RAMP

Data challenges with modularization and code submission

Lessons learned

Balázs Kégl

Université Paris-Saclay / CNRS

Who am I?

Balázs Kégl

- Directeur de recherche **CNRS**
  - machine learning (20 years)
    interfacing with particle physics (10 years)
- Director of the **Paris-Saclay Center for Data Science**
  - interfacing with biology, economy, climatology, chemistry, etc. (4 years)
- Data science **consulting and training** (4 years)

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
- How can **you use it**?
  - in a **classroom**: to **teach ML**
  - as a **domain scientist**: to **crowdsource your predictive problem**
  - as a **data scienctis**: to **benchmark your new techniques**

UNIVERSITÉ PARIS-SACLAY

19 *fondateurs*	**11 000** *chercheurs et enseignants-chercheurs*
60 000 *étudiants*	**300** *laboratoires*
6 000 *doctorants*	**8 000** *publications /an*
15 000 *étudiants en master*	**15 %** *de la recherche publique française*
8 *Schools*	**10** *départements*

+ horizontal multi-disciplinary and multi-partner initiatives to create cohesion

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

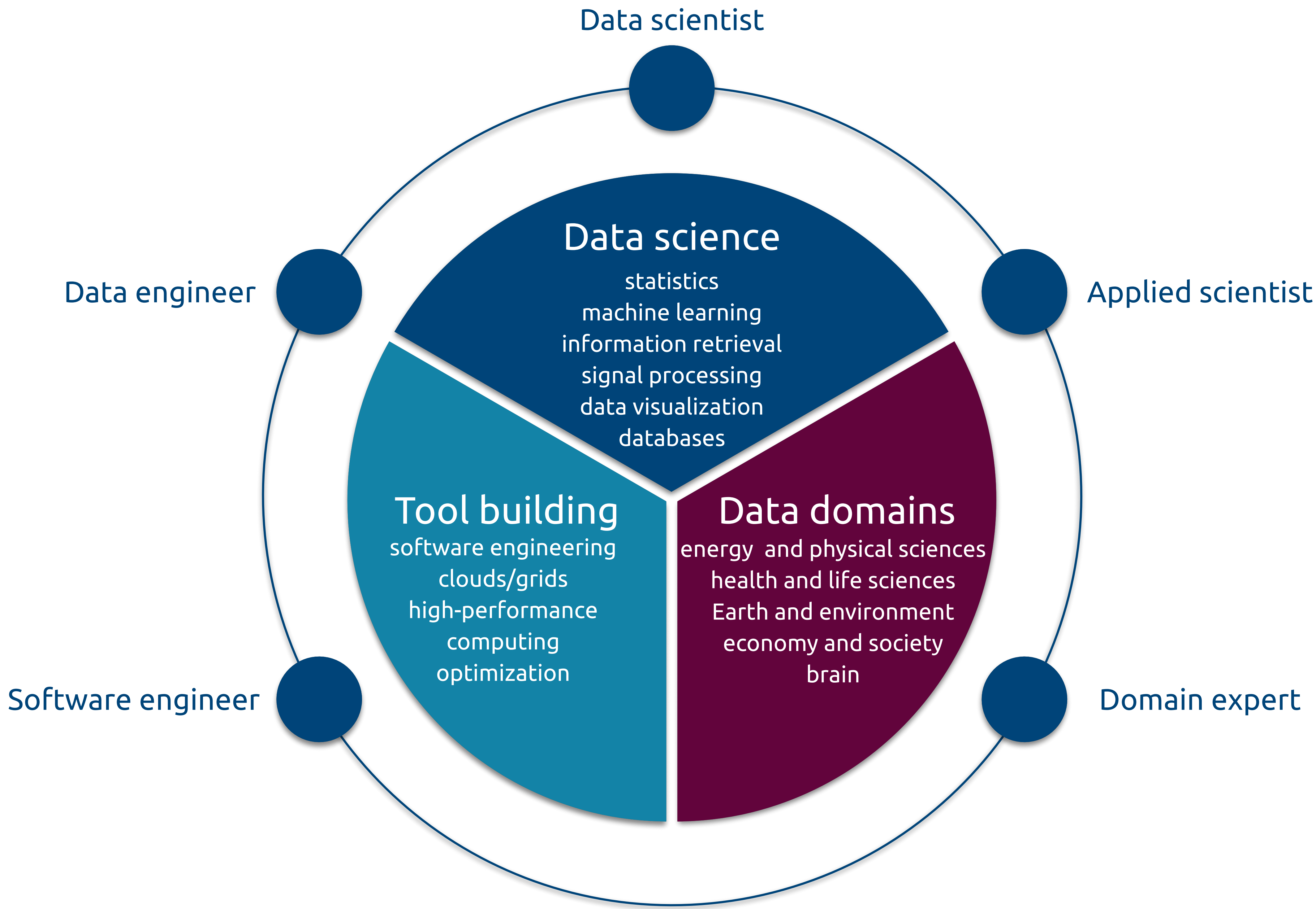
Visualization
INRIA
LIMSI

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

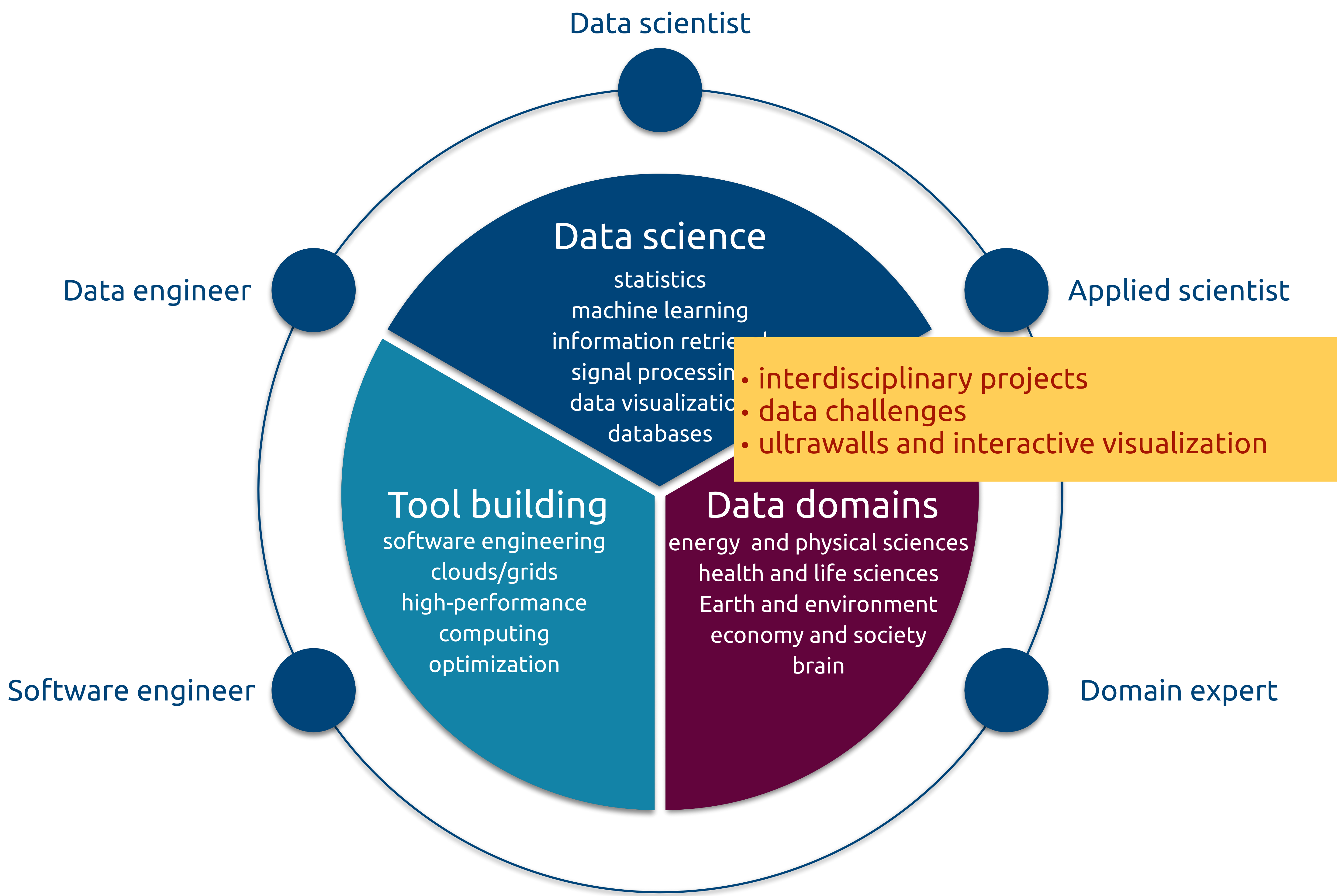
The data science ecosystem

https://medium.com/@balazskegl



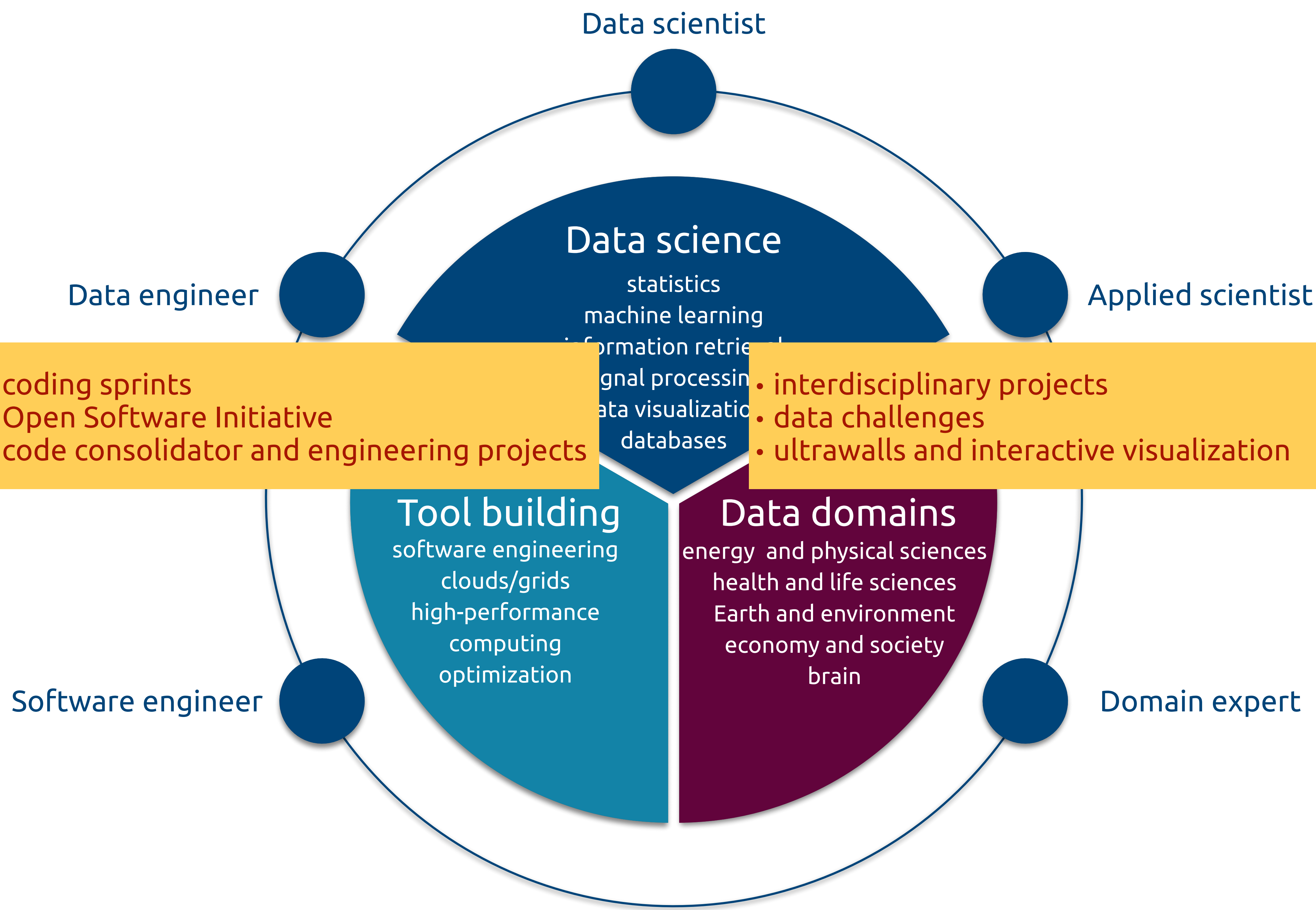
THE DATA SCIENCE ECOSYSTEM

https://medium.com/@balazskegl



THE DATA SCIENCE ECOSYSTEM

https://medium.com/@balazskegl

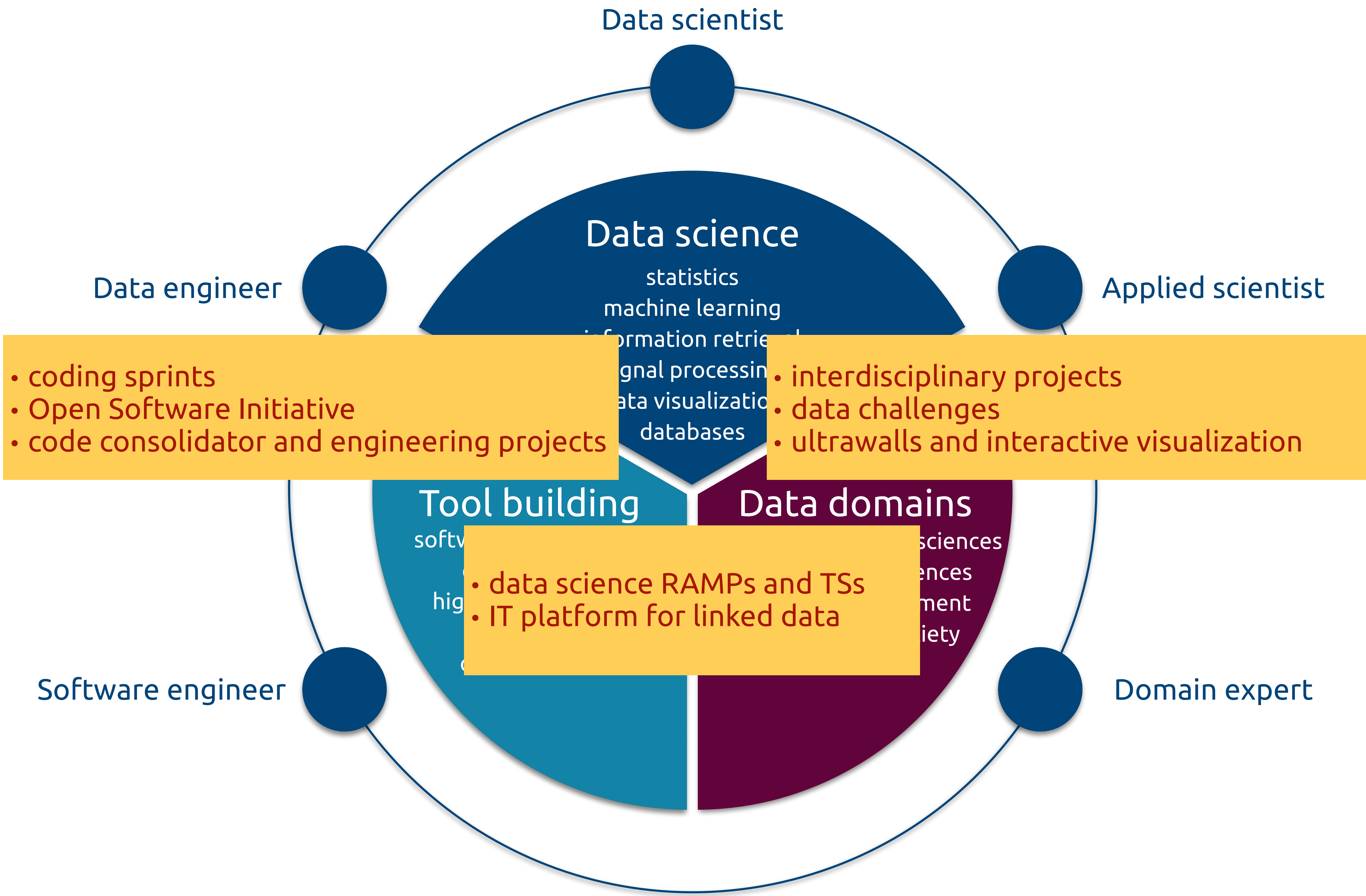


- coding sprints
- Open Software Initiative
- code consolidator and engineering projects

- interdisciplinary projects
- data challenges
- ultrawalls and interactive visualization

THE DATA SCIENCE ECOSYSTEM

https://medium.com/@balazskegl



Université Paris-Saclay
Paris-Saclay Center for Data Science

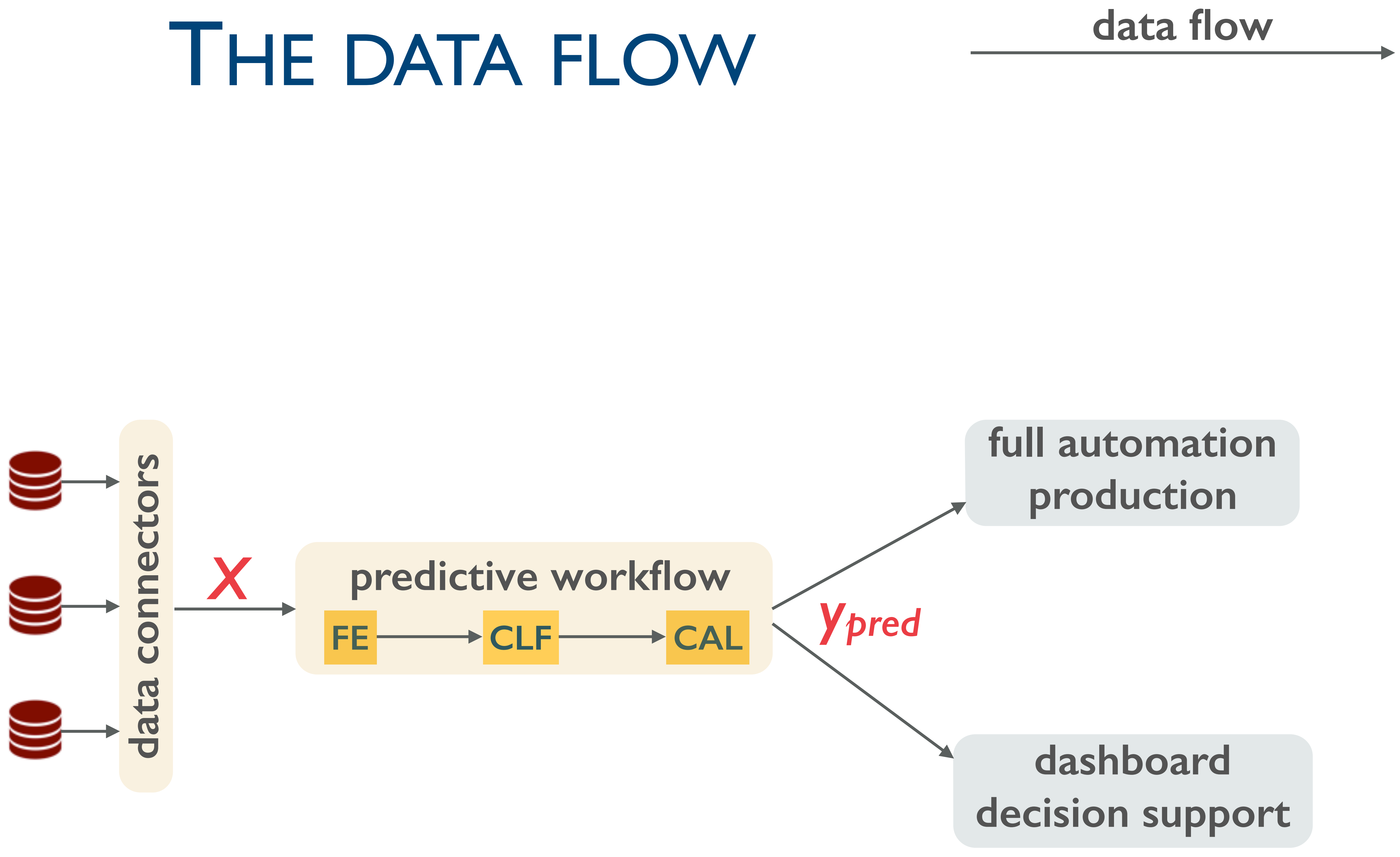
LACK OF TOOLS

- 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

GOAL

- Design a data science pipeline development tool that
  - hides heavy computational details and provides a simple interface to data scientists to experiment with algorithms
  - promotes collaboration and the rapid propagation of ideas
  - modularizes complex pipelines so the different expertise can be applied without having to understand all the details of the full workflow
  - allows the challenge organizer to walk away with a working prototype

THE DATA FLOW



THE IDEAL SEQUENCE

data flow

“works on”

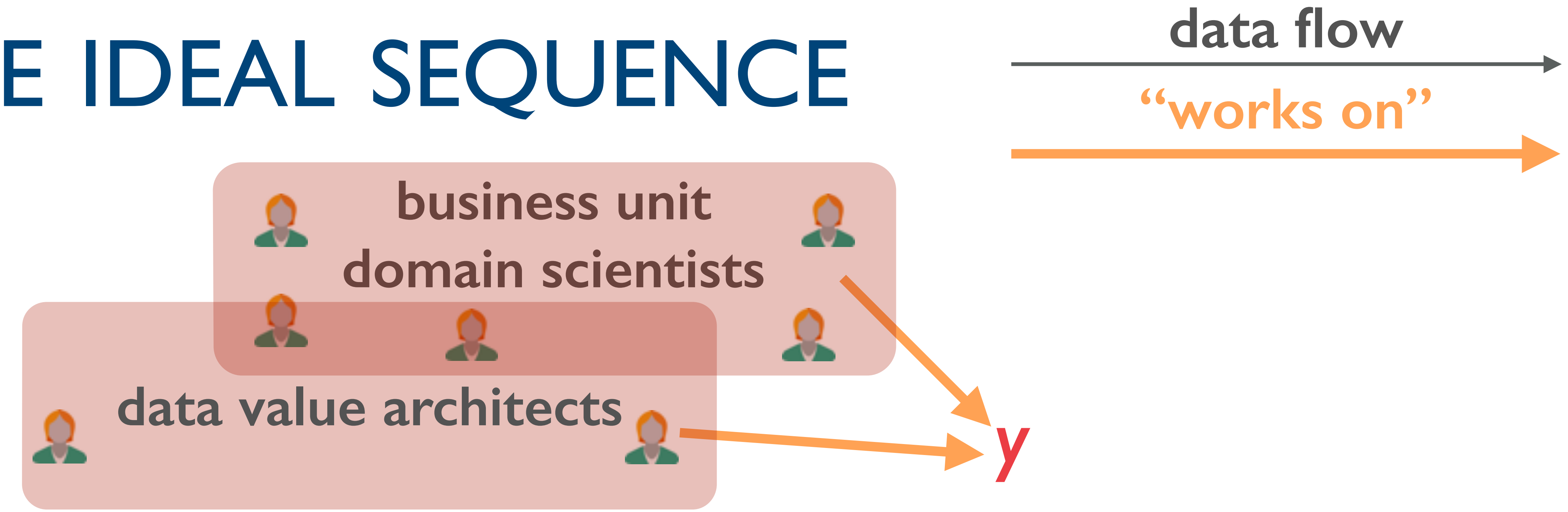
The ideal sequence

data flow

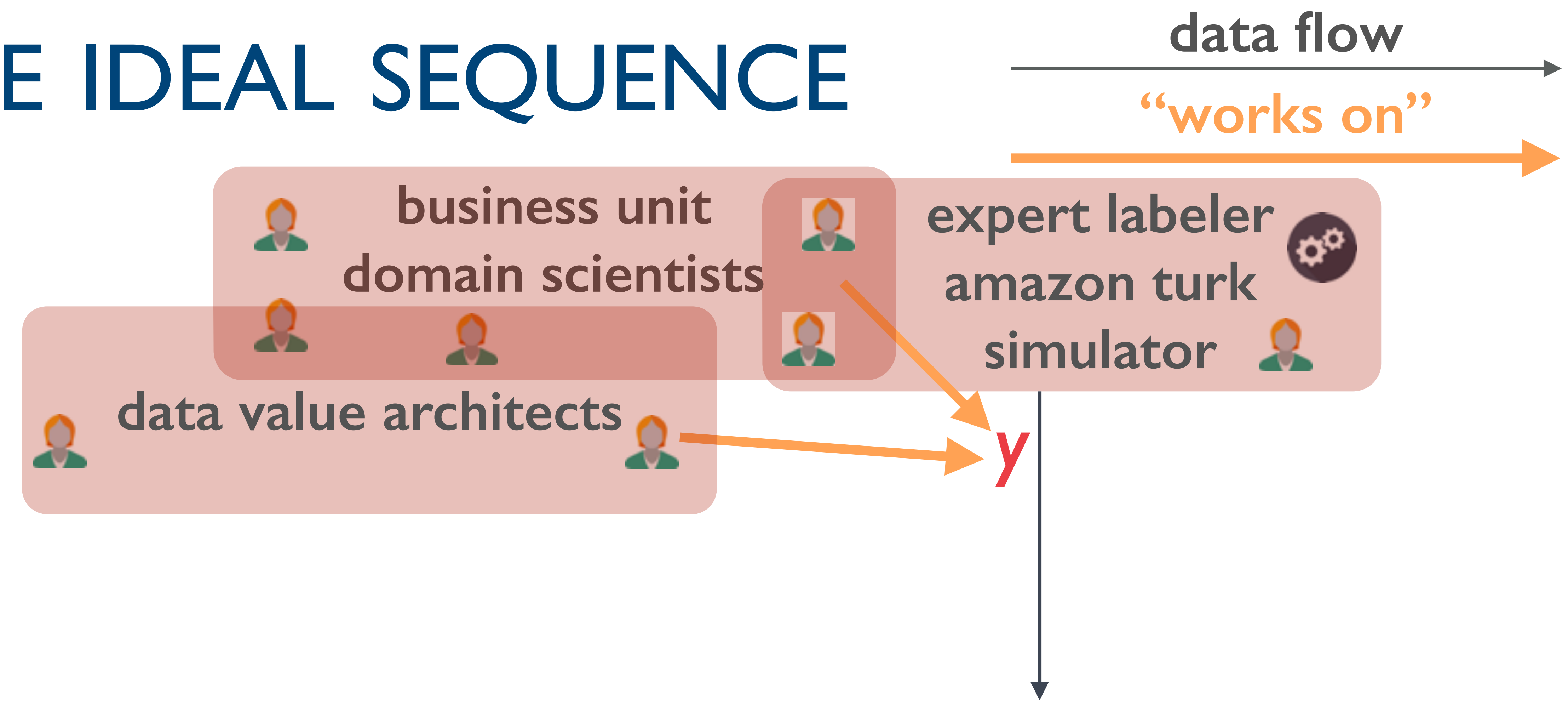
"works on"

y

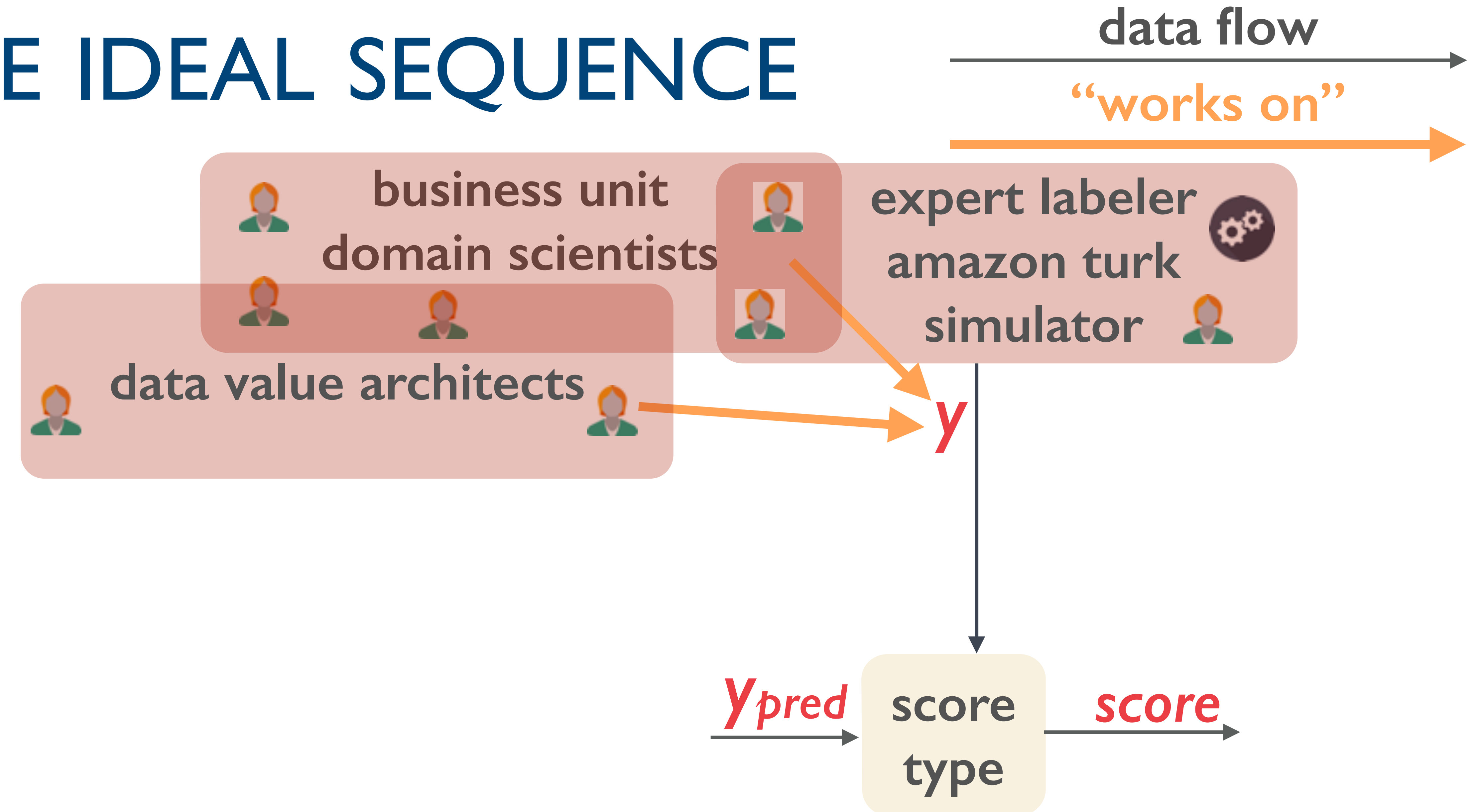
THE IDEAL SEQUENCE



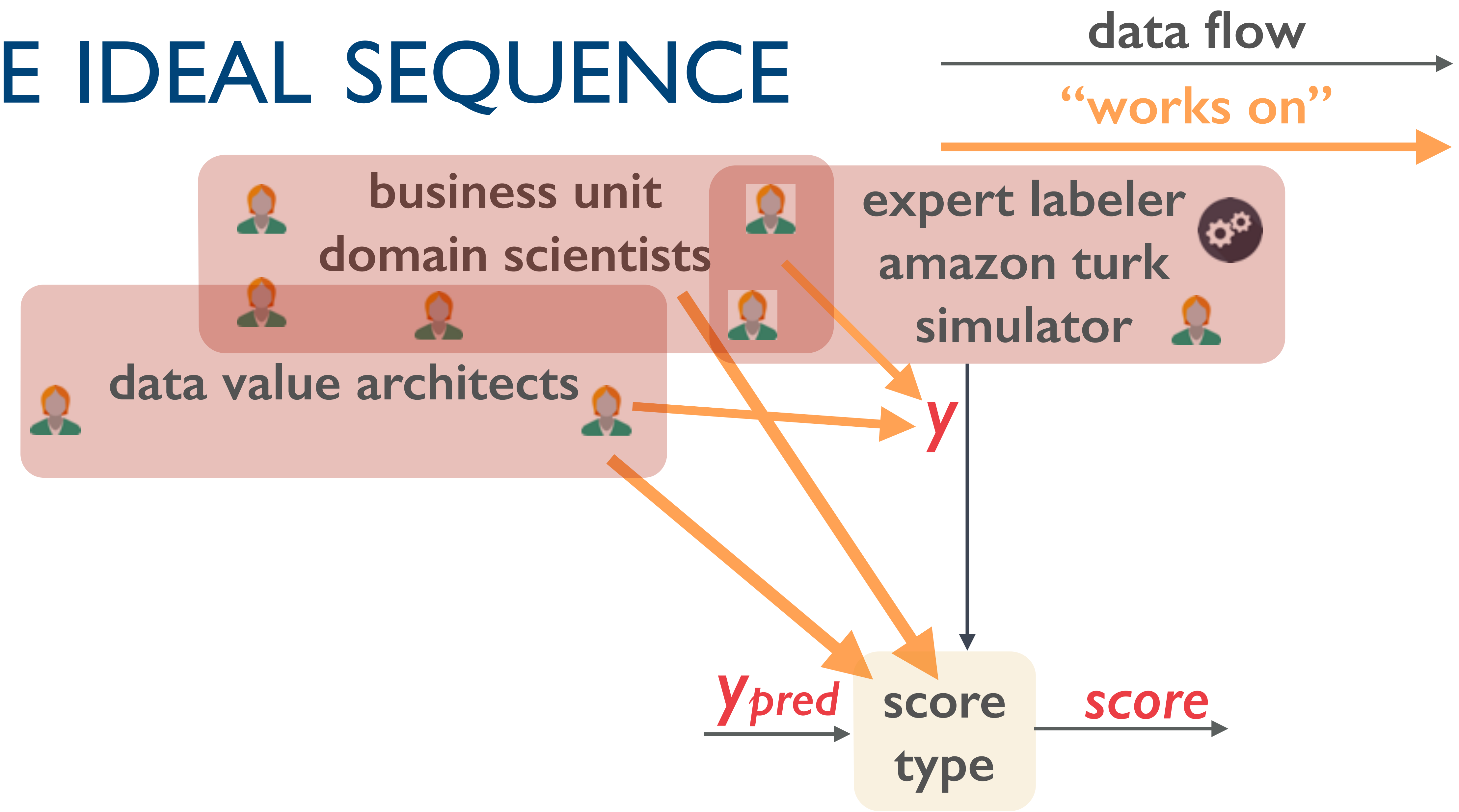
THE IDEAL SEQUENCE



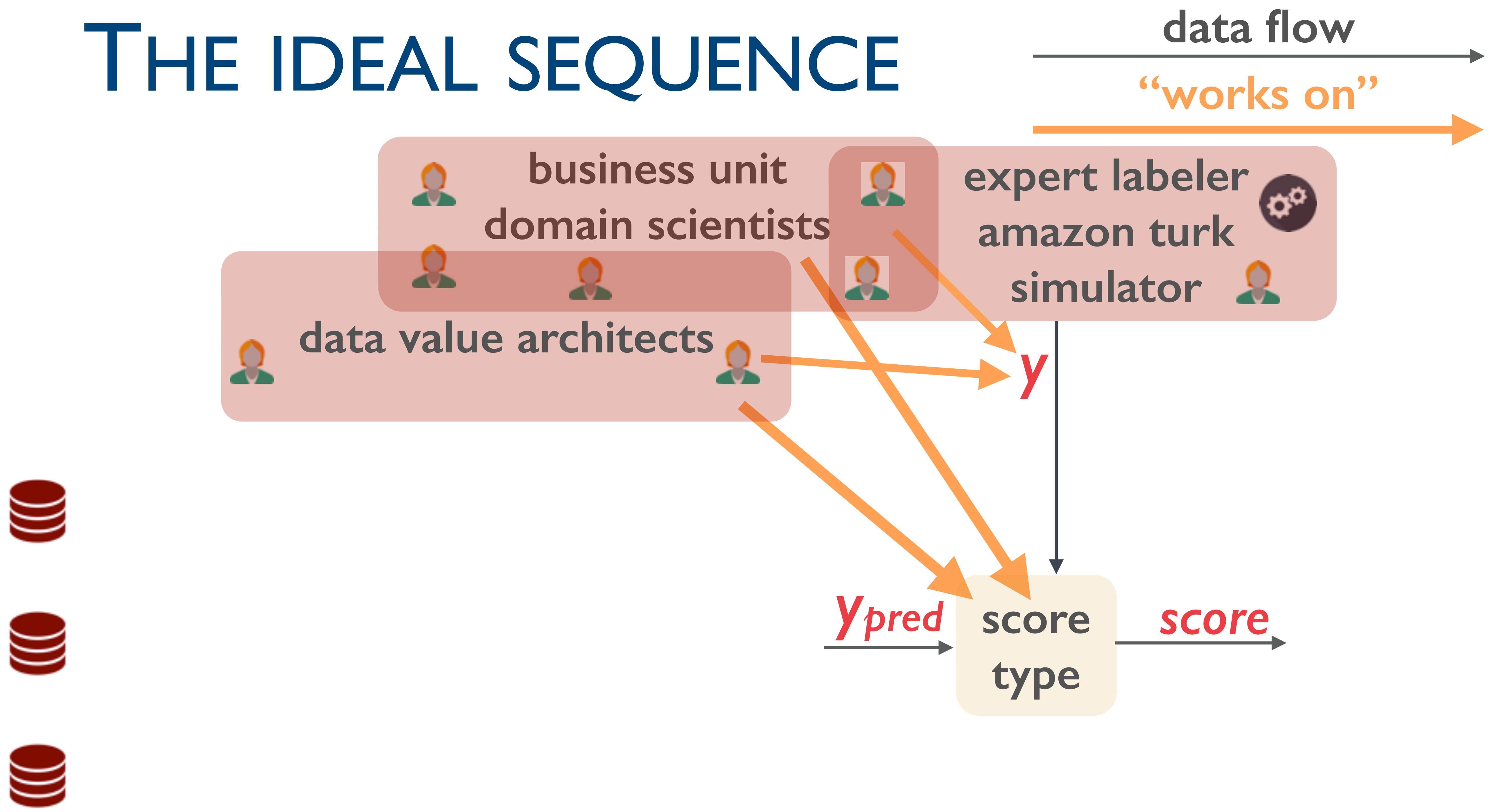
THE IDEAL SEQUENCE



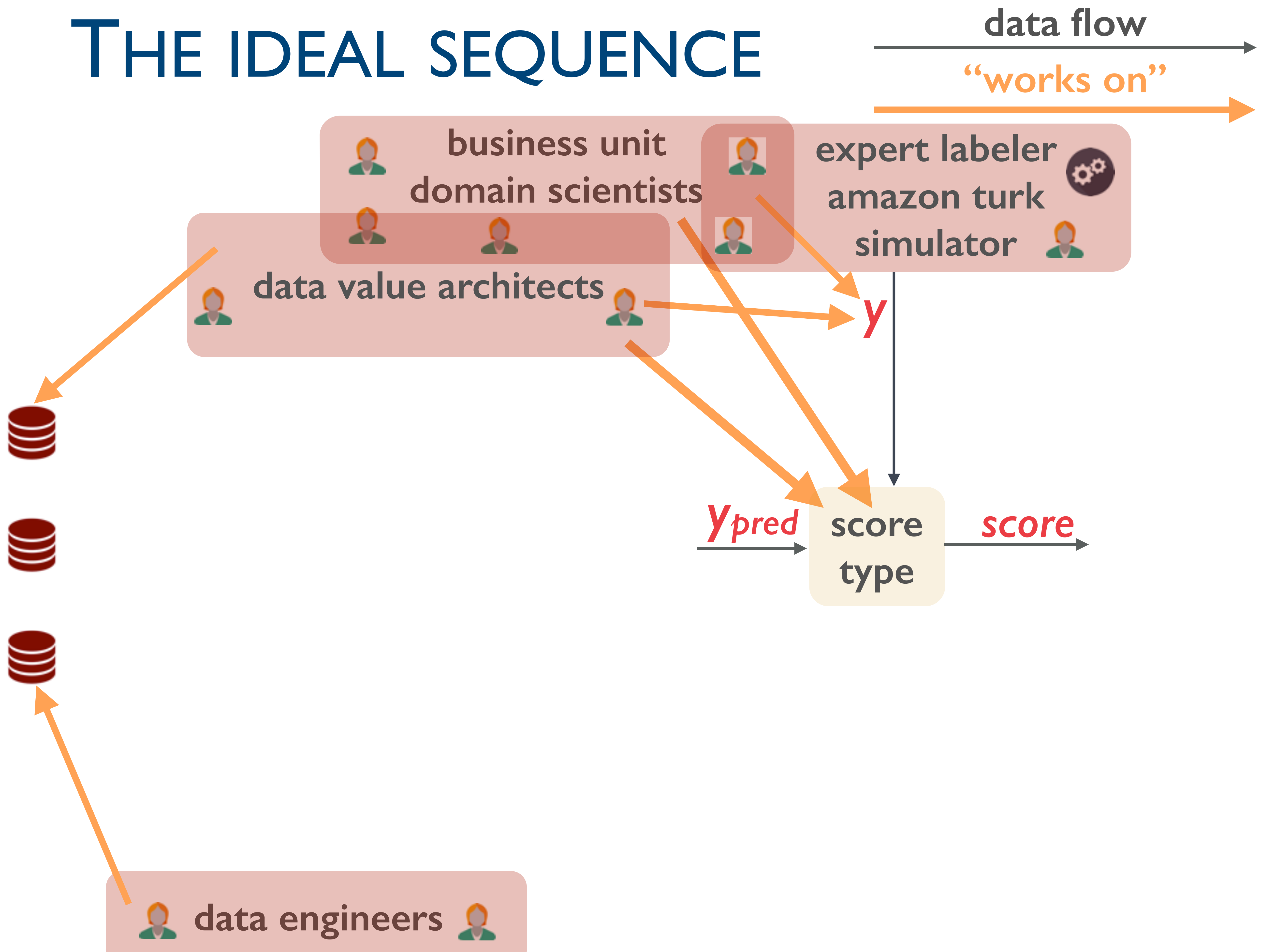
THE IDEAL SEQUENCE



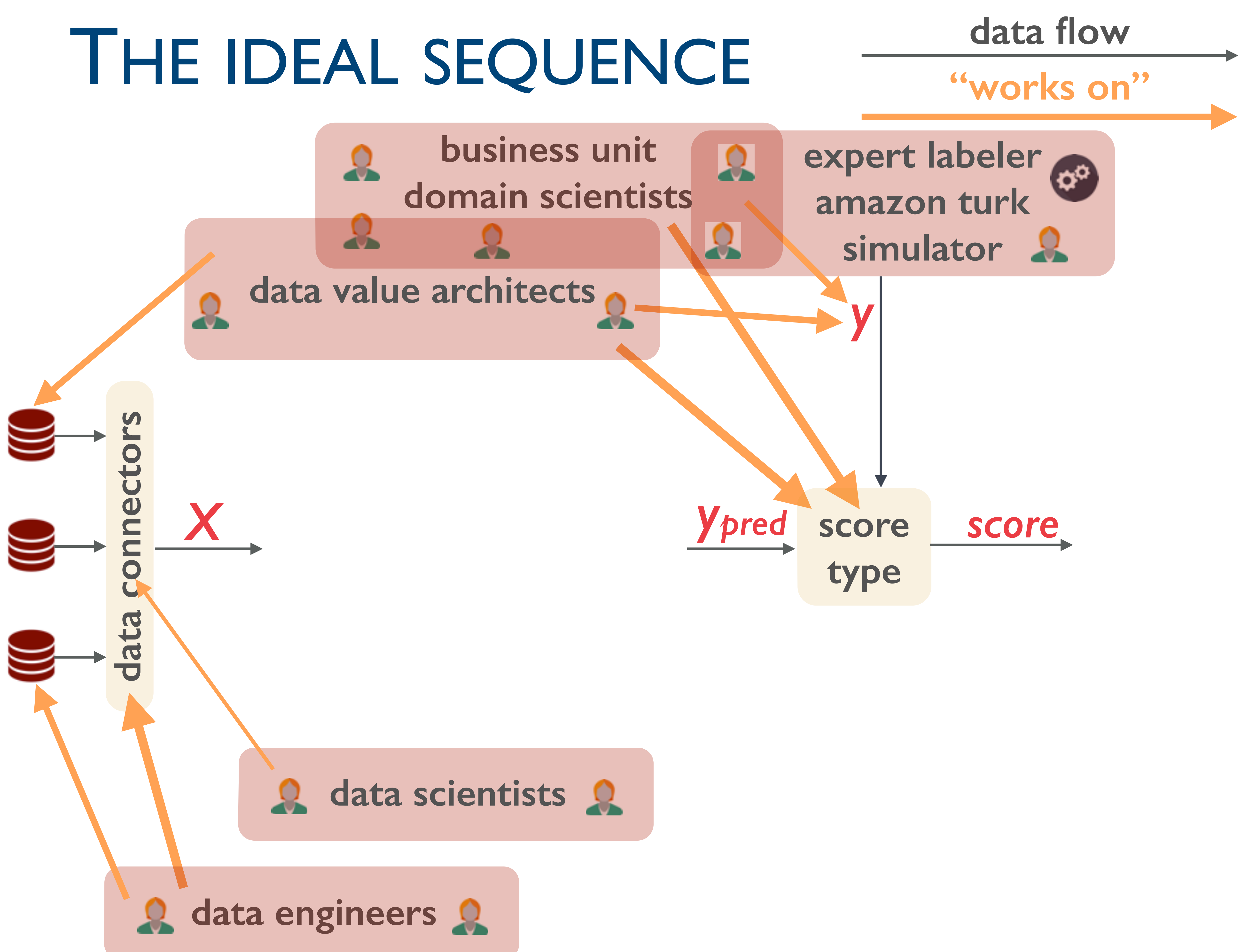
THE IDEAL SEQUENCE



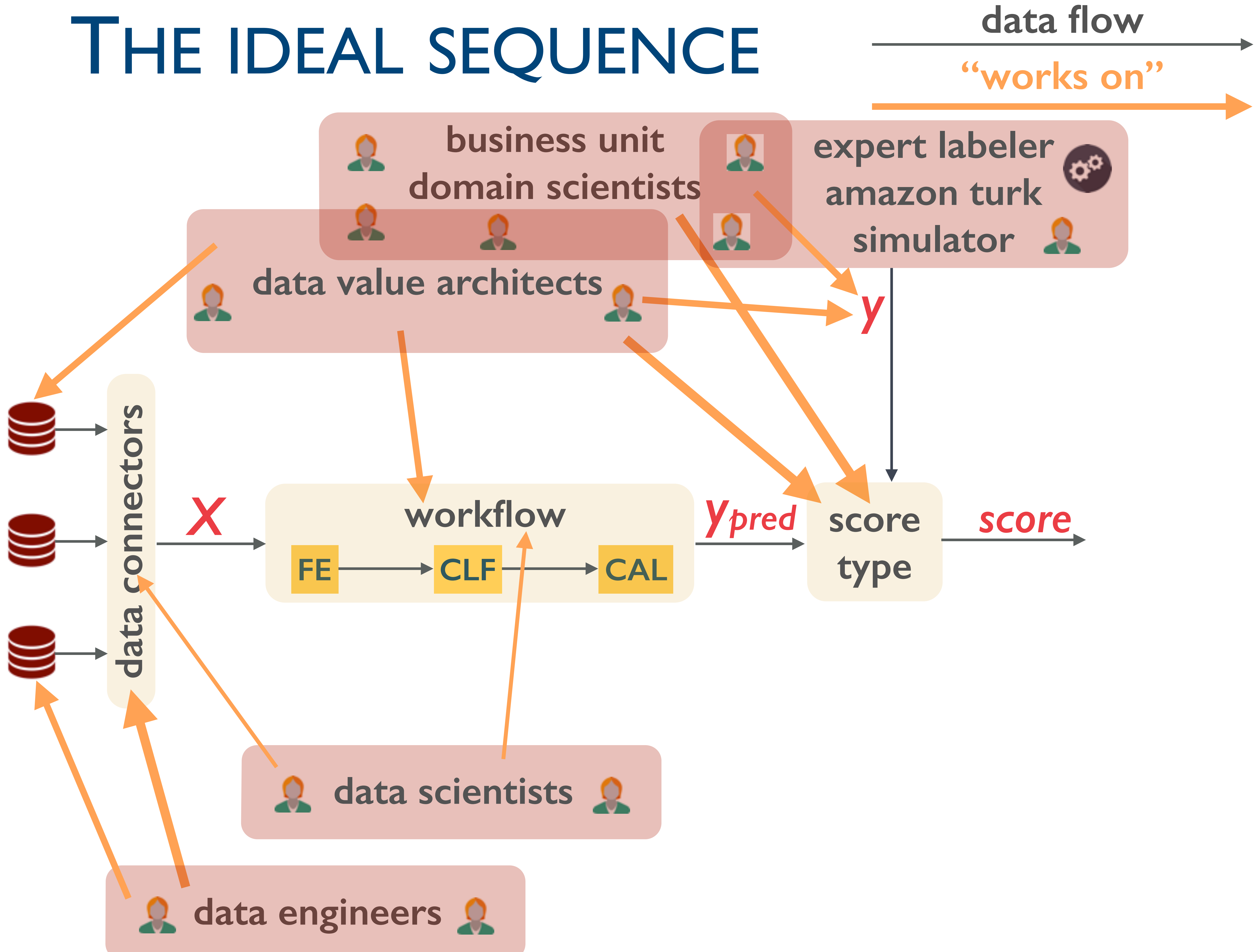
THE IDEAL SEQUENCE



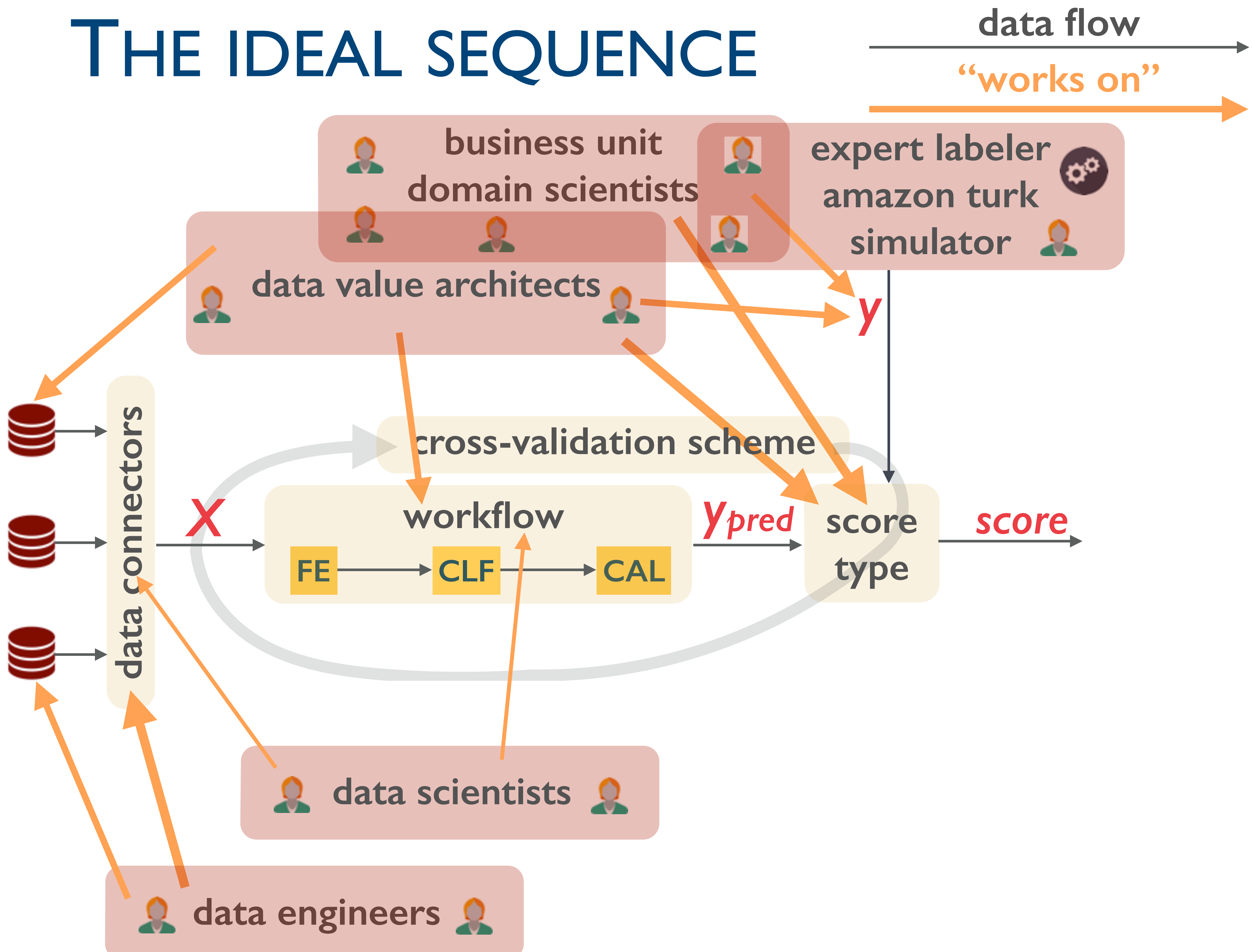
THE IDEAL SEQUENCE



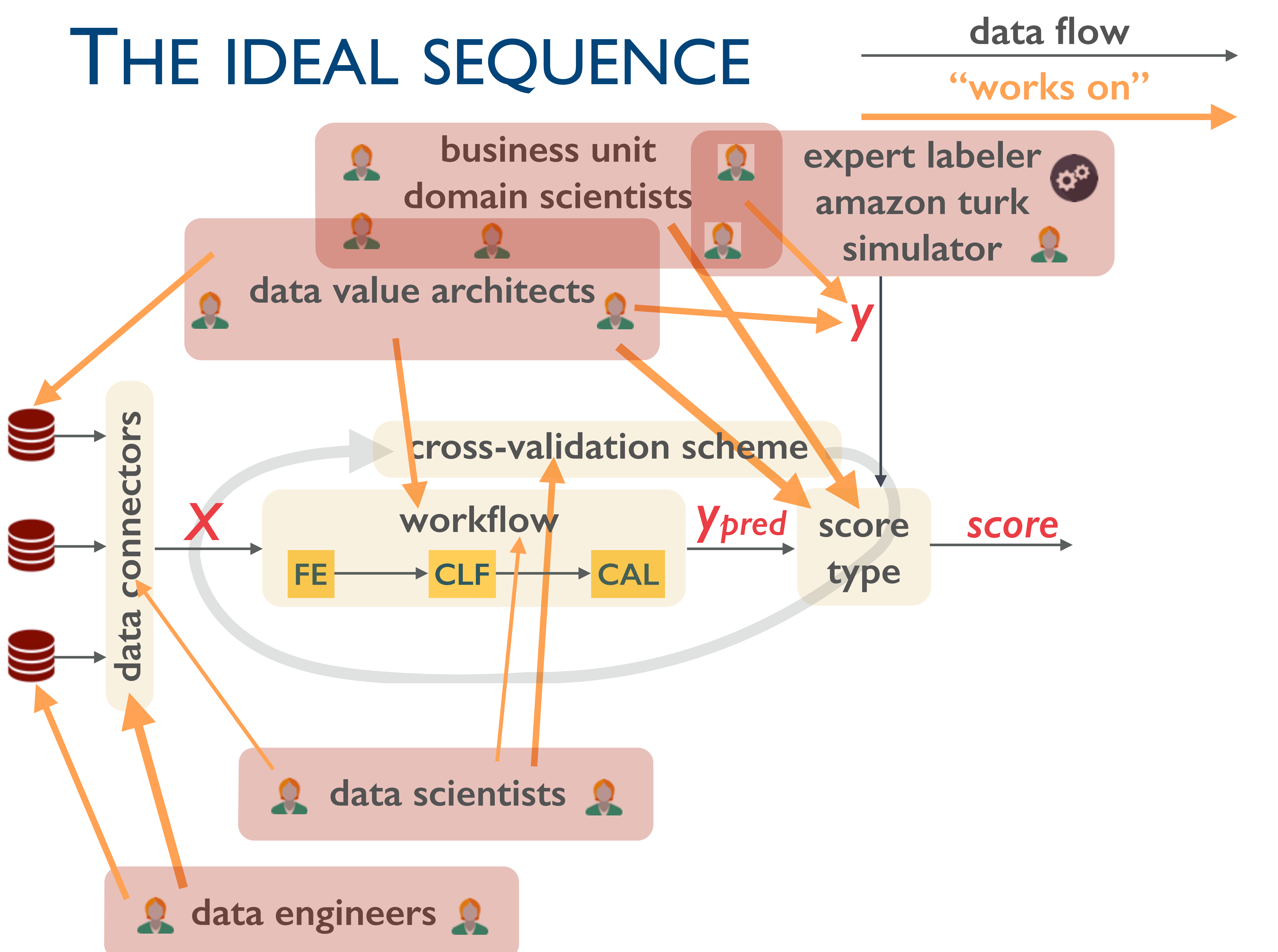
THE IDEAL SEQUENCE



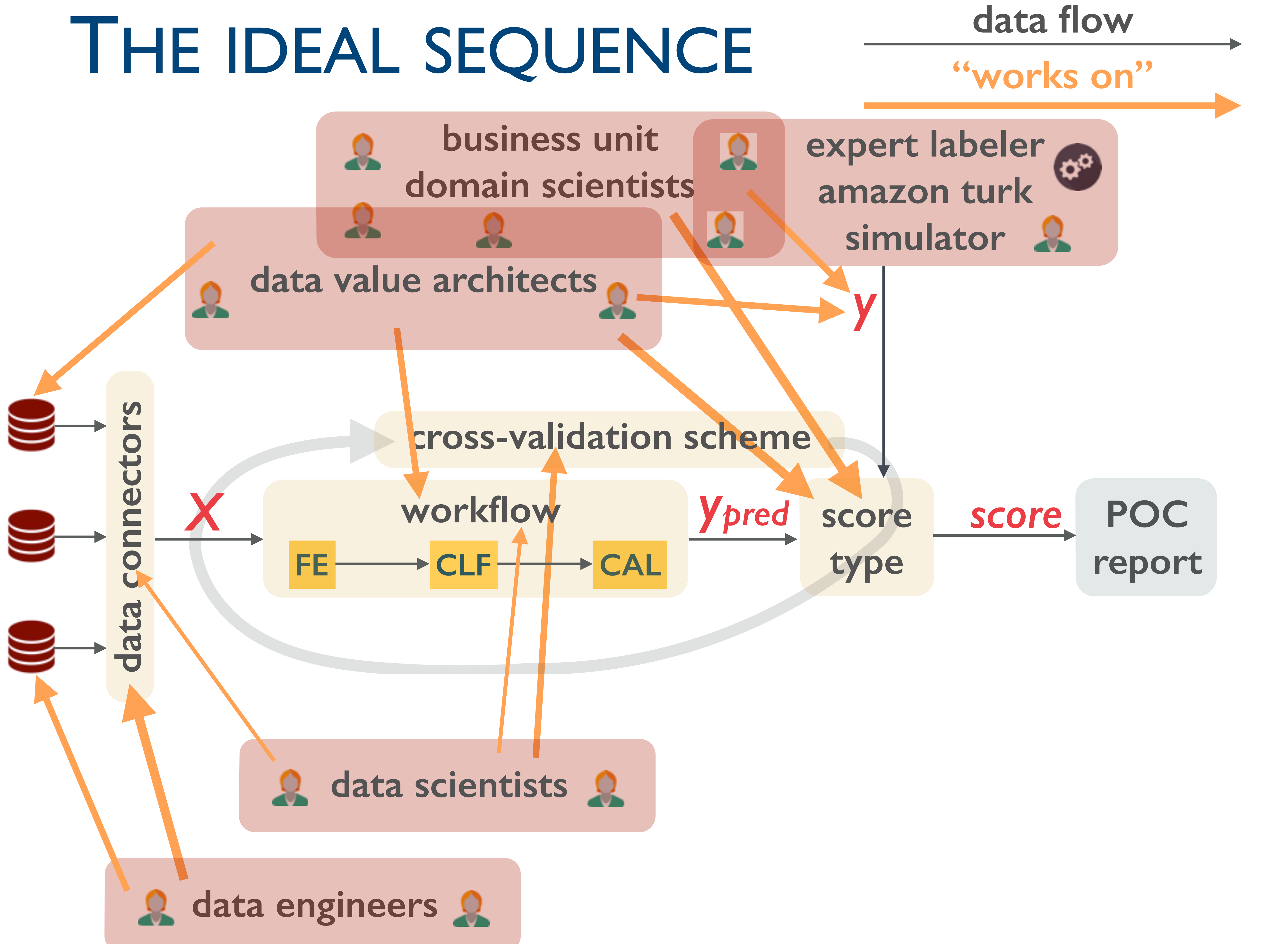
THE IDEAL SEQUENCE



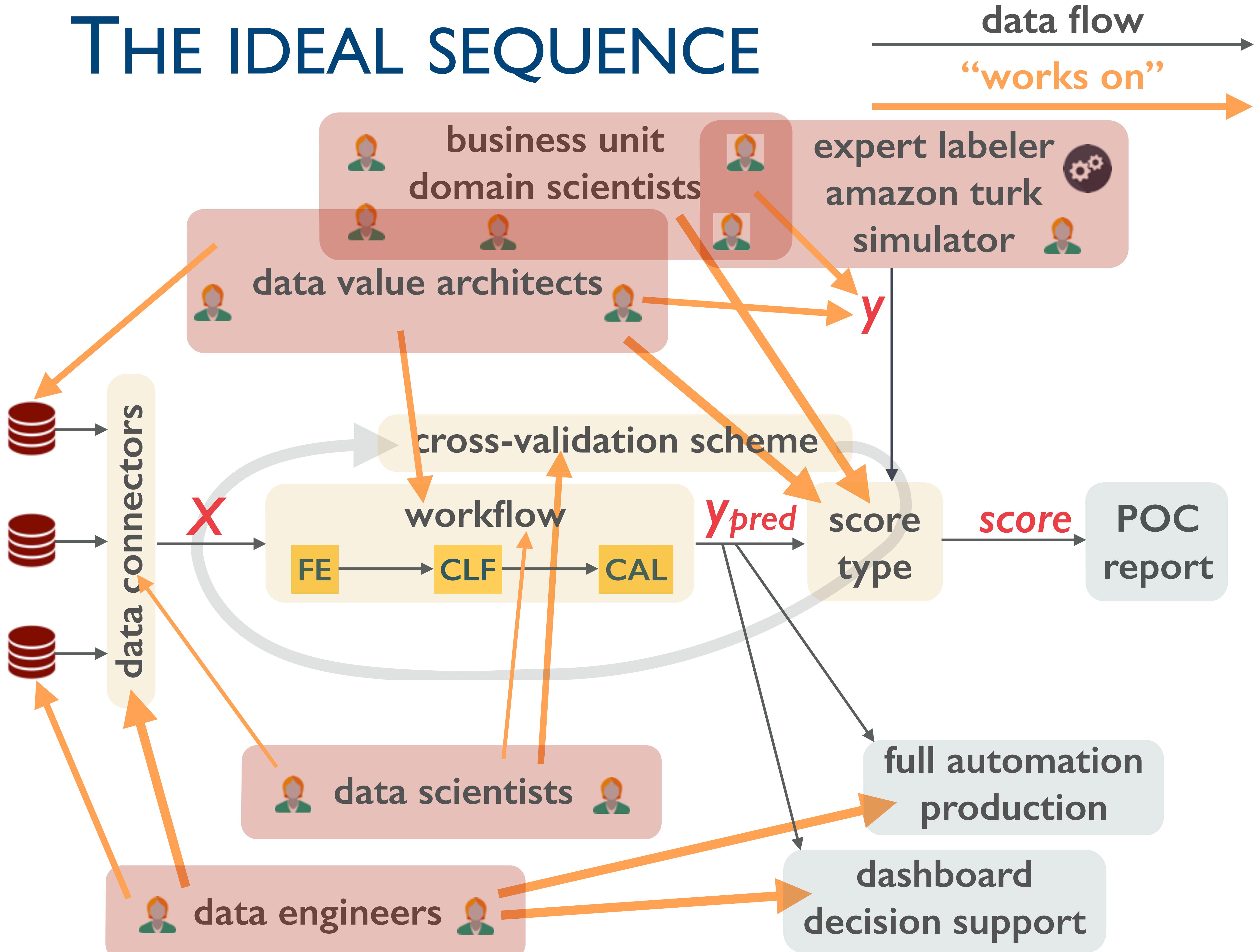
THE IDEAL SEQUENCE



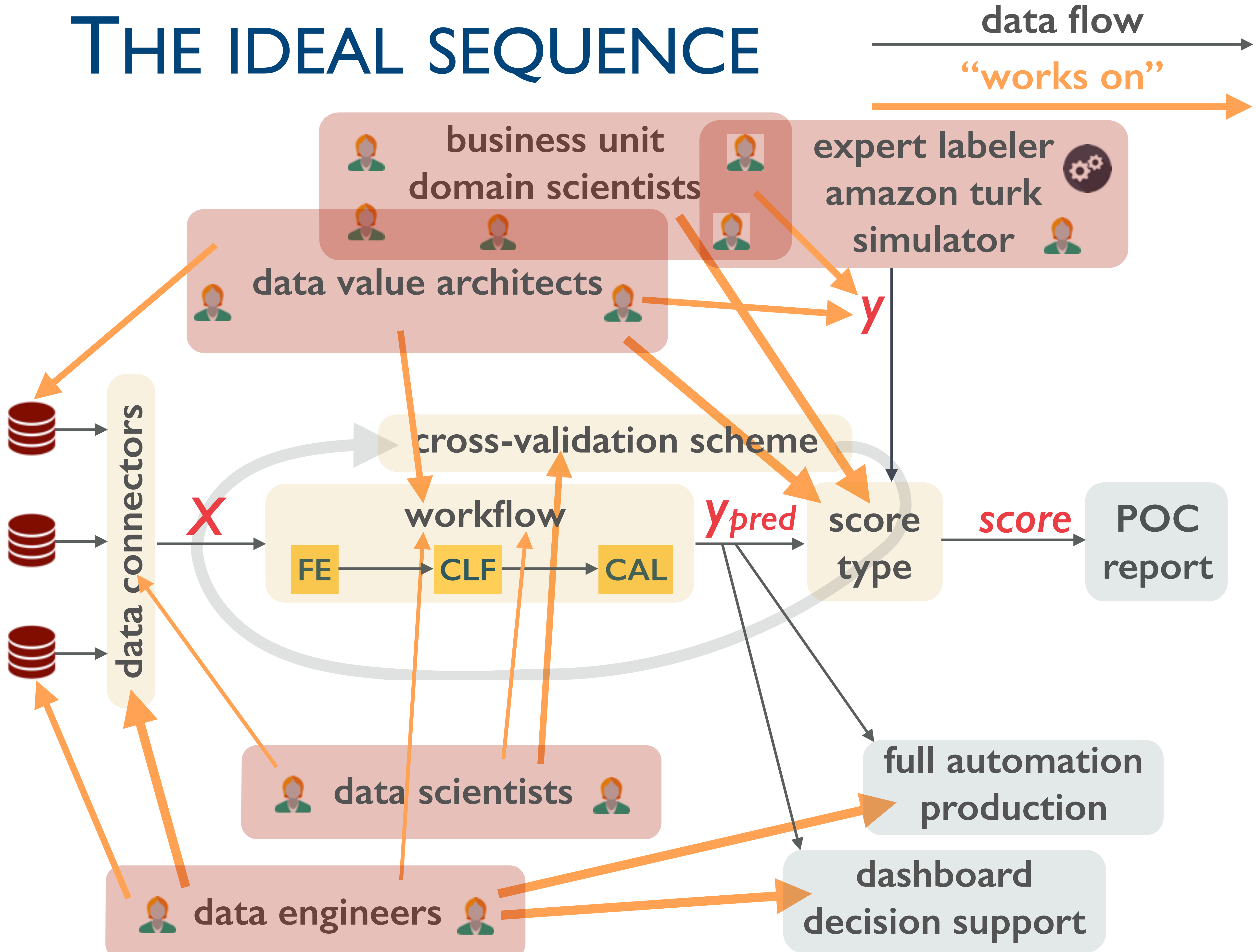
THE IDEAL SEQUENCE



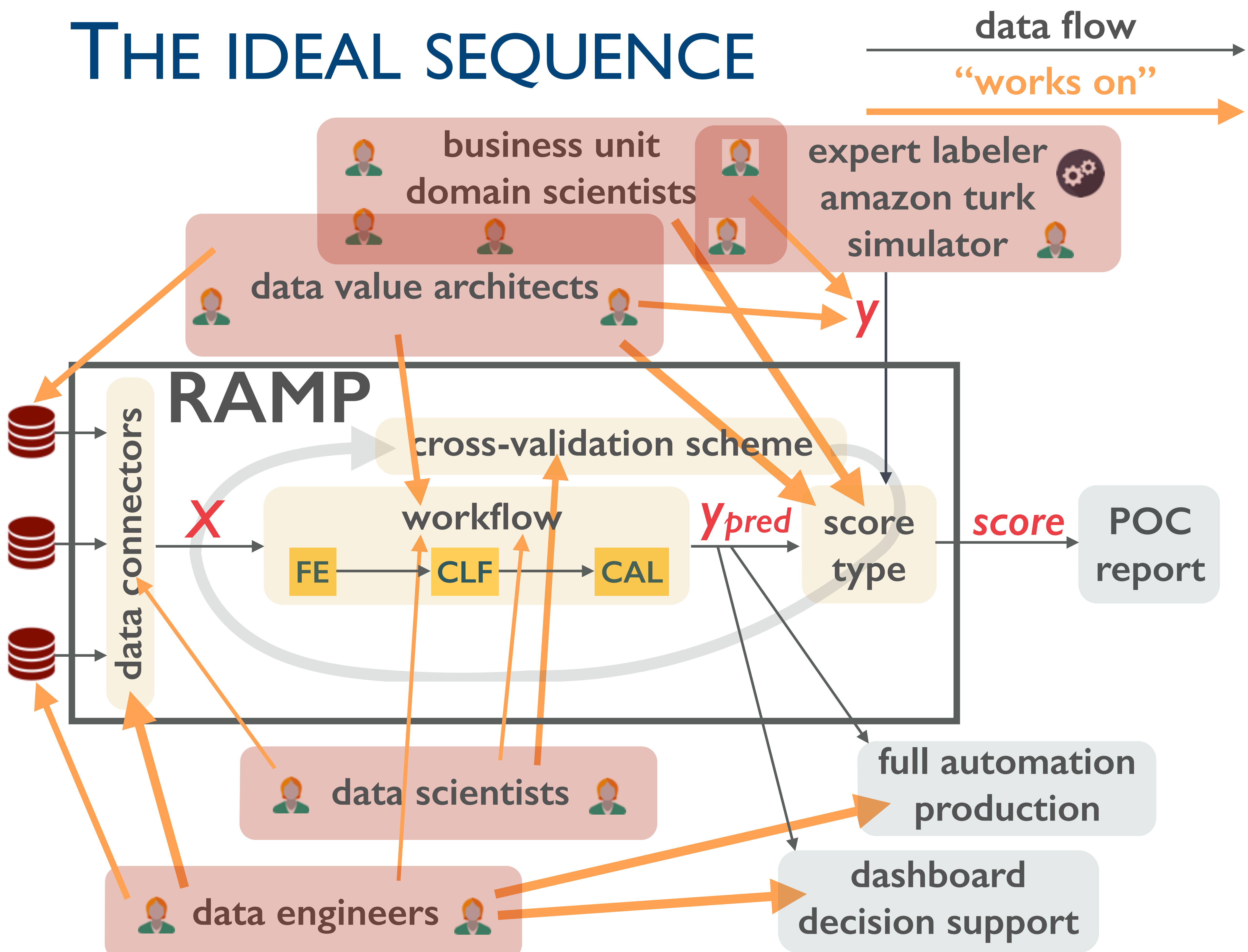
THE IDEAL SEQUENCE



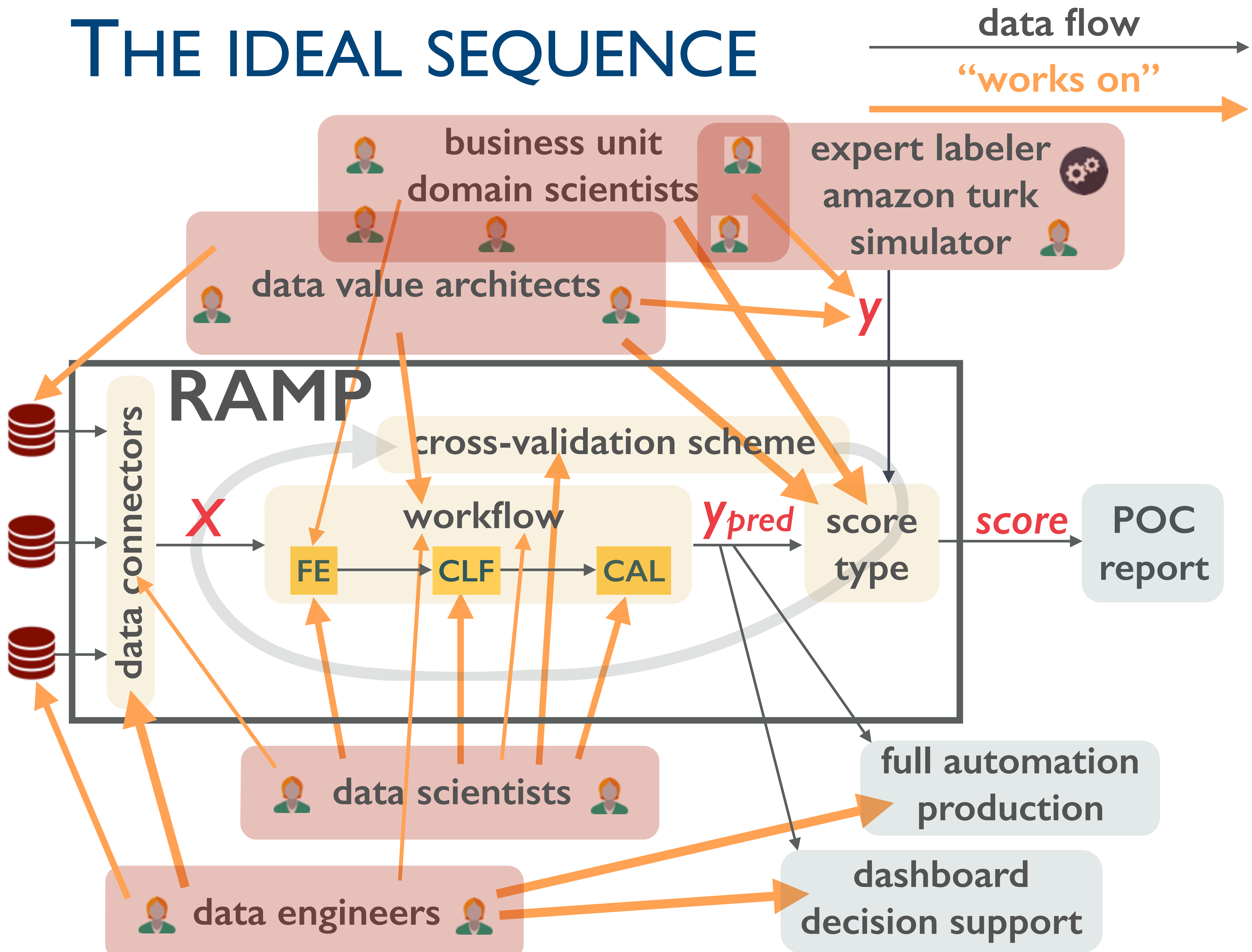
THE IDEAL SEQUENCE



THE IDEAL SEQUENCE

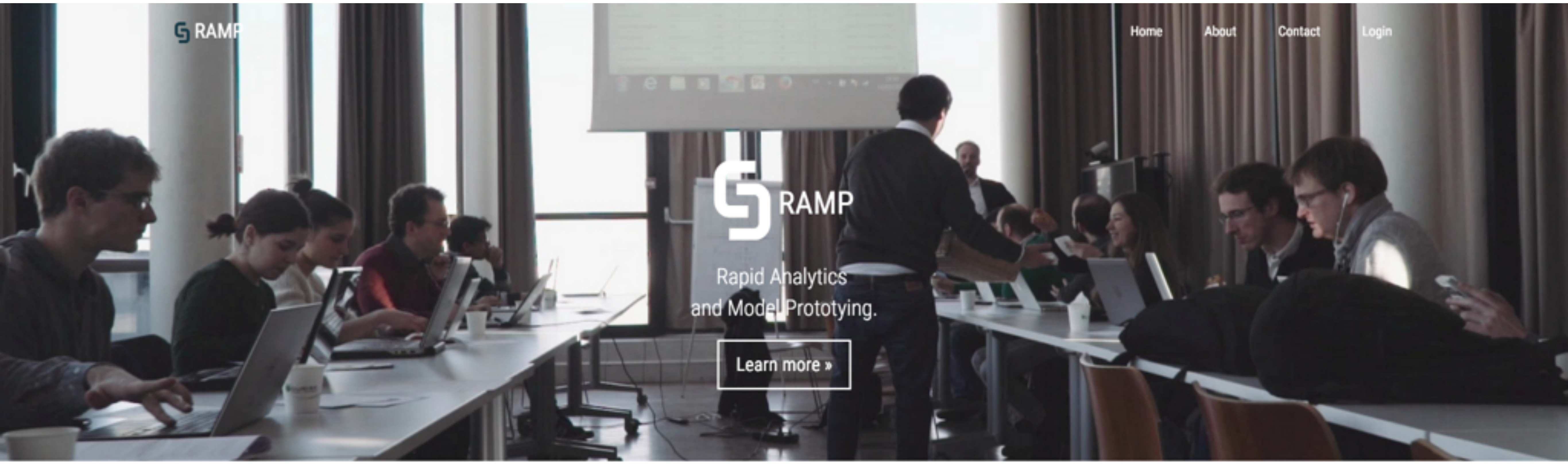


The ideal sequence



RAPID ANALYTICS AND MODEL PROTOTYPING (RAMP)

http://www.ramp.studio



Collaborative prototyping

During the RAMP, the participants submit predictive solutions (code). The models are trained on our back-end. The scores are displayed on a leaderboard. All participants have access to all code, and they are encouraged to look at and to reuse each other's solutions. This accelerates the development process since good ideas spread fast.

Training

A great tool to learn data science! RAMPs are used in the MS Big Data at Telecom ParisTech, in three UPSaclay M2 programs (Data Science, AIC, Data and Knowledge), in a course on Machine Learning for Finance and Economics at Université Panthéon-Assas, in a graduate course in the Data analysis and decision program at Ecole Centrale de Lille.

Networking

Each RAMP attracts about 30-50 participants, coming from different backgrounds and carrier stages, who usually meet for the first time. They develop a working relationship in a relaxed environment, and sometimes keep working together after the event.

RAPID ANALYTICS AND MODEL PROTOTYPING (RAMP)

http://www.ramp.studio

Team


Balázs Kégl


Alex Gramfort


Akin Kazakçi


Mehdi Cherti


Yohann Sitruk

Alumni


Djalel Benbouzid


Camille Marini

RAMP

RAPID ANALYTICS AND MODEL PROTOTYPING

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

Current RAMPs

www.ramp.studio/problems

RAMP

Hi Balazs!

- **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

www.ramp.studio/data_science_themes

RAMP

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

www.ramp.studio/data_science_themes

RAMP

- **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

www.ramp.studio/data_domains

RAMP

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

DATA CHALLENGE WITH CODE SUBMISSION

frontend

backend

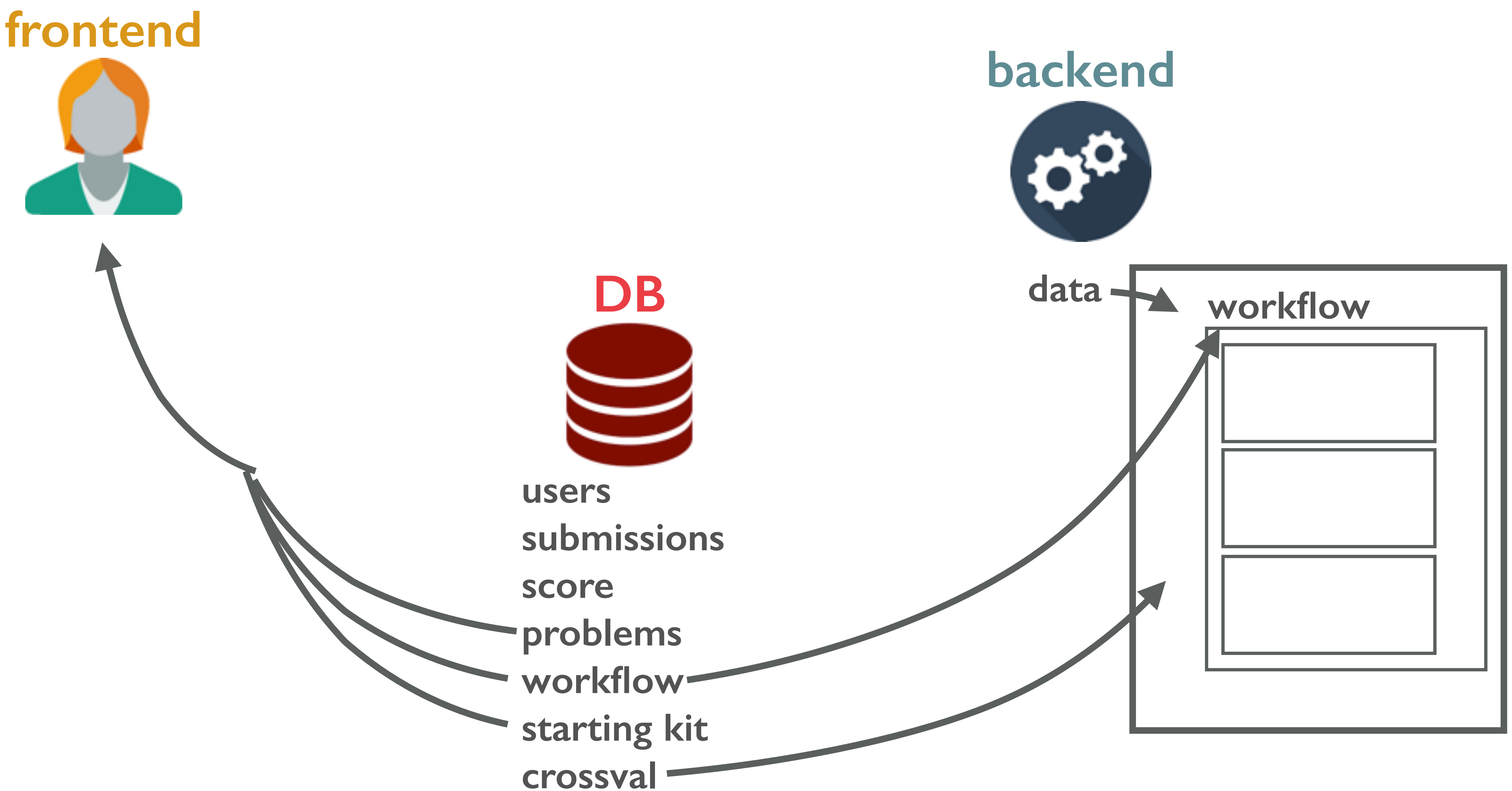
data

DB

users
submissions
score
problems
workflow
starting kit
crossval

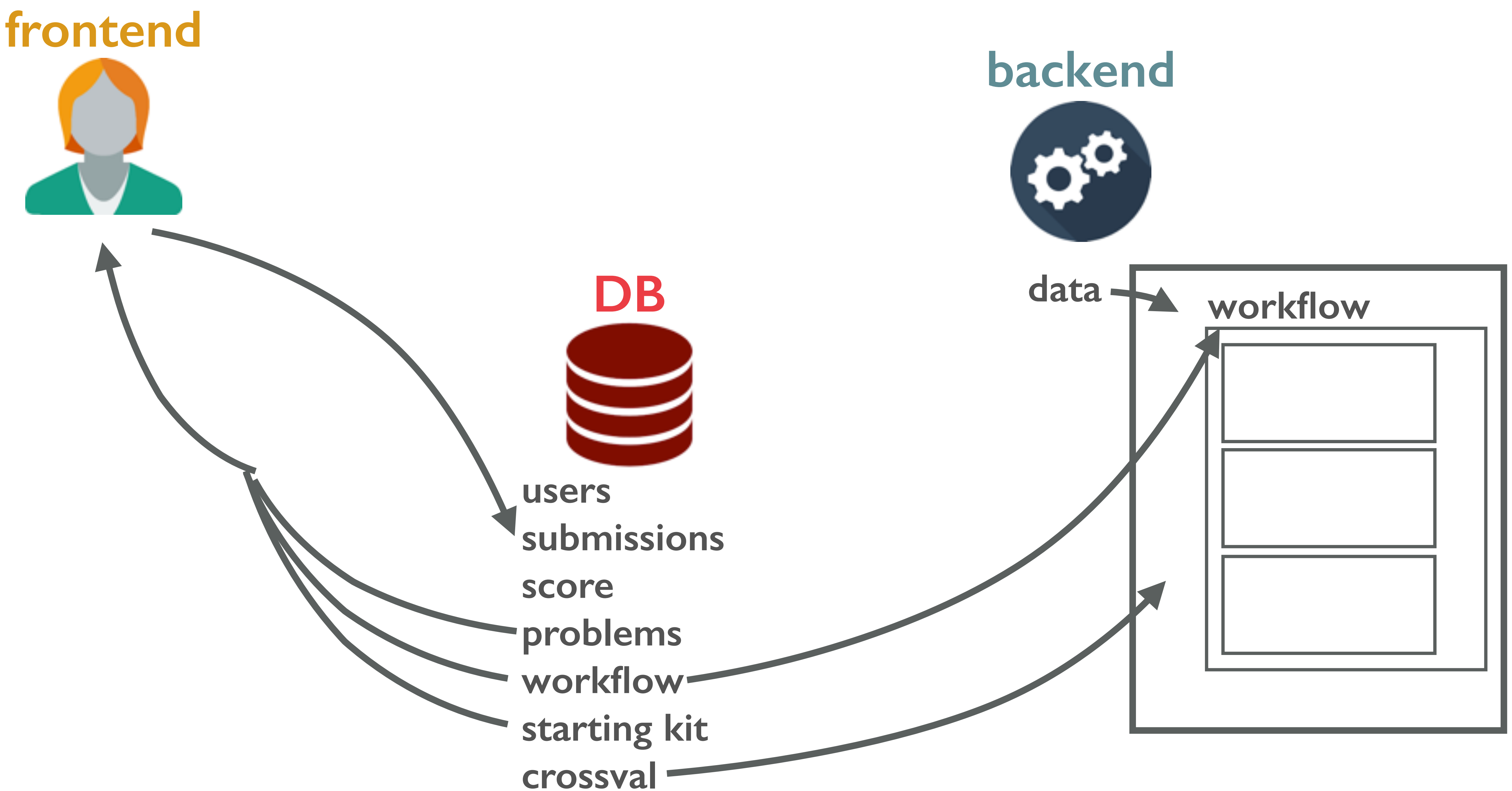
RAMP

DATA CHALLENGE WITH CODE SUBMISSION



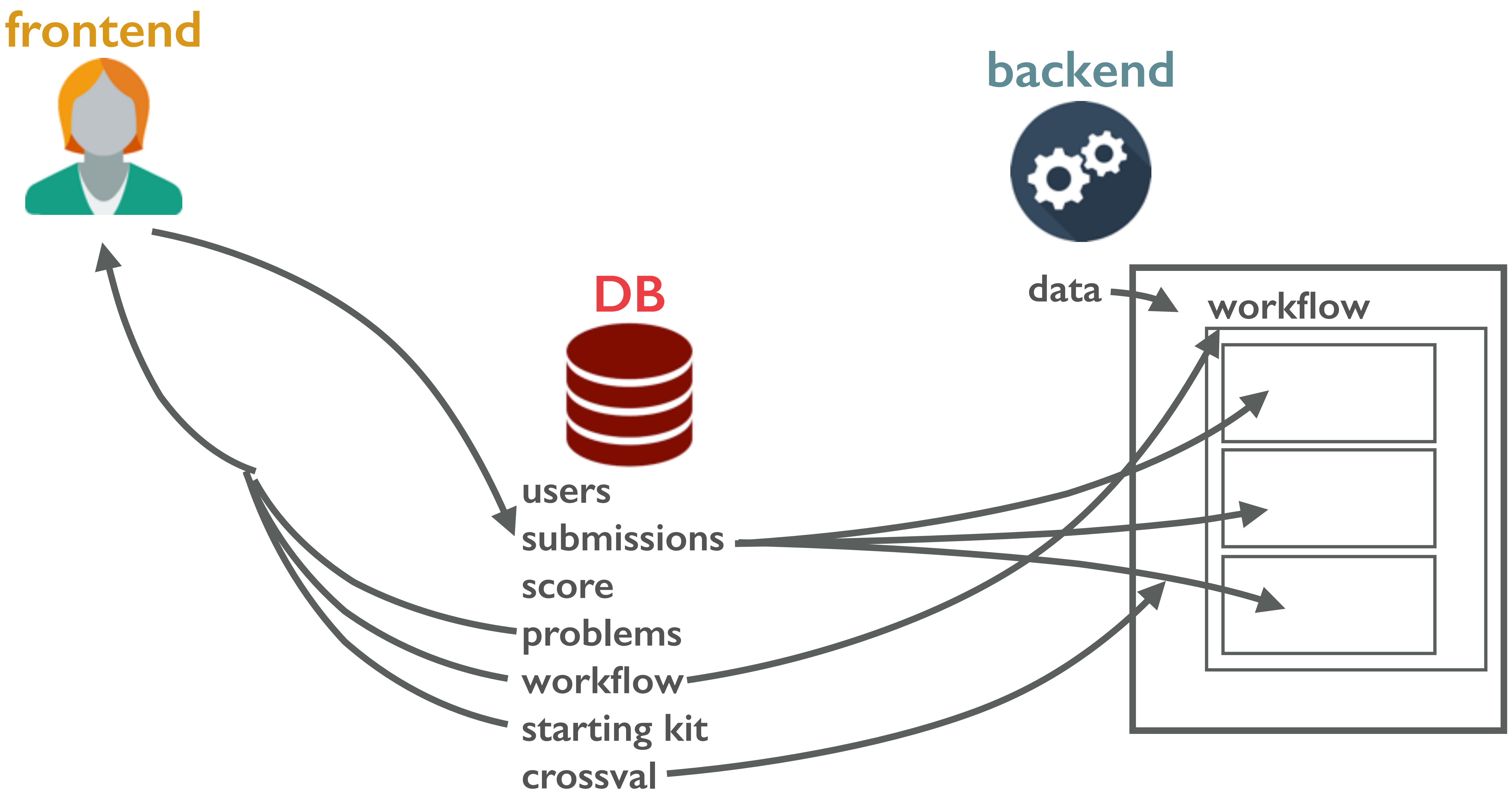
RAMP

DATA CHALLENGE WITH CODE SUBMISSION



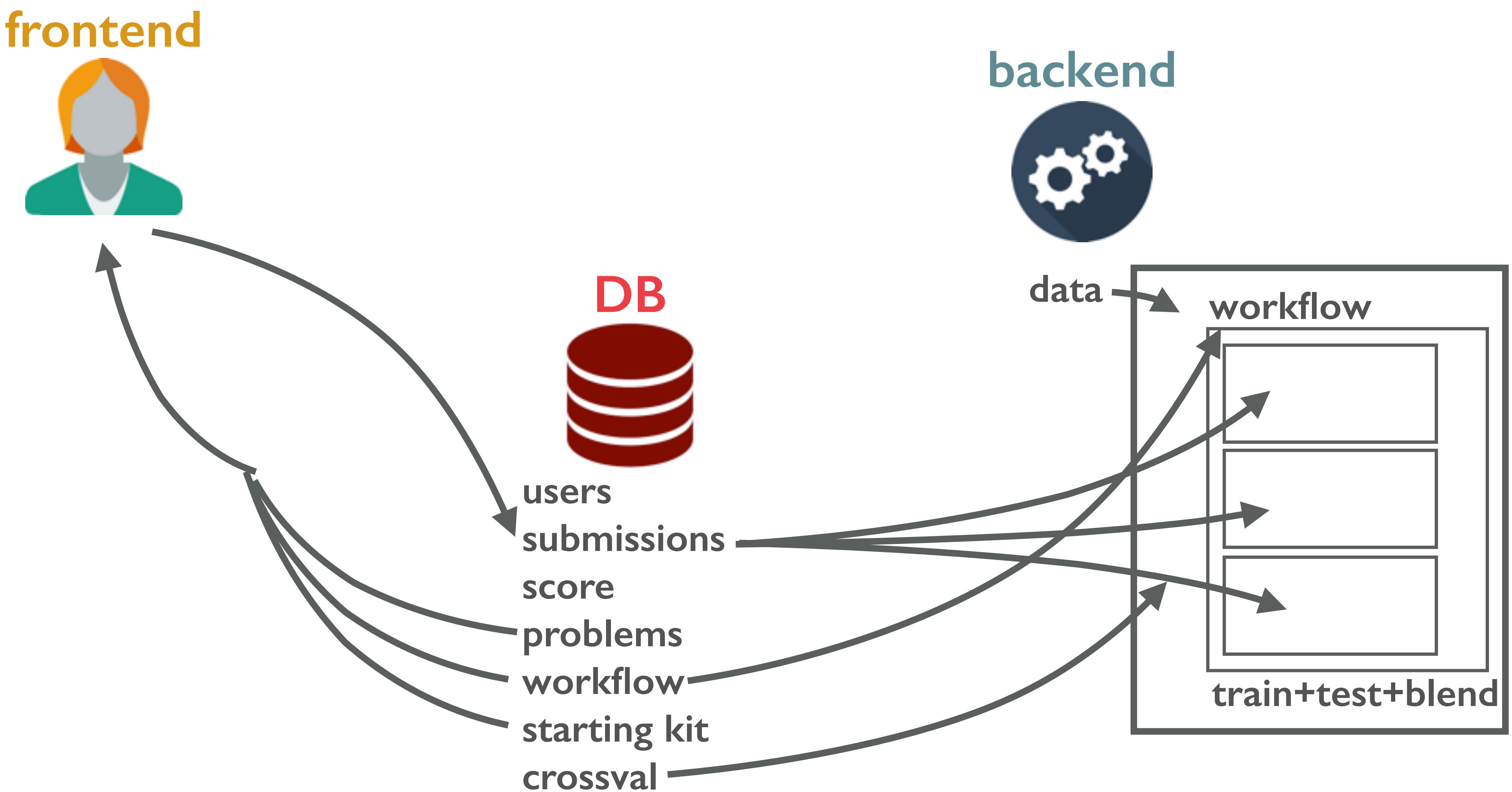
RAMP

DATA CHALLENGE WITH CODE SUBMISSION



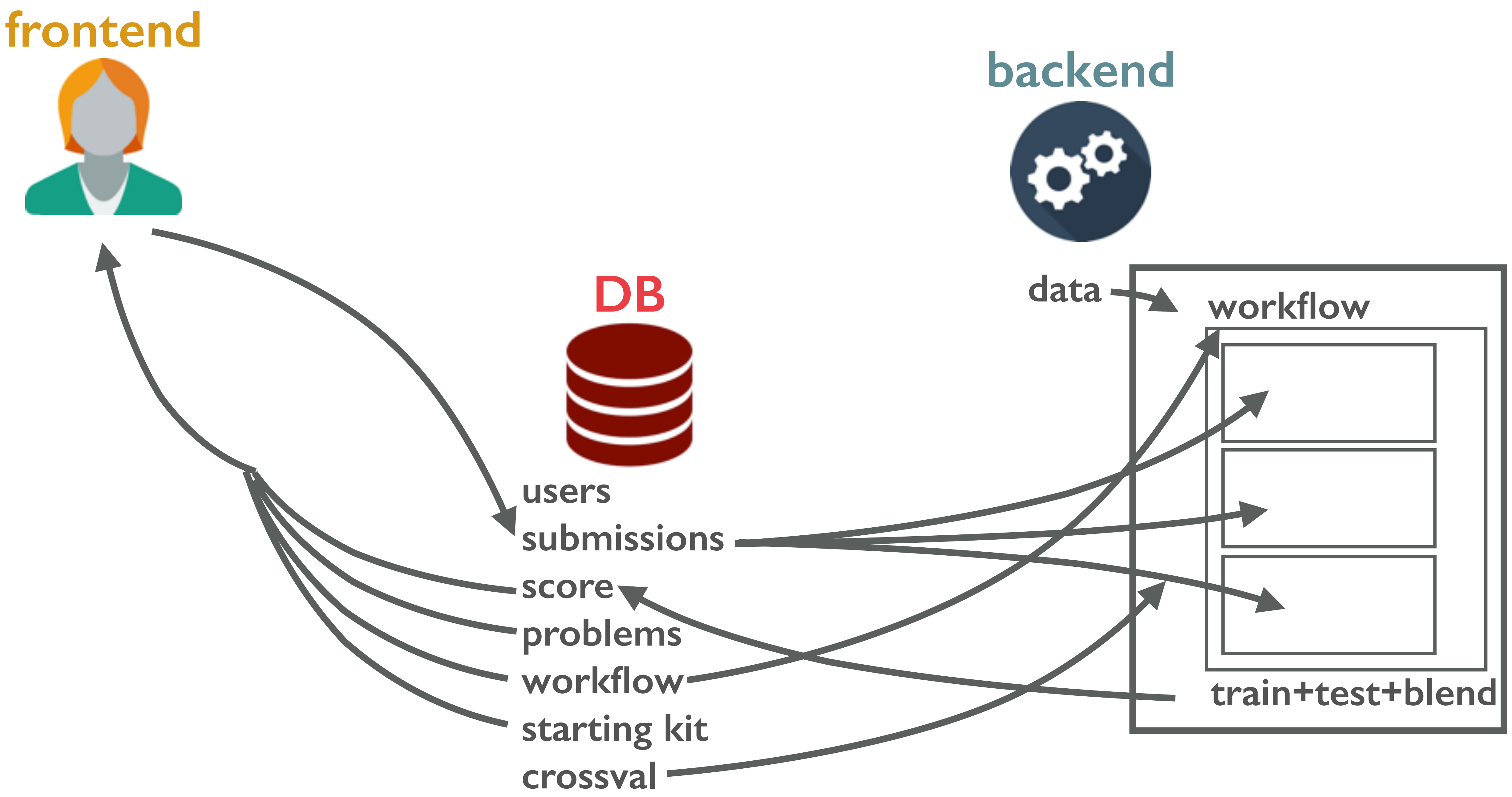
RAMP

DATA CHALLENGE WITH CODE SUBMISSION



RAMP

DATA CHALLENGE WITH CODE SUBMISSION



RAMP

www.ramp.studio/events/sea_ice_M1XMAP583_201617

RAMP

Hi Balazs!

sea_ice_M1XMAP583_201617

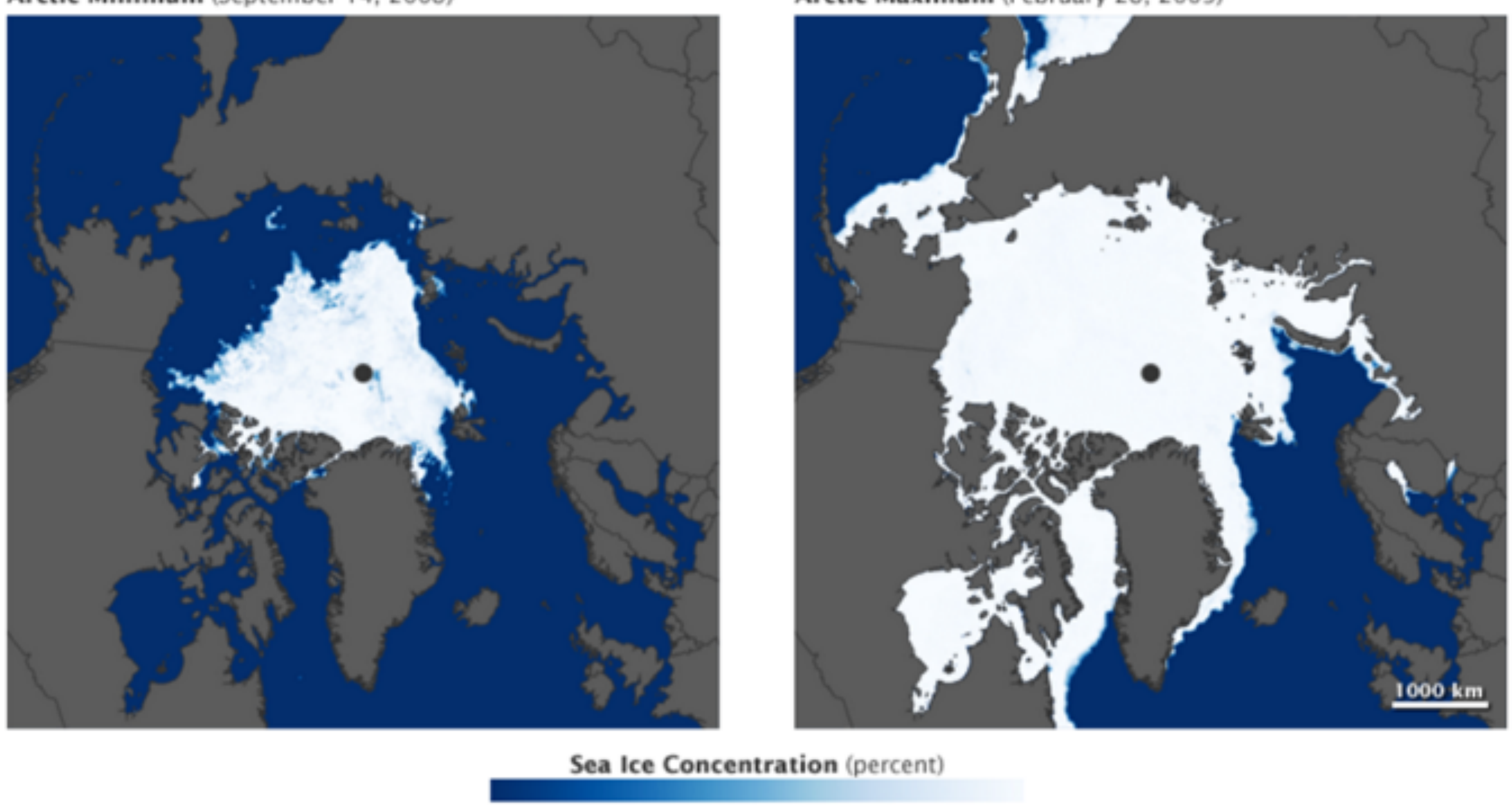
Description

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

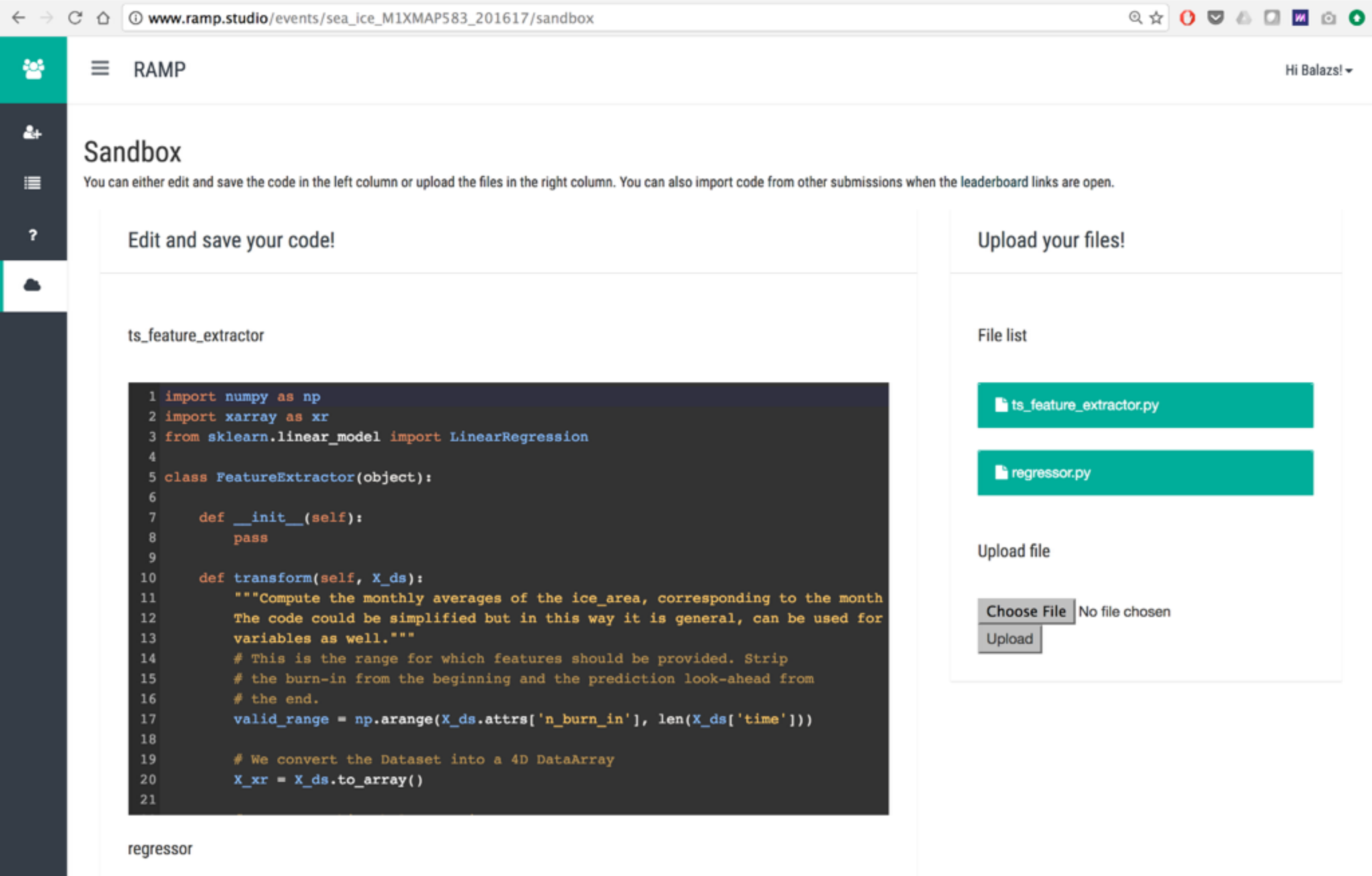
Introduction

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.

Arctic Minimum (September 14, 2008) **Arctic Maximum** (February 28, 2009)



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
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

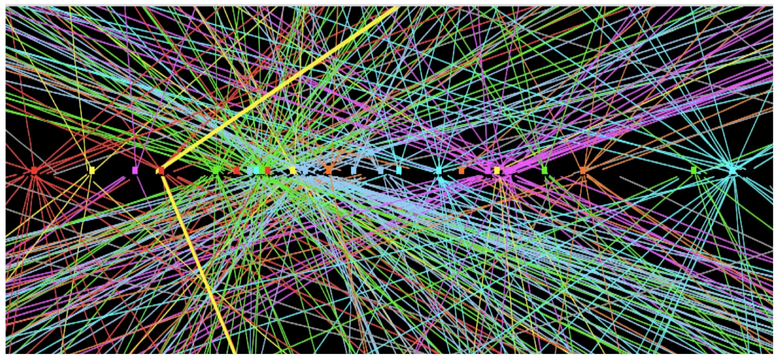
Search:

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

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)

CLASSIFYING AND REGRESSING ON MOLECULAR SPECTRA

chemotherapy drug in elastic pocket

CLASSIFYING AND REGRESSING ON MOLECULAR SPECTRA

chemotherapy drug in elastic pocket

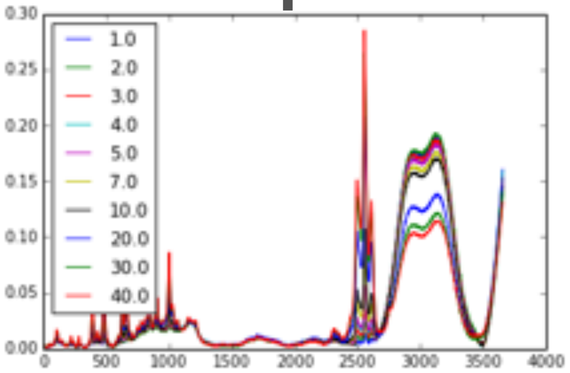
laser spectrometer

CLASSIFYING AND REGRESSING ON MOLECULAR SPECTRA

chemotherapy drug in elastic pocket

laser spectrometer

molecular spectra

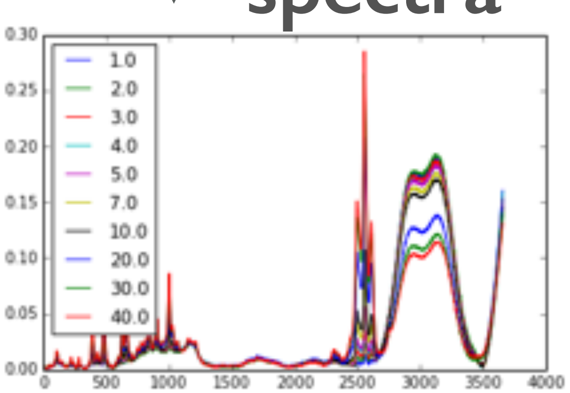


CLASSIFYING AND REGRESSING ON MOLECULAR SPECTRA

chemotherapy drug in elastic pocket

laser spectrometer

molecular spectra



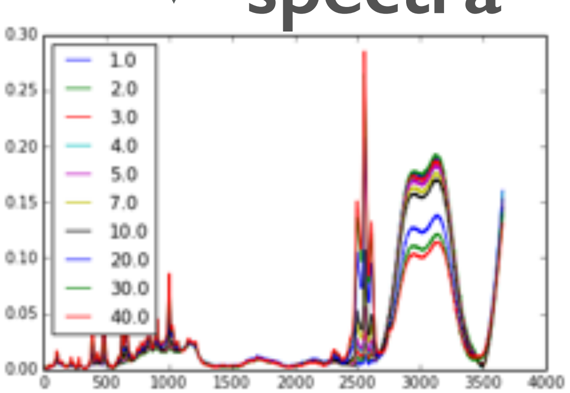
feature extractor 1

CLASSIFYING AND REGRESSING ON MOLECULAR SPECTRA

chemotherapy drug in elastic pocket

laser spectrometer

molecular spectra



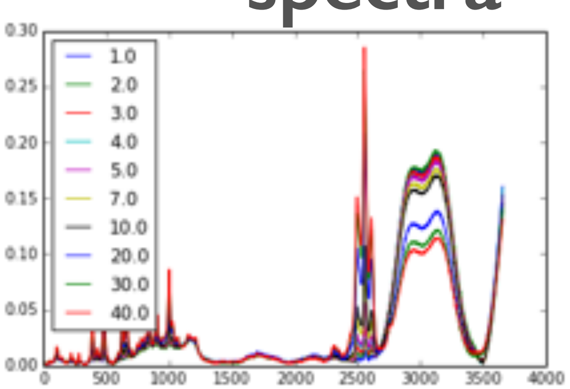
feature extractor 1 → **classifier** → drug type

CLASSIFYING AND REGRESSING ON MOLECULAR SPECTRA

chemotherapy drug in elastic pocket

laser spectrometer

molecular spectra



feature extractor 1

classifier

drug type

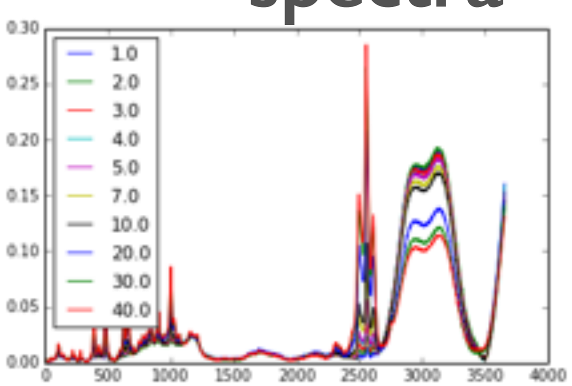
feature extractor 2

CLASSIFYING AND REGRESSING ON MOLECULAR SPECTRA

chemotherapy drug in elastic pocket

laser spectrometer

molecular spectra

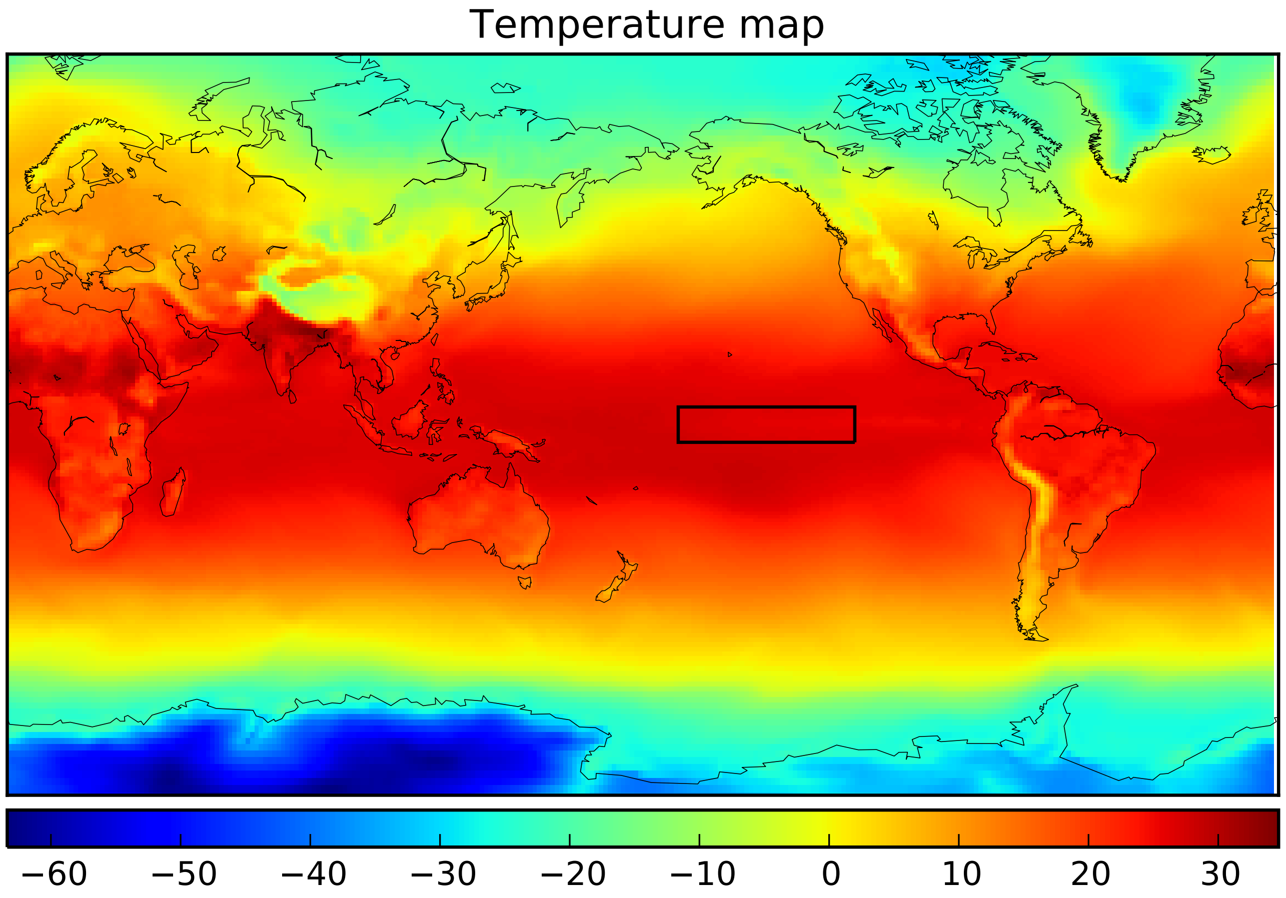


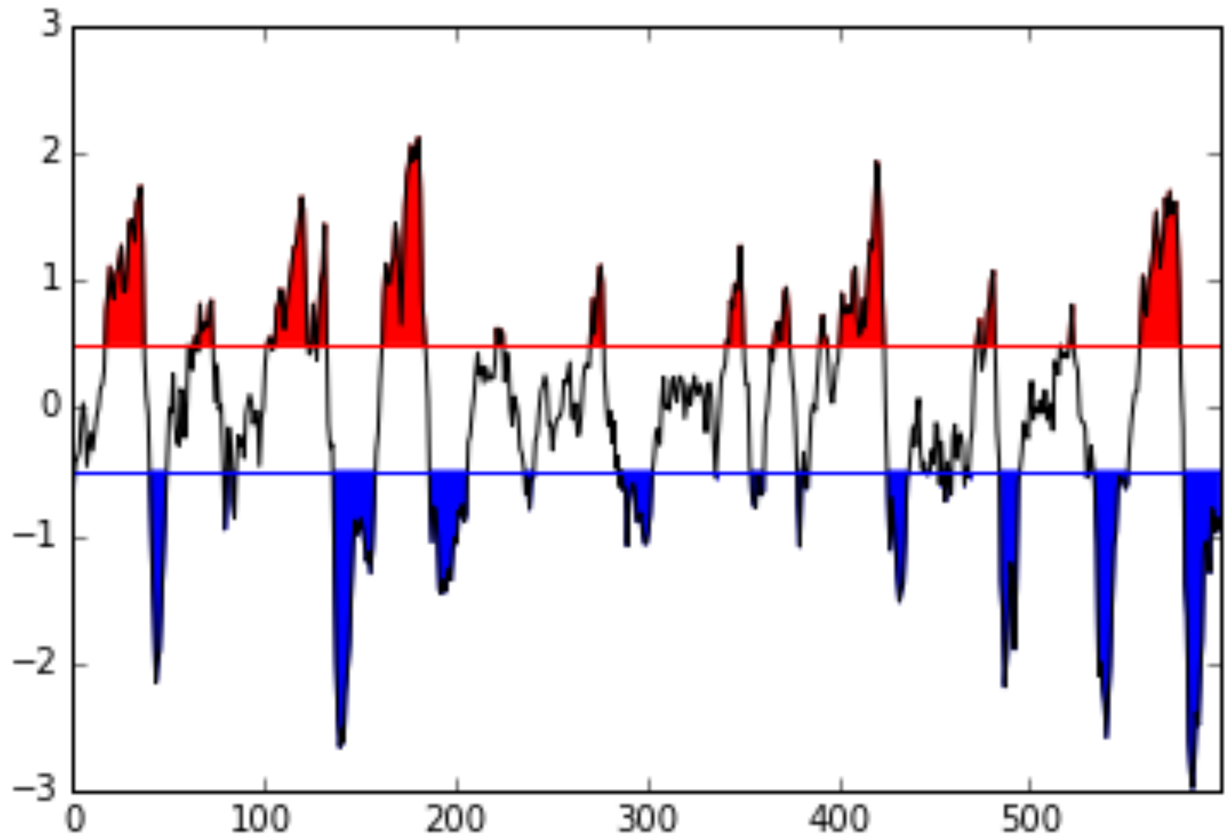
feature extractor 1 → classifier → drug type

classifier → feature extractor 2

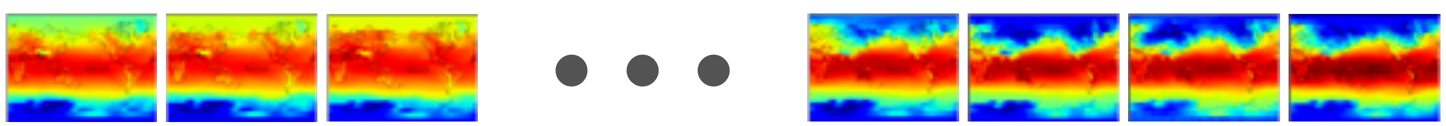
feature extractor 2 → regressor → concentration

FORECASTING EL NINO SIX MONTHS AHEAD

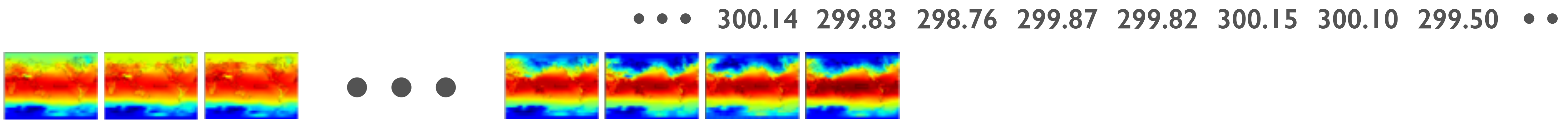




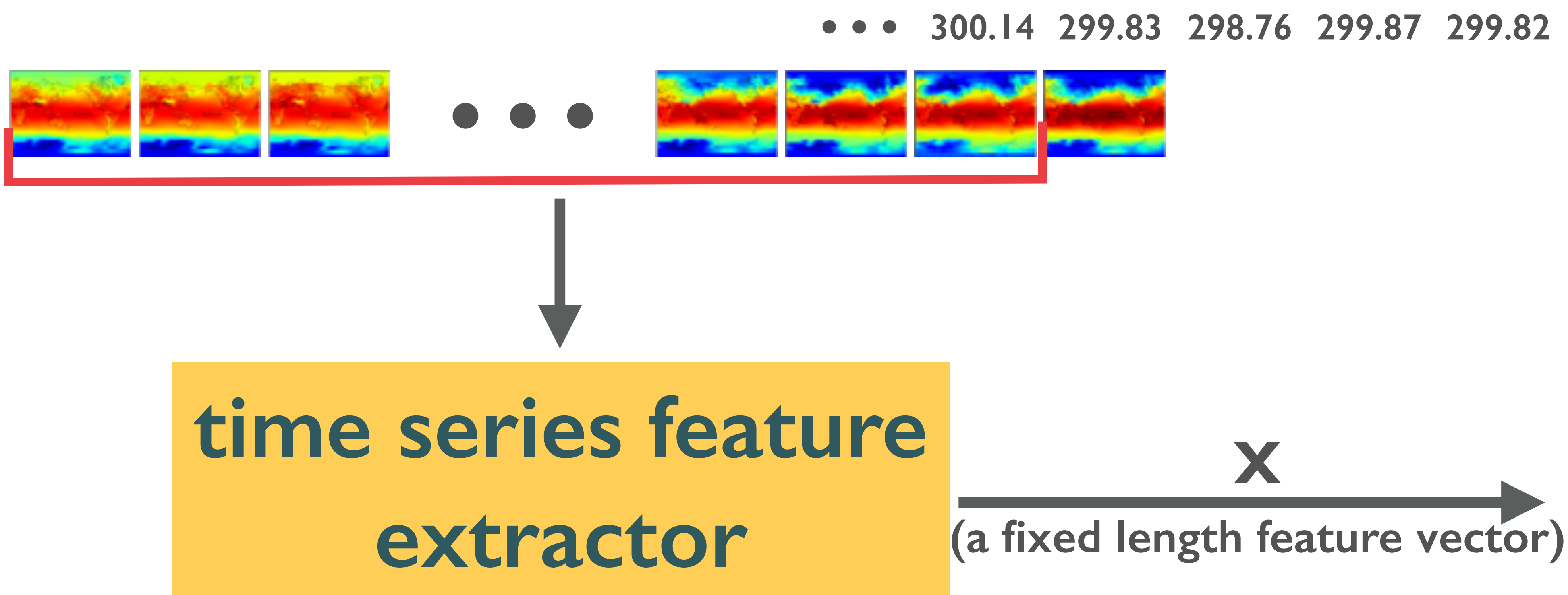
Forecasting El Nino six months ahead



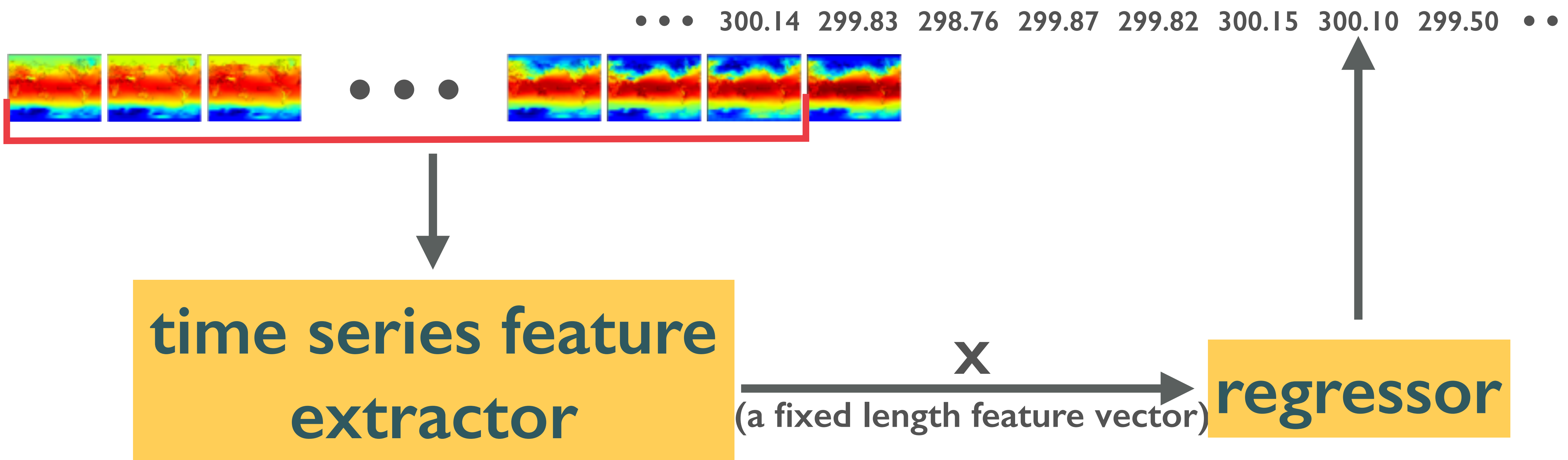
Forecasting El Nino six months ahead



FORECASTING EL NINO SIX MONTHS AHEAD

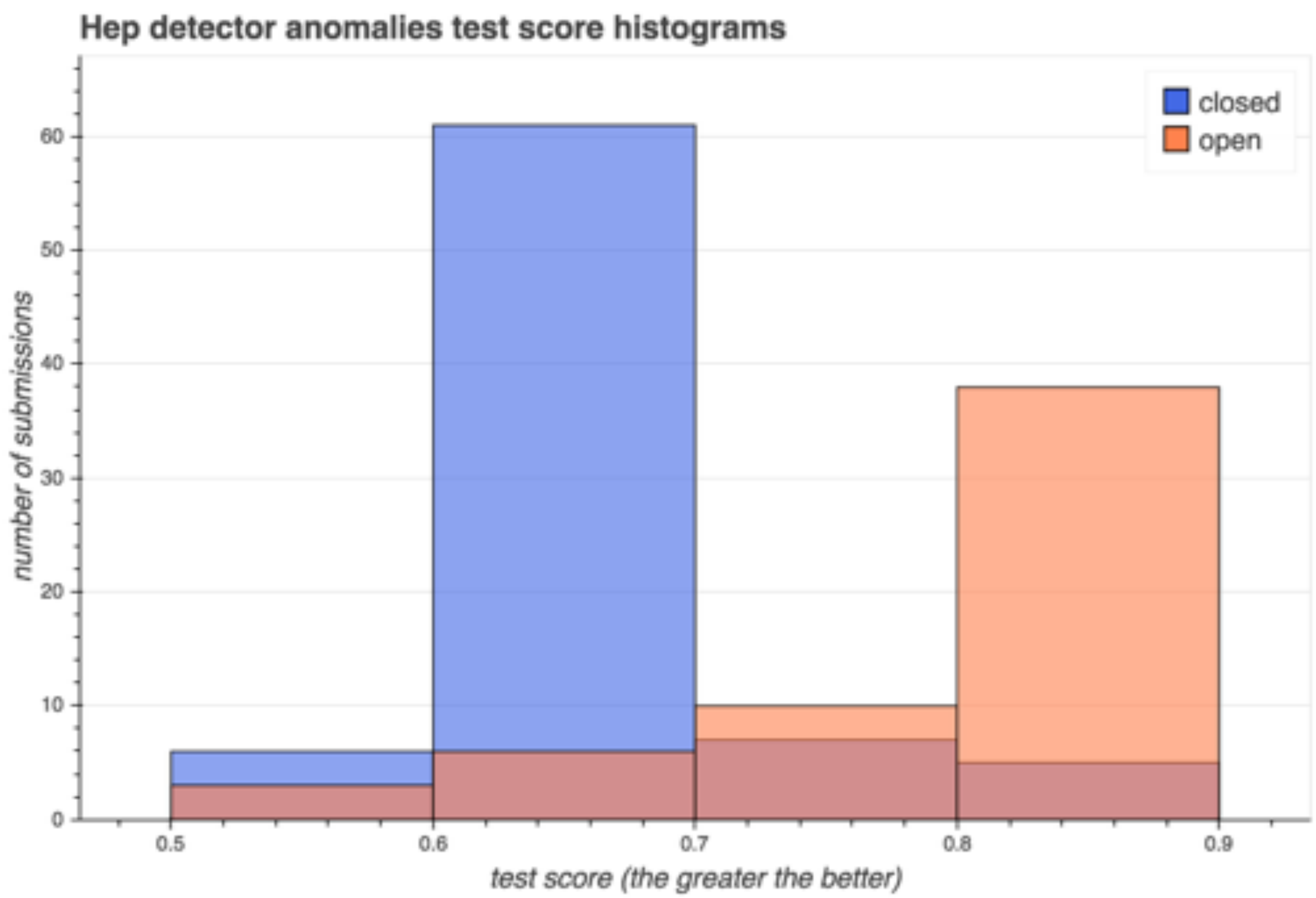


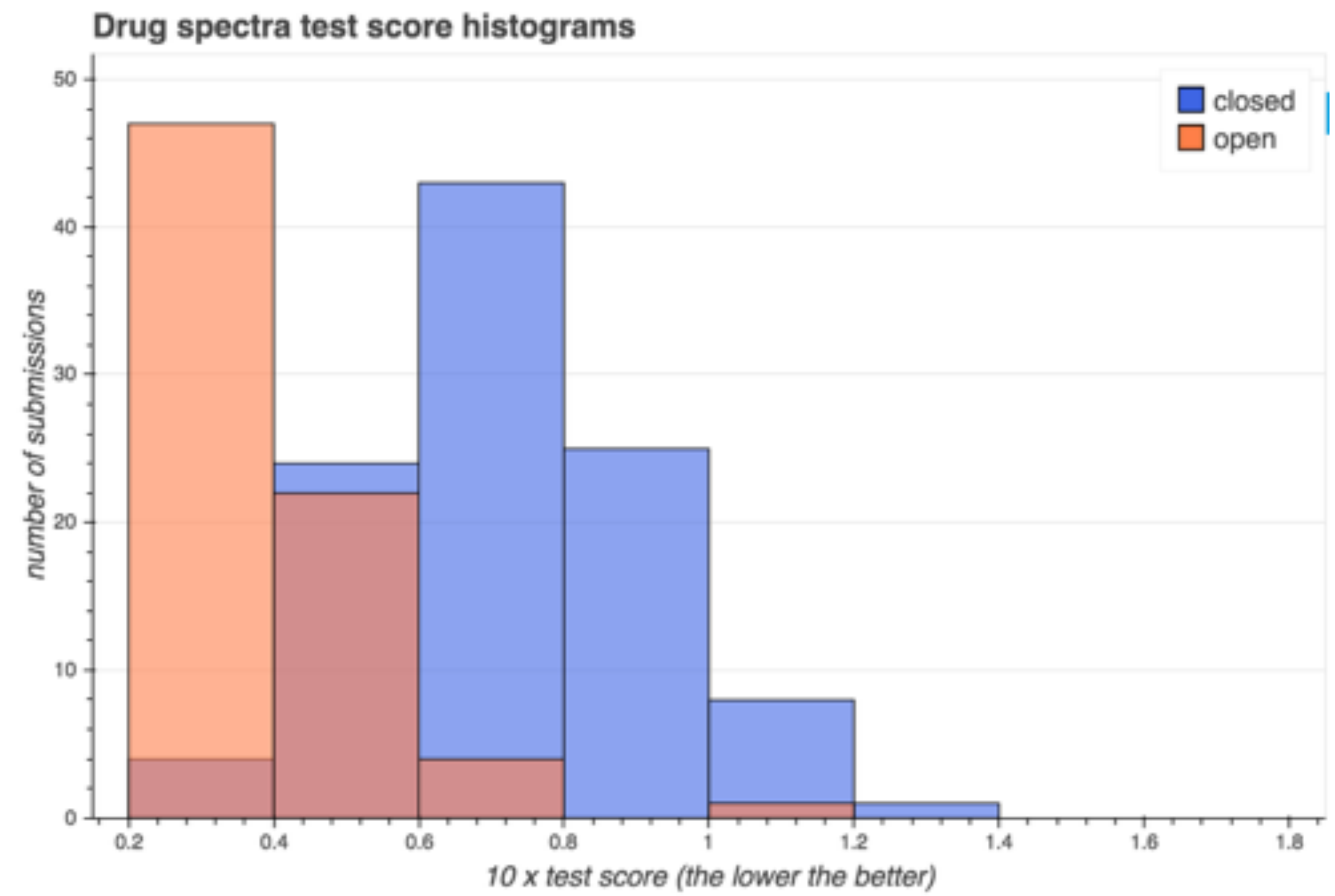
FORECASTING EL NINO SIX MONTHS AHEAD

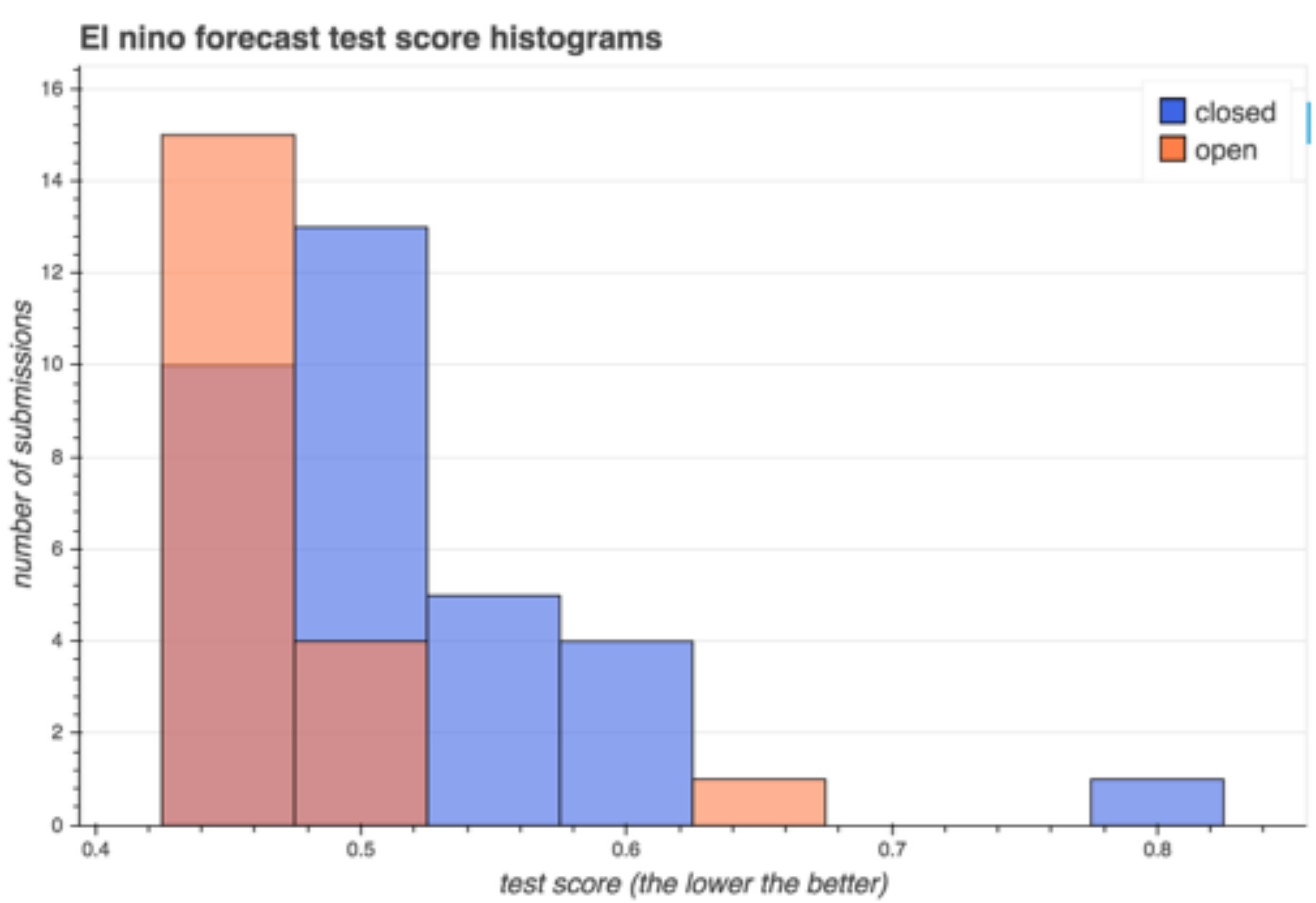


Analyzing the process

OPEN PHASE LETS PARTICIPANTS CATCH UP



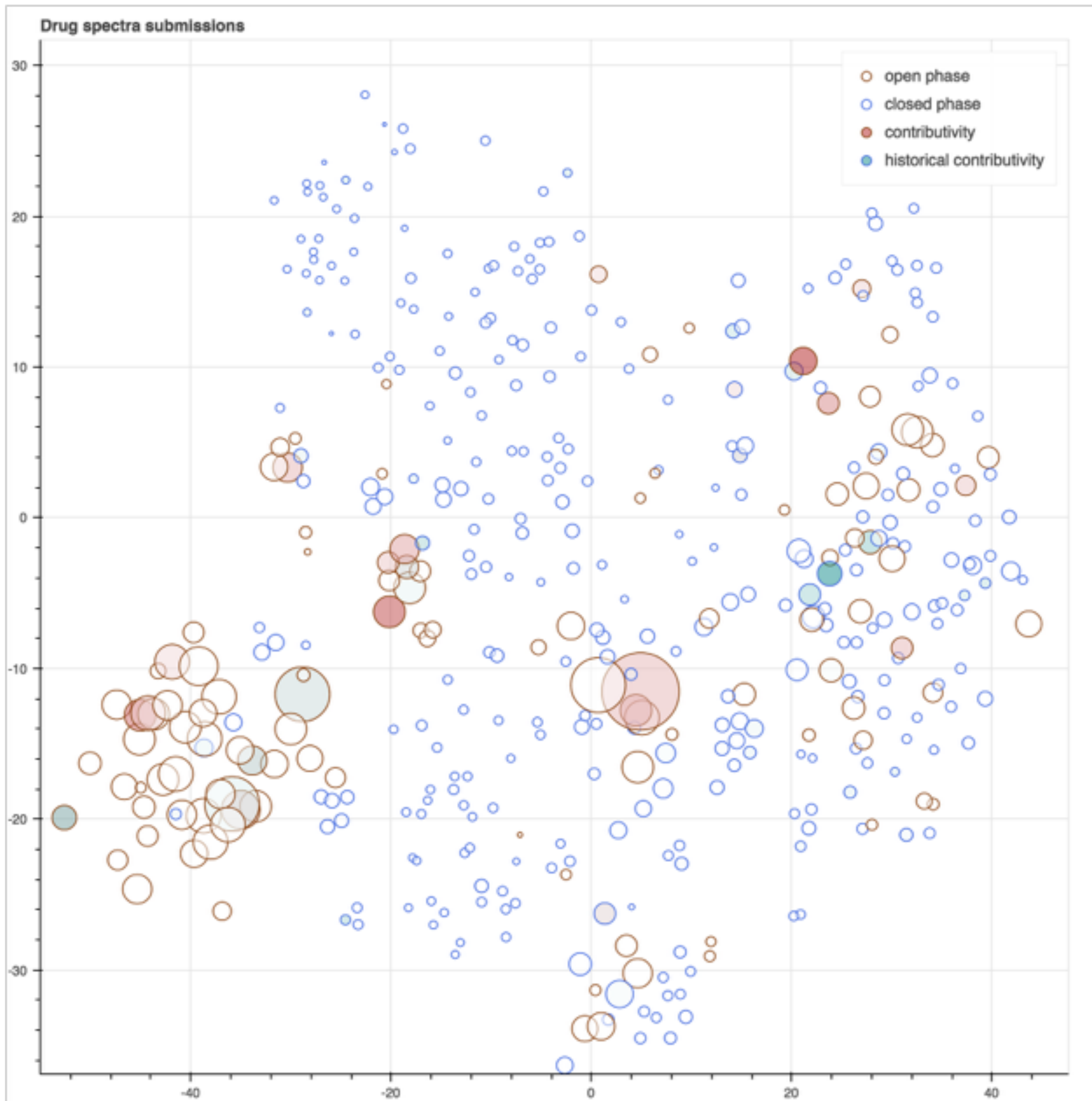




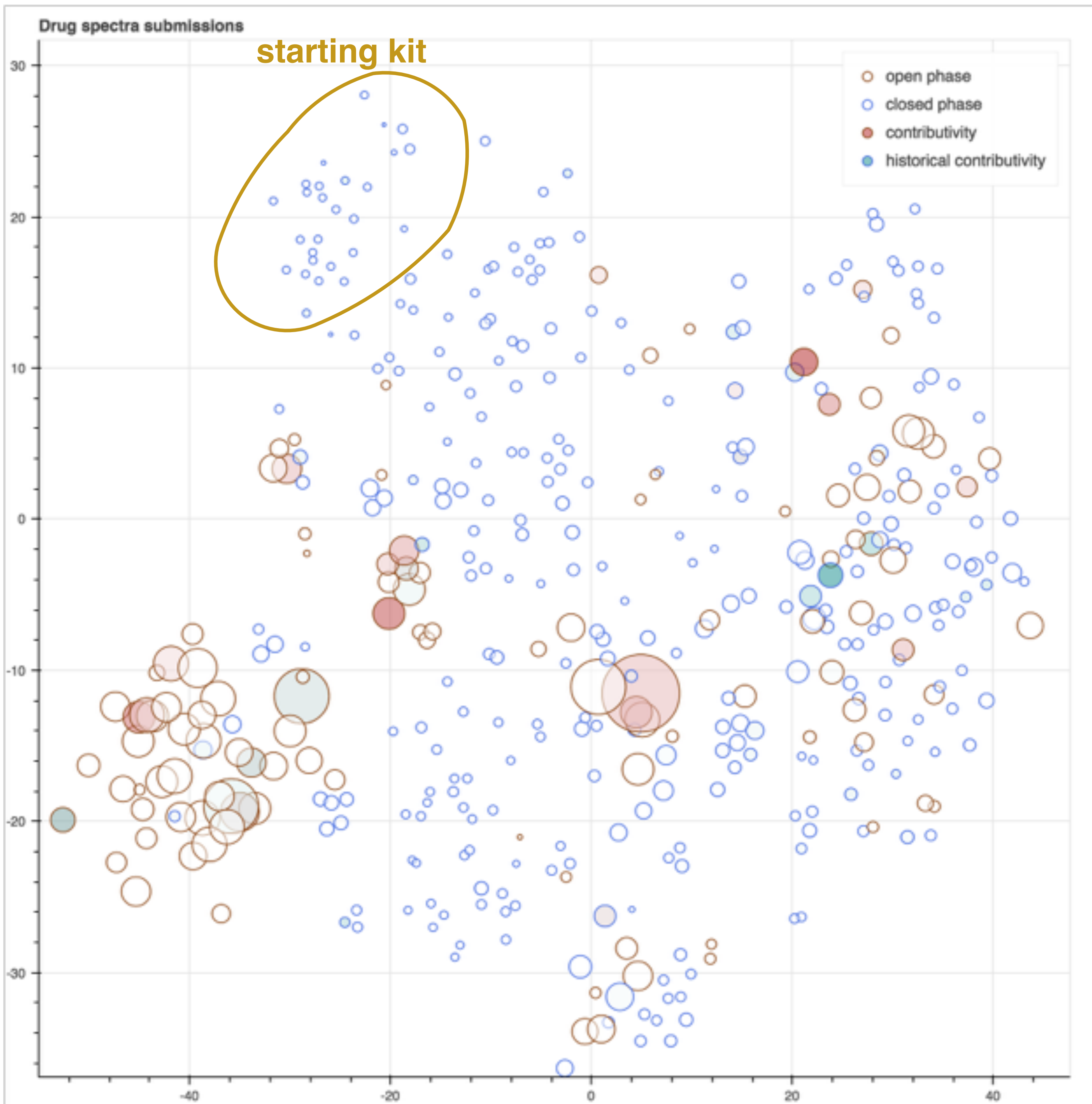
universite PARIS-SACLAY

Paris-Saclay Center for Data Science

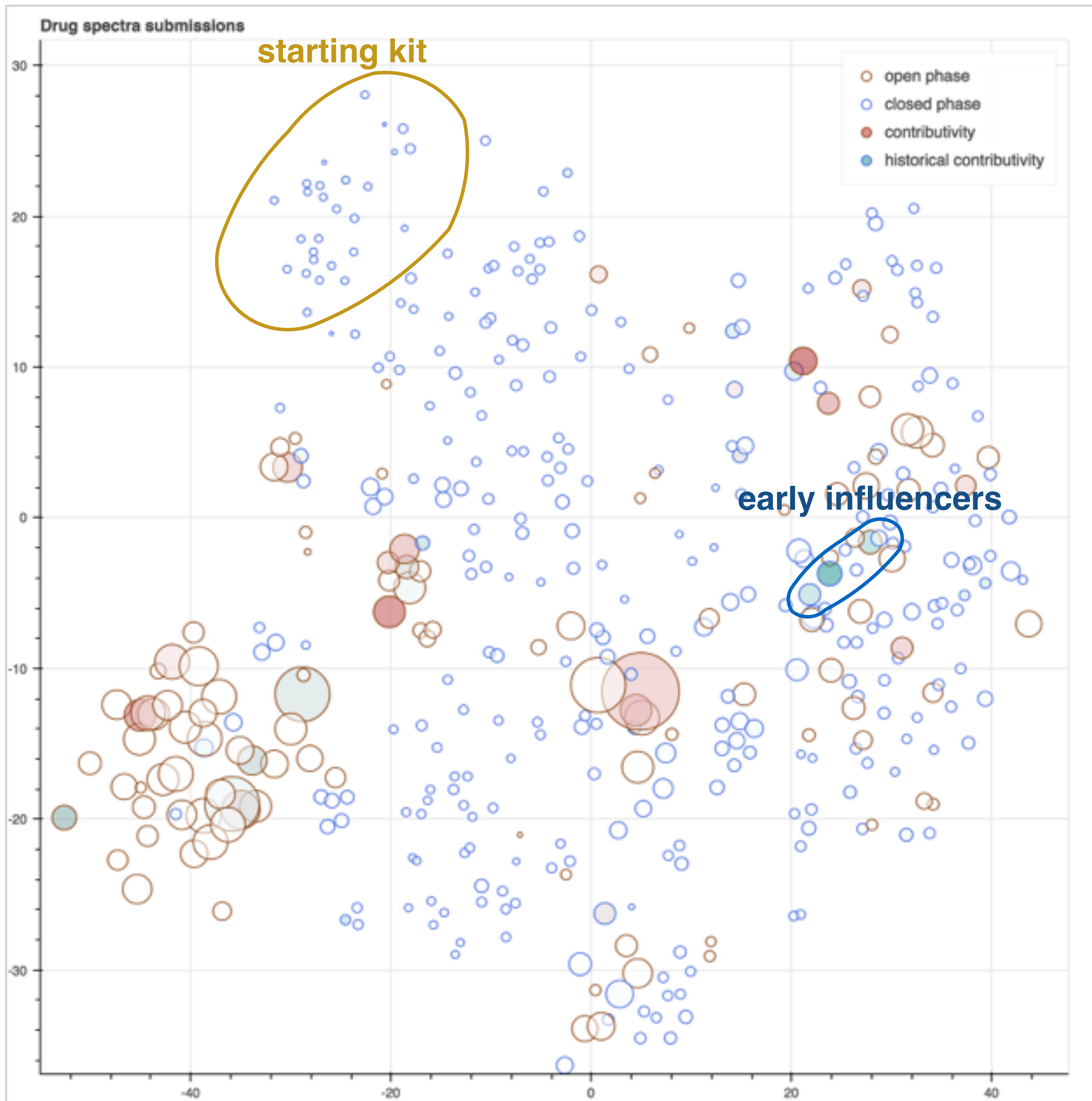
The dynamics of collaboration



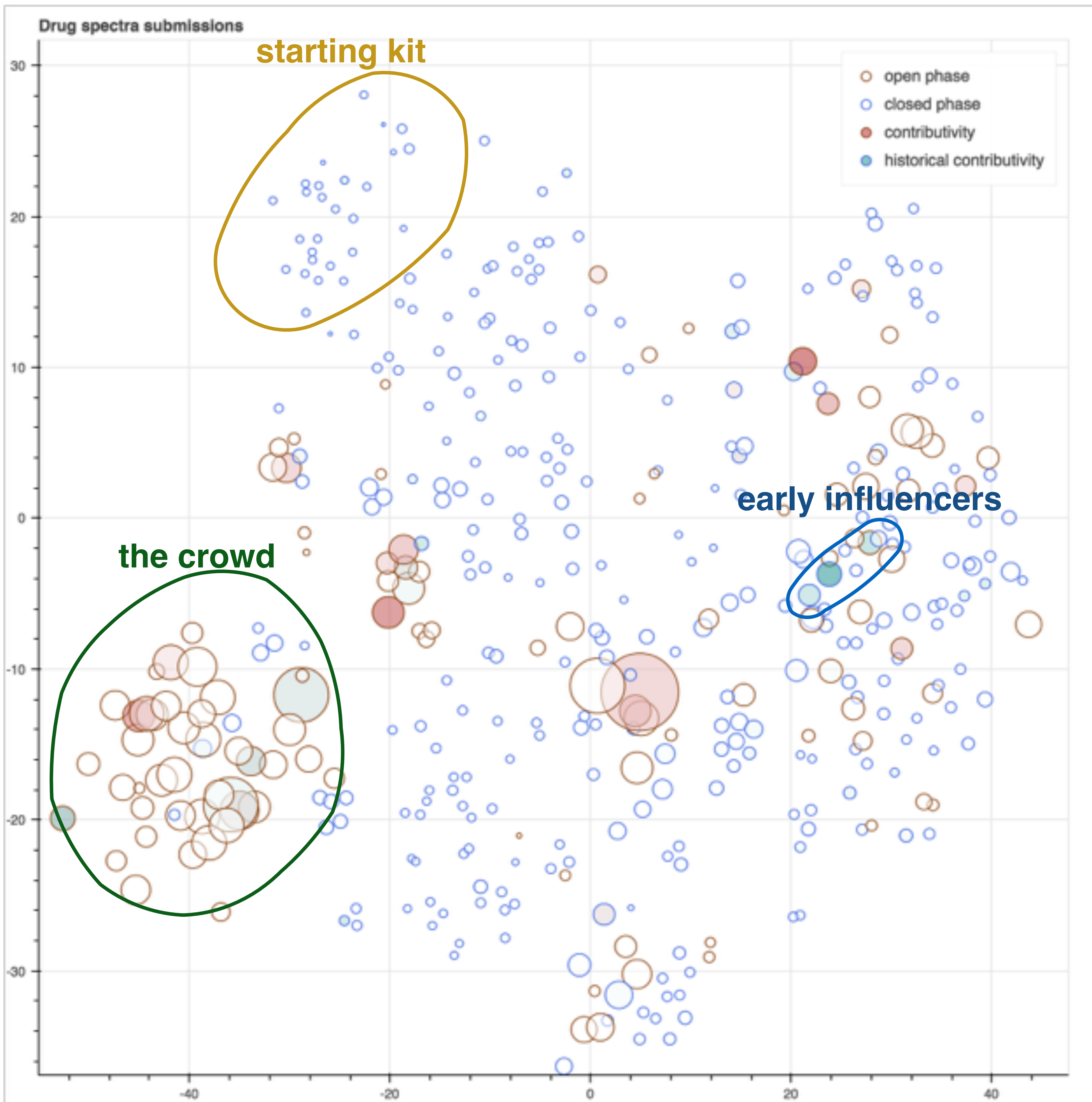
The dynamics of collaboration



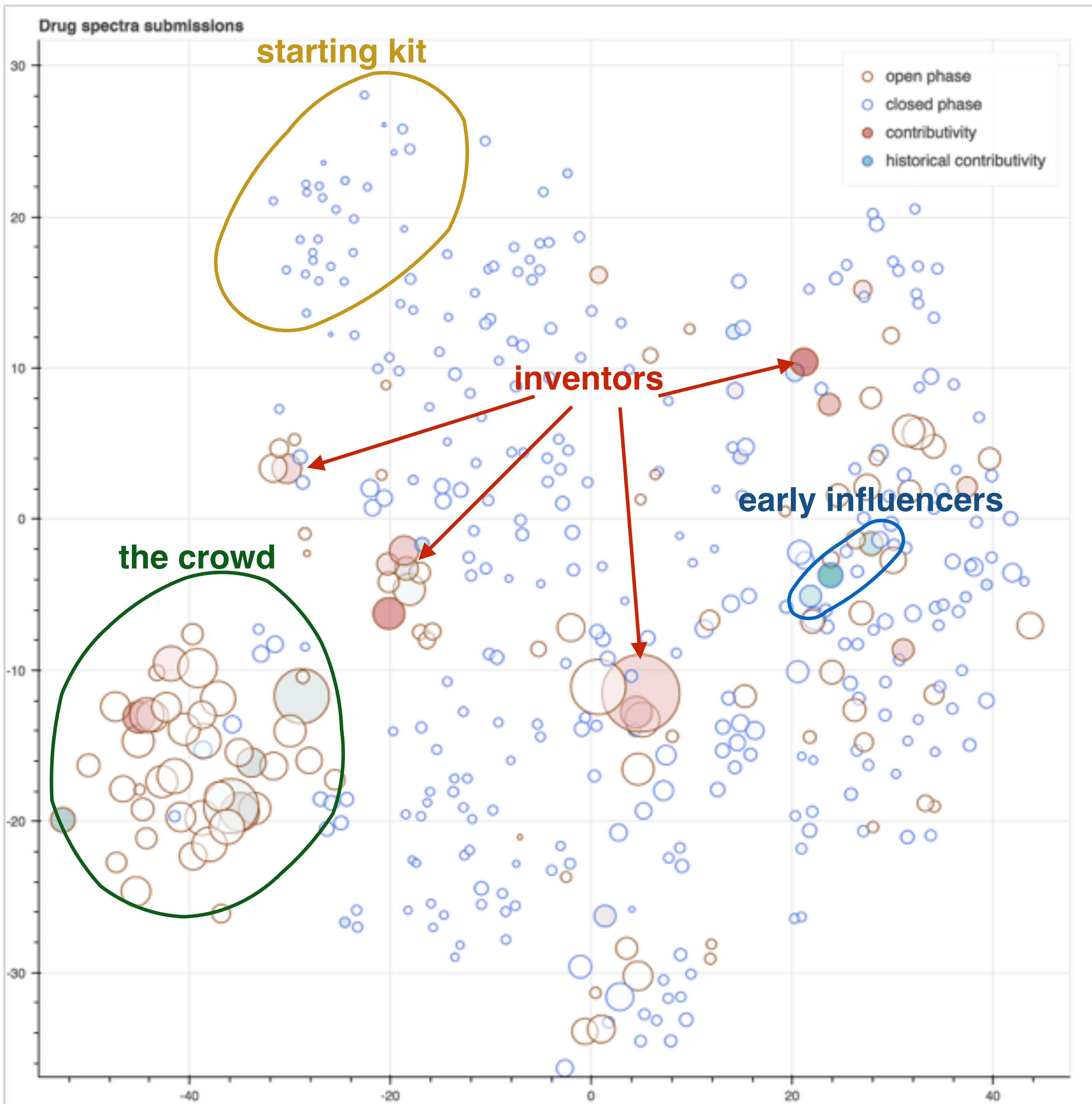
THE DYNAMICS OF COLLABORATION



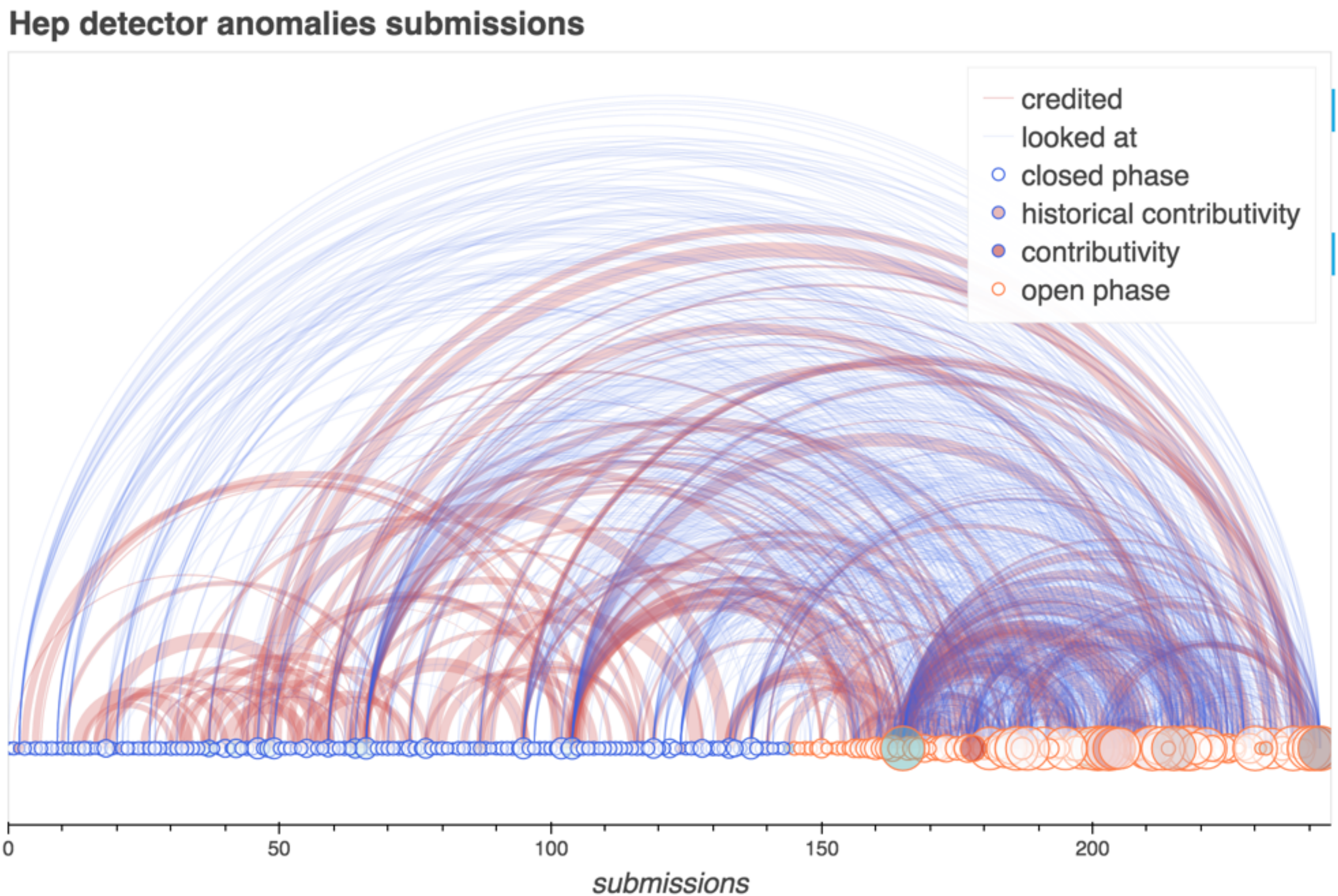
THE DYNAMICS OF COLLABORATION



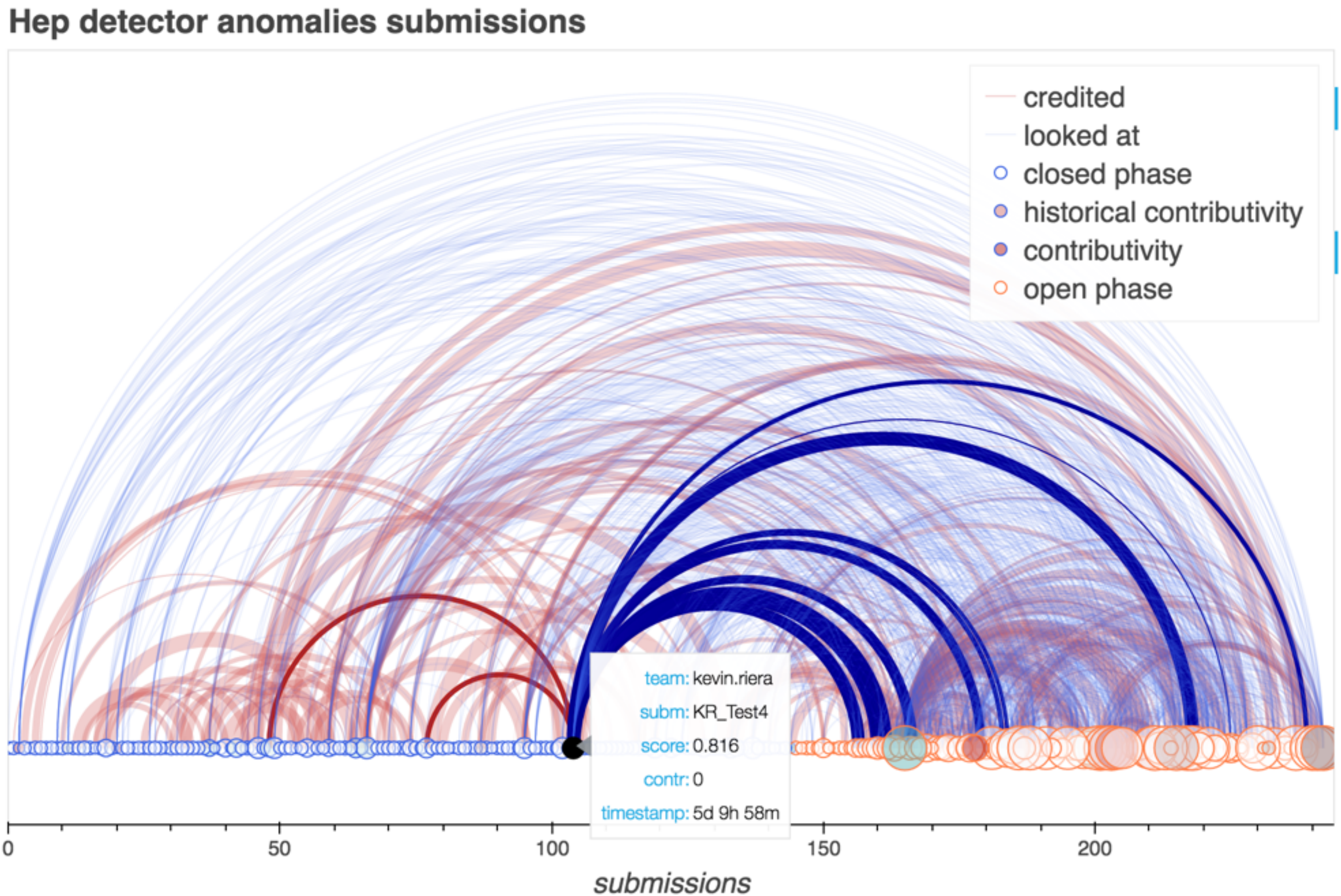
The dynamics of collaboration



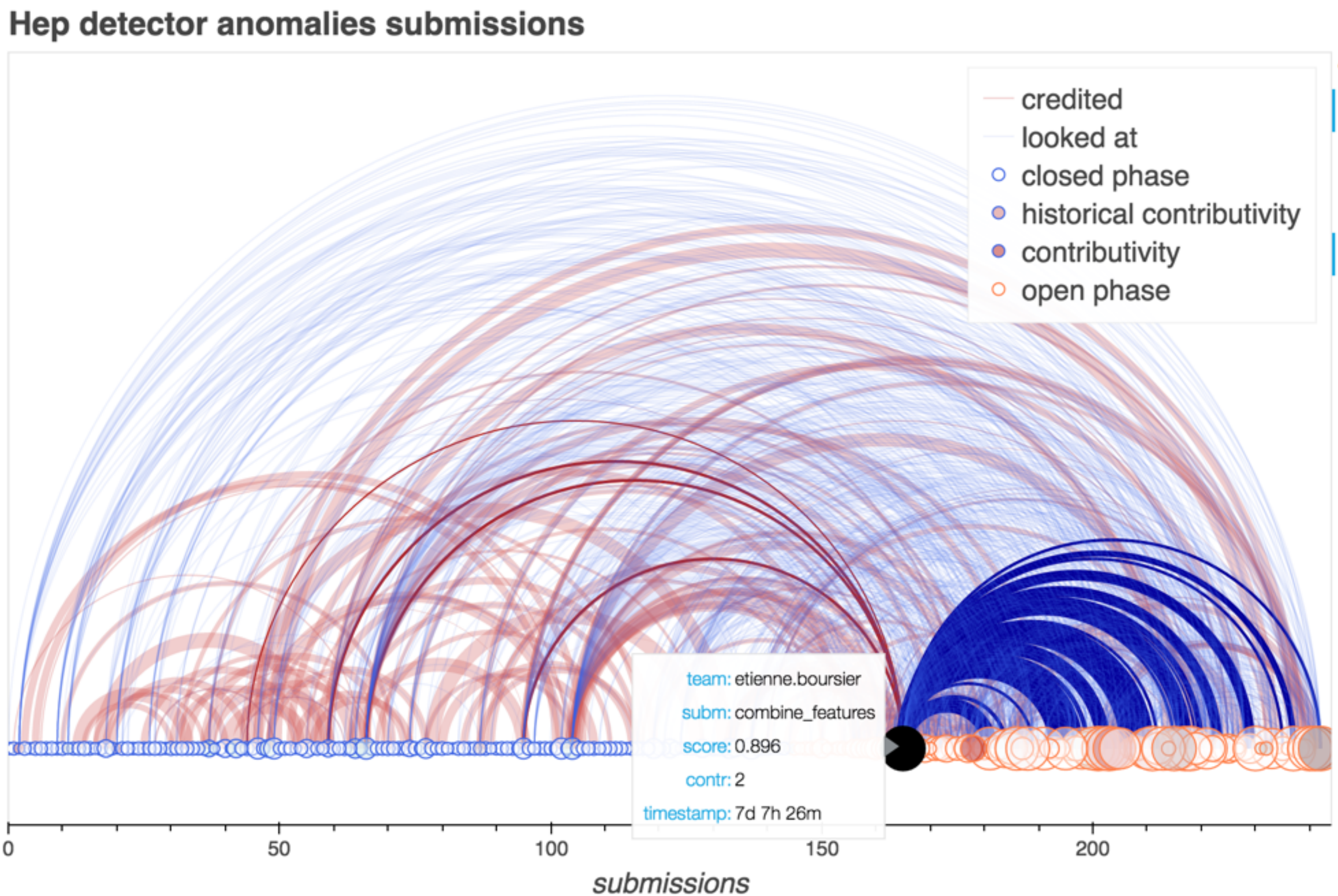
THE DYNAMICS OF COLLABORATION



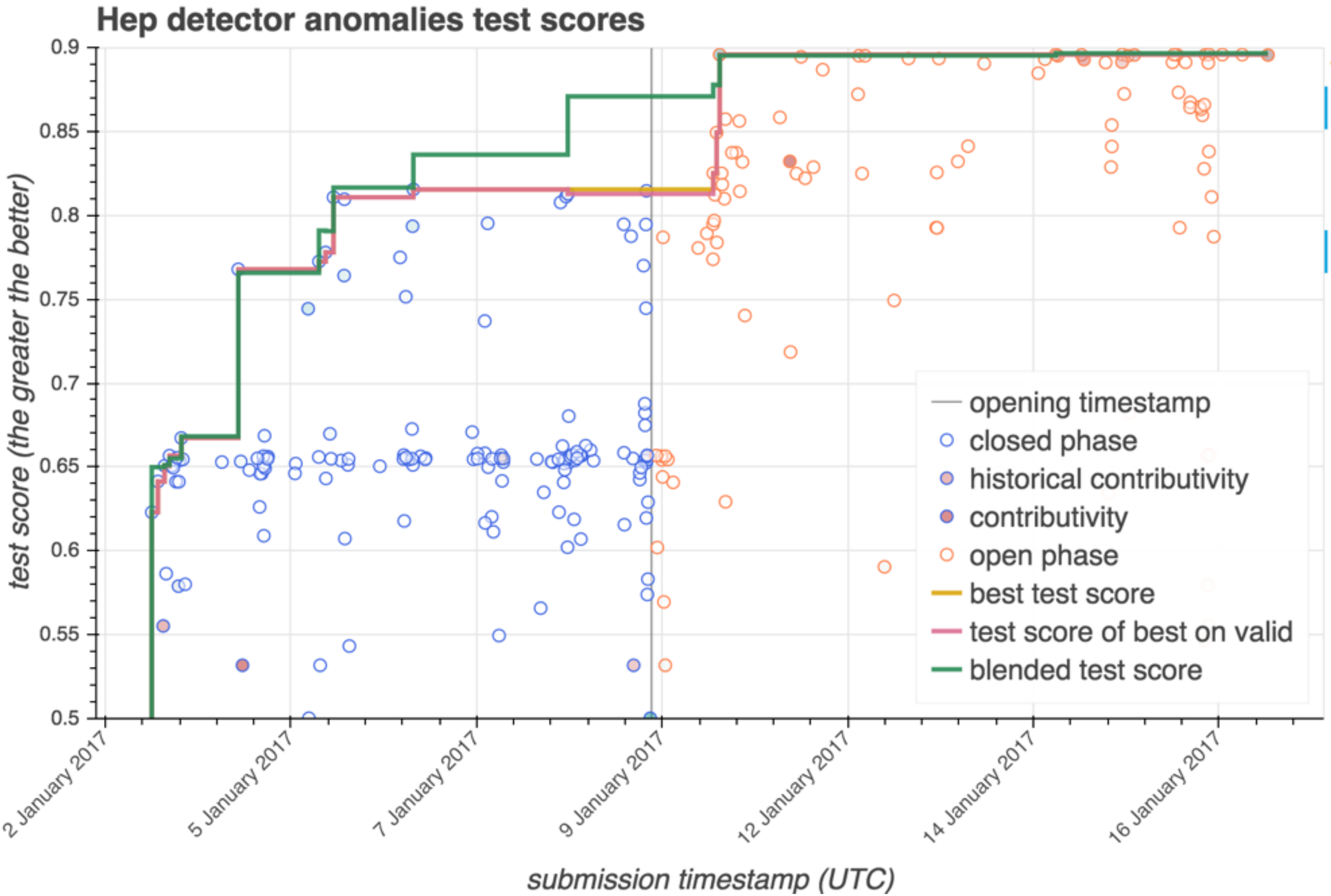
THE DYNAMICS OF COLLABORATION

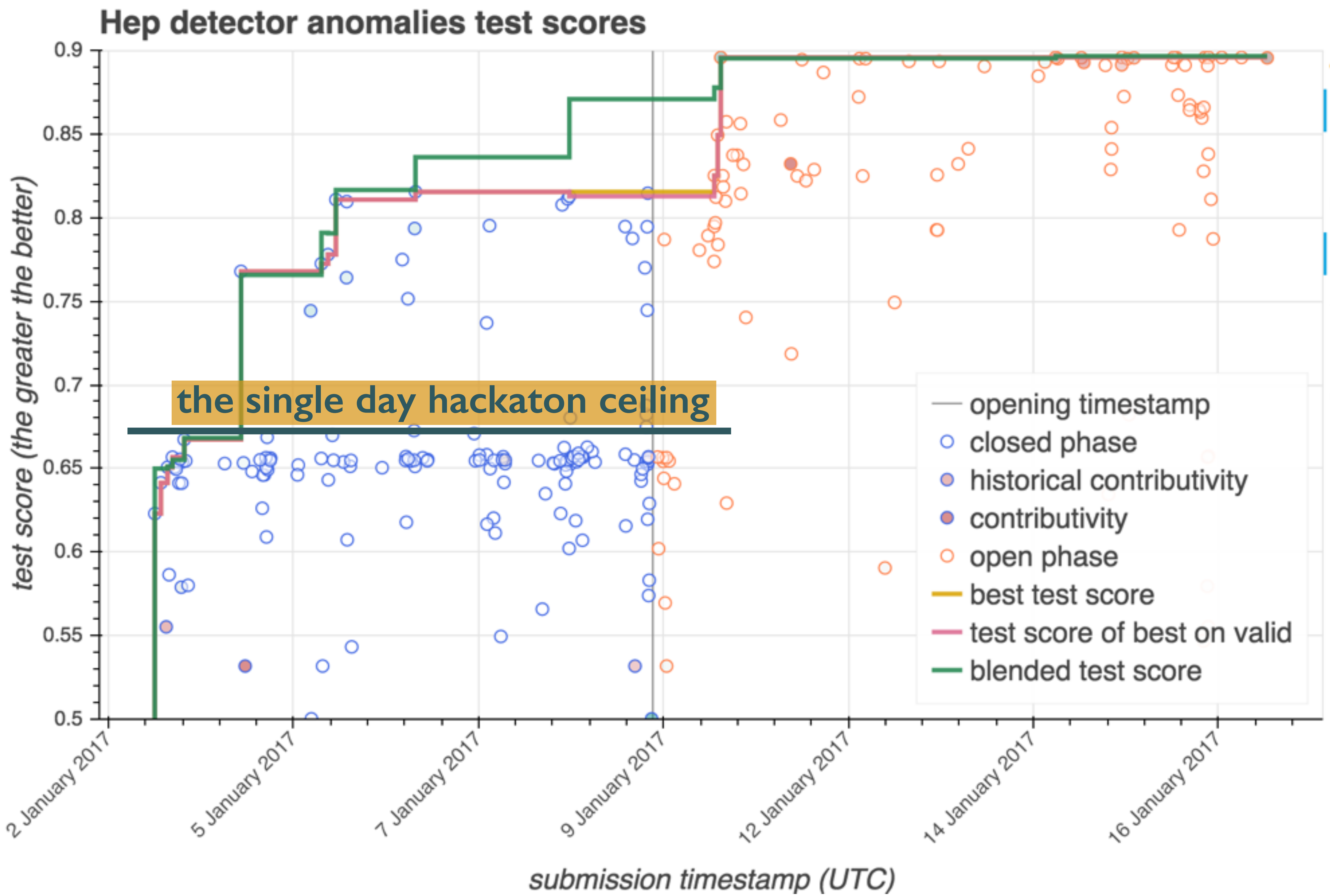


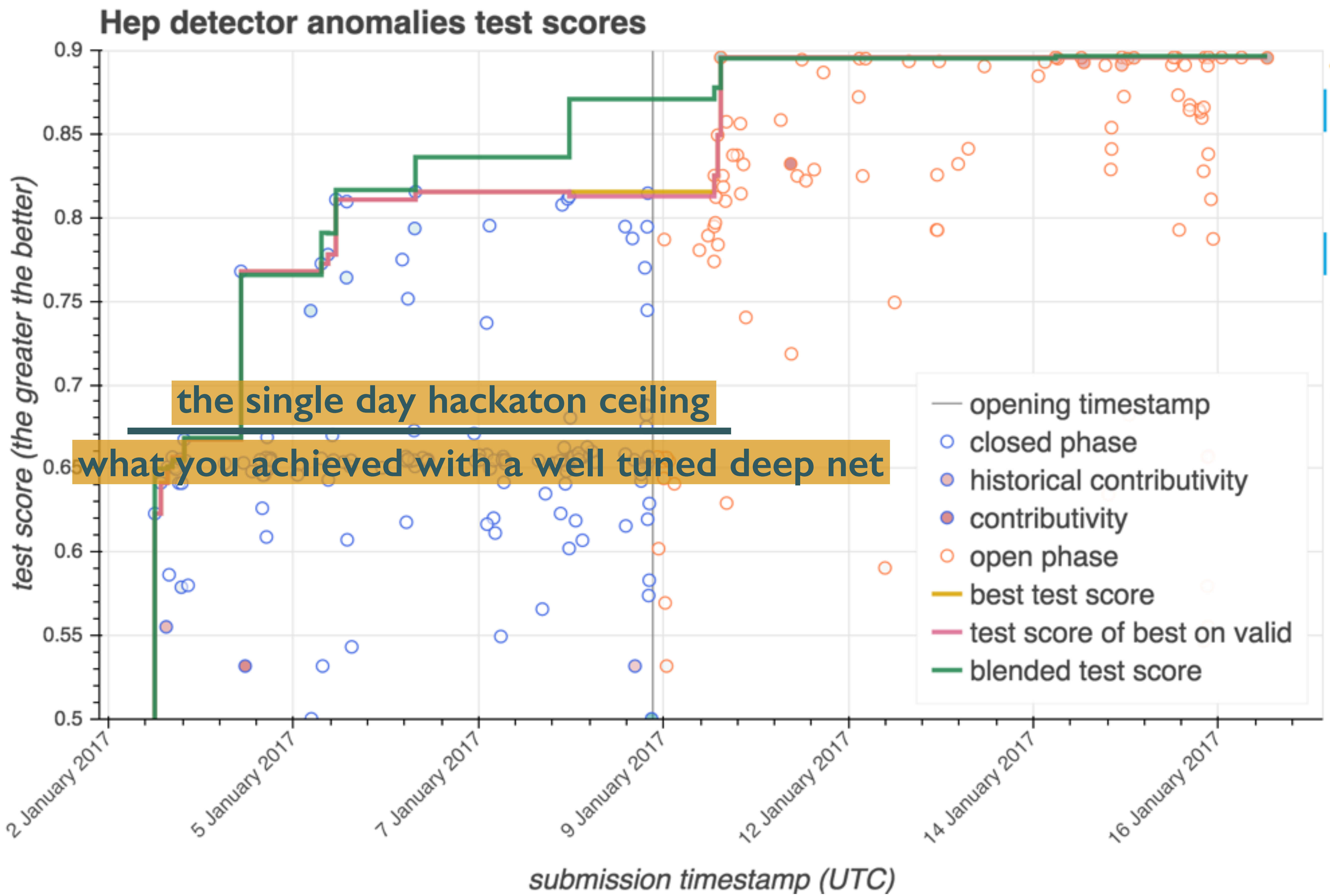
THE DYNAMICS OF COLLABORATION



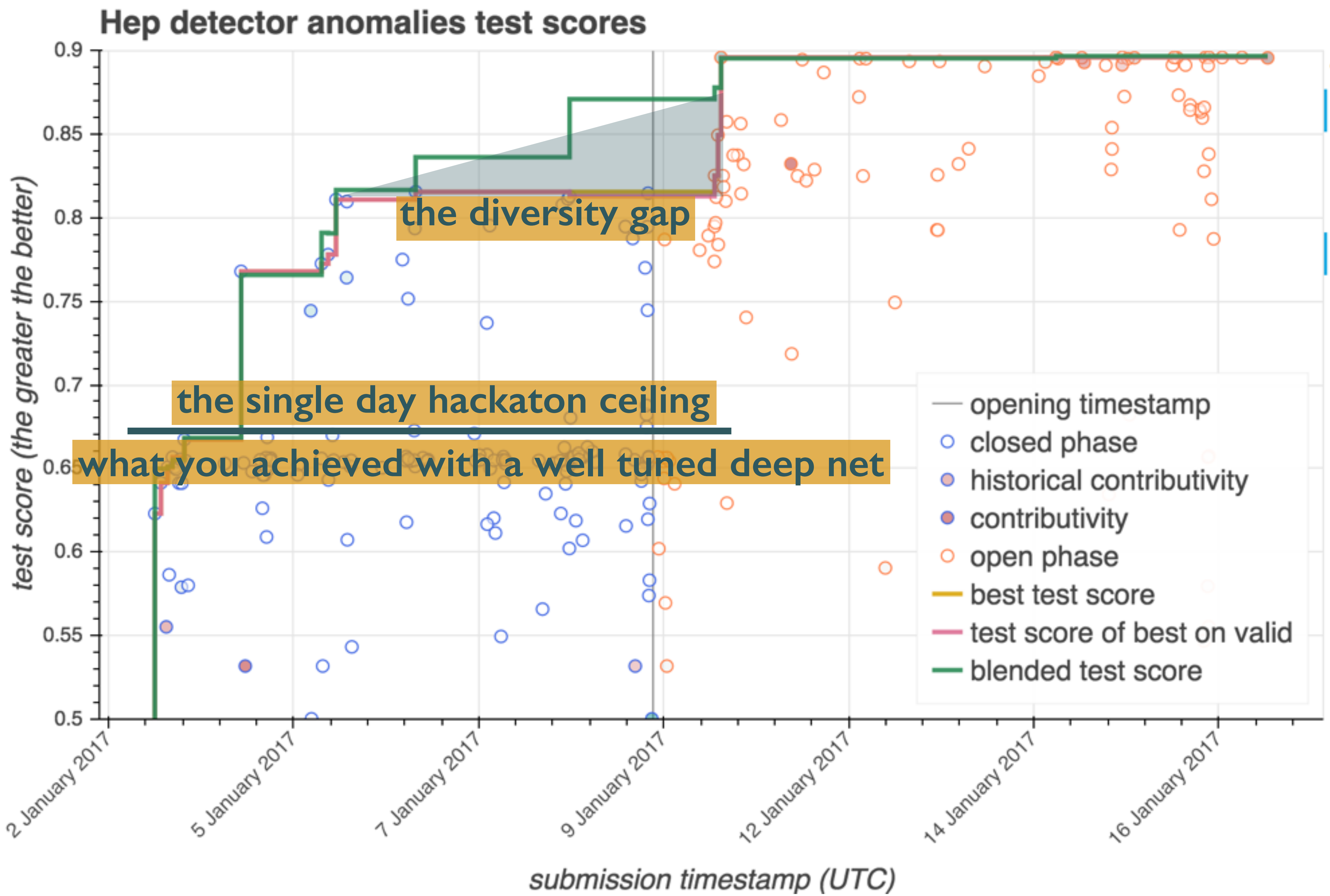
université PARIS-SACLAY Paris-Saclay Center for Data Science


Hep detector anomalies test scores
test score (the greater the better)
0.9
0.85
0.8
0.75
0.7
0.65
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)
opening timestamp
closed phase
historical contributivity
contributivity
open phase
best test score
test score of best on valid
blended test score


Hep detector anomalies test scores
test score (the greater the better)
0.9
0.85
0.8
0.75
0.7
0.65
0.6
0.55
0.5
the single day hackaton ceiling
opening timestamp
closed phase
historical contributivity
contributivity
open phase
best test score
test score of best on valid
blended test score
2 January 2017
5 January 2017
7 January 2017
9 January 2017
12 January 2017
14 January 2017
16 January 2017
submission timestamp (UTC)

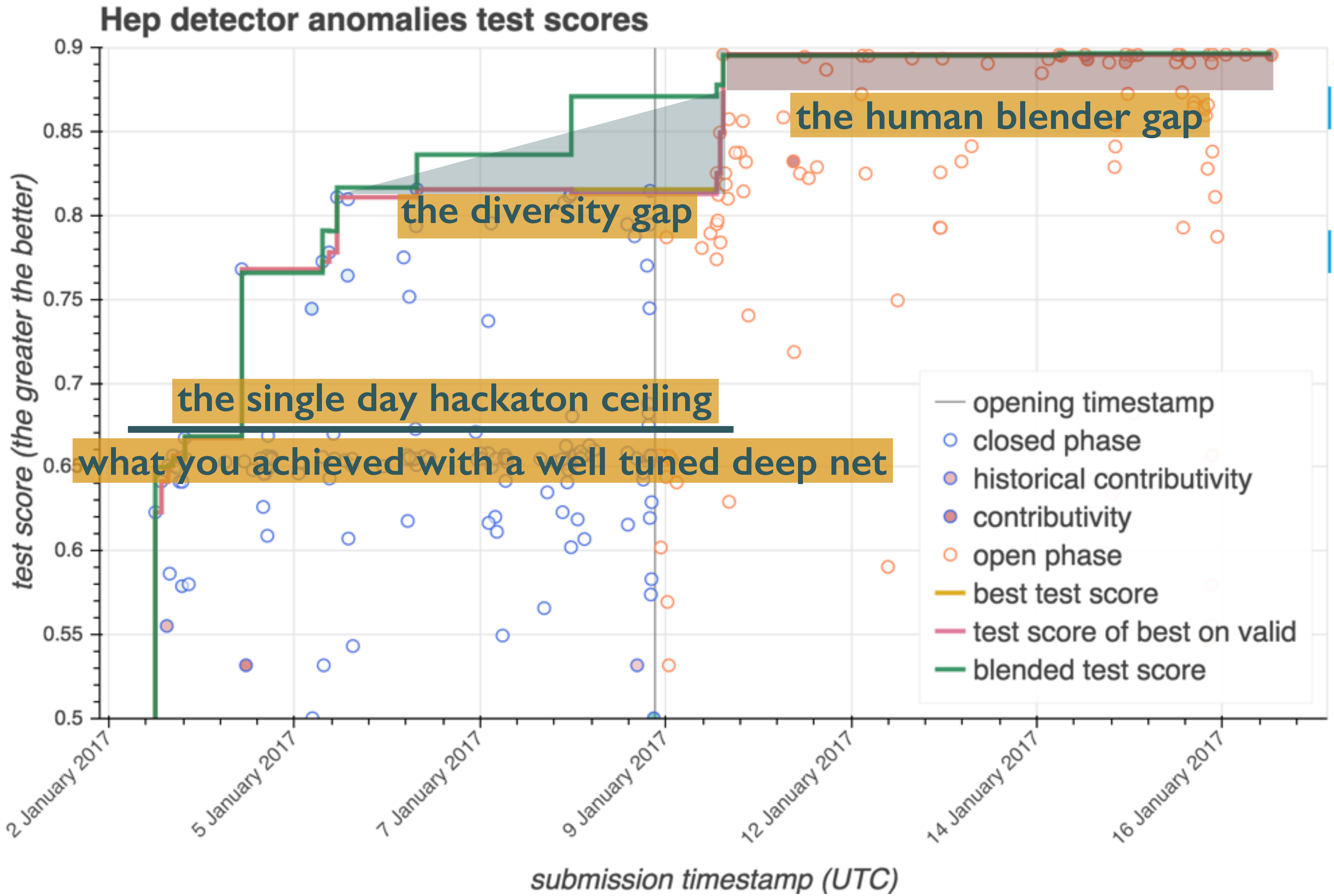


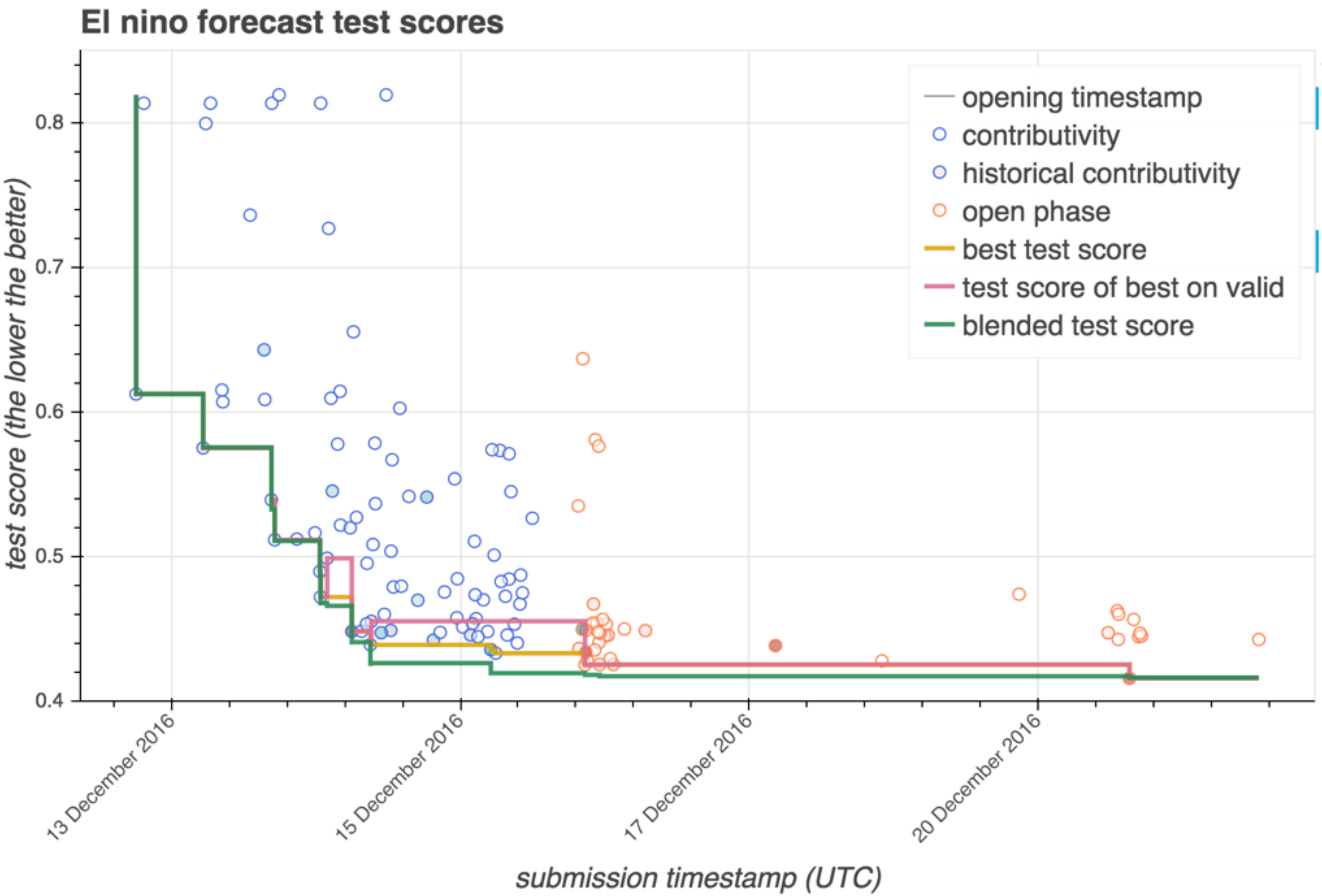
universite PARIS-SACLAY
Paris-Saclay Center for Data Science

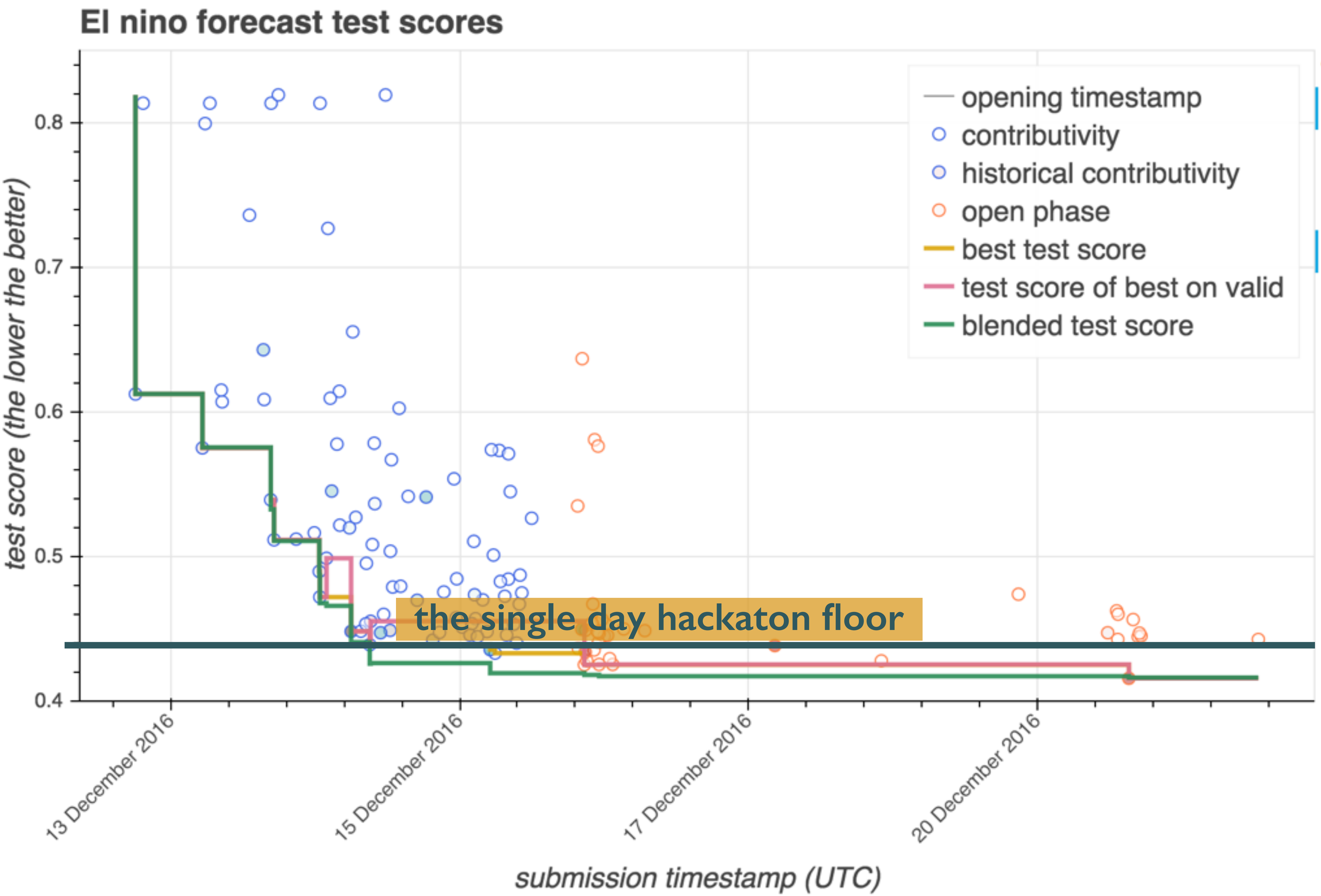


universite PARIS-SACLAY

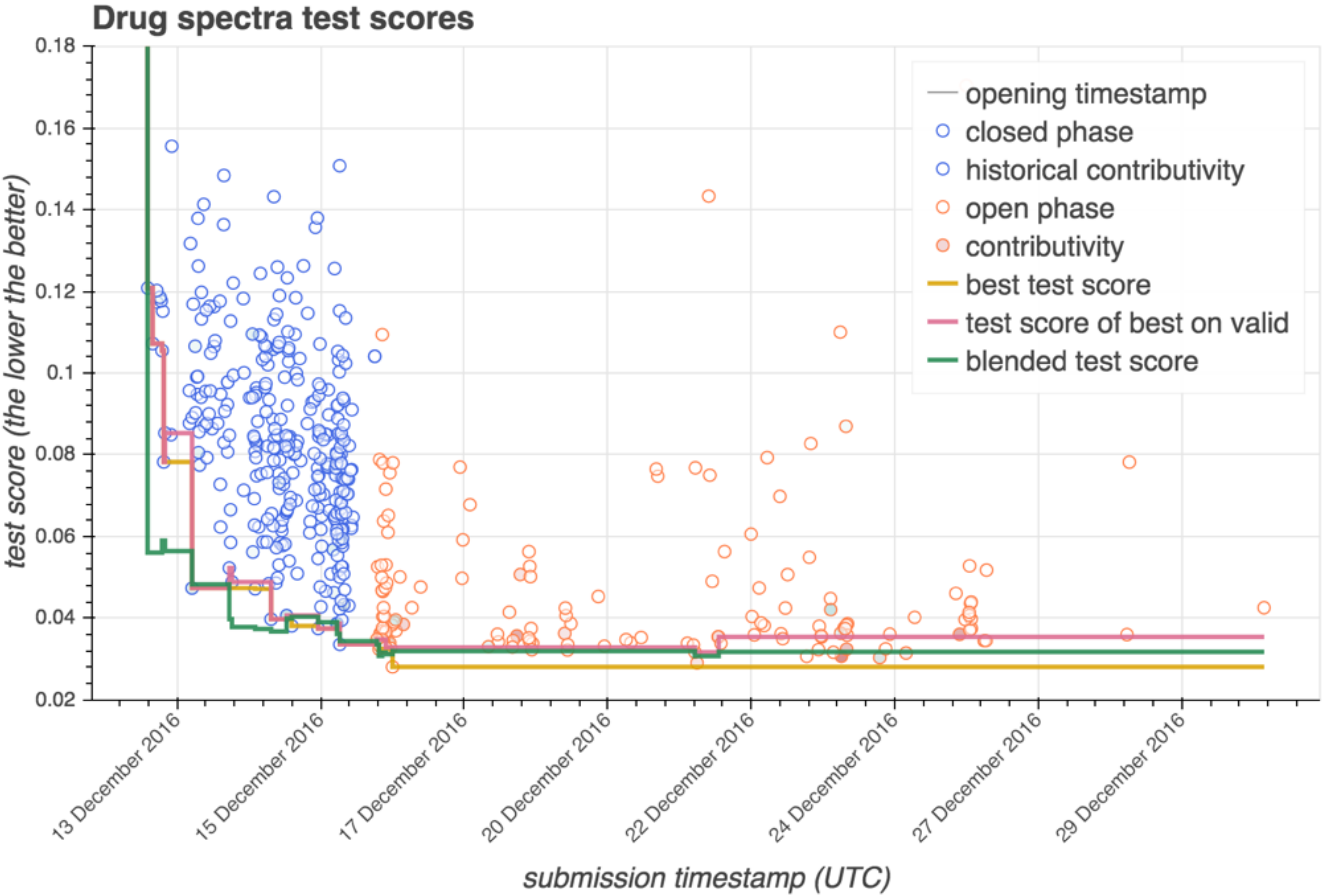
Paris-Saclay Center for Data Science

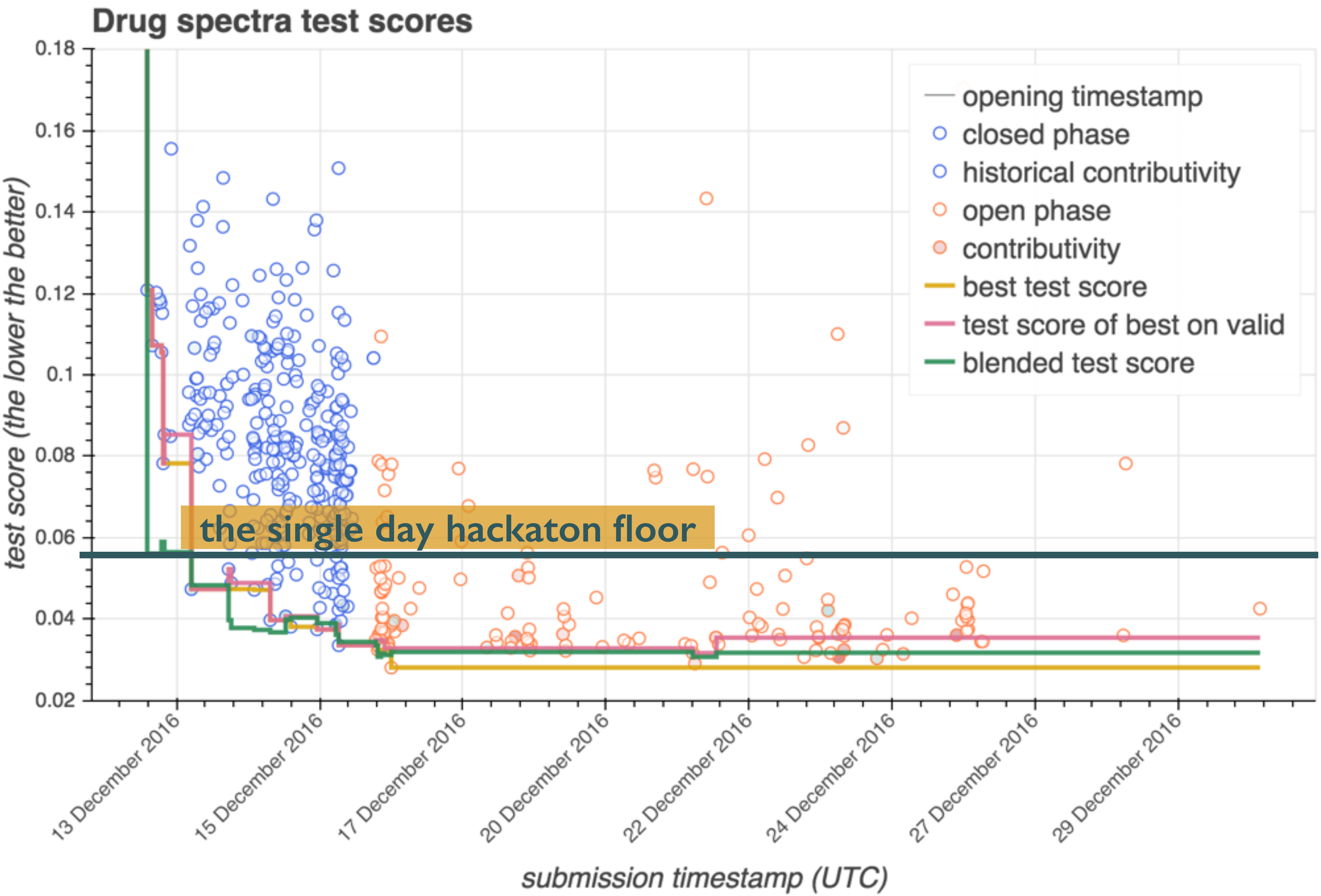

Hep detector anomalies test scores
test score (the greater the better)
0.9
0.85
0.8
0.75
0.7
0.65
0.6
0.55
0.5
submission timestamp (UTC)
2 January 2017
5 January 2017
7 January 2017
9 January 2017
12 January 2017
14 January 2017
16 January 2017
the human blender gap
the diversity gap
the single day 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


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



El nino forecast test scores
test score (the lower the better)
0.8
0.7
0.6
0.5
0.4
opening timestamp
contributivity
historical contributivity
open phase
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
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
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)


Drug spectra test scores
test score (the lower the better)
0.18
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
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

What we learned

- **Open phase helps novice participants to catch up: the goal of teaching!**
  - Sometimes also makes the best and blended score better

What we learned

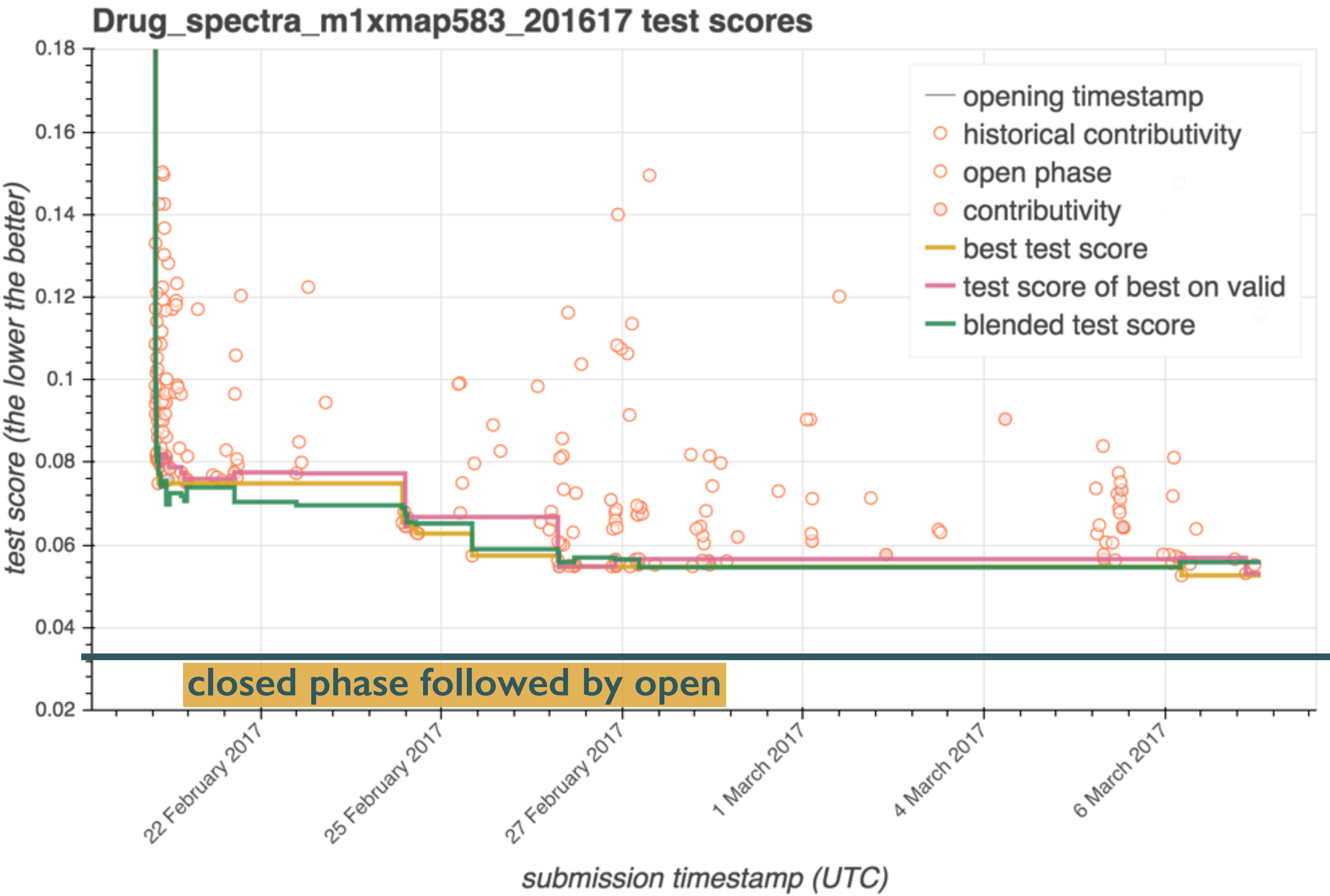
- Open phase helps novice participants to catch up: the goal of teaching!
  - Sometimes also makes the best and blended score better
- Human blending often beats machine blending

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

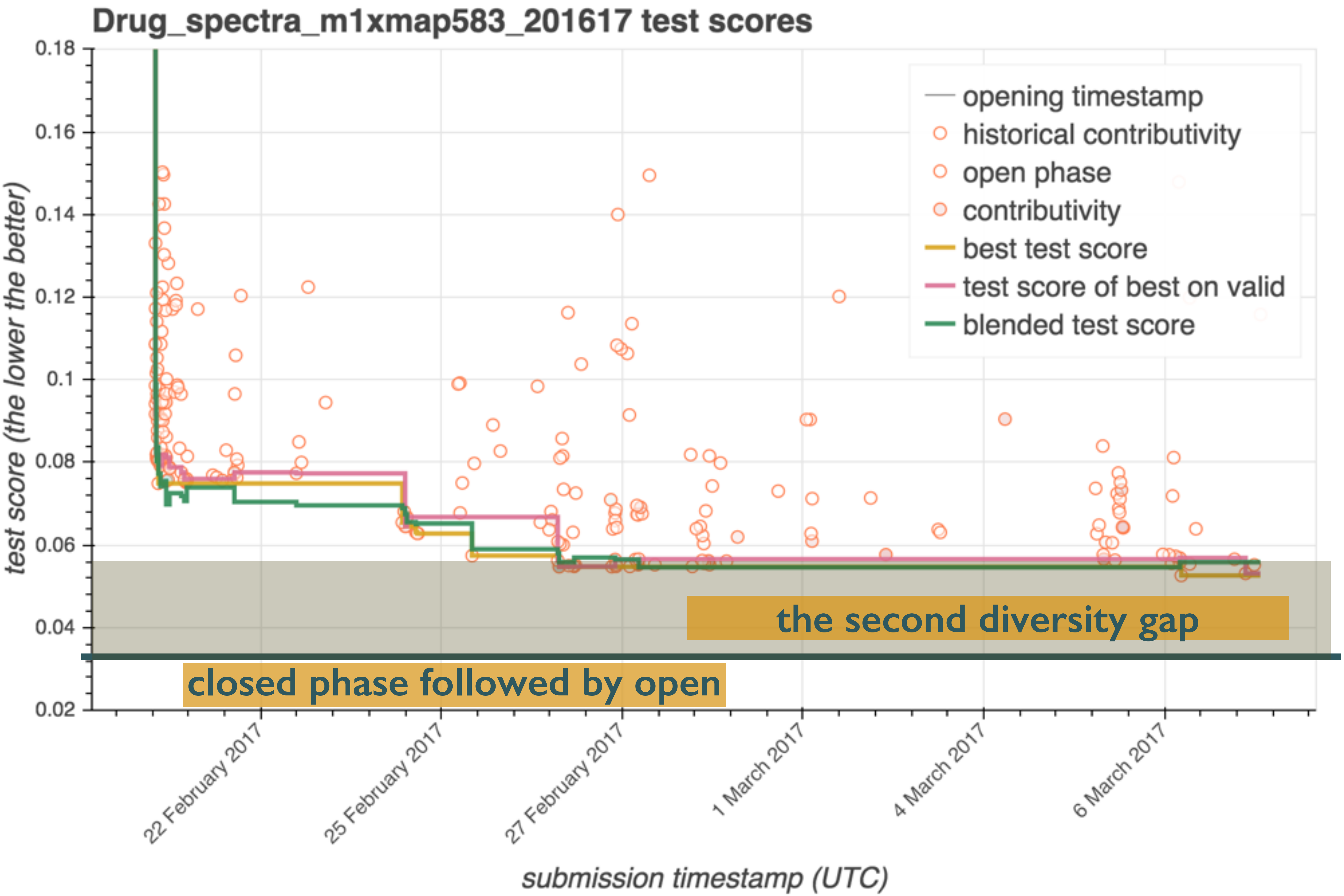
WHAT WE LEARNED

- **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?

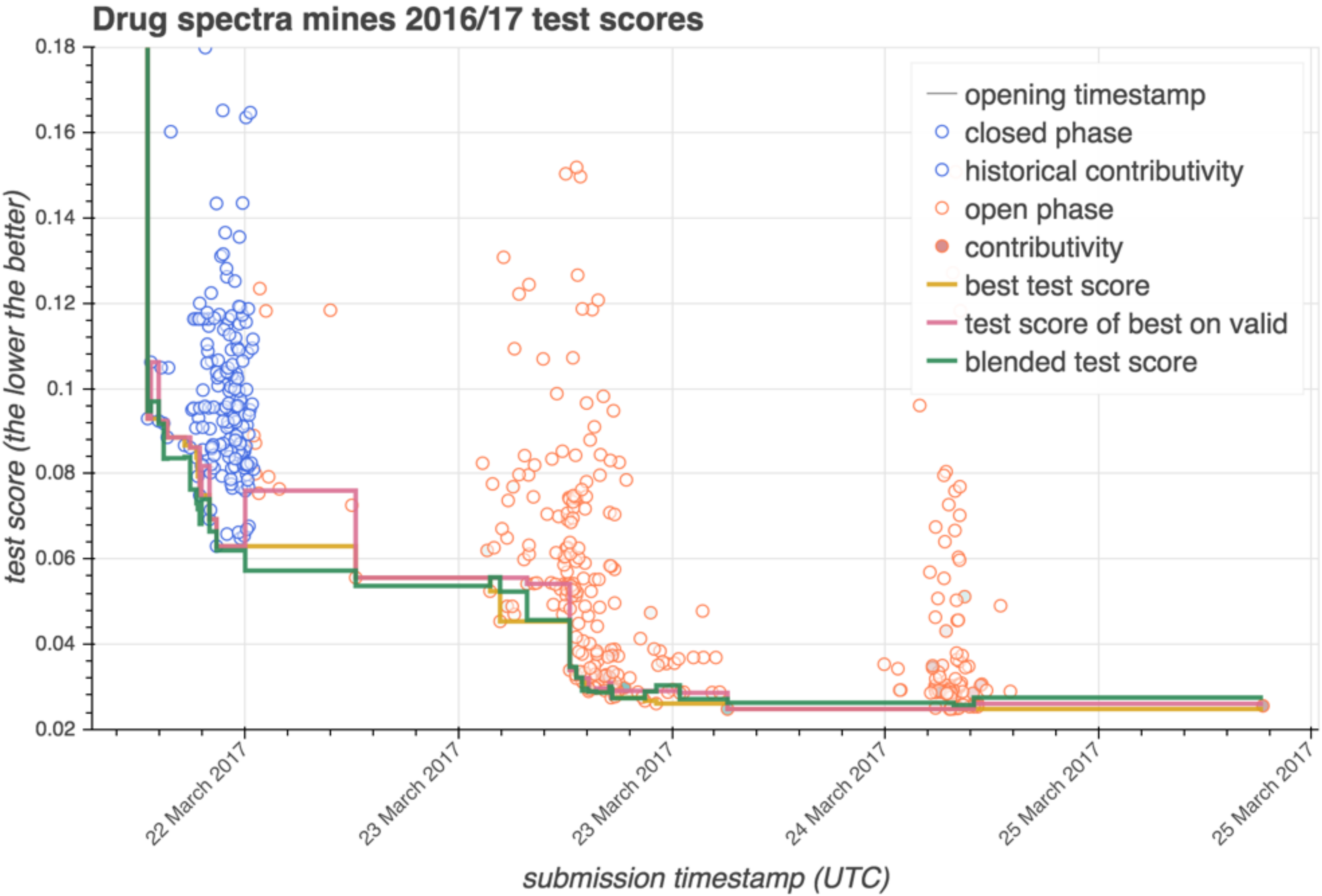


closed phase followed by open

université PARIS-SACLAY
Paris-Saclay Center for Data Science



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

What's next

WHAT'S NEXT

- More RAMPs
  - galaxy morphology, detecting autism from brain fMRI, detecting Mars craters, forecasting space weather (solar storm early warning)

What's next

- More RAMPs
  - galaxy morphology, detecting autism from brain fMRI, detecting Mars craters, forecasting space weather (solar storm early warning)
- More courses
  - >1000 students next year

What's next

- **More RAMPs**
  - galaxy morphology, detecting autism from brain fMRI, detecting Mars craters, forecasting space weather (solar storm early warning)
- **More courses**
  - >1000 students next year
- **Build your own RAMPs**

What is in it for you

- If you are a **data science teacher**
  - we have been using **classroom RAMPs in different formats**: homework, final project, data camp
  - students **love it and work** their butt off, they **learn from each other** and **collaborate**
- If you are a **domain science researcher or data scientist**
  - We can **solve your predictive problems** better than any single researcher in a classical project
- If you are a **data science researcher or engineer**
  - You can **benchmark your new techniques** on a variety of problems

sign up: www.ramp.studio

contact us: balazs.kegl@gmail.com