# Rapid Analytics and Model Prototyping

Akın Kazakçı

MdC / Mines ParisTech
CGS

Alexandre Gramfort

MdC / Telecom ParisTech
LTCI

Balázs Kégl

DR / CNRS - UPSaclay
LAL & LRI

# OUTLINE

- Paris-Saclay Center for Data Science
  - the data science ecosystem
- Analytics tools
  - data challenges
  - rapid analytics and model prototyping

# DATA SCIENCE

Design of automated methods

to analyze massive and complex data

to extract useful information

# Data Science
# ≠
# Big Data

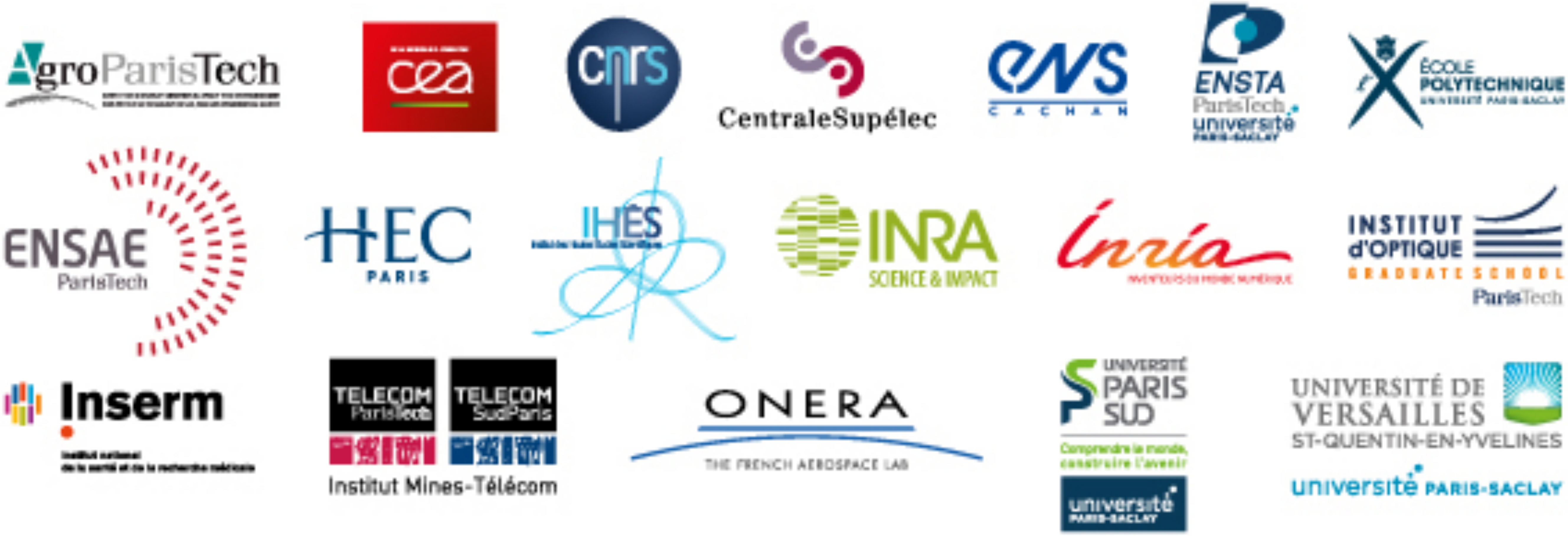
We are focusing on inference:

data → knowledge

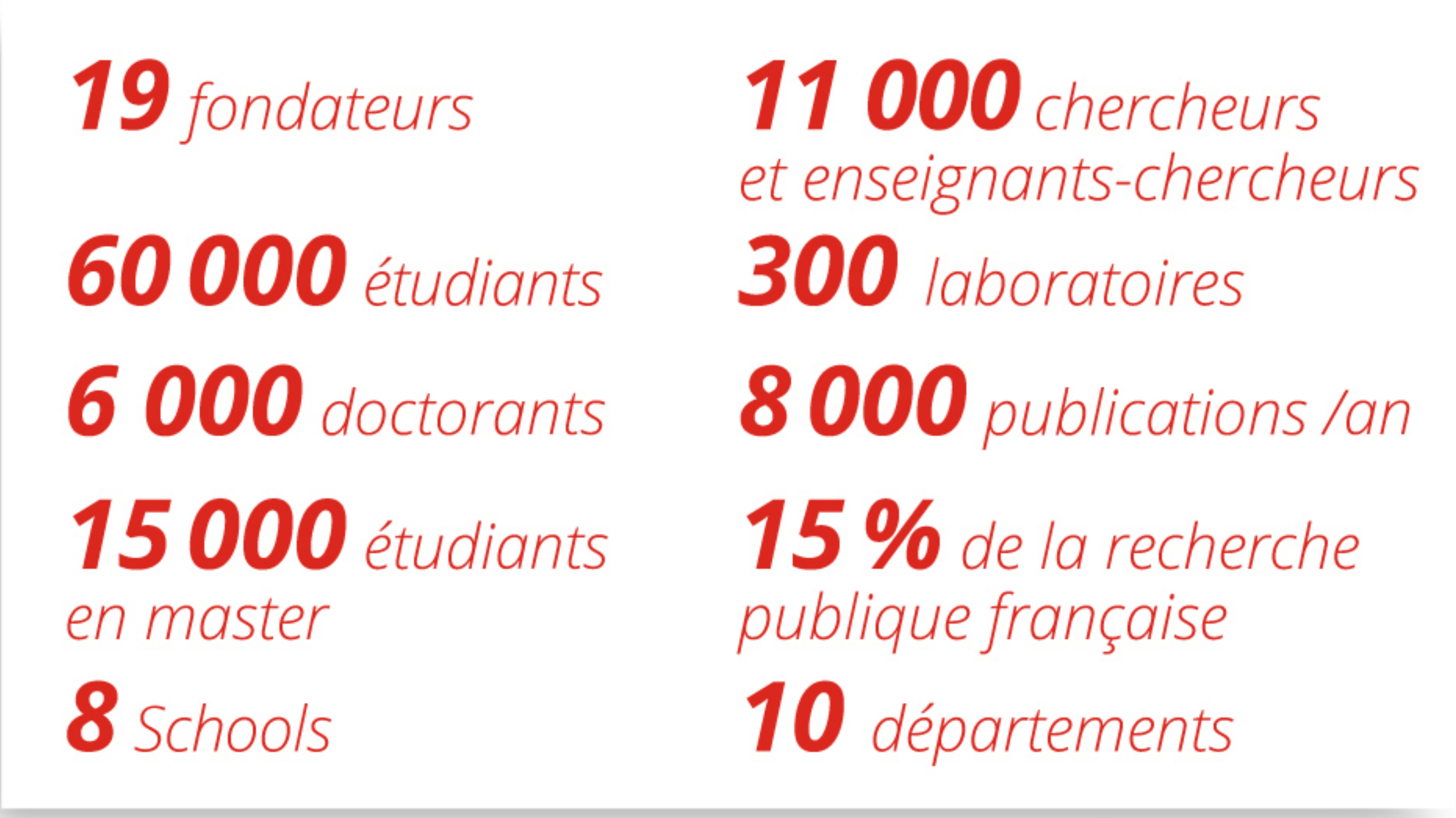
Interfacing with infrastructure, security, production

# UNIVERSITÉ PARIS-SACLAY

## 19 founding partners



# 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 ("lidex") to create cohesion**

universite PARIS-SACLAY

# 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

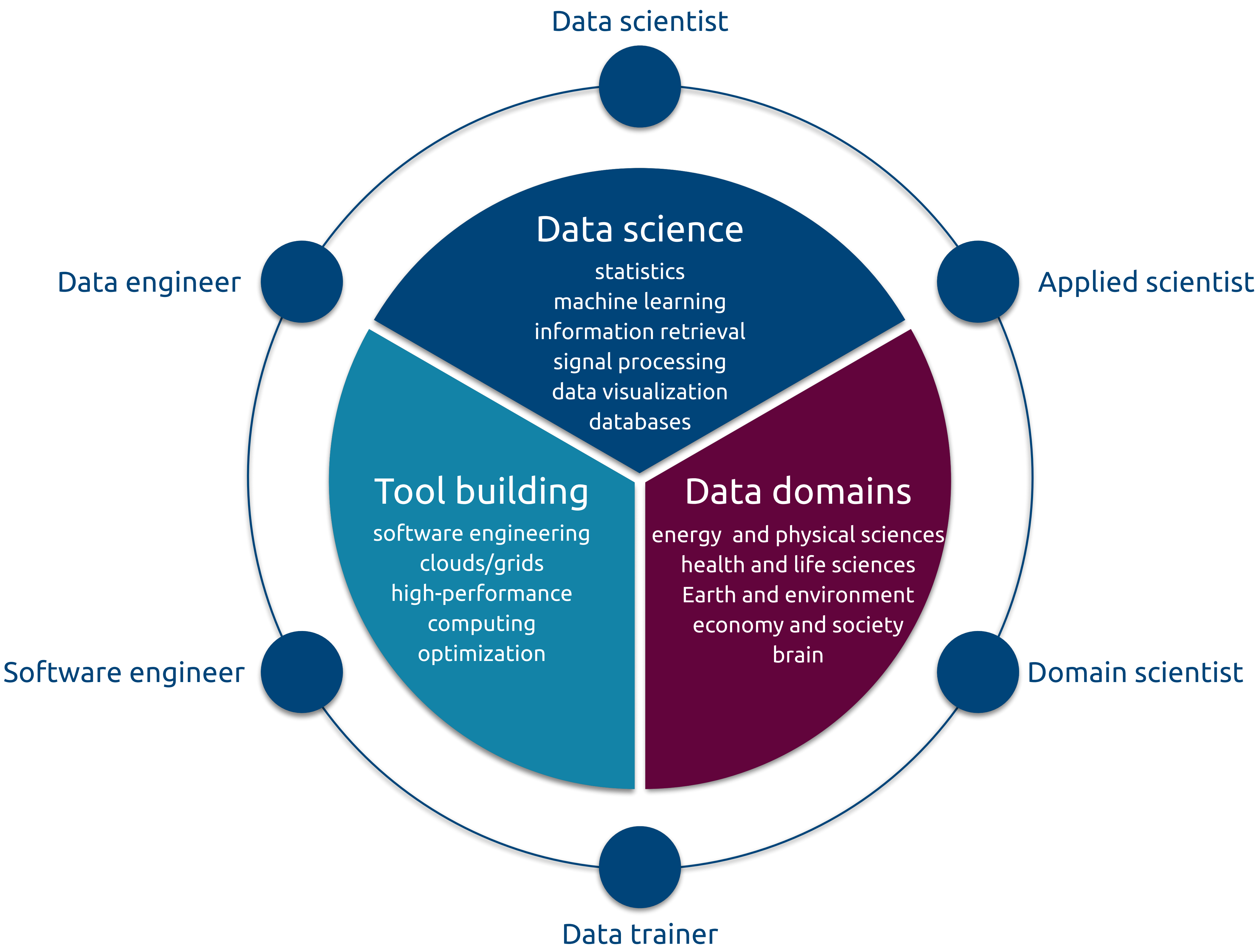
**Machine learning**
LRI/UPSud
LTCI/Telecom
CMLA/Cachan
LS/ENSAE
LIX/Polytechnique
MIA/Agro
CMA/Polytechnique
LSS/Supélec
CVN/Centrale
LMAS/Centrale
DTIM/ONERA
IBISC/UEvry
LIST/CEA

**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

# The data science ecosystem



Data scientist

Data engineer

Applied scientist

Software engineer

Domain scientist

Data trainer

## Data science

statistics
machine learning
information retrieval
signal processing
data visualization
databases

## Tool building

software engineering
clouds/grids
high-performance computing
optimization

## Data domains

energy and physical sciences
health and life sciences
Earth and environment
economy and society
brain

# Tools

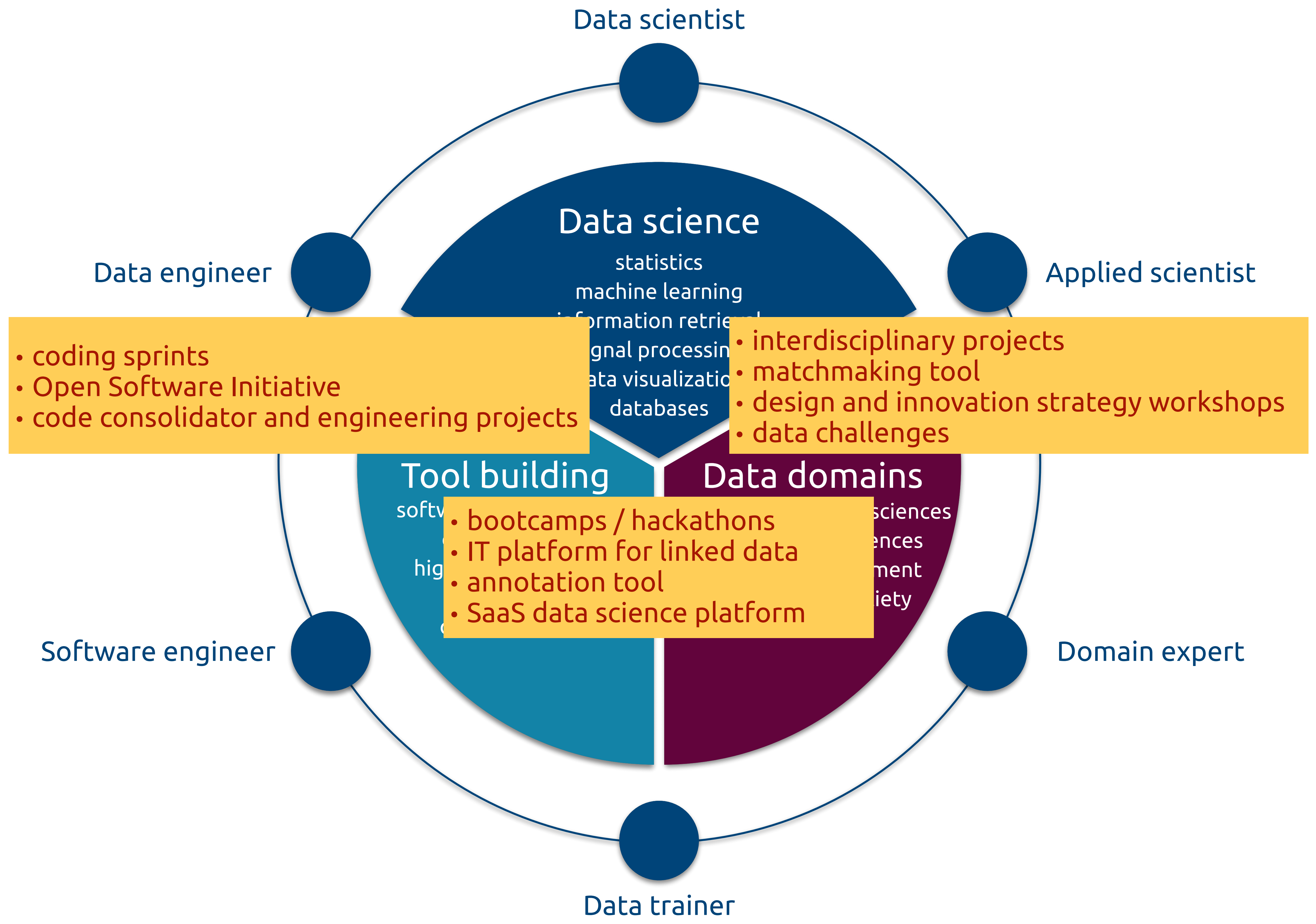
We are designing and learning to manage

tools

to accompany data science projects

with different needs

# TOOLS: LANDSCAPE TO ECOSYSTEM



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

- interdisciplinary projects
- matchmaking tool
- design and innovation strategy workshops
- data challenges

- bootcamps / hackathons
- IT platform for linked data
- annotation tool
- SaaS data science platform

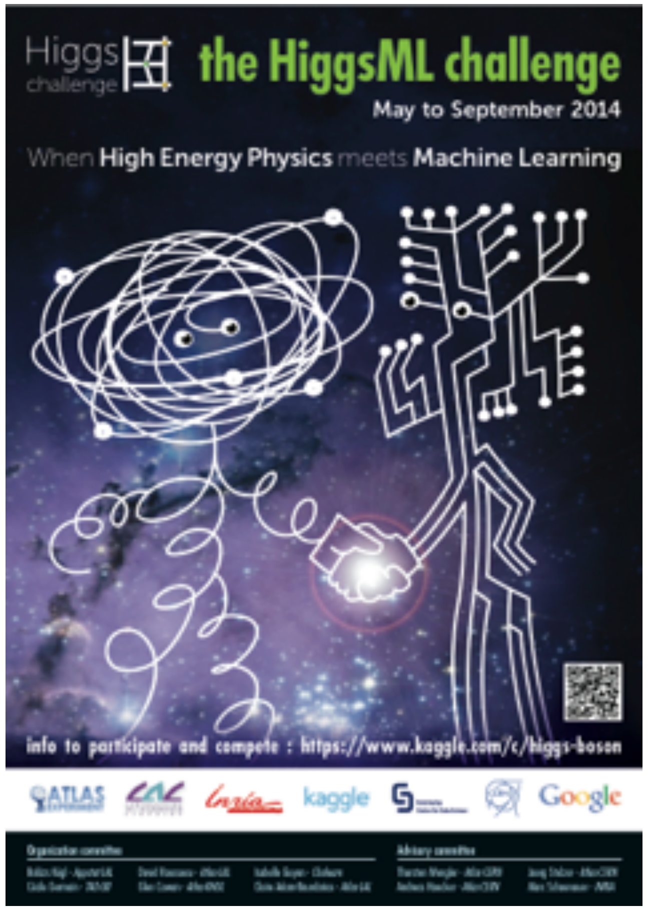
# Two analytics tools

## Data challenges

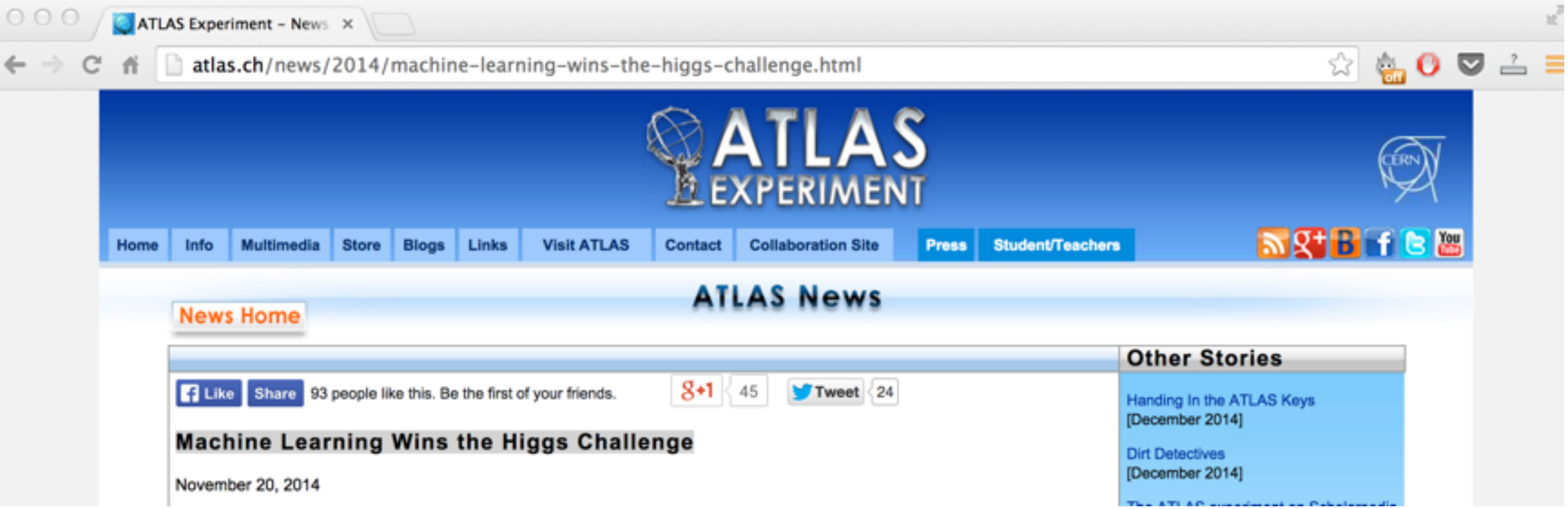
## Rapid Analytics and Model prototyping

# 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**


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

# 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.

Machine Learning Wins the Higgs Challenge

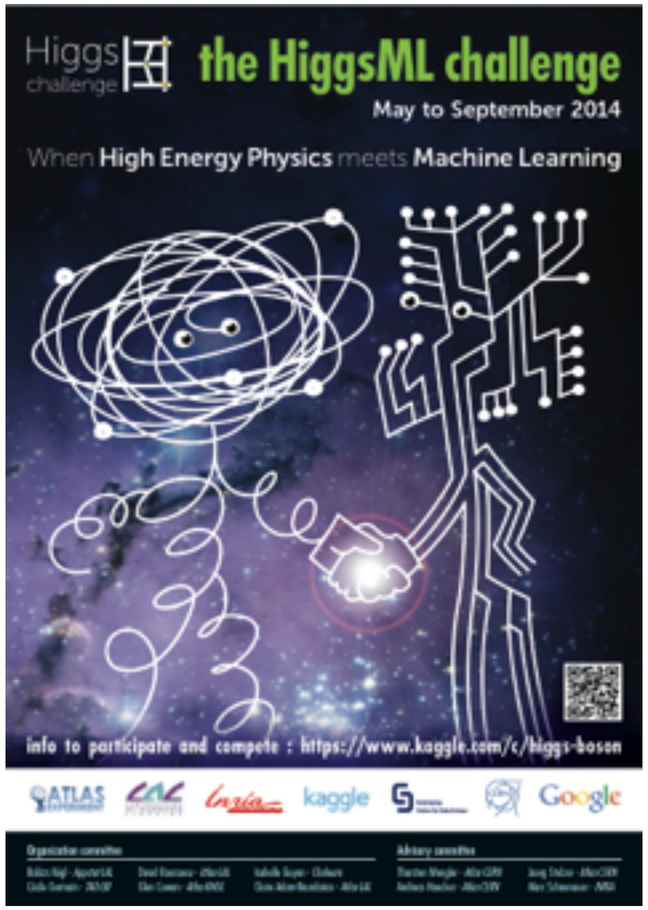
November 20, 2014

Other Stories

Handing In the ATLAS Keys
[December 2014]

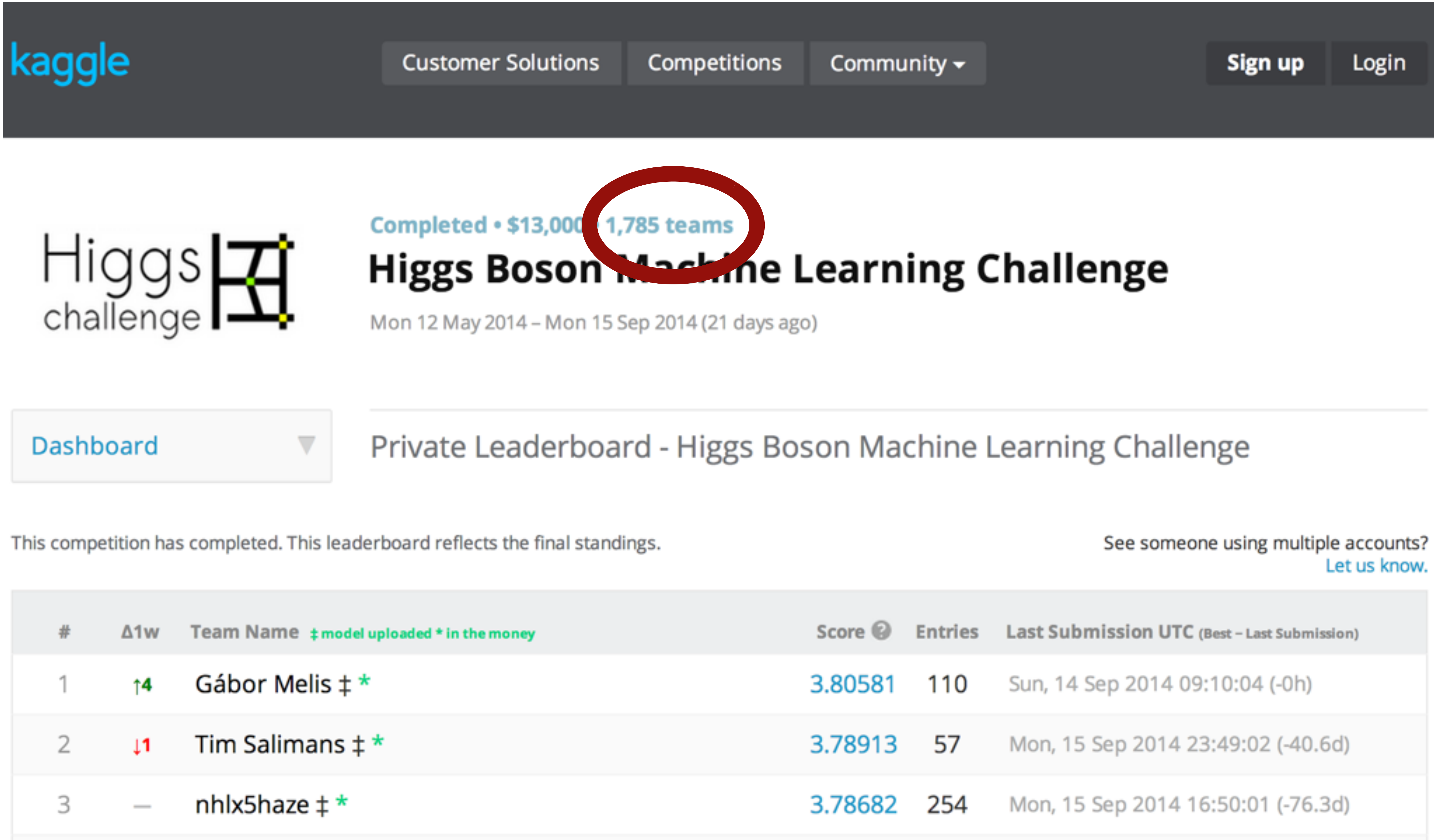
Dirt Detectives
[December 2014]

- 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

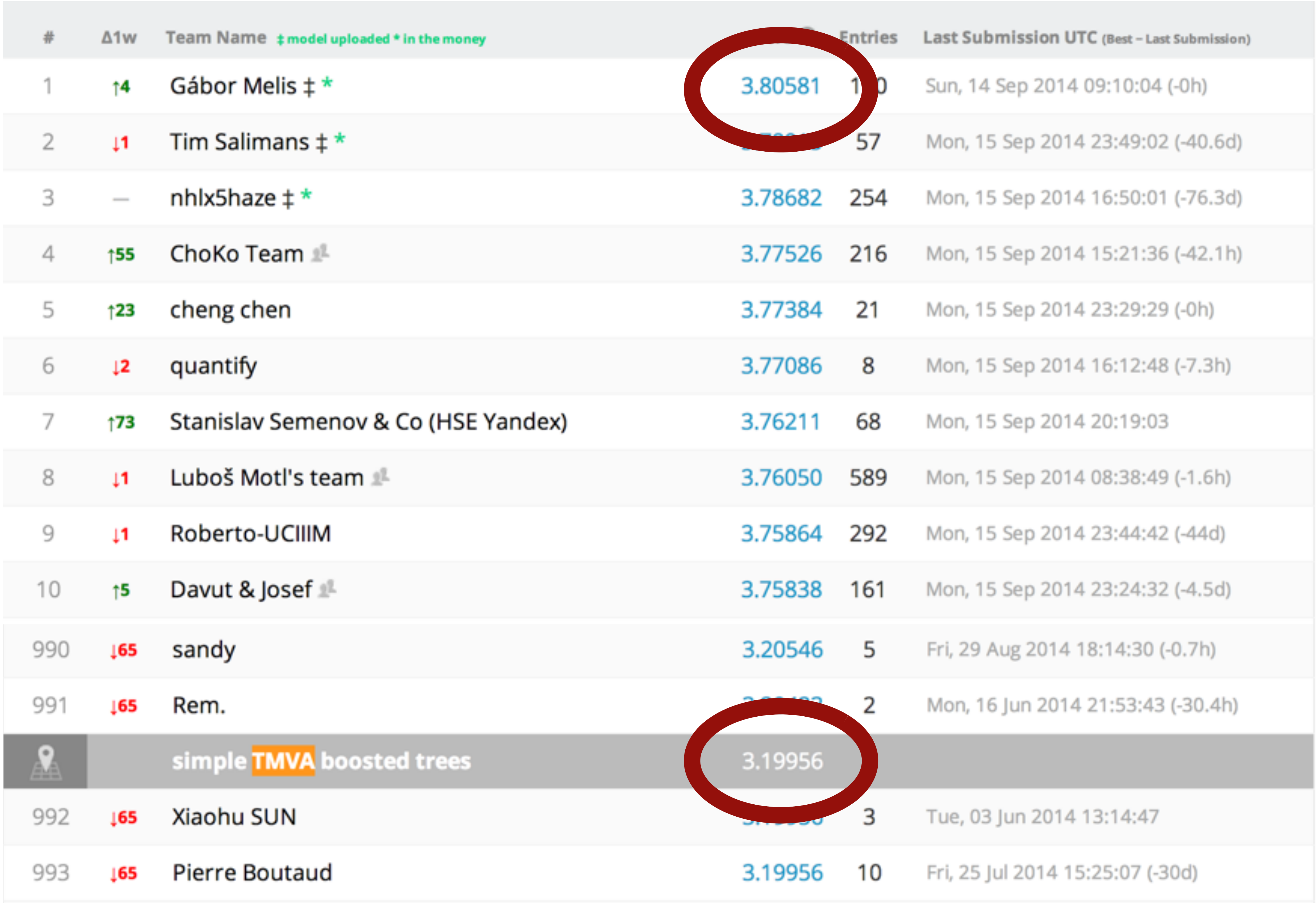
This competition has completed. This leaderboard reflects the final standings.

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) |

Paris-Saclay Center for Data Science

# 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) |

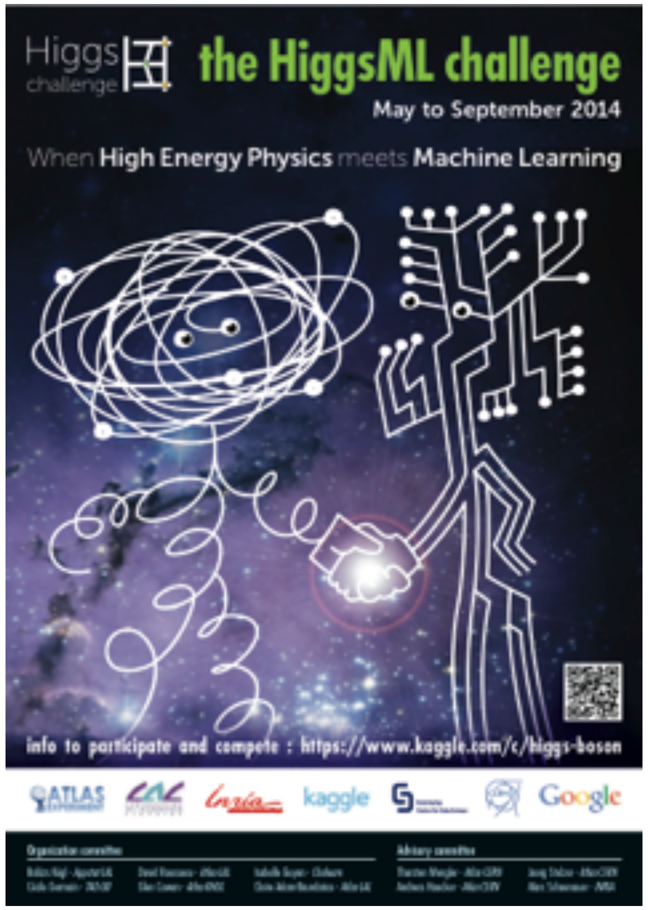
## Huge publicity

## 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


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

We decided to design something better

# RAPID ANALYTICS AND MODEL PROTOTYPING

- Single-day coding sessions
  - 20-30 participants
  - preparation is similar to challenges
- Goals
  - focusing and motivating top talents
  - promoting collaboration, speed, and efficiency
  - solving (prototyping) real problems

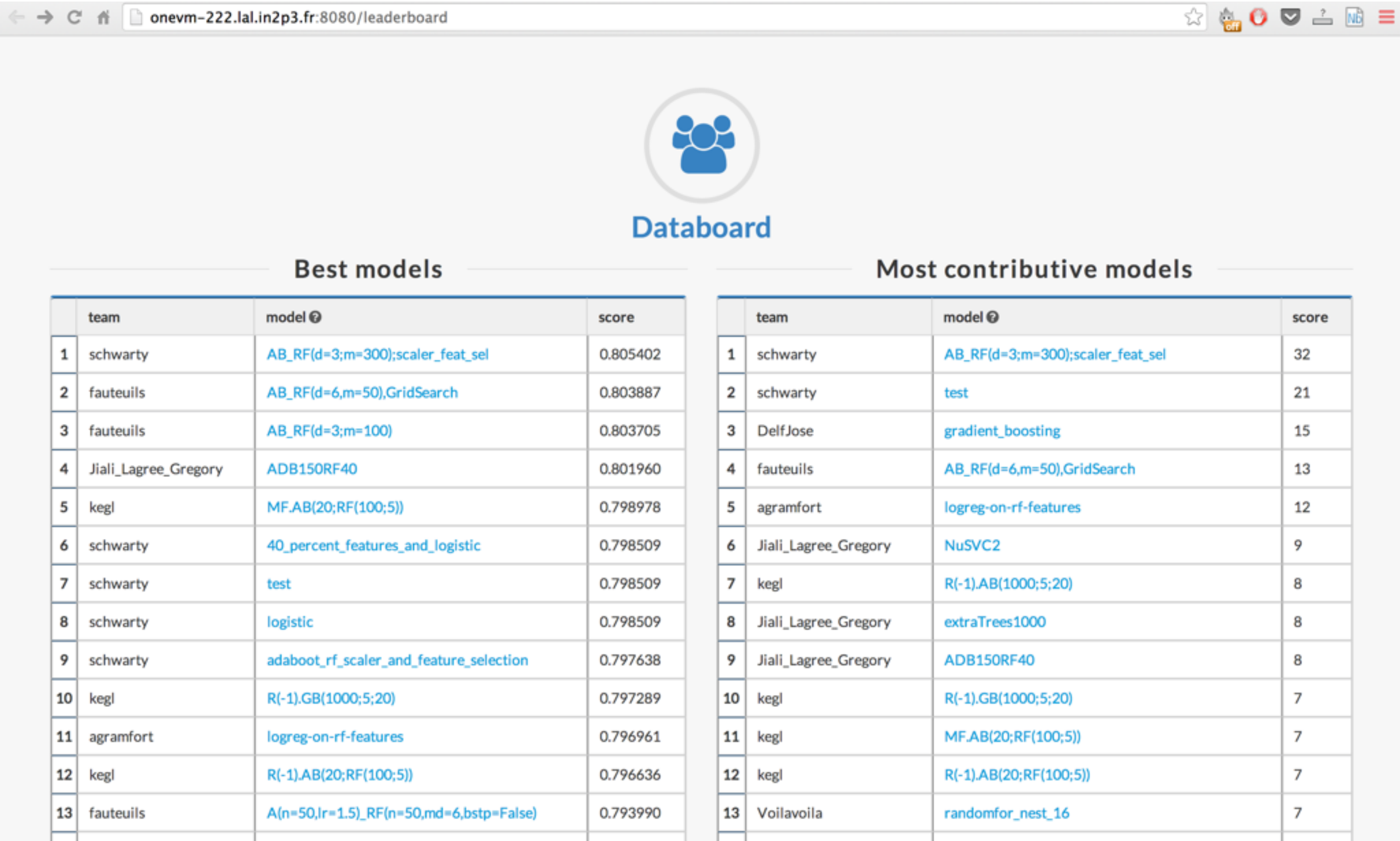
# RAPID ANALYTICS AND MODEL PROTOTYPING

# Rapid analytics and model prototyping

# Rapid analytics and model prototyping

# Rapid analytics and model prototyping

# ANALYTICS TOOLS TO PROMOTE COLLABORATION AND INNOVATION



onevm-222.lal.in2p3.fr:8080/leaderboard

## Databoard

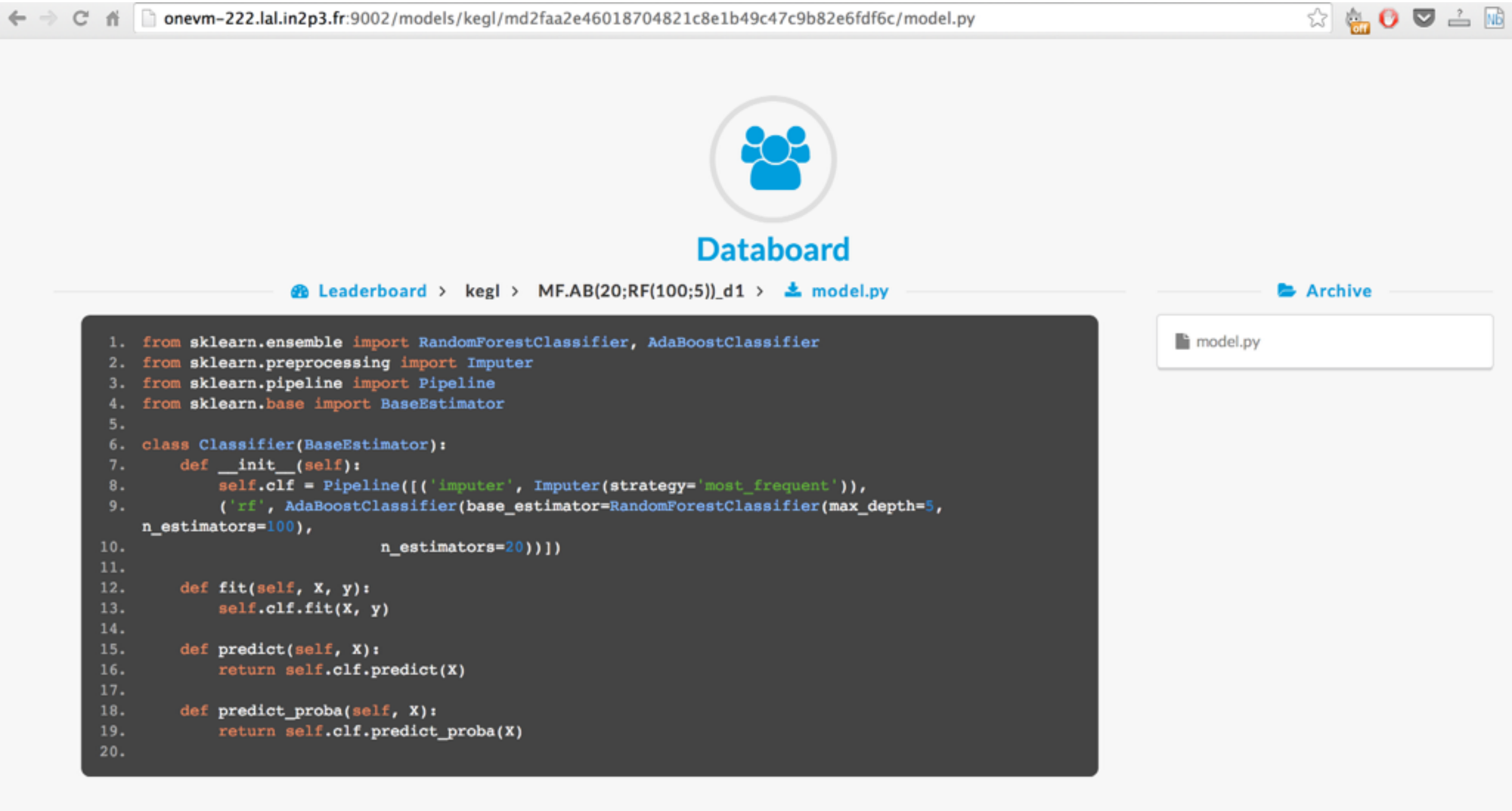
### Best models

| | team | model ? | score |
|---|---|---|---|
| 1 | schwarty | AB_RF(d=3;m=300);scaler_feat_sel | 0.805402 |
| 2 | fauteuils | AB_RF(d=6,m=50),GridSearch | 0.803887 |
| 3 | fauteuils | AB_RF(d=3;m=100) | 0.803705 |
| 4 | Jiali_Lagree_Gregory | ADB150RF40 | 0.801960 |
| 5 | kegl | MF.AB(20;RF(100;5)) | 0.798978 |
| 6 | schwarty | 40_percent_features_and_logistic | 0.798509 |
| 7 | schwarty | test | 0.798509 |
| 8 | schwarty | logistic | 0.798509 |
| 9 | schwarty | adaboot_rf_scaler_and_feature_selection | 0.797638 |
| 10 | kegl | R(-1).GB(1000;5;20) | 0.797289 |
| 11 | agramfort | logreg-on-rf-features | 0.796961 |
| 12 | kegl | R(-1).AB(20;RF(100;5)) | 0.796636 |
| 13 | fauteuils | A(n=50,lr=1.5)_RF(n=50,md=6,bstp=False) | 0.793990 |

### Most contributive models

| | team | model ? | score |
|---|---|---|---|
| 1 | schwarty | AB_RF(d=3;m=300);scaler_feat_sel | 32 |
| 2 | schwarty | test | 21 |
| 3 | DelfJose | gradient_boosting | 15 |
| 4 | fauteuils | AB_RF(d=6,m=50),GridSearch | 13 |
| 5 | agramfort | logreg-on-rf-features | 12 |
| 6 | Jiali_Lagree_Gregory | NuSVC2 | 9 |
| 7 | kegl | R(-1).AB(1000;5;20) | 8 |
| 8 | Jiali_Lagree_Gregory | extraTrees1000 | 8 |
| 9 | Jiali_Lagree_Gregory | ADB150RF40 | 8 |
| 10 | kegl | R(-1).GB(1000;5;20) | 7 |
| 11 | kegl | MF.AB(20;RF(100;5)) | 7 |
| 12 | kegl | R(-1).AB(20;RF(100;5)) | 7 |
| 13 | Voilavoila | randomfor_nest_16 | 7 |

# ANALYTICS TOOLS TO PROMOTE COLLABORATION AND INNOVATION


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

Databoard

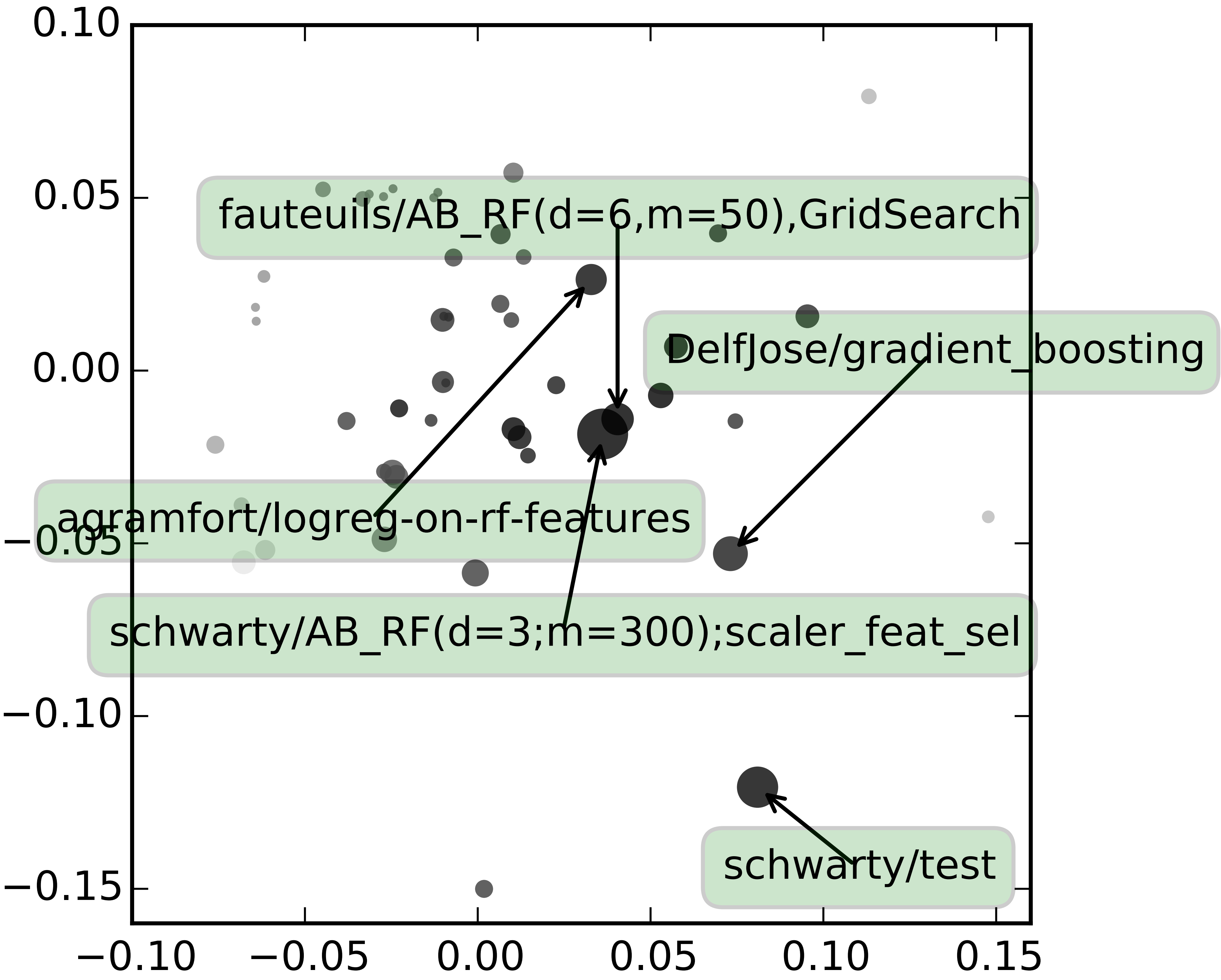
Leaderboard › kegl › MF.AB(20;RF(100;5))_d1 › model.py

Archive

model.py

```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)

```

# ANALYTICS TOOLS TO MONITOR PROGRESS



# RESEARCH (BEYOND SOLVING PROBLEMS)

- Algorithm selection and hyperparameter optimization
  - studying human problem-solving
  - combining human solutions with automatic tools
  - comparing and tuning hyperparameter optimizers
  - meta-learning: embedding data sets and models, collaborative optimization

# Rapid analytics and model prototyping

**2015 Jan 15**

**replaying the**

**HiggsML**

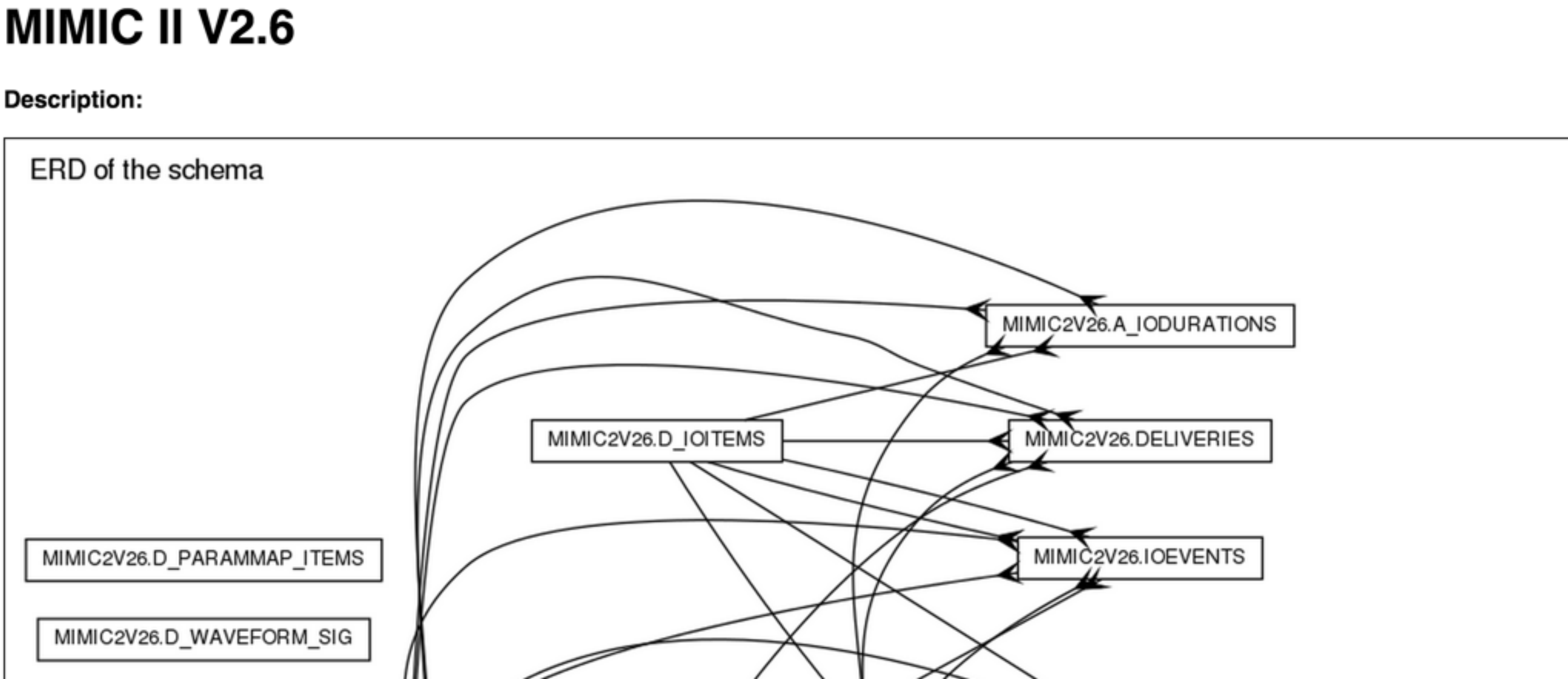
**challenge**

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
Inria
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

# Rapid analytics and model prototyping

## 2015 Feb 9

## Mortality prediction in septic patients



MIMIC II V2.6

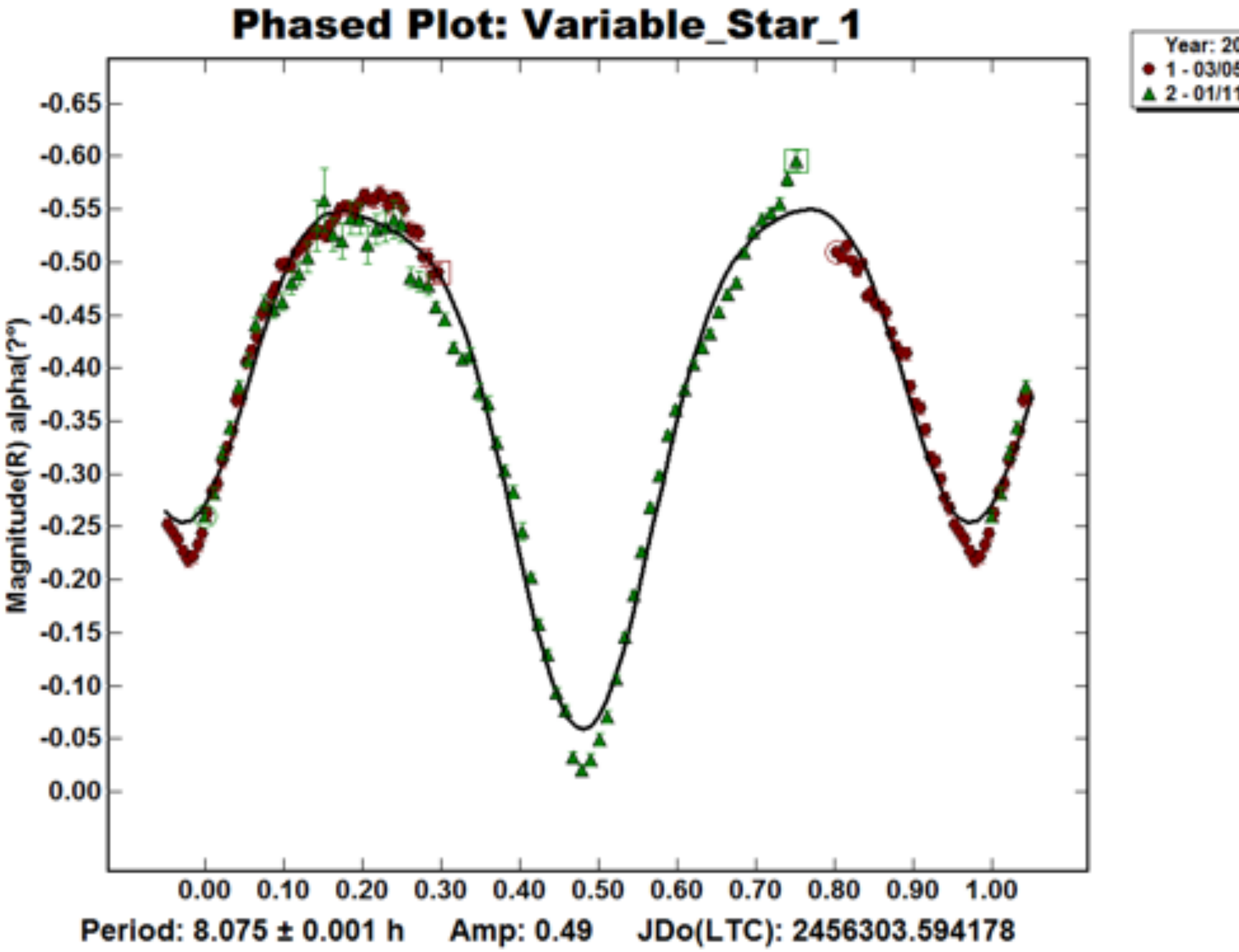
Description:

ERD of the schema

# Rapid analytics and model prototyping

## 2015 Apr 10

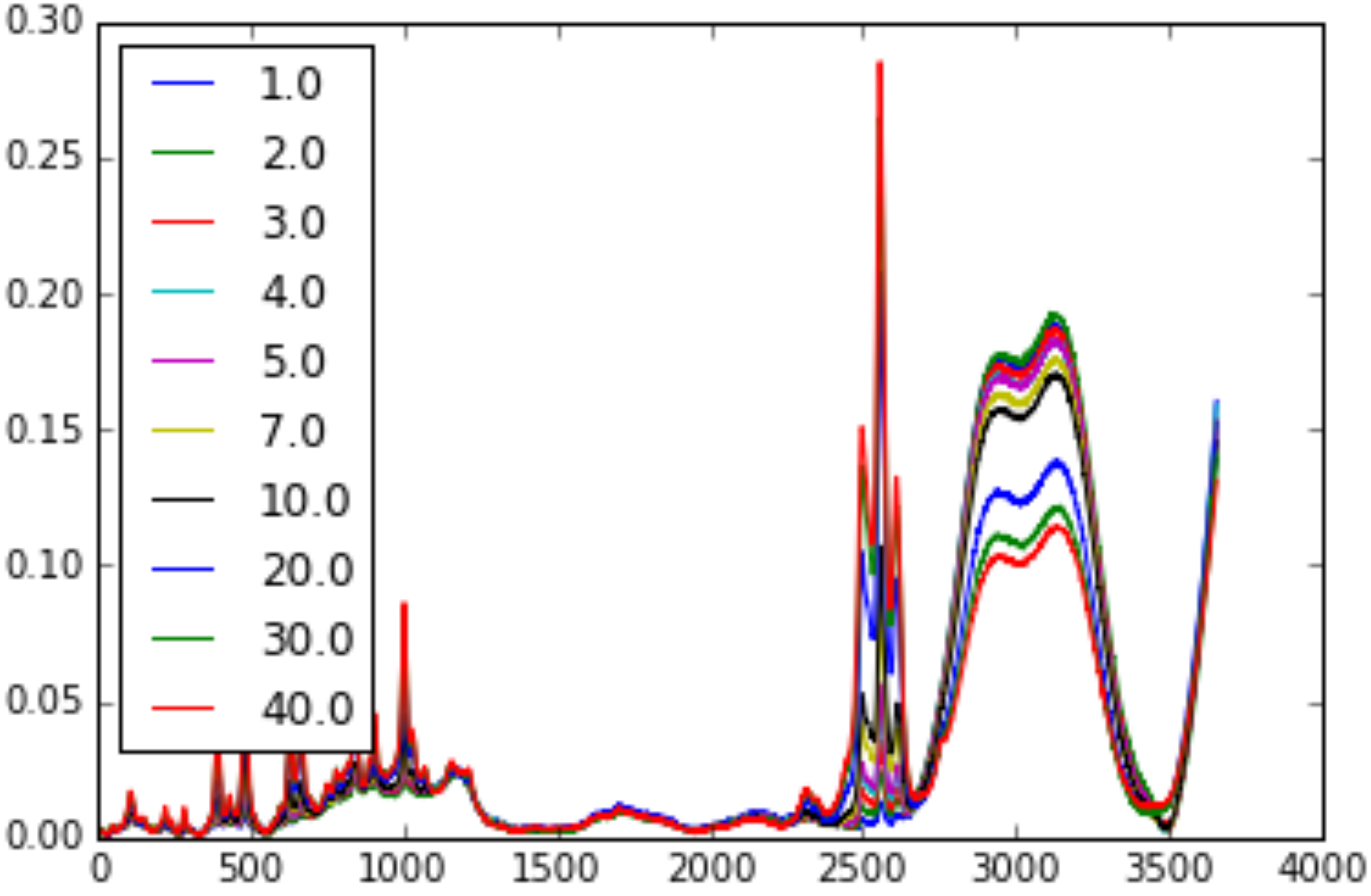
## Classifying variable stars



# Rapid analytics and model prototyping

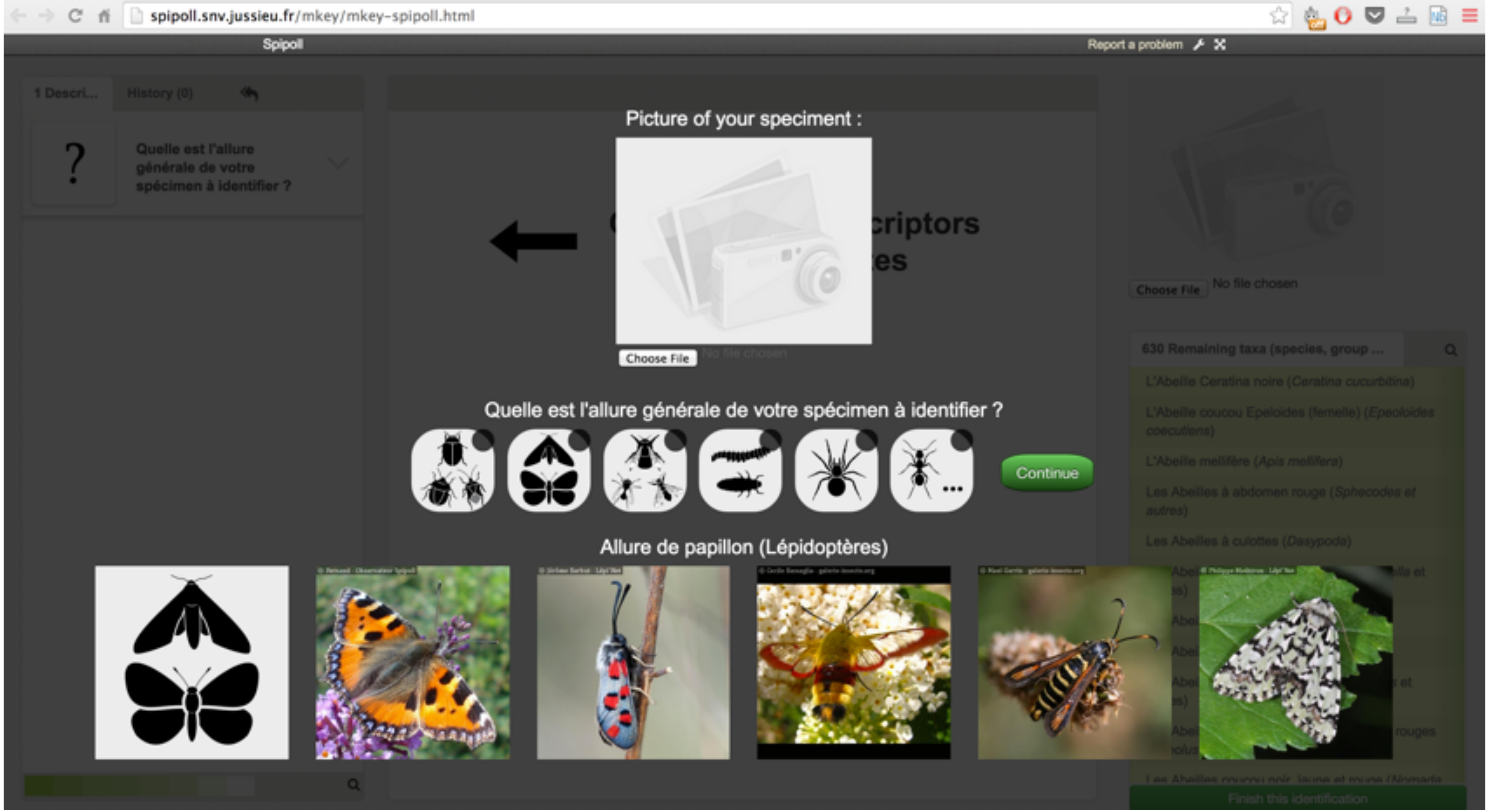
## 2015 May

## Drug identification from spectra



# 2015 June

# Insect classification



spipoll.snv.jussieu.fr/mkey/mkey-spipoll.html

Spipoll

Report a problem

1 Descri... History (0)

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

Picture of your speciment :

criptors
es

Choose File No file chosen

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

Continue

Allure de papillon (Lépidoptères)

Choose File No file chosen

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*)

Les Abeilles coucou noir, jaune et rouge (*Nomada*

Finish this identification

# IMPLEMENTING RAMPS IN AN INDUSTRIAL CONTEXT

- Short-term, **ad-hoc teams** assembled for a given task
- **Low-engagement** consulting job
- **Efficient** use of scarce data scientists (i.e., your time)
- Developing and practicing **marketable skills**
  - fast-feedback experimentation is **also useful in research**
- Networking
- Management meta-tools to **track** your performance, to guide your **training**

# THANK YOU!