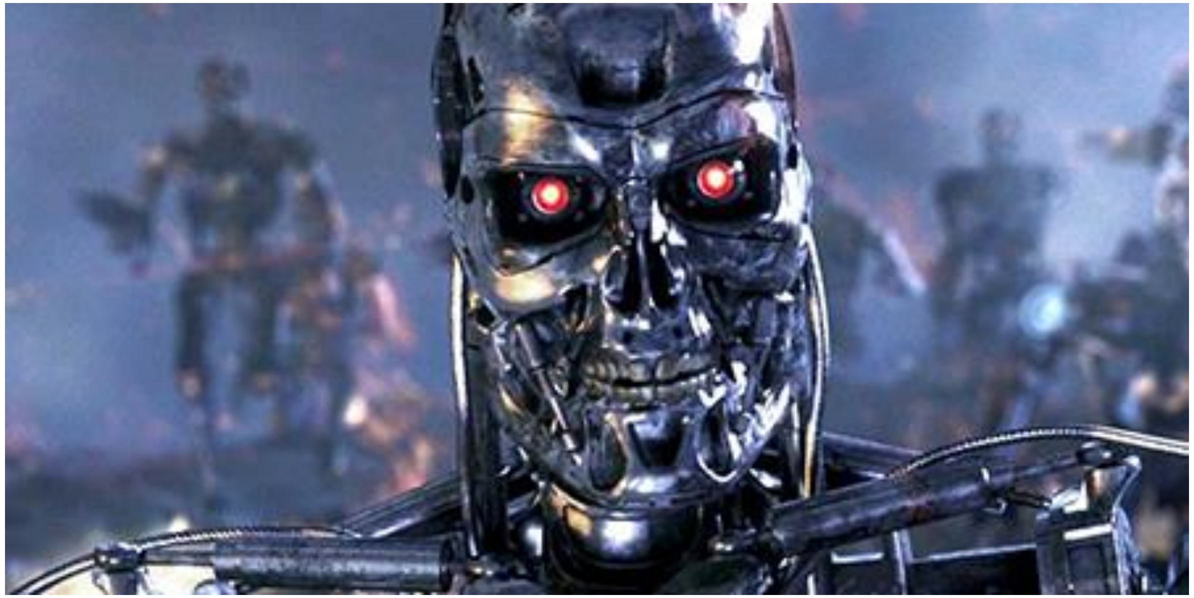
Machine Learning and Data science

By Mohcine Boudhane

What is machine learning?

- When we hear about machine learning, or more generally artificial intelligence - which machine learning is a subdomain, we usually think about it:



But experts in the field are formal: despite all the concerns raised in the media, machine learning, and more generally artificial intelligence, do not pose a real threat. In the current state, we are really very far from having reached a level of sufficient intelligence in the machines to have something to worry about. In any case, the ethical issues around machine learning will not be addressed in this course.

What is machine learning?

What is machine learning

- Learning 1: bright and yellow mangoes are sweeter than pale and yellow ones.
- Learning 2: the smaller, bright yellow mangoes are sweet only half the time.
- Learning 3: softer ones are juicier.
- Learning 4: Green mangoes are tastier than yellow ones.
- Learning 5: I don’t need mangoes any more.

What is machine learning

For machine:

- if is bright yellow and size is big and sold by x: mango is sweet.
- if (soft): mango is juicy

Machine learning: Definition

- Machine Learning is a concept which allows the machine to learn from examples and experience, and that without being explicitly programmed. So instead of you writing the code, ….you feed data …., and the algorithm/machine builds the logic based on the given data.

- This view of machine learning can be traced back to Arthur Samuel's definition from 1959:

"Machine Learning: Field of study that gives computers the ability to learn without being explicitly programmed".

Arthur Samuel is one of the pioneers of machine learning. While at IBM he developed a program that learned how to play checkers better than him.

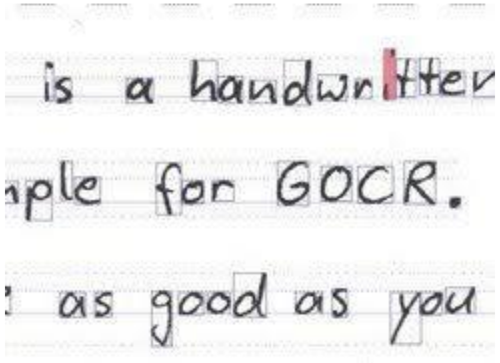
Why using machine learning?

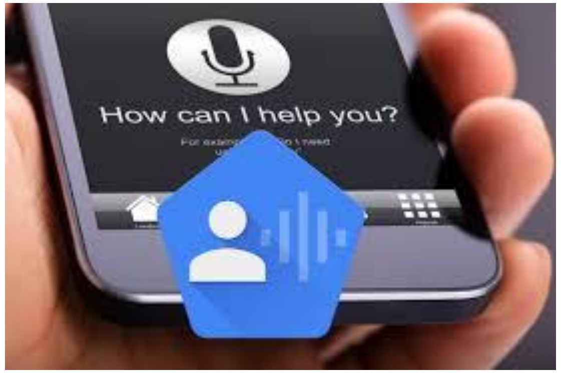
- Do we need really machine learning algorithms in our life

Data science and machine learning are two words that are very popular when talking about the **Big Data revolution**, the **prediction of behavior** or simply the **digital transformation of companies**. And as for all innovative areas, it is sometimes difficult to understand what it is.

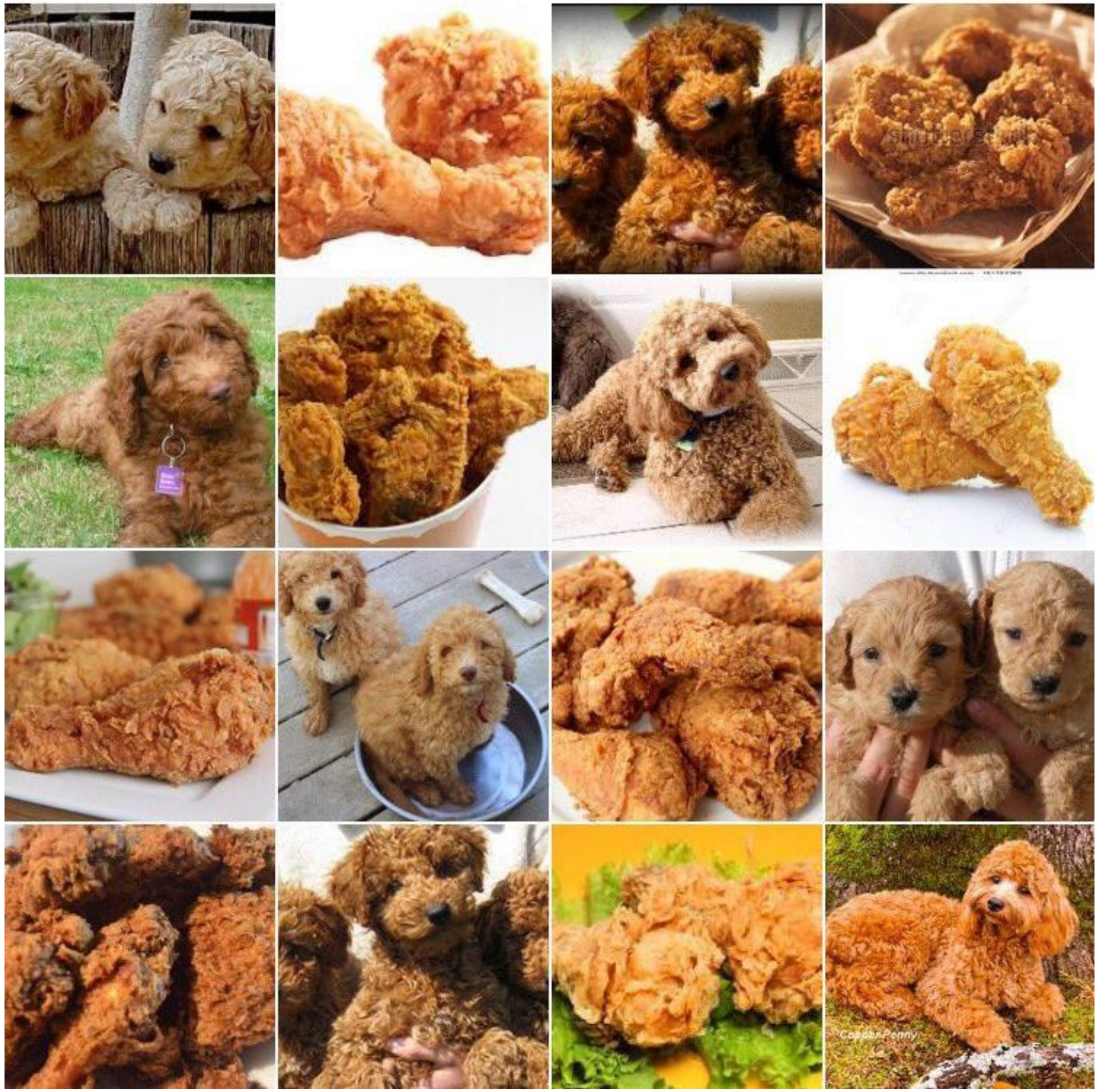
The reason of using machine learning is:
The presence of huge amount of data produced and collected by humans.
The improvement and greater accessibility of machine learning algorithms.
The exponential increase in computing capabilities of computers.

Machine learning: usecases

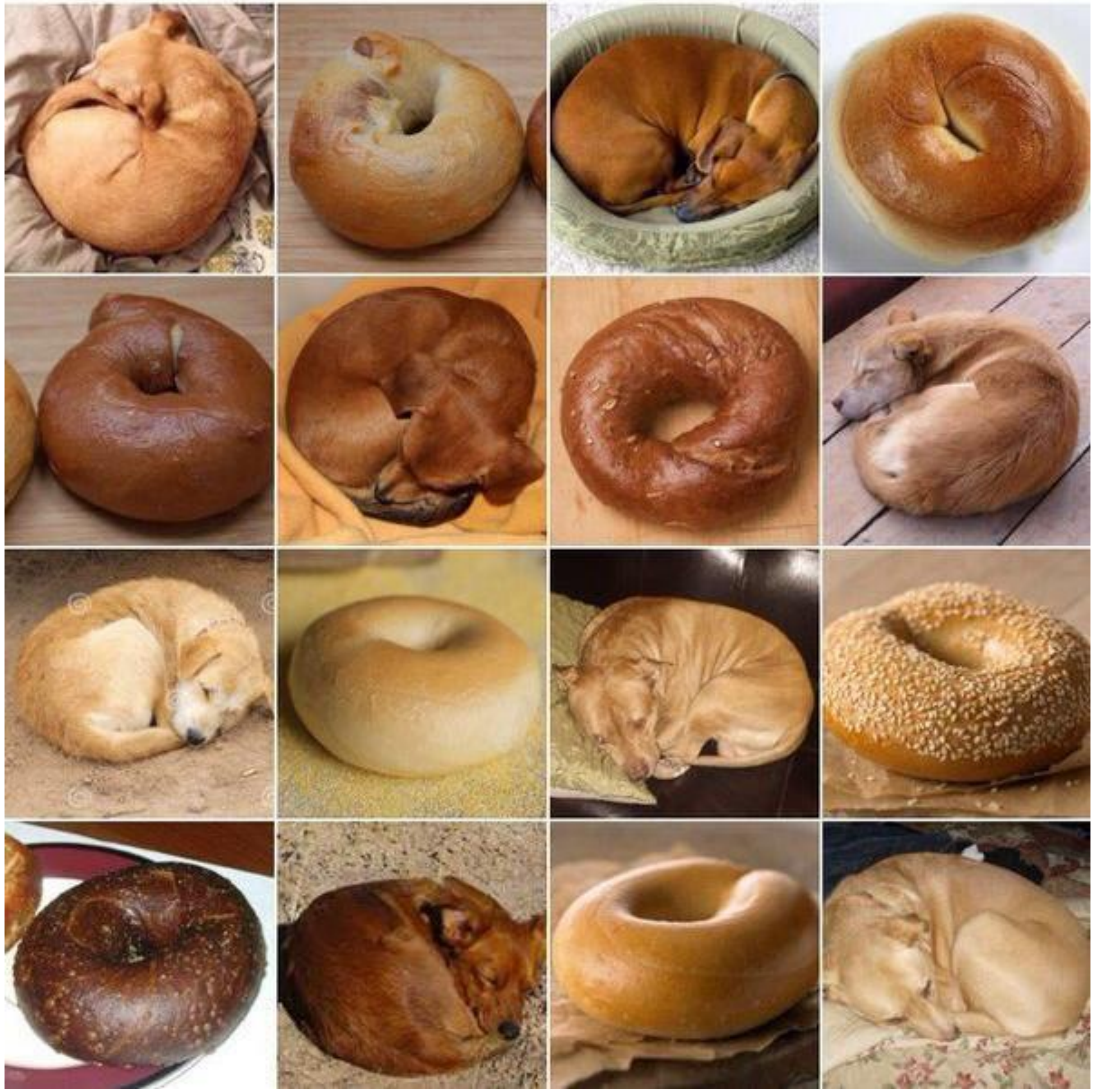




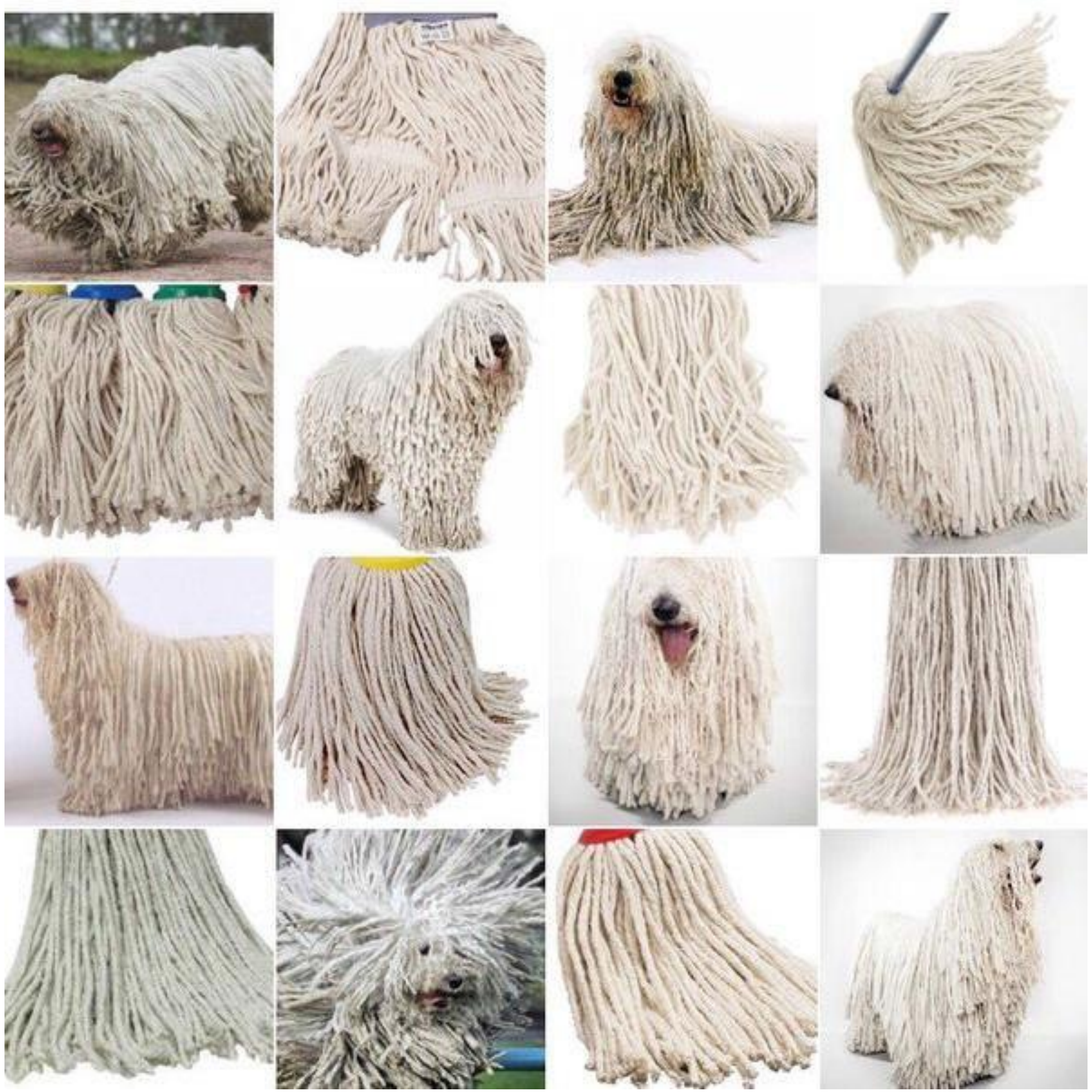
Labradoodle or fried chicken



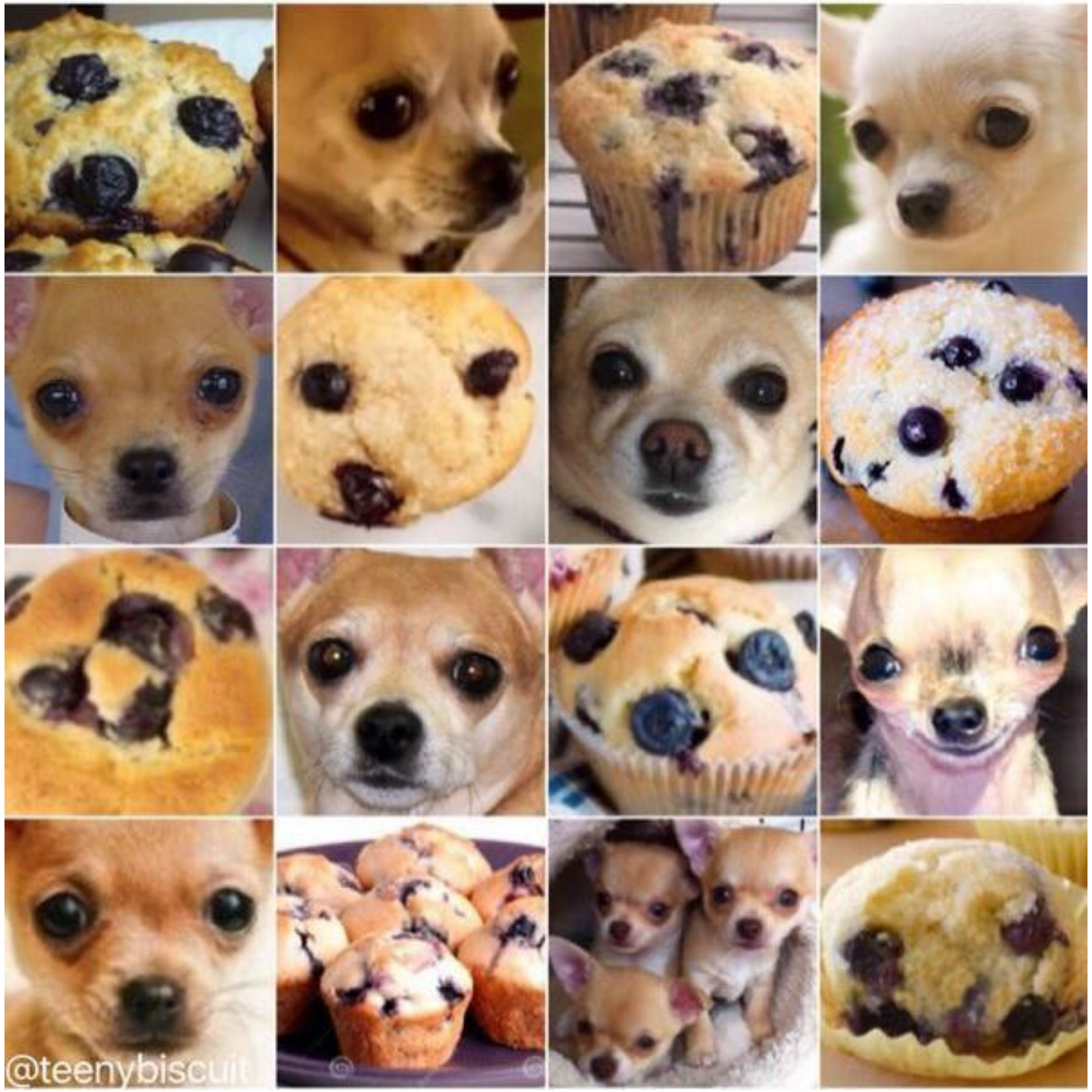
Puppy or bagel



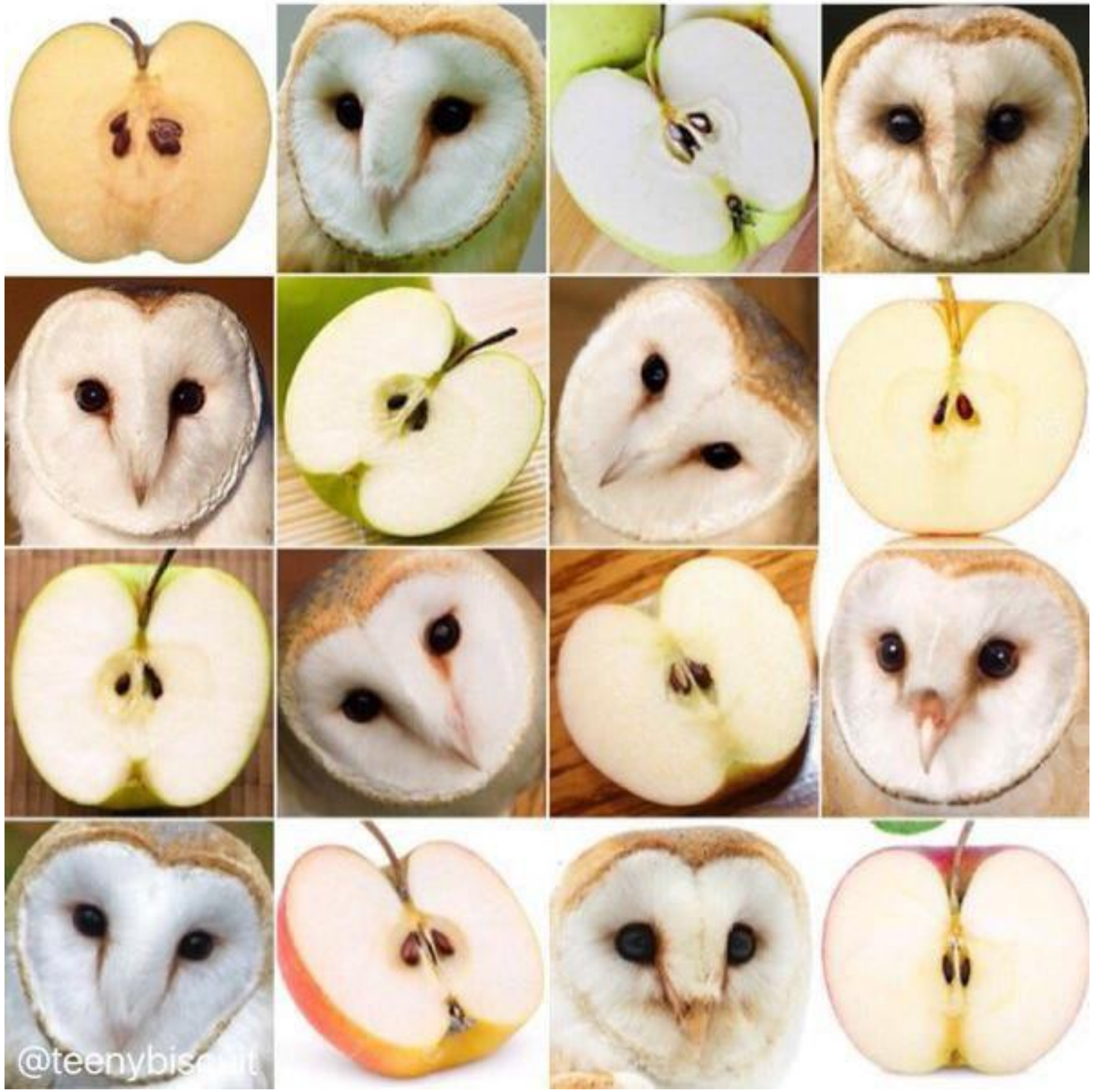
Sheepdog or mop



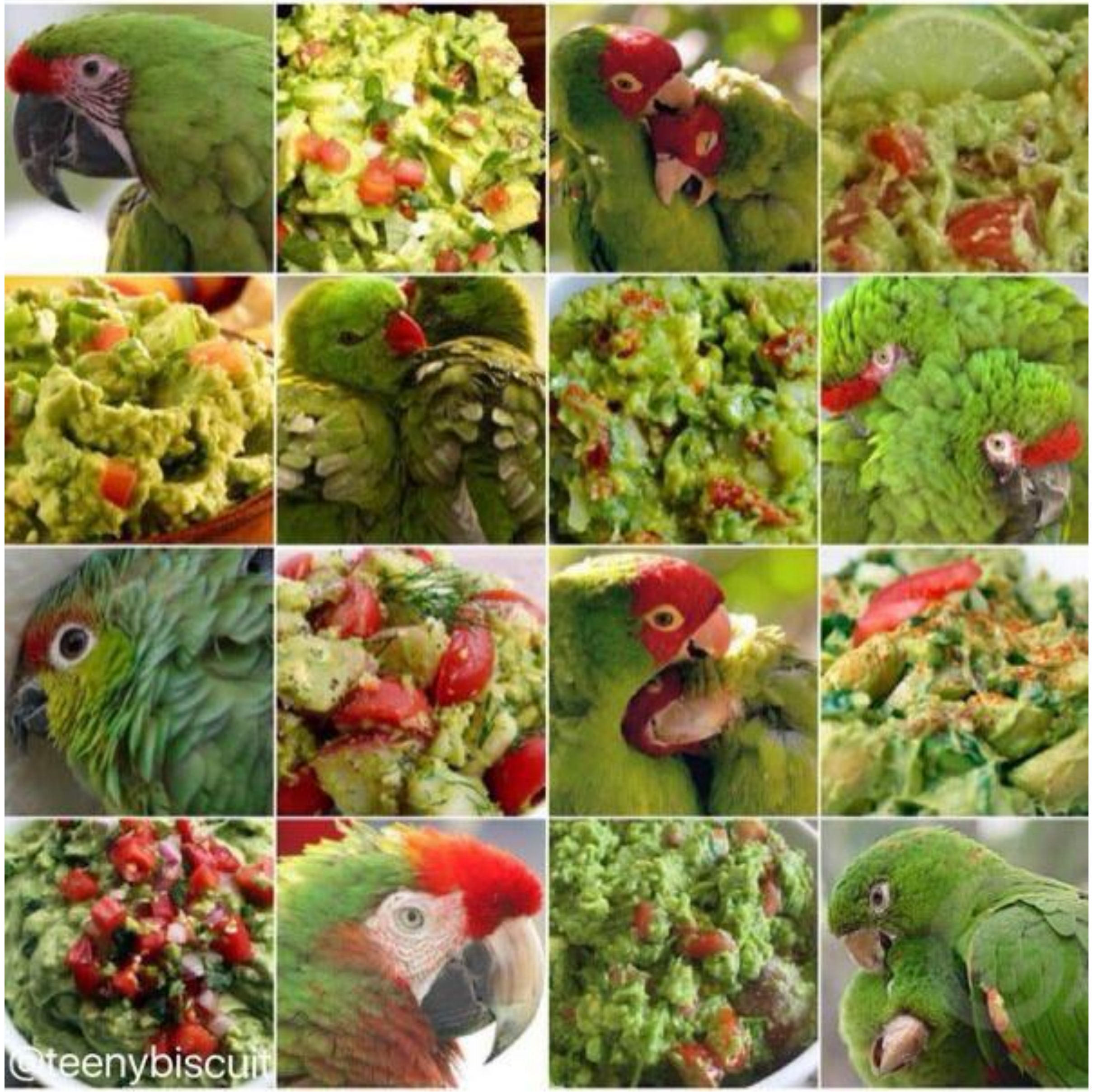
Chihuahua or muffin



Barn owl or apple



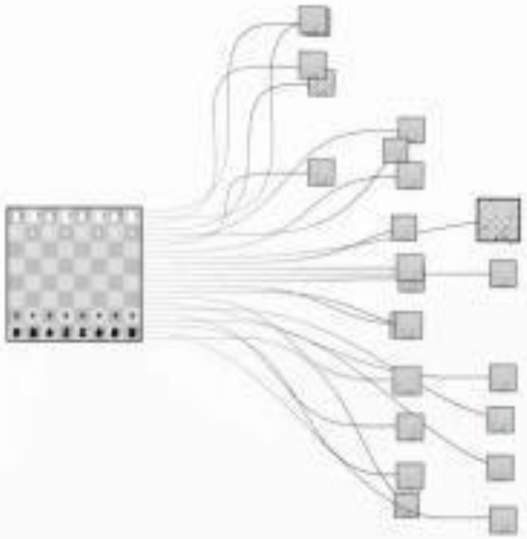
Parrot or guacamole



Raw chicken or Donald Trump

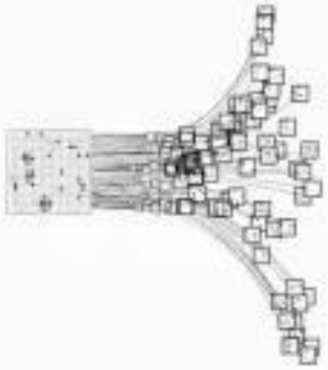
But, we human actually lose!

- If we want to make it hard for bots, it has to be hard for human as well.



Chess: 10^{47}

Deep Blue, Feb 10, 1996



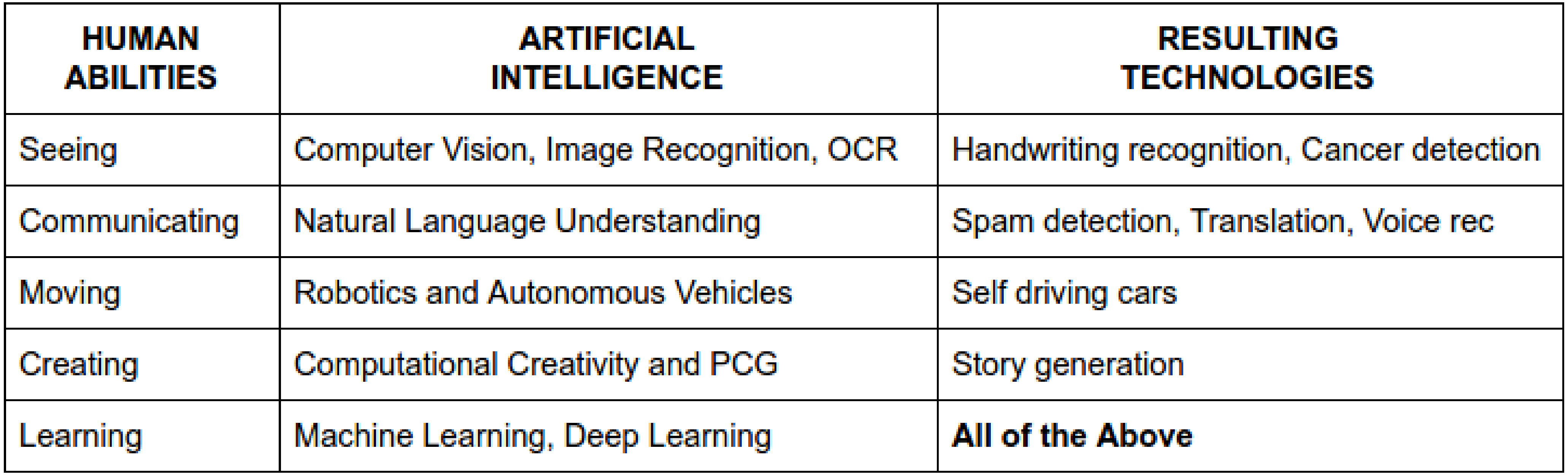
Go: 10^{170}

AlphaGo, March, 2016

We (will) lose on many specific tasks!

- Speech recognition
- Translation
- Self-driving
- …

Humans abilities vs Artificial intelligent



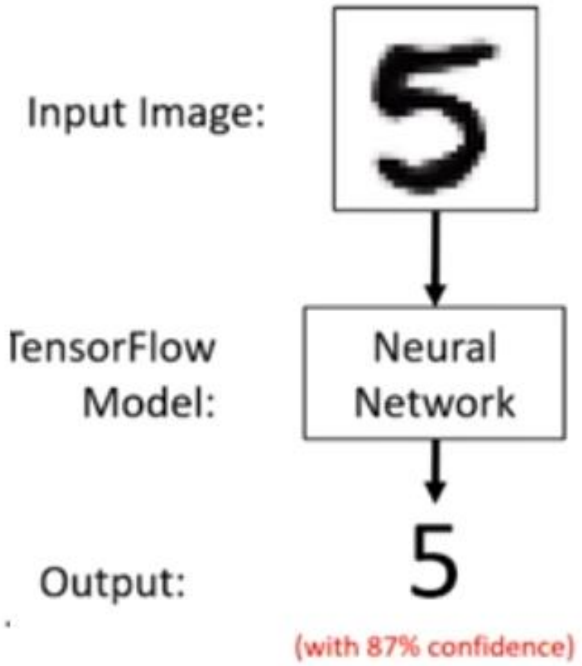
HUMAN ABILITIES	ARTIFICIAL INTELLIGENCE	RESULTING TECHNOLOGIES
Seeing	Computer Vision, Image Recognition, OCR	Handwriting recognition, Cancer detection
Communicating	Natural Language Understanding	Spam detection, Translation, Voice rec
Moving	Robotics and Autonomous Vehicles	Self driving cars
Creating	Computational Creativity and PCG	Story generation
Learning	Machine Learning, Deep Learning	**All of the Above**

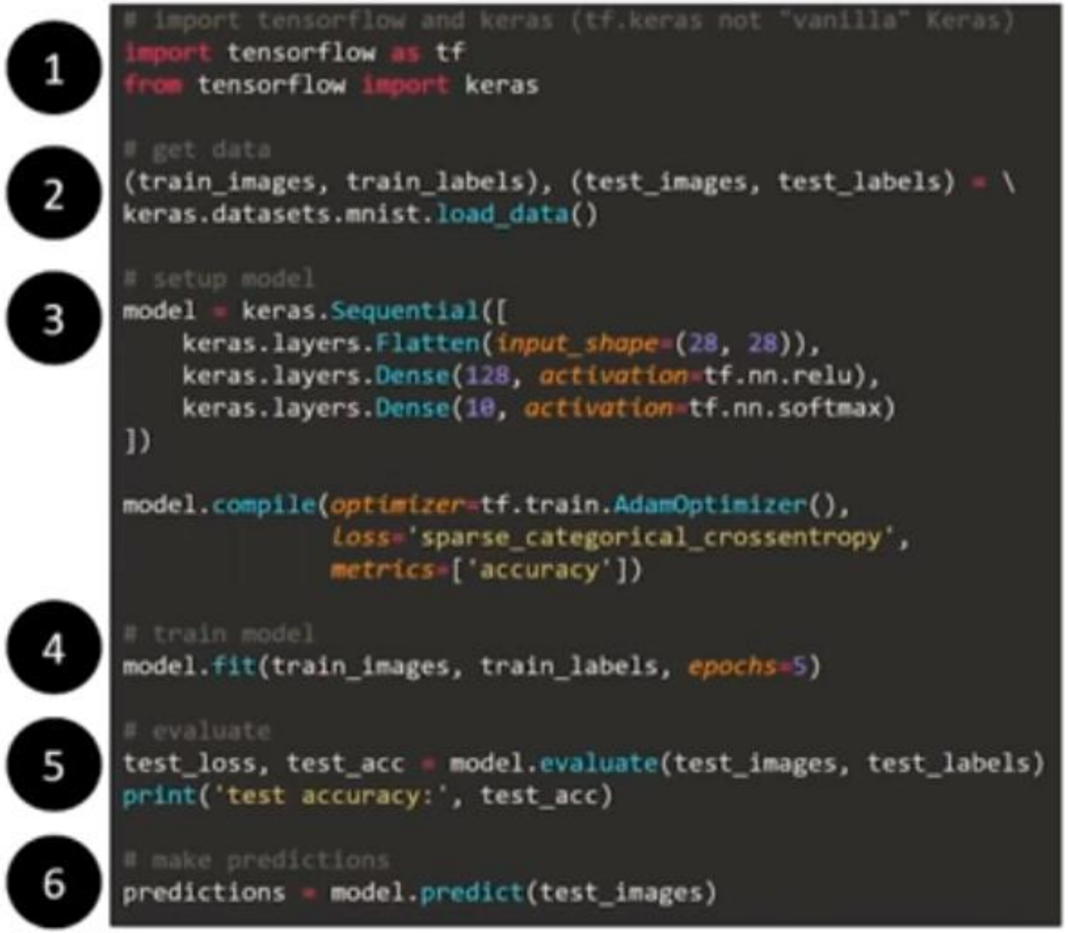
Machine Learning: How it looks like in the reality !

- Human can learn from very few examples
- Machine (in most cases) need thousands / million of examples

Machine learning is so cool for so many problems…

An algorithm of Machine learning





1

```
# import tensorflow and keras (tf.keras not "vanilla" Keras)
import tensorflow as tf
from tensorflow import keras
```

2

```
# get data
(train_images, train_labels), (test_images, test_labels) = \
keras.datasets.mnist.load_data()
```

3

```
# setup model
model = keras.Sequential([
    keras.layers.Flatten(input_shape=(28, 28)),
    keras.layers.Dense(128, activation=tf.nn.relu),
    keras.layers.Dense(10, activation=tf.nn.softmax)
])

model.compile(optimizer=tf.train.AdamOptimizer(),
              loss='sparse_categorical_crossentropy',
              metrics=['accuracy'])
```

4

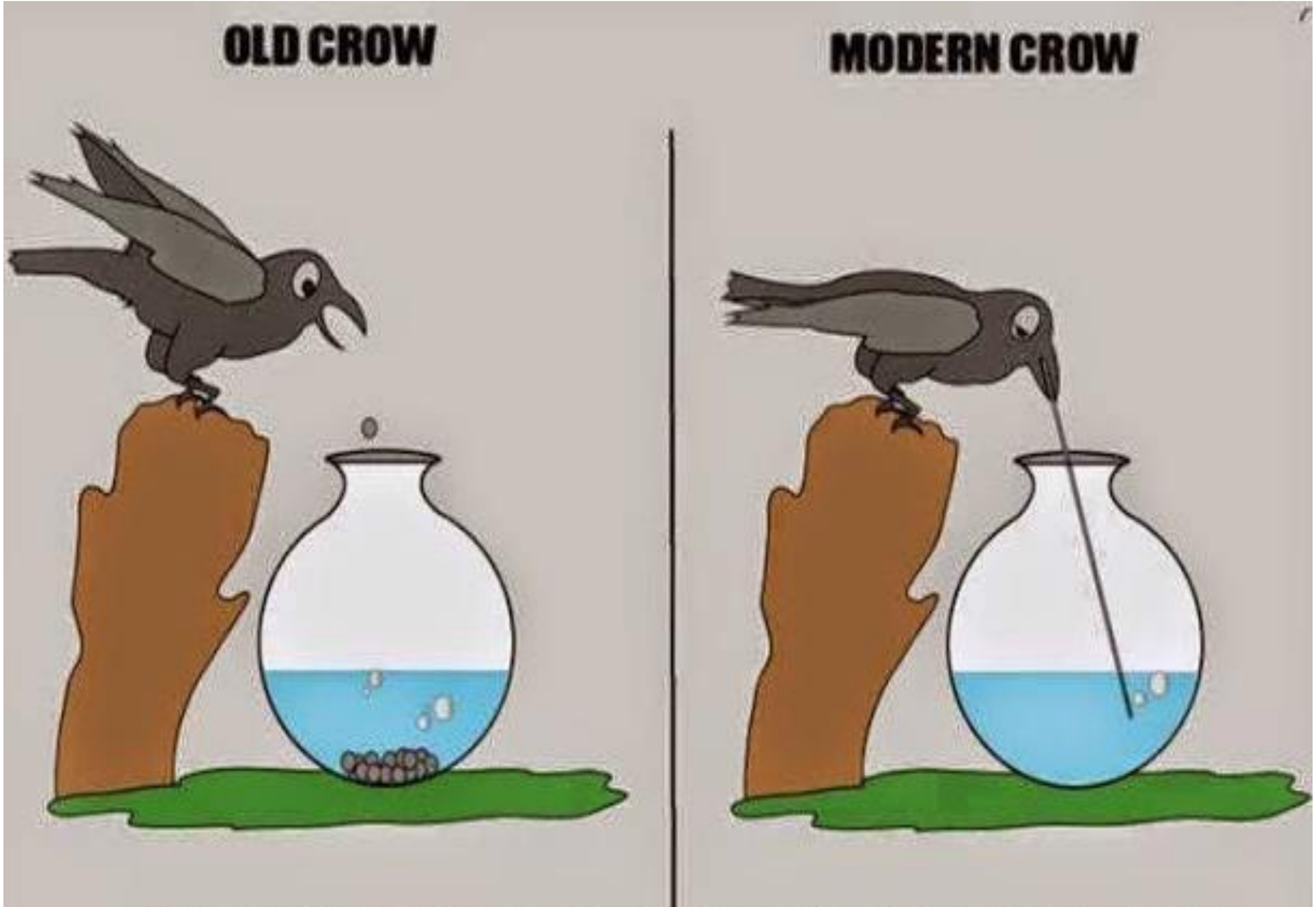
```
# train model
model.fit(train_images, train_labels, epochs=5)
```

5

```
# evaluate
test_loss, test_acc = model.evaluate(test_images, test_labels)
print('test accuracy:', test_acc)
```

6

```
# make predictions
predictions = model.predict(test_images)
```



OLD CROW

MODERN CROW

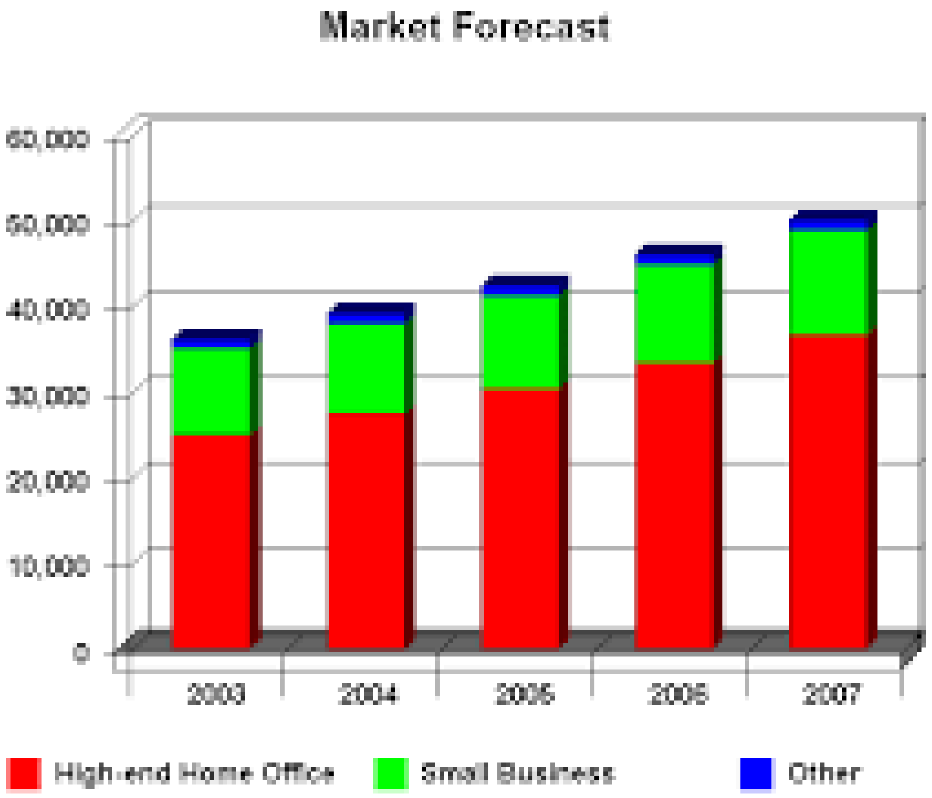
Update Yourself - It saves a lot of extra effort

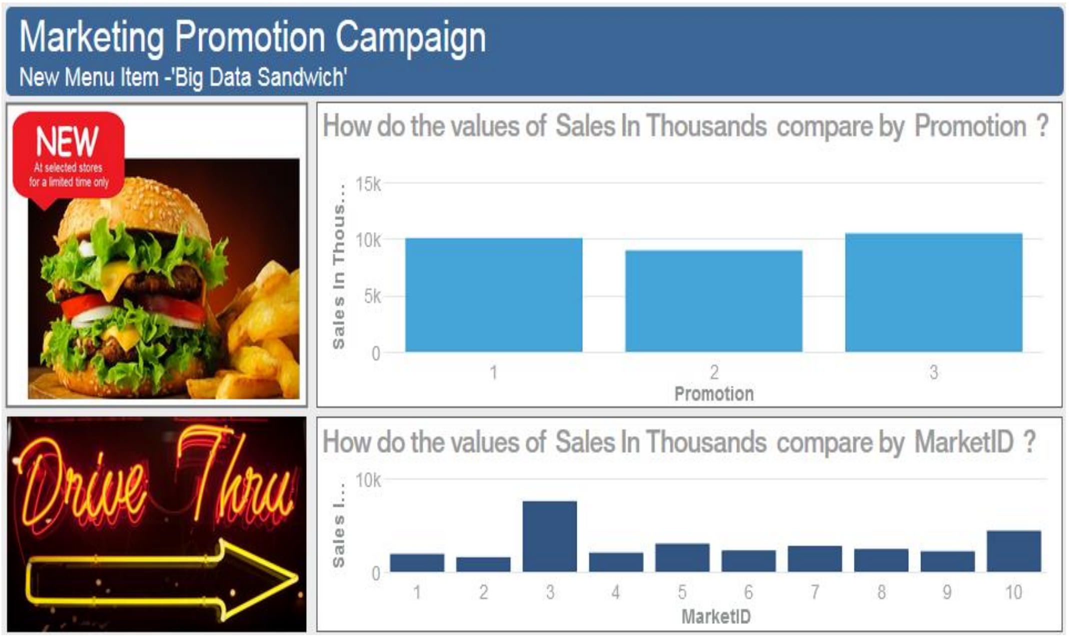
Data science components

Two components are needed to begin to ask whether data science can, yes or no, bring value and help to solve a problem: **data** and a well-defined **problem**.

Data science: realistic examples

- Predict sales of a marketing campaign




Marketing Promotion Campaign

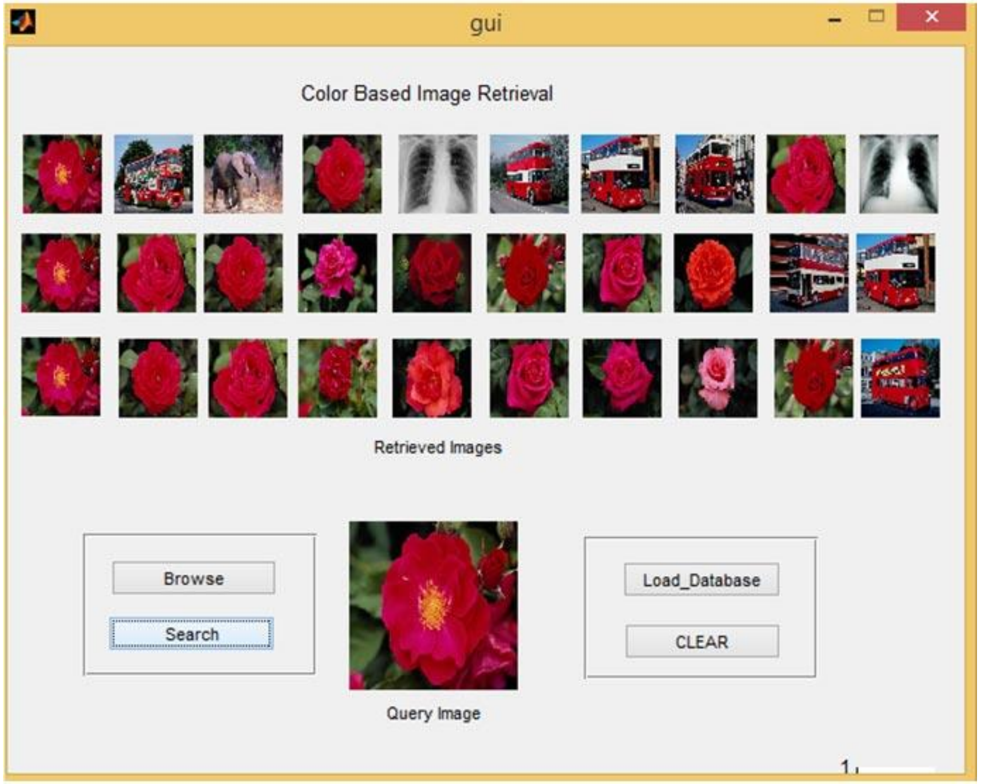
New Menu Item -'Big Data Sandwich'

How do the values of Sales In Thousands compare by Promotion ?

How do the values of Sales In Thousands compare by MarketID ?

Data science: realistic examples

- Identify if an image is already present in an existing image bank



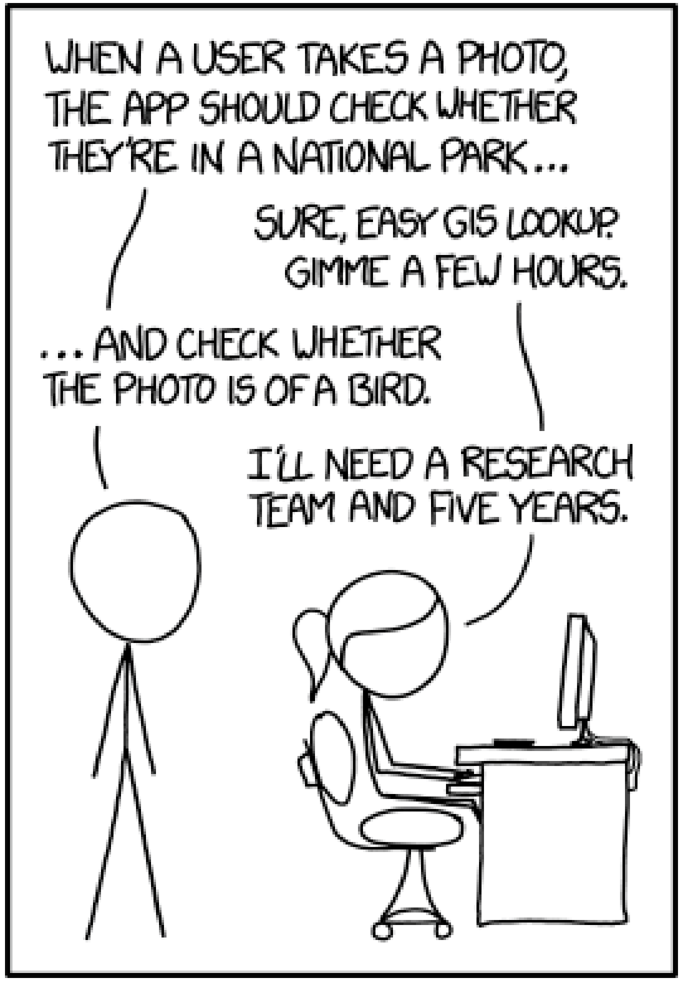
Data science: realistic examples

- Segment users of a website into several groups based on their behavior on it.

Data science: other examples

- Detect irony in a sentence

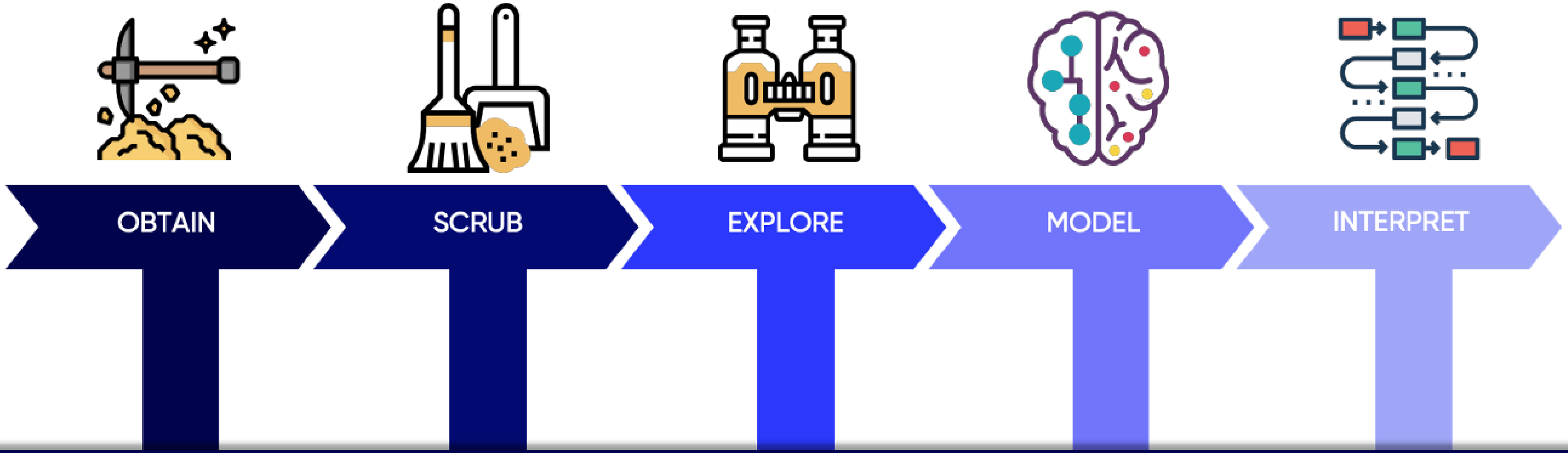
- Whether a correlation between two variables is causal or not


WHEN A USER TAKES A PHOTO, THE APP SHOULD CHECK WHETHER THEY'RE IN A NATIONAL PARK...
SURE, EASY GIS LOOKUP. GIMME A FEW HOURS.
...AND CHECK WHETHER THE PHOTO IS OF A BIRD.
I'LL NEED A RESEARCH TEAM AND FIVE YEARS.

IN CS, IT CAN BE HARD TO EXPLAIN THE DIFFERENCE BETWEEN THE EASY AND THE VIRTUALLY IMPOSSIBLE.

Data Science Process

LEAD



OBTAIN

SCRUB

EXPLORE

MODEL

INTERPRET

O

Gather data from relevant sources

S

Clean data to formats that machine understands

E

Find significant patterns and trends using statistical methods

M

Construct models to predict and forecast

N

Put the results into good use

First step : finding data 🔍

- **Mission**: explore all possible paths to recover the data.

- Everything must be sifted! Existing databases, alternative raw data (image, sound), and even the creation of new data acquisition channels. Try to find all the variables that directly or indirectly affect the phenomenon that interests you.

Data

- **From Database:**

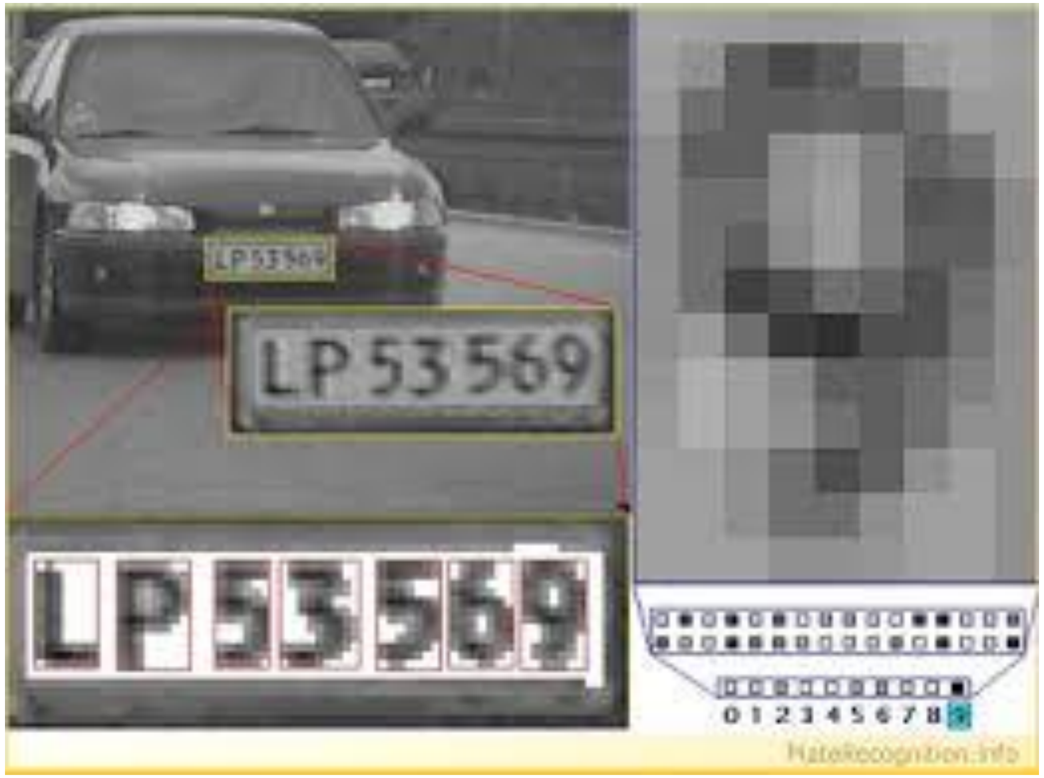
There is a plethora of technologies (from Hadoop to SQL) to ensure the recovery, storage and robustness of this data. These databases may include different types of information, many of which are generally specific to the business activity.

- **Using raw data**:

Other raw data, often more complex and requiring specific preprocessing to make them manipulable by the algorithms, that can serve as sources for a modeling problem.

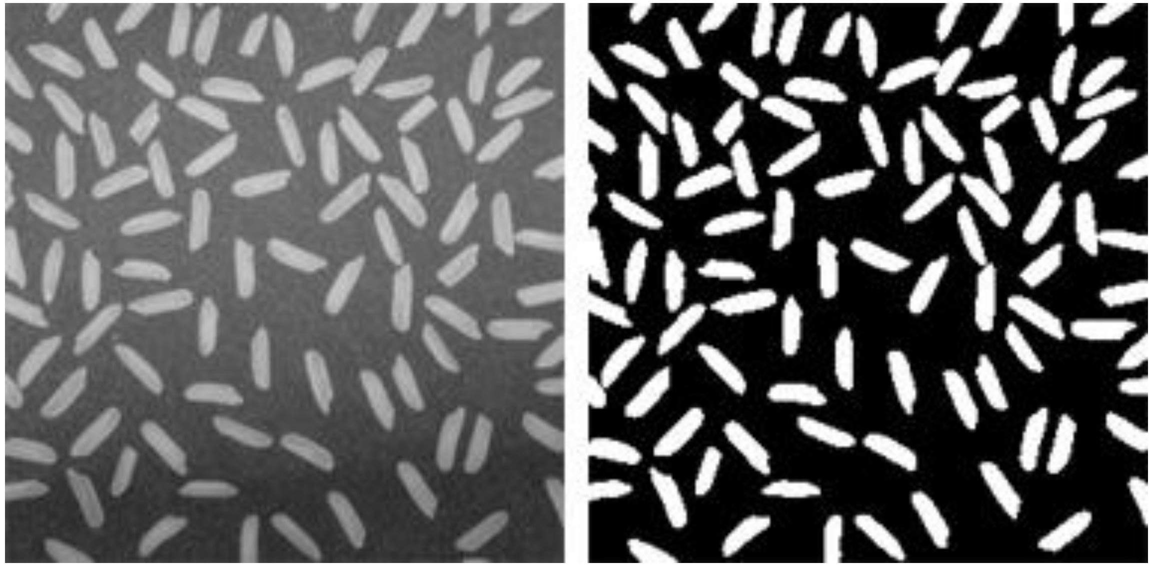
Data: types of data

- **Text / voice (NLP as Natural Language Processing)**
- Image/ sequences (Computer vision)
- Iot (Internet of things)



Data: examples of raw data

- Text / voice (NLP as Natural Language Processing)
- **Image/ sequences (Computer vision)**
- Iot (Internet of things)



Data: examples of raw data

- Text / voice (NLP as Natural Language Processing)
- Image/ sequences (Computer vision)
- **Iot (Internet of things)**

An example of a connected object: the Nest Business Smart Thermostat (credits: Nest) that optimizes electricity consumption by monitoring temperature, resident presence..

Step 2: Cleaning data

It must be ensured that the data are consistent, without outliers or missing values.

In this step, we :

- **Deal with missing values** (“?”, “N/A”, 0 or just a blank cell)
- **Data normalization**: is a technique that deal with how to adjust data to be more useful (same scale..)
- **Bin data:** is to classify data into bins (into different groups)

Step 3: Data exploration

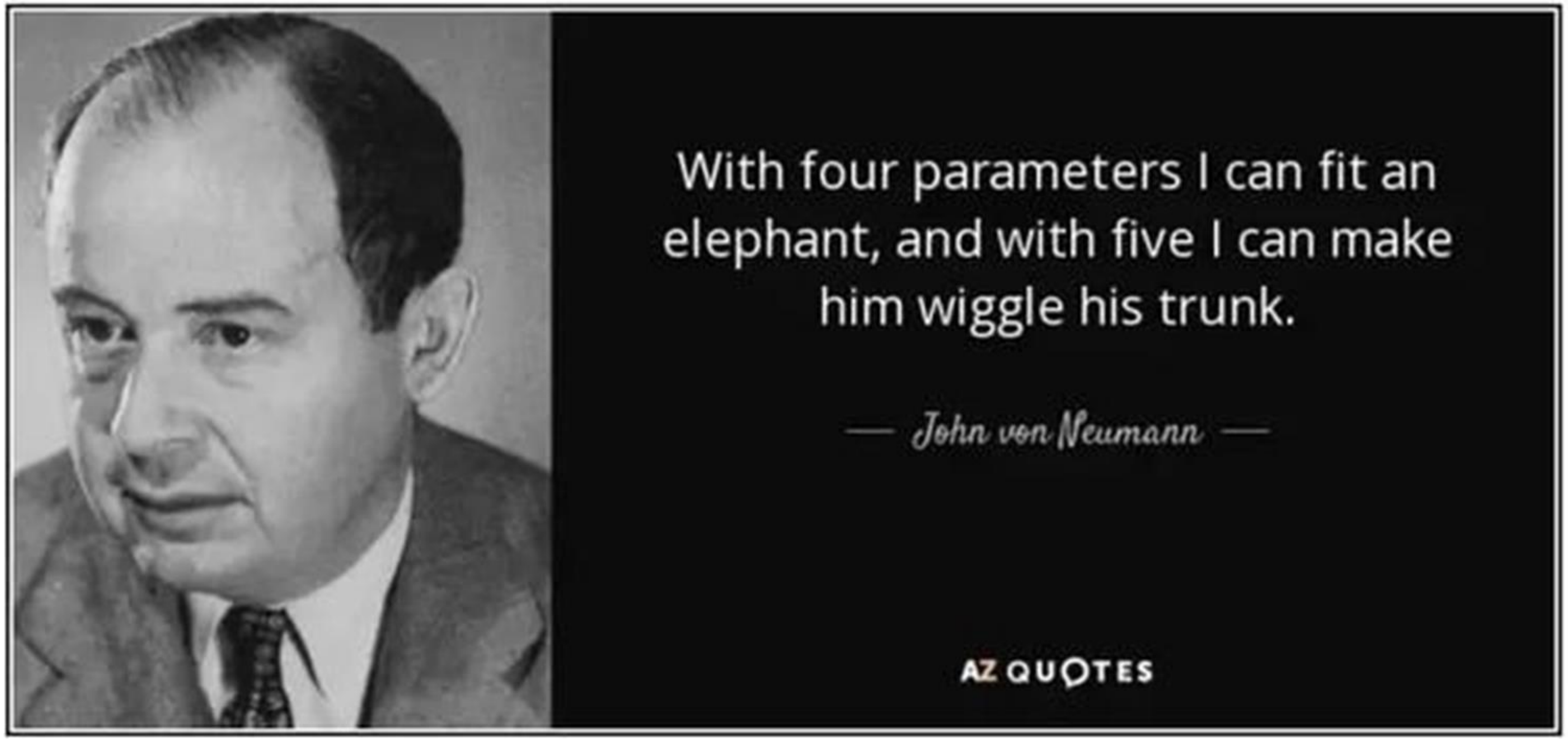
Do not hesitate to display all kinds of graphs, compare the different variables to each other, test correlation hypotheses, etc.

At the end of the exploration, you will need to be able to:

- Propose several hypotheses about the causes underlying the generation of the dataset: "following the exploration, there is clearly a relationship between X and Y".
- Propose several possible statistical modeling of the data (we will study this part in detail later in the course), which will solve the problem of departure considered.
- Propose, if necessary, new sources of data that would help to better understand the phenomenon (feature engineering)

Feature Engineering vs. Learning

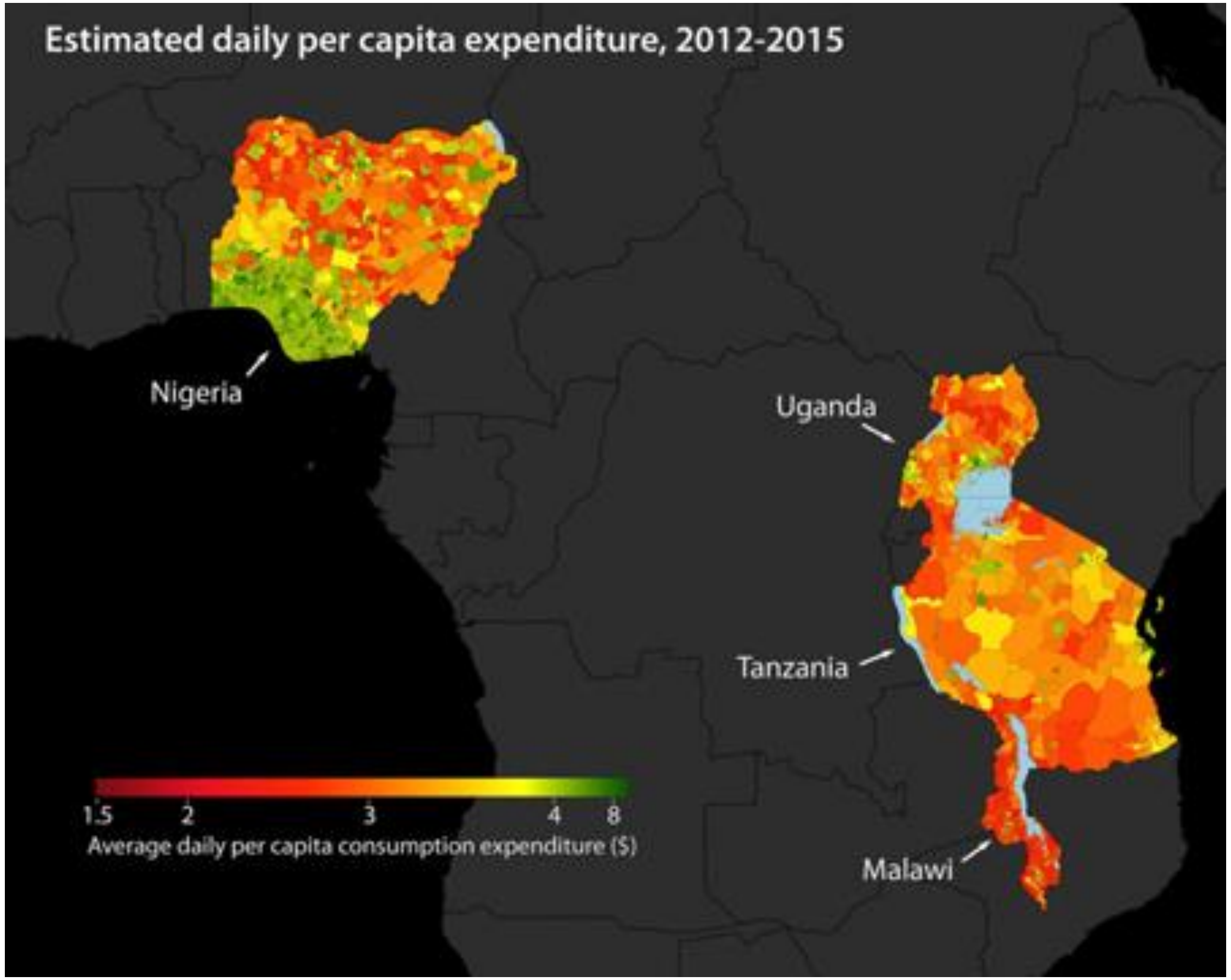
- Feature engineering is the process of using domain knowledge of the data to create features that make machine learning algorithms work.



With four parameters I can fit an elephant, and with five I can make him wiggle his trunk.

— John von Neumann —

Example: Poverty in Africa



Estimated daily per capita expenditure, 2012-2015
Nigeria
Uganda
Tanzania
Malawi
1.5
2
3
4
8
Average daily per capita consumption expenditure ($)

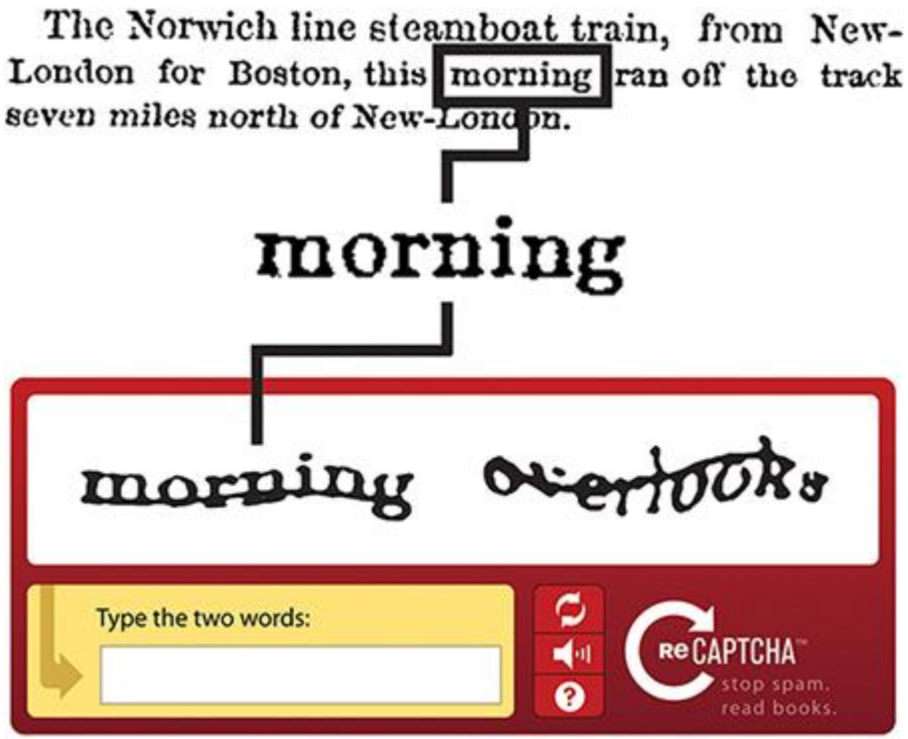
Researchers have used machine learning to map poverty areas automatically, simply from satellite images!

Credit. Neal Jean et al.

Example: CAPTCHAs for the automatic digitization of books

Luis von Ahn, entrepreneur and researcher, created a famous reCAPTCHA system that allowed both websites to validate that the forms were well filled by humans, and that fed at the same time the database of an algorithm of digitalization of books.

Thanks to the many examples provided directly by humans, the algorithm has finally had enough sample data to succeed then only to transcribe in text scanned images of books, with a very low error rate.

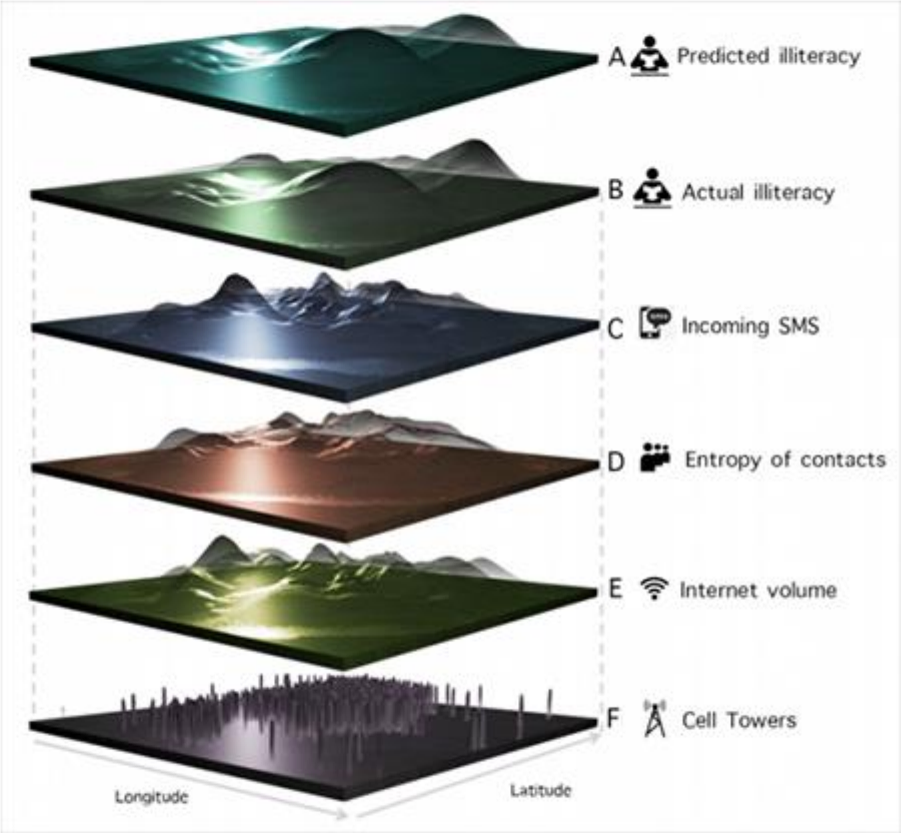


The Norwich line steamboat train, from New-London for Boston, this morning ran off the track seven miles north of New-London.

Example 3: illettrisme detection

The habits of use mobile phones to detect illiteracy

A Norwegian researcher used several types of mobile data (such as SMS, number of contacts, etc.) to detect illiterate people in developing countries.
Check out this article to learn more.



Each plan represents a different feature retrieved from a phone and captioned to the right. Source: arxiv.org/abs/1607.01337

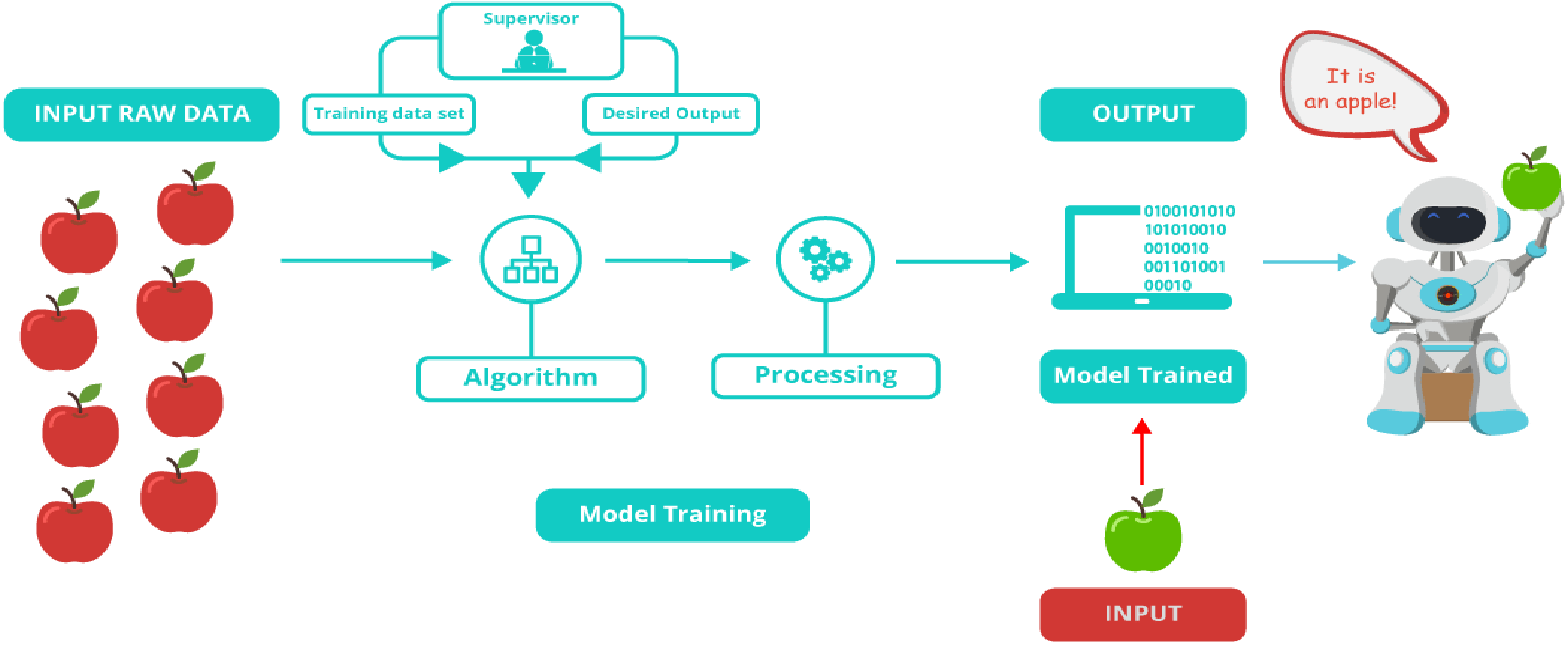
Step 4: Modelling using ML algorithms

There are four categories of ML:

- **Supervised learning**
- **Unsupervised learning.**
- **Semi-supervised Learning**
- **Reinforcement Learning**

Supervised learning

- In supervised learning, you will retrieve annotated data from their outputs to train the model, ie you have already associated a label or a target class and you want the algorithm to be able to predict it on new non-annotated data once trained.



Features : color, size, shape, grown in which part of the country, sold by which vendor, etc.
Output variables: the sweetness, juiciness, ripeness of that mango
Testing mode: Next time when you go shopping, you will measure the characteristics of the mangoes which you are purchasing(test data)and feed it to the Machine Learning algorithm. It will use the model which was computed earlier to predict if the mangoes are sweet, ripe and/or juicy.

Supervised learning Use case

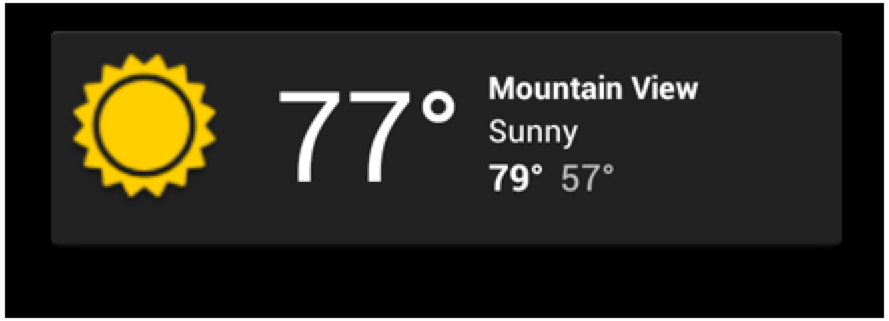
Hi. I'm Cortana.

Ask me a question!

Cortana or any speech automated system in your mobile phone trains your voice and then starts working based on this training.

Cortana
Cortana or any speech automated system in your mobile phone trains your voice and then starts working based on this training. This is an application of Supervised Learning

Weather Apps
Predicts the upcoming weather by analyzing the parameters for a given time on some prior knowledge (when its sunny, temperature is higher; when its cloudy, humidity is higher, etc.).



77°

Mountain View

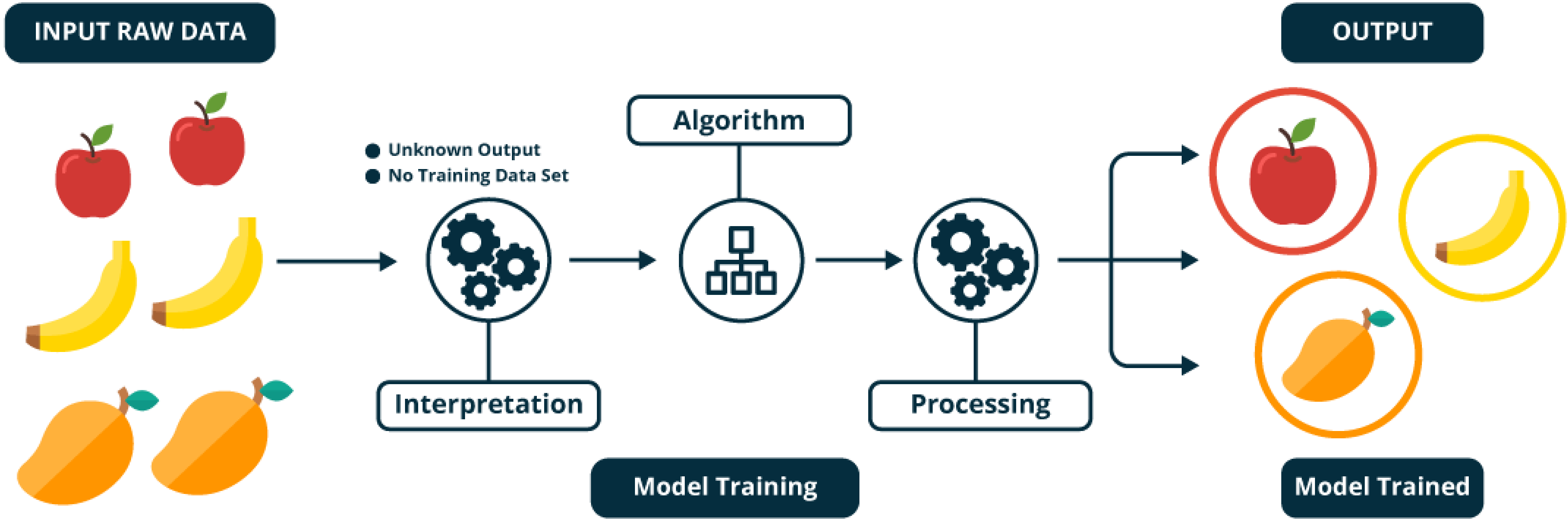
Sunny

79° 57°

Biometric Attendance

In Biometric Attendance you can train the machine with inputs of your biometric identity – it can be your thumb, iris or ear-lobe, etc. Once the machine is trained it can validate your future input and can easily identify you.

Unsupervised learning

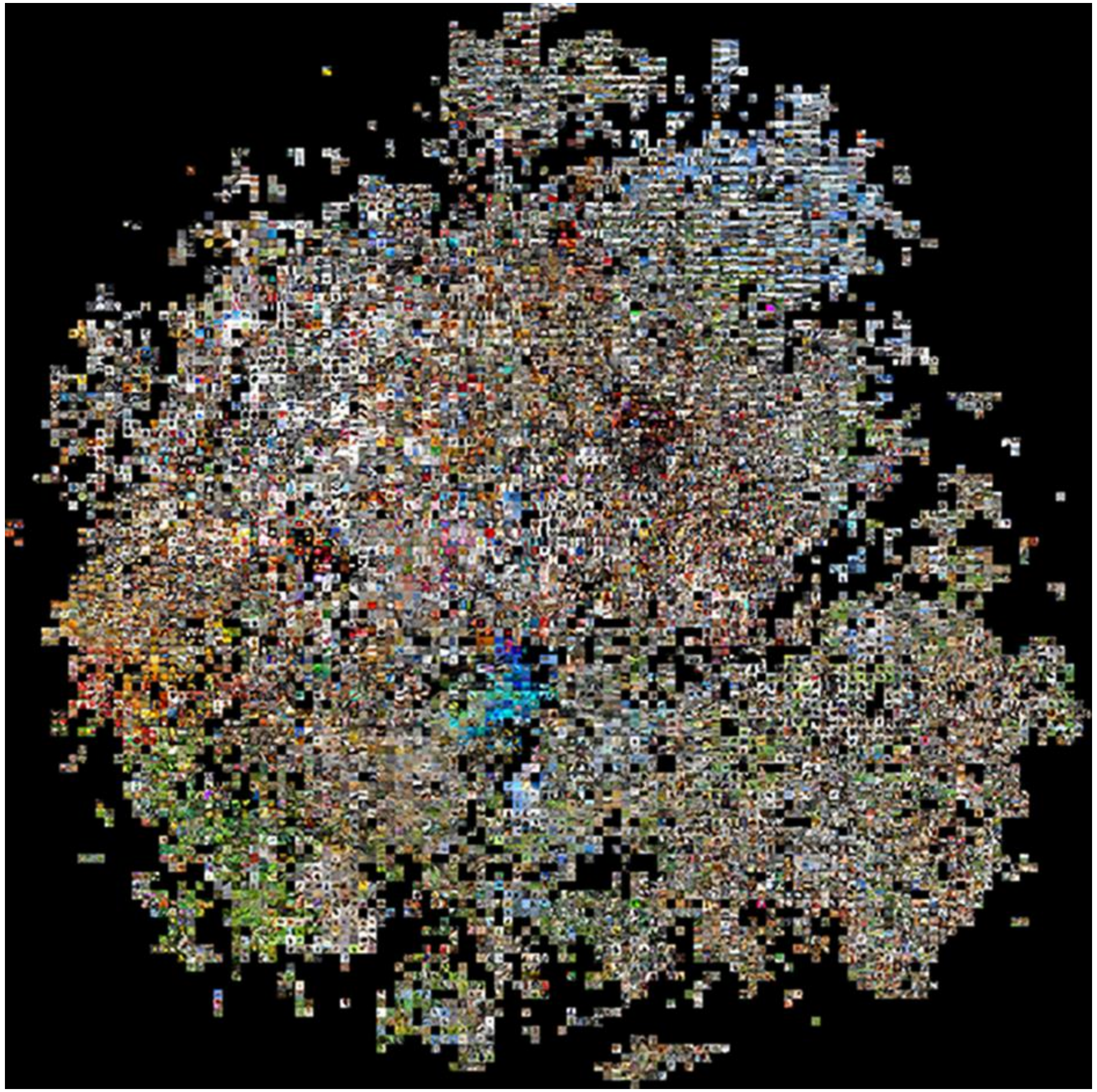


- In unsupervised learning, the input data is not annotated.

The training algorithm applies in this case to find only the similarities and distinctions within these data, and to group together those that share common characteristics. In our example, similar photos would be automatically grouped together within the same category.

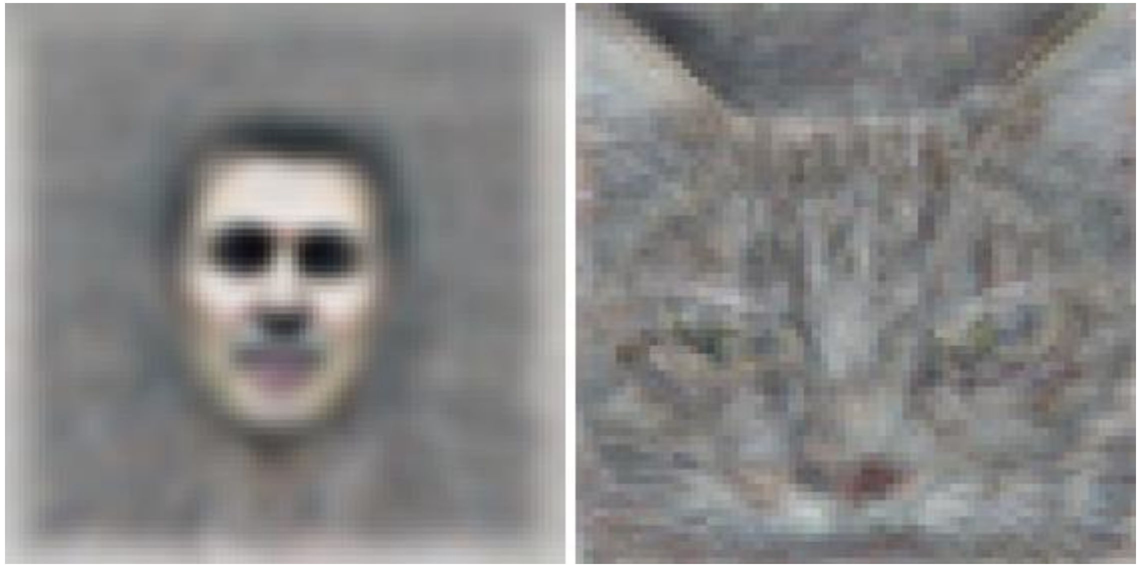
Unsupervised learning

- Automated image classifications by category.
- (link) https://cs.stanford.edu/people/karpathy/cnnembed/cnn_embed_6k.jpg



Unsupervised learning: example

- Researchers at Google Brain applied unsupervised learning algorithms a few years ago on Youtube videos, to see what this algorithm would learn to learn.
- Link



This image transcribes the internal representation of the concepts "**face**" and "**cat**" learned by an unsupervised algorithm from images extracted from **millions** of youtube videos (credits: Quoc V. Le et al.)

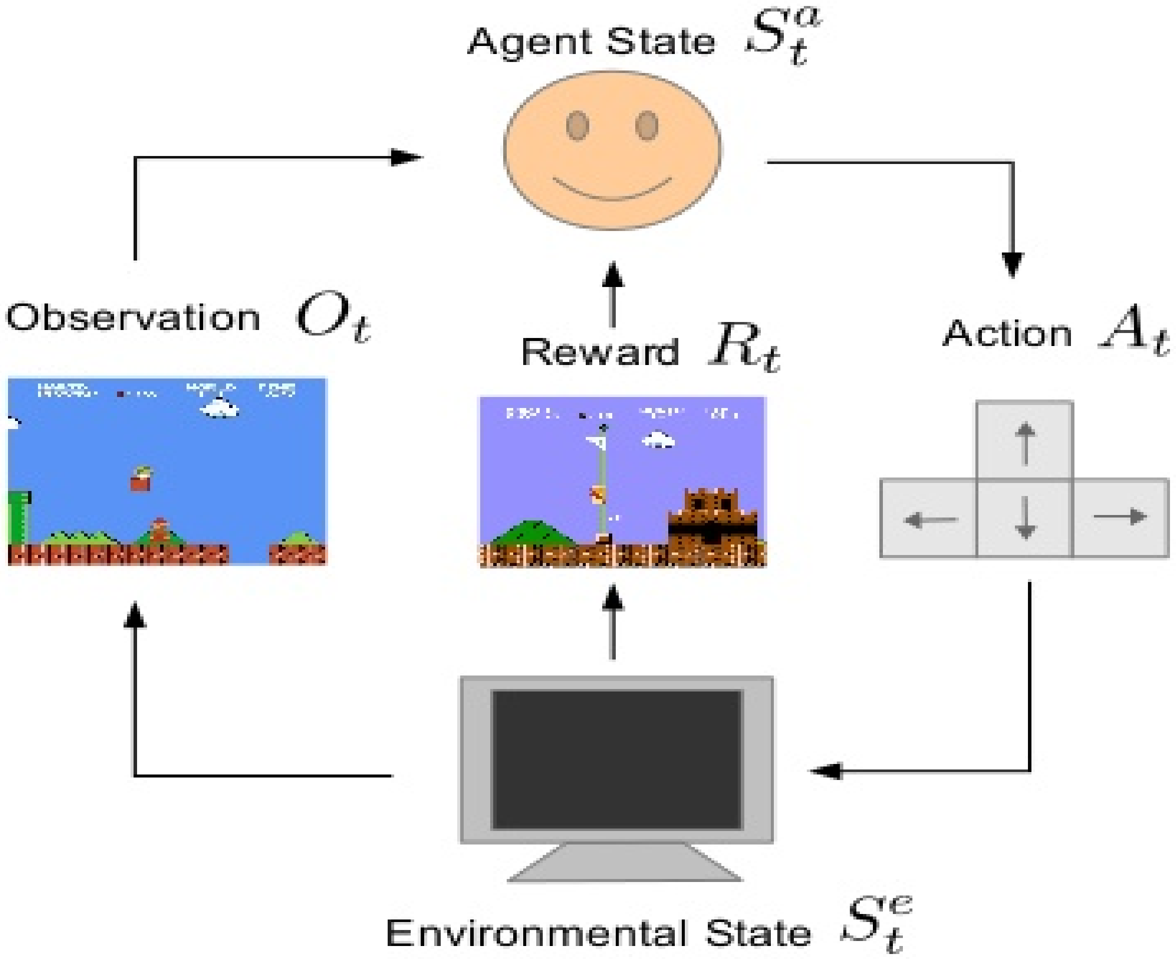
Other categories

There are actually two other families of algorithms that we will not detail (little used at the moment in practice), but you are free to inquire about:

- **The semi-supervised learning** which takes as input some annotated data and some not. These are very interesting methods that take advantage of both worlds (supervised and unsupervised), but of course bring their share of difficulties.

- **Reinforcement learning** which is based on a **cycle of experience / reward** and **improves performance** at each iteration. An analogy often cited is that of the dopamine cycle: a "good" experiment increases dopamine and therefore increases the likelihood that the agent will repeat the experiment.

Reinforcement Learning



- Rules of the game are unknown.
- No supervisor, only a reward signal.
- Feedback is delayed.
- Agent’s actions affect the subsequent data it receives.

Reinforcement learning: usecases

- Self driving

Robots

- Gaming

Recieved direct reward and discounted expected future reward (gamma=0.90)
State

Recommendation systems

NETFLIX

ML algorithms

The learning algorithm is the method by which the statistical model will be parameterized from the example data. There are many different algorithms! We will choose a particular type of **algorithm depending on the type of task** we want to perform and the **type of data available**

Exemples:

- Linear regression
- Knn
- Support Vector Machine (SVM)
- Neural networks
- Randon forests, decision tree..
- etc.

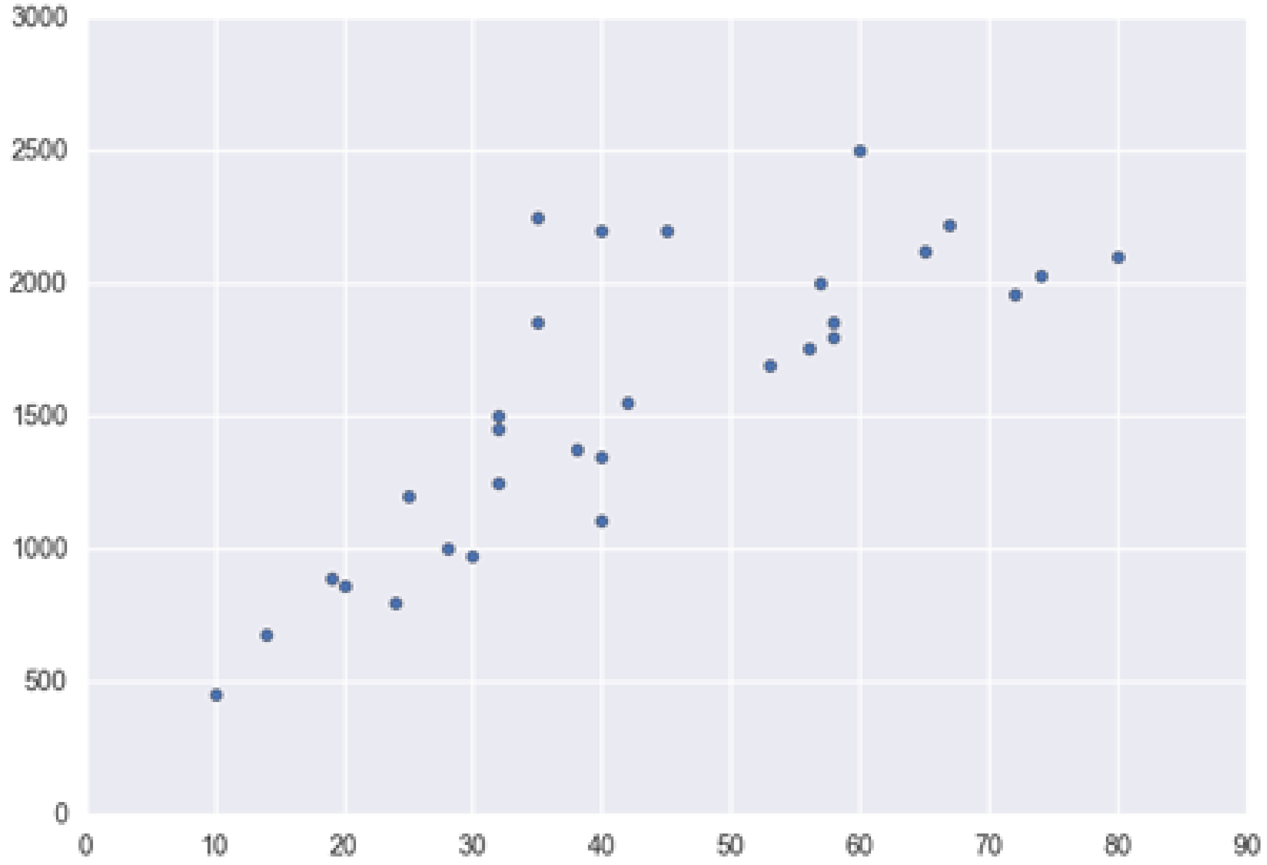
Linear regression: Rent modeling

- Imagine that you want to know if you are paying too much for your rent. You have recovered from a rental site about thirty prices of available rentals, as well as the associated area:

Rent (€ /month)	surface (m2)
1500	32
2120	65
2500	60

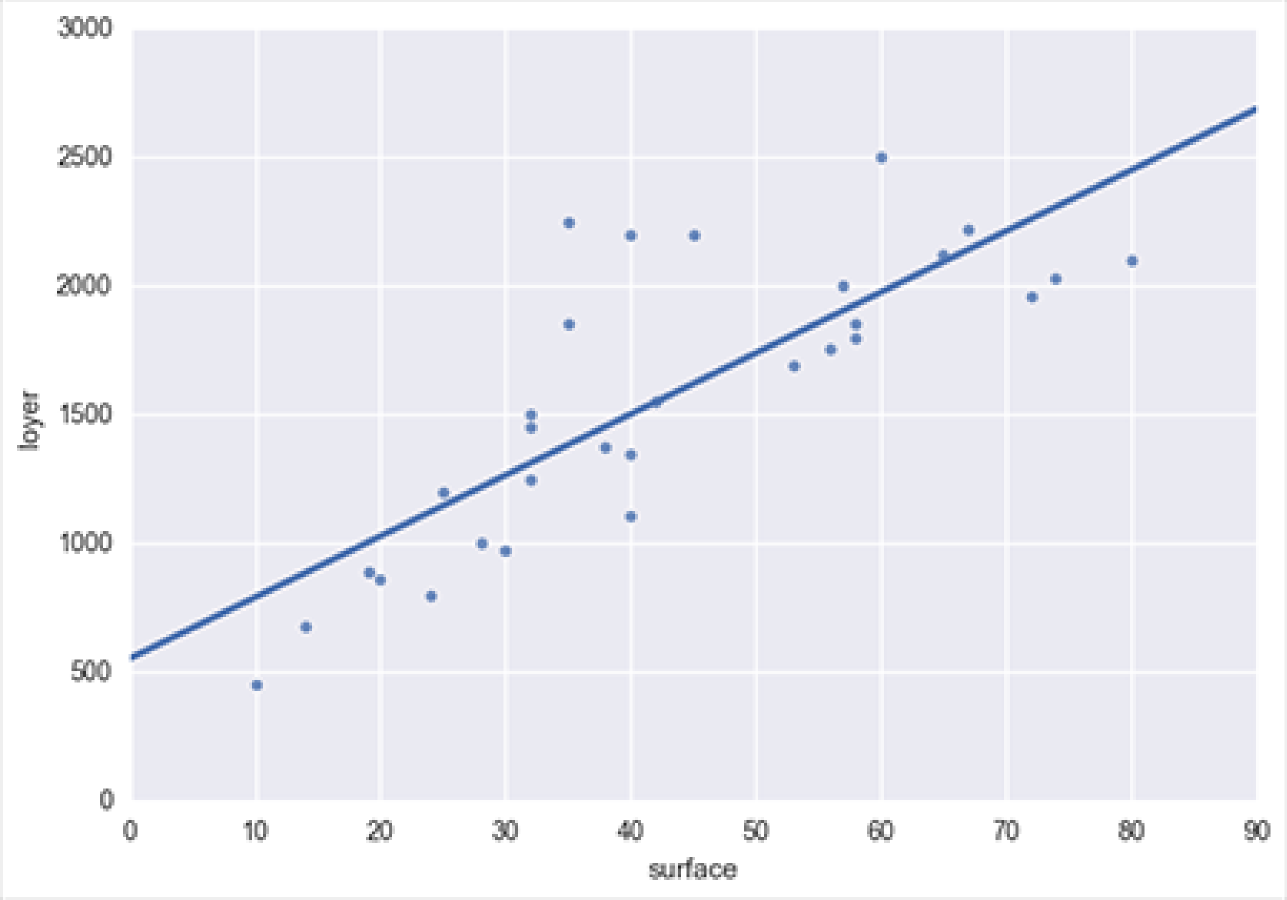
Example: Rent modelling

- As expected, there is a relatively linear increase in rent relative to the surface of the apartment. A first simple modeling of the phenomenon (the price of rent) would therefore simply be to consider the right "closest" to all points.



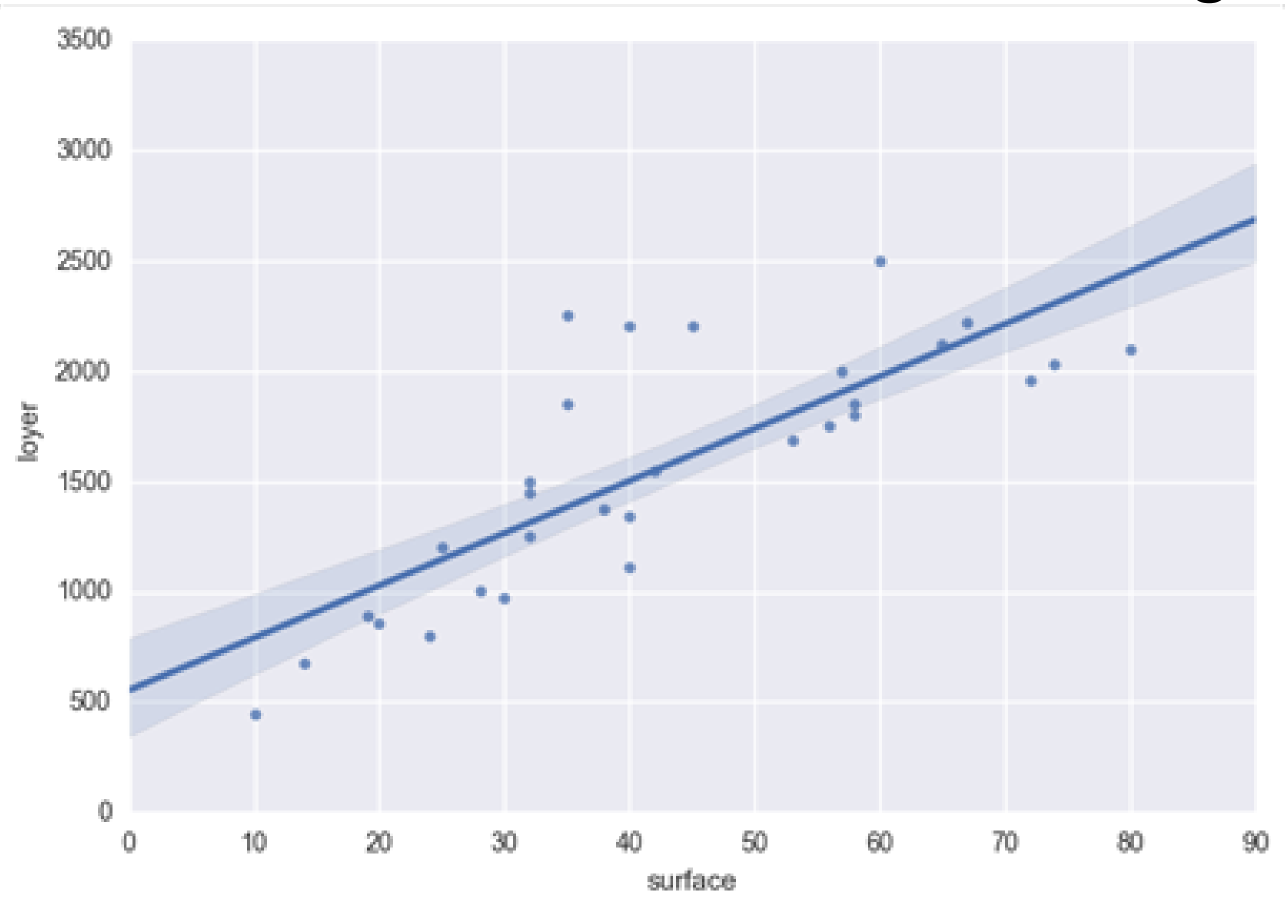
Example: Rent modelling

- The line represents our model of the phenomenon. Here, we can add the confidence interval in which we think the right is.



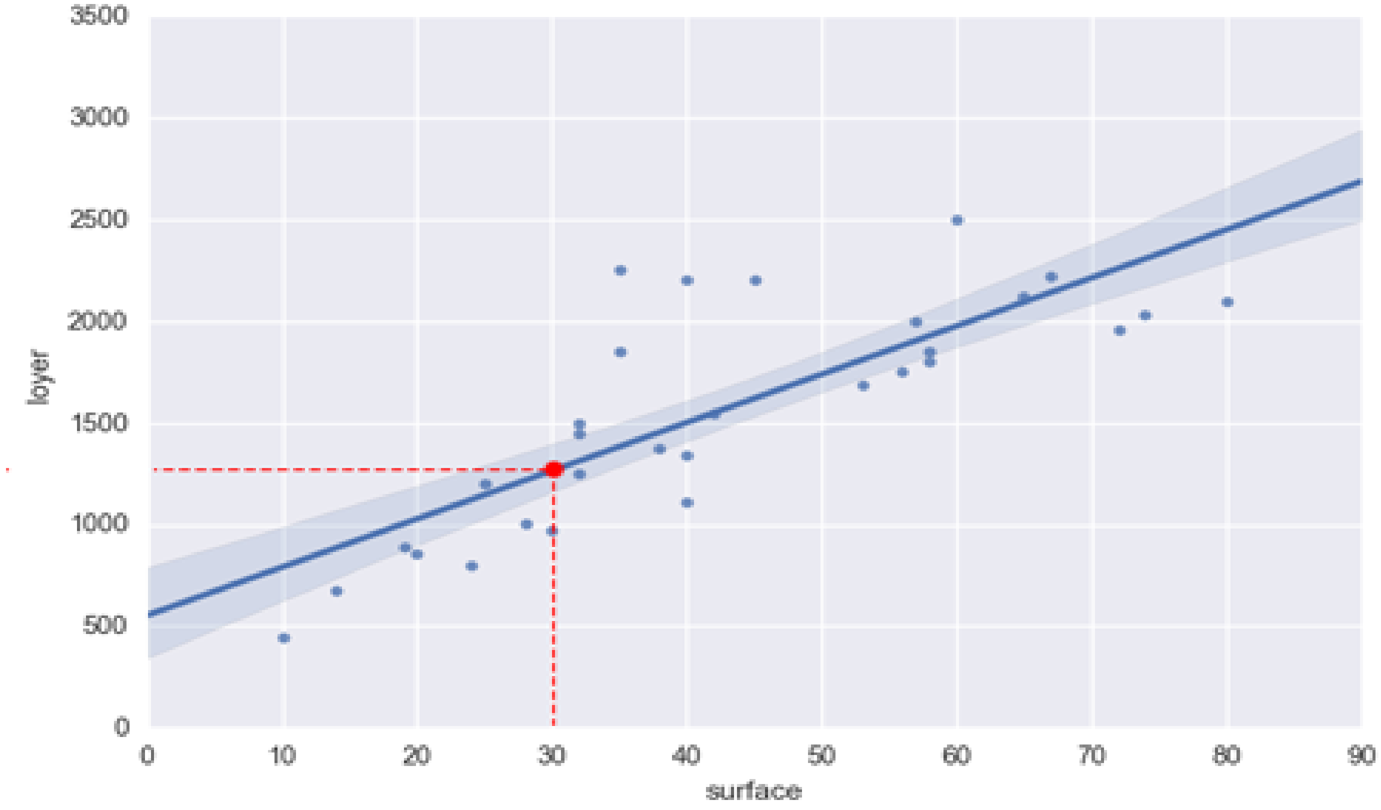
Example: Rent modelling

- The line represents our model of the phenomenon, to which we can add the confidence interval in which we think the right is.



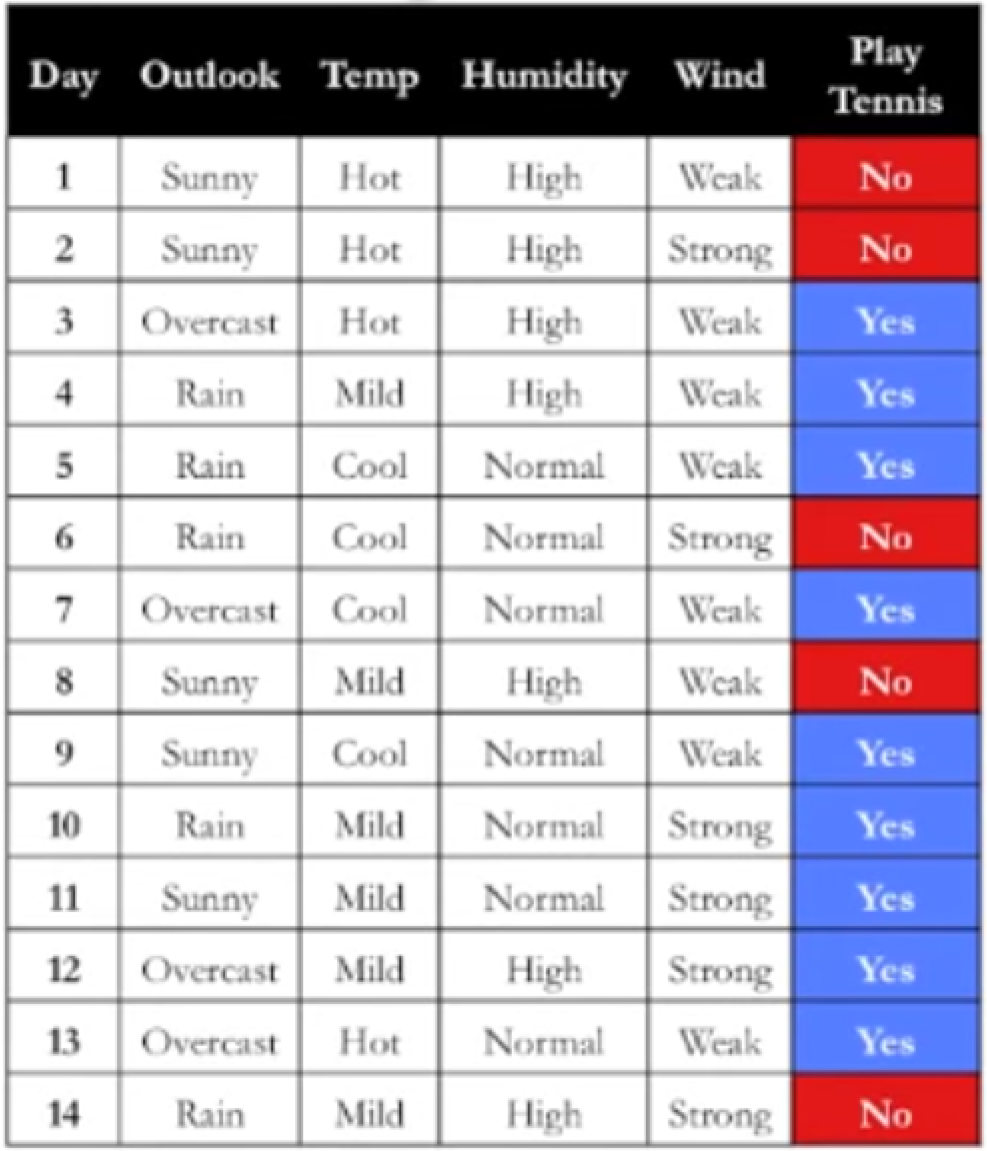
Exemple: Rent modeling

- According to our model, an apartment has a surface of 30 meters square (point in red), a legitimate estimation of the rent would be around 1300 euros



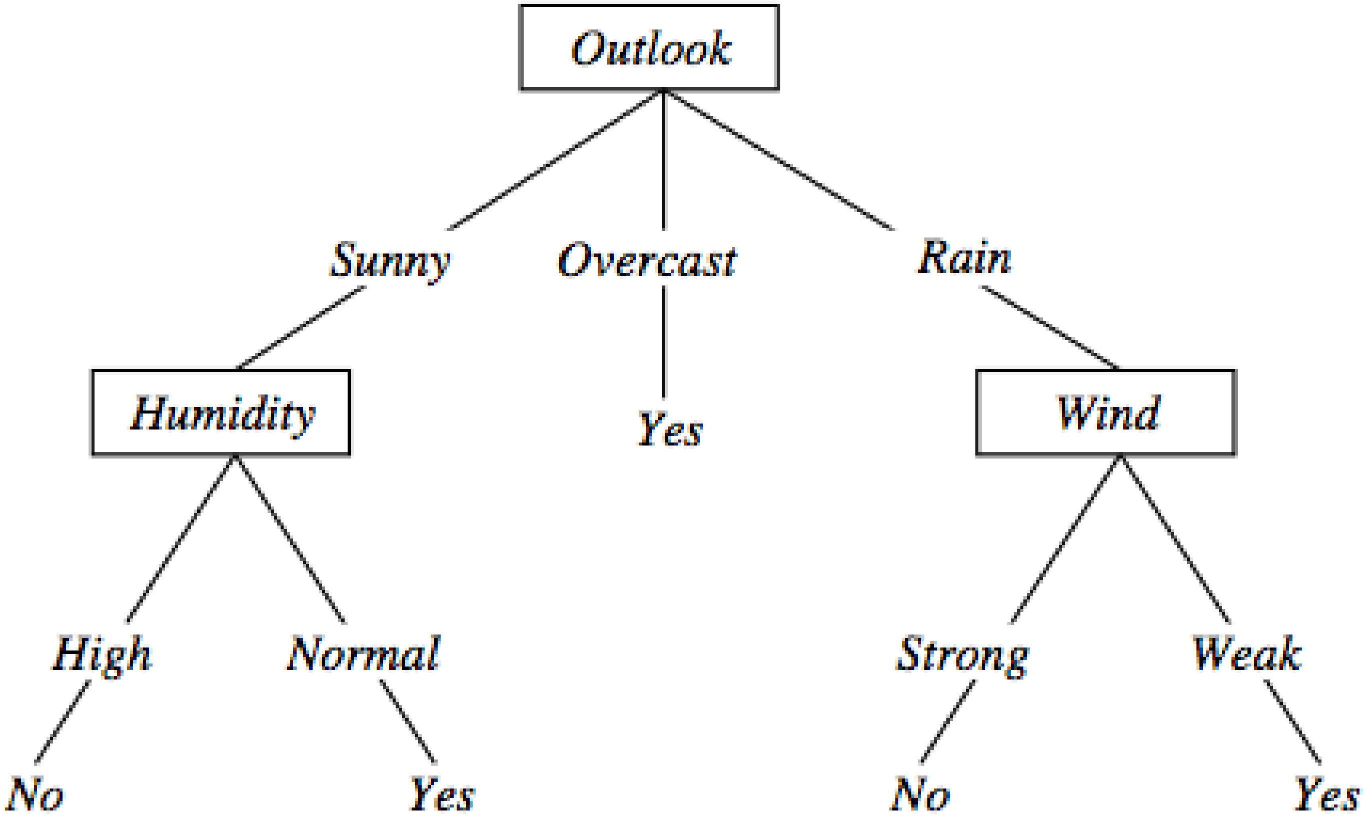
Decision tree: example

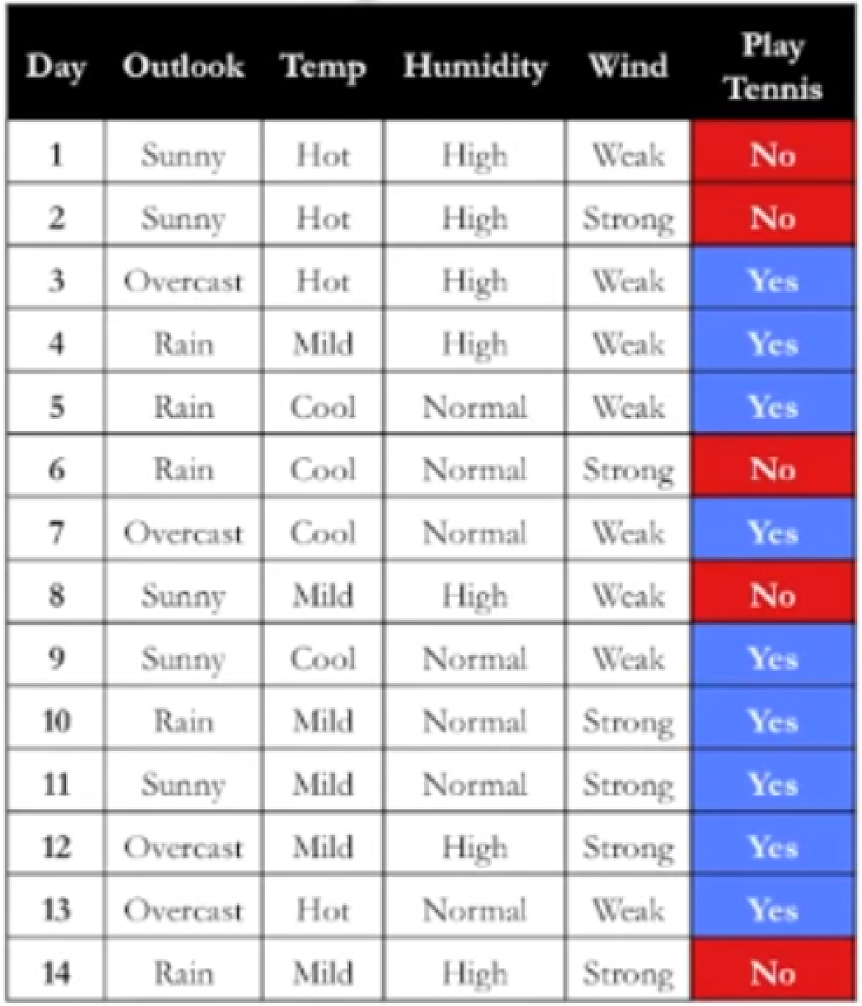
- Federer will play the game or not according to the **weather**?



Day	Outlook	Temp	Humidity	Wind	Play Tennis
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Weak	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Strong	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

Decision tree representation (PlayTennis)





Day	Outlook	Temp	Humidity	Wind	Play Tennis
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Weak	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Strong	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

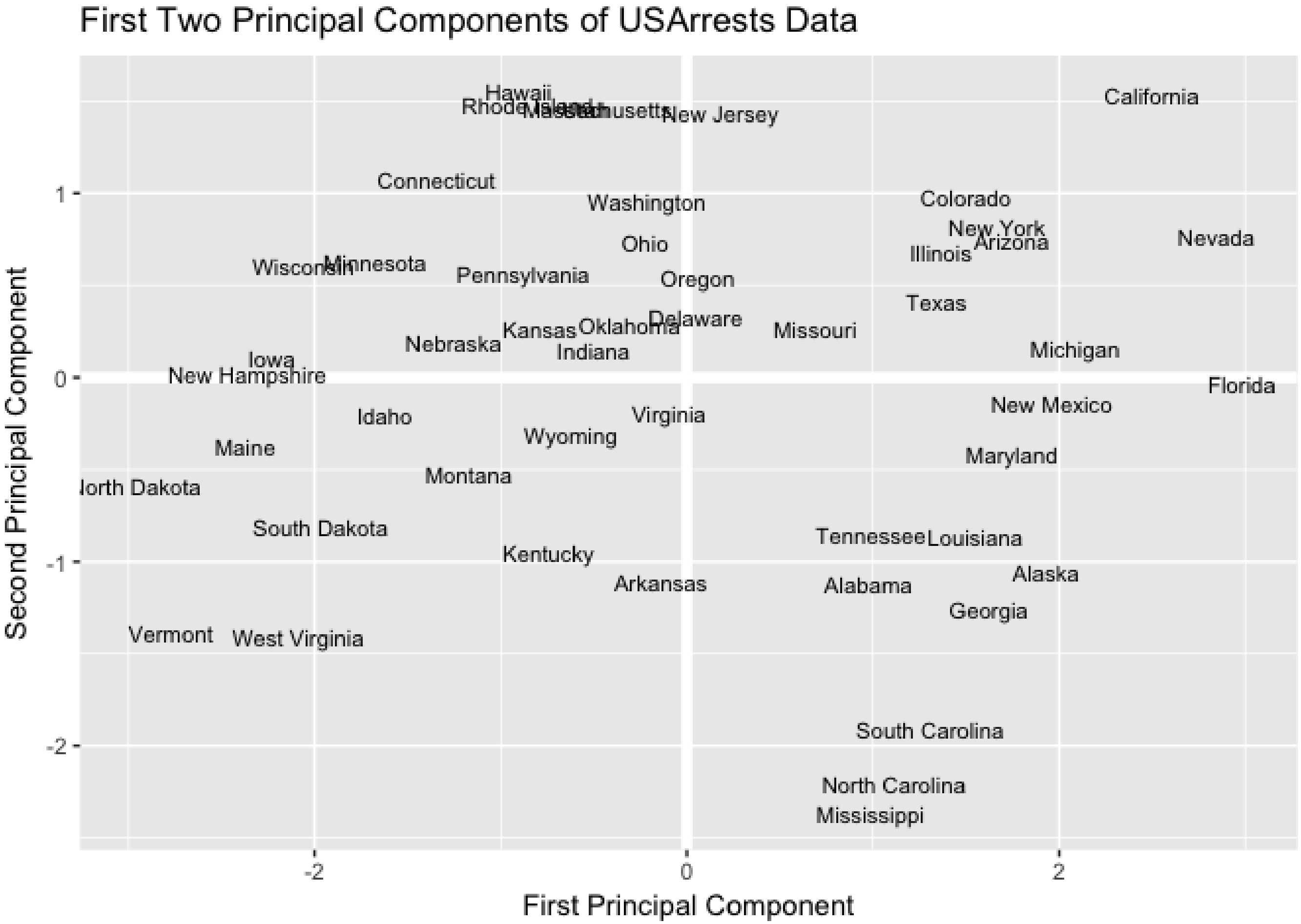
$\langle \mathit{Outlook}=\mathit{Sunny}, \mathit{Temp}=\mathit{Hot}, \mathit{Humidity}=\mathit{High}, \mathit{Wind}=\mathit{Strong} \rangle$ No

Principal component Analysis (PCA)

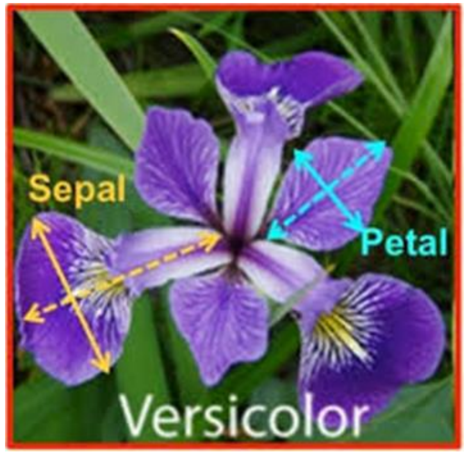
Principal Component Analysis, or PCA, is a dimensionality-reduction method that is often used to reduce the dimensionality of large data sets, by transforming a large set of variables into a smaller one that still contains most of the information in the large set.

Exemple: **Crimes in USA**

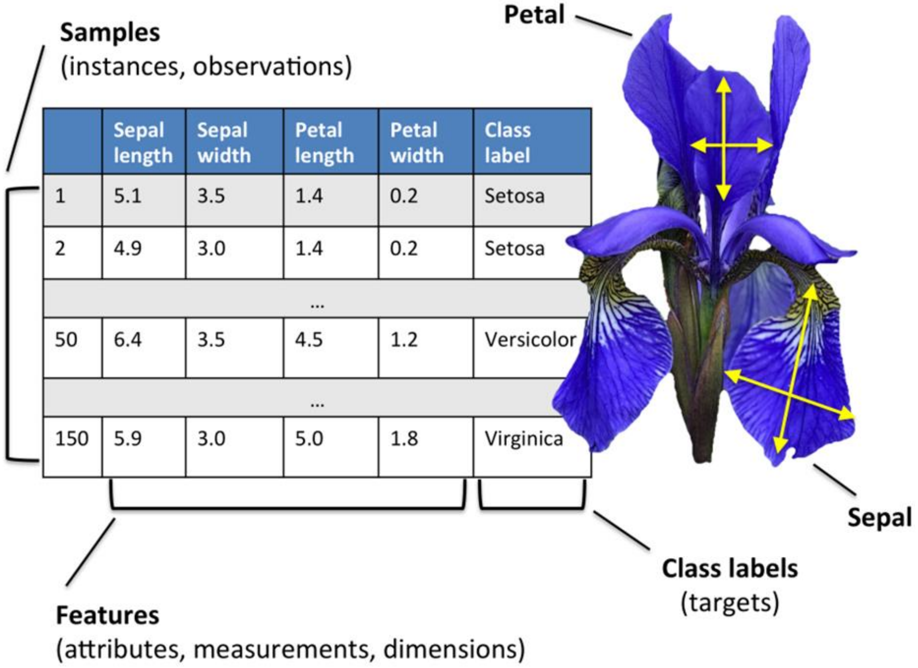
DEPARTMENT OF DEFENSE
UNITED STATES OF AMERICA
RESOLUTION OF THE DEPARTMENT OF DEFENSE OF THE UNITED STATES OF AMERICA
AND THE MINISTRY OF NATIONAL DEFENSE OF THE REPUBLIC OF KOREA:
GUIDING PRINCIPLES FOLLOWING THE TRANSITION OF WARTIME OPERATIONAL CONTROL



Support Vector Machine (SVM)



Example: Classification of iris flowers on the iris dataset (sepal and petal leafs).



Samples
(instances, observations)

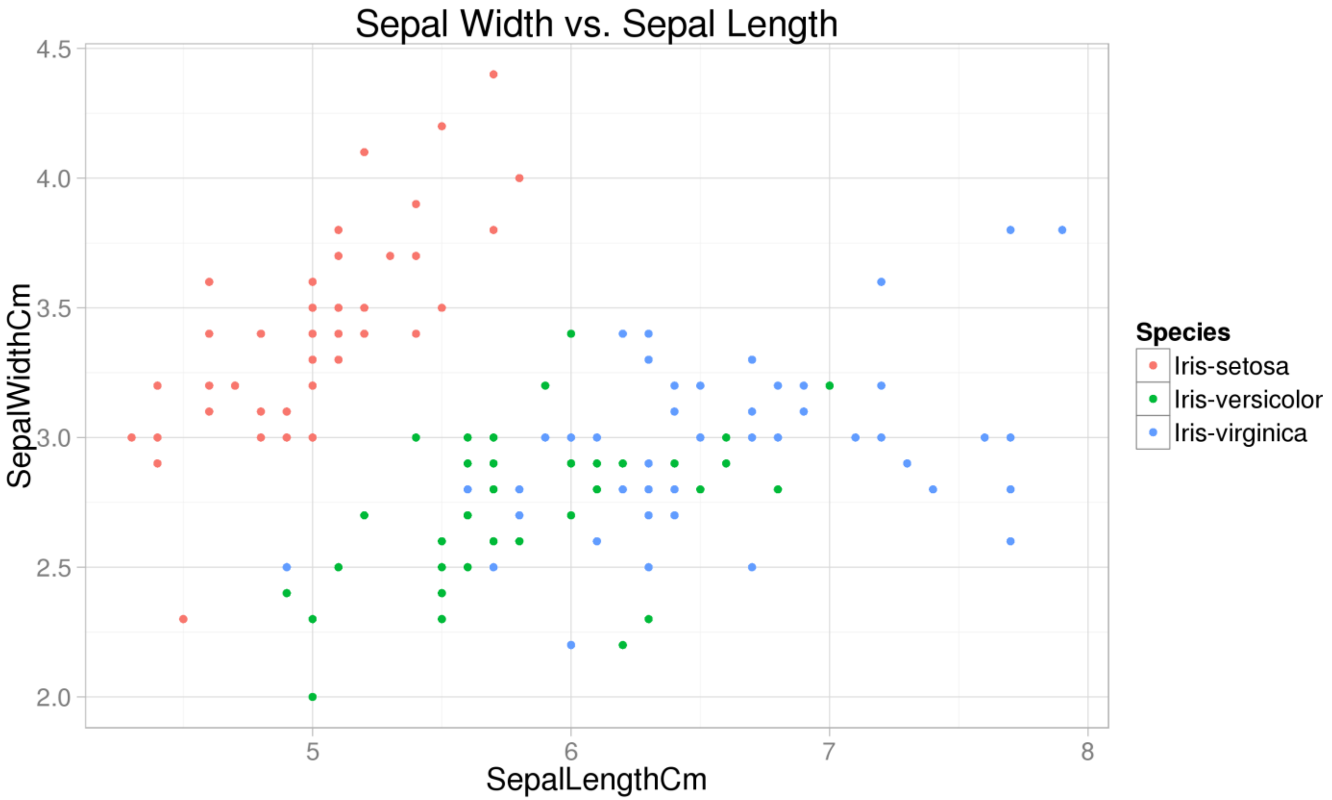
Petal

	Sepal length	Sepal width	Petal length	Petal width	Class label
1	5.1	3.5	1.4	0.2	Setosa
2	4.9	3.0	1.4	0.2	Setosa
			...		
50	6.4	3.5	4.5	1.2	Versicolor
			...		
150	5.9	3.0	5.0	1.8	Virginica

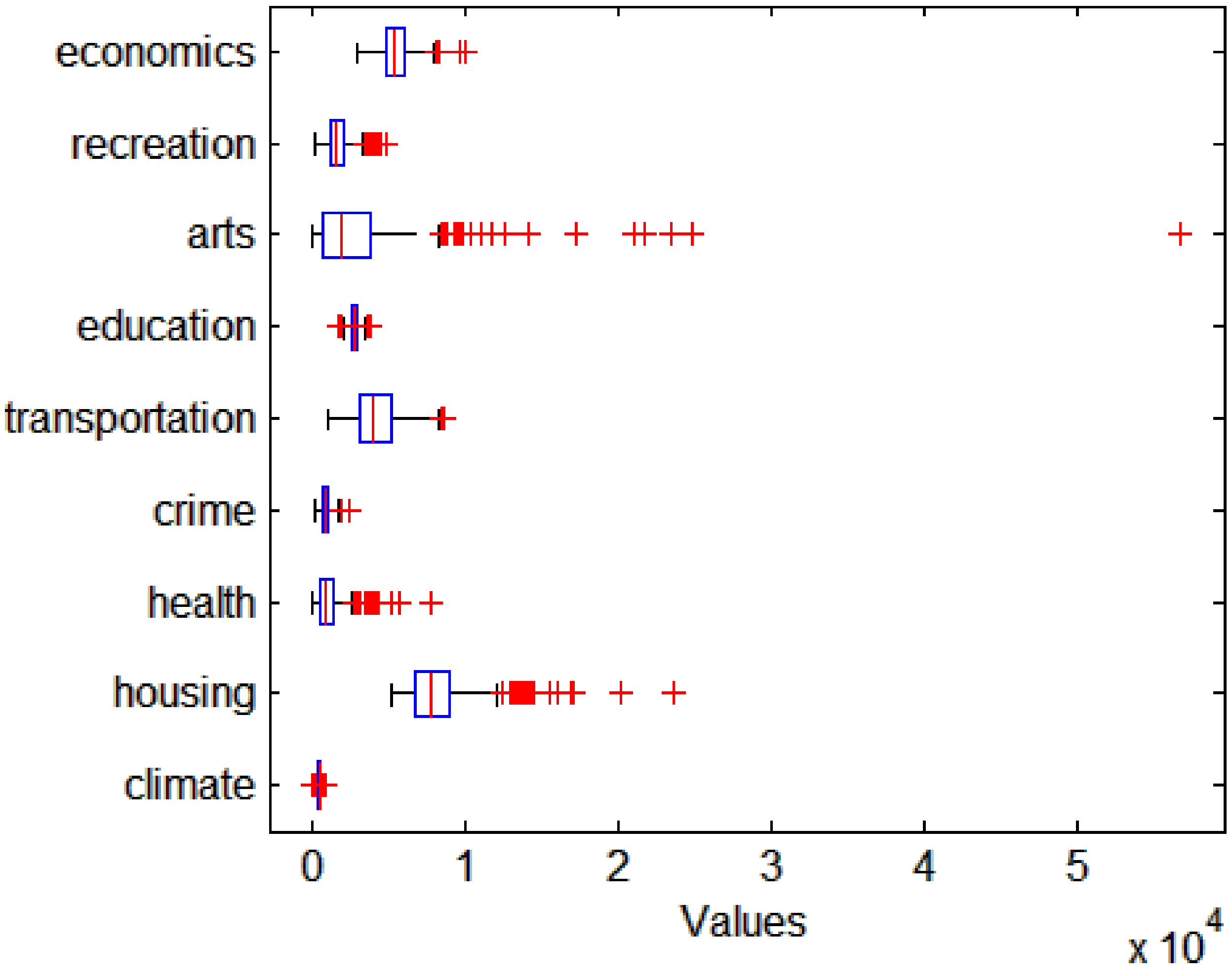
Sepal

Class labels
(targets)

Features
(attributes, measurements, dimensions)

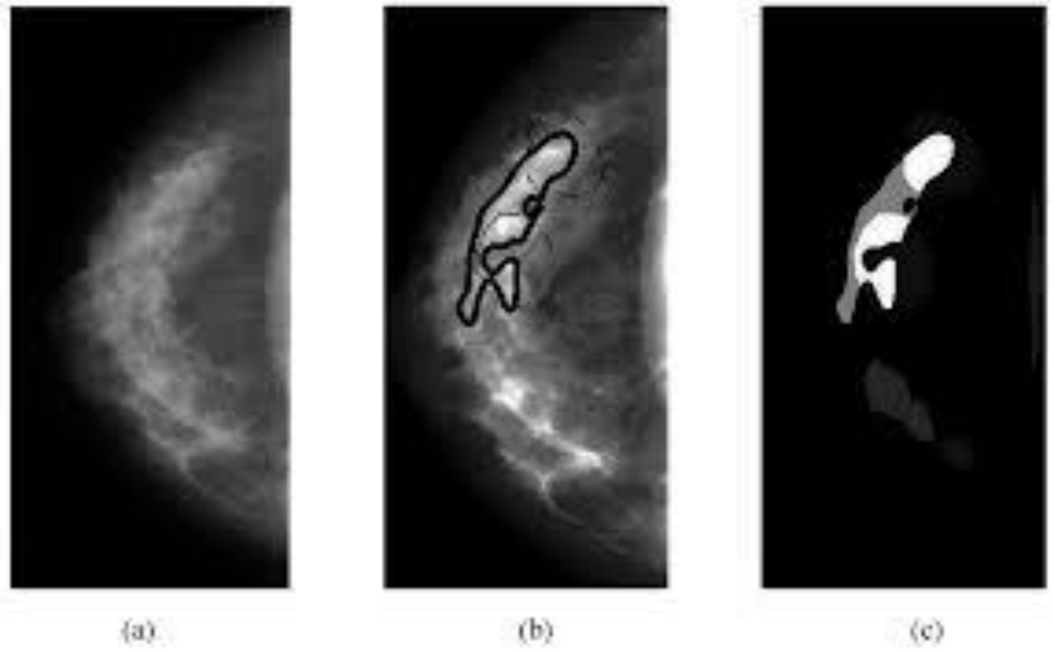


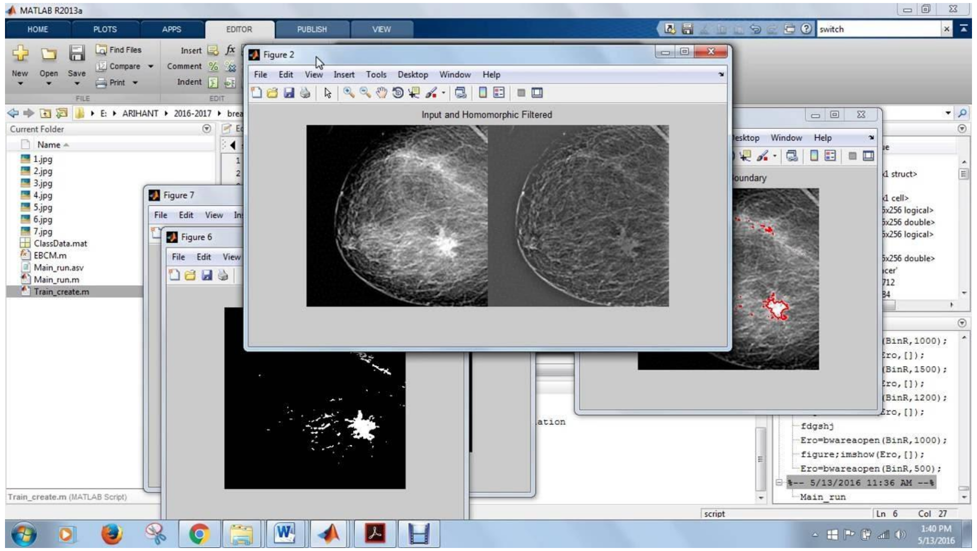
Estimation of life quality in US



Medicine: Brest cancer detection

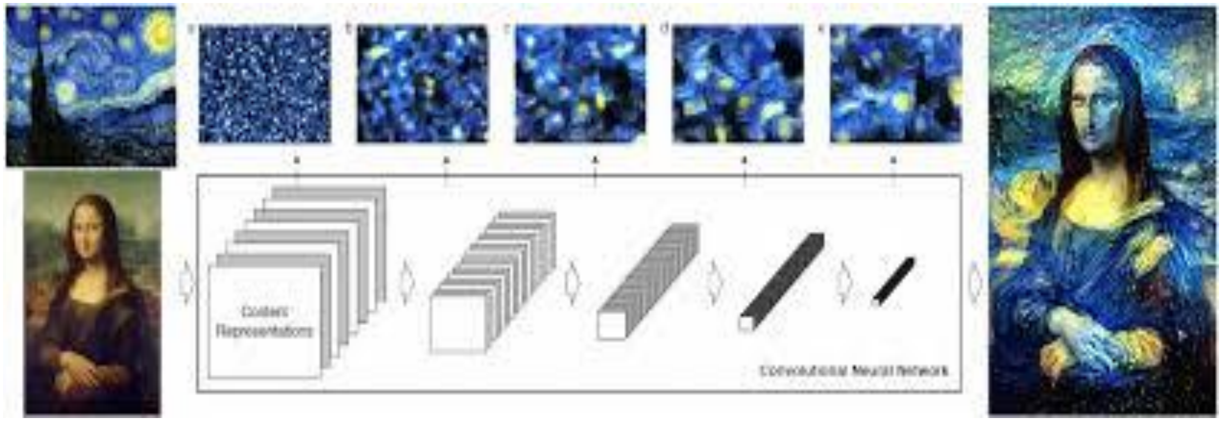
- KNN K-nearest neighbors


(a) (b) (c)



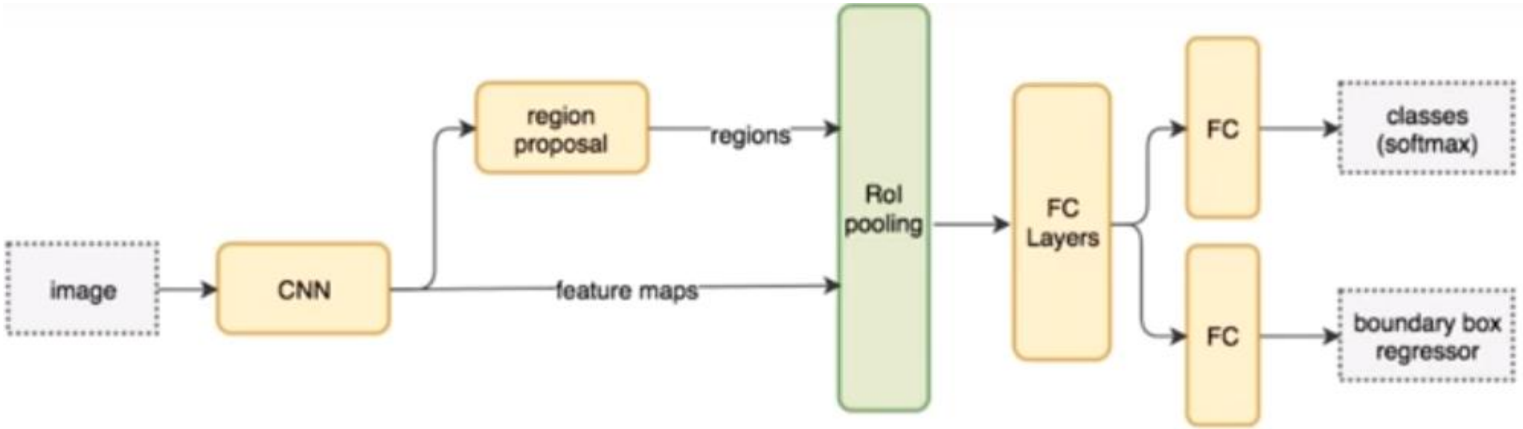
```
(BinR,1000);
Ero,[]);
(BinR,1500);
Ero,[]);
(BinR,1200);
Ero,[]);
fdgshj
Ero=bwareaopen(BinR,1000);
figure;imshow(Ero,[]);
Ero=bwareaopen(BinR,500);
%-- 5/13/2016 11:36 AM --%
Main_run
```

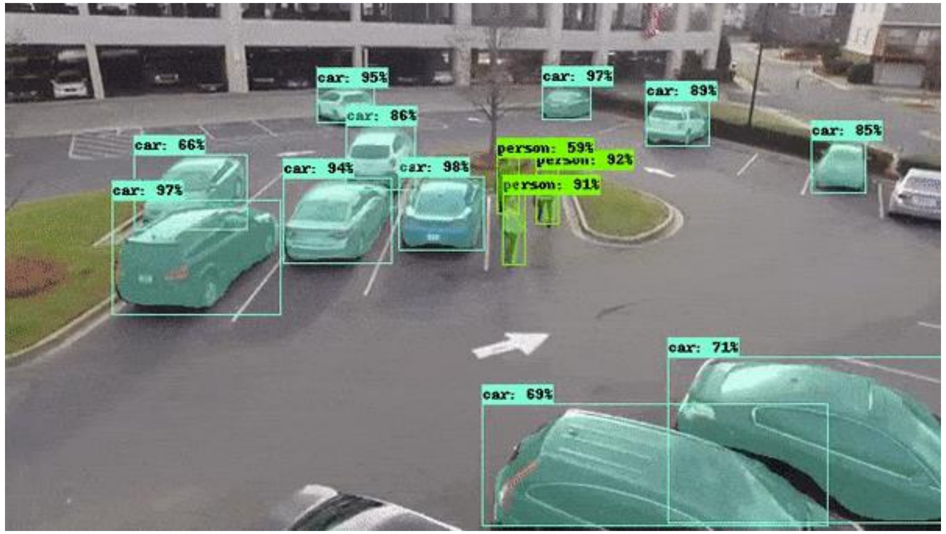
Artificial Neural networks: Machine painting



Machine Learning: Object recognition

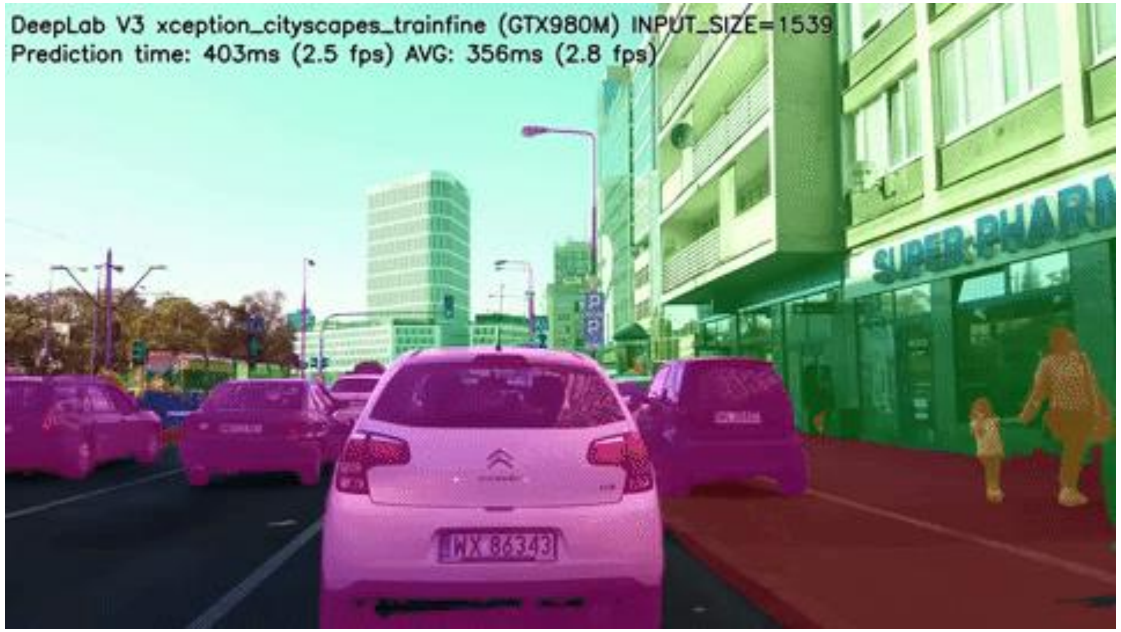
- **Object detection and localization**





Machine Learning: Semantic segmentation

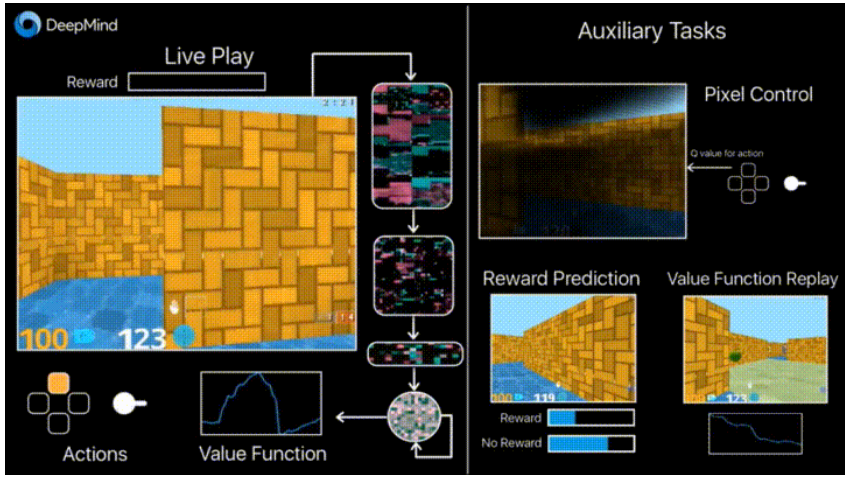
- **Fully convoluted Neural Networks**



```
DeepLab V3 xception_cityscapes_trainfine (GTX980M) INPUT_SIZE=1539
Prediction time: 403ms (2.5 fps) AVG: 356ms (2.8 fps)
```

Machine learning: *other cases*

- Deep reinforcement learning



THANK YOU