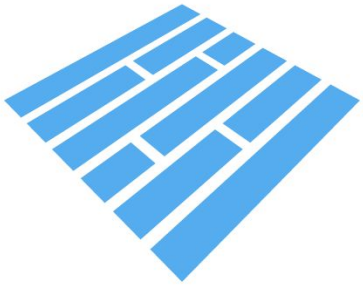
Big data science
Day 1 - Hands on

F. Legger - INFN Torino

https://github.com/leggerf/MLCourse-INFN-2021

What we will use

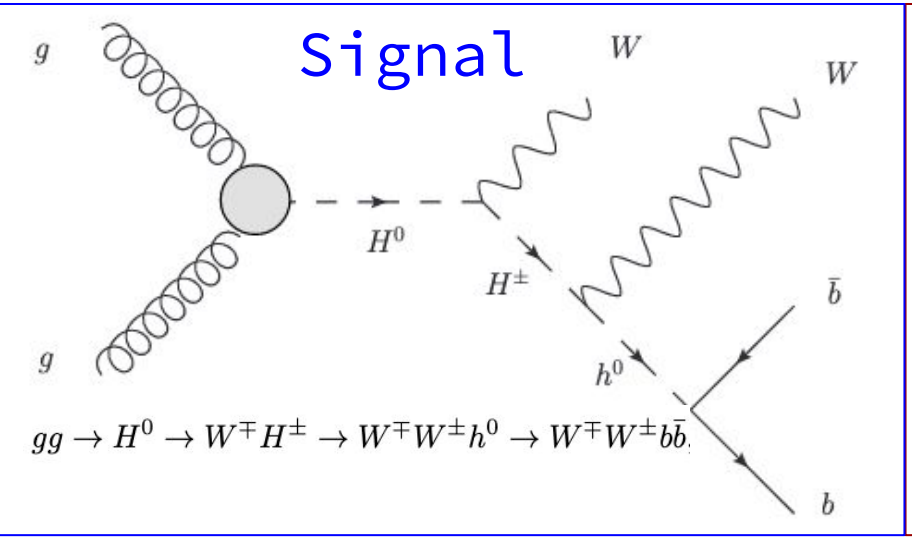


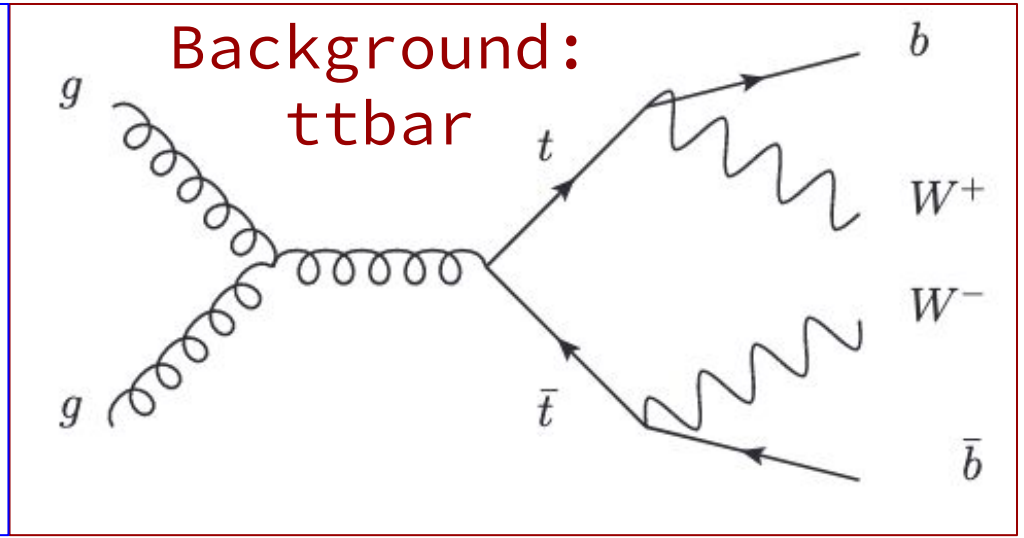
- **Python** with Jupyter notebooks
- **Day 1:** familiarise with ML dataset, **parquet** files
- **Day 2:** Gradient Boosting Trees **GBT MLlib**
- **Day 3: Neural networks**
  - Multilayer Perceptron Classifier **MCP MLlib**
  - **Keras** Sequential model

Input dataset for hands-on

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

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

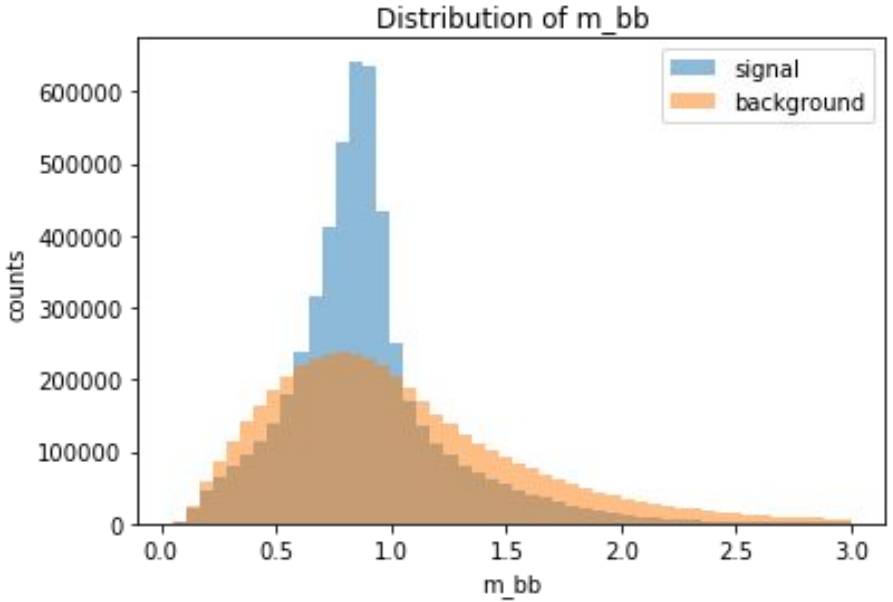




Baldi, Sadowski, and Whiteson. "Searching for Exotic Particles in High-energy Physics with Deep Learning." Nature Communications 5

Input dataset for hands-on

- Monte Carlo events
  - 21 low level features
    - pt's, angles, MET, b-tag, …
  - 7 high level features
    - Invariant masses (m(jj), m(jjj), …)



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/leggerf/MLCourse-INFN-2021

How to start

1. **Point your browser to the JHub link you received by email**
2. **Authenticate** through github
3. **Open a terminal:**
   - git clone https://github.com/leggerf/MLCourse-INFN-2021.git
4. **From JupyterHub Home tab:**
   - Notebooks/Day1/*inputForML.ipynb*
   - *You will receive the solutions tomorrow*

Start/stop jupyterHub