hillel
IT school

Lesson 01
Introduction to Machine Learning

План лекції

- Привітання та знайомство
- Огляд курсу
- Що таке AI/DS/ML/DL?
- Задачі, що вирішує ML
- Етапи реалізації проєктів з ML
- Збір та обробка даних в ML, джерела даних та їх особливості
- Огляд основних інструментів в ML з використанням Python

Знайомство

Олег Коменчук

Machine Learning Engineer (4 years of commercial experience)

PhD Student in Information Systems and Technologies (2021 - Present)

Kaggle Notebooks Expert

LinkedIn | Kaggle | GitHub

Знайомство зі студентами

- Інформація про себе
- Освіта та сфера інтересів
- Робочий досвід (якщо є)
- Попередні курси чи проєкти, досвід програмування на Python
- Очікування від курсу

Комунікація

Telegram-група: https://t.me/+RVnrUzqmZZcxMGUy

Огляд курсу

23 заняття
2 заняття на тиждень в Zoom
Вівторок, П’ятниця (19:15 - 21:15)

Домашні завдання:

- Практичні завдання
- Тест
- Дедлайни
- Перездачі
- Оцінювання

- Introduction to Machine Learning
- NumPy, Pandas, Matplotlib
- Linear & Logistic Regression
- Logistic Regression for Multiclass Classification
- Regularization
- Tree Based Models
- Scikit-Learn Workflow
- Intro to Deep Learning: PyTorch
- Intro to Deep Learning: Layers
- Intro to Deep Learning: Optimization
- Computer Vision: Intro
- Computer Vision: Classification Models
- Computer Vision: Segmentation Models
- Computer Vision: Object Detection
- Natural Language Processing: Intro
- Natural Language Processing: Embeddings
- Natural Language Processing: RNNs
- Transformers
- LLMs, LLM Prompting
- Recommender Systems
- Autoencoders
- Generative Adversarial Networks (GANs)
- Machine Learning Model Deployment

What is Artificial Intelligence, Machine Learning, Deep Learning, and Data Science?

AI/ML: Self-Driving Cars

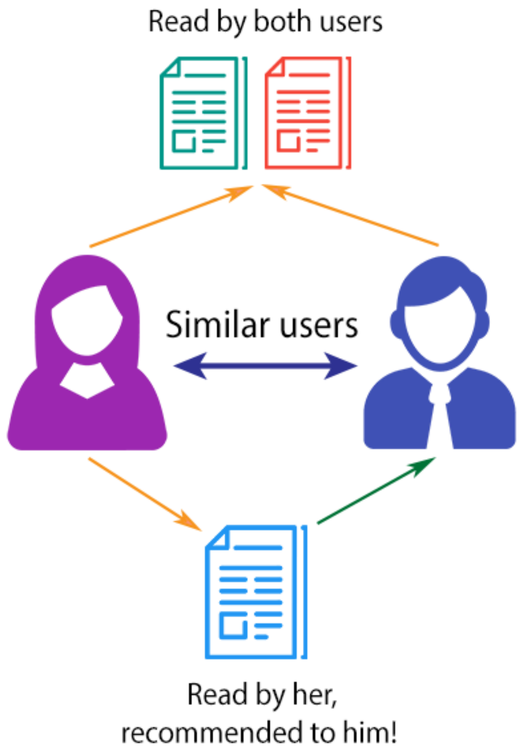
Tesla Autopilot

Waymo

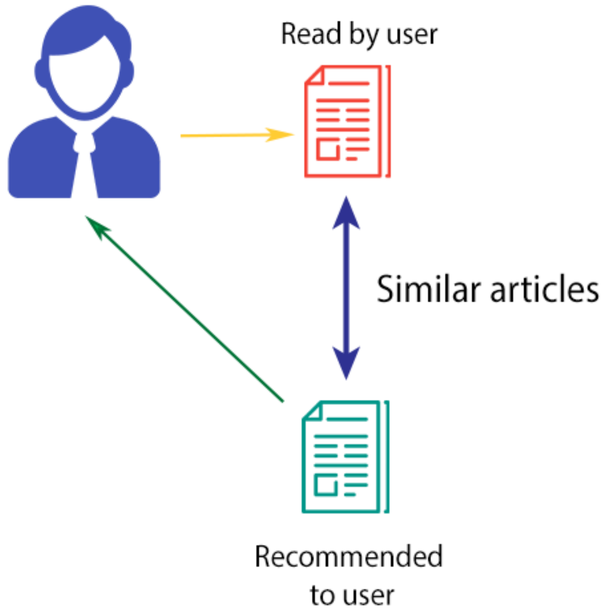
Mobileye

AI/ML: Recommender Systems

COLLABORATIVE FILTERING



CONTENT-BASED FILTERING



YouTube

Spotify®

NETFLIX

Google

Instagram

AI/ML: Smart Assistants

Apple Siri

Siri is a virtual assistant that is part of Apple's iOS, watchOS, macOS, and tvOS operating systems. The assistant uses voice queries and a natural-language user interface to answer questions, make recommendations, and perform actions by delegating requests to a set of Internet services.

Amazon Alexa

Alexa is a virtual assistant developed by Amazon. It is capable of voice interaction, music playback, making to-do lists, etc.

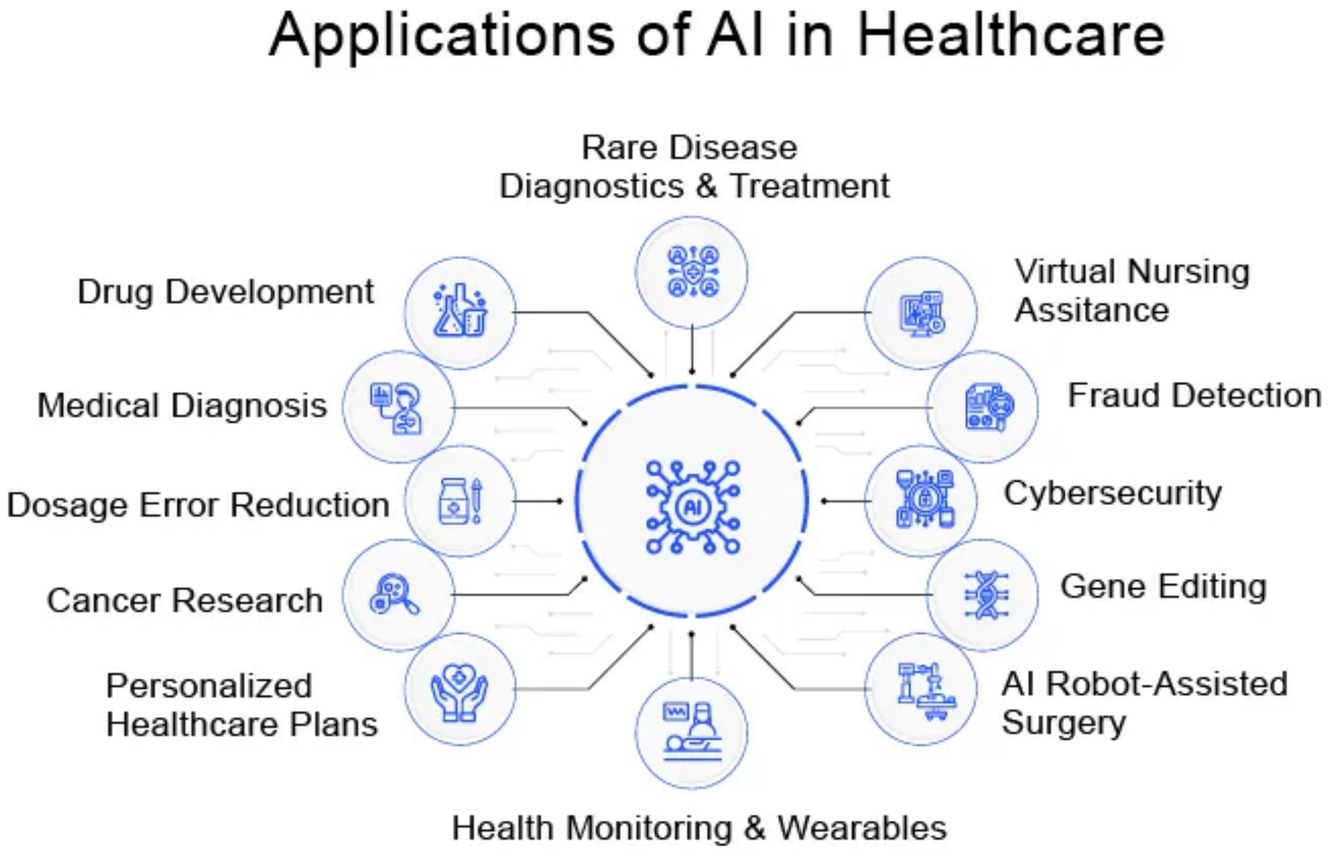
Google Assistant

Google Assistant is an AI-powered virtual assistant developed by Google that is primarily available on mobile and smart home devices, and can engage in two-way conversations.

Microsoft Cortana

Cortana is a virtual assistant created by Microsoft for Windows 10, Windows 10 Mobile, Windows Phone 8.1. It can set reminders, recognize natural voice without the requirement for keyboard input, and answer questions using information from the Bing search engine.

AI/ML: Healthcare



AI/ML: Generative AI

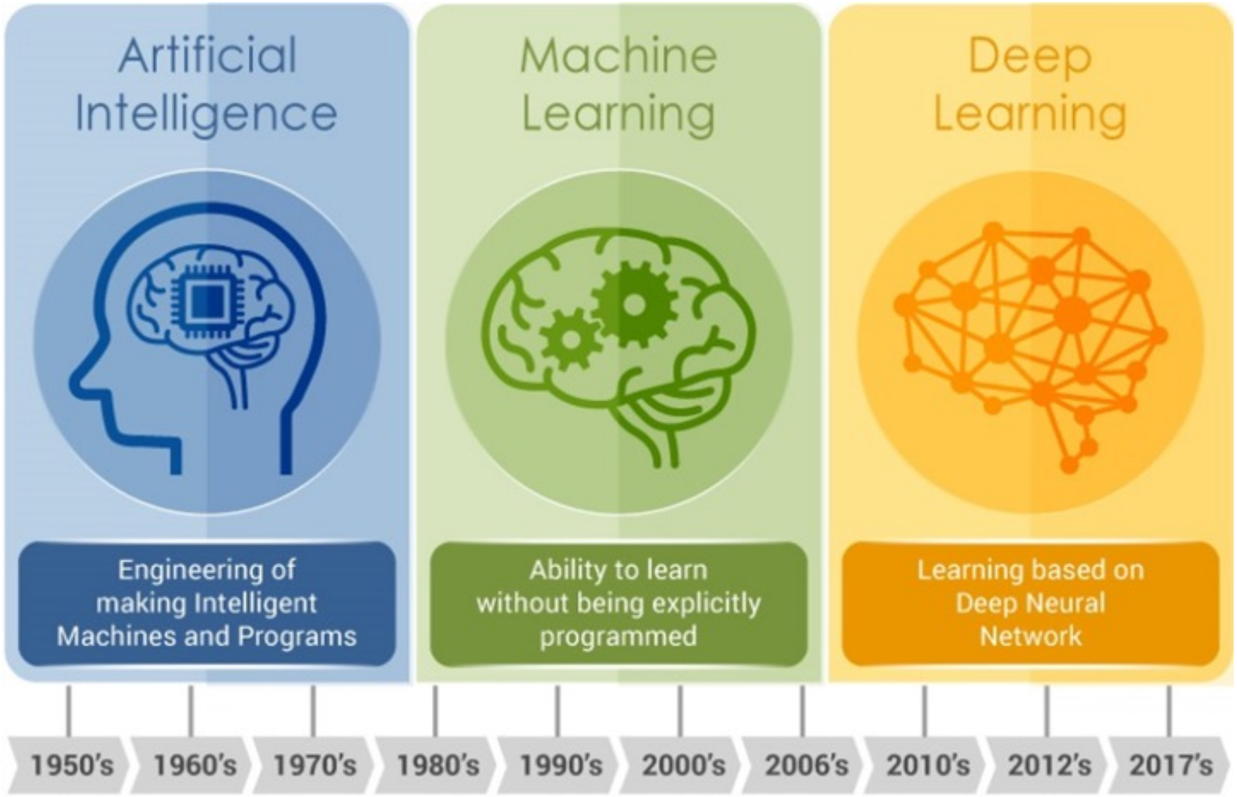
ChatGPT is a chatbot developed by OpenAI and launched on November 30, 2022. Based on a large language model, it enables users to refine and steer a conversation towards a desired length, format, style, level of detail, and language.

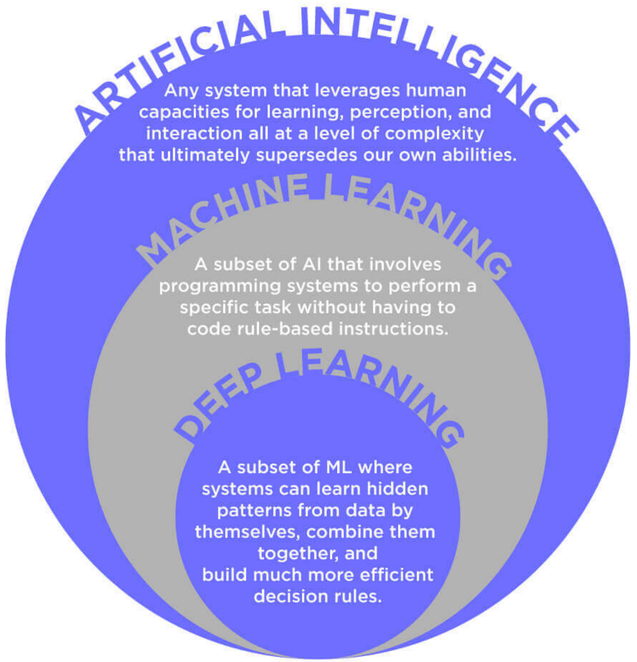
DALL·E, DALL·E 2, and DALL·E 3 are text-to-image models developed by OpenAI using deep learning methodologies to generate digital images from natural language descriptions, called "prompts."

Gemini, formerly known as Bard, is a generative artificial intelligence chatbot developed by Google.

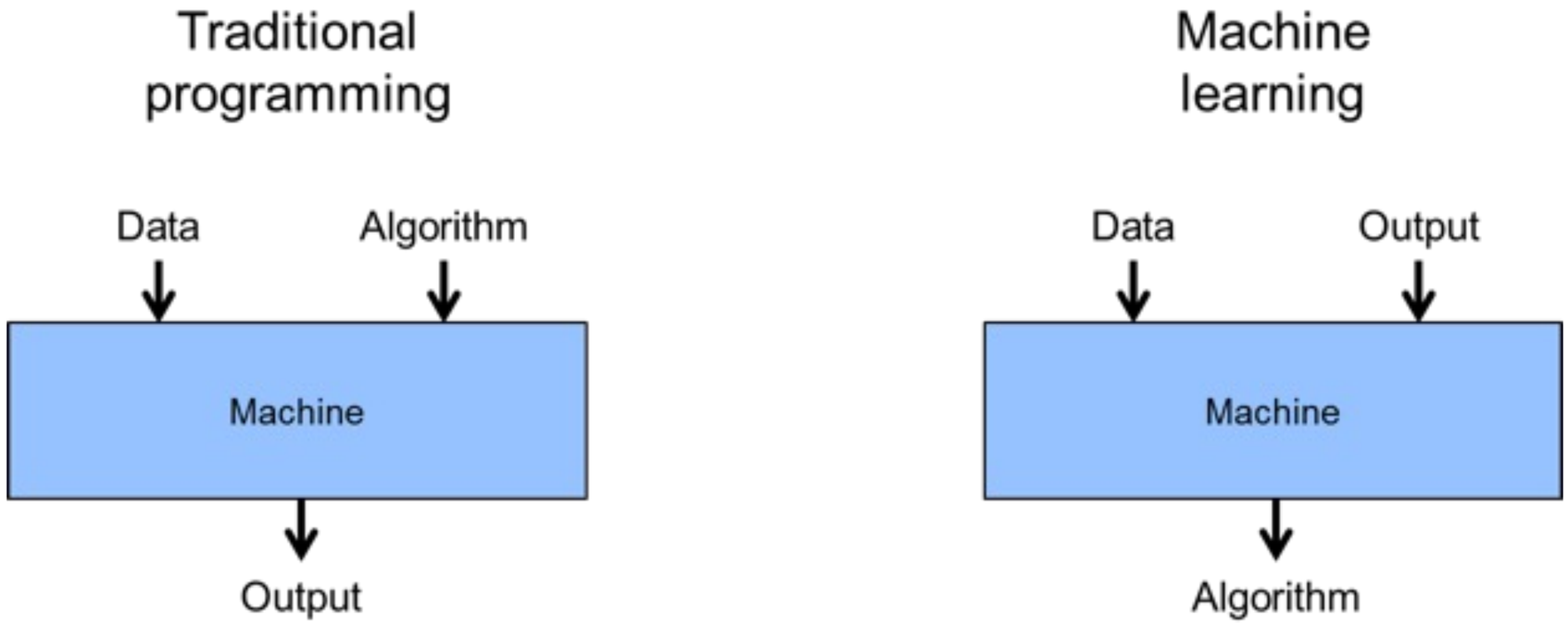
Midjourney is a generative artificial intelligence program and service created and hosted by the San Francisco–based independent research lab Midjourney, Inc. Midjourney generates images from natural language descriptions, called prompts, similar to OpenAI's DALL-E and Stability AI's Stable Diffusion.

AI, ML, DL


Artificial Intelligence
Machine Learning
Deep Learning
Engineering of making Intelligent Machines and Programs
Ability to learn without being explicitly programmed
Learning based on Deep Neural Network
1950's
1960's
1970's
1980's
1990's
2000's
2006's
2010's
2012's
2017's


ARTIFICIAL INTELLIGENCE
Any system that leverages human capacities for learning, perception, and interaction all at a level of complexity that ultimately supersedes our own abilities.
MACHINE LEARNING
A subset of AI that involves programming systems to perform a specific task without having to code rule-based instructions.
DEEP LEARNING
A subset of ML where systems can learn hidden patterns from data by themselves, combine them together, and build much more efficient decision rules.

The difference between traditional programming and machine learning

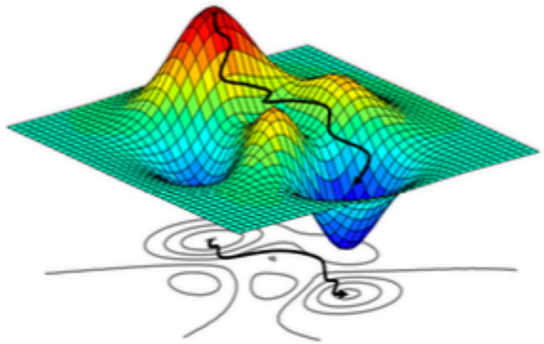


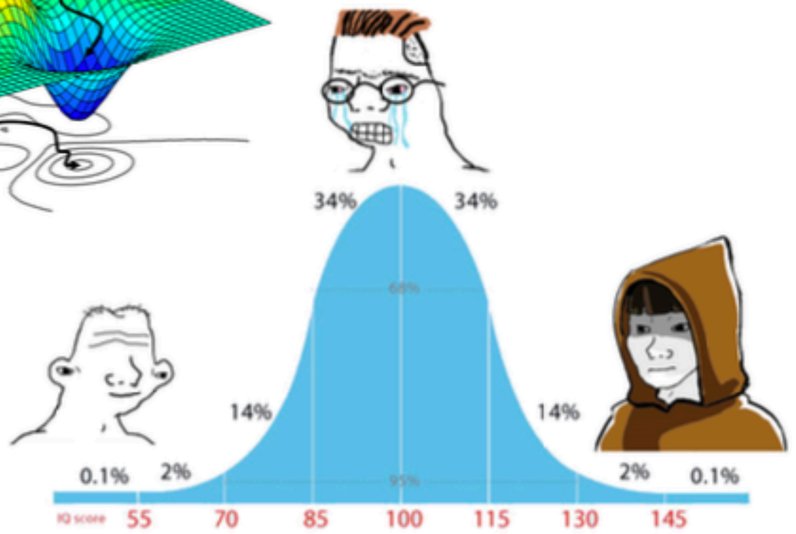
Math in ML

- Linear Algebra
- Mathematical Analysis
- Calculus
- Statistics

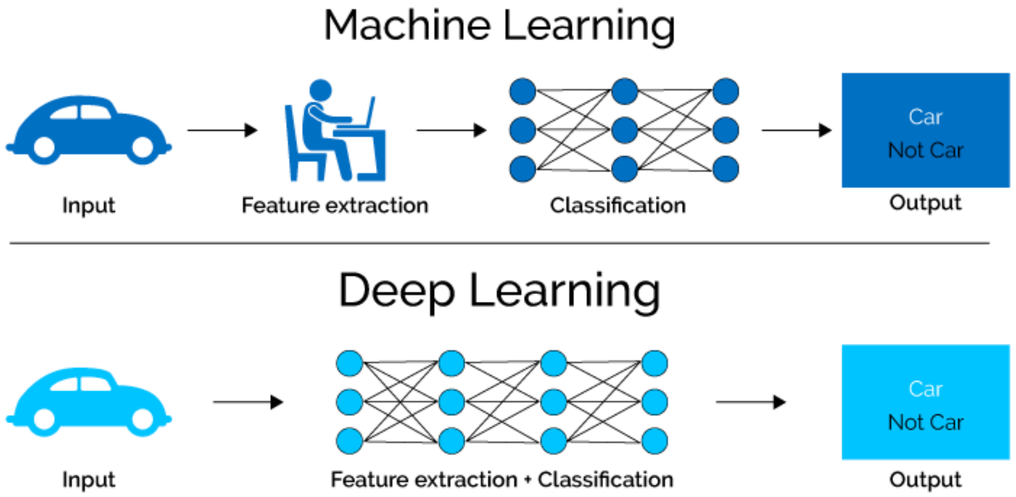
$$\begin{bmatrix} a_1 & a_2 & a_3 \\ a_4 & a_5 & a_6 \\ a_7 & a_8 & a_9 \end{bmatrix} \begin{bmatrix} b_1 & b_2 & b_3 \\ b_4 & b_5 & b_6 \\ b_7 & b_8 & b_9 \end{bmatrix} = \begin{bmatrix} c_1 & c_2 & c_3 \\ c_4 & c_5 & c_6 \\ c_7 & c_8 & c_9 \end{bmatrix}$$

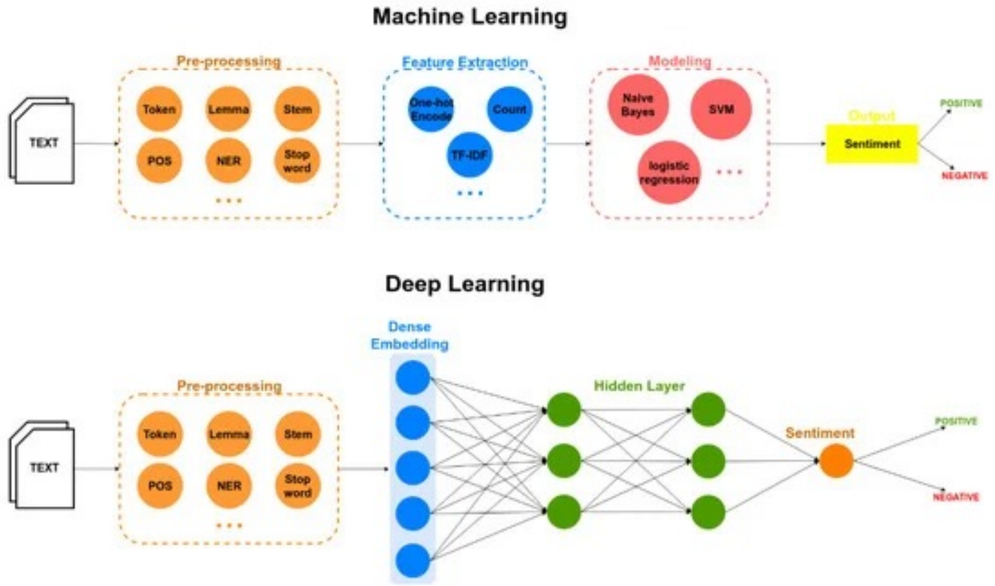
$$\frac{d(sinx)}{x} = cosx$$
$$\frac{d(cosx)}{x} = -sinx$$
$$\frac{d(tanx)}{x} = -sec^2$$
$$\frac{d(cotx)}{x} = cosec^2$$
$$\frac{d(secx)}{x} = secx.tanx$$
$$\frac{d(cosecx)}{x} = -cosecx.cotx$$



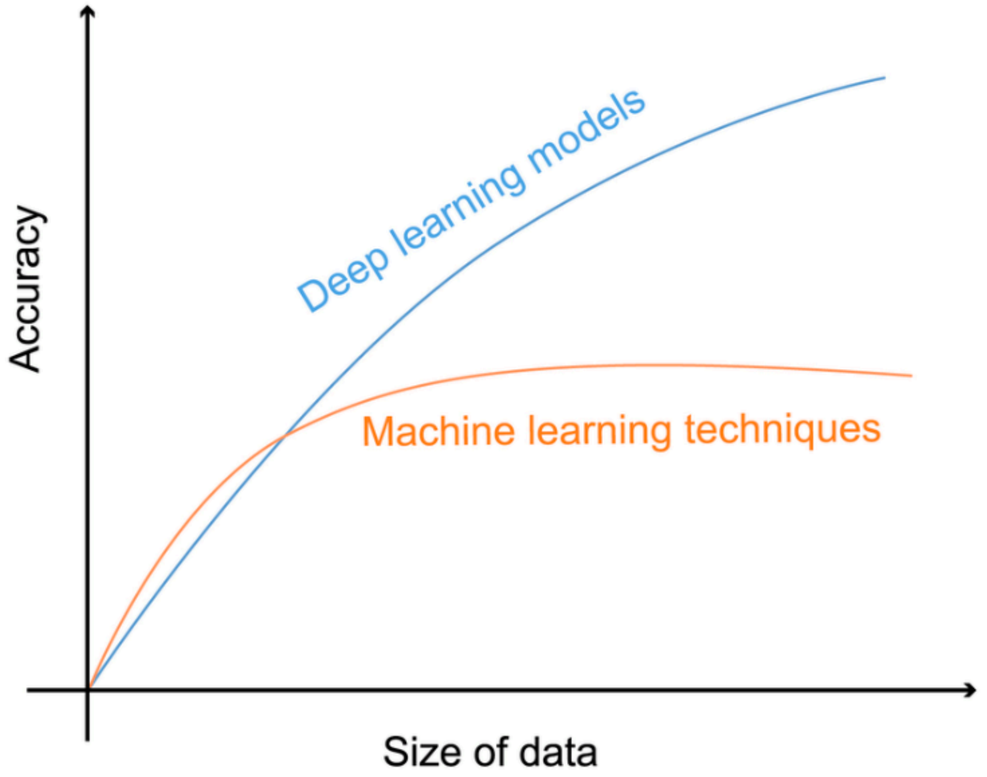


Deep Learning refers to training Artificial Neural Networks, sometimes very large neural networks. Today, the terms Neural Networks and Deep Learning are used almost interchangeably.

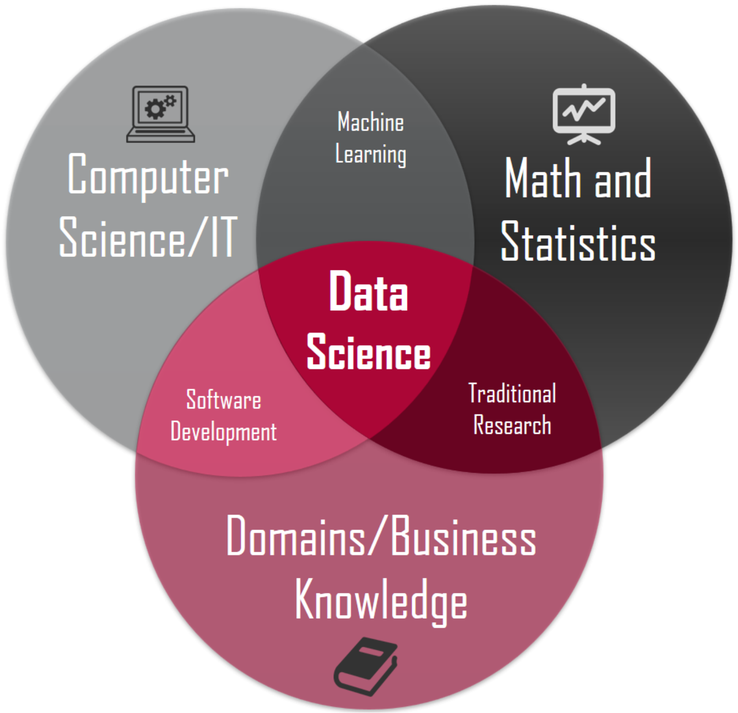




Why DL ?


Accuracy
Deep learning models
Machine learning techniques
Size of data

What is Data Science?



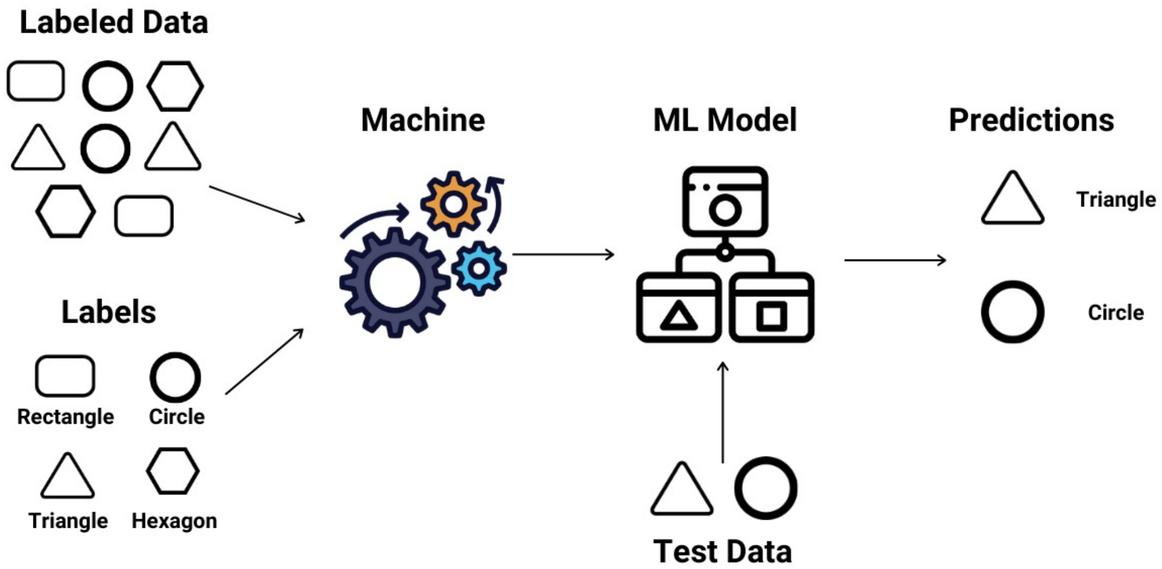
Data science is the domain of study that deals with vast volumes of data using modern tools and techniques to find unseen patterns, derive meaningful information, and make business decisions. Data science uses complex machine learning algorithms to build predictive models. The data used for analysis can come from many different sources and be presented in various formats.

Machine Learning

In general, any machine learning problem can be assigned to one of three broad classifications:

- Supervised Learning
- Unsupervised Learning
- Reinforcement Learning

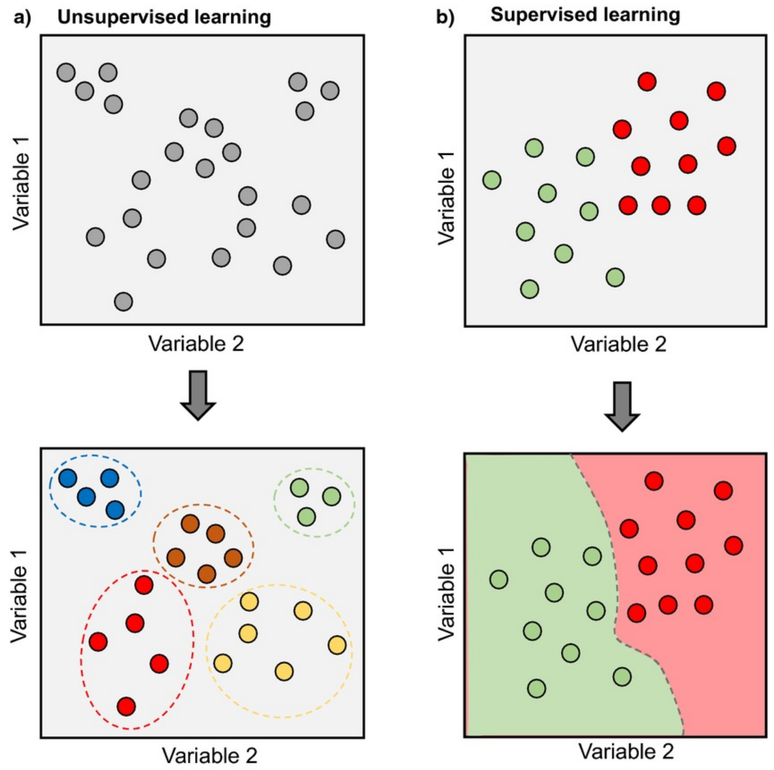
In **supervised learning**, we are given a data set and **already know what our correct output should look like**, having the idea that there is a relationship between the input and the output.
Supervised learning problems are categorized into ***regression*** and ***classification*** problems:

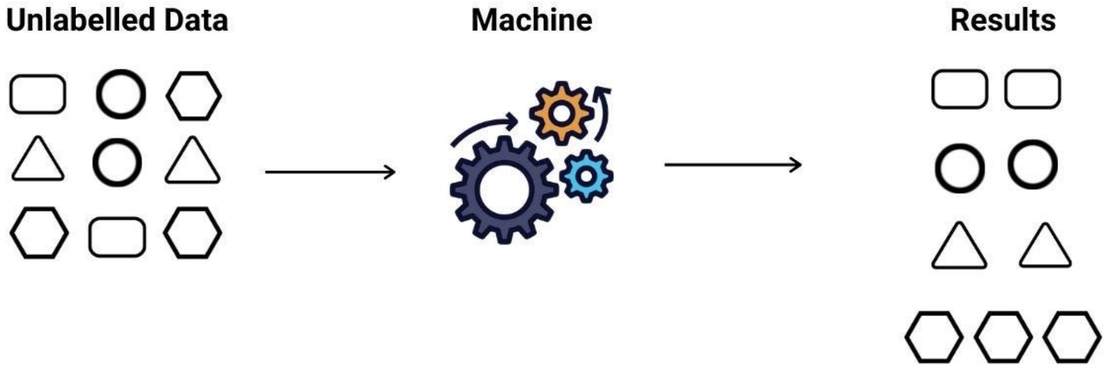


Classification - predict a class label
Regression - predict a number

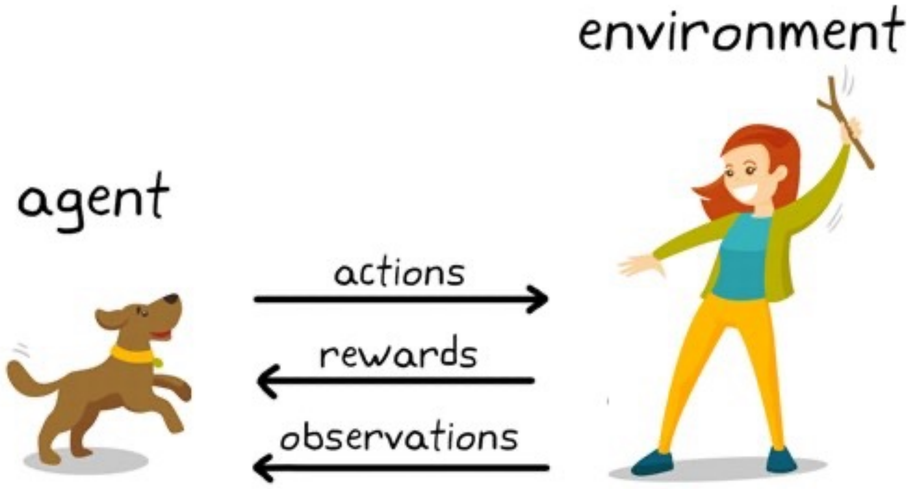
Unsupervised Learning

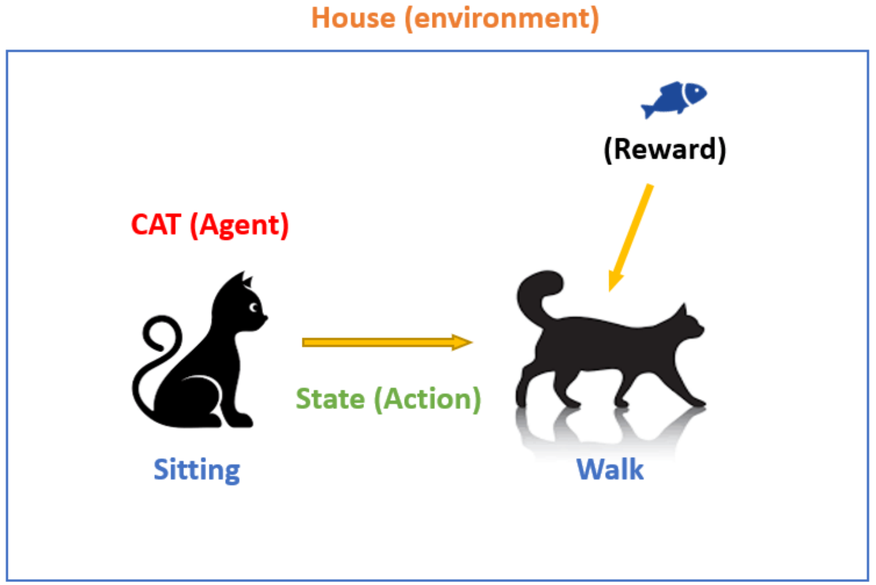
Unsupervised learning, on the other hand, allows us to approach problems with little or no idea what our results should look like. We can derive structure from data where we don't necessarily know the effect of the variables. We can derive this structure by clustering the data based on relationships among the variables in the data.





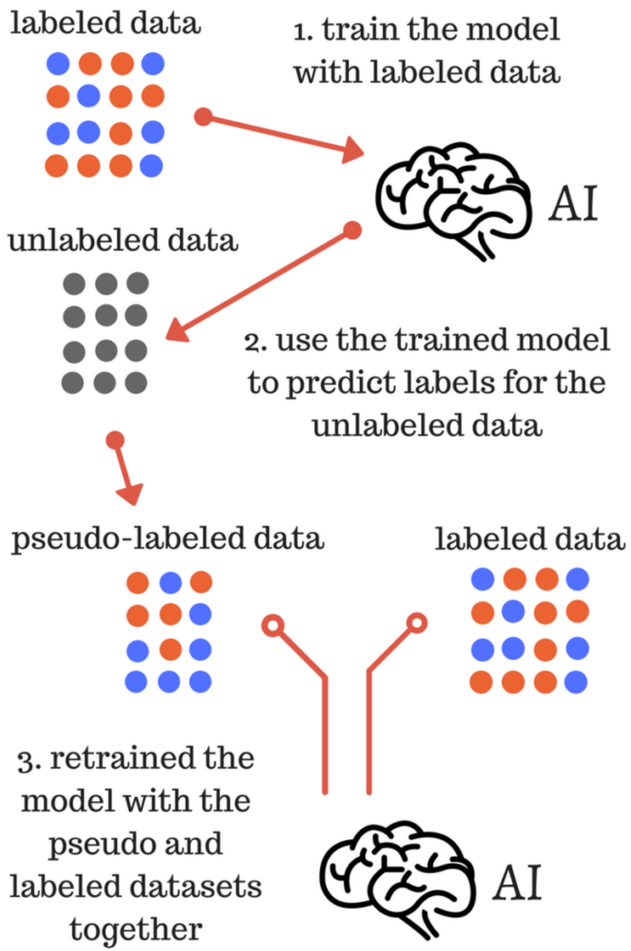
Reinforcement learning is learning what to do—how to map situations to actions—so as to maximize a numerical reward signal. The learner is not told which actions to take, but instead must discover which actions yield the most reward by trying them. In the most interesting and challenging cases, actions may affect not only the immediate reward but also the next situation and, through that, all subsequent rewards. These two characteristics—trial-and-error search and delayed reward—are the two most important distinguishing features of reinforcement learning.





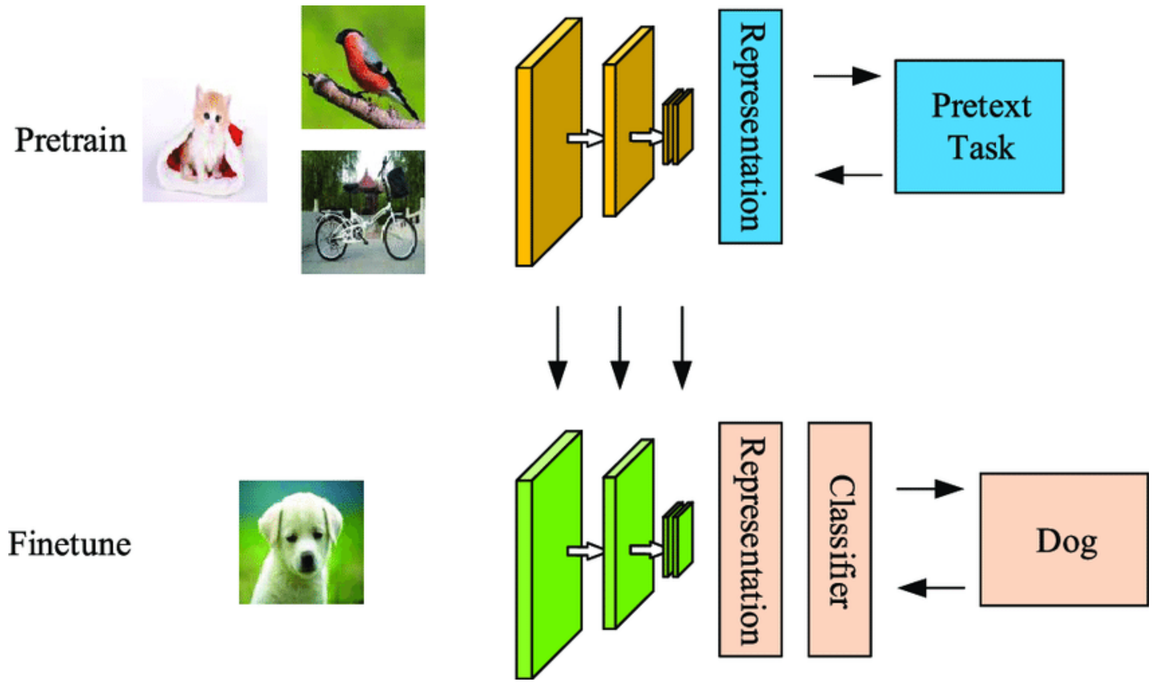
Semi-supervised learning

- Semi-supervised learning (you have some amount of data with labels and other data without labels)



Self-supervised learning

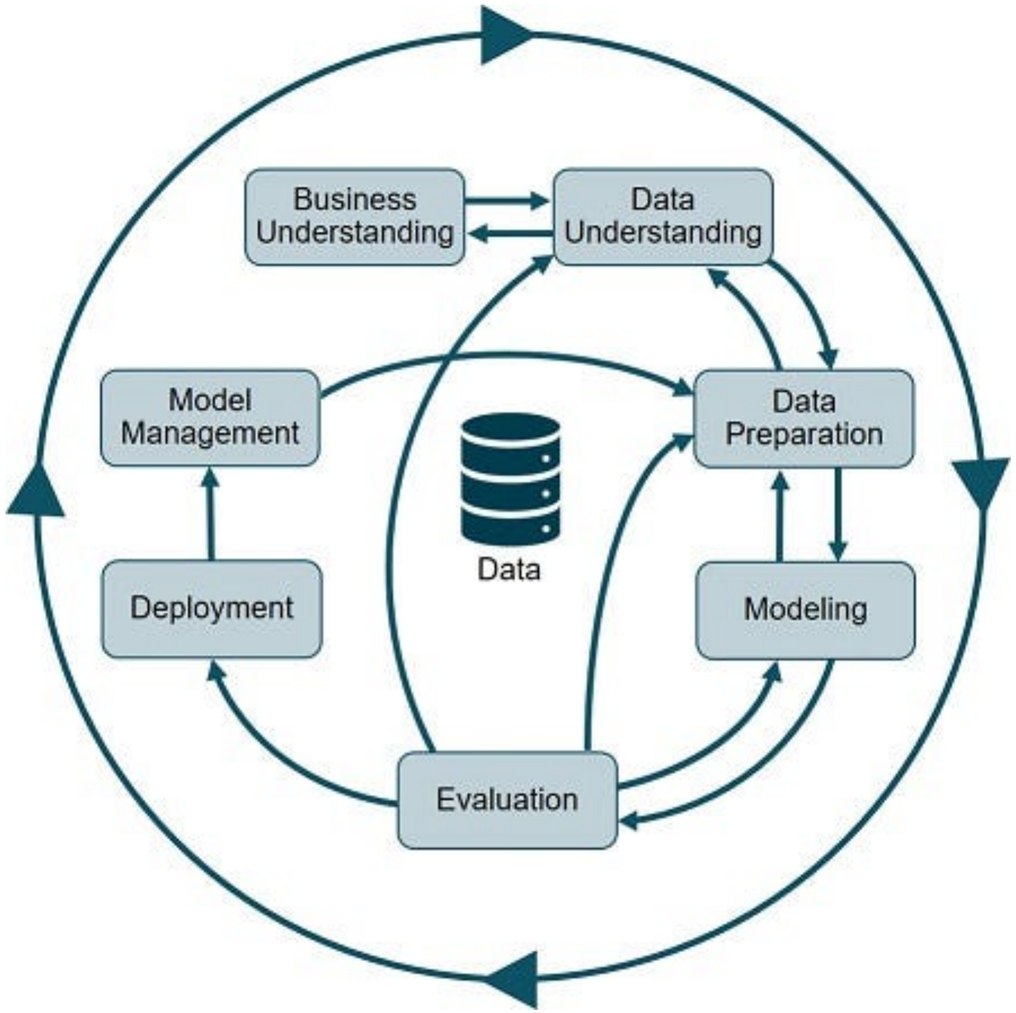
- Self-supervised learning (unsupervised learning problem that is framed as a supervised learning problem)



ML Project Stages

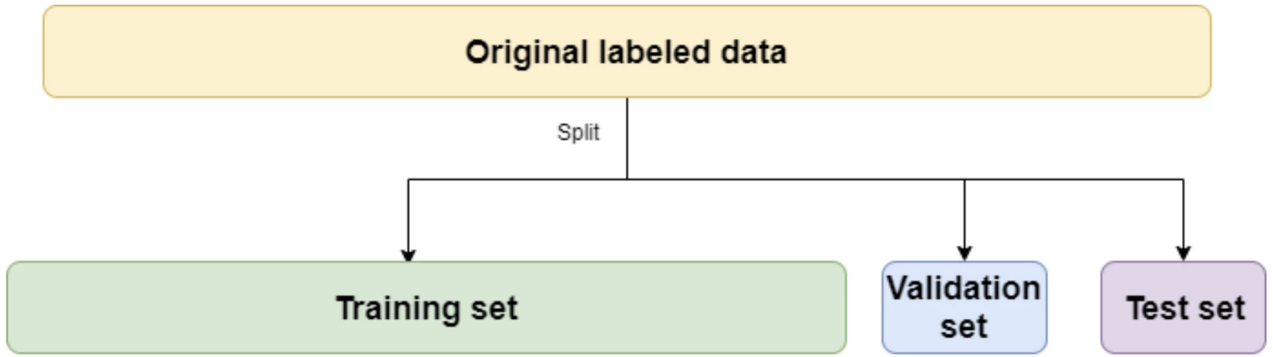
- Initiation – problem understanding, vision, goals
  - Data collection and exploratory data analysis
    - Data cleaning & pre-processing
  - Model development, training and evaluation
    - Model deployment and maintenance

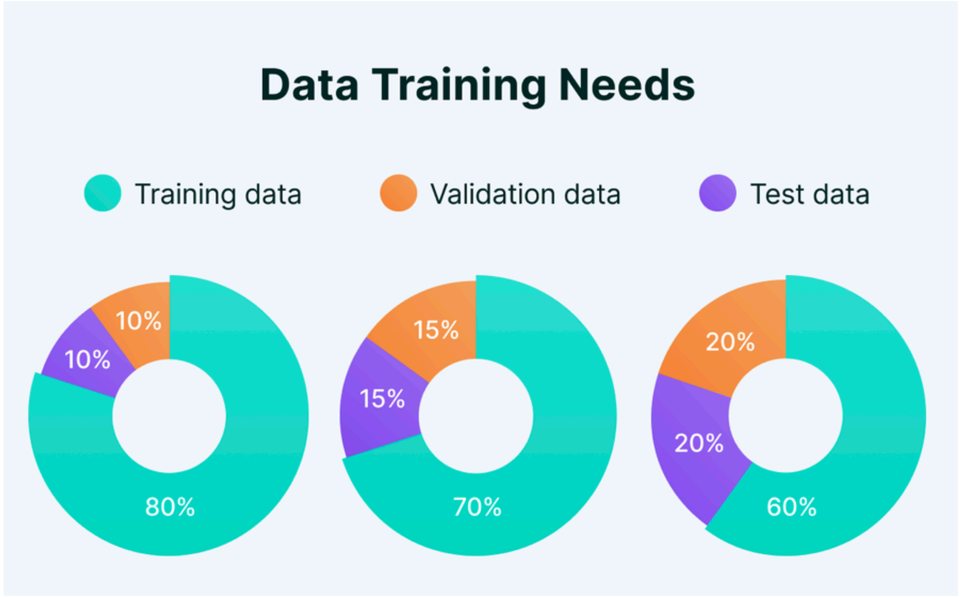
CRISP-DM in ML



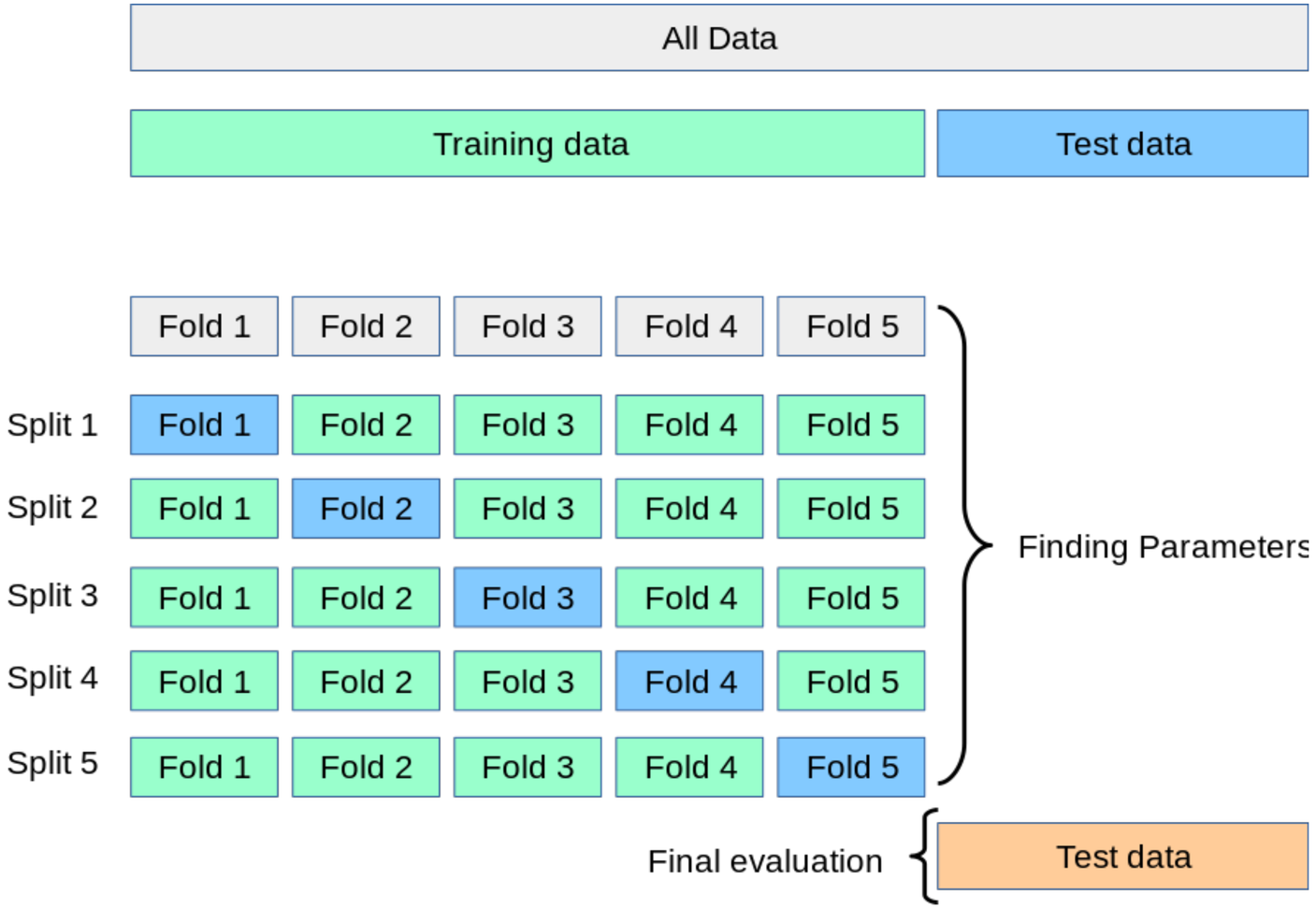
CRoss Industry Standard Process for Data Mining (CRISP-DM)

Data Split


Original labeled data
Split
Training set
Validation set
Test set


Data Training Needs
Training data
Validation data
Test data
10%
10%
80%
15%
15%
70%
20%
20%
60%

Evaluation Strategies

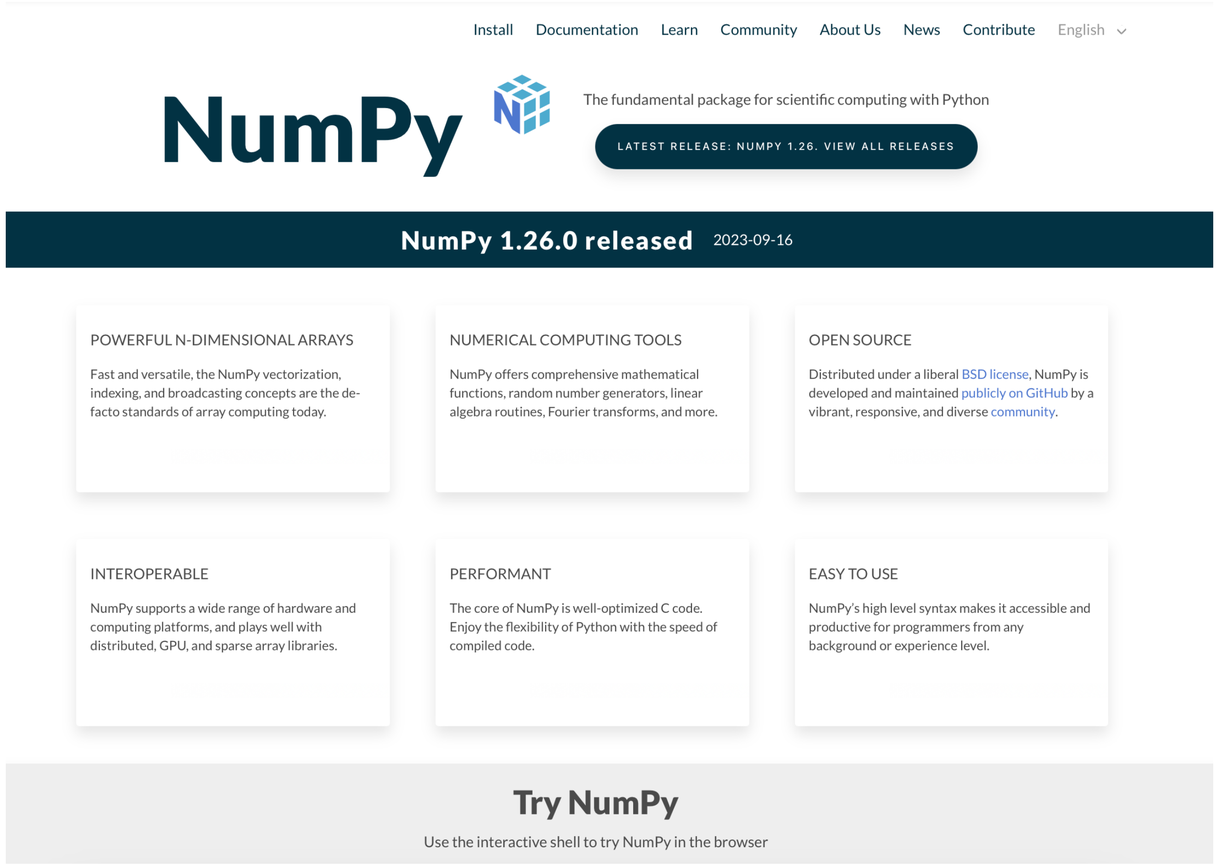


Datasets

- Kaggle
- Papers With Code
- 🤗 HF Datasets

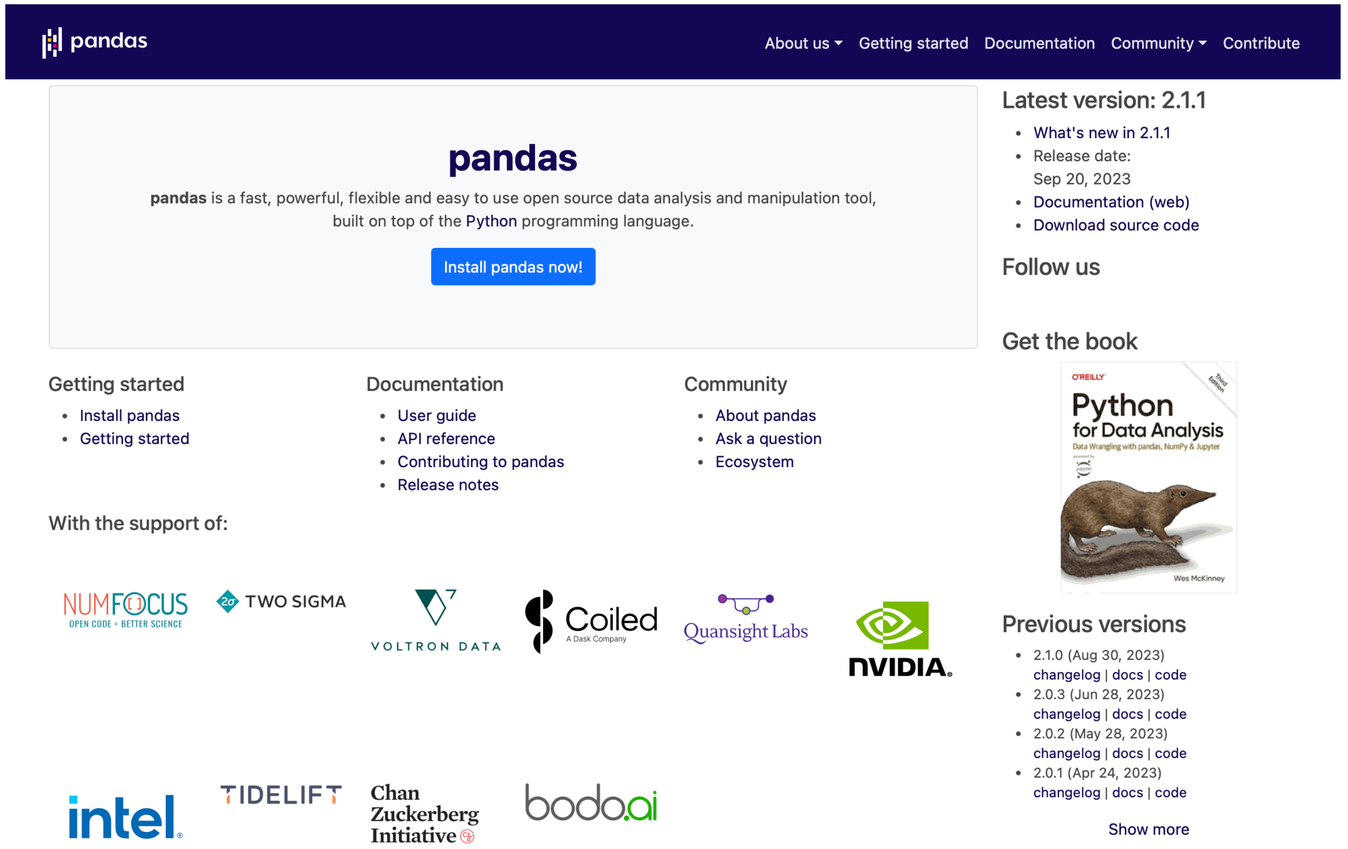
Python ML Tools Basics Review. NumPy

Released as Numeric in 1995, **NumPy** (since 2005) is a library for the Python programming language, adding support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays.



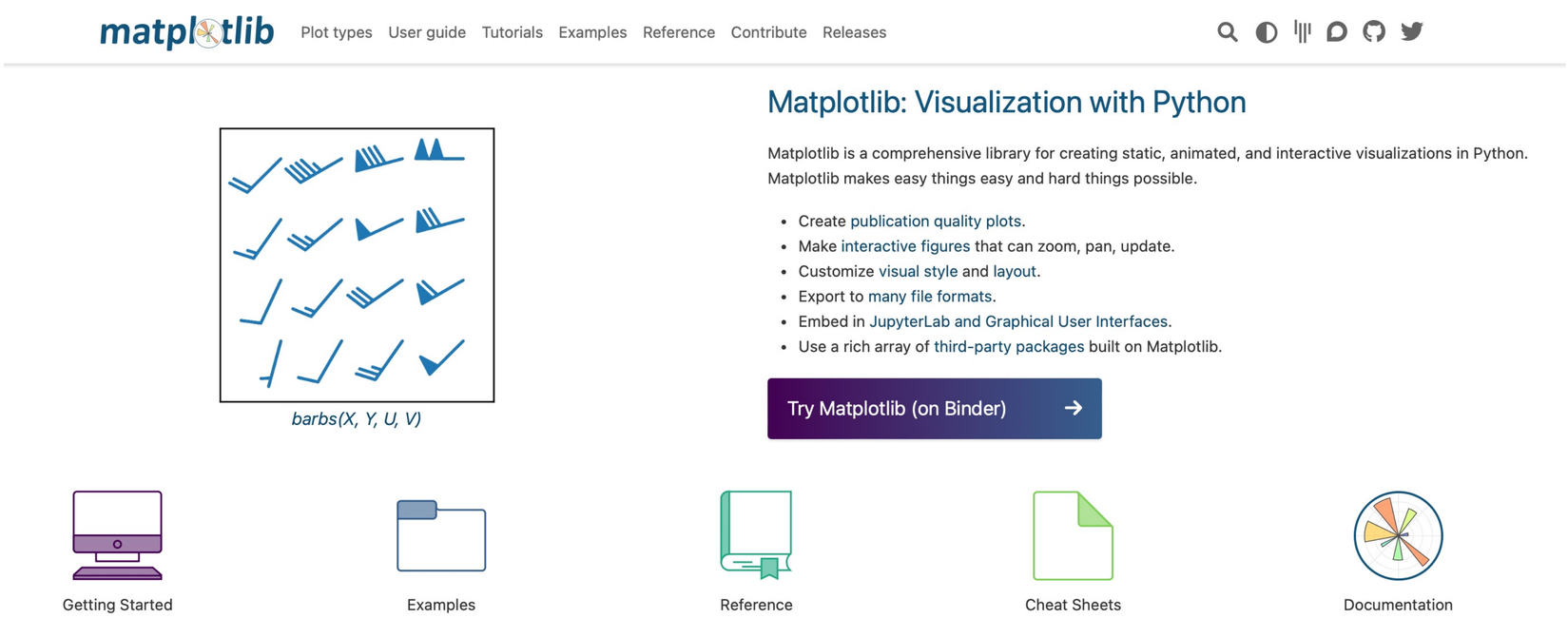
Python Basics Review. Pandas

pandas is a fast, powerful, flexible and easy to use open source data analysis and manipulation tool, built on top of the Python programming language.



Python Basics Review. Matplotlib

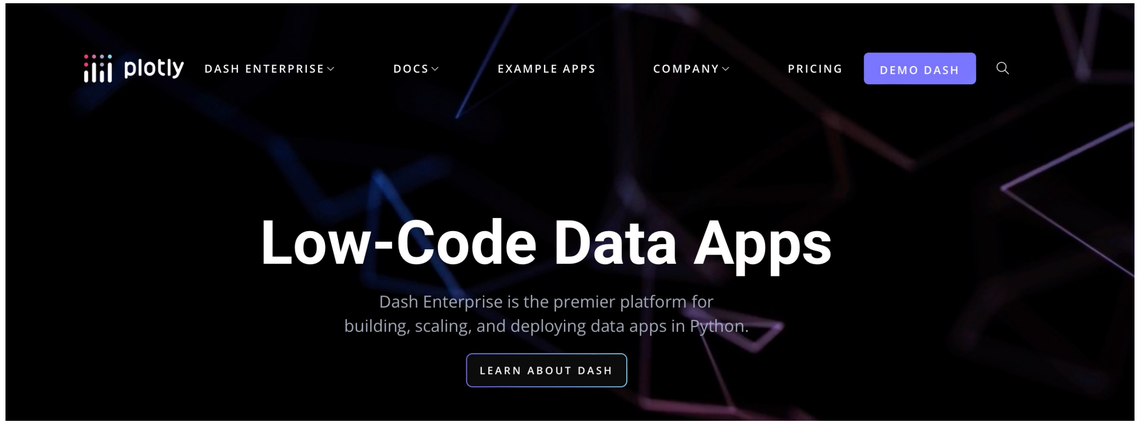
Initially released in 2003, **Matplotlib** is a plotting library for the Python programming language and NumPy.



Seaborn & Plotly

Seaborn is a Python data visualization library based on matplotlib. It provides a high level interface for drawing attractive and informative statistical graphics.

Plotly makes interactive, publication-quality graphs: line plots, scatter plots, area charts, bar charts, error bars, box plots, histograms, heatmaps, subplots, multiple-axes, polar charts, bubble charts, etc.



hillel
IT school

Thanks for your attention