

# ***Machine Learning primer***



***Machine learning*** is a method of data analysis that automates analytical model building. Using algorithms that iteratively learn from data, ***machine learning*** allows computers to find hidden insights without being explicitly programmed where to look (source: SAS)

# A word about myself



- My name is Gilles, I was born in France, spend my time between the US, France and Japan
- I hold M.Sc, B.Sc in Applied Mathematics and a Business degree
- I'm using mainly: Python, Scikit-Learn, R, Keras, Theano (for Neural networks) – although C++ was a long time favorite of mine
- Been doing Machine Learning and Analytics for more than 5 years – mainly in Finance and Marketing
- I am currently working on connected services for a large manufacturer
- My favorite games so far are Uncharted 3 and Last of Us, I like reading about history during my free time and to make a mess in my kitchen for the purpose of cooking nice food. Last dish was grilled salmon in mango salsa.

# Overview

## Objective:

- Develop an intuition of what Machine Learning is and does

## Audience:

- Interest in understanding the underlying mechanisms
- Interest in using and practicing Machine Learning
- Interest in exploring the potential of Machine Learning

## Content:

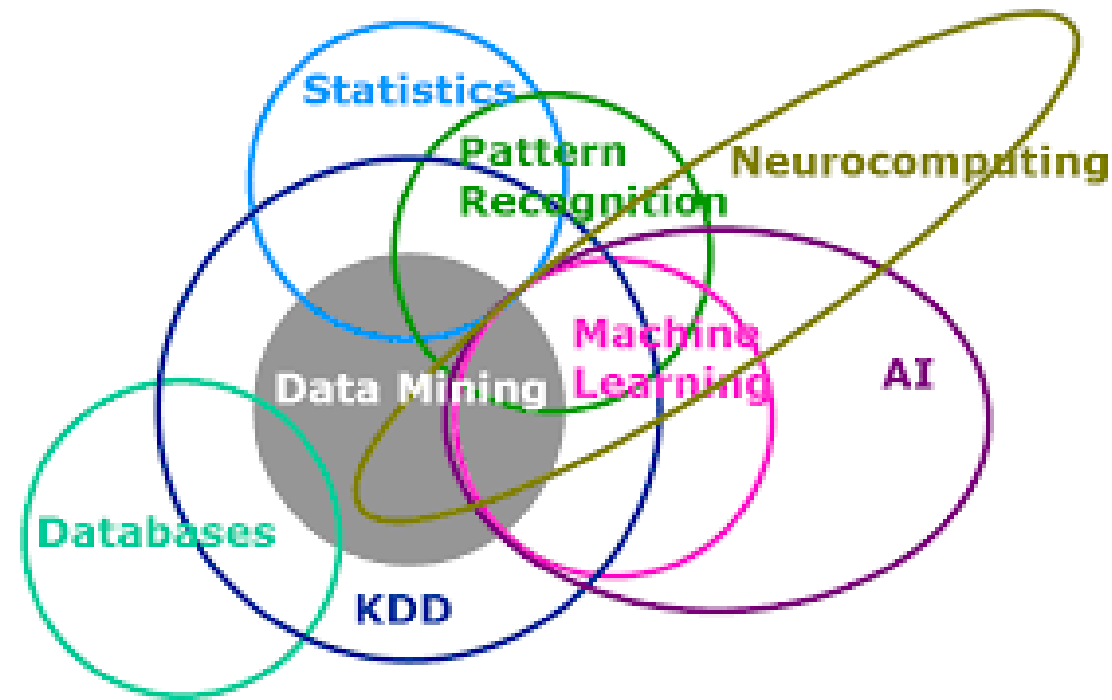
- Intro some comparison
- Math part:
  - Conventions
  - How it works at the core
- What it can do
- What it takes
- Challenges

# What do they say on the internet:

## Very different way of thinking

	<i>MACHINE LEARNERS</i>	<i>STATISTICIANS</i>
<i>Network/Graphs vs. Models</i>	<i>Network/Graphs to train and test data</i>	<i>Models to create predictive power</i>
<i>Weights vs. Parameters</i>	<i>Weights used to maximize accuracy scoring and hand tuning</i>	<i>Parameters used to interpret real-world phenomena - stress on magnitude</i>
<i>Confidence Interval</i>	<i>There is no notion of uncertainty</i>	<i>Capturing the variability and uncertainty of parameters</i>
<i>Assumptions</i>	<i>No prior assumption (we learn from the data)</i>	<i>Explicit a-priori assumptions</i>
<i>Distribution</i>	<i>Unknown a priori</i>	<i>A-priori well-defined distribution</i>
<i>Fit</i>	<i>Best fit to learning models (generalization)</i>	<i>Fit to the distribution</i>

What do they say on the internet:  
Very different techniques



# What do they say on the internet: How is it done?

## Supervised Machine Learning v. Econometrics/Statistics Lit. on Causality

### Supervised ML

- ▶ Well-developed and widely used nonparametric prediction methods that work well with big data
  - ▶ Used in technology companies, computer science, statistics, genomics, neuroscience, etc.
  - ▶ Rapidly growing in influence
- ▶ Cross-validation for model selection
- ▶ Focus on prediction and applications of prediction
- ▶ Weaknesses
  - ▶ Causality (with notable exceptions, e.g. Pearl, but not much on data analysis)

### Econometrics/Soc Sci/Statistics

- ▶ Formal theory of causality
  - ▶ Potential outcomes method (Rubin) maps onto economic approaches
- ▶ “Structural models” that predict what happens when world changes
  - ▶ Used for auctions, anti-trust (e.g. mergers) and business decision-making (e.g. pricing)
- ▶ Well-developed and widely used tools for estimation and inference of causal effects in exp. and observational studies
  - ▶ Used by social science, policy-makers, development organizations, medicine, business, experimentation
- ▶ Weaknesses
  - ▶ Non-parametric approaches fail with many covariates
  - ▶ Model selection unprincipled

# What do they (still say) on the internet:

## Very briefly

Machine learning	Statistics
network, graphs	model
weights	parameters
learning	fitting
generalization	test set performance
supervised learning	regression/classification
unsupervised learning	density estimation, clustering



# Vocabulary 1/3

- $X$  : observations, they can have  $k$  features (parameters)
- $Y$  : observed outcomes for observations
- Example:
  - $X$  (average km per day,  
#days since installation,  
#maintenance days)
  - $Y$  (useful life of part)



# Vocabulary 2/3

- $X$  : Data
- $X_{.,j} : X_{.,1} X_{.,2} \dots X_{.,k}$   $k$  Features that describe each data point
- $X \Rightarrow \left. \begin{array}{c} X_{1,.} \\ X_{2,.} \\ X_{3,.} \\ \dots \\ X_{n,.} \end{array} \right\} \text{N observations}$

# Vocabulary 3/3

- $F$  : Model
- $\Theta$  : Weights of the model
- $Y'$  : output of the Model
- Example:
  - $Y'(\text{useful life of part}) =$   
 $12 + \Theta_1 \times \text{\#Km/day} + \Theta_2 \times \text{\#days} + \Theta_3 \times \text{\#maintenance}$

# What is Machine learning all about?

- We want to build a model  $F$ , such as  $Y' = F(X)$  so we can generalize for any new entry  $X$  an expected outcome  $Y'$
- For this, we are going to train the model  $F$  by using all the outcomes  $Y$  we know about for observations  $X$
- We will look for the Weights  $\Theta_i$  of the model such that given  $Y$  and  $X$ , we have  $Y \sim F_{\Theta_i}(X)$

# In other words

- We want to minimize the expression:

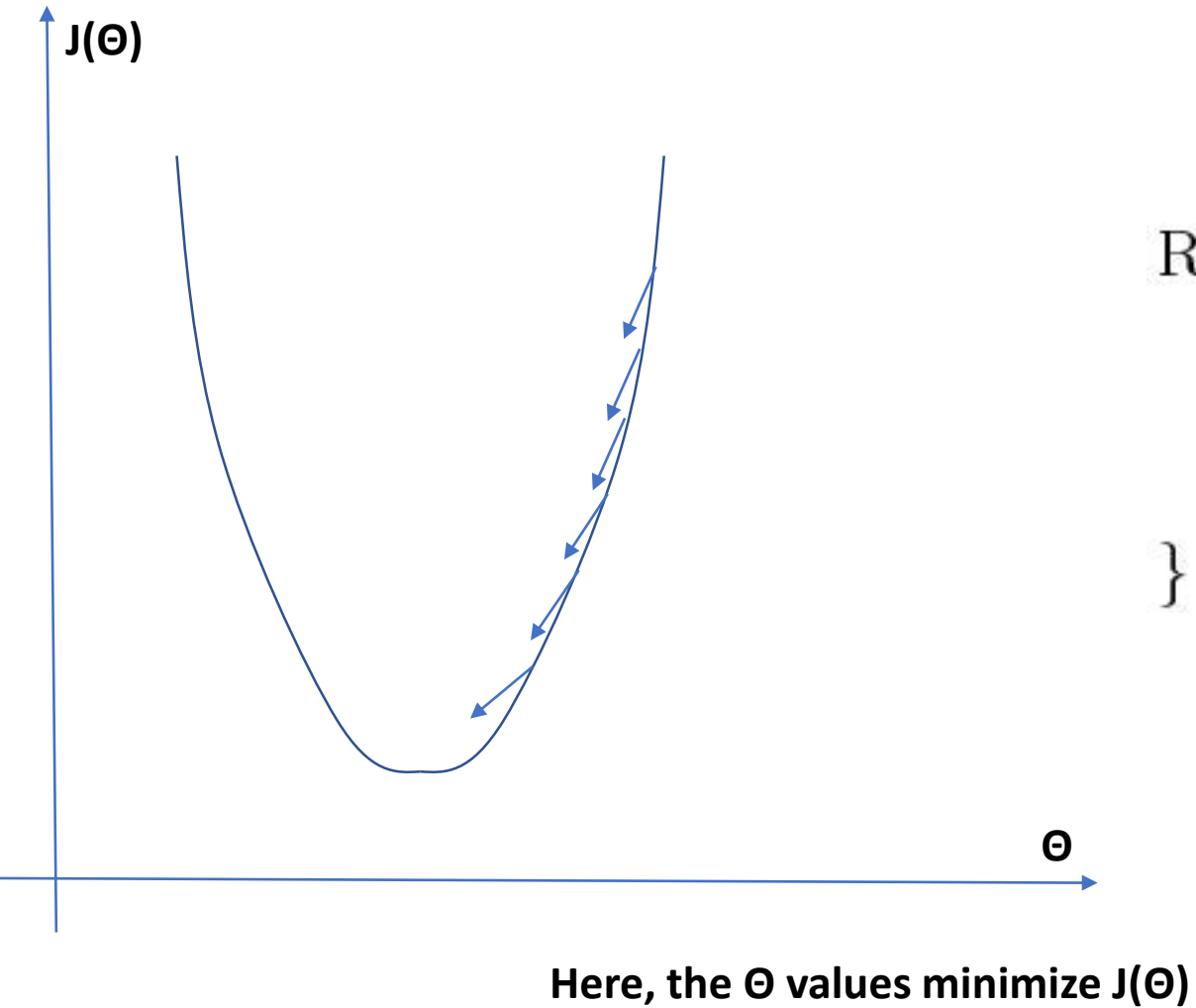
$$\underbrace{||Y - F_{\Theta_i}(X)||}_{J(\Theta_i)} \leq \text{Epsilon}$$

- Given that we know  $X$  and  $Y$  while we train the model and that we are trying to determine the parameters  $\Theta_i$  of  $F$ , this is equivalent to a minimization problem on  $\Theta$

$$J(\Theta) \leq \text{Epsilon}$$

solved by **gradient descent** algorithm

# Gradient descent



Repeat until convergence {

$$\theta_j \leftarrow \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta)$$

}

By iterating the algorithm on all the data points  $X$  and known outcomes  $Y$ , we can determine the model

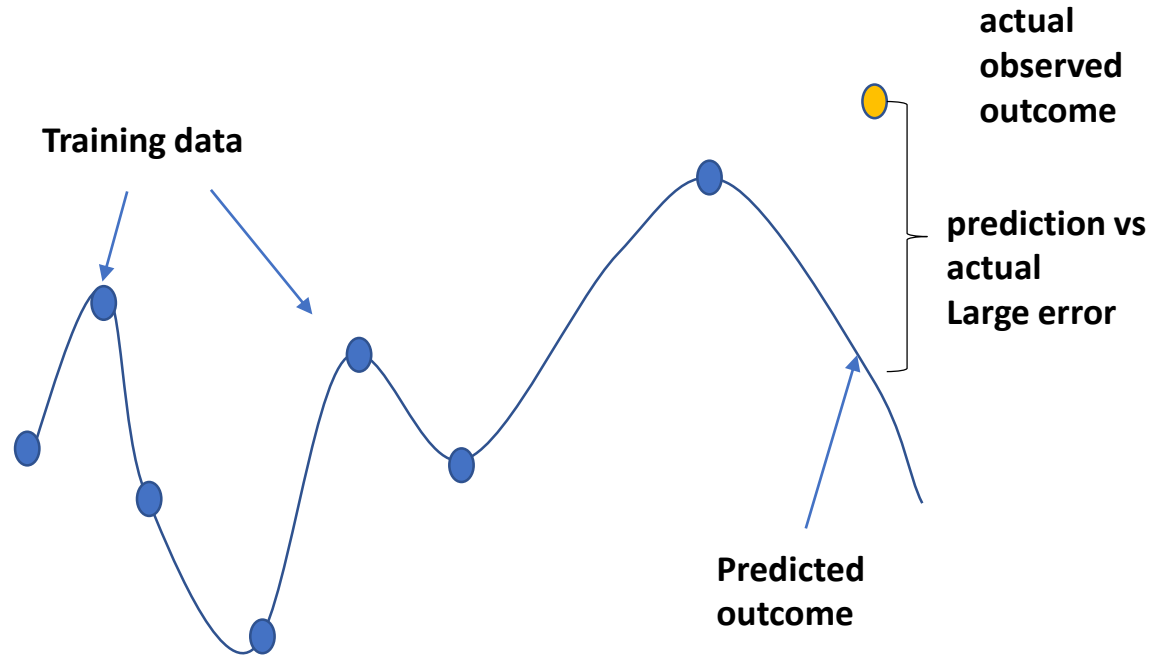
# Why is it (really) state of the art now?

- We have the ability to collect a lot of data
- We have computing power
- It is a mathematical solution that can extract information quickly, elegantly, optimally without (too much) prior knowledge of the underlying phenomena

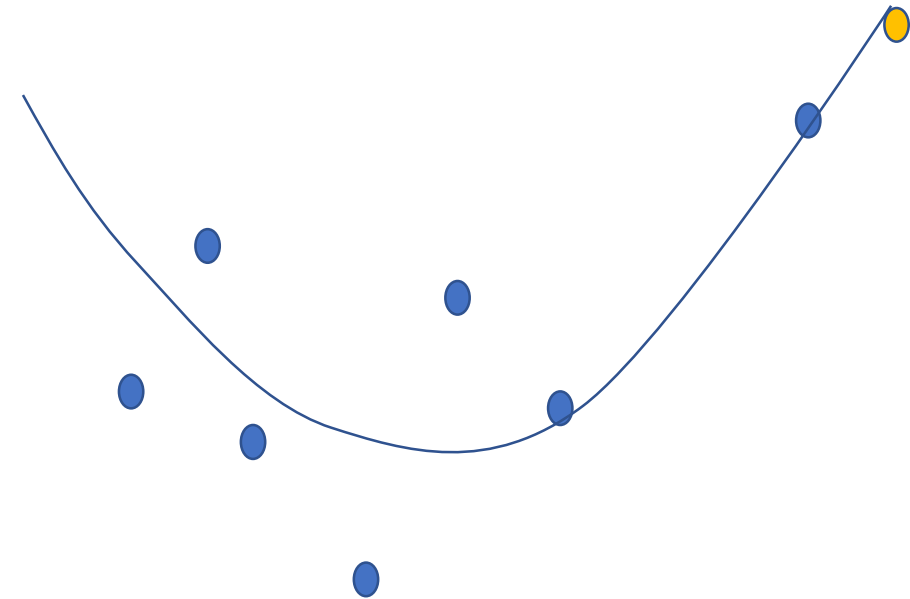
**In short, this is big data**

# Pitfalls

- Overfitting (model is too sophisticated and doesn't generalize well)



**Overfitting: model very sophisticated and fits too well the training data**

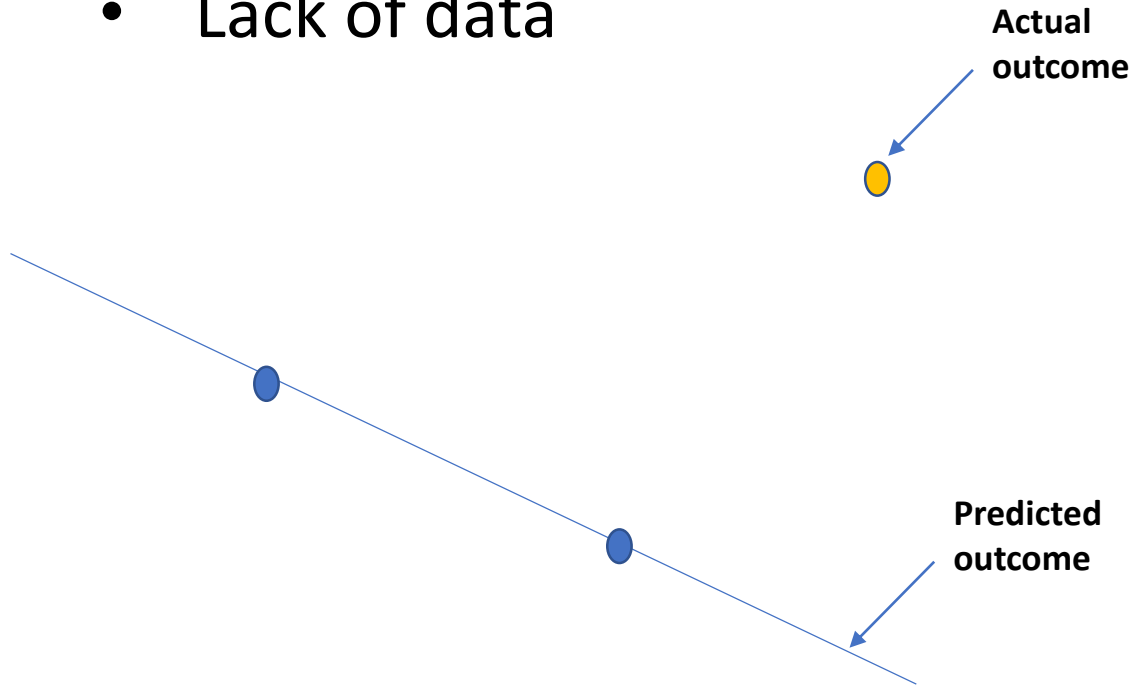


**Model has a larger training error but generalize better the prediction of new observations**

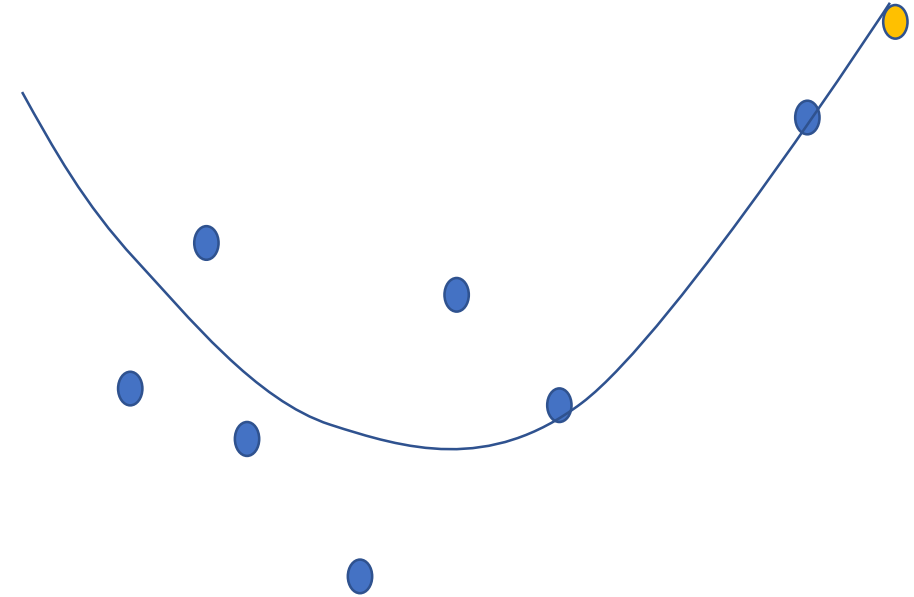


# Pitfalls

- Lack of data



**Not enough data, models can be poor**



**Good amount of data, models are better  
(5,000 pts can be nice, 1,000 starts to be challenging)**

# In practice

- We use 70% of the data to train the model and 30% to test its validity (confirm we do not have overfitting and model can generalize well to new data)
- Machine learners cannot accept overfitting and fortunately a few techniques exist to avoid it

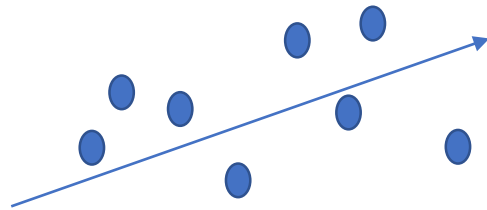
**We constantly check the training error vs generalization error**

# A few points

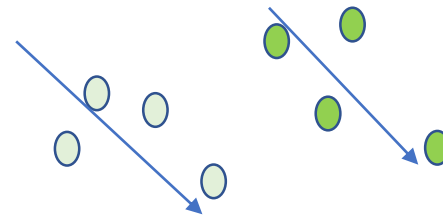
- Preferable to have numerical data (most algorithms are mathematical)
- Preferable to have lots of data (law of big numbers)
- Very important work of pre-processing the data (scaling, cleaning, missing values, handling of categorical data, dimension reduction)

# Where's the game?

- Being able to formulate the objective (it is not trivial at all!)
- Picking the most appropriate model
- Feature engineering, if we have  $X$ :
  - we may use also  $X$ ,  $\sin(X)$ ,  $X^2$ , ...
  - Momentum, velocity, filters, pre-training, etc...
- Constantly keep in check and reduce the model error
- And be careful



Demand is increasing  
in total population...



Same data, if we look at females and males separately  
demand is actually decreasing!

# Models are used (mainly) for

- Classification (sort data according to classes)
  - customer C will be more interested in owning a car
  - customer D will be more interested in sharing a car
- Regression (predict a value)
  - Number of cars a highway restaurant is likely to receive between 6pm and 9pm the first week of August
- Clustering (automatically group customers with same underlying characteristics)
  - Customers who purchased white Leaf cars are likely to subscribe to infant service media programs
  - Automated segmentation through underlying similarities among people who showed interest in mini cars

Keep in mind the words “supervised” and “unsupervised” (basically does the training requires a Y or not? Typically clusters don’t) that are used in Machine Learning (question of costs. Labeled data is generally more expensive).

# A word about Neural networks

- A very nice set of tools to have:
  - It is *not* “standard Machine Learning” vs NN
  - Again best model should be chosen for the task at hand
- Can do (best of range and my favorites)
  - Image recognition (CNN convolutional network) – it works
  - Time series and prediction (LSTM – Long short term memory) – it works
  - Dimension reduction, anomaly detection (SOM, Auto-encoders) – it works
- Can do also
  - NLP (Natural language processing) – great field for automated human machine communication (chatbots), requires a lot of expertise, still a lot of research is being done but I will wait for the libraries
  - Generative networks: create new digital material from past data (not much experience on this)

# Challenges

- Models can be (painfully) more accurate and faster than humans
- Mathematics and AI are difficult to explain, especially why they work (if you have temperature, you are sick. It is the same with ML,... it measures the symptoms of a system to understand a situation)
- Focus of ML is on the outcome, the why is complex and not primary – keeping in mind that the “real” human why cannot but be biased (by experience and education in the best cases) and is actually not always super accurate

**Big Data is challenging, many possibilities but requires a different way of thinking**



# Cultural difference

**Traditional Statistics**

**Machine Learning**



**A Data Science Continuum**

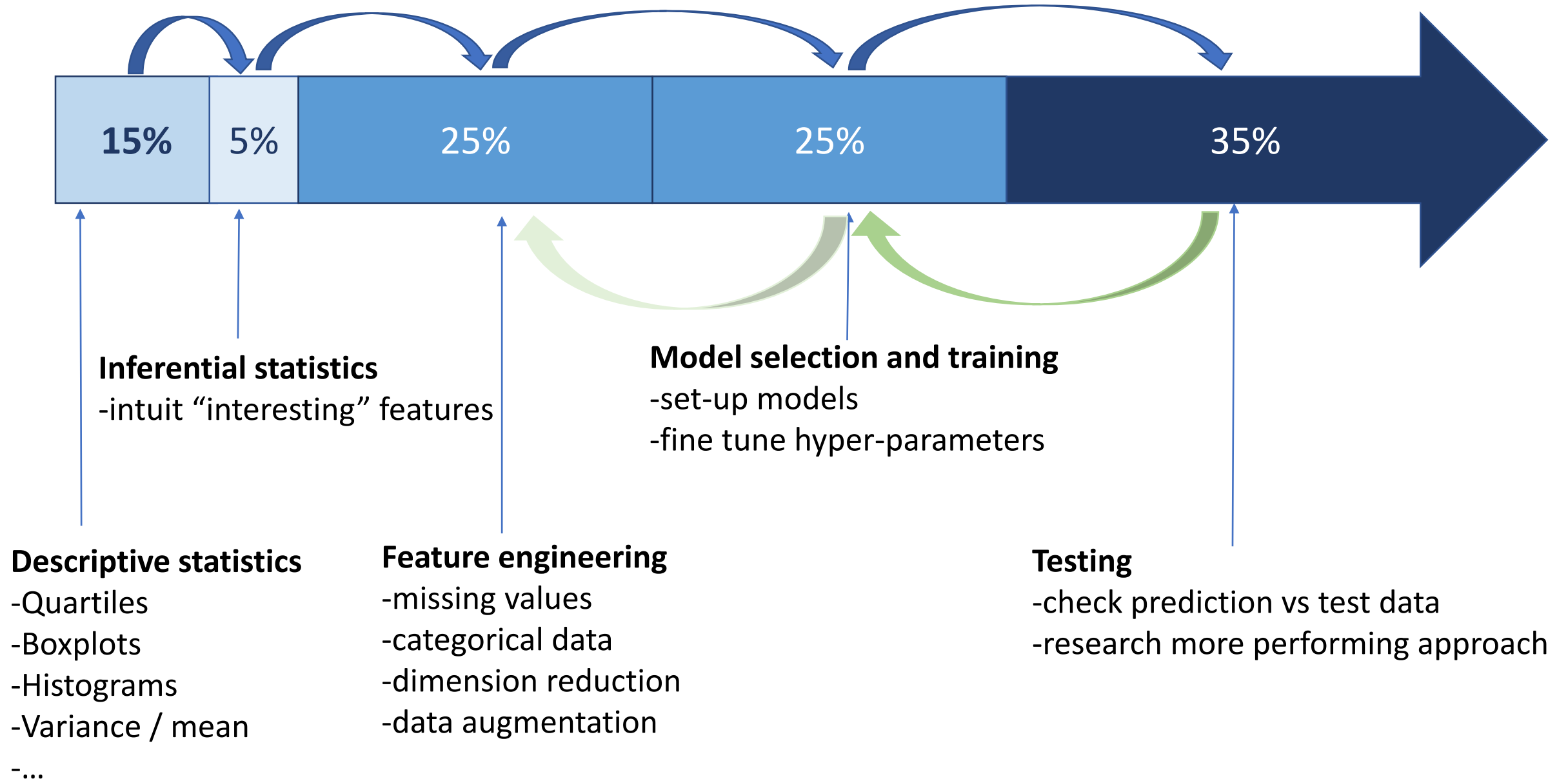
**White-box modelling**

simpler computation, emphasis on introspection, form, causal effects and processes, finding a 'correct' model

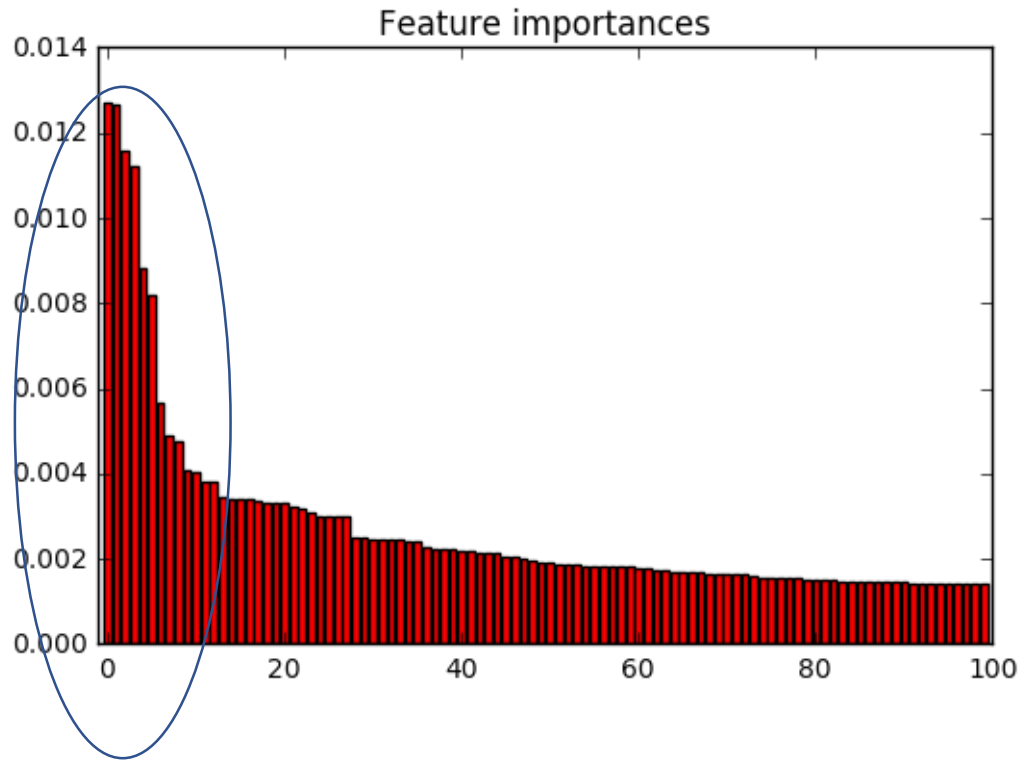
**Black-box modelling**

high computational complexity, emphasis on speed and quality of prediction, finding a 'performant' model

# Machine Learner mental workload

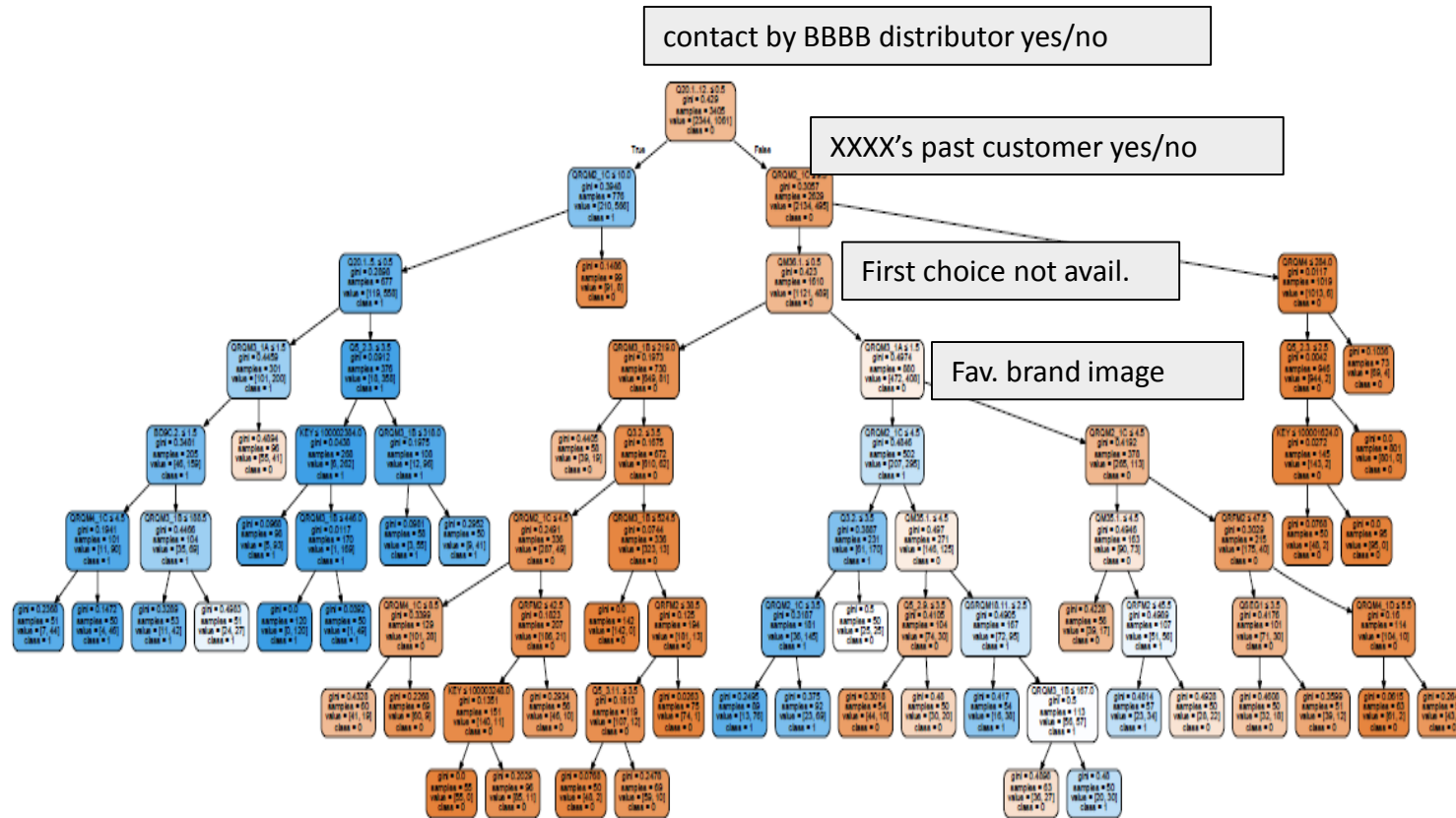


# Purchase prediction of XXXX product



- Possible to predict purchase of XXXX product with **84% accuracy** (random tree forest algorithm used)
- Less than 20 parameters really significant for prediction (dimension reduction algorithm)

# Main determining factors



## Determining factors\*:

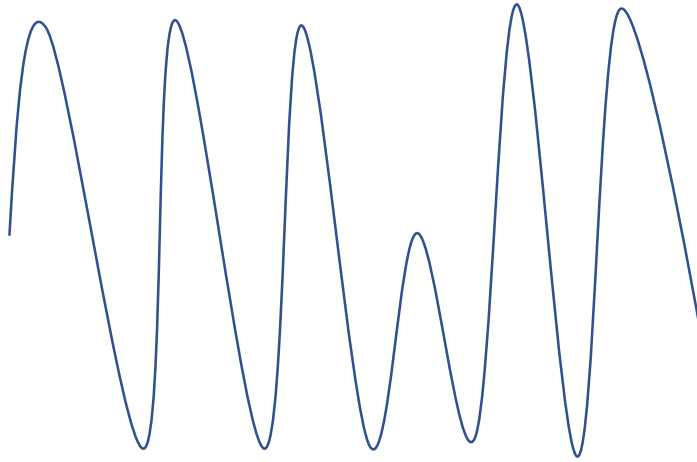
- Contact by BBBB distributor
- XXXX's past customer
- First choice of product not available
- Favorable brand image

(Blue likelihood to purchase XXXXX product)

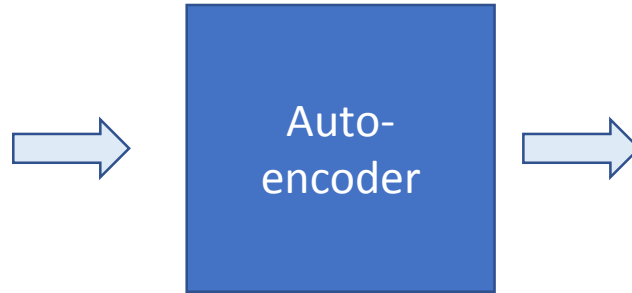
\*data may actually be misleading (strong warning!!! questionnaire needs clarification)

# Anomaly detections

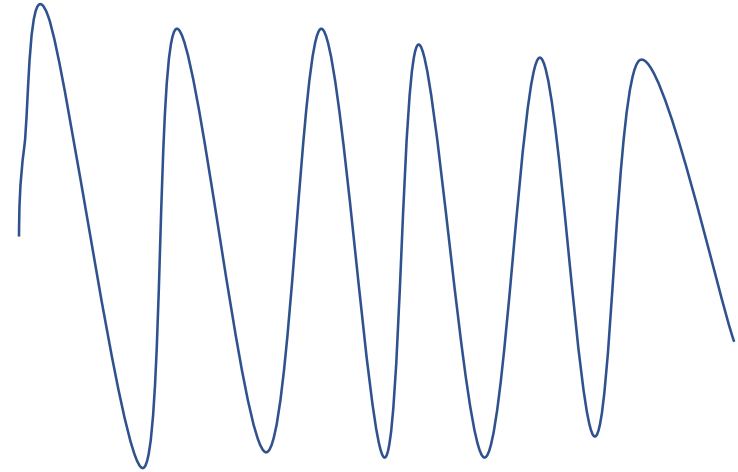
1 Input signal



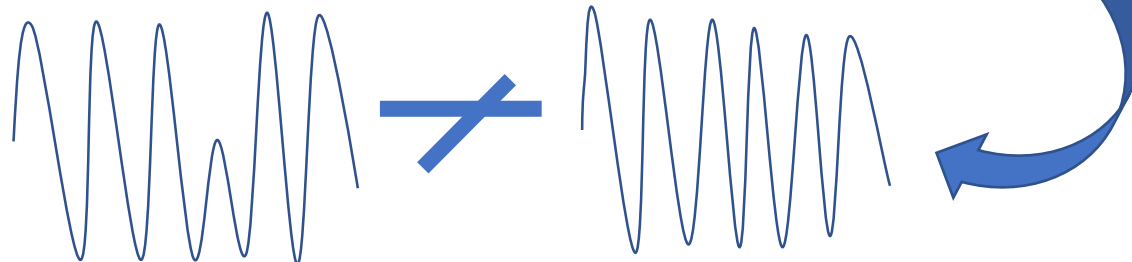
2 Auto-encoding



3 Expected reconstructed signal



4 Comparison



Anomaly (e.g. breaks not reacting as expected. Possible wear)

# Many terms (view from SAS)

## MACHINE LEARNING AND SOME OTHER TERMS YOU OFTEN HEAR

