

# 0. Introduction

Network Data Analysis - NDA (2022–2023)

Anastasios Giovanidis

Sorbonne-LIP6



## Course (main) Bibliography

- 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.

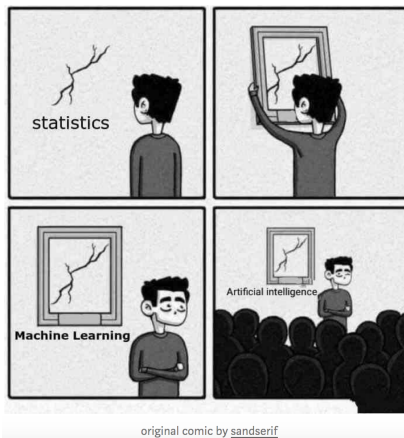
## Surveys - Overview

A. Giovanidis 2022

- 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 2022



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

# 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.

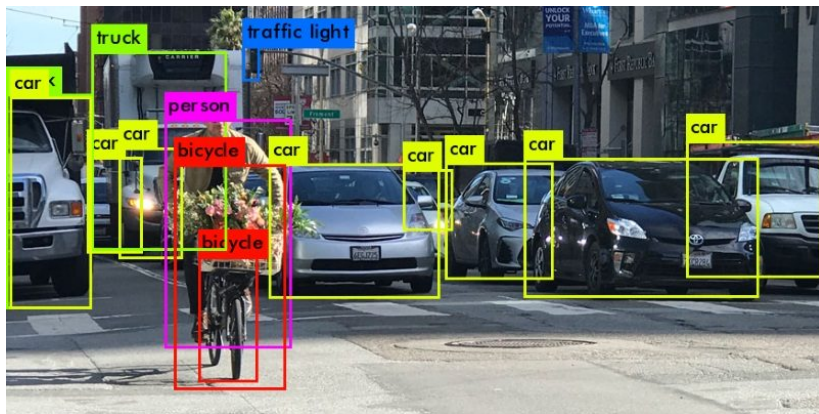
Been around since a very long time...

- ▶ *Statistics is a branch of mathematics dealing with the collection, analysis, interpretation and presentation of massive 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...

# Object detection

A. Giovanidis 2022



# Speech recognition

A. Giovanidis 2022

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

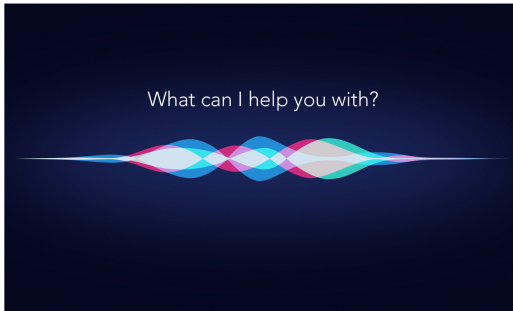
ABHISHEK SHARMA - FEB 16, 2018

f

t

in

What can I help you with?



## Useful recommendations

A. Giovanidis 2022

## Frequently Bought Together



+

Total price: **\$83.09**[Add both to Cart](#)[Add both to List](#)

- ☒ **This Item:** Structure and Interpretation of Computer Programs - 2nd Edition (MIT Electrical Engineering and... by Harold Abelson Paperback **\$50.50**
- ☒ **The Pragmatic Programmer: From Journeyman to Master** by Andrew Hunt Paperback **\$32.59**

## Customers Who Bought This Item Also Bought

Page 1 of 13



**The Little Schemer** - 4th Edition  
 > Daniel P. Friedman  
 ★★★★★ 64  
 Paperback  
**\$36.00** ✓Prime



**Structure and Interpretation of Computer Programs...**  
 > Gerald Jay Sussman  
 ★★★★★ 5  
 Paperback  
**\$28.70** ✓Prime



**The Pragmatic Programmer: From Journeyman to Master**  
 > Andrew Hunt  
 ★★★★★ 328  
 Paperback  
**\$32.59** ✓Prime



**Introduction to Algorithms**, 3rd Edition (MIT Press)  
 > Thomas H. Cormen  
 ★★★★★ 313  
**#1 Best Seller** in Computer Algorithms  
 Hardcover  
**\$66.32** ✓Prime



**An Introduction to Functional Programming Through Lambda Calculus**  
 > Greg Michaelson  
 ★★★★★ 23  
 Paperback  
**\$20.70** ✓Prime



**Purely Functional Data Structures**  
 > Chris Okasaki  
 ★★★★★ 19  
 Paperback  
**\$40.74** ✓Prime



**Code: The Hidden Language of Computer Hardware and Software**  
 > Charles Petzold  
 ★★★★★ 334  
**#1 Best Seller** in Machine Theory  
 Paperback  
**\$17.99** ✓Prime



**The Little Prover** (MIT Press)  
 > Daniel P. Friedman  
 ★★★★★ 4  
 Paperback  
**\$31.78** ✓Prime





# Taxonomy of ML methods

A. Giovanidis 2022

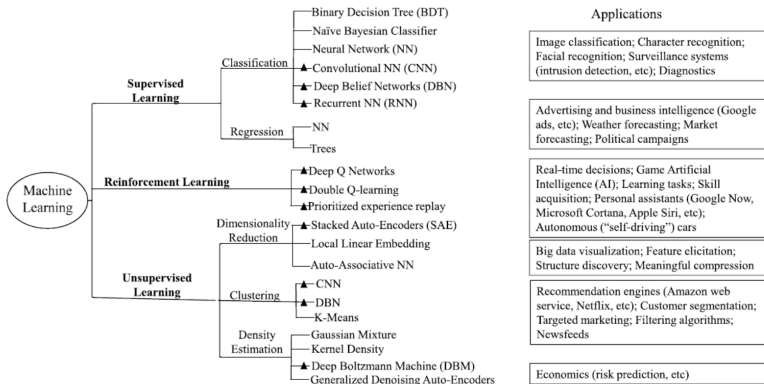


Figure: Taxonomy and applications ( Fadlullah, et al (IEEE, 2017)).

## 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.

# 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 2022

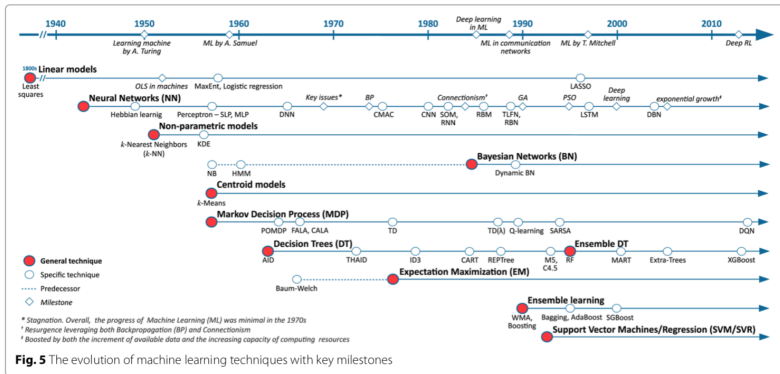


Figure: ML historical evolution (from [S.1]).

# 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 2022

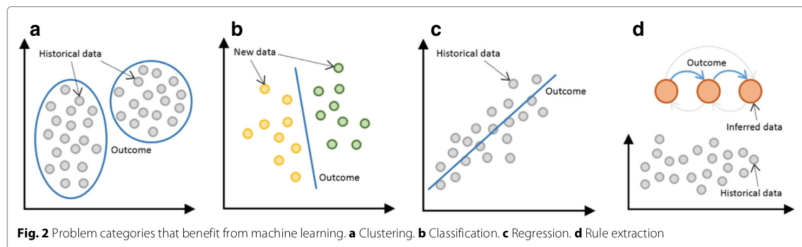


Figure: Task examples (from [S.1]).

# General methodology

A. Giovanidis 2022

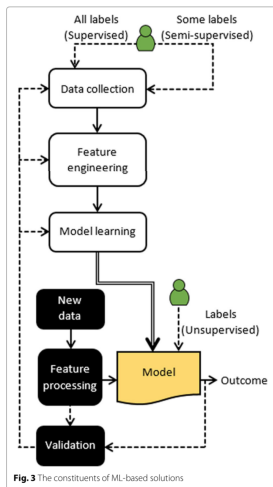


Figure: ML historical evolution (from [S.1]).

# Telecom Network science and Data

☞ 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, adaptive 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 2022

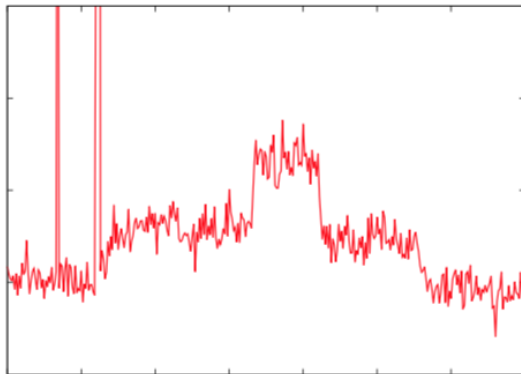


Figure: *image from thesis Audrey Wilmet.*

# 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!*

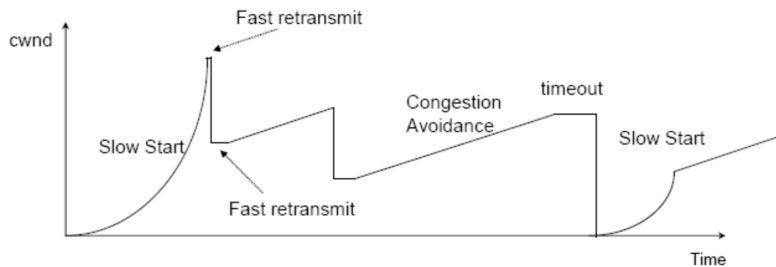
## ► Routing

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 2022



# TCP congestion control

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!

# 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

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

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

## Part I: Statistics

- ▶ C1. Intro to NDA / Probability basics (20 September 2022)
- ▶ C2. Frequentist Estimation
- ▶ C3. Hypothesis Tests
- ▶ C4. Bayes Rule

# Structure of the course II

## Part II: Machine Learning

### a. Supervised

- ▶ C5. Regression pt.1
- ▶ C6. Regression pt.2
- ▶ C7. Cross-Validation
- ▶ C8. Classification
- ▶ C9. Trees-Forests
- ▶ C10. Regularisation or SVM

### b. Unsupervised

- ▶ C11. Clustering
- ▶ C12. PCA and Anomaly Detection

# Structure of the course III

## Part III: Time-series

### ► C13-C14. Time-Series

📅 End January – Begin February 2023 final exam.

Final Note:

**50% Python code from all TPs**

**50% Final exam.**



# Teaching material

A. Giovanidis 2022

Course Material (slides):

**<https://github.com/yokaiAG/DataNets-Course>**

Course responsible:

▶ Anastasios Giovanidis

Contact / Questions:

✉ **[anastasios.giovanidis@lip6.fr](mailto:anastasios.giovanidis@lip6.fr)**

**END**