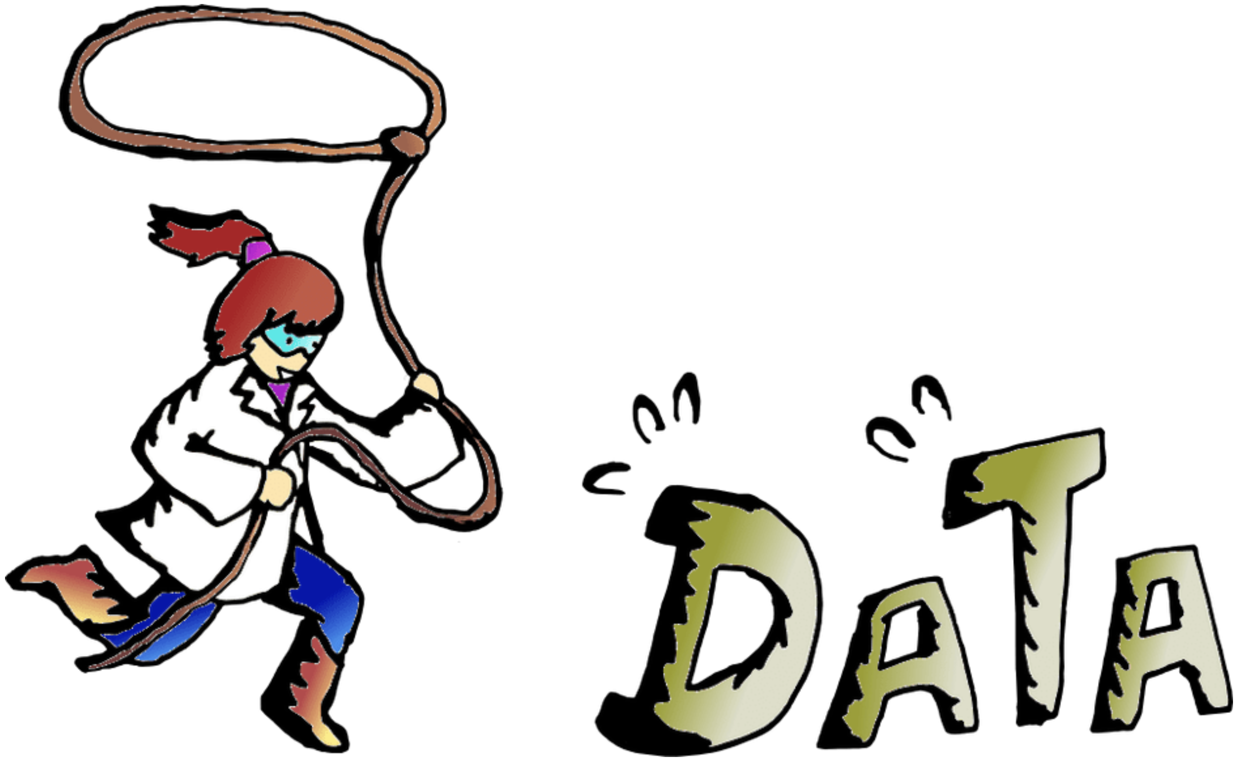
# Data wrangling/manipulation



# Overview

Data wrangling/manipulation with dplyr

Brief history of data visualization

# Announcements

A practice midterm exam will be posted by next class

Slides with answers will also be posted soon

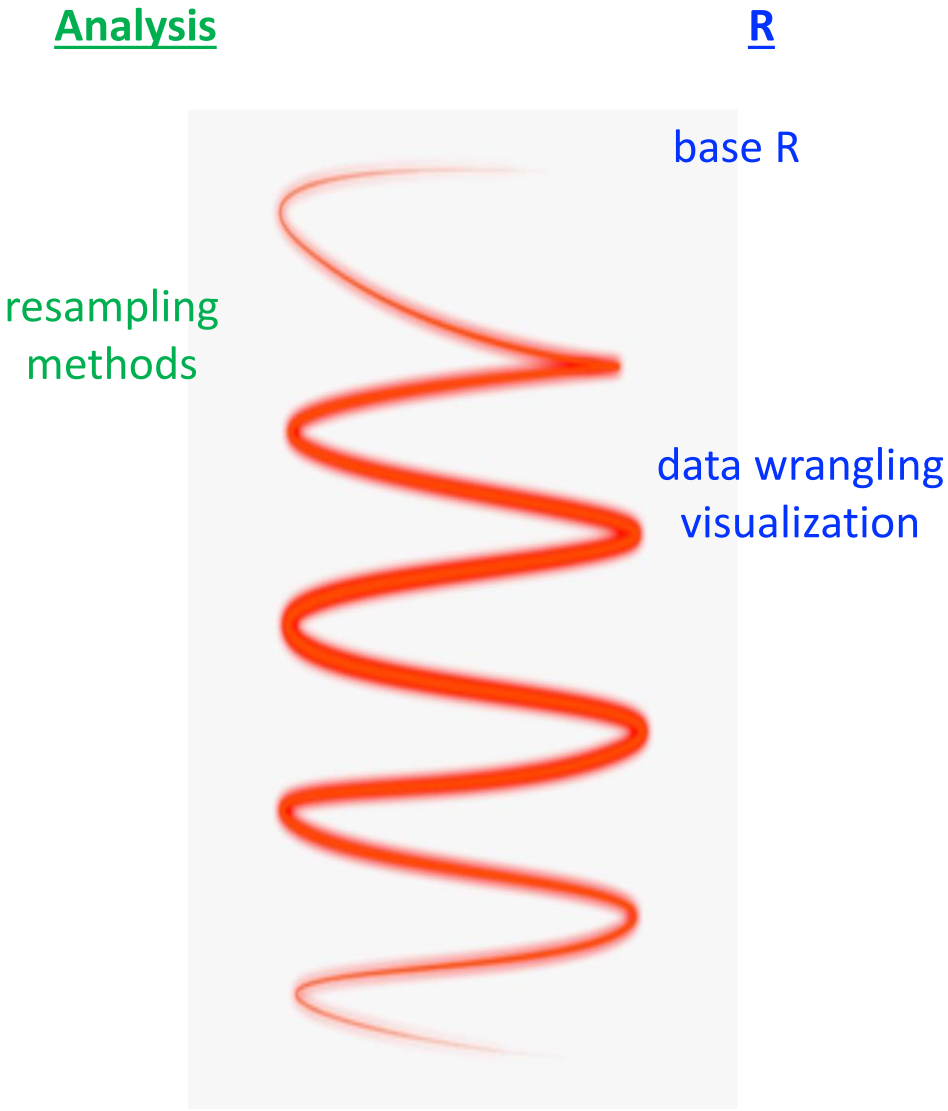
Get started on homework 5 early

- I recommend you do the dplyr exercises prior to next class!

Any other questions about class logistics?

# Plan for the semester

| | | |
|---|---|---|
| 1 | Sep 2 | Course overview, introduction to R, descriptive statistics |
| 2 | Sep 7-9 | Review of central statistical concepts and exploratory analysis using R |
| 3 | Sep 14-16 | Confidence Intervals and the bootstrap |
| 4 | Sep 21-23 | Review of hypothesis tests and permutation tests in R |
| 5 | Sep 28-30 | Parametric, non-parametric and theories of hypothesis testing |
| 6 | Oct 5-7 | Data manipulation and visualization |
| 7 | Oct 12-14 | Mapping, review and midterm exam |
| 8 | Oct 22 | October break |



**Analysis**

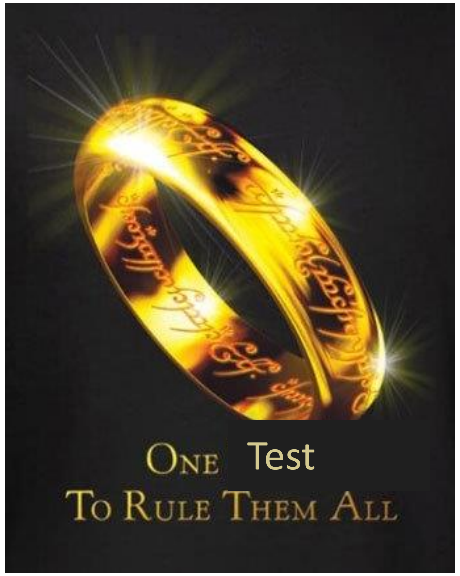
**R**

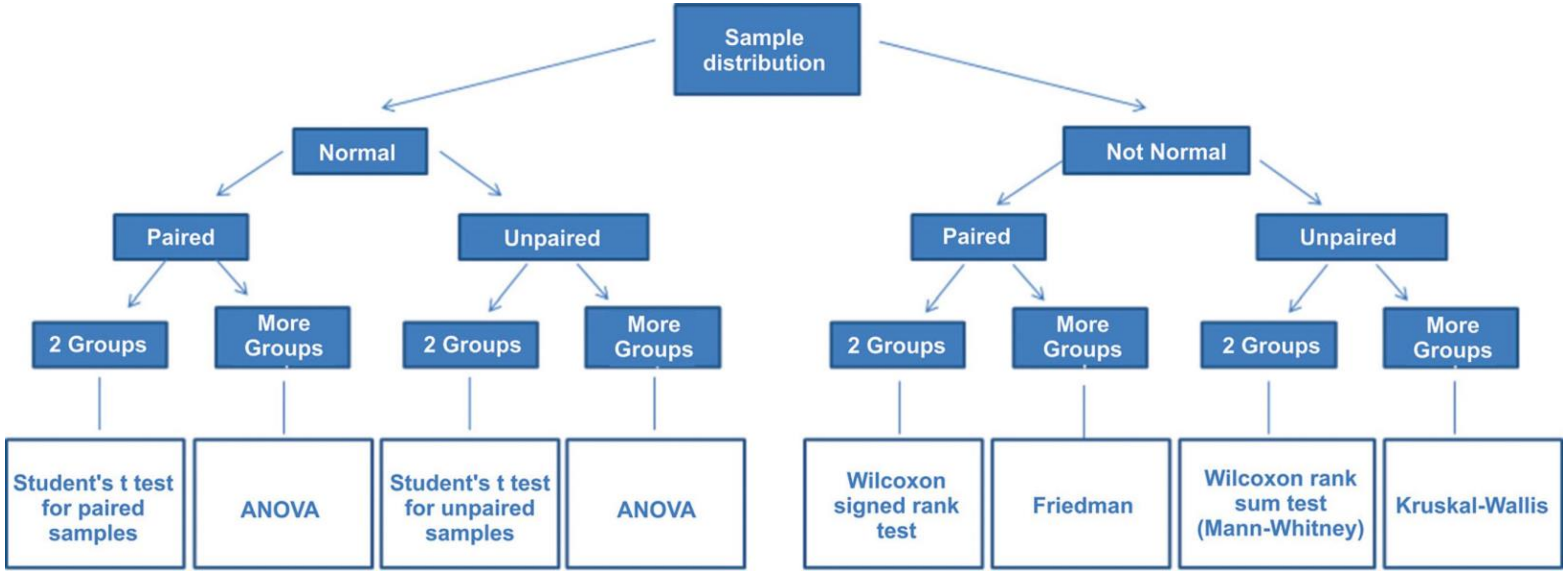
base R

resampling methods

data wrangling visualization

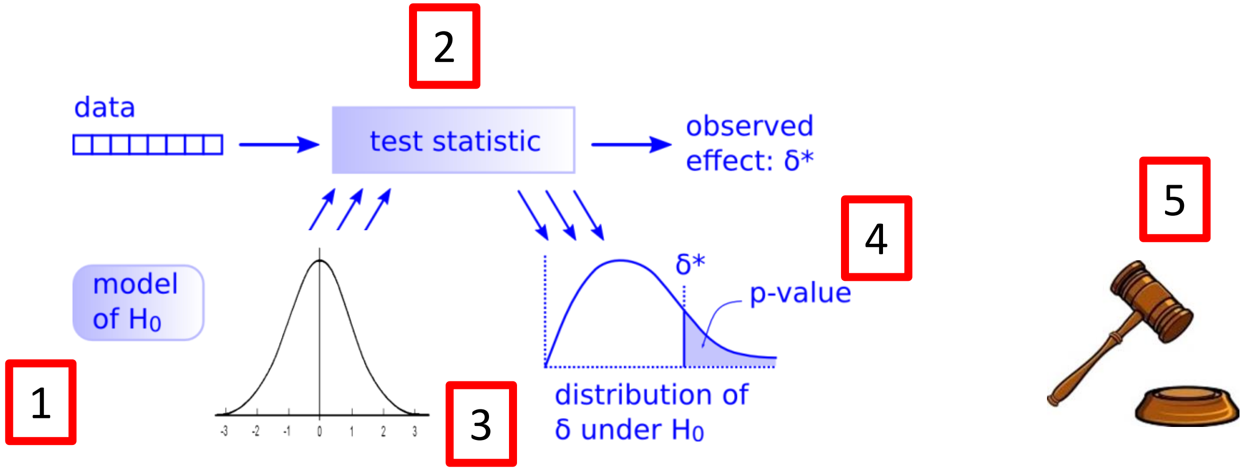
# Very quick review



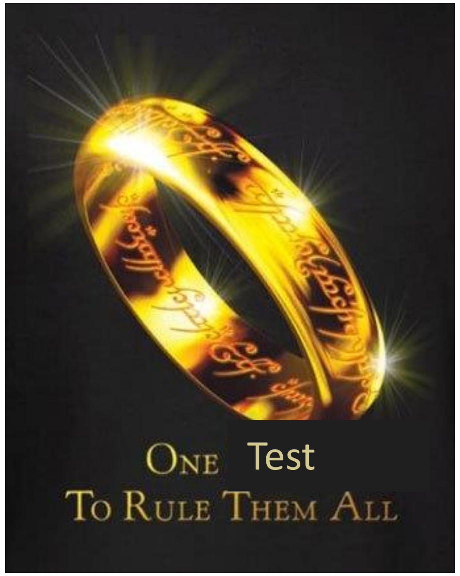


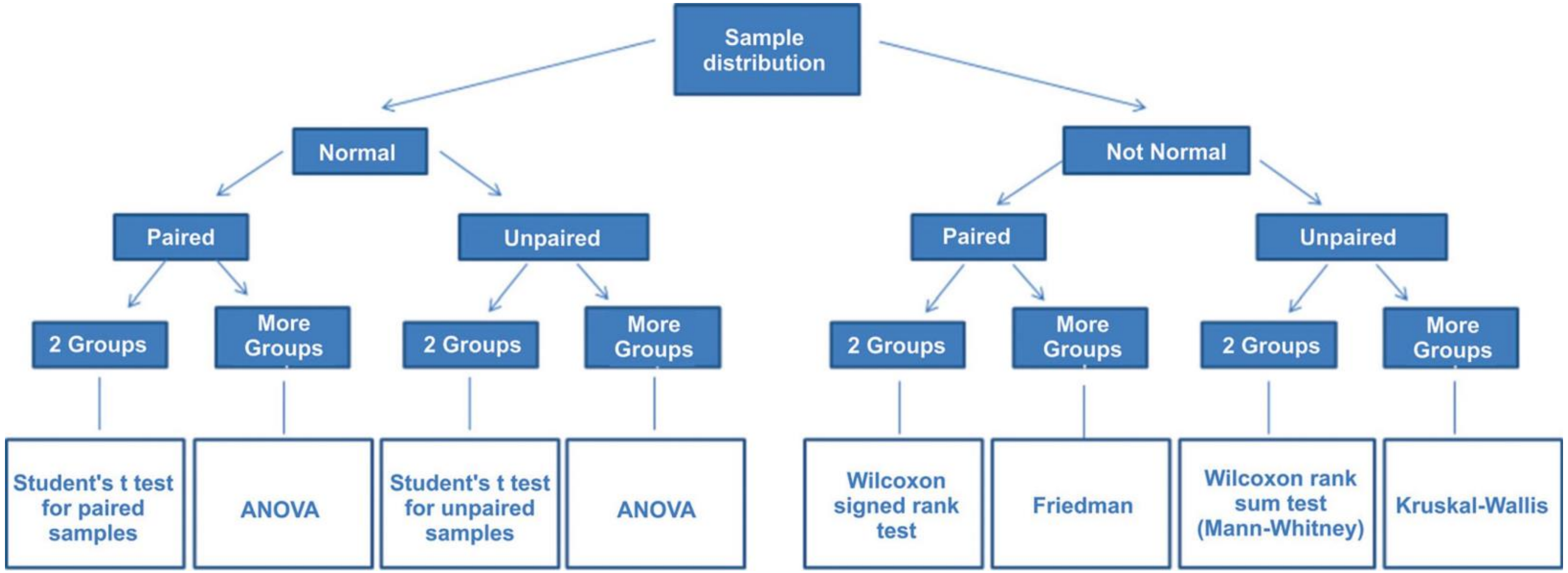
- Sample distribution
  - Normal
    - Paired
      - 2 Groups
        - Student's t test for paired samples
      - More Groups
        - ANOVA
    - Unpaired
      - 2 Groups
        - Student's t test for unpaired samples
      - More Groups
        - ANOVA
  - Not Normal
    - Paired
      - 2 Groups
        - Wilcoxon signed rank test
      - More Groups
        - Friedman
    - Unpaired
      - 2 Groups
        - Wilcoxon rank sum test (Mann-Whitney)
      - More Groups
        - Kruskal-Wallis

Just need to follow 5 steps!



# Very quick review





To select the appropriate parametric test, focus on the parameters being tested in the null hypothesis

- E.g., $H_0: \pi = 0.5$ $H_0: \mu = 0.5$ $H_0: \mu_T = \mu_C$ $H_0: \mu_1 = \mu_2 = \ldots = \mu_k$

Parametric tests are derived from particular mathematical assumptions

- E.g., data from the two samples comes from normal populations with the same variance
- Some hypothesis tests are "robust" to violations of these assumptions
  - The robustness can be evaluated this through computer simulations

# Very quick review: theories of hypothesis testing

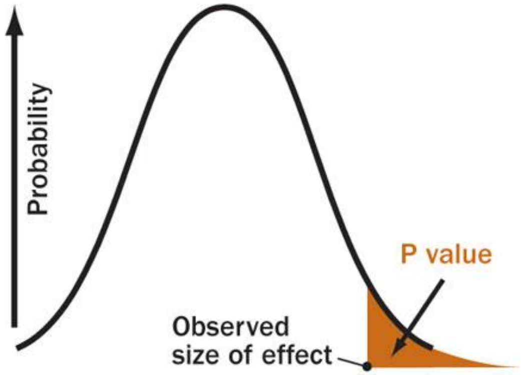
Fisher (1890-1962)

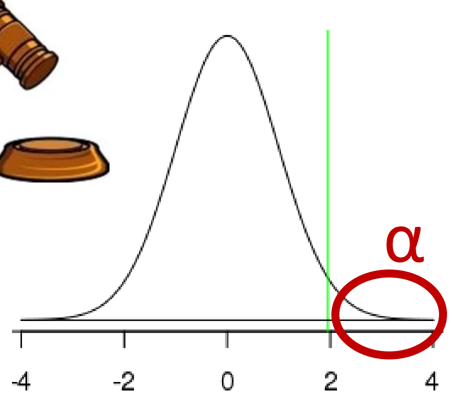
Neyman (1894-1981)

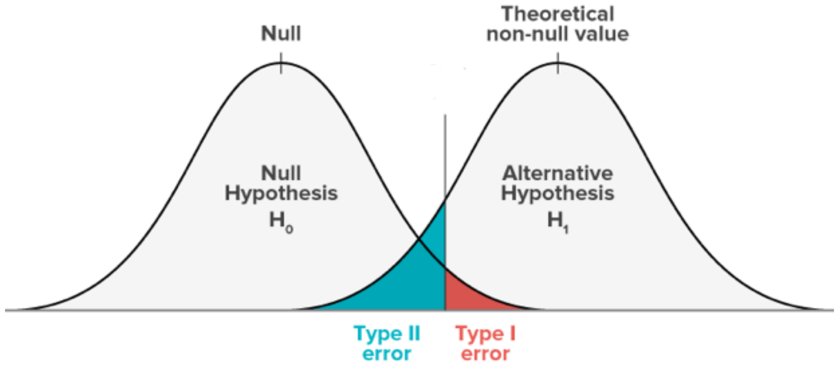
Pearson (1895-1980)

p-value a strength of evidence

Use p-value to make a decision







# Questions?

# The tidyverse and dplyr

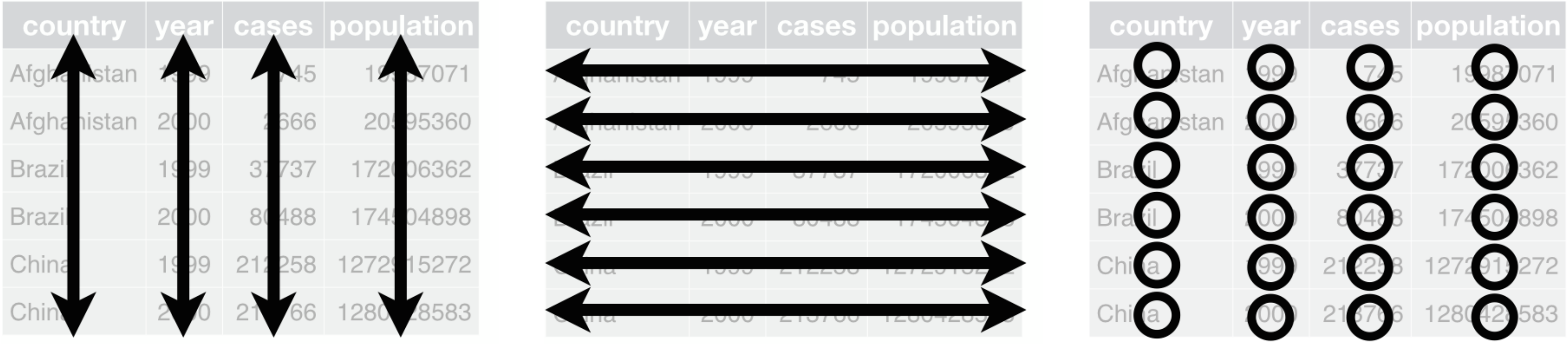
# The ‘tidyverse’

The tidyverse is set of R packages that operate ‘tidy data’

- i.e., that operate on data frames (or tibbles)

Tidy data is data where:

- Each variable must have its own column
- Each observation must have its own row
- Each value must have its own cell



variables

observations

values

# Messy data…

What would be an example of data that is not tidy?

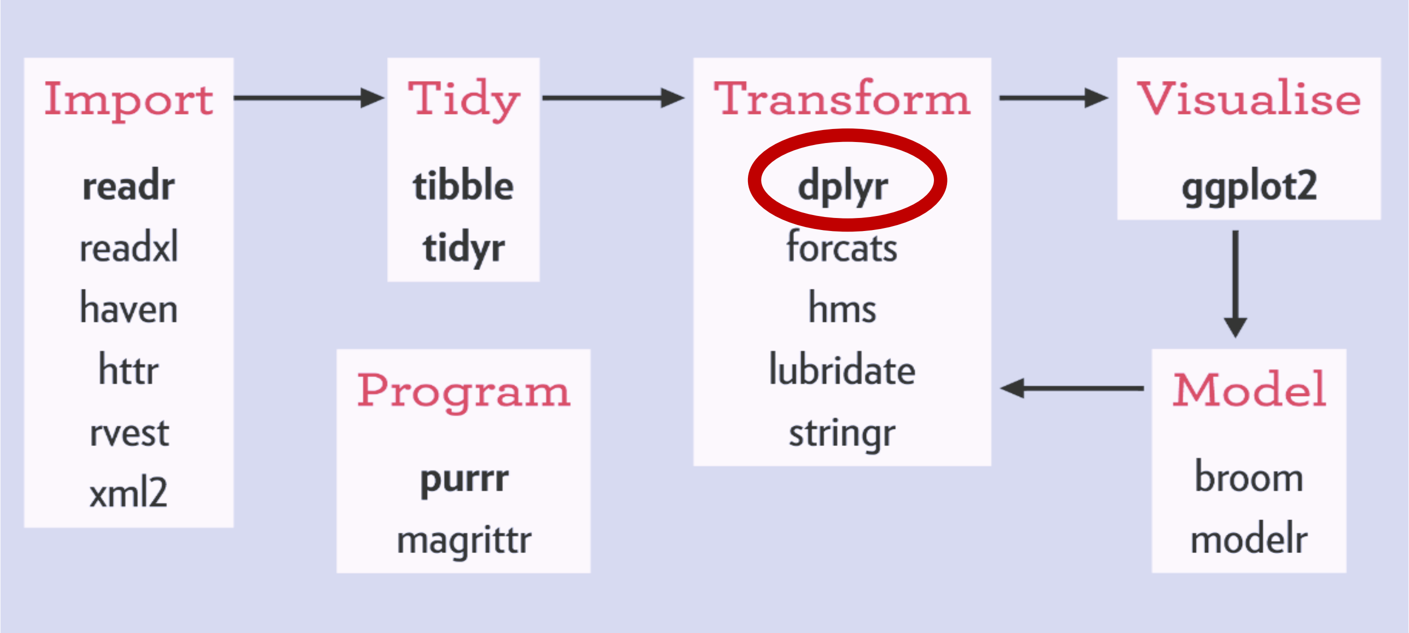
# Messy data…

“Happy families are all alike; every unhappy family is unhappy in its own way.” –
– Leo Tolstoy

# The ‘tidyverse’

The packages share a common design philosophy

- Most written by Hadley Wickham



# dplyr:  A grammar for data wrangling

**Grammar**:  a set of components that can be combined to achieve a goal

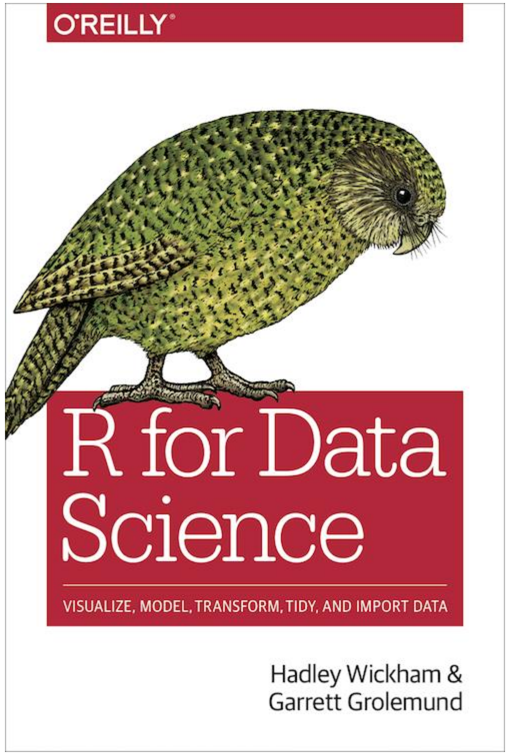
**dplyr** is a package that has a set of verbs that are useful for transformations data:

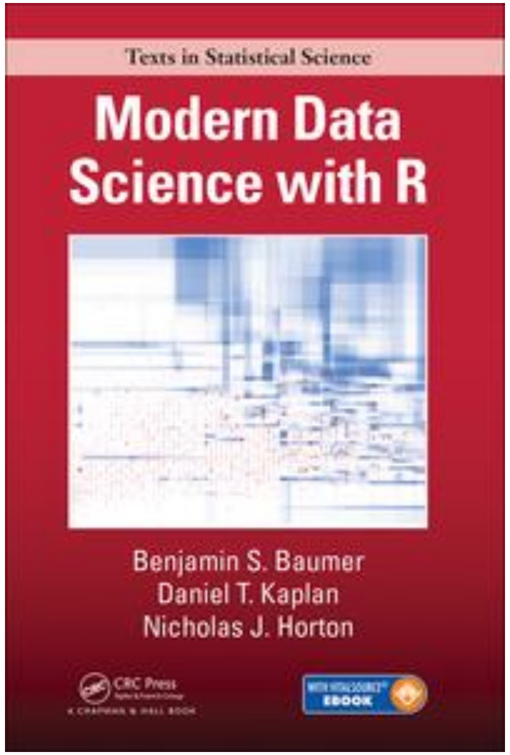
1. filter()
2. select()
3. mutate()
4. arrange()
5. summarize()
6. group_by()

All these function **take a data frame** and other arguments and **return a data frame**

```
> library(dplyr)   # load the dplyr package
```

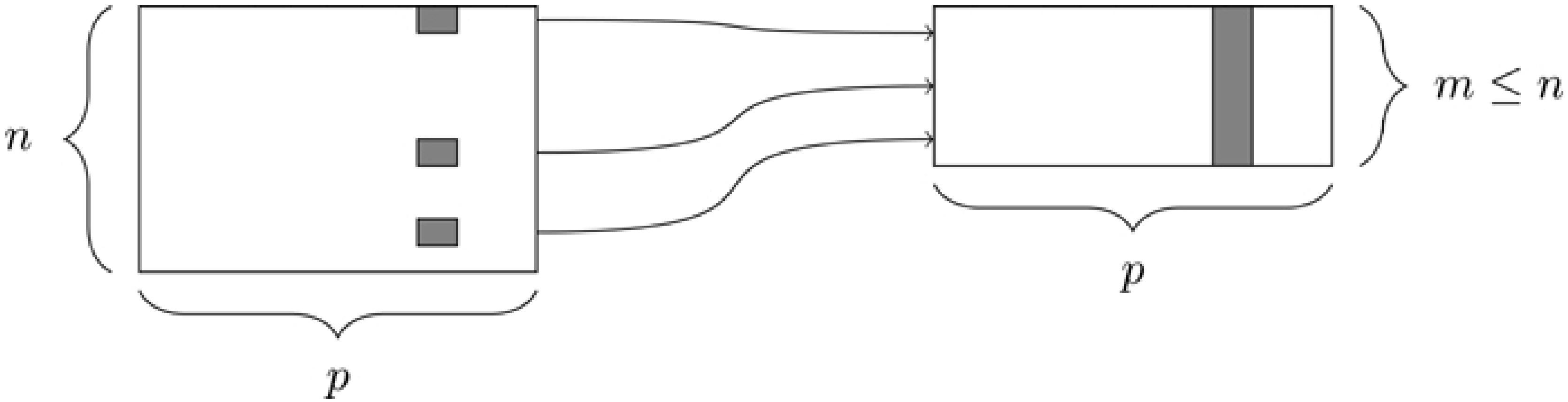
# Quick overview of the dplyr functions





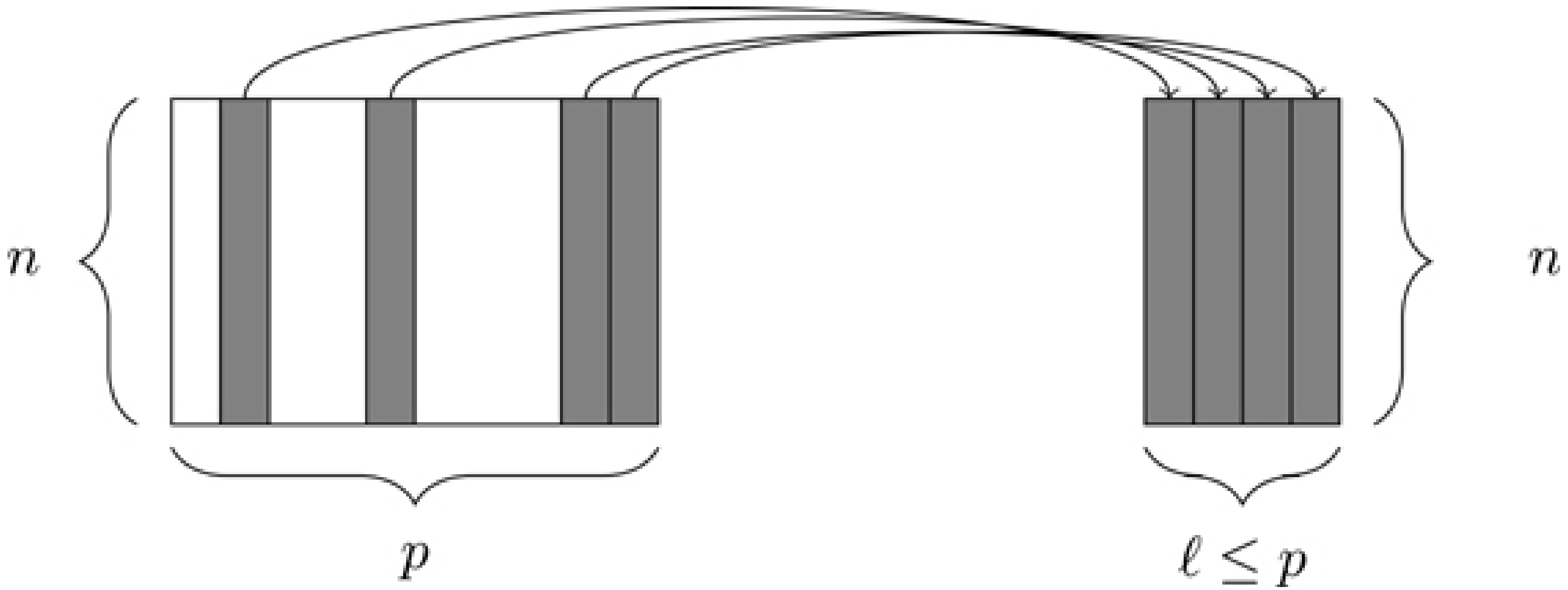
# 1. filter()

The filter() function allows you to select a subset of rows in data frame



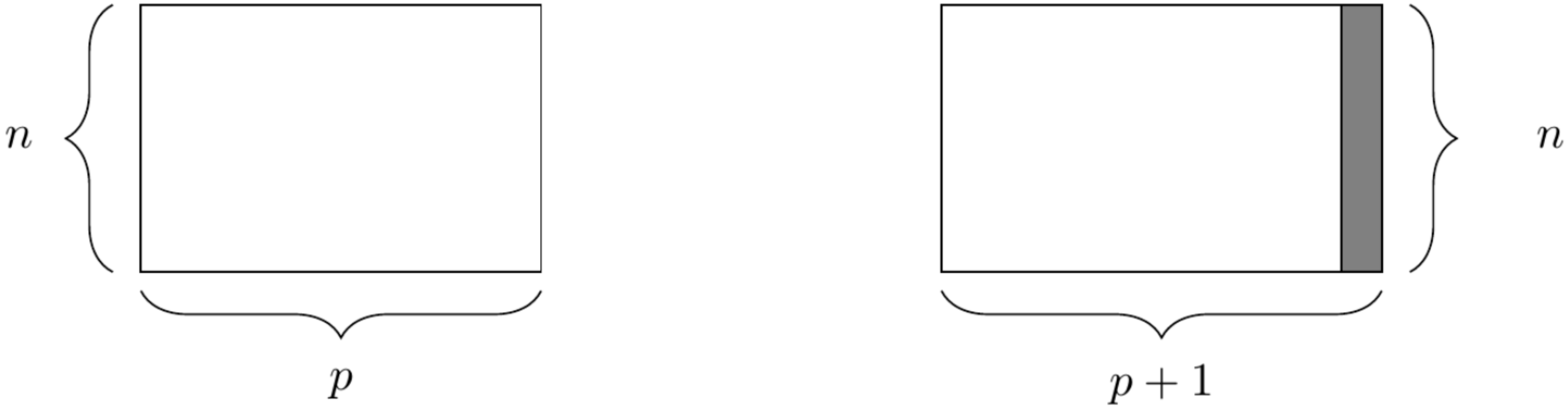
# 2. select()

The select() function allows you to select a subset of columns



# 3. mutate()

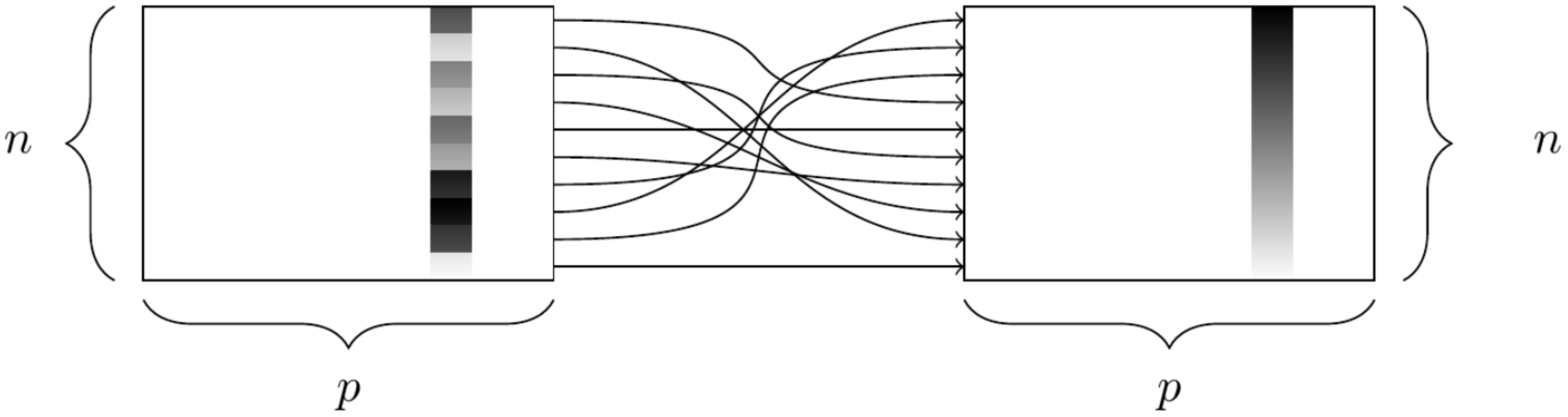
The mutate() function allows you to create new columns that are functions of existing columns



# 4. arrange()

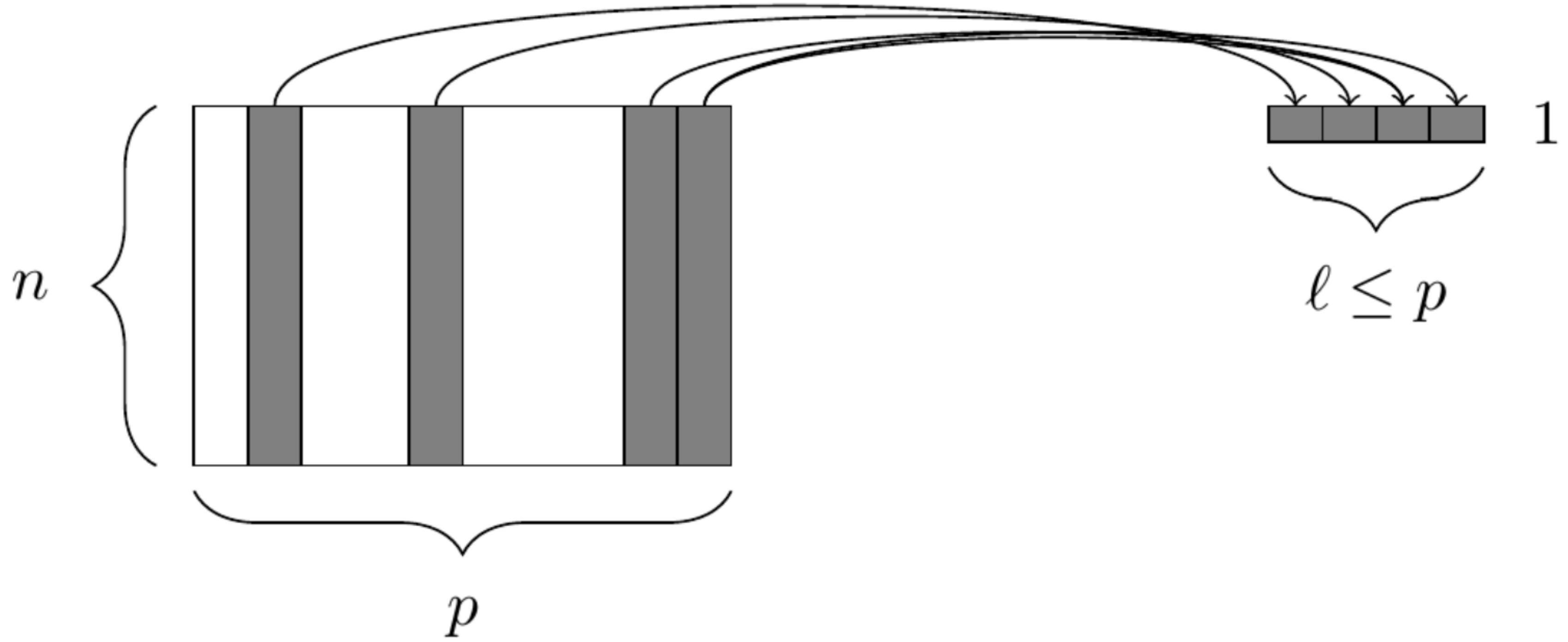
The arrange() function arranges the rows based values in a column

- arrange(dec()) arranges from largest to smallest



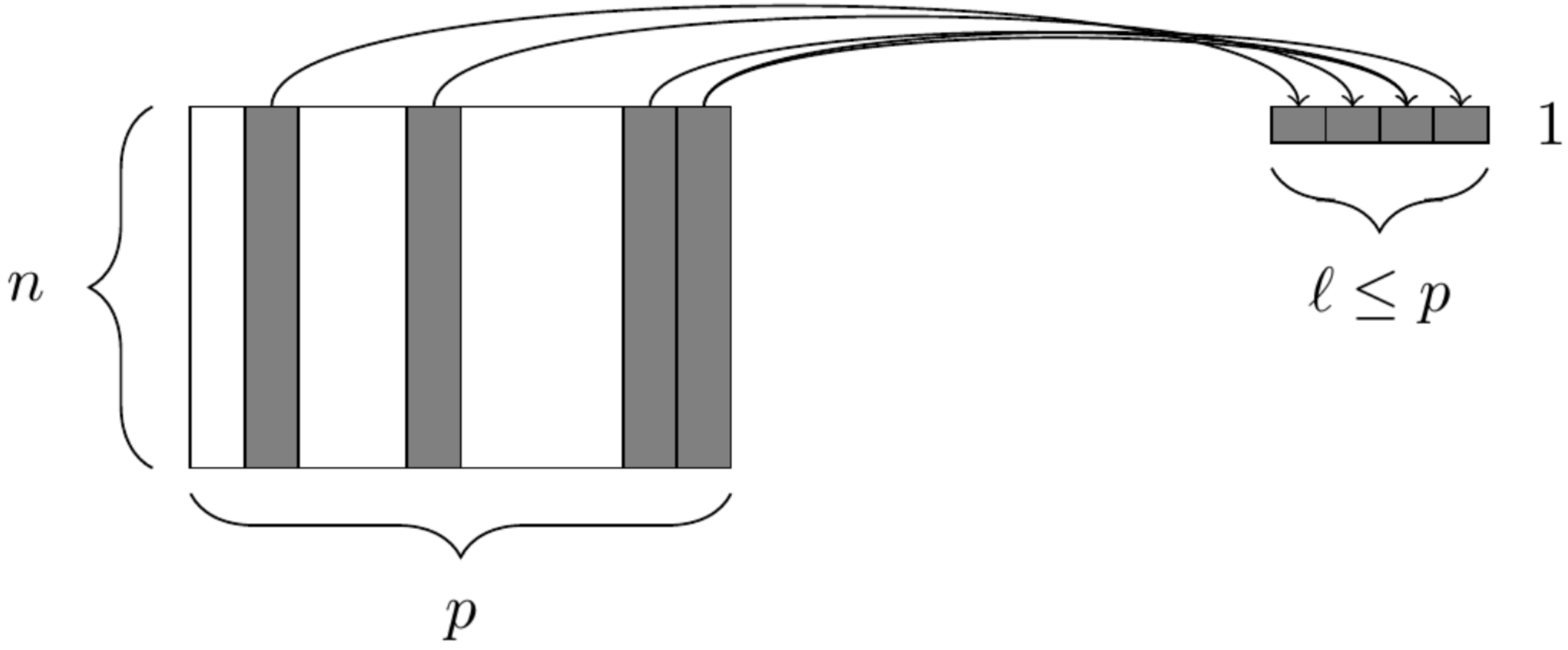
# 5. summarize()

The summarize() function reduces values in many rows into single values



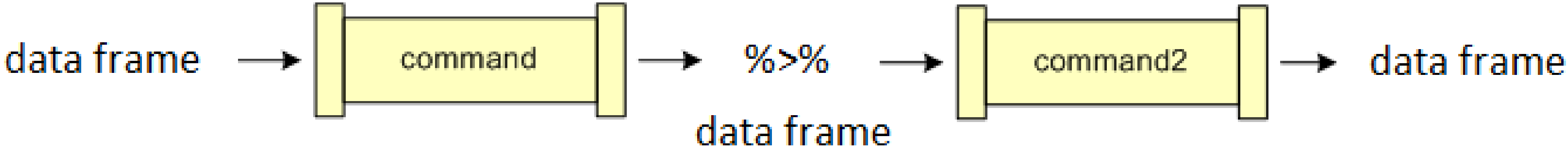
# 6. The group_by() function

The group_by() function groups variables for future operations



# The pipe operator

The pipe operator %>% allows us to chain commands together



data frame → command → %>% data frame → command2 → data frame

# Let’s try it out!

# A very brief history of data visualization

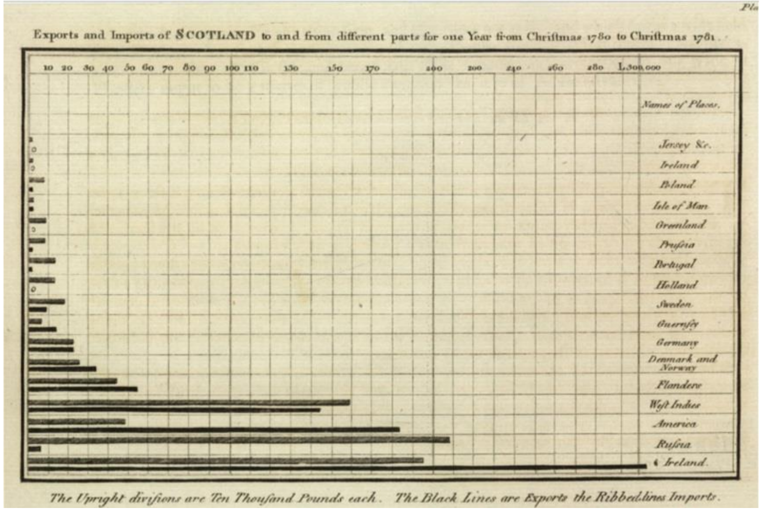
# Data visualization

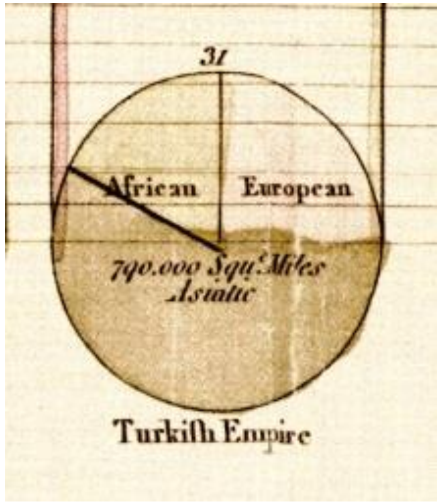
What are some reasons we visualize data rather than just reporting statistics?

# A very brief history of data visualization

The age of modern statistical graphs began around the beginning of the 19th century

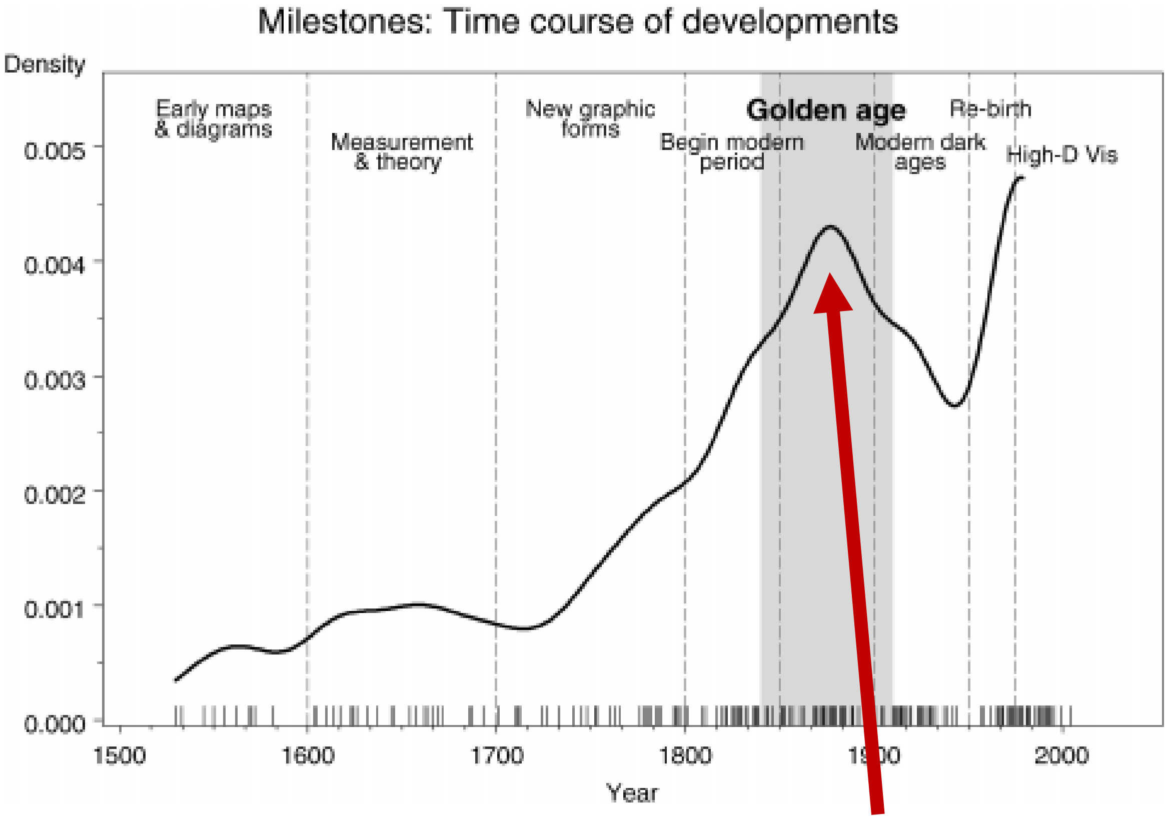
William Playfair (1759-1823) credited with inventing the line graph, bar chart and pie chart





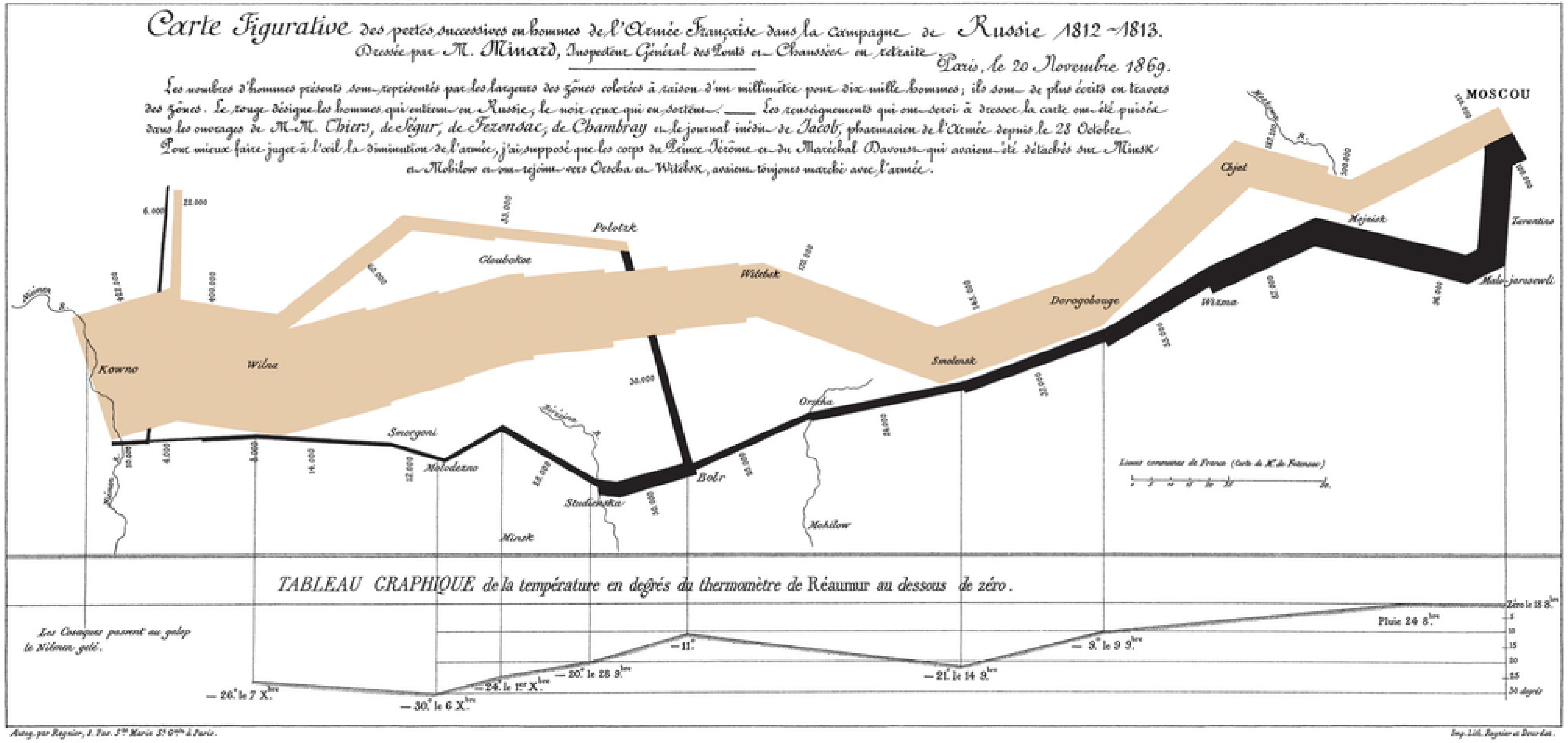
# A very brief history of data visualization

According to Friendly, statistical graphics researched its golden age between 1850-1900



# A very brief history of data visualization

Joseph Minard (1781-1870)



Carte Figurative des pertes successives en hommes de l'Armée Française dans la campagne de Russie 1812-1813.
Dressée par M. Minard, Inspecteur Général des Ponts et Chaussées en retraite.
Paris, le 20 Novembre 1869.

Les nombres d'hommes présents sont représentés par les largeurs des zones colorées à raison d'un millimètre pour dix mille hommes; ils sont de plus écrits en travers des zones. Le rouge désigne les hommes qui entrent en Russie, le noir ceux qui en sortent. —— Les renseignements qui ont servi à dresser la carte ont été puisés dans les ouvrages de M.M. Thiers, de Ségur, de Fezensac, de Chambray et le journal inédit de Jacob, pharmacien de l'Armée depuis le 28 Octobre.
Pour mieux faire juger à l'œil la diminution de l'armée, j'ai supposé que les corps du Prince Jérôme et du Maréchal Davoust qui avaient été détachés sur Minsk et Mohilow et ont rejoint vers Orscha et Witebsk, avaient toujours marché avec l'armée.

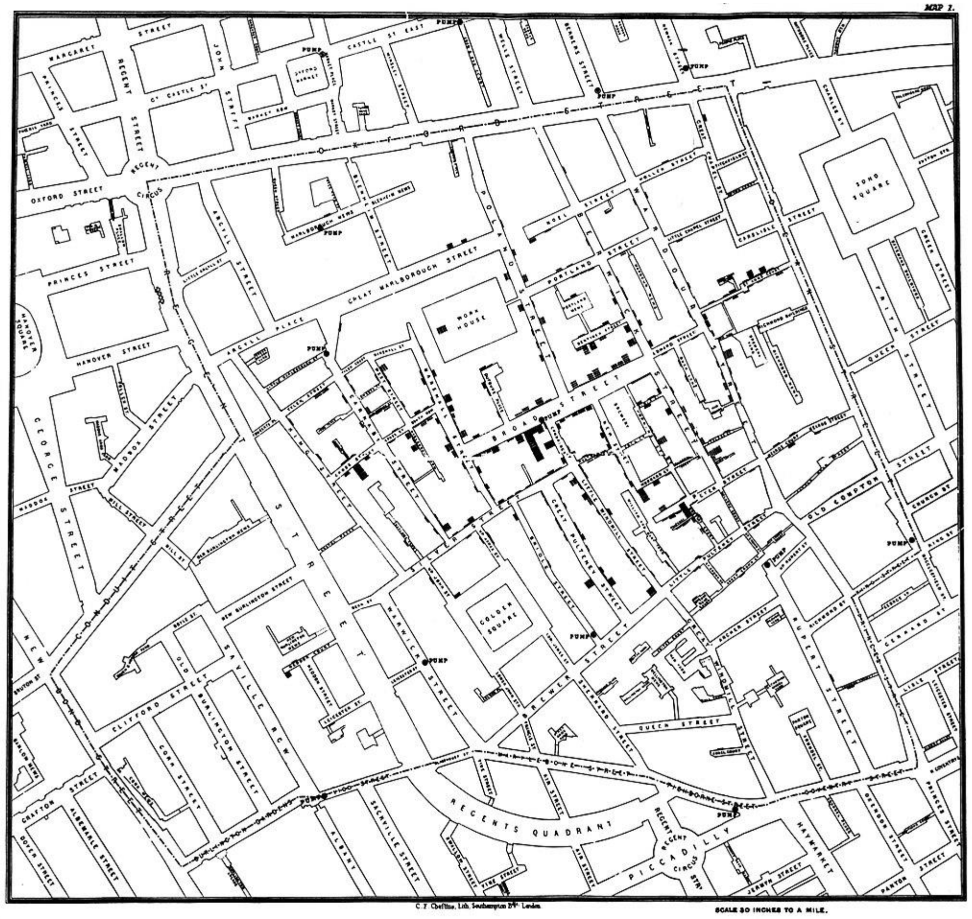
TABLEAU GRAPHIQUE de la température en degrés du thermomètre de Réaumur au dessous de zéro.

Les Cosaques passent au galop le Niémen gelé.

Map of Napoleon's march on Russia

# A very brief history of data visualization

John Snow (1813-1858)



Clusters of cholera cases in London epidemic of 1854

# A very brief history of data visualization

Florence Nightingale (1820-1910)

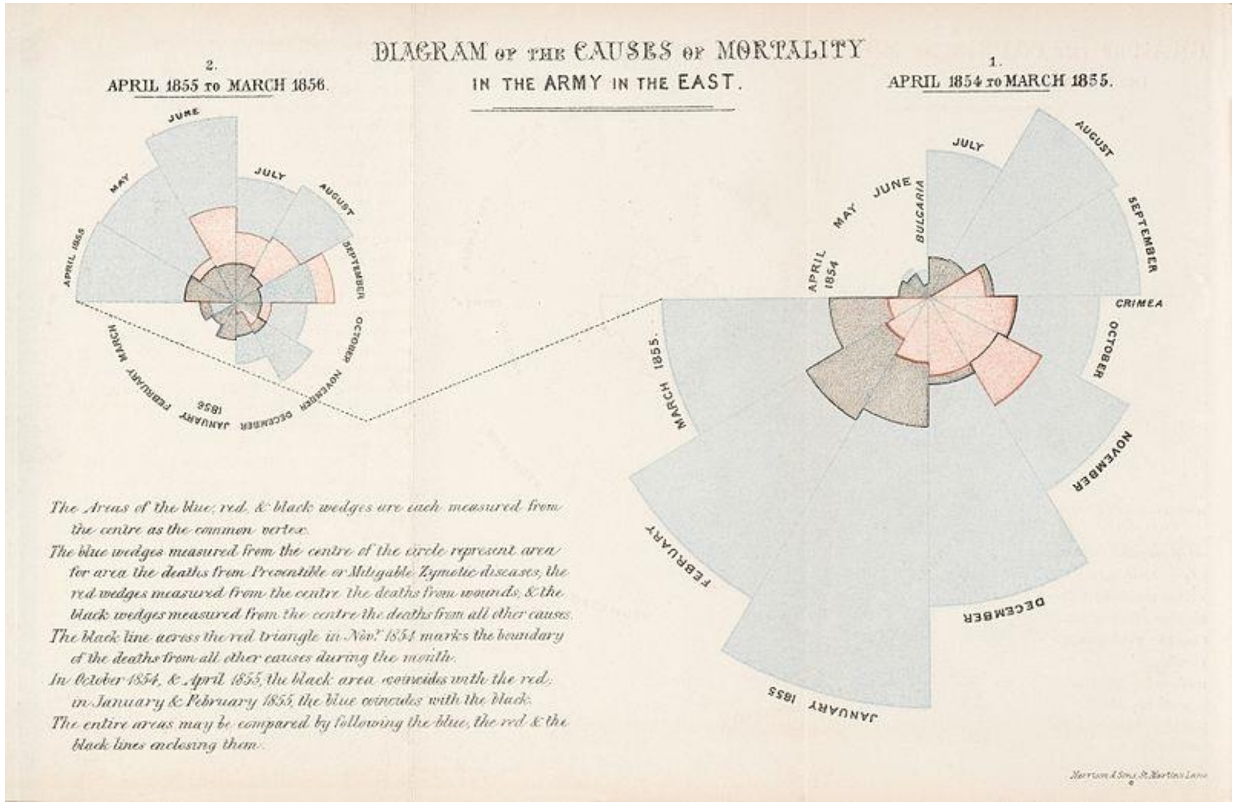


DIAGRAM OF THE CAUSES OF MORTALITY
IN THE ARMY IN THE EAST.

The Areas of the blue, red, & black wedges are each measured from the centre as the common vertex.

The blue wedges measured from the centre of the circle represent area for area the deaths from Preventible or Mitigable Zymotic diseases; the red wedges measured from the centre the deaths from wounds, & the black wedges measured from the centre the deaths from all other causes.

The black line across the red triangle in Nov.$^r$ 1854 marks the boundary of the deaths from all other causes during the month.

In October 1854, & April 1855, the black area coincides with the red; in January & February 1855, the blue coincides with the black.

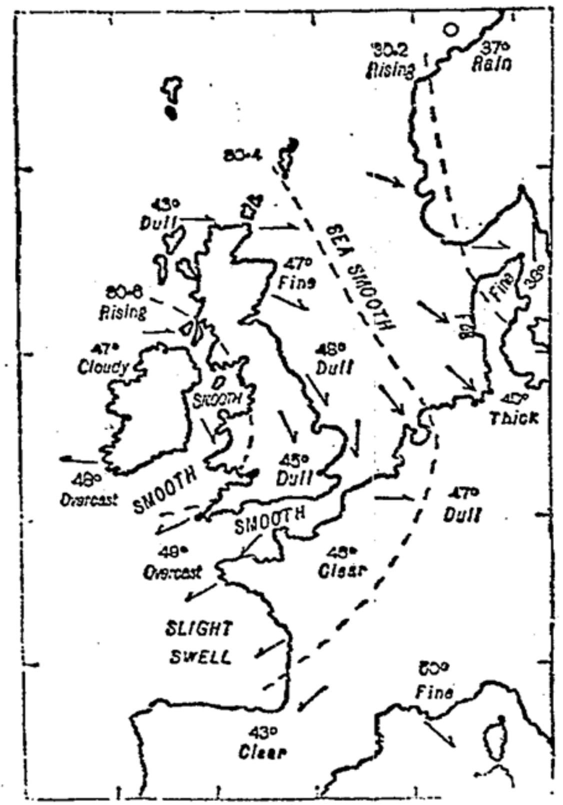
The entire areas may be compared by following the blue, the red & the black lines enclosing them.

Diagram of the causes of mortality in the army in the east

# A very brief history of data visualization

Francis Galton  (1822-1911)

WEATHER CHART, MARCH 31, 1875.

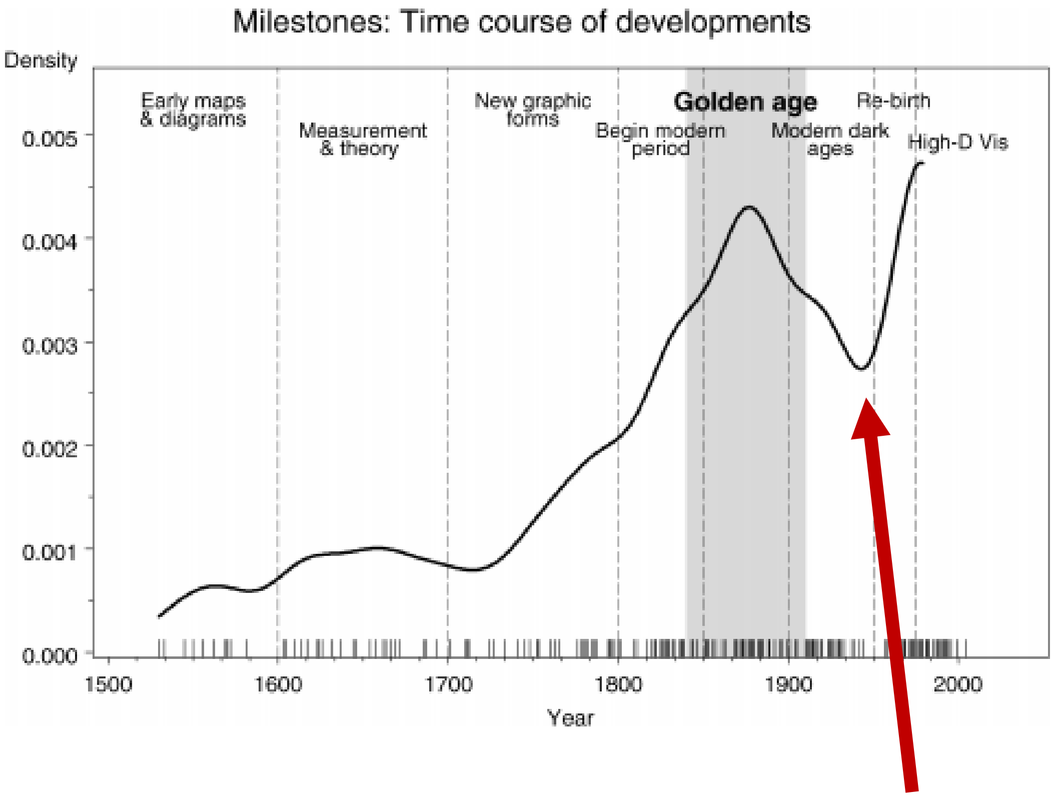


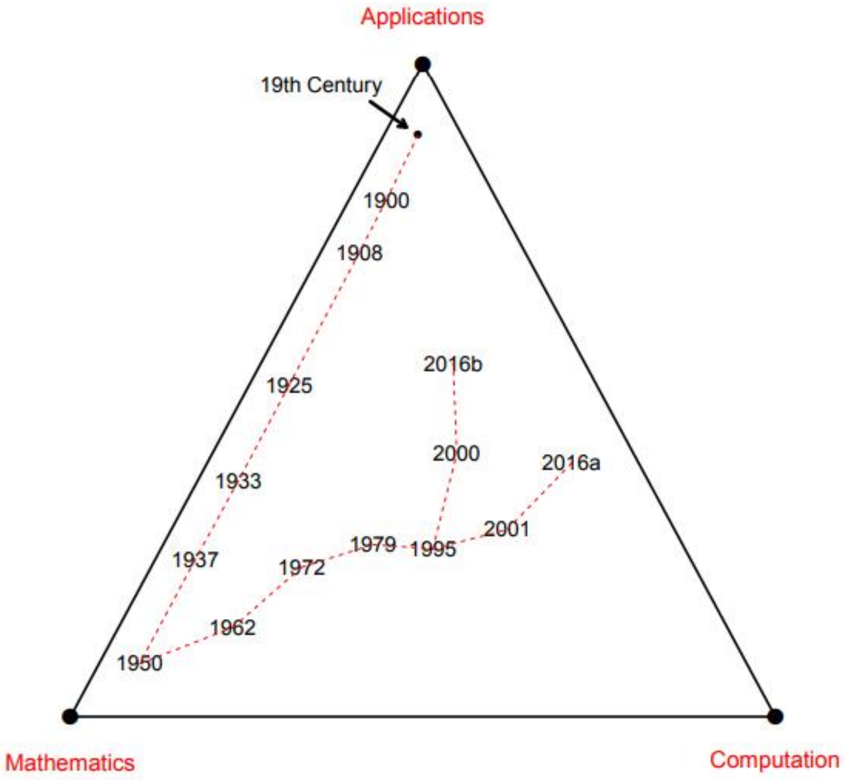
The dotted lines indicate the gradations of barometric pressure The variations of the temperature are marked by figures, the state of the sea and sky by descriptive words, and the direction of the wind by arrows—barbed and feathered according to its force. ⊙ denotes calm.

First weather map published in a newspaper (1875)

# A very brief history of data visualization

## “Graphical dark ages” around 1950

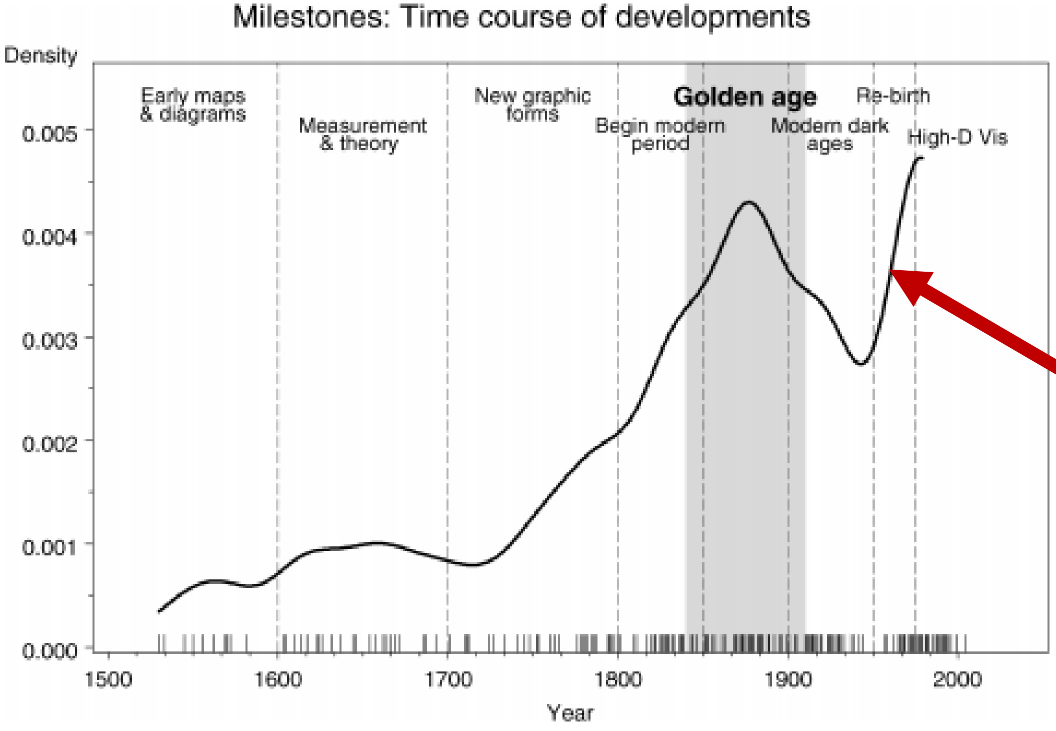




Computer Age Statistical Inference, Efron and Hastie

# A very brief history of data visualization

Currently undergoing a “Graphical re-birth”



166 CHARTING STATISTICS

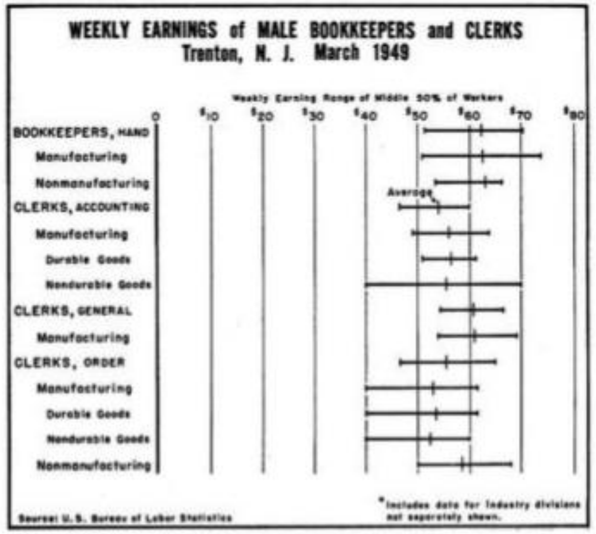


Fig. 6-23. The range bar and symbol.

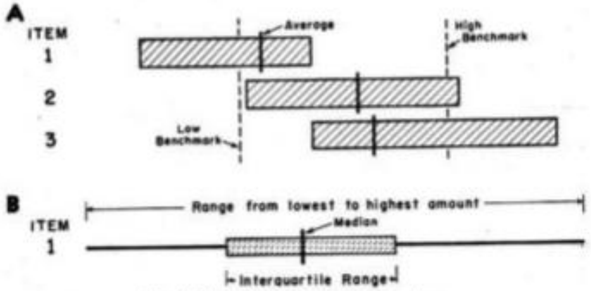
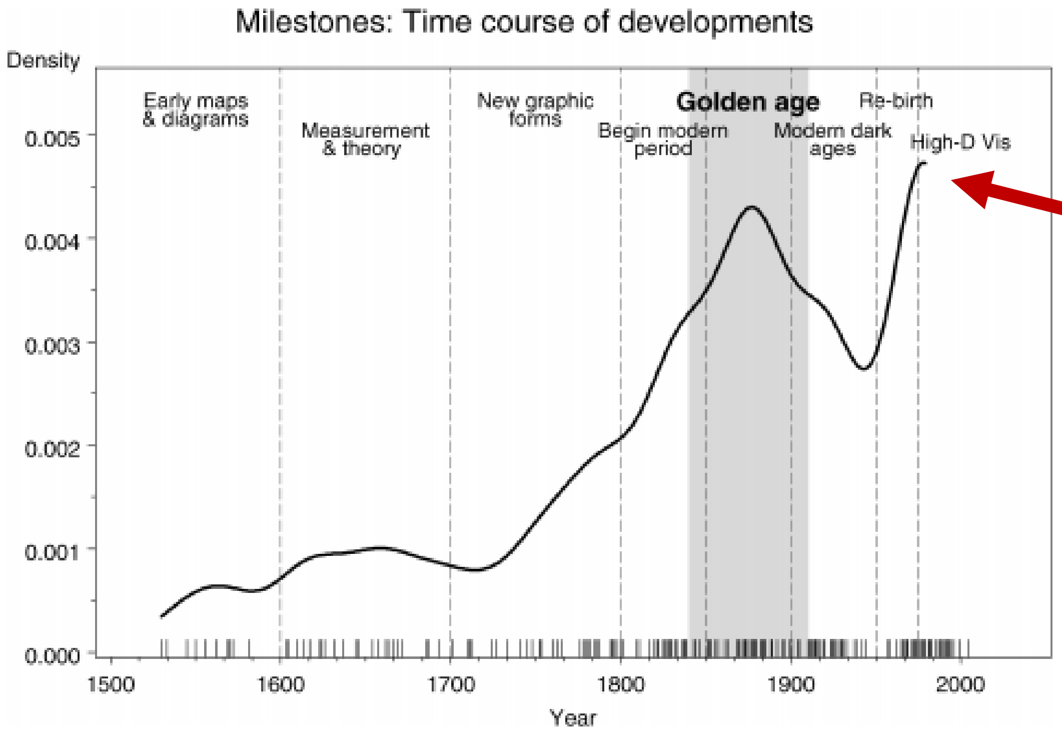


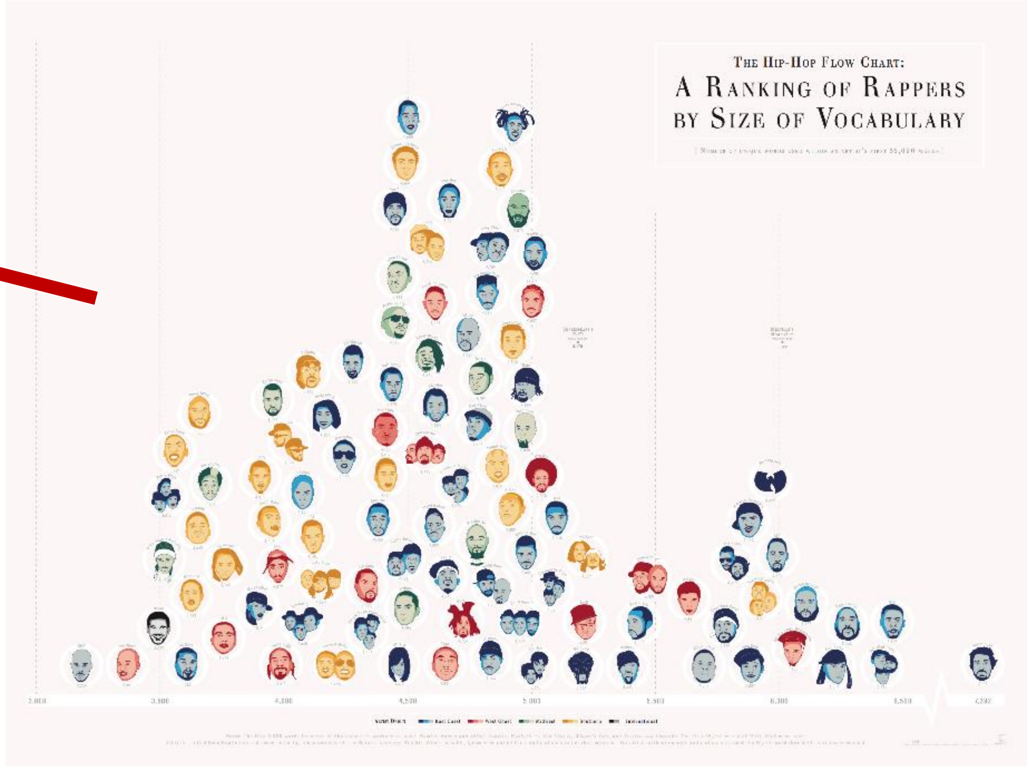
Fig. 6-24. Various uses of the range bar.

Spear 1952, Tukey 1970

# A very brief history of data visualization

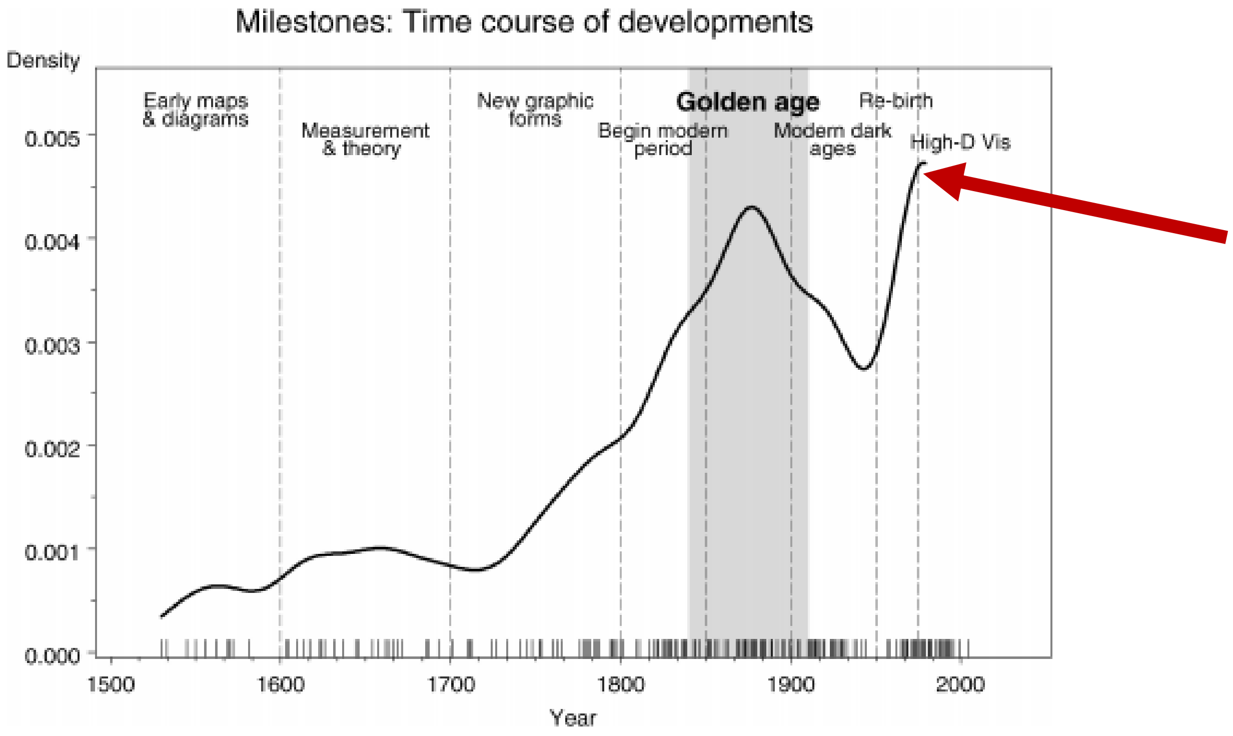
Currently undergoing a “Graphical re-birth”





# A very brief history of data visualization

Currently undergoing a “Graphical re-birth”



Hans Rosling’s gapminder

- Simple version
- TV special effects
- Ted Talk

Gapminder tools:

https://www.gapminder.org/tools

> library('gapminder')

# Next class: a grammar of graphics and ggplot

Start on homework 5 early!

Question : Find an interesting data visualization

- https://www.reddit.com/r/dataisbeautiful/
- https://flowingdata.com/