Universitat
de les Illes Balears

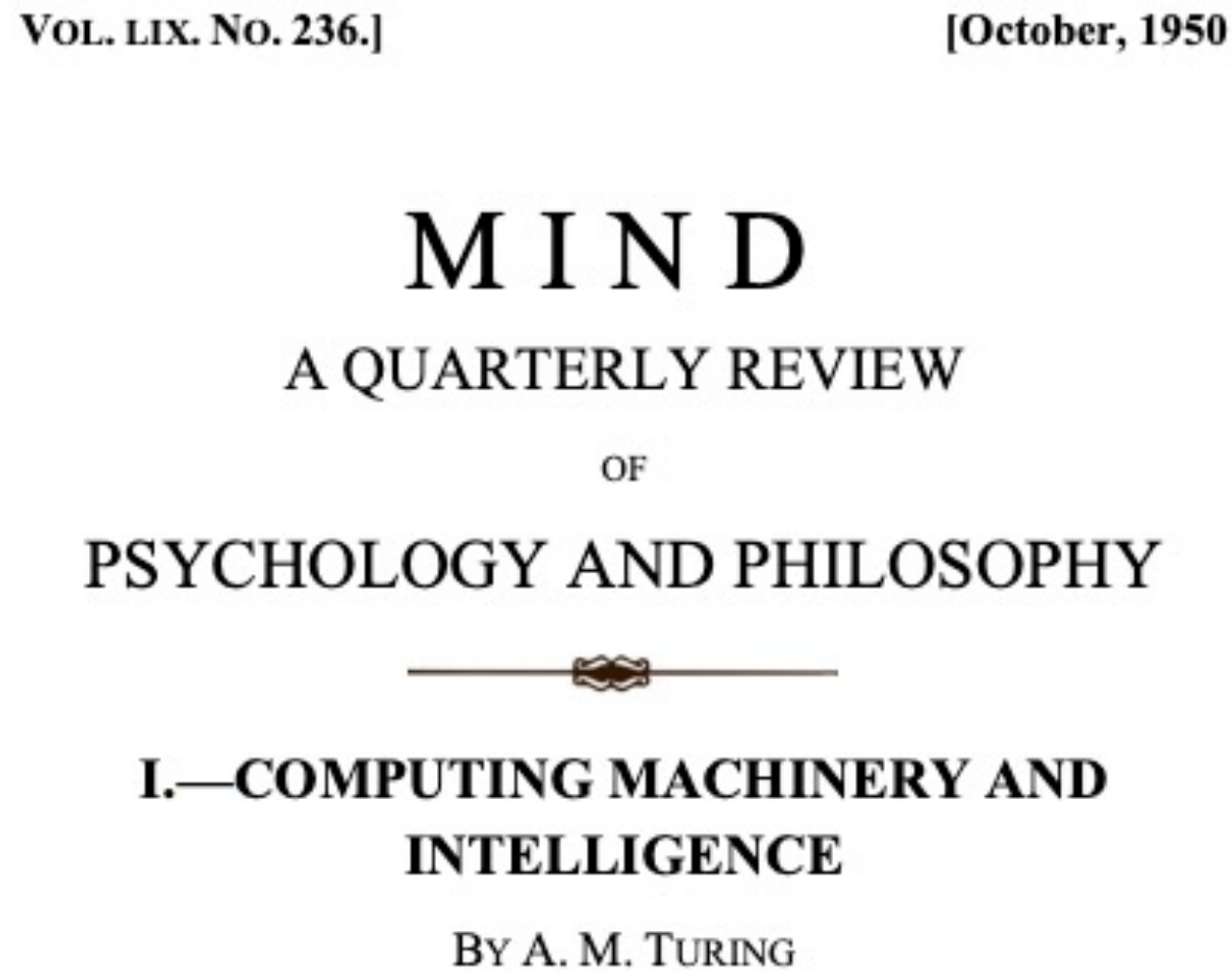
Machine Learning

Lesson 1: Introduction

Machine Learning in the context of Artificial Intelligence

I propose to consider the question:

Can machines think?



VOL. LIX. NO. 236.] [October, 1950

MIND

A QUARTERLY REVIEW

OF

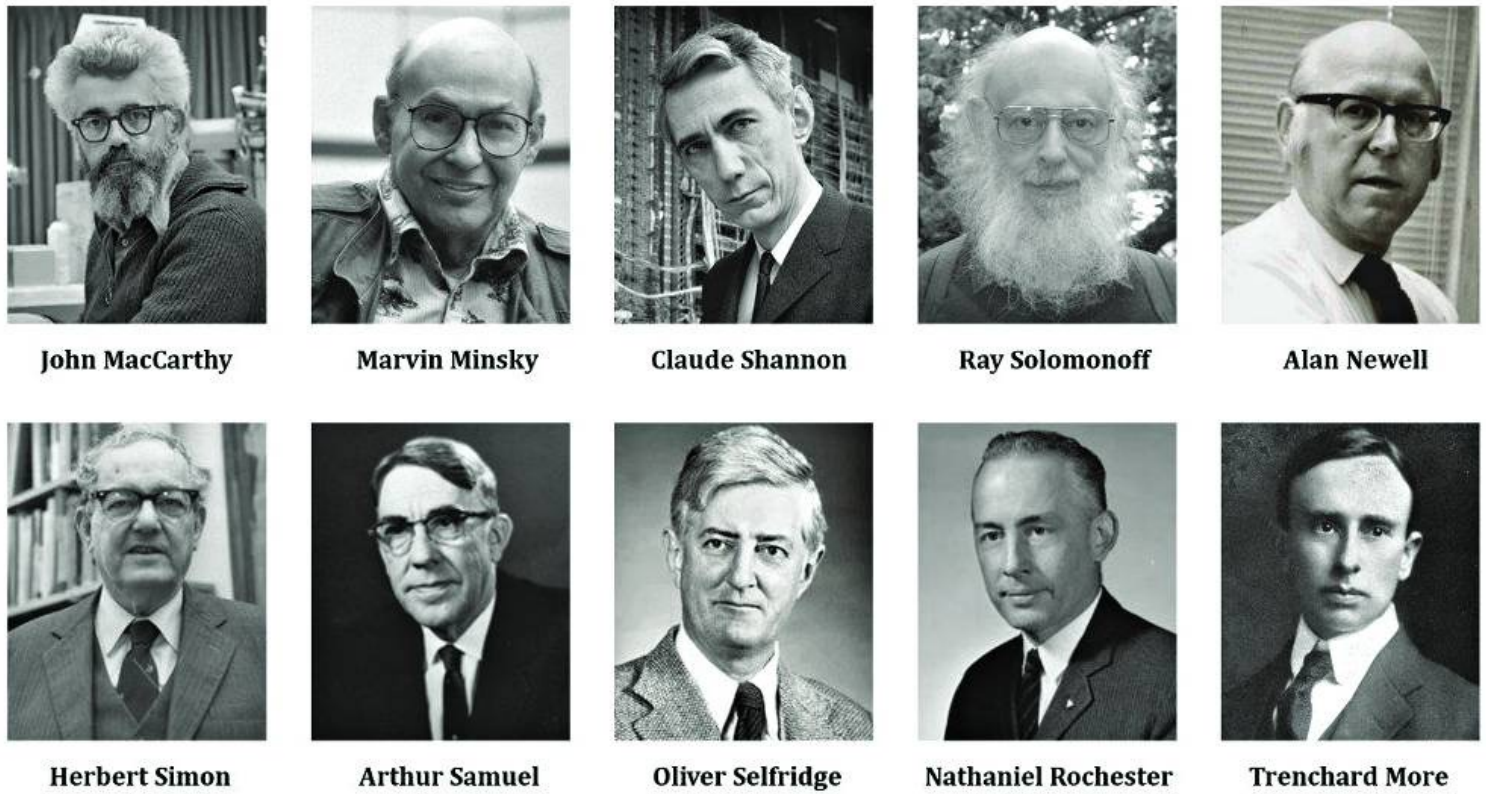
PSYCHOLOGY AND PHILOSOPHY

I.—COMPUTING MACHINERY AND INTELLIGENCE

BY A. M. TURING

Full paper

1956 Dartmouth Conference: The Founding Fathers of AI



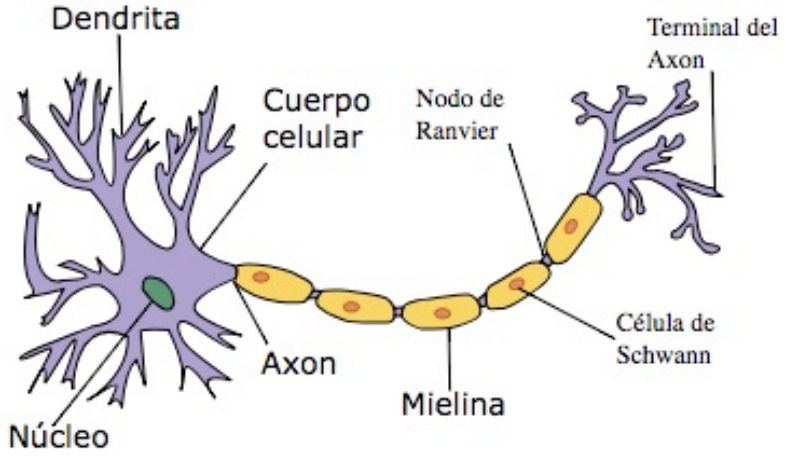
The study is to proceed on the basis of the conjecture that **every aspect of learning** or any other feature of **intelligence can in principle be so precisely described that a machine can be made to simulate it**.

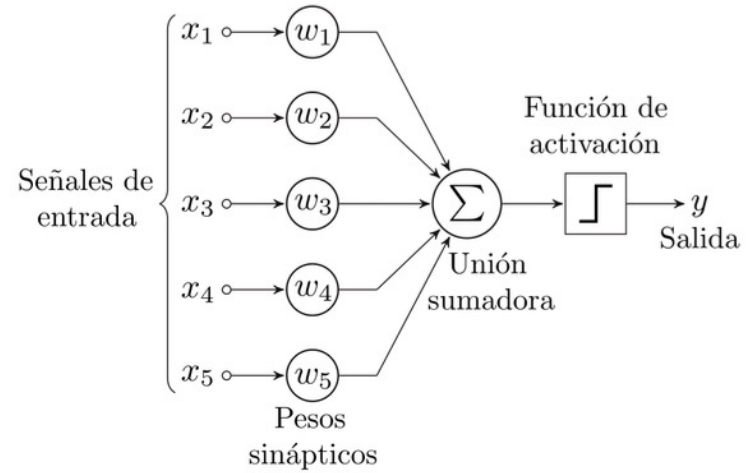
Psychological Review
Vol. 65, No. 6, 1958

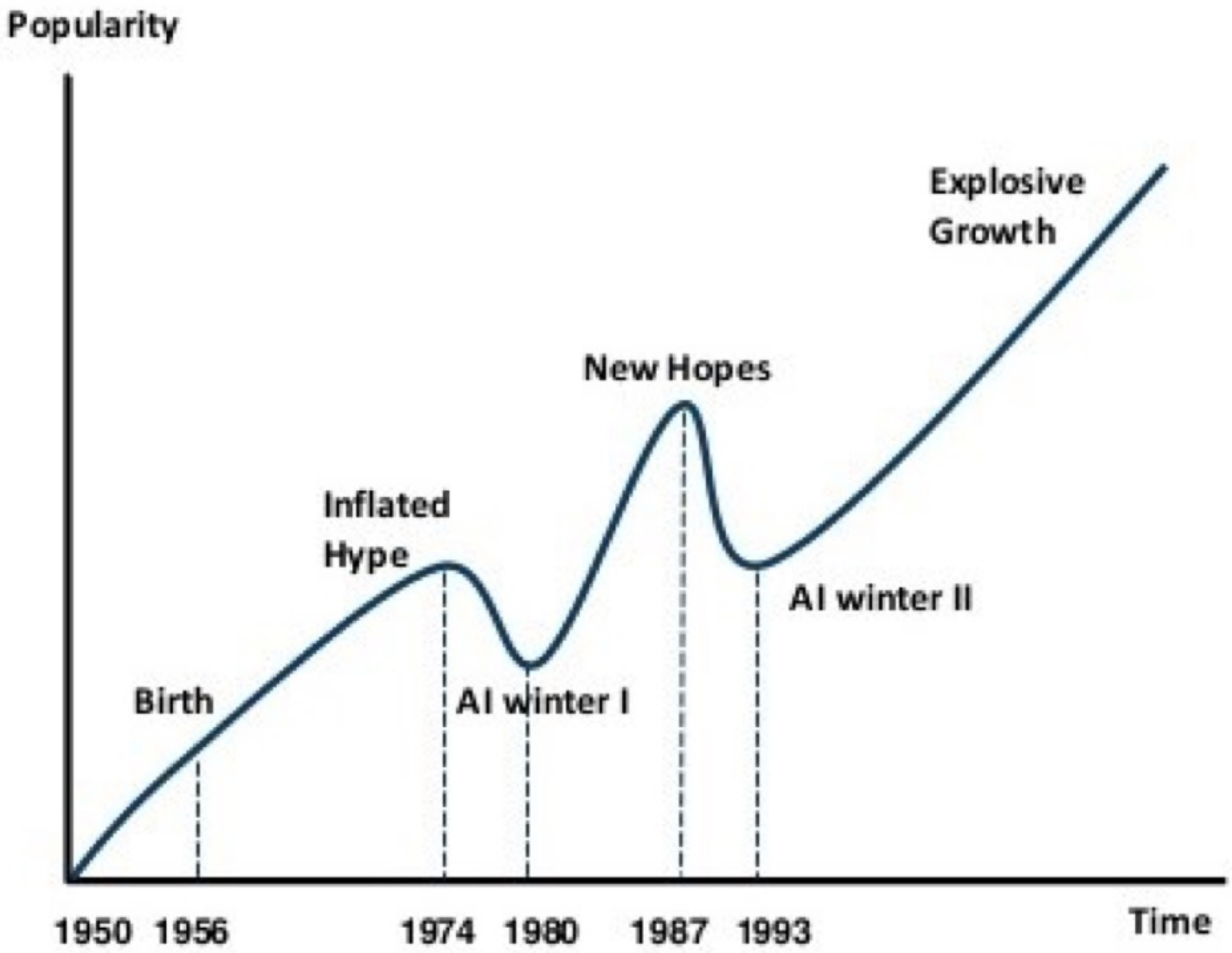
THE PERCEPTRON: A PROBABILISTIC MODEL FOR INFORMATION STORAGE AND ORGANIZATION IN THE BRAIN[1]

F. ROSENBLATT

Cornell Aeronautical Laboratory







Timeline of AI Development

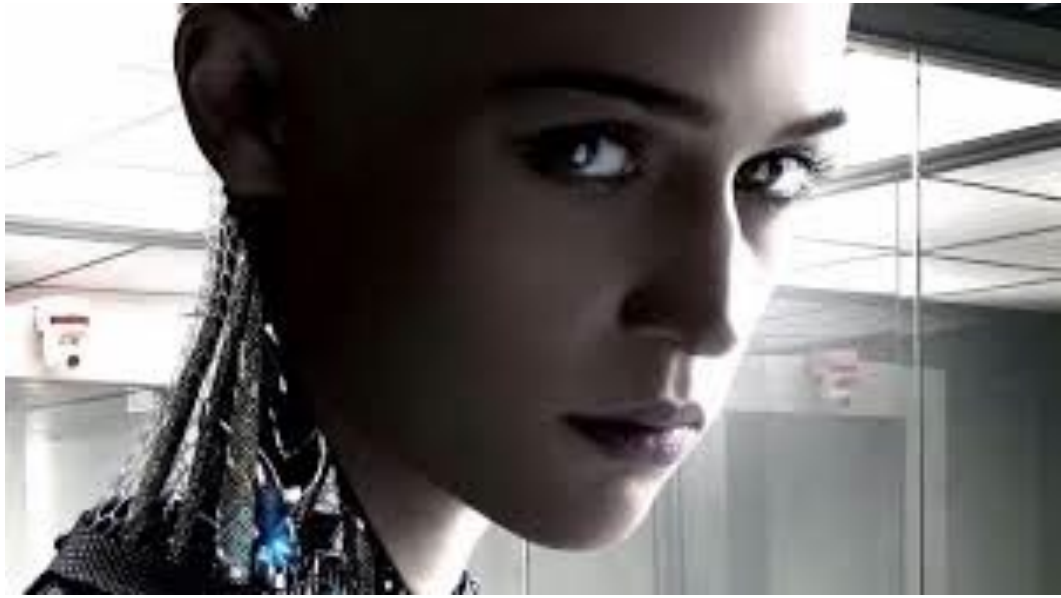
- **1950s-1960s**: First AI boom - the age of reasoning, prototype AI developed
- **1970s**: AI winter I
- **1980s-1990s**: Second AI boom: the age of Knowledge representation (appearance of expert systems capable of reproducing human decision-making)
- **1990s**: AI winter II
- **1997**: Deep Blue beats Gary Kasparov
- **2006**: University of Toronto develops Deep Learning
- **2011**: IBM's Watson won Jeopardy
- **2016**: Go software based on Deep Learning beats world's champions

Artificial Intelligence

Definition

1. A branch of computer science dealing with the simulation of intelligent behavior in computers

2. The capability of a machine to imitate intelligent human behavior

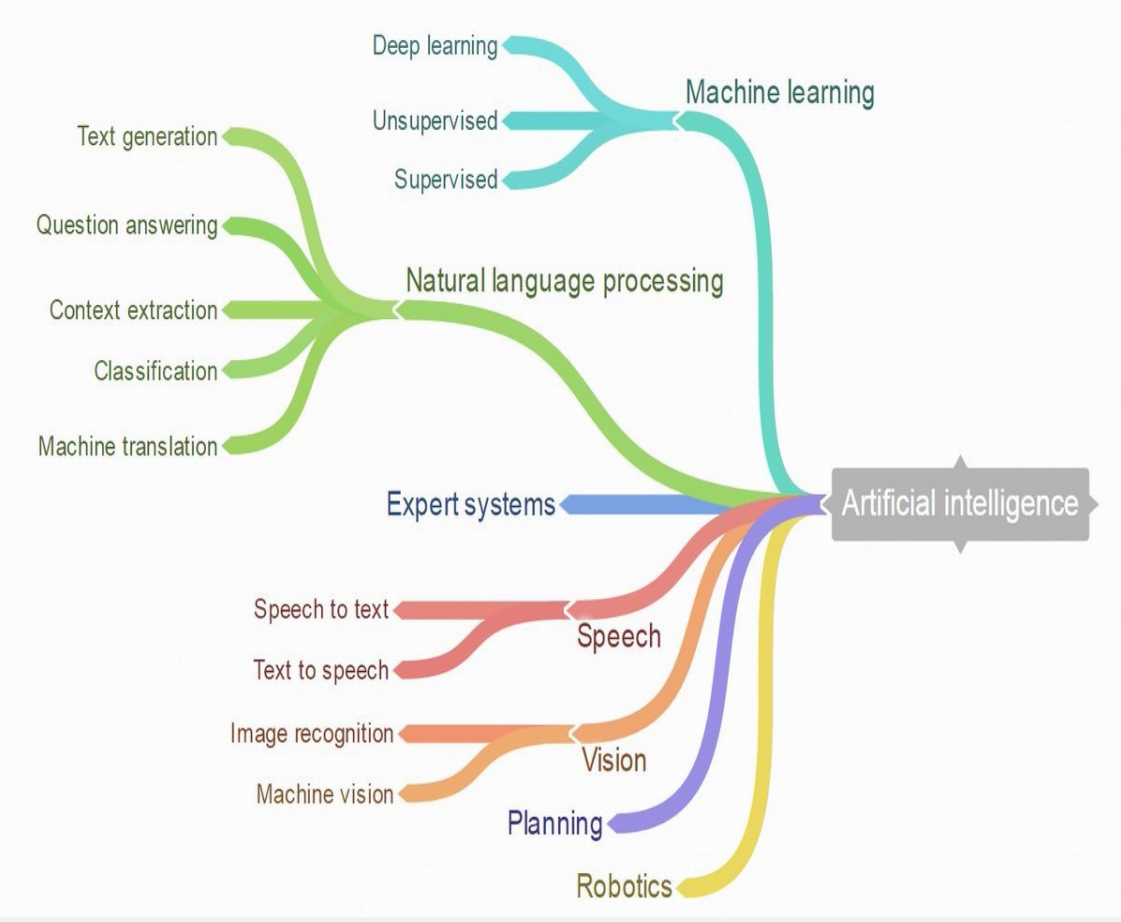


AI technologies

It is a set of rich sub-disciplines and methods (machine learning, pattern recognition, computer vision,…)

We have to consider all these different disciplines and methods in seeking true solutions in delivering value to human beings and organizations

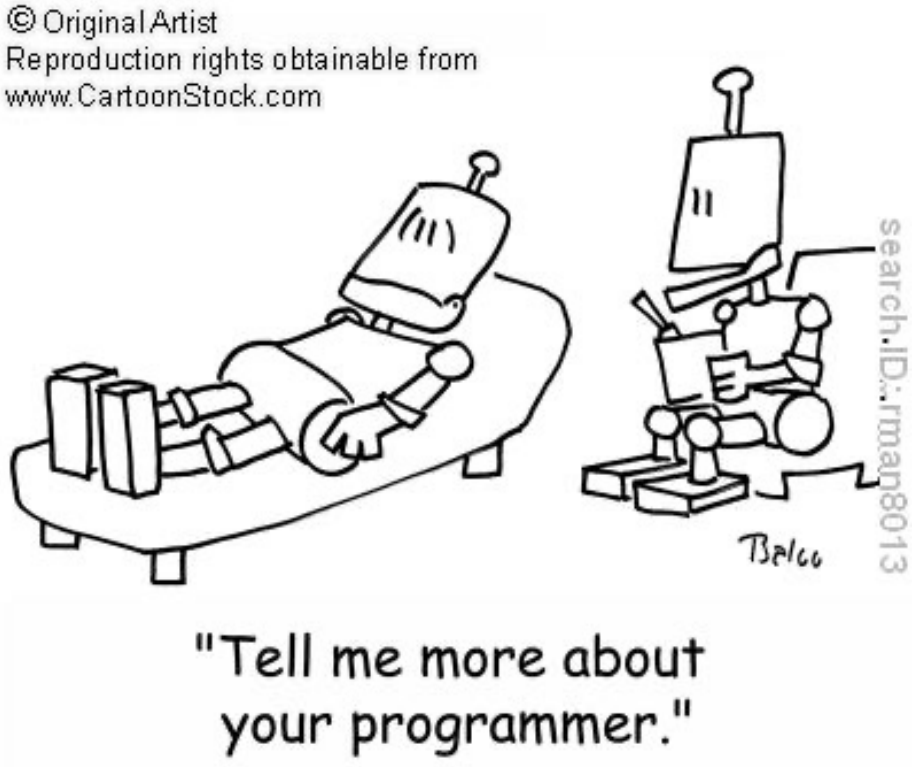
Eric Horvitz (Microsoft Research)

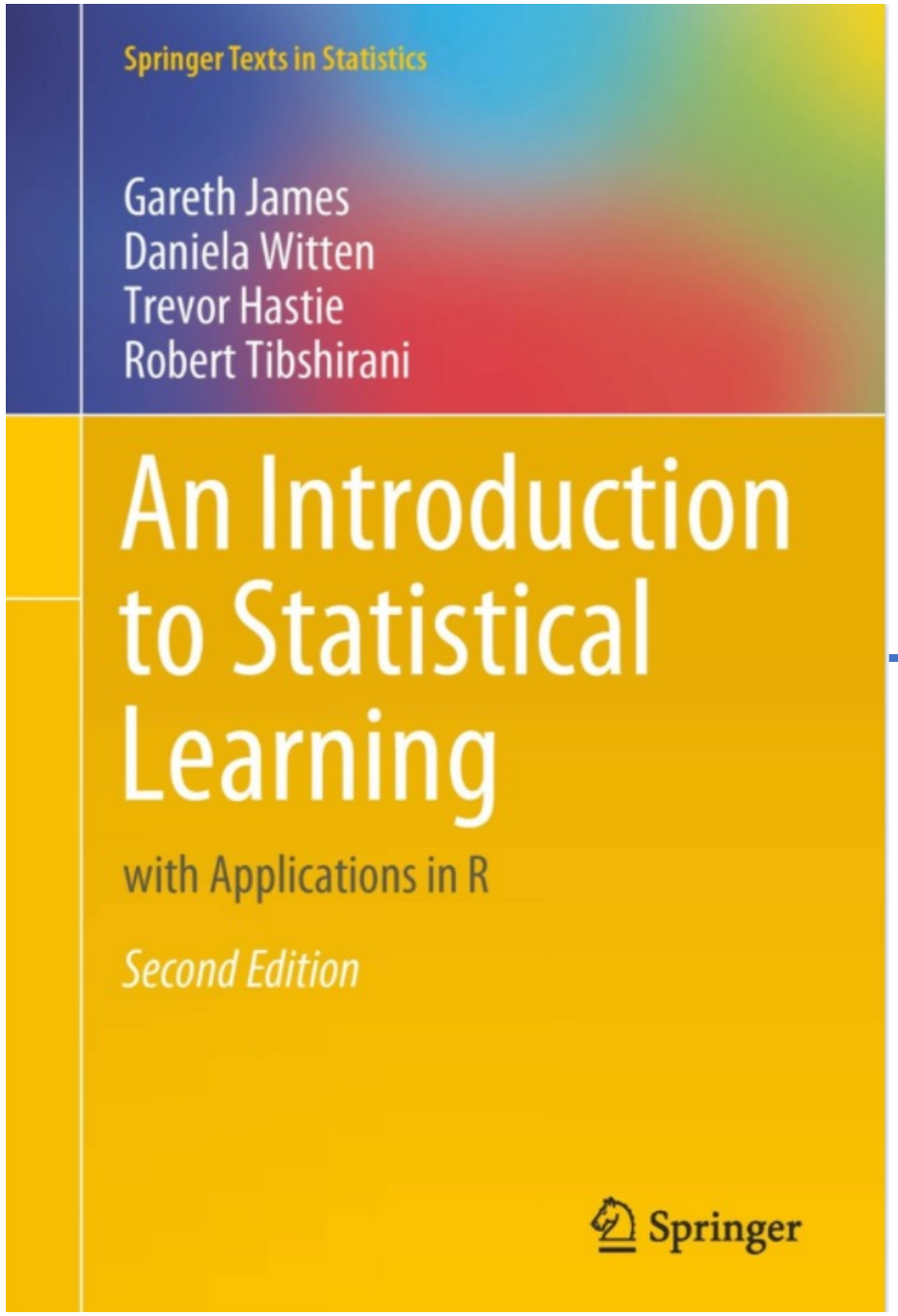


Machine Learning

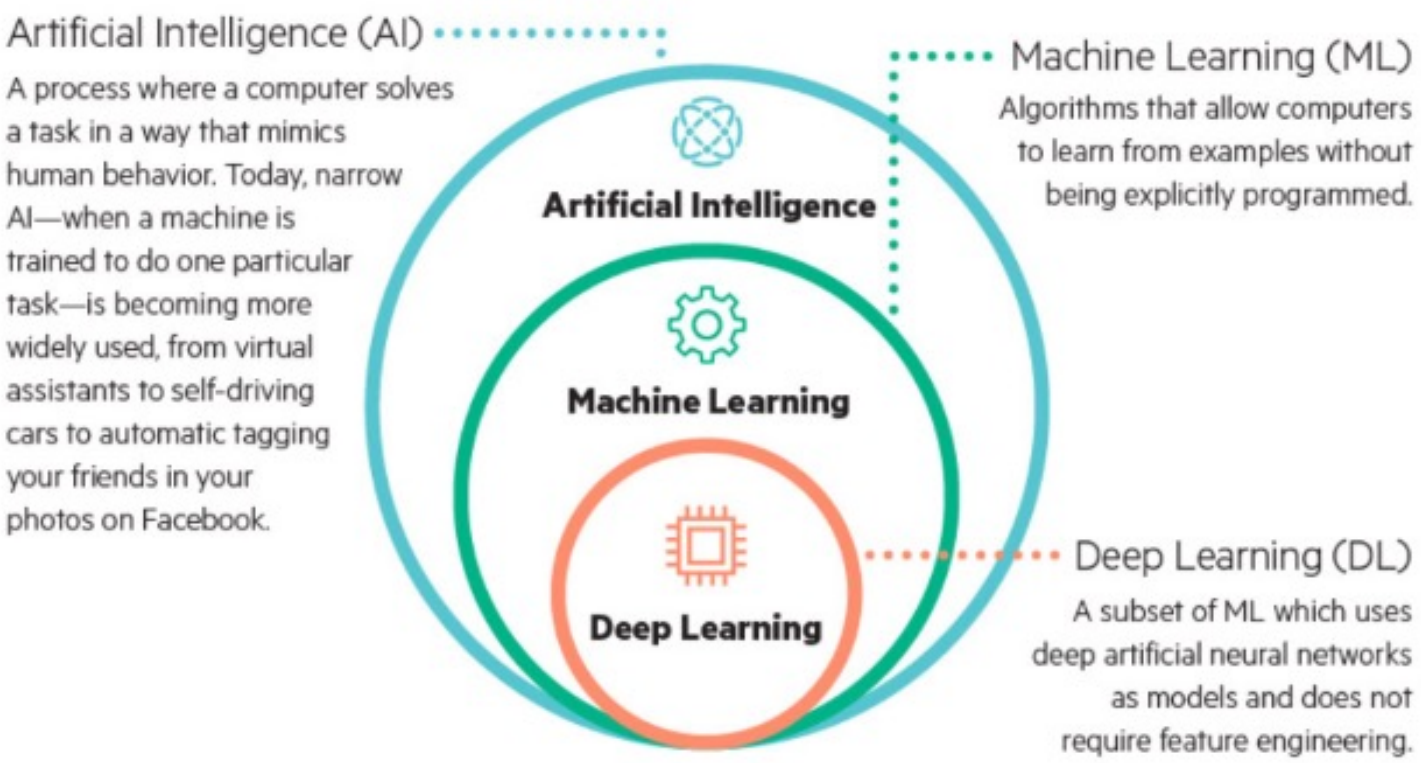
Definition

Machine learning is a subset of artificial intelligence in the field of computer science that often uses statistical techniques to give computers the ability to "learn" (i.e., progressively improve performance on a specific task) **with data**, without being explicitly programmed.

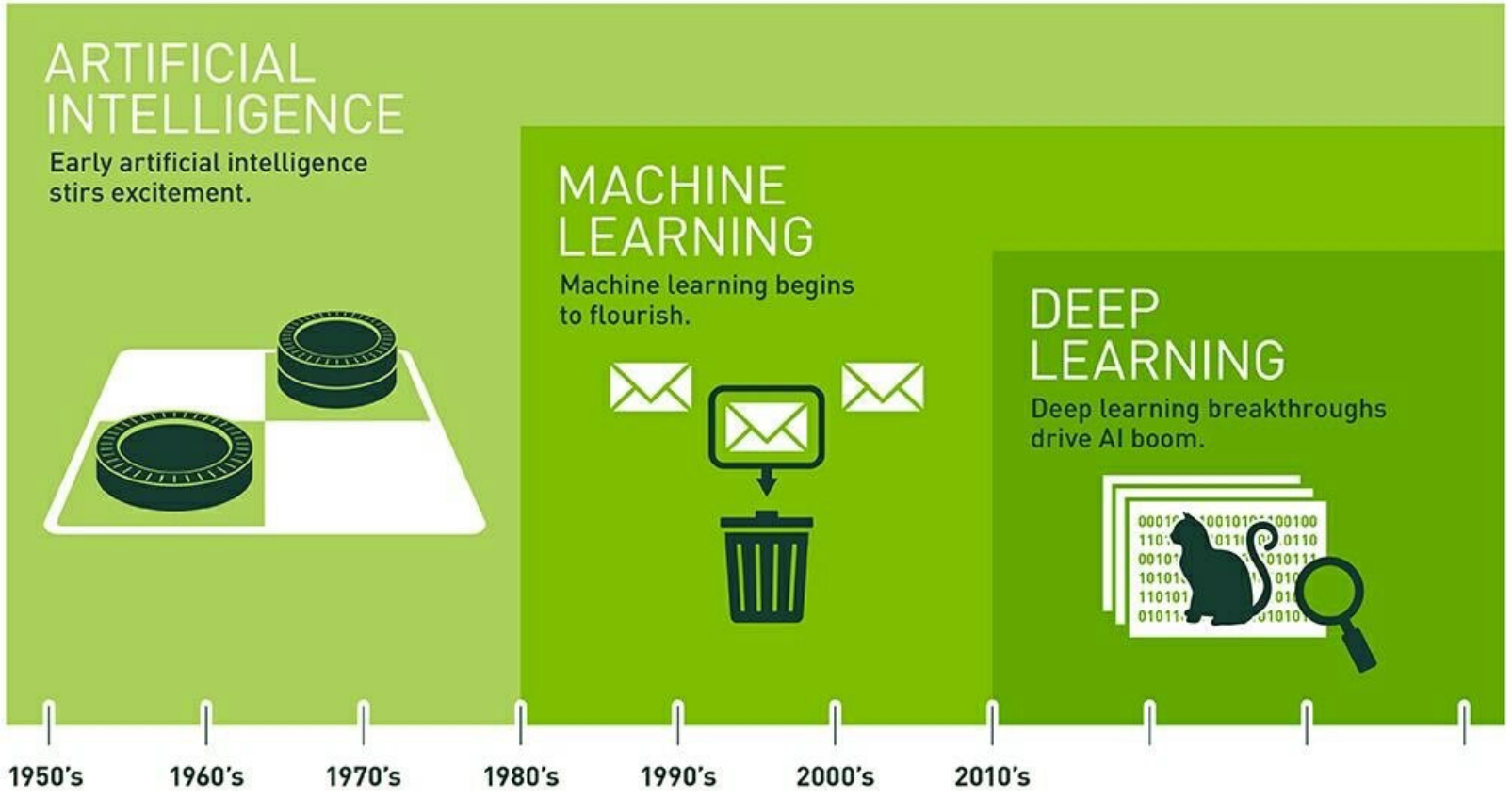



Springer Texts in Statistics
Gareth James
Daniela Witten
Trevor Hastie
Robert Tibshirani
An Introduction to Statistical Learning
with Applications in R
Second Edition
Springer

Machine Learning in the context of AI



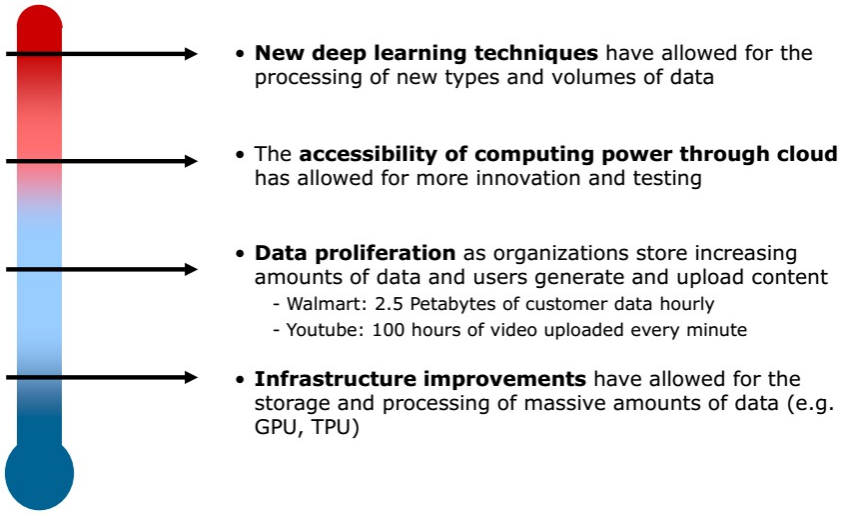
Machine Learning in the context of AI



Since an early flush of optimism in the 1950s, smaller subsets of artificial intelligence – first machine learning, then deep learning, a subset of machine learning – have created ever larger disruptions.

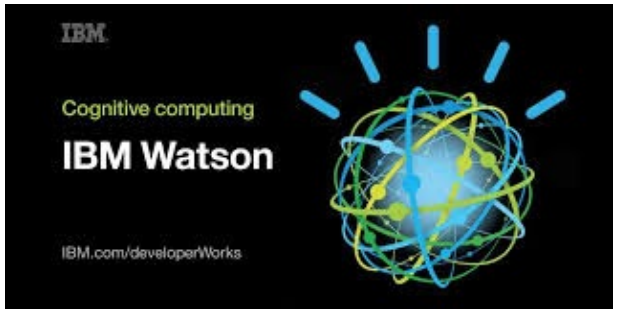
Why now ?

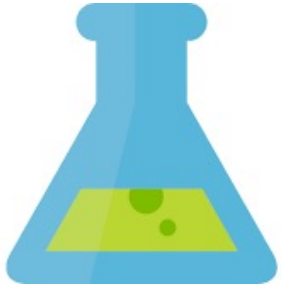
- Computational power
  - cloud computing, hardware (gpu, tpu), availability, cost…
- Big data
  - Over the last years 90 percent of the data in the world was generated.
- Democratization of tools:
  - Open source toolbelt
  - Wide community adoption
- Research and applied research:
  - Economical value created from AI → more funding → more applications

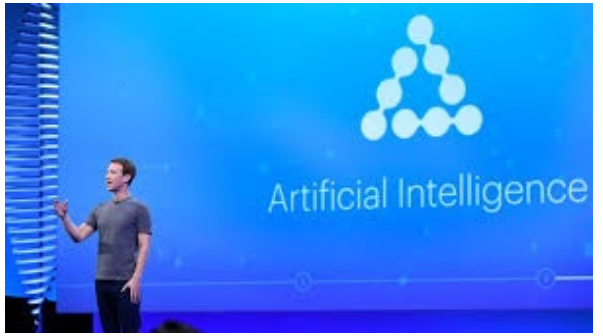


Big data, the big bang

ML expansion: companies







Machine Learning?

Without learning, the agents must be AI programmed during the design. Therefore the machine intelligence depends from the **designer** and his ability to predict the diversity of the environment and its possible changes

It is possible?

ML formalization

A computer program is said to learn from experience **E** with respect to some class of tasks **T** and performance measure **P**, if its performance at tasks in **T**, as measured by **P**, improves with experience **E**.

Tom Mitchell (CMU, 1998)

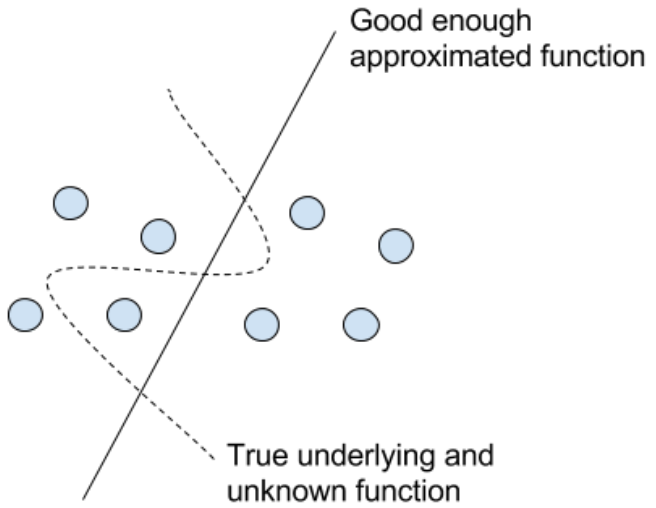
Se dice de un programa que aprende de la experiencia **E** con respecto a la tarea **T** y una medida de rendimiento **P** si el rendimiento **P** sobre **T** aumenta con la experiencia **E**

T. Mitchel, 1997

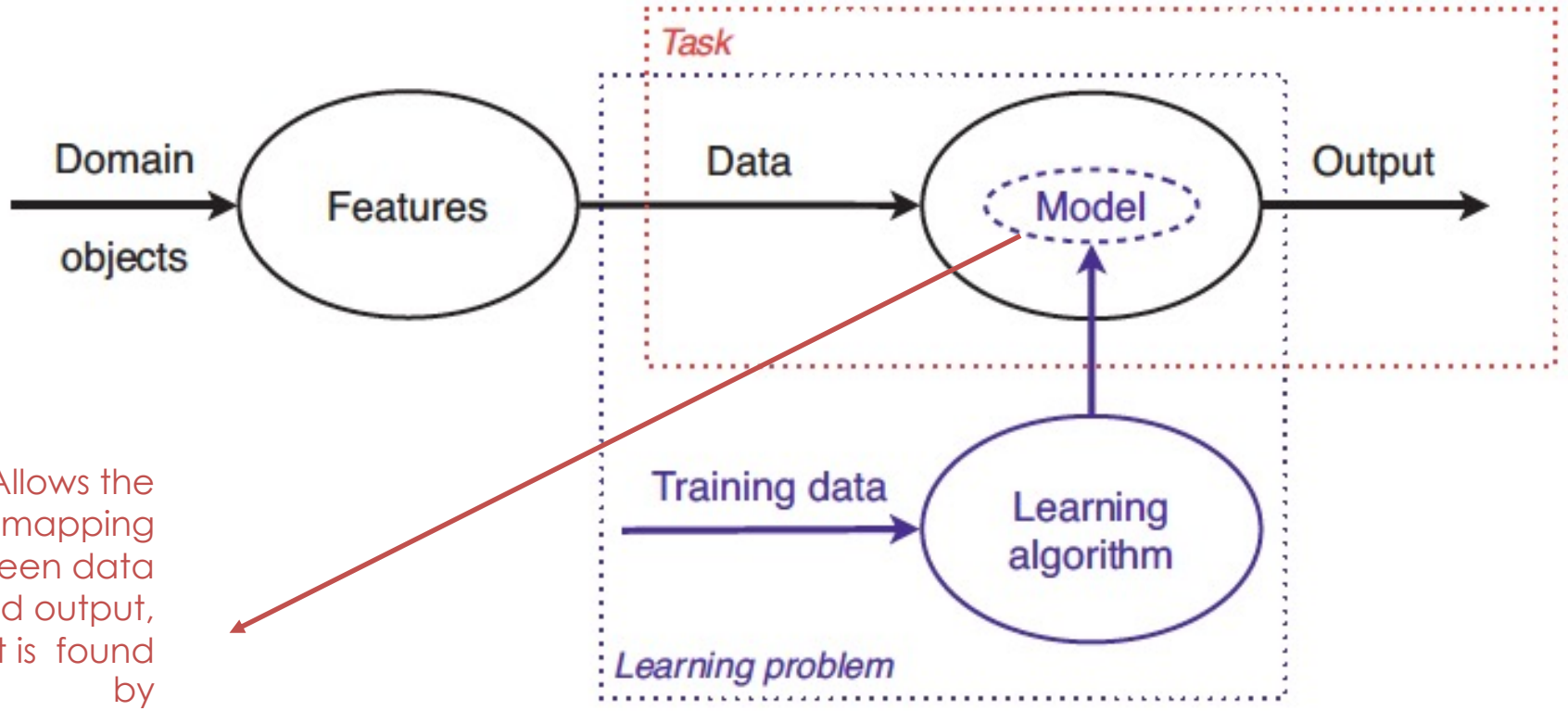
Campo de estudio que da a los ordenadores la capacidad de aprender sin ser programados de forma específica

A. Samuel, 1959

Podemos entender un algoritmo de ML como **un algoritmo que cambia su estado interno** para encontrar el major **mapeo** entre una variable de salida y sus características de entrada



Task: machine learning model



Allows the mapping between data and output, and it is found by computational learning algorithms

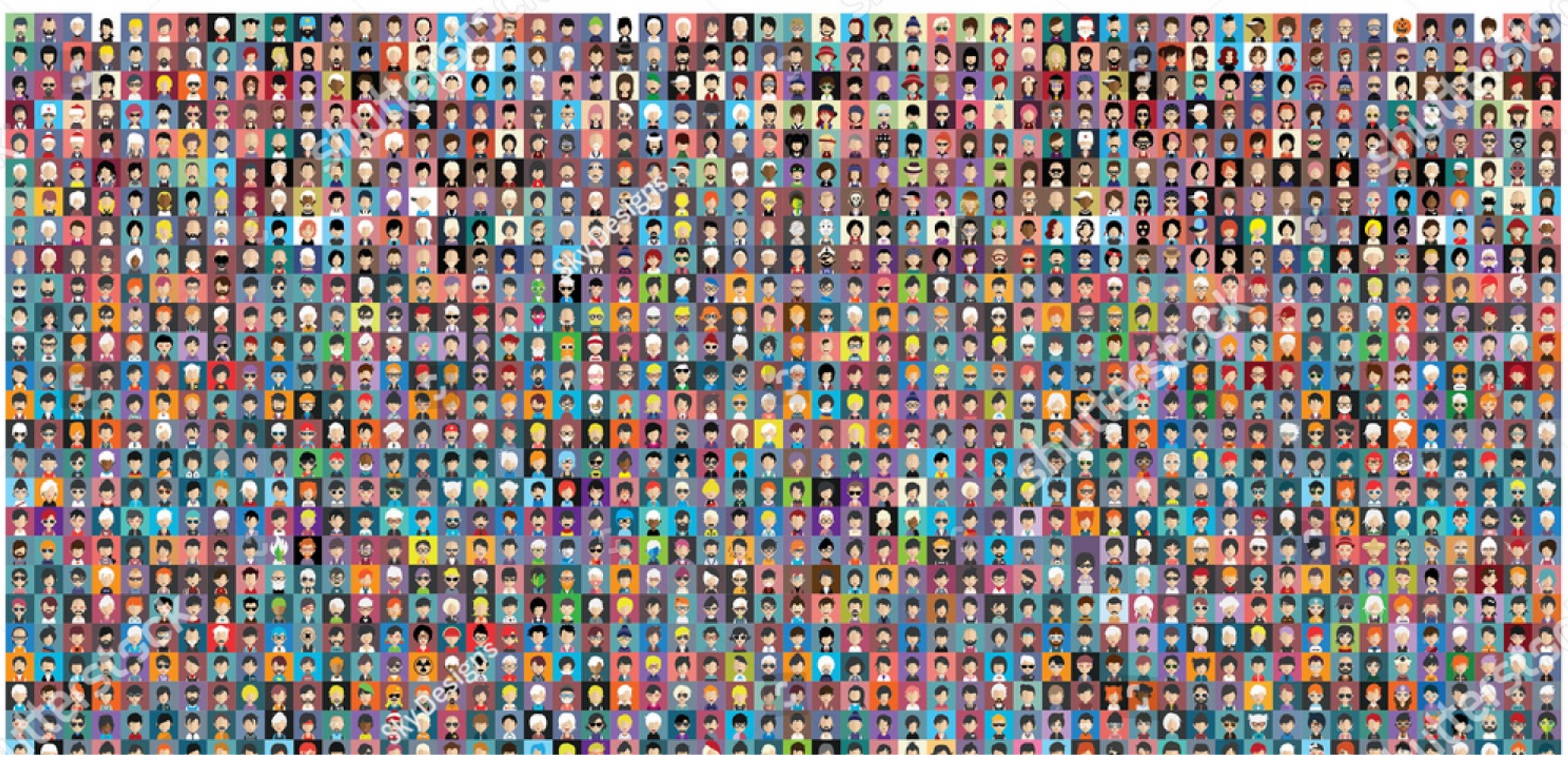
Task: defining the problem

Example: Gender recognition

male

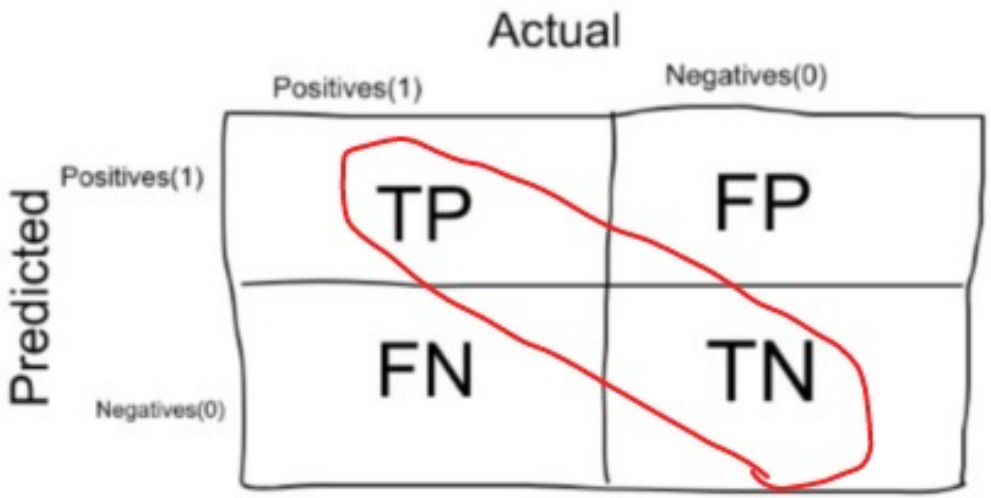
female

Dataset (Experience)



Performance

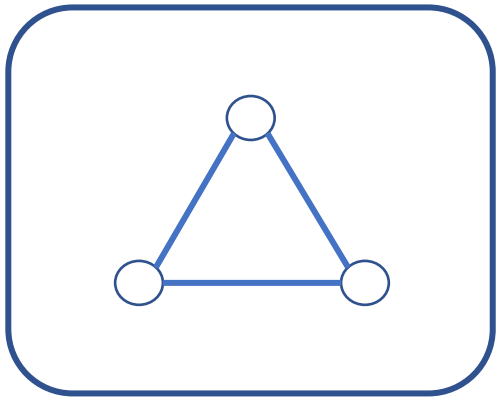
Classification Task



$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN}$$

Training

updating model…

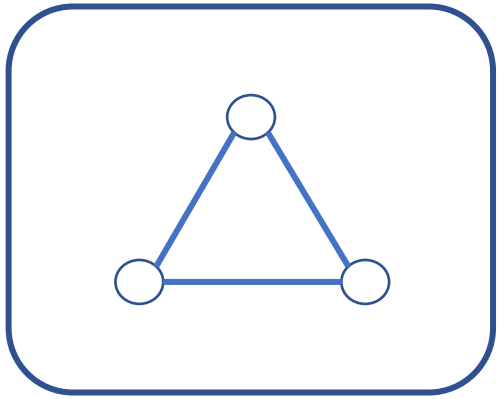


updating model…

Testing: (probability)

?

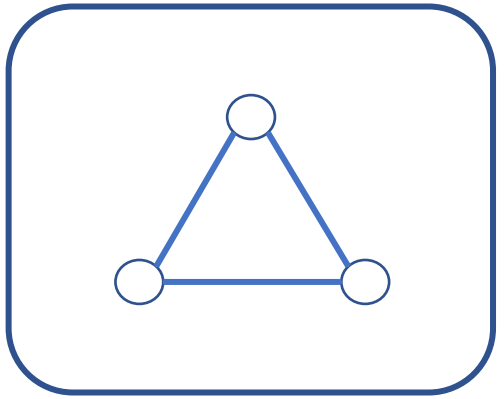
This is a female
(98% sure)



Testing (Performance)

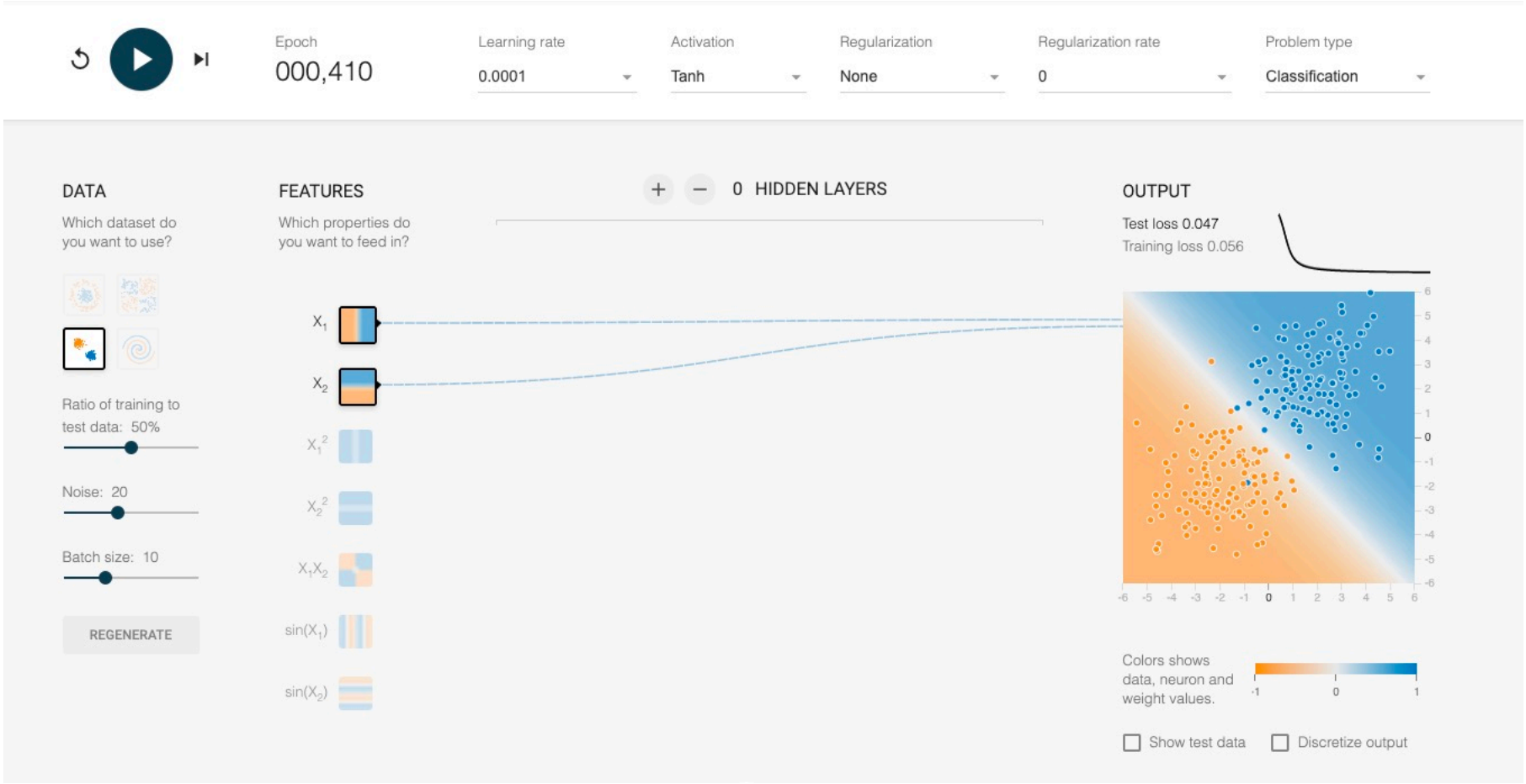
?

This is a male
(70% sure)



A more detailed example…

https://playground.tensorflow.org/



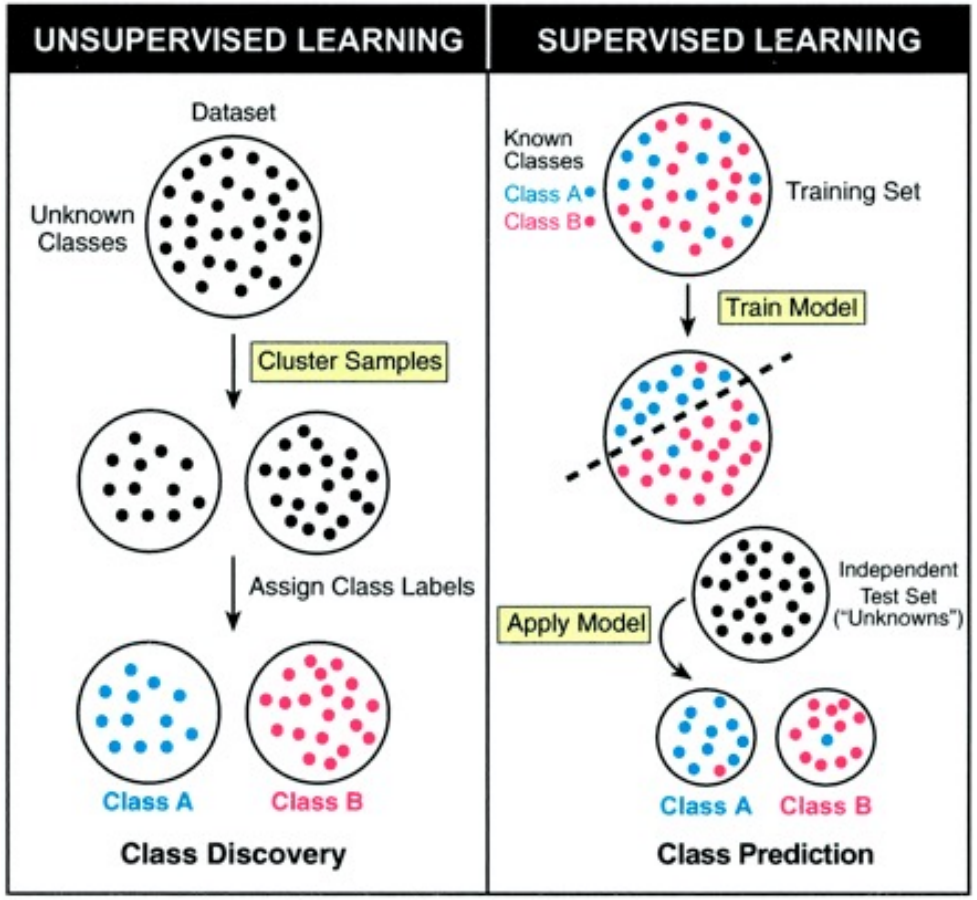
Learning main types

Supervised

Data is provided (through a "master") as pairs of cases-labels, indicating whether or not they belong to the type of association to be learned (classification, regression). Practical case: prediction

Unsupervised

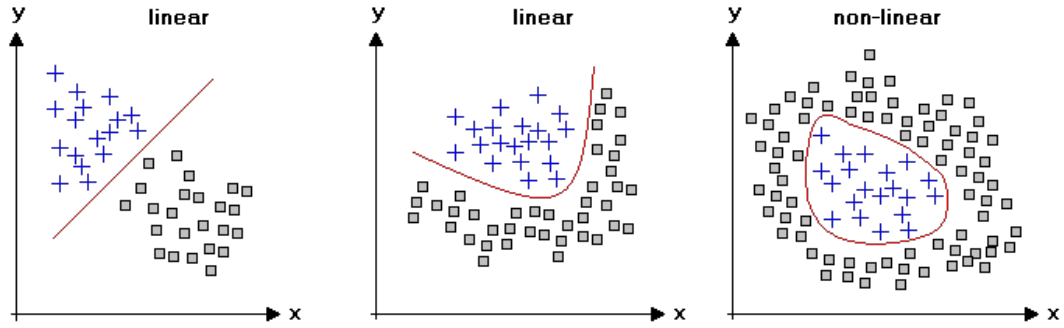
The algorithm must use other funds to obtain a "feedback" that tells him whether he does it well or not, since he does not have a "master" (clustering). Practical case: segmentation



Task types

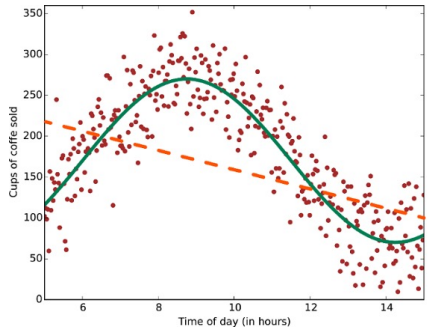
Classification

Classification algorithms are used when the desired output is a discrete label.

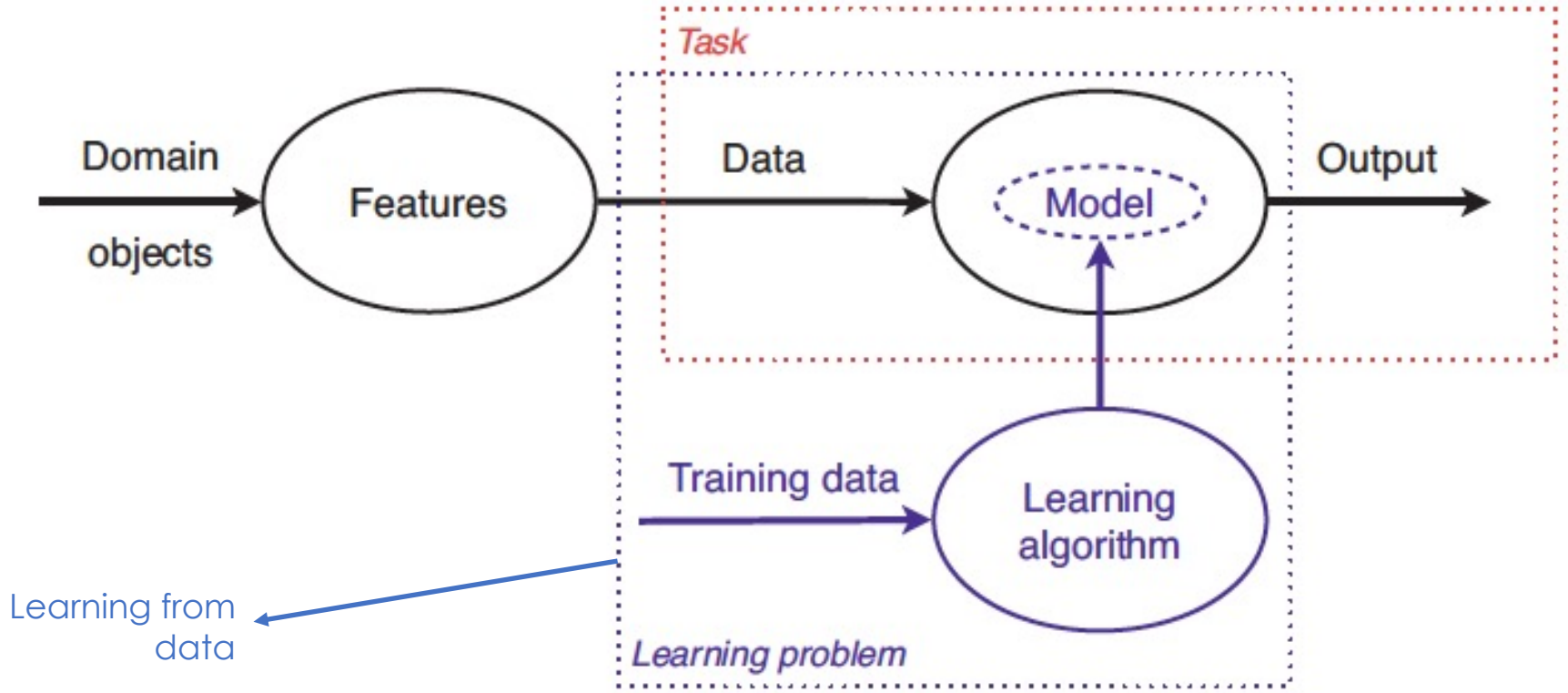


Regression

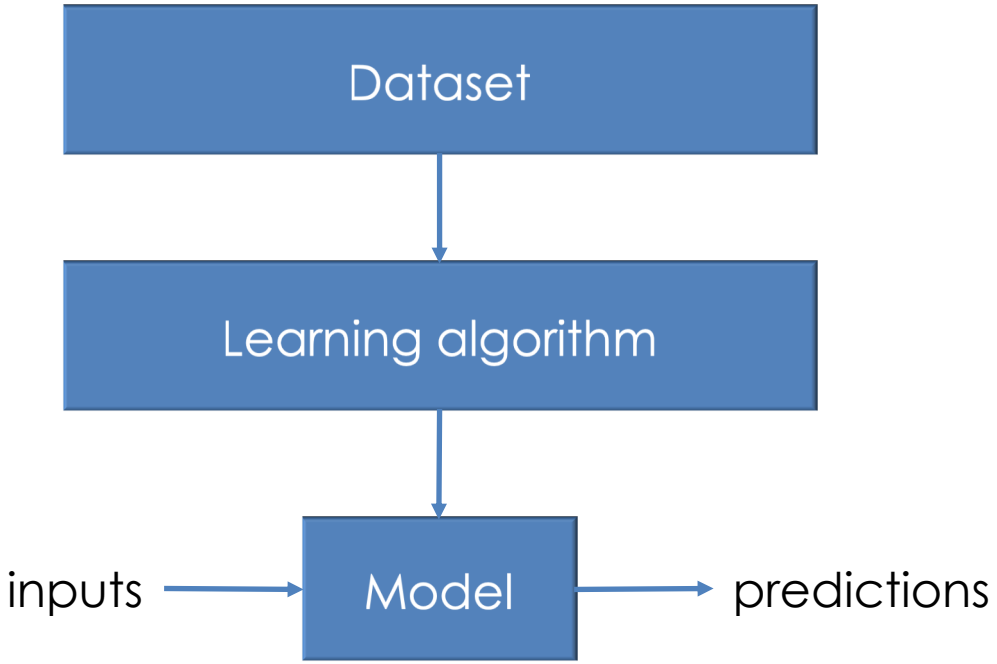
Approximate a continuous function (predicting outputs that are continuous)



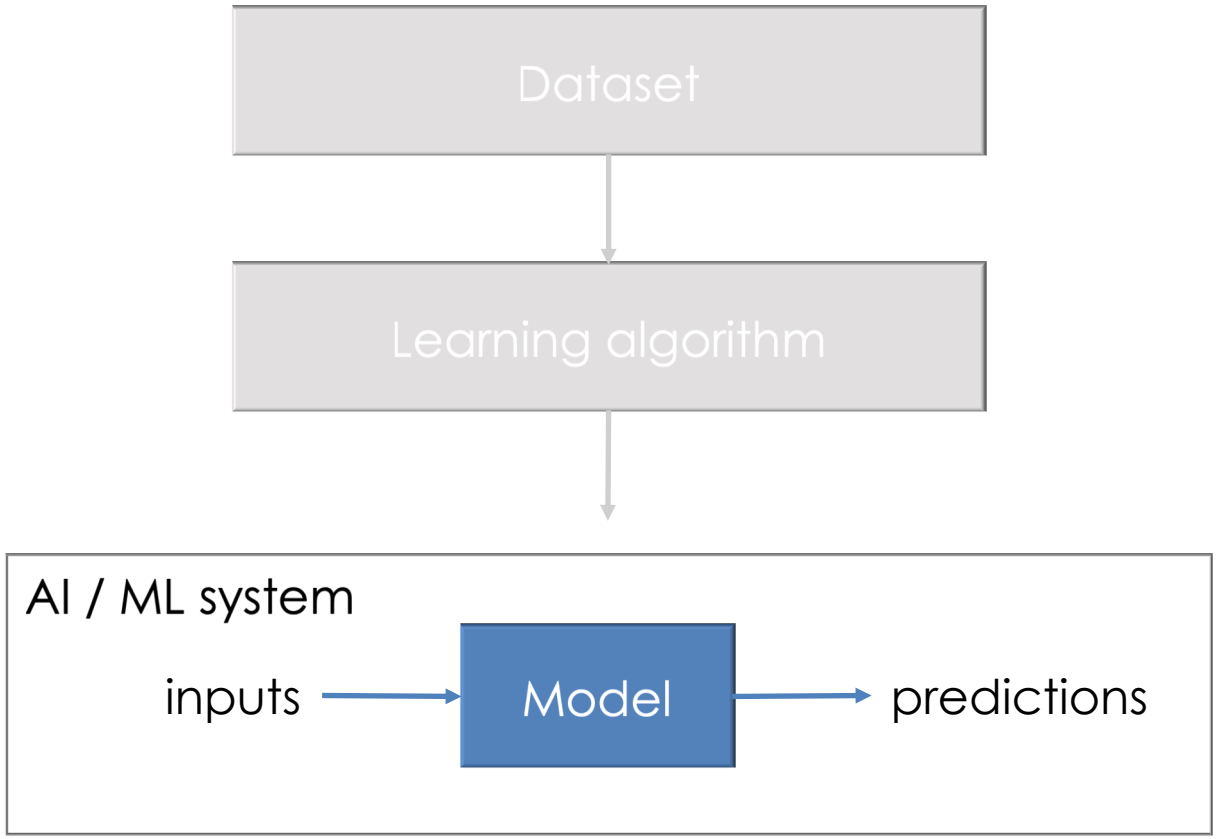
Computational learning: training



Model.train



Model.predict



Why ML ?

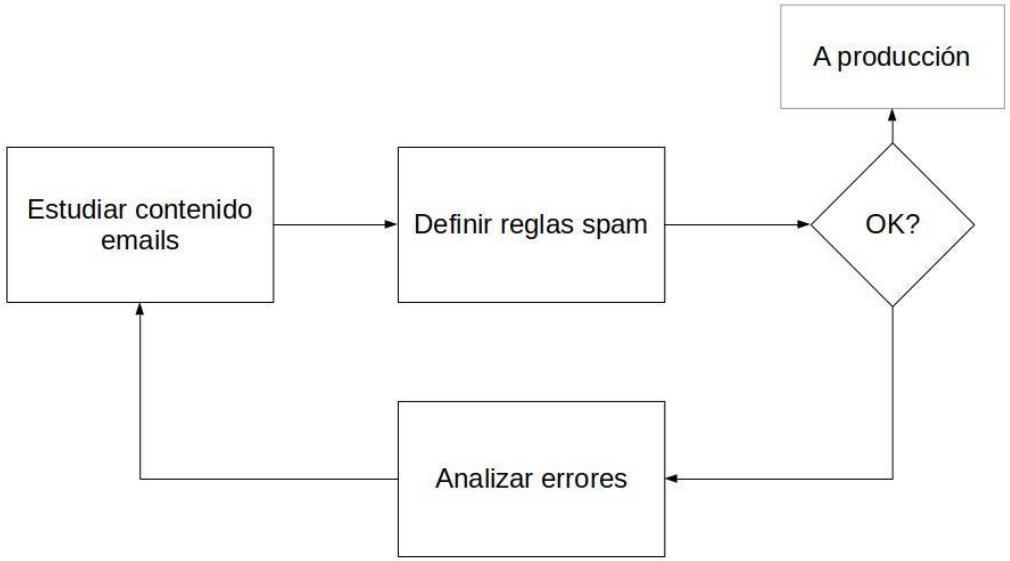
Imaginemos que queremos construir un filtro de spam.

Como diseñaríais un programa que detectase si un mail es spam o no?

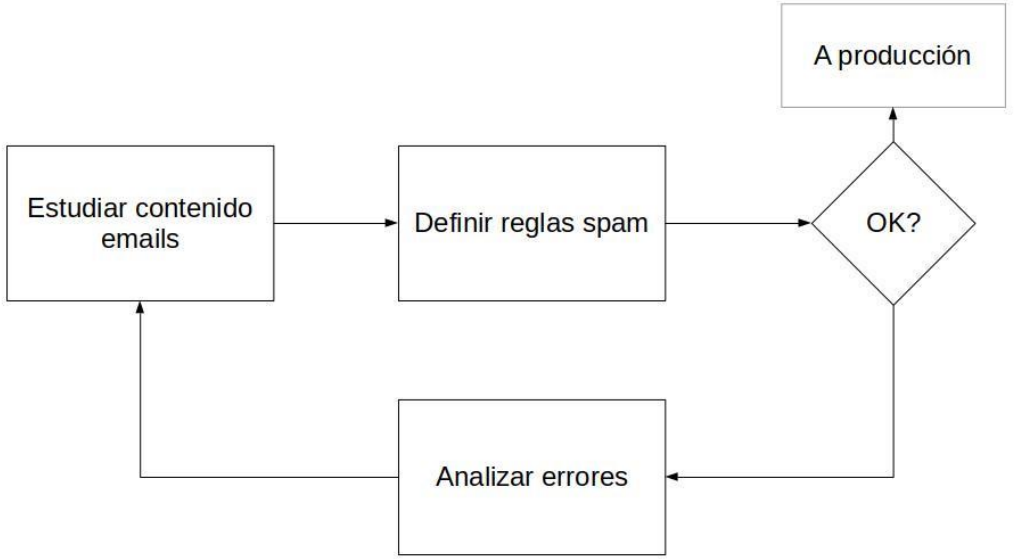
Why ML ?

Aproximación tradicional…

- Estudiar contenido de los emails
  - Definir reglas en base a expresiones regulares
- Crear una lista negra de ips
- Crear una lista negra de ”spammers”



Why ML ?



Pero...

- Los spammers se pueden adaptar...
- ..las reglas cambiar...
- ...obligarnos a iterar sobre los distintos pasos del sistema montado...
- ...con la contra de que, aparte de consumir mucho tiempo, el sistema se vuelve inmantenible.

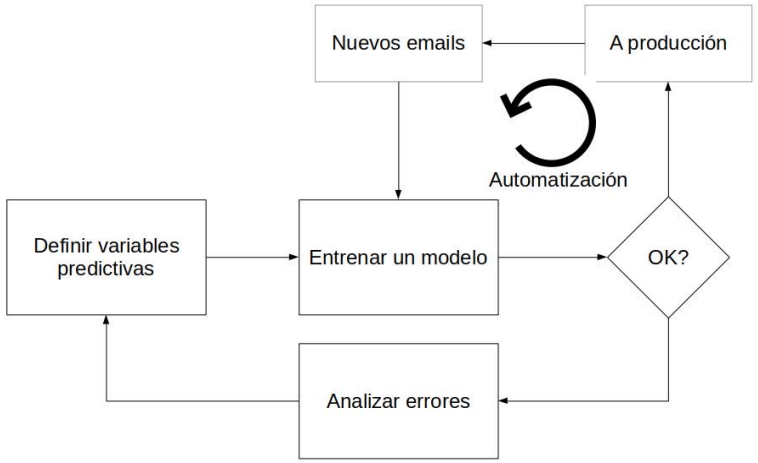

[sad developer noise]

The ML way

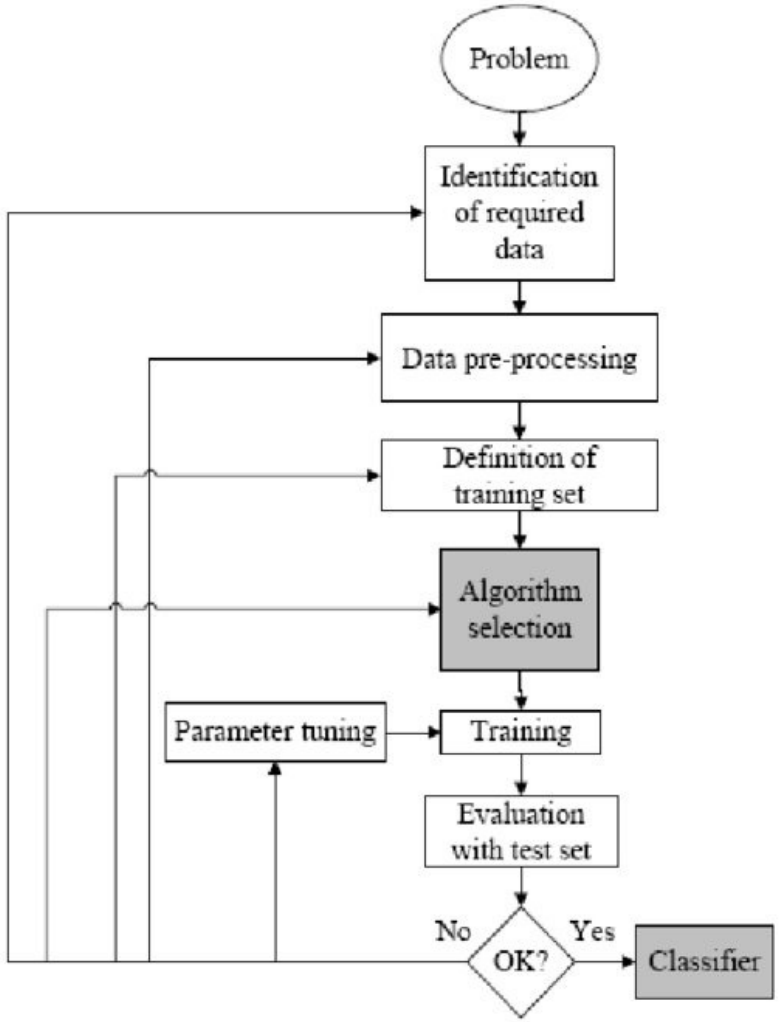
Si en vez de intentar definir las reglas a mano cada vez que queramos modificar nuestro sistema de detección de spam utilizamos un clasificador, podemos redefinir el problema como

- Estudiar el problema desde el punto de vista de que **variables** me afectan a que sea spam
  - por ejemplo patrones de palabras frecuentes en emails normales o spam (bag of words, tf-idf)

- Entrenar un clasificador sobre las variables seleccionadas:
  - e.g. naive-bayes, logistic regression...



How to develop a ML solution



Framing the problem:

From Real world problem to ML problem

Getting the data

Do I have data to solve this problem?

Explore the data

Is the data representative?

Train and select the model

Model selection and feature engineering

Deploy

How it will be consumed?

When ML ?

Los sistemas de ML son buenos para:

- Problemas que requieren mucho trabajo manual o largas listas de reglas
- Entornos que evolucionan
- Conseguir insights en datos complejos

Real world applications

CARS

Self-driving cars will be in place by the end of the next decade, 2030.

HEALTH

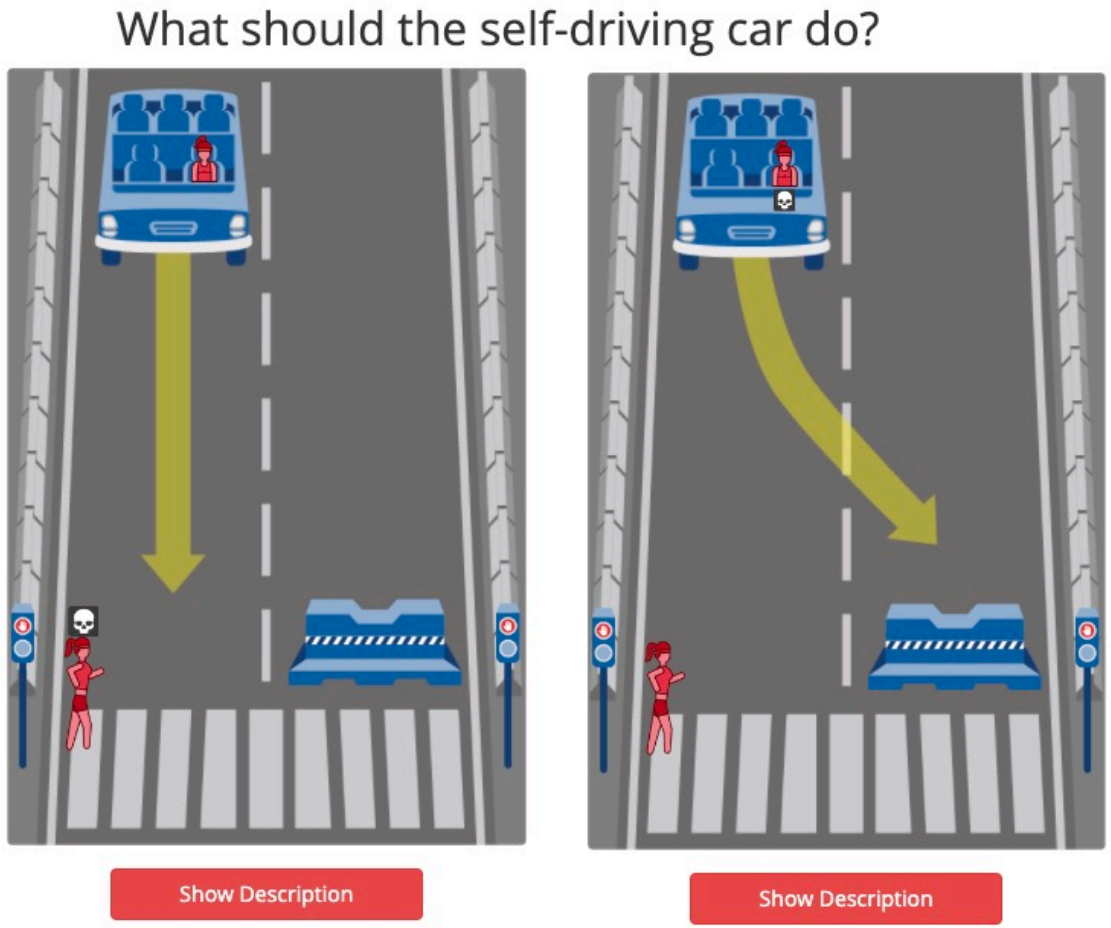
Genetic research, medical assistance, and personalized treatments.

ROBOTICS

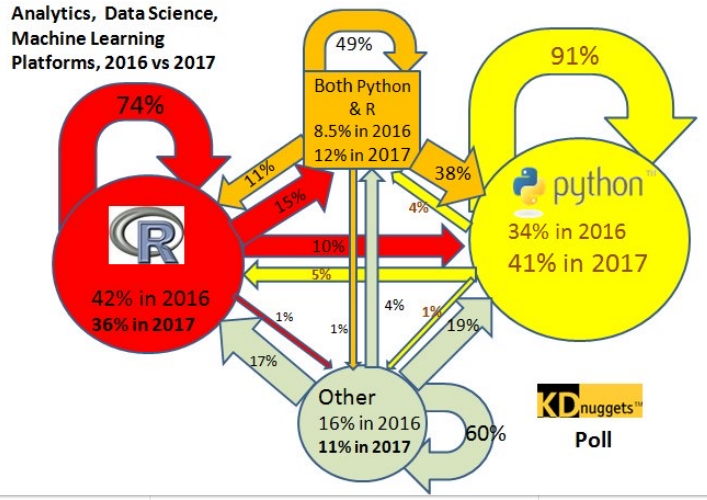
Autonomous robots to assist people.

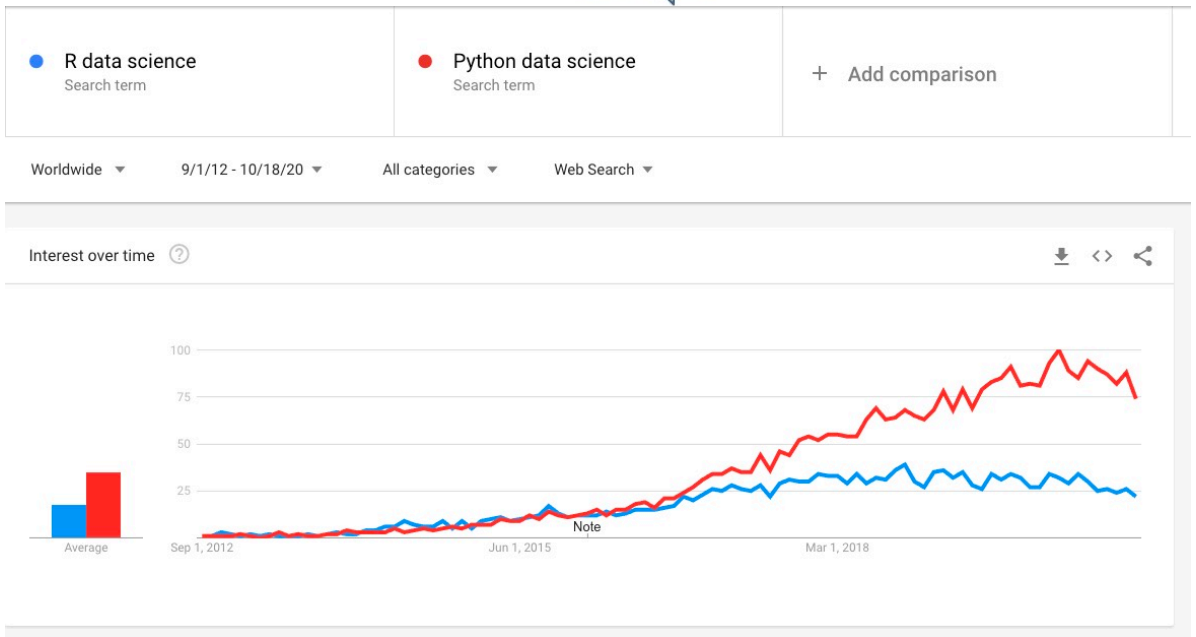
Ethics!

1. The purpose specification
2. Limitations of data: bias
3. Limitations of data: privacy
4. Autonomous?
5. Education
6. People: Explainable AI



Python for data science and ML

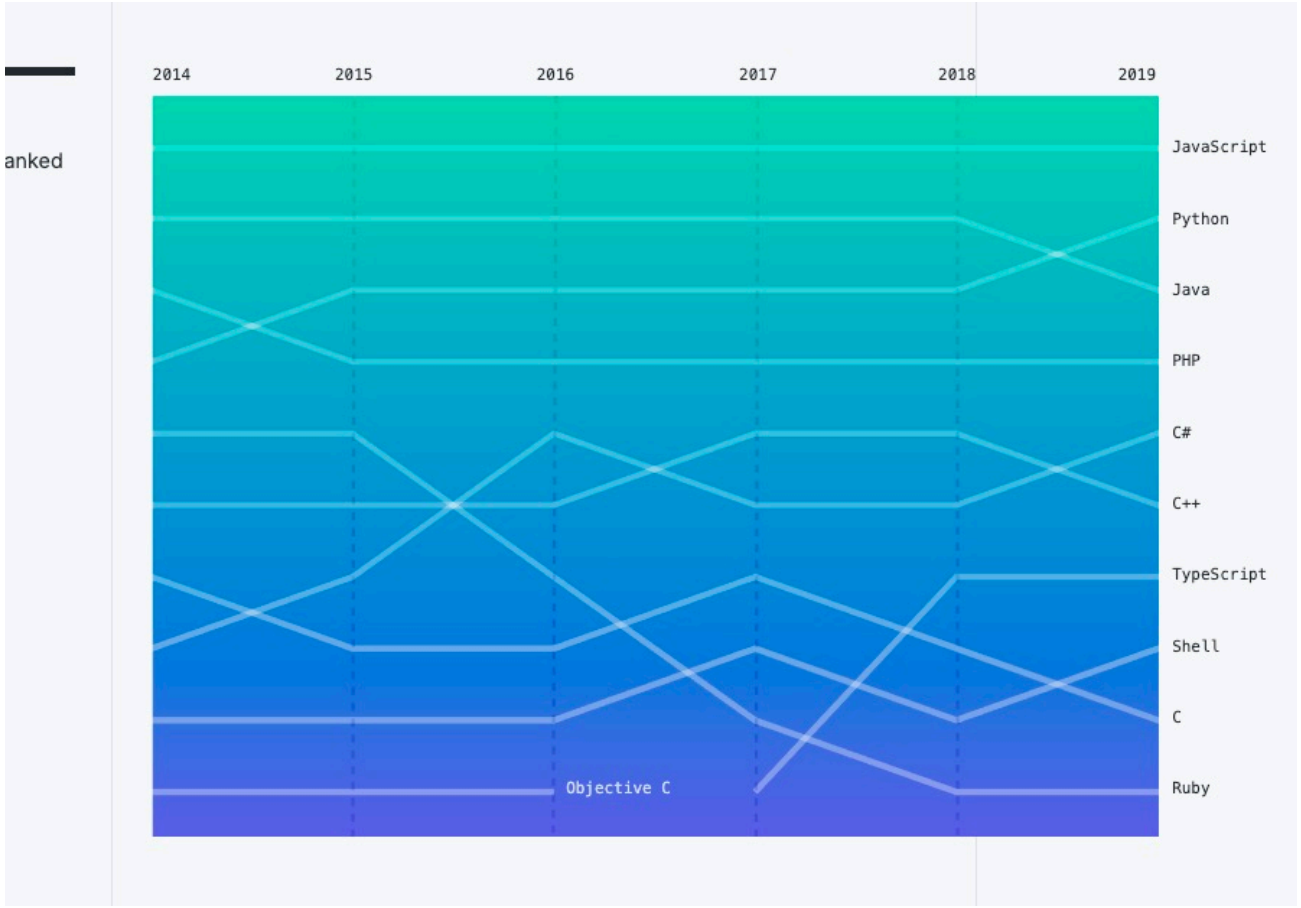




- Python es un lenguaje de programación interpretado cuya filosofía hace hincapié en una sintaxis que favorezca un código legible.

- Como cualquier otra matería prima, el dato debe pasar por una serie de etápas y transformaciones desde que es "extraído" (APIs, scrapping, encuestas...) hasta que es "consumido" (aplicaciones, sistemas de recomendación, ...)

Python for data science and ML

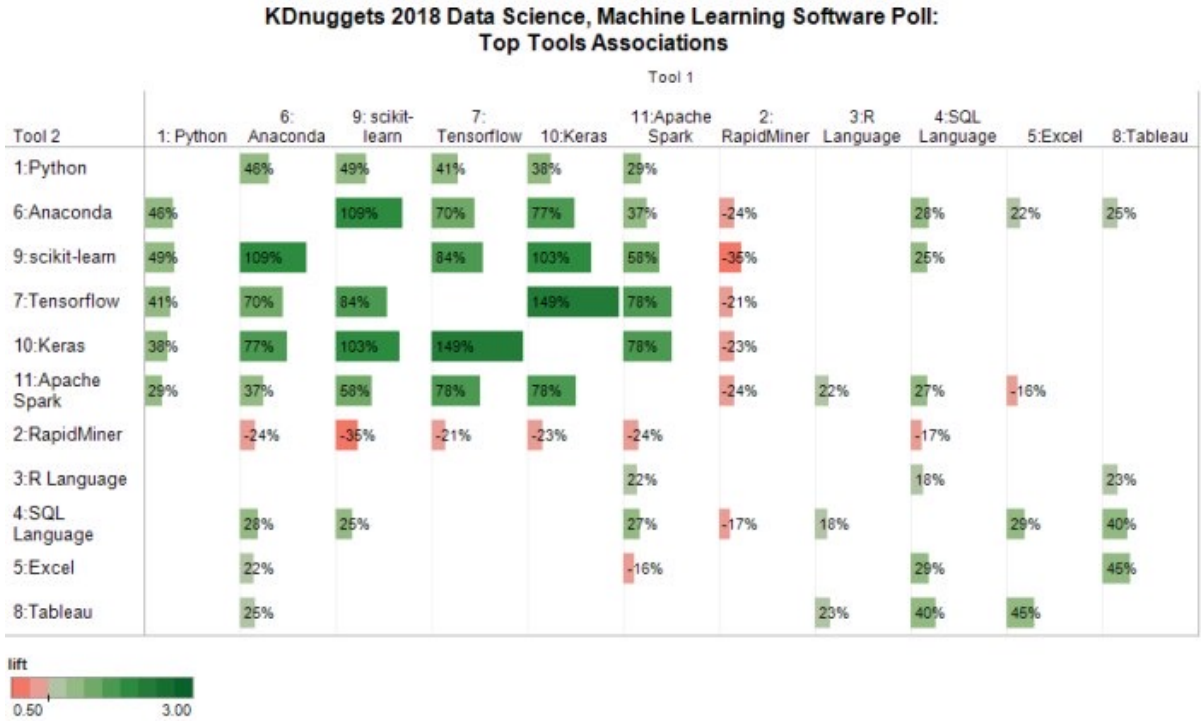


- Although GitHub has traditionally been home to software developers, the world's code is evolving. **Behind Python's growth is a speedily-expanding community of data science professionals and hobbyists—and the tools and frameworks they use every day**. These include the many core data science packages powered by Python that are both lowering the barriers to data science work and proving foundational to projects in academia and companies alike.

- Beyond Python, repositories with **topics** like **"deep learning", "natural language processing", and "machine learning"** have become more popular over the years with growing communities focused on data science. Among the most popular (based on star counts) public repositories labelled with the topic, over half of them **are built on numpy, and many of them depend on scipy, scikit-learn, and TensorFlow**. We've also seen non-code contributions from the data science field, including academic papers.

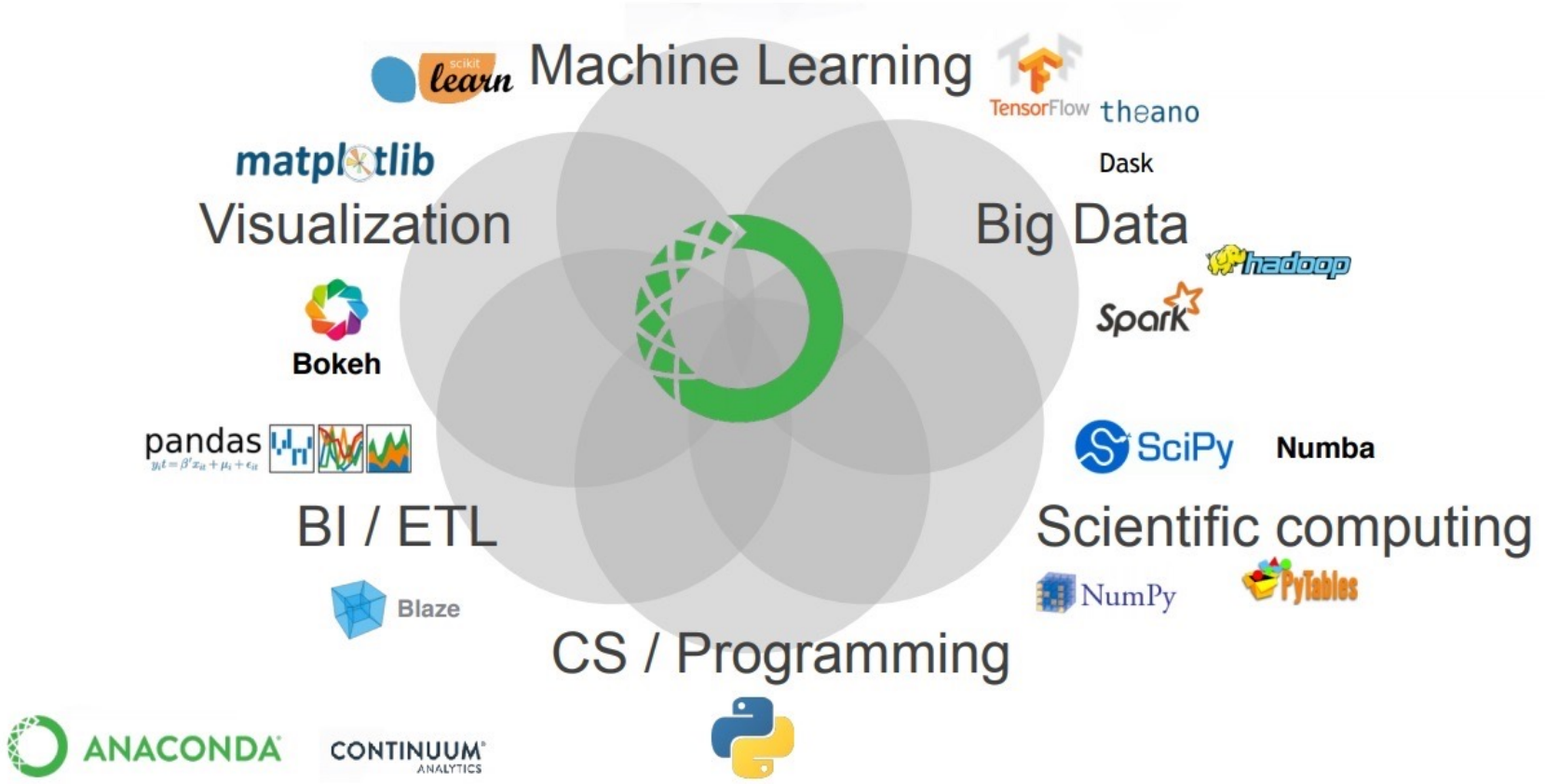
source

Python for data science and ML



6 primary tools that together make the modern open source data science ecosystem: **Python, Anaconda, scikit-learn, Tensorflow, Keras, and Apache Spark.**

Python for data science and ML



To do…

- Create github account:
  - where code happens
  - Own portfolio
  - Code repository

- Install data science working environment:
  - Anaconda installation (easiest way: bundled and curated list of libraries for data science and python development)