Computing challenges from a high energy physicist point of view

J.P. Araque
On behalf of the exotics searches group @ LIP
jp.araque@cern.ch

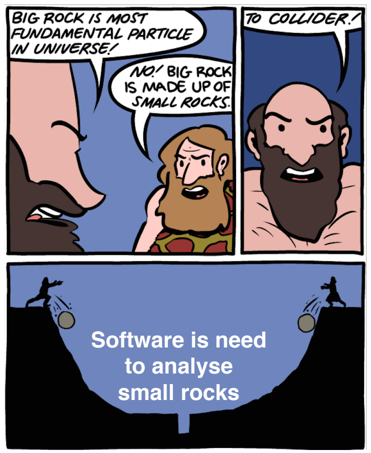
http://minho.lip.pt

Presenting myself

- Physics degree in Universidad de Granada (2010)
- Master degree in Universidad de Granada (2012)
- Ph.D. student in the ATLAS collaboration (CERN) within the Portuguese group (From 2012 to first half of 2015)
- Working on:
  - Physics analysis in search for vector-like quarks production and the LHC.
  - Noise description in the TileCal calorimeter in the ATLAS detector.
- Paper published: http://arxiv.org/abs/arXiv:1409.5500

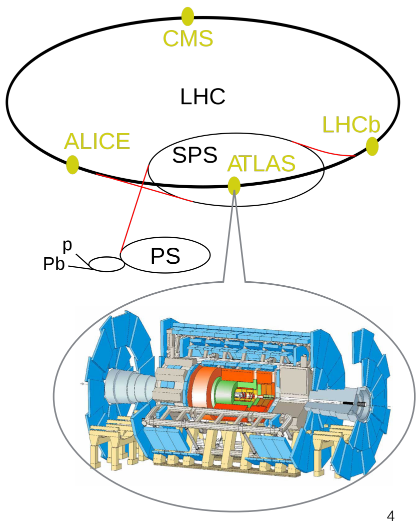
Science and tools

When science was…	Tools were…
Counting sheeps	Fingers
Follow the stars	Telescope
Study natural forces	Magnets, masses, springs…
Study the deep universe	Satellites
Particle physics	Colliders and detectors



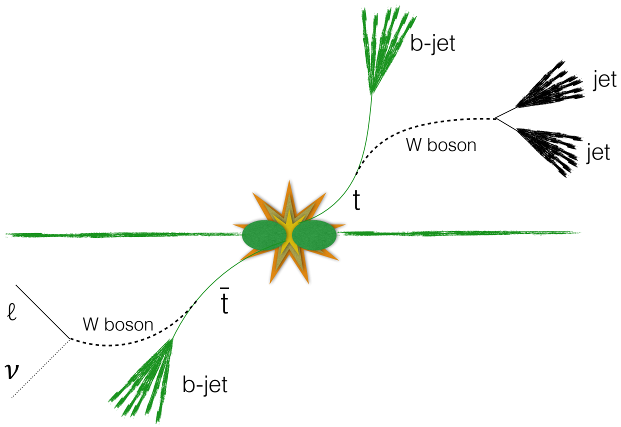
ATLAS challenges

- The LHC accelerates proton bunches at 7TeV (v~99.9999991% of c).
- ATLAS:
  - Has to deal with 40 x 10^6 collisions per second.
  - Produce 1Pb of data per second.
- **This will be increased next year… Why?**
  - **Because we, physicists, don't know when enough is enough.**
  - **Higher energy and luminosity -> higher sensitivity to new physics and statistics -> high technical challenges.**



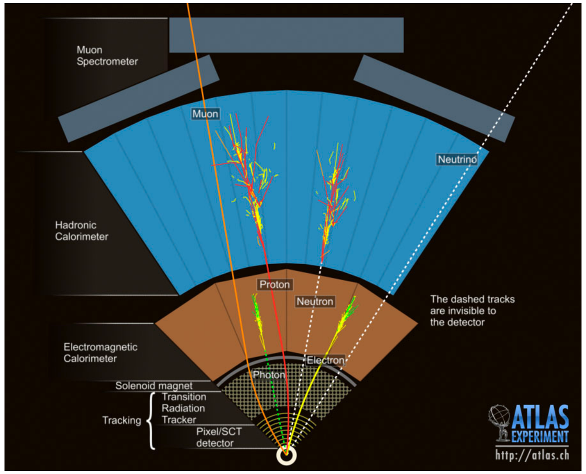
What do we want?

Data measurements



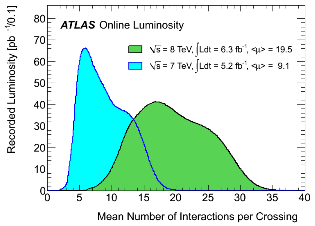
How do we get it?

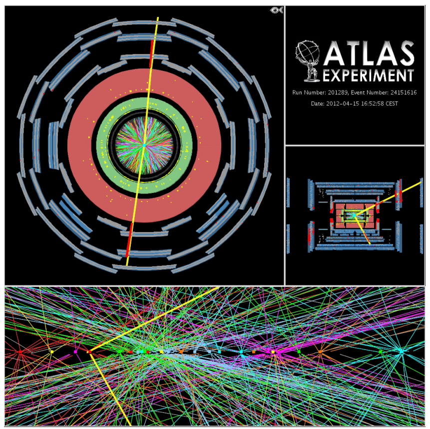
Data measurements



A crowded place

Data measurements





- The detector has to:
  - Pick up the interesting events among all possible collisions.
  - Distinguish different interaction points.
  - Get useful information from them to make the decision.

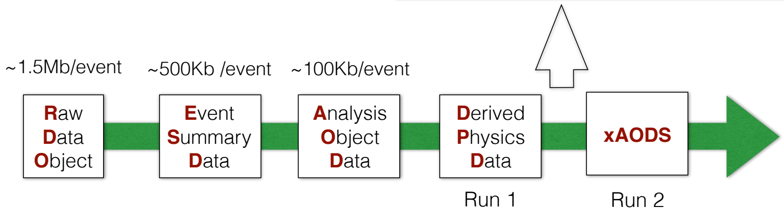
Data model in ATLAS

Data measurements

Project	C/C++ code	C/C++ headers	PYTHON code	Total code
Core	390,000	43,000	240,000	860,000
Event	200,000	110,000	16,000	350,000
Conditions	280,000	90,000	21,000	620,000
Detector	38,000	6,100	8,400	140,000
Sum	910,000	250,000	280,000	2,000,000

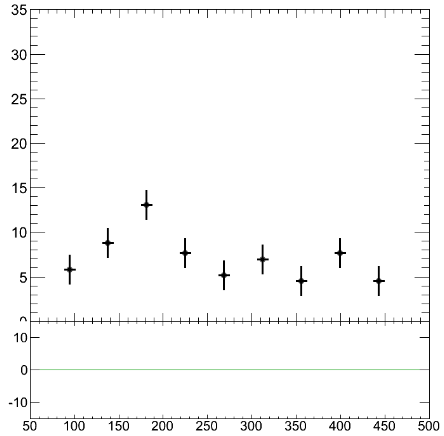
ATLAS software framework
ATHENA (2010)

- **DPD**
  - Linearised event information for further analysis. Group or analysis dependent.
- **xAODS**
  - Keep event and object structure of AODs but make different flavours (derivations) for each group or analysis.



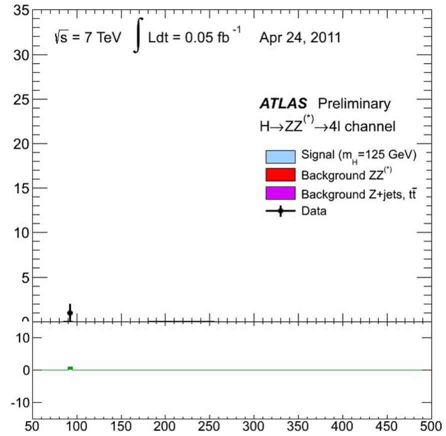
Data is not everything

MC simulation



Data is not everything

MC simulation



High energy physics simulation

MC simulation

~1.5Mb/event ~500Kb /event ~100Kb/event

~~Raw Data Object~~ → Event Summary Data → Analysis Object Data → Derived Physics Data → xAODS

Run 1 Run 2

Event generation

MadGraph Mc@NLO

PROTOS

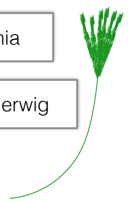
Sherpa AcerMC

Alpgen

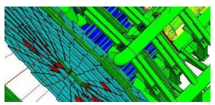
Hadronisation

Pythia

Herwig



Detector simulation



Geant 4

PGS

Delphes

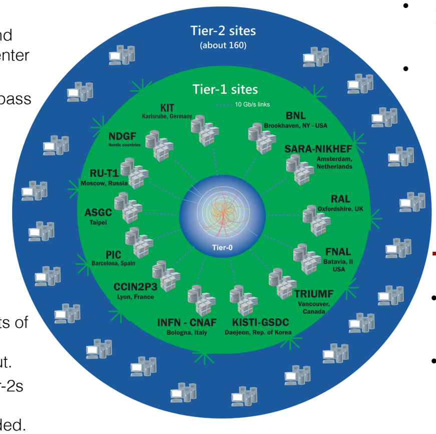
The Worldwide LCH Computer Grid (WLCG)

Tier 0

- Located at CERN and Wigner Research center in Budapest.
- Store raw data and pass first reconstruction.
- Send reconstruction data to tier-1 sites.

Tier1

- There are 13 tier-1s around the world.
- Safe-keeping of parts of the raw data and reconstruction output.
- Comunicate with tier-2s and massive reprocessing if needed.



Tier 2

- About 160 around the world. Typically Universities or research centers.
- Enough storage capacities and computing power for specific analysis task, event generation or reconstruction.

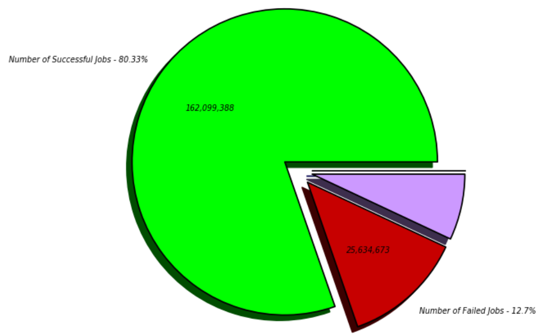
Tier 3

- Small computer centres for data analysis.
- Typically local clusters where physicists run analyses and has no formal engagement with the WLCG.

Performance of the WLCG

dashboard

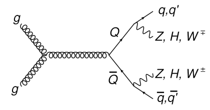
Number of Successful and Failed Jobs (Sum: 201,780,915)

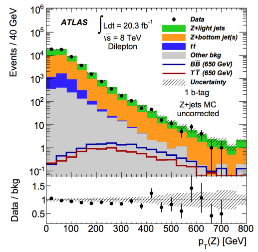


Number of Successful Jobs - 80.33% (162,099,388)
Number of Cancelled Jobs - 6.96% (14,046,853)
Number of Failed Jobs - 12.7% (25,634,674)
Number of Unknown-Status Jobs - 0.0% (0.00)

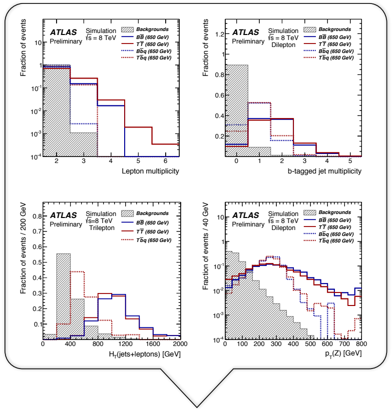
What do I really do?

Vector-like quarks search as an example



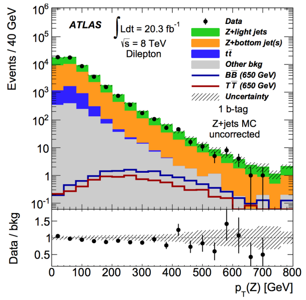


~50k events bkg.
vs
~20 events signal
After some selection is done

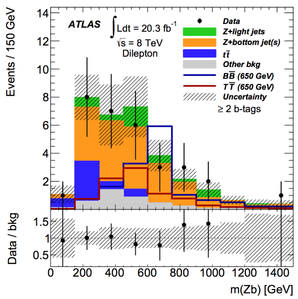


Find different behaviour between signal and background and
make event selection based on those variables.

What do we really do?



~50k events bkg.
vs
~20 events signal
After some selection is done



~30 events bkg.
vs
~15 events signal

How do we do it?

Very simple structure

Initialization

- Different input format.
- Special normalisation needed?
- Special tools needed? Dependencies satisfied?
- Is any correction needed?

The devil lies in the detail

Output generation

- Do you need a simple root file with histogram?
- Do you need to perform any modification?
- Do you need to organise the output in a given way for an external tool?

Loop

- How many times do you need to loop?
- Systematic estimation not homogeneous.
- Any PDF need to be derived?
- Any reconstruction is needed? If so, what is information is necessary?

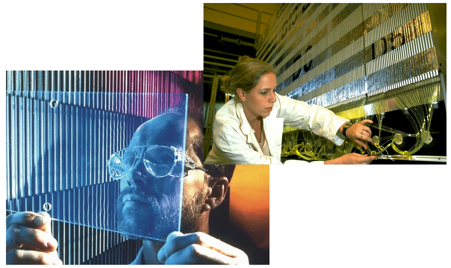
Example (in a local cluster)

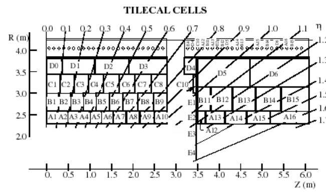
- Vector-like quark analysis steps until publication:
    - Small code to validate MC simulation (~10s to run).
    - Start studying data and background after a first skimming in WLCG (~7min to run).
    - Increased luminosity (~15min to run).
    - Define 3 different signal regions (3x times previous analysis, ~1h to run).
    - Study different corrections to dominant background (up to 4 different corrections, ~3h to run).
    - Decide that it is starting to take too much time and rewrite the code to accommodate what wasn't needed when it was written (~1h to run).
    - Add systematics… 50 of them! (~1 day to run).

This doesn't include plots production or limit calculation.

The yesterday factor: The analysis code development cannot be disconnected from the analysis process itself. No much time to think about optimisations. You never know what is to come.

ATLAS hadronic calorimeter: TileCal



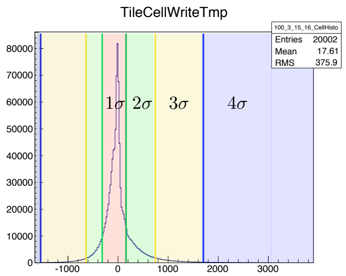


- Technical work requires reading low level data from detector to better understand the measurements.
- TileCal consist of:
    - 4 partitions.
    - 4 layers.
    - 16 towers.
    - 64 modules.

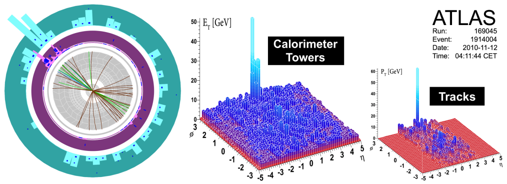
Total of 16384 cells! Each one contains information from each event measured.

Detector analysis

- Each cell measures the energy deposition of strong interacting particles.



- We don't only have a lot of events to read but a lot of information per event.
- Differences on running the code for data or MC or for a given detector configuration.
- For each cell we need to extract shape information for energy deposition of pileup events (non "interesting" events)

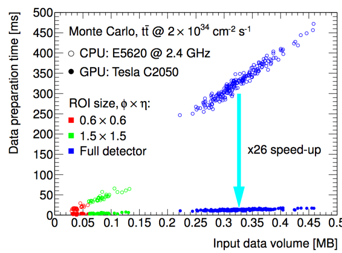


- It is important to know where to start looking for cluster of energy and reconstruct jets.

Software optimisation: a very alive subject

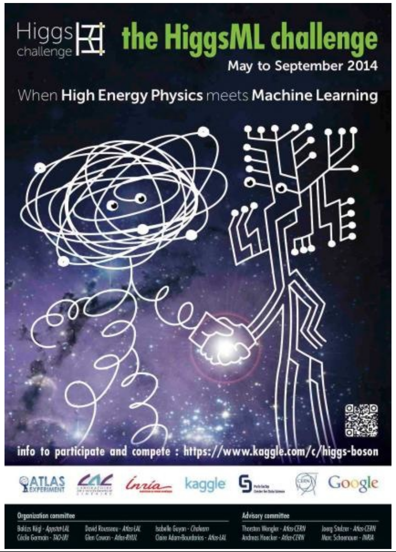
- Work is being done to improve the software used:
- **For Analyses:**
    - ROOT is always evolving to improve both speed and memory usage.
    - All analysis groups are focused preparing run-II software tools to reduce disk space and execution time.
    - Parallel analyses have been used in the past (PROOF, OpenMP…) and are being improved.
- **For detector physics:**
    - GPUs are being studied in triggers.
    - Vectorisation is being also considered and some detectors have already implemented them.

GPUs for high level trigger in ATLAS



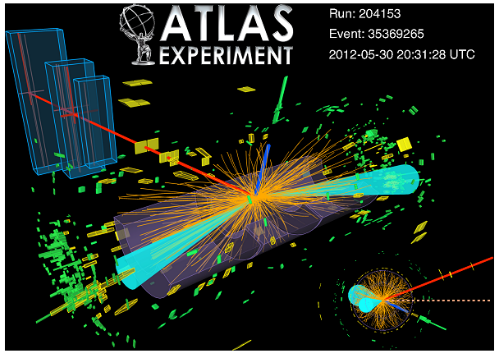
D Emeliyanov and J Howard 2012 J. Phys.: Conf. Ser. 396 012018

When HEP mets Machine Learning



Higgs learning machine challenge

#	Prize
1	7000 $
2	4000 $
3	2000 $
ML	Trip to meet ATLAS collaboration with $2700 for expenses



https://www.kaggle.com/c/higgs-boson/

- Given a set of data for Higgs events train a machine to recognise H->tau-tau events and optimise the ratio S/√B.
- Techniques used by winners:
  - Set of deep neural network.
  - Complex C++ code, GPU parallel code and OpenMP.

Not always a totally correct answer is a totally valid one