Departamento de Eletrónica, Telecomunicações e Informática

Machine Learning

LECTURE 1 : INTRODUCTION

Petia Georgieva
(petia@ua.pt)

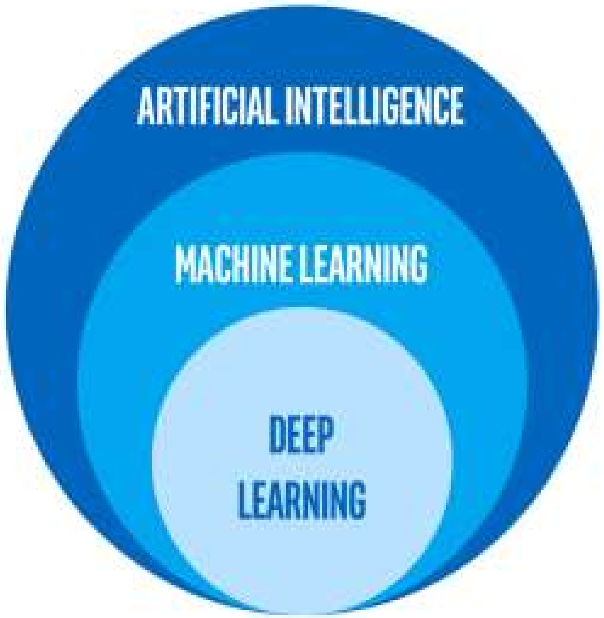
universidade de aveiro

Artificial Intelligence (AI)

AI is a general purpose technology that may influence every industry (similar to electricity, internet) .

AI is based on

Machine Learning (ML) & Deep Learning (DL) algorithms



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)

Deep Learning

Deep Learning architectures :

- CNN (Convolutional Neural Networks);
- LSTM (Lond Short Term Memory) neural network

- Multivariate Gaussian approach for Anomaly Detection
- Recommender Systems

Evaluation

Lectures & labs: 3 hours per week.

<u>Practical component - 50% of the final grade</u>

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

The first project is evaluated based on a submitted report (IEEE format) and a short (10-15 min.) oral presentation.

The second project 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/) .

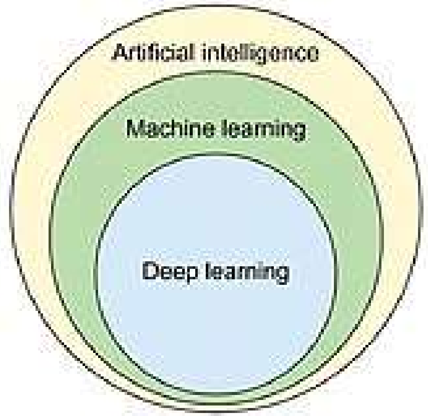
<u>Theoretical Component – 50% of the final grade (</u>Final exam).

universidade de aveiro

Why Machine/Deep Learning

- **Sensors** get cheaper (e.g. widely available IoT devices)

- Exponential **growth of data** – WSN/IoT, medical records, biology, engineering, etc.
- **Data sources**: sound, vibration, image, electrical signals, accelerometer, temperature, pressure, LIDAR etc.
- Increasing **computational resources**.



- **Complex Applications:**
  - ✓ Autonomous driving;
  - ✓ Intelligent robotics;
  - ✓ Computer Vision;
  - ✓ Natural Language Processing (Speech recognition, Machine translation)
  - ✓ 5G+ networks

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

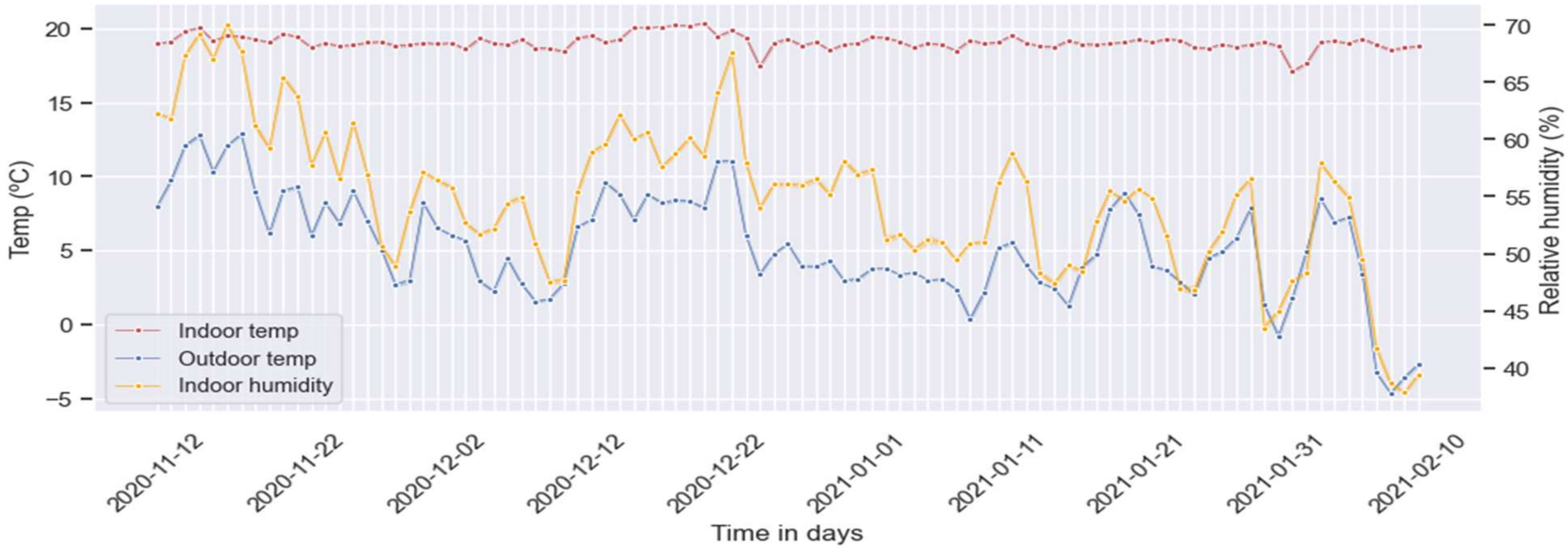
Image classification	Classification & Localization	Detection
	 b_x, b_y, b_h, b_w	

Image classification: input a picture into ML/DL 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.

Object Detection: outputs the position and labels of several objects.

Time Series (TS) Data



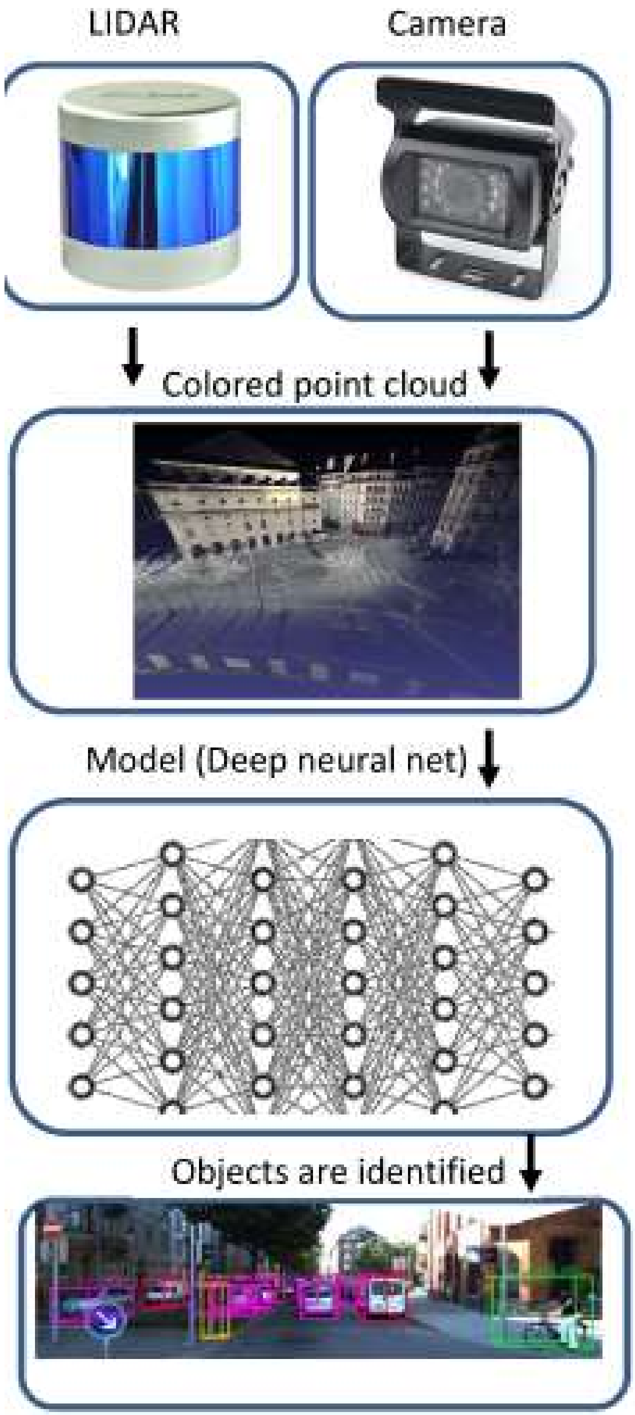
Time Series (TS) - collection of samples recorded at a sequence of time intervals

TS forecasting (prediction) => based on past samples, predict future trends, seasonality, anomalies, etc. Many applications:

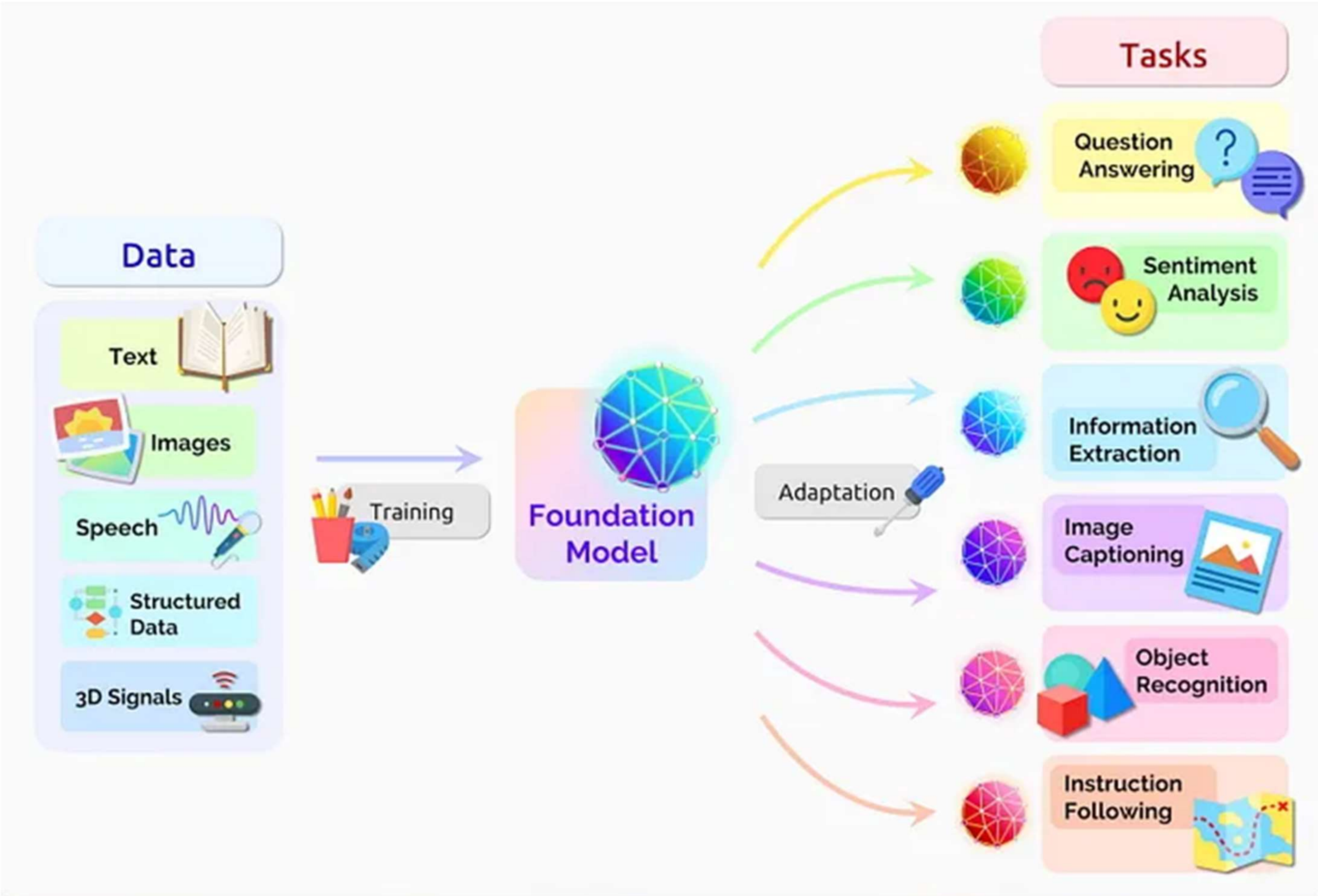
- Key Performance Indicators (KPIs) : network traffic prediction
- Smart Homes – predict indoor temp., heating set-point, thermal comport
- Weather forecast – heat waves, flooding
- WSN physical layer – channel modelling / estimation

universidade de aveiro

Multimodal Object Detection

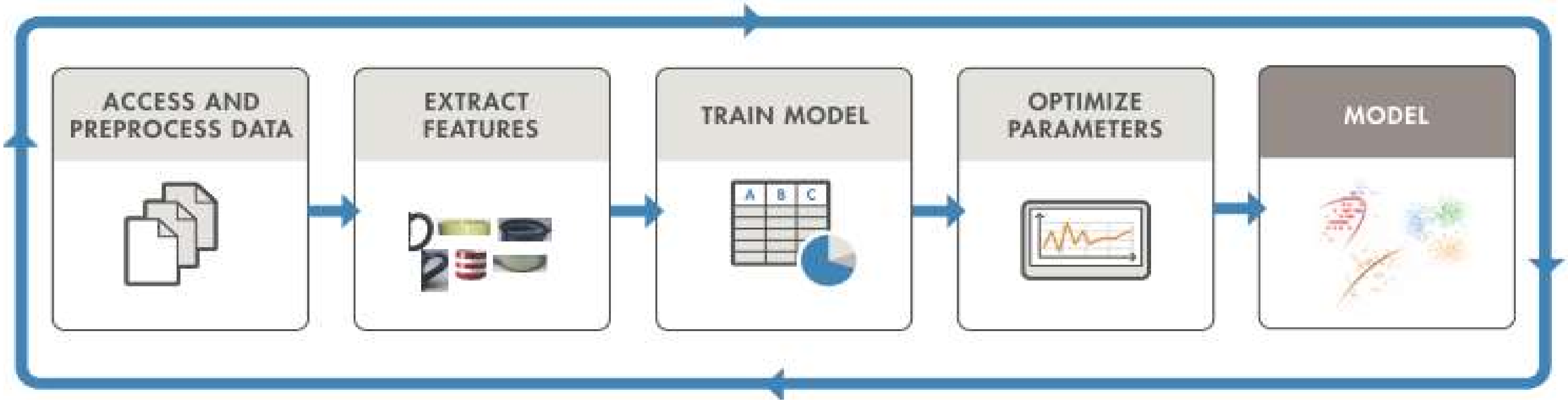


Multimodal generative AI models

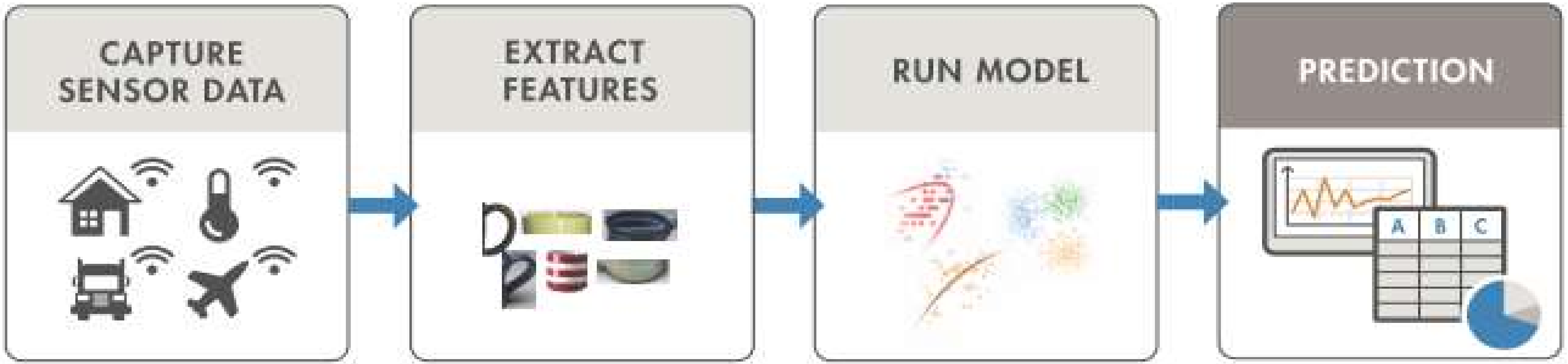


ML workflow

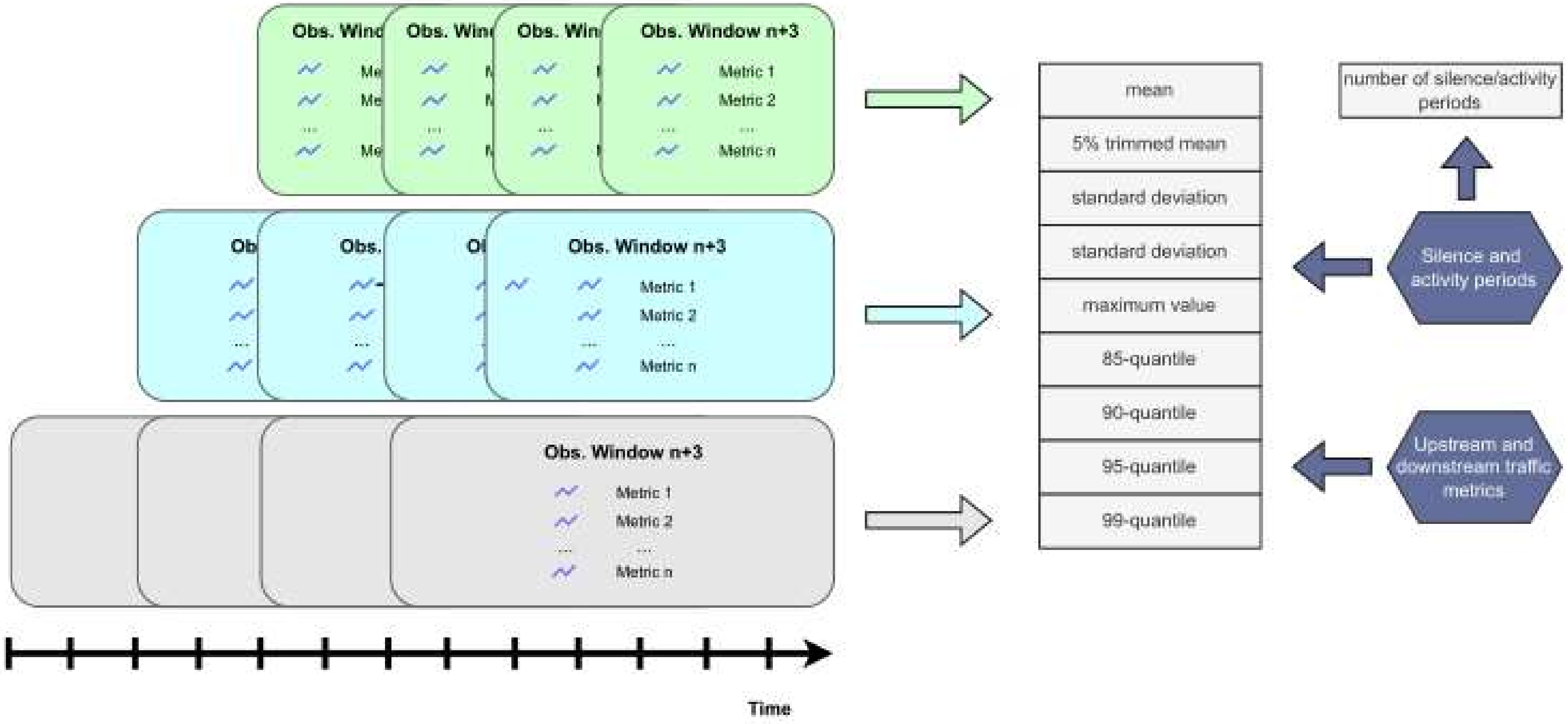
Train: Iterate until achieve satisfactory performance **(off-line)**



Predict: Integrate trained models into applications **(real time**)



From Raw data to Hand-crafted features



Raw data:
collected upstream/downstream network traffic metrics; sensor measurements
uploaded packets (#, Bytes), downloaded packets (#, Bytes), silence/activity periods

Feature extraction (input vector x) - e.g. statistical metrics
mean, max, min, standard deviation, different quantiles, over multiple sub-windows

Class (label y) : Network traffic OK (0) / NOT OK (1)

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

Applications: intelligent robotics, autonomous systems

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^2)	#bedrooms	Price (1000$s)
2104	3	400
1600	3	330
2400	3	369
1416	2	232
3000	4	540
⋮	⋮	⋮

Classification: The Labels are categorical values (class 1, class 2, etc.)

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

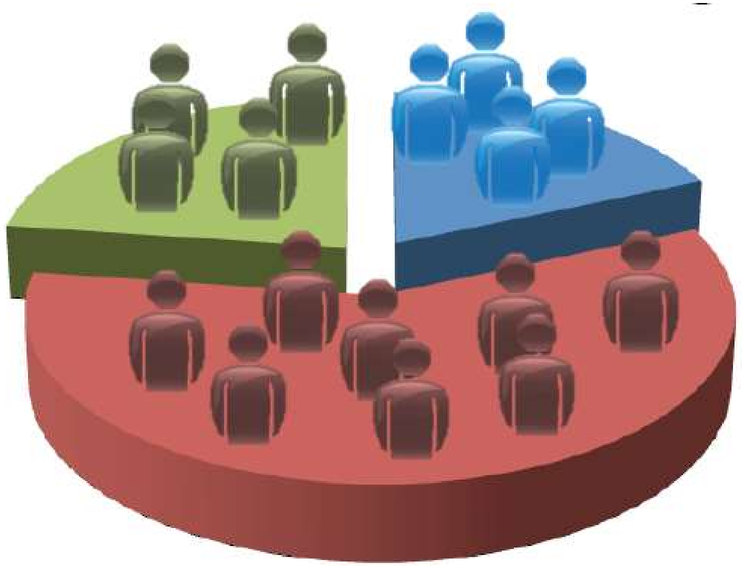
Features: memory use of computer ; number of disc accesses /sec; CPU load ; network traffic; silence

Unsupervised Learning

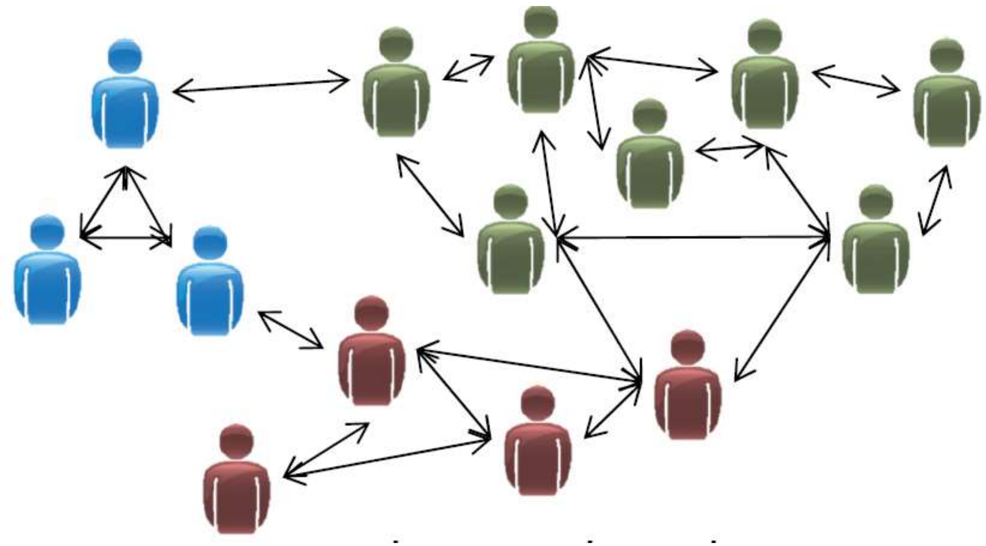
Given unlabeled data (NO answers)

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

Market segmentation



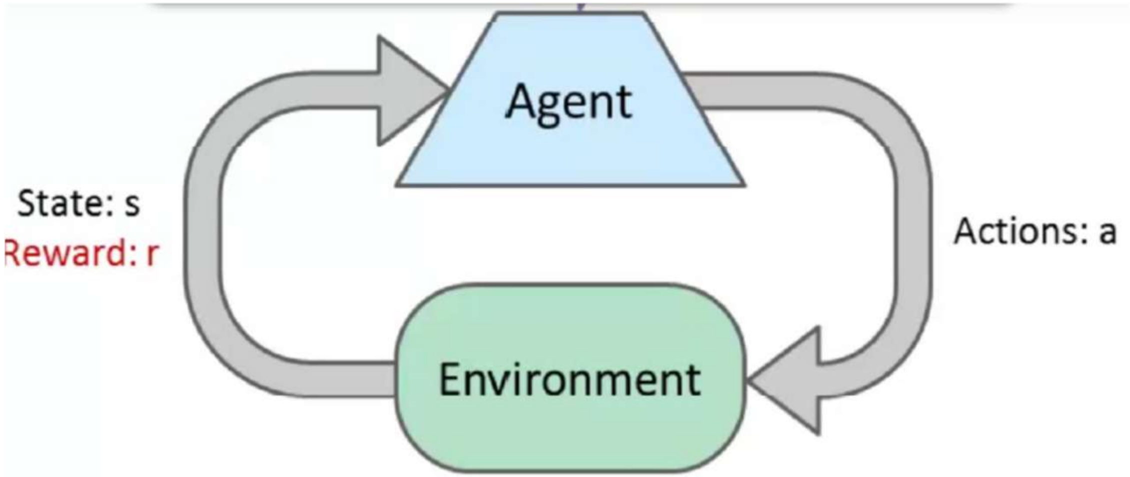
Social network analysis



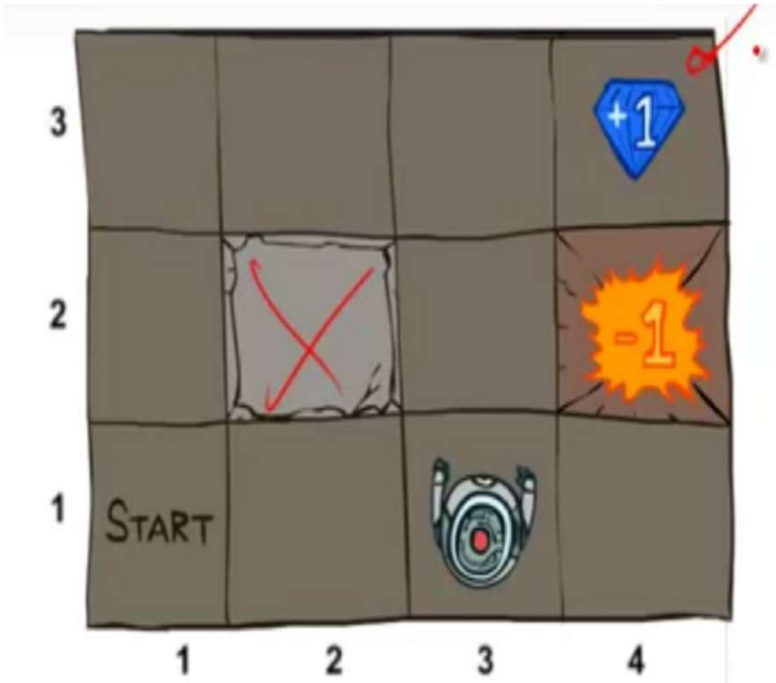
Clustering: Given a collection of examples (e.g. user profiles with a number of features). Each example is a point in the multidimensional space of features. Find a similarity measure that separates the points into clusters.

-***K-means clustering***

Reinforcement Learning



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



Learn to act so as to maximize
expected rewards
Learning is based on observed episodes

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.

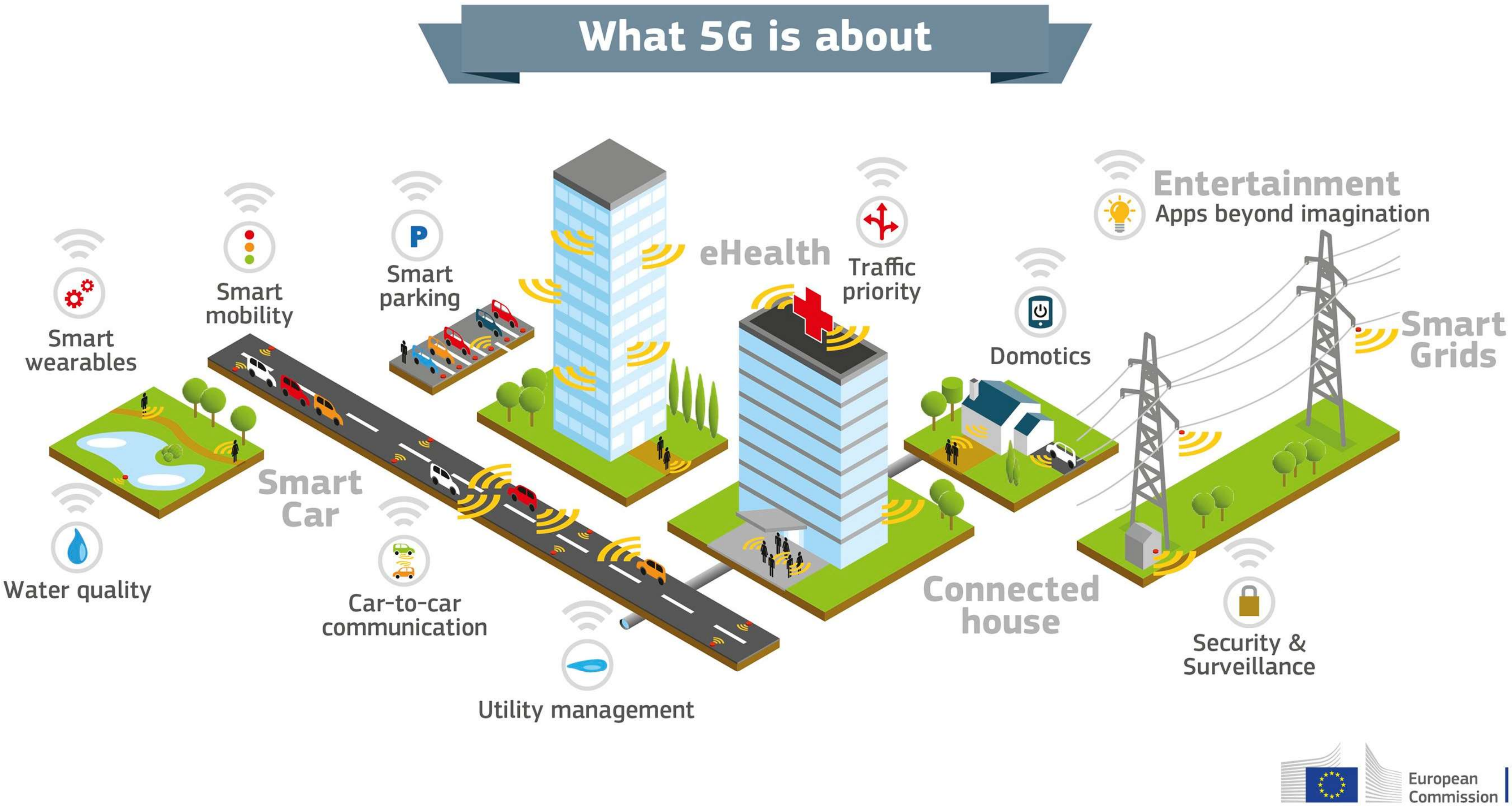
Big Data: Exponential growth of data, (IoT, medical records, biology, engineering, etc.)

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)

DL for 5G+ networks

Data traffic forecast – a key mechanism to automate 5G Network



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

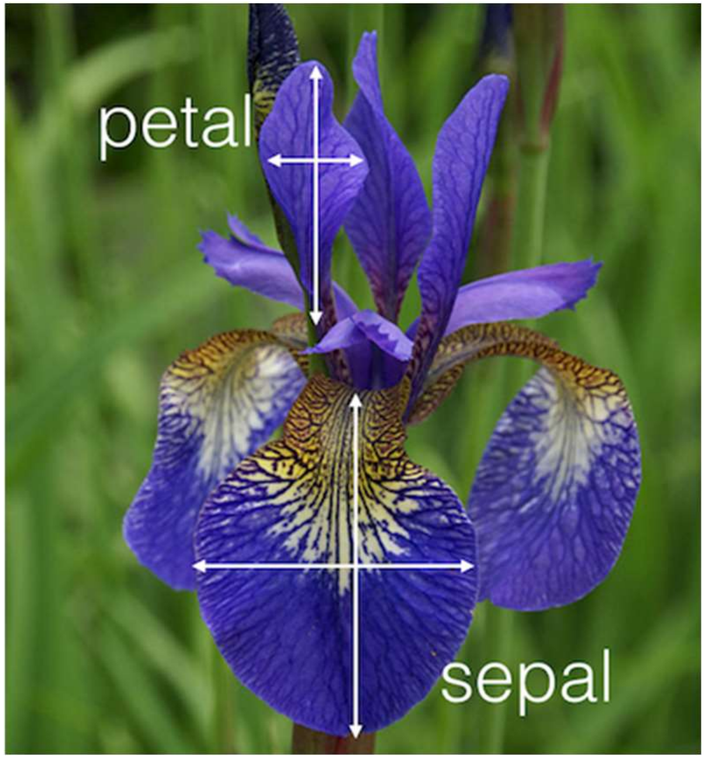
Red	Yellow	Green
1	0	0
1	0	0
0	1	0
0	0	1

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

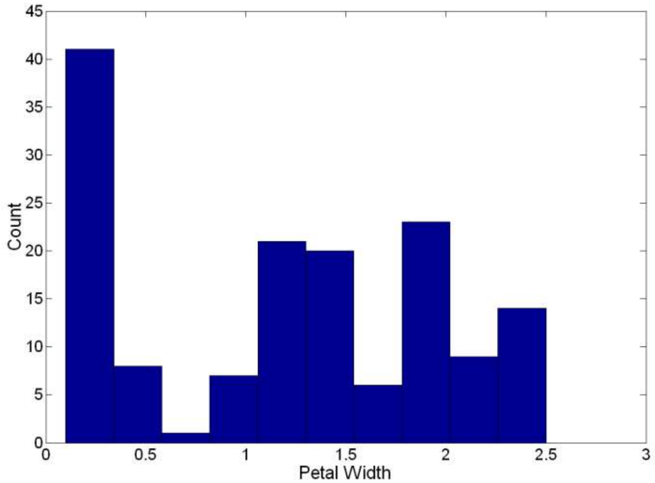


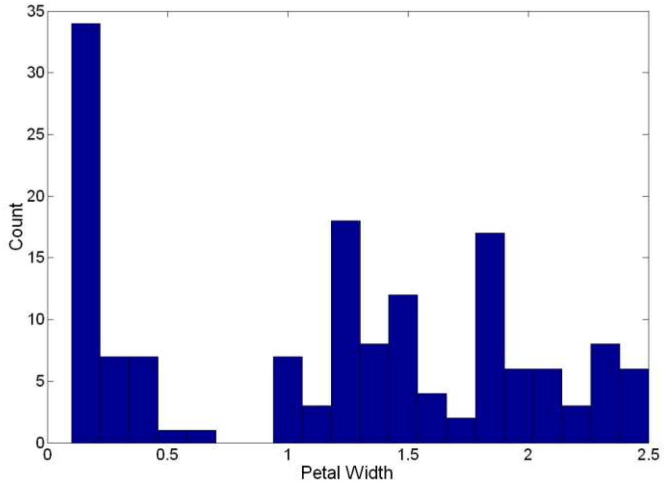
universidade de aveiro

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)

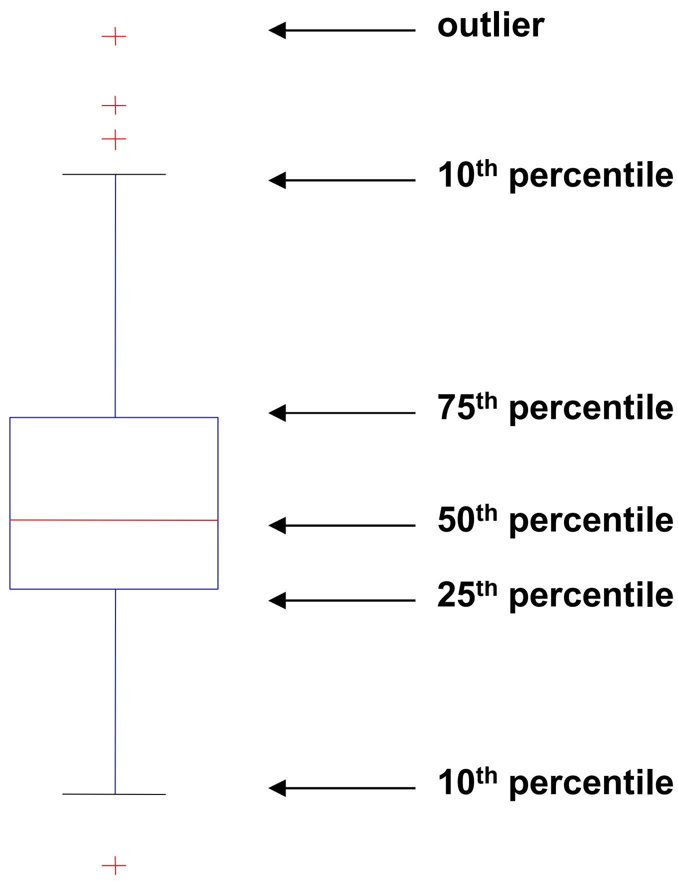




universidade de aveiro

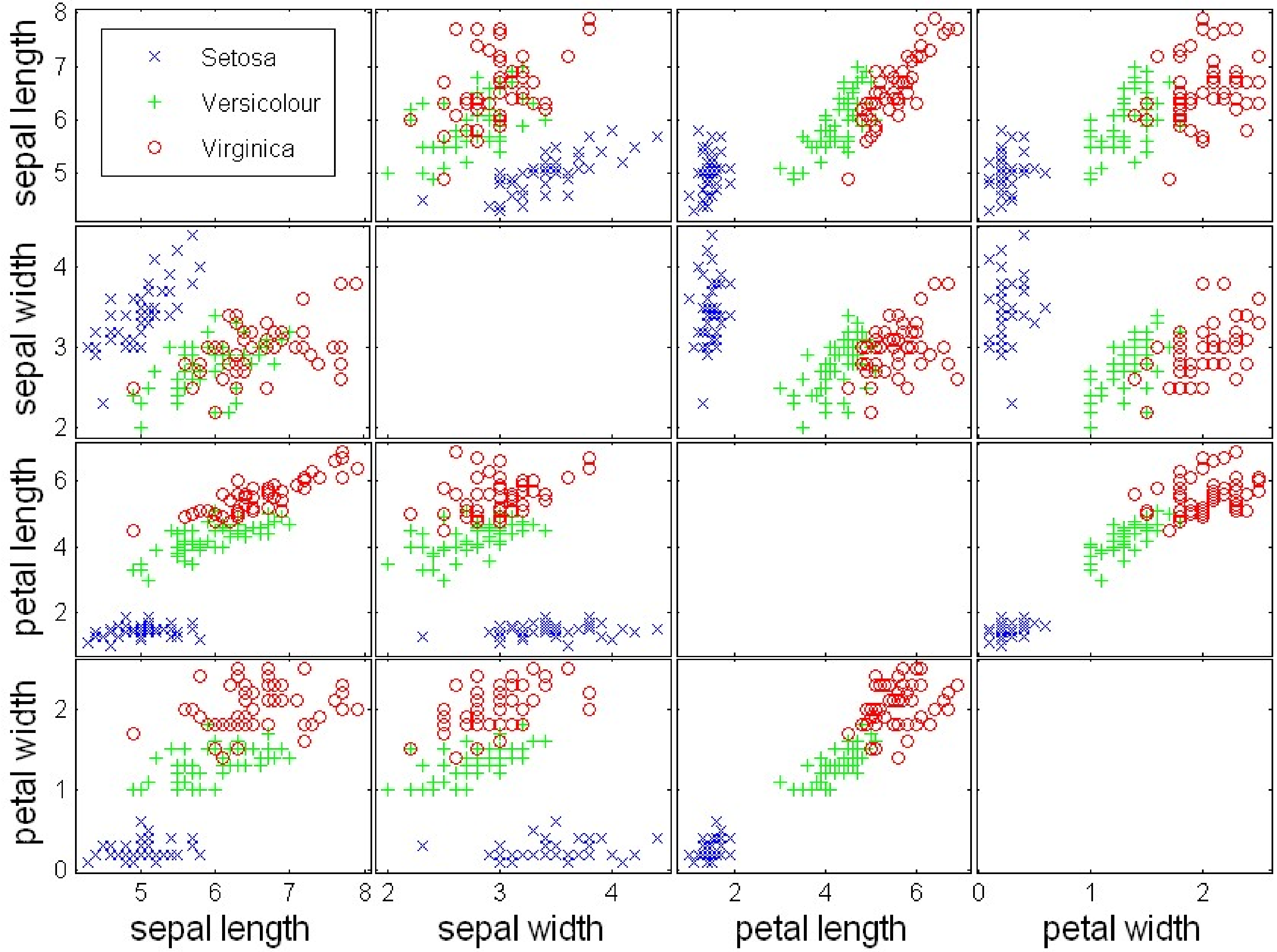
Data Visualization (2)

- **Box Plots**
  - Another way of displaying the distribution of data



Data Visualization (3)

Scatter Plot Array



universidade de aveiro

RECOMMENDED BIBLIOGRAPHY

- Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems. Aurélien Géron. O'Reilly, 2019

- François Chollet. Deep Learning with Python, Manning, 2018. (on-line)

- Andrew Ng, Machine Learning Yearning, 2017.

- Tom Mitchell, Machine Learning. McGraw-Hill, 1997.

- http://cs229.stanford.edu/

- MOOC (Massive Open Online Courses)

  e.g. https://www.coursera.org/

ANACONDA 3

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

https://docs.anaconda.com/anaconda/install/

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

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

Comment: If use higher versions than python 3.11 problems with tensorflow/ kerras libraries may arise.
Try to keep for now python version below 3.11.