OPEN SOFTWARE INITIATIVE (OSI)

AT

Paris-Saclay Center for Data Science

université PARIS-SACLAY

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

@SaclayCDS

OCT. 26 2015

université PARIS-SACLAY
Paris-Saclay Center for Data Science

DATA SCIENCE **IN THE WORLD**

NYU CENTER FOR DATA SCIENCE

UNIVERSITY of WASHINGTON

UNIVERSITY of WASHINGTON

UC BERKELEY INSTITUTE FOR DATA SCIENCE

ADVANCING DATA-INTENSIVE DISCOVERY IN ALL FIELDS

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

AgroParisTech

cea

cnrs

CentraleSupélec

ens CACHAN

ENSTA ParisTech

École Polytechnique

ENSAE ParisTech

HEC PARIS

IHÉS

INRA SCIENCE & IMPACT

Inria

Institut d'Optique Graduate School ParisTech

Inserm

Institut Mines-Télécom

ONERA THE FRENCH AEROSPACE LAB

Université Paris Sud

Université de Versailles St-Quentin-en-Yvelines

université PARIS-SACLAY

université PARIS-SACLAY

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

université PARIS-SACLAY

Paris-Saclay Center for Data Science

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

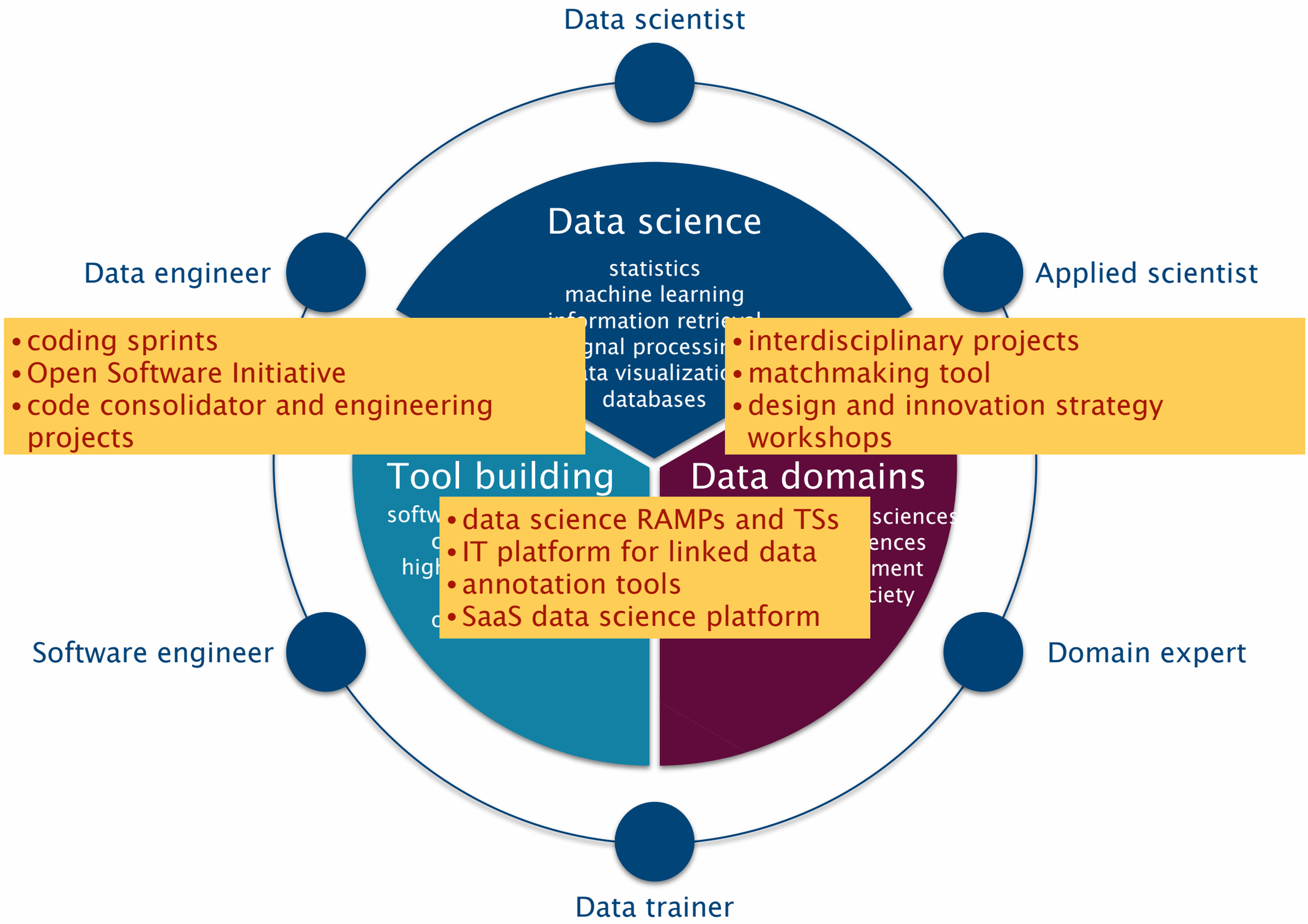
Paris-Saclay Center for Data Science

WHAT IS THE CHALLENGE?

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

NYU-CDS manifesto

TOOLS: LANDSCAPE TO ECOSYSTEM



https://medium.com/@balazskegl

universite PARIS-SACLAY

Paris-Saclay Center for Data Science

Open Software Initiative (OSI)

Coding Sprints

PHD STUDENTS PROJECTS

Objective: Strengthen students skills in software engineering by contributing to scientific open source software

ENGINEERING PROJECTS

Objective: Consolidate academic software, foster good engineering practice between labs, improve the tools for research

Data Challenges

DATA CHALLENGES

kaggle

Host | Competitions | Scripts | Jobs | Community | agramfort | Logout

Welcome to Kaggle's data science competitions.

New to Data Science?
Tutorials on the Titanic competition »

Want to learn from other's code?
Kaggler's top rated scripts »

Download

Choose a competition & download the training data.

Build

Build a model using whatever methods and tools you prefer.

Submit

Upload your predictions. Kaggle scores your solution and shows your score on the leaderboard.

Active Competitions

All Competitions

Active Competitions

Springleaf Marketing Response
Determine whether to send a direct mail piece to a customer
35 days
1365 teams
734 scripts
$100,000

Western Australia Rental Prices
Predict rental prices for properties across Western Australia
2 months
28 teams
$100,000

Coupon Purchase Prediction
16 days
878 teams

Beyond Data Challenges

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

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**

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

Databoard

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

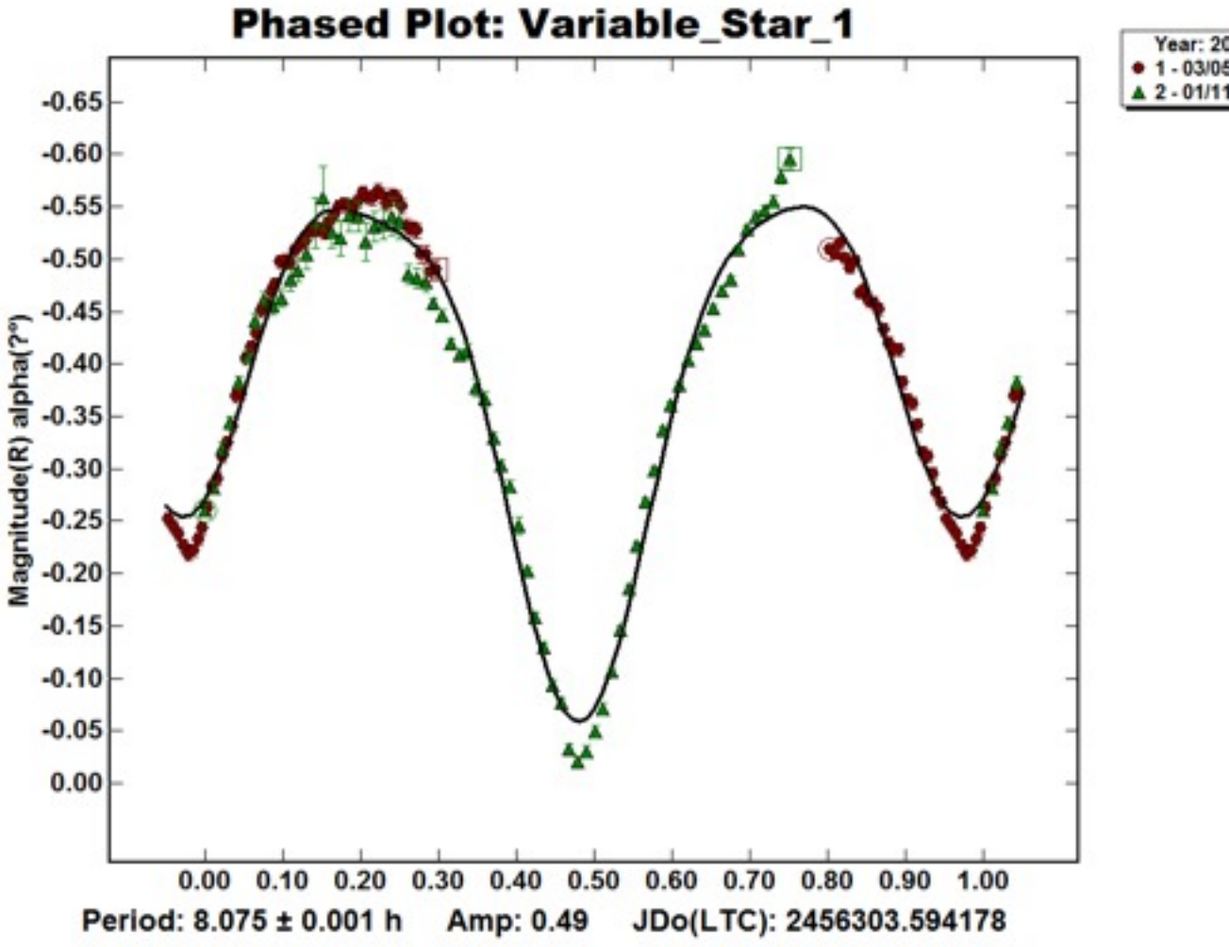
Archive

```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)
```

model.py

2015 Apr 10

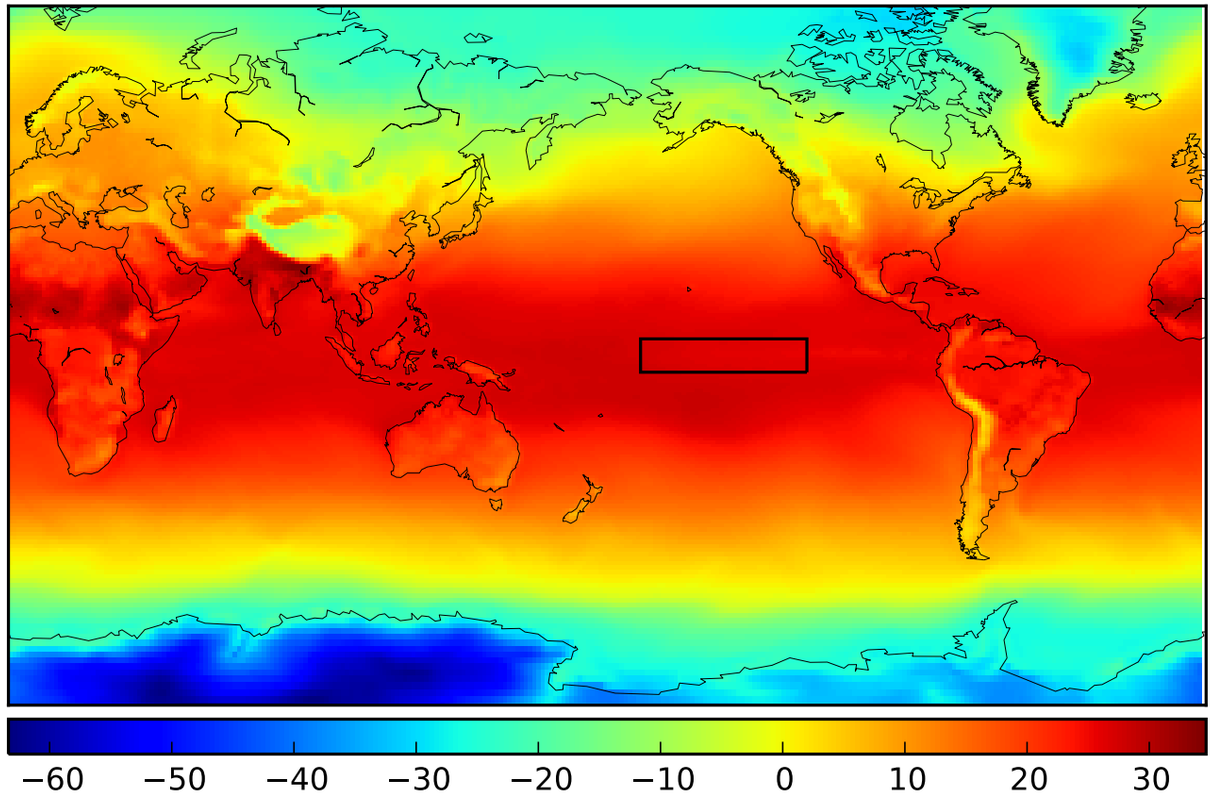
Classifying variable stars

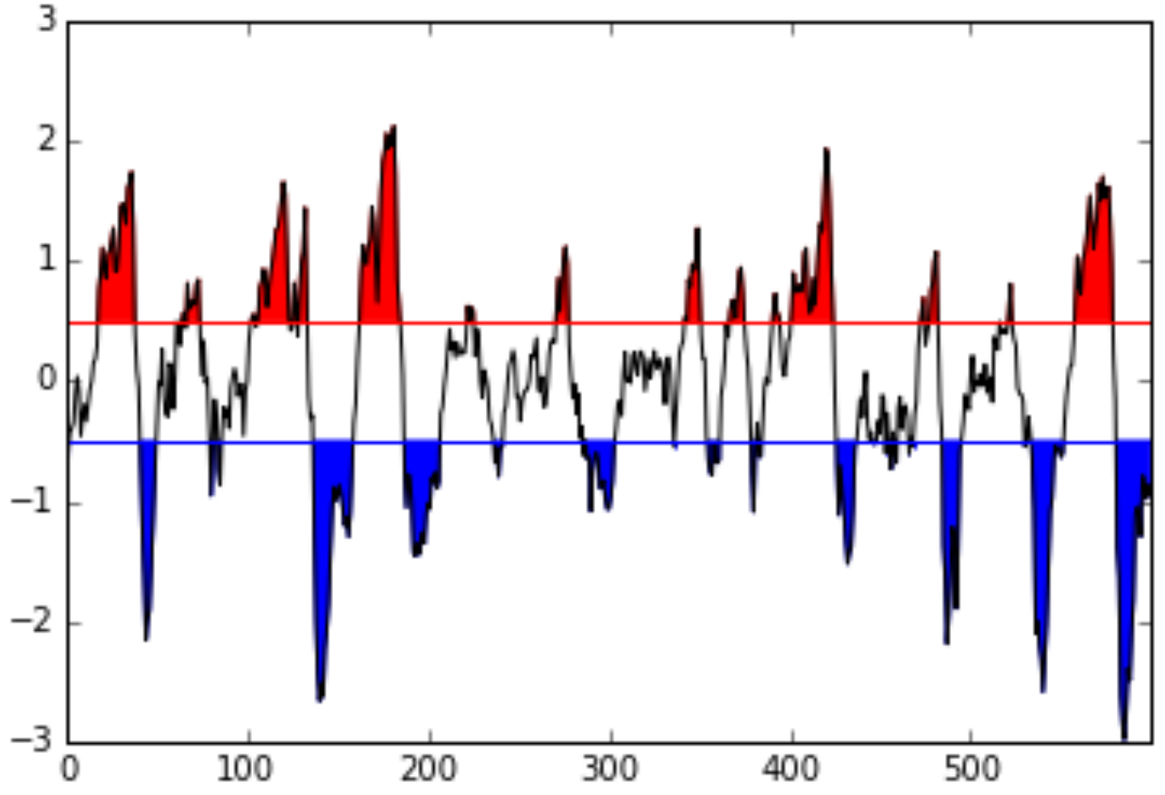


2015 June 16 and Sept 26

Predicting El Nino

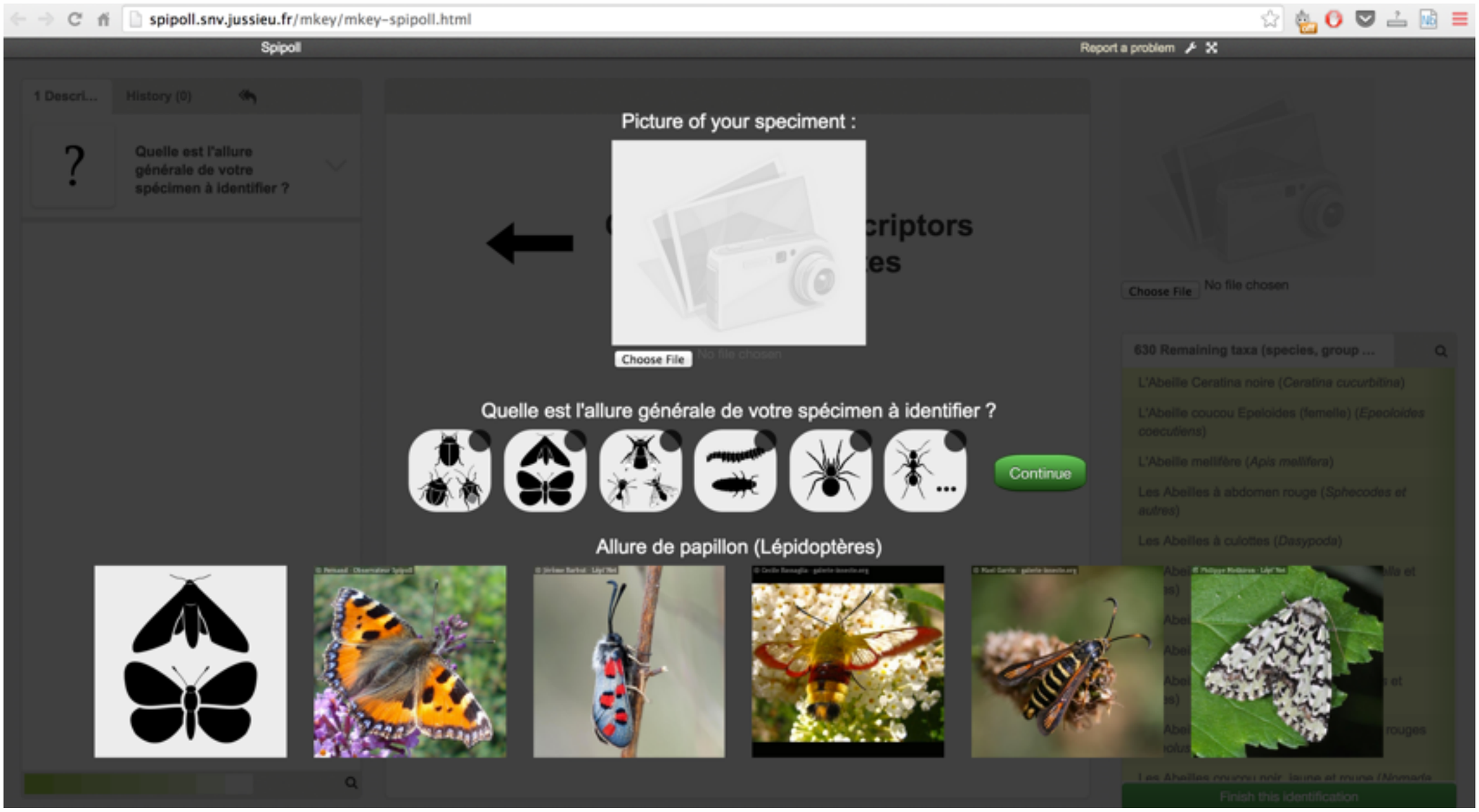
Temperature map





2015 October 8

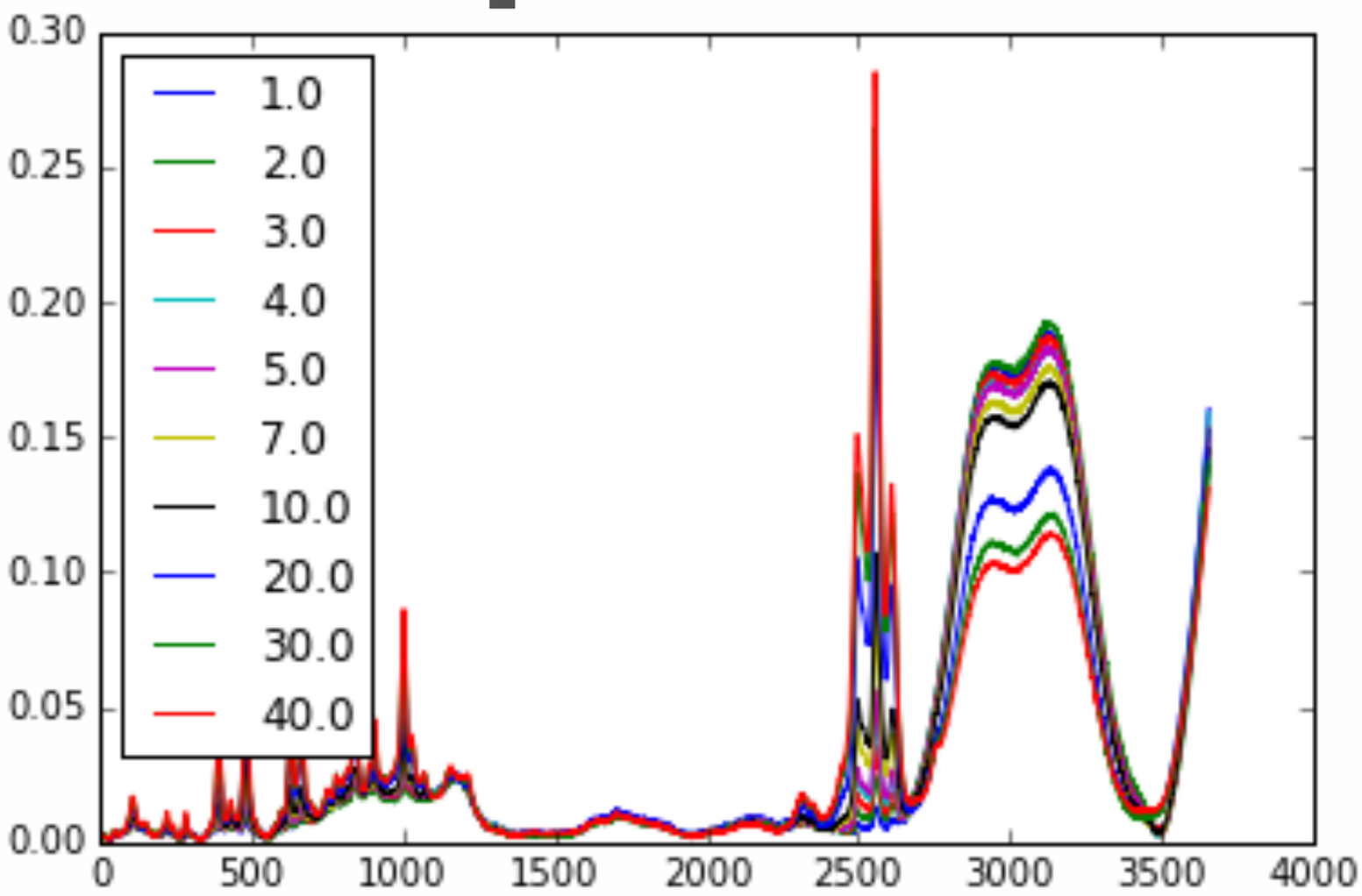
Insect classification



RAPID ANALYTICS AND MODEL PROTOTYPING

2015 Fall

Drug identification from spectra



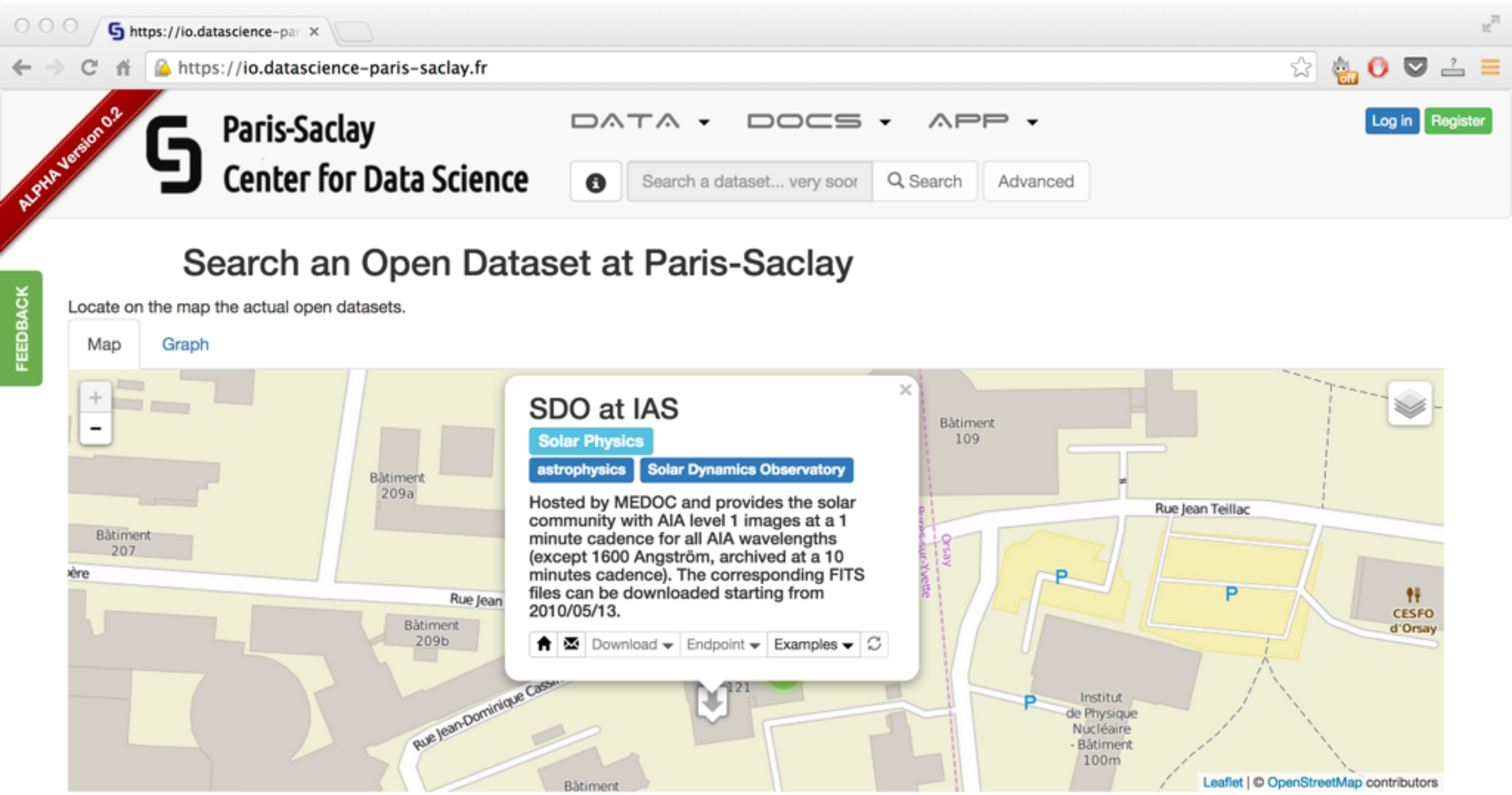
Giving access to data

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



Today's keynotes

Fernando Perez
Univ. California Berkeley

Andreas Mueller
New York University

TODAY'S PROGRAM

9h00 - 9h15 Alexandre Gramfort, Welcome talk
9h15 - 10h15 Fernando Perez, UC Berkeley
10h15 - 10h30 Oscar Najera, *Sphinx-Gallery: Package documentation made easy*
10h30 - 11h00 Coffee Break
11h00 - 11h15 Mehdi Cherti: *Py-Earth: Multivariate Adaptive Regression Splines in Python*
11h15 - 11h30 Lorenzo Desantis, *MNE from shell scripts and Unix commands to Python*
11H30 - 12h00 Sarah Cohen Boulakia, *Reproducible research in Bioinformatics*
12h00 - 12h15 Karin Dassas, *LoOPS Network for developers at Paris-Saclay*
12h15 - 12h30 Cécile Germain, The io.datascience DaaS platform

12h30 - 14h00 Lunch break at Proto 204

14h00 - 14h20 Miguel Colom, *A service-oriented platform for online physiological data analysis*
14h20 - 14h40 Diem BUI THI, *Automation for chemical data analysis techniques*
14h40 - 15h00 Estelle Chaix, *Information Extraction Challenge for Gene Regulation Network in plant*
15h00 - 15h30 Coffee Break
15h30 - 15h50 Loic Estève, *Mining brain imaging data: lessons learned from nilearn and joblib*
15h50 - 16h10 J. Duperrier & D. Guarino, *A customizable framework for neurophysiology data management and provenance tracking*
16h10 - 16h25 Nicolas Goix, *Anomaly detection algorithms in Scikit-Learn*
16h25 - 16h40 Romain Brault, *Structured Prediction with Opera-Lib*
16h40 - 17h30 Andreas Mueller, *Software and engineering efforts at NYU Center for Datascience*