

# Big data science Hands-on

F. Legger - INFN Torino

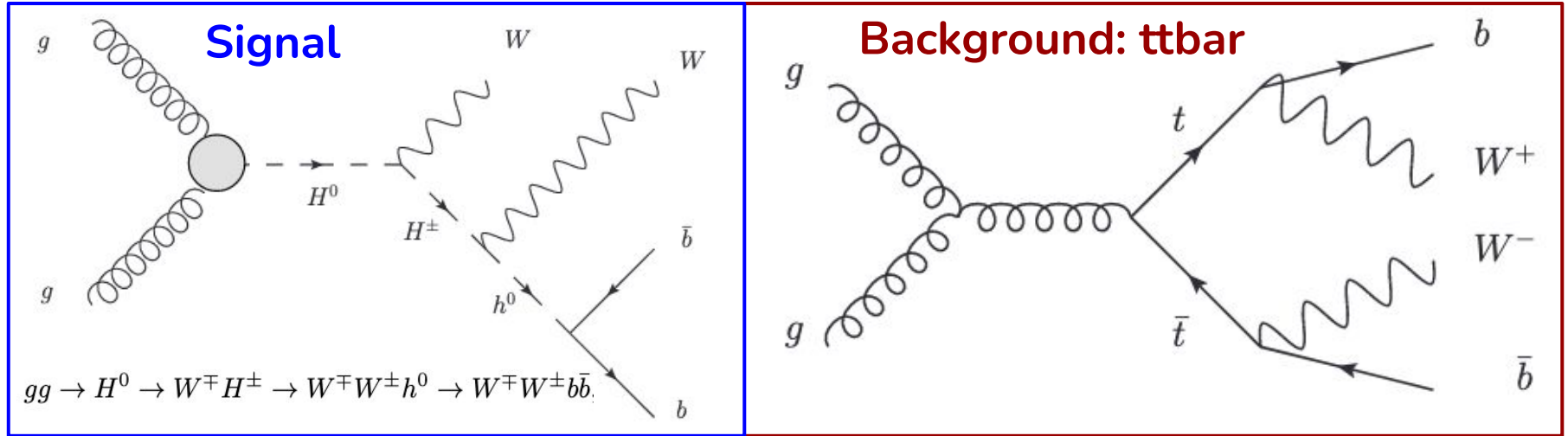
<https://github.com/Course-bigDataAndML/MLCourse-2425>

- So long, and thanks for all the images!
  - Taken freely from the web
  - Credits go to original creators



# Classification problem

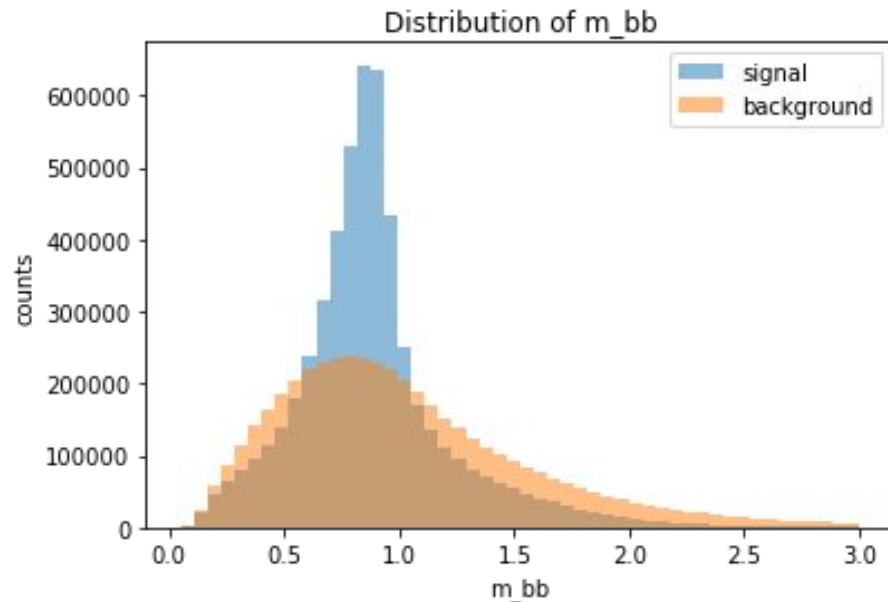
- Open HEP dataset @UCI
- Signal (heavy Higgs) + background (ttbar)



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 (.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 (100k, 1M)



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

Exercise

# Hands-on

- You will familiarize with *jupyter notebooks*, *numpy*, *pandas*, *spark*, *kubernetes*
- 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](#)?
- Distributed data analysis with Spark on top of Kubernetes
- Visualization
  - *explore dataset, plot features, correlation matrix*

# What we will use



- *Slides and notebooks available on github*

<https://github.com/Course-bigDataAndML/MLCourse-2425>

# 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-2425.git>
  - cp MLCourse-2425/Notebooks/Day1/\* .
4. **From JupyterHub Home tab:**
  - start and run *inputForML\_exercises.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 sub-menu for 'New' is also visible, listing Console, Notebook, Terminal, Text File, Markdown File, Python File, and R File. The main workspace area contains a grid of icons for different environments: Python 3 (ipykernel), PySpark, R, and spylon-kernel. Below this, there is a section for 'Other' with icons for Console, Terminal, Text File, Markdown File, Python File, R File, and Show Contextual Help. Red circles highlight the 'Python 3 (ipykernel)' icon in the top row and the 'Terminal' icon in the 'Other' section. Red text annotations are present: 'Open Notebook' in red, 'Open terminal' in red, and 'Start/stop jupyterHub' in red. The 'Hub Control Panel' and 'Log Out' options in the File menu are also circled in red.

File Edit View Run Kernel Spark Tabs Settings Help

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

Log Out

Console

Notebook

Terminal

Text File

Markdown File

Python File

R File

Python 3 (ipykernel)

PySpark

R

spylon-kernel

Console

Open Notebook

Python 3 (ipykernel)

PySpark

R

spylon-kernel

Other

Open terminal

Terminal

Text File

Markdown File

Python File

R File

Show Contextual Help

# Correlation matrix

