

Intro to ML Naïve Bayes

Vladislav Goncharenko
Radoslav Neychev



MSU, spring 2024



Outline

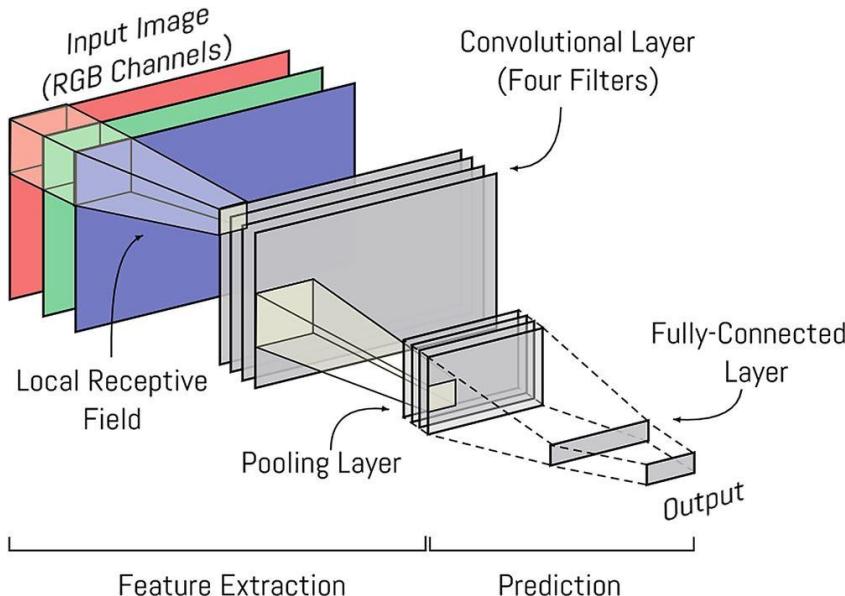
- 
1. ML and AI overview
 2. What is Machine Learning
 3. Thesaurus and notation
 4. Datasets
 5. Exploratory data analysis
 6. Maximum Likelihood Estimation
 7. Some Machine Learning problems
 8. Naïve Bayes classifier
 9. Machine learning libraries

ML and AI overview

girafe
ai

01

Computer Vision



Basics:

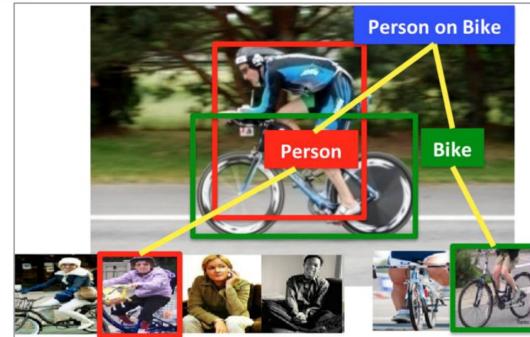
- Classical CV (filters, border detectors)
- Convolutional Neural Networks



Computer Vision

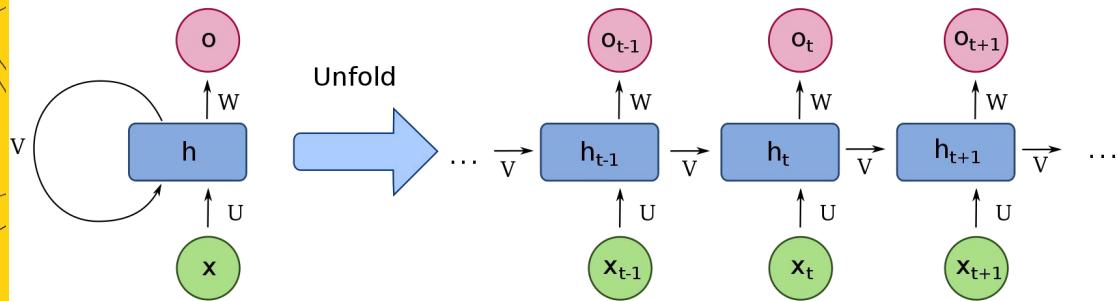
Some achievements:

- Object detection
- Semantic segmentation
- Generative models





Natural Language Processing



Basics:

- Language models
- Recurrent Neural Networks
- Attention module



Natural Language Processing

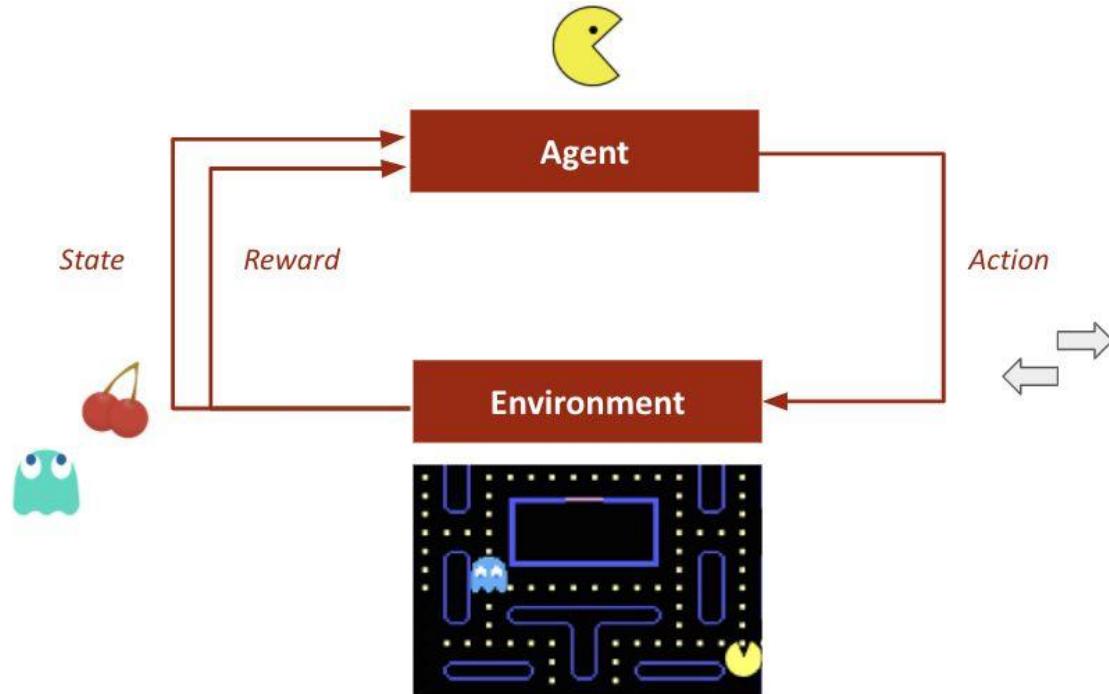
Some achievements:

- Machine translation
- Texts classification
- Texts generation





Reinforcement Learning



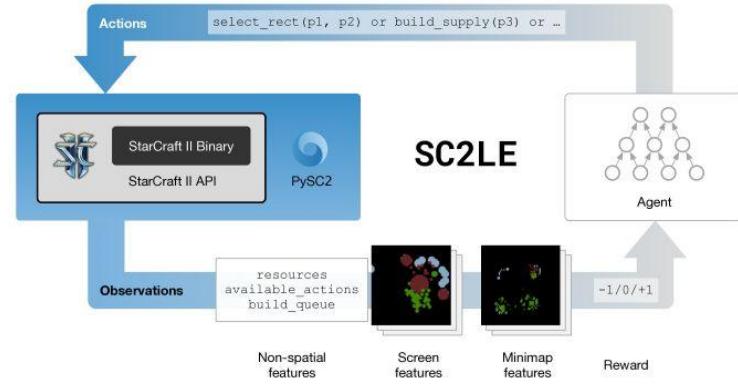
Basics:

- Q-learning
- DQN
- REINFORCE

Reinforcement Learning

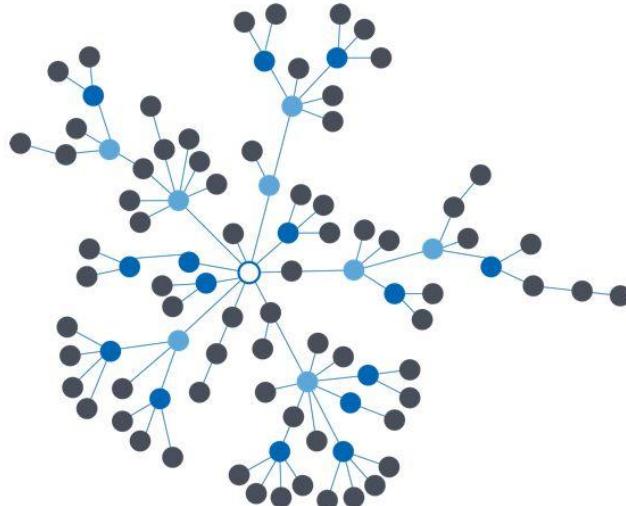
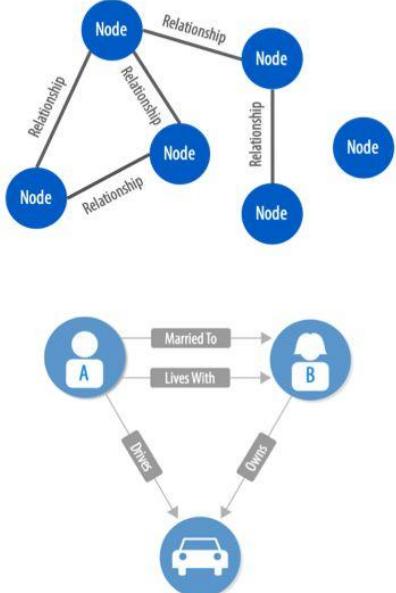
Achievements:

- Alpha Go
- OpenAI Five
- DeepMind Star Craft 2





Machine Learning on Graphs



Basics:

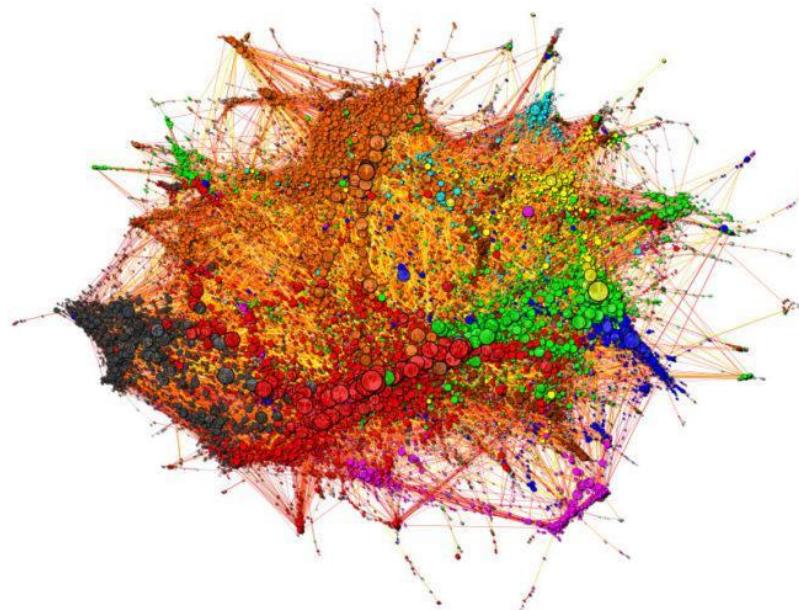
- Random graphs
- Small world model
- Graphs convolutions



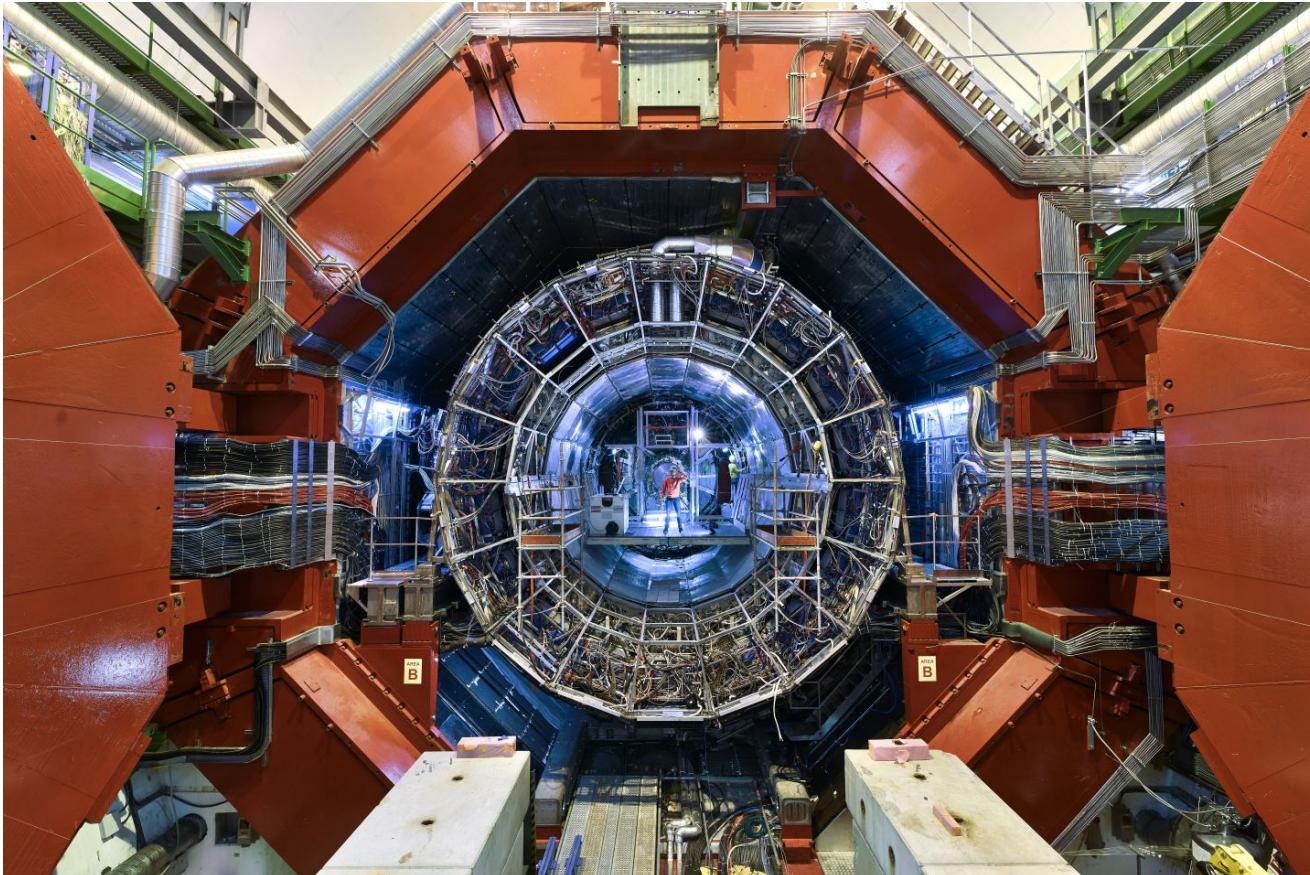
Machine Learning on Graphs

Some achievements:

- Communities detection
- Recommender system

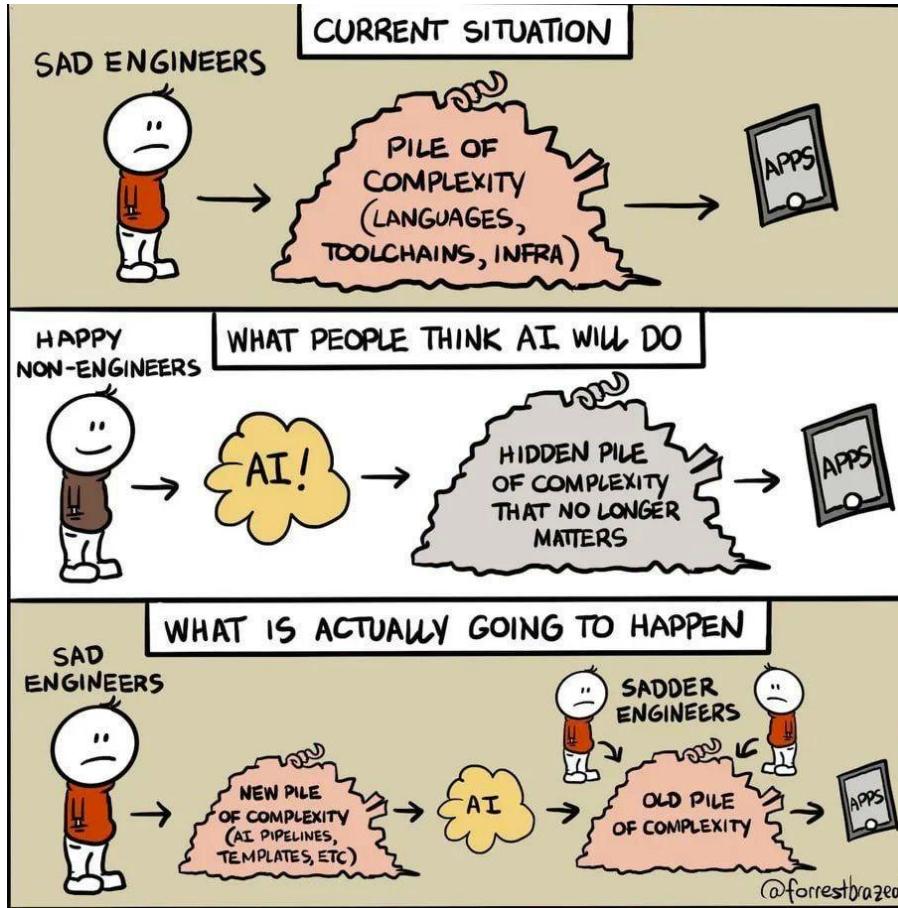


Machine Learning applications





What is AI and ML for you

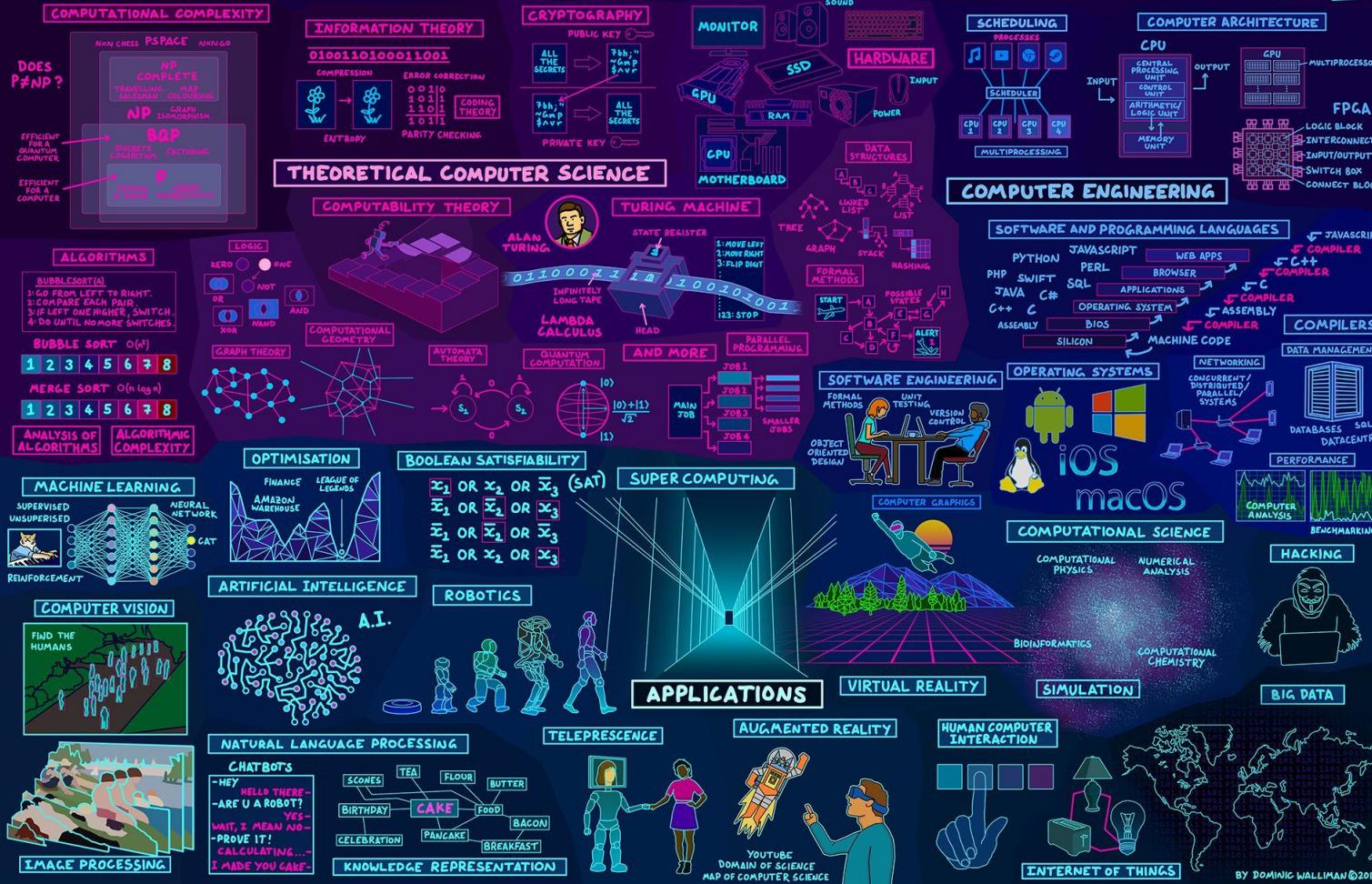


What is Machine Learning?

girafe
ai

02

MAP OF COMPUTER SCIENCE

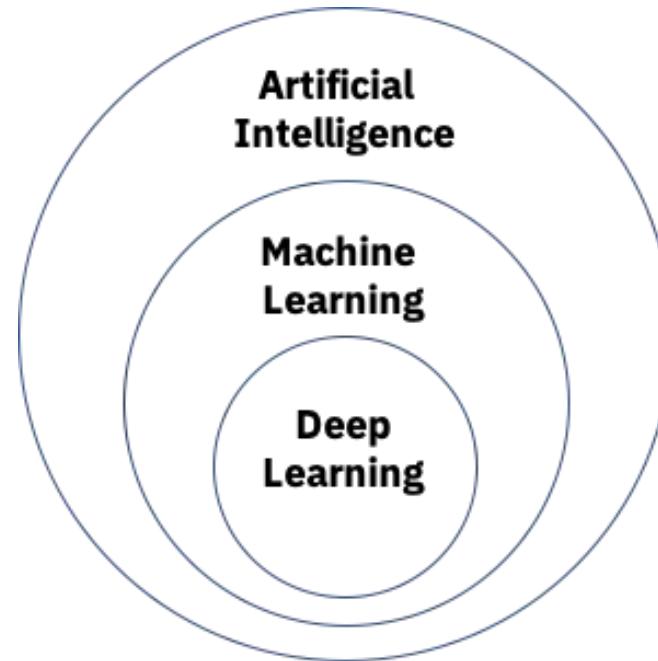


Terms around ML



There are different words that occur in similar contexts, but mean different things:

- Artificial Intelligence
- Machine Learning
- Deep Learning
- Data Science
- Big Data



Artificial Intelligence



Any system that perceives its environment and takes actions that maximize its chances of achieving its goals. [Wikipedia](#)

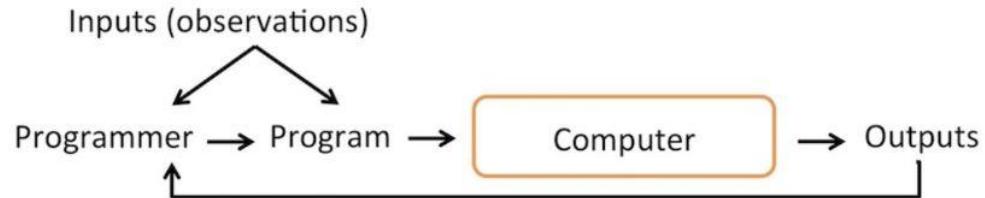


The electric kettle is already part of Skynet!

Machine Learning

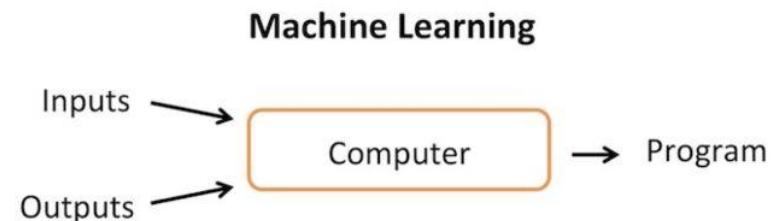


The study of computer algorithms that improve automatically through experience and by the use of data. [Wikipedia](#)



Machine Learning is the field of study that gives computers the ability to learn without being explicitly programmed
– Arthur Samuel (1959)

The most important distinguishing feature of machine learning is the statistical nature of the models



Machine Learning vs Traditional Programming

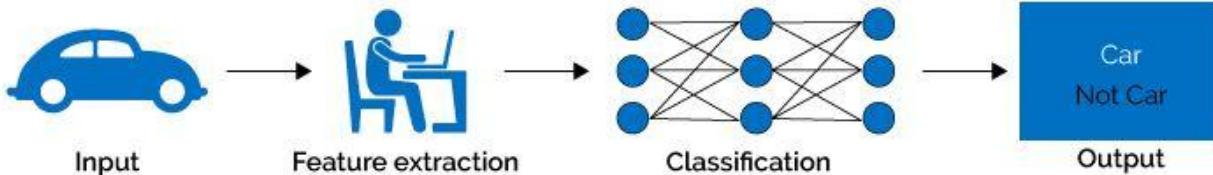


Deep Learning

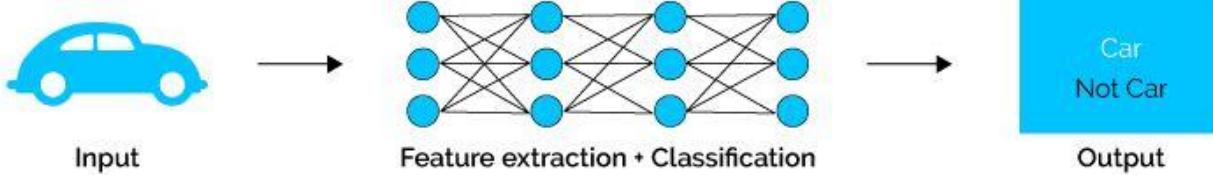
A specific type of models that has shown high results in many applied tasks in the last decade

One of the key success factors is compatibility with modern computing systems (GPUs), simplicity and flexibility of the architecture of a particular model

Machine Learning



Deep Learning



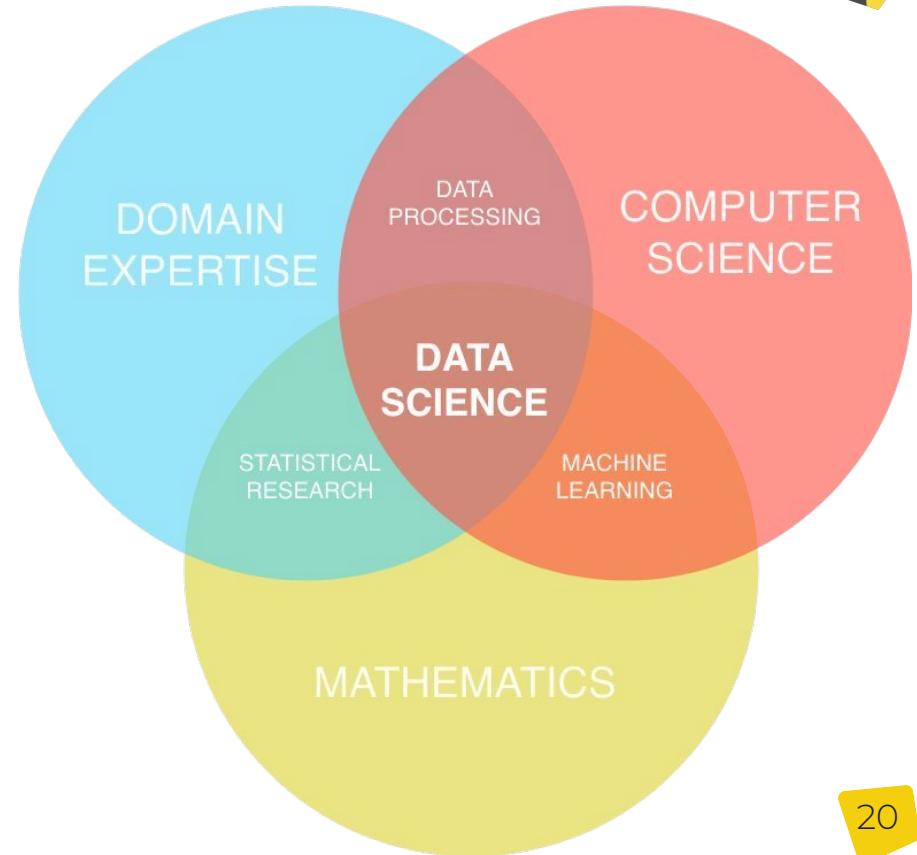
Data Science



An interdisciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from structured and unstructured data, and apply knowledge and actionable insights from data across a broad range of application domains.

[Wikipedia](#)

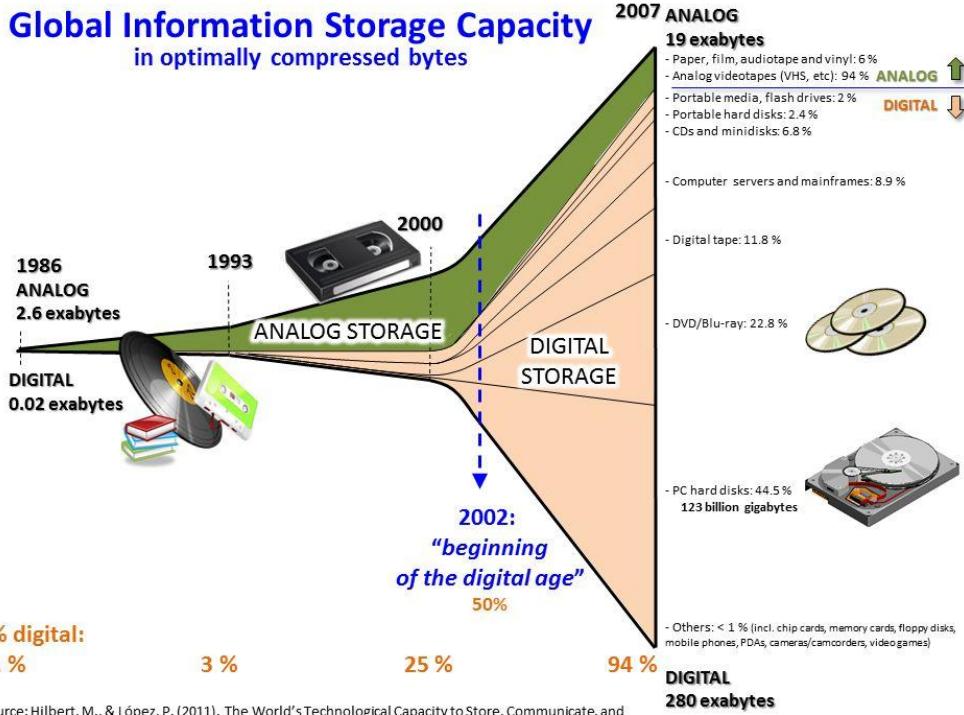
That is, it is the application of mathematical methods to data to extract useful information (so-called insights)



Big Data



Global Information Storage Capacity in optimally compressed bytes

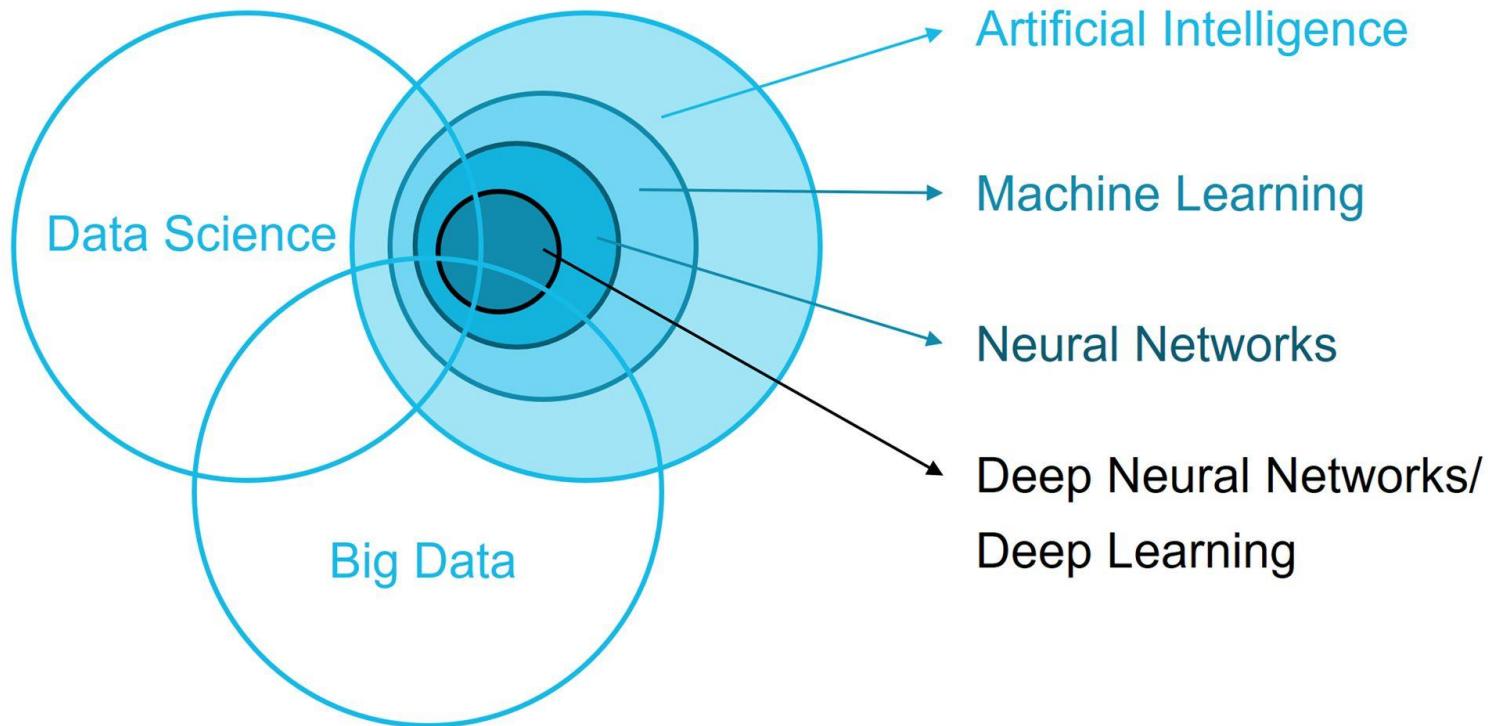


Source: Hilbert, M., & López, P. (2011). The World's Technological Capacity to Store, Communicate, and Compute Information. *Science*, 332(6025), 60–65. <http://www.martinhilbert.net/WorldInfoCapacity.html>

A field that treats ways to analyze, systematically extract information from, or otherwise deal with data sets that are too large or complex to be dealt with by traditional data-processing application software. [Wikipedia](#)

The main distinguishing feature is the processing of a hitherto unprecedented amount of information and the development of methods for the effective processing of such data for a variety of purposes

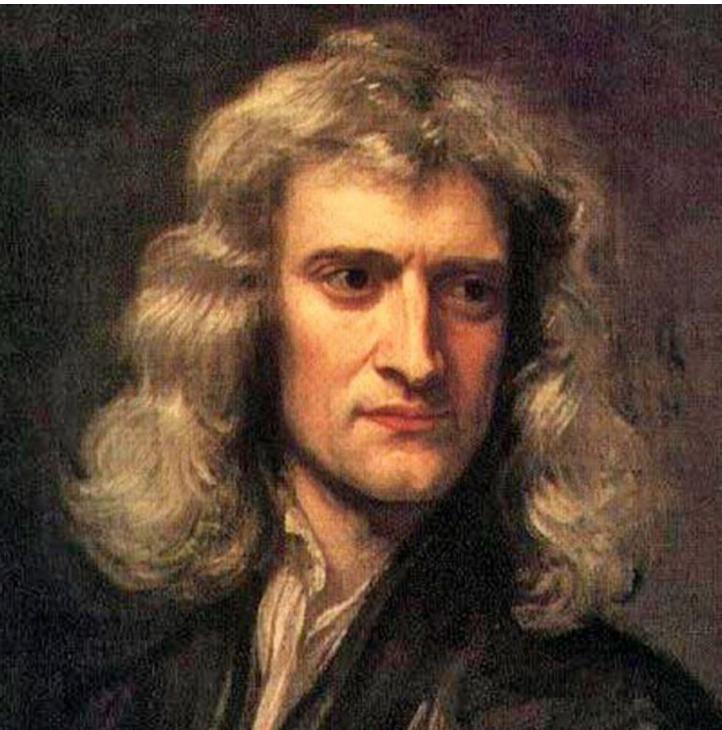
Relation of terms and spheres



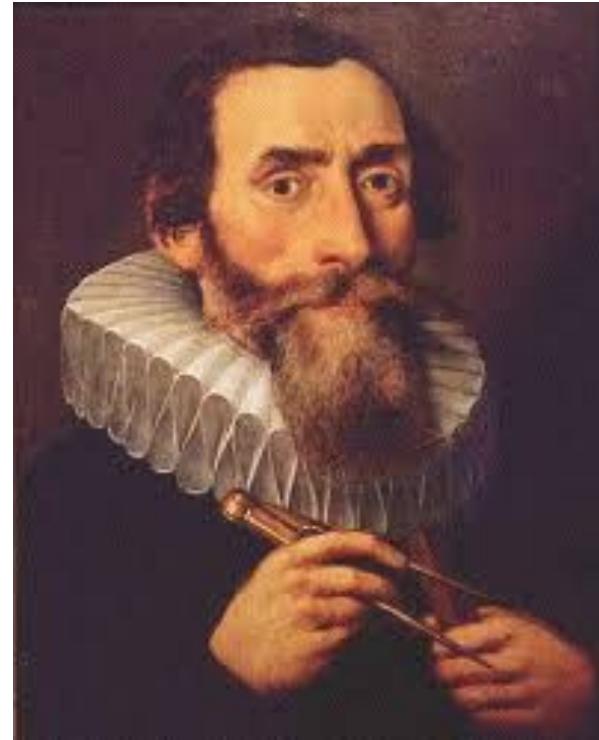


Data → Knowledge

Long before the ML

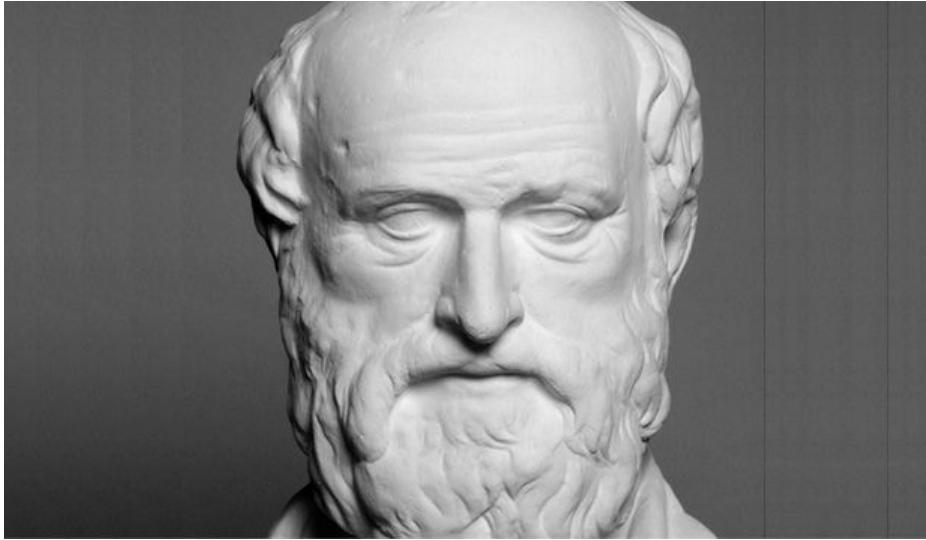


Isaac Newton



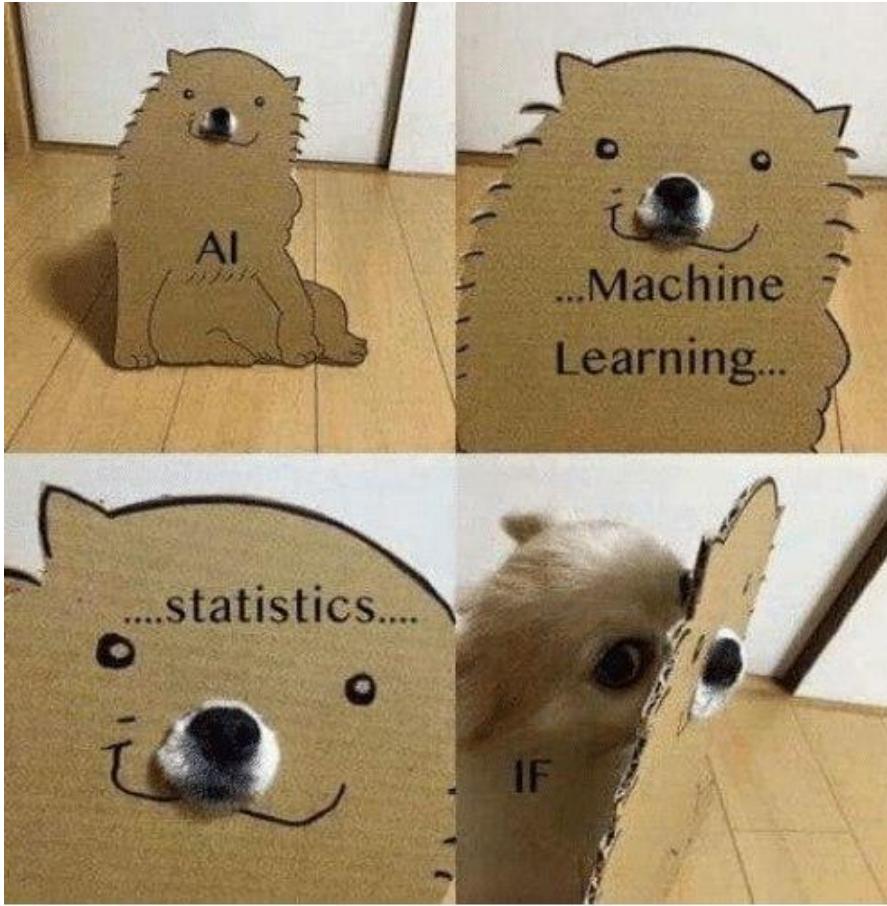
Johannes Kepler

Long before the ML



Eratosthenes

What it is actually



ML thesaurus

girafe
ai

03



ML thesaurus

Denote the **dataset**.

Name	Age	Statistics (mark)	Python (mark)	Eye color	Native language	Target (mark)	Target (passed)
John	22	5	4	Brown	English	5	TRUE
Aahna	17	4	5	Brown	Hindi	4	TRUE
Emily	25	5	5	Blue	Chinese	5	TRUE
Michael	27	3	4	Green	French	5	TRUE
Some student	23	3	3	NA	Esperanto	2	FALSE



ML thesaurus

Observation (or datum, or data point) is one piece of information.

Name	Age	Statistics (mark)	Python (mark)	Eye color	Native language	Target (mark)	Target (passed)
John	22	5	4	Brown	English	5	TRUE
Aahna	17	4	5	Brown	Hindi	4	TRUE
Emily	25	5	5	Blue	Chinese	5	TRUE
Michael	27	3	4	Green	French	5	TRUE
Some student	23	3	3	NA	Esperanto	2	FALSE

In many cases the **observations** are supposed to be ***i.i.d.***

- ***independent***
- ***identically distributed***



ML thesaurus

Feature (or predictor) represents some special property.

Name	Age	Statistics (mark)	Python (mark)		Eye color	Native language	Target (mark)	Target (passed)
John	22	5	4	Brown	English		5	TRUE
Aahna	17	4	5	Brown	Hindi		4	TRUE
Emily	25	5	5	Blue	Chinese		5	TRUE
Michael	27	3	4	Green	French		5	TRUE
Some student	23	3	3	NA	Esperanto		2	FALSE



ML thesaurus

These all are features

Name	Age	Statistics (mark)	Python (mark)	Eye color	Native language	Target (mark)	Target (passed)
John	22	5	4	Brown	English	5	TRUE
Aahna	17	4	5	Brown	Hindi	4	TRUE
Emily	25	5	5	Blue	Chinese	5	TRUE
Michael	27	3	4	Green	French	5	TRUE
Some student	23	3	3	NA	Esperanto	2	FALSE



ML thesaurus

These all are features

Name	Age	Statistics (mark)	Python (mark)	Eye color	Native language	Target (mark)	Target (passed)
John	22	5	4	Brown	English	5	TRUE
Aahna	17	4	5	Brown	Hindi	4	TRUE
Emily	25	5	5	Blue	Chinese	5	TRUE
Michael	27	3	4	Green	French	5	TRUE
Some student	23	3	3	NA	Esperanto	2	FALSE



ML thesaurus

These all are features

Name	Age	Statistics (mark)	Python (mark)	Eye color	Native language	Target (mark)	Target (passed)
John	22	5	4	Brown	English	5	TRUE
Aahna	17	4	5	Brown	Hindi	4	TRUE
Emily	25	5	5	Blue	Chinese	5	TRUE
Michael	27	3	4	Green	French	5	TRUE
Some student	23	3	3	NA	Esperanto	2	FALSE



ML thesaurus

These all are features

Name	Age	Statistics (mark)	Python (mark)	Eye color	Native language	Target (mark)	Target (passed)
John	22	5	4	Brown	English	5	TRUE
Aahna	17	4	5	Brown	Hindi	4	TRUE
Emily	25	5	5	Blue	Chinese	5	TRUE
Michael	27	3	4	Green	French	5	TRUE
Some student	23	3	3	NA	Esperanto	2	FALSE



ML thesaurus

And even the name is a **feature**

Name	Age	Statistics (mark)	Python (mark)	Eye color	Native language	Target (mark)	Target (passed)
John	22	5	4	Brown	English	5	TRUE
Aahna	17	4	5	Brown	Hindi	4	TRUE
Emily	25	5	5	Blue	Chinese	5	TRUE
Michael	27	3	4	Green	French	5	TRUE
Some student	23	3	3	NA	Esperanto	2	FALSE



ML thesaurus

The **feature matrix or design matrix** contains all the observations and their features

Name	Age	Statistics (mark)	Python (mark)	Eye color	Native language	Target (mark)	Target (passed)
John	22	5	4	Brown	English	5	TRUE
Aahna	17	4	5	Brown	Hindi	4	TRUE
Emily	25	5	5	Blue	Chinese	5	TRUE
Michael	27	3	4	Green	French	5	TRUE
Some student	23	3	3	NA	Esperanto	2	FALSE

Features can even be multidimensional, we will discuss it later in this course

Matrix notation: features



Name	Age	Statistics (mark)	Python (mark)	Eye color	Native language	Target (mark)	Target (passed)
John	22	5	4	Brown	English	5	TRUE
Aahna	17	4	5	Brown	Hindi	4	TRUE
Emily	25	5	5	Blue	Chinese	5	TRUE
Michael	27	3	4	Green	French	5	TRUE
Some student	23	3	3	NA	Esperanto	2	FALSE

Feature matrix is usually denoted as $X \in R^{n \times p}$

where n is number of objects in dataset and p is number of properties



ML thesaurus

Target represents the information we are interested in.

Name	Age	Statistics (mark)	Python (mark)	Eye color	Native language	Target (mark)	Target (passed)
John	22	5	4	Brown	English	5	TRUE
Aahna	17	4	5	Brown	Hindi	4	TRUE
Emily	25	5	5	Blue	Chinese	5	TRUE
Michael	27	3	4	Green	French	5	TRUE
Some student	23	3	3	NA	Esperanto	2	FALSE

Target can be either a **number** (real, integer, etc.) – for **regression** problem



ML thesaurus

Target represents the information we are interested in.

Name	Age	Statistics (mark)	Python (mark)	Eye color	Native language	Target (mark)	Target (passed)
John	22	5	4	Brown	English	5	TRUE
Aahna	17	4	5	Brown	Hindi	4	TRUE
Emily	25	5	5	Blue	Chinese	5	TRUE
Michael	27	3	4	Green	French	5	TRUE
Some student	23	3	3	NA	Esperanto	2	FALSE

Or a **label** – for **classification** problem



ML thesaurus

Target represents the information we are interested in.

Name	Age	Statistics (mark)	Python (mark)	Eye color	Native language	Target (mark)	Target (passed)
John	22	5	4	Brown	English	5	TRUE
Aahna	17	4	5	Brown	Hindi	4	TRUE
Emily	25	5	5	Blue	Chinese	5	TRUE
Michael	27	3	4	Green	French	5	TRUE
Some student	23	3	3	NA	Esperanto	2	FALSE

Mark can be treated as a label too (due to finite number of labels: 1 to 5)



ML thesaurus

Further we will work with the numerical target (mark)

Name	Age	Statistics (mark)	Python (mark)	Eye color	Native language	Target (mark)
John	22	5	4	Brown	English	5
Aahna	17	4	5	Brown	Hindi	4
Emily	25	5	5	Blue	Chinese	5
Michael	27	3	4	Green	French	5
Some student	23	3	3	NA	Esperanto	2



ML thesaurus

Target represents the information we are interested in.

Name	Age	Statistics (mark)	Python (mark)	Eye color	Native language	Target (mark)	Target (passed)
John	22	5	4	Brown	English	5	TRUE
Aahna	17	4	5	Brown	Hindi	4	TRUE
Emily	25	5	5	Blue	Chinese	5	TRUE
Michael	27	3	4	Green	French	5	TRUE
Some student	23	3	3	NA	Esperanto	2	FALSE

Target can be either a **number** (real, integer, etc.) – for **regression** problem

Matrix notation: target



Name	Age	Statistics (mark)	Python (mark)	Eye color	Native language	Target (mark)
John	22	5	4	Brown	English	5
Aahna	17	4	5	Brown	Hindi	4
Emily	25	5	5	Blue	Chinese	5
Michael	27	3	4	Green	French	5
Some student	23	3	3	NA	Esperanto	2

Target matrix is usually denoted as $Y \in R^n$

where n is number of objects in dataset



ML thesaurus

The ***prediction*** contains values we predicted using some ***model***.

Name	Age	Statistics (mark)	Python (mark)	Eye color	Native language	Target (mark)	Predicted (mark)
John	22	5	4	Brown	English	5	4.5
Aahna	17	4	5	Brown	Hindi	4	4.5
Emily	25	5	5	Blue	Chinese	5	5
Michael	27	3	4	Green	French	5	3.5
Some student	23	3	3	NA	Esperanto	2	3

One could notice that prediction just averages of Statistics and Python marks. So our ***model*** can be represented as follows:

$$\hat{\text{mark}}_{ML} = \frac{1}{2}\text{mark}_{Statistics} + \frac{1}{2}\text{mark}_{Python}$$



ML thesaurus

The ***prediction*** contains values we predicted using some ***model***.

Name	Age	Statistics (mark)	Python (mark)	Eye color	Native language	Target (mark)	Predicted (mark)
John	22	5	4	Brown	English	5	4.5
Aahna	17	4	5	Brown	Hindi	4	4.5
Emily	25	5	5	Blue	Chinese	5	5
Michael	27	3	4	Green	French	5	3.5
Some student	23	3	3	NA	Esperanto	2	3

Different models can provide different predictions:

$$\hat{\text{mark}}_{ML} = \frac{1}{2}\text{mark}_{Statistics} + \frac{1}{2}\text{mark}_{Python}$$



ML thesaurus

The ***prediction*** contains values we predicted using some ***model***.

Name	Age	Statistics (mark)	Python (mark)	Eye color	Native language	Target (mark)	Predicted (mark)
John	22	5	4	Brown	English	5	1
Aahna	17	4	5	Brown	Hindi	4	5
Emily	25	5	5	Blue	Chinese	5	2
Michael	27	3	4	Green	French	5	4
Some student	23	3	3	NA	Esperanto	2	3

Different models can provide different predictions:

$$\hat{\text{mark}}_{ML} = \text{random}(\text{integer from } [1; 5])$$



ML thesaurus

The ***prediction*** contains values we predicted using some ***model***.

Name	Age	Statistics (mark)	Python (mark)	Eye color	Native language	Target (mark)	Predicted (mark)
John	22	5	4	Brown	English	5	1
Aahna	17	4	5	Brown	Hindi	4	5
Emily	25	5	5	Blue	Chinese	5	2
Michael	27	3	4	Green	French	5	4
Some student	23	3	3	NA	Esperanto	2	3

Different models can provide different predictions.

Usually some ***hypothesis*** lies beneath the model choice.



ML thesaurus

Loss function measures the error rate of our model.

Square deviation	Target (mark)	Predicted (mark)
16	5	1
1	4	5
9	5	2
1	5	4
1	2	3

- **Mean Squared Error** (where \mathbf{y} is vector of targets):

$$MSE(\mathbf{y}, \hat{\mathbf{y}}) = \frac{1}{N} \|\mathbf{y} - \hat{\mathbf{y}}\|_2^2 = \frac{1}{N} \sum_i (y_i - \hat{y}_i)^2$$

ML thesaurus



Loss function measures the error rate of our model.

Absolute deviation	Target (mark)	Predicted (mark)
4	5	1
1	4	5
3	5	2
1	5	4
1	2	3

- **Mean Absolute Error** (where \mathbf{y} is vector of targets):

$$MAE(\mathbf{y}, \hat{\mathbf{y}}) = \frac{1}{N} \|\mathbf{y} - \hat{\mathbf{y}}\|_1 = \frac{1}{N} \sum_i |y_i - \hat{y}_i|$$



ML thesaurus

To learn something, our **model** needs some degrees of freedom:

Name	Age	Statistics (mark)	Python (mark)	Eye color	Native language	Target (mark)	Predicted (mark)
John	22	5	4	Brown	English	5	4.5
Aahna	17	4	5	Brown	Hindi	4	4.5
Emily	25	5	5	Blue	Chinese	5	5
Michael	27	3	4	Green	French	5	3.5
Some student	23	3	3	NA	Esperanto	2	3

$$\hat{\text{mark}}_{ML} = w_1 \cdot \text{mark}_{Statistics} + w_2 \cdot \text{mark}_{Python}$$



ML thesaurus

To learn something, our **model** needs some degrees of freedom:

Name	Age	Statistics (mark)	Python (mark)	Eye color	Native language	Target (mark)	Predicted (mark)
John	22	5	4	Brown	English	5	4.447
Aahna	17	4	5	Brown	Hindi	4	4.734
Emily	25	5	5	Blue	Chinese	5	5.101
Michael	27	3	4	Green	French	5	3.714
Some student	23	3	3	NA	Esperanto	2	3.060

$$\hat{\text{mark}}_{ML} = w_1 \cdot \text{mark}_{Statistics} + w_2 \cdot \text{mark}_{Python}$$



ML thesaurus

To learn something, our **model** needs some degrees of freedom:

Name	Age	Statistics (mark)	Python (mark)	Eye color	Native language	Target (mark)	Predicted (mark)
John	22	5	4	Brown	English	5	1
Aahna	17	4	5	Brown	Hindi	4	5
Emily	25	5	5	Blue	Chinese	5	2
Michael	27	3	4	Green	French	5	4
Some student	23	3	3	NA	Esperanto	2	3

$$\hat{\text{mark}}_{ML} = \text{random}(\text{integer from } [1; 5])$$

ML thesaurus



Last term we should learn for now is **hyperparameter**.

Hyperparameter should be fixed before our model starts to work with the data.

We will discuss it later with kNN as an example.



Main concepts

- Dataset (both X and Y)
 - Design matrix (only X)
 - Feature matrix (only X)
- Observation
 - object
 - datum
 - row
- Feature
 - column
- Target
 - Label
- Model
- Prediction

Datasets and where to find them

girafe
ai

04



Datasets search

Nowadays there are tons of data available on the Internet.

It covers most of the cases you can think of.

So the main problem is to search for the right one!

Let's overview some ways to collect datasets.

Google dataset search



<https://datasetsearch.research.google.com/>

Contains main info about dataset

and links to sources.

Not all datasets are easily available

The screenshot shows the Google Dataset Search interface. At the top, there's a search bar with the query "cancer cell segmentation". Below the search bar are several filter buttons: "Last updated", "Download format", "Croissant", "Usage rights", "Topic", "Provider", and "Free". The main results area displays "100+ data sets found". The first result is a dataset from Zenodo titled "Cell Colony Image Segmentation Dataset 1 for T-47D Breast Cancer Cells". It includes links to zenodo.org and explore.openaire.eu, and download options for txt and zip files. The second result is "CCAgT: Images of Cervical Cells with AgNOR Stain Technique" from data.mendeley.com, also with download links for txt and zip files. The third result is "Segmentation of organelles in isotropic electron microscopy..." from janelia.figshare.com, with a download link for bin files. The results are presented in a clean, modern style with clear navigation and filtering options.

Kaggle

<https://www.kaggle.com/datasets>



Datasets

Explore, analyze, and share quality data. [Learn more](#) about data types, creating, and collaborating.

+ New Dataset

Your Work

Search datasets

All datasets Computer Science Education Classification Computer Vision NLP Data Visualization Pre-Trained Model

Trending Datasets

SI - 100	Moderate
101 - 150	Unhealthy for S
151 - 200	Unhealthy

2023 Air Quality Data for CBSAs

Nikki Perry · Updated 20 hours ago
Usability 10.0 · 166 kB

▲ 16



Smite Item Statistics Data

Matt OP · Updated a day ago
Usability 10.0 · 15 kB
1 File (CSV)

▲ 16



Machine Learning Engineer Salary in 2024

Chopper53 · Updated 21 hours ago
Usability 10.0 · 110 kB
1 File (CSV)

▲ 8



FuelConsumption

Krupa Dharamshi · Updated a month ago
Usability 10.0 · 6 kB
1 File (CSV)

▲ 15

Hugging Face



<https://huggingface.co/datasets>

 Hugging Face

Models Datasets Spaces Posts Docs Solutions Pricing Log In Sign Up

Main Tasks Libraries Languages Licenses Other

Modalities

3D Audio Geospatial Image
Tabular Text Time-series Video

Size (rows)

<1K >1T

Format

json csv parquet imagefolder
soundfolder webdataset text arrow

Datasets 223,270 Filter by name Full-text search Sort: Trending

google/frames-benchmark
Viewer Updated 1 day ago 824 840 126

fka/awesome-chatgpt-prompts
Viewer Updated Sep 4 170 8.42k 5.83k

FBK-MT/mosel
Viewer Updated 6 days ago 51.1M 120 47

HackerNoon/where-startups-trend
Preview Updated 8 days ago 70 38

glaiveai/reflection-v1
Viewer Updated 19 days ago 60.1k 428 39

KingNish/reasoning-base-20k
Viewer Updated 3 days ago 19.9k 178 35

k-mktr/improved-flux-prompts-photoreal-portrait
Viewer Updated 5 days ago 20k 131 65

openai/MMMLU
Viewer Updated 5 days ago 393k 6.21k 377

nvidia/OpenMathInstruct-2
Viewer Updated about 17 hours ago 22M 173 29

lmms-lab/LLaVA-Video-178K
Viewer Updated 2 days ago 1.63M 146 34

argilla/FinePersonas-v0.1
Viewer Updated 20 days ago 21.1M 399 306

5CD-AI/Viet-LAION-Gemini-VQA
Viewer Updated 5 days ago 844k 149 25

migtissera/Synthia-v1.5-I
Viewer

40umov/dostoevsky

Papers with code



<https://paperswithcode.com/datasets>



9761 dataset results

Search for datasets ×

Best match ▼

Filter by Modality

Images	2661
Texts	2546
Videos	862
Audio	394
Medical	336
3D	303

CIFAR-10 (Canadian Institute for Advanced Research, 10 classes)
The CIFAR-10 dataset (Canadian Institute for Advanced Research, 10 classes) is a subset of the Tiny Images dataset and consists of 60000 32x32 color images. The images are labelled...
14,073 PAPERS • 98 BENCHMARKS

ImageNet
The ImageNet dataset contains 14,197,122 annotated images according to the WordNet hierarchy. Since 2010 the dataset is used in the ImageNet Large Scale Visual Recognition...
13,418 PAPERS • 40 BENCHMARKS

MS COCO (Microsoft Common Objects in Context)
The MS COCO (Microsoft Common Objects in Context) dataset is a large-scale object detection, segmentation, key-point detection, and captioning dataset. The dataset consists of...
10,128 PAPERS • 92 BENCHMARKS

UCI ML repository



<https://archive.ics.uci.edu/>

Historical source of datasets.

Most of them collected in 90s

and 00s.

Welcome to the UC Irvine Machine Learning Repository

We currently maintain 664 datasets as a service to the machine learning community. Here, you can donate and find datasets used by millions of people all around the world!

[VIEW DATASETS](#) [CONTRIBUTE A DATASET](#)

Popular Datasets

- Iris**
A small classic dataset from Fisher, 1936. One of the earliest known datasets used for...
[Classification](#) 150 Instances 4 Features
- Heart Disease**
4 databases: Cleveland, Hungary, Switzerland, and the VA Long Beach
[Classification](#) 303 Instances 13 Features
- Dry Bean**
Images of 13,611 grains of 7 different registered dry beans were taken with a high-res...
[Classification](#) 13.61K Instances 16 Features
- Rice (Cammao and Osmancik)**
A total of 3810 rice grain's images were taken for the two species, processed and feat...
[Classification](#) 3.81K Instances 7 Features

New Datasets

- PhiUSIIL Phishing URL (Website)**
PhiUSIIL Phishing URL Dataset is a substantial dataset comprising 134,850 legitimate ...
[Classification](#) 135.8K Instances 54 Features
- RT-IoT2022**
The RT-IoT2022, a proprietary dataset derived from a real-time IoT infrastructure, is in...
[Classification, Regres...](#) 123.12K Instances 84 Features
- Regensburg Pediatric Appendicitis**
This repository holds the data from a cohort of pediatric patients with suspected app...
[Classification](#) 782 Instances 59 Features
- National Poll on Healthy Aging (NPHA)**
This is a subset of the NPHA dataset filtered down to develop and validate machine le...
[Classification](#) 714 Instances 15 Features

Standardization initiative



Croissant dataset standard

<https://mlcommons.org/working-groups/data/croissant/>

<https://research.google/blog/croissant-a-metadata-format-for-ml-ready-datasets/>

The collage includes:

- A screenshot of the Hugging Face Datasets interface showing a dataset card for "titanic-survival".
- A screenshot of the Kaggle website showing a dataset card for "titanic_1".
- A screenshot of the OpenML platform showing a dataset card for "titanic".
- A screenshot of a Google search results page for "titanic" where the "Croissant" download format is highlighted.
- A screenshot of the Croissant Editor interface, titled "Create", showing a form to build a dataset named "Titanic".
- A screenshot of the Croissant Editor interface, titled "Load", showing a preview of a croissant icon.
- A code snippet in a terminal window demonstrating the use of TensorFlow's `tfds` library to build a Croissant dataset from the original Titanic dataset.

Exploratory data analysis (EDA)

girafe
ai

05

Dataset analysis

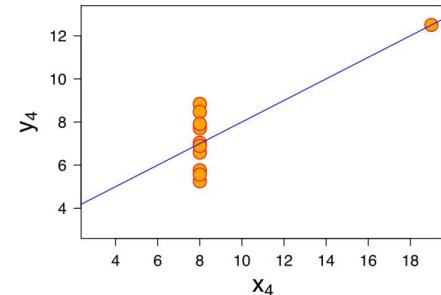
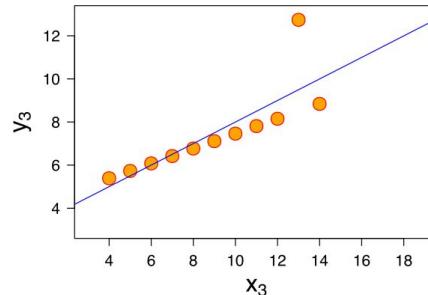
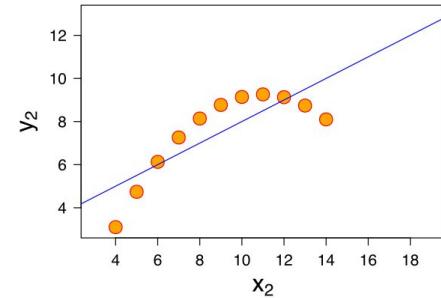
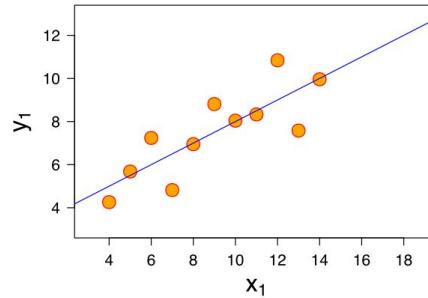


Before you build any model it's important to analyse data you have with various methods because all the models depend on data and its structure.

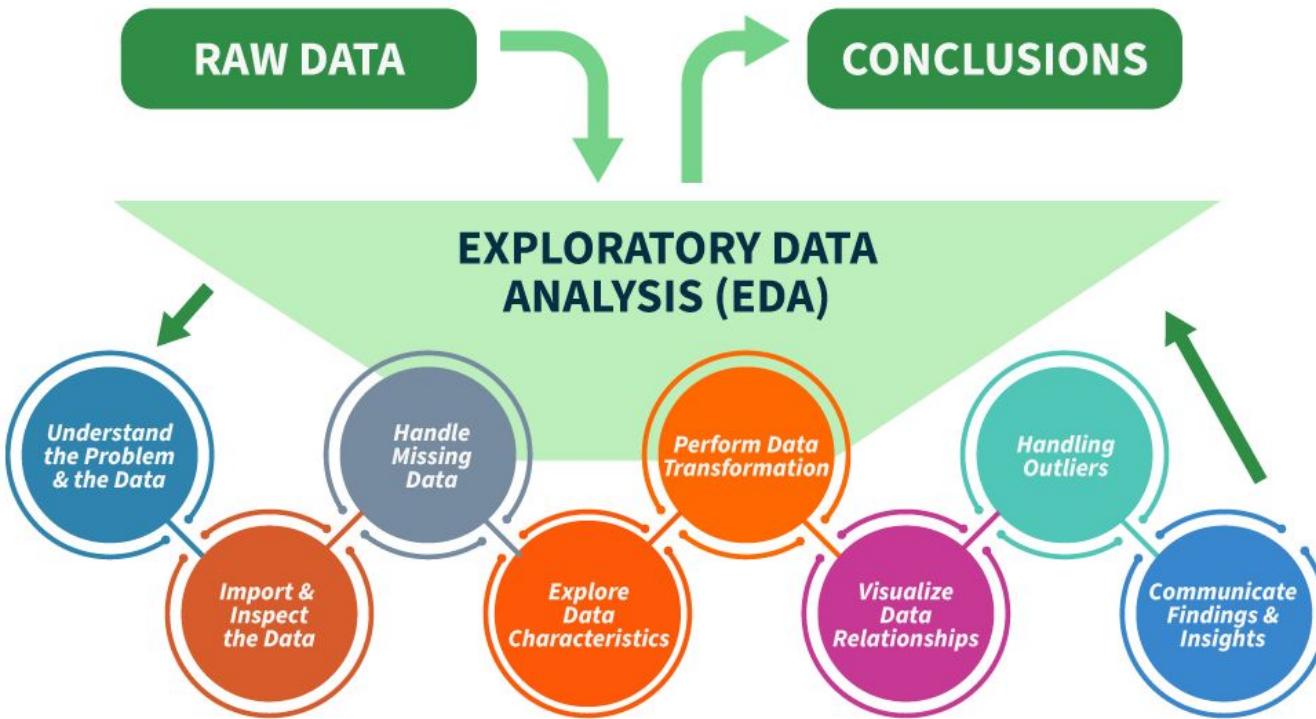
See [Anscombe's quartet](#) for example:

four datasets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed

This process is called
Exploratory Data Analysis (EDA)



Steps of EDA



<https://www.geeksforgeeks.org/what-is-exploratory-data-analysis/>

Categorical features encoding



Great overview

https://github.com/Dyakonov/PZAD/blob/master/2020/PZAD2020_042featureengineering_07.pdf

Whole ebook on feature engineering <https://feaz-book.com/>

Maximum Likelihood Estimation

girafe
ai

06



Parametric and nonparametric models

Nonparametric statistics is a type of statistical analysis that makes minimal assumptions about the underlying distribution of the data being studied. Often these models are infinite-dimensional, rather than finite dimensional, as is parametric statistics.

Nonparametric statistics can be used for descriptive statistics or statistical inference. Nonparametric tests are often used when the assumptions of parametric tests are evidently violated.

[© Common knowledge site](#)



Likelihood maximization

Consider the most simple case of discrete features and target.

Denote dataset X, Y generated by distribution with parameter θ

Likelihood of a parameter is defined as probability of sampling this particular data in case underlying distribution is defined by this parameter.

Maximization of likelihood means we choose the most probable parameters having this particular dataset

$$L(\theta|X, Y) = P(X, Y|\theta) \rightarrow \max_{\theta}$$

Note that likelihood is not probability function of θ



i.i.d. property

We can employ i.i.d property of data samples to split probability of the whole dataset into independent problems

$$P(X, Y | \theta) = \prod_i P(x_i, y_i | \theta)$$

Then we apply logarithm function to both parts of equation above

$$\log P(X, Y | \theta) = \sum_i \log P(x_i, y_i | \theta)$$

The latter expression is easier to operate with:
later we will predict log-probability of each object directly

Log-likelihood equivalence



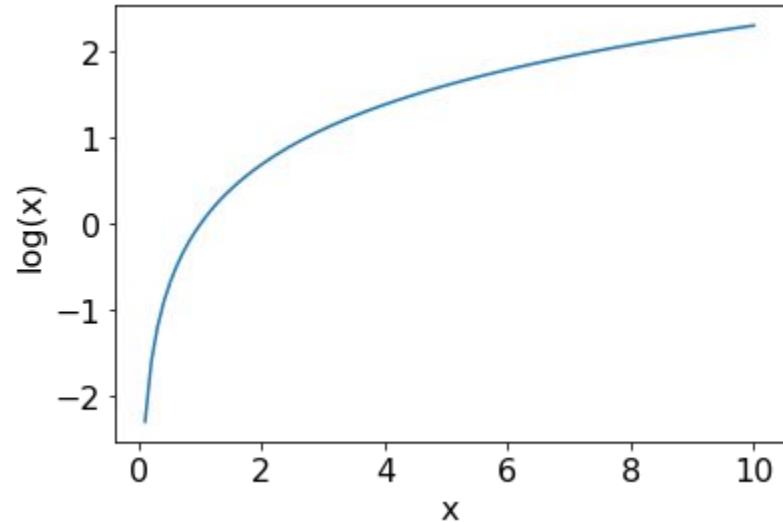
Since logarithm is a convex function on open set, it preserves maximum of expression when applied, so that

$$L(\theta|X, Y) \rightarrow \max_{\theta}$$

and

$$\log L(\theta|X, Y) \rightarrow \max_{\theta}$$

have the same solutions in terms of θ



Maximum Likelihood Estimation



$$\hat{\theta} = \arg \max_{\theta} L(\theta | X, Y)$$

is called maximum likelihood estimation of model parameters.

In optimization theory functions are usually minimized, so the same problem could be reformulated using **Negative Log-Likelihood (NLL)** loss

$$\hat{\theta} = \arg \min_{\theta} - \sum_i \log P(x_i, y_i | \theta)$$



Note

Here we formulate MLE in terms of probability which means we assume finite number of parameter values.

Defining MLE for infinite parameters set is left as easy exercise for you.

$$\hat{\theta} = \arg \min_{\theta} - \sum_i \log P(x_i, y_i | \theta)$$



Not only MLE

Generally in statistics plenty of ways to estimate parameters exist:

- Maximum likelihood estimators
- Bayes estimators
- Method of moments estimators
- Cramér–Rao bound
- Maximum a posteriori (MAP or EM algorithm)
- Particle filter
- Markov chain Monte Carlo (MCMC)
- Kalman filter, and its various derivatives
- Wiener filter

see https://en.wikipedia.org/wiki/Estimation_theory#Estimators

<https://en.wikipedia.org/wiki/Estimator>

https://en.wikipedia.org/wiki/Expectation%20maximization_algorithm

Machine Learning problems overview

girafe
ai

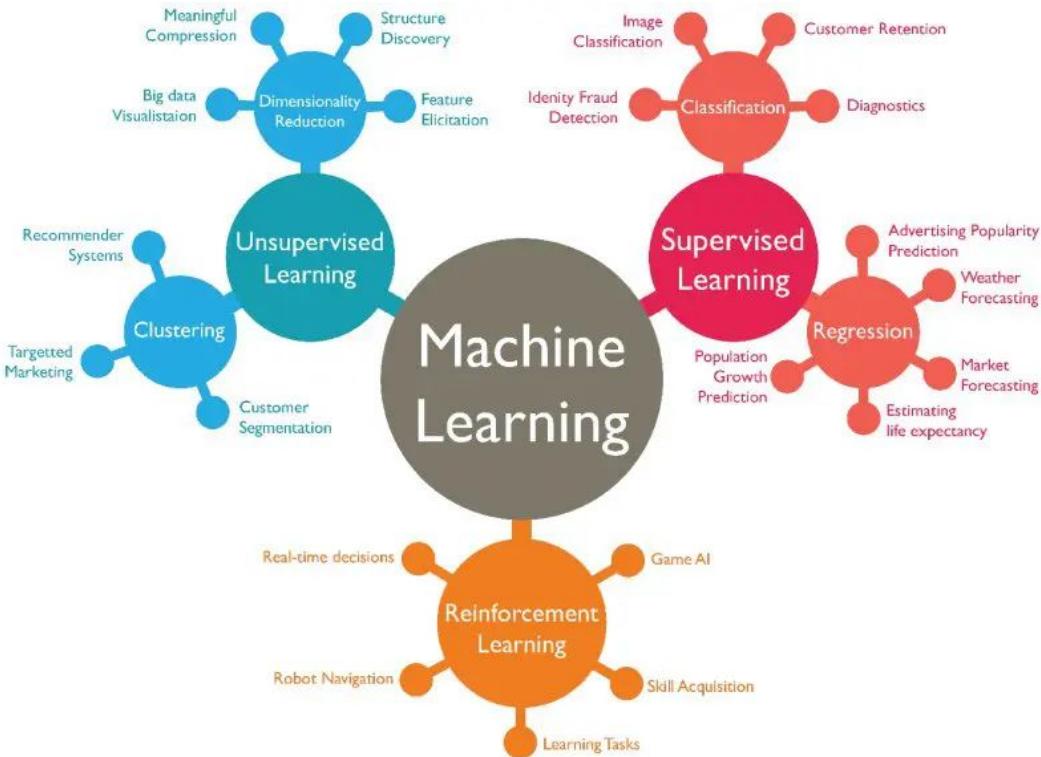
07

Types of problems in ML

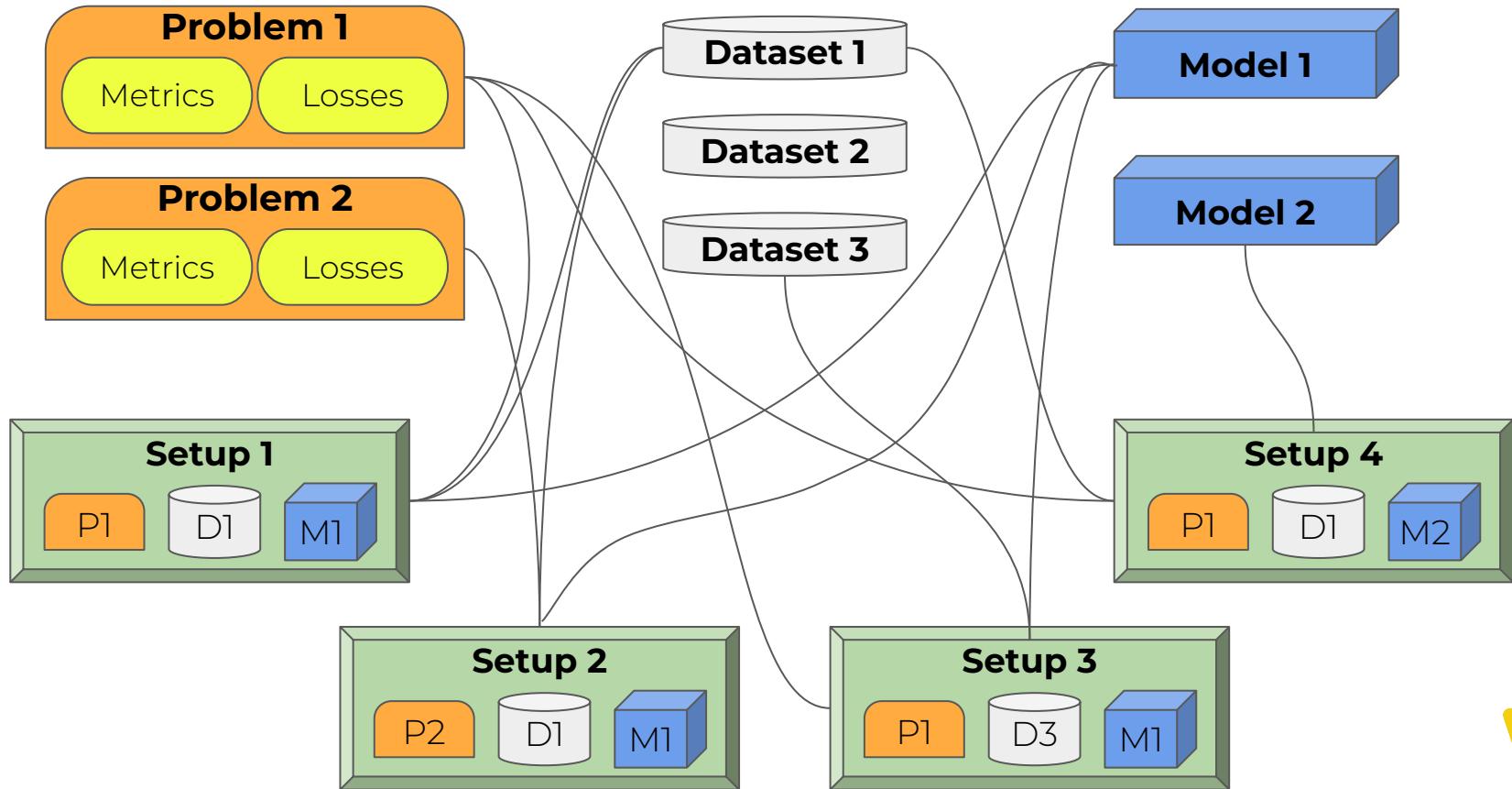


- Supervised
 - true answers are given
- Unsupervised
 - no true answers
- Semi-supervised
- Self-supervised
- Reinforced
- Generative
- etc...

There's no full and non-intersecting taxonomy of ML tasks, but terms above are used in practice



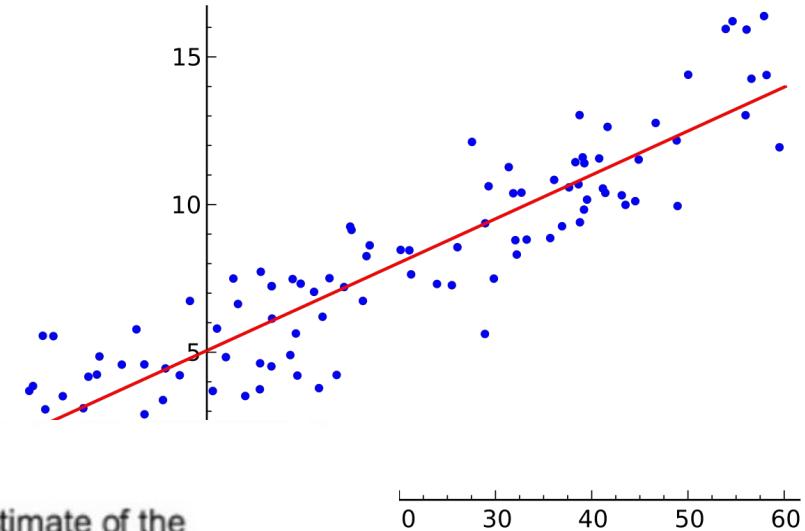
Problems and models



Problems covered in this course



- Regression problem



Estimated
(or predicted)
Y value for
observation i

Estimate of
the regression
intercept

Estimate of the
regression slope

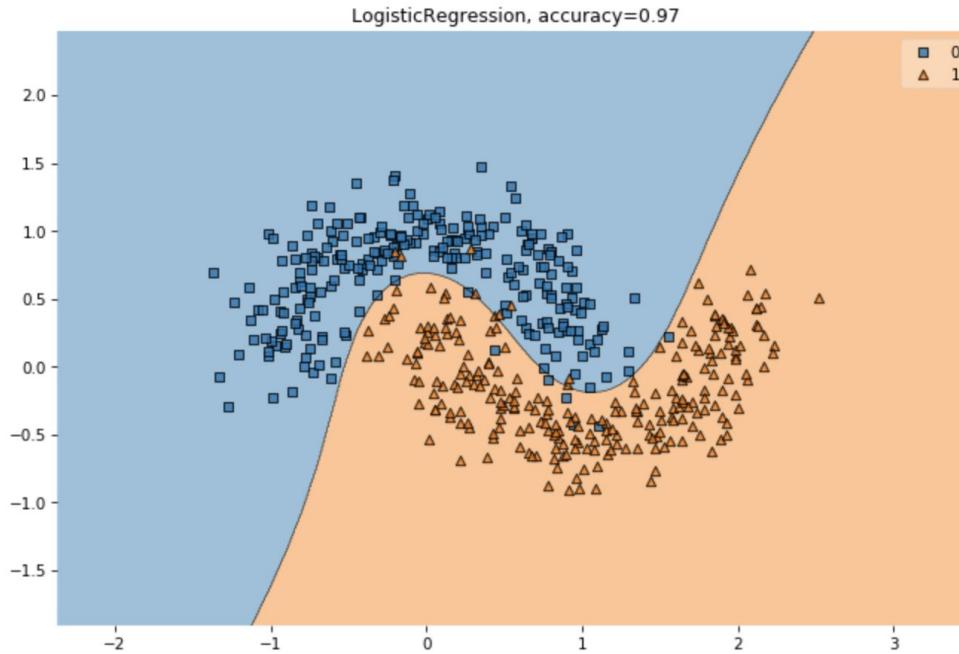
$$\hat{Y}_i = b_0 + b_1 X_i$$

Value of X for
observation i



Problems covered in this course

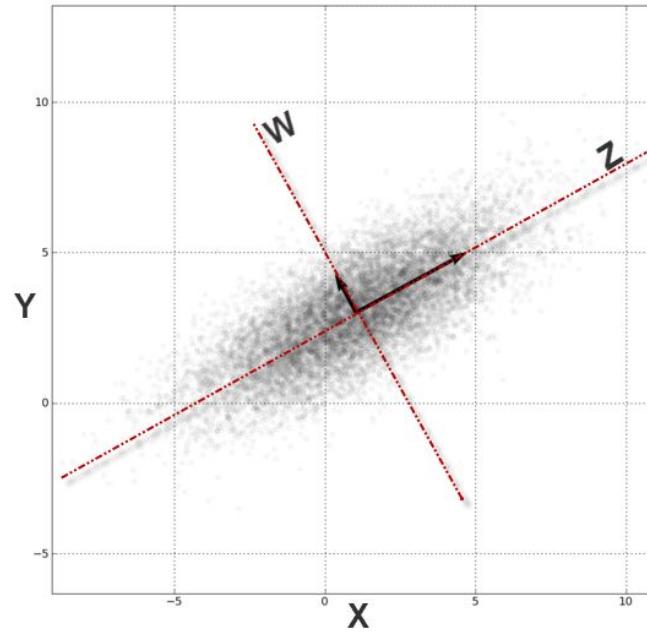
- Regression problem
- Classification problem



Problems covered in this course



- Regression problem
- Classification problem
- Dimensionality reduction
- Clustering
- Anomaly detection



Naïve Bayes classifier

girafe
ai

08



Bayes' theorem

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

or, in our case

$$P(y_i = C_k | \mathbf{x}_i) = \frac{P(\mathbf{x}_i | y_i = C_k)P(y_i = C_k)}{P(\mathbf{x}_i)}$$



Naïve Bayes classifier

Let's denote:

$\mathbf{x}_i \in \mathbb{R}^p$, $y_i \in \{C_1, \dots, C_K\}$ for K-class classification

$$P(y_i = C_k | \mathbf{x}_i) = \frac{P(\mathbf{x}_i | y_i = C_k) P(y_i = C_k)}{P(\mathbf{x}_i)}$$

Naïve assumption: features are **independent**



Naïve Bayes classifier

$$P(y_i = C_k | \mathbf{x}_i) = \frac{P(\mathbf{x}_i | y_i = C_k) P(y_i = C_k)}{P(\mathbf{x}_i)}$$

Naïve assumption: features are **independent**:

$$P(\mathbf{x}_i | y_i = C_k) = \prod_{l=1}^p P(x_i^l | y_i = C_k)$$



Naïve Bayes classifier

$$P(y_i = C_k | \mathbf{x}_i) = \frac{P(\mathbf{x}_i | y_i = C_k) P(y_i = C_k)}{\cancel{P(\mathbf{x}_i)}}$$

Optimal class label:

$$C^* = \arg \max_k P(y_i = C_k | \mathbf{x}_i)$$

To find maximum we even do not need the denominator

But we need it to get probabilities

Practical example



No	Statistics	Python	Target: Machine Learning
1	5	5	5
2	5	3	5
3	3	4	4
4	4	3	4
5	4	5	3
6	2	3	3

Class	Prob
5	0.33
4	0.33
3	0.33

feature	feature value	target	P
Statistics	2	5	0
Statistics	3	5	0
Statistics	4	5	0
Statistics	5	5	1

Machine learning libraries

girafe
ai

09

Sklearn



is the biggest library of ML methods implemented in one place.

Despite it is great tool for learning, it is not made for production

e.g. see <https://alexanderdyakonov.wordpress.com/2021/03/04/ml-scikit-learn/>

During this course we will give you tools which can be used for contemporary problems.



Pandas

is very handful tool but it is quite slow.

For better experience you can get alternatives:

- [Polars](#) (highly recommended!)
- [dask](#)
- [cudf](#)
- [swifter](#)



How to store data?

Straightforward solution is a .csv file. But it has many disadvantages such as filesize, sequential reading and others.

Better alternatives:

- [parquet](#) (supported by pandas and others) - column based
- [feather](#) - column based
- [ORC](#)
- [hdf5](#) - scientific format for datasets

Great resources on machine learning



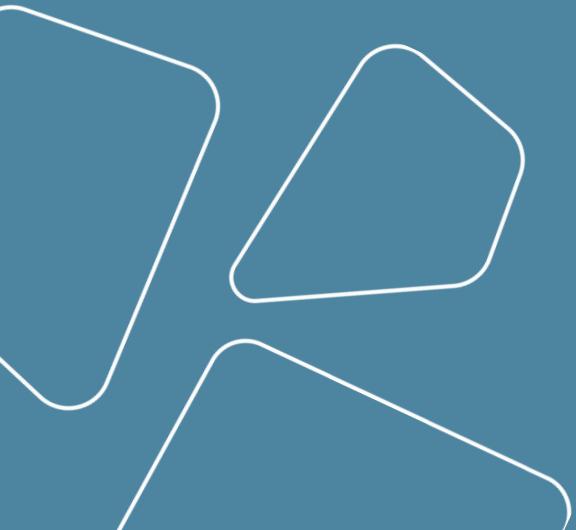
- [ru] <https://alexanderdyakonov.wordpress.com/>
- [ru] <https://github.com/esokolov/ml-course-hse>
- [en] <https://cs231n.stanford.edu/>
- [ru] <https://education.yandex.ru/handbook/ml>
- [ru] <http://iitp.ru/upload/publications/6256/vyugin1.pdf>
- [en, ru] <https://github.com/rasbt/python-machine-learning-book-3rd-edition>



Takeouts

- To find parameters use MLE
- Remember the i.i.d. property
- The first dimension of a dataset corresponds to the number of objects
- The second (and so on) to the features/time/...
- Even the naïve assumptions may be suitable in some cases
- Simple models provide great baselines

Revise

- 
- 1. ML and AI overview
 - 2. Thesaurus and notation
 - 3. Maximum Likelihood Estimation
 - 4. Some Machine Learning problems
 - a. Classification
 - b. Regression
 - c. Dimensionality reduction
 - 5. Naïve Bayes classifier

Thanks for attention!

Questions?



girafe
ai

