

Big data science

Day 1 - Hands on

F. Legger - INFN Torino

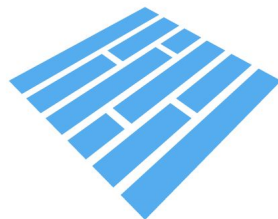
<https://github.com/Course-bigDataAndML/MLCourse-2223>

What we will use

- **Python** with Jupyter notebooks
- **Day 1:** familiarise with **Kubernetes, spark, parquet** files
- **Day 2:** Gradient Boosting Trees
GBT MLlib
- **Day 3: Neural networks**
 - Multilayer Perceptron Classifier
MCP MLlib
 - **Keras** Sequential model
- **Day 4: bigDL** Sequential model



kubernetes



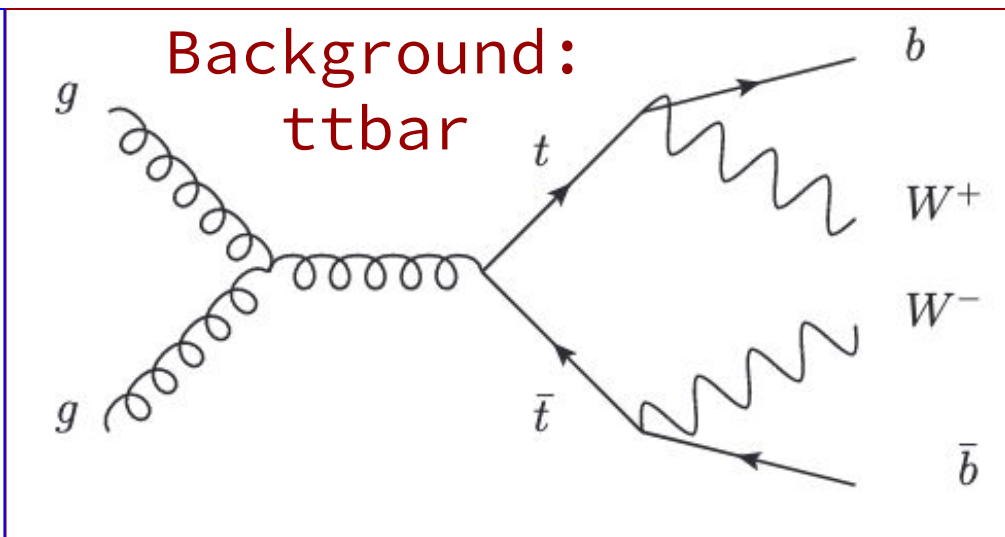
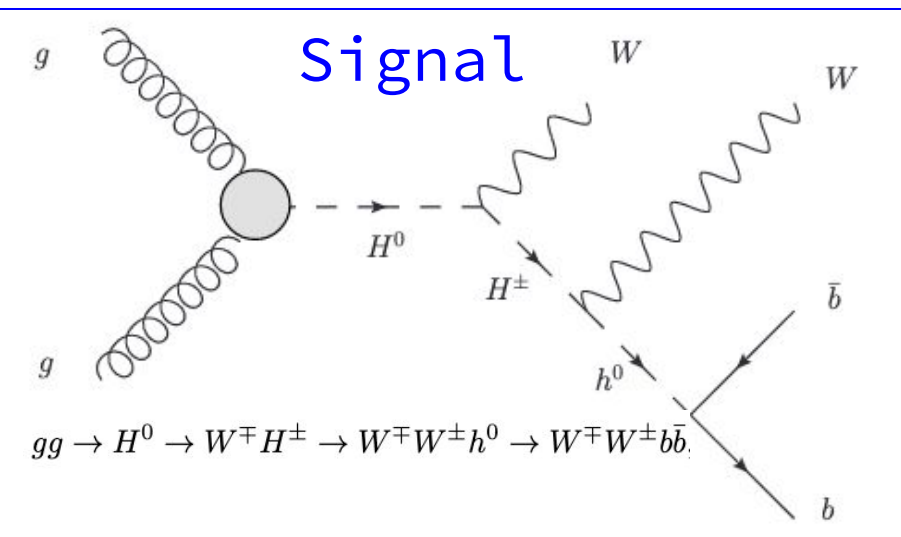
Parquet



Input dataset for hands-on

<https://archive.ics.uci.edu/ml/datasets/HIGGS>

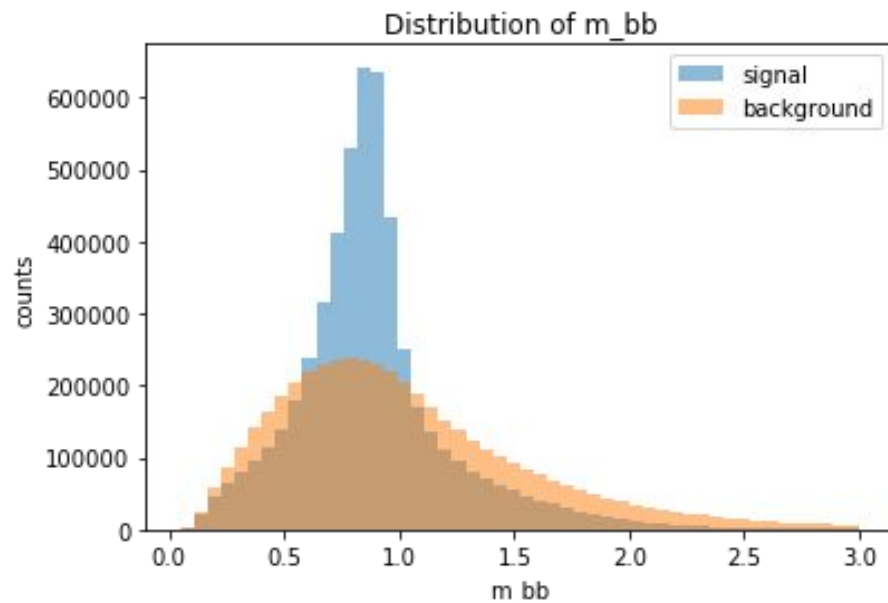
- Open HEP dataset @UCI
- Signal (heavy Higgs) + background (ttbar)



Baldi, Sadowski, and Whiteson. "Searching for Exotic Particles in High-energy Physics with Deep Learning." *Nature Communications* 5

Input dataset for hands-on

- 10M Monte Carlo events
(7GB .csv)
 - 21 low level features
 - pt's, angles, MET, b-tag, ...
 - 7 high level features
 - Invariant masses ($m(jj)$, $m(jjj)$, ...)
- Smaller datasets for code testing (1M, 100k)



↑
Exercise 3

Hands-on today

- You will familiarize with *jupyter notebooks*, *numpy*, *pandas*
- Input data:
 - efficient format: convert **CSV to Parquet**
 - A comma-separated values (CSV) *file* is a delimited text *file* that uses a comma to separate values
 - And [Apache parquet](#)?
 - Create input for ML. Format depends on chosen ML library, in our case MLlib from Apache
- Visualization
 - *explore dataset, plot features, correlation matrix*
- ***Slides and notebooks available on github***

<https://github.com/Course-bigDataAndML/MLCourse-2223>

How to start

1. **Point your browser to:** <https://yoga.to.infn.it>
2. **Authenticate** through github
3. **Open a terminal:**
 - git clone
<https://github.com/Course-bigDataAndML/MLCourse-2223.git>
 - cp MLCourse-2223/Notebooks/Day1/* .
4. **From JupyterHub Home tab:**
 - start and run *inputForML.ipynb*
 - *You will receive the solutions tomorrow*

The screenshot shows the JupyterLab application window. The top menu bar includes File, Edit, View, Run, Kernel, Spark, Tabs, Settings, and Help. The File menu is open, displaying options such as New, New Launcher, Open from Path..., New View for, New Console for Activity, Close Tab, Close and Shutdown, Close All Tabs, Save, Save As..., Save All, Reload from Disk, Revert to Checkpoint, Rename..., Download, Save and Export Notebook As..., Save Current Workspace As..., Save Current Workspace, Print..., Hub Control Panel, and Log Out. A red circle highlights the 'Hub Control Panel' option. The main workspace area contains a grid of icons for opening new files or notebooks. The 'Python 3 (ipykernel)' icon is circled in red. Other icons include PySpark, R, spylon-kernel, Console, Other, Terminal, Text File, Markdown File, Python File, R File, and Show Contextual Help. Red text annotations are present: 'Open Notebook' is written in red above the 'Console' icon; 'Open terminal' is written in red above the 'Terminal' icon; and 'Start/stop jupyterHub' is written in red below the 'Hub Control Panel' option in the File menu.

File Edit View Run Kernel Spark Tabs Settings Help

New

New Launcher $\uparrow \mathbb{M} L$

Open from Path...

New View for

New Console for Activity

Close Tab $\mathbb{C} W$

Close and Shutdown $\wedge \uparrow \mathbb{Q}$

Close All Tabs

Save $\mathbb{M} S$

Save As... $\uparrow \mathbb{M} S$

Save All

Reload from Disk

Revert to Checkpoint

Rename...

Download

Save and Export Notebook As...

Save Current Workspace As...

Save Current Workspace

Print... $\mathbb{M} P$

Hub Control Panel

Log Out

Console

Notebook

Terminal

Text File

Markdown File

Python File

R File

Python 3 (ipykernel)

PySpark

R

spylon-kernel

Open Notebook

Console

Python 3 (ipykernel)

PySpark

R

spylon-kernel

Other

Terminal

Text File

Markdown File

Python File

R File

Show Contextual Help

Open terminal

Start/stop jupyterHub

Correlation matrix

Exercise 4

