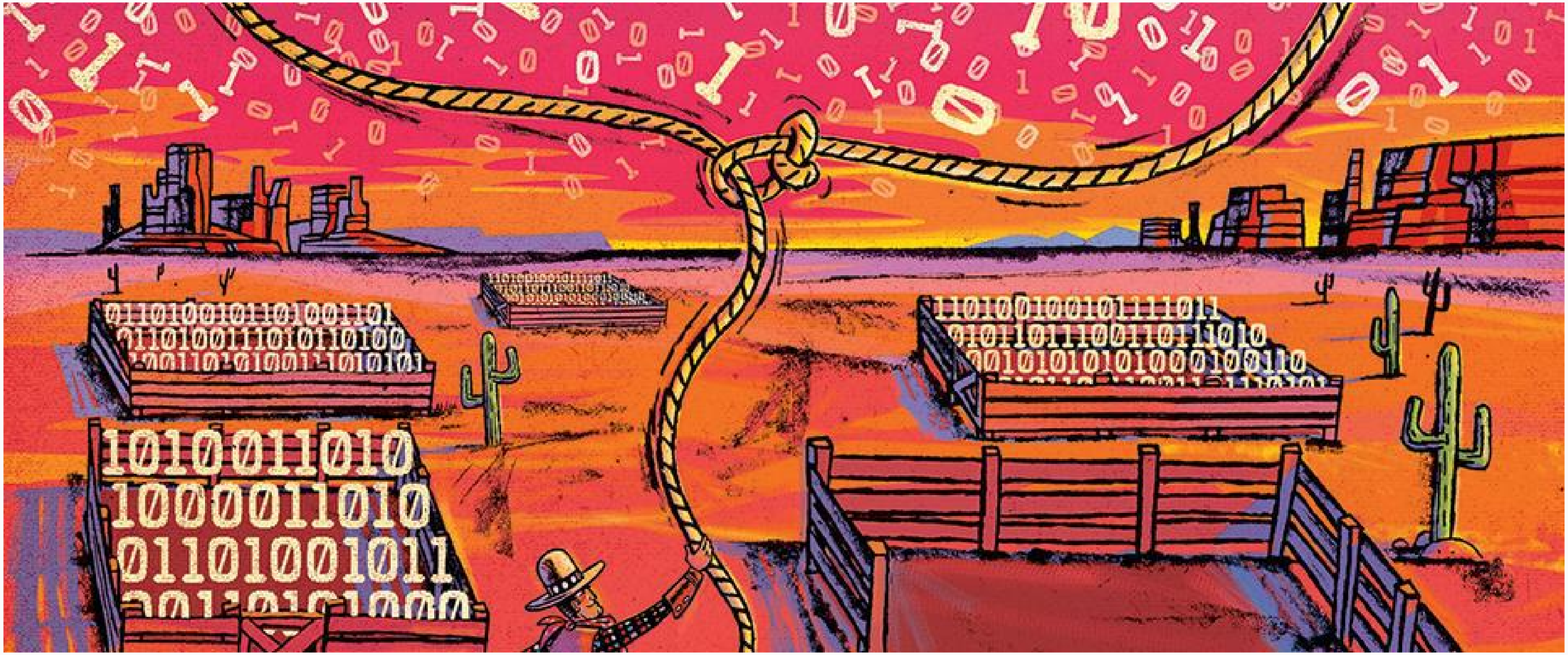
> *verb*
>
> *to tend or round up (cattle, horses, or other livestock).*
>
> — dictionary.com

- So, by analogy, "wrangling data" means to collect, clean, and organize digital information (tend and round up)

- Informal word, but data scientists will understand what you mean if you use it

- In practice, data wrangling requires the use of data transformations to accomplish tasks related to processing, structuring, and analysis.

The word "wrangle"



Source: Digital image of a cowboy wrangling data, Digital image on *likelihoodlog.com*, accessed September 20, 2017, www.likelihoodlog.com/?p=1151

`ggplot2` needs clean/tidy datasets

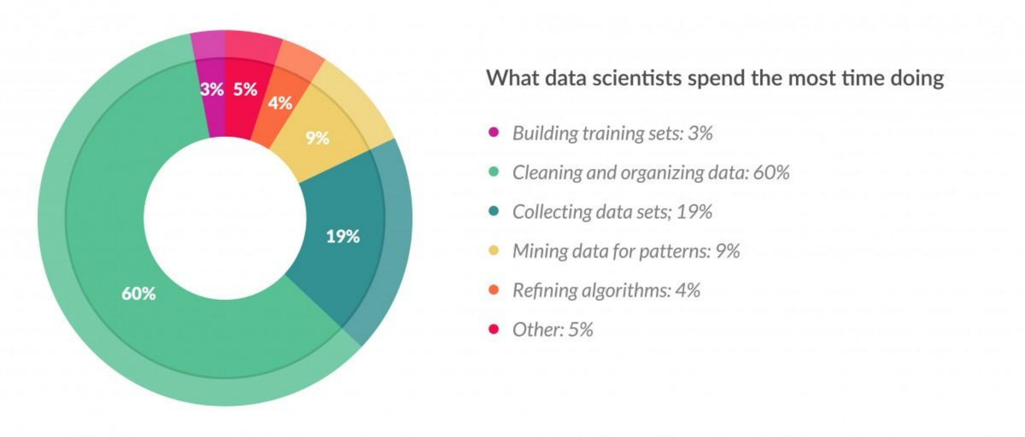
- Datasets such as the `mpg` dataset are small and nicely organized

`ggplot2` needs clean/tidy datasets

- Datasets such as the `mpg` dataset are small and nicely organized
- It would be nice if all datasets were like this! ...but they're the exceptions to the rule

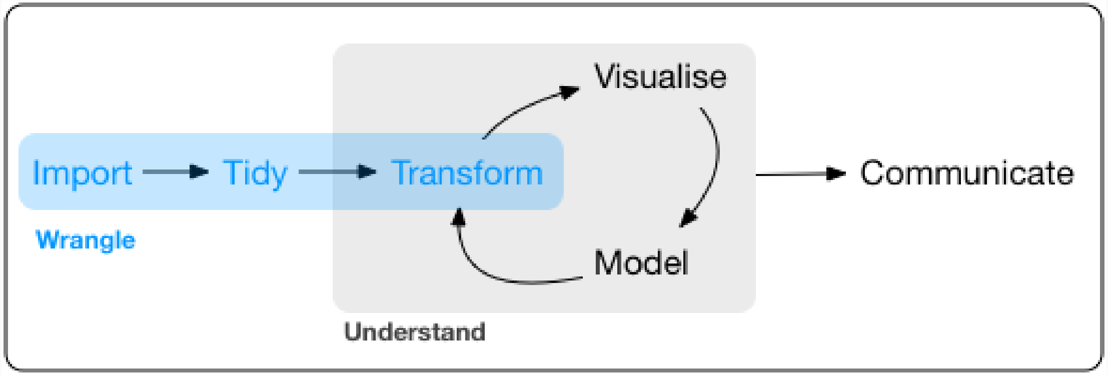
`ggplot2` needs clean/tidy datasets

- Datasets such as the `mpg` dataset are small and nicely organized
- It would be nice if all datasets were like this! ...but they're the exceptions to the rule
- Most raw datasets need cleaning, and this is where data scientists will spend **most** of their time



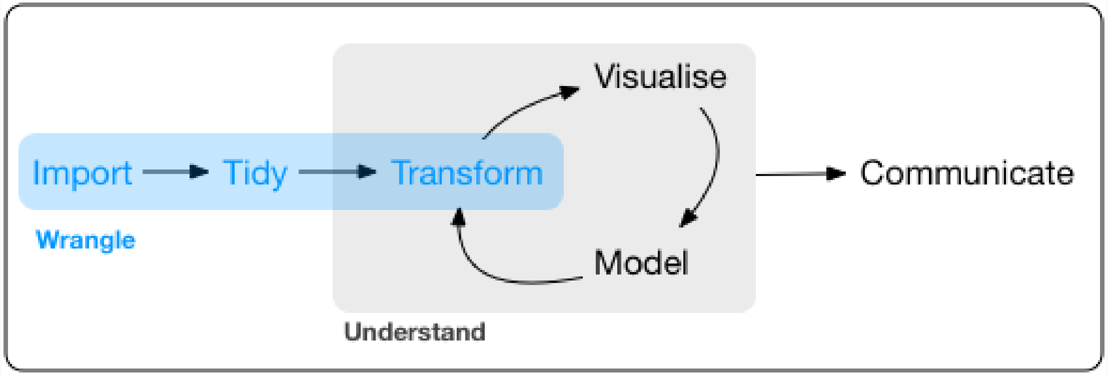
Source: Cleaning Big Data: Most Time-Consuming, Least Enjoyable Data Science Task, Survey Says, Digital image on *forbes.com*, accessed September 20, 2017, www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/

The "data wrangling" pipeline



Source: Data wrangling pipeline, Digital image on *r4ds.had.co.nz*, accessed September 20, 2017, r4ds.had.co.nz/wrangle-intro.html

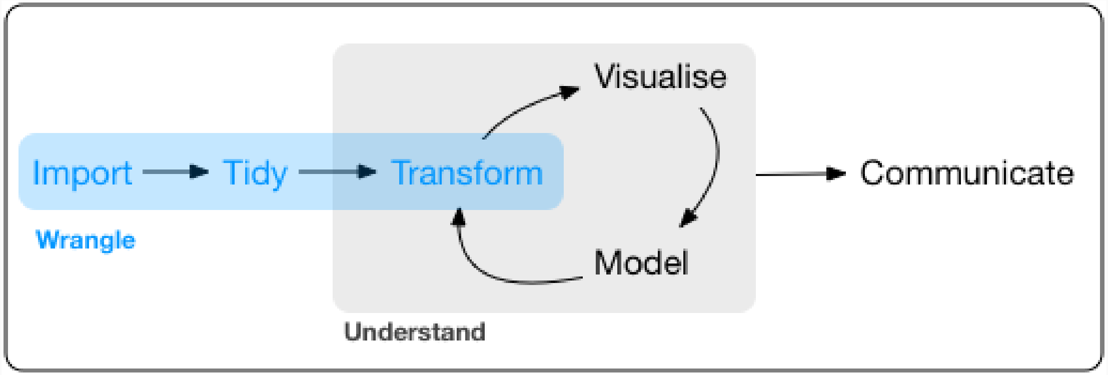
The "data wrangling" pipeline



- **import** → obtain data and get it into R

Source: Data wrangling pipeline, Digital image on *r4ds.had.co.nz*, accessed September 20, 2017, r4ds.had.co.nz/wrangle-intro.html

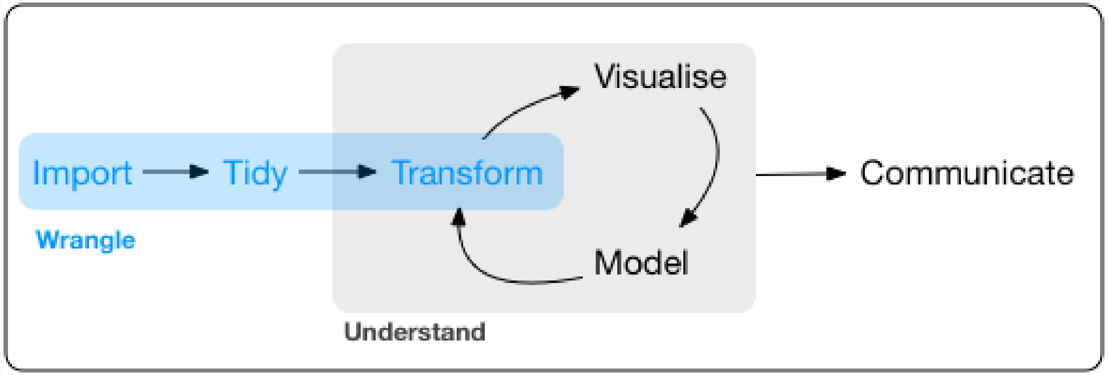
The "data wrangling" pipeline



- **import** → obtain data and get it into R
- **tidy** → reshape rows and columns to follow the Tidy data rules

Source: Data wrangling pipeline, Digital image on *r4ds.had.co.nz*, accessed September 20, 2017, r4ds.had.co.nz/wrangle-intro.html

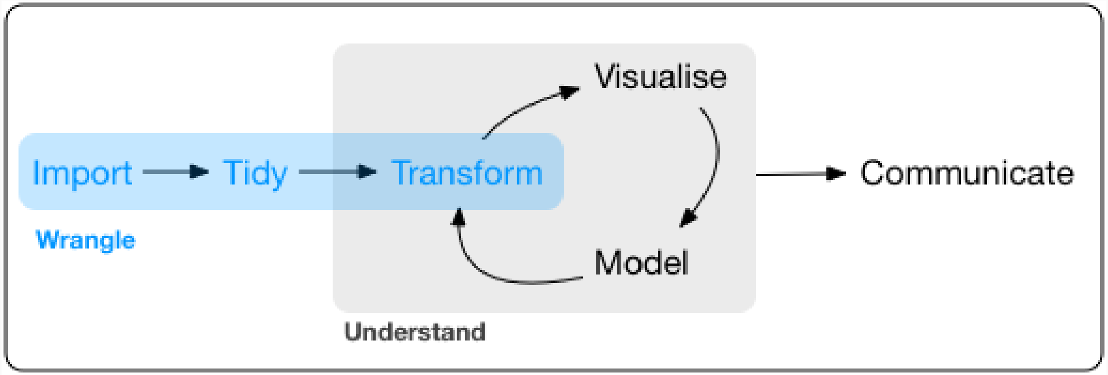
The "data wrangling" pipeline



- **import** → obtain data and get it into R
- **tidy** → reshape rows and columns to follow the Tidy data rules
- **transform** → cleaning the dataset (not the same as tidying) as well as "slicing and dicing" the dataset for exploration and analysis.

Source: Data wrangling pipeline, Digital image on *r4ds.had.co.nz*, accessed September 20, 2017, r4ds.had.co.nz/wrangle-intro.html

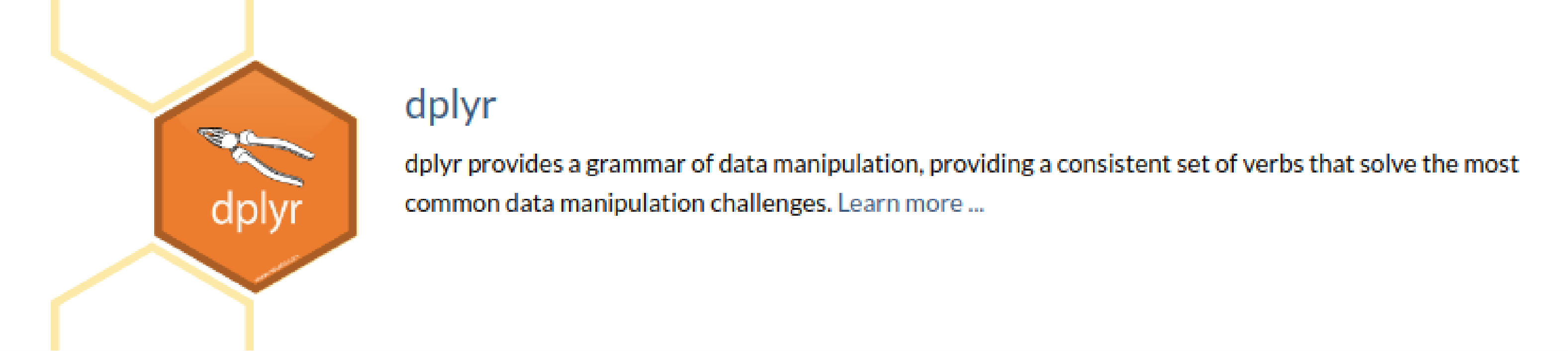
The "data wrangling" pipeline



- **import** → obtain data and get it into R
- **tidy** → reshape rows and columns to follow the Tidy data rules
- **transform** → cleaning the dataset (not the same as tidying) as well as "slicing and dicing" the dataset for exploration and analysis.

Source: Data wrangling pipeline, Digital image on *r4ds.had.co.nz*, accessed September 20, 2017, r4ds.had.co.nz/wrangle-intro.html

Meet the dplyr package



dplyr

dplyr provides a grammar of data manipulation, providing a consistent set of verbs that solve the most common data manipulation challenges. Learn more ...

Source: dplyr logo with blurb, Digital image on packages page of *tidyverse.org*, accessed on September 27, 2018, www.tidyverse.org/packages/

Credits

License