A data science observatory

Akin Kazakci, Mines ParisTech
Balazs Kégl, CNRS

Team

Balázs Kégl
CNRS

Alexandre Gramfort
Télécom ParisTech

Akın Kazakçı
Mines ParisTech

Djalel Benbouzid
UPMC

Camille Marini
Télécom ParisTech

Mehdi Cherti
UP Saclay

Yohann Sitruk
Mines ParisTech

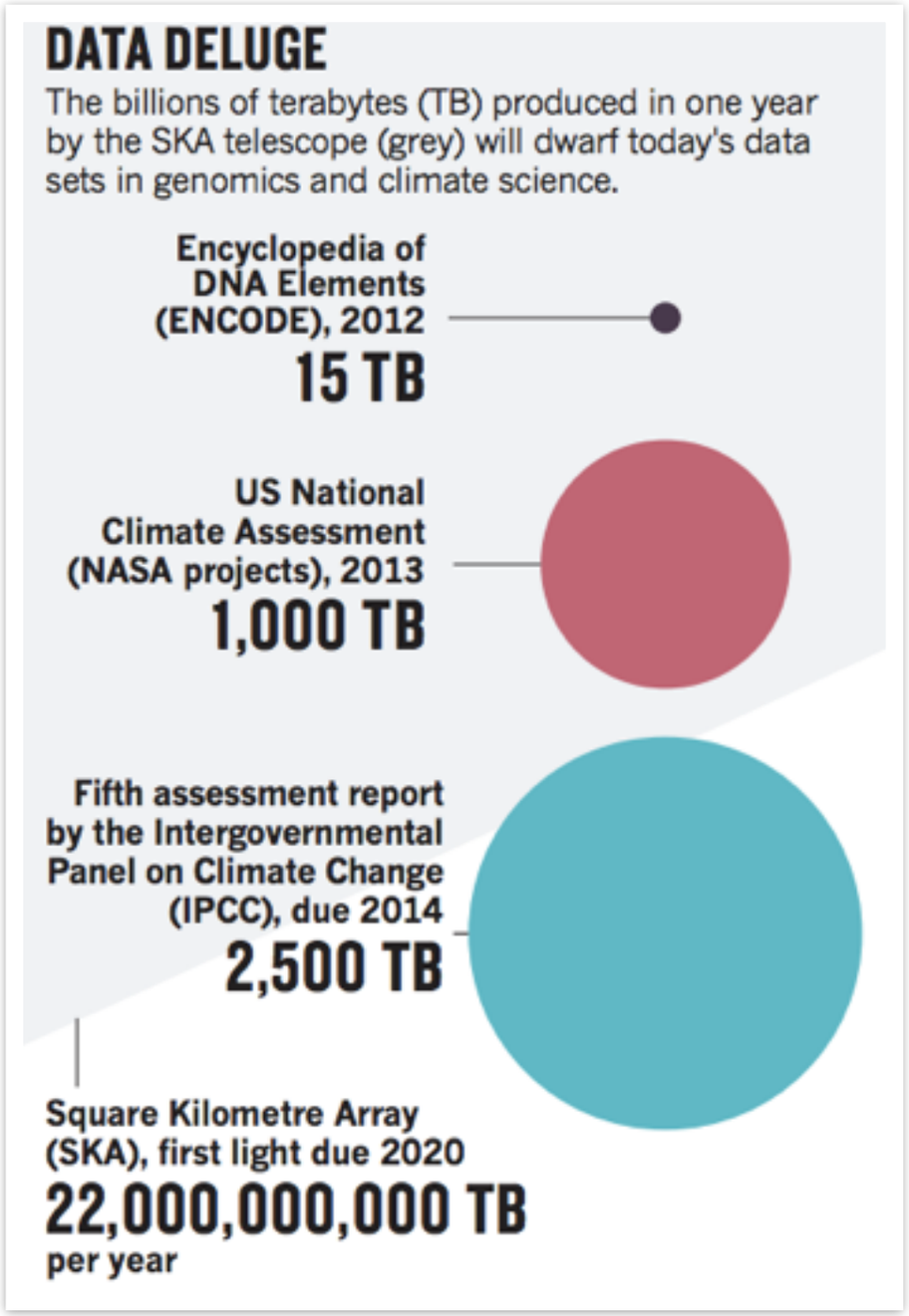
The research objective & questions

Enough with the chairs

- Design research is falling behind in dealing with **contemporary challenges**
  - Claim 1a: Too much in-breeding and repetition
  - Claim 1b: Huge amount of work is based on ideas from 80s'

- Design is **not about objects, but about reasoning**

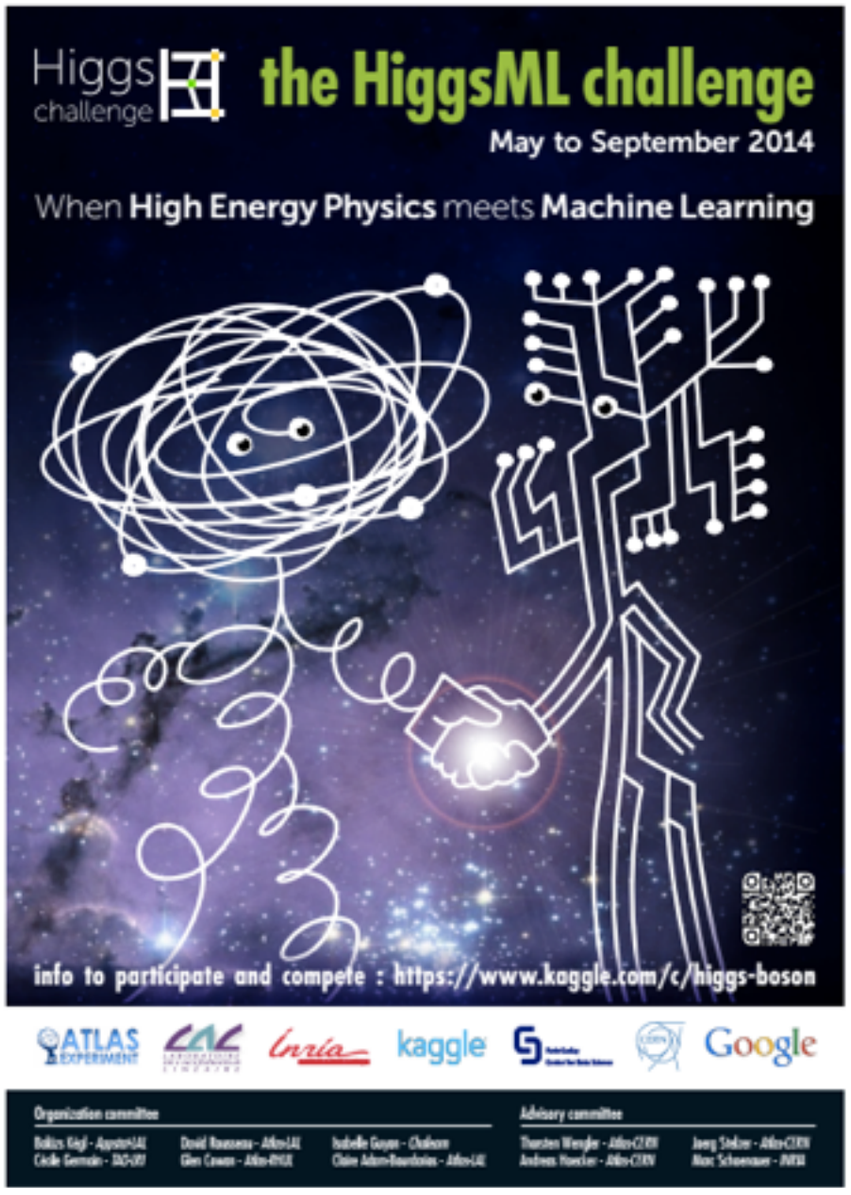
Revealing the potential of data: what role for design?

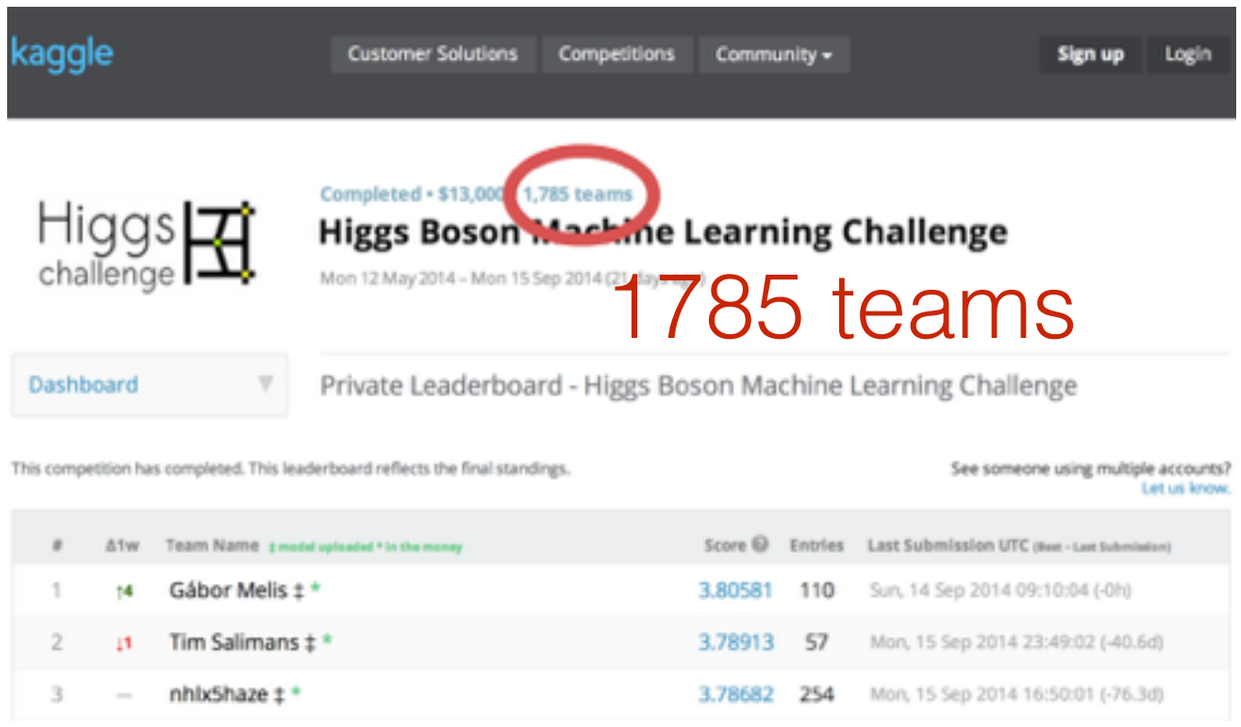


- Physics (Particle physics, Plasma physics, astrophysics…)
- Biology (Genetics, Epidemiology…)
- Chemistry
- Economics, Finance, Banking
- Manufacturing, Industrial Internet
- Internet of things, Connected Devices
- Social media
- Transport & Mobility
- …

There is not enough data scientists to handle this much data

Last year: Crowdsourcing data challenges





1785 teams

Reasonable doubts about the effectiveness of data science contests

Kazakci, A., Data science as a new frontier for design, ICED'15, Milan

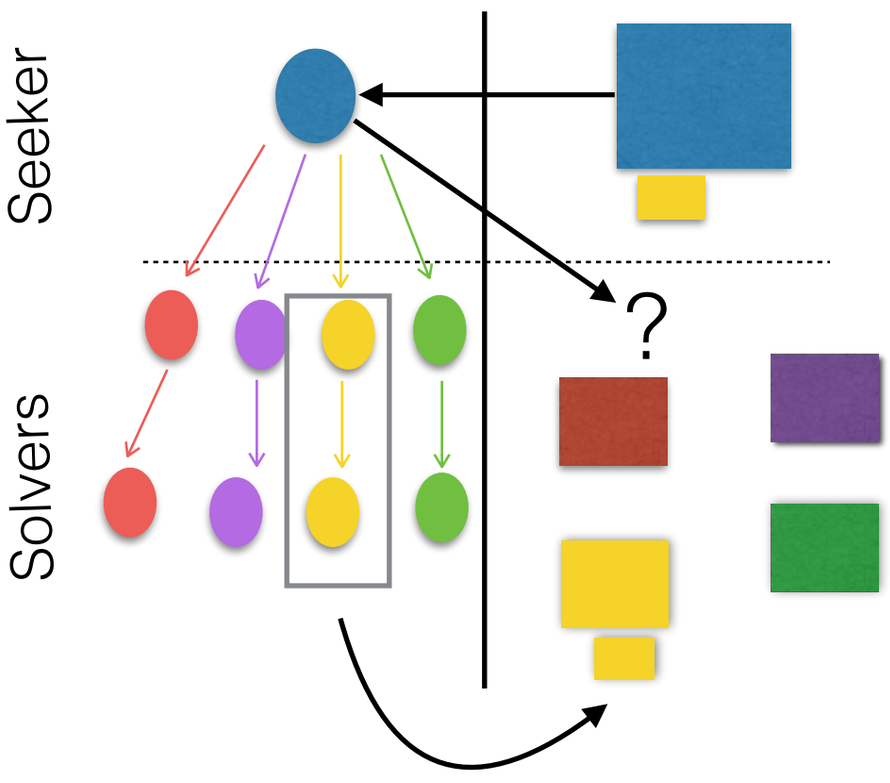
Crowdsourcing /?/ Design

crowd
/kraʊd/ noun
1.a large number of people gathered together in a *disorganized or unruly way*

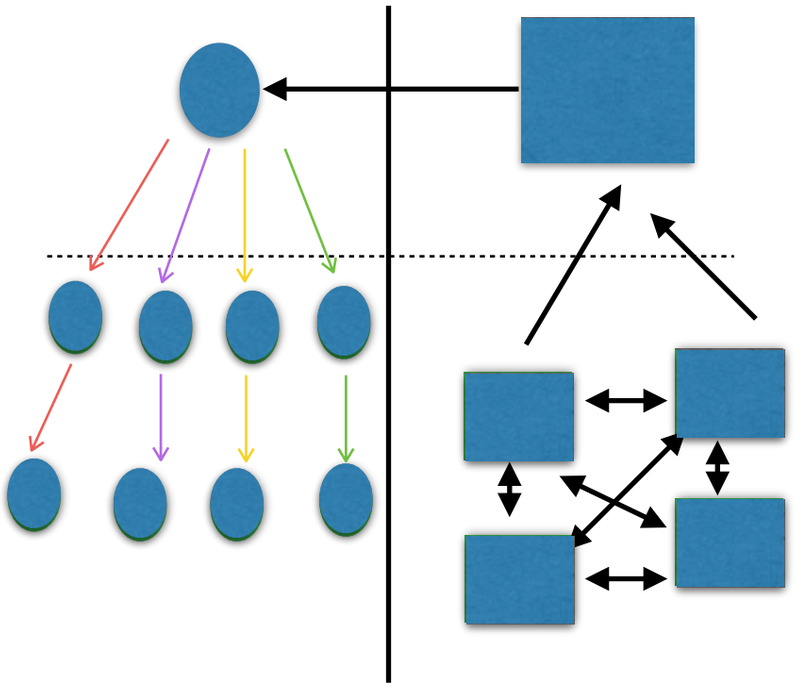
1. How to study the design process of a crowd?
2. How to manage the design process of a crowd?

Crowdsourcing: C-K dynamics

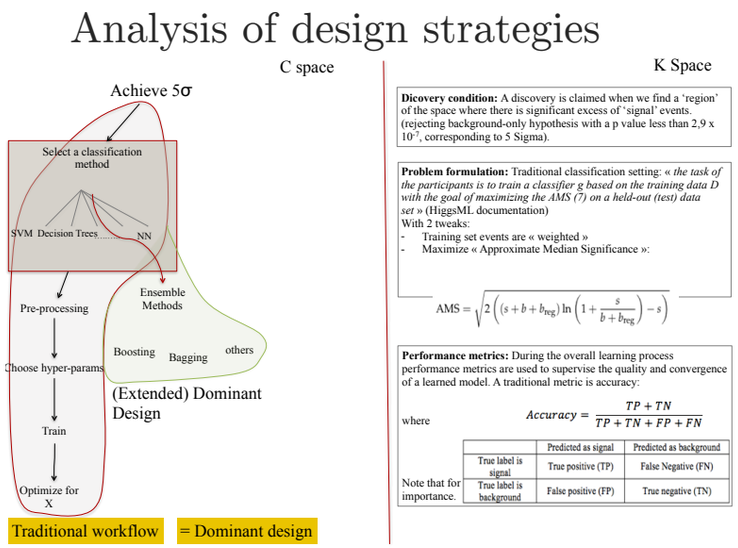
Crowdsourced contests

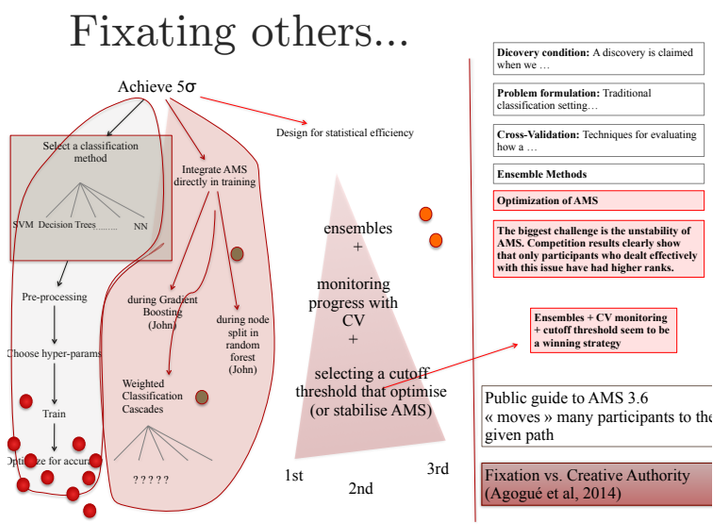


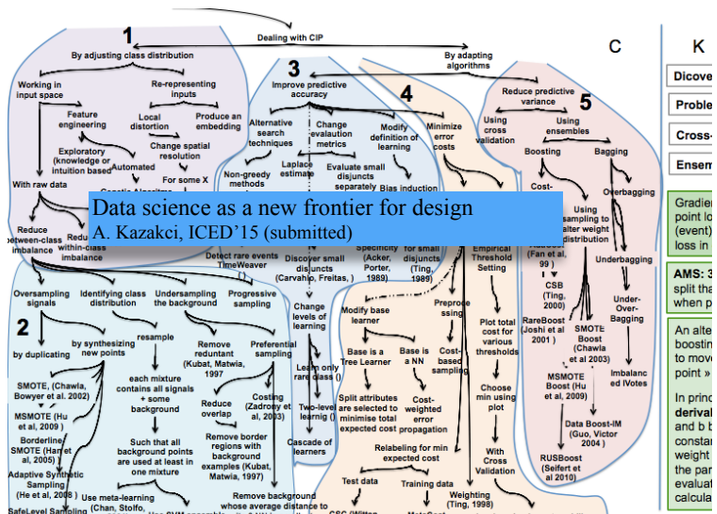
Crowdsourced collaboration



Data challenges are hard to analyze







- Available data for HiggsML
  - Forums ➔ 136 topics, 1400+ posts
  - Documentation
  - Participants' blog entries
  - GitHub codes
- Qualitative interpretation combined with C-K modelling of participants' strategies

How do you put a crowd under a microscope?

2 The research instrument

RAMP - Rapid Analytics and Model Prototyping

A Collaborative Development Platform for Data Science

RAMP
Rapid Analytics and Model Prototyping

Number of air passengers prediction

my submissions
new submission
leaderboard
log out

Instant access to all submitted code - for participants & organizers

Combined score: 0.227

Combined test score: 0.225

Private leaderboard

team	submission	public bag
mouhcine_taoufiq	First submition	0.244
robin_vogel	sparse and spurious	0.237
baptiste_roziere	INOPA	0.263
nicolas_legroux	second_try	0.260
youcef_sebiat	last_sub	0.242
youcef_sebiat	HMM_2	0.241
pierre_fournier	firstTryOKLM	0.257
youcef_sebiat	HMM_3	0.241

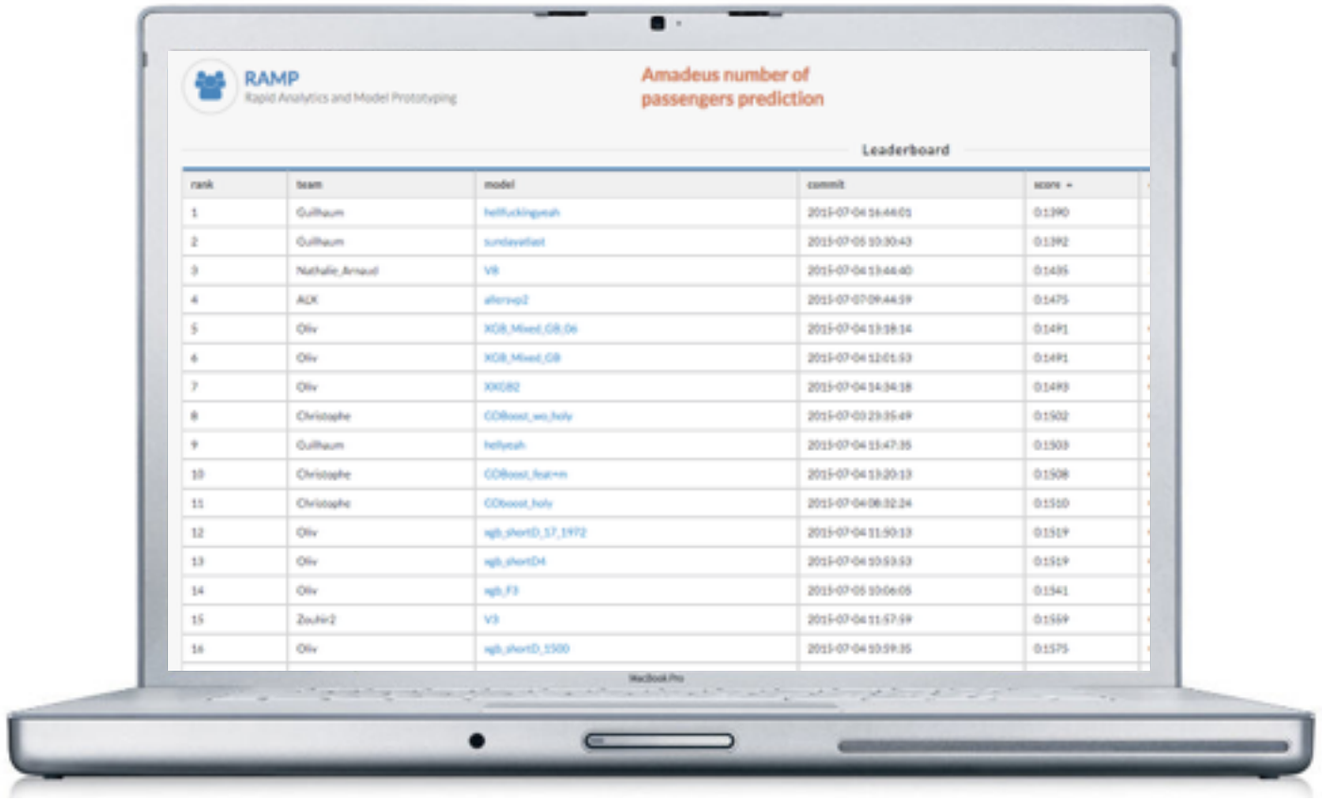
Databoard

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)
```

RAMP - Rapid Analytics and Model Prototyping

A Collaborative Development Platform for Data Science



RAMP allows us to collect data on the data science model development process

1. We prepare a ‘starting kit’
2. Continuous access to code:
   - Participants can analyse and build on every submission
   - Organizers can follow real-time what’s happening - and react
3. Submissions are trained and performances are displayed
4. Users actions and interactions are recorded
5. Main Output: Dozens of predictive models and performance benchmark

Collecting data with RAMP

RAMP
Rapid Analytics and Model Prototyping

Number of air passengers prediction

my submissions
new submission
leaderboard
log out

timestamp (UTC)	user	team	interaction	submission	file	code similarity	diff
2016-01-31 16:26:06 Sun	akazakci	akazakci	looking at user				
2016-01-31 16:24:56 Sun	kegl	kegl	looking at leaderboard				
2016-01-31 16:24:09 Sun	akazakci	akazakci	looking at user				
2016-01-31 16:23:47 Sun	akazakci	akazakci	looking at user				
2016-01-29 18:00:39 Fri	kegl	kegl	looking at submission	mouhcine_taoufiq/First submition	feature_extractor		
2016-01-29 15:27:56 Fri	eliot_barril	eliot_barril	looking at submission	mouhcine_taoufiq/First submition	regressor		
2016-01-29 15:27:38 Fri	eliot_barril	eliot_barril	looking at leaderboard				
2016-01-29 15:27:38 Fri	eliot_barril	eliot_barril	looking at leaderboard				
2016-01-29 14:05:12 Fri	eliot_barril	eliot_barril	looking at submission	eliot_barril/aftercompetition2	regressor		
2016-01-29 14:05:00 Fri	eliot_barril	eliot_barril	looking at submission	eliot_barril/GlouGlou_2	regressor		
2016-01-29 1							
2016-01-29 1				mou			
2016-01-29 1				mou			
2016-01-29 1							
2016-01-29 1							
2016-01-29 0							
2016-01-28 1			board				
2016-01-28 1							
2016-01-28 0				ivay			
2016-01-28 0				ivay			
2016-01-28 0				tian			
2016-01-28 09:14:31 Thu	eliot_barril	eliot_barril	looking at leaderboard				
2016-01-27 14:26:37 Wed	basile_calderan	basile_calderan	looking at submission	robin_vogel/sparse and spurious	feature_extractor		
2016-01-27 14:26:24 Wed	basile_calderan	basile_calderan	looking at private leaderboard				

- Number of submissions
- Frequency of submissions
- Timing of submissions
- User interactions
- Performance of submissions
- Submitted code
- …

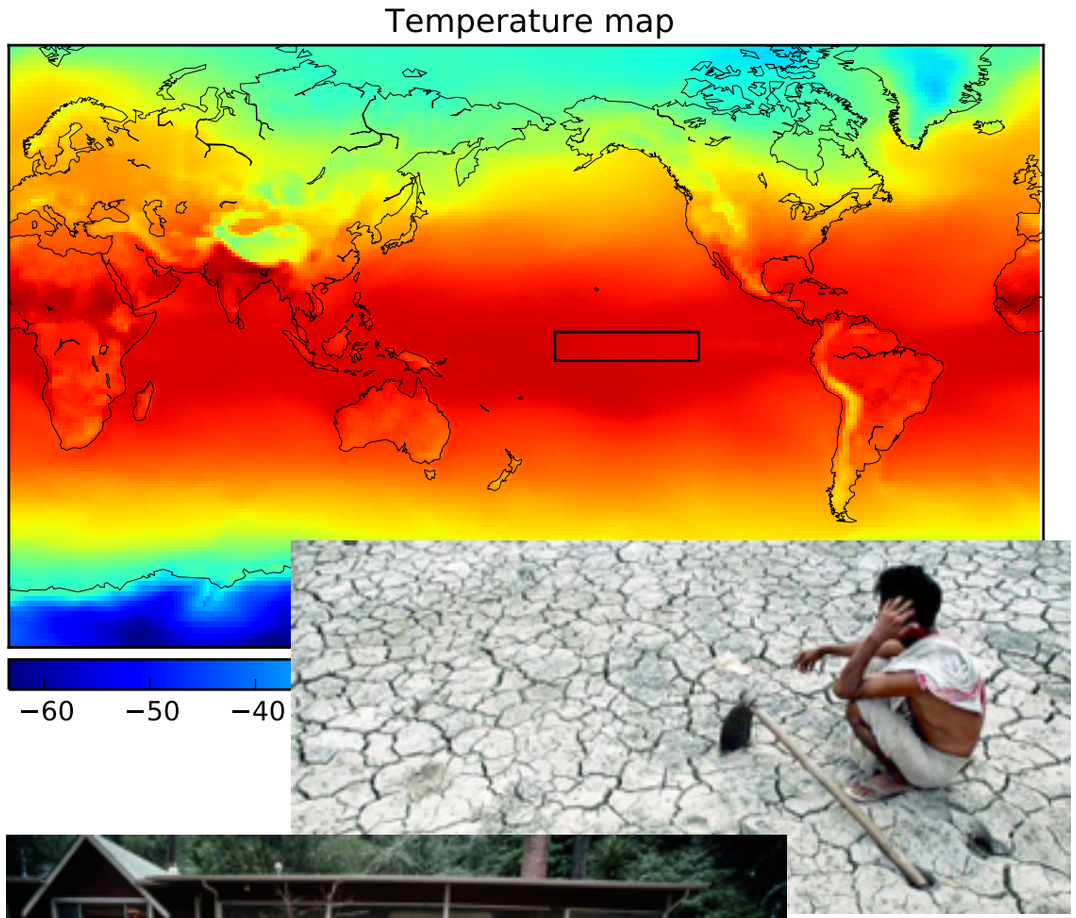
We are interested in

- the variety (code space + prediction space)
- the mutual influences and inheritance (code space)
- score and delta score (impact)
- …

3

Some applications, preliminary observations & findings

Climatology



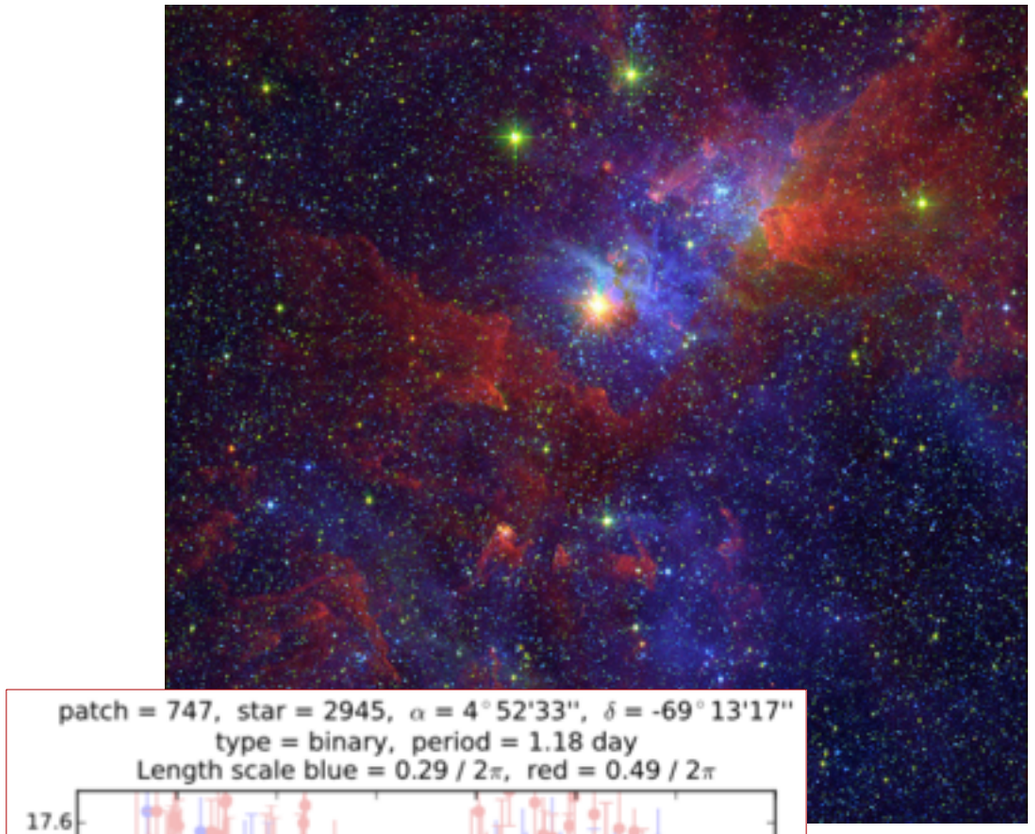
Time Series Based Event Prediction on Geo-tagged data

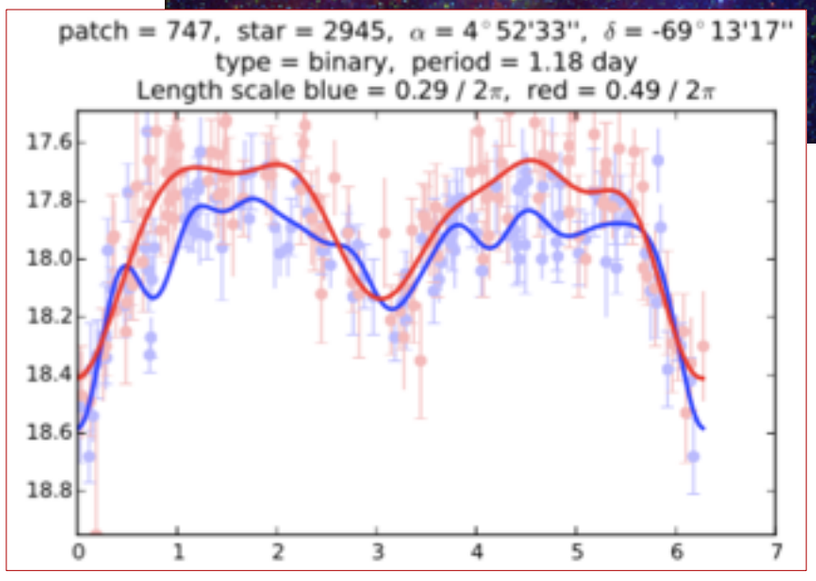
George Washington University
George Mason University

EI Nino Prediction
- Temperature data
- Geo-tagged time series
- Prediction: 6 months ahead

Two workshops: Improvement in RMSE score: from 0.90 to 0.43

Astrophysics





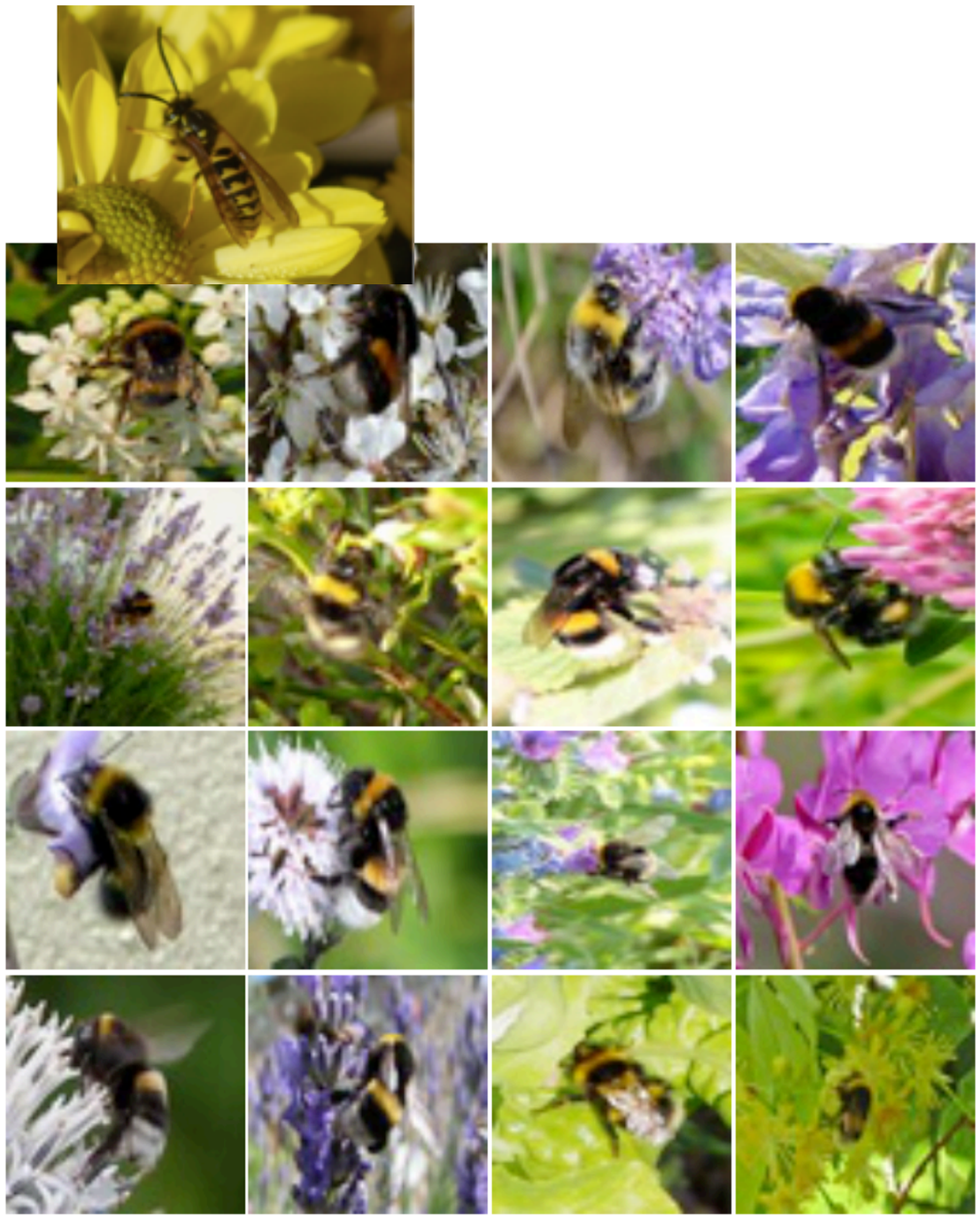
Classification of variable stars

Marc Monier (LAL),
Gilles Faÿ (Centrale-Supelec)

Light curves (luminosity vs time profiles)
- Static features
- Functional data

One day workshop: Accuracy improvement: %89 to %96

Ecology



Finding & Classifying Insects from Image Data

Paris Museum of Natural History, SPIPOLL.org, NVDIA, Université de Champagne-Ardenne ROMEO HPC Center

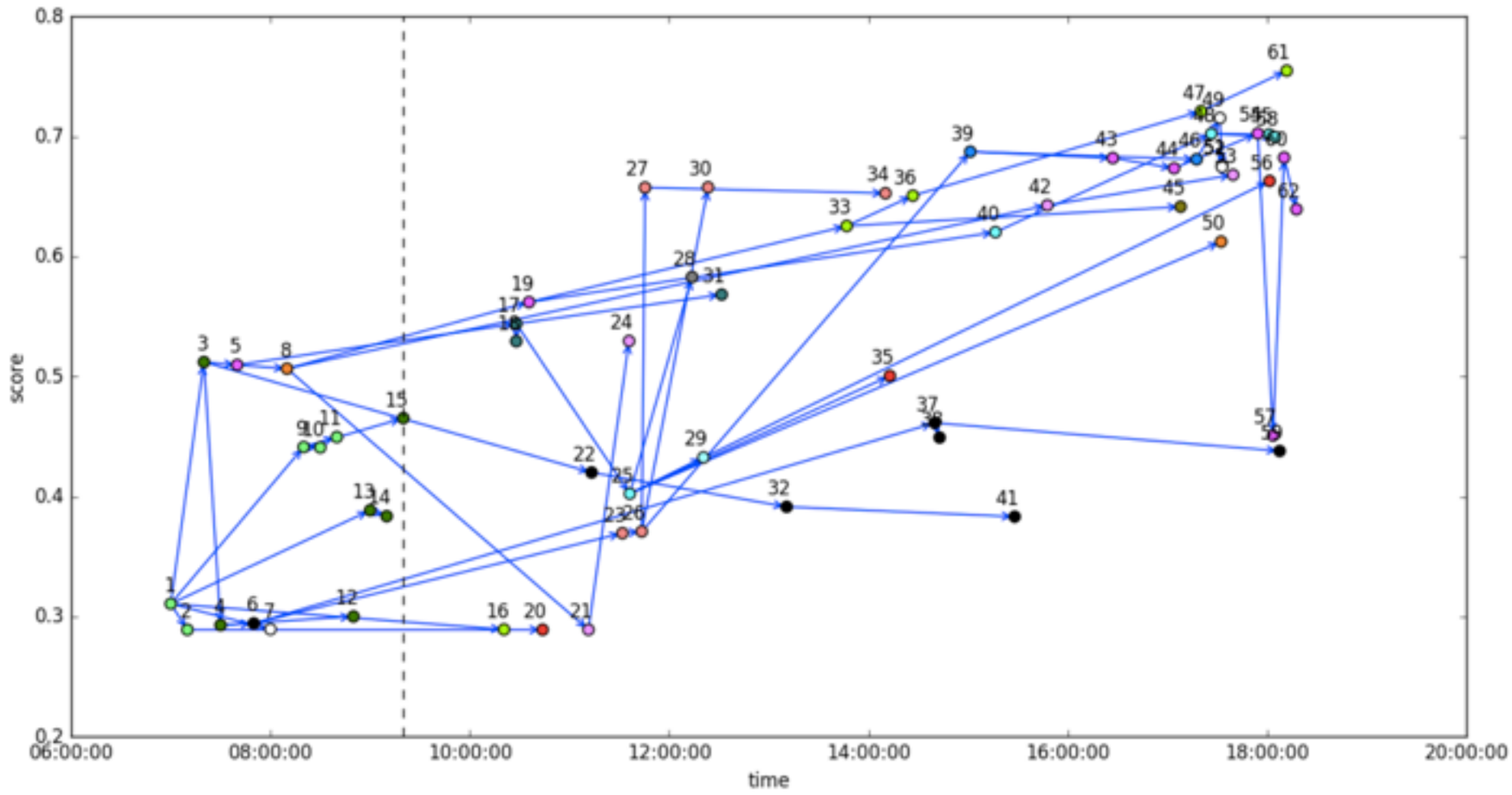
Pollenating Insects
- Image data (20K images)
- 18 types of insects
- Deep neural net models

One day event:
Improvement in prediction accuracy: from 0.31 to 0.70

Cancer Research

Epidemium est un programme de recherche scientifique participatif et ouvert dédié à la compréhension du cancer grâce aux Big Data, qui se concrétisera sous la forme d'un *data challenge*.

A graph of model similarities



Some observations

- **Steady progression**. They have built systematically on a submission they previously created, without being influenced by the others. Their performance may go either up (constantly) or down (constantly).
- **Breakdowns or jumps**. There are other groups, where the performance increased or decreased strongly from one modification submission to the next). There may be some robustness/vulnerability issue with their approach - to be further investigated.
- **Successful expansions.** An important “break” has happened at 12:00. This corresponds to “cropping” idea. Strangely, two very similar submissions (small distance) have been submitted at the same time - one of them did not improve the score at all (around 0.35, while the leader was around 0.55), whereas the other improved considerably (0.65).
- Currently, **we see no dependency between this break and the winning solution**. This might be related to the way we have measured the code similarity.

"Note that, following the RAMP approach, this model is the result of **a succession of small improvements made on top of other participants' contributions**. We did not reach a prediction score of 0.71 in one shot, but after applying several tricks and manually tuning some parameters."

Heuritech,
Winner of Insect Challenge
Blog entry

How to compare design concepts - represented as code?

RAMP
Rapid Analytics and Model Prototyping

Number of air passengers prediction

my submissions
new submission
leaderboard
log out

Leaderboard > hien-long-bruno_ly > GBYTES > feature_extractor.py

```python
import pandas as pd
import os


class FeatureExtractor(object):
    def __init__(self):
        pass

    def fit(self, X_df, y_array):
        pass

    def transform(self, X_df):
        X_encoded = X_df
        path = os.path.dirname(__file__)
        data_weather = pd.read_csv(os.path.join(path,
"external_data.csv"))
        X_weather = data_weather[['Date', 'AirPort', 'Max
TemperatureC']]
        X_weather = X_weather.rename(
            columns={'Date': 'DateOfDeparture', 'AirPort': 'Arrival'})
        X_encoded = X_encoded.set_index(['DateOfDeparture', 'Arrival'])
        X_weather = X_weather.set_index(['DateOfDeparture', 'Arrival'])
        X_encoded = X_encoded.join(X_weather).reset_index()

        X_encoded = X_encoded.join(pd.get_dummies(
            X_encoded['Departure'], prefix='d'))
        X_encoded = X_encoded.join(
            pd.get_dummies(X_encoded['Arrival'], prefix='a'))
        X_encoded = X_encoded.drop('Departure', axis=1)
        X_encoded = X_encoded.drop('Arrival', axis=1)

        X_encoded = X_encoded.drop('DateOfDeparture', axis=1)
        X_array = X_encoded.values
        return X_array
```

~

?

RAMP
Rapid Analytics and Model Prototyping

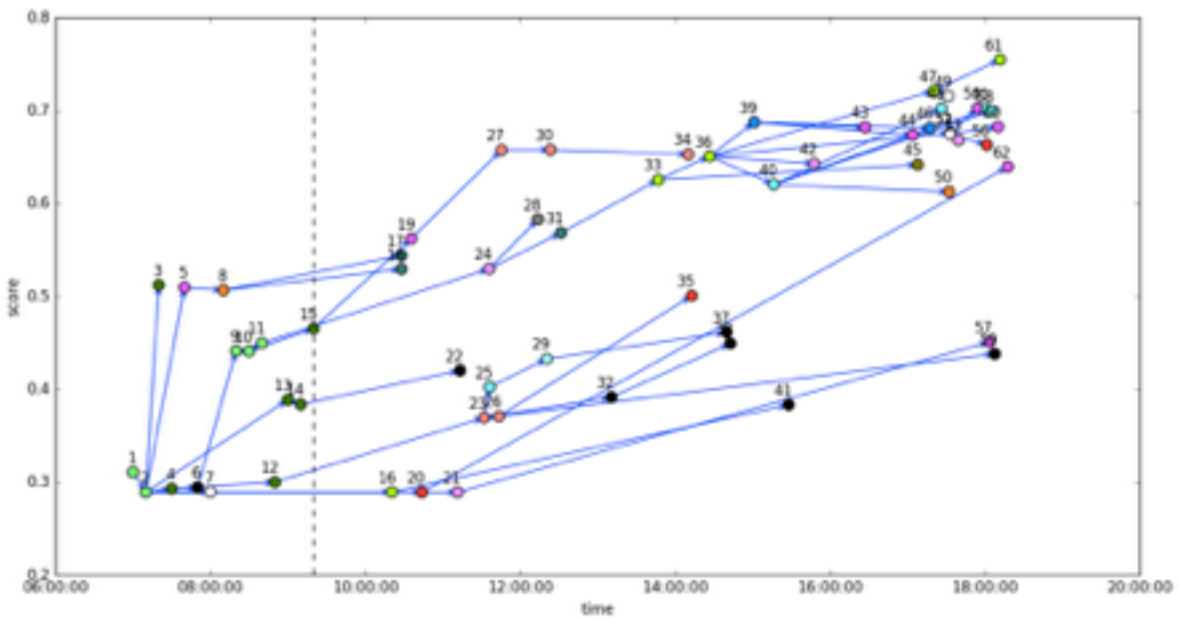
Number of air passengers prediction

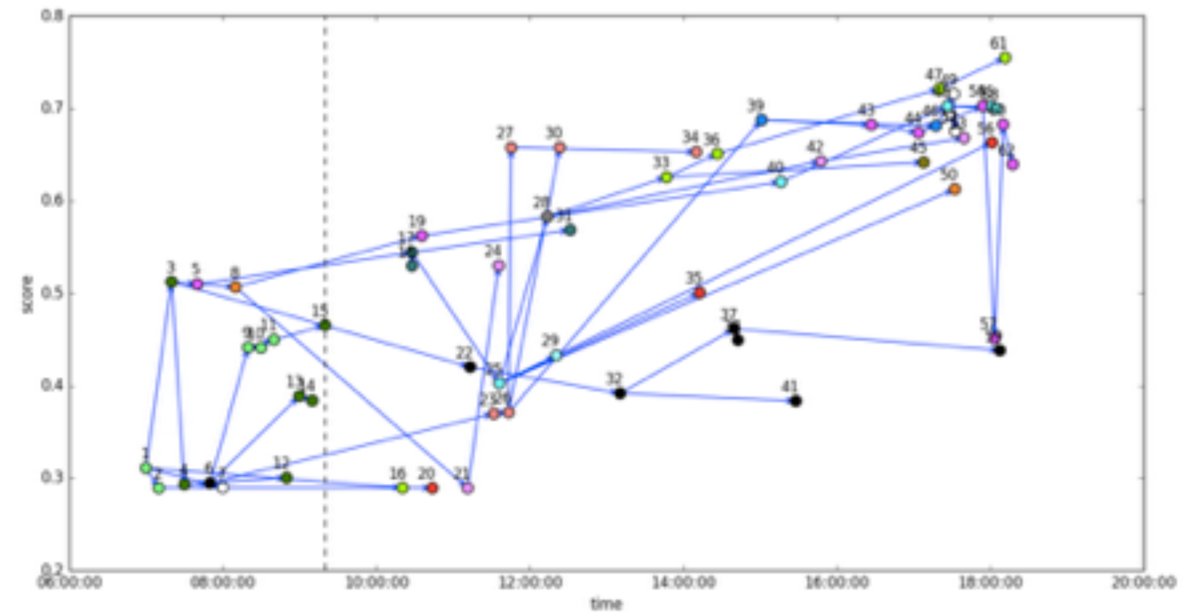
my submissions
new submission
leaderboard
log out

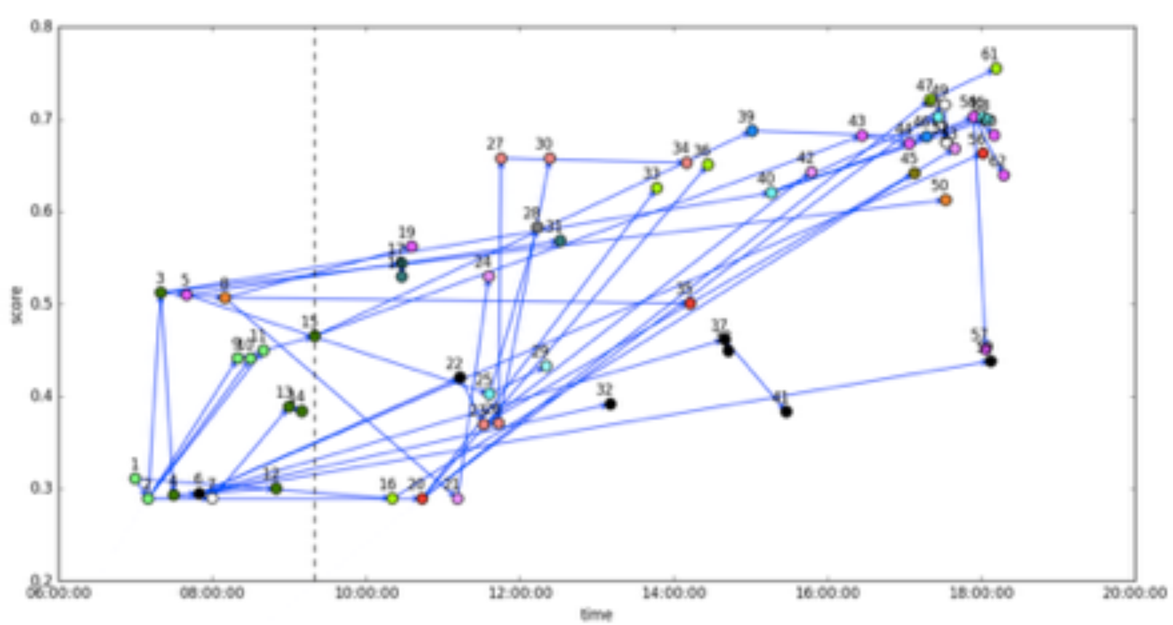
Leaderboard > djabbz > NdirouUnJob > feature_extractor.py

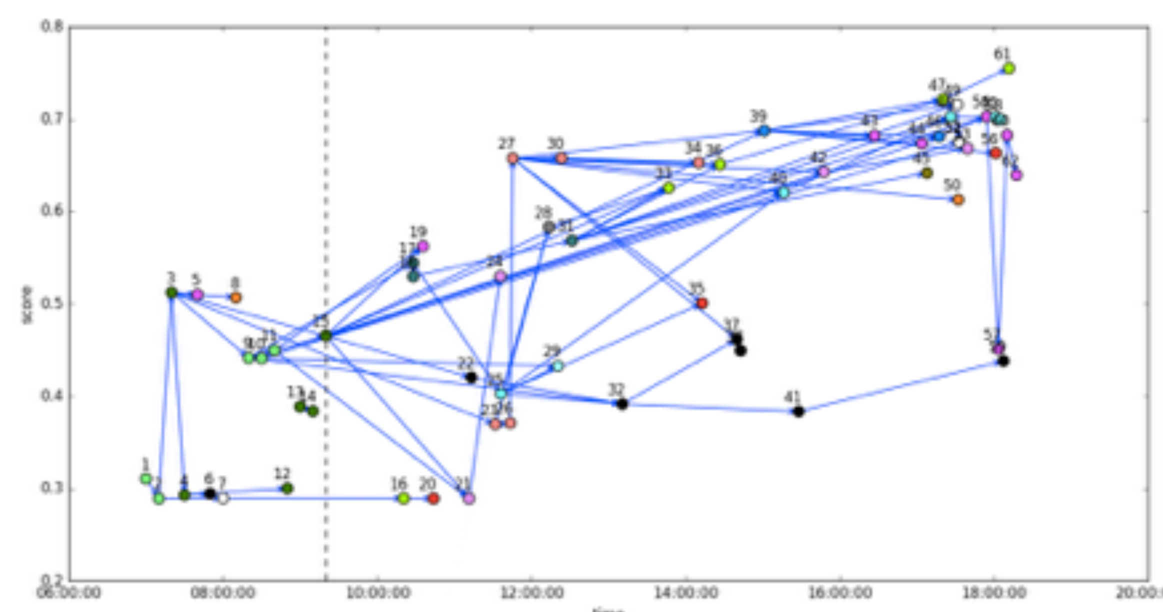
```python
import pandas as pd
import os

def get_binary_var(data, prefix=None, prefix_sep='_'):
    from pandas.tools.merge import concat
    import numpy as np

    data = data.astype('category')
    length = data.cat.categories.size
    length = int(np.ceil(np.log2(length)))
    code_string = data.cat.codes.map(lambda x:
'{0:b}'.format(x).rjust(length, '0'))

    codes = []
    for i in range(length):
        codes.append(code_string.str[i])

#     levels = data.cat.categories
    if prefix is None:
        prefix = data.name
    cols_name = ['%s%s' %(prefix, prefix_sep, i) for i in range(1,
length + 1)]
    df = concat(codes, axis=1)
    df.columns = cols_name
    return df.astype(np.int8)

class FeatureExtractor(object):
    def __init__(self):
        pass

    def fit(self, X_df, y_array):
        pass

    def transform(self, X_df):

        data_encoded = X_df

        state_pop = pd.DataFrame([('ATL', 9.9),
                                  ('BOS', 6.6),
                                  ('CLT', 9.7),
                                  ('DEN', 5.2),
                                  ('DFW', 26.6),
                                  ('DTW', 9.9),
                                  ('EWR', 8.9),
                                  ('IAH', 26.6),
                                  ('JFK', 19.5),
                                  ('LAS', 2.7),
                                  ('LAX', 38),
                                  ('LGA', 19.5),
                                  ('MCO', 19.3),
                                  ('MIA', 19.3),
                                  ('MSP', 5.4),
                                  ('ORD', 12.9),
                                  ('PHL', 12.8),
                                  ('PHX', 6.9),
                                  ('SEA', 6.9),
                                  ('SFO', 38)], columns=['airports',
'pop'])
        data_encoded = data_encoded.merge(state_pop, how='outer',
left_on='Departure', right_on='airports').drop('airports', axis=1)
        data_encoded = data_encoded.merge(state_pop, how='outer',
left_on='Arrival', right_on='airports').drop('airports', axis=1)

        data_encoded =
data_encoded.join(get_binary_var(data_encoded['Departure'],
prefix='dep'))
        data_encoded =
data_encoded.join(get_binary_var(data_encoded['Arrival'],
prefix='arr'))
        data_encoded = data_encoded.drop('Departure', axis=1)
        data_encoded = data_encoded.drop('Arrival', axis=1)

        data_encoded['DateOfDeparture'] =
pd.to_datetime(data_encoded['DateOfDeparture'])
        data_encoded['year'] = data_encoded['DateOfDeparture'].dt.year
        data_encoded['month'] =
data_encoded['DateOfDeparture'].dt.month
        data_encoded['day'] = data_encoded['DateOfDeparture'].dt.day
        data_encoded['weekday'] =
data_encoded['DateOfDeparture'].dt.weekday
        data_encoded['week'] = data_encoded['DateOfDeparture'].dt.week
        data_encoded['n_days'] =
data_encoded['DateOfDeparture'].apply(lambda date: (date -
pd.to_datetime('1970-01-01')).days)

        data_encoded = data_encoded.drop('DateOfDeparture', axis=1)

        data_encoded =
data_encoded.join(get_binary_var(data_encoded['year'], prefix='y'))
        data_encoded =
data_encoded.join(get_binary_var(data_encoded['month'], prefix='m'))
        data_encoded =
```

Various distance measures

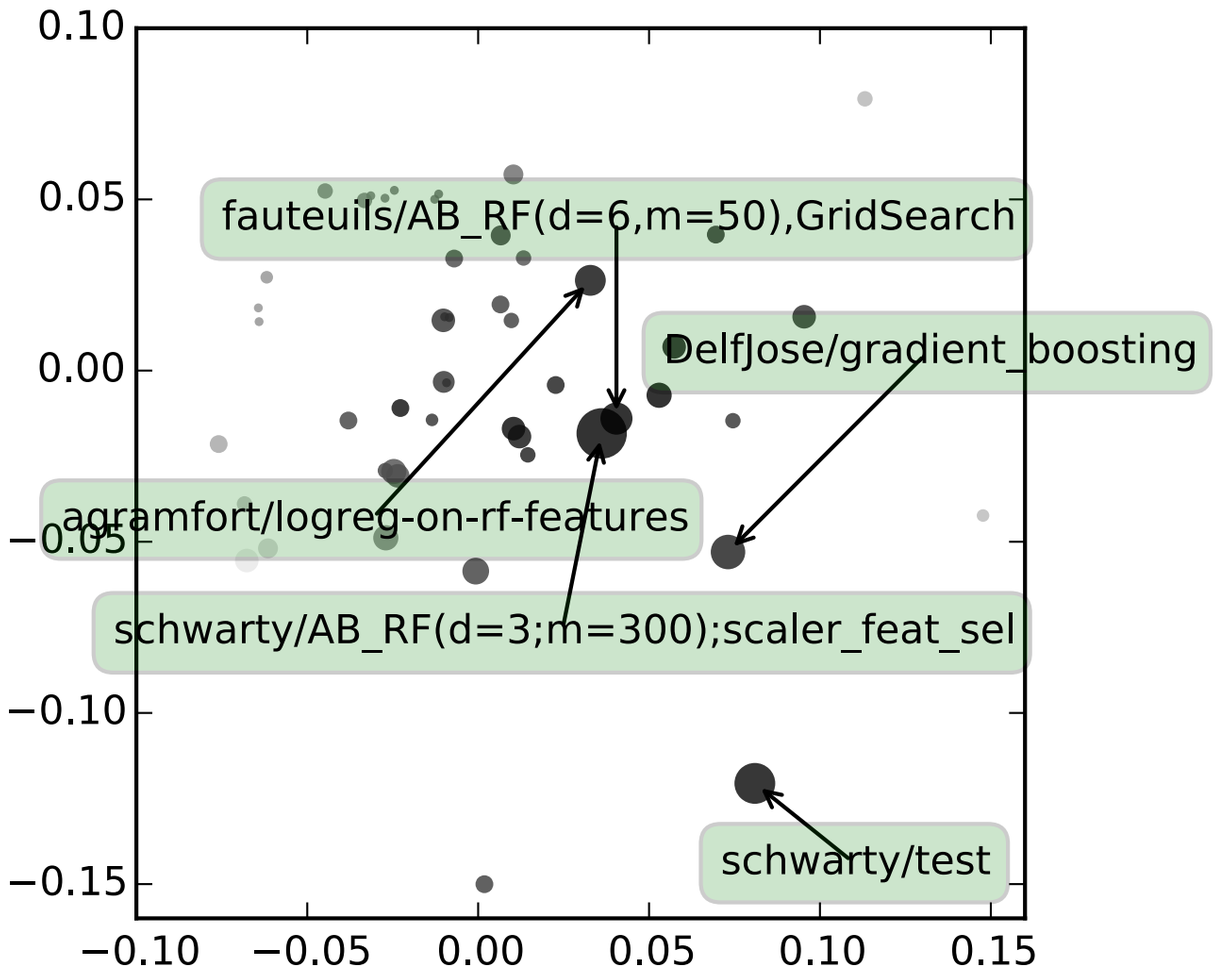








Comparing performance profiles: Promoting novelty search



2D projection (MDS) of model's prediction profiles

- Greyness: Model's raw score
- Size: model's contribution
- Position: similarity/dissimilarity in predictions

We Found (to be validated by further studies):

- Gravitation: following a given submission, others are hovering around the same coordinates, by incremental adjustments
- Repulsion: new submission using out-of-the-box code to explore the white space (no previous close-by submissions exist)
- Hybridation: opportunistic integration of previous submissions, involving/inspired by at least two different source of code.

In progress:

- Monitoring & Modelling “contribution”
  (Pierre Fleckinger, Economic Agents & Incentive Theory)
- Pushing towards “Novelty Search”
  (Jean-Bastiste Mouret, Novelty Search)
- Controlled experiments

RAMP platform

- RAMP platform is meant to be a free tool for researchers and students; this opens up new perspectives (pedagogy & research) and hopefully brings closer different communities

Thank you

Akin Kazakci, Mines ParisTech
akin.kazakci@mines-paristech.fr

MINES ParisTech

d t m i design theory and methods for innovation

Paris-Saclay Center for Data Science

cnrs