Exploratory Data Analysis and Visualization

1. Introduction

LI Xinke

Department of Data Science

City University of Hong Kong

http://xinke.li

LI Xinke

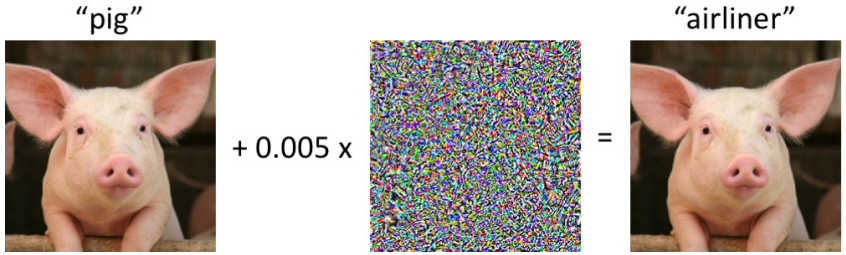
- Global Research Assistant Professor
  - Department of Data Science, City University of Hong Kong

香港城市大學
City University of Hong Kong

- Ph.D., Industry System Engineering, National University of Singapore

NUS National University of Singapore

- Research: 3D Computer Vision, Trustworthy Geometry Learning, Medical AI applications.



This course

Major objectives:

- Introduce exploratory data analysis (EDA).
- Introduce popular visualization techniques.

Tools

- The main coding environment is Python (consistent with other courses – popular).
- Tableau will also be introduced.
- There are several lab sessions right after the lecture (for Python and Tableau only).

This course

Tentative Schedule

- Mid-term exam in Week 10. No final.
- There is a group project
  - Find your friends before Week 9. (6-8 members)
  - Project presentation in Week 11-13.

This course

- Grading:
  - Group project: 40%
  - Individual coursework: 25%
    - Quizzes: on time submission - 2pt,
      late submission - 1pt
    - HWs: based on performance,
      full grade - 10pt
  - Midterm test: 35%

This course

- Canvas: 202509SDSC5002C62
  (not 202509SDSC5002 !!!)

  Same content as 202509SDSC5002C61 Prof. Lijia Wang

- You can find:
  - Announcement
  - Slides
  - Tutorials
  - Recordings
  - Quiz
  - Assignment
  - Project details

  …

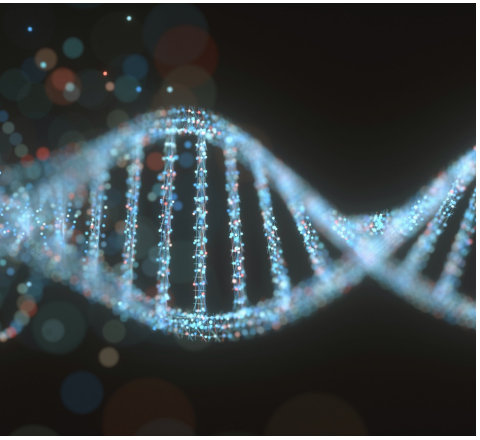
  If you need more information to be posted on Canvas, contact me!

This course

Tentative Schedule:

- Weeks 1-3: key concepts of exploratory data analysis and data visualization
- Week 4: data analytics and data visualization for machine learning
- Week 5: National day
- Weeks 6-7: data analytics and data visualization for machine learning
- Weeks 8: High-dimensional data visualization
- Weeks 9: Chung Yeung Festival (Project Proposal Submission)
- Week 10: mid-term
- Weeks 11: Advanced topics in EDA, start of group project presentation (tentative)
- Week 12: group project presentation
- Week 13: group project presentation

- **Instructor:** LI, Xinke
  - xinkeli@cityu.edu.hk
- **TAs:**
  - Mr. PAN Jiming(Organize Group Project)
  - jmpan3-c@my.cityu.edu.hk
  - Mr. ZHANG Wenlin (Python)
  - wl.z@cityu.edu.hk
  - Ms. LI Yanru(Tableau)
  - yanru.li@my.cityu.edu.hk

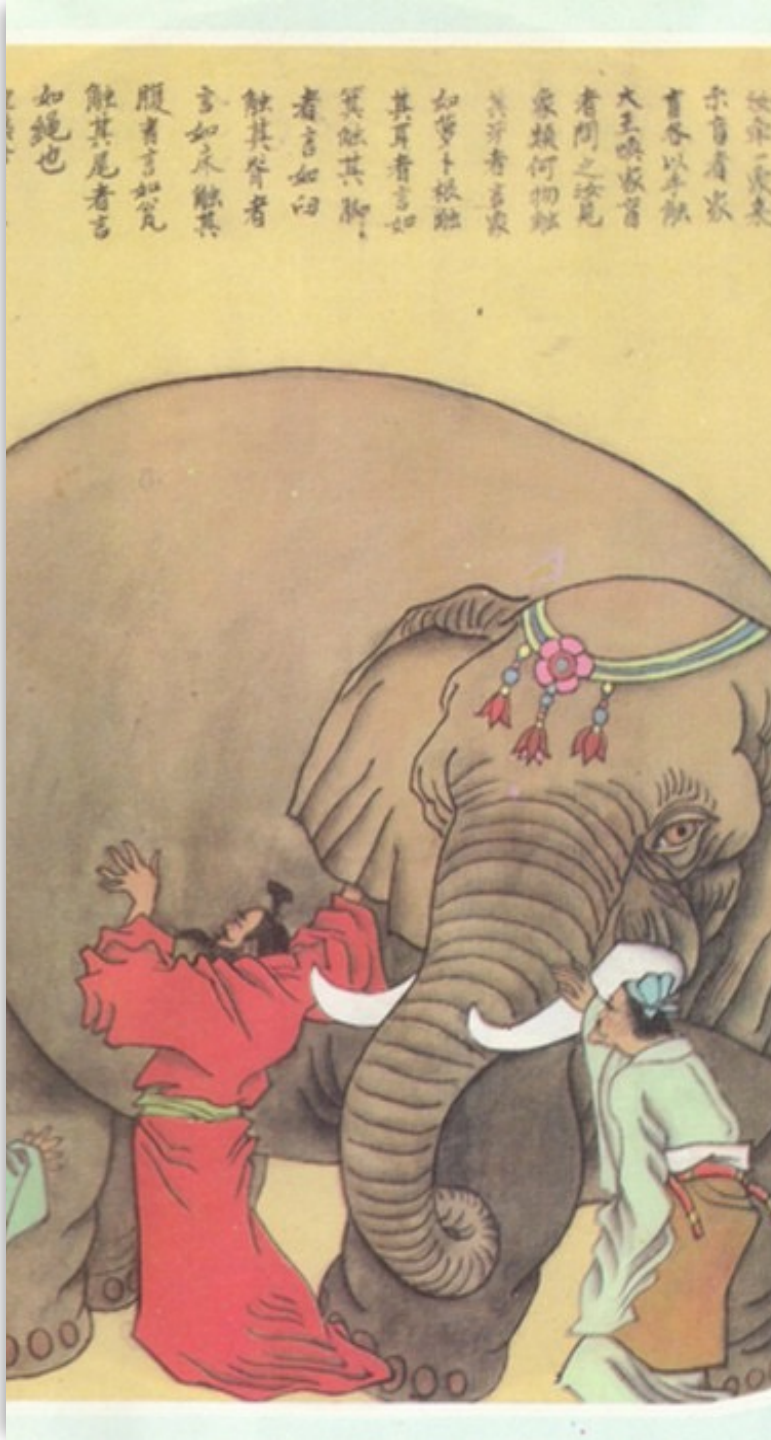


Data Analytics

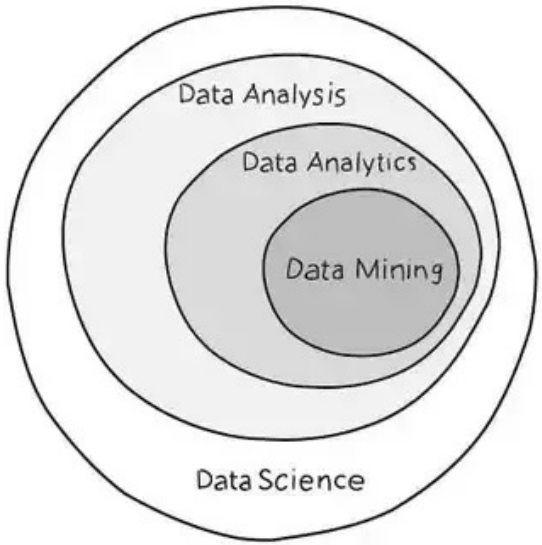
- Systems (social, economic, ecological, biological, political, educational, transportation and other types of systems) generate huge amount of data.
- In today's world, we have the necessary hardware and software to utilize this data effectively.
- The time has come to capitalize on the potential of data analytics.

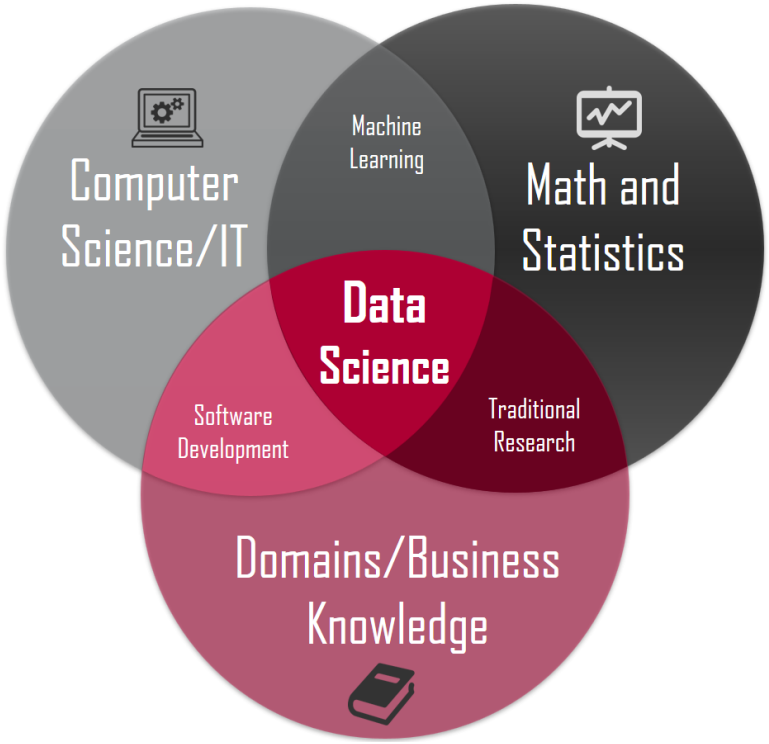
Data Analytics

- The world is a mystery for everyone.
  - e.g., animals/humans remain insufficiently studied in the realms of biology, psychology, and social sciences.
- Animals/Humans' behaviors are stochastic and SUPER complex. It is difficult to build a PHYSICAL model!
- Fortunately, we can extract data, analyze the data, and generate insights and knowledge about the system.
- "盲人摸象" (Blind men and an elephant)



Data Analytics





- https://www.mo-data.com/what-is-the-difference-between-data-analytics-data-analysis-data-mining-data-science-machine-learning-big-data-and-predictive-analytics/

Question

- What is your current primary smart phone?
  - Android
  - iPhone
  - Others
- What was your last primary smart phone?
  - Android
  - iPhone
  - Others

Very simple model

- # of people using iPhone, x_1; # of people using Android, x_2
- # of people used iPhone,y_1; # of people used Android, y_2

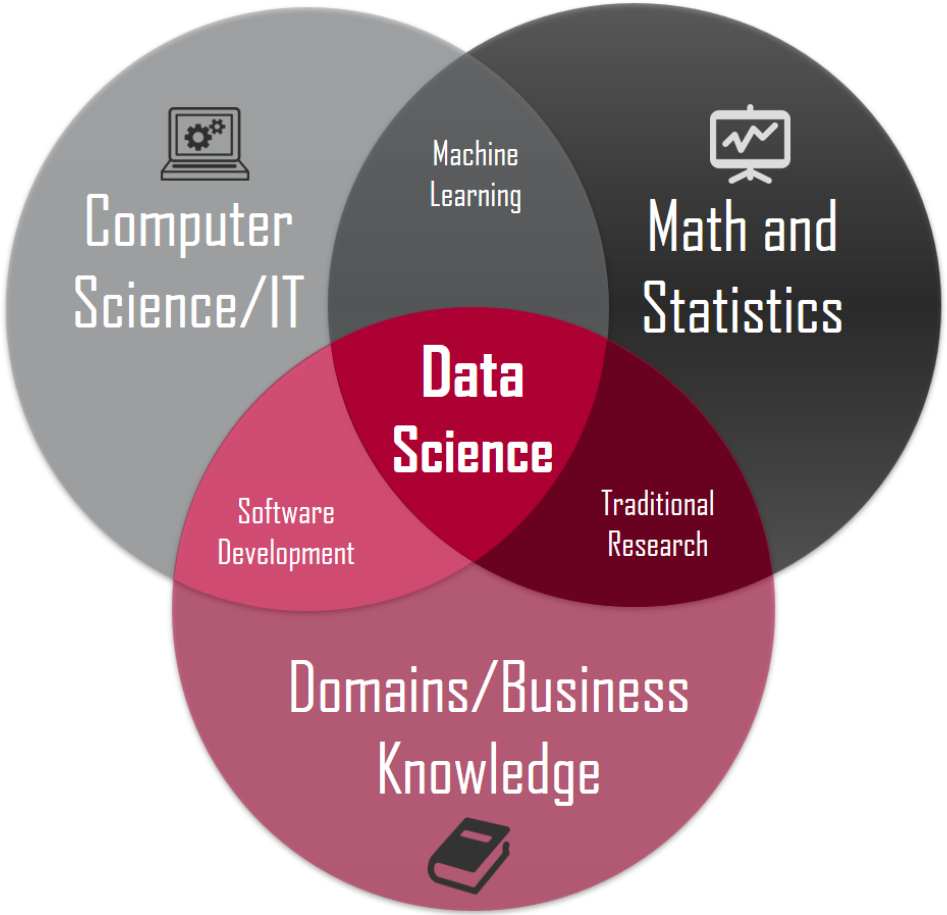
- # of people who switched to iPhone from Android, $x_{1,2}$
  - %: $p_{1,2}=x_{1,2}/ y_2$
- # of people who switched from Android to iPhone, $x_{2,1}$
  - %: $p_{2,1}= x_{2,1}/y_1$
- % of people who remained (stickiness of the product)
  - iPhone: $p_{1,1}=(x_1 – x_{1,2})/ y_1$
  - Android: $p_{2,2}=(x_2 – x_{2,1})/ y_2$

Prediction

- Next time when you upgrade your phone, how many of you will be using
  - iPhone: $x_1' = x_1 p_{1,1} + x_2 p_{2,1}$
  - Android: $x_2' = x_2 p_{2,2} + x_1 p_{1,2}$
  - $x_1' + x_2' = x_1 + x_2$

Domain Knowledge

- Domain knowledge is important!



If we have domain knowledge

- We would ask for more questions:
  - Gender, g
  - Income, m
  - Interest in digital products, d
- Then, we build a regression model:
  - $y_{t+1} = \beta_0 + \beta_1 y_t + \beta_2 g + \beta_3 m + \beta_4 d + \epsilon$

Additional domain knowledge

- Social influence
  - % of friends who are using Android, a
- $y_{t+1} = \beta_0 + \beta_1 y_t + \beta_2 g + \beta_3 m + \beta_4 d + \beta_5 a + \epsilon$

- The more the better “多多益善”?
  - How to evaluate the performance of the prediction?
  - EDA and data visualization
- If too much – we need to “engineer” features
  - Data mining

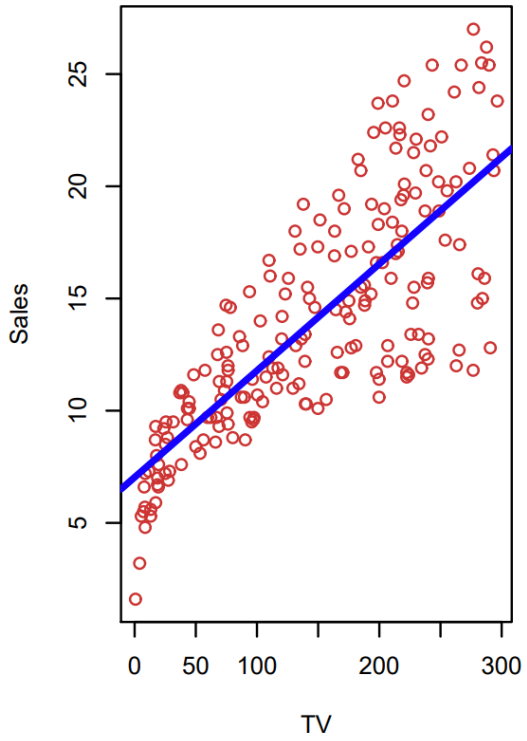
AI and Statistics

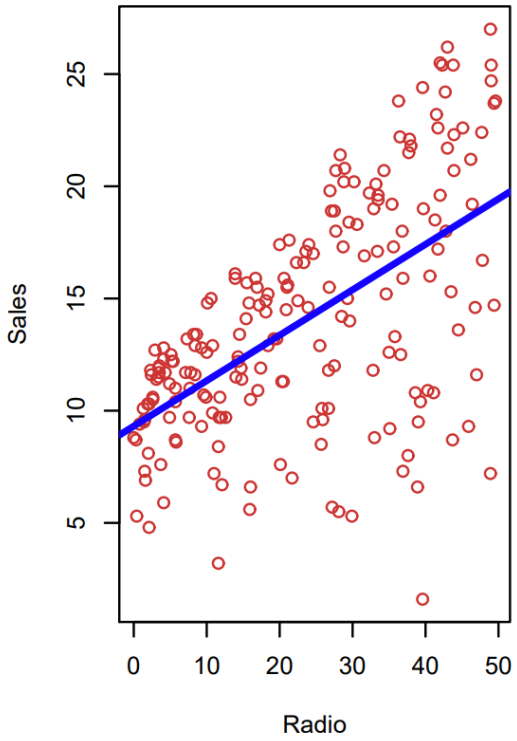
- Statistics is the key tool for machine learning. (Key philosophy)
- Machine learning is packaged as “AI” for now.

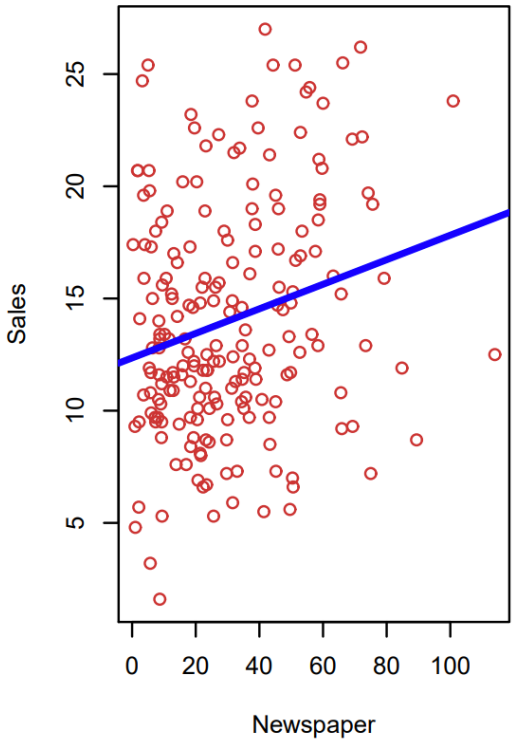
Let’s build a skyscraper step by step!

Statistical Learning

- Example (Advertising data)



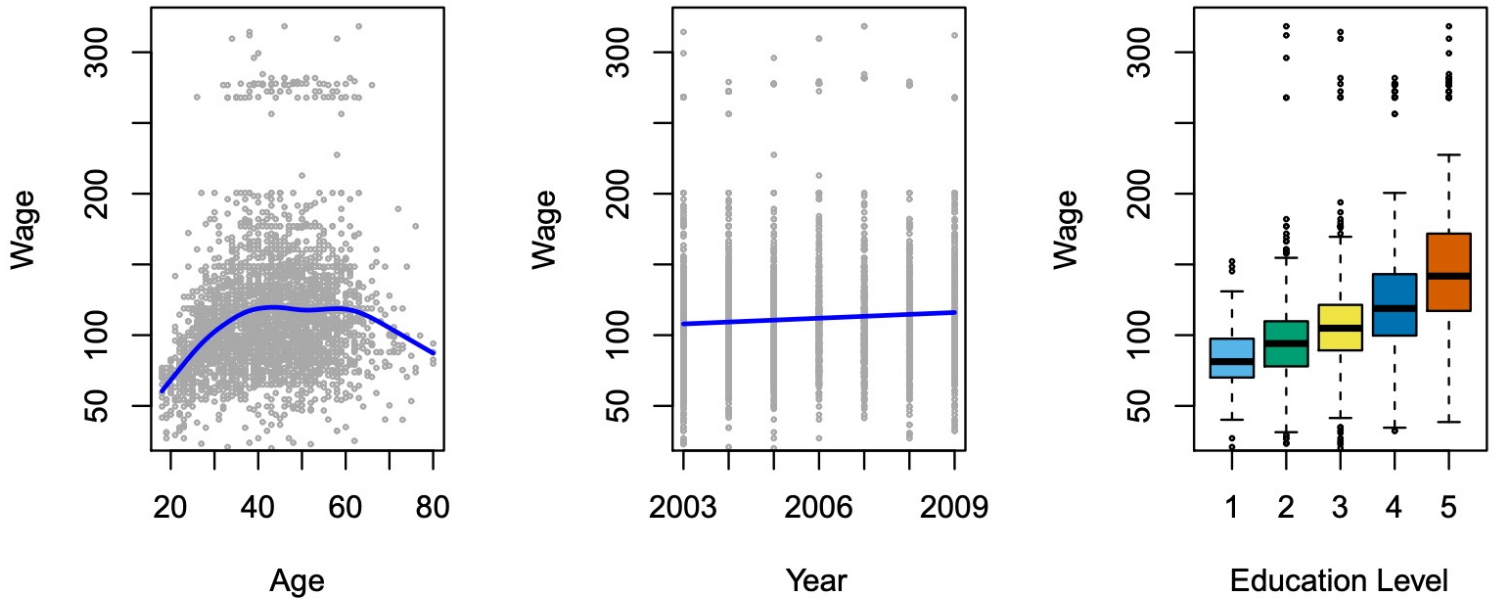




https://www.statlearning.com/

Statistical Learning

- Example

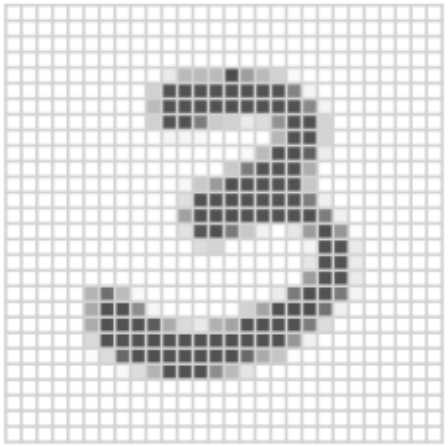


Income survey data for males from the central Atlantic region of the USA in 2009.

https://www.statlearning.com/

Statistical Learning

- Example



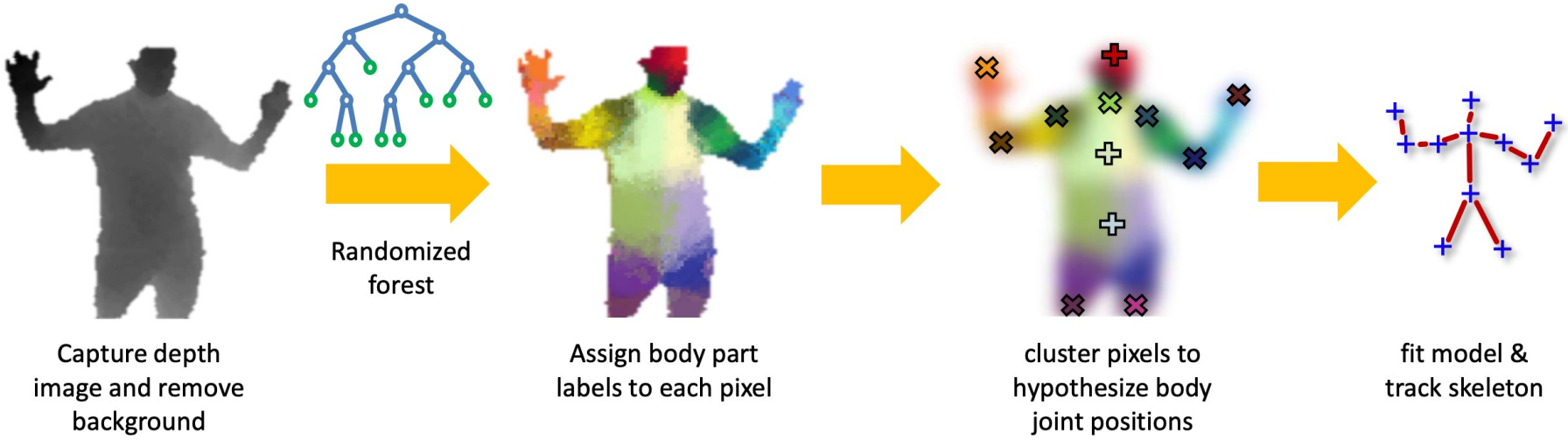
https://en.wikipedia.org/wiki/MNIST_database

Statistical Learning

- Example

Statistical Learning

- Kinect
- https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/ks_book_2012.pdf



Supervised Learning

- Outcome measurement Y
  - dependent variable, response, target
- Vector of p predictor measurements X
  - inputs, regressors, covariates, features, independent variables
- In regression problem, Y is quantitative
  - price, blood pressure
- In classification problem, Y takes discrete values
  - survived/died, digit 0-9, cancer class of tissue sample
- We have training data $(x_1, y_1), \ldots, (x_N, y_N)$. These are observations (examples, instances) of these measurements.

Supervised Learning

Objective

- On the basis of the training data we would like to:
  - Accurately predict unseen test cases.
  - Understand which inputs affect the outcome, and how.
  - Assess the quality of our predictions and inferences.

Supervised Learning

A few points

- It is important to **understand the ideas behind the various questions**, in order to know how and when to use them.
  - Domain knowledge – IE/Biz students are good at industrial/biz problems

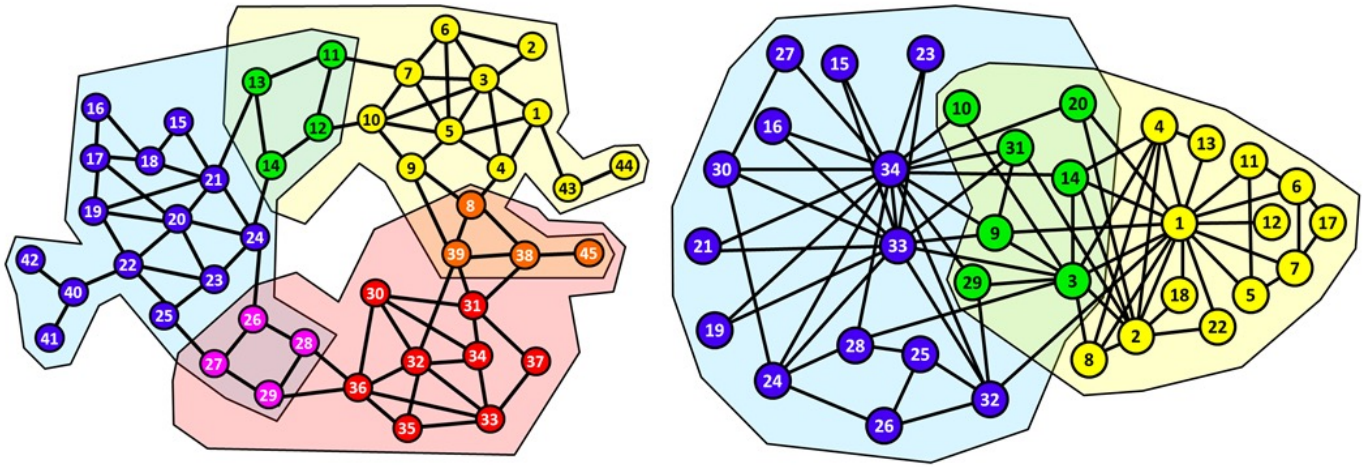
- One has to understand **the simpler methods first**, in order to master the more complicated ones.
- It is important to accurately **assess the performance** of a method, to know how well or how badly it is working
  - Simpler methods often perform as well as fancier ones!
- This is an exciting research area, having important applications in science, industry and finance.

Unsupervised Learning

- **No outcome variable**, just a set of predictors (features) measured on a set of samples.
- Objective is more general –
  - find groups of samples that behave similarly,
  - find features that behave similarly,
  - find linear combinations of features with the most variation.
- Difficult to know how well you are doing.
- Different from supervised learning but can be useful as a pre-processing step for supervised learning.

Unsupervised Learning

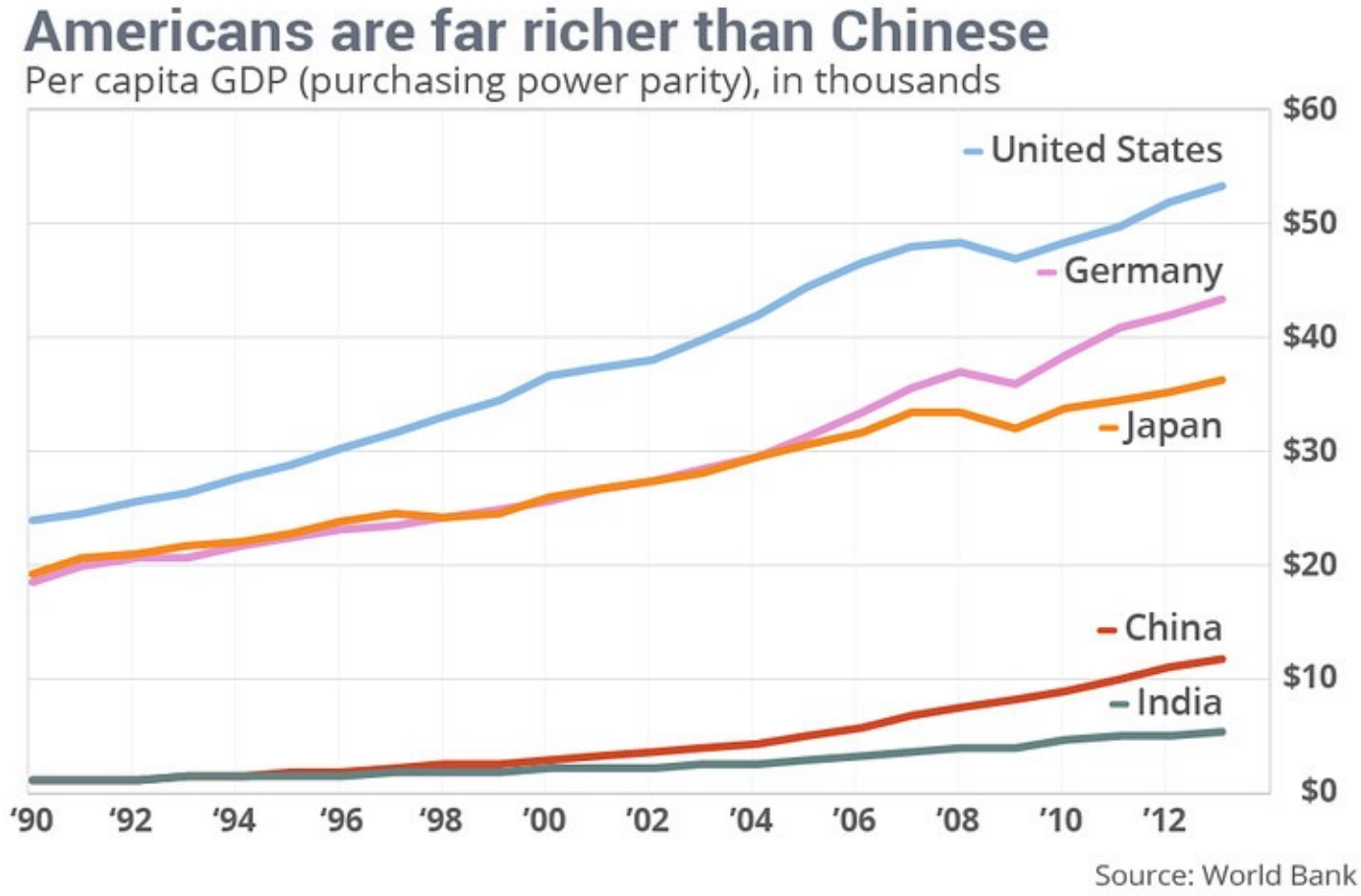
- Community detection in social networks



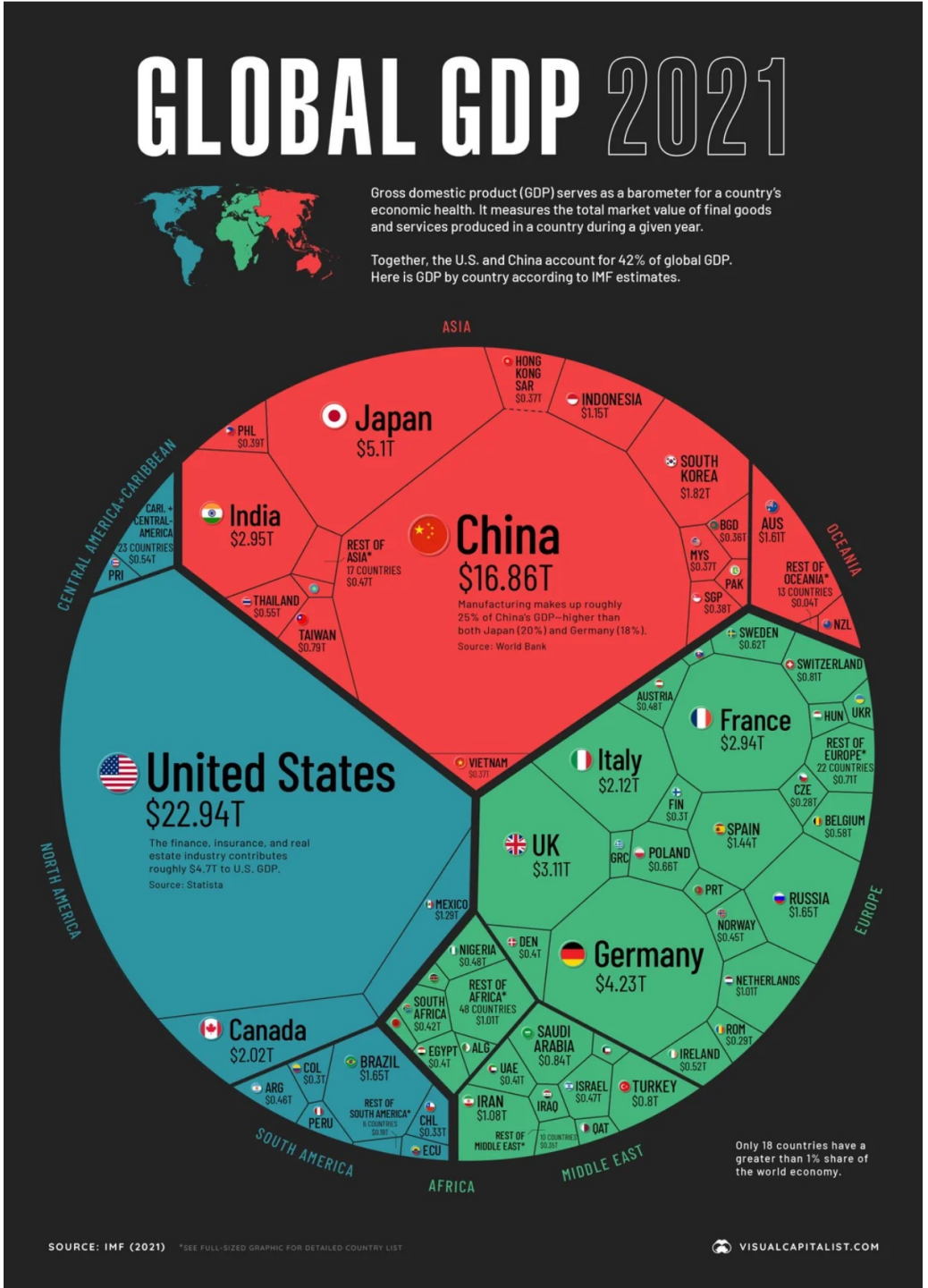
Focus of this course

- Exploratory data analysis and visualization
  - Exploratory data analysis
  - Data visualization

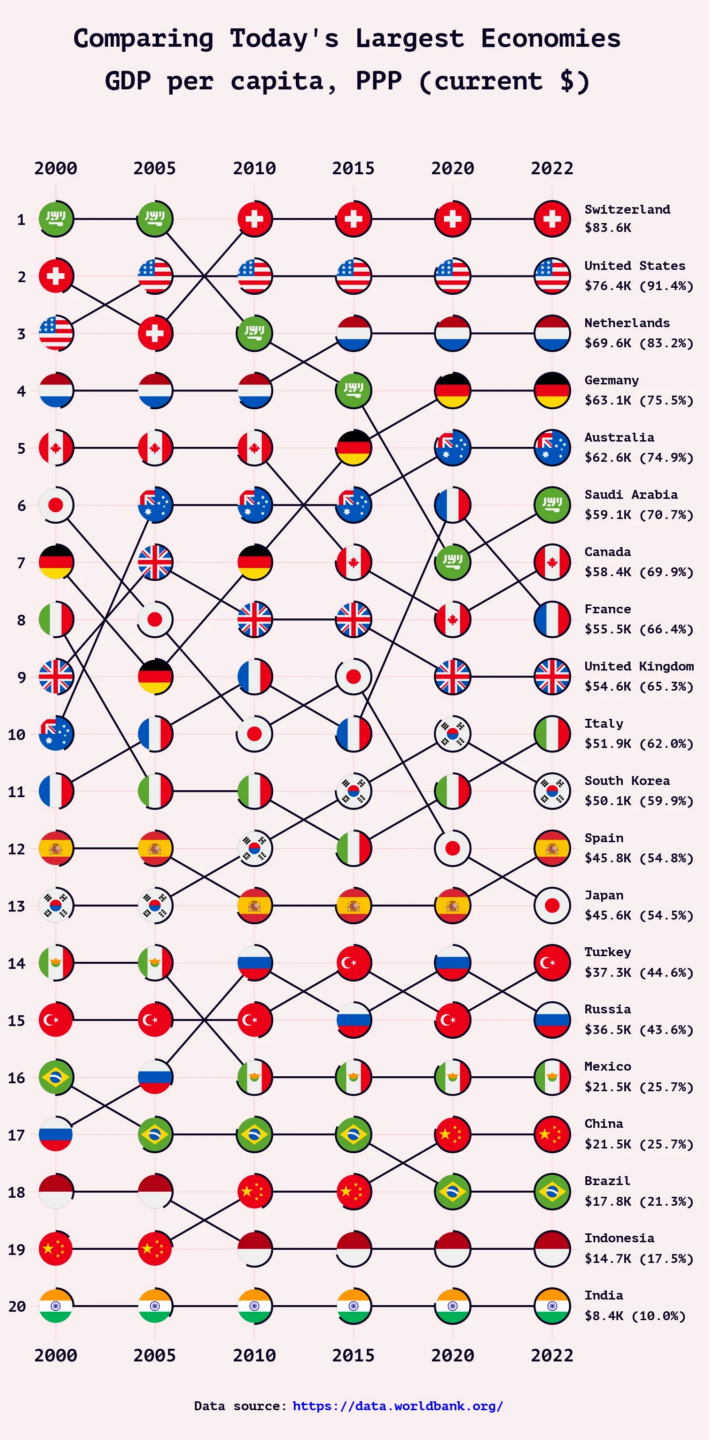
Example: GDP

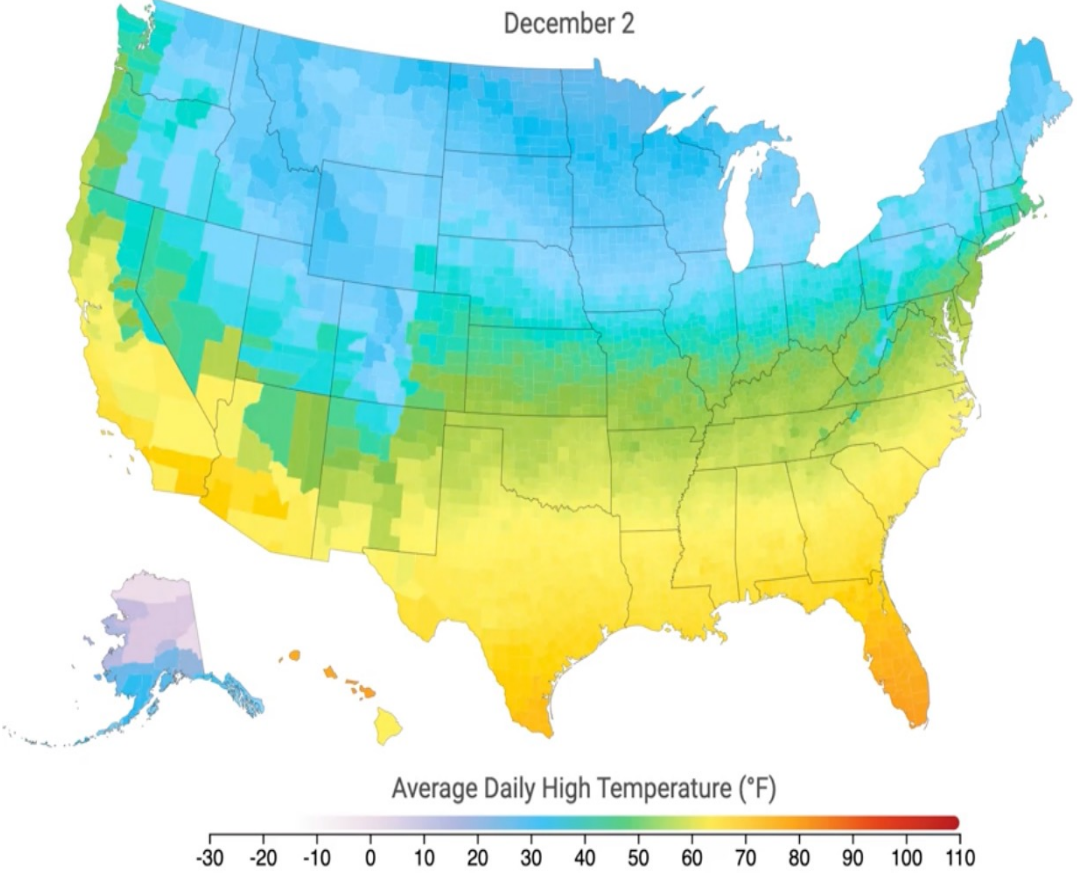

Americans are far richer than Chinese
Per capita GDP (purchasing power parity), in thousands
United States
Germany
Japan
China
India
$60
$50
$40
$30
$20
$10
$0
'90
'92
'94
'96
'98
'00
'02
'04
'06
'08
'10
'12
Source: World Bank

Example: GDP

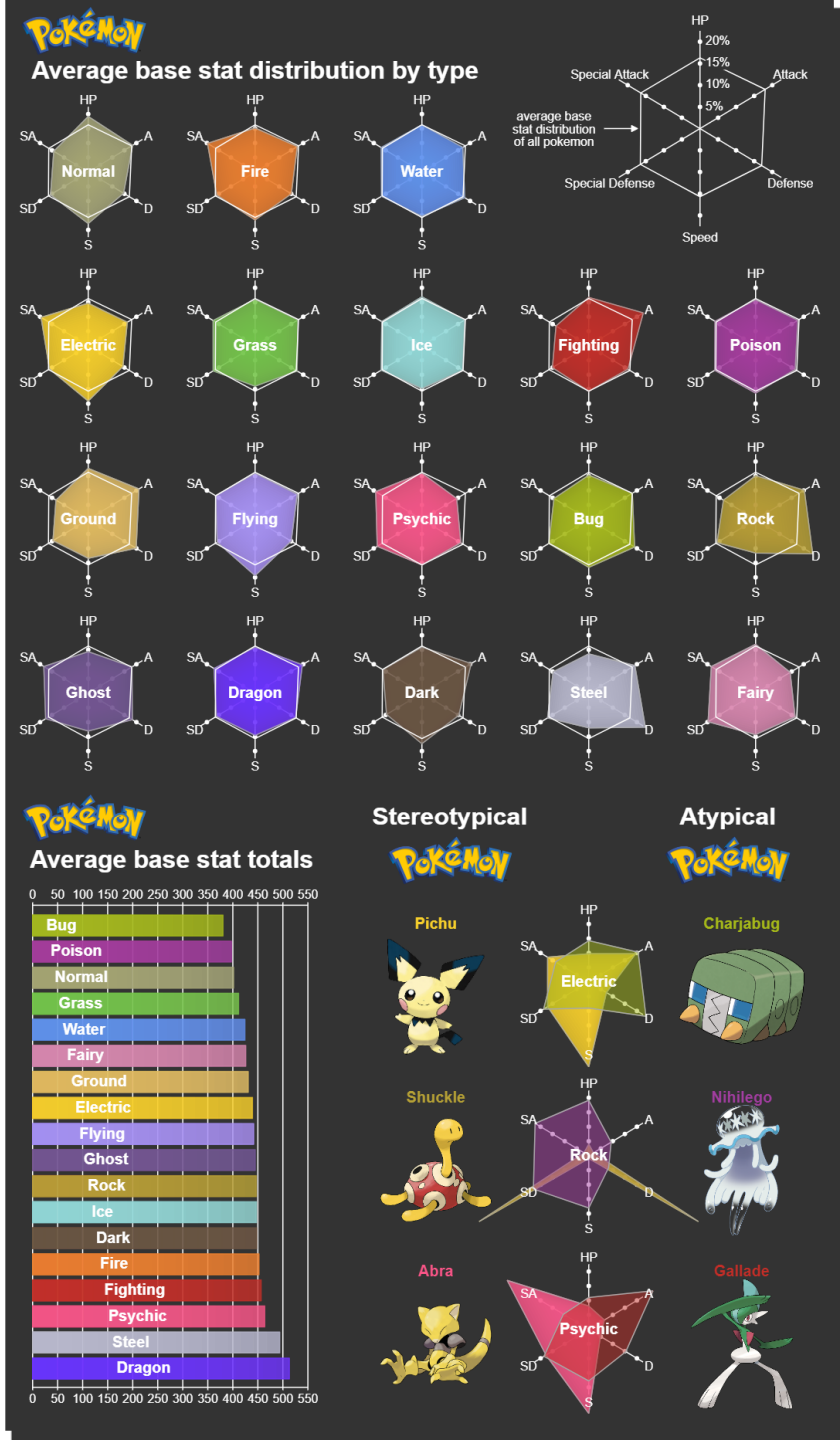

GLOBAL GDP 2021
Gross domestic product (GDP) serves as a barometer for a country's economic health. It measures the total market value of final goods and services produced in a country during a given year.
Together, the U.S. and China account for 42% of global GDP. Here is GDP by country according to IMF estimates.
ASIA
Japan $5.1T
China $16.86T
Manufacturing makes up roughly 25% of China's GDP—higher than both Japan (20%) and Germany (18%).
Source: World Bank
India $2.95T
HONG KONG SAR $0.37T
INDONESIA $1.15T
SOUTH KOREA $1.82T
PHL $0.39T
REST OF ASIA* 17 COUNTRIES $0.47T
THAILAND $0.55T
TAIWAN $0.79T
VIETNAM $0.37T
BGD $0.36T
MYS $0.37T
PAK
SGP $0.38T
CENTRAL AMERICA+CARIBBEAN
CARI. + CENTRAL AMERICA 23 COUNTRIES $0.54T
PRI
OCEANIA
AUS $1.61T
REST OF OCEANIA* 13 COUNTRIES $0.04T
NZL
EUROPE
France $2.94T
Italy $2.12T
Germany $4.23T
UK $3.11T
SWEDEN $0.62T
SWITZERLAND $0.81T
AUSTRIA $0.48T
HUN
UKR
REST OF EUROPE* 22 COUNTRIES $0.71T
CZE $0.28T
FIN $0.3T
SPAIN $1.44T
BELGIUM $0.58T
GRC
POLAND $0.66T
PRT
RUSSIA $1.65T
NORWAY $0.45T
DEN $0.4T
NETHERLANDS $1.01T
ROM $0.29T
IRELAND $0.52T
NORTH AMERICA
United States $22.94T
The finance, insurance, and real estate industry contributes roughly $4.7T to U.S. GDP.
Source: Statista
MEXICO $1.29T
Canada $2.02T
SOUTH AMERICA
BRAZIL $1.65T
COL $0.3T
ARG $0.46T
PERU
REST OF SOUTH AMERICA* 6 COUNTRIES $0.39T
CHL $0.33T
ECU
AFRICA
NIGERIA $0.48T
SOUTH AFRICA $0.42T
REST OF AFRICA* 48 COUNTRIES $1.01T
EGYPT $0.4T
ALG
MIDDLE EAST
SAUDI ARABIA $0.84T
UAE $0.41T
IRAN $1.08T
IRAQ
REST OF MIDDLE EAST* 10 COUNTRIES $0.35T
ISRAEL $0.47T
TURKEY $0.8T
QAT
Only 18 countries have a greater than 1% share of the world economy.
SOURCE: IMF (2021)
*SEE FULL-SIZED GRAPHIC FOR DETAILED COUNTRY LIST
VISUALCAPITALIST.COM

Example: GDP


Comparing Today's Largest Economies
GDP per capita, PPP (current $)
2000
2005
2010
2015
2020
2022
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
Switzerland
$83.6K
United States
$76.4K (91.4%)
Netherlands
$69.6K (83.2%)
Germany
$63.1K (75.5%)
Australia
$62.6K (74.9%)
Saudi Arabia
$59.1K (70.7%)
Canada
$58.4K (69.9%)
France
$55.5K (66.4%)
United Kingdom
$54.6K (65.3%)
Italy
$51.9K (62.0%)
South Korea
$50.1K (59.9%)
Spain
$45.8K (54.8%)
Japan
$45.6K (54.5%)
Turkey
$37.3K (44.6%)
Russia
$36.5K (43.6%)
Mexico
$21.5K (25.7%)
China
$21.5K (25.7%)
Brazil
$17.8K (21.3%)
Indonesia
$14.7K (17.5%)
India
$8.4K (10.0%)
Data source: https://data.worldbank.org/


December 2
Average Daily High Temperature (°F)
-30 -20 -10 0 10 20 30 40 50 60 70 80 90 100 110


3361.42
3314.28
3267.14
3220.00
3172.86
3125.72
3078.58
3031.44
2984.30
2937.16
2890.02
MA10:3129.74
MA30:3074.52
09-11-18
12-03
12-21
10-01-07
01-25
02-10
03-05
03-23
04-09
2502116
2001692
1501269
1000846
500423
×100
MA10:1118822
MA30:980143
9u8u.com


Average base stat distribution by type
HP
SA
A
SD
D
S
Normal
Fire
Water
Electric
Grass
Ice
Fighting
Poison
Ground
Flying
Psychic
Bug
Rock
Ghost
Dragon
Dark
Steel
Fairy
20%
15%
10%
5%
Special Attack
Attack
Special Defense
Defense
Speed
average base stat distribution of all pokemon
Average base stat totals
0 50 100 150 200 250 300 350 400 450 500 550
Bug
Poison
Normal
Grass
Water
Fairy
Ground
Electric
Flying
Ghost
Rock
Ice
Dark
Fire
Fighting
Psychic
Steel
Dragon
Stereotypical
Atypical
Pichu
Charjabug
Shuckle
Nihilego
Abra
Gallade

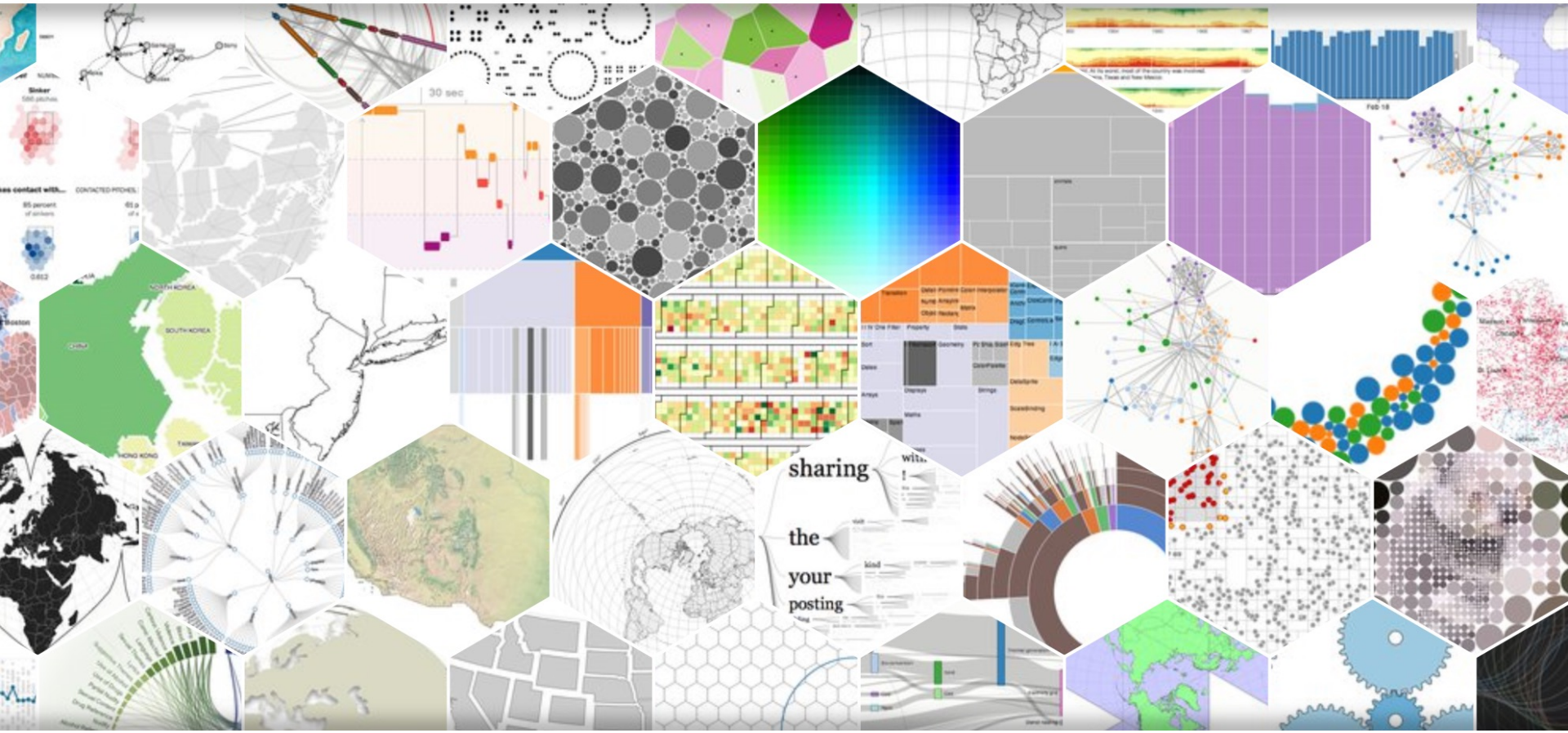
Core

- **Data Foundations** - information in numbers
- **Context and Understanding** - contributes to problem solving
- A visualization has a mission:
- Put it simple,
  - we can characterize the data from a straightforward perspective.
  - We would like to visualize the analytics results to help others understand.

Tools

- **Excel** - the business standard
- **SPSS** and **SAS** – commercial
  - Heavily used in social sciences
- **R** – statistician (data scientist) standard
- **Python** – data scientist standard
- **D3** – computer scientist standard
  - Based on JavaScript - interactive
- **Tableau** – a beautiful enterprise standard

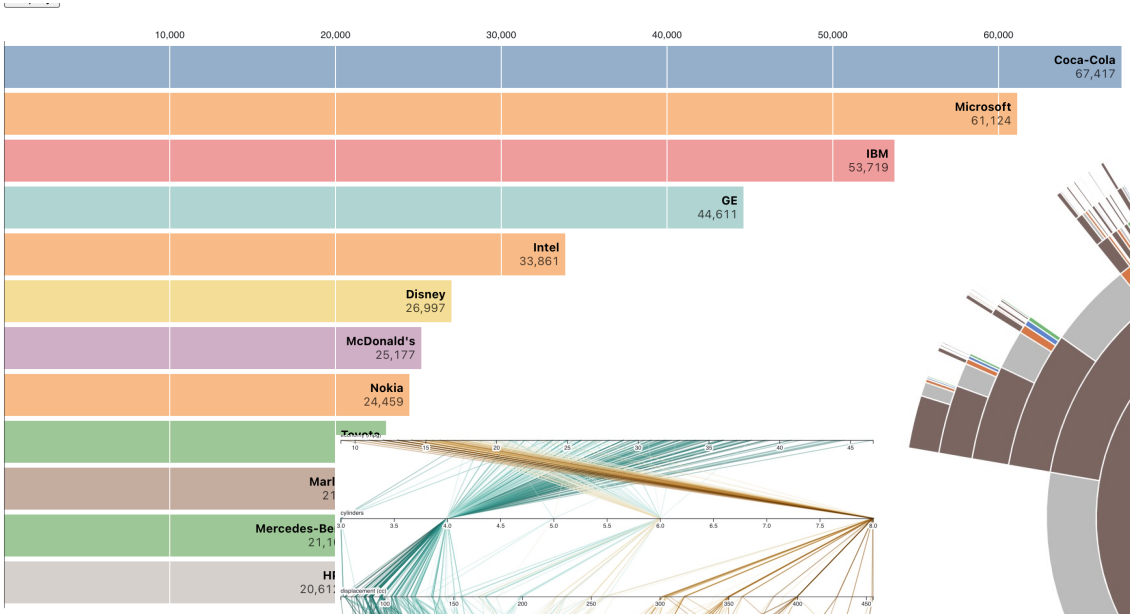
Data-Driven Documents

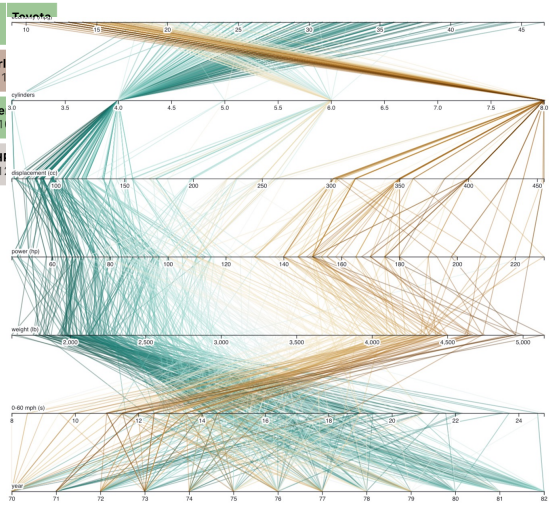


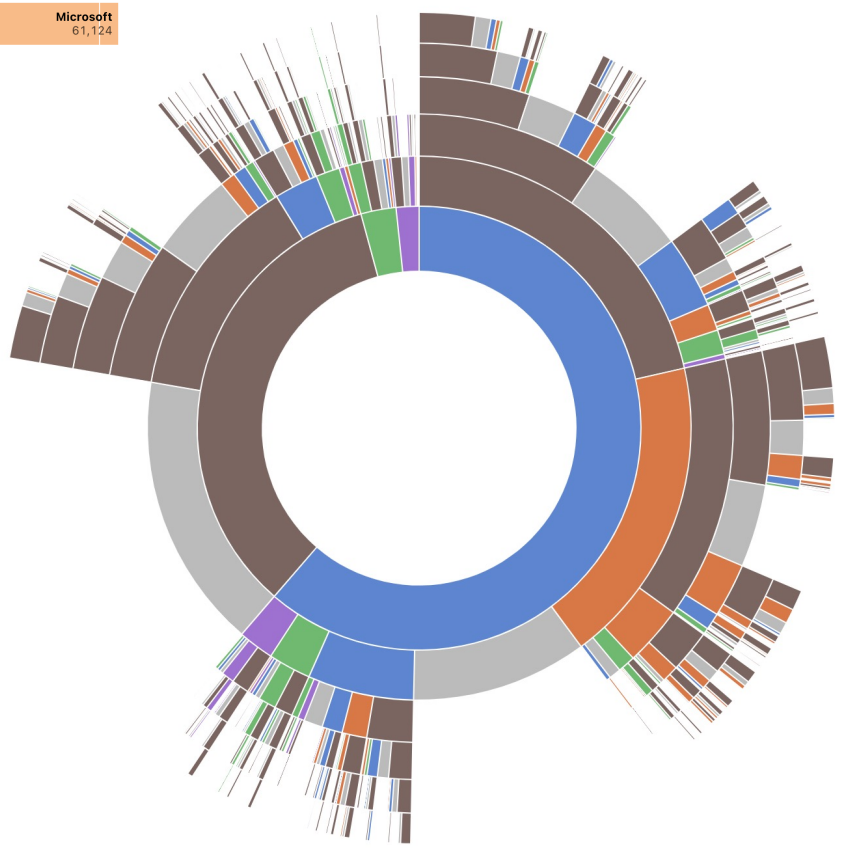
https://d3js.org/

Example:

- https://observablehq.com/@d3/gallery?utm_source=d3js-org&utm_medium=hero&utm_campaign=try-observable

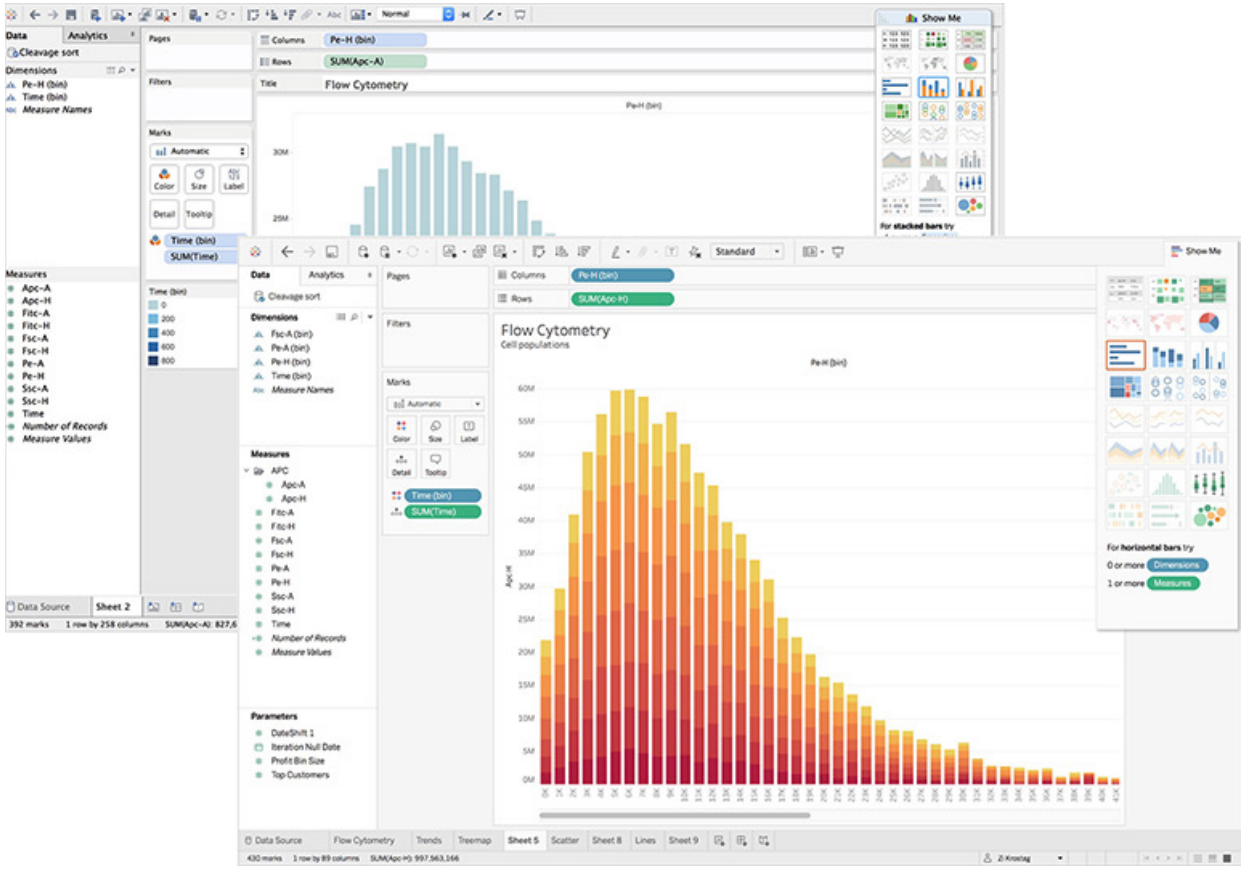




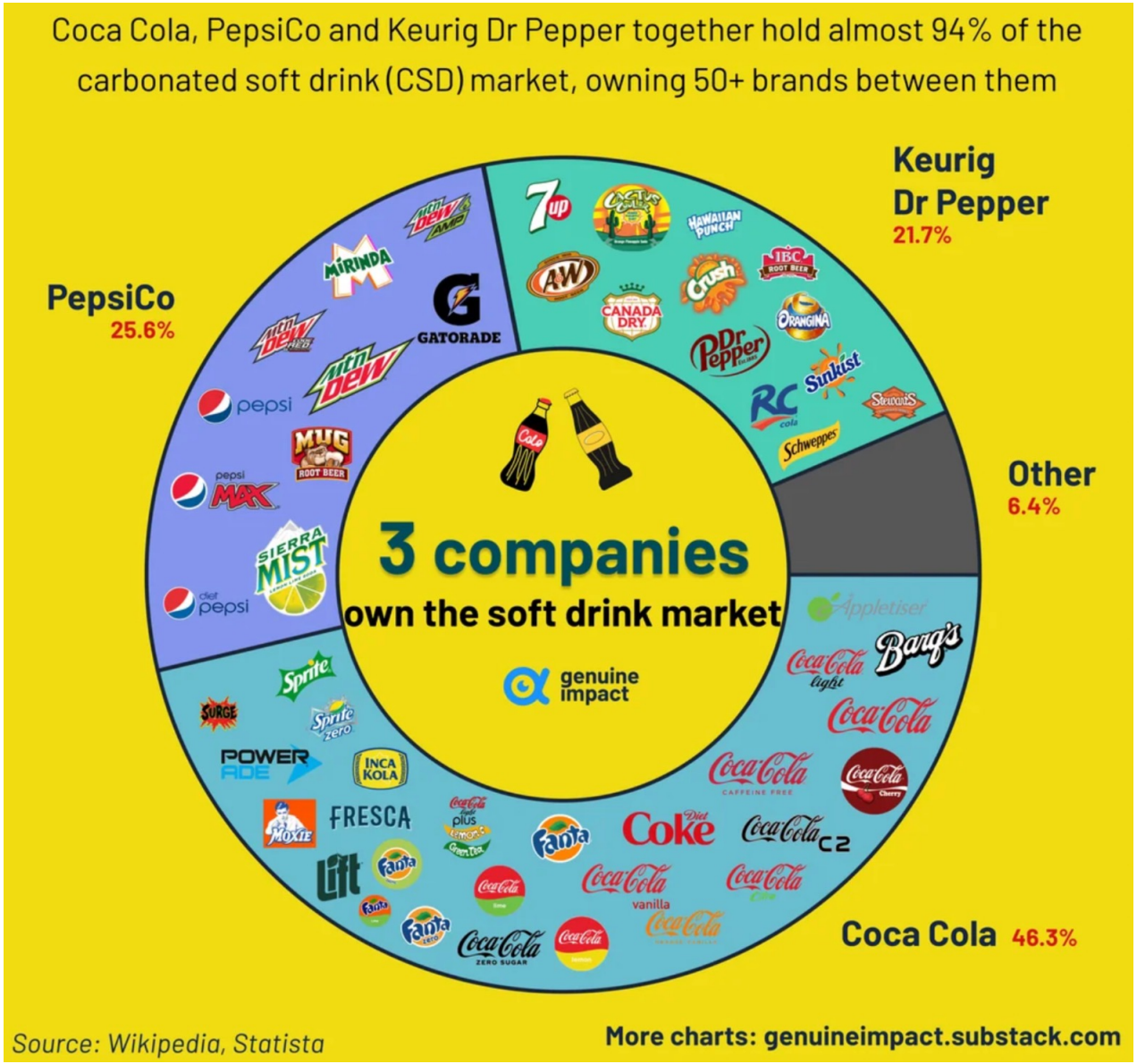


Commercial software

- Tableau (Free for student)

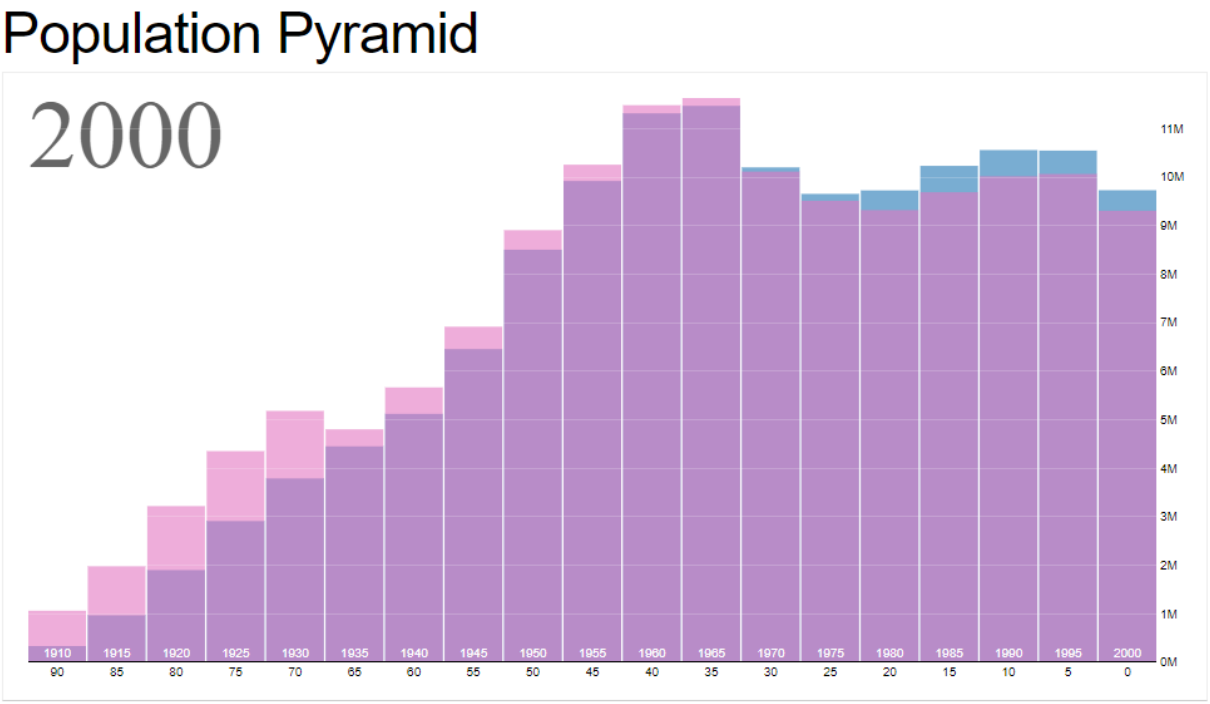


Pie Chart


Coca Cola, PepsiCo and Keurig Dr Pepper together hold almost 94% of the carbonated soft drink (CSD) market, owning 50+ brands between them
Keurig Dr Pepper 21.7%
PepsiCo 25.6%
Other 6.4%
Coca Cola 46.3%
3 companies
own the soft drink market
genuine impact
Source: Wikipedia, Statista
More charts: genuineimpact.substack.com

Bar Chart

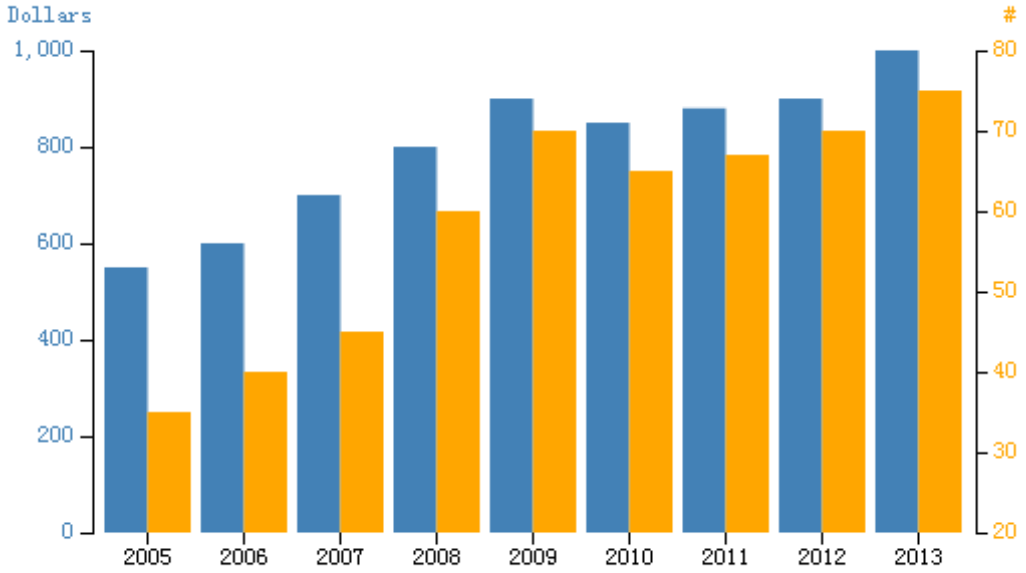
- This diagram shows the distribution of age groups in the United States.



- https://bl.ocks.org/mbostock/4062085

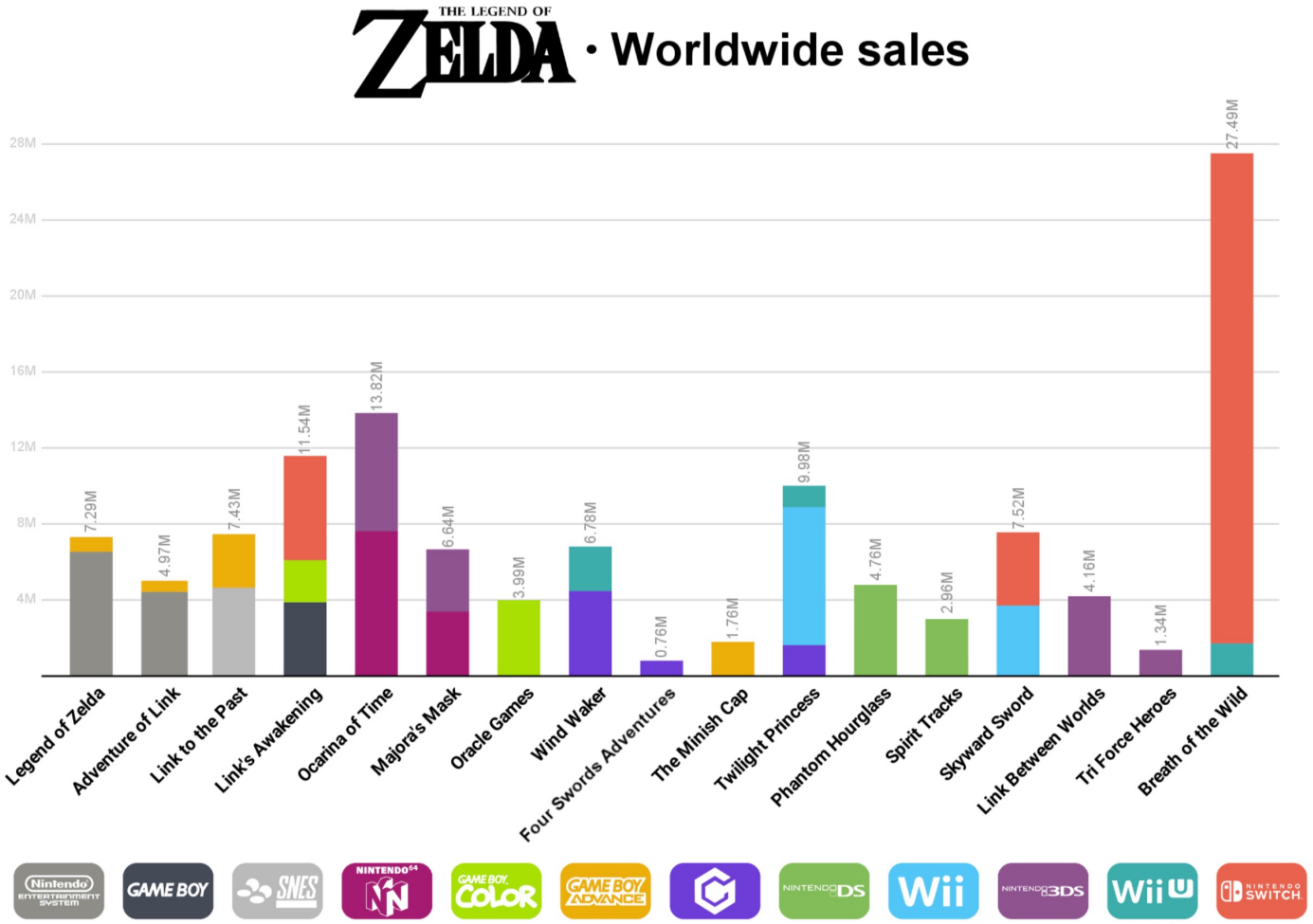
Bar Chart

- Double Bar Char



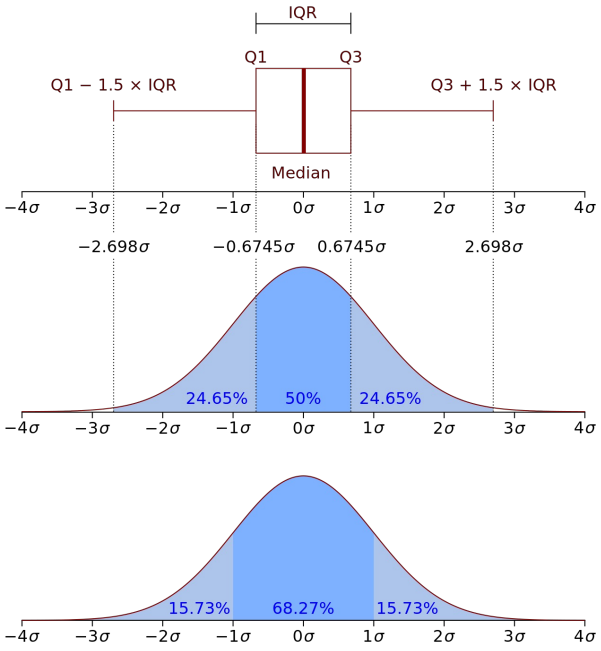
- https://github.com/liufly/Dual-scale-D3-Bar-Chart

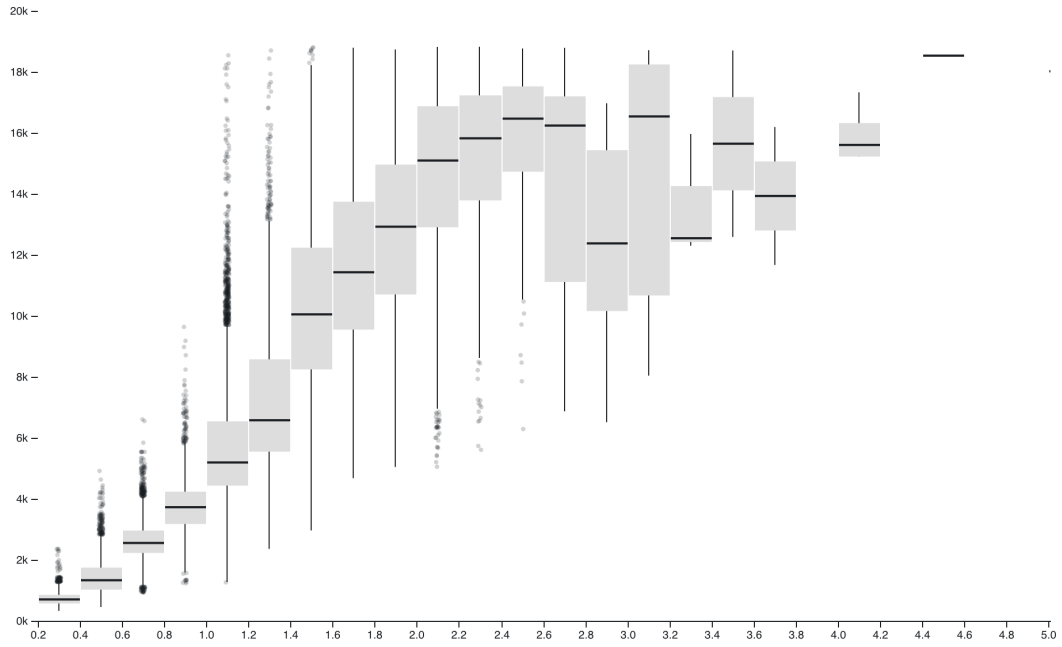
Bar Chart


THE LEGEND OF ZELDA · Worldwide sales
28M
24M
20M
16M
12M
8M
4M
7.29M
4.97M
7.43M
11.54M
13.82M
6.64M
3.99M
6.78M
0.76M
1.76M
9.98M
4.76M
2.96M
7.52M
4.16M
1.34M
27.49M
Legend of Zelda
Adventure of Link
Link to the Past
Link's Awakening
Ocarina of Time
Majora's Mask
Oracle Games
Wind Waker
Four Swords Adventures
The Minish Cap
Twilight Princess
Phantom Hourglass
Spirit Tracks
Skyward Sword
Link Between Worlds
Tri Force Heroes
Breath of the Wild

Box Plot

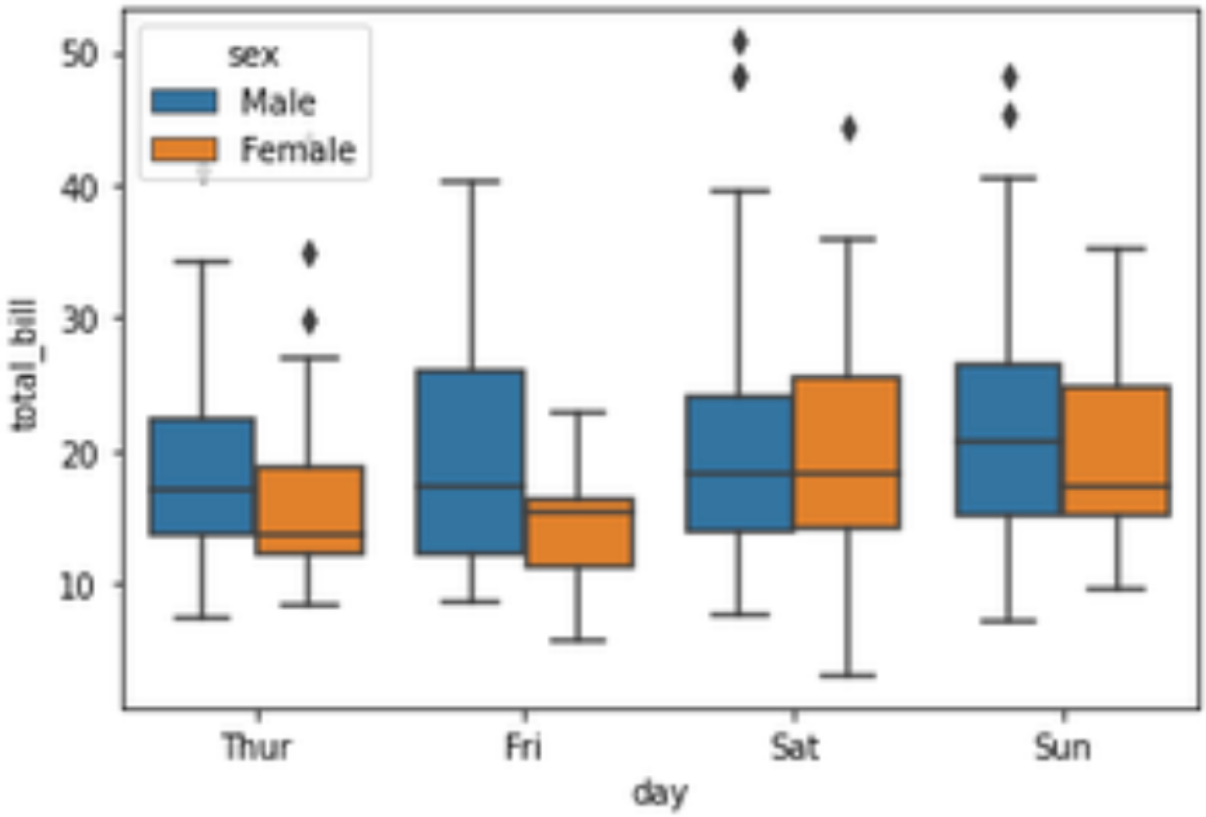
- In descriptive statistics, a box plot or boxplot is a method for graphically depicting groups of numerical data through their quartiles.





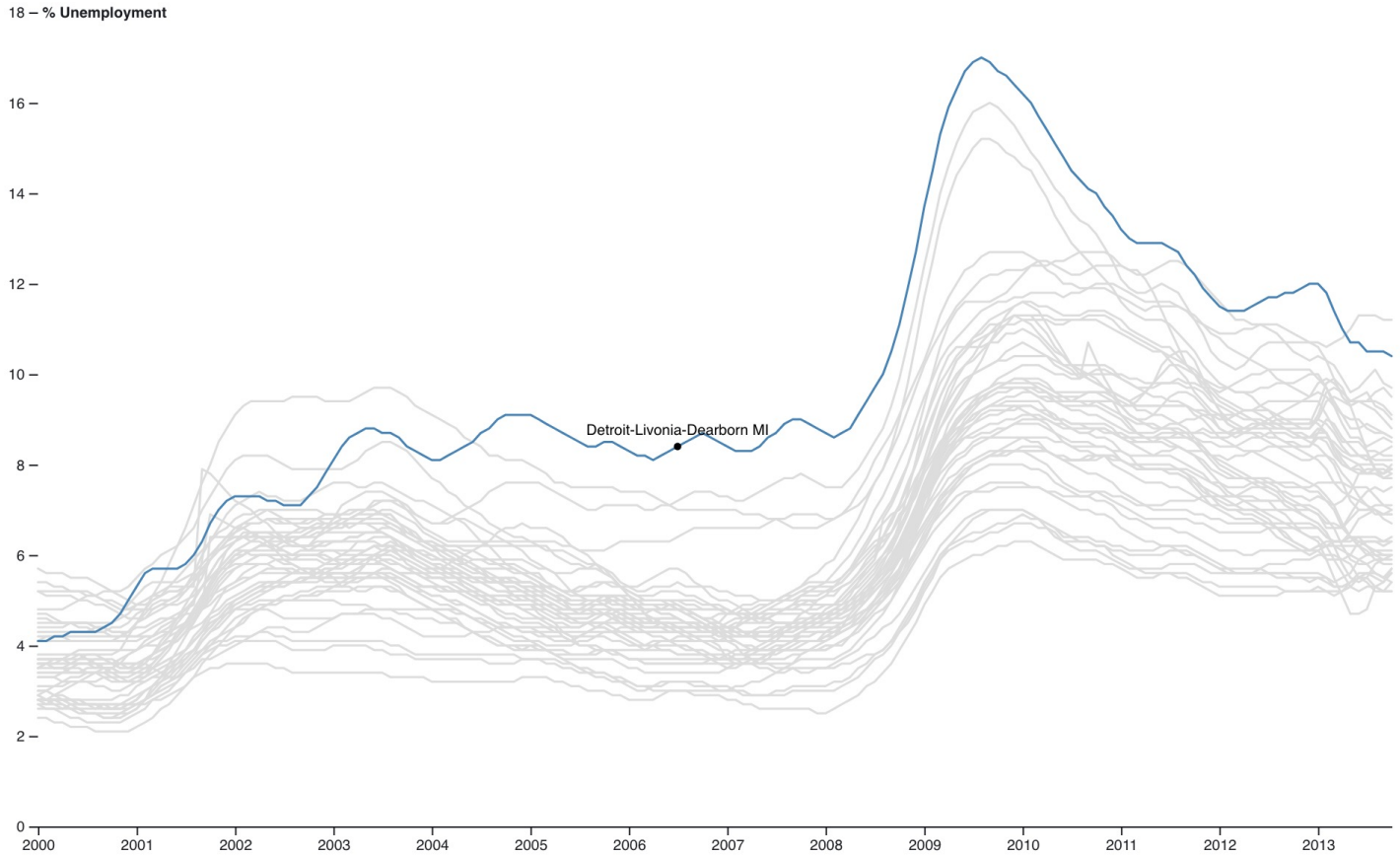
- https://beta.observablehq.com/@mbostock/d3-box-plot

Box Plot



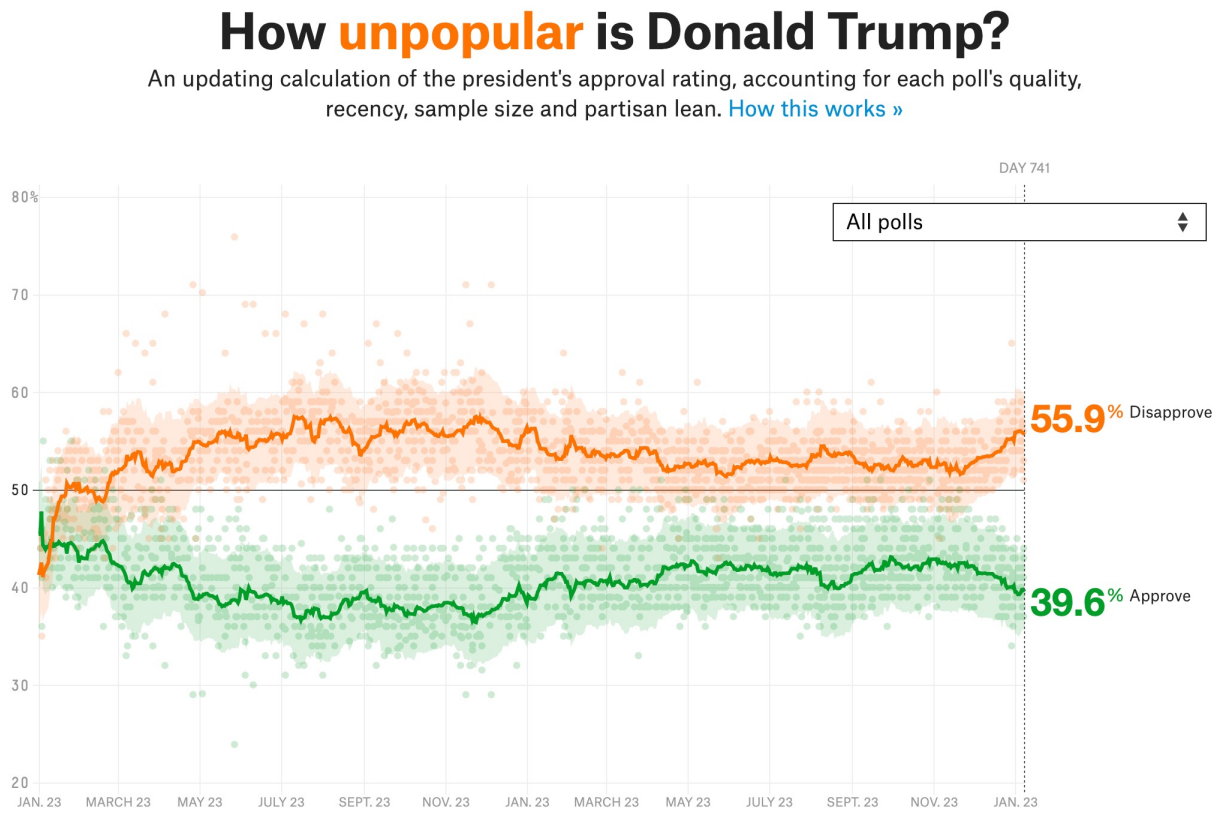
https://rdrr.io/cran/reshape2/man/tips.html

Time Series



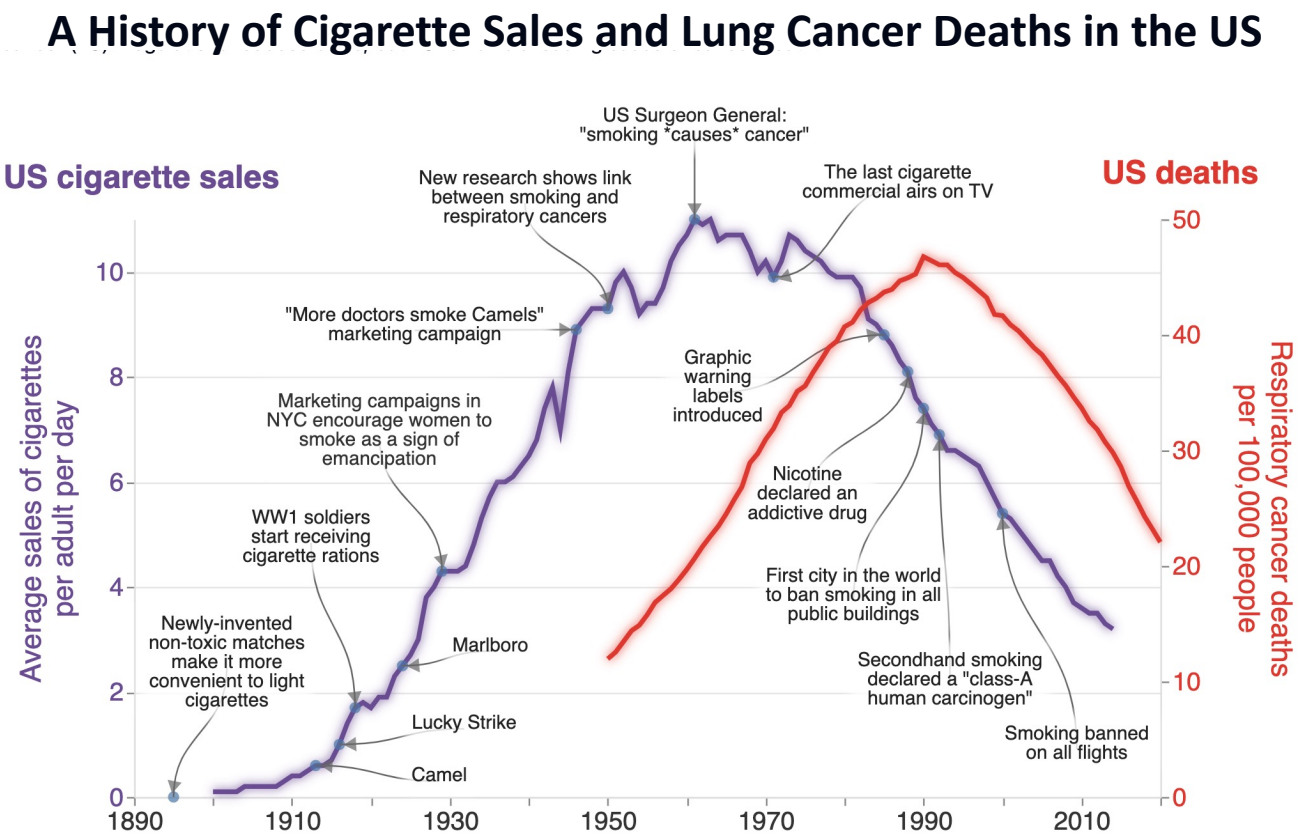
- https://beta.observablehq.com/@mbostock/d3-multi-line-chart

Time Series



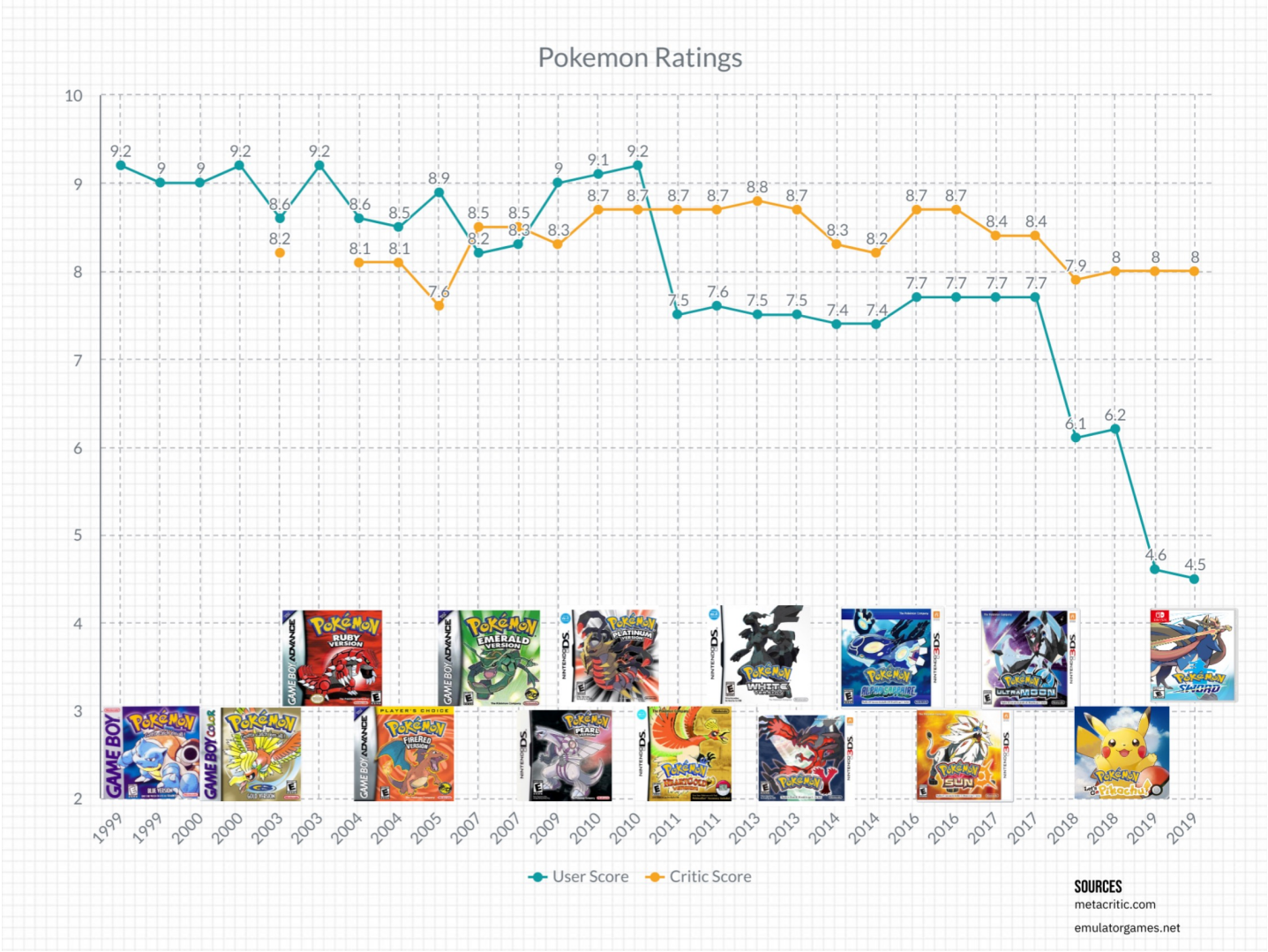
- https://projects.fivethirtyeight.com/trump-approval-ratings

Time Series

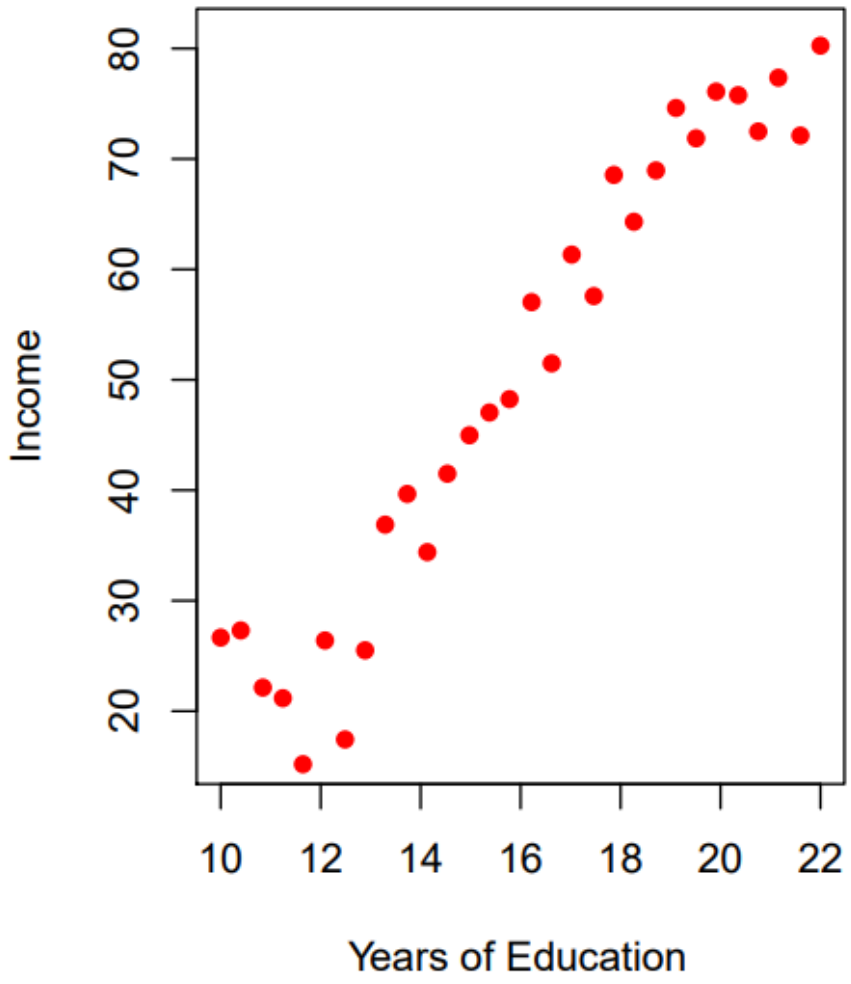


https://www.contextualize.ai/mpereira/a-history-of-cigarette-sales-and-lung-cancer-deaths-in-the-us-ce8dea7a

Time Series

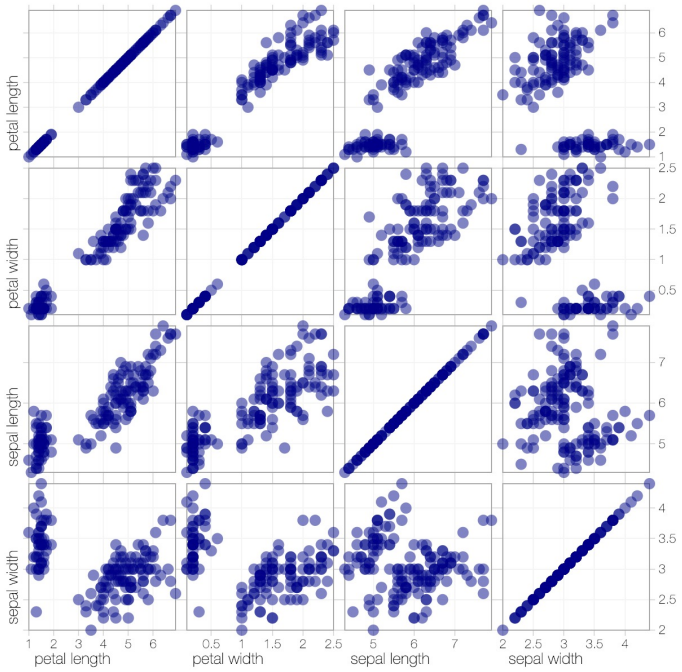

Pokemon Ratings
10
9
8
7
6
5
4
3
2
9.2 9 9 9.2 8.6 9.2 8.6 8.5 8.9 8.2 8.5 8.3 9 9.1 9.2 7.5 7.6 7.5 7.5 7.4 7.4 7.7 7.7 7.7 7.7 6.1 6.2 4.6 4.5
8.2 8.1 8.1 7.6 8.5 8.3 8.3 8.7 8.7 8.7 8.7 8.8 8.7 8.3 8.2 8.7 8.7 8.4 8.4 7.9 8 8 8
1999 1999 2000 2000 2003 2003 2004 2004 2005 2007 2007 2009 2010 2010 2011 2011 2013 2013 2014 2014 2016 2016 2017 2017 2018 2018 2019 2019
User Score
Critic Score
SOURCES
metacritic.com
emulatorgames.net

Scatterplot



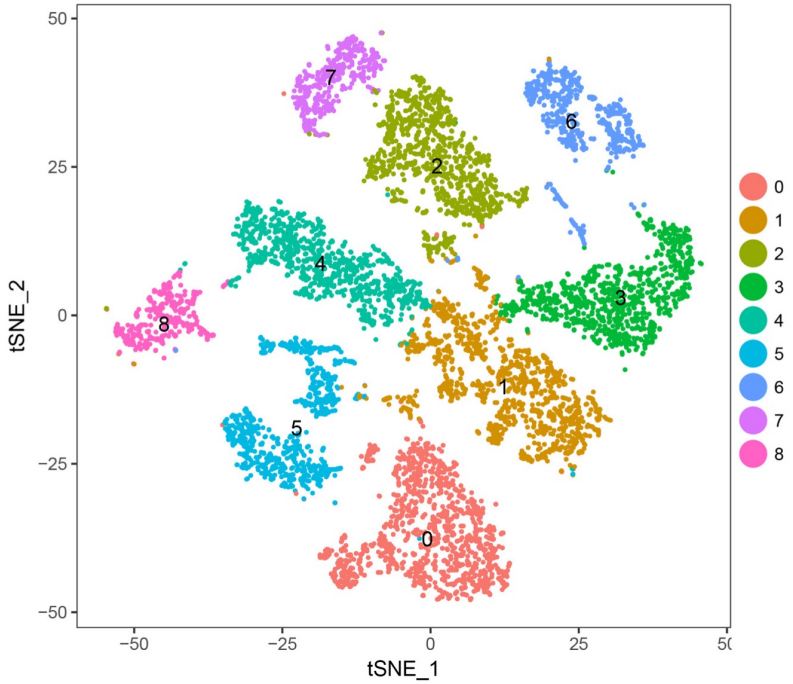
Scatterplot Matrix

- Pairwise correlations for multi-dimensional data; (Thus, the diagonal cells are diagonal lines.)
- This matrix shows Anderson's data on iris flowers on the Gaspé Peninsula.

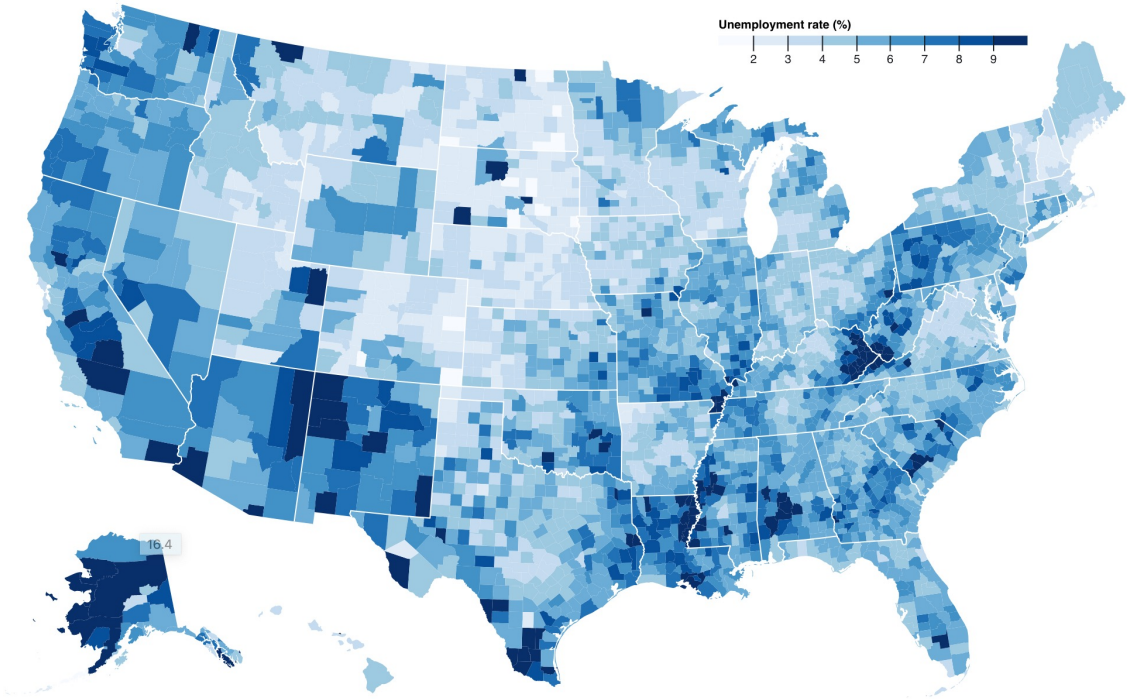


- http://benjiec.github.io/scatter-matrix/demo/demo.html

Scatter Plot for clustering

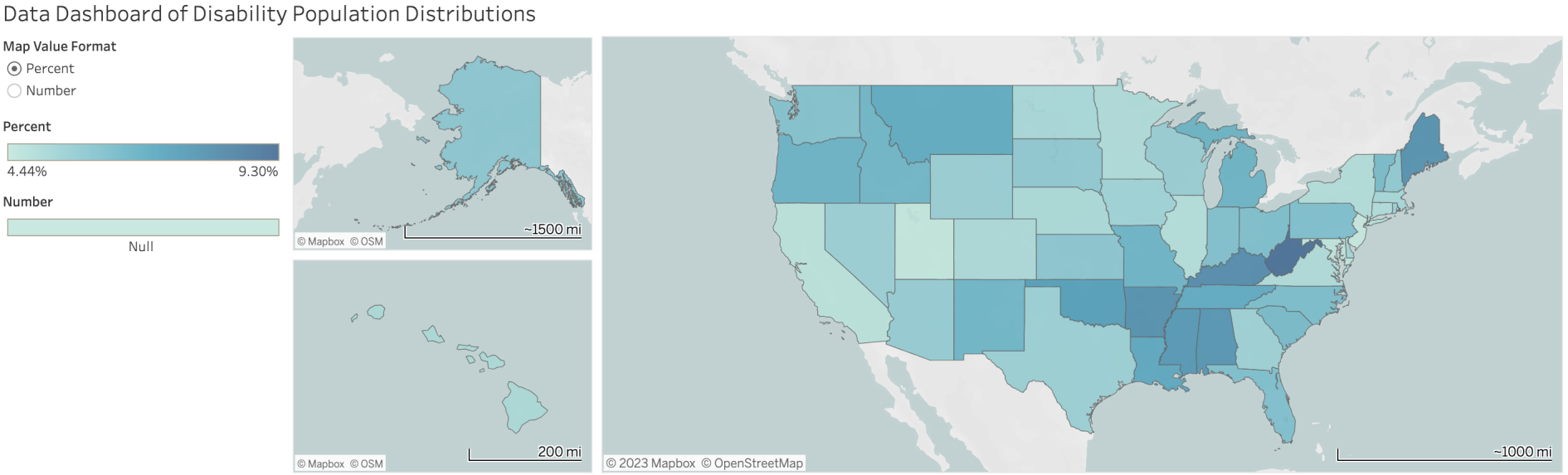


Map: Choropleth map



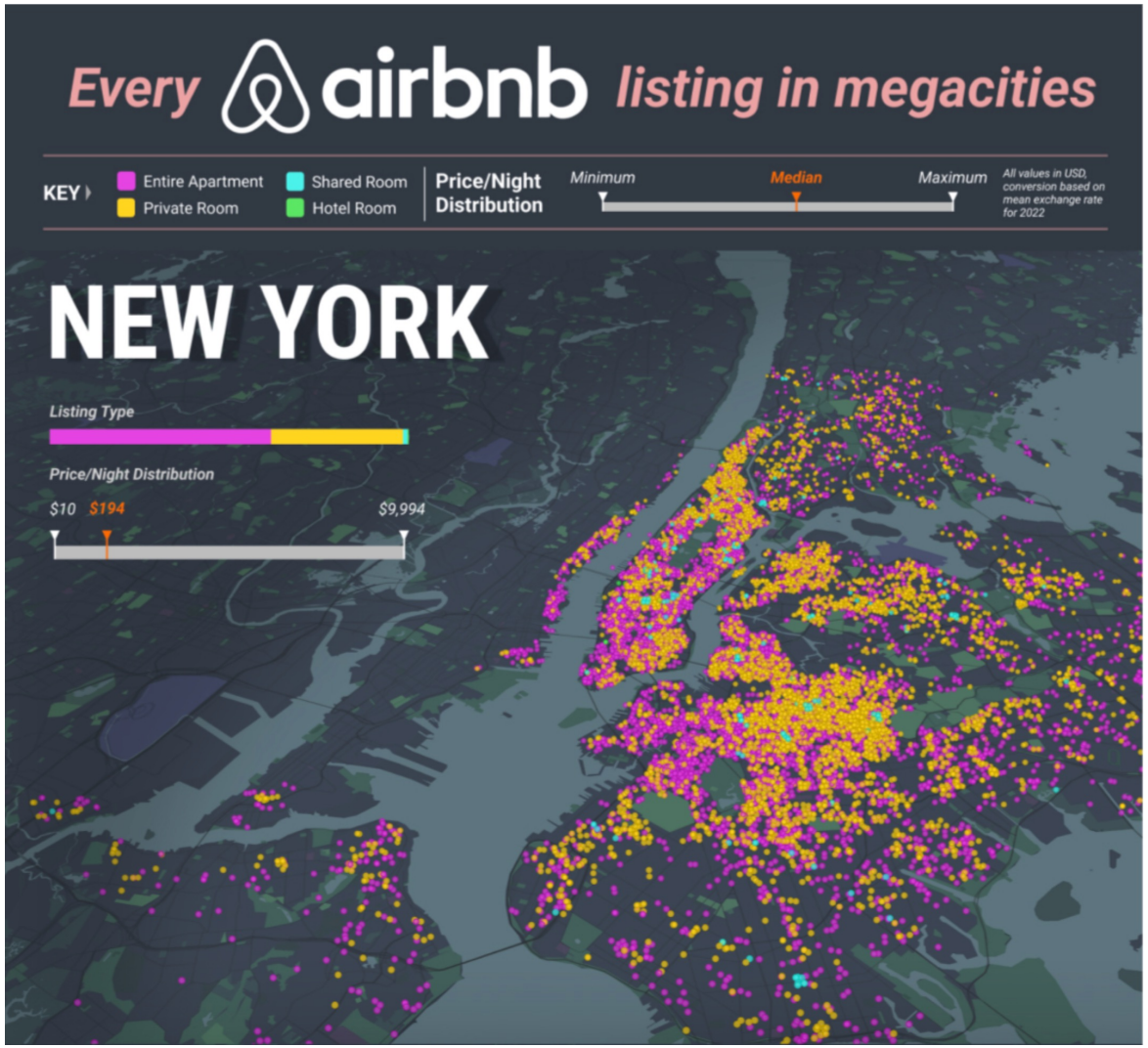
- https://beta.observablehq.com/@mbostock/d3-choropleth

Choropleth map: U.S. Disability Population Statistics



https://public.tableau.com/app/profile/jiawei.wang8332/viz/DisbilityPopulationStatistics-Year11-14/MainDashboard

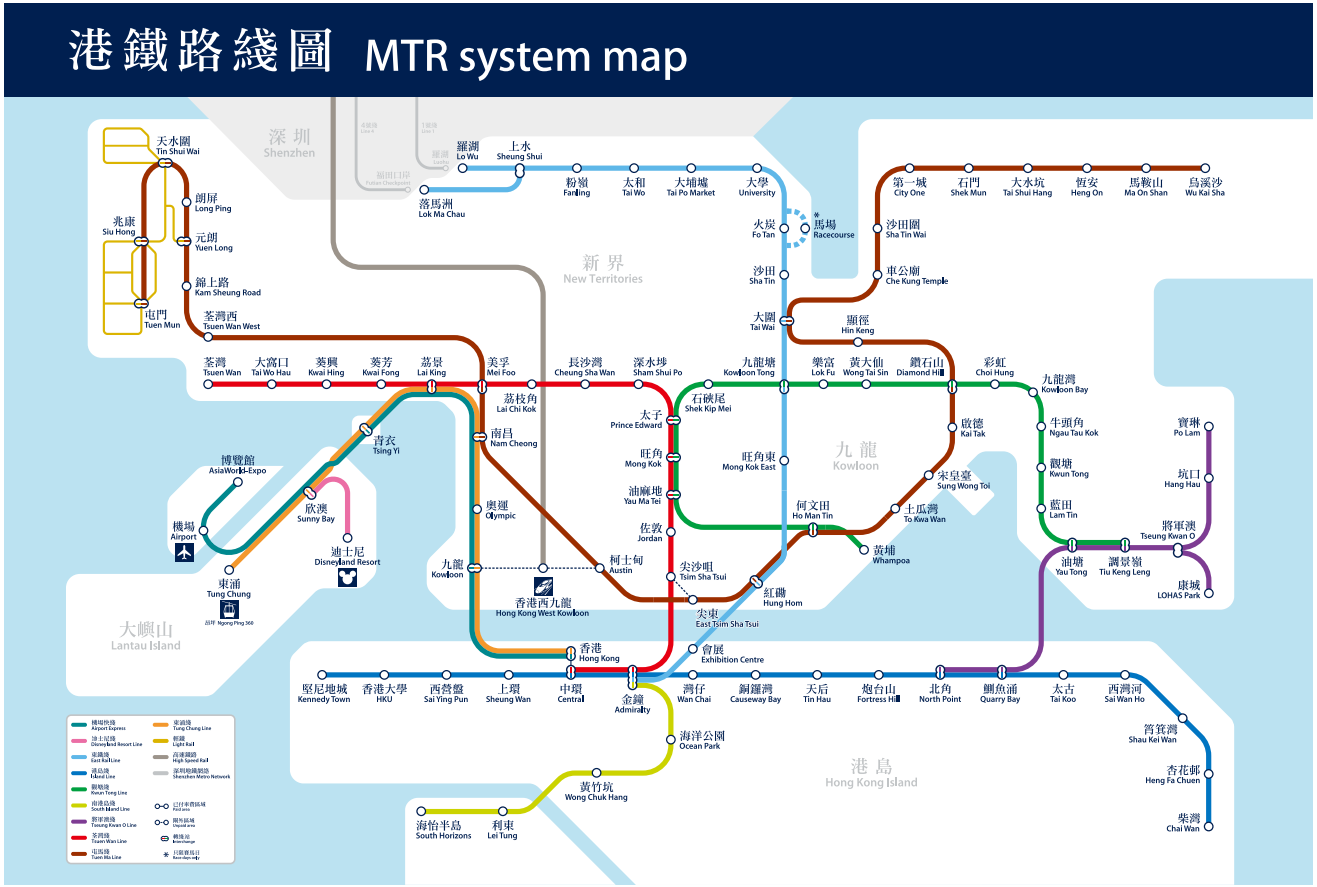
Map plot


Every airbnb listing in megacities
KEY
Entire Apartment
Private Room
Shared Room
Hotel Room
Price/Night Distribution
Minimum
Median
Maximum
All values in USD, conversion based on mean exchange rate for 2022
NEW YORK
Listing Type
Price/Night Distribution
$10
$194
$9,994

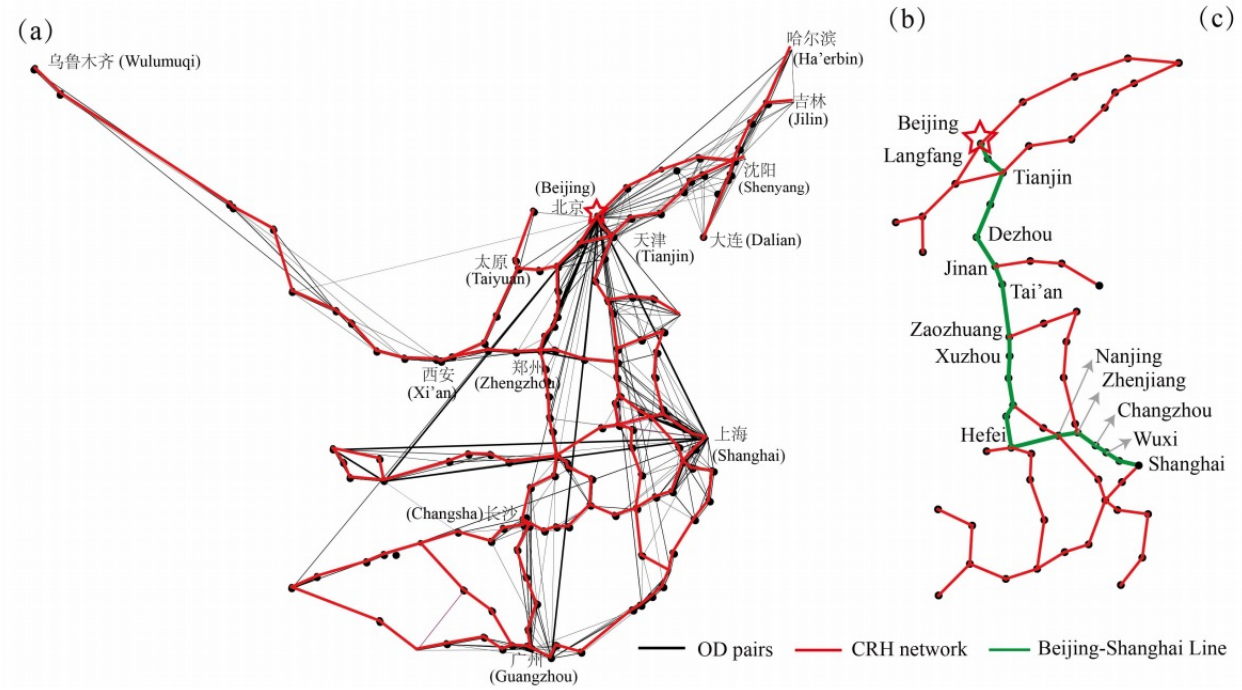
Transportation Map

https://www.cathaypacific.com/cx/en_HK.html

Transportation Map



Transportation Map



(c)

From CITY To Beijing

CITY	# of trips
Langfang	7220
Baoding	6144
Tianjin	8532
Guangzhou	7465
Chengdu	7278
Shanghai	7021
Wuhan	6609
Shijiazhuang	6549
Zhengzhou	7829
Shenzhen	7370

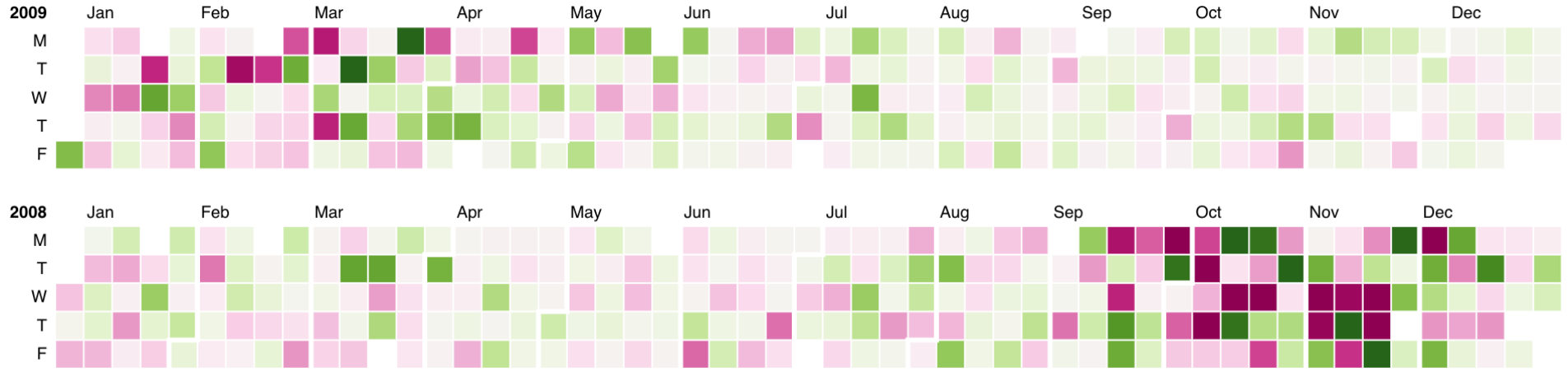
From Beijing To CITY

CITY	# of trips
Langfang	8577
Tianjin	7675
Guangzhou	7619
Baoding	7219
Shenzhen	6585
Chengdu	6551
Shanghai	9626
Wuhan	9630
Shijiazhuang	6856
Zhengzhou	5688

- https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=8710615

Calendar View

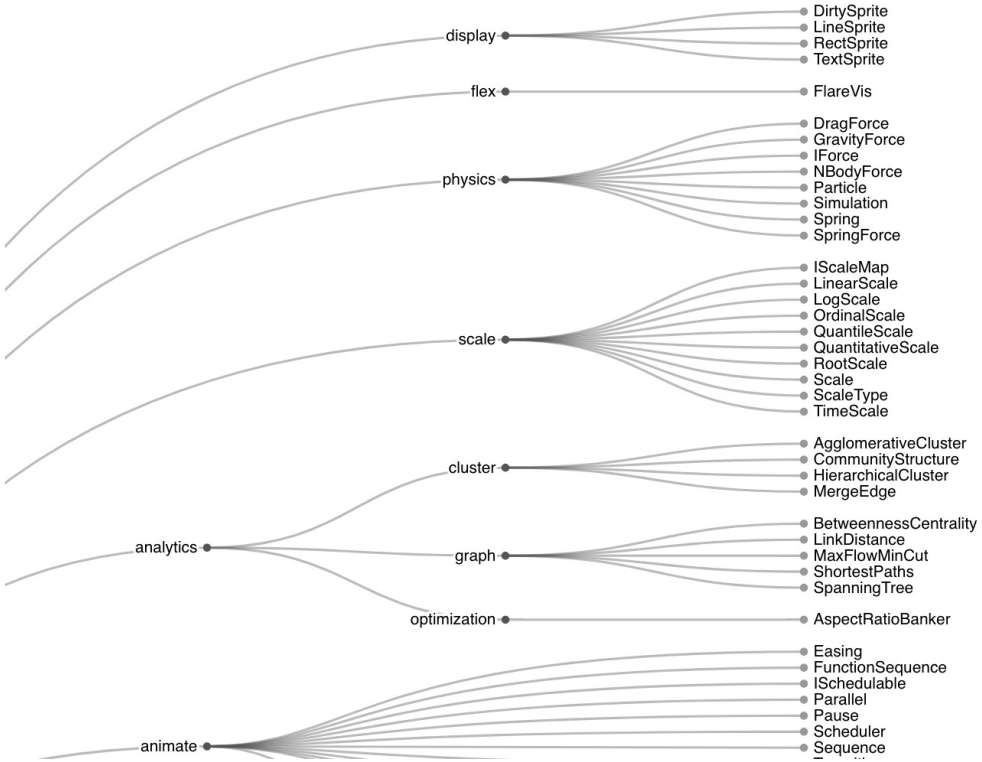
- This chart shows daily changes of the Dow Jones Industrial Average from 1990 to 2010. Days the index went up are green; days the index went down are pink.



- https://beta.observablehq.com/@mbostock/d3-calendar-view

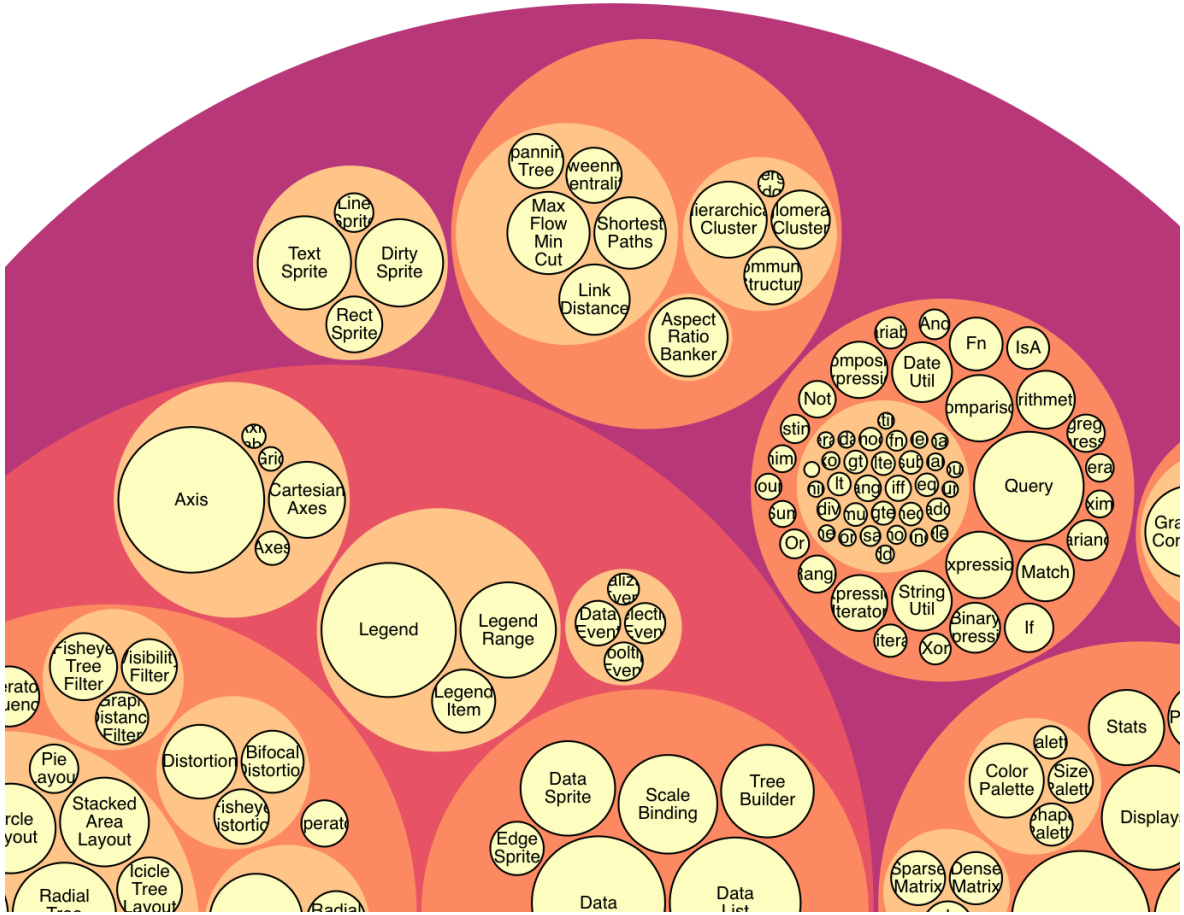
Hierarchical Diagram

- Tree – shows the organizational structure



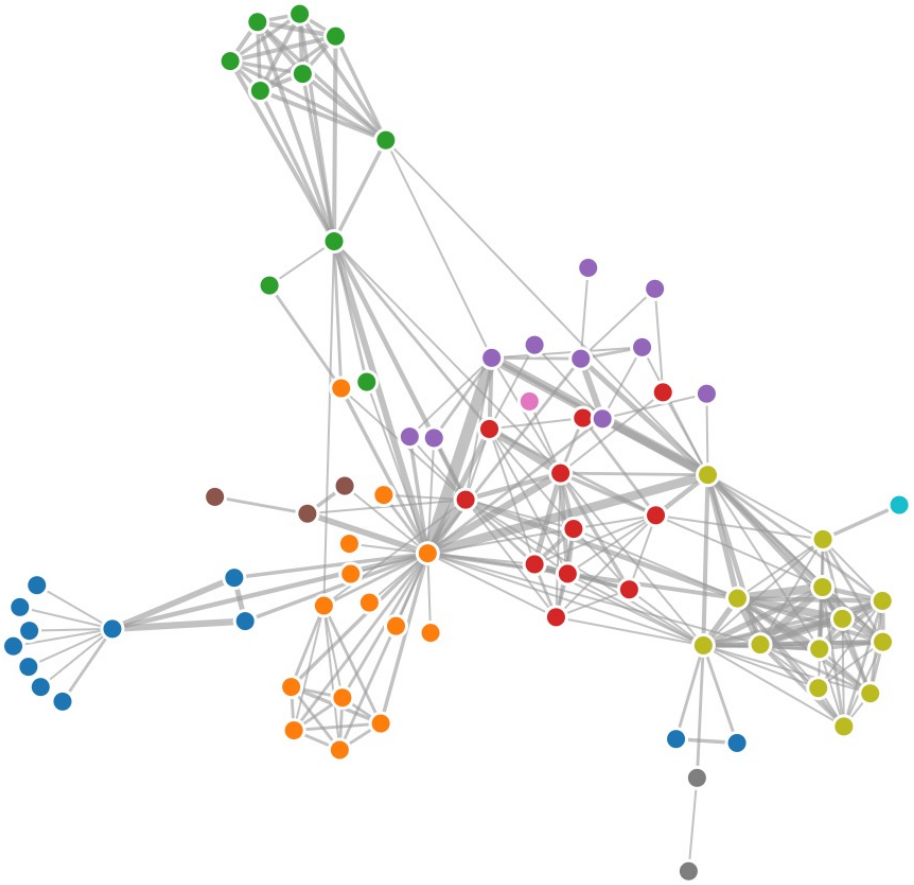
- https://beta.observablehq.com/@mbostock/d3-cluster-dendrogram

Circle Packing



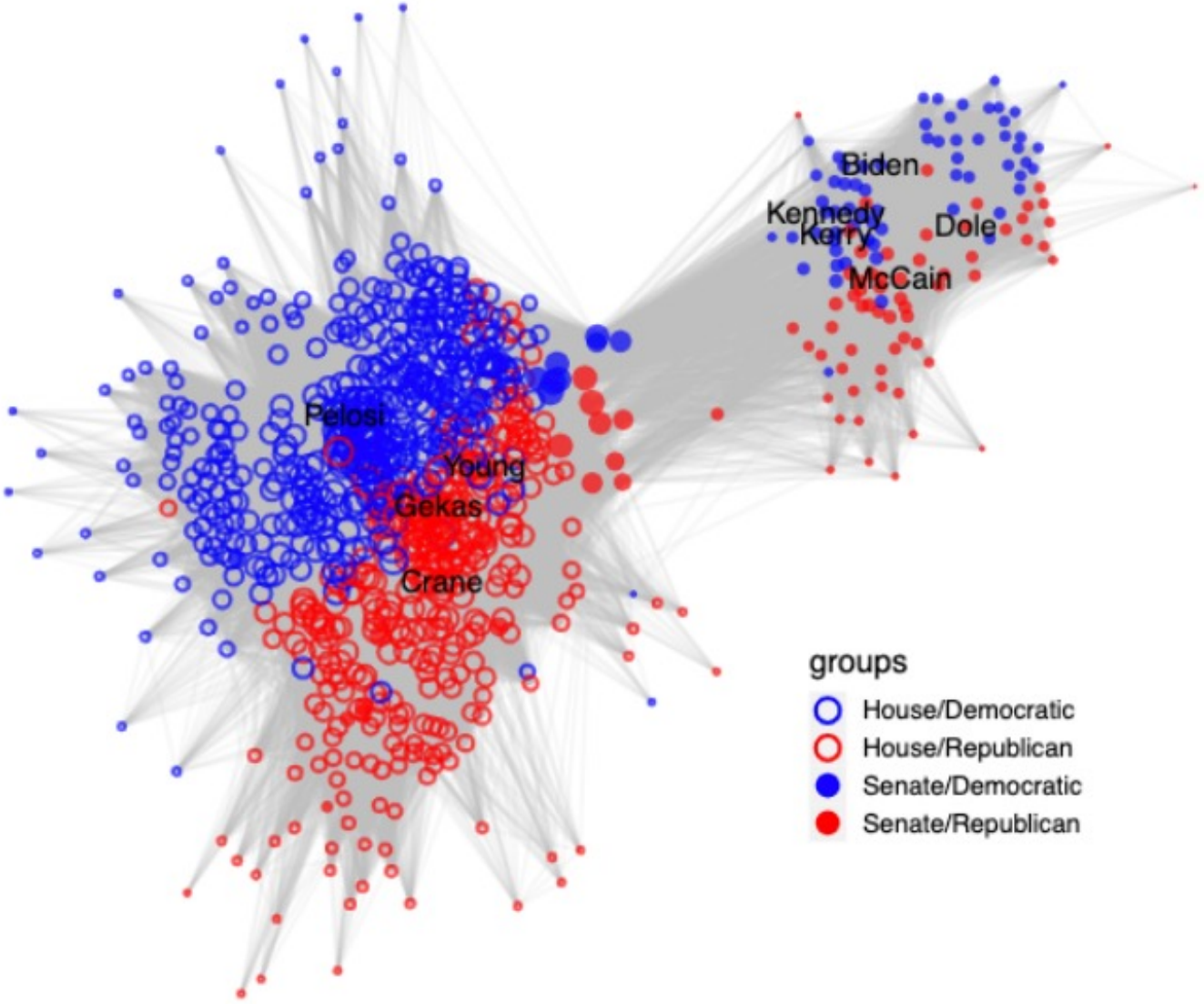
- https://beta.observablehq.com/@mbostock/d3-circle-packing

Social Network

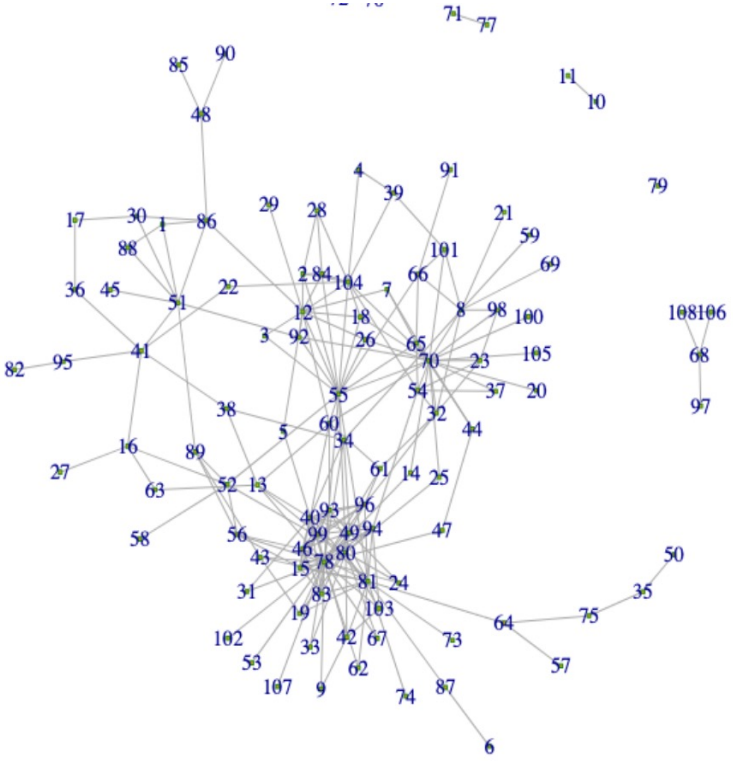


- https://beta.observablehq.com/@mbostock/d3-force-directed-graph
- http://vax.herokuapp.com/game

Social Network

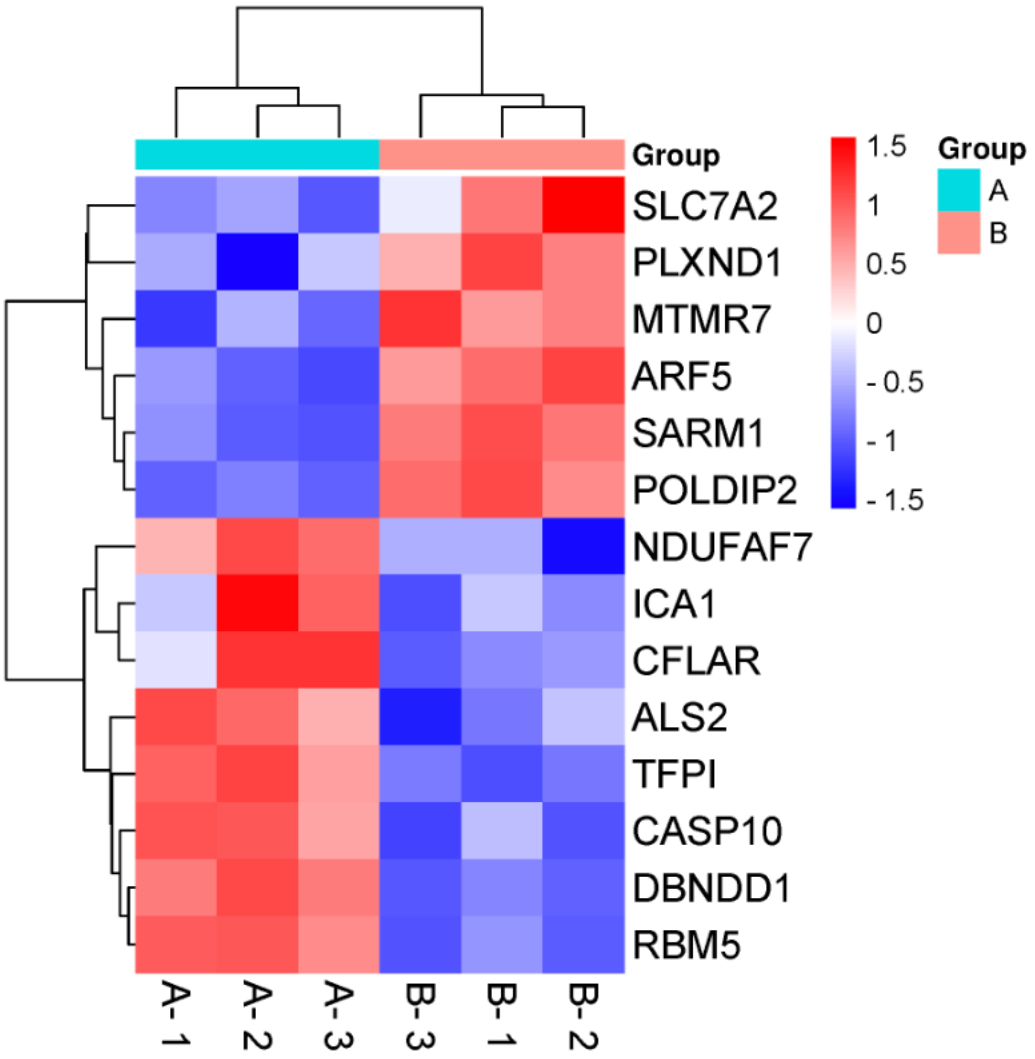


Citation Network



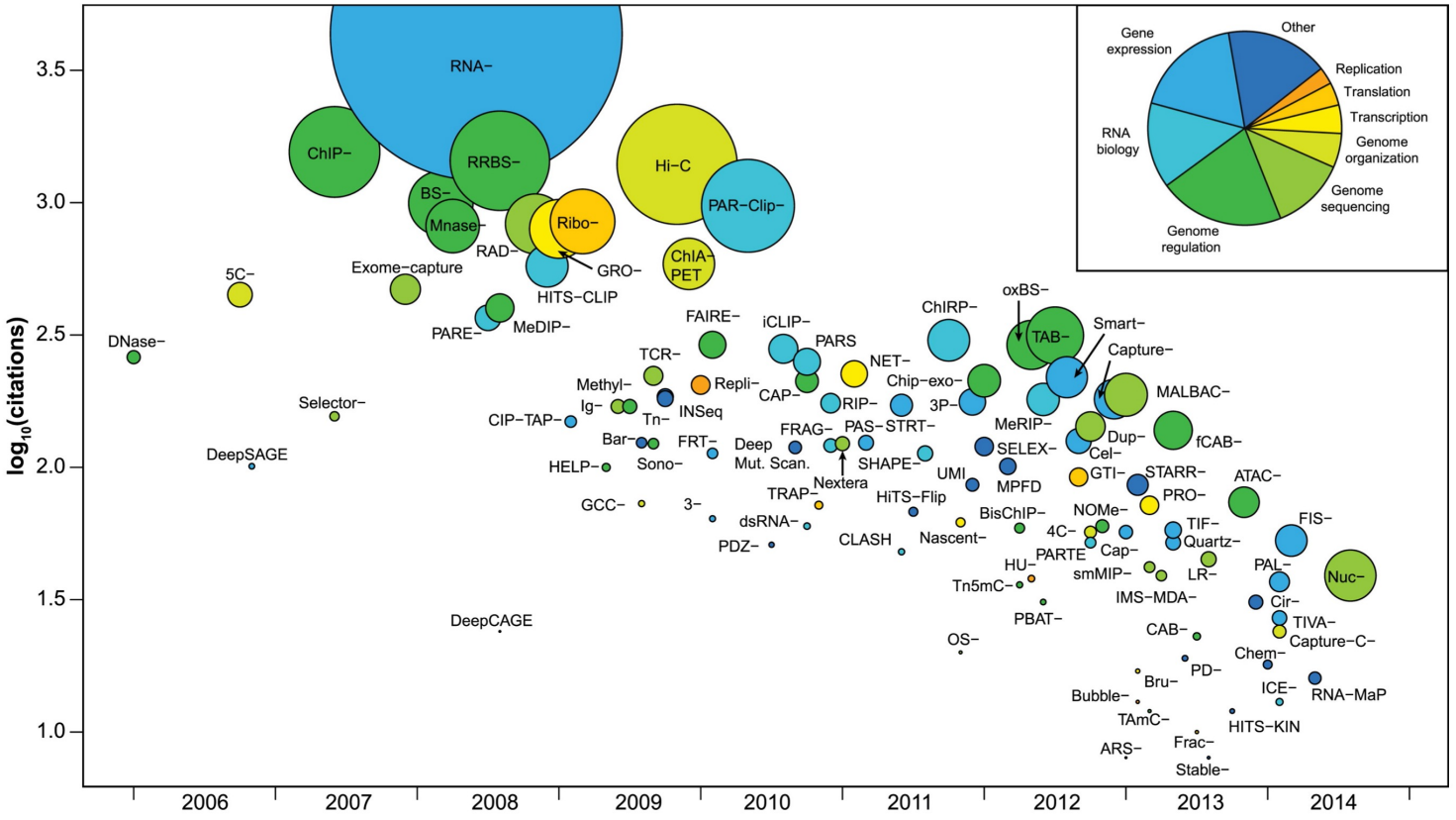
https://www.sciencedirect.com/science/article/pii/S2666389922001295

Heat map



https://www.bioinformatics.com.cn/plot_basic_cluster_heatmap_plot_024_en

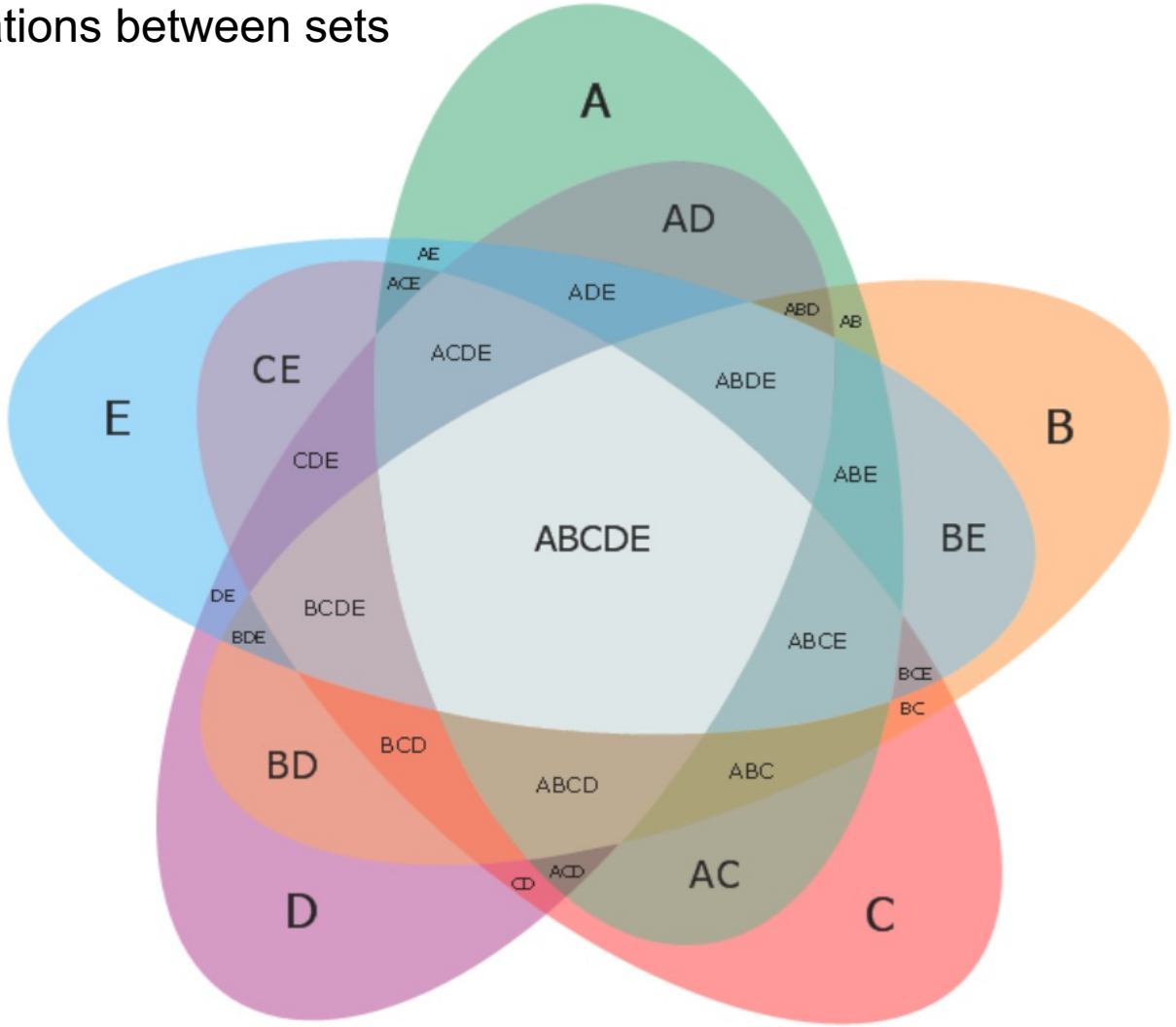
Bubble plots



https://www.cell.com/molecular-cell/pdf/S1097-2765(15)00340-8.pdf

Venn diagram

- Presents relations between sets



Question

- Have you learnt Machine Learning Courses Before?
- Are you familiar with Python?

Getting prepared

- Install Tableau academic version
  - https://www.tableau.com/academic
  - Optional: using the practice datasets to generate some plots, e.g., bar, line and pie plots.

- Setup Python programming environment.
  - Preparing for the tutorial in week 2

- Finish **QIZE 1** before Wednesday 11pm.