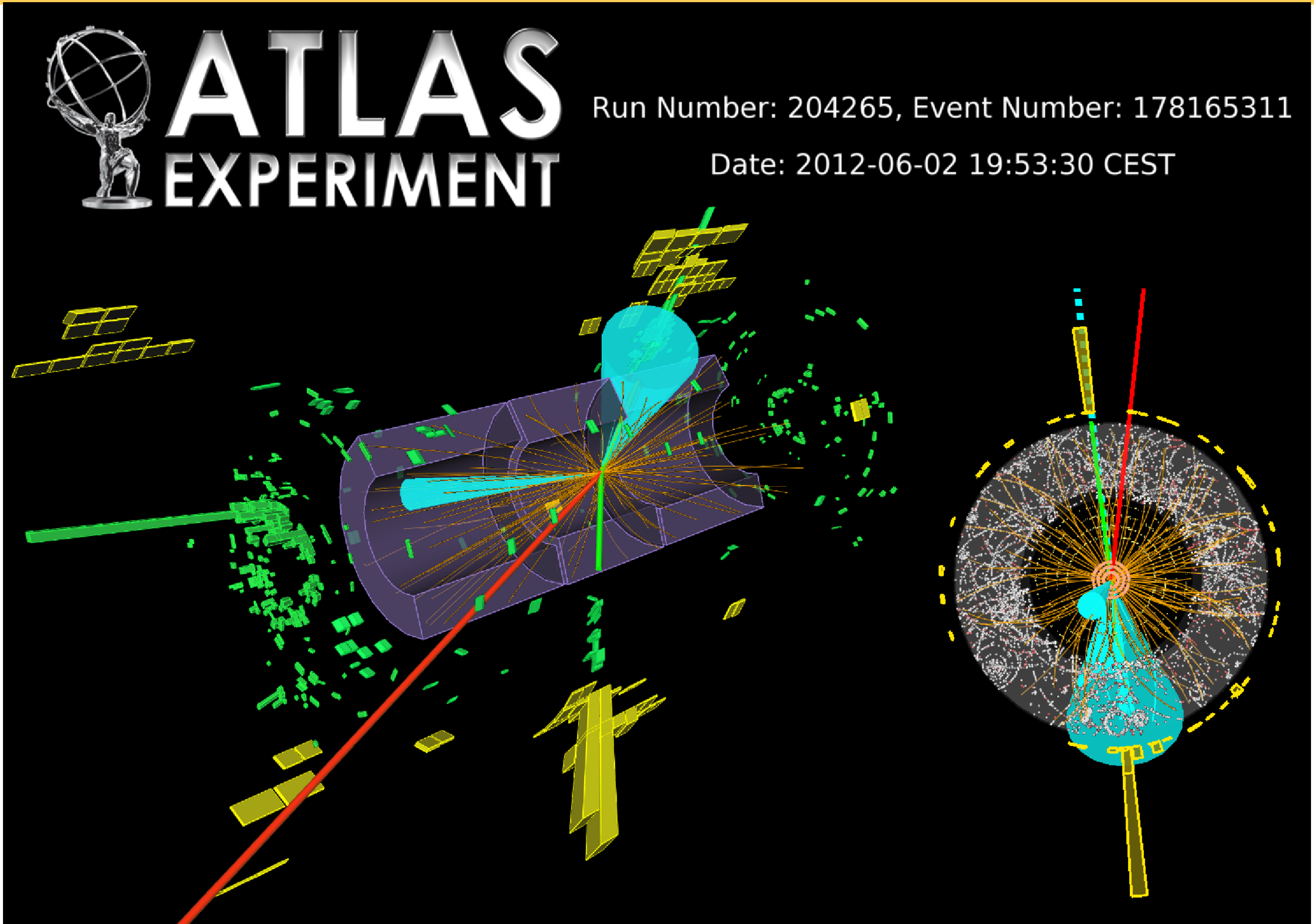
Introduction to the HEPML workshop and the HiggsML challenge

Balázs Kégl
David Rousseau, Cécile Germain, Isabelle Guyon, Glen Cowan

CNRS/IN2P3/University Paris-S{ud,aclay}, ChaLearn, Royal Holloway

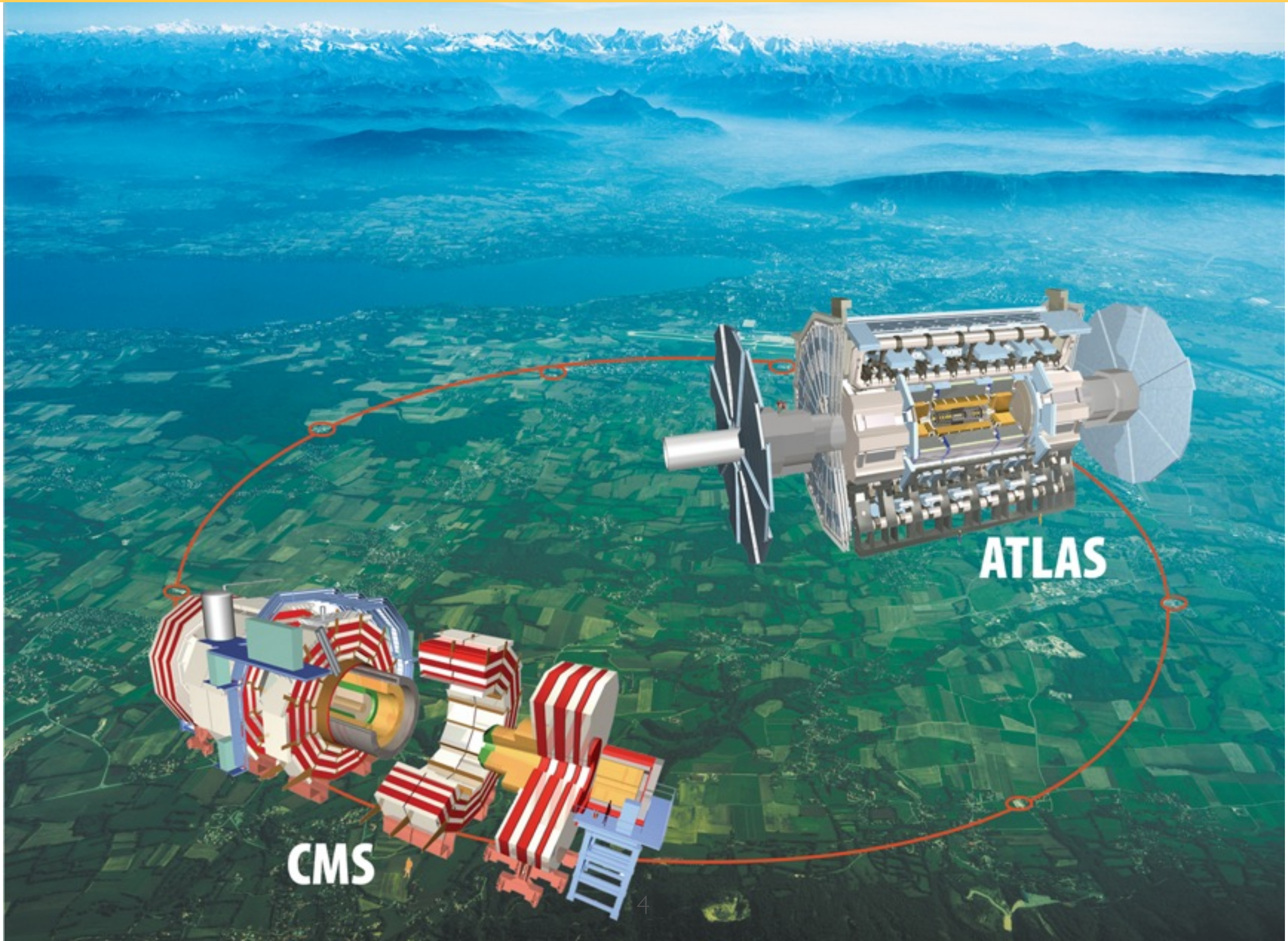
HEPML NIPS'14 workshop
December 13, 2014

Classification for discovery
The HiggsML challenge

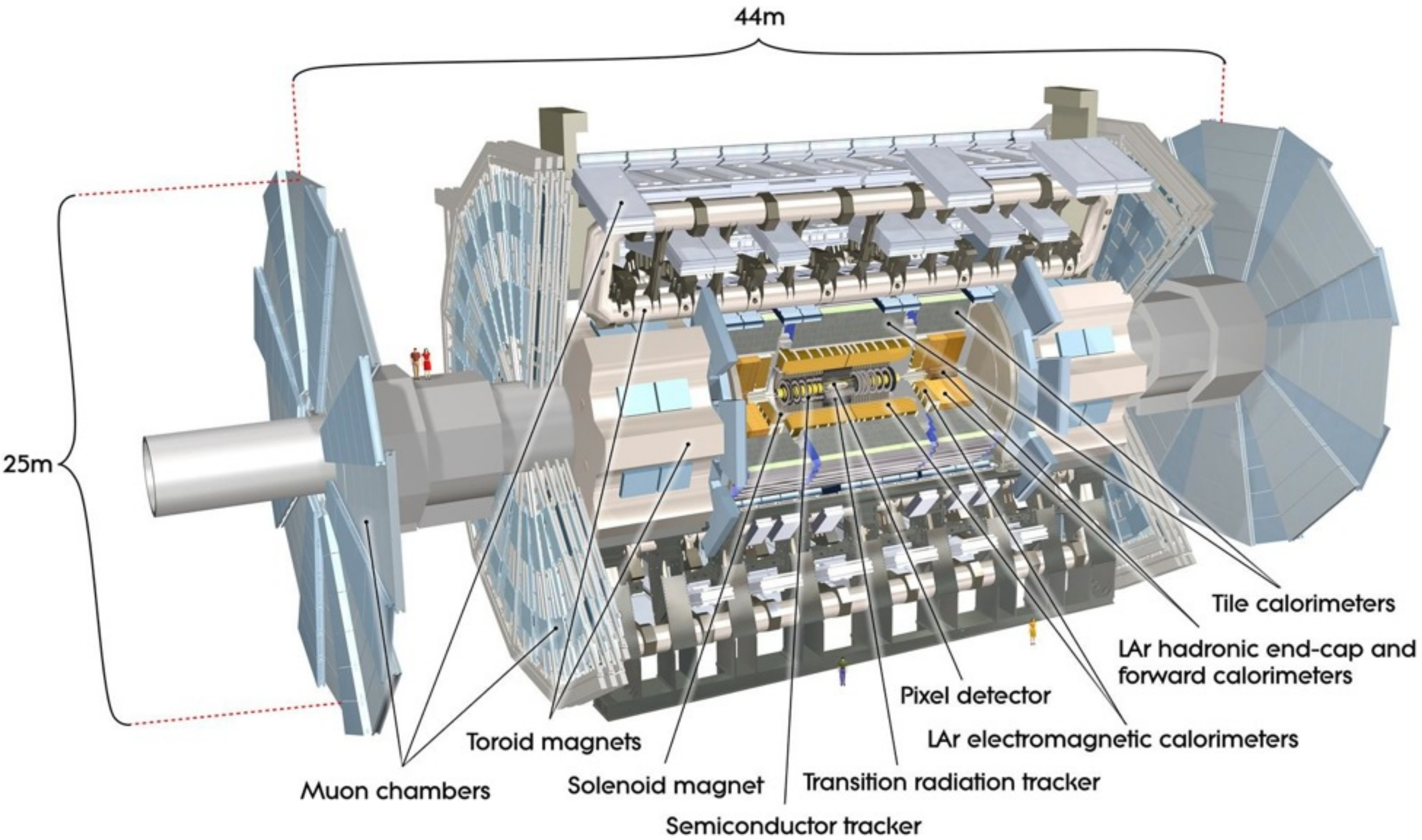

ATLAS EXPERIMENT
Run Number: 204265, Event Number: 178165311
Date: 2012-06-02 19:53:30 CEST

Data collection

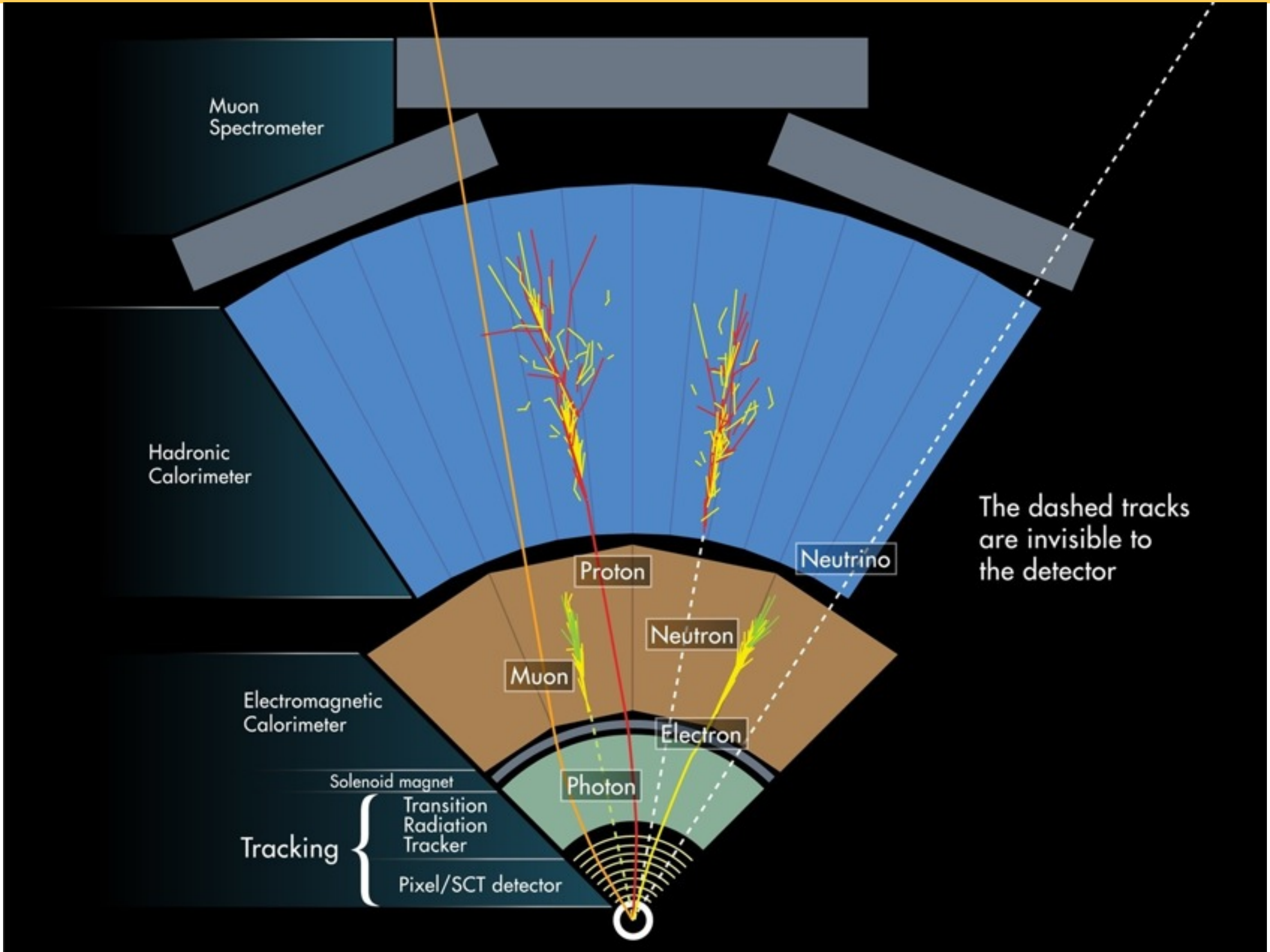
The LHC in Geneva



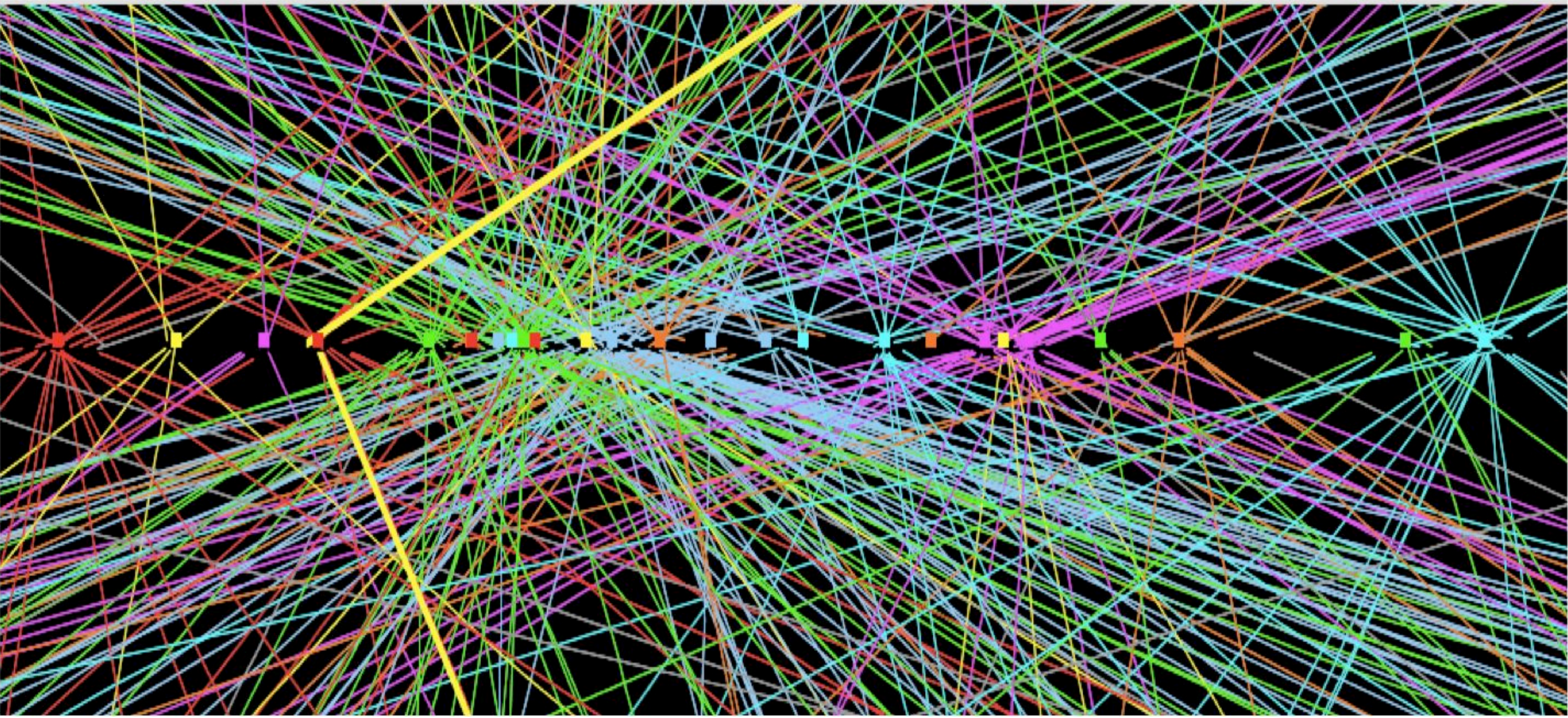
The ATLAS detector


44m
25m
Tile calorimeters
LAr hadronic end-cap and forward calorimeters
Pixel detector
LAr electromagnetic calorimeters
Toroid magnets
Muon chambers
Solenoid magnet
Transition radiation tracker
Semiconductor tracker

THE ATLAS DETECTOR



The ATLAS detector



Data Collection

- **Hundreds of millions** of proton-proton collisions **per second**
- Filtered down to **400 events per second**
  - still **petabytes per year**
  - **real-time** (budgeted) classification: trigger
  - a research theme on its own

Feature engineering

Feature engineering

- Each collision is an event
  - hundreds of particles: decay products
  - hundreds of thousands of sensors (but sparse)
  - for each particle: type, energy, direction is measured
  - a fixed-length list of ~30-40 extracted features: *x*
  - e.g., angles, energies, directions, reconstructed mass
  - based on 50 years of accumulated domain knowledge

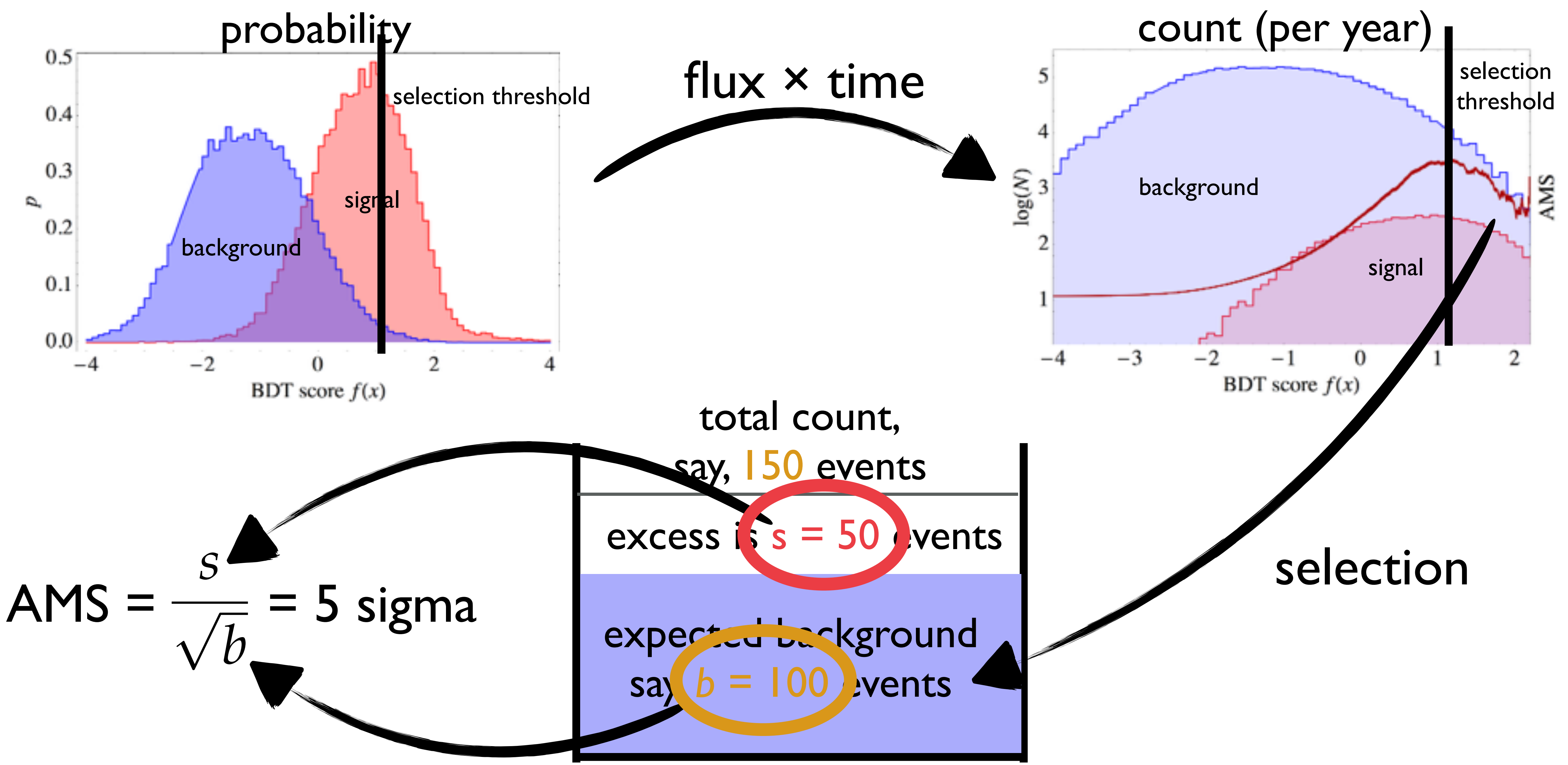
THE HIGGS TO TAU-TAU CHANNEL

- Highly **unbalanced** data
  - we expect to see **<100** Higgs bosons per year in $\approx 10^{10}$ events
  - after pre-selection, we will have
    500K **background (negative)** and
    1K **signal (positive)** events per year (2012)
  - Training on **simulated data**

The metric

Classification for discovery

Goal: optimize the expected discovery significance



probability

flux × time

count (per year)

selection

total count, say, 150 events

excess is s = 50 events

expected background say, b = 100 events

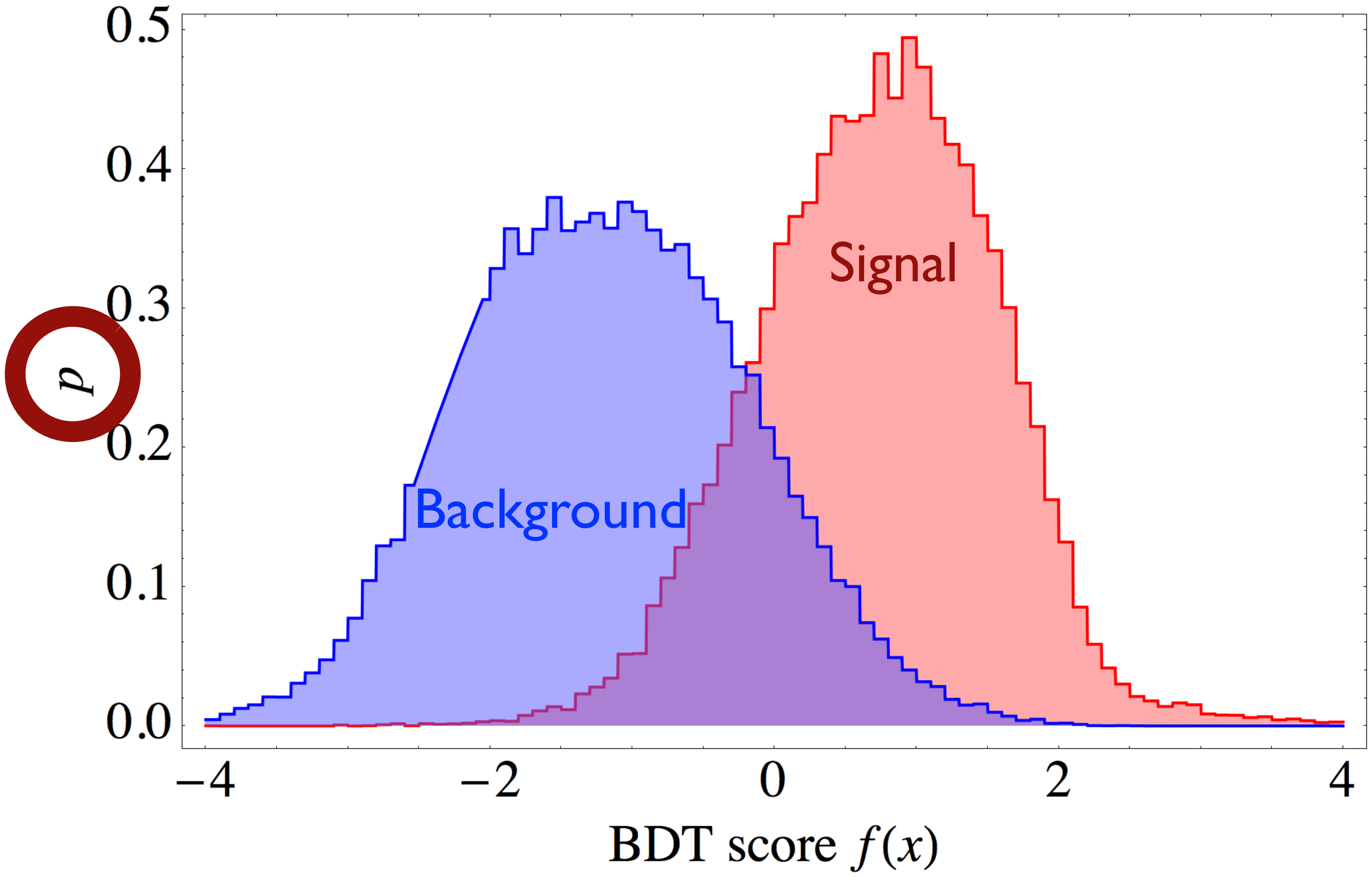
$$\text{AMS} = \frac{s}{\sqrt{b}} = 5 \text{ sigma}$$

CLASSIFICATION FOR DISCOVERY

How to design g to maximize the sensitivity?

- A two-stage approach

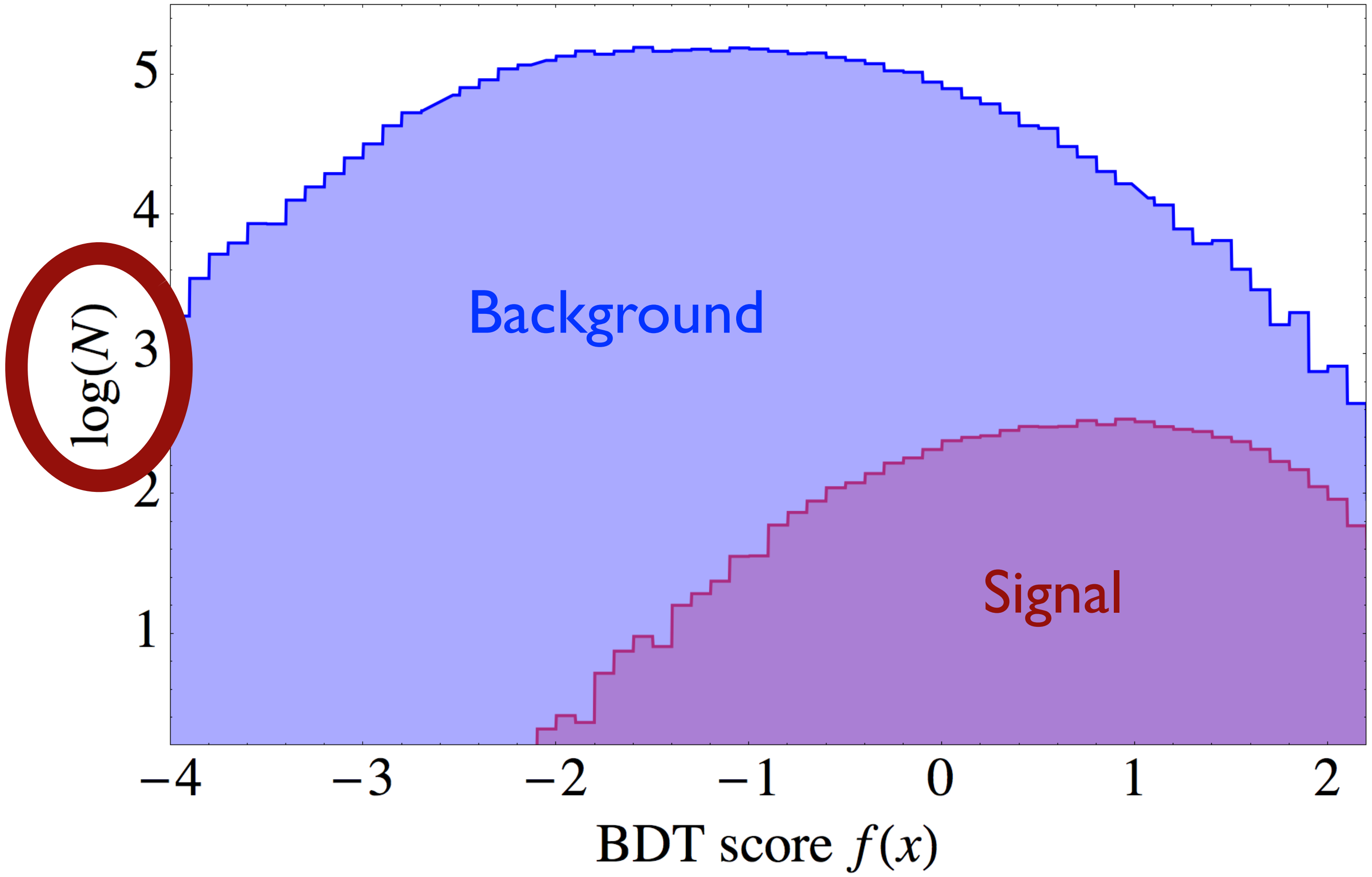
  1. optimize a discriminant (score) function $f : \mathbb{R}^d \to \mathbb{R}$ using a classical learning algorithm (BDT, NN)



Classification for discovery

How to design g to maximize the sensitivity?

- A two-stage approach
  1. optimize a discriminant (score) function $f : \mathbb{R}^d \rightarrow \mathbb{R}$ using a classical learning algorithm (BDT, NN)

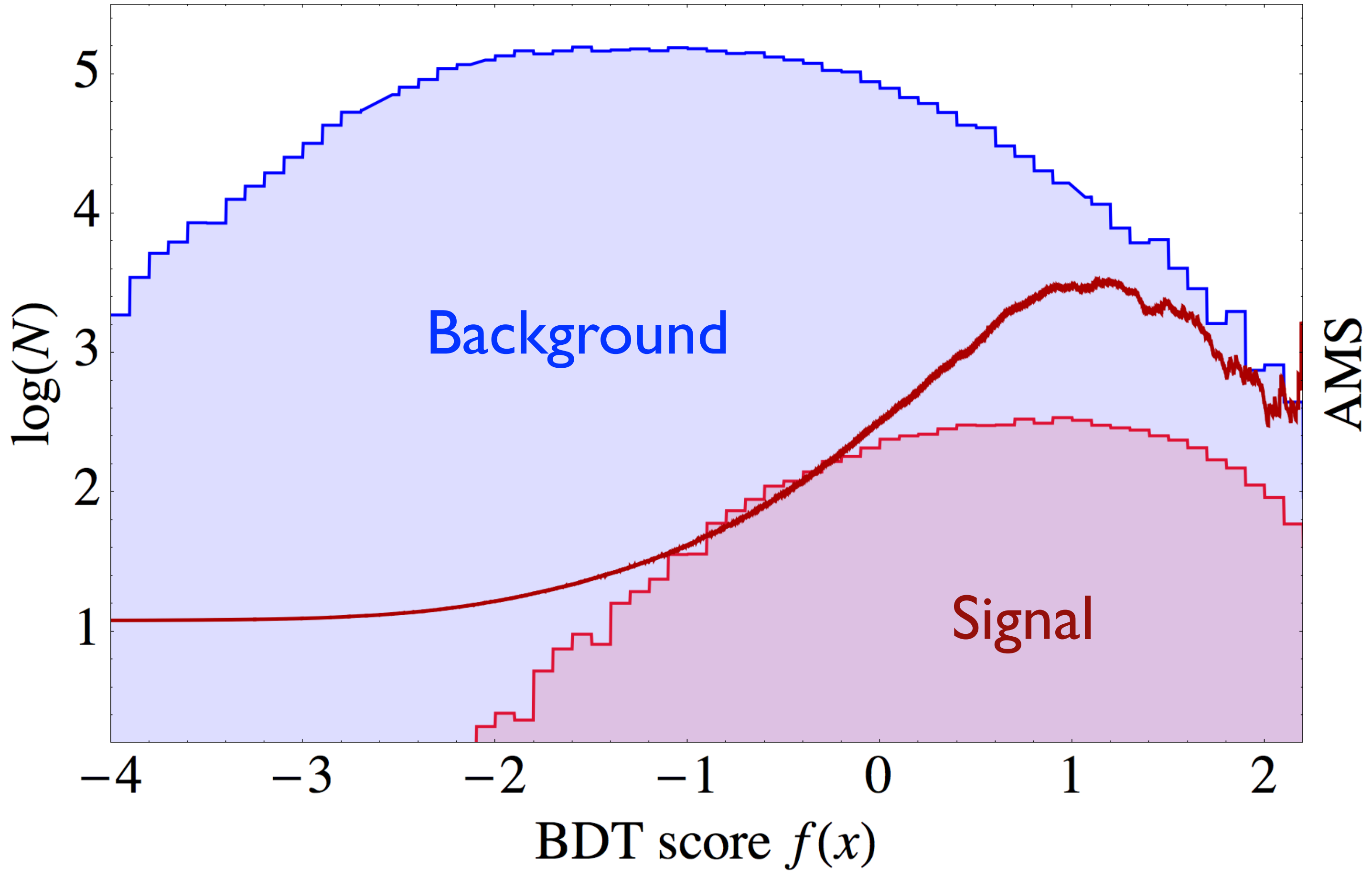


CLASSIFICATION FOR DISCOVERY

How to design g to maximize the sensitivity?

- A two-stage approach

1. optimize a discriminant (score) function $f : \mathbb{R}^d \to \mathbb{R}$ using a classical learning algorithm (BDT, NN)

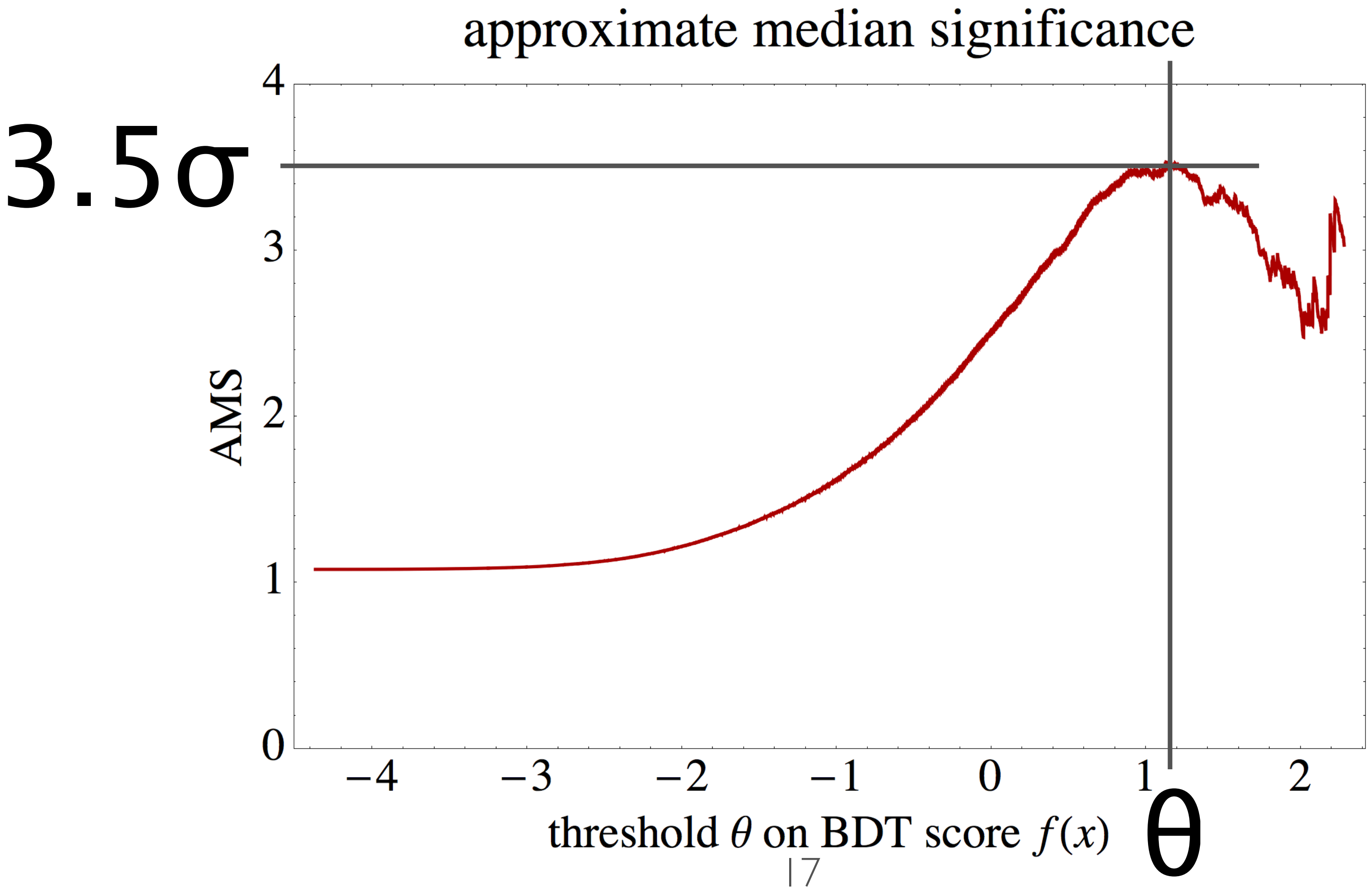


CLASSIFICATION FOR DISCOVERY

How to design g to maximize the sensitivity?

- A two-stage approach (make figure with score)

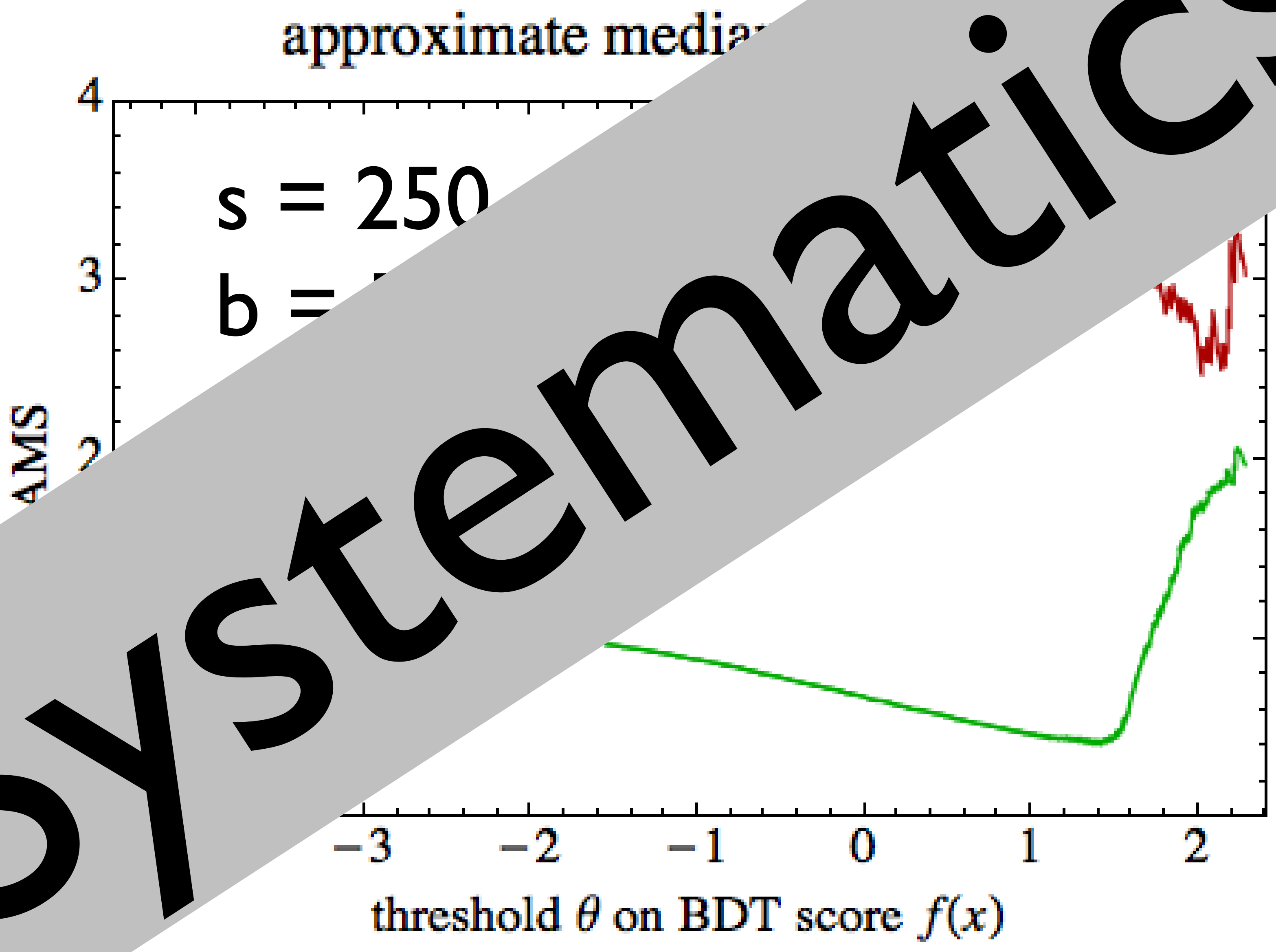
1. optimize a discriminant (score) function $f : \mathbb{R}^d \to \mathbb{R}$ using a classical learning algorithm (BDT, NN)
2. define $g(\mathbf{x}) = \text{sign}\big(f(\mathbf{x}) - \theta\big)$ and optimize θ for maximizing the AMS



Classification for discovery

Comparing with Atlas analysis

- Atlas does a manual pre-selection (category), th of the AMS is completely eliminated. Why?



Systematics!

CLASSIFICATION FOR DISCOVERY

How to handle systematic (model) uncertainties?

- OK, so let's design an objective function that can take background systematics into consideration

  - Likelihood with unknown background $b \sim \mathcal{N}(\mu_\mathsf{b}, \sigma_\mathsf{b})$

$$L(\mu_\mathsf{s}, \mu_\mathsf{b}) = P(n, b | \mu_\mathsf{s}, \mu_\mathsf{b}, \sigma_\mathsf{b}) = \frac{(\mu_\mathsf{s} + \mu_\mathsf{b})^n}{n!} e^{-(\mu_\mathsf{s}+\mu_\mathsf{b})} \frac{1}{\sqrt{2\pi}\sigma_\mathsf{b}} e^{-(b-\mu_\mathsf{b})^2/2{\sigma_\mathsf{b}}^2}$$

  - Profile likelihood ratio $\lambda(0) = \frac{L(0, \hat{\hat{\mu}}_\mathsf{b})}{L(\hat{\mu}_\mathsf{s}, \hat{\mu}_\mathsf{b})}$

  - The new Approximate Median Significance (by Glen Cowan)

$$\mathsf{AMS} = \sqrt{2\left((s+b)\ln\frac{s+b}{b_0} - s - b + b_0\right) + \frac{(b-b_0)^2}{{\sigma_\mathsf{b}}^2}}$$

where

$$b_0 = \frac{1}{2}\left(b - {\sigma_\mathsf{b}}^2 + \sqrt{(b - {\sigma_\mathsf{b}}^2)^2 + 4(s+b){\sigma_\mathsf{b}}^2}\right)$$

Classification for discovery

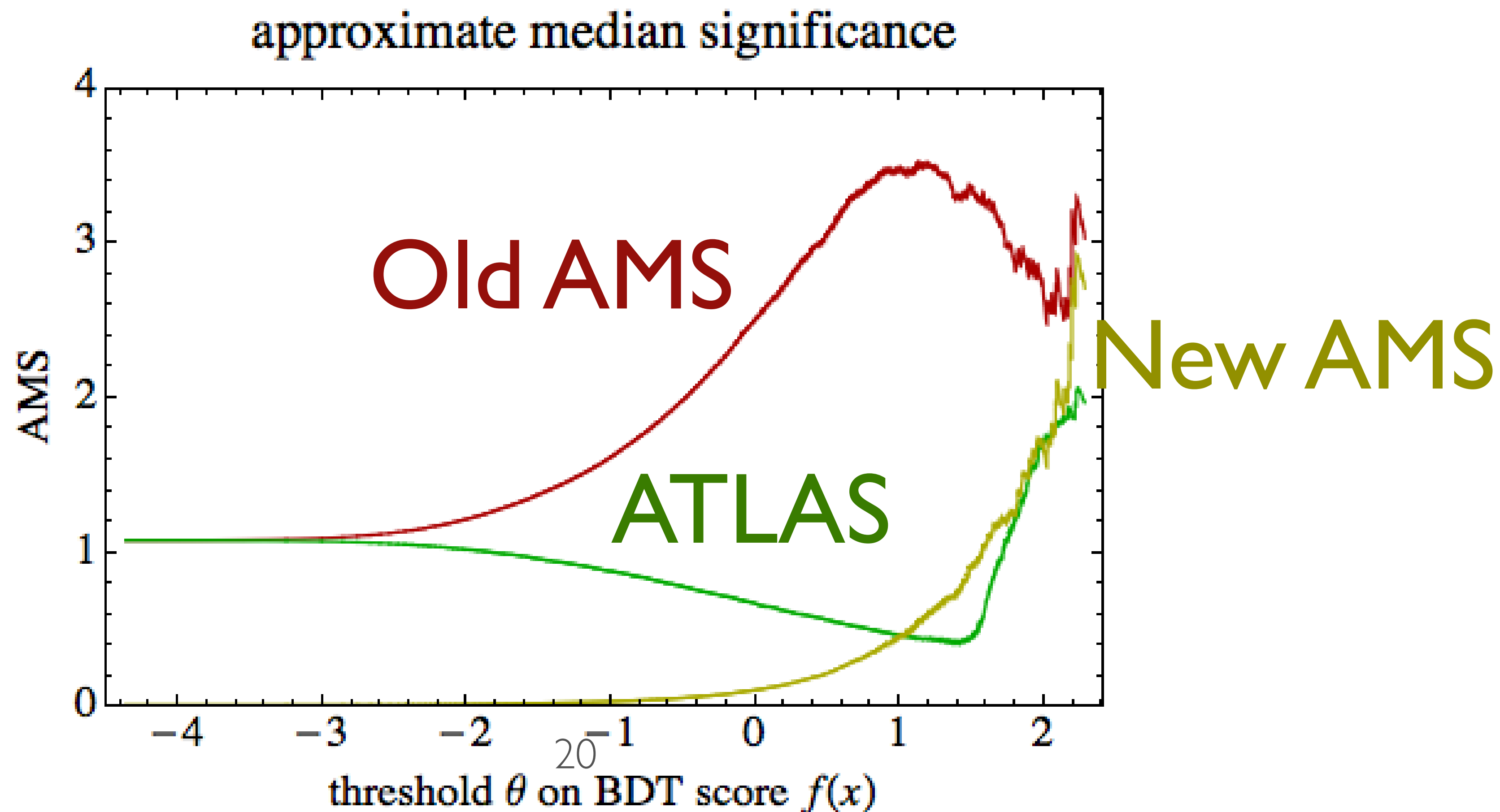
How to handle systematic (model) uncertainties?

- The new Approximate Median Significance

$$\text{AMS} = \sqrt{2\left((s+b)\ln\frac{s+b}{b_0} - s - b + b_0\right) + \frac{(b-b_0)^2}{\sigma_\text{b}{}^2}}$$

where

$$b_0 = \frac{1}{2}\left(b - \sigma_\text{b}{}^2 + \sqrt{(b-\sigma_\text{b}{}^2)^2 + 4(s+b)\sigma_\text{b}{}^2}\right)$$

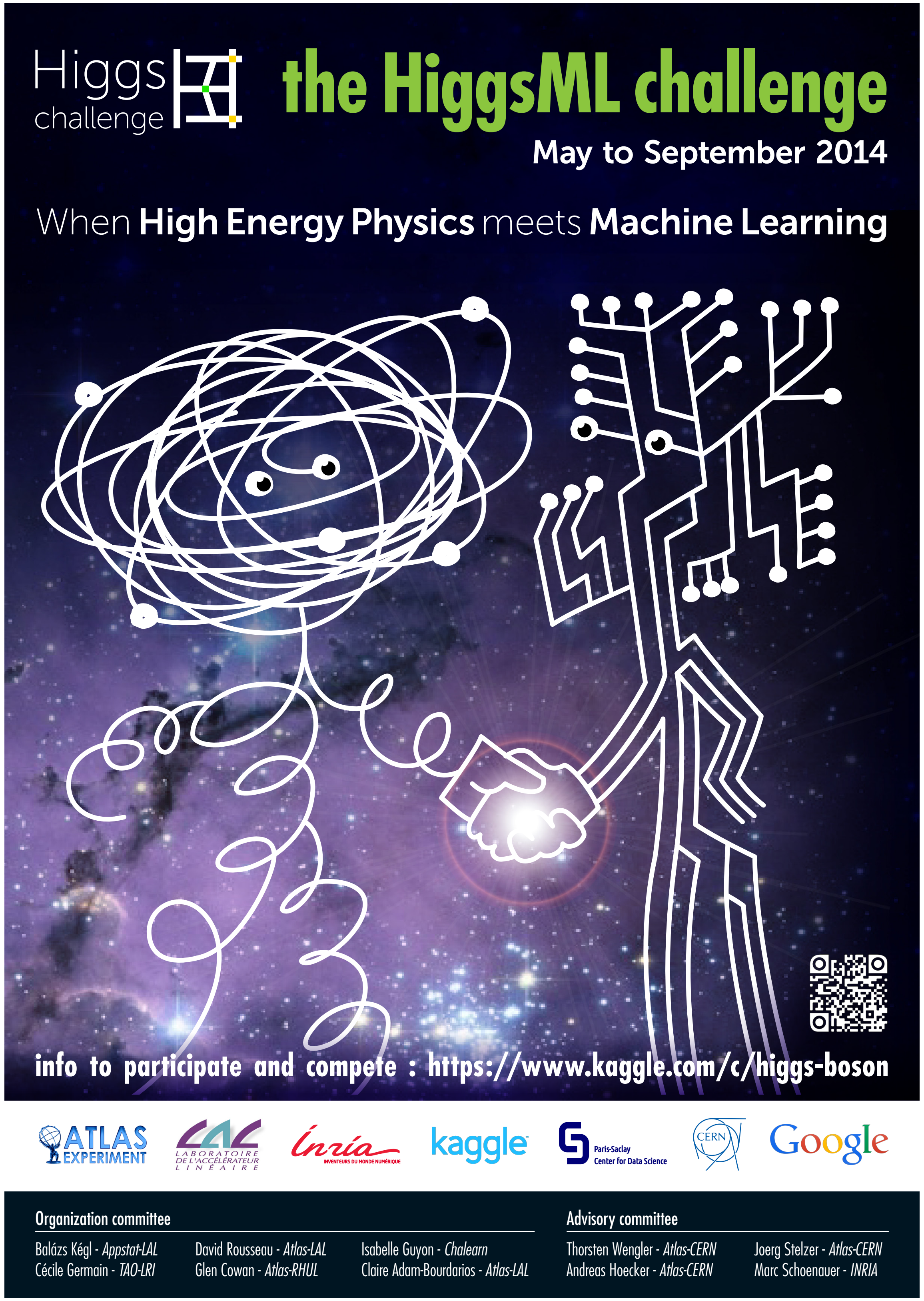


CLASSIFICATION FOR DISCOVERY

- Exciting physics
    - The Higgs to tau-tau excess is not yet at five sigma
      Tech. Rep. ATLAS-CONF-2013-108
- Exciting data science
    - What is the theoretical relationship between classification and test sensitivity?
    - What is the quantitative criteria to optimize?
    - How to formally include systematic uncertainties?
    - How to design (or redesign classical) algorithms for optimizing the criteria?
    - Redesign the counting test?

Classification for discovery

We organized a
data challenge
to answer some of these
questions



Higgs challenge

the HiggsML challenge

May to September 2014

When **High Energy Physics** meets **Machine Learning**

info to participate and compete : https://www.kaggle.com/c/higgs-boson

ATLAS EXPERIMENT · LAL LABORATOIRE DE L'ACCÉLÉRATEUR LINÉAIRE · Inria INVENTEURS DU MONDE NUMÉRIQUE · kaggle · Paris-Saclay Center for Data Science · CERN · Google

Organization committee

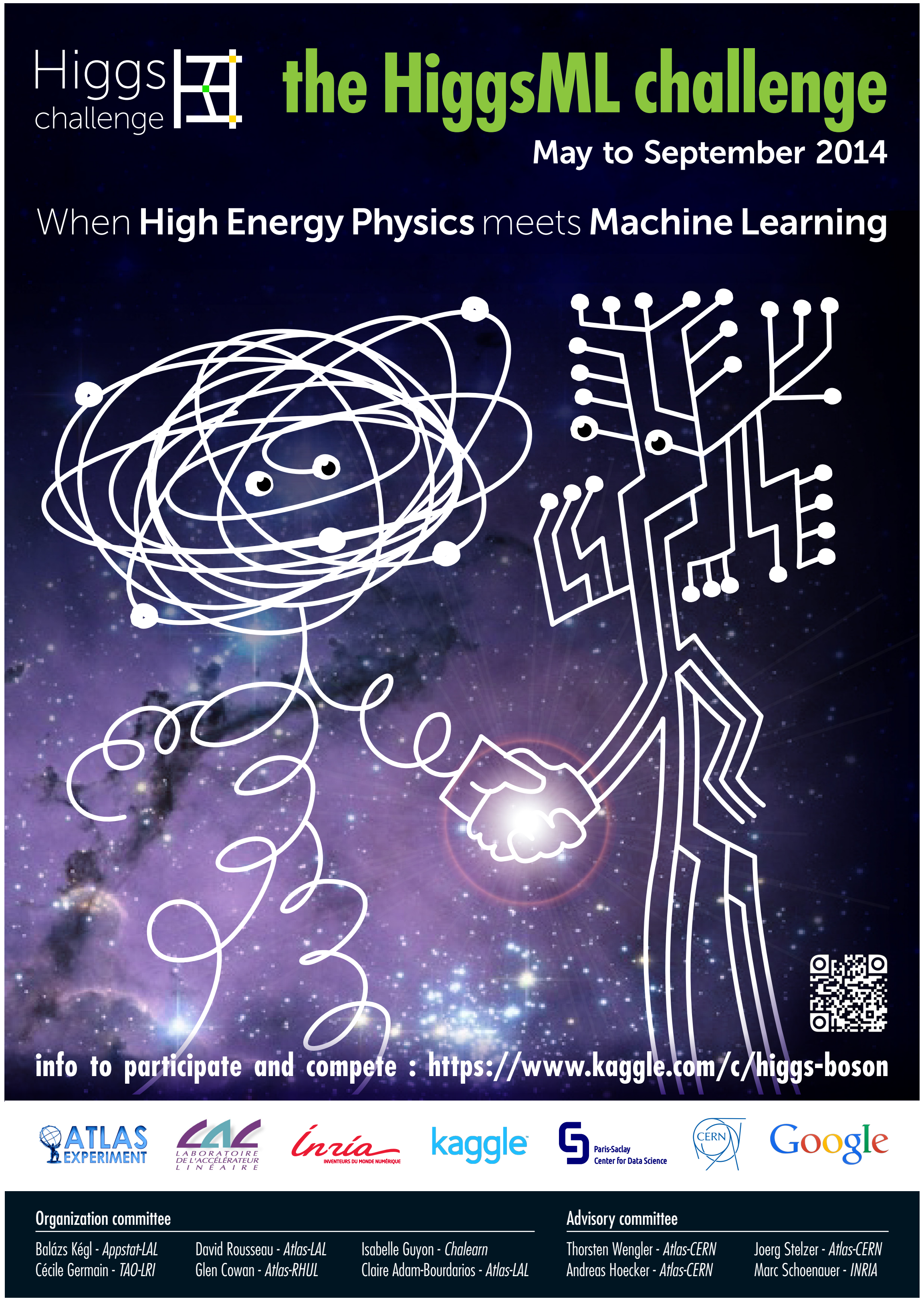
Balázs Kégl - *Appstat-LAL*
Cécile Germain - *TAO-LRI*
David Rousseau - *Atlas-LAL*
Glen Cowan - *Atlas-RHUL*
Isabelle Guyon - *Chalearn*
Claire Adam-Bourdarios - *Atlas-LAL*

Advisory committee

Thorsten Wengler - *Atlas-CERN*
Andreas Hoecker - *Atlas-CERN*
Joerg Stelzer - *Atlas-CERN*
Marc Schoenauer - *INRIA*

Classification for Discovery

- Organizing committee
  - David Rousseau (ATLAS / LAL)
  - Balázs Kégl (AppStat / LAL)
  - Cécile Germain (LRI / UPSud)
  - Glen Cowan (ATLAS / Royal Holloway)
  - Claire Adam Bourdarios (ATLAS / LAL)
  - Isabelle Guyon (ChaLearn)



Higgs challenge

the HiggsML challenge

May to September 2014

When **High Energy Physics** meets **Machine Learning**

info to participate and compete : https://www.kaggle.com/c/higgs-boson

ATLAS EXPERIMENT · LABORATOIRE DE L'ACCÉLÉRATEUR LINÉAIRE · Inria INVENTEURS DU MONDE NUMÉRIQUE · kaggle · Paris-Saclay Center for Data Science · CERN · Google

Organization committee

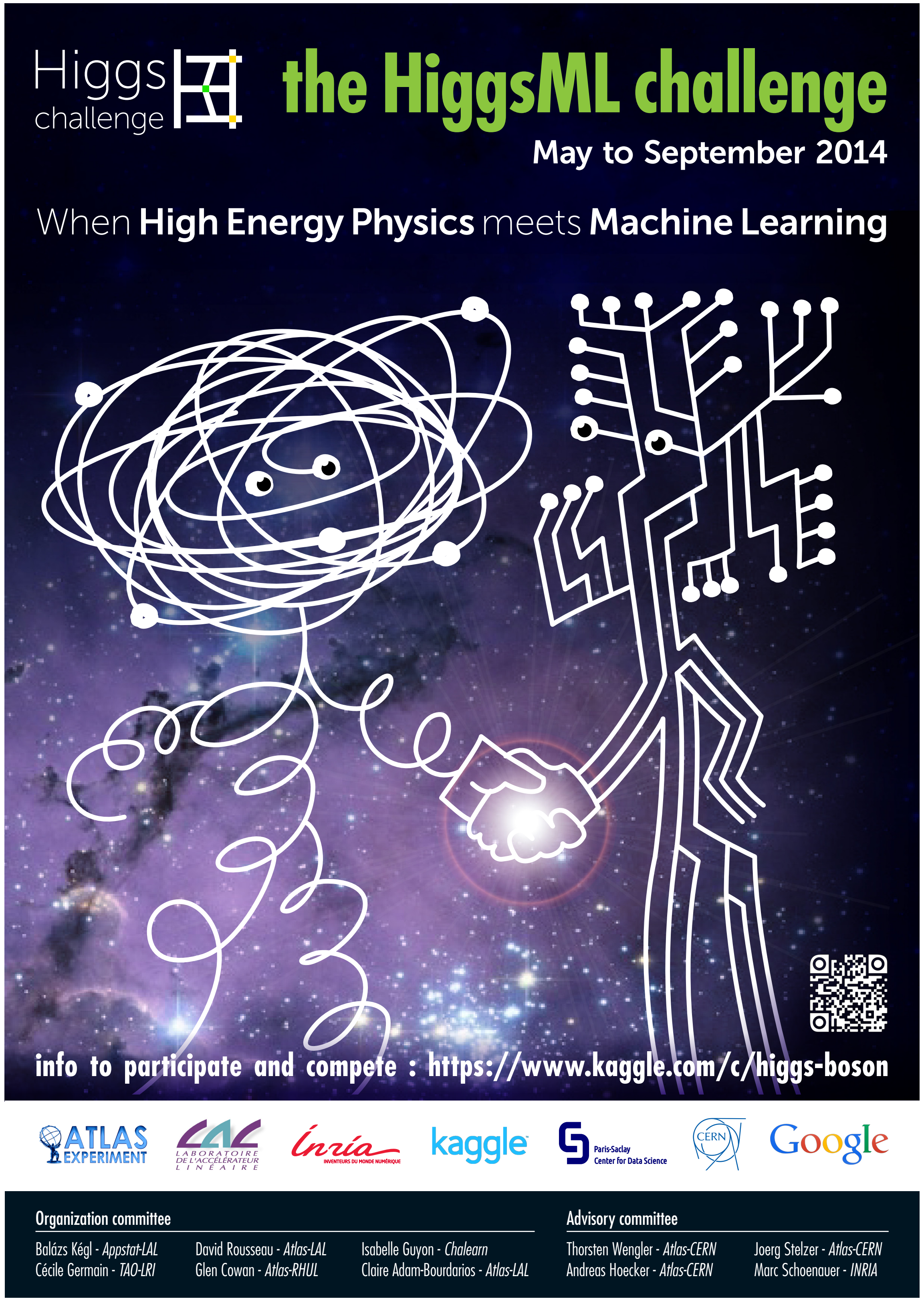
Balázs Kégl - *Appstat-LAL*
Cécile Germain - *TAO-LRI*
David Rousseau - *Atlas-LAL*
Glen Cowan - *Atlas-RHUL*
Isabelle Guyon - *Chalearn*
Claire Adam-Bourdarios - *Atlas-LAL*

Advisory committee

Thorsten Wengler - *Atlas-CERN*
Andreas Hoecker - *Atlas-CERN*
Joerg Stelzer - *Atlas-CERN*
Marc Schoenauer - *INRIA*

Classification for discovery

- 16K$ prize pool
  - 7-4-2K$ for the top three
  - HEP meets ML award for the most useful model, decided by the ATLAS members of the organizing committee



Higgs challenge

the HiggsML challenge

May to September 2014

When High Energy Physics meets Machine Learning

info to participate and compete : https://www.kaggle.com/c/higgs-boson

Organization committee

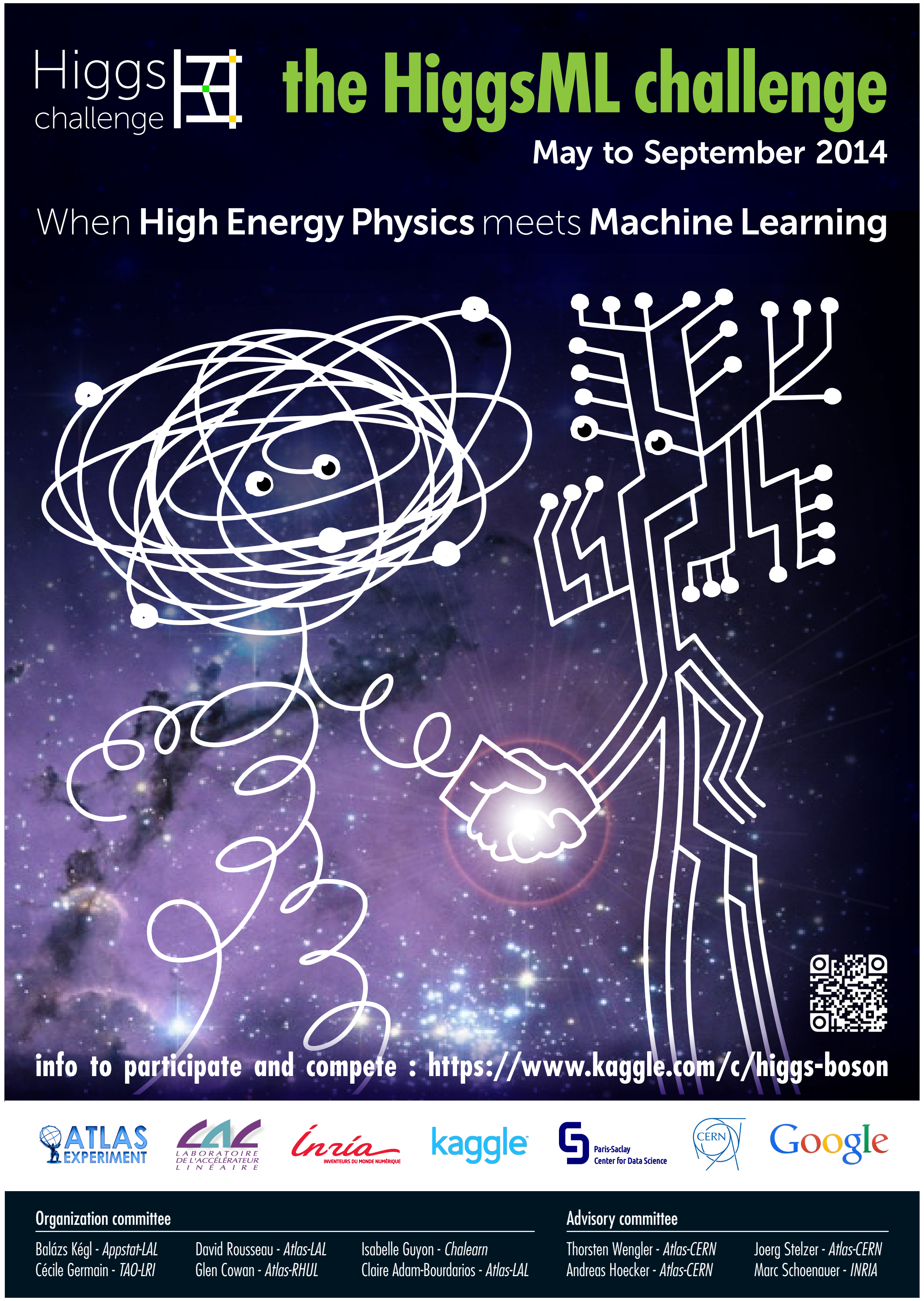
Balázs Kégl - *Appstat-LAL*
Cécile Germain - *TAO-LRI*
David Rousseau - *Atlas-LAL*
Glen Cowan - *Atlas-RHUL*
Isabelle Guyon - *Chalearn*
Claire Adam-Bourdarios - *Atlas-LAL*

Advisory committee

Thorsten Wengler - *Atlas-CERN*
Andreas Hoecker - *Atlas-CERN*
Joerg Stelzer - *Atlas-CERN*
Marc Schoenauer - *INRIA*

CLASSIFICATION FOR DISCOVERY

- Official ATLAS GEANT4 simulations
  - 30 features (variables)
  - 250K training: input, label, weight
  - 100K public test (AMS displayed real-time), only input
  - 450K private test (to determine the winner after the closing of the challenge), only input
  - public and private tests are shuffled, participants submit a vector of 550K labels



Higgs challenge

the HiggsML challenge

May to September 2014

When High Energy Physics meets Machine Learning

info to participate and compete : https://www.kaggle.com/c/higgs-boson

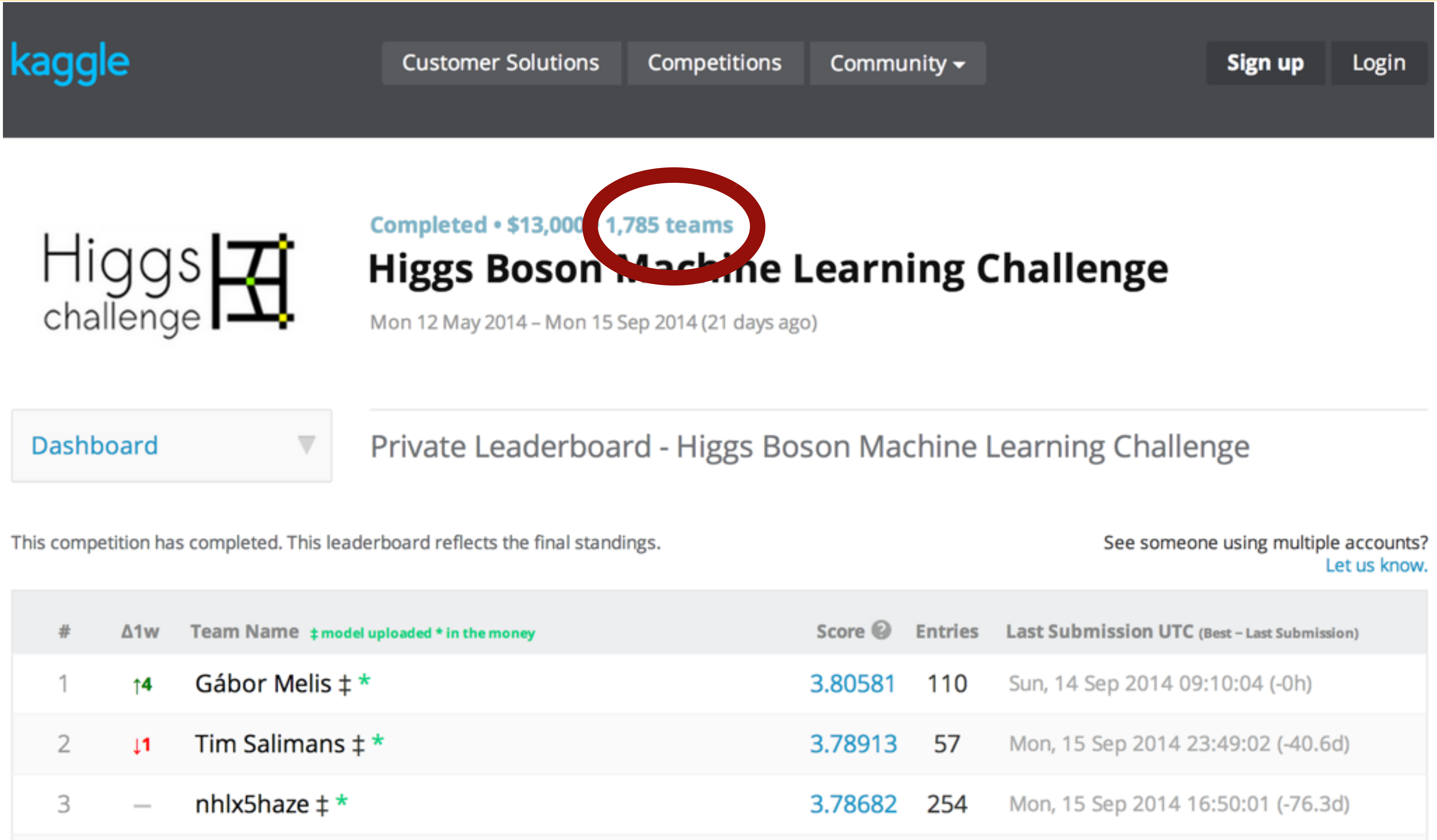
Organization committee

Balázs Kégl - *Appstat-LAL*
Cécile Germain - *TAO-LRI*
David Rousseau - *Atlas-LAL*
Glen Cowan - *Atlas-RHUL*
Isabelle Guyon - *Chalearn*
Claire Adam-Bourdarios - *Atlas-LAL*

Advisory committee

Thorsten Wengler - *Atlas-CERN*
Andreas Hoecker - *Atlas-CERN*
Joerg Stelzer - *Atlas-CERN*
Marc Schoenauer - *INRIA*

CLASSIFICATION FOR DISCOVERY



Completed • $13,000 • 1,785 teams

Higgs Boson Machine Learning Challenge

Mon 12 May 2014 – Mon 15 Sep 2014 (21 days ago)

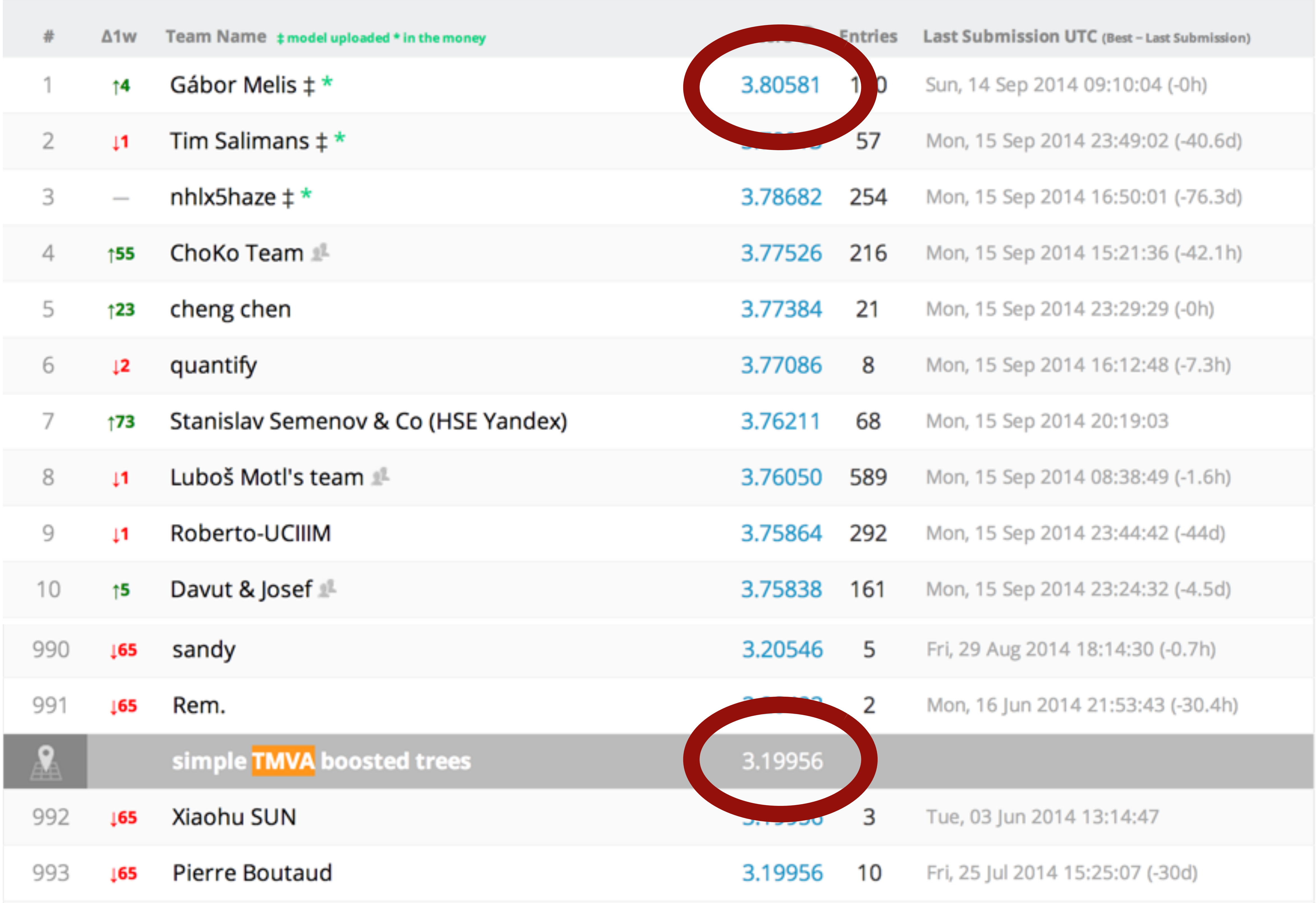
Private Leaderboard - Higgs Boson Machine Learning Challenge

This competition has completed. This leaderboard reflects the final standings.

See someone using multiple accounts?
Let us know.

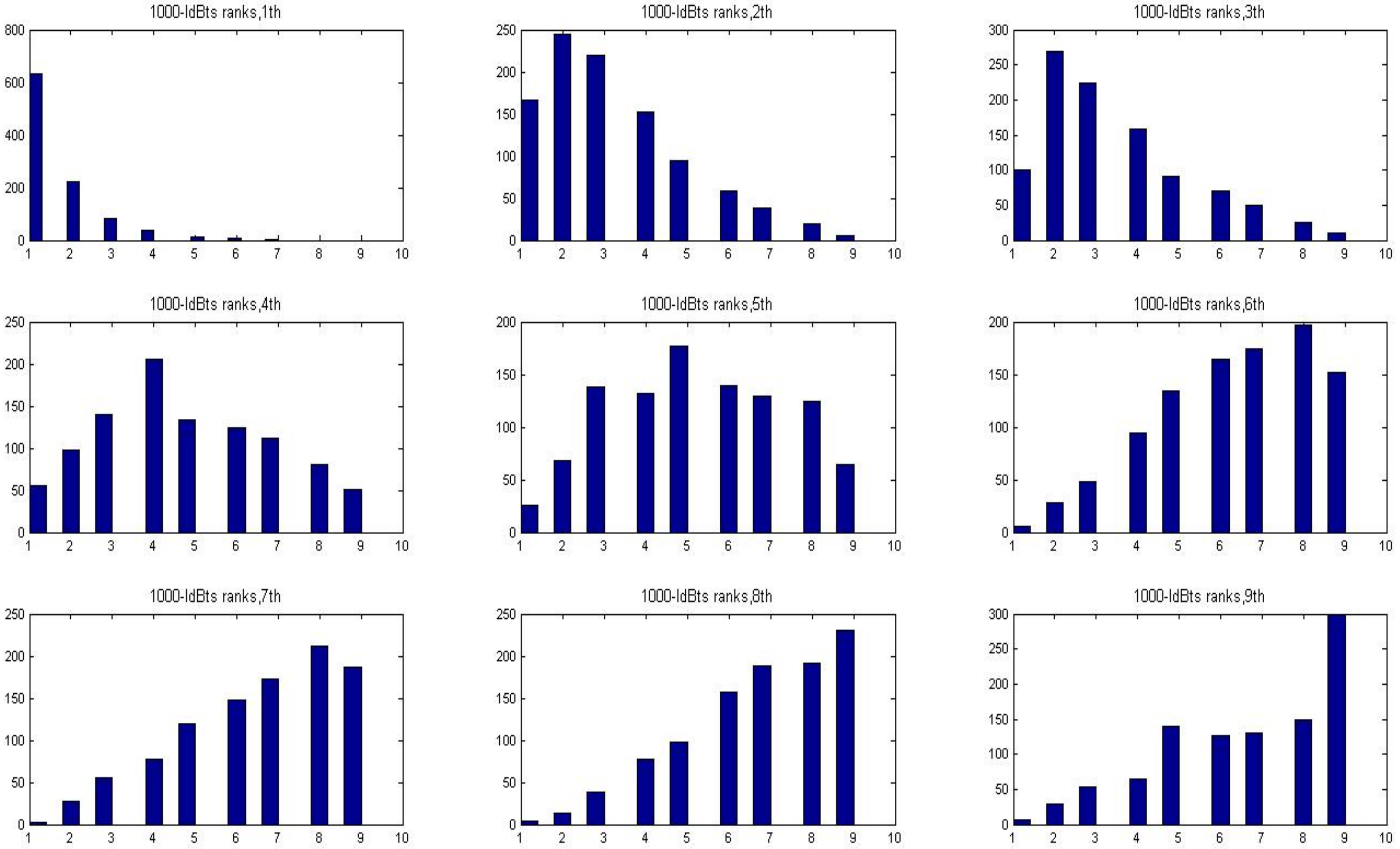
#	Δ1w	Team Name ‡ model uploaded * in the money	Score	Entries	Last Submission UTC (Best – Last Submission)
1	↑4	Gábor Melis ‡ *	3.80581	110	Sun, 14 Sep 2014 09:10:04 (-0h)
2	↓1	Tim Salimans ‡ *	3.78913	57	Mon, 15 Sep 2014 23:49:02 (-40.6d)
3	—	nhlx5haze ‡ *	3.78682	254	Mon, 15 Sep 2014 16:50:01 (-76.3d)

Classification for discovery

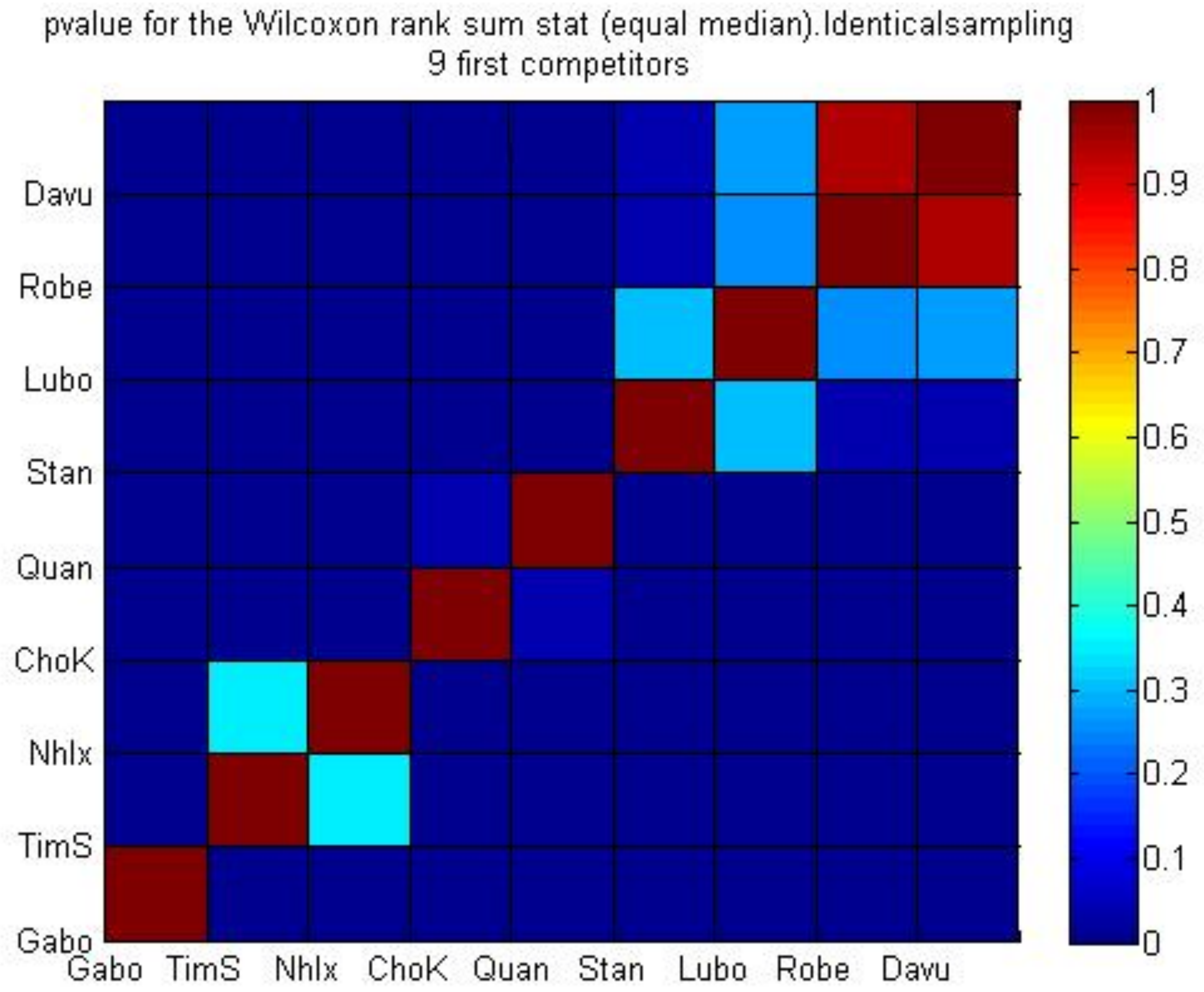


#	Δ1w	Team Name ‡ model uploaded * in the money	Score	Entries	Last Submission UTC (Best – Last Submission)
1	↑4	Gábor Melis ‡ *	3.80581	[illegible]	Sun, 14 Sep 2014 09:10:04 (-0h)
2	↓1	Tim Salimans ‡ *	[illegible]	57	Mon, 15 Sep 2014 23:49:02 (-40.6d)
3	—	nhlx5haze ‡ *	3.78682	254	Mon, 15 Sep 2014 16:50:01 (-76.3d)
4	↑55	ChoKo Team	3.77526	216	Mon, 15 Sep 2014 15:21:36 (-42.1h)
5	↑23	cheng chen	3.77384	21	Mon, 15 Sep 2014 23:29:29 (-0h)
6	↓2	quantify	3.77086	8	Mon, 15 Sep 2014 16:12:48 (-7.3h)
7	↑73	Stanislav Semenov & Co (HSE Yandex)	3.76211	68	Mon, 15 Sep 2014 20:19:03
8	↓1	Luboš Motl's team	3.76050	589	Mon, 15 Sep 2014 08:38:49 (-1.6h)
9	↓1	Roberto-UCIIIM	3.75864	292	Mon, 15 Sep 2014 23:44:42 (-44d)
10	↑5	Davut & Josef	3.75838	161	Mon, 15 Sep 2014 23:24:32 (-4.5d)
990	↓65	sandy	3.20546	5	Fri, 29 Aug 2014 18:14:30 (-0.7h)
991	↓65	Rem.	[illegible]	2	Mon, 16 Jun 2014 21:53:43 (-30.4h)
		simple TMVA boosted trees	3.19956		
992	↓65	Xiaohu SUN	3.19956	3	Tue, 03 Jun 2014 13:14:47
993	↓65	Pierre Boutaud	3.19956	10	Fri, 25 Jul 2014 15:25:07 (-30d)

ARE THE WINNING SCORES SIGNIFICANTLY DIFFERENT?



ARE THE WINNING SCORES SIGNIFICANTLY DIFFERENT?

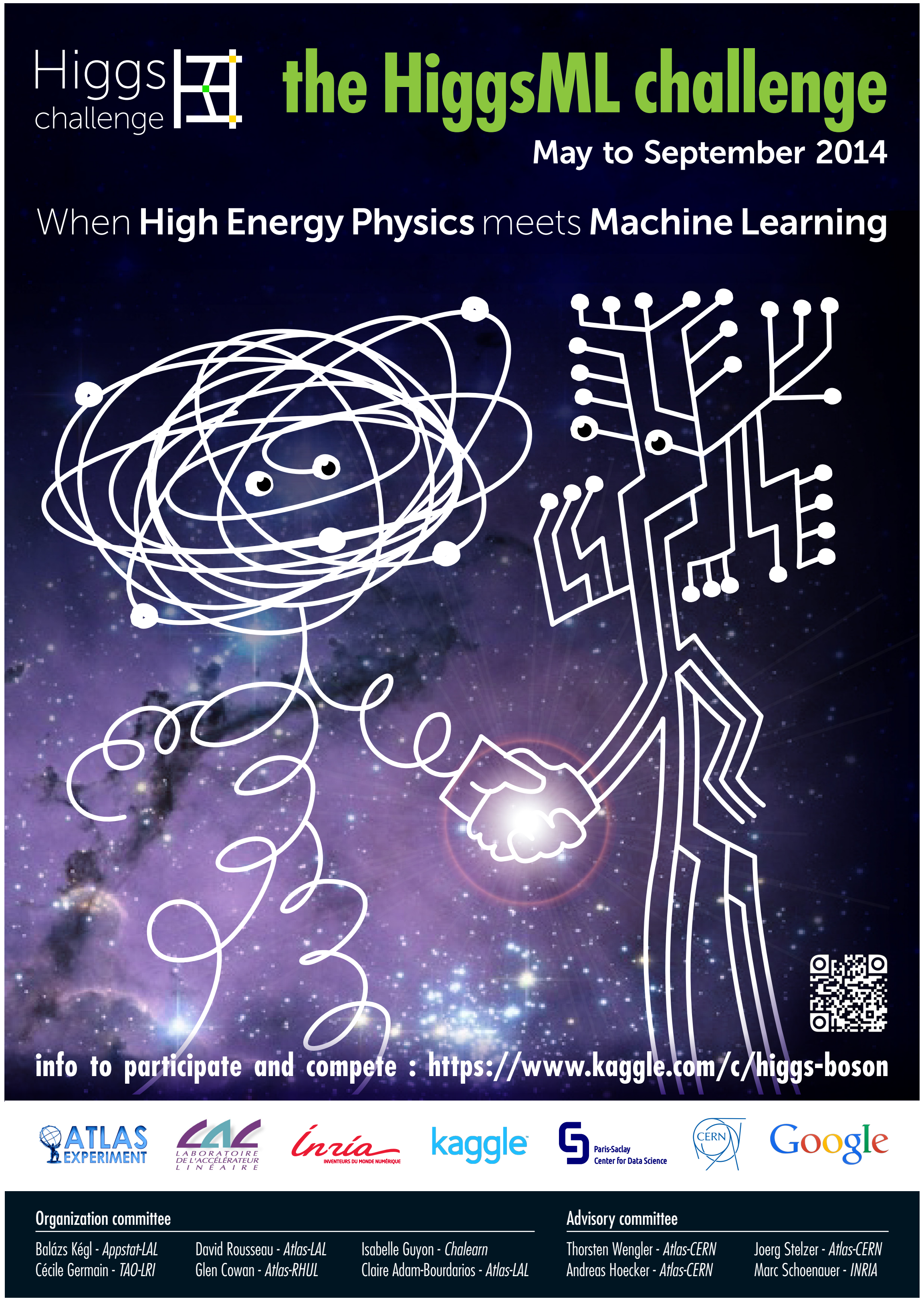


Classification for discovery

- 18 months to organize it

- 4 months to run it

- ?? months to transfer to HEP what we learned



Higgs challenge

the HiggsML challenge

May to September 2014

When High Energy Physics meets Machine Learning

info to participate and compete : https://www.kaggle.com/c/higgs-boson

Organization committee

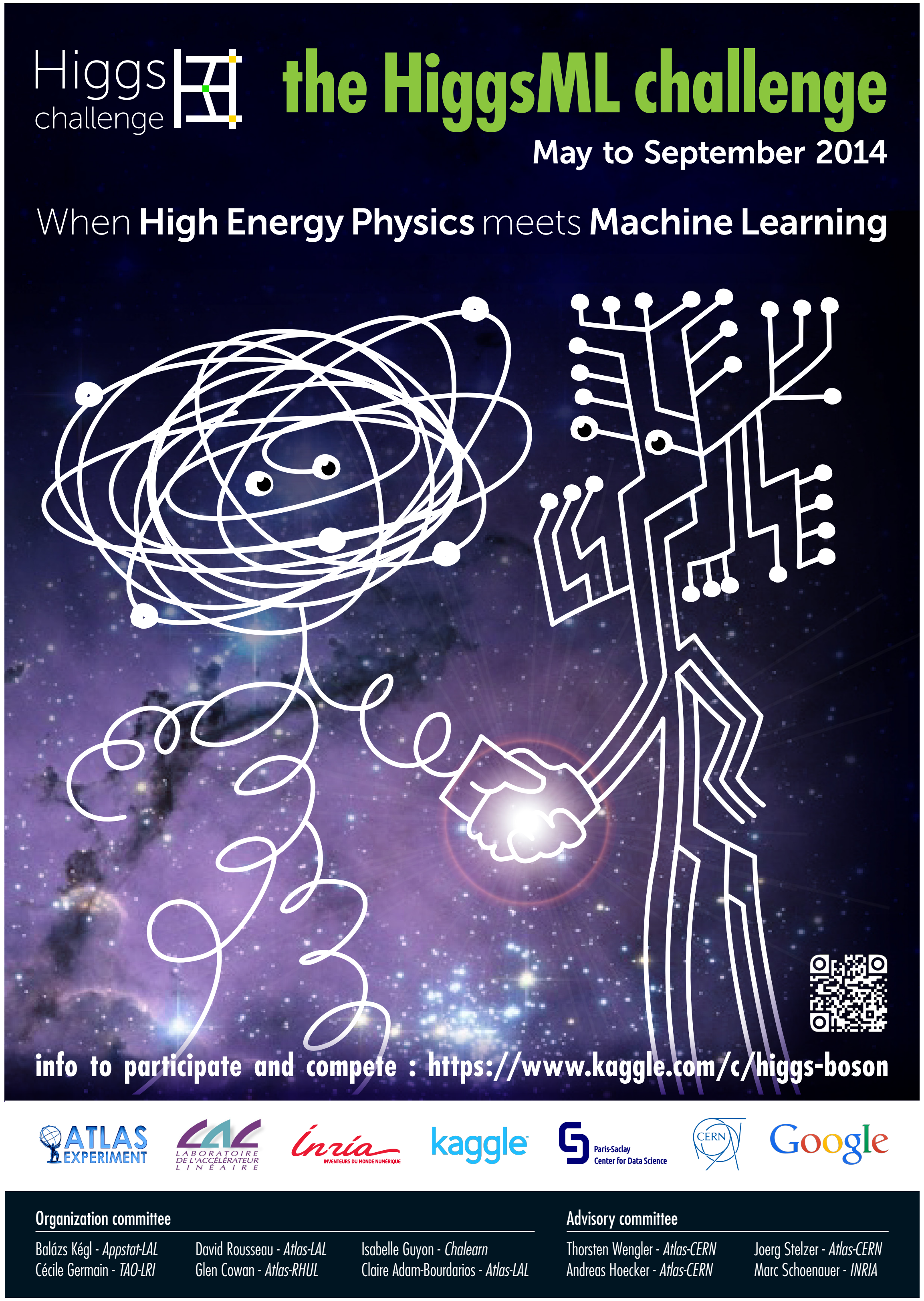
Balázs Kégl - *Appstat-LAL*
Cécile Germain - *TAO-LRI*
David Rousseau - *Atlas-LAL*
Glen Cowan - *Atlas-RHUL*
Isabelle Guyon - *Chalearn*
Claire Adam-Bourdarios - *Atlas-LAL*

Advisory committee

Thorsten Wengler - *Atlas-CERN*
Andreas Hoecker - *Atlas-CERN*
Joerg Stelzer - *Atlas-CERN*
Marc Schoenauer - *INRIA*

What have we learned so far?

- **Neural nets** (dropout, RLU, etc.) rule (although no slam dunk)
- **Ensemble methods** (random forest, boosting) rule
- **Meta-ensembles** of diverse models rule
- 800K points is small for this task: **careful cross-validation** rules



Higgs challenge

the HiggsML challenge

May to September 2014

When **High Energy Physics** meets **Machine Learning**

info to participate and compete : https://www.kaggle.com/c/higgs-boson

ATLAS EXPERIMENT · LAL LABORATOIRE DE L'ACCÉLÉRATEUR LINÉAIRE · Inria INVENTEURS DU MONDE NUMÉRIQUE · kaggle · Paris-Saclay Center for Data Science · CERN · Google

Organization committee

Balázs Kégl - *Appstat-LAL* · David Rousseau - *Atlas-LAL* · Isabelle Guyon - *Chalearn*
Cécile Germain - *TAO-LRI* · Glen Cowan - *Atlas-RHUL* · Claire Adam-Bourdarios - *Atlas-LAL*

Advisory committee

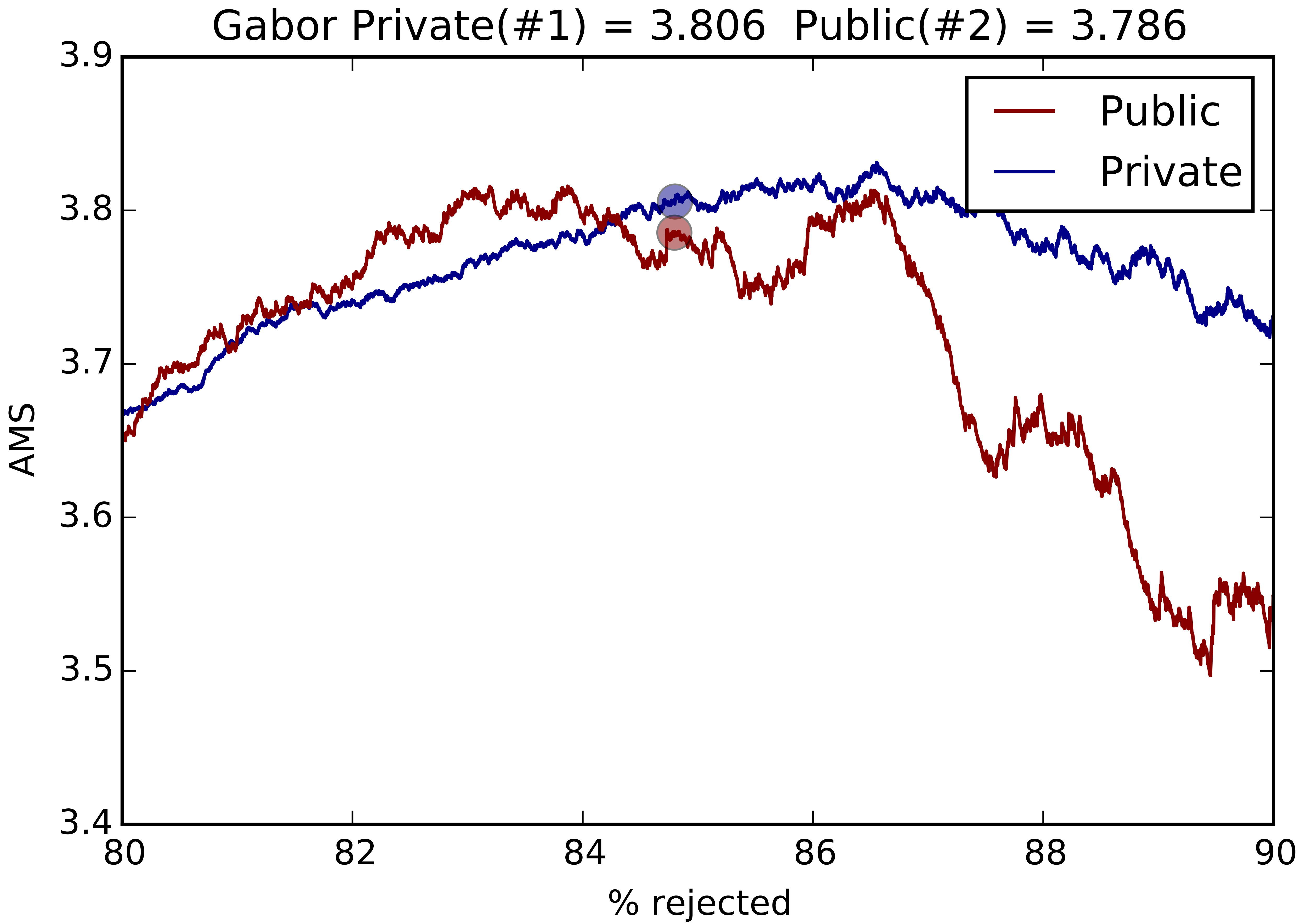
Thorsten Wengler - *Atlas-CERN* · Joerg Stelzer - *Atlas-CERN*
Andreas Hoecker - *Atlas-CERN* · Marc Schoenauer - *INRIA*

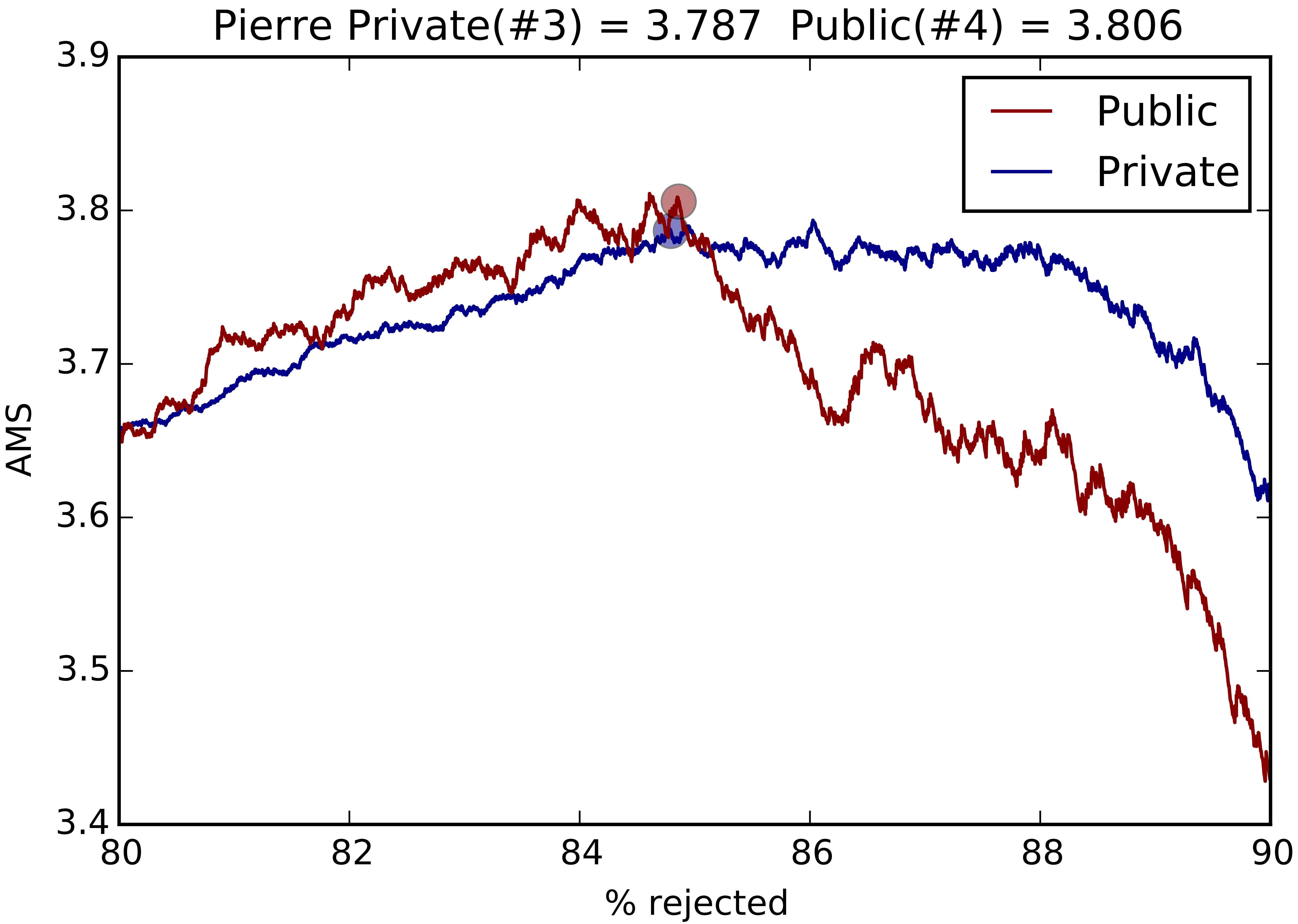
800K IS A SMALL?

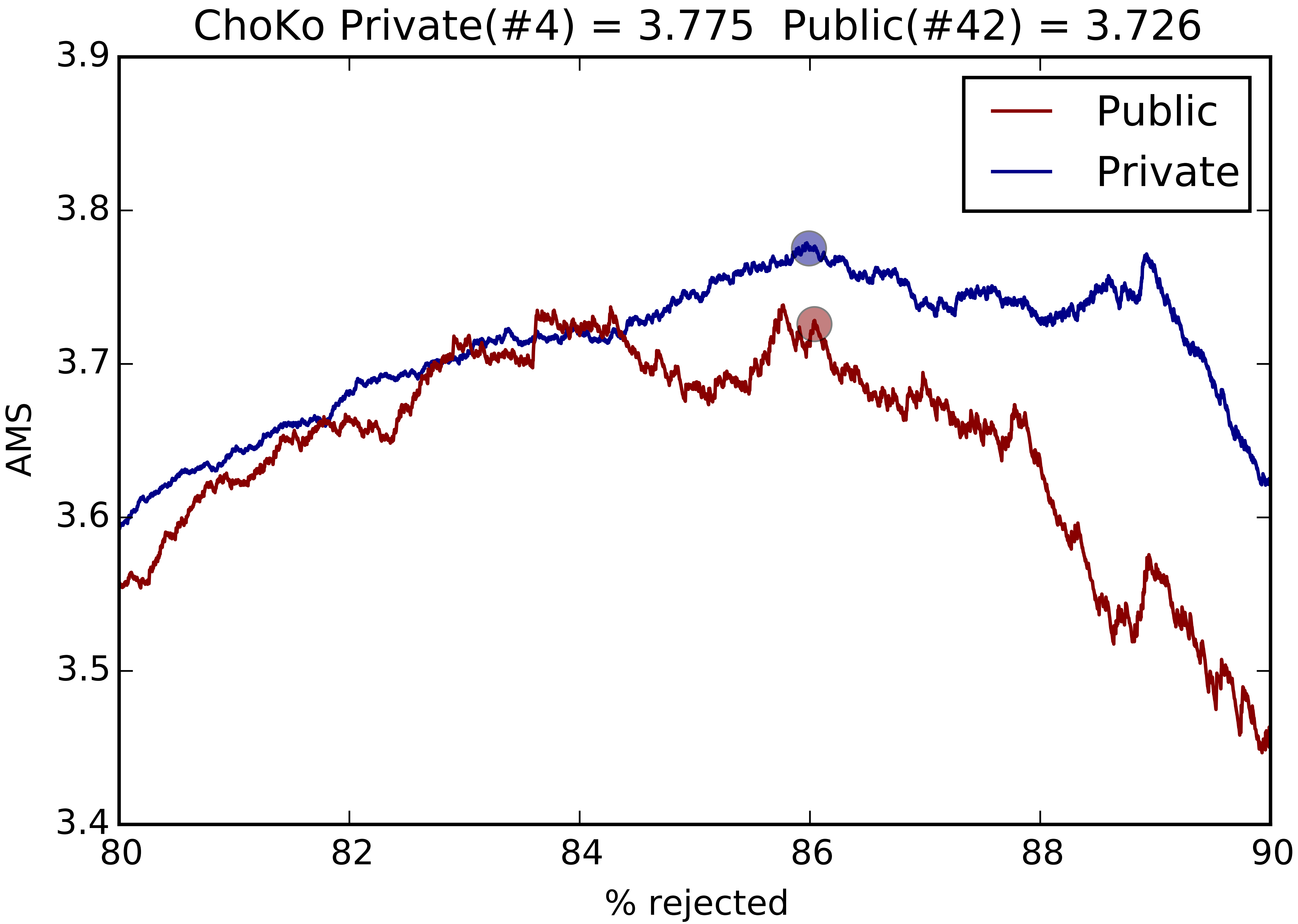
- We asked participants to find good classifiers (in "smooth" AUC sense) but also to **come up with the best selection threshold**
- Optimal region contains about ~15% of the points
- Standard deviation of AMS (given the classifier and the threshold) is about **0.04** (0.08 on the public leaderboard)

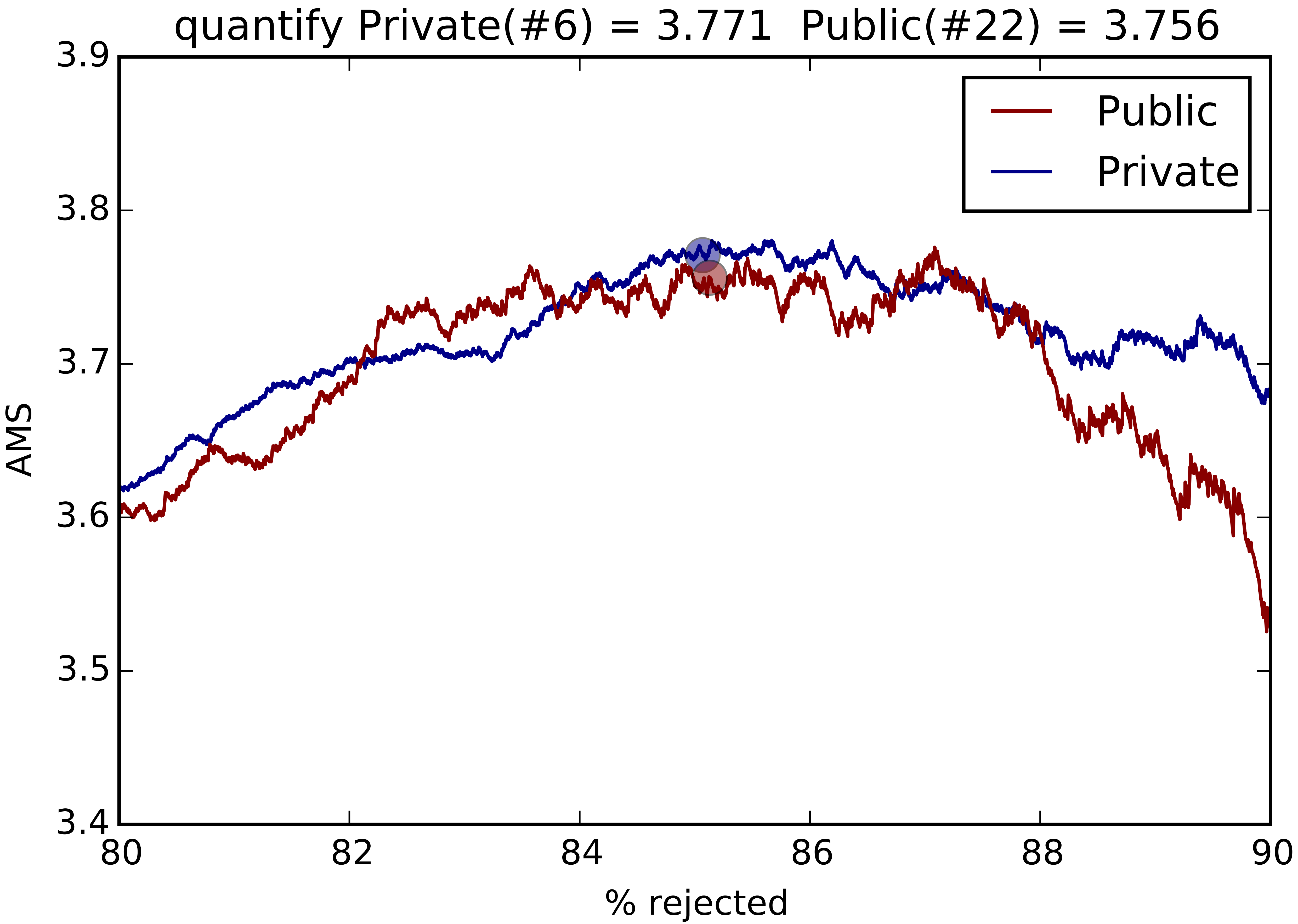
800K IS A SMALL?

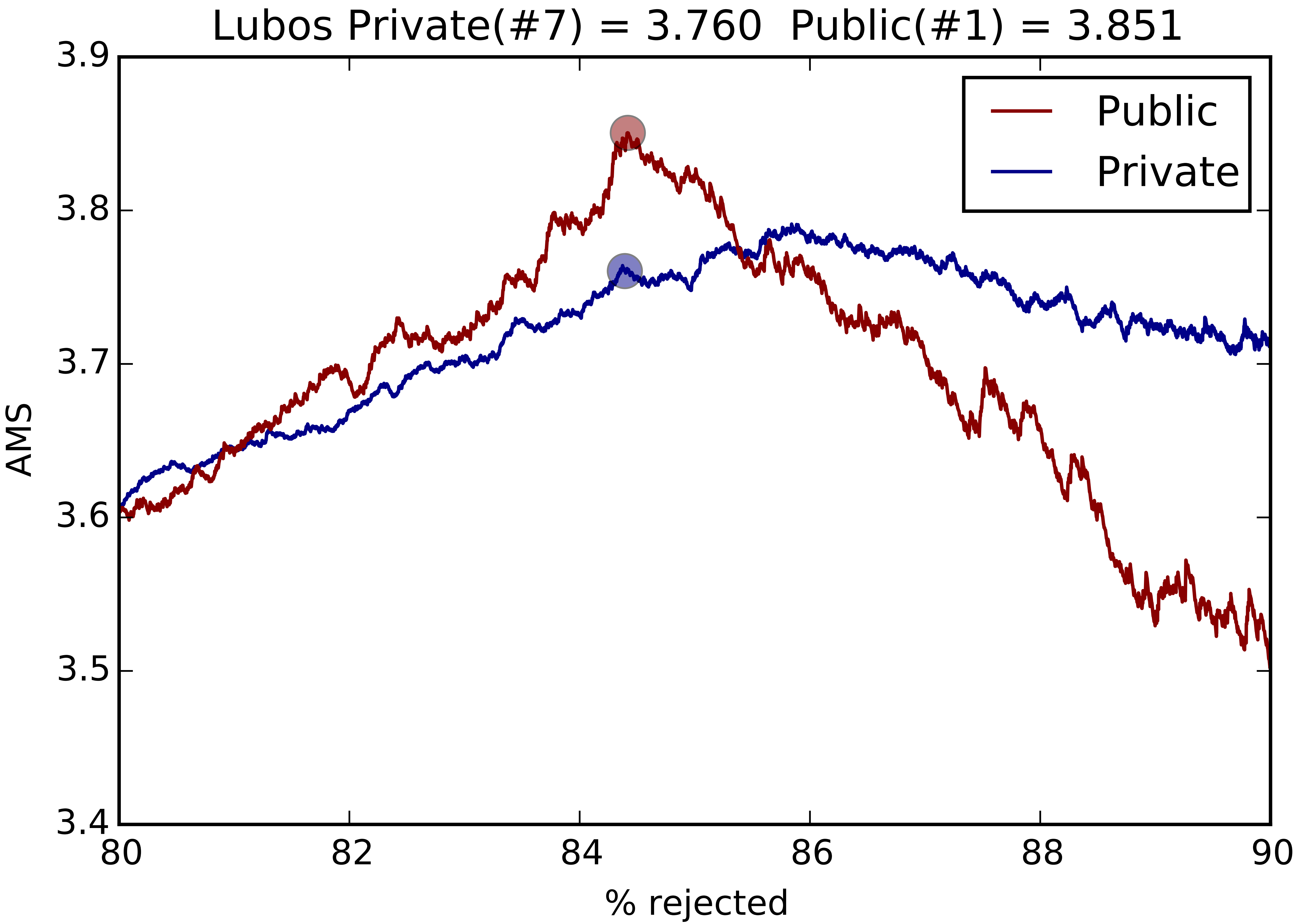
Find the maximum of a noisy diffusion-like process

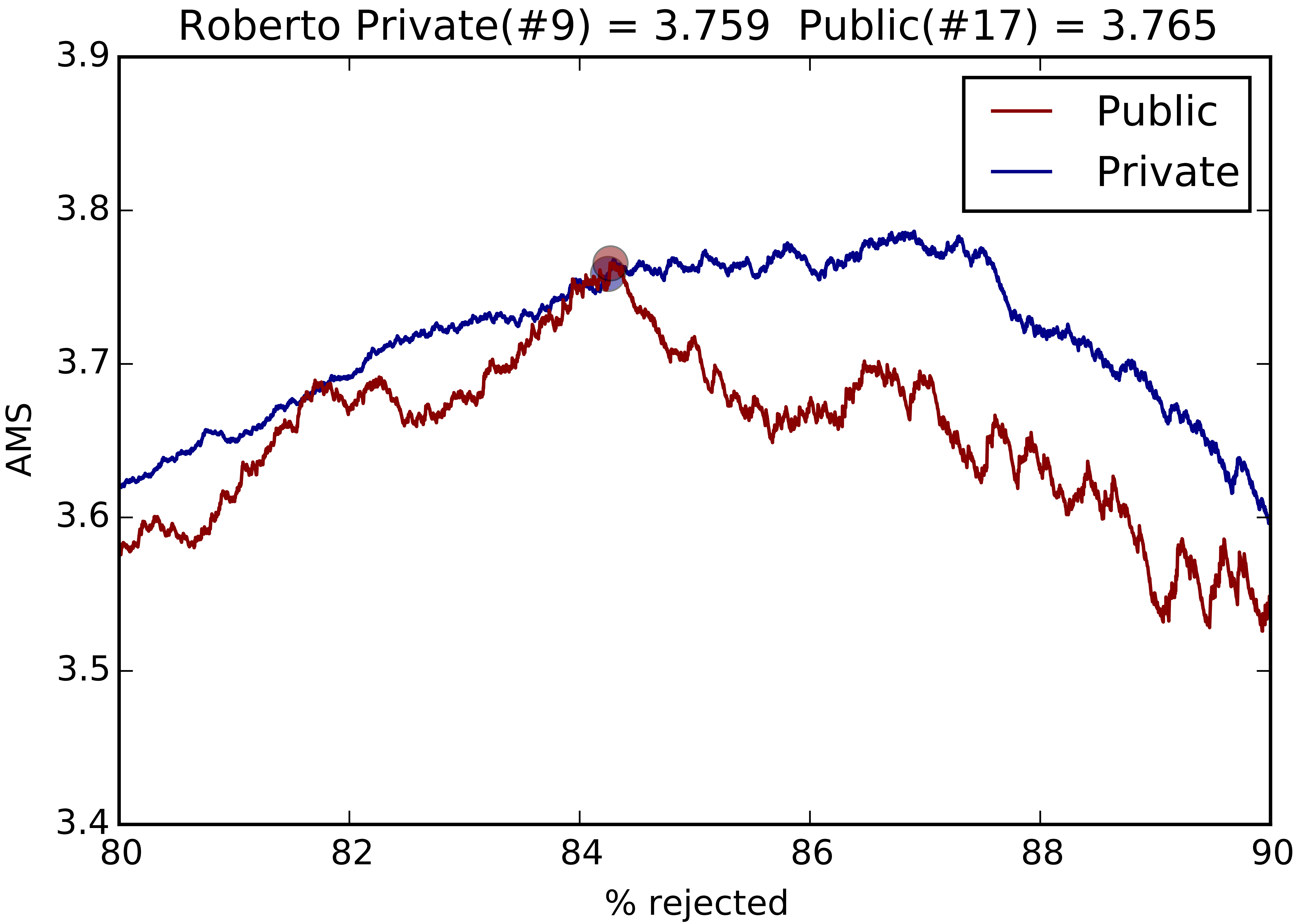

Gabor Private(#1) = 3.806 Public(#2) = 3.786
AMS
3.9
3.8
3.7
3.6
3.5
3.4
80
82
84
86
88
90
% rejected
Public
Private

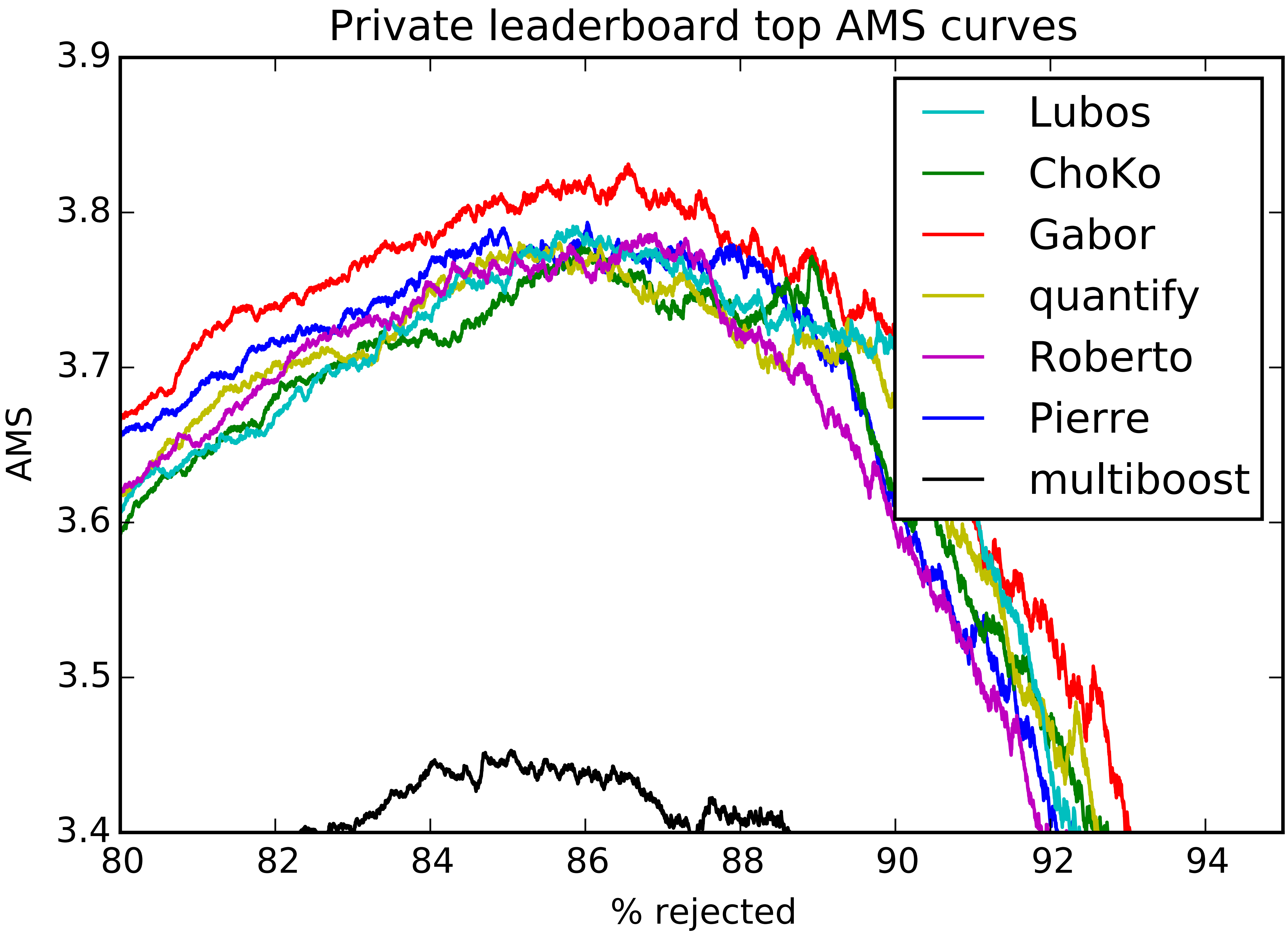

Pierre Private(#3) = 3.787 Public(#4) = 3.806
Public
Private
AMS
3.4
3.5
3.6
3.7
3.8
3.9
% rejected
80
82
84
86
88
90


ChoKo Private(#4) = 3.775 Public(#42) = 3.726
AMS
3.4
3.5
3.6
3.7
3.8
3.9
% rejected
80
82
84
86
88
90
Public
Private

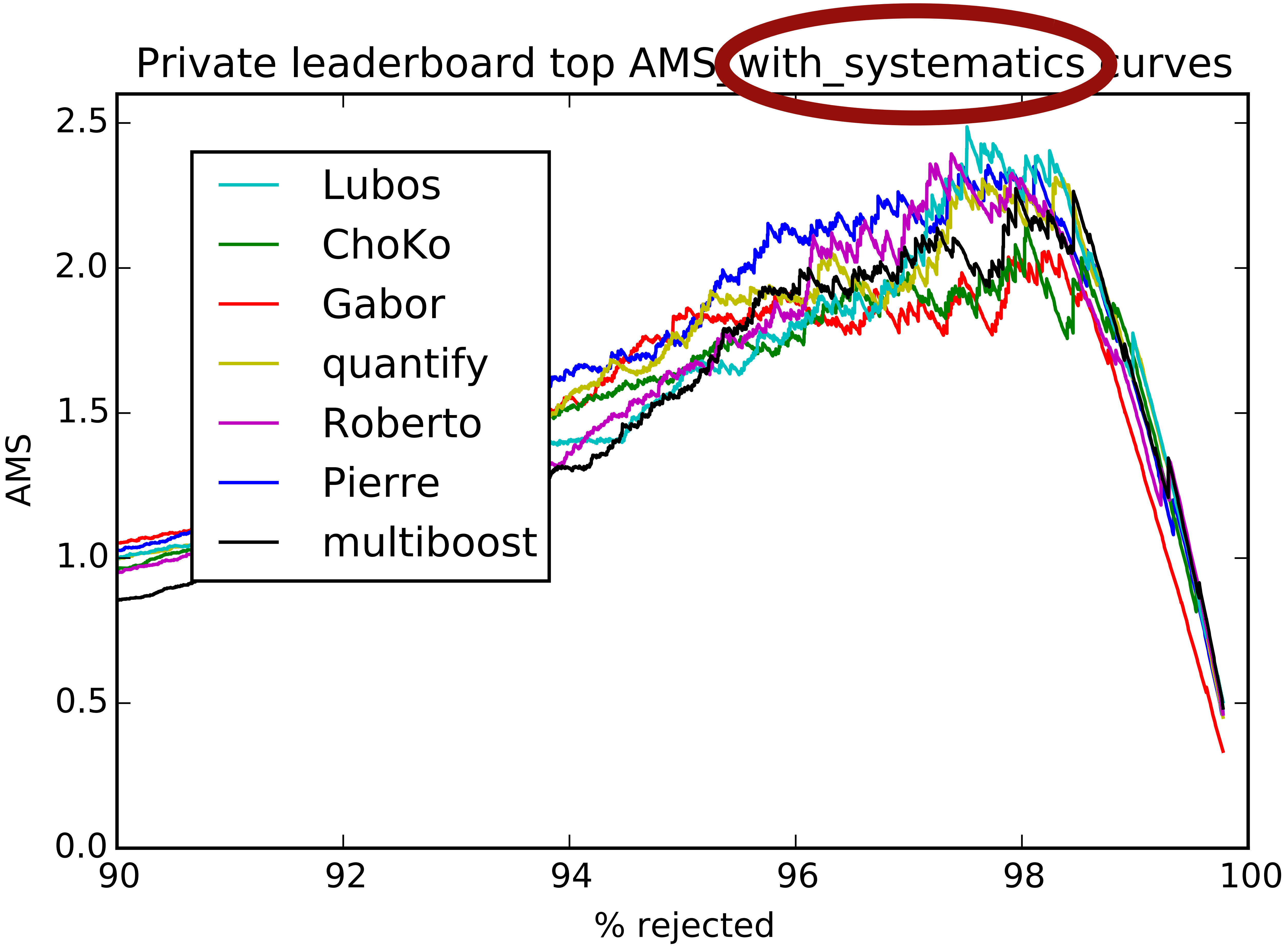

quantify Private(#6) = 3.771 Public(#22) = 3.756
AMS
3.9
3.8
3.7
3.6
3.5
3.4
80
82
84
86
88
90
% rejected
Public
Private


Lubos Private(#7) = 3.760 Public(#1) = 3.851
Public
Private
AMS
% rejected
3.9
3.8
3.7
3.6
3.5
3.4
80
82
84
86
88
90


Roberto Private(#9) = 3.759 Public(#17) = 3.765
Public
Private
AMS
3.9
3.8
3.7
3.6
3.5
3.4
80
82
84
86
88
90
% rejected



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


Private leaderboard top AMS_with_systematics curves
Lubos
ChoKo
Gabor
quantify
Roberto
Pierre
multiboost
AMS
0.0
0.5
1.0
1.5
2.0
2.5
90
92
94
96
98
100
% rejected

META

- A data challenge is a great way to
  - generate visibility
  - human resources
  - optimize a tiny segment of the complete workflow
- Limitations
  - technical constraints (e.g., no server-side execution)
  - sociological constraints (should not be too far from an off-the-shelf problem)
  - emphasizes competition instead of collaboration

Data will be available soon

http://opendata.cern.ch/education/ATLAS

ML IN HIGH-ENERGY PHYSICS

- Budgeted classification for online triggers
- Maximizing the discovery significance and other exotic metrics
- Deep learning for getting closer to raw data
- How to be robust to systematic errors

HEPML workshop at NIPS14

Saturday 13 December 2014

Session 1 - Level 5, room 511 c (08:30-10:00)

time	title	presenter
08:30	Welcome (00h15')	KÉGL, Balázs
08:45	HEP&ML and the HiggsML challenge (00h35')	KÉGL, Balázs
09:20	Embedding ML in Classical Statistical tests used in HEP (invited talk) (00h40')	CRANMER, Kyle

Coffee break - Level 5, room 511 c (10:00-10:30)

Session 2 - Level 5, room 511 c (10:30-12:10)

time	title	presenter
10:30	Presentation of the winner of the HiggsML challenge (00h20')	MELIS, Gábor
10:50	Presentation of the runner up of the HiggsML challenge (00h20')	SALIMANS, Tim
11:10	Presentation of the winner of the HEP meets ML prize (00h20')	CHEN, Tianqi
11:30	Real time data analysis at the LHC : present and future (00h40')	GLIGOROV, Vava

Session 3 - Level 5, room 511 c (15:00-16:30)

time	title	presenter
15:00	Machine Learning for Ultra-High-Energy Physics (invited talk) (00h40')	WHITESON, Daniel
15:40	Weighted Classification Cascades for Optimizing Discovery Significance in the HiggsML Challenge (00h20')	MACKEY, Lester
16:00	Consistent optimization of AMS by logistic loss minimization (00h20')	KOTLOWSKI, Wojciech

Coffee break - Level 5, room 511 c (16:30-17:00)

Session 4 - Level 5, room 511 c (17:00-18:30)

time	title	presenter
17:00	Ensemble of maximied Weighted AUC models for the maximization of the median discovery significance (00h20')	MORALES, Roberto Diaz
17:20	Deep Learning In High-Energy Physics (invited talk) (00h40')	BALDI, Pierre
18:00	Panel discussion (00h30')	