A. Giovanidis 2019

Introduction

Data Analysis for Networks - NDA (2019-2020)

Anastasios Giovanidis

Sorbonne-LIP6

September 11, 2019

Course (main) Bibliography

A. Giovanidis 2019

B.1 Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. "An introduction to statistical learning: with applications in R". Springer Texts in Statistics.

ISBN 978-1-4614-7137-0 (DOI 10.1007/978-1-4614-7138-7)

B.2 C. Bishop, "Pattern Recognition and Machine Learning", Springer 2006.

ISBN 978-0387-31073-2

B.3 H. Pishro-Nik, "Introduction to probability, statistics, and random processes", available at https://www.probabilitycourse.com, Kappa Research LLC, 2014.

Surveys - Overview

A. Giovanidis 2019

S.1 Raouf Boutaba et al. - "A comprehensive survey on machine learning for networking: evolution, applications and research opportunities", Journal of Internet Services and Applications, Springer (2018) 9:16
DOI 10.1186/s13174-018-0087-2

Stats VS Machine Learning

A. Giovanidis 2019

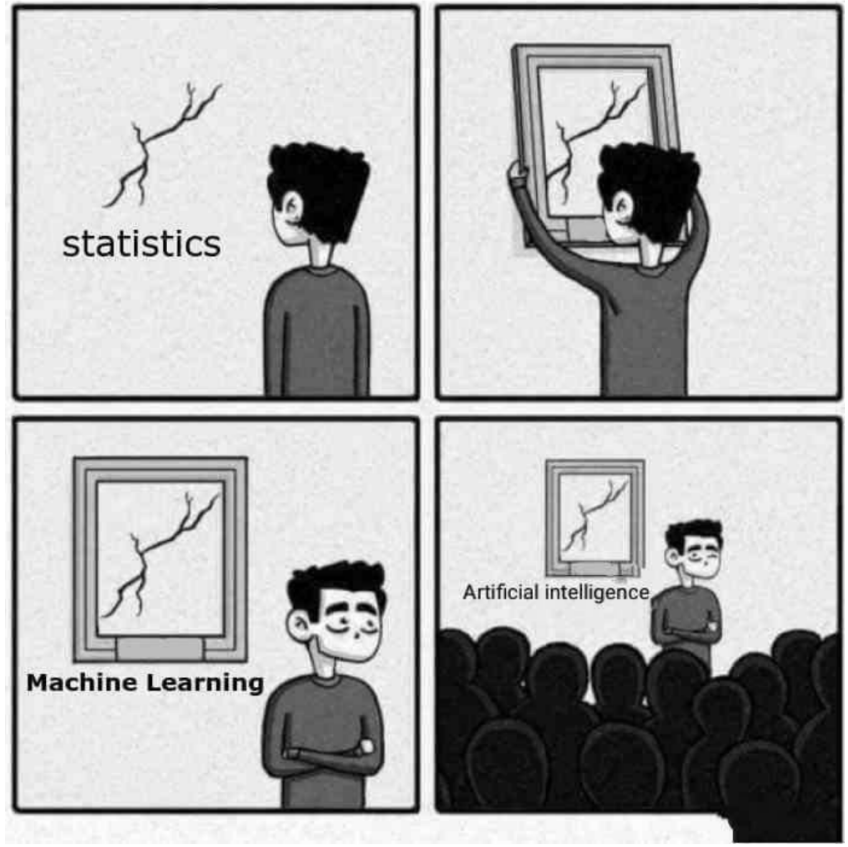


Figure: "When you're fundraising, it's AI. When you're hiring, it's ML. When you're implementing, it's logistic regression."

A. Giovanidis 2019

Intro

Data Analysis and Machine Learning (ML) revolutionise our world!

- Computer Vision (CV) and Natural Language Processing (NLP): classifying images, facial recognition, automatic translation.
- Recommendation engines: Amazon, Netflix, or Youtube.

Been around since a very long time...

- *Statistics is a branch of mathematics dealing with the collection, analysis, interpretation, and presentation of masses of numerical data.*
- *"Machine Learning, is the field of study that gives computers the ability to learn without being explicitly programmed" (Arthur Samuel, 1959)*

Why now? Sufficient and cheap computational power & lots, lots, lots of (labeled) data available e.g. Facebook and Google photos, WWW...

Highlights I

A. Giovanidis 2019

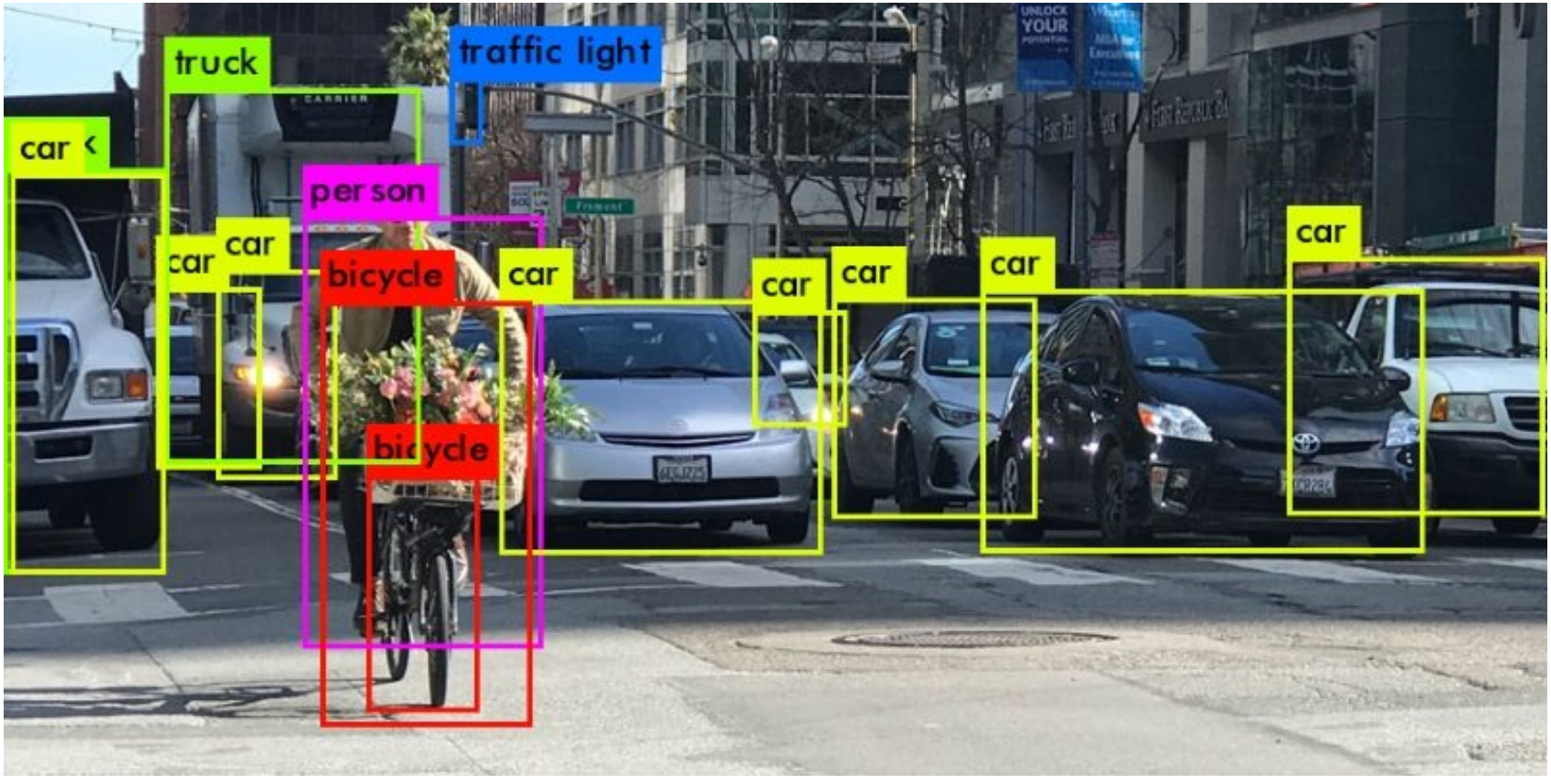


Figure: Object detection and recognition for driverless cars.

Highlights II

Behind Hey Siri: How Apple's AI-Powered Personal Assistant Uses DNN

ABHISHEK SHARMA - FEB 16, 2018

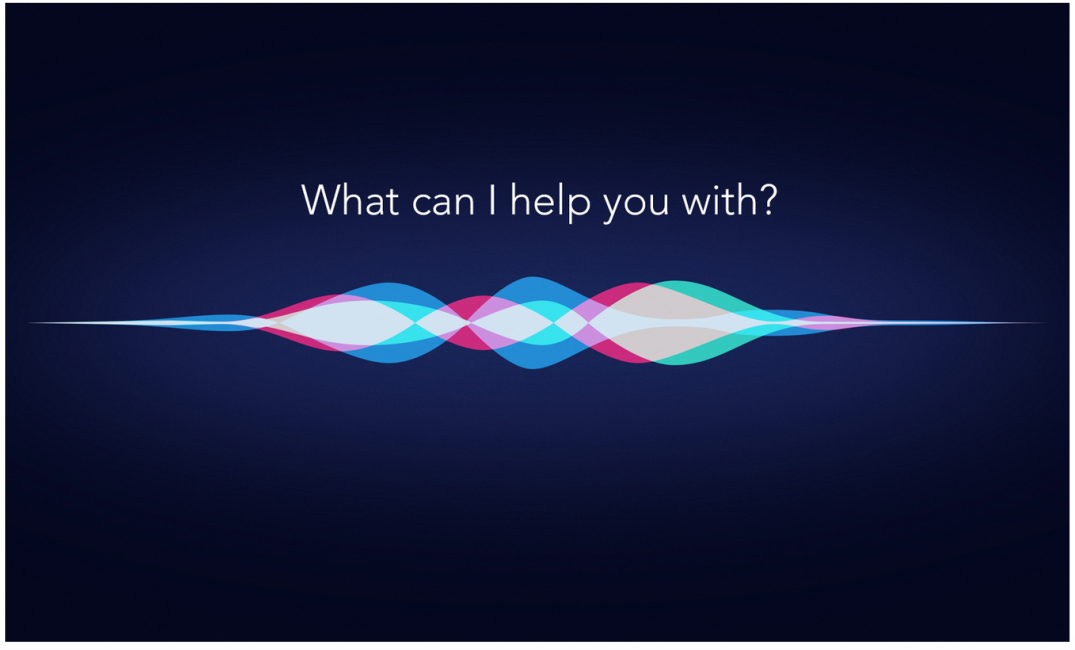


Figure: Speech recognition.

Highlights III

A. Giovanidis 2019

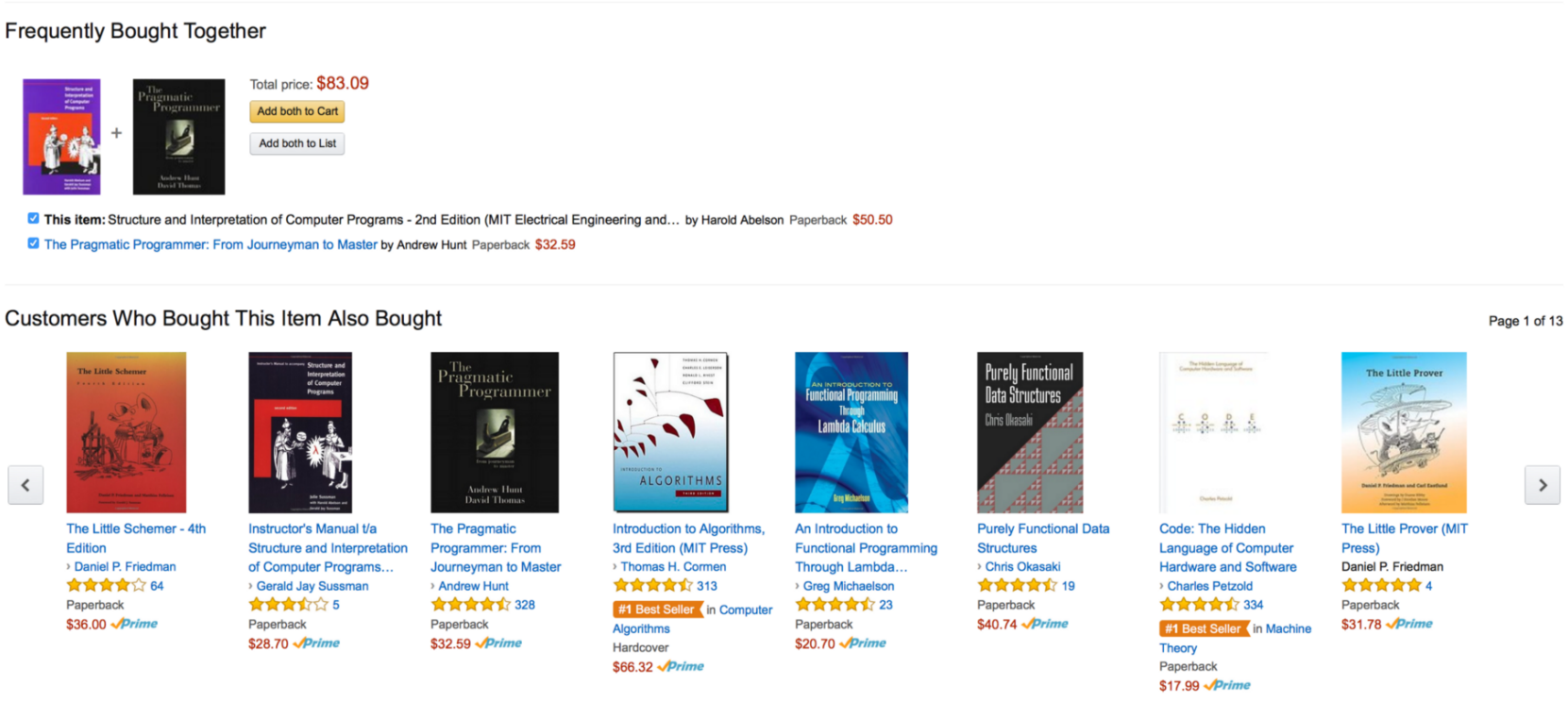


Figure: Useful recommendations.

Taxonomy of ML methods

A. Giovanidis 2019

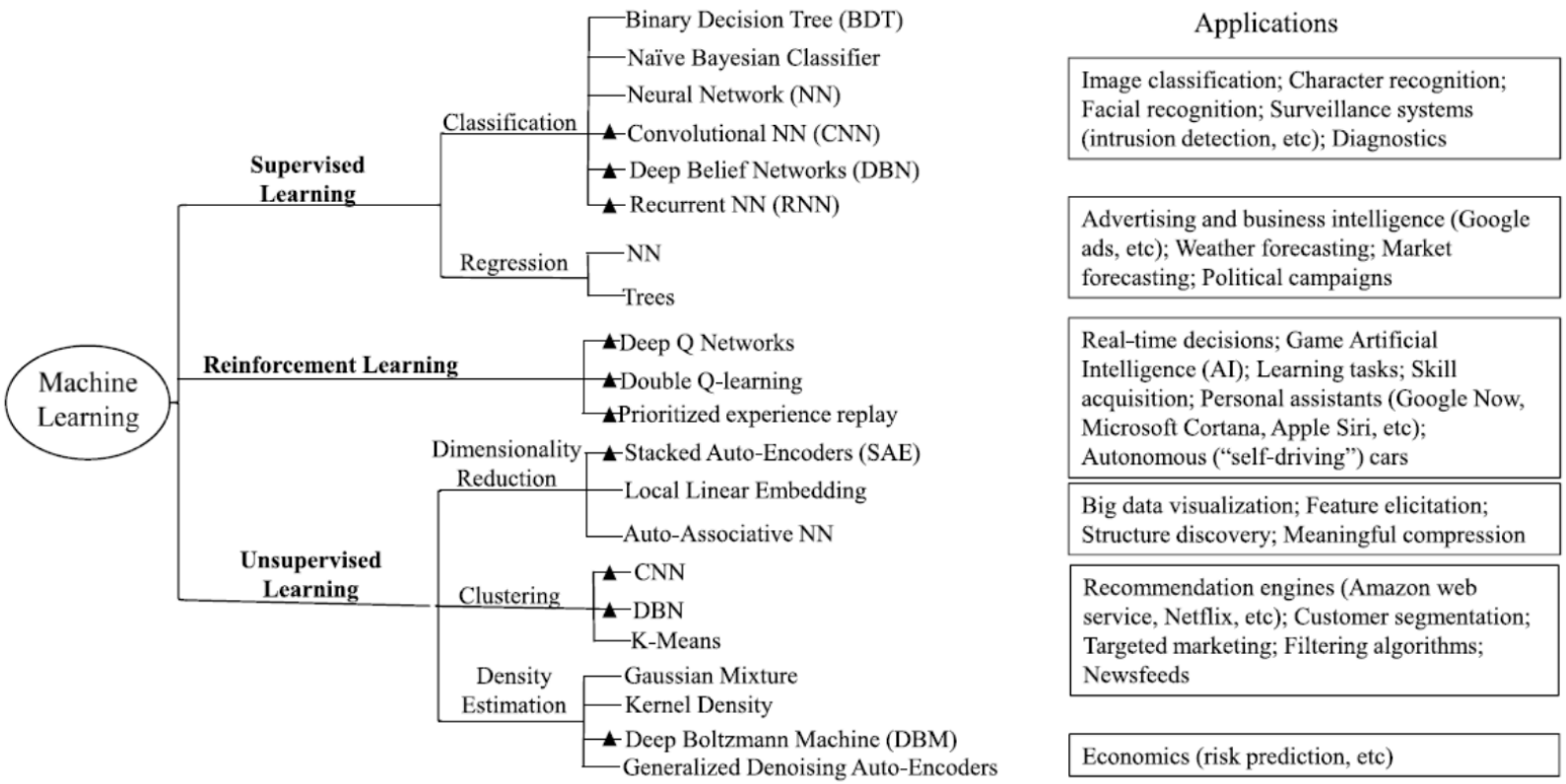


Figure: Taxonomy and applications (Fadlullah, et al (IEEE, 2017)).

A. Giovanidis 2019

Method differences

All three methods require a common element to work:

DATA!!!

The difference is the type of data available or collected:

- Supervised: Labelled data, model learning.
- Unsupervised: Unlabelled data (majority of telecom data).
- Reinforcement: Exploration-exploitation. Data is the rewards collected by application of an action.

Labeling is a non-trivial process to establish the ground-truth. Often hand-made by experts.

A. Giovanidis 2019

Method differences

All three methods require a common element to work:

DATA!!!

The difference is the type of data available or collected:

- Supervised: Labelled data, model learning.
- Unsupervised: Unlabelled data (majority of telecom data).
- Reinforcement: Exploration-exploitation. Data is the rewards collected by application of an action.

Labeling is a non-trivial process to establish the ground-truth. Often hand-made by experts.

☞ Make a distinction between static and dynamic environments: Data from the first are n-dimensional points, from the second **time-series**.

History of Data Analysis and ML methods

A. Giovanidis 2019

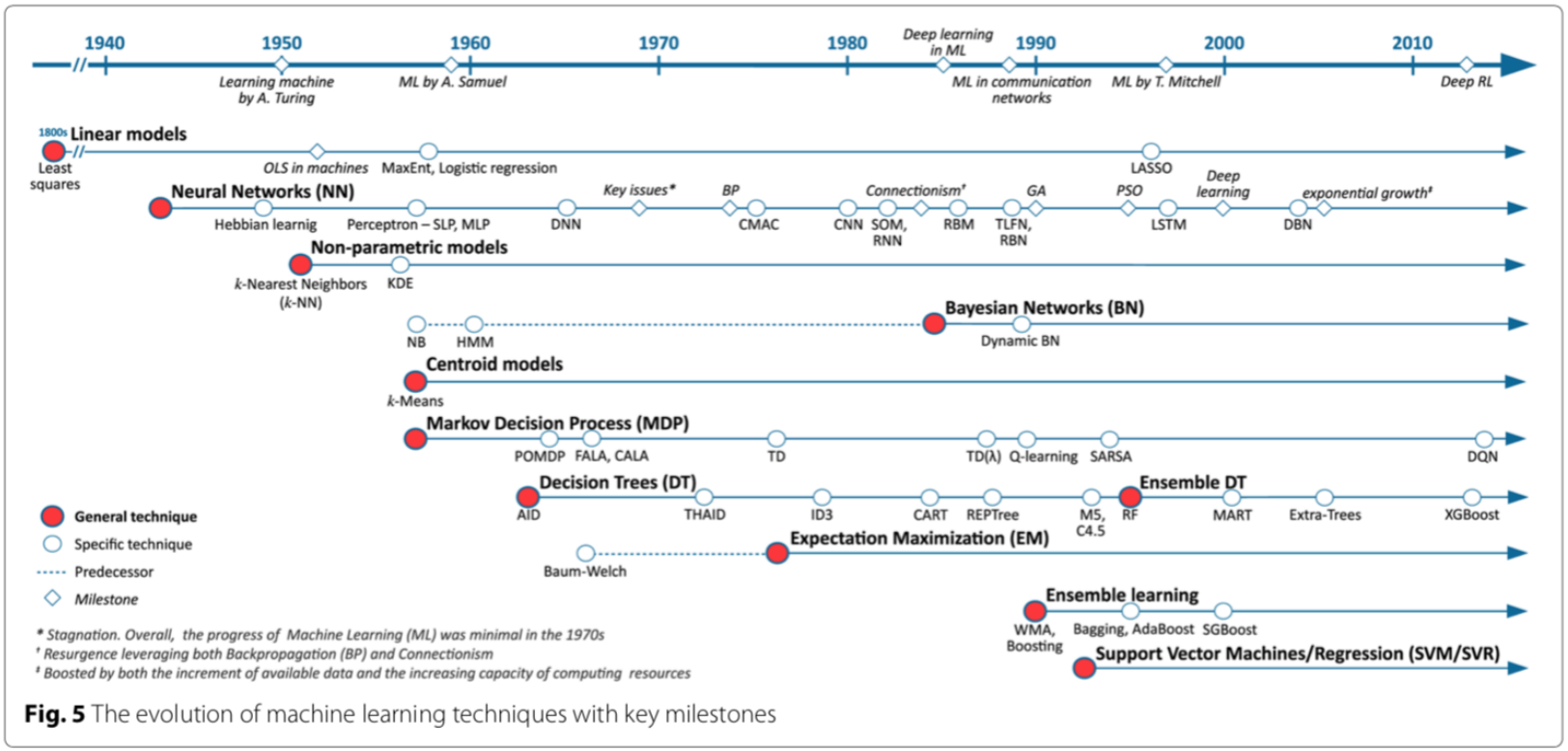


Fig. 5 The evolution of machine learning techniques with key milestones

Figure: ML historical evolution (from [S.1]).

A. Giovanidis 2019

Main tasks to perform

What can we do with all these methods?

- Estimation: quantify unknown parameters from observations.
- Inference: guess the unknown underlying statistics.
- Regression: guess an underlying model and predict possible outcomes of an experiment.
- Classification: decide on the class of an object.
- Dimensionality Reduction: compress the information contained in several features to easier describe an object.
- Clustering: group objects based on affinity.

Some Tasks

A. Giovanidis 2019

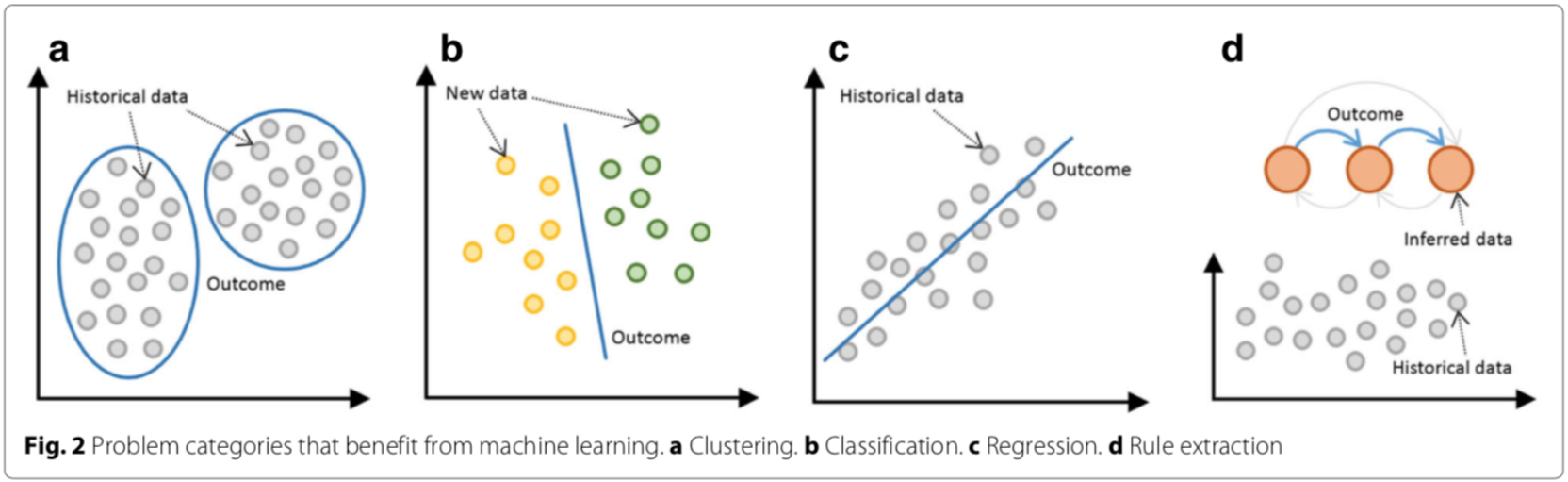


Fig. 2 Problem categories that benefit from machine learning. **a** Clustering. **b** Classification. **c** Regression. **d** Rule extraction

Figure: Task examples (from [S.1]).

General methodology

A. Giovanidis 2019

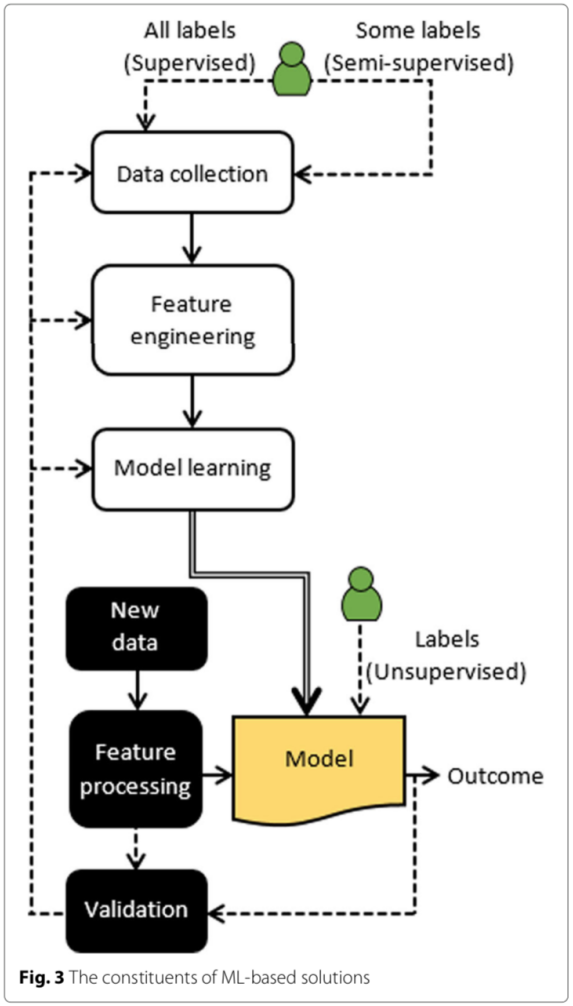


Figure: ML historical evolution (from [S.1]).

Telecom Network science and Data

A. Giovanidis 2019

☞ Telecommunication networks offer the infrastructure for ML.

But! Their design and functionality can profit from data analysis and ML, through Telemetry: massive data availability about QoS, QoE, KPIs...

Main possibilities:

1. Traffic: prediction, classification, routing.
2. Performance: congestion control, resource management, fault management, QoS/QoE management.
3. Anomaly detection: hardware/software failure.
4. Security: Intrusion detection, DoS or DDoS Attacks.

Traffic IP

A. Giovanidis 2019

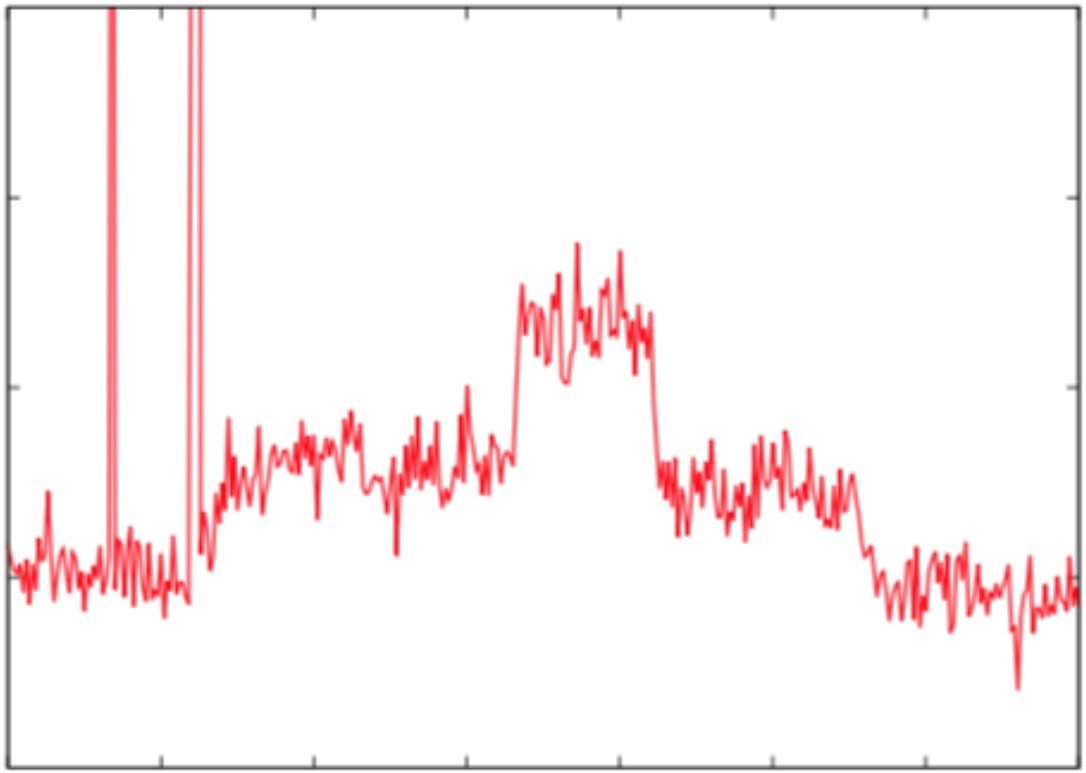


Figure: *image from thesis Audrey Wilmet*.

A. Giovanidis 2019

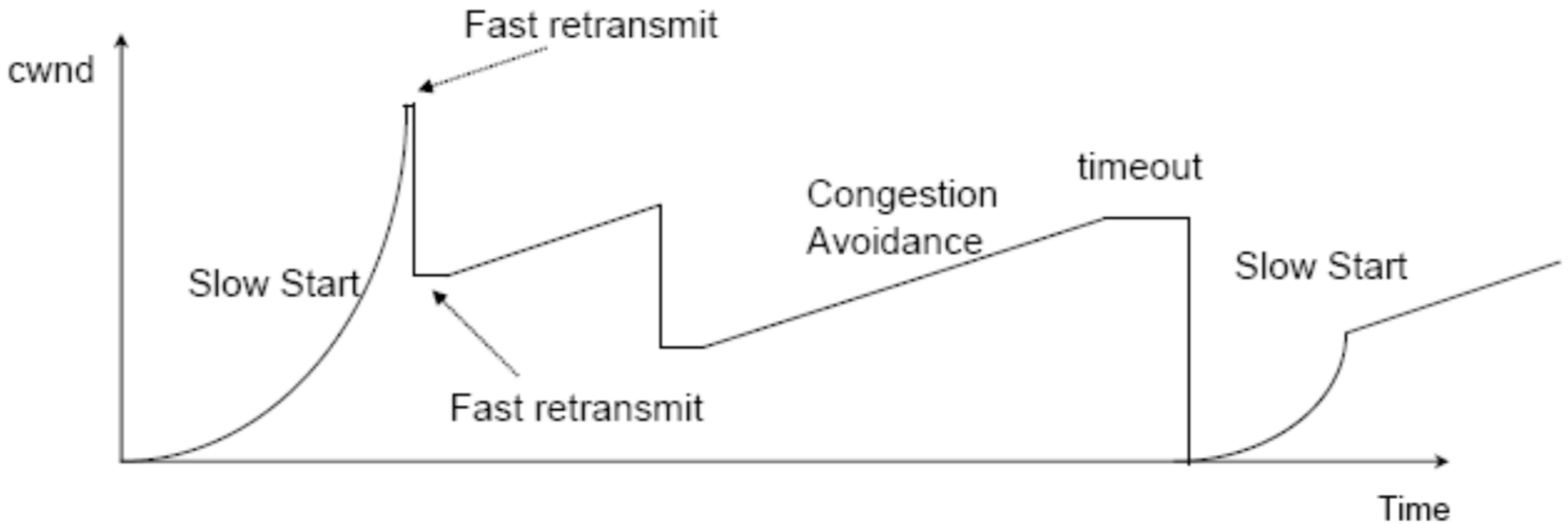
Traffic

- **Prediction**
  Forecast future traffic from previously observed data: Time series forecasting through ARMA models (auto-regressive moving average)
- **Classification**
  Associate network traffic to pre-defined classes, e.g. HTTP, FTP, WWW, DNS, P2P
  or applications, e.g. Skype, YouTube, Netflix...
  Features: port number, packet payload, host behaviour, flow features, QoS requirements.
  Traffic can be encrypted! Rely on stochastic characteristics.
- **Routing**
  Select a path for packet transmission with an objective: cost minimisation, link utilisation, QoS provisioning, etc.

  Use of Reinforcement Learning techniques, to explore the environment without supervision (trial-and-error learning).

TCP

A. Giovanidis 2019



TCP congestion control

A. Giovanidis 2019

TCP protocol limits the packet sending rate when congestion is detected.

But! TCP recognizes and handles all packet losses as network congestion (buffer overflow).
A packet loss can be due to other reasons:

- Packet reordering.
- Fading and shadowing in wireless.
- Wavelength contention in optical.

Solution: Classify the cause of packet loss and reduce TCP transmission rate only when congestion.

Features: inter-arrival time, round-trip time, one-way delay.

☞ Also, learn the appropriate window reduction per congestion event!

A. Giovanidis 2019

Network security

Protect the network against cyber-threats.
Attacks can compromise the network's availability and resources.

☞ Businesses are under security threats → cost billions in damage and recovery, may have impact on their reputation.

Current Security measures include :

- Encryption of network traffic, Anti-viruses, Firewalls, etc.

☞ Extra protection:

- Intrusion Detection/Prevention: phishing, DoS, DDoS, ...

Monitor the network for malicious / anomalous activities, find patterns (=attack signatures) in big datasets that deviate from normal behaviour.

What is normal? Unsupervised learning, clustering methods.

Structure of the course I

A. Giovanidis 2019

☞ Methods from statistics, machine-learning and stochastic processes.

Each course on Wednesdays: 2 hours Theory + 2 hours Python Lab

Part I: Statistics

- C1. Intro to NDA / Probability basics (18 September 2019)
- C2. Frequentist Estimation (25 September 2019)
- C3. Hypothesis Tests (02 October 2019)
- C4. Bayes Rule (09 October 2019)

A. Giovanidis 2019

Structure of the course II

Part II: Machine Learning

a. Supervised

- C5. Regression (16 October 2019)
- C6. Mode Selection / Validation (23 October 2019)
- C7. Classification (06 November 2019)
- C8. Feature Selection / Regularisation (27 November 2019)
- C9. Tree-based methods (04 December 2019)

b. Unsupervised

- C10. Clustering (11 December 2019)
- C11. Anomaly Detection / PCA (18 December 2019)

Structure of the course III

A. Giovanidis 2019

Part III: Time-series

- C12. Principles of Time-Series (08 January 2020)
- C13. Modelling with Time-Series (15 January 2020)
- C14. Forecasting using Time-Series (22 January 2020)

☞ End January / begin February 2020 final exam.

Final Note:
40% Python code from all TPs
60% Final exam.

Teaching material

Course Material (slides):
https://github.com/yokaiAG/DataNets-Course

People:

- Anastasios Giovanidis (responsible)
- Maximilien Danisch, Lionel Tabourier.

A. Giovanidis 2019

END