The ATLAS Higgs Machine Learning Challenge

CHEP, Okinawa, Japan
16 April 2015

Claire Adam-Bourdarios[1], Glen Cowan[2], Cécile Germain-Renaud[3], Isabelle Guyon[4], Balázs, Kégl[1], David Rousseau[1]

[1] Laboratoire de l'Accélérateur Linéaire, Orsay, France
[2] Royal Holloway, University of London, UK
[3] Laboratoire de Recherche en Informatique, Orsay, France
[4] Chalearn, California, USA

Outline

Multivariate analysis in High Energy Physics

The ATLAS Higgs Machine Learning Challenge

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

http://higgsml.lal.in2p3.fr/

C. Adam-Bourdarios et al., Learning to discover: the Higgs boson machine learning challenge, CERN Open Data Portal, DOI: 10.7483 OPENDATA.ATLAS.MQ5J.GHXA

The Problem

The Solutions

Future challenges

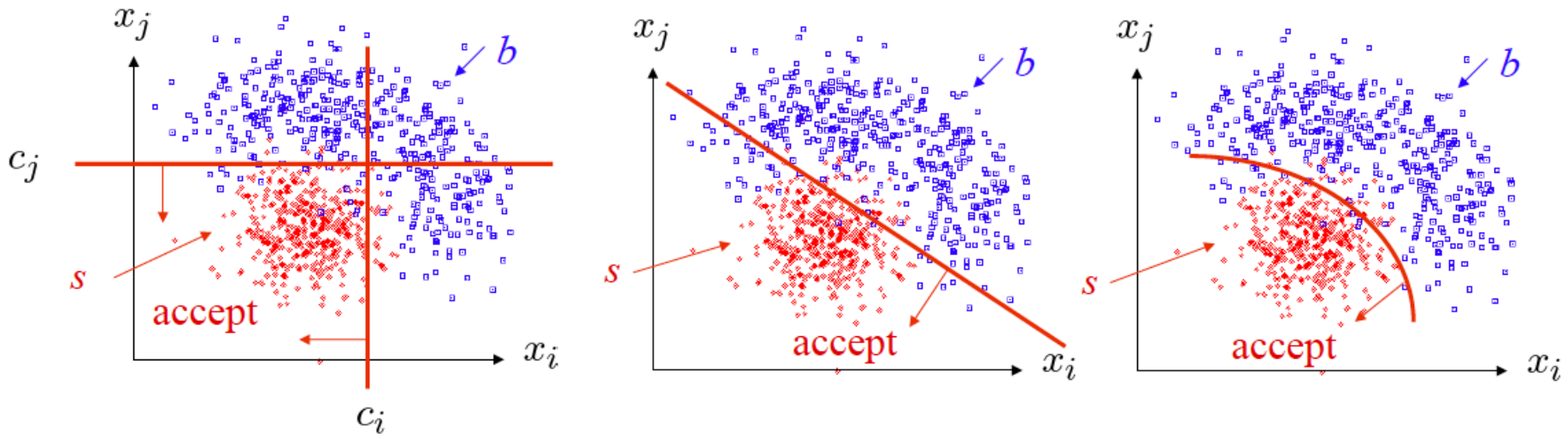
Prototype analysis in HEP

Each event yields a collection of numbers $\vec{x} = (x_1, \ldots, x_n)$

x_1 = number of muons, x_2 = p_t of jet, ...

$\vec{x}$ follows some n-dimensional joint pdf, which depends on the type of event produced, i.e., signal or background.

1) What kind of decision boundary best separates the two classes?



2) What is optimal test of hypothesis that event sample contains only background?

Machine Learning in HEP

Optimal analysis uses information from all (or in any case many) of the measured quantities → Multivariate Analysis (MVA)

Long history of cut-based analyses, followed by:

1990s	Fisher Discriminants, Neural Networks
Early 2000s	Boosted Decision Trees, Support Vector Machines

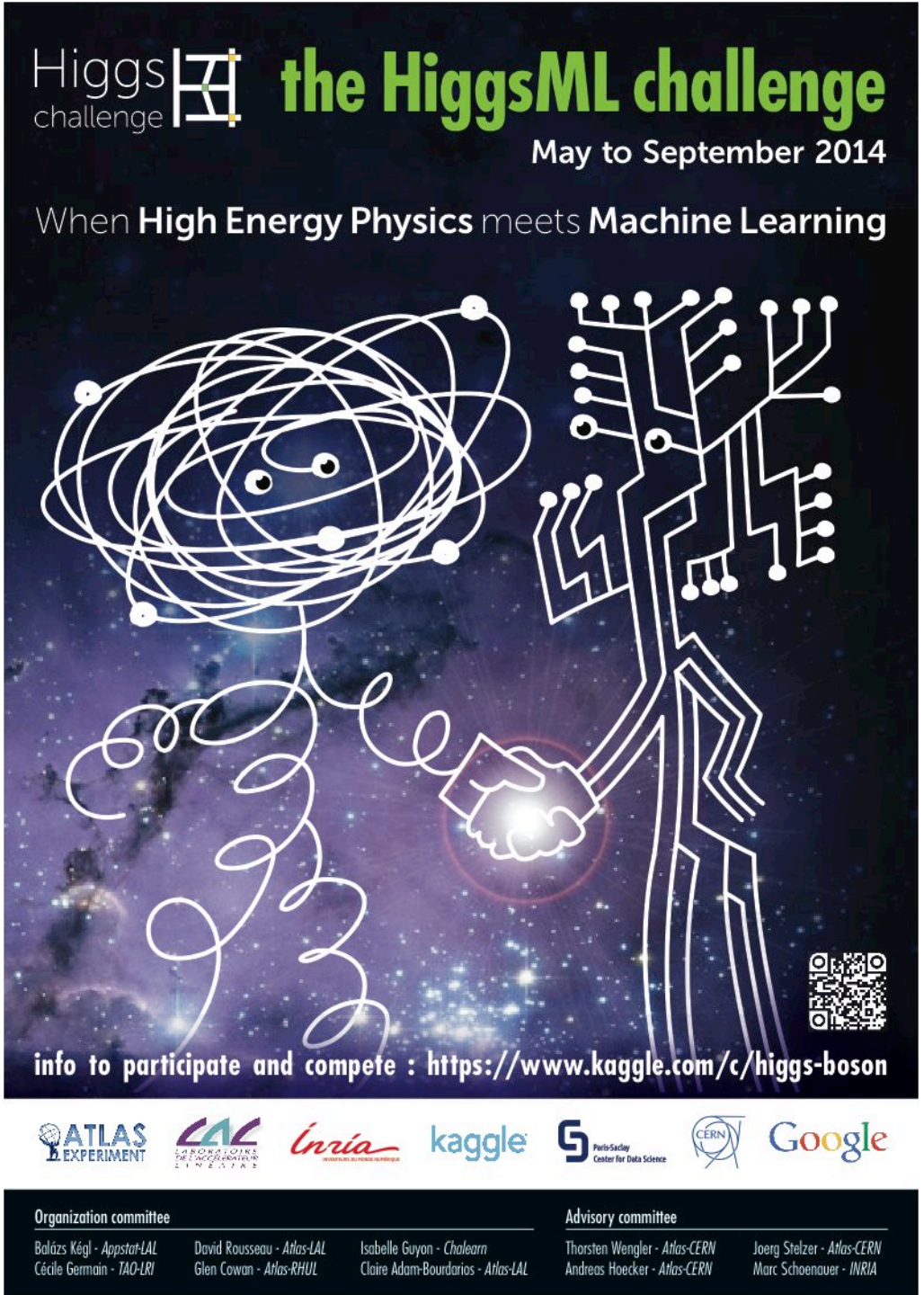
But much recent work in Machine Learning only slowly percolating into HEP (deep neural networks, random forests,...)

Therefore try to promote transmission of ideas from ML into HEP using a Data Challenge.

Challenge ?

- Challenges have become in the last 10 years a common way of working for the machine learning community
- Machine learning scientists are eager to test their algorithms on real life problems; more valuable (= publishable) than artificial problems
- Company or academics want to outsource a problem to machine learning scientist, but also geeks, etc. The company sets up a challenge like:
  - Netflix : predict movie preference from past movie selection
  - NASA/JPL mapping dark matter through (simulated) galaxy distortion
- Some companies makes a business from organising challenges: datascience.net, kaggle

The Higgs Machine Learning Challenge



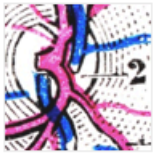
… in a nutshell

- 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 a 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 sufficiently close to reality to still be useful for us.
- Also try to foster long-term collaborations between HEP and ML.

The Host

Competition hosted by Kaggle, which provides platform for many data science challenges, e.g.,

www.kaggle.com

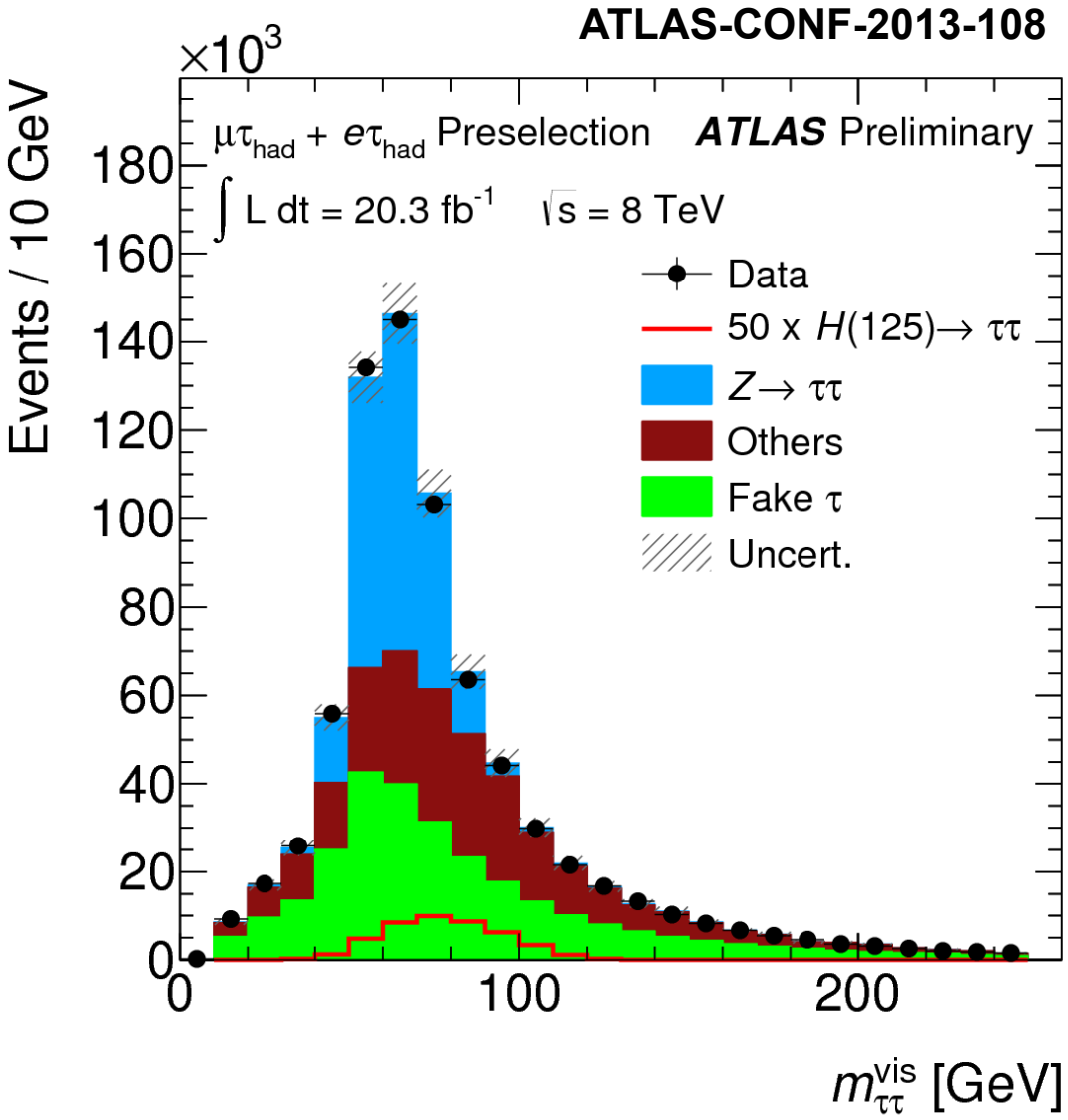
Competition Name	Reward	Teams	Deadline
 **Diabetic Retinopathy Detection** Identify signs of diabetic retinopathy in eye images	$100,000	228	3 months
 **Restaurant Revenue Prediction** Predict annual restaurant sales based on objective measurements	$30,000	1638	19 days
 **Microsoft Malware Classification Challenge (BIG 2015)** Classify malware into families based on file content and characteristics	$16,000	383	2.4 days
 **Otto Group Product Classification Challenge** Classify products into the correct category	$10,000	2247	33 days
 **How much did it rain?** Predict probabilistic distribution of hourly rain given polarimetric radar measurements	$500	225	30 days

Sponsors

This competition is brought to you by

Additional support from:

The signal process: Higgs → $\tau^+\tau^-$



4.1 σ evidence

Now superseded by ATLAS paper: *Evidence for the Higgs-boson Yukawa coupling to tau leptons with the ATLAS detector*, arXiv:1501.04943

ATLAS Monte Carlo Data

ASCII csv file, with mixture of Higgs to ττ signal and corresponding background, from official GEANT4 ATLAS simulation

250k training sample (event type s or b given)

100k public + 450k private test samples (event type hidden)

30 variables (derived and "primitive")

+ event weight (given for training sample only):

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
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

Objective Function

Typical Machine Learning goal is event classification; try to minimize e.g. classification error rate.

Goal in HEP search is to establish whether event sample contains only background; rejecting this hypothesis ≈ discovery of signal.

Often approach in HEP is to use distribution of MVA classifier. Simplest case, use classifier to define "search region" and count:

s = expected number of signal events (assuming it exists)

b = expected number of background events

Goal: Minimize Approximate Median Significance of discovery:

$$\mathrm{AMS} = \sqrt{2\left((s+b)\ln\left(1+\frac{s}{b}\right)-s\right)}$$

(Modified in the Challenge to prevent small search region where estimate of b may fluctuate very low: $b \to b + b_{\mathrm{reg}}$.)

Real analysis vs challenge

	Real analysis		challenge
1.	Systematics	1.	No systematics
2.	2 categories x n BDT score bins	2.	No categories, one signal region
3.	Background estimated from data (embedded, anti tau, control region) and some MC	3.	Straight use of ATLAS G4 MC
4.	Weights include all corrections. Some negative weights (tt)	4.	Weights only include normalisation and pythia weight. Neg. weight events rejected.
5.	Potentially use any information from all 2012 data and MC events	5.	Only use variables and events preselected by the real analysis
6.	Few variables fed in two BDT	6.	All BDT variables + categorisation variables + primitives 3-vector
7.	Significance from complete fit with NP etc…	7.	Significance from "regularised Asimov"
8.	MVA with TMVA BDT	8.	MVA "no-limit"

Simpler, but not too simple!

Participation & Outcome

Competition ran 12 May to 15 September 2014

Kaggle's most popular challenge ever!

1785 teams (1942 people) made submissions

(6517 people downloaded the data)

35772 solutions uploaded

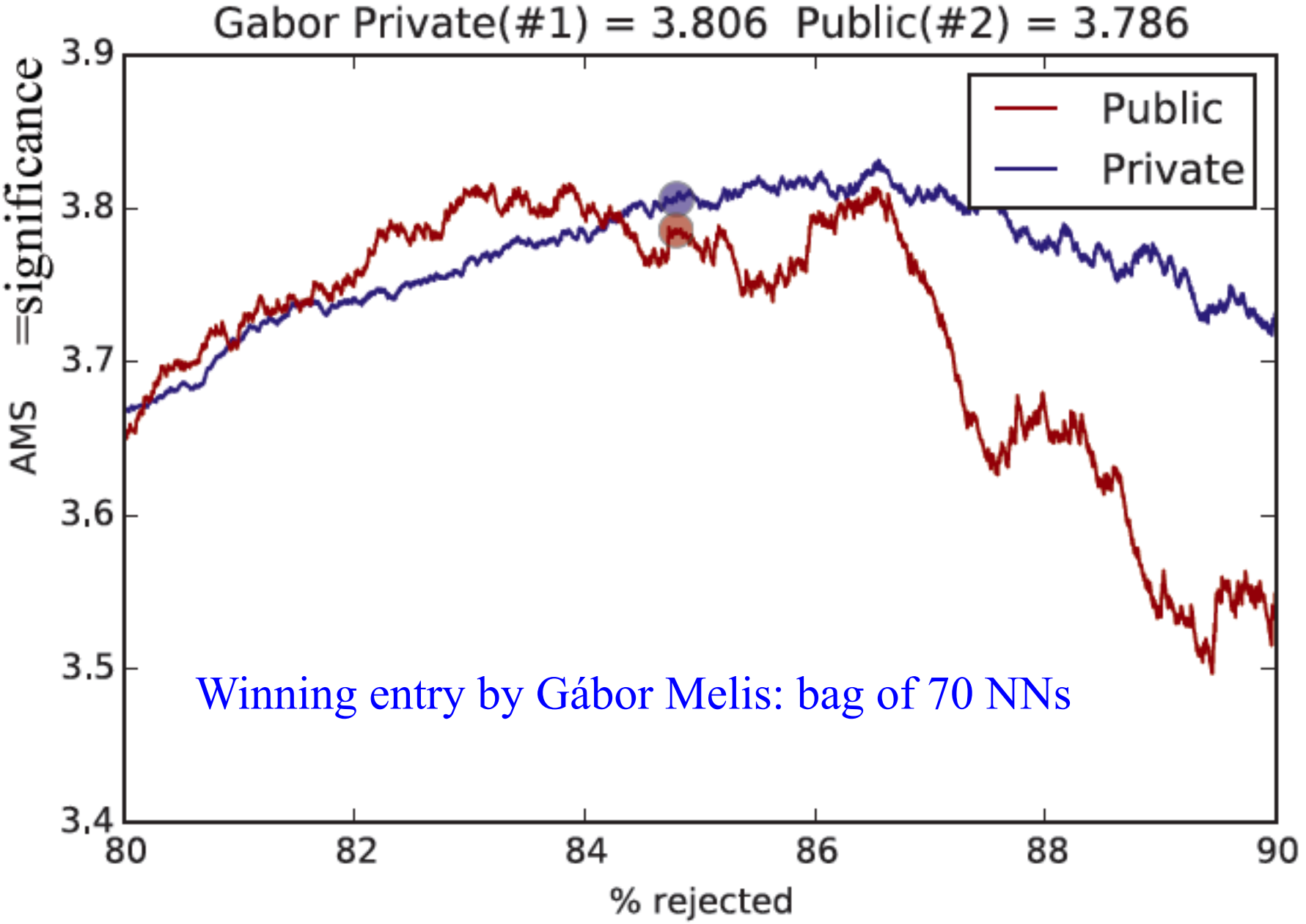
136 forum topics with 1100 posts

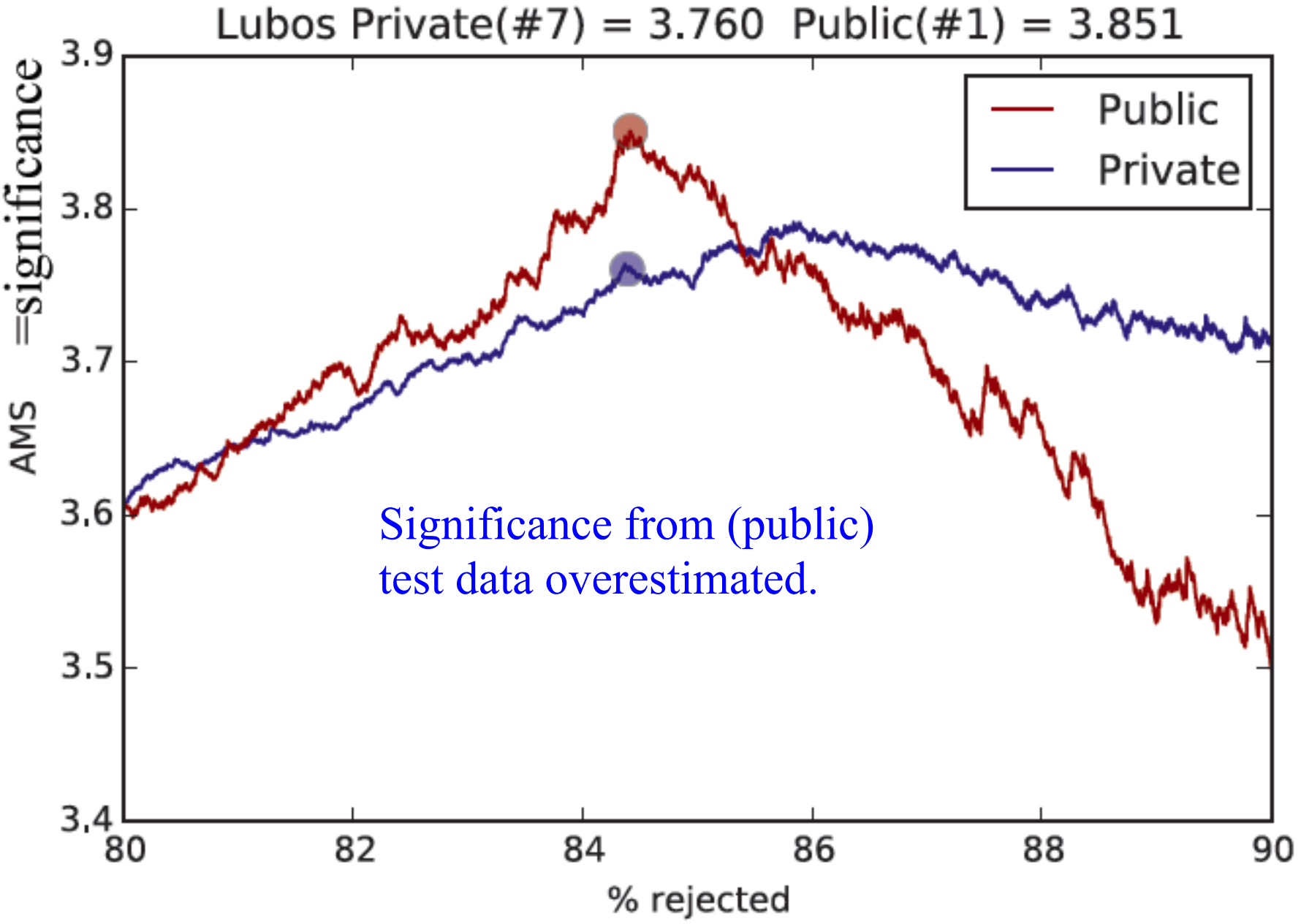
The winners:

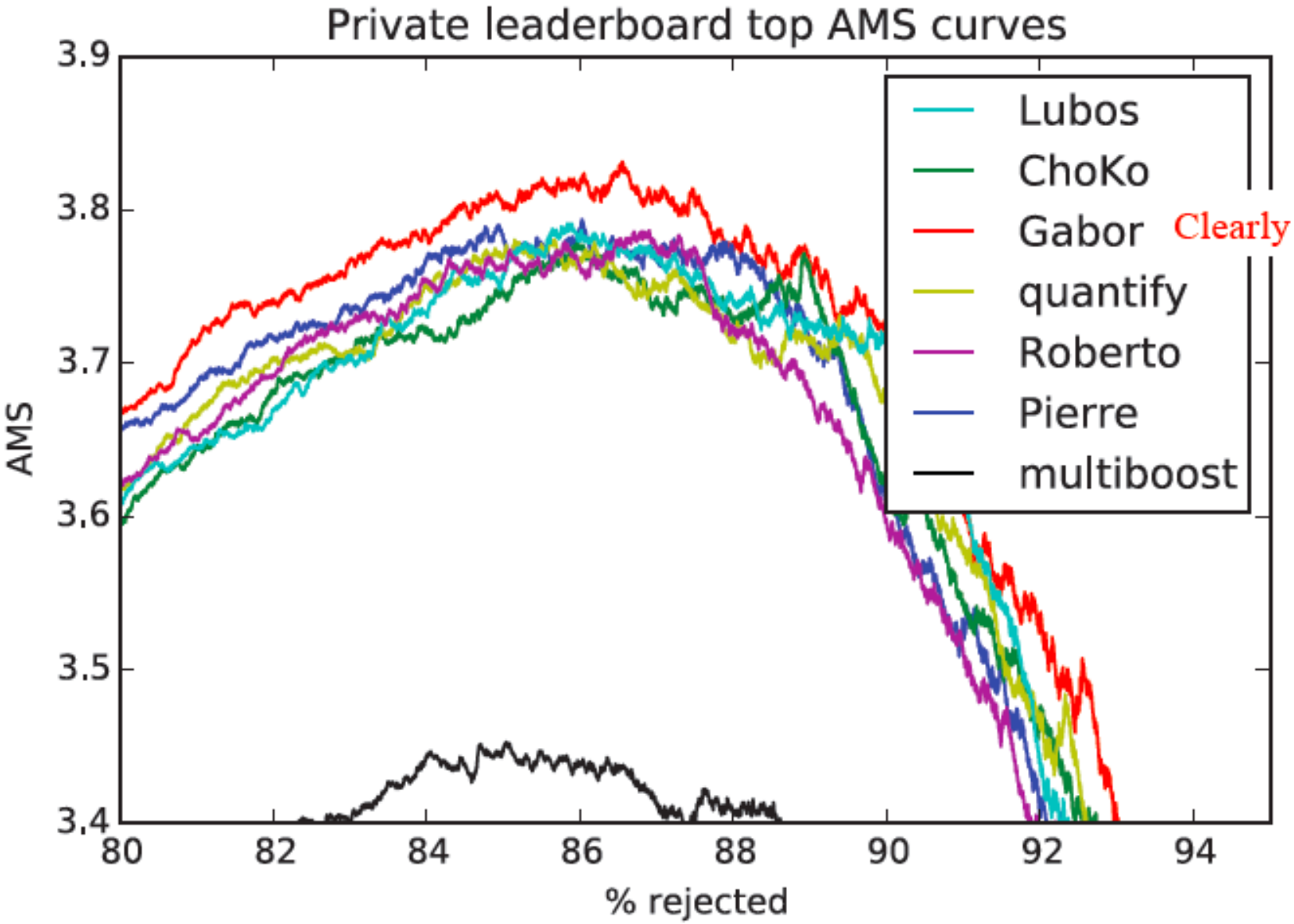
Gabor Melis (3.806)	$7000
Tim Salimans (3.789)	$4000
Pierre Courtiol (3.787)	$2000
Tianqi Chen and Tong He	"HEP meets ML" award

Final leaderboard

#	Δrank	Team Name ‡ model uploaded * in the money		Score	Entries	Last Submission UTC (Best – 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	Best physicist	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	TMVA expert, with TMVA improvements	3.49945	29	Mon, 15 Sep 2014 07:26:13 (-46.1h)

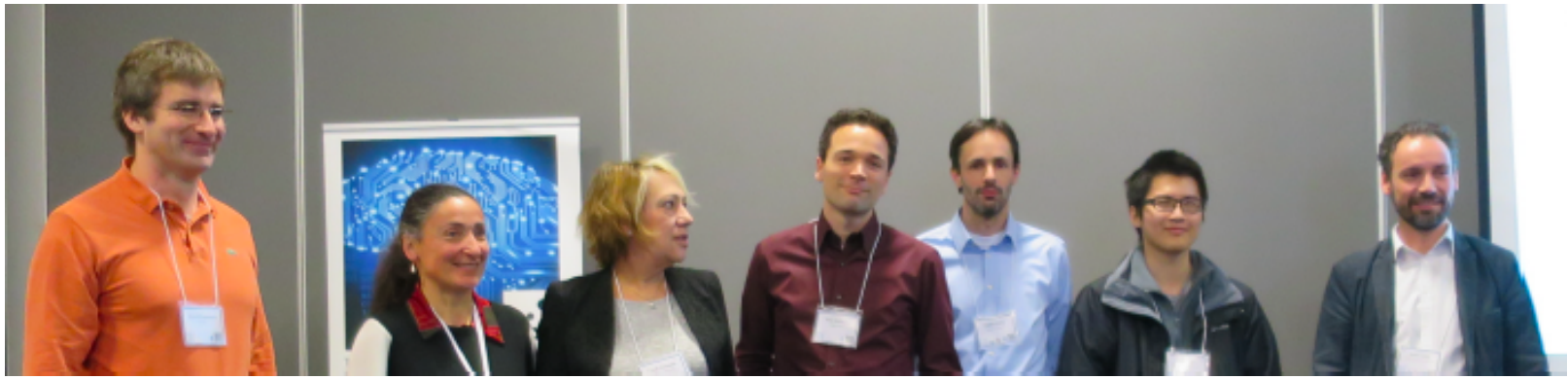

Gabor Private(#1) = 3.806 Public(#2) = 3.786
AMS =significance
3.9
3.8
3.7
3.6
3.5
3.4
80
82
84
86
88
90
% rejected
Public
Private
Winning entry by Gábor Melis: bag of 70 NNs


Lubos Private(#7) = 3.760 Public(#1) = 3.851
AMS =significance
3.9
3.8
3.7
3.6
3.5
3.4
80
82
84
86
88
90
% rejected
Public
Private
Significance from (public)
test data overestimated.


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 at the top!

What we've learned

- Very successful satellite workshop at NIPS in Dec 2014 @ Montreal: https://indico.lal.in2p3.fr/event/2632/



20% gain w.r.t. to untuned TMVA

Deep Neural nets

Ensemble methods (random forest, boosting)

Meta-ensembles of diverse models

careful cross-validation (250k training sample really small)

Complex software suites using routinely multithreading, GPU, etc…

Some techniques (e.g. meta-ensembles) too complex to be practical, and marginal gain, others appear practical and useful

Next steps

Re-importing into HEP all the ML developments (will take time!); e.g., discussions on-going with TMVA experts.

Dataset will remain on CERN Open Data Portal with citeable d.o.i.:
http://opendata.cern.ch/collection/ATLAS-Higgs-Challenge-2014

~800k events with full truth info

HEPML@NIPS contributions to be published in Proceedings of Machine Learning Research 42

Award winners at CERN (authors of XGboost and HEP meets ML winners Tianqi Chen and Tong He, and overall winner Gabor Melis)

Mini workshop 19th May 2015, 3 pm in CERN Auditorium,
http://cern.ch/higgsml-visit (will be webcast)

Related:

1) Data Science @ LHC workshop at CERN 9-13 Nov 2015
2) New mailing list: HEP-data-science@googlegroups.com

Extra slides

How did it work ?

- First idea in Sep 2012
- Challenge ran from May to September 2014
- People register to Kaggle web site hosted https://www.kaggle.com/c/higgs-boson . (additional info on https://higgsml.lal.in2p3.fr)
- Open to almost any one
  - Data scientist
  - HEP physicists
  - Students, geeks,
  - Except LAL-Orsay employees (for legal reasons)
- ...download training dataset (with label) with 250k events
- ...train their own algorithm to optimise the significance (à la s/sqrt(b))
- ...download test dataset (without labels) with 550k events
- ...upload their own classification
- The site automatically calculates significance. Public (100k events) and private (450k events) leader boards update instantly.
- Competition closes mid september 2014. People are asked to provide their code and methods. Best 1 2 3 from private leaderboard win 7k€ 4k€ 2k€

Funded by: Paris Saclay Center for Data Science, Google, INRIA

Cross validation

Common practice in HEP has been to divide the available MC data into a training sample and test sample:

Training sample used to train classifier
Test sample used to estimate its performance

But then only ~half of the expensive MC data is used for each task.

In k-fold cross validation, divide sample into k subsets or "folds" (say, $k = 10$), then:

Use all but the jth fold for training, jth fold for testing
→ get performance measure ε_j.

Repeat for all k folds, average resulting ε_j and use this to optimize classifier and estimate performance.

Train final classifier using all of the available events.

Many flavours, see e.g. Cross Validation wikipedia page
http://en.wikipedia.org/wiki/Cross-validation_(statistics)

Committees

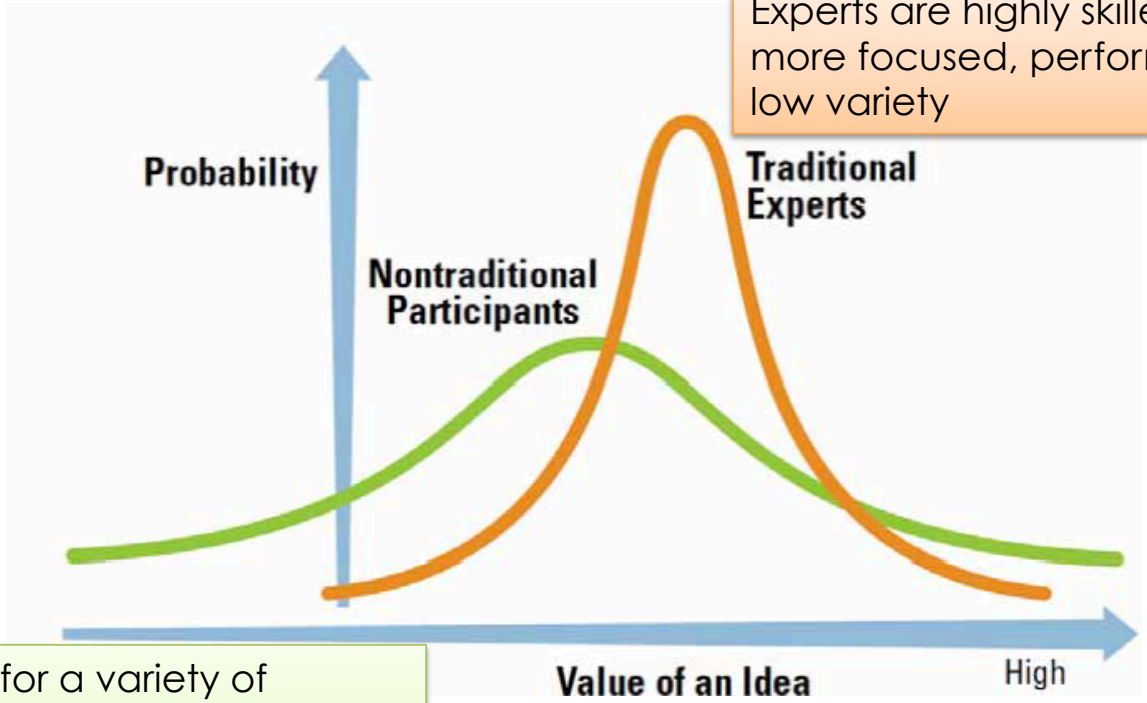
- Organization committee:
  - ATLAS:
    - David Rousseau ATLAS-LAL
    - Claire Adam-Bourdarios ATLAS-LAL (outreach, legal matters)
    - Glen Cowan ATLAS-RHUL (statistics)
  - Machine Learning:
    - Balázs, Kégl Appstat-LAL
    - Cécile Germain TAO-LRI
    - Isabelle Guyon Chalearn (challenges organization)
- Advisory committee:
  - Andreas Hoecker ATLAS-CERN (PC,TMVA)
  - Joerg Stelzer ATLAS-CERN (TMVA)
  - Thorsten Wengler ATLAS-CERN (ATLAS management)
  - Marc Schoenauer INRIA (French computer science institute)

Why challenges work

MOTIVATION OF ORGANIZING CONTESTS: EXTREME VALUE

Courtesy : Lakhani 2014

Experts are highly skilled, trained - > more focused, performed solution, low variety



OI is suitable for a variety of nonconventional surprising ideas that are « far » from traditional expertise - > high volatility

Not just ML, but a general trend:
Open Innovation

From domain to challenge and back