THE SYSTEMIC CHALLENGES OF DATA SCIENCE INITIATIVES

BALÁZS KÉGL

CNRS & University Paris Saclay

université PARIS-SACLAY

Paris-Saclay Center for Data Science

I will not talk about science

I will talk about

management
(of) (data) science

Where does it come from?

- My **eight-year of experience** interfacing between **high-energy physics** and **data science**
- Our **two-year** experience of **running PS-CDS**
- **Extensive collaboration** with **management scientist**

DATA SCIENCE IN THE WORLD

NYU CENTER FOR DATA SCIENCE

UNIVERSITY of WASHINGTON

UNIVERSITY of WASHINGTON

UC BERKELEY INSTITUTE FOR DATA SCIENCE

UNIVERSITY OF ROCHESTER

INSTITUTE FOR DATA SCIENCE

Amsterdam Data Science

THE UNIVERSITY of EDINBURGH

DATA SCIENCE

Data Science

universite PARIS-SACLAY

Paris-Saclay Center for Data Science

UNIVERSITÉ PARIS-SACLAY

19 founding partners

Paris-Saclay Center for Data Science

UNIVERSITÉ PARIS-SACLAY

19 *fondateurs*

60 000 *étudiants*

6 000 *doctorants*

15 000 *étudiants en master*

8 *Schools*

11 000 *chercheurs et enseignants-chercheurs*

300 *laboratoires*

8 000 *publications /an*

15 % *de la recherche publique française*

10 *départements*

+ horizontal multi-disciplinary and multi-partner initiatives to create cohesion

Paris-Saclay Center for Data Science

A multi-disciplinary initiative to define, structure, and manage the data science ecosystem at the Université Paris-Saclay

http://www.datascience-paris-saclay.fr/

250 researchers in **35** laboratories

Biology & bioinformatics
IBISC/UEvry
LRI/UPSud
Hepatinov
CESP/UPSud-UVSQ-Inserm
IGM-I2BC/UPSud
MIA/Agro
MIAj-MIG/INRA
LMAS/Centrale

Chemistry
EA4041/UPSud

Earth sciences
LATMOS/UVSQ
GEOPS/UPSud
IPSL/UVSQ
LSCE/UVSQ
LMD/Polytechnique

Economy
LM/ENSAE
RITM/UPSud
LFA/ENSAE

Neuroscience
UNICOG/Inserm
U1000/Inserm
NeuroSpin/CEA

Particle physics astrophysics & cosmology
LPP/Polytechnique
DMPH/ONERA
CosmoStat/CEA
IAS/UPSud
AIM/CEA
LAL/UPSud

Machine learning
LRI/UPSud
LTCI/Telecom
CMLA/Cachan
LS/ENSAE
LIX/Polytechnique
MIA/Agro
CMA/Polytechnique
LSS/Supélec
CVN/Centrale
LMAS/Centrale
DTIM/ONERA
IBISC/UEvry
LIST/CEA

Visualization
INRIA
LIMSI

Signal processing
LTCI/Telecom
CMA/Polytechnique
CVN/Centrale
LSS/Supélec
CMLA/Cachan
LIMSI
DTIM/ONERA

Statistics
LMO/UPSud
LS/ENSAE
LSS/Supélec
CMA/Polytechnique
LMAS/Centrale
MIA/AgroParisTech

DATA SCIENCE

Design of automated methods

to analyze massive and complex data

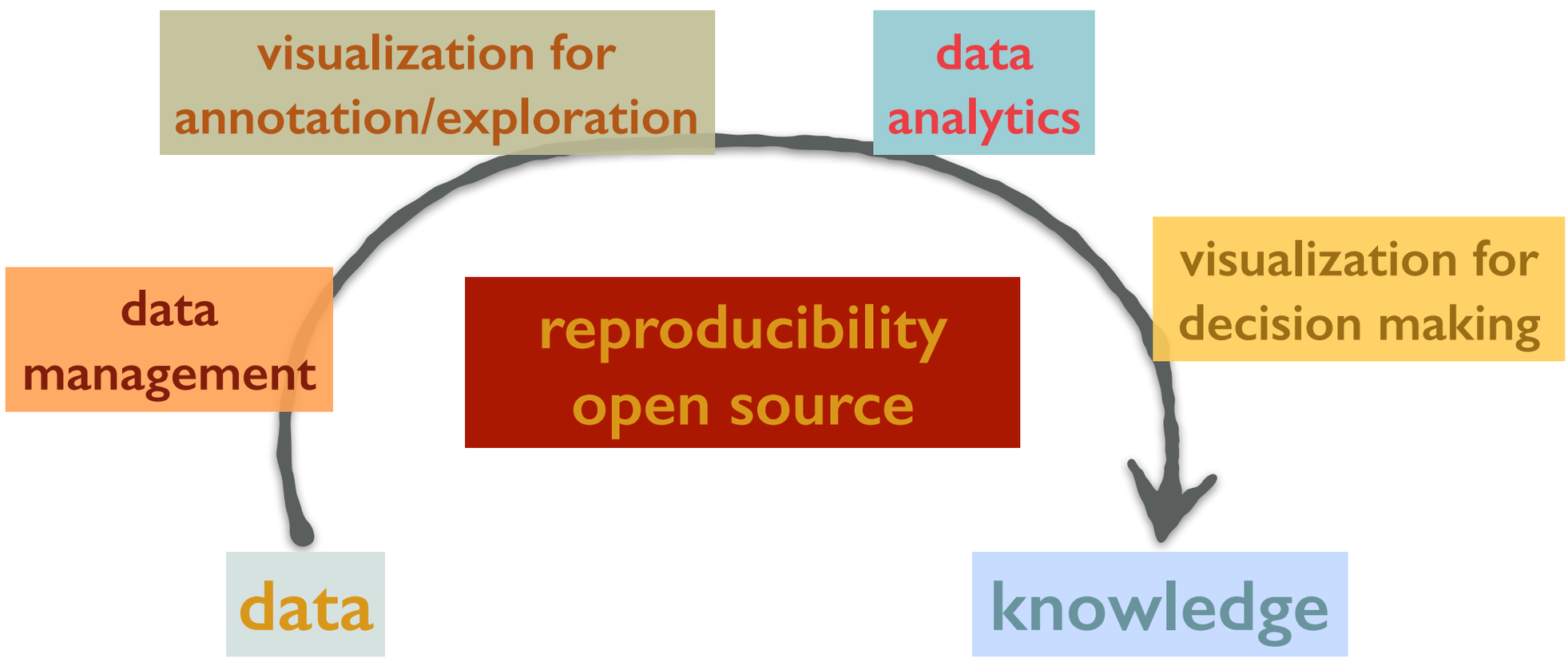
to extract useful information

Center for Data Science

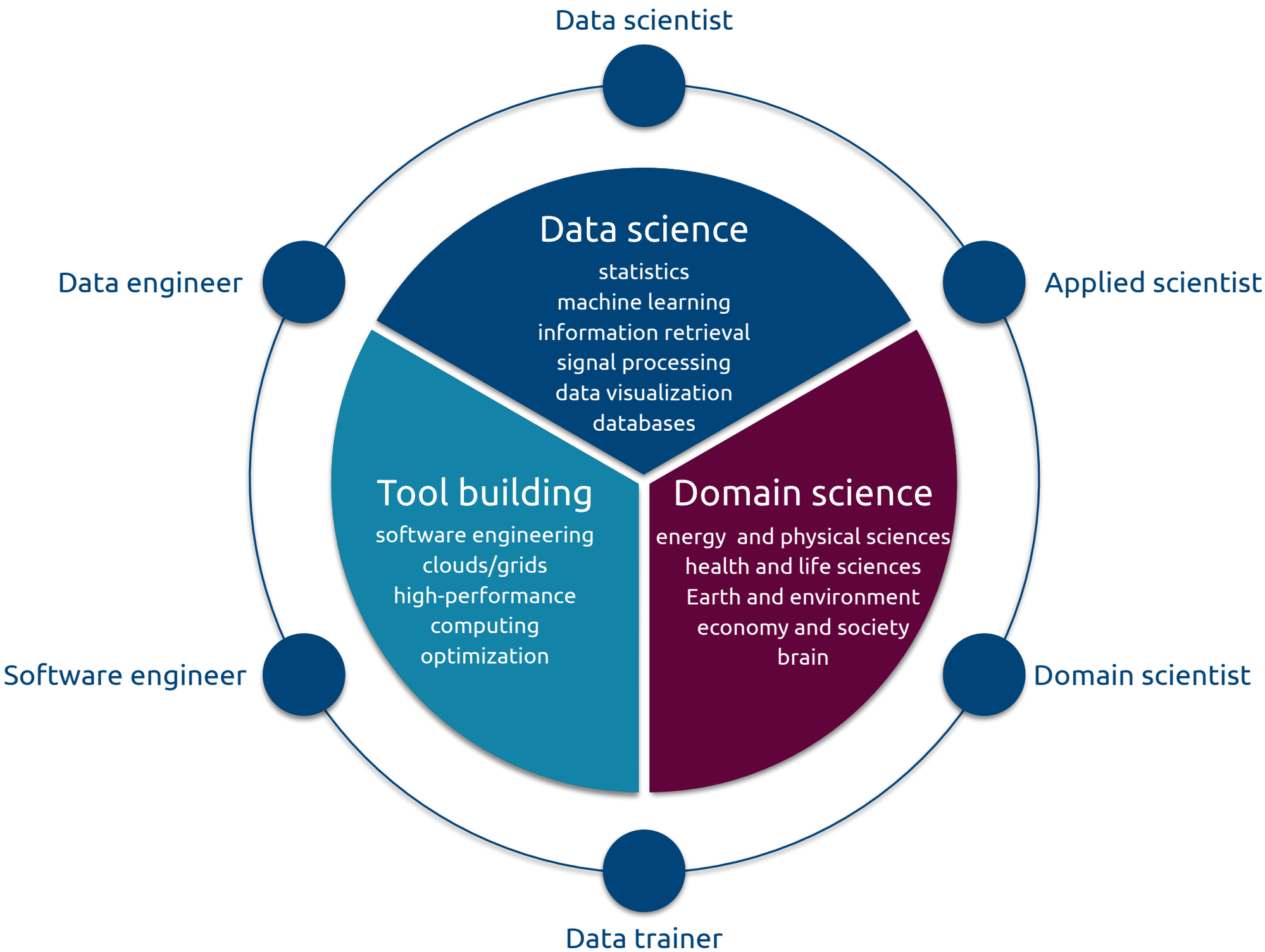
≠

Data Center

We are focusing on inference: data → knowledge



THE DATA SCIENCE LANDSCAPE

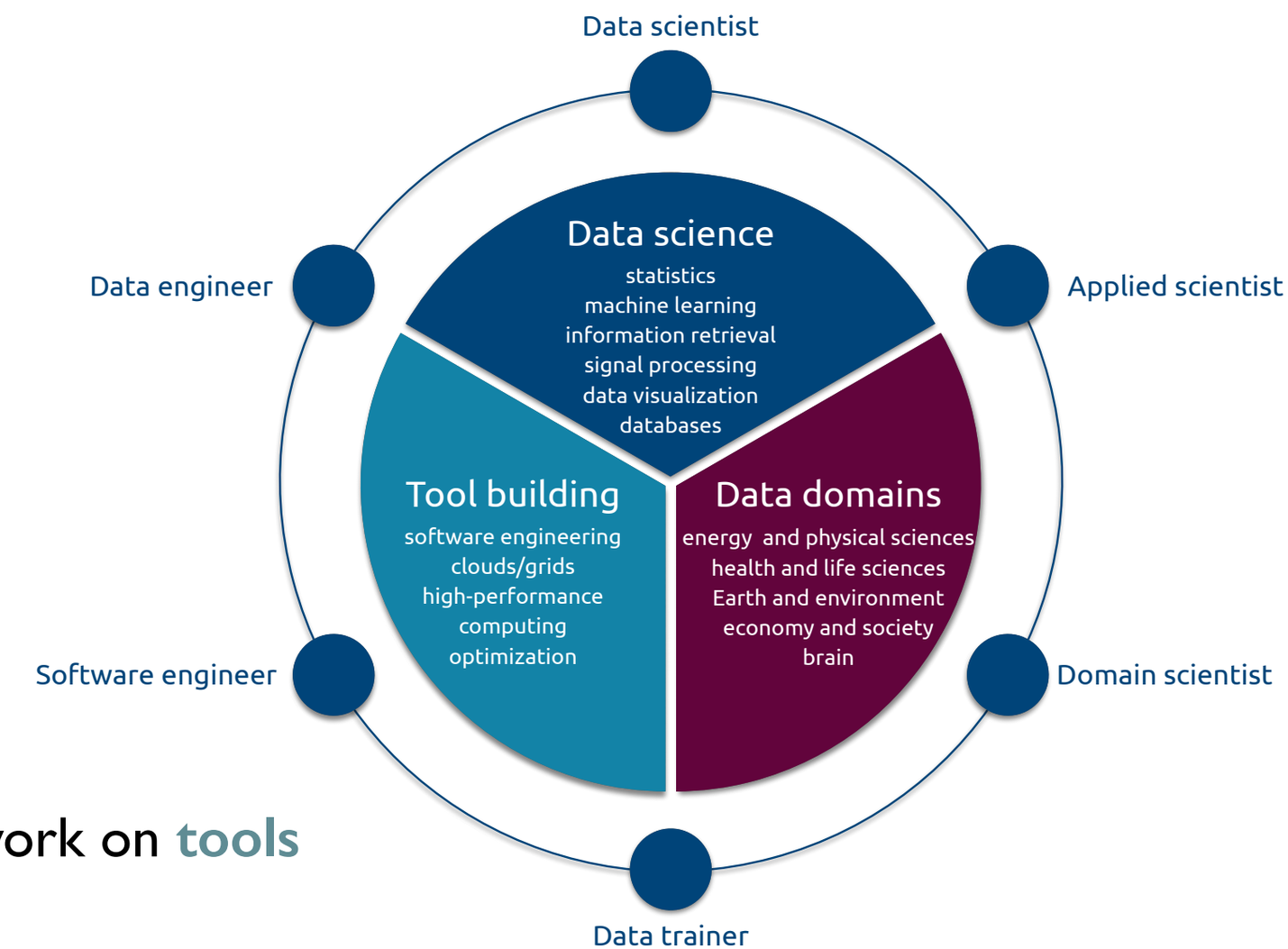


THE DATA SCIENCE LANDSCAPE

https://medium.com/@balazskegl

CHALLENGES

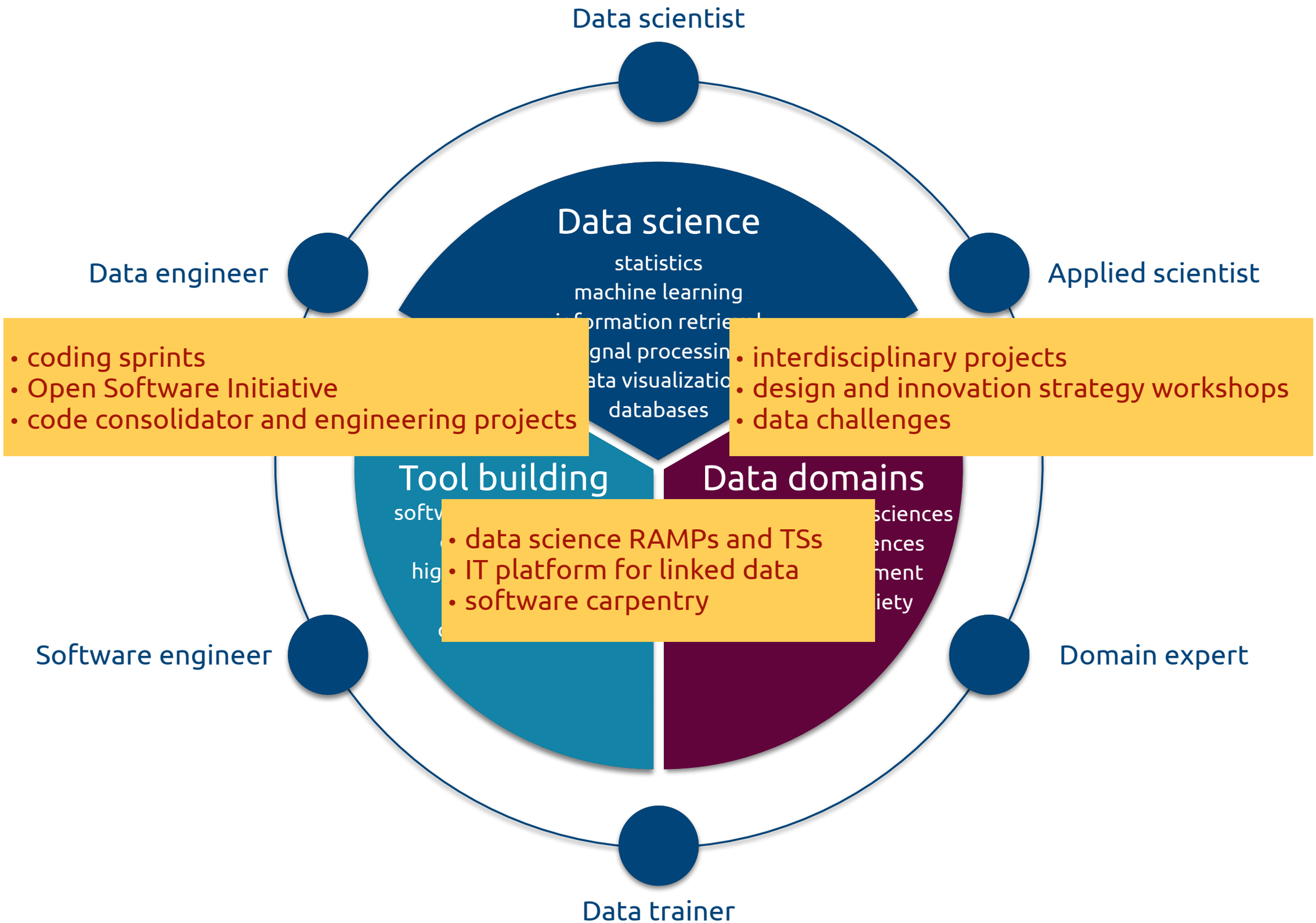
- (The lack of) manpower
  - especially at the **interfaces**
  - industrial **brain-drain**
- Incentives
  - data scientists are **not incentivized** to work on **domain science**
  - scientists are **not incentivized** to work on **tools**
- Access
  - no well-developed channels to **identify the right experts** for a given problem
- Tools
  - few **tools** that can help domain scientists and data scientists to **collaborate efficiently**



Tools

We are designing and learning to manage tools to accompany data science projects with different needs

TOOLS: LANDSCAPE TO ECOSYSTEM



Two analytics tools for initiating domain-data science interactions

Data challenges

Rapid analytics and model prototyping (RAMP)

DATA CHALLENGES

kaggle | Host | Competitions | Scripts | Jobs | Community | Balazs Kegl | Logout

We are making our URLs prettier -- Claim your personal URL now!

Active Competitions

Springleaf	**Springleaf Marketing Response** Determine whether to send a direct mail piece to a customer	7.4 days 2193 teams 1213 scripts $100,000
Deloitte.	**Western Australia Rental Prices** Predict rental prices for properties across Western Australia	49 days 48 teams $100,000
AI2	**The Allen AI Science Challenge** Is your model smarter than an 8th grader?	4 months 92 teams $80,000
	Rossmann Store Sales Forecast sales using store, promotion, and competitor data	2 months 856 teams 305 scripts $35,000
CERN	**Flavours of Physics: Finding τ → μμμ** Identify a rare decay phenomenon	10 hours 677 teams 736 scripts $15,000
Dato	**Truly Native?**	2.4 days

Balazs Kegl
View / Edit Profile

Is your company hiring?
Are you on the job market?
Visit our jobs board >>

Recent Jobs

AWOK.com - Senior Data Scientist (Big Data) (Dubai - UAE, Bengaluru - India)

Zynga - Senior Product Manager, Data Science (San Francisco)

DataRobot - Data Scientist (Japan)

trivago - Data Scientist – Amsterdam Office (Düsseldorf)

Red Ventures - Director, Data Science (Charlotte, NC)

BBC-Group - CTO - Software Engineer Machine Learning for a new business unit (Start-Up Division) (Zurich, Switzerland)

On the Forums

DATA CHALLENGES

- A **data challenge** is a recently developed unconventional **dissemination** and **communication** 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**

universite PARIS-SACLAY
Paris-Saclay Center for Data Science

DATA CHALLENGES

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

HUGE PUBLICITY

kaggle | Customer Solutions | Competitions | Community | Sign up | Login

Higgs challenge

Completed • $13,000 • 1,785 teams

Higgs Boson Machine Learning Challenge

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

Dashboard

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	1[illegible]0	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)

Huge publicity

Significant improvement over the baseline

yet partially missing the objectives

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)

- Prototyping
- Training
- Human resources
- Collaboration building, networking
- Social science observatory

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

SIGNIFICANT IMPROVEMENT OVER THE BASELINE

#	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)
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.	3.20423	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)

Analytics tool to promote collaboration and Code Reuse

onevm-222.lal.in2p3.fr:9002/models/kegl/md2faa2e46018704821c8e1b49c47c9b82e6fdf6c/model.py

Databoard

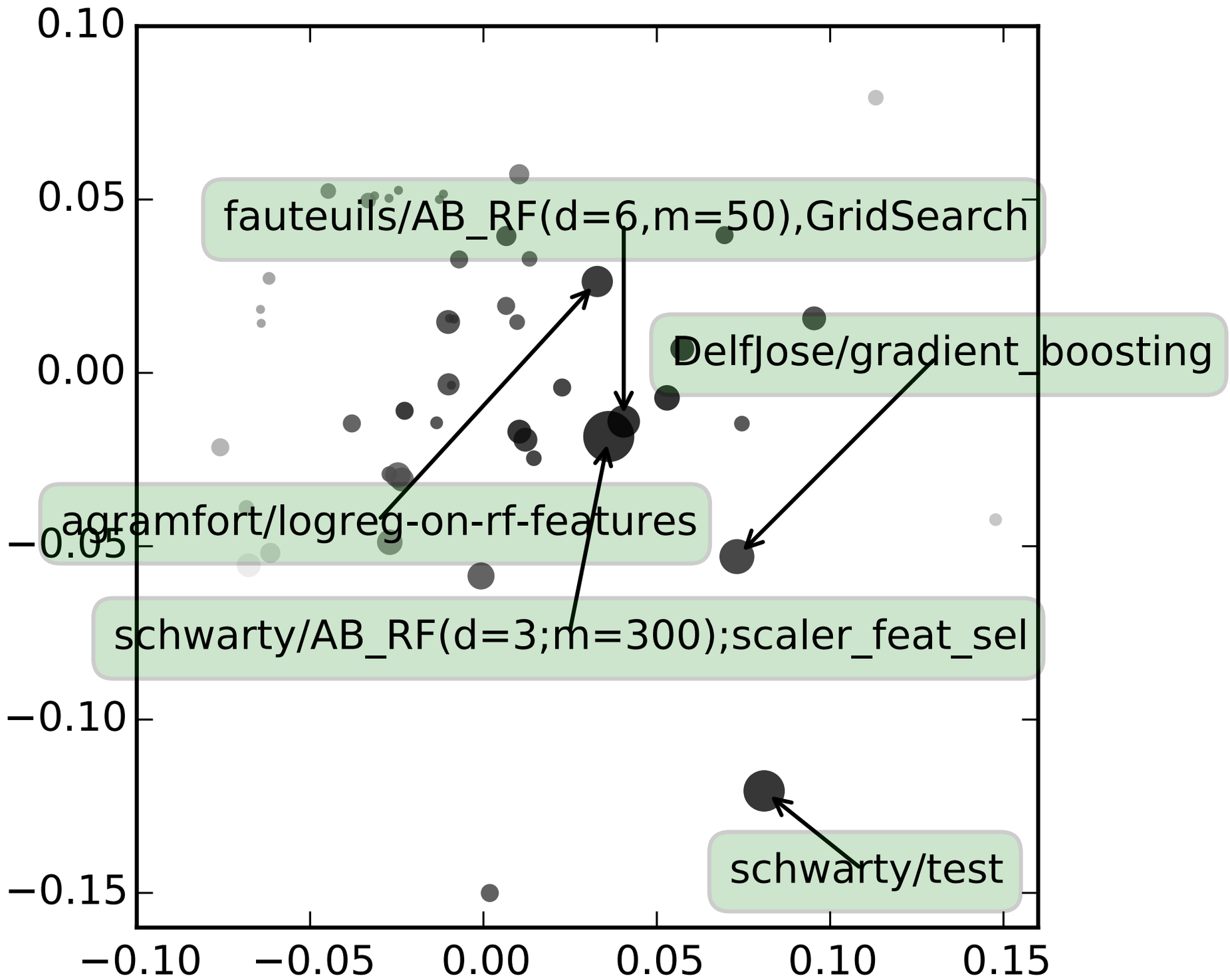
Leaderboard > kegl > MF.AB(20;RF(100;5))_d1 > model.py

Archive

model.py

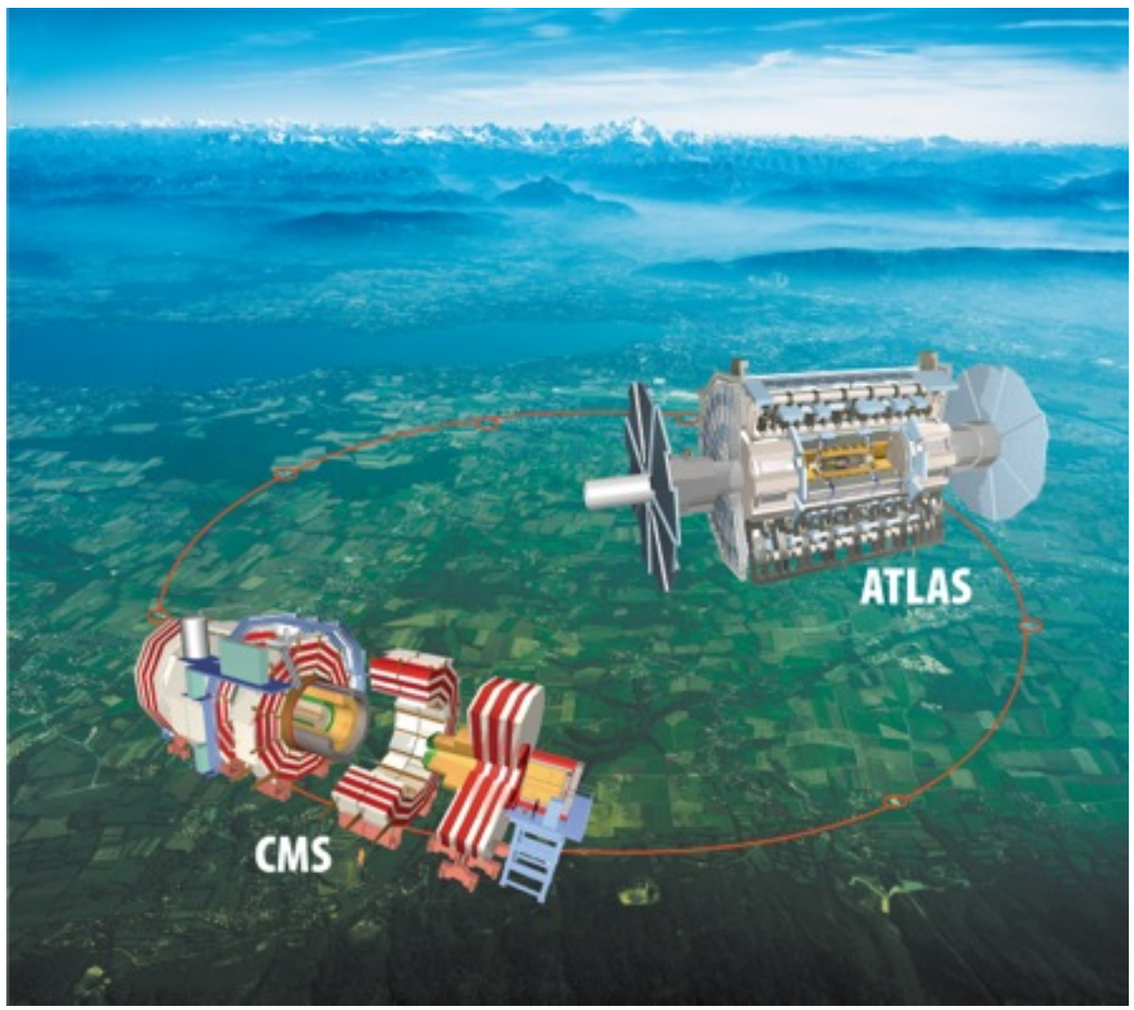
```
1.  from sklearn.ensemble import RandomForestClassifier, AdaBoostClassifier
2.  from sklearn.preprocessing import Imputer
3.  from sklearn.pipeline import Pipeline
4.  from sklearn.base import BaseEstimator
5.
6.  class Classifier(BaseEstimator):
7.      def __init__(self):
8.          self.clf = Pipeline([('imputer', Imputer(strategy='most_frequent')),
9.          ('rf', AdaBoostClassifier(base_estimator=RandomForestClassifier(max_depth=5,
    n_estimators=100),
10.                                n_estimators=20))])
11.
12.     def fit(self, X, y):
13.         self.clf.fit(X, y)
14.
15.     def predict(self, X):
16.         return self.clf.predict(X)
17.
18.     def predict_proba(self, X):
19.         return self.clf.predict_proba(X)
20.
```

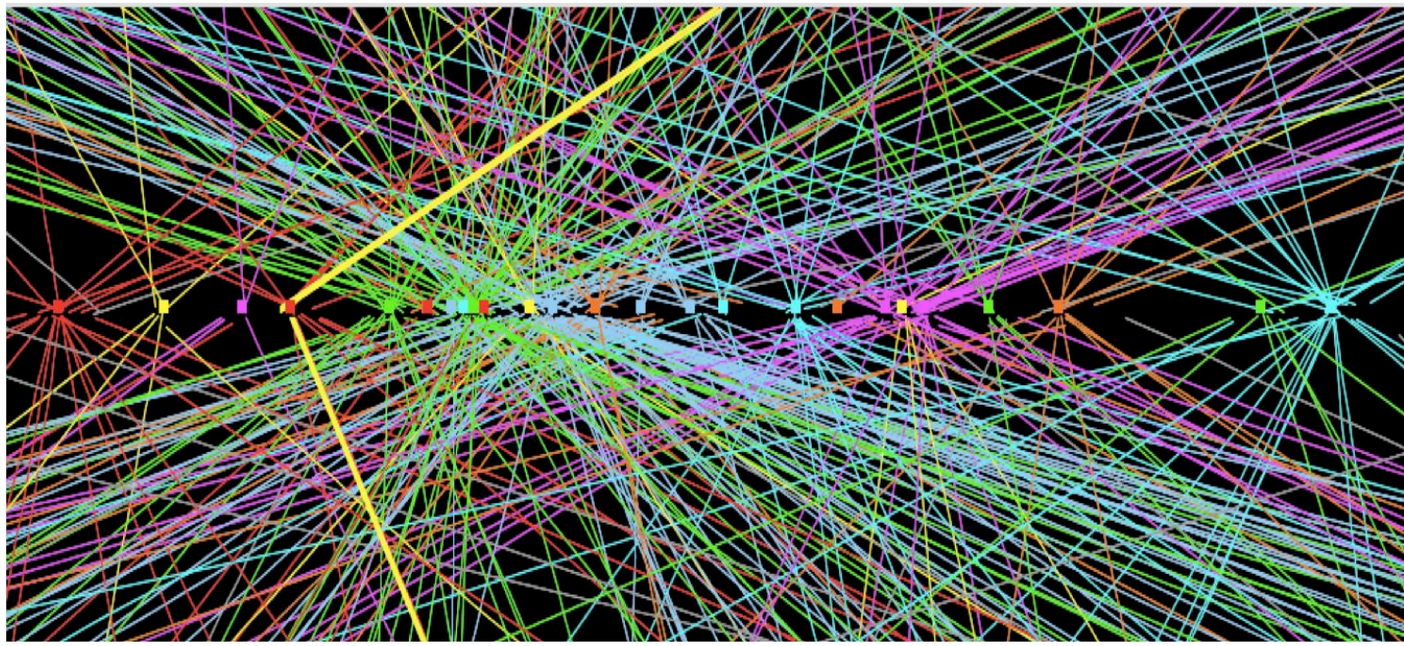
ANALYTICS TOOLS TO MONITOR PROGRESS



2015 Jan 15

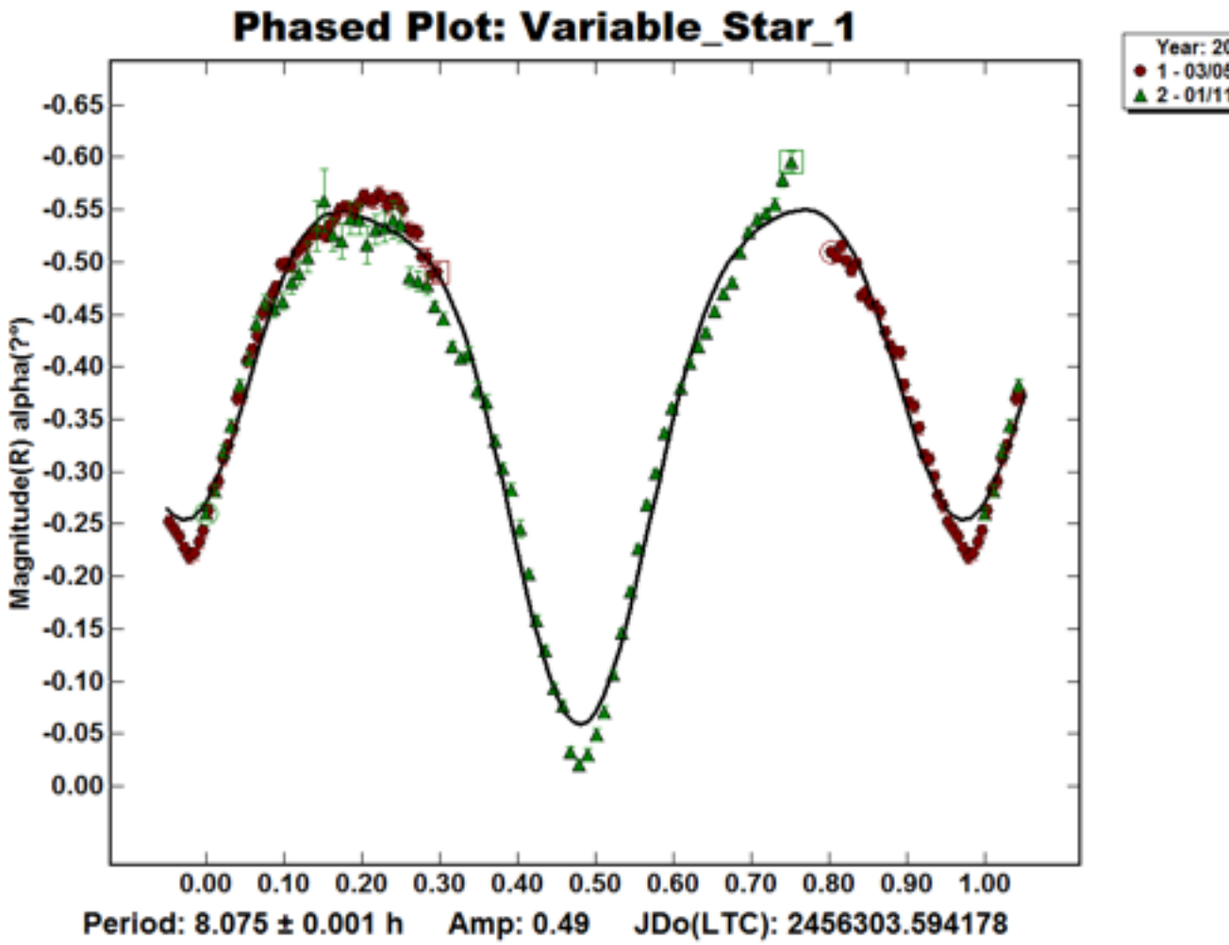
The HiggsML challenge



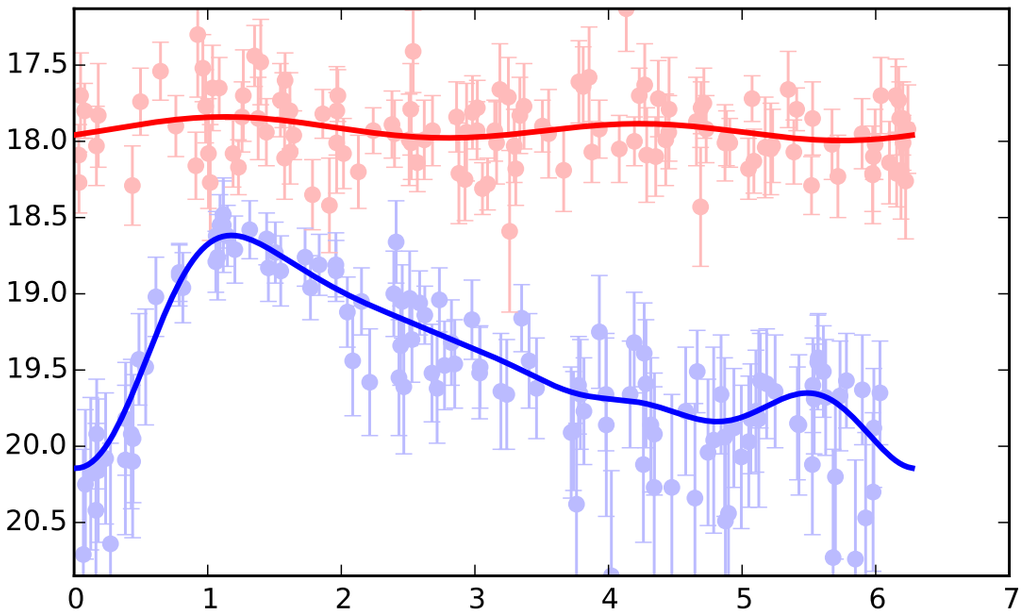


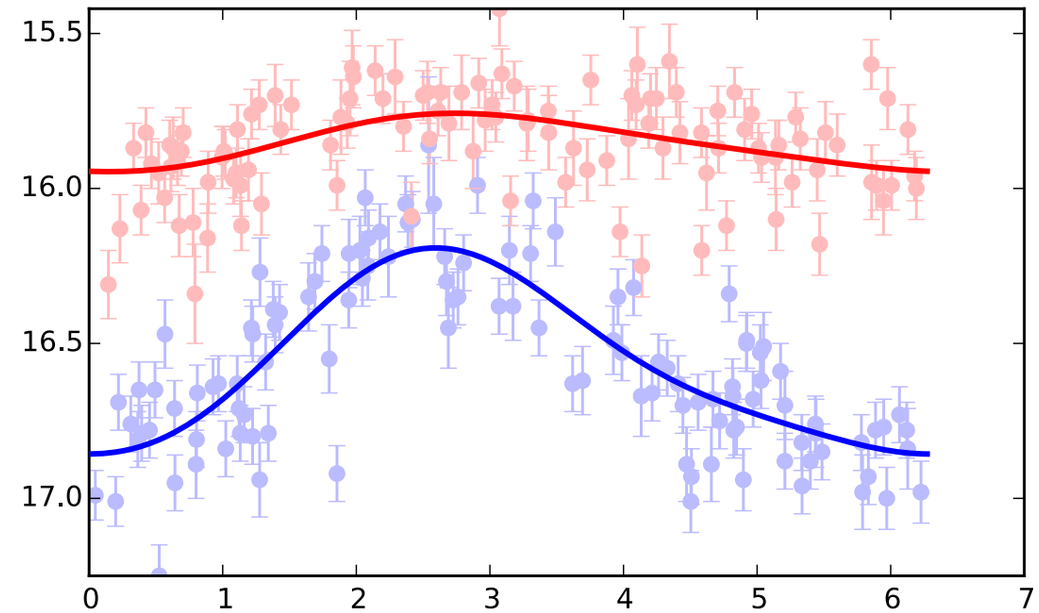
2015 Apr 10

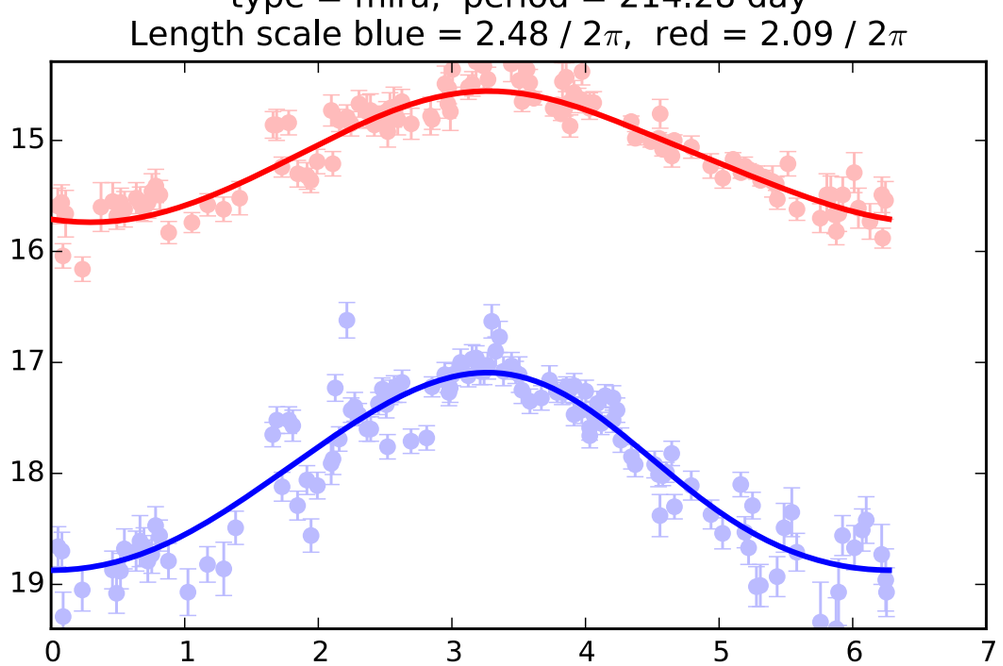
Classifying variable stars

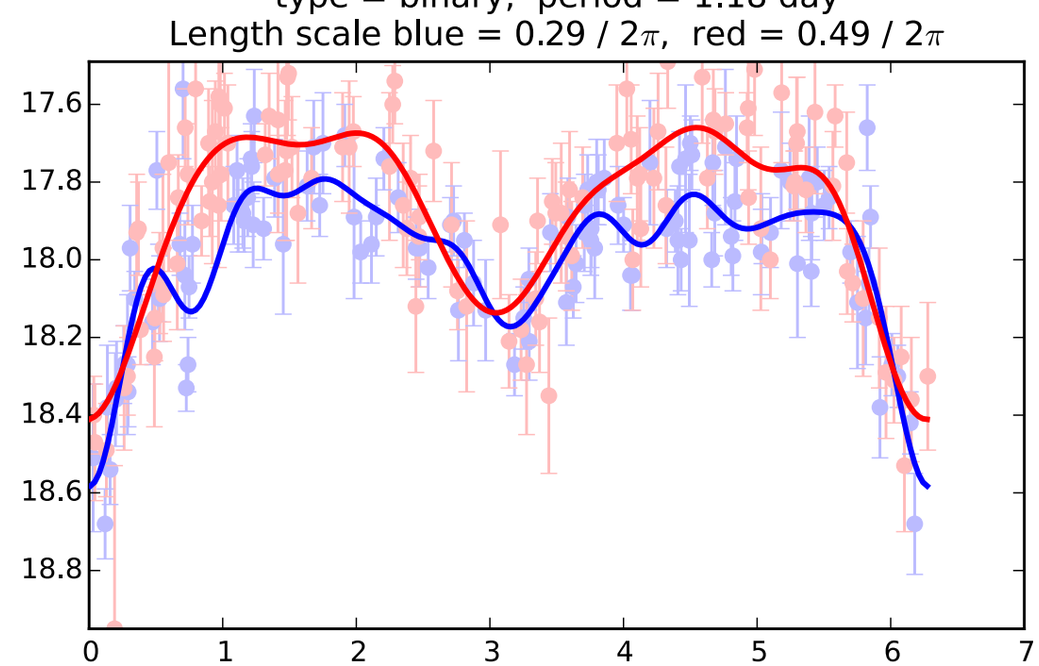


VARIABLE STARS









VARIABLE STARS

RAMP
Rapid Analytics and Model Prototyping

Variable star type prediction

Leaderboard

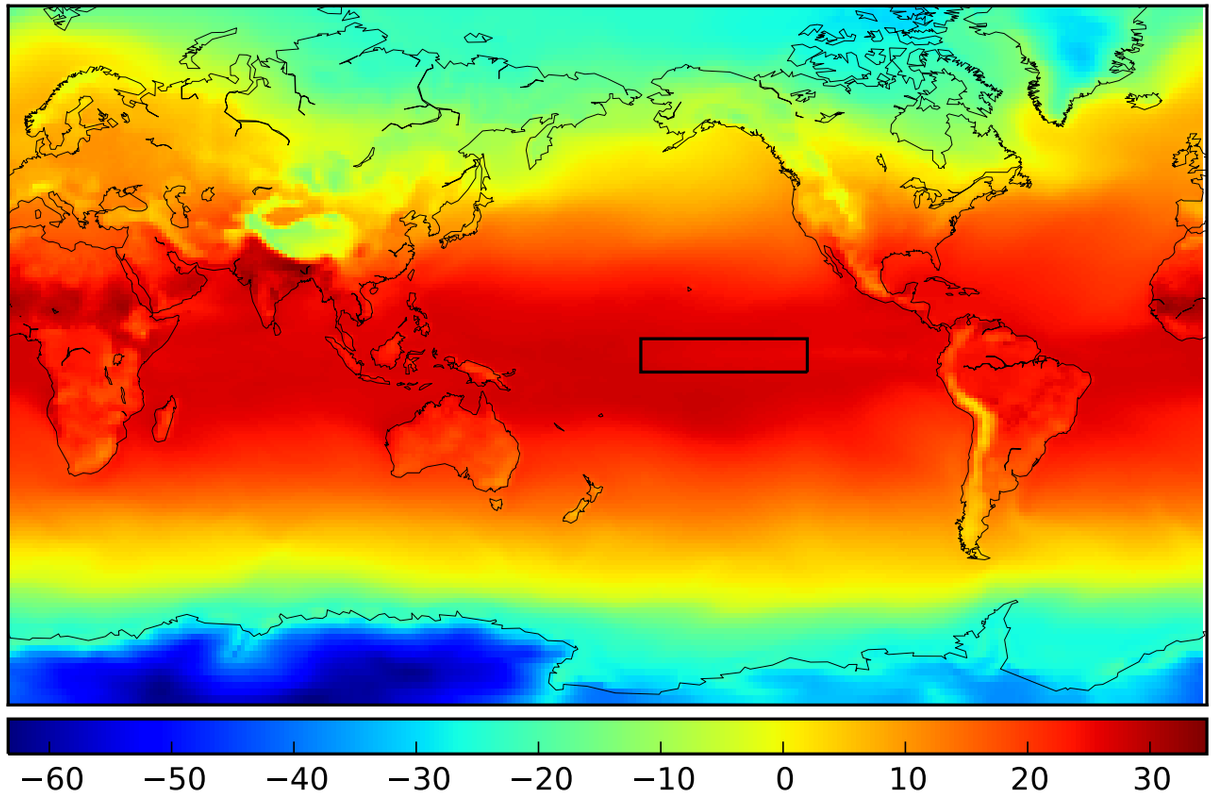
rank	team	model	commit	score ▲	contributivity	train time	test time
1	LesTortuesNinja	gp_fixed_3	2015-04-11 00:48:59	0.9621	19	117	103
2	agramfort	gp_rf30_adaboost10_v2	2015-04-10 14:30:50	0.9596	3	117	104
3	Overfitters	stack_wavelet	2015-04-10 17:03:27	0.9588	6	313	132
7	delphine	feature_selection	2015-04-10 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

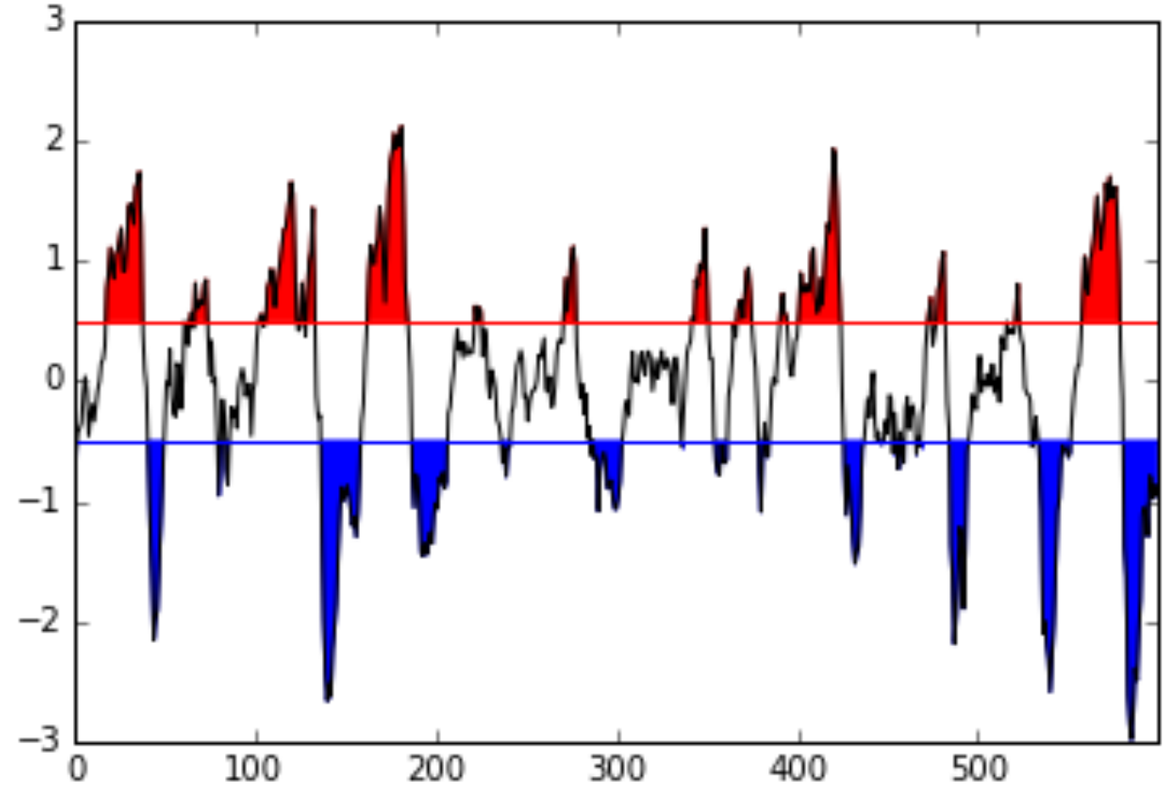
accuracy improvement: 89% to 96%

2015 June 16 and Sept 26

Predicting El Nino

Temperature map





RAPID ANALYTICS AND MODEL PROTOTYPING

RAMP
Rapid Analytics and Model Prototyping

El Nino prediction

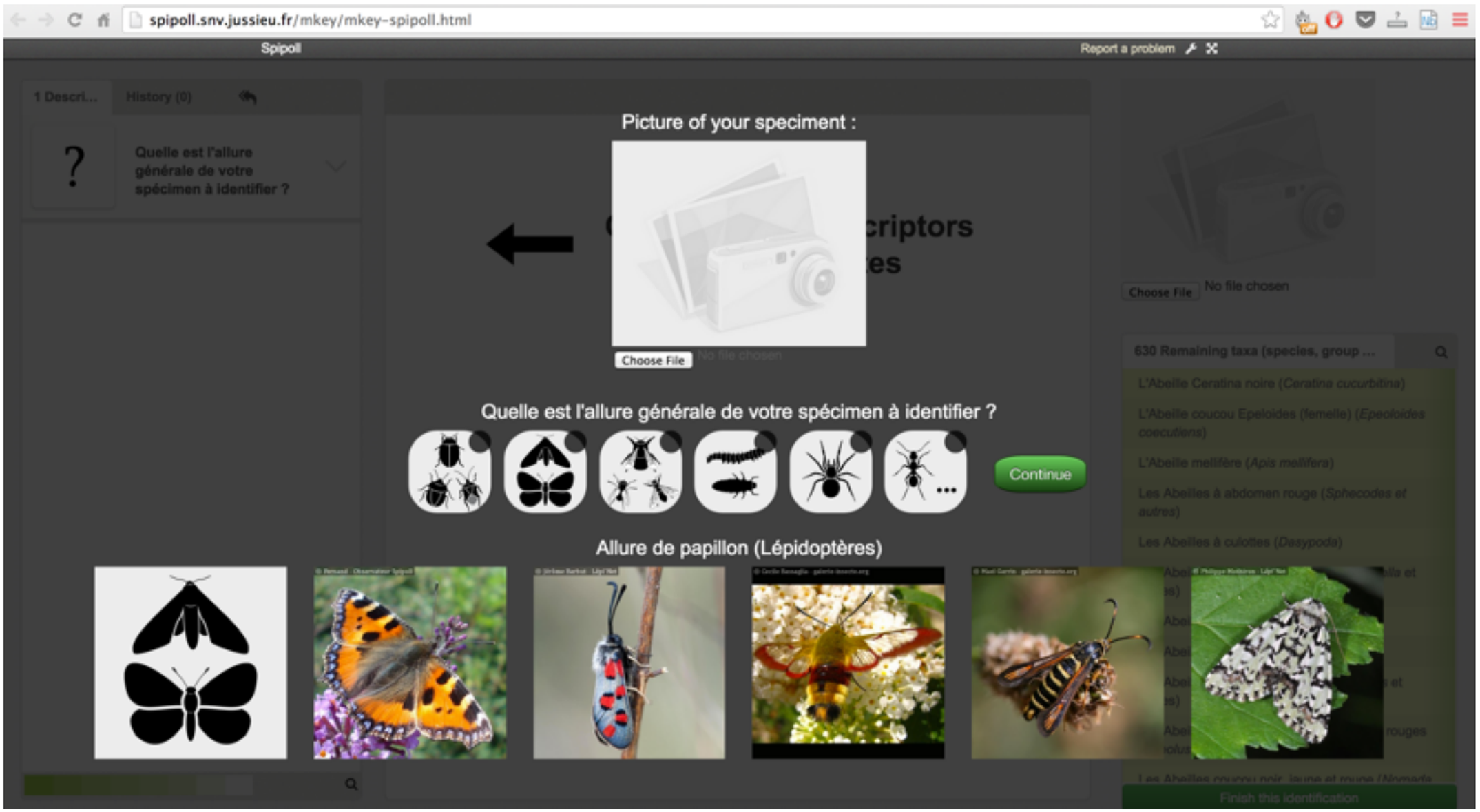
Leaderboard

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 19:37:48	0.4756	0	11	0

RMSE improvement: 0.9°C to 0.4°C

2015 October 8

Insect classification



RAPID ANALYTICS AND MODEL PROTOTYPING

RAMP
Rapid Analytics and Model Prototyping

Pollenating insect classification

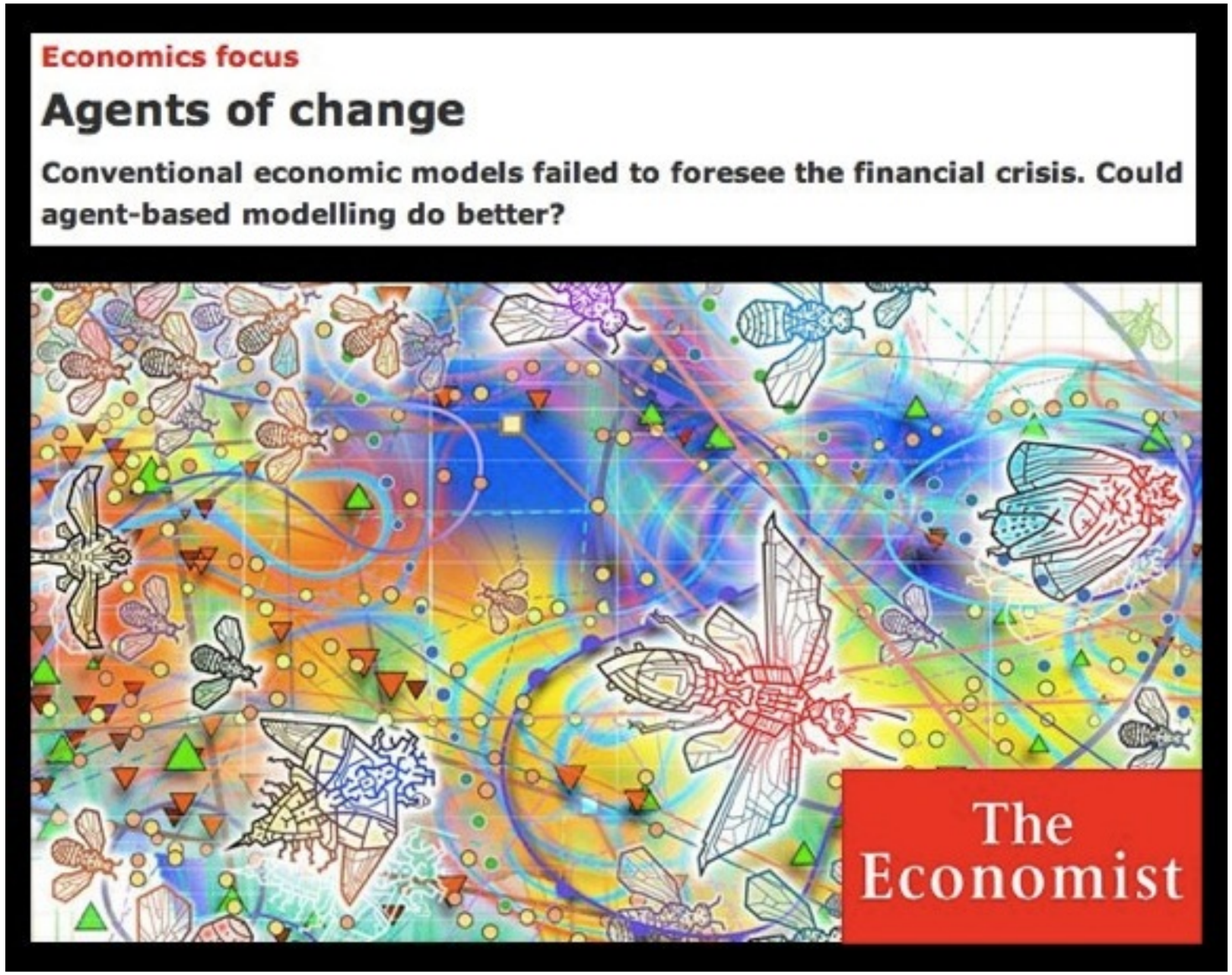
Leaderboard

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%

2016 February 10

Macroeconomic agent-based models



Rapid analytics and model prototyping

RAMP
Rapid Analytics and Model Prototyping

Macroeconomic ABM surrogate

my submissions
new submission
leaderboard
log out

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

f1-score improvement: 0.57 to 0.63

2016 February 13

Epidemium cancer survival rate

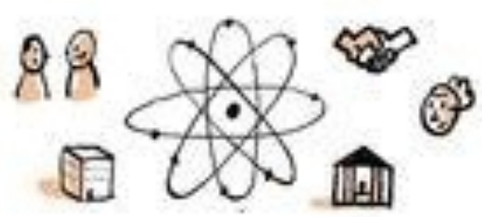
RAMP | Rapid Analytics & Model Prototyping

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

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

python learn jupyter

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

RAMP
Rapid Analytics and Model Prototyping

Epidemium cancer rate prediction

my submissions
new submission
leaderboard
log out

Combined score: 331.0

Combined test score: 260.0

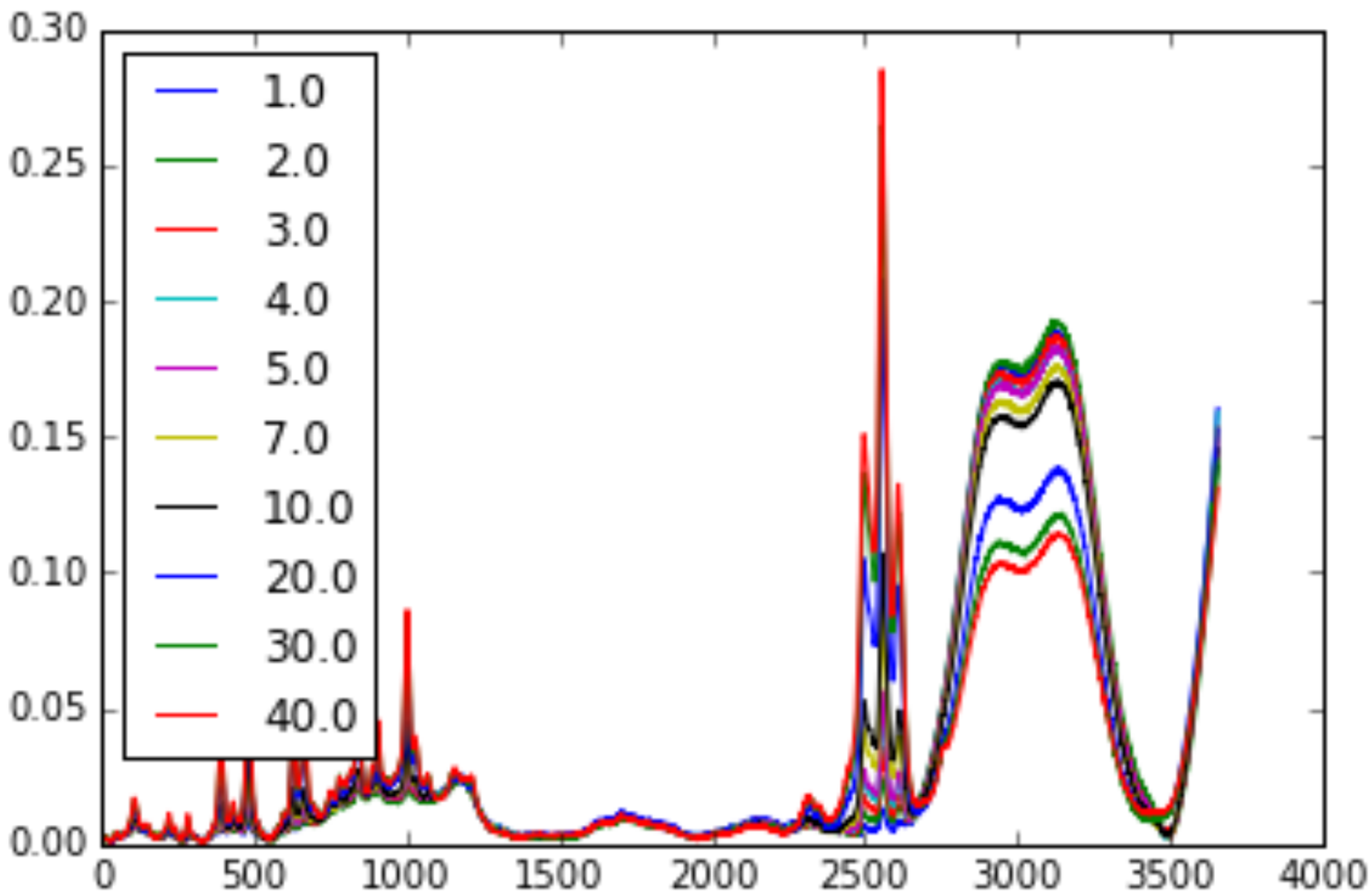
Leaderboard

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
philippe						
moham						
philippe						
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

2016 May 11

Drug identification from spectra



THE RAMP TOOL

A prototyping tool for collaborative development of data science workflows

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

TAKE HOME MESSAGES

- We have cool tools for **collaborative data analytics**
- **Data management is a big part** of the data analytics workflow
- **Big data is rare**: our problems are more about **flexible organization of heterogeneous data**
  - we especially need **collaborative** and **crowdsourcing** tools

Thank you!

Designing data science projects

- Efficient exploration of the space of innovative ideas
- Communication, knowledge sharing
- Project building

DESIGNING DATA SCIENCE PROJECTS

Data value

Exploration of value

- design theory
- data-based prospection
- innovation workshops

Data analytics

Problem formulation
Problem solving

- specialized teams
- RAMPs / training sprints
- data challenges

Design and Innovation strategy workshops

- Putting **domain scientists**, **data scientists**, and **management scientist** in the same room
- Getting them **understand** each other
- Keeping them **collectively creative**
- The goal: **identifying** and **defining projects**
  - low-hanging fruits
  - breakthrough projects
  - long-term vision

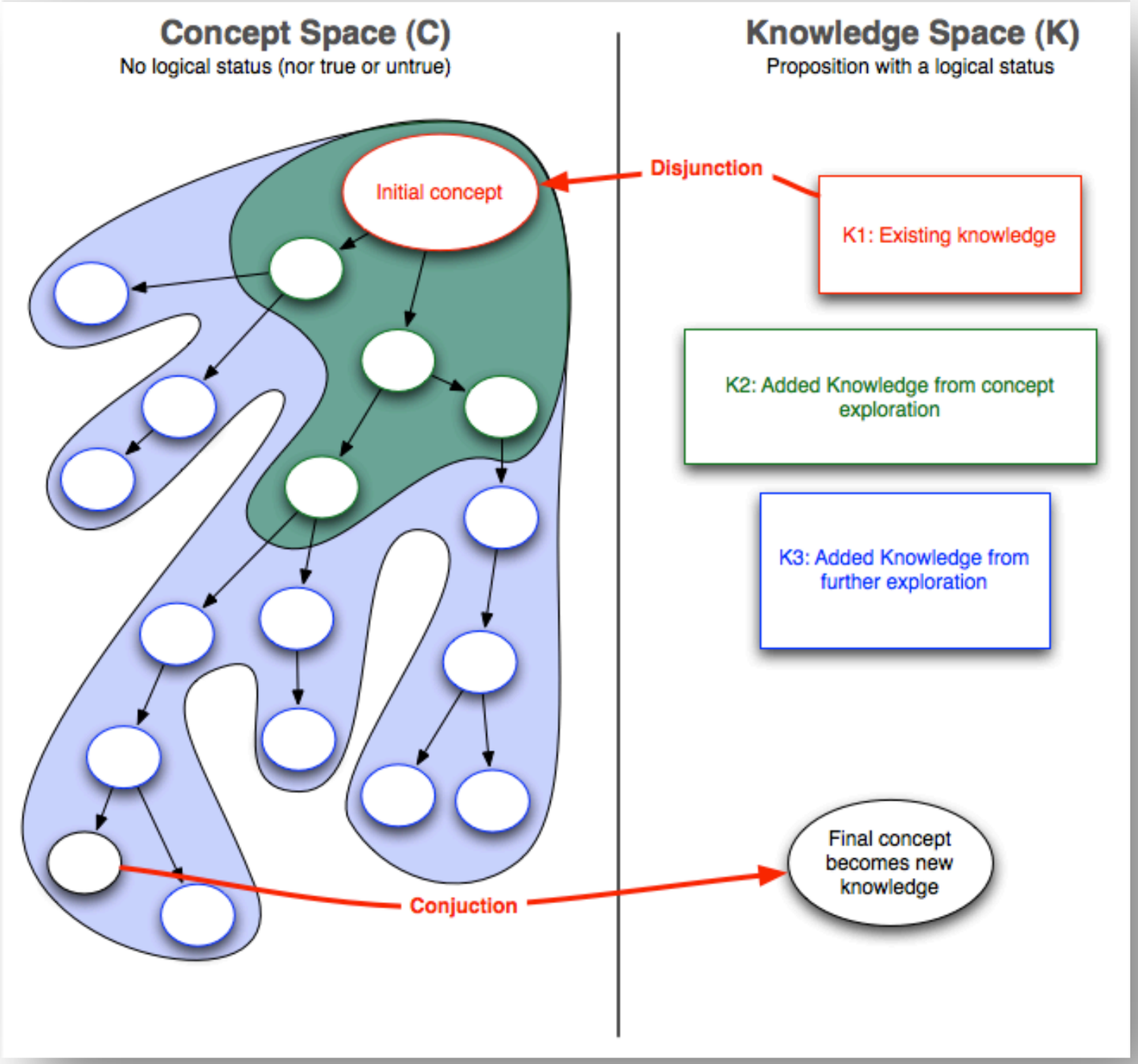
Design and Innovation strategy workshops

C/K design theory

innovative design

=

interaction and joint expansion of **concepts** and **knowledge**



DESIGN AND INNOVATION STRATEGY WORKSHOPS

DKCP process: linearizing C-K dynamics

RÉAU Prix de l'innovation brevetée 2013 du groupe Safran

Ils feront voler les hélicos avec moins de carburant

Romain Thiriet *(à gauche)* et Patrick Marconi, ingénieurs chez Turbomeca, ont eu l'idée de mettre deux moteurs de puissance différente et capables de démarrer en quatre secondes sur les hélicoptères pour réduire jusqu'à 15 % de leur consommation de carburant.

IT PLATFORM FOR LINKED DATA

http://io.datascience-paris-saclay.fr/

- A **window** to **open data** at Paris-Saclay
- We are **not storing** or handling existing large data sets
- Rather **indexing**, **linking**, and **mapping**, embedding in the worldwide linked data (RDF) ecosystem
- Storing **small data sets** of small teams is possible
- Subsets of large sets for **prototyping**
- Or simply store **metadata plus pointer**

IT PLATFORM FOR LINKED DATA

https://io.datascience-par ×

https://io.datascience-paris-saclay.fr

ALPHA Version 0.2

Paris-Saclay
Center for Data Science

DATA DOCS APP

Log in Register

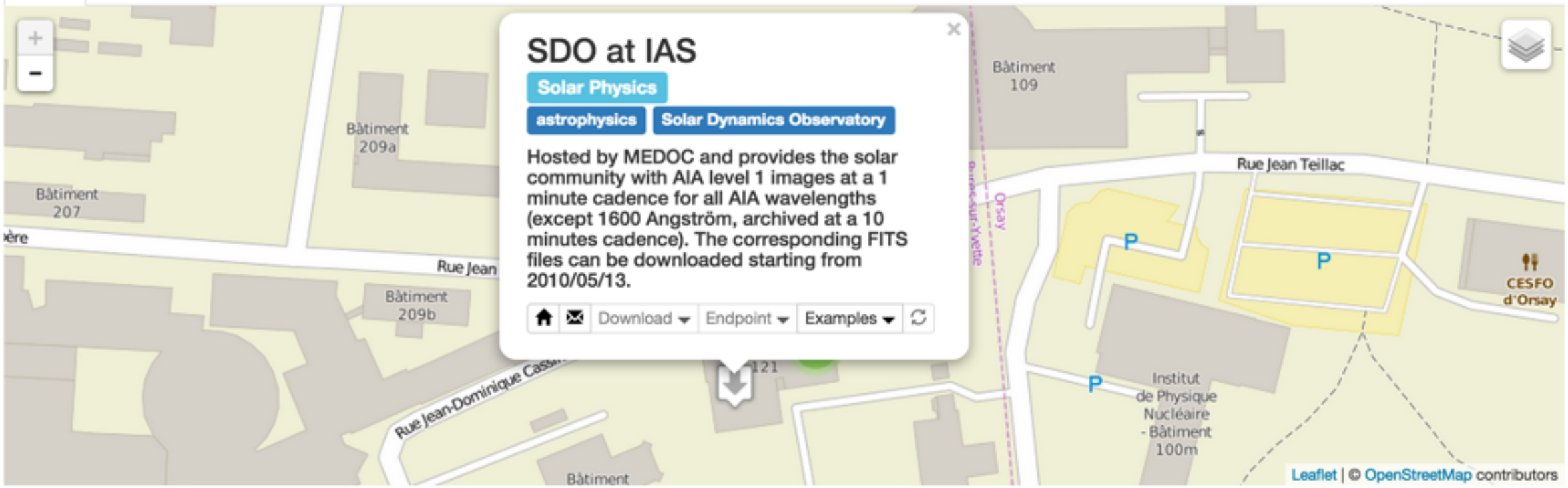
Search a dataset... very soor Search Advanced

FEEDBACK

Search an Open Dataset at Paris-Saclay

Locate on the map the actual open datasets.

Map Graph



Follow @SaclayCDS 62 followers Tweet 2 g+1 0 Follow @SaclayCDS

Contact | About | Jobs | Privacy & Cookies | Accessibility | Developer Docs | Issues & new features

BorderCloud

What is new?

"As the flow of data increases, it is increasingly processed, analyzed, and acted upon by machines, not humans."

NYU-CDS manifesto

What is new?

- We have the **data**
  - statistical / physical modeling is less important
  - data-driven prediction
- We have the **computational power**
- We have the **algorithms**
  - deep learning breakthrough: image, speech, language
  - closing on AI, step by step

TRAINING SPRINTS

- Single-day **training sessions**
  - **20-40** participants
  - focusing on a **single subject** (deep learning, model tuning, functional data, etc.)
  - preparing RAMPs