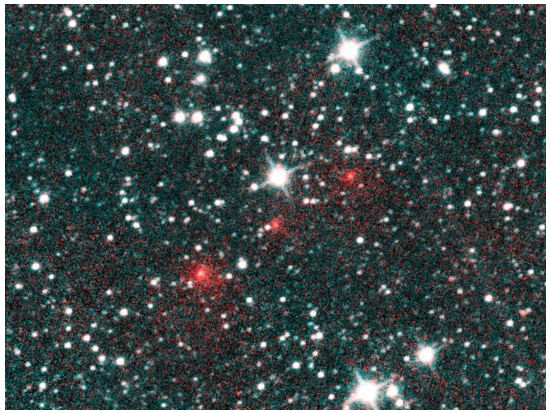
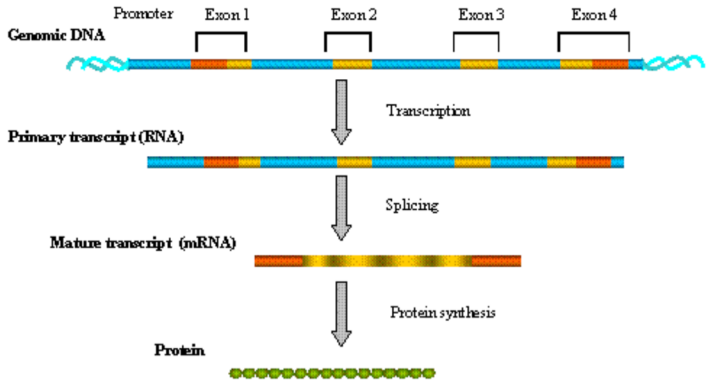
Data Science with Machine Learning

Guillem & Roderic, Summer 2025

Biological School of Data Science

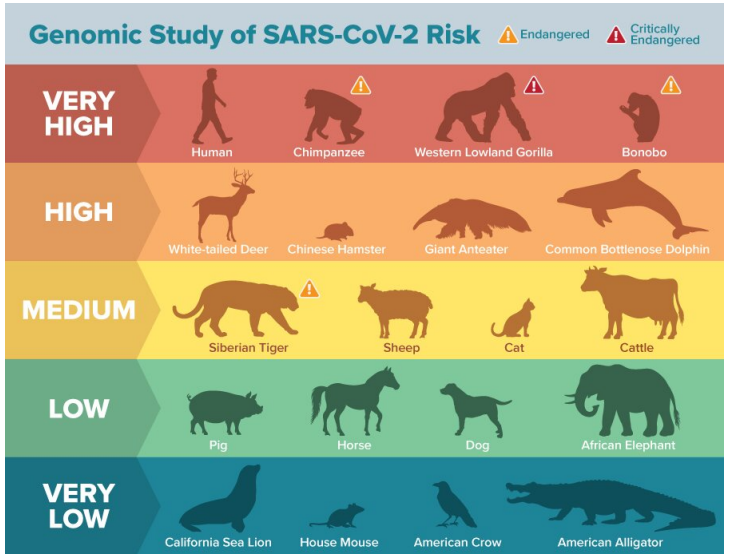
- Biology – branch of science that studies living beings
- Data science – interdisciplinary field that uses scientific methods to extract or extrapolate knowledge from data
- Data – collection of values that contain information

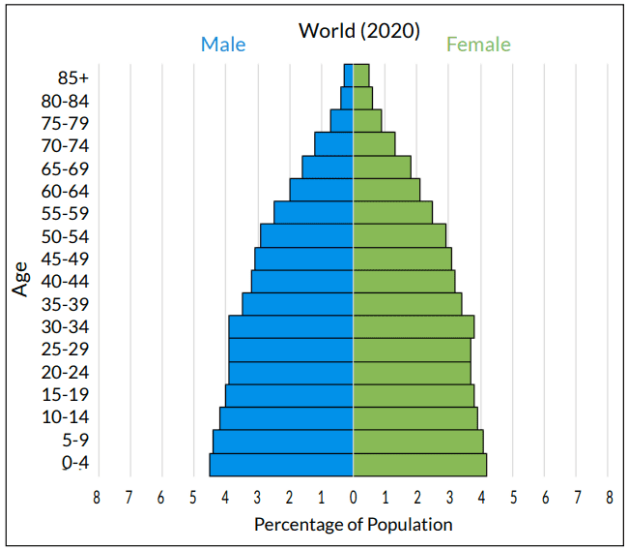




What can we do with data?

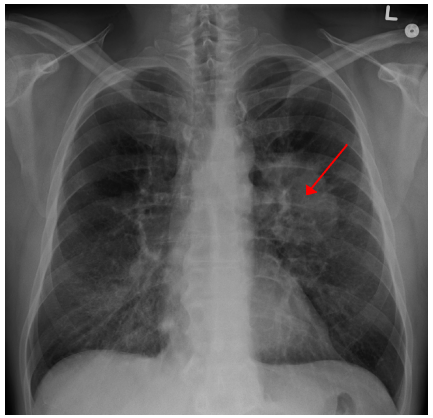
- Understand something about the real world

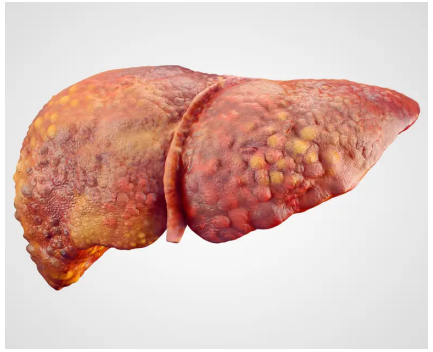




What can we do with data?

- Draw conclusions $\rightarrow$ B is caused by A



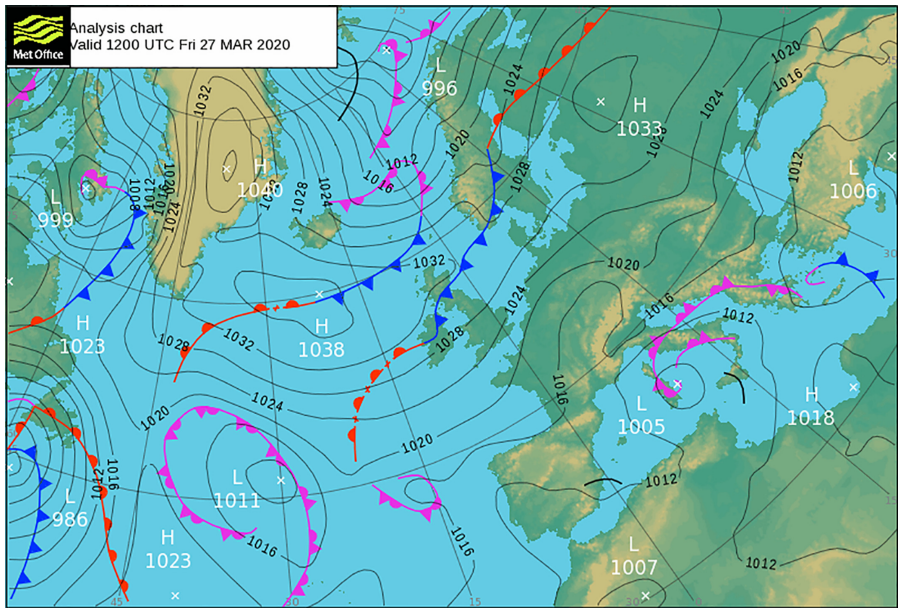


What can we do with data?

- Make predictions about future events → modeling

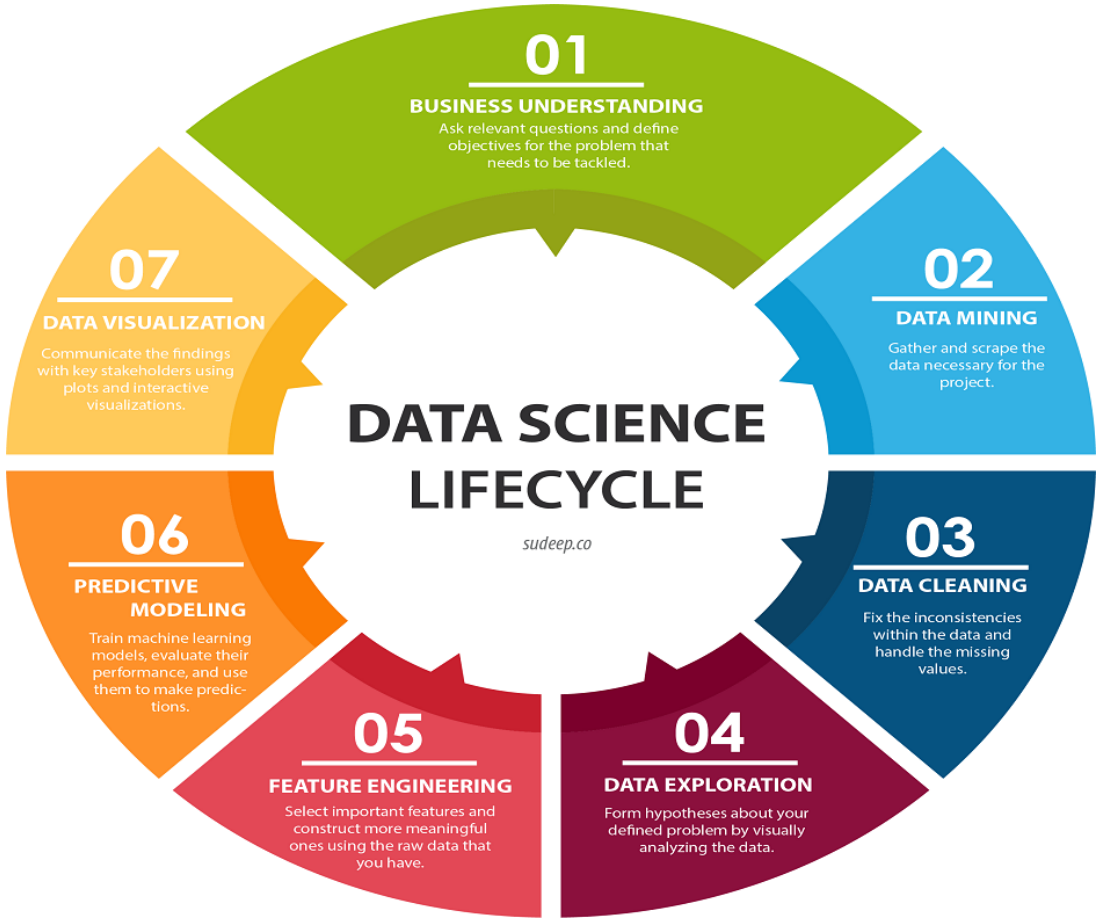
→ Model → Cat

→ Model → Dog



Data Science

- Data collection
- Data analysis
- Descriptive statistics
- Data processing
- Algorithms
- Model training
- Inferential statistics
- Deployment



Computation

- Systematic approach to manipulate data
- Observations → from qualitative to quantitative descriptions
- Computer: machine that performs logical and arithmetic operations following instructions – programmable
- Algorithm: sequence of well-defined instructions

Data representation

- Data is represented by numbers → *D* dimensional vectors

Features

Observations

Height	Weight
178	77
186	82
168	66

$D = 2$

$N = 3$

$x^1 = [178, 77]$

$x^2 = [186, 82]$

$x^3 = [168, 66]$

$D = 3$

$N = 4$

m^2	*n* rooms	year
80	2	2014
75	3	1987
92	3	1999
130	4	2005

$$\begin{bmatrix} 80 & 2 & 2014 \\ 75 & 3 & 1987 \\ 92 & 3 & 1999 \\ 130 & 4 & 2005 \end{bmatrix}$$

4x3 matrix

$D = 5$

$N = 5$

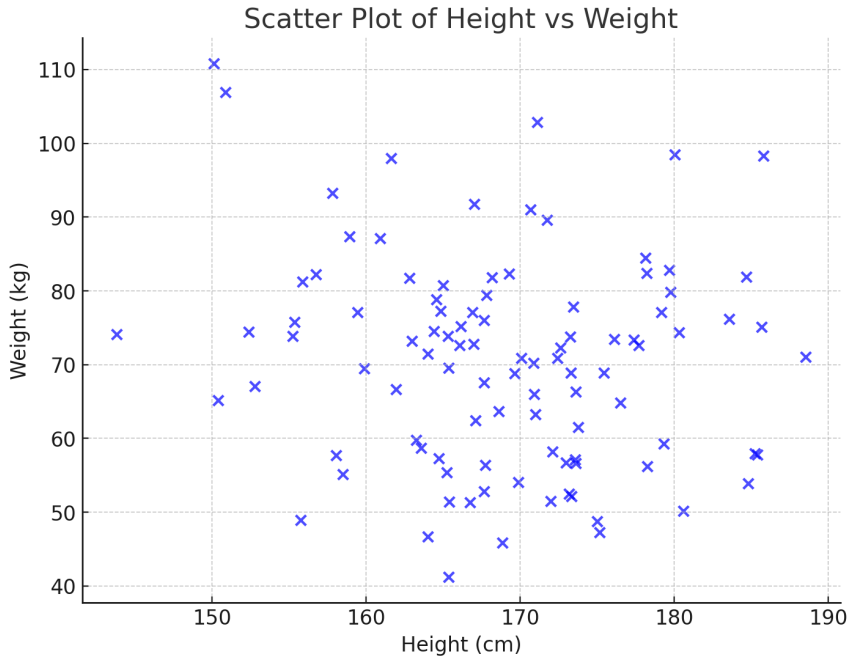
x_1	x_2	x_3	x_4	x_5
0.01	0.41	0.78	0.93	0.53
0.86	0.09	0.61	0.07	0.39
0.22	0.02	0.54	0.47	0.80
0.28	0.30	0.56	0.73	0.64
0.42	0.66	0.76	0.04	0.24

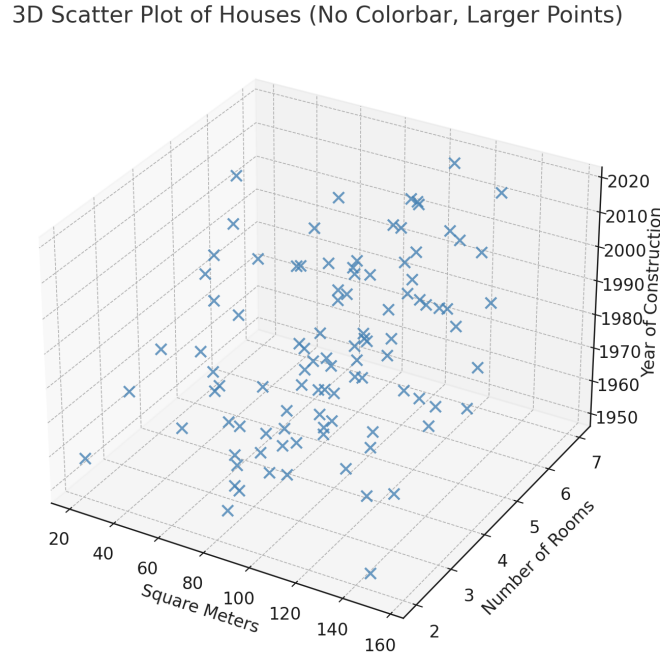
$x_i^j \rightarrow i$th feature of the jth observation

Dataset → Set of all observations

Data representation

- Each observation is represented by a point in the dimension of the feature space

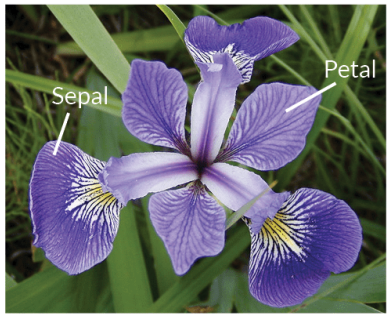




Euclidean distance $\rightarrow$ $d(p,q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_n - q_n)^2}.$

Example: Iris dataset

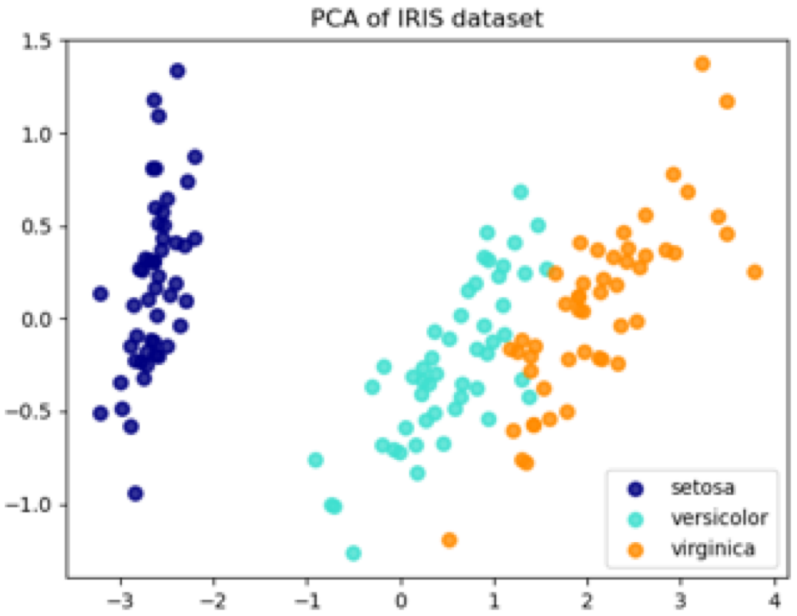
Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species
1	5.1	3.5	1.4	0.2	Iris-setosa
2	4.9	3	1.4	0.2	Iris-setosa
3	4.7	3.2	1.3	0.2	Iris-setosa
4	4.6	3.1	1.5	0.2	Iris-setosa
5	5	3.6	1.4	0.2	Iris-setosa
6	5.4	3.9	1.7	0.4	Iris-setosa
7	4.6	3.4	1.4	0.3	Iris-setosa
8	5	3.4	1.5	0.2	Iris-setosa
9	4.4	2.9	1.4	0.2	Iris-setosa
10	4.9	3.1	1.5	0.1	Iris-setosa
11	5.4	3.7	1.5	0.2	Iris-setosa
12	4.8	3.4	1.6	0.2	Iris-setosa
13	4.8	3	1.4	0.1	Iris-setosa
14	4.3	3	1.1	0.1	Iris-setosa
15	5.8	4	1.2	0.2	Iris-setosa
16	5.7	4.4	1.5	0.4	Iris-setosa
17	5.4	3.9	1.3	0.4	Iris-setosa
18	5.1	3.5	1.4	0.3	Iris-setosa
19	5.7	3.8	1.7	0.3	Iris-setosa



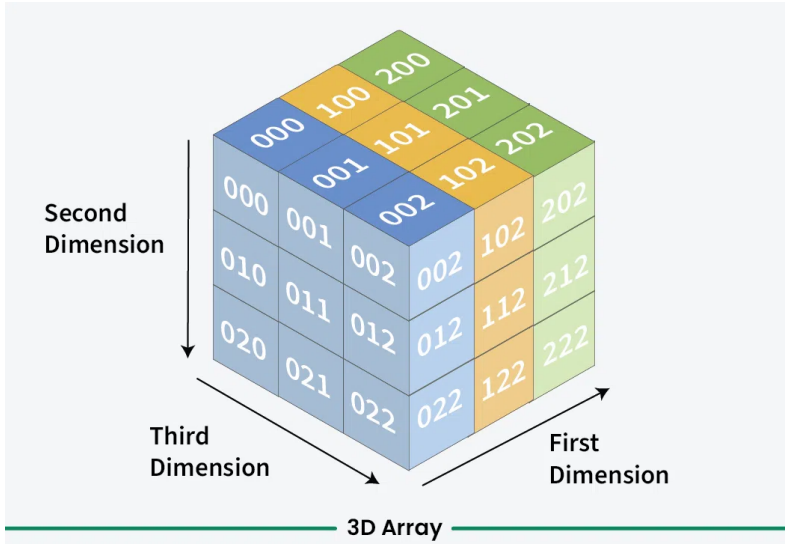
Iris Versicolor

Iris Setosa

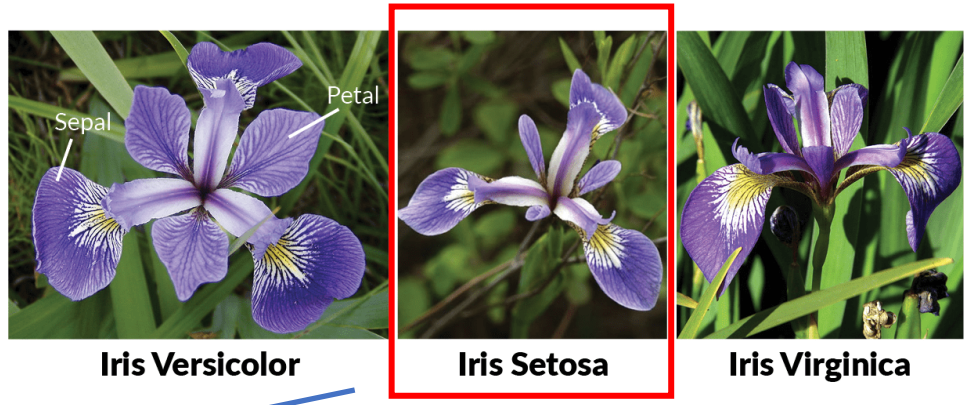
Iris Virginica

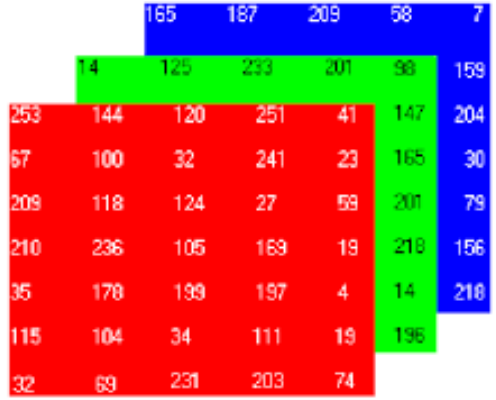


Example: Iris dataset



Tensor

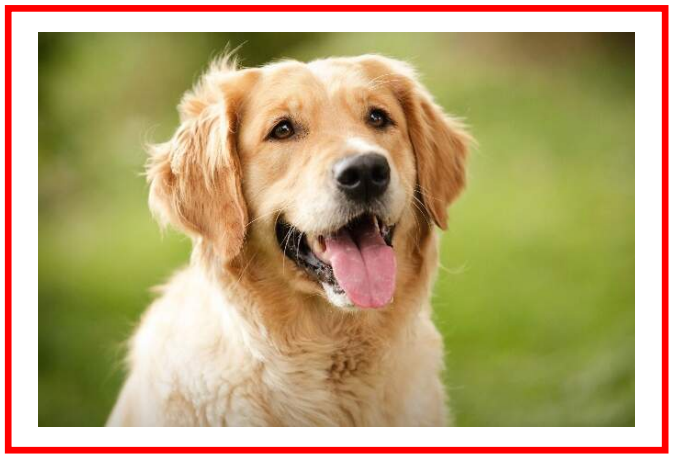




RGB

$$X_1 = [0.99, 0.64, 0.83, 0.37, 0.56, 0.92, 0.90, 0.58, 0.27, 0.09, \ldots]$$

Data representation



→	1	3	4	8	0
→	0	4	12	20	1

Labels

	Pointy ears	Eye separation	Snout length	Mouth width	Label
→	yes	3cm	4cm	8cm	Cat
→	no	4cm	12cm	20cm	Dog

Data representation

- The way we process the available data is determined by what we want to achieve

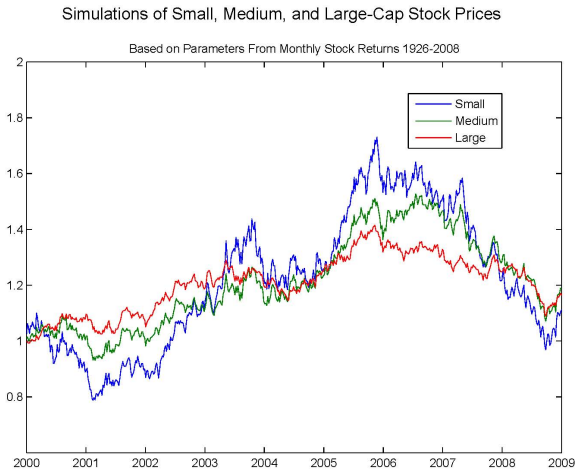
$$\begin{bmatrix} RBC \\ HB \\ HCT \\ WBC \\ MCV \\ \dots \end{bmatrix}$$

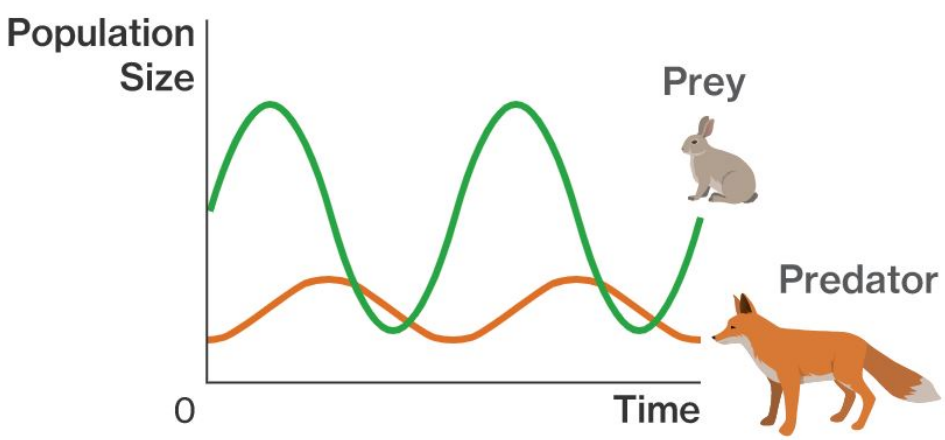
Leucemia or healthy?

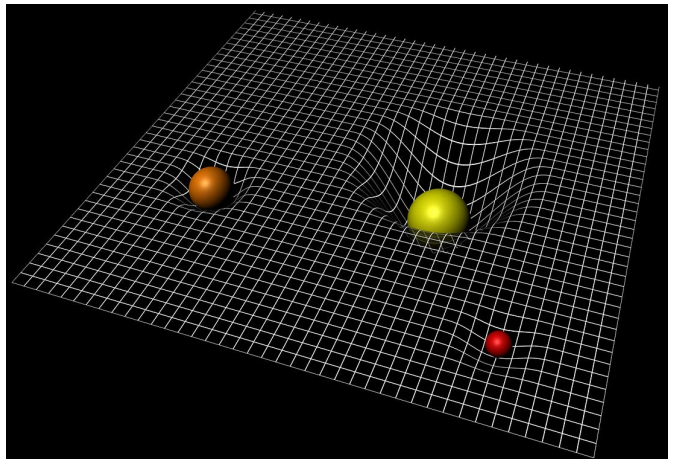
Katy Perry or Miley Cirus?

Modeling

- A mathematical model is a representation of a real-world system using mathematical concepts and language
- We usually make assumptions about how these systems behave
- Models are described by mathematical functions

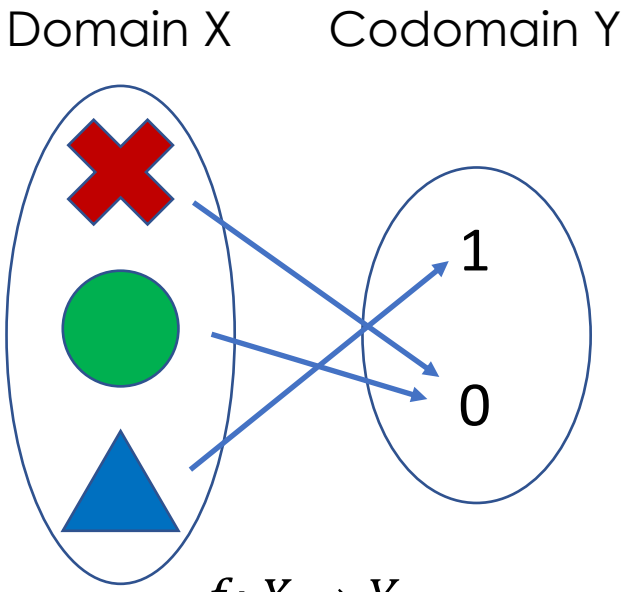




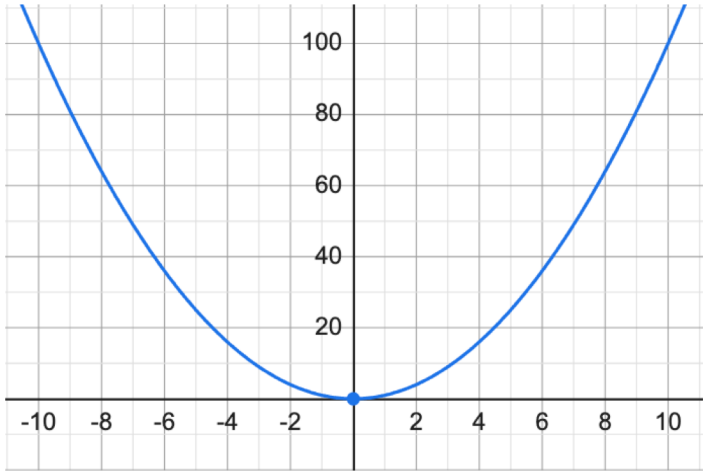


Functions

- Functions describe relationships between sets
- A function from set A to set B assigns to each element of A exactly one element of B → map
- Continuous or discontinuous
- Graph: set of ordered pairs (x, y) where $f(x) = y$
- $g\colon \mathbb{R} \to \mathbb{R}^+$, $g\colon \mathbb{C} \to \mathbb{C}$



$f: X \to Y$



$g(x) = x^2$

Predictions from data

- A model applies a transformation to the input data to generate a prediction

Transformation

m^2	*n* rooms	year
80	2	2014
75	3	1987
92	3	1999
130	4	2005

Model of house price

$$price = 100 \cdot m^2 + 5000 \cdot n\,rooms + 5 \cdot year + 10000$$

Predictions from data

$f(m^2, n\ rooms, year) \rightarrow f: \mathbb{R}^3 \rightarrow \mathbb{R}^1$

Regression: prediction of a continuous value

(cat and dog feature representation)

$f(x_1, x_2, x_3 x_4) \rightarrow f: \mathbb{R}^4 \rightarrow \mathbb{R}^1$

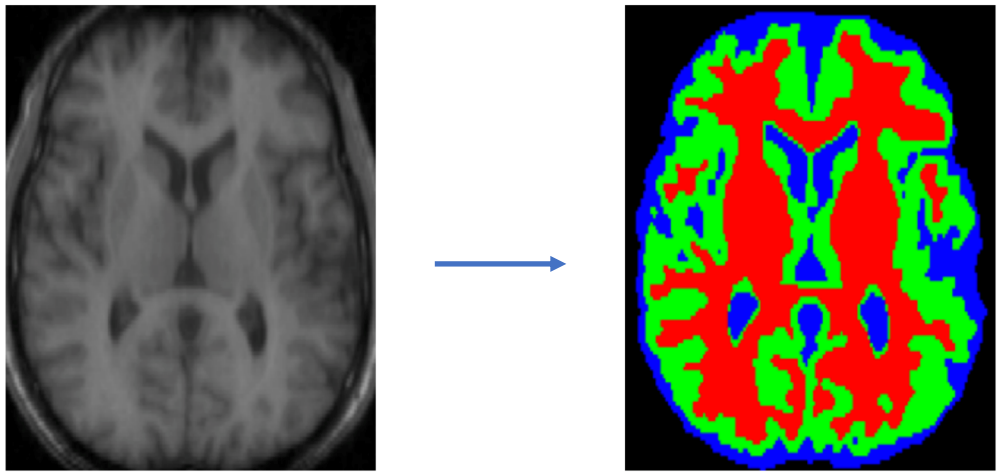
Classification: prediction of a discrete category

(cat and dog image representation)

$f(P) \rightarrow f: \mathbb{R}^{H \times W \times 3} \rightarrow \mathbb{R}^1$

Brain tissue segmentation

$f(P) \rightarrow f: \mathbb{R}^{H \times W} \rightarrow \mathbb{R}^{H \times W}$



(a) Axial slice (b) Tissue segmentation

How do we come up with the correct model?

- How can we find the transformation that for a given input returns the correct output?

Model → Cat

Model → Dog

Parametrization

- Parameters are constants that define a particular transformation
- In a multivariate linear model (for example the price of a house):

  $f(x_1, x_2) = ax_1 + bx_2 + c$

  x_1, x_2 are variables a, b, c are parameters
- f is parametrized by x_1, x_2
- In the following cubic model:

  $g(x) = 2x^3 - 1x^2 + 3x - 4$

  x is a variable $2, -1, 3, -4$ are parameters

- In machine learning, training a model means finding the parameters that minimize an objective function

Example: gravitational model

- Newton's universal law of gravitation: "every particle in the universe attracts every other particle with a force that is proportional to the masses of both particles and to the distance between their centers"
- Derived from empirical observations by inductive reasoning
- Principles, generalization
- We can make predictions about the behavior of bodies affected by gravity

$$F = G\frac{m_1 m_2}{r^2}$$

m_1, m_2 and r are variables

G is a parameter which depends on the units (kg, m, s)

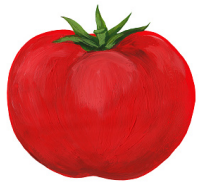
$$G = 6.67 \times 10^{-11}\frac{Nm^2}{kg^2}$$

Example: tomato price model

- How is the amount of tomatoes of a certain variety that I buy related to the price that I pay?
- Assumptions:

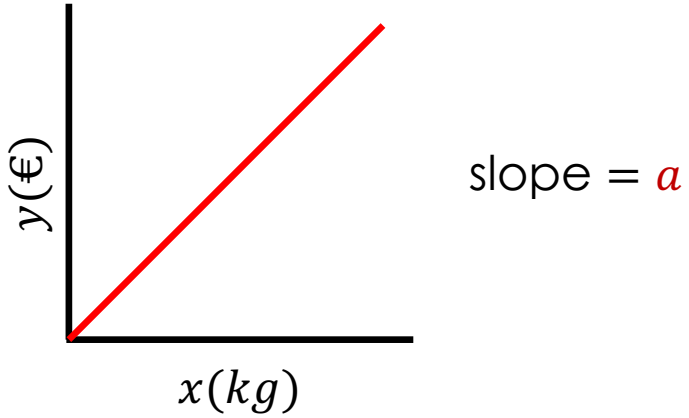
1. Each kg of tomatoes is worth equal
2. There is no flat rate to pay when buying tomatoes

How can we find a? by looking at the price tag



$$y = f(x) = ax$$

E.g. $a = \frac{2€}{kg}$



slope = a

Example: exponential growth

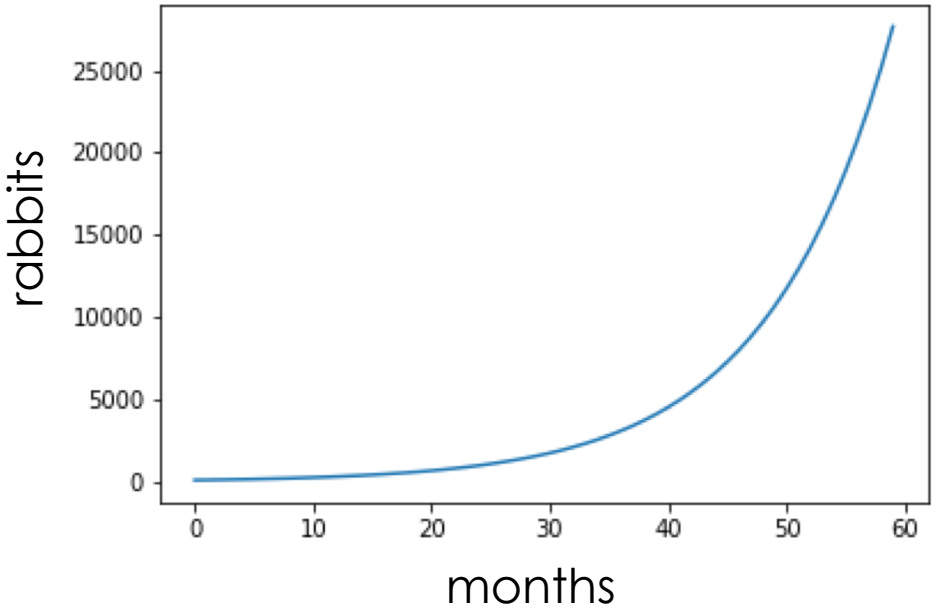
- How does a rabbit population grow with time?
- Suppose we start with a population of 100 rabbits and it increases by 10% each month. The rabbit population population after 5 years is given by

$$P(n) = 100 \cdot (1.1)^{60} = 27680$$

- The general model of exponential growth is

$$P(n) = a(1 + r)^t,$$

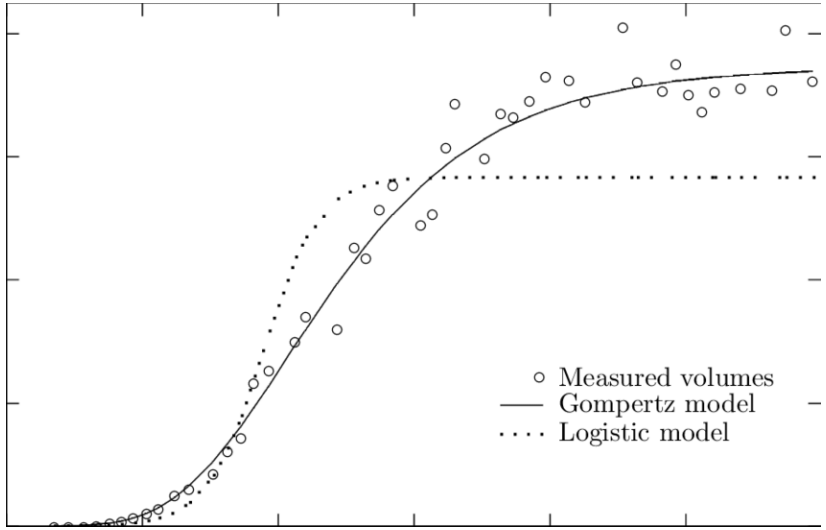
Where a is the initial population, r the growth rate and t the time



Example: logistic growth

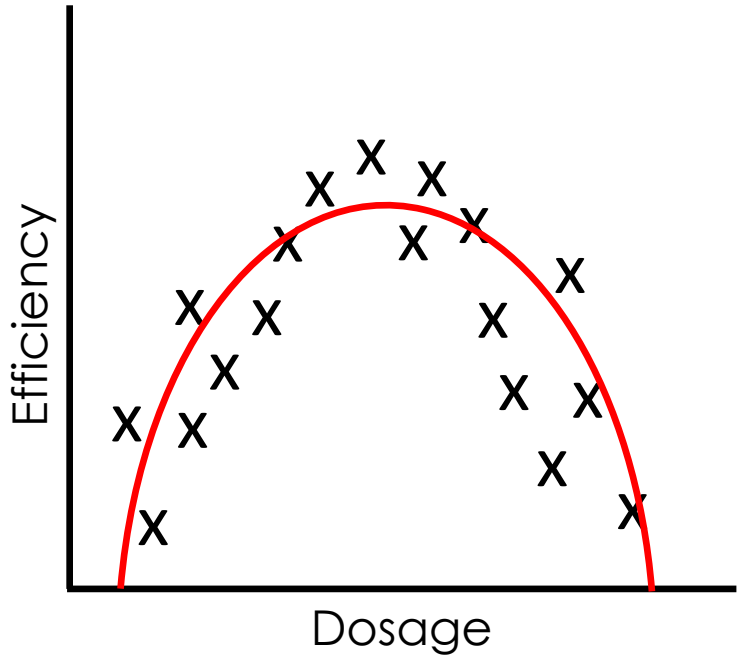
- The population cannot grow indefinitely
- L is the carrying capacity of the system
- k is the logistic growth rate
- x_0 is the sigmoid midpoint
- We find L, k and x_0 by minimizing some error function in our observations

$$f(x) = \frac{L}{1 + e^{-k(x-x_0)}}$$



Example: drug efficiency model

- The effectiveness of a drug may be unnoticeable for small doses, ideal in intermediate doses and harmful in high doses
- This can be modelled by a quadratic function: $E(x) = ax^2 + bx + c$
- Again, a, b and c are the parameters of the model, in this case the coefficients of a quadratic equation, and we find them using the available data

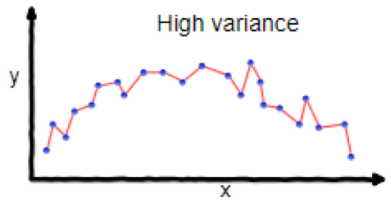


How did we build these models?

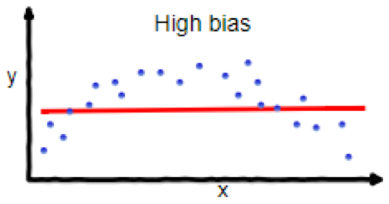
- In the more general sense, our assumption is that there exists a function that maps each input to its correct output
- If we had all the information, we would theoretically be able to correctly predict the output for any arbitrary input
- This is impossible in practice
- We use the training data to infer characteristics about the population
- Goal: approximate the mapping function with the available information

Machine learning

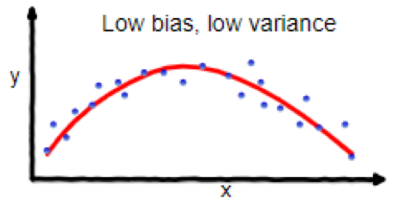
- We have some freedom to choose how we want the function that describes our model to be
- The more degrees of freedom (parameters), the less assumptions we do about the structure of our data



overfitting

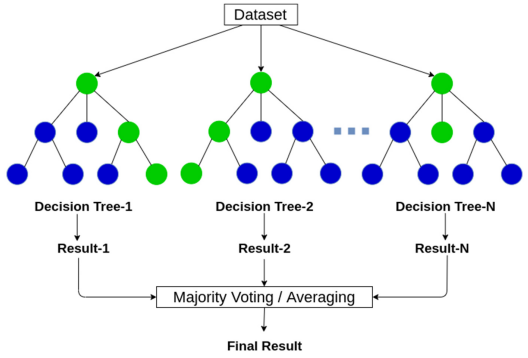


underfitting



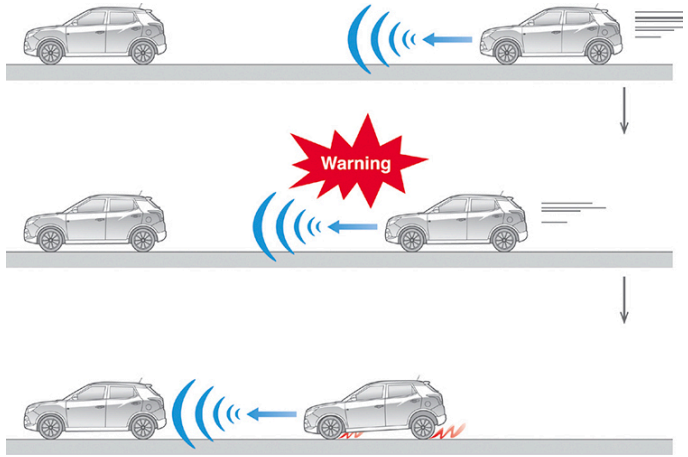
Good balance

Random Forest



Artifical Intelligence

- Intelligence displayed by machines
- ChatGPT: “Intelligence is a complex and multifaceted concept that refers to the ability of an individual or system to acquire and apply knowledge, reason, solve problems, make decisions, adapt to new situations, learn from experience, and exhibit general cognitive skills”
- Natural intelligence: displayed by animals and some humans
- Artificial Intelligence: ability to mimic human actions and behavior

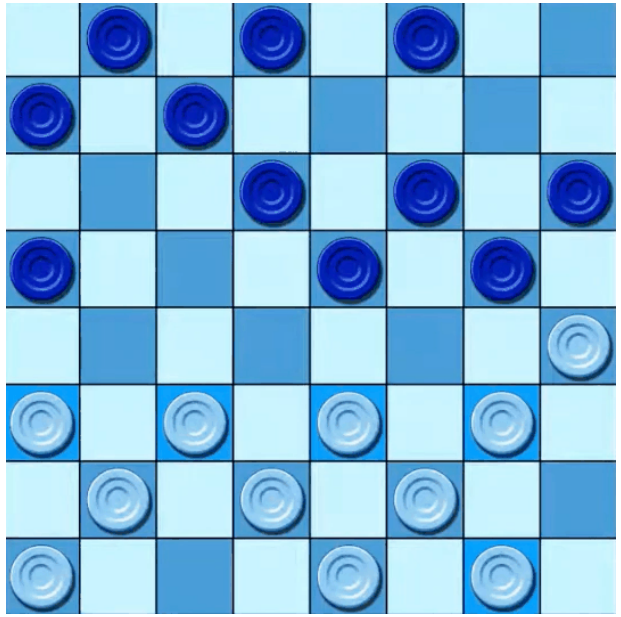


Learning

- The acquisition of knowledge or skills through study, experience, or being taught – *Oxford English Dictionary*
- Memory
- Repetition
- Improvement
- Understanding

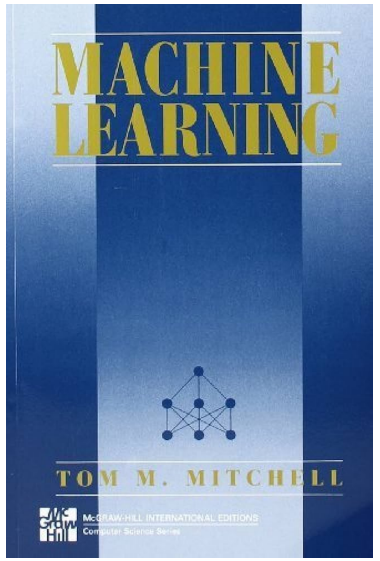
Machine learning

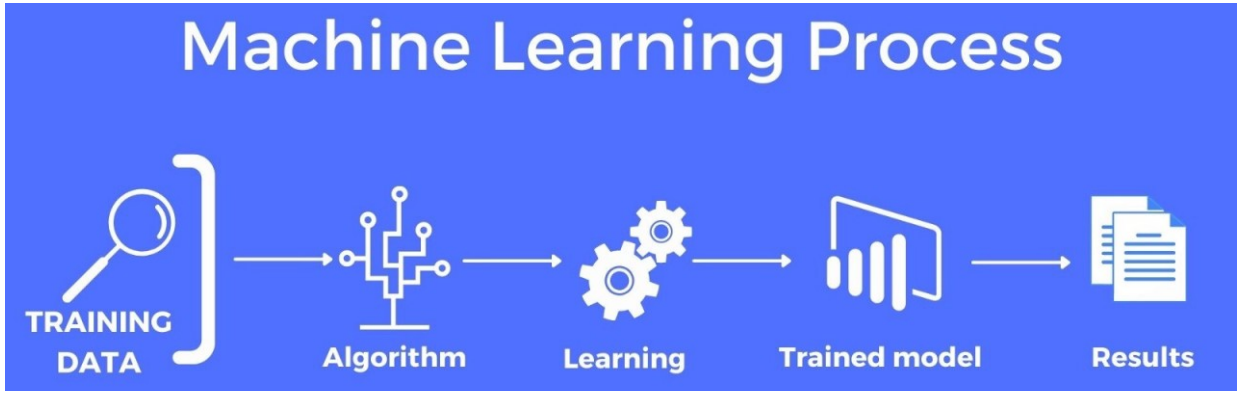
- Arthur Samuel in 1959: Computer that plays checkers that improves with experience
- Mini-max algorithm with alpha-beta pruning
- Rote learning: news inputs were compared with previously stored input-output pairs, amplifying the search depth of the tree
- Machine learning: “Field of study that gives computers the ability to learn without being explicitly programmed”

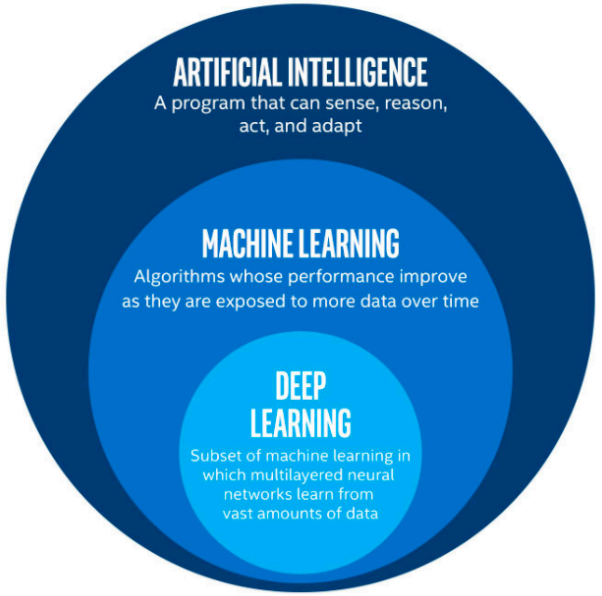


Machine learning

- "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E " – Tom Mitchell 1997



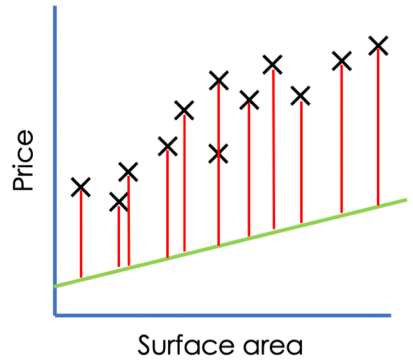


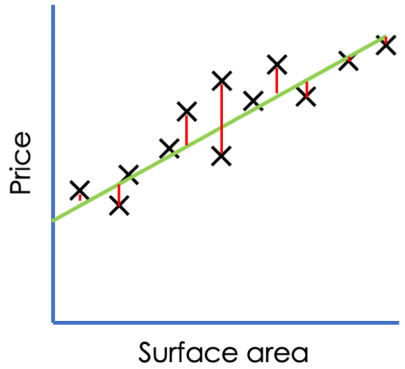


Model training

- Metric that we want to optimize
- Algorithm → procedural method to find the parameters that minimize that metric
- Model → set of parameters & functions that define the transformation from the input data to prediction

$$\hat{y} = ax + b$$





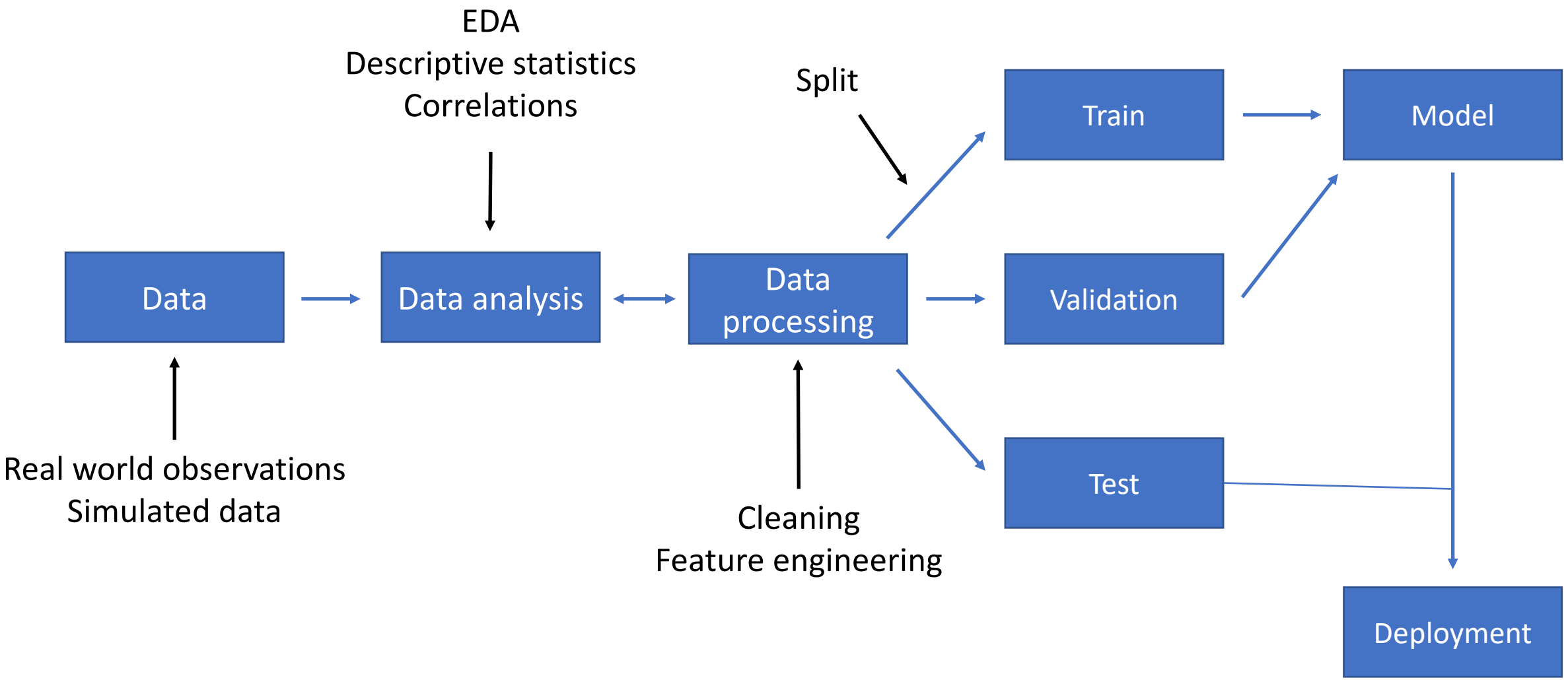
Dog

Supervised learning

UMAP projection

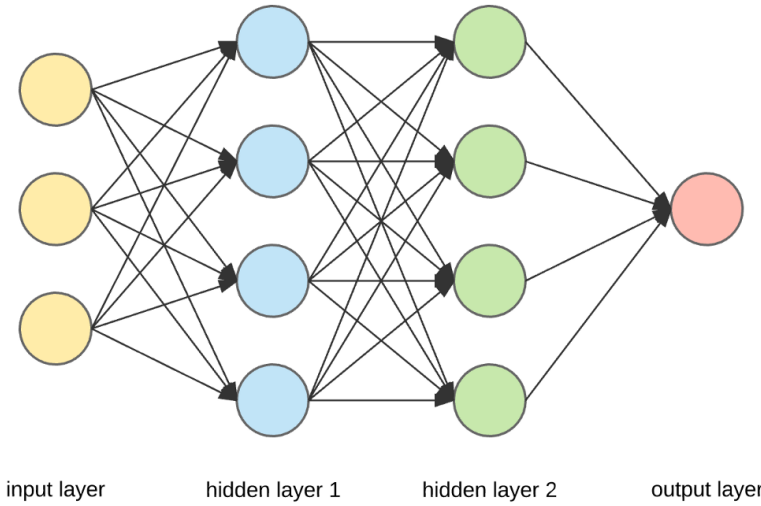
Unsupervised learning

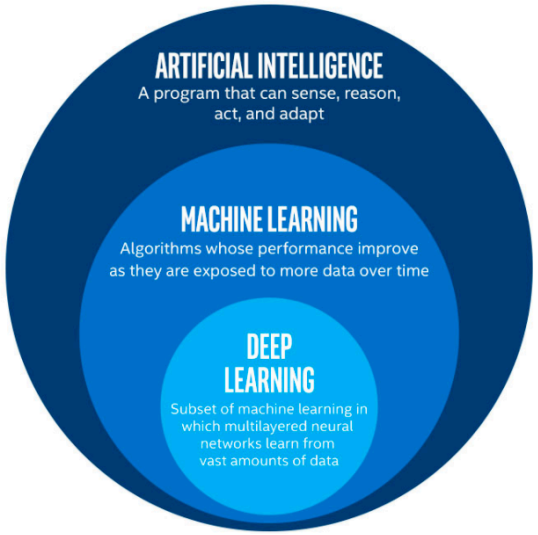
General Machine Learning procedure



Deep Learning

- Sub-field of Machine Learning that uses deep neural networks inspired by the human brain
- Very powerful models when large amounts of data are available





Vladyslav Ostash
National Technical University of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute”, Ukraine

Roman Dzhikirba
University of London

Roderic Guigó Corominas
Harvard University, USA

Guillem Guigo Corominas
Universitat de Girona, Spain
Expert in machine learning

Kateryna Pantiukh
University of Tartu

Nazar Shevchuk
University of Tartu

Summary of the Week

- **Day 1 - Sunday**
  - Data types, Data Transformations, Vectorization
- **Day 2 - Monday**
  - Dimensionality Reduction Algorithms
- **Day 3 - Tuesday**
  - Logistic Regression, Clustering
- **Day 4 - Wednesday**
  - Machine Learning Models
- **Day 5 - Thursday**
  - Linear and Convolutional Neural Networks
- **Day 6 - Friday**
  - Transformers

Format and Tools

Theory / Practical

GitHub

Jupyter Notebooks

NumPy, Pandas, Matplotlib

For Machine Learning: Keras

Google Colab

Welcome to UBDS3!

UBDS3, 2023