1. Introduction to Machine Learning

Machine Learning definitions:

- Arthur Samuel (1959). Machine Learning: Field of study that gives computers the ability to learn without being explicitly programmed.

- From the book “An Introduction to Statistical Learning”: Statistical learning refers to a vast set of tools for understanding data.

- From the book “Machine Learning in action”:
  - Machine learning is turning data into information.
  - Machine learning lies at the intersection of computer science, engineering, and statistics and often appears in other disciplines.
  - It’s a tool that can be applied to many problems. Any field that needs to interpret and act on data can benefit from machine learning techniques.
  - Machine learning uses statistics.

- From the book “Introduction to Machine Learning” by Ethem Alpaydin:
  - Machine learning is not just a database problem; it is also a part of artificial intelligence.
  - Machine learning also helps us find solutions to many problems in vision, speech recognition, and robotics.
  - Machine learning is programming computers to optimize a performance criterion using example data or experience.

https://youtu.be/1iqh1B1OZAg

1.1 Machine Learning Applications

- Face recognition
- Emotion recognition
- Identifying spam emails
- Weather forecasting
- Object recognition
- Recommendation systems
- Medical data analysis

In addition, computer prorams have been able to beat the best humans in:

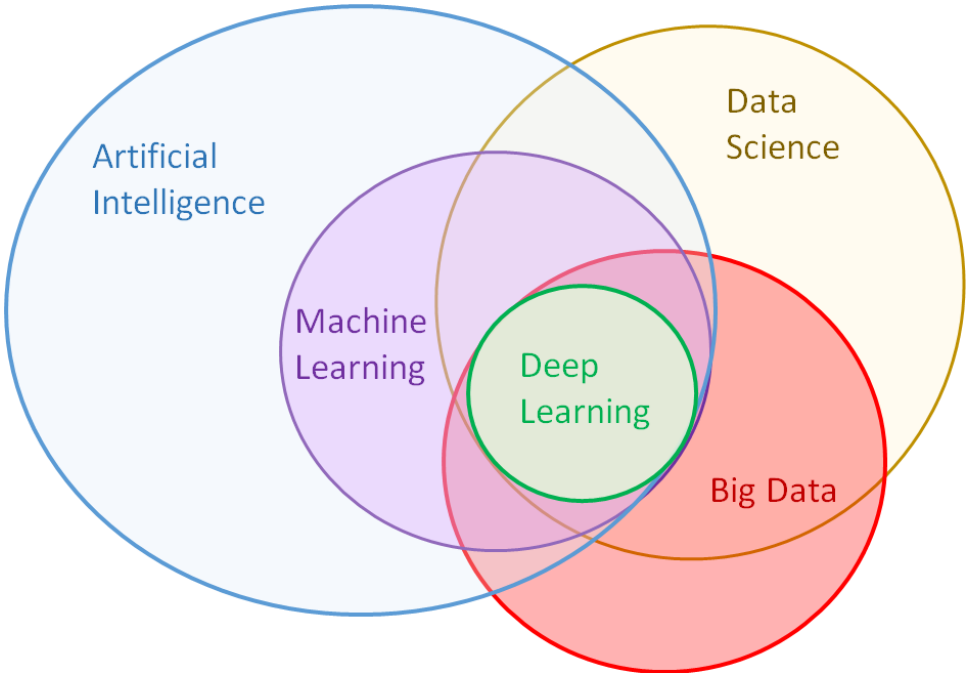
- Chess
- Go

1.2 Why is Machine Learning now capturing much attention?

Because there are:

- A massive amount of data. See the images
- Computer resources (hardware) are cheaper, faster, and more powerful
- Better understanding of algorithms
- We have platforms to share code (for example, GitHub)

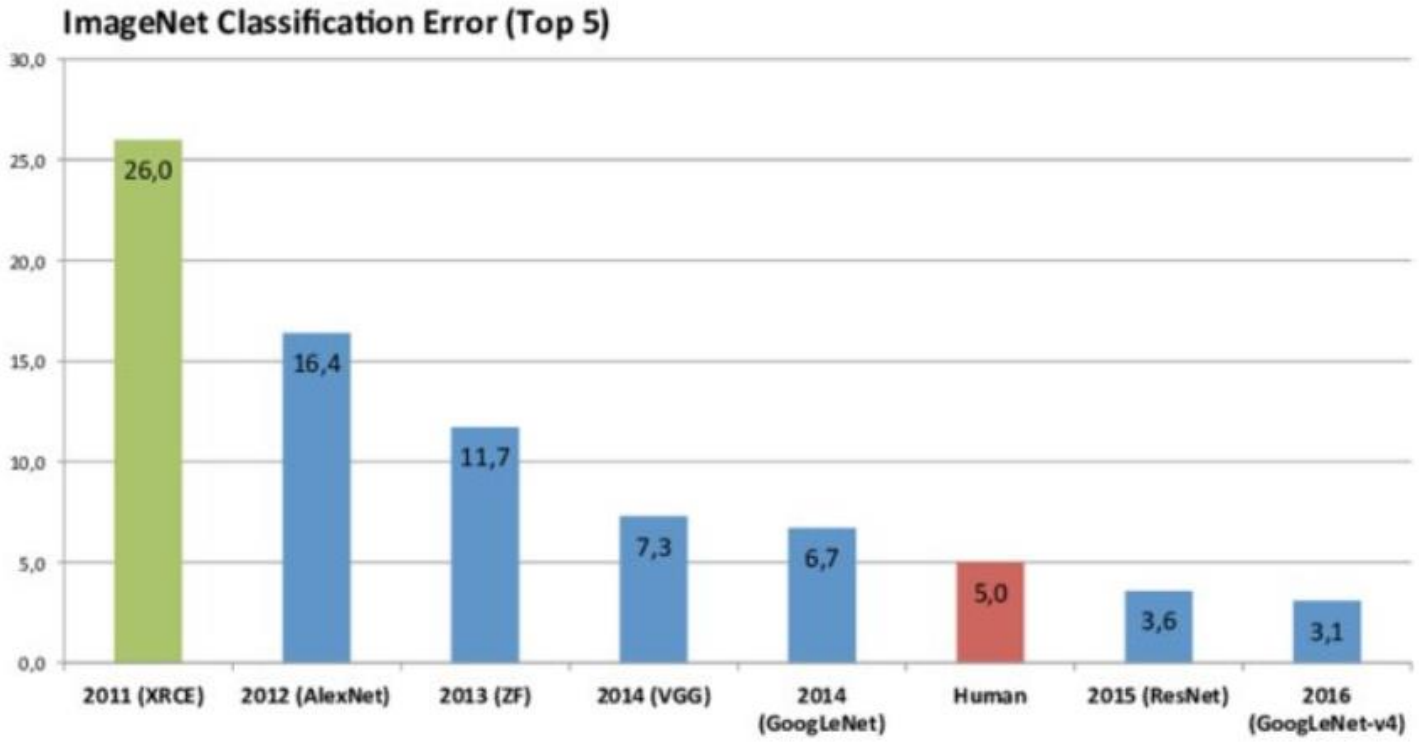
1.3 Some concepts: AI, Data Science, Machine Learning and Deep Learning



- **Artificial Intelligence** is a subfield of computer science. It is the ability of a digital computer to perform tasks commonly associated with intelligent beings.
- **Machine learning** is a branch of Artificial Intelligence. The goal is to turn data into information.
- **Deep learning** is one kind of machine learning (neural networks) that's very popular now. It has been given very impressive results. It needs many data and computational resources to work.
- **Data science** deals with unstructured and structured data. It is a field that comprises everything related to data cleaning, preparation, and analysis. It combines statistics, mathematics, programming, problem-solving, and capturing data in ingenious ways.

1.4 Recent amazing applications with the use of Deep Learning

ImageNet is a challenge that consists of recognizing a target of **1,000 different categories**. The dataset is composed of more than **1.2 million images**. It had been a difficult problem for computers until 2015, where AI algorithms' results improve human results. The following image shows the error of the challenge over the years.



Other applications are:

- Multiple object recognition
- Image segmentation
- Image description
- Face Aging
- Interpolation of faces
- Art style transfer
- Self driving cars

1.5 Programming languages

	**Python** • Expressive • General purpose computer programming language • Interpreted • Free and open source • There are a lot scientific libraries. For example: Machine Learning (numpy, scipy, scikit-learn) and Computer vision (OpenCV)
	**R** • It is focus on statistics. • Free • There are Machine Learning libraries.
	**Matlab** • It is focus on mathematics. • Programming language: M • It is not free **GNU Octave** is an open source version of Matlab
	**Java/Weka** • Java: General purpose computer programming language. • Weka: Collection of algorithms of Machine Learning for Java.

1.6 Public data sources for Machine Learning

http://archive.ics.uci.edu/ml/

https://competitions.codalab.org/

https://www.kaggle.com/

1.7 Types of algorithms in Machine Learning

- **Exploratory data analysis** is an approach to analyzing data sets to summarize their main characteristics, often using statistical graphics and other data visualization methods. It is recommended to perform an EDA before fitting learning models.

- **Supervised Learning** (labeled data): It creates models where a variable (or several variables) guides the learning process.

    - **Regression**. Labels are continuous. Examples: weather forecasting, and grades predictions.

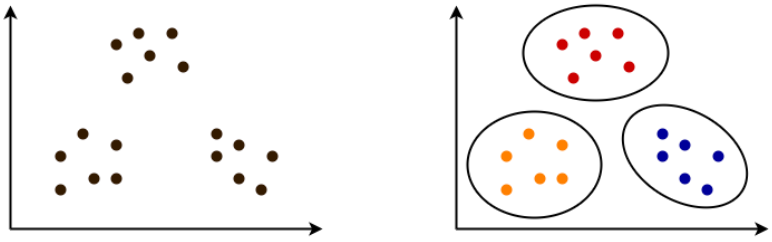
fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	density	pH	sulphates	alcohol	quality
7.4	0.7	0	1.9	0.076	0.9978	3.51	0.56	9.4	5
7.8	0.88	0	2.6	0.098	0.9968	3.2	0.68	9.8	5
7.8	0.76	0.04	2.3	0.092	0.997	3.26	0.65	9.8	5
11.2	0.28	0.56	1.9	0.075	0.998	3.16	0.58	9.8	6
7.4	0.7	0	1.9	0.076	0.9978	3.51	0.56	9.4	5
7.4	0.66	0	1.8	0.075	0.9978	3.51	0.56	9.4	5
7.9	0.6	0.06	1.6	0.069	0.9964	3.3	0.46	9.4	5
7.3	0.65	0	1.2	0.065	0.9946	3.39	0.47	10	7
7.8	0.58	0.02	2	0.073	0.9968	3.36	0.57	9.5	7

    - **Classification**. Labels are discrete. Examples: image classification, disease dignossis, digit recognition, and spam detection.

sepal_length	sepal_width	petal_length	petal_width	Iris_class
5	2	3.5	1	versicolor
6	2.2	4	1	versicolor
6.2	2.2	4.5	1.5	versicolor
6	2.2	5	1.5	virginica
4.5	2.3	1.3	0.3	setosa
5.5	2.3	4	1.3	versicolor

- **Unsupervised learning** (unlabeled data)

    - **Clustering**. Focus on grouping the data. Examples: recommendation systems, customers segmentation.



    - **Dimensionality reduction**. Focus on reducing the number of features of variables.