Higgs Machine Learning Challenge

Claire Adam-Bourdarios, Glen Cowan, Cécile Germain, Isabelle Guyon, Balazs Kegl, David Rousseau

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

HiggsML visits CERN, 19th May 2015

Outline

- Machine Learning, Challenges …
- The Higgs Machine Learning challenge
- What's next

Machine Learning and HEP

- Machine Learning is more or less what we call Multi Variate Analysis
- Neural Nets used somewhat in the 90'ies (e.g. LEP), Fisher Discriminant as well
- BDT (Adaboost) invented in 97
- MVA techniques have been used extensively at D0/CDF (mostly BDT, but not only) in the 00'ies
- ATLAS/CMS less eager to adopt MVA at LHC starts for some good reasons:
  - Need to understand well the input variables first
  - Still a lot to gain by improving input variables
  - Systematics more difficult to evaluate
  - Collected luminosity was increasing fast
- But lot of work recently with MVA techniques
  - Competition
  - Best use of available data
- Meanwhile Neural Net re-appear in their "deep" incarnation

Machine Learning and HEP (2)

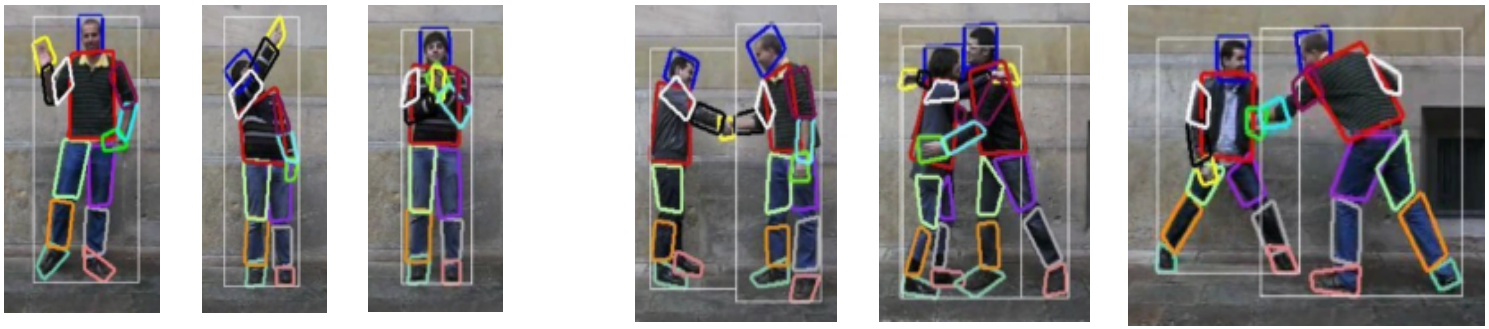
- However:
  - TMVA, within Root, has been instrumental in popularising MVA technique within HEP
  - Most HEP people using TMVA, most people using BDT in TMVA
  - Although getting a reasonable answer from TMVA is quick and easy, it takes time to really become an expert with e.g. BDT
  - People are focussing on the choices of input variables and the evaluation of systematics (which of course are excellent things to do)
- Not much work within LHC experiments on studying possible better MVA techniques, for which you need the software and the know-how
- Enormous development of Machine Learning in the outside world in the last 10 years ("Big Data", "Data Science" buzz words, even "Artificial Intelligence" is back)

Machine Learning 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(=publisheable) than artificial problems
- Company or academics want to outsource a problem to machine learning scientist, but also geeks etc.
- Some companies make a business from organising challenges: kaggle
- A few recent examples now…

Looking at People (2012-14)

Actions **Interactions**



Wave Point Clap Shake Hands Hug Fight

RGB Depth User mask Skeletal model

http://chalearn.org/

CodaLab

Microsoft Research

HuPBA

UAB Universitat Autònoma de Barcelona

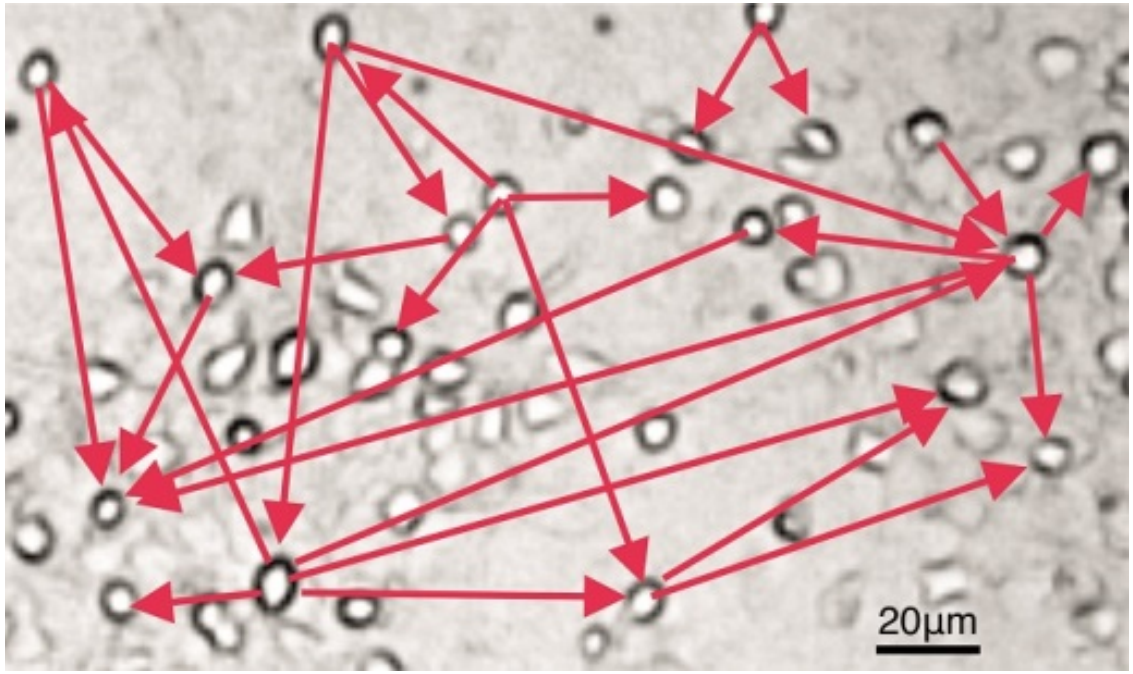
UOC Universitat Oberta de Catalunya www.uoc.edu

CVC Computer Vision Center

INAOE

Universitat de Barcelona

Neural connectomics (2015)

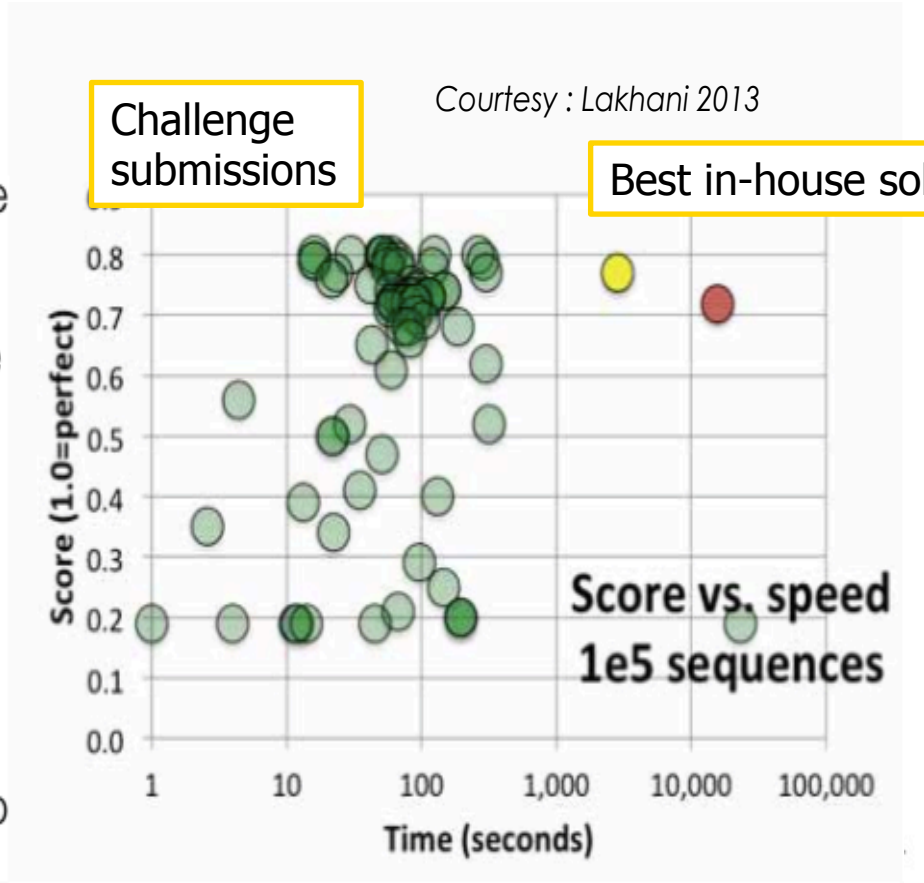


http://chalearn.org/

Genomics

Harvard Medical School Contest for Biology Big Data Problem in Genomics
Two week long competition - $2000 prize pot x 3 on TopCoder.com

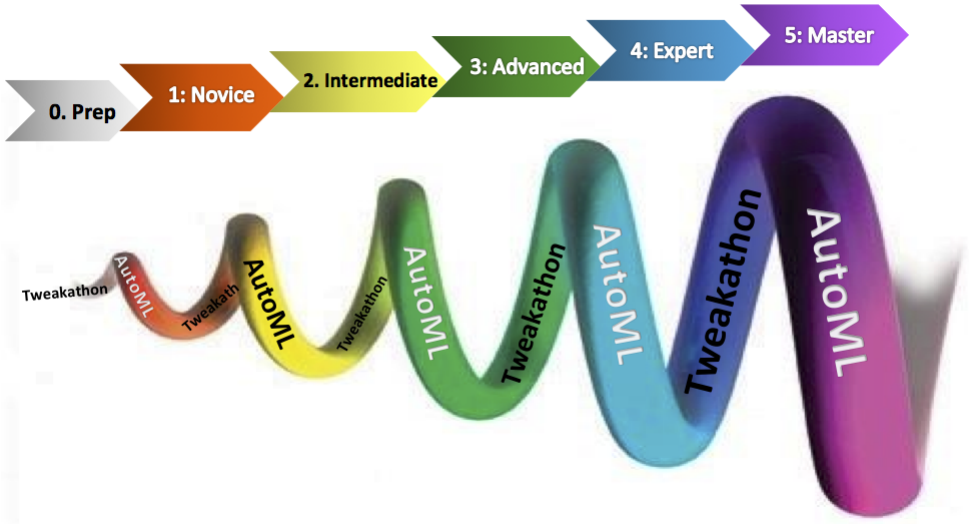
- 122 coders submitted 654 submissions
- 34 coders exceeded state of the art by 10^2 - 10^5
- 89 different approaches to solve problem identified
- Winners from Russia, France, Egypt, Belgium & US
- Annotate 10 million sequences in < 3 mins; Quarter billion sequences in ~ 1 hour on laptop



Olga Kokshagina 2015

NEW: AutoML challenge (2015)

Fully automatic machine learning without ANY human intervention



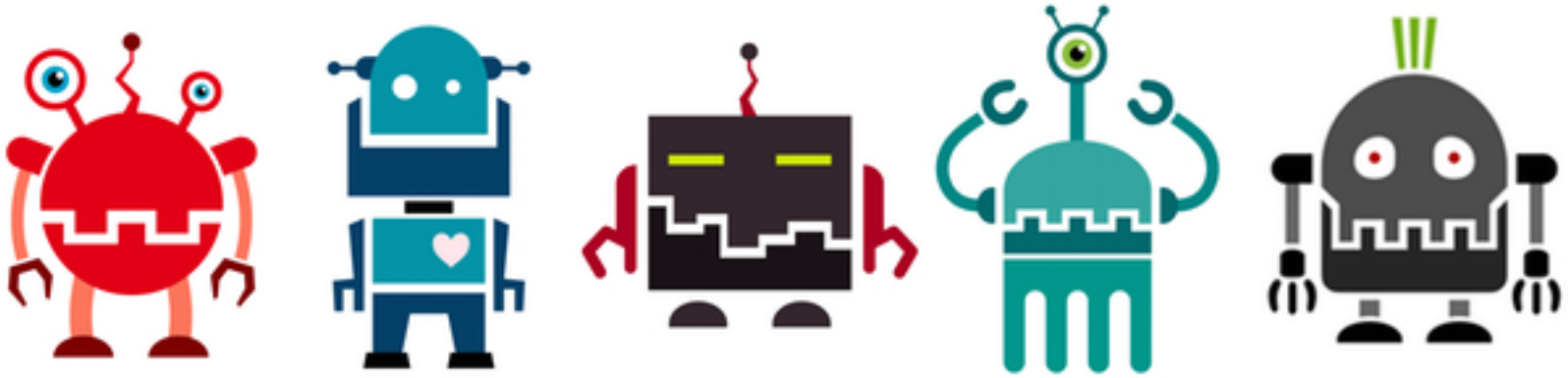
http://codalab.org/AutoML

December 2014 – May 2015

Some problems in the outside world look like what we are doing in HEP

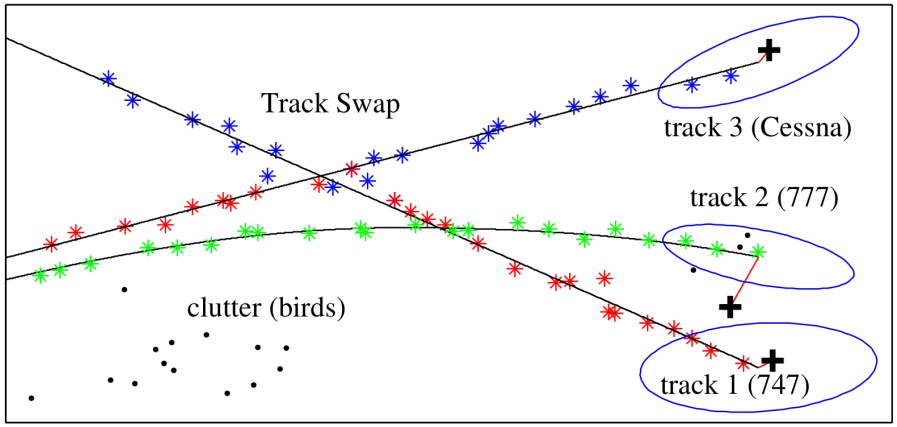
Classification (obviously)

- Facebook challenge on Kaggle (https://www.kaggle.com/c/facebook-recruiting-iv-human-or-bot, ends 8th June 2015)
  - Predict if an online bid is made by a machine or a human

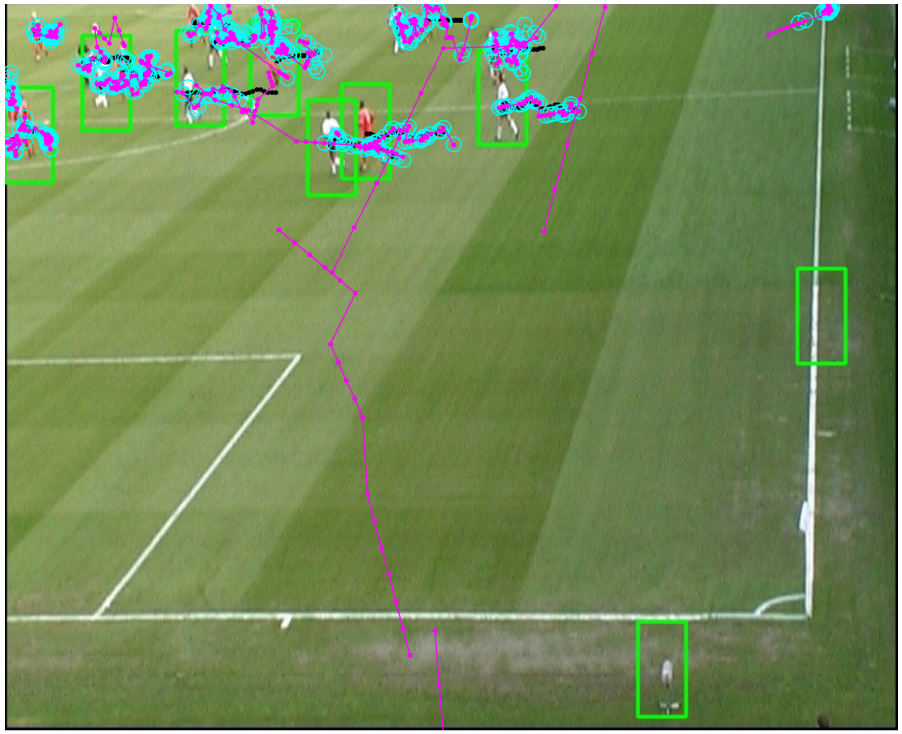


Something like tracking

- one example among many from NIPS 2014 :
  http://papers.nips.cc/paper/5572-a-complete-variational-tracker.pdf

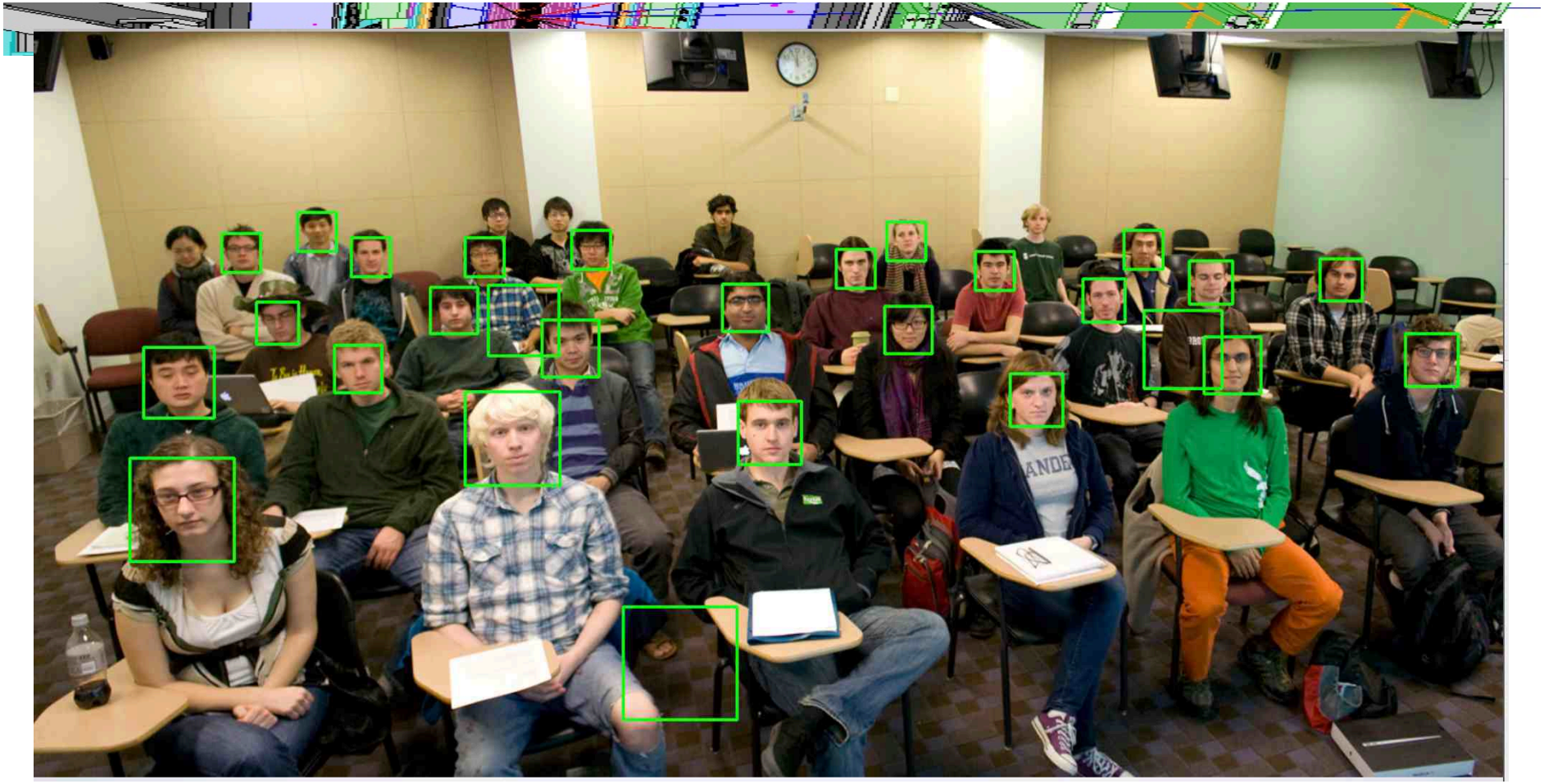


Also:



- Note that these are real-time applications, with CPU constraints
- Worry about efficiency, "track swap"

Something like trigger



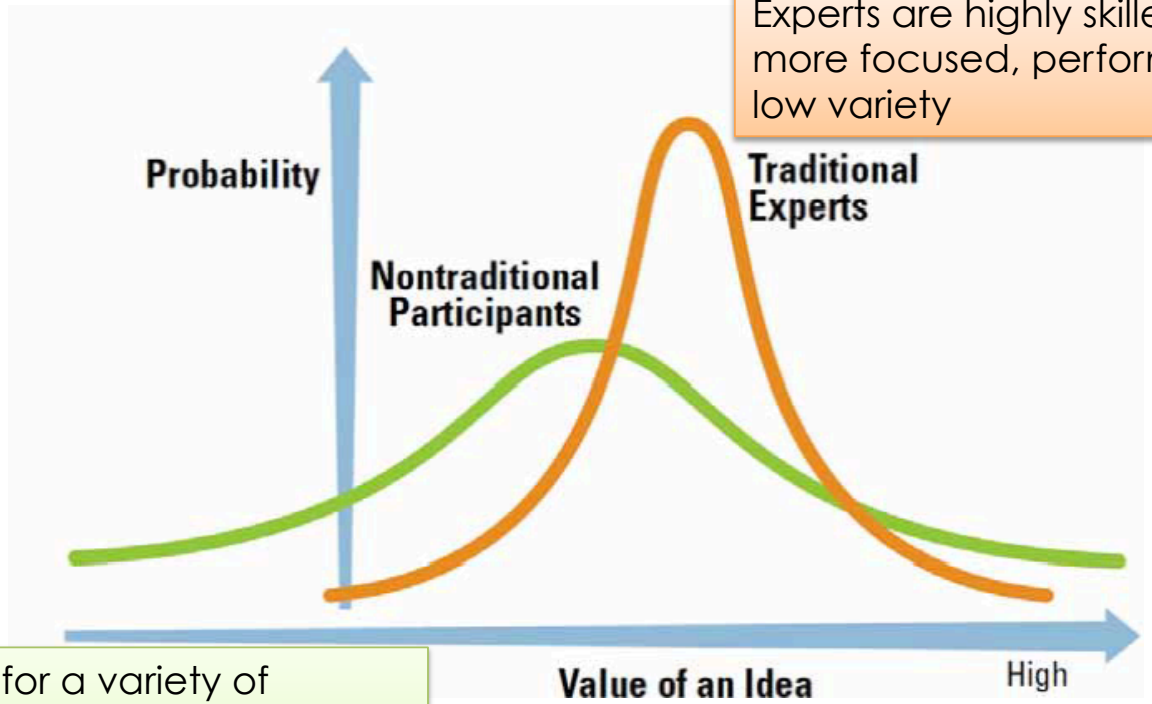
Real-time face recognition : efficiency, fake, CPU time…

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

Domain e.g. HEP

Challenge organisation

Challenge

Problem

simplify

Problem

Domain experts solve the domain problem

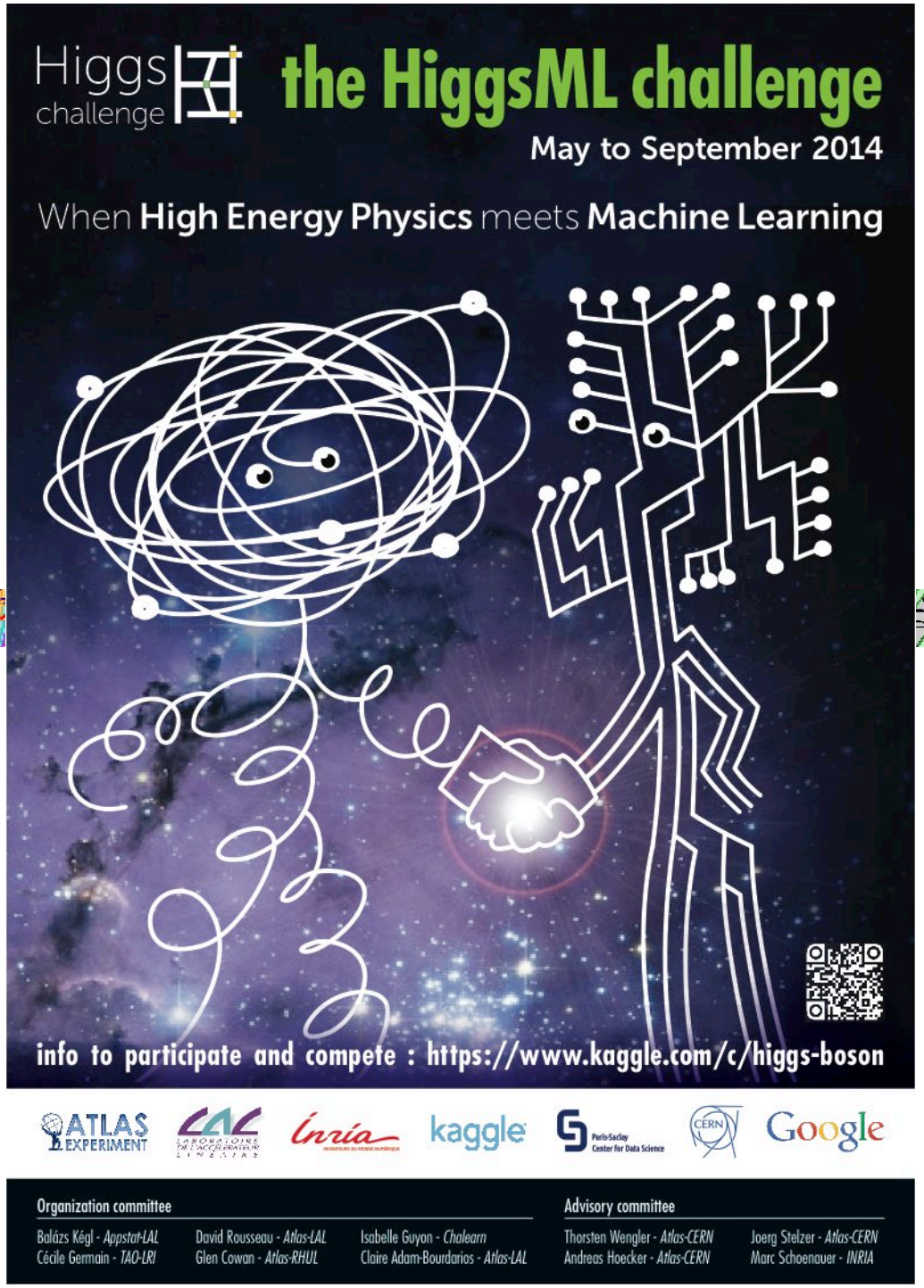
The crowd solves the challenge problem

Solution

reimport

Solution

Higgs Machine Learning Challenge



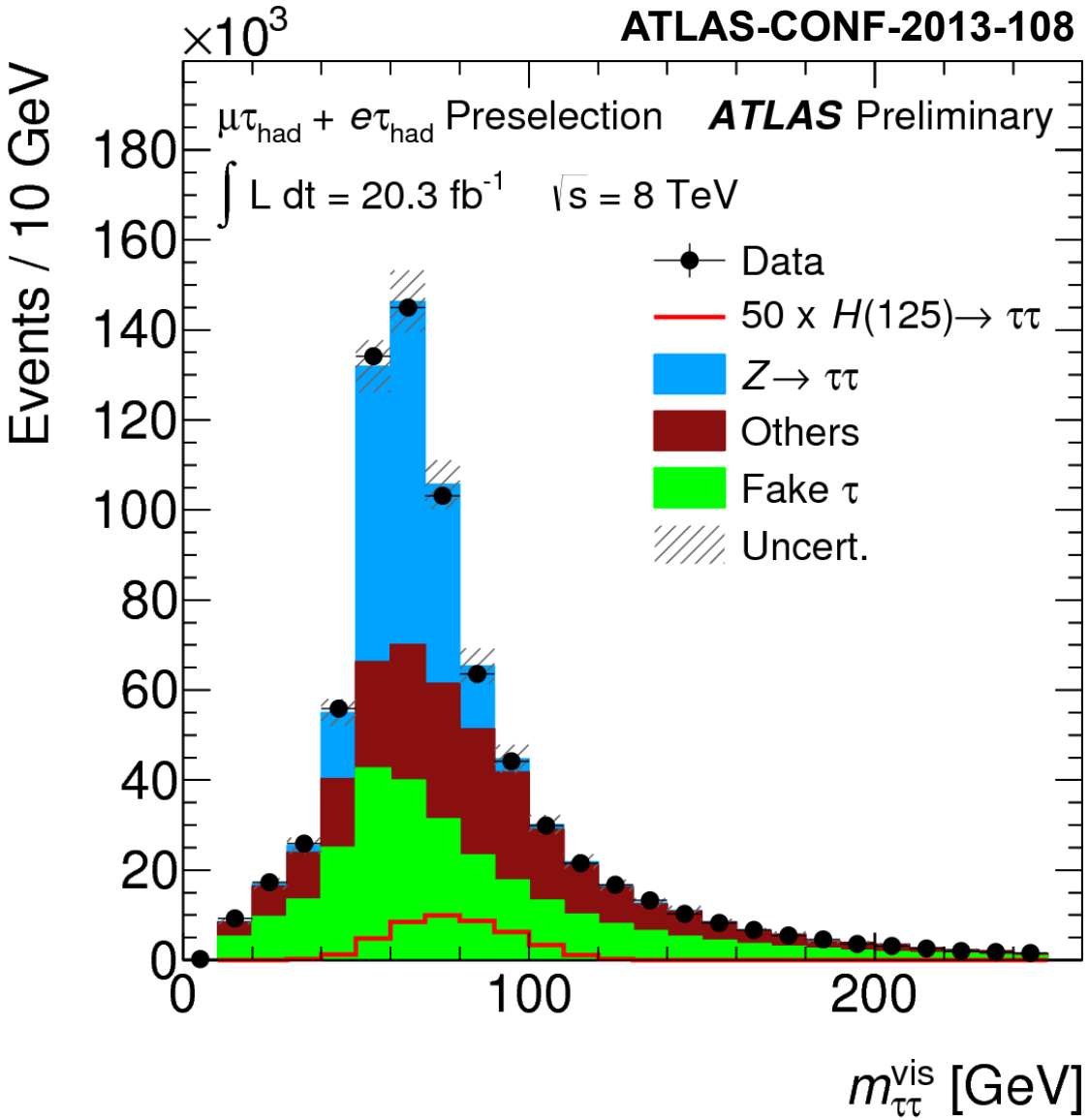
… in a nutshell

- 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 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

Committees

- Organization committee:
  - ATLAS
    - Claire Adam-Bourdarios : ATLAS-LAL
    - Glen Cowan : ATLAS-RHUL
    - David Rousseau : ATLAS-LAL
  - Machine Learning
    - Cécile Germain : TAO-LRI
    - Isabelle Guyon : Chalearn (challenges organisation)
    - Balazs Kegl : Appstat-LAL
- Advisory committee:
  - Andreas Hoecker : ATLAS-CERN
  - Marc Schoenauer : INRIA
  - Joerg Stelzer : ATLAS-CERN
  - Thorsten Wengler : ATLAS-CERN

H tautau



4.1 σ evidence

(now superseded by paper **arXiv:1501.049** tbp in JHEP. Analysis methods essentially unchanged.)

Dataset

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:

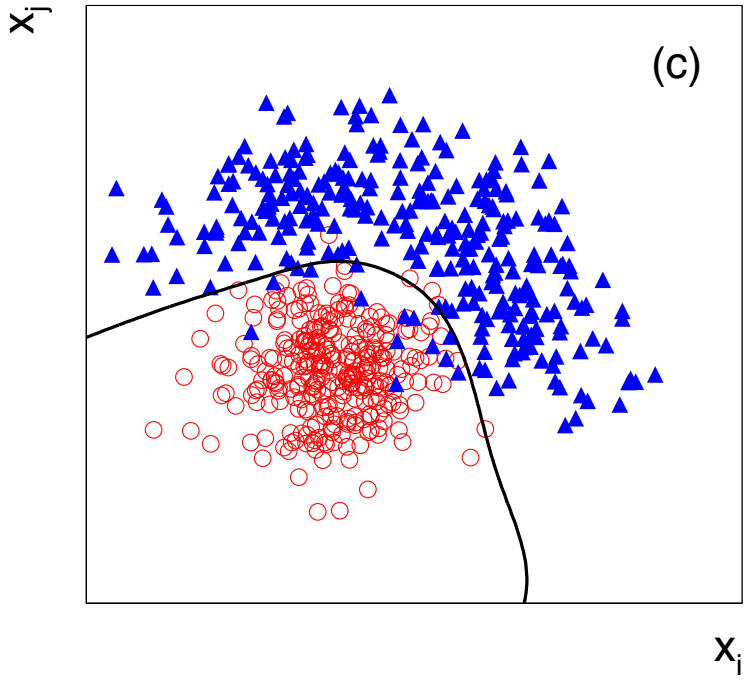
DER_mass_MMC
DER_mass_transverse_met_lep
DER_mass_vis
DER_pt_h
DER_deltaeta_jet_jet
DER_mass_jet_jet
DER_prodeta_jet_jet
} VBF signature
DER_deltar_tau_lep
DER_pt_tot
DER_sum_pt
DER_pt_ratio_lep_tau
DER_met_phi_centrality
DER_lep_eta_centrality

Primitive 3-vectors allowing to compute the 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
} VBF signature
PRI_jet_all_pt

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.

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. (Only the public is visible)
- Competition closed mid september 2014. Private leaderboard is disclosed. People are asked to provide their code and methods. Best 1 2 3 win 7k$ 4k$ 2k$
- In addition, the potentially most interesting one gets the "HEP meets ML award"

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

From domain to challenge and back

Domain e.g. HEP

Challenge organisation

Challenge

18 months

Problem

simplify

Problem

Domain experts solve the domain problem

The crowd solves the challenge problem

Solution

reimport

Solution

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!

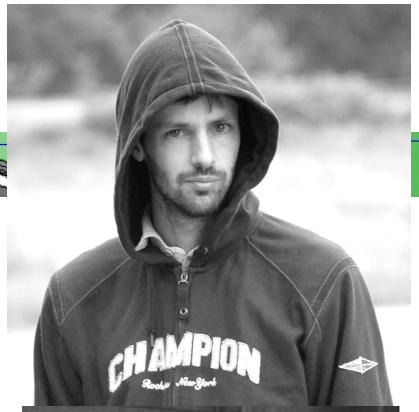
Participation

- 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 a few weeks ago)
  - 35772 solutions uploaded
- 136 forum topics with 1100 posts

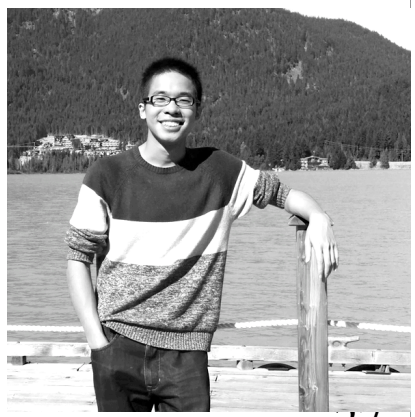
Final leaderboard

#	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		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		

Meet the winners



?

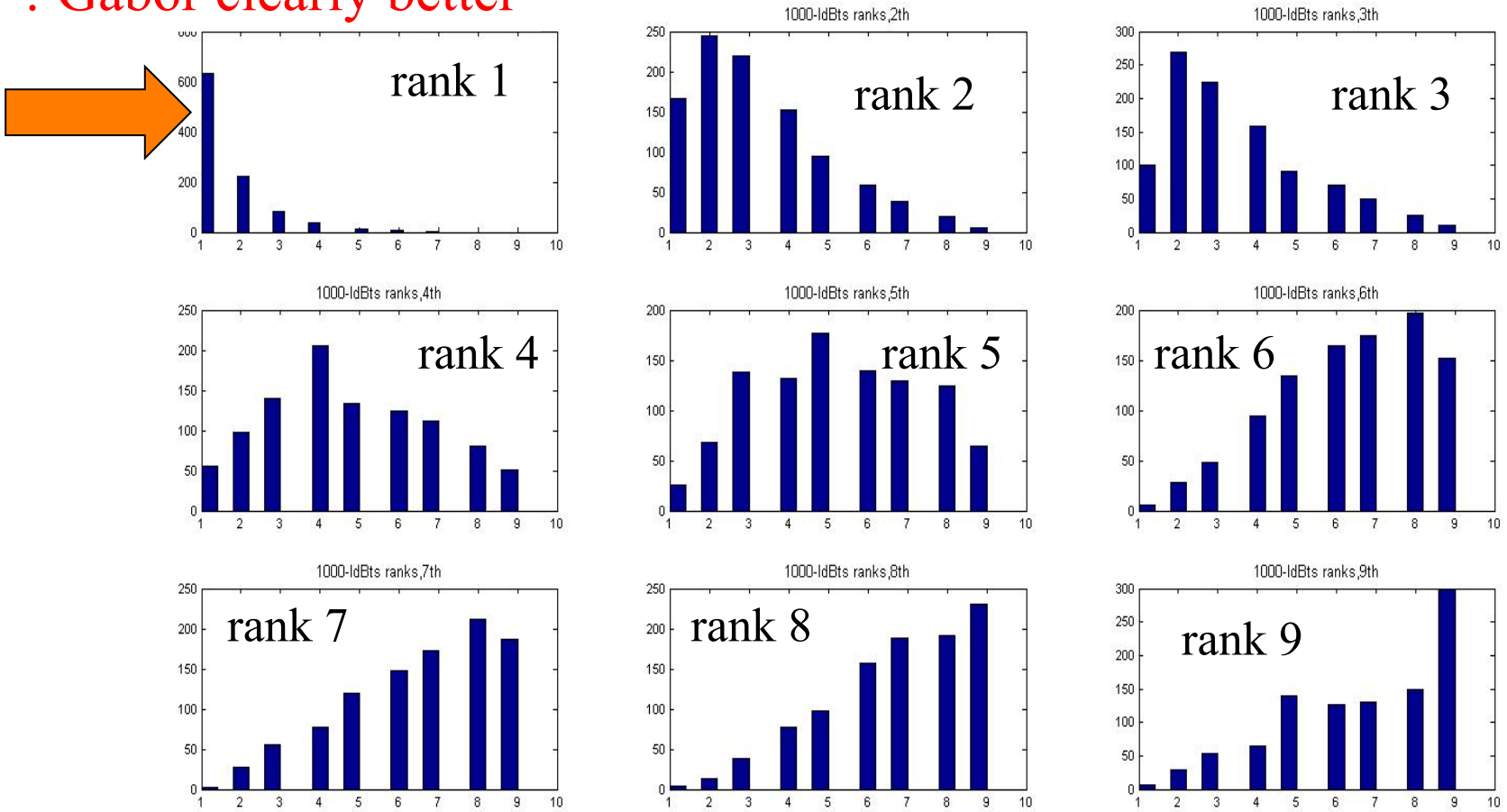


- See http://atlas.ch/news/2014/machine-learning-wins-the-higgs-challenge.html
- 1 : Gabor Melis (Hungary) sw developer and consultant : wins 7000$. 3rd talk.
- 2 : Tim Salimans (Neitherland) data science consultant: wins 4000$
- 3 : Pierre Courtiol (nhlx5haze) (France) ? : wins 2000$
- HEP meets ML award: (team crowwork), Tianqi Chen (U of Washington PhD student in Data Science) and Tong He (graduate student Data Science SFU). Provided XGBoost used by many participants. Win a free trip and visit to CERN in 2015, this is today! 2nd talk.

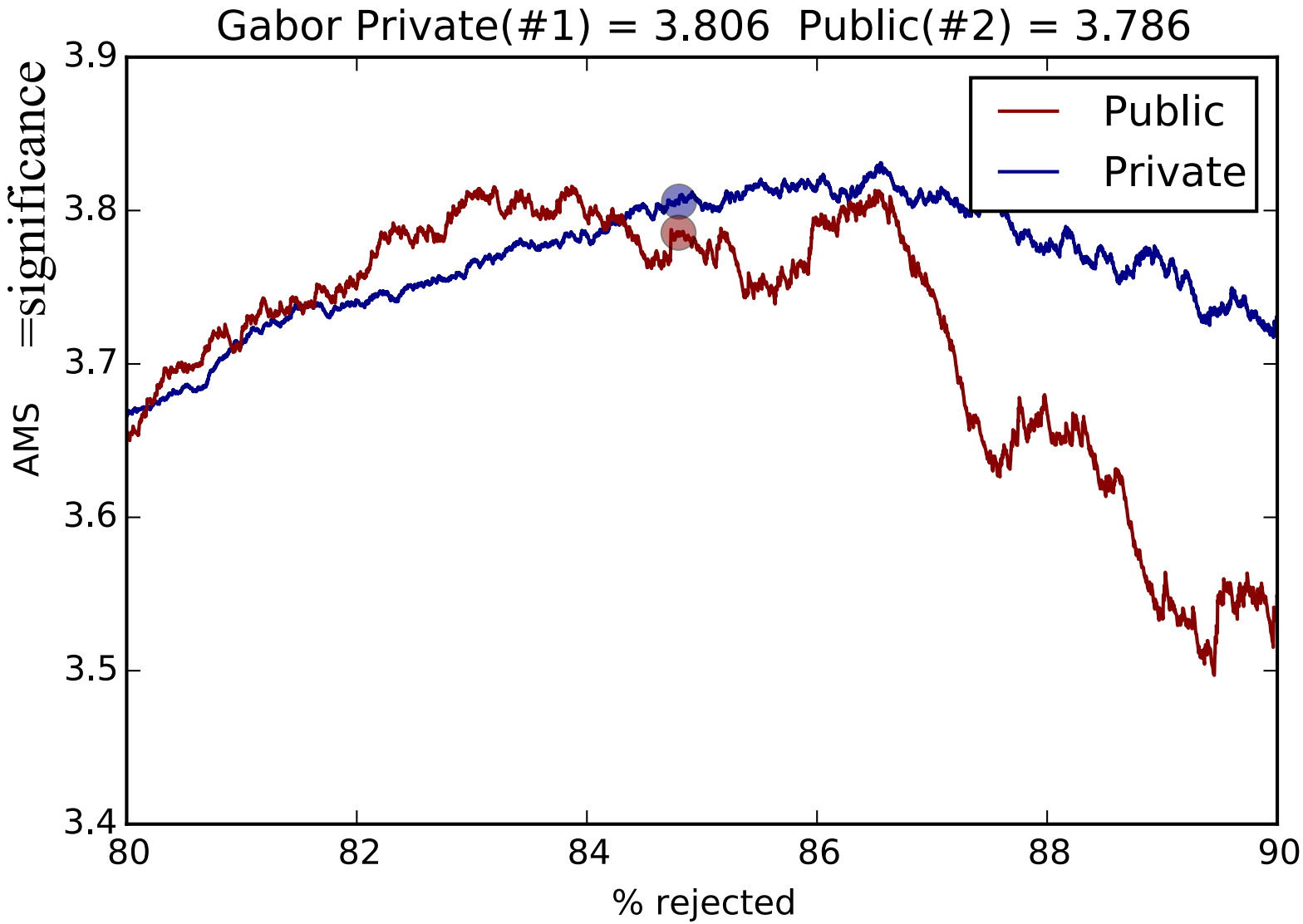
Rank distribution after bootstrap

Distribution of rank of participant of rank i after 1000 bootstraps of the test sample.

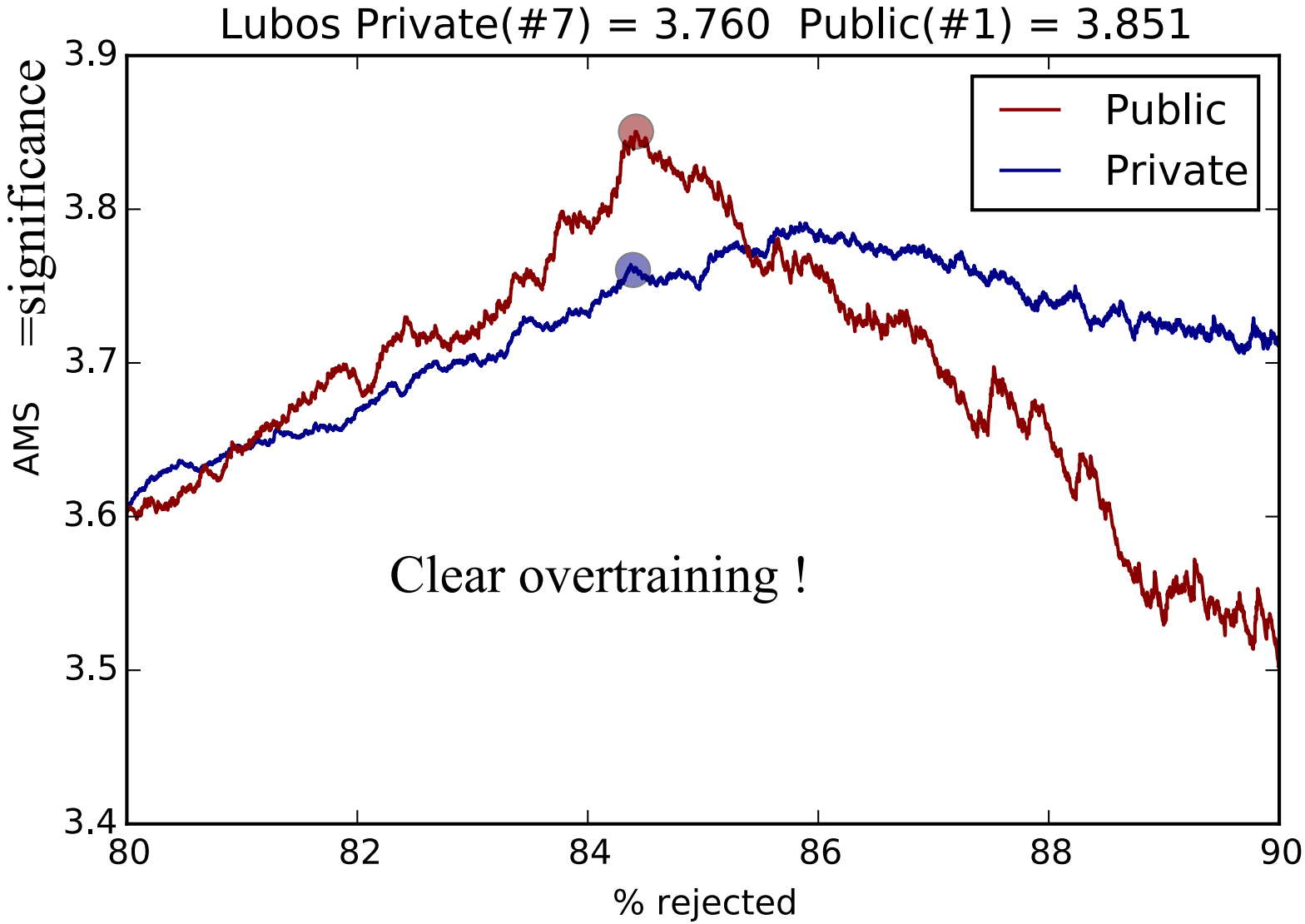
! Gabor clearly better



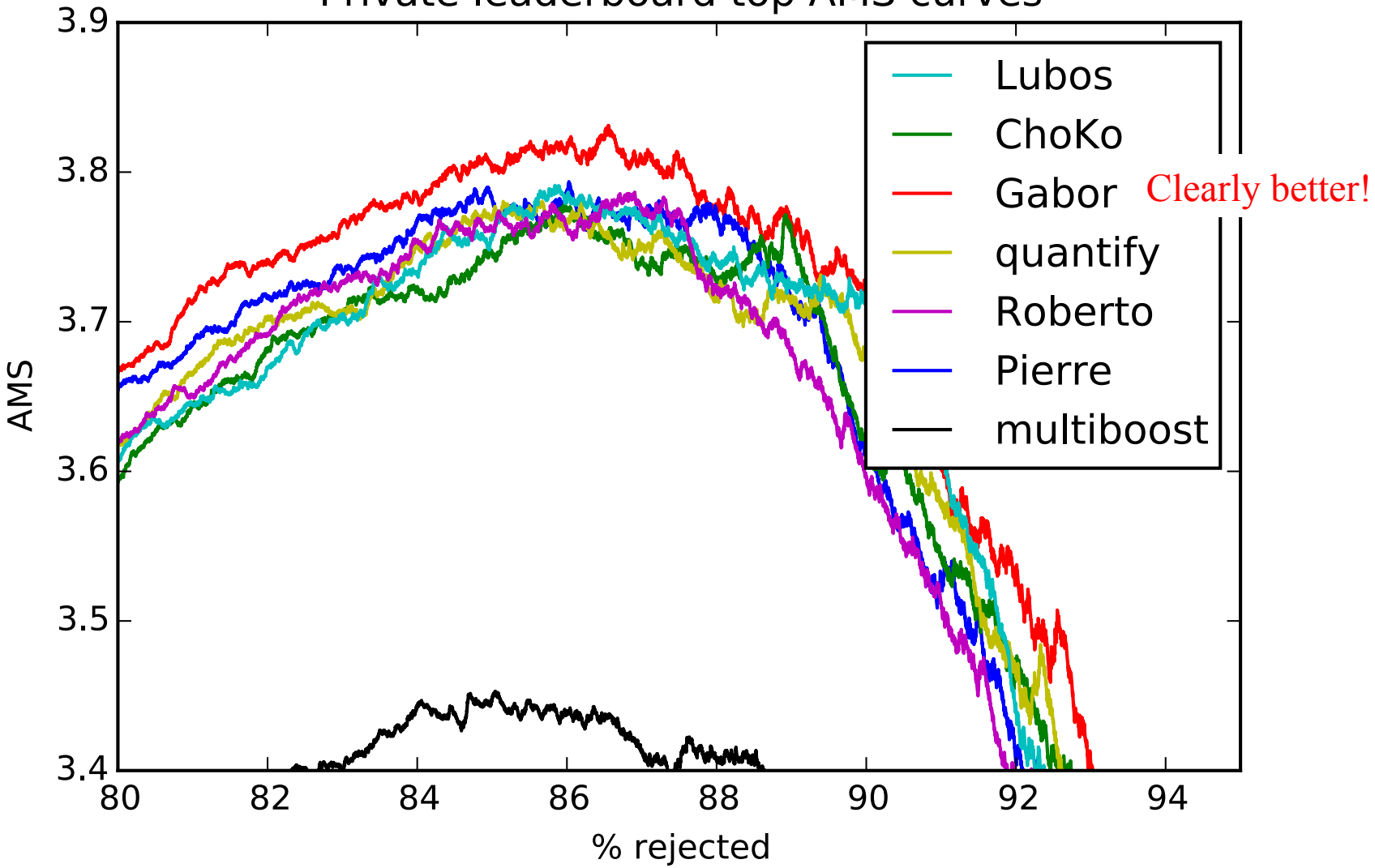
Private score vs public: Gabor



Private score vs public: Lubos

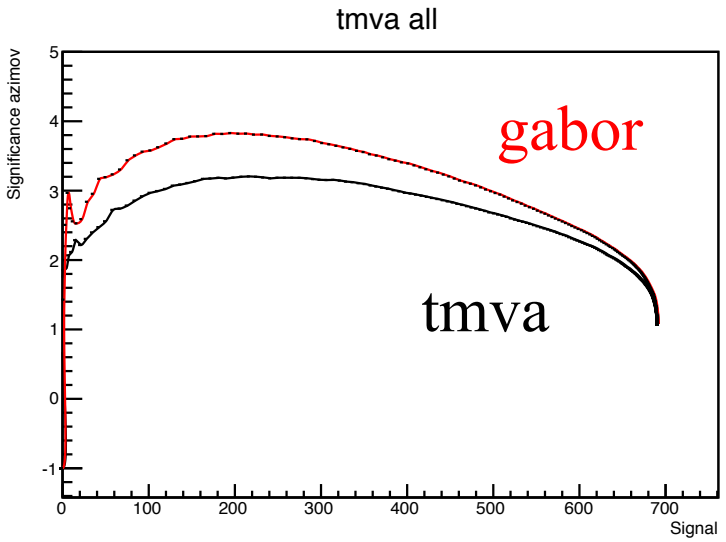


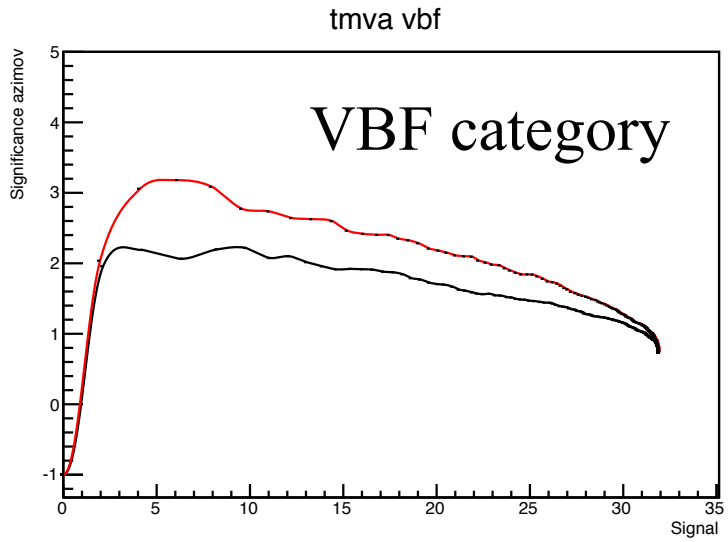
Private leaderboard top AMS curves

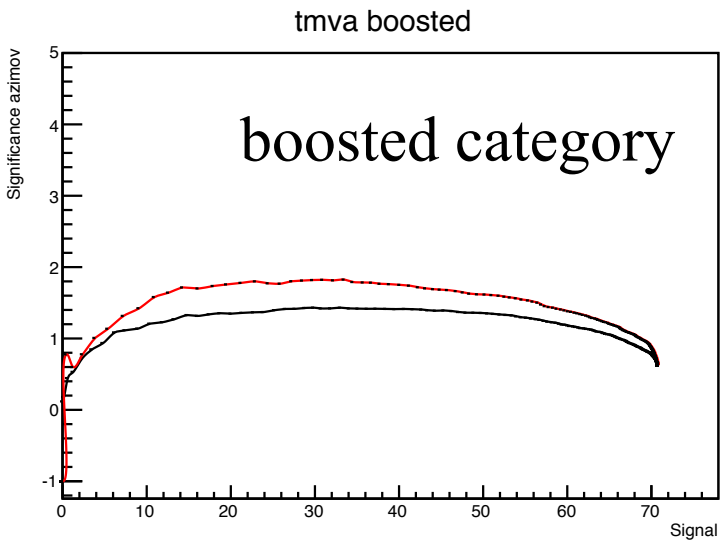


Clearly better!

TMVA vs Gabor







- vbf, boosted categories as is ATLAS note (no ATLAS insider information)
- tmva, gabor are trained without categories, on full 30 variables (not directly comparable to ATLAS analysis)
- (also significance is simple asimov, no bin, no systematics (and fake tau missing))
- Gabor improves more significantly in VBF categories (2 jets ➔events more complex)

Intermezzo

QM computation with ML

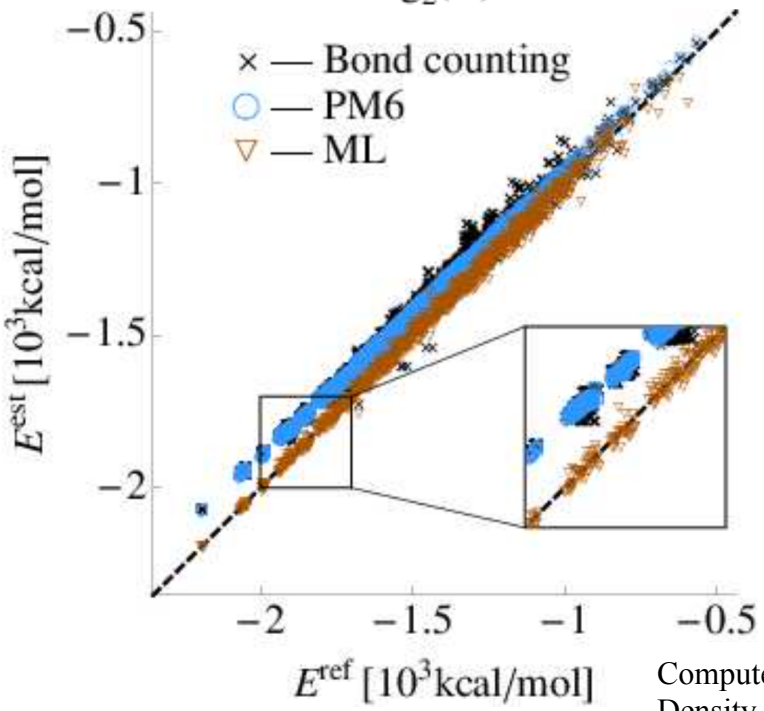
Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning Phys. Rev. Lett. **108**, 058301 — Published 31 January 2012

Matthias Rupp,[1,2] Alexandre Tkatchenko,[3,2] Klaus-Robert Müller,[1,2] and O. Anatole von Lilienfeld[4,2,*]

[1] *Machine Learning Group, Technical University of Berlin, Franklinstr 28/29, 10587 Berlin, Germany*
[2] *Institute of Pure and Applied Mathematics, University of California Los Angeles, Los Angeles, CA 90095, USA*
[3] *Fritz-Haber-Institut der Max-Planck-Gesellschaft, 14195 Berlin, Germany*
[4] *Argonne Leadership Computing Facility, Argonne National Laboratory, Argonne, Illinois 60439, USA*

Solving the Schrödinger equation (SE), $H\Psi = E\Psi$, for assemblies of atoms is a fundamental problem in quantum mechanics. Alas, solutions that are exact up to numerical precision are intractable for all but the smallest systems with very few atoms. Hierarchies of approximations have evolved, usually trading accuracy for computational efficiency [1]. Conventionally, the external potential, defined

- One day NN…NLO computation with Machine Learning?



Computed with hybrid Density-Functional Theory
(accurate but computer intensive)

Outlook

What did we learn

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

- In short (much more in next two talks):
  - 20% gain w.r.t. to untuned TMVA
  - **deep Neural nets** (but marginally better than BDT) rules
  - **Ensemble methods** (random forest, boosting) rule
  - **Meta-ensembles** of diverse models
  - **careful cross-validation** (250k training sample really small)
  - Complex software suites using routinely multithreading, GPU, etc… (e.g. XGBoost, scikit-learn)
  - Some techniques (e.g. meta-ensembles) too complex to be practical, and marginal gain…
  - …Others appear practical and useful

From domain to challenge and back

Domain e.g. HEP

Challenge organisation

Challenge

18 months

simplify

Problem

Problem

Domain experts solve the domain problem

4 months

The crowd solves the challenge problem

>n months/years ?

reimport

Solution

Solution

Outlook

- Now working on re-importing into HEP all the ML developments
- Dataset has been released on CERN Open Data Portal http://opendata.cern.ch/collection/ATLAS-Higgs-Challenge-2014, to remain available until the end of time (citeable with a d.o.i)
  - Release with the full truth
- NIPS proceedings write-up (with detailed description of "how they did it ?") to be released before the summer
- Many ideas/software/methodology to be digested and tried out possibly on different problems. Drop us a line higgsml@lal.in2p3.fr if your are interested.
- Mailing list just opened to any one with an interest in both Data Science and High Energy Physics : HEP-data-science@googlegroups.com
- Workshop "Data Science @ LHC" at CERN, 9th-13th Nov. 2015 (under the umbrella of LHC Physics Center at CERN http://lpcc.web.cern.ch/LPCC/)
- Rumors about other HEP challenges being prepared

Spares

Machine Learning?

- Machine Learning is the part of computer science which, in particular, deals among other with automatic classification:
  - E.g. neural nets to read handwritten digits (~30 years ago)
  - In HEP we call it MVA (Multi Variate Analysis)
- Developing rapidly
  - Lots of data to deal with
  - Lots of CPU power too
  - Big money involved:
    - Google/Facebook advertisements based on your searches, your gmail messages, your +1/likes
    - amazon : "we recommend for you" based on what you already bought
  - "Big data" buzz word
  - New field "Data Science"

What make a good challenge?

Isabelle Guyon http://chalearn.org/

- Good problem
- Good data + enough data
- Clear objective + good metric
- Simple rules + no Information Protection contingencies
- Prizes (~ $5000)
- Starter kit
- On-line feed-back
- Publication outlets
- Avoid many pitfalls, in particular leaking the truth, or a large luck factor in the ranking
- (see http://ciml.chalearn.org/schedule a NIPS satellite workshop on organising challenges, in particular Ben Hammer's, Kaggle chief scientist, on do's and don'ts)

What data did we release ?

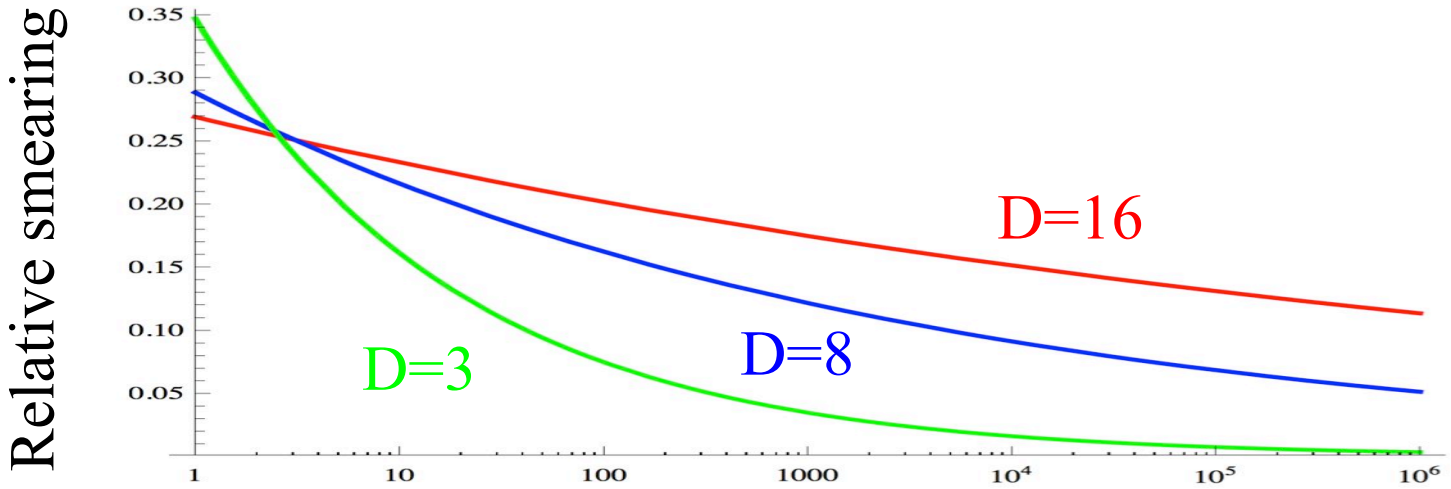
- From ATLAS full sim Geant4 MC12 production
- 30 variables
- Signal is H➔tautau, Background a mixture of : Z, top, W
- Based on November 2013 ATLAS Htautau conf note ATLAS-CONF-2013-108
- Preselection for lep-had topology : single lepton trigger, one lepton identified, one hadronic tau identified
- ➔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)

The ATLAS hacker

- In principle, since the data we release is available to all ATLAS members, one ATLAS hacker could:
  - use the variables to match back (using lepton/hadron/jet Pt eta phi) to the original MC event
  - check whether it is signal or background
  - cheat to get best significance
- (this would be discovered at the software release stage)
- This was transparently explained to ATLAS :it would be bad for our image and counterproductive if the public leaderboard is cluttered by ATLAS hackers
- ➔no evidence/no hint there was such attempt

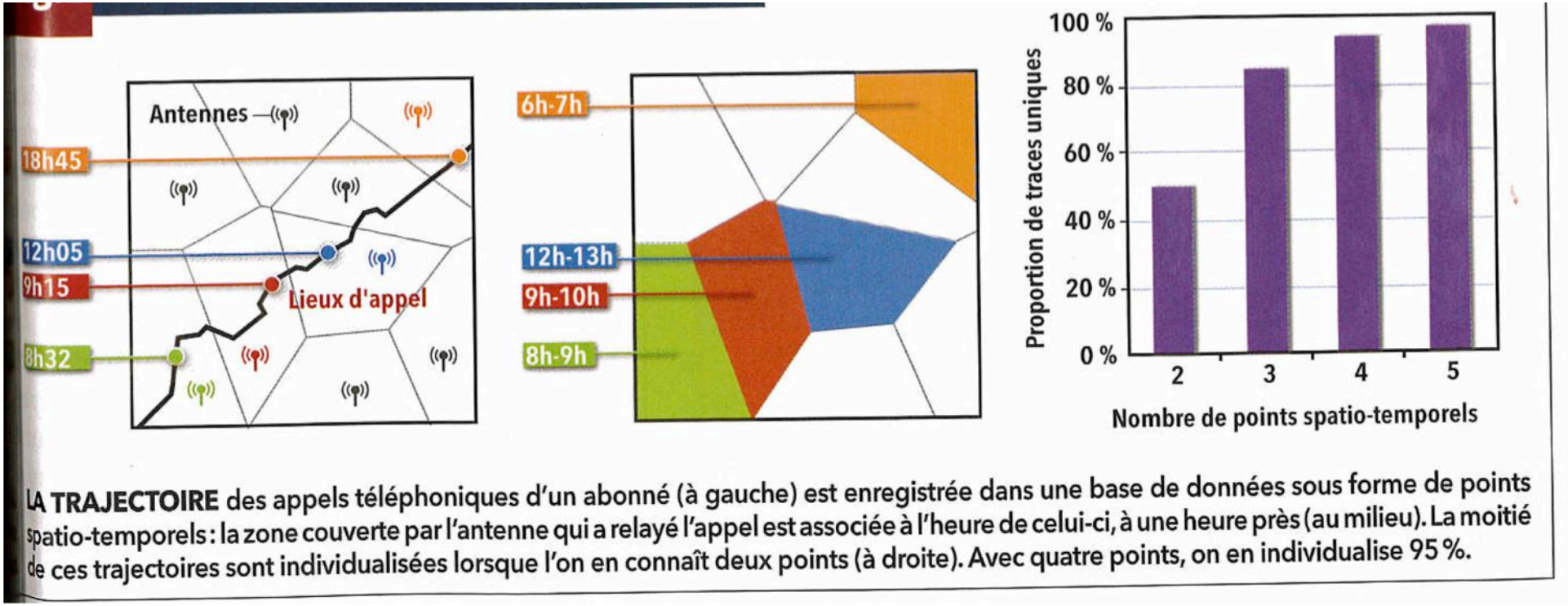
ATLAS hacker (2)

- We thought of smearing the parameters to prevent the matching
- Back of an envelope calculation: by how much should D variables be smeared so that the original can be matched with 50% probability among N entries
- ➔smearing should be at least 15%
- ➔events become meaningless
- ➔bad idea! Not pursued



ATLAS hacker (3)

- Hiding the origin of an entry in a DB is called "sanitizing", this is a notoriously difficult, and very hot topic, e.g:
  - Finding owner of a medical record given anonymized parameters (gender, zip code, and birth date uniquely identifies 83% of US people)
  - Finding owner of a mobile phone given hour and area of a few calls (Pour La Science June 2014):



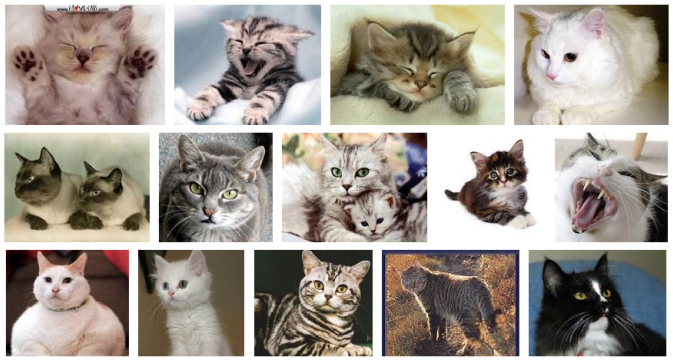
LA TRAJECTOIRE des appels téléphoniques d'un abonné (à gauche) est enregistrée dans une base de données sous forme de points spatio-temporels : la zone couverte par l'antenne qui a relayé l'appel est associée à l'heure de celui-ci, à une heure près (au milieu). La moitié de ces trajectoires sont individualisées lorsque l'on en connaît deux points (à droite). Avec quatre points, on en individualise 95 %.

Licensing issues

- Anyone participating to the challenge had to agree to the rules (https://www.kaggle.com/c/higgs-boson/rules) in particular:
- Software
  - Participants can use whatever software they like, but to win a price, they have to release it under an OS license (so that we can 1) verify it 2) use it eventually)
- Simulated data:
  - All ATLAS real and simulated data belongs (legally) to CERN
  - We've had the signature of S. Bertolucci, CERN Director of Research, to release the data
  - Data was initially made available **only** for the duration of the challenge and **only** for the challenge ("Can I use the data for a master thesis ?" "Sorry no.")➔finally agreement later to release the data on CERN ODP➔(got the same question again from someone else, "Yes please!")

Transfer Learning (2011)

- Learn a data representation from one task



- Use it for another task:

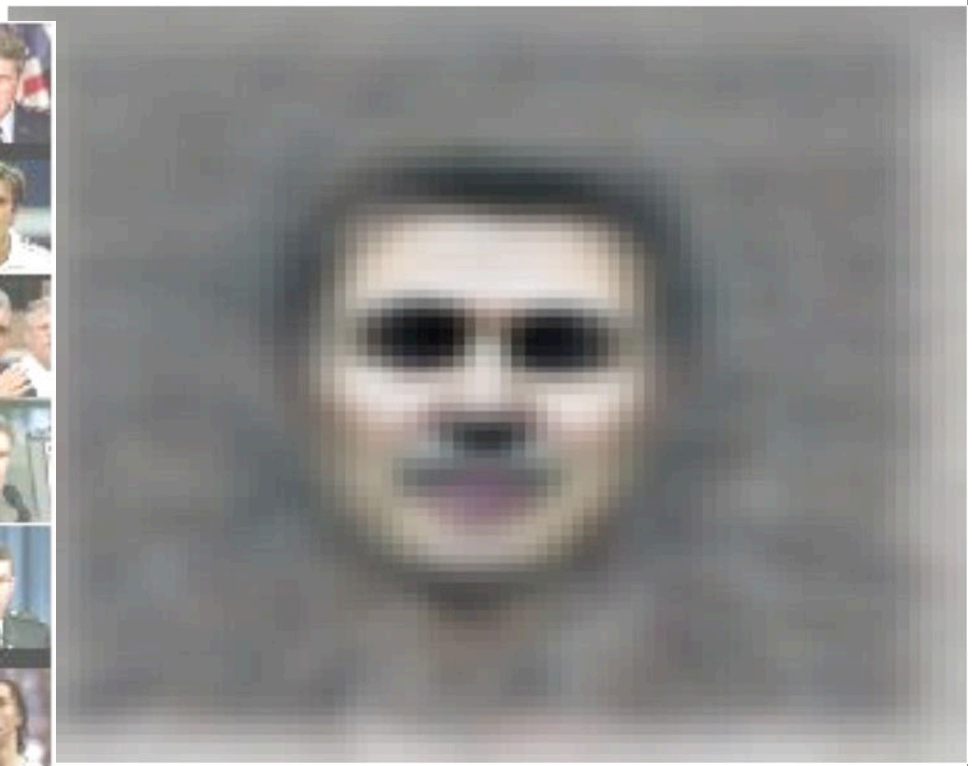
http://chalearn.org/

Million $ prize challenges

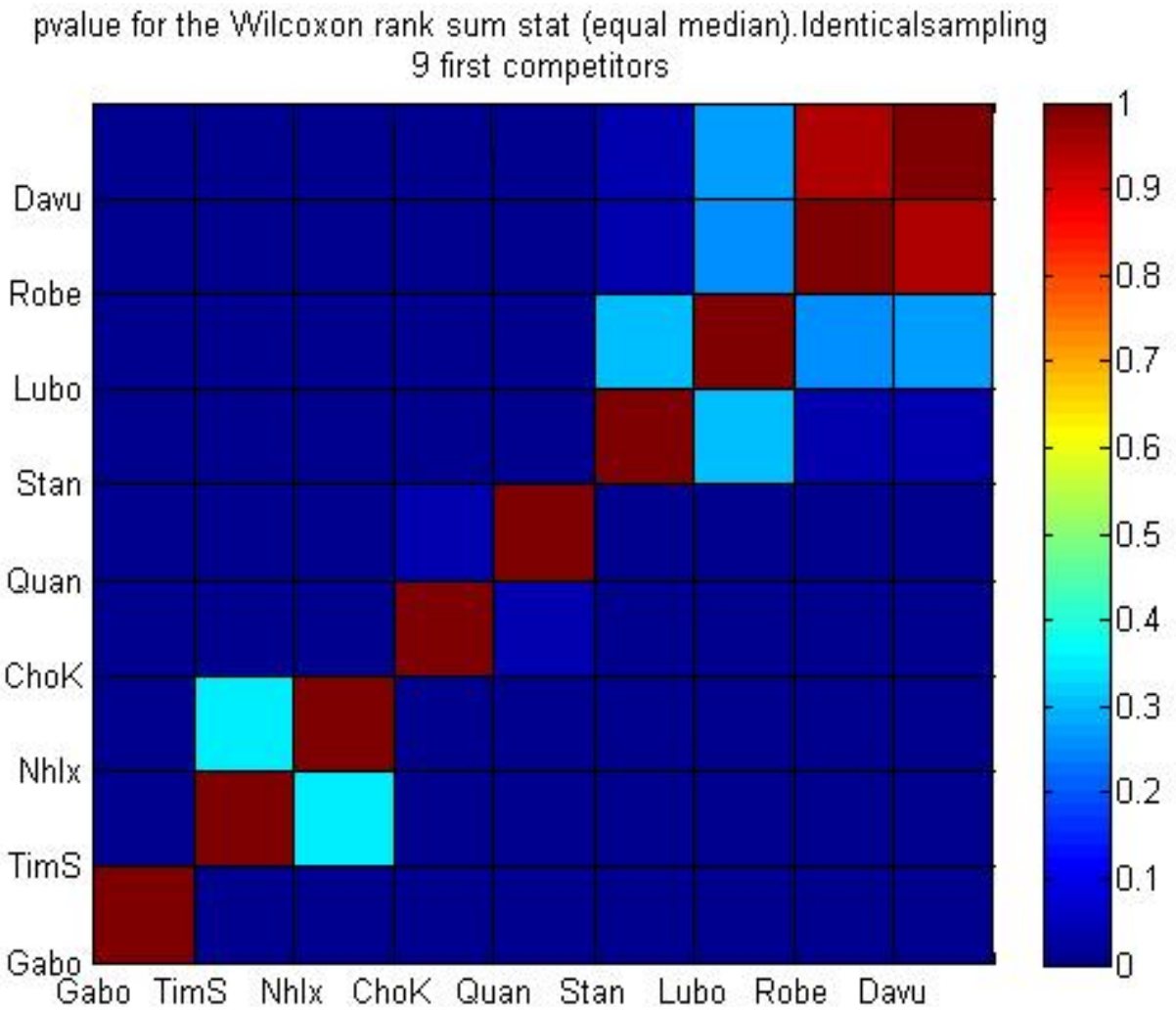
2009 $1,000,000

BDT and NN

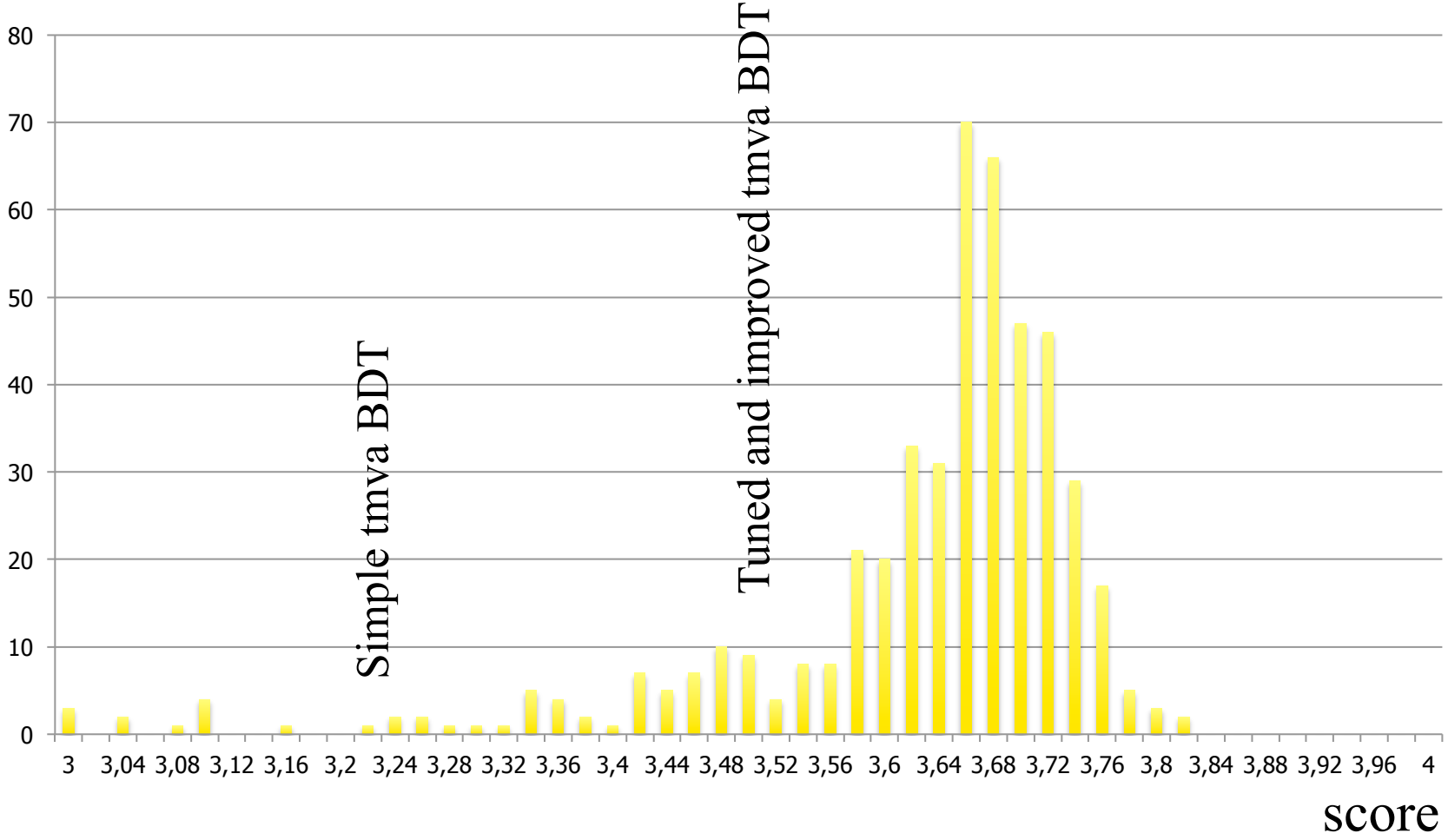
- BDT (Boosted Decision Tree) which is by far the most used technique in Atlas/CMS is actually an old technique (Adaboost 1997)
- More recent techniques, just an example:
  - Unsupervised neural network (Example: The Google cat: deep learning technique running on 16K cores for three days, watching 10M random YouTube video stills [Le et al., ICML'12]



Are winning score different ?



Best private scores



Private score vs public: nhlx5haze