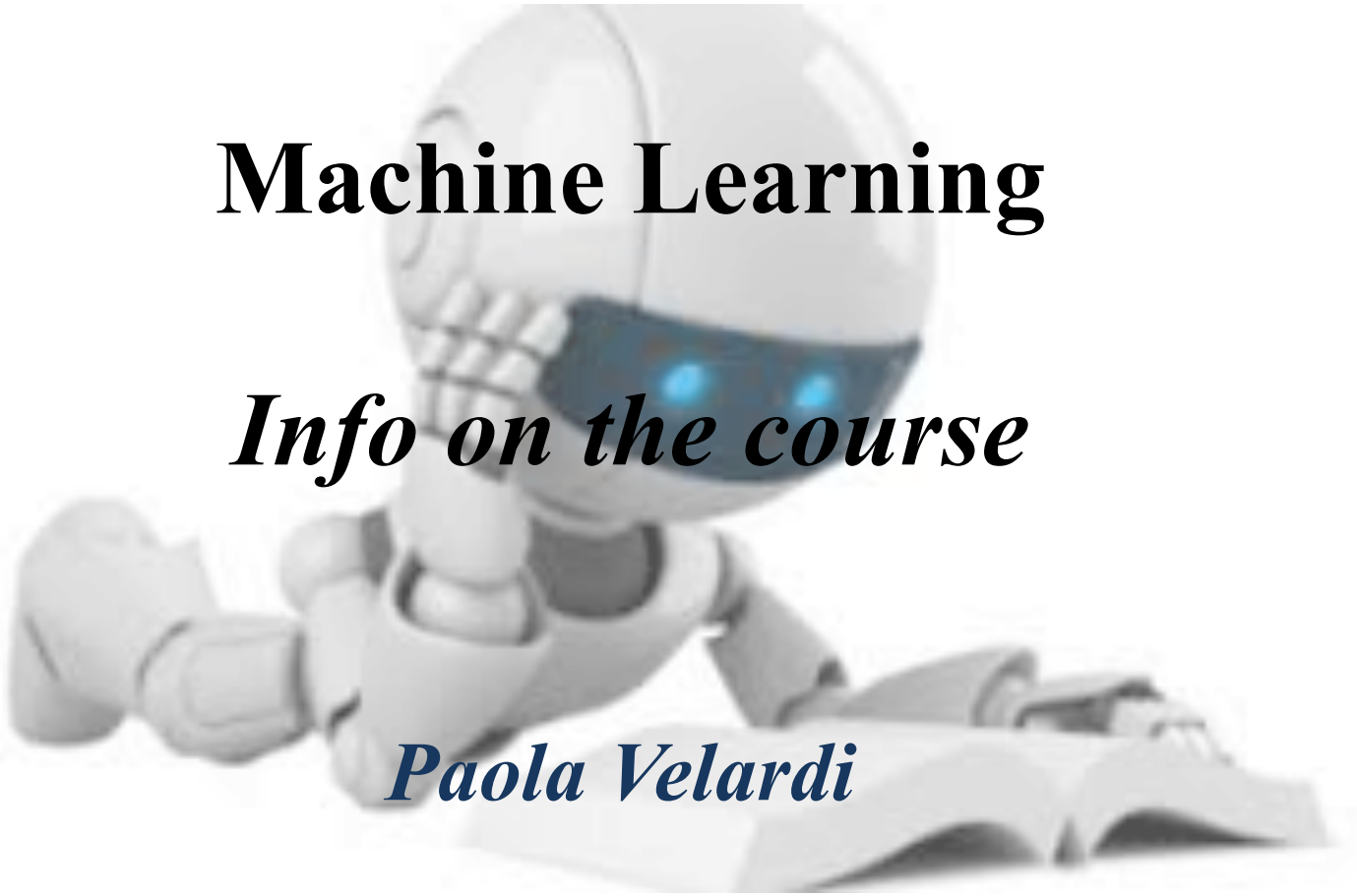
Machine Learning

Info on the course



Paola Velardi

What is machine learning (in a nutshell)

- A set of methodologies to find regularities in data
- These findings are use to predict future outcomes and/or to classify unseen data

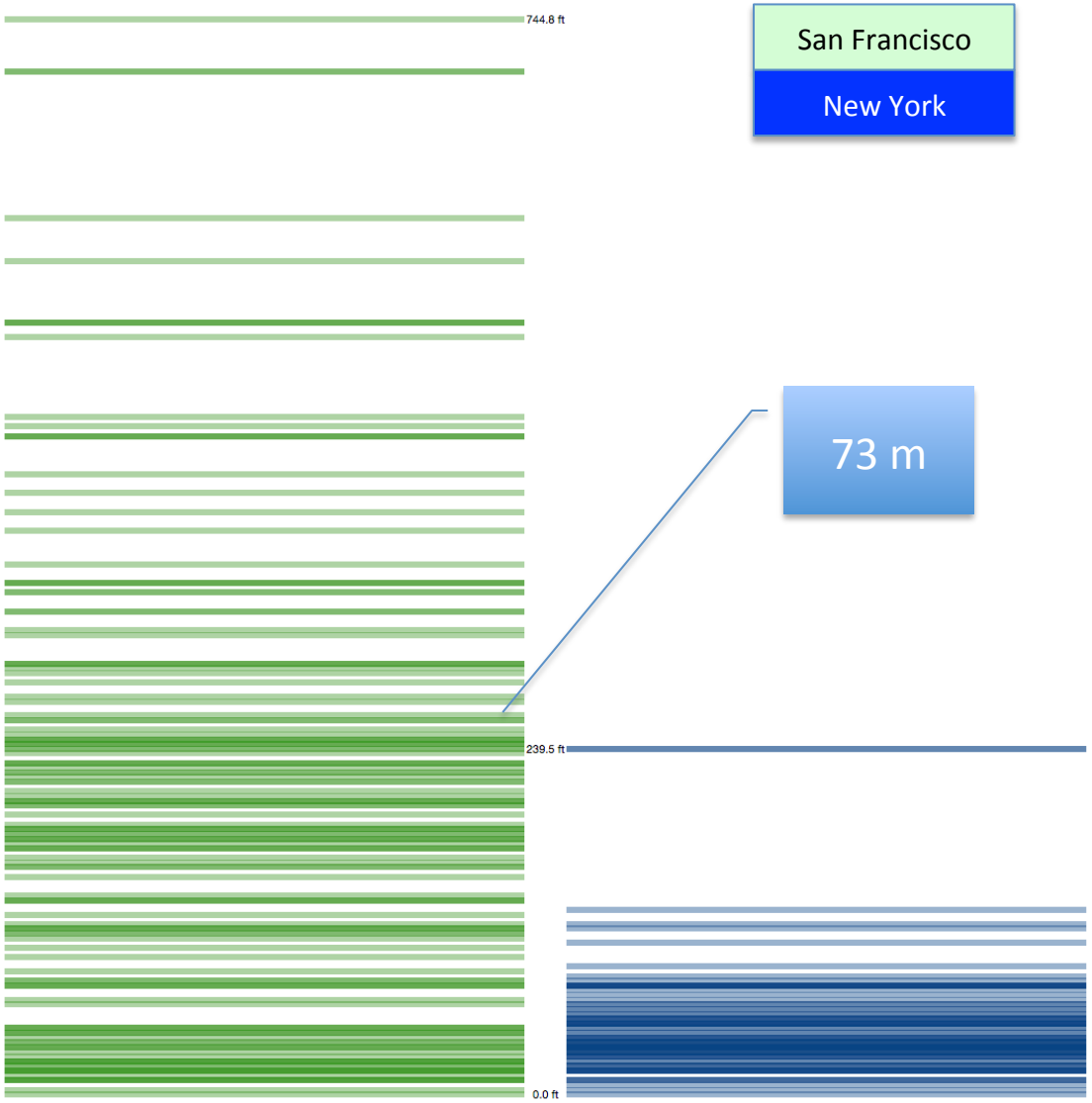
A (VERY) simple example

- Using a data set about homes, we wish to create a machine learning model to distinguish homes in New York from homes in San Francisco.
- Let's say you had to determine whether a home is in **San Francisco** or in **New York**. In machine learning terms, categorizing data points is a **classification** task.

To begin, we consider building elevation

- Since San Francisco is relatively hilly, the elevation of a home may be a good way to distinguish the two cities.
- Based on the home-elevation data to the right, you could argue that a home above 73 meters should be **classified** as one in San Francisco.
- So we can infer the following rule:

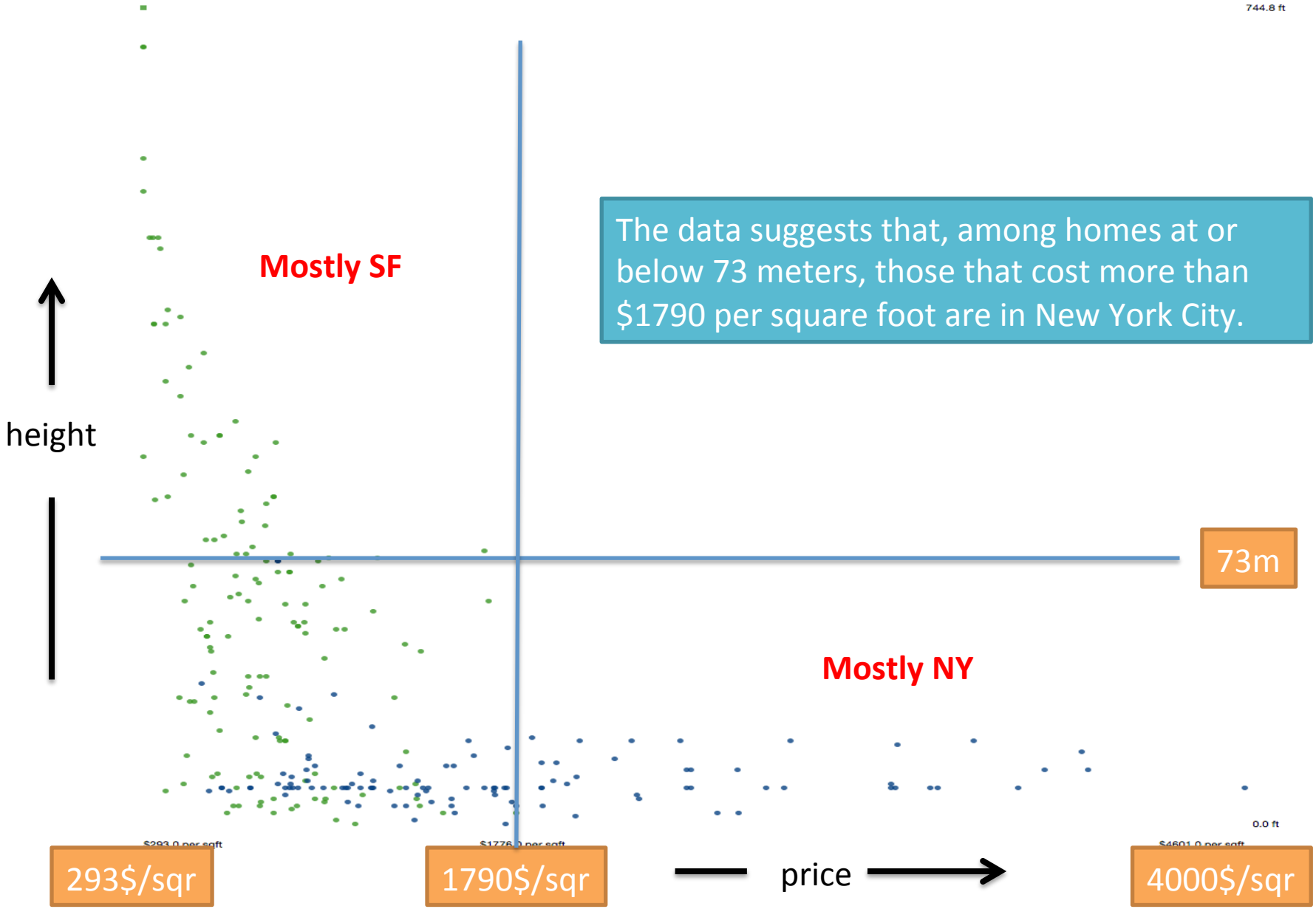
IF elevation>73m THEN city=SF

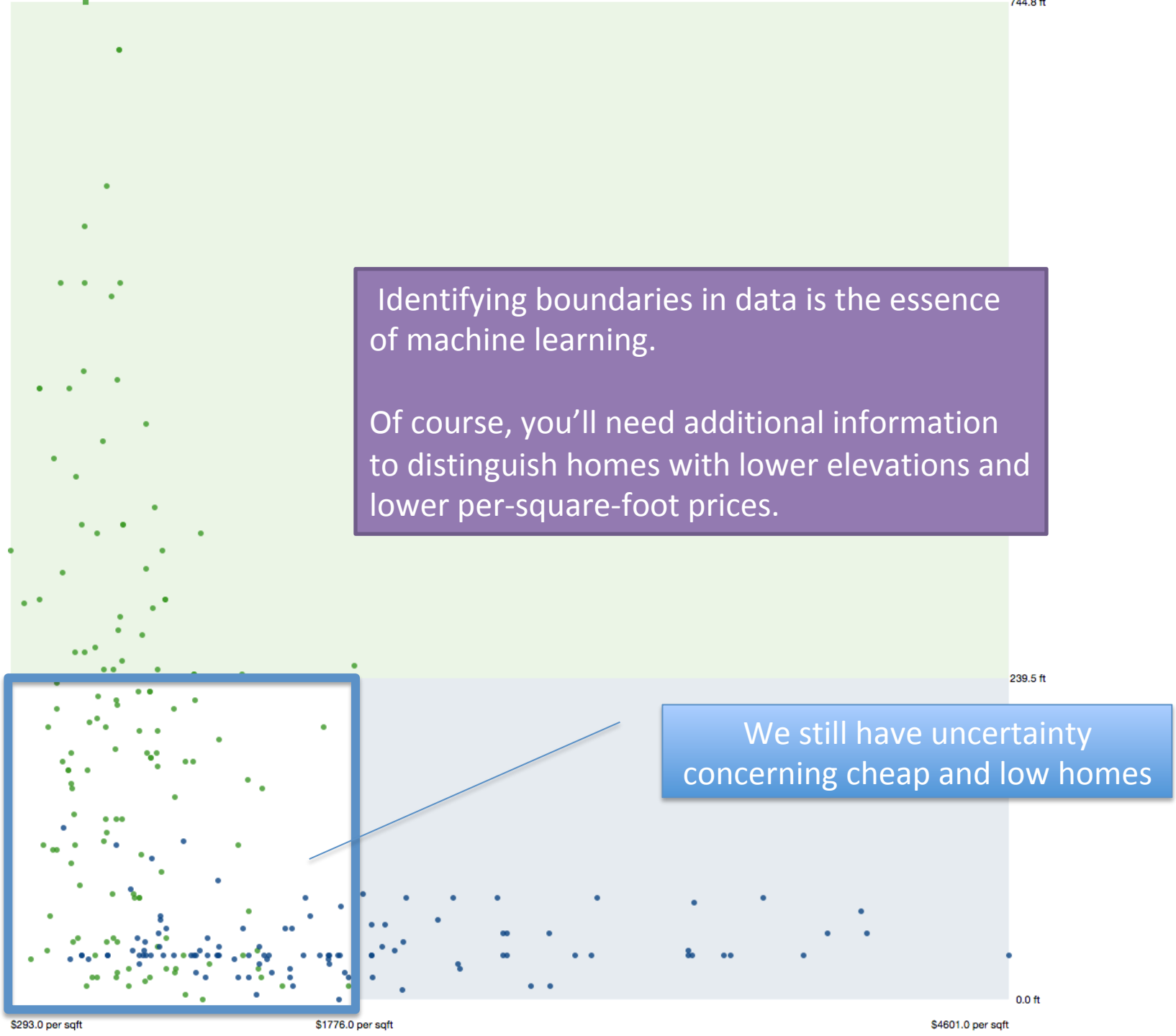


Is this enough?

- What if elevation <73? We cannot tell..
- Adding another **dimension** allows for more nuance. For example, New York apartments can be extremely expensive per square foot.
- Dimensions (such as height, price) are denoted as **FEATURES** in machine learning (= what may characterize the objects we wish to classify). They are also called attributes, descriptors, dimensions..

Elevation *and* price per square foot **scatterplot**

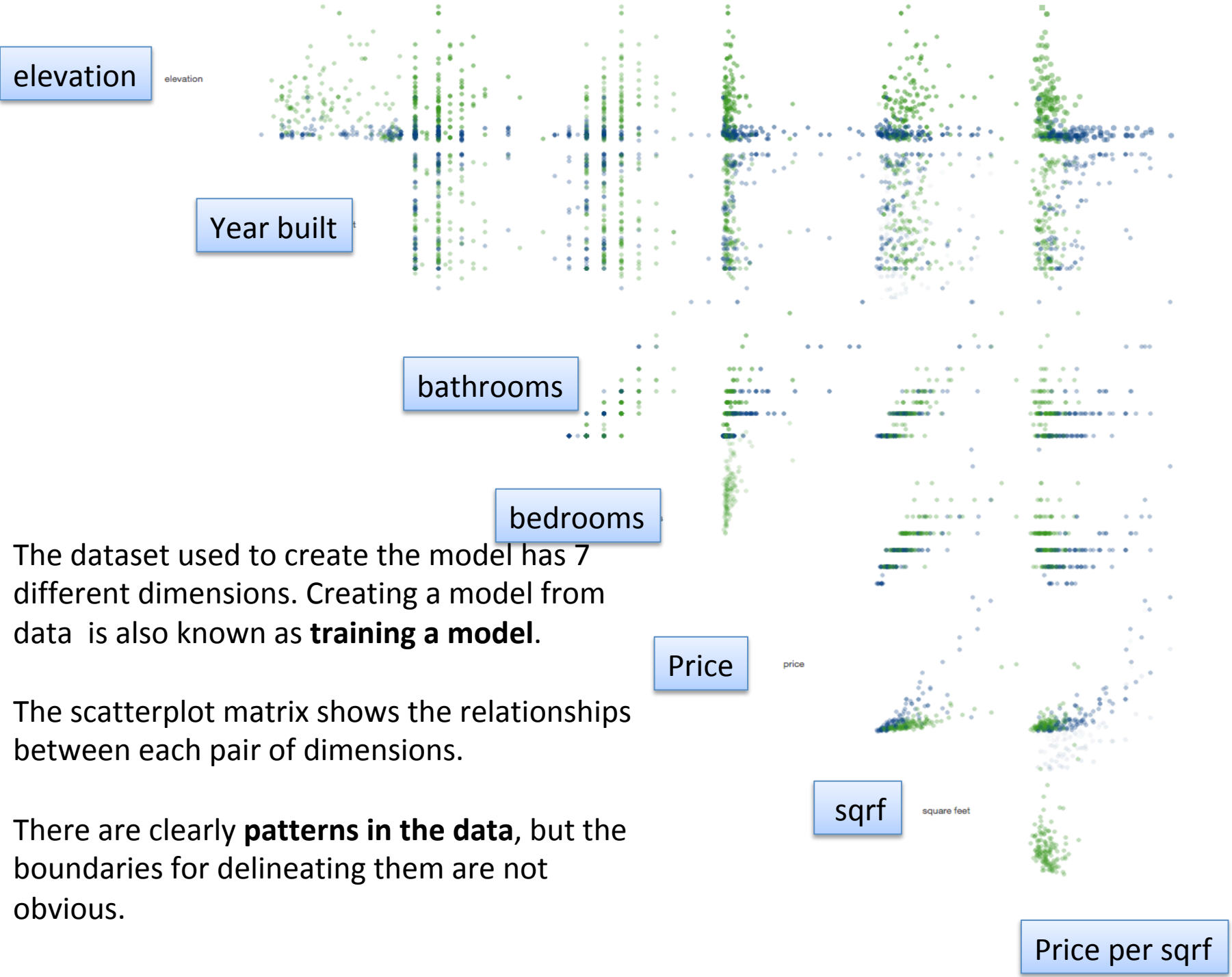




Identifying boundaries in data is the essence of machine learning.

Of course, you'll need additional information to distinguish homes with lower elevations and lower per-square-foot prices.

We still have uncertainty concerning cheap and low homes



The dataset used to create the model has 7 different dimensions. Creating a model from data is also known as **training a model**.

The scatterplot matrix shows the relationships between each pair of dimensions.

There are clearly **patterns in the data**, but the boundaries for delineating them are not obvious.

So, machine learning is…

- Finding patterns in data
- Machine learning methods use learning algorithms to identify regularities and boundaries
- “Boundaries” may come in several forms, they can be probabilities, separating lines, or rules (if height >73 then SF)

Issues in Machine Learning

So, what is Machine Learning?

What is Machine Learning?

- Machine Learning is a scientific discipline that addresses the following question: '*How can we program systems to automatically learn and to improve with experience*? '
- **What** is learning?
- **How** do we learn?
- **What** can we learn?
- **How** can we "improve", and over what??
- **What** is "experience"??

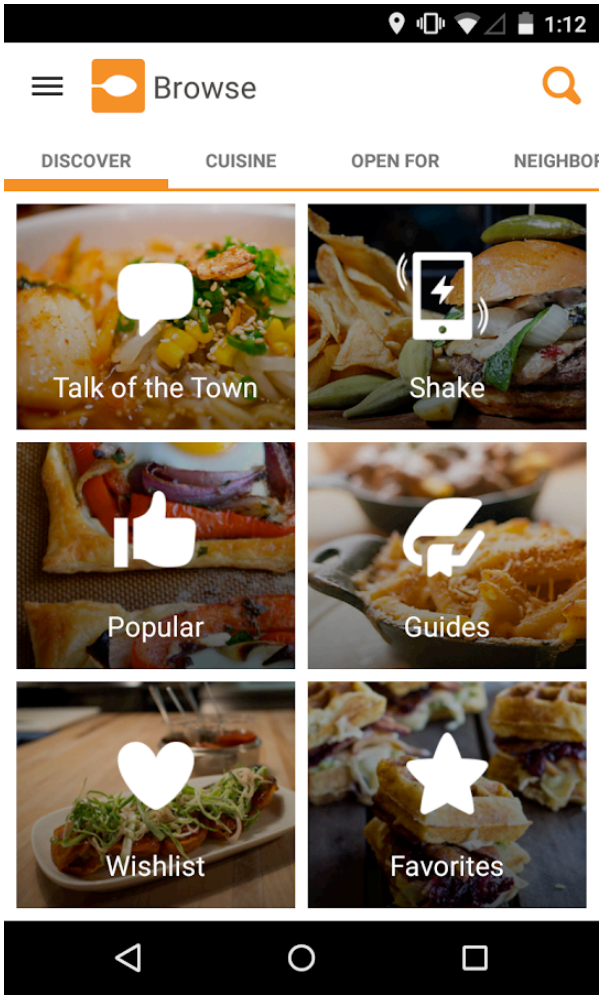
What is learning??

Fire burns!

But we learned using it

you can study (learn) Machine learning

An then build an app to learn your favorite food and advise you on restaurants



So, what is learning?

- **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.

- http://www.skillsyouneed.com/general/learning.html#ixzz3QtQQTKvg

What is learning?

UNDERSTAND + GAIN KNOWLEDGE + USE NEW KNOWLEDGE TO DO SOMETHING

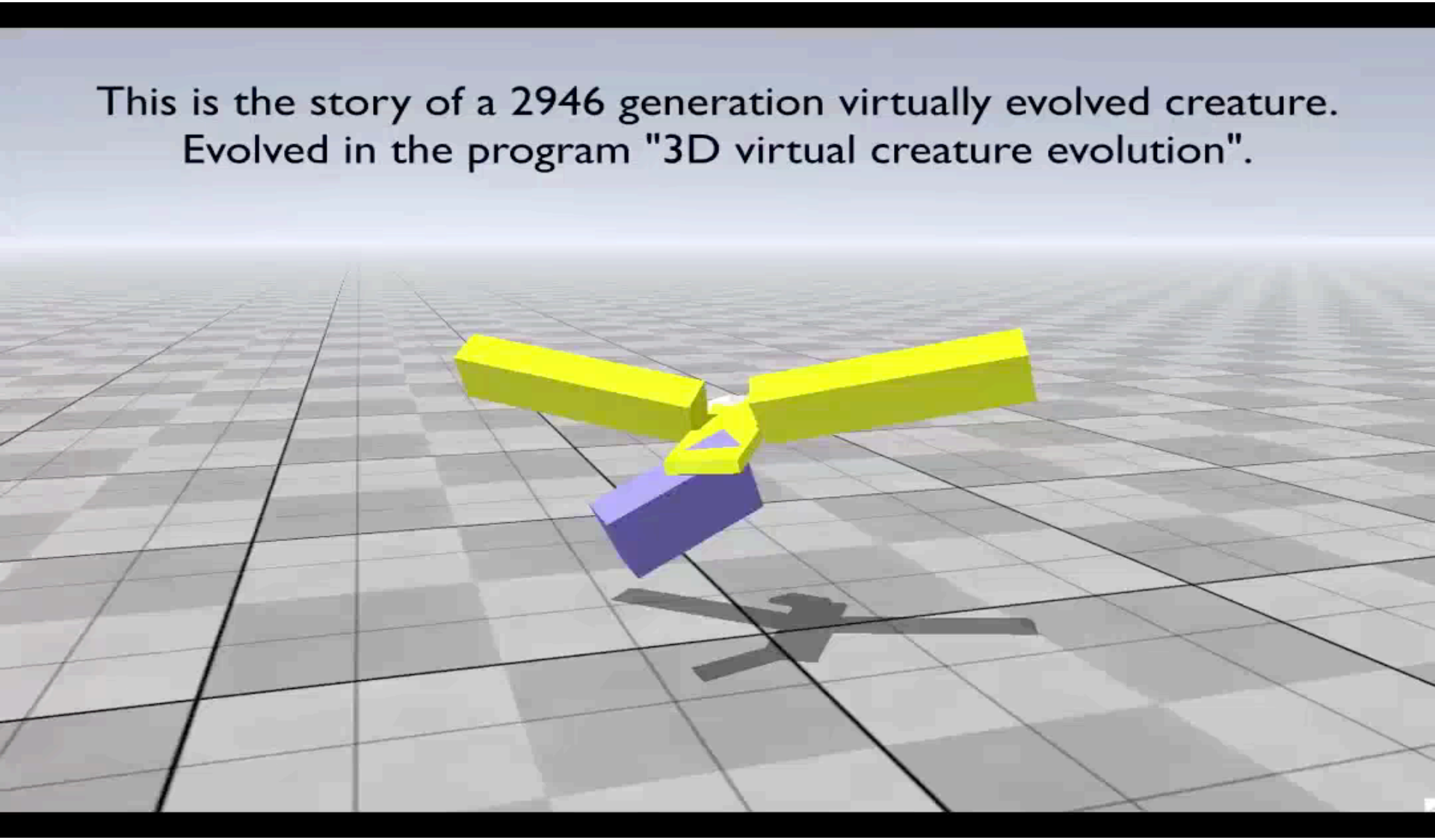
But, how do we learn??

How do humans learn?

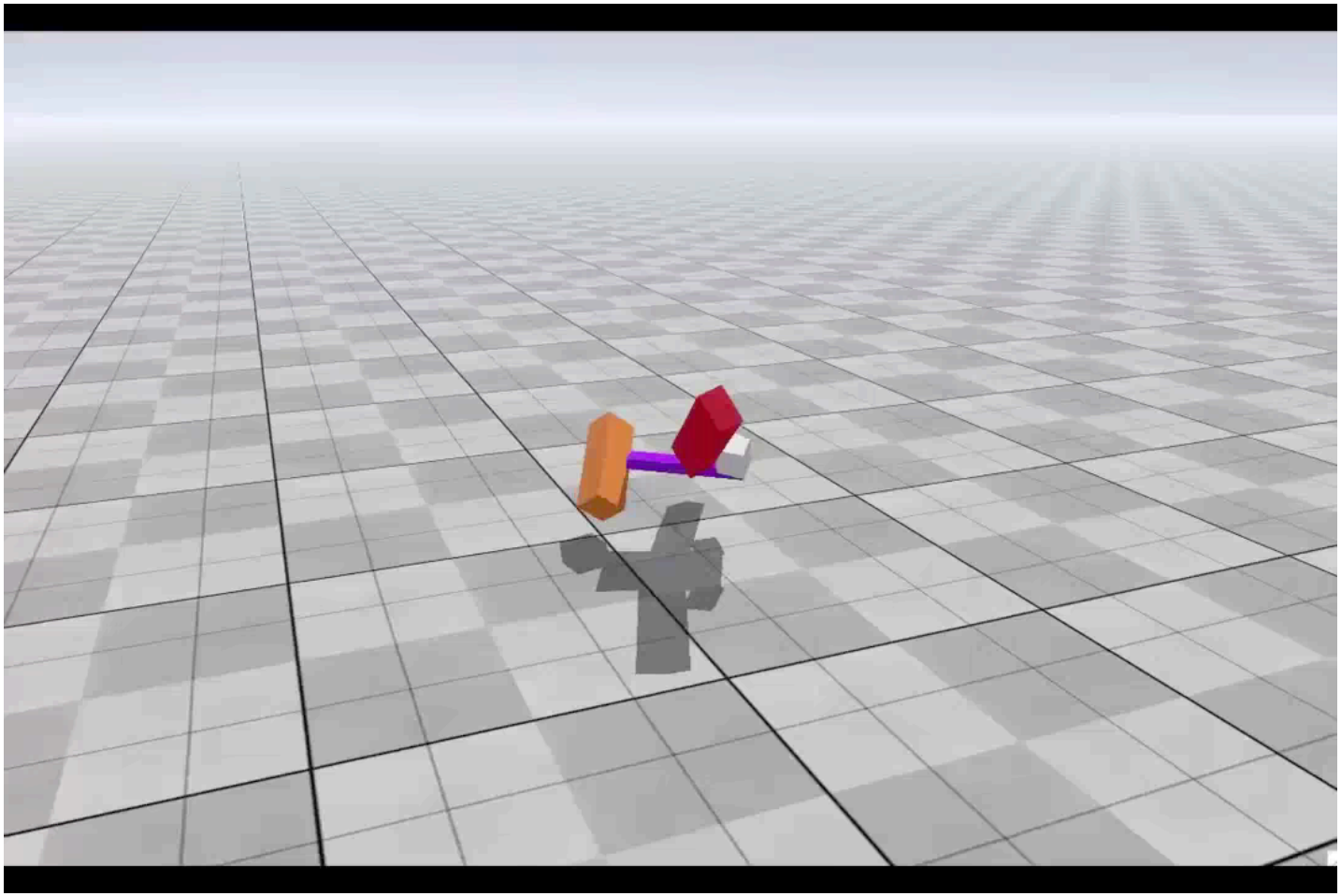
- Someone tell us (teacher, or whatching others)
- Try and test (learning by doing) - As in the fire example -

Is there something humans cannot learn??

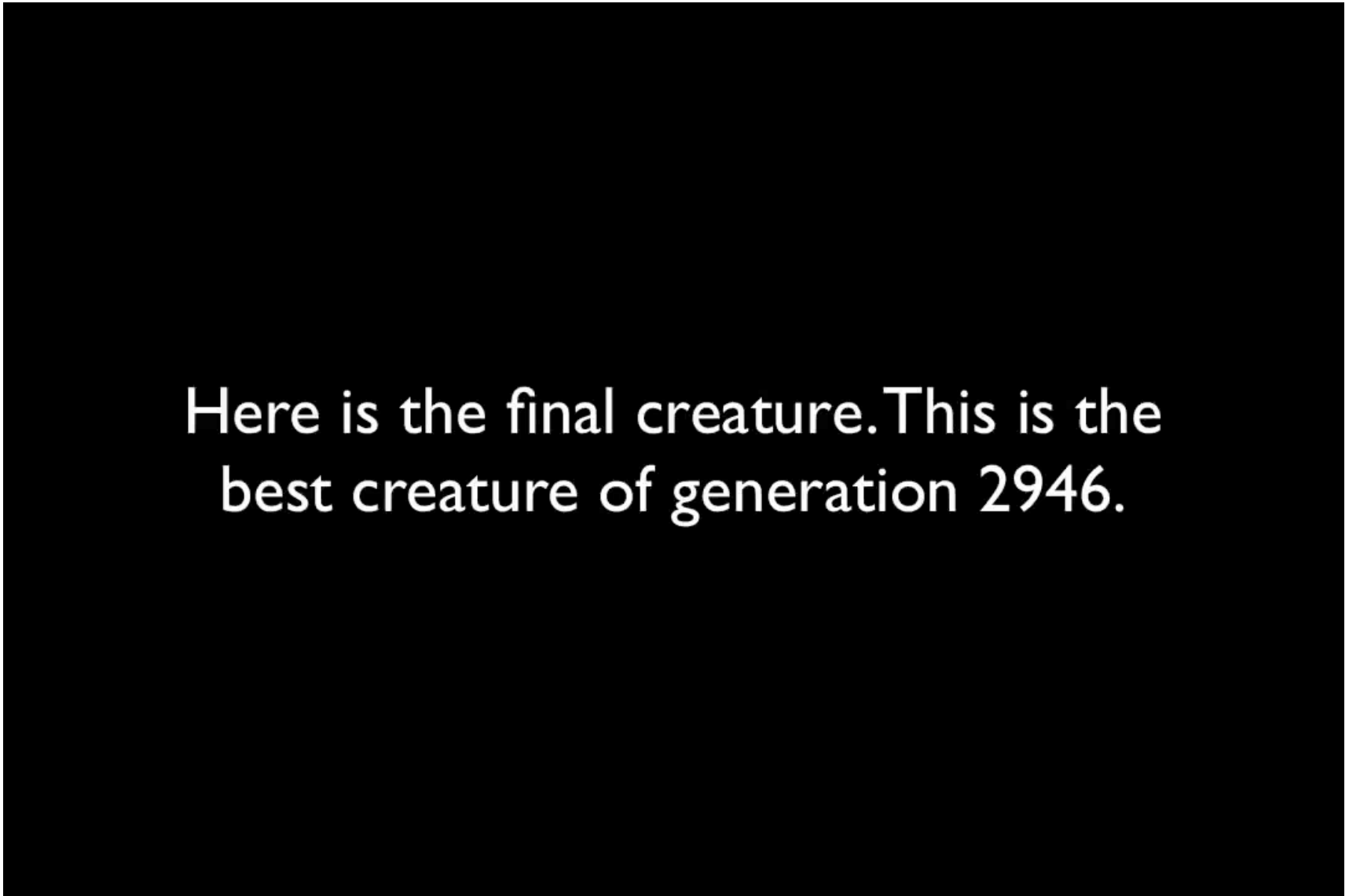
Machines that learn to fly



.. After some time..



After some more time..



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? (1)

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

When is it difficult for humans to learn? (2)

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

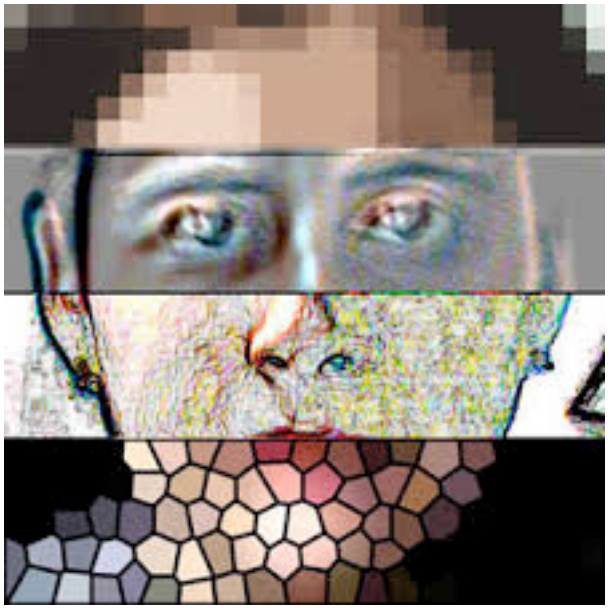
Stock market values
And quotes

When is it difficult for humans to learn? (3)

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

When is it difficult for humans to teach? (4)

- 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? (1)

- 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:
    - solution changes in time (market data for market forecast)
    - Solution needs to be adapted to particular cases (personalized systems for recommendation, diagnosis, etc.)

So when is it advisable to use Machine Learning? (2)

- 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 filter
  - Personalized tutoring
  - Recommenders
- Discover new knowledge from large databases (***data mining***).
  - Market basket analysis (e.g. diapers and beer)
  - Medical text mining (e.g. migraines to calcium channel blockers to magnesium)
  - Twitter mining

An interdisciplinary topic: many related disciplines!

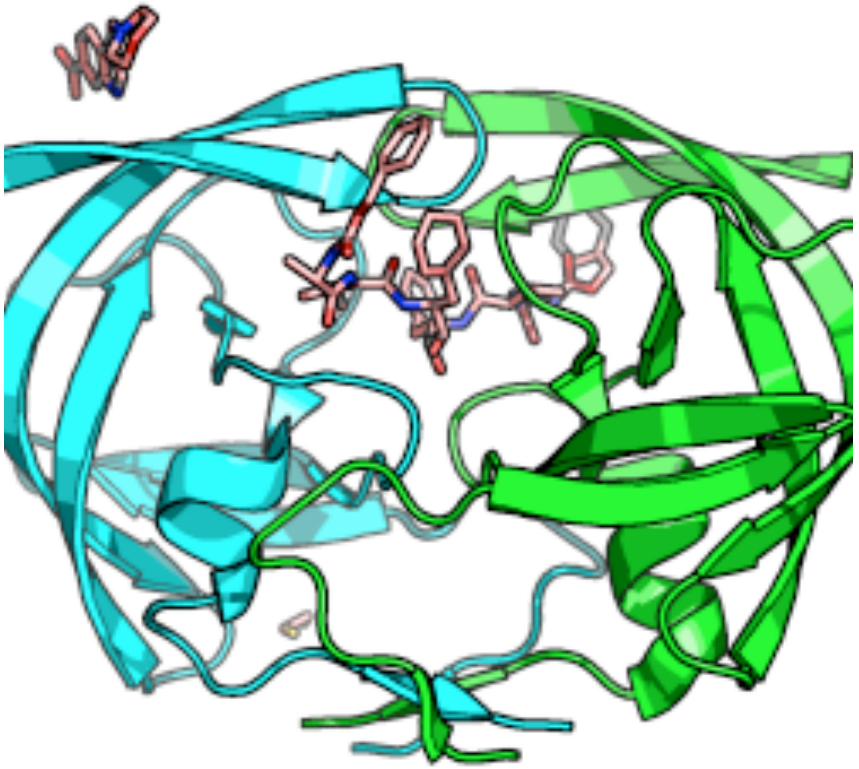
- Artificial Intelligence
- Data Mining
- Probability and Statistics
- Information theory
- Numerical optimization
- Computational complexity theory
- Control theory (adaptive)
- Psychology (developmental, cognitive)
- Neurobiology
- Linguistics
- Philosophy

ML is perhaps the most interdisciplinar of CS areas!!

Some “real hot” ML applications

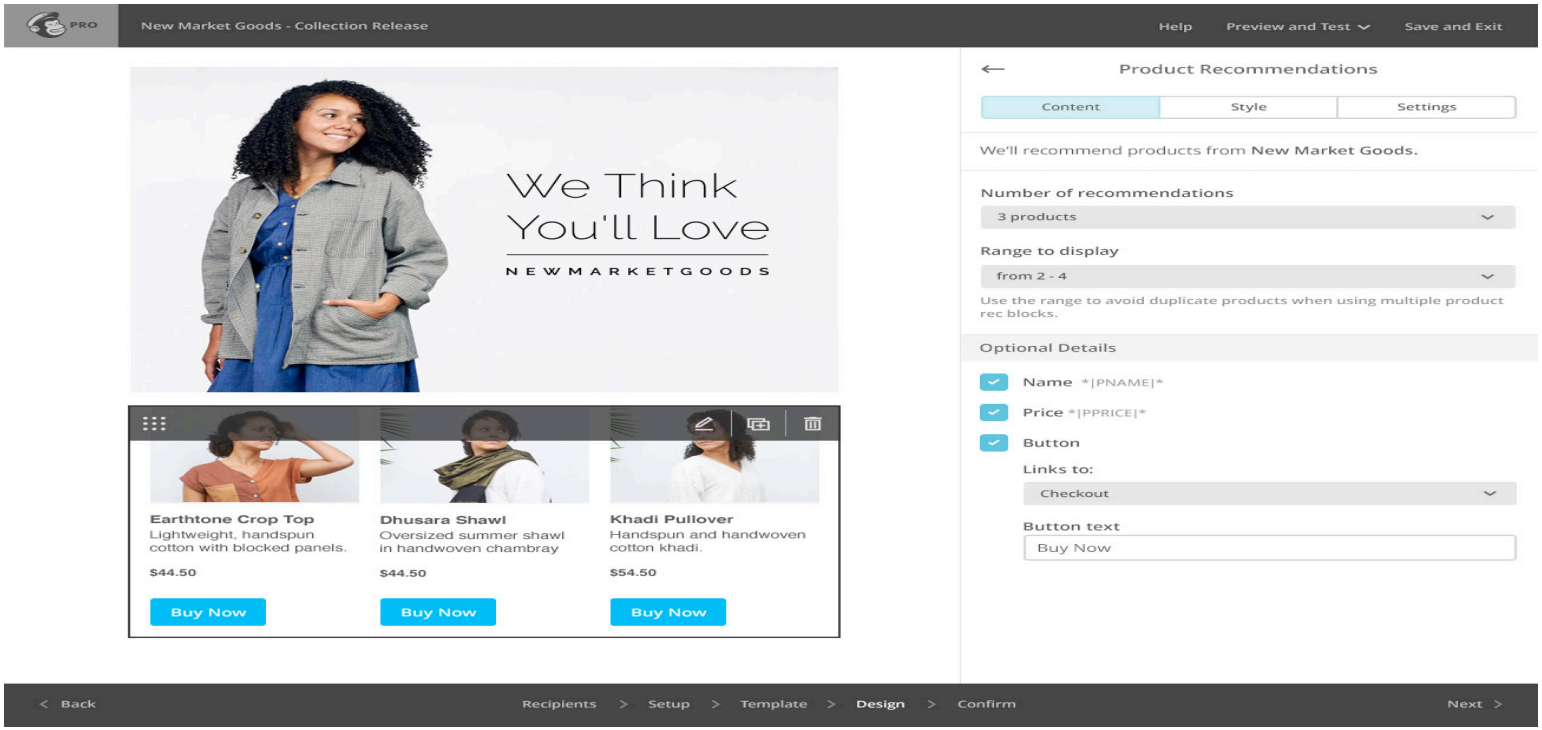
- It is really hard to find a problem where machine learning is not already applied -- machine learning is practically everywhere, in business applications and science!
- Here is a list of “hot” applications:

Computational Biology & E-health



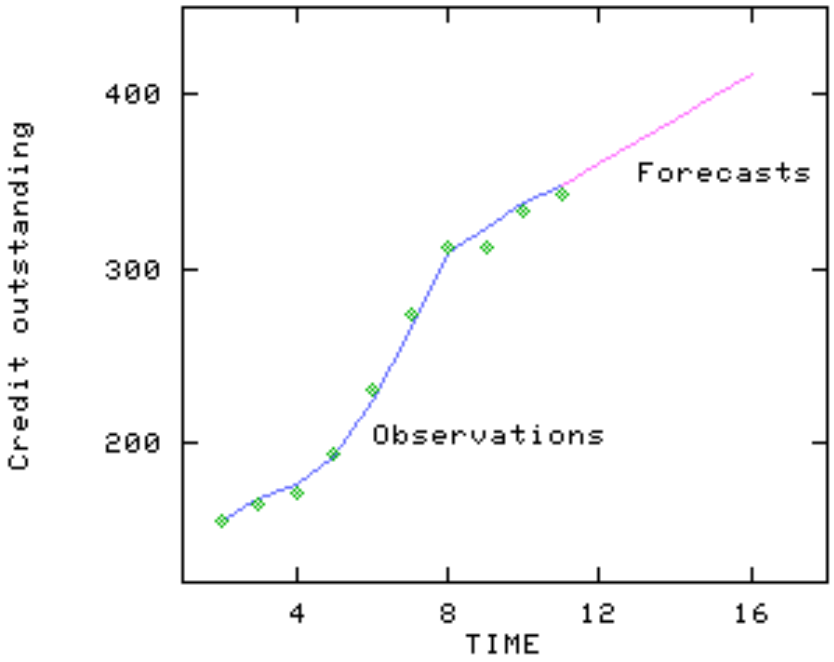
- Predicting diseases and complications from patient's health records
- Predicting disease genes trough the analysis of biological networks (e.g. interactions between proteins)
- Predicting epidemics through the analysis of human interaction data (e.g., population density, data on population moves, climatic 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 fraud
- 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

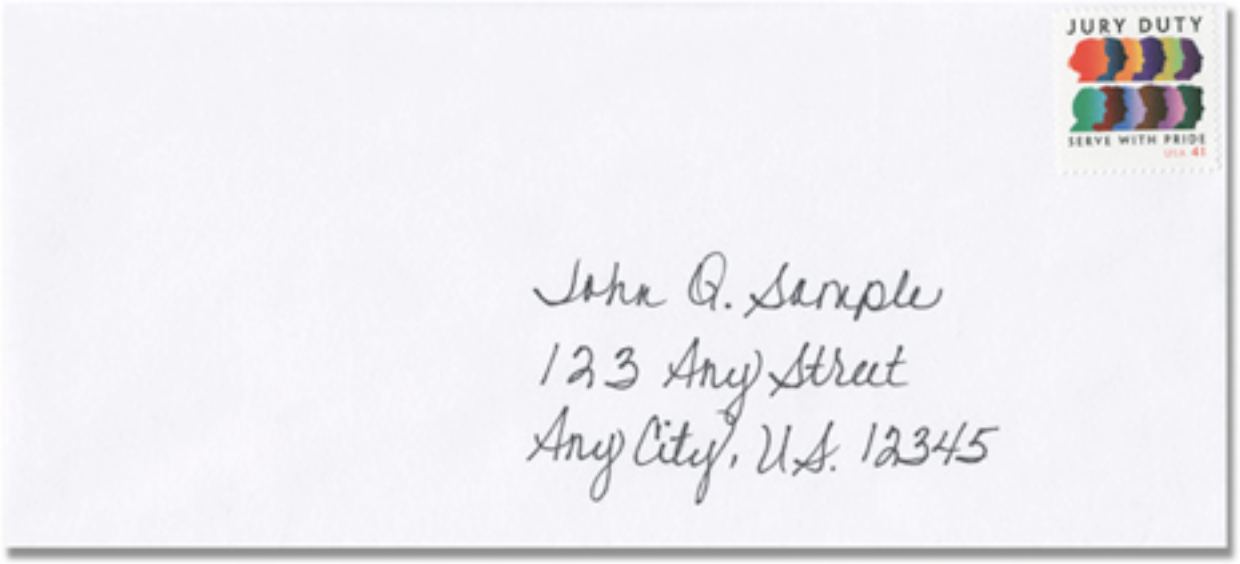
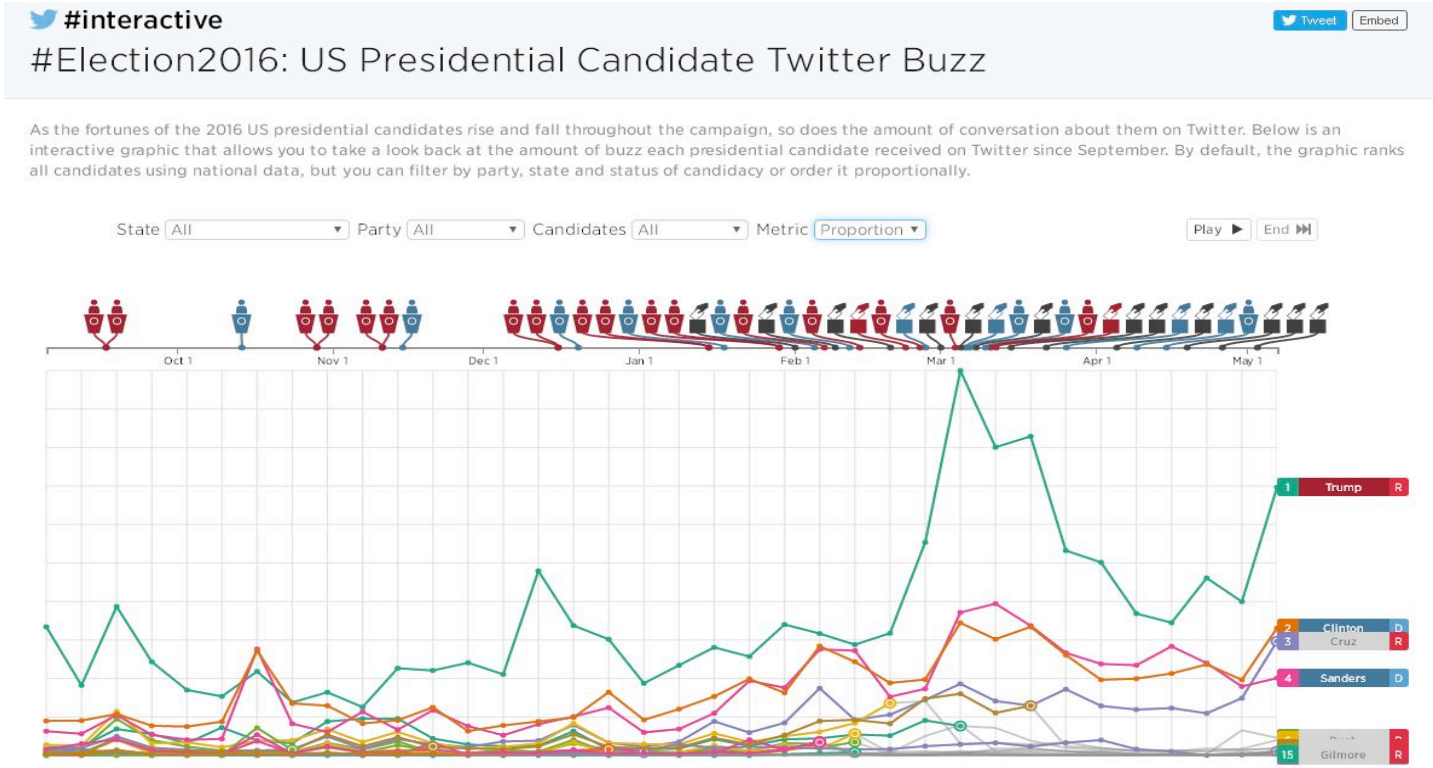


Image Understanding and Robotics

- Indentifcation 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 ORGANIZATION AND SYLLABUS

Course material

- Slides (partly) from: http://www.cs.utexas.edu/users/mooney/cs391L/ and many other sources
- Textbook: Tom Mitchell, Machine Learning, McGraw Hill, 1997 (new 2017 chapters on http://www.cs.cmu.edu/~tom/NewChapters.html)
- Introduction to machine learning ETHEM ALPAYDIN (on line book)
- Deep learning (MIT press): https://www.deeplearningbook.org/
- Course twiki http://twiki.di.uniroma1.it/twiki/view/ApprAuto

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

Course Syllabus

1. Concept learning and Decision Tree Learning
2. Practical Machine learning: feature selection and feature engineering
3. Evaluation methods: experimental and theoretical methods
4. Artificial Neural Networks and Deep learning (convolutional and denoising)
5. Support Vector Machines
6. Probabilistic Learning: Naive Bayes
7. Ensamble Methods
8. Unsupervised learning:
    1. Rule learning (Apriori, FP-Growth)
    2. Clustering methods
9. Reinforcement learning: Q-learning and Genetic Algorithms

Course labs

- Algorithms experimented on Weka toolkit http://www.cs.waikato.ac.nz/ml/weka/ (only for those really weak with programming) AND on http://scikit-learn.org/stable/ (a python platform). Other libraries can be used (e.g. TensorFlow for deep methods)
- Objective of labs is learning **practical ML**: selecting features, processing real life datasets, choosing algorithms, hyper-parameter tuning, evaluation experiments.

Caveat: Coverage of ML topics is limited!

- This is a first-level “basic” ML course
- On year 2 there is an advanced course
- ML algorithms for specific applications (NLP, security..) are also taught in other courses

Exam and Project

- Written exam on course material (60%)
- + Course Project: free choice of a topic (40%)
- Projects can be carried on by teams of 2 students
- How should the project be:
  - Not a trivial problem. Choose a **real life** problem, or invent one
  - Sufficiently **large data**set (many repositories available). Best if merging different datasets
  - Dataset must need some **feature engineering** (some non trivial 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**

How is the course organized

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