UNIVERSITÀ
DEGLI STUDI
DI TRIESTE

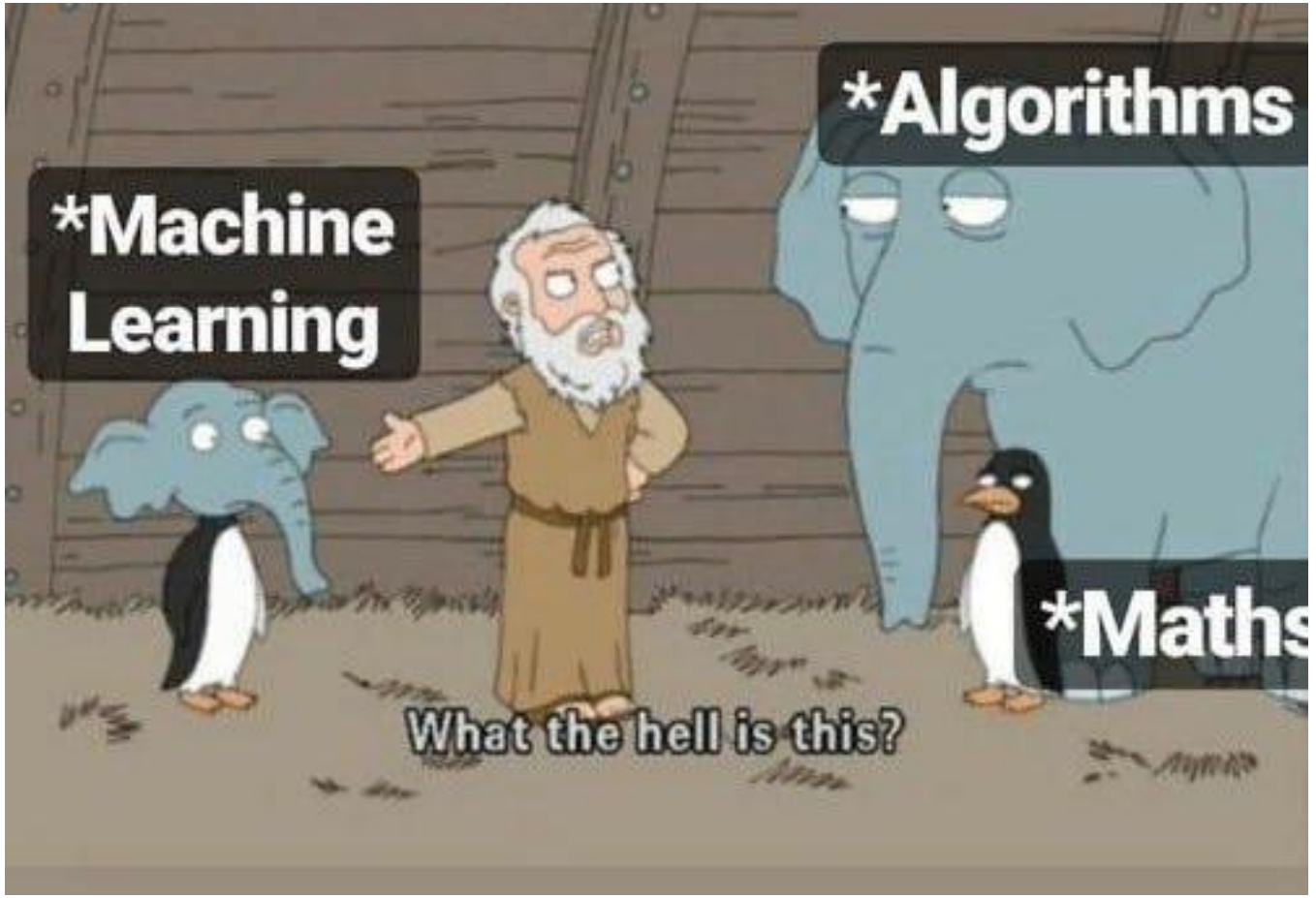
Statistics

Introduction to Machine Learning

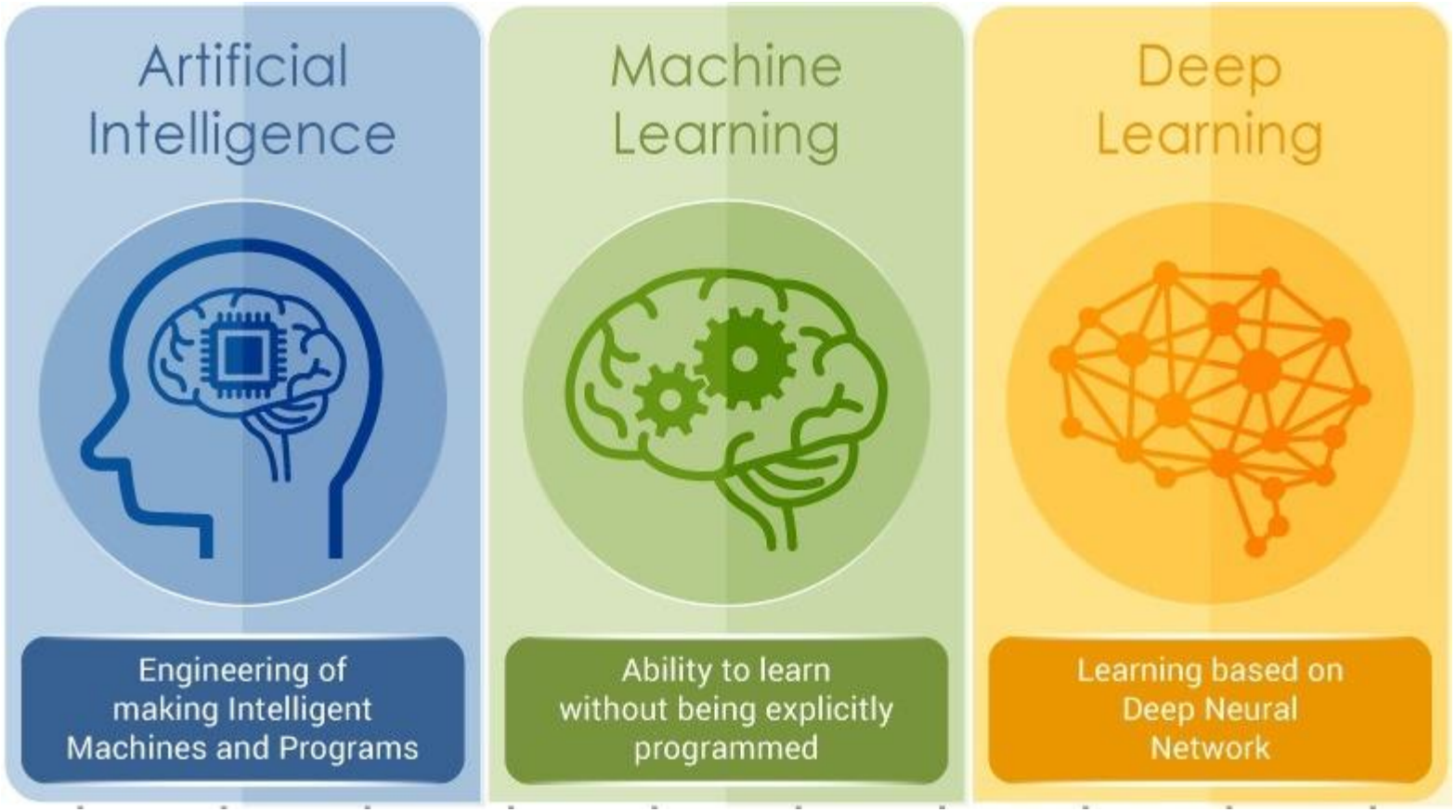
Luca Pennella

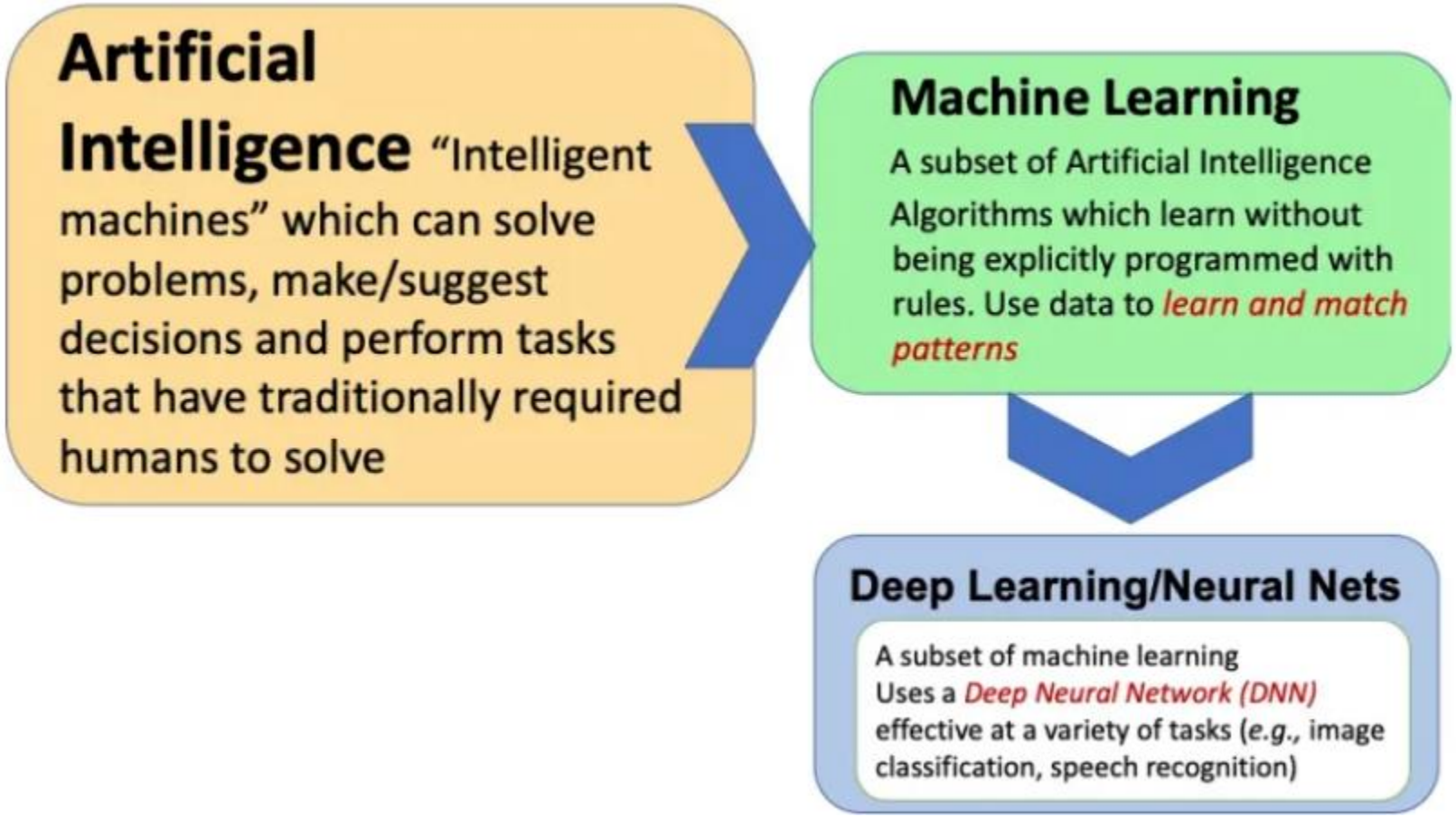
September 17th, 2024

What is machine learning?



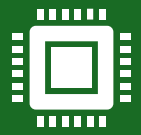
Machine learning is part of artificial intelligence that deals with building methods that '**learn**', that is, methods that use data to improve performance on some set of tasks (e.g. image recognition). It is seen as a part of artificial intelligence.




Artificial Intelligence "Intelligent machines" which can solve problems, make/suggest decisions and perform tasks that have traditionally required humans to solve
Machine Learning
A subset of Artificial Intelligence
Algorithms which learn without being explicitly programmed with rules. Use data to learn and match patterns
Deep Learning/Neural Nets
A subset of machine learning
Uses a Deep Neural Network (DNN)
effective at a variety of tasks (e.g., image classification, speech recognition)

What is machine learning?

Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn from data without being explicitly programmed.

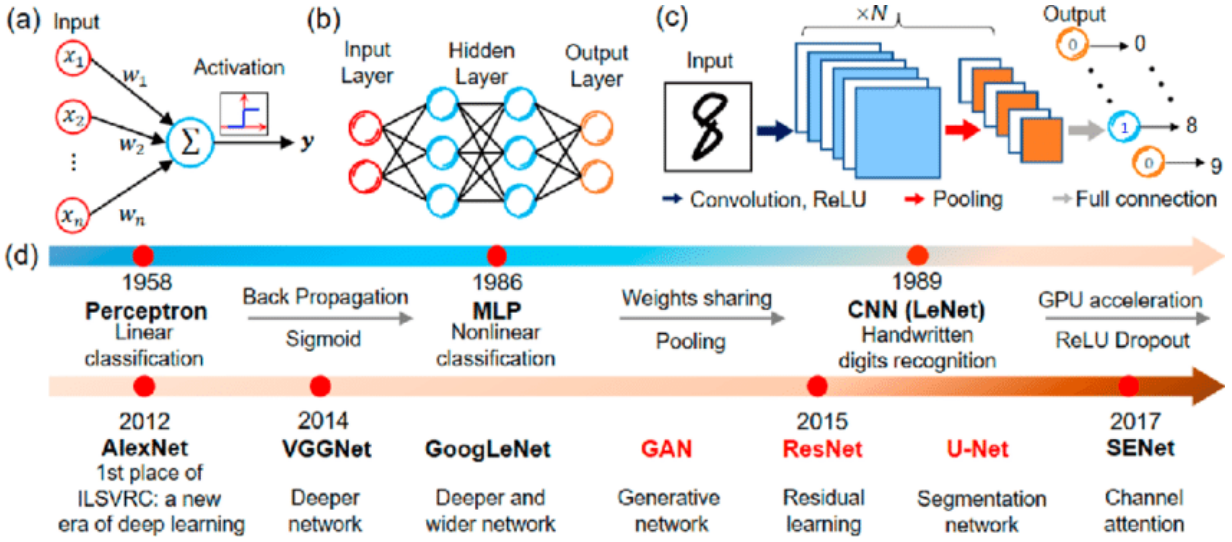


Main difference with traditional Computer Science: learn a program that deals with data (ML) vs have a coded program that run the data (CS). *Optimization not logic*.

What made possible the ML success?

- Data abundance

- Algorithmic development



- Better computers

Data abundance: images

IMAGENET

14,197,122 images, 21841 synsets indexed

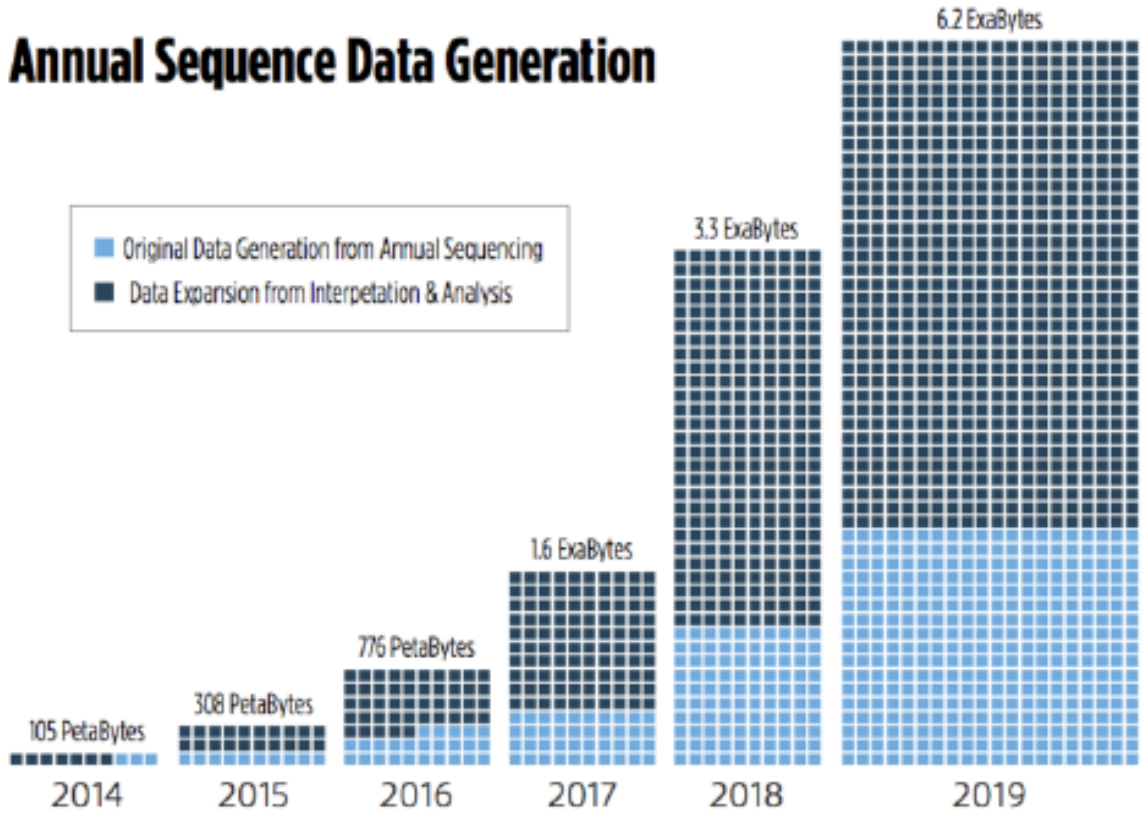
Home Download Challenges About

Not logged in. Login | Signup

An Update to the ImageNet Website and Dataset

Data abundance: genetics

Easier and simpler sequencing techniques

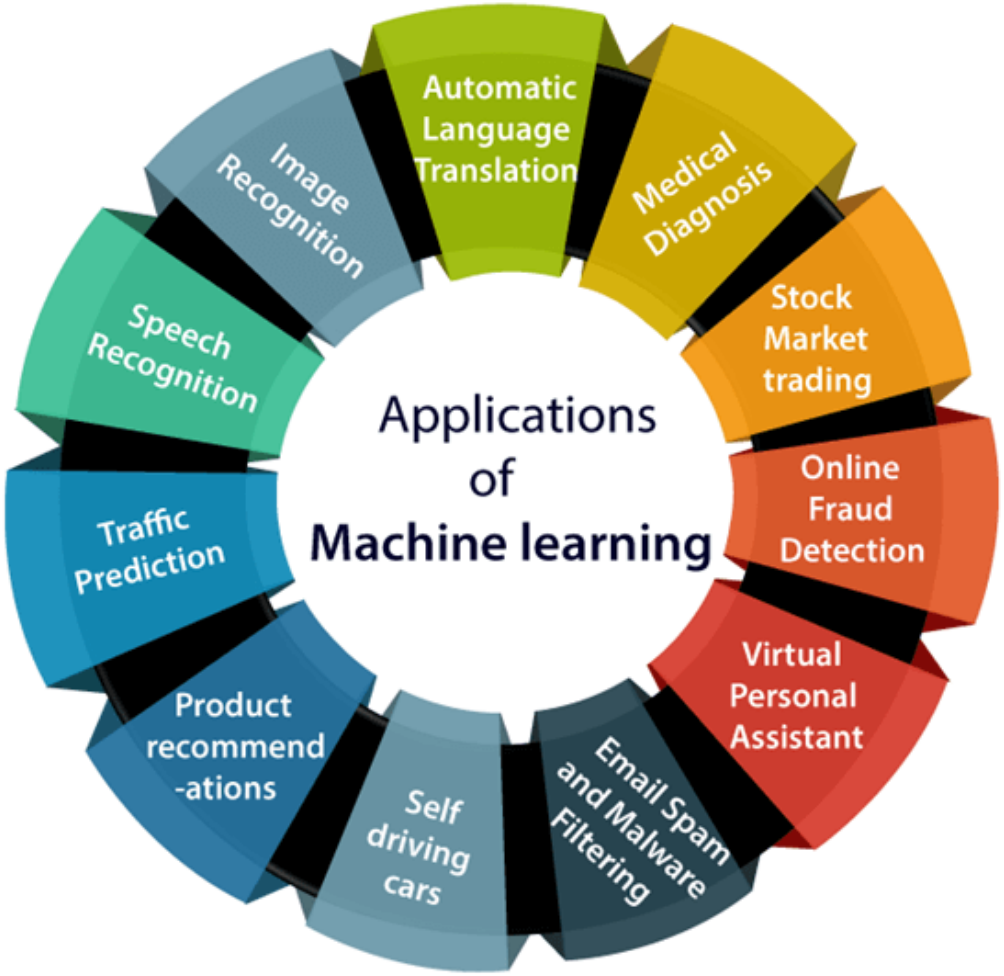


Data abundance: chemistry

ChemSpider

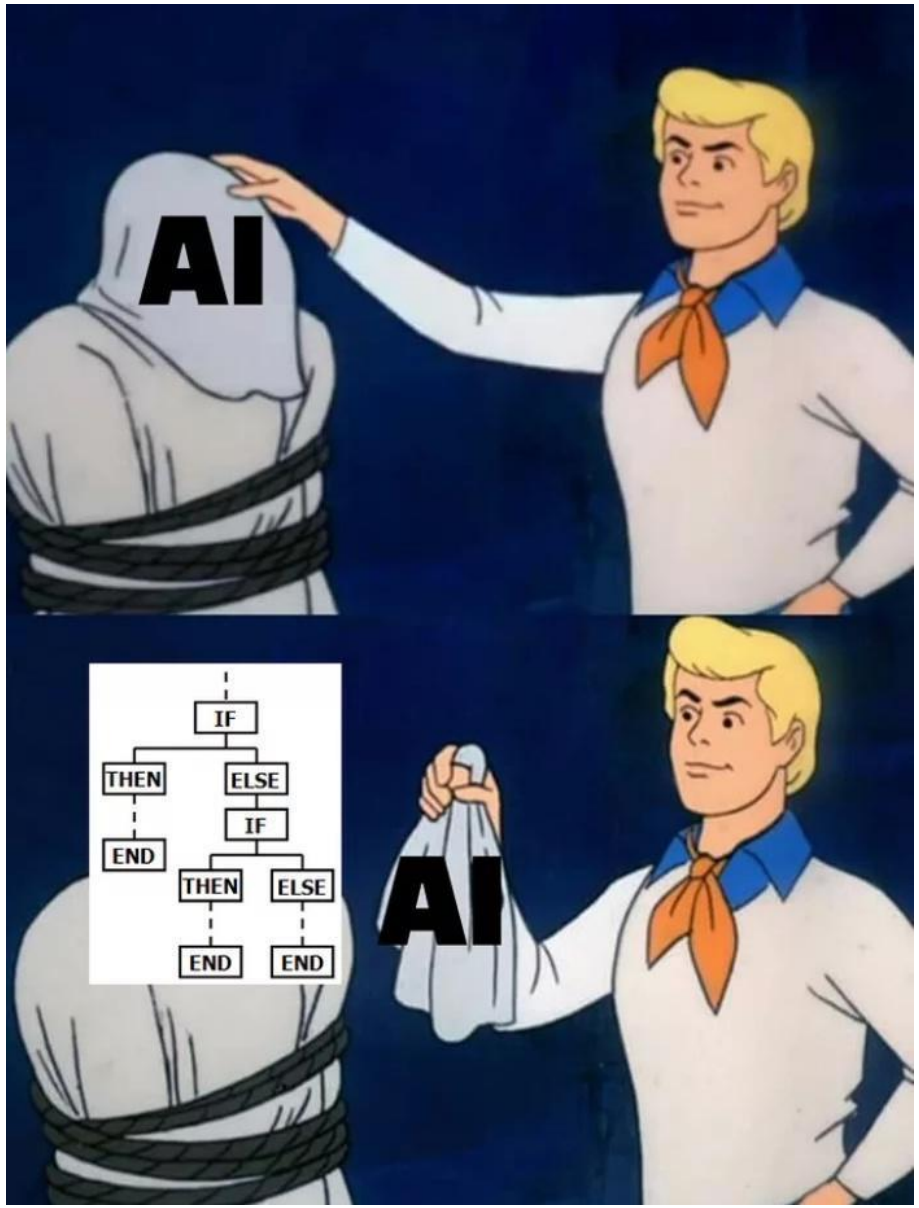
Content	
Description	A chemical structure database providing fast access to over 100 million structures, properties and associated information.
Contact	
Research center	Raleigh, North Carolina, United States
Laboratory	Antony J. Williams Royal Society of Chemistry[1]
Access	
Website	www.chemspider.com
Tools	
Standalone	https://itunes.apple.com/us/app/chemspider/id458878661
Miscellaneous	

Many many applications

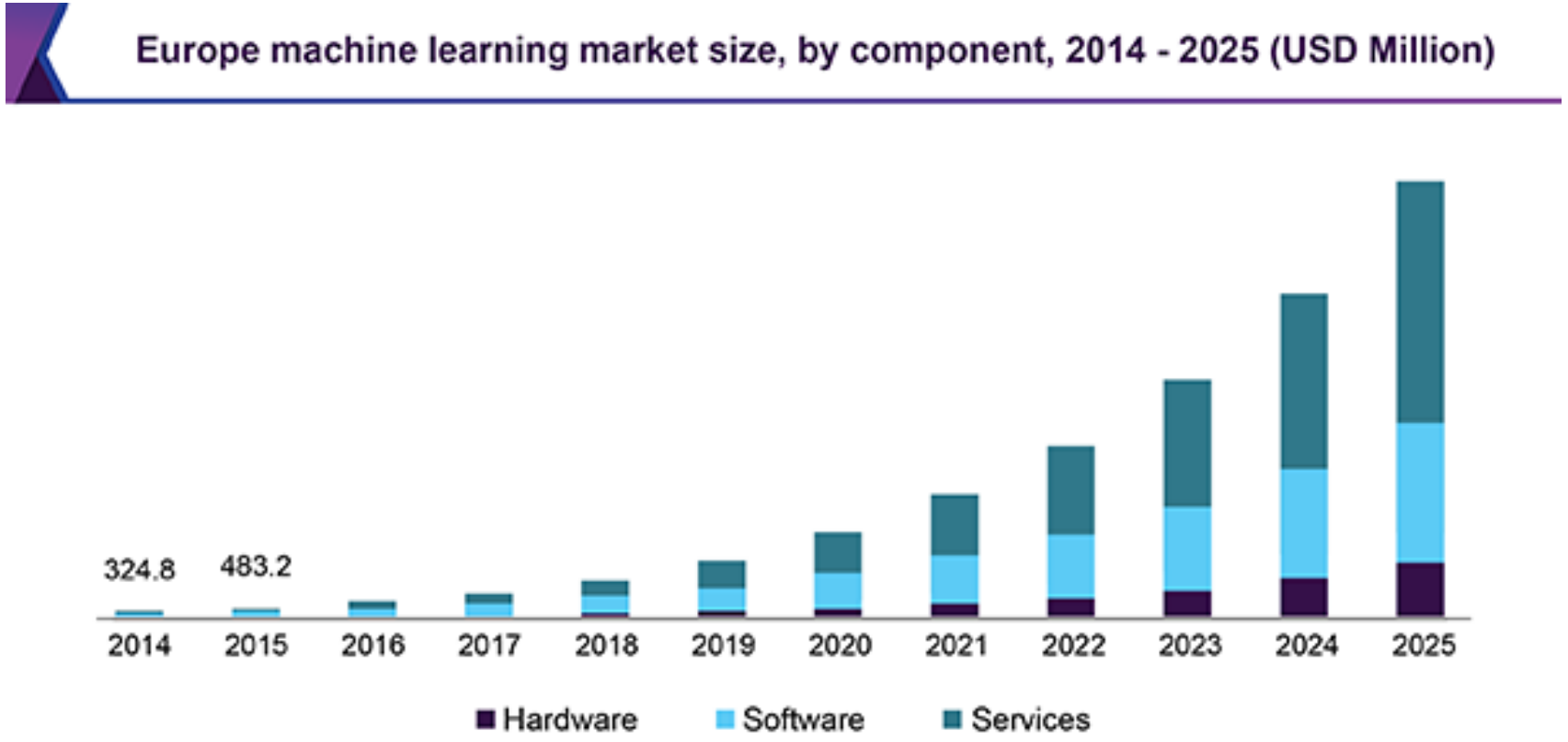


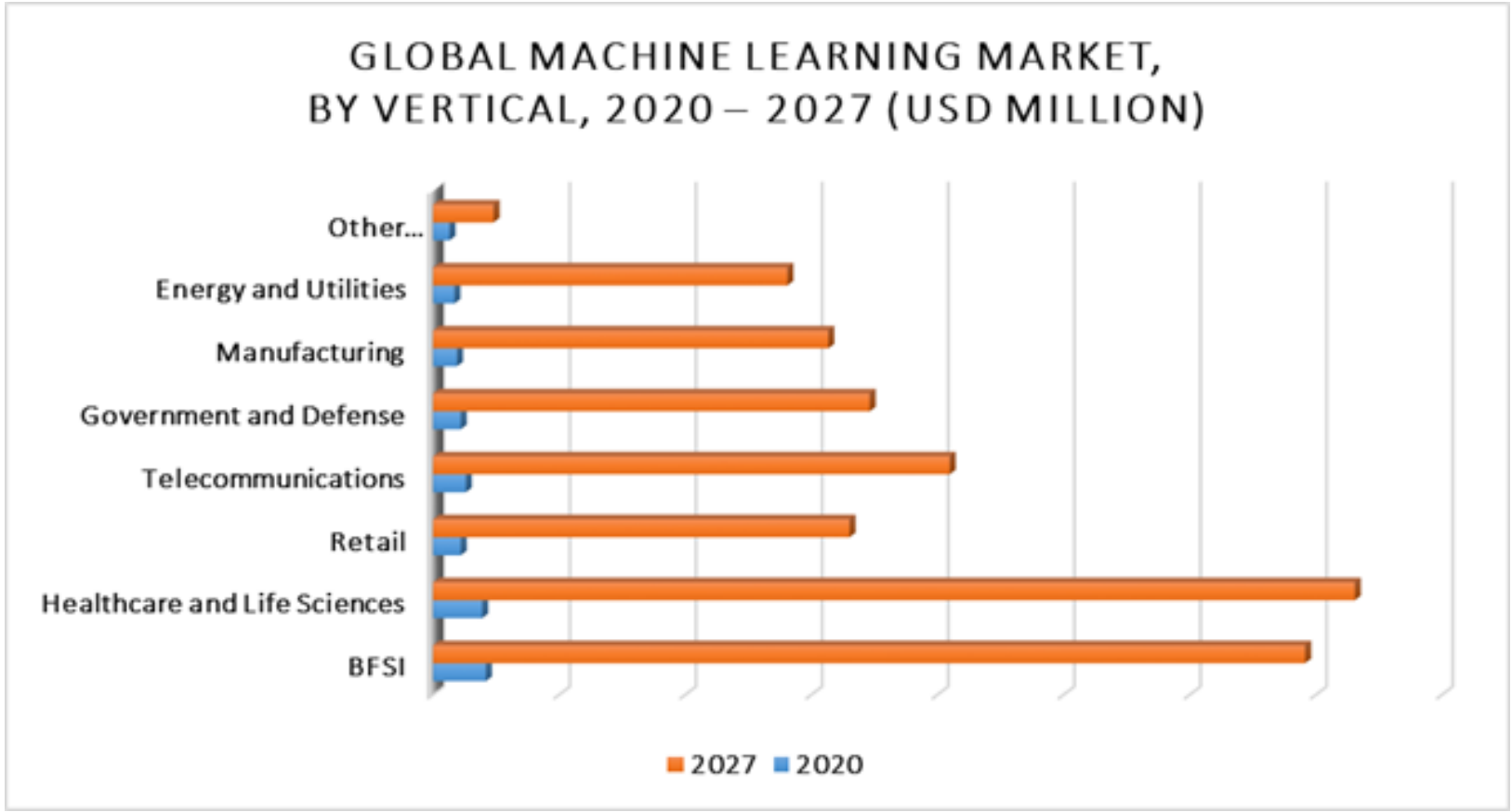
Why study machine learning?

+ APPLICATIONS



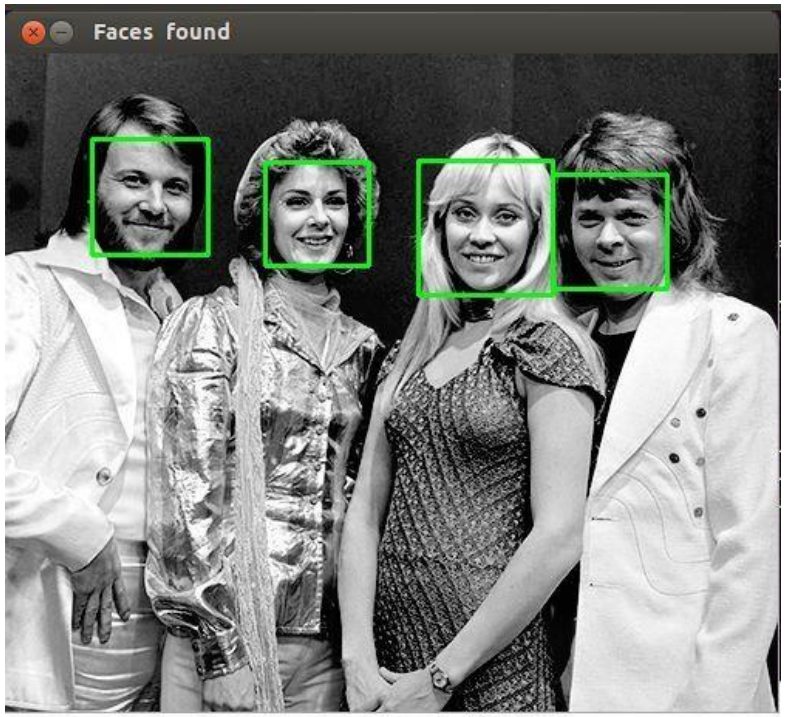
Market size Europe




GLOBAL MACHINE LEARNING MARKET,
BY VERTICAL, 2020 – 2027 (USD MILLION)
Other...
Energy and Utilities
Manufacturing
Government and Defense
Telecommunications
Retail
Healthcare and Life Sciences
BFSI
2027
2020

APPLICATIONS: some examples

Image processing: face recognition…



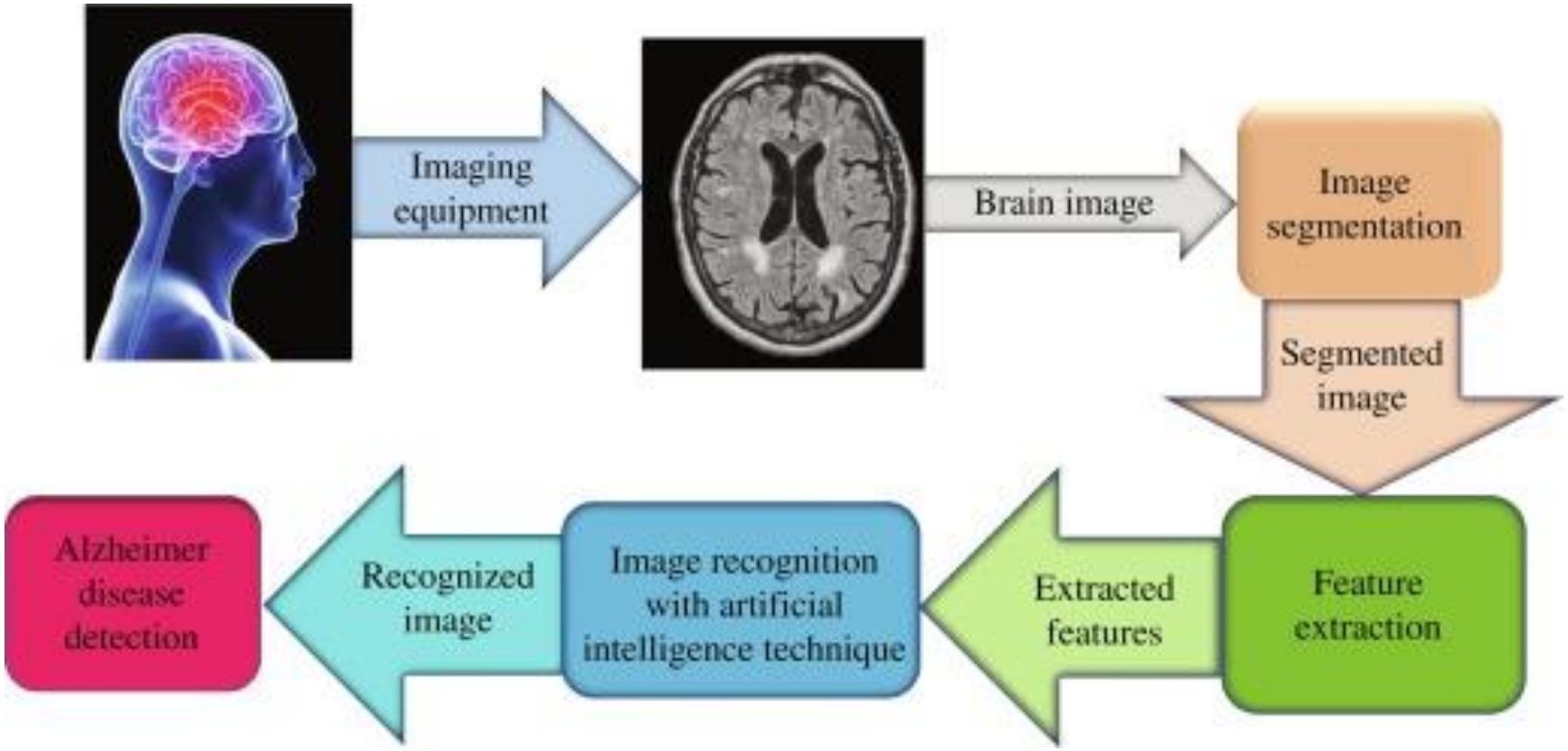
From automatic friend tag…

… to surveillance

Image processing: style changing

Medicine

Detection of Alzheimer disease through brain imaging



Mathematics

nature › articles › article

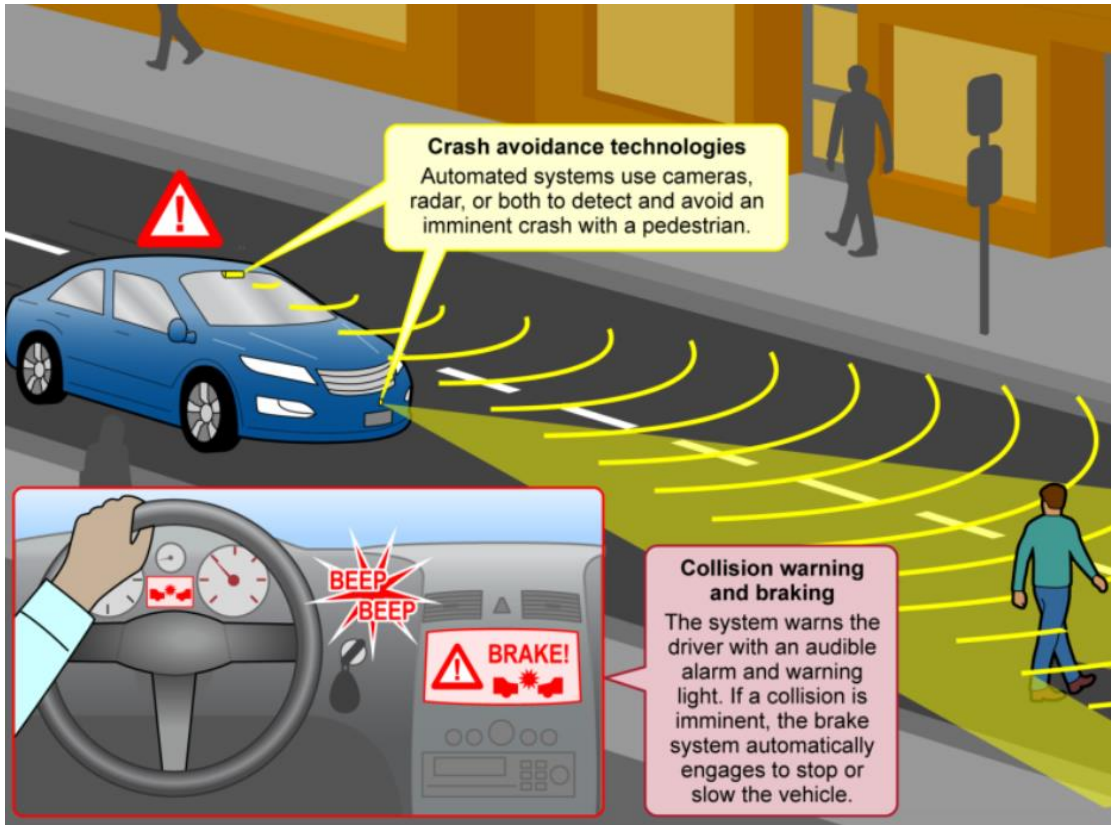
Article | Open Access | Published: 01 December 2021

Advancing mathematics by guiding human intuition with AI

Alex Davies, Petar Veličković, Lars Buesing, Sam Blackwell, Daniel Zheng, Nenad Tomašev, Richard Tanburn, Peter Battaglia, Charles Blundell, András Juhász, Marc Lackenby, Geordie Williamson, Demis Hassabis & Pushmeet Kohli

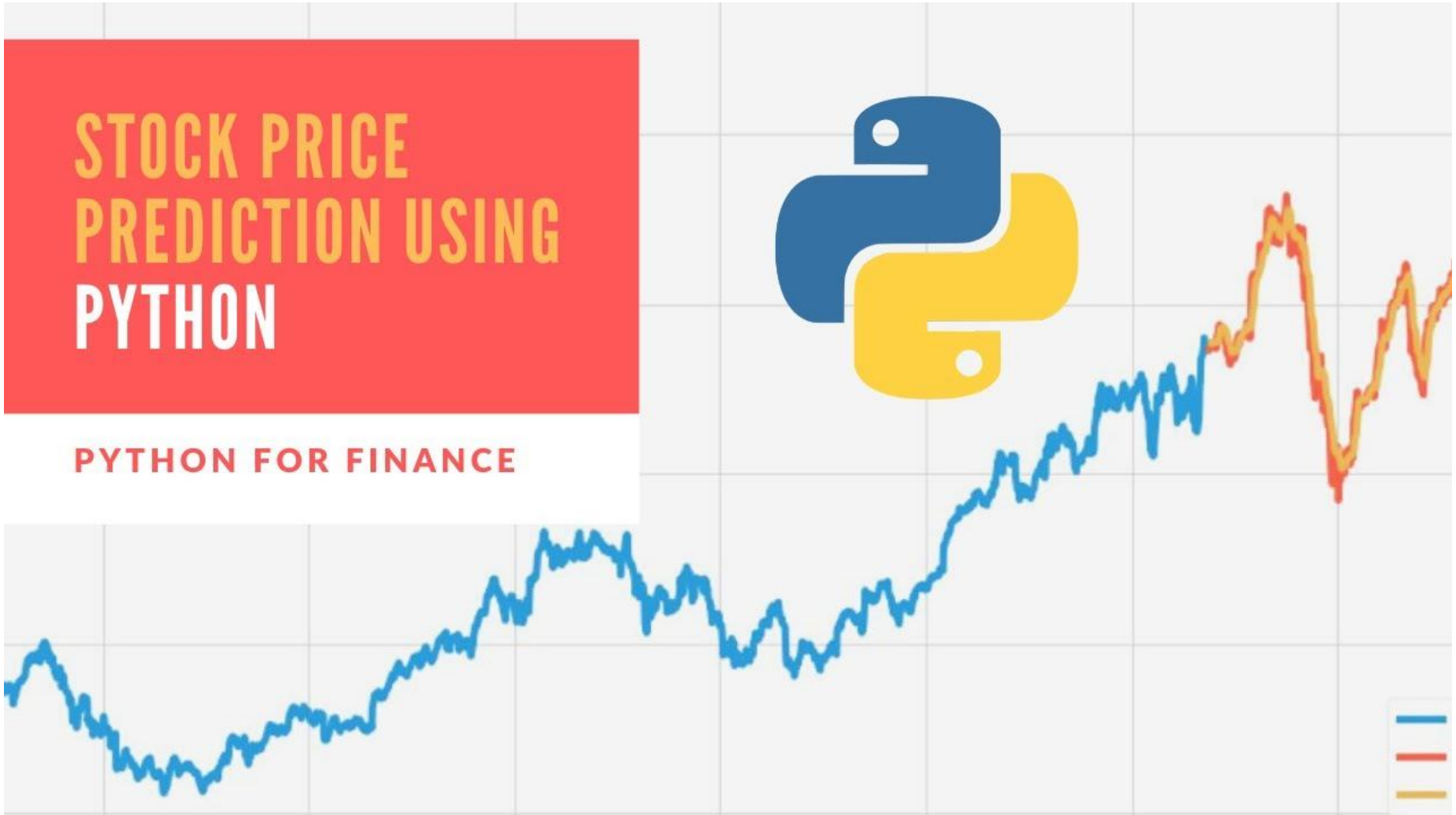
Nature **600**, 70–74 (2021) | Cite this article

Self driving cars

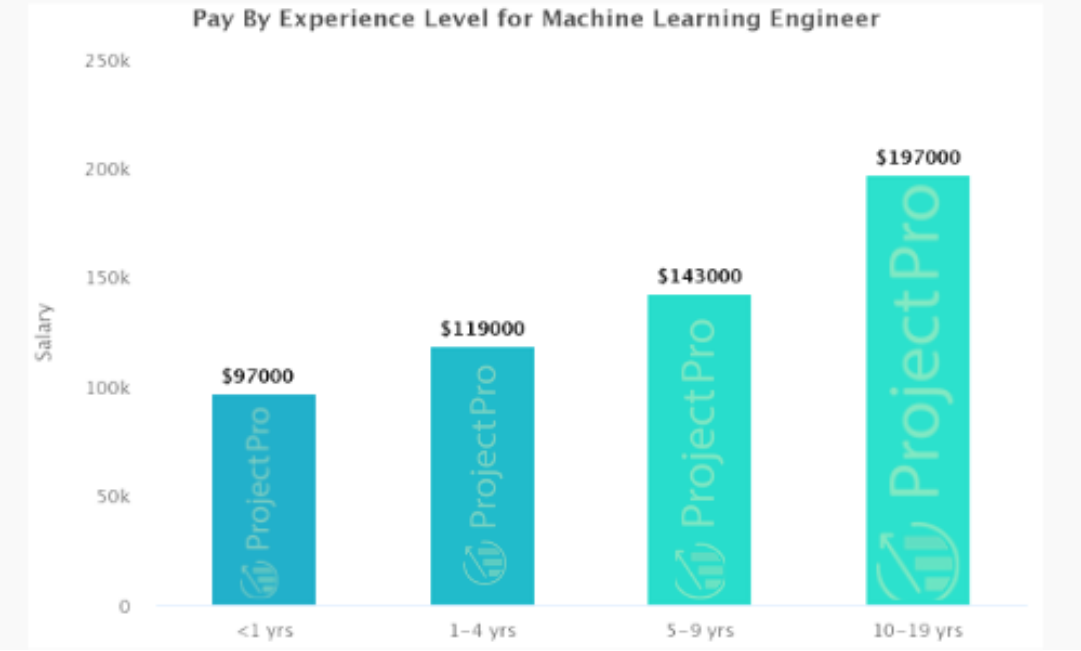


STOCK PRICE PREDICTION USING PYTHON

PYTHON FOR FINANCE



Also...

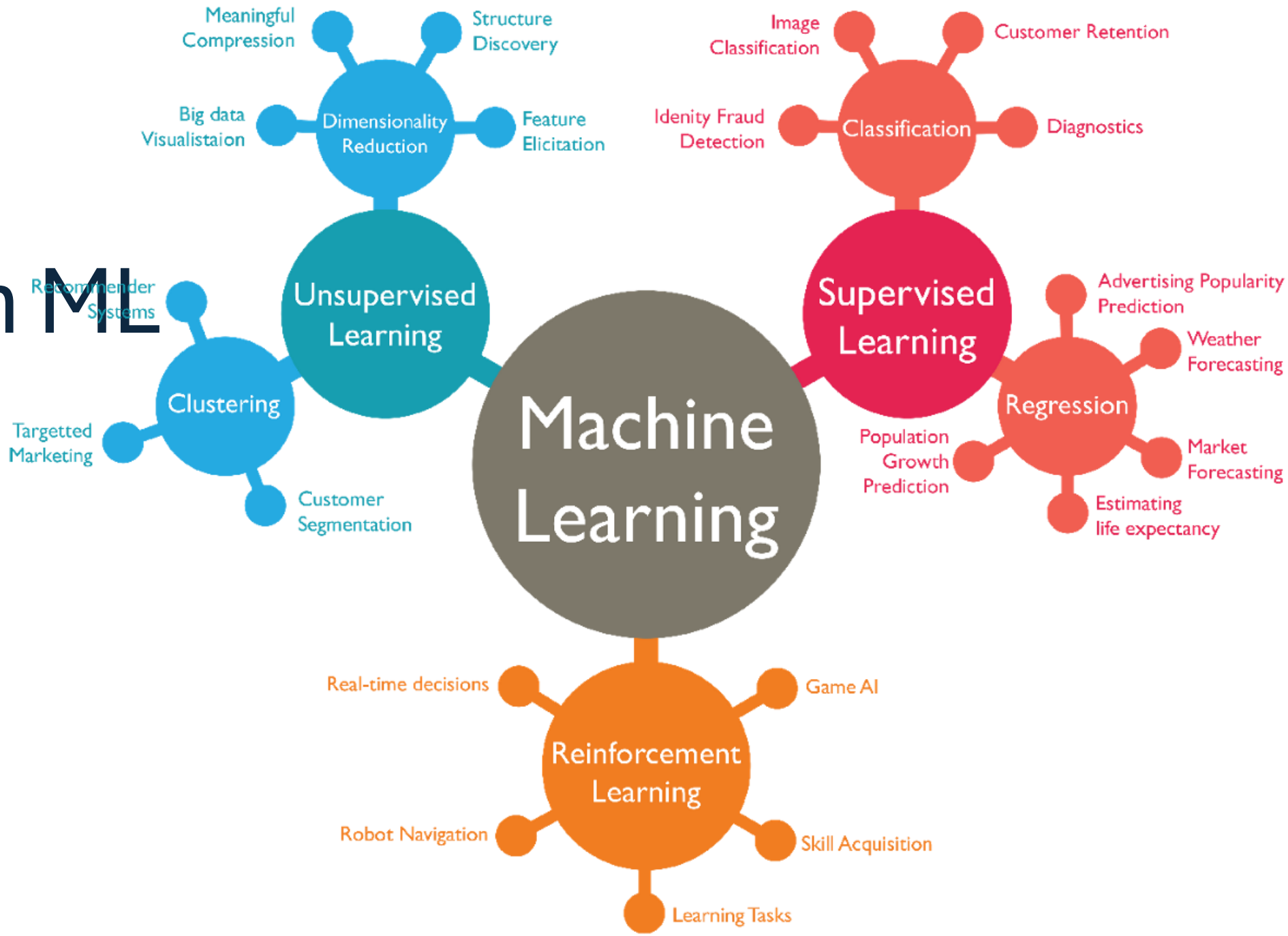


Popular Employer Salaries for Machine Learning Engineer

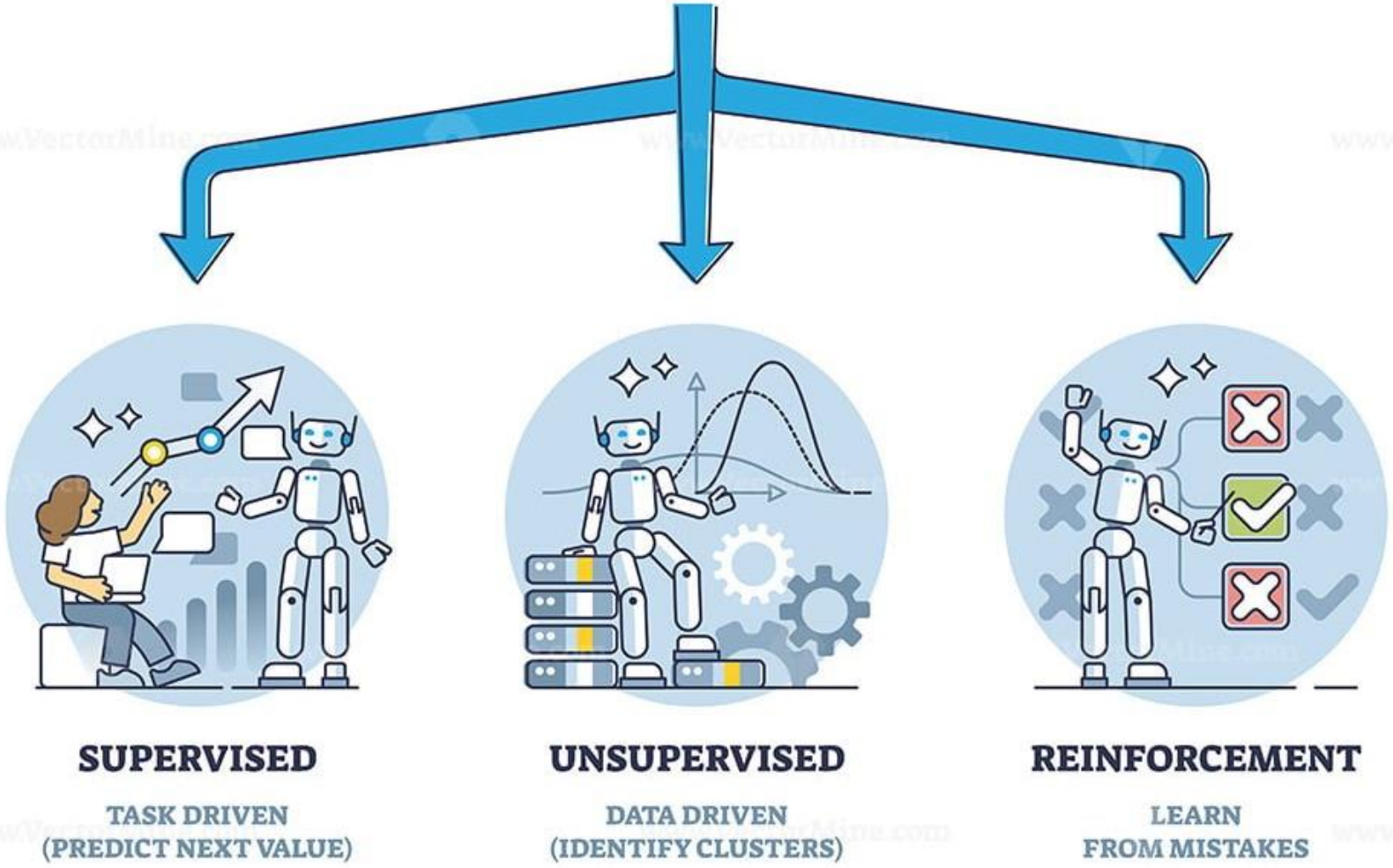
Amazon.com Inc	$130k
Apple Computer, Inc	$126k
Autodesk, Inc.	$130k
Nike, Inc.	$102k
Pitchbook	$100k

ML learning modes

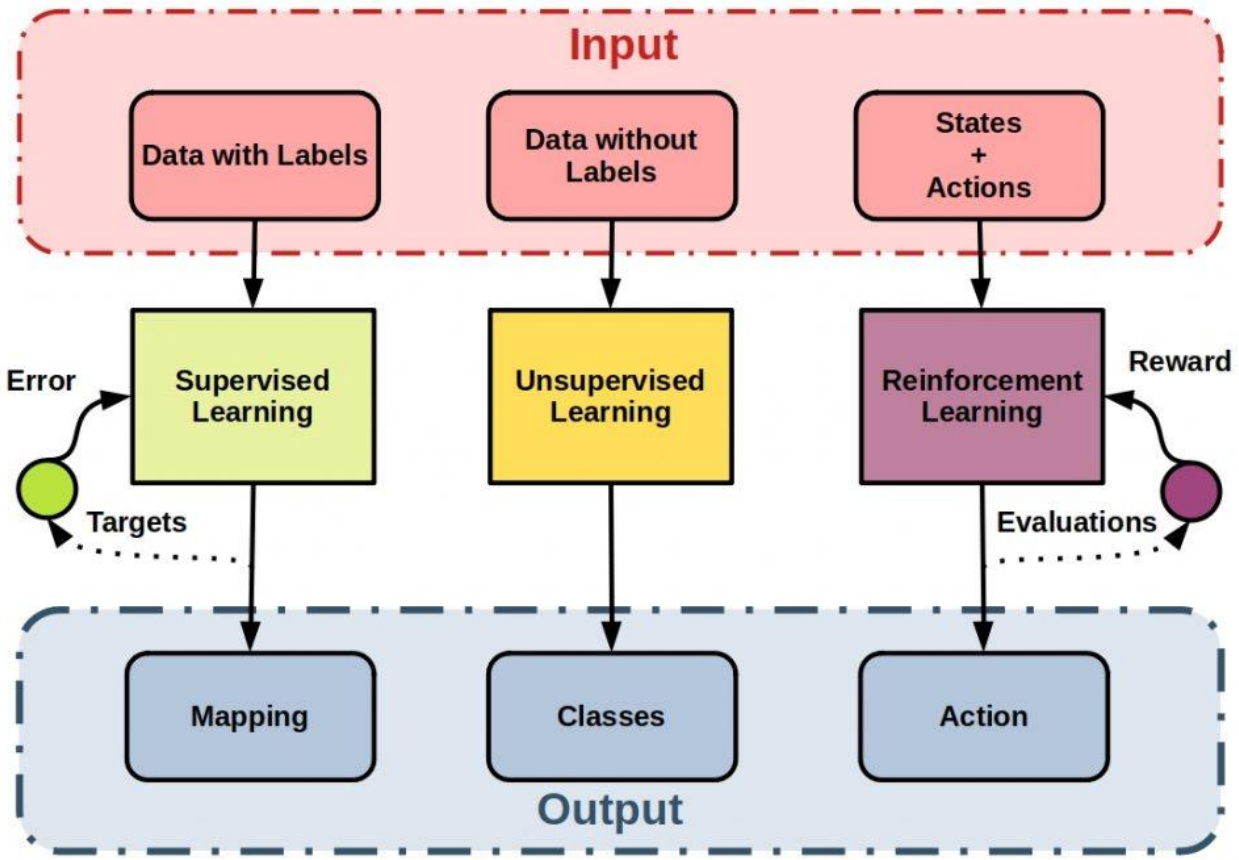
Learning modes in ML



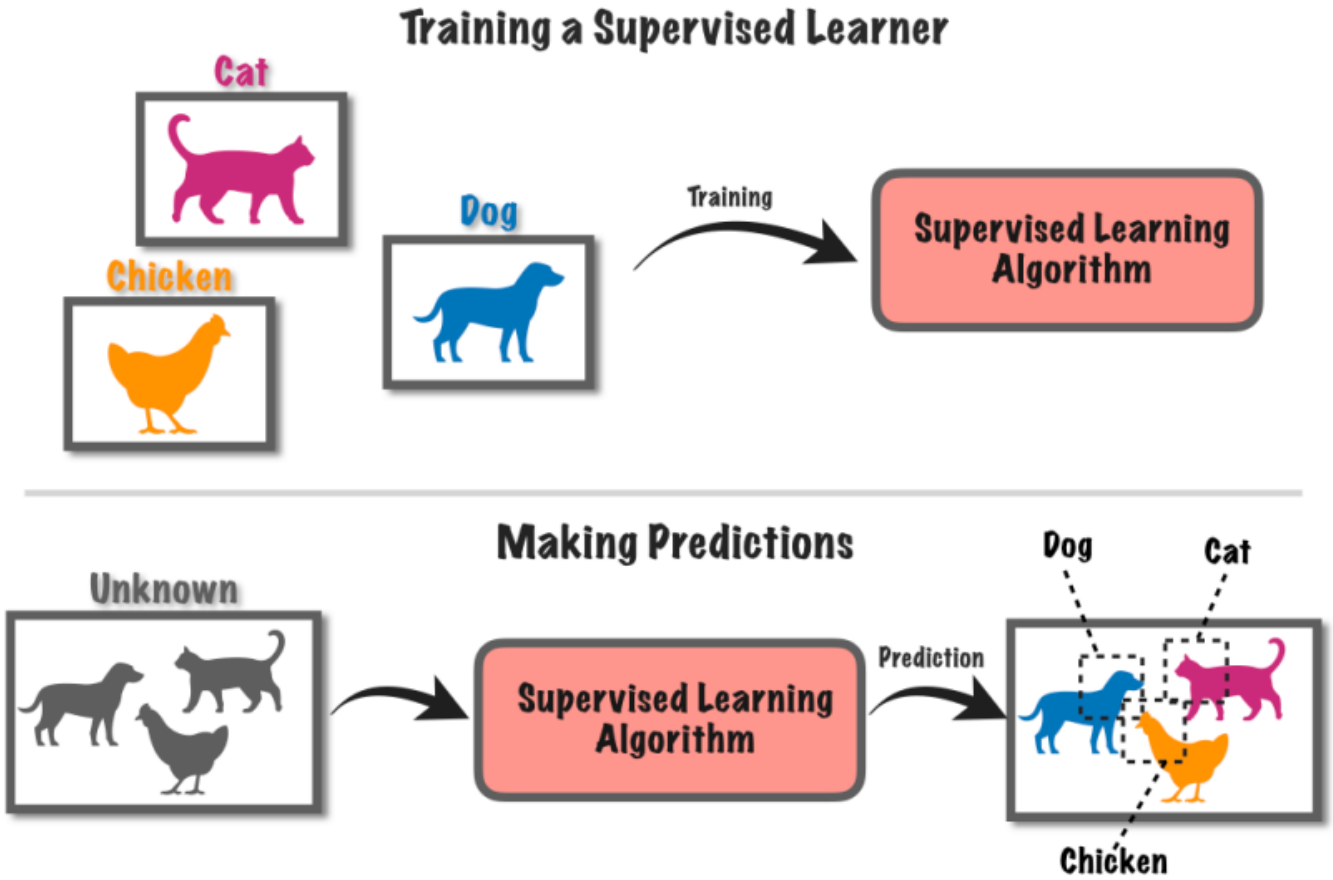
Learning modes in ML: keep simple



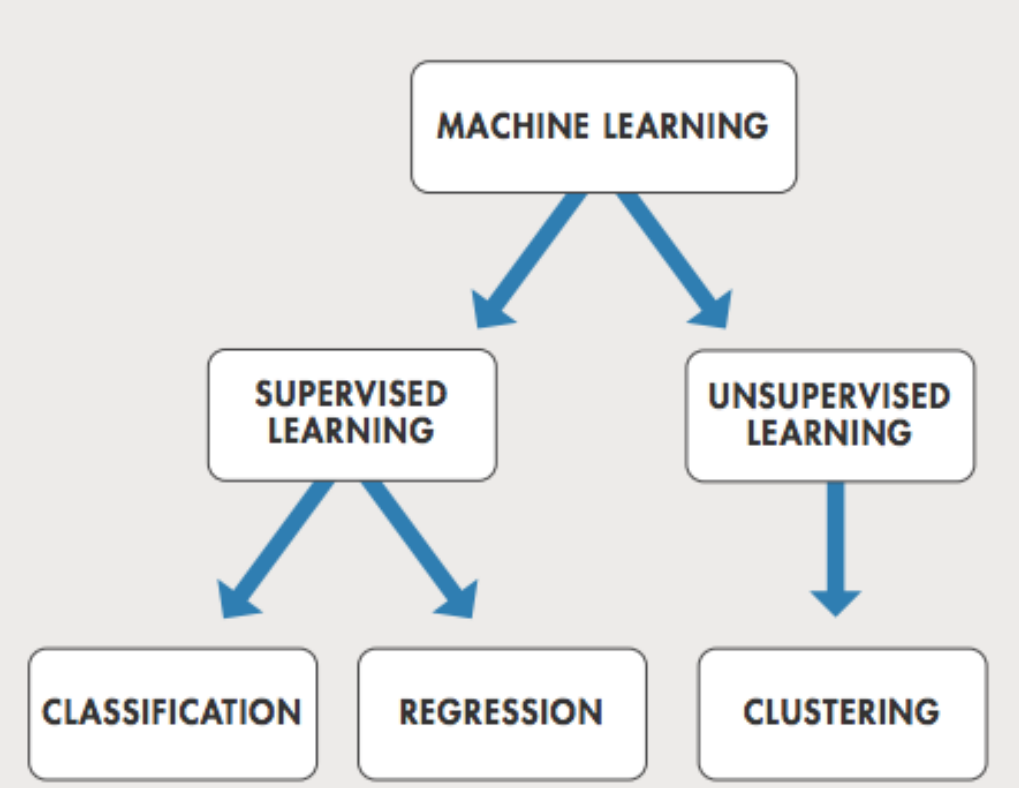
Learning modes in ML



Learning modes in ML: Supervised



Supervised and unsupervised methods



Supervised learning: regression.

Predicting House pricing

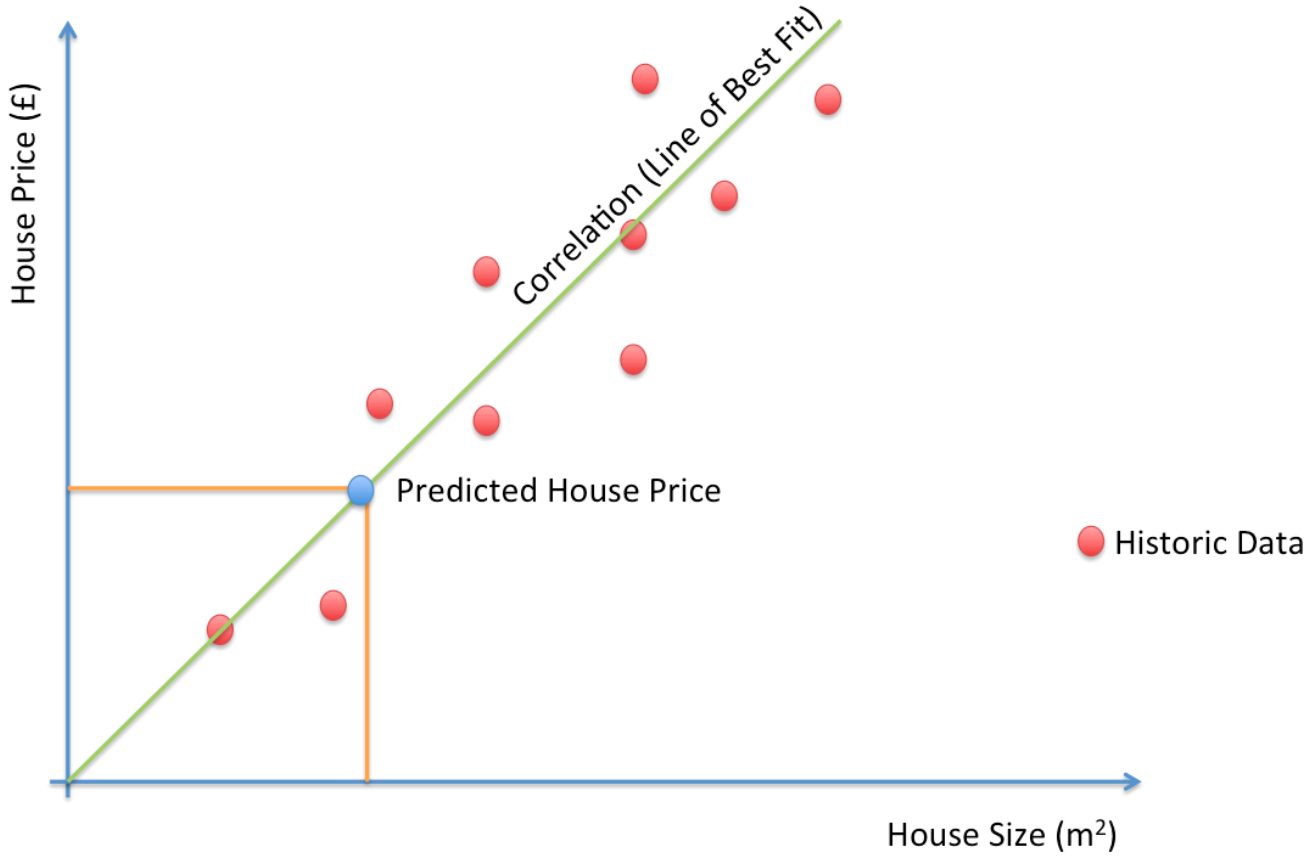
Supervised learning: regression

+ For example: we want to find a relation between Sqft and price

Sr. No.	Details	Price	Bedrooms	Bathrooms	Living Room Sqft	Floors	View	Waterfront	Grade	Basement sqft
1	23534368	221456	3	2	1008	1.00	0	0	6	410
2	89756456	321234	4	3	1342	2.00	0	0	7	700
3	45767857	134000	2	2	2001	1.00	0	0	6	0
4	25756756	214679	3	1	1200	1.00	0	0	6	0
5	23445466	213245	3	1	980	1.00	0	0	8	0

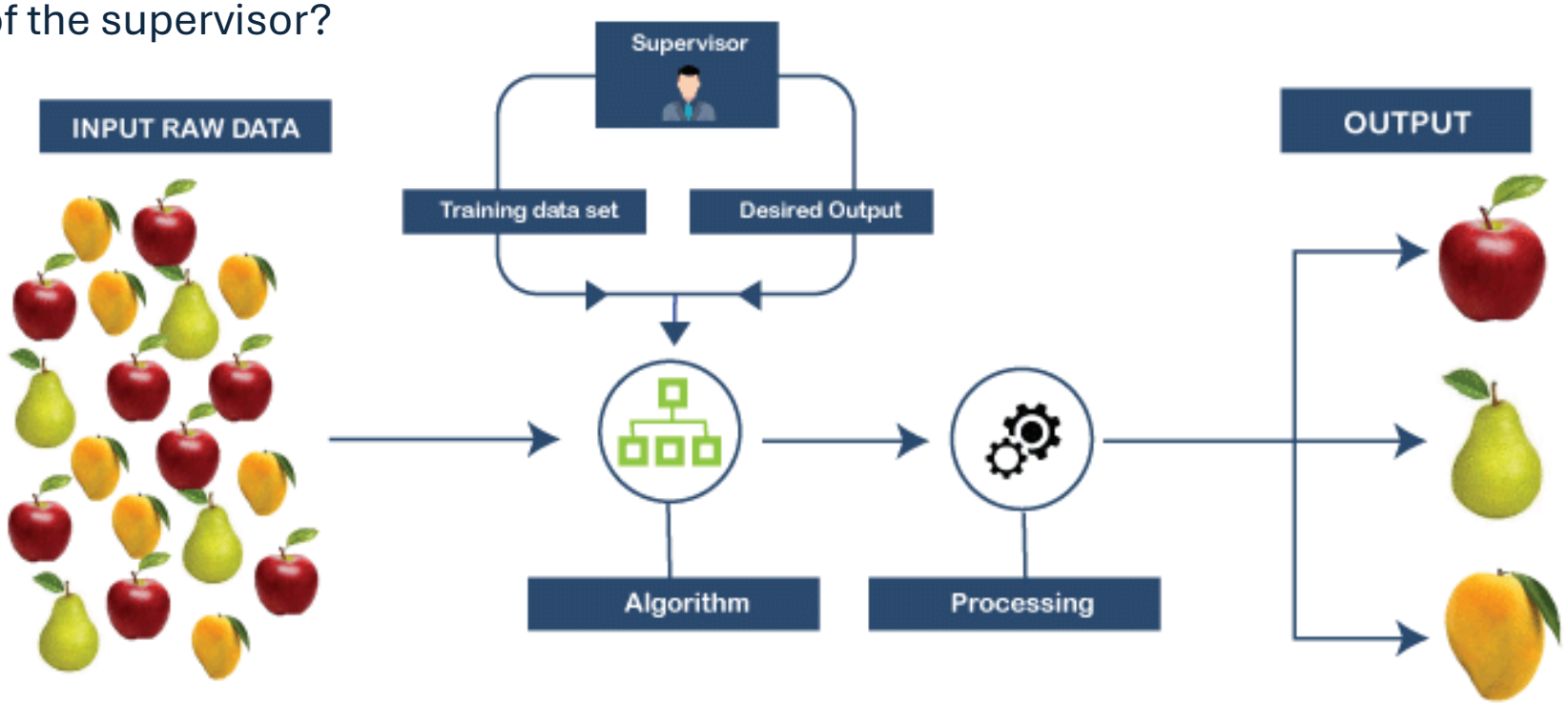
Supervised learning: regression

+ Which type of correlation between input and output?



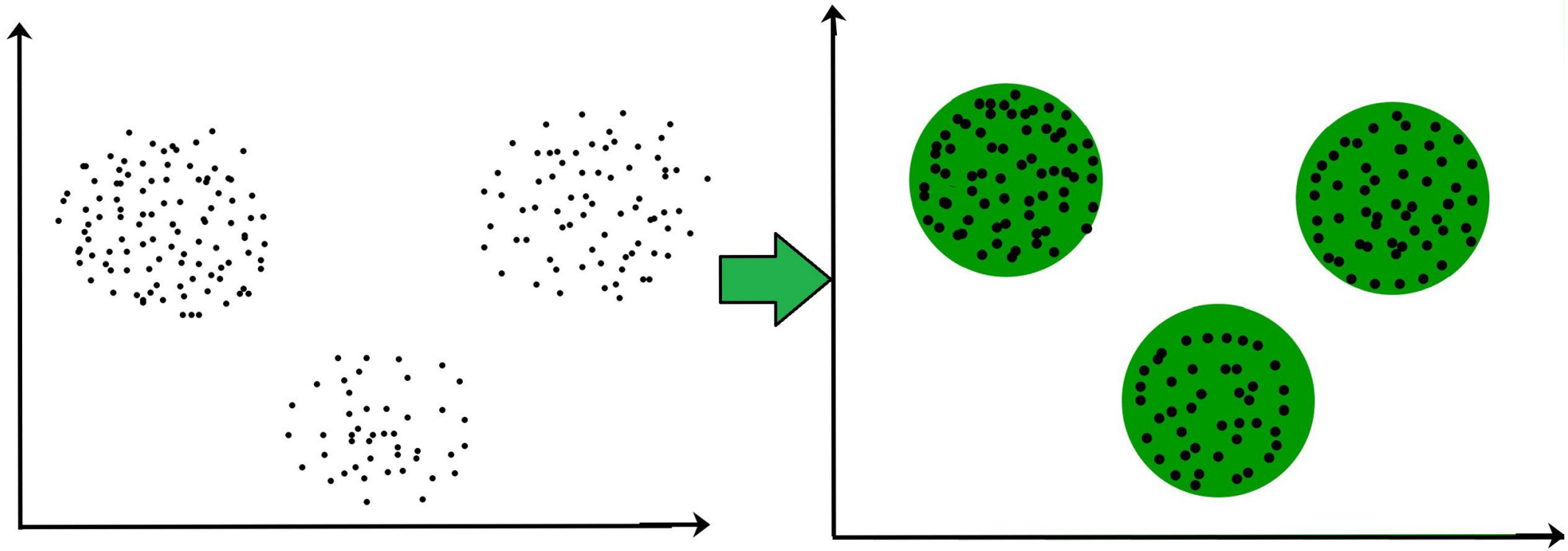
Supervised learning: classification

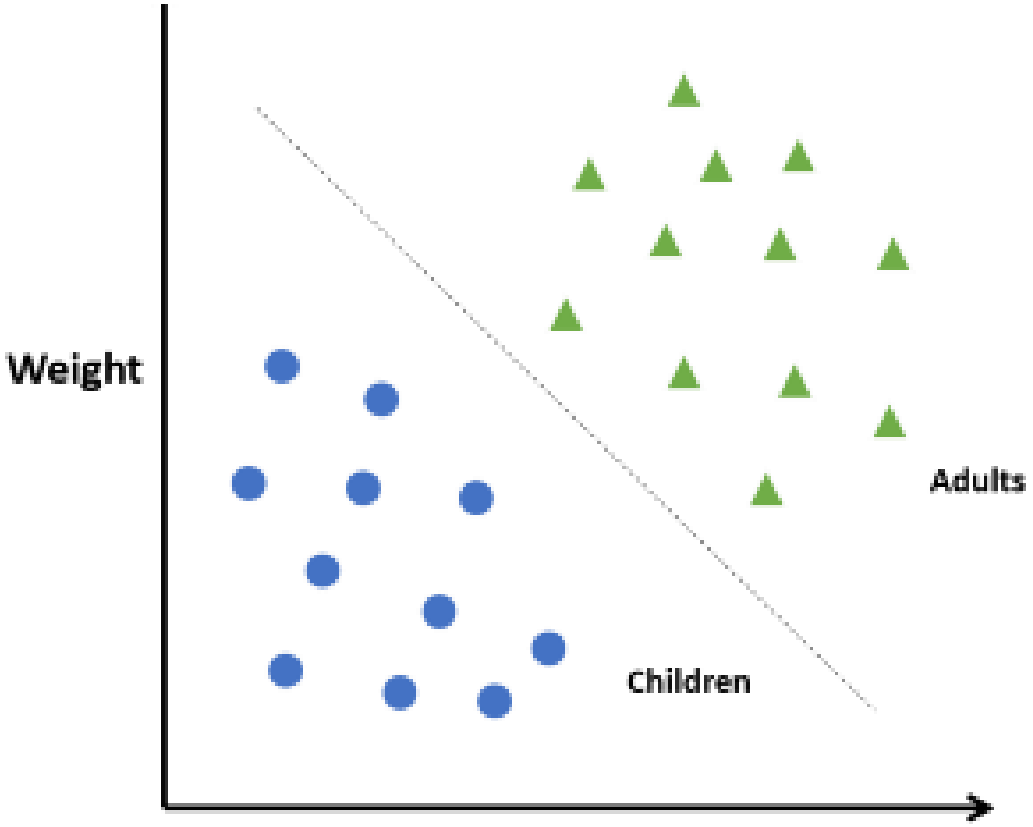
+ What's the role of the supervisor?

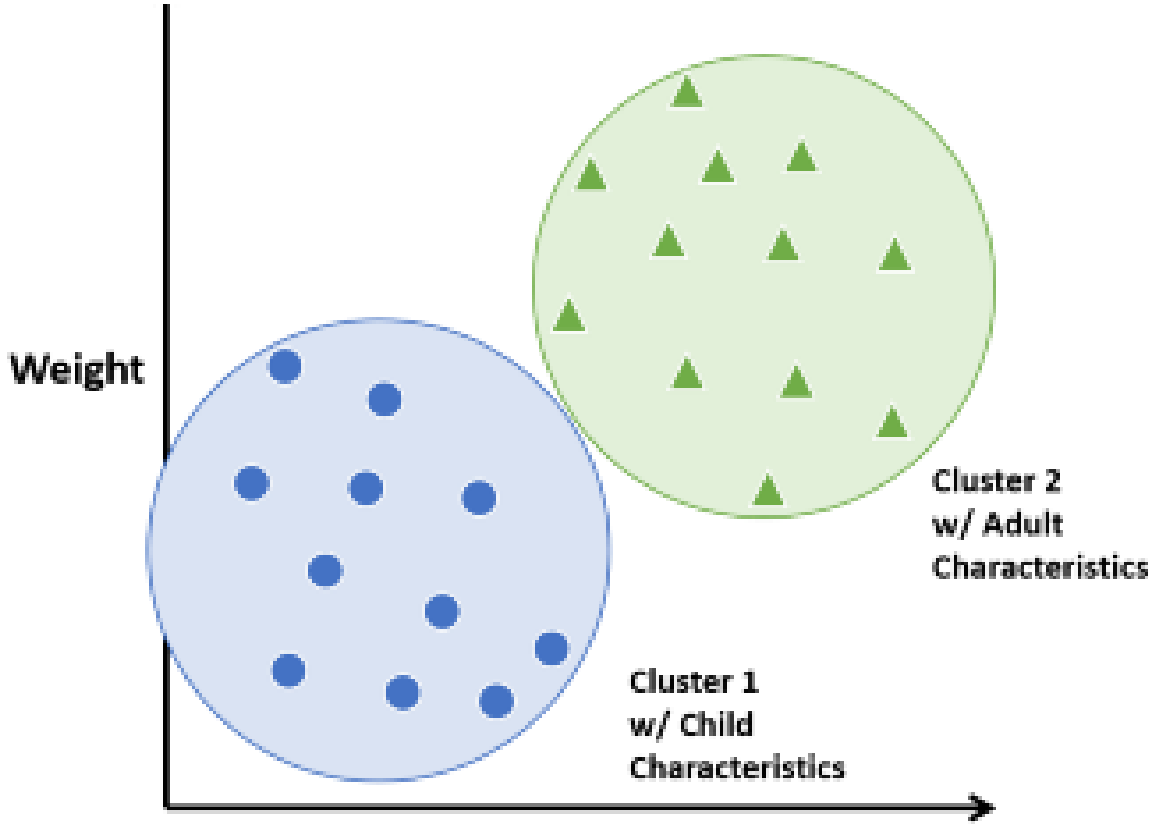


Unsupervised learning: clustering

+ Do you have an intuition of the principle to cluster data?

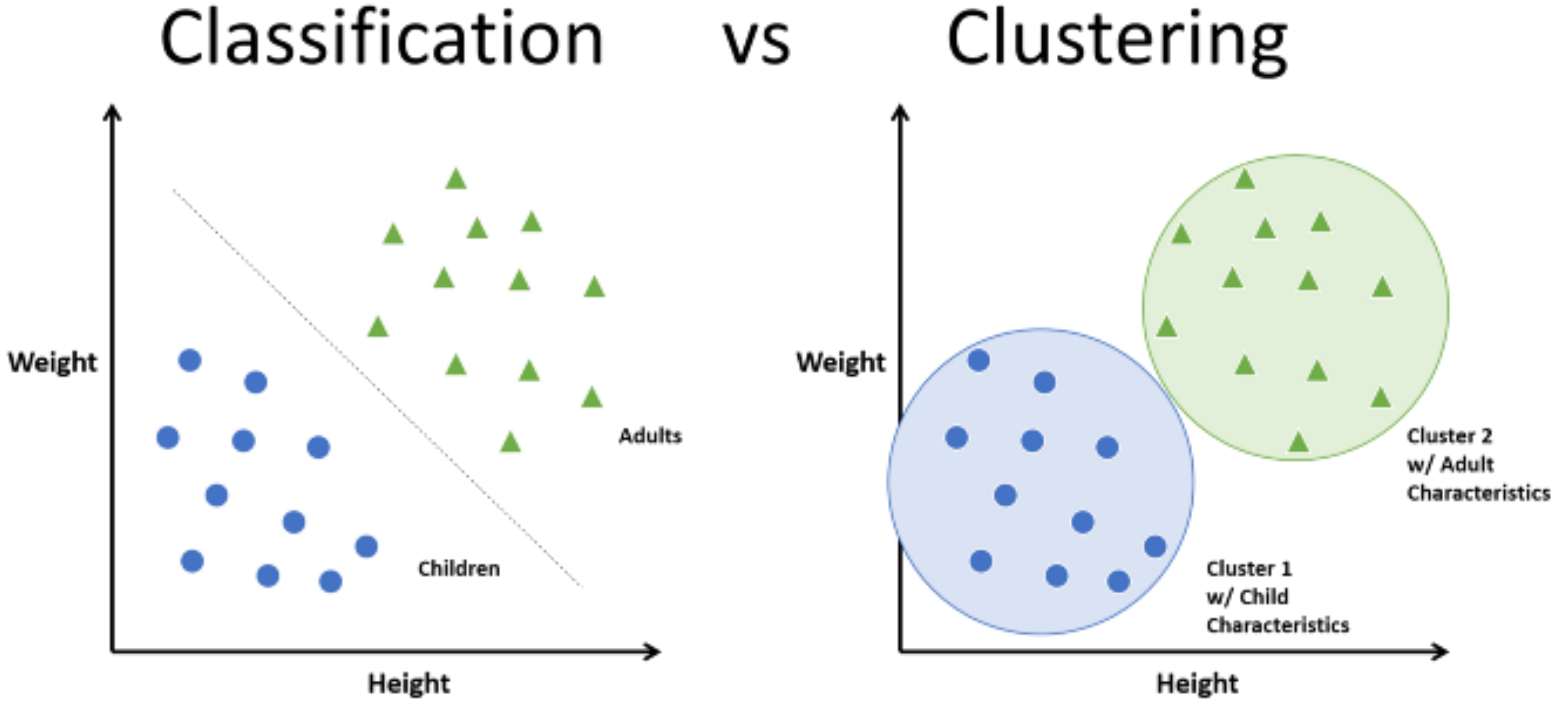






What is the difference between clustering and classification?

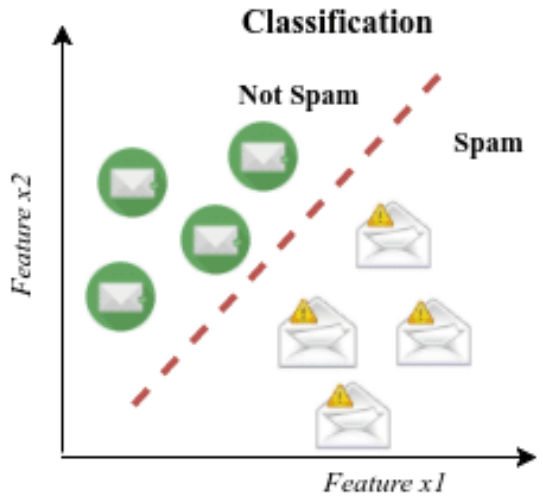
What is the difference between clustering and classification?



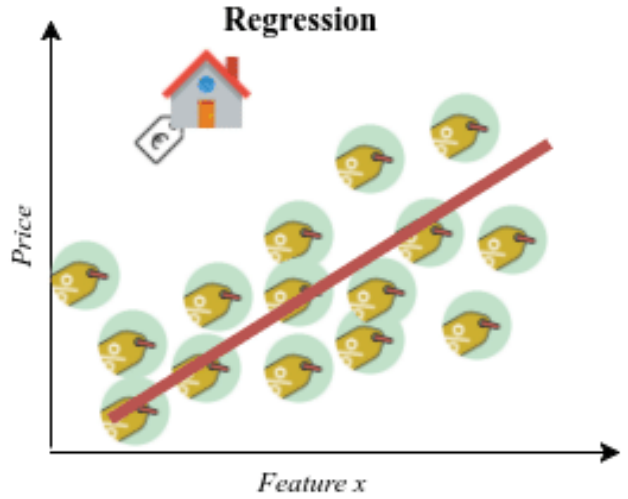
Classification: models the **relation** between labels and features

Clustering: assigns labels based on the **geometry** of the points

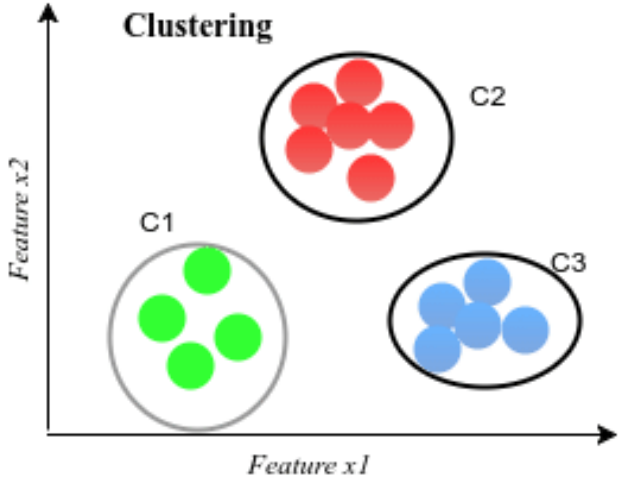
Summarizing



Predicts relation label/input; discrete quantity

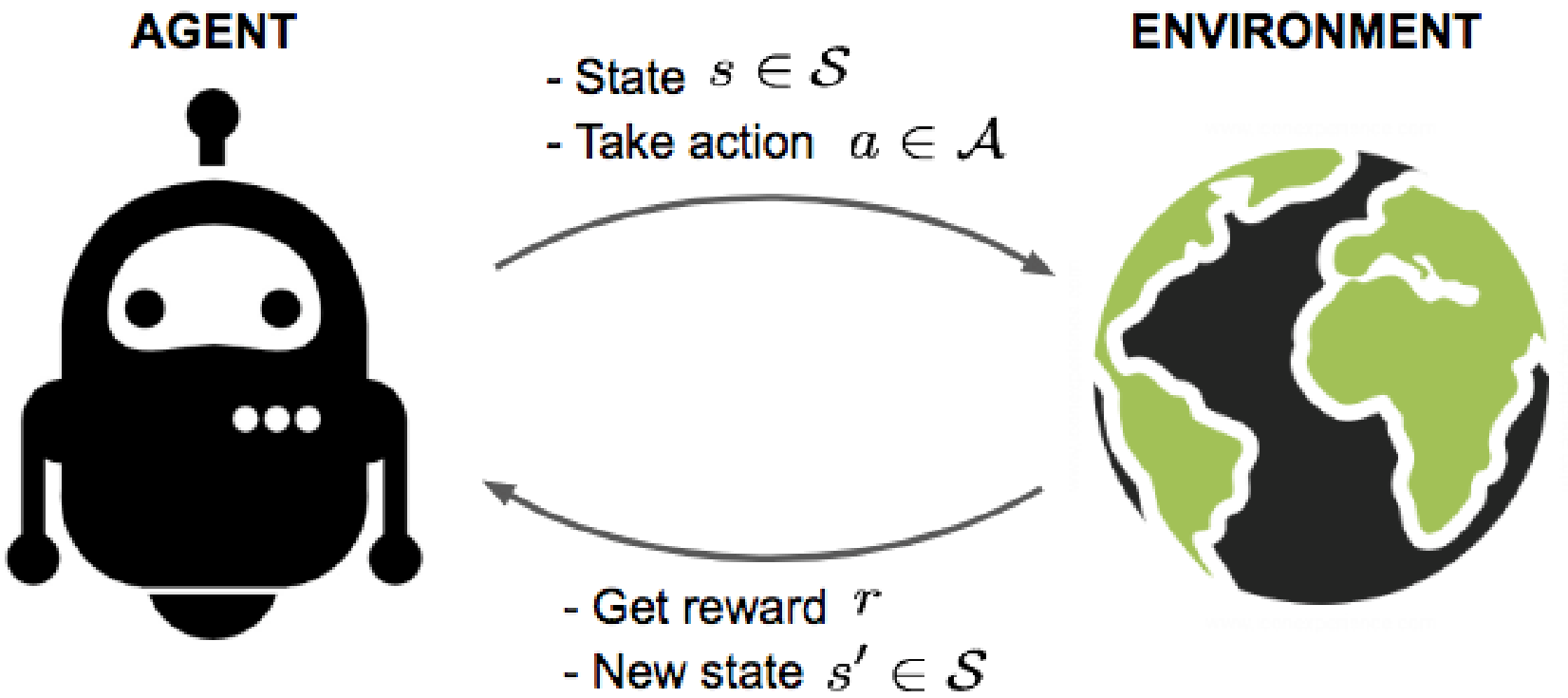


Predicts Continuous quantity

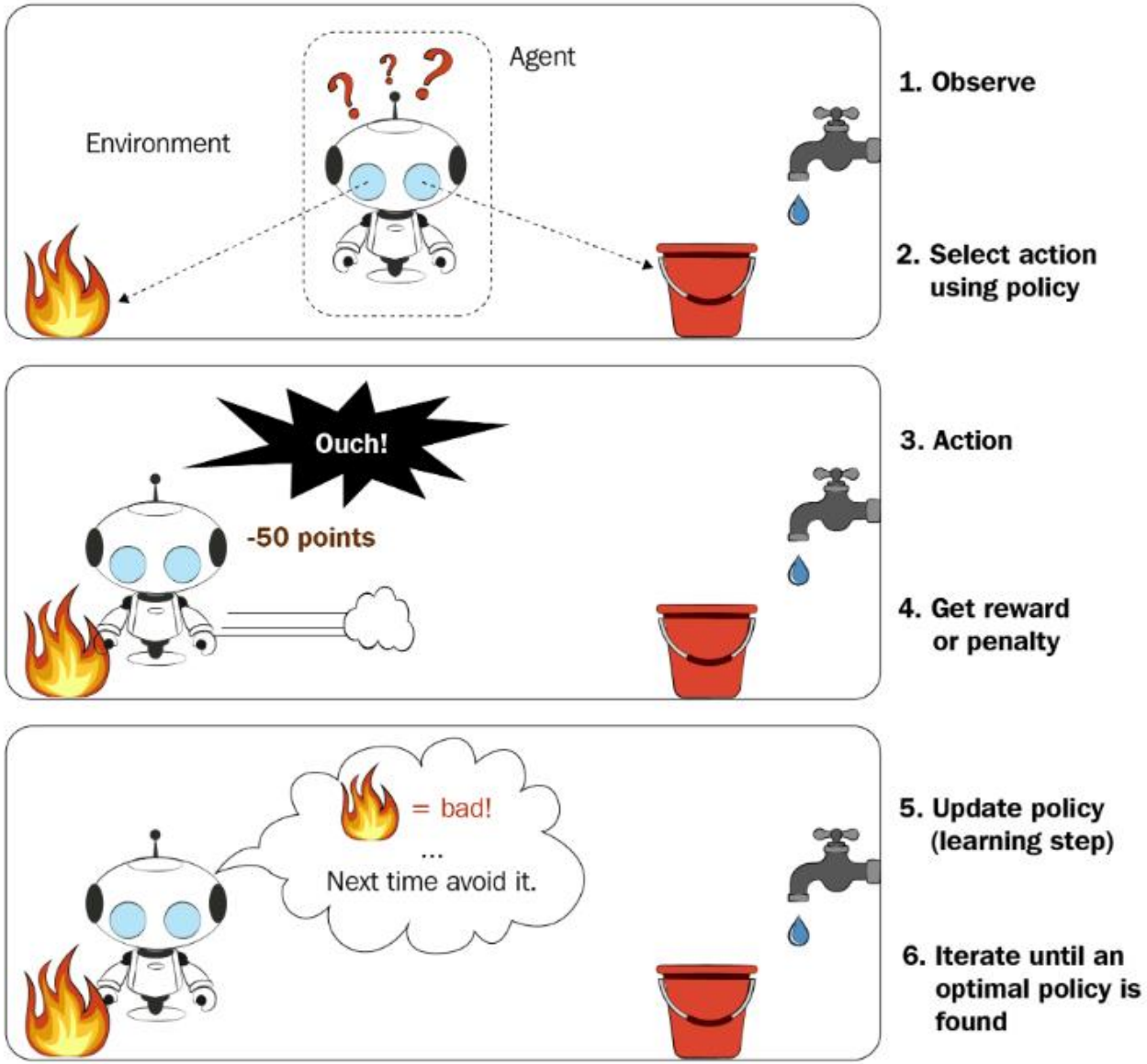


Predict class based on Geometry of points

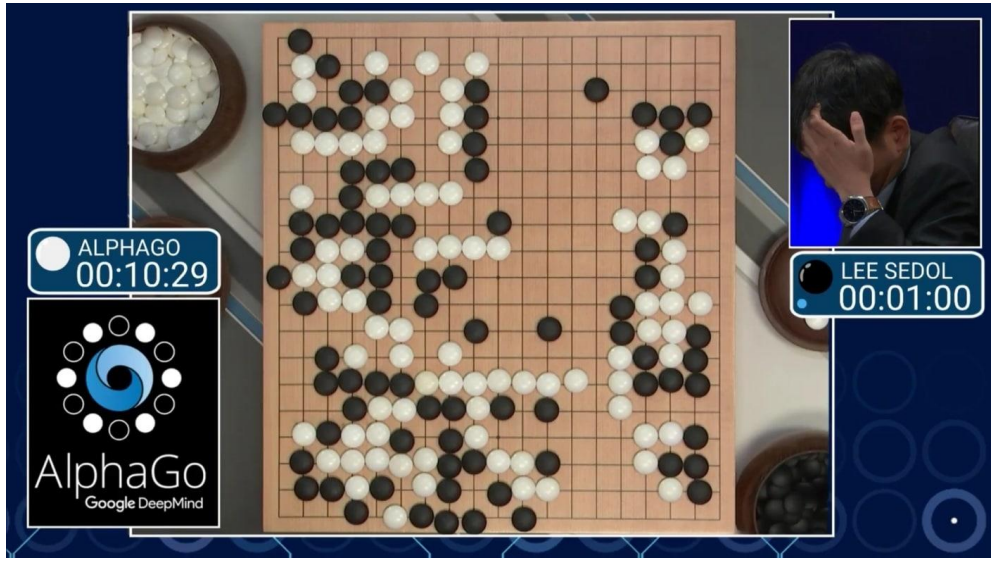
Reinforcement Learning

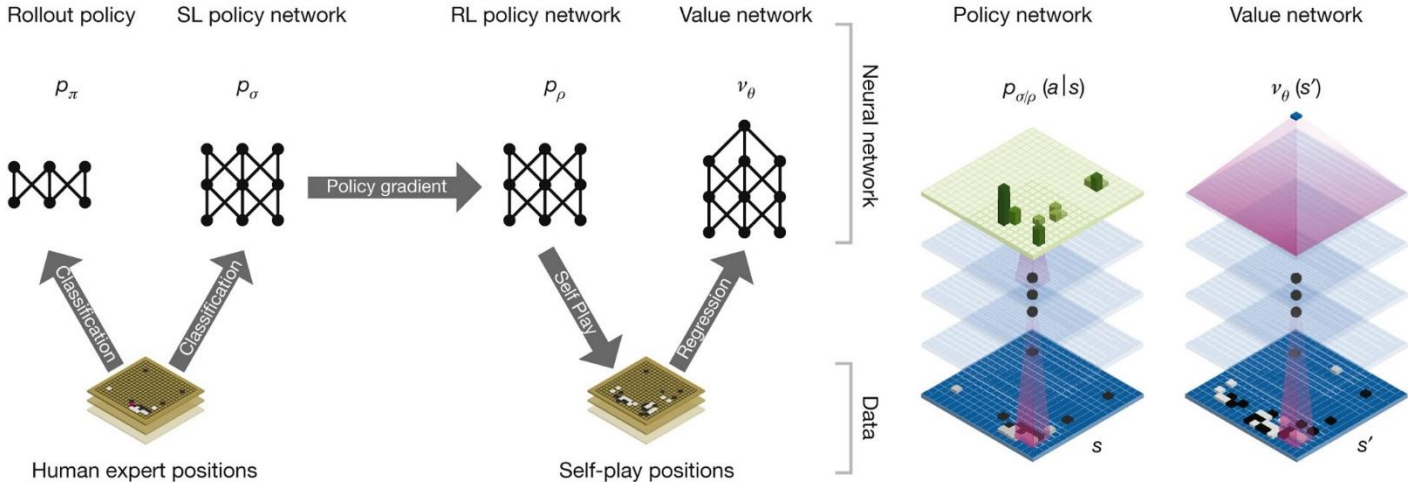


Reinforcement Learning: example



Reinforcement Learning: AlphaGo (DeepMind)





Notation: data space

Let us formalize the supervised machine learning setup. Our training data comes in pairs of inputs $(\mathbf{x}, y)$, where $\mathbf{x} \in \mathcal{R}^d$ is the input instance and y its label. The entire training data is denoted as

$$D = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\} \subseteq \mathcal{R}^d \times \mathcal{C}$$

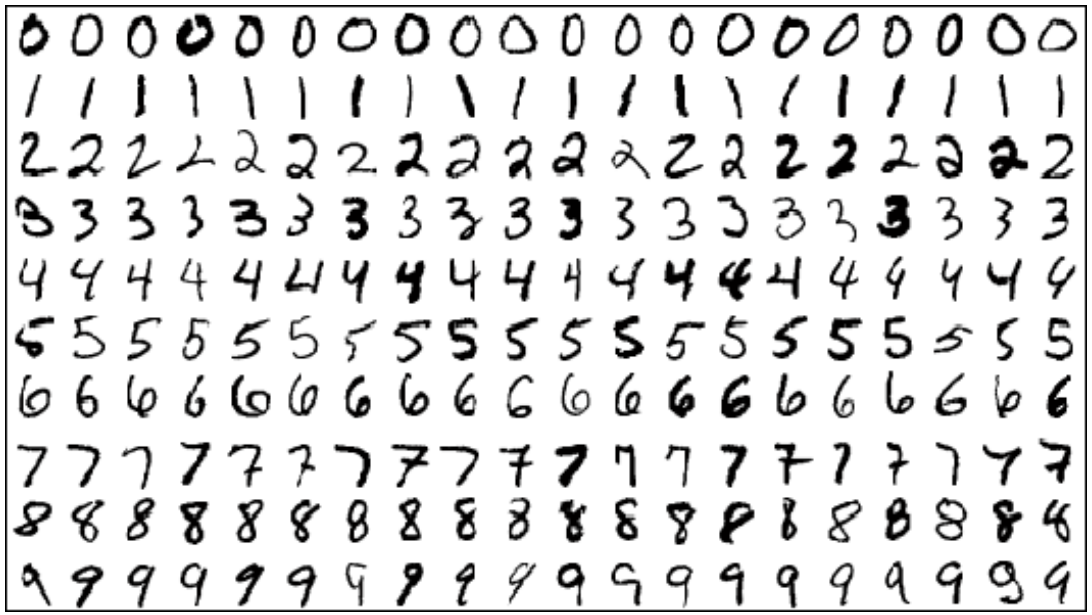
where:

- $\mathcal{R}^d$ is the d-dimensional feature space
- $\mathbf{x}_i$ is the input vector of the i^{th} sample
- y_i is the label of the i^{th} sample
- $\mathcal{C}$ is the label space

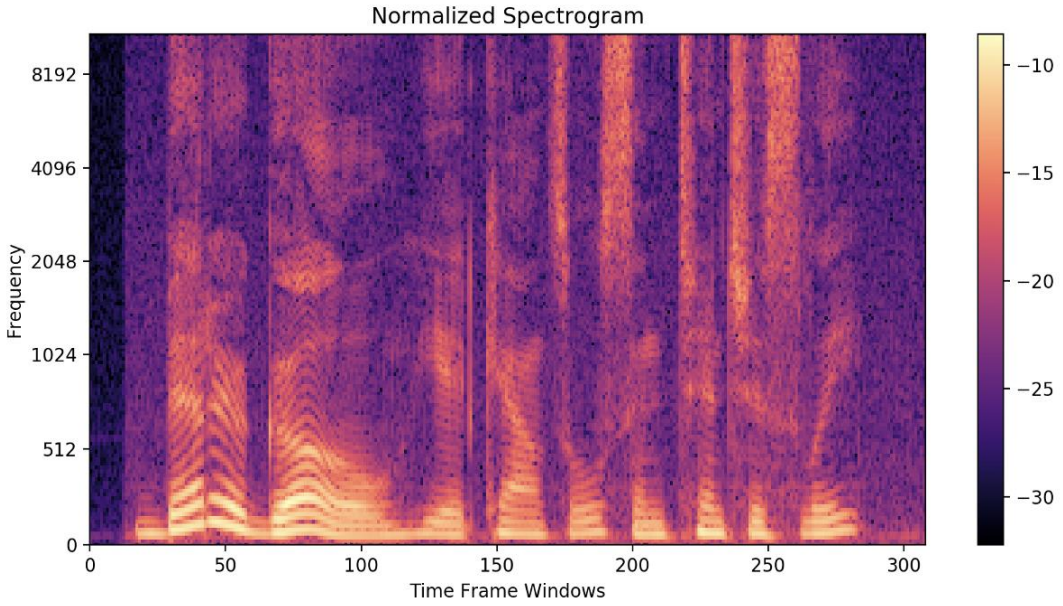
The data points $(\mathbf{x}_i, y_i)$ are drawn from some (unknown) distribution $\mathcal{P}(X, Y)$. Ultimately we would like to learn a function h such that for a new pair $(\mathbf{x}, y) \sim \mathcal{P}$, we have $h(\mathbf{x}) = y$ with high probability (or $h(\mathbf{x}) \approx y$). We will get to this later. For now let us go through some examples of X and Y.

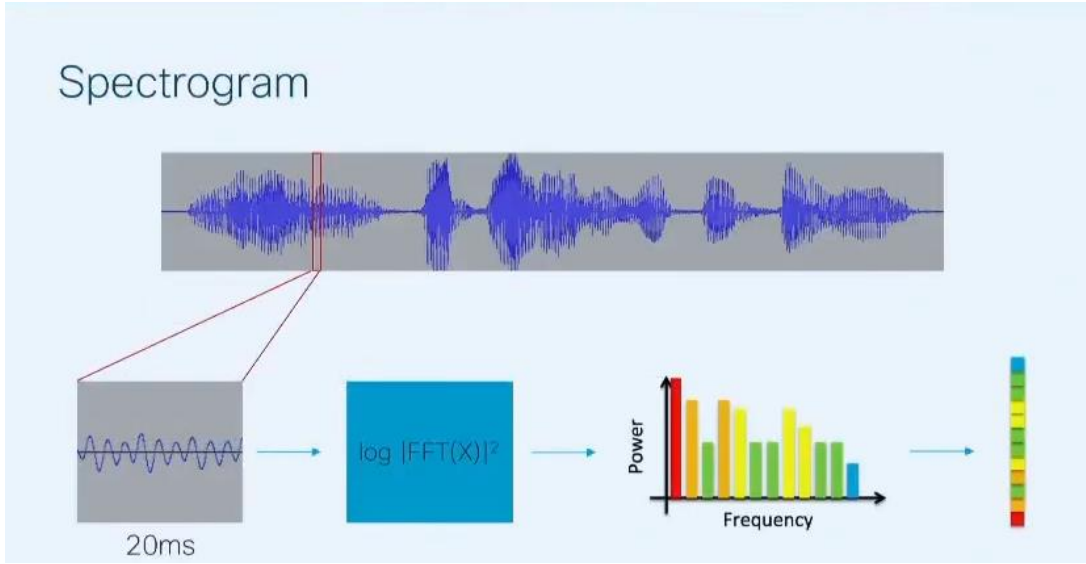
Some examples of x_i

Datasets: images

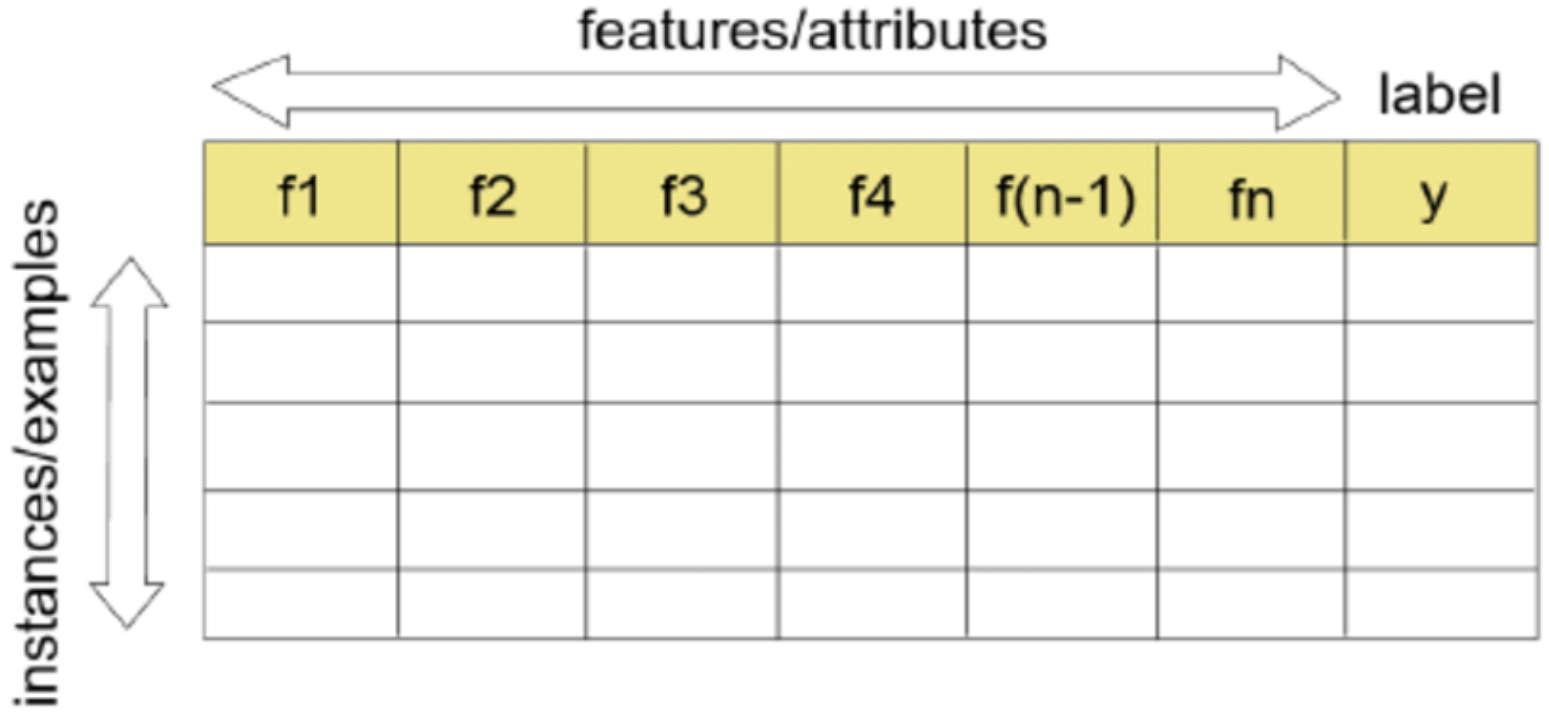


Datasets: sounds





Datasets: numbers and categories



Notation: label space (what about y_i?)

+ What about unsupervised learning?
+ How do you transform many classes labels into vectors?

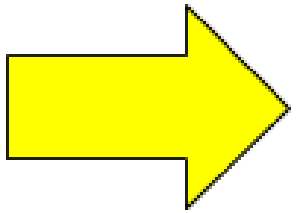
Examples of Label Spaces

There are multiple scenarios for the label space $\mathcal{C}$:

Binary classification	$\mathcal{C} = \{0, 1\}$ or $\mathcal{C} = \{-1, +1\}$.	Eg. spam filtering. An email is either spam $(+1)$, or not (-1).
Multi-class classification	$\mathcal{C} = \{1, 2, \cdots, K\}$ $(K \geq 2)$.	Eg. face classification. A person can be exactly one of K identities (e.g., 1="Barack Obama", 2="George W. Bush", etc.).
Regression	$\mathcal{C} = \mathbb{R}$.	Eg. predict future temperature or the height of a person.

One hot encoding

Color
Red
Red
Yellow
Green
Yellow



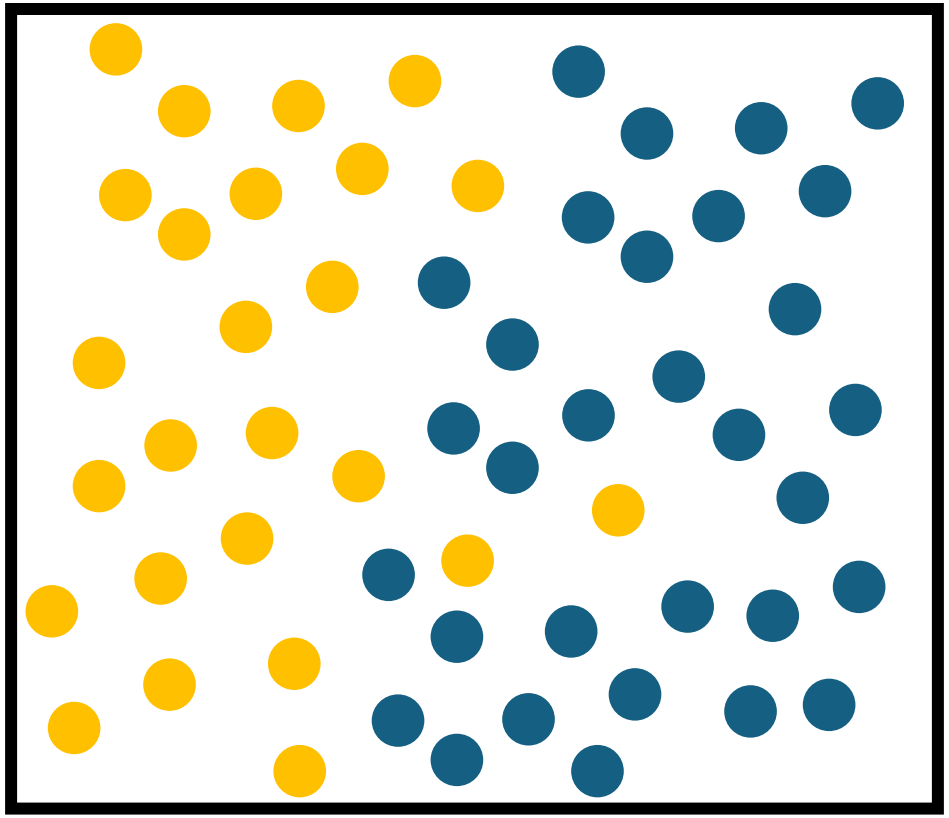
Red	Yellow	Green
1	0	0
1	0	0
0	1	0
0	0	1

How do we extract meaningful information from data?

We need to choose a model of the data

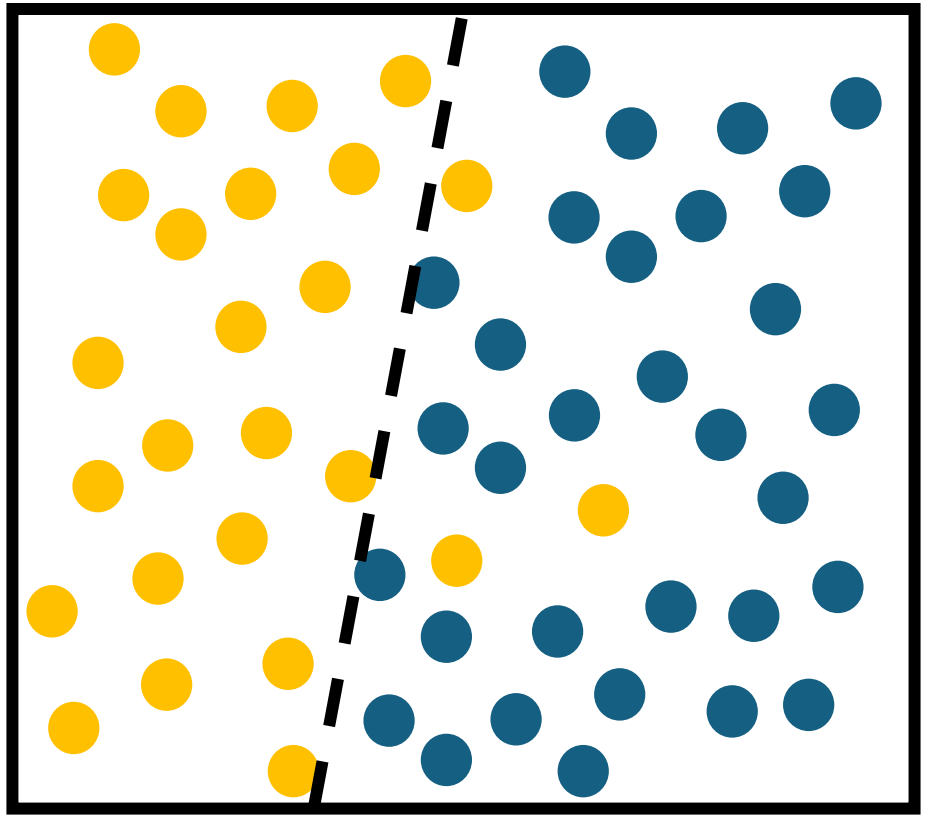
Task and model

- My task would be to discriminate between orange and blue points.
- My model would be the function that encodes the line at the border



Hypothesis space and model complexity

- Hypothesis: The border is a straight line (small hypothesis space).
- What do you think about its complexity?



Hypothesis space and model complexity

+Hypothesis: The border is a general line (big space).

+Complexity: High.