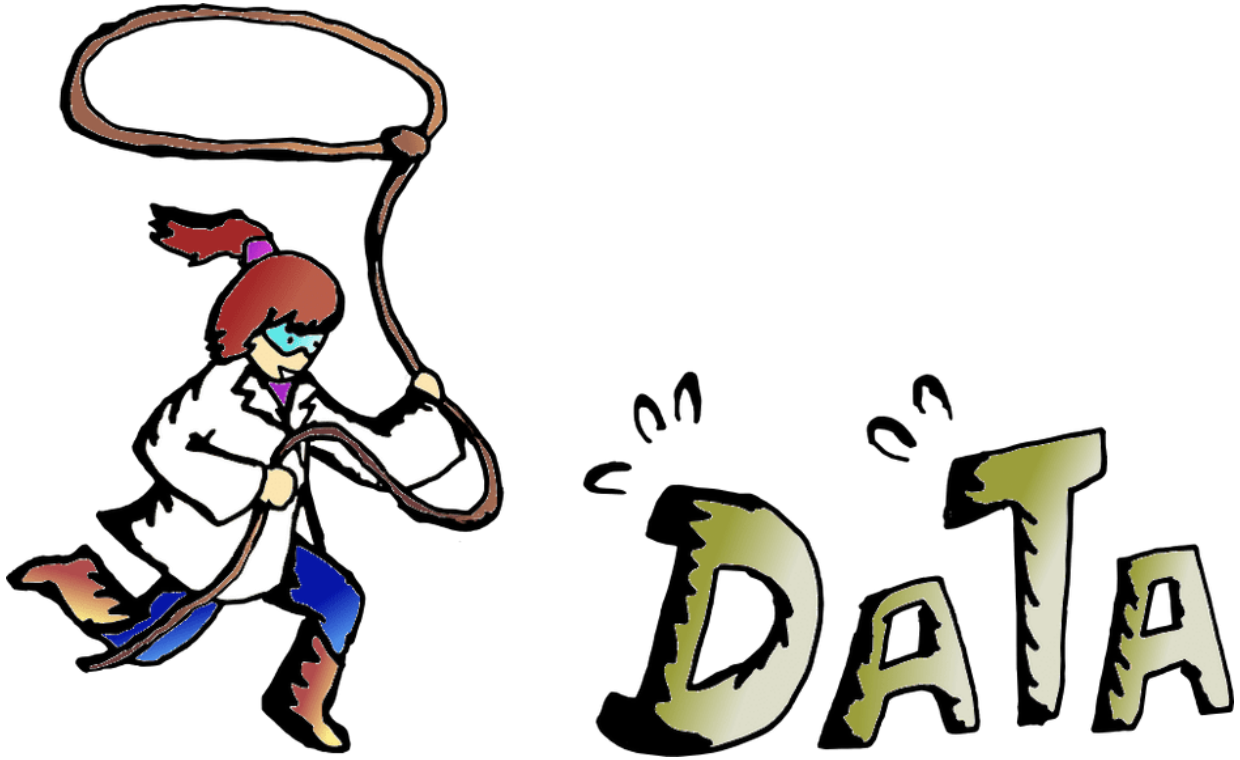
Data wrangling/manipulation



Overview

Very quick review of connections in hypothesis tests and confidence intervals

Data wrangling/manipulation with dplyr

Brief history of data visualization

Announcements

A practice midterm exam has been posted

- Midterm will be a written exam taken in class on Thursday 10/13

Homework 5 has been posted

- I strongly recommend you do the first two parts prior to next class

Kickoff event for Yale's Institute for Foundations of Data Science

Event will take place from 1-4pm of Friday, October 14 O. C. in Marsh Lecture Hall

The technical content will consist of two one-hour rounds of rapid-fire talks illuminating the breadth of data science research at Yale, beginning with remarks from Peter Salovey and Scott Strobel.

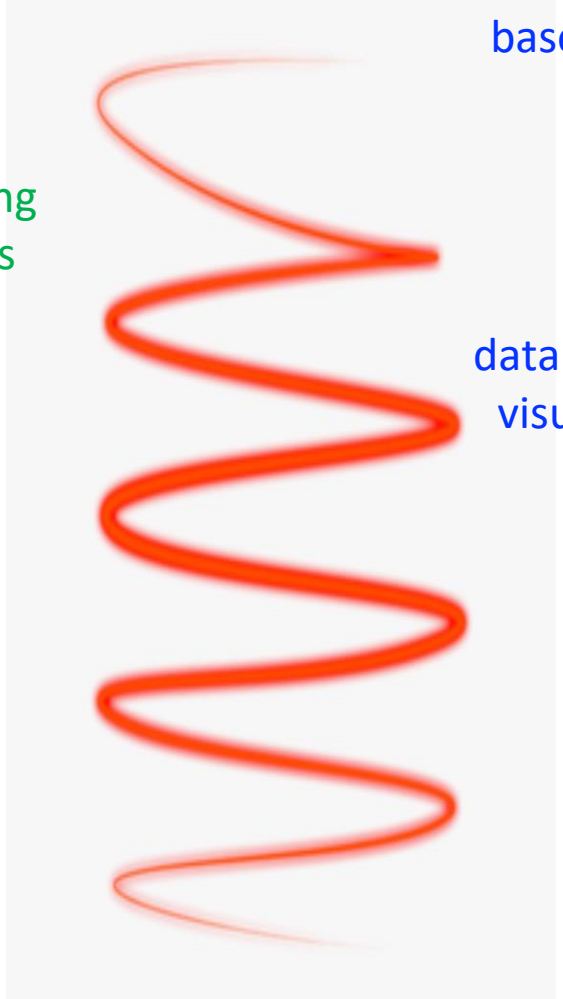
- It will be fast and fun!
- If you would like to attend the kickoff, you can RSVP using this link

If you are interested in helping run the kickoff, there is funding to be paid to help ($15 / hour)

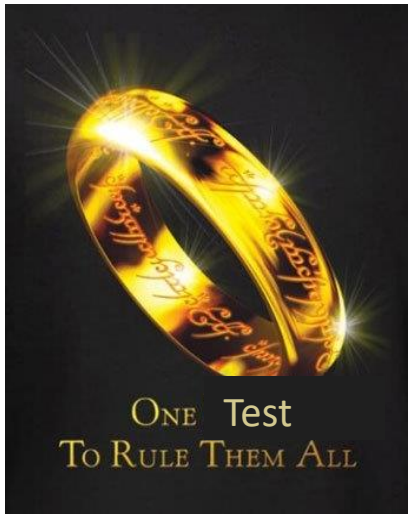
- Please email Emily Hau (emily.hau@yale.edu) if you're interested.

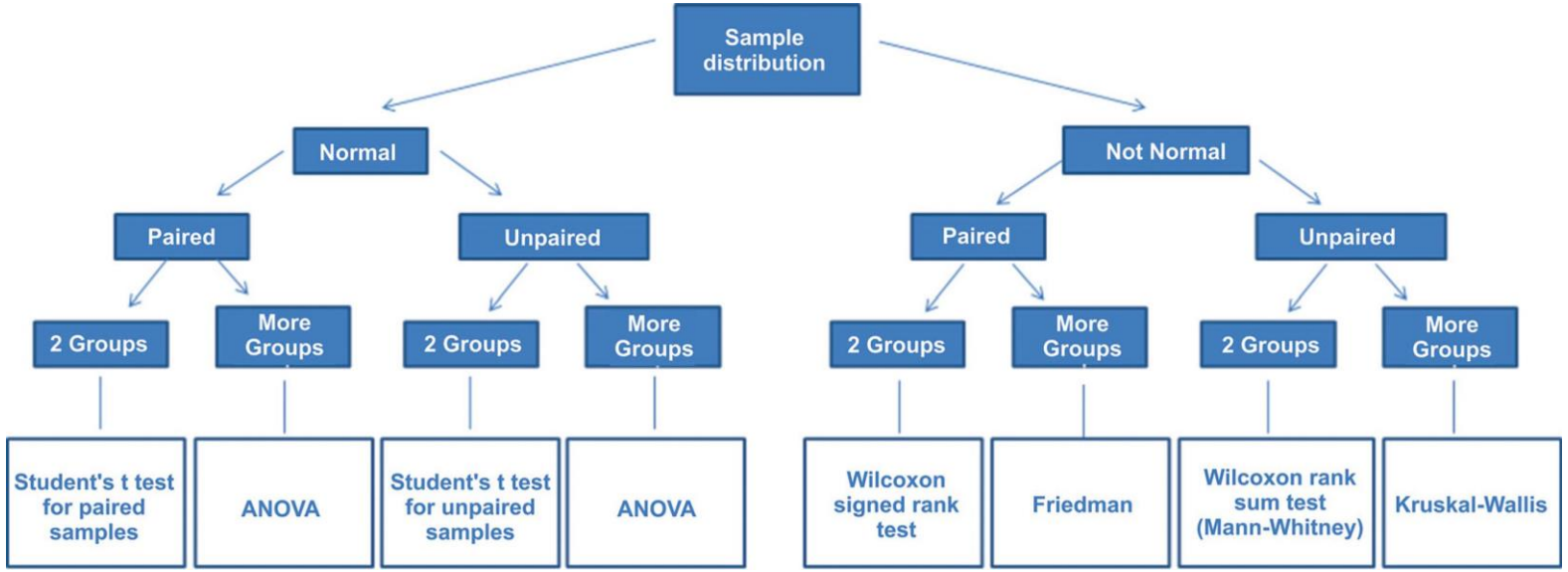
Plan for the semester

			Analysis	R
1	Sep 1	Course overview, introduction to R, descriptive statistics		base R
2	Sep 6-8	Review of central statistical concepts and exploratory analysis using R	resampling methods	
3	Sep 13-15	Confidence Intervals and the bootstrap		
4	Sep 20-22	Review of hypothesis tests and permutation tests in R		data wrangling visualization
5	Sep 27-29	Parametric, non-parametric and theories of hypothesis testing		
6	Oct 4-6	Data manipulation and visualization		
7	Oct 11-13	Review and midterm exam		
8	Oct 18-22	Joining and mapping, October break		

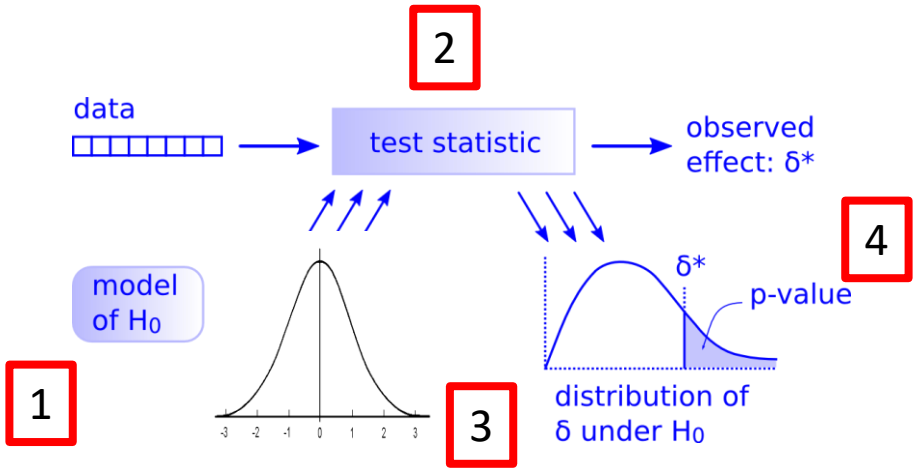


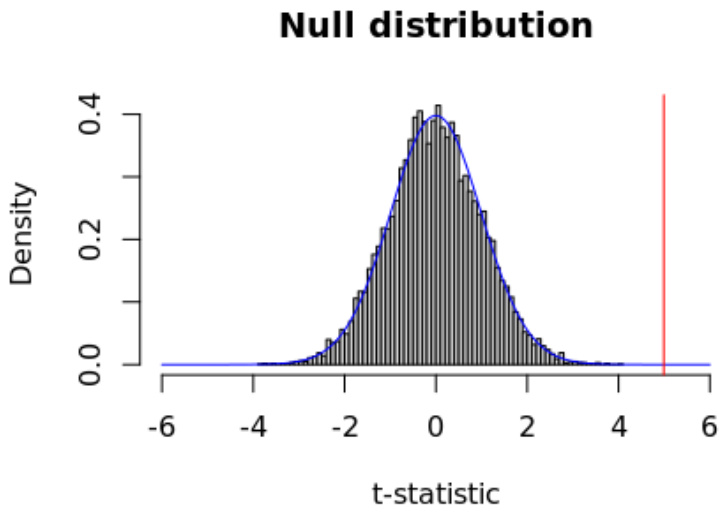
Very quick review



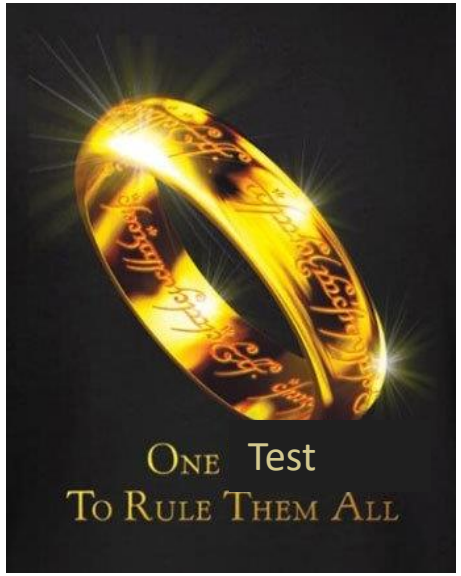


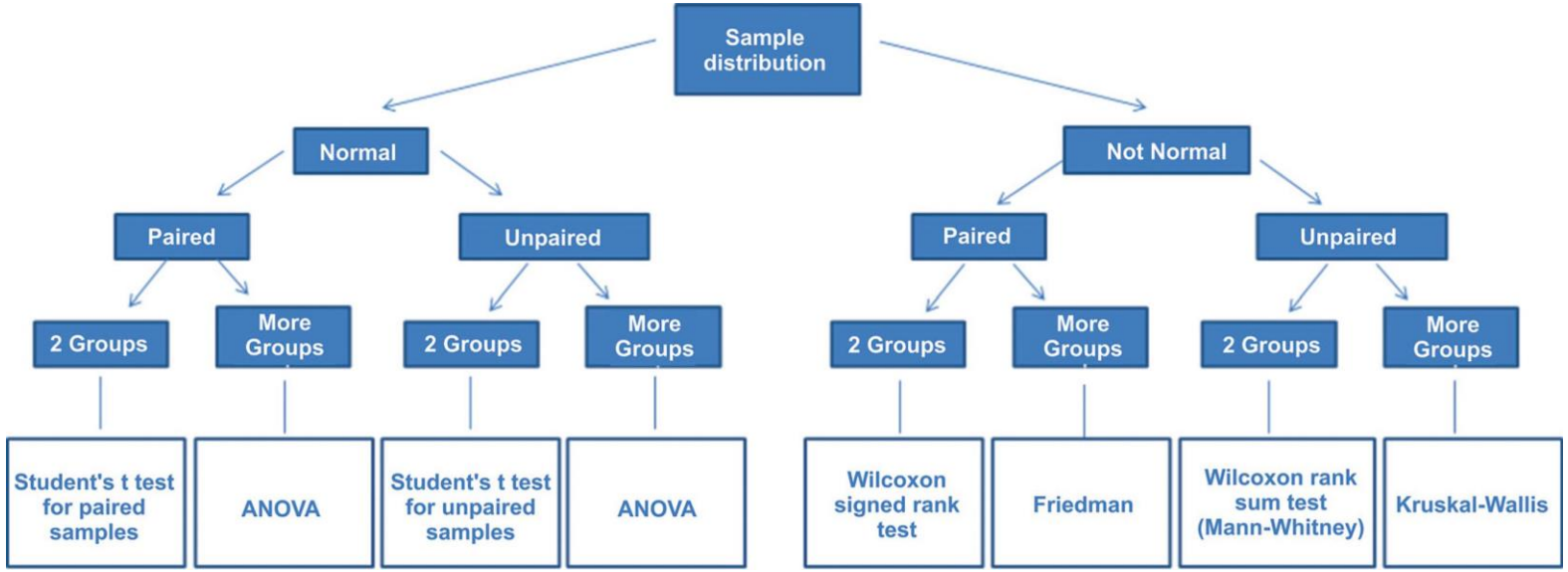
Just need to follow 5 steps!





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

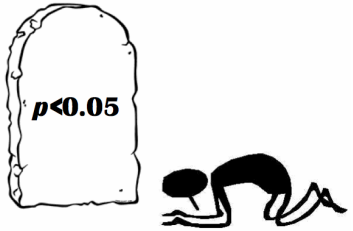
- E.g., data from the two samples comes from normal populations with the same variance
- Some hypothesis tests are "robust" to violations of these assumptions
  - The robustness can be evaluated this through computer simulations

Very quick review: theories of hypothesis testing

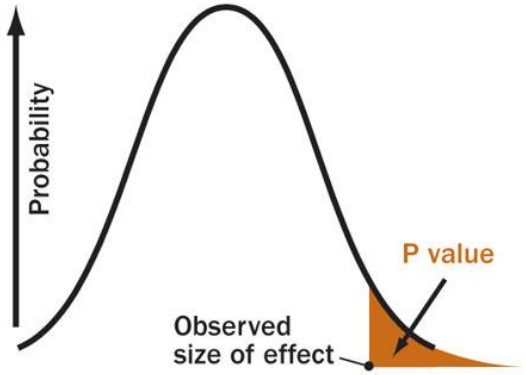
Fisher (1890-1962)

Neyman (1894-1981)

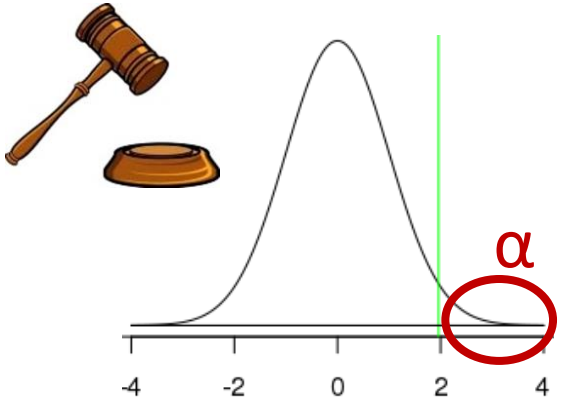
Pearson (1895-1980)

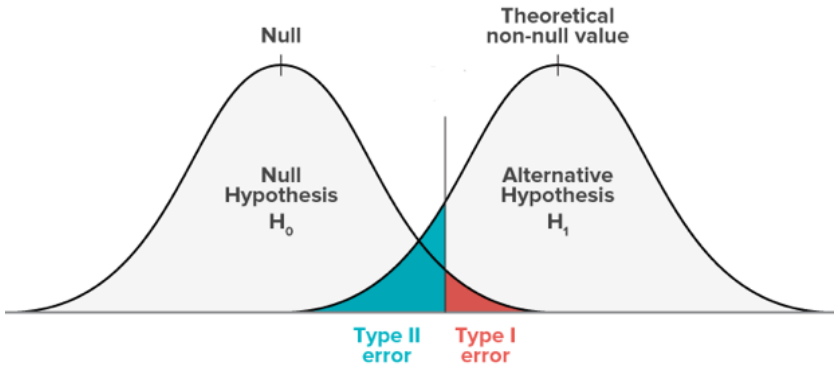


p-value a strength of evidence

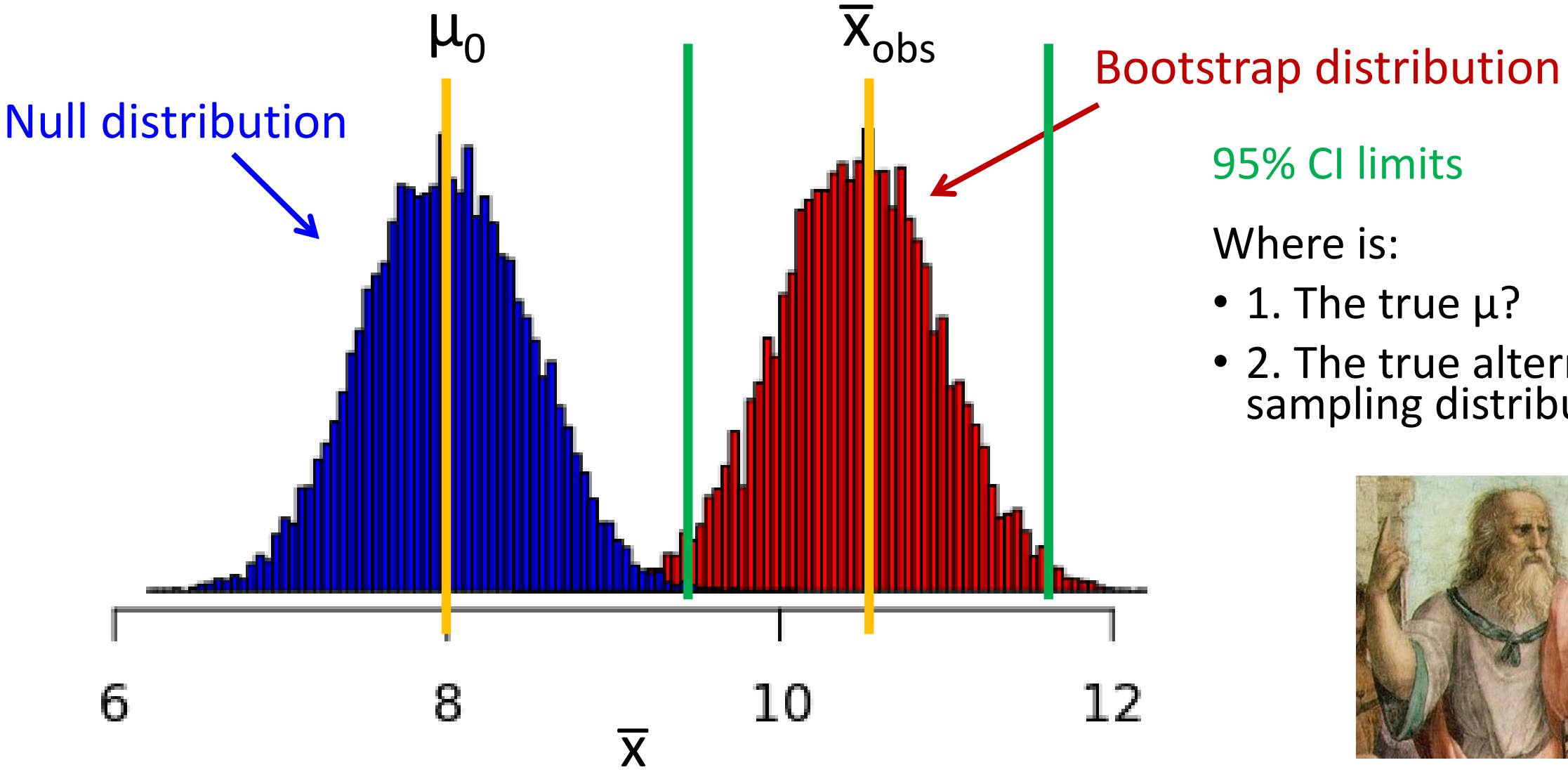


Use p-value to make a decision



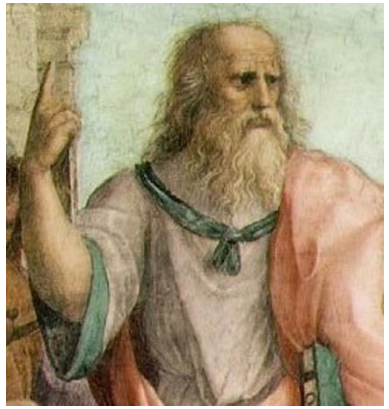


Relationship between null and bootstrap distributions



Where is:

- 1. The true μ?
- 2. The true alternative sampling distribution?



Questions?

The tidyverse and dplyr

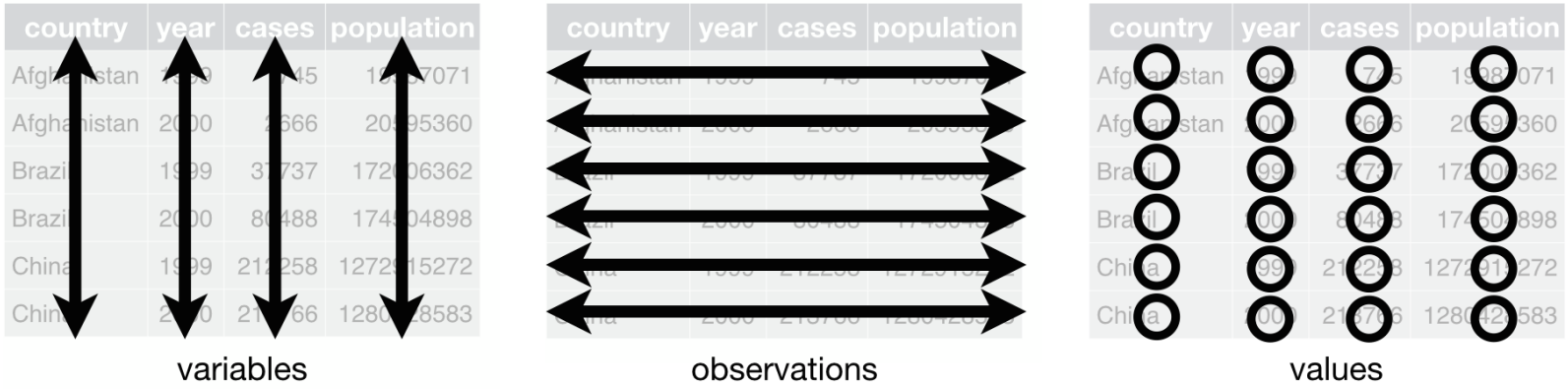
The ‘tidyverse’

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…

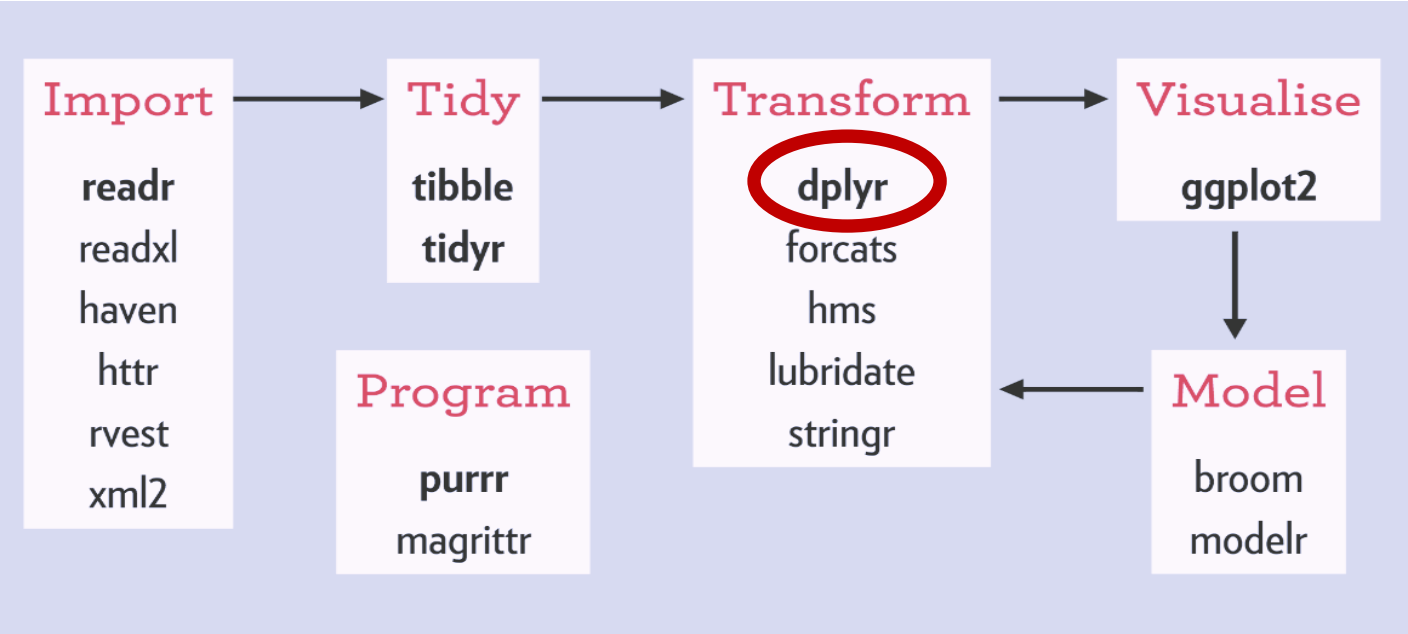
What would be an example of data that is not tidy?

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			

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

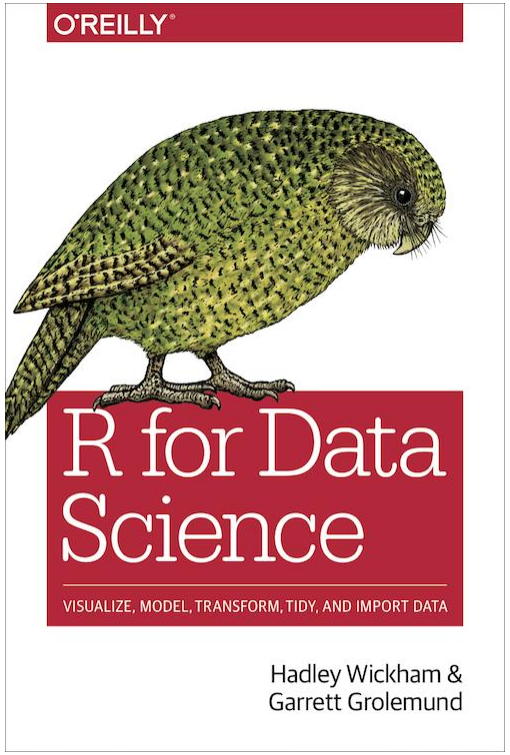
dplyr is a package that has a set of verbs that are useful for transformations data:

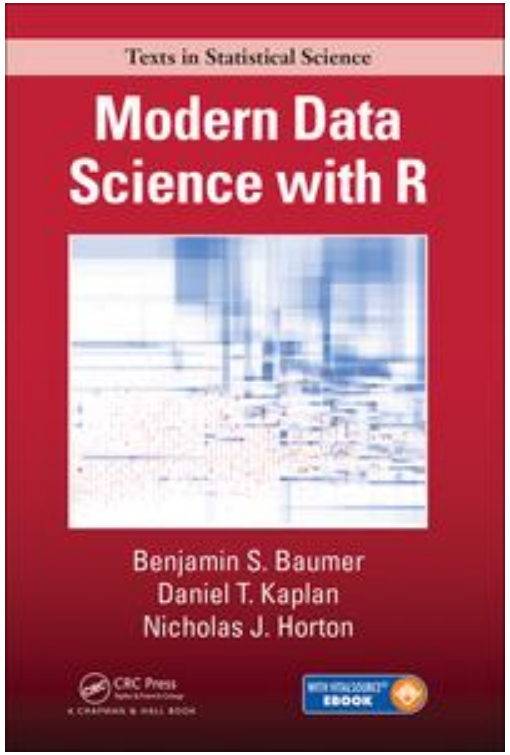
1. filter()
2. select()
3. mutate()
4. arrange()
5. group_by()
6. summarize()

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

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

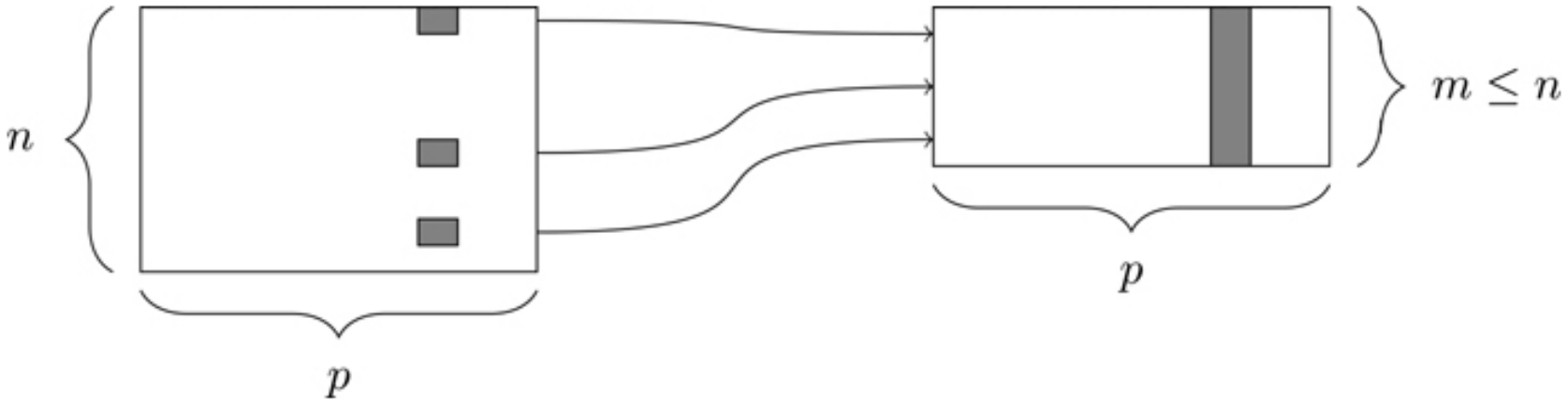
Quick overview of the dplyr functions





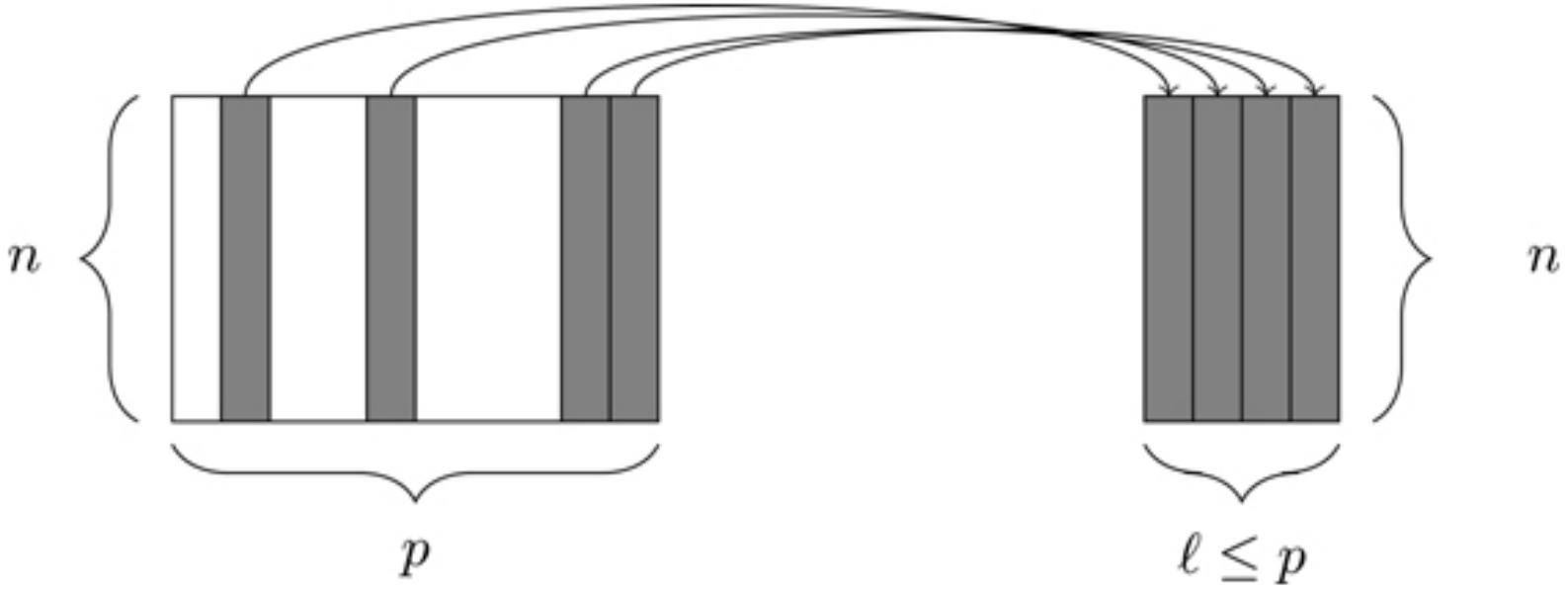
1. filter()

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



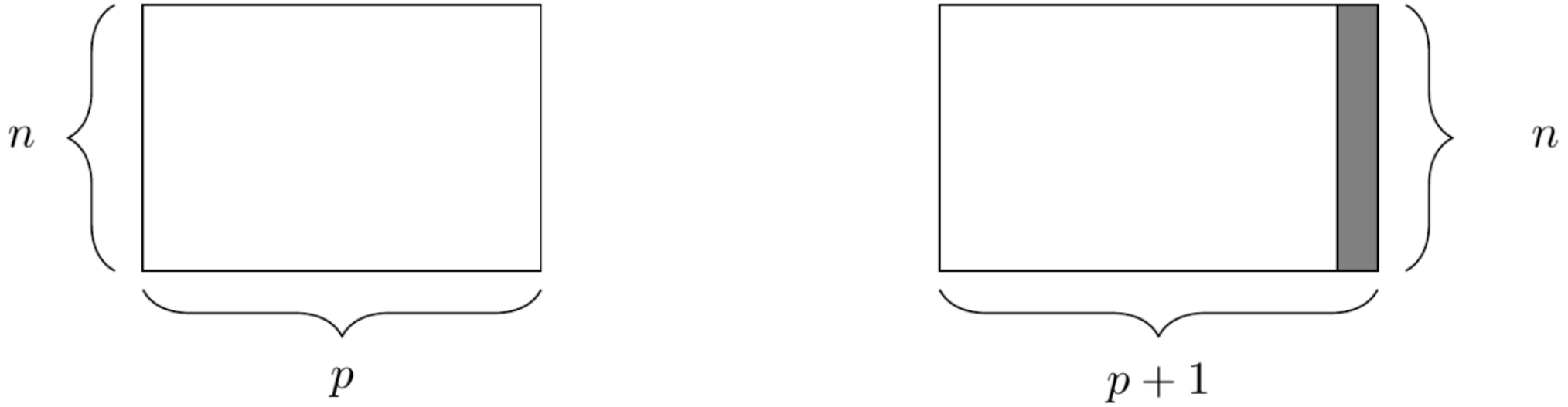
2. select()

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



3. mutate()

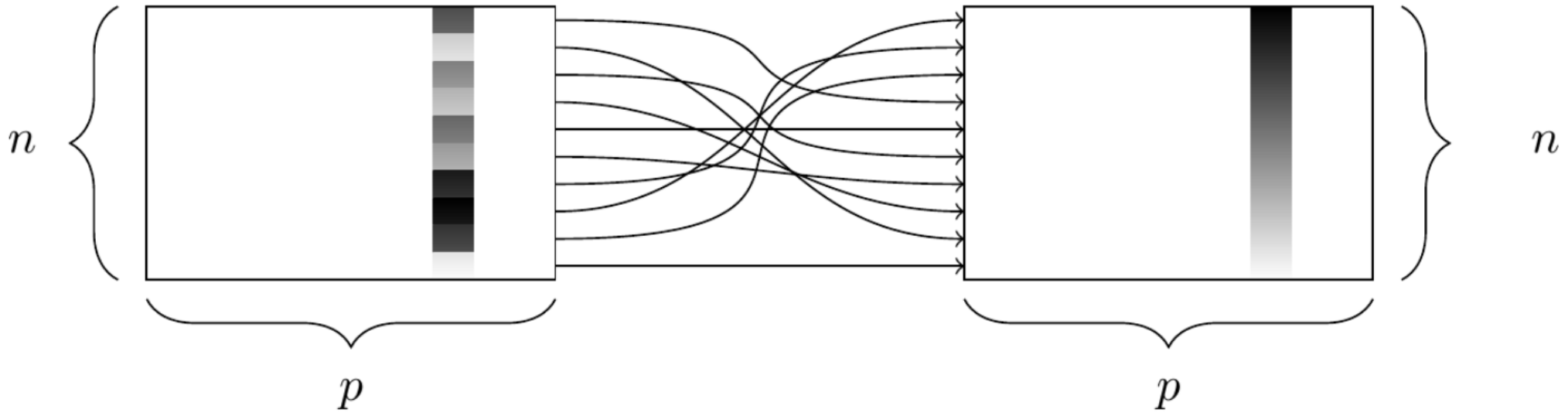
The mutate() function allows you to create new columns that are functions of existing columns



4. arrange()

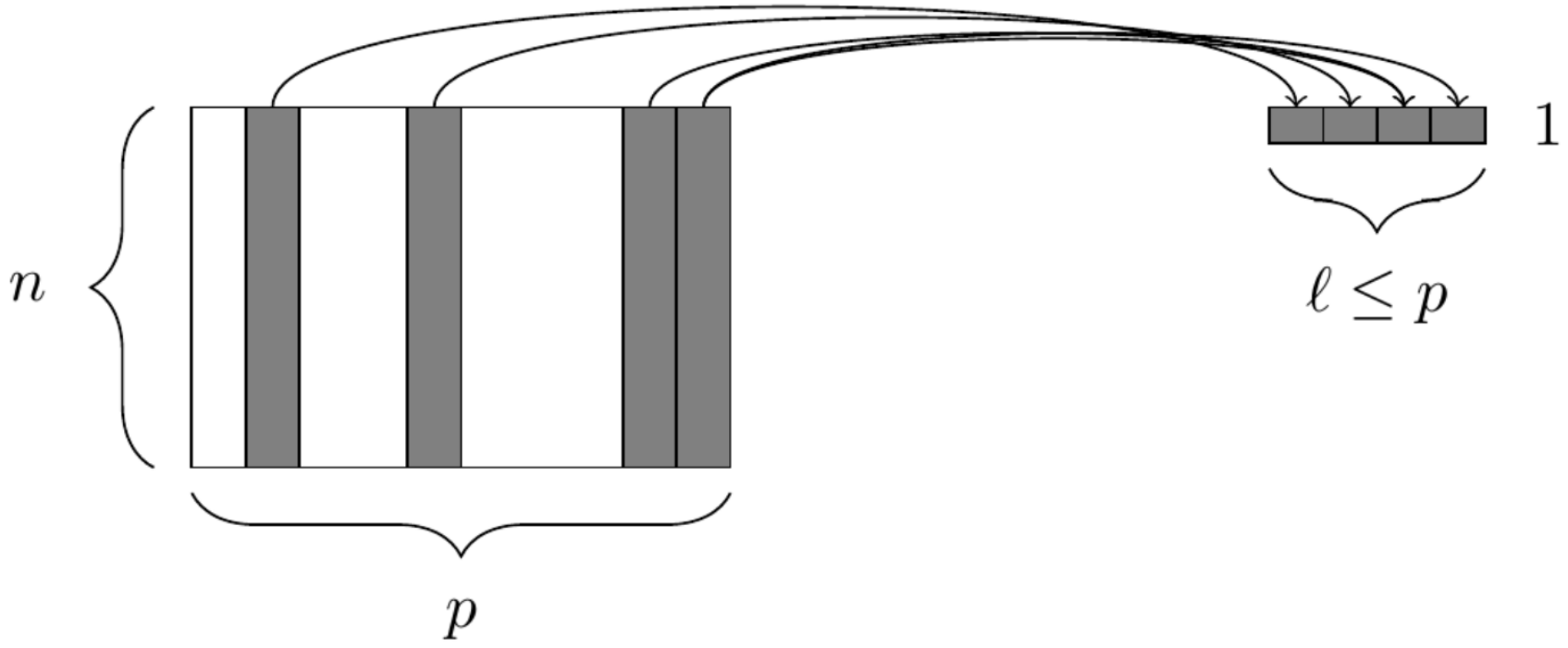
The arrange() function arranges the rows based values in a column

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



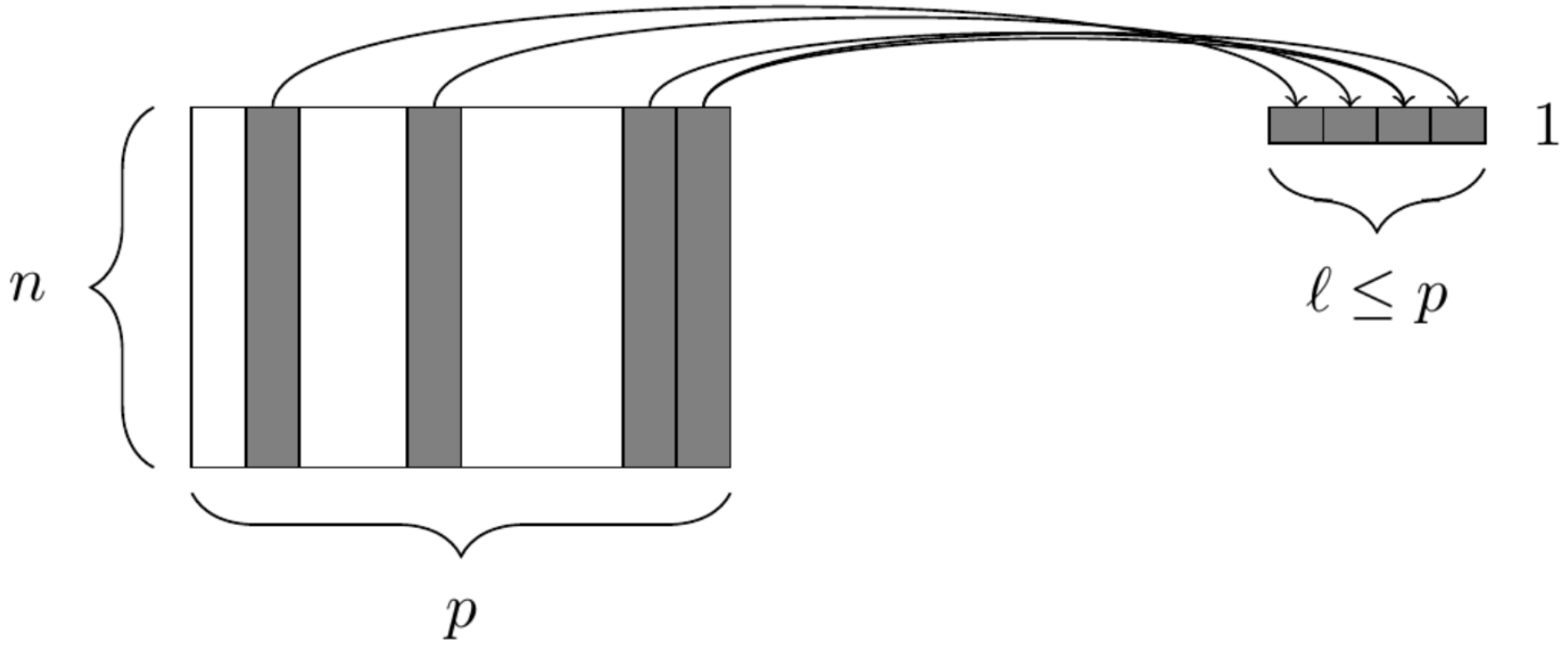
5. summarize()

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



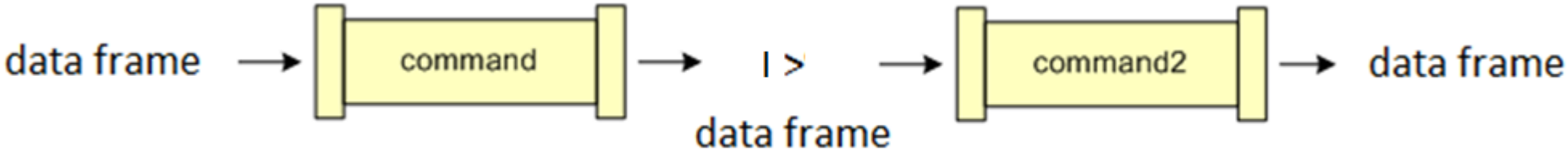
6. The group_by() function

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

The Golden Age of Statistical Graphics

Michael Friendly

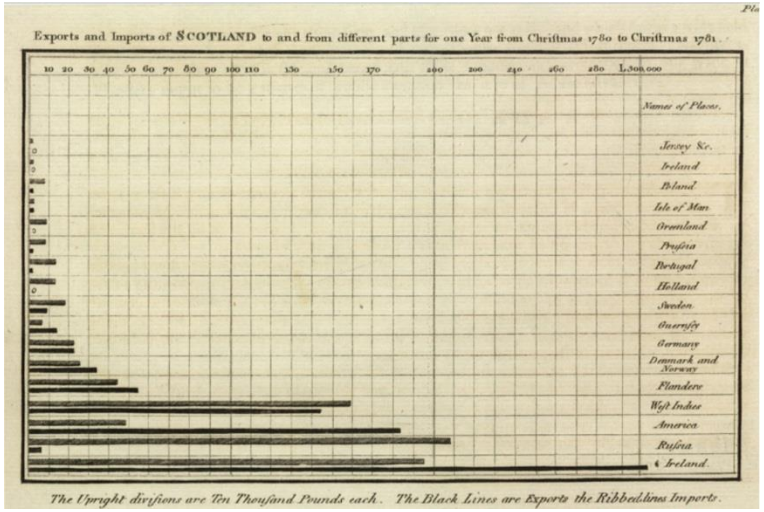
Data visualization

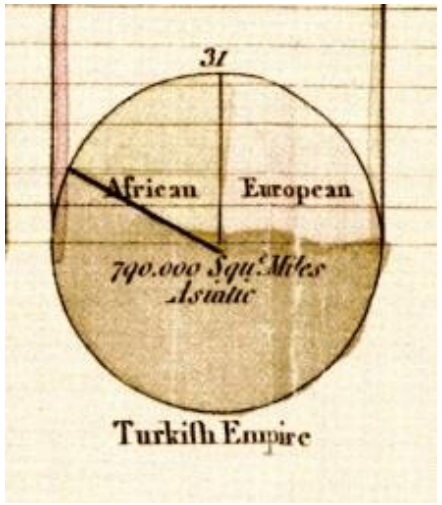
Q: What are some reasons we visualize data rather than just reporting statistics?

A very brief history of data visualization

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

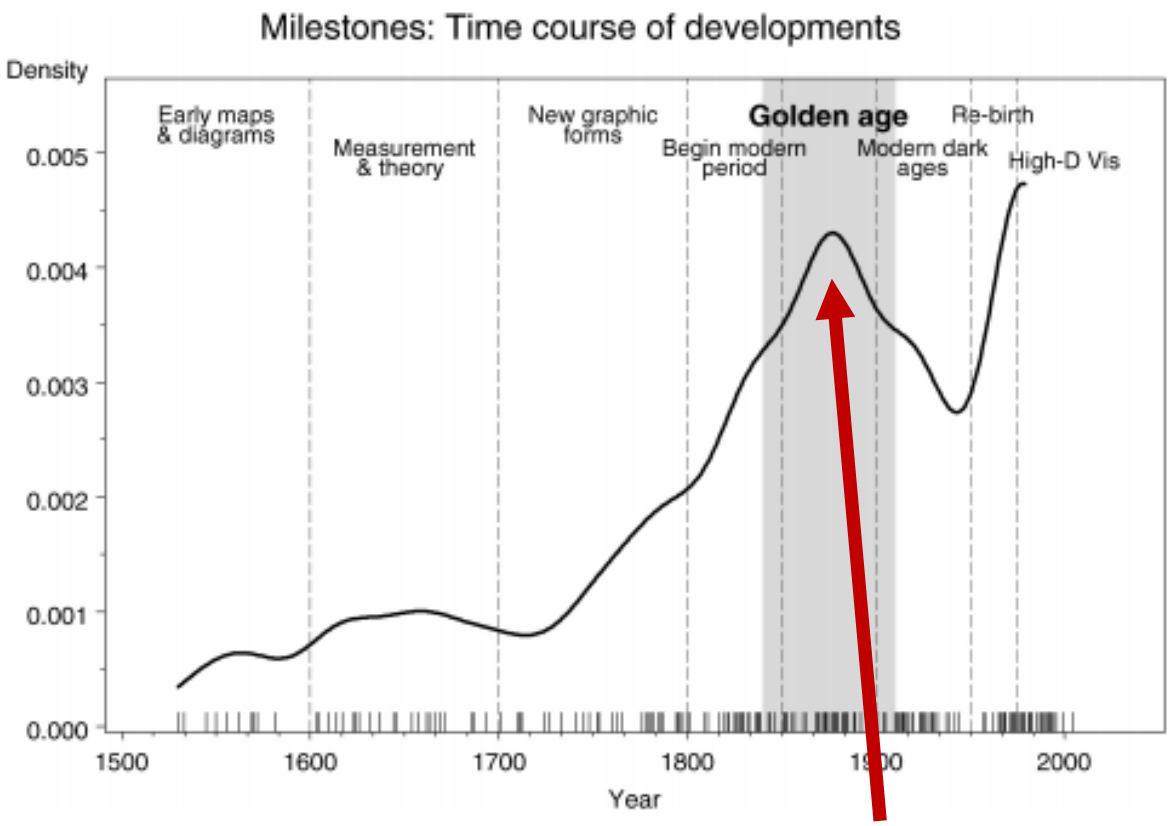
William Playfair (1759-1823) credited with inventing the line graph, bar chart and pie chart





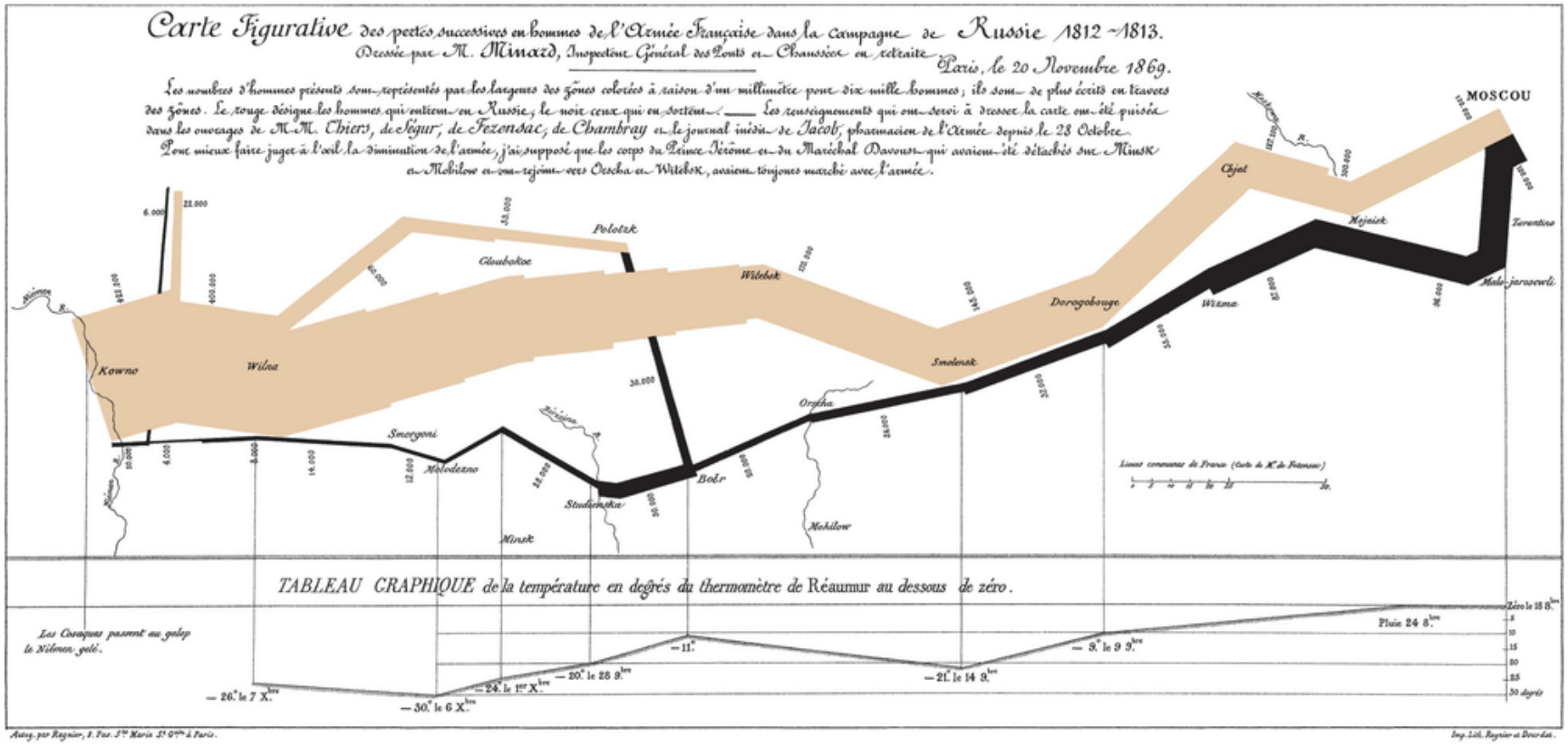
A very brief history of data visualization

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



A very brief history of data visualization

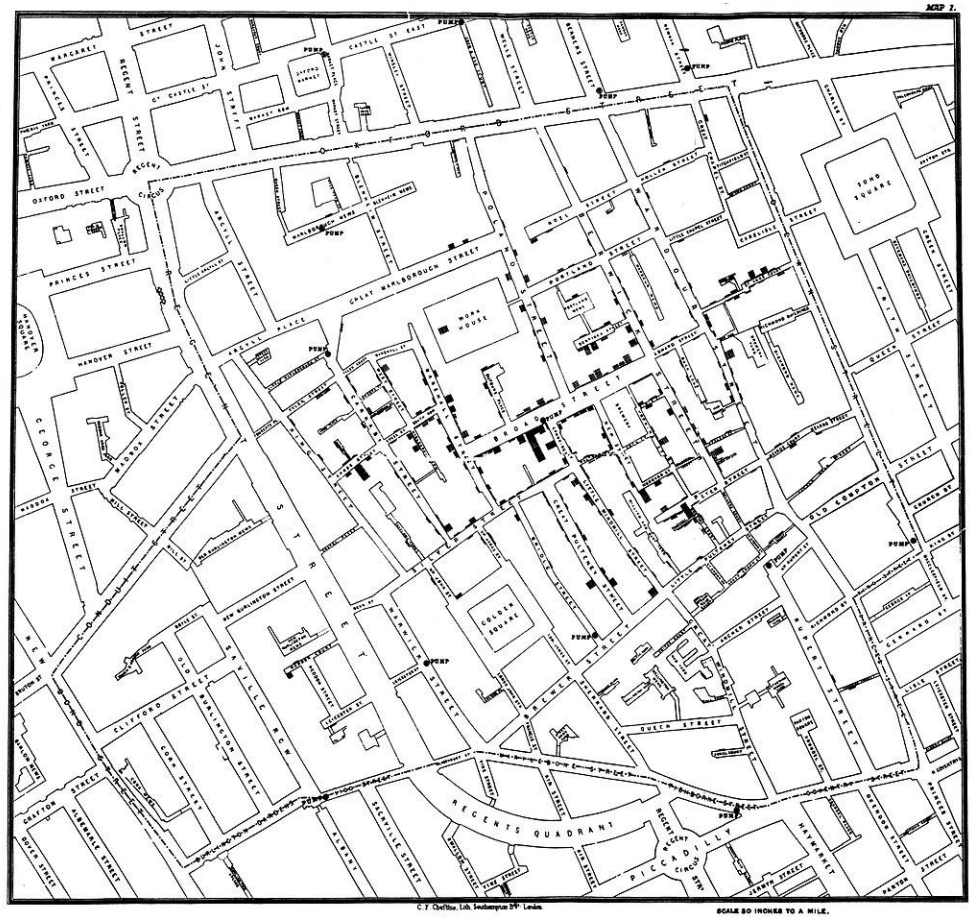
[Joseph Minard](#) (1781-1870)



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)

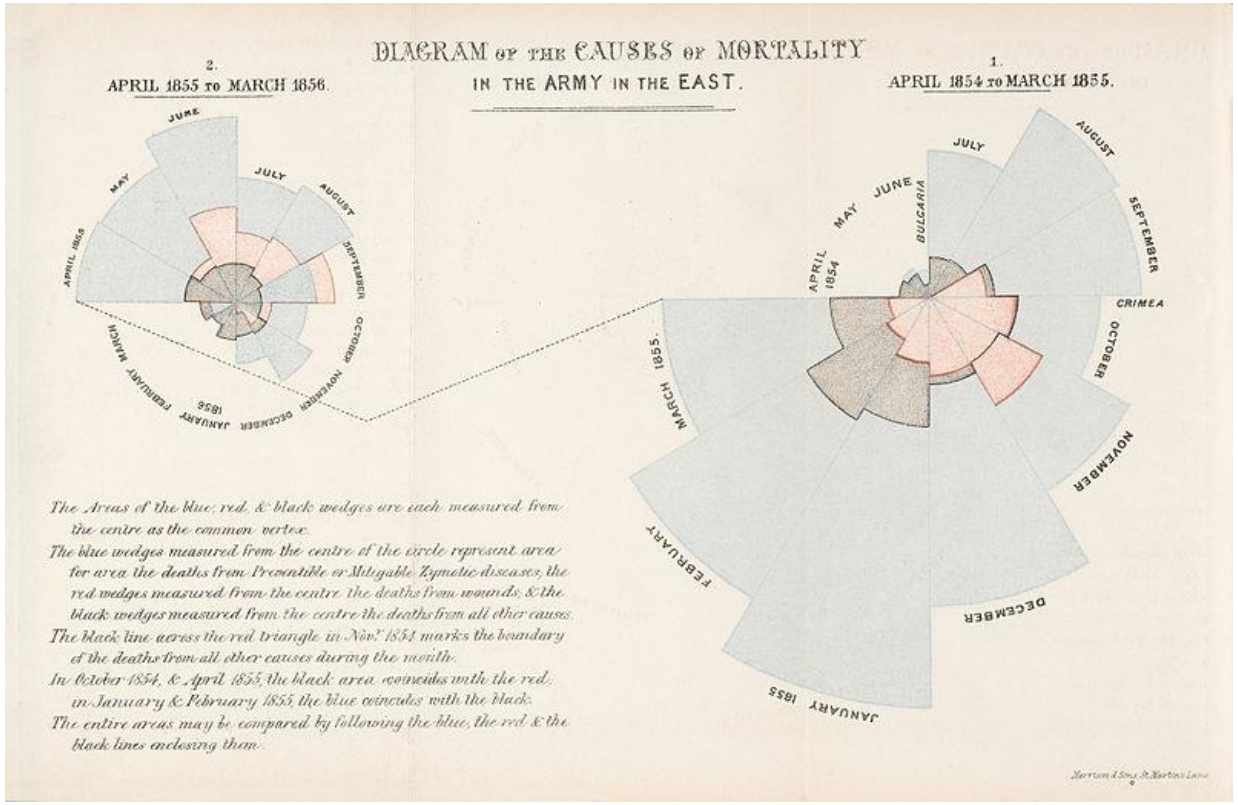
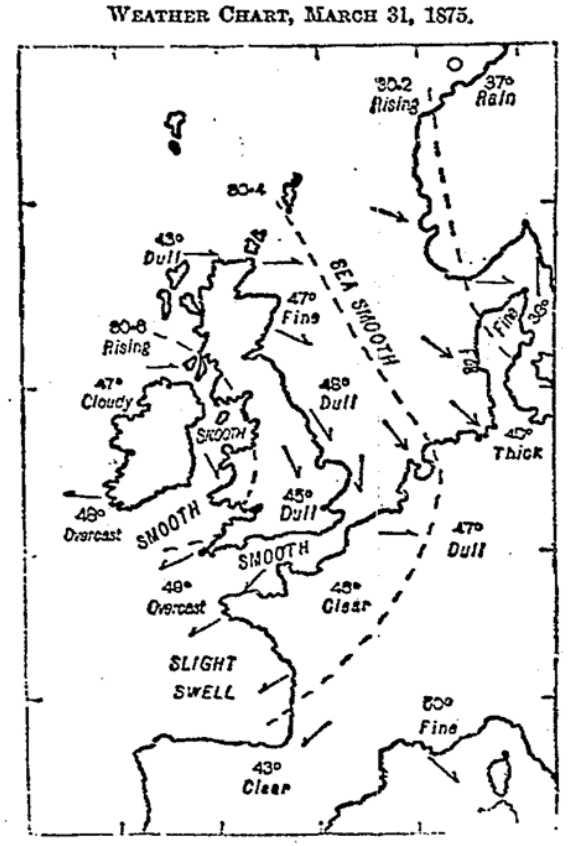


Diagram of the causes of mortality in the army in the east

A very brief history of data visualization

Francis Galton (1822-1911)

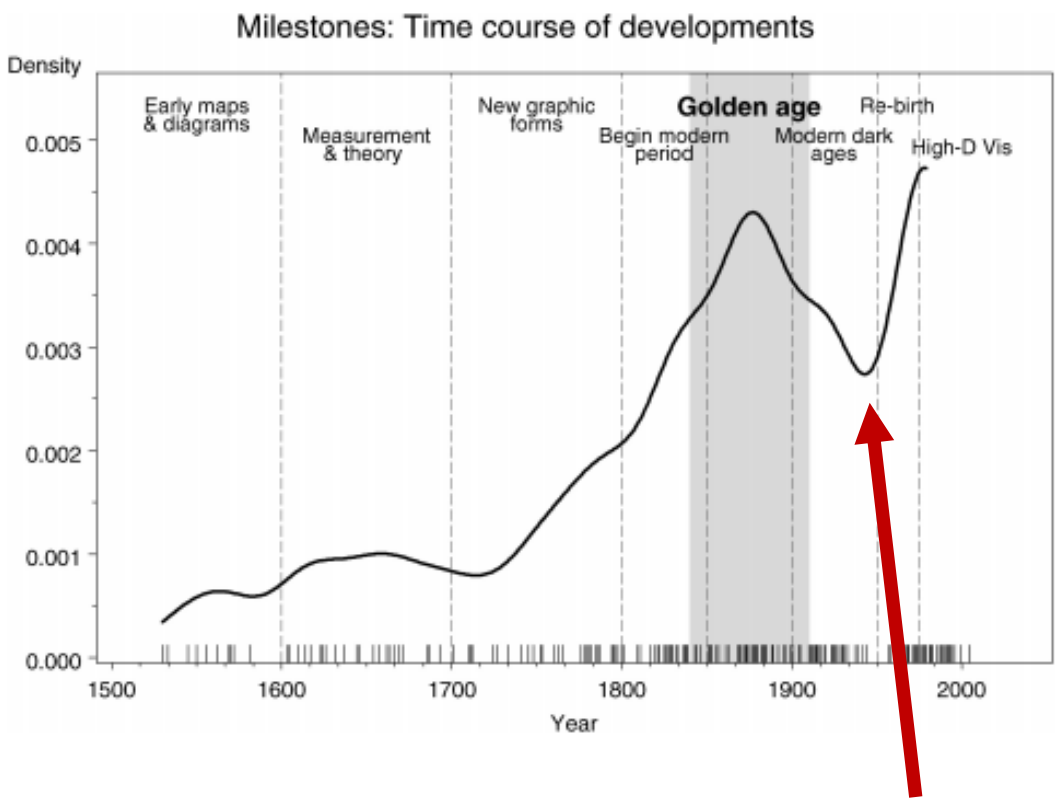


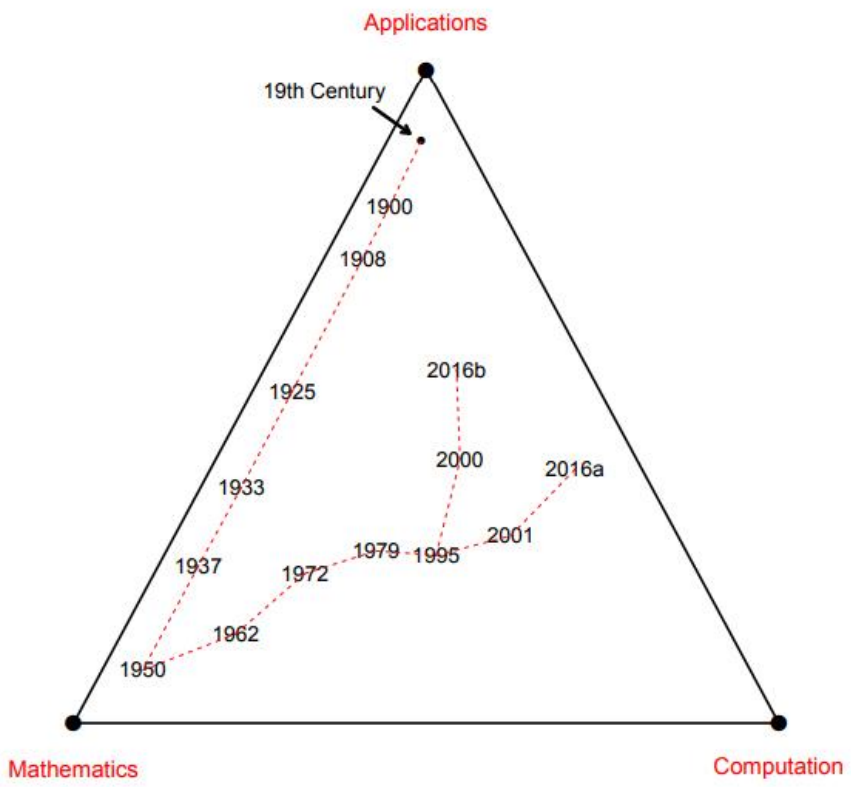
The dotted lines indicate the gradations of barometric pressure The variations of the temperature are marked by figures, the state of the sea and sky by descriptive words, and the direction of the wind by arrows—barbed and feathered according to its force. ⊙ denotes calm.

First weather map published in a newspaper (1875)

A very brief history of data visualization

“Graphical dark ages” around 1950

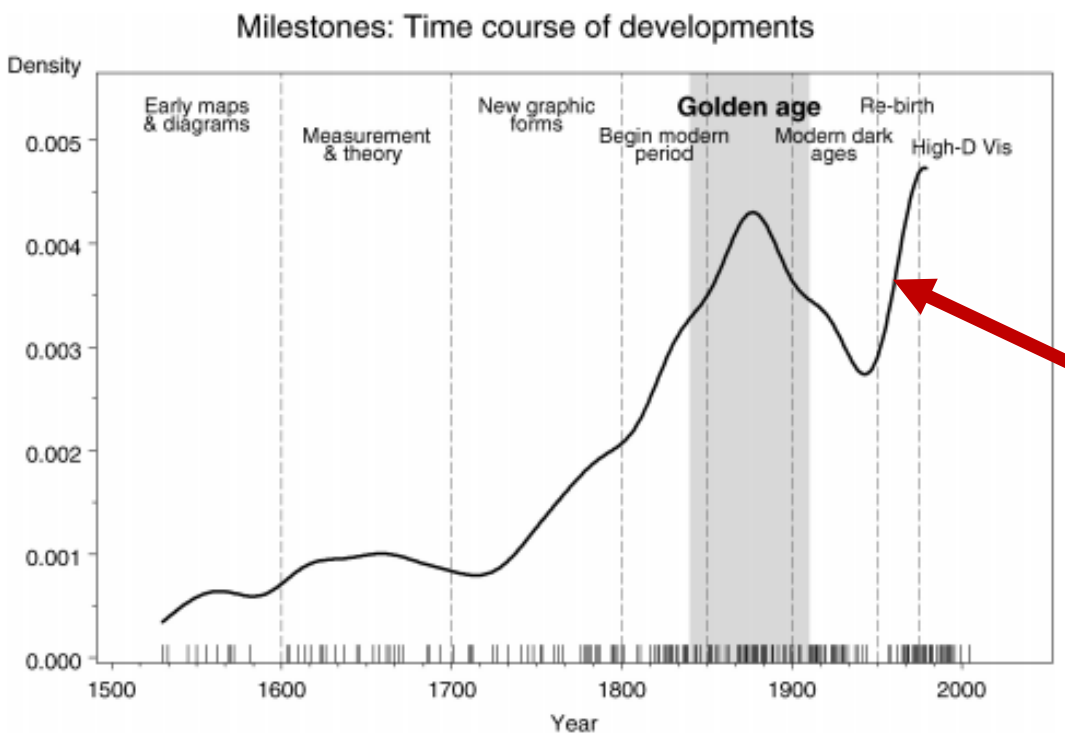




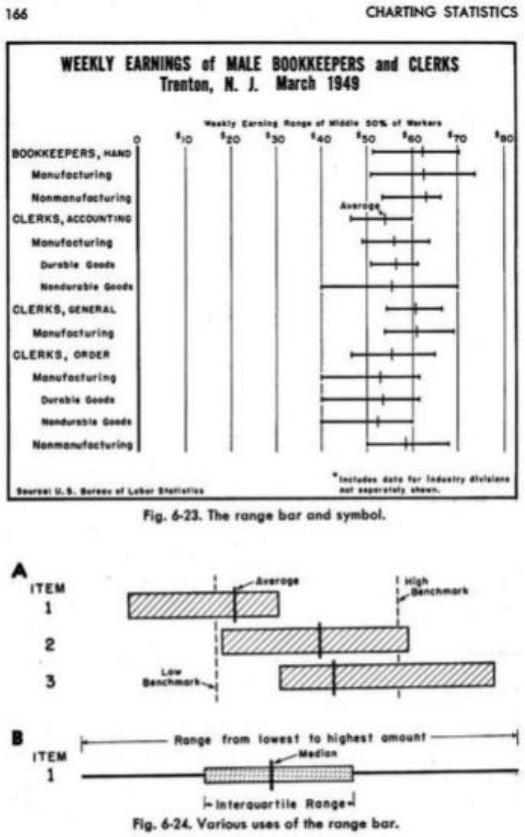
Computer Age Statistical Inference, Efron and Hastie

A very brief history of data visualization

Currently undergoing a “Graphical re-birth”



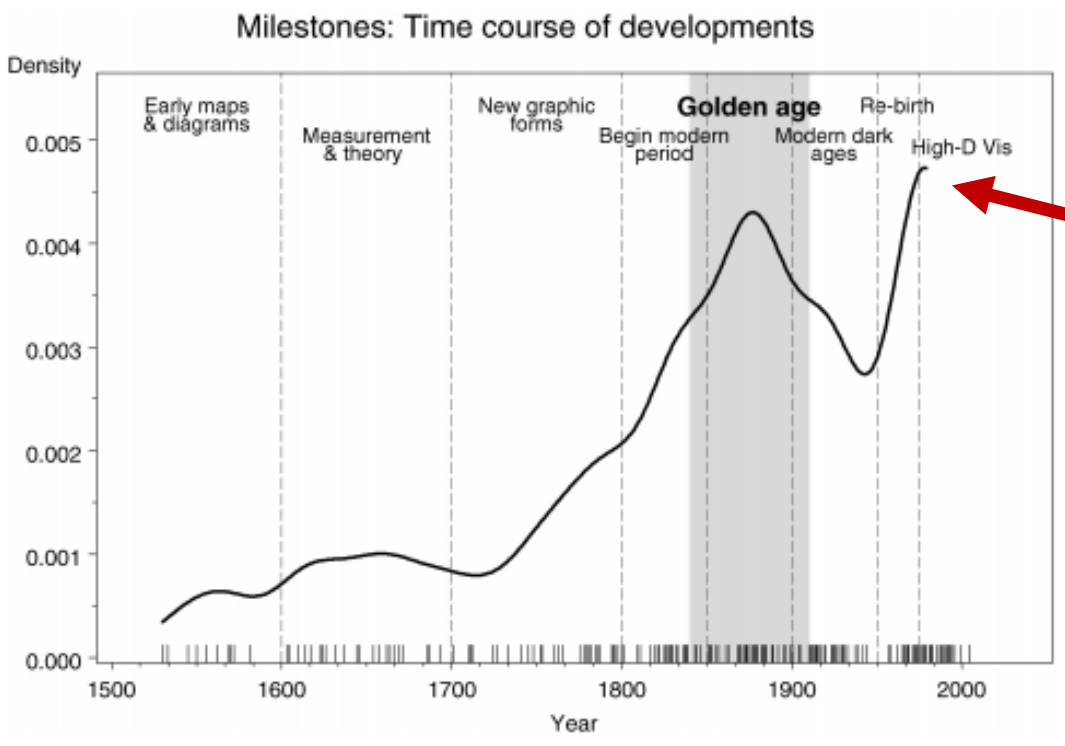
Box plot

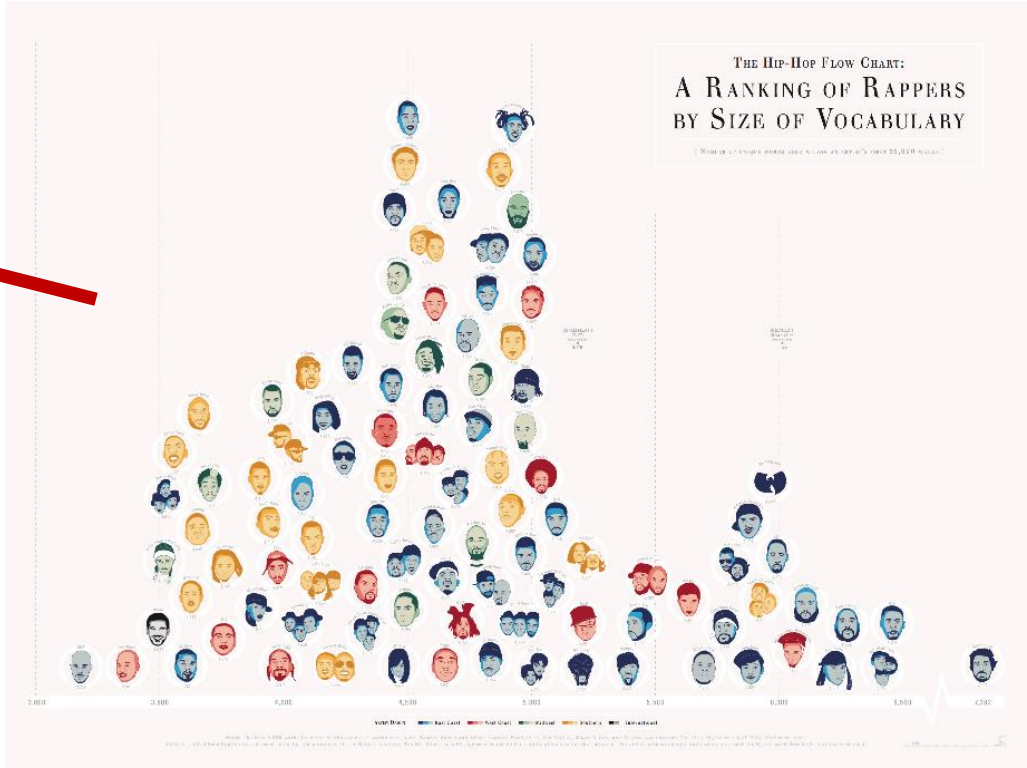


Spear 1952, Tukey 1970

A very brief history of data visualization

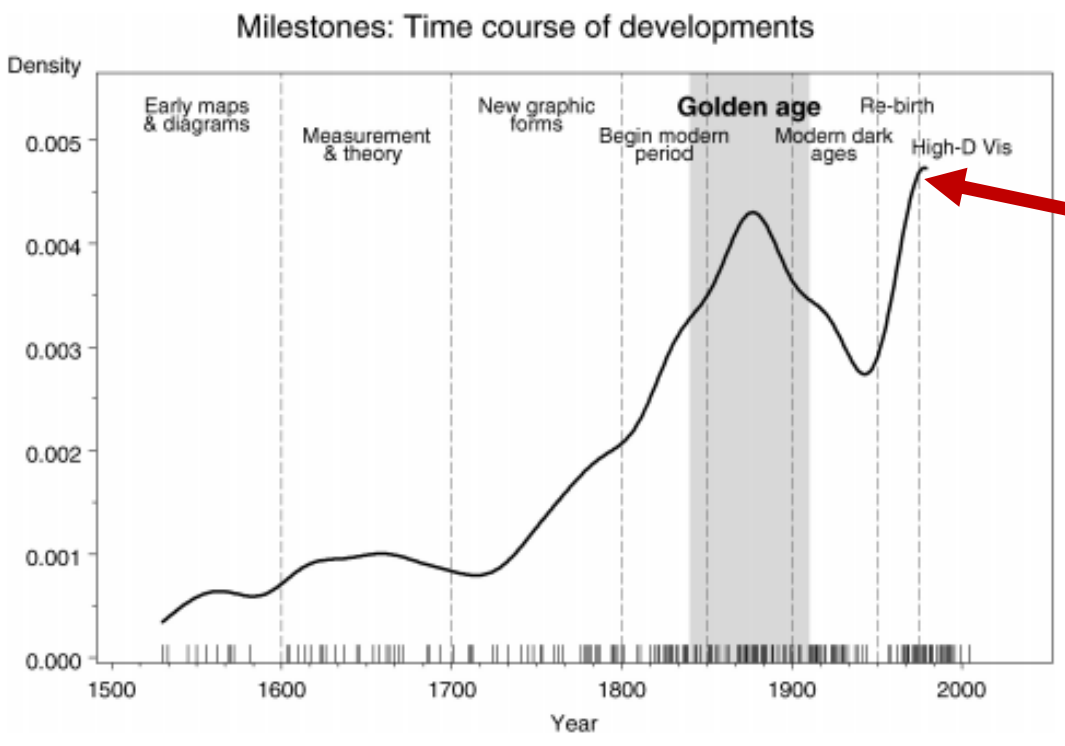
Currently undergoing a “Graphical re-birth”





A very brief history of data visualization

Currently undergoing a “Graphical re-birth”



Hans Rosling’s gapminder

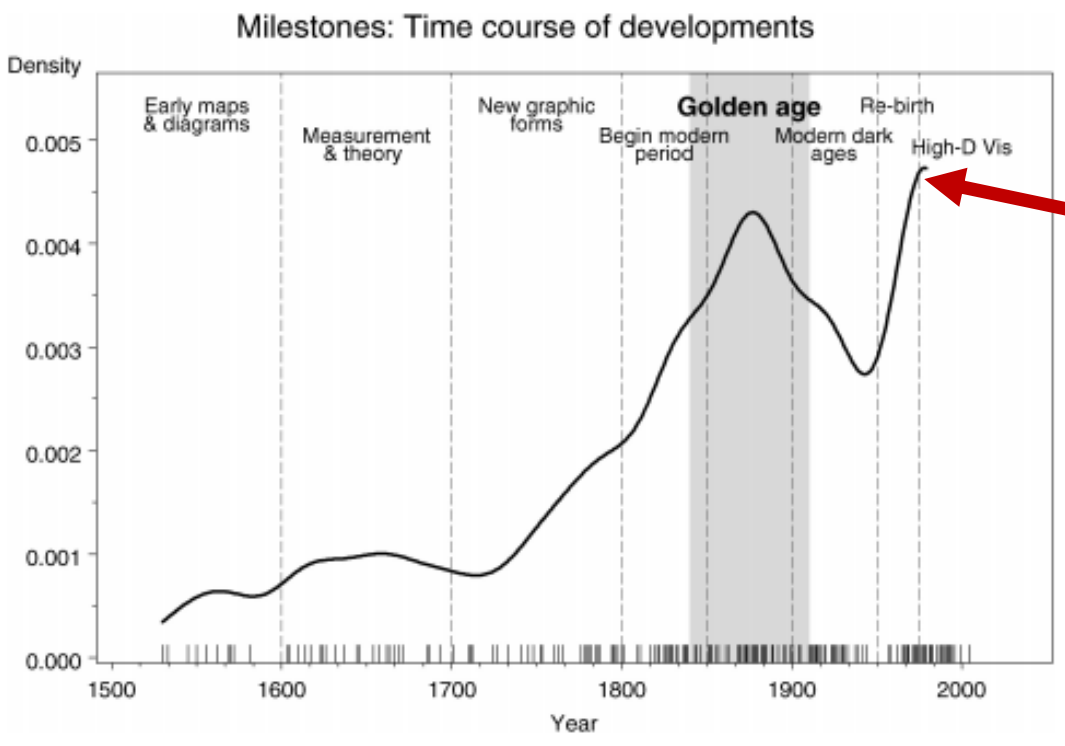
- Simple version
- TV special effects
- Ted Talk

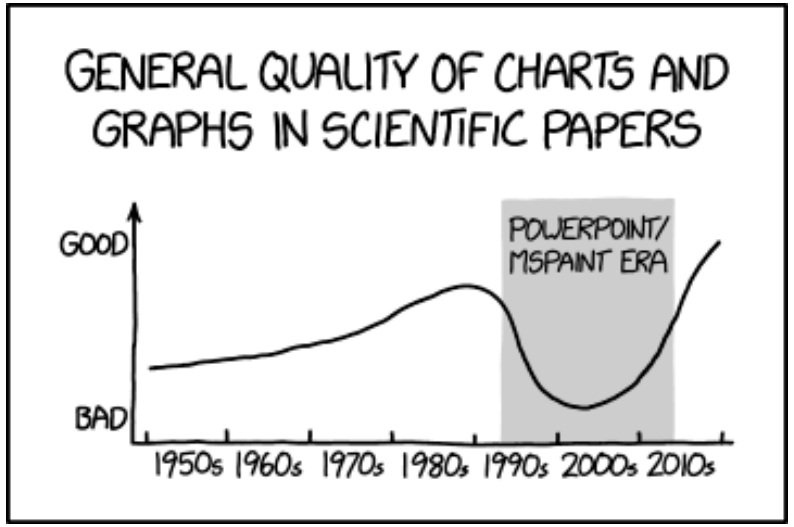
Gapminder tools:
https://www.gapminder.org/tools

```
> library('gapminder')
```

A very brief history of data visualization

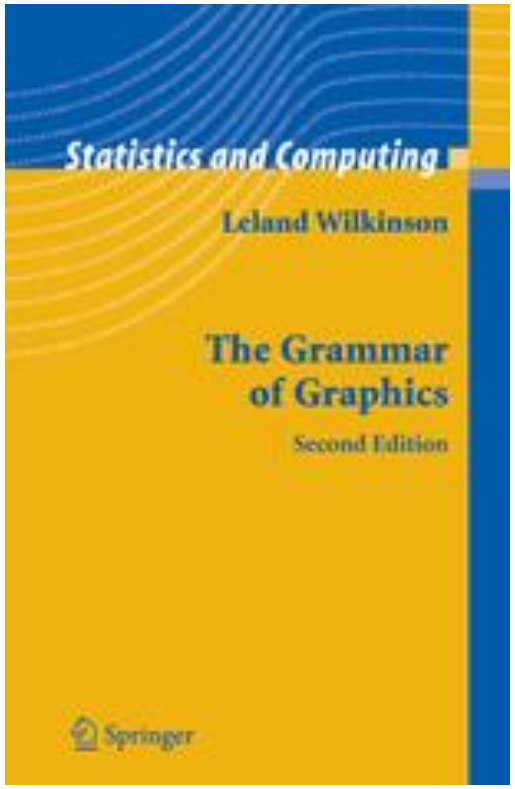
Currently undergoing a “Graphical re-birth”

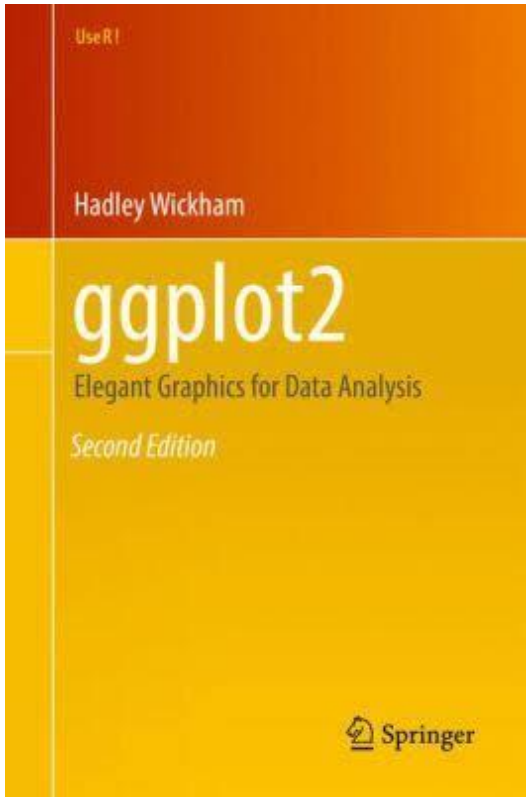




Next class: a grammar of graphics and ggplot

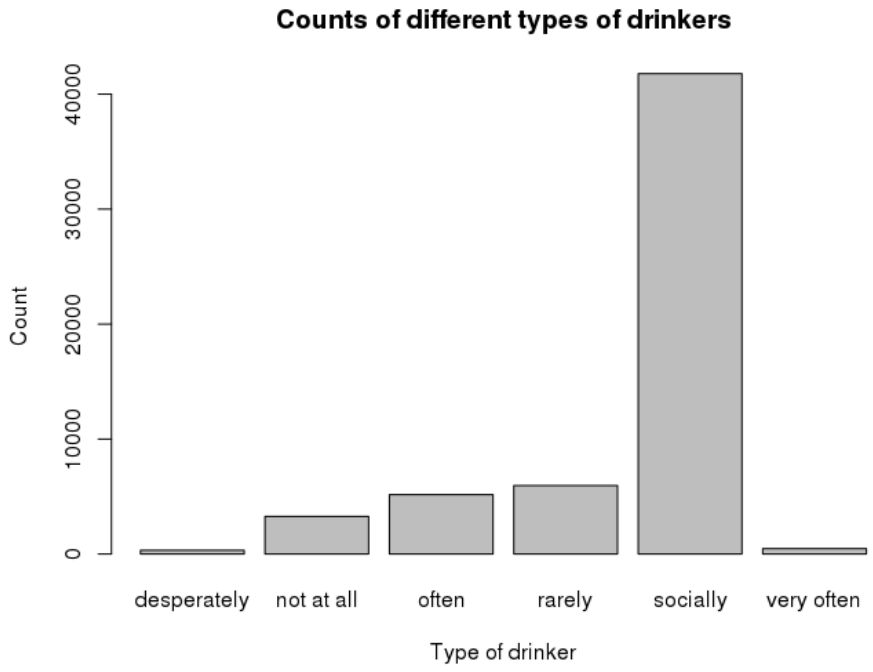
A grammar of graphics and ggplot



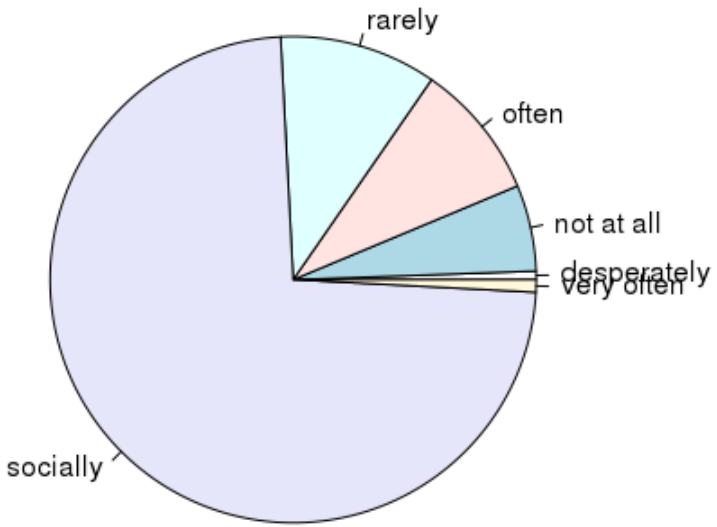


Review: plots of categorical data

Bar plot

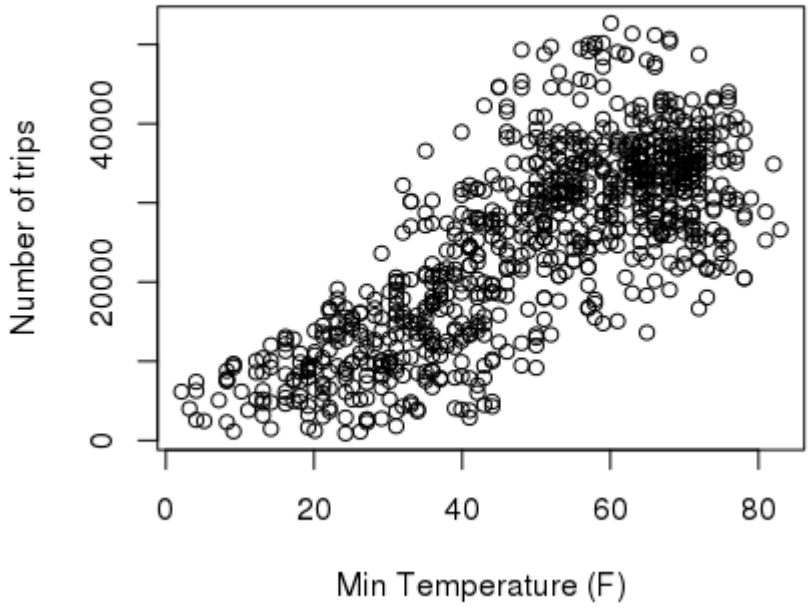


Pie chart

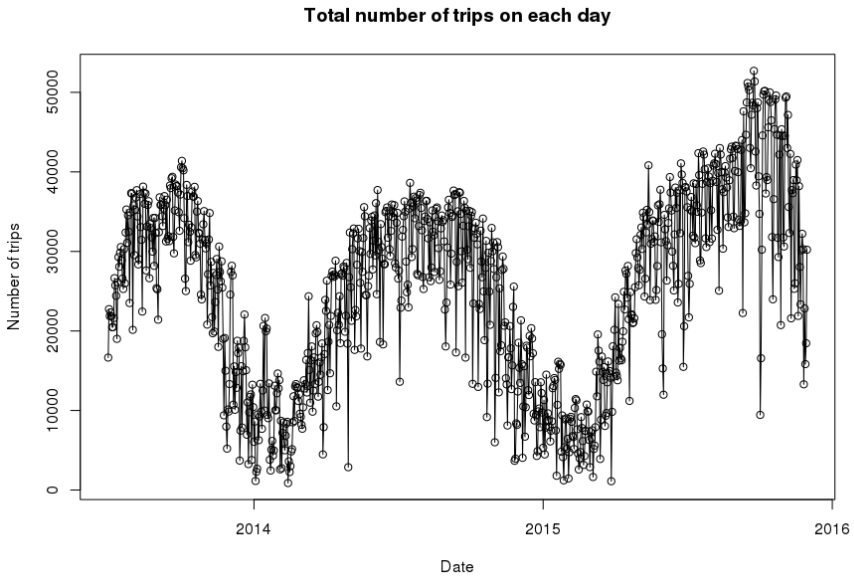


Review: plots of quantitative data

Scatter plots

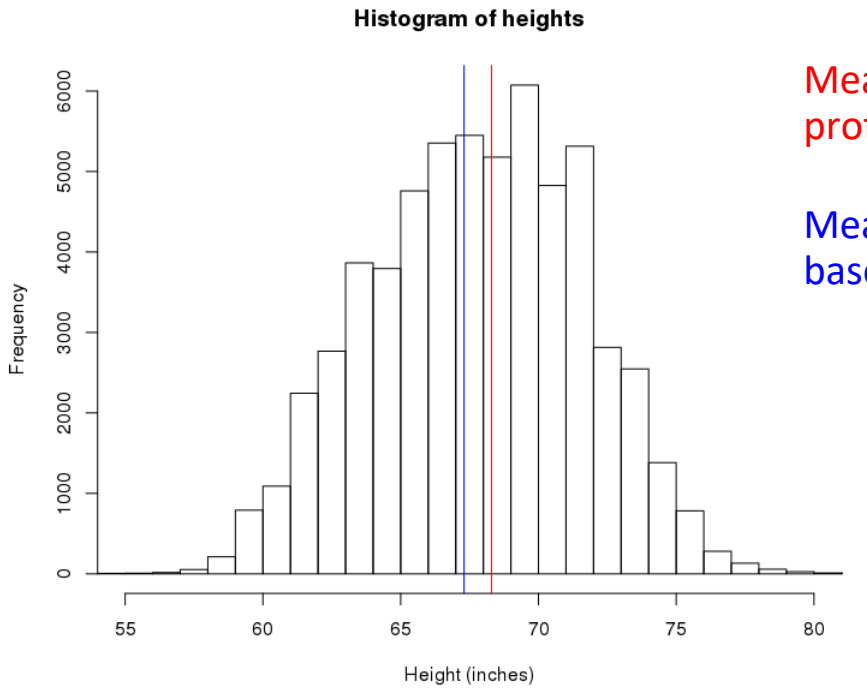


Line chart



Review: plots of quantitative data

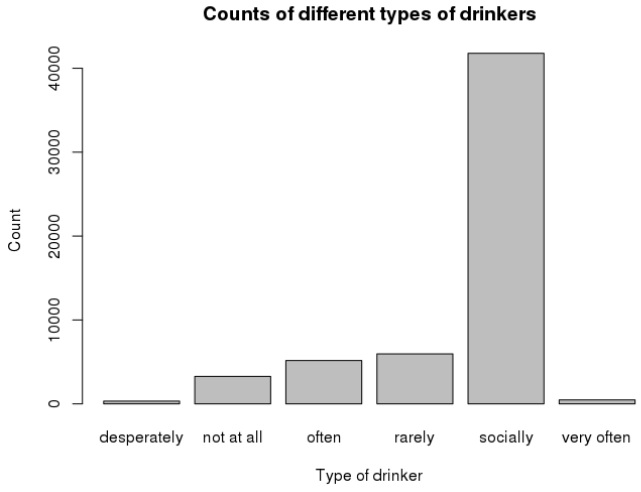
Histograms

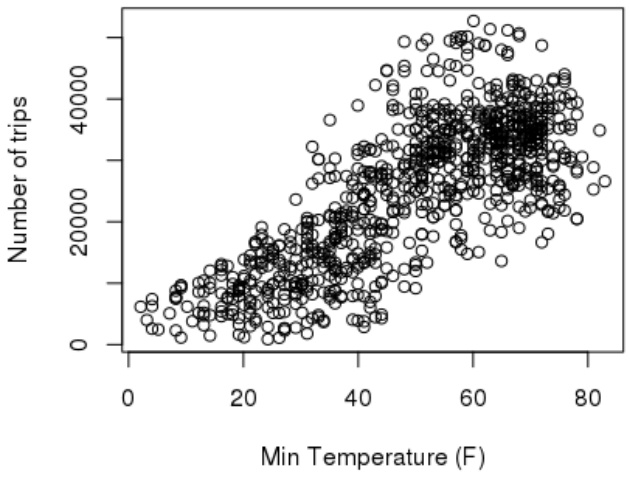


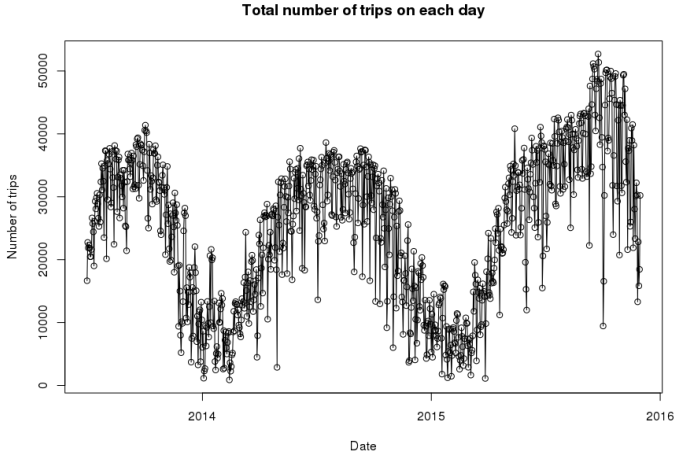
Mean height in the profiles data

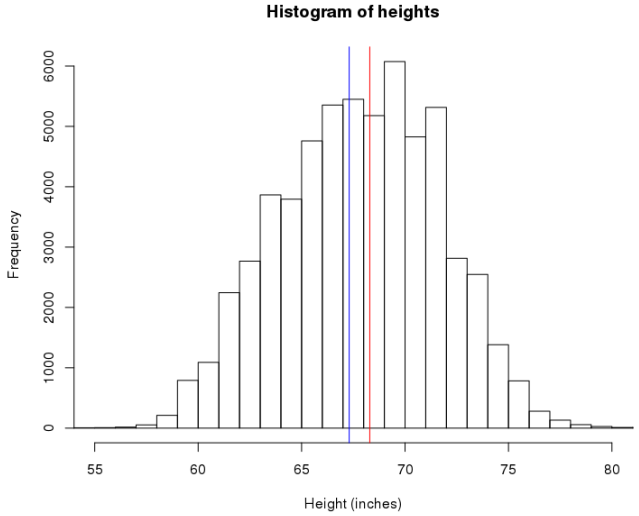
Mean height expected based on US population

What are some similarities between these graphs?







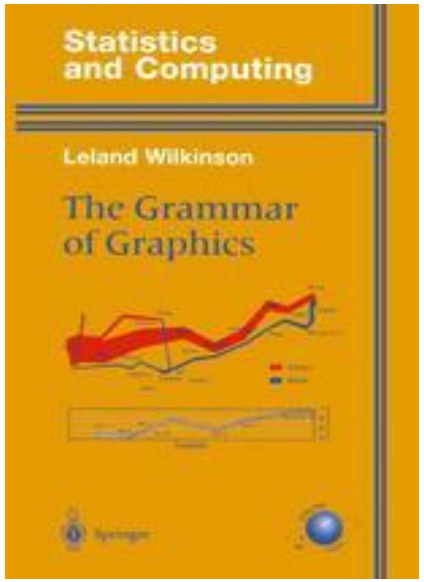


The grammar of graphics

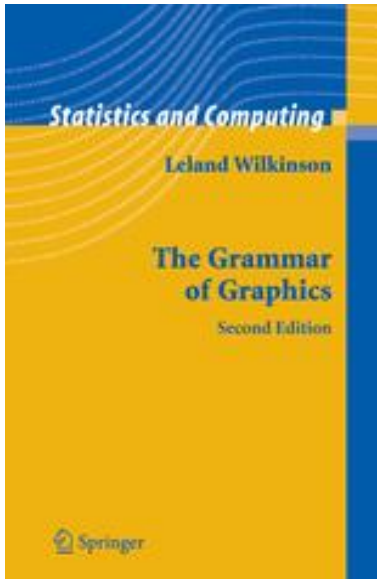
Leland Wilkinson noticed similarities between many graphs and tried to generate a ‘grammar’ that could be used to express a graph

- i.e., a list elements that can be combined together to create a graph

First edition



Second edition



Graphs are composed of…

A Frame: Coordinate system on which data is placed

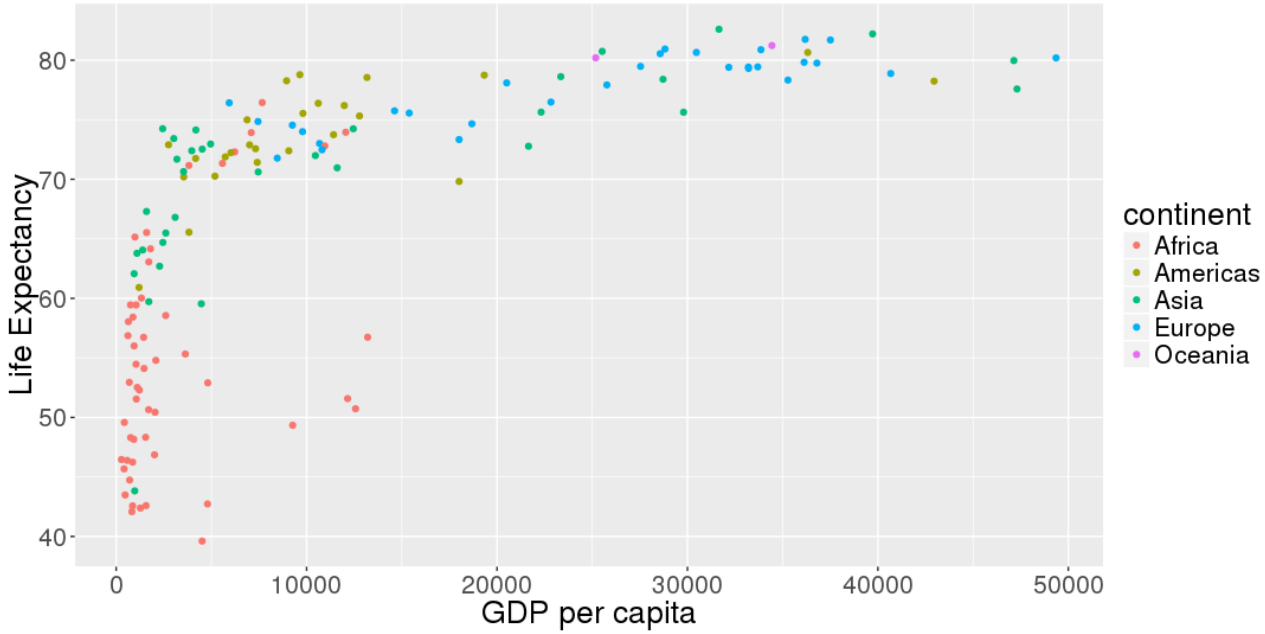
- E.g., Cartesian coordinate system, polar coordinates, etc.

Glyphs: basic graphic unit representing cases or statistics

- Contains visual properties (aesthetics) such as: shape, color, size, etc.
- Need to specify how properties of the data are **mapped** onto these aesthetics

Scales and guides: shows how to interpret axes and other properties of the glyphs

- i.e., gives information about how the data values were mapped into glyph properties



Plots can also contain…

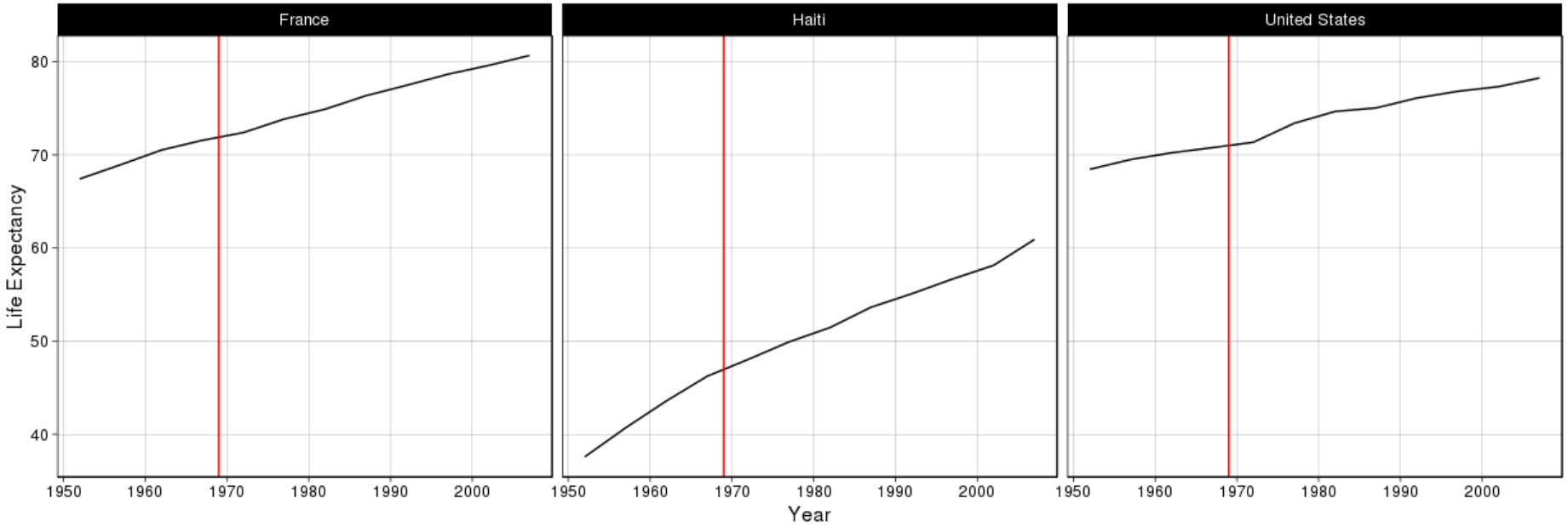
Facets: allows for multiple side-by-side graphs based on a categorical variable

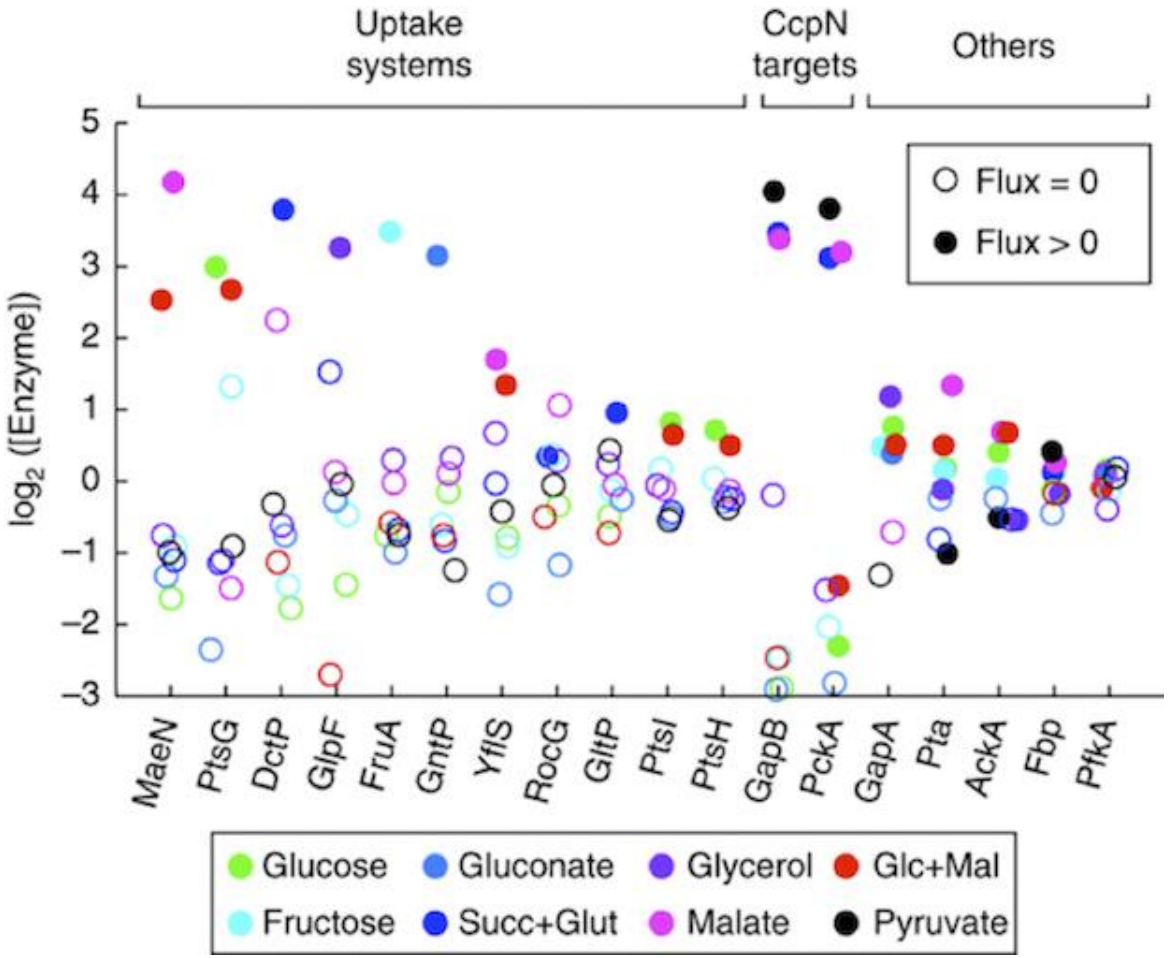
- Makes it easier to compare different conditions

Layers: allows for more than one types of data to be mapped onto the same figure

Theme: contains finer points of display

- E.g., font size, background color, etc.





The variables are:

- Log enzyme concentration
  - -3 to 5
- Target
  - CcpN, Uptake,…
- Flux
  - Zero or positive
- Gene
  - MaeN, PtsG, …
- Molecule:
  - Glocose, Fructose, …

Question: What all the mapping between variables and visual attributes?

- i.e., see if you can list the mappings from all variables to visual attributes.

Also, can sketch out the data frame that underlies this figure on a piece of paper?

Competitive States	NYT Aug 31	538 Aug 4	Cook Aug 22	Roth. Aug 29	Sabato Aug 27	WaPo Aug 29
New Hampshire	84% Dem.	90% Dem.	Leaning	Likely	Likely	>99% Dem.
Michigan	74% Dem.	65% Dem.	Tossup	Leaning	Likely	99% Dem.
Colorado	57% Dem.	60% Dem.	Tossup	Tossup	Leaning	65% Dem.
Iowa	53% Dem.	55% Dem.	Tossup	Tossup	Tossup	63% Rep.
Alaska	52% Dem.	Even	Tossup	Tossup	Tossup	66% Dem.
North Carolina	51% Rep.	Even	Tossup	Tossup	Tossup	91% Dem.
Louisiana	60% Rep.	55% Rep.	Tossup	Tossup	Tossup	51% Dem.
Arkansas	66% Rep.	60% Rep.	Tossup	Tossup	Tossup	65% Rep.
Georgia	82% Rep.	75% Rep.	Tossup	Likely	Leaning	83% Rep.
Kentucky	86% Rep.	80% Rep.	Tossup	Leaning	Likely	94% Rep.

* Rothenberg ratings are converted from a nine-category scale to a seven-category scale to make comparisons easier.

Solid Dem. | Likely Dem. | Leaning Dem. | Tossup | Leaning Rep. | Likely Rep. | Solid Rep.

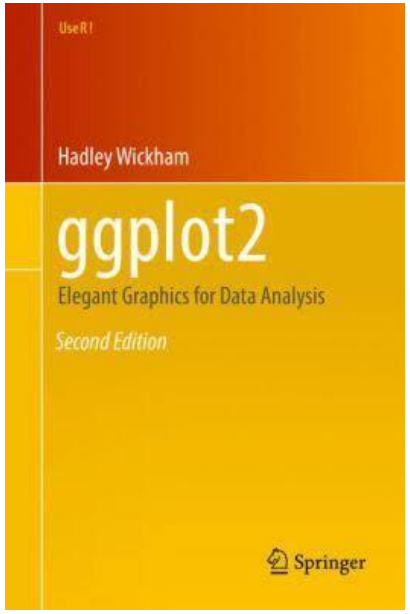
1. What variables define the frame?
2. What is the glyph and its graphical attributes
3. What sets the order for the vertical variable?

ggplot

ggplot2 is an R package that implements the grammar of graphics

- It builds up graphics by starting with a frame, adding glyphs, etc.

```
# load the ggplot2 library
> library('ggplot2')
```



Get the book on GitHub

Example data: mtcars

ROAD TEST By Jim Brokaw

THE LUXURY CARS

Imperial Palace
Fortress Fleetwood
Castle Continental

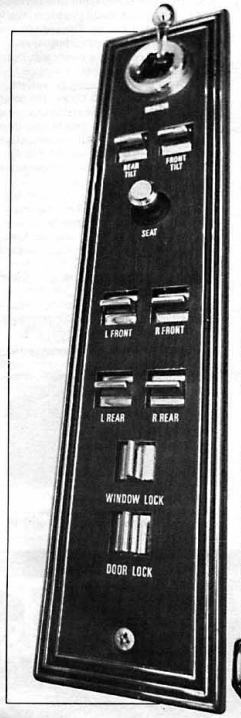
Crippled by the fuel flap and sniggered at lewdly by those smug Mercedes owners, today it seems that these great mastadons are dismissed as symbols of an ancient aristocracy whose strata was marked by expanse of wheelbase, and the heavenly quantity of gross cubic mass able to be shouldered by four beleagured tires. As the sole surviving heirs of princely Packards, dynosaurean Duesenbergs and the Brobdingnagian Bugattis, these marvels of grand proportion should be headed for the Smithsonian Institute by way of the mucky La Brea tar pits.

But are they?

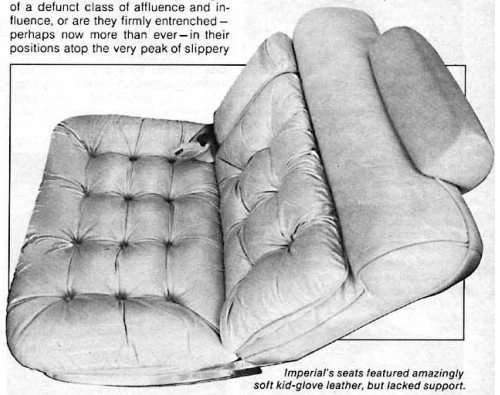
Are these behemoths simply vestiges of a defunct class of affluence and influence, or are they firmly entrenched – perhaps now more than ever – in their positions atop the very peak of slippery Mount Status, whose crags we scale daily, whether we wish to acknowledge that fact or not?

Ah, but welcome to the current state of American affairs: beef prices manipulated by withholding the animals from market; endless rounds of strikes by unions whose members are engaged in serving the public; gasoline supplies that rise and fall in mysterious coincidence with rising prices; dry rot, manipulation and partisan hatcheteering in government. They leave us little to believe in, and less to trust.

The very qualities that appear to condemn these sail-less luxury liners will very likely ensure their perpetuation. In days past, the block-long Cadillac and shiny Lincoln with paint jobs three feet deep flaunted a socio-economic position we could never hope to achieve.

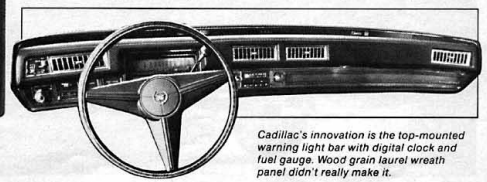


Lincoln's placement of seat controls on arm rest panel is less desireable than lower side-of seat location of Cad and Imperial.



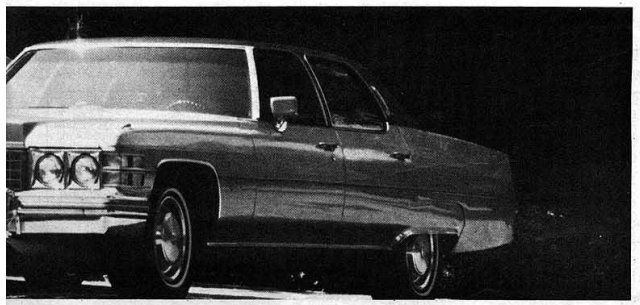
Imperial's seats featured amazingly soft kid-glove leather, but lacked support.

They did, however, constantly remind us that such positions, such wealth, such power, did, in fact, exist. The gliding specter of the shiny Imperial eagle stirred within a few heretical souls the bold idea that if such positions of power and wealth existed, there must be some means of attainment. More than a few of the haughty, distant drivers of the velvet tanks clawed their way up from the very pavement they now whisper over to



Cadillac's innovation is the top-mounted warning light bar with digital clock and fuel gauge. Wood grain laurel wreath panel didn't really make it.

PERFORMANCE	CADILLAC	LINCOLN	IMPERIAL
Acceleration			
0-30 mph	4.30	3.97	4.2
0-50 mph	8.49	8.00	9.15
0-60 mph	12.00	9.50	12.1
Standing Start 1/4-mile			
Mph	77.05	77.65	80.28
Elapsed time	17.98	17.82	17.42
Passing speeds			
40-60 mph	6.58	5.9	7.1
50-70 mph	7.00	6.8	6.8
Stopping distance			
From 30 mph	32'1"	31'4"	27'5"
From 60 mph	182'7"	153'10"	129'3"
Gas mileage range	10.43	10.42	14.7
Width – in.	79.8	80.0	79.7
Front Track – in.	63.5	64.3	64
Rear Track – in.	63.3	64.3	63.7
Wheelbase – in	133.0	127.0	124.0
Overall length – in.	233.7	232.6	231.1
Height – in.	55.6	55.4	54.7
Curb Weight – lbs.	5,250	5,425	5,345
Fuel Capacity – gals.	27	22.5	25
Oil Capacity – qts.	4 (1)	4 (1)	4 (1)
Storage Capacity – cu. ft.	19.27	20.9	20+
Base Price	$9,312	$7,637	$7,062
Price as tested	$11,435	$9,452	$8,737
Engine:	OHV V-8	OHV V-8	OHV V-8
Bore & Stroke – ins.	4.3x4.06	4.36x3.85	4.32x3.75
Displacement – cu. in.	472	460	440
HP @ RPM	205 @ 3600	215 @ 4000	230 @ 4000
Torque: lbs.-ft. @ rpm	365 @ 2000	350 @ 2600	350 @ 3200
Compression Ratio	8.25:1	NA	8.2:1
Carburetion	4V	4V	4V
Transmission	Auto. Turbo Hydra-Matic	Auto. Select Shift	Auto. Torqueflite
Final Drive Ratio	2.93	3.00	3.23 (?)
Steering Type	Recirculating Ball & Nut Power	Recirculating Ball & Nut With Integral Power Unit	Recirculating Ball Power
Steering Ratio	17.8-9.0	21.6 To 1	18.9:1
Turning Diameter (curb-to-curb-ft.)	(Wall To Wall) 24.54'	46.7'	44.69'
Wheel Turns (lock-to-lock)	2.83	3.99	3.5
Tire Size	LR78X15 Steel Belted Radials	LR78X15 Steel Belted Radials	LR78X15 Steel Belted Radial Ply
Brakes	Power Disc/Drum	Power Disc/Drum	Power Disc/Disc
Front Suspension	Coils/Shocks Front Diagonal Tie Struts Stabilizer	Coils/Shocks Axial Strut Stabilizer	Torsion Bar Shocks Stabilizer
Rear Suspension	4 Link, Coils/ Shocks	Three Link, Rubber Cushioned Pivots Coils/Shocks	Leaf Springs Shocks
Body/Frame Construction	Perimeter Frame	Body On Perimeter Frame	Unitized Construction



mtcars data frame

How can you determine what variables are in a data frame?

```
> View(mtcars)      # only works in Rstudio, not in Markdown
> glimpse(mtcars)
> ? mtcars      # this data frame as a code book
```

[, 1]	mpg	Miles/(US) gallon
[, 2]	cyl	Number of cylinders
[, 4]	hp	Gross horsepower
[, 6]	wt	Weight (1000 lbs)
[, 9]	am	Transmission (0 = automatic, 1 = manual)

Do cars that weigh more use more fuel?

Question: do cars that weigh more use more fuel?

What variables in the mtcars data frame are of interest?

- mpg
- wt

We can create a scatter plot using base graphics…

```
> plot(mtcars$wt, mtcars$mpg)
```

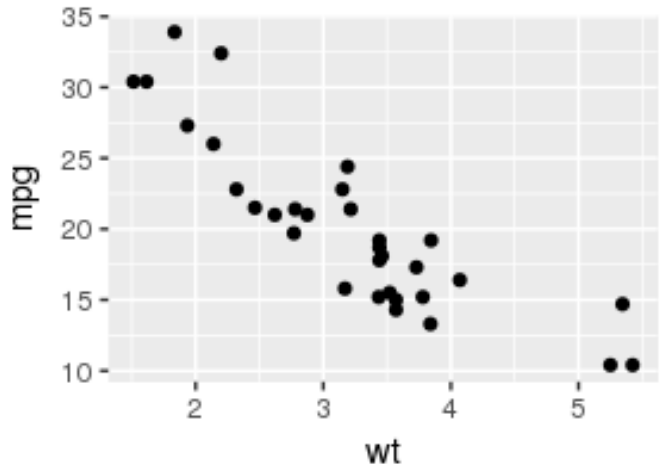
Creating a scatter plot in ggplot

Data frame to be used

Aesthetic mapping

```
> ggplot(data = mtcars, mapping = aes(x = wt, y = mpg)) +
    geom_point()
```

Adds a layer with glyphs



	wt	cyl	hp	mpg	disp
Mazda RX4	2.620	6	110	21.0	160.0
Mazda RX4 Wag	2.875	6	110	21.0	160.0
Datsun 710	2.320	4	93	22.8	108.0
Hornet 4 Drive	3.215	6	110	21.4	258.0
Hornet Sportabout	3.440	8	175	18.7	360.0

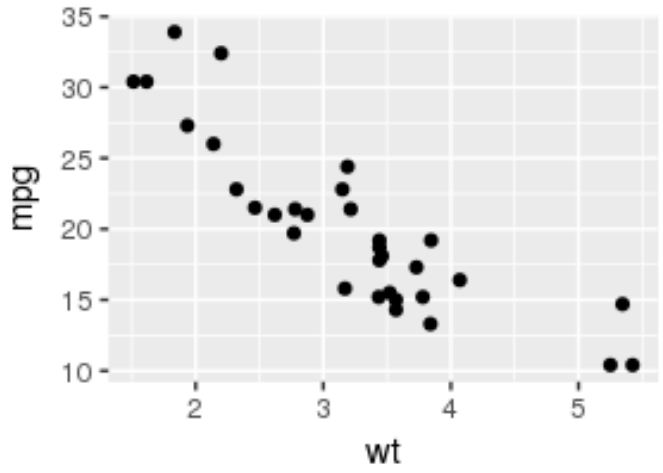
Creating a scatter plot in ggplot

Data frame to be used

Aesthetic mapping

```
> ggplot(mtcars, aes(x = wt, y = mpg)) +
    geom_point()
```

Adds a layer with glyphs



	wt	cyl	hp	mpg	disp
Mazda RX4	2.620	6	110	21.0	160.0
Mazda RX4 Wag	2.875	6	110	21.0	160.0
Datsun 710	2.320	4	93	22.8	108.0
Hornet 4 Drive	3.215	6	110	21.4	258.0
Hornet Sportabout	3.440	8	175	18.7	360.0

A lot more that ggplot can do!

- More aesthetic mapping
- Multiple glyphs/layers
- Axis labels
- Facets
- Visual themes
- Different coordinate systems
- Etc.

The R Graph Gallery

Next class: trying out ggplot!