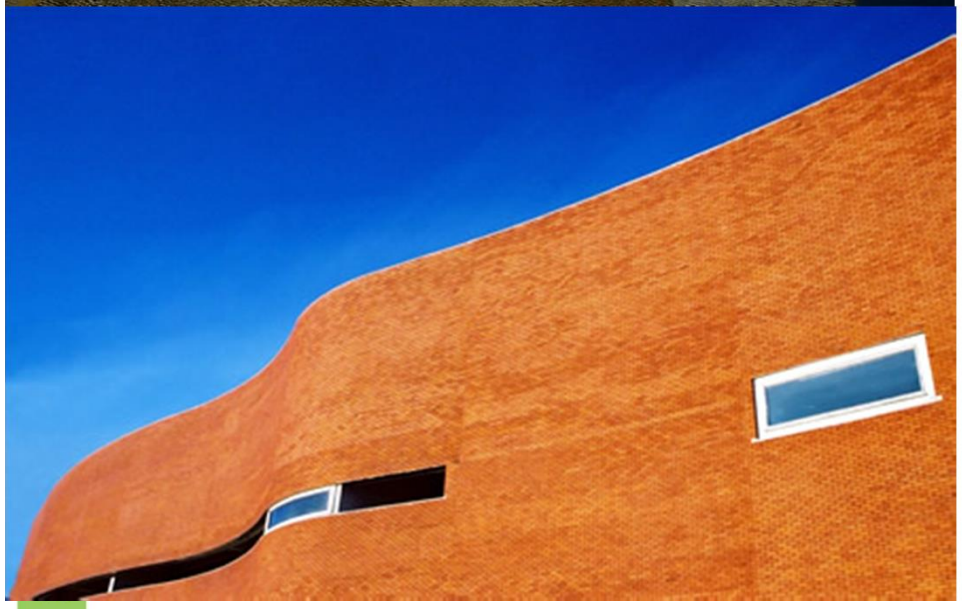
Aprendizagem Aplicada à Segurança

(Mestrado em Cibersegurança-DETI-UA)



Petia Georgieva
(petia@ua.pt)

DETI/IEETA
University of Aveiro

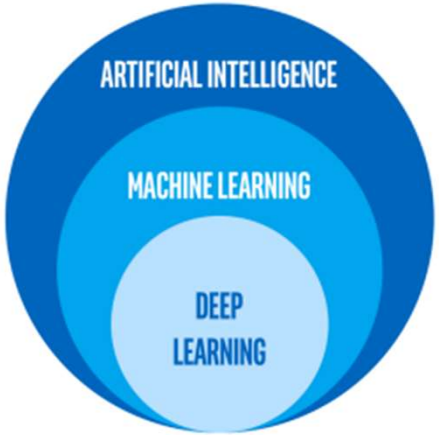
AI - the new Electricity

Artificial Intelligence (AI) will influence every industry .

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

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

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

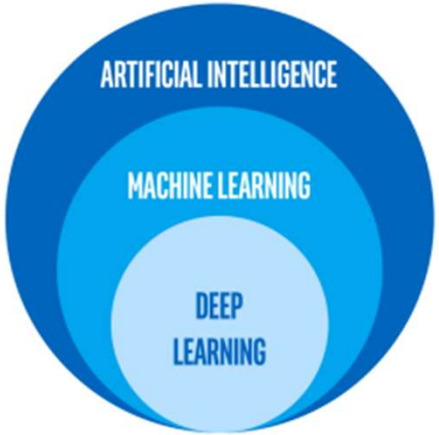


universidade de aveiro

AAS - Machine Learning (ML) module

Starts on 26/Nov/2021 (7th week)

- Anomaly Detection
- Supervised learning –regression
- Supervised learning – classification
- Unsupervised learning – clustering
- Deep Learning - introduction



universidade de aveiro

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 programed by hand.

  - ✓ Autonomous driving;
  - ✓ Computer Vision;
  - ✓ Natural Language Processing (Speech recognition, Machine translation)
  - ✓ User behaviour monitoring (Sentiment classification, Video activity recognition) .

ML advance is due to the rise of

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

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 (1)

Learning to detect & recognize faces

Computer Vision (2)

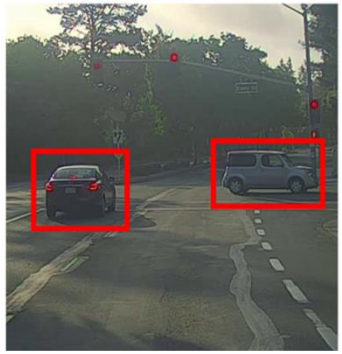
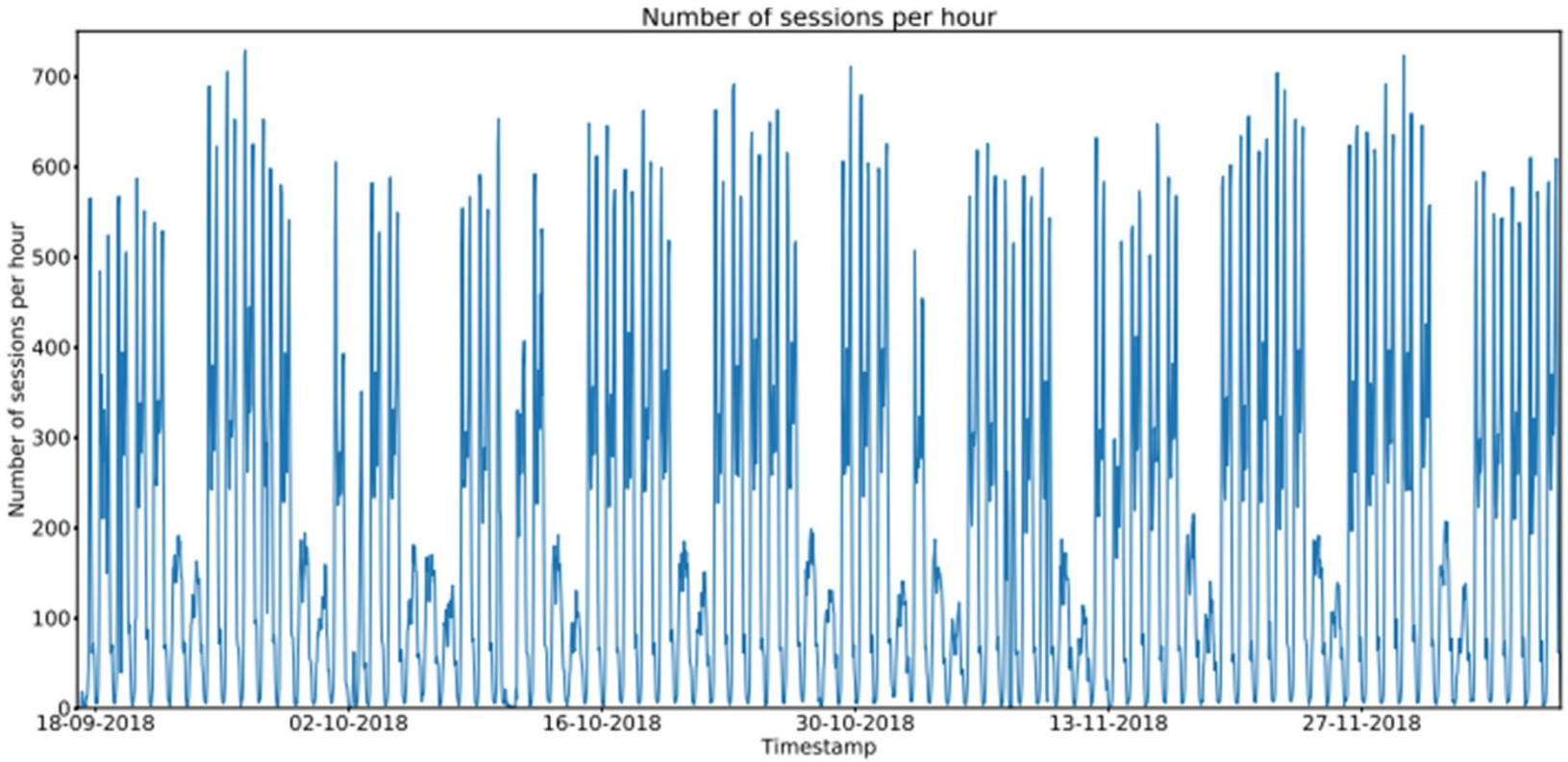
Image classification	Classification & Localization	Detection
	 b_x, b_y, b_h, b_w	

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

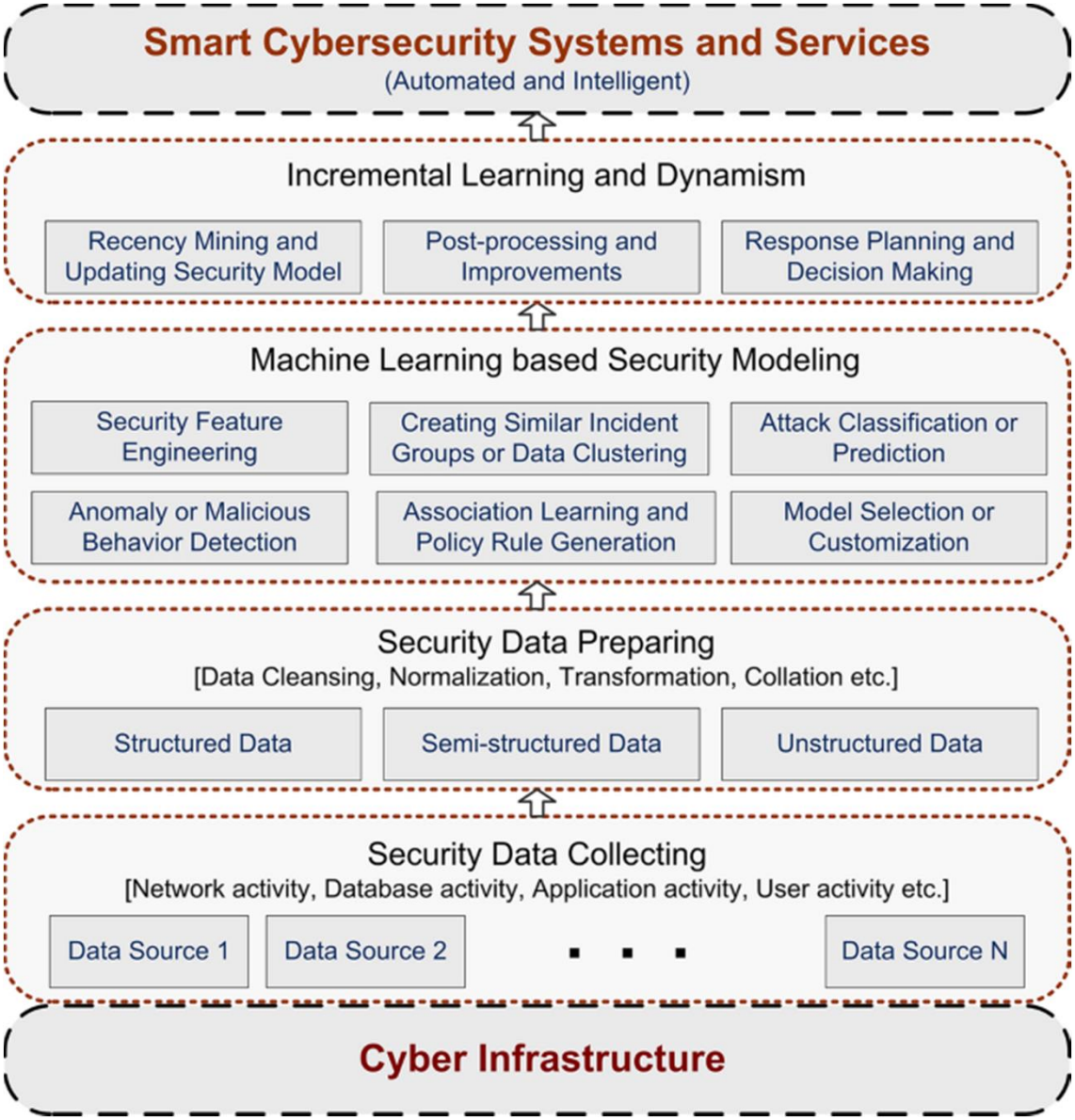
Classification & localization: the model outputs not only the class label of the object but also draws a bounding box (the coordinates) of its position in the image.

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

Time Series (TS) Data



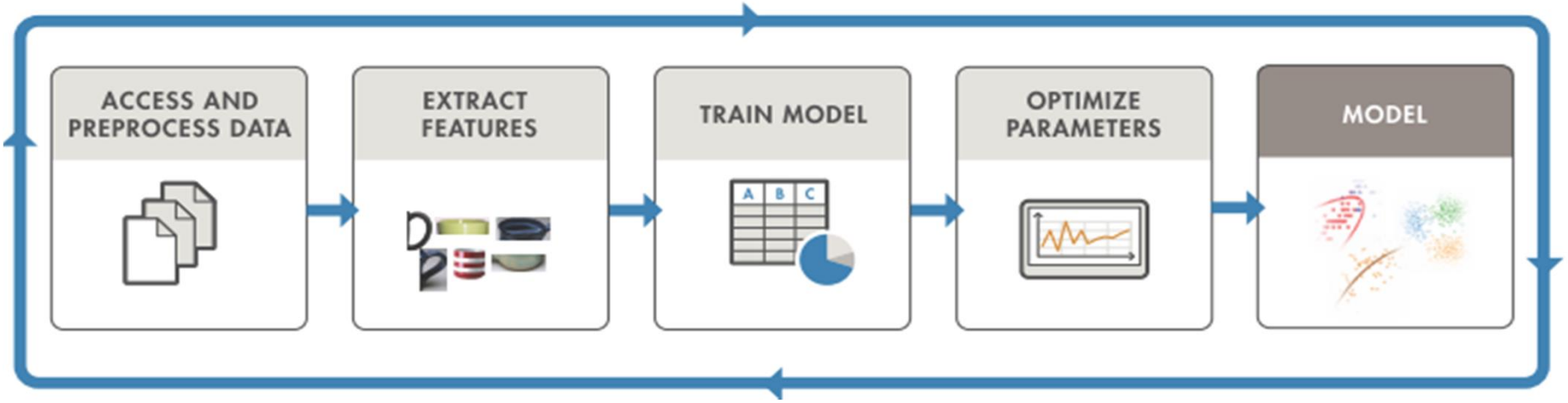
Time Series - collection of data points indexed based on the time they were collected . Most often, data are recorded at regular time intervals.



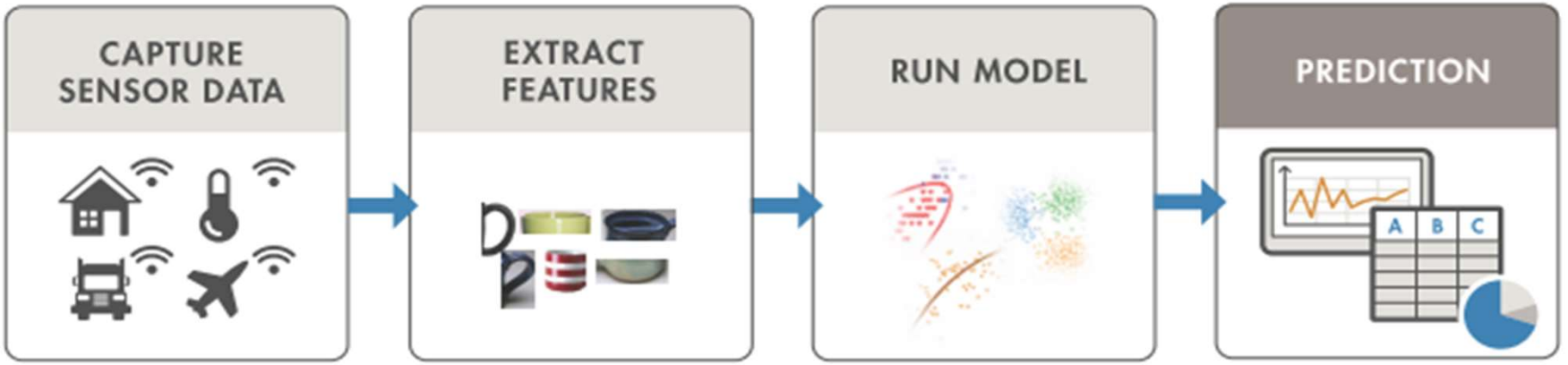
Sarker, I.H., Kayes, A.S.M., Badsha, S. *et al.* Cybersecurity data science: an overview from machine learning perspective. *Journal of Big Data* 7, 41 (2020). https://doi.org/10.1186/s40537-020-00318-5

ML workflow

Train: Iterate until achieve satisfactory performance.



Predict: Integrate trained models into applications.



All starts with data collection

- **Plan data recording** to cover all sources of variation related to

✓ target (different events, anomalies)
✓ background (noise, different environments and conditions)
✓ equipment (different sensors, placement).

- **Collect iteratively**

ML works best as an iterative process.

✓ Collect data to build a basic model that proves the effectiveness of the technique, even if not yet for the full range of expected variations.
✓ Take more data and build a model to get acceptable accuracy in wide range of variations.
✓ Take more data and focus on model optimization to get the best possible performance.

All starts with data collection

Data collection and labelling is the most expensive, most time-consuming aspect of any ML project.

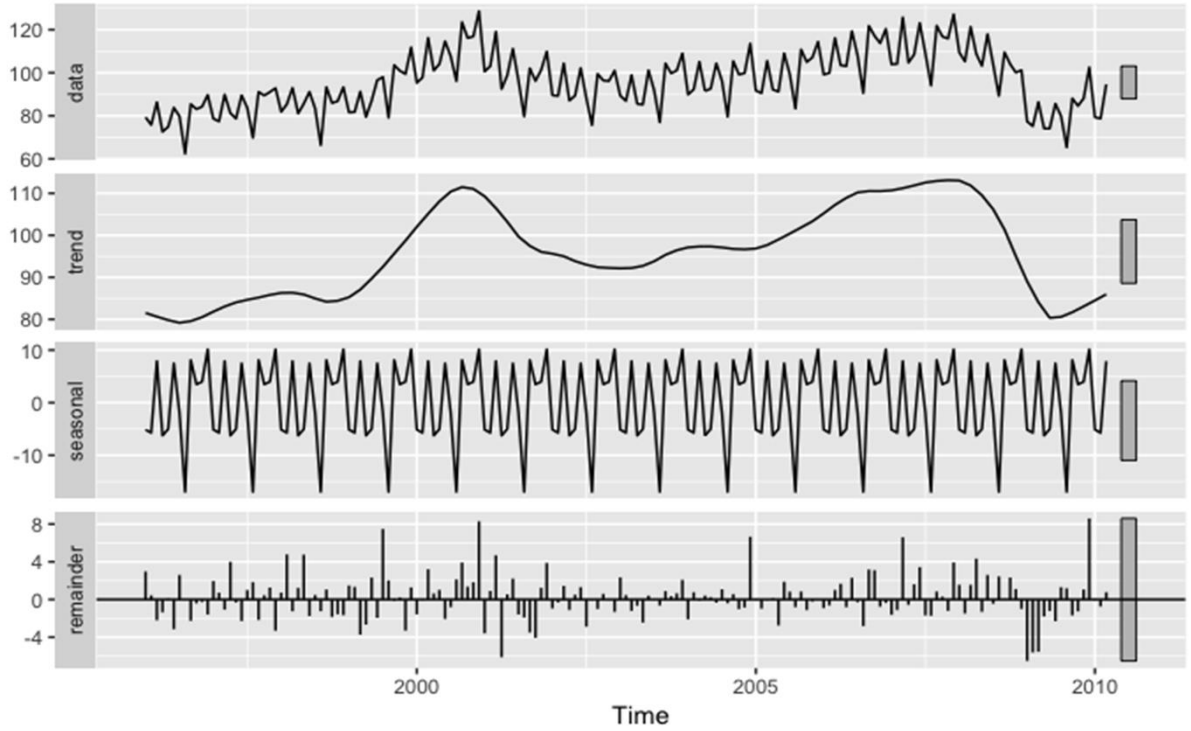
- ✓ **Use rich data.** Collect the least pre-processed signal data from sensors. Compression algorithms often discard important information.

- ✓ **Use too high sample rates.** Collect at the highest sample rate possible in the early stages. Down-sampling in software is easy and cheap. Going back and recollecting data is difficult and expensive.

Data Preprocessing & Feature Extraction

Data preprocessing
Remove outliers, impute missing data, normalize data

Feature extraction (hand-crafted features by domain experts)
Traditional statistics: mean, median, standard deviation, auto-correlation
More parameters : Skewness, Kurtosis
Frequency domain: power spectrum, dominant frequency
Tme series: time related features trend, seasonality, residuals per month, day of week, hour.



Feature Engineering

Ex. Monitoring computers in a data center

Basic features:

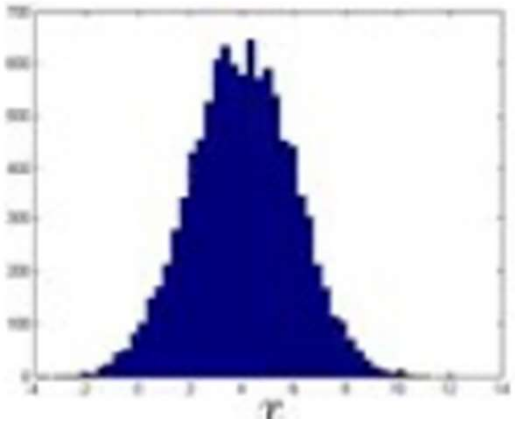
- memory use of computer
- number of disc accesses /sec
- CPU load
- network traffic

New features:

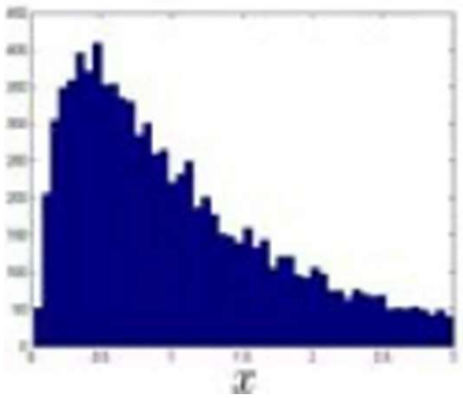
- CPU load /network traffic
- (CPU load)^2 /network traffic

Feature Engineering

Anomaly detection often assume the feature has a Gaussian distribution



….and if not ?

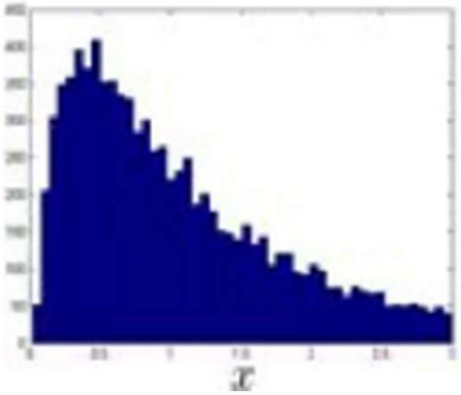


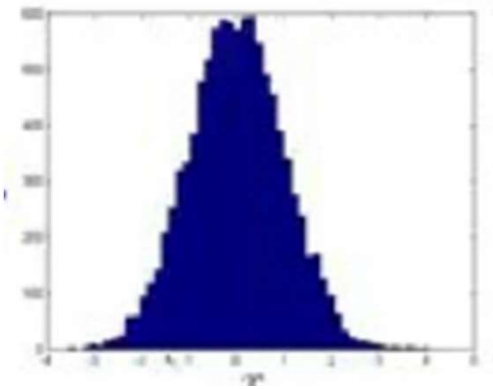
Feature Engineering

Popular feature transformations =>

- $log(x)$
- $log(x + c)$
- $\sqrt{x}$

for example: x => log(x) => better Gaussian curve





Machine Learning Approaches

Supervised Learning

Given examples with “correct answer” (labeled examples)

(e.g. given dataset with spam/not-spam labeled emails)

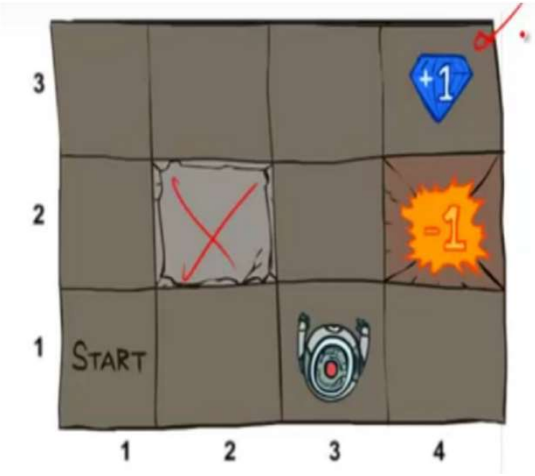
Unsupervised Learning

Given examples without answers (no labels).

Reinforcement Learning

On-line learning by taking actions and getting rewards/penalties.

(intelligent robotics =>



Deep Learning

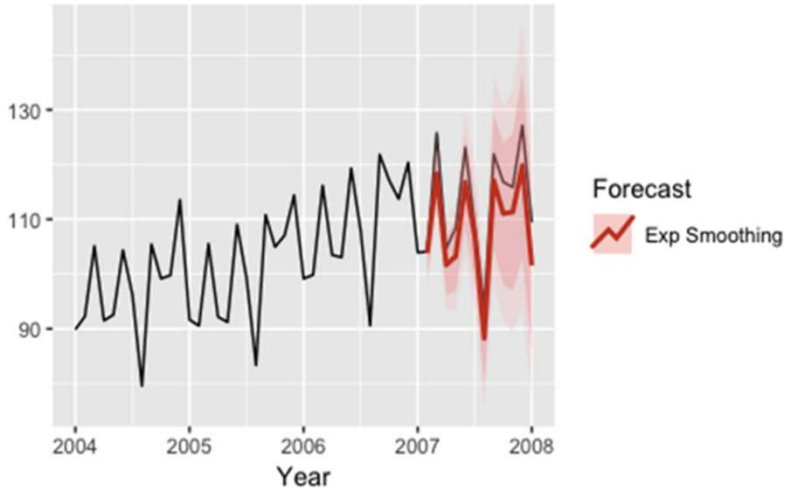
Automatically extract hidden features (in contrast to hand-crafted features). Need a lot of data (Big data) . Need for very high computational resources (GPUs).

Supervised Learning

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

Regression: The model output is a real number

Ex. Time series forecasting (predict the network traffic)



Classification: The model output is a label (e.g. 0, 1).

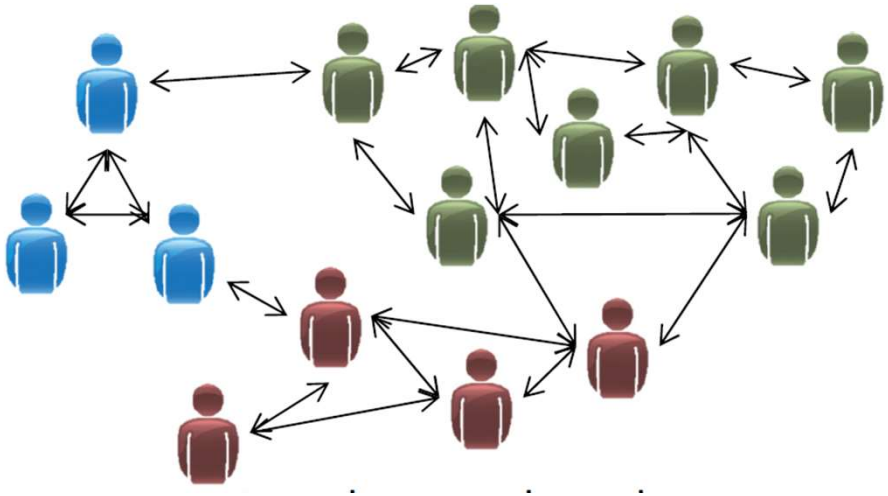
Ex. Learn to predict normal (0) or abnormal (1) state of data center computers:

- memory use of computer ; number of disc accesses /sec;
- CPU load ; network traffic

Unsupervised Learning

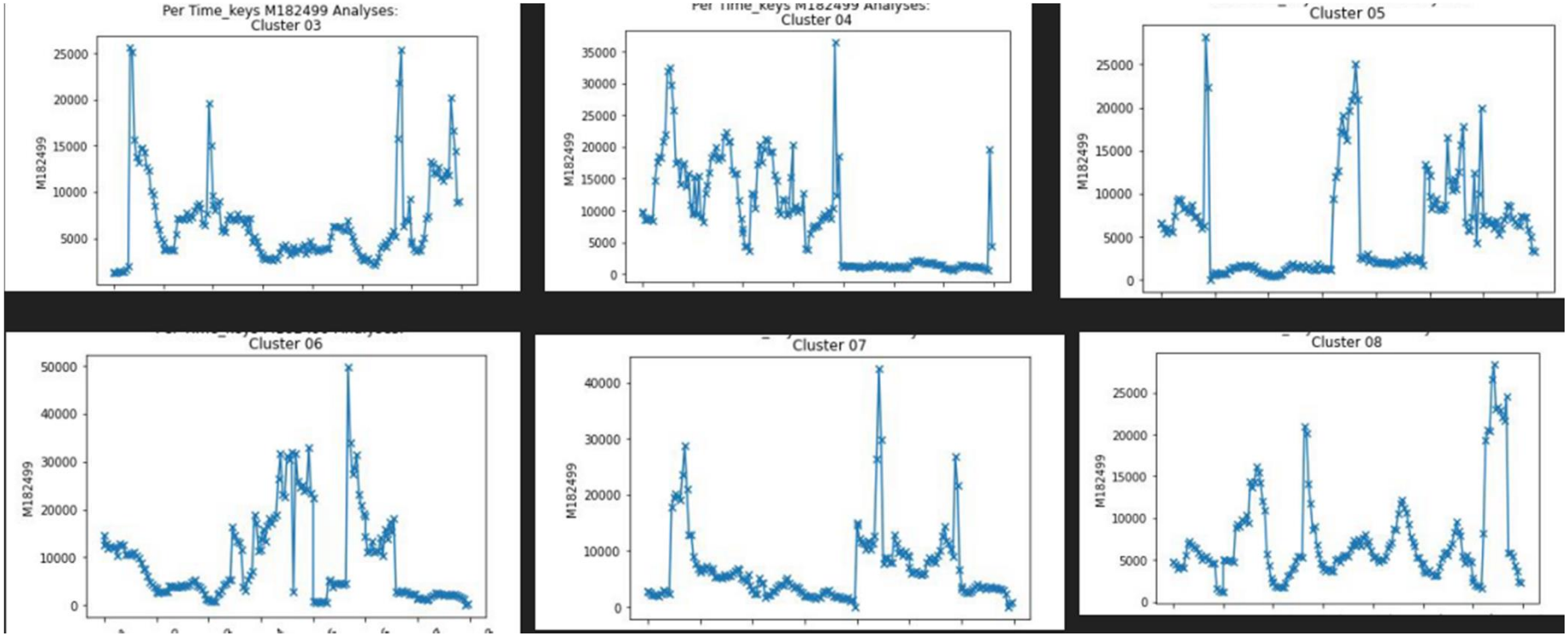
Given unlabeled data (examples without answers).

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

Raw Data Clustering



Preprocessing : Group data into clusters with similar patterns

Data Types

Numeric (Quantitative) features - Integer /Real numbers

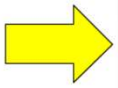
Boolean – True/False

Categorical features - days of the week, seasons, country, colors, etc.

How to deal with categorical features ? –

One-hot encoding (1,0) transforms *n* categories into *n* features

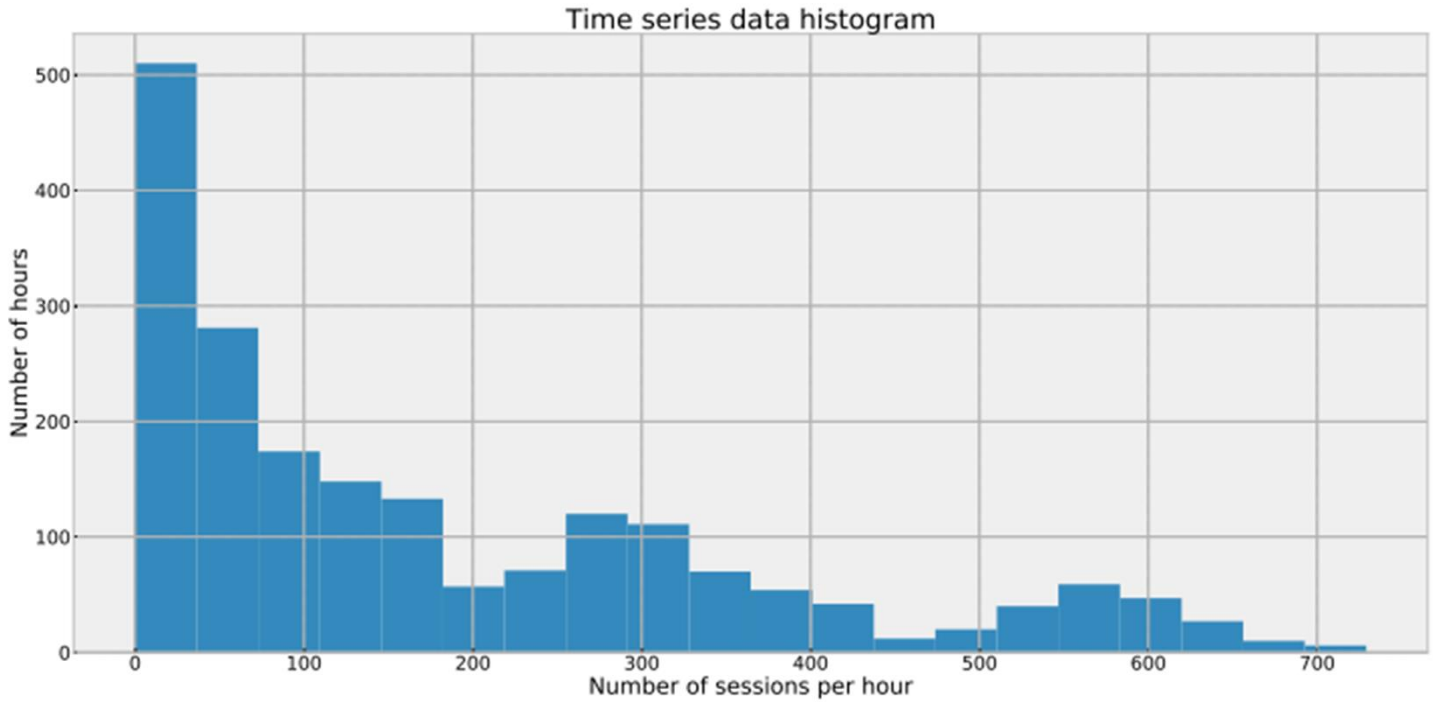
Color
Red
Red
Yellow
Green
Yellow



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

Data Visualization

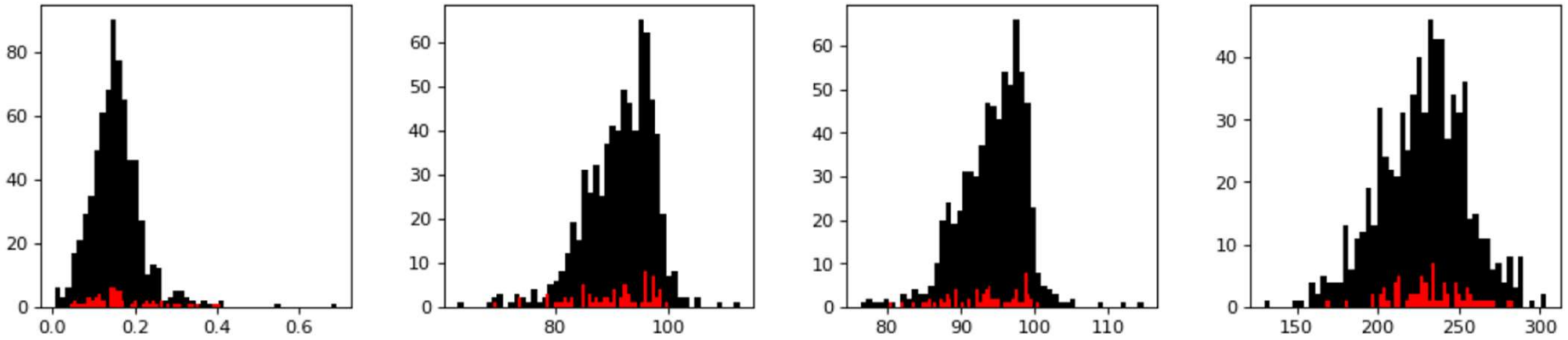
- **Histograms**
  - Show the distribution of values of a single feature
  - Divide the range of values of a single feature into bins and show bar plots of the number of examples in each bin.
  - Histogram shape depends on the number of bins



universidade de aveiro

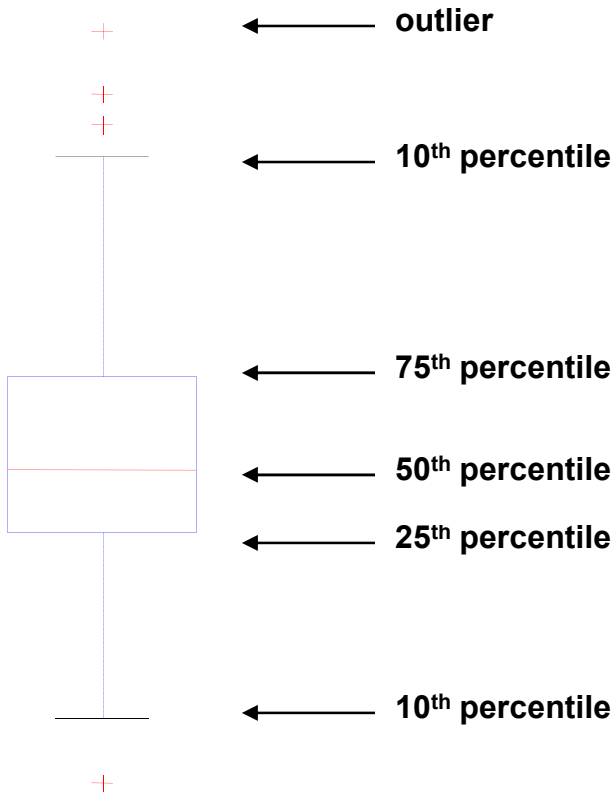
Data Visualization

- **Histograms with** class distribution notation



Data Visualization

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



ML Lab framework

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/

Recommended BIBLIOGRAPHY

Books:

- Andrew Ng, "Machine Learning Yearning", 2018
  https://www.deeplearning.ai/machine-learning-yearning
- Ian Goodfellow, Yoshua Bengio, Deep Learning, MIT Press, 2016.
  http://faculty.neu.edu.cn/yury/AAI/Textbook/DeepLearningBook.pdf

ML courses:

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