

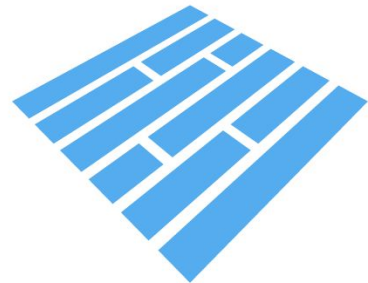
# Big data science Day 1 - Hands on

F. Legger - INFN Torino



# What we will use

- **Python** with Jupyter notebooks
- Prerequisites: some familiarity with numpy and pandas
  - **Day 1:** familiarise with ML dataset, **parquet** files
- ML libraries
  - **Day 2: MLlib**
    - Gradient Boosting Trees **GBT**
    - Multilayer Perceptron Classifier **MCP**
  - **Day 3: Keras**
    - Sequential model
  - **Day 4: bigDL**
    - Sequential model

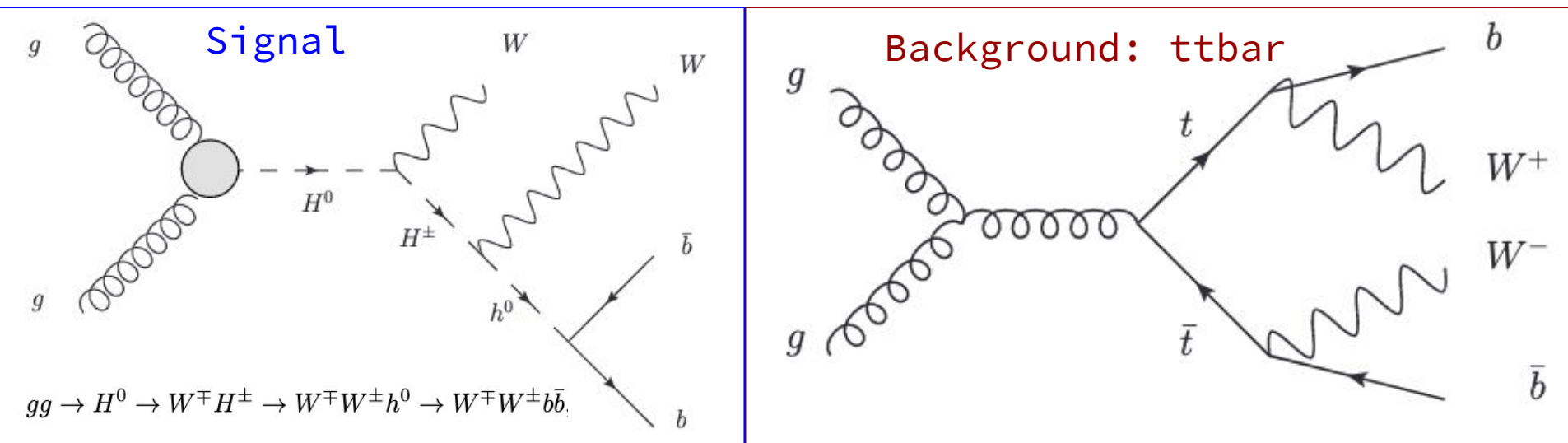


Parquet



# Input dataset

- Open HEP dataset @UCI <https://archive.ics.uci.edu/ml/datasets/HIGGS>
- 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

- Open HEP dataset @UCI, 7GB (.csv)
- 10M Monte Carlo events
  - 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)
- You'll find them on HDFS

# Hands-on today

- You will familiarize with *jupyter notebooks*, *numpy* and *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*