Choosing hosting platforms

PLAsTiCC Workshop, July/2017

Emille E. O. Ishida

Laboratoire de Physique de Clermont - Université Clermont-Auvergne
Clermont Ferrand, France

Goals:

1)Increase the participation of non-astronomers

2)Facilitate posterior usage of results

3)Answer multiple questions

Option 1: Kaggle

www.kaggle.com

How it works:

- proposers work with Kaggle team to check the data (~ 6 weeks)
- participants submit **responses**
- metric is calculated by the back-end for each entry
- there is a kernel tool, with pre-loaded packages in R and Python which users **can** use
- for scientific challenges, the code must be *open source*
- prizes are not mandatory, but advisable
- it is possible to make Amazon cores available to the users
  (Kaggle team might be willing to provide it without extra charge if they think the challenge is of their interest)

Option 1: Kaggle

www.kaggle.com

How it works:

- proposers work with Kaggle team to check the data (~ 6 weeks)
- participants submit **responses**
- metric is calculated by the back-end for each entry
- there is a kernel tool, with pre-loaded packages in R and Python which users **can** use
- for scientific challenges, the code must be *open source*
- prizes are not mandatory, but advisable
- it is possible to make Amazon cores available to the users
  (Kaggle team might be willing to provide it without extra charge if they think the challenge is of their interest)

Pros:

- High visibility among non-astronomers
- already built community
- platform has already been tested in many different situations (expertise)
- Kaggle team provides support: data preparation, development of Q&A, metric
- integrated forum to talk with the participants
- participants can use any language they want

Option 1: Kaggle

www.kaggle.com

How it works:

- proposers work with Kaggle team to check the data (~ 6 weeks)
- participants submit **responses**
- metric is calculated by the back-end for each entry
- there is a kernel tool, with pre-loaded packages in R and Python which users **can** use
- for scientific challenges, the code must be *open source*
- prizes are not mandatory, but advisable
- it is possible to make Amazon cores available to the users
  (Kaggle team might be willing to provide it without extra charge if they think the challenge is of their interest)

Pros:

- High visibility among non-astronomers
- already built community
- platform has already been tested in many different situations (expertise)
- Kaggle team provides support: data preparation, development of Q&A, metric
- integrated forum to talk with the participants
- participants can use any language they want

Cons:

- using the provided kernels is **not obligatory**
- participants can use any language they want
- 1 challenge → 1 question
- good ideas might go unnoticed if they are not top listed

Option 1: Kaggle

www.kaggle.com

How it works:

- proposers work with Kaggle team to check the data (~ 6 weeks)
- participants submit **responses**
- metric is calculated by the back-end for each entry
- there is a kernel tool, with pre-loaded packages in R and Python which users **can** use
- for scientific challenges, the code must be *open source*
- prizes are not mandatory, but advisable
- it is possible to make Amazon cores available to the users
  (Kaggle team might be willing to provide it without extra charge if they think the challenge is of their interest)

Pros:

- High visibility among non-astronomers
- already built community
- platform has already been tested in many different situations (expertise)
- Kaggle team provides support: data preparation, development of Q&A, metric
- integrated forum to talk with the participants
- participants can use any language they want

Cons:

- using the provided kernels is **not obligatory**
- participants can use any language they want
- 1 challenge → 1 question
- good ideas might go unnoticed if they are not top listed

3 questions

U$ 15 000,00

Option 2: RAMP

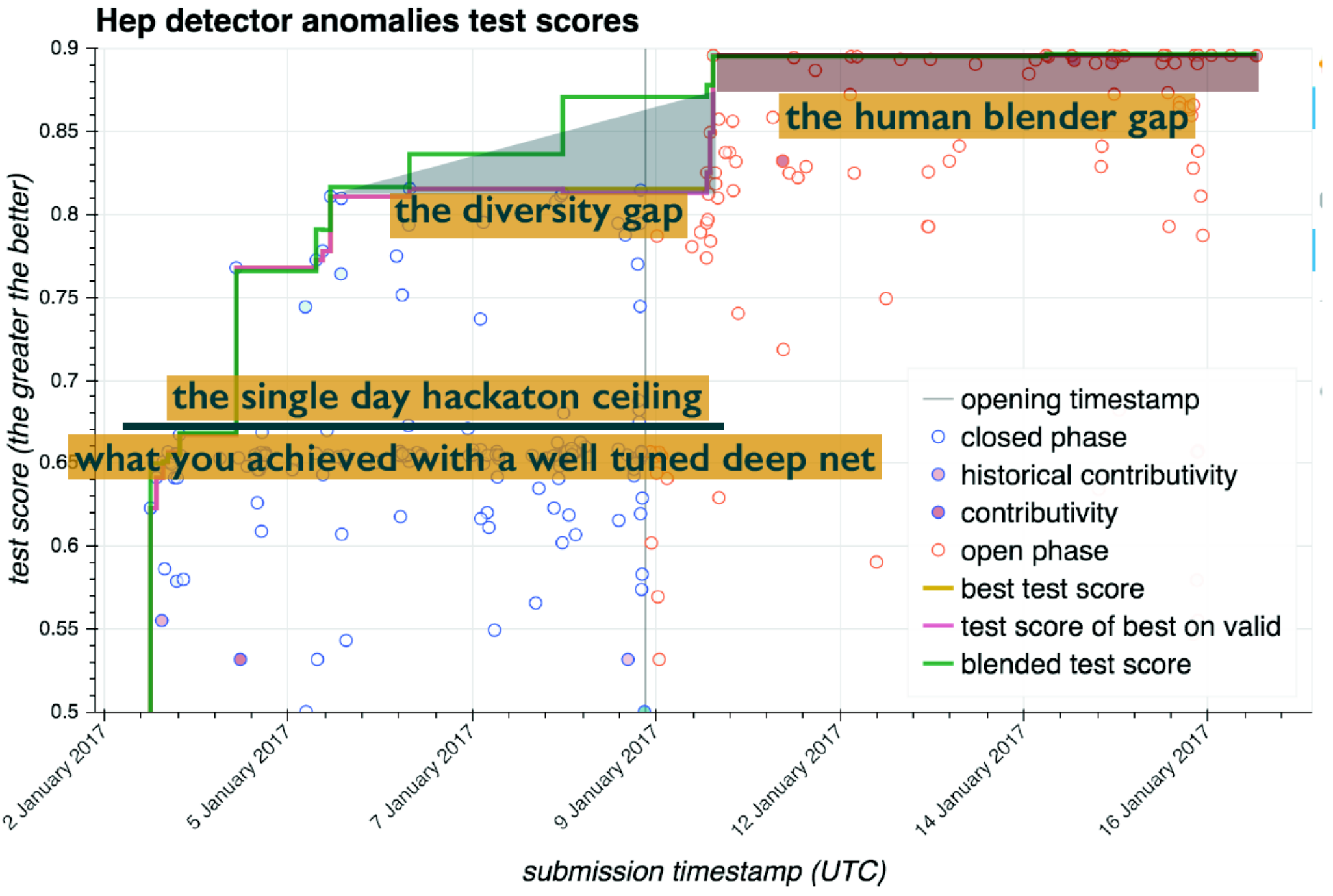
Rapid Analytics and Model Prototyping

Paris-Saclay Center for Data Science

www.ramp.studio

How it works:

- proposers work with RAMP team to check the data and prepare a starting kit
- participants submit **codes**
- entire calculation is done by the back-end, forums are hosted at Slack
- all code must be *open source*, and in **Python**
- 3 phases:
  - 1St – hackday: the RAMP team and proposers meet with the participants
  - 2nd - close: normal challenge (prizes should be given at the end of this stage)
  - 3Rd - open: everyone see each others code and collaborate to improve the result



Option 2: RAMP

Rapid Analytics and Model Prototyping

Paris-Saclay Center for Data Science

www.ramp.studio

How it works:

- proposers work with RAMP team to check the data
- participants submit **codes**
- entire calculation is done by the back-end, forums are hosted at Slack
- all code must be *open source*
- 3 phases:
  - 1St – hackday: the RAMP team and proposers meet with the participants
  - 2nd - close: normal challenge (prizes should be given at the end of this stage)
  - 3Rd - open: everyone see each others code and collaborate to improve the result

Pros:

- Submitted codes are guarantee to run
- The open phase allows good ideas to be used, even if the complete entry is not a top one
- No limit in the number of questions per competition
- **We can use computational resources at CC-IN2P3, Lyon as a back-end** *(under negotiation)*
- RAMP team is willing to provide the assistance (including implementation at CC-IN2P3)

Option 2: RAMP

Rapid Analytics and Model Prototyping

Paris-Saclay Center for Data Science

www.ramp.studio

How it works:

- proposers work with RAMP team to check the data
- participants submit **codes**
- entire calculation is done by the back-end, forums are hosted at Slack
- all code must be *open source*
- 3 phases:
  - 1St – hackday: the RAMP team and proposers meet with the participants
  - 2nd - close: normal challenge (prizes should be given at the end of this stage)
  - 3Rd - open: everyone see each others code and collaborate to improve the result

Pros:

- Submitted codes are guarantee to run
- The open phase allows good ideas to be used, even if the complete entry is not a top one
- No limit in the number of questions per competition
- **We can use computational resources at CC-IN2P3, Lyon as a back-end** *(under negotiation)*
- RAMP team is willing to provide the assistance (including implementation at CC-IN2P3)

Cons:

- code must be in **Python**
- less visibility – it would require a larger effort on dissemination to reach non-astronomers
- *it should be allowed for outsiders to also participate in the open phase (?)*

Option 2: RAMP

Rapid Analytics and Model Prototyping

Paris-Saclay Center for Data Science

www.ramp.studio

How it works:

- proposers work with RAMP team to check the data
- participants submit **codes**
- entire calculation is done by the back-end, forums are hosted at Slack
- all code must be *open source*
- 3 phases:
  - 1St – hackday: the RAMP team and proposers meet with the participants
  - 2nd - close: normal challenge (prizes should be given at the end of this stage)
  - 3Rd - open: everyone see each others code and collaborate to improve the result

Pros:

- Submitted codes are guarantee to run
- The open phase allows good ideas to be used, even if the complete entry is not a top one
- No limit in the number of questions per competition
- **We can use computational resources at CC-IN2P3, Lyon as a back-end** *(under negotiation)*
- RAMP team is willing to provide the assistance (including implementation at CC-IN2P3)

Cons:

- code must be in **Python**
- less visibility – it would require a larger effort on dissemination to reach non-astronomers
- *it should be allowed for outsiders to also participate in the open phase (?)*

Next steps:

1) **Characterize the data:**
   - how big the data will be?
   - can be estimate the magnitude of computer resources necessary for 1 participant?

2) **Define our expectations:**
   - how many questions we want to answer?
   - how many prizes we will give?

3) **Choose a hosting platform:**
   3.a) If we want more information on **Kaggle**:
      - estimate computer resources and possibility to use Amazon servers
      - get a more realistic estimation of costs

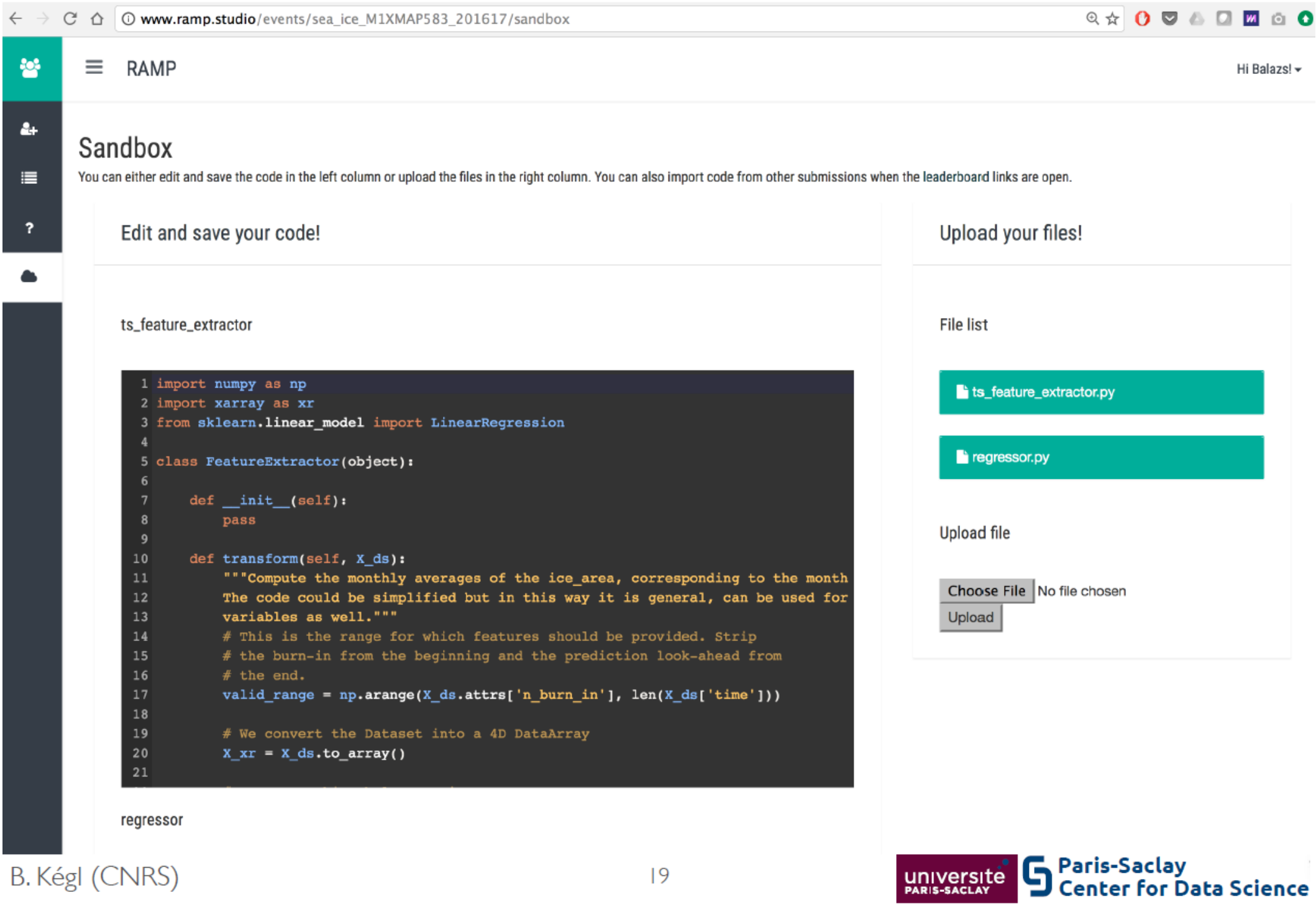
   3.b) If we want more information on **RAMP**:
      - invite someone from the RAMP team to one of our telecons
      - estimate computer resources necessary (ASAP)
      - formalized our request to use CC-IN2P3 and RAMP (there will be work in implementation for them)

   3.c) Other options?

4) **Prepare for post-challenge phase:**
   - what do we expect to have as a product from these efforts?
     If papers, how many? In what time scale? With what focus?
   - how to adapt the results for subsequent analysis with real data?

Extra slide

RAMP



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

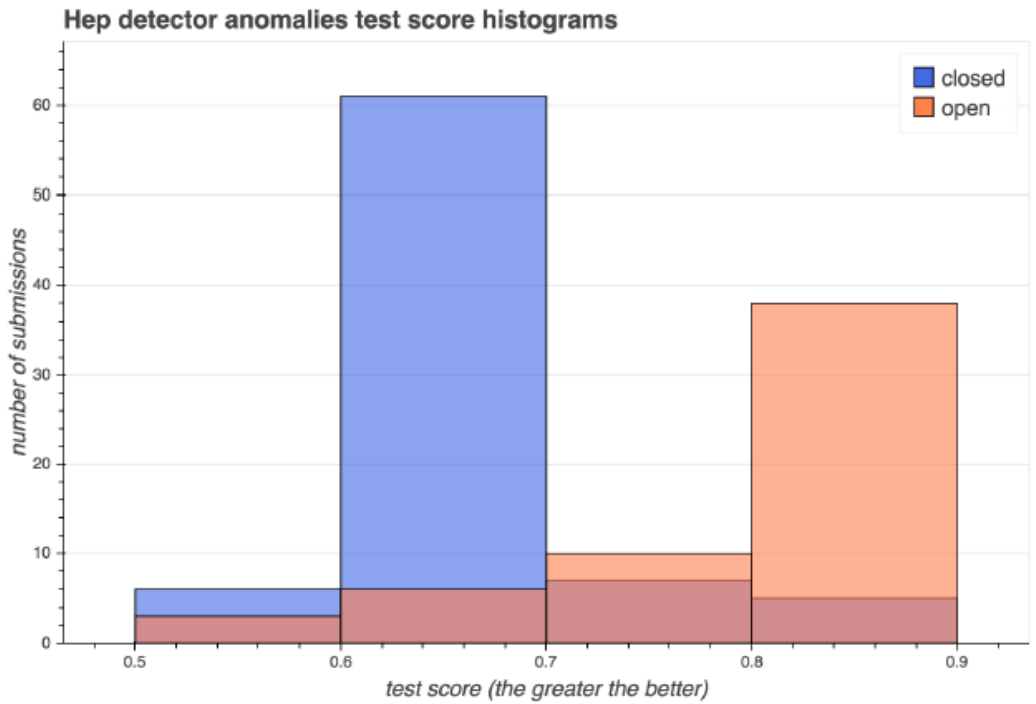
Search:

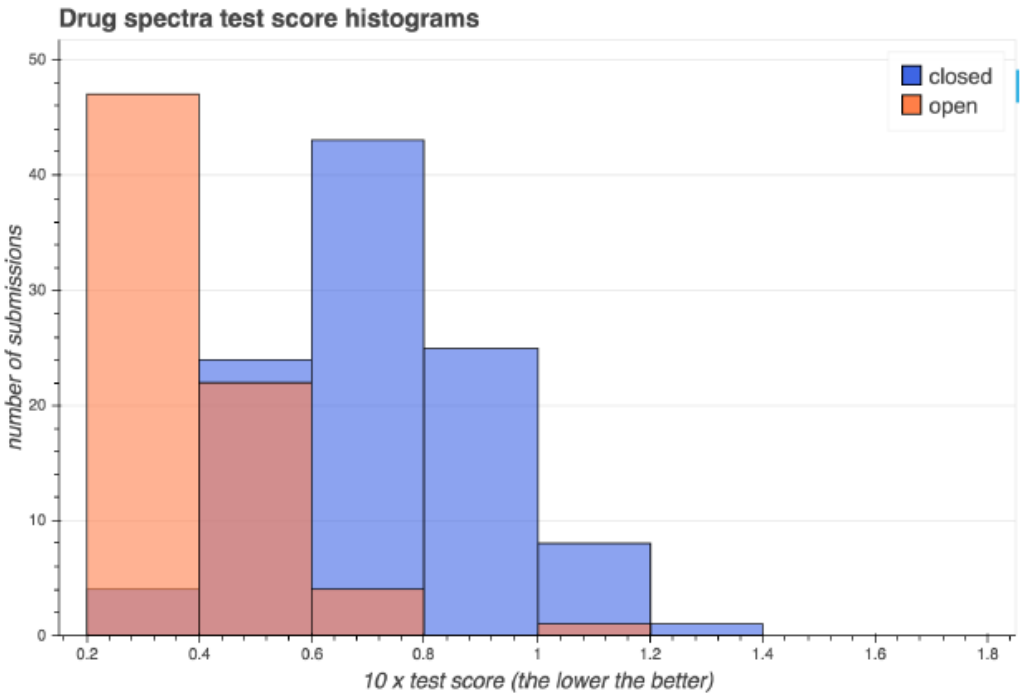
team	submission	contributivity	historical contributivity	rmse	train time	test time	submitted at (UTC)
joseph.budin	noName	26	3	0.279	286	3	2017-02-13 11:36:28 Mon
alexis.thual	timeseries	16	16	0.296	1	1	2017-02-13 17:48:47 Mon
julien.habis	try_hard3	11	8	0.300	475	3	2017-02-13 19:45:35 Mon
kangzheng.liang	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

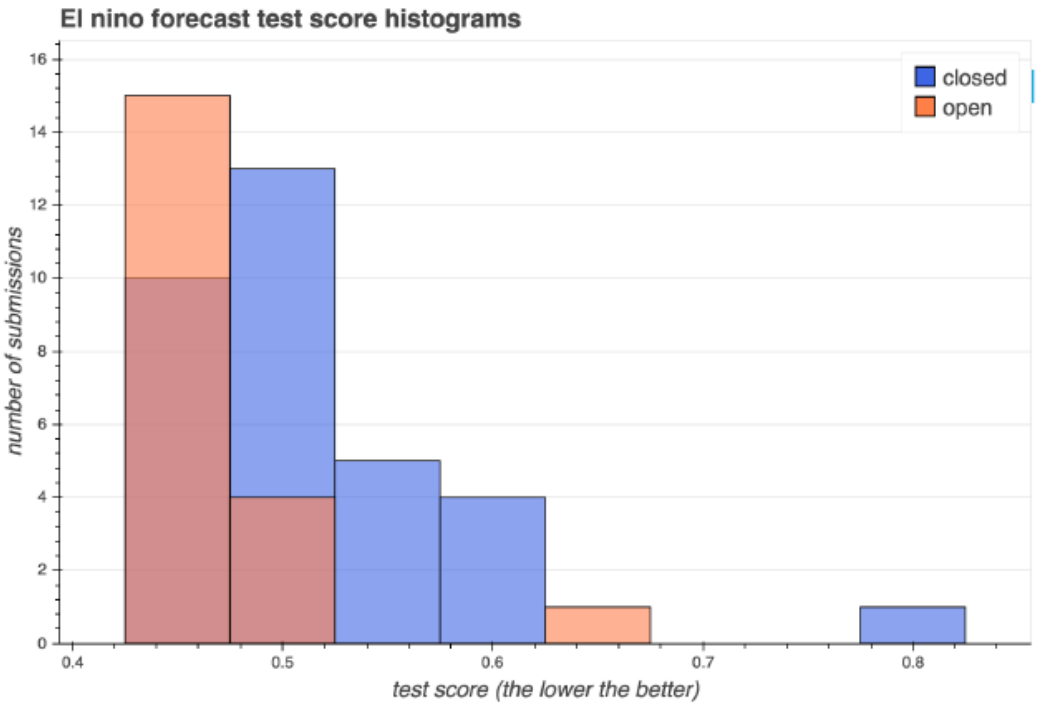
Showing 1 to 10 of 172 entries

Previous 1 2 3 4 5 … 18 Next

OPEN PHASE LETS PARTICIPANTS CATCH UP







Hep detector anomalies submissions