Section 001 CDS 101 *Introduction to Computational and Data Sciences*

Class 8: Data wrangling I

February 15, 2018

General

Annoucements

- Reading for next class: *R for Data Science* - chapters 4 (short) and 5
- Homework 1 posted on website, http://spring18.cds101.com, due Friday, February 23rd by 11:59pm
- RStudio cheatsheet resource
  - Will post cheatsheets on website soon

What is data wrangling?

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)

- Also encompasses the act of transforming data as a processing step to facilitate analysis

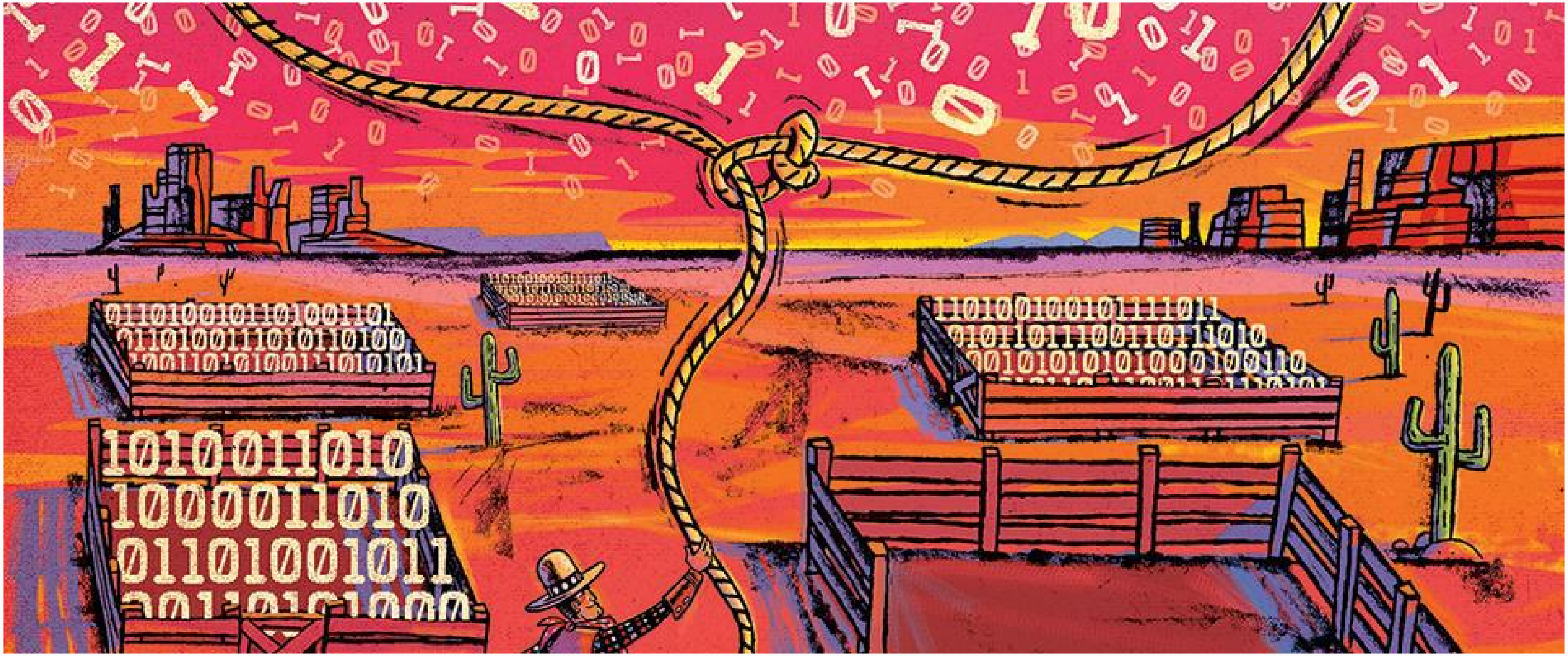
The word "wrangle"

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)
- Also encompasses the act of transforming data as a processing step to facilitate analysis
- Informal word, but data scientists will understand what you mean if you use it

The word "wrangle"



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

`ggplot2` needs clean/tidy datasets

- Datasets such as `mpg` or `rail_trail` (Assignment 1) are small and nicely organized

`ggplot2` needs clean/tidy datasets

- Datasets such as `mpg` or `rail_trail` (Assignment 1) are small and nicely organized

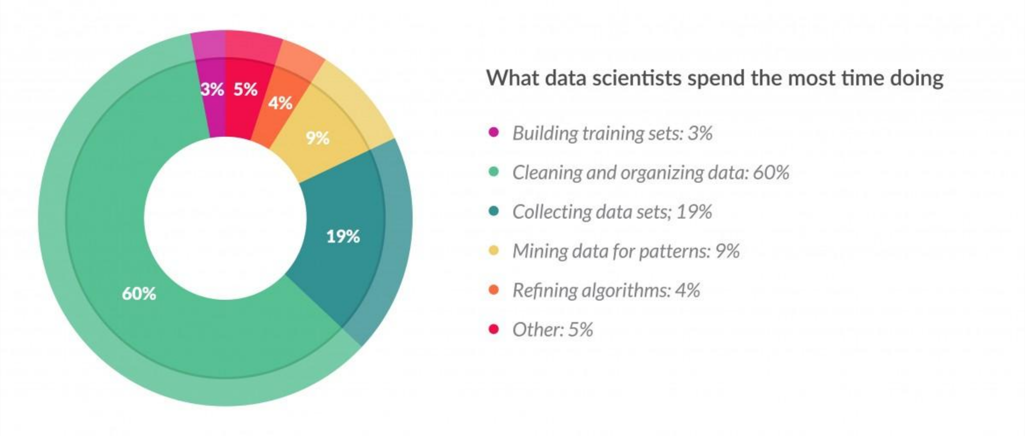
- It would be nice if all datasets were like this! ...but they're the exceptions to the rule

`ggplot2` needs clean/tidy datasets

- Datasets such as `mpg` or `rail_trail` (Assignment 1) are small and nicely organized

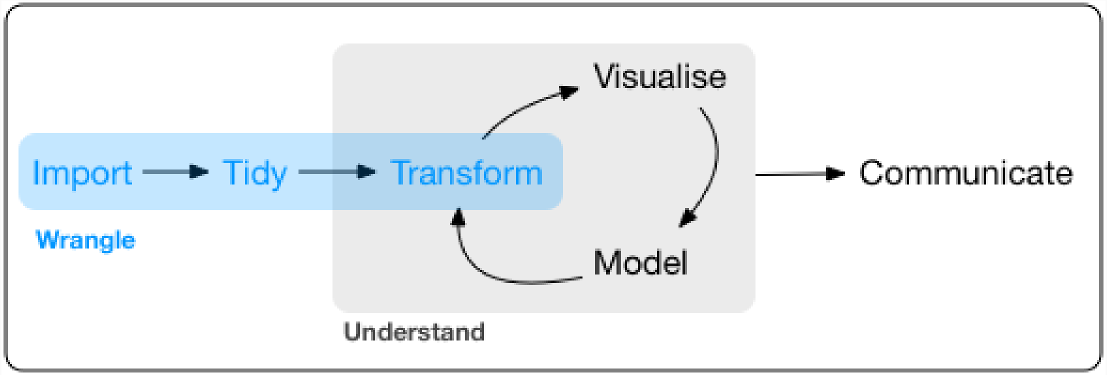
- It would be nice if all datasets were like this! ...but they're the exceptions to the rule

- Most raw datasets need cleaning, and this is where data scientists will spend **most** of their time

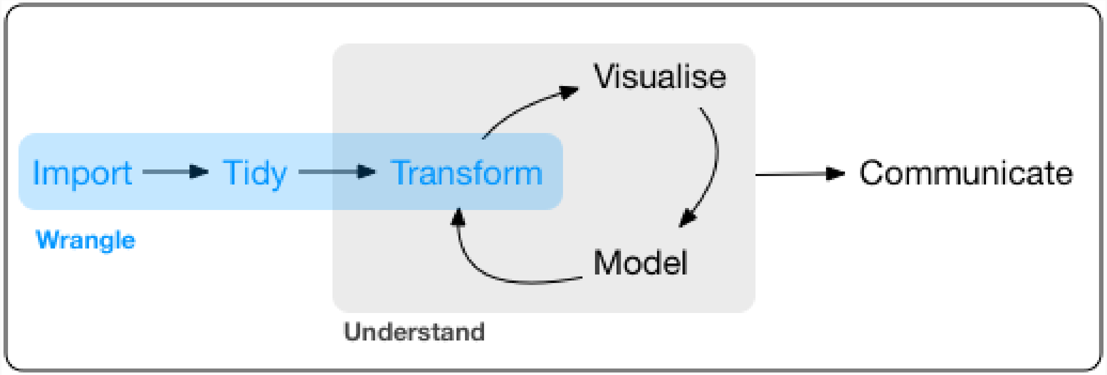


Source: Cleaning Big Data: Most Time-Consuming, Least Enjoyable Data Science Task, Survey Says, Digital image on *forbes.com*, accessed September 20, 2017, https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/

The "data wrangling" pipeline

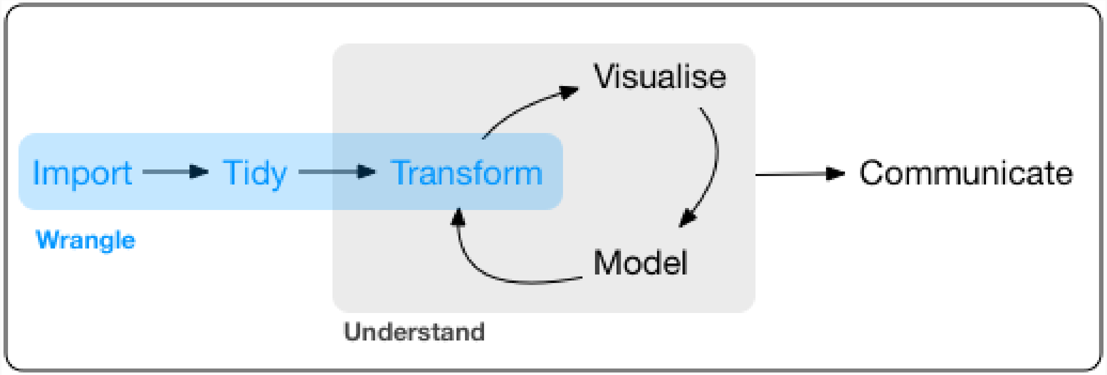


The "data wrangling" pipeline



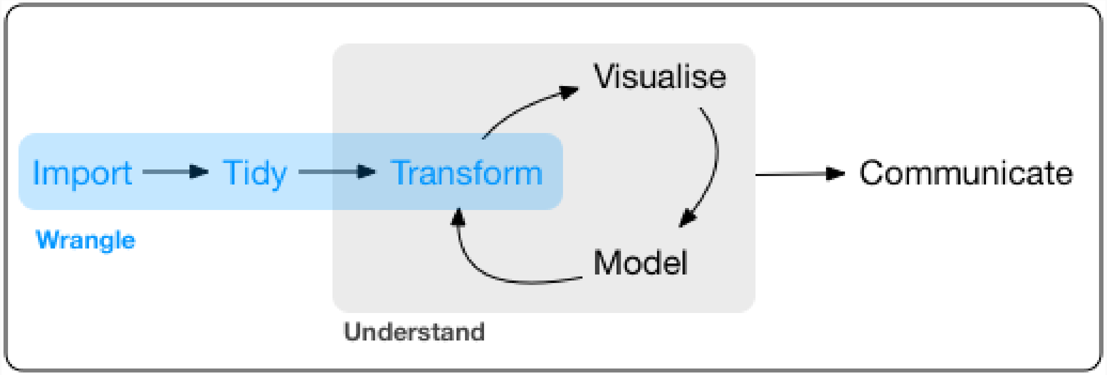
- **import** → obtain data and get it into R

The "data wrangling" pipeline



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

The "data wrangling" pipeline

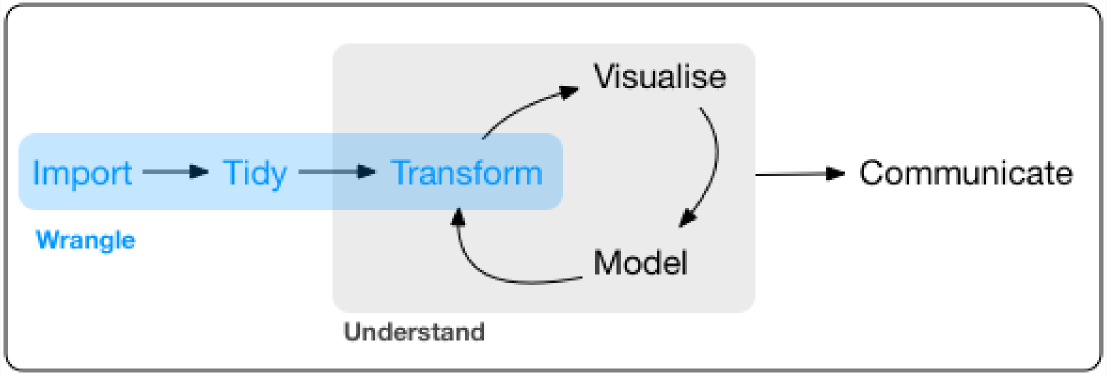


- **import** → obtain data and get it into R

- **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.

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.

Data wrangling in R

A few bits of R history

- The first stable version of R, v1.0.0, was released on February 29, 2000.
- R itself is an implementation of the S programming language, which was designed at Bell Laboratories in the mid-1970s.
- *Base R* was built for statisticians and for doing data analysis, but not necessarily for modern Data Science
- It's age and legacy brings along old implementations of data structures and abbreviated function (commands) names

Source: David Smith, *Over 16 years of R project history*, Revolutions blog, last updated on March 4, 2016, accessed September 20, 2017, http://blog.revolutionanalytics.com/2016/03/16-years-of-r-history.html

Modernizing R with `tidyverse`

- Over the last 3 years, chief scientist at RStudio, Hadley Wickham, has brought R into the modern era with the `tidyverse`.

> *The tidyverse is an opinionated collection of R packages designed for data science. All packages share an underlying philosophy and common APIs.*
>
> — Front page of the Tidyverse website

- In practice, this meant reducing everything to a small, core set of commands that all behave in a similar way.

Core `tidyverse`

- `ggplot2`: ggplot2 is a system for declaratively creating graphics, based on The Grammar of Graphics. You provide the data, tell ggplot2 how to map variables to aesthetics, what graphical primitives to use, and it takes care of the details.
- `dplyr`: dplyr provides a grammar of data manipulation, providing a consistent set of verbs that solve the most common data manipulation challenges.
- `tidyr`: tidyr provides a set of functions that help you get to tidy data. Tidy data is data with a consistent form: in brief, every variable goes in a column, and every column is a variable.

Source: Tidyverse packages, *tidyverse.com*, accessed on September 20, 2017, https://www.tidyverse.org/packages/

Core `tidyverse`

- `readr`: readr provides a fast and friendly way to read rectangular data (like csv, tsv, and fwf). It is designed to flexibly parse many types of data found in the wild, while still cleanly failing when data unexpectedly changes.

- `purrr`: purrr enhances R's functional programming (FP) toolkit by providing a complete and consistent set of tools for working with functions and vectors. Once you master the basic concepts, purrr allows you to replace many for loops with code that is easier to write and more expressive.

- `tibble`: tibble is a modern re-imaginging of the data frame, keeping what time has proven to be effective, and throwing out what it has not. Tibbles are data.frames that are lazy and surly: they do less and complain more forcing you to confront problems earlier, typically leading to cleaner, more expressive code.

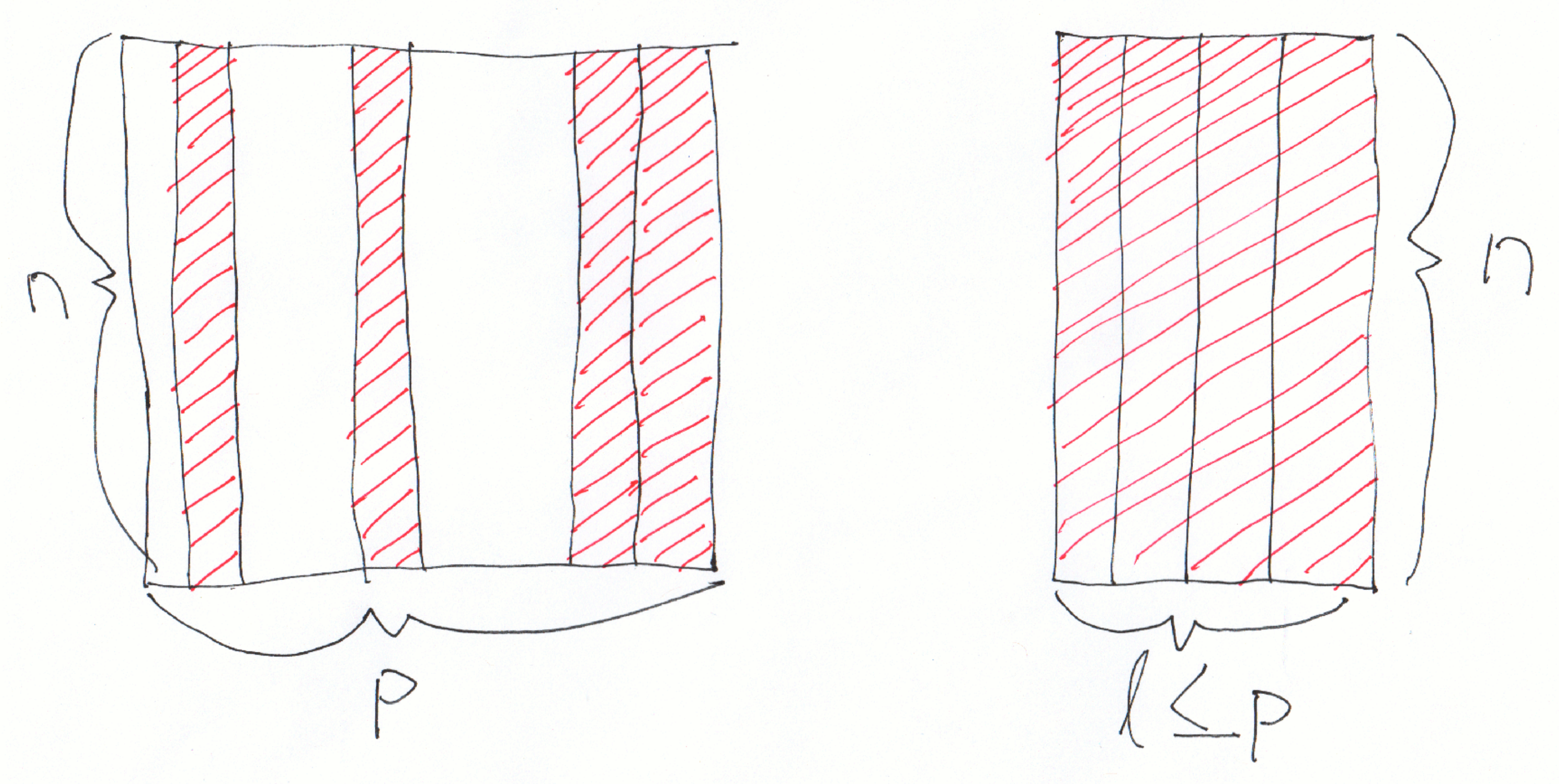
Source: Tidyverse packages, *tidyverse.com*, accessed on September 20, 2017, https://www.tidyverse.org/packages/

dplyr package

Get copy of `dplyr` demo repository

- On website, spring18.cds101.com, click Materials → Class 8
- Obtain a copy of the linked repository
- Load in RStudio, and follow along in demos

select()



`select()` demo

Follow along in RStudio

%>% aside

Instead of this:

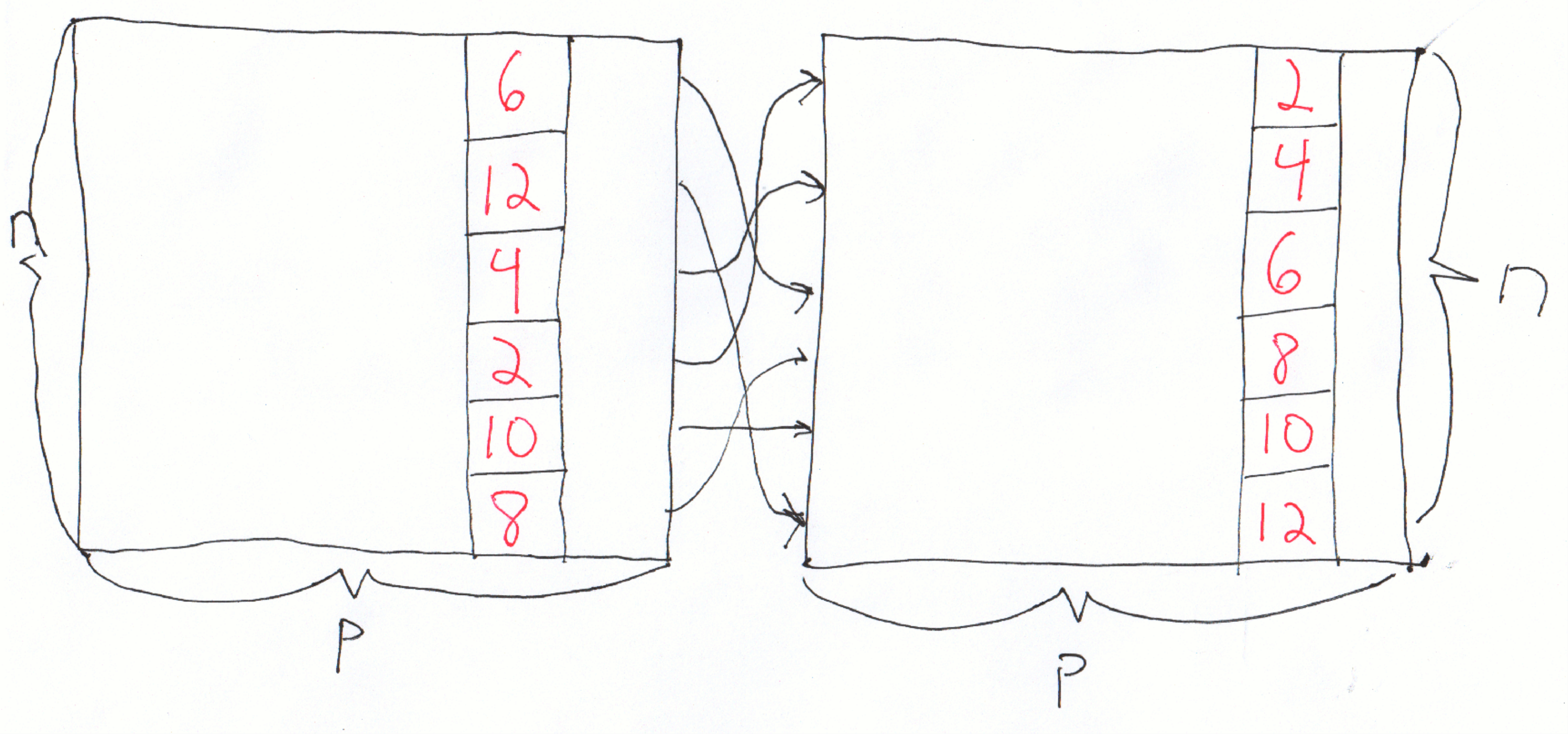
```
select(presidential, name, party)
```

We write this:

```
presidential %>%
  select(name, party)
```

- Show the order of transformations
- Useful when we have to chain together many transformations!

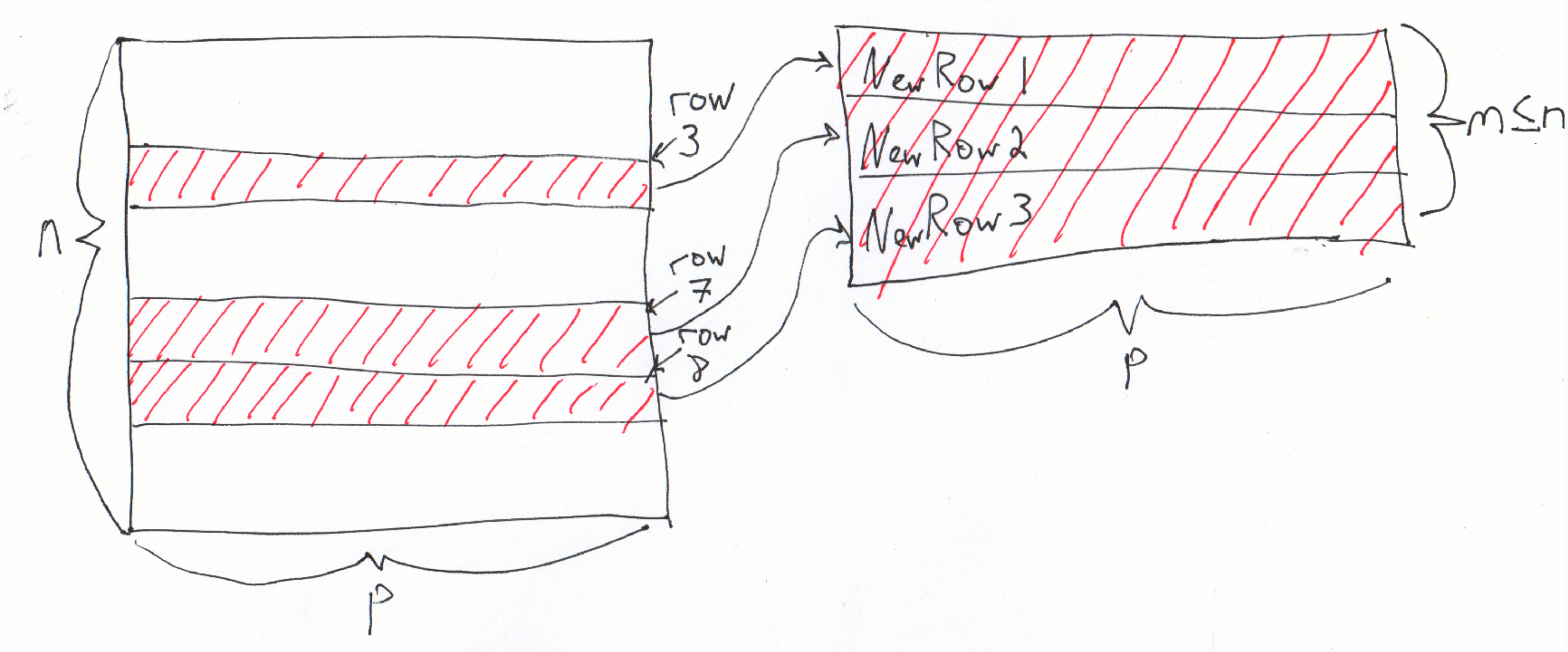
arrange()



`arrange()` demo

Follow along in RStudio

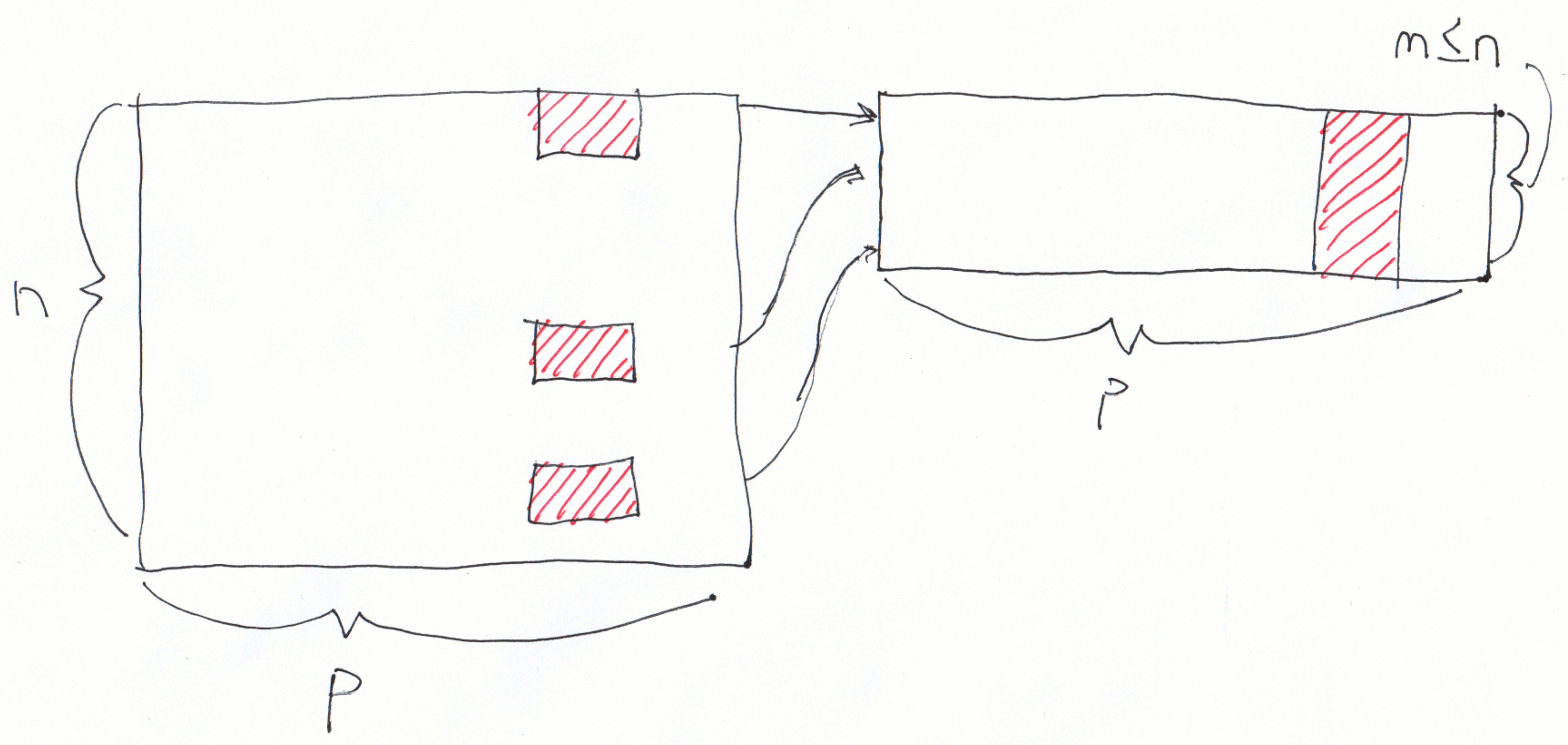
slice()



`slice()` demo

Follow along in RStudio

filter()

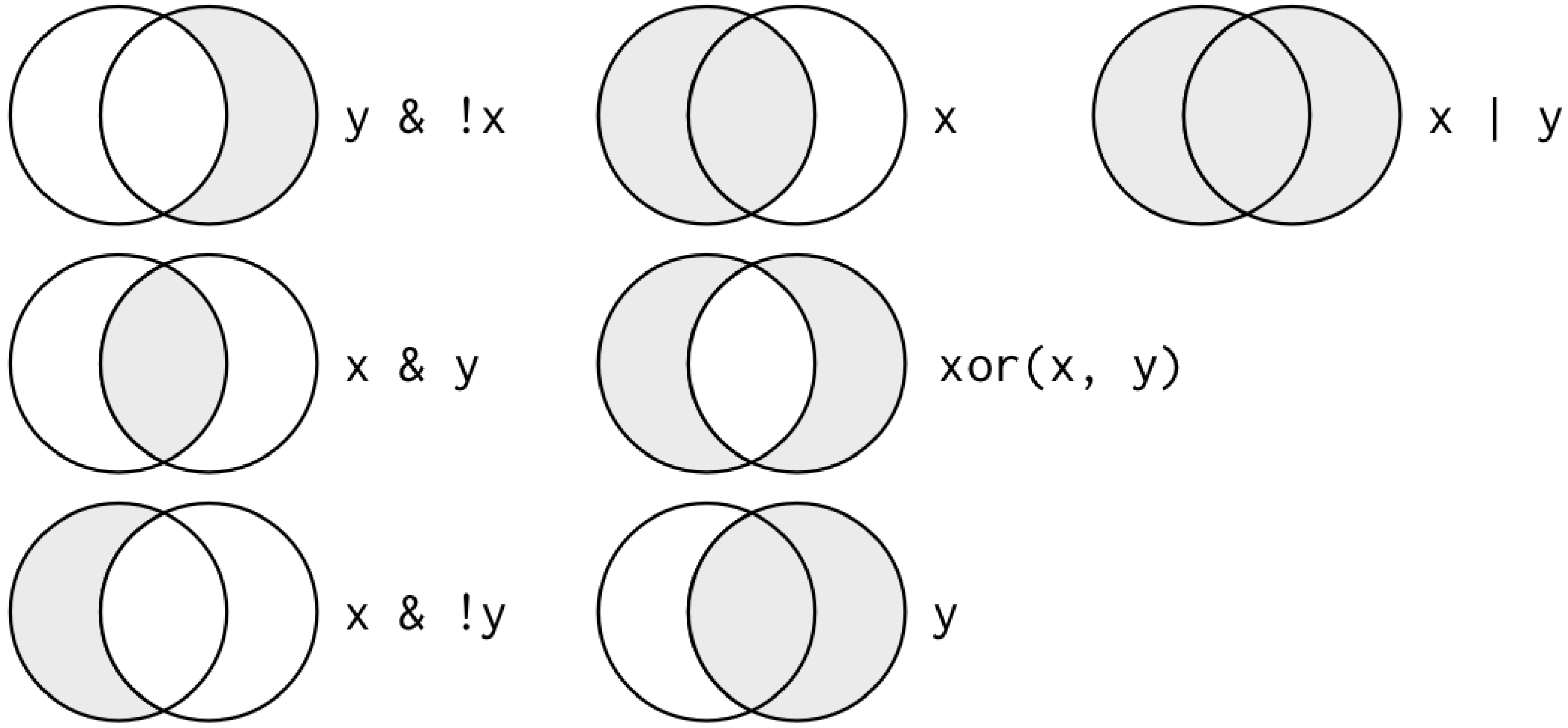


Comparisons

Simple comparisons can be made using the following symbols:

- `>` : greater than
- `>=` : greater than or equal to
- `<` : less than
- `<=` : less than or equal to
- `!=` : not equal
- `==` : equal

Logical operators



Source: Digital image of logical operations, *R for Data Science website*, accessed September 20, 2017, http://r4ds.had.co.nz/transform.html#logical-operators

`filter()` demo

Follow along in RStudio