Big Data Challenges in Particle Physics

Fabrizio Salvatore, Lily Asquith, on behalf of EPP and TPP research groups

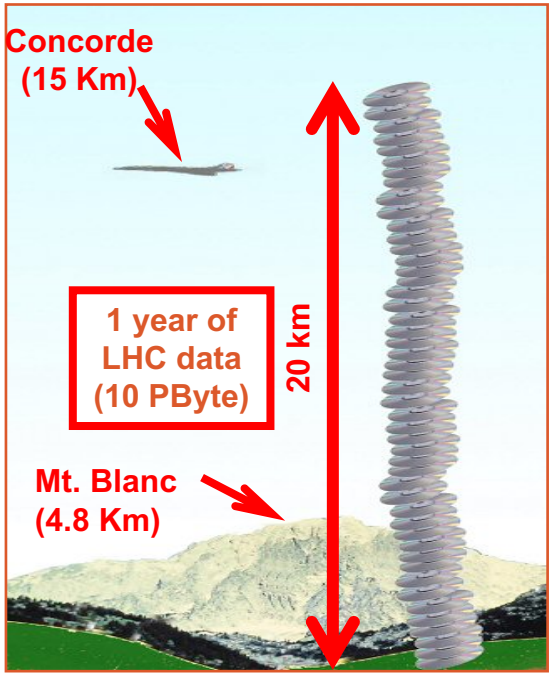
UNIVERSITY OF SUSSEX

Dealing with large data sets...

- In every single research discipline there is more and more discussion about big data
  - Try googling ‘Big Data’ !!

- In High Energy Physics there is a long tradition of dealing with large data sets
  - Big HEP Experiments – e.g. LHC, Lep (CERN), Tevatron (Fermilab), etc.
  - Simulation of Monte Carlo events for future studies – e.g. Linear Collider, future neutrino programme, etc.
  - Grid computing – interconnected computers used to analyse large data sets

- During the years there has been a lot of development of advanced statistical techniques to deal with the analysis of large data sets

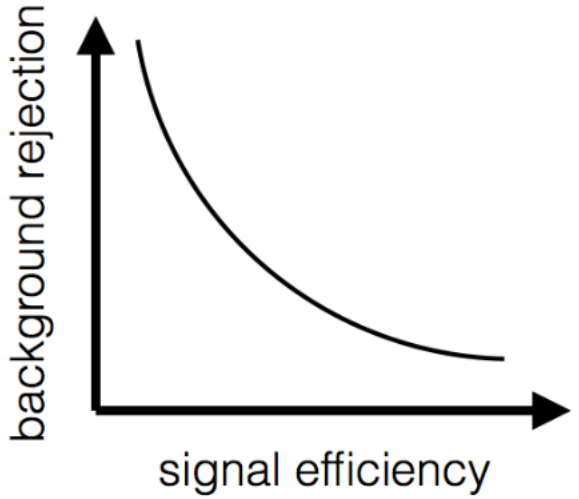


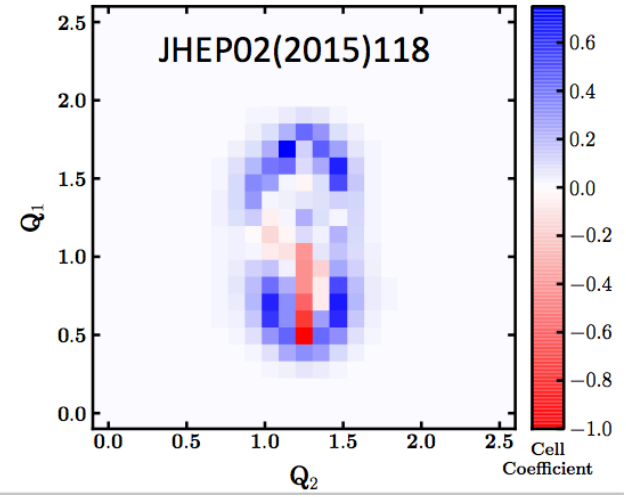
UNIVERSITY OF SUSSEX

Using advanced analysis techniques

We use Machine Learning (ML) techniques in all aspects of our analyses

- Optimisation of our signal selection
- Teaching to the algorithms how to learn
  - using physics knowledge to improve performances
- Learning from what the algorithm has learned
  - extract information about whether or not the algorithm is learning
  - use this information to design new variables
    - Visualisation can be a key component

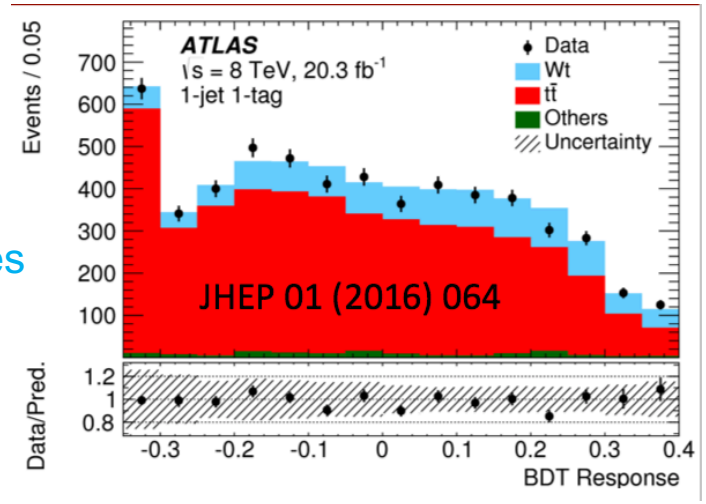


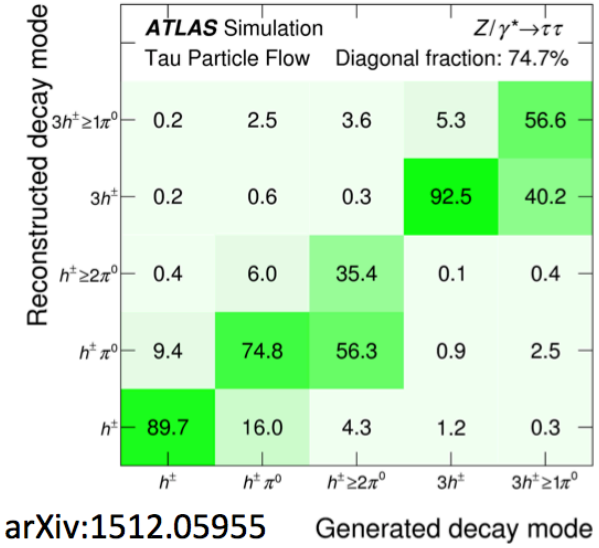


UNIVERSITY OF SUSSEX

Many applications of ML in HEP

- Analysis –
  - separation of signals from backgrounds
  - reconstruction/identification of complex states
    - e.g. unstable particles, long decay chains
- Detector reconstruction –
  - Energy calibration
  - Particle identification
- Computing –
  - Estimate needed number of simulated data for meaningful statistical comparison with ‘real’ data

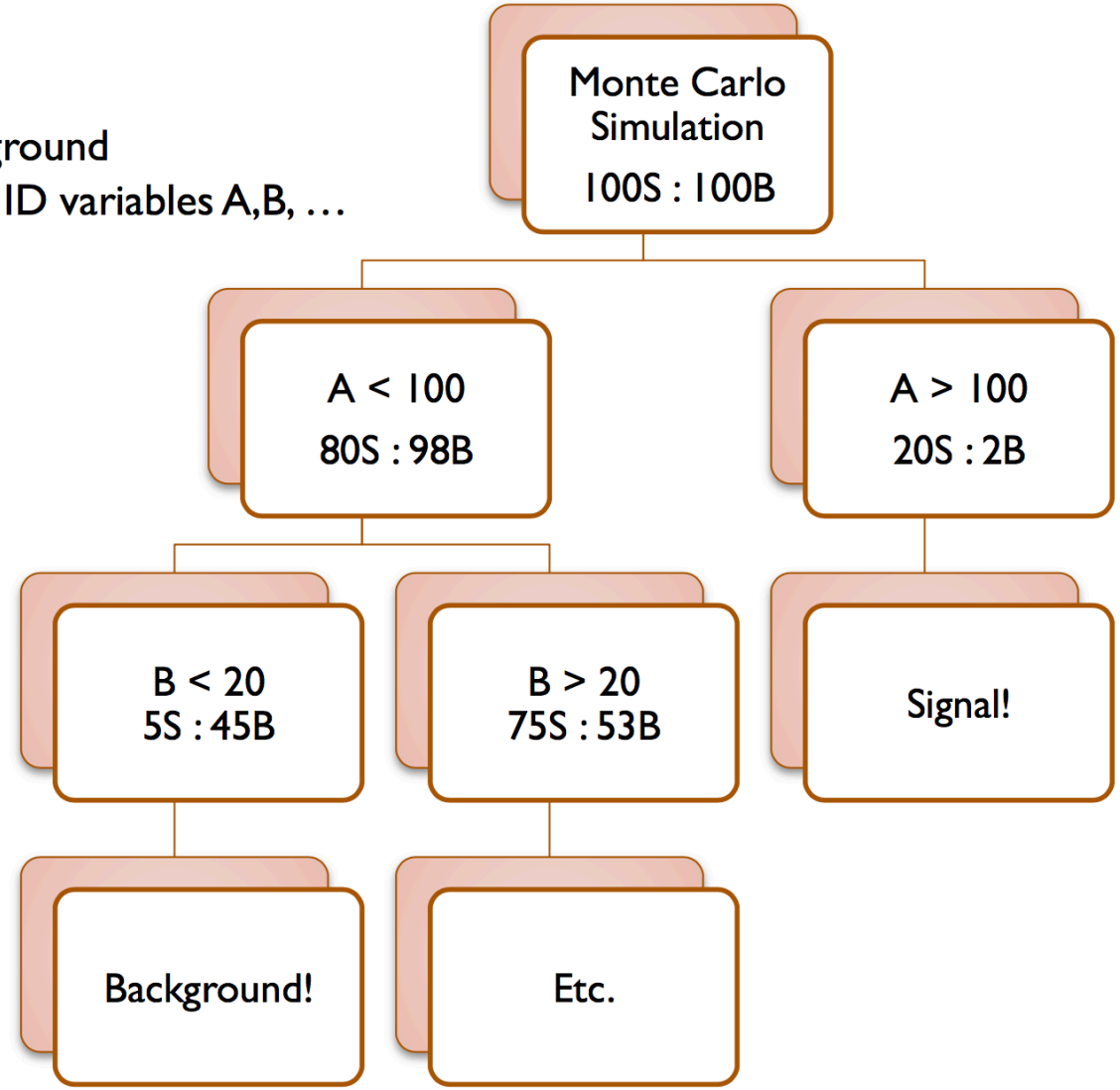


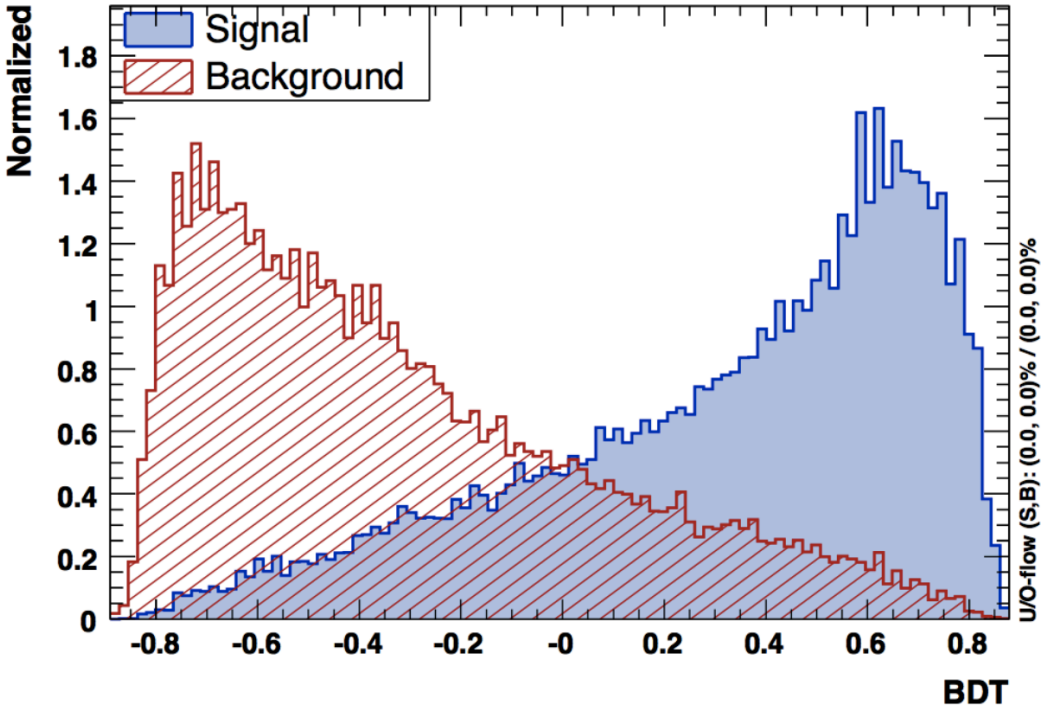


UNIVERSITY OF SUSSEX

Boosted Decision Trees (BDT)

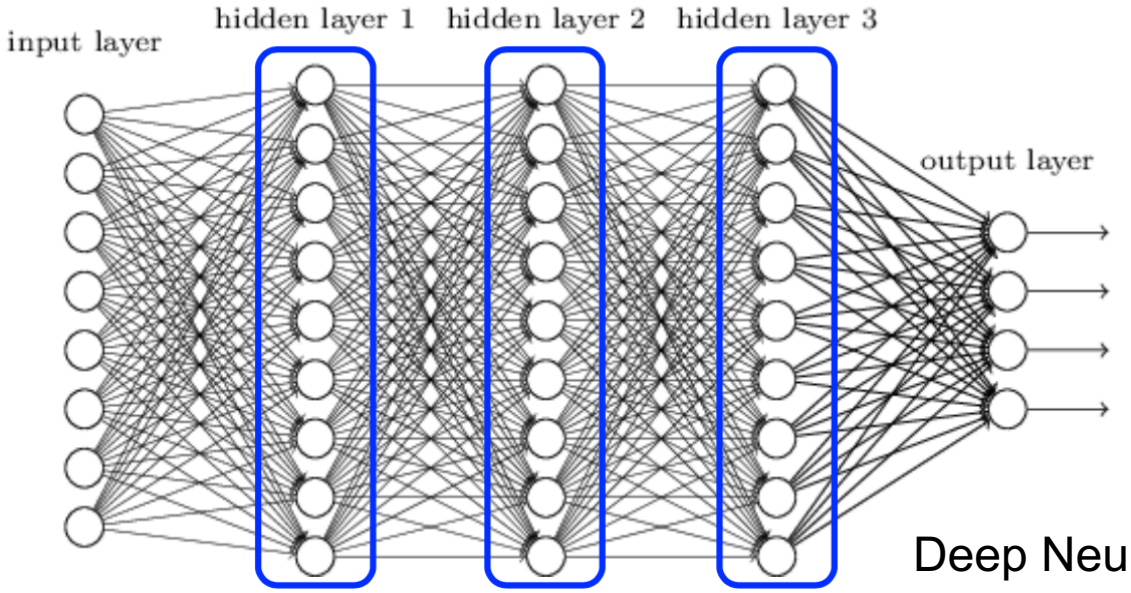
S: Signal
B: Background
Particle ID variables A,B, …



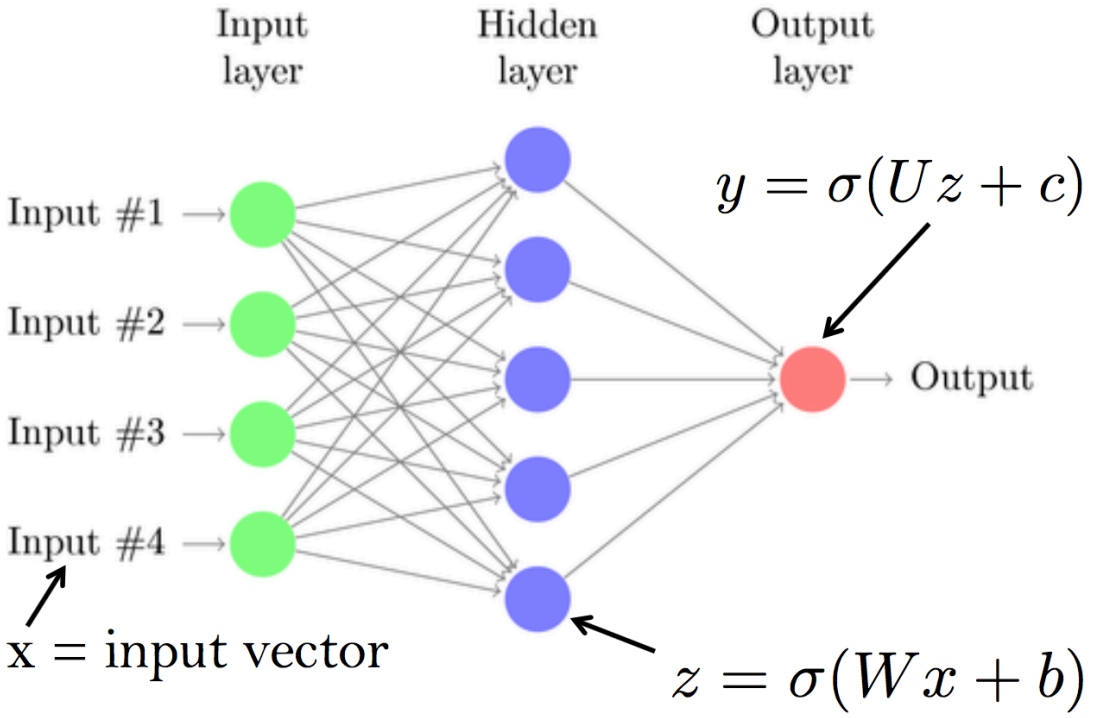


Neural Networks (NN)

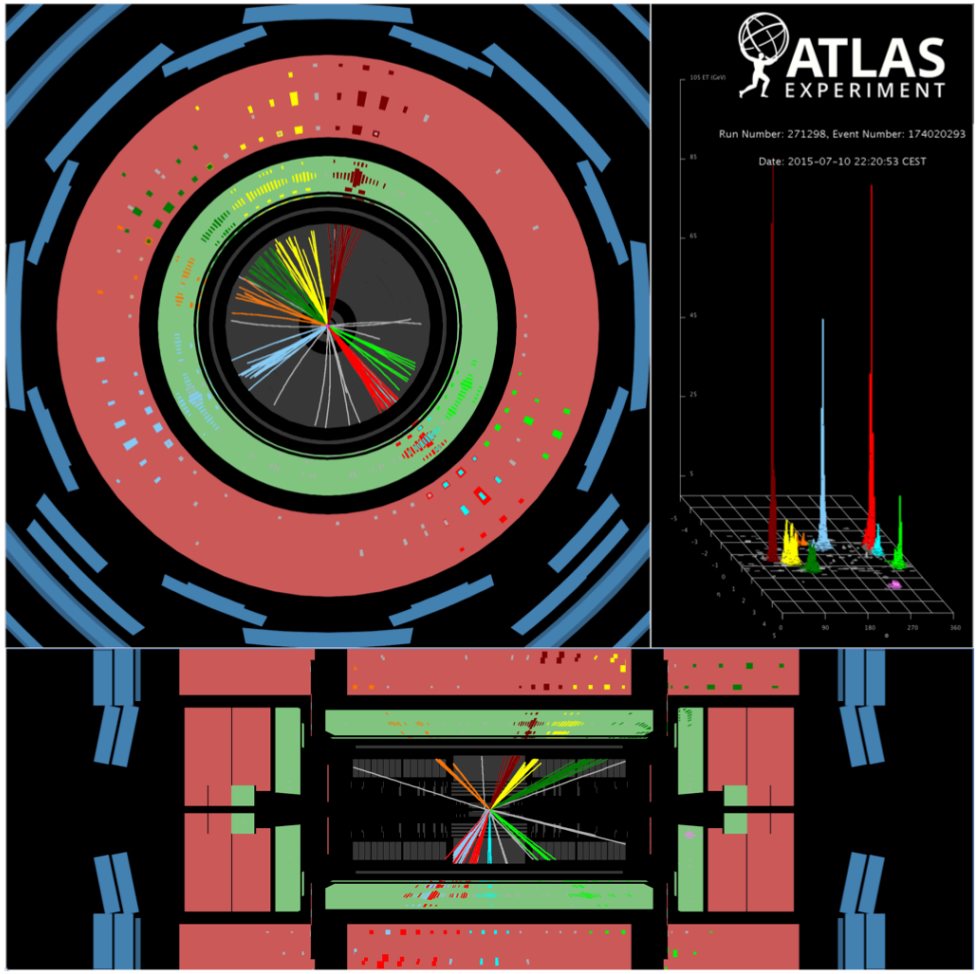
- Not a novel idea (I used them in my PhD thesis in 1998!)
- Care must be made to choose the correct variables to use
  - How many? Are they correlated?
- Important to know how many nodes in intermediate layer to avoid over-training

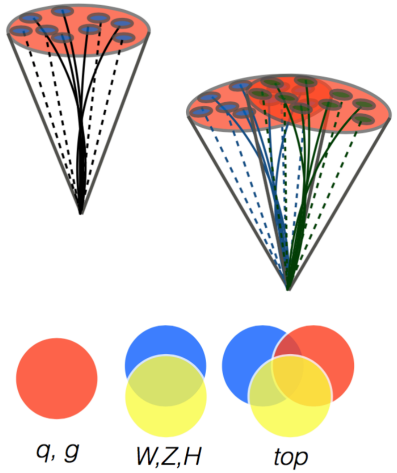


Deep Neural Networs (DNN)



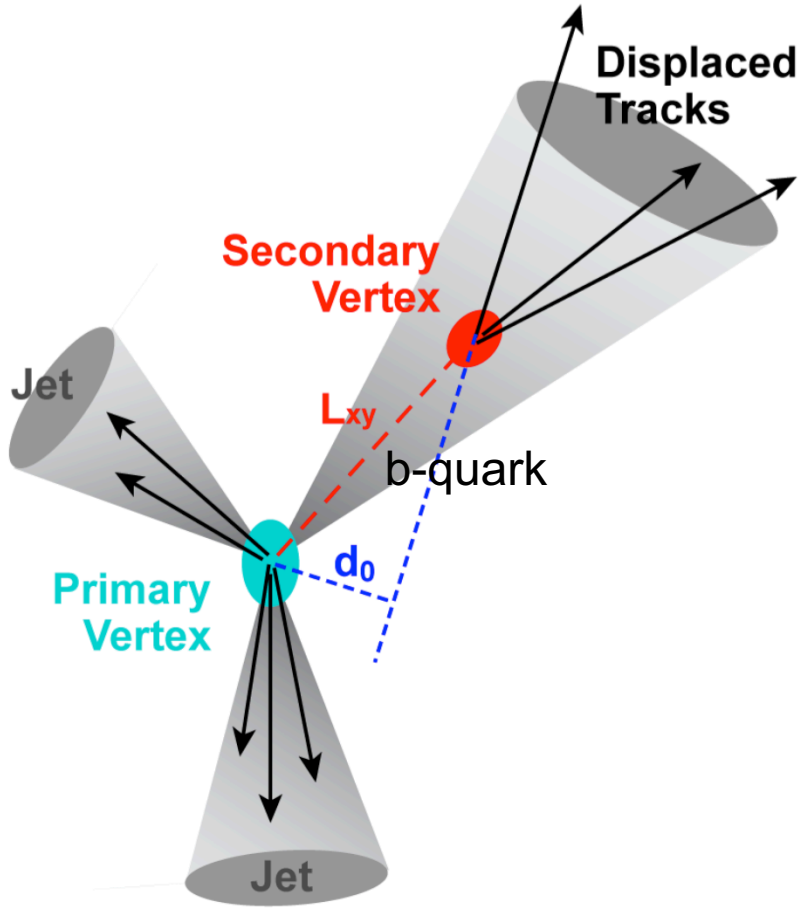
Example: finding jet of particles at LHC





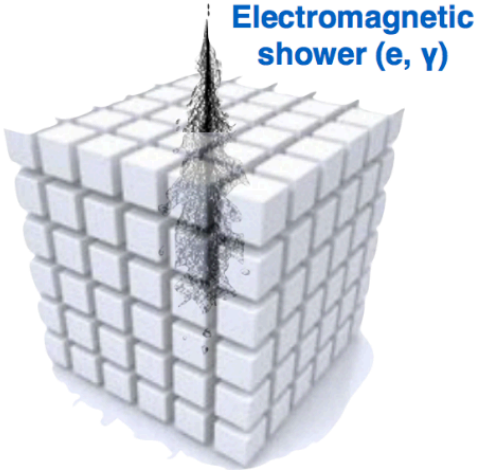
UNIVERSITY OF SUSSEX

Identifying b-quark jets



Where can ML go next?

- A lot is being developed for using ML in HEP
  - use computer vision and imaging techniques to track particles very precisely within a detector volume

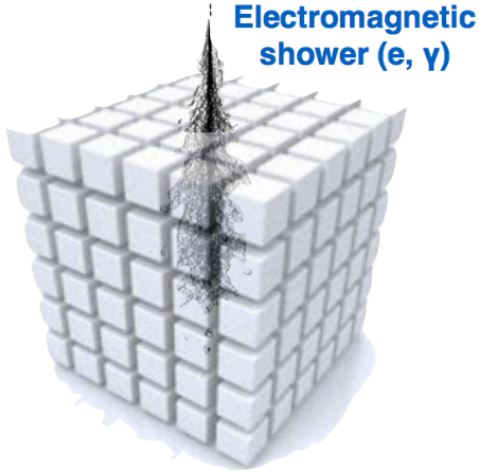


UNIVERSITY OF SUSSEX

Where can ML go next?

- A lot is being developed for using ML in HEP
  - use computer vision and imaging techniques to track particles very precisely within a detector volume

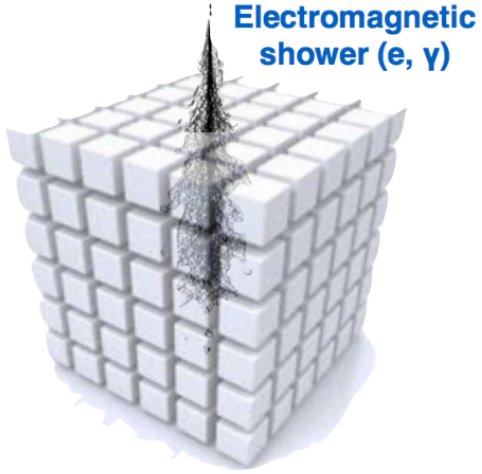
- Application to other disciplines
  - Anywhere were analysis of large datasets is required
    - health (malaria, HIV/AIDS, etc.)
    - environment (volcano watching, water quality, etc.)
    - electoral system data analysis
    - ……



UNIVERSITY OF SUSSEX

Where can ML go next?

- A lot is being developed for using ML in HEP
  - use computer vision and imaging techniques to track particles very precisely within a detector volume



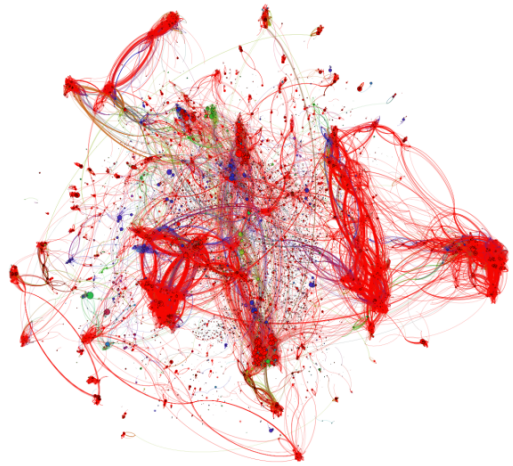
- Application to other disciplines
  - Anywhere were analysis of large datasets is required
    - health (malaria, HIV/AIDS, etc.)
    - environment (volcano watching, water quality, etc.)
    - electoral system data analysis
    - ……

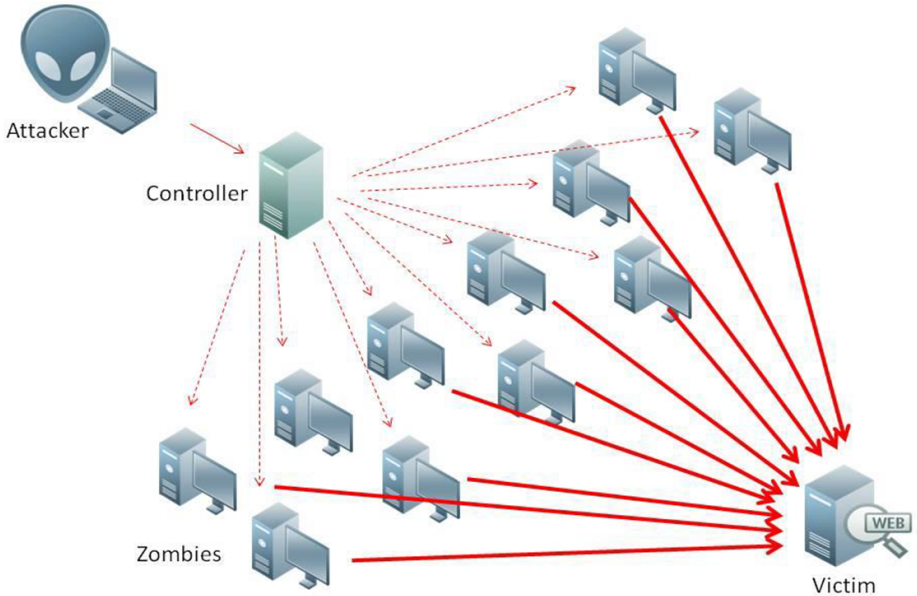
- Application to industry
  - Image recognition (CCTV, Security screening, etc)
  - tests of combustion engines
  - Network security (see next slide)

UNIVERSITY OF SUSSEX

From the Higgs to Network security

- The ATLAS detector at the Large Hadron Collider collects data at enormous rates, then filters and saves it at 2 Gbps
- Sussex developed techniques and software (aka Trigger) to process data arriving at 200 Gbps
  - Innovative architecture, algorithms and machine learning techniques





- Similar data-rates are experienced in the largest cybersecurity DDOS attacks
  - Botnets are used to attack and flood computer systems
- With industry partners we are translating our expertise and IP to this pressing problem for society
  - Just obtained funding from STFC to pursue this line of research with our industrial partners

UNIVERSITY OF SUSSEX

Conclusion

- There is vast expertise in the PP group in handling and analysing large data set
- We use cutting edge statistical techniques that could find application in many other areas of reseach
  - Machine learning, Neural Networks, Grid, ….
- We are already looking to apply these techniques to pressing everyday problems in society
  - Network security
- We are confident that there are opportunities to start collaborating with other research group in the University
  - Hope to find some common ground at this meeting

UNIVERSITY OF SUSSEX