

Big data science Hands-on

F. Legger - INFN Torino

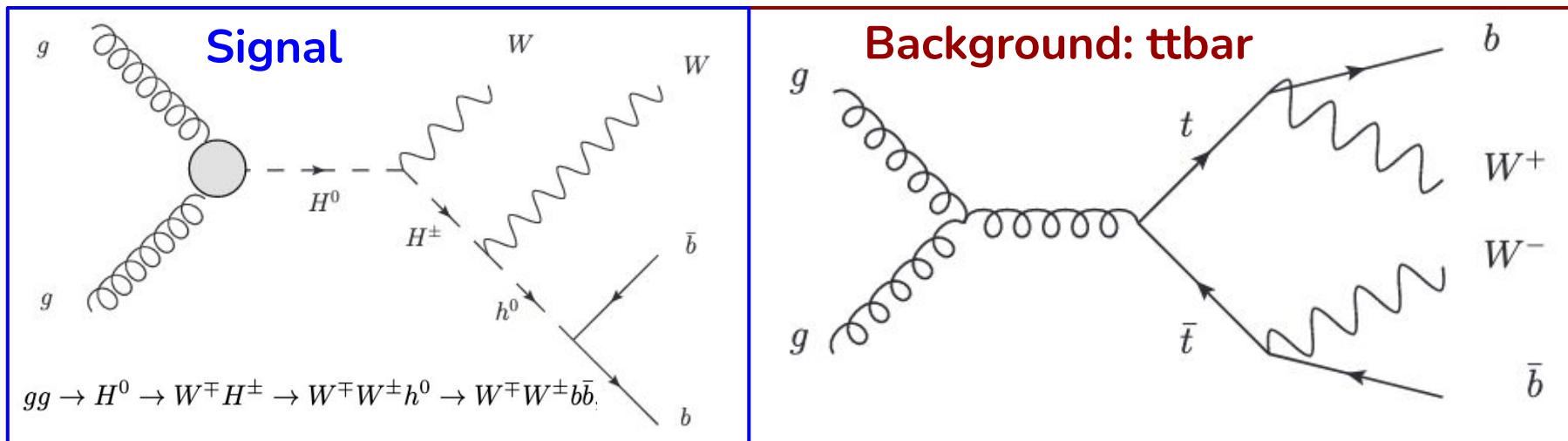
<https://github.com/Course-bigDataAndML/MLCourse-2425>

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



Classification problem

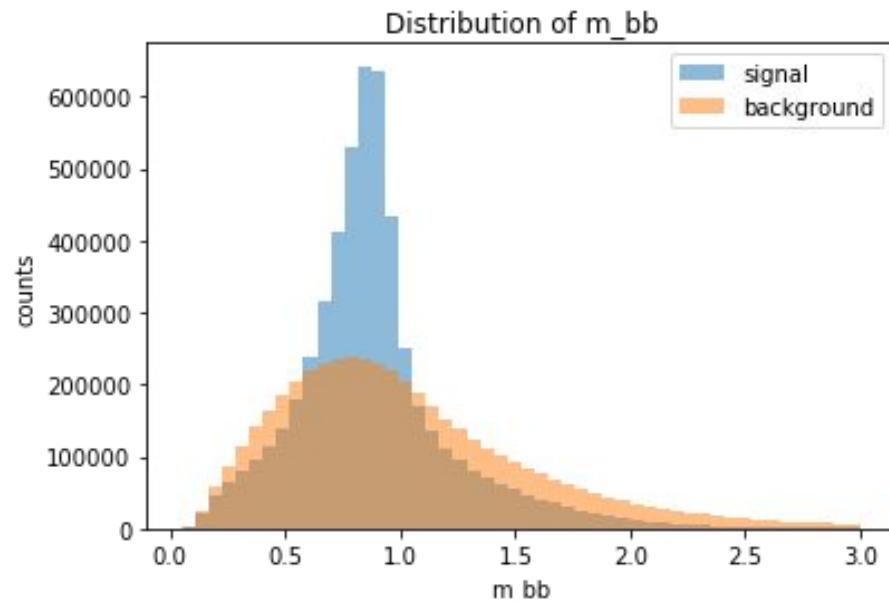
- Open HEP dataset @UCI
- Signal (heavy Higgs) + background (ttbar)



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)



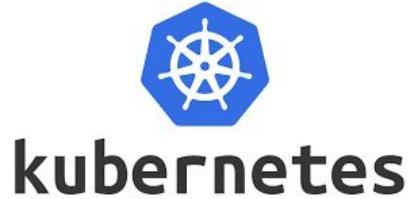
Exercise

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

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

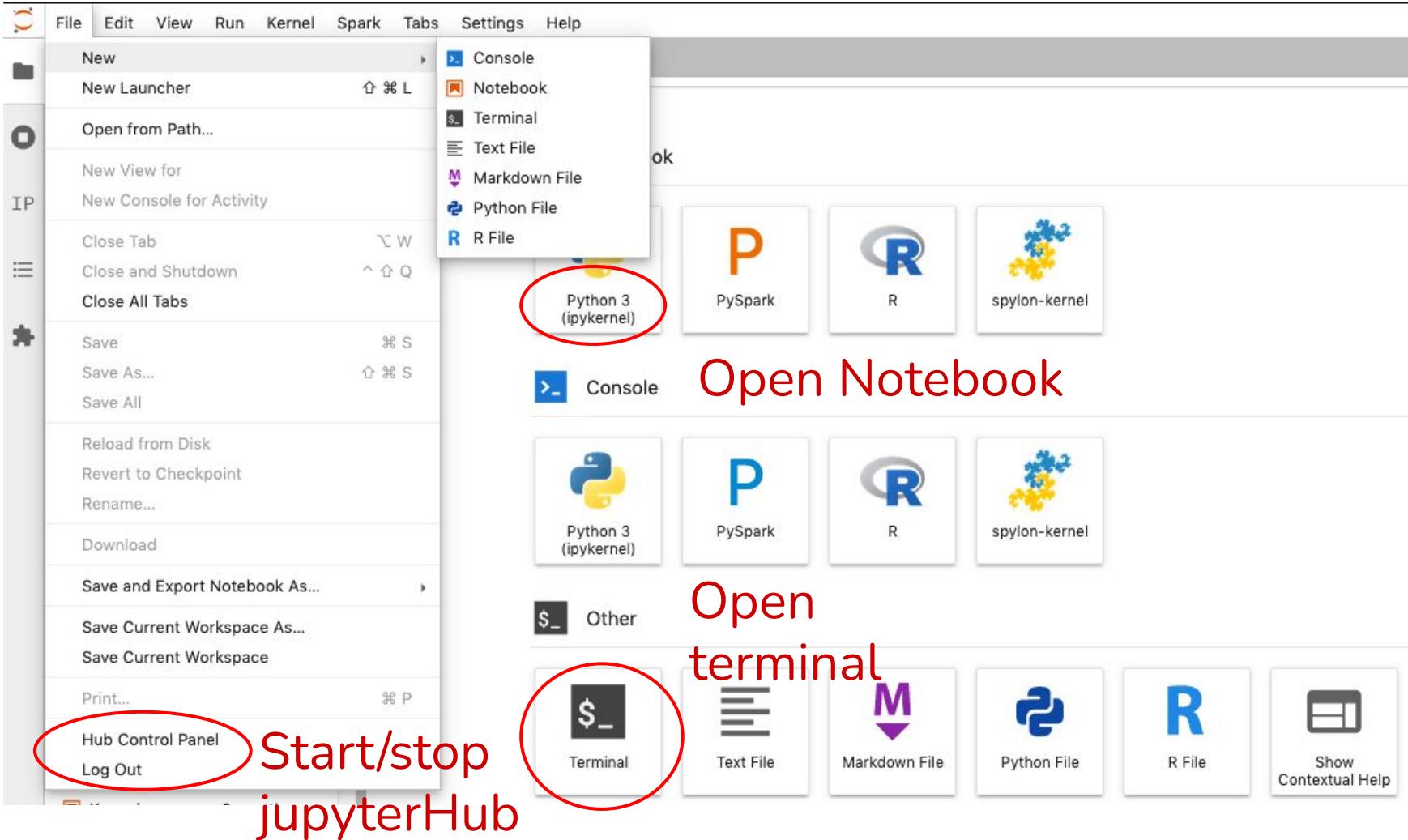


- *Slides and notebooks available on github*

<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:
 - o git clone
<https://github.com/Course-bigDataAndML/MLCourse-2425.git>
 - o cp MLCourse-2425/Notebooks/Day1/* .
4. From JupyterHub Home tab:
 - o start and run *inputForML_exercises.ipynb*
 - o You will receive the solutions tomorrow



Correlation matrix

