MACHINE LEARNING IN HIGH ENERGY PHYSICS

PRACTICAL CLASS #1

Alex Rogozhnikov, 2015

COMMON RULES

- Theoretical tasks (deadline — next lecture)
- New practical task after each seminar (submission of tasks via email)
- Materials are published

WHAT IS ML ABOUT?

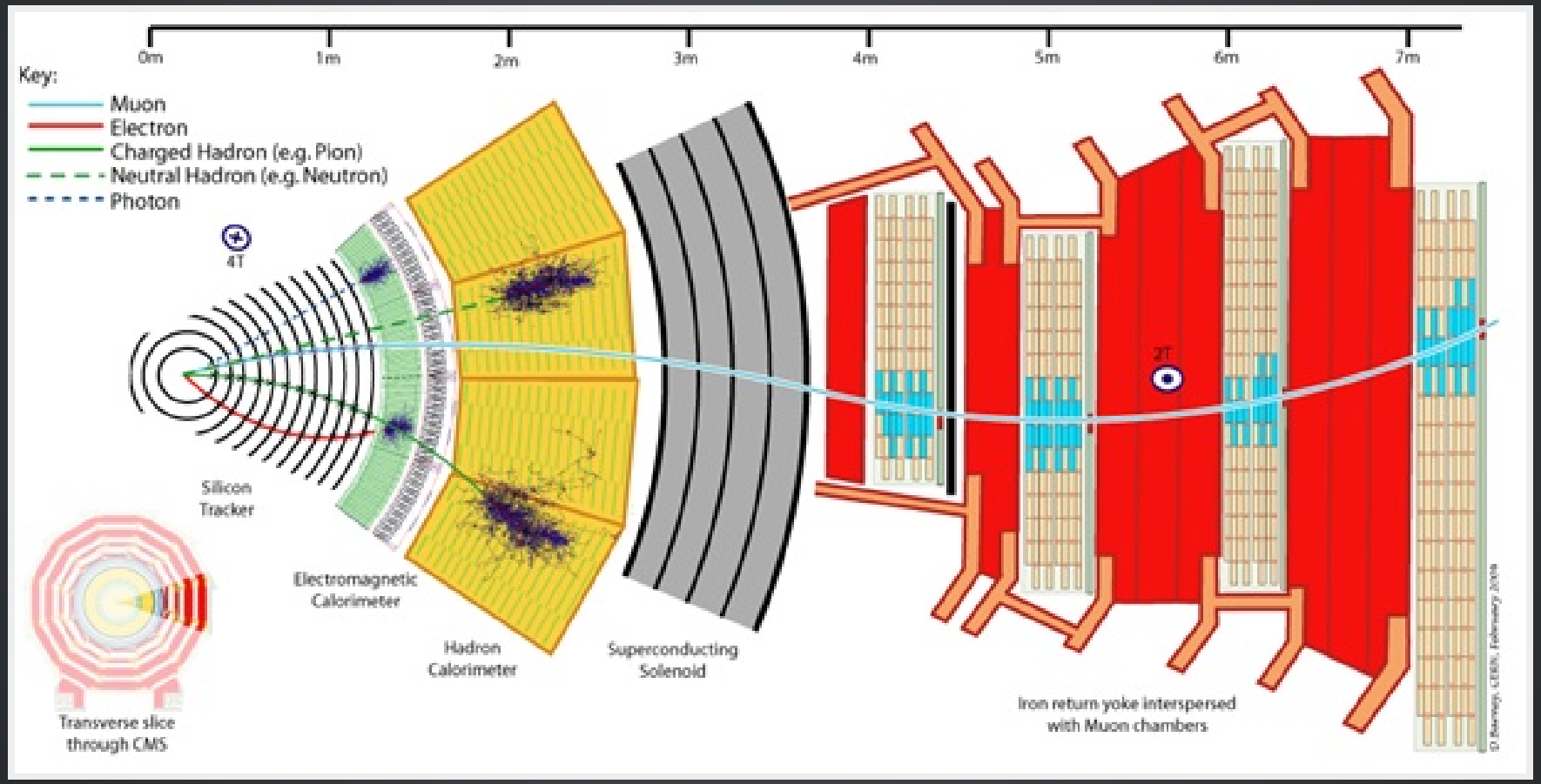
Inference of statistical dependencies which give us ability to predict

Data is cheap, knowledge is precious

WHERE ML IS CURRENTLY USED?

- Search engines, spam detection
- Security: virus detection, DDOS defense
- Computer vision and speech recognition
- Market basket analysis, Customer relationship management (CRM)
- Credit scoring, fraud detection
- Health monitoring
- NASA: star identification
- ... and hundreds more

MACHINE LEARNING IN HIGH ENERGY PHYSICS



ML IN HIGH ENERGY PHYSICS

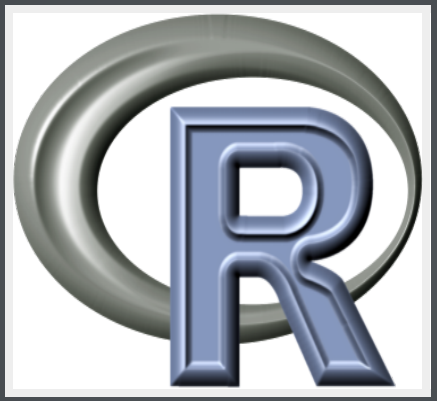
- High-level triggers (ATLAS trigger system: $10^8 \rightarrow 400$ events)
- Particle identification
- Calibration
- Tagging (W-tagging, b-tagging)
- Stripping line
- Analysis

Data used:

- kinematic variables (masses, momenta, decay angles, …)
- event properties (jet/lepton multiplicity, sum of charges, …)
- event shape (sphericity, Fox-Wolfram moments, …)
- detector response (silicon hits, dE/dx, Cherenkov angle, shower profiles, muon hits, …)

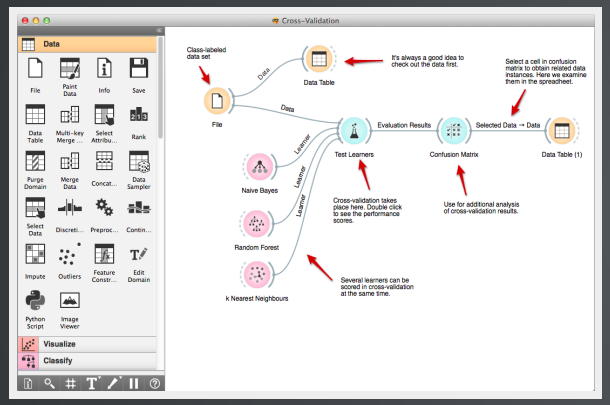
TOOLS FOR DATA ANALYSIS

Languages



TOOLS FOR DATA ANALYSIS

ML Libraries with GUI and dataflow



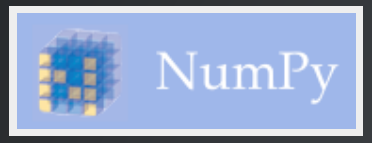
- Orange
- RapidMiner

POPULAR ML LIBRARIES:

- Weka
- Scikit-learn
- Torch
- Vowpal Wabbit
- LibSVM
- H2O
- XGBoost
- MultiBoost
- and many-many others

IPYTHON DEMONSTRATION

LIBRARIES FOR PYTHON

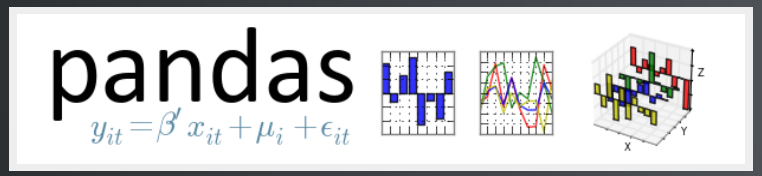


NumPy

vectorized computations in python

Matplotlib

for drawing



Pandas

for data manipulation and analysis (based on numpy)

LIBRARIES FOR PYTHON

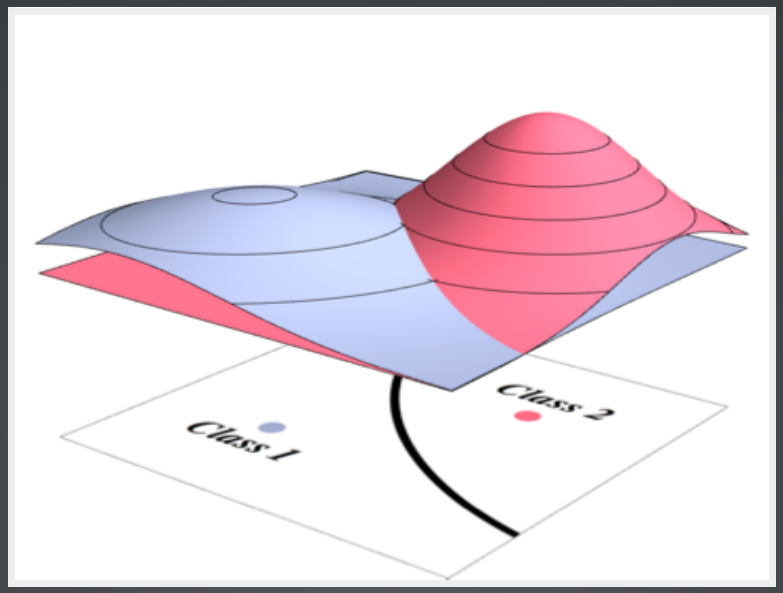
Scikit-learn

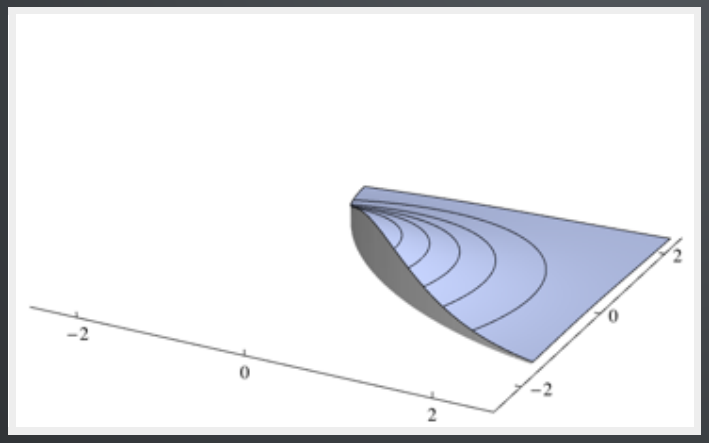
most popular library for machine learning

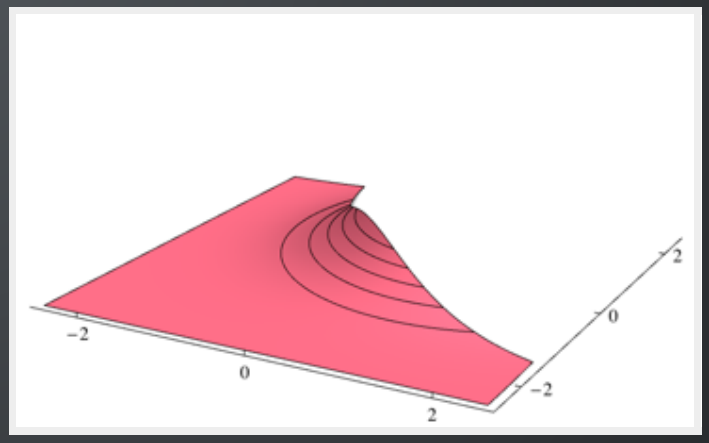
Scipy

libraries for science and engineering

OPTIMAL BAYES CLASSIFIER



Error = 

+ 

EXAMPLE: BINARY CLASSIFICATION PROBLEM

$$p(\omega_1|x) = 0.4, \;\; p(\omega_2|x) = 0.6$$

Now let's add following information:
ω_1 - regular letter, ω_2 - spam letter.

WHAT IS THE PROBLEM WITH USING PROBABILITIES?

- Hardly can be reconstructed
- Specially in high-dimensional spaces
- So, we switch to discriminative approach


THE END