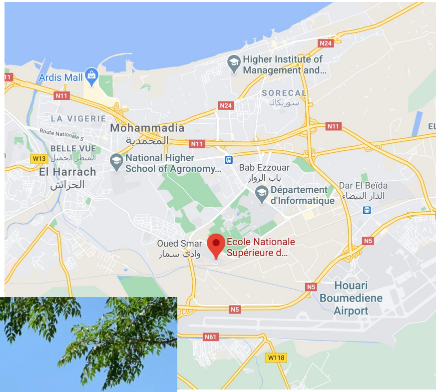
Machine Learning for High Energy Physics

Second HEP Graduate Workshop, 3th April 2021

Dalila Salamani (CERN) , Sabrina Amrouche (UniGe)

About Us

Ecole Nationale Supérieure d'Informatique



2016

ML 4

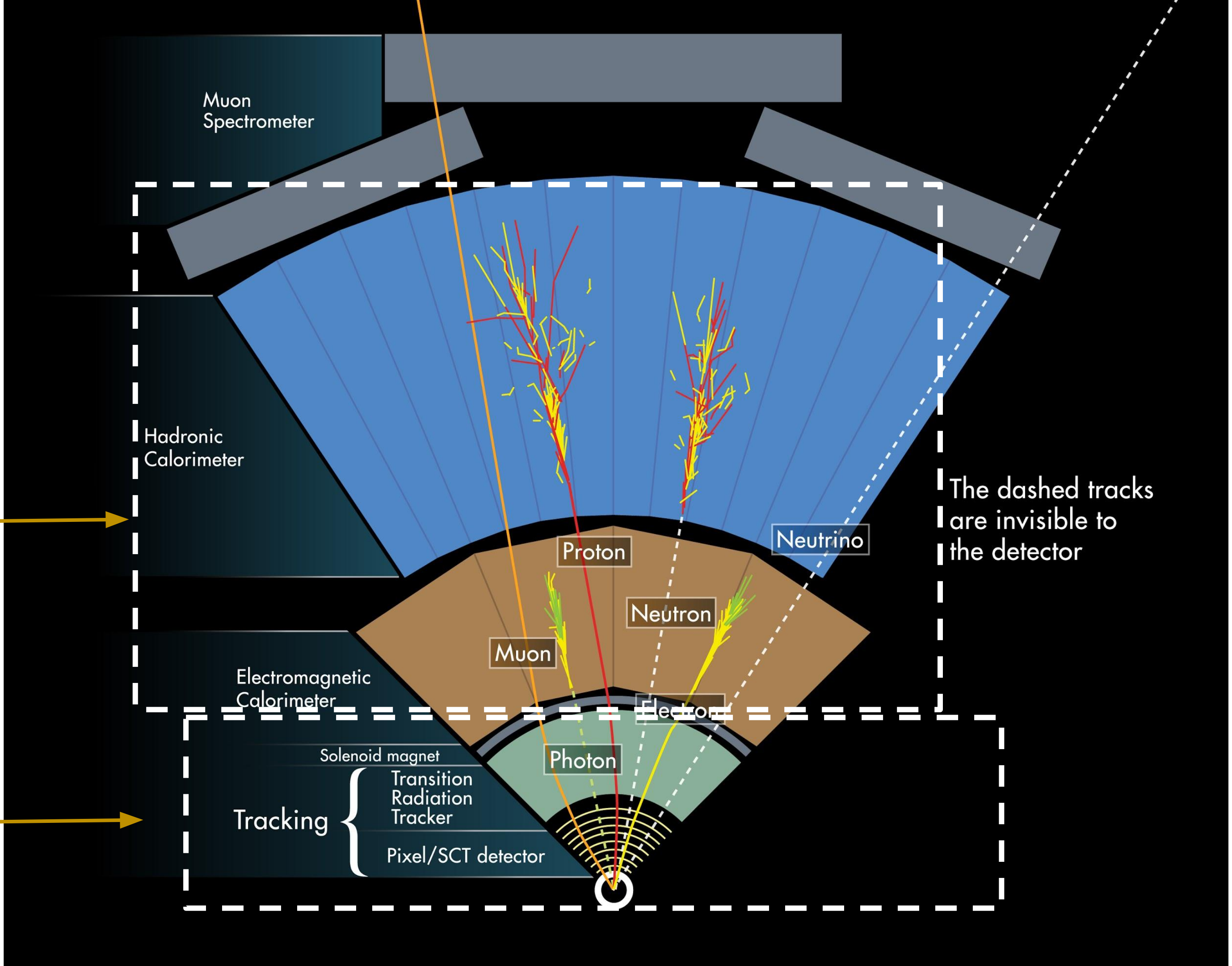
Particle Tracking

Shower Simulation

PhD

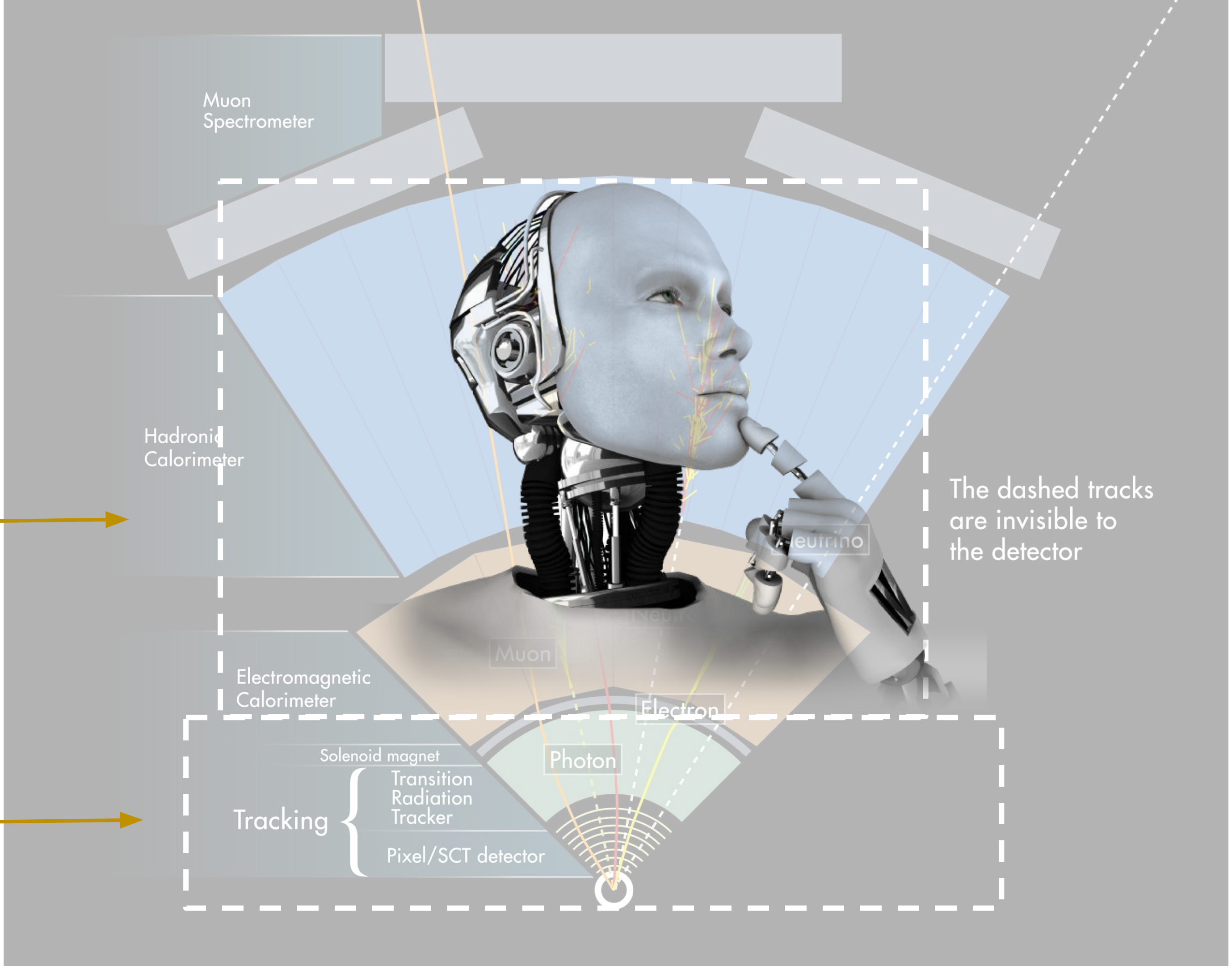
Dalila

Sabrina



Dalila

Sabrina



Glossary

Data Mining

Artificial Intelligence

Deep Learning

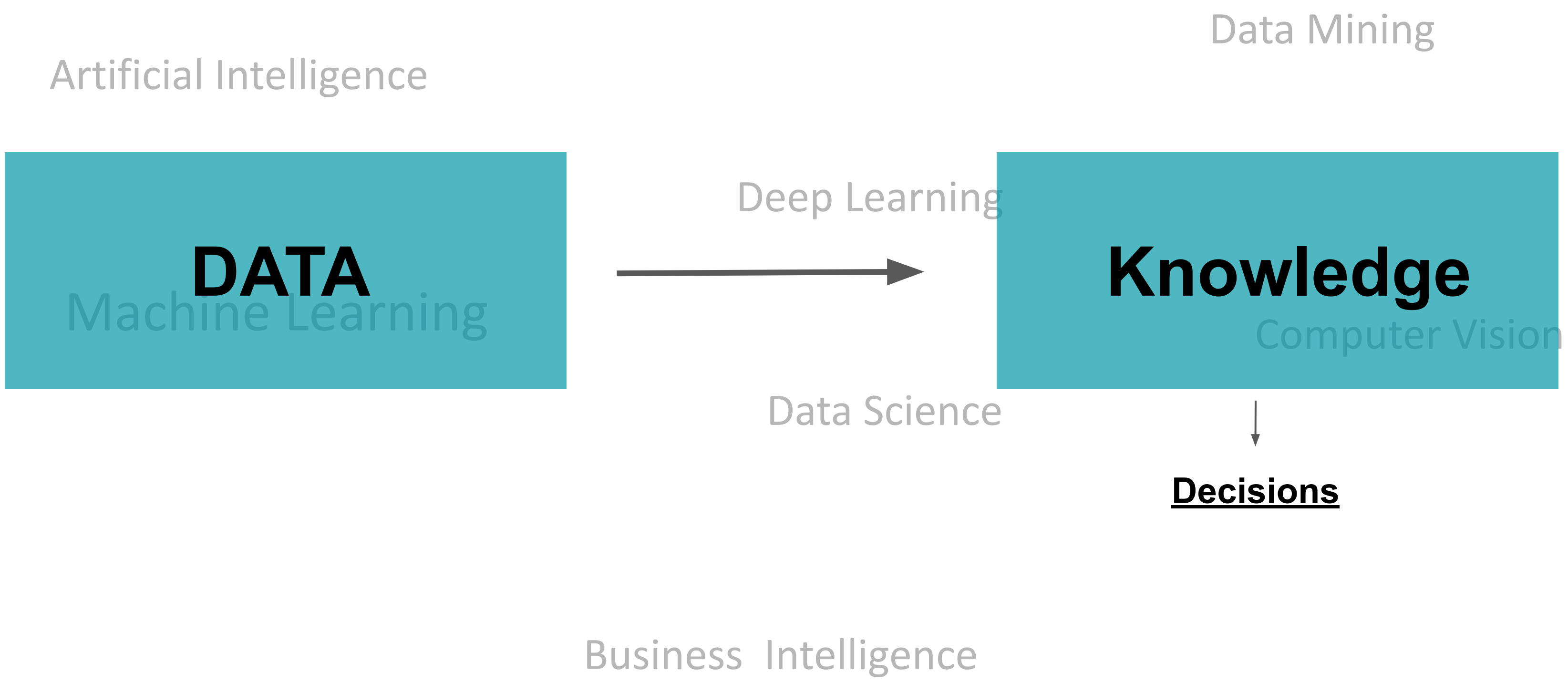
Machine Learning

Computer Vision

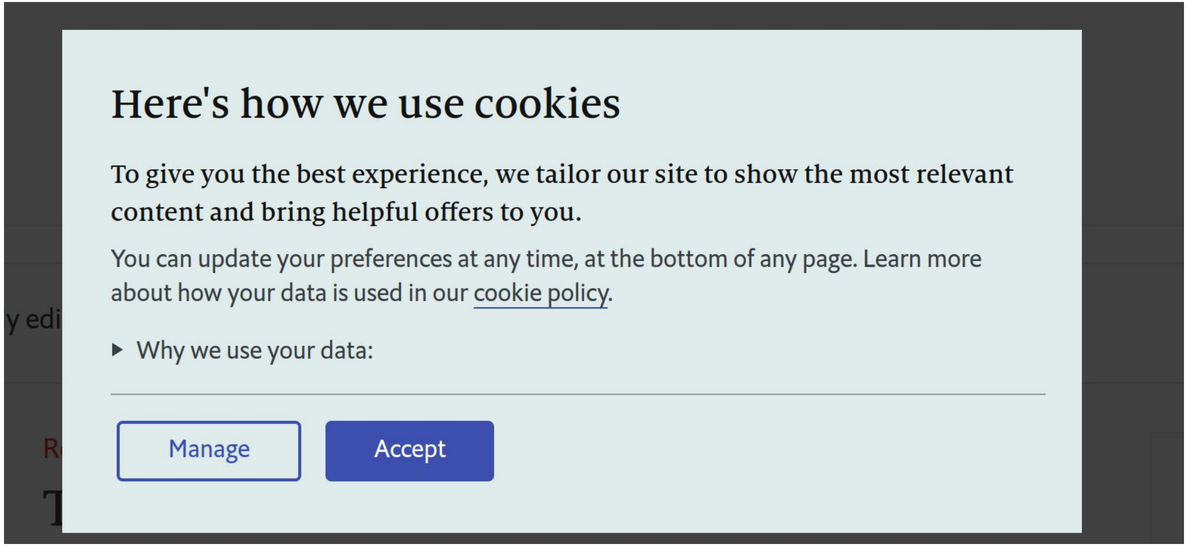
Data Science

Business Intelligence

Glossary

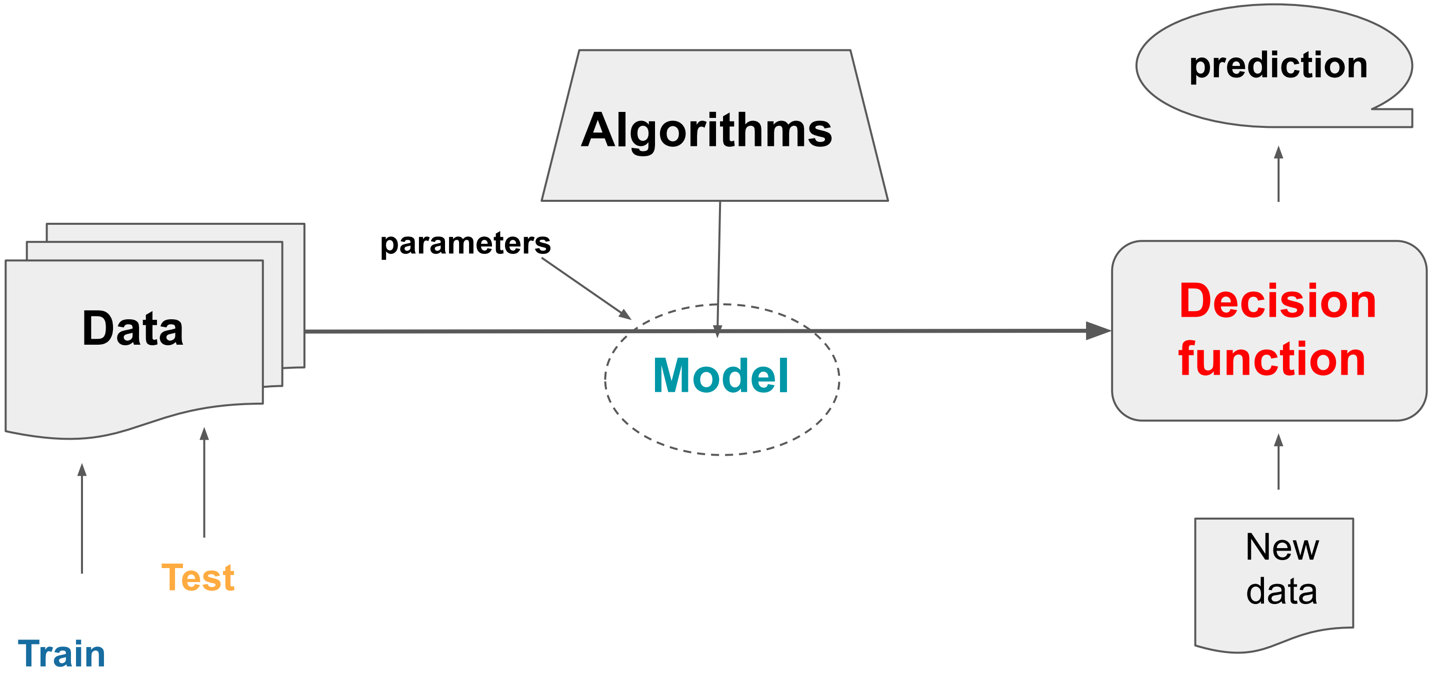

Data Mining
Artificial Intelligence
Deep Learning
DATA
Machine Learning
Knowledge
Computer Vision
Data Science
Decisions
Business Intelligence

Data

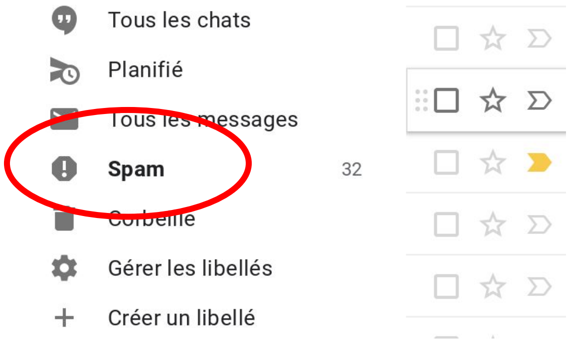


Value (data) > value (oil) or value (gold)

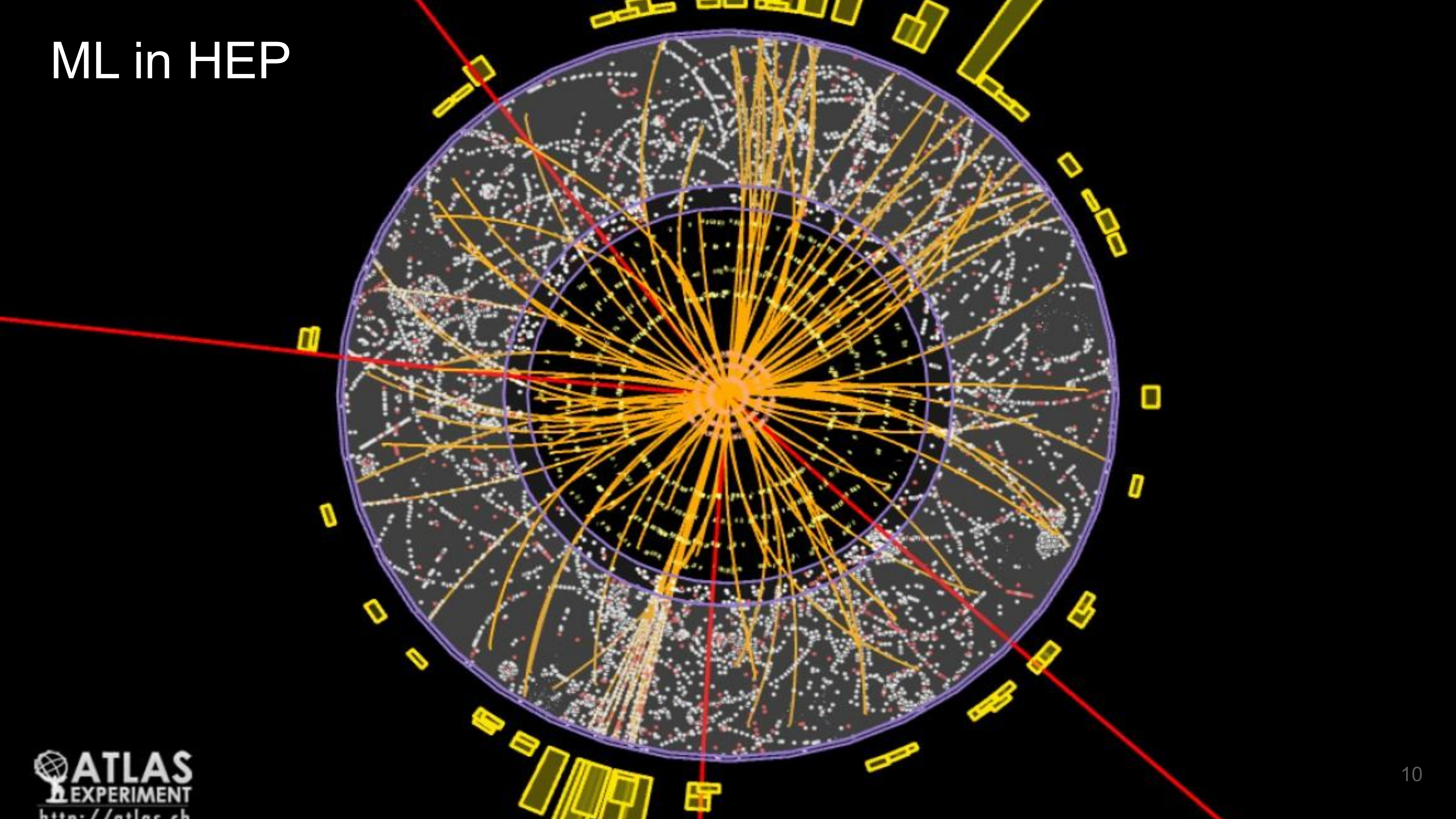
Machine learning in a nutshell



ML in the world: where you already use it



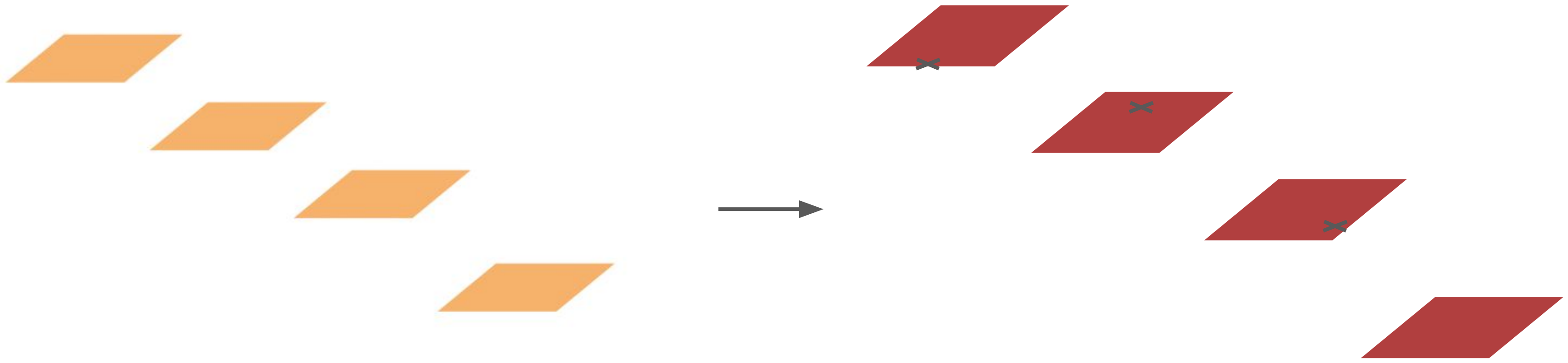
ML in HEP



ML in HEP

- Overview here : arXiv:1807.02876
- Most frequently used ML : Boosted Decision Trees (BDTs) and Neural Networks (NN)
- Estimate of a particle's energy using multiple detectors measurements
- Neural network for merged pixel clusters

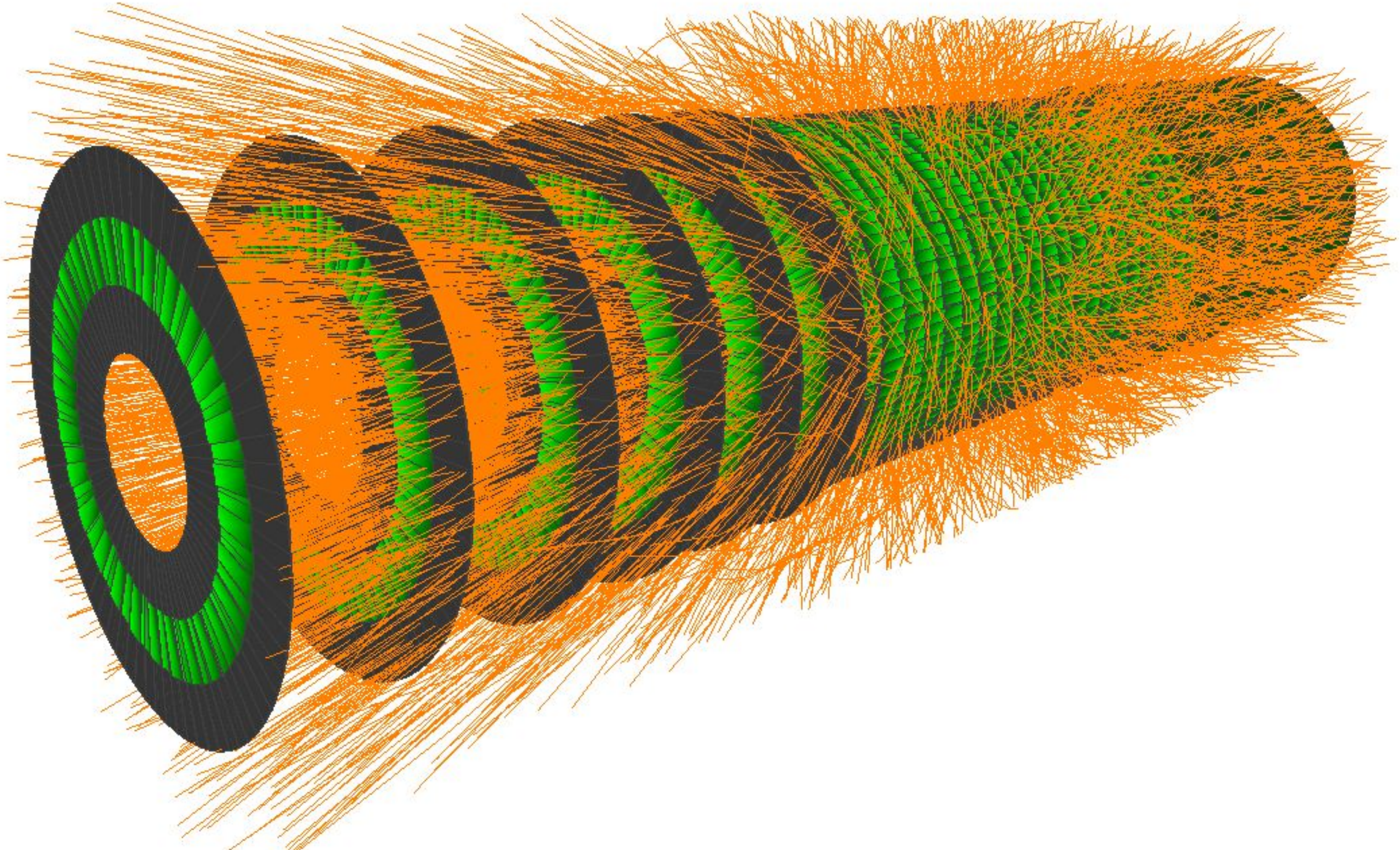
Particle Tracking with Machine Learning



(1) The actual particle

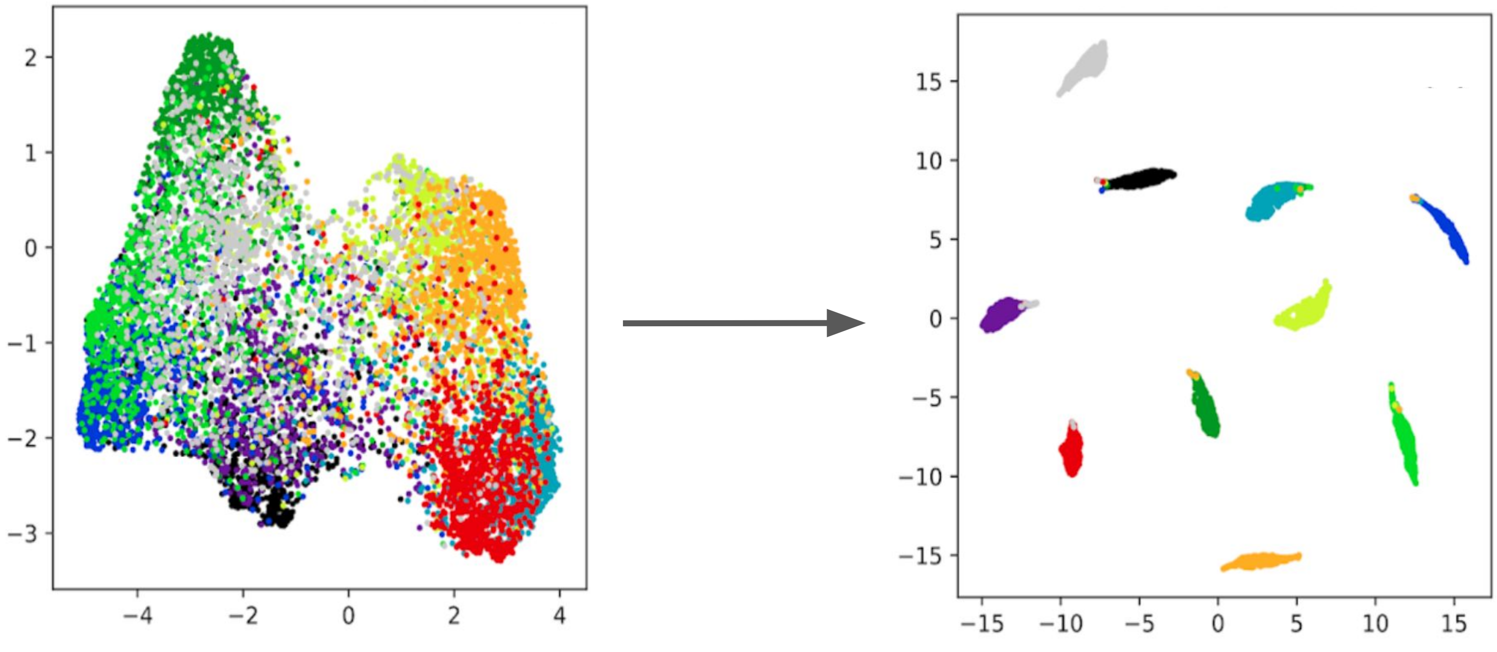
(2) The measurements

Particle Tracking with Machine Learning



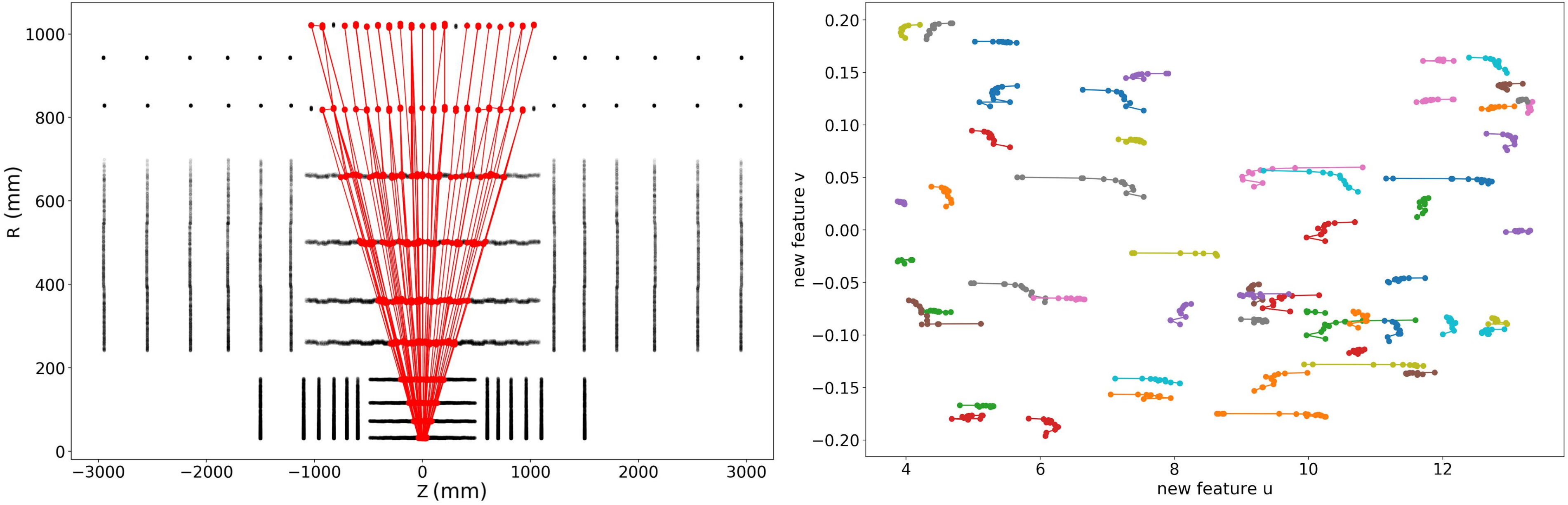
Metric (similarity) Learning

We know truth association from simulation



Metric Learning

$$\mathcal{D}(x_1, x_5) < \mathcal{D}(x_1, x_2)$$



- The ML model : State of the art [UMAP](#) (Uniform Manifold Approximation and Projection)

Generative model for fast simulation

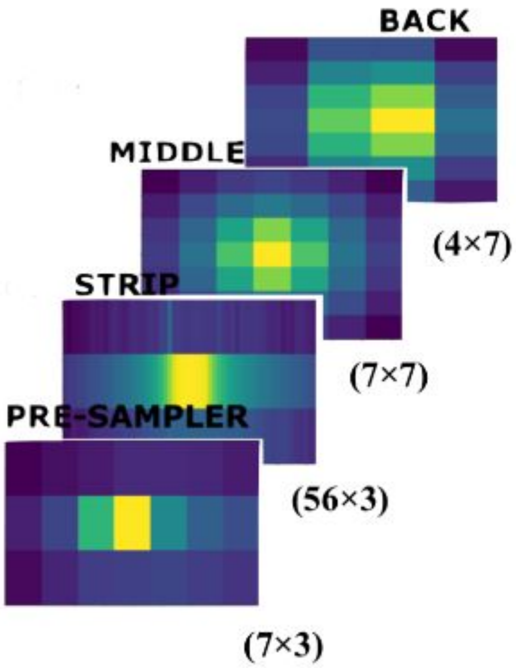
```
def FastCaloMLSim(type,energy,eta):
    return P(shower|type,energy,eta)
```

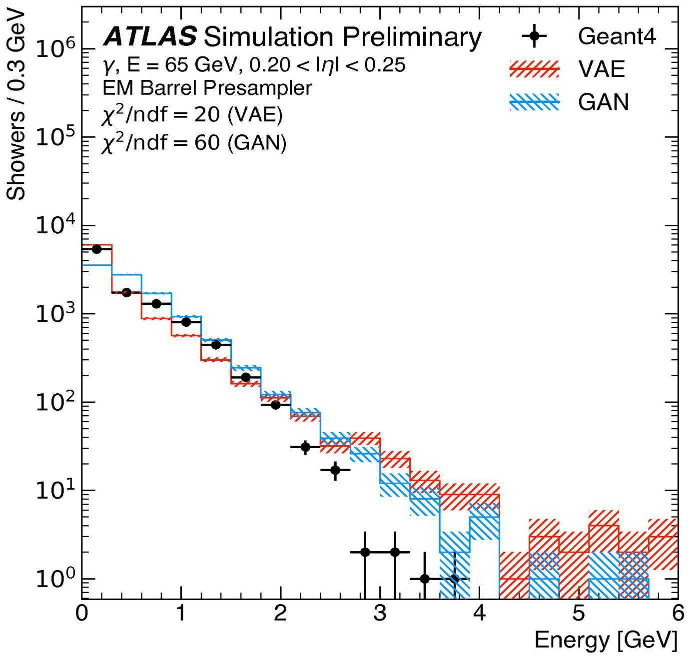
Variational Autoencoders (VAE)

Generative Adversarial Network (GAN)

Generative model for fast simulation

Example of a photon shower development
in the electromagnetic calorimeter





Machine learning in a nutshell

Reinforcement

- Decision by reward
- Select best option
- Learn behaviour

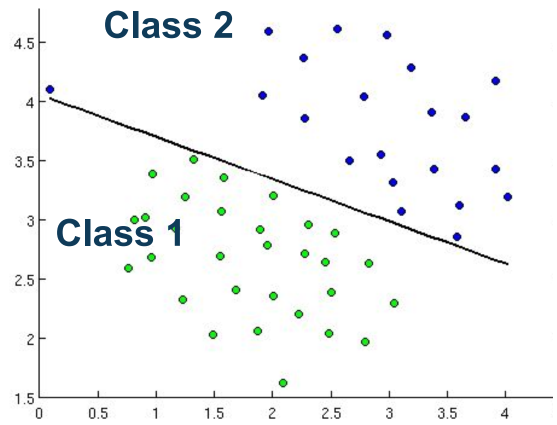
Supervised

- Labeled data
- Classification (discrete values)
- Regression (real values)

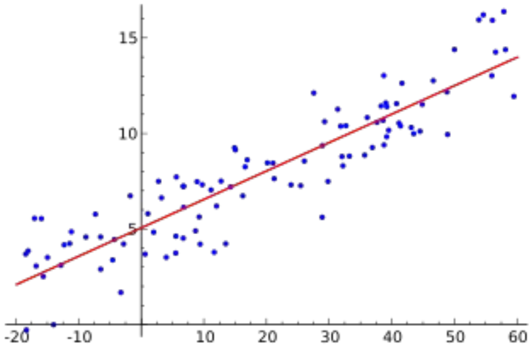
Unsupervised

- No data labels
- Clustering

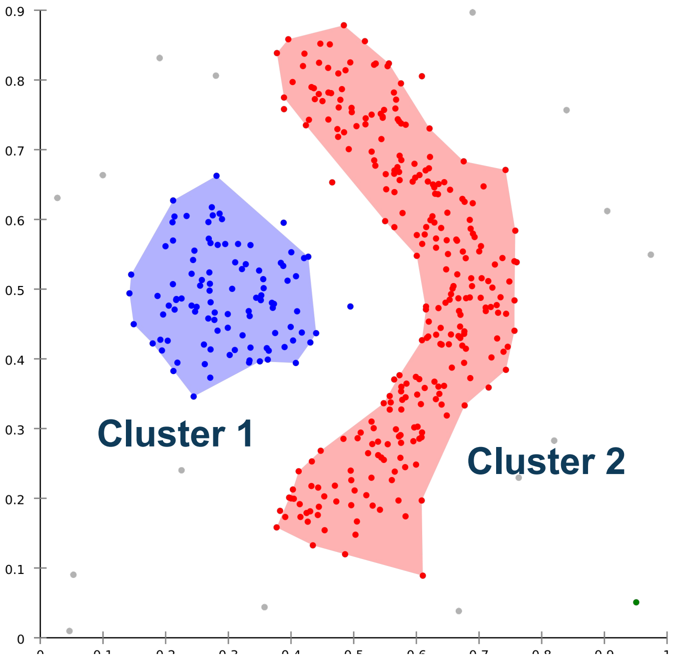
Classification / Regression / Clustering



- Classify future observations
- Known classes



- Predict continuous attribute



- No prior knowledge
- Discover patterns

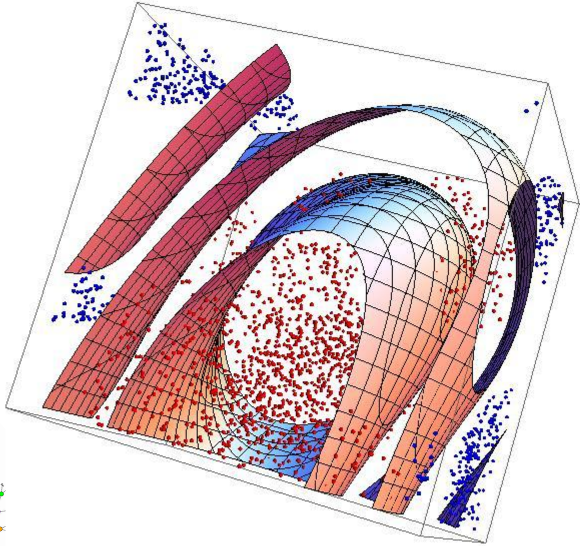
Interactive Session

slido.com

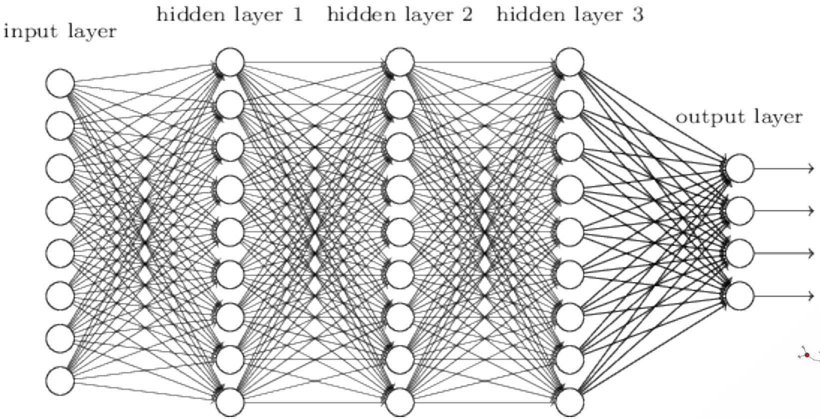
#736251

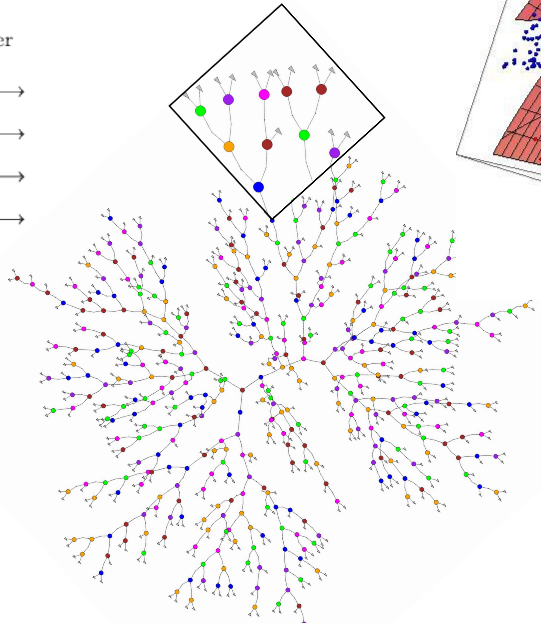
Popular ML models

Support Vector Machines



Neural Networks

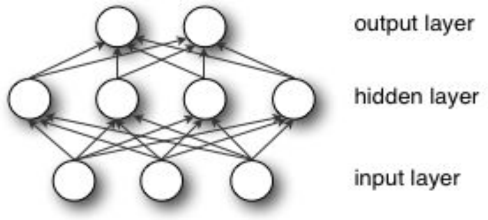


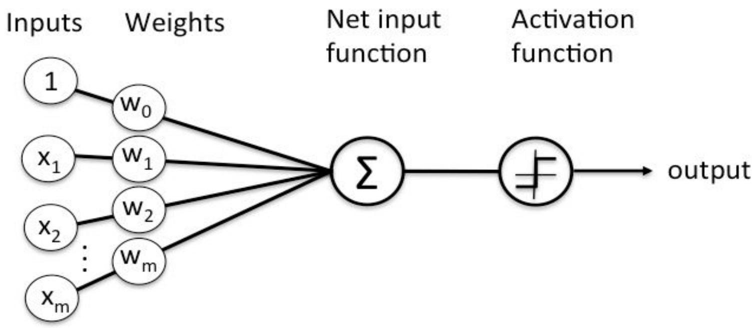


Random Forests

Neural networks

- Inspired by biological neurons
- Designed to recognise patterns
- Key components
  - **Node**: represents an artificial neuron
  - **Weight**: importance of the node in the learning process
  - **Layer**: a set of nodes.
    - Input
    - Hidden: learns different aspects about the data by minimizing an error/cost function
    - Output
  - **Activation function**:
- Learning from sample observations by adjusting the weights to improve the accuracy of the result.
- Learning is done by minimizing the observed errors.





Images from [Ref]

Why Python ?

- General purpose language
- Easy to use
- Popular in data science community
- Integrated packages : data processing, ML, data structure saving/loading

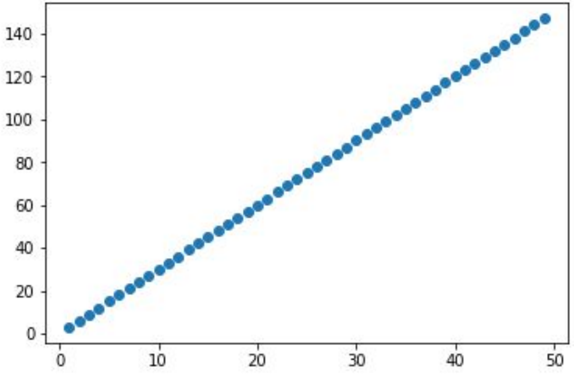
Jupyter notebook

Open-source web application to create live code, equations, visualizations and text [Ref]

In [3]:
```python
import matplotlib.pyplot as plt

X = range(1, 50)
Y = [value * 3 for value in X]

plt.scatter(X,Y)
plt.show()
```



The Lorenz Equations ¶

```latex
\begin{align}
\dot{x} & = \sigma(y-x) \\
\dot{y} & = \rho x - y - xz \\
\dot{z} & = -\beta z + xy
\end{align}
```

$$\begin{aligned}\dot{x} &= \sigma(y-x) \\ \dot{y} &= \rho x - y - xz \\ \dot{z} &= -\beta z + xy\end{aligned}$$

Numpy

Package for scientific computing with Python [Ref]

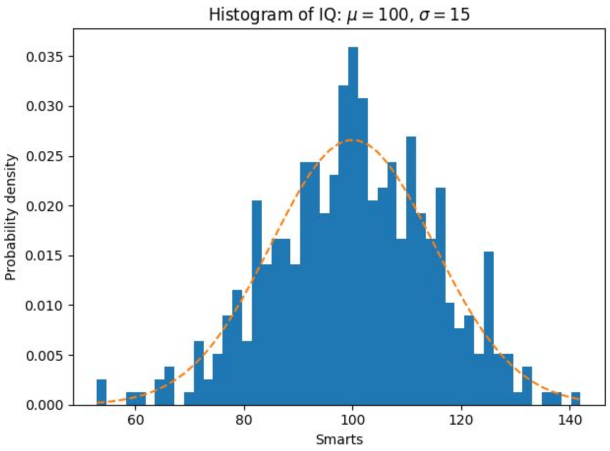
- N-dimensional array manipulation
- sophisticated (broadcasting) functions
- tools for integrating C/C++ and Fortran code
- useful linear algebra, Fourier transform, and random number capabilities ...

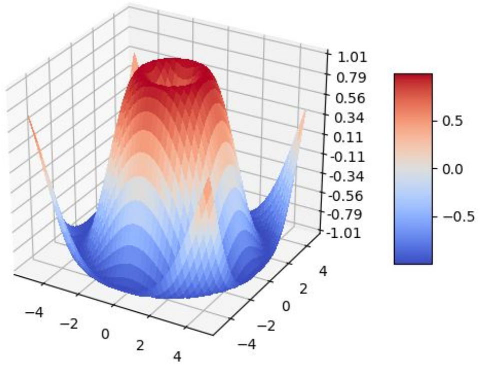
Uproot

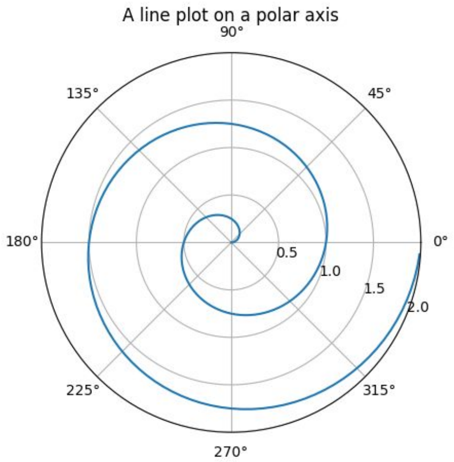
- Reader/Writer of the ROOT file format using **only Python and Numpy** [Ref]
- No dependence on C++ ROOT
- Uses Numpy to cast blocks of data from the ROOT file as Numpy arrays.
- Designed to stream data into machine learning libraries in Python

Matplotlib

Plotting library [Ref]







File structures: CSV, HDF5

- CSV
    - A CSV is a comma-separated values file
    - It is a plain-text file

```
ID, Energy, Angle, Mass
1, 10, 0, 0.1
2, 20, 10, 0.2
3, 30, 20, 0.3
4, 40, 30, 0.4
5, 50, 90, 0.5
6, 60, 25, 0.1
7, 25, 35, 0.5
8, 45, 90, 1
9, 100, 10, 2
10, 500, 50, 4
```

- HDF5
    - Hierarchical Data Format (HDF)
    - It is a binary file format
    - Supports large, complex, heterogeneous data

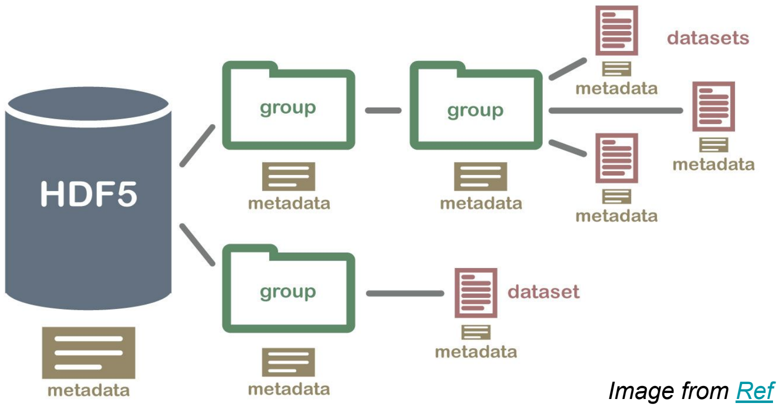


Image from Ref

Interactive Session

Part I : Hands-On session Python libraries

Github
[Colab]