Data challenges and RAMPs

Balázs Kégl
LAL / CNRS

Alexandre Gramfort
LTCI / Telecom ParisTech

Isabelle Guyon
LRI / UPSud

Akin Kazakci
Ecole des Mines

Camille Marini
LTCI / CNRS

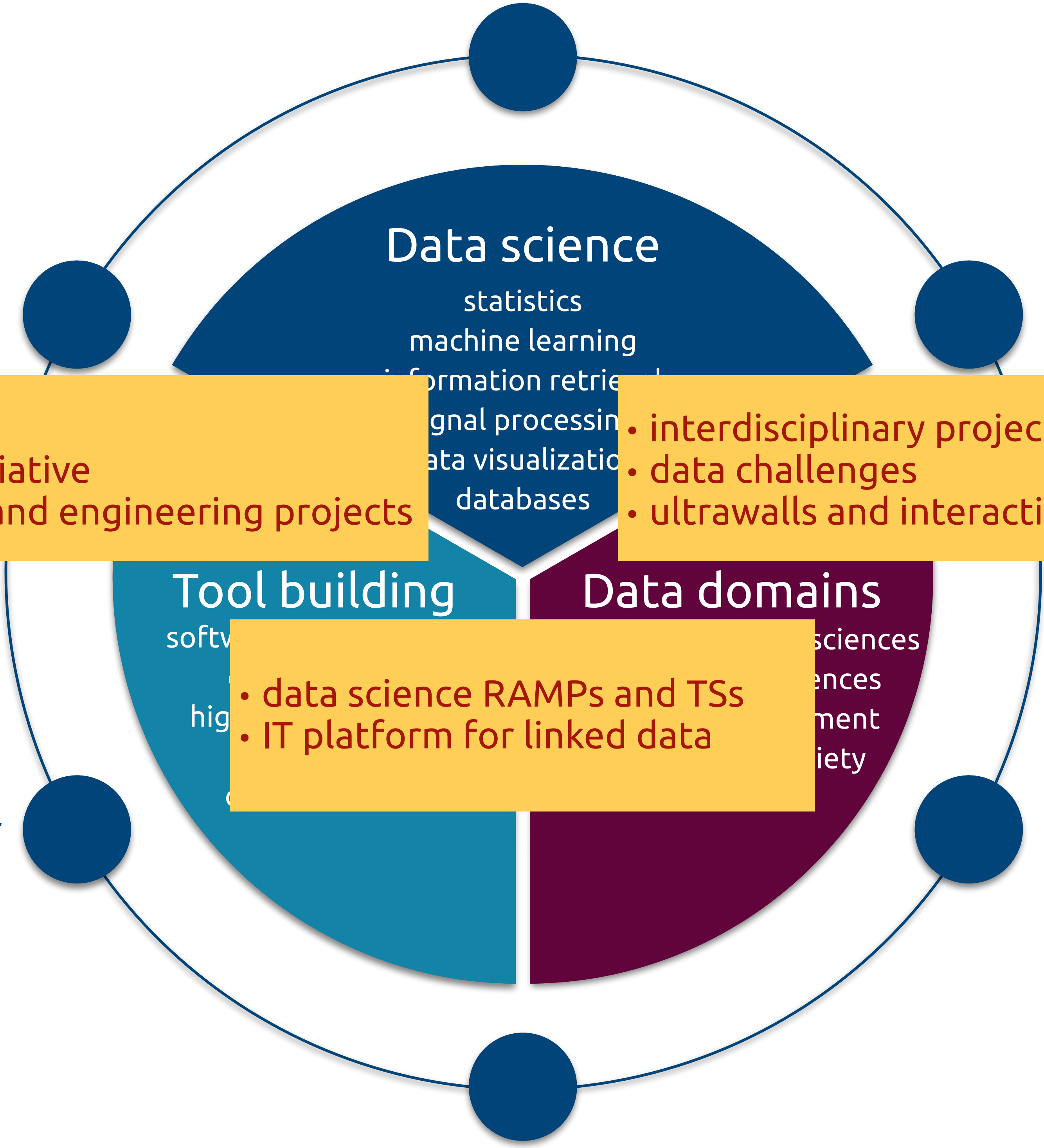
Mehdi Cherti
LAL / CNRS

CDS: A SET OF INNOVATIVE TOOLS AND PROCESSES TO CONNECT COMMUNITIES, TO LAUNCH AND ACCOMPANY PROJECTS

Data scientist

Data engineer

Applied scientist



- coding sprints
- Open Software Initiative
- code consolidator and engineering projects

- interdisciplinary projects
- data challenges
- ultrawalls and interactive visualization

- data science RAMPs and TSs
- IT platform for linked data

Software engineer

Domain expert

Data trainer

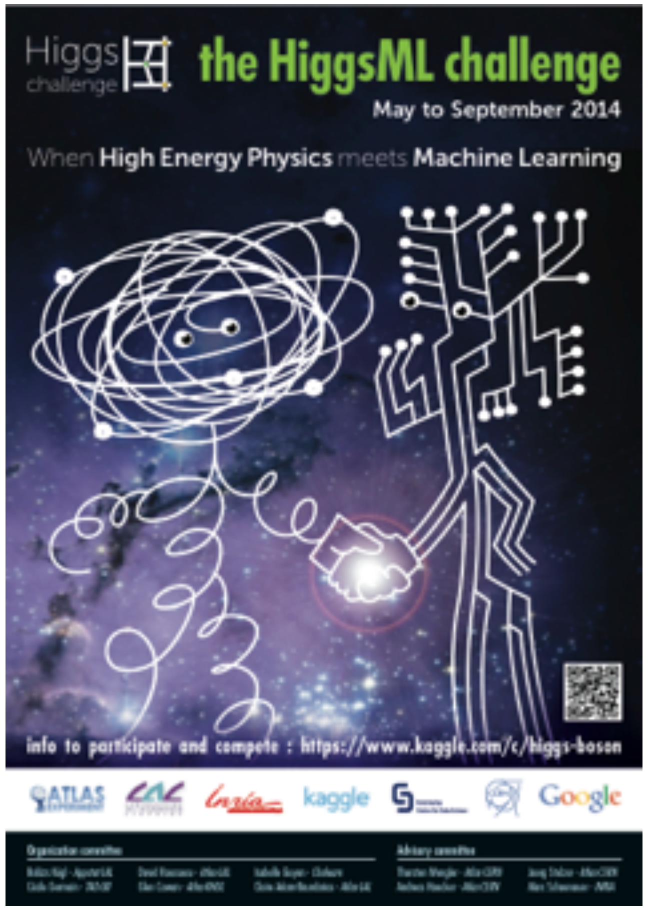
Two analytics tools for initiating domain-data science interactions

Data challenges

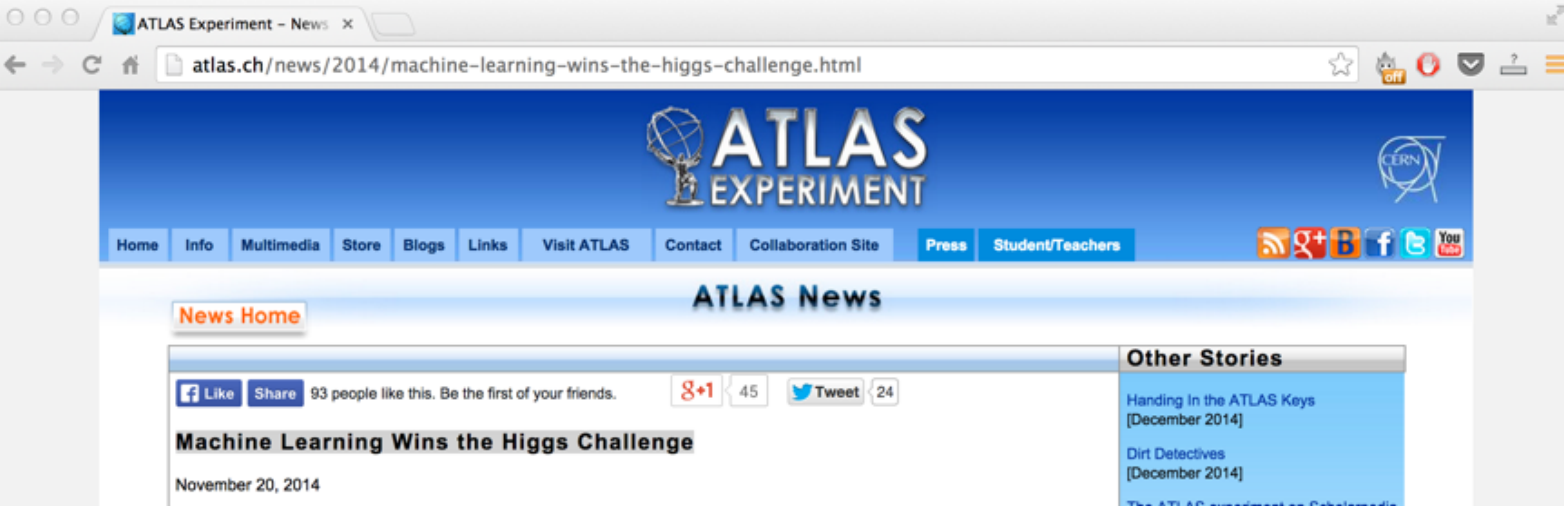
Rapid analytics and model prototyping (RAMP)

DATA CHALLENGES

- A data challenge is a dissemination/communication/crowdsourcing tool
  - a scientific or industrial data producer arrives with a well-defined problem and a corresponding annotated data set
  - defines a quantitative goal
  - makes the problem and part of the data set (the training set) public on a dedicated site
  - data science experts then take the public training data and submit solutions (predictions) for a test set with hidden annotations
  - submissions are evaluated numerically using the quantitative measure
  - contestants are listed on a leaderboard
  - after a predefined time, typically a couple of months, the final results are revealed and the winners are awarded


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
kaggle
Google

DATA CHALLENGES



ATLAS Experiment – News

atlas.ch/news/2014/machine-learning-wins-the-higgs-challenge.html

ATLAS News

93 people like this. Be the first of your friends.

+1 45 · Tweet 24

Machine Learning Wins the Higgs Challenge

November 20, 2014

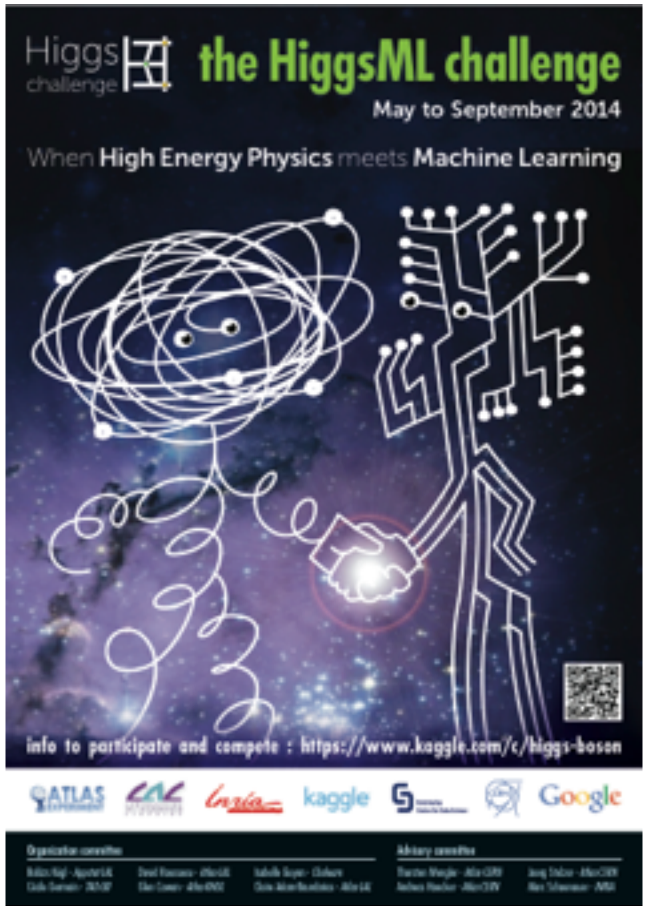
Other Stories

Handing In the ATLAS Keys
[December 2014]

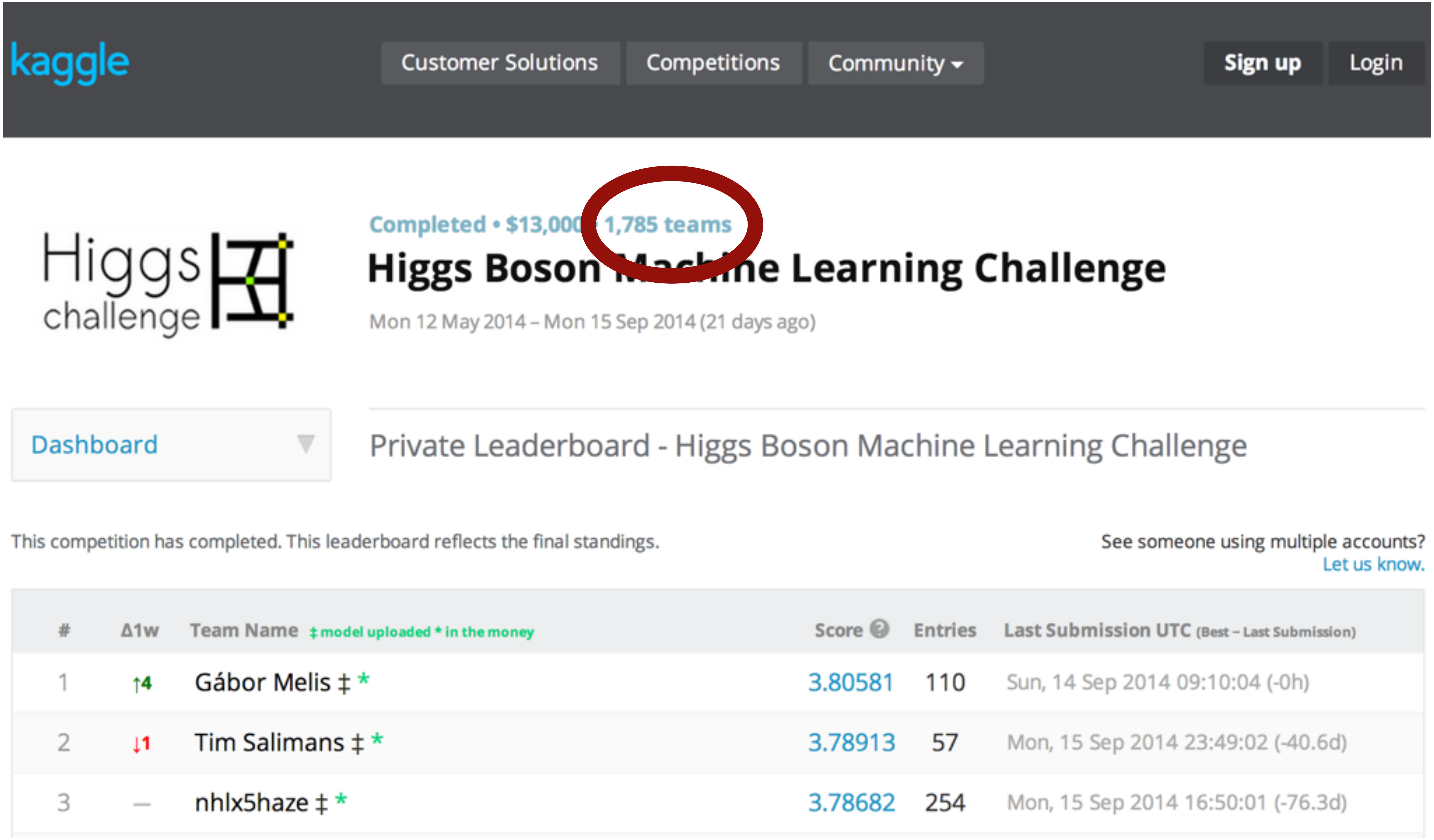
Dirt Detectives
[December 2014]

The ATLAS experiment on Schoolmedia

- The HiggsML challenge on Kaggle
  - https://www.kaggle.com/c/higgs-boson


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
kaggle
Google
Organization committee
Advisory committee

HUGE PUBLICITY



Completed • $13,000 • 1,785 teams

Higgs Boson Machine Learning Challenge

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

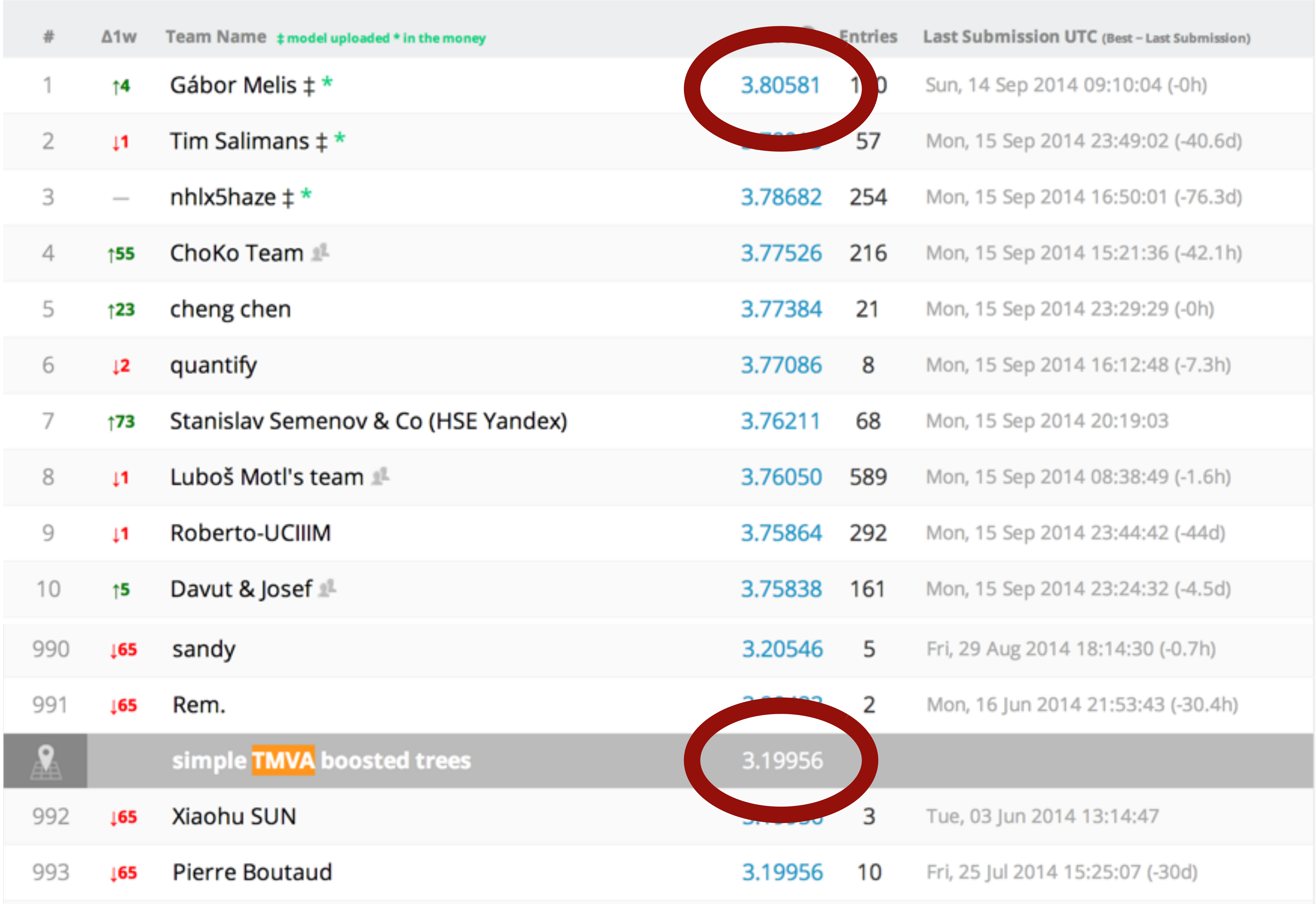
Private Leaderboard - Higgs Boson Machine Learning Challenge

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)

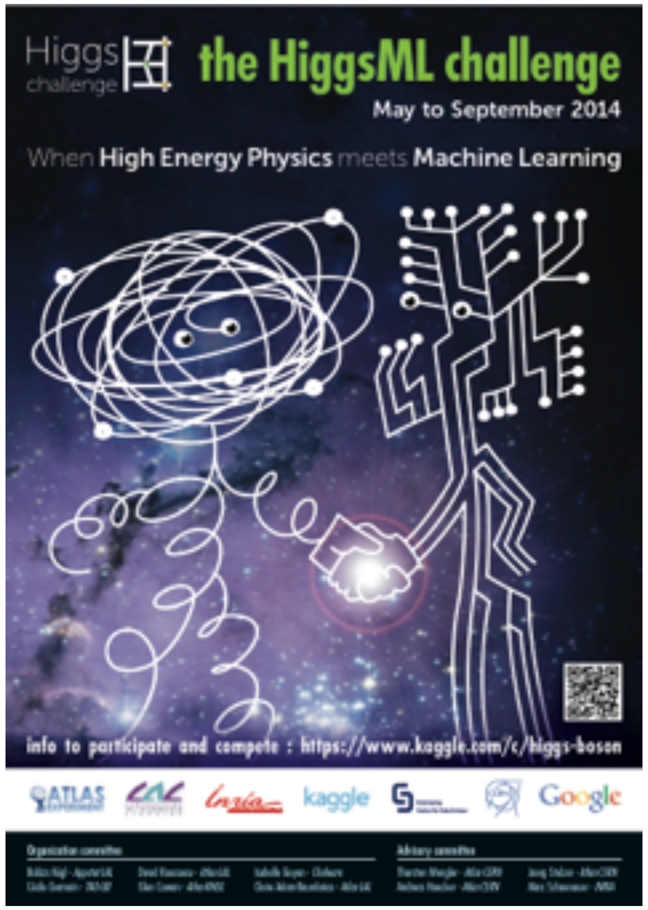
SIGNIFICANT IMPROVEMENT OVER THE BASELINE



#	Δ1w	Team Name ‡ model uploaded * in the money		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	[illegible]	3	Tue, 03 Jun 2014 13:14:47
993	↓65	Pierre Boutaud	3.19956	10	Fri, 25 Jul 2014 15:25:07 (-30d)

DATA CHALLENGES

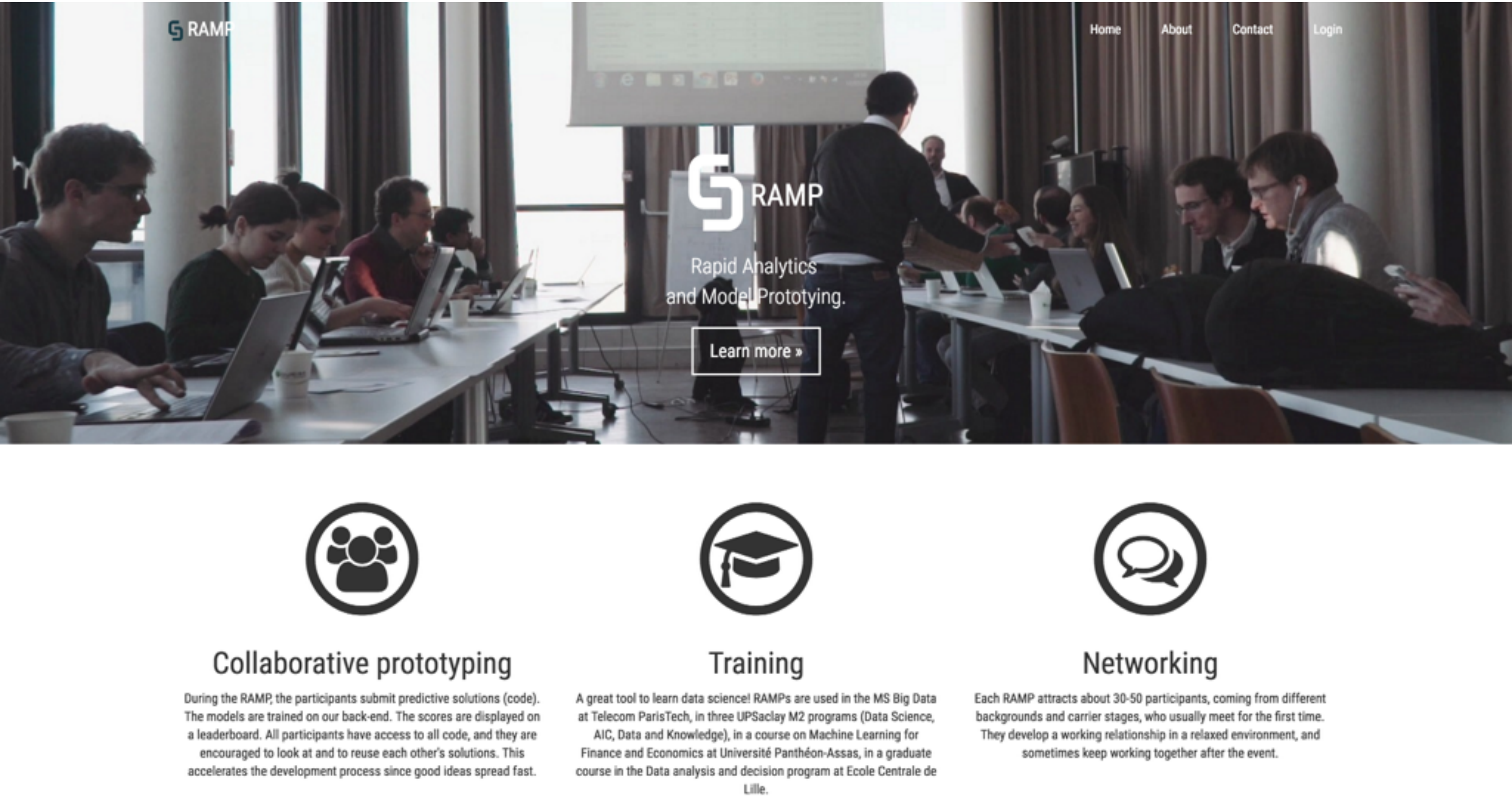
- Challenges are useful for
  - generating visibility in the data science community about novel application domains
  - benchmarking in a fair way state-of-the-art techniques on well-defined problems
  - finding talented data scientists
- Limitations
  - not necessary adapted to solving complex and open-ended data science problems in realistic environments
  - no direct access to solutions and data scientist
  - emphasizes competition



We decided to design something better

Rapid analytics and model prototyping (RAMP)

http://www.ramp.studio



Collaborative prototyping

During the RAMP, the participants submit predictive solutions (code). The models are trained on our back-end. The scores are displayed on a leaderboard. All participants have access to all code, and they are encouraged to look at and to reuse each other's solutions. This accelerates the development process since good ideas spread fast.

Training

A great tool to learn data science! RAMPs are used in the MS Big Data at Telecom ParisTech, in three UPSaclay M2 programs (Data Science, AIC, Data and Knowledge), in a course on Machine Learning for Finance and Economics at Université Panthéon-Assas, in a graduate course in the Data analysis and decision program at Ecole Centrale de Lille.

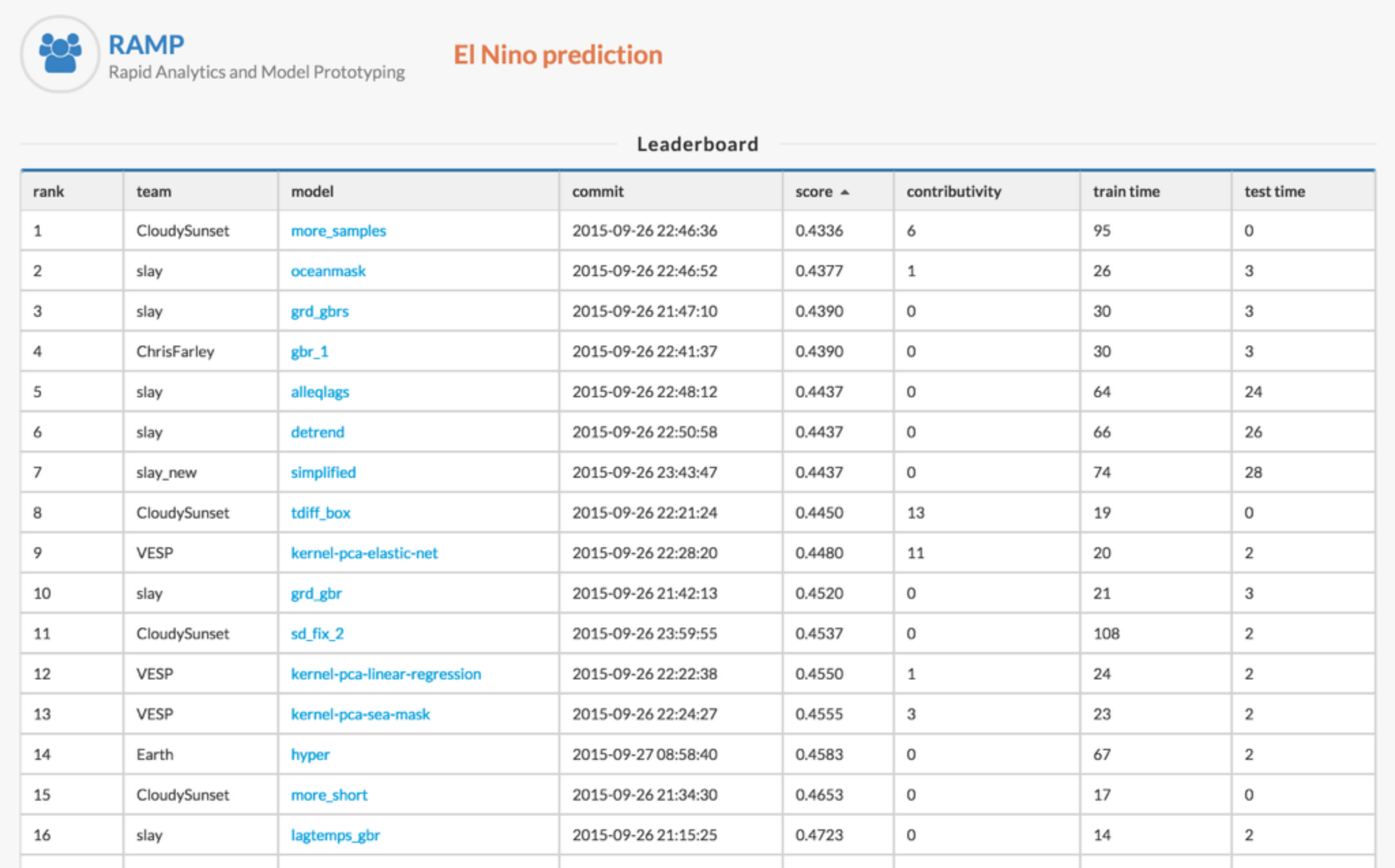
Networking

Each RAMP attracts about 30-50 participants, coming from different backgrounds and carrier stages, who usually meet for the first time. They develop a working relationship in a relaxed environment, and sometimes keep working together after the event.

RAMPs

- Single-day **coding sessions**
  - **20-40** participants
  - **preparation** is similar to challenges
- Goals
  - **focusing** and **motivating** top talents
  - promoting **collaboration**, **speed**, and **efficiency**
  - **solving** (prototyping) **real** problems

ANALYTICS TOOLS TO PROMOTE COLLABORATION AND CODE REUSE



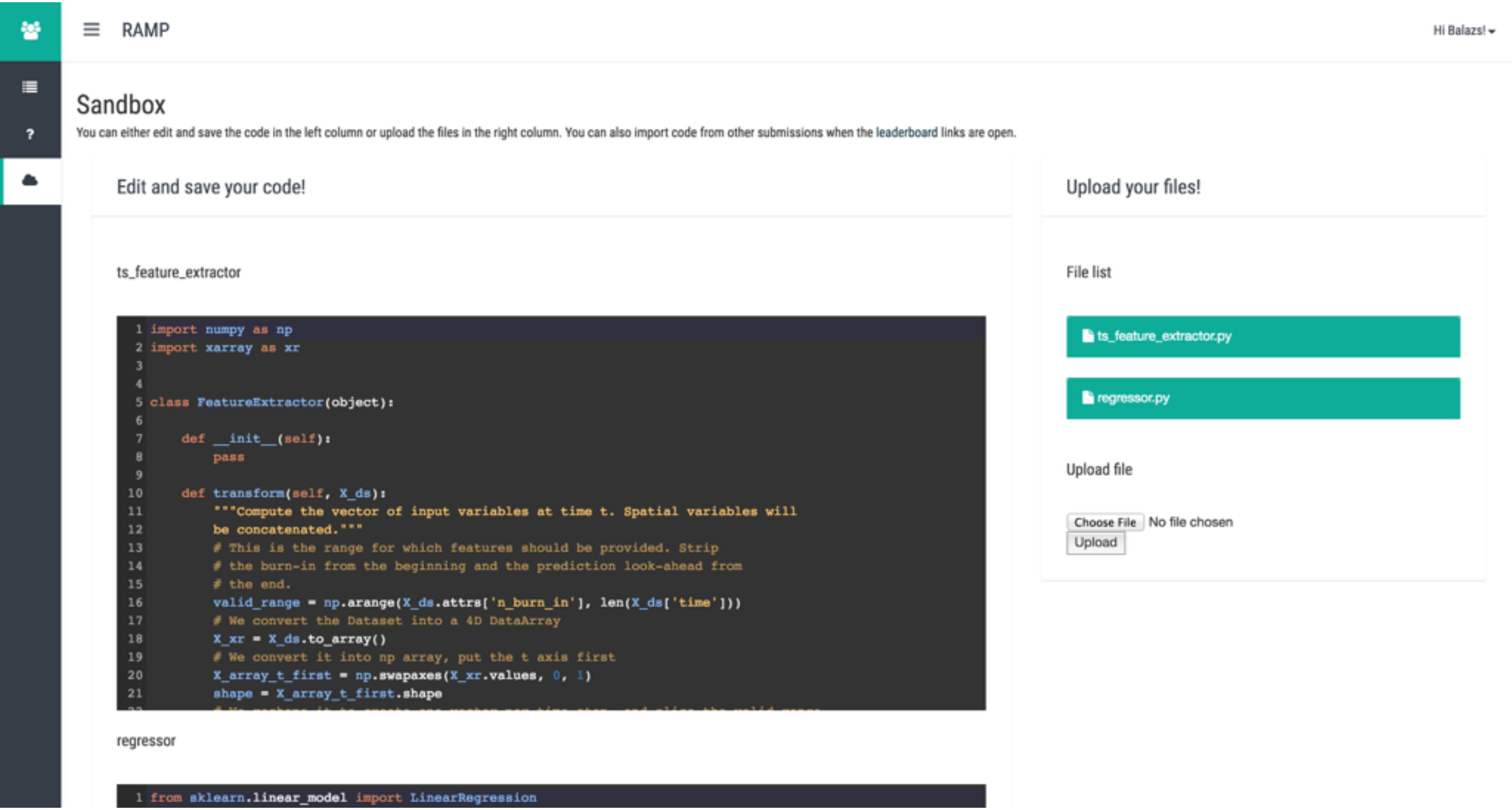
RAMP
Rapid Analytics and Model Prototyping

El Nino prediction

Leaderboard

rank	team	model	commit	score ▲	contributivity	train time	test time
1	CloudySunset	more_samples	2015-09-26 22:46:36	0.4336	6	95	0
2	slay	oceanmask	2015-09-26 22:46:52	0.4377	1	26	3
3	slay	grd_gbrs	2015-09-26 21:47:10	0.4390	0	30	3
4	ChrisFarley	gbr_1	2015-09-26 22:41:37	0.4390	0	30	3
5	slay	alleqlags	2015-09-26 22:48:12	0.4437	0	64	24
6	slay	detrend	2015-09-26 22:50:58	0.4437	0	66	26
7	slay_new	simplified	2015-09-26 23:43:47	0.4437	0	74	28
8	CloudySunset	tdiff_box	2015-09-26 22:21:24	0.4450	13	19	0
9	VESP	kernel-pca-elastic-net	2015-09-26 22:28:20	0.4480	11	20	2
10	slay	grd_gbr	2015-09-26 21:42:13	0.4520	0	21	3
11	CloudySunset	sd_fix_2	2015-09-26 23:59:55	0.4537	0	108	2
12	VESP	kernel-pca-linear-regression	2015-09-26 22:22:38	0.4550	1	24	2
13	VESP	kernel-pca-sea-mask	2015-09-26 22:24:27	0.4555	3	23	2
14	Earth	hyper	2015-09-27 08:58:40	0.4583	0	67	2
15	CloudySunset	more_short	2015-09-26 21:34:30	0.4653	0	17	0
16	slay	lagtemps_gbr	2015-09-26 21:15:25	0.4723	0	14	2

ANALYTICS TOOL TO PROMOTE COLLABORATION AND CODE REUSE



Sandbox

You can either edit and save the code in the left column or upload the files in the right column. You can also import code from other submissions when the leaderboard links are open.

Edit and save your code!

ts_feature_extractor

```python
import numpy as np
import xarray as xr


class FeatureExtractor(object):

    def __init__(self):
        pass

    def transform(self, X_ds):
        """Compute the vector of input variables at time t. Spatial variables will
        be concatenated."""
        # This is the range for which features should be provided. Strip
        # the burn-in from the beginning and the prediction look-ahead from
        # the end.
        valid_range = np.arange(X_ds.attrs['n_burn_in'], len(X_ds['time']))
        # We convert the Dataset into a 4D DataArray
        X_xr = X_ds.to_array()
        # We convert it into np array, put the t axis first
        X_array_t_first = np.swapaxes(X_xr.values, 0, 1)
        shape = X_array_t_first.shape
```

regressor

```python
from sklearn.linear_model import LinearRegression
```

Upload your files!

File list

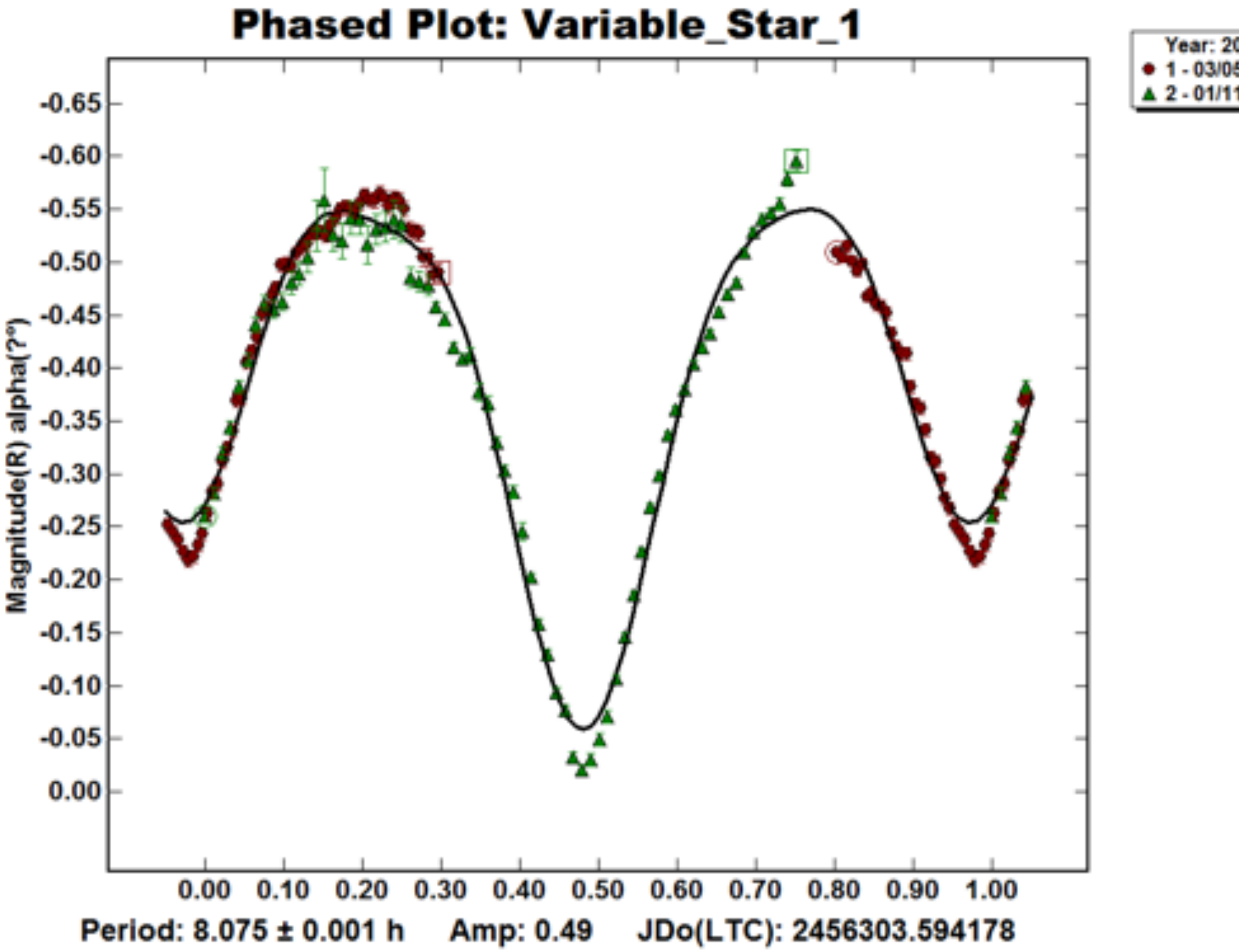
- ts_feature_extractor.py
- regressor.py

Upload file

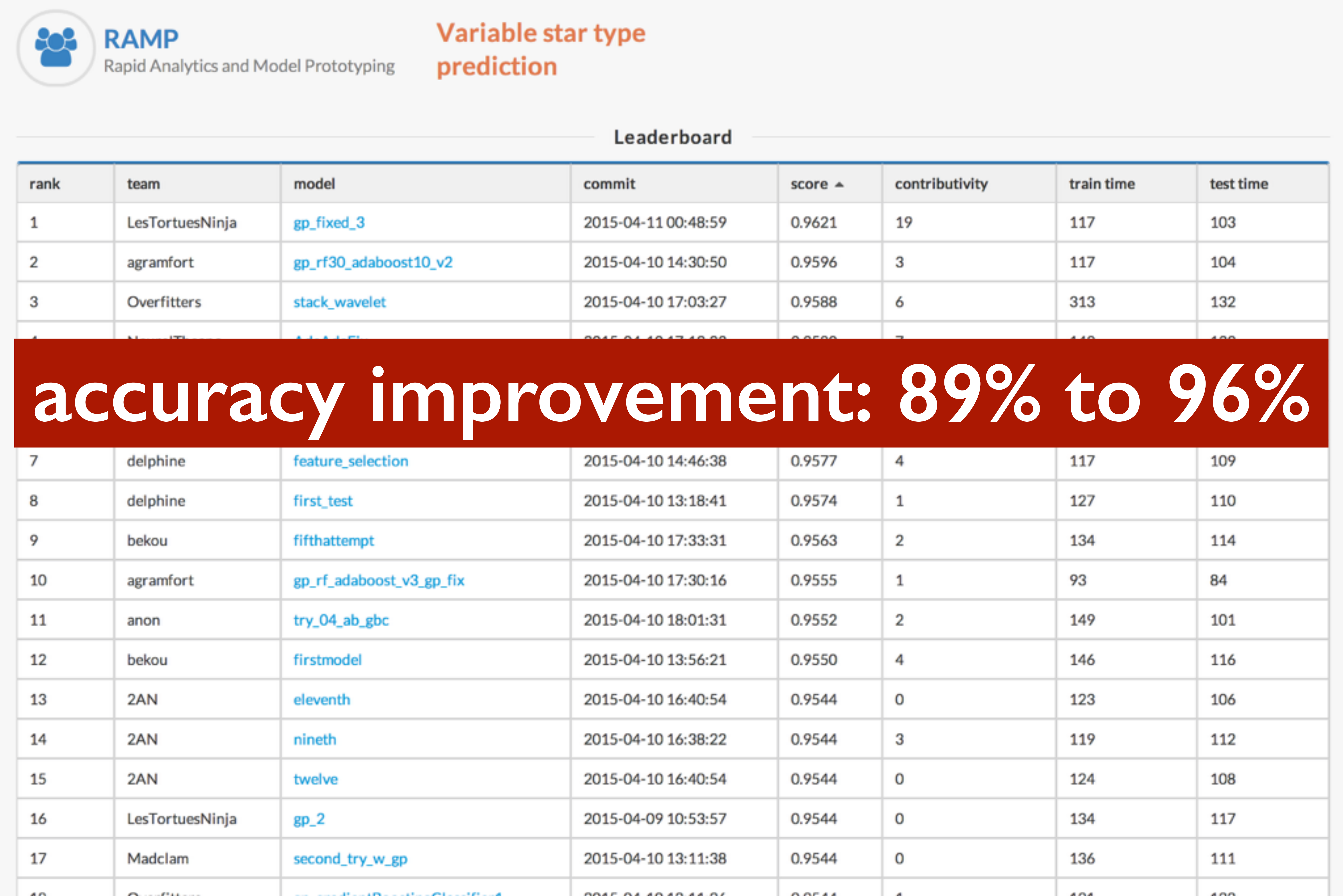
RAPID ANALYTICS AND MODEL PROTOTYPING

2015 Apr 10

Classifying variable stars



VARIABLE STARS



RAMP
Rapid Analytics and Model Prototyping

Variable star type prediction

Leaderboard

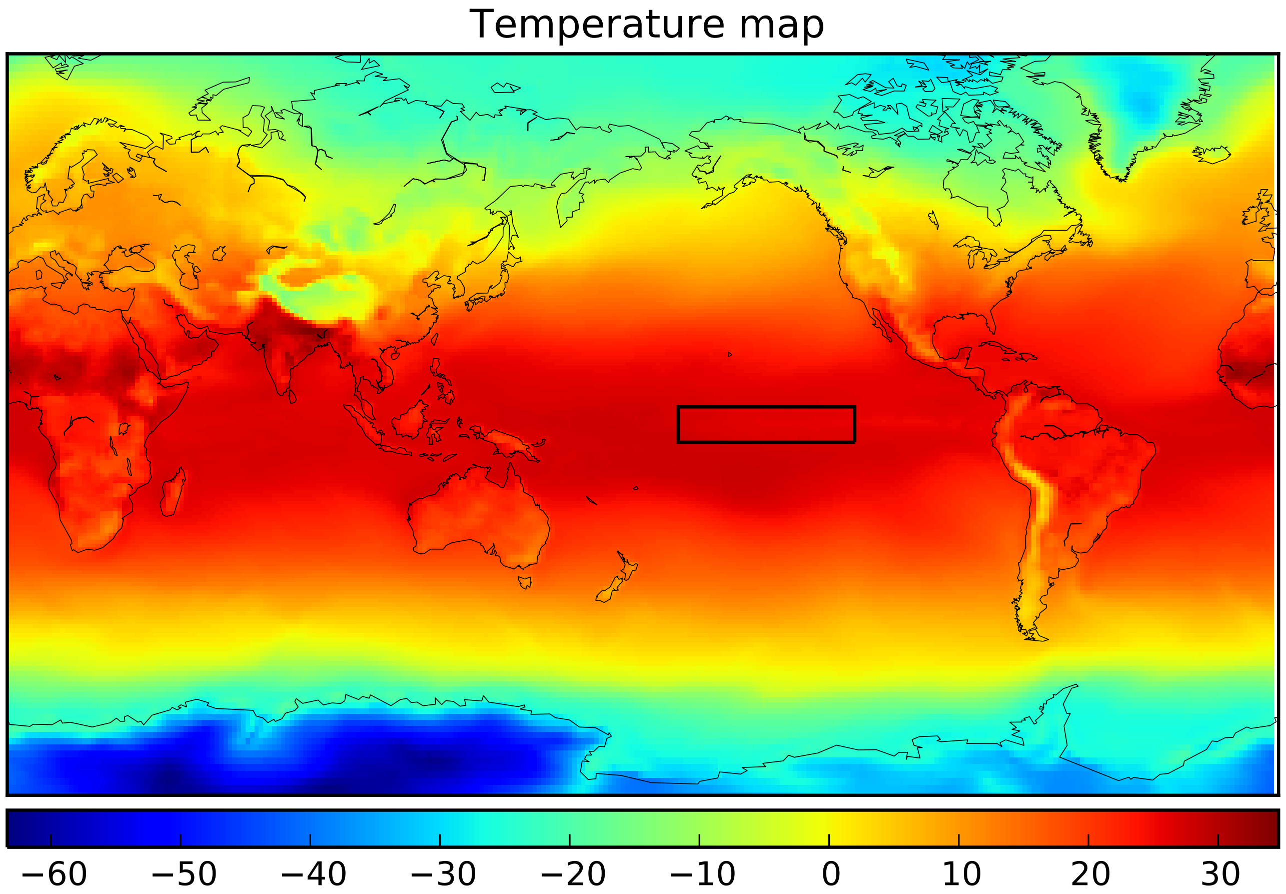
rank	team	model	commit	score ▲	contributivity	train time	test time
1	LesTortuesNinja	gp_fixed_3	2015-04-11 00:48:59	0.9621	19	117	103
2	agramfort	gp_rf30_adaboost10_v2	2015-04-10 14:30:50	0.9596	3	117	104
3	Overfitters	stack_wavelet	2015-04-10 17:03:27	0.9588	6	313	132
7	delphine	feature_selection	2015-04-10 14:46:38	0.9577	4	117	109
8	delphine	first_test	2015-04-10 13:18:41	0.9574	1	127	110
9	bekou	fifthattempt	2015-04-10 17:33:31	0.9563	2	134	114
10	agramfort	gp_rf_adaboost_v3_gp_fix	2015-04-10 17:30:16	0.9555	1	93	84
11	anon	try_04_ab_gbc	2015-04-10 18:01:31	0.9552	2	149	101
12	bekou	firstmodel	2015-04-10 13:56:21	0.9550	4	146	116
13	2AN	eleventh	2015-04-10 16:40:54	0.9544	0	123	106
14	2AN	nineth	2015-04-10 16:38:22	0.9544	3	119	112
15	2AN	twelve	2015-04-10 16:40:54	0.9544	0	124	108
16	LesTortuesNinja	gp_2	2015-04-09 10:53:57	0.9544	0	134	117
17	Madclam	second_try_w_gp	2015-04-10 13:11:38	0.9544	0	136	111

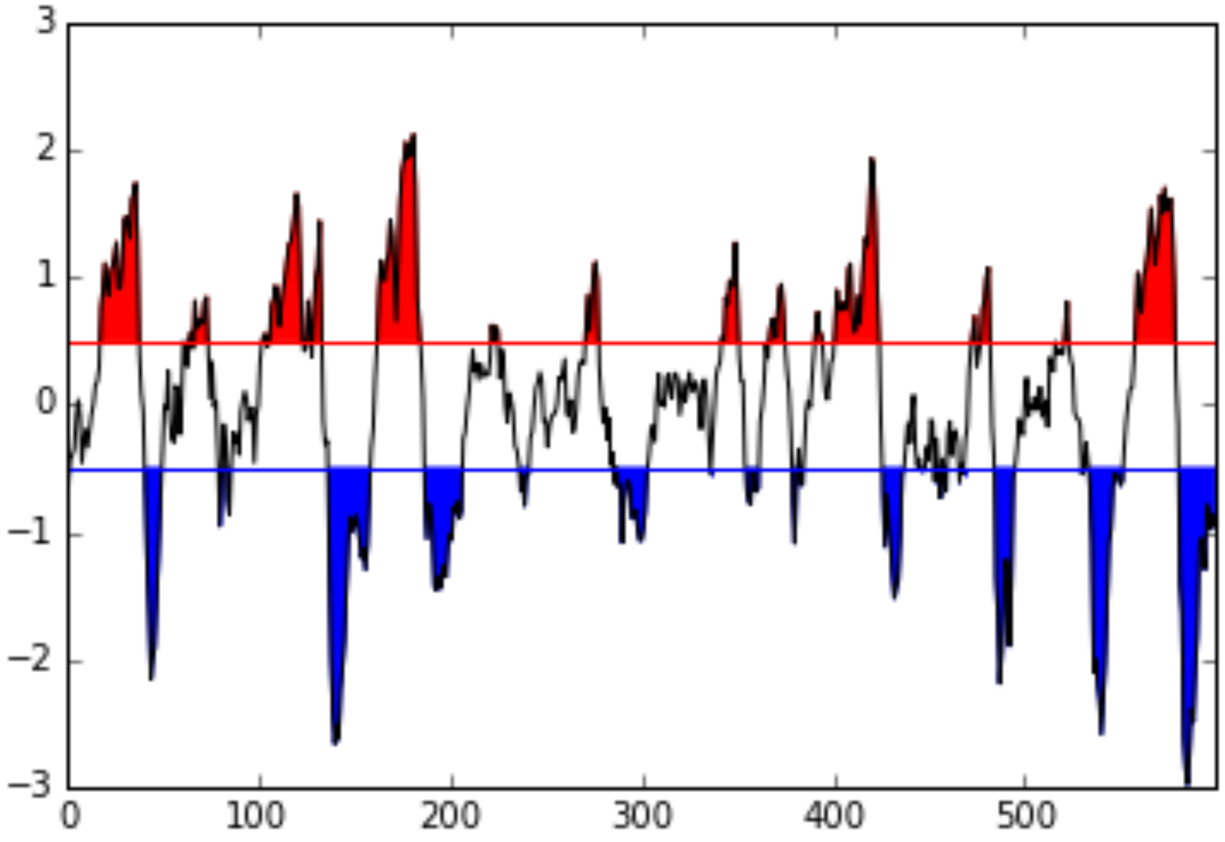
accuracy improvement: 89% to 96%

Rapid analytics and model prototyping

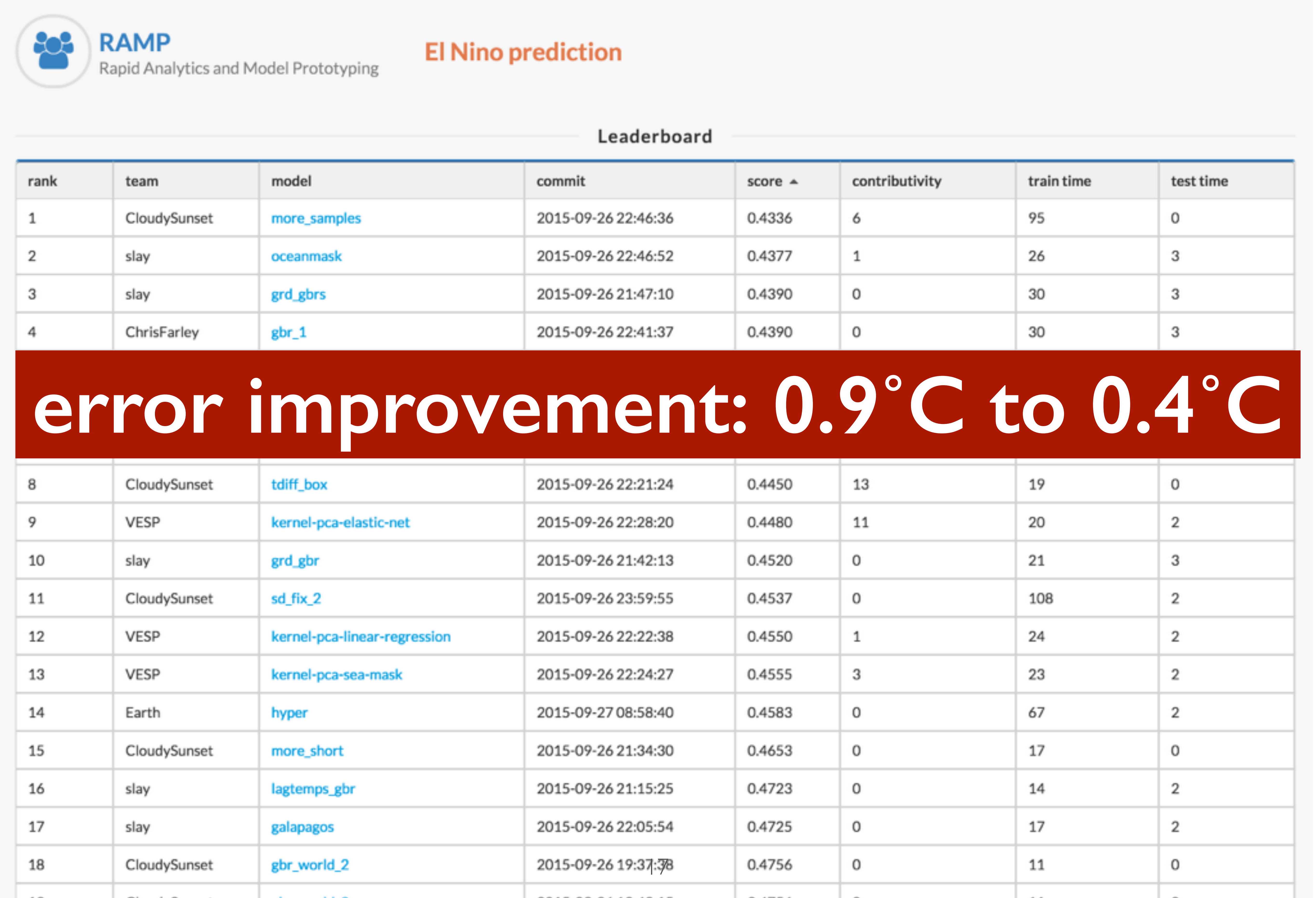
2015 June 16 and Sept 26

Predicting El Nino





Rapid analytics and model prototyping



RAMP

Rapid Analytics and Model Prototyping

El Nino prediction

Leaderboard

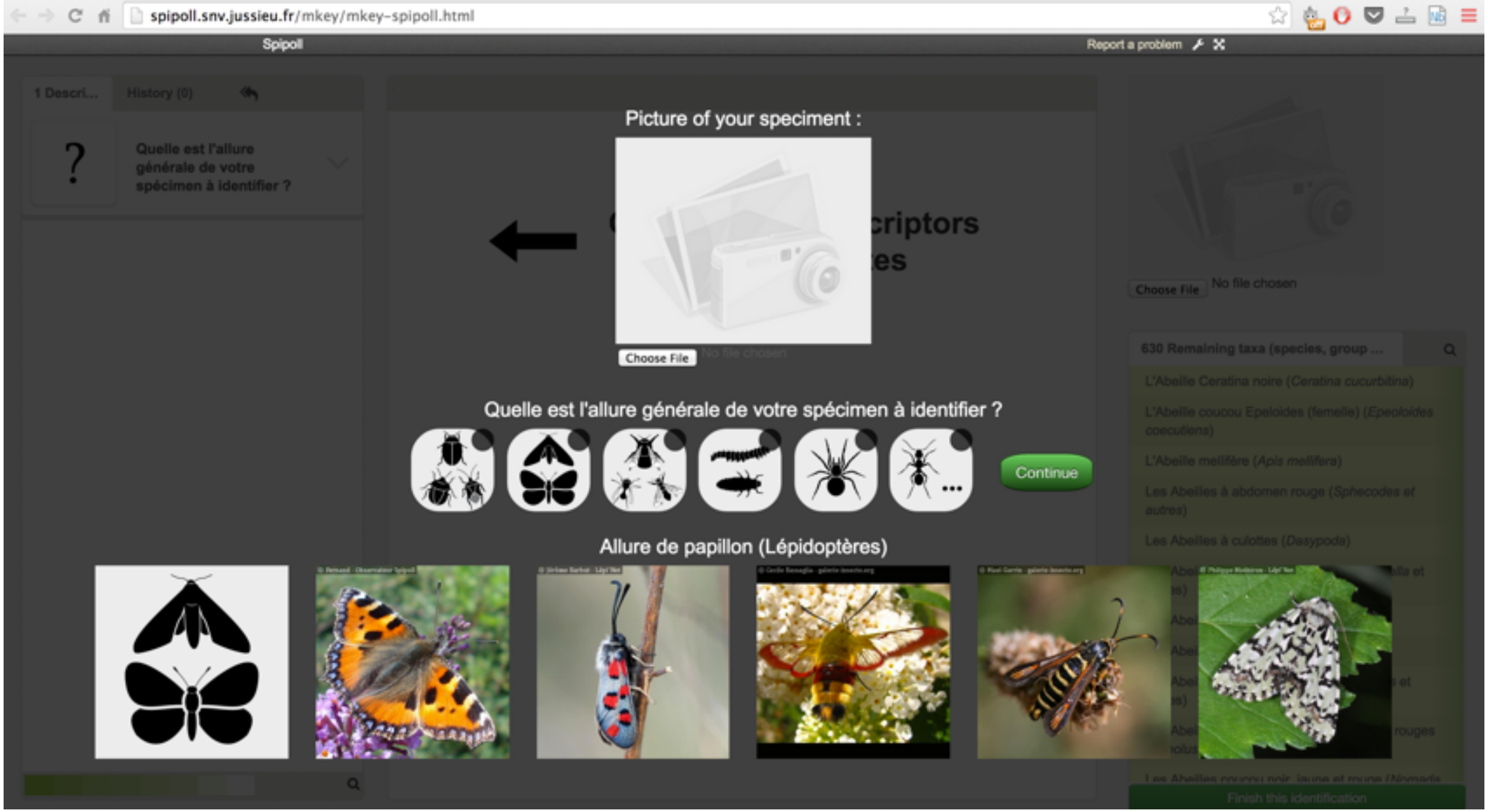
rank	team	model	commit	score ▲	contributivity	train time	test time
1	CloudySunset	more_samples	2015-09-26 22:46:36	0.4336	6	95	0
2	slay	oceanmask	2015-09-26 22:46:52	0.4377	1	26	3
3	slay	grd_gbrs	2015-09-26 21:47:10	0.4390	0	30	3
4	ChrisFarley	gbr_1	2015-09-26 22:41:37	0.4390	0	30	3
8	CloudySunset	tdiff_box	2015-09-26 22:21:24	0.4450	13	19	0
9	VESP	kernel-pca-elastic-net	2015-09-26 22:28:20	0.4480	11	20	2
10	slay	grd_gbr	2015-09-26 21:42:13	0.4520	0	21	3
11	CloudySunset	sd_fix_2	2015-09-26 23:59:55	0.4537	0	108	2
12	VESP	kernel-pca-linear-regression	2015-09-26 22:22:38	0.4550	1	24	2
13	VESP	kernel-pca-sea-mask	2015-09-26 22:24:27	0.4555	3	23	2
14	Earth	hyper	2015-09-27 08:58:40	0.4583	0	67	2
15	CloudySunset	more_short	2015-09-26 21:34:30	0.4653	0	17	0
16	slay	lagtemps_gbr	2015-09-26 21:15:25	0.4723	0	14	2
17	slay	galapagos	2015-09-26 22:05:54	0.4725	0	17	2
18	CloudySunset	gbr_world_2	2015-09-26 19:37:38	0.4756	0	11	0

error improvement: 0.9˚C to 0.4˚C

Rapid analytics and model prototyping

2015 October 8

Insect classification



Quelle est l'allure générale de votre spécimen à identifier ?

Picture of your speciment :

Quelle est l'allure générale de votre spécimen à identifier ?

Allure de papillon (Lépidoptères)

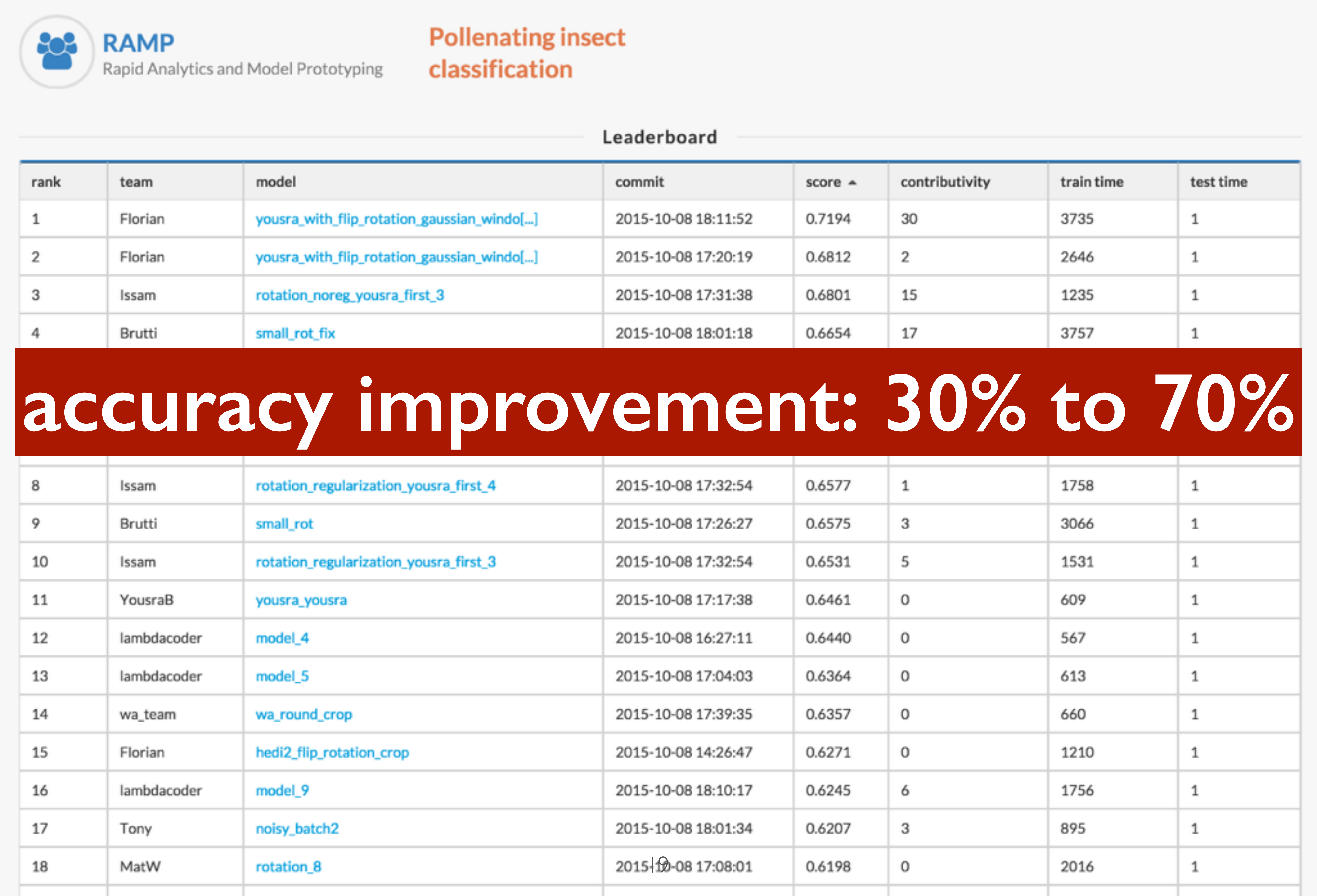
630 Remaining taxa (species, group ...

- L'Abeille Ceratina noire (*Ceratina cucurbitina*)
- L'Abeille coucou Epeloides (femelle) (*Epeoloides coecutiens*)
- L'Abeille mellifère (*Apis mellifera*)
- Les Abeilles à abdomen rouge (*Sphecodes et autres*)
- Les Abeilles à culottes (*Dasypoda*)

université PARIS-SACLAY

Paris-Saclay Center for Data Science

RAPID ANALYTICS AND MODEL PROTOTYPING



RAMP
Rapid Analytics and Model Prototyping

Pollenating insect classification

Leaderboard

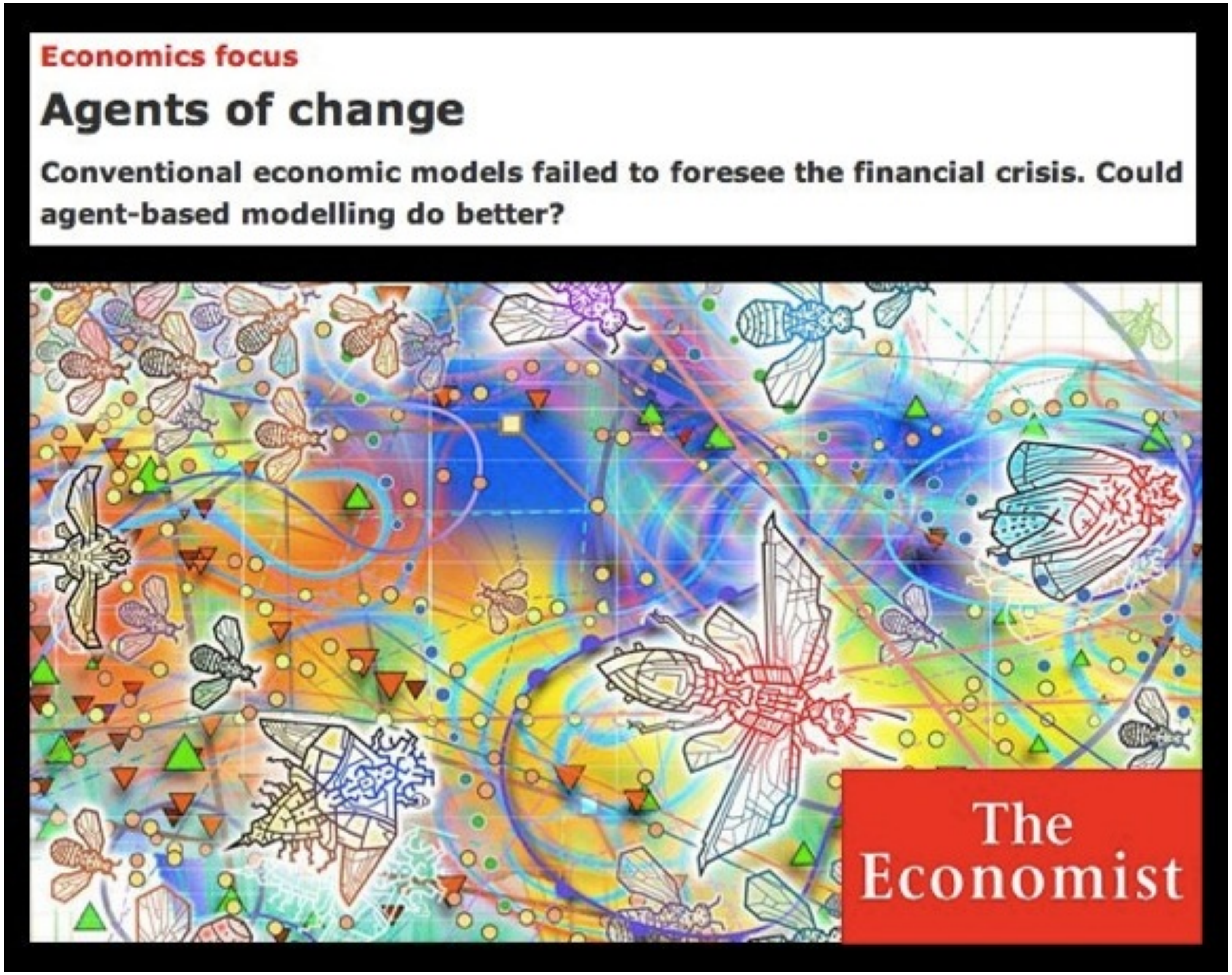
rank	team	model	commit	score ▲	contributivity	train time	test time
1	Florian	yousra_with_flip_rotation_gaussian_windo[...]	2015-10-08 18:11:52	0.7194	30	3735	1
2	Florian	yousra_with_flip_rotation_gaussian_windo[...]	2015-10-08 17:20:19	0.6812	2	2646	1
3	Issam	rotation_noreg_yousra_first_3	2015-10-08 17:31:38	0.6801	15	1235	1
4	Brutti	small_rot_fix	2015-10-08 18:01:18	0.6654	17	3757	1
8	Issam	rotation_regularization_yousra_first_4	2015-10-08 17:32:54	0.6577	1	1758	1
9	Brutti	small_rot	2015-10-08 17:26:27	0.6575	3	3066	1
10	Issam	rotation_regularization_yousra_first_3	2015-10-08 17:32:54	0.6531	5	1531	1
11	YousraB	yousra_yousra	2015-10-08 17:17:38	0.6461	0	609	1
12	lambdacoder	model_4	2015-10-08 16:27:11	0.6440	0	567	1
13	lambdacoder	model_5	2015-10-08 17:04:03	0.6364	0	613	1
14	wa_team	wa_round_crop	2015-10-08 17:39:35	0.6357	0	660	1
15	Florian	hedi2_flip_rotation_crop	2015-10-08 14:26:47	0.6271	0	1210	1
16	lambdacoder	model_9	2015-10-08 18:10:17	0.6245	6	1756	1
17	Tony	noisy_batch2	2015-10-08 18:01:34	0.6207	3	895	1
18	MatW	rotation_8	2015-10-08 17:08:01	0.6198	0	2016	1

accuracy improvement: 30% to 70%

Rapid analytics and model prototyping

2016 February 10

Macroeconomic agent-based models

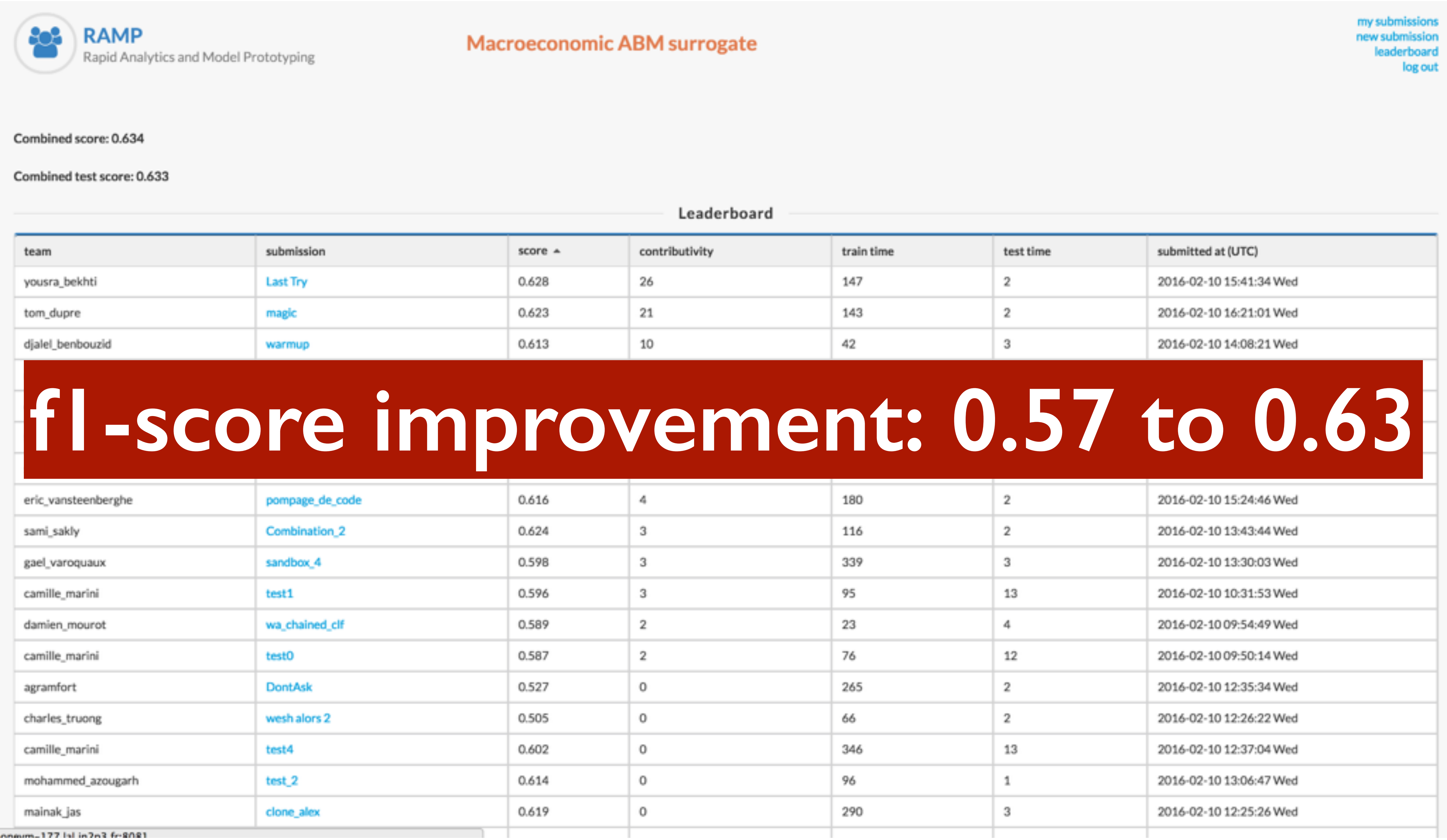


Economics focus

Agents of change

Conventional economic models failed to foresee the financial crisis. Could agent-based modelling do better?

Rapid analytics and model prototyping



RAMP
Rapid Analytics and Model Prototyping

Macroeconomic ABM surrogate

Combined score: 0.634

Combined test score: 0.633

Leaderboard

team	submission	score ▲	contributivity	train time	test time	submitted at (UTC)
yousra_bekhti	Last Try	0.628	26	147	2	2016-02-10 15:41:34 Wed
tom_dupre	magic	0.623	21	143	2	2016-02-10 16:21:01 Wed
djalel_benbouzid	warmup	0.613	10	42	3	2016-02-10 14:08:21 Wed
eric_vansteenberghe	pompage_de_code	0.616	4	180	2	2016-02-10 15:24:46 Wed
sami_sakly	Combination_2	0.624	3	116	2	2016-02-10 13:43:44 Wed
gael_varoquaux	sandbox_4	0.598	3	339	3	2016-02-10 13:30:03 Wed
camille_marini	test1	0.596	3	95	13	2016-02-10 10:31:53 Wed
damien_mourot	wa_chained_clf	0.589	2	23	4	2016-02-10 09:54:49 Wed
camille_marini	test0	0.587	2	76	12	2016-02-10 09:50:14 Wed
agramfort	DontAsk	0.527	0	265	2	2016-02-10 12:35:34 Wed
charles_truong	wesh alors 2	0.505	0	66	2	2016-02-10 12:26:22 Wed
camille_marini	test4	0.602	0	346	13	2016-02-10 12:37:04 Wed
mohammed_azougarh	test_2	0.614	0	96	1	2016-02-10 13:06:47 Wed
mainak_jas	clone_alex	0.619	0	290	3	2016-02-10 12:25:26 Wed

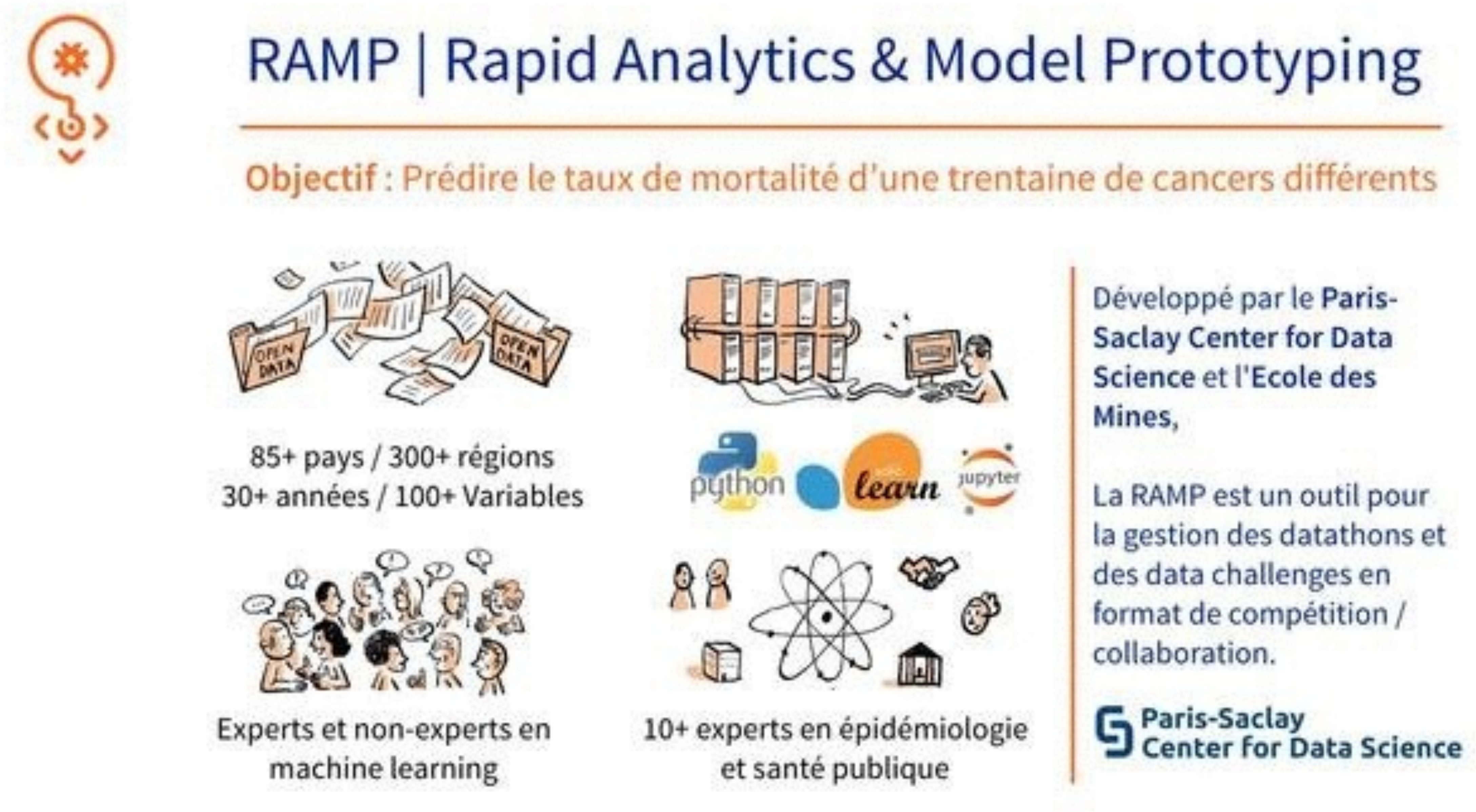
onevm-177.lal.in2p3.fr:8081

f1-score improvement: 0.57 to 0.63

RAPID ANALYTICS AND MODEL PROTOTYPING

2016 February 13

Epidemium cancer survival rate



RAMP | Rapid Analytics & Model Prototyping

Objectif : Prédire le taux de mortalité d'une trentaine de cancers différents

85+ pays / 300+ régions
30+ années / 100+ Variables

Experts et non-experts en machine learning

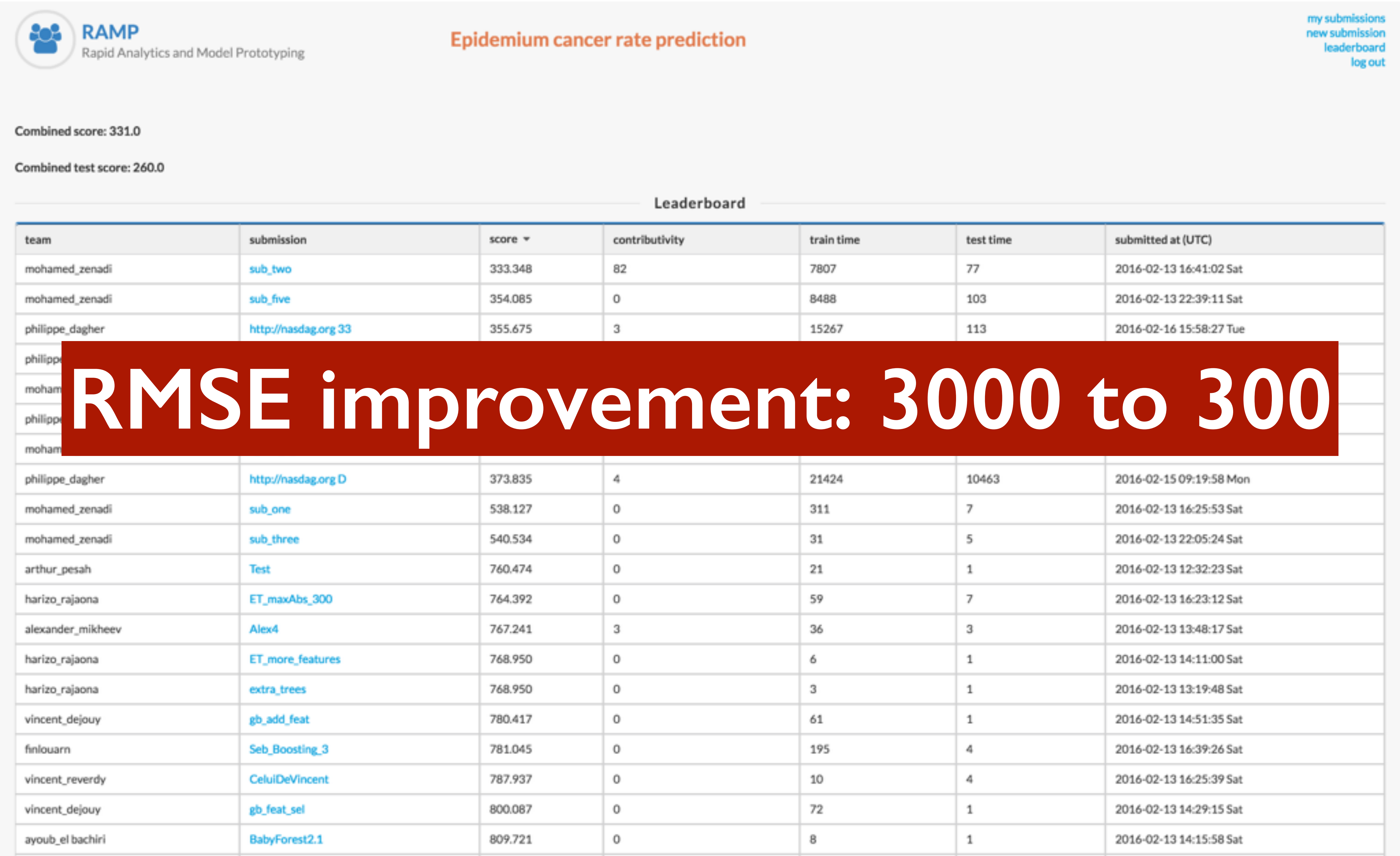
10+ experts en épidémiologie et santé publique

Développé par le **Paris-Saclay Center for Data Science** et l'**Ecole des Mines**,

La RAMP est un outil pour la gestion des datathons et des data challenges en format de compétition / collaboration.

Paris-Saclay Center for Data Science

RAPID ANALYTICS AND MODEL PROTOTYPING



Epidemium cancer rate prediction

Combined score: 331.0

Combined test score: 260.0

Leaderboard

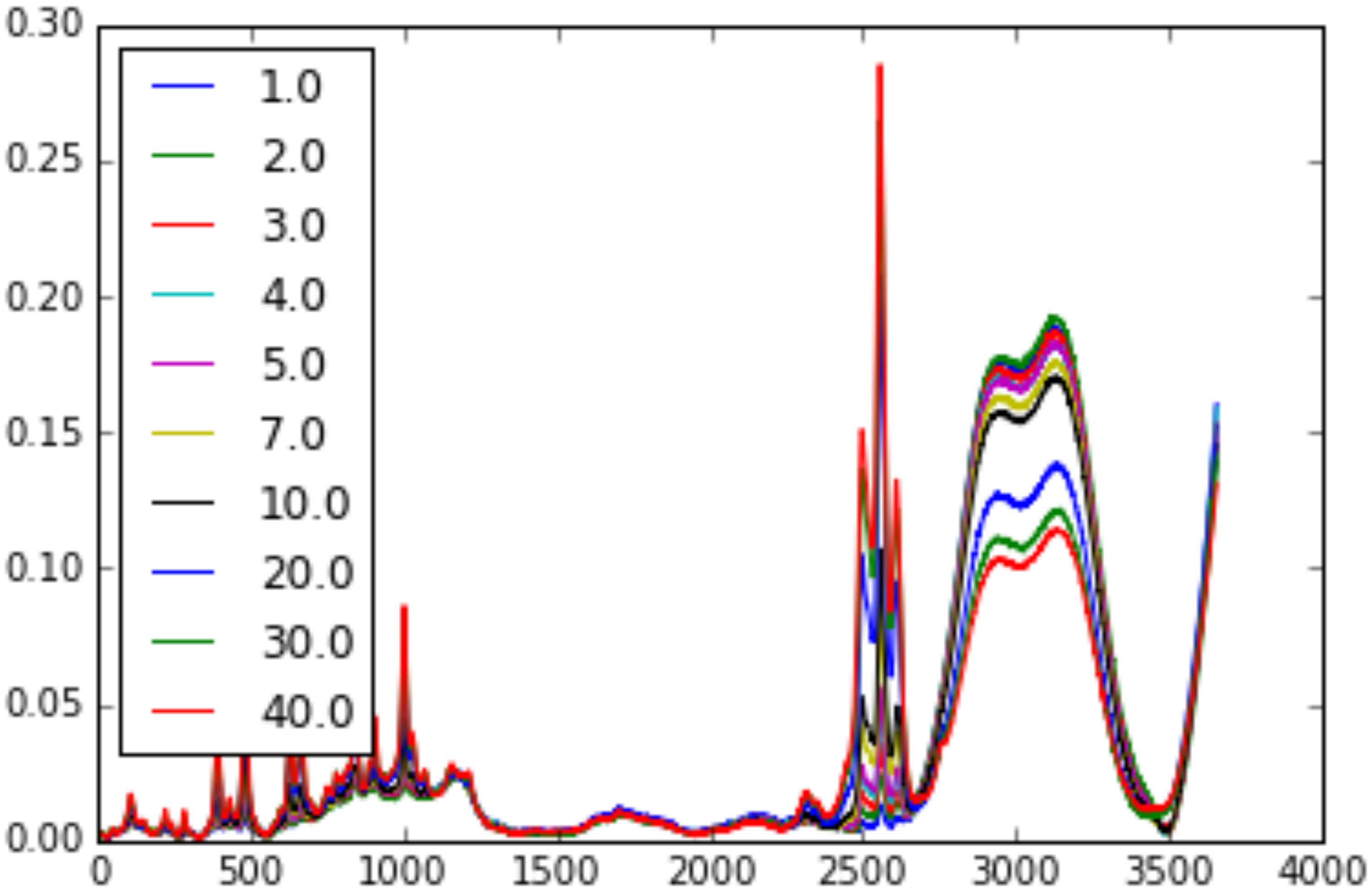
team	submission	score ▾	contributivity	train time	test time	submitted at (UTC)
mohamed_zenadi	sub_two	333.348	82	7807	77	2016-02-13 16:41:02 Sat
mohamed_zenadi	sub_five	354.085	0	8488	103	2016-02-13 22:39:11 Sat
philippe_dagher	http://nasdag.org 33	355.675	3	15267	113	2016-02-16 15:58:27 Tue
philipp						
moham						
philipp						
moham						
philippe_dagher	http://nasdag.org D	373.835	4	21424	10463	2016-02-15 09:19:58 Mon
mohamed_zenadi	sub_one	538.127	0	311	7	2016-02-13 16:25:53 Sat
mohamed_zenadi	sub_three	540.534	0	31	5	2016-02-13 22:05:24 Sat
arthur_pesah	Test	760.474	0	21	1	2016-02-13 12:32:23 Sat
harizo_rajaona	ET_maxAbs_300	764.392	0	59	7	2016-02-13 16:23:12 Sat
alexander_mikheev	Alex4	767.241	3	36	3	2016-02-13 13:48:17 Sat
harizo_rajaona	ET_more_features	768.950	0	6	1	2016-02-13 14:11:00 Sat
harizo_rajaona	extra_trees	768.950	0	3	1	2016-02-13 13:19:48 Sat
vincent_dejouy	gb_add_feat	780.417	0	61	1	2016-02-13 14:51:35 Sat
finlouarn	Seb_Boosting_3	781.045	0	195	4	2016-02-13 16:39:26 Sat
vincent_reverdy	CeluiDeVincent	787.937	0	10	4	2016-02-13 16:25:39 Sat
vincent_dejouy	gb_feat_sel	800.087	0	72	1	2016-02-13 14:29:15 Sat
ayoub_el bachiri	BabyForest2.1	809.721	0	8	1	2016-02-13 14:15:58 Sat

RMSE improvement: 3000 to 300

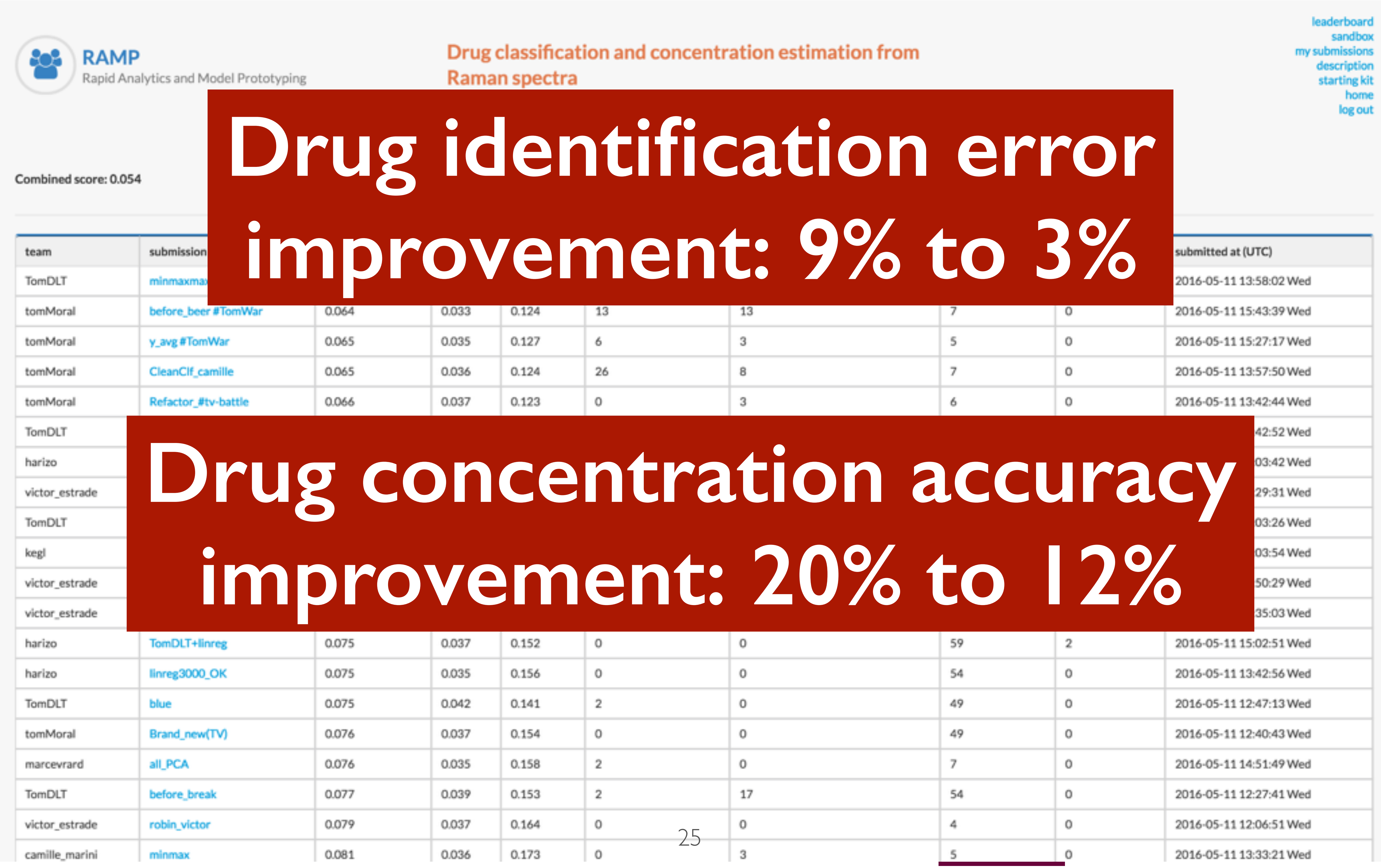
Rapid analytics and model prototyping

2016 May 11

Drug identification from spectra



RAPID ANALYTICS AND MODEL PROTOTYPING



RAMP
Rapid Analytics and Model Prototyping

Drug classification and concentration estimation from Raman spectra

Combined score: 0.054

team	submission								submitted at (UTC)
TomDLT	minmaxma								2016-05-11 13:58:02 Wed
tomMoral	before_beer #TomWar	0.064	0.033	0.124	13	13	7	0	2016-05-11 15:43:39 Wed
tomMoral	y_avg #TomWar	0.065	0.035	0.127	6	3	5	0	2016-05-11 15:27:17 Wed
tomMoral	CleanClf_camille	0.065	0.036	0.124	26	8	7	0	2016-05-11 13:57:50 Wed
tomMoral	Refactor_#tv-battle	0.066	0.037	0.123	0	3	6	0	2016-05-11 13:42:44 Wed
TomDLT									42:52 Wed
harizo									03:42 Wed
victor_estrade									29:31 Wed
TomDLT									03:26 Wed
kegl									03:54 Wed
victor_estrade									50:29 Wed
victor_estrade									35:03 Wed
harizo	TomDLT+linreg	0.075	0.037	0.152	0	0	59	2	2016-05-11 15:02:51 Wed
harizo	linreg3000_OK	0.075	0.035	0.156	0	0	54	0	2016-05-11 13:42:56 Wed
TomDLT	blue	0.075	0.042	0.141	2	0	49	0	2016-05-11 12:47:13 Wed
tomMoral	Brand_new(TV)	0.076	0.037	0.154	0	0	49	0	2016-05-11 12:40:43 Wed
marcevrard	all_PCA	0.076	0.035	0.158	2	0	7	0	2016-05-11 14:51:49 Wed
TomDLT	before_break	0.077	0.039	0.153	2	17	54	0	2016-05-11 12:27:41 Wed
victor_estrade	robin_victor	0.079	0.037	0.164	0	0	4	0	2016-05-11 12:06:51 Wed
camille_marini	minmax	0.081	0.036	0.173	0	3	5	0	2016-05-11 13:33:21 Wed

Drug identification error improvement: 9% to 3%

Drug concentration accuracy improvement: 20% to 12%

The RAMP tool

A prototyping tool for collaborative development of data science workflows

- Fast development of analytics solutions
- Teaching support
- Networking
- Support for collaborative team work

TAKE HOME MESSAGE

- We have **expertise** and **tools** to build and run data challenges
- It’s not magic
  - needs **publicly available annotated data**
  - needs a **use case** and a **prediction pipeline**
- It gives you
  - **dissemination**/**communication**
  - **access** to (the time and expertise) of **data scientists**
  - **prototype** of optimized pipeline