

Module 1. Introduction to Machine Learning

Lecture Plan:

1. Brief History of Machine and Deep Learning.
2. Standard Machine Learning Tasks.
3. Approach to Solving Machine Learning Tasks.
4. Machine Learning Toolkit and Tech Stack.

1. Brief History of Machine and Deep Learning.



IN THIS BUILDING DURING THE SUMMER OF 1956
JOHN MCCARTHY (DARTMOUTH COLLEGE), MARVIN L. MINSKY (MIT),
NATHANIEL ROCHESTER (IBM), AND CLAUDE SHANNON (BELL LABORATORIES)
CONDUCTED
**THE DARTMOUTH SUMMER RESEARCH PROJECT
ON ARTIFICIAL INTELLIGENCE**
FIRST USE OF THE TERM "ARTIFICIAL INTELLIGENCE"
FOUNDING OF ARTIFICIAL INTELLIGENCE AS A RESEARCH DISCIPLINE
"To proceed on the basis of the conjecture
that every aspect of learning or any other feature of intelligence
can in principle be so precisely described that a machine can be made to simulate it."
IN COMMEMORATION OF THE PROJECT'S 50th ANNIVERSARY
JULY 13, 2006

Trenchard More (PL/US), John McCarthy (LT), Marvin Minsky (US), Oliver Selfridge (EN) and Ray Solomonoff (US/BY)



From left: Yann LeCun, Geoffrey Hinton and Yoshua Bengio

Everything is moving super fast


Therefore, I recommend following the news of the world of Machine and Deep Learning:

- Yann LeCun - <https://twitter.com/ylecun>
- Geoffrey Hinton - <https://twitter.com/geoffreyhinton>
- Ian Goodfellow - https://twitter.com/goodfellow_ian
- Andrej Karpathy - <https://www.youtube.com/@AndrejKarpathy>
- Yannik Kilcher - <https://www.youtube.com/@YannickKilcher>
- Sam Altman - <https://twitter.com/sama>

How It All Started...


How It All Started...

1940
- 1950

- 
- Creation of the first primitive computers
 - Statement on the Artificial Intelligence scope creation
 - Alan Turing invented the Turing test
 - The first program that learned to play checkers
 - Frank Rosenblatt creates the first model of the Perceptron

How It All Started...

1960 - 1970

- 
- The emergence of the first machine learning algorithms (Nearest Neighbors and its Modifications)
 - The first Multilayer Neural Network. Creation of the Backpropagation method
 - Separation of the Machine Learning and Artificial Intelligence as independent areas

How It All Started...


1980-1990



- The first “AI Winter”. Stagnation and the emergence of a huge number of systems based on rules
- Origin of the recurrent neural networks (RNNs)
- Origin of the SVM algorithm
- Creation of the boosting methods
- Late 90s boom and the emergence of neural networks for working with sound and image data


How It All Started...

1990-2000

- 
- Second “AI Winter”
 - Improvements on Boosting Algorithms
 - The creation and popularity of the World Wide Web and Computers
 - The release of Boltzmann Machines algorithms
 - Formation of the Reinforcement Learning (RL)


How It All Started...

2000-2010

- 
- The rise of the World Wide Web and computers popularity leads to an increase in data and more opportunities for computation and data processing
 - Renaissance of the neural networks: processing images, video data, sound data, speech, text and etc.
 - The victory of the neural network in GO game (Google's DeepMind Competition)
 - The emerge of the Deep Fake technologies
 - Cloud Computing popularity

How It All Started...

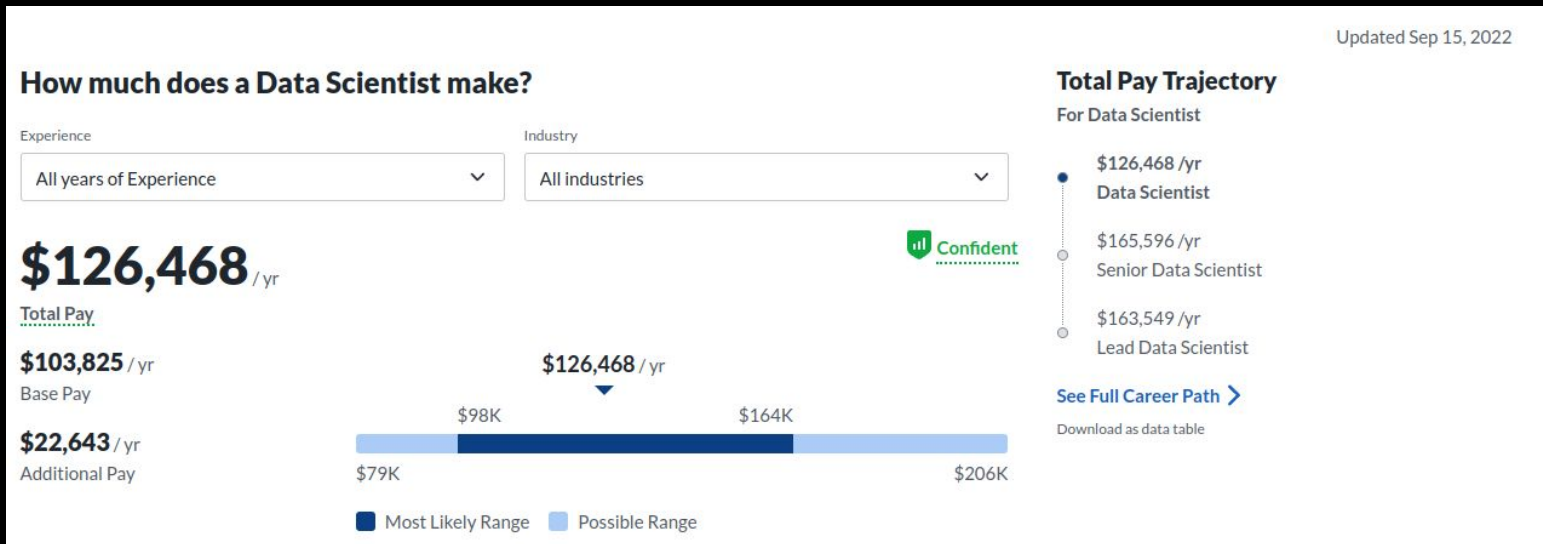
2010-H.B.

- 
- Deep Learning development
 - Transformer architecture (Attention is all you need)
 - Quantum Computing
 - Distribution of the machine learning and deep learning algos in the IT industry and manufacturing (e.g. Copilot, GANs, GPT-3, Bert, ChatGPT etc.)
 - Self-driving cars, self-supervised systems, robotics
 - True AI (AGI - artificial general intelligence)
 - LLMs (Large Language Models)
 - Multi-modal models

Use-cases and examples

- <https://thispersondoesnotexist.com/>
- <https://talktotransformer.com/>
- <http://www.r2d3.us/visual-intro-to-machine-learning-part-1/>
- <https://www.youtube.com/watch?v=S3F1vZYpH8c>
- <https://www.youtube.com/watch?v=7atZfX85nd4>

Why is it worth doing it

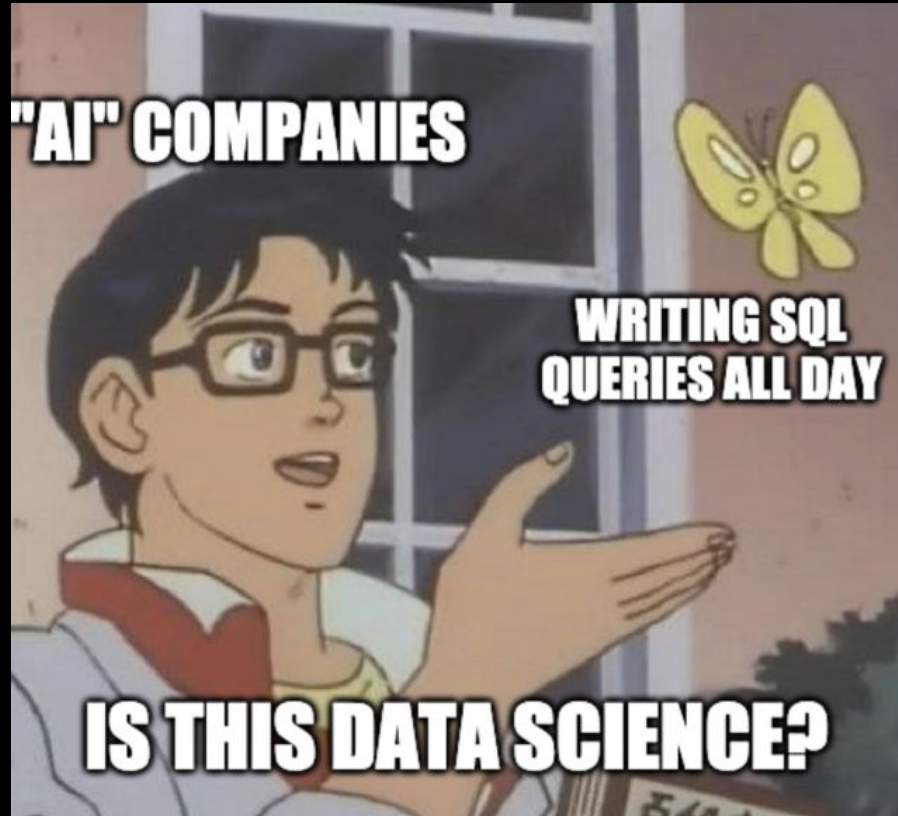


https://www.glassdoor.com/Salaries/data-scientist-salary-SRCH_KO0.14.htm

Pros and Cons of ML

- Multidisciplinary
- In Demand Profession
- Variety of challenges
- Big Industry
- Self-Learning. Lots of self-learning...
- Research Speed
- Efficiency
- Very fast industry development. You need to run fast in order just to stay in place.

Pros and Cons of ML

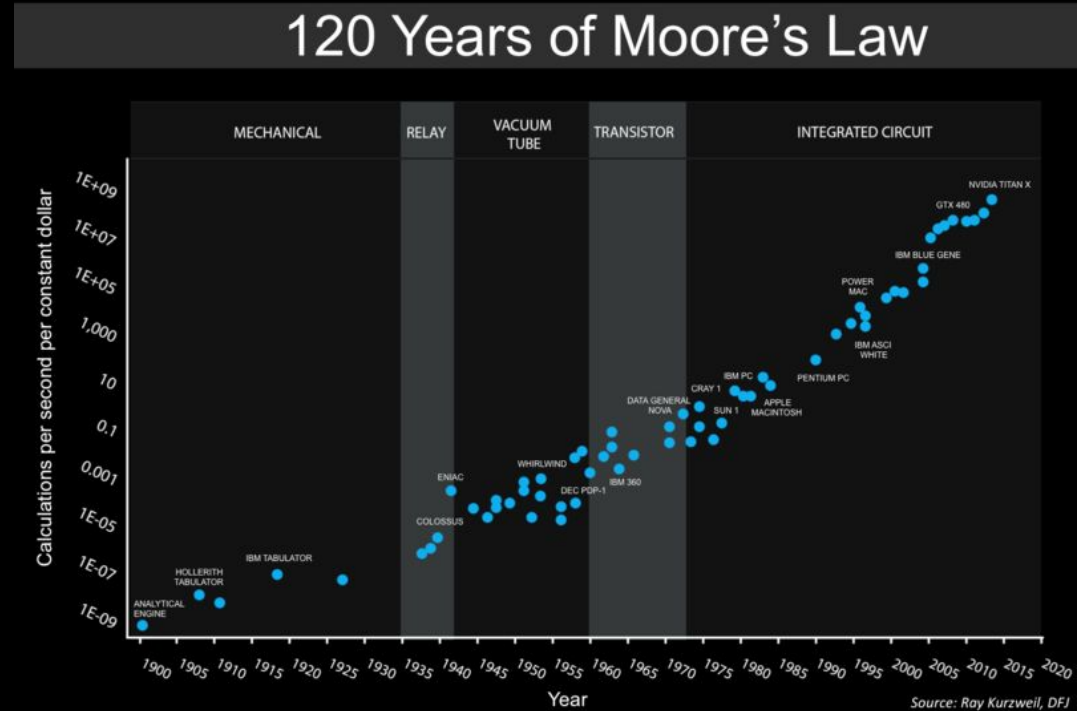


Computation Speed

Moore's Law (1965) is an empirical observation originally made by Gordon Moore (IBM) which states that the number of transistors placed on an integrated circuit chip will double every 24 months.

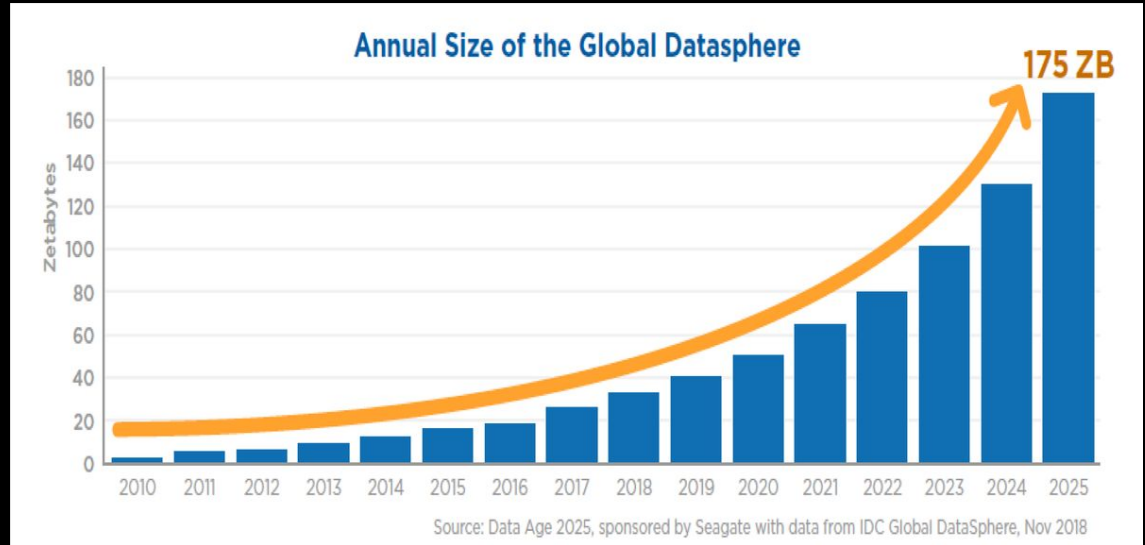
The oft-quoted 18-month interval is related to a prediction by David House (Intel) that processor performance should double every 18 months due to a combination of increasing transistor count and increasing processor clock speeds.

This law has now been refuted.



World Wide Web Development and Data Accumulation

Thanks to the development of the Internet and the comprehensive penetration of computers and mobile devices into all spheres of human life, there is an increase in the amount of data on the network.



According to the IDC forecast, the share of the global information sphere subjected to analysis by 2025 will grow 50 times or more, reaching 5.2 Zbytes, and the amount of data analyzed with the participation of cognitive systems will grow 100 times, amounting to 1.3 Zbytes. Cognitive systems will allow more frequent and more flexible data analysis in many industries.

Big Data

1880s - processing of information and the presentation of census data in Europe and the United States took more than 8 years.

At the same time, according to forecasts, the processing of census data in 10 years would have taken even longer, and the results would not have been ready even before the next census (1900).

Then the problem was solved by the tabulating machine, invented by Herman Hollerith in 1881.

1997 - The term Big Data was first introduced in 1997 by Michael Cox and David Ellsworth at the 8th IEEE Visualization Conference.

In 1998, SGI research leader John Mashy at the USENIX conference used the term Big Data in its modern form.

Big data is data, the processing of which, under given conditions and restrictions, requires a huge amount of computing resources and the use of new technical approaches.

Big Data



The turning point was in 2003.

The more data were generated than for the entire previous time.

The Google File System Conference.

MapReduce computing system formed the basis of Apache Hadoop.

Hadoop has become a full-fledged exodus for storing and processing Big Data.

Interrelation of Data Mining, Big Data and Data Science

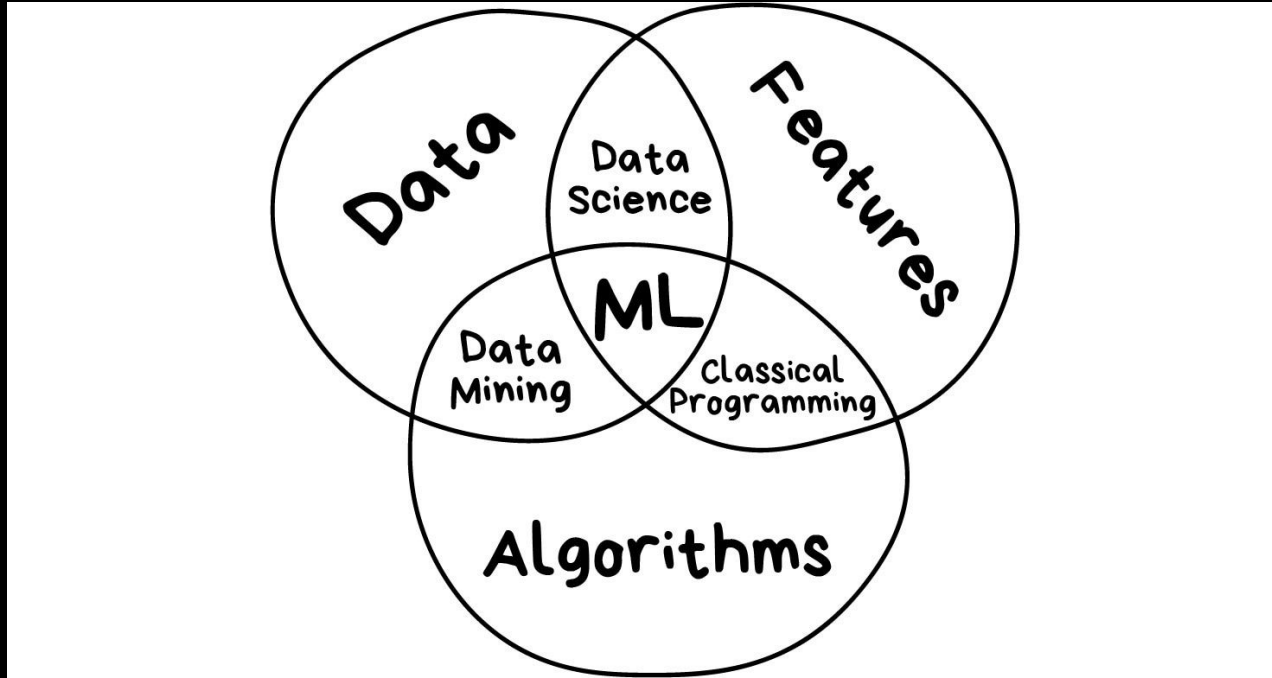
Data Mining (DM) - extraction, procession and interpretation of data.

Methods:

- **Exploratory Data Analysis** (EDA) - mining for useful information using data visualization tools and statistical methods.
- **Classical Machine Learning Algorithms** - classification methods, statistical methods (decision trees, correlation and regression analyses, factor analysis and etc.)

One of the most important purposes of data mining methods is to visualize the results of computations. Give the result in human-accessible manner.

Interrelation of Data Mining, Big Data and Data Science



2. Standard Machine Learning Tasks.

Data

- To solve ML/DL problems we need data.
- Lots of data (but there are exceptions)
- Data can be divided into 2 groups
 - Structured data
 - Unstructured data



Structured Data

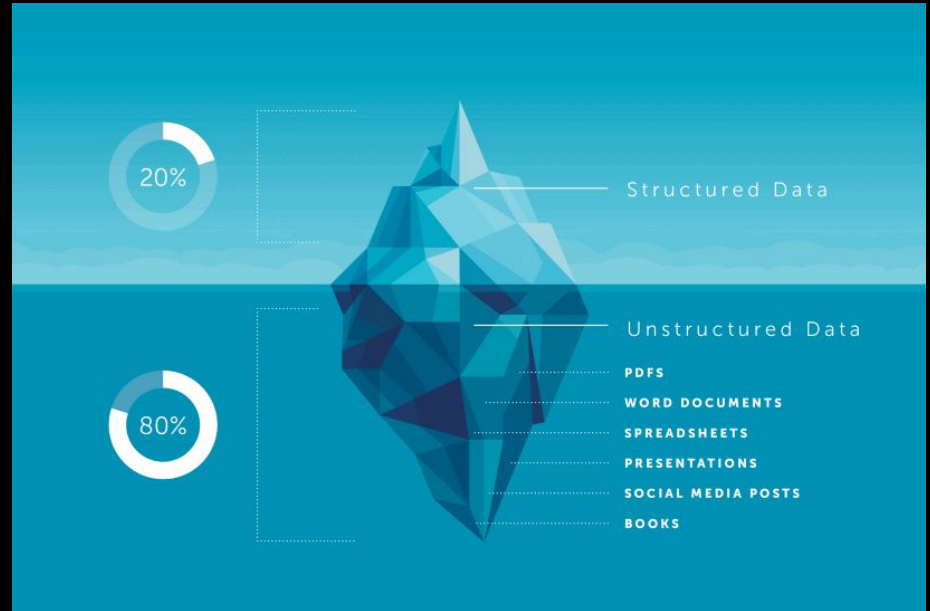
Columns

Rows

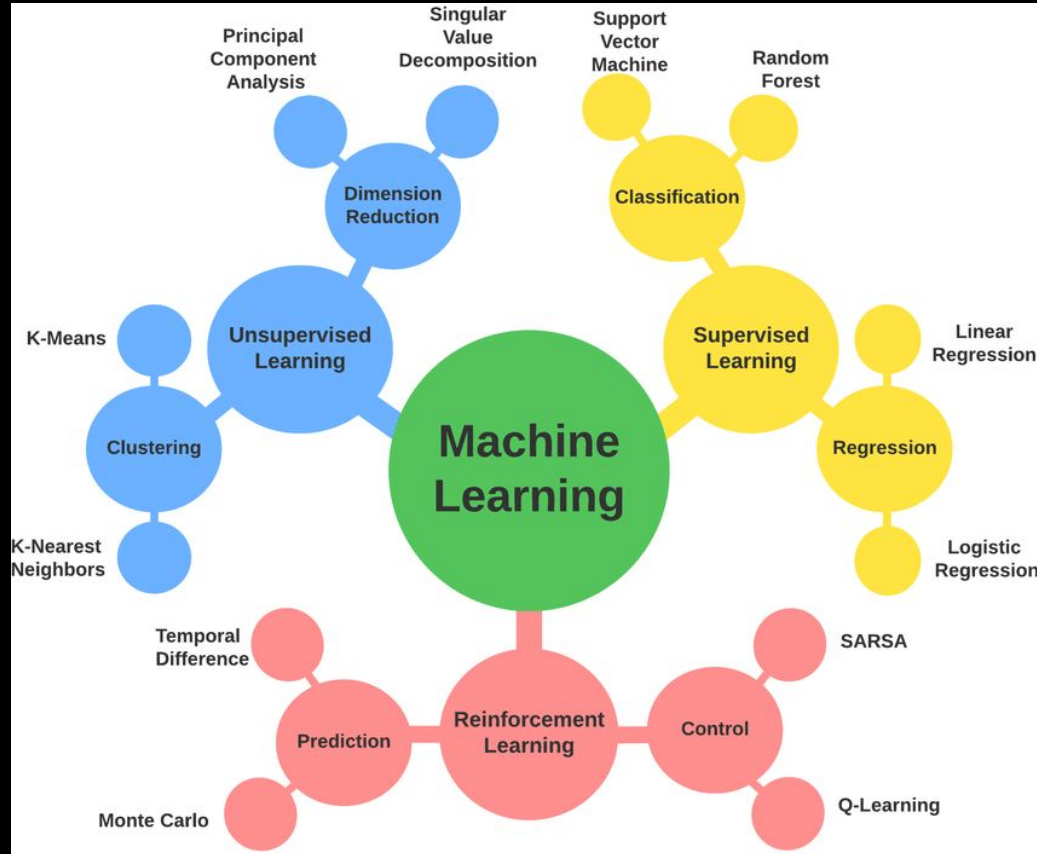
name	region	sales	expenses
William	East	50000	42000
Emma	North	52000	43000
Sofia	East	90000	50000
Markus	South	34000	44000
Edward	West	42000	38000
Thomas	West	72000	39000
Ethan	South	49000	42000
Olivia	West	55000	60000
Arun	West	67000	39000
Anika	East	65000	44000
Paulo	South	67000	45000

Unstructured Data

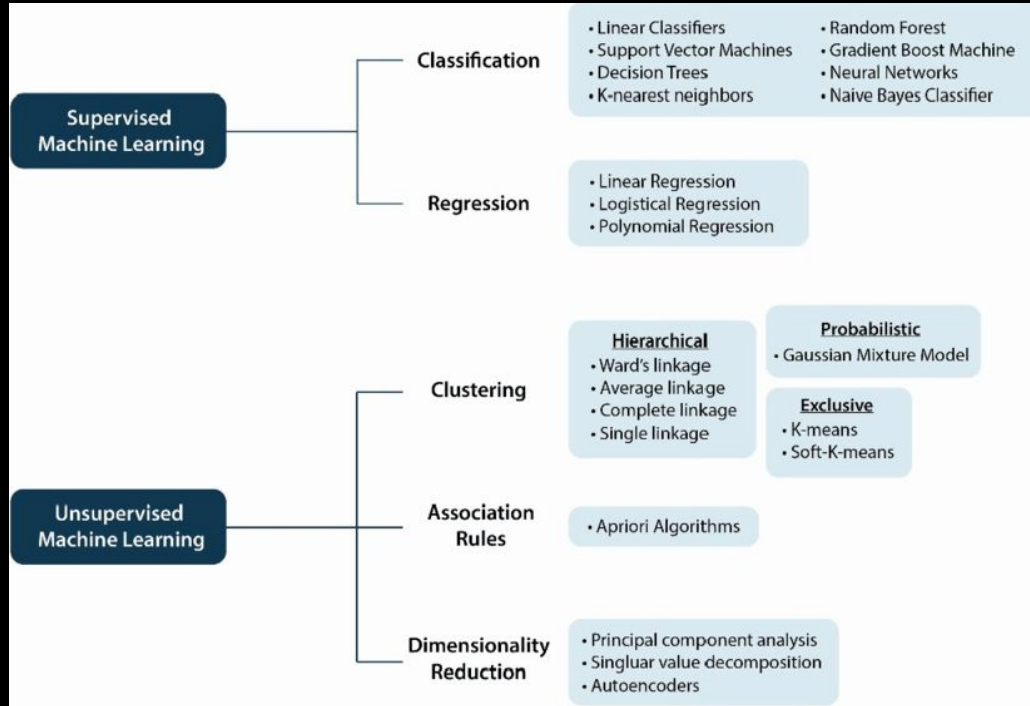
- Text Data
- Image Data
- Video Data
- Audio Data
- Graph Data
- Streaming Data
- Metadata
- Geospatial Data



Types Of Machine Learning Tasks



Classical Machine Learning



Supervised Learning

- Data has a label (Label, Feature Label, Y , target variable)
- There are observations (positive and negative) that need to be predicted
- If the target variable is categorical (e.g. 0 or 1) then we have a **classification** task
- If the target variable is a continuous value (e.g. price of a security, the level of air pollution and etc.) then we have a **regression** task

Unsupervised Learning

- Data has no label (**Unlabeled** data)
- But it is necessary to **extract useful information** (patterns, clusters, similar behaviour or insights)
- For example: Which objects (observations) are similar to each other and why?
This is a **clustering** task
- Which objects (observations) are not similar to any in the sample. This is the **outlier detection** task
- We need to change the number of features in the data without losing useful information. This is the **dimensionality reduction** task

Reinforcement Learning

- “True AI”
- There are no correct answers, but there is an **environment** in which **agents interact**. There is a **reward** for actions and a **punishment** for an incorrect action. The task is to develop an optimal plan of behavior in accordance with the task.
- Examples: Dota AI, DeepMind etc.

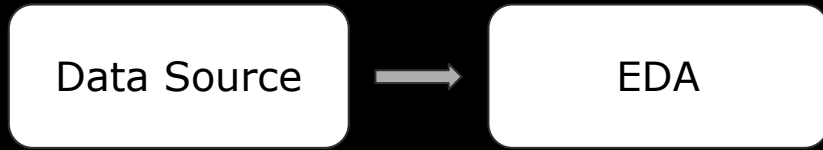
3. Approach to Solving Machine Learning Tasks

Approach to Solving Machine Learning Tasks

Data Source

- What type of data
- Data source
- Content
- Volume
- Structure

Approach to Solving Machine Learning Tasks



Exploratory Data Analysis (EDA)

- Data exploration
- Statistical analysis
- Data visualization
- Search for patterns and correlation
- Potential model selection

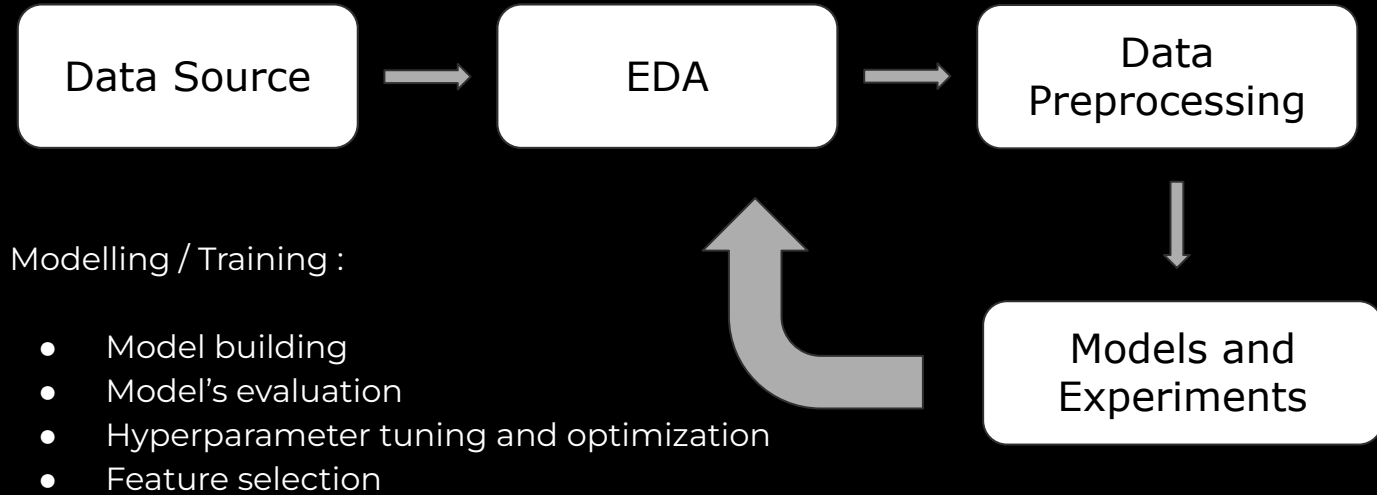
Approach to Solving Machine Learning Tasks



Data Preprocessing:

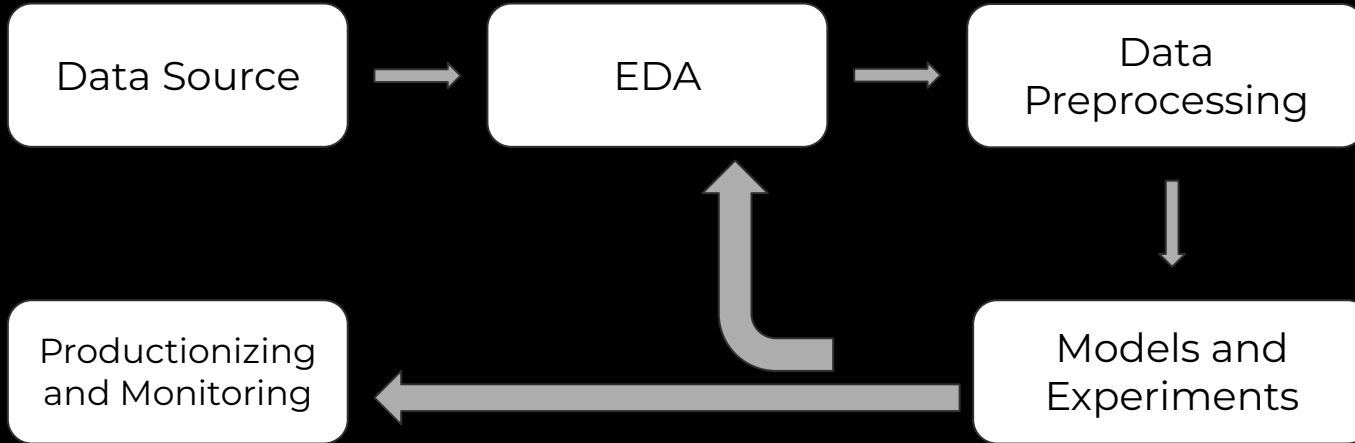
- Converting data into a suitable format for machine learning
- Encoding of the categorical features
- Outliers and anomaly detection
- Data cleaning
- Feature creation
- Feature selection
- Dimensionality reduction

Approach to Solving Machine Learning Tasks



Repeat the cycle if everything is bad.

Approach to Solving Machine Learning Tasks

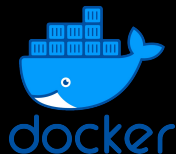
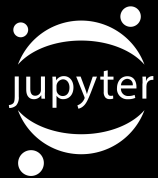


Approach to Solving Machine Learning Tasks

Now you know the main steps in the methodology called **CRISP-DM** (Cross Industry Standard Process for Data Mining).

4. Machine Learning Toolkit and Tech Stack

5. ML Toolkit



6. Python / Anaconda / Jupyter Notebook / Google Collab

Main instrument: Python



Where to download

<https://www.python.org/downloads/>

6. Python / Anaconda / Jupyter Notebook / Google Collab

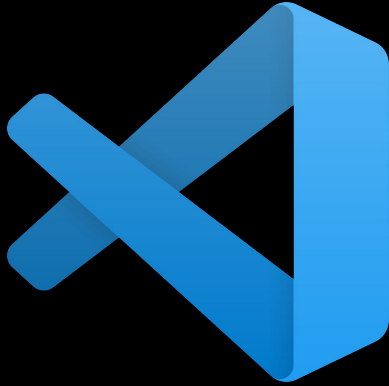


Installation:

- Anaconda: <https://www.anaconda.com/>
- Google Colab: <https://colab.research.google.com/>

6. Python / Anaconda / Jupyter Notebook / Google Collab

IDE or Code Editor



VS Code

<https://code.visualstudio.com/>

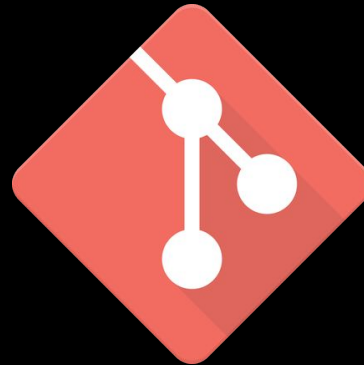


PyCharm

<https://www.jetbrains.com/pycharm/download>

6. Python / Anaconda / Jupyter Notebook / Google Collab

Version Control System (VCS)

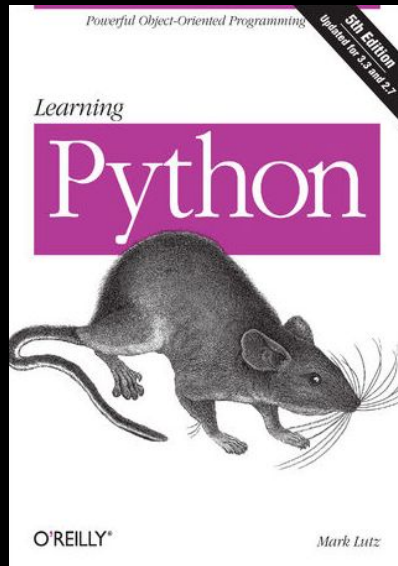


- GitHub <https://github.com/>
- Download git bash <https://git-scm.com/downloads>

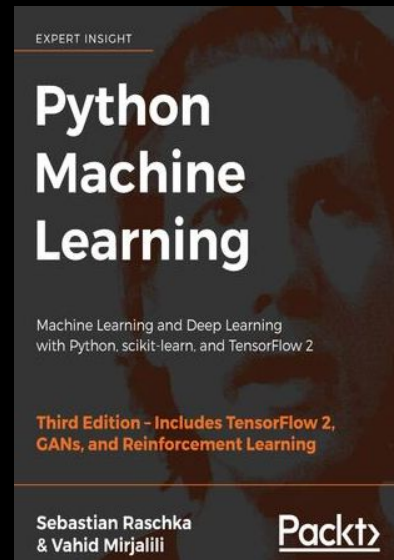
Resume

- Machine learning is one of the most demanded areas of the present and near future.
- Data Science combines many interdisciplinary areas, including business in the form of domain knowledge
- Cross Disciplinarity allows you not to be tied to a particular industry, but requires large doses of self-education and independent search for information on emerging issues

Recommended books



**Mark Lutz - Learning
Python**



**Sebastian Raschka -
Python Machine
Learning**