

Departamento de Eletrónica, Telecomunicações e
Informática

Machine Learning

INTRODUCTION

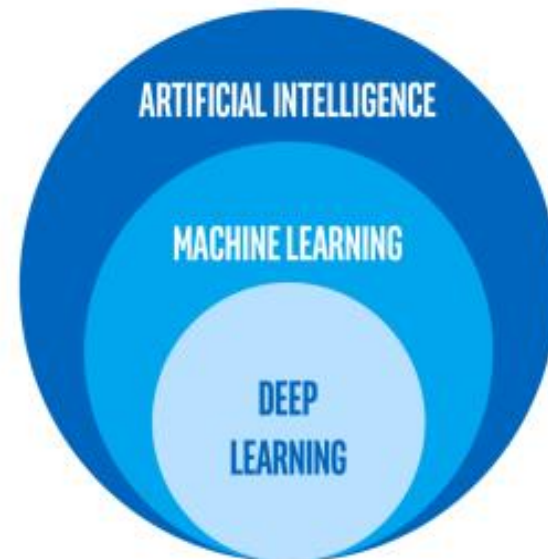
Author: Petia Georgieva

Edited by: Susana Brás (susana.bras@ua.pt)

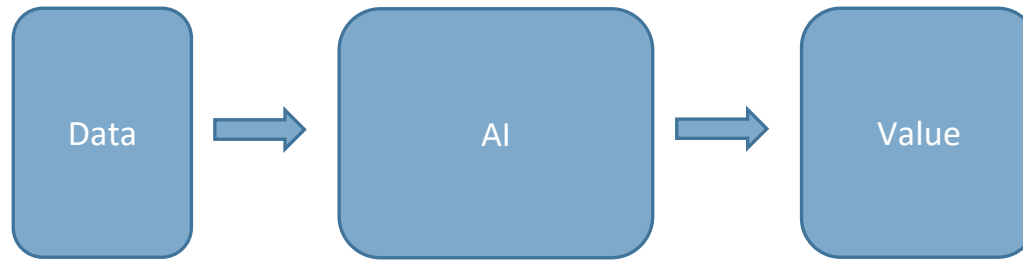
Artificial Intelligence (AI)

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

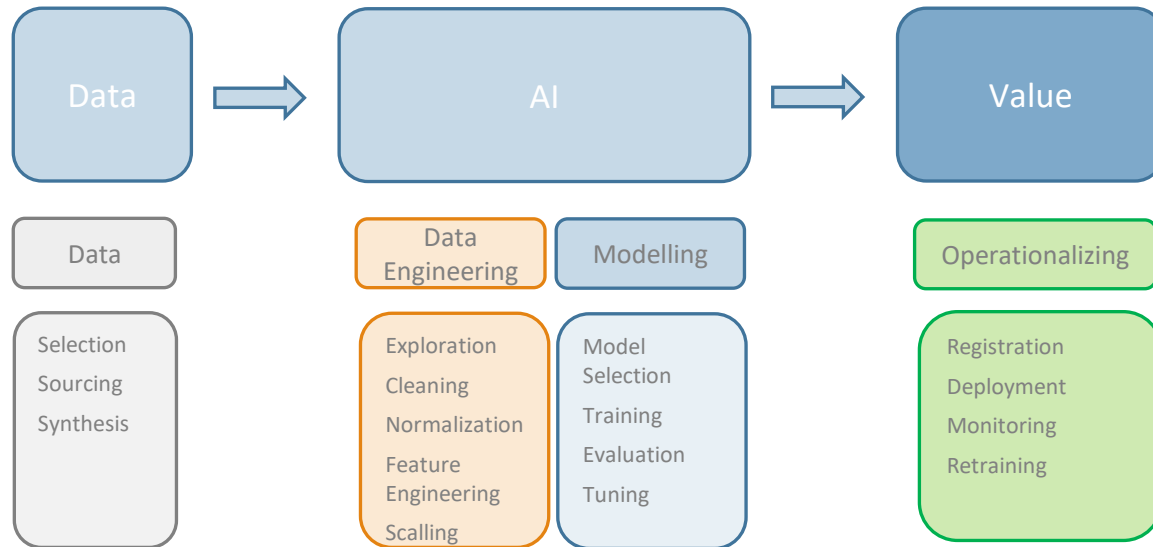
AI is based on Machine Learning (ML) & Deep Learning (DL) algorithms



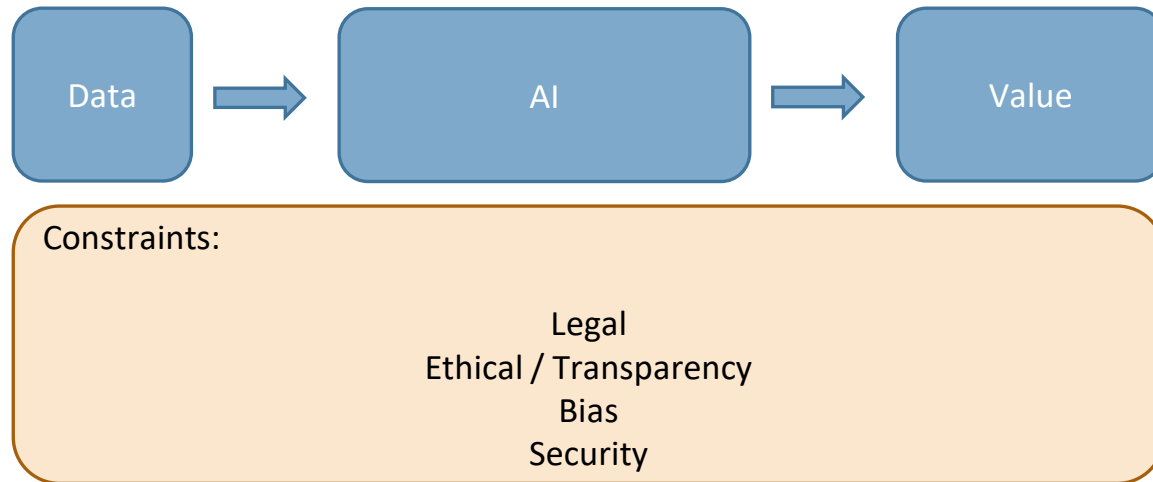
The process



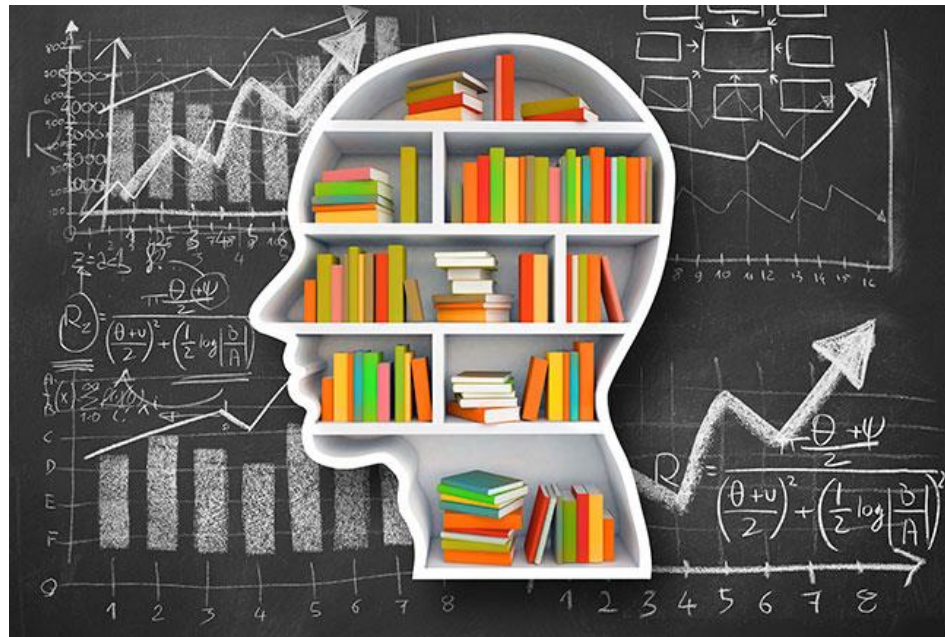
The process



The process



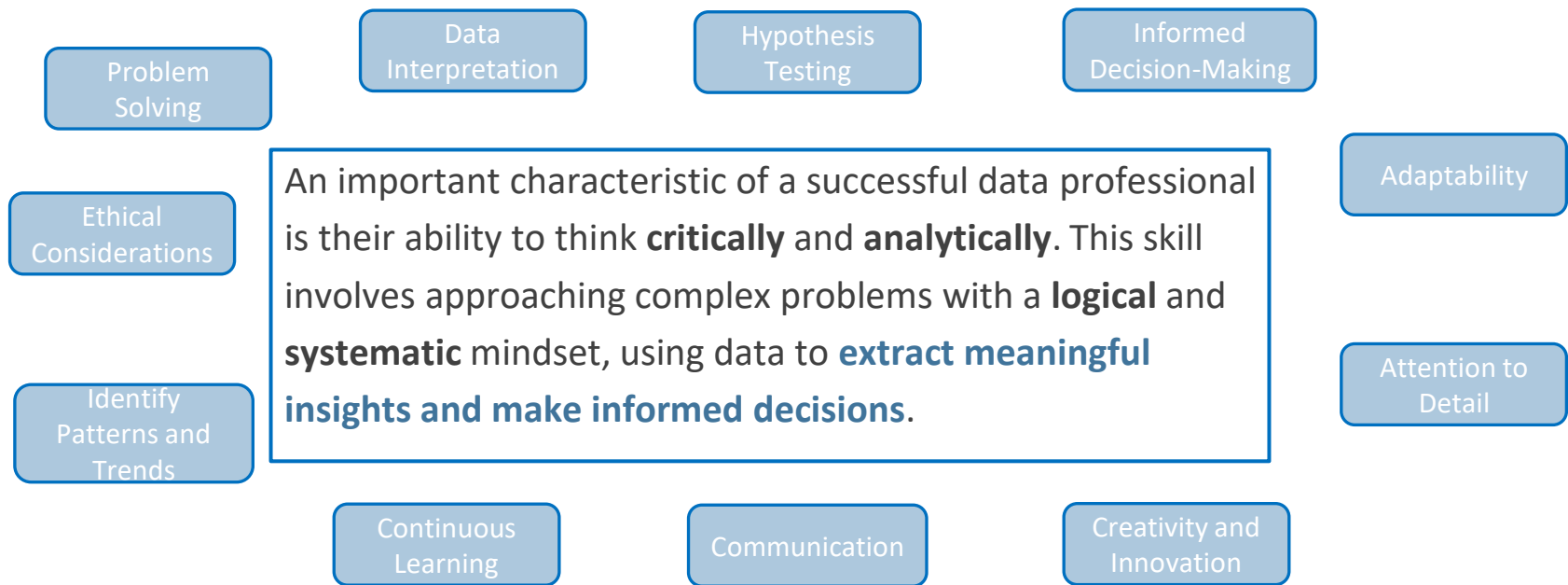
What is an important characteristic of a Data Professional?



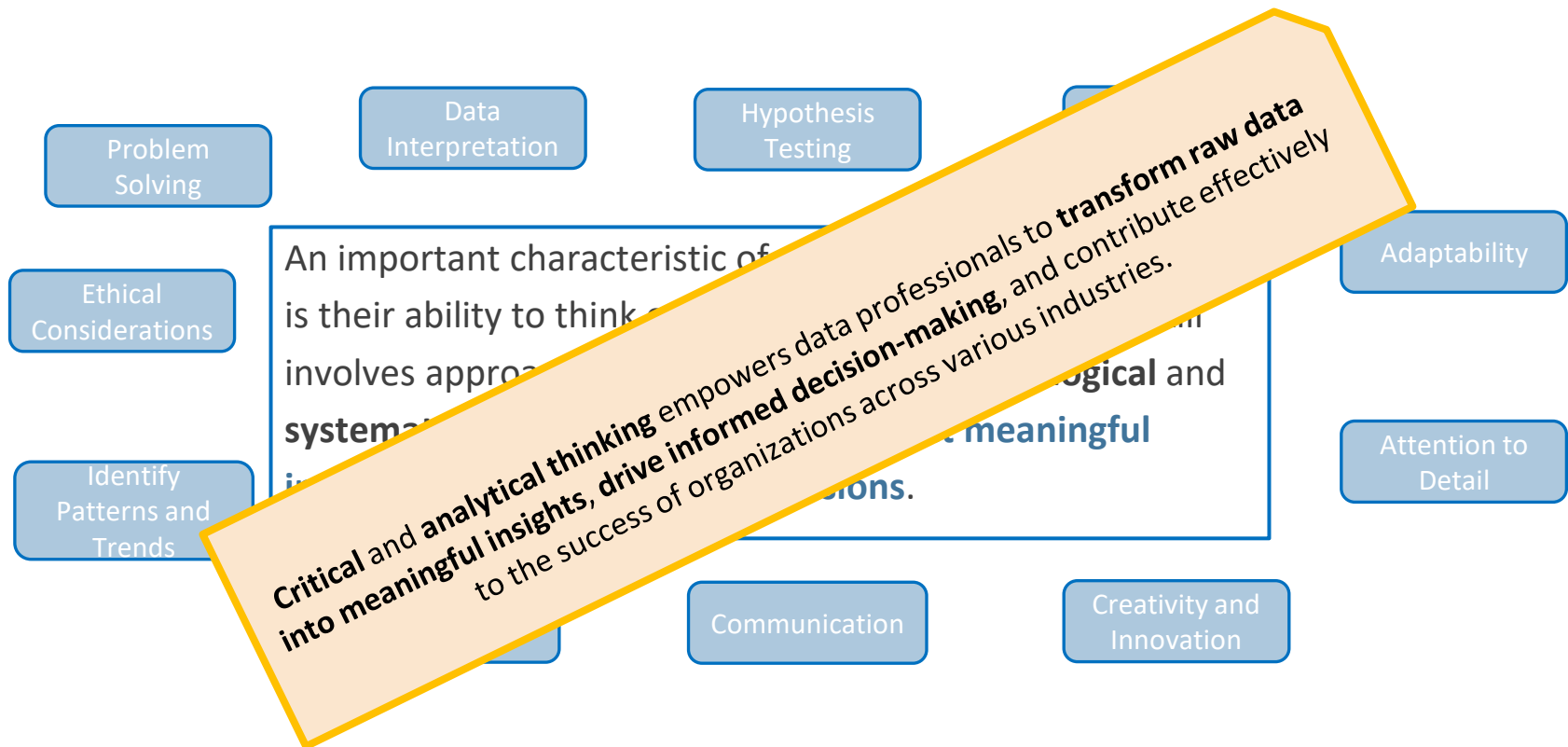
What is an important characteristic of a Data Professional?

An important characteristic of a successful data professional is their ability to think **critically** and **analytically**. This skill involves approaching complex problems with a **logical** and **systematic** mindset, using data to **extract meaningful insights and make informed decisions**.

What is an important characteristic of a Data Professional?



What is an important characteristic of a Data Professional?



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 (Long Short Term Memory) neural network
- Multivariate Gaussian approach for Anomaly Detection
- Recommender Systems

Avaliação

- Trabalho 1 (T1):
 - Trabalho individual
 - Trabalho escrito de cariz teórico/ técnico, onde os alunos de forma individual deverão explorar um tema a nível de dados, no final do semestre serão compilados os vários trabalhos formando um portfolio da turma sobre “Data Compliance, Quality and Exploration”.
 - Permitido o uso de ferramentas de AI, desde que mencionadas, privilegiando as que permitem fazer tracking das referências (<https://scite.ai/> , <https://consensus.app/> , <https://elicit.com/>) .
 - Avaliação: qualidade do documento, rigor técnico/ científico, capacidade critica do tema. Máximo 4 páginas, formato IEEE.
 - Não haverá lugar a apresentação.
 - Entrega na última semana de março.
- Trabalho 2 (T2):
 - Trabalho de grupo (3 elementos)
 - Trabalho prático, onde os alunos deverão resolver um problema de machine learning, que será proposto pelo grupo, recorrendo a bases de dados disponíveis nas plataformas web, fazendo uso das ferramentas lecionadas.
 - 3 entregas programadas
 - Entrega 1 – primeira semana de abril – Definição de grupo, tema, titulo, objetivos
 - Entrega 2 – primeira semana de maio - Resultados preliminares (algoritmo selecionado, metodologia de tratamento de dados)
 - Entrega 3 – última semana de maio – Documento final, código e apresentação
 - Apresentação – 2 últimas aulas do semestre
 - Avaliação:
 - Entrega 1 e 2: sem avaliação quantitativa, apenas qualitativa em caso de necessidade de ajuste
 - Documento: qualidade do documento, rigor técnico/ científico, dificuldade do problema, capacidade critica. Máximo 6 páginas, formato IEEE.
 - Apresentação: qualidade da apresentação, capacidade de discussão do tema
 - Avaliação entre pares dentro do grupo
 - Avaliação pelos colegas da turma
 - Documento (40%), apresentação (40%), Avaliação dentro do grupo (10%), Avaliação turma (10%)
- Exame (E): exame escrito.
- $NF = 0.3 \cdot T1 + 0.3 \cdot T2 + 0.4 \cdot E$

Evaluation

- Work 1 (T1):
 - Individual Assignment
 - A theoretical/technical written assignment where students, individually, must explore a data-related topic. At the end of the semester, the assignments will be compiled into a class portfolio on "Data Compliance, Quality, and Exploration."
 - The use of AI tools is allowed, they should be mentioned in the document, with preference given to those that enable reference tracking (e.g., <https://scite.ai/>, <https://consensus.app/>, <https://elicit.com/>).
 - Assessment: Based on document quality, technical/scientific rigor, and critical analysis of the topic. Maximum 4 pages, IEEE format.
 - There will be no presentation.
 - Submission: Last week of March.
- Work 2 (T2):
 - Group Assignment (3 members)
 - A practical assignment where students must solve a machine learning problem proposed by the group, using datasets available on web platforms and the tools taught in class.
 - Three scheduled submissions:
 - Submission 1 – First week of April: Group definition, topic, title, goals.
 - Submission 2 – First week of May: Preliminary results (selected algorithm, data processing methodology).
 - Submission 3 – Final week of May: Final document, code and presentation.
 - Presentation: During the last two classes of the semester.
 - Assessment:
 - Submissions 1 and 2: No quantitative assessment, only qualitative feedback if adjustments are needed.
 - Final Document: Assessed for quality, technical/scientific rigor, problem difficulty, and critical analysis. Maximum 6 pages, IEEE format.
 - Presentation: Evaluated for quality and discussion of the topic.
 - Peer Evaluation: Within the group.
 - Class Peer Evaluation: By classmates.
 - Document (40%), presentation (40%), Peer Evaluation (10%), Class Peer Evaluation (10%)
- Exam (E): written exam.

$$NF = 0.3 \cdot T1 + 0.3 \cdot T2 + 0.4 \cdot E$$

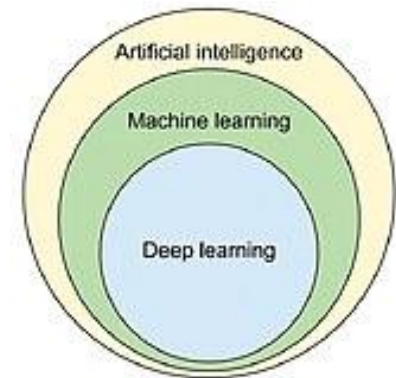


Why Machine/Deep Learning

- **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 Intelligence (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

The Paradigm Shifts in Artificial Intelligence

Communications of the ACM

21/10/2024

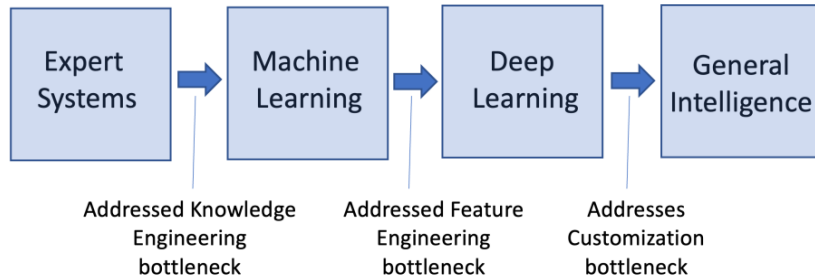


Figure. The history of artificial intelligence.

Table. The Paradigm Shifts in AI.

	DATA	EXEMPLAR	SCOPE	CURATION
Expert Systems	Human	Rules	Follows	High
Machine Learning	+ Databases	Rules/networks	+ Discovers relationships	Medium
Deep Learning	+ Sensory	Deep neural networks	+ Senses relationships	Low
General Intelligence	+ Everything	Pre-trained deep neural networks	+ Understands the world	Minimal

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


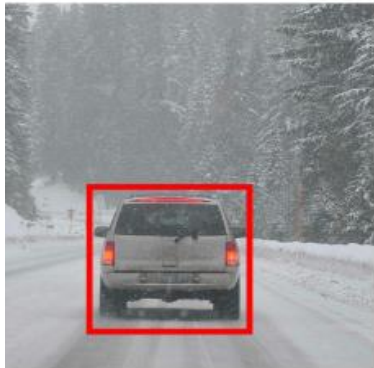

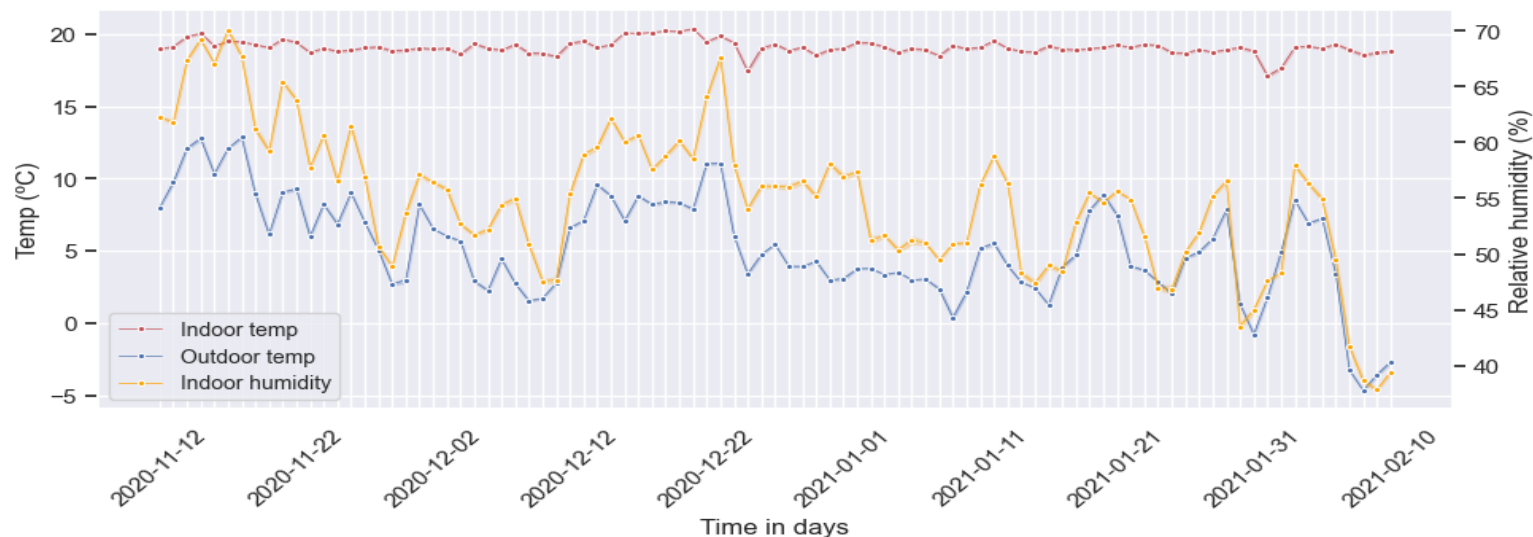
Image classification	Classification & Localization	Detection
	 b_x, b_y, b_h, b_w	

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

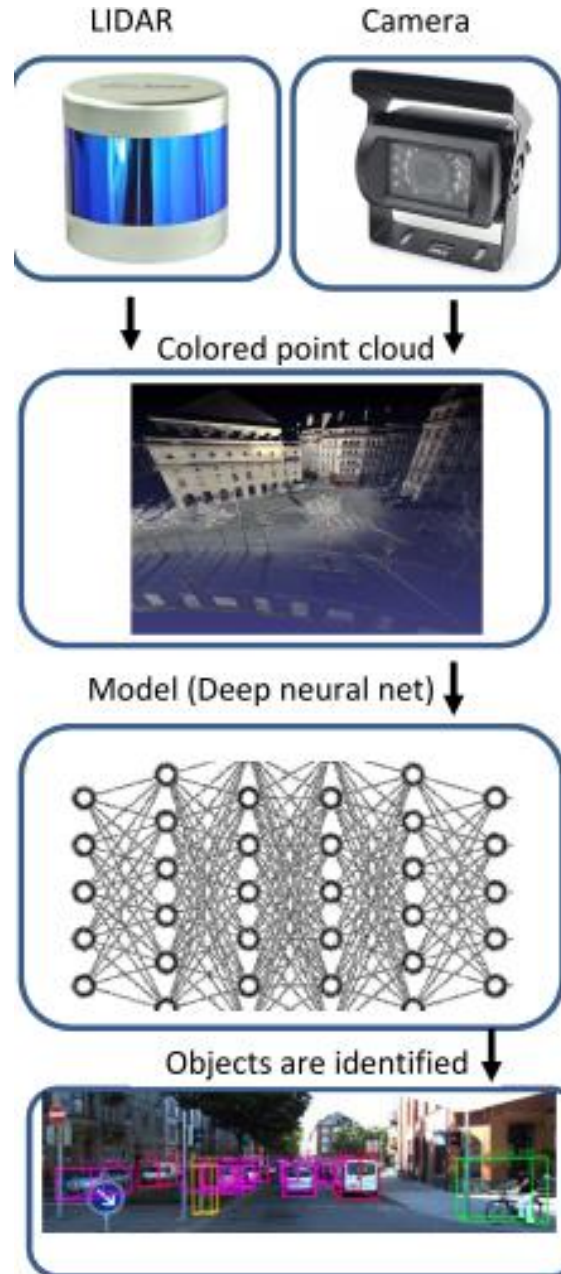


Time Series (TS) - collection of samples recorded at a sequence of time intervals

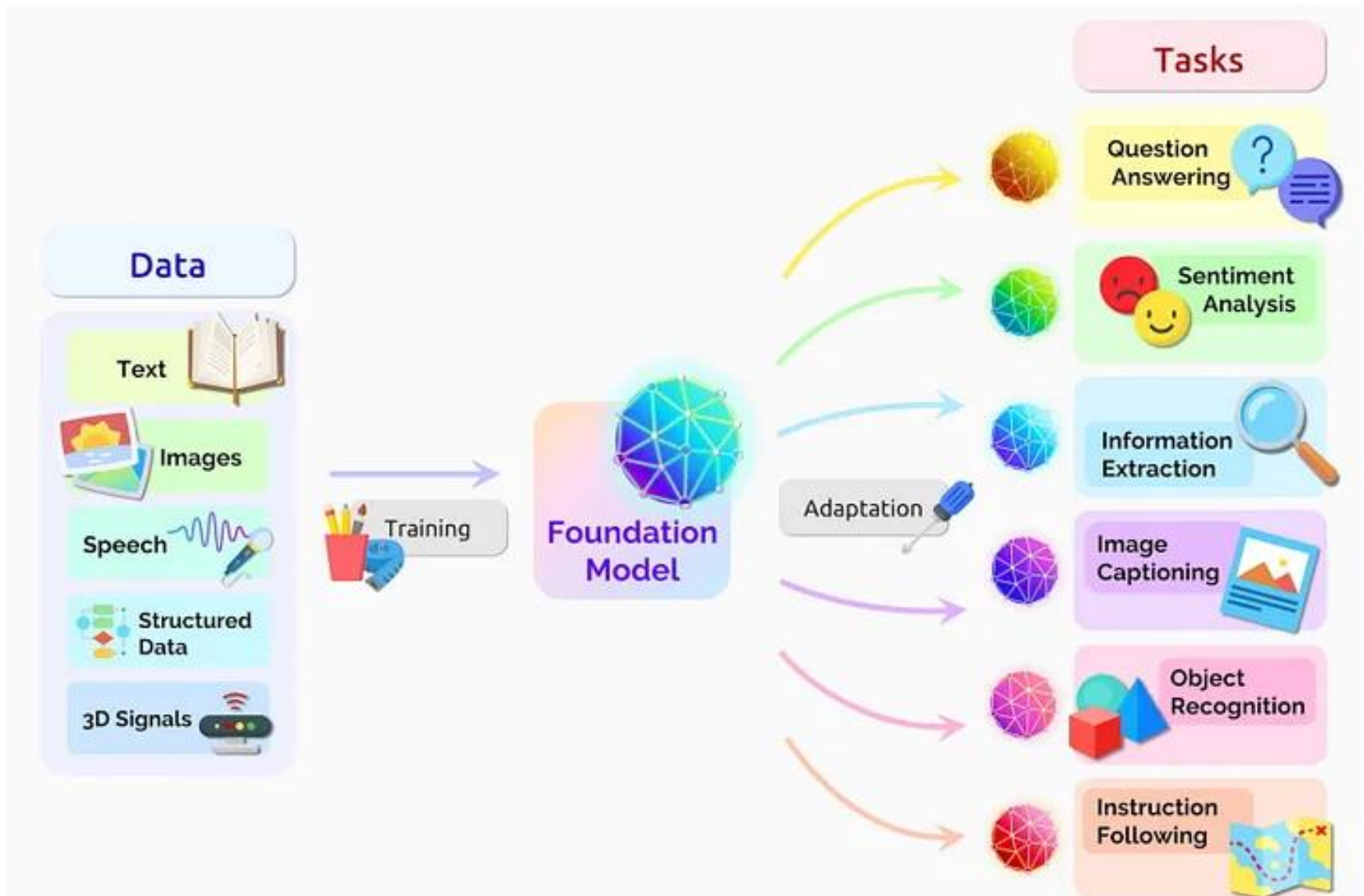
TS forecasting (prediction) => based on past samples, predict future trends, seasonality, anomalies, etc. Many applications:

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

Multimodal Object Detection

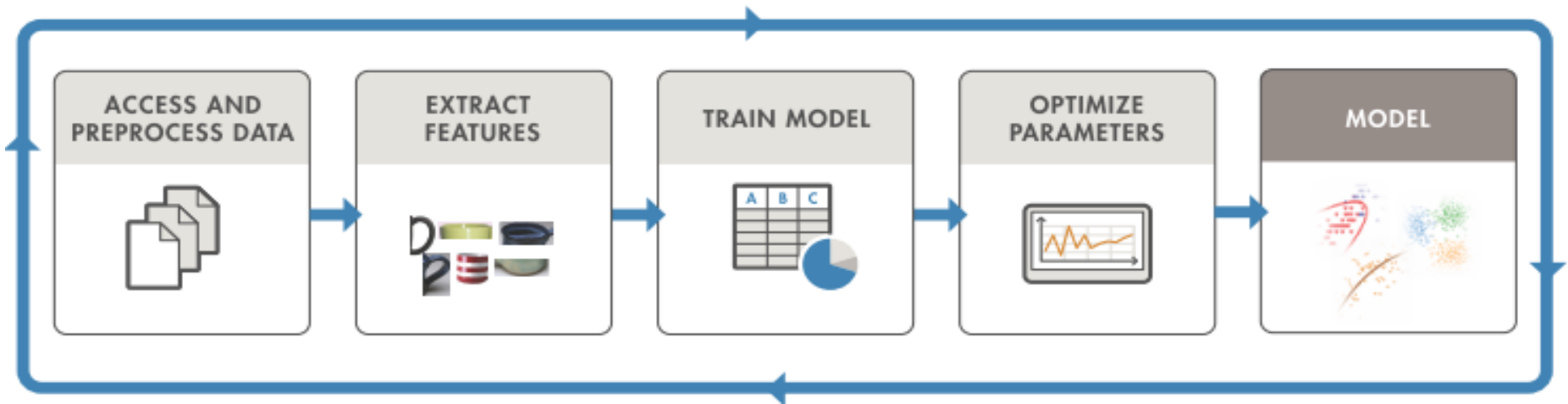


Multimodal generative AI models

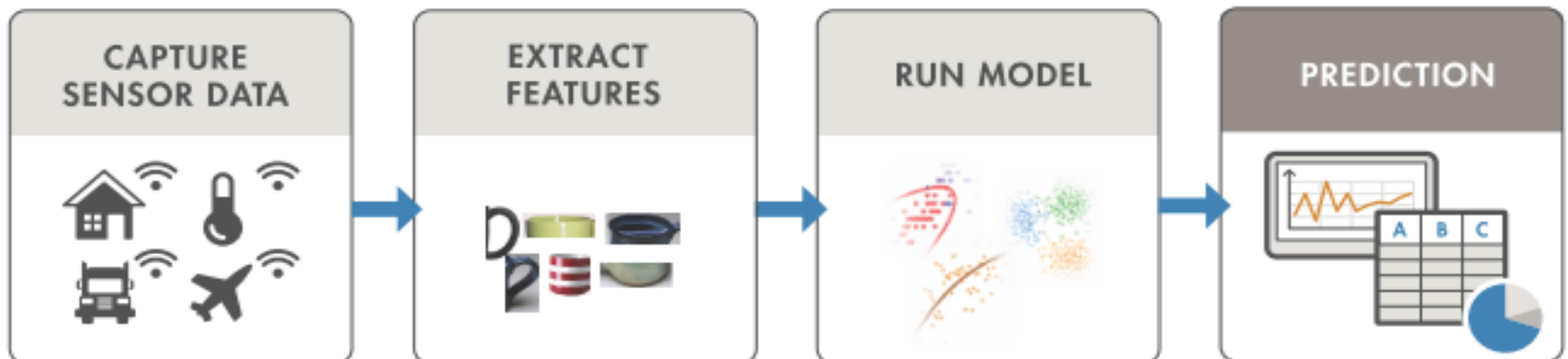


ML workflow

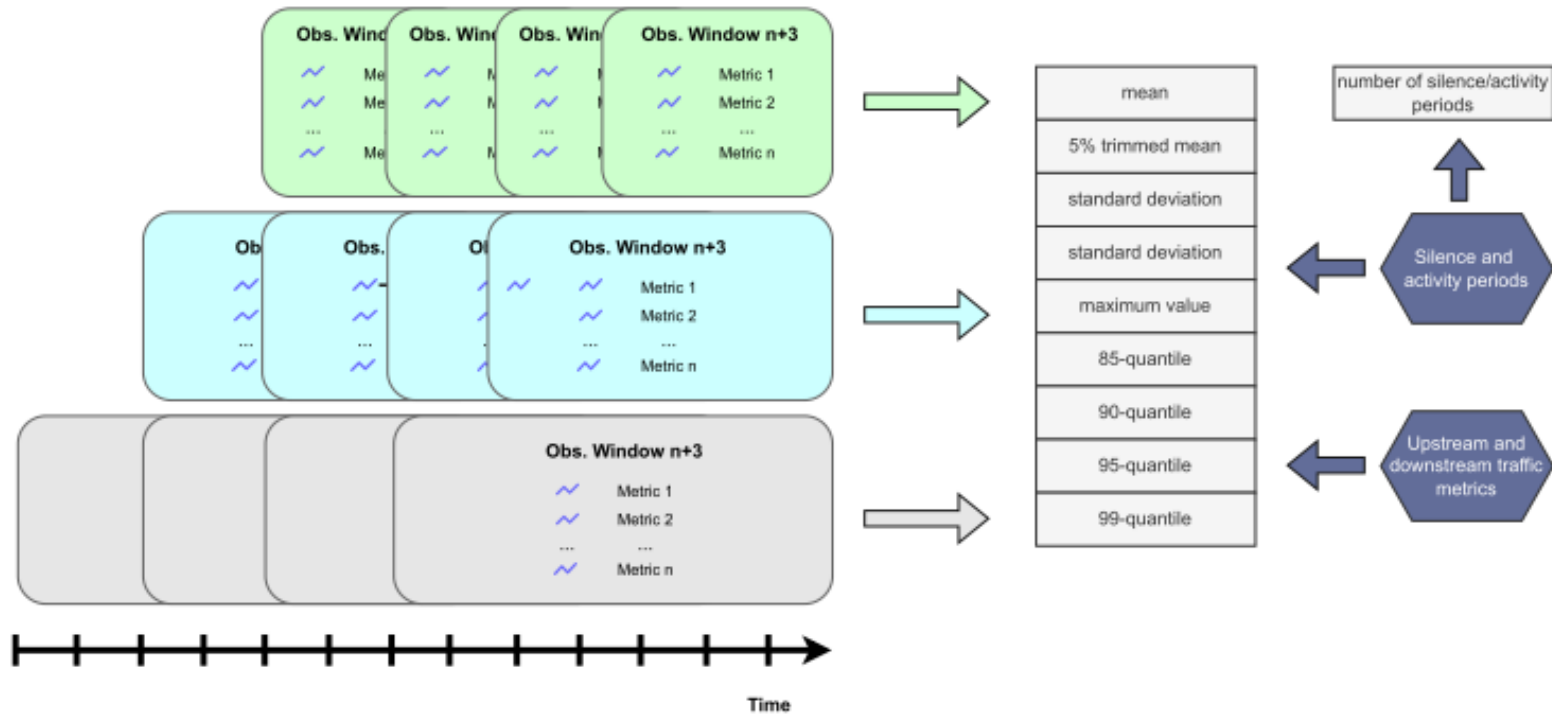
Train: Iterate until achieve satisfactory performance (**off-line**)



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 \mathbf{x}) - e.g. statistical metrics

mean, max, min, standard deviation, different quantiles, over multiple sub-windows

Class (label \mathbf{y}) : Network traffic OK (0) / NOT OK (1)

Terminologia

- The terms *sample*, *data point*, *observation*, or *instance* refer to a single, independent unit of data, such as a customer, patient, or compound. The term *sample* can also refer to a subset of data points, such as the training set sample. The text will clarify the appropriate context when this term is used.
- The *training set* consists of the data used to develop models while the *test* or *validation* sets are used solely for evaluating the performance of a final set of candidate models.
- The *predictors*, *independent variables*, *attributes*, or *descriptors* are the data used as input for the prediction equation.
- *Outcome*, *dependent variable*, *target*, *class*, or *response* refer to the outcome event or quantity that is being predicted.
- *Continuous* data have natural, numeric scales. Blood pressure, the cost of an item, or the number of bathrooms are all continuous. In the last case, the counts cannot be a fractional number, but is still treated as continuous data.
- *Categorical* data, otherwise known as *nominal*, *attribute*, or *discrete* data, take on specific values that have no scale. Credit status (“good” or “bad”) or color (“red,” “blue,” etc.) are examples of these data.
- *Model building*, *model training*, and *parameter estimation* all refer to the process of using data to determine values of model equations.



Pre- Processing and Data preparation

- Data pre-processing techniques generally refer to the addition, deletion, or transformation of training dataset.
- Different models have different sensitivities to the type of predictors in the model; how the predictors enter the model is also important.
- Transformations of the data to reduce the impact of data skewness or outliers can lead to significant improvements in performance.
- Simpler strategies such as removing predictors based on their lack of information content can also be effective.
- The need for data pre-processing is determined by the type of model being used. Some procedures, such as tree-based models, are notably insensitive to the characteristics of the predictor data. Others, like linear regression, are not.



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:

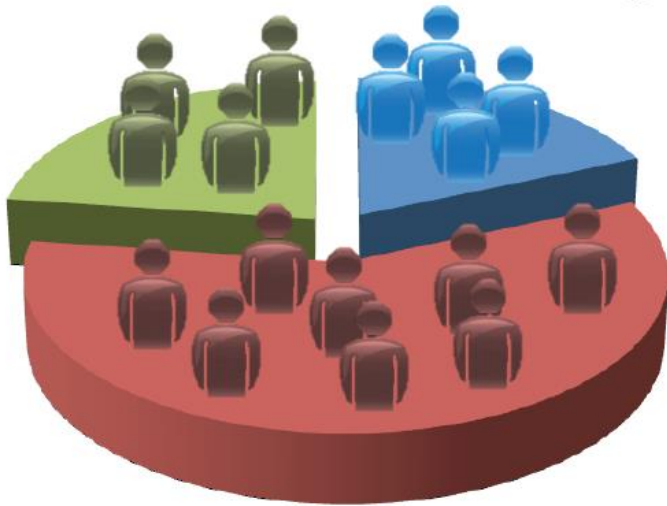
Features: memory use of computer ; number of disc accesses /sec; CPU load ; network traffic; silence

Unsupervised Learning

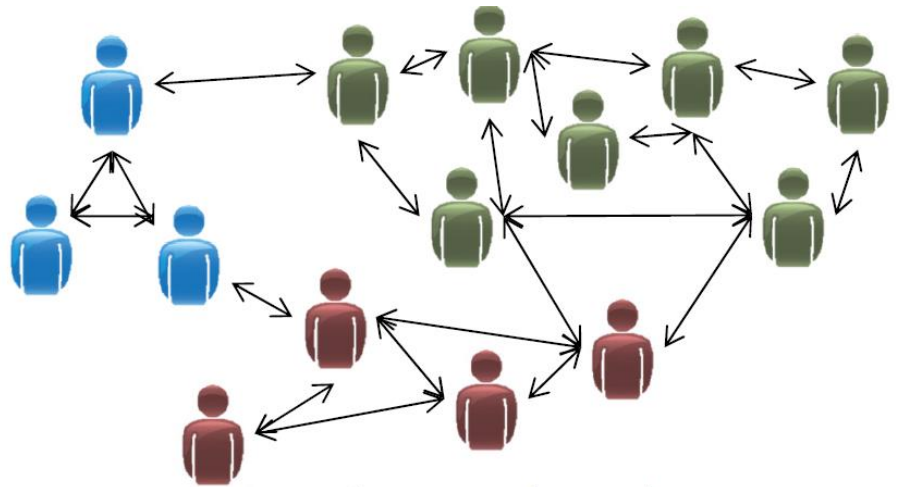
Given unlabeled data (NO answers)

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

Market segmentation



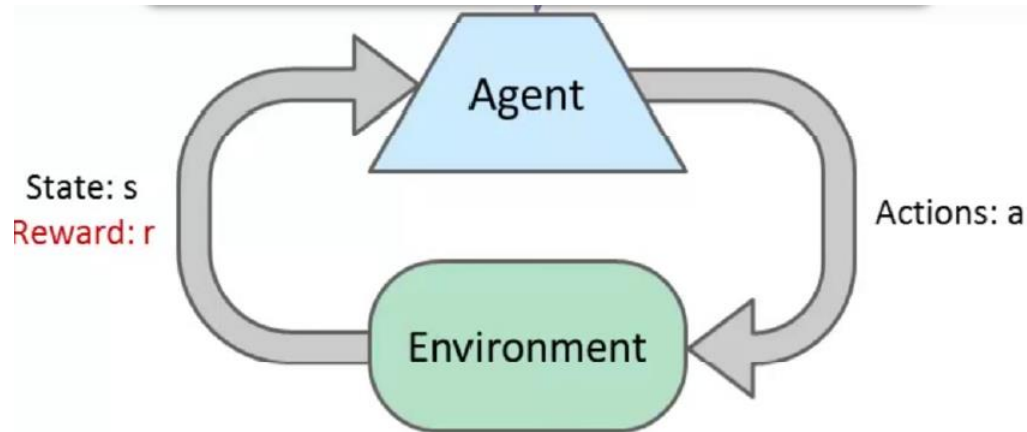
Social network analysis



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

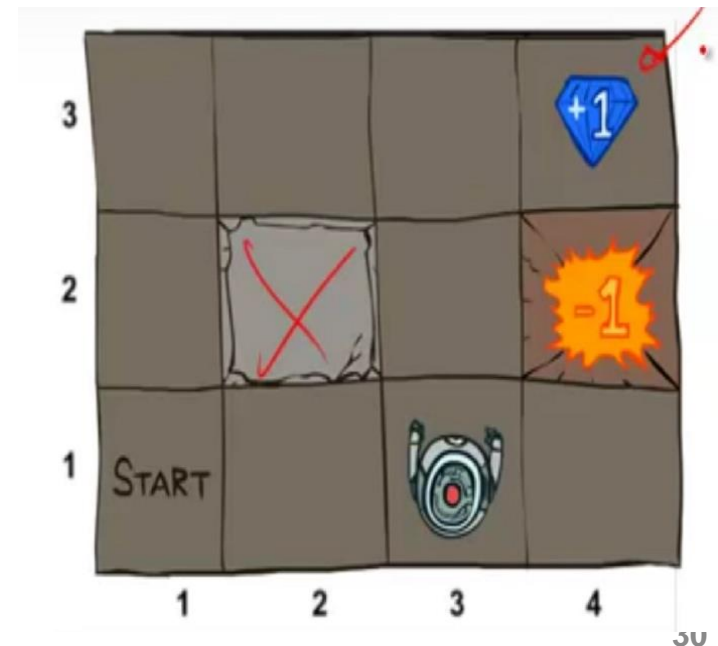
Reinforcement Learning



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

What 5G is about



Data Types

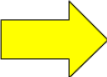
1. Numeric (Quantitative) features

- 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

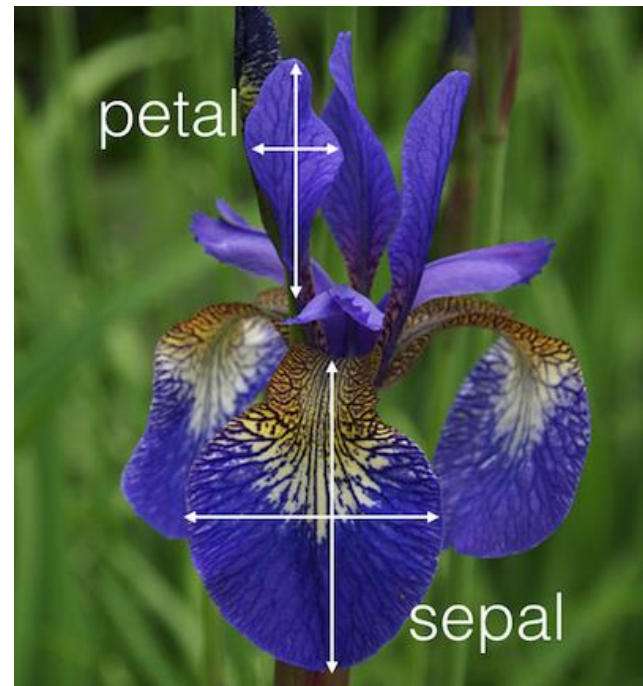


Color
Red
Red
Yellow
Green
Yellow

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

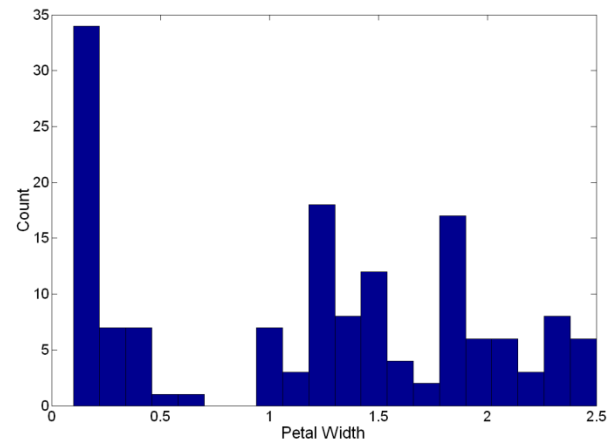
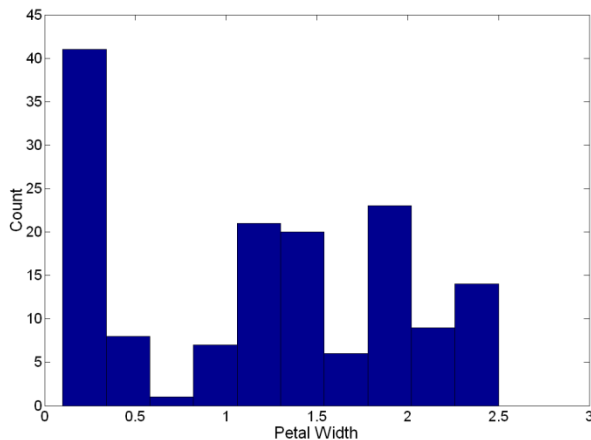


Data Visualization (1)

- **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

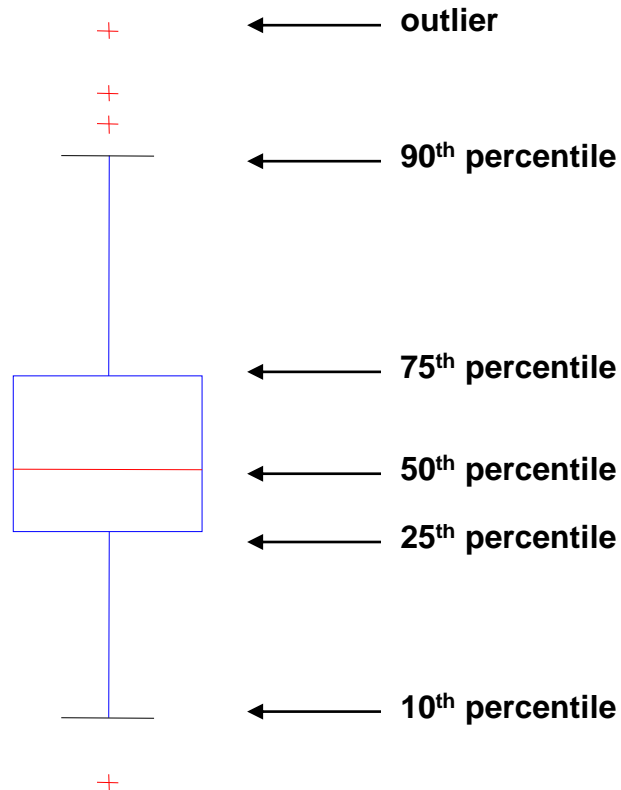
- Example: Petal Width (10 and 20 bins, respectively)



Data Visualization (2)

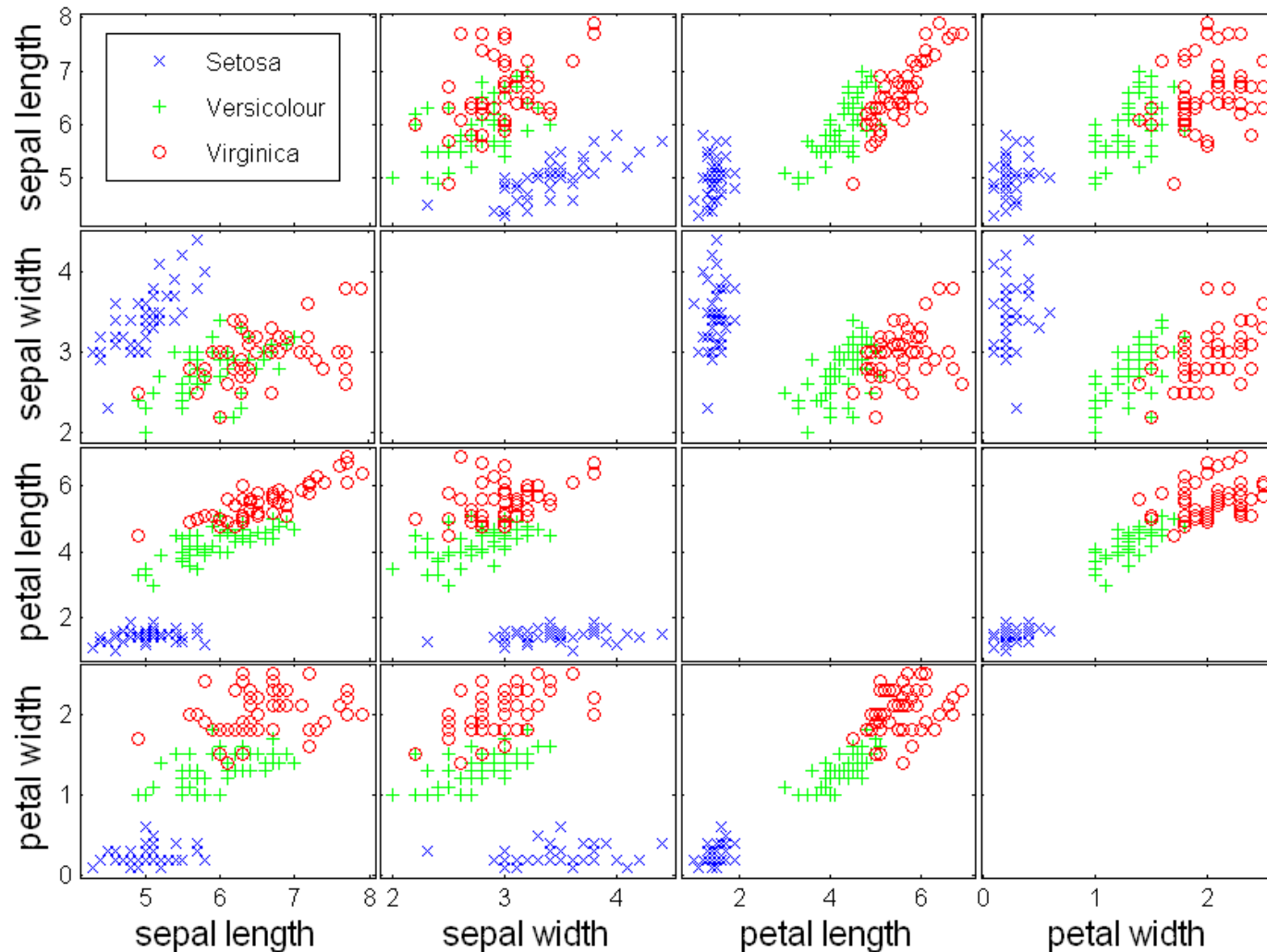
- **Box Plots**

- Another way of displaying the distribution of data



Data Visualization (3)

Scatter Plot Array



RECOMMENDED BIBLIOGRAPHY

- 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.

