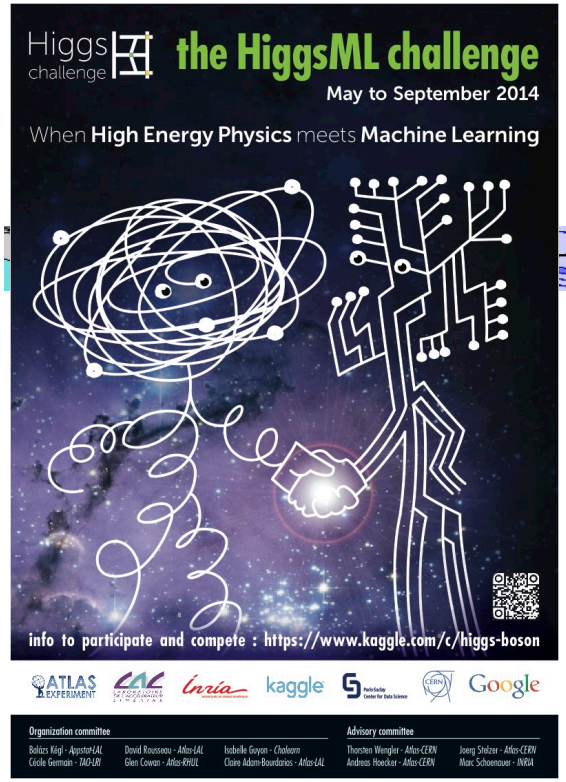
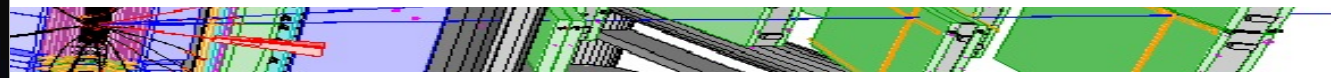
Higgs Machine Learning Challenge what now?





David Rousseau
LAL-Orsay
rousseau@lal.in2p3.fr

4th Dec 2015, Interexperiment Machine Learning WG

… in a nutshell

- (almost same talk I did in September, I'll be quick)
- Why not put some ATLAS simulated data on the web and ask data scientists to find the best machine learning algorithm to find the Higgs ?
  - Instead of HEP people browsing machine learning papers, coding or downloading possibly interesting algorithm, trying and seeing whether it can work for our problems
- Challenge for us : make a full ATLAS Higgs analysis simple for non physicists, but not too simple so that it remains useful
- Also try to foster long term collaborations between HEP and ML

Participation

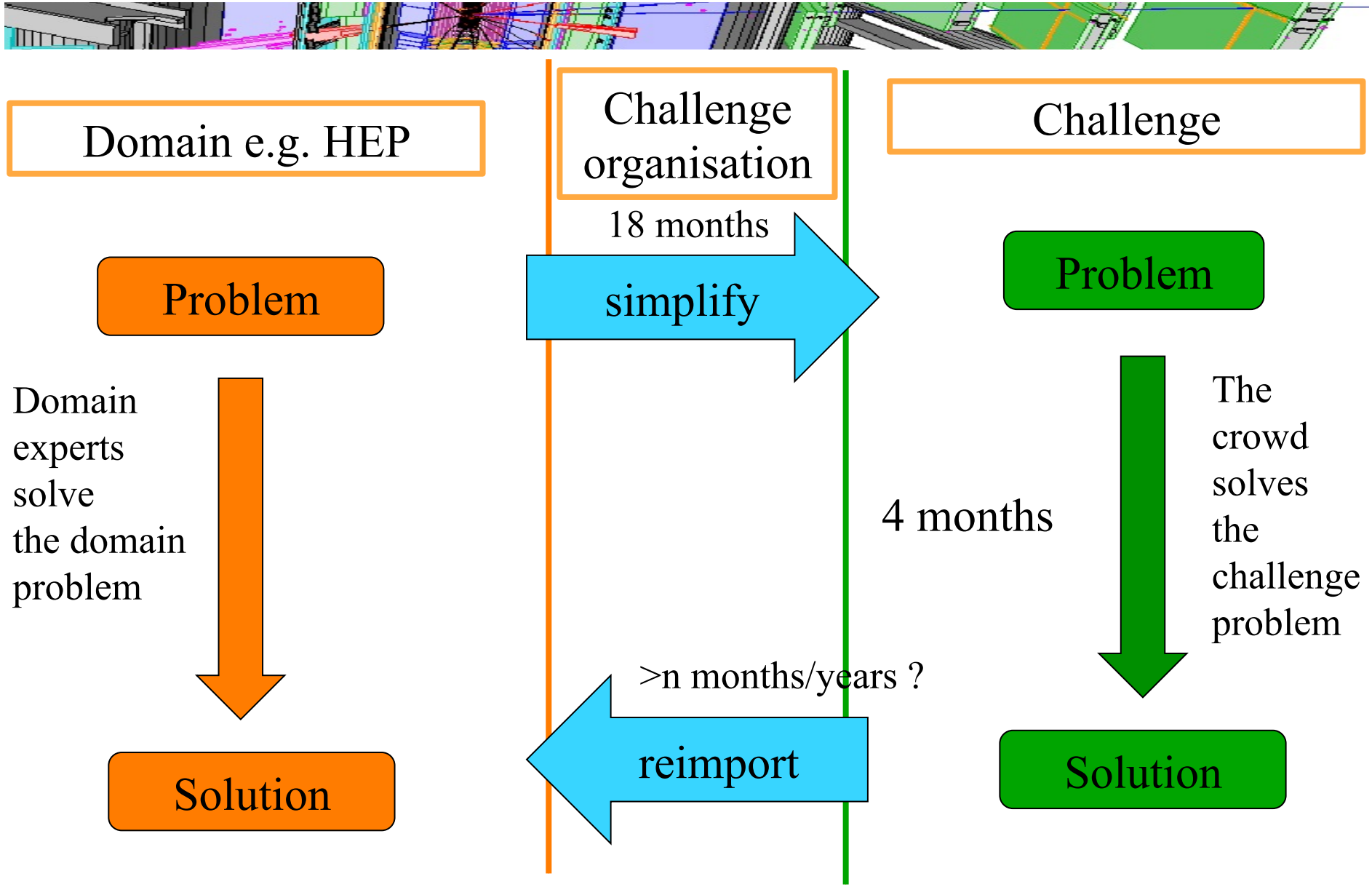
- (see spares for a description of the challenge)
- Big success !
- 1785 teams (1942 people) have participated (participation=submission of at least one solution)
  - (6517 people have downloaded the data)
  - ➔most popular challenge on the Kaggle platform (until spring 2015)
  - 35772 solutions uploaded
- 136 forum topics with 1100 posts
- Many participants have worked very hard

Final leaderboard

Also hired by DeepMind this summer!

#	Δrank	Team Name ‡ model uploaded * in the money				Last Submission)
1	↑1	Gábor Melis ‡ *	7000$	3.80581	110	Sun, 14 Sep 2014 09:10:04 (-0h)
2	↑1	Tim Salimans ‡ *	4000$	3.78913	57	Mon, 15 Sep 2014 23:49:02 (-40.6d)
3	↑1	nhlx5haze ‡ *	2000$	3.78682	254	Mon, 15 Sep 2014 16:50:01 (-76.3d)
4	↑38	ChoKo Team		3.77526	216	Mon, 15 Sep 2014 15:21:36 (-42.1h)
5	↑35	cheng chen		3.77384	21	Mon, 15 Sep 2014 23:29:29 (-0h)
6	↑16	quantify		3.77086	8	Mon, 15 Sep 2014 16:12:48 (-7.3h)
7	↑1	Stanislav Semenov & Co (HSE Yandex)		3.76211	68	Mon, 15 Sep 2014 20:19:03
8	↓7	Luboš Motl's team		3.76050	589	Mon, 15 Sep 2014 08:38:49 (-1.6h)
9	↑8	Roberto-UCIIIM		3.75864	292	Mon, 15 Sep 2014 23:44:42 (-44d)
10	↑2	Davut & Josef		3.75838	161	Mon, 15 Sep 2014 23:24:32 (-4.5d)
45	↑5	crowwork ‡	HEP meets ML award XGBoost authors Free trip to CERN	3.71885	94	Mon, 15 Sep 2014 23:45:00 (-5.1d)
782	↓149	Eckhard	Tuned TMVA	3.49945	29	Mon, 15 Sep 2014 07:26:13 (-46.1h)
991	↑4	Rem.		3.20423	2	Mon, 16 Jun 2014 21:53:43 (-30.4h)
		simple TMVA boosted trees		3.19956		

From domain to challenge and back



What did we learn

- Very successful full day satellite workshop at NIPS (one of the two major Machine Learning conferences) in Dec 2014 @ Montreal:

  https://indico.lal.in2p3.fr/event/2632/
- Proceedings published (August 2015 : JMLR Workshop and Proceedings Vol 42 http://jmlr.org/proceedings/papers/v42/) Contributions from some of the top players, plus summary from organisers
- Many additional piece of information in kaggle forum or random blog or github repository
- Each participant have used a range of ideas selected by trial and error ➔ difficult to decipher what really worked best at the end

Algorithms

- "deep" Neural Nets win (Gabor Melis).
  - Gabor's words : "deep" in 2014 because using 3 hidden layers, would not qualified as deep nowadays
- BDT marginally behind (number 2 was 0.02 behind in significance)
  - Gabor's words : NN not worth it, too much work/too many possibilities to tune the training
- Meta-ensemble (combining BDT or NN with different hyper parameters) marginally better, but much more complex, not worth it
- Conclusion : for a typical HEP problem, BDT should be the default choice (OK we sort of knew about it)

Software

- Lots of development of Machine Learning Open Source software outside HEP
- In particular:
  - XGBoost (eXtreme Gradient Boosting): released for the HiggsML challenge, used by many participants. Now used in other challenges as well. One of the best software on the market for BDT/BRT. Good performance out of the box. Also fast (multithreaded)
  - SciKit-learn : large developer/user base. Toolbox like TMVA. Used already a bit in ATLAS (e.g. Htautau hadhad channel at least)
  - Note that both software were improved thanks to the challenge, in particular to handle event weights
  - Note that quite often these Open Source software are routinely multithreaded, and sometimes even run on GPU (NN, not BDT)

Feature Engineering

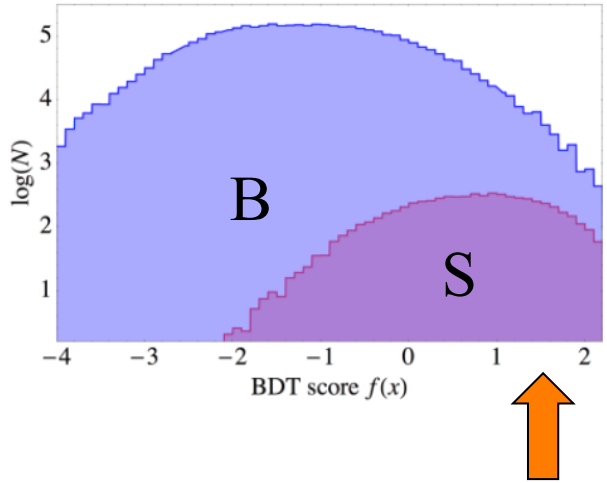
- In ML jargon, this is the building of new variables from the original ones
- We (HEP) have been doing this since the beginning of times
- Given enough training data, ML techniques could "discover" these features (e.g. invent the concept of transverse mass)
- It did not work at all for HiggsML (significance less than 3 (wrt 3.8) if removing the high level variables provided
- There are techniques to automatically generate new features
- Not clear they would beat HEP expertise

Cross Validation

- Cross Validation (CV) are techniques to measure MVA performance independently of the training
- Goal is to build an optimisation curve (e.g. significance, ROC,..) with the smallest variance (despite lack of data), for a better optimisation of hyper parameters or choice of techniques
- Default TMVA CV (one fold CV):
  - split sample in two halves A and B.
  - train on A, test on B
- Two-fold CV (e.g. ATLAS Htautau analysis)
  - Split sample in two halves A and B
  - Train on A, test on B; train on B test A
  - ➔test statistics = total statistics➔double test statistics wrt one fold CV (double training time of course)
- Five-fold CV (e.g. Gabor)
  - Split sample in 5 equal pieces A,B,C,D and E
  - Train on ABCD, test on E;train on ABCE, test on D; etc…
  - ➔same test statistics wrt Two-fold CV, but larger training statistics 4/5 over ½ (larger training time as well)
- CV à la Gabor (he did not invent it but no better name)
  - Redo Five-fold CV e.g. 4 times with a different random splitting
  - For each event, one has now 4 different (but similar) scores. Then
    - Average these scores (with whatever definition of "average")
    - Or build directly the optimisation curve from the 4*N scores, with additional weight ¼➔smoother curve

Focussed learning

- By default, classification algorithm will optimise the overall ROC curve (typically the Area Under roc Curve, AUC)
- However we often have more specific figure of merit, like the significance à la s/sqrt(b) (which depend of the size of the subsample) (not to mention full blown RooFit !)



- We want to focus on a specific region of the ROC curve (not AUC), with background rejection >95% and signal efficiency ~10-20%
- Different techniques have been used by participants to handle this:
  - Hyper parameter optimisation maximising significance (but quickly overtraining)
  - Chose internal training parameter as a function of the optimisation functional
  - Prescription to modify the event weights and iterate learning (Weighted Classification Cascade, see arxiv 1409.2655)
  - ➔need more study to understand what works best
- Note : systematics were deliberately ignored for the challenge, but these methods could be used with a significance including systematics

Random additional ideas

- The following were also mentioned in brainstorming at the NIPS workshop or in other discussions triggered by the challenge (not complete!)
  - Use of deep learning (see papers by Baldi, Sadowski, Whiteson): able to guess high level feature on some problem but not others ?
  - Handle the lack of training statistics (in particular for deep learning) by combining in a clever way full/fast sim (almost accurate but expensive) and super fast sim (~Delphes) (inaccurate but cheap). ML jargon "Transfer Learning"
  - Transform a classification problem in a regression problem, easier to train, with a surrogate function. E.g. build a very sophisticated deep NN, train it, then emulate it with a simple BDT
    - The simple BDT does not work better than the very sophisticated deep NN, but it is faster and better than training directly the simple BDT
  - Handling of systematics, data vs MC differences, is completely foreign to Machine Learning community (well, not any more actually)

Summary and outlook

- Wealth of disorganised input from the challenge participants. What we could decipher:
  - BDT still the algorithm of choice
  - Better software out there (XGBoost, SciKitLearn) than *current* TMVA
  - Many techniques beyond just BDT training (Cross Validation, focussed training etc…)
  - Lots of expertise in ML community we should tap into
- Pointer collection:
  - https://www.kaggle.com/c/higgs-boson
  - https://higgsml.lal.in2p3.fr
  - http://opendata.cern.ch/collection/ATLAS-Higgs-Challenge-2014: permanent home of the challenge dataset
  - https://indico.lal.in2p3.fr/event/2632/ NIPS 2014 workshop agenda and NEW proceedings http://jmlr.org/proceedings/papers/v42/
  - http://cern.ch/higgsml-visit mini workshop at CERN

Spares

Dataset

Permanently available and usable by anyone (also non ATLAS) on CERN Open Data:

http://opendata.cern.ch/collection/ATLAS-Higgs-Challenge-2014

ASCII csv file, with mixture of Higgs to tautau (lephad) signal and corresponding backgrounds, from official GEANT4 ATLAS simulation

Weight and signal/background (for training dataset only)

weight (fully normalised)

label : « s » or « b »

Conf note variables used for categorization or BDT:

DER_mass_MMC
DER_mass_transverse_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

} VBF signature

Primitive 3-vectors allowing to compute the conf note variables (mass neglected),

16 independent variables:

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 (0,1,2,3, capped at 3)
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

} VBF signature

Real analysis vs challenge

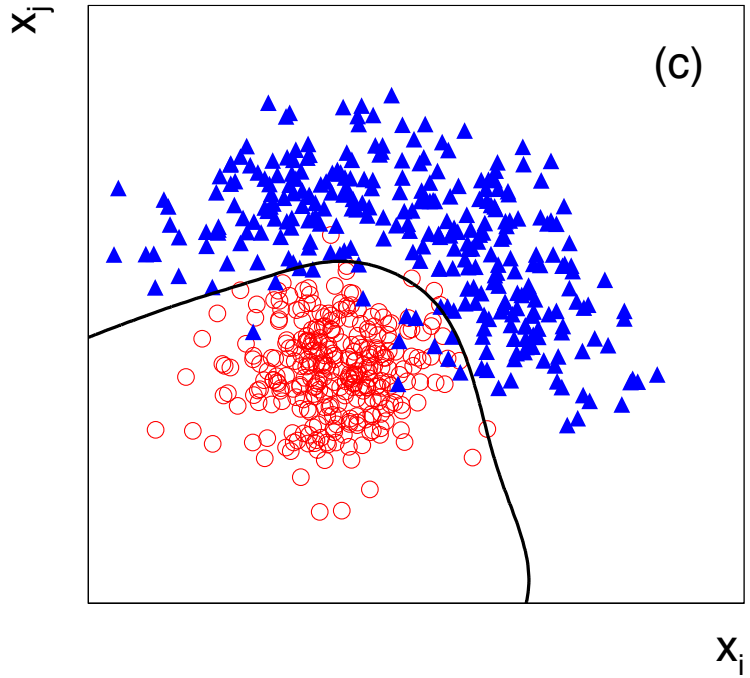
1. Systematics (and data vs MC)
2. 2 categories x n BDT score bins
3. Background estimated from data (embedded, anti tau, control region) and some MC
4. Weights include all corrections. Some negative weights (tt)
5. Potentially use any information from all 2012 data and MC events
6. Few variables fed in two BDT
7. Significance from complete fit with NP etc…
8. MVA with TMVA BDT

1. No systematics
2. No categories, one signal region
3. Straight use of ATLAS G4 MC
4. Weights only include normalisation and pythia weight. Neg. weight events rejected.
5. Only use variables and events preselected by the real analysis
6. All BDT variables + categorisation variables + primitives 3-vector
7. Significance from "regularised Asimov"
8. MVA "no-limit"

Simpler, but not too simple!

Significance

- Need to have one robust estimator of the quality of the classification algorithm
- Decided to use the well known (in HEP) "Asimov" formula (G. Cowan, K. Cranmer, E. Gross, and O. Vitells, "Asymptotic formulae for likelihood-based tests of new physics", *EPJCC*, vol. 71, pp. 1–19, 2011.) with regularization on top
  - $\sqrt{(2*((s+b')*\log(1+s/b')-s))}$ $\sim s/\sqrt{b'}$
  - with s and b'=b+10 normalised to 2012 data taking luminosity:
  - $s=\sum$(selected signal) weights_i
  - $b=\sum$(selected background) weights_i

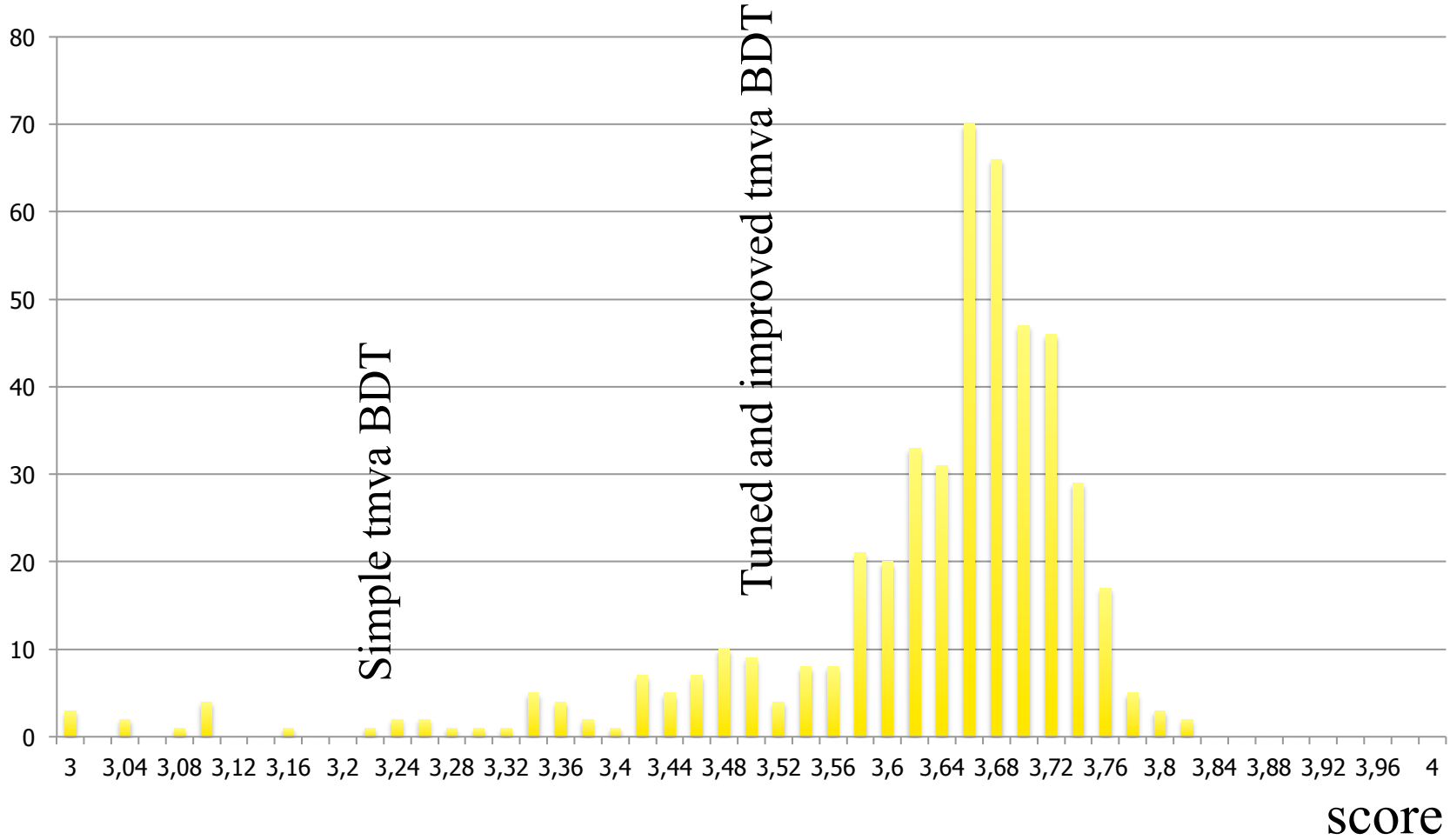


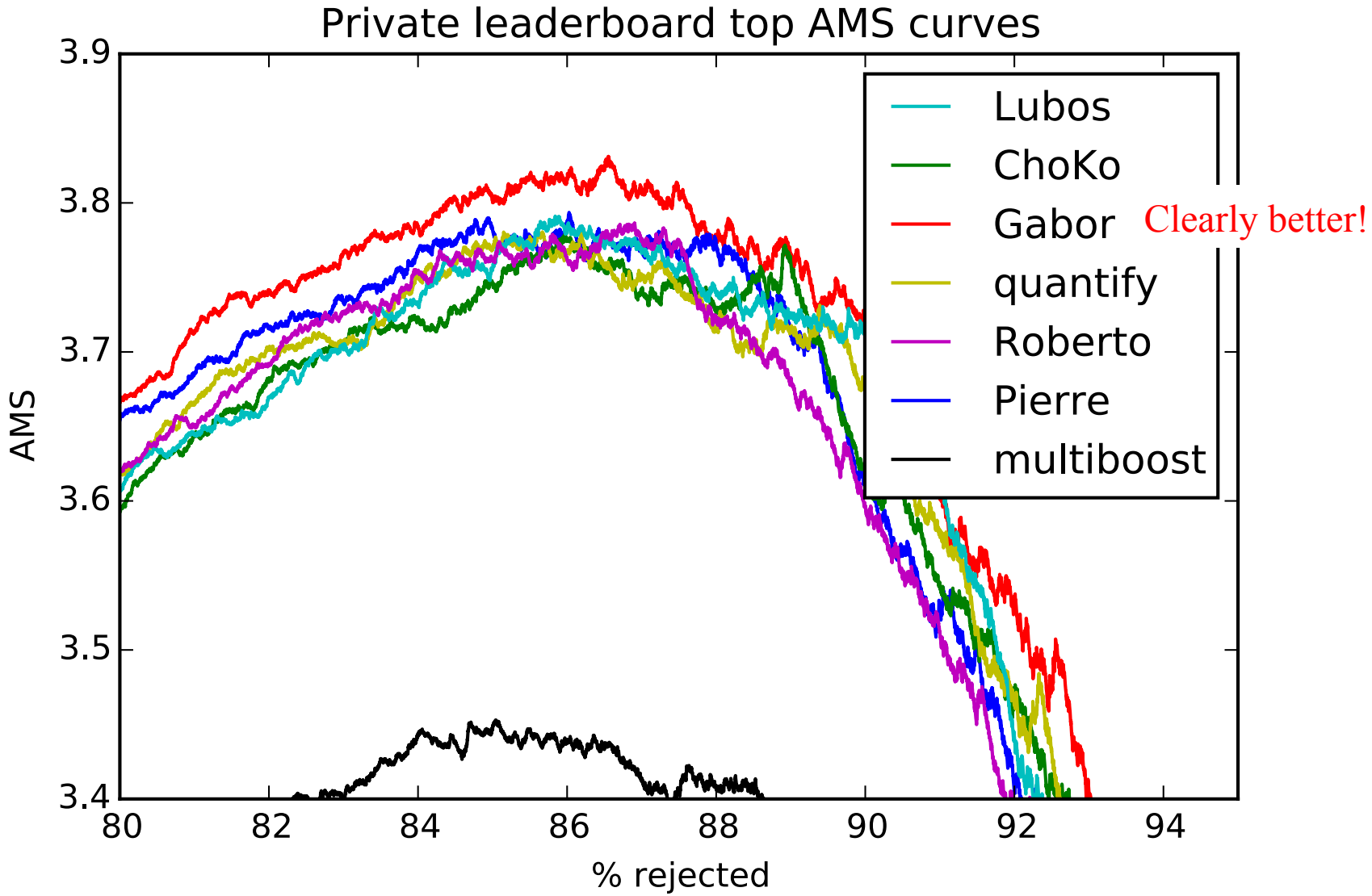
- Why b'=b+10 ("regularisation") : practical way to avoid large significance fluctuation when small phase space region with very few background events is chosen. Do not want to pick winners on their luck.
- Note that normalisation already included in the weights : no need to explain integrated luminosity and cross-section
- Glen Cowan has derived a new version of Asimov formula including a sigma_b from systematics or statistics ➔However in our case this leads to favour small region➔large variance.

What data did we release ?

- From ATLAS full sim Geant4 MC12 production
- 30 variables
- Signal is H➔tautau, Background a mixture of : Z, top, W
- Based on November 2013 ATLAS Htautau conf note ATLAS-CONF-2013-108
- Preselection for lep-had topology : single lepton trigger, one lepton identified, one hadronic tau identified
- ➔800.000 events:
  - 250.000 training data set
  - 550.000 test data set without label and weight
- Reproduces reasonably well (~20%) content of 3 highest sensitivity bins (x 2 categories) in conf note
- (some background and many correction factors deliberately omitted so that the sample cannot be used for physics, only for machine learning studies)

Best private scores




Private leaderboard top AMS curves
AMS
% rejected
3.9
3.8
3.7
3.6
3.5
3.4
80
82
84
86
88
90
92
94
Lubos
ChoKo
Gabor
quantify
Roberto
Pierre
multiboost
Clearly better!

Imputation

- In ML jargon, this is the handling of missing variables, a very hot topic
- In HiggsML, we provided leading and sub-leading jet 4-momenta, plus variable based on these (e.g. di-jet mass), but many events with just one or zero jet
  - In addition MMC would fail in a few percent of the cases
- No clear winning strategy among:
  - not doing anything special
  - Replace missing variables by average on other events
  - Separate training samples according to available variables