Data Science Fundamentals

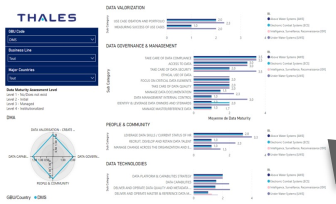
Yann Gouedo
Data Scientist Leader – Machine Learning / Artificial Intelligence
Marketing / Risk / Fraud / Maintenance / Pricing
Distinguished Data Scientist, Open Group Certification

ENGAGEMENT APPROACH

How to engage a data driven digital transformation?

To answer to business expectations, the methodology is based on a co-working approach, named Data Thinking , focused to address specific clients needs and context

***# 1 Workshop**: Evaluate **Data maturity** within the organization*



#2 Workshop**: Define the **ambitions** related to the **data strategy

***# 3 Workshop**: Identify **use cases** answering to operational needs with "**data thinking**"*

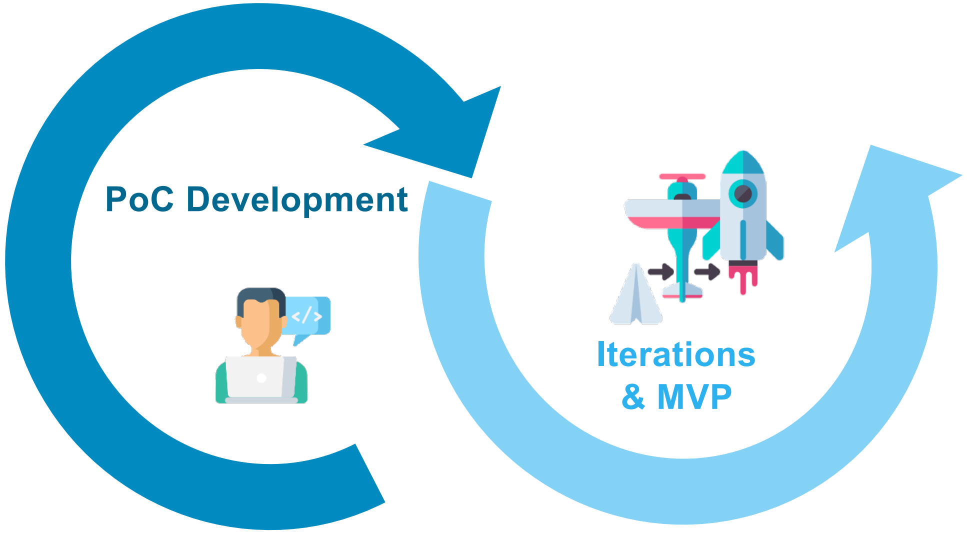
***# 4 Workshop** : **Prioritize** the use cases*

***# 5 Workshop** : Define functional and technical needs with stakeholders*

Once the use cases have been determined, the creation of a first prototype allowing you to carry out tests before industrialization

DATA EXPLORATION

IMPLEMENTATION



THE DEVELOPMENT OF AI MODELS presupposes the establishment of an **EFFICIENT** and **ROBUST ENVIRONMENT**

Some key steps before your **MVP goes to PRODUCTION** !

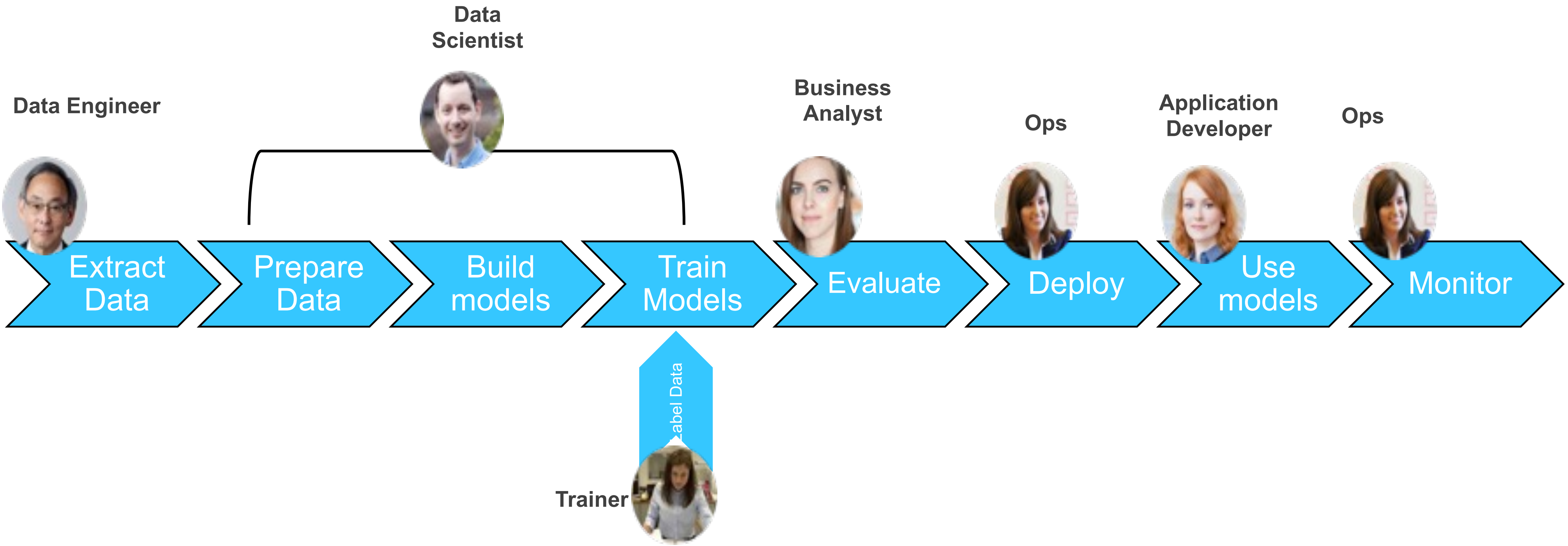
- « **Design to operate** »: developments compatible with the production environment
- **Execution of a PoC** that anticipates the visualization of data (data Analytics / AI)
- **Evolution of the PoC** (Proof of concept / "fictional" test) **into an MVP** (Minimum Viable Product / real case on a small perimeter)
- **Go / No Go** on the interest of exploiting data and going into industrialization

A use case prototyping is key to find and implement new sources of competitive advantage

Identify a Quick Hit opportunity	To identify a quick-hit analytics opportunity using our specialized selection methodology
Real data	To port a real, actionable data set – even messy data - into a unique toolset and platform enabled with a cloud or on premise platform
Data Scientists	Data scientists use special techniques to analyze the data that doesn't require traditional data models or schema
Fast turnaround	The prototype is finished in a matter of weeks (not months or years)
Actionable insights	Actionable findings and outcomes are ready for business consumption
ROI	Business and economic value are realized as the first real bite of analytics outcomes are pursued and won

COMPETENCIES

Data Science is a Team Sport



- Building Machine Learning Models infused apps requires multiple skillsets:

  - Define an ML model
  - Store, manage, update training data
  - Manage lifecycle of the trained model
  - Ability to do inferencing on the trained model(s)

A Data Science team requires a large variety of competencies

Behavior

- Ability to synthetize / Simplify
- Communication skills (oral, written)
- Client focused
- Ability to share / pass on knowledge
- Creativity et innovation/ Problem solving
- Ability to negotiate
- Decision making

Transversal

- Analytical ability
- Capacity to manage a project
- Capacity to develop & improve skills
- Capacity to anticipate business / strategic evolution
- Capacity to develop & leverage networks

Functional and Technical Competencies

- Business / IT Relationship
- Descriptive Analytics
- Model exploration
- Applied Mathematics and Algorithms
- Domains of competencies (risk and fraud management, predictive maintenance, digital marketing, supply chain,…)
- Data knowledge

- Statistics
- Predictive Analytics
- Simulation
- Optimization – Prescriptive Analytics
- Cognitive computing
- NLP – Text mining
- Robotics

Methodologies/ Tools

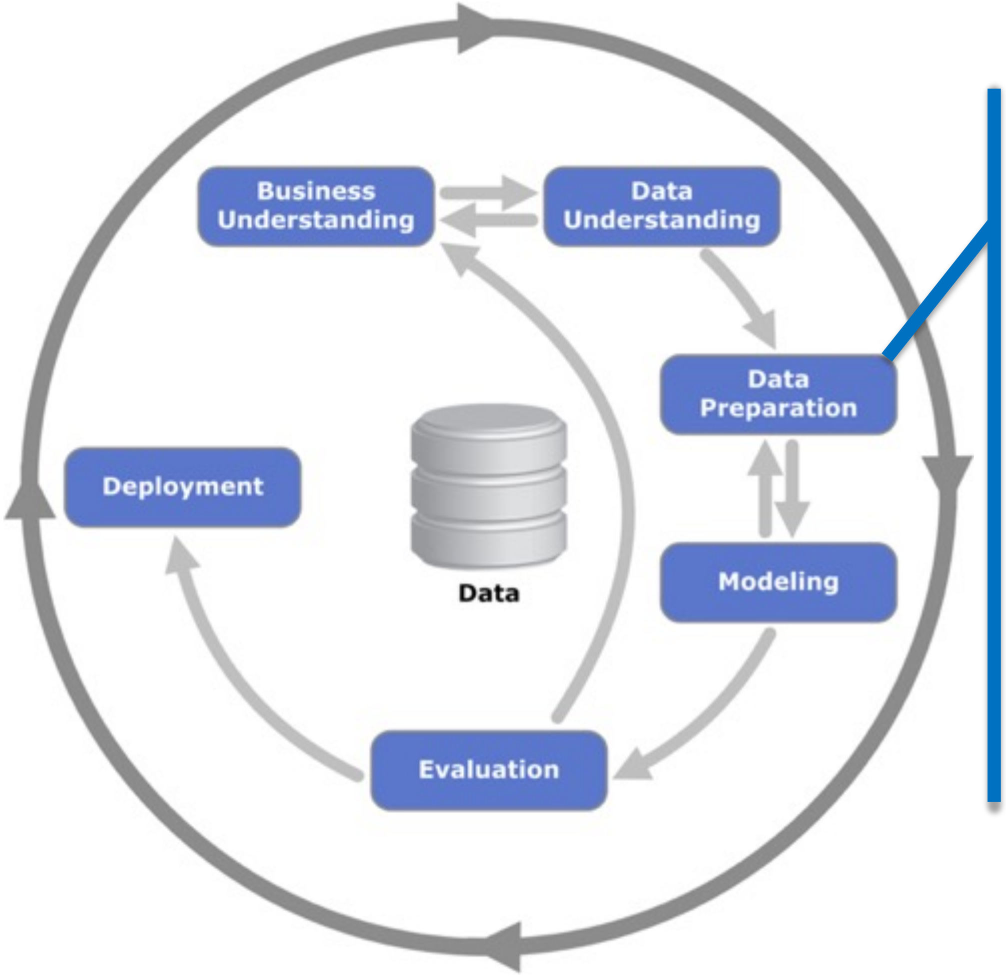
- Languages IT : SQL/R/PYTHON/SPARK/JS/ SCALA
- CRISP - DM
- Data Science Platforms
- Data exploration
- API interfacing

METHODOLOGY

The Data Science methodology consisted to cover the following steps

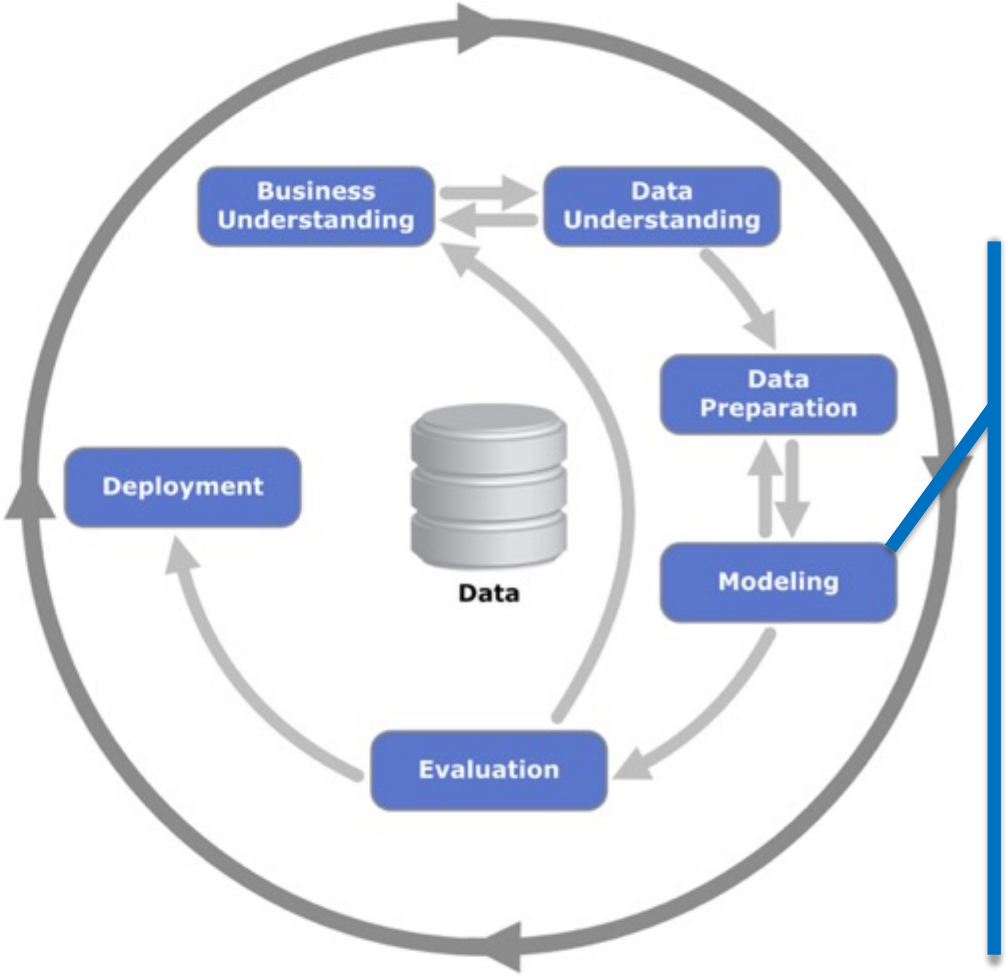
- A known Data Science methodology (Cross Industry Standard Process for Data Mining)
- Criteria of success (Business, Technical, Change Management, …)
- A collaborative work between the CLIENT and the Data Science Expert Team, which secures the success
- A Data Science platform

The Data Preparation step



- Technical Data Cleaning (Types of variables, …)
- Statistical Data Cleaning
- Creation of Variables/Indicators (features engineering)
- Selection of the dataset
- Building of the Information Foundation

The Modelling step: a robust approach that helped to ensure to work in a virtuous cycle



- The measurement (continuous or categorical target)
- A cross validation technique in order to validate the best algorithm and to minimize over-fitting
- An auto-modelling: that helped to know the best algorithm to use

MEASUREMENTS

Measurement (numerous target) – RMSLE

- **RSS, for Residual Sum of Squares**. Deviations predicted from actual empirical values of data.

$$RSS = \sum_{i=1}^{n} (f(x_i) - y_i)^2$$

- **MSE, for Mean Squared Error**. The RSS is generally normalized (by the number of observations) to avoid to have a huge number (in case there is a big number of observations.

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (f(x_i) - y_i)^2$$

- **RMSE, for Root Mean Squared Error**. Squared error to have the same unit than y=f(xi).

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (f(x_i) - y_i)^2}$$

- **RMSLE for Root Mean Squared Log Error**. In case the empirical values are on a large scale, it is necessary to do a logarithmic transformation. (a error of 10 units on the value of 4 is not the same than on a value of 100!)

$$RMSLE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (\log(f(x_i) + 1) - \log(y_i + 1))^2}$$

Measurement (numerous target) - MAE

- **Prediction Error = Actual Value - Predicted Value**. subtraction of Predicted value from Actual Value

- **Absolute Error → |Prediction Error|**

- **MAE, for Mean Absolute Error.** Mean for all recorded absolute errors (Average sum of all absolute errors). Refers to the measurement of the difference between two continuous variables.

$$mae = \frac{\sum_{i=1}^{n} abs\left(y_i - \lambda(x_i)\right)}{n}$$

- **MAE with the logarithm**

Prediction Error = log (1+Actual Value) – log(1+Predicted Value)

Measurement (categorical target) – Confusion Matrix

- In the field of machine learning and specifically the problem of statistical classification , A confusion matrix, also known as an error matrix, is a specific table layout that allows visualization of the performance of an algorithm. It is a table with two rows and two columns that reports the number of *false positives*, *false negatives*, *true positives*, and *true negatives*.

- The pattern built from all machines data provided the following Confusion matrix:

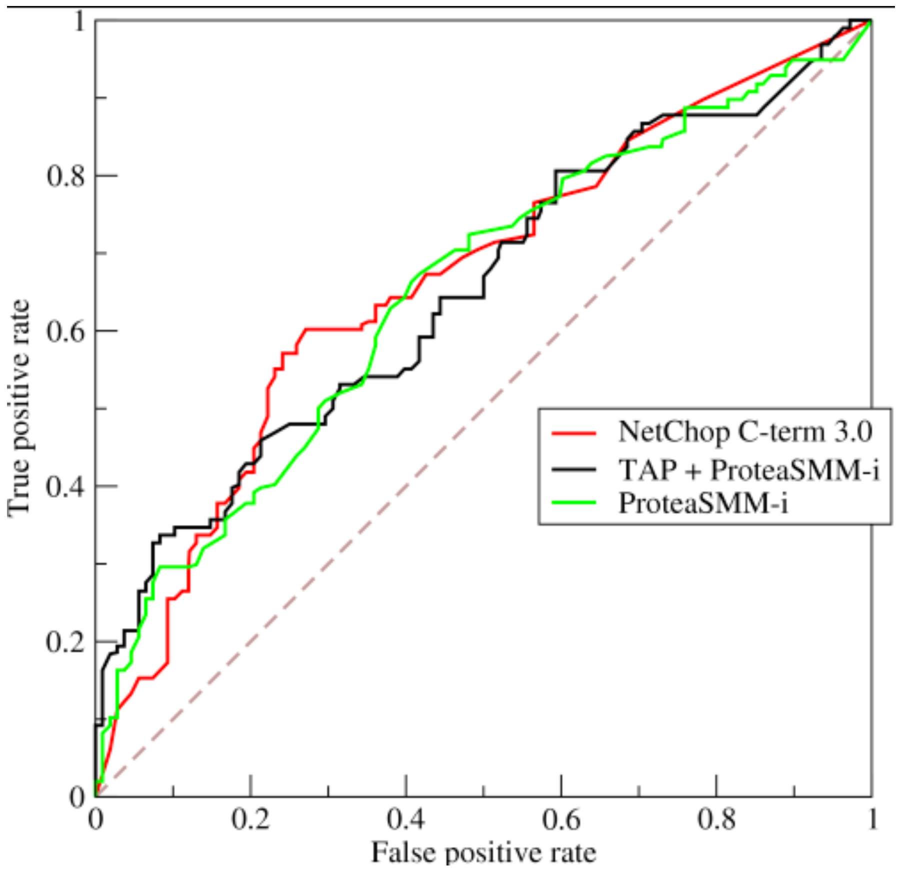
  - **True positives (TP):** These are cases in which we predicted Yes (1) (the failure occurs), and the failure occured.
  - **True negatives (TN):** We predicted No (0), and the failure did not occur.
  - **False positives (FP):** We predicted Yes (1), but the failure did not occur
  - **False negatives (FN):** We predicted No (0), but the failure occured

Measurement (categorical target) – Extended Confusion Matrix

		True condition				
	Total population	Condition positive	Condition negative	Prevalence $= \frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$	Accuracy (ACC) $= \frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$	
Predicted condition	Predicted condition positive	**True positive**, Power	**False positive**, Type I error	Positive predictive value (PPV), Precision $= \frac{\Sigma \text{ True positive}}{\Sigma \text{ Predicted condition positive}}$	False discovery rate (FDR) $= \frac{\Sigma \text{ False positive}}{\Sigma \text{ Predicted condition positive}}$	
	Predicted condition negative	**False negative**, Type II error	**True negative**	False omission rate (FOR) $= \frac{\Sigma \text{ False negative}}{\Sigma \text{ Predicted condition negative}}$	Negative predictive value (NPV) $= \frac{\Sigma \text{ True negative}}{\Sigma \text{ Predicted condition negative}}$	
		True positive rate (TPR), Recall, Sensitivity, probability of detection $= \frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$	False positive rate (FPR), Fall-out, probability of false alarm $= \frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$	Positive likelihood ratio (LR+) $= \frac{TPR}{FPR}$	Diagnostic odds ratio (DOR) $= \frac{LR+}{LR-}$	F_1 score $= \frac{2}{\frac{1}{\text{Recall}} + \frac{1}{\text{Precision}}}$
		False negative rate (FNR), Miss rate $= \frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$	Specificity (SPC), Selectivity, True negative rate (TNR) $= \frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$	Negative likelihood ratio (LR−) $= \frac{FNR}{TNR}$		

Confusion Matrix

ROC curve (receiver operating characteristic) and Sensitivity



sensitivity, **recall**, **hit rate**, or **true positive rate** (TPR)

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{P}} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} = 1 - \mathrm{FNR}$$

fall-out or **false positive rate** (FPR)

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{N}} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}} = 1 - \mathrm{TNR}$$

MODEL PERFORMANCE (SUPERVISED TECHNIQUES)

Overfitting and Underfitting with Machine Learning Algorithms

- **Supervised machine learning** is best understood as approximating a target function (f) that maps input variables (X) to an output variable (Y) (**Y = f(X))**

- An important consideration in learning the target function from the training data is **how well the model generalizes to new data**. Generalization is important because the data we collect is only a sample, it is incomplete and noisy. Generalization refers to how well the concepts learned by a machine learning model apply to specific examples not seen by the model when it was learning.

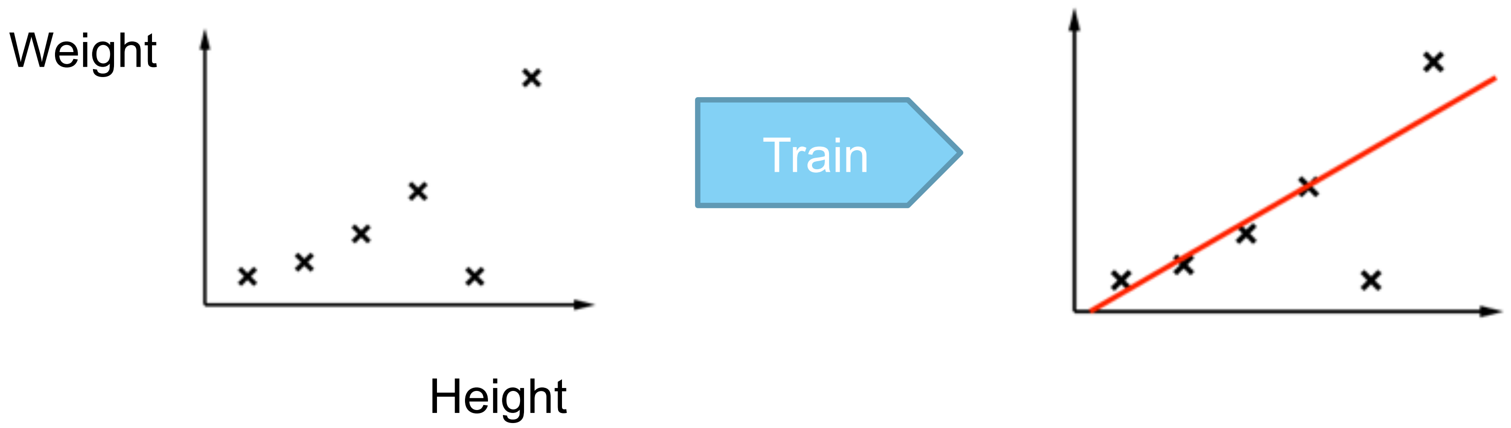
- The **goal of a good machine learning model is to generalize well** from the training data to any data from the problem domain. This allows us to make predictions in the future on data the model has never seen.

- The cause of **poor performance** in machine learning is either **overfitting** or **underfitting** the data.
  - **Overfitting** refers to a model that models the training data too well (nonparametric and nonlinear models)
  - **Underfitting** refers to a model that can neither model the training data nor generalize to new data (obvious to detect – no discussion)

Example 1: height and weight of people

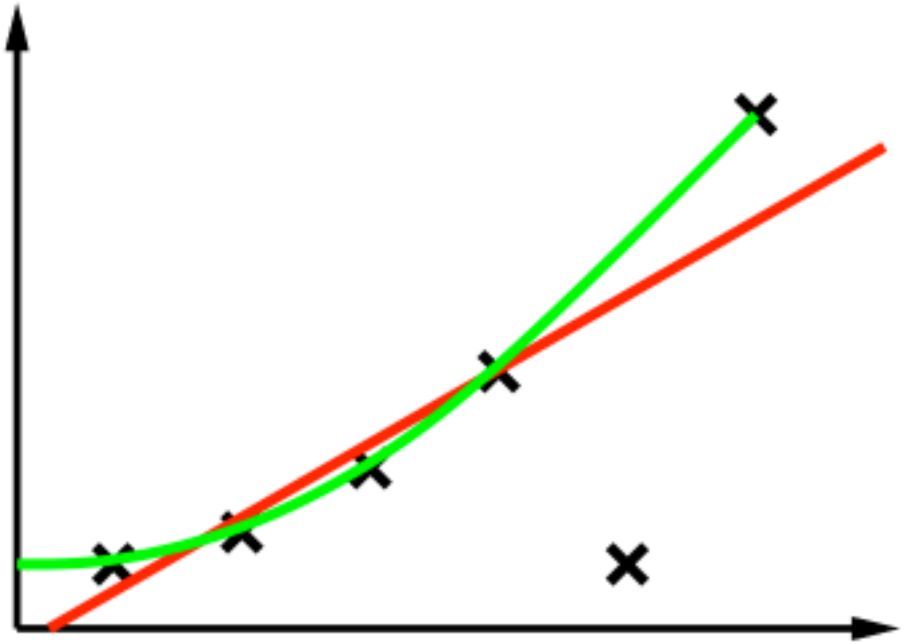
- Used for prediction: Linear regression

- Examples with only two variables: height and weight of people
  – We want to learn how to predict weight as a function of height



Example 1: height and weight of people
Feature Engineering

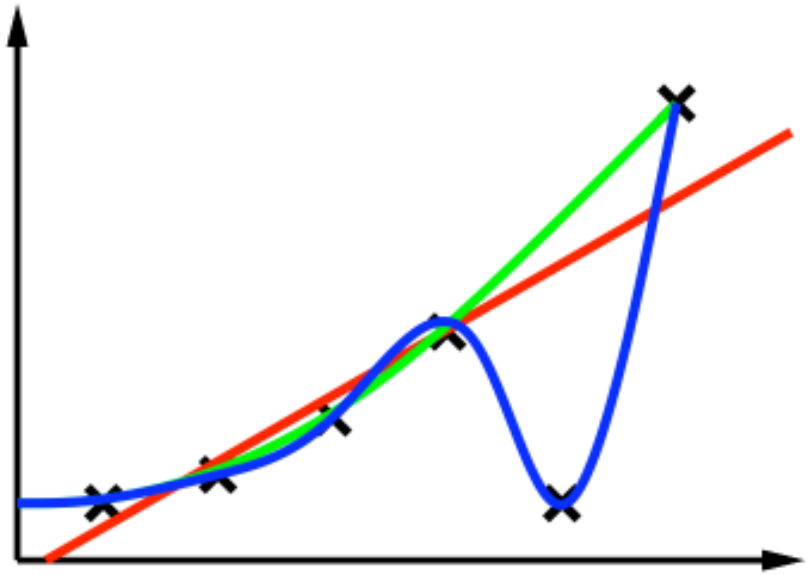
- Consider square of height
- Ignore outliers



Example 1: height and weight of people
Model training

- Balance two goals
  - Fit train data correctly, i.e. lead to good predictions on train data
  - Have a model as simple as possible to avoid overfitting.

- Overfitting can occur with regression, see blue curve below



Example 2: identify people as good or bad from their appearance

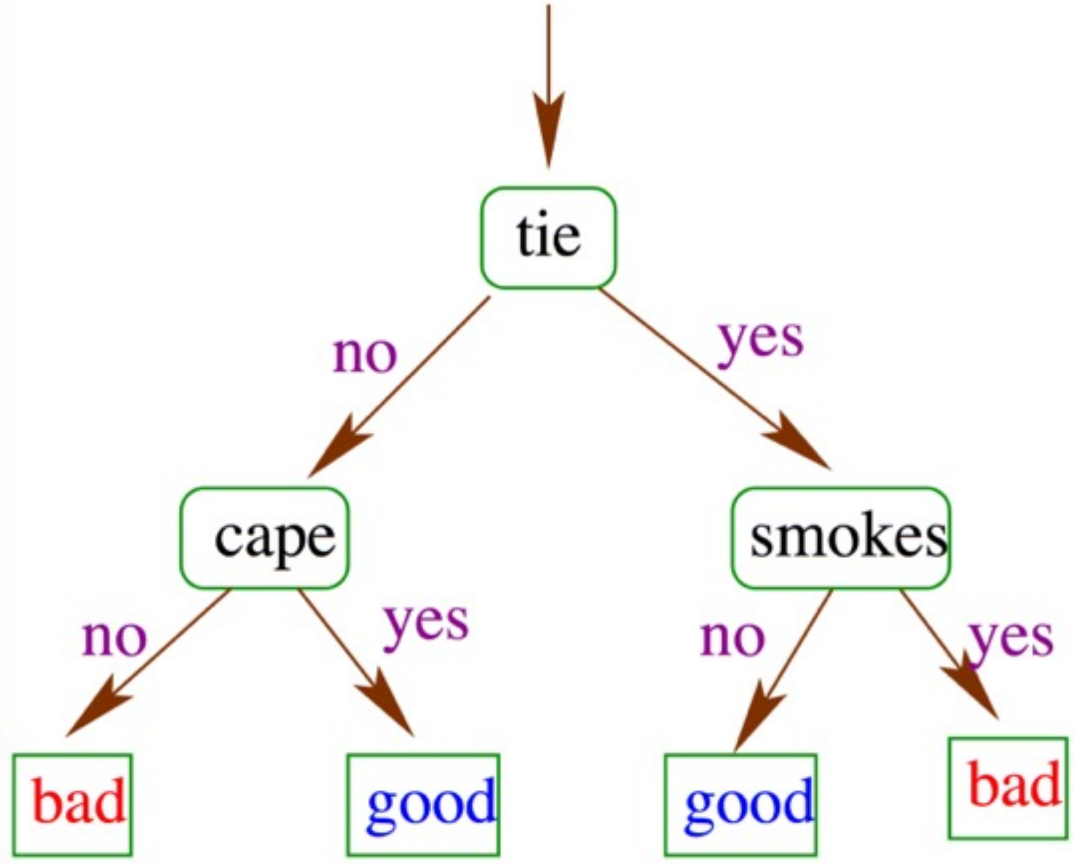
- Used for prediction: Logistic regression, Decision Tree, Support Vector Machines, …

	sex	mask	cape	tie	ears	smokes	class
			training data				
batman	male	yes	yes	no	yes	no	Good
robin	male	yes	yes	no	no	no	Good
alfred	male	no	no	yes	no	no	Good
penguin	male	no	no	yes	no	yes	Bad
catwoman	female	yes	no	no	yes	no	Bad
joker	male	no	no	no	no	no	Bad
			test data				
batgirl	female	yes	yes	no	yes	no	??
riddler	male	yes	no	no	no	no	??

Example 2: identify people as good or bad from their appearance

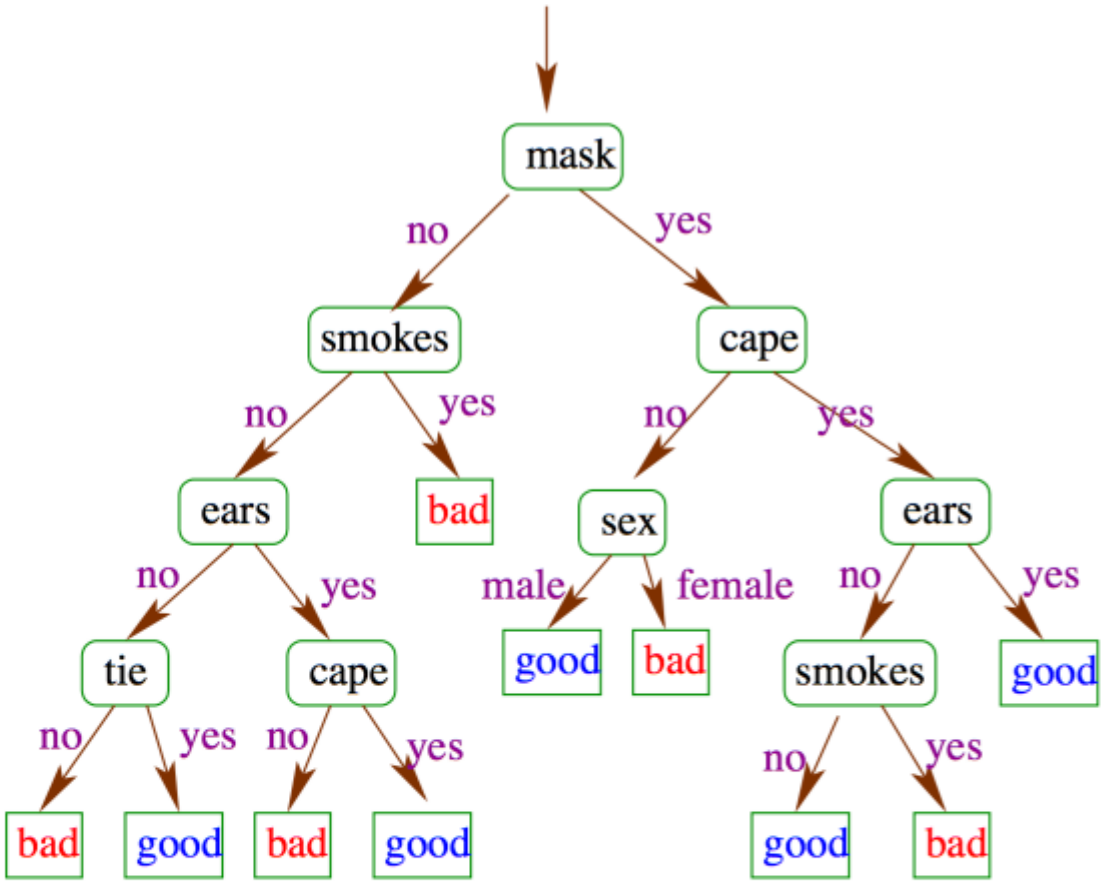
- A serie of tests

- Class assigned at the leaves of the tree



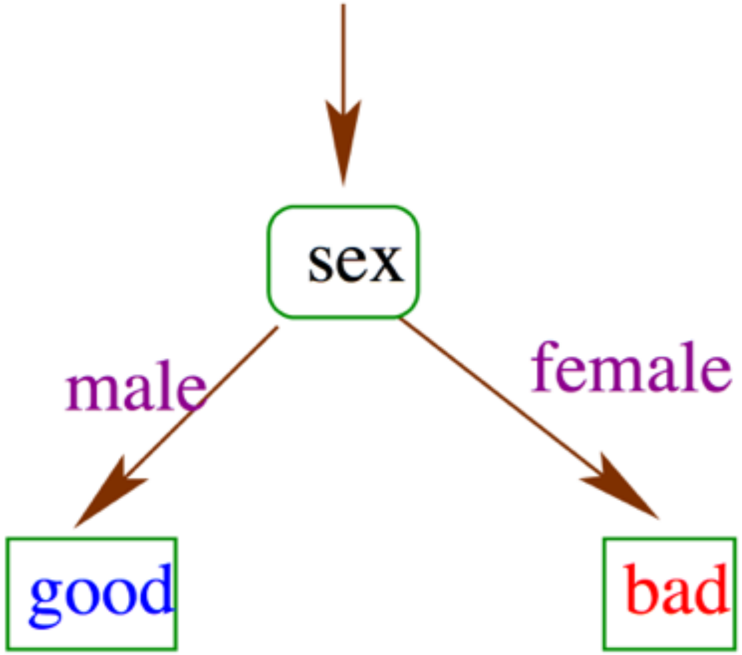
Example 2: identify people as good or bad from their appearance

- Classifies test data perfectly
- Performs poorly on new data
- Why?
  - Almost rote learning of train data
- We must favor simpler models
  - Trees with less nodes



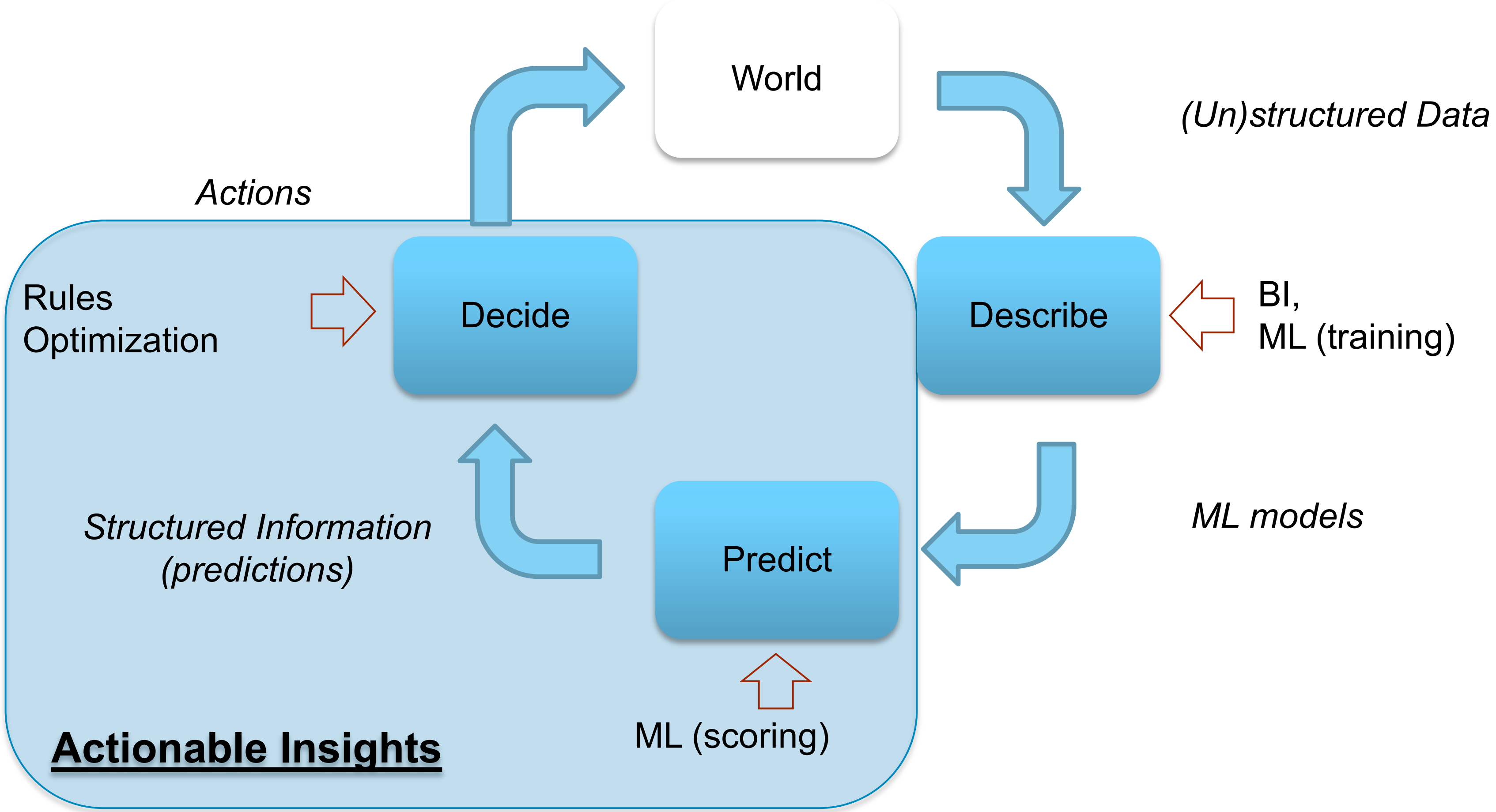
Example 2: Underfitting

- Too simple models do not have good predictive power either

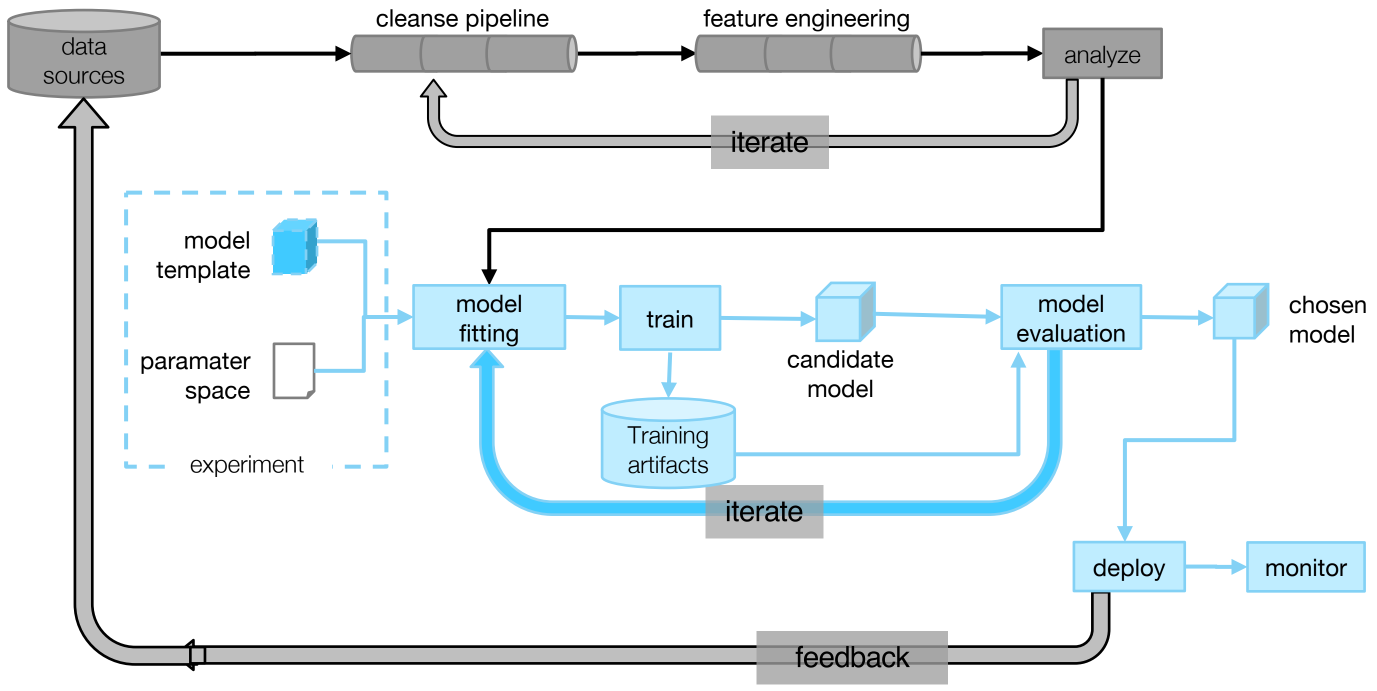


ARCHITECTURE

Machine Learning and Decision Making



Machine Learning architecture



Thank You

Yann Gouedo
Data Scientist Leader – Machine Learning / Artificial Intelligence
Marketing / Risk / Fraud / Maintenance / Pricing
Distinguished Data Scientist, Open Group Certification