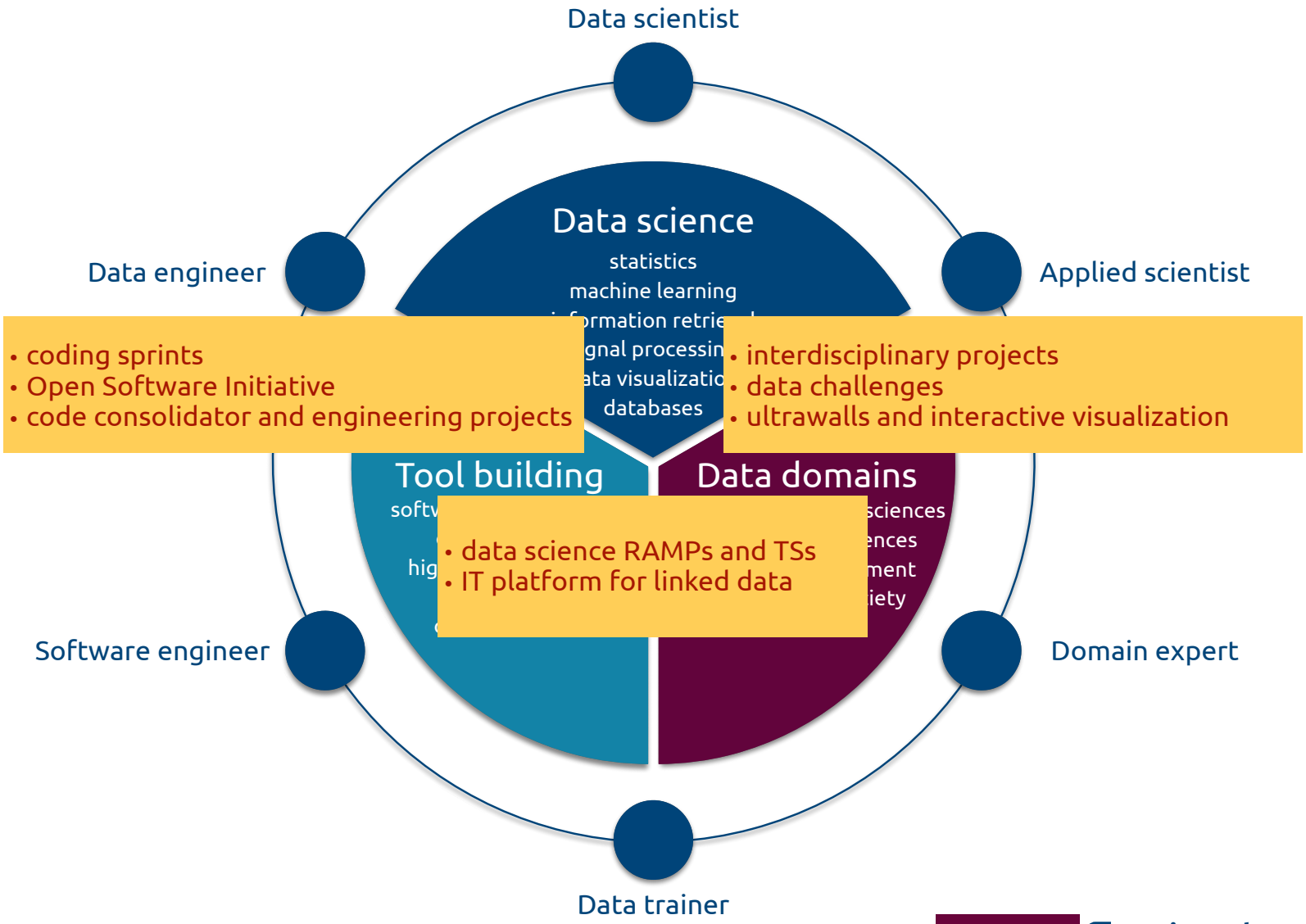
Rapid Analytics & Model Prototyping (RAMP)

goals and lessons

Yetkin Yilmaz
on behalf of
AppStat - LAL

CDS: A SET OF INNOVATIVE TOOLS AND PROCESSES TO CONNECT DATA SCIENCE AND DOMAIN SCIENCE COMMUNITIES

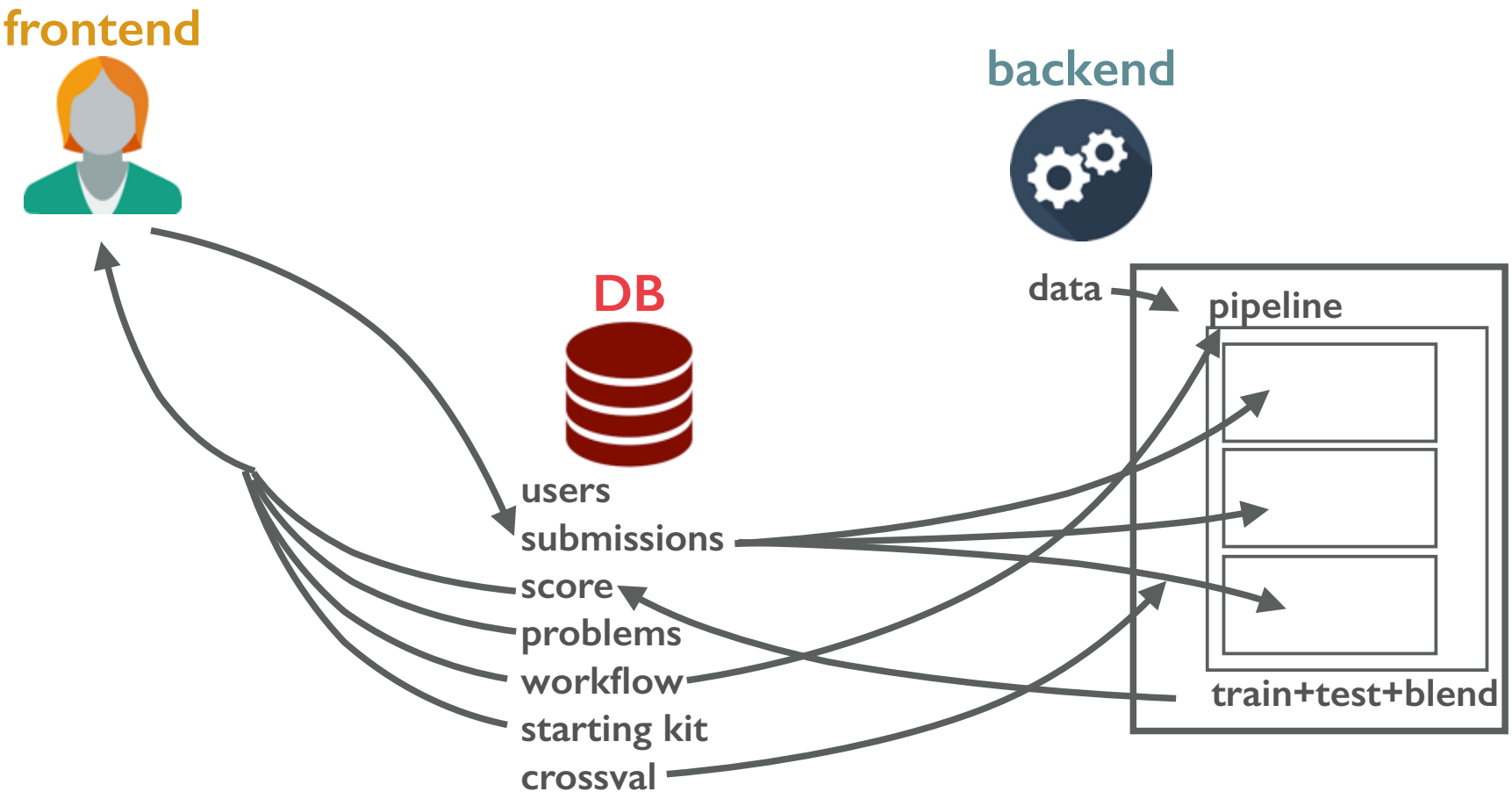


Paris-Saclay Center for Data Science

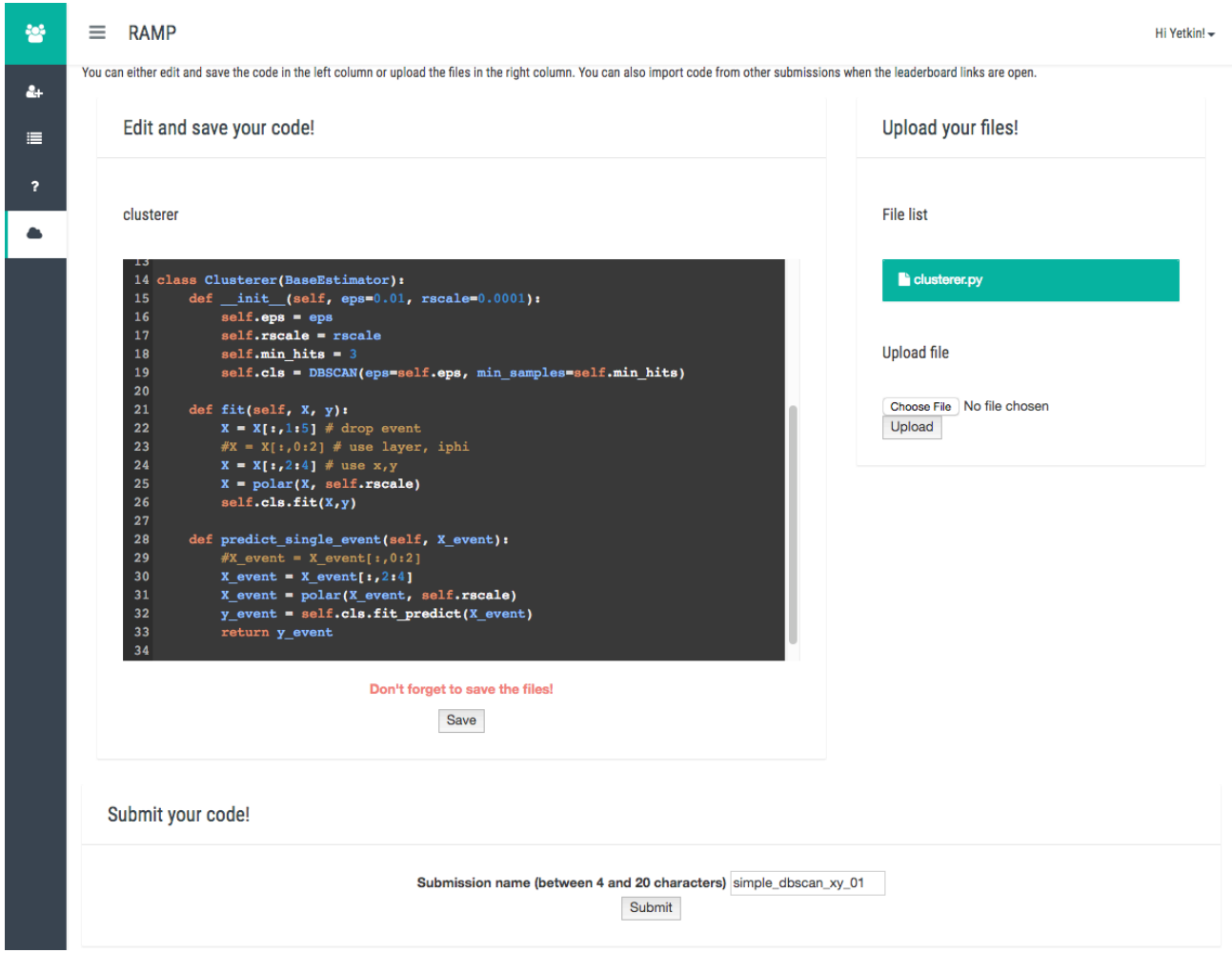
Why RAMP?

- Interface domain sciences and data scientists
- Streamline the innovation process
- Provide computing resources
- Modularize analysis workflow
- Collaborative development

Modularized workflow



Submission


RAMP
Hi Yetkin!
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!
clusterer
class Clusterer(BaseEstimator):
def __init__(self, eps=0.01, rscale=0.0001):
self.eps = eps
self.rscale = rscale
self.min_hits = 3
self.cls = DBSCAN(eps=self.eps, min_samples=self.min_hits)
def fit(self, X, y):
X = X[:,1:5] # drop event
#X = X[:,0:2] # use layer, iphi
X = X[:,2:4] # use x,y
X = polar(X, self.rscale)
self.cls.fit(X,y)
def predict_single_event(self, X_event):
#X_event = X_event[:,0:2]
X_event = X_event[:,2:4]
X_event = polar(X_event, self.rscale)
y_event = self.cls.fit_predict(X_event)
return y_event
Don't forget to save the files!
Save
Upload your files!
File list
clusterer.py
Upload file
Choose File
No file chosen
Upload
Submit your code!
Submission name (between 4 and 20 characters)
simple_dbscan_xy_01
Submit

Contrasts to other data challenges

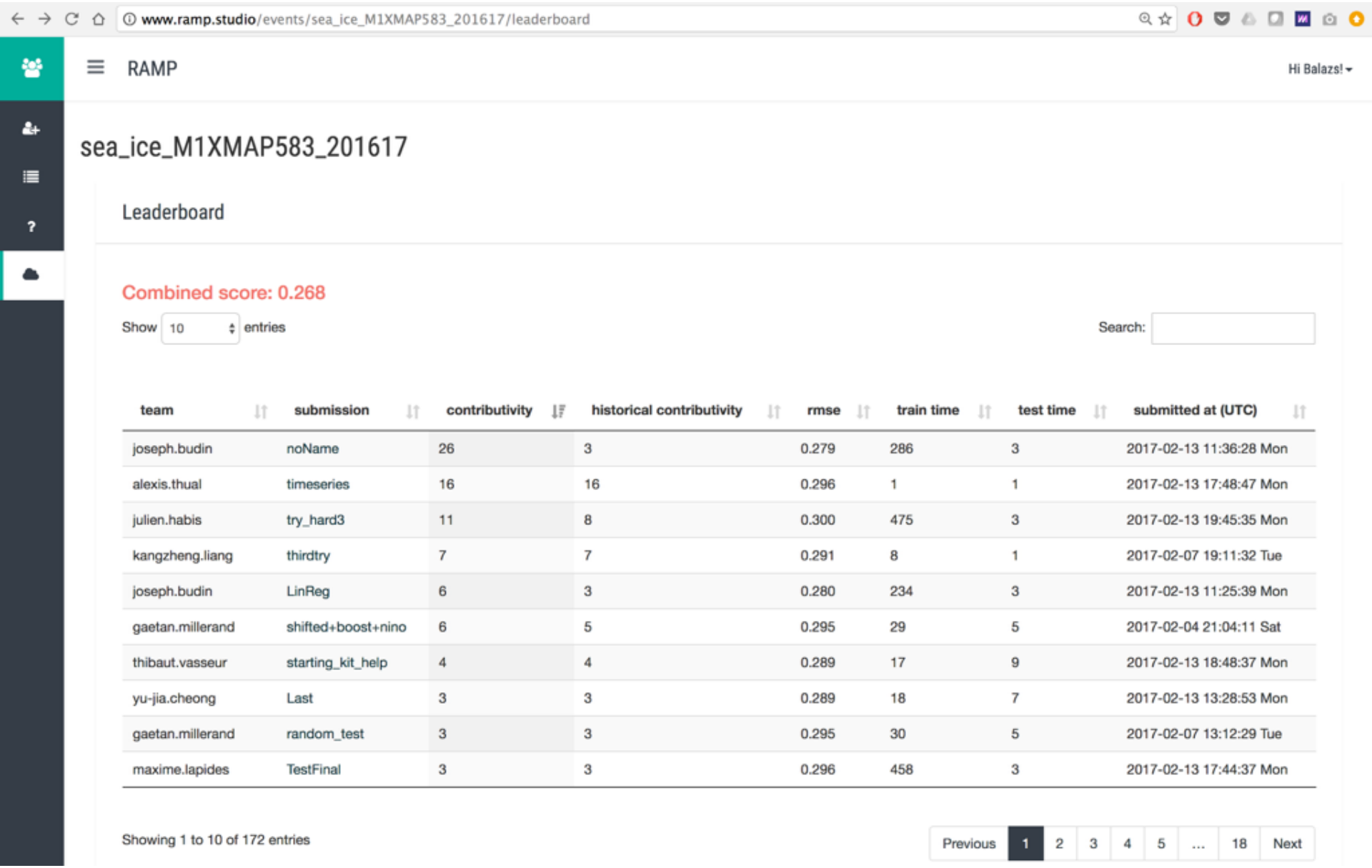
- Submit code instead of solution
- Build on each other's code
- Development process is available – intermediate ideas visible

Data challenges with RAMP

- Single day hackathons :
  - 20-50 participants
  - open leaderboard

- 1-3 week course challenges :
  - up to 150 students
  - limited (1-3) submissions per day
  - closed leaderboard phase for individual development
  - followed by open leaderboard for collaborative development
  - discussion using Slack

- 500+ users, 3000+ models


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

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.

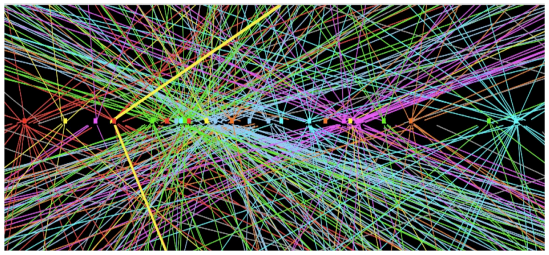
submission	credit
self credit	10
el_nino/raphael.berdugo/end_submit	0
el_nino/domitille.coulomb/Third Trial	40
el_nino/bkabid/second_attempt	0
el_nino/domitille.coulomb/Second Trial	50
el_nino/domitille.coulomb/First Trial	0
el_nino/domitille.coulomb/starting_kit	0

Submit

Diverse range of scientific fields

- Epidemium cancer mortality rate prediction
- Drug classification and concentration estimation from Raman spectra
- Pollenating insects
- El Nino forecast
- Northern hemisphere sea ice prediction
- Number of air passengers prediction
- Detecting anomalies in the LHC ATLAS detector
- Particle tracking in the LHC ATLAS detector

Automatized Data-Quality Monioring



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

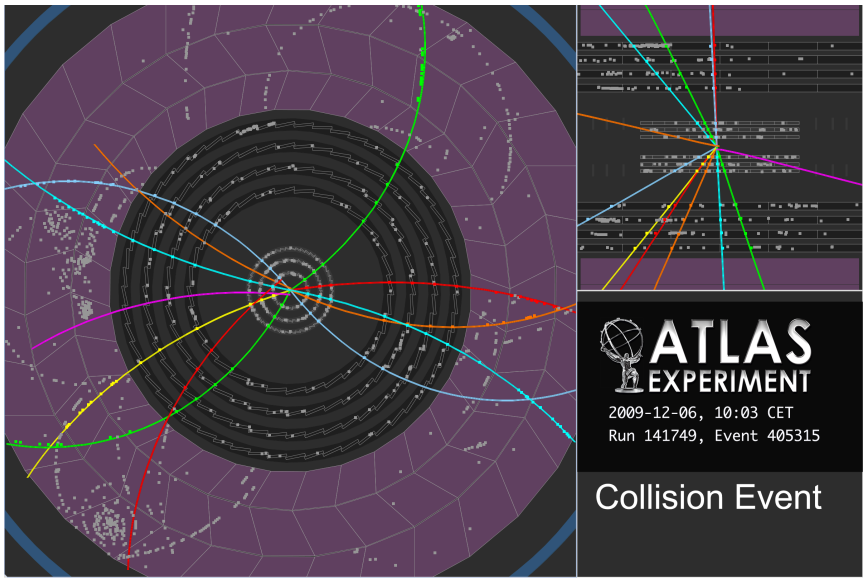
classifier

?

correct
(isSkewed = 0)

anomaly
(isSkewed = 1)

Tracking in 2D



http://atlas.web.cern.ch/Atlas/public/EVTDISPLAY/events.html

Assign a predicted cluster id

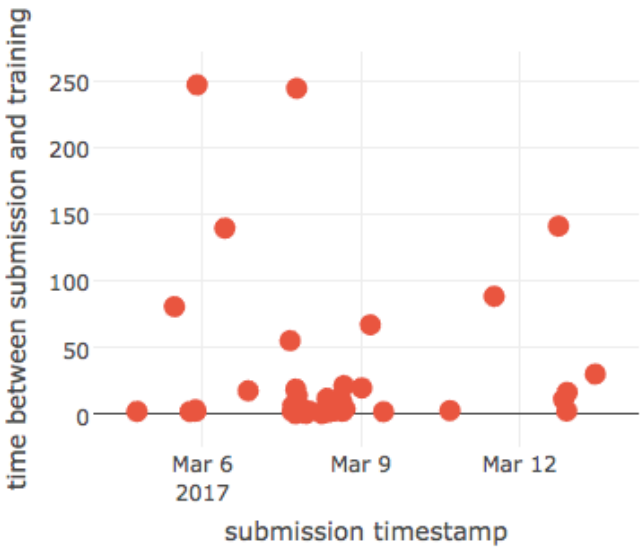
```
In [10]: df.sort_values(by=['event_id','cluster_id'])
```

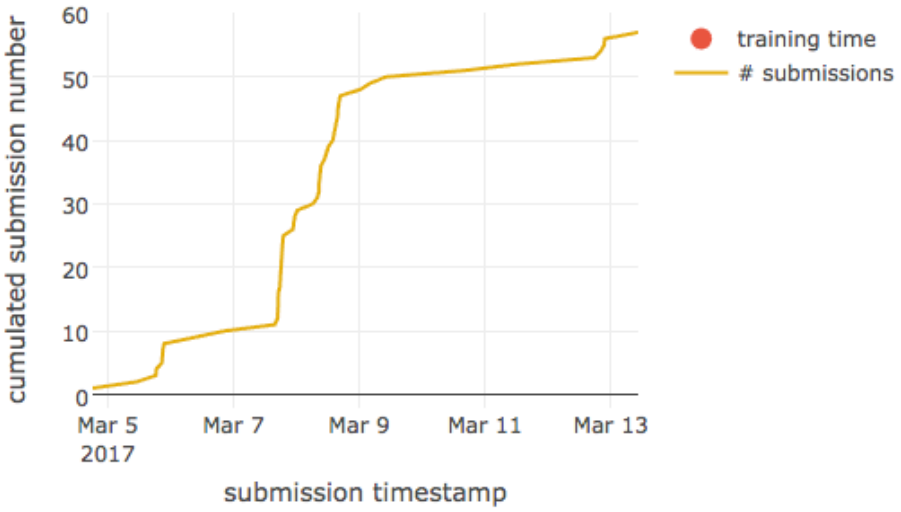
Out[10]:

	event_id	cluster_id	layer	iphi	x	y
12	3	0	5	30498	-328.776740	-236.497051
14	3	0	1	12805	-68.954047	-49.702510
21	3	0	7	57478	-615.753064	-448.878786
26	3	0	4	40796	-220.267647	-157.870718
27	3	0	6	42351	-455.354264	-329.388060
29	3	0	0	5891	-31.407037	-23.121376
31	3	0	3	32062	-173.165994	-124.026363
34	3	0	2	23333	-125.991539	-90.283620
54	3	0	8	75522	-805.369064	-592.773709
1	3	1	5	37216	-47.614439	-402.191329
2	3	1	0	7181	-4.253919	-38.767308
11	3	1	6	51612	-67.634707	-557.915358
28	3	1	4	49832	-31.193083	-269.198796
30	3	1	8	91683	-127.898414	-991.787273
42	3	1	1	15644	-9.436681	-84.474547

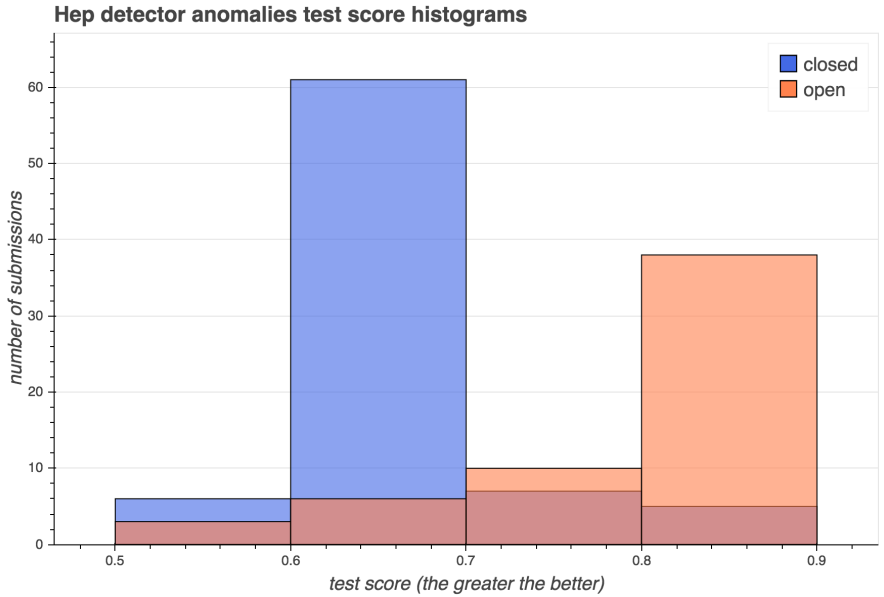
Submission statistics

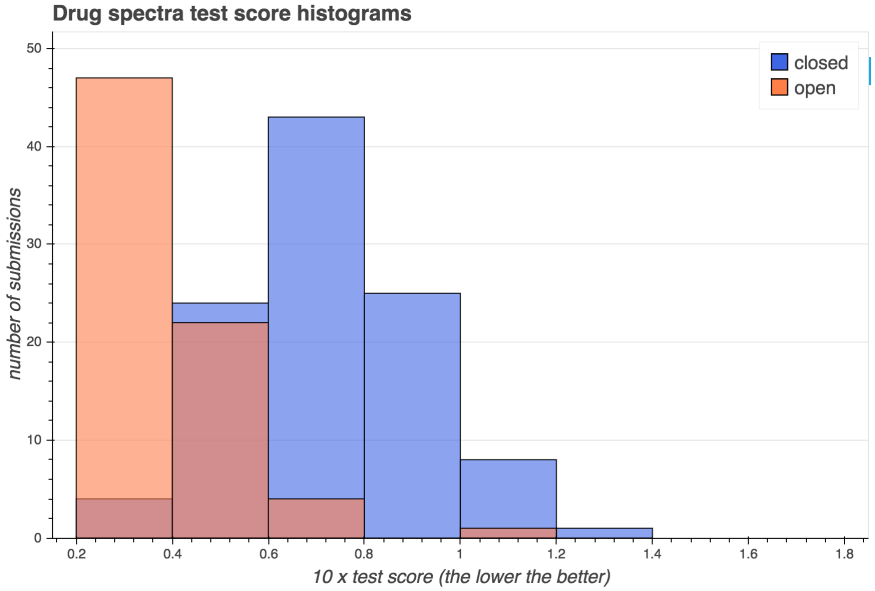
2D Tracking Hackathon

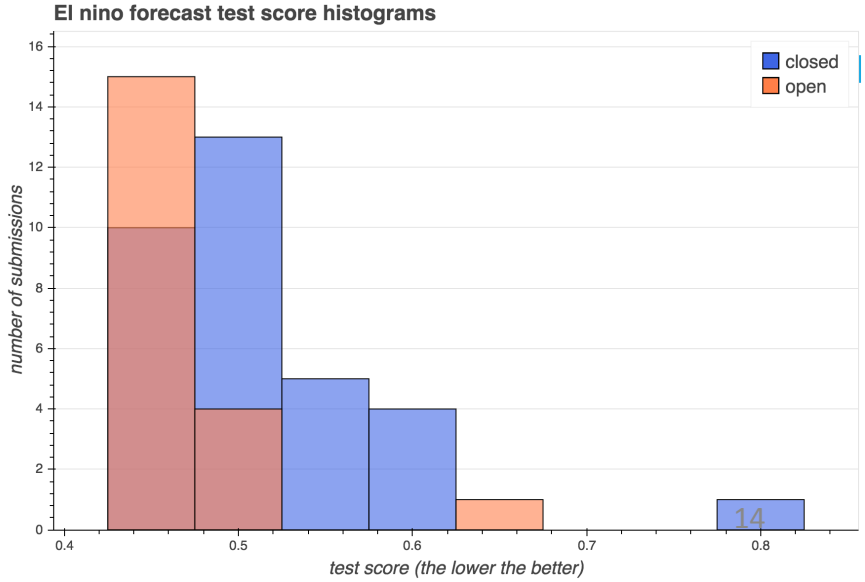




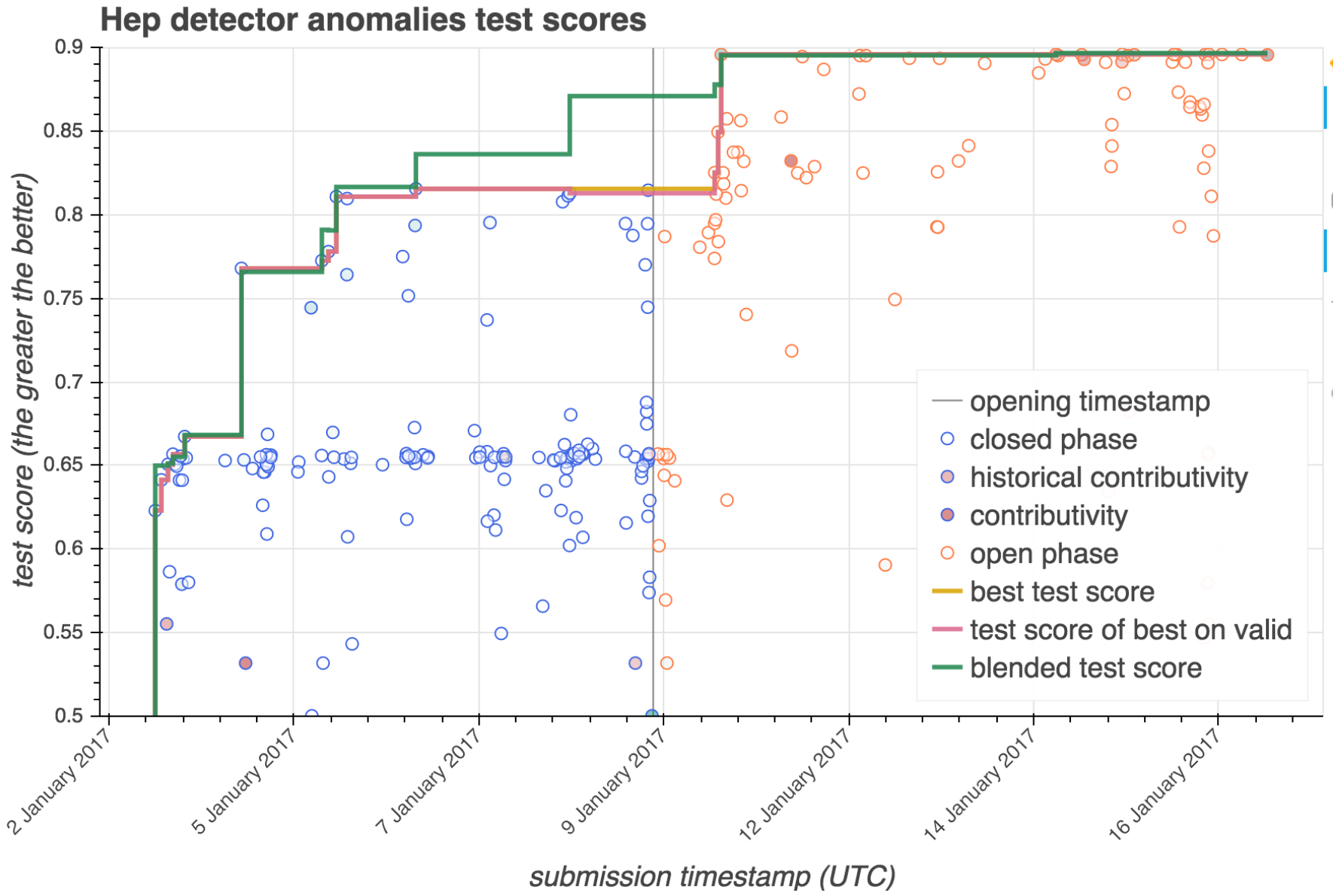
Score statistics



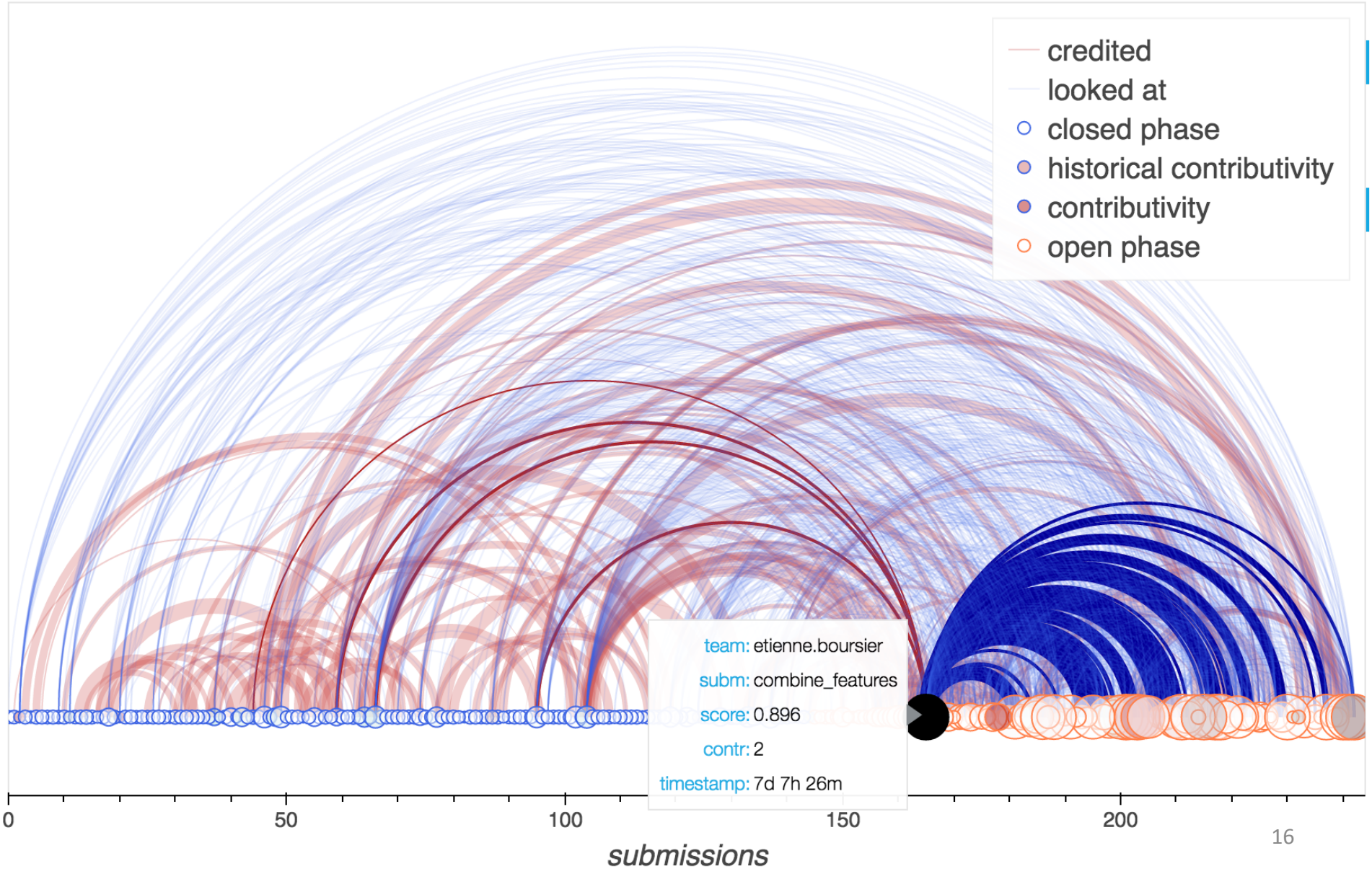




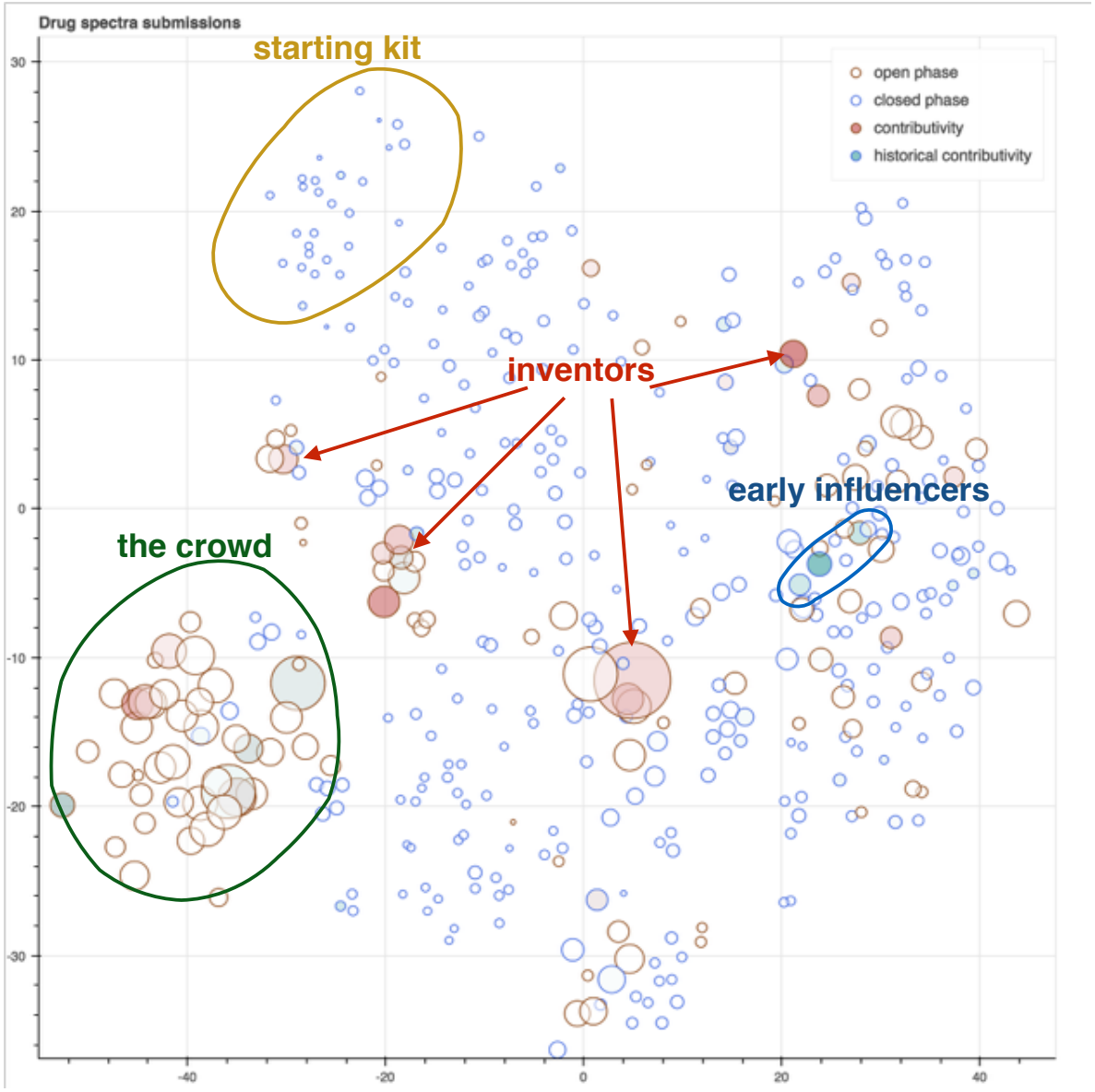
Score progress



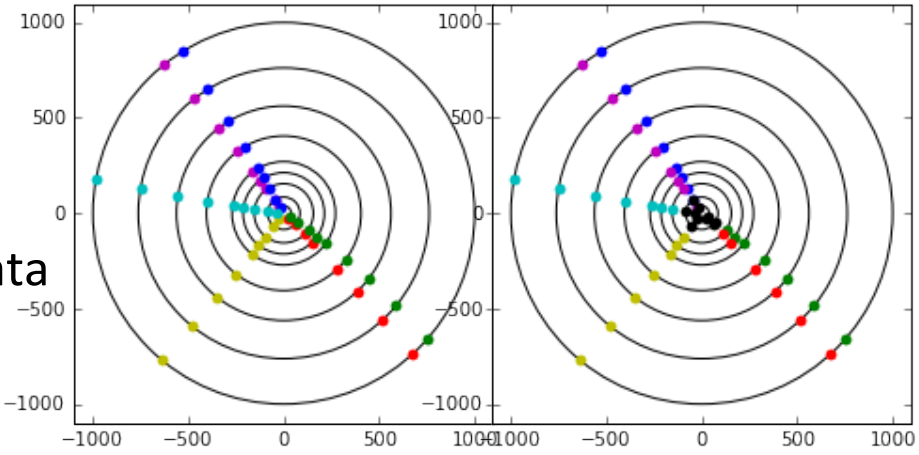
Hep detector anomalies submissions



Innovation analysis



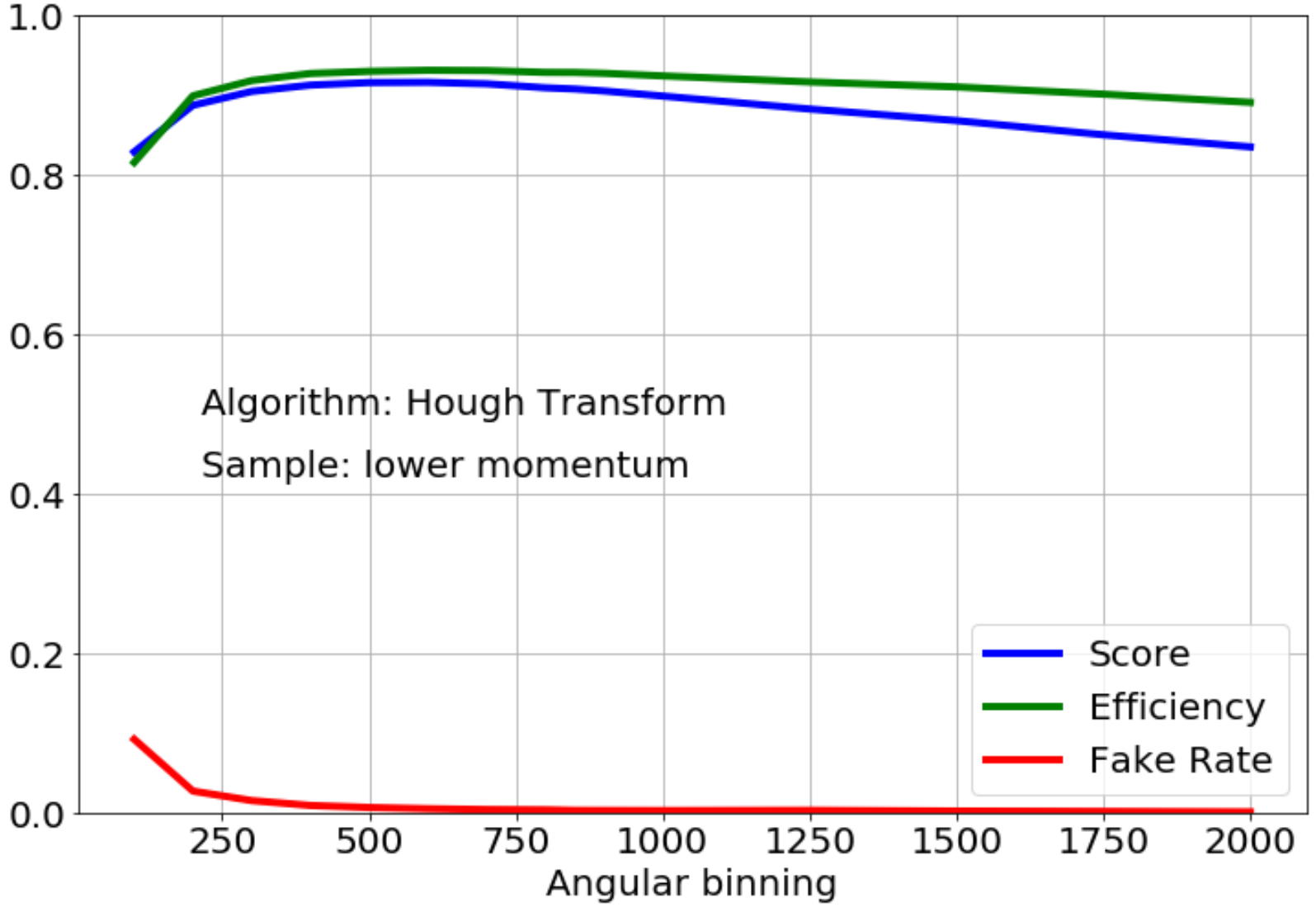
Types of scoring



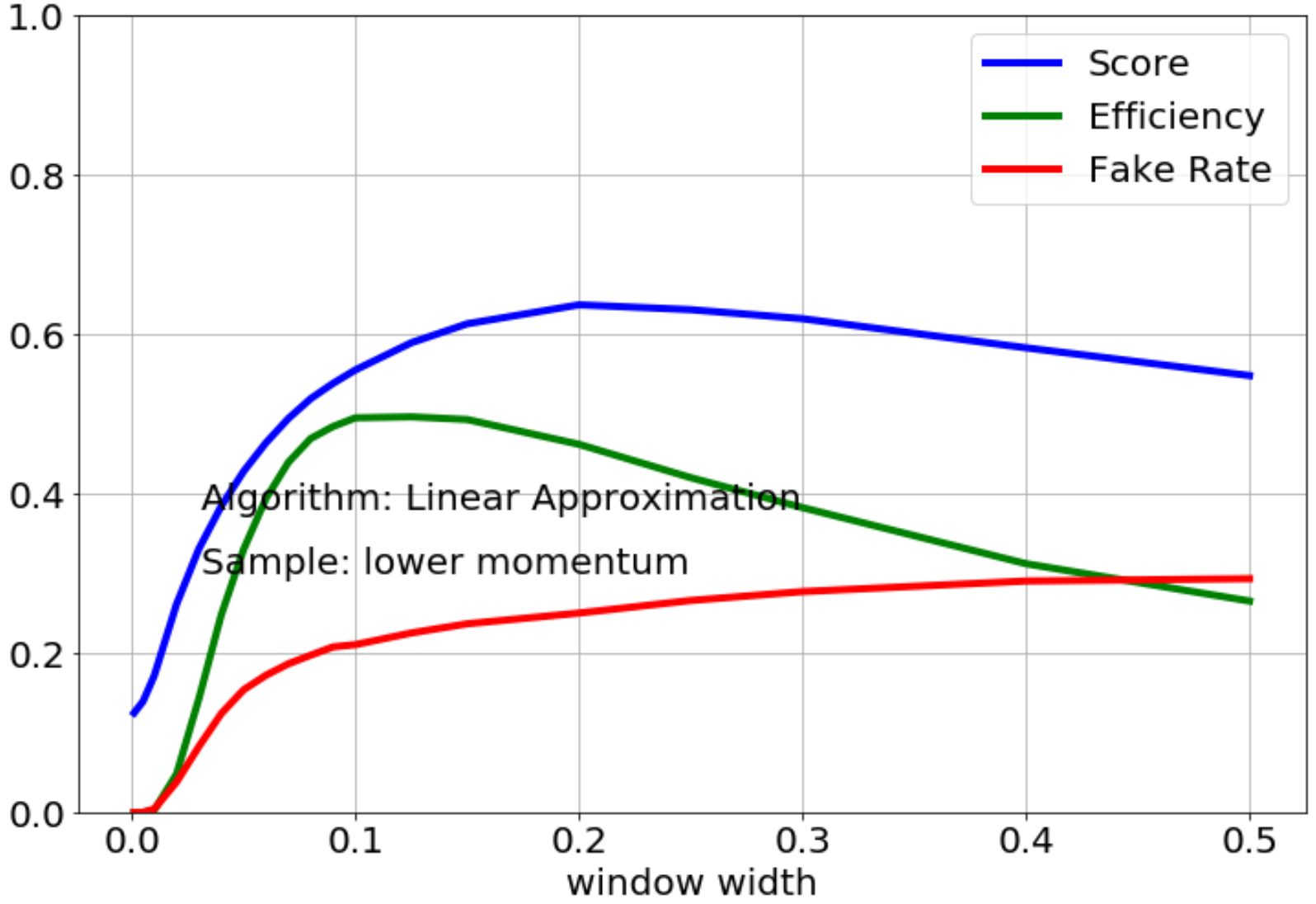
- Classification score
  - ~ fraction of correctly identified data
- Regression score
  - ~ distance from the correct values
- Clustering score (example in HEP_tracking context)
  - ~ fraction of identified objects? (efficiency and fake)
  - ~ distance of derived parameters from the correct ones? (momentum resolution)
  - ~ how well the data is segmented? (the choice in HEP_tracking)
  - ~ correct labeling? (labels arbitrary)
  - ~ overlaps? (neglected)
  - ~ how to combine solutions, assign contributivity (not solved yet)

Defining the score defines a significant part of the problem!

Score vs qualities



Score vs qualities



Rapid Analytics and Model Prototyping

for more, go to :
http://www.ramp.studio