Application of Machine Learning to ATLAS Physics Analysis

BDP(Bachelor's degree project) case of study

Samuel Campo

Overview

What is about Machine Learning and models?

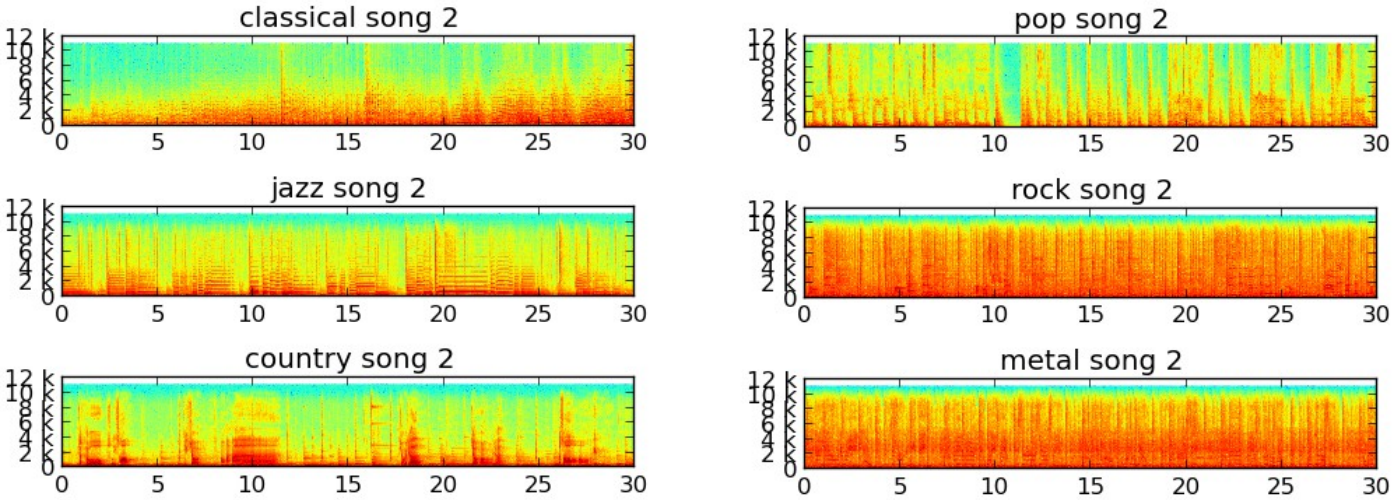
How is related data and model for choosing wisely?

Which are the Machine Learning analysis options?

Machine Learning?

To start:

- **Recognise complex patterns**



- **Ask to data and wait for an understandable answer**

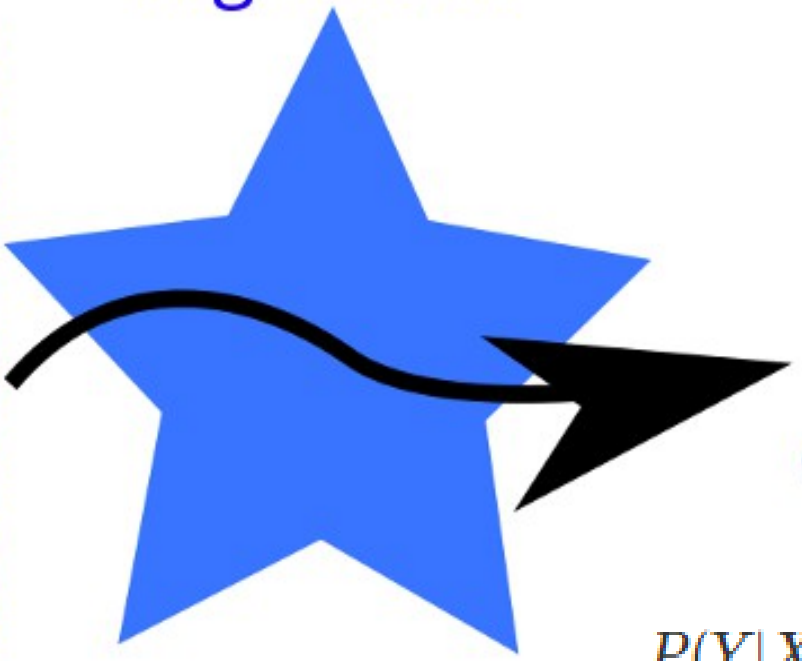
What is a model?

LEARNING

Data

Algorithm

Model

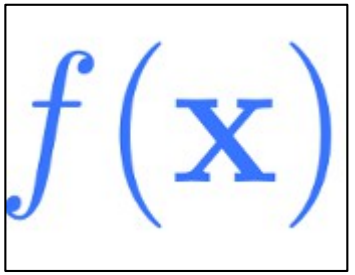


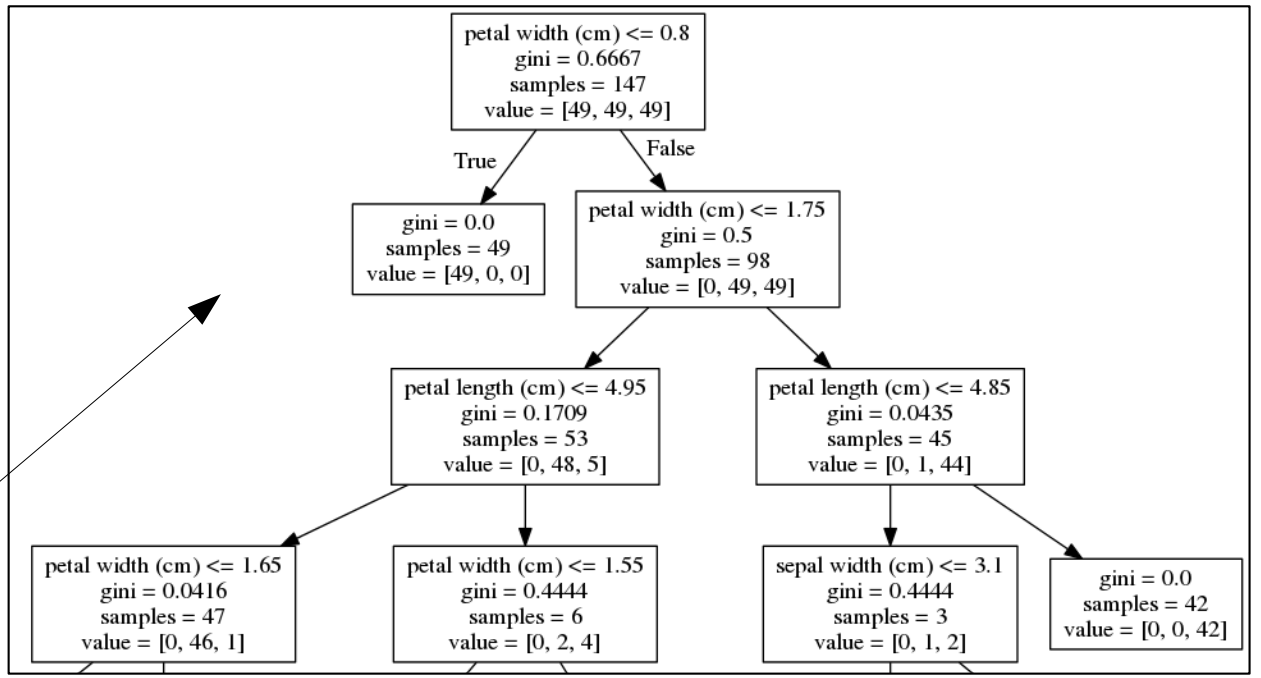
$f(\mathbf{x})$

$$P(Y|X) = \theta_1 x_1 + \theta_2 x_2 + \ldots$$

What makes a model?

ANALYSIS





Which model should I choose?

Depends on:

- Which **data type** you have

- What **information** you want to extract

Data types

Label Data

sepal length(cm)	sepal width(cm)	petal length(cm)	petal width (cm)	type
5,0	3,3	1,4	0,2	Iris-setosa
7,0	3,2	4,7	1,4	Iris-versicolor
5,9	3,0	5,1	1,8	Iris-virginica

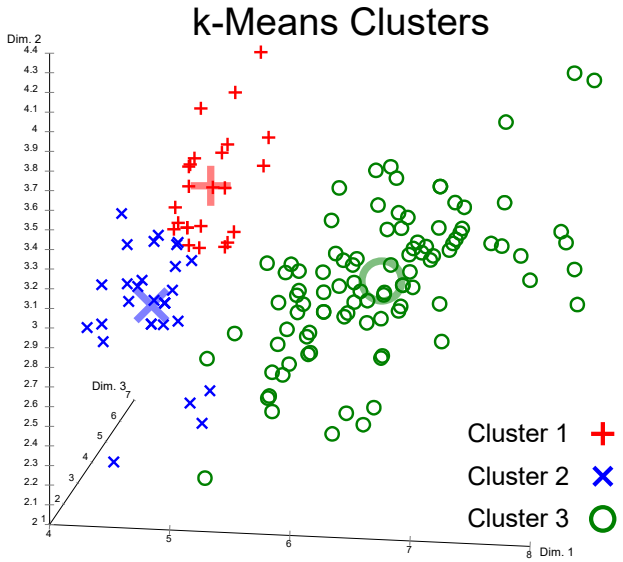
Unlabel Data

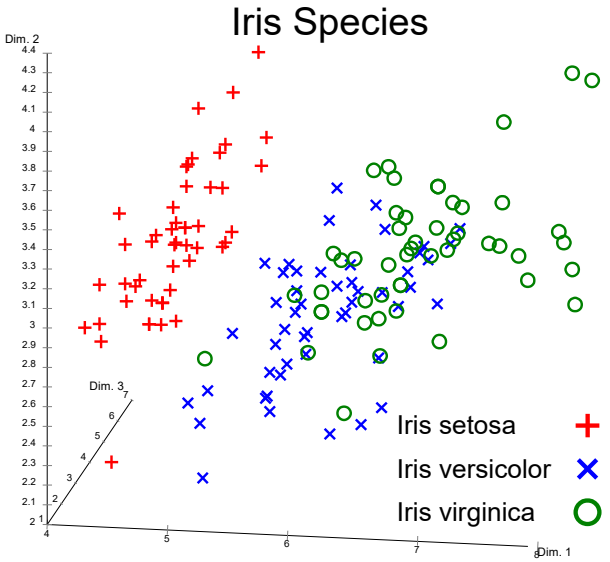
Height (cm)	Weight (kg)	Age (years)	Buys (€)
172	80	14	20
150	43	50	400
162	46	34	1000
160	65	89	10

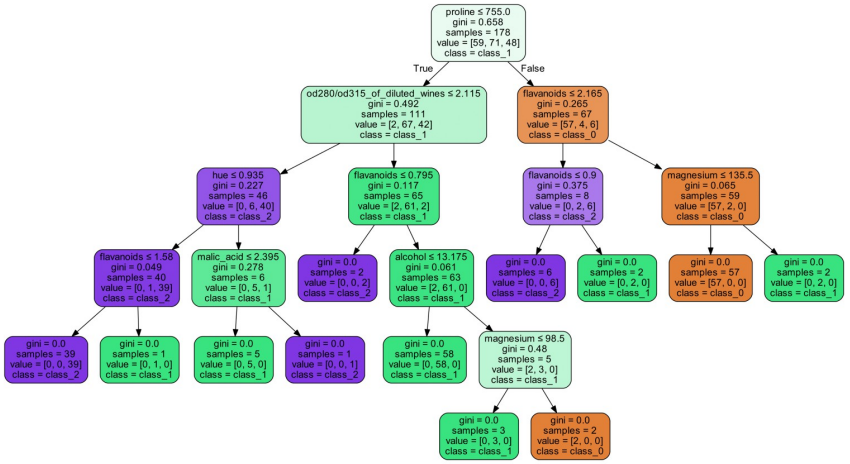
Data types – Learning model type

Label data ⟶ Supervised learning

Unlabel data ⟶ Unsupervised learning



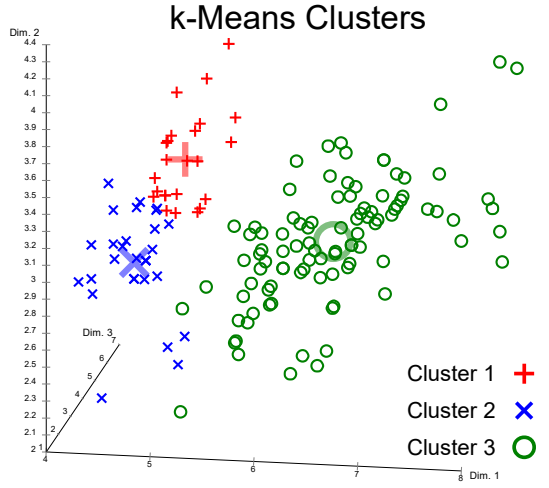


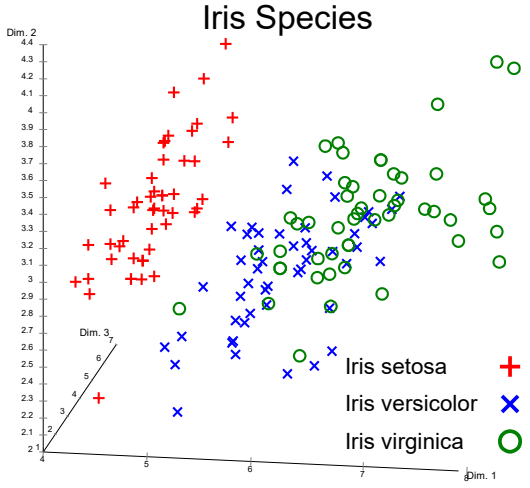


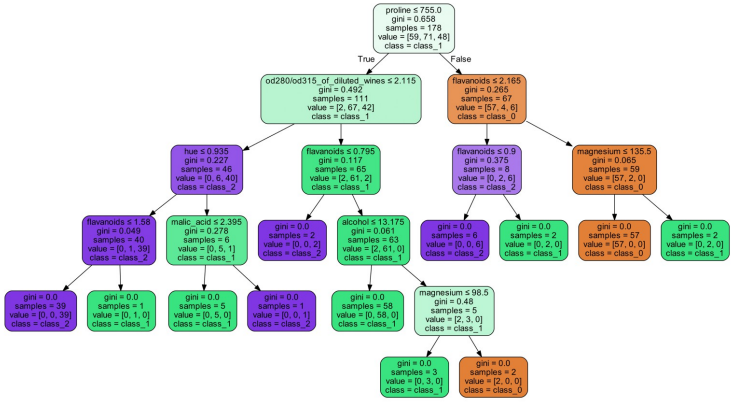
Unsupervised and Supervised Learning

Unsupervised: *interpreting and processing models*
Supervised: *predictive and descriptive models*

	Unsupervised Learning	Supervised Learning
CONCLUSION	-	+
TEST	-	+
FEATURES ANALYSIS	+	-
PROCESS DATA	+	-
CONTROL	-	+

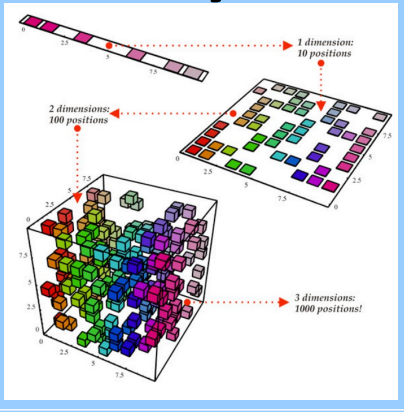




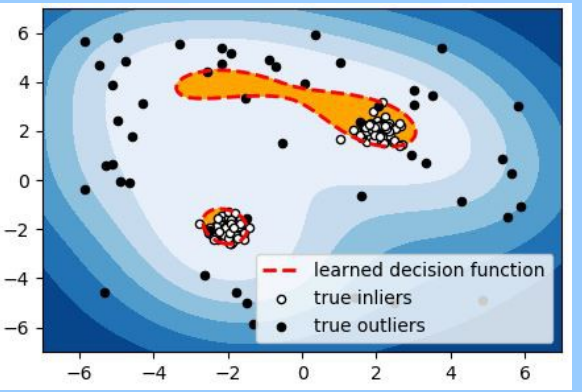


Information from data by ML

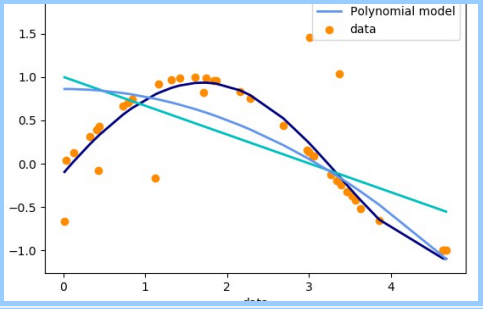
Dimensionality Reduction



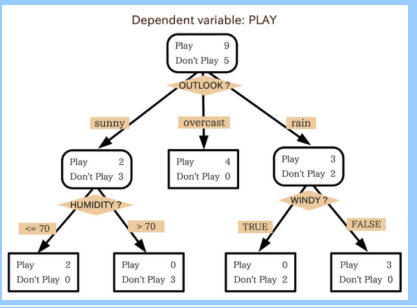
Anomaly detection



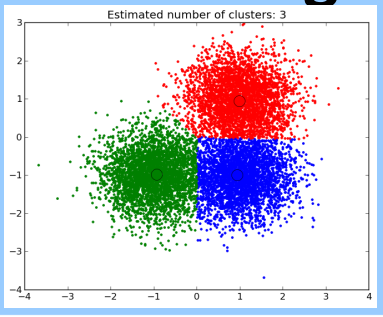
Regression



Classification



Clustering

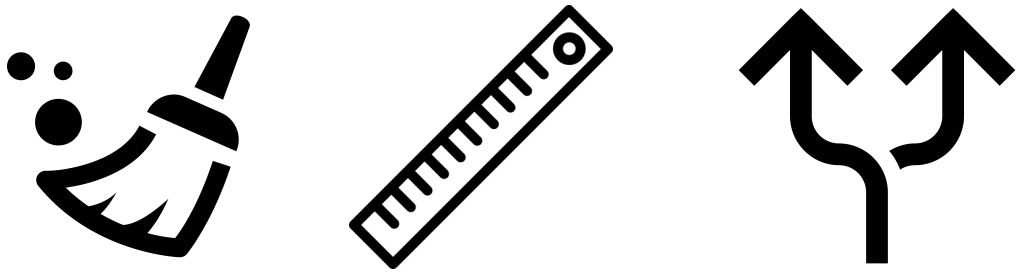


Generation

Machine Learning Workflow

Gather Data > Prepare Data > Choose Model (Train > Evaluate > Hyperparameter Tuning) > Prediction

Prepare Data: *Clean, scale and train/test split*



Choose Model: *Train, evaluate and maximize accuracy*

Our case of study

Pseudorapidity	Missing Momentum	Angle	Invariant mass	Model
2,1245	1,5123	-2,7362	1500	1
0,5787	0,1913	-2,4253	500	1
-1,4872	0,1008	2,9731	750	0
-0,4587	-2,0445	2,8722	1000	1
1,0124	-0,6498	-2,5371	1250	0
0,7652	-1,0118	2,2763	1000	1
-0,1241	1,6875	2,7352	1000	1

Models we have used:

Classification: *Neural Network, Decision Tree, Logistic Regression, etc.*

Models we are going to use:

Clustering: *K-means*

Generation: Generative Neural Network

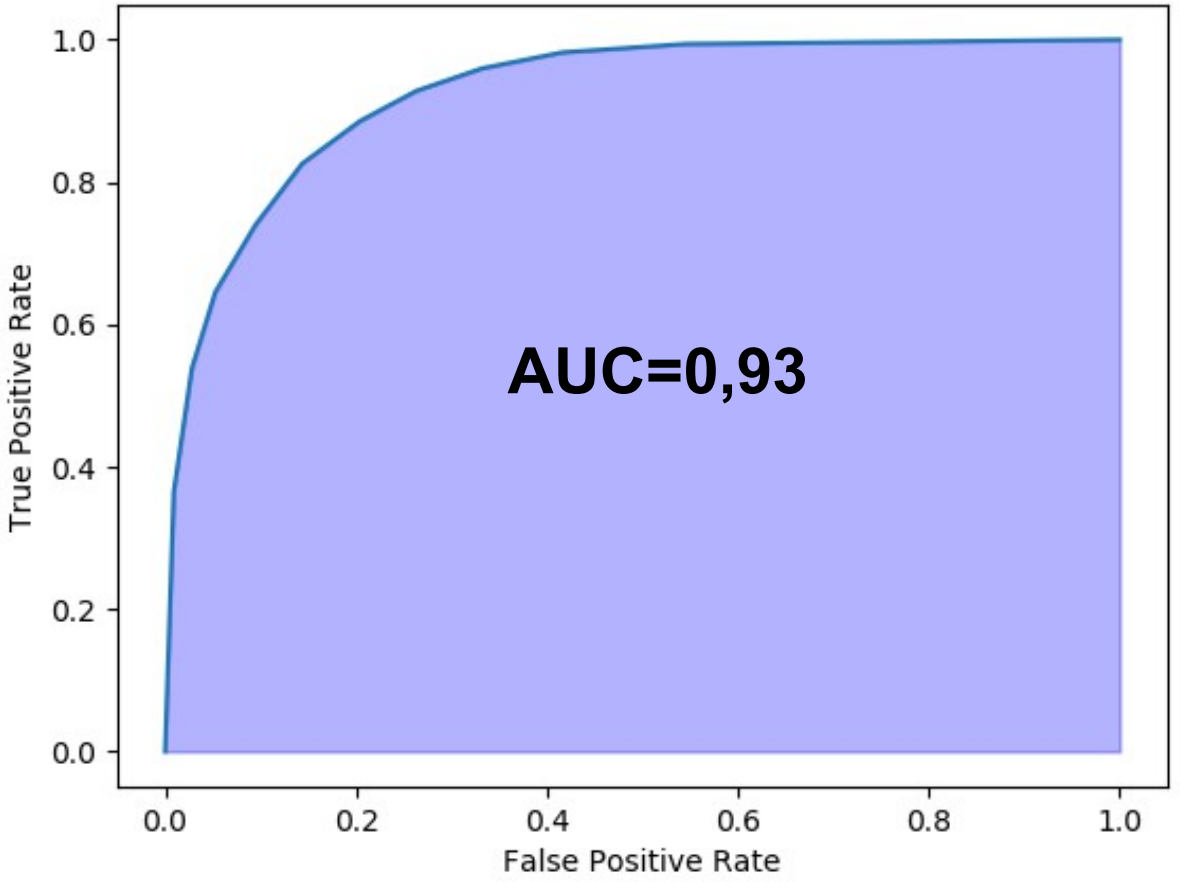
Showcase of Machine Learning

- 28 features to analyse with.

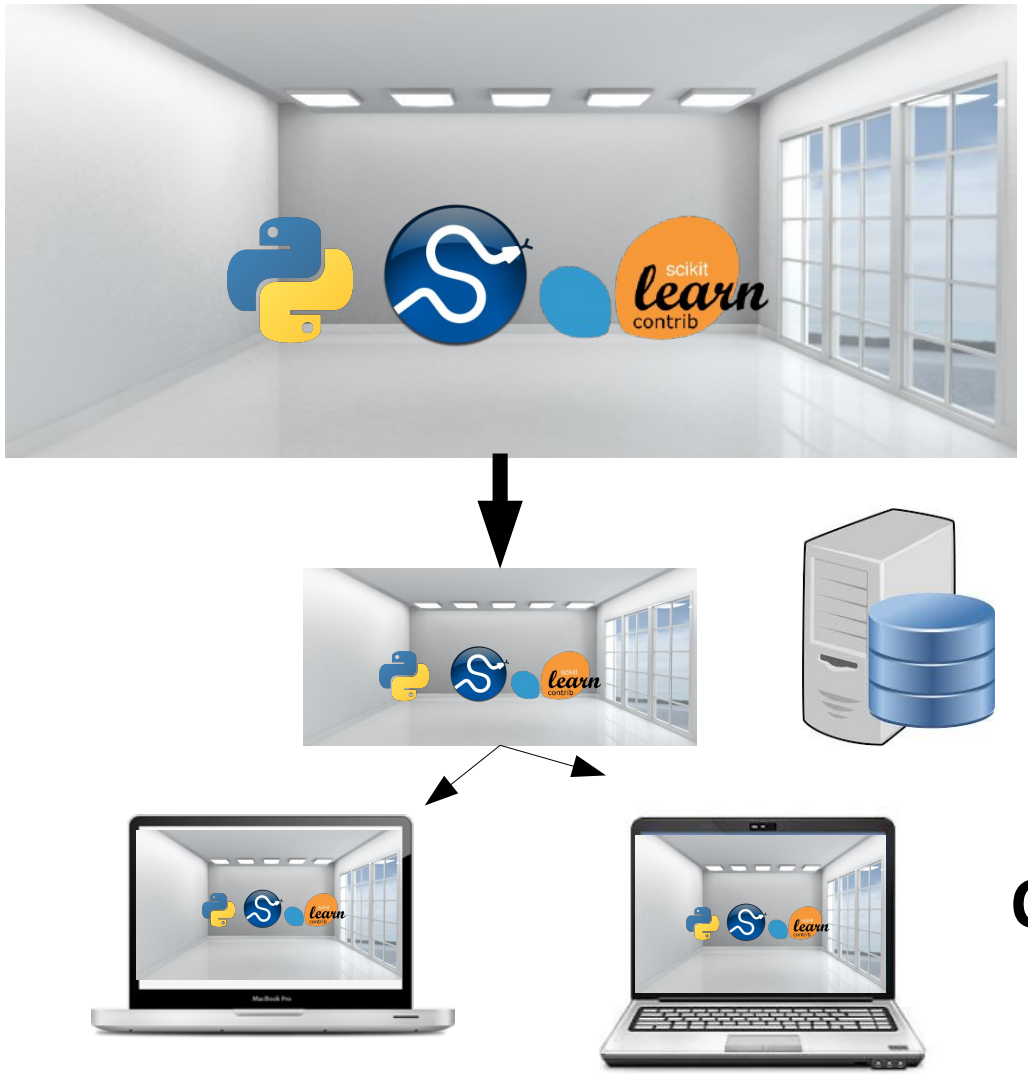
- More than 7,000,000 events.

- 2 Labels.

Classification for 2 labels among 196,000,000 non-integer numbers



Code Laboratories: Enviroments and Dockers



Local computer

Shared enviroment

Companion local computer

Code Laboratories

Anaconda 

Enviroments which can be copied and shared, but you **need some installation**.

Easiest and more common

Docker 

Containers that can be copied and shared, and you have **NO need of installing** nothing at all.

Lightest and more powerful

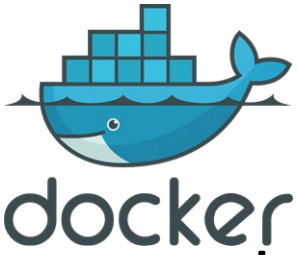
Open Fronts

Nowadays our project has these open fronts:

- Make real use of **SWAN** on our study

- Put together ML models trained in **Python and ROOT**

- Implement **Docker** as a common tool on our development



- Improve our analysis on **Neural Network using GPU's**

To sum up:

Machine Learning is used for **pattern analysis** and **prediction**

Docker and **Anaconda** are really good friends for **Science research** using Machine Learning

References to learn more...

Machine Learning explanations and examples:
https://machinelearningmastery.com/
http://scikit-learn.org
https://stackoverflow.com/

Datasets to practise:
https://archive.ics.uci.edu/ml/index.php
https://www.kaggle.com/datasets

Thanks for your attention!