The Higgs Boson Machine Learning Challenge

ATLAS Collaboration

HEP

Machine Learning

Doing Big Data for 30 years

Using MVA for 20 years (neural net, boosted decision trees, and more)

HiggsML challenge

Large players in industry, and academics

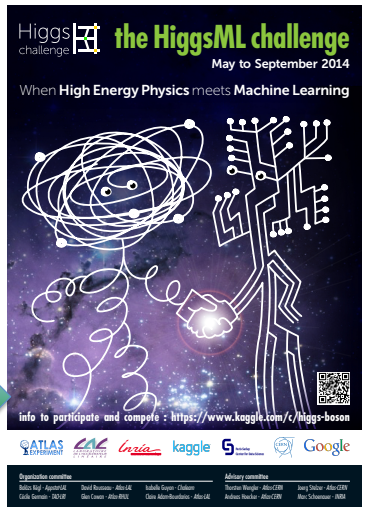
Large computing power

Complex problems

Exponential growth of the field. New algorithms and techniques

Let's put some ATLAS simulated data on the web and ask data scientists to develop the best machine learning algorithm to find the Higgs !

Jointly organised by ATLAS physicists and data scientists:



How does it work ?

- Participant register to kaggle
- Download *training* dataset (250 kEvt) **with** weight and sig or bkg label
- Train their best algorithm, adapt it to the specificities of the Challenge
- Download test dataset (550 kEvt) **without** weight and label
- Apply trained algorithm to test dataset, determine s/b classification

Public leaderboard from 100 kEvt

Opened 12th May
Snapshot 26th June
811 participants
3959 downloads

Private leaderboard from 450 kEvt

15th Sep

Best scores

Best method

Invation to CERN/ATLAS

Choice of figure of merit

Need one and only one robust estimator of the quality of the classification algorithm : Approximate Medium Significance AMS

Given s and b expected number of signal and background normalised to 2012 luminosity:

s=Σ(selected signal) weights$_i$

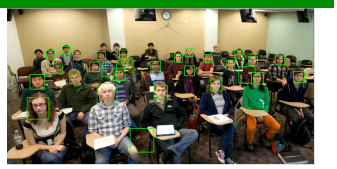
b=Σ(selected bkg) weights$_i$

Decided to use the "Asimov" formula (G. Cowan, K. Cranmer, E. Gross, and O. Vitells, "Asymptotic formulae for likelihood-based tests of new physics", *EPJC*, vol. 71, pp. 1–19, 2011.) with « regularization » : b'=b+10 (avoid large fluctuations in small regions)

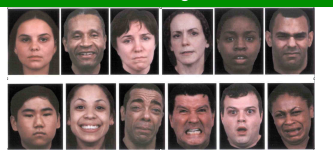
$$AMS = \sqrt{2}\sqrt{(s+b')\log(1+s/b') - s}$$

Character recognition

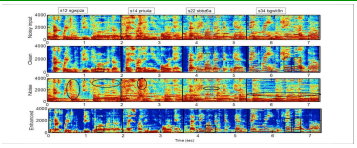
Real time face detection



Emotion recognition



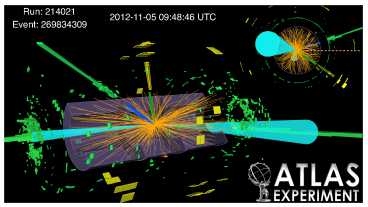
Speech recognition

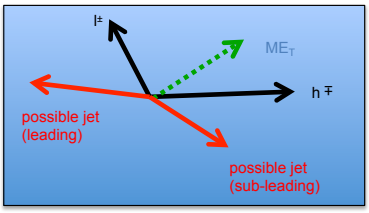


Choice of ATLAS H→ττ analysis

- "Evidence for Higgs Boson Decays to $\tau^+\tau^-$ Final State with the ATLAS detector" ATLAS CONF-2013-108
- 4.1 σ observed (3.2 σ expected)
- Direct evidence of Higgs coupling to leptons
- Complex analysis (e.g. signal H mass ~one σ from dominant Z→$\tau^+\tau^-$ background, VBF signature…)

lepton-hadron topology





Real analysis vs Challenge

	Real analysis	Challenge
1.	Systematics	No systematics
2.	2 categories x n BDT score bins	No categories, one signal region
3.	Background estimated from data (embedded, anti tau, control region) and some MC	Straight use of ATLAS Geant4 MonteCarlo : signal Higgs, backgrounds Z, W, top
4.	Weights include all corrections. Some negative weights.	Weights include normalisation and generator weight. Negative weight events rejected.
5.	Potentially use any information from all 2012 data and MonteCarlo events	Only use variables and events preselected by the real analysis
6.	Few variables fed in two BDT	All BDT variables + categorisation variables + primitives 3-vector
7.	Significance from complete fit with Nuisance Parameters, Control Regions, etc…	Significance from "regularised Asimov"
8.	MVA with TMVA BDT	MVA "no-limit"

Simpler, but not simple!

Variables provided

Weight and signal/background label

weight
label

Conference note DER ived variables used for categorization or BDT (VBF and Boosted categories):

DER_mass_MMC
DER_mass_trans_met_lep
DER_mass_vis
DER_pt_h
DER_deltaeta_jet_jet
DER_mass_jet_jet
DER_prodeta_jet_jet

DER_deltar_tau_lep
DER_pt_tot
DER_sum_pt
DER_pt_ratio_lep_tau
DER_met_phi_centrality
DER_lep_eta_centrality

PRImitive 3-vectors allowing to compute the DER variables (mass neglected)

PRI_tau_pt
PRI_tau_eta
PRI_tau_phi
PRI_lep_pt
PRI_lep_eta
PRI_lep_phi
PRI_met
PRI_met_phi
PRI_met_sumet

PRI_jet_num
PRI_jet_leading_pt
PRI_jet_leading_eta
PRI_jet_leading_phi
PRI_jet_subleading_pt
PRI_jet_subleading_eta
PRI_jet_subleading_phi
PRI_jet_all_pt

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

Contact : higgsml@lal.in2p3.fr

ATLAS EXPERIMENT