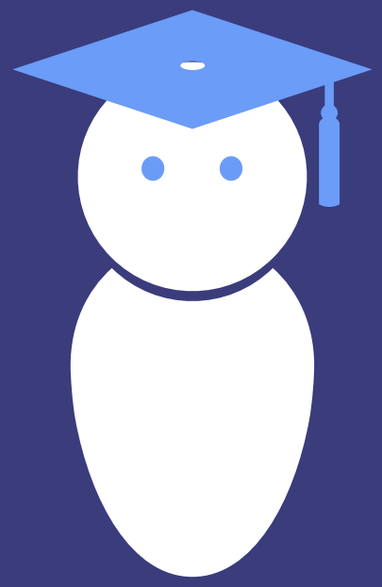
OpenML

machine learning, better, together

Joaquin Vanschoren & the OpenML team

www.openml.org

@open_ml



The Science of Machine Learning

Understand how to learn effectively from data
How to design/select algorithms (meta-learning)

… has a ***lot*** of friction…

Datasets

scattered, unclear tasks, unparseable

Algorithms

undocumented, crashing

Models, Scores

inaccessible, unorganized

Reproducibility

missing details, different results

The same algorithm can learn to walk in wildly different ways. YUVAL TASSA

Missing data hinder replication of artificial intelligence studies

By Matthew Hutson | Feb. 15, 2018 , 12:30 PM

website called OpenML. It hosts not only algorithms, but also data sets and more than 8 million experimental runs with all their attendant details. "The exact way that you run your experiments is full of undocumented assumptions and decisions," Vanschoren says. "A lot of this detail never makes it into papers."

What if…
we could map the entire universe?

What if…
we could organize the world’s
machine learning information?

For every model, get the *exact* dataset and algorithm used

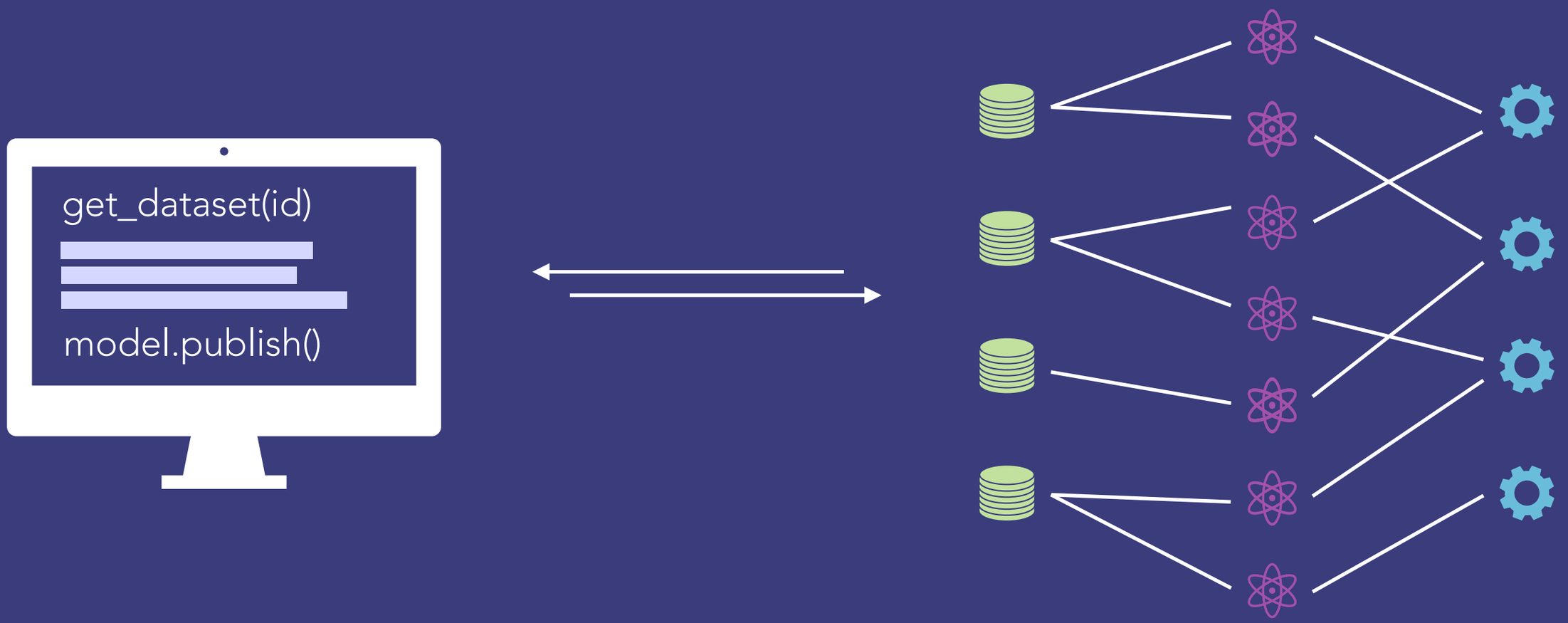
For every dataset, find all models built (and how good they are)

For every algorithm, find how useful it is for every dataset

Networked Science

Move data out of our labs and minds, into online tools that help us organize the information, and reuse it in unexpected new ways.

(M. Nielsen)



Only scales if all sources of friction are removed

Frictionless machine learning

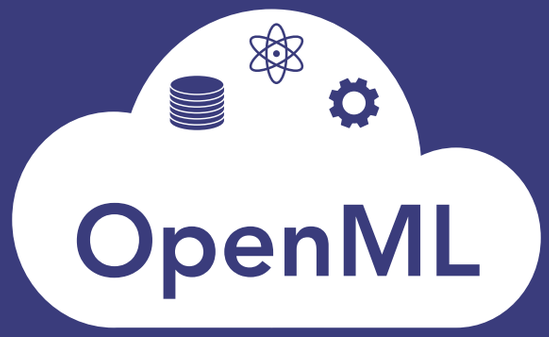
- Share datasets directly from existing environments
- Auto-compute uniform, rich meta-data
- Frictionless discovery and import for any dataset

- Share algorithms/pipelines from existing environments
- Generate uniform, tool-independent descriptions

- Link produced models to exact data and algorithms
- Objective evaluations, organized online

- *Automatically* extract *all* details for reproducibility
- Clean versioning of data and code

- Auto-track contributions, give credit where it's due



www.openml.org Python API R API Java API C# API CLI ...

Find datasets/algorithms/models, and share your own

APIs import/export everything (semi)automatically

Servers organize all information, evaluate models

OpenML Search HELP SIGN IN

www.openml.org

OpenML beta_2

Machine learning, better, together

20215 data sets

Find or add **data** to analyse

68210 tasks

Download or create scientific **tasks**

6173 flows

Find or add data analysis **flows**

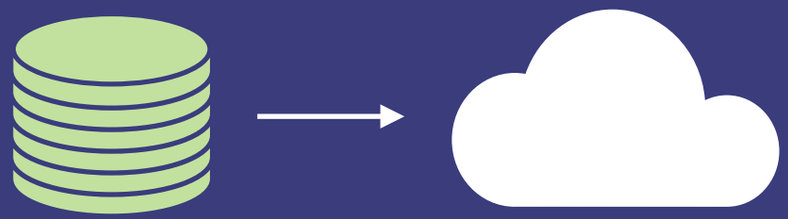
9494287 runs

Upload and explore all **results** online.

HACKATHON

Bring your own data, bring your own algorithms, or build cool new features.

It starts with data



Share data

```
import openml
my_data = openml.datasets.functions.create_dataset(
    data=df, name='', description='', licence='CC0')
my_data.publish()
```

```
Data published with ID 42521
https://www.openml.org/d/41521
```

Data can be pandas dataframe, numpy array, file, or URL

20+ pieces of meta-data by user, 100+ computed by server

auto-versioned, auto-analysed, organised

OpenML Search

19587 results

Only showing active datasets (public or shared with you).

FILTERS SORT: **MOST RUNS** ID'S TABLE + ADD NEW

credit-g (1)
This dataset classifies people described by a set of attributes as good or bad c...
★ **439801 runs** ♥ **2 likes** **39 downloads** **41 reach** **11 impact**
1000 instances - 21 features - 2 classes - 0 missing values

blood-transfusion-service-c...
Data taken from the Blood Transfusion Service Center in Hsin-Chu City in Taiw...
★ **424540 runs** ♥ **1 likes** **21 downloads** **22 reach** **13 impact**
748 instances - 5 features - 2 classes - 0 missing values

blood-transfusion-service-c... Data taken from the Blood Transfusion Service Center in Hsin-Chu City in Taiw...
★ 424540 runs ♥ 1 likes 21 downloads 22 reach 13 impact
748 instances - 5 features - 2 classes - 0 missing values

wilt (1) High-resolution Remote Sensing data set (Quickbird). Small number of training ...
★ 326436 runs ♥ 0 likes 32 downloads 32 reach 13 impact
4839 instances - 6 features - 2 classes - 0 missing values

steel-plates-fault (1) A dataset of steel plates' faults, classified into 7 different types. The goal was t...
★ 234209 runs ♥ 1 likes 23 downloads 24 reach 13 impact
1941 instances - 34 features - 2 classes - 0 missing values

qsar-biodeg (1) QSAR biodegradation Data Set * Abstract: Data set containing values for 41 att...
★ 215860 runs ♥ 1 likes 12 downloads 13 reach 13 impact
1055 instances - 42 features - 2 classes - 0 missing values

Australian (3) This dataset was retrieved 2014-11-14 from the libSVM site. It was normalized ...
★ 188737 runs ♥ 0 likes 11 downloads 11 reach 4 impact
690 instances - 15 features - 2 classes - 0 missing values

wdbc (1) Current dataset was adapted to ARFF format from the UCI version. Sample cod...
★ 186776 runs ♥ 1 likes 28 downloads 29 reach 13 impact
569 instances - 31 features - 2 classes - 0 missing values

kr-vs-kp (1) 1. Title: Chess End-Game -- King+Rook versus King+Pawn on a7 (usually abbre...
★ 180569 runs ♥ 0 likes 24 downloads 24 reach 12 impact
3196 instances - 37 features - 2 classes - 0 missing values

kc2 (1) One of the NASA Metrics Data Program defect data sets. Data from software f...

auto-versioned, auto-analysed, organised

OpenML Search

cardiotocography

V. 1

active ARFF Publicly available Visibility: public Uploaded 21-05-2015 by Rafael G. Mantovani Edit

0 likes downloaded by 21 people , 31 total downloads 0 issues 0 downvotes 21 reach 45 impact

OpenML100 × study_14 × study_34 × study_7 × + Add tag

Help us complete this description → Edit

Author: J. P. Marques de Sá, J. Bernardes, D. Ayers de Campos.
Source: UCI
Please cite: Ayres de Campos et al. (2000) SisPorto 2.0 A Program for Automated Analysis of Cardiotocograms. J Matern Fetal Med 5:311-318, UCI

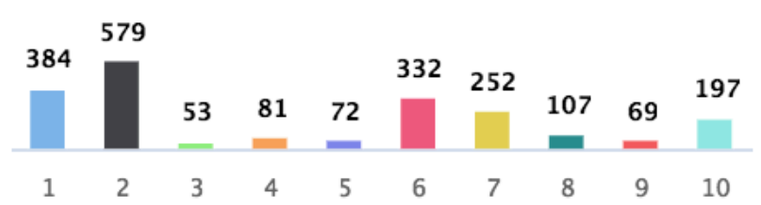
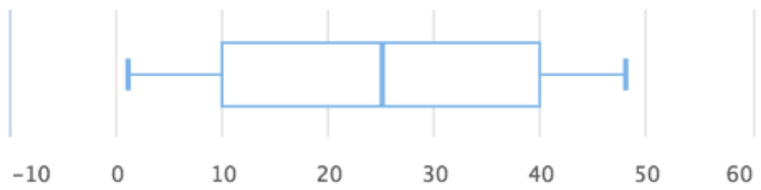
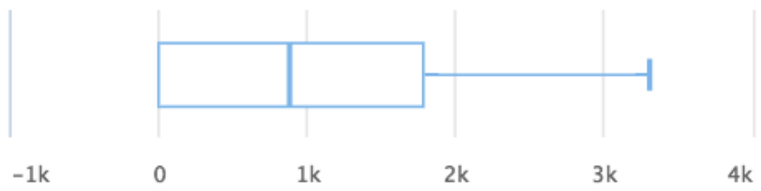
Help us complete this description → Edit

Author: J. P. Marques de Sá, J. Bernardes, D. Ayers de Campos.
Source: UCI
Please cite: Ayres de Campos et al. (2000) SisPorto 2.0 A Program for Automated Analysis of Cardiotocograms. J Matern Fetal Med 5:311-318, UCI

2126 fetal cardiotocograms (CTGs) were automatically processed and the respective diagnostic features measured. The CTGs were also classified by three expert obstetricians and a consensus classification label assigned to each of them. Classification was both with respect to a morphologic pattern (A, B, C. ...) and to a fetal state (N, S, P). Therefore the dataset can be used either for 10-class or 3-

▾ Show all

36 features

Class **(target)**	nominal	10 unique values 0 missing	
V1	numeric	48 unique values 0 missing	
V2	numeric	979 unique values 0 missing	

▾ Show all 36 features

Search data

```
openml.datasets.list_datasets(filters**)
```

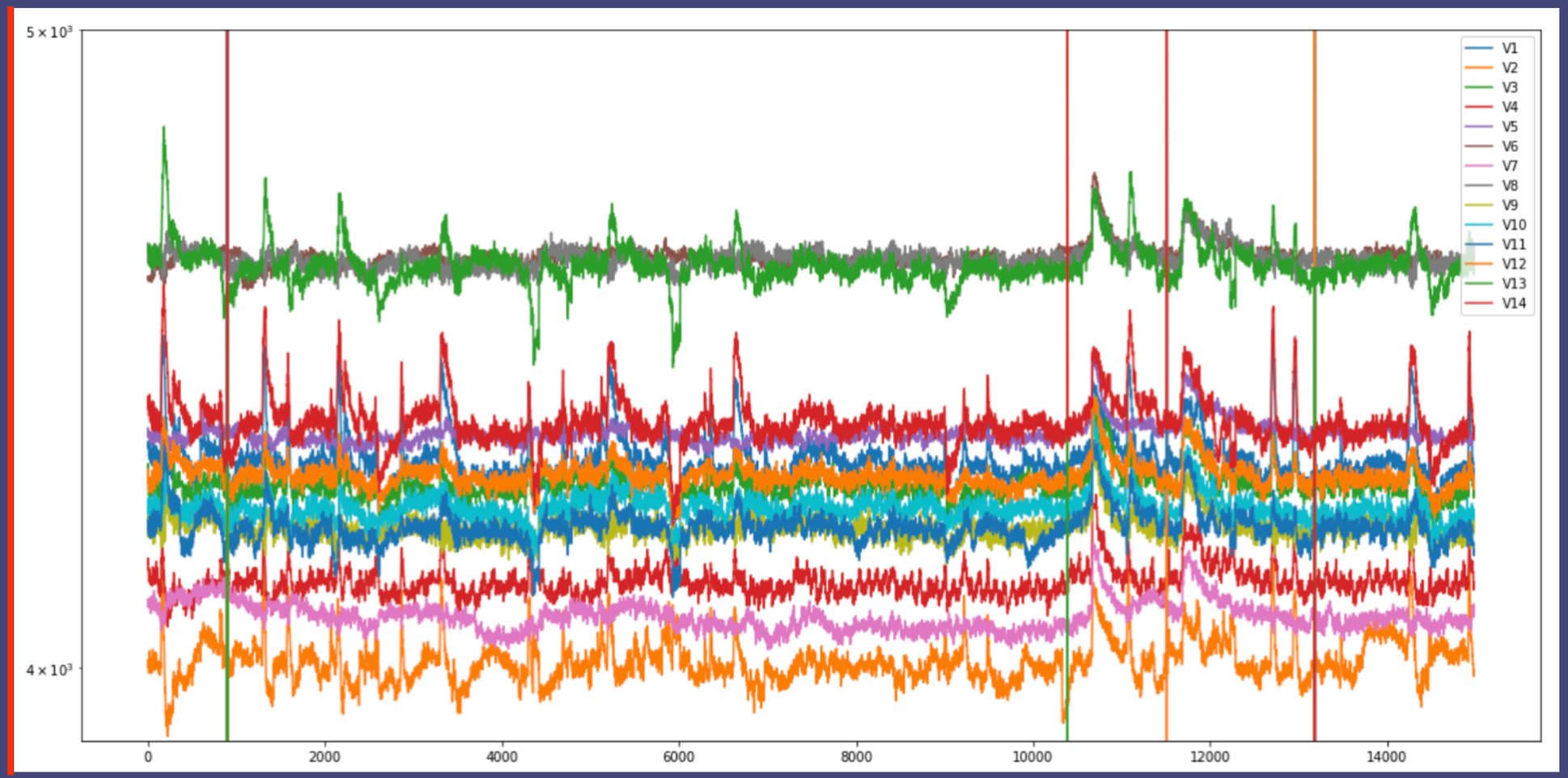
```
Found 2522 datasets
```

data id

	did	name	NumberOfInstances	NumberOfFeatures	NumberOfClasses
2	2	anneal	898	39	5
3	3	kr-vs-kp	3196	37	2
4	4	labor	57	17	2
5	5	arrhythmia	452	280	13
6	6	letter	20000	17	26
7	7	audiology	226	70	24

Download data

```
dataset = openml.datasets.get_dataset(1471)
df = dataset.get_data(options**)
df.plot()
```



Download + model data via sklearn

Works natively:

```
from sklearn.datasets import fetch_openml
X, y = fetch_openml(name='miceprotein',
                    return_X_y=True)
```

Fit an sklearn model:

```
clf = neighbors.KNeighborsClassifier()
clf.fit(X, y)
```

Tasks

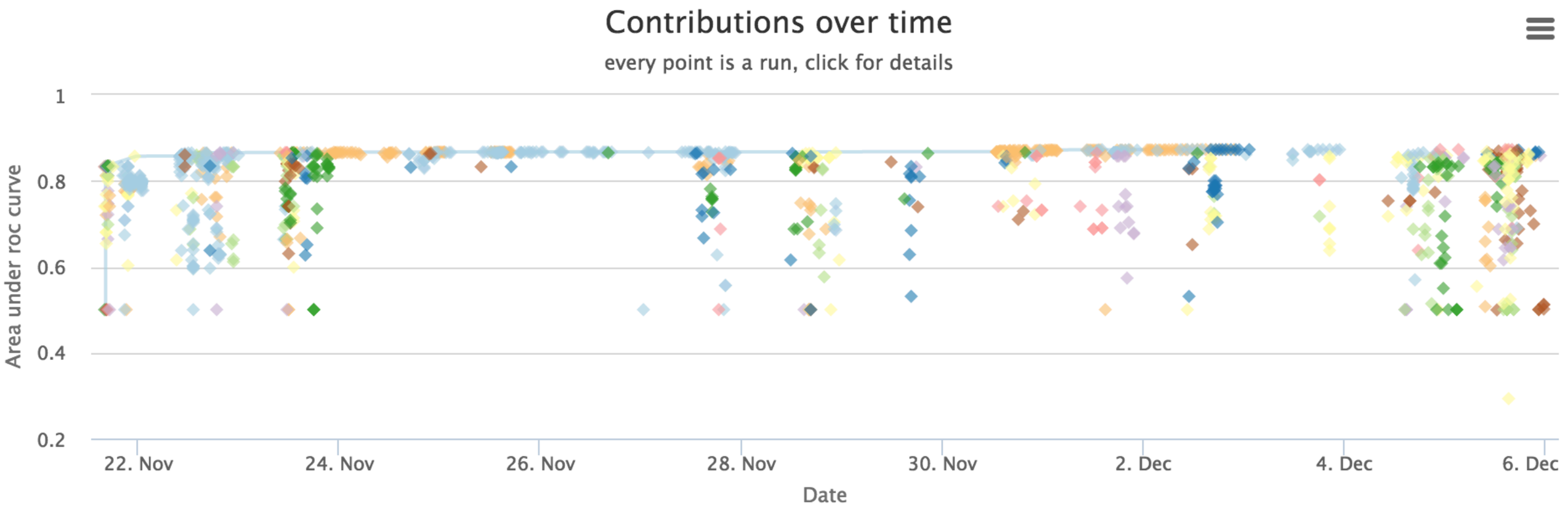
auto-benchmarking and collaboration

Tasks contain data, goals, procedures.

Auto-build + evaluate models correctly

All evaluations are directly comparable

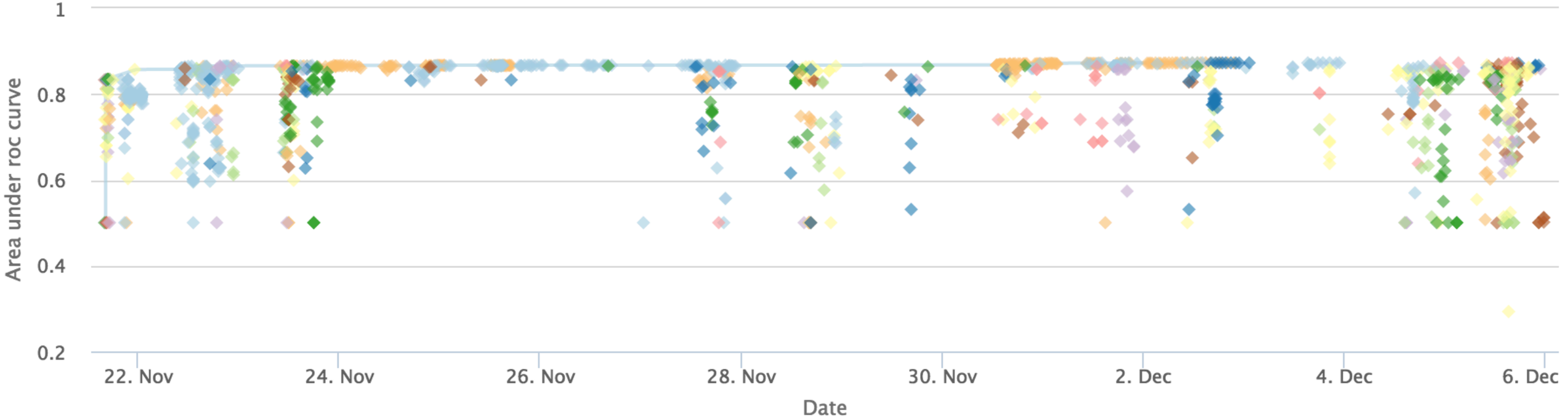
Collaborate in real time online



Predict target T
optimize accuracy
10-fold Xval

Contributions over time

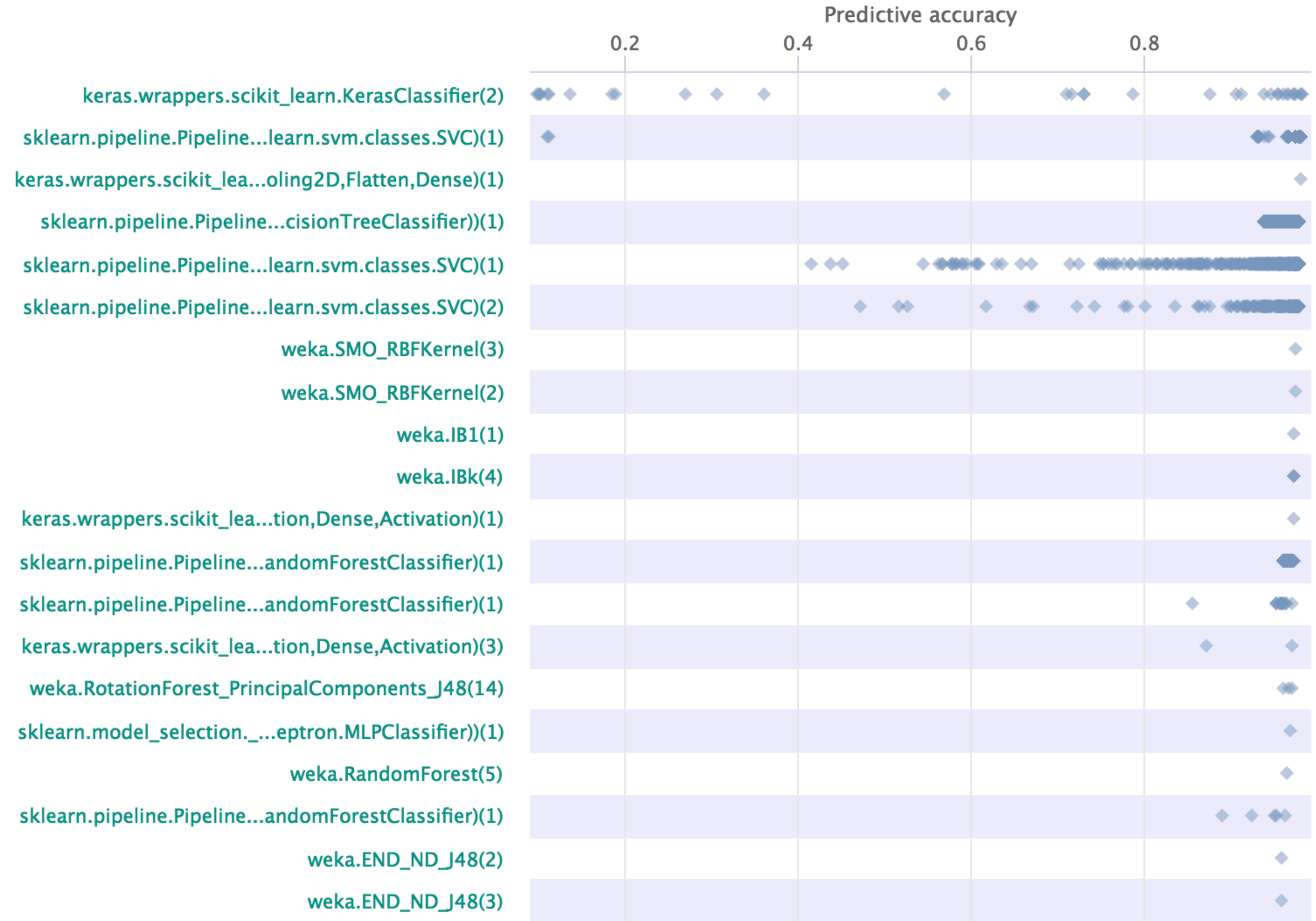
every point is a run, click for details



frontier Armand Duijn Daan de Graaf M. van Nijnatten Adriaan Knapen Evertjan Peer Maarten Visscher
a Hilde Weerts Jeroen Besems J Rensen T Peters Peter Gerards Marten Kenbeek Kagan Akbas
Khanh Nguyen Maarten Platenburg Laurence Keijzer Kuijpers Jasper Adegeest Yiming Lin Gijs Walravens
r f Kevin Chang Francisco Lozano George Park Gerson Foks Martijn Noordhof Jeroen Donners
Sven Arends van der Linde Chris Jansen Lujaina Abuerban Wouter Verlaek Michiel Verburg Luca Weibel
Adriaan Vast Luuk Godtschalk Anuvab Sahoo Bram Mulders Sajid Mohideen Pepijn Obels Ton Matton
B. Elfrink Marleen Hillen Julia Hofs Anne K Joaquin Vanschoren Sander doesnotmatter Casper Siksma
Johann Slabber Romek Vinke Jelte Dirks Aleksandr Popov Mert Zararsiz koen de Raad J KT
Daan Luttik Sjef van Loo Wijnands Leenen

Evaluations per flow (multiple parameter settings)

every point is a run, click for details



Search tasks

```
task_list = openml.tasks.list_tasks(size=5000)
```

First 5 of 5000 tasks:

	tid	did	name	task_type	estimation_procedure	evaluation_measures
2	2	2	anneal	Supervised Classification	10-fold Crossvalidation	predictive_accuracy
3	3	3	kr-vs-kp	Supervised Classification	10-fold Crossvalidation	predictive_accuracy
4	4	4	labor	Supervised Classification	10-fold Crossvalidation	predictive_accuracy
5	5	5	arrhythmia	Supervised Classification	10-fold Crossvalidation	predictive_accuracy
6	6	6	letter	Supervised Classification	10-fold Crossvalidation	predictive_accuracy

task id

Flows

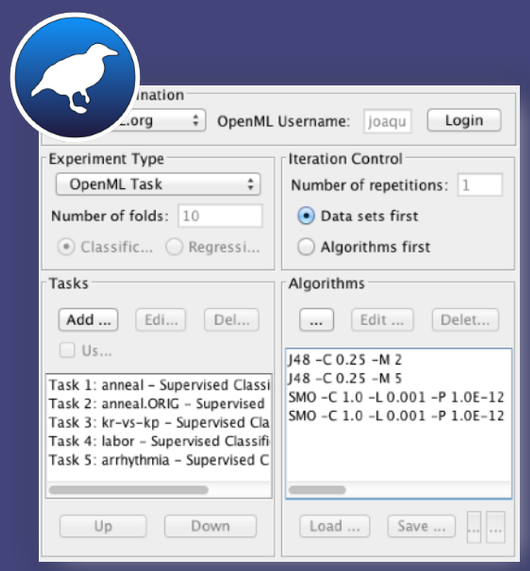
Run experiments ***locally***, share them ***globally***

Auto-run algorithms/workflows on any task

Integrated in many machine learning tools (+ APIs)

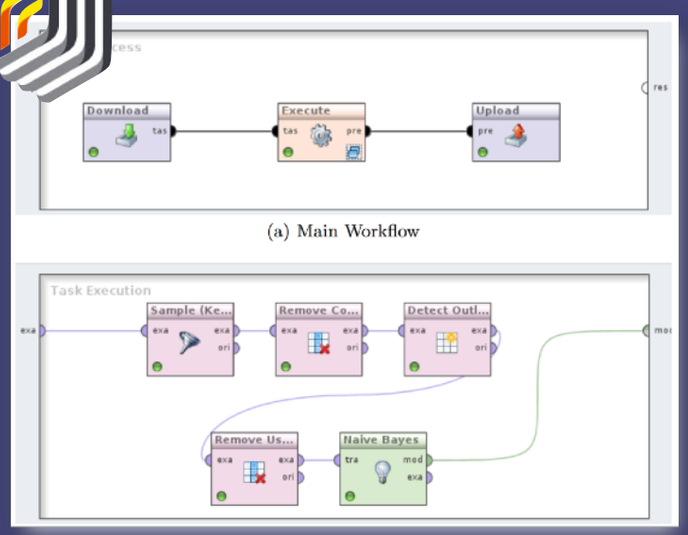
R mlR scikit learn H_2O.ai dmlc XGBoost

Integrated in many machine learning tools (+ APIs)




```python
import openml as oml
from sklearn import tree

task = oml.tasks.get_task(14951)
clf = tree.ExtraTreeClassifier()
flow = oml.flows.sklearn_to_flow(clf)
run = oml.runs.run_flow_on_task(task, flow)
myrun = run.publish()
```



(a) Main Workflow

```r
library(OpenML)
library(mlr)
task = getOMLTask(10)
lrn = makeLearner("classif.rpart")
res = runTaskMlr(task, lrn)
run.id = uploadOMLRun(res)
```

Fit and share (sklearn example)

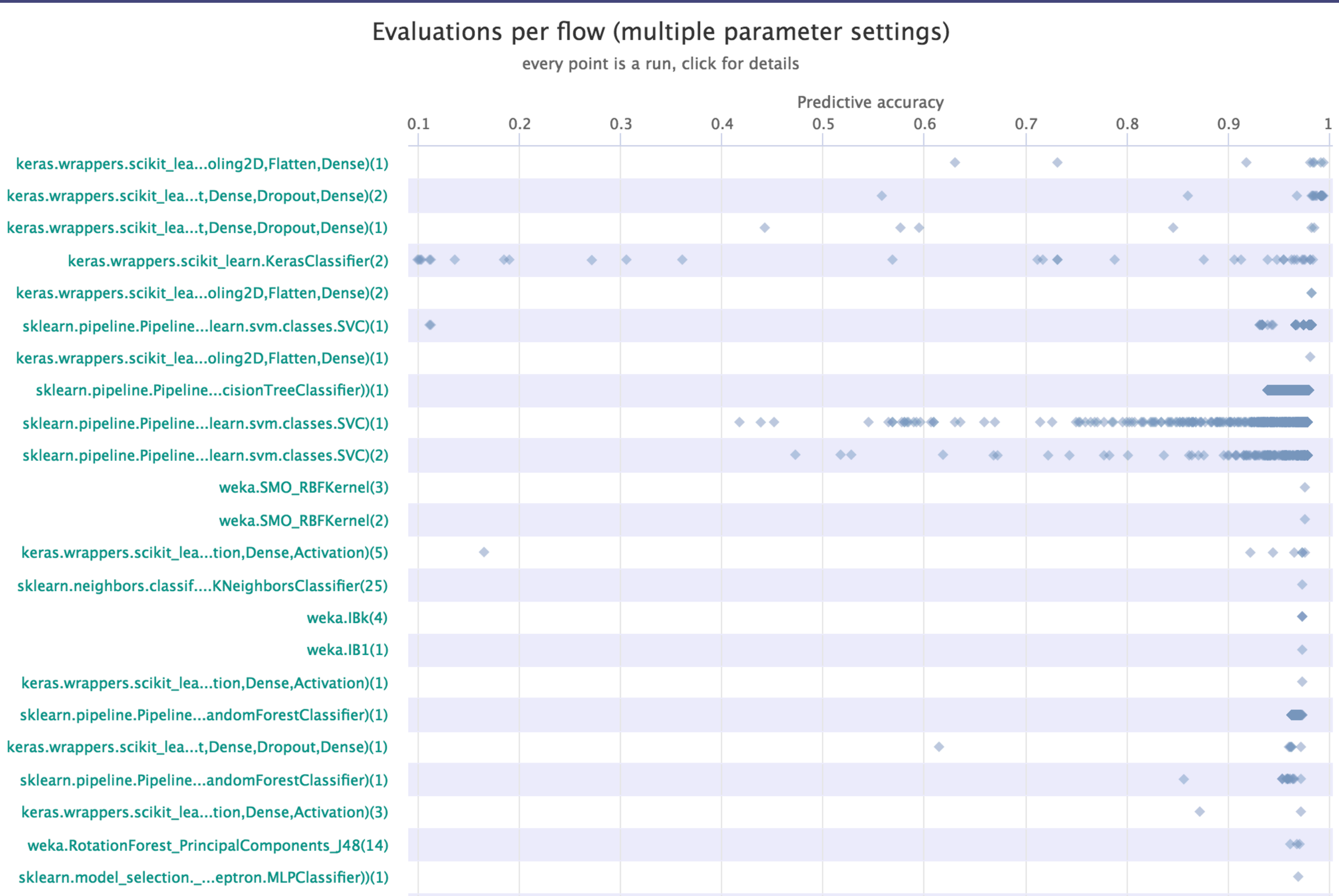
```
task = openml.tasks.get_task(14951)
clf = sklearn.tree.ExtraTreeClassifier()
flow = openml.flows.sklearn_to_flow(clf)
run = openml.runs.run_flow_on_task(flow,task)
myrun = run.publish()
```

```
Run published with ID 9204488
https://www.openml.org/r/9204488
```

Any sklearn pipeline can be run on any OpenML task

Similar capabilities for other Python, R, Java libraries

Compare to state-of-the-art



Runs

Share and reuse results

Experiments auto-uploaded, evaluated online

reproducible, linked to **data, flows, authors**

and **all other experiments**

Experiments auto-uploaded, evaluated online

Result files

Description

XML file describing the run, including user-defined evaluation measures.

Model readable

A human-readable description of the model that was built.

Model serialized

A serialized description of the model that can be read by the tool that generated it.

Predictions

ARFF file with instance-level predictions generated by the model.

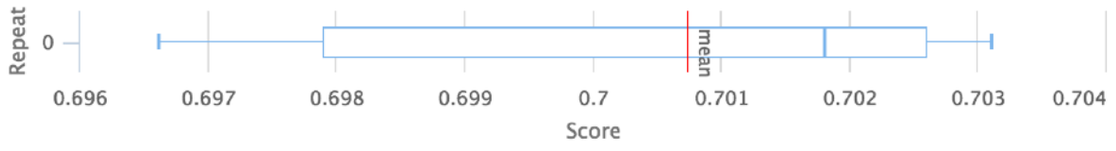
Area under ROC curve

0.7007 ± 0.0023

Per class

0	1
0.7007	0.7007

Cross-validation details (10-fold Crossvalidation)



Download, reuse runs

```
evals = openml.evaluations.list_evaluations(
    task=[145677], function='area_under_roc_curve')
scores =[{"flow":e.flow_name, "score":e.value} for id
sns.violinplot(x="score", y="flow", data= pd.DataFrame
```

Publishing, impact tracking

Heidi Seibold

PhD student in Computational Biostatistics at the University of Zurich. I am into R, open science and reproducible research.

University of Zurich Joined 2016-01-27

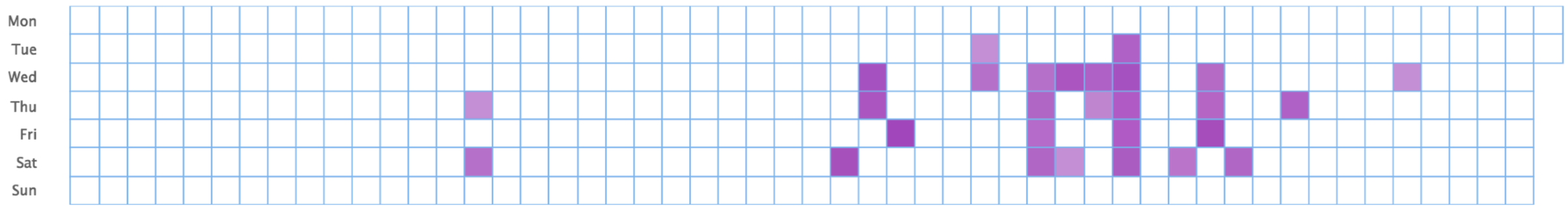
Activity 1649.5 | Reach 12 | Impact 195 | Uploads 1 | 35 | 0 | 1605

EDIT PROFILE

	Activity	Reach		Impact
Data Sets	1	4	0	0
Flows	35	7	0	195
Tasks	0	0	0	
Runs	1605	1	0	

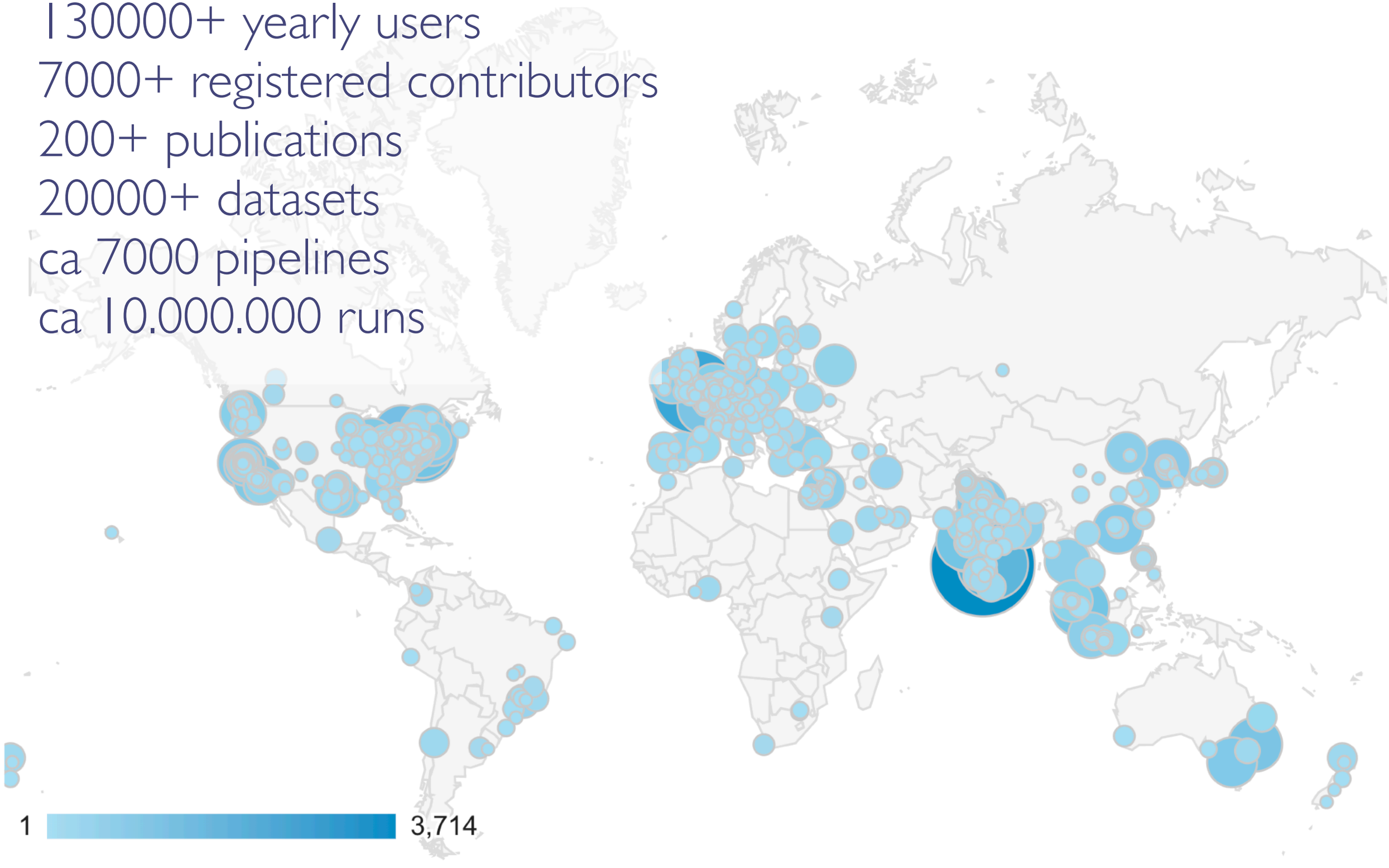
Activity: 1.65K | 1.64K | 0 | 17

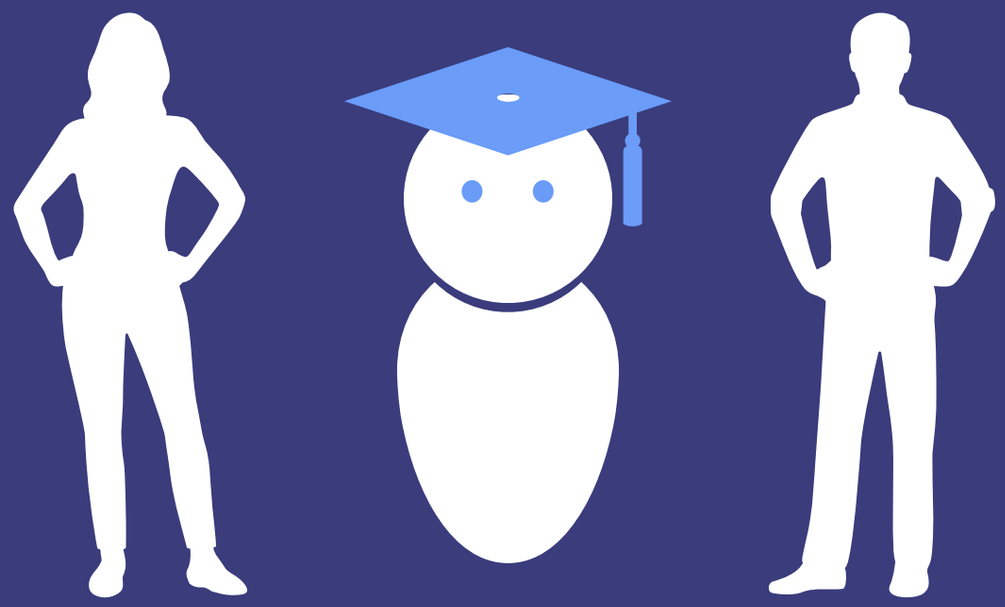
Activity from Sunday 2016-05-29 to Monday 2017-05-29



OpenML Community

130000+ yearly users
7000+ registered contributors
200+ publications
20000+ datasets
ca 7000 pipelines
ca 10.000.000 runs





Collaborations and unexpected outcomes

Using data correctly

ARFF CSV JSON XML RDF

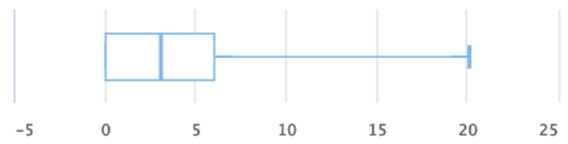
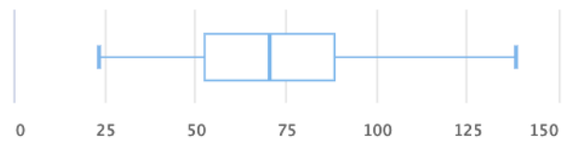
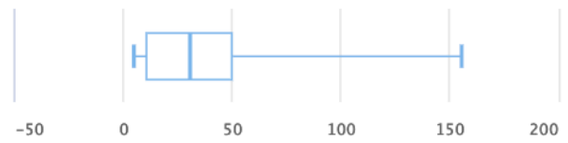
liver-disorders

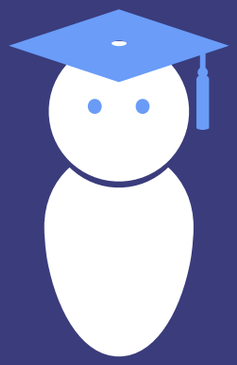
active ARFF Publicly available Visibility: public Uploaded 06-04-2014 by Jan van Rijn

2 likes downloaded by 29 people , 43 total downloads 0 issues 0 downvotes

study_127 study_50 study_88 uci + Add tag

7 features

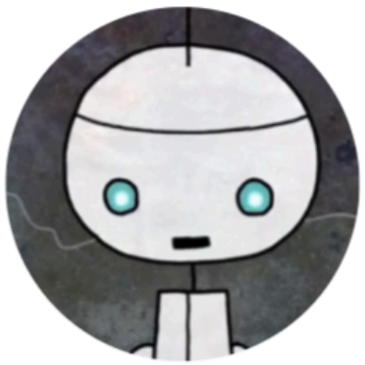
drinks **(target)**	numeric	16 unique values 0 missing	
selector **(ignore)**	nominal	2 unique values 0 missing	
alkphos	numeric	78 unique values 0 missing	
sgpt	numeric	67 unique values 0 missing	
sgot	numeric	47 unique values 0 missing	



Learning to learn

Algorithms (bots) learn from models shared by humans

Humans learn from models built by humans and bots



OpenML_Bot R

Can I help? You can find me here: https://github.com/ja-thomas/OMLbots

Joined 2017-03-07

Uploads

0 40 0 5692477

Data Sets	0	0	0	0
Flows	40	33	1	2
Tasks	0	0	0	
Runs	5692477	8	1	

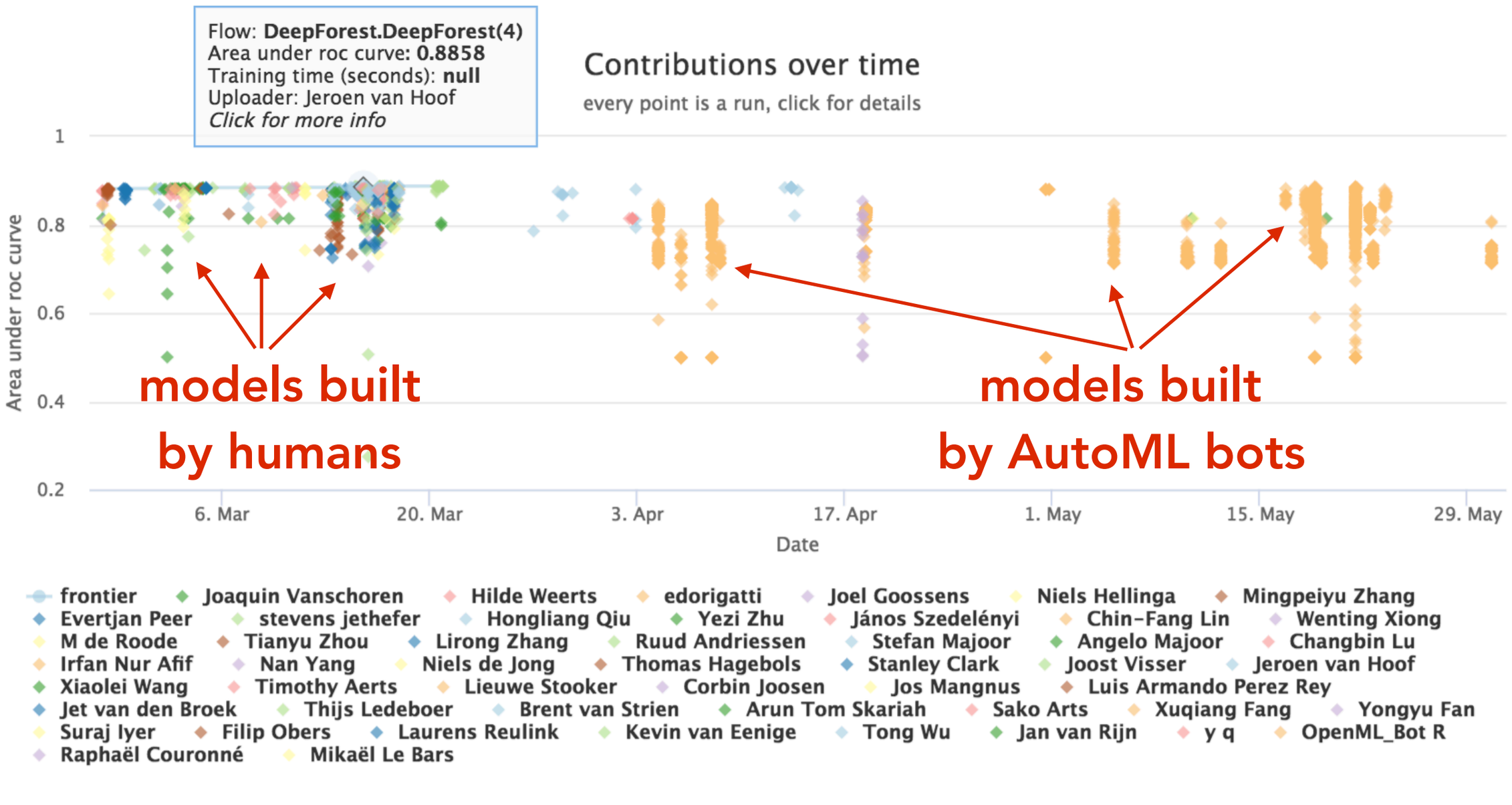
Learning to learn

Algorithms (bots) learn from models shared by humans

Humans learn from models built by humans and bots

Timeline

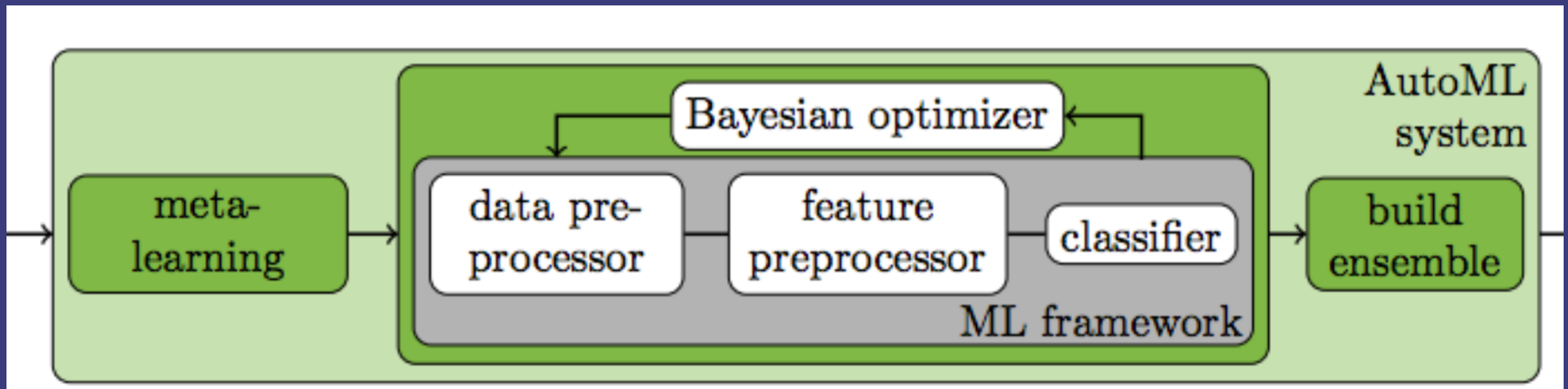
Metric: AREA UNDER ROC CURVE

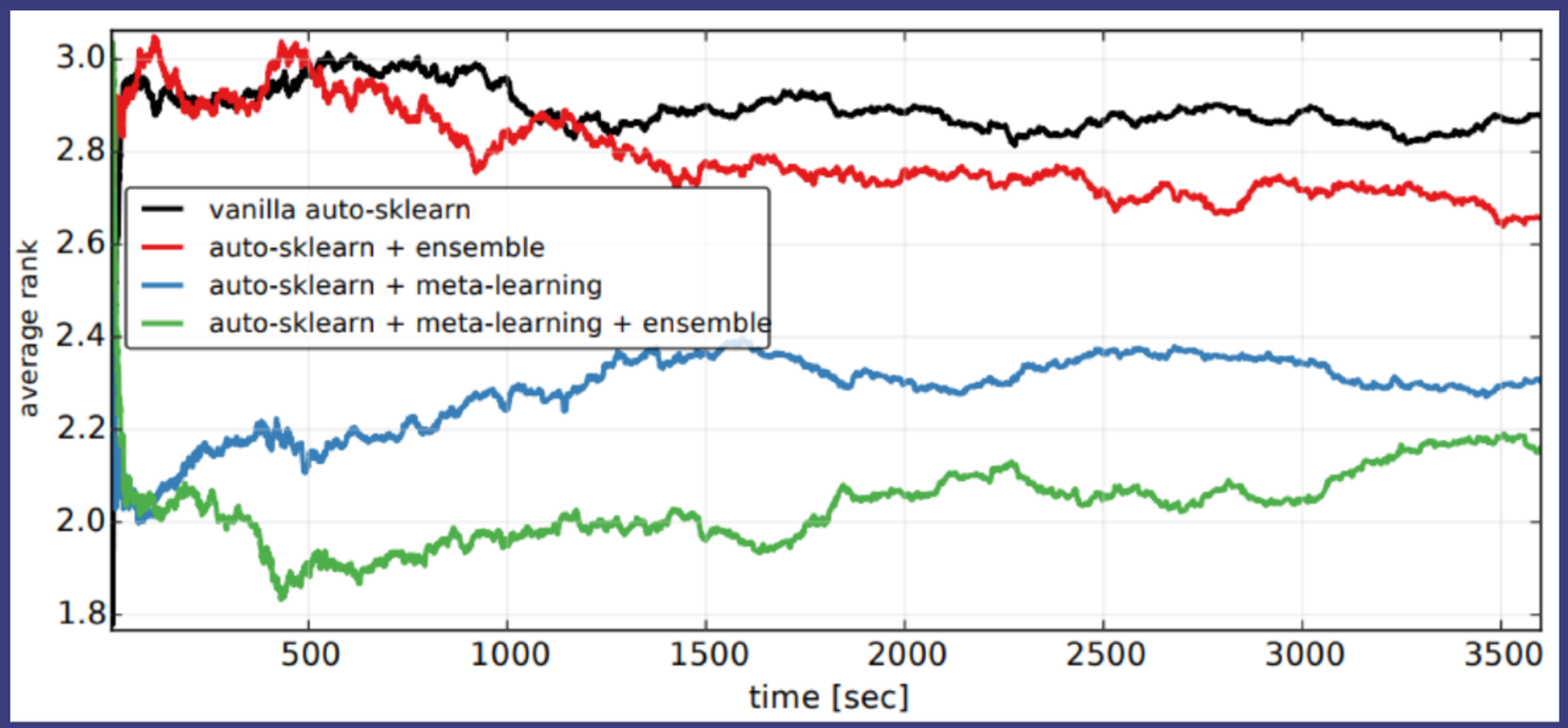


frontier Joaquin Vanschoren Hilde Weerts edorigatti Joel Goossens Niels Hellinga Mingpeiyu Zhang
Evertjan Peer stevens jethefer Hongliang Qiu Yezi Zhu János Szedelényi Chin-Fang Lin Wenting Xiong
M de Roode Tianyu Zhou Lirong Zhang Ruud Andriessen Stefan Majoor Angelo Majoor Changbin Lu
Irfan Nur Afif Nan Yang Niels de Jong Thomas Hagebols Stanley Clark Joost Visser Jeroen van Hoof
Xiaolei Wang Timothy Aerts Lieuwe Stooker Corbin Joosen Jos Mangnus Luis Armando Perez Rey
Jet van den Broek Thijs Ledeboer Brent van Strien Arun Tom Skariah Sako Arts Xuqiang Fang Yongyu Fan
Suraj Iyer Filip Obers Laurens Reulink Kevin van Eenige Tong Wu Jan van Rijn y q OpenML_Bot R
Raphaël Couronné Mikaël Le Bars

Meta-learning for automated machine learning

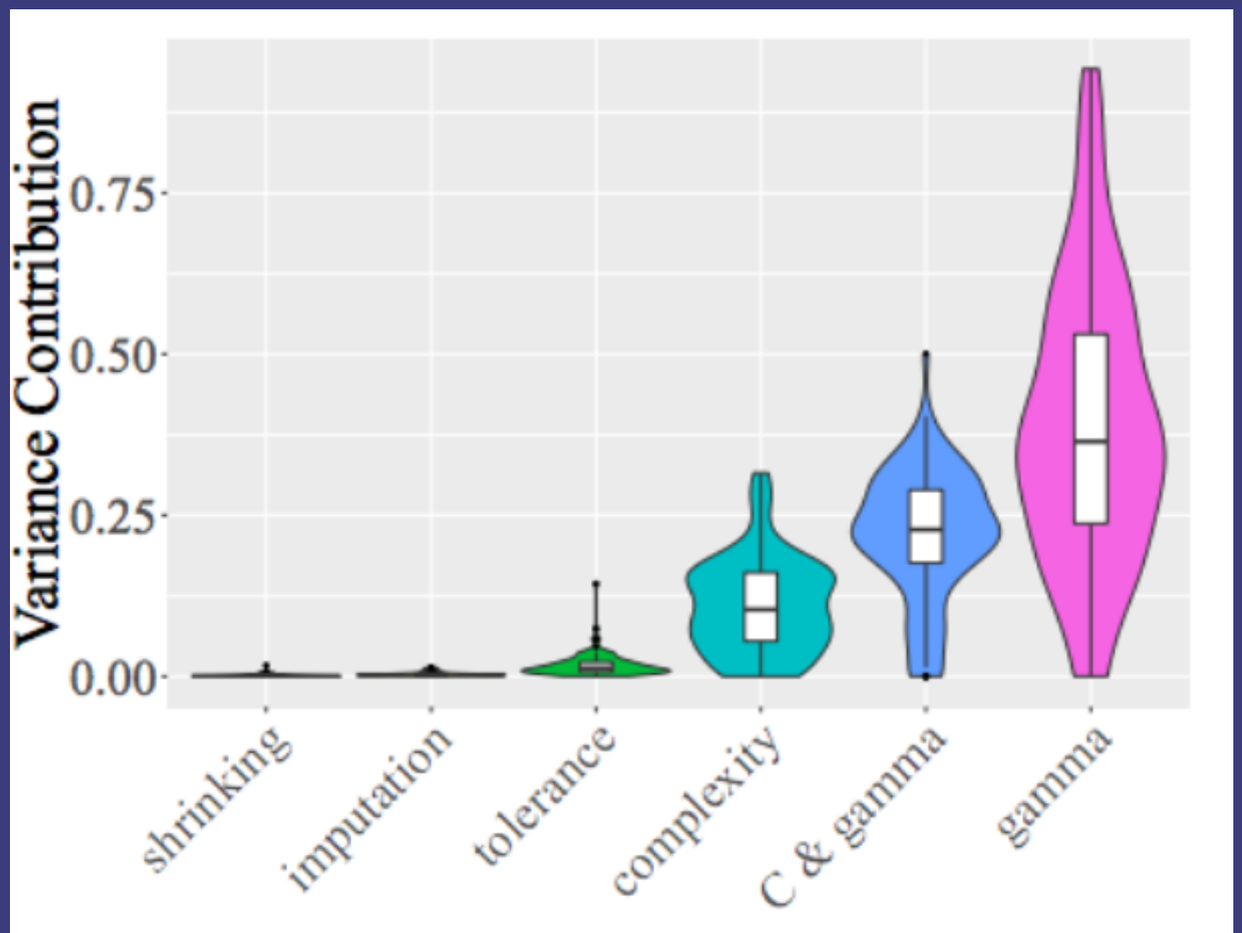
- **Auto-sklearn** (AutoML challenge winner, 2016-2017)
  - Lookup similar datasets on OpenML, start with best pipelines

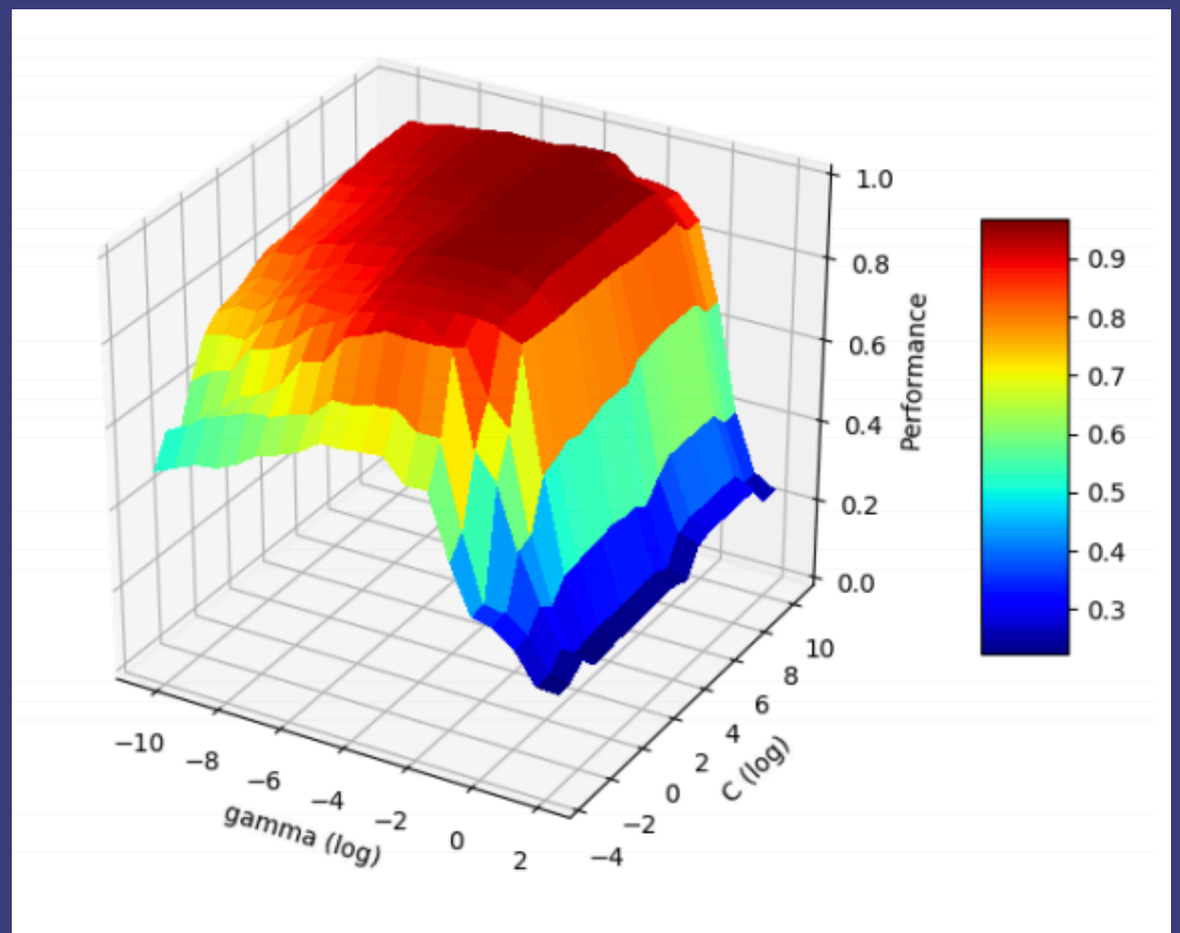




Making machine learning easier/faster

- Leaning better (data-specific) hyper parameter defaults
- Leaning hyperparameter importance for faster tuning





Benchmarking suites

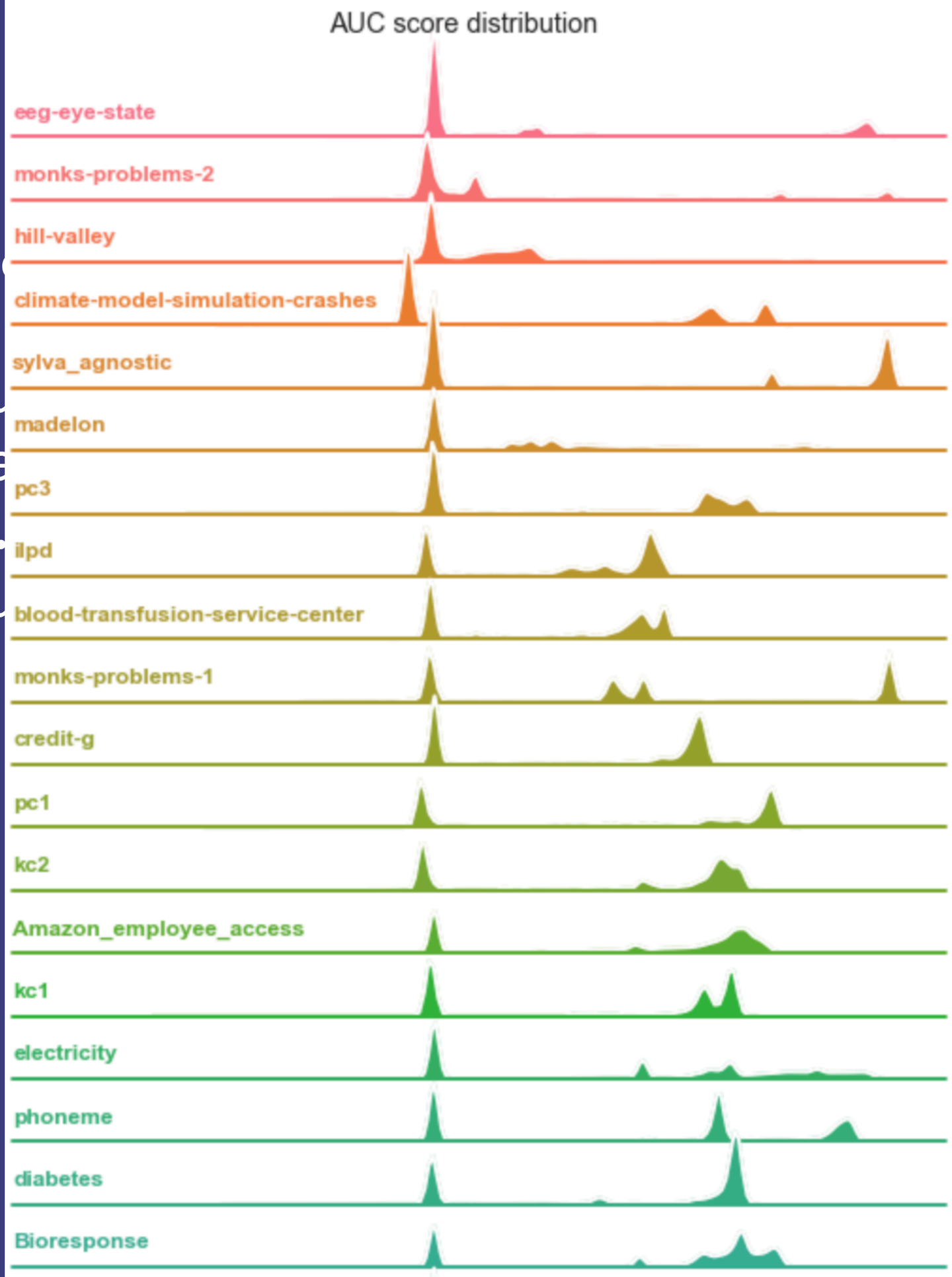
- Comprehensive, curated, easy-to-use benchmarking

```
benchmark = openml.study.get_study('CC18','tasks')
for task_id in benchmark.tasks:
   task = openml.tasks.get_task(task_id)
   run = openml.runs.run_model_on_task(clf, task)
   run.publish()
```

Benchmarking suites

- Comprehensive, curated,

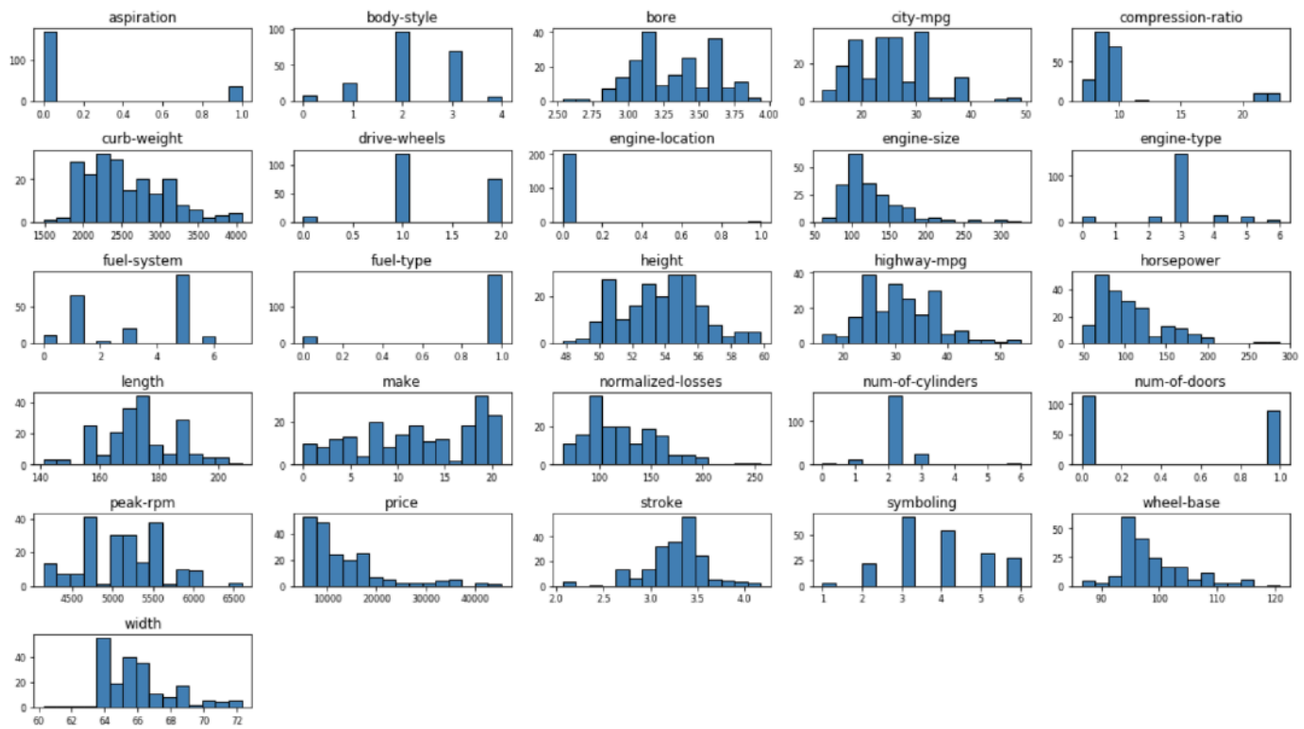
```
benchmark = openml.stu
for task_id in benchma
    task = openml.tasks.
    run = openml.runs.ru
    run.publish()
```



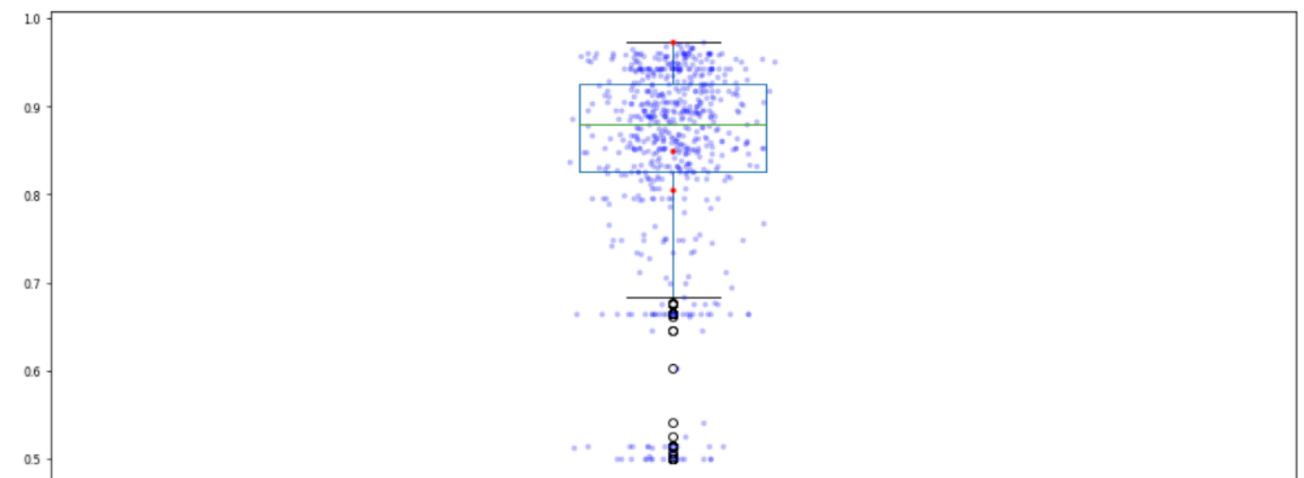
Auto-notebooks

- Basic insight into any OpenML dataset

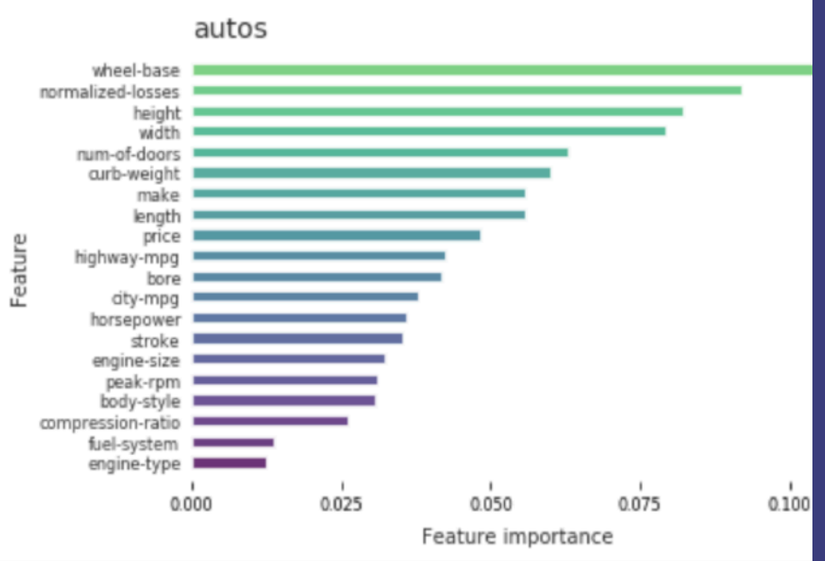
```
In [3]: from scripts.dataVisualization import *
        show1DHist(data)
```




```
In [12]: from scripts.relativePerformance import *
         showRelativePerformanceBoxplot(scores, topList, settings.strats, maxBaseline)
```



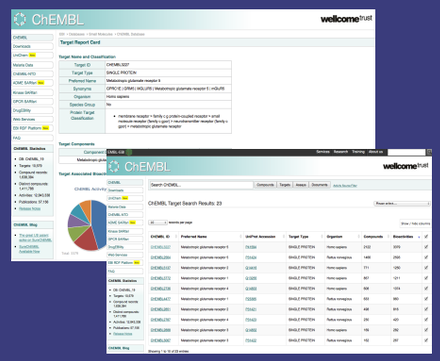
```
In [7]: from scripts.featureImportance import *
        featureImportance(data)
```



	0	1	2	3	4	5	6	7	8	9	anomaly_score
68	1.53125	10.73	0	2.1	69.81	0.58	13.3	3.15	0.28	1	-0.182281
46	1.51514	14.01	2.68	3.5	69.89	1.68	5.87	2.2	0	4	-0.135589
4	1.53393	12.3	0	1	70.16	0.12	16.19	0	0.24	1	-0.123574
105	1.51316	13.02	0	3.04	70.48	6.21	6.96	0	0	4	-0.119413
163	1.51321	13	0	3.02	70.7	6.21	6.93	0	0	4	-0.117612
32	1.51115	17.38	0	0.34	75.41	0	6.65	0	0	5	-0.0969372
24	1.52058	12.85	1.61	2.17	72.18	0.76	9.7	0.24	0.51	4	-0.0899955
102	1.52365	15.79	1.83	1.31	70.43	0.31	8.61	1.68	0	6	-0.0823851
173	1.51831	14.39	0	1.82	72.86	1.41	6.47	2.88	0	6	-0.0820739
150	1.51131	13.69	3.2	1.81	72.81	1.76	5.43	1.19	0	6	-0.0762463
21	1.52475	11.45	0	1.88	72.19	0.81	13.24	0	0.34	1	-0.07546
117	1.51838	14.32	3.26	2.22	71.25	1.46	5.79	1.63	0	6	-0.0734898
164	1.52739	11.02	0	0.75	73.08	0	14.96	0	0	1	-0.0397467
126	1.51653	11.95	0	1.19	75.18	2.7	8.93	0	0	6	-0.0365798
86	1.52777	12.64	0	0.67	72.02	0.06	14.4	0	0	1	-0.0336708
160	1.52664	11.23	0	0.77	73.21	0	14.68	0	0	1	-0.0302277
56	1.52247	14.86	2.2	2.06	70.26	0.76	9.76	0	0	6	-0.0189952

Drug discovery

- Build better models that recommend drug candidates for rare diseases



ChEMBL DB: 1.4M compounds,
10k proteins,12.8M activities

Molecule representations

MW	LogP	TPSA	b1	b2	b3	b4	b5	b6	b7	b8	b9	
377.435	3.883	77.85	1	1	0	0	0	0	0	0	0	
341.361	3.411	74.73	1	1	0	1	0	0	0	0	0	...
197.188	-2.089	103.78	1	1	0	1	0	0	0	1	0	
346.813	4.705	50.70	1	0	0	1	0	0	0	0	0	
							.					
							:					

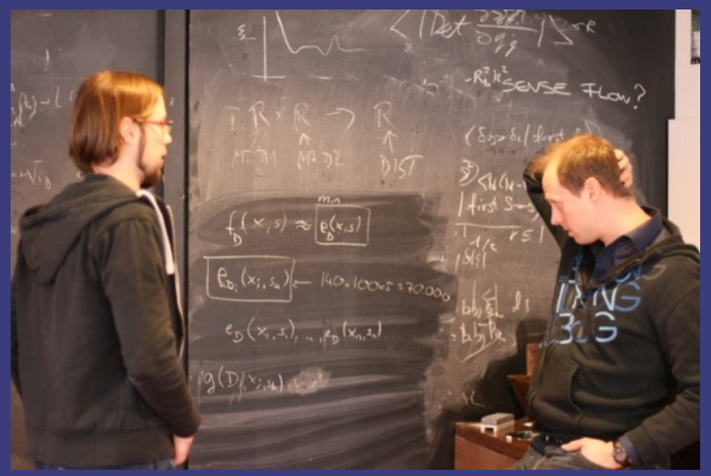
16,000 regression datasets
x52 pipelines (on OpenML)

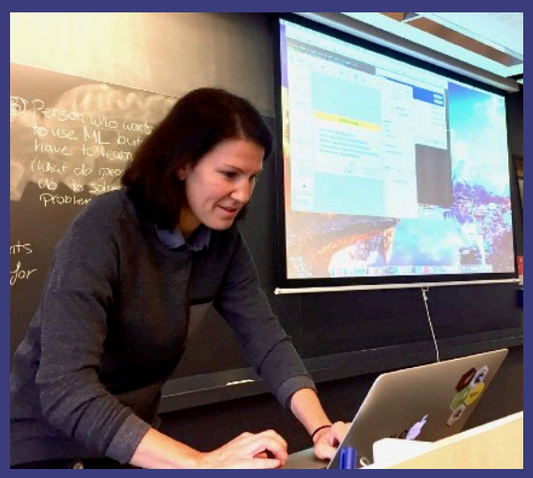
all data on new protein → meta-model → **optimal models to predict activity**

(Olier et al., Machine Learning 107(1), 2018)

Building an open source team

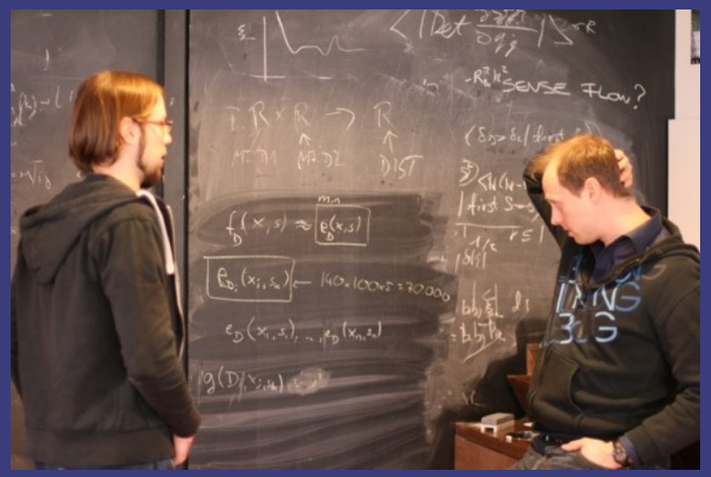
20+ active developers
Hackathons + GitHub
Open to existing projects
Academia + industry

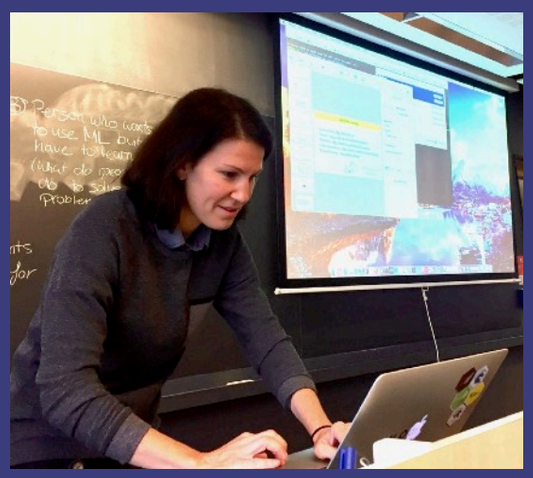




Join us! (and change the world)

Active open source community
Hackathons 2-3x a year
We need bright people
- ML, Devs, UX
OpenML Foundation





Thanks to the entire OpenML team

Jan
van Rijn

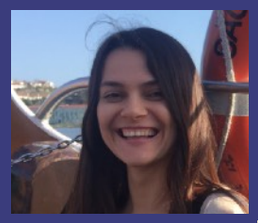
Guiseppe
Casalicchio

Mattias
Feurer

Bernd
Bischl

Andreas
Mueller

Heidi
Seibold



Bilge
Celik

Pieter
Gijsbers

Andrey
Ustyuzhanin

William
Raynaut

Erin
LeDell

Tobias
Glasmachers

Jakub
Smid

And many more!

Thank you!

谢谢

@open_ml

OpenML

www.openml.org