Machine Learning: *a gentle introduction & info about the course*

Paola Velardi

What is machine learning (in a nutshell)

A set of methodologies to find regularities in data

→

These findings are used to **predict** future outcomes and/or to **prescribe** optimal strategies

What is the task?

- Predictive:
  - Given previous historical «labelled» data, learn a model to predict future outcomes (e.g., see what happened to past credit applicants, or to past patients, and learn what may happen to new applicants or new patients)
  - Examples: predict patients’ risk of a complication, predict future sales of a new product, users’ satisfaction in a market campaign..
- Prescriptive:
  - Given «unlabelled» data, or given an environment and some stimuli, prescribe «how to», e.g., best actions to be performed
  - Example: customer segmentation according to their profiles, best strategy to win a game, bets way for a robot to execute a given task – e.g., drive a car - …

Example of predictive learning: Credit risk assessment

Customer ID	AGE	INCOME	EDUCATION	DEFAULT
ID1	27	30.000	YES	1
ID2	50	45.000	NO	0
ID3	60	46.000	YES	0
……				
ID1348	32	55.000	YES	0

- Credit scoring is a fairly widespread practice in banking institutions, whose main objective is to discriminate between borrowers, based on their *credit worthiness*.
- Decision on whether granting credit to new customers is based on **past data on customers who experienced a default or not**
- Machine learning can help assessing the risk of default of new customers based on a «risk model» learned from **past data**

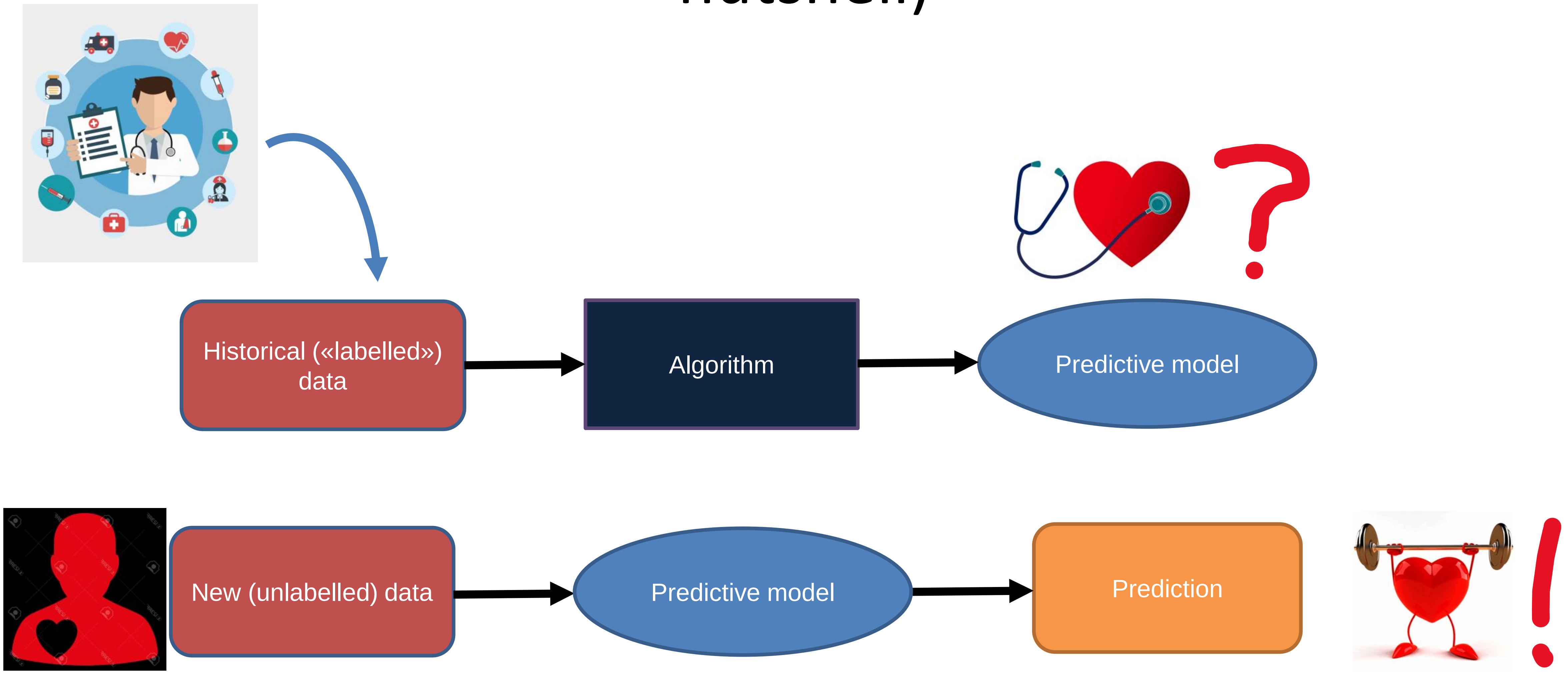
***** We say that data are «**labelled**» to mean that historical data include the label of the variable we want to predict, «default» in this example.

Example 2: cardiovascular risk assessment

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1
5	57	1	0	140	192	0	1	148	0	0.4	1	0	1	1
6	56	0	1	140	294	0	0	153	0	1.3	1	0	2	1
7	44	1	1	120	263	0	1	173	0	0.0	2	0	3	1
8	52	1	2	172	199	1	1	162	0	0.5	2	0	3	1
9	57	1	2	150	168	0	1	174	0	1.6	2	0	2	1

- Electronic patient records are now widely available. They collect the «history» of patients, their clinical tests, treatments and diseases
- Doctors can be supported in deciding the best therapy, or in estimating a specific risk of complications (e.g., cardiovascular risk) by machine learning systems, based on the analysis of historical (labelled) data of previous patients

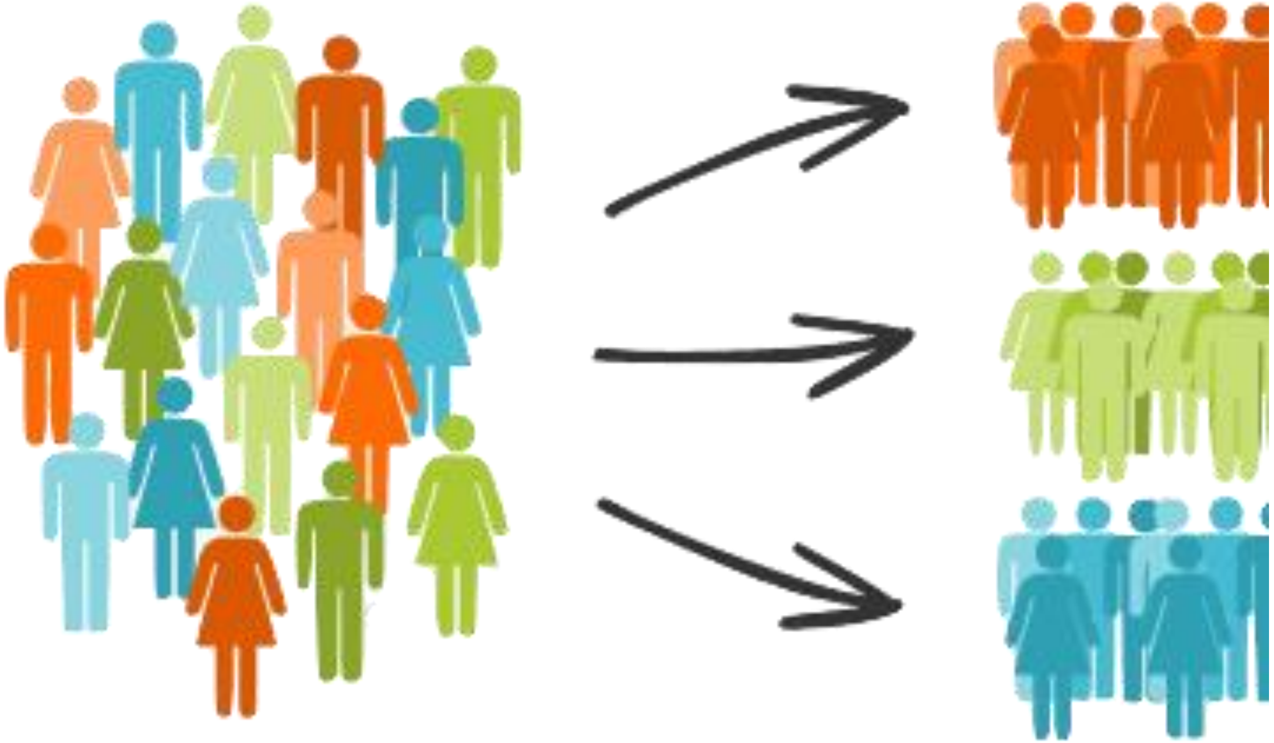
Basic workflow of a predictive ML system (in a nutshell)



What is the task?

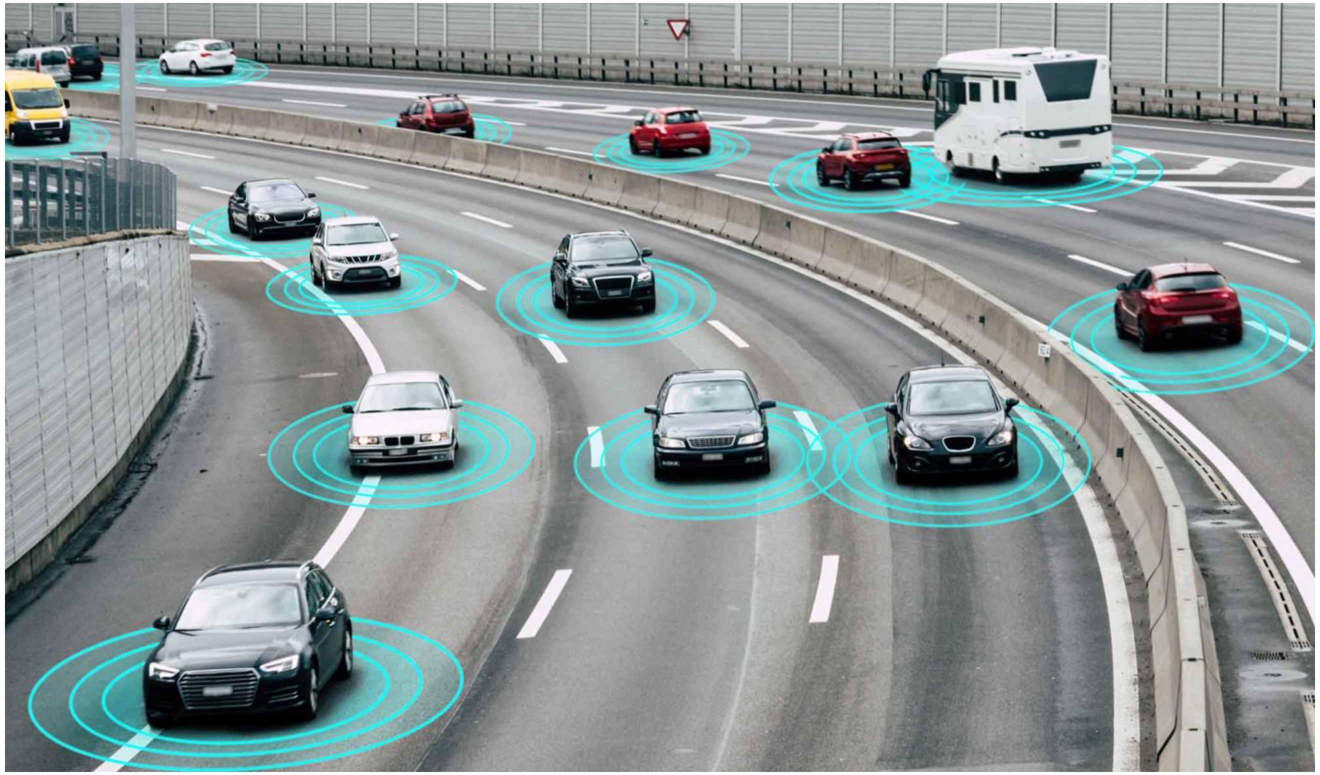
- Predictive:
  - Given previous historical «labelled» data, learn a model to predict future outcomes (e.g., see what happened to past credit applicants, or to past patients, and learn what may happen to new applicants or new patients)
  - Examples: predict patients' risk of a complication, predict future sales of a new product, users' satisfaction in a market campaign..
- Prescriptive:
  - Given «unlabelled» data, or given an environment and some stimuli, prescribe «how to», e.g., best actions to be performed
  - Example: customer segmentation according to their profiles, best strategy to win a game, bets way for a robot to execute a given task – e.g., drive a car - …

Example of predictive analytics: customer segmentation



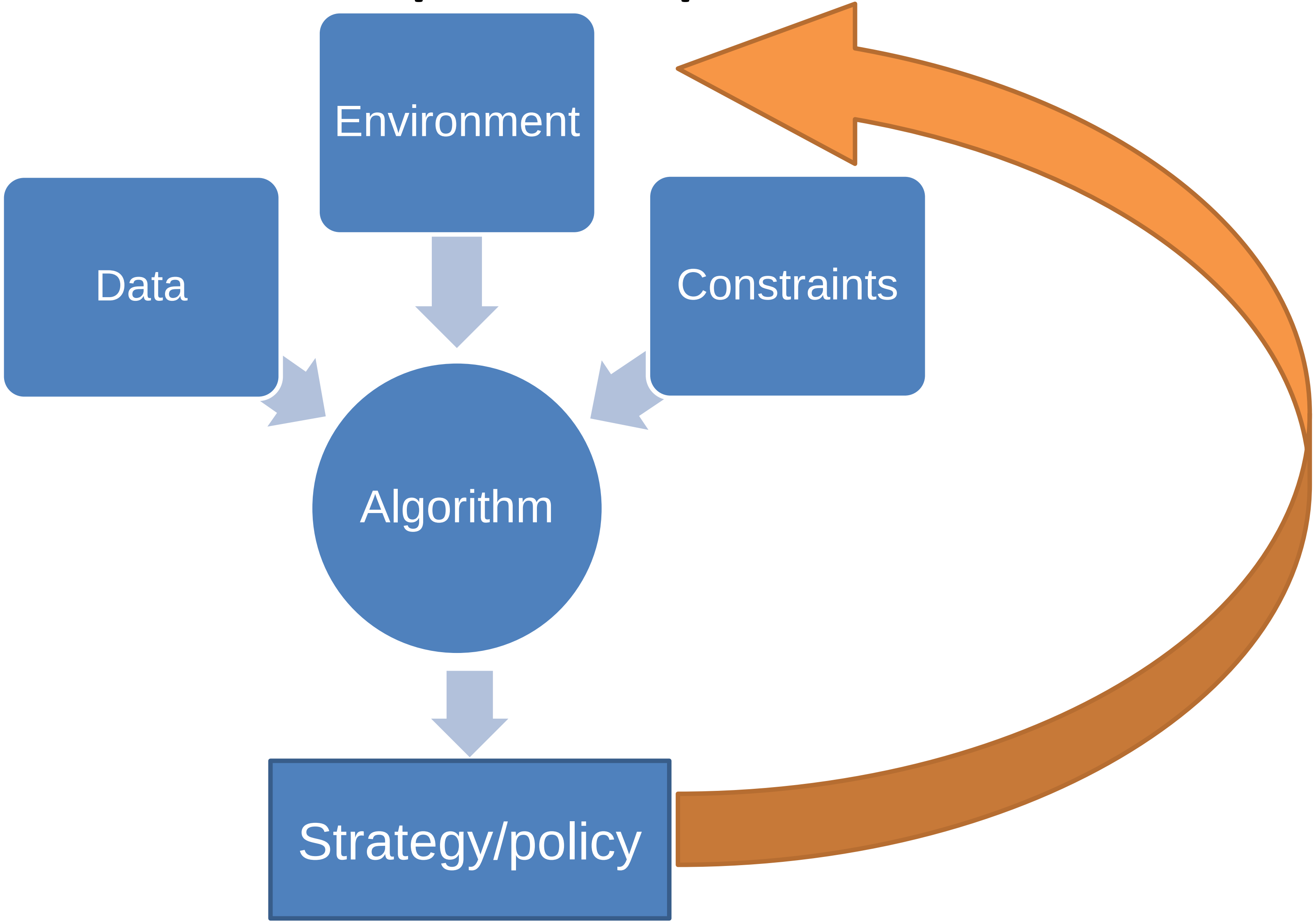
- Given data on customer profiles, cluster them into groups of «similar ones»
- Then, use these groups to identify best personalized marketing campaigns to optimize revenues

Example 2: self driving cars



- Analyse driving behaviours of million drivers
- Learn best strategy to react to the environment in any condition

Workflow of prescriptive ML



Data is the fuel of ML

- Data may come in different forms (tables, images, text, videos..)
- As we will see, it takes a lot of hard work to make data «ingestible» by ML algorithms
- Whatever it takes, it is worth: without the fuel of «good» data, algorithms just don’t work

Issues in Machine Learning

Issues in Machine Learning

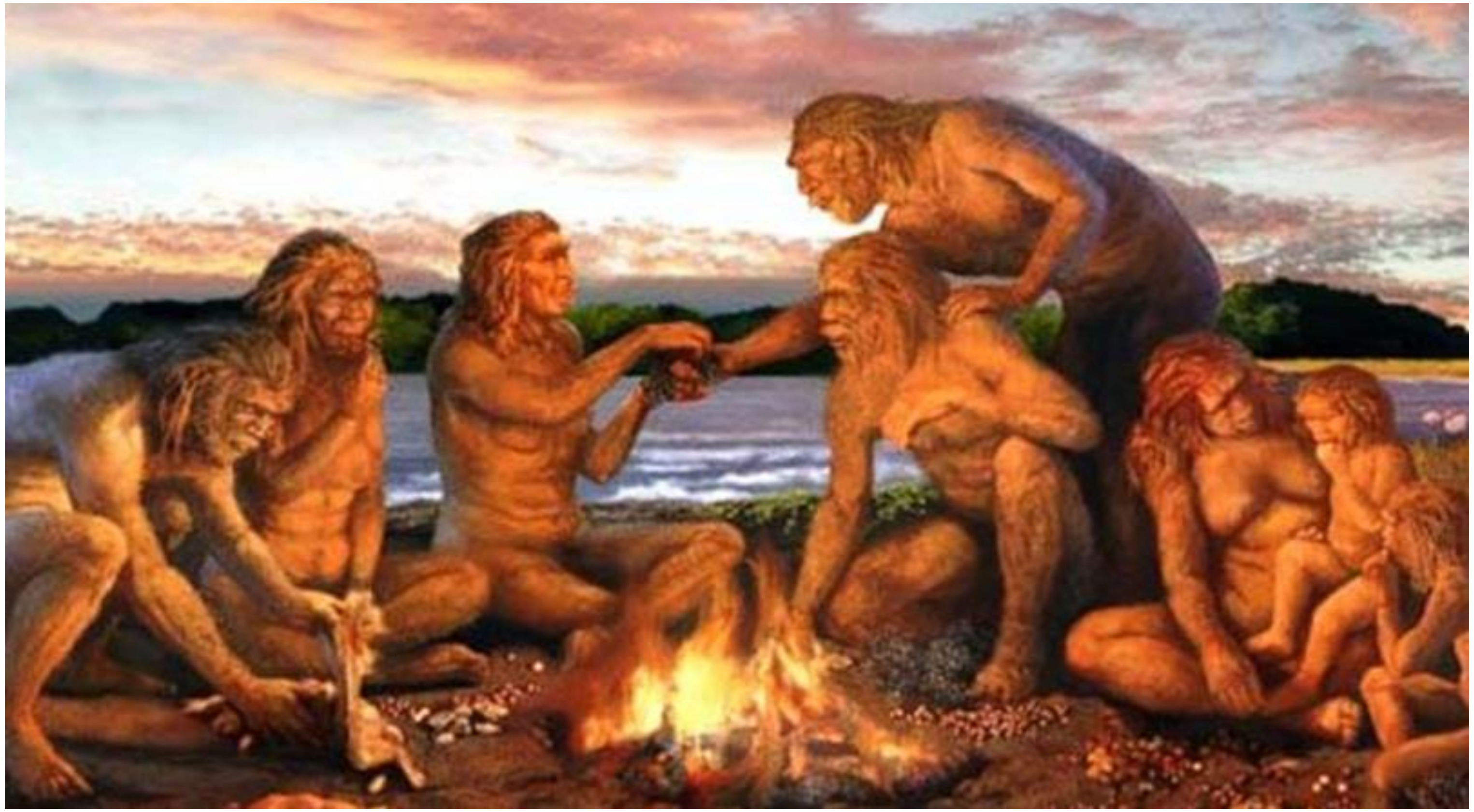
“*How can we program systems to automatically learn from «data» and to improve with experience*? “

- **What** is learning?
- **What** can we learn?
- **What** is “experience”??
- **How** do we learn?
- **How** can we “improve”, and over what??

What is learning??

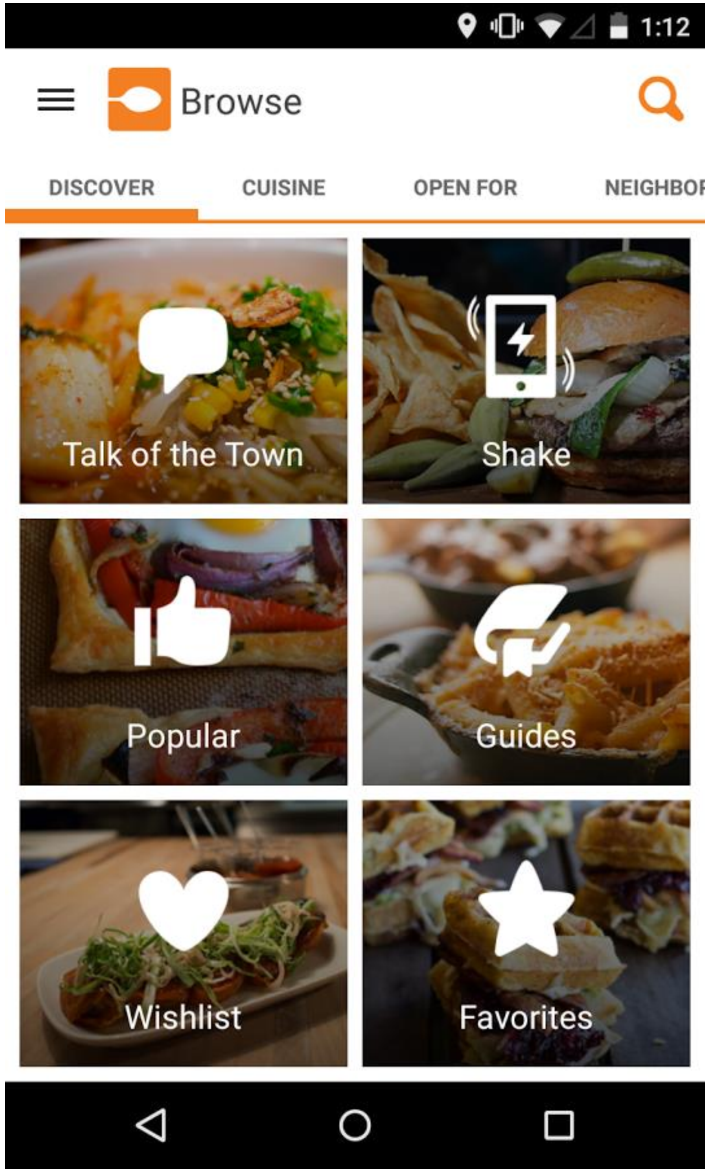
Fire burns!

But we learned using it



You can study (learn) Machine Learning

And then build an app to reccomend best restaurants based on people’s preferences



So, what is learning (for humans)?

- **Make sense** of a **subject**, **event** or **feeling** by interpreting it into our own words or actions

- **Use** our newly acquired ability or knowledge - in conjunction with skills and understanding we already possess - **to do something useful** with the new knowledge or skill

What is learning?

UNDERSTAND

+

GAIN KNOWLEDGE

+

USE NEW KNOWLEDGE TO DO SOMETHING

But, how do we learn??

How do humans learn?

- Someone tell us (teacher, or watching others)

- Try and test (learning by doing) as in the fire example

Is there something humans cannot learn??

As a matter of dacts, machines can learn to fly, swim, run..

- Surprisingly, with rather different strategies …

https://www.youtube.com/watch?v=4ZqdvYrZ3ro&feature=emb_imp_woyt

Besides things that we cannot learn, there are others that are either..

- Difficult to learn
- Difficult to teach

When is it difficult for humans to learn?

If there are **too many data**, humans cannot easily make sense of them (e.g. finding regularities in the human genome, learning to recognize one among millions of objects, market analysis and forecasts)

When is it difficult for humans to learn?

If data **change too frequently**, humans might be unable to continuously adapt their knowledge (e.g. personalized recommendations, market analysis forecast)

Stock market values
And quotes

When is it difficult
for humans to learn?

If the environment is dangerous, “learning by doing” cannot be applied (e.g. rescue systems)

When is it difficult for humans to teach?

If there is not enough information or previous expertise to “understand and gain knowledge”

(we actually**do not understand** the image and speech recognition process by humans – it is not “teachable”)

So when is it advisable to use Machine Learning?

ML is used when:

- No expertise
  - Human expertise does not exist (navigating on Mars), or there is a danger
  - Humans are unable to explain their expertise (speech/image recognition)
- Too many data, data change frequently:
  - A solution changes in time (market data for market forecast)
  - A solution needs to be adapted to particular cases (personalized systems for a recommendation, diagnosis, etc.)

So when is it advisable to use Machine Learning?

- Develop systems that are too difficult/expensive to construct manually because they require specific detailed skills or knowledge tuned to a specific task (**knowledge engineering bottleneck**).
  - ➢ Expert systems

- Develop systems that can automatically adapt and **customize** themselves to individual users.
  - ➢ Personalized news or mail filters
  - ➢ Personalized tutoring
  - ➢ Recommenders

- Discover new knowledge from large databases (**data mining**).
  - ➢ Customer preferences (learn from large samples of customers' shopping behaviours)
  - ➢ Medical text mining (electronic health records)
  - ➢ Social network mining (messages and friendship relations)
  - ➢ Emotion detection (from large datasets of people's images)

An interdisciplinary topic: many related disciplines!

- Artificial Intelligence
- Data Mining
- Probability and Statistics
- Information theory
- Numerical optimization
- Computational complexity theory
- Control theory (adaptive)
- Psychology (developmental, cognitive)
- Neurobiology
- Linguistics
- Philosophy

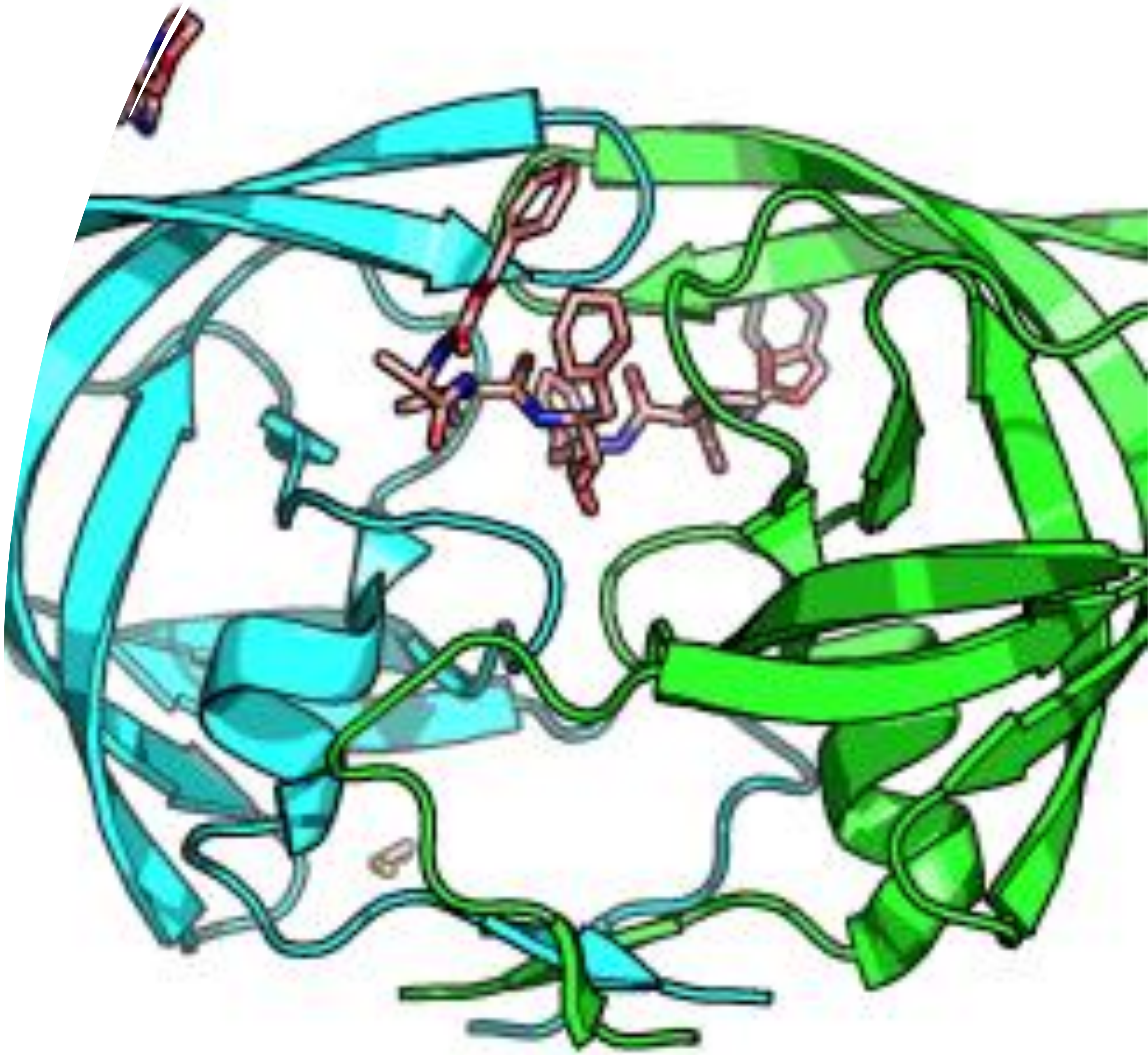
ML is perhaps the most interdisciplinar of CS areas!!

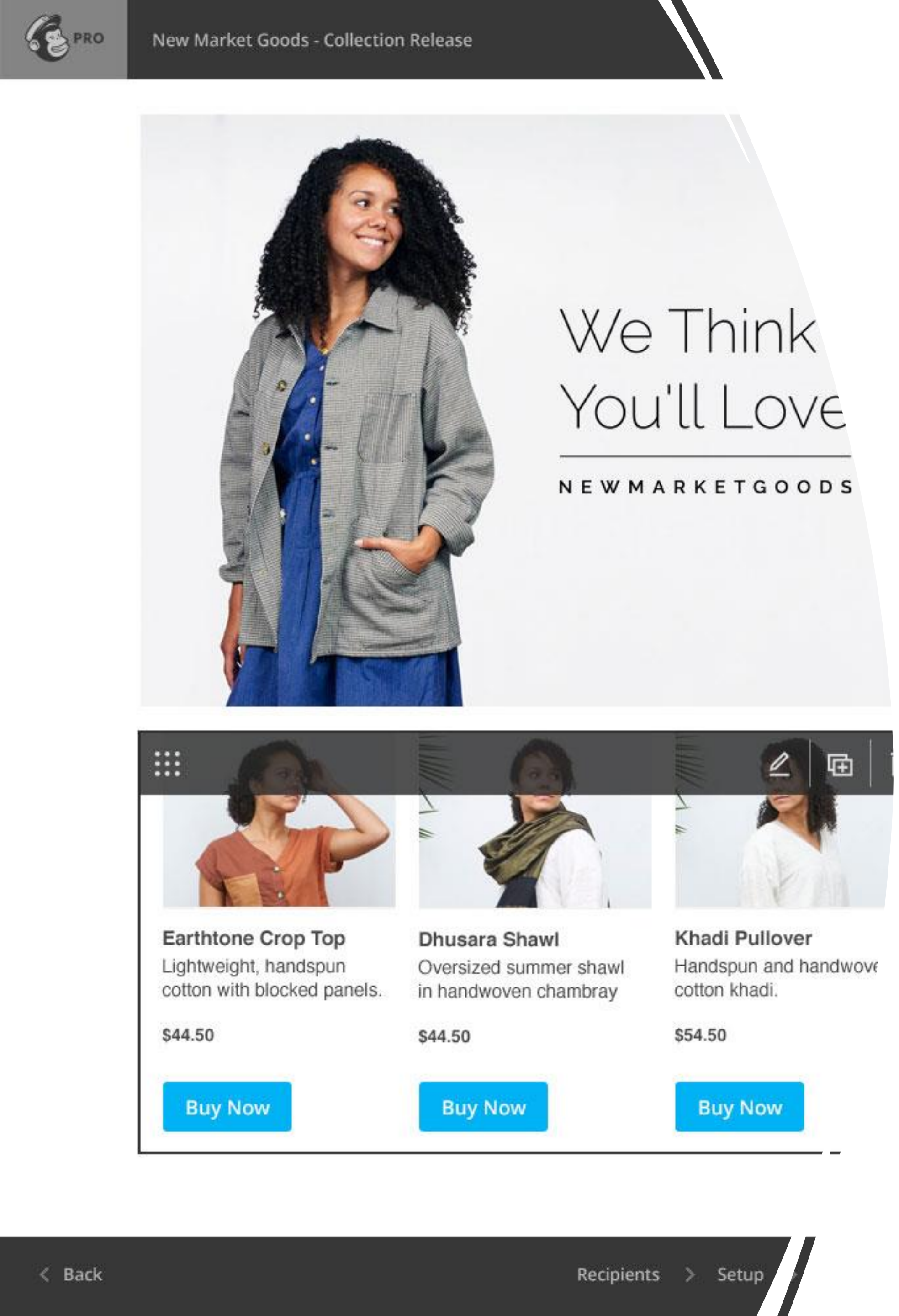
Some “real hot” ML applications

- It is really hard to find a problem where machine learning is not already applied -- machine learning is practically everywhere, in business applications and science!
- Here is a list of “hot” applications…

Computational Biology & E-health

- Predicting diseases and complications from the patient's health records
- Drug repurposing through the analysis of biological networks (e.g. interactions between proteins)
- Predicting epidemics through the analysis of human interaction data (e.g., population density, data on population movements, climatic data, etc.)



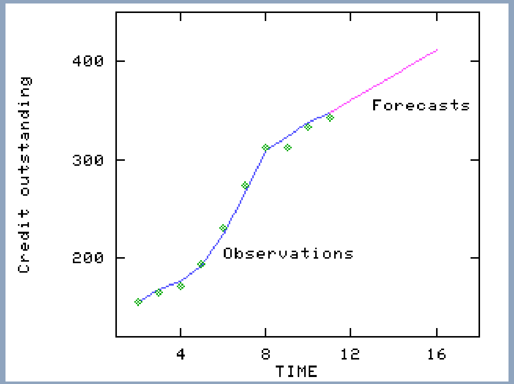


Web Search and Recommendation Engines

- Find relevant searches, predict which results are most relevant to us, return a ranked output (Google)

- Recommend similar products (e.g., Netflix, Amazon, etc.)

Finance



- Predict if an applicant is credit-worthy
- Detect credit card frauds
- Find promising trends on the stock market

Text and Speech Recognition

- Handwritten digit and letter recognition at the post office
- Voice assistants (Siri)
- Language translation services

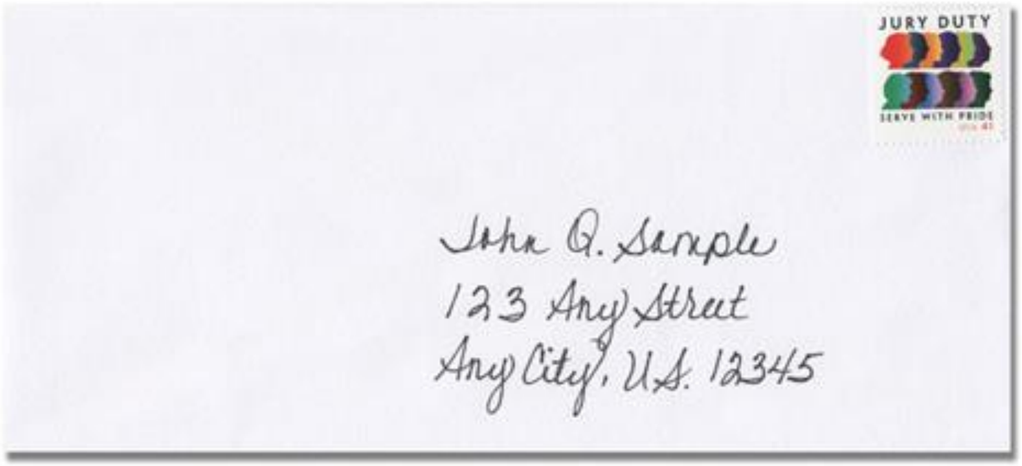
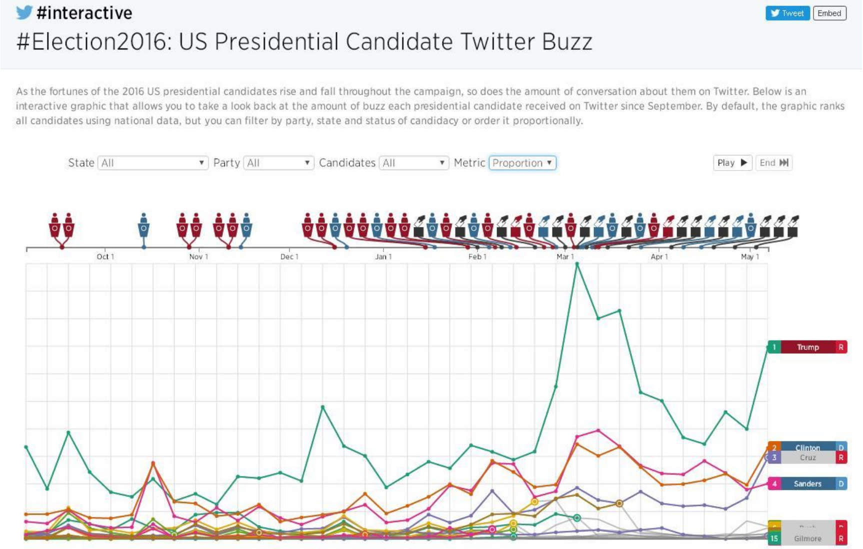


Image Understanding and Robotics

- Identification of relevant information (objects) in large amounts of Astronomy data
- Robotics for industry, energy saving, and smart cities
- Self-driving cars



Social Networks and Advertisement

- Social data mining:
  - data mining of personal information
  - Predict/analyze opinions, political choices, purchase behaviors

COURSE OBJECTIVES, ORGANIZATION AND SYLLABUS

OBJECTIVES

1. EXPLANATION OF THE MACHINE LEARNING WORKFLOW (steps required for a successful ML project, from data engineering, to selection of algorithms to evaluation)
2. COVERAGE OF MACHINE LEARNING ALGORITHMS (off-the-shelf and deep, supervised and unsupervised)
3. LABS & USAGE OF ML POPULAR PLATFORMS

Course material

https://twiki.di.uniroma1.it/twiki/view/ApprAuto/WebHome

- Slides (partly) from: link and many other sources
- Textbook: Tom Mitchell, Machine Learning, McGraw Hill, 1997 (new 2017 chapters on link)
- Introduction to machine learning ETHEM ALPAYDIN (online book)
- Deep learning (MIT press): link
- Course twiki: link

THERE IS PLENTY OF MATERIAL ON THE WEB, AND PLENTY OF DATASETS AND LIBRARIES

mathematical details are provided «to some extent» (coverage and practice are the focus here). For those who are interested, references are provided

Course labs

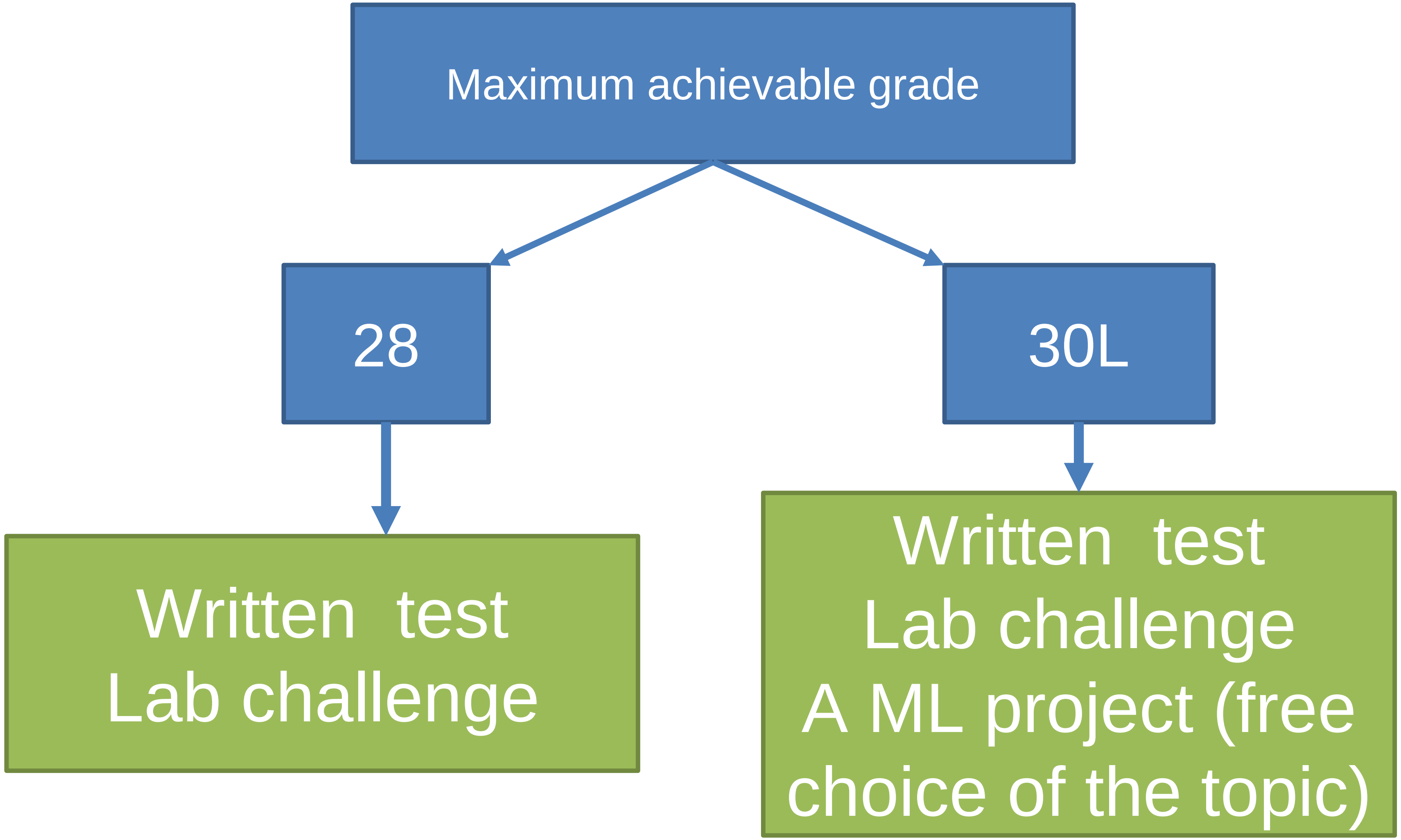
- Algorithms experimented on Keras and Scikit-Learn TensorFlow. Other libraries can be used

- The objective of labs is learning **practical ML building workflow**: data selection, data preparation and cleaning, choosing algorithms, hyper-parameter tuning, evaluation experiments.

Caveat: Coverage of ML topics is limited!

- This is a first-level “basic” ML course
- On the second semester there is an advanced course (more insight on deep learning, especially for image processing)
- ML algorithms for specific applications (NLP, security, etc.) are also taught in other courses

Exam and Project



Project

- Projects can be carried on by teams of **2** students (no extra grade if you are alone)

- How should the project be:
  - Not a trivial problem. Choose a **real-life** problem, or invent one
  - Use a sufficiently **large dataset** (many repositories are available). Best if merging different datasets
  - The dataset must need some feature engineering (some nontrivial pre-processing of data)
  - **More than one algorithm** tested, hyper-parameter tuning
  - **Evaluation** and analysis of results must show that you understand **why** you get a given result
  - I don’t care if you get very good performances (in complex problems results might not be so good) but rather that you **understand what is going on**

Deadlines and important issues

- There are two written tests on january-february (winter session) two on june-july (summer session) one on september
- I open ONE INFOSTUD call per SESSION (not per exam, per session!!!)
- Written test dates are published on the Department web site for all exams!

Deadlines and important issues

- You CAN’T deliver the project and the challenge when you want!!
- You will be given 1 DEADLINE for each session (winter, summer and september)
- Projects and challenges delivered after that date will not be considered and will shift to the subsequent session

How is the course organized

- Theoretical lessons + labs
- After every (or so) lesson, self-assessments are provided
- Self-assessments are useful to test your understanding of the subject. Very useful to pass the written test
- **PLEASE DO SUBSCRIBE TO GOOGLE GROUP** (use your Sapienza email and don't forgive to check it often, or redirect to your main email– don't miss my mails!)
- Google group is **also useful to discuss self-assessment solutions among students**!

  (peer evaluation)

Please be aware!

- Make sure you **read carefully** what is written of the course web site
- Make sure you **don’t miss may emails** on the Google group
- I will NOT answer email where you ask me things that I have explained already..
- Although this happens all the time!