M2 Data Camp 2017

Balázs Kégl
Université Paris-Saclay
CNRS

Alexandre Gramfort
Université Paris-Saclay
INRIA

université PARIS-SACLAY

Paris-Saclay Center for Data Science

Most classical data challenges are HR and publicity events

We decided to turn them into a tool for

1. Collaborative prototyping
2. Teaching aid
3. Data science process management

Funded by Université Paris-Saclay

Team


Balázs Kégl


Alex Gramfort


Akin Kazakçi


Mehdi Cherti


Yohann Sitruk


Guillaume Lemaître


Alexandre Boucaud


Joris Van den Bossche

Alumni


Djalel Benbouzid


Camille Marini

RAMP.STUDIO

DATA CHALLENGE WITH CODE SUBMISSION

RAMP

Hi Balázs!

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!

classifier

```
from sklearn.base import BaseEstimator
from sklearn.ensemble import RandomForestClassifier


class Classifier(BaseEstimator):
    def __init__(self):
        pass

    def fit(self, X, y):
        self.clf = RandomForestClassifier(
            n_estimators=2, max_leaf_nodes=3, random_state=61)
        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)
```

Upload your files!

File list

classifer.py

Upload file

Choose File No file chosen

Upload

RAMP

Leaderboard

Combined score: 0.899

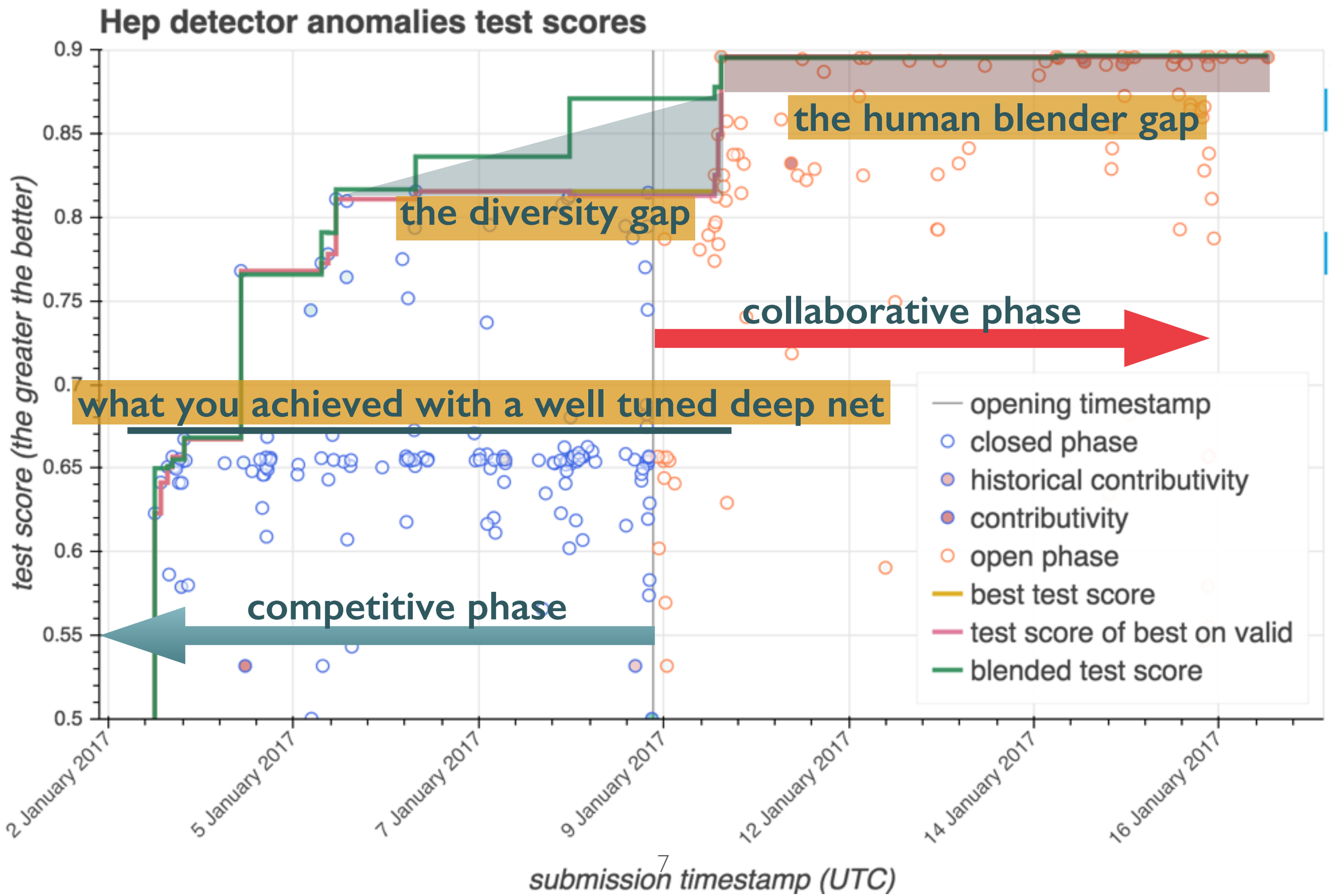
Show 10 entries

team	submission	contributivity	historical contributivity	auc	accuracy	nll	train time	test time	submitted at (UTC)
diego.souza	tuning_xgboost3	9	5	0.896	0.820	0.385	3074	30	2017-01-17 11:34:53 Tue
ndeye-fatou.diop	kit_from_all	5	1	0.896	0.819	0.382	1167	10	2017-01-14 20:03:00 Sat
diego.souza	tuning_xgboost2	4	2	0.896	0.819	0.385	4900	17	2017-01-15 19:35:03 Sun
ndeye-fatou.diop	kit_from_all_cleaner	3	0	0.896	0.819	0.384	1175	10	2017-01-15 03:45:44 Sun
etienne.boursier	combine_features	2	7	0.896	0.820	0.383	2712	3	2017-01-10 15:26:21 Tue
clement.vignac	boursier_improved_1	1	0	0.896	0.819	0.385	2499	4	2017-01-16 08:21:55 Mon

Code submission

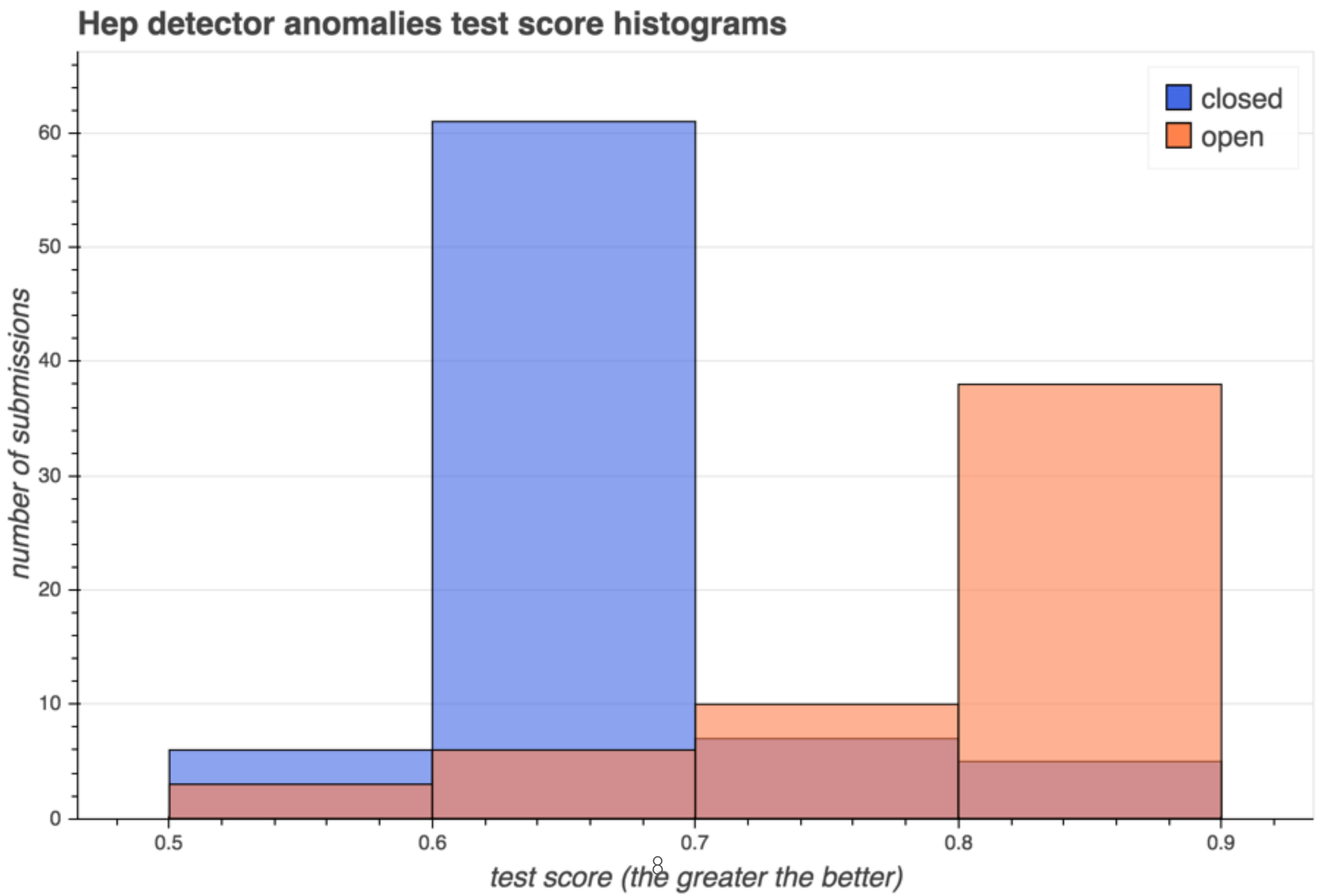
1. lets us deliver a working prototype
2. lets the participants collaborate
3. makes the backend challenging to run (cloud management)

THE POWER OF THE (COLLABORATING) CROWD

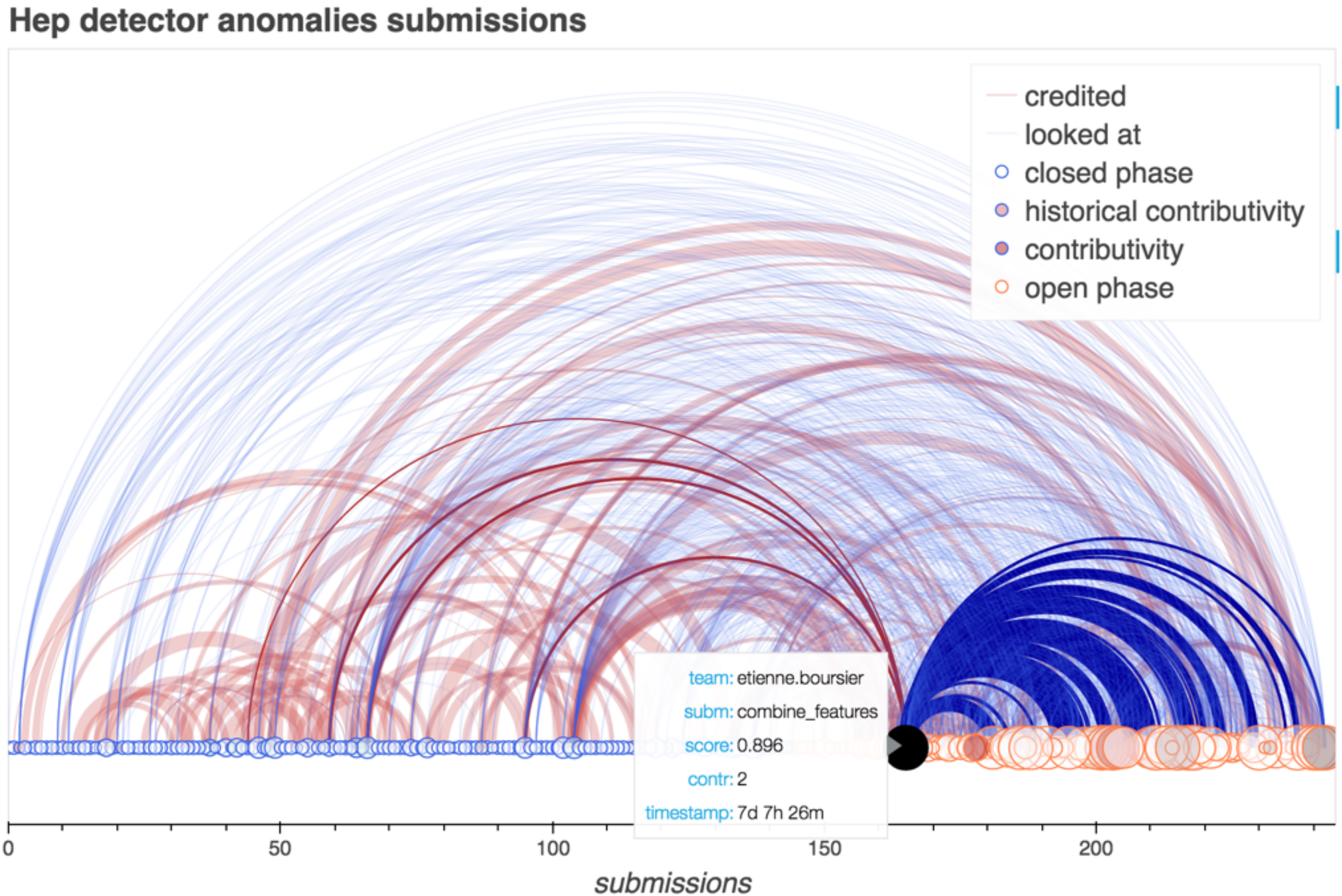


OPEN PHASE LETS PARTICIPANTS CATCH UP

THE GOAL OF TEACHING

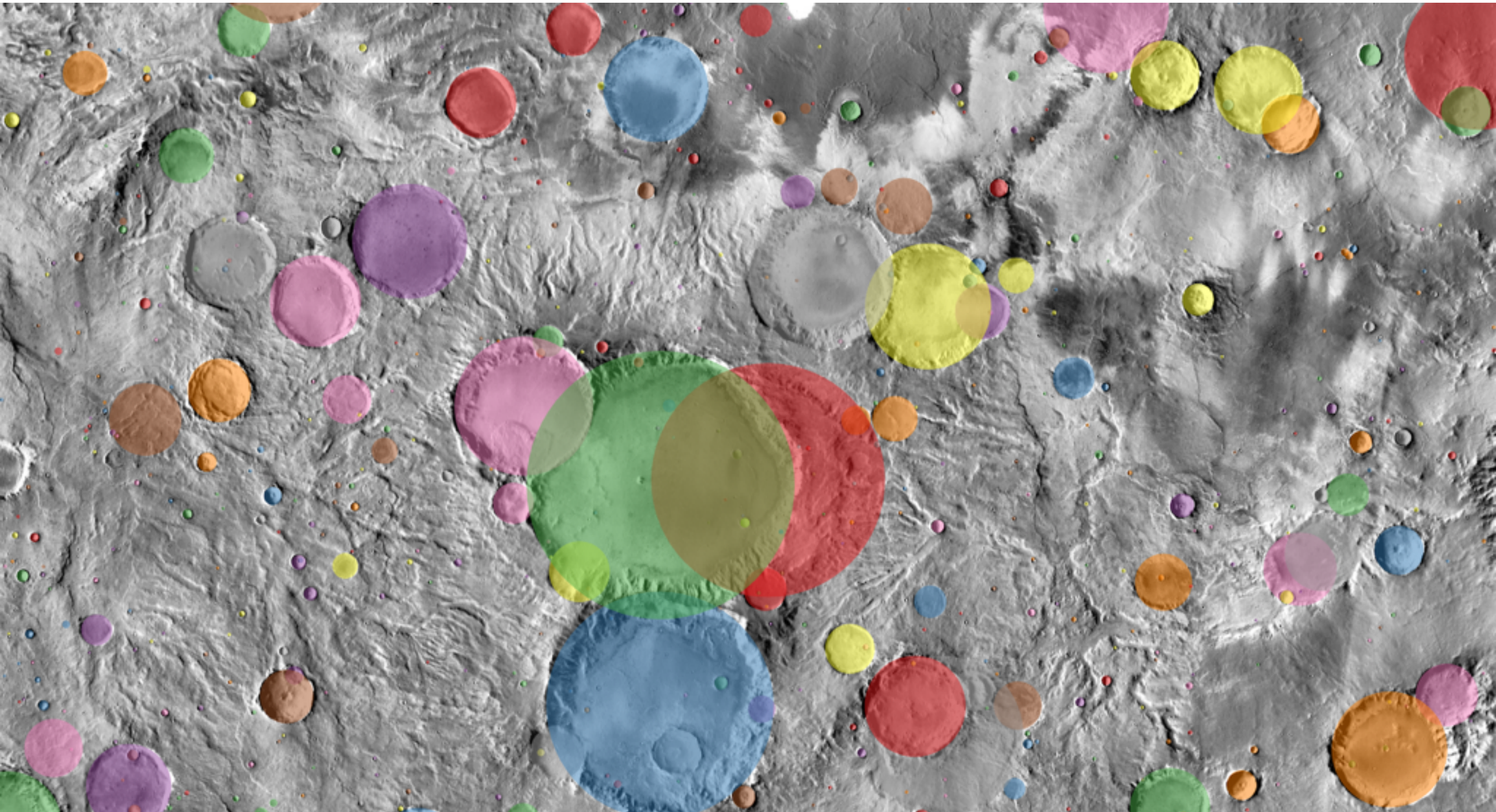


Communication and reuse



Choose one of three challenges

Challenge 1
Detect Mars craters

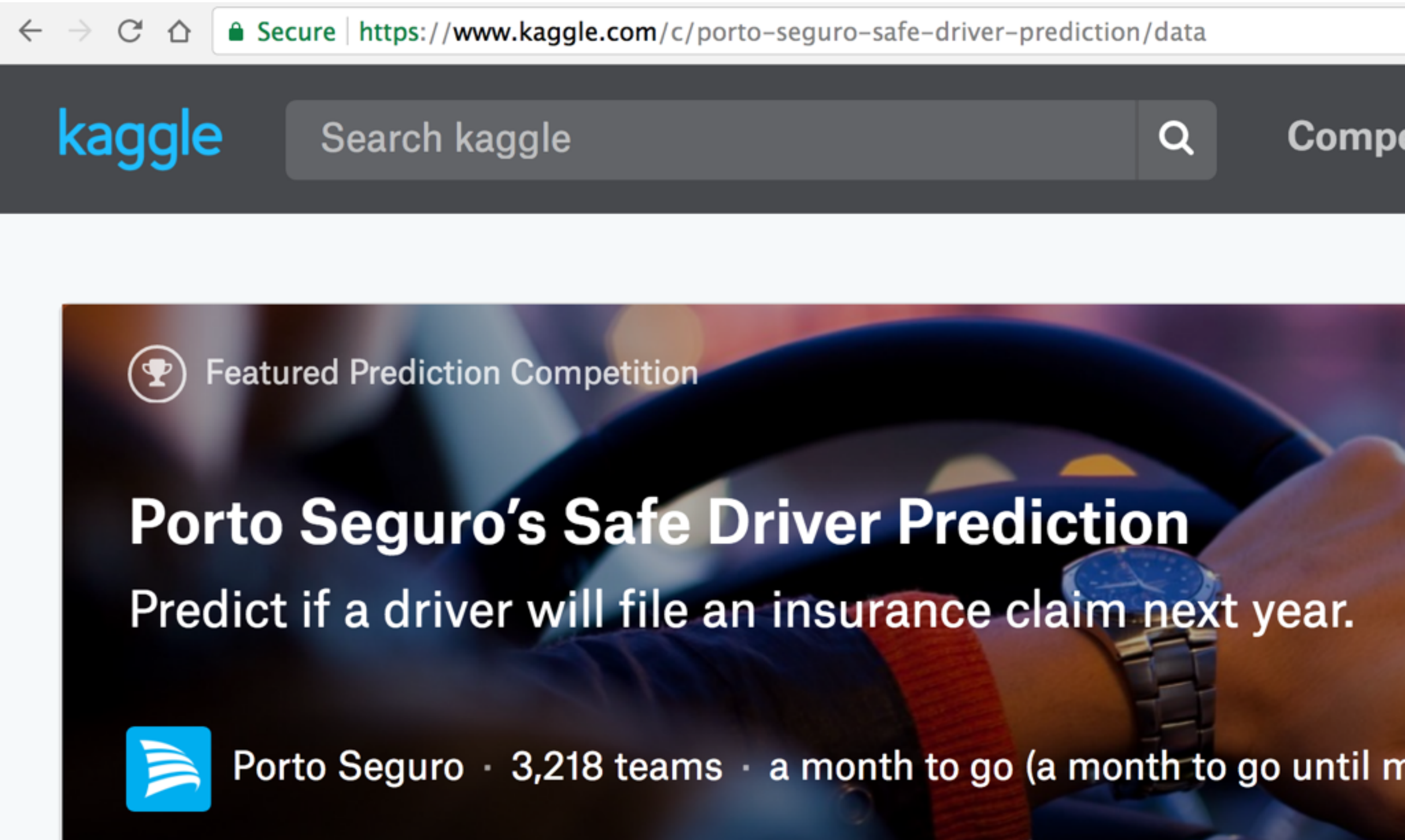


Challenge 2: fake news

Predict the truthfulness of news

Challenge 3: Kaggle Seguro

Predict insurance claims



RULES

- **Choose one challenge**
  - You will be graded on your **best** score across the three challenges
  - But **you can sign up** for all three
- **Competitive phase until December 17, 20h**
  - except for Kaggle Seguro
  - you will work **on your own**
  - mainly on your computer
  - submit **max once a day** on ramp.studio
  - a total of **50h training time**
    - single CPU
    - on Mars crater, you will be able to request GPUs

RULES

- **Collaborative phase until January 30, 20h**
  - You will be incentivized to **look at and reuse each other's codes**
  - You will be **graded on your influence** (credits you receive from your fellow students)
  - You will be graded on the **marginal improvements of the combined score** (to incentivize good but also diverse models)
  - Incentivizing the jumps also means that **this phase will probably finish before Christmas**
  - A **fresh 50h** of computational time
  - Grading of the competitive/collaborative phases will be closer to 5/5 than to 8/2 originally announced

Special rules for the Kaggle challenge

- **You can only sign up if**
  - You have your Kaggle account
  - You submit a solution at Kaggle with a **public score of at least 0.28**, before **Nov 6 20h**, when the competitive RAMP phase starts
  - You should **not have more than 5 submissions at Kaggle**, but try even less
  - You agree to **enter the RAMP team on Kaggle** and stop submitting to Kaggle on your own
- **Timeline**
  - The **competitive RAMP phase** will end on **Nov 20 at 20h**
  - The **collaborative phase** will run until the **Kaggle submission deadline on November 29**
  - Grading will be similar to the normal challenges, but this challenge will be **open to non-datacamp participation**
  - In case we win money, 50% goes to the CDS, 50% will be shared among the participants, according to a similar scheme we're using to grade you (contribution to jumps and influence)

How do I choose

- Challenge 1:
  - **Image detection**, **deep learning**, hot but complicated
  - we will help but you will need to **manage your GPU-equipped machine** (either your own or on AWS)
  - you will **learn a lot** an will be **red hot on the job market**
  - potentially a **large margin of improvement over the baseline**
  - we're planning a **collective ICML paper**
- Challenge 2:
  - **NLP** + **categorical** variables, **small data**, computationally relatively simple
  - it is possible that there is **little margin over the baseline**
  - we are planning to run a **high-profile money-prize challenge on this later**
- Challenge 3:
  - **tabular** but quite **big data** (~1M instances)
  - **tight timeline**
  - may be quite high profile if we end up close to the top

platform:

www.ramp.studio

toolkit:

github.com/paris-saclay-cds/ramp-workflow

examples:

github.com/ramp-kits

slack:

https://join.slack.com/t/datacamp2017/signup

course syllabus:

http://bit.ly/datacamp2017