@open_ml

OpenML

Be a part of this talk:

- Log in / create an account on www.openml.org
  - *You also need a GitHub account*
- Click the R or Python icon
- Click ‘Launch Demo’

JOAQUIN VANSCHOREN, TU EINDHOVEN, 2016

OpenML

DEMOCRATIZING AND AUTOMATING MACHINE LEARNING

JOAQUIN VANSCHOREN, TU EINDHOVEN, 2017

@joavanschoren

Research different.

We want to empower everybody to do great machine learning

Find interesting datasets and use them immediately.
Or share your own.

View problems that people are working on.
Or crowdsource your own problems.

Use open-source tools to try many algorithms.
Automate drudge work (with smart robots)

Reproducible, transparent, reusable results
Organized for easy analysis and reuse

(Increasingly) frictionless online collaboration
Easy sharing of data and results

WHAT IF WE CAN ANALYSE DATA

COLLABORATIVELY

WHAT IF WE CAN ANALYSE DATA

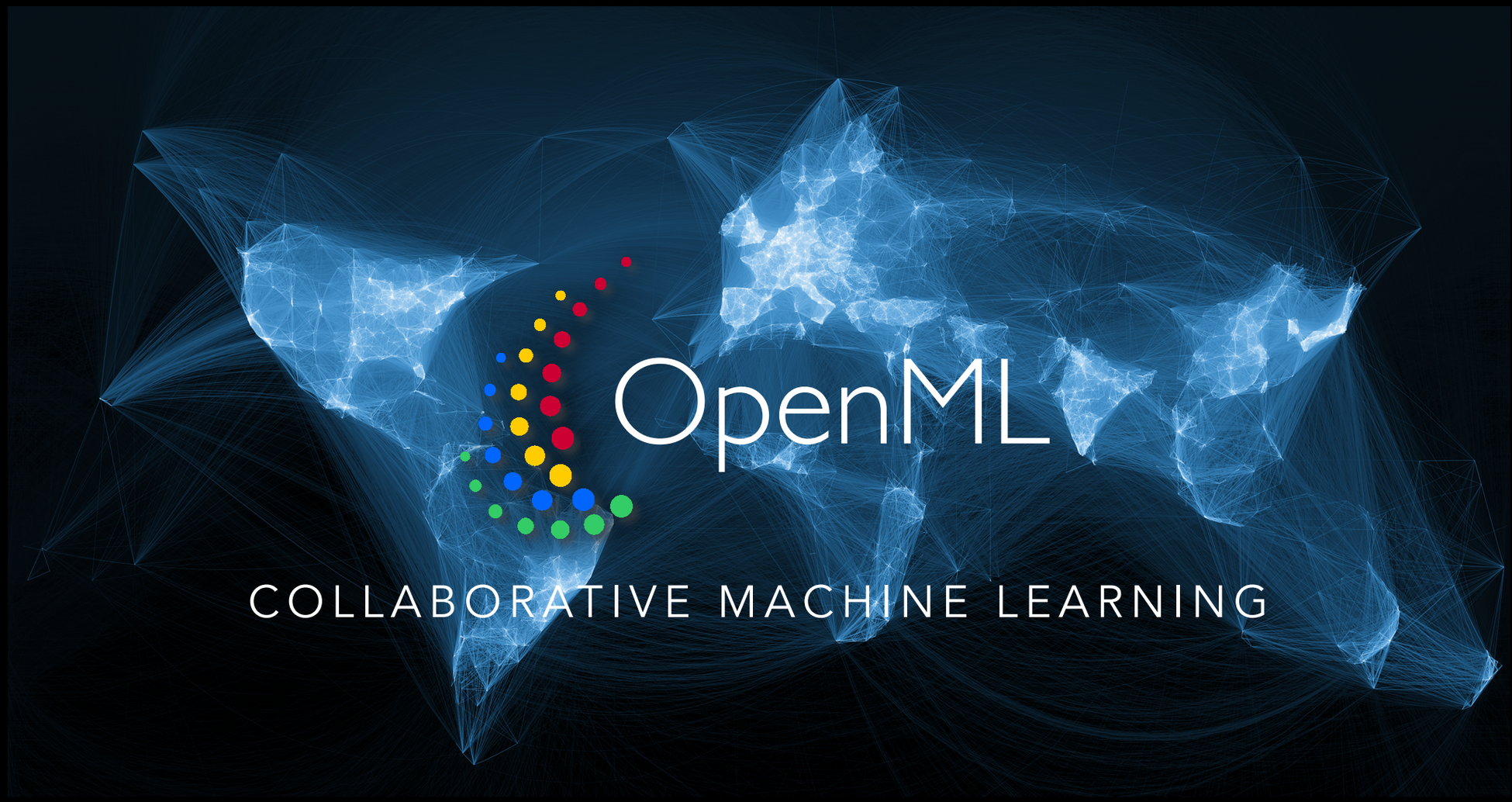
COLLABORATIVELY

ON WEB SCALE

WHAT IF WE CAN ANALYSE DATA

COLLABORATIVELY

ON WEB SCALE IN REAL TIME



Easy to use: Integrated in many ML tools/environments

Easy to contribute: Automated sharing of data, code, results

Organized data: Meta-data, reproducible models, link to people

Reward structure: Track your impact, build reputation

Self-learning: Learn from many experiments to help people

It starts with data

UCI Machine Learning Repository

mldata.org machine learning data set repository

figshare

zenodo

kaggle

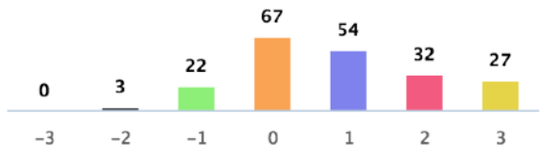
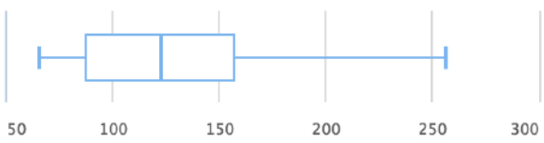
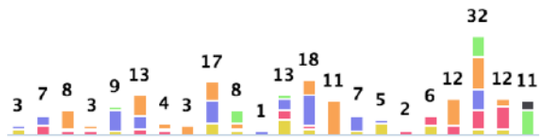
It starts with data

Data (ARFF) uploaded or referenced (URL)

auto-versioned, analysed, meta-data extracted, organised online

auto-versioned, analysed, meta-data
extracted, organised online

26 features

symboling **(target)**	nominal	6 unique values 0 missing	
normalized-losses	numeric	51 unique values 41 missing	
make	nominal	22 unique values 0 missing	

Show all 26 features

72 properties

DefaultAccuracy	**0.33**	The predictive a
NumberOfClasses	**7**	The number of c
NumberOfFeatures	**26**	The number of f
NumberOfInstances	**205**	The number of i
NumberOfMissingVal...	**59**	Counts the total

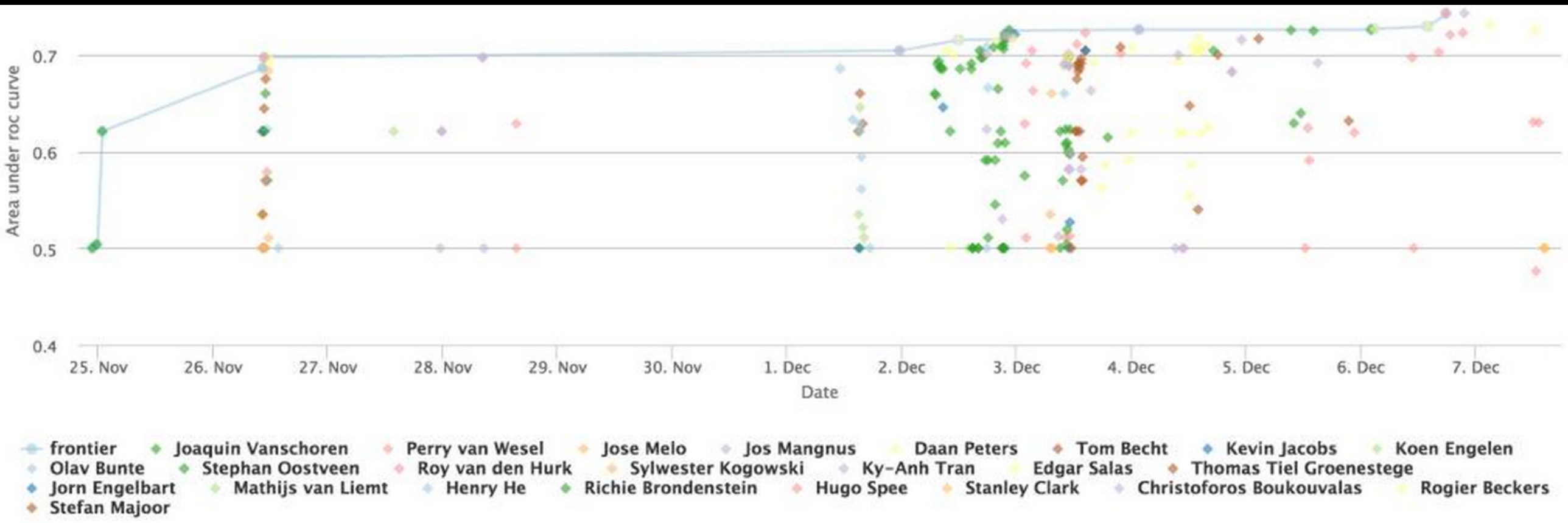
Set your goals, find help

Tasks contain data, goals, procedures.

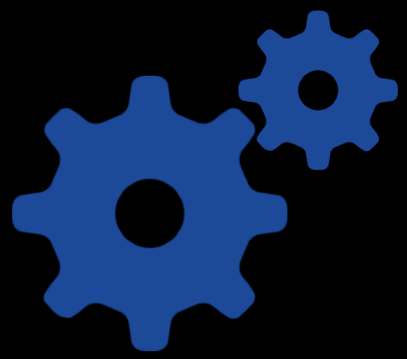
Readable by tools, automates experimentation

All results organized online: **realtime overview**

All results organized online: **realtime overview**

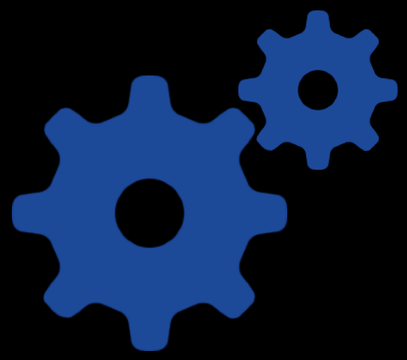


Explore all possible solutions

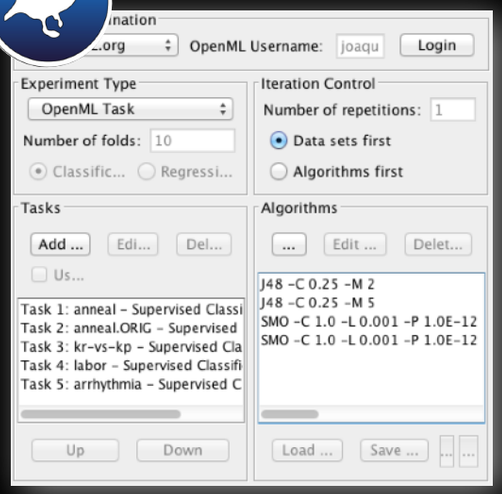


Flows (workflows, scripts) can run anywhere (locally)

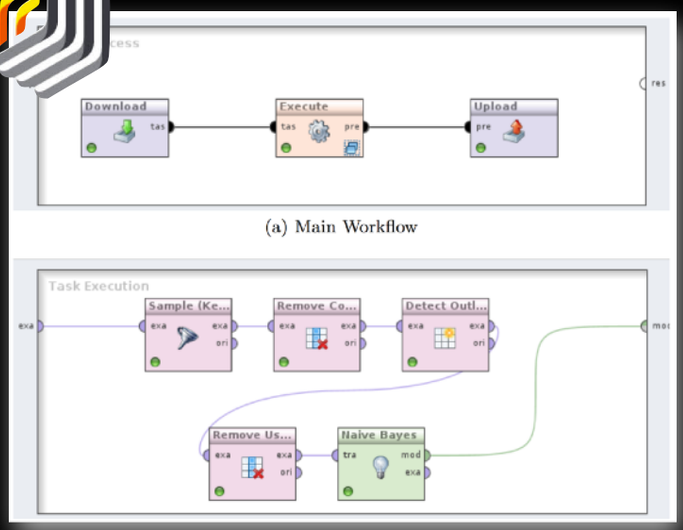
Tool integrations + APIs (REST, R, Python, Java,…)



Integrations + APIs (REST, R, Python, Java,...)

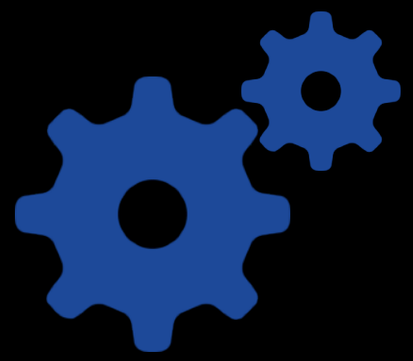


```python
from sklearn import tree
from openml import tasks,runs
task = tasks.get_task(14951)
clf = tree.DecisionTreeClassifier()
run = runs.run_task(task, clf)
return_code, response = run.publish()
```



(a) Main Workflow

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



Integrations + APIs (REST, R, Python, Java,…)

Run locally (or wherever you want)

```
from sklearn import tree
from openml import tasks,runs
task = tasks.get_task(14951)
clf = tree.DecisionTreeClassifier()
run = runs.run_task(task, clf)
return_code, response = run.publish()
```

Analyse results objectively

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.

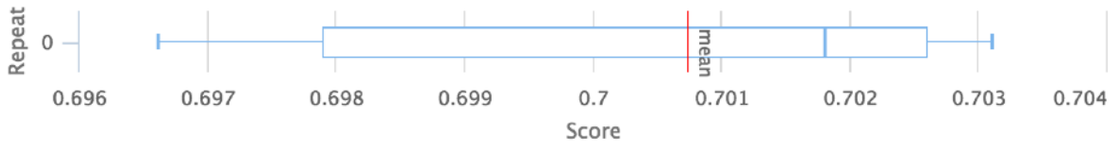
Area under ROC curve

0.7007 ± 0.0023

Per class

0	1
0.7007	0.7007

Cross-validation details (10-fold Crossvalidation)



Publish, and track your impact

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

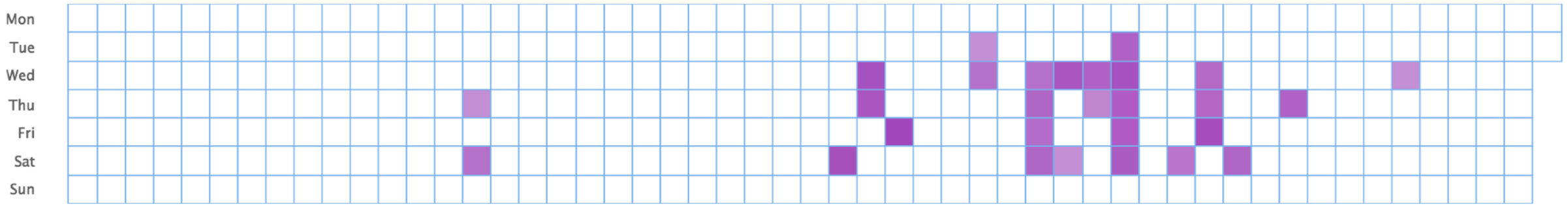
Activity	Reach	Impact	Uploads			
1649.5	12	195	1	35	0	1605

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



Demo

OpenML beta

Exploring machine learning better, together

19626 data sets

Find or add **data** to analyse

25941 tasks

Download or create scientific **tasks**

3353 flows

Find or add data analysis **flows**

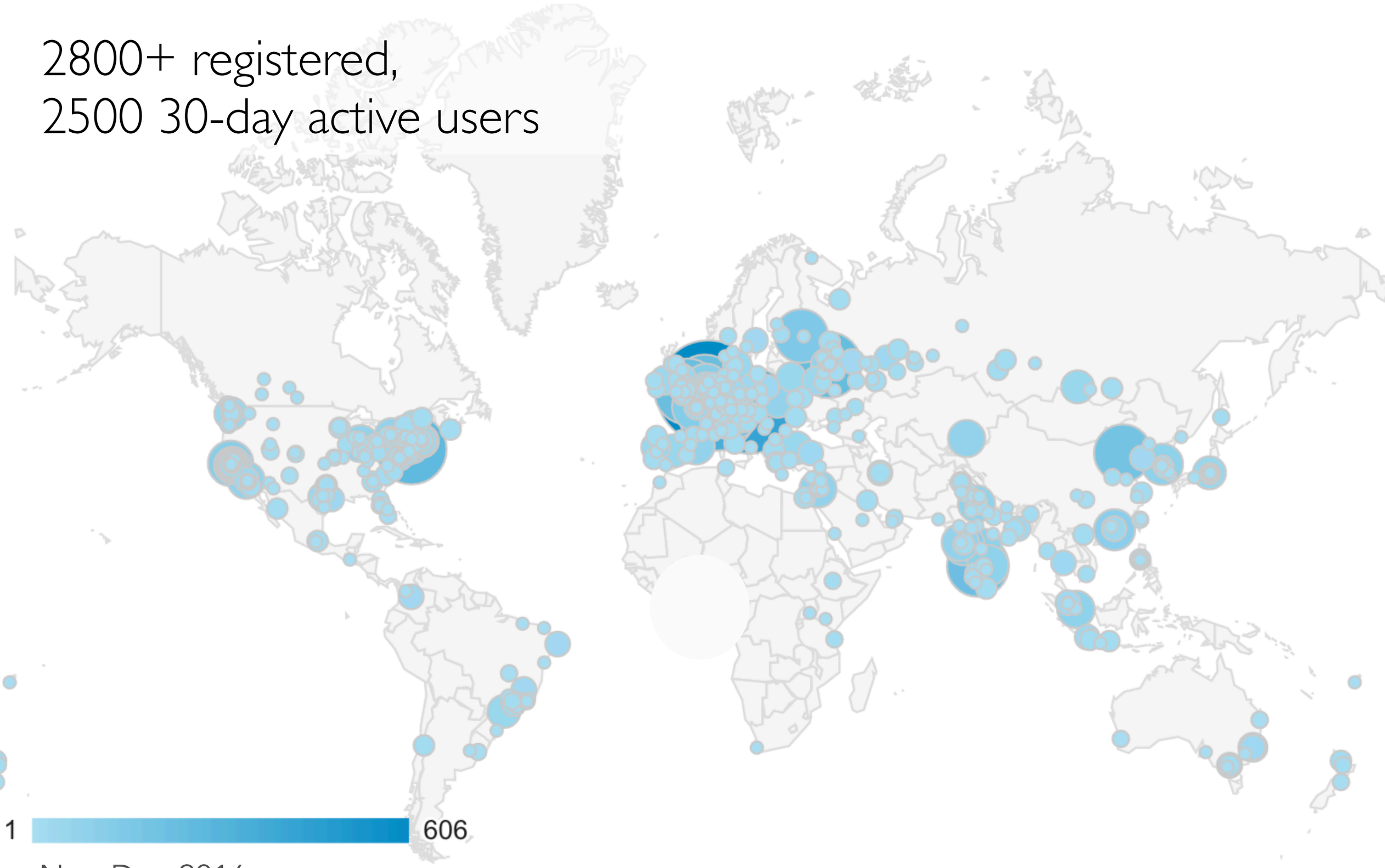
1698629 runs

Upload and explore all **results** online.

Download and share data, flows and runs through:

OpenML Community

2800+ registered,
2500 30-day active users



Collaboration tools (in progress)

Sharing settings
Share datasets, flows, studies with certain people.
Easily publish them later.

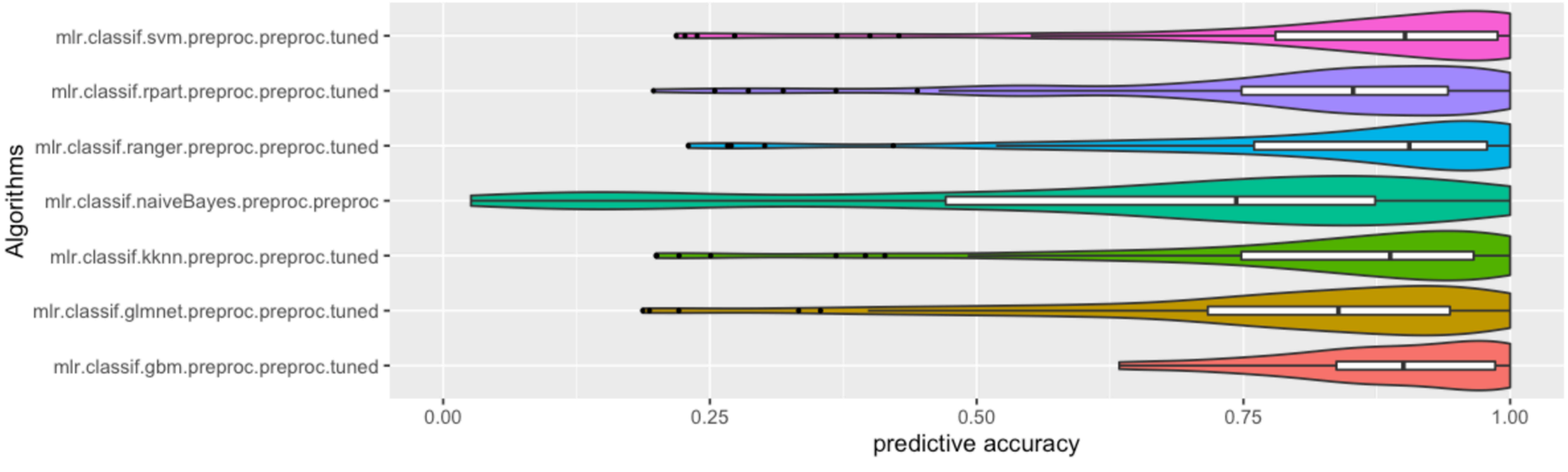
Studies
Online counterpart of a paper
Linked to GitHub, Jupyter notebooks

Code submissions
Sharing versioned code, docker images, archiving
GitHub integration

For data scientists

Instead of isolated benchmarking results in papers, have an open online benchmark that everyone can extend

- Curated list of benchmark datasets (e.g. OpenML100)
- Reuse available benchmarks
  - Algorithms from WEKA, R, scikit-learn,…
  - Includes data cleaning, hyperparameter optimization
- Code for analysis, comparison (notebooks)



For e

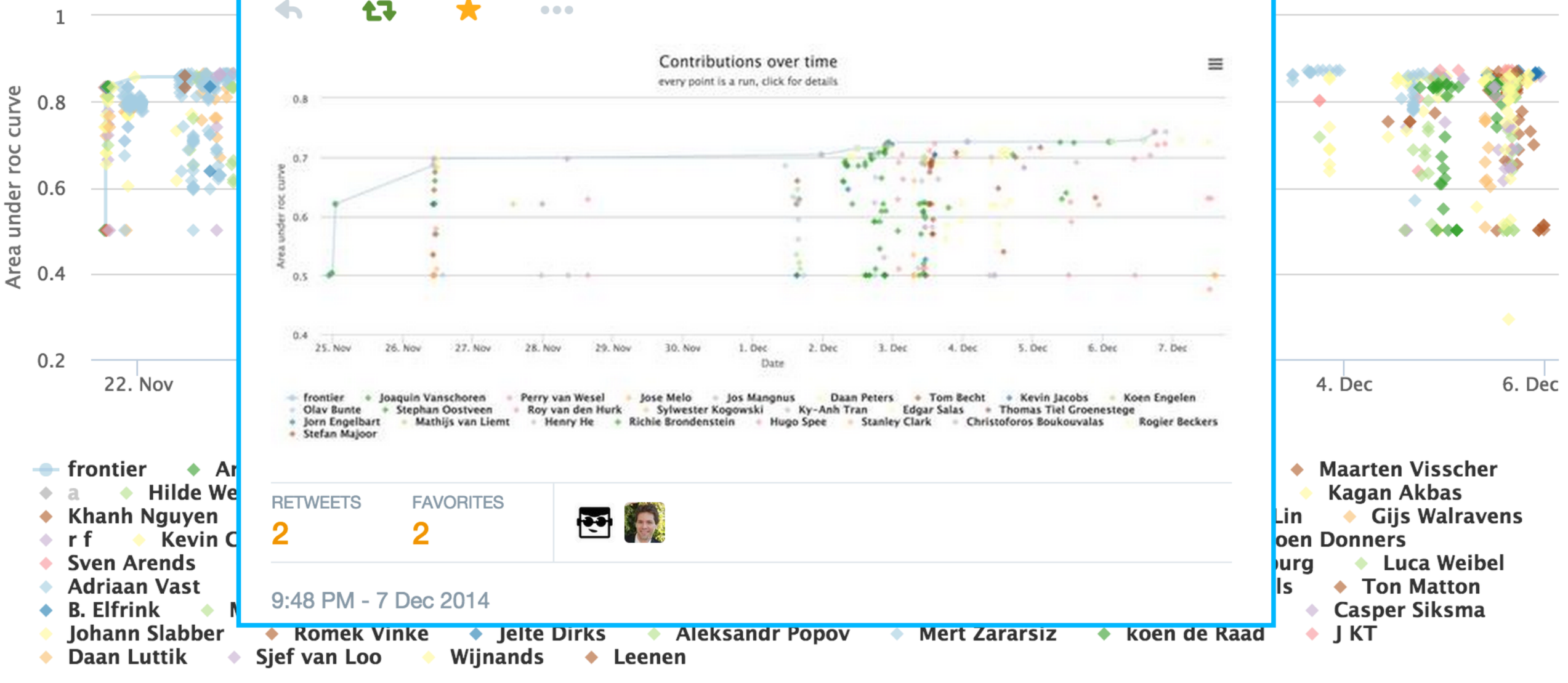
Real-time c

Students ca

Simple resu

Hidden tes

tter



Rogier Beckers
@RogierBeckers

Follow

Het bewijs dat ik studeer op zondag!
“@joavanschoren: #Machinelearning students on a #collaborative data mining ”

View translation

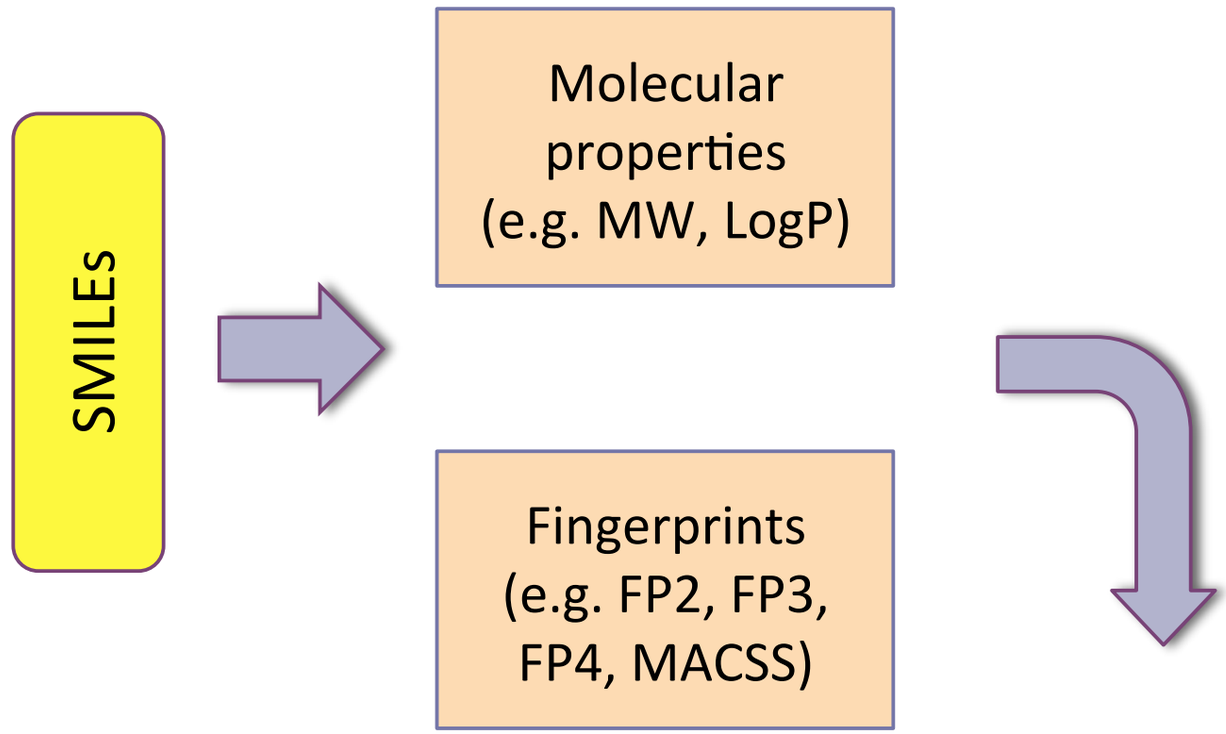
Lauradorp, Landgraaf

RETWEETS	FAVORITES
2	2

9:48 PM - 7 Dec 2014

For scientists

Example: predict which drugs will inhibit certain proteins (and hence viruses, parasites,…)



ChEMBL database

1.4M compounds,10k proteins,
12,8M activities

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+ QSAR datasets

2750 targets (proteins), x 6 feature representations

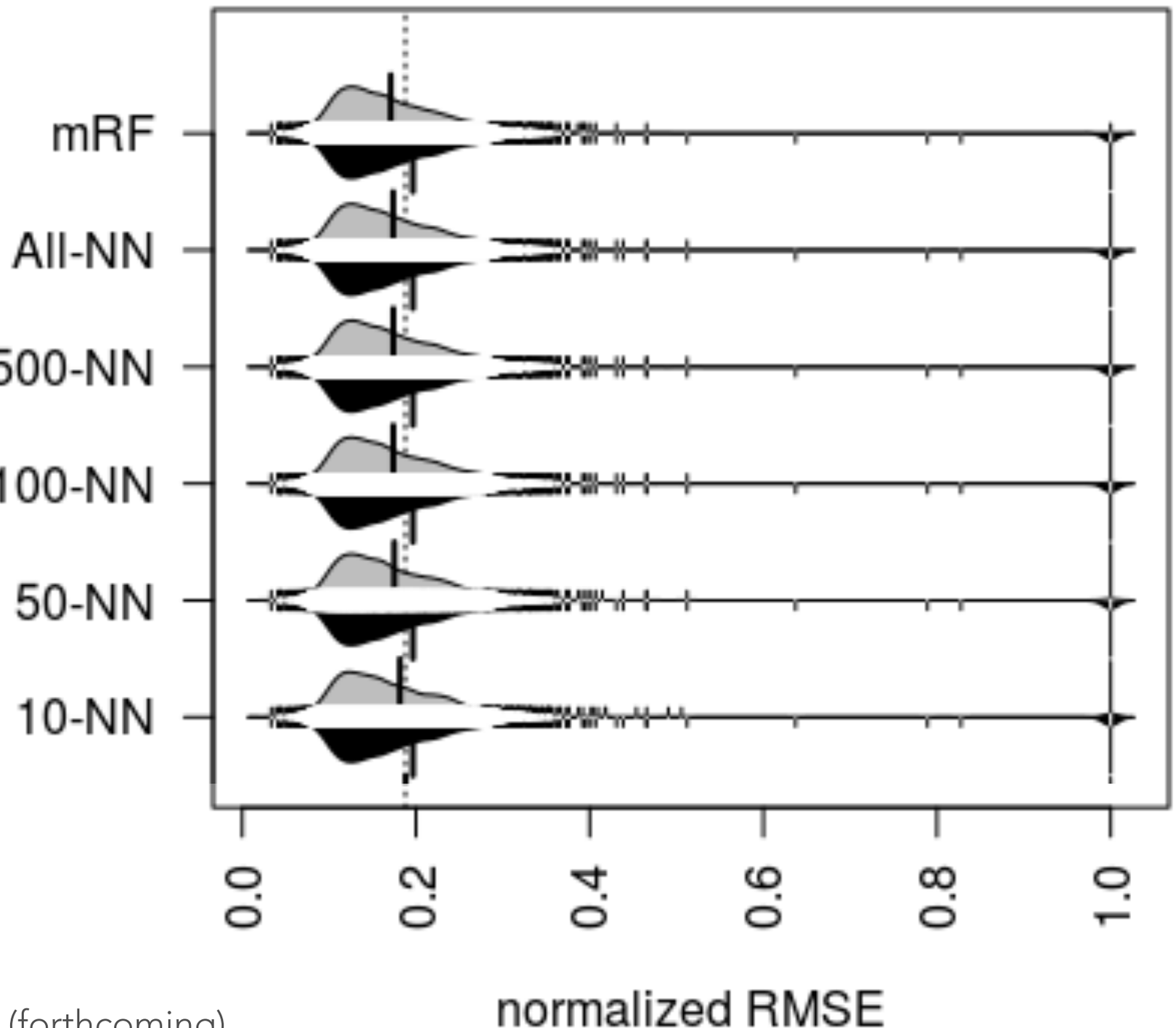
For scientists

Meta-learning: meta-models (grey) that outperform the state-of-the-art (black)

Meta-learner:

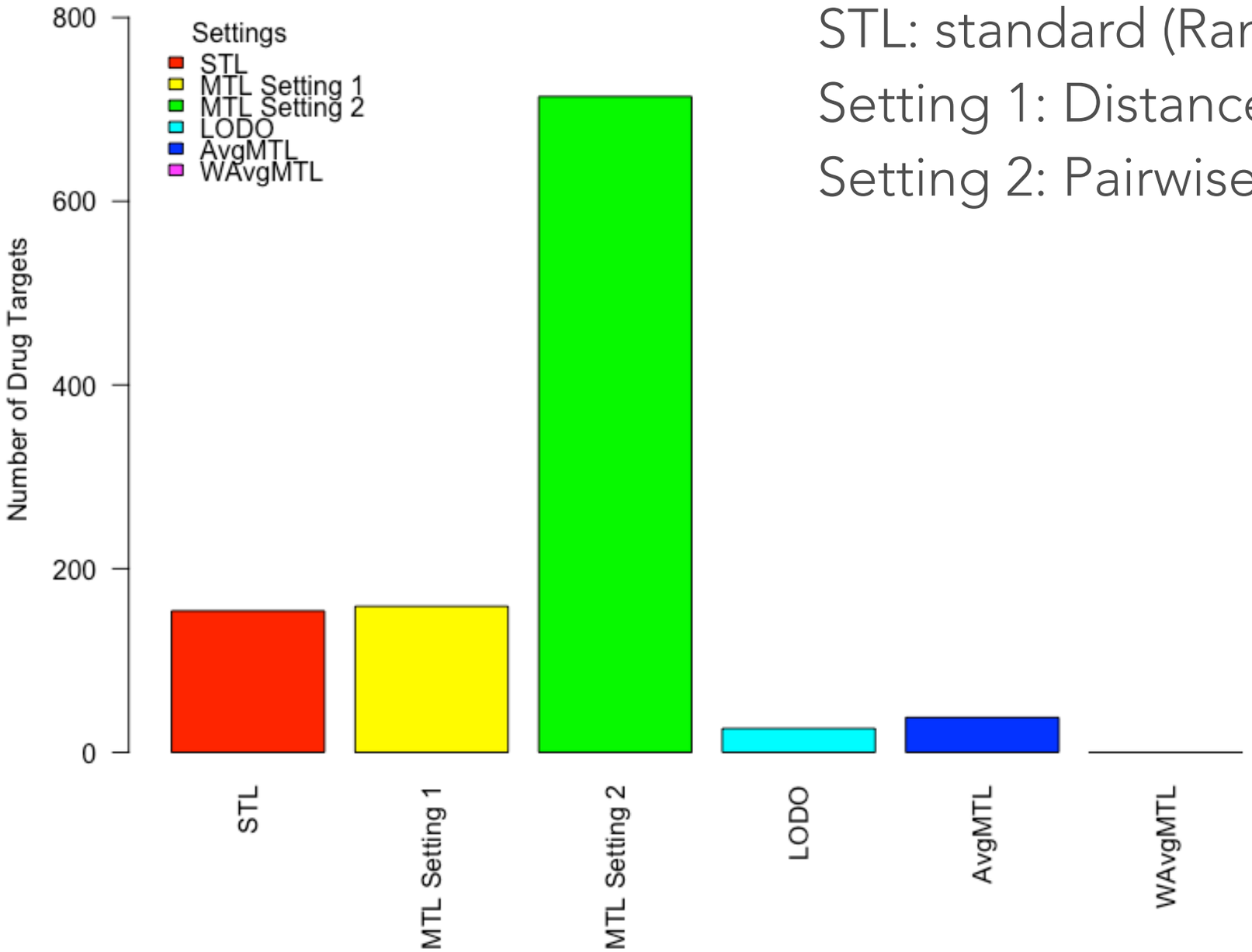
mRF: RandomForest

k-NN: kNN



I. Olier et al. (2017) Machine Learning (forthcoming)

For scientists



When few drugs are tested on a given target, include data on *related* targets:

STL: standard (Random Forests)

Setting 1: Distance in taxonomy (ChEMBL)

Setting 2: Pairwise sequence alignment

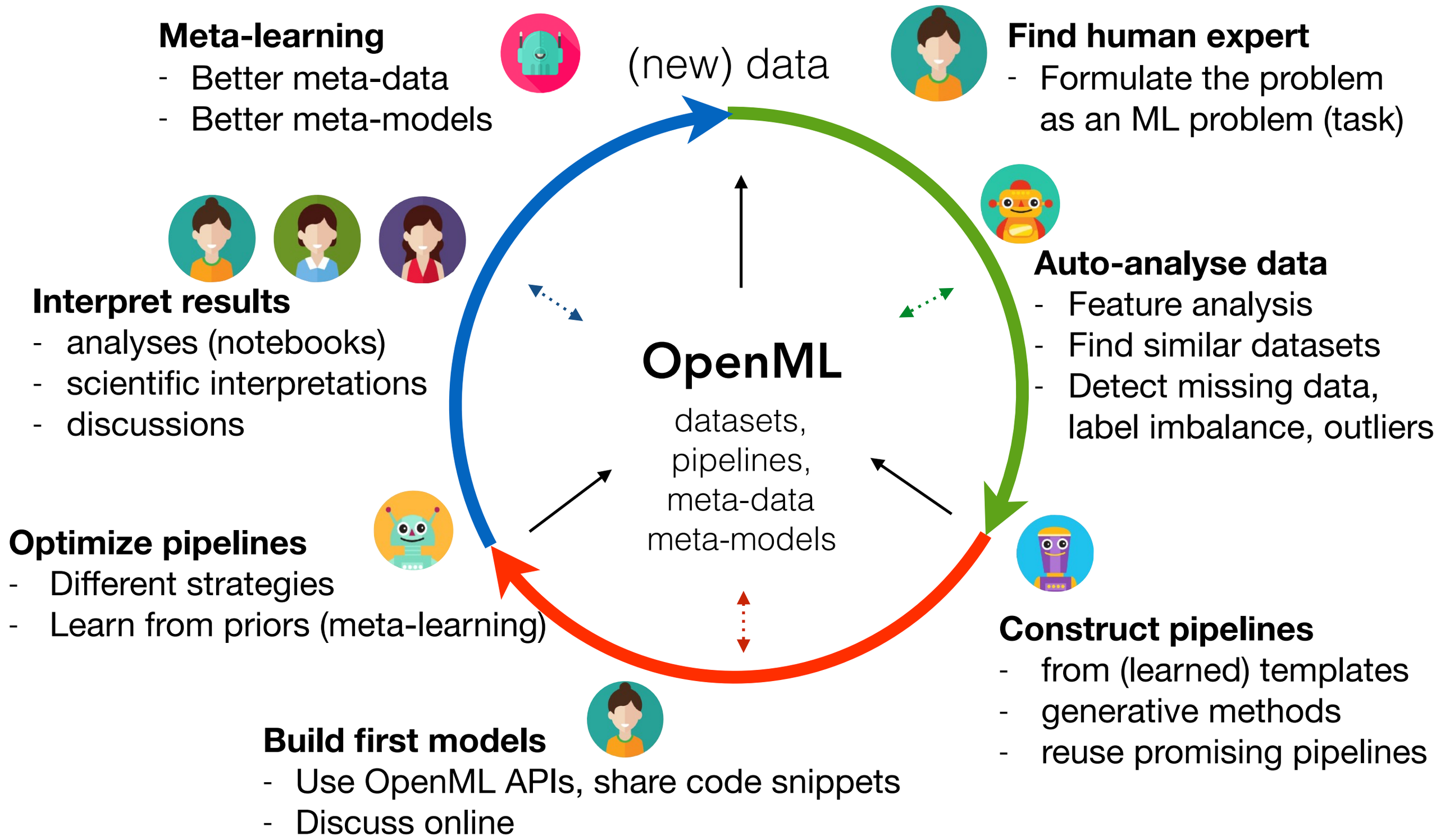
We just scratched the surface. All data is available on OpenML.

Automating machine learning

Data-driven modelling
Learn how to build models based on many prior experiments

AI-human interaction
Bots/services to simplify work
Find data, construct pipelines, optimize hyperparameters,…

Automating machine learning: a human-robot symbiosis



Meta-learning
- Better meta-data
- Better meta-models

Find human expert
- Formulate the problem as an ML problem (task)

Interpret results
- analyses (notebooks)
- scientific interpretations
- discussions

Auto-analyse data
- Feature analysis
- Find similar datasets
- Detect missing data, label imbalance, outliers

Optimize pipelines
- Different strategies
- Learn from priors (meta-learning)

Construct pipelines
- from (learned) templates
- generative methods
- reuse promising pipelines

Build first models
- Use OpenML APIs, share code snippets
- Discuss online

Auto-experimentation bot

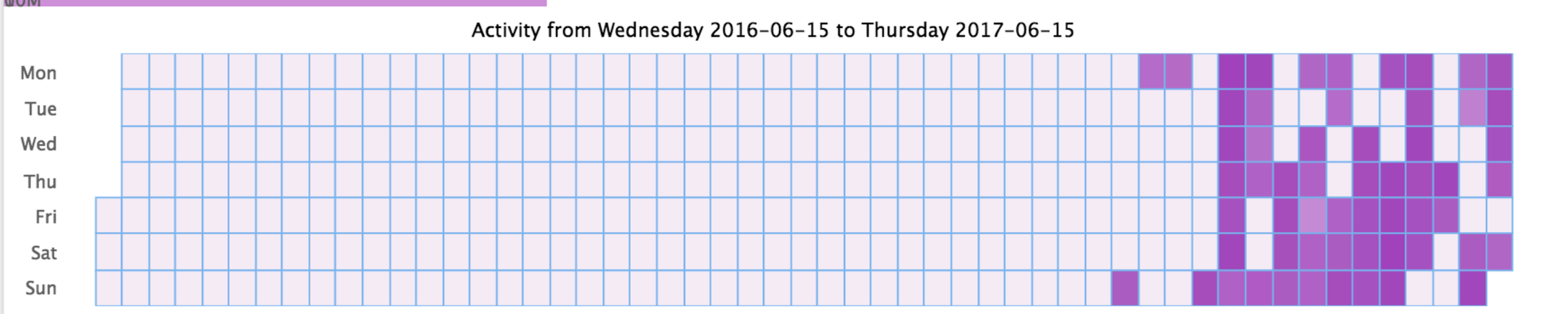
OpenML_Bot R

Joined 2017-03-07

Activity	Reach	Impact	Uploads				
307034.5	4	0	0	30	0	307004	EDIT PROFILE

	Activity	Reach	Impact
Data Sets	0	0	0 0
Flows	30	1	0 0
Tasks	0	0	0
Runs	307004	3	0

Activity: 307K 307K 0 1



Pipeline construction bot

- Tree-based pipeline optimization (TPOT)
  - Genetic programming to construct pipelines out of most successful ones

```
from tpot import TPOTClassifier
from sklearn.datasets import load_digits
from sklearn.model_selection import train_test_split

digits = load_digits()
X_train, X_test, y_train, y_test = train_test_split(digits.data, digits.target,
                                                    train_size=0.75, test_size=0.25)

tpot = TPOTClassifier(generations=5, population_size=50, verbosity=2, n_jobs=-1)
tpot.fit(X_train, y_train)
```

```
Optimization Progress:   0%|          | 0/300 [00:00<?, ?pipeline/s]
```

```
print(tpot.score(X_test, y_test))
```

Runtime prediction bot

- Regression model (Random Forest) trained on 100k+ experiments
- Predicts runtime accurate within a few seconds

Regressor	RF	SVC	Tree	NB	Boosting	LR	kNN
Median	0.39	0.88	0.74	0.83	0.86	0.57	0.72
Mean	0.4	0.88	0.75	0.86	0.86	0.57	0.73
Extrapolate	0.44	0.80	0.45	1.33	0.33	0.84	1.09
RR	0.12	0.15	0.43	0.54	0.1	0.21	0.3
RF	**0.07**	**0.12**	**0.28**	**0.47**	**0.07**	**0.11**	**0.16**
SVR	0.39	0.87	0.72	0.86	0.84	0.56	0.71

Table 3: Mean Absolute Deviation of $\log_{10}$ runtime predictions. Best results are bold.

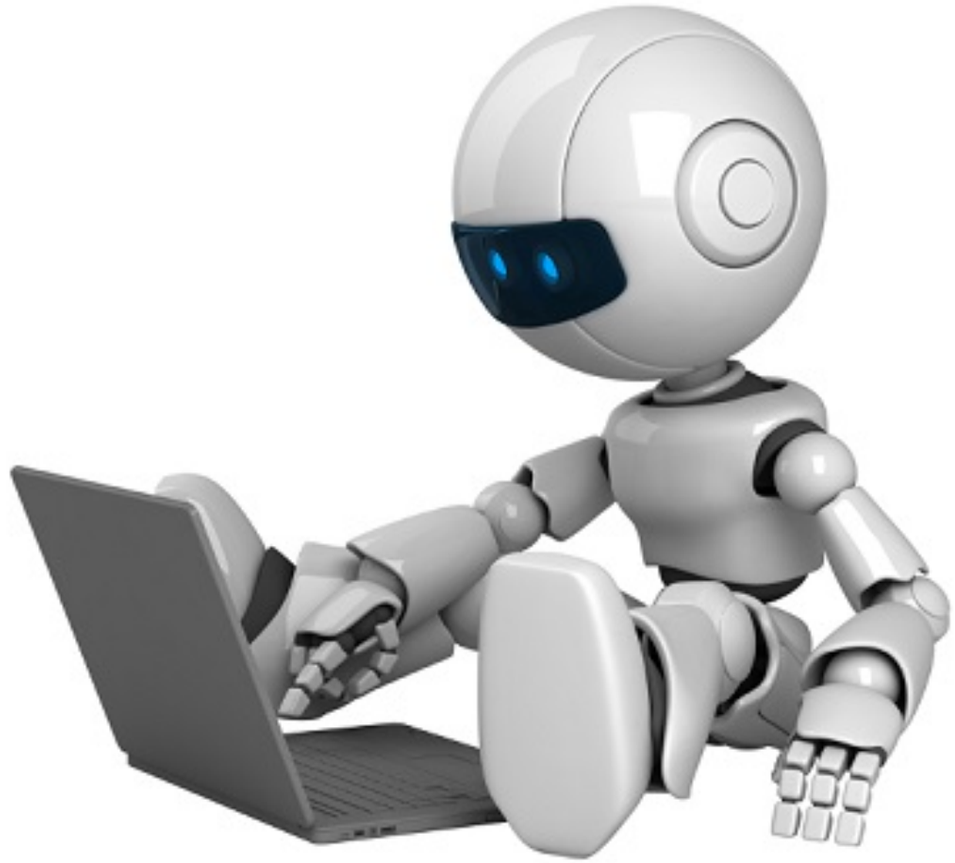
PhD vacancy

Topic: Automatic Machine Learning

- Automatically generate pipelines: preprocessing + modeling
- Work with OpenML
- Develop bots

Location: TU Eindhoven

Start: As soon as possible :)

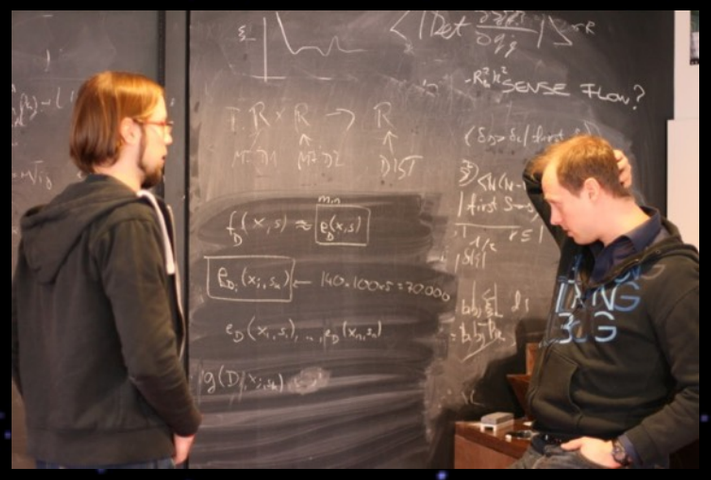


Join Us!

www.openml.org

Next hackathon:

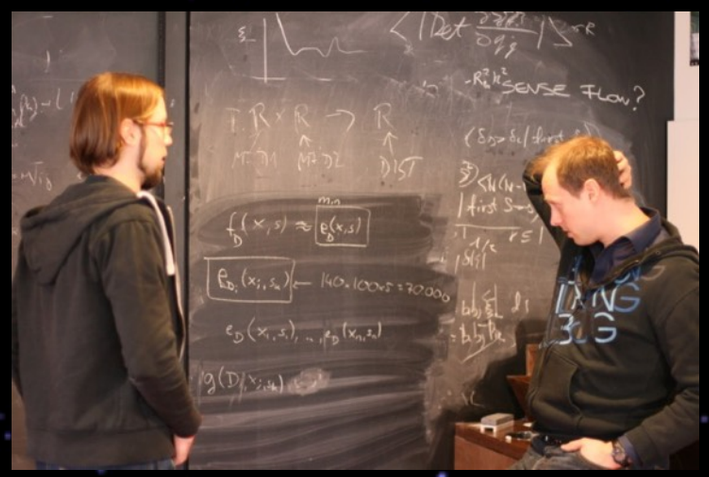
- October 9-13
- Lorentz Center, Leiden



Help us :)

We are always looking for:

- Code contributions (open source)
  - New tool/platform integrations
    - E.g. Keras/TensorFlow
  - New bots
  - Your own ideas
- Interesting datasets
- Computing resources
- Funding ideas



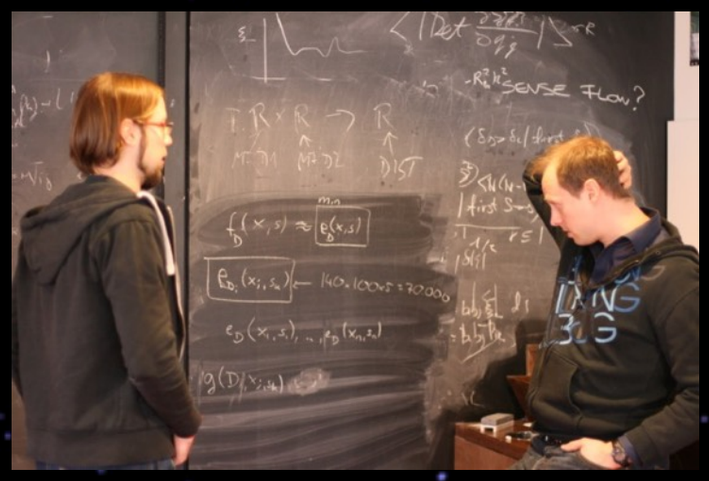
@open_ml

OpenML

Thank You

Now try it yourself :)










OpenML hands on

Register on OpenML.org

Explore datasets, results via website or scripts

Build you own ML models and share them

https://www.openml.org/guide#!r
-> Tutorial

https://www.openml.org/guide#!python
-> Quickstart Notebook

https://www.openml.org/guide#!plugin_weka
-> Quickstart