DEEP LEARNING AND THE SYSTEMIC CHALLENGES OF DATA SCIENCE INITIATIVES

BALÁZS KÉGL

CNRS & University Paris Saclay

I’m not going to explain deep learning in detail

Rather: give an overview of **what** you can do with it

DEEP LEARNING COURSES

- <u>Vincent Vanhoucke</u> (Google)
- <u>Hugo Larochelle</u> (Twitter)
- <u>Andrew Ng</u> (Baidu)
- <u>Nando de Freitas</u> (Oxford, Google DeepMind)

Your challenges are
not technological
but organizational

WHY CHALLENGES ARE ORGANIZATIONAL?

- Technology is **disruptive**
- The **current organization of research** is **half broken** and changing
  - Misplaced incentives, interdisciplinarity, peer-reviewed publications, code vs papers, funding, reproducibility, questions around data-driven scientific method
- We are using **few of the tools** developed mainly in industry to **manage disruptive innovation**

Outline

- Intro to **deep learning**
- The **PS-CDS**
- The data science ecosystem: **challenges**
- Some **tools**

DATA-DRIVEN INFERENCE

- You have a prediction or inference problem
  $y = f(x)$
  - X: photo, spectrum, y: galaxy/star and redshift
  - X: calorimetric image, y: particle parameters
  - X: particle parameters, y: calorimetric image

DATA-DRIVEN INFERENCE

- You have a **prediction** or **inference** problem
  $y = f(x)$
- You have **no model to fit**, but a **large set** of (x, y) pairs
  - The source is (typically) either
    - **observation** + human **labeling**
    - **simulation**
- And a **loss function** $L(y, y_{pred})$

The shallow learning paradigm

- The solution
  - Design/define a lot of application/domain-dependent cues/ features $h_j(x)$
  - Learn a **linear function** $f(x) = \sum_j w_j h_j(x)$
    - shallow neural nets, ensemble methods, kernel methods
  - **Works well for most** of the practical problems (but **not all**)

Your most important question is:

are you in the “not all” part?

THE DEEP LEARNING PARADIGM

- The solution
  - Parametrize $f(x) = f(x, w)$
  - w is very **high-dimensional**, f has a **lot of capacity**
  - make everything **quasi-differentiable** (L and $f(.,w)$)
  - regularize (L_1, L_2, dropout, etc.)
  - learn w using **stochastic gradient descent**

Shallow to Deep Learning

- From a design (user) point of view
  - Instead of hand-crafting (families of) informative features, you will design a system of reusable blocks of differentiable functions
  - Close to the data, domain knowledge is important
  - Deeper layers are rather general
  - A lot of partly reusable trial-end-error tricks
  - Pre-trained and saved networks/blocks,
  - “dark knowledge”

State of the Art

- Computer vision
  - close to the data: convolutional layers, max pooling
- Sequential data (speech, language)
  - recurrent nets, networks with memory (LSTM)
- Multi-modal embeddings (eg: caption generation)
- (Half) future: robotics, Turing machine, reasoning, neural simulators

THE DEEP LEARNING PARADIGM

- Tools, techniques
  - deep learning libraries (Theano, TensorFlow, Caffe, Torch)
  - automatic differentiation
  - stochastic gradient descent
  - hyperparameter optimization
  - lots of data and machines (GPUs)

I will stop talking about science

Well, not really

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

UNIVERSITÉ PARIS-SACLAY

19 founding partners

AgroParisTech, CEA, CNRS, CentraleSupélec, ENS Cachan, ENSTA ParisTech, École Polytechnique, ENSAE ParisTech, HEC Paris, IHÉS, INRA, Inria, Institut d'Optique Graduate School, Inserm, Telecom ParisTech, Telecom SudParis, Institut Mines-Télécom, ONERA, Université Paris-Sud, Université de Versailles St-Quentin-en-Yvelines

UNIVERSITÉ PARIS-SACLAY

19 *fondateurs*

11 000 *chercheurs et enseignants-chercheurs*

60 000 *étudiants*

300 *laboratoires*

6 000 *doctorants*

8 000 *publications /an*

15 000 *étudiants en master*

15 % *de la recherche publique française*

8 *Schools*

10 *départements*

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

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

universite PARIS-SACLAY
Paris-Saclay Center for Data Science

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

Interfacing with HPC, cloud, storage, production, privacy, security

What is new?

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

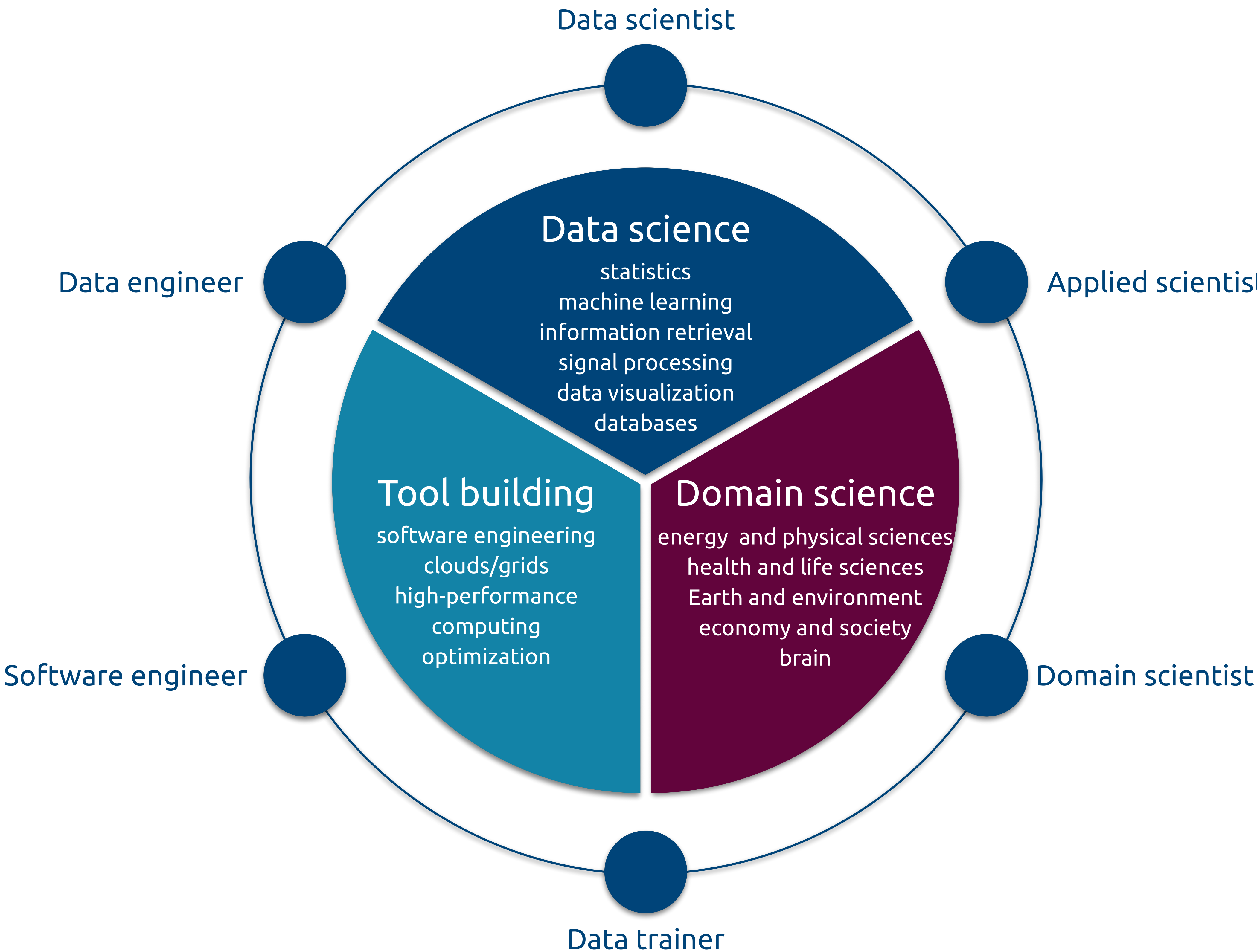
NYU-CDS manifesto

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

https://medium.com/@balazskegl

THE DATA SCIENCE LANDSCAPE



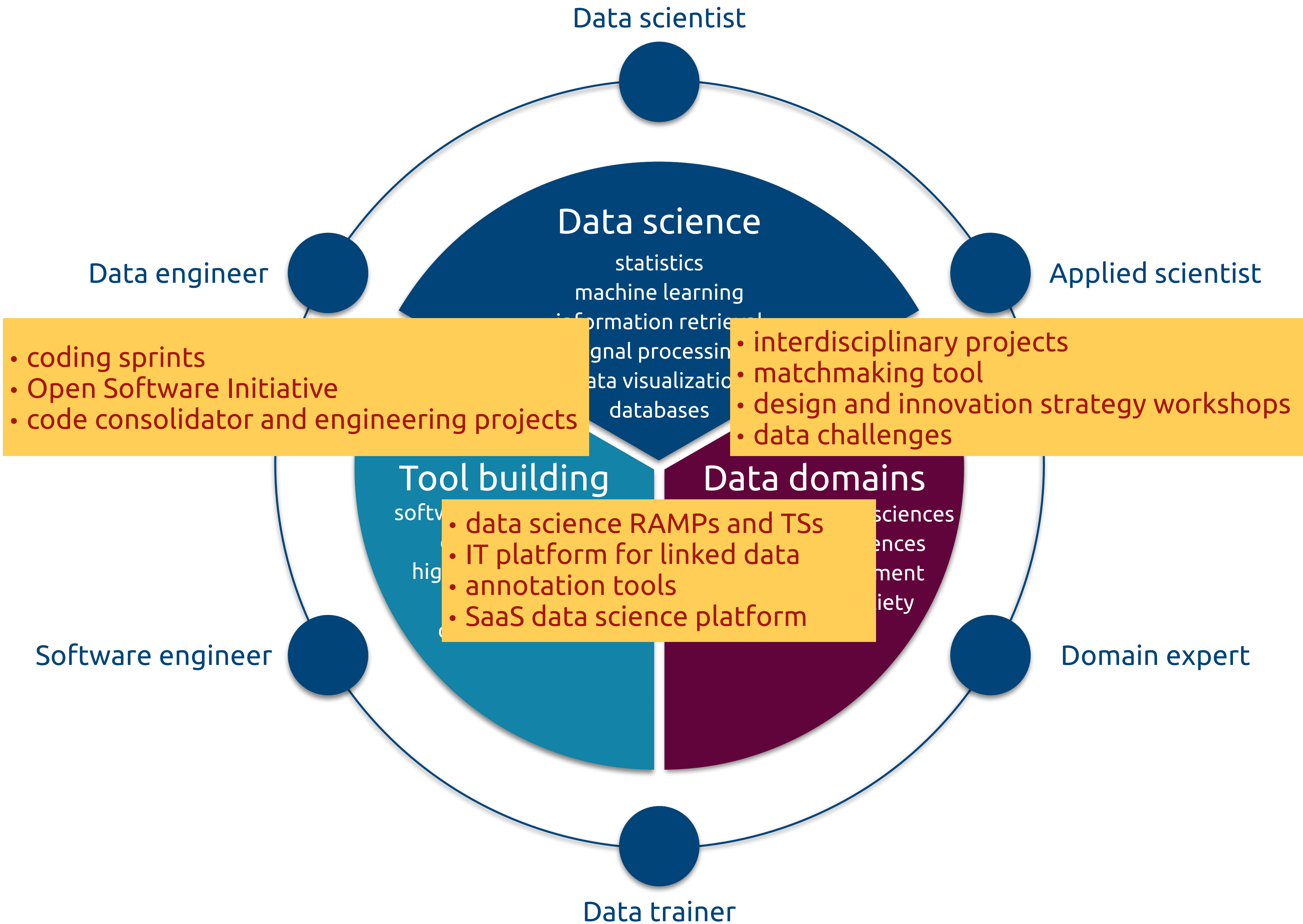
CHALLENGES

- (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



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

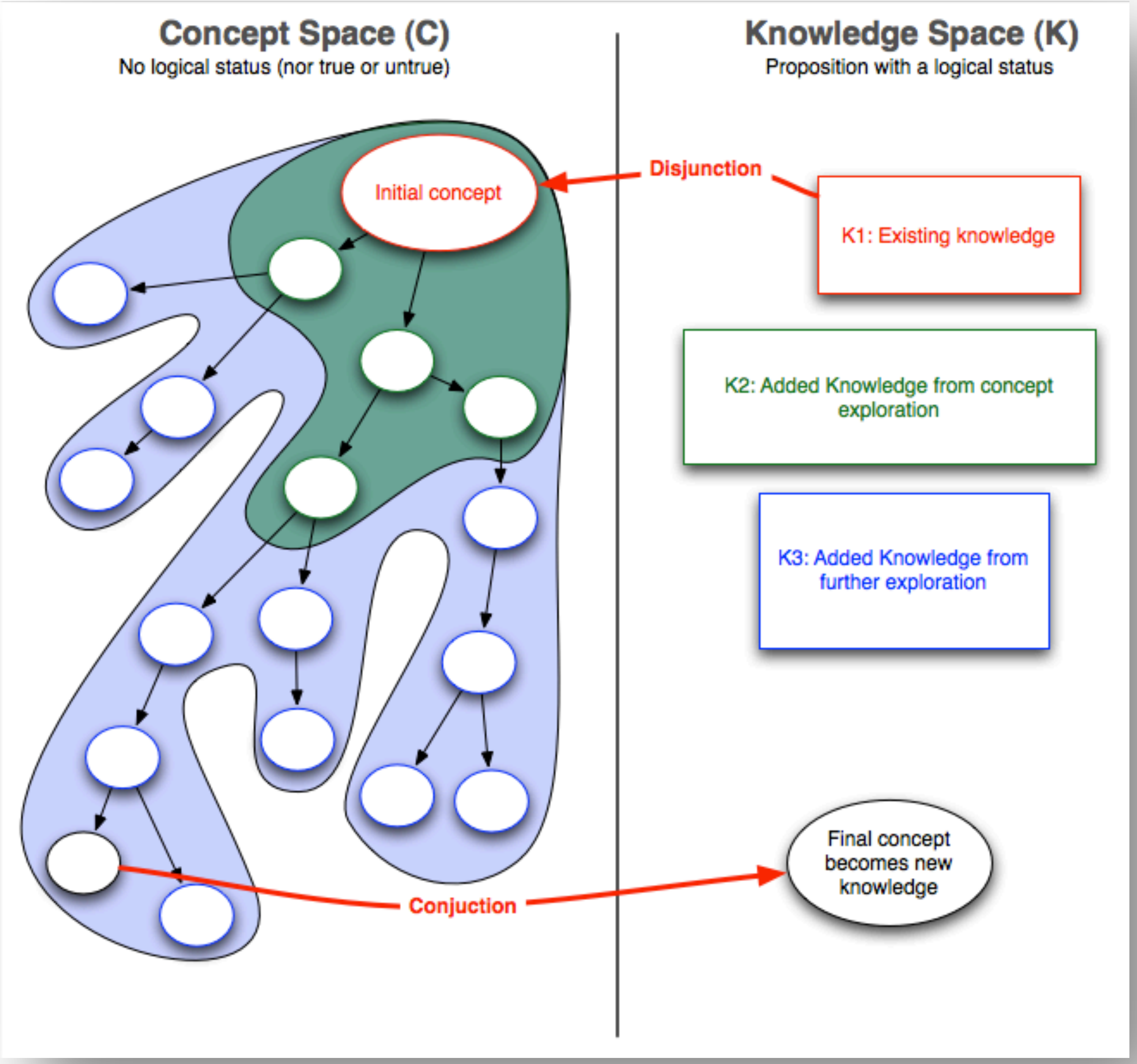
Design and Innovation strategy workshops

C/K design theory

innovative design

=

interaction and joint expansion of concepts and knowledge



DKCP process: linearizing C-K dynamics

Initialisation → [K] Knowledge sharing Workshops → [C] IFM-Design Workshops → [P] Project building → [RUN]

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.

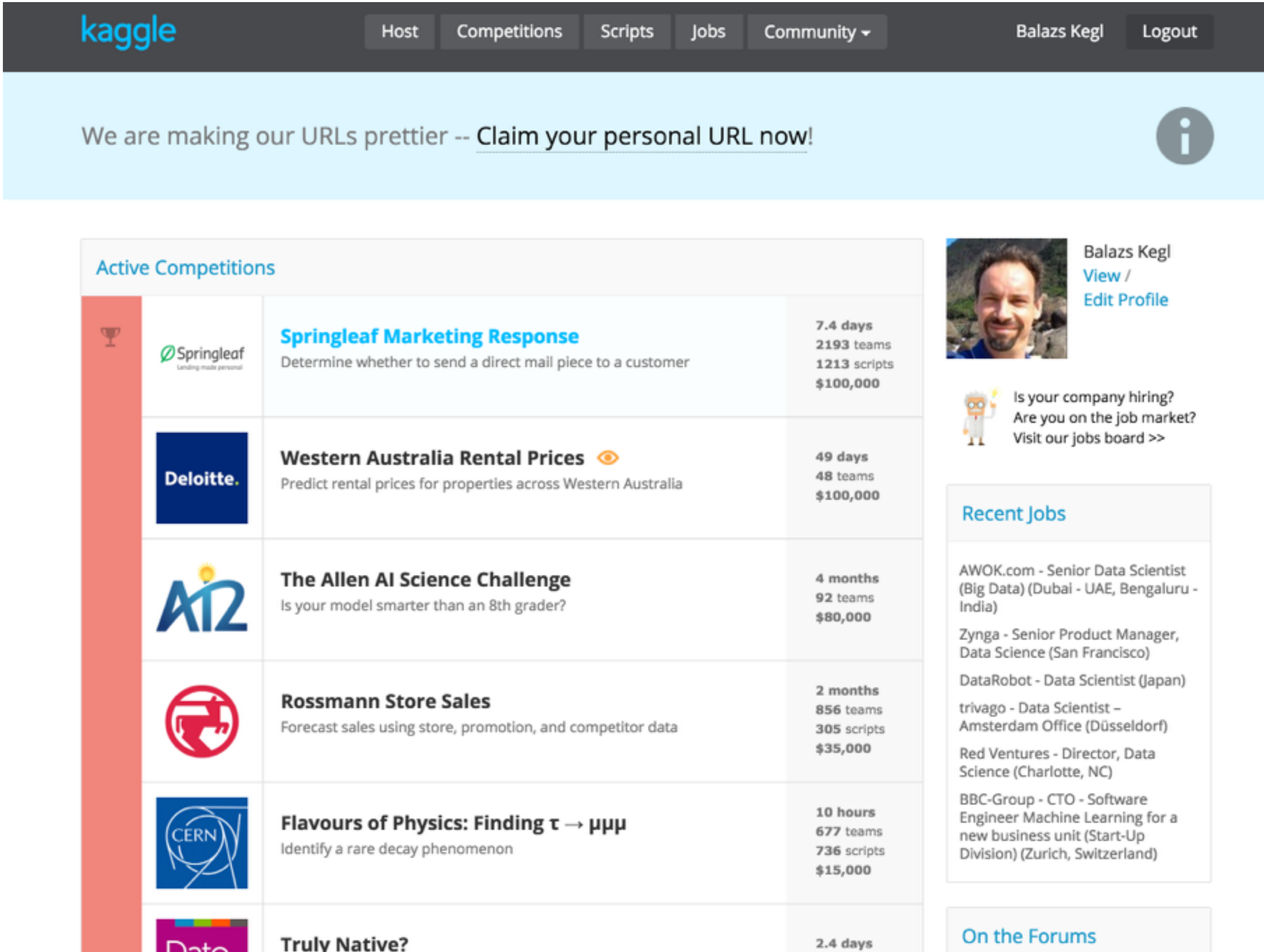
Three analytics tools for initiating domain-data science interactions

Data challenges

Rapid analytics and model prototyping (RAMP)

Training sprints (TS)

DATA CHALLENGES

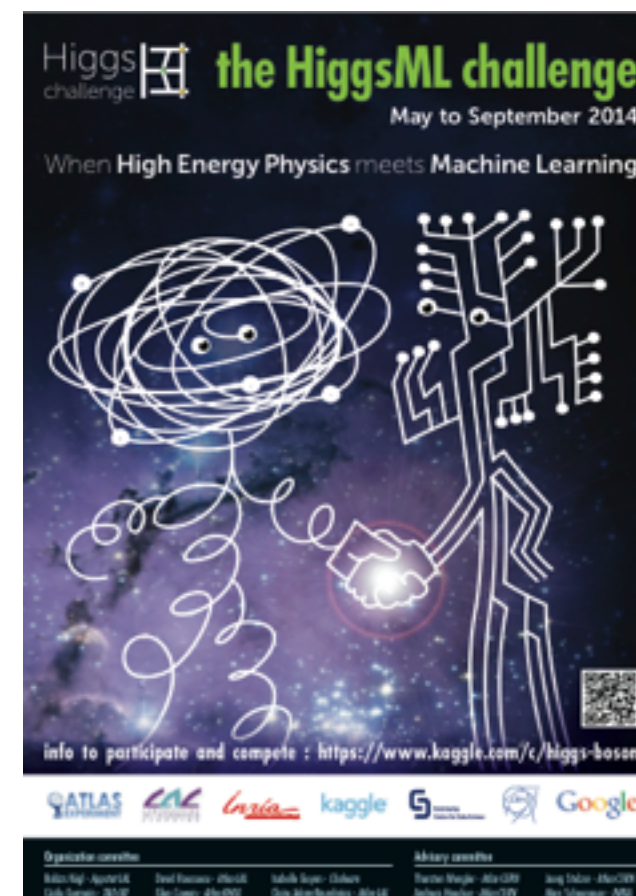


cience

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**

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	[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)

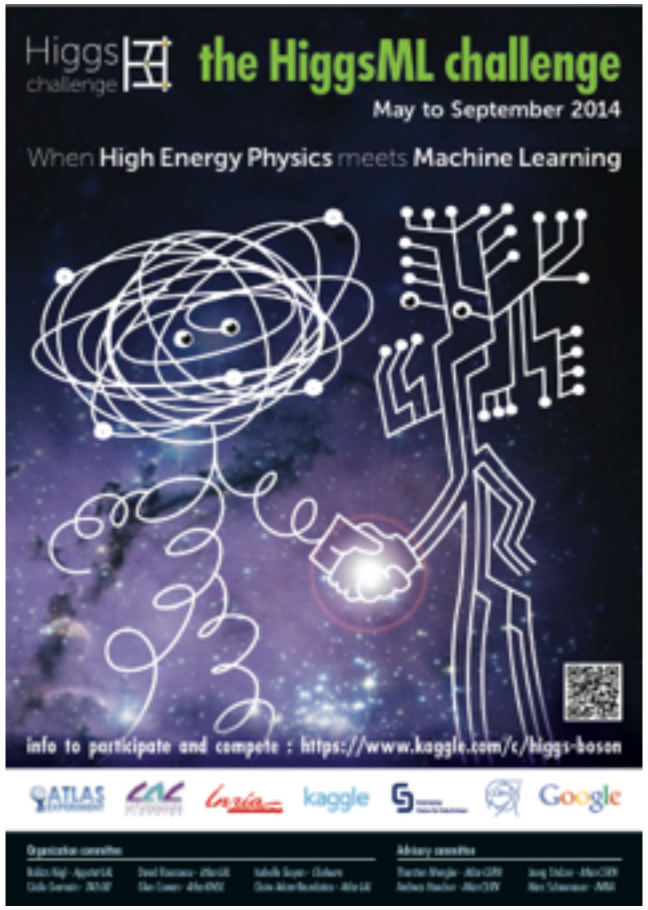
HUGE PUBLICITY

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**
- **Collaboration building**

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

Training sprints

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

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

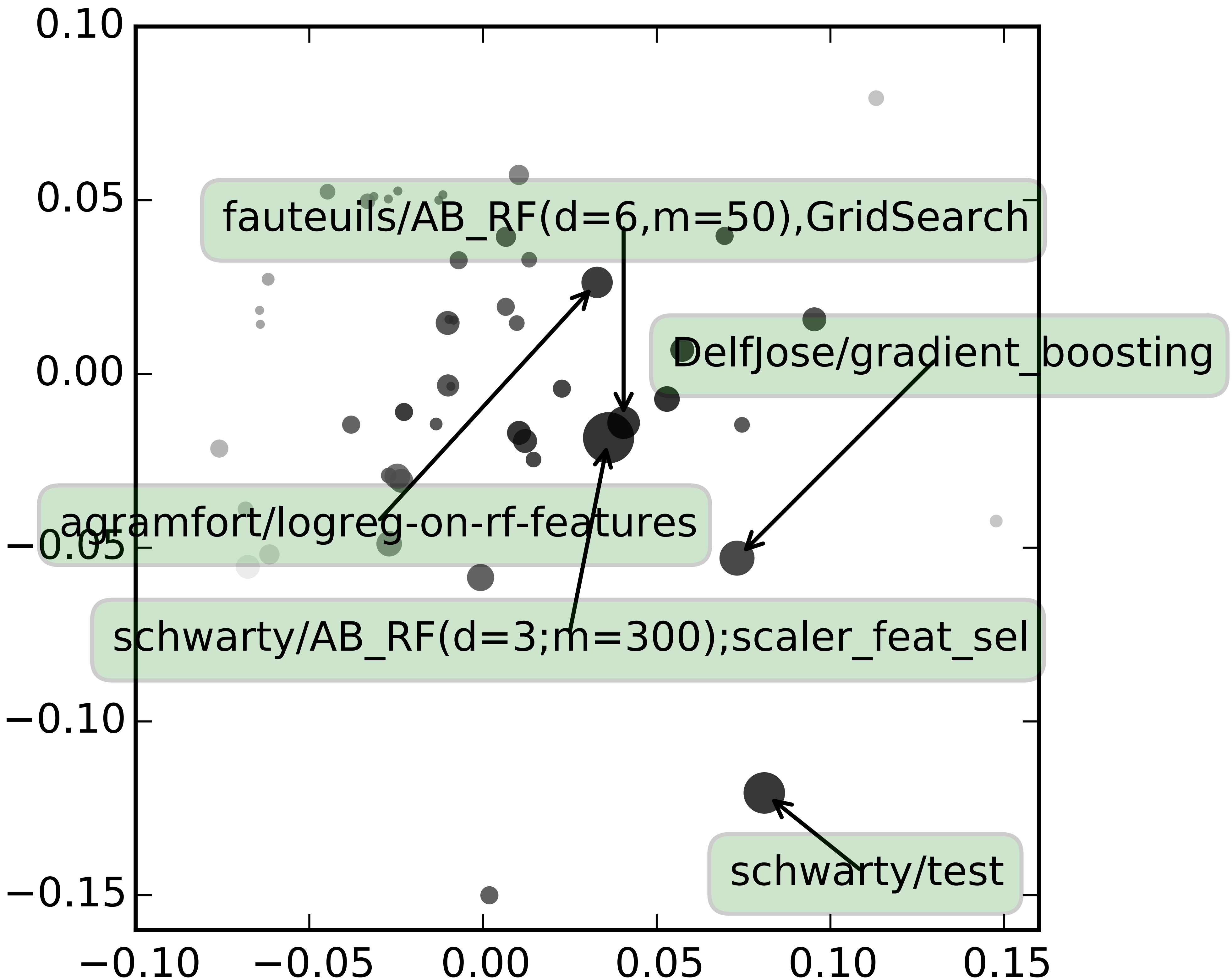
Leaderboard › kegl › MF.AB(20;RF(100;5))_d1 › model.py

```python
from sklearn.ensemble import RandomForestClassifier, AdaBoostClassifier
from sklearn.preprocessing import Imputer
from sklearn.pipeline import Pipeline
from sklearn.base import BaseEstimator

class Classifier(BaseEstimator):
    def __init__(self):
        self.clf = Pipeline([('imputer', Imputer(strategy='most_frequent')),
        ('rf', AdaBoostClassifier(base_estimator=RandomForestClassifier(max_depth=5,
n_estimators=100),
                                  n_estimators=20))])

    def fit(self, X, y):
        self.clf.fit(X, y)

    def predict(self, X):
        return self.clf.predict(X)

    def predict_proba(self, X):
        return self.clf.predict_proba(X)
```

Archive

model.py

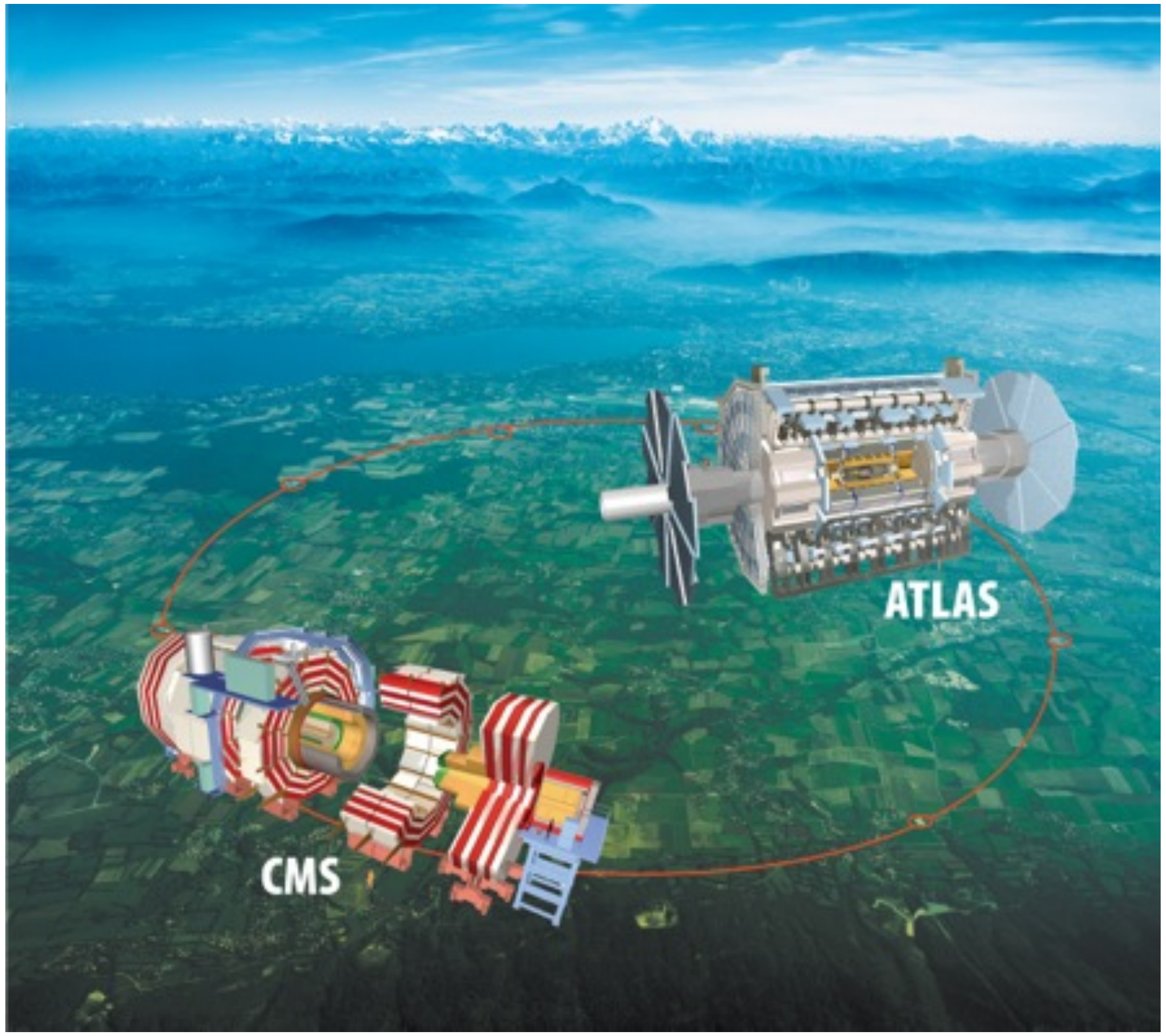
ANALYTICS TOOLS TO MONITOR PROGRESS

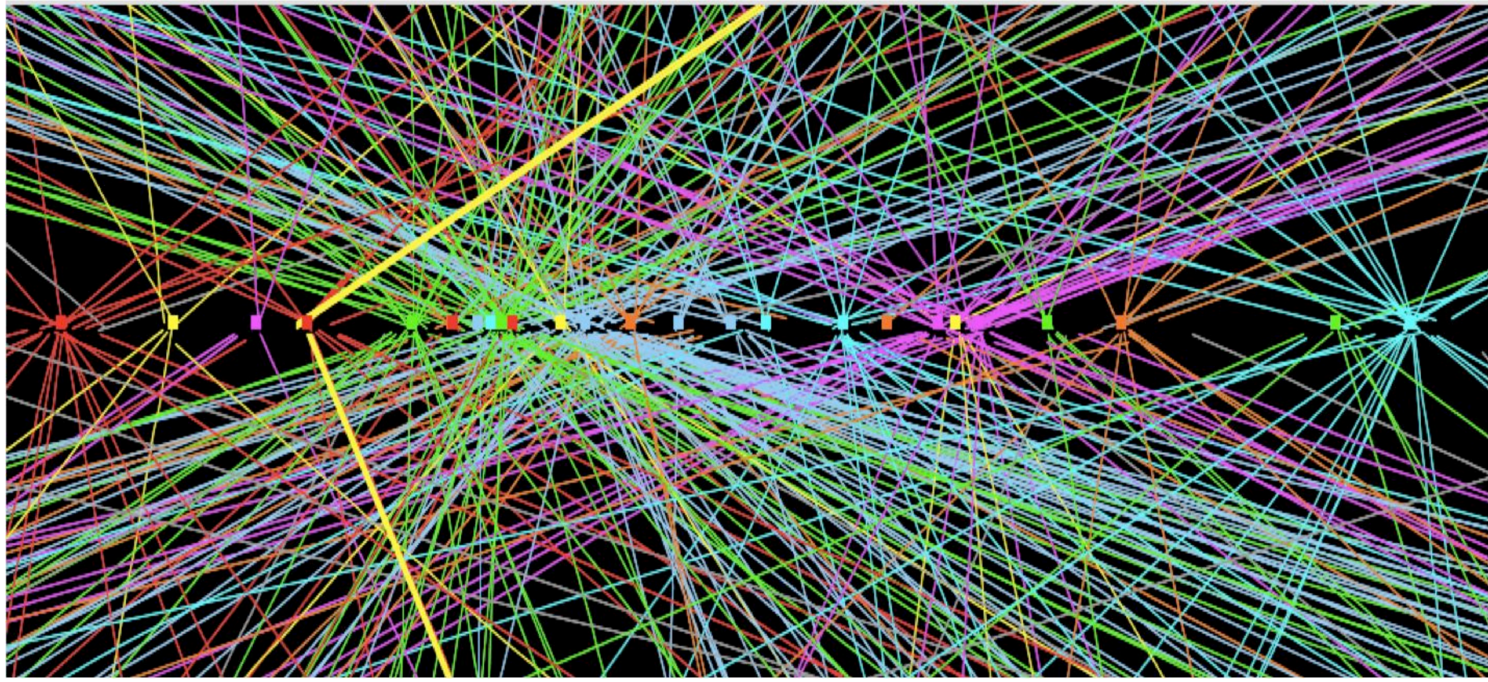


Rapid analytics and model prototyping

2015 Jan 15

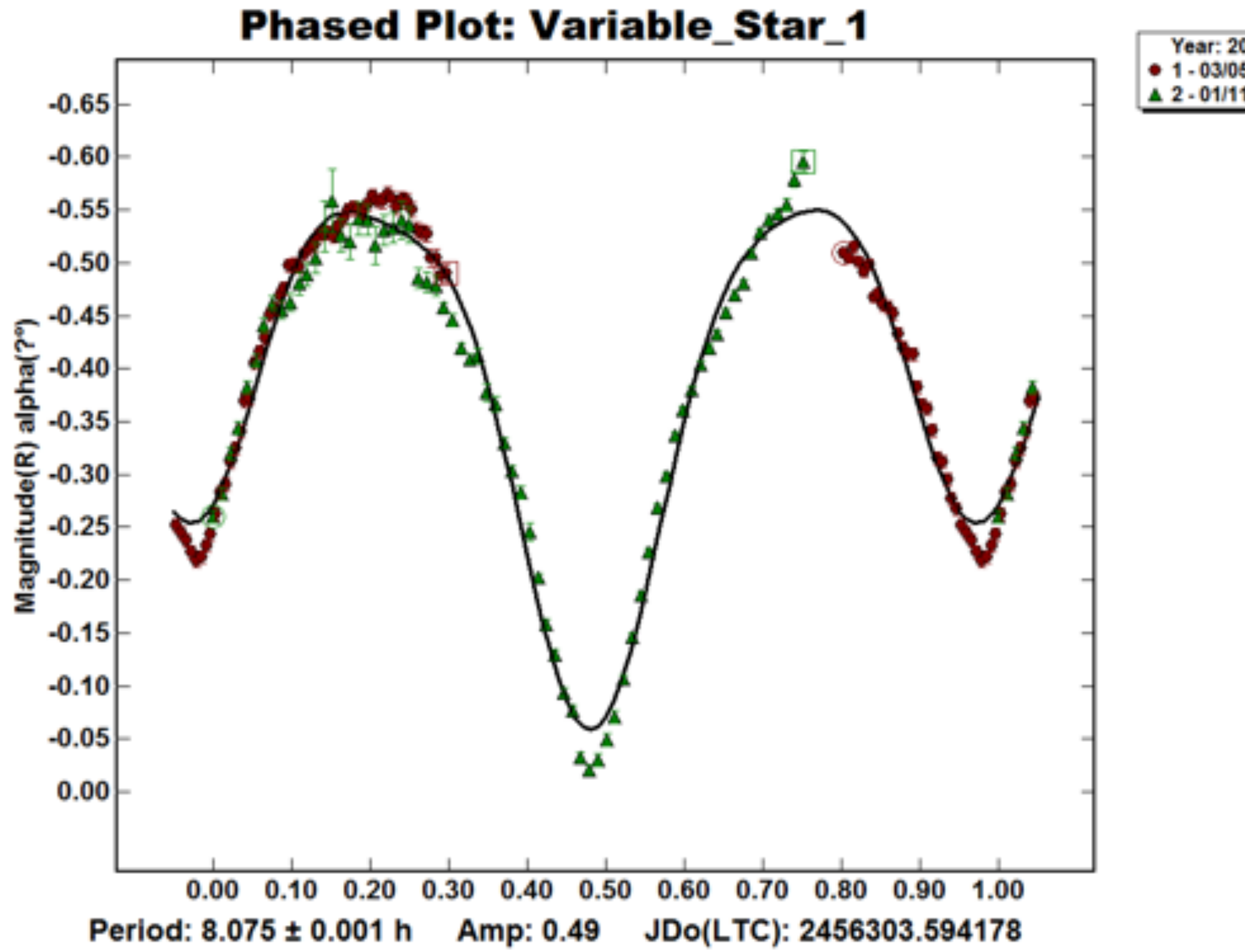
The HiggsML challenge



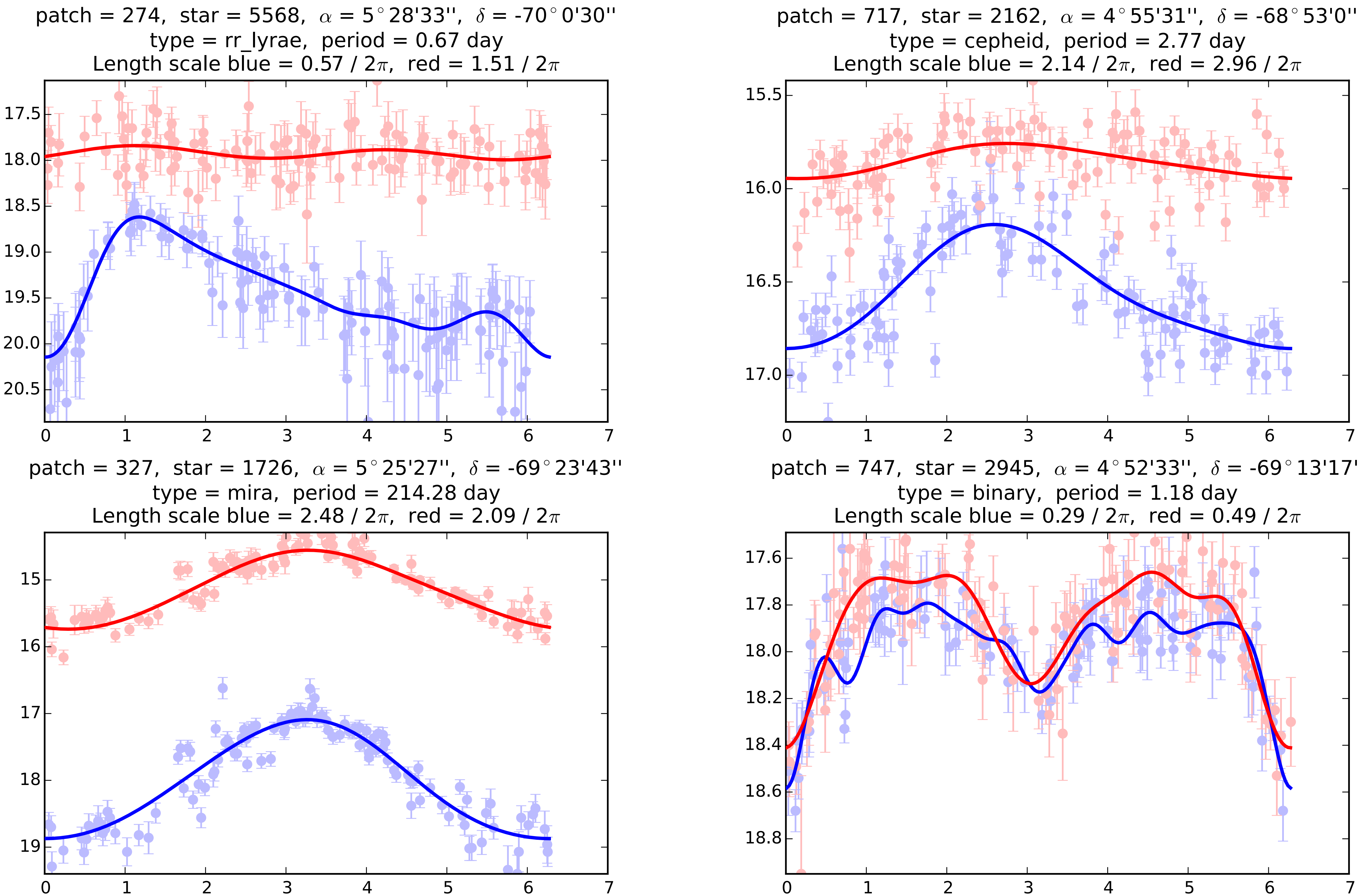


2015 Apr 10

Classifying variable stars



VARIABLE STARS



VARIABLE STARS

RAMP
Rapid Analytics and Model Prototyping

Variable star type prediction

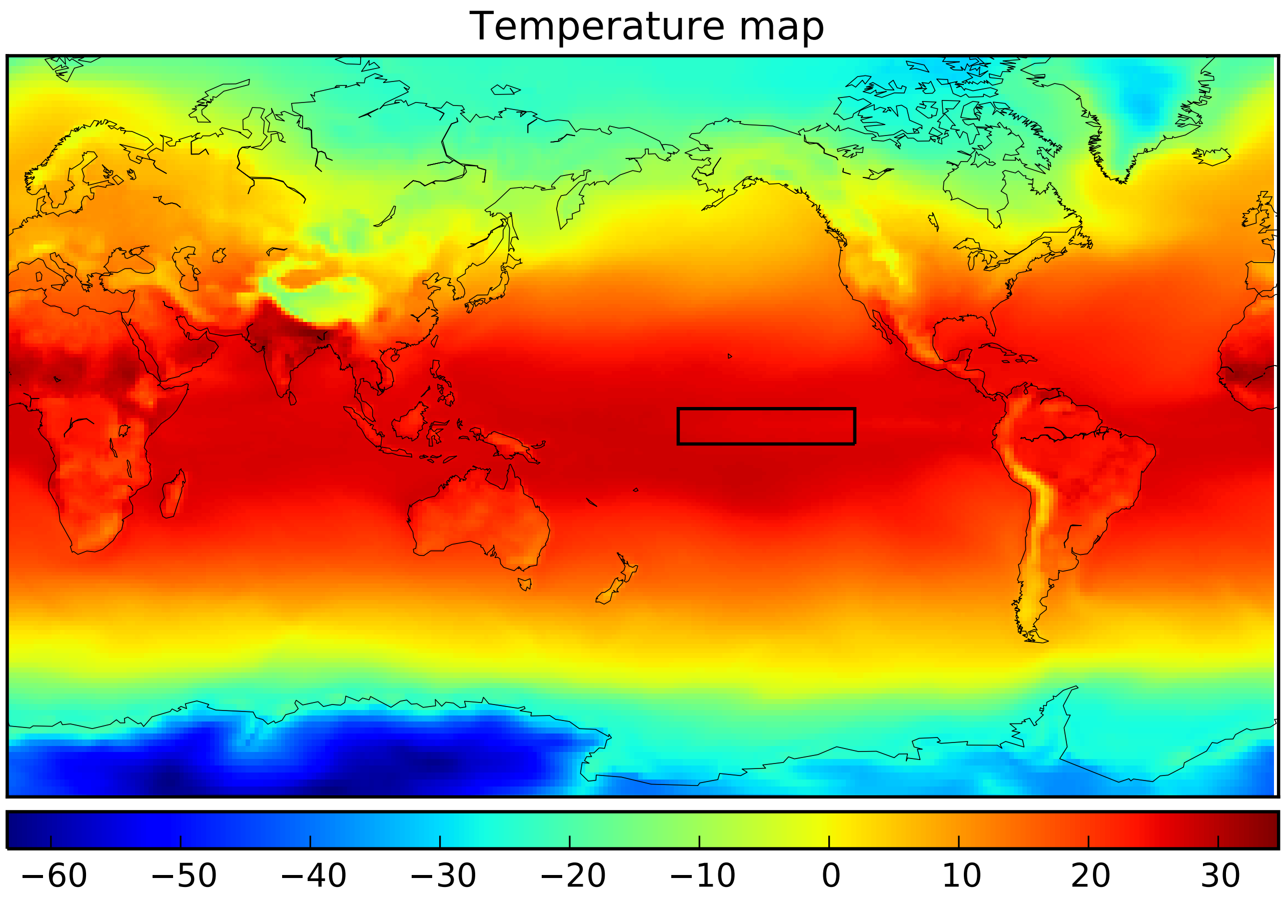
Leaderboard

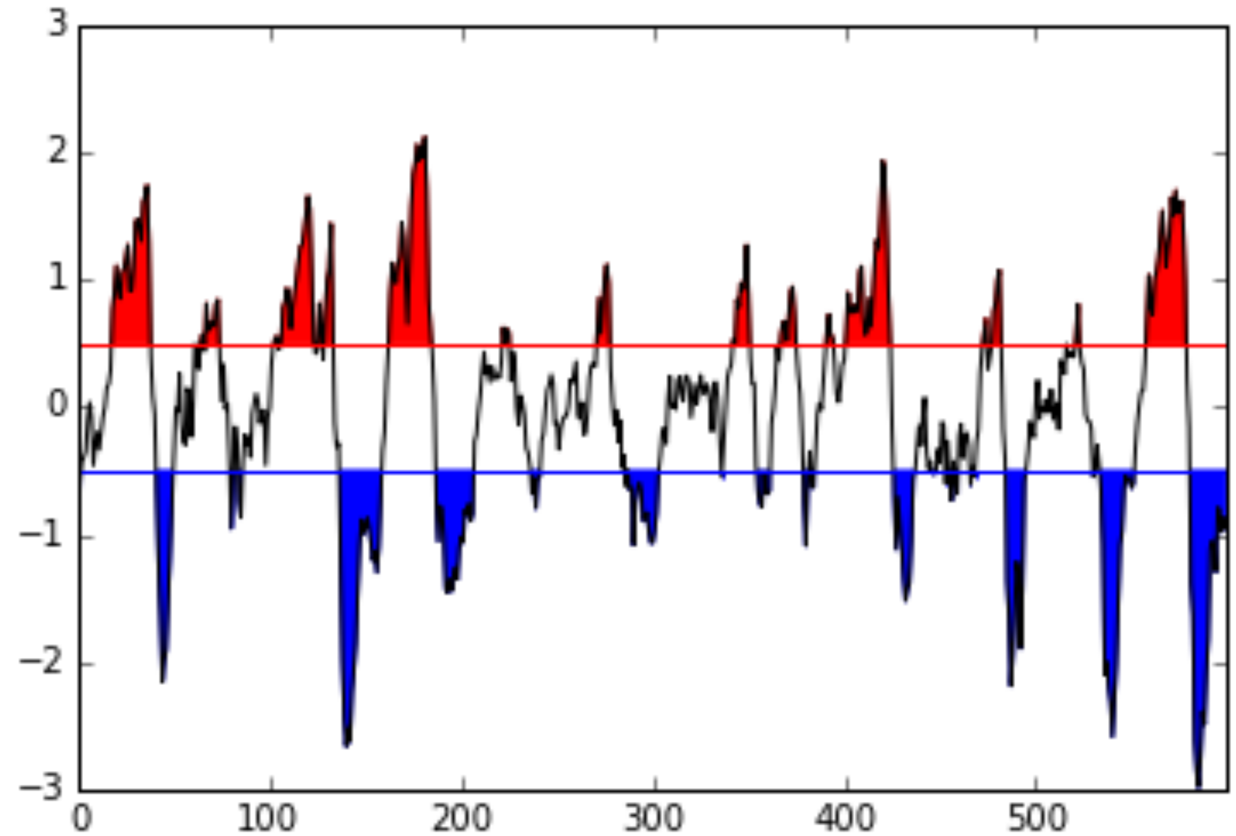
rank	team	model	commit	score ▲	contributivity	train time	test time
1	LesTortuesNinja	gp_fixed_3	2015-04-11 00:48:59	0.9621	19	117	103
2	agramfort	gp_rf30_adaboost10_v2	2015-04-10 14:30:50	0.9596	3	117	104
3	Overfitters	stack_wavelet	2015-04-10 17:03:27	0.9588	6	313	132
7	delphine	feature_selection	2015-04-10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





RAPID ANALYTICS AND MODEL PROTOTYPING

RAMP
Rapid Analytics and Model Prototyping

El Nino prediction

Leaderboard

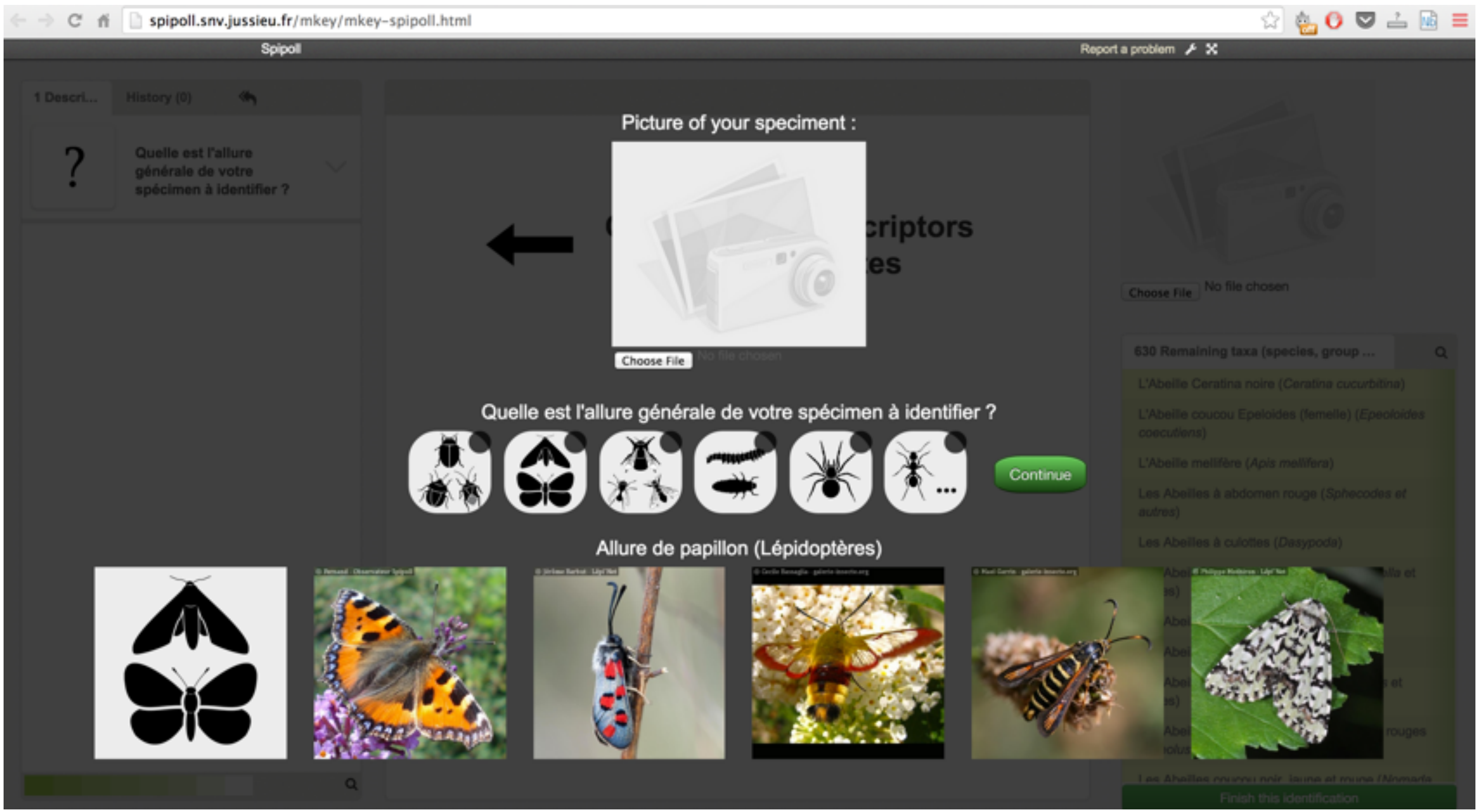
rank	team	model	commit	score ▲	contributivity	train time	test time
1	CloudySunset	more_samples	2015-09-26 22:46:36	0.4336	6	95	0
2	slay	oceanmask	2015-09-26 22:46:52	0.4377	1	26	3
3	slay	grd_gbrs	2015-09-26 21:47:10	0.4390	0	30	3
4	ChrisFarley	gbr_1	2015-09-26 22:41:37	0.4390	0	30	3

RMSE improvement: 0.9˚C to 0.4˚C

rank	team	model	commit	score	contributivity	train time	test time
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

2015 October 8

Insect classification



RAPID ANALYTICS AND MODEL PROTOTYPING

RAMP
Rapid Analytics and Model Prototyping

Pollenating insect classification

Leaderboard

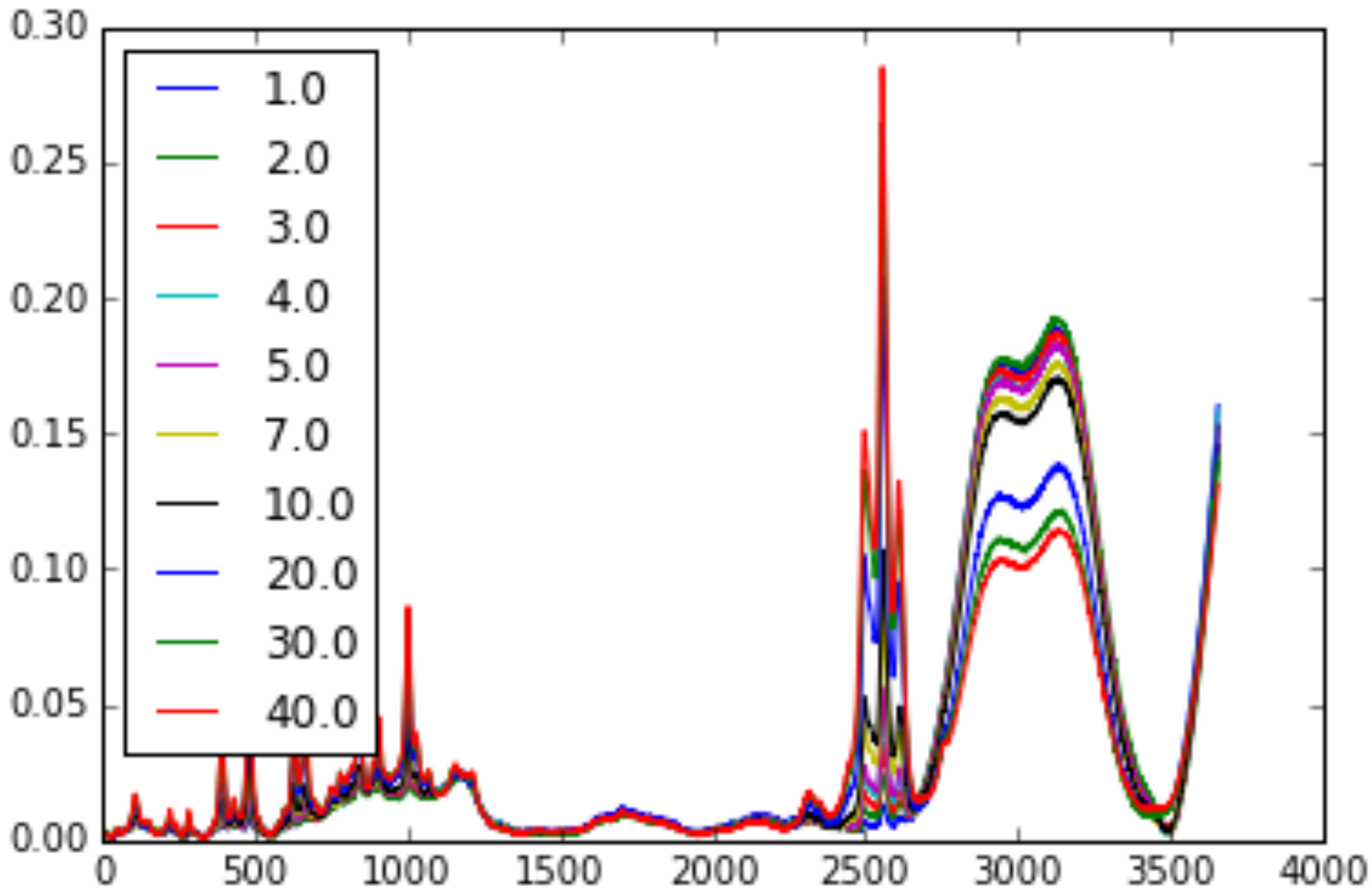
rank	team	model	commit	score ▲	contributivity	train time	test time
1	Florian	yousra_with_flip_rotation_gaussian_windo[...]	2015-10-08 18:11:52	0.7194	30	3735	1
2	Florian	yousra_with_flip_rotation_gaussian_windo[...]	2015-10-08 17:20:19	0.6812	2	2646	1
3	Issam	rotation_noreg_yousra_first_3	2015-10-08 17:31:38	0.6801	15	1235	1
4	Brutti	small_rot_fix	2015-10-08 18:01:18	0.6654	17	3757	1

accuracy improvement: 30% to 70%

rank	team	model	commit	score ▲	contributivity	train time	test time
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

2015 Fall

Drug identification from spectra



The RAMP tool

A prototyping tool for collaborative development of data science workflows

- Teaching support
- Networking and HR support
- Support for collaborative team work

Thank you!

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