

# Lecture 2

Introduction to Data Science



# Questions

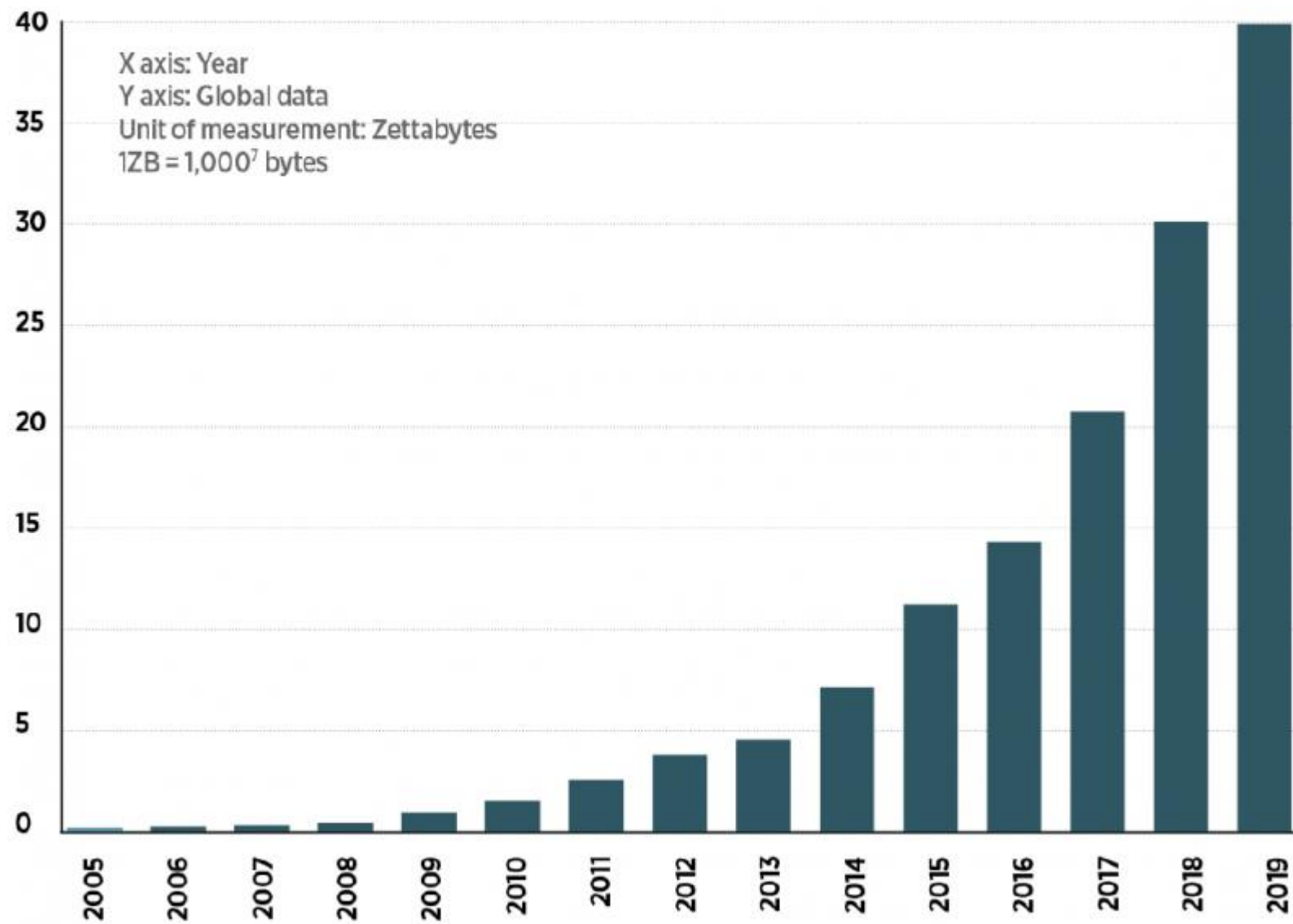
- Good to record the lectures and put them on Canvas?
- Have you programmed in Python before?
- Do you have a lab partner?
  - If not, you may want to stay online after the lecture and give a short presentation of yourself?

# 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

# Data sources



# Data sources



# Data sources



U.S. INTERNATIONAL CANADA ESPAÑOL 中文

PLAY THE CROSSWORD Account

Thursday, January 14, 2021  
Today's Paper

World U.S. Politics N.Y. Business Opinion Tech Science Health Sports Arts Books Style Food Travel Magazine T Magazine Real Estate Video

**The New York Times**

Listen to 'The Daily' Impeached, again.

Listen to 'Sway' Anna Wintour on the Kamala Harris Vogue cover.

The Book Review Podcast Charles Yu on his National Book Award-winning novel, "Interior Chinatown."

## Trump, Impeached Twice, Now Faces Another Senate Trial

**A Conviction Could Mean He Would Never Hold Office Again**

The House of Representatives impeached President Trump for inciting a violent insurrection against the Capitol, just one week before he was to leave office.

A small but significant number of Republicans joined Democrats to charge him with high crimes and misdemeanors for an unprecedented second time.

Senator Mitch McConnell will not bring the Senate back before Jan. 19, meaning a trial is unlikely until around President-elect Joe Biden's inauguration.



**Jim Jordan** Republican of Ohio

"It's always been about getting the president, no matter what. It's an obsession, an obsession that has now broadened."

See how every representative voted.

**News Analysis: A Preordained Coda to a Presidency**

The impeachment of President Trump for a second time seemed like the almost

BBC Sign in Home News Sport Reel Worklife Travel

**NEWS**

Home Coronavirus Video World UK Business Tech Science Stories Entertainment & Arts

## Trump faces Senate trial after historic charge

The US president could be barred from holding public office again after his second impeachment.

1h US & Canada



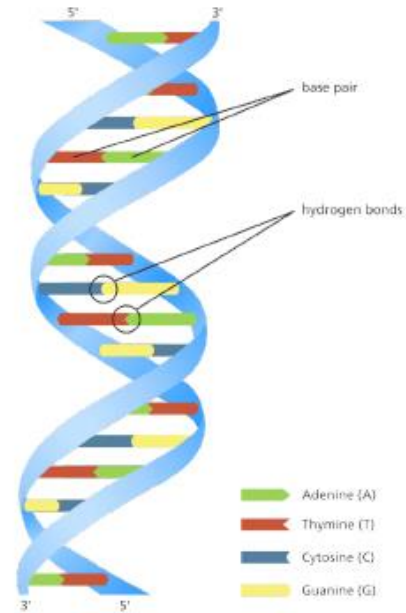
► Moment Trump impeached - again • Voters react: Symbolic but necessary • What this means for America



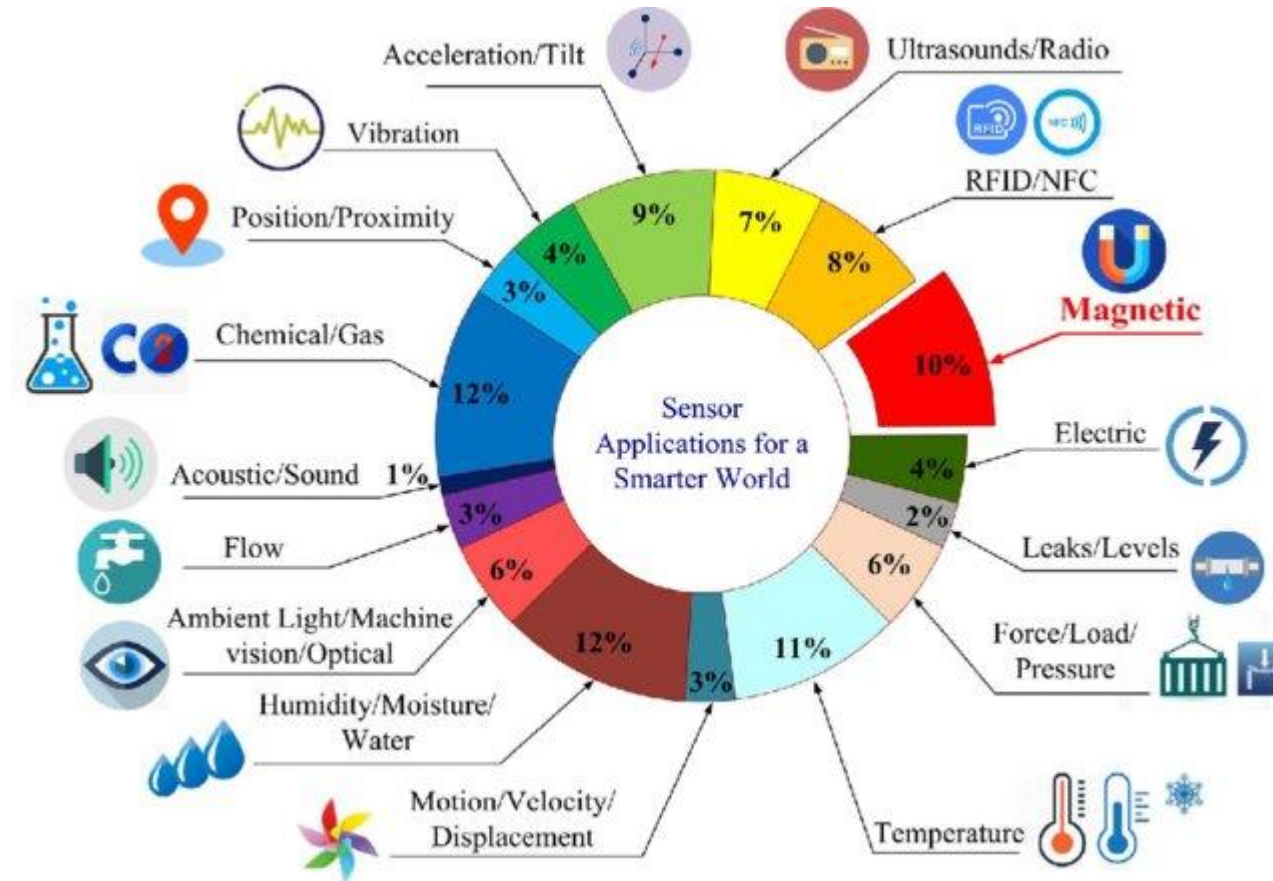
<https://www.bbc.com/news/health>



# Data sources



# Data sources



Common sensor categories. Image: Liu Xuyang

Machine learning

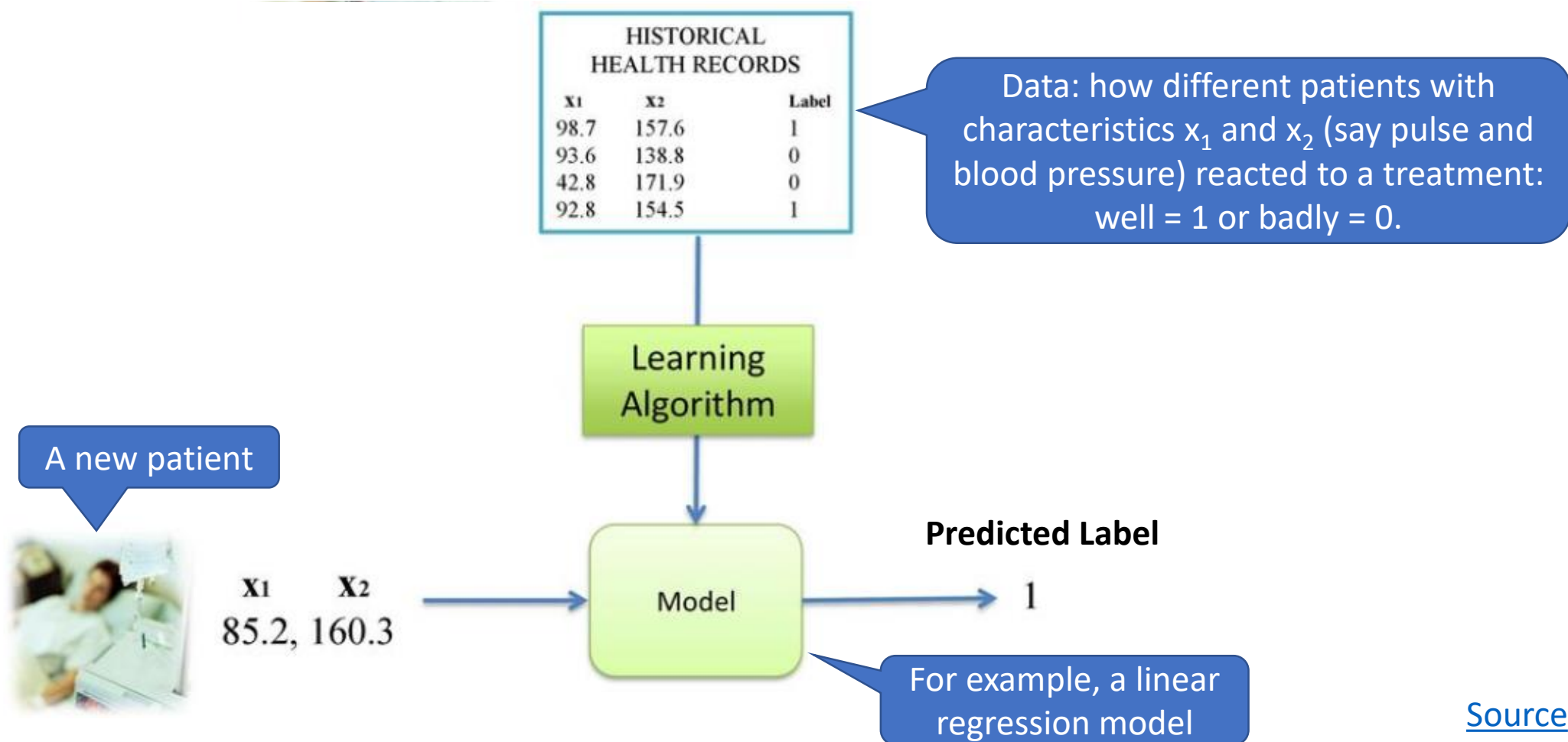
# What is Machine Learning?

- Tom Mitchell: A computer program *learns* from experience  $E$  with respect to some task  $T$  and some performance measure  $P$ , if its performance on  $T$ , as measured by  $P$ , improves with experience  $E$ .
- ML is the study of computer programs that learn from experience.

In other words: ML studies  
algorithms that learn from data

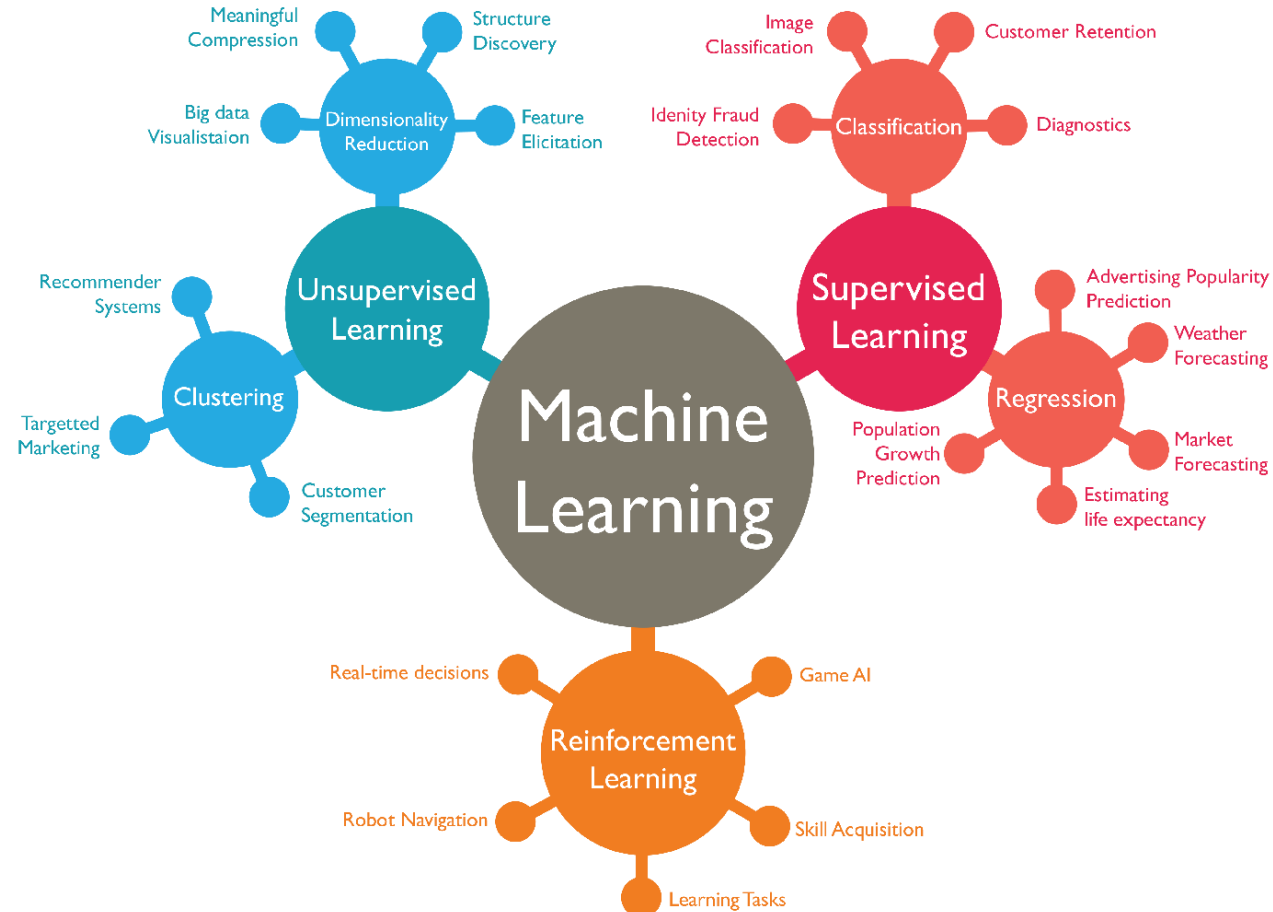
Here is an example

# Learning from data

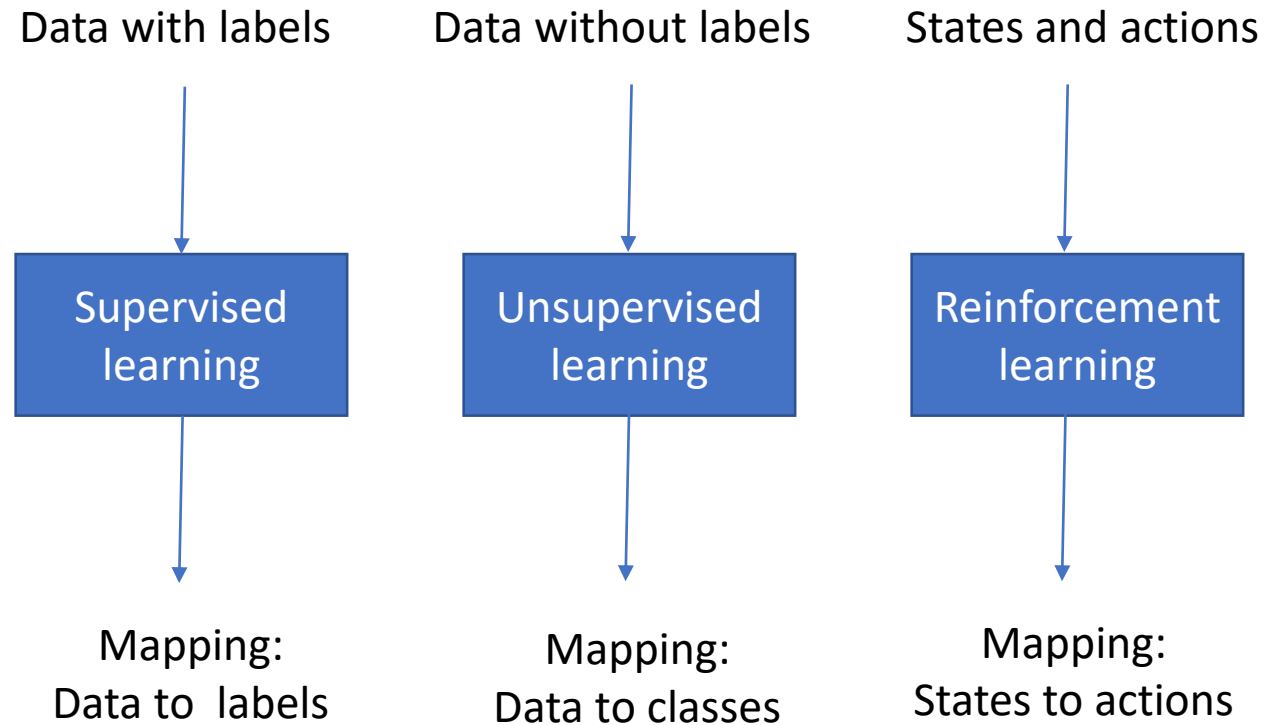


[Source](#)

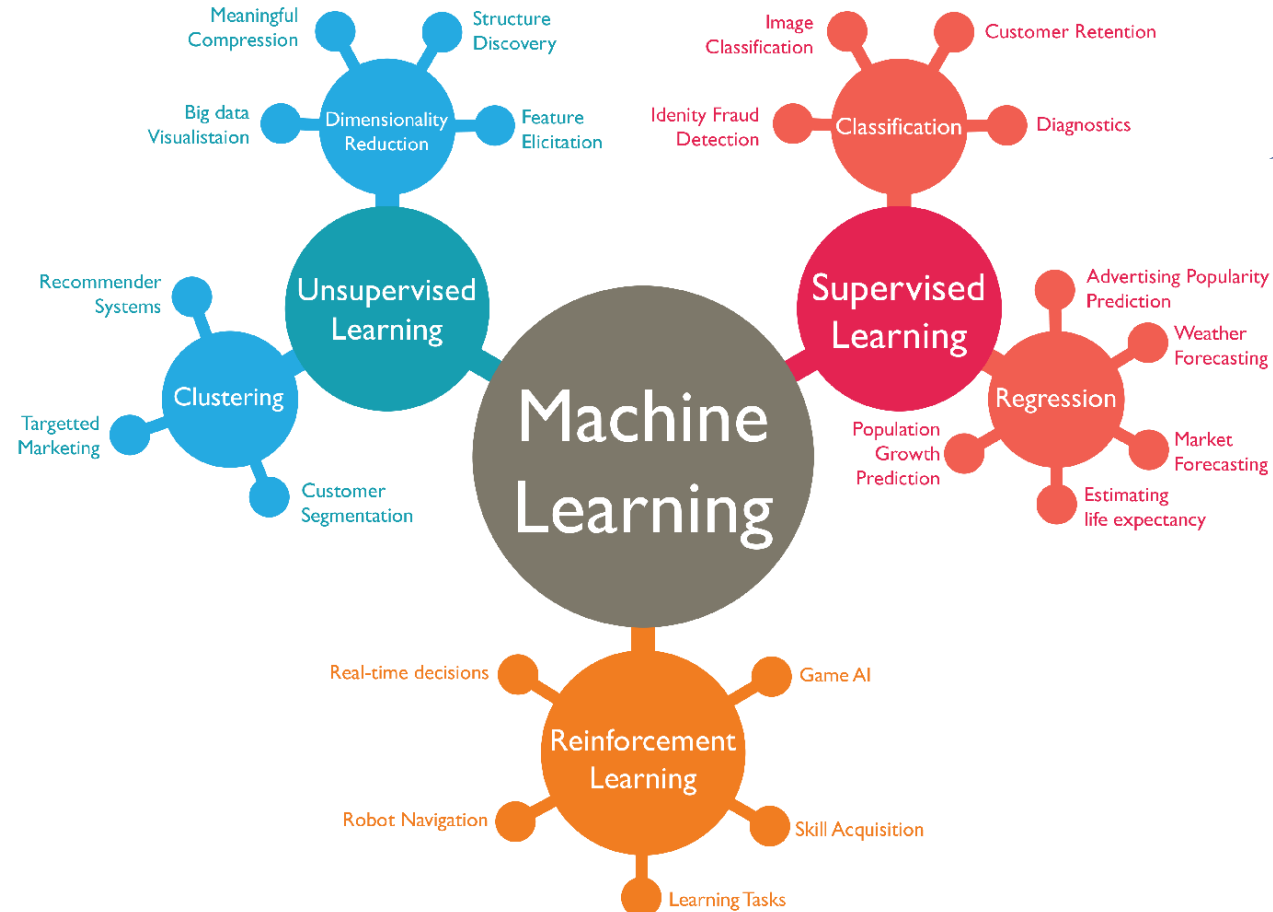
# Types of machine learning



# Main types of ML



# Types of machine learning

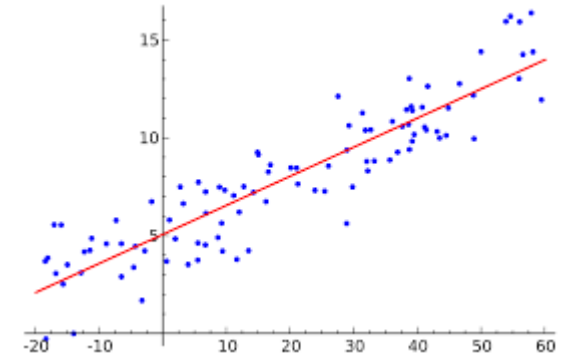


Let's zoom in a bit



# Regression

Map data points to numbers

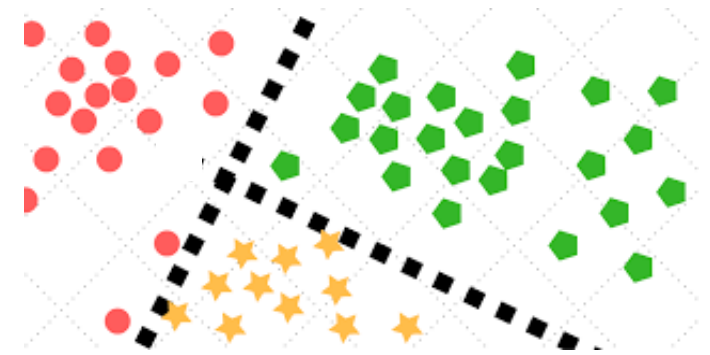


- What is the market price of that house?
- How much will it snow tomorrow?
- How many people will retweet that tweet?
- What will the price of this stock be in one hour?
- What is the temperature in that room?
- How much will we harvest next year?
- How much will we sell next month?

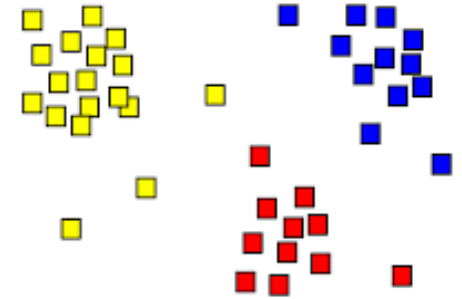
# Classification

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?



# 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?

# 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



# Applications of Machine Learning

# Identify objects

FLOWER



STRELITZIA

BIRD



EASTERN MEADOWLARK

TREE

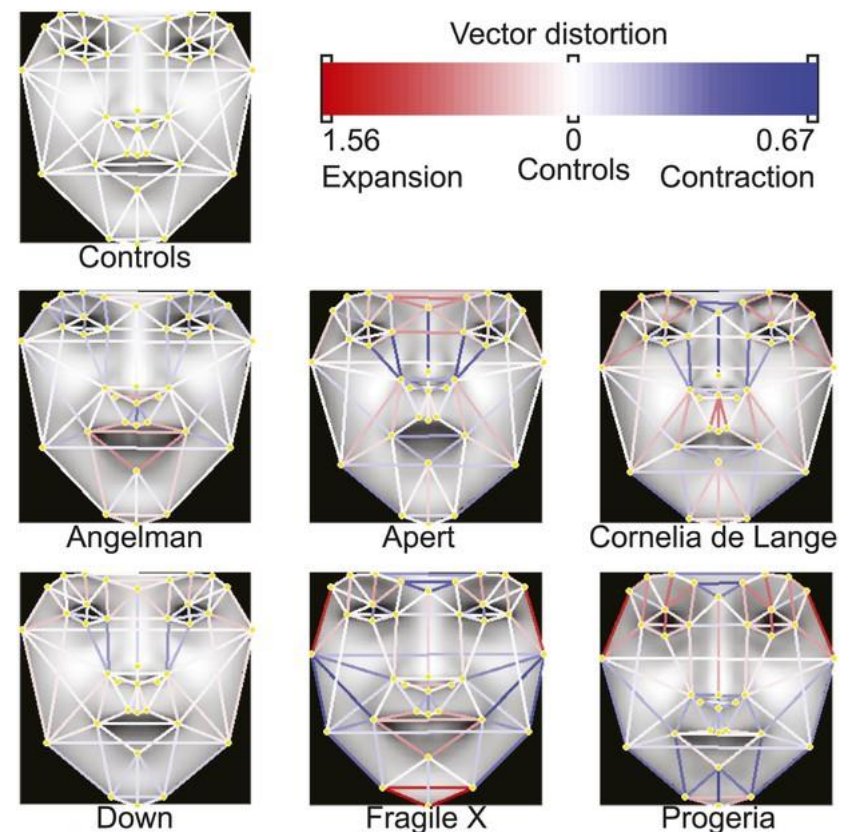


SALICACEAE

[Source](#)

# Diagnose diseases

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





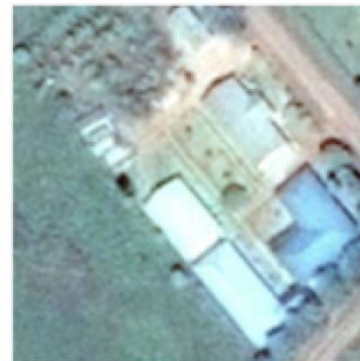
# Identify poverty

Identifying poverty by  
analyzing satellite images  
and classifying roofs

Photo



Satellite image



[Source](#)



# Literary analysis

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

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

# Translate text

Svenska ▼

↔

Engelska ▼

×

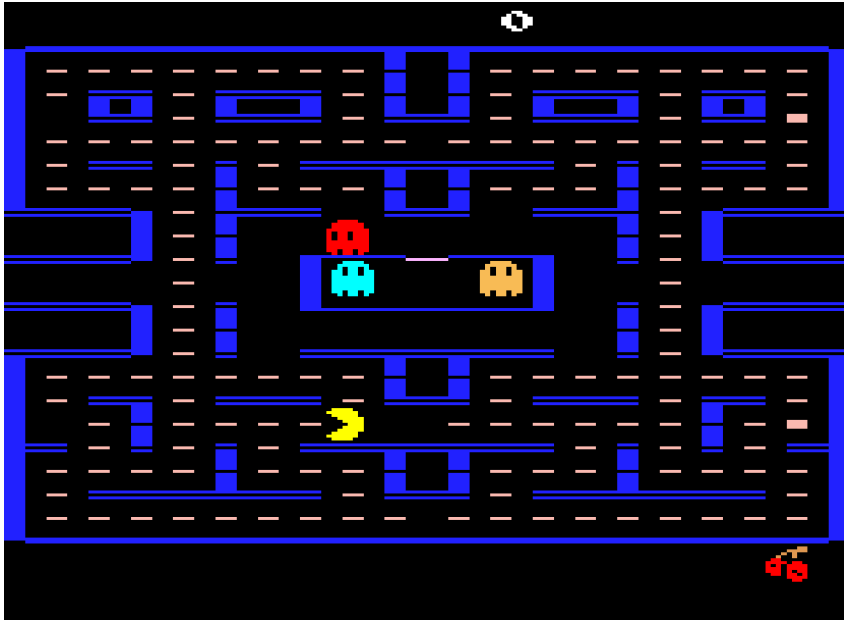
Men där var ingen vindsdörr. Där var endast en vanlig trappa, likadan som de andra.

Jag hade alltså räknat fel; jag hade ännu en trappa kvar.

But there was no attic door. There was only an ordinary staircase, like the others.

So I had calculated incorrectly; I still had one flight of stairs left.

# Play games



Pac-Man



StarCraft 2

# Drive a car



[Source](#)

Break?

Data science

# Data science

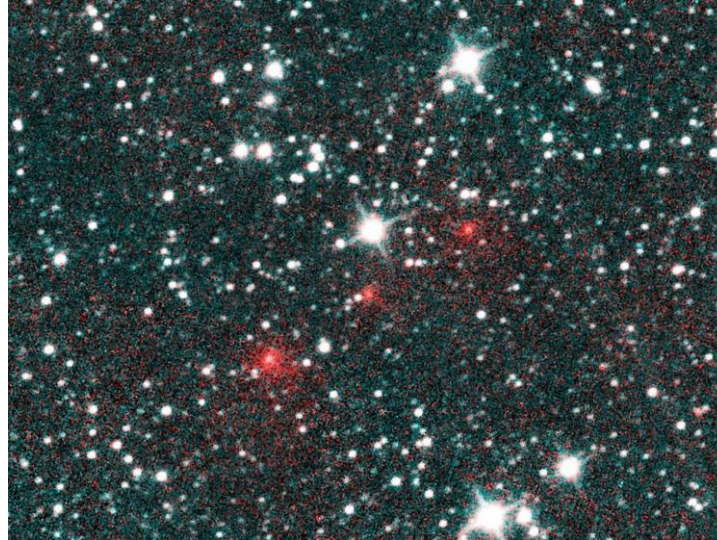
- The term “Data Science” was used for the first time in 1985 by Jeff Wu as an alternative name for statistics [\[source\]](#)
- It is often associated with the combination of big data, high performance computing, and machine learning



# Something new?

Turing award winner Jim Gray:

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

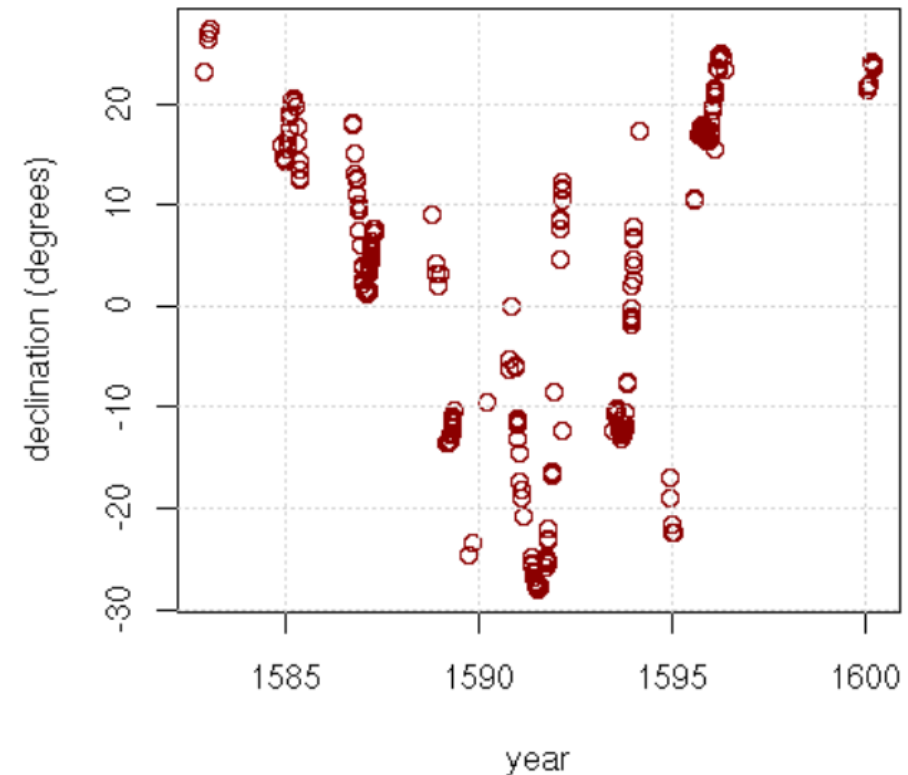


Data science for scientific discovery.  
Discovery of Comet NEOWISE (series of red dots). NASA/JPL-Caltech

# Something old?



**Tycho Brahe's Mars Observations**



source: Tychonis Brahe Dani Opera Omnia

# 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 is the process of

1. collecting data
2. cleaning data
3. analyzing data [\[source\]](#)



We'll use this definition

Data comes in many different forms

# Collecting data

	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	2,556.55
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	835.01
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
						319.20
						802.00

File Edit Format View Help

This is a .TXT file open in Microsoft Notepad.

© FileInfo.com

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Vivamus condimentum sagittis lacus, laoreet luctus ligula laoreet ut. Vestibulum ullamcorper accumsan velit vel vehicula. Proin tempor lacus eros. Nunc at ullam condimentum, semper nisi et, condimentum et. In venenatis blandit nibh at sollicitudin. Vestibulum dapibus mauris et orci sodales pellentesque. Nulla id elementum ipsum. Suspendisse cursus lobortis viverra. Proin et erat at mauris tincidunt porttitor vitae ac dui.

Donec vulputate lorem tortor, nec fermentum nibh bibendum vel. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Praesent dictum luctus massa, non euismod lacus. Pellentesque condimentum dolor est, ut dapibus lectus luctus ac. Ut sagittis commodo arcu. Integer nisi nulla, facilisis sit amet nulla quis, eleifend suscipit purus. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos himenaeos. Aliquam euismod ultrices lorem, sit amet imperdiet est tincidunt vel. Phasellus dictum justo sit amet ligula varius aliquet auctor et metus. Fusce vitae tortor et nisi pulvinar vestibulum eget in risus. Donec ante ex, placerat a lorem eget, ultricies bibendum purus. Nam sit amet neque non ante laoreet rutrum. Nullam aliquet commodo urna, sed ullamcorper sedlo feugiat id. Mauris nisi sapien, porttitor in condimentum nec, venenatis eu urna. Pellentesque feugiat diam est, et rhoncus orci porttitor non.

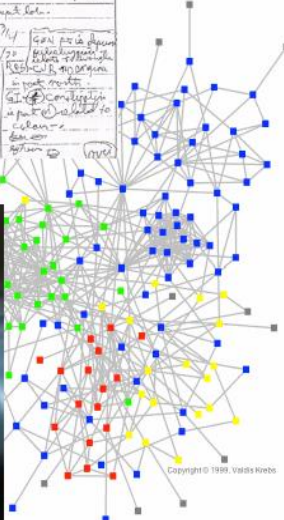
Nulla luctus sem sit amet nisi consequat, et ornare ipsum dignissim. Sed elementum elit nibh, eu condimentum orci viverra quis. Aenean suscipit vitae felis non suscipit. Suspendisse pharetra turpis non eros semper dictum. Etiam tincidunt venenatis venenatis. Praesent eget gravida lorem, ut congue diam. Etiam facilisis elit at porttitor agestas. Praesent consequat, velit non vulputate conwallis, ligula diam sagittis urna, in venenatis nisi justo et mauris. Vestibulum posuere sollicitudin et, et vulputate nisi fringilla non. Nulla ornare pretium velit a euismod. Nunc sagittis venenatis vestibulum. Nunc sodales libero a est ornare ultricies. Sed sed leo sed orci pellentesque ultrices. Mauris sollicitudin, sem quis placerat ornare, velit arcu conwallis ligula, pretium finibus nisi sapien vel sem. Vivamus sit amet tortor id lorem consequat hendrerit. Nulla et dui risus.

Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia Curae; Lorem ipsum dolor sit amet, consectetur adipiscing elit. Sed feugiat semper velit consequat facilisis. Etiam facilisis justo non laculis dictum. Nunc turpis nunc, pharetra et odio eu, hendrerit rhoncus lacus. Nunc orci felis, imperdiet vel interdum quis, porta eu ipsum. Pellentesque dictum sed lacinia, auctor dui in, malesuada nunc. Maecenas sit amet mollis eros. Proin fringilla viverra ligula, sollicitudin viverra ante sollicitudin congue. Donec mollis felis eu libero malesuada, et lacinia risu interdum.

Etiam vitae accumsan augue. Ut urna orci, malesuada ut nisi a, condimentum gravida magna. Nulla bibendum ex in vulputate sagittis. Nulla facilisi. Nullam faucibus et metus ac consequat. Quisque tempus eros velit, id semper nibh aliquet a. Aenean tempus elit et finibus auctor. Sed at imperdiet mauris. Vestibulum pharetra non lacus sed pulvinar. Sed pellentesque magna a eros vulputate ullamcorper. In hac habitasse platea dictumst. Donec ipsum et, feugiat in eros sed, varius lacinia turpis. Donec vulputate tincidunt dui et laoreet. Sed in eros dui. Pellentesque placerat tristique ligula in finibus. Proin nec faucibus felis, eu commodo ipsum.

Integer eu hendrerit diam, sed consequat nunc. Aliquam a sem vitae leo fermentum faucibus quis at sem. Etiam blandit, quam quis fermentum varius, ante urna ultricies luctus, vel pellentesque ligula arcu nec elit. Donec placerat ante in ante scelerisque pretium. Donec at rhoncus erat. Aenean tempus nisi vitae augue tincidunt luctus. Nam condictum dictum ante, et laoreet neque pellentesque id. Curabitur consectetur cursus neque aliquam porta. Ut interdum nunc nec nibh vestibulum, in sagittis metus facilisis. Pellentesque feugiat condimentum metus. Etiam venenatis quam at ante rhoncus vestibulum. Maecenas suscipit congue pellentesque. Vestibulum suscipit scelerisque.

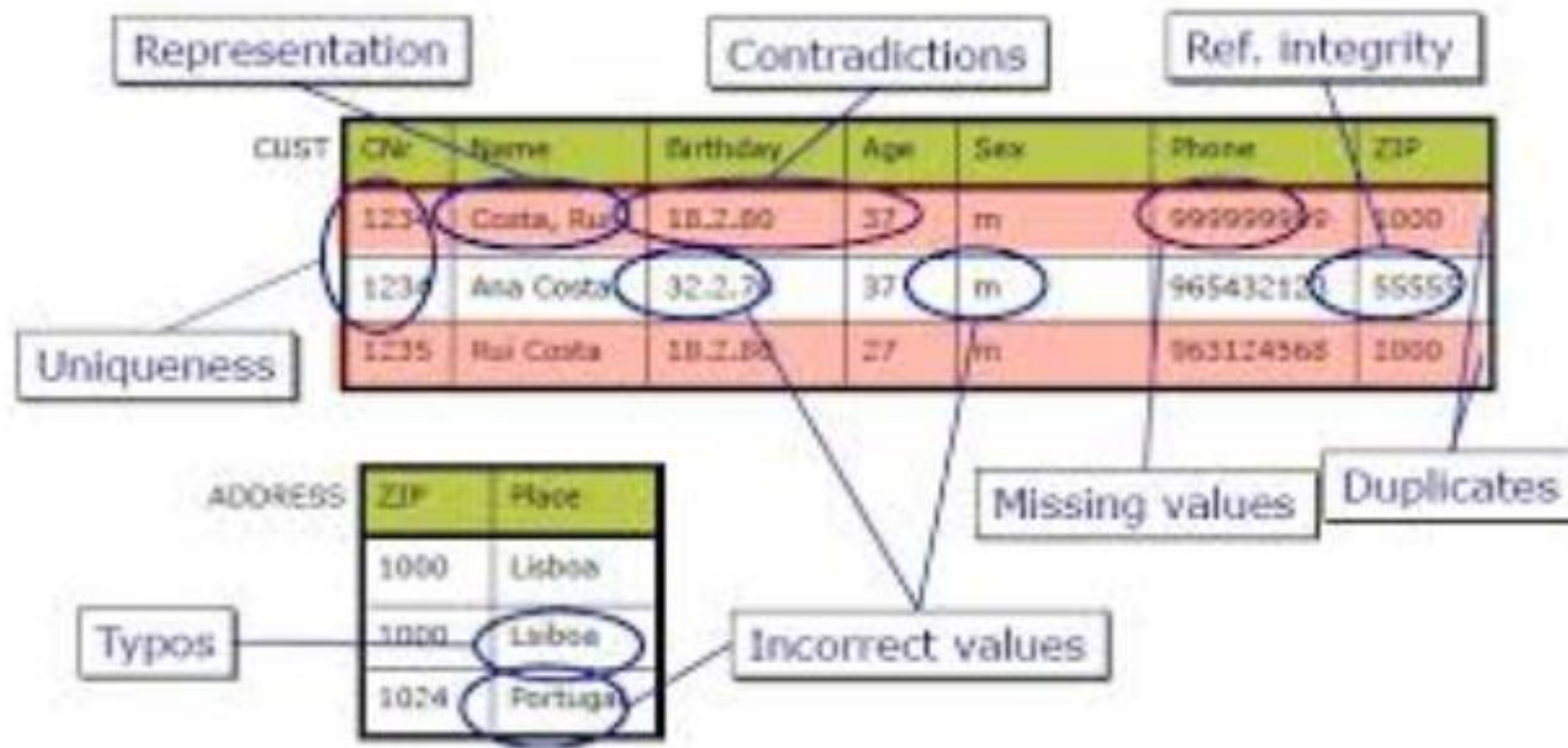
Handwritten notes and data on a lined paper, including dates, names, and numerical values.





Data commonly  
needs pre-processing  
before it can be used

# Cleaning data



[Source](#)

Analyze the data using statistics,  
machine learning, and visualization

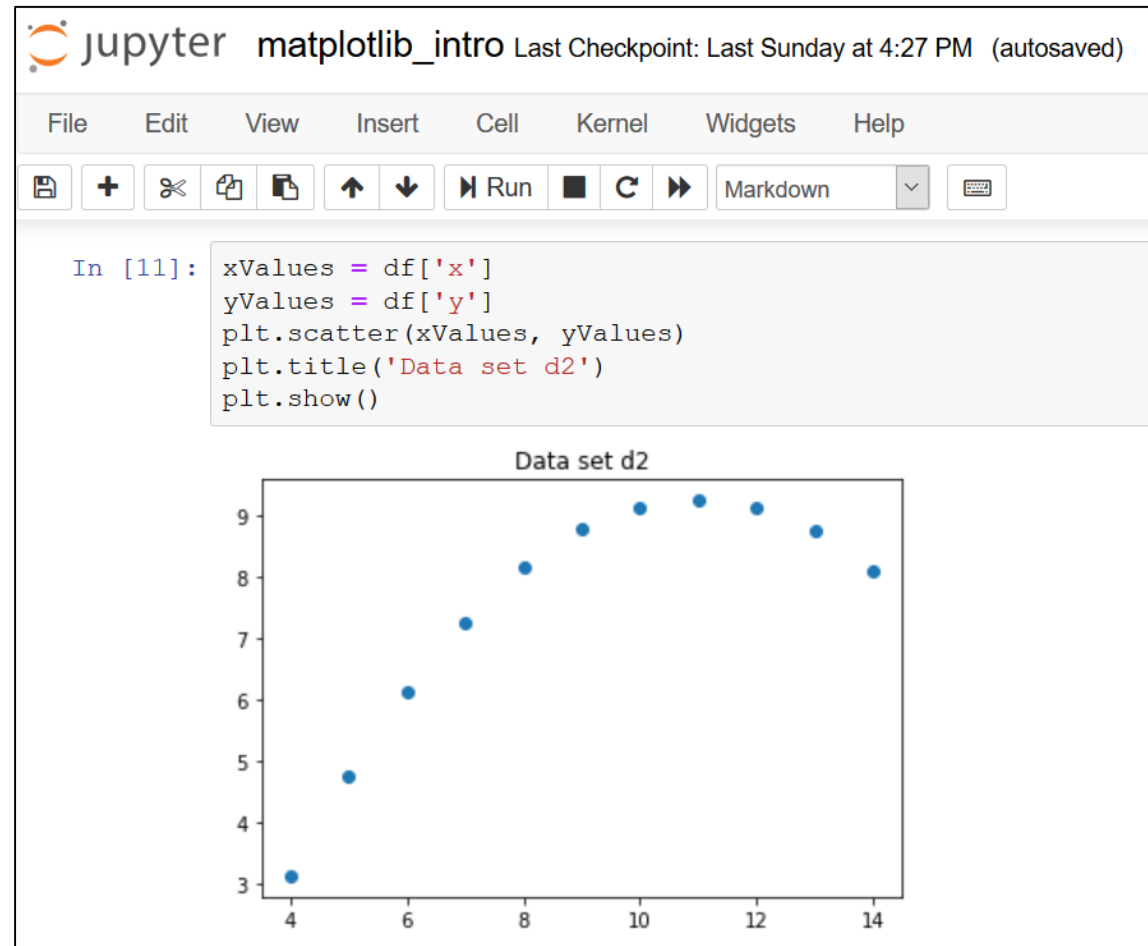
# Analyzing data

Programming languages:  
Python, R, Mathematica, Excel...

Notebook environments:  
Mathematica or Jupyter

Many specialised  
software packages

data formats:  
CSV, XML, SQL, JSON,...



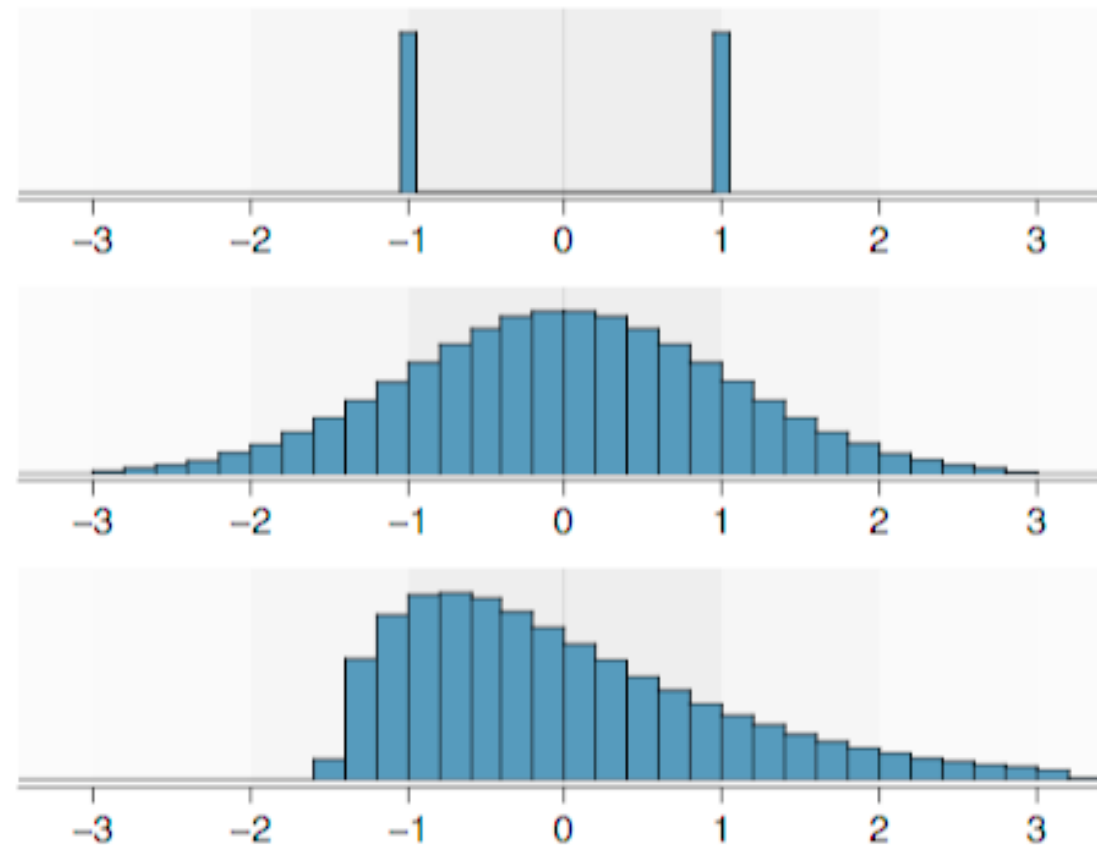
# Visualizing data





# Limits of statistics

A *statistic* is a number that measures a property of a dataset. Examples include mean and standard deviation



Very different data sets with the same mean and standard deviation

Illustrates the need to look at the data!

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

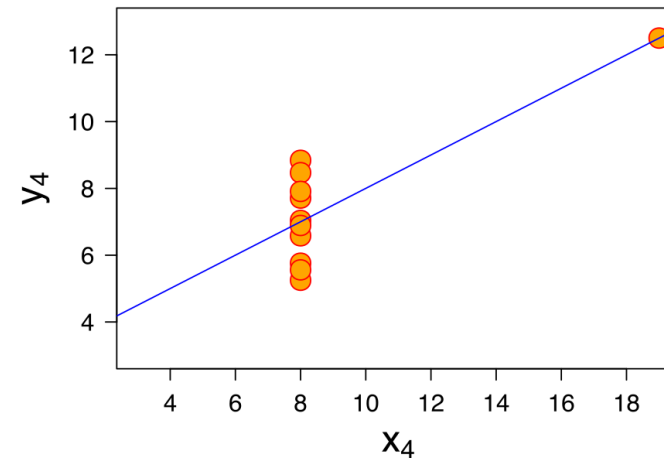
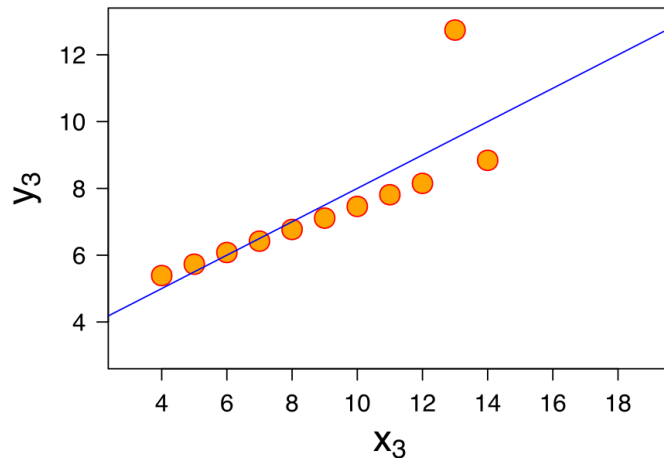
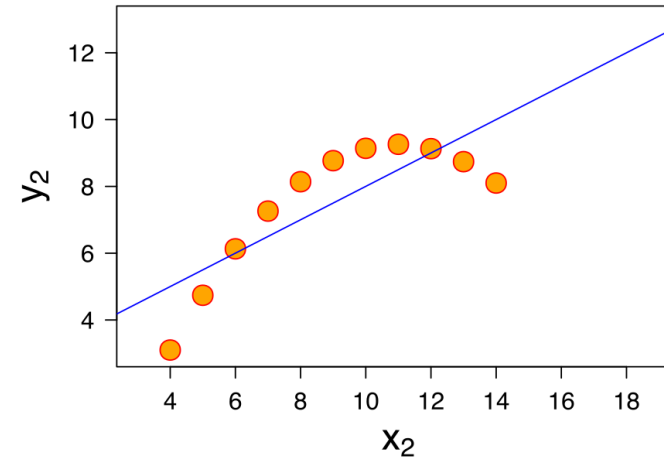
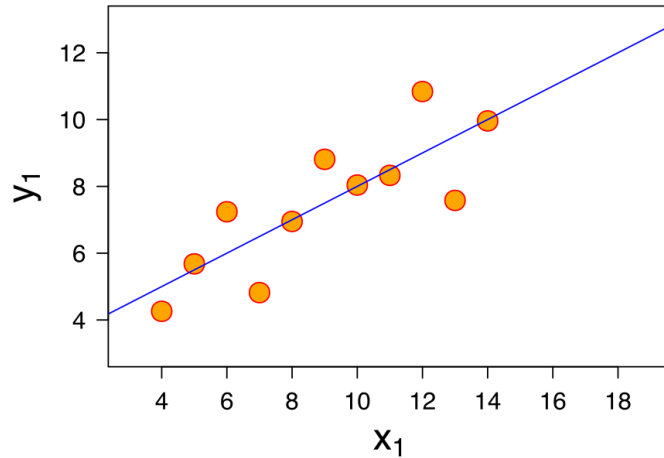
# Limits of mathematical models

Anscombe's quartet

Very different data sets have the same regression line

Very different realities have the same mathematical model

Illustrates the need to look at the data!



# 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 sheet of paper can be regarded as 2D, but we need 3D to represent a pencil drawing on it: (x,y,color)
- On the computer, a 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

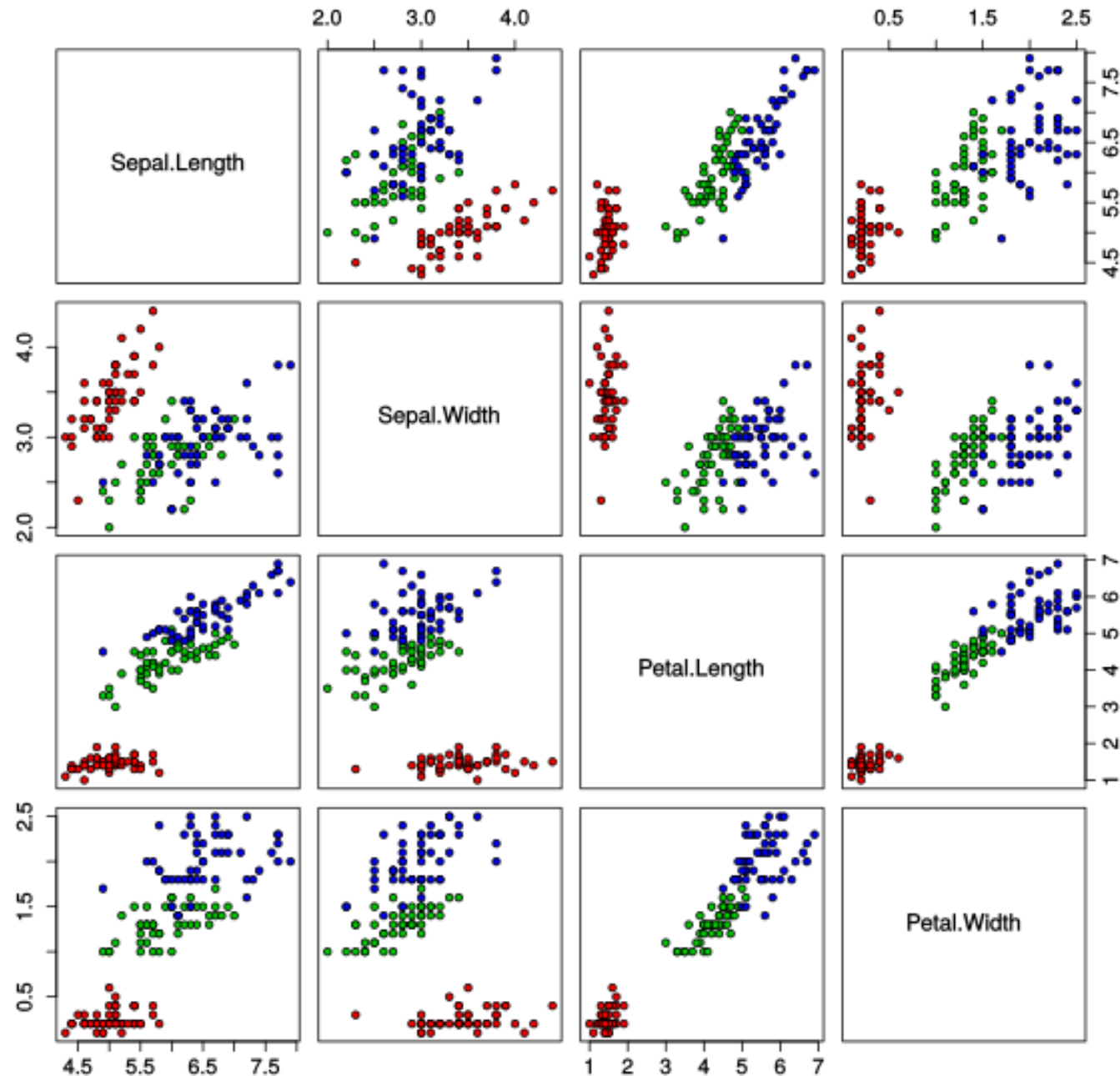


Iris  
virginica



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

# Iris Data (red=setosa,green=versicolor,blue=virginica)



We can't visualize 5D datapoints, but we can do this

Everything times everything

# Scatter plot

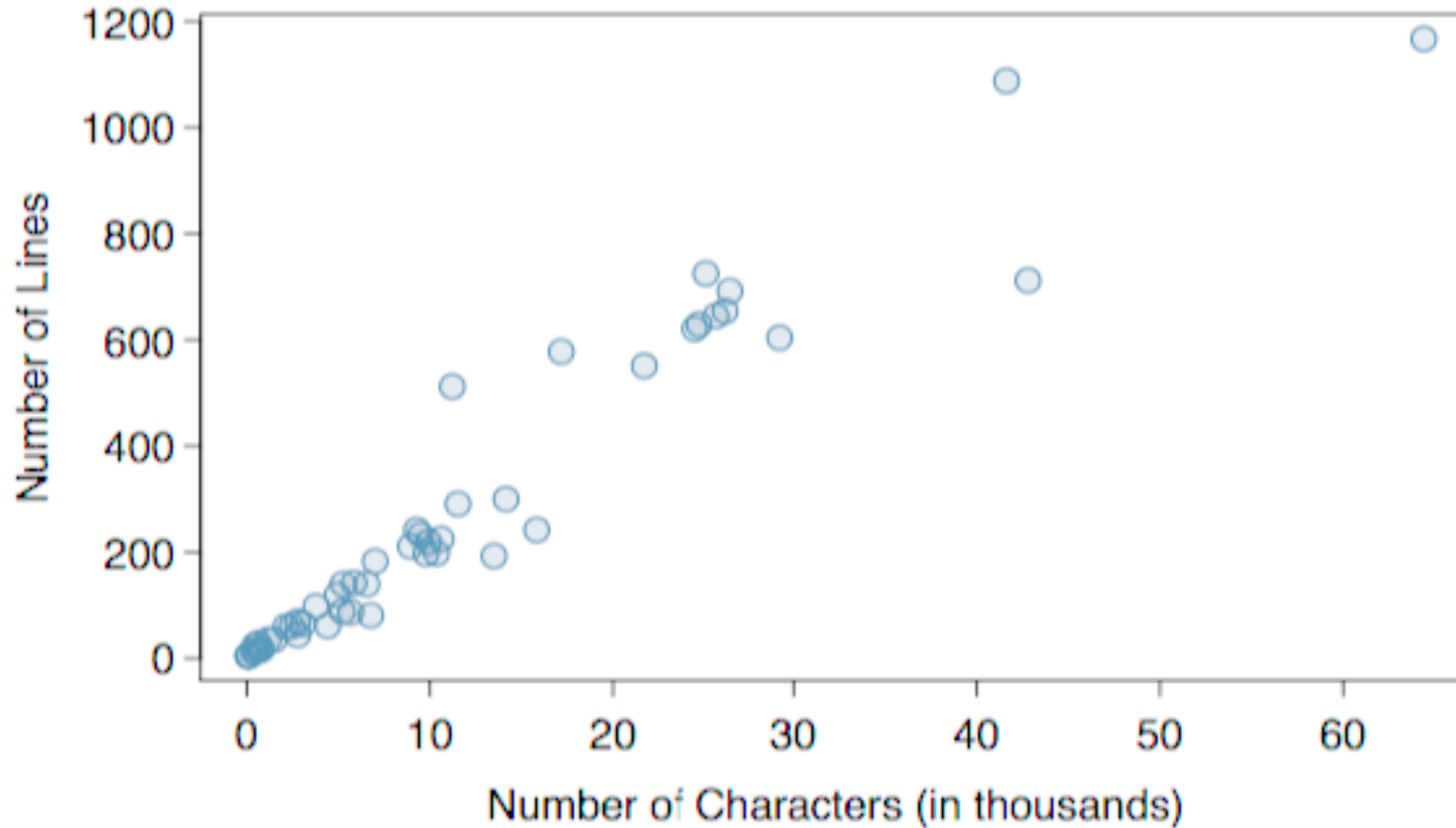
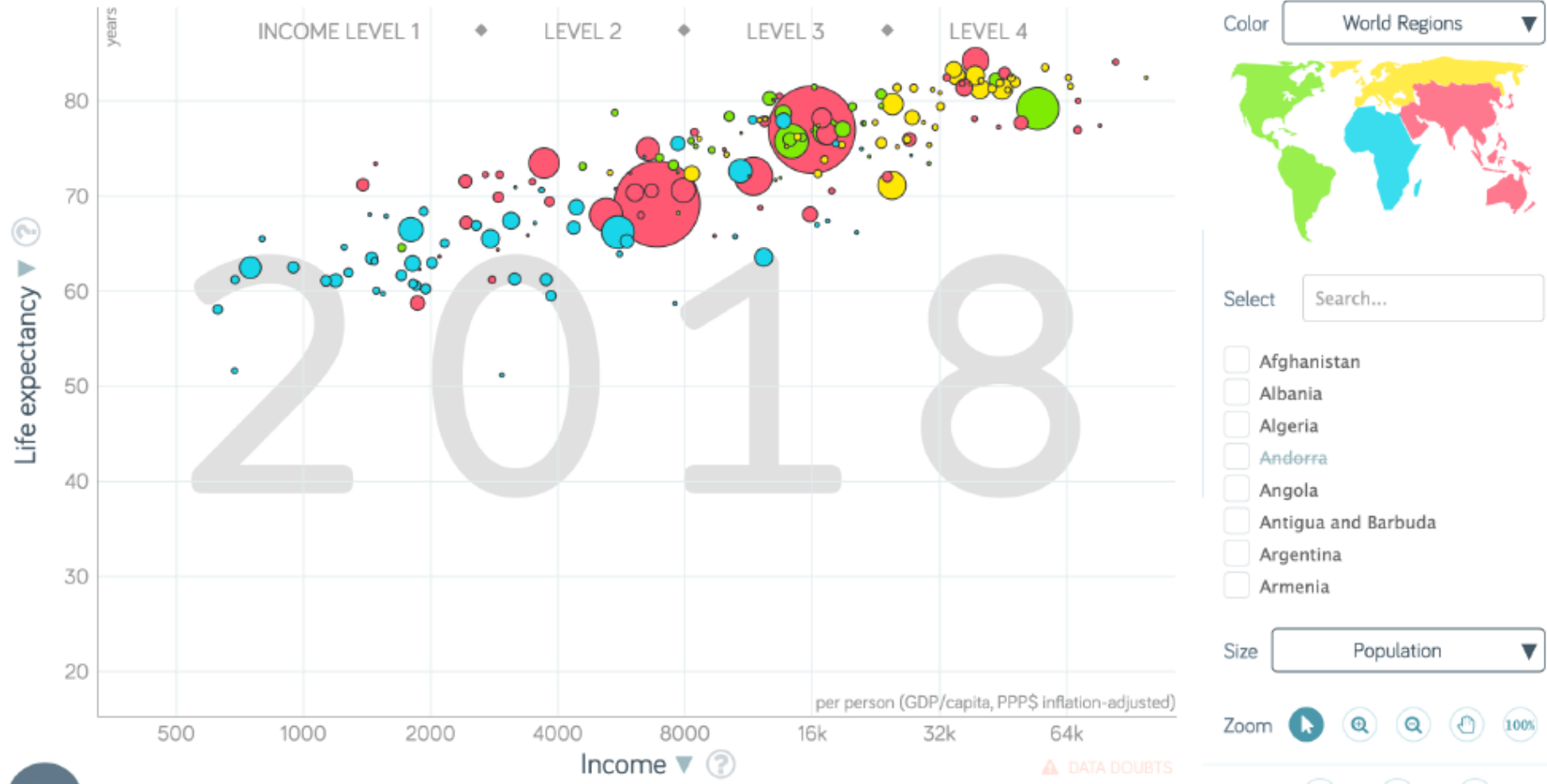


Figure 1.17: A scatterplot of `line_breaks` versus `num_char` for the `email150` data.

# Scatter plot



Here colors add a third dimension

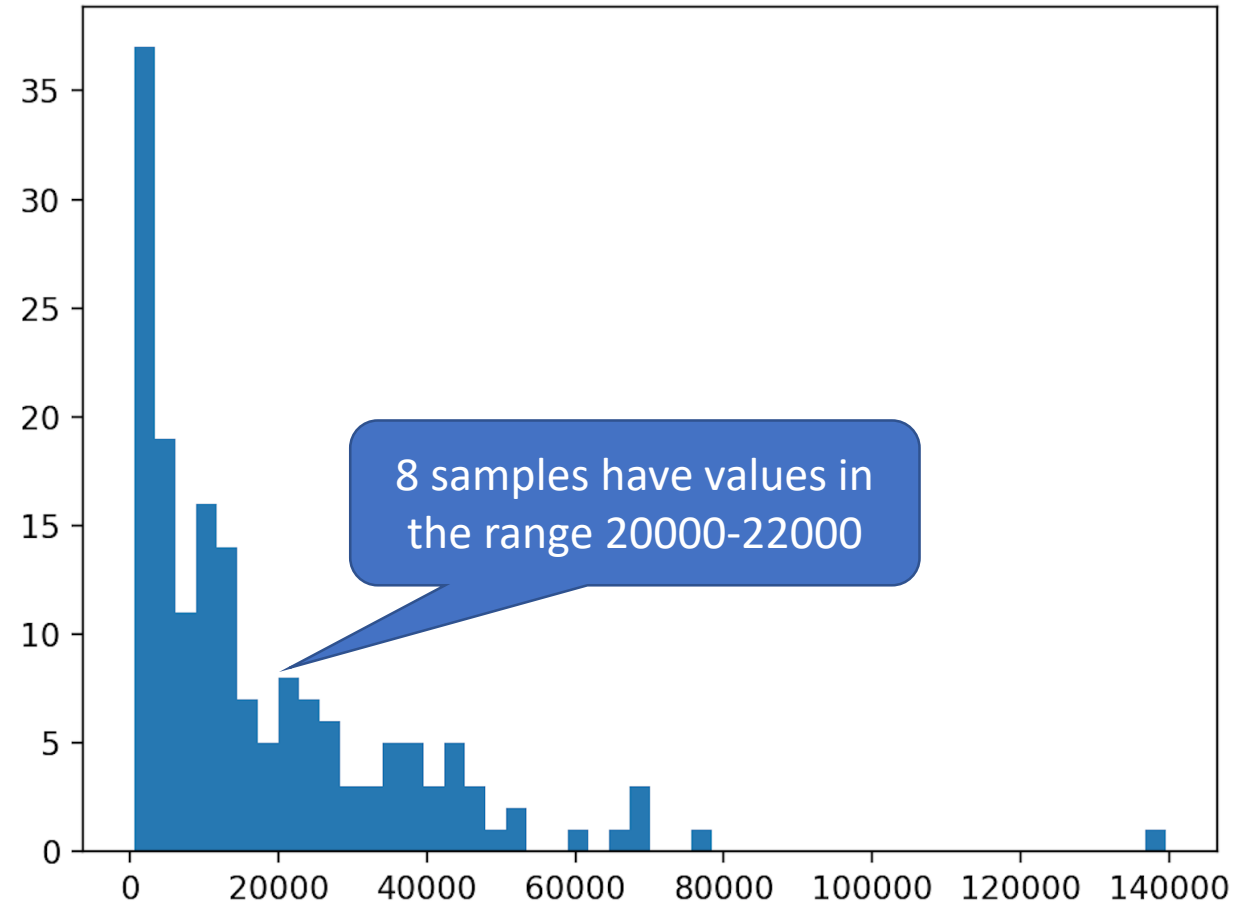
Disk size a fourth

With movies we can represent time as a fifth dimension

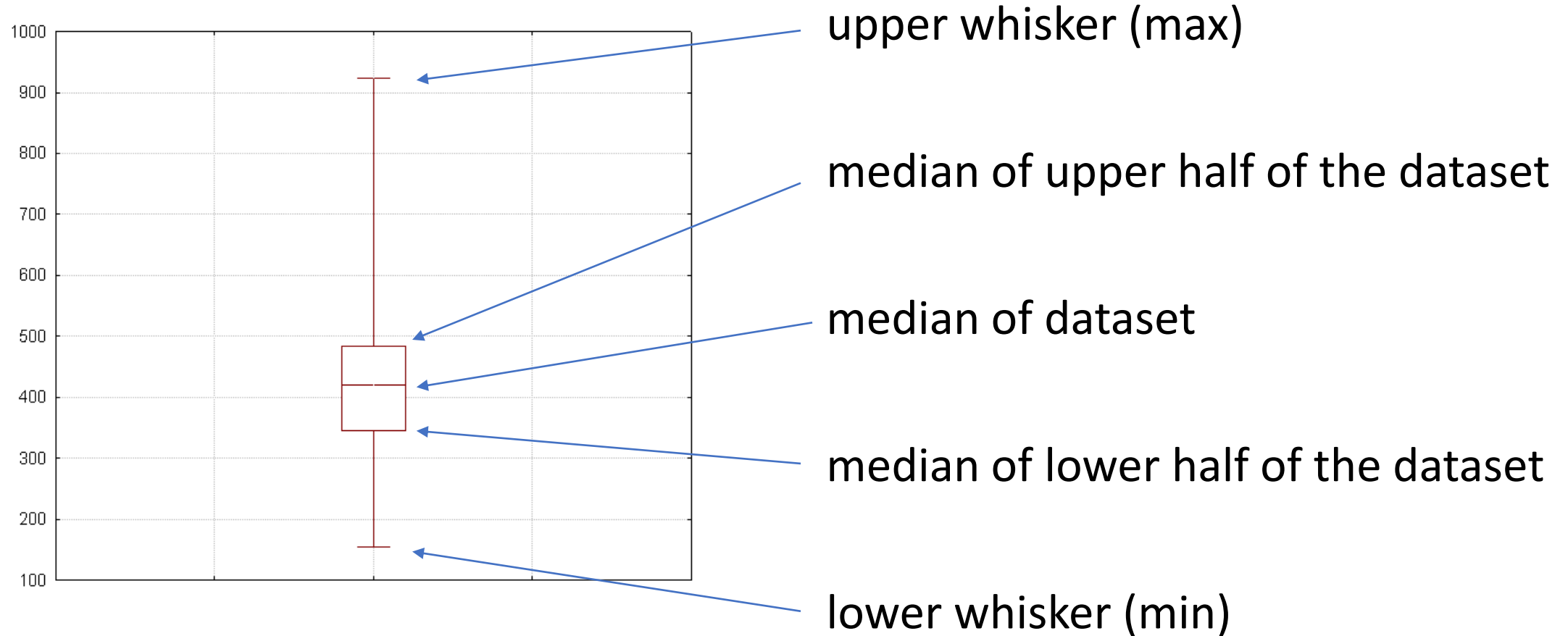
[Gapminder](https://www.gapminder.org)



# Histogram



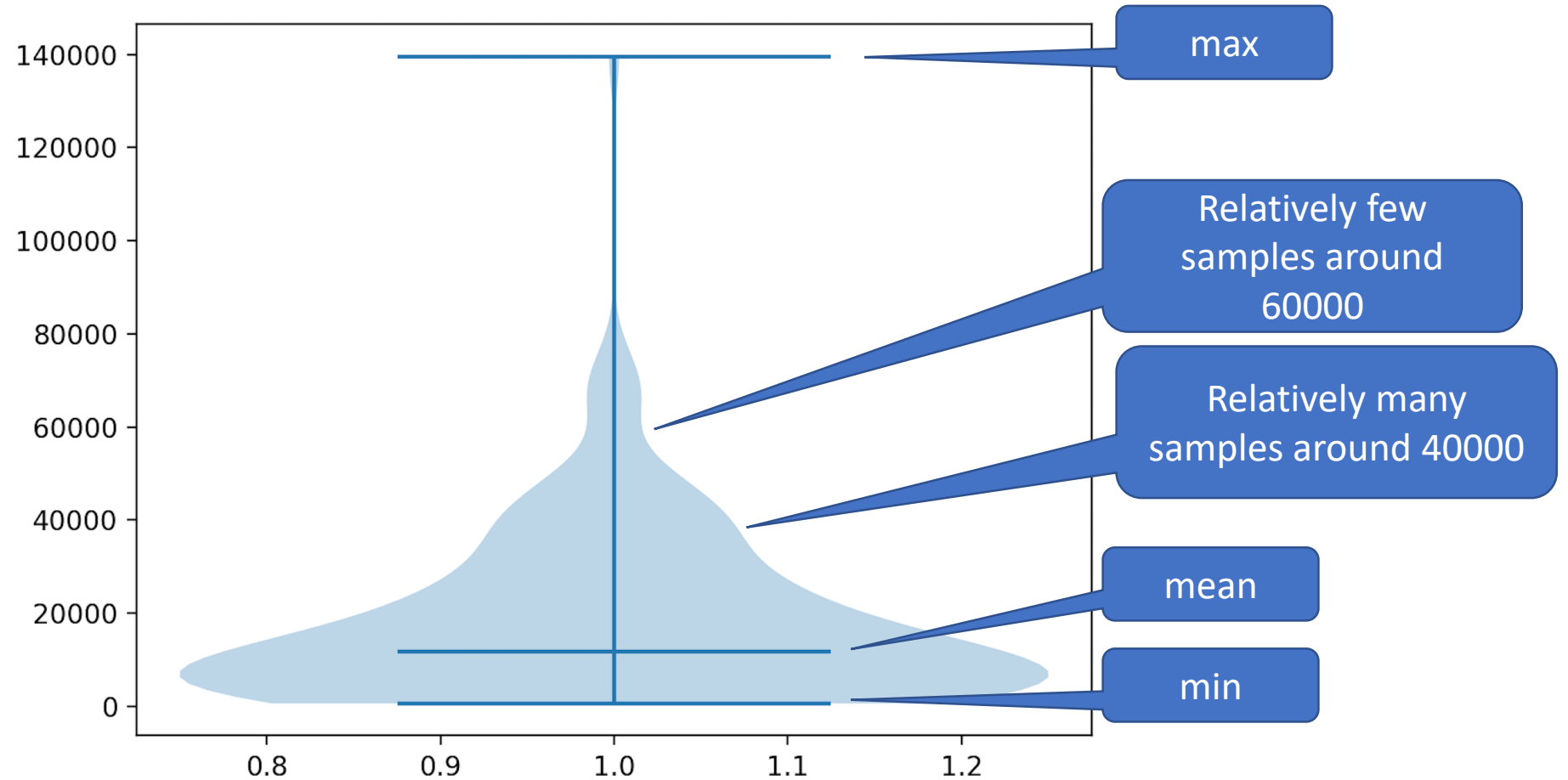
# Boxplot



A boxplot is a way of displaying a set of numbers as a five-number summary

# Violin plot

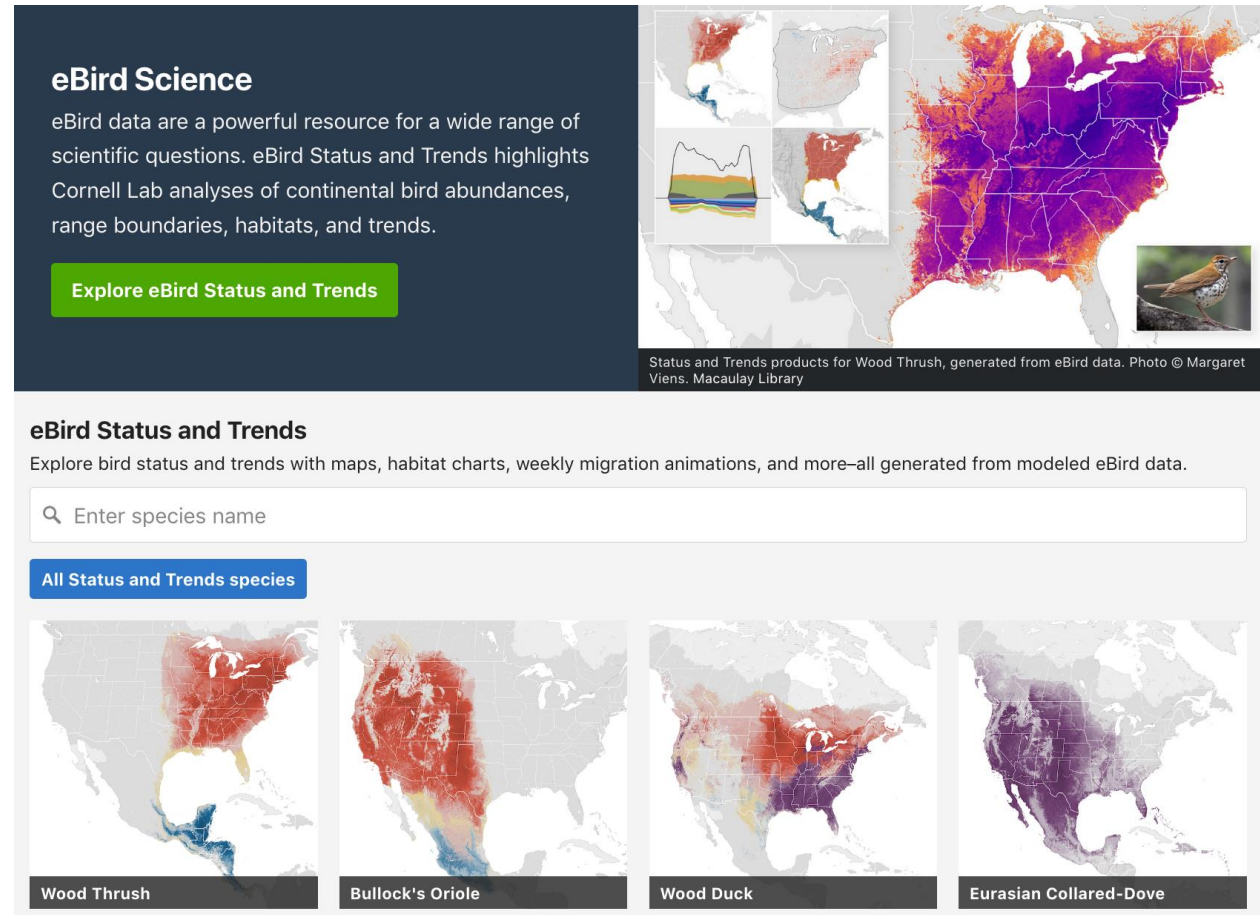
Like a box plot with probability density information added



Where are  
the birds?

# Ordinary map

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



[Source](#)

# Heat map

Time spent eyeballing  
different spots of a web page.



Scrolling speed for different  
parts of a web page.



[Source](#)

# Tink about the visualisation aesthetic

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

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 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/>

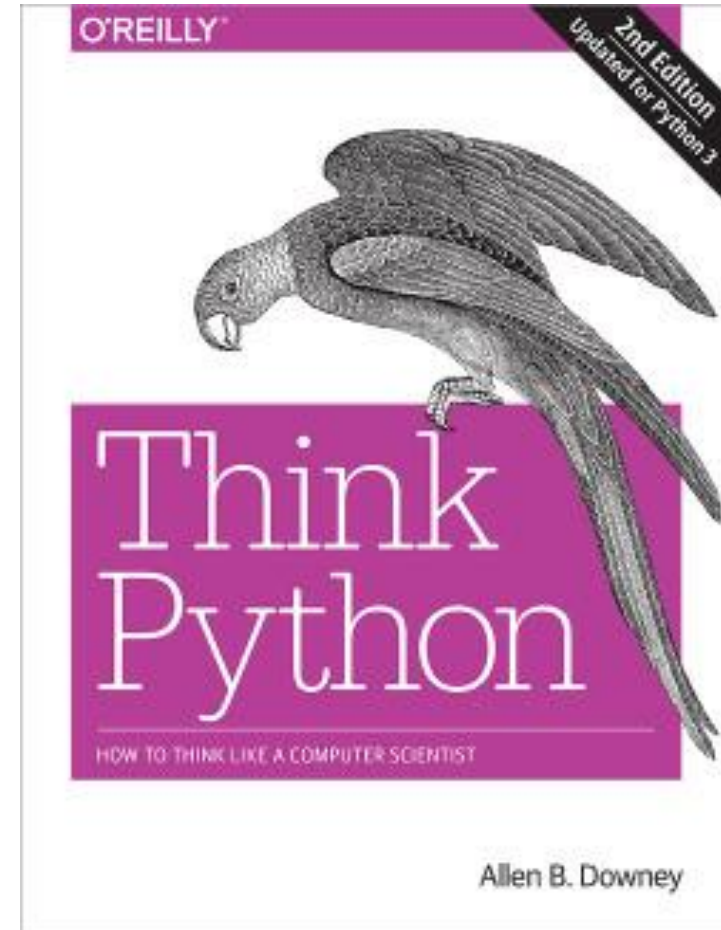
# Open a Jupyter notebook

- Make a new directory and save the course notebooks from Canvas (@Modules) there.
- Open the program Jupyter Notebook. Then you get a “File explorer” tab in your web browser.
- Open some notebook, e.g., jupyter\_intro from this “File explorer”.

Python programming

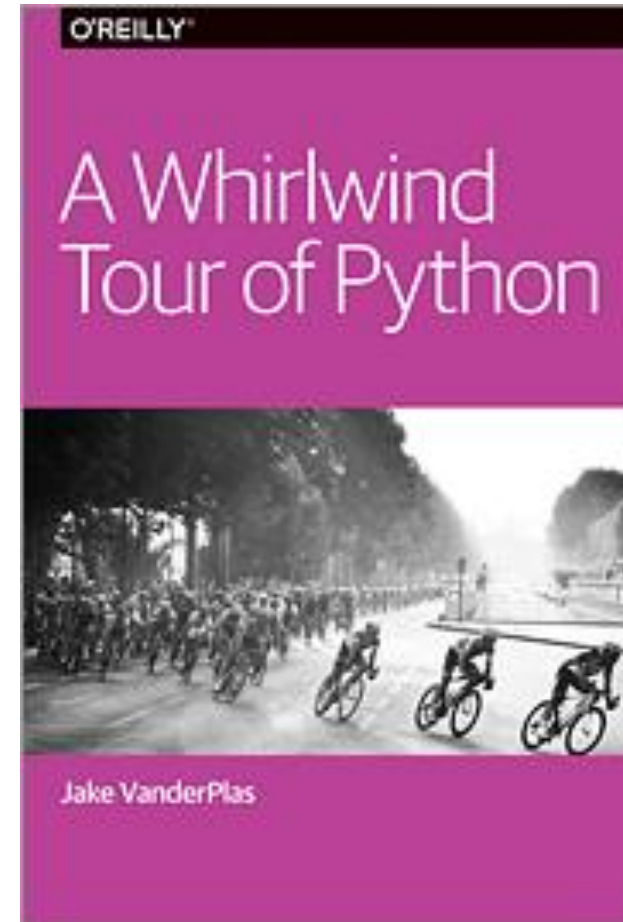
# Python programming

- 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
```

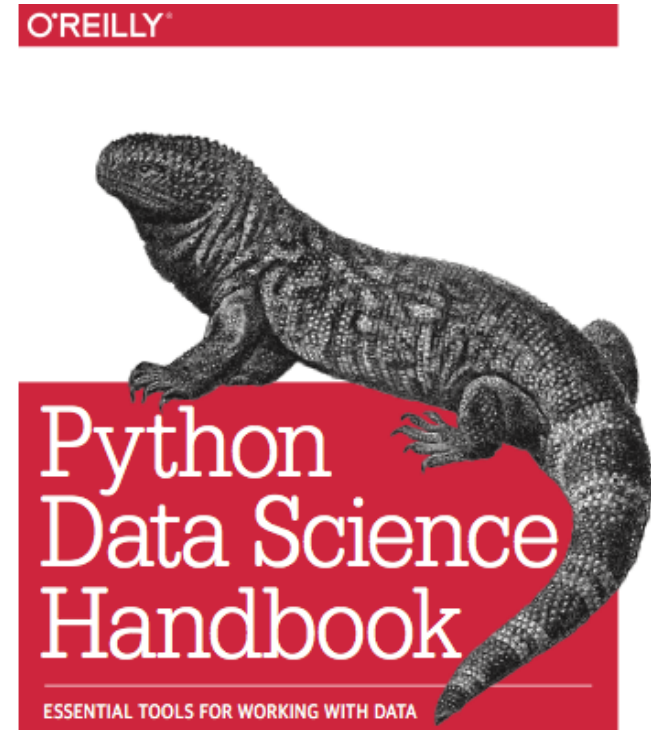
```
import numpy as np
```

```
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](#)



Jake VanderPlas

# Quick introductions

- Let's have a look at some notebooks (@Modules):
  - jupyter\_intro
  - python\_intro
  - pandas\_intro
  - numpy\_intro
  - matplotlib\_intro