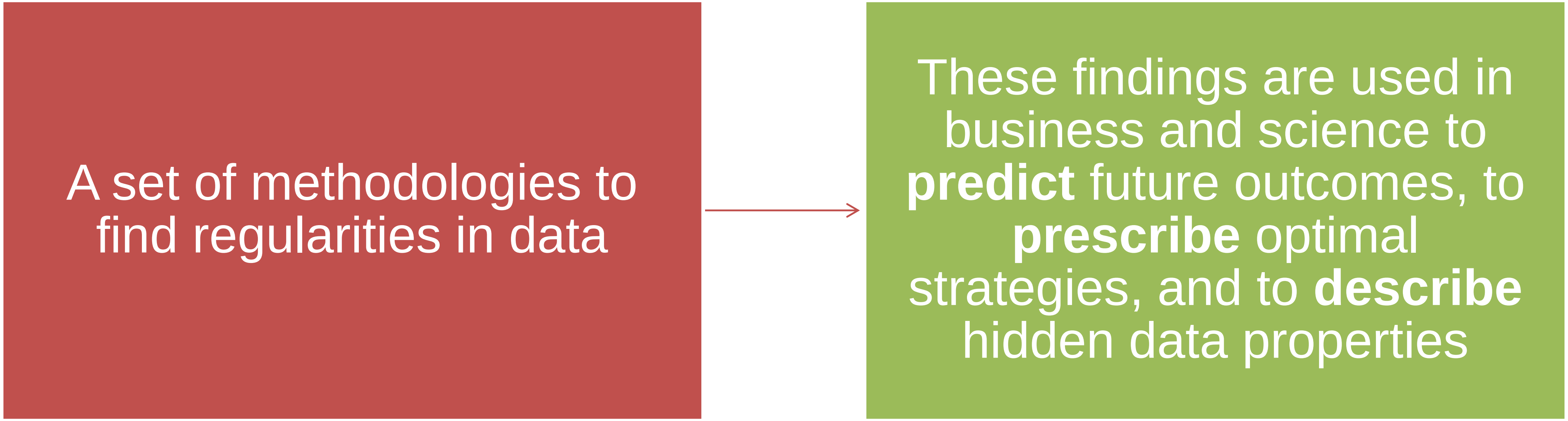
Machine Learning: *a gentle introduction & info about the course*

Paola Velardi

What is machine learning (in a nutshell)



A set of methodologies to find regularities in data

These findings are used in business and science to **predict** future outcomes, to **prescribe** optimal strategies, and to **describe** hidden data properties

What is the task?

- Traditional ML learning tasks are applied to virtually all business and human activities such as health, sport, finance, supply chain, resource optimization & repair, economy, management, marketing, planning, social analysis.. They can be classified into:

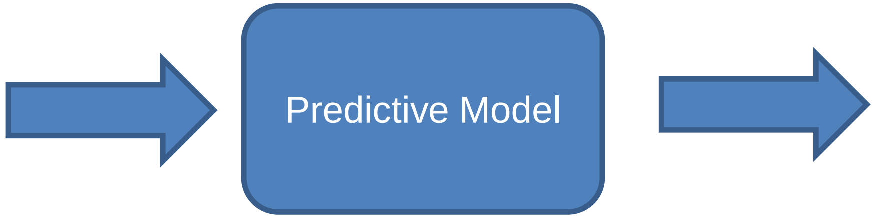
- Predictive:
  - Examples: predict patients’ risk of a complication (e.g., cardiovascular risk), predict future sales of a new product, predict
    users’ satisfaction in a market campaign, predict future value of a financial asset,
  - **ML Task:** Given available data, learn a model to predict future outcomes (e.g., study what happened to past patients, and learn to predict what may happen to new patients based on the gained knowledge)

- Prescriptive/descriptive:
  - Examples: (descriptive)
    customer segmentation according to their profiles (prescriptive)
    best strategy to win a game, (prescriptive) best way for a robot
    to execute a given task – e.g., drive a car – (prescriptive)
    recommending items to buy in e-commerce
  - **ML Task:** Given available data, or given an environment and some stimuli, learn to prescribe «how to», e.g., best actions to be performed

Prescriptive model example

Task: predicting the risk of cardiological complications

Input: A patient’s data:clinical exams, imaging, drugs data, electronic health records, genomic data..

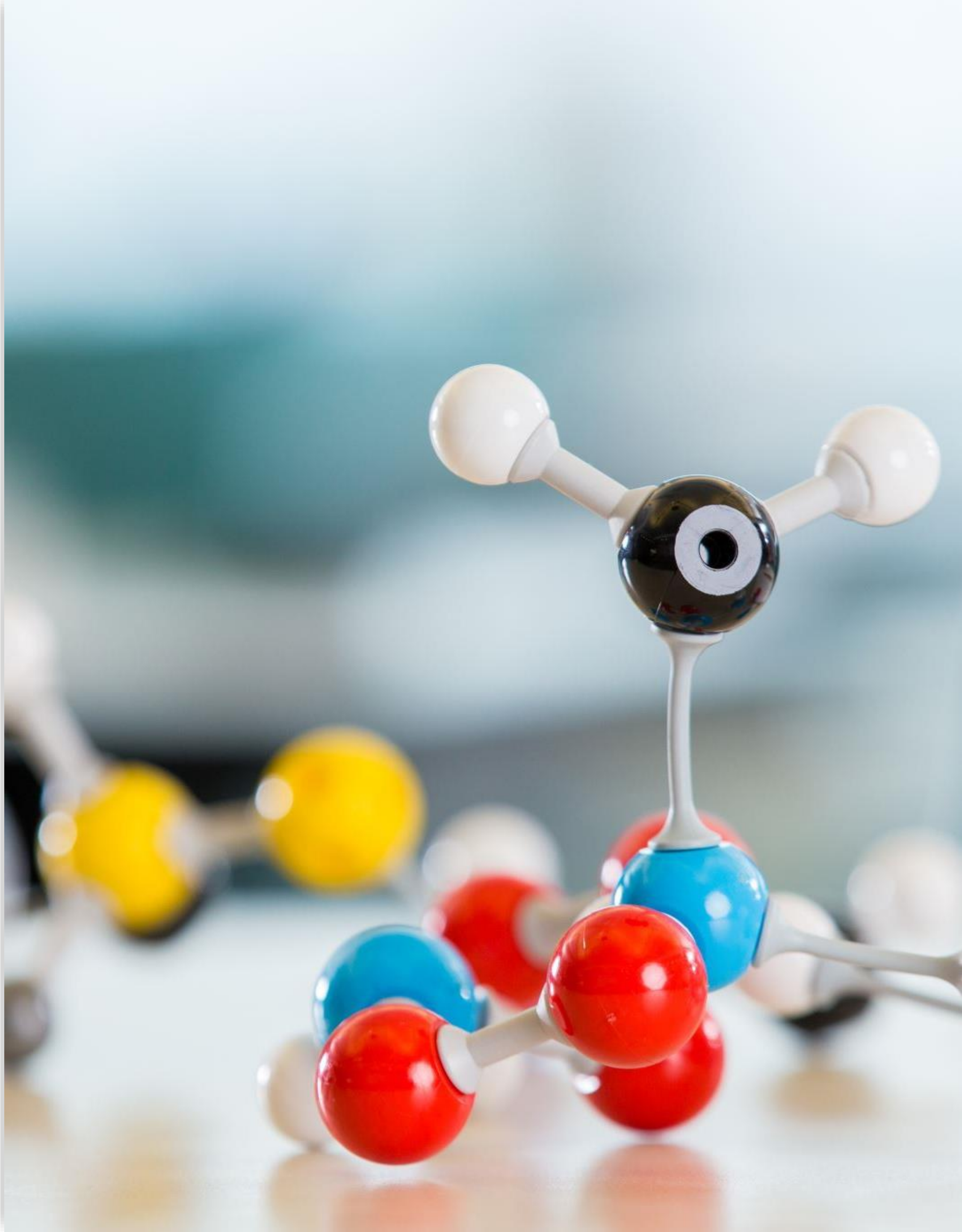


Predictive Model

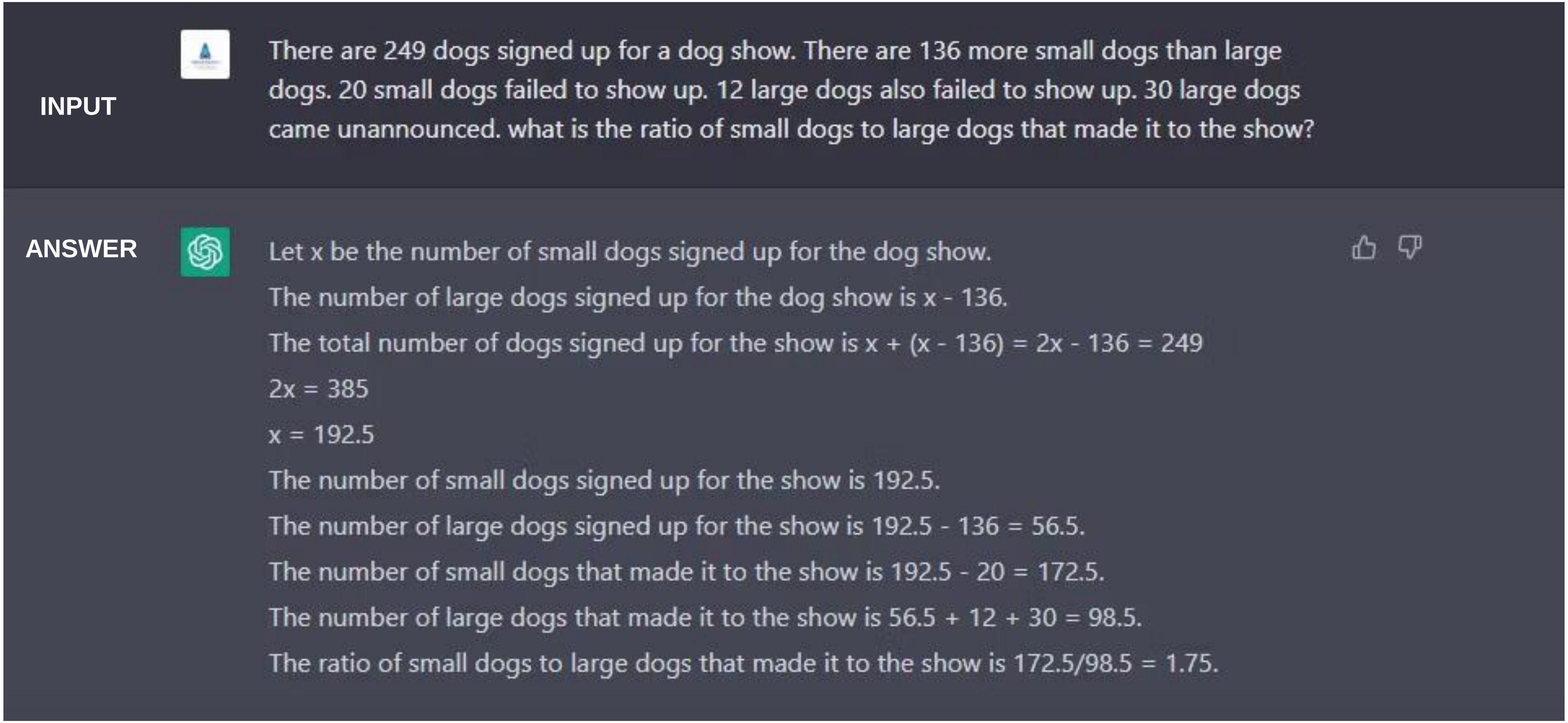
Answer: % risk of the input patient, given evidence provided by available data

Is this classification still appropriate?

- ML is now at the basis of Vision and NLP applications.
- Recent ML systems (such as GTP/OpenAI) perform a variety of language and vision-related tasks **that cannot be classified as prescriptive/descriprive/predictive**
- Rather, these systems learn **transforming**/translating one type of information/data into another type of information/data, examples:
  - Machine translation
  - Question answering
  - Summarization
  - Automated code generation
  - Image generation from NL text and viceversa
  - Visual charts from datasets
  - A growing number of complex applications were the task is a transformation from "some" input to "some" output
- In these systems input is not «data» but natural language
- Transformers will be introduced in the NLP course (2nd semester)



Example GTP: solving mathematical problems



INPUT

There are 249 dogs signed up for a dog show. There are 136 more small dogs than large dogs. 20 small dogs failed to show up. 12 large dogs also failed to show up. 30 large dogs came unannounced. what is the ratio of small dogs to large dogs that made it to the show?

ANSWER

Let x be the number of small dogs signed up for the dog show.

The number of large dogs signed up for the dog show is x - 136.

The total number of dogs signed up for the show is x + (x - 136) = 2x - 136 = 249

2x = 385

x = 192.5

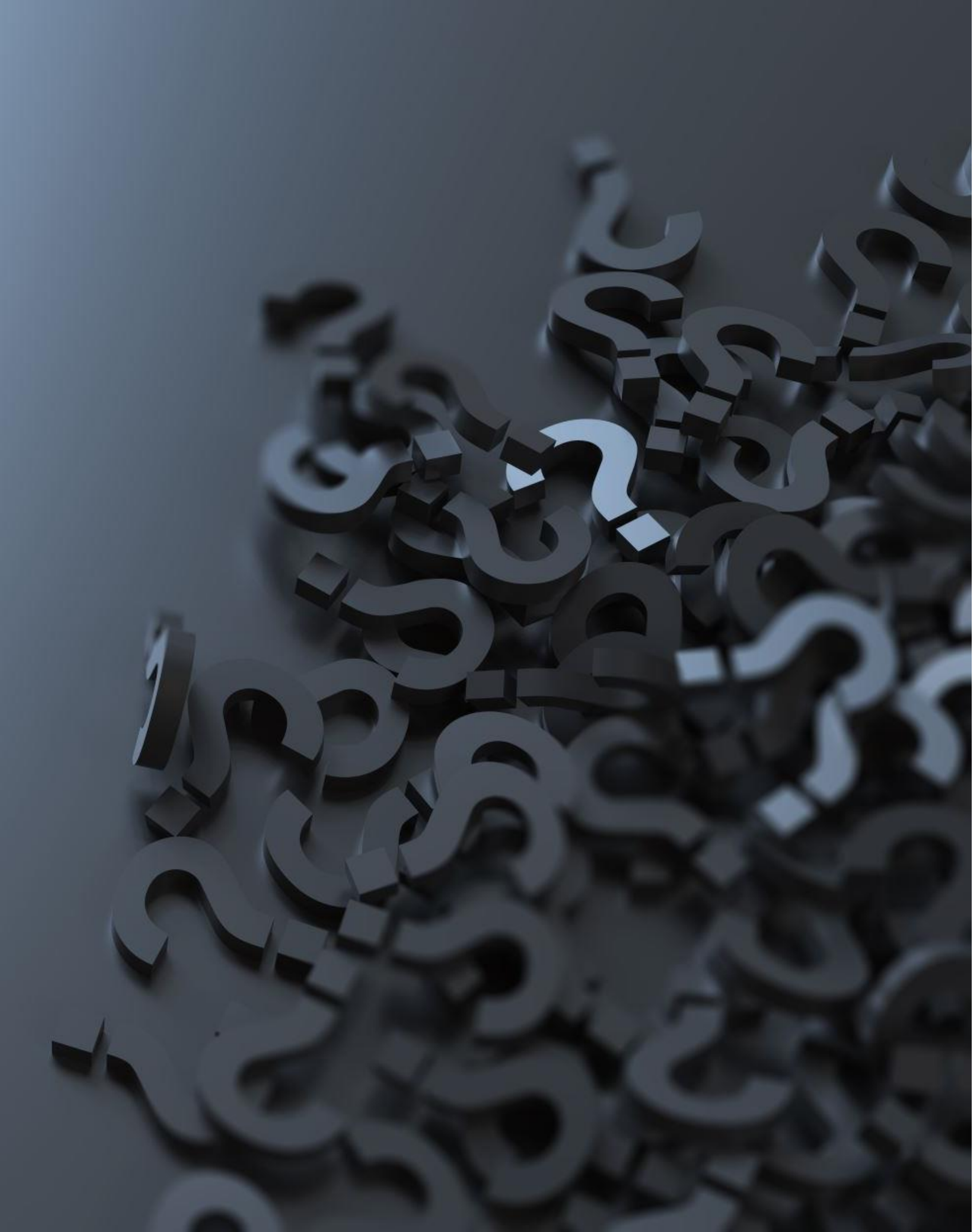
The number of small dogs signed up for the show is 192.5.

The number of large dogs signed up for the show is 192.5 - 136 = 56.5.

The number of small dogs that made it to the show is 192.5 - 20 = 172.5.

The number of large dogs that made it to the show is 56.5 + 12 + 30 = 98.5.

The ratio of small dogs to large dogs that made it to the show is 172.5/98.5 = 1.75.



What is the difference?

- "Business oriented" ML systems handle **structured data** (in the form of vectors or matrixes), possibly evolving in time – like sequences of stock market prices, both in input and output;
- Data can be also originated from unstructured sources (like text or images) but they are presented to the ML model in a structured form, possibly after complex human-driven transformations (e.g., from words to vectors)
- Transformers handle unstructured input (queries in natural language, images, ..) as they are, possibly along with instructions (prompts) in NL, and transform them into another unstructured output (answer, translation, chart, summary..)
- In this course, we consider business-oriented systems since they are at the basis of the more complex transformers

ML systems learn from data

- Both Transformers and traditional ML systems need to be TRAINED to learn performing a given task
- They are both trained using available, possibly very large, quantities of **historical data** related to the task to be performed
- NOTE: transformers like GTP and-the-like are PRE-TRAINED, just like traditional models. After pre-training, they can be adapted («fine-tuned») for specific tasks where they process NLP input and instructions.
- So: all ML systems learn from DATA

What data are used to learn?

- Historical («*labelled*») data: data collected in the past, for which the outcome is known (example: patient histories where we know if a cardiovascular event has occurred or not; bank customer’s histories for which we know is they have been defaulters or not; mathematical problems along with their solution..)
- *Unlabelled* data: data with no labels, for example the sequence of purchases of a user on an e-commerce web site, or sequence of actions on flight actuators by a human pilot of an airplain

What is «labelled»? Usually the task is learning to predict the value of some variable (e.g., cardiovascular risk). Historical data provide examples of such values.

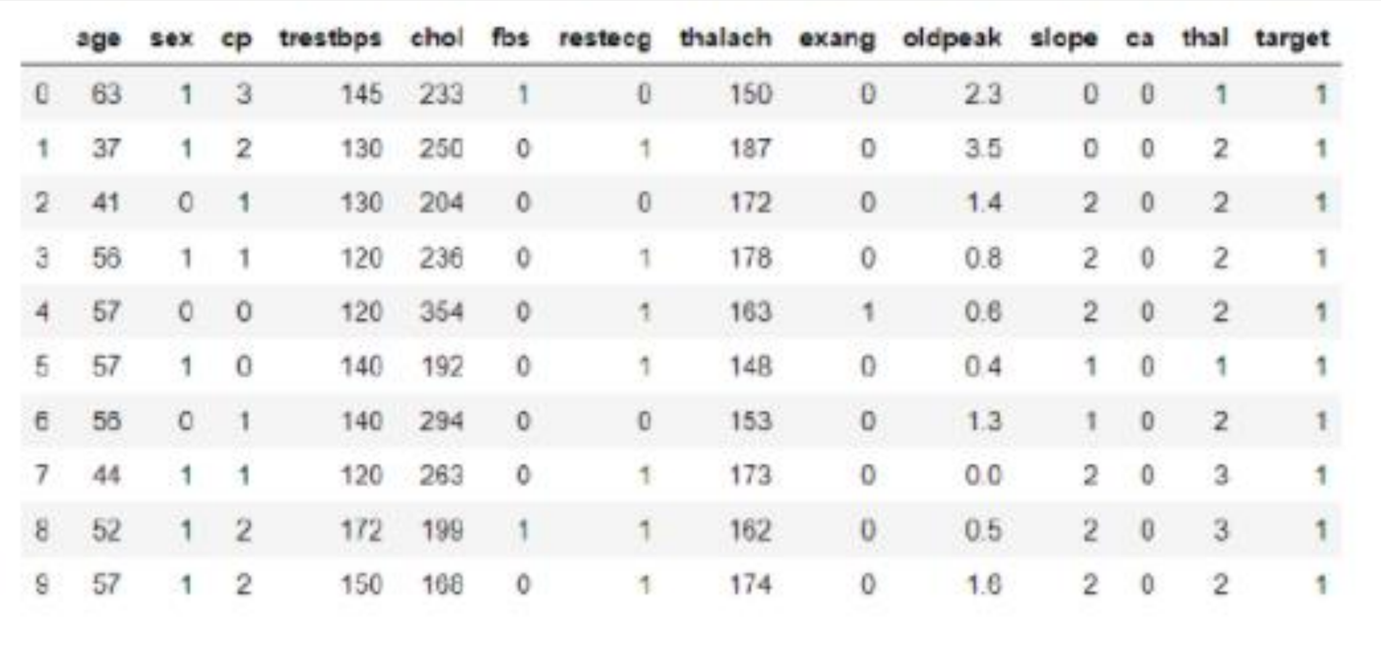
Example of training data in predictive ML systems: Credit risk assessment

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**

***** Here data are «**labelled**», to mean that historical data include the label (value) of the **variable we want to predict**, «Default» in this example. Note that Default is a binary variable, but as we will see, we can predict either discrete or continuos variables.

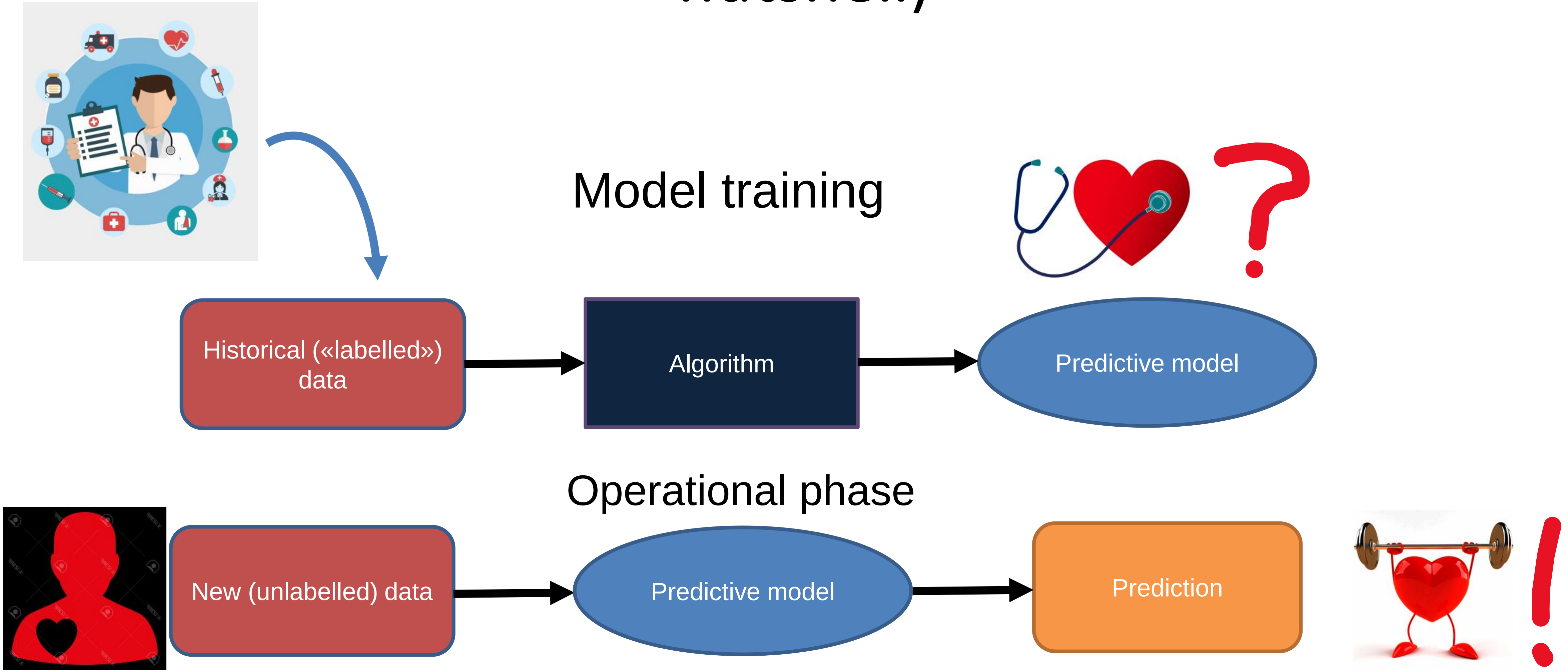
Example 2: cardiovascular risk assessment



	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 data of previous patients

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



Model training

Historical («labelled») data

Algorithm

Predictive model

Operational phase

New (unlabelled) data

Predictive model

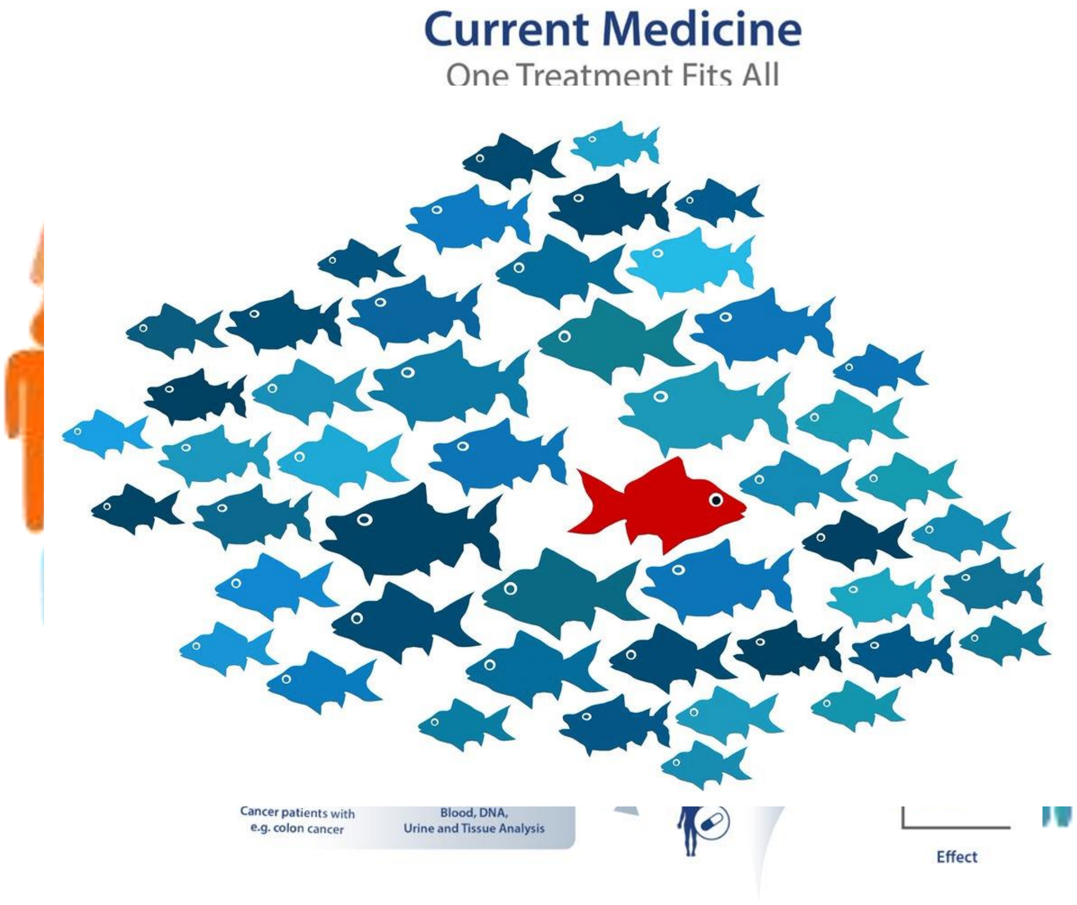
Prediction

Note, **not all ML systems work in this way**. This is the most popular category of ML systems, named **Supervised Machine Learning**.

What is the task?

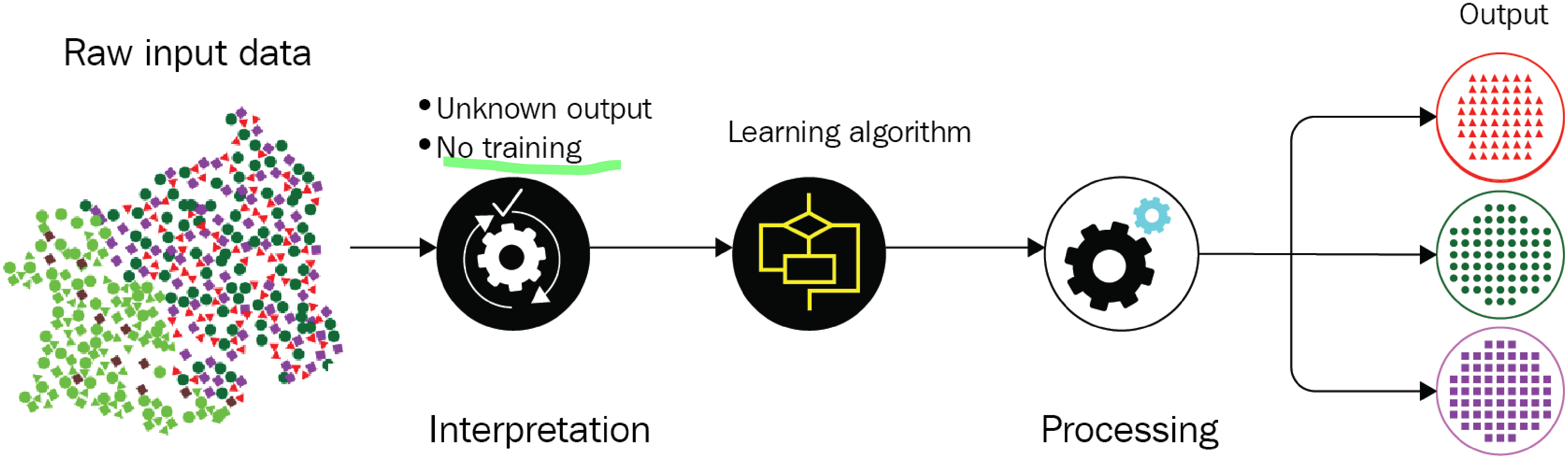
- Predictive:
  - Given previous historical «labelled» data, learn a model to predict future outcomes (e.g., see what happened to past credit applicants, or to past patients, and learn what may happen to new applicants or new patients)
  - Examples: predict patients’ risk of a complication, predict future sales of a new product, users’ satisfaction in a market campaign..
- **Prescriptive/Descriptive:**
  - **Given available 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, best way for a robot to execute a given task – e.g., drive a car – how to improve on-line sales by recommending the right items to customers**

Example of descriptive analytics: customer segmentation, precision therapies, anomaly detection


Current Medicine
One Treatment Fits All
Cancer patients with e.g. colon cancer
Blood, DNA, Urine and Tissue Analysis
Effect

- 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

Workflow of descriptive analytics



Output

Raw input data

- Unknown output
- No training

Learning algorithm

Interpretation

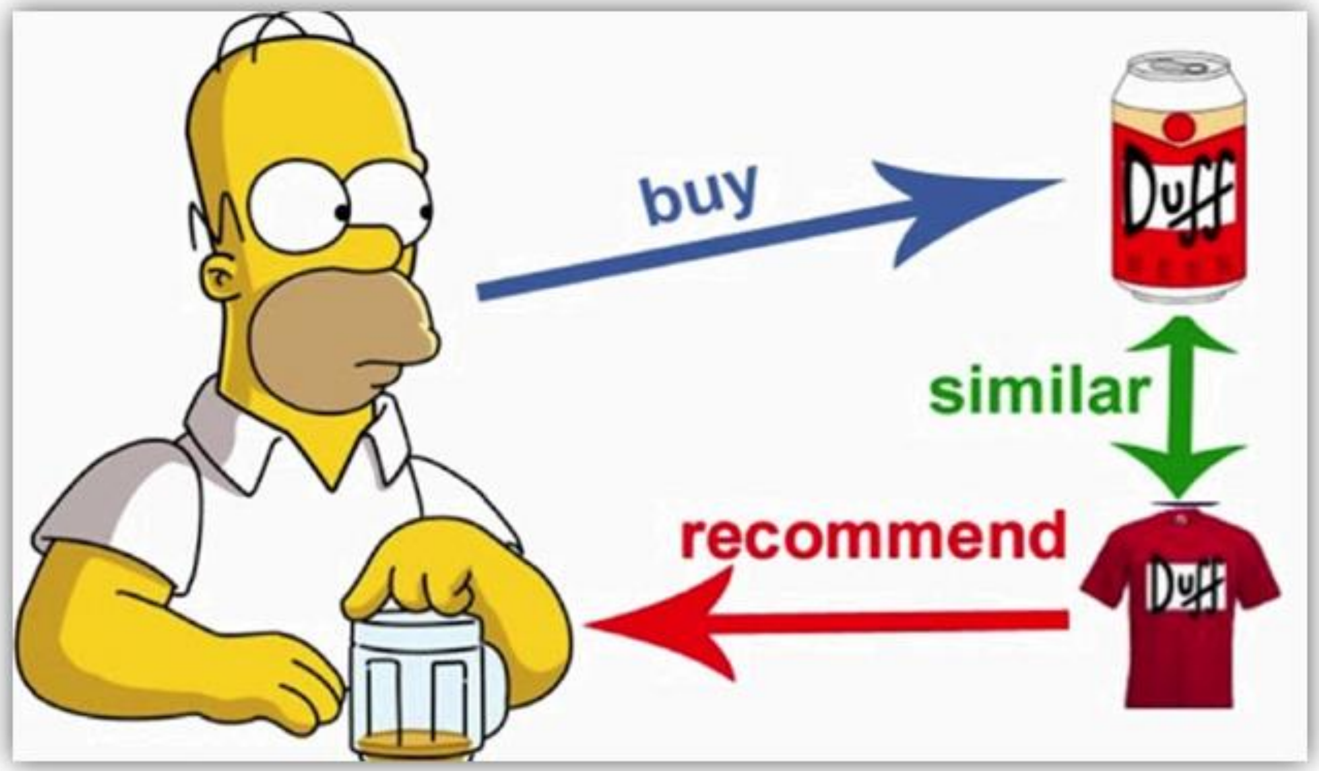
Processing

Note: system learns from unlabelled data, these ML models are called **Unsupervised Machine Learning** models

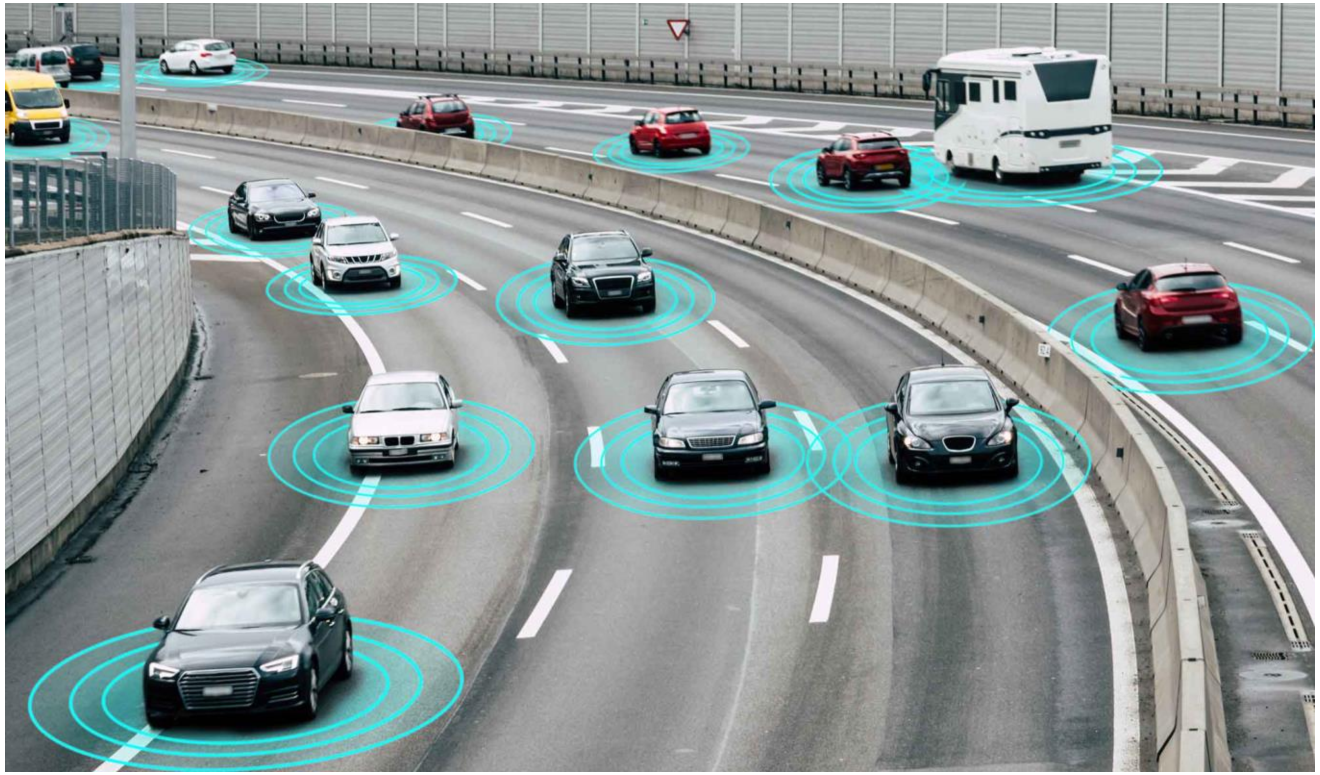
Example of prescriptive analytics

Recommender systems

- Observe a behavior and "recommend" items to buy, music to listen, people to follow on social,..

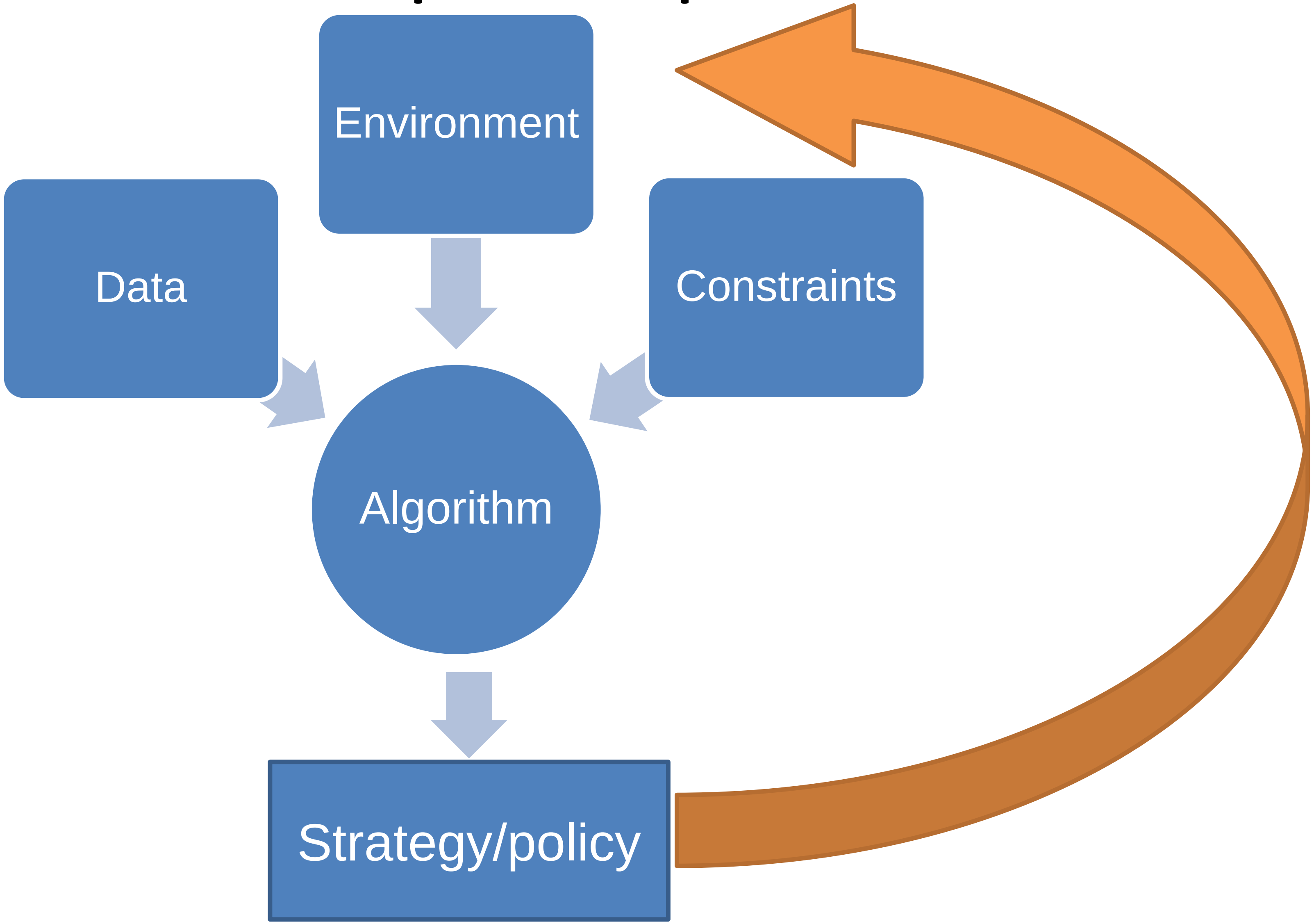


Example of prescriptive analytics: self driving cars



- Analyse driving behaviours of million «human» drivers
- Learn best strategy to react to the environment (driving strategies) in any condition

Workflow of prescriptive ML



Environment

Data

Constraints

Algorithm

Strategy/policy

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 select good data and make it «ingestable» by ML algorithms
- Whatever it takes, it is worth: without the fuel of «good» data, algorithms **just don’t work**
- **We will dedicate several lessons to the «data» issue: which types, how to find, represent and filter good data for a given ML task**
- **Keep in mind:** also Transformers need pre-training from raw data. Next, they can be adapted to specific tasks in which they accept NLP instructions /questions

Issues in Machine Learning

Issues in Machine Learning

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

Need to ponder on how human beings learn..

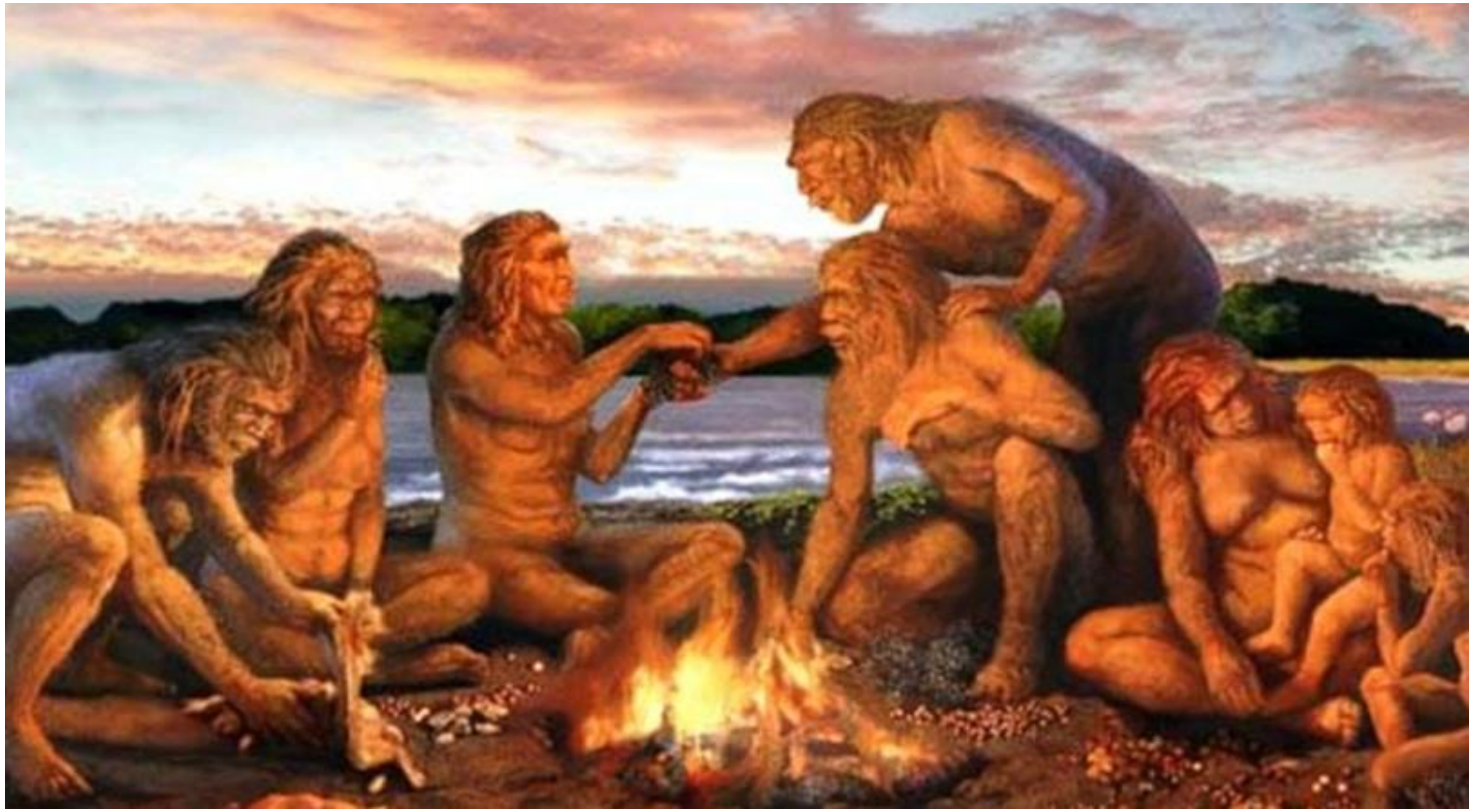
- **What** is learning?
- **What** can we learn?
- **What** is “experience”??

- **How** do we learn?
- **How** can we “improve”, and over what??

What is learning??

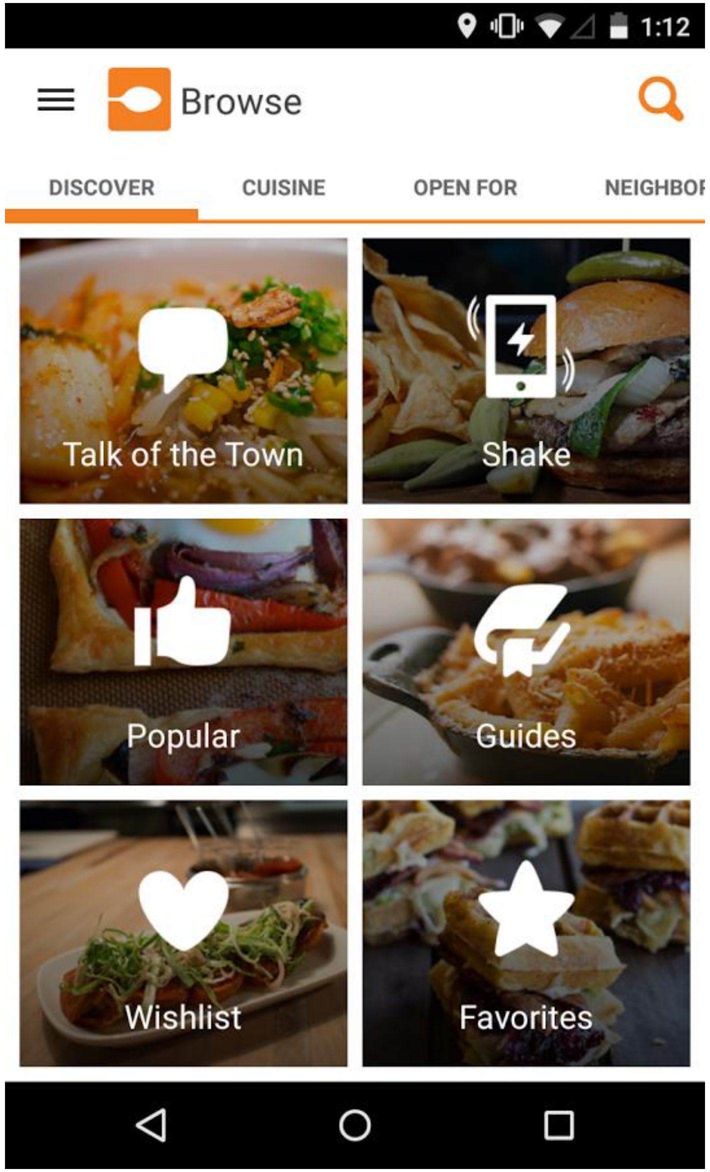
Fire burns!

But we eventually 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?

COLLECT AVAILABLE DATA (*ingest*)

+

GAIN KNOWLEDGE (*understand*, interpret data and transform it into knowledge)

+

USE NEW KNOWLEDGE

TO DO SOMETHING (*act*)

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

Basically, ML systems learn in one of these two ways

There is only one thing more painful than learning from experience, and that is not learning from experience.
Laurence J. Peter

Besides things that humans cannot learn (but possibly machines can..), 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)

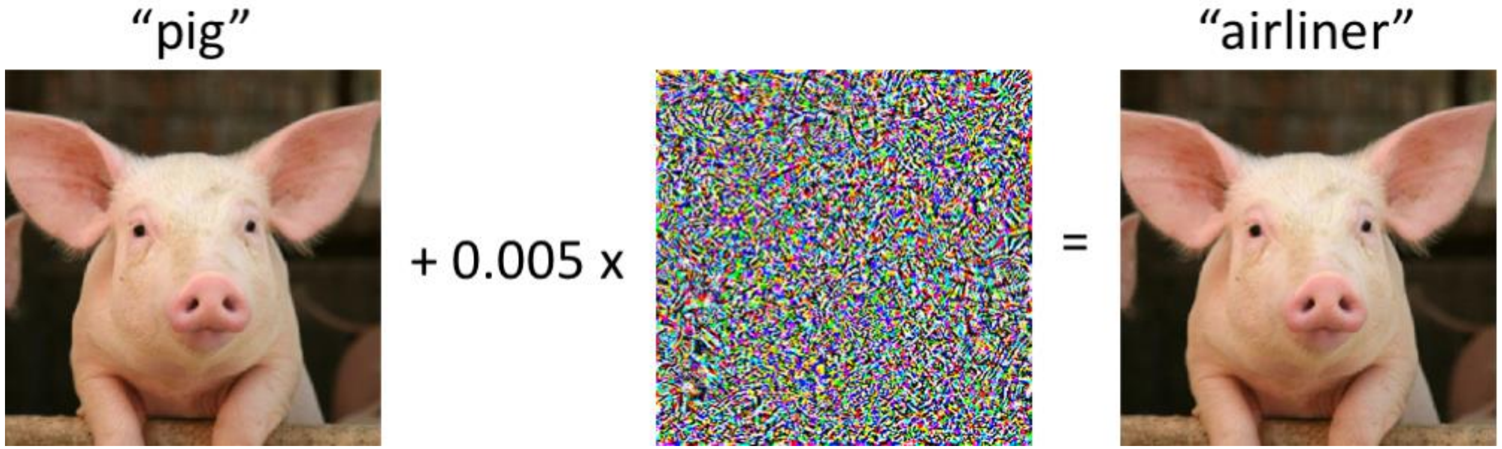
When is it difficult for humans to teach?

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”)

Are machines better than humans?

- Yes in some applications and «to some extent» (e.g., games, precision surgery, image understanding, recently NLP is «booming» as an application)



https://gradientscience.org/intro_adversarial/

- Major limitations to date: black-box (lack of explainability) issue; computing power and data availability (only very few big players dominate the field)

ML is 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

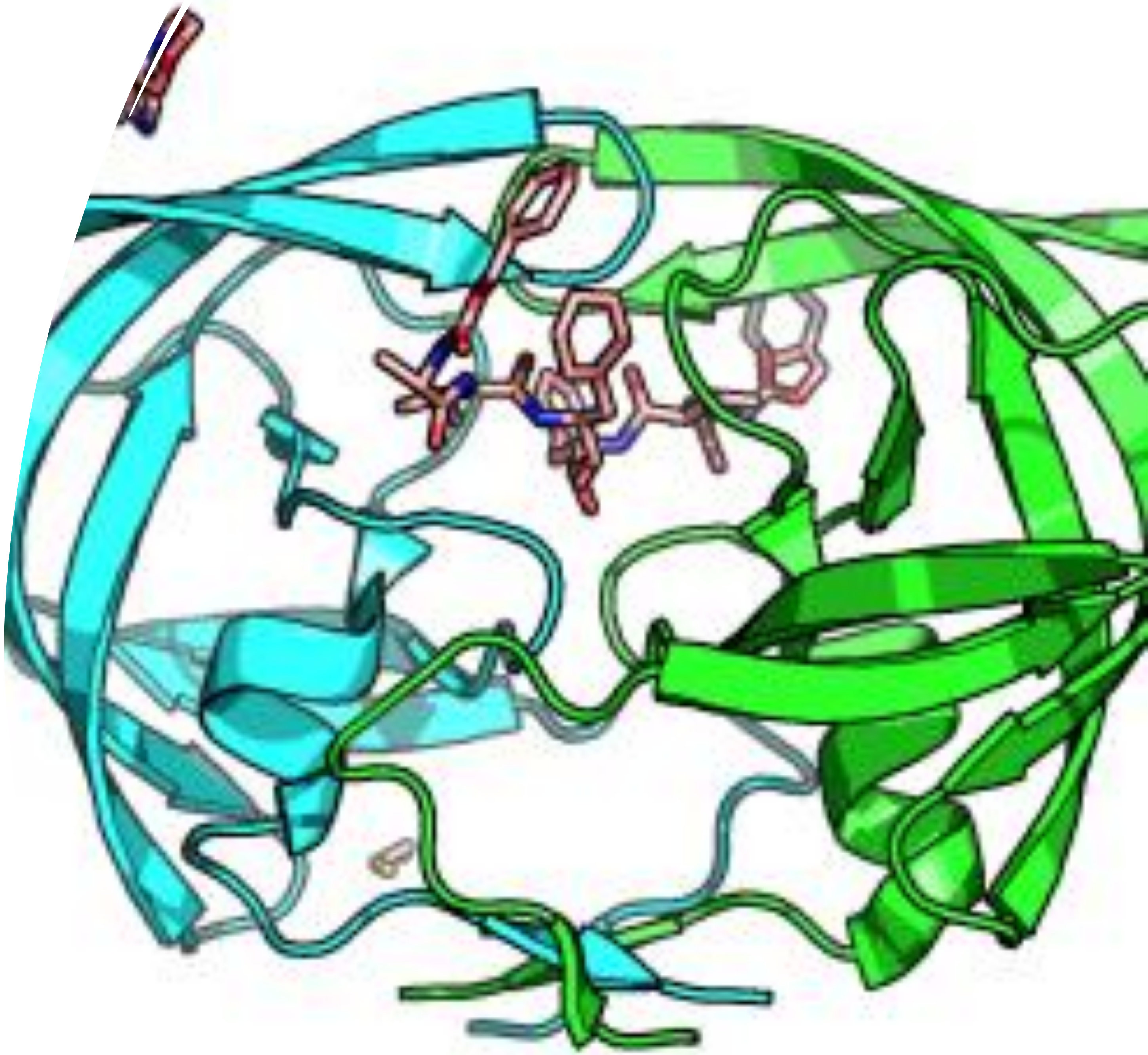
- It is really hard to find a problem where machine learning is not already applied -- machine learning is practically everywhere, in business applications and science!

- Let’s see a list of (truly) “hot” applications…

INTELLIGENT
INTELLIGENT
Natural Enzymes Children Toothpaste
INTELLIGENT
Natural Enzymes Children Toothpaste

Computational Biology & E-health

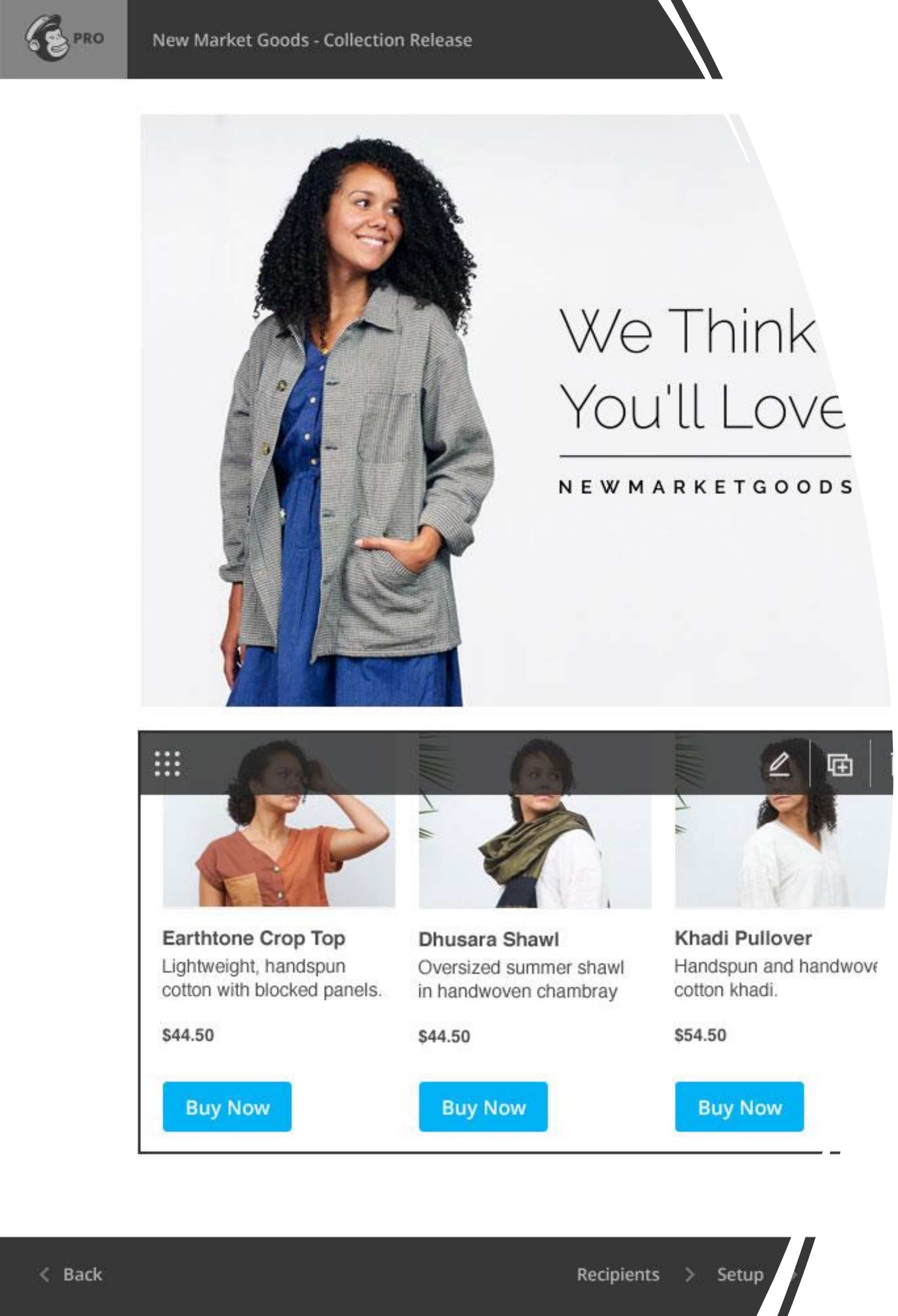
- Predicting diseases and complications from genomic data (metabolic, gene-disease relations, ..)
- Drug repurposing through the analysis of biological networks (e.g. interactions between proteins)
- Predicting epidemics through the analysis of human interaction data (e.g., population density, data on 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.)



New Market Goods - Collection Release

We Think You'll Love

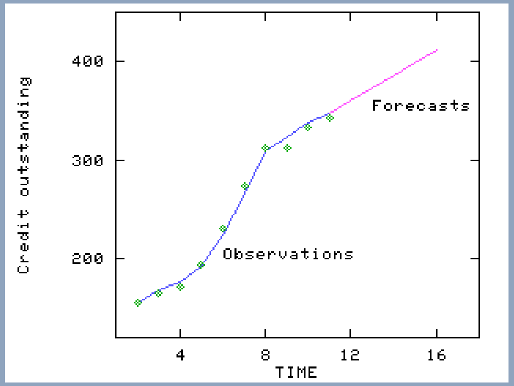
NEWMARKETGOODS

Earthtone Crop Top
Lightweight, handspun cotton with blocked panels.
$44.50

Dhusara Shawl
Oversized summer shawl in handwoven chambray
$44.50

Khadi Pullover
Handspun and handwove
cotton khadi.
$54.50

Finance



- Predict if an applicant is credit-worthy
- Detect credit card frauds
- Find promising trends on the stock market (*algorithmic trading*)

Text and Speech Recognition

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

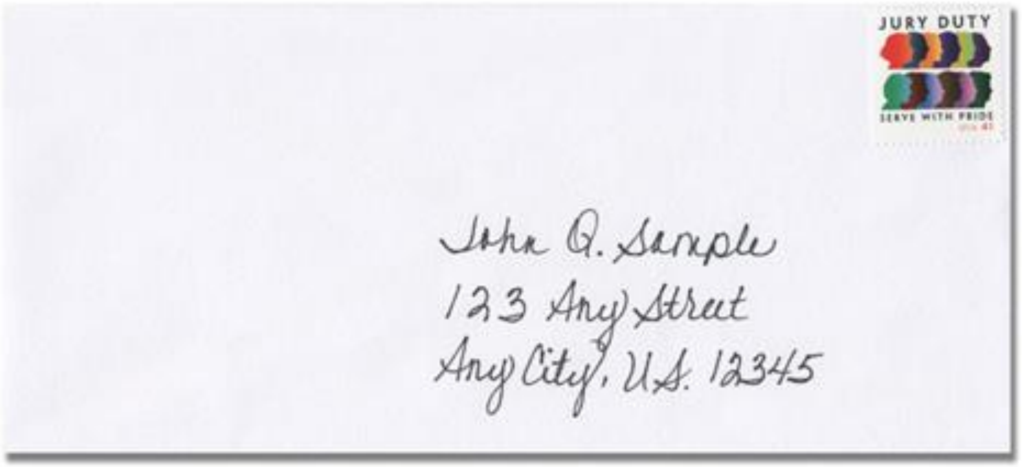
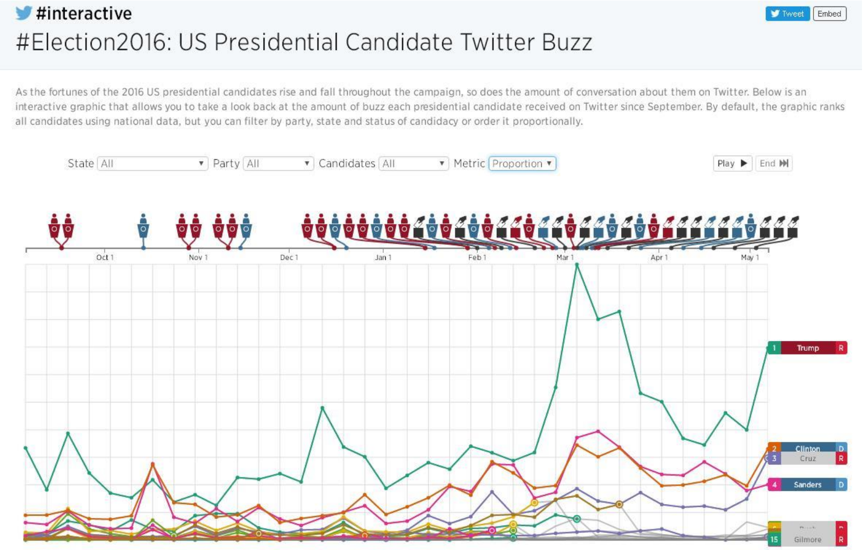


Image Understanding and Robotics

- 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



#interactive

#Election2016: US Presidential Candidate Twitter Buzz

As the fortunes of the 2016 US presidential candidates rise and fall throughout the campaign, so does the amount of conversation about them on Twitter. Below is an interactive graphic that allows you to take a look back at the amount of buzz each presidential candidate received on Twitter since September. By default, the graphic ranks all candidates using national data, but you can filter by party, state and status of candidacy or order it proportionally.

State All ▾ Party All ▾ Candidates All ▾ Metric Proportion ▾

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

And the latest

- Generating code from NLP instructions (e.g., NL2SQL)
- Domain-general question answering
- Visual recommender systems: given data, recommend best charts to gain insight from data (DATA2Viz)
- Generating artistic images from NLP descriptions (NL2Images)
- Automated image captioning (IMAGE2NL)
- Summarization
- Automated problem solving
- ….exponential growth of applications

COURSE OBJECTIVES, ORGANIZATION AND SYLLABUS

OBJECTIVES

1. IN BREADTH COVERAGE OF MACHINE LEARNING ALGORITHMS («some relevant» off-the-shelf algorithms, deep architectures, supervised and unsupervised)
2. MACHINE LEARNING WORKFLOW (steps required for a successful ML project, from **data engineering**, to selection and tuning of algorithms, up to performance evaluation and visual reporting)
3. PUTTING IT ALL TOGETHER IN LABS & USAGE OF ML POPULAR PLATFORMS
4. Focus in more on learning the entire business intelligence pipeline, in depth study of many algorithms in other advanced ML courses

Syllabus (23-24)

- What is ML, types of ML algorithms, ML systems workflow
- Simple ML algorithms: decision trees, perceptron
- Neural networks and the backpropagation algorithm
- Convolutional Neural Networks, Denoising Autoencoders, Sequential machine learning (RNN, LSTM)
- Ensambles: bagging, boosting, gradient boosting
- Basics of probabilistic learning
- Clustering
- Evaluating ML systems
- Data sources identification, data preparation (structured, unstructured, symbolic, numeric, sequential) and data pre-processing (missing data, unbalanced data, heterogeneous data..)
- Explainability of ML systems
- ML Labs: Put it all together on real data (e-health, finance): data preprocessing, model fitting, evaluation

Course material

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

- Slides (partly) from: link and many other sources
- Textbook: No textbook! Plenty of on-line resources, tutorials, papers..
- Deep learning (MIT press): link
- Course twiki: link

Course labs

- Algorithms experimented on different libraries (keras, tensorFlow, sckitlearn..)

- The objective of labs is learning **practical ML building workflow**: data selection, data preparation and cleaning, choosing algorithms, hyper-parameter tuning, evaluation experiments, visualization.
- This year lab focus: **MEDICAL ML, FINANCE**

Caveat: Coverage of ML topics is limited!

- This is a first-level “basic” ML course
- On the second semester and first semester, second year there are other advanced courses (more insight on deep learning, especially for image processing)
- ML algorithms for specific applications (NLP, security, etc.) are also taught in other courses
- Caveat: Impossible to avoid overlappings (students with different background, and different master programs – cybersecurity, data science –
- **Please read programs carefully, and CHOOSE ML courses in a way that best fits your background and interests**

Exam rules

- Written test (60%) + challenge (40%)
- Self assessment are provided at the end of every introduced topic. SA are very useful to pass the test
- Challenge: a project on real data implying the entire ML pipeline (data cleaning, preprocessing, model selection, model fitting, evaluation, reporting & visualization)
- Challenge topic will be presented towards the end of November
- If numbers allow, a mid-term exam in november (3° week)

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**!!!) BEFORE the session opens
- Written test dates are published on the Department web site for all exams!

Deadlines and important issues

- You CAN’T deliver the challenge when you want!!
- You will be given 1 DEADLINE for each session (winter, summer and september)
- 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, on which I sent previous email to the g-group, or written on the course web page**
- Although this happens all the time!