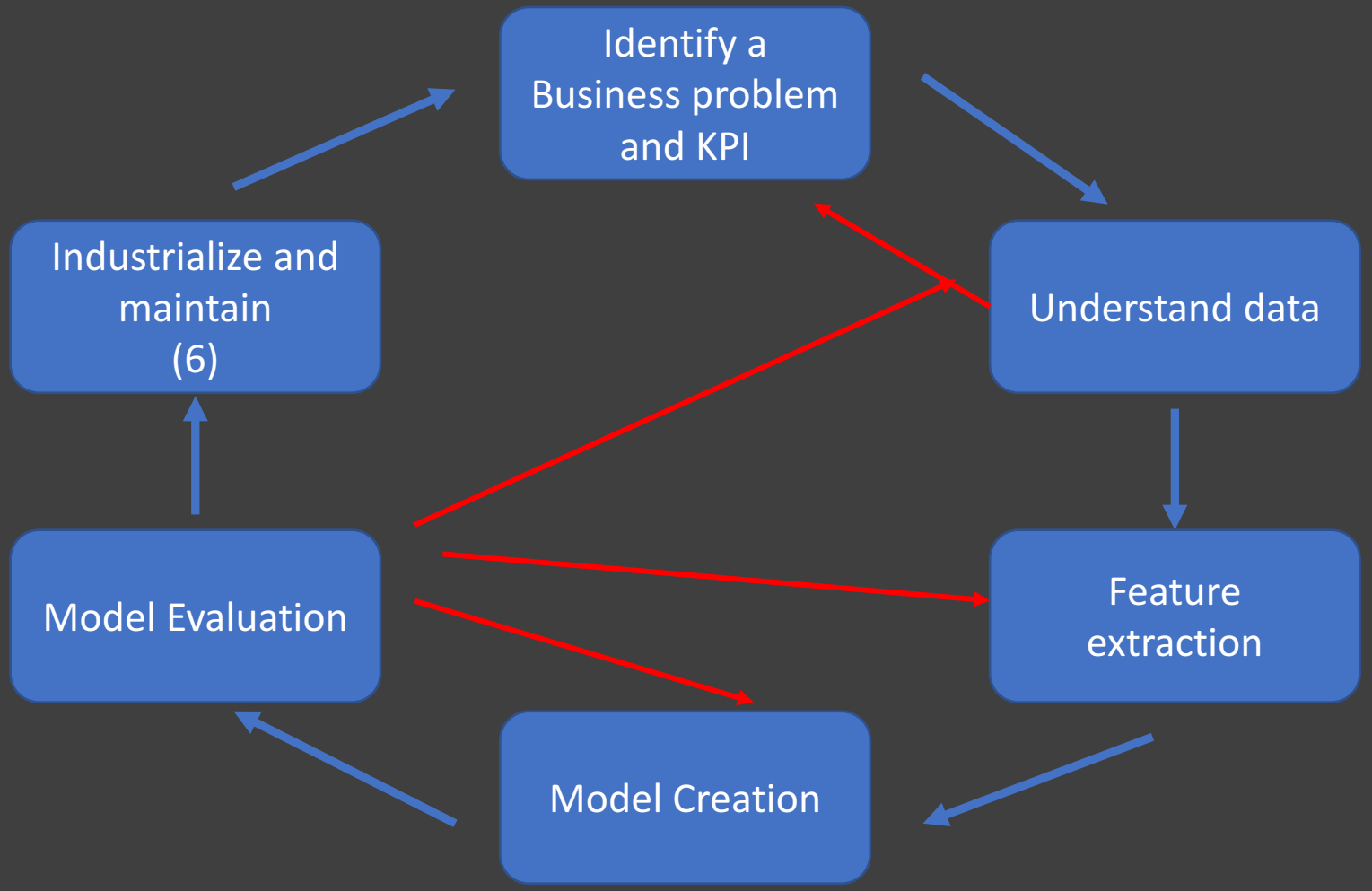
Modeling Methodologies

Sebastien Perez Vasseur
Igor Arambasic

Steps of Modeling

Translating a Business Problem to Machine Learning



What is Machine Learning ?

"Machine learning is a field of computer science that uses statistical techniques to give computer systems the ability to **"learn" with data**, without being explicitly programmed.”

“El aprendizaje automatizado o aprendizaje de máquinas es el subcampo de las ciencias de la computación y una rama de la inteligencia artificial, cuyo objetivo es desarrollar técnicas que permitan que las computadoras aprendan.”

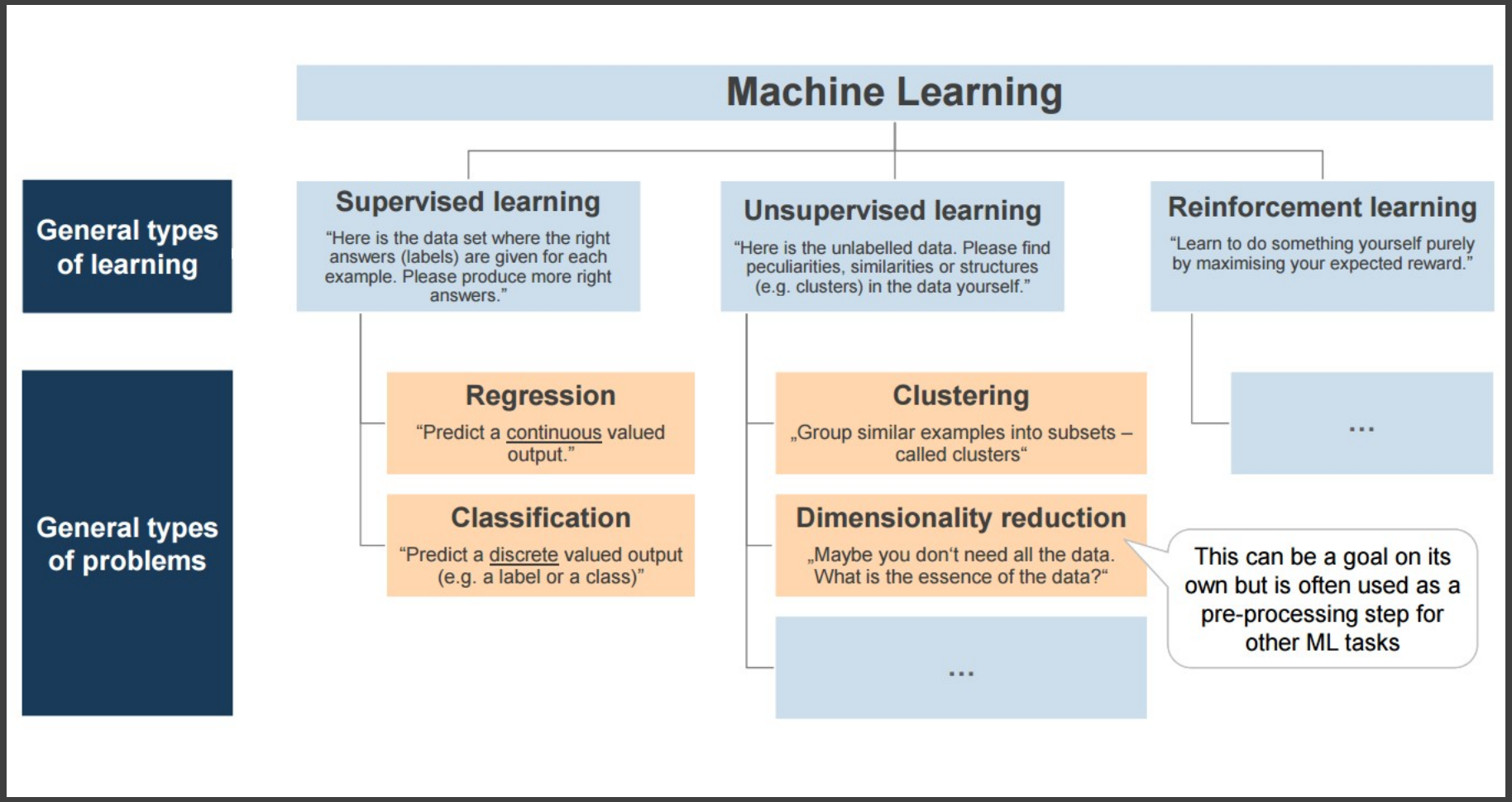
Wikipedia

Cognitive Tasks

In the same way, physical tasks were automated in the 70s, cognitive tasks are automated with Machine Learning.

The tasks are simple and need to be assembled in order to produce results.

Which types of cognitive tasks ?



Which types of cognitive tasks ?

- Supervised Learning (Learning from past elements):
  - Regression: Predicting a number
  - Classification: Predicting a label

- Unsupervised (Learning by comparison):
  - Clustering: Finding elements alike
  - Dimensionality Reduction: Explaining elements with less attributes

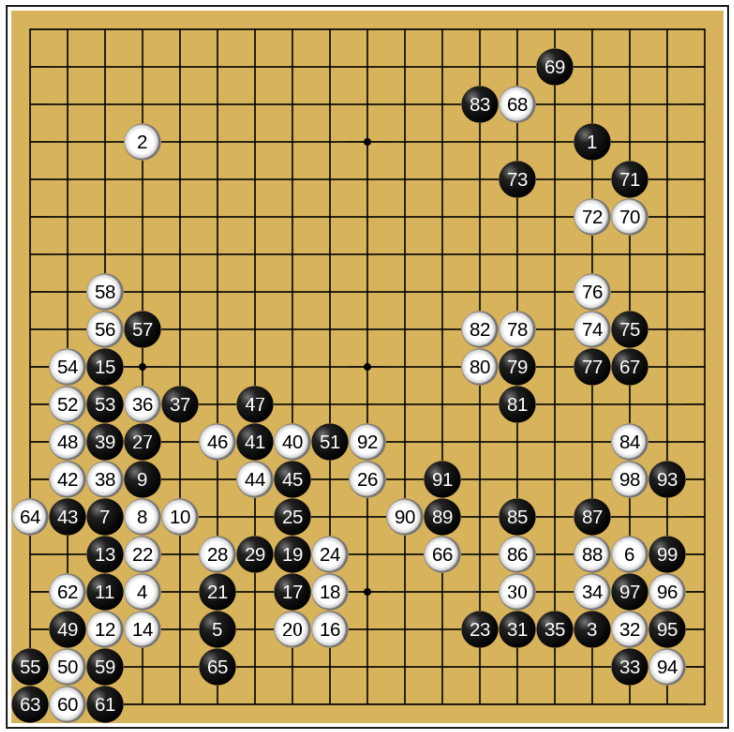
Machine Learning

- Vs Artificial Intelligence:
  - ML = “El aprendizaje automatizado o aprendizaje de máquinas es el subcampo de las ciencias de la computación y una rama de la inteligencia artificial …”
  - Strong AI or Broad AI is area of computer science that makes machines seem like they have human intelligence.
  - Weak AI or Narrow AI is every sort of machine intelligence that surrounds us today.
    - It operates within a pre-determined, pre-defined range.
    - **This is actually ML which appears to be much more sophisticated than it is.**
    - It lacks the self-awareness, consciousness, and genuine intelligence to match human intelligence.

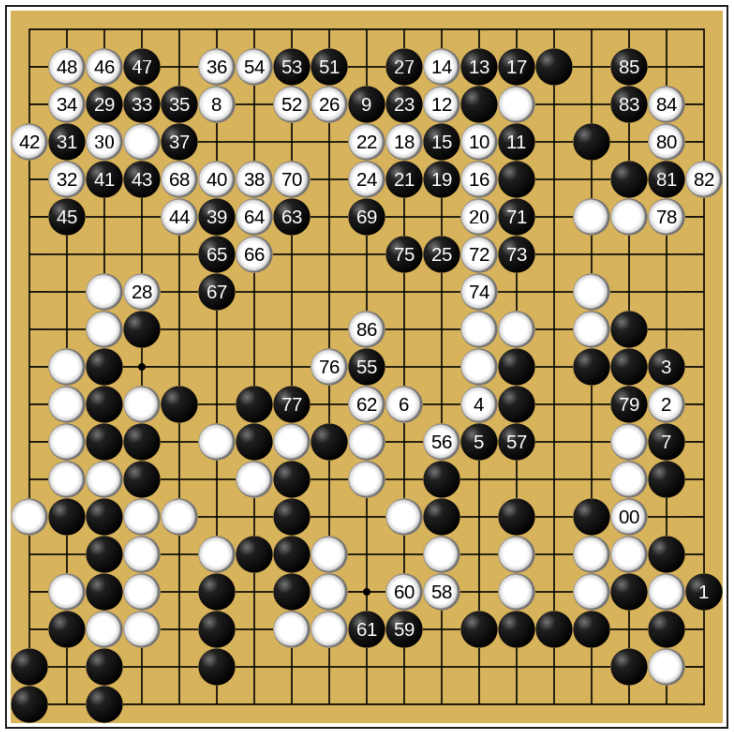
Alpha Go

Example game [edit]

AlphaGo Master (white) v. Tang Weixing (31 December 2016), AlphaGo won by resignation. White 36 was widely praised.



First 99 moves



Moves 100–186 (149 at 131, 150 at 130)

19x19 board

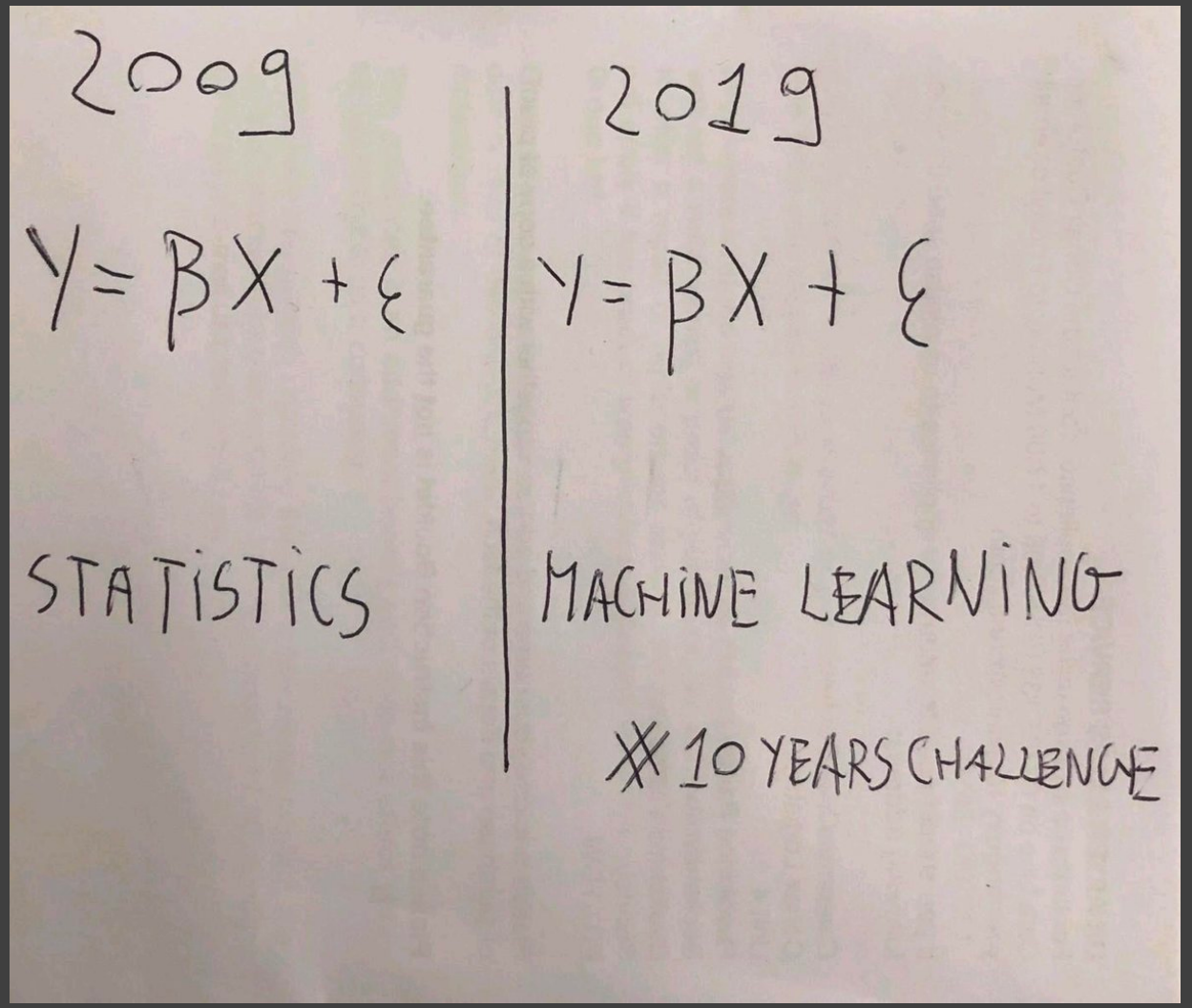
What will happen on 20x20 board?

Machine Learning

- Vs Deep Learning
  - Deep Learning is also known as **deep** neural **learning** or **deep neural network**
  - Deep Learning is a type of machine learning models

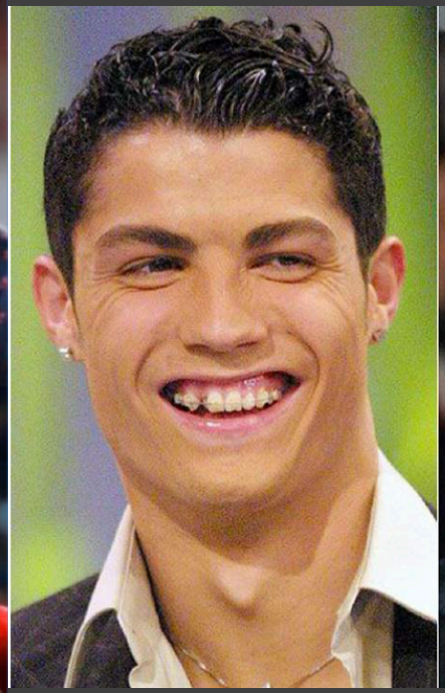
- Vs Statistics
  - Statistics are a field of Mathematics. Some aspects of Statistics have been renamed as Machine Learning.

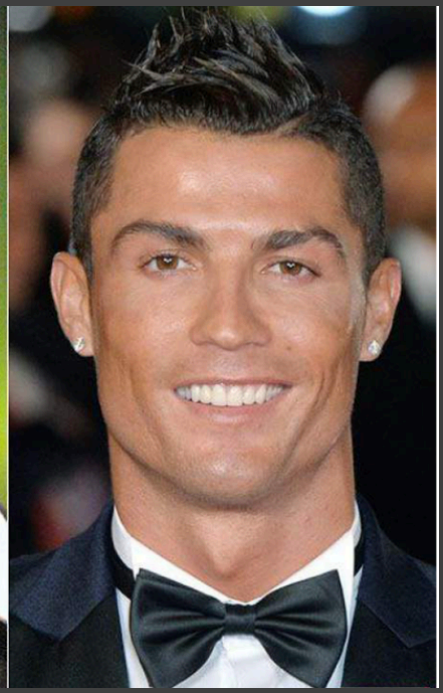
#10yearchallenge



#10yearchallenge

Who is building facial recognition data set?





What is ML about ?

ML Example: Predicting House Prices

Problem Statement:

We would like to predict the price of a house according to its characteristics.

Available Data:

- Catastro
- House Websites: Idealista, Fotocasa, …
- …

Machine Learning is about features

- Machine Learning requires explaining reality as a table of features

Id	LotArea	GrLivArea	GarageArea	PoolArea	Sale Price
0	8450	196.0	1710	548	208500
1	9600	0.0	1262	460	181500
2	11250	162.0	1786	608	223500
3	9550	0.0	1717	642	140000
4	14260	350.0	2198	836	250000

Features Description

Id	LotArea	GrLivArea	GarageArea	PoolArea	Sale Price
0	8450	196.0	1710	548	208500
1	9600	0.0	1262	460	181500
2	11250	162.0	1786	608	223500
3	9550	0.0	1717	642	140000
4	14260	350.0	2198	836	250000

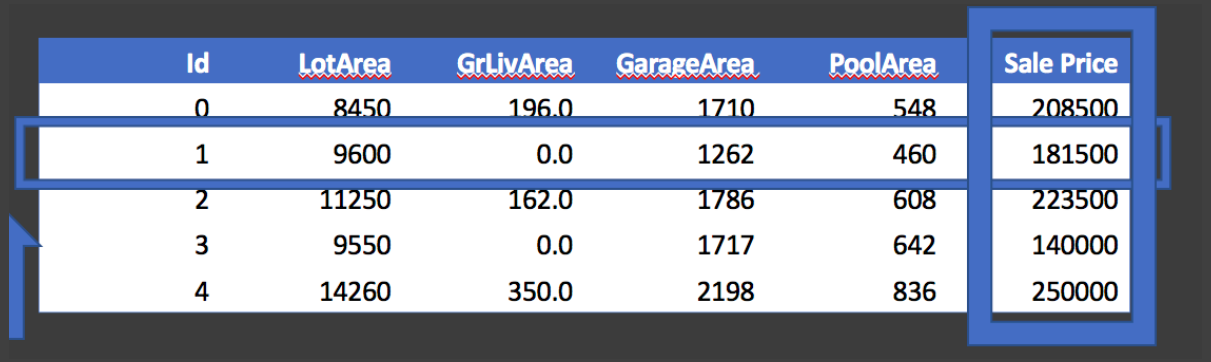
Each line represents a house

Target

Machine Learning is about models

- Machine Learning use models to explain relationships between features.

- In this case, we look for a function that explains the saleprice:



Id	LotArea	GrLivArea	GarageArea	PoolArea	Sale Price
0	8450	196.0	1710	548	208500
1	9600	0.0	1262	460	181500
2	11250	162.0	1786	608	223500
3	9550	0.0	1717	642	140000
4	14260	350.0	2198	836	250000

SalePrice = f(Features)

Example 1 of Model: Linear Regression

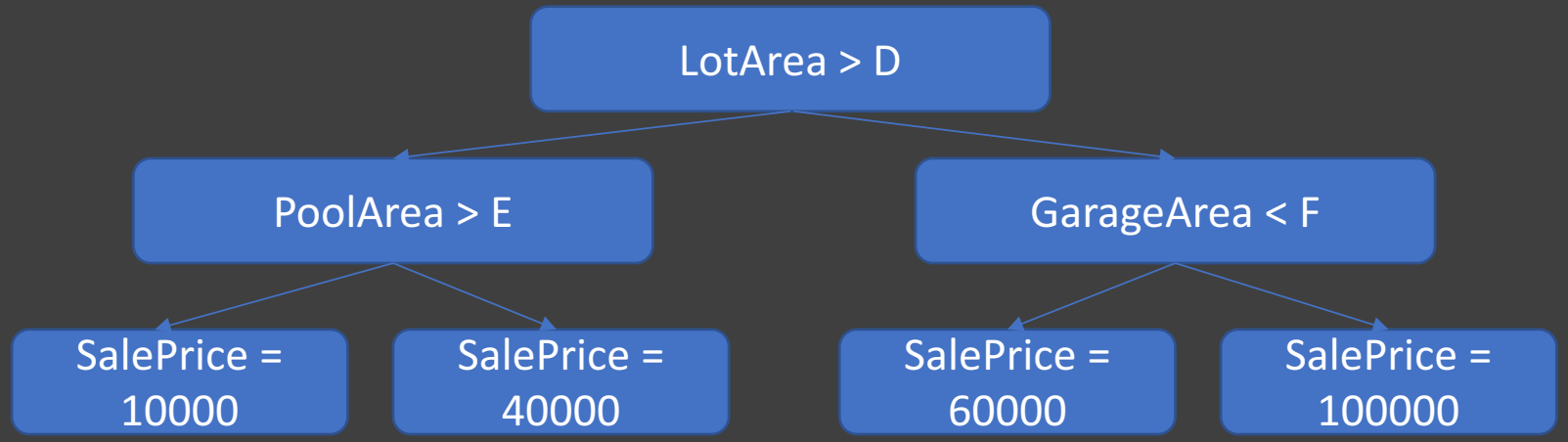
- A linear regression tries to find a linear function, in this case:

$$\text{SalePrice} = a \times \text{LotArea} + b \times \text{GrLivArea} + c \times \text{GarageArea} + d \times \text{PoolArea}$$

- We need to find the best a, b, c and d to have the best relationship.

Example 2 of Model: Decision Tree

- A decision tree tries to find a path to explain the target:



- We need to find the best splits for our predictions.

And there are many more …

- Every Data Scientist has different types of models for a given cognitive task:
  - Linear Regression
  - Decision Tree
  - Random Forest
  - K-neighbors
  - SVM
  - Neural Networks, Deep Learning, ...
  - And the list is growing thanks to the active research.

Parameters of the model

The Data Scientist works to find the best parameters for the model.

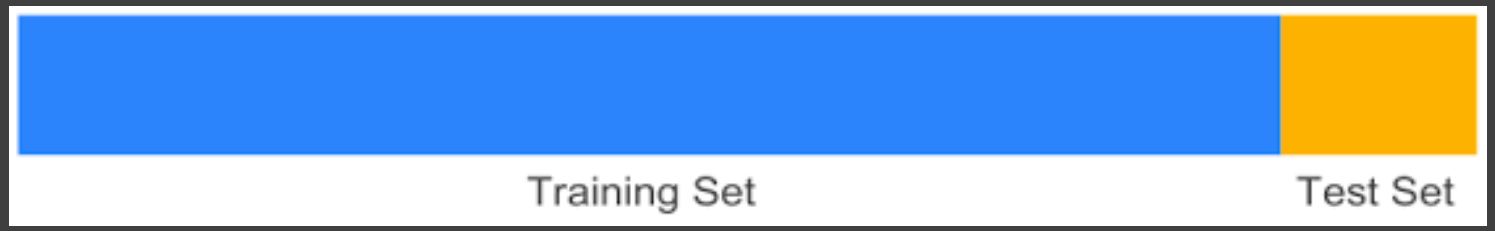
Linear Regression: a,b,c and d.

Decision Tree: Number of Splits, Max Depth, Min leaf samples…

How do you find the best parameters ?

Training

- Each model has a particular way to find the best parameters.
- We say the model is trained when it has found the best parameters with the available data
- We called this Data the Training Data

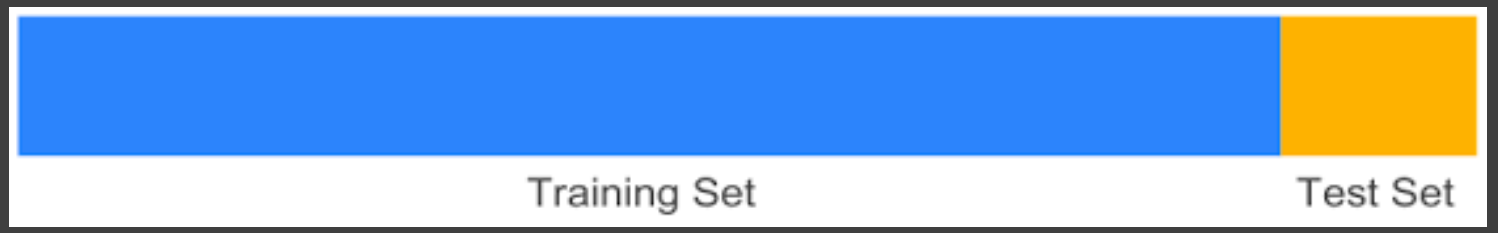


Evaluate Models: Training – Test Split

- We take the available features table and we split in 2 sets:

  1 set to calculate parameters : Training Set
  1 set to calculate the metrics : Test Set

  In the Test Set, we compare the real price of the house with the predicted price.



What are we missing?

- What do we have?
  - We have our problem defined
  - We have our features selected/created
  - We have our models selected
  - We have our training and evaluation set

- What are we missing?
  - A Metric
  - A way to select the best parameters of the model
  - A way to compare the results of different models

Machine Learning is about Metrics

There are several metrics that can be chosen for this task. For example:

- Bias: Average of the errors
- MAE: Average of the absolute value of errors
- RMSE: Square root of Average of the square of errors …

We chose the model parameters that provide the best METRIC. In this case, let's use MAE:

$$\frac{|\text{RealPrice1}-\text{PredictedPrice1}| + |\text{Real Price2}-\text{PredictedPrice2}| + \ldots}{\text{Number Of Houses in Test Set}}$$

Final Model

- Once we have the best parameters for a type of model, we can rank the models with the best metric they had:

- Linear Regression - MAE: 15
- Decision Tree – MAE: 10

- And we ultimately chose the model with the best metric.

Model Usage

- We can now use this model to “predict” the SalePrice of a new house from its characteristics:

$$\text{SalePrice} = F(\text{House features})$$

Summary: Machine Learning is about features, models and metrics

- Machine Learning requires explaining reality as a table of features
- We then use models to explain relationships between features for the given ML task
- Each model is evaluated according to a metric

- The correct metric depends on the use case!!!

ML Example: Detecting diabetes in patients

Classification

- Problem: Detecting diabetes in patients
- Features:

Each line represents a patient

Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	Diabetes Pedigree Function	Age	Outcome
6	148	72	35	0	33.6	0.627	50	1
1	85	66	29	0	26.6	0.351	31	0
8	183	64	0	0	23.3	0.672	32	1
1	89	66	23	94	28.1	0.167	21	0
0	137	40	35	168	43.1	2.288	33	1
5	116	74	0	0	25.6	0.201	30	0
3	78	50	32	88	31	0.248	26	1
10	115	0	0	0	35.3	0.134	29	0

Target

- Metric: Accuracy (Percentage of correctly guessed patient's disease)

ML Example: Detecting groups of similar customers

Segmentation

- Problem: Detecting groups of similar customers
- Features:

Each line represents a customer

CustomerID	Departure Date	Age	Distance	Country of POS
1	22/12/18	20	1000	Spain
2	21/04/18	25	100	France
3	24/05/18	30	5000	Germany
4	30/03/19	55	500	Spain
5	14/03/20	32	200	Germany
6	22/12/18	45	400	Bulgaria
7	20/01/18	43	10000	France
8	12/06/19	12	5000	United Kingdom
9	25/03/18	39	1300	Spain
10	31/07/18	67	700	France

- Method: We find groups of similar customers based on their proximity
- Metric: Silhouette Index (how far are customers from other clusters than theirs)

ML Canvas

Translating a Business Problem to Machine Learning

Any Business Problem can be analyzed through the following axes:

- Business Value and Measure of Success (1)
- Data Sources (2)
- Machine Learning task + Metric (3)
- Features Extraction + Model Creation (4)
- Industrialization and Maintenance (5)

Opportunity:

Estimated Value:

Estimated Cost:

Value Proposition

Business Description
Resulting Action
Measure of Success (KPI)
ML Initiative Cost

Machine Learning

ML task
Methods of evaluation

Data

Sources
Frequency
Legal
History

Acceptable Quality
Main Features

Ranking of Models

List of the models and their metric

Industrialization

Description
Implementation Cost
Maintenance Cost
Model Specific Cost

By:

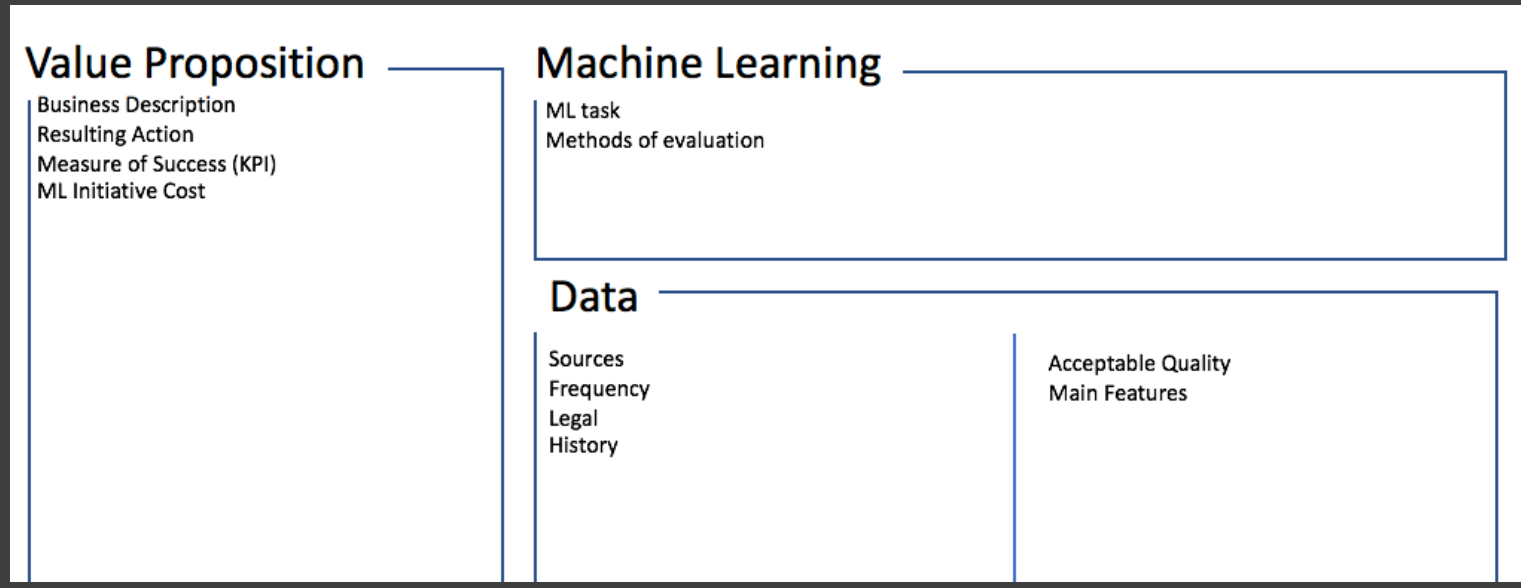
Iteration:

Date:

Collaboration with DS on creating iteration 1

Interaction with DS involves discussing about data:

- Data Sources
- History
- Features
- Frequency
- Quality
- Volume



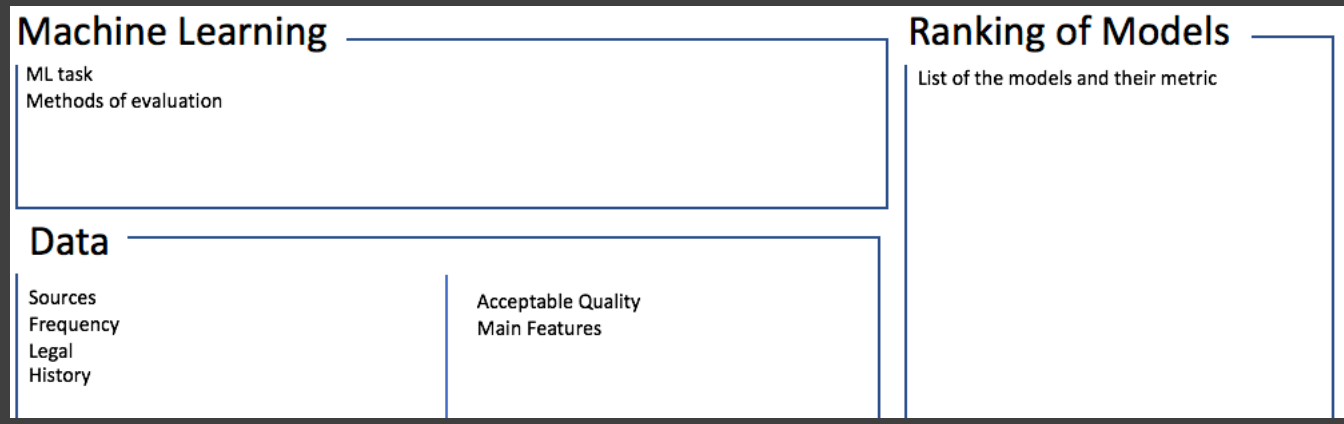
and discussing bout measuring success:

- Evaluation of success
- How does it translate to a metric

Data Science Process

Once the first iteration of the canvas is created, DS perform the following actions:

- Extraction of features (2)
- Selection of ML task (3)
- Selection of evaluation methods(3)
- Model Creation (4)
- Quantity of Data > Parameters
- Loop until satisfied



The whole process is iterative

- Product Managers and Data Scientists can revise their initial plans based on discoveries made during the different steps:
  - Insufficient Data
  - Metric not good enough for the task
  - …

Opportunity:

Estimated Value:

Estimated Cost:

Value Proposition

Business Description
Resulting Action
Measure of Success (KPI)
ML Initiative Cost

Machine Learning

ML task
Methods of evaluation

Data

Sources
Frequency
Legal
History

Acceptable Qu
Main Features

DS + PM

Ranking of Models

List of the models and their metric

Industrialization

Description
Implementation Cost
Maintenance Cost
Model Specific Cost

By:

Iteration:

Date:

Opportunity:

Estimated Value:

Estimated Cost:

Value Proposition

Business Description
Resulting Action
Measure of Success (KPI)
ML Initiative Cost

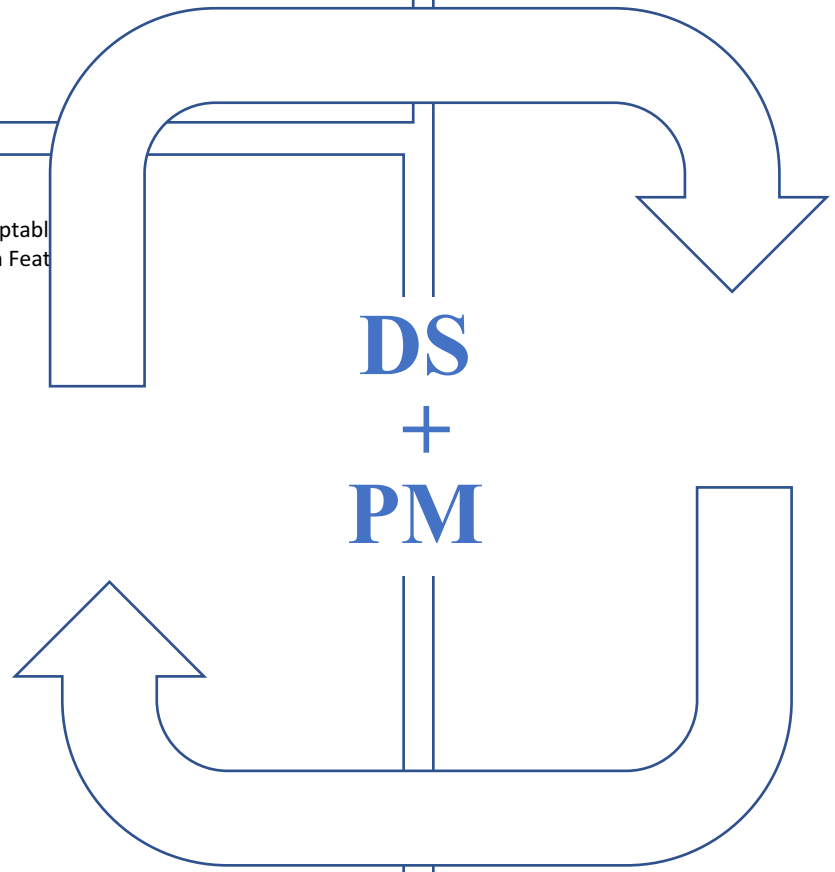
Machine Learning

ML task
Methods of evaluation

Data

Sources
Frequency
Legal
History

Acceptabl
Main Feat



Ranking of Models

List of the models and their metric

Industrialization

Description
Implementation Cost
Maintenance Cost
Model Specific Cost

By:

Iteration:

Date:

Deliverables

- Model(s) with satisfactory metric:
  - A suitable model (or type of model) is chosen for the task at hand.
  - This model can now do the necessary predictions, but needs to be assembled …
- Industrial Implementation Discussion:
  - Product Managers, Data Scientists and Development can start discussing how the model will be used, trained, …

Opportunity:

Estimated Value:

Estimated Cost:

Value Proposition

Business Description
Resulting Action
Measure of Success (KPI)
ML Initiative Cost

Machine Learning

ML task
Methods of evaluation

Data

Sources
Frequency
Legal
History

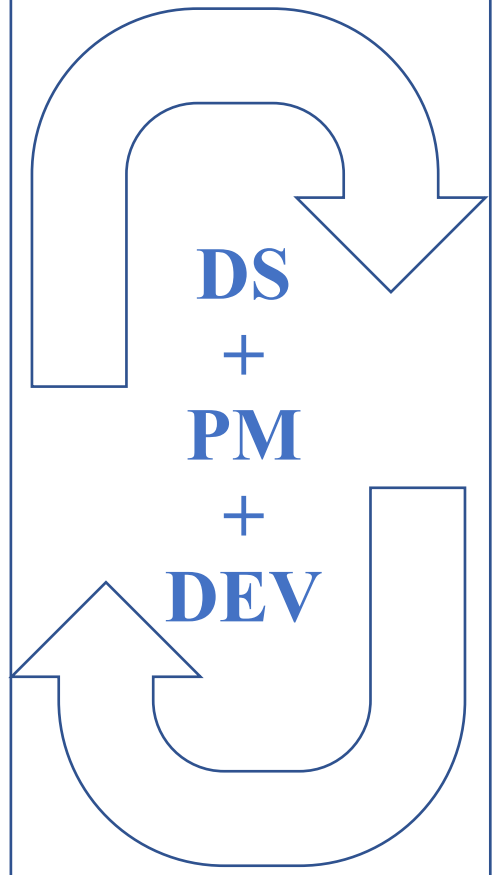
Acceptable Quality
Main Features

Ranking of Models

List of the models and their metric

Industrialization

Description
Implementation Cost
Maintenance Cost
Model Specific Cost



By:

Iteration:

Date:

Opportunity: House Price Estimation

Estimated Value:

Estimated Cost:

Value Proposition

Business Description
Resulting Action
Measure of Success (KPI)
ML Initiative Cost

We are Real estate brokerage agency like Tecnocasa or Redpiso.

Business Description:
The goal is to offer our services only to clients with houses that can be sold easily with high margin

Resulting Action:
Daily report of the new houses with offered price and estimated price.

Measure of Success (KPI):

- Increased: Number of houses sold with high margin
- Decreased: time needed to sell the house

ML Initiative Cost:
1 month study

Machine Learning

ML task
Methods of evaluation

Regression
Method of evaluation: MAE

Data

Sources
Frequency
Legal
History

Data Sources:
Idealista, Kaggle Data, …

Frequency:
Weekly, some daily

Legal:
Public Data

History:
1 year

Acceptable Quality
Main Features

Acceptable Quality:
Yes

Main Features:
Area of house,
Bathrooms, …

Ranking of Models

List of the models and their metric

Decision Tree
MAE:

Linear Regression:
MAE:

Industrialization

Description
Implementation Cost
Maintenance Cost
Model Specific Cost

Weekly job that retrains and produces new model.

Every morning, we run the predictions against a set of newly available houses.

Yearly evaluation to check changes

By:

Iteration:

Date:

Hands on Exercises

Hands on Exercises

Let's try to do that exercise for those 5 use cases:

- Weather Forecast
- Movie Recommendation
- Email Spam Filters
- Pollution Prediction
- Customer Segmentation

Hands on Exercises

- 40’ Business:
  - 10’: Think about business/company that is related to the use case. This would be your audience at presentation
  - 10’: Define Business Problem
  - 10’: What would be the action with the results
  - 10’: KPI or Objective reached (how to measure the success of ML approach)
- 5’ Machine Learning:
  - Define ML Problem
  - Methods of evaluation

- 45’ Data:
  - 15’: Find data sources (Example: https://toolbox.google.com/datasetsearch)
  - 15’: Analyze the data --> Extract the column/features
  - 15’: Propose features which will be used for modeling
- 30’: Industrialization:
  - 15’: Practical Use Case
  - 15’: Draw the Data Flow
- 60’: Presentation:
  - 20’: Create Presentation of your DS project proposal
  - 40’: Present (5min per group + feedback)