RAMP

Data challenges with modularization and code submission

Lessons learned

Balázs Kégl

Université Paris-Saclay / CNRS

OUTLINE

- A **short history** of RAMPs
  - **motivations**, **design principles**, and the **current tool**
- Three data challenges
  - **anomaly detection** in the LHC ATLAS detector
  - **classifying** and **quantifying** drug preparations for cancer therapy
  - **time series forecasting** of El Niño
- What have we learned?
  - **number of participants**, **incentives**?
  - **open** vs **closed**?
  - **blending** vs **human ingenuity**

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, building interfaces, matching people, helping them launching projects

345 affiliated researchers, 50 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

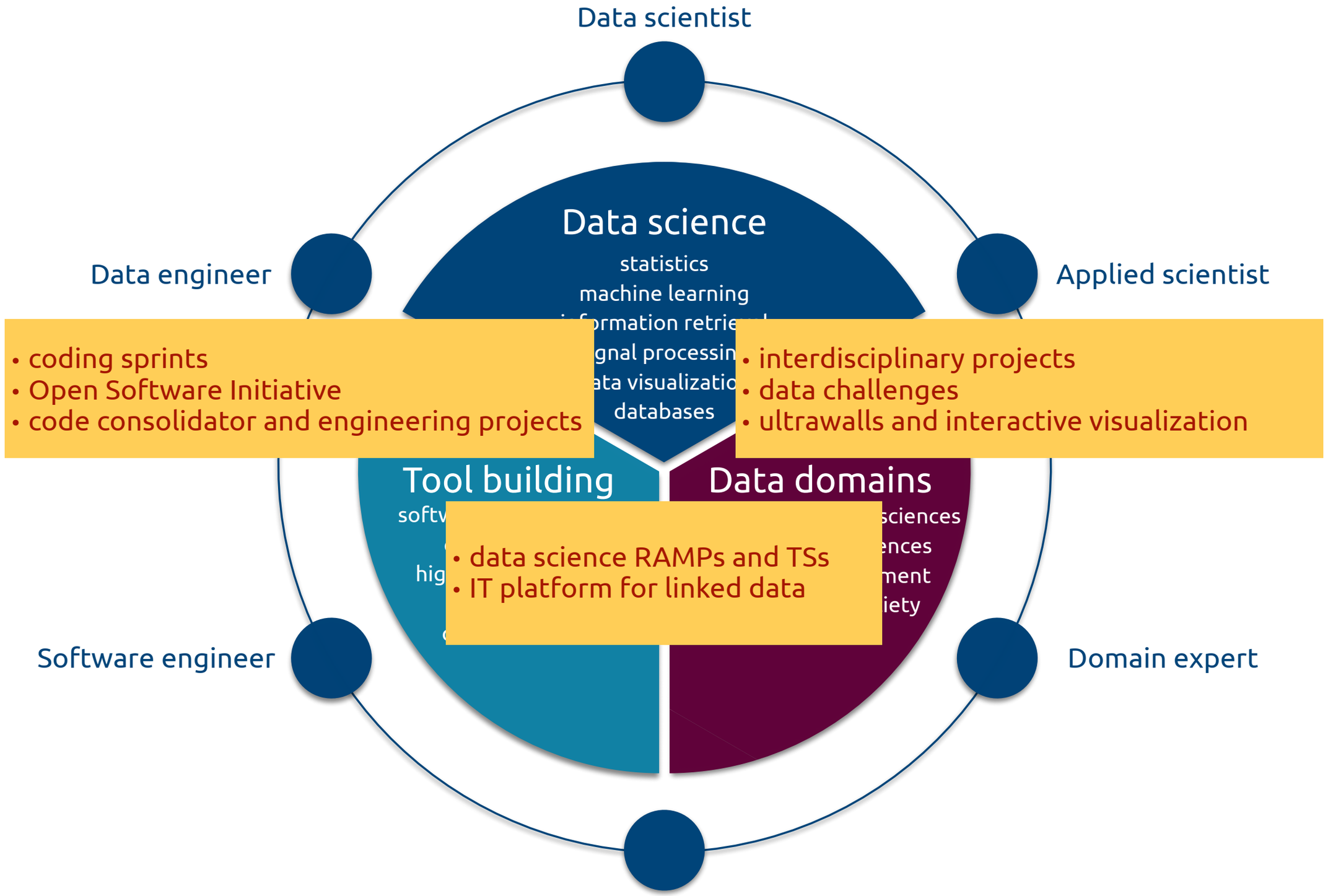
Visualization
INRIA
LIMSI

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

https://medium.com/@balazskegl/the-data-science-ecosystem-678459ba6013

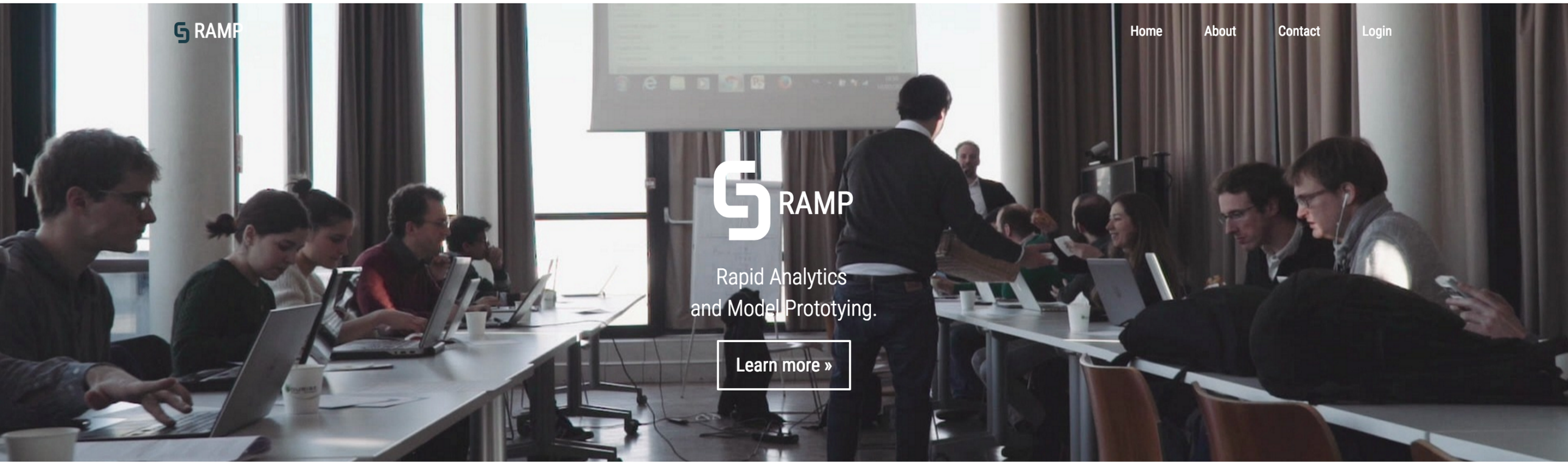


LACK OF TOOLS

- We have realized how much **data scientists are ill-equipped to manage the data science development process**
  - **collaborating with domain scientists or business units** on **data-driven problem/product formulation**
  - bench for **logging experiments**, guiding (human) **model search and tuning**
  - **managing data science teams** and **collaborating with each other**
  - **collaborating with AI** (out-of-the-box AutoML does not work: the search space is too big)
  - managing the **productionalization** of a prototype model

RAPID ANALYTICS AND MODEL PROTOTYPING (RAMP)

http://www.ramp.studio



Team


Balázs Kégl


Alex Gramfort


Akin Kazakçi


Camille Marini


Mehdi Cherti


Yohann Sitruk

LIMITATIONS OF DATA CHALLENGES

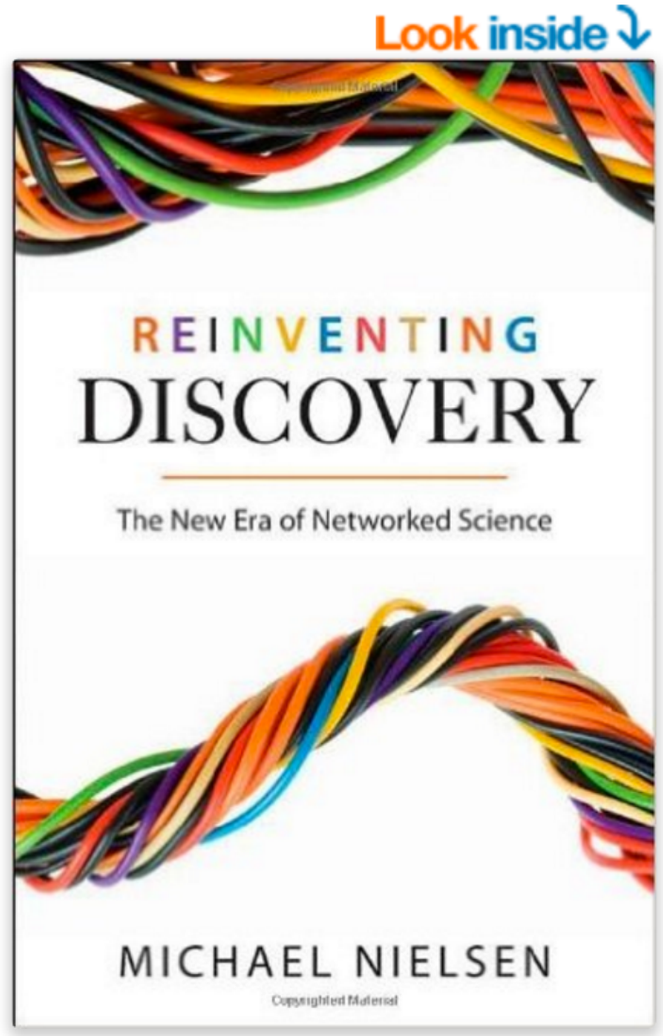
- Organizers have **no direct access to solutions**

- Emphasize **competition**: participants **cannot build on each other's solutions**

- **No modularization**: **ideas go unnoticed** unless packaged into a top submission

RAMP

- Challenge with **code submission**
- Following Nielsen's three crowdsourcing principles:
  - **modularity**: workflows are sliced into **workflow element modules** that can be tackled independently
  - **encourage small contributions**: e.g., copy another submission, add features, change the hyperparameters, resubmit
  - rich and well structured **information commons**: **open and download** each other's code, discuss on **slack**



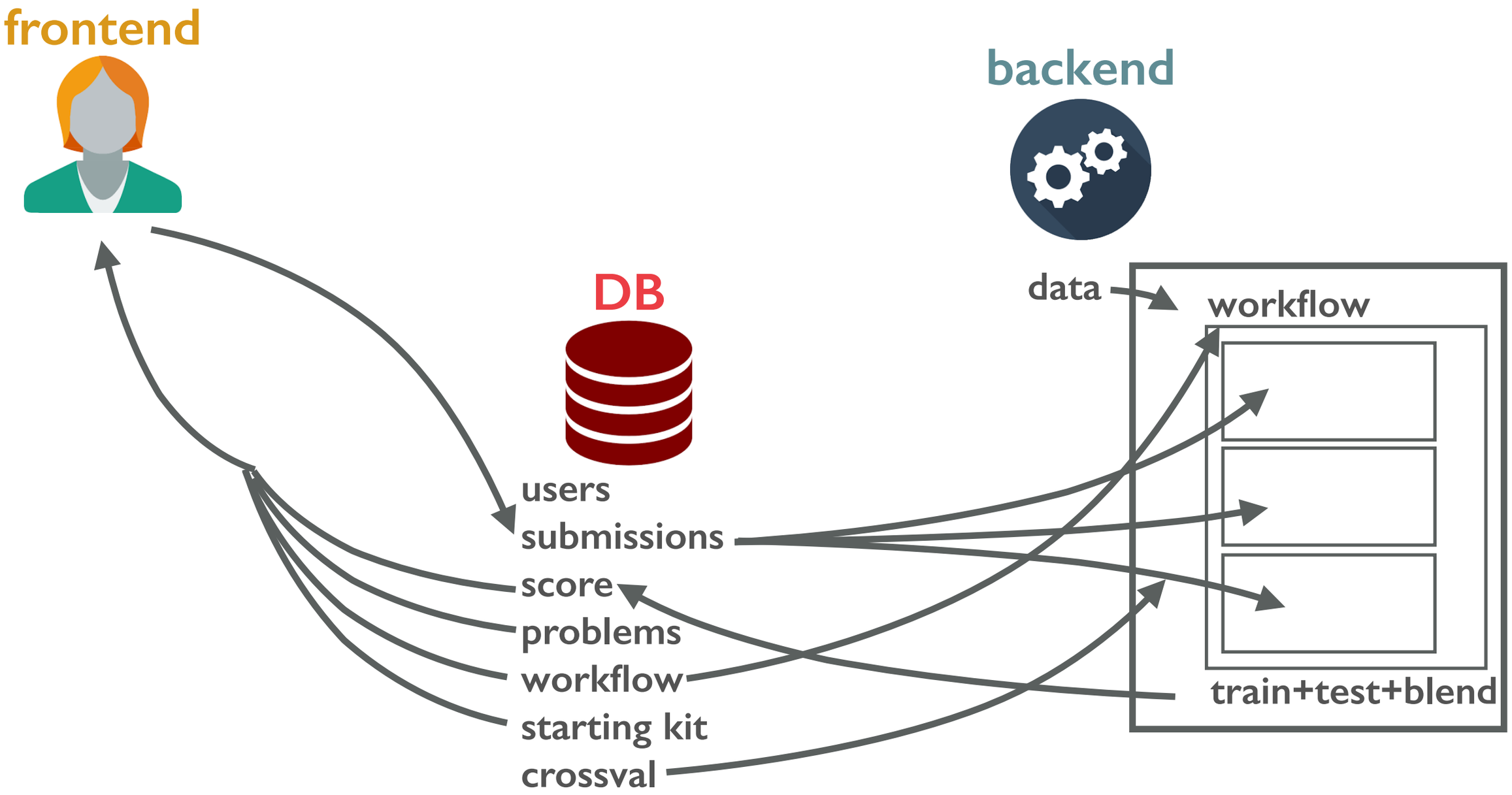
RAMP

RAPID ANALYTICS AND MODEL PROTOTYPING

- Roughly two formats
  - **single day hackatons** with 20-50 participants, **open leaderboard**, 15 minute timeout
  - 1-3 week **course challenges** up 150 students (but no limit really): **closed phase** with 1-3 submissions per day **followed by an open phase** with 15 minute timeout
- **600+ users**, **5000+ models**

RAMP

RAPID ANALYTICS AND MODEL PROTOTYPING



RAMP

www.ramp.studio/events/sea_ice_M1XMAP583_201617

RAMP

Hi Balazs!

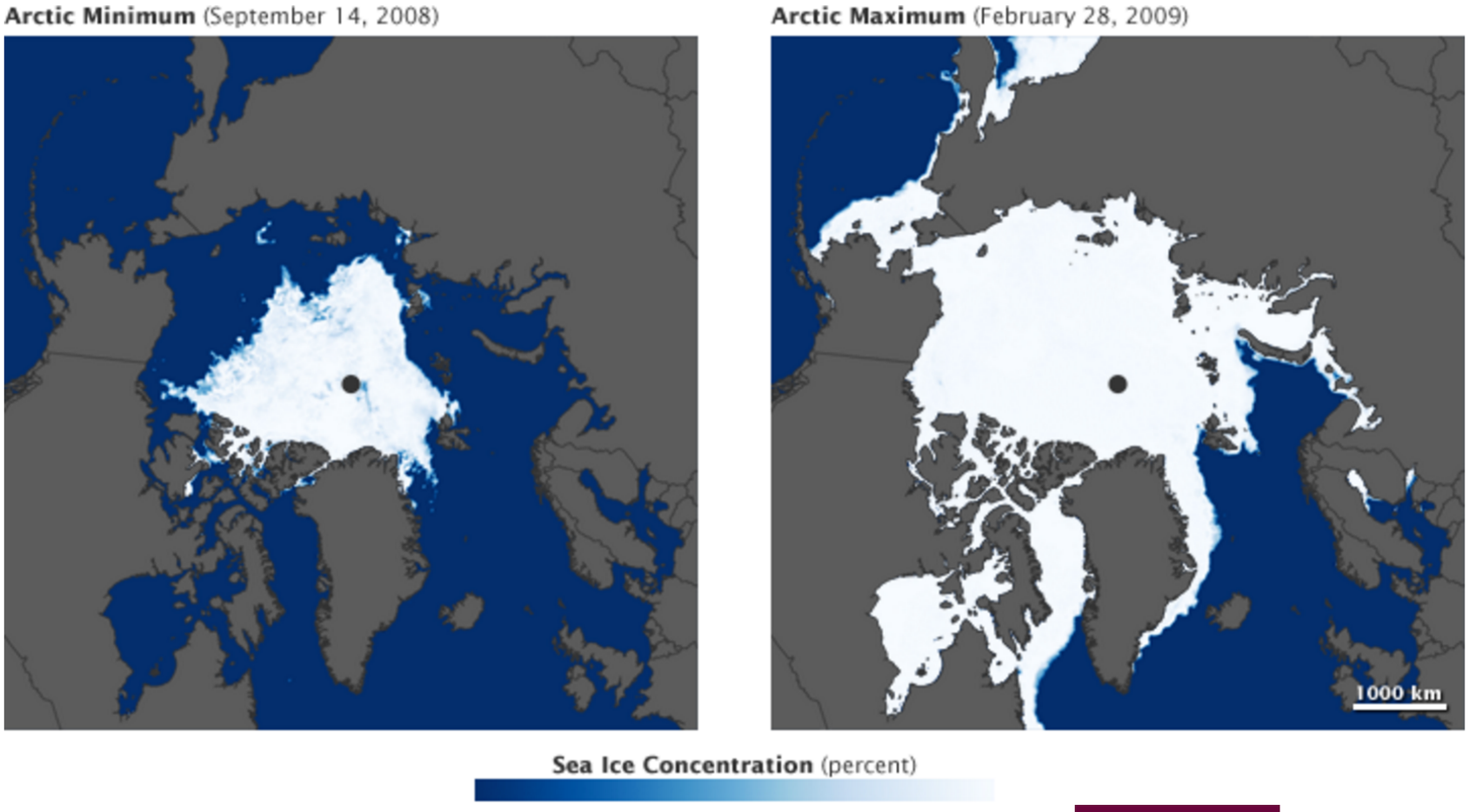
sea_ice_M1XMAP583_201617

Description

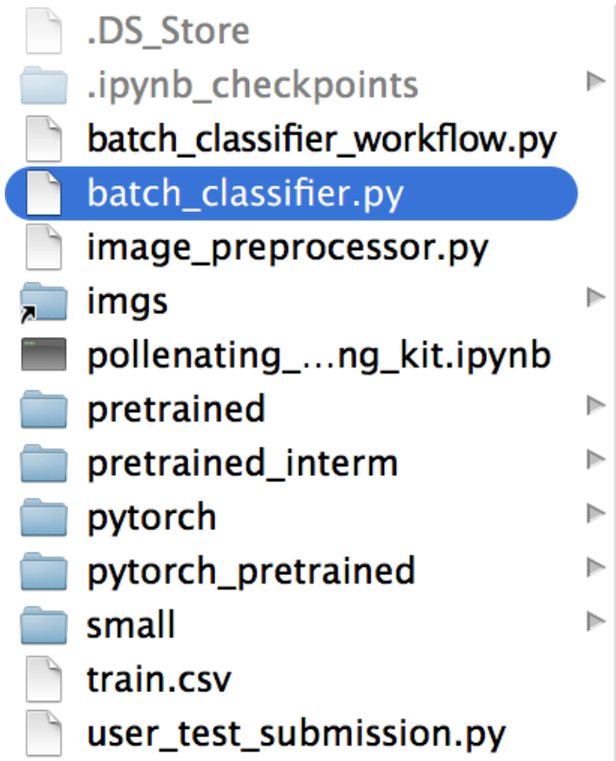
Balázs Kégl (CNRS), Camille Marini (CNRS), Andy Rhines (UW), Jennifer Dy (NEU), Arindam Banerjee (UMN)

Introduction

Arctic sea ice cover is one of the most variable features of Earth's climate. Its annual cycle peaks at around 15 million square kilometers in early spring, melting back to a minimum of about 6 million square kilometers in September. These seasonal swings are important for Earth's energy balance, as ice reflects the majority of sunlight while open water absorbs it. Changes in ice cover are also important for marine life and navigation for shipping.



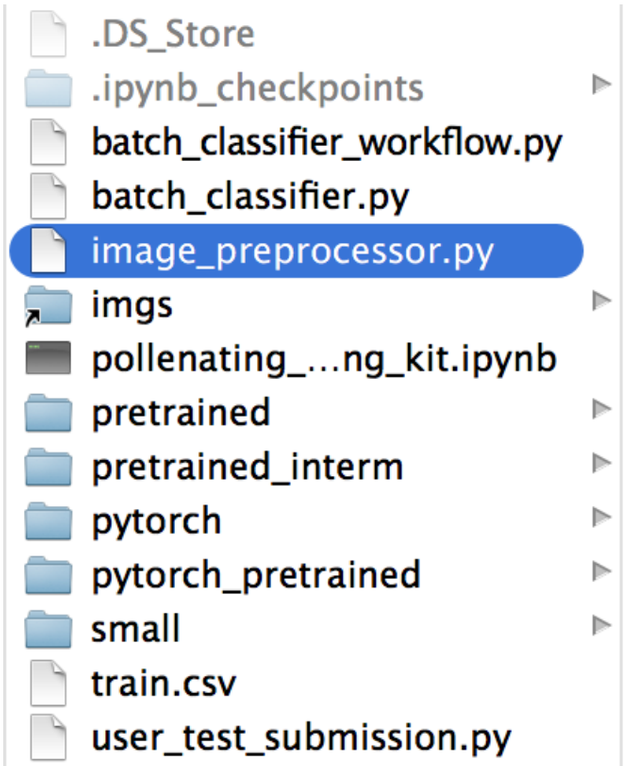
RAMP




```
from keras.models import Model
from keras.layers import Input
from keras.layers import Dense
from keras.layers import Convolution2D
from keras.layers import ZeroPadding2D
from keras.layers import MaxPooling2D
from keras.layers import Flatten
from keras.optimizers import Adam

class BatchClassifier(object):

    def __init__(self):
        self.model = build_model()

    def fit(self, gen_builder):
        gen_train, gen_valid, nb_train, nb_valid =
gen_builder.get_train_valid_generators(batch_size=64,
valid_ratio=0.1)
        self.model.fit_generator(
                gen_train,
                samples_per_epoch=nb_train,
                nb_epoch=3,
                # In parallel to training, a CPU process
loads and preprocesses data from disk and put
                # it into a queue in the form of mini-
batches of size `batch_size`.`max_q_size` controls
                # the maximum size of that queue.
                # The size of the queue should be big
enough so that the training process (GPU) never
                # waits for data (the queue should be
never be empty).
                # The CPU process loads chunks of 1024
images each time, and
                # 1024/batch_size mini-batches from that
chunk are put into the queue.
                # Assuming training the model on those
1024/batch_size mini-batches is slower than
```

Name	batch_classifier.py
Kind	Python Source
Size	5 KB
Created	Friday, 2017March 10, at 15:16

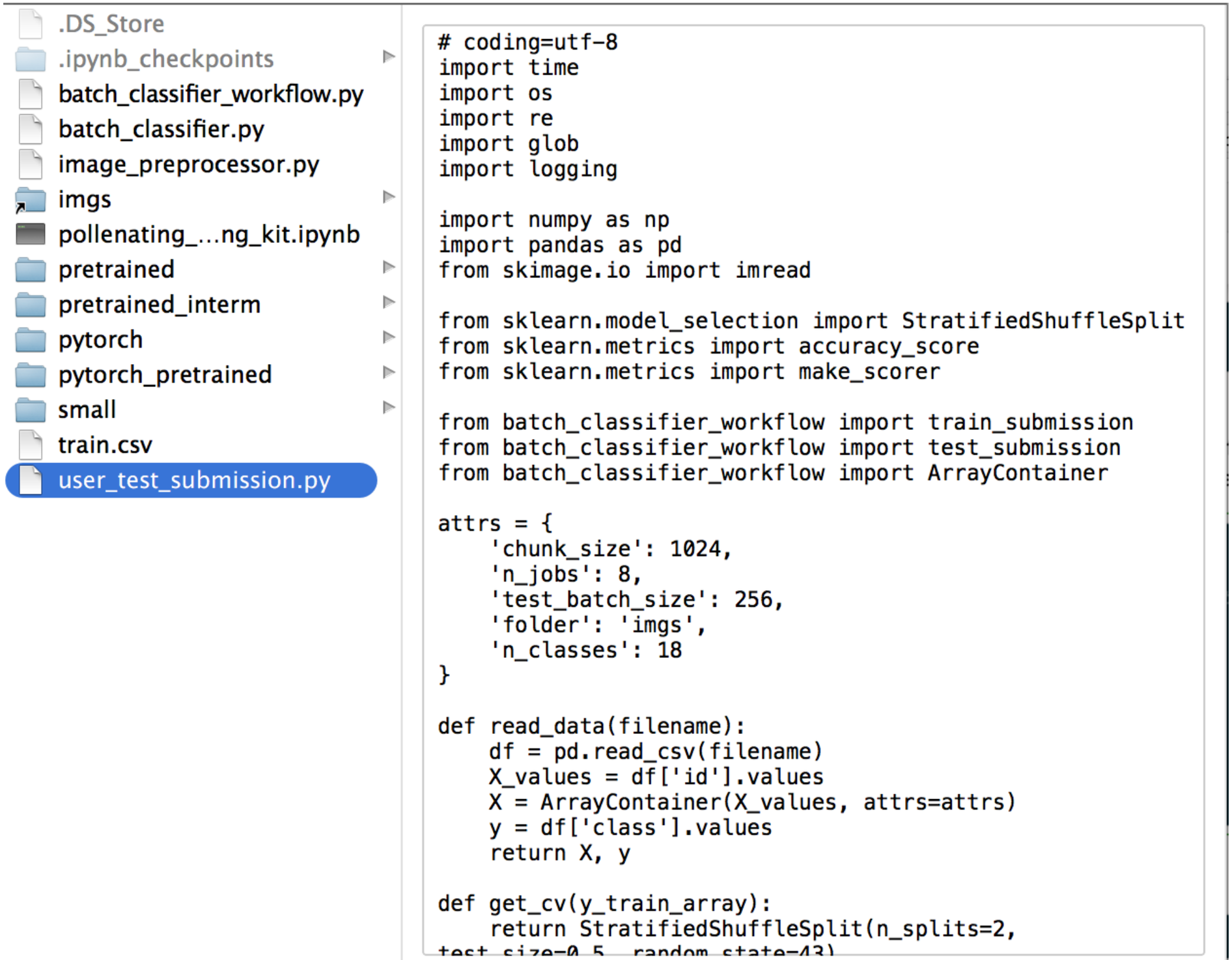
RAMP




```
import numpy as np
from skimage.transform import resize

def transform(x):
    if x.shape[2] == 4:
        x = x[:, :, 0:3]
    x = resize(x, (64, 64), preserve_range=True)
    x = x / 255.
    x = x.transpose((2, 0, 1))
    return x
```

Name image_preprocessor.py
Kind Python Source
Size 234 bytes
Created Thursday, 2017March 9, at 16:17

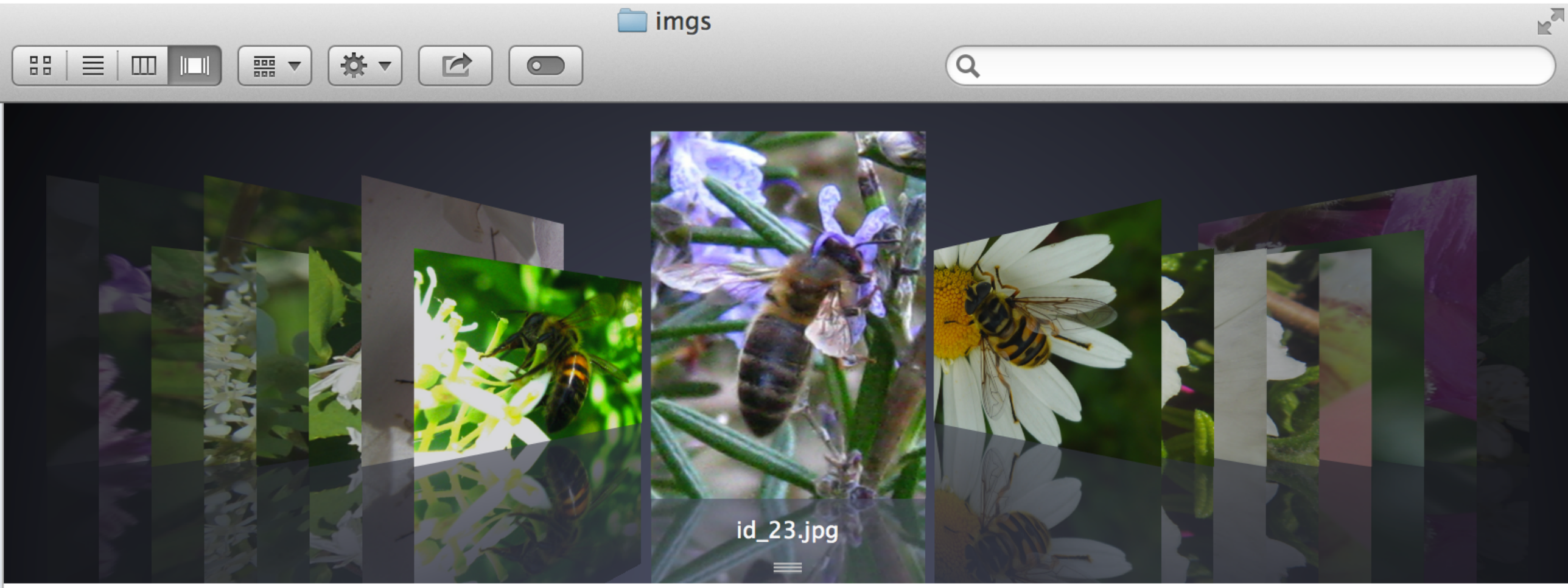
RAMP


.DS_Store
.ipynb_checkpoints
batch_classifier_workflow.py
batch_classifier.py
image_preprocessor.py
imgs
pollenating_...ng_kit.ipynb
pretrained
pretrained_interm
pytorch
pytorch_pretrained
small
train.csv
user_test_submission.py
# coding=utf-8
import time
import os
import re
import glob
import logging
import numpy as np
import pandas as pd
from skimage.io import imread
from sklearn.model_selection import StratifiedShuffleSplit
from sklearn.metrics import accuracy_score
from sklearn.metrics import make_scorer
from batch_classifier_workflow import train_submission
from batch_classifier_workflow import test_submission
from batch_classifier_workflow import ArrayContainer
attrs = {
'chunk_size': 1024,
'n_jobs': 8,
'test_batch_size': 256,
'folder': 'imgs',
'n_classes': 18
}
def read_data(filename):
df = pd.read_csv(filename)
X_values = df['id'].values
X = ArrayContainer(X_values, attrs=attrs)
y = df['class'].values
return X, y
def get_cv(y_train_array):
return StratifiedShuffleSplit(n_splits=2,
Name user_test_submission.py
Kind Python Source
Size 1 KB
Created Sunday, 2017March 12, at 10:33

RAMP

```
silver6:starting_kit 17 kegl$ python user_test_submission.py
length of training array: 2304 months = 192 years
length of test array: 1296 months = 108 years
length of burn in: 120 months = 10 years
length of common block: 1032 months = 86 years
length of validation block: 1152 months = 96 years
length of each cv block: 144 months = 12 years
/Users/kegl/anaconda/lib/python2.7/site-packages/scipy/linalg/basic.py:884: RuntimeWarning: internal gelsd driver lwork query error, re
quired iwork dimension not returned. This is likely the result of LAPACK bug 0038, fixed in LAPACK 3.2.2 (released July 21, 2010). Fall
ing back to 'gelss' driver.
  warnings.warn(mesg, RuntimeWarning)
train RMSE = 0.554 ; valid RMSE = 0.599 ; test RMSE = 0.654 ; train RMSESept = 0.545 ; valid RMSESept = 0.565 ; test RMSESept = 0.651
train RMSE = 0.552 ; valid RMSE = 0.593 ; test RMSE = 0.646 ; train RMSESept = 0.555 ; valid RMSESept = 0.578 ; test RMSESept = 0.662
train RMSE = 0.547 ; valid RMSE = 0.599 ; test RMSE = 0.651 ; train RMSESept = 0.552 ; valid RMSESept = 0.598 ; test RMSESept = 0.674
train RMSE = 0.553 ; valid RMSE = 0.577 ; test RMSE = 0.639 ; train RMSESept = 0.56 ; valid RMSESept = 0.569 ; test RMSESept = 0.653
train RMSE = 0.549 ; valid RMSE = 0.596 ; test RMSE = 0.645 ; train RMSESept = 0.546 ; valid RMSESept = 0.615 ; test RMSESept = 0.65
train RMSE = 0.554 ; valid RMSE = 0.572 ; test RMSE = 0.628 ; train RMSESept = 0.546 ; valid RMSESept = 0.597 ; test RMSESept = 0.636
train RMSE = 0.553 ; valid RMSE = 0.578 ; test RMSE = 0.623 ; train RMSESept = 0.564 ; valid RMSESept = 0.477 ; test RMSESept = 0.619
train RMSE = 0.549 ; valid RMSE = 0.659 ; test RMSE = 0.625 ; train RMSESept = 0.56 ; valid RMSESept = 0.443 ; test RMSESept = 0.622
mean train RMSE = 0.551 ± 0.0025
mean valid RMSE = 0.596 ± 0.0255
mean test RMSE = 0.639 ± 0.0113
mean train RMSESept = 0.553 ± 0.007
mean valid RMSESept = 0.555 ± 0.0578
mean test RMSESept = 0.646 ± 0.0179
silver6:starting_kit 17 kegl$
```

RAMP



Name

- .DS_Store
- id_0.jpg
- id_1.jpg
- id_3.jpg
- id_4.jpg
- id_6.jpg
- id_7.jpg
- id_8.jpg
- id_13.jpg
- id_17.jpg
- id_23.jpg
- id_29.jpg

RAMP

localhost:8953/notebooks/sea_ice_starting_kit.ipynb

jupyter sea_ice_starting_kit (unsaved changes)

n_burn 4 of 8

File Edit View Insert Cell Kernel Help Python 2

Code Cell Toolbar: None

In [120]:
```
X_ds = xr.open_dataset('sea_ice_X_public_train.nc', decode_times=False)
y_array = np.load('sea_ice_y_public_train.npy')
```

y_array on the disk is already shifted by n_lookahead = 4 months. n-burnin = 120 (months) is the length of the prefix for which no prediction is required. If your feature extractor only uses these ten years of the past to extract features from, you don't need to worry about missing data in the beginning of the sequence. Otherwise you should take care of the issue "manually".

In [122]:
```
n_lookahead = 4
n_burn_in = 120  # 10 years;
# When cutting the data into train, test, and validation blocks, make sure that each block starts
# in January
# otherwise the
n_train = 200 * 12
n_test = len(X_ds['time']) - n_train
X_train_ds = X_ds.isel(time=slice(None, n_train))
y_train_array = y_array[:n_train]
X_test_ds = X_ds.isel(time=slice(n_train, None))
y_test_array = y_array[n_train:]
```

Printing it, you can see that it contains all the data, indices, and other metadata.

In [123]:
```
X_train_ds
```

Out[123]:
```
<xarray.Dataset>
Dimensions:   (lat: 39, lon: 58, time: 2400)
Coordinates:
  * time      (time) int64 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 ...
  * lat       (lat) float64 -90.0 -85.29 -80.58 -75.86 -71.15 -66.44 -61.73 ...
  * lon       (lon) float64 0.0 6.25 12.5 18.75 25.0 31.25 37.5 43.75 50.0 ...
Data variables:
    ice area  (time) float64 14.06 14.87 14.84 13.89 12.31 10.41 8.343 6.773 ...
```

RAMP

www.ramp.studio/events/sea_ice_M1XMAP583_201617/sandbox

RAMP

Hi Balazs!

Sandbox

You can either edit and save the code in the left column or upload the files in the right column. You can also import code from other submissions when the leaderboard links are open.

Edit and save your code!

ts_feature_extractor

```python
import numpy as np
import xarray as xr
from sklearn.linear_model import LinearRegression

class FeatureExtractor(object):

    def __init__(self):
        pass

    def transform(self, X_ds):
        """Compute the monthly averages of the ice_area, corresponding to the month
        The code could be simplified but in this way it is general, can be used for
        variables as well."""
        # This is the range for which features should be provided. Strip
        # the burn-in from the beginning and the prediction look-ahead from
        # the end.
        valid_range = np.arange(X_ds.attrs['n_burn_in'], len(X_ds['time']))

        # We convert the Dataset into a 4D DataArray
        X_xr = X_ds.to_array()

```

regressor

Upload your files!

File list

ts_feature_extractor.py

regressor.py

Upload file

Choose File No file chosen

Upload

RAMP

www.ramp.studio/events/sea_ice_M1XMAP583_201617/leaderboard

RAMP

Hi Balazs!

sea_ice_M1XMAP583_201617

Leaderboard

Combined score: 0.268

Show 10 entries

Search:

team	submission	contributivity	historical contributivity	rmse	train time	test time	submitted at (UTC)
joseph.budin	noName	26	3	0.279	286	3	2017-02-13 11:36:28 Mon
alexis.thual	timeseries	16	16	0.296	1	1	2017-02-13 17:48:47 Mon
julien.habis	try_hard3	11	8	0.300	475	3	2017-02-13 19:45:35 Mon
kangzheng.liang	thirdtry	7	7	0.291	8	1	2017-02-07 19:11:32 Tue
joseph.budin	LinReg	6	3	0.280	234	3	2017-02-13 11:25:39 Mon
gaetan.millerand	shifted+boost+nino	6	5	0.295	29	5	2017-02-04 21:04:11 Sat
thibaut.vasseur	starting_kit_help	4	4	0.289	17	9	2017-02-13 18:48:37 Mon
yu-jia.cheong	Last	3	3	0.289	18	7	2017-02-13 13:28:53 Mon
gaetan.millerand	random_test	3	3	0.295	30	5	2017-02-07 13:12:29 Tue
maxime.lapides	TestFinal	3	3	0.296	458	3	2017-02-13 17:44:37 Mon

Showing 1 to 10 of 172 entries

Previous 1 2 3 4 5 … 18 Next

RAMP

www.ramp.studio/credit/1cf3f4ba0a6b9b522ea4edaa476536ff6dc2d2c1

RAMP

Hi Balazs!

Quantify credits for el_nino/domitille.coulomb/Last Chance

Please take a couple of minutes to credit the sources of this submission: what percentage of it is new? what percentage of it is coming from or inspired by other submissions?

The numbers should add up to 100.

The list contains all submissions by team domitille.coulomb and submissions that team members have looked at.

The numbers will be used for computing the total contributivity of the submissions by propagating the current contributivity backwards.

Be honest and fair as much as possible.

submission	credit
self credit	10
el_nino/raphael.berdugo/end_submit	0
el_nino/domitille.coulomb/Third Trial	40
el_nino/bkabid/second_attempt	0
el_nino/domitille.coulomb/Second Trial	50
el_nino/domitille.coulomb/First Trial	0
el_nino/domitille.coulomb/starting_kit	0

Submit

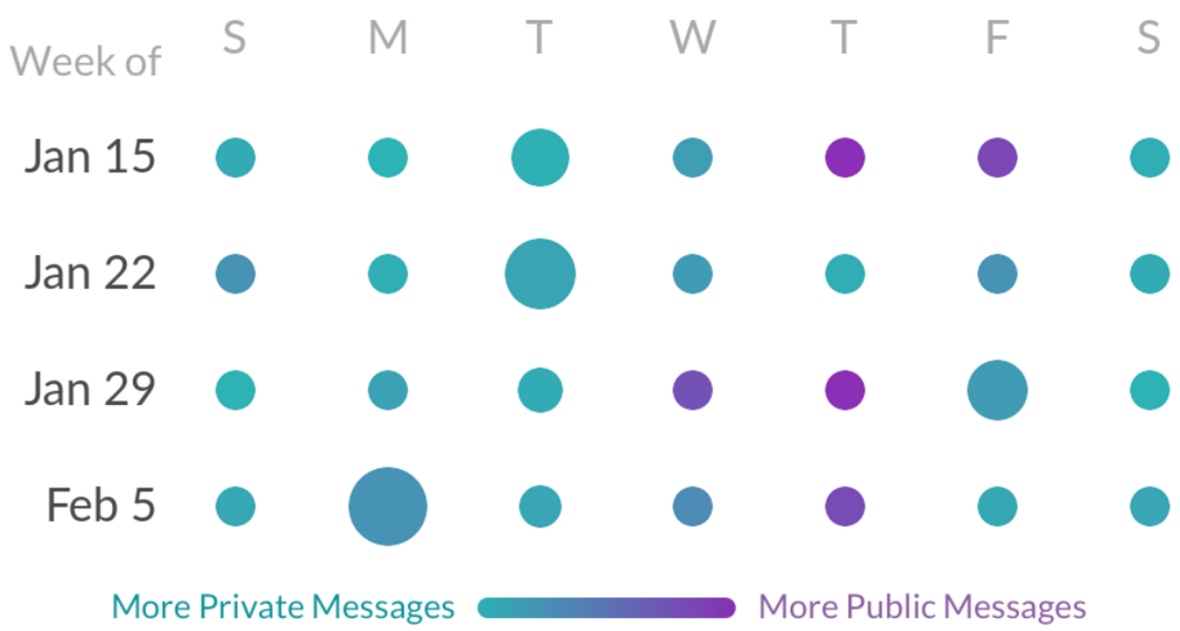
RAMP

slack

X MAP583 2016/17's Weekly Summary

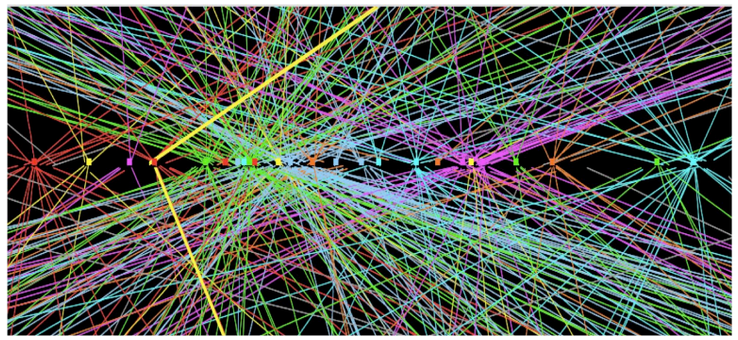
Sunday, February 5th – Saturday, February 11th

Your team sent a total of **326 messages** last week (that's 70 more than the week before). Of those, **24% were in public channels**, **6% were in private channels**, and **70% were direct messages**. Your team also uploaded **6 files** (that's 1 more than the week before).



Three recent RAMPs

ANOMALY DETECTION IN THE LHC ATLAS DETECTOR



reconstruction
+simulated anomalies

DER_mass_transverse_met_lep	1.937
DER_mass_vis	64.546
DER_pt_h	41.791
DER_deltar_tau_lep	2.301
DER_pt_tot	7.975
DER_sum_pt	105.305
DER_pt_ratio_lep_tau	0.926
DER_met_phi_centrality	1.087
PRI_tau_pt	36.259
PRI_tau_eta	-2.248
PRI_tau_phi	-2.239
PRI_lep_pt	33.582
PRI_lep_eta	-1.893
PRI_lep_phi	0.035
PRI_met	19.872
PRI_met_phi	-0.040
isSkewed	0.000

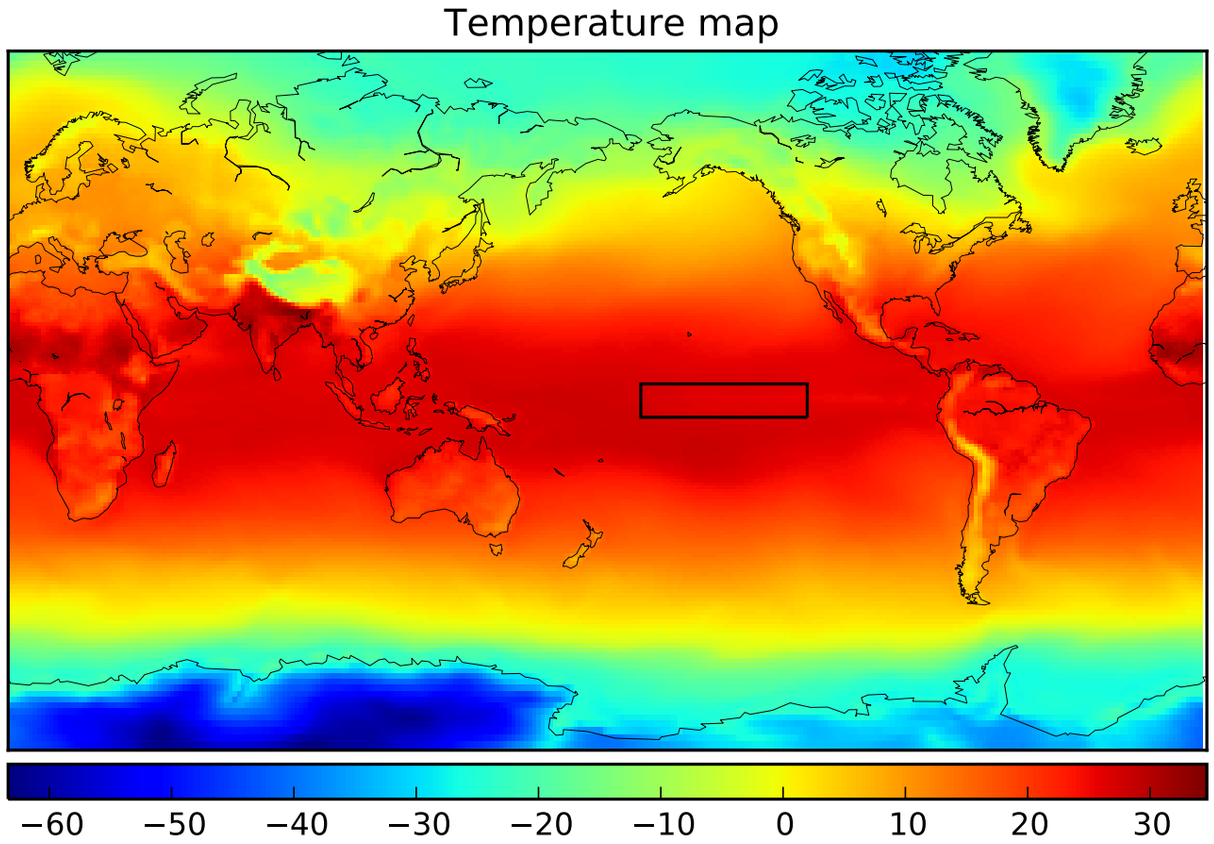
classifier

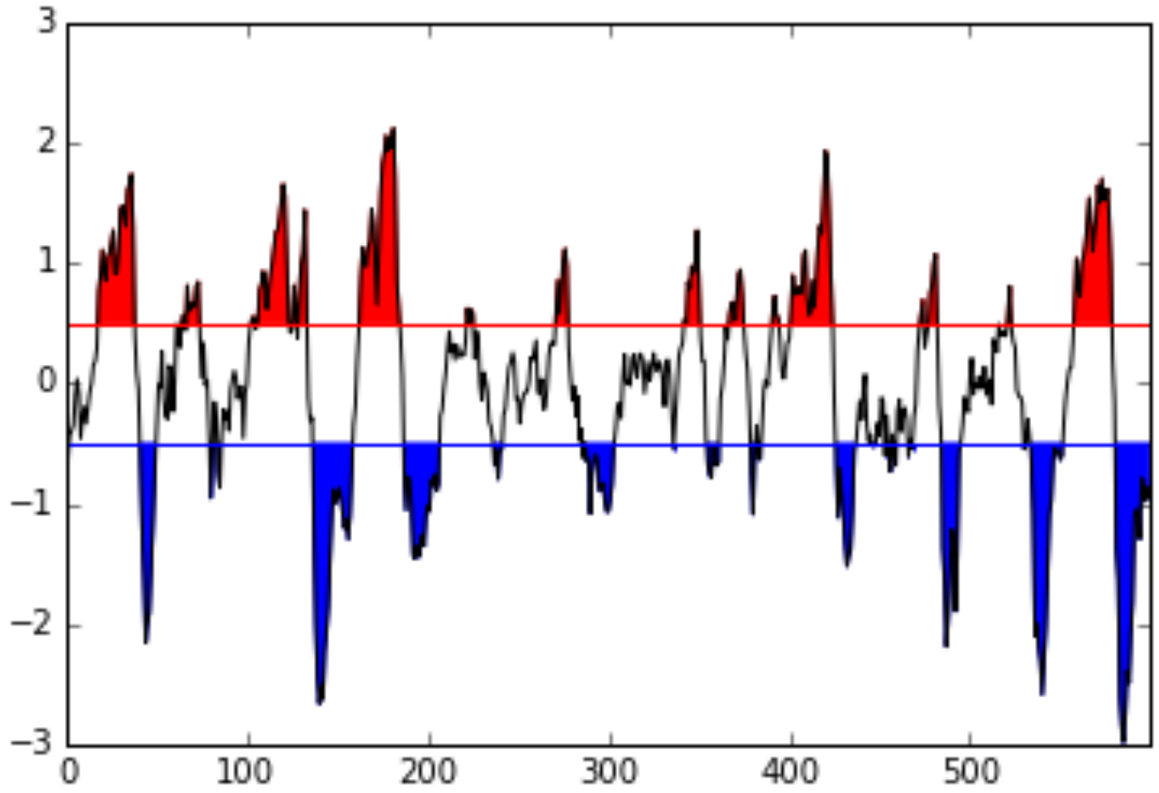
correct
(isSkewed = 0)

?

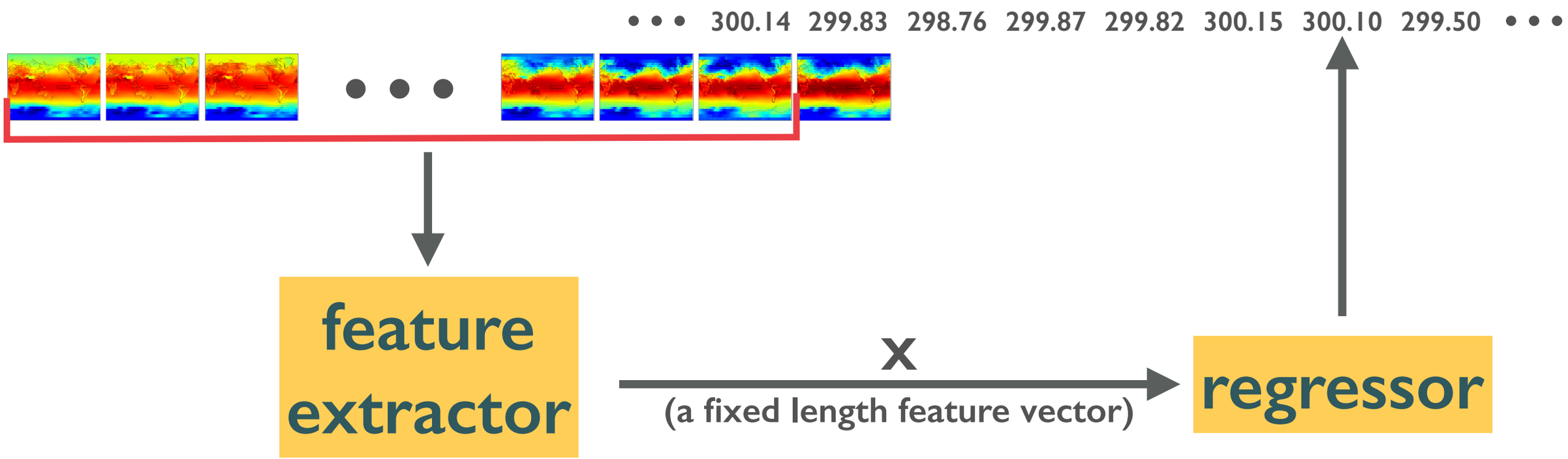
anomaly
(isSkewed = 1)

Forecasting El Nino six months ahead





FORECASTING EL NINO SIX MONTHS AHEAD



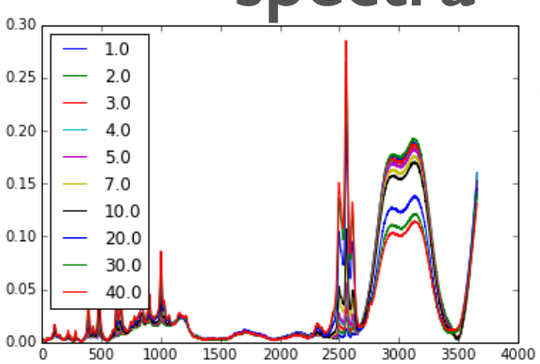
- We give the full series to the feature extractor
- It **could look ahead** in the future (even inadvertently)
- Checking lookahead by a **randomized test**

CLASSIFYING AND REGRESSING ON MOLECULAR SPECTRA

chemotherapy drug in elastic pocket

laser spectrometer

molecular spectra



feature extractor 1

classifier

drug type

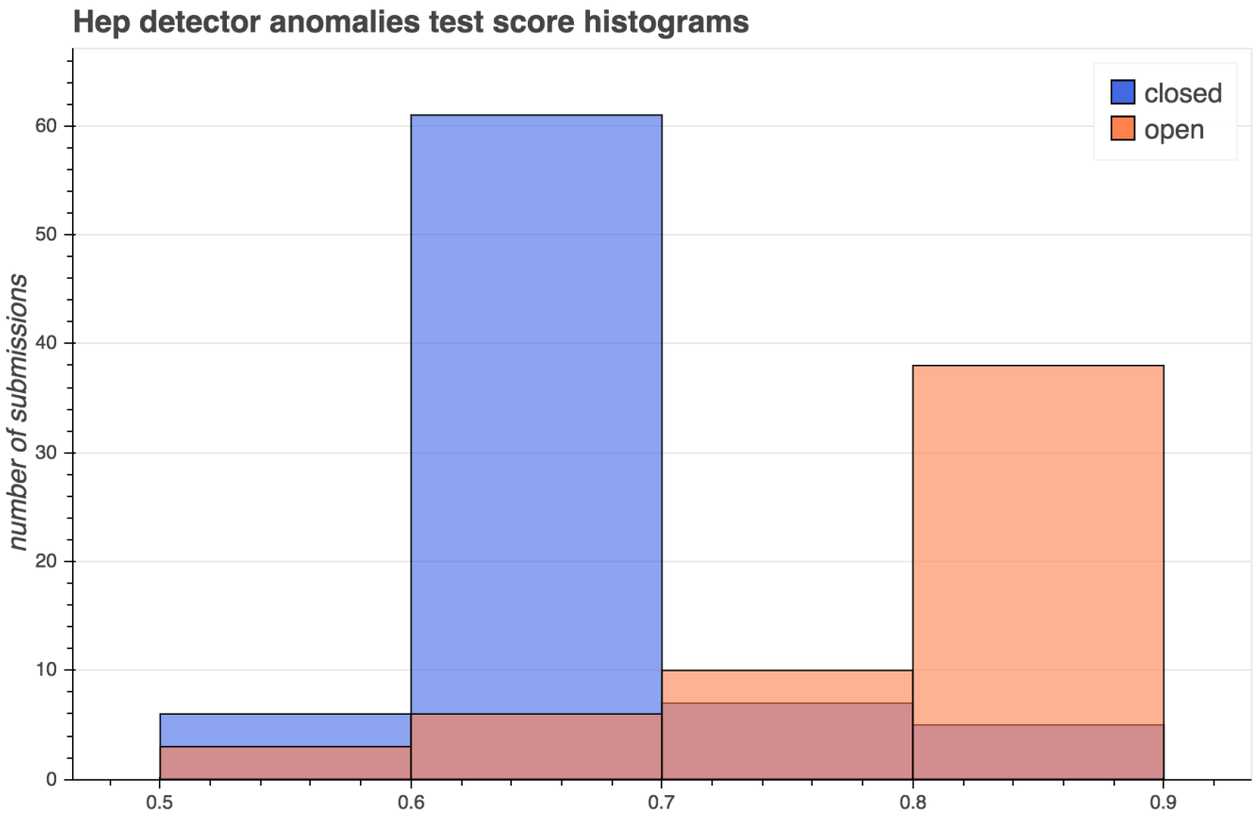
feature extractor 2

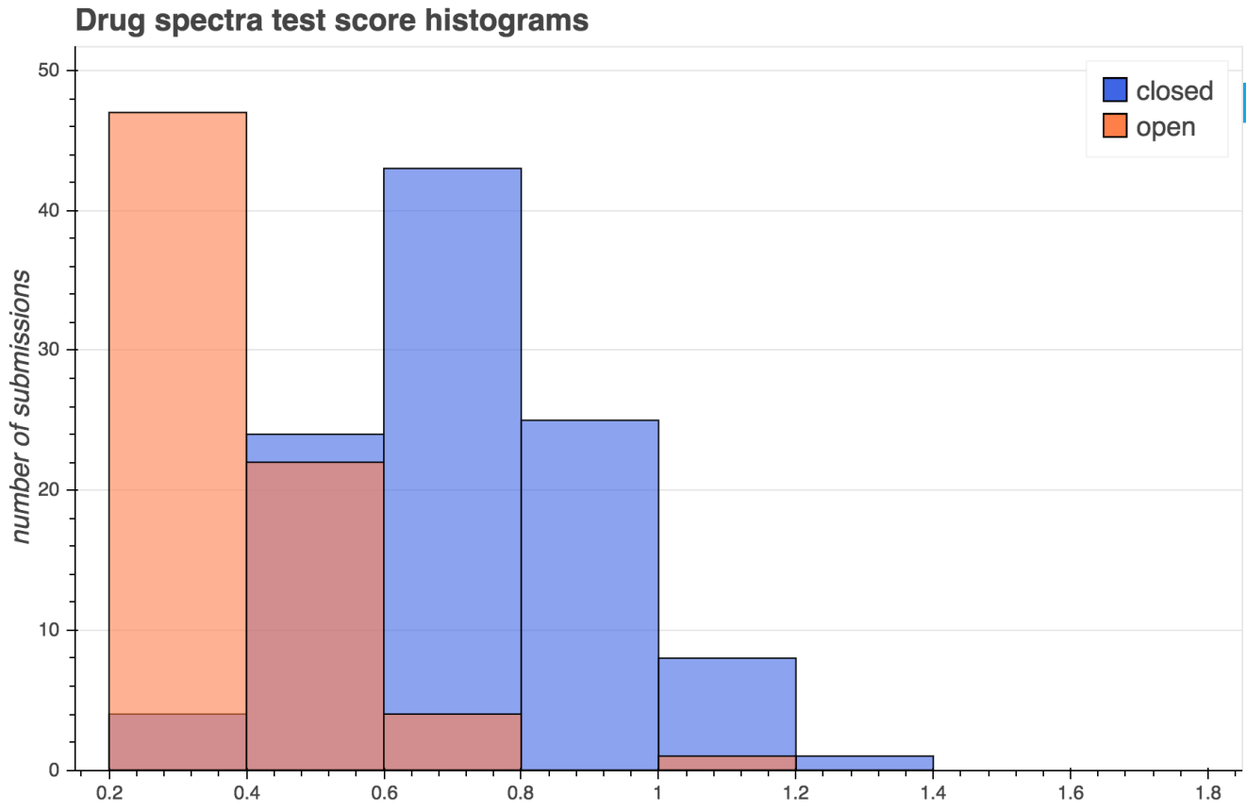
regressor

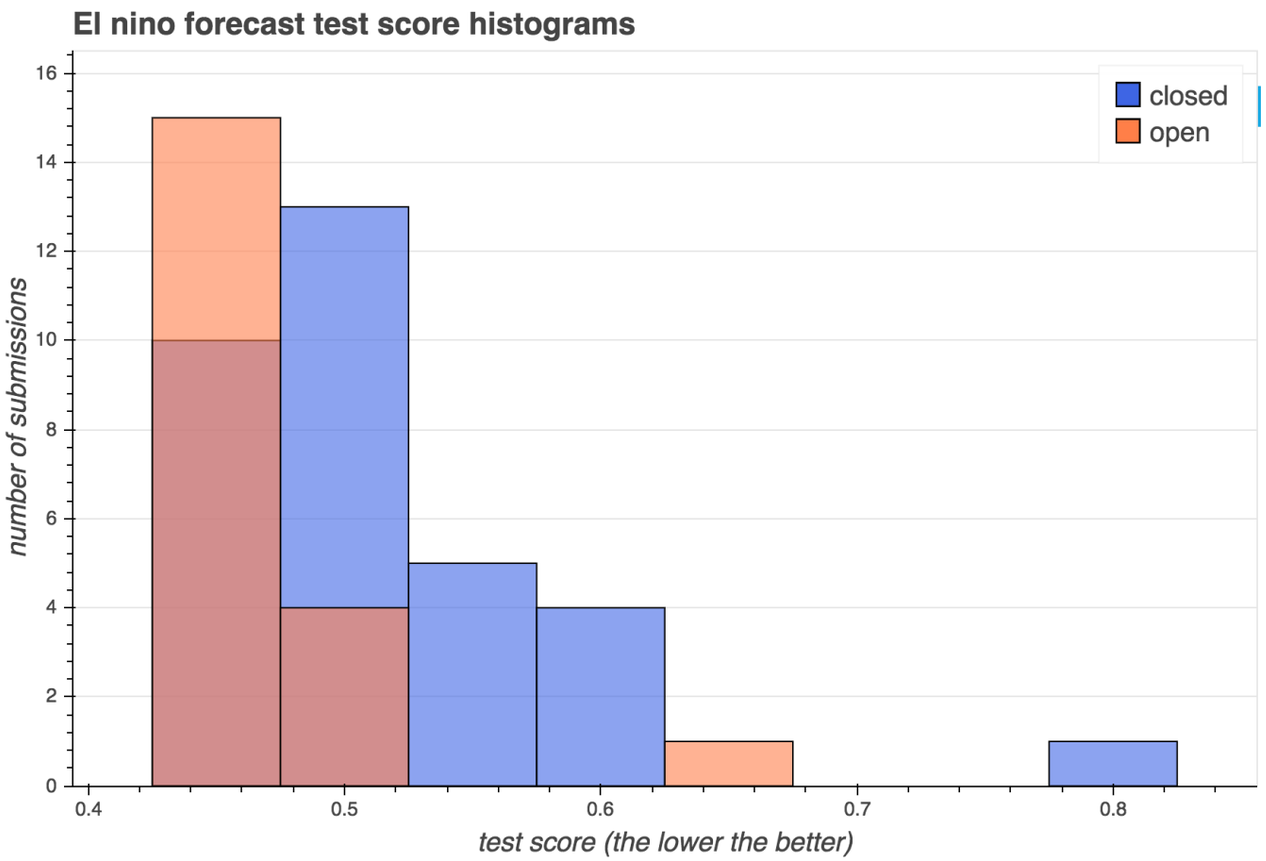
concentration

Analyzing the analysis

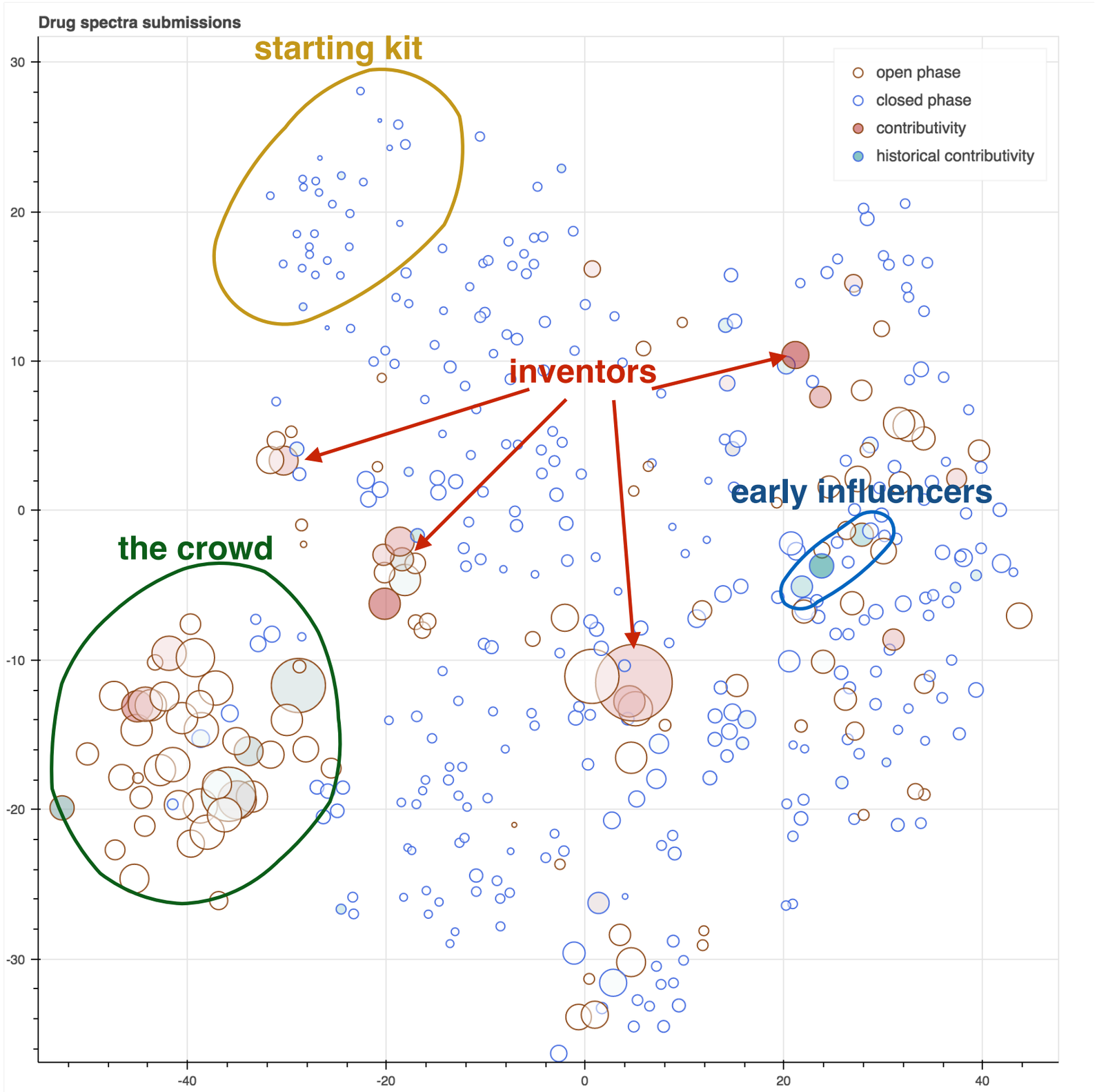
OPEN PHASE LETS PARTICIPANTS CATCH UP



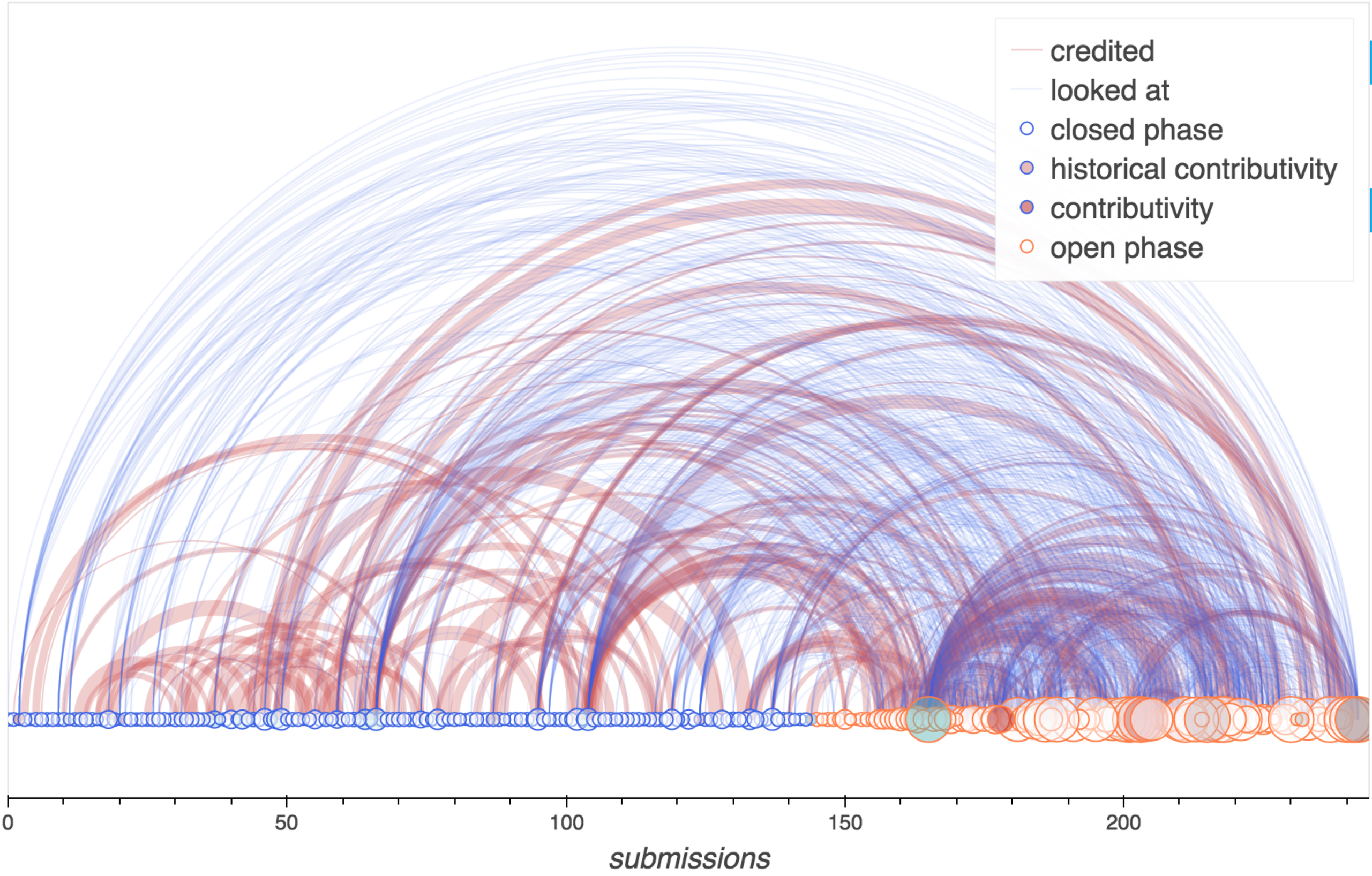




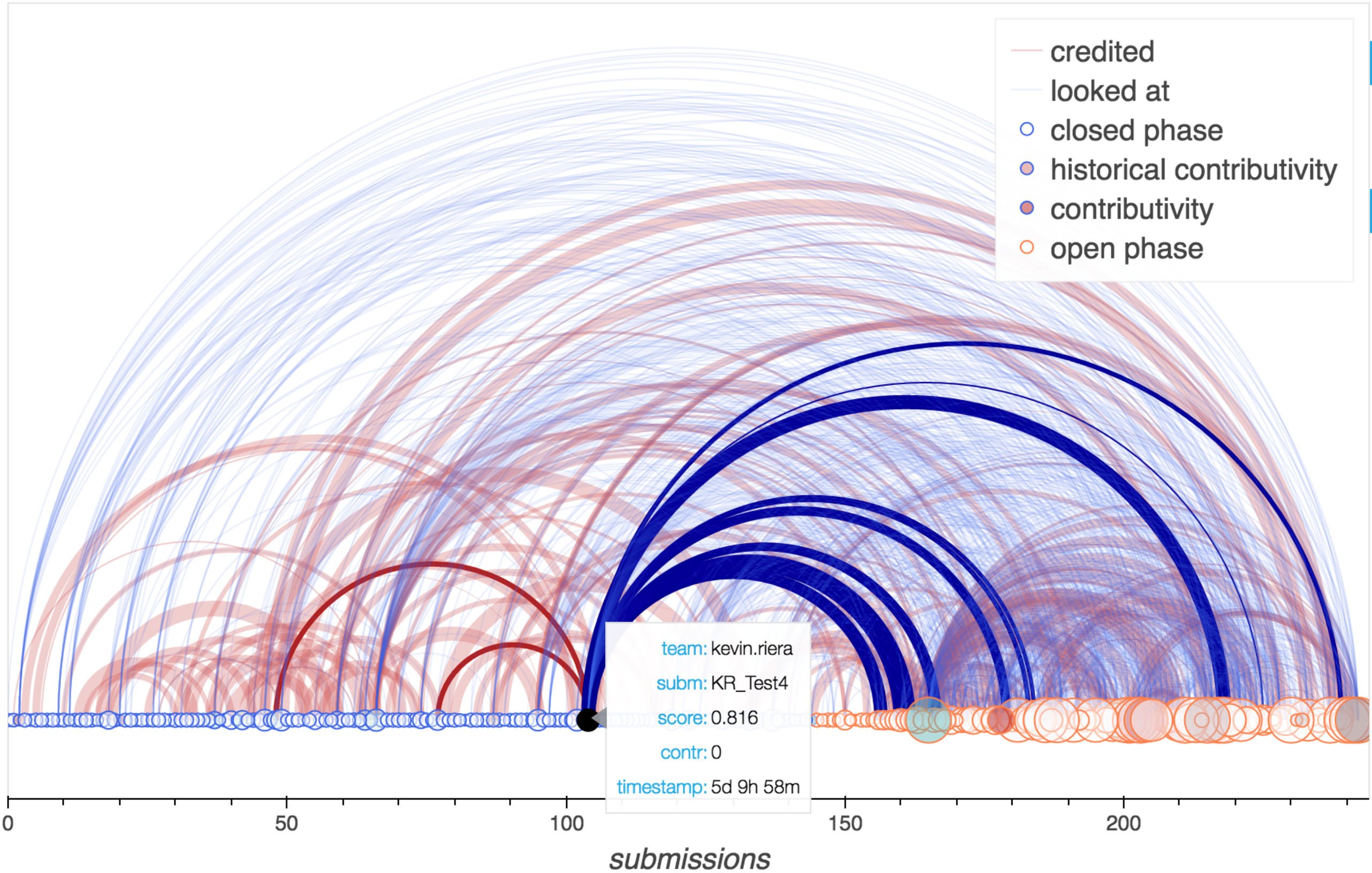
T-SNE ON TEST PREDICTIONS



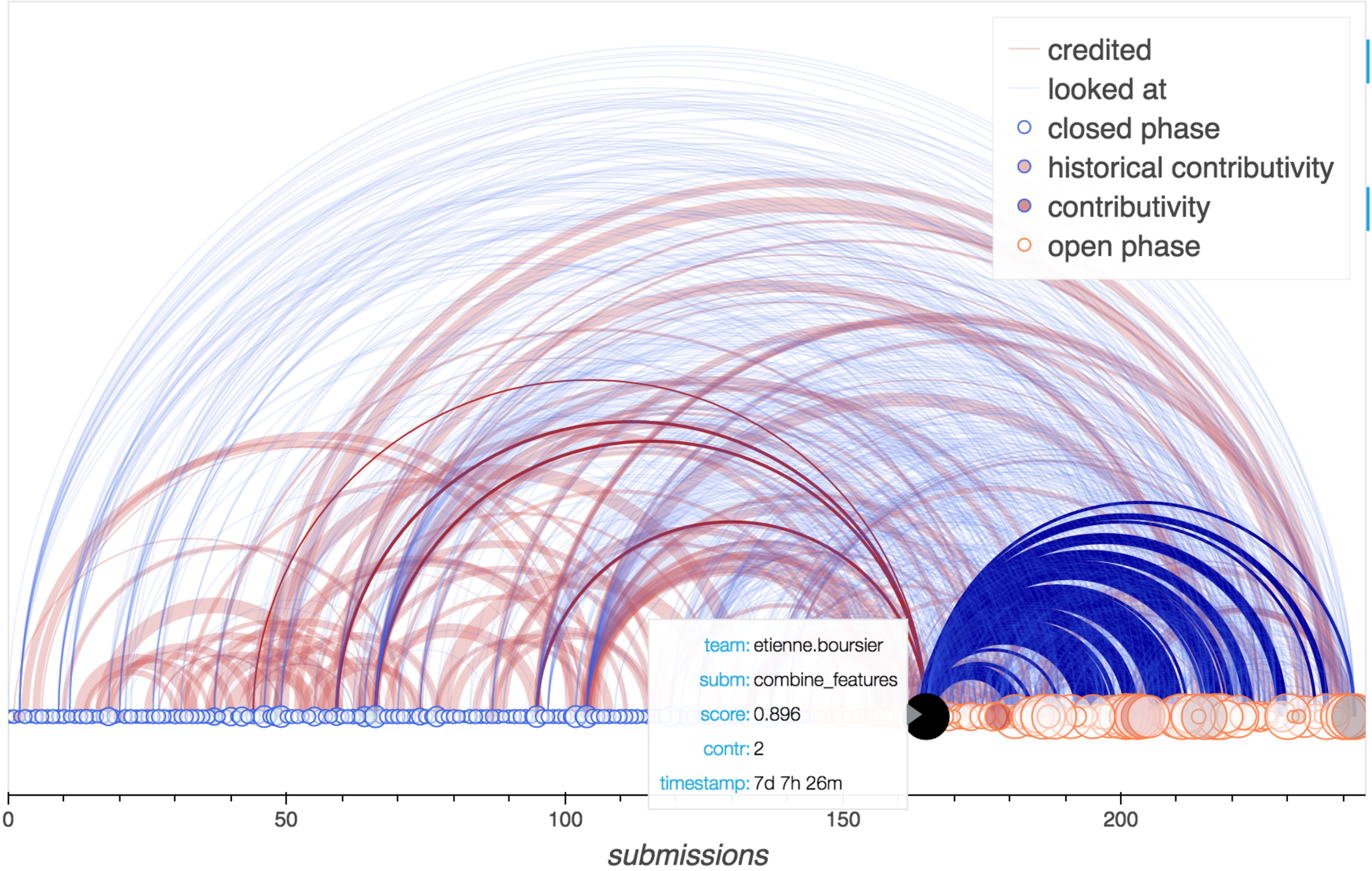
Hep detector anomalies submissions



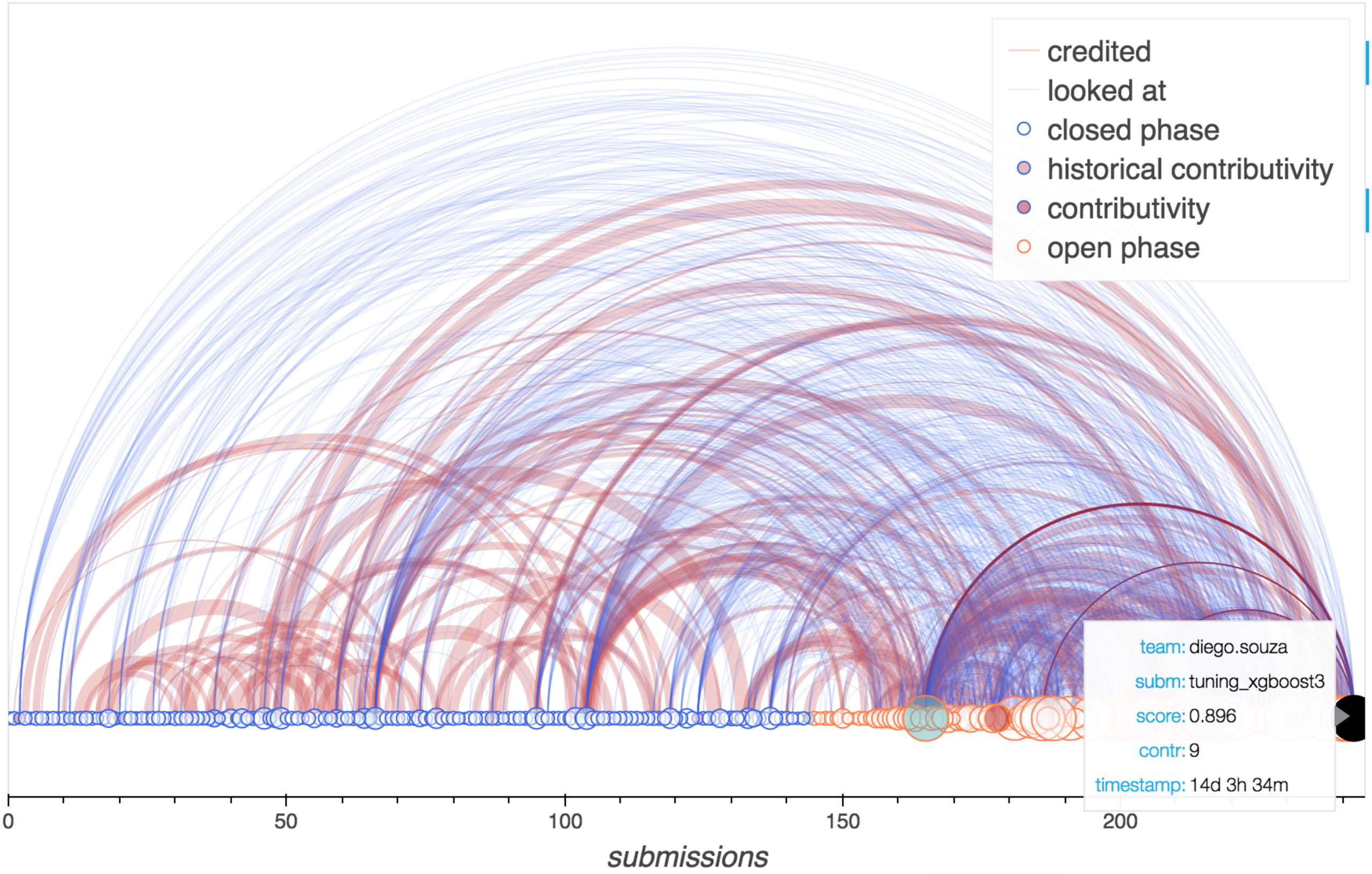
Hep detector anomalies submissions

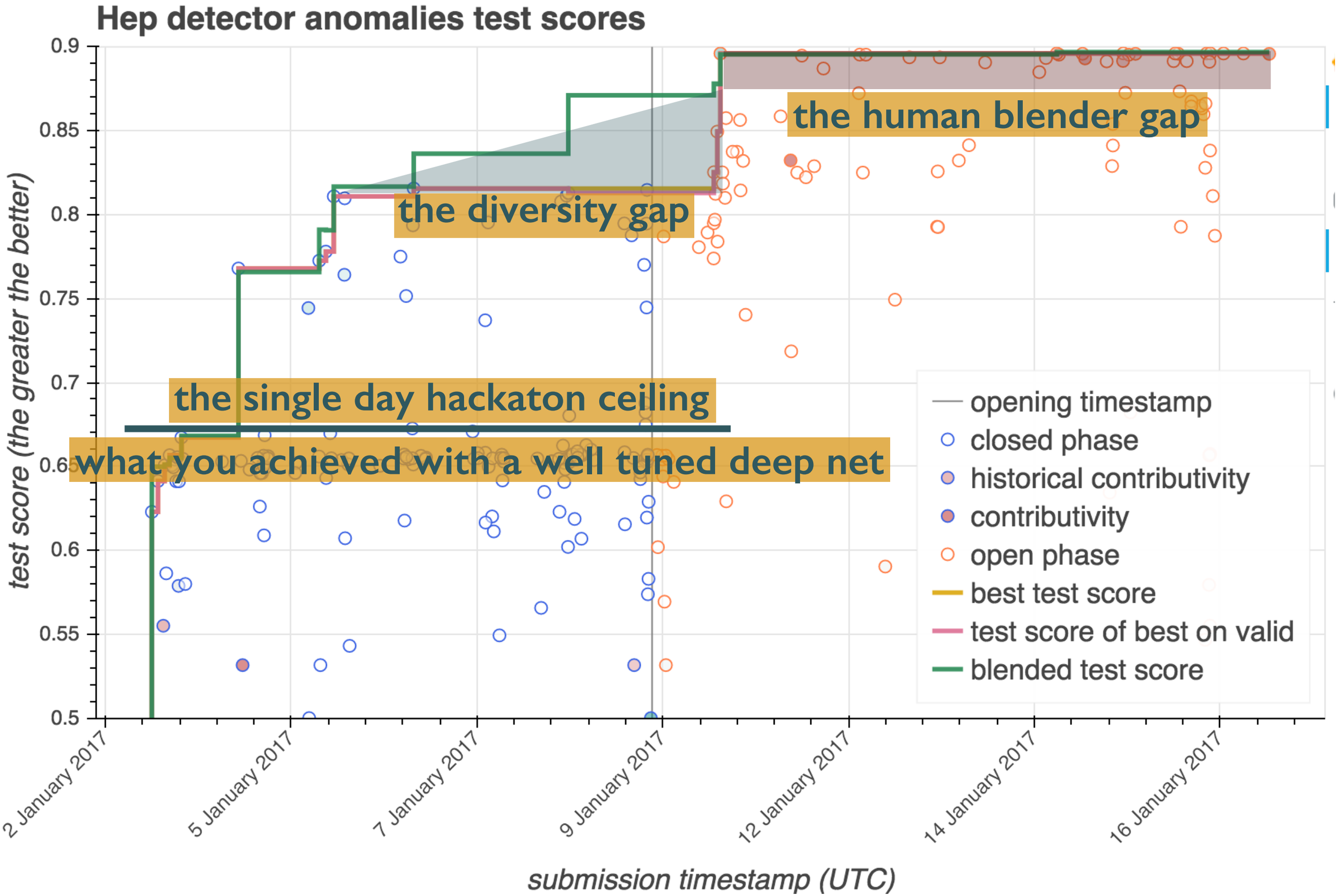


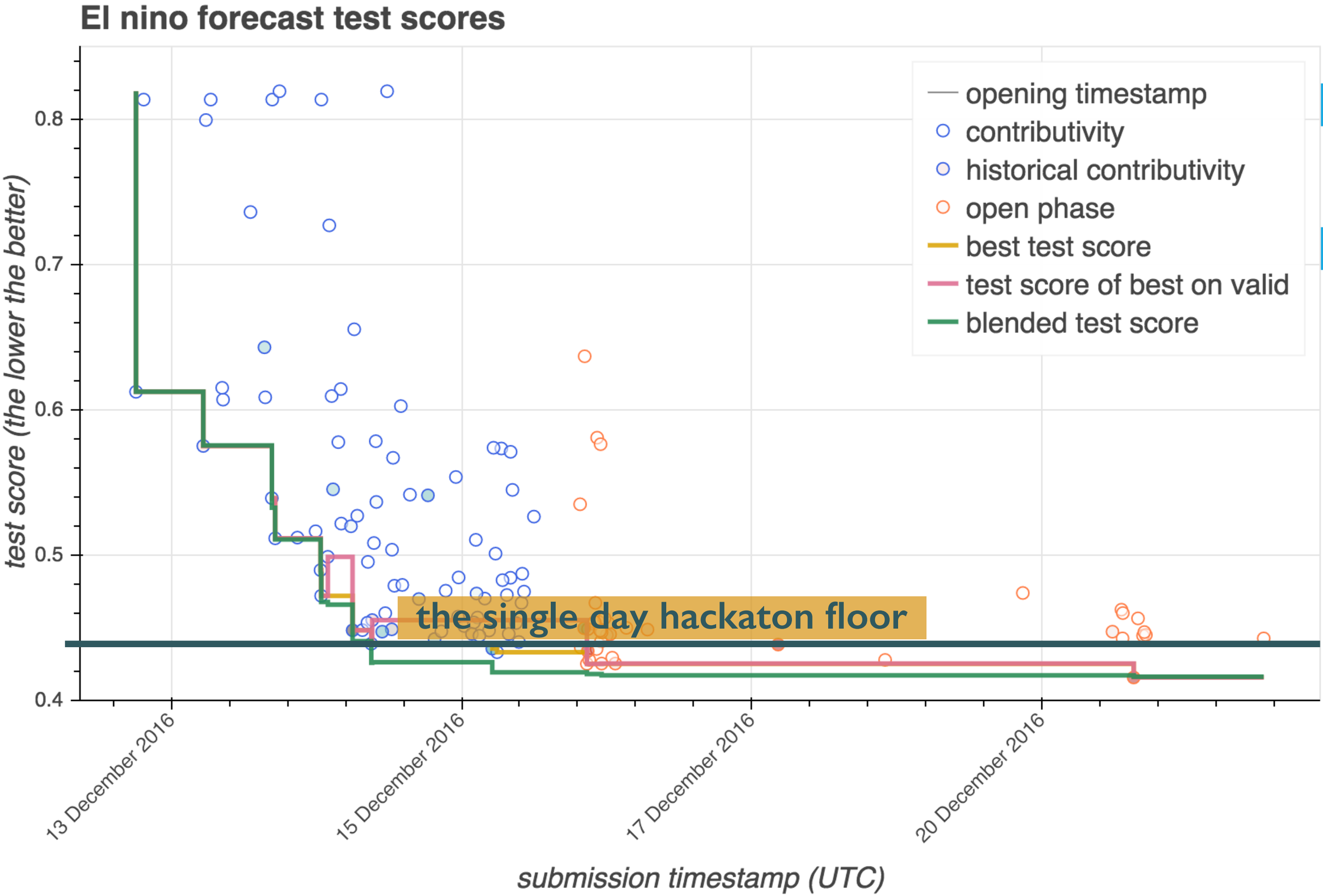
Hep detector anomalies submissions

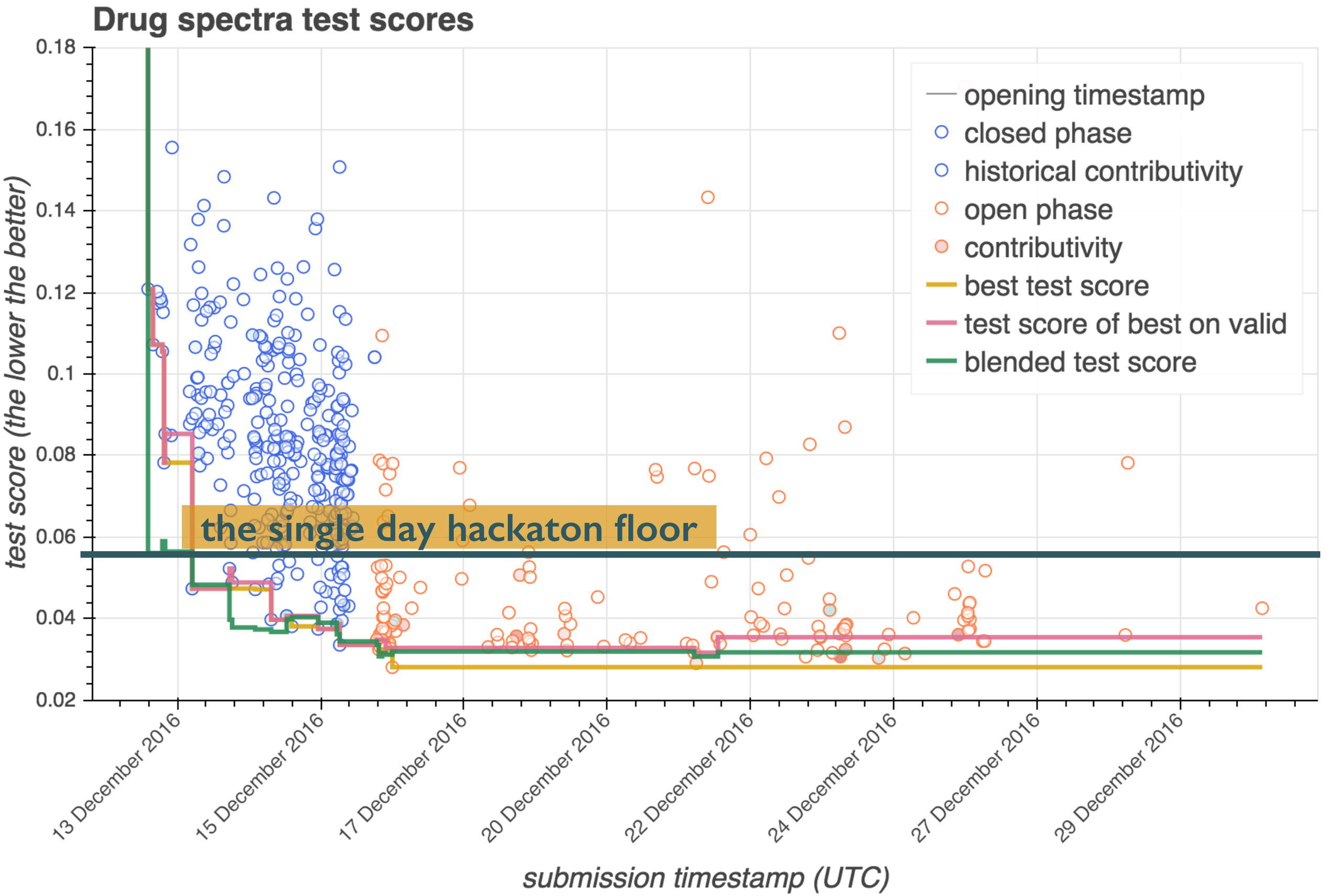


Hep detector anomalies submissions



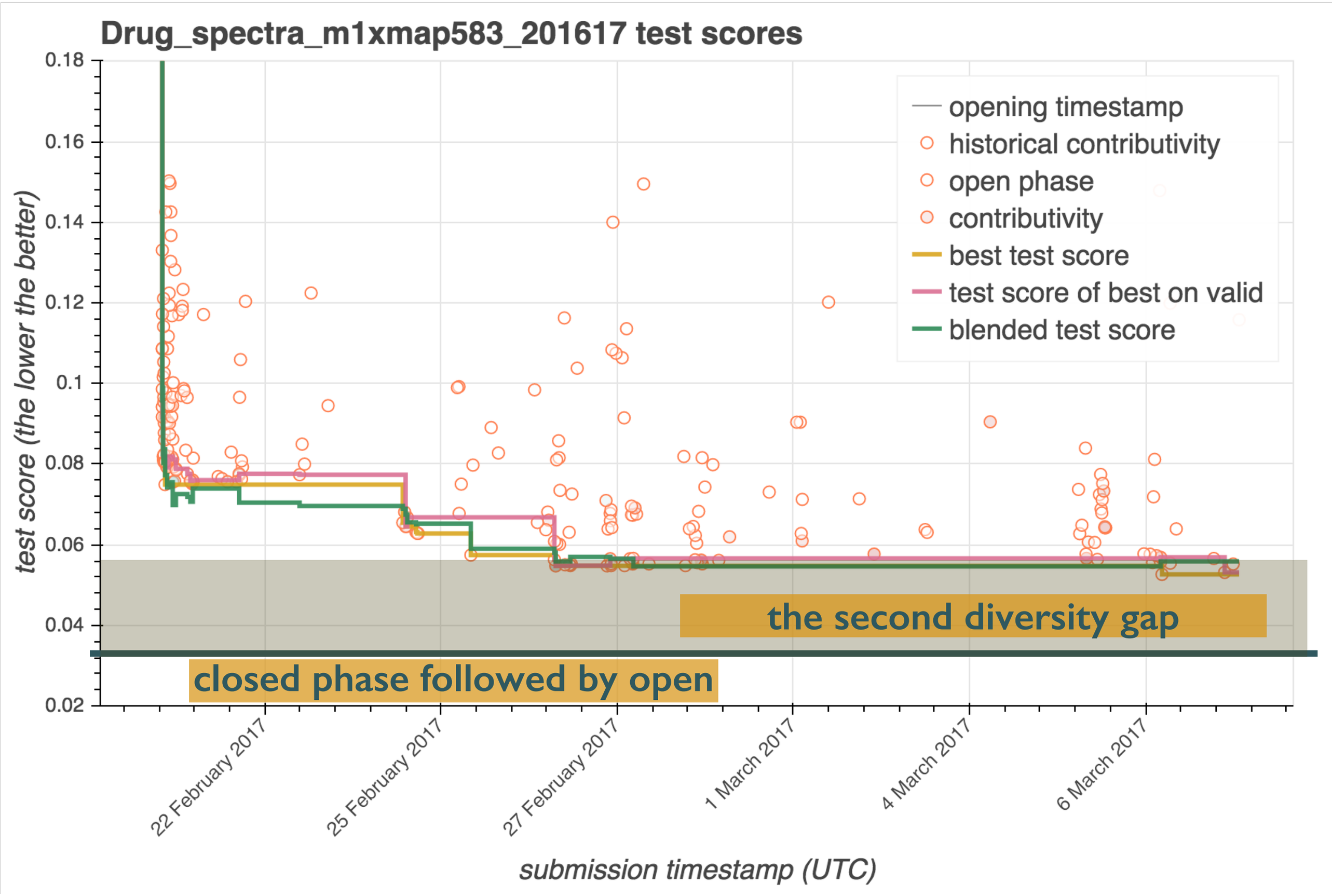

Hep detector anomalies test scores
the human blender gap
the diversity gap
the single day hackaton ceiling
what you achieved with a well tuned deep net
opening timestamp
closed phase
historical contributivity
contributivity
open phase
best test score
test score of best on valid
blended test score
test score (the greater the better)
0.9
0.85
0.8
0.75
0.7
0.6
0.55
0.5
2 January 2017
5 January 2017
7 January 2017
9 January 2017
12 January 2017
14 January 2017
16 January 2017
submission timestamp (UTC)

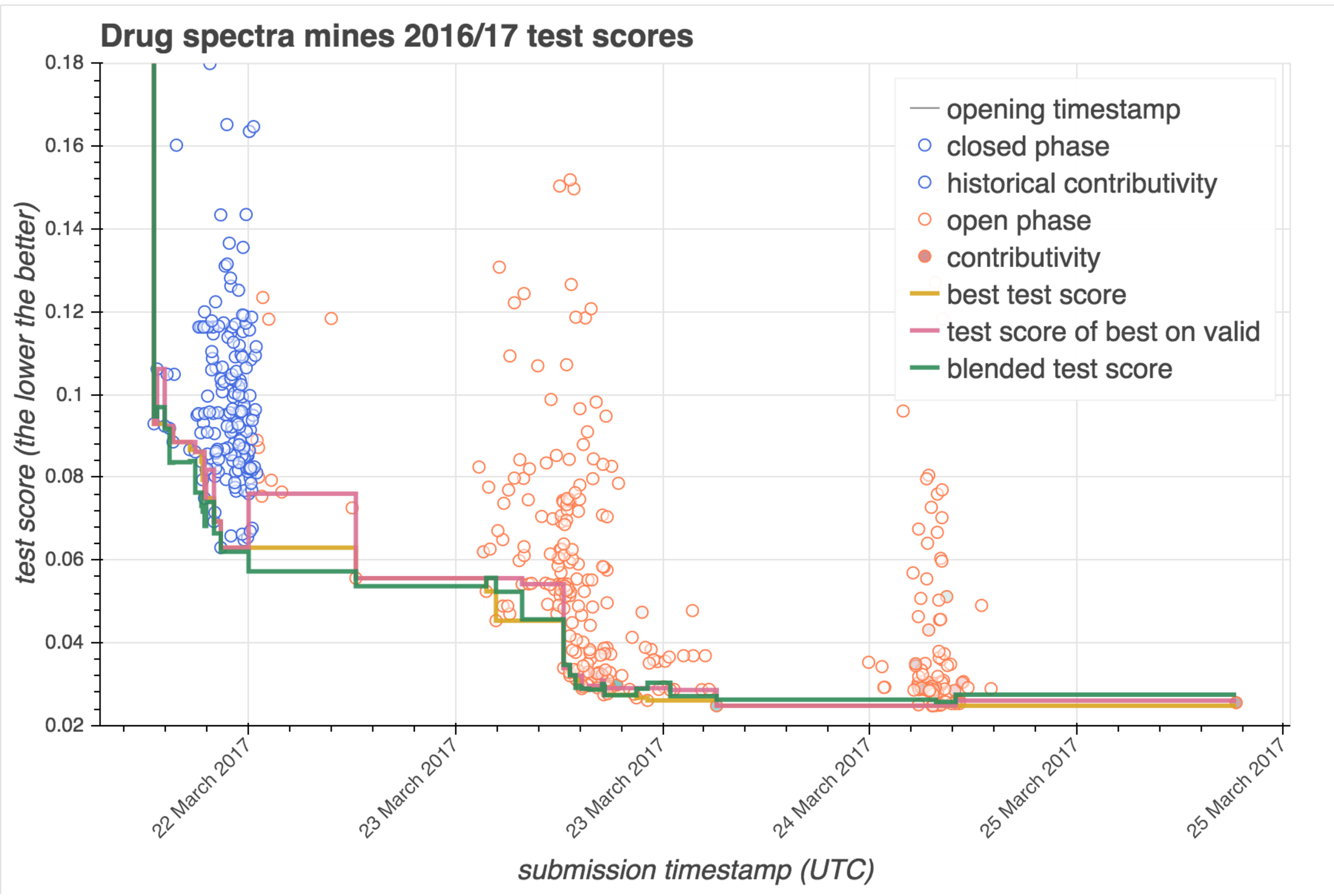

El nino forecast test scores
opening timestamp
contributivity
historical contributivity
open phase
best test score
test score of best on valid
blended test score
0.8
0.7
0.6
0.5
0.4
test score (the lower the better)
the single day hackaton floor
13 December 2016
15 December 2016
17 December 2016
20 December 2016
submission timestamp (UTC)


Drug spectra test scores
0.18
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
test score (the lower the better)
opening timestamp
closed phase
historical contributivity
open phase
contributivity
best test score
test score of best on valid
blended test score
the single day hackaton floor
13 December 2016
15 December 2016
17 December 2016
20 December 2016
22 December 2016
24 December 2016
27 December 2016
29 December 2016
submission timestamp (UTC)

What we learned

- **Open phase helps novice participants to catch up: the goal of teaching!**
  - Sometimes also makes the best and blended score better
- **Human blending often beats machine blending**
- **Human feature engineering easily beats deep learning on some data**
- **Course RAMPs beat single day hackatons** significantly
  - larger number of students?
  - longer RAMPs?
  - novice and master-level students are better than data science researchers?
  - stronger incentives?
  - closed phase preceding an open phase (vs pure open RAMP) helps to create diversity?


Drug_spectra_m1xmap583_201617 test scores
0.18
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
test score (the lower the better)
opening timestamp
historical contributivity
open phase
contributivity
best test score
test score of best on valid
blended test score
the second diversity gap
closed phase followed by open
22 February 2017
25 February 2017
27 February 2017
1 March 2017
4 March 2017
6 March 2017
submission timestamp (UTC)


Drug spectra mines 2016/17 test scores
0.18
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
test score (the lower the better)
opening timestamp
closed phase
historical contributivity
open phase
contributivity
best test score
test score of best on valid
blended test score
22 March 2017
23 March 2017
23 March 2017
24 March 2017
25 March 2017
25 March 2017
submission timestamp (UTC)

Classifying and quantifying monoclonal antibody preparations for cancer therapy using machine learning

Laetitia Le [ab], Camille Marini [ce], Alexandre Gramfort [cfg], David Nguyen [a], Mehdi Cherti [ch], Sana Tfaili [b], Ali Tfayli [b], Arlette Baillet-Guffroy [b], Eric Caudron [ab], Balázs Kégl [ch]

[a] European Georges Pompidou Hospital (AP-HP), Pharmacy department, Paris, France
[b] Lip(Sys) Chimi Analytique Pharmaceutique, Univ. Paris-Sud, Universit Paris Saclay, F92290 Chatenay-Malabry, France (EA4041 Groupe de Chimie Analytique de Paris Sud)
[c] Center of Data Science, Université Paris-Saclay
[d] Université Paris-Sud
[e] CMAP, Ecole Polytechnique, Palaiseau, France
[f] INRIA, Parietal team, Saclay, France
[g] LTCI, Télécom ParisTech
[h] LAL, CNRS, France

THE RAMP TOOL

A prototyping tool for collaborative development of data science workflows

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

WHAT'S NEXT

- **More RAMPs**, sign up at **http://www.ramp.studio** if interested

- Tools to **manage the crowd**, redirect group attention, etc.

- **Human - AI interaction:** "hyperopt this module for me, would you?"

- Streamlining deployment