



ieeta instituto de engenharia electrónica e telemática de aveiro

# **Foundations of Machine Learning**

(winter semester 2021/2022)

## **LECTURE 1 : INTRODUCTION**

**Petia Georgieva**  
**(petia@ua.pt)**

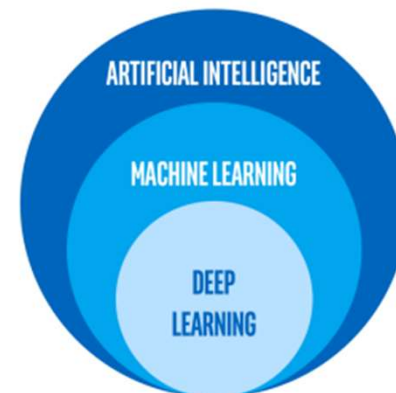
# AI - the new Electricity

**Artificial Intelligence (AI) will influence every industry .**

McKinsey estimated 13 trillion dollars of global GDP value creation by 2030 due to AI.

**Software Industry** (strongly affected by AI) : Web Search; On-line Advertysing; Language translation; Social Media

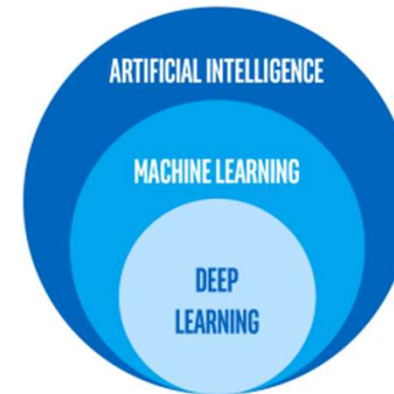
**Non-Software Industry** (still long way to go): Manufacture, Agriculture, Retail, Transportation, Logistics, etc.



# PROGRAM

## **Supervised learning**

Linear (univariate/ multivariate) regression  
Logistic regression. Regularization  
Artificial Neural Networks (ANN)  
Support Vector Machines (SVM)  
Decision Tree (DT);  
Naive Bayes classifier  
k-Nearest Neighbor (k-NN) classifier



## **Unsupervised learning**

K-means clustering  
Data dimensionality reduction  
Principal components analysis (PCA)  
Multivariate Gaussian approach for Anomaly Detection

## **Deep Learning**

Deep Learning architectures :  
CNN (Convolutional Neural Networks);  
YOLO (You Only Look Once) neural network  
LSTM (Long Short Term Memory) neural network

# Evaluation

Lectures & labs: 3 hours per week.

## **Practical component - 50% of the final grade**

Practical component consists of 2 projects, developed in a group of two students.

**Project 1** is evaluated based on a submitted report (IEEE format) and a short (10-15 min.) oral presentation.

**Project 2** is evaluated based on a submitted report (IEEE format).

The students are encouraged to use Latex text editor.

Overleaf is a convenient platform for collaborative writing and publishing using Latex (<https://www.overleaf.com/>) .

## **Theoretical Component – 50% of the final grade**

Final individual exam.

# Why ML ?

- Grew out of work in Artificial Intelligence and increasing computational resources.
- Exponential growth of data – need for data mining (IoT, medical records, biology, engineering, etc.)
- Applications can't be explicitly programmed by hand.
  - ✓ Autonomous driving;
  - ✓ Computer Vision;
  - ✓ Natural Language Processing (Speech recognition, Machine translation)
  - ✓ User behaviour monitoring (Sentiment classification, Video activity recognition) .

# AI/ML is advancing due to:

- **Data** (lots of sensors) + Computational resources
- **Talents** (Easy to access AI courses on MOOC, Coursera, University)
- **Ideas** (100 AI papers/per day)
- **Tools** (open source platforms Pytorch, Keras, Tensorflow, mxnet, etc.)

# A bit of history

- **1950**, Alan Turing: "Computing Machinery and Intelligence" define the question "Can machines think?"  
=>Turing test.
- **1956** –The field of Artificial Inteligente (AI) formally established at the conference in Dartmouth College.
- **1959**, Arthur Samuel: “ Field of study that gives computers the ability to learn without being explicitly programmed ”.
- **1998**, Tom M. Mitchell: “ Can the computer program learn from experience ? “.

# Machine Learning – “definition”

„A computer program is said to learn from experience  $E$  with respect to some task  $T$  and some performance measure  $P$ , if its performance on  $T$ , as measured by  $P$ , improves with experience  $E$ .“  
(**T. Mitchell 1998**)

- **Given**

- a task  $T$  (e.g. classify spam/regular emails)
- a performance measure  $P$  (weighted sum of mistakes)
- some experience  $E$  with the task (e.g. hand-sorted emails)

- **Goal**

- generalize the experience in a way that allows to improve the machine performance on the task



# Learning to classify documents



## Web page:

Company, Personal, University, etc.

## Articles:




Sport, Political, History, etc.

# Computer Vision

Learning to detect & recognize faces



# Computer Vision Tasks

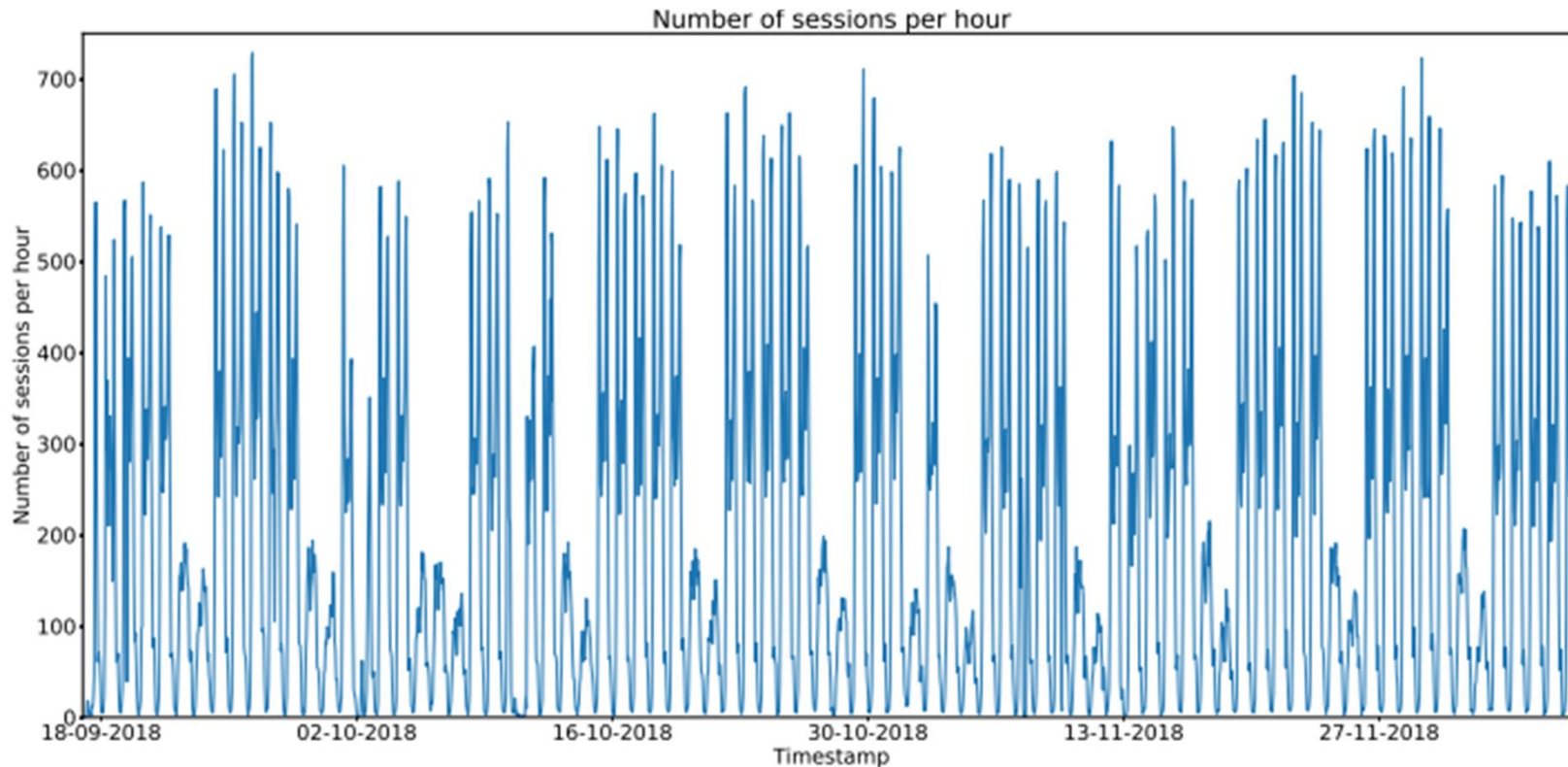
Image classification	Classification & Localization	Detection
	 $b_x, b_y, b_h, b_w$	

**Image classification:** input a picture into the model and get the class label (e.g. person, bike, car, background, etc.)

**Classification & localization:** the model outputs not only the class label of the object but also draws a bounding box (the coordinates) of its position in the image.

**Detection:** the model detects and outputs the position of several objects.

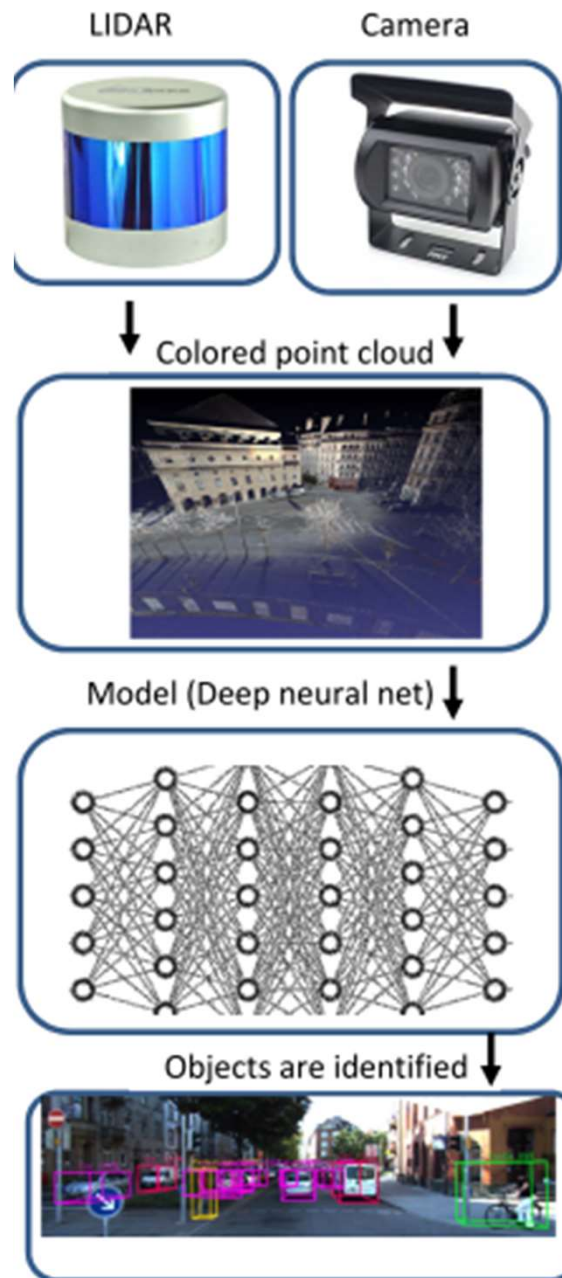
# Time Series (TS) Forecasting



Time Series - collection of data points indexed based on the time they were collected . Most often, data are recorded at regular time intervals.

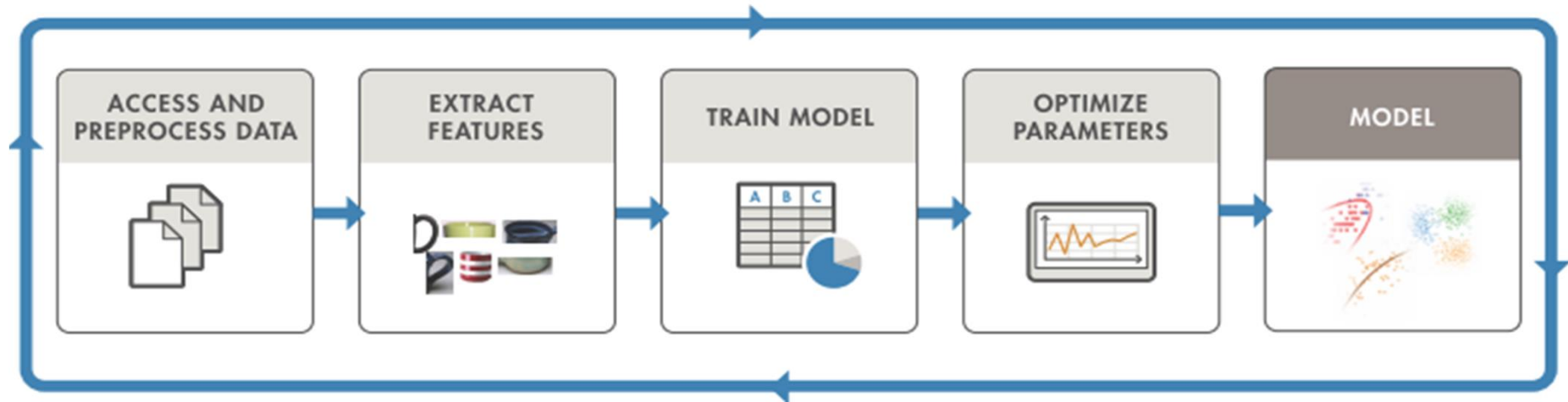
Based on past data, predict future trends, seasonality, anomalies, etc.

# Object Detection (sensor fusion)

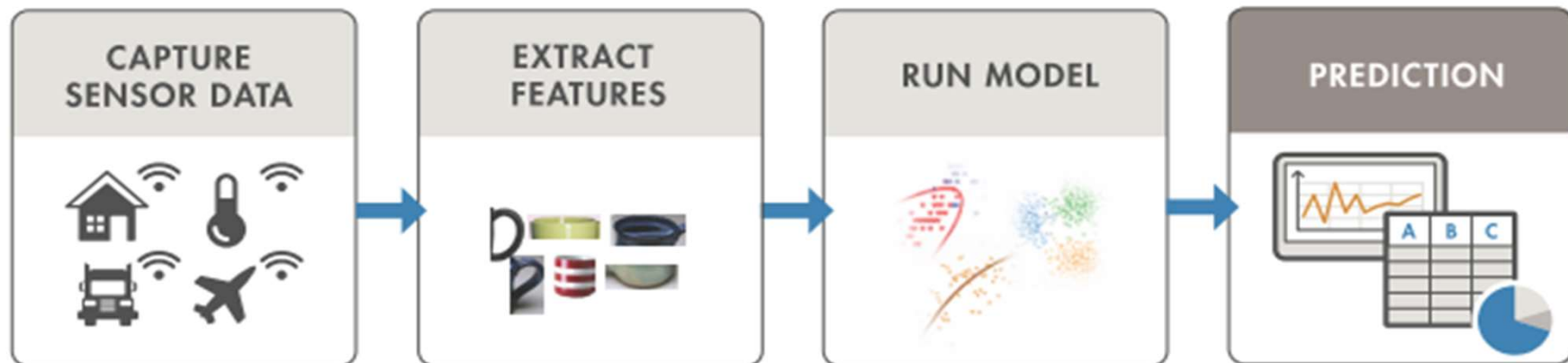


# ML workflow

**Train:** Iterate until achieve satisfactory performance (**usually off-line**)



**Predict:** Integrate trained models into applications ( **on-line** )





# Machine Learning Approaches

## **Supervised Learning**

Given examples with “correct answer” (labeled examples)  
(e.g. given dataset with spam/not-spam labeled emails)

## **Unsupervised Learning**

Given examples without answers (no labels).

## **Deep Learning**

Automatically extract hidden features (in contrast to hand-crafted features). Need a lot of data (Big data) . Need for very high computational resources (GPUs).

## **Reinforcement Learning**

On-line (on the fly) learning, by trial and error.

# Supervised Learning

Requires labeled data (examples with “correct answer”).

**Regression:** The Labels are real numbers.

**Ex.** Predict the house price (output) based on data for the house area and number of bedrooms (features).

Living area (feet <sup>2</sup> )	#bedrooms	Price (1000\$s)
2104	3	400
1600	3	330
2400	3	369
1416	2	232
3000	4	540
⋮	⋮	⋮

**Classification:** The Labels are integer numbers (class 1, class 2, etc.)

**Ex.** Predict normal (0) or abnormal (1) state of data center computers:

**Features:** memory use of computer ; number of disc accesses /sec; CPU load ; network traffic; silence

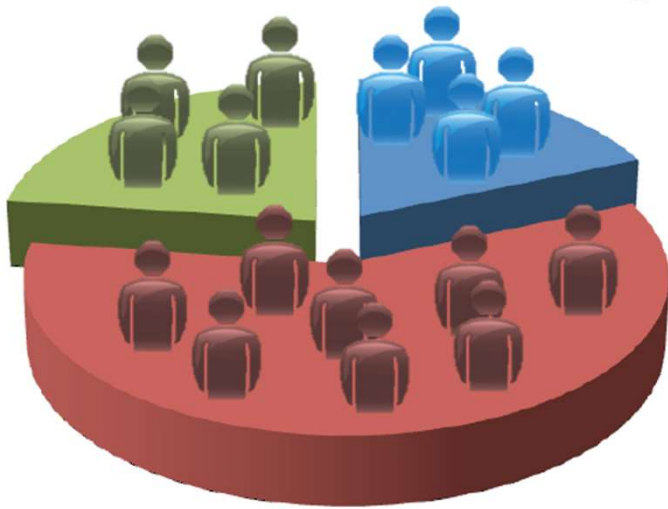


# Unsupervised Learning

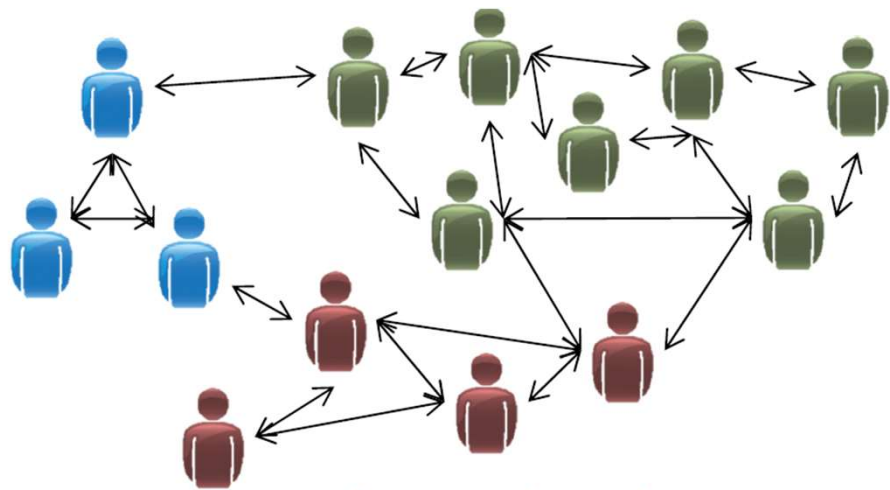
Given unlabeled data (NO answers)

**Features:** education, job, age, marital status, etc.

Market segmentation



Social network analysis



**Clustering:** Given a collection of examples (e.g. user profiles with a number of features). Each example is a point in the multidimensional space of features. Find a similarity measure that separates the points into clusters.

-K-means clustering

# Why Deep Learning ?

Hardware get smaller.

Sensors get cheaper, widely available IoT devices with high sample-rate.

Data sources: sound, vibration, image, electrical signals, accelerometer, temperature, pressure, LIDAR, etc.

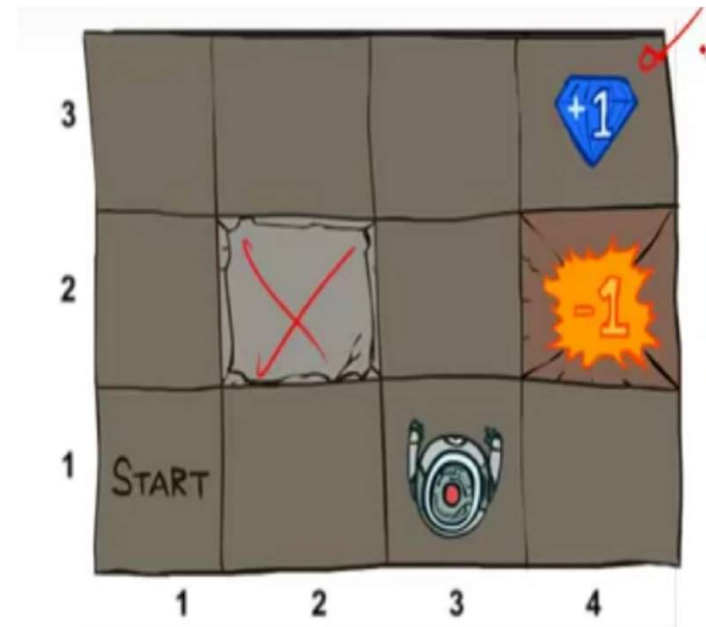
**Big Data:** Exponential growth of data, (IoT, medical records, biology, engineering, etc.)

How to deals with **unstructured data** (image, voice, text, EEG, ECG, etc.) =>  
What are the best feature ?

Deep Neural Networks: first extract (automatically) the hidden features, then solve ML tasks (classification, regression)

# Reinforcement Learning

On-line learning by taking actions  
and getting rewards/penalties.  
intelligent robotics =>



# Little vs. Lots of Data

Most ML applications lay somewhere in this spectrum:

**Little data** <-----> **Lots of data**

We have lots of data for speech recognition;  
Reasonable large data for image recognition (cats or dogs) ;  
and much less data for object detection (bounding boxes) .

**If Lots of data:** the best way to get good performance is to build deep models (several layers), playing with network architectures, but less hand-engineering.

**If Little data:** the best way to get good performance is hand-engineering – very difficult and skilful task that requires a lot of inside (expert) knowledge.

# Data Types

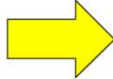
## 1. Numeric (Quantitative) features

- Integer numbers
- Floats (decimals) - temperature, height, weight, humidity, etc.

## 2. Boolean – True/False

## 3. Categorical features - gender, days of the week, seasons, country of birth, colors, etc.

How to deal with categorical features ? - One-hot encoding (1,0) transforms  $n$  categories into  $n$  features

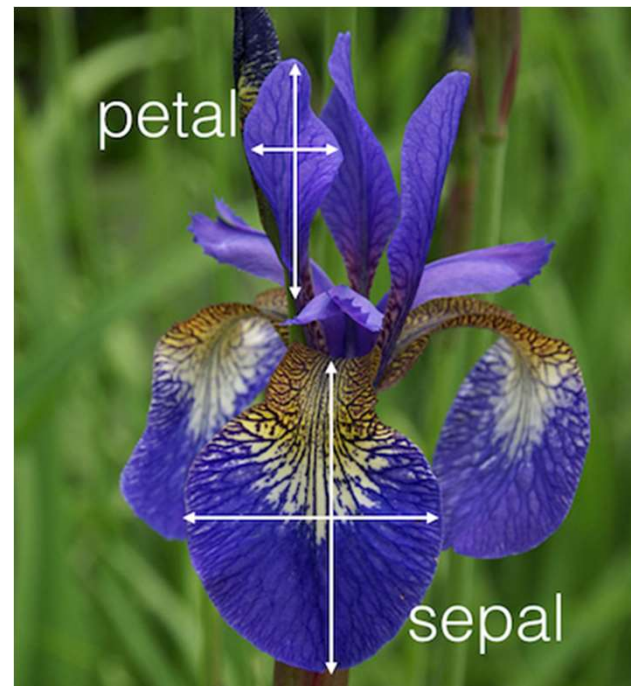


Color
Red
Red
Yellow
Green
Yellow

Red	Yellow	Green
1	0	0
1	0	0
0	1	0
0	0	1

# Iris Plant data

- Iris Plant data – benchmark dataset for illustration of ML methods.
  - UCI Machine Learning Repository  
<http://www.ics.uci.edu/~mlearn/MLRepository.html>
  - 3 flower types (classes):
    - Setosa
    - Virginica
    - Versicolour
  - 4 attributes (features)
    - Sepal width and length
    - Petal width and length

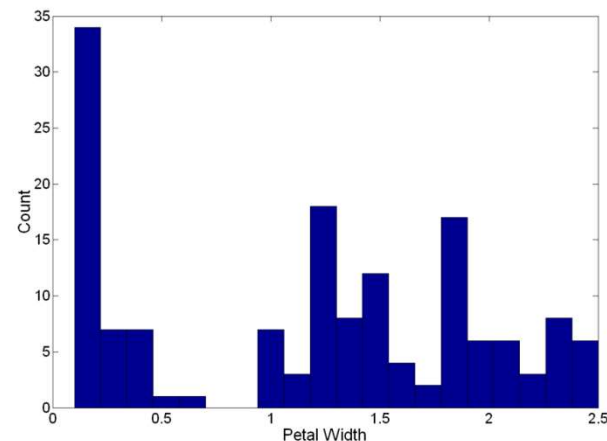
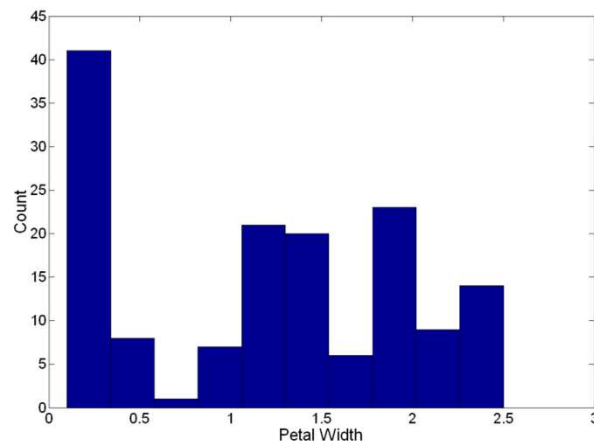


# Data Visualization (1)

- **Histograms**

- Show the distribution of values of a single feature
- Divide the range of values of a single feature into bins and show bar plots of the number of examples in each bin.
- Histogram shape depends on the number of bins

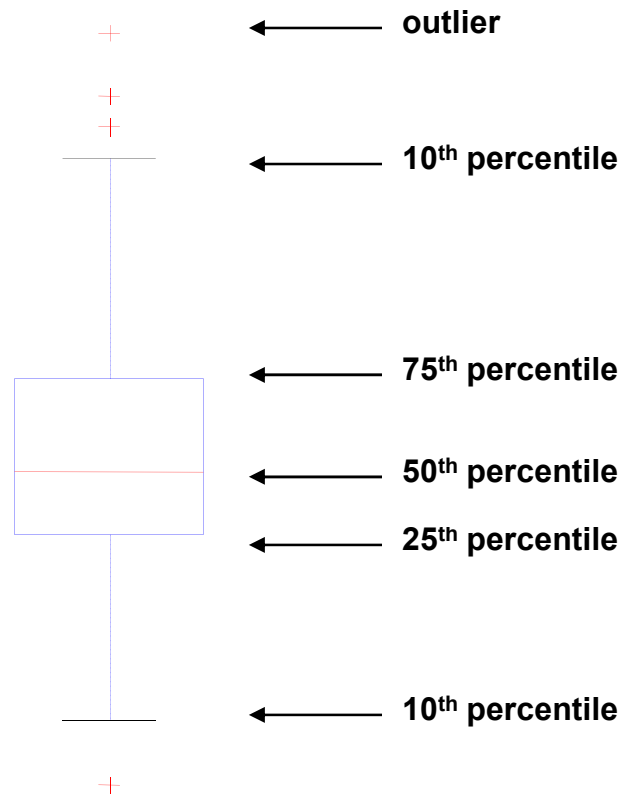
- Example: Petal Width (10 and 20 bins, respectively)



# Data Visualization (2)

- **Box Plots**

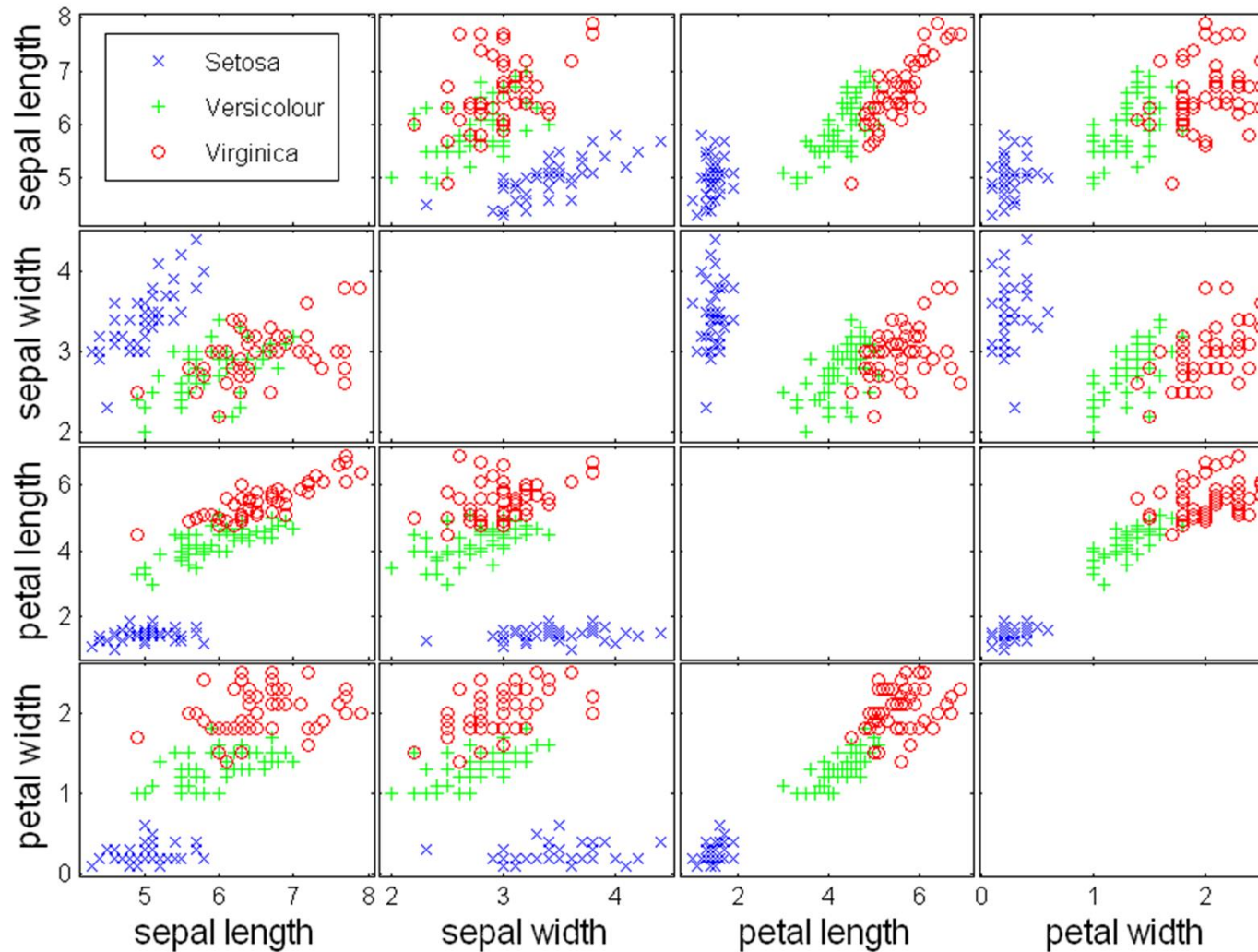
- Another way of displaying the distribution of data





# Data Visualization (3)

## Scatter Plot Array



# RECOMMENDED BIBLIOGRAPHY

## Books:

- Tom Mitchell, Machine Learning. McGraw-Hill, 1997.
- Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems. Aurélien Géron. O'Reilly
- Python Machine Learning. Sebastian Raschka. Packt Publishing
- F. Chollet. Deep Learning in Python, Manning, 2018.
- I. Goodfellow, Y. Bengio, A. Courville. Deep Learning, MIT Press, 2016.

## Courses:

- <http://cs229.stanford.edu/>
- MOOC (Massive Open Online Courses) e.g. <https://www.coursera.org/>

# Food for Thought

Ethical issues of Artificial Intelligence /ML !!!

How ethics and human values could be embedded into the ML algorithms used in AI ?

One hundred year study of AI, Stanford University, August, 2016  
<https://ai100.stanford.edu/>

Google, Facebook, Amazon, IBM and Microsoft created Partnership on Artificial Intelligence to Benefit People and Society

<https://www.partnershiponai.org/> => **Socially responsible AI**

# AI challenges for the society

## **Presence of AI** in mainstream technologies:

- robot motion planning and navigation
- computer vision (e.g. object recognition)
- natural language processing and speech recognition

## **Recent future of AI:**

- autonomous vehicles (e.g. drones, self-driving cars)
- medical diagnosis and treatment
- physical assistance for elderly

## **AI challenges for economy & society:**

- Potential threat to humankind (?)
- AI experts have different opinions (?)
- Jobs are missing due to AI (!)
- Militarized AI is a commonly shared concern (!)

# Are today's AI systems intelligent ?

Michael I. Jordan:

**- Artificial Intelligence—The Revolution Hasn't Happened Yet, 2019**

<https://hdsr.mitpress.mit.edu/pub/wot7mkc1/release/9>

**- Stop Calling Everything AI, Machine-Learning Pioneer Says, 2021**

[https://spectrum.ieee.org/the-institute/ieee-member-news/stop-calling-everything-ai-machinelearning-pioneer-says?fbclid=IwAR2pVqysg0u\\_Mp4BDhxhf-x8T\\_iBPAfOV2yCj70VgabKYlClAPQkIBtK1IU#.YGUEzjDr0JA.facebook](https://spectrum.ieee.org/the-institute/ieee-member-news/stop-calling-everything-ai-machinelearning-pioneer-says?fbclid=IwAR2pVqysg0u_Mp4BDhxhf-x8T_iBPAfOV2yCj70VgabKYlClAPQkIBtK1IU#.YGUEzjDr0JA.facebook)

“People are getting confused about the meaning of AI in discussions of technology trends—that there is some kind of intelligent thought in computers that is responsible for the progress and which is competing with humans. We don't have that, but people are talking as if we do.”

# ANACONDA 3

**1) Install Anaconda 3 for Python 3:**

**<https://www.anaconda.com/distribution/>**

**2) Learn how to use Jupyter Notebook (part of Anaconda)**

**<https://www.dataquest.io/blog/jupyter-notebook-tutorial/>**