RAMP

Data challenges with modularization and code submission

Balázs Kégl

Université Paris-Saclay / CNRS

universite PARIS-SACLAY

Paris-Saclay Center for Data Science

Who am I?

Balázs Kégl

- Senior researcher CNRS
  - machine learning (20 years)
    interfacing with particle physics (10 years)
- Head of the Paris-Saclay Center for Data Science
  - interfacing with biology, economy, climatology, chemistry, etc. (4 years)

universite
PARIS-SACLAY

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

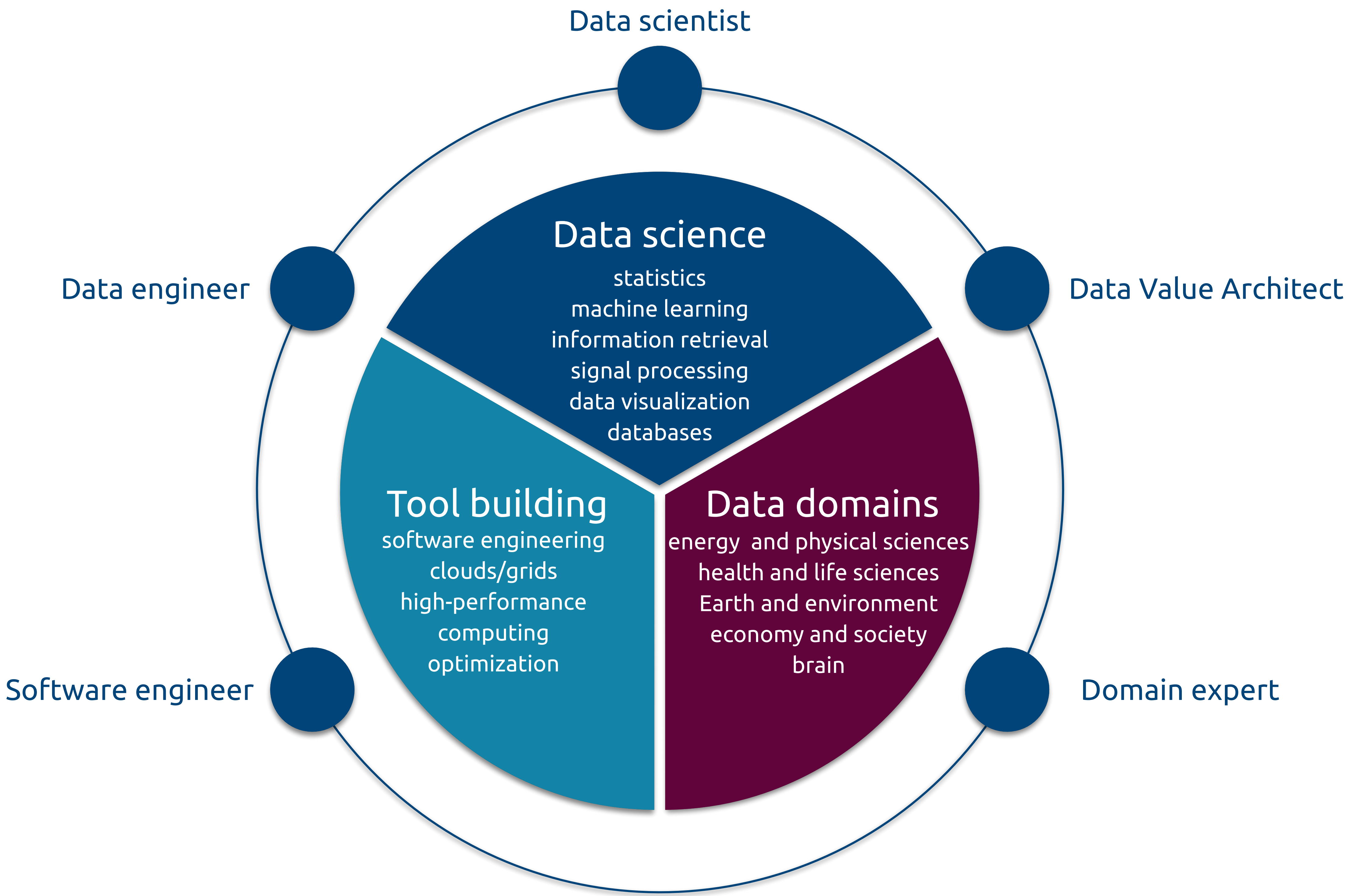
Visualization
INRIA
LIMSI

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

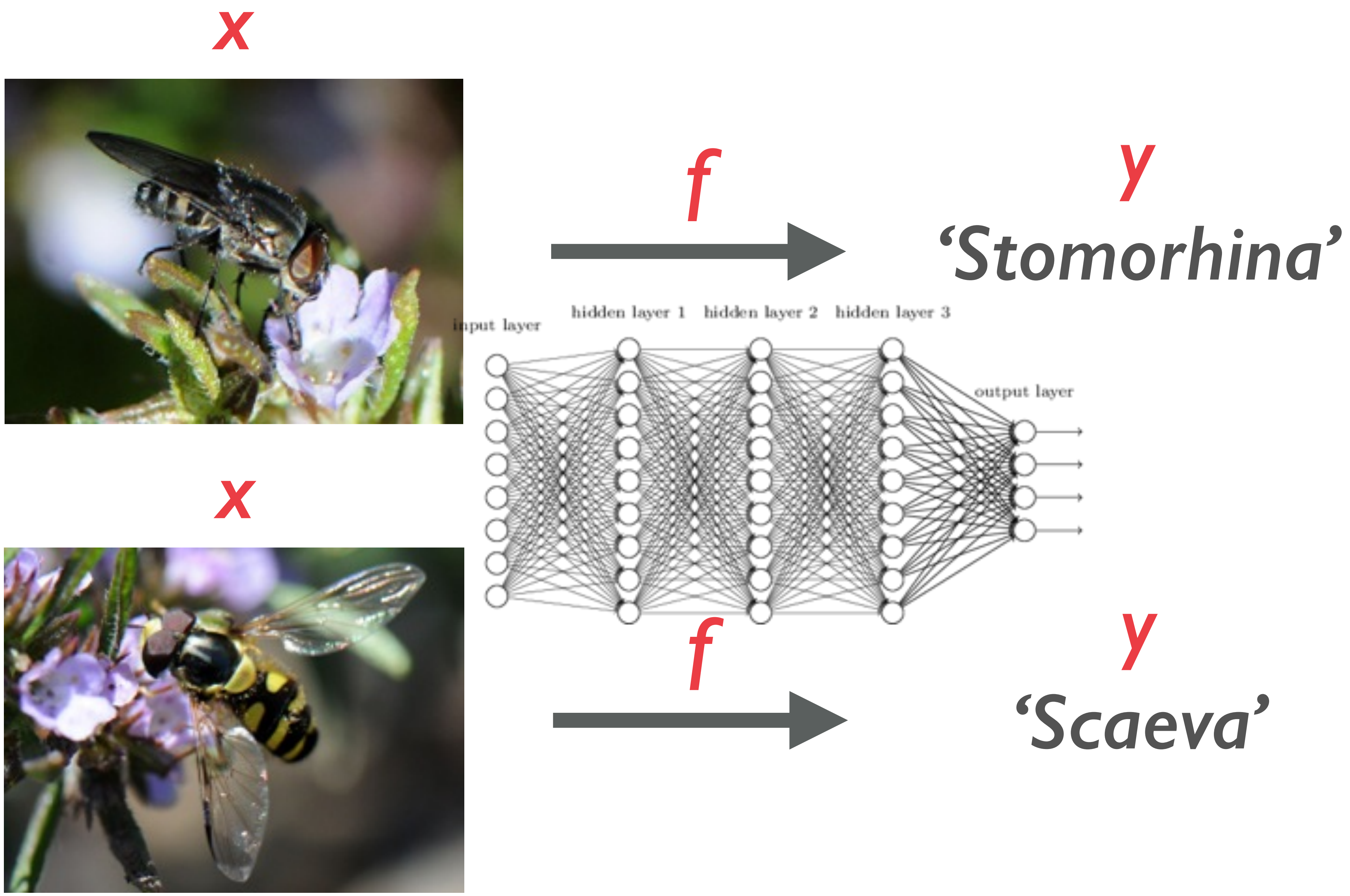
THE DATA SCIENCE ECOSYSTEM

https://medium.com/@balazskegl



WHAT DOES MACHINE LEARNING DO

- **Classification** problem $y = f(x)$



x

f

y

'Stomorhina'

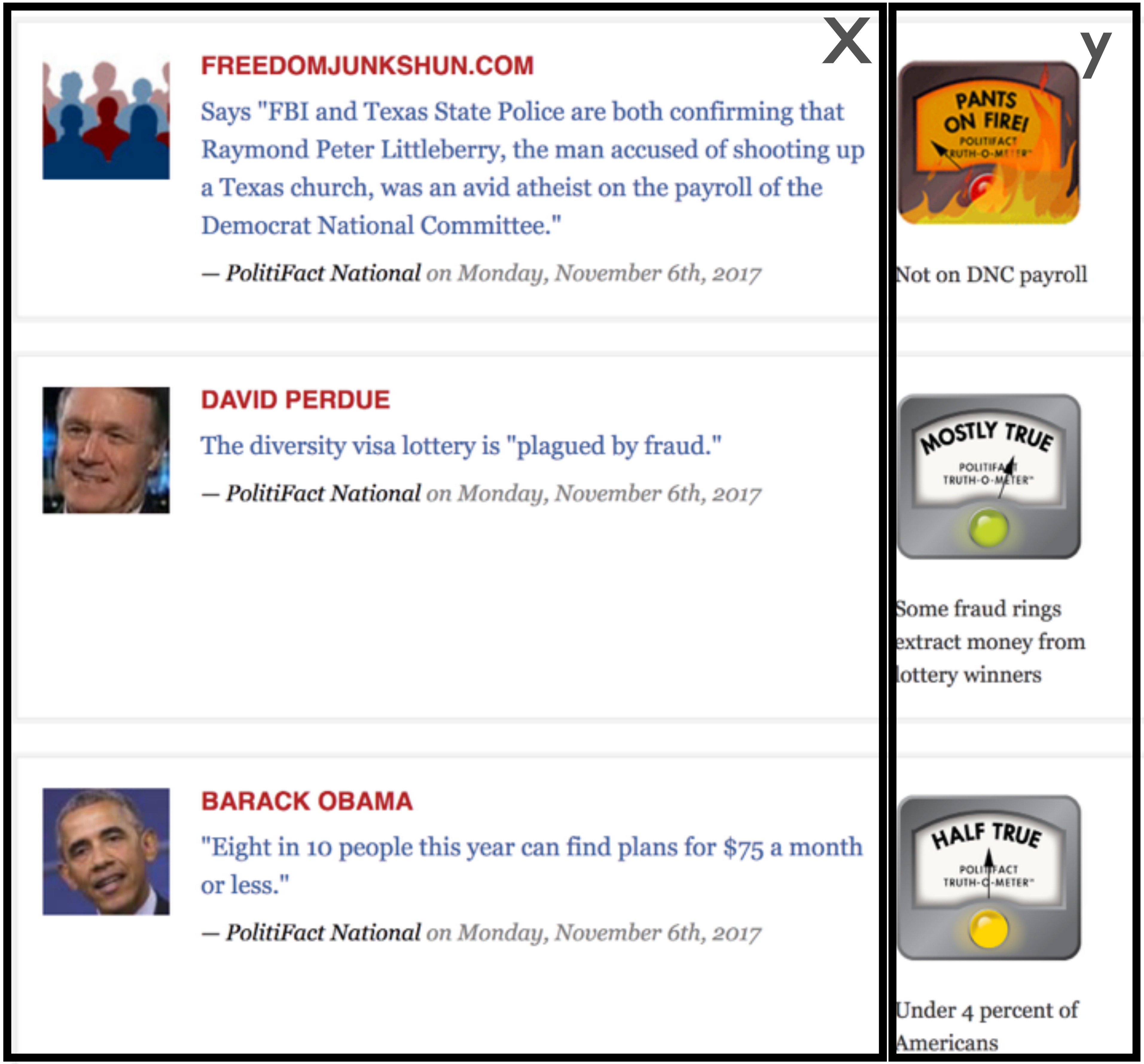
x

f

y

'Scaeva'

What does machine learning do


X
y
PANTS ON FIRE!
POLITIFACT TRUTH-O-METER™
MOSTLY TRUE
POLITIFACT TRUTH-O-METER™
HALF TRUE
POLITIFACT TRUTH-O-METER™

FREEDOMJUNKSHUN.COM

Says "FBI and Texas State Police are both confirming that Raymond Peter Littleberry, the man accused of shooting up a Texas church, was an avid atheist on the payroll of the Democrat National Committee."

— *PolitiFact National* on *Monday, November 6th, 2017*

Not on DNC payroll

DAVID PERDUE

The diversity visa lottery is "plagued by fraud."

— *PolitiFact National* on *Monday, November 6th, 2017*

Some fraud rings extract money from lottery winners

BARACK OBAMA

"Eight in 10 people this year can find plans for $75 a month or less."

— *PolitiFact National* on *Monday, November 6th, 2017*

Under 4 percent of Americans

ML IN NATURAL SCIENCES

- Inference: to invert the generative model
  - “predict” a particle, detect an anomaly, infer a parameter y, from observation x
- Generation: to replace expensive simulations
  - “learn” a GEANT4 simulation with a neural net
- Hypothesis generation: to replace theoreticians :)
  - learn, represent structural knowledge and generate novelty in model space, e.g., molecule generation in drug discovery

RAMP is a tool for

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

Funded by Université Paris-Saclay and CNRS

Team

Balázs Kégl

Alex Gramfort

Akin Kazakçi

Mehdi Cherti

Yohann Sitruk

Guillaume Lemaître

Alexandre Boucaud

Joris Van den Bossche

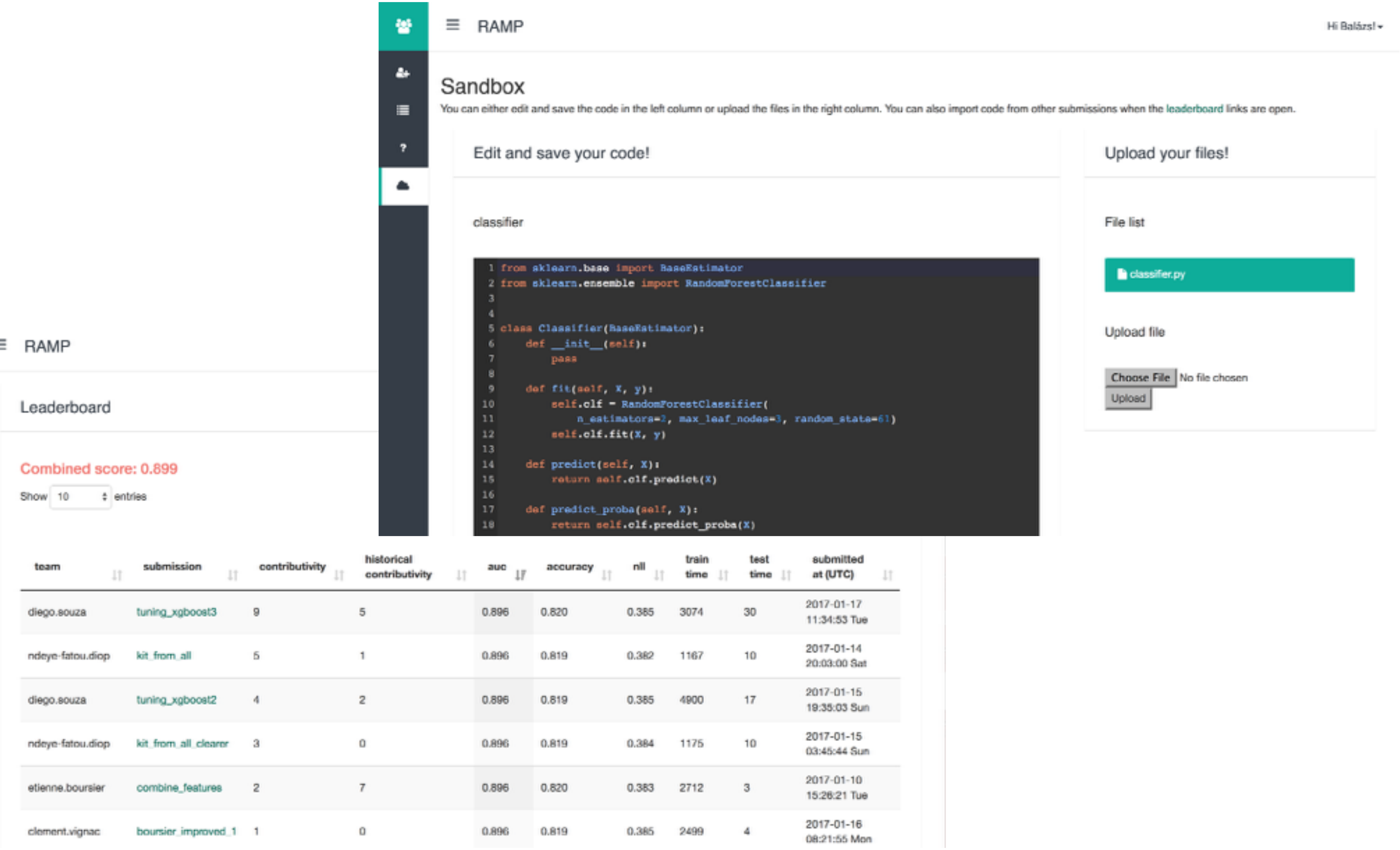
Alumni

Djalel Benbouzid

Camille Marini

RAMP.STUDIO

DATA CHALLENGE WITH CODE SUBMISSION



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

```python
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

Upload file

Leaderboard

Combined score: 0.899

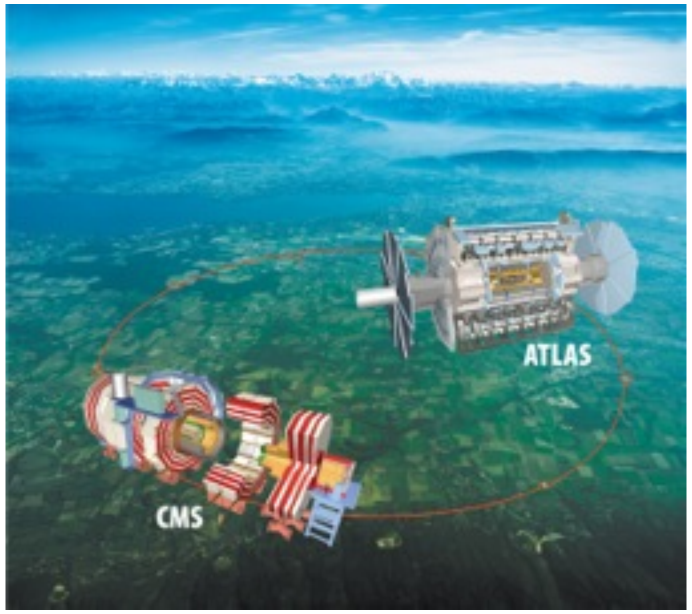
Show 10 entries

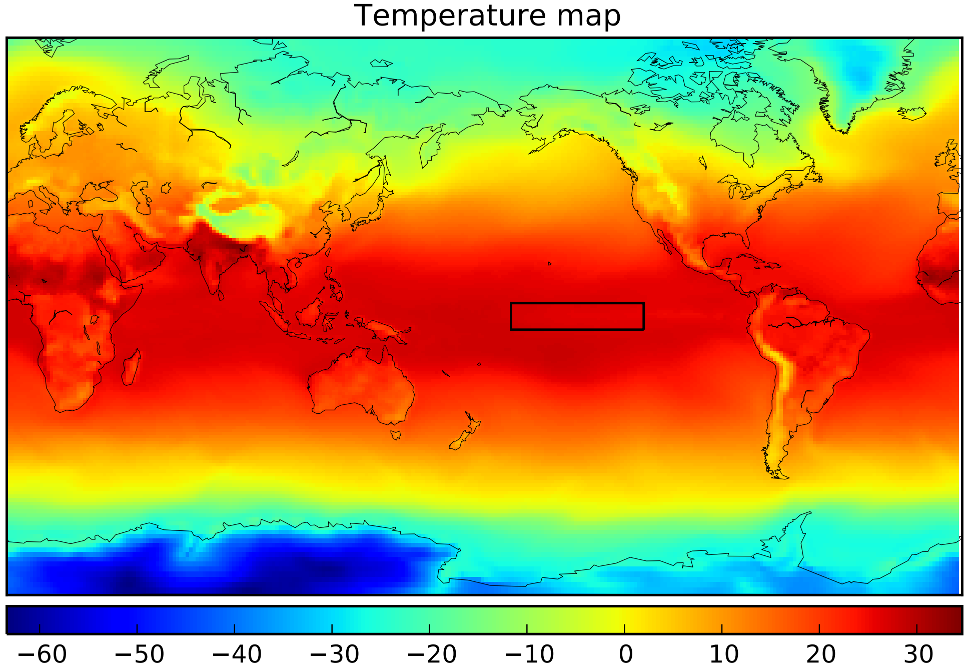
team	submission	contributivity	historical contributivity	auc	accuracy	nll	train time	test time	submitted at (UTC)
diego.souza	tuning_xgboost3	9	5	0.896	0.820	0.385	3074	30	2017-01-17 11:34:53 Tue
ndeye-fatou.diop	kit_from_all	5	1	0.896	0.819	0.382	1167	10	2017-01-14 20:03:00 Sat
diego.souza	tuning_xgboost2	4	2	0.896	0.819	0.385	4900	17	2017-01-15 19:35:03 Sun
ndeye-fatou.diop	kit_from_all_clearer	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

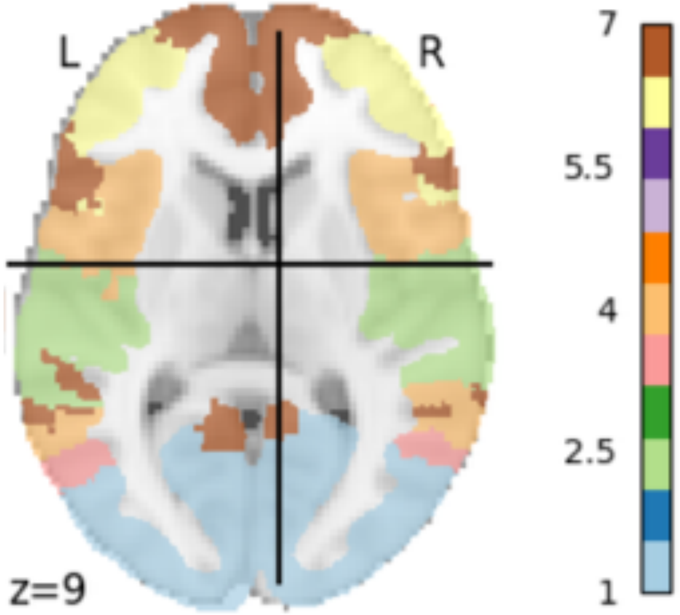
Code submission

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

Rapid analytics and model prototyping

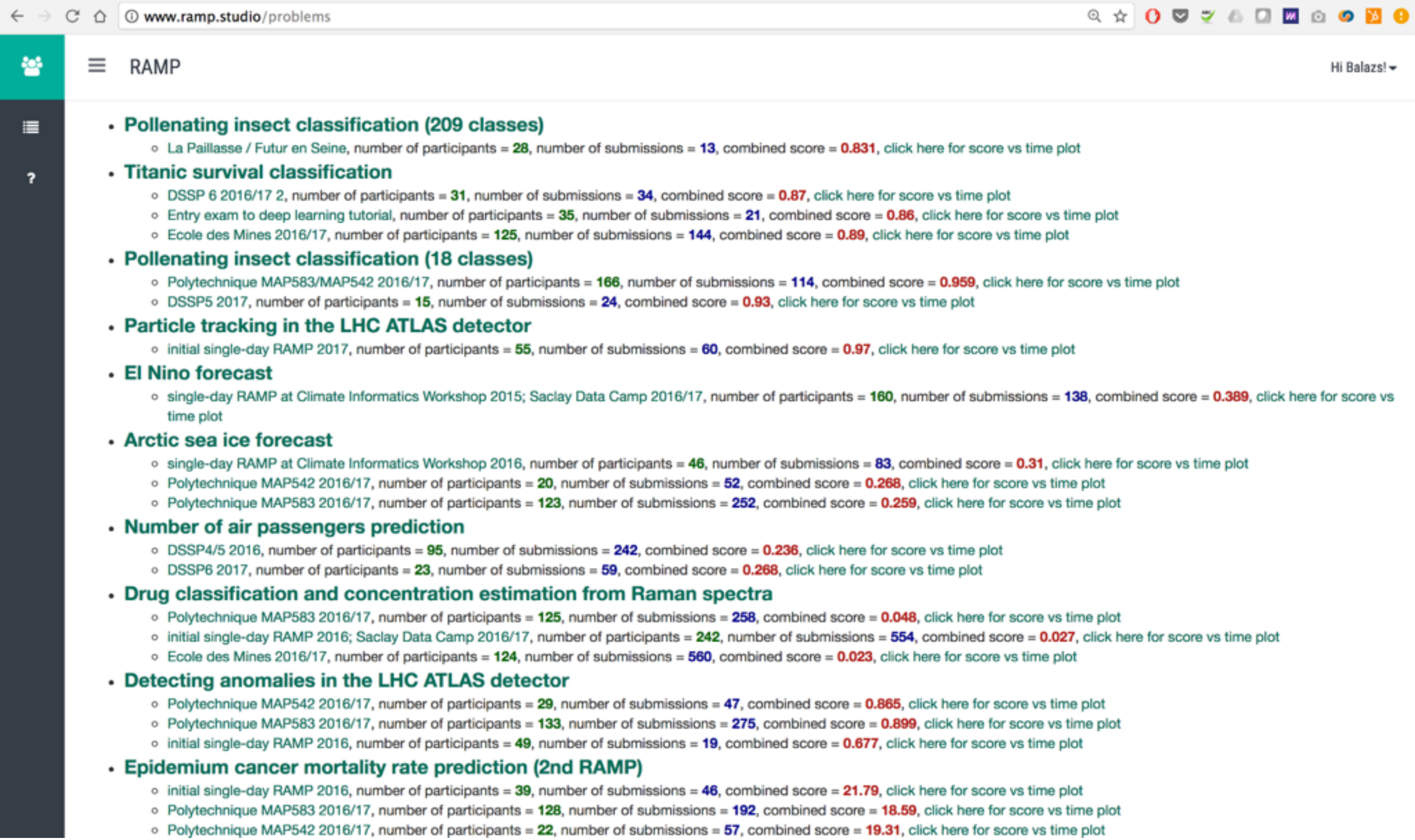






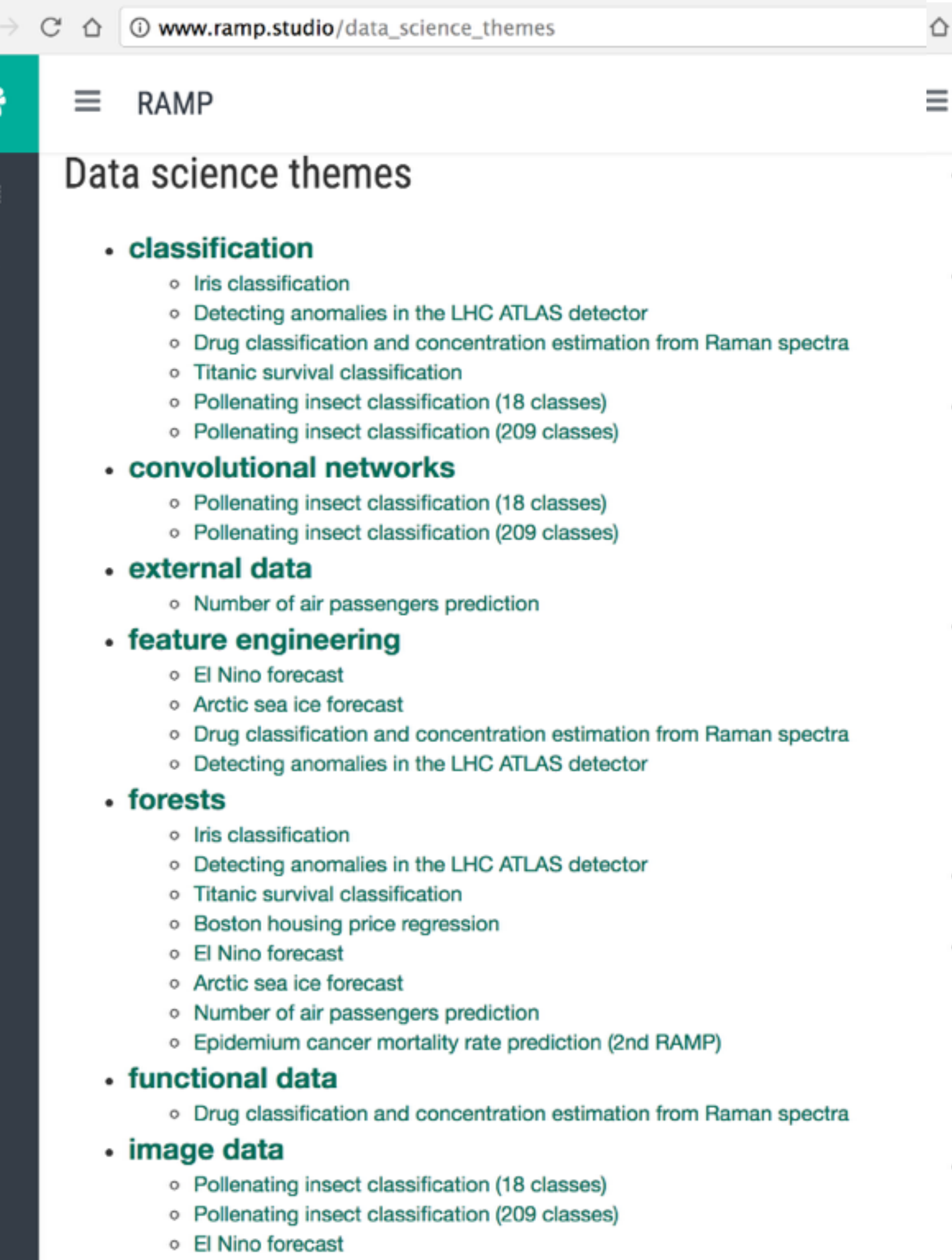
functional data, time series, data augmentation, deep learning, learning on simulations, nonstandard and multi-objective losses

CURRENT RAMPs



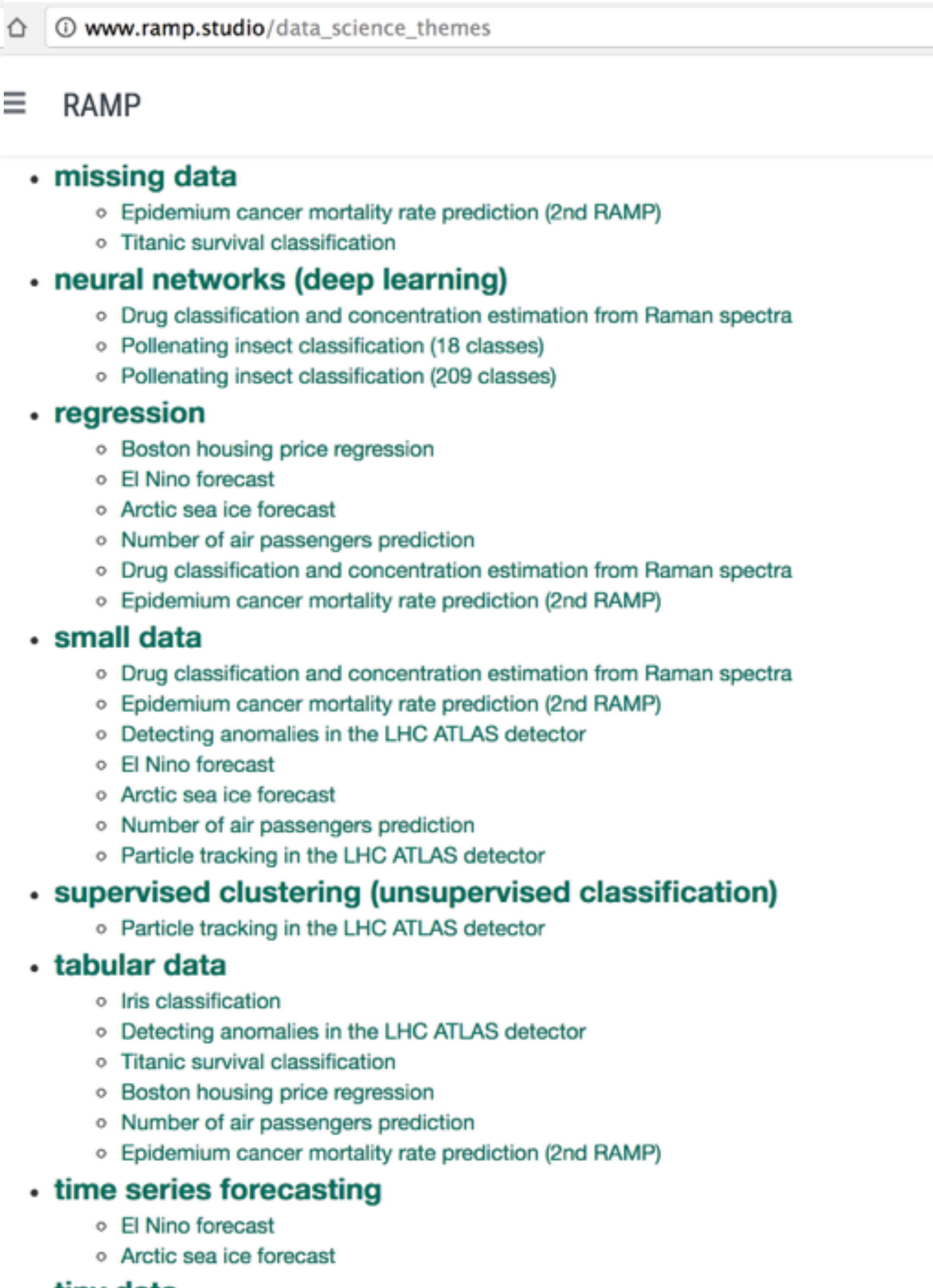
- **Pollenating insect classification (209 classes)**
  - La Paillasse / Futur en Seine, number of participants = 28, number of submissions = 13, combined score = 0.831, click here for score vs time plot
- **Titanic survival classification**
  - DSSP 6 2016/17 2, number of participants = 31, number of submissions = 34, combined score = 0.87, click here for score vs time plot
  - Entry exam to deep learning tutorial, number of participants = 35, number of submissions = 21, combined score = 0.86, click here for score vs time plot
  - Ecole des Mines 2016/17, number of participants = 125, number of submissions = 144, combined score = 0.89, click here for score vs time plot
- **Pollenating insect classification (18 classes)**
  - Polytechnique MAP583/MAP542 2016/17, number of participants = 166, number of submissions = 114, combined score = 0.959, click here for score vs time plot
  - DSSP5 2017, number of participants = 15, number of submissions = 24, combined score = 0.93, click here for score vs time plot
- **Particle tracking in the LHC ATLAS detector**
  - initial single-day RAMP 2017, number of participants = 55, number of submissions = 60, combined score = 0.97, click here for score vs time plot
- **El Nino forecast**
  - single-day RAMP at Climate Informatics Workshop 2015; Saclay Data Camp 2016/17, number of participants = 160, number of submissions = 138, combined score = 0.389, click here for score vs time plot
- **Arctic sea ice forecast**
  - single-day RAMP at Climate Informatics Workshop 2016, number of participants = 46, number of submissions = 83, combined score = 0.31, click here for score vs time plot
  - Polytechnique MAP542 2016/17, number of participants = 20, number of submissions = 52, combined score = 0.268, click here for score vs time plot
  - Polytechnique MAP583 2016/17, number of participants = 123, number of submissions = 252, combined score = 0.259, click here for score vs time plot
- **Number of air passengers prediction**
  - DSSP4/5 2016, number of participants = 95, number of submissions = 242, combined score = 0.236, click here for score vs time plot
  - DSSP6 2017, number of participants = 23, number of submissions = 59, combined score = 0.268, click here for score vs time plot
- **Drug classification and concentration estimation from Raman spectra**
  - Polytechnique MAP583 2016/17, number of participants = 125, number of submissions = 258, combined score = 0.048, click here for score vs time plot
  - initial single-day RAMP 2016; Saclay Data Camp 2016/17, number of participants = 242, number of submissions = 554, combined score = 0.027, click here for score vs time plot
  - Ecole des Mines 2016/17, number of participants = 124, number of submissions = 560, combined score = 0.023, click here for score vs time plot
- **Detecting anomalies in the LHC ATLAS detector**
  - Polytechnique MAP542 2016/17, number of participants = 29, number of submissions = 47, combined score = 0.865, click here for score vs time plot
  - Polytechnique MAP583 2016/17, number of participants = 133, number of submissions = 275, combined score = 0.899, click here for score vs time plot
  - initial single-day RAMP 2016, number of participants = 49, number of submissions = 19, combined score = 0.677, click here for score vs time plot
- **Epidemium cancer mortality rate prediction (2nd RAMP)**
  - initial single-day RAMP 2016, number of participants = 39, number of submissions = 46, combined score = 21.79, click here for score vs time plot
  - Polytechnique MAP583 2016/17, number of participants = 128, number of submissions = 192, combined score = 18.59, click here for score vs time plot
  - Polytechnique MAP542 2016/17, number of participants = 22, number of submissions = 57, combined score = 19.31, click here for score vs time plot

DATA SCIENCE THEMES



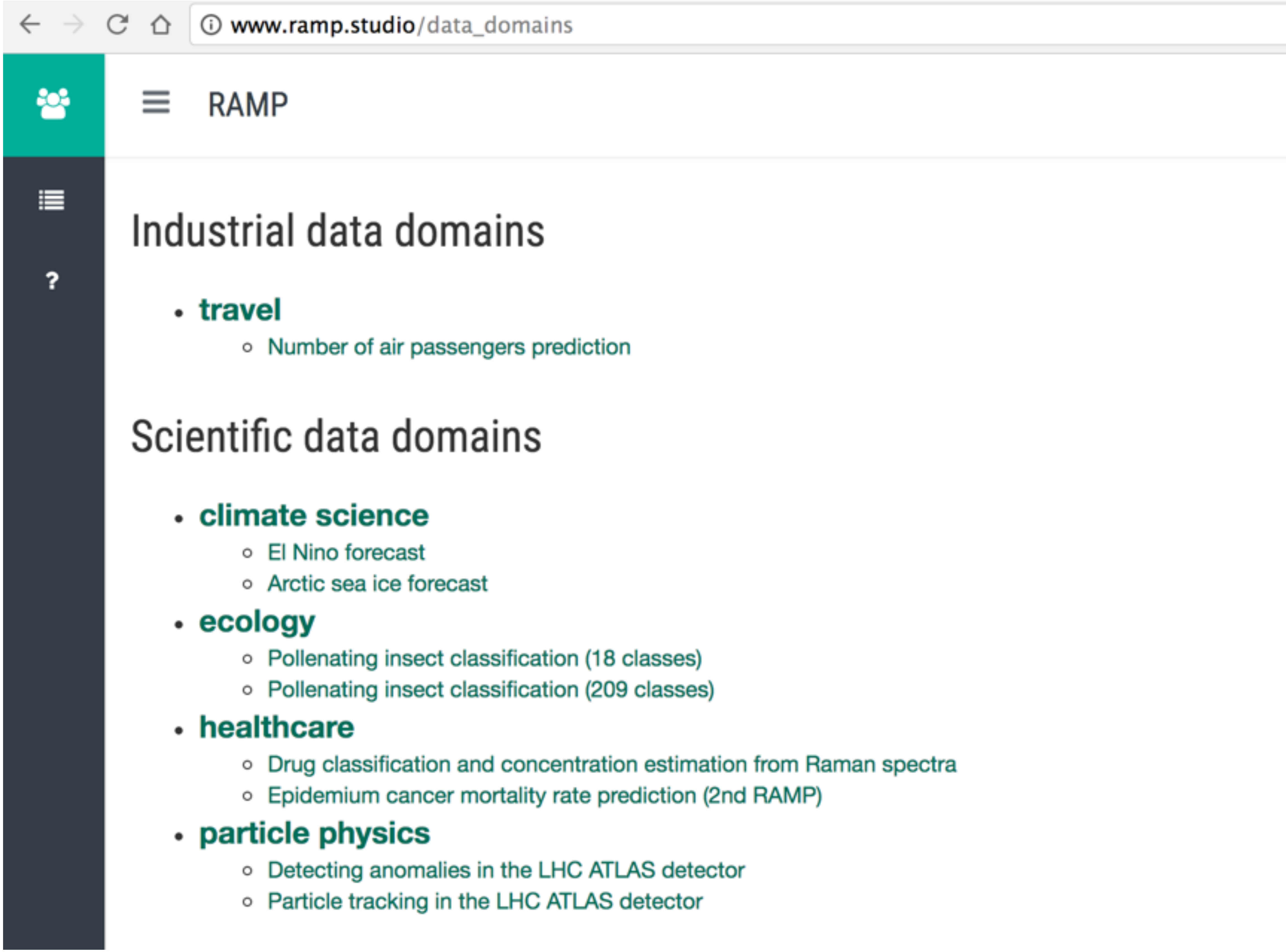
Data science themes

- **classification**
  - Iris classification
  - Detecting anomalies in the LHC ATLAS detector
  - Drug classification and concentration estimation from Raman spectra
  - Titanic survival classification
  - Pollenating insect classification (18 classes)
  - Pollenating insect classification (209 classes)
- **convolutional networks**
  - Pollenating insect classification (18 classes)
  - Pollenating insect classification (209 classes)
- **external data**
  - Number of air passengers prediction
- **feature engineering**
  - El Nino forecast
  - Arctic sea ice forecast
  - Drug classification and concentration estimation from Raman spectra
  - Detecting anomalies in the LHC ATLAS detector
- **forests**
  - Iris classification
  - Detecting anomalies in the LHC ATLAS detector
  - Titanic survival classification
  - Boston housing price regression
  - El Nino forecast
  - Arctic sea ice forecast
  - Number of air passengers prediction
  - Epidemium cancer mortality rate prediction (2nd RAMP)
- **functional data**
  - Drug classification and concentration estimation from Raman spectra
- **image data**
  - Pollenating insect classification (18 classes)
  - Pollenating insect classification (209 classes)
  - El Nino forecast



- **missing data**
  - Epidemium cancer mortality rate prediction (2nd RAMP)
  - Titanic survival classification
- **neural networks (deep learning)**
  - Drug classification and concentration estimation from Raman spectra
  - Pollenating insect classification (18 classes)
  - Pollenating insect classification (209 classes)
- **regression**
  - Boston housing price regression
  - El Nino forecast
  - Arctic sea ice forecast
  - Number of air passengers prediction
  - Drug classification and concentration estimation from Raman spectra
  - Epidemium cancer mortality rate prediction (2nd RAMP)
- **small data**
  - Drug classification and concentration estimation from Raman spectra
  - Epidemium cancer mortality rate prediction (2nd RAMP)
  - Detecting anomalies in the LHC ATLAS detector
  - El Nino forecast
  - Arctic sea ice forecast
  - Number of air passengers prediction
  - Particle tracking in the LHC ATLAS detector
- **supervised clustering (unsupervised classification)**
  - Particle tracking in the LHC ATLAS detector
- **tabular data**
  - Iris classification
  - Detecting anomalies in the LHC ATLAS detector
  - Titanic survival classification
  - Boston housing price regression
  - Number of air passengers prediction
  - Epidemium cancer mortality rate prediction (2nd RAMP)
- **time series forecasting**
  - El Nino forecast
  - Arctic sea ice forecast

DATA DOMAINS



Industrial data domains

- travel
  - Number of air passengers prediction

Scientific data domains

- climate science
  - El Nino forecast
  - Arctic sea ice forecast
- ecology
  - Pollenating insect classification (18 classes)
  - Pollenating insect classification (209 classes)
- healthcare
  - Drug classification and concentration estimation from Raman spectra
  - Epidemium cancer mortality rate prediction (2nd RAMP)
- particle physics
  - Detecting anomalies in the LHC ATLAS detector
  - Particle tracking in the LHC ATLAS detector

RAMP.STUDIO

DATA CHALLENGE WITH CODE SUBMISSION

20+ challenges

40+ events

1200+ users

7000+ predictive models

RAMP.STUDIO

DATA CHALLENGE WITH CODE SUBMISSION

12 hackathons

6 remote data challenges

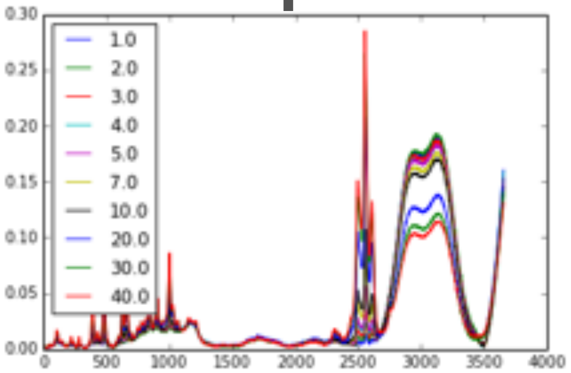
11 course data camps

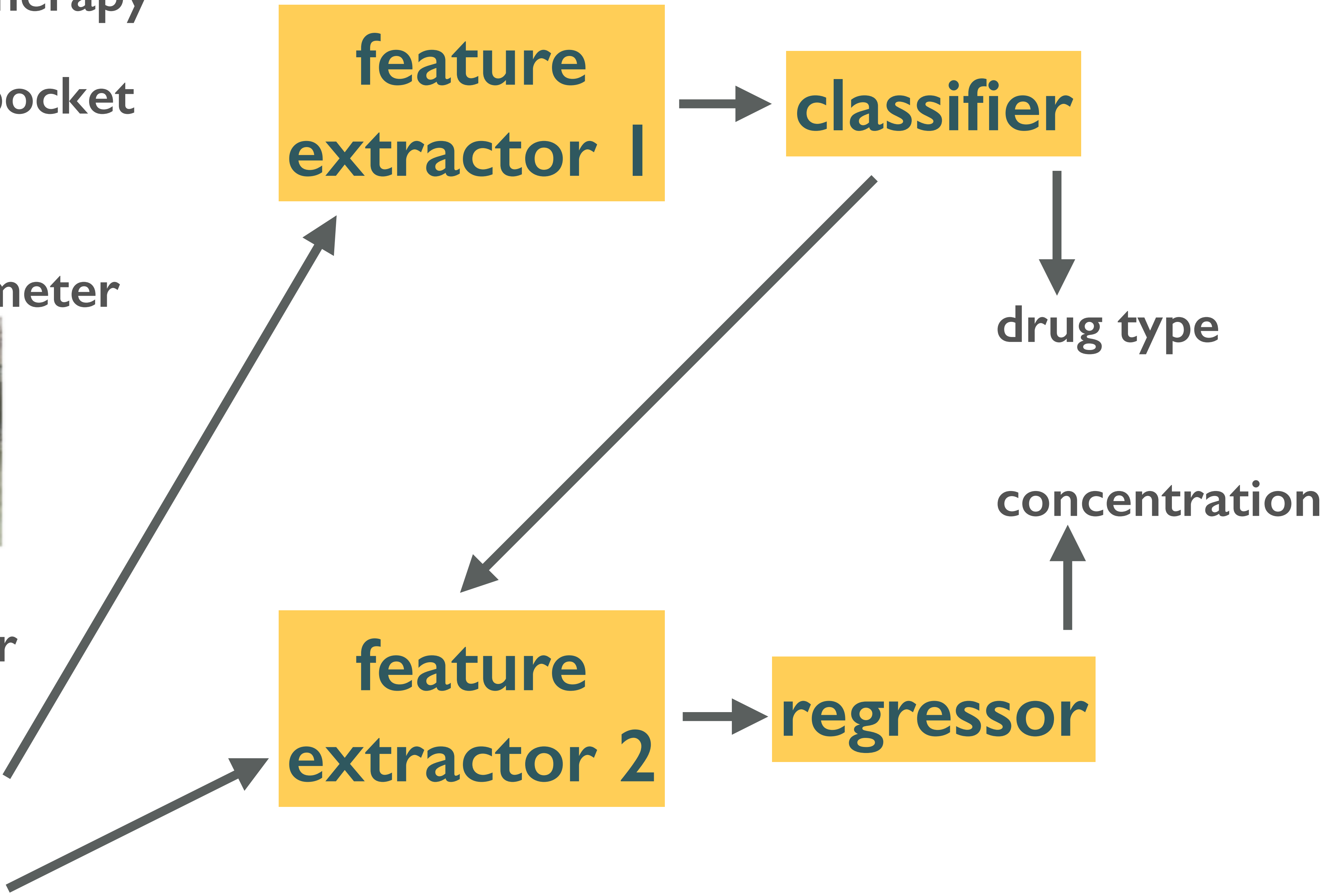
CLASSIFYING AND REGRESSING ON MOLECULAR SPECTRA

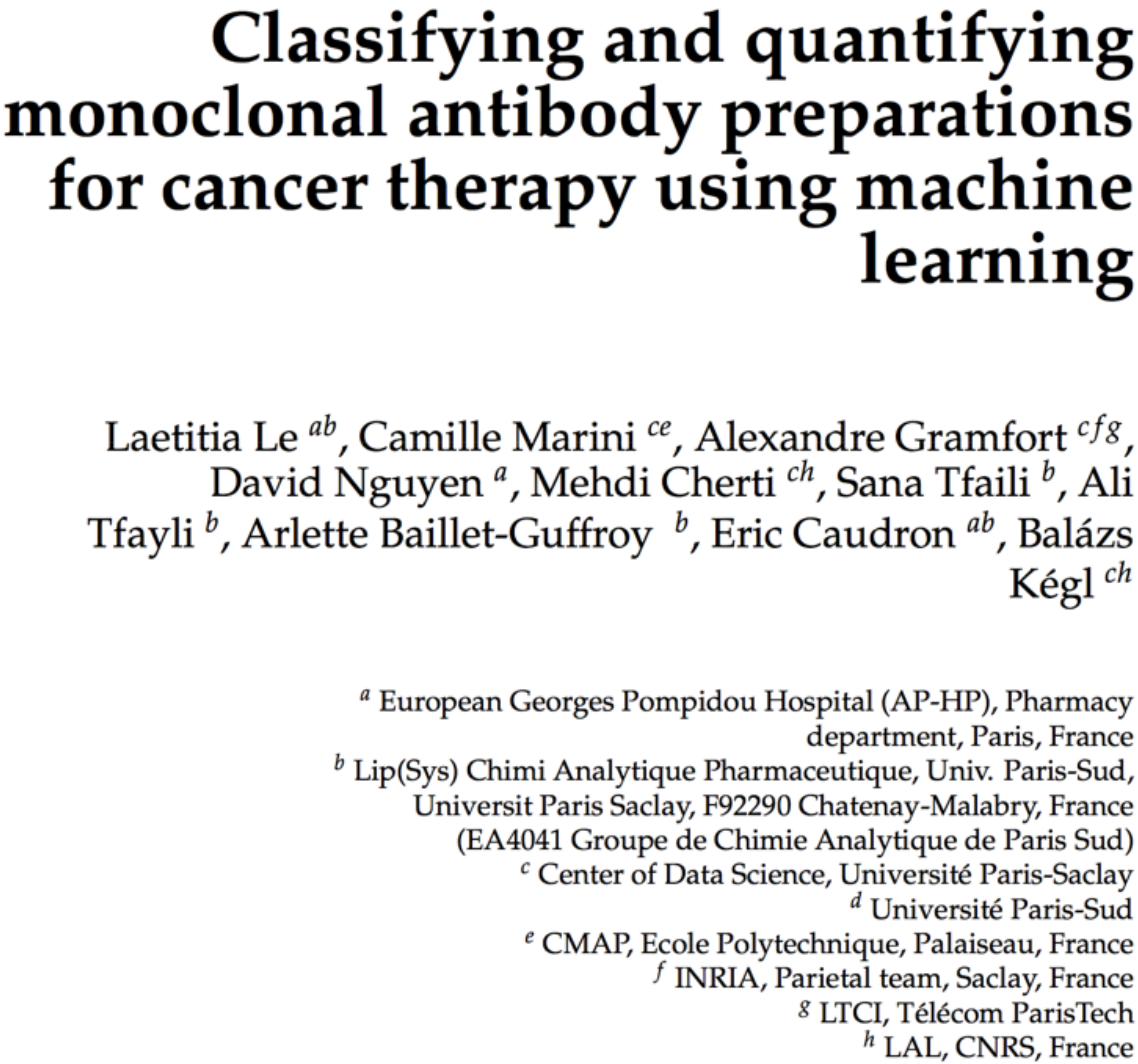
chemotherapy drug in elastic pocket

laser spectrometer

molecular spectra





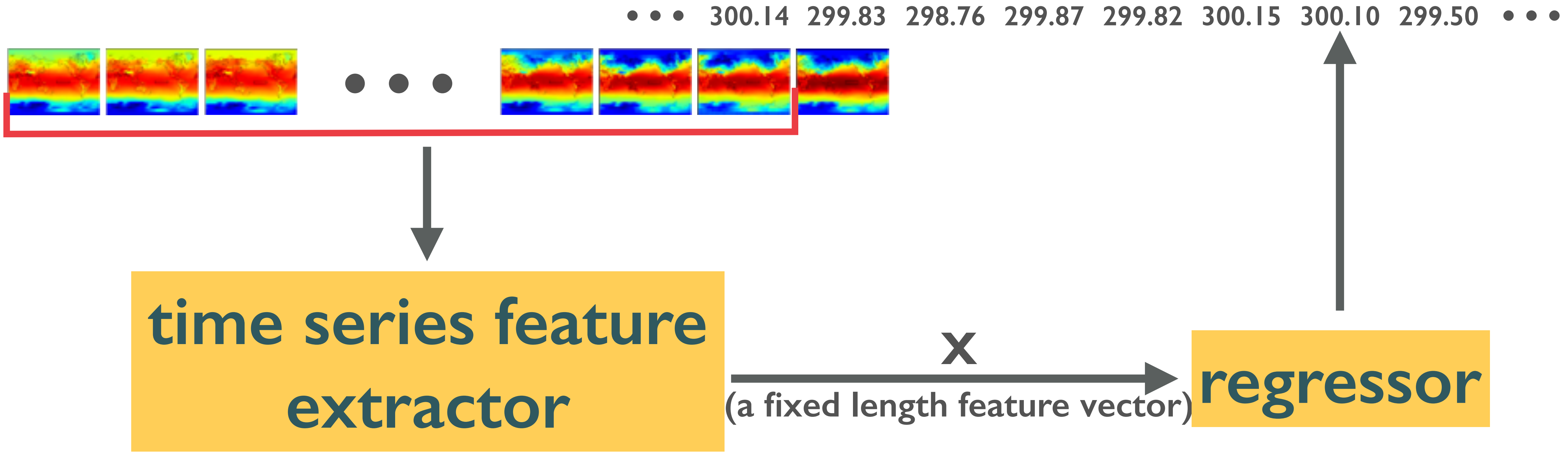


Classifying and quantifying monoclonal antibody preparations for cancer therapy using machine learning

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

Forecasting El Nino six months ahead



• • • 300.14 299.83 298.76 299.87 299.82 300.15 300.10 299.50 • • •

• • •

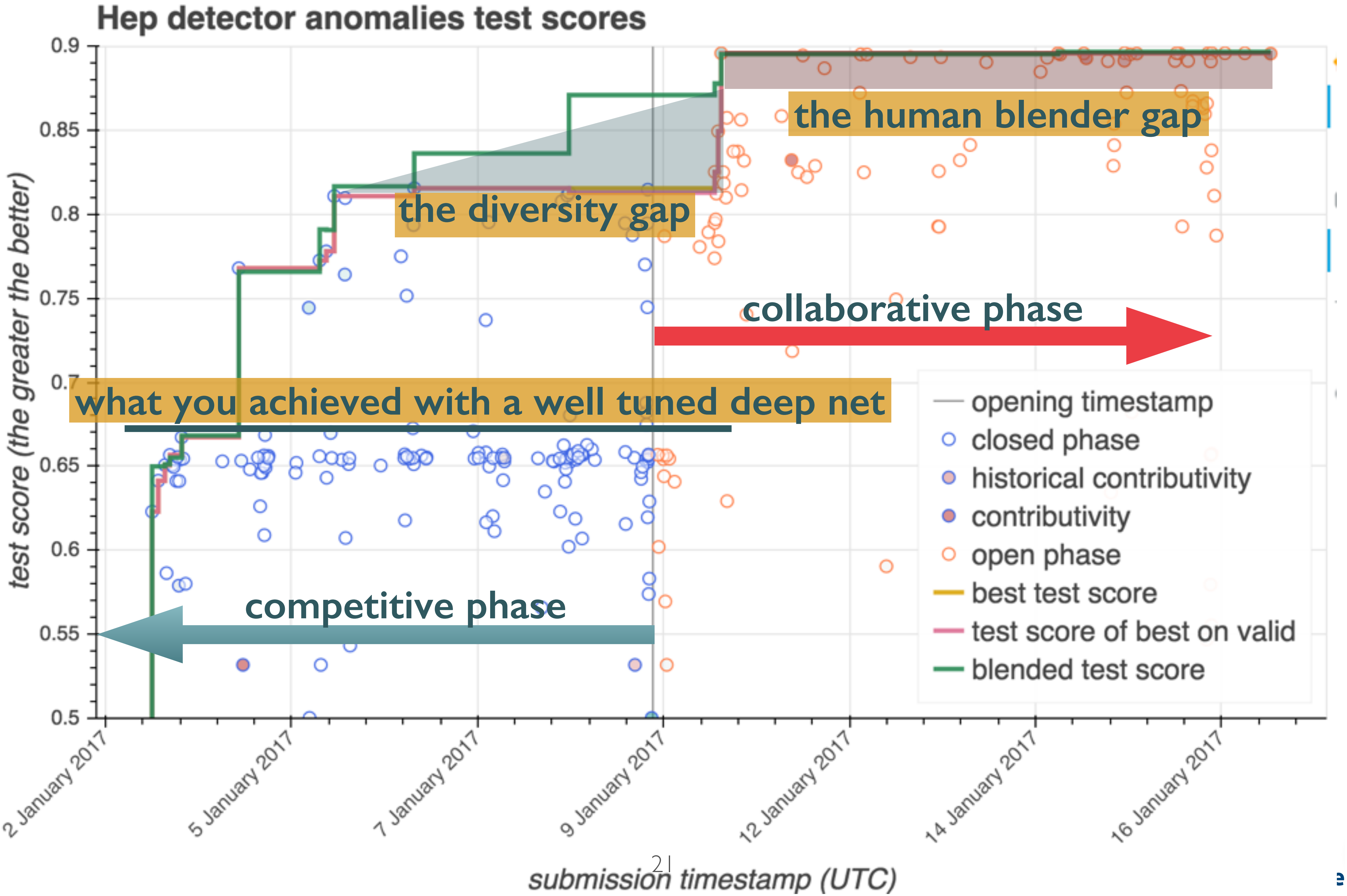
time series feature extractor

x

(a fixed length feature vector)

regressor

The power of the (collaborating) crowd



the human blender gap

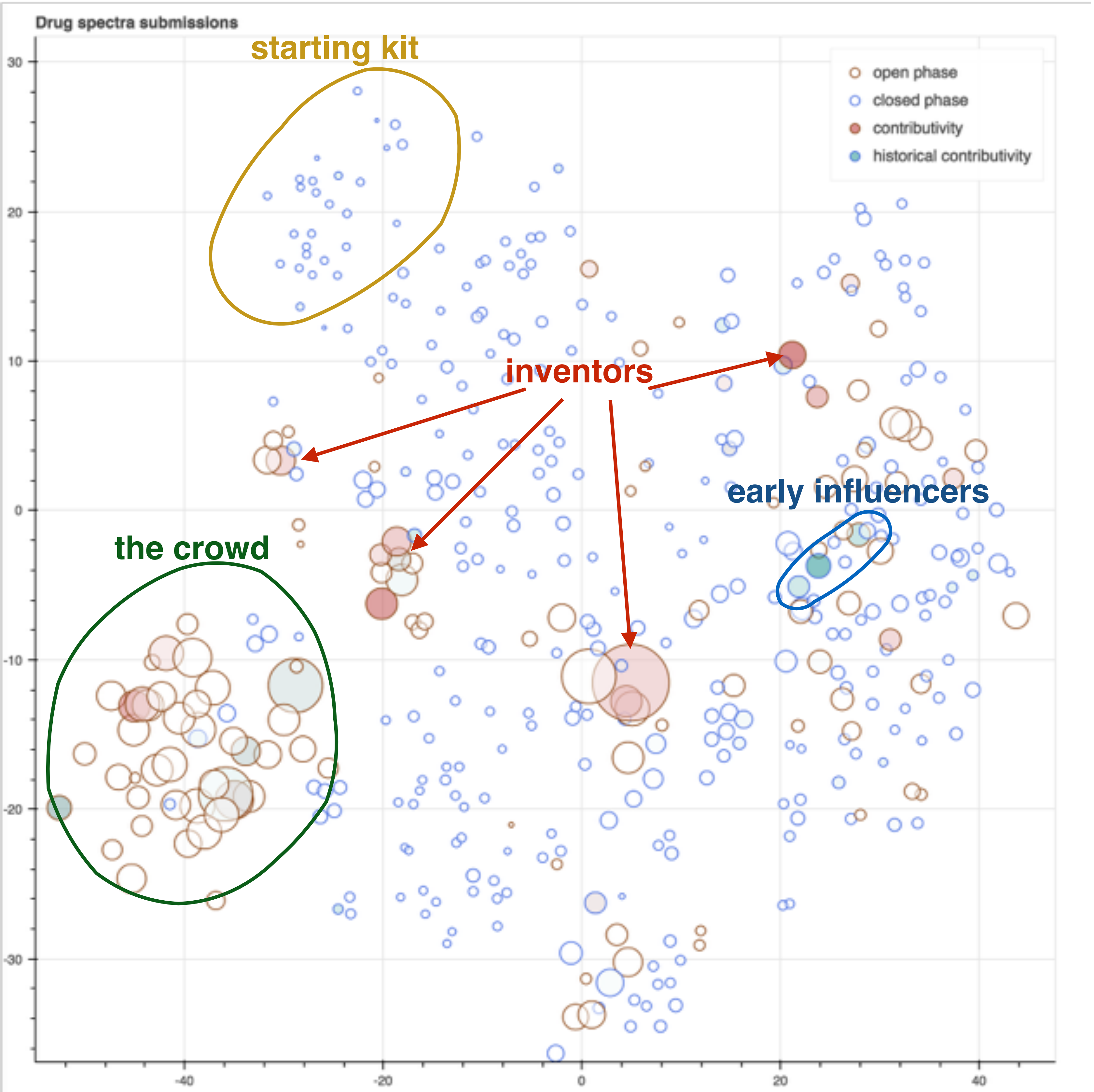
the diversity gap

collaborative phase

what you achieved with a well tuned deep net

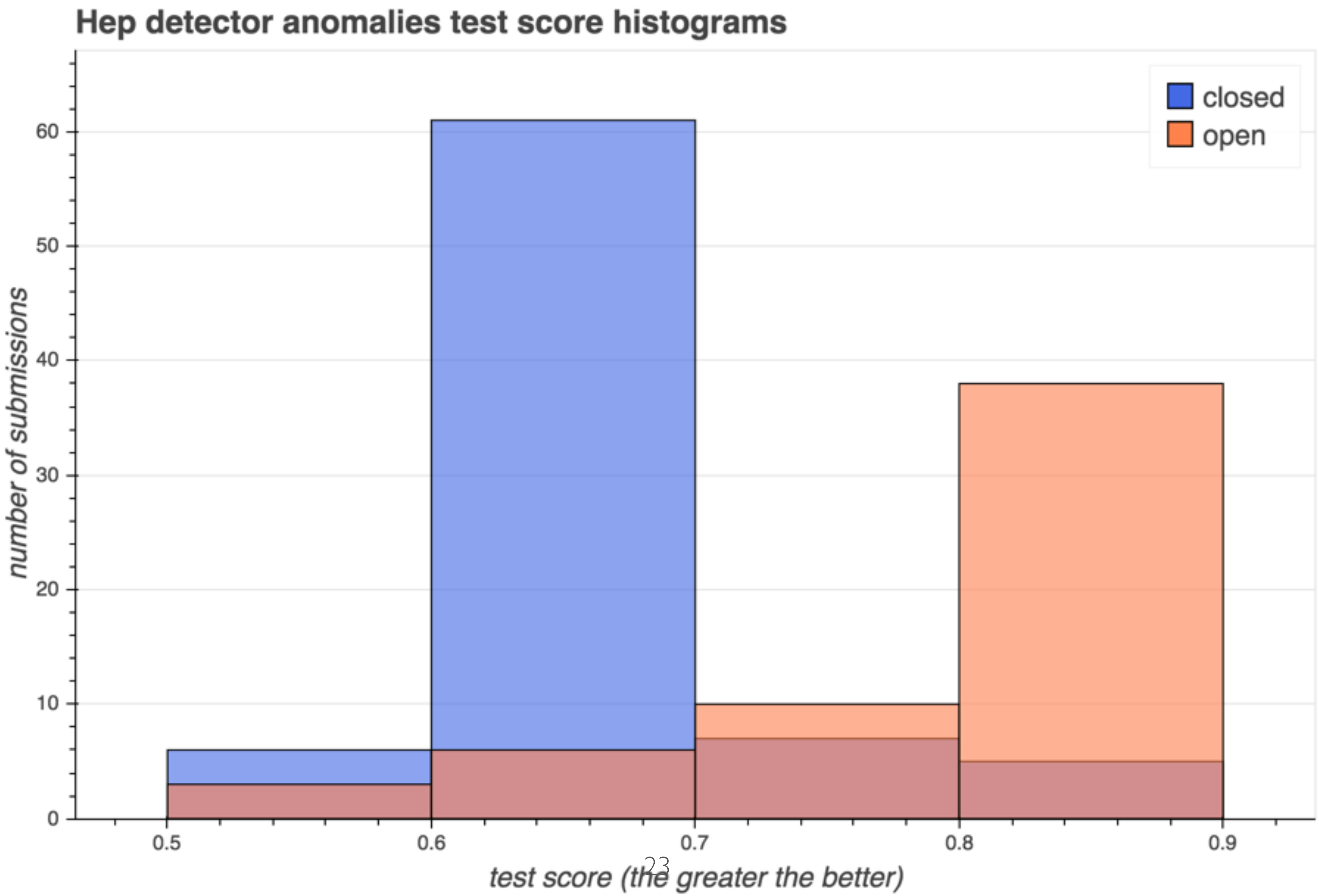
competitive phase

THE DYNAMICS OF COLLABORATION

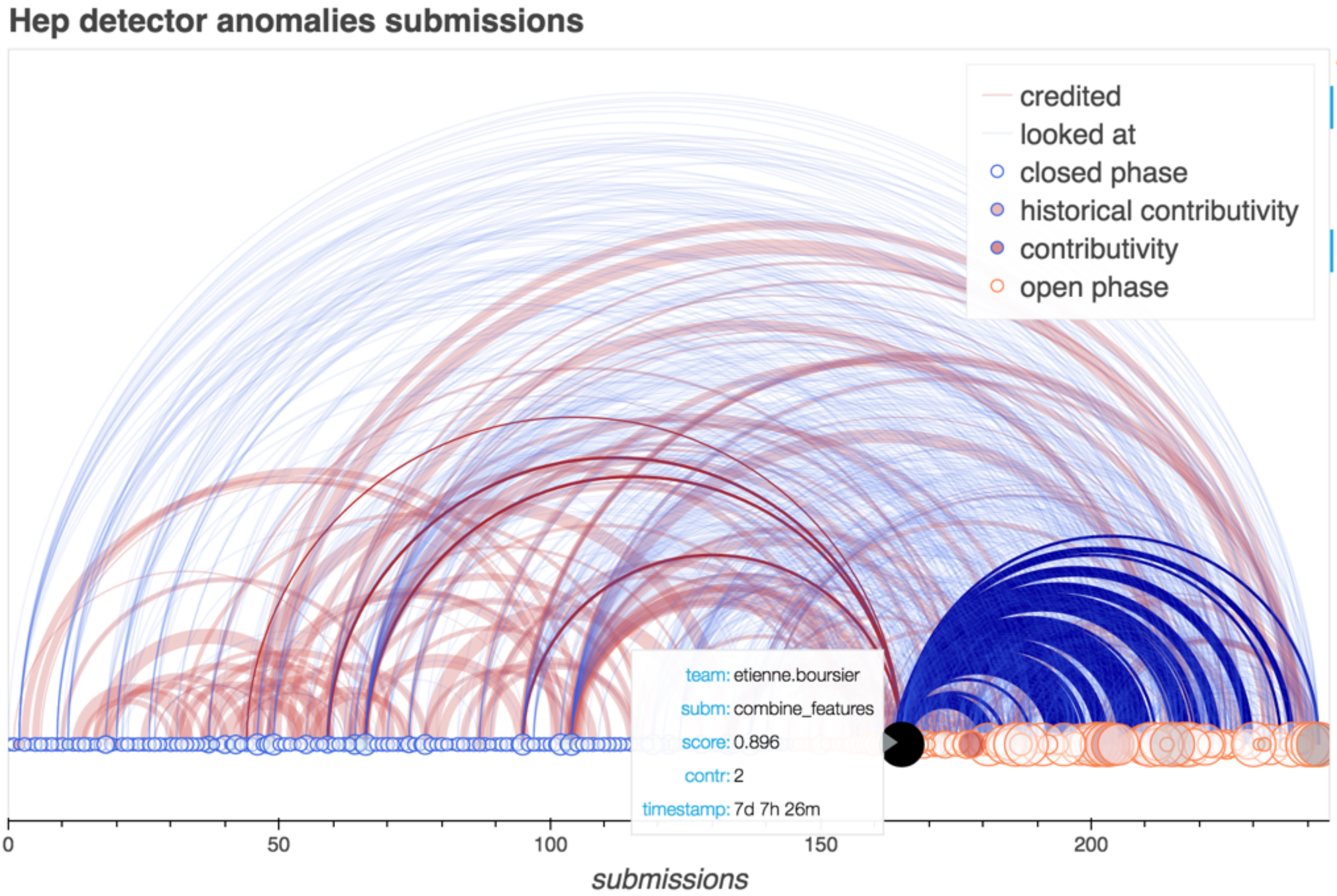


OPEN PHASE LETS PARTICIPANTS CATCH UP

THE GOAL OF TEACHING



COMMUNICATION AND REUSE



team: etienne.boursier
subm: combine_features
score: 0.896
contr: 2
timestamp: 7d 7h 26m

CLASSIFY POLLENATING INSECTS

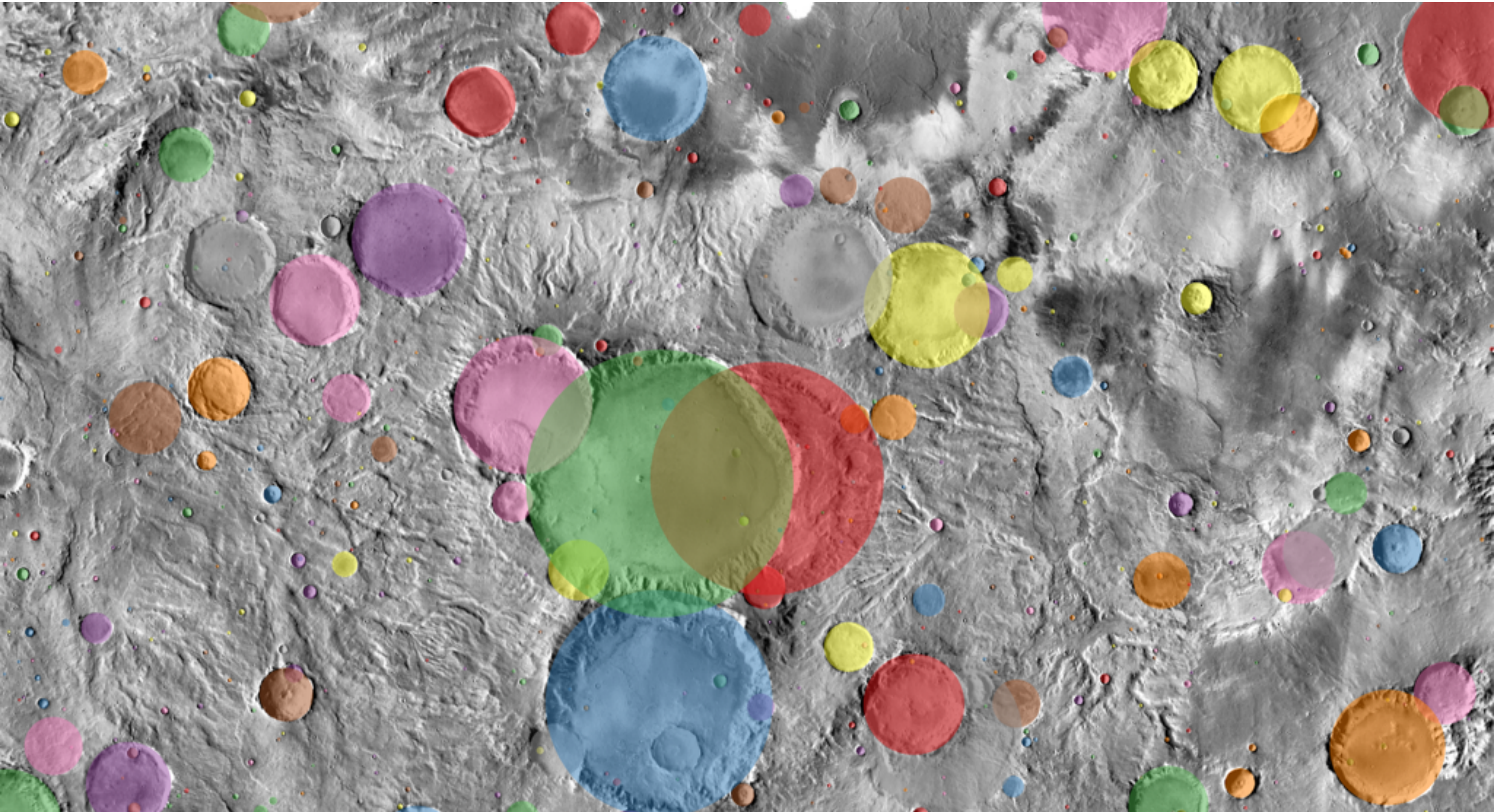
https://www.ramp.studio/events/pollenating_insects_3_JNI_2017

4.5K€ for the competitive phase

3K€ for the collaborative phase

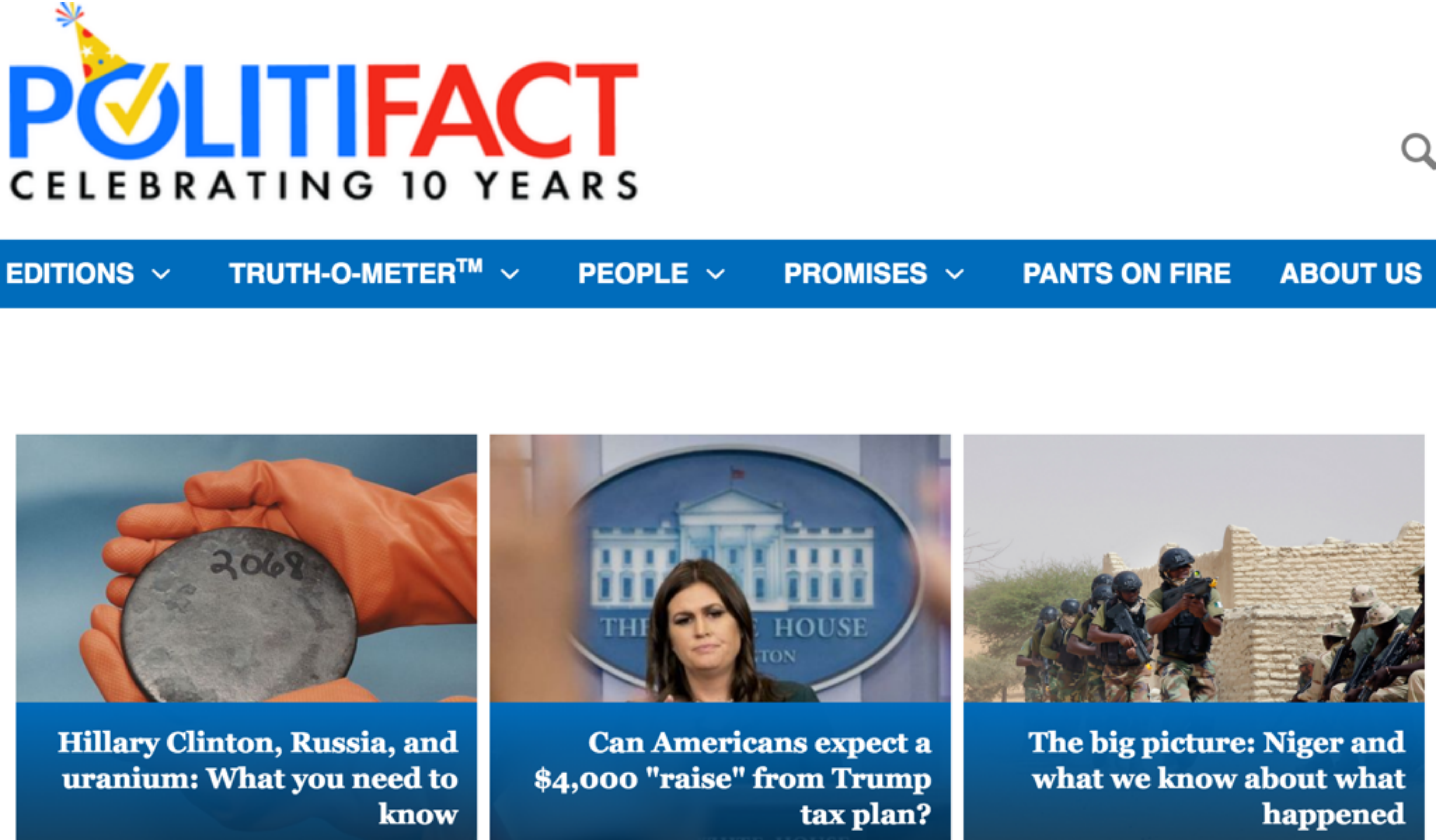
50 GPU hours per participant

Detecting Mars craters



FAKE NEWS

Predict the truthfulness of news

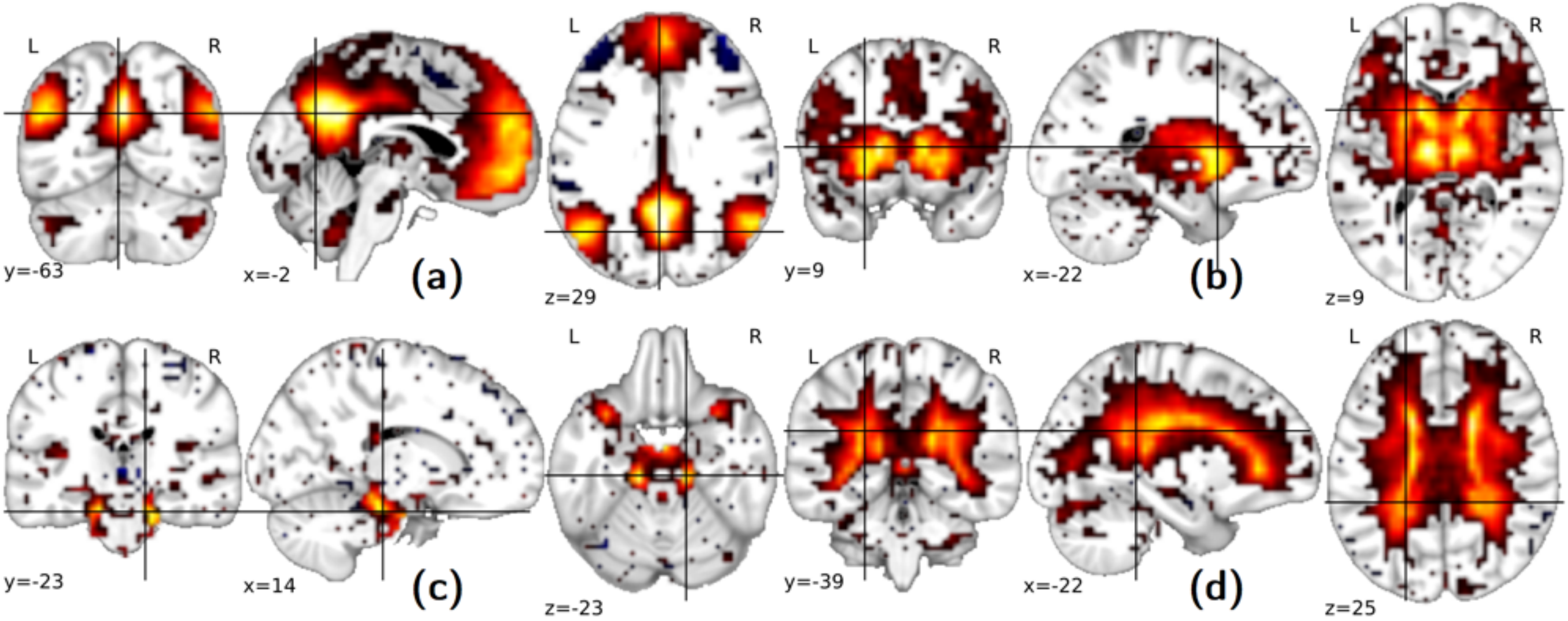


Hillary Clinton, Russia, and uranium: What you need to know

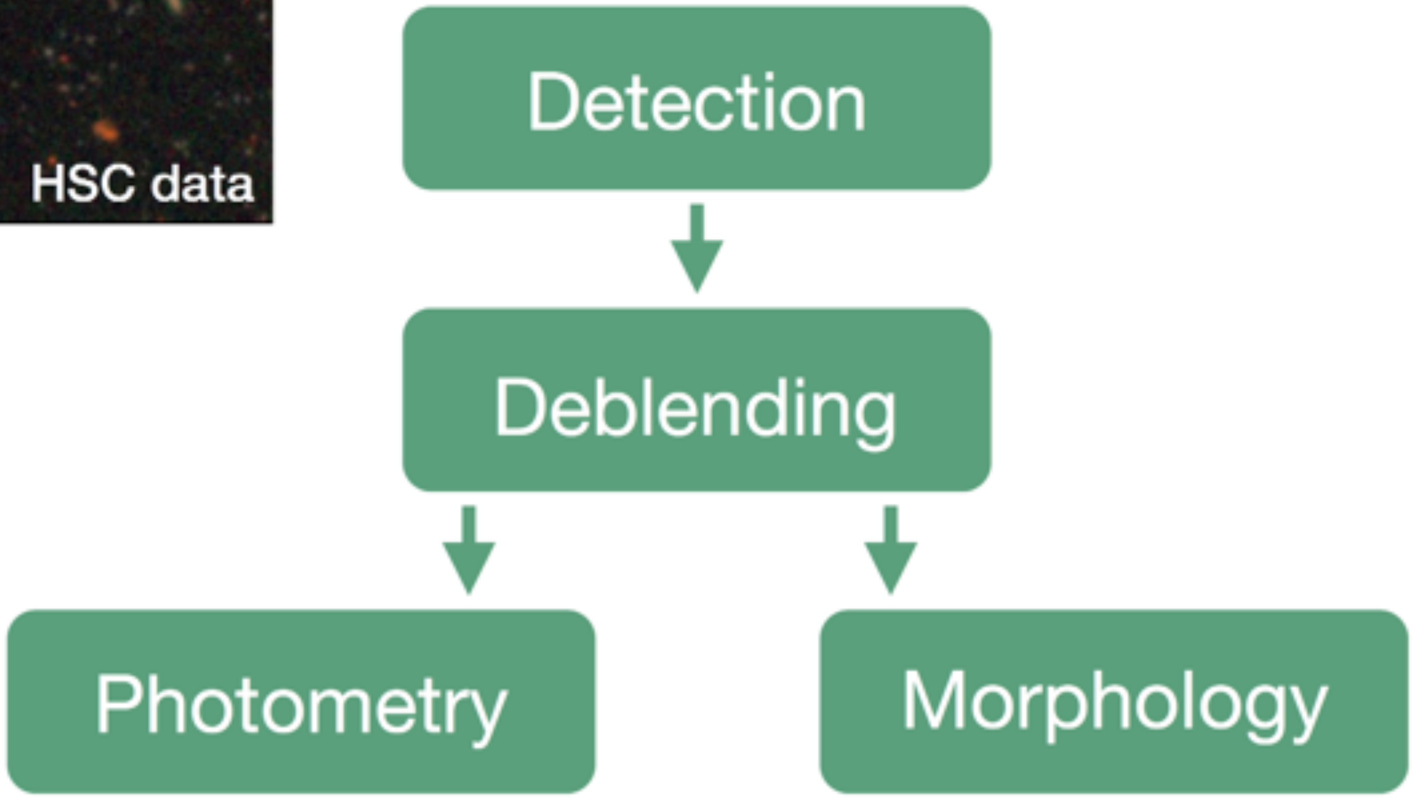
Can Americans expect a $4,000 "raise" from Trump tax plan?

The big picture: Niger and what we know about what happened

Upcoming challenge
Predict autism from brain scans



Upcoming challenge(s)
Astroimaging



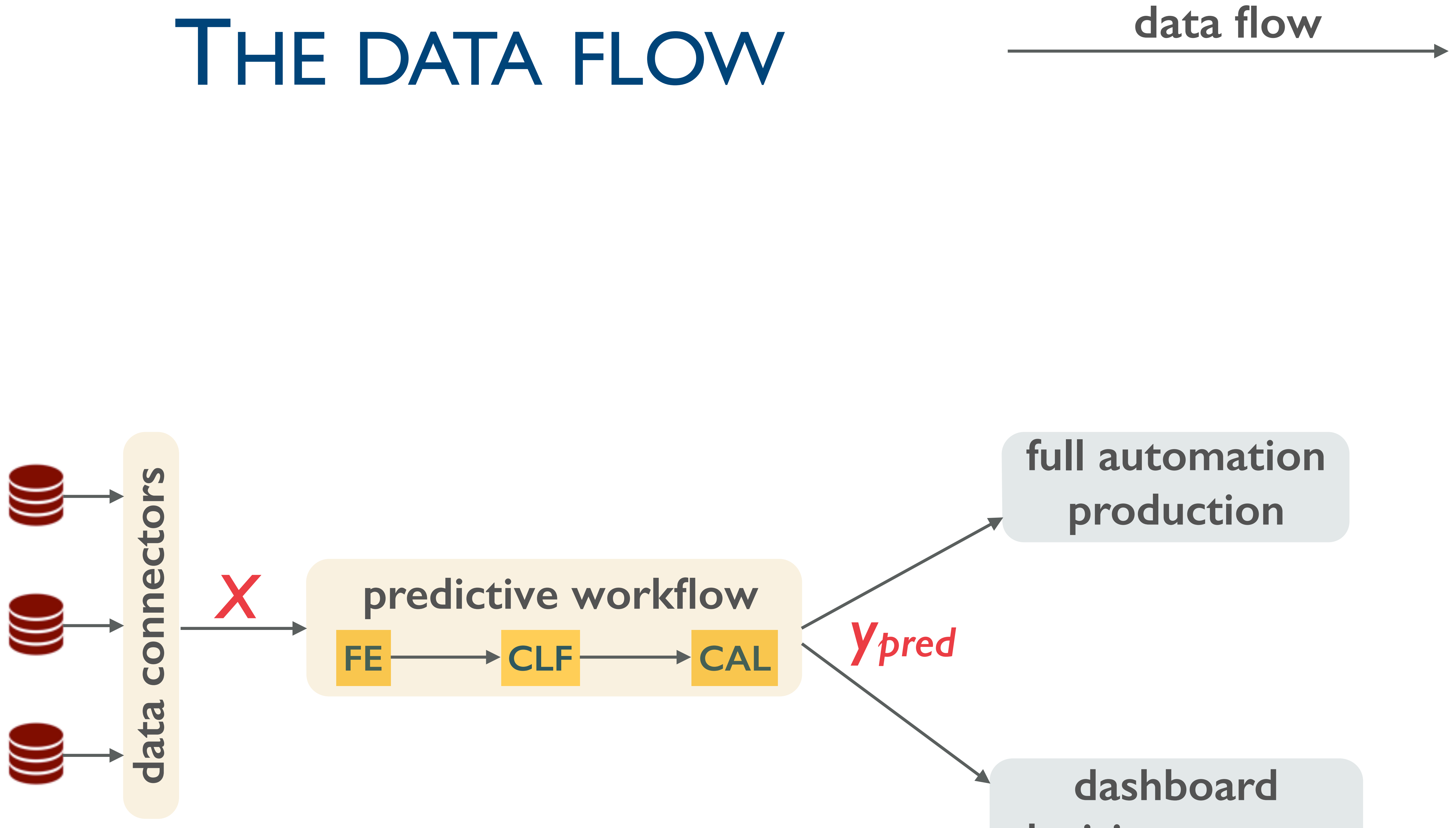
Detection

Deblending

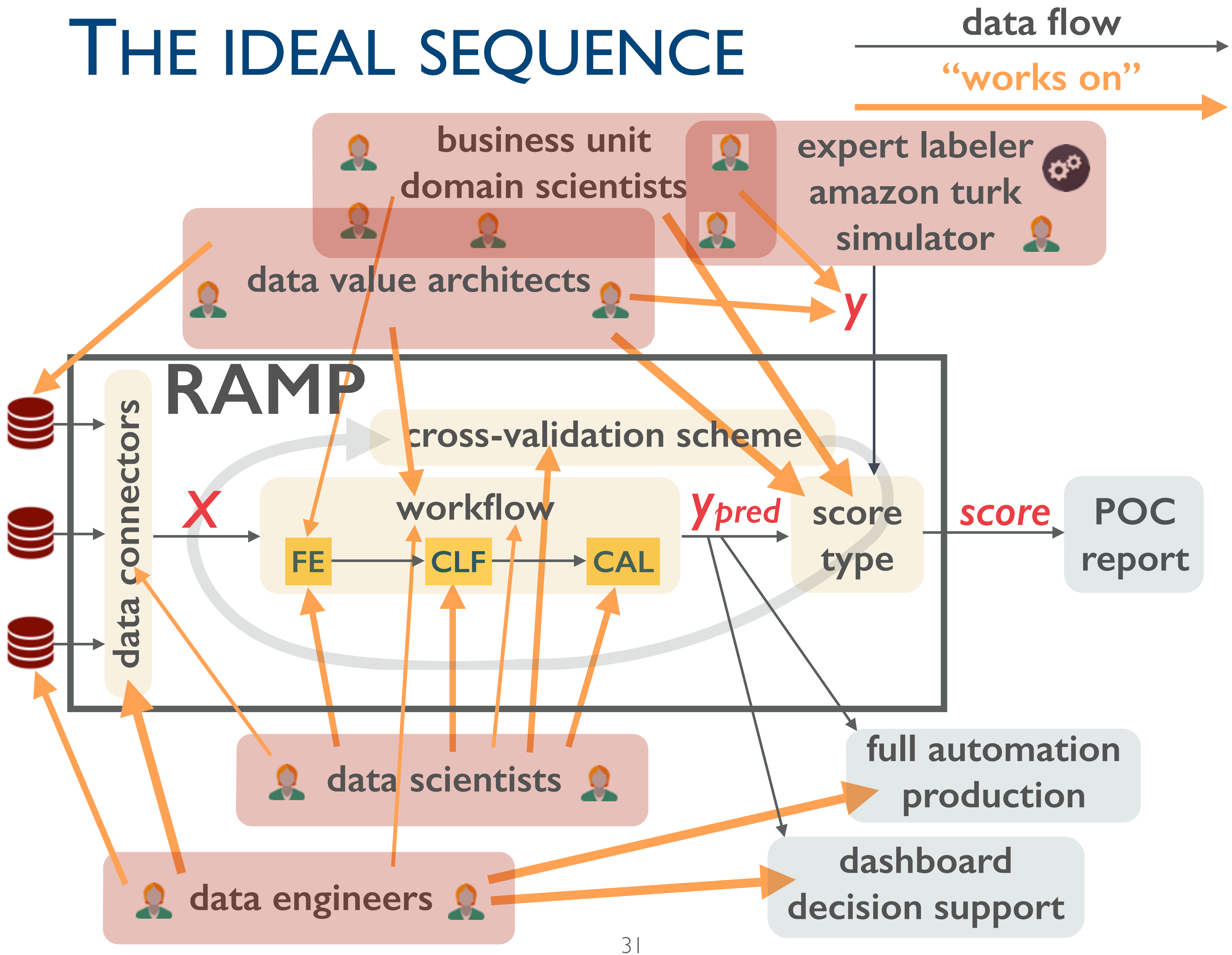
Photometry

Morphology

THE DATA FLOW



THE IDEAL SEQUENCE



Before solving the problem,

set it up

(even put it into production)

Setting up the RAMP
~~is~~ was
long and hard.

Separate workflow building and workflow optimization

RAMP

~~Data challenges with~~

~~modularization and code submission~~

Frugal data science

process management

Balázs Kégl

Université Paris-Saclay / CNRS

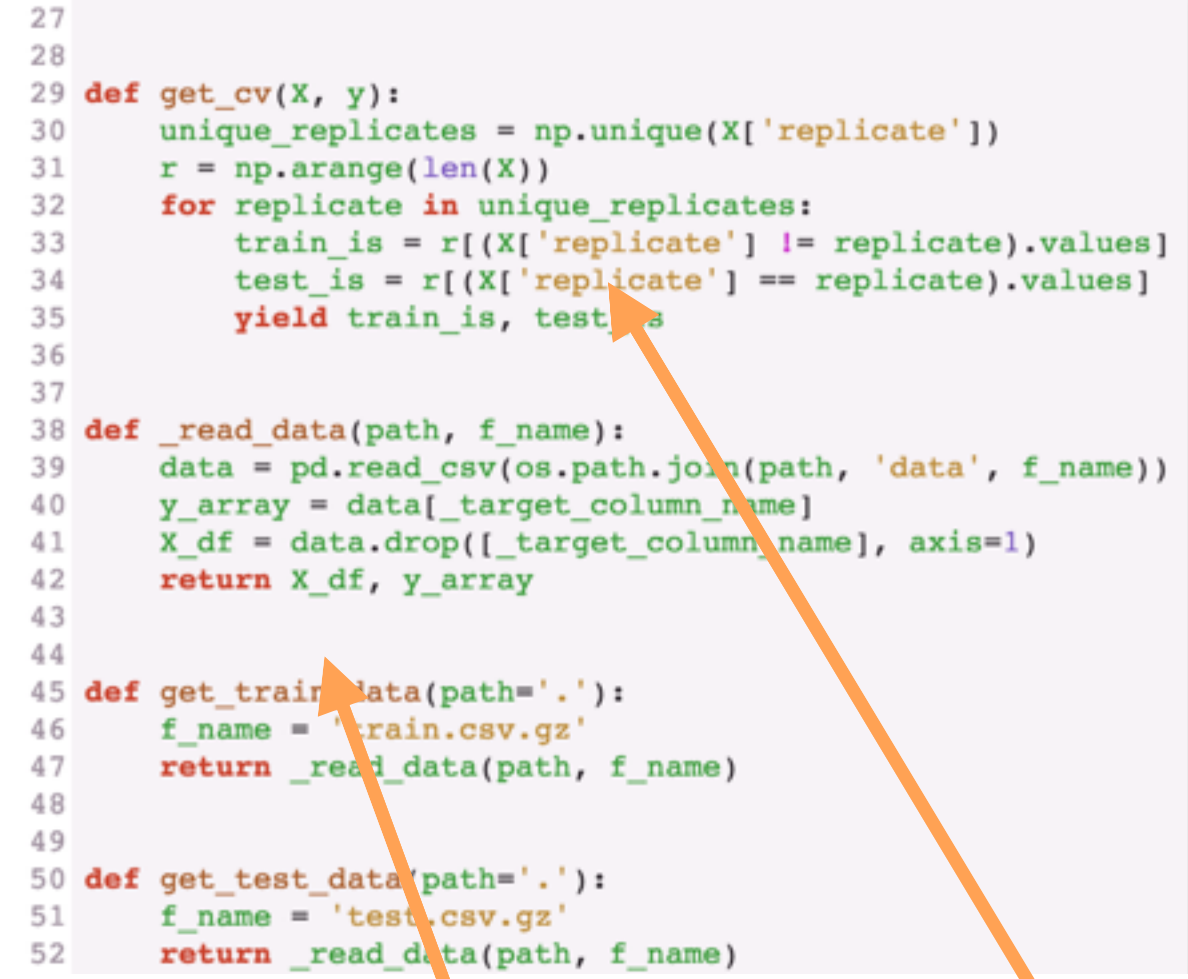
universite PARIS-SACLAY

Paris-Saclay Center for Data Science

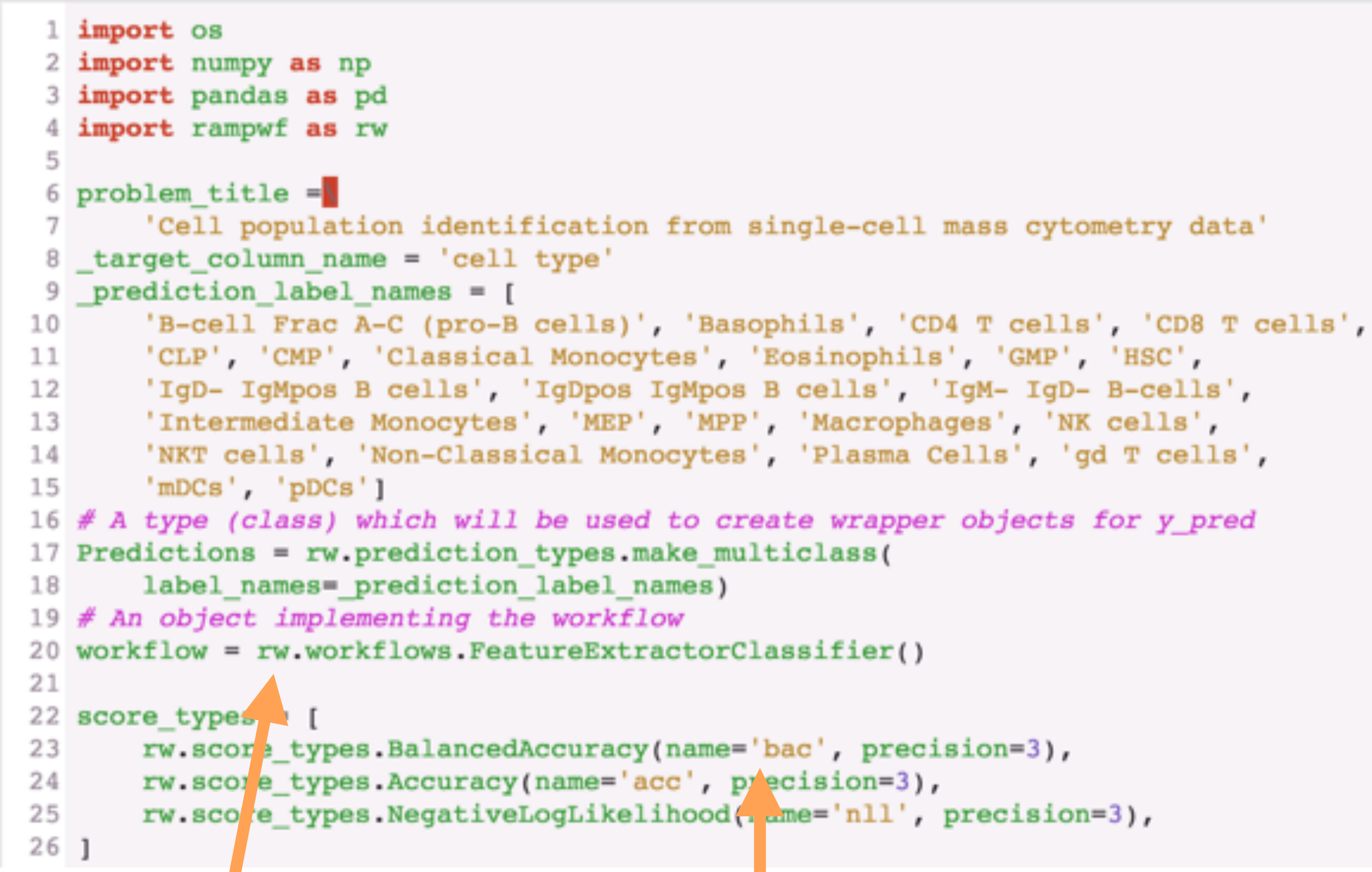
RAMP-WORKFLOW & RAMP-KITS

- toolkit: https://github.com/paris-saclay-cds/ramp-workflow
  - for designing workflows
  - set of ready-made metrics, workflows, CV schemes, data readers
  - unique command-line test script
- examples: https://github.com/ramp-kits
  - a zoo of problems, experiments, workflows
  - (at least) one initial solution

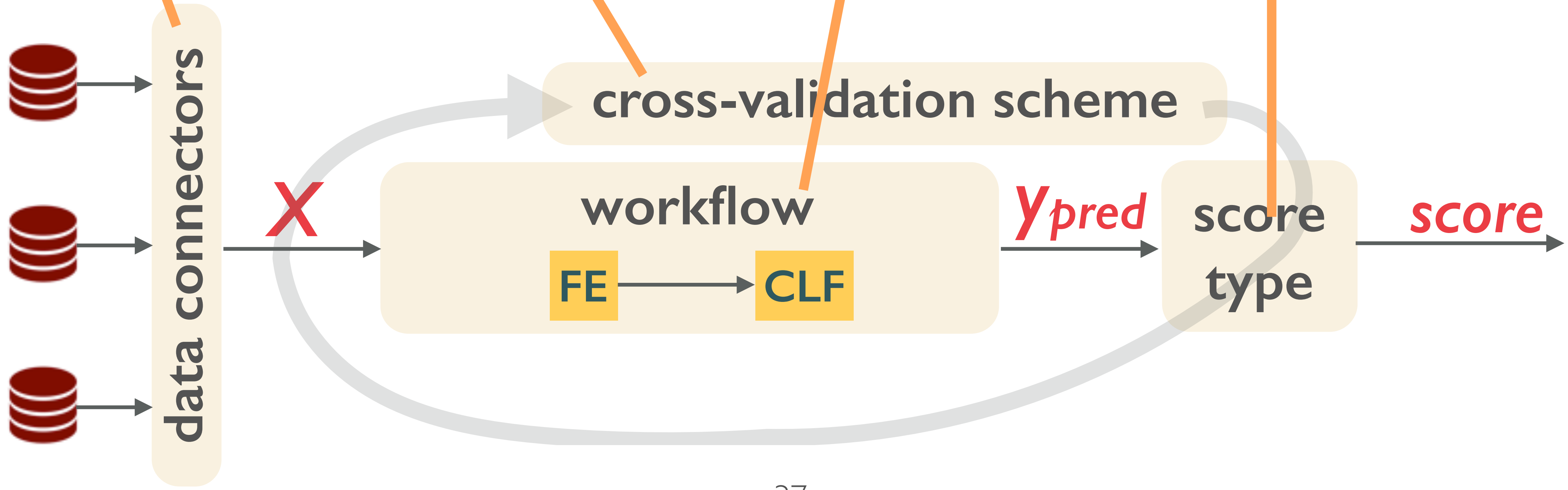
A SINGLE SCRIPT TO DEFINE THE BUNDLE



```python


def get_cv(X, y):
    unique_replicates = np.unique(X['replicate'])
    r = np.arange(len(X))
    for replicate in unique_replicates:
        train_is = r[(X['replicate'] != replicate).values]
        test_is = r[(X['replicate'] == replicate).values]
        yield train_is, test_is


def _read_data(path, f_name):
    data = pd.read_csv(os.path.join(path, 'data', f_name))
    y_array = data[_target_column_name]
    X_df = data.drop([_target_column_name], axis=1)
    return X_df, y_array


def get_train_data(path='.'):
    f_name = 'train.csv.gz'
    return _read_data(path, f_name)


def get_test_data(path='.'):
    f_name = 'test.csv.gz'
    return _read_data(path, f_name)
```

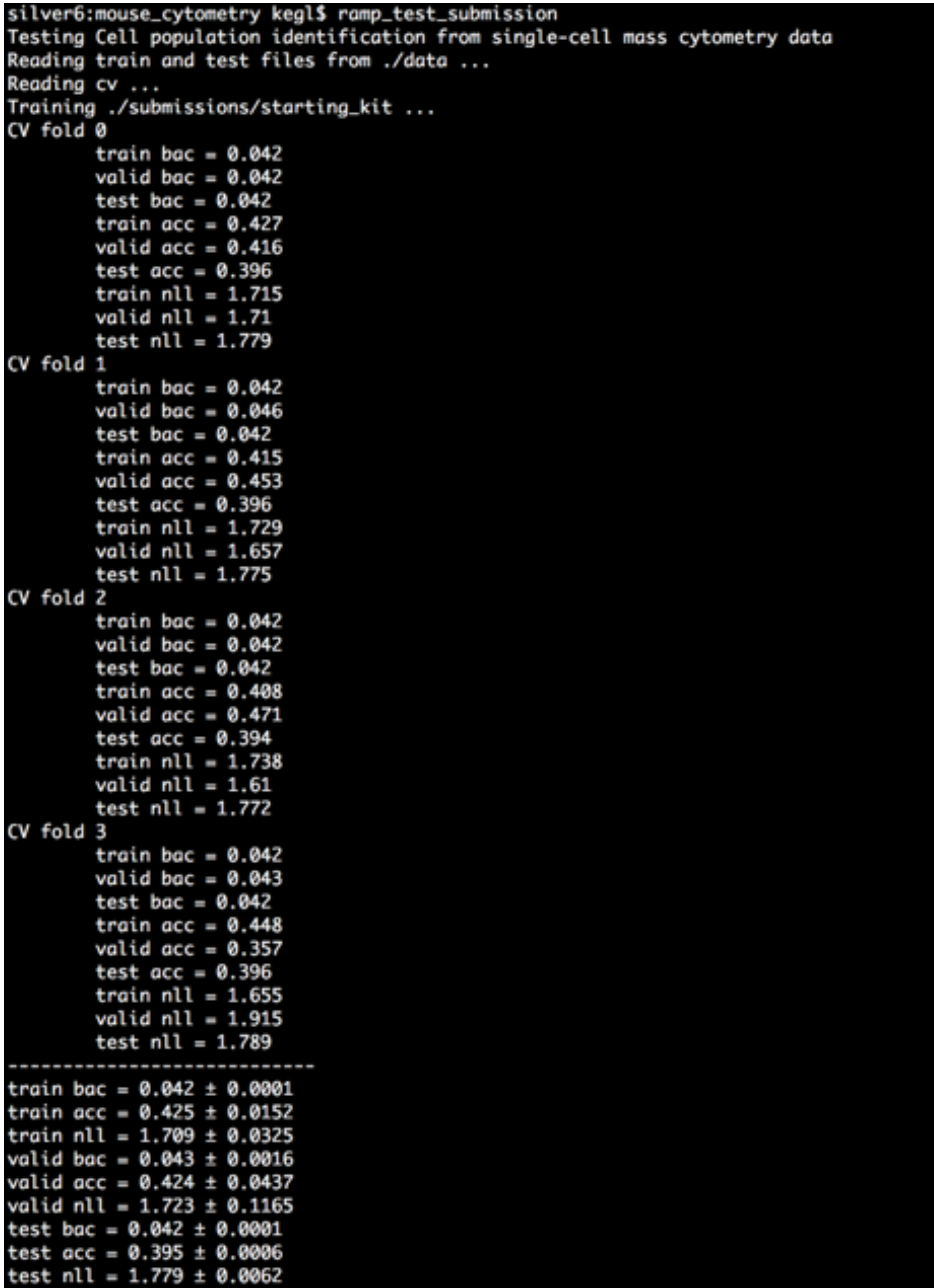


```python
import os
import numpy as np
import pandas as pd
import rampwf as rw

problem_title = \
    'Cell population identification from single-cell mass cytometry data'
_target_column_name = 'cell type'
_prediction_label_names = [
    'B-cell Frac A-C (pro-B cells)', 'Basophils', 'CD4 T cells', 'CD8 T cells',
    'CLP', 'CMP', 'Classical Monocytes', 'Eosinophils', 'GMP', 'HSC',
    'IgD- IgMpos B cells', 'IgDpos IgMpos B cells', 'IgM- IgD- B-cells',
    'Intermediate Monocytes', 'MEP', 'MPP', 'Macrophages', 'NK cells',
    'NKT cells', 'Non-Classical Monocytes', 'Plasma Cells', 'gd T cells',
    'mDCs', 'pDCs']
# A type (class) which will be used to create wrapper objects for y_pred
Predictions = rw.prediction_types.make_multiclass(
    label_names=_prediction_label_names)
# An object implementing the workflow
workflow = rw.workflows.FeatureExtractorClassifier()

score_types = [
    rw.score_types.BalancedAccuracy(name='bac', precision=3),
    rw.score_types.Accuracy(name='acc', precision=3),
    rw.score_types.NegativeLogLikelihood(name='nll', precision=3),
]
```



A SINGLE EXECUTABLE TO TEST THE SUBMISSIONS

- Keep your different submissions in a **simple file structure**
- Communicate them on **git**
- Execute them also from the **notebook**



```
silver6:mouse_cytometry kegl$ ramp_test_submission
Testing Cell population identification from single-cell mass cytometry data
Reading train and test files from ./data ...
Reading cv ...
Training ./submissions/starting_kit ...
CV fold 0
        train bac = 0.042
        valid bac = 0.042
        test bac = 0.042
        train acc = 0.427
        valid acc = 0.416
        test acc = 0.396
        train nll = 1.715
        valid nll = 1.71
        test nll = 1.779
CV fold 1
        train bac = 0.042
        valid bac = 0.046
        test bac = 0.042
        train acc = 0.415
        valid acc = 0.453
        test acc = 0.396
        train nll = 1.729
        valid nll = 1.657
        test nll = 1.775
CV fold 2
        train bac = 0.042
        valid bac = 0.042
        test bac = 0.042
        train acc = 0.408
        valid acc = 0.471
        test acc = 0.394
        train nll = 1.738
        valid nll = 1.61
        test nll = 1.772
CV fold 3
        train bac = 0.042
        valid bac = 0.043
        test bac = 0.042
        train acc = 0.448
        valid acc = 0.357
        test acc = 0.396
        train nll = 1.655
        valid nll = 1.915
        test nll = 1.789
----------------------------
train bac = 0.042 ± 0.0001
train acc = 0.425 ± 0.0152
train nll = 1.709 ± 0.0325
valid bac = 0.043 ± 0.0016
valid acc = 0.424 ± 0.0437
valid nll = 1.723 ± 0.1165
test bac = 0.042 ± 0.0001
test acc = 0.395 ± 0.0006
test nll = 1.779 ± 0.0062
```

You can

1. Participate in upcoming RAMPs
2. Use RAMP in teaching or training
3. Use the toolkit for your own workflows
4. Submit it to us if you want to run a data challenge

frontend:

www.ramp.studio

toolkit:

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

examples:

github.com/ramp-kits

slack:

ramp-studio.slack.com

blogs:

medium.com/@balazskegl

mail:

balazs.kegl@gmail.com