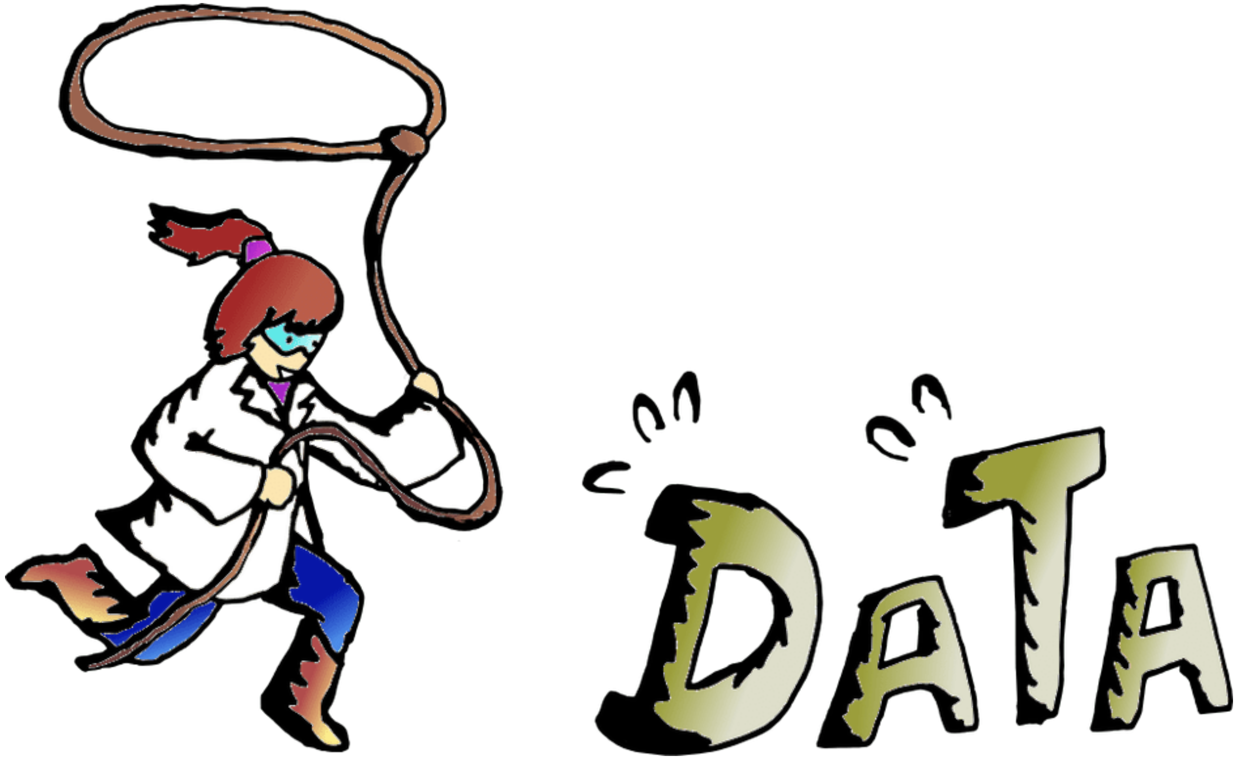
Data wrangling/manipulation



Overview

Data wrangling/manipulation with dplyr

Brief history of data visualization

Announcements

A practice midterm exam and slides with the answers will be posted by next class

- Exam format: multiple choice, short essays, short coding

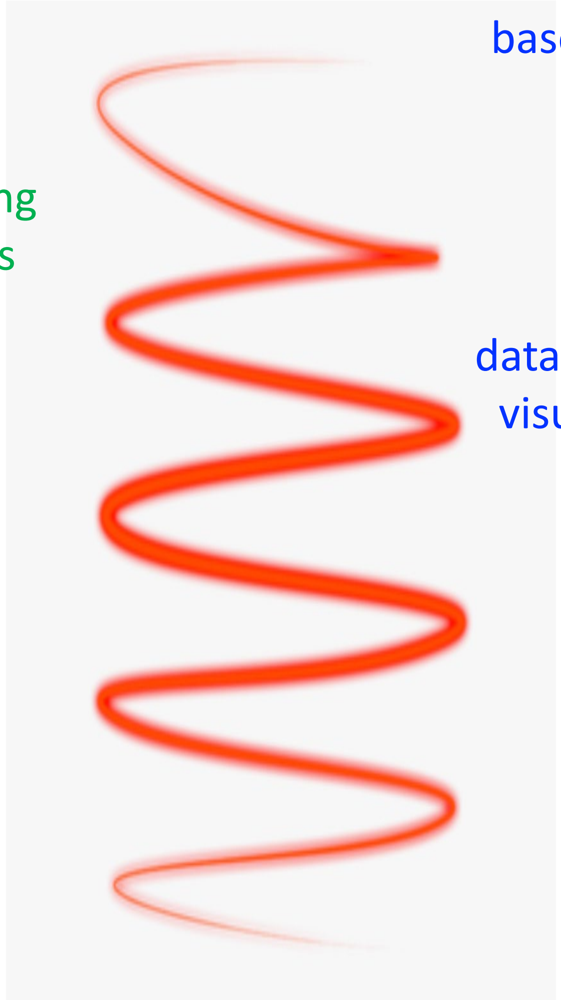
Get started on homework 5 early!!!

- I strongly 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

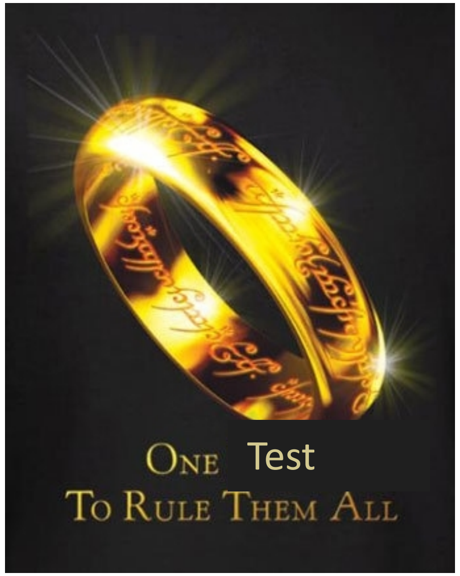
R

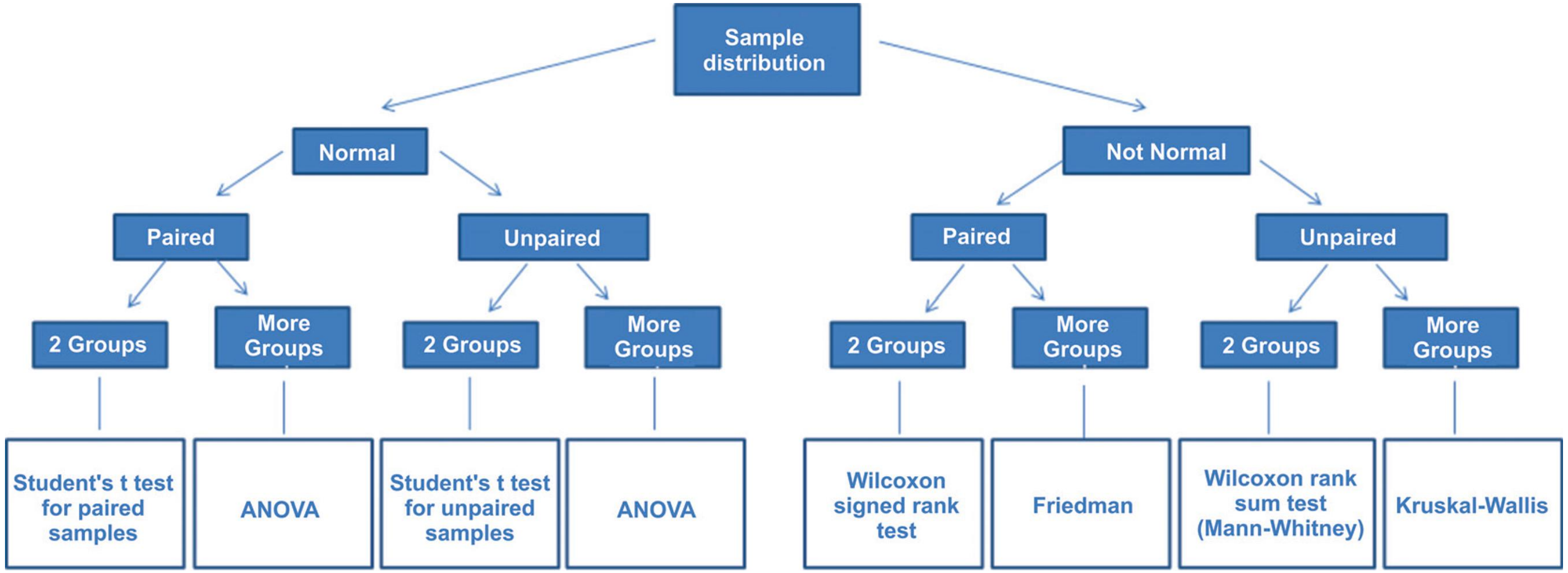
base R

resampling methods

data wrangling visualization

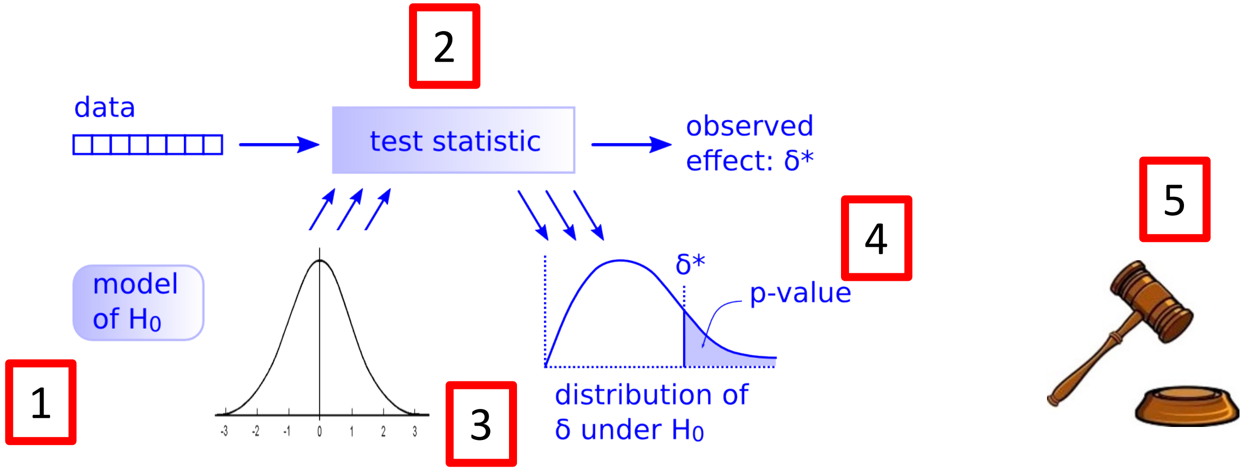
Very quick review



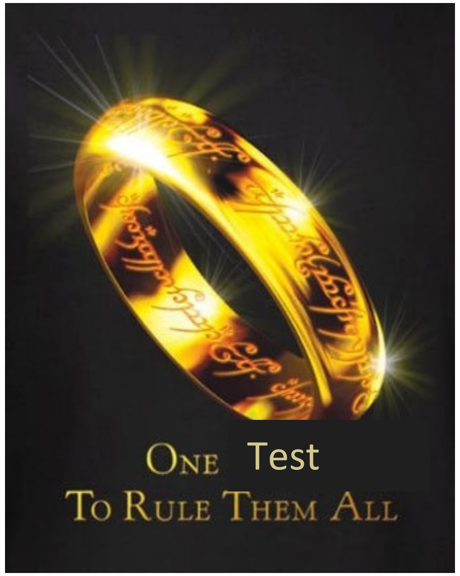


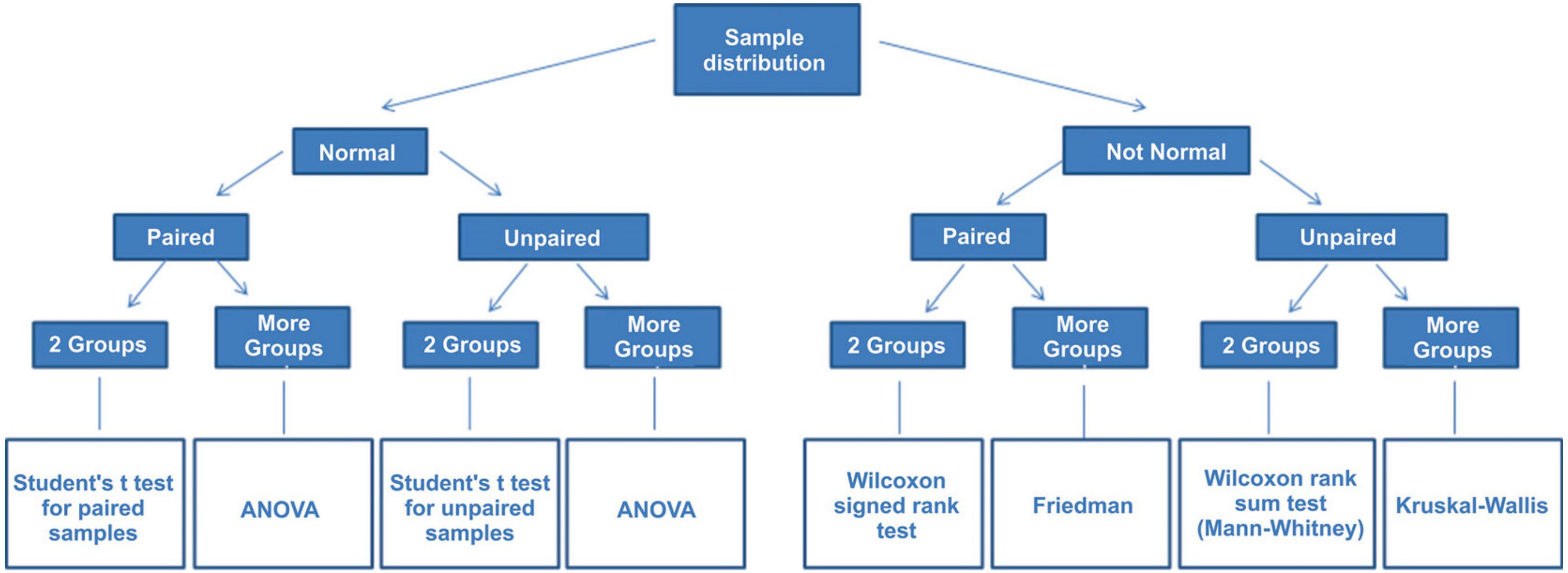
- Sample distribution
  - Normal
    - Paired
      - 2 Groups
        - Student's t test for paired samples
      - More Groups
        - ANOVA
    - Unpaired
      - 2 Groups
        - Student's t test for unpaired samples
      - More Groups
        - ANOVA
  - Not Normal
    - Paired
      - 2 Groups
        - 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: $\mu_1 = \mu_2 = \ldots = \mu_k$

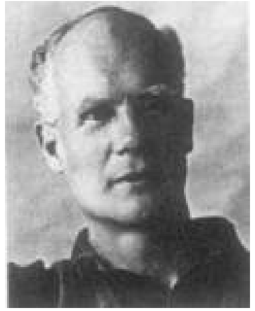
Parametric tests are derived from particular mathematical assumptions

- 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

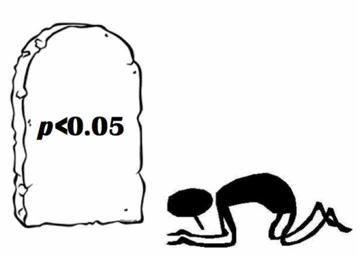
Very quick review: theories of hypothesis testing

Fisher (1890-1962)

Neyman (1894-1981)

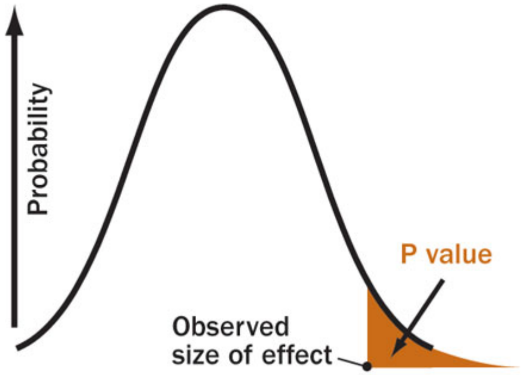


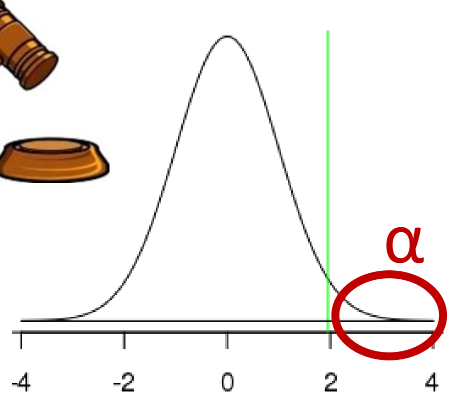
Pearson (1895-1980)

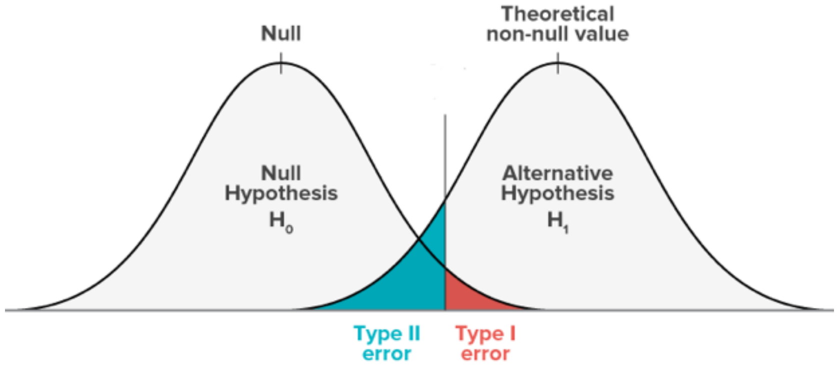


p-value a strength of evidence

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’

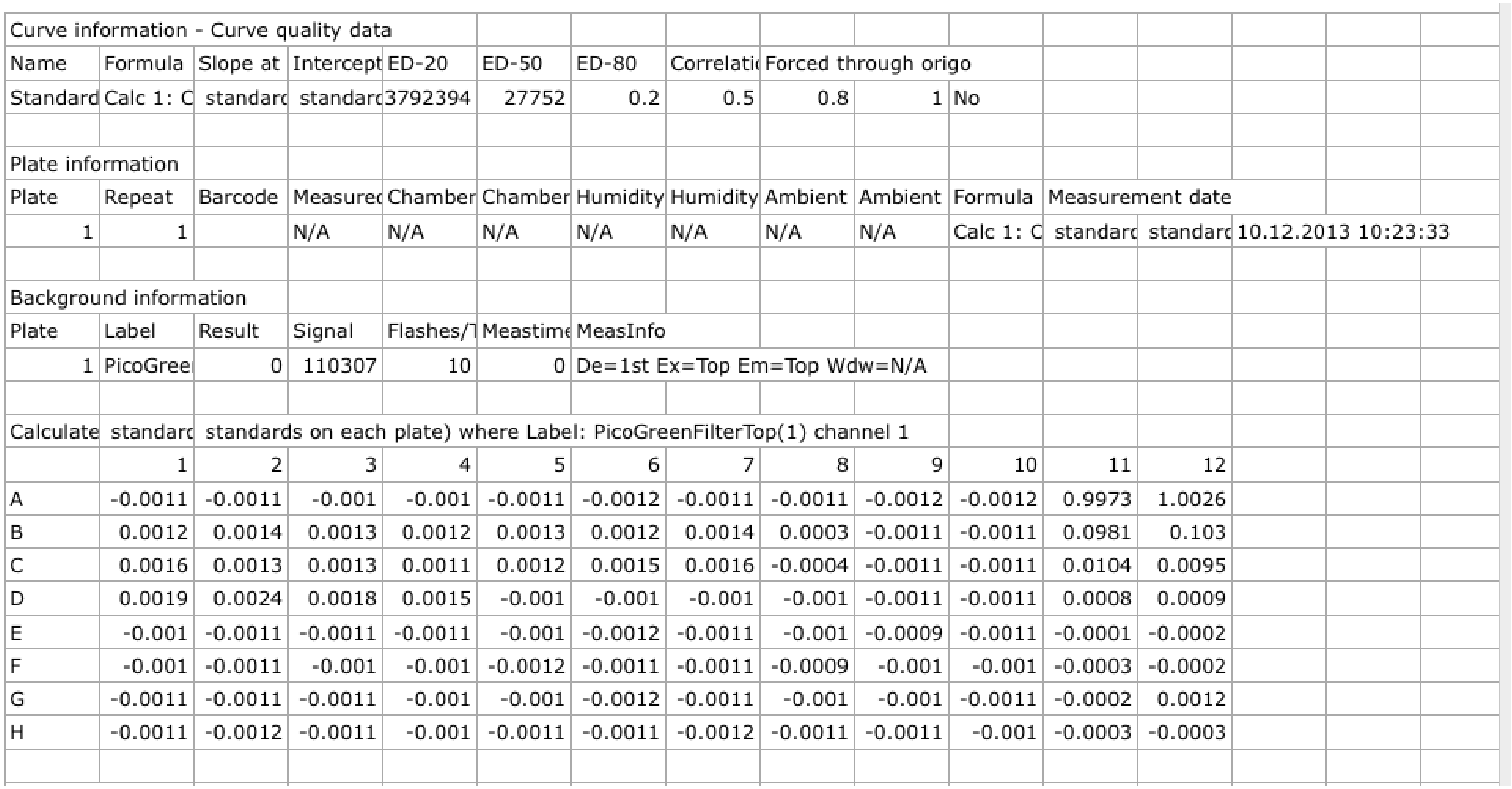
- i.e., that operate on data frames   (or tibbles)

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

Messy data…

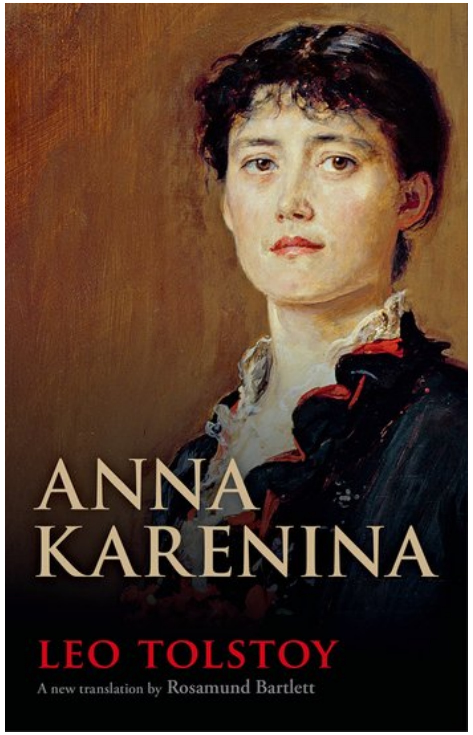
What would be an example of data that is not tidy?

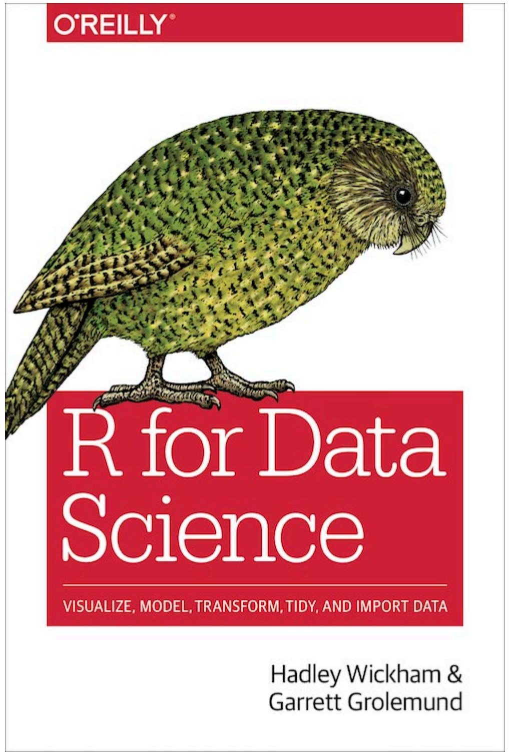


Curve information - Curve quality data													
Name	Formula	Slope at	Intercept	ED-20	ED-50	ED-80	Correlati	Forced through origo					
Standard	Calc 1: C	standarc	standarc	3792394	27752	0.2	0.5	0.8	1	No			
Plate information													
Plate	Repeat	Barcode	Measurec	Chamber	Chamber	Humidity	Humidity	Ambient	Ambient	Formula	Measurement date		
1	1		N/A	N/A	N/A	N/A	N/A	N/A	N/A	Calc 1: C	standarc	standarc	10.12.2013 10:23:33
Background information													
Plate	Label	Result	Signal	Flashes/T	Meastime	MeasInfo							
1	PicoGree	0	110307	10	0	De=1st Ex=Top Em=Top Wdw=N/A							
Calculate	standarc	standards on each plate) where Label: PicoGreenFilterTop(1) channel 1											
	1	2	3	4	5	6	7	8	9	10	11	12	
A	-0.0011	-0.0011	-0.001	-0.001	-0.0011	-0.0012	-0.0011	-0.0011	-0.0012	-0.0012	0.9973	1.0026	
B	0.0012	0.0014	0.0013	0.0012	0.0013	0.0012	0.0014	0.0003	-0.0011	-0.0011	0.0981	0.103	
C	0.0016	0.0013	0.0013	0.0011	0.0012	0.0015	0.0016	-0.0004	-0.0011	-0.0011	0.0104	0.0095	
D	0.0019	0.0024	0.0018	0.0015	-0.001	-0.001	-0.001	-0.001	-0.0011	-0.0011	0.0008	0.0009	
E	-0.001	-0.0011	-0.0011	-0.0011	-0.001	-0.0012	-0.0011	-0.001	-0.0009	-0.0011	-0.0001	-0.0002	
F	-0.001	-0.0011	-0.001	-0.001	-0.0012	-0.0011	-0.0011	-0.0009	-0.001	-0.001	-0.0003	-0.0002	
G	-0.0011	-0.0011	-0.0011	-0.001	-0.001	-0.0012	-0.0011	-0.001	-0.001	-0.0011	-0.0002	0.0012	
H	-0.0011	-0.0012	-0.0011	-0.001	-0.0011	-0.0011	-0.0012	-0.0011	-0.0011	-0.001	-0.0003	-0.0003	

Messy data…

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

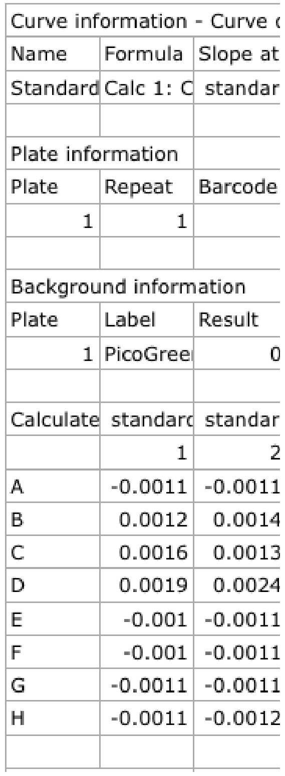



O'REILLY
R for Data Science
VISUALIZE, MODEL, TRANSFORM, TIDY, AND IMPORT DATA
Hadley Wickham & Garrett Grolemund

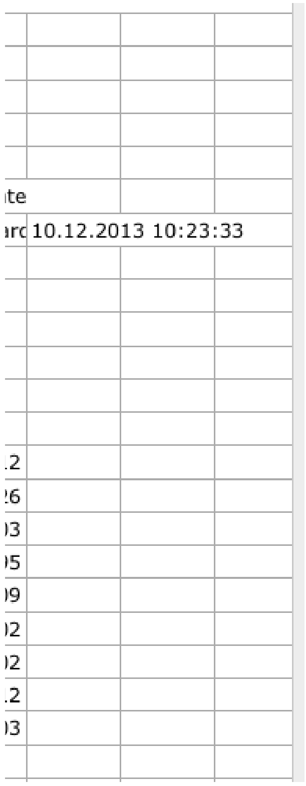
“Tidy datasets are all alike, but every messy dataset is messy in its own way.” –
– Hadley Wickham

Messy data…

Messy data can be difficult to deal with



Curve information - Curve c		
Name	Formula	Slope at
Standard	Calc 1: C	standar
Plate information		
Plate	Repeat	Barcode
1	1	
Background information		
Plate	Label	Result
1	PicoGree	0
Calculate	standard	standar
	1	2
A	-0.0011	-0.0011
B	0.0012	0.0014
C	0.0016	0.0013
D	0.0019	0.0024
E	-0.001	-0.0011
F	-0.001	-0.0011
G	-0.0011	-0.0011
H	-0.0011	-0.0012

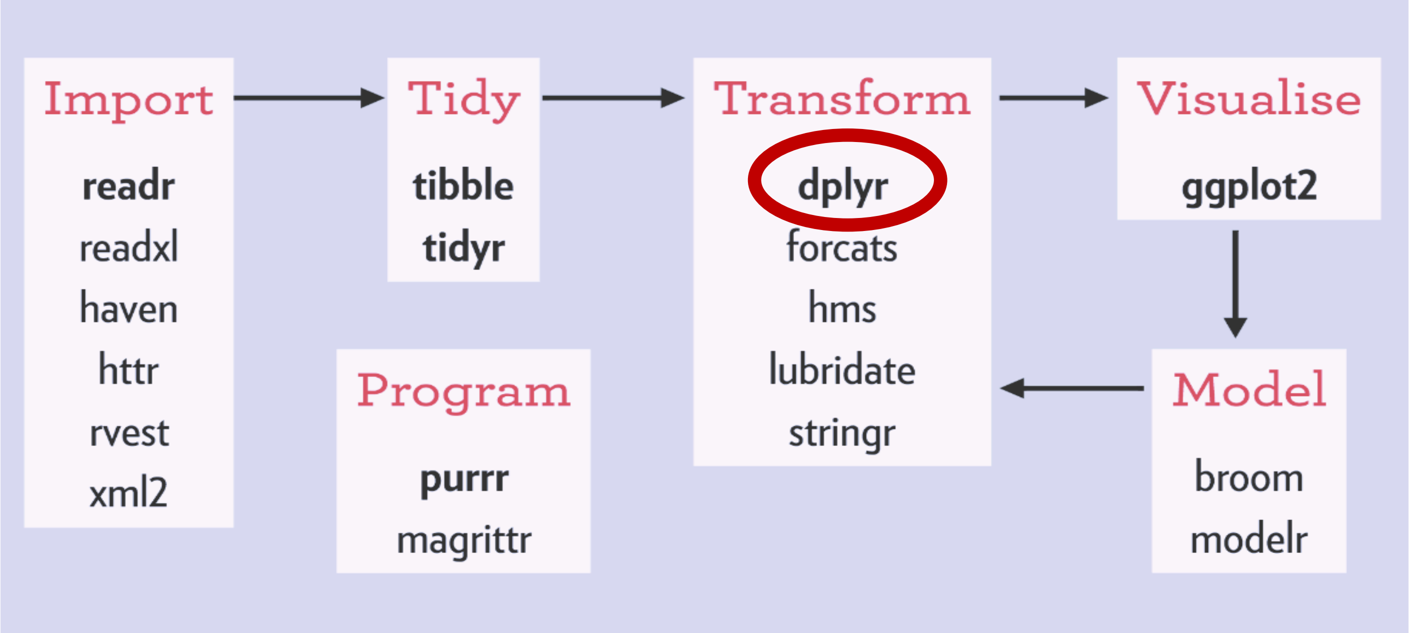


te			
arc	10.12.2013 10:23:33		
.2			
26			
03			
05			
09			
02			
02			
.2			
03			

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

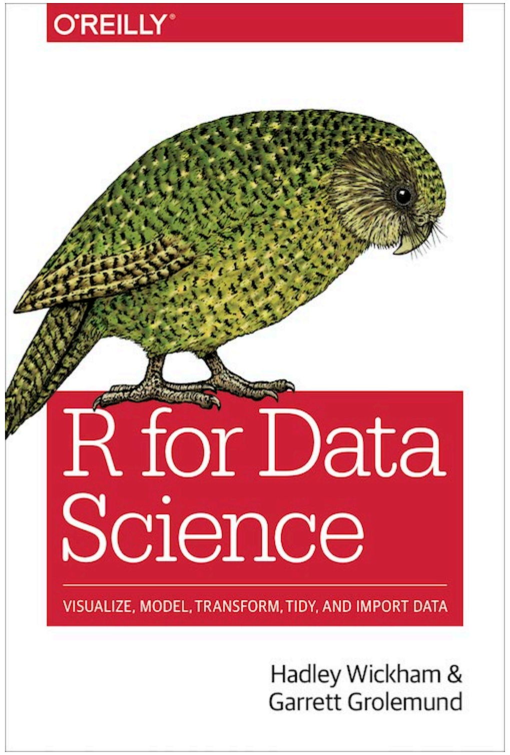
dplyr is a package that has a set of verbs that are useful for transformations data:

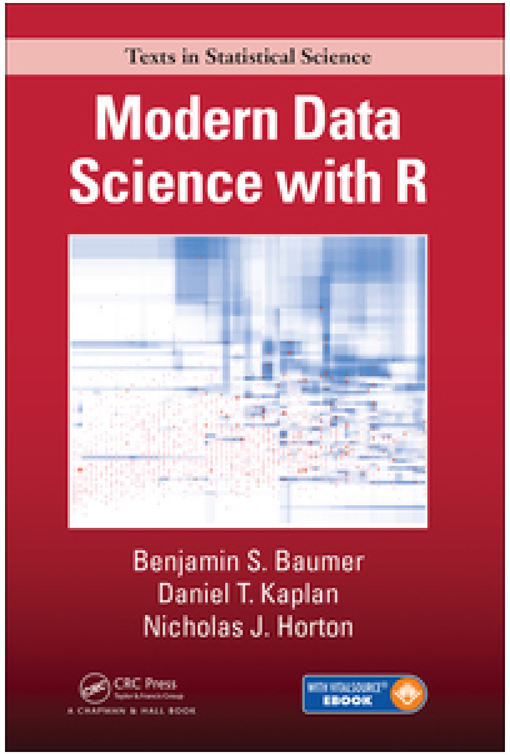
1. filter()
2. select()
3. mutate()
4. arrange()
5. summarize()
6. group_by()

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

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

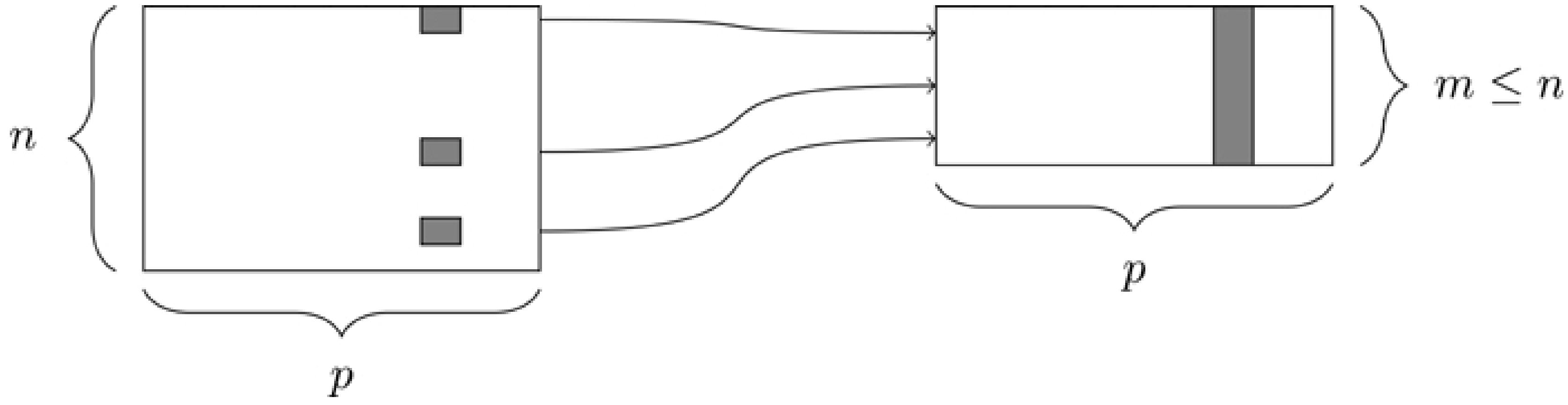
Quick overview of the dplyr functions





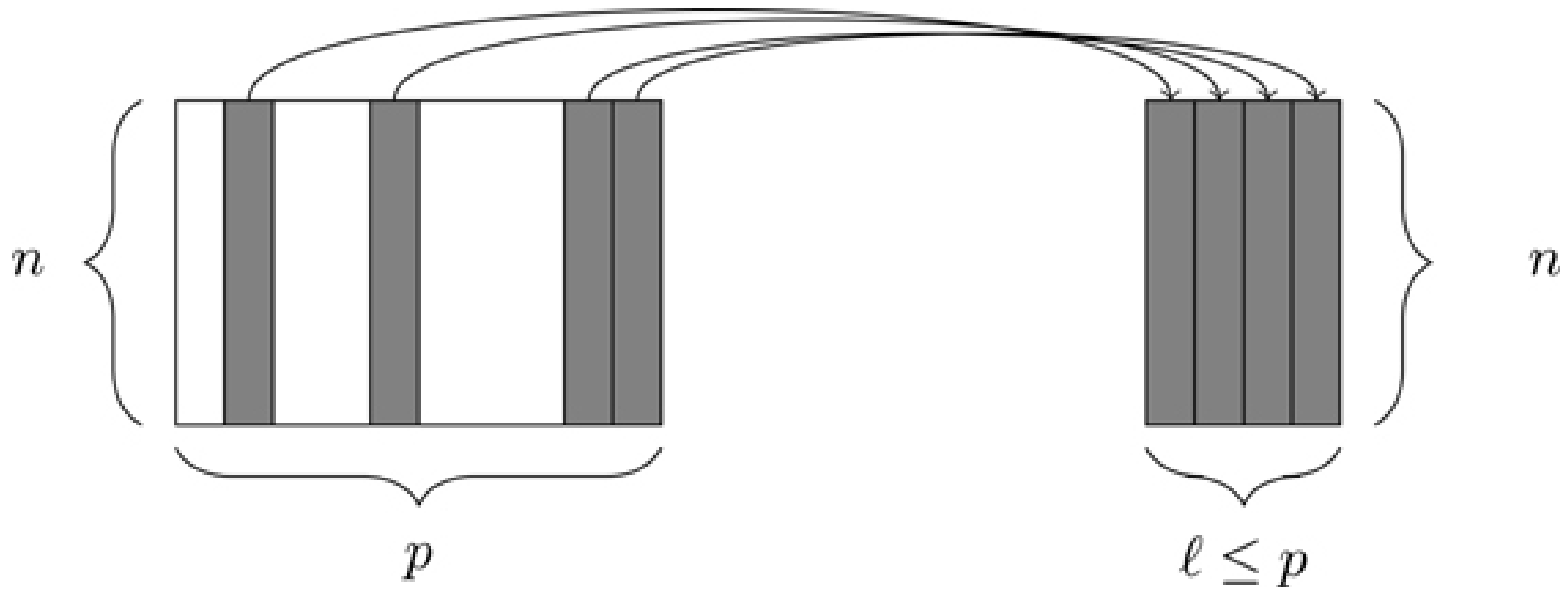
1. filter()

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



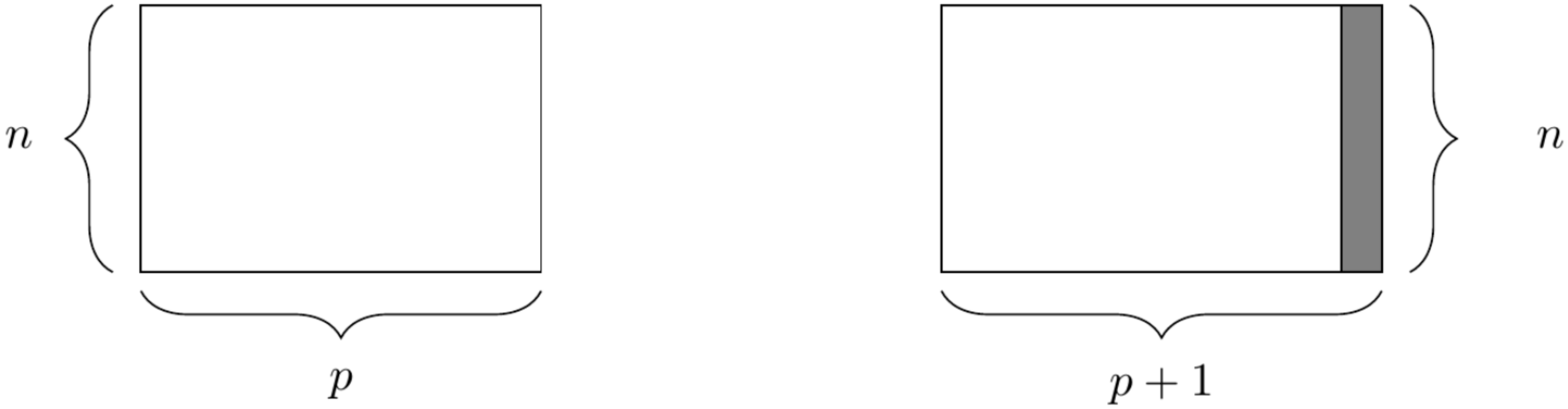
2. select()

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



3. mutate()

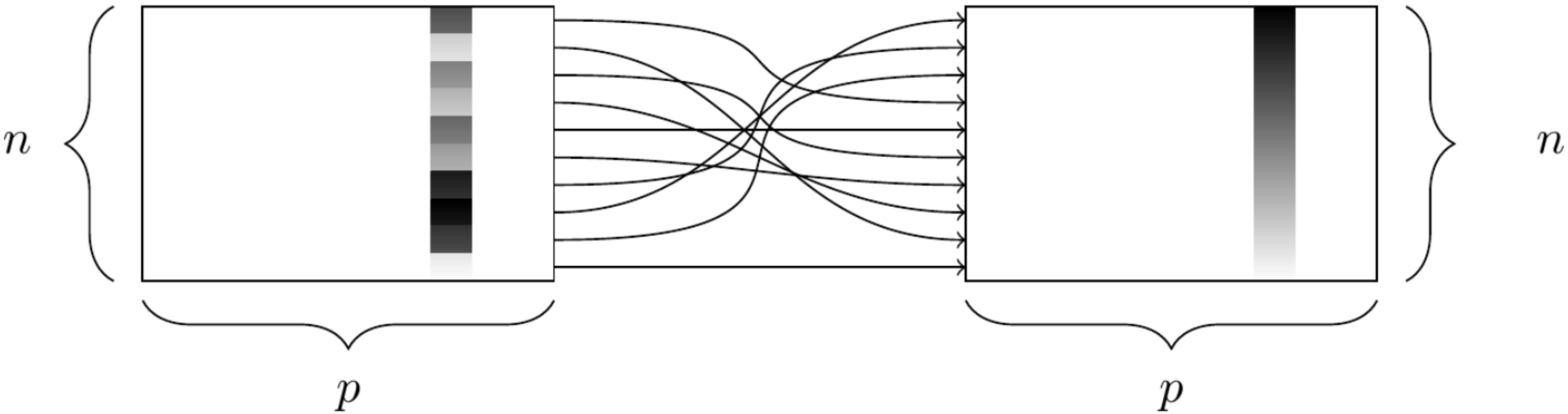
The mutate() function allows you to create new columns that are functions of existing columns



4. arrange()

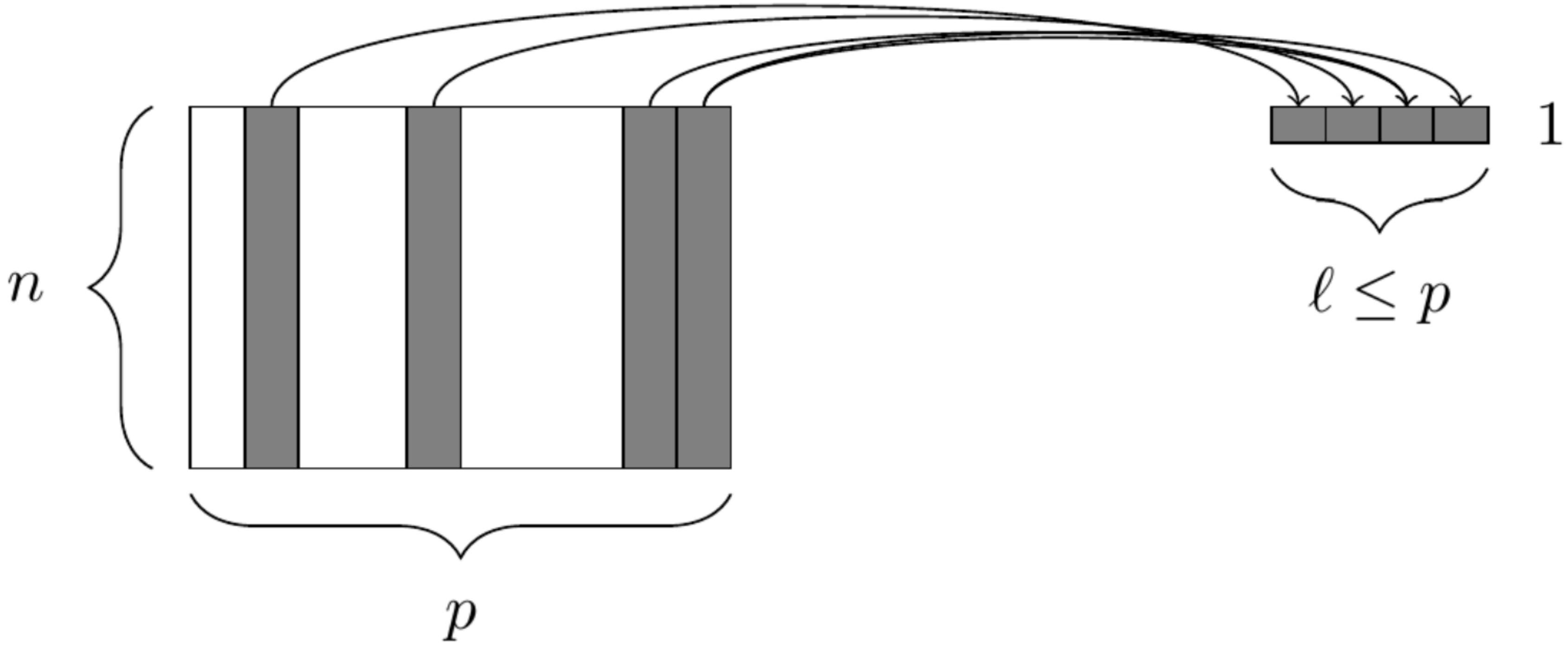
The arrange() function arranges the rows based values in a column

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



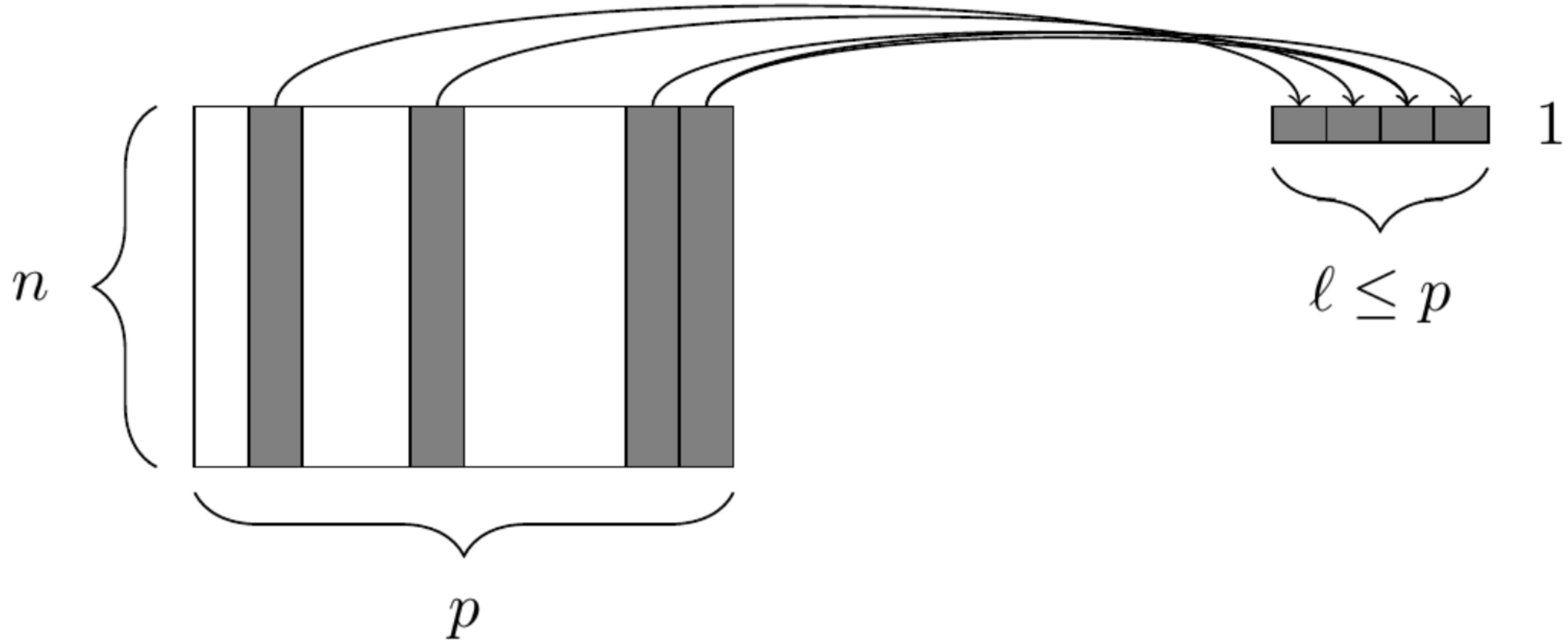
5. summarize()

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



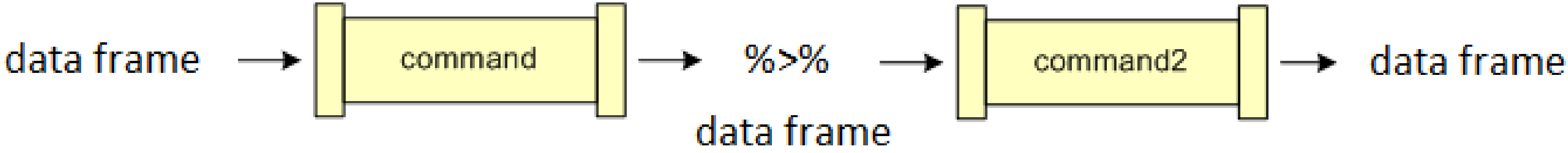
6. The group_by() function

The group_by() function groups variables for future operations



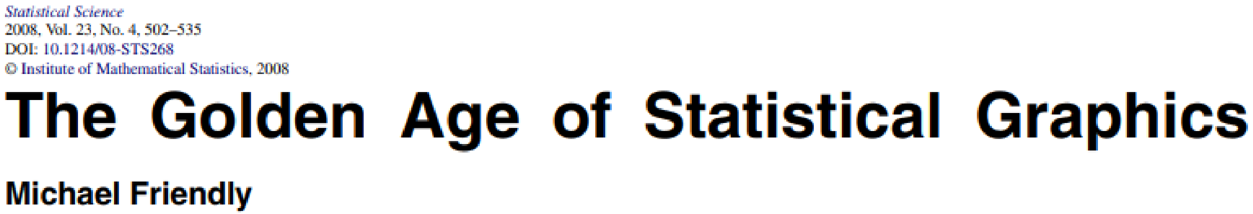
The pipe operator

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



Let’s try it out!

A very brief history of data visualization



Statistical Science
2008, Vol. 23, No. 4, 502–535
DOI: 10.1214/08-STS268
© Institute of Mathematical Statistics, 2008

The Golden Age of Statistical Graphics

Michael Friendly

Data visualization

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

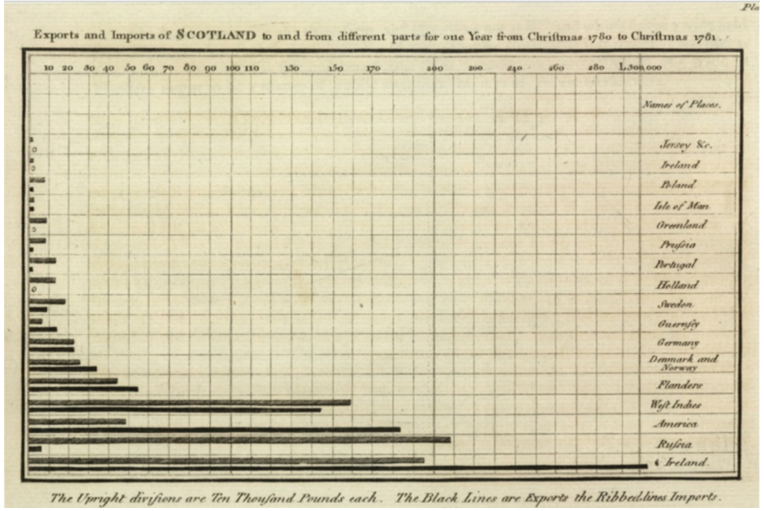
Whatever relates to extent and quantity may be represented by geometrical figures. Statistical projections which speak to the senses without fatiguing the mind, possess the advantage of fixing the attention on a great number of important facts.

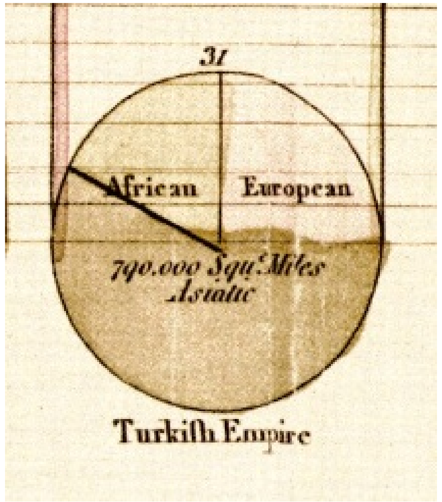
—Alexander von Humboldt, 1811

A very brief history of data visualization

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

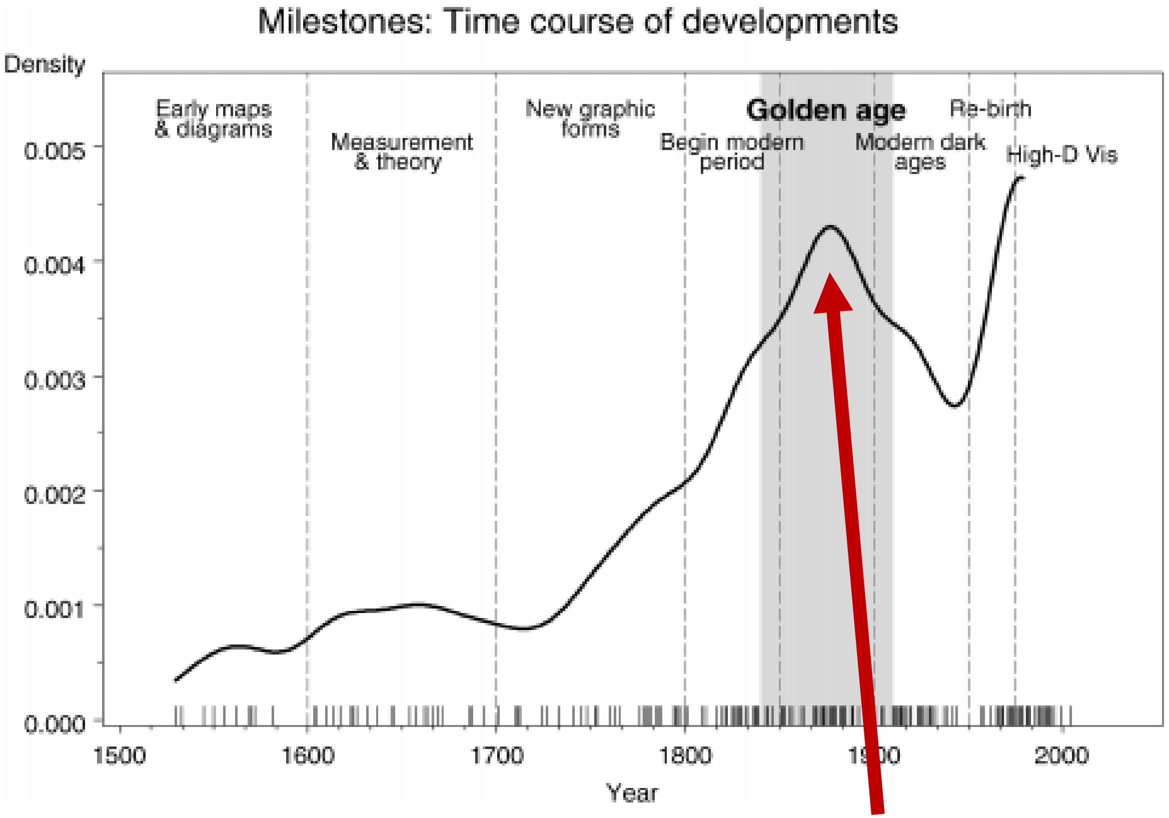
William Playfair (1759-1823) credited with inventing the line graph, bar chart and pie chart





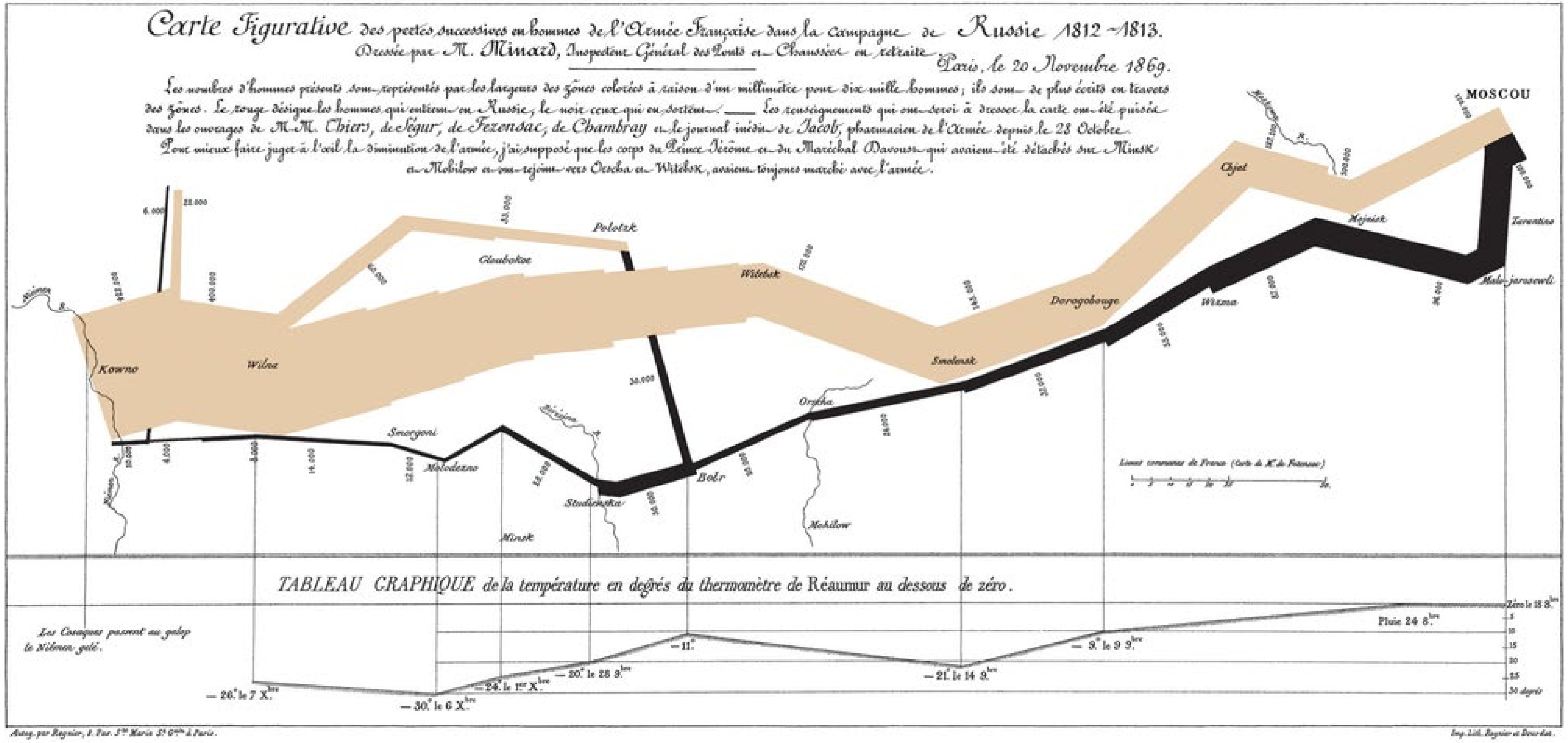
A very brief history of data visualization

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



A very brief history of data visualization

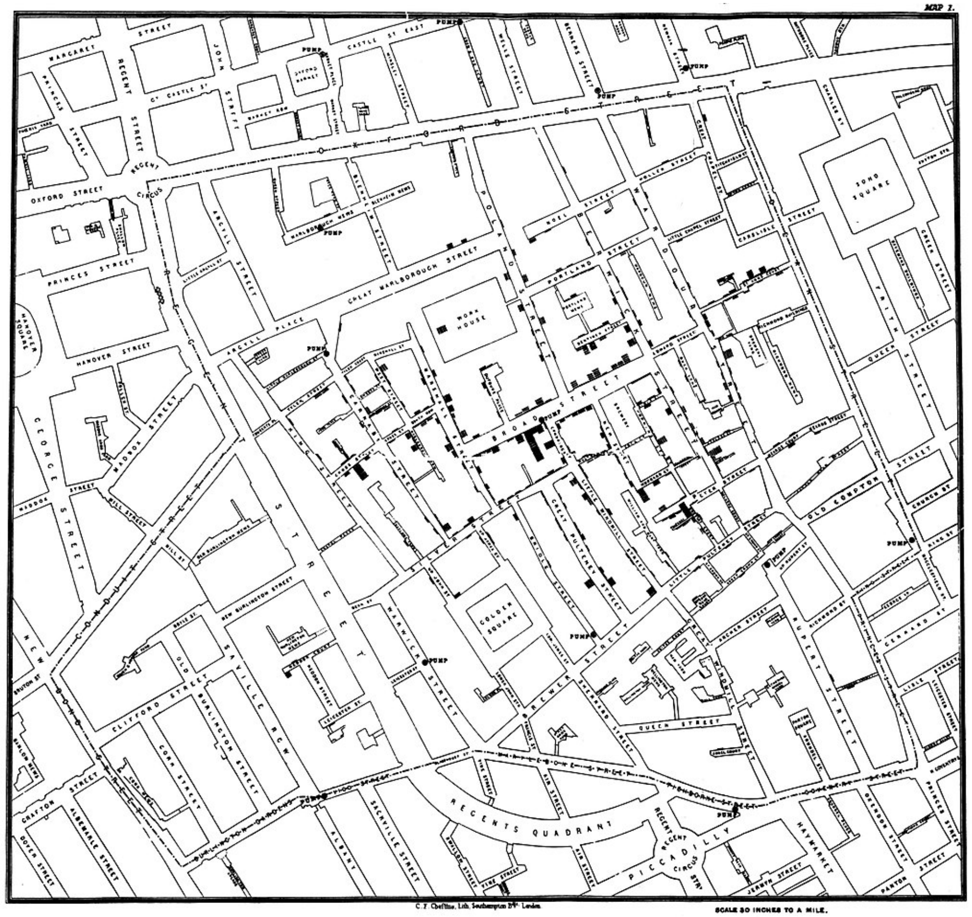
Joseph Minard (1781-1870)


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.
MOSCOU
Niemen R.
Kowno
Wilna
Smorgoni
Molodezno
Glubokoe
Polotzk
Witebsk
Smolensk
Dorogobouge
Wizma
Chjat
Mojaisk
Tarantino
Malo-jarosewli
Moskowa R.
Bérézina R.
Studienska
Minsk
Botr
Orscha
Mohilow
422.000
400.000
6.000
22.000
60.000
33.000
175.000
145.000
127.100
100.000
96.000
87.000
55.000
37.000
34.000
24.000
20.000
50.000
28.000
30.000
12.000
14.000
8.000
4.000
10.000
Lieues communes de France (Carte de M. de Fezensac)
TABLEAU GRAPHIQUE de la température en degrés du thermomètre de Réaumur au dessous de zéro.
Les Cosaques passent au galop le Niemen gelé.
— 26.° le 7 X.bre
— 30.° le 6 X.bre
— 24.° le 1.er X.bre
— 20.° le 28 9.bre
— 11.°
— 21.° le 14 9.bre
— 9.° le 9 9.bre
Pluie 24 8.bre
Zéro le 18 8.bre
5
10
15
20
25
30 degrés
Autog. par Regnier, 8. Pas. S.te Marie S.t G.ain à Paris.
Imp. Lith. Regnier et Dourdet.

Map of Napoleon's march on Russia

A very brief history of data visualization

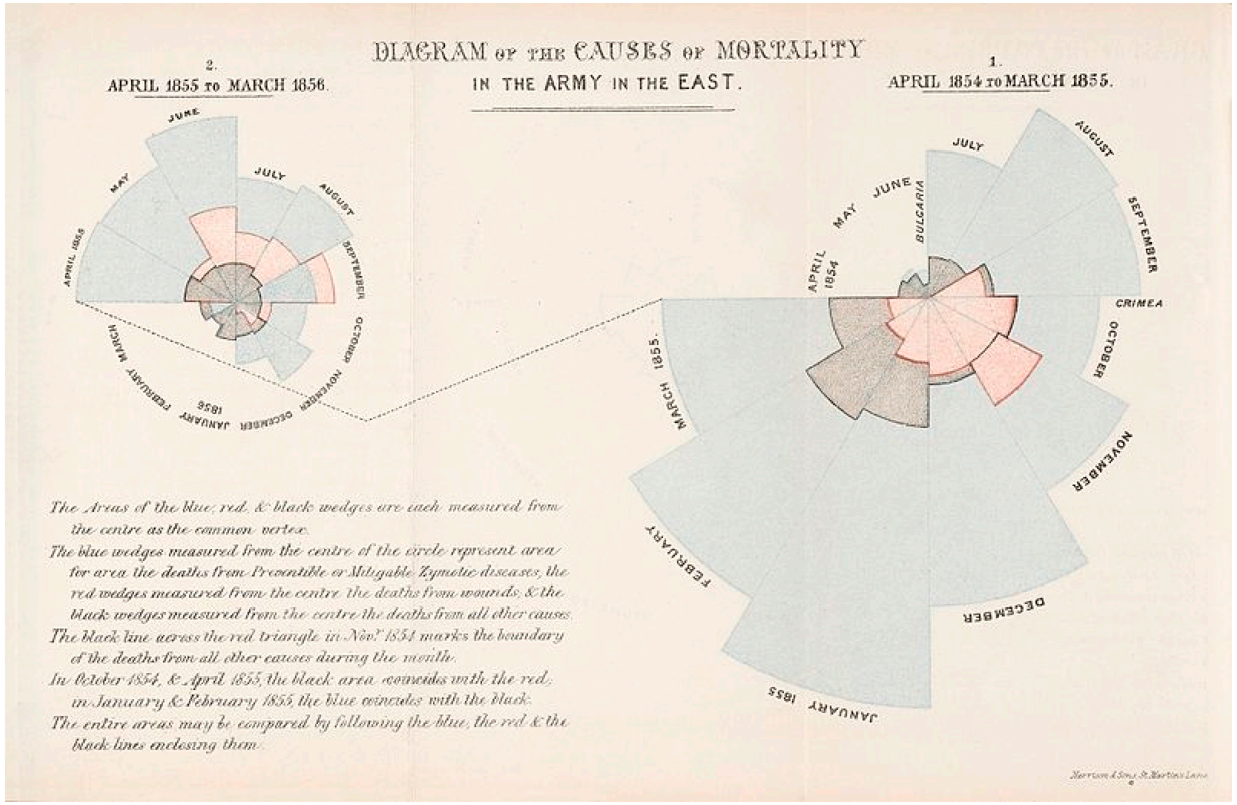
John Snow (1813-1858)



Clusters of cholera cases in London epidemic of 1854

A very brief history of data visualization

Florence Nightingale (1820-1910)



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.

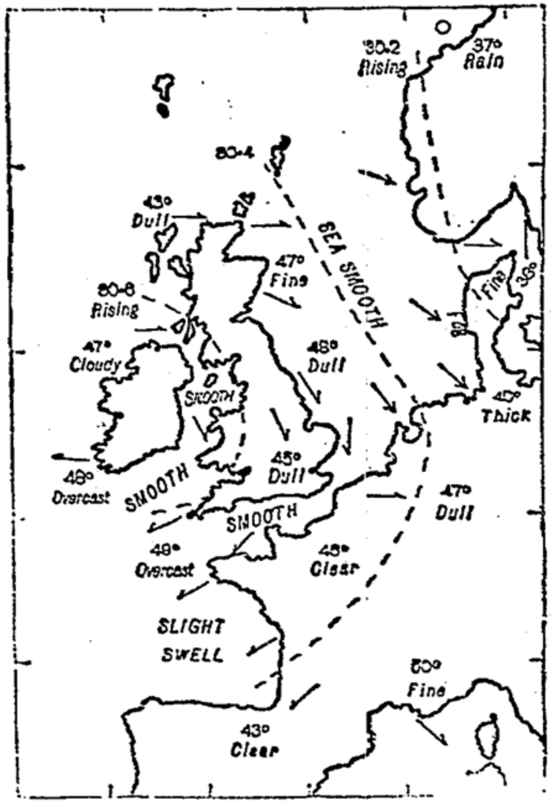
The entire areas may be compared by following the blue, the red & the black lines enclosing them.

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.

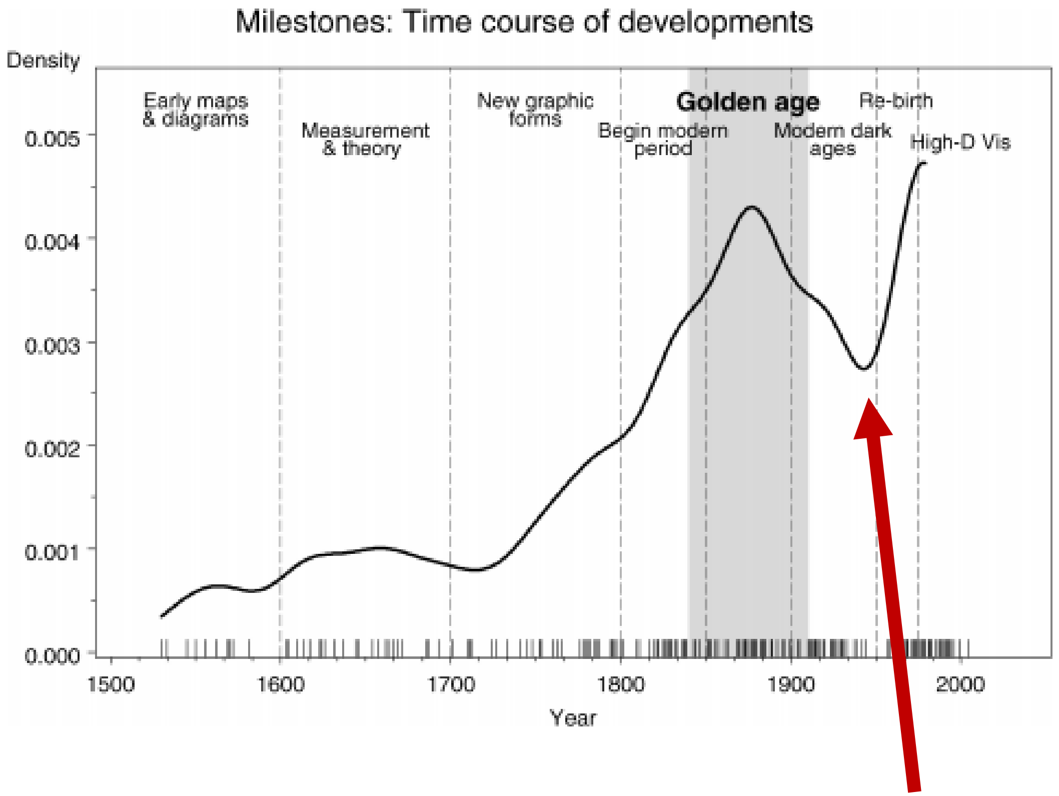


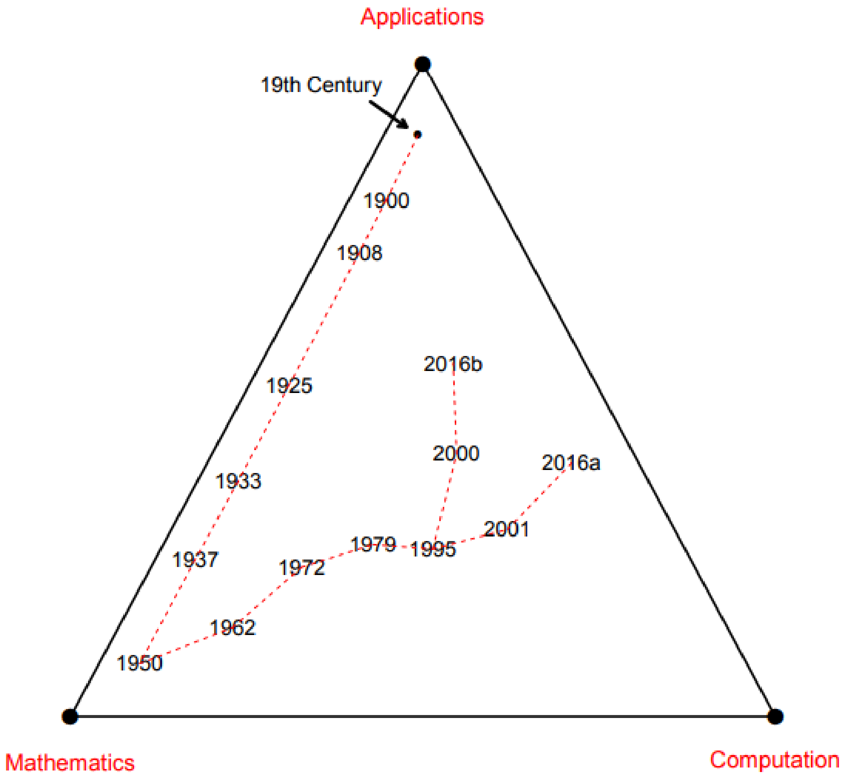
The dotted lines indicate the gradations of barometric pressure The variations of the temperature are marked by figures, the state of the sea and sky by descriptive words, and the direction of the wind by arrows—barbed and feathered according to its force. ⊙ denotes calm.

First weather map published in a newspaper (1875)

A very brief history of data visualization

“Graphical dark ages” around 1950

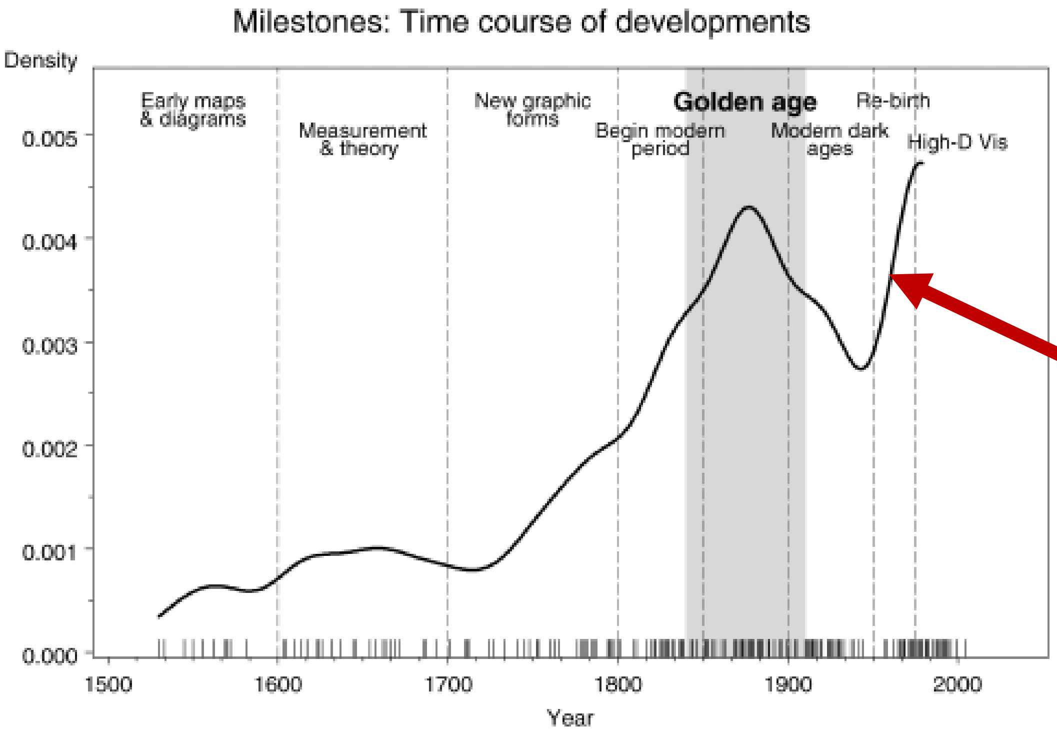




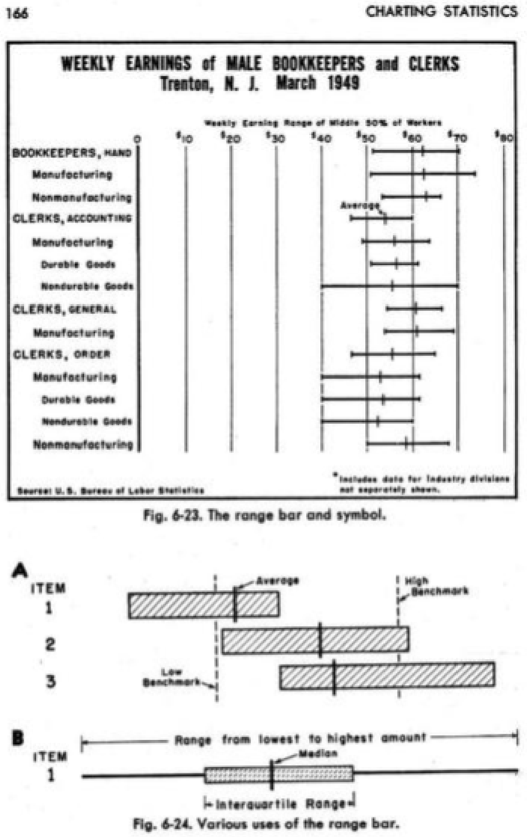
Computer Age Statistical Inference, Efron and Hastie

A very brief history of data visualization

Currently undergoing a “Graphical re-birth”



Box plot



166 CHARTING STATISTICS

Source: U.S. Bureau of Labor Statistics

*Includes data for industry divisions not separately shown.

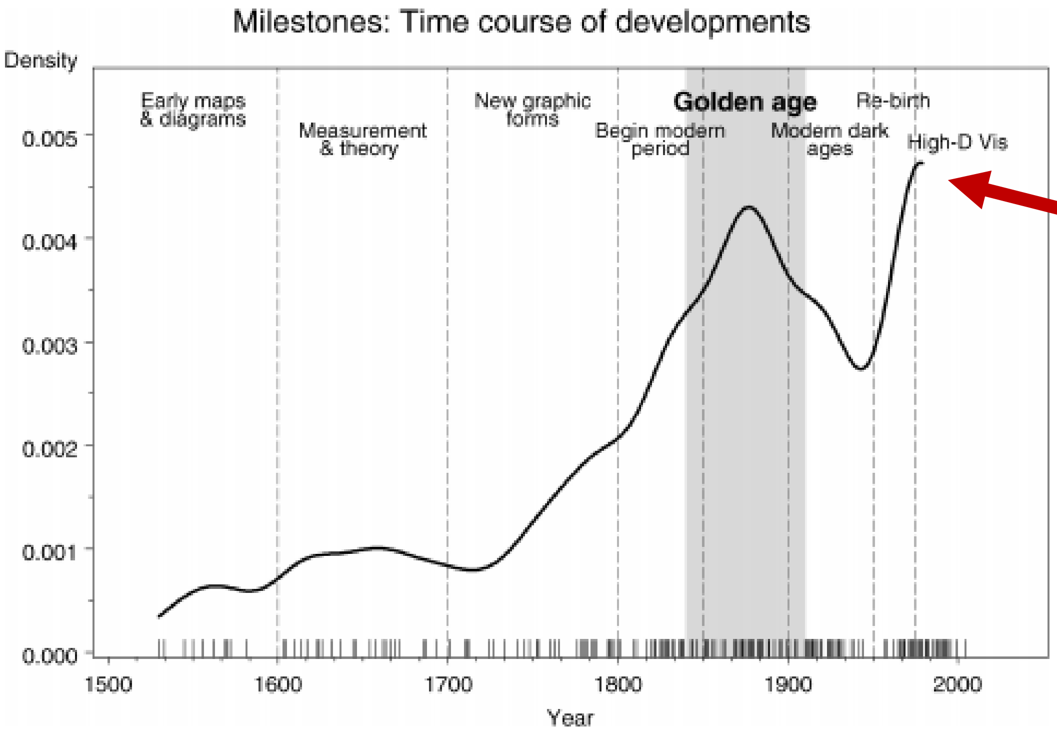
Fig. 6-23. The range bar and symbol.

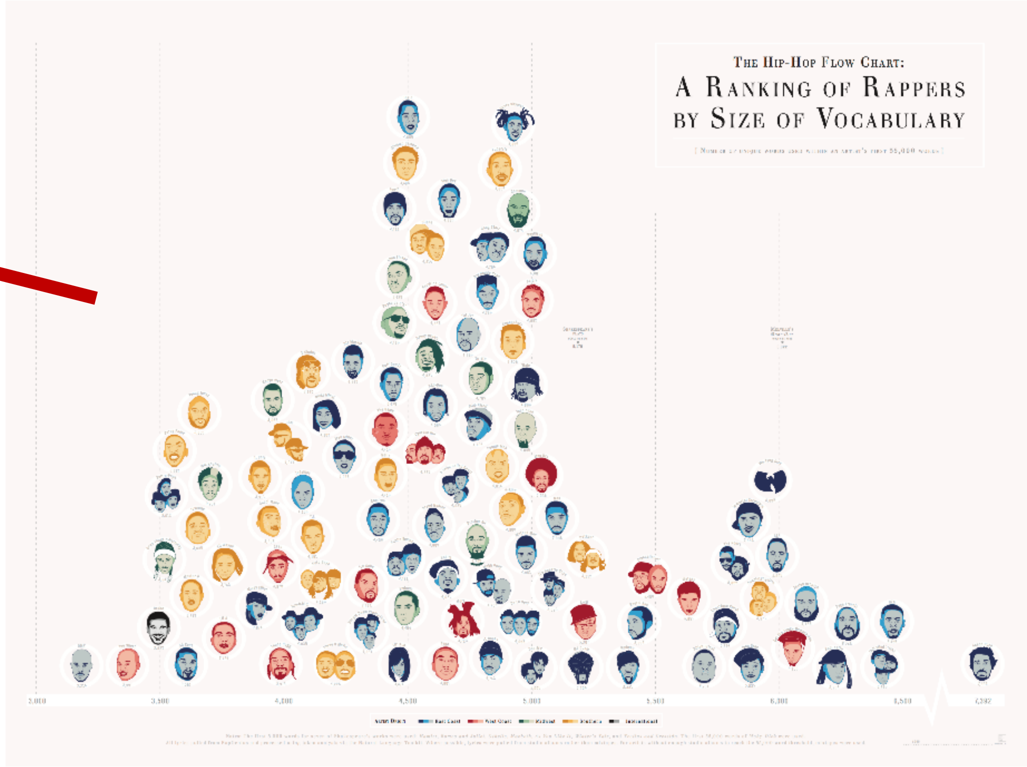
Fig. 6-24. Various uses of the range bar.

Spear 1952, Tukey 1970

A very brief history of data visualization

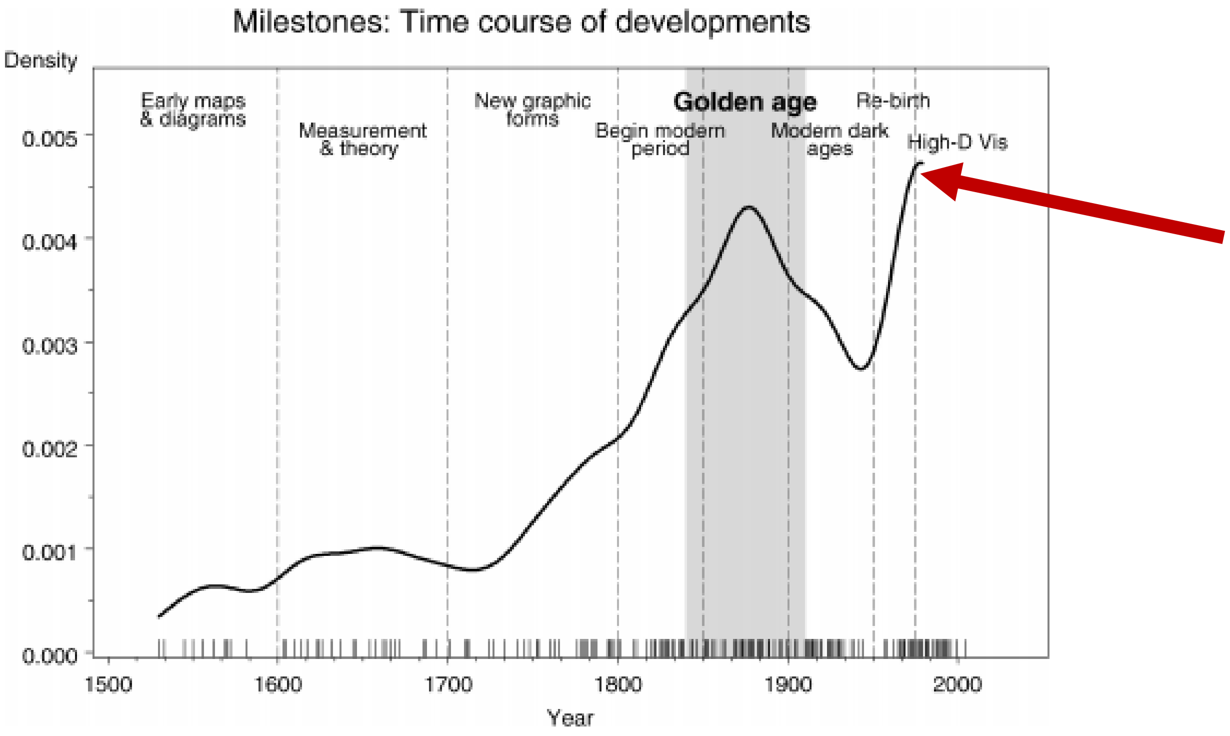
Currently undergoing a “Graphical re-birth”





A very brief history of data visualization

Currently undergoing a “Graphical re-birth”



Hans Rosling’s gapminder

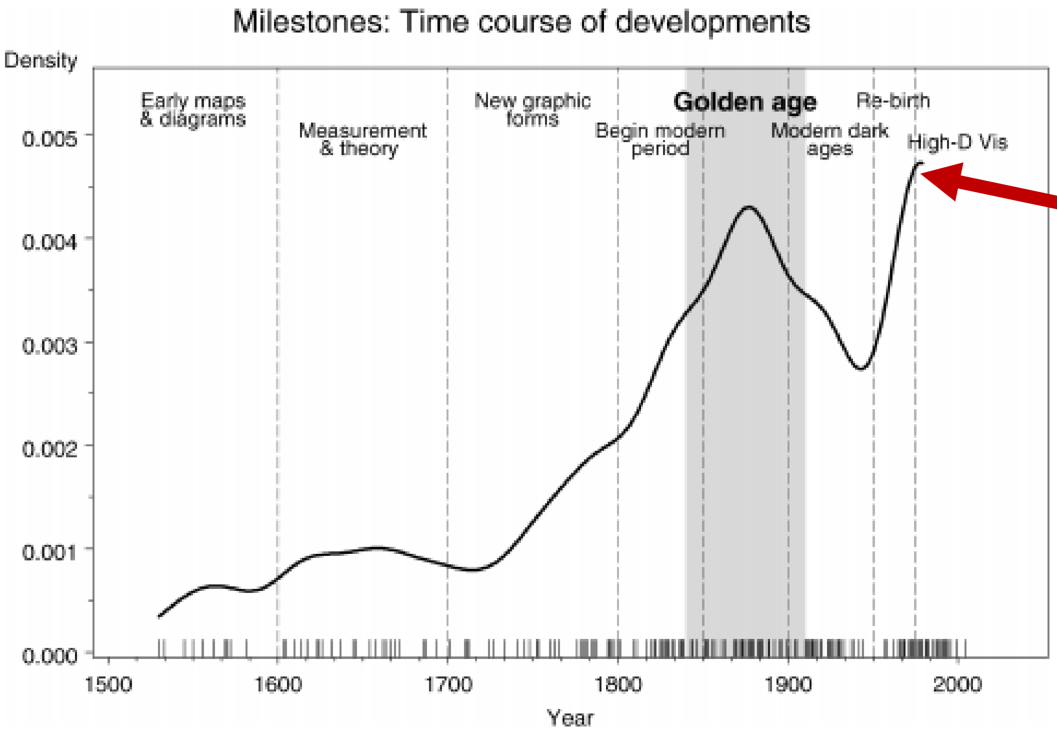
- Simple version
- TV special effects
- Ted Talk

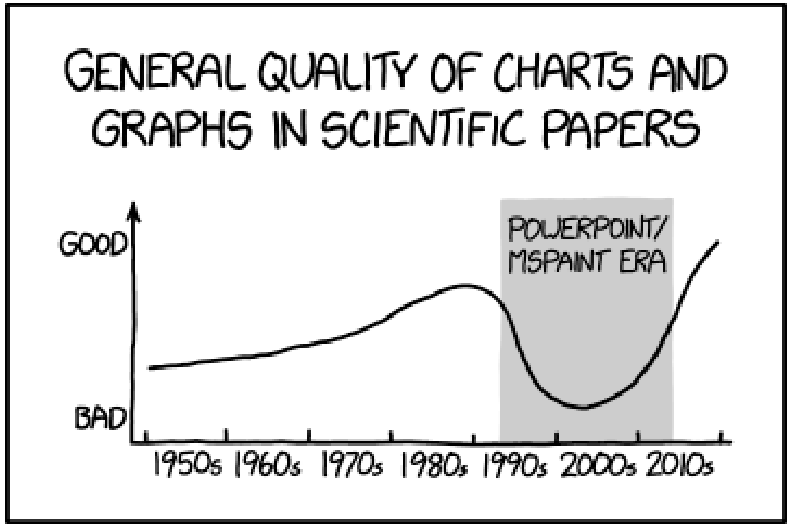
Gapminder tools:
https://www.gapminder.org/tools

> library('gapminder')

A very brief history of data visualization

Currently undergoing a “Graphical re-birth”





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/