Big data science
Day 1 - Hands on

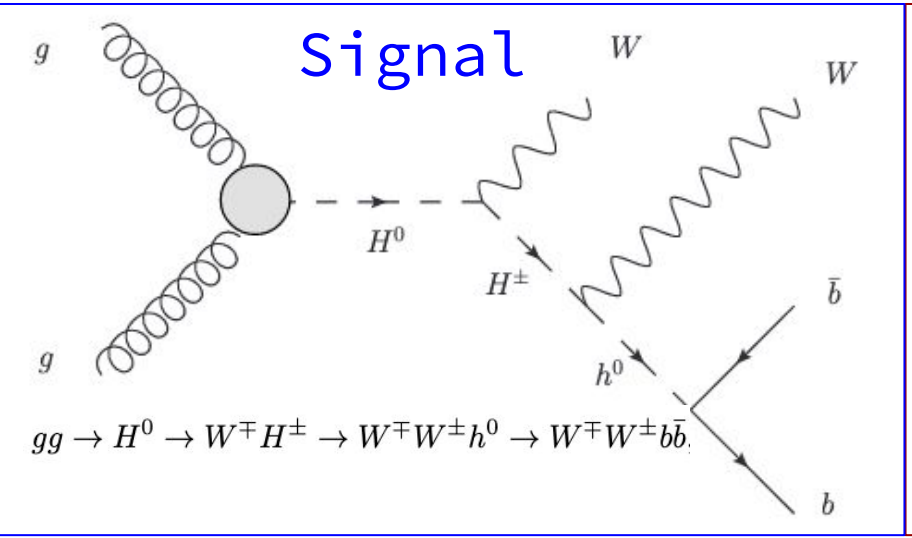
F. Legger - INFN Torino

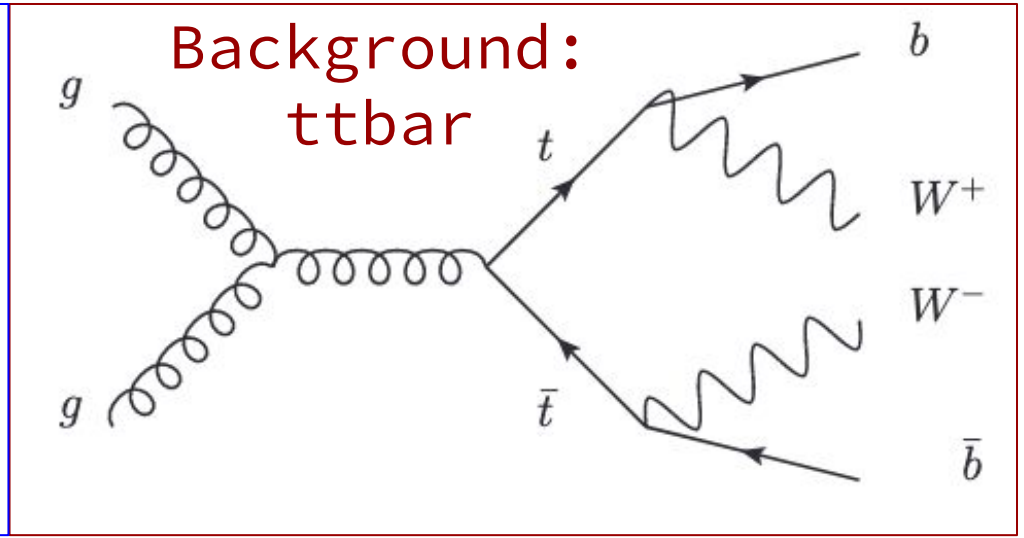
https://github.com/Course-bigDataAndML/MLCourse-2324

Input dataset for hands-on

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

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

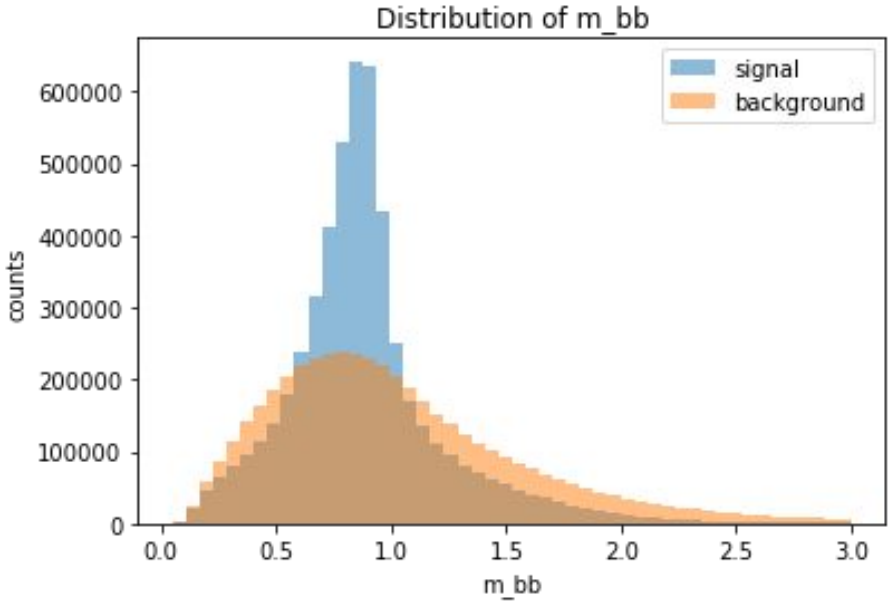




Baldi, Sadowski, and Whiteson. "Searching for Exotic Particles in High-energy Physics with Deep Learning." Nature Communications 5

Input dataset for hands-on

- 10M Monte Carlo events (.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

Hands-on today

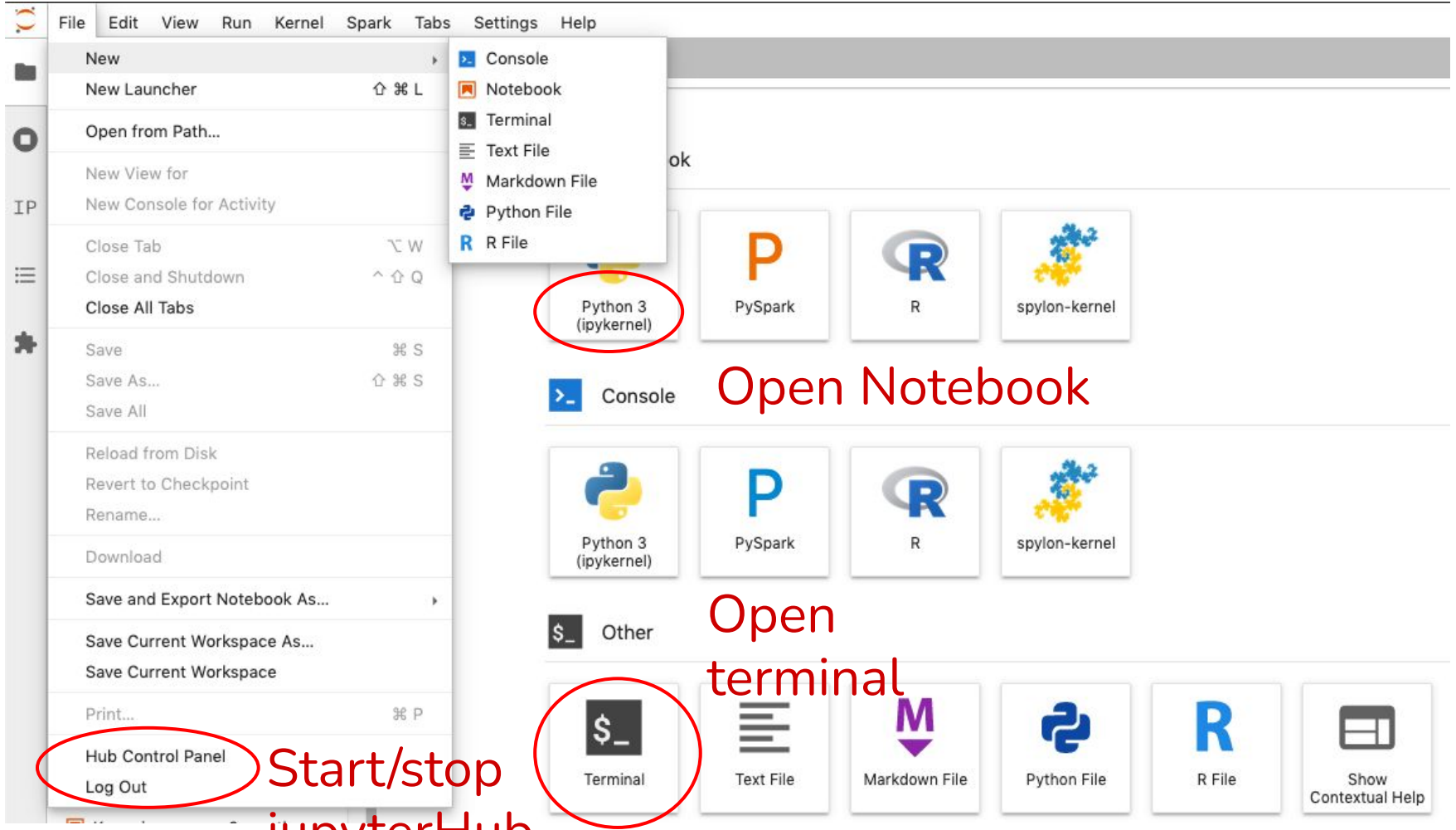
- You will familiarize with *jupyter notebooks*, *numpy*, *pandas*, *spark*, *kubernetes*
- Input data:
  - efficient format: convert **CSV to Parquet**
    - A comma-separated values (*CSV*) *file* is a delimited text *file* that uses a comma to separate values
    - And Apache parquet?
- Distributed 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-2324

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-2324.git
   - cp MLCourse-2324/Notebooks/Day1/* .
4. **From JupyterHub Home tab:**
   - start and run *inputForML_exercises.ipynb*
   - *You will receive the solutions tomorrow*


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
Python 3 (ipykernel)
PySpark
R
spylon-kernel
Other
Terminal
Text File
Markdown File
Python File
R File
Show Contextual Help
Open Notebook
Open terminal
Start/stop jupyterHub

Correlation matrix