Lecture 2

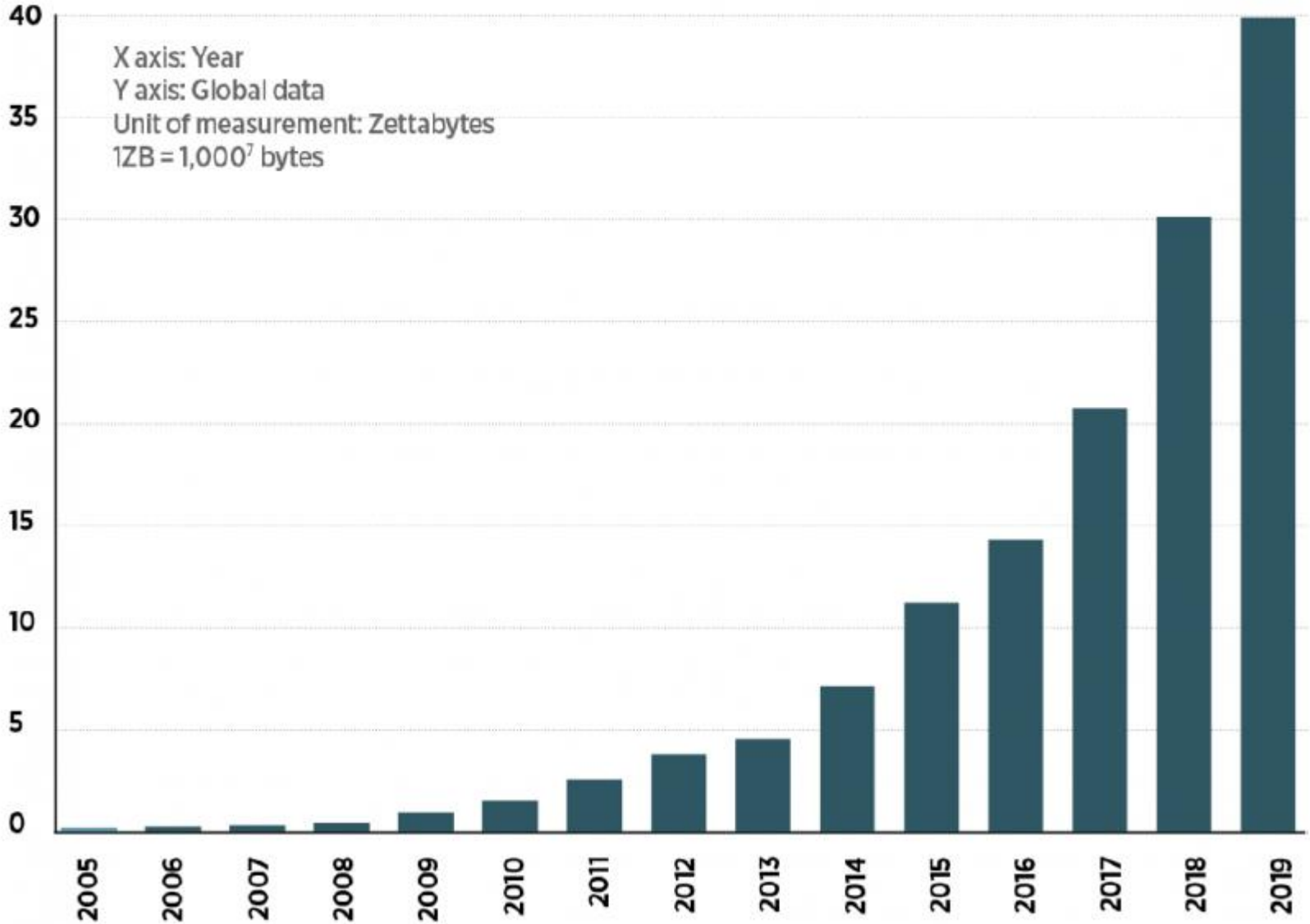
Introduction to Data Science

Today

- Data
- Machine Learning
- Applications of Machine Learning
- Data Science
- Visualizing data
- Jupyter Notebook
- Python programming
- Python packages

Data

DATA GROWTH



Note: Post-2013 figures are predicted. Source: UNECE

Artistic works

Data sources

Social media

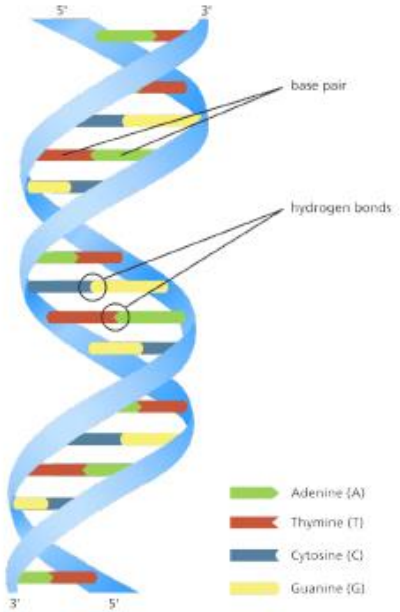
Data sources

Mass media

Data sources

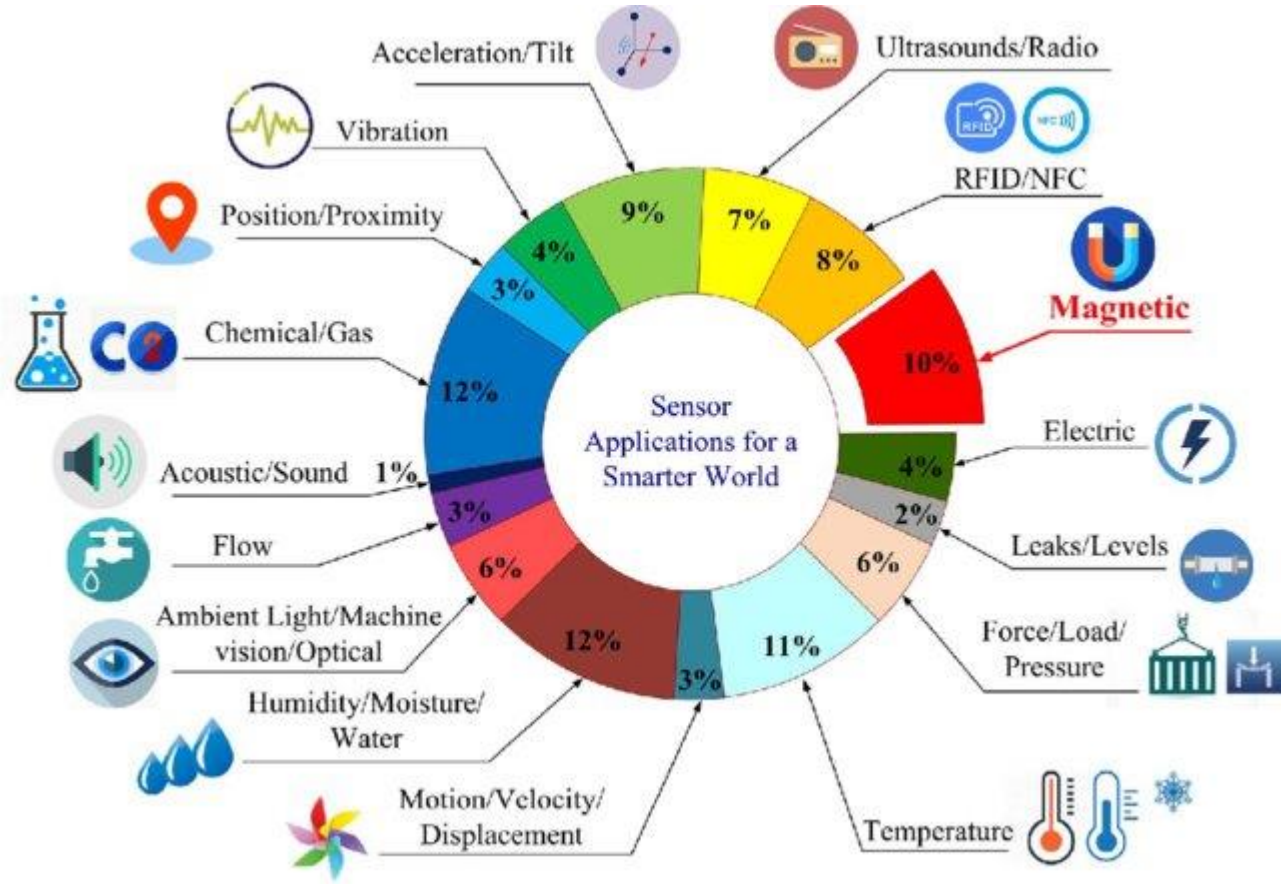
Scientific data

Data sources



Sensor data (for control systems and IoT)

Data sources



Common sensor categories. Image: Liu Xuyang

Machine learning

Learning from data

ML is about learning from data. Here is an example.

HISTORICAL HEALTH RECORDS

x1	x2	Label
98.7	157.6	1
93.6	138.8	0
42.8	171.9	0
92.8	154.5	1

Data: how different patients with characteristics x1 and x2 reacted to a treatment: well = 1 or badly = 0.

Learning Algorithm

A new patient

x1 x2
85.2, 160.3

Model

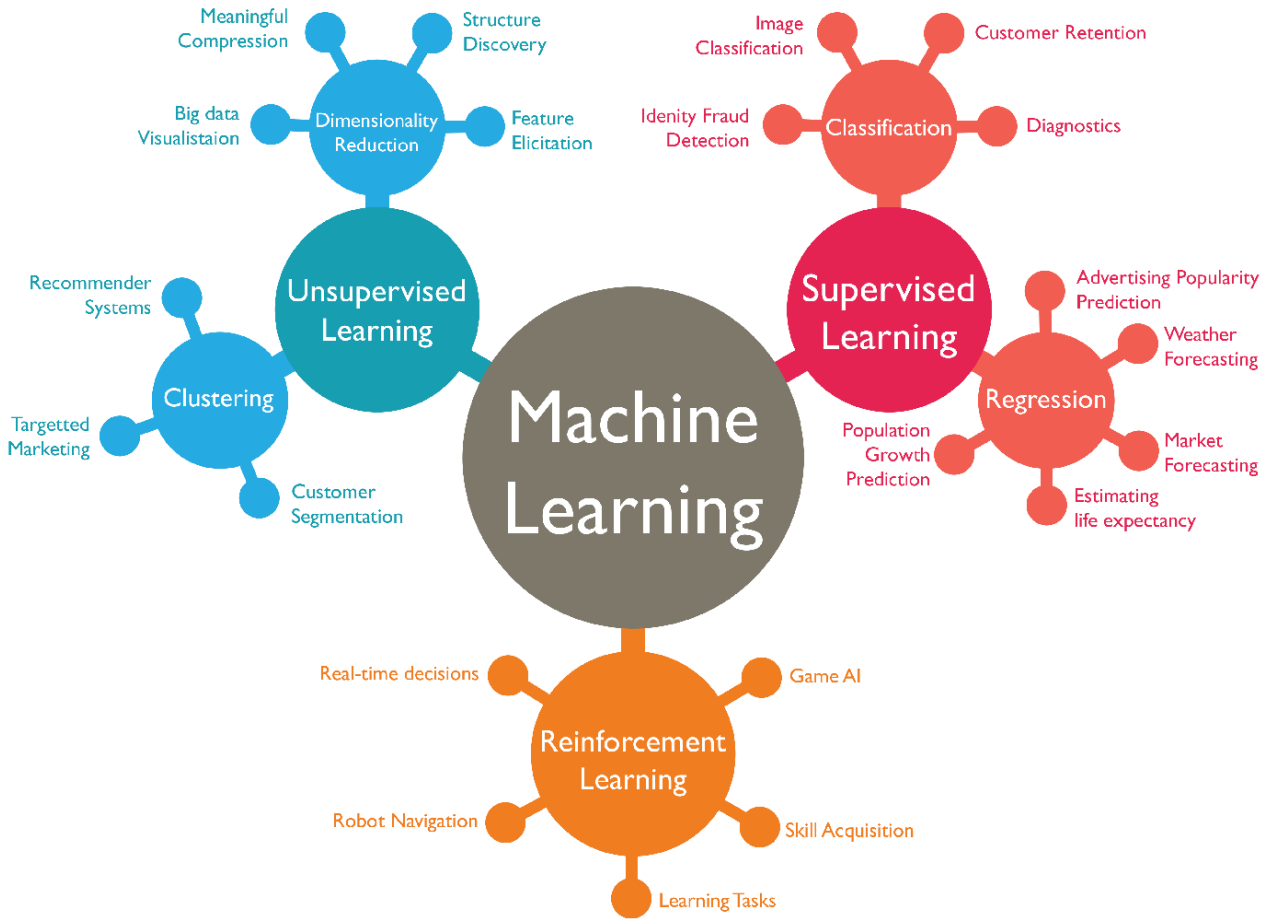
Predicted Health Status

1

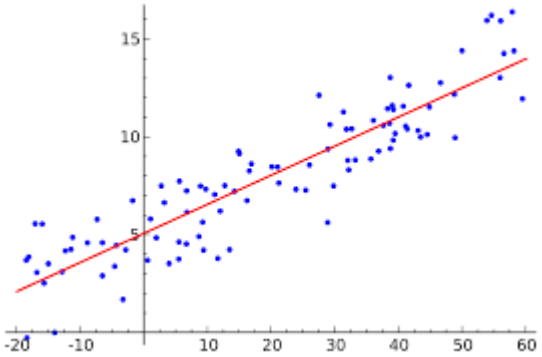
Source

Types of machine learning

Different ways of learning from data



Regression

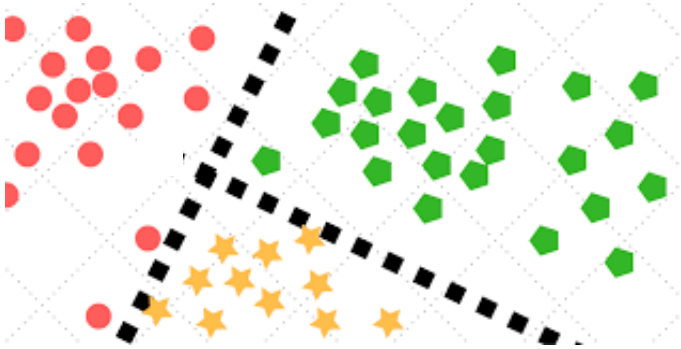


Map data points to numbers

- What is the market price of that house?
- How much will it snow tomorrow?
- How many people will retweet that tweet?
- What will the price of this stock be in one hour?
- What is the temperature in that room?
- How much will we harvest next year?
- How much will we sell next month?

Coursera

Classification

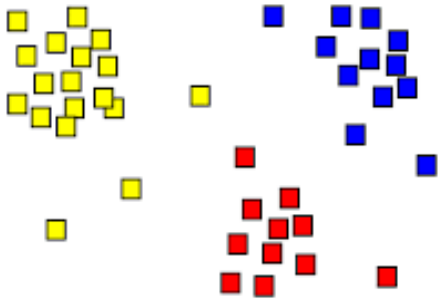


Map data points to labels (classes)

- Will this treatment help that person?
- Will this person pay back that loan?
- Will this person like that book?
- Is this email spam or not?
- Is this review positive or negative or neutral?
- What musical genre does this song belong to?
- What breed of dog does this picture show?

Coursera

Clustering



Form groups (clusters) of data points

- What distinct groups are there in your customer base?
- Who likes who on that social medium?
- Are there some suspect cases of credit card fraud (outlier detection)?
- What kind of microbes are there in this sample?
- Which animals are related to each other?
- Which molecules have similar properties?
- What types of land use do those satellite images show?

Tutorialspoint

Reinforcement learning

Learn (rewarded) behavior from experience

- Drive an autonomous car
- Decide the next treatment step for a Sepsis patient
- Control the cooling system of a Data Center
- Play a game of chess
- Work as a financial trader
- Recommend news items
- Pick and place physical objects

Neptune AI

Applications of Machine Learning

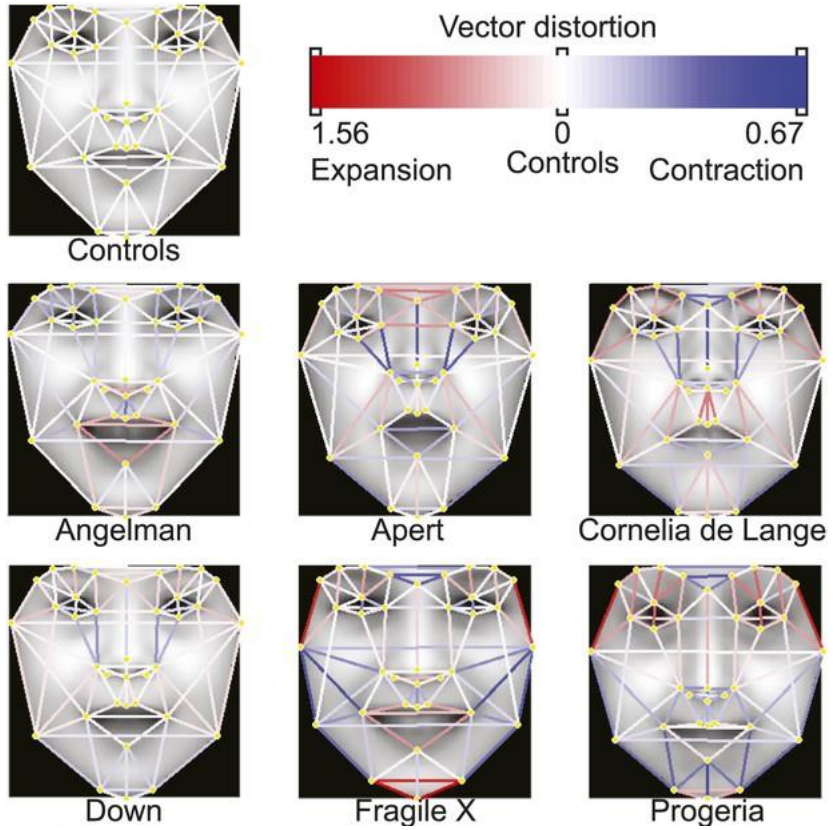
Identify objects

Source

Diagnose diseases

ML for diagnosing rare genetic syndromes based on photos

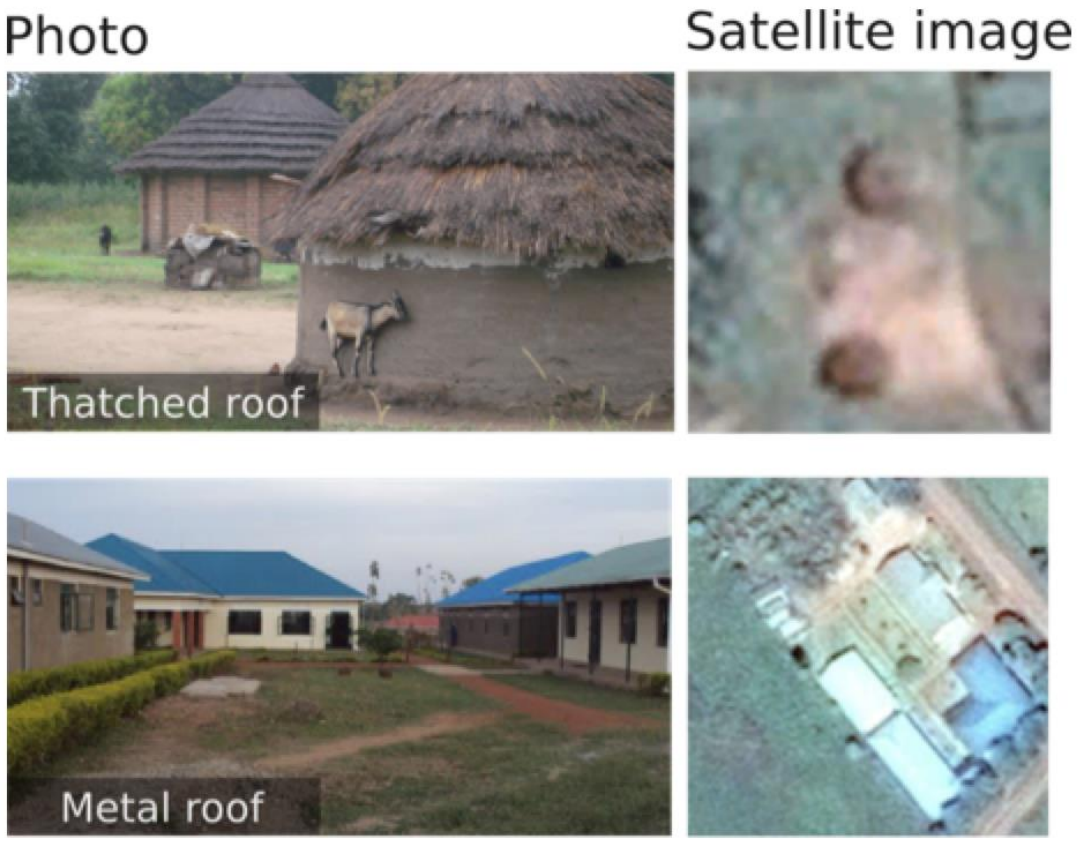
- Rare genetic syndroms are heavily underdiagnosed
- Early diagnosis improves health and quality of life
- Data: Photos of faces of patients with different genetic syndromes and controls



Source

Identify poverty

Identifying poverty by analyzing satellite images and classifying roofs



Source

Find similar novels

A distance measure between texts was used. The distances between several English novels were computed. Who influenced who?

Author	*Title*	Distance	
Austen, Jane	*Pride and Prejudice*	0.000000	1813
Austen, Jane	*Emma*	1.260236	
Austen, Jane	*Sense and Sensibility*	1.268725	
Austen, Jane	*Mansfield Park*	1.421373	
Austen, Jane	*Northanger Abbey*	1.600394	
Austen, Jane	*Persuasion*	1.673071	
Gaskell, Elizabeth	*Ruth*	1.716687	1853
Craik, Dinah Maria	*Olive*	1.745832	1850
Church A. B. Mrs.	*Greymore a Story of Country Life*	1.747513	1860
Grant, Louisa	*Charles Stanley*	1.765758	1854
Tainsh, Edward Campbell	*One Maiden Only*	1.767951	1870

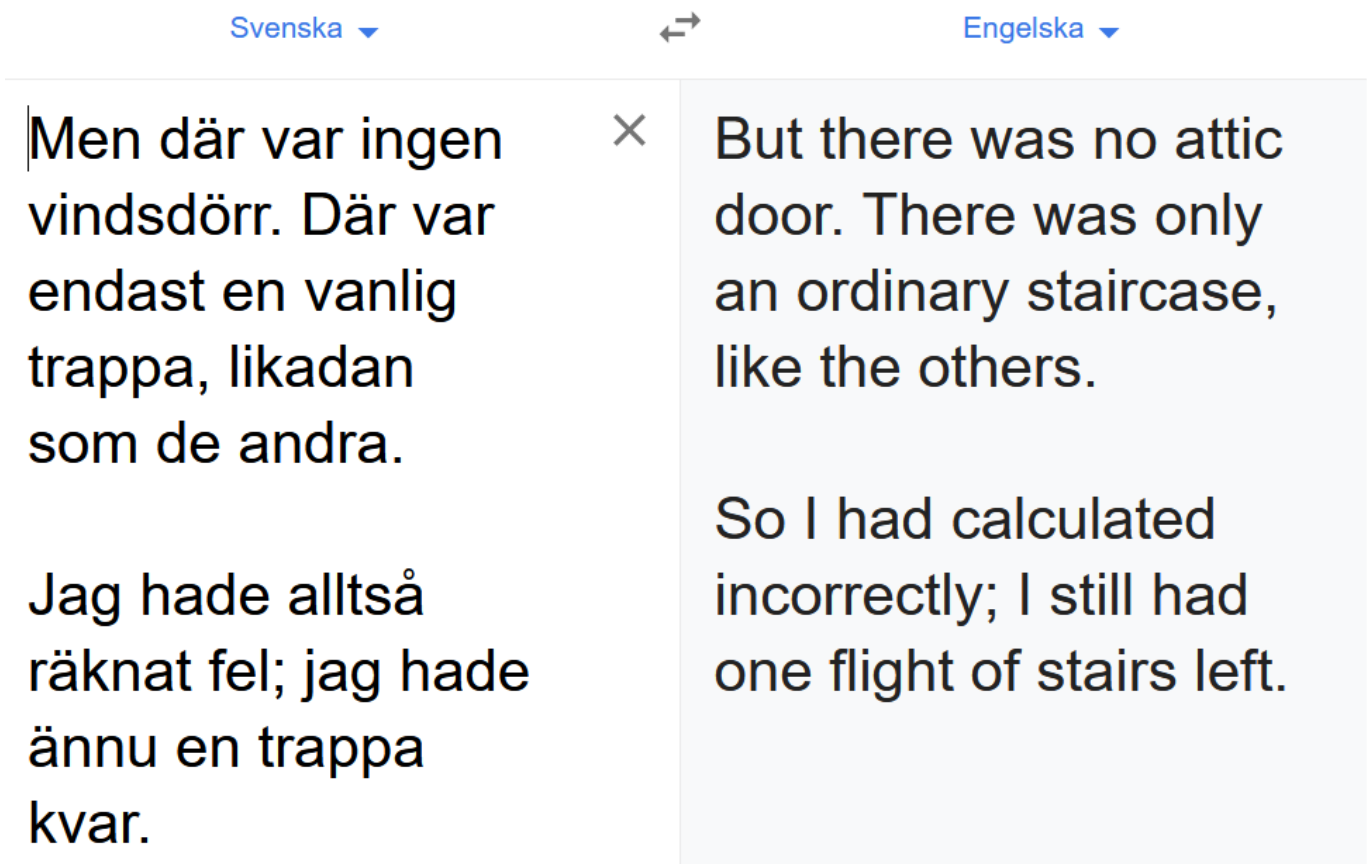
Source

Answer questions

Question	Answer
Where is the Louvre Museum located?	in Paris, France
What's the abbreviation for limited partnership?	L.P.
What are the names of Odin's ravens?	Huginn and Muninn
What currency is used in China?	the yuan
What kind of nuts are used in marzipan?	almonds
What instrument does Max Roach play?	drums
What's the official language of Algeria?	Arabic
How many pounds are there in a stone?	14

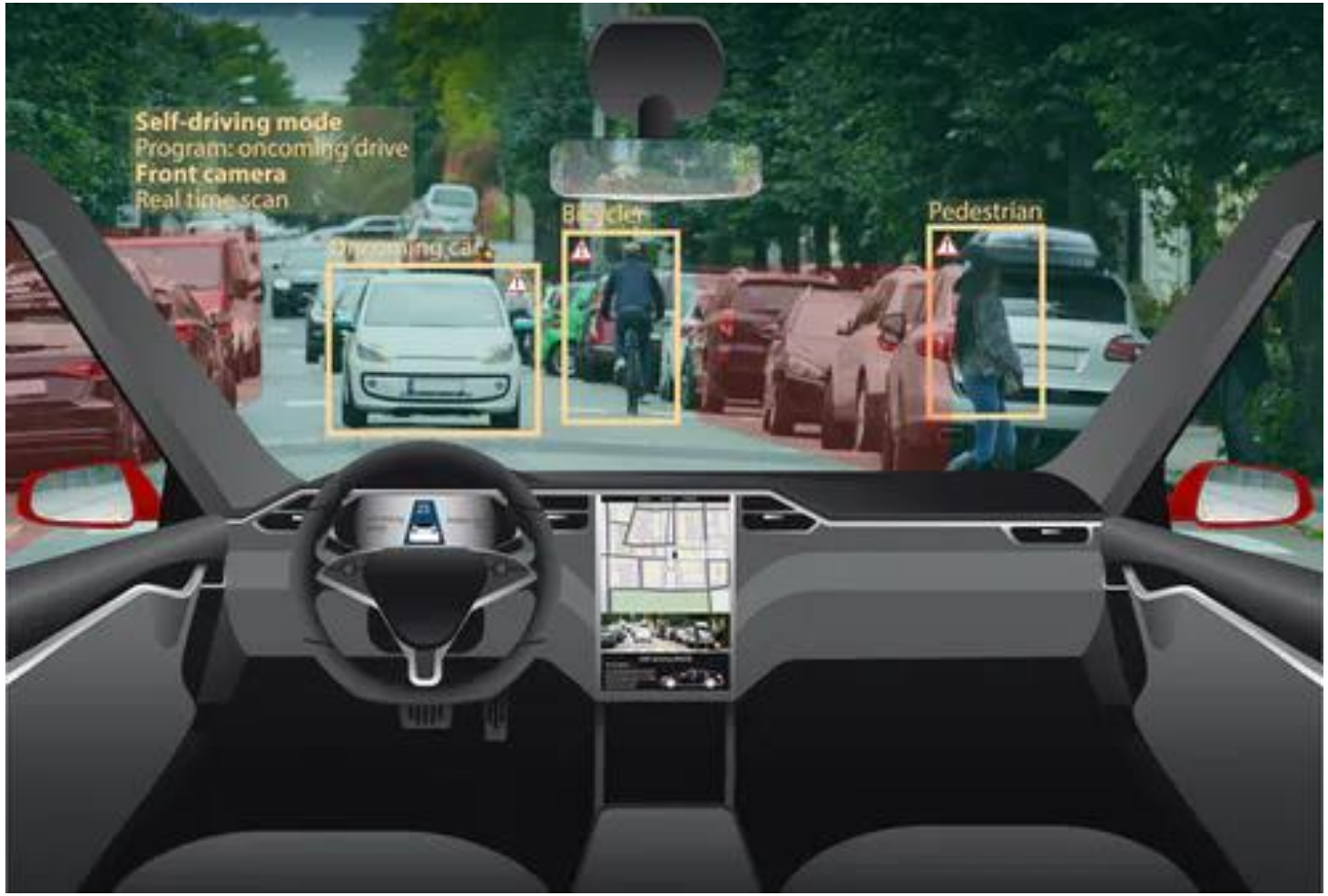
Source

Translate text



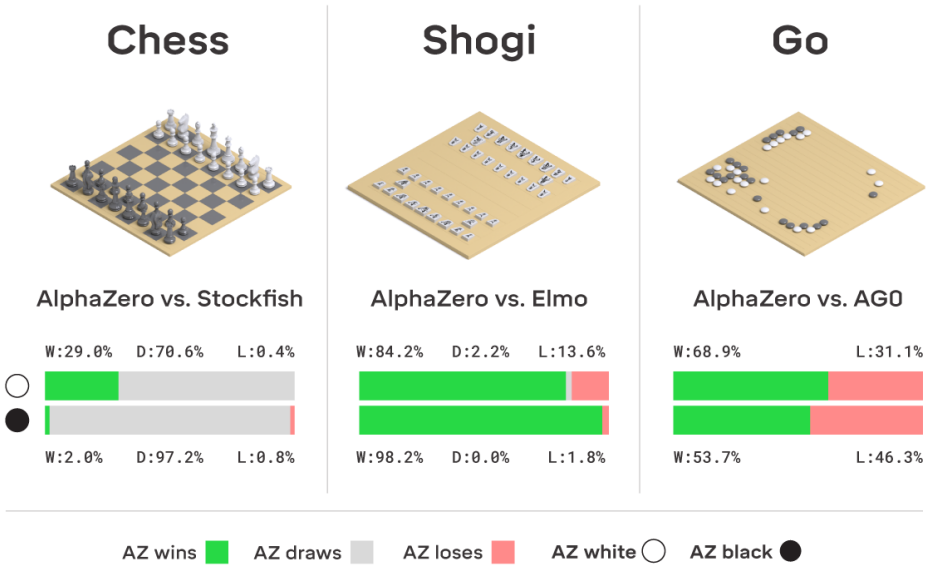
Google Translate

Drive a car



Source

Play games


Chess
Shogi
Go
AlphaZero vs. Stockfish
AlphaZero vs. Elmo
AlphaZero vs. AG0
W:29.0% D:70.6% L:0.4%
W:2.0% D:97.2% L:0.8%
W:84.2% D:2.2% L:13.6%
W:98.2% D:0.0% L:1.8%
W:68.9% L:31.1%
W:53.7% L:46.3%
AZ wins
AZ draws
AZ loses
AZ white
AZ black

Data science

Data science

- The term “Data Science” was used for the first time in 1985 by Jeff Wu as an alternative name for statistics [source]

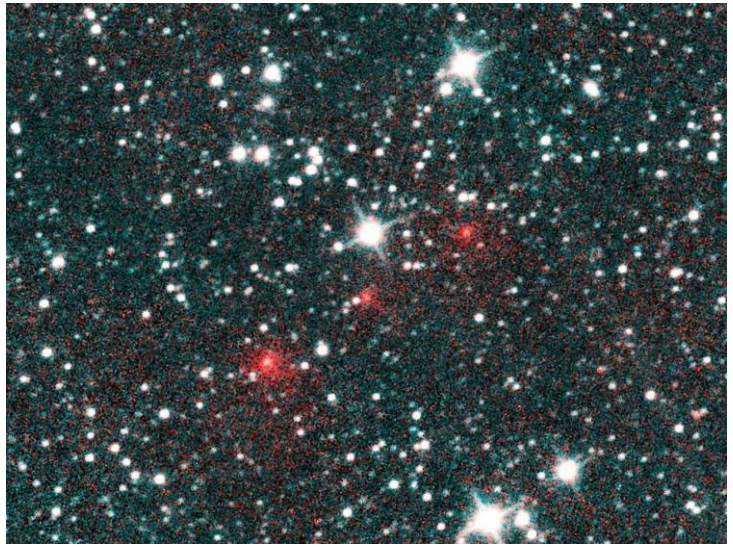
- It is often associated with the combination of big data, high performance computing, and machine learning

Data Science in the 21st century

Something new?

Turing award winner Jim Gray:

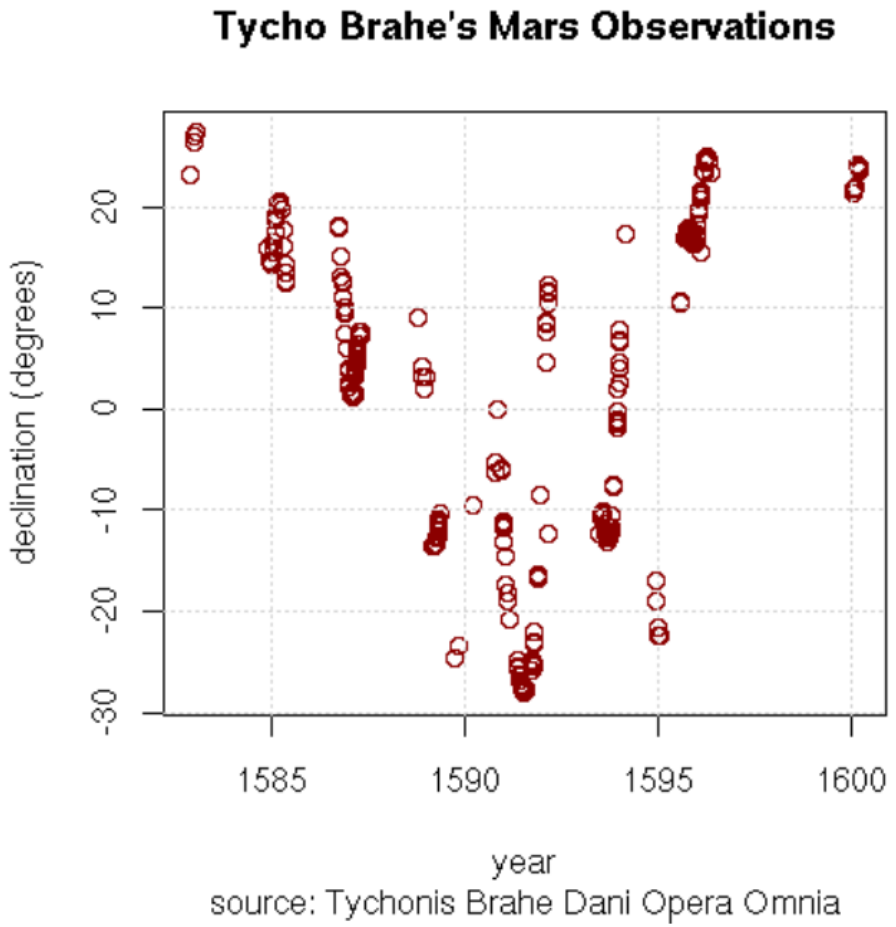
> Data science is a "fourth paradigm" of science (empirical, theoretical, computational and now data-driven) [source]



Data science for scientific discovery.
Discovery of Comet NEOWISE (series of red dots). NASA/JPL-Caltech

Something old?

Data Science in the 16th century



Just a new name?

- There is still no consensus on the definition of data science and it is considered by some to be a buzzword [source]

- David Donoho: Data science is not distinguished from statistics by the size of datasets or use of computing [source]

- Nate Silver: Just another name for statistics [source]

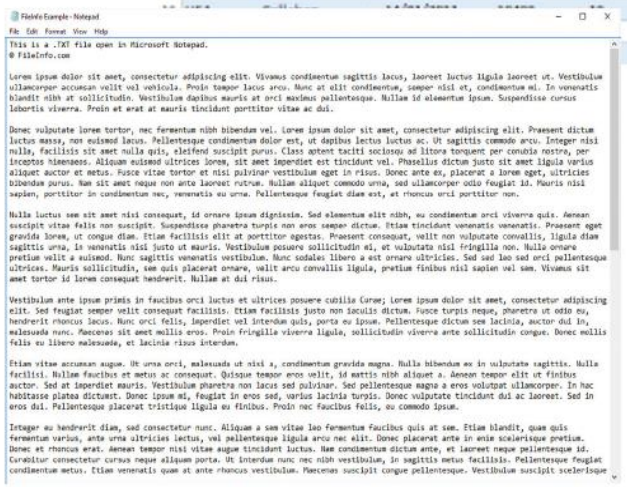
Working definition

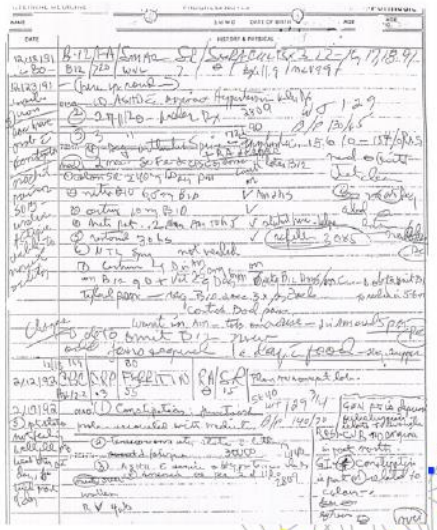
- Data Science studies the process of
  - collecting data
  - cleaning data
  - analyzing data [source]

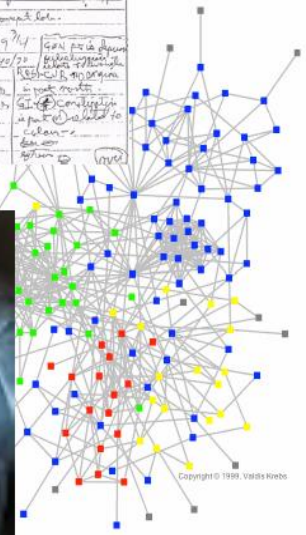
Collecting data

Data comes in many different forms

	A	B	C	D	E	F
1	Country	Salesperson	Order Date	OrderID	Units	Order Amount
2	USA	Fuller	1/01/2011	10392	13	1,440.00
3	UK	Gloucester	2/01/2011	10397	17	716.72
4	UK	Bromley	2/01/2011	10771	18	344.00
5	USA	Finchley	3/01/2011	10393	16	2,556.95
6	USA	Finchley	3/01/2011	10394	10	442.00
7	UK	Gillingham	3/01/2011	10395	9	2,122.92
8	USA	Finchley	6/01/2011	10396	7	1,903.80
9	USA	Callahan	8/01/2011	10399	17	1,765.60
10	USA	Fuller	8/01/2011	10404	7	1,591.25
11	USA	Fuller	9/01/2011	10398	11	2,505.60
12	USA	Coghill	9/01/2011	10403	18	855.01
13	USA	Finchley	10/01/2011	10401	7	3,868.60
14	USA	Callahan	10/01/2011	10402	11	2,713.50
15	UK	Rayleigh	13/01/2011	10406	15	1,830.78
						1,622.40
						319.20
						802.00

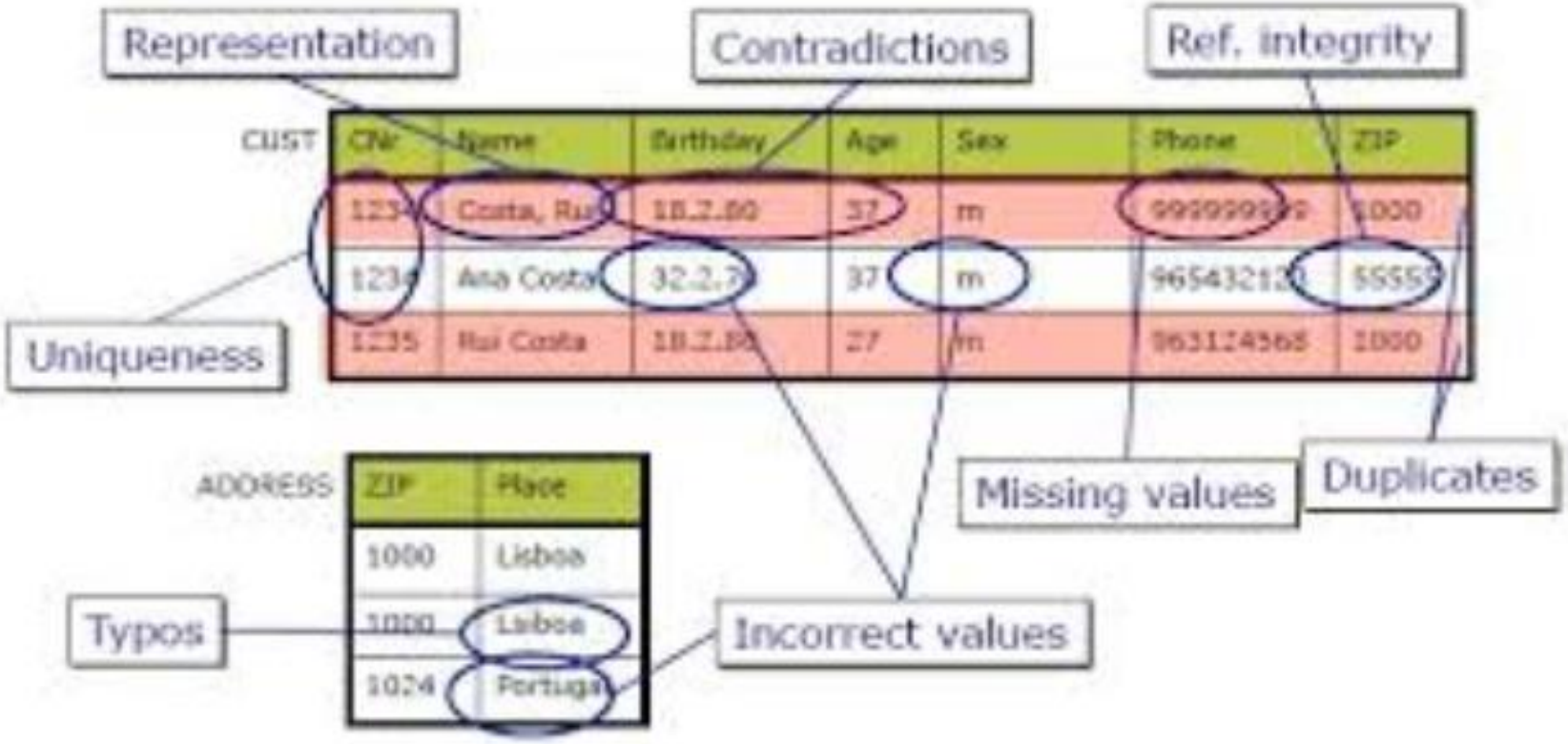






Cleaning data

Data commonly needs pre-processing before it can be used



Source

Analyzing data

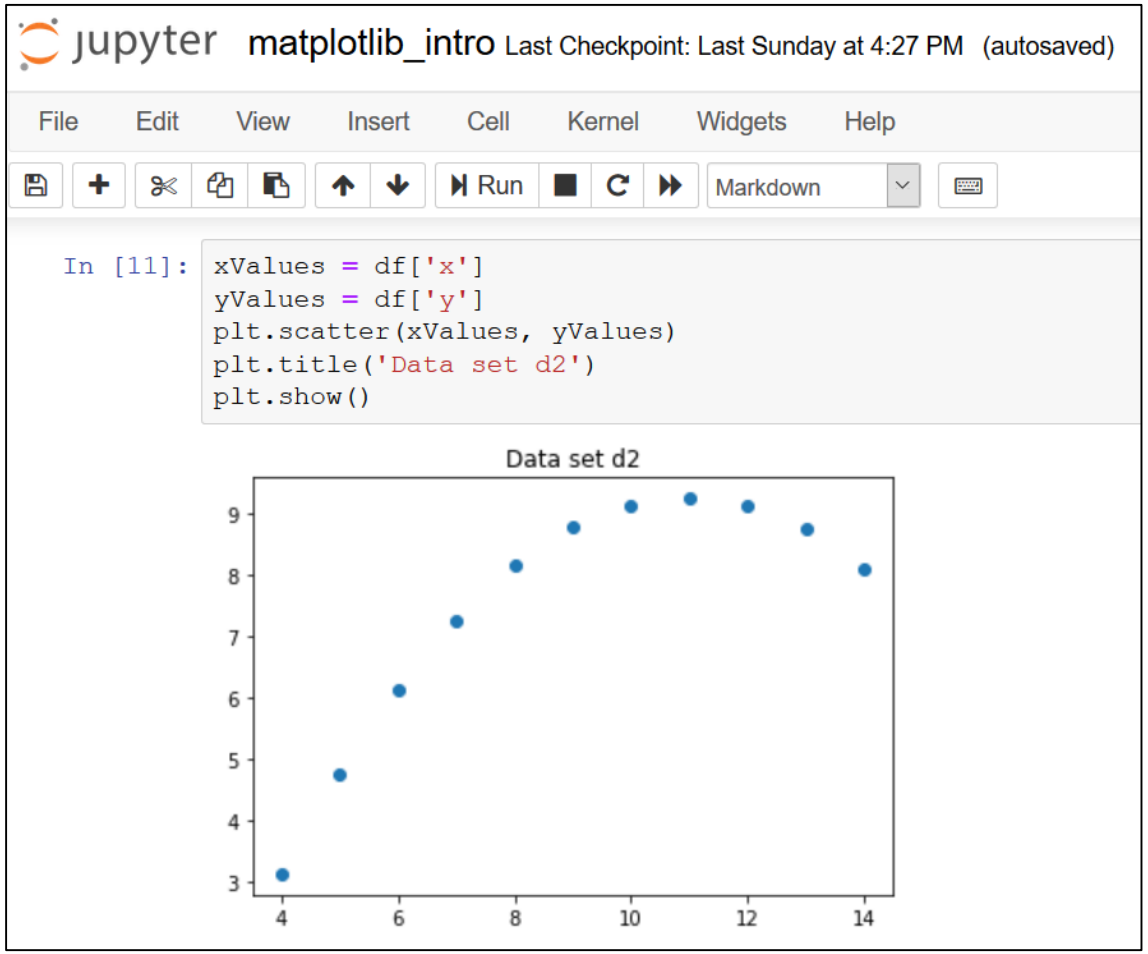
Analyze the data using statistics, machine learning, and visualization

Programming languages:
Python, R, Mathematica, Excel…

Notebook environments:
Mathematica or Jupyter

Many specialised software packages

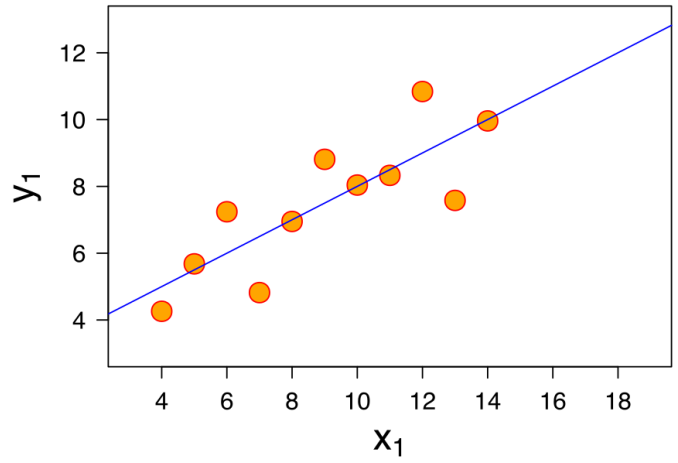
data formats:
CSV, XML, SQL, JSON,…

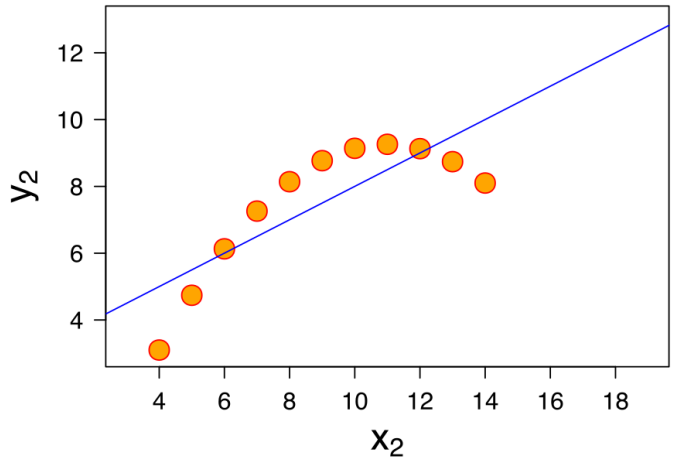


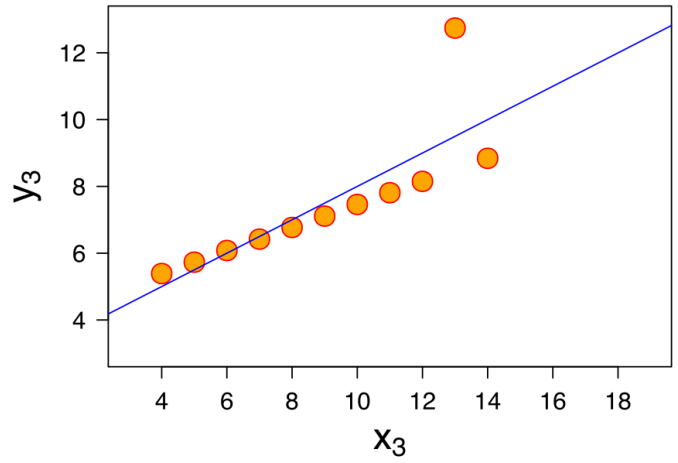
Break?

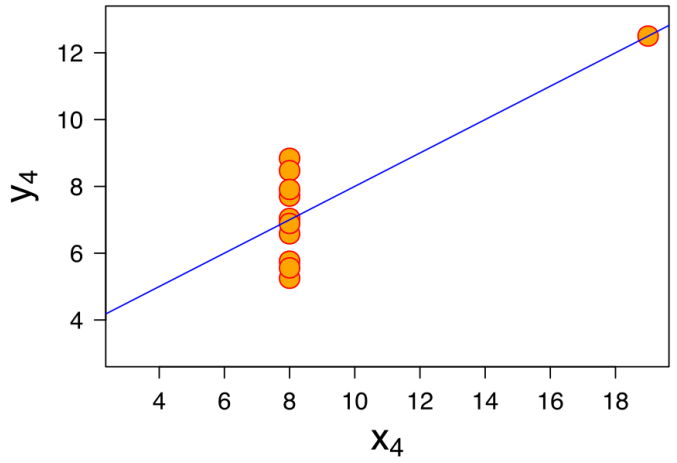
Visualizing data

Limits of mathematical models









Anscombe's quartet

Very different data sets have the same regression line

Very different realities have the same mathematical model

Ilustrates the need to look at the data!

Limits of statistics

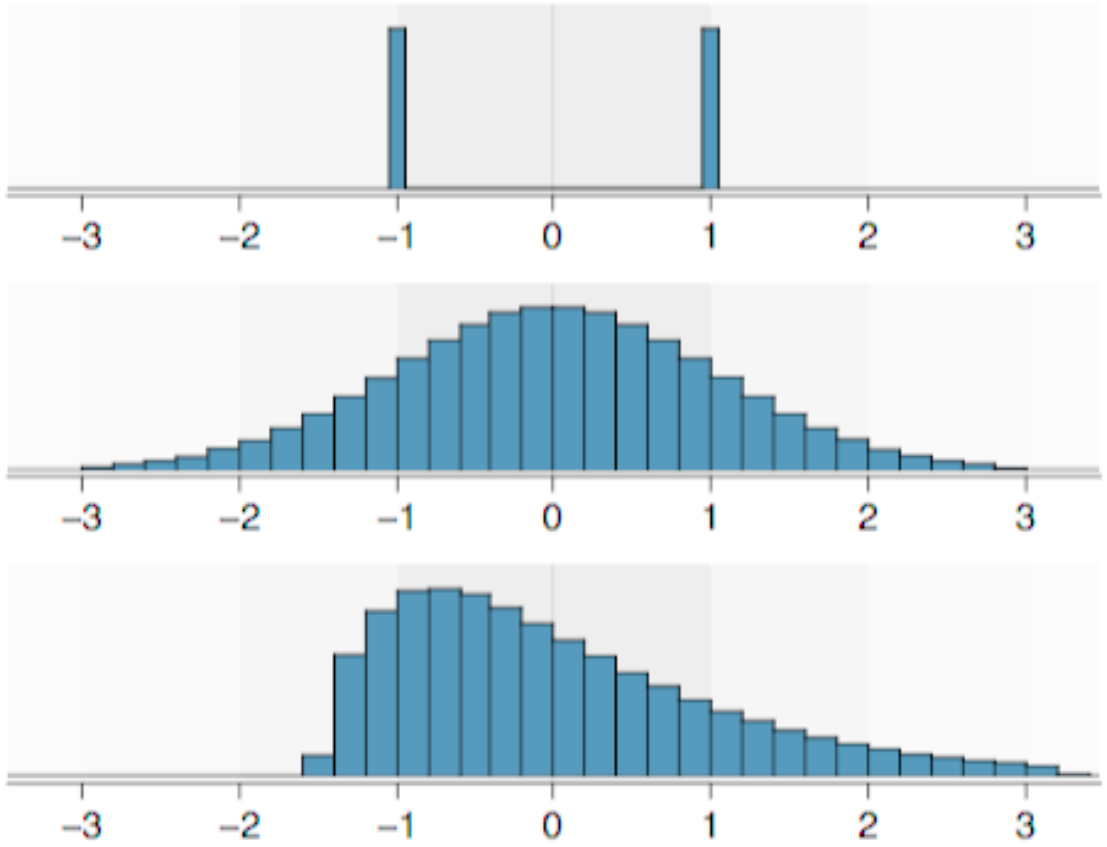


Figure 1.25: Three very different population distributions with the same mean $\mu = 0$ and standard deviation $\sigma = 1$.

Very different data sets with the same mean and variance

Ilustrates the need to look at the data!

Image: OpenIntro Statistics

Living in a multi-dimensional world

- Our perception is in thousands of dimensions
  - Tastes are 5-tuples (sweet, sour, bitter, salt, umami)
  - Smells are 300-tuples
  - Tactile stimuli target thousands of tactile receptors

- Fruits as 6D objects

Nutritional Information

Fruit	Serving Size	Calories	Carbs	Protein	Fiber	Fat	Sodium
Apples*	1 Medium Apple	80	22g	0g	5g	0g	0mg
Peaches	1 Medium Peach	40	10g	.06g	1.5g	0g	0mg
Nectarines	1 Medium Nectarine	70	16g	1g	3g	1g	0mg
Plums	1 Medium Plum	36	8.6g	0.52g	1.0g	0.41g	0mg
Asian Pears	1 Medium Pear	59	13g	0.9g	4g	0.1g	0mg
Strawberries	8 Medium Berries	70	17g	1g	3g	0.5g	0mg
Raspberries	10 Raspberries	10	2.3g	0.2g	1.2g	0.1g	0.2mg
Blueberries	1 Cup Blueberries	83	21.0g	1.1g	3.5g	0.5g	1mg
Pumpkins**	1 Cup	49	12g	2g	3g	0g	0mg

*NOTE: Slight variation depending on variety; figures reflect an overall average for the fruit.
**NOTE: Figures are based on pumpkin being cooked, boiled, drained, without salt.

Representing multi-dimensional objects

- A spreadsheet table with one row per object and one column per feature can be used to represent N-dimensional objects
- A 3D object can be represented as a 2D object (perspective drawing)
- A piece of paper can be regarded as 2D, but we need 3D to represent a pencil drawing on it: (x,y,color)
- On the computer, the drawing can be represented as a 100x100 matrix. Or as a 10000D vector. Or as a set of 10000 triplets of the form (x,y,color).

Visualizing multi-dimensional objects

- Iris data set

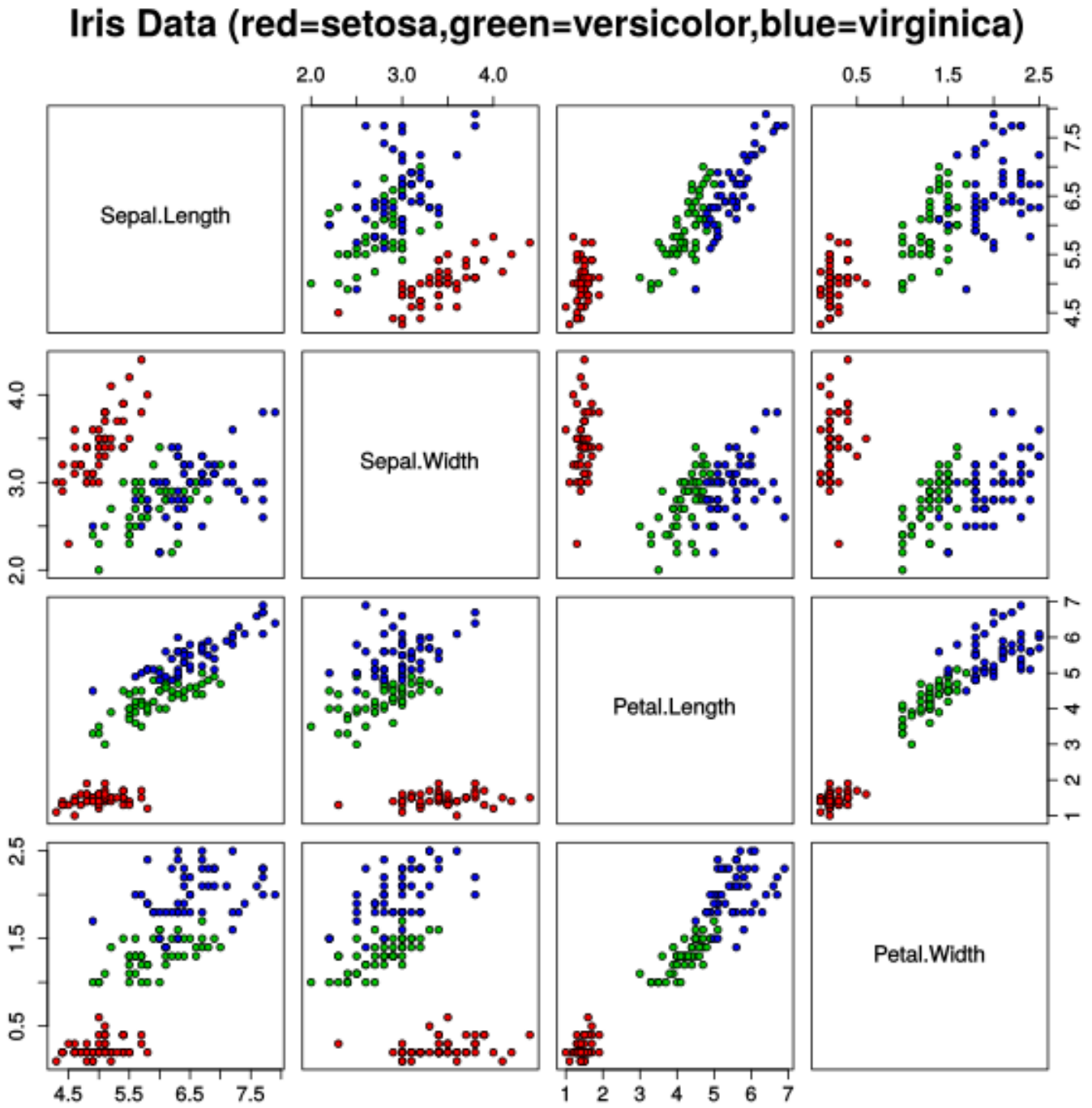
- Each datapoint has 5 dimensions:
  - Petal length
  - Petal width
  - Sepal length
  - Sepal width
  - Species

Iris setosa

Iris versicolor

Iris virginica

*R. A. Fisher (1936). "The use of multiple measurements in taxonomic problems". Annals of Eugenics. **7** (2): 179–188.*



Everything times everything

https://commons.wikimedia.org/wiki/File:Iris_dataset_scatterplot.svg (User:Nicoguaro)

Scatter plot

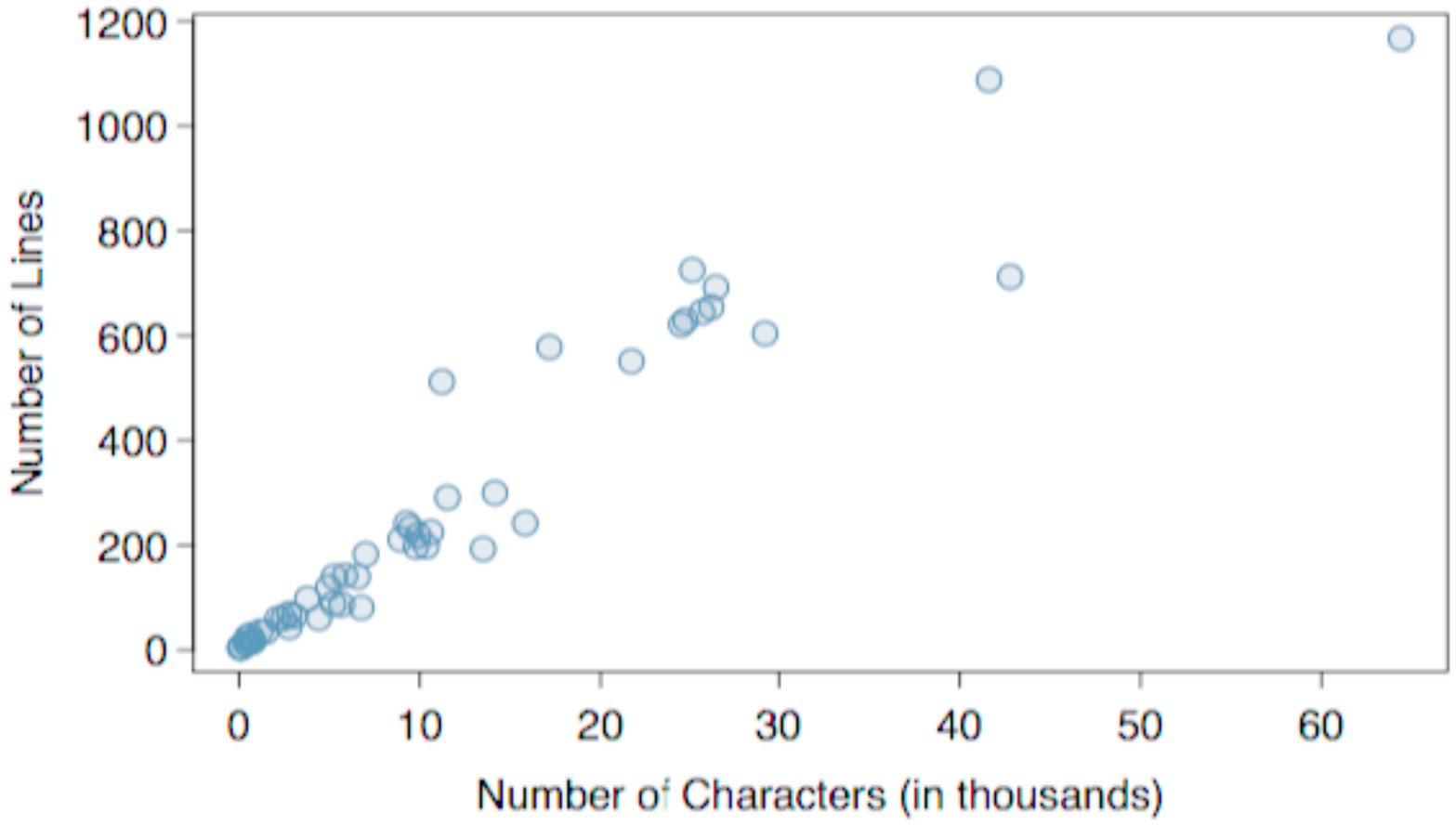
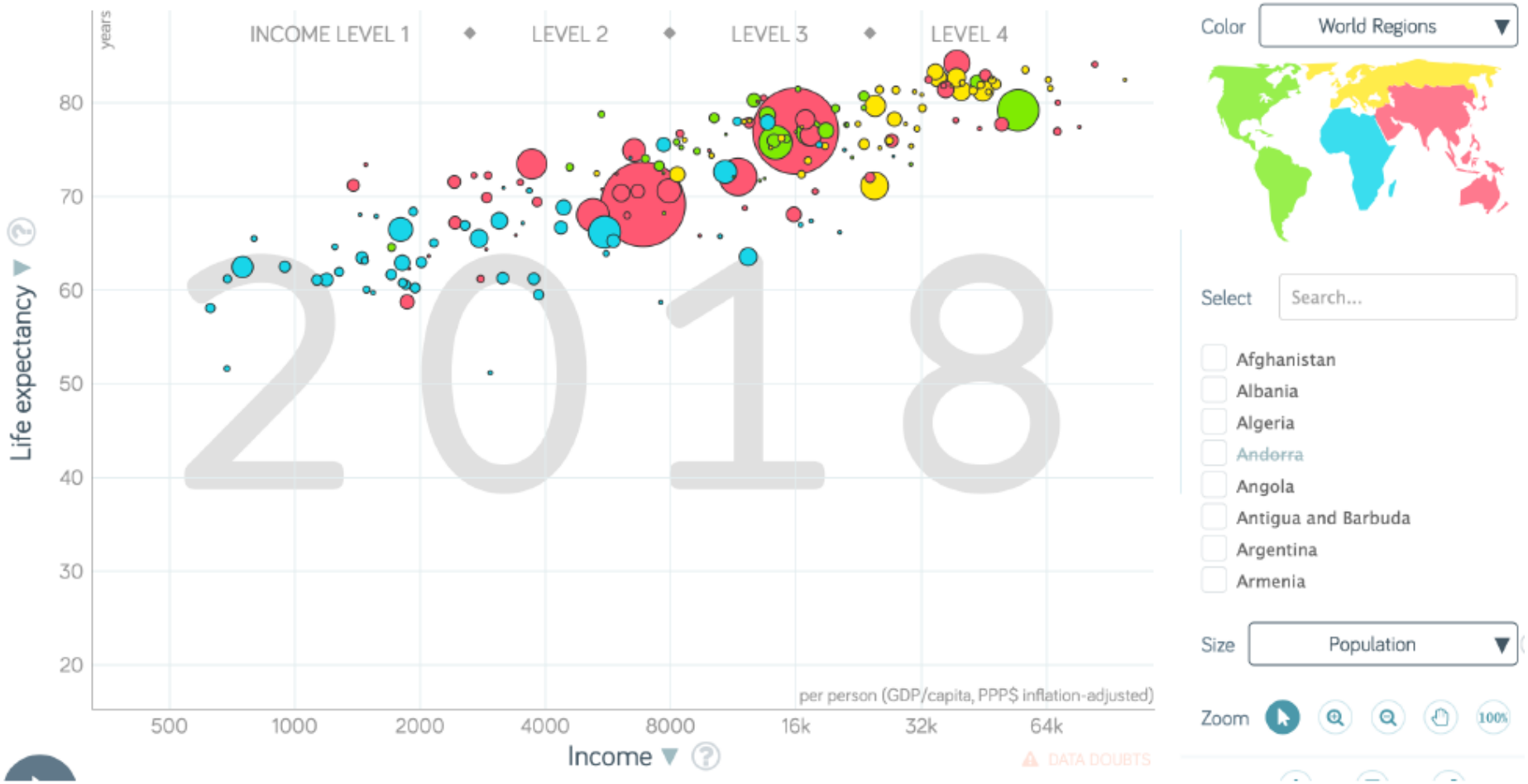


Figure 1.17: A scatterplot of `line_breaks` versus `num_char` for the `email50` data.

Image: OpenIntro Statistics

Scatter plot



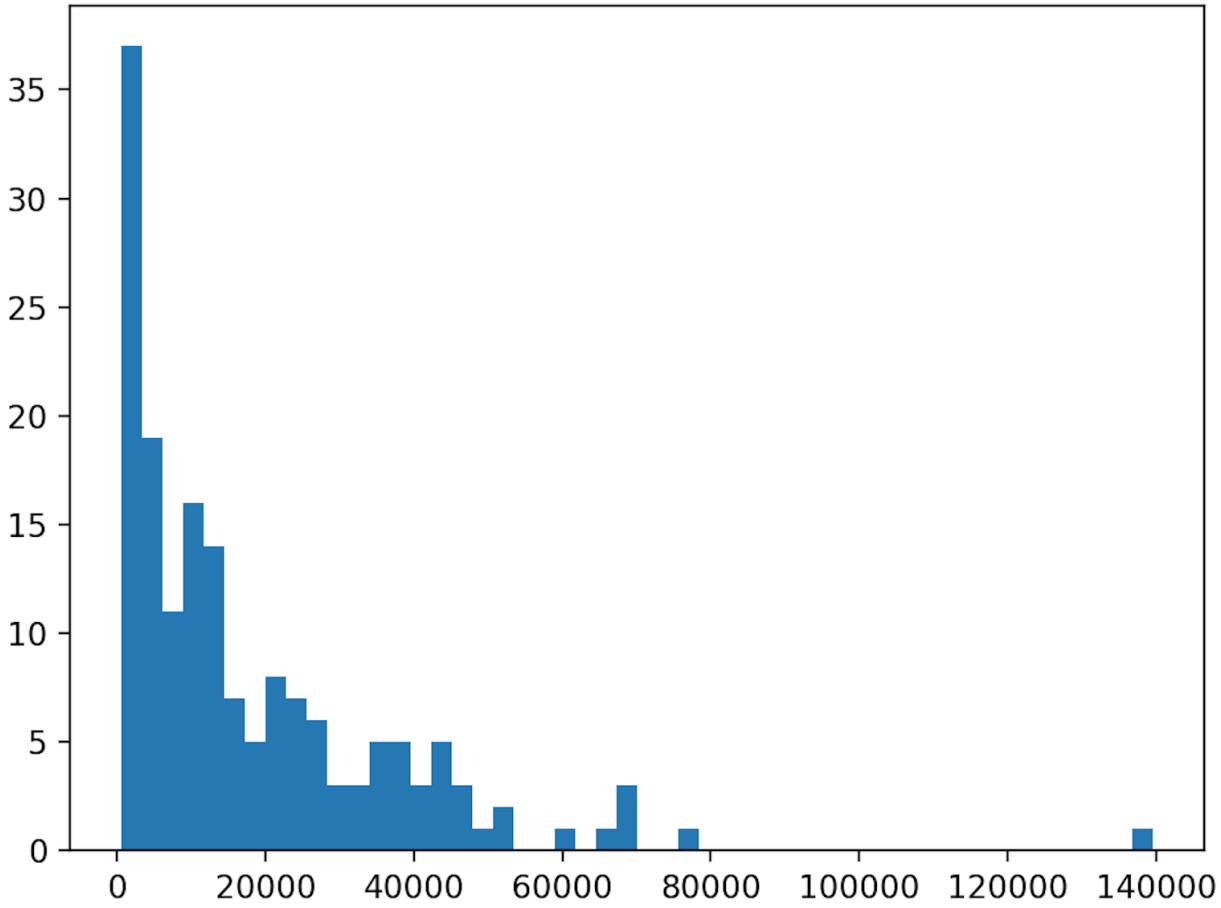
Here colors add a dimension

Point size too!

And movies showing development over time!

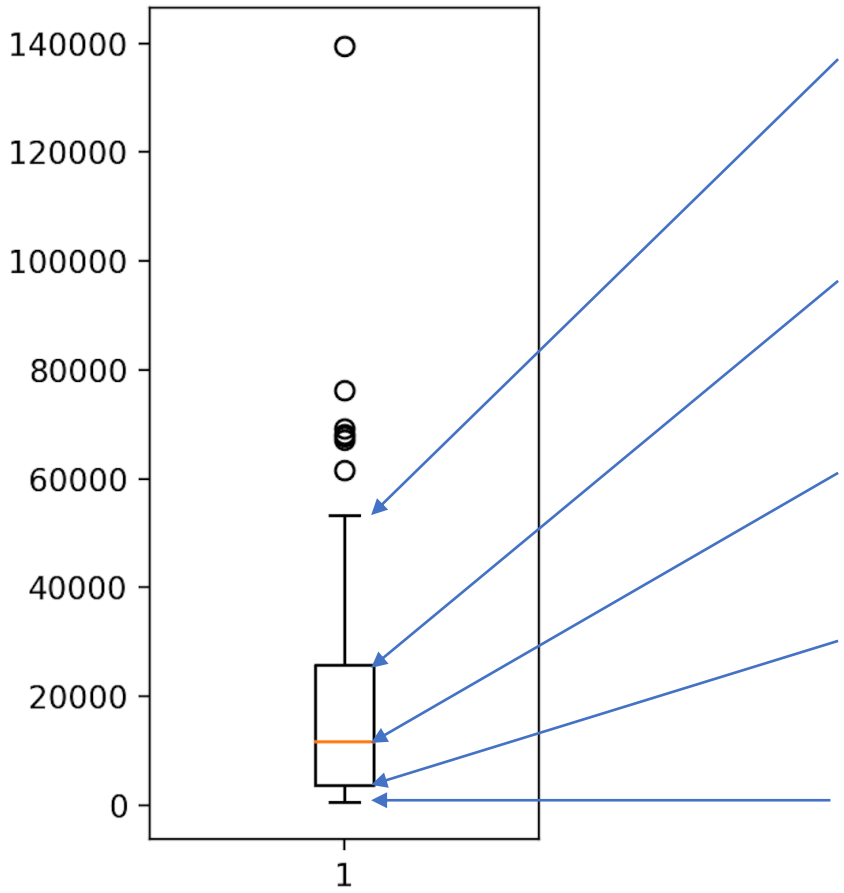
Figure from [Gapminder]

Histogram


35
30
25
20
15
10
5
0
0
20000
40000
60000
80000
100000
120000
140000

Boxplot

What aspects of the dataset are not visible here? How does it compare to the histogram?



Upper whisker, e.g. largest data value lower than Q3 + 1.5 * (interquartile range)

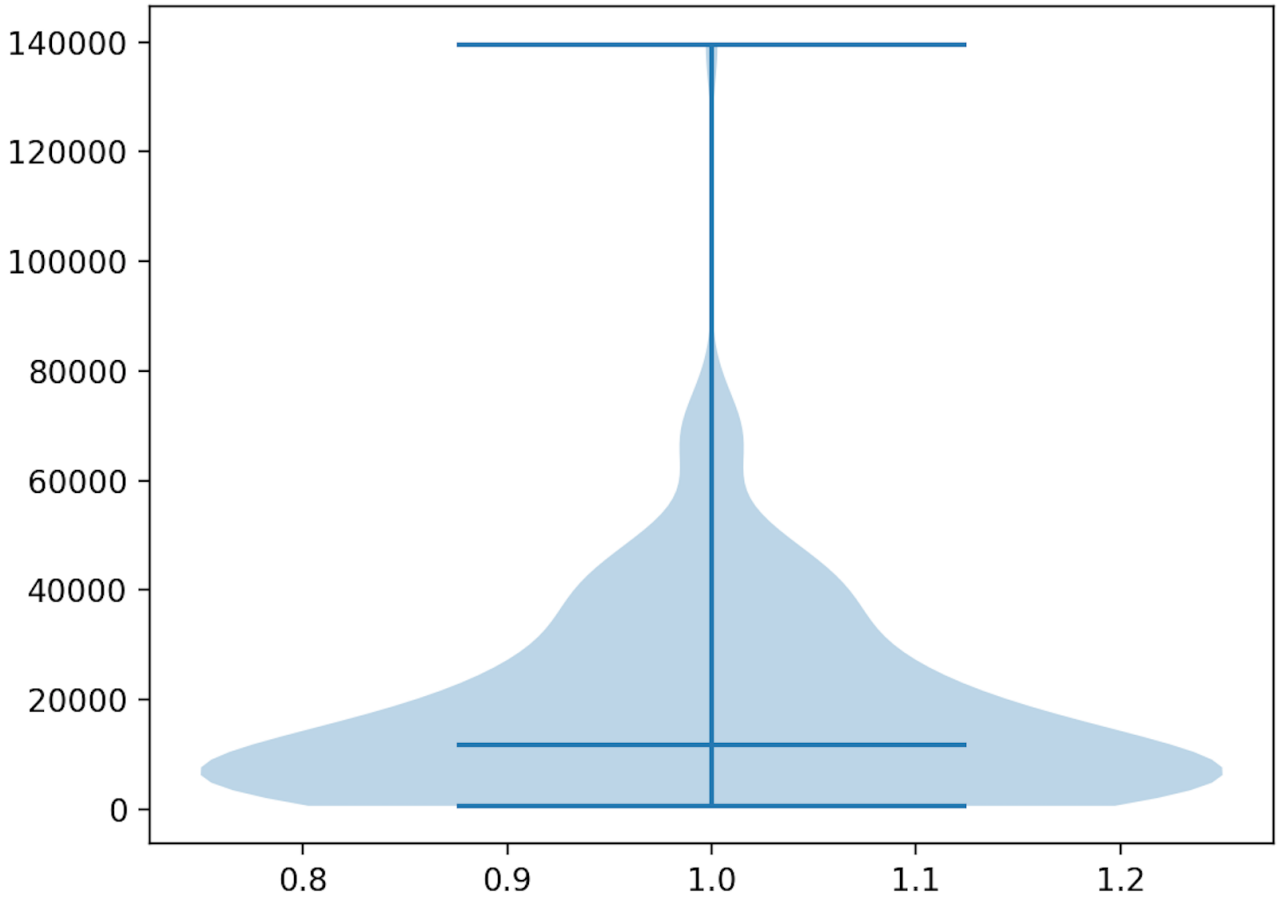
Q3 (3rd quartile)

median

Q1 (1st quartile)

Lower whisker

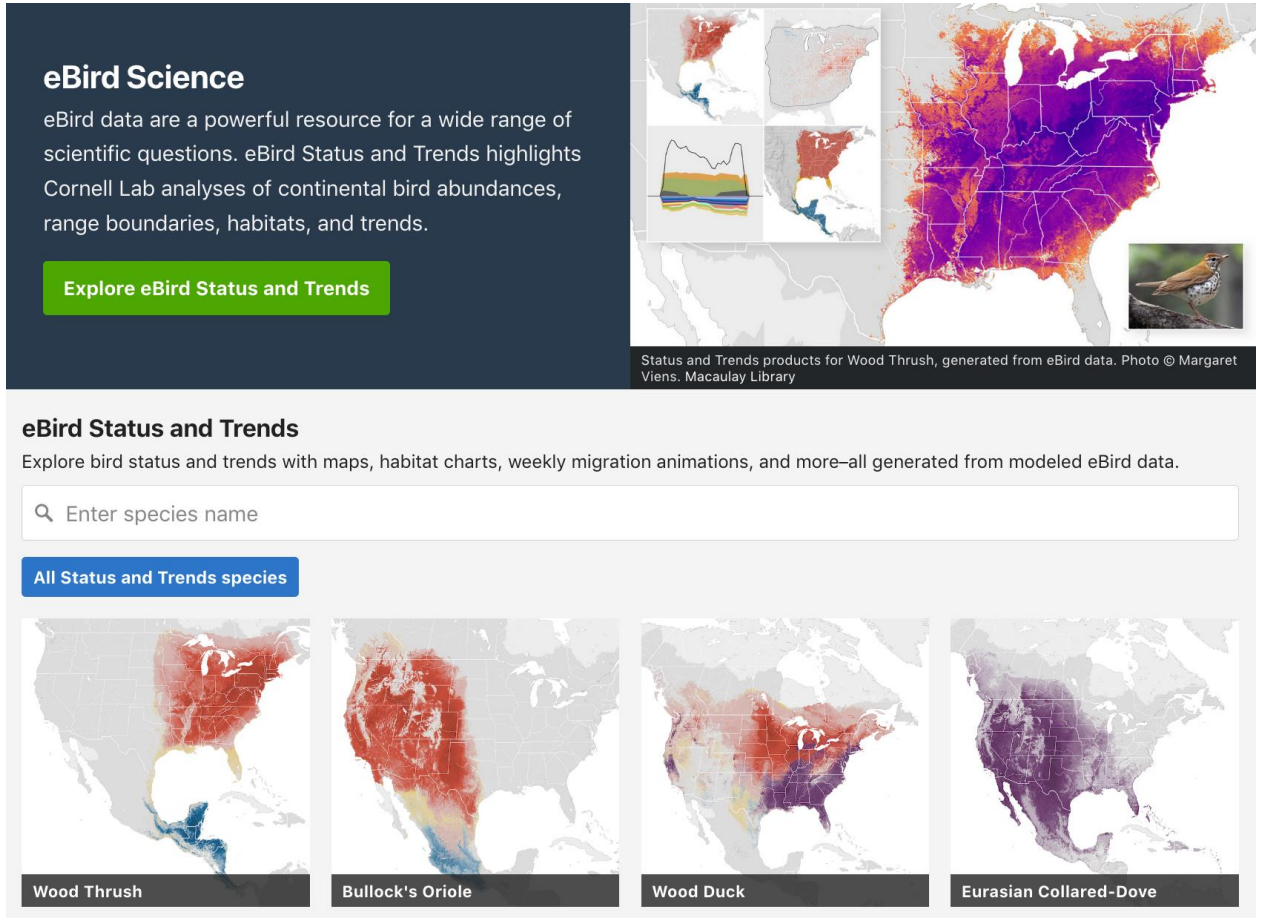
Violin plot



Ordinary map

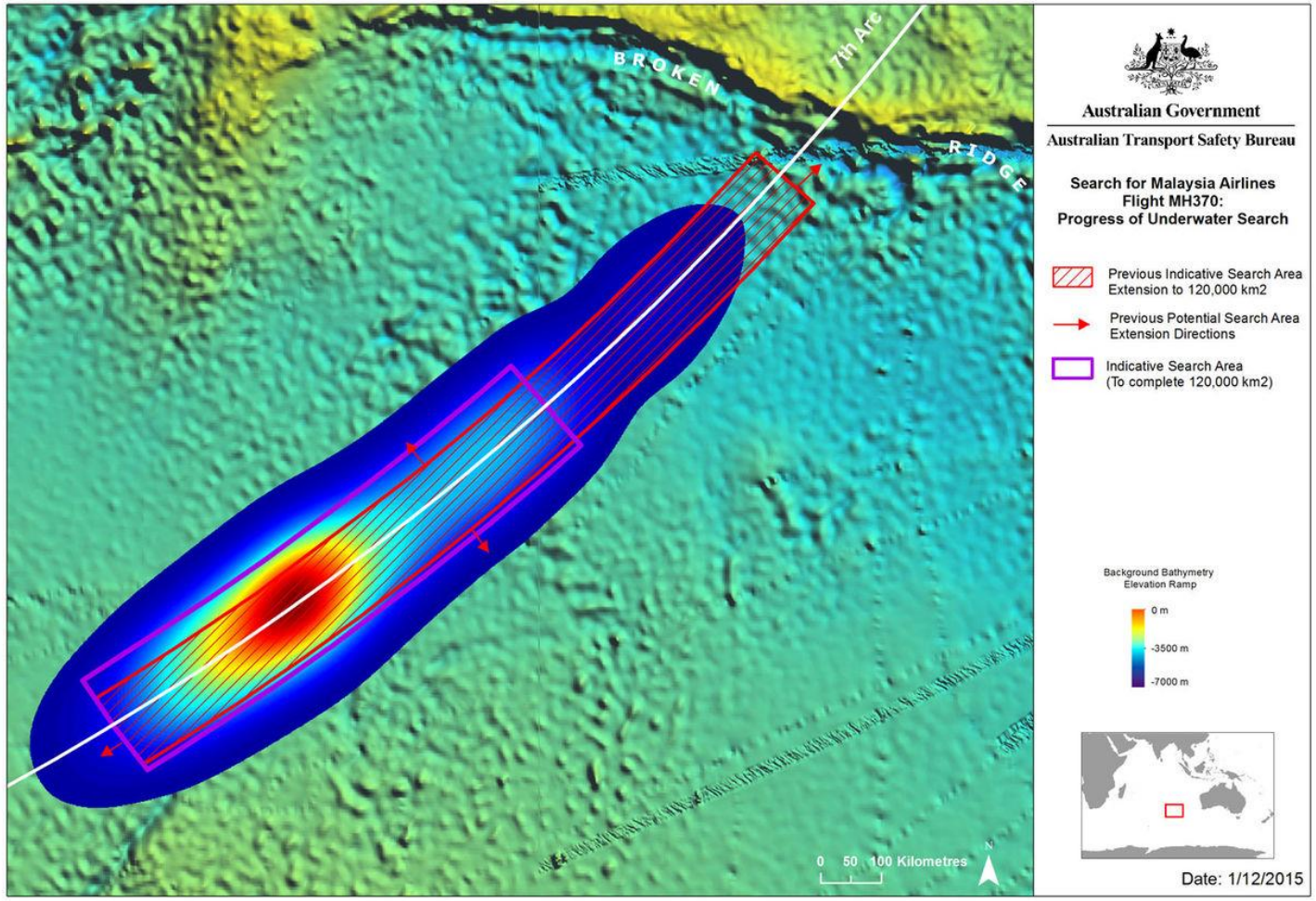
Where are the birds?

- eBird: Quantified Bird Watching
- Bird watchers as “sensors”
- Citizen Science



Source

Heat map


BROKEN
RIDGE
7th Arc
0 50 100 Kilometres
Australian Government
Australian Transport Safety Bureau
Search for Malaysia Airlines Flight MH370: Progress of Underwater Search
Previous Indicative Search Area Extension to 120,000 km2
Previous Potential Search Area Extension Directions
Indicative Search Area (To complete 120,000 km2)
Background Bathymetry Elevation Ramp
0 m
-3500 m
-7000 m
Date: 1/12/2015

Develop a visualisation aesthetic

- Maximize data-ink ratio
- Minimize the lie factor
- Minimize chart junk
- Use proper scales and clear labeling
- Make effective use of colour

Jupyter Notebook

Install Anaconda (already done)

- Please install Anaconda (not just Miniconda)
- That will give you Python 3.8 and Jupyter Notebook
- You will also get several packages:
  - Pandas (data science)
  - NumPy (math)
  - Matplotlib (plots)
- Also please install Pytorch and Tensorflow (for neural networks)

We will be using Anaconda: a platform for data science

- Free and open source distribution of Python and R
- Over 1500 packages
- Anaconda Navigator includes:
  - Jupyter Notebook
  - Spyder – an integrated development environment (IDE) for Python

https://www.anaconda.com/

Open a Jupyter notebook

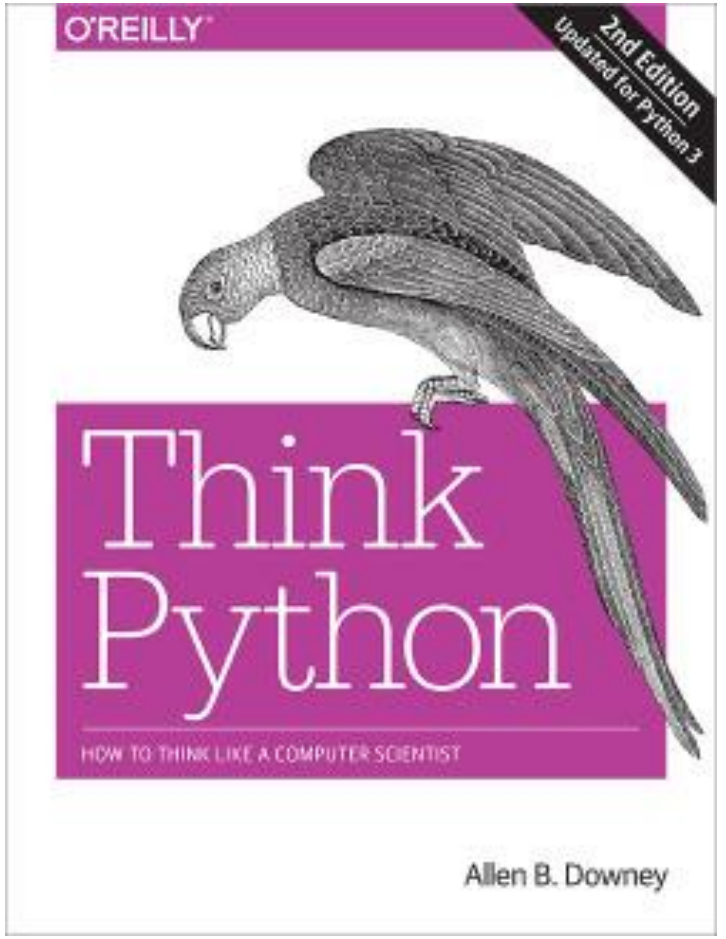
- Make a new directory and save the course notebooks from Canvas there.
- Open the program Jupyter Notebook. Then you get a “File explorer” tab in your web browser.
- Open the notebook jupyter_intro from this “File explorer”.
- Now let’s take a look at it together!

Python programming

Python programming

- Good introduction. No previous programming experience needed.

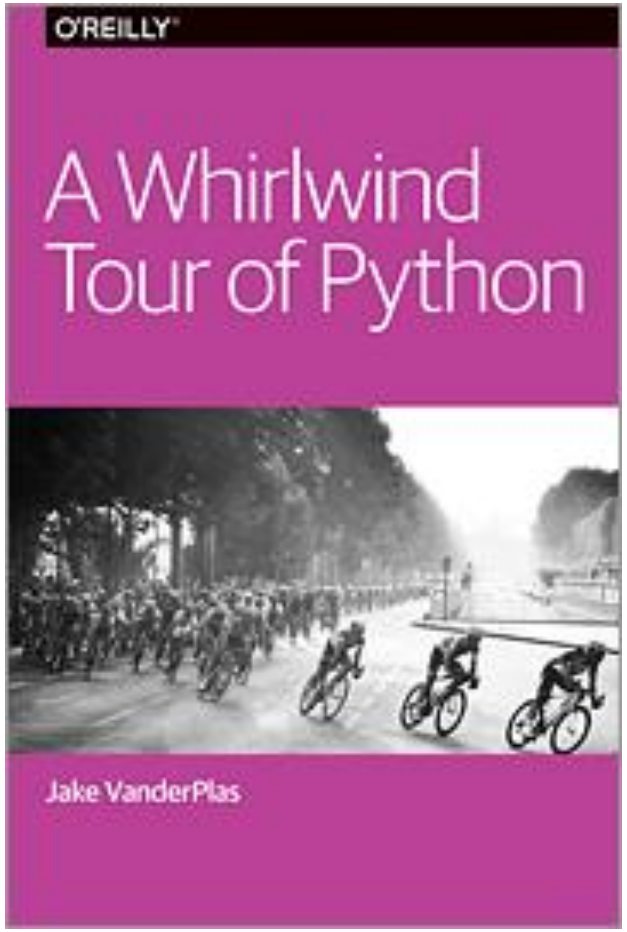
- Free online version



Python programming

- Faster pace than “Think Python”

- Free online version



Quick introduction

- If you are new to Python, you may want to take a look at the notebook
  - python_intro (@Modules)

Python packages

Python packages

Lots! including:

- Pandas
- NumPy
- SciPy
- Matplotlib
- Seaborn
- Scikit-learn

To use the functions in a module or a package, these have to be imported, e.g.

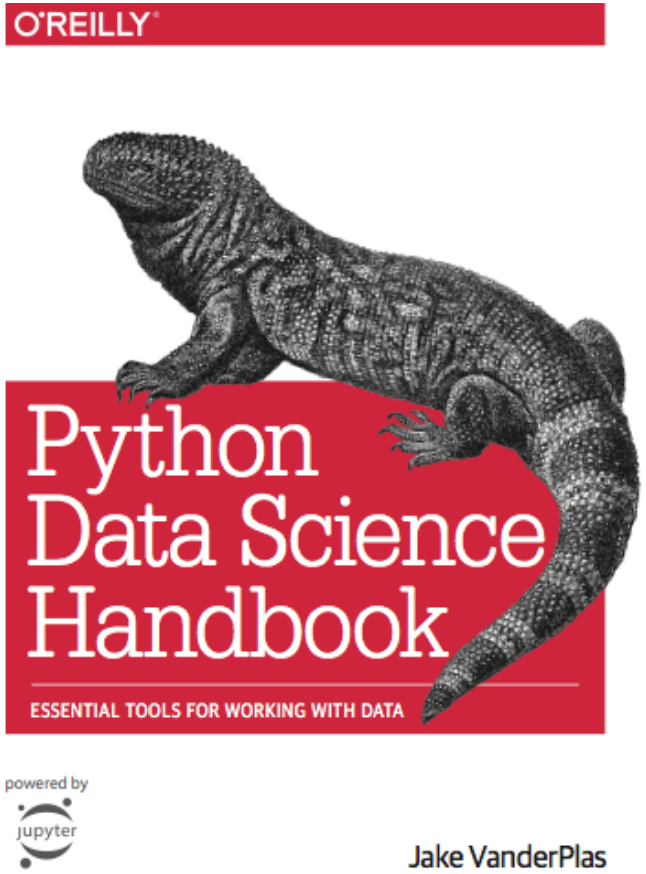
```
import pandas

import numpy as np

from sklearn.linear_model import LinearRegression
```

Python packages

- Assumes some knowledge of Python
- Focuses on using packages like NumPy, Pandas, Matplotlib, Scikit-learn
- Free online version



Quick introductions

- Finally, let’s have a look at some notebooks (@Modules):
  - numpy_intro
  - pandas_intro
  - matplotlib_intro