@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

@open_ml

OpenML

DEMOCRATIZING AND AUTOMATING MACHINE LEARNING

JOAQUIN VANSCHOREN, TU EINDHOVEN, 2016

Research different.

Polymaths: Solve math problems
by massive **online** collaboration

Broadcast question, combine
many minds to solve it

Networked Science

Serendipity: what's hard for one person is easy for another

Collaboration only scales if **all friction is eliminated**

Easy, organized, access to data, code, and results

WHAT IF WE CAN EXPLORE DATA

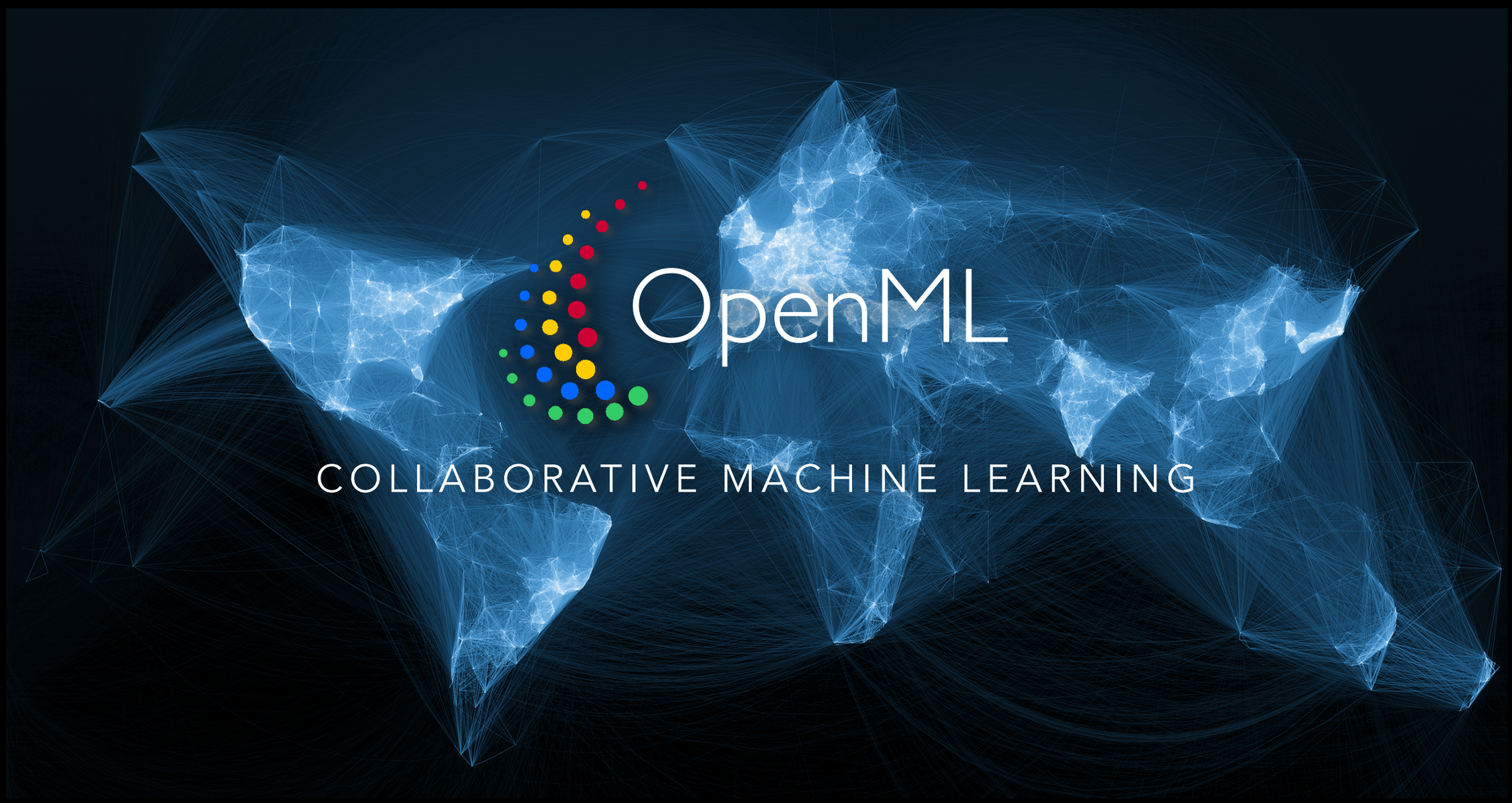
COLLABORATIVELY

WHAT IF WE CAN EXPLORE DATA

COLLABORATIVELY

ON WEB SCALE


WHAT IF WE CAN EXPLORE DATA
COLLABORATIVELY
ON WEB SCALE
IN REAL TIME



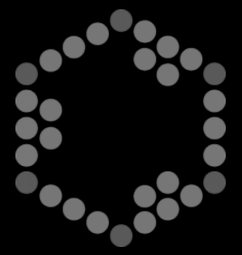
Easy to use: Integrated in many ML environments

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

Organized data: Reproducible, connected to data, code, people

Reward structure: Build reputation and trust

Self-learning: Learn from millions of experiments to help users



Data (ARFF) uploaded or referenced, versioned

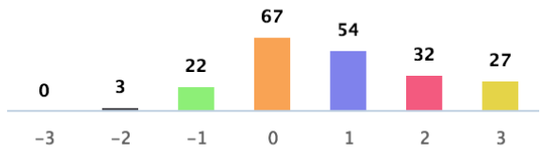
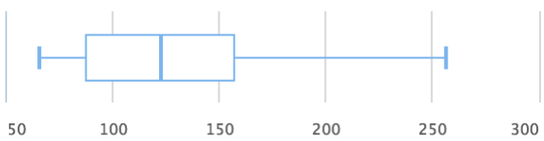
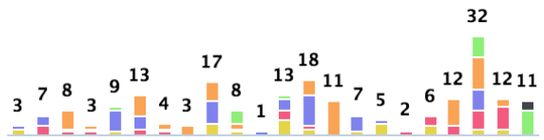
analyzed, characterized, organized online

analyzed, characterized, organized online

\+ visualizations, statistics, landmarkers, error checking,

queryable through website + API

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

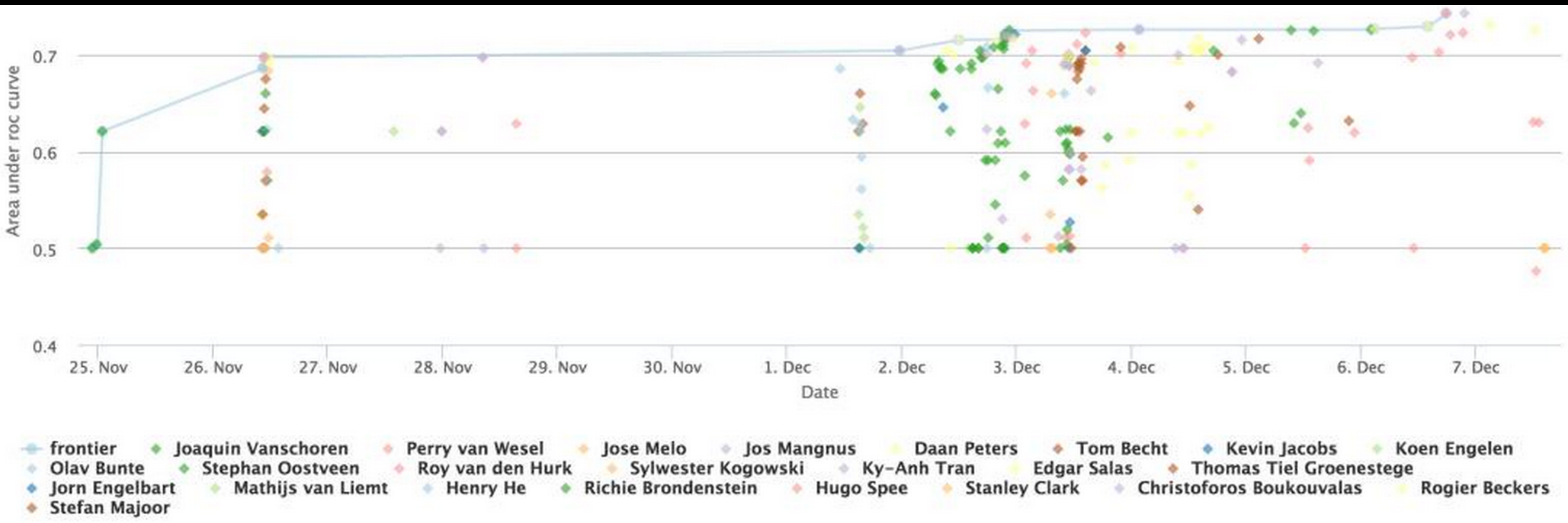
DefaultAccuracy	**0.33**	The predictive a
NumberOfClasses	**7**	The number of c
NumberOfFeatures	**26**	The number of f
NumberOfInstances	**205**	The number of i
NumberOfMissingVal...	**59**	Counts the total

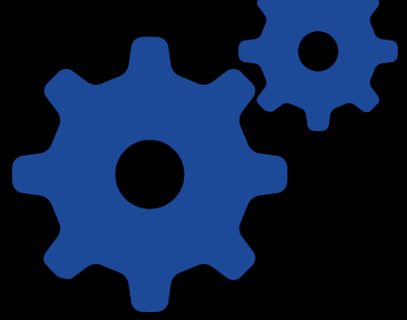
Tasks contain data, goals, procedures.

Readable by tools, automates experimentation

All results organized online: **realtime overview**

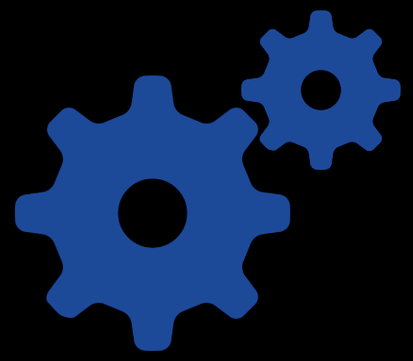
All results organized online: **realtime overview**



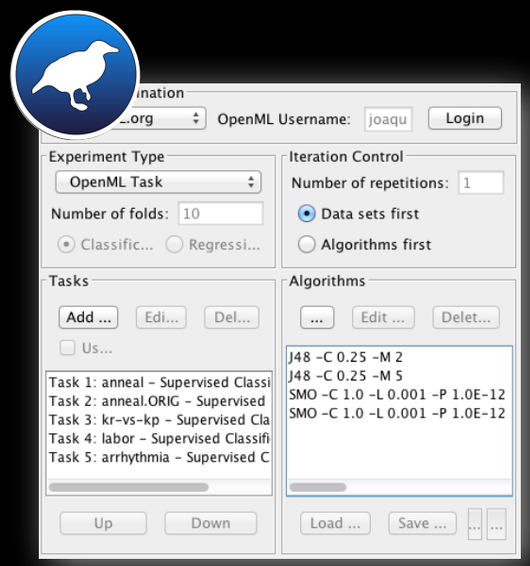


Flows (code) run anywhere, using your favorite tools

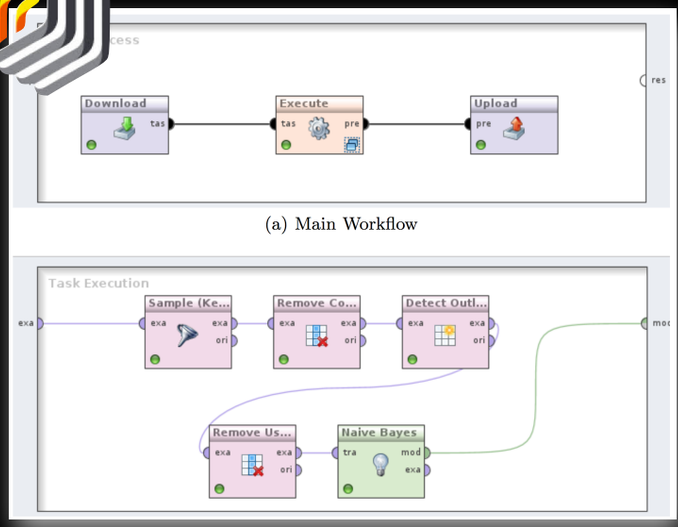
Integrations + APIs (REST, R, Python, Java,…)



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

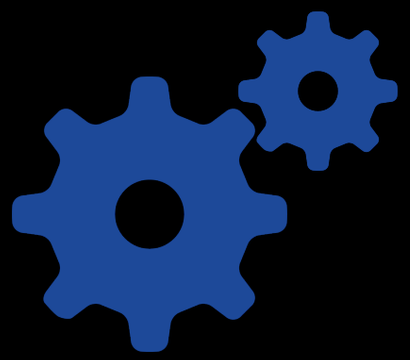


```
from sklearn import tree
from openml import tasks,runs
task = tasks.get_task(14951)
clf = tree.DecisionTreeClassifier()
run = runs.run_task(task, clf)
return_code, response = run.publish()
```



(a) Main Workflow

```
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,...)

```
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()
```

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.

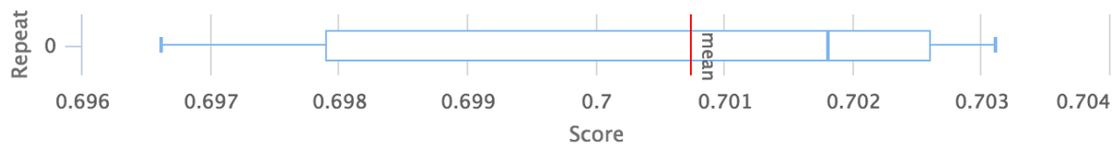
Area under ROC curve

0.7007 ± 0.0023

Per class

0	1
0.7007	0.7007

Cross-validation details (10-fold Crossvalidation)



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**

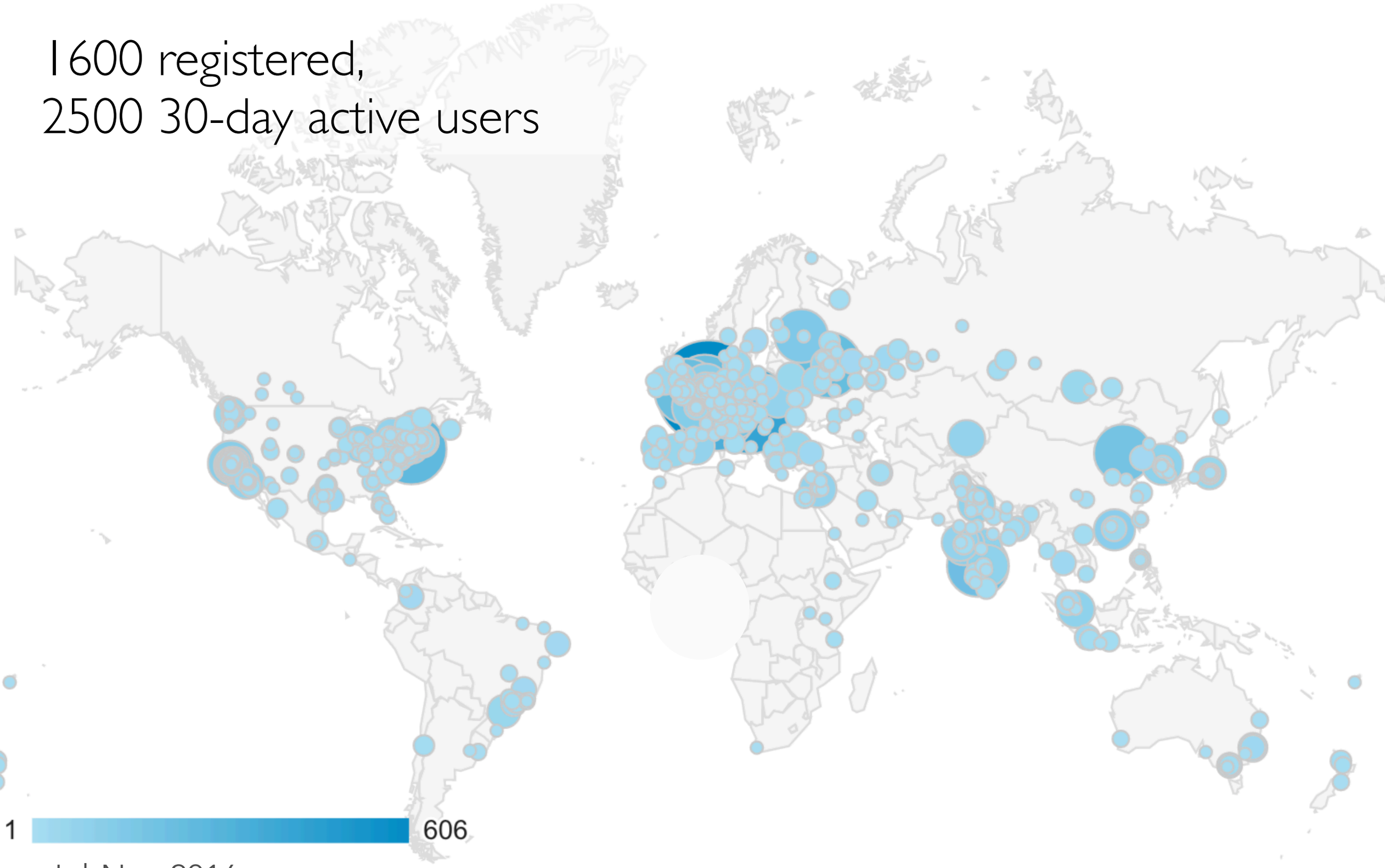
1698629 runs

Upload and explore all **results** online.

Download and share data, flows and runs through:

OpenML Community

1600 registered,
2500 30-day active users



Collaboration tools (in progress)

Circles
Create collaborations with trusted researchers

Studies (e-papers)
Online counterpart of a paper, linkable

Reputation
Auto-tracking of your activity, reach, impact

Notebooks
Easy online collaboration on data analysis scripts

Classroom challenges

Results a

Students

Simple re

Hidden t

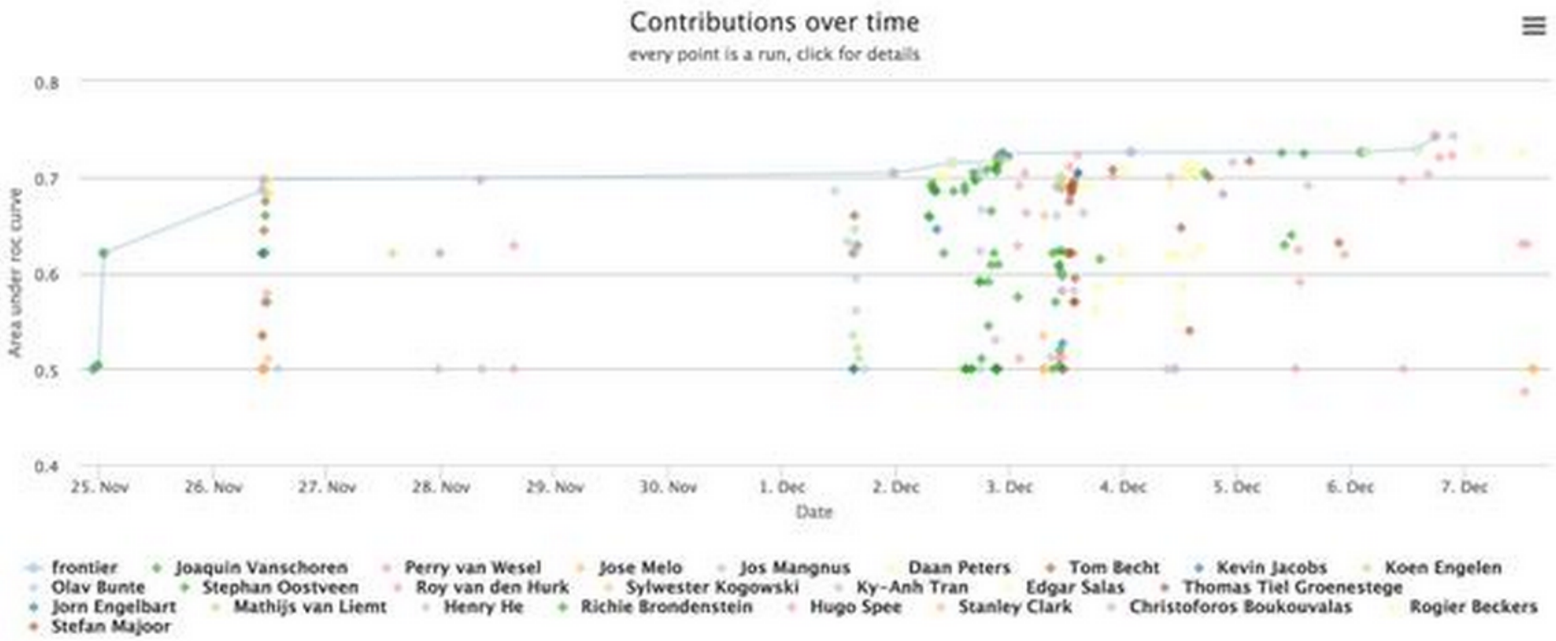
Rogier Beckers
@RogierBeckers

Follow

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

View translation

Lauradorp, Landgraaf

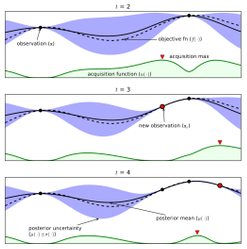


RETWEETS 2

FAVORITES 2

9:48 PM - 7 Dec 2014

Upcoming realtime challenges

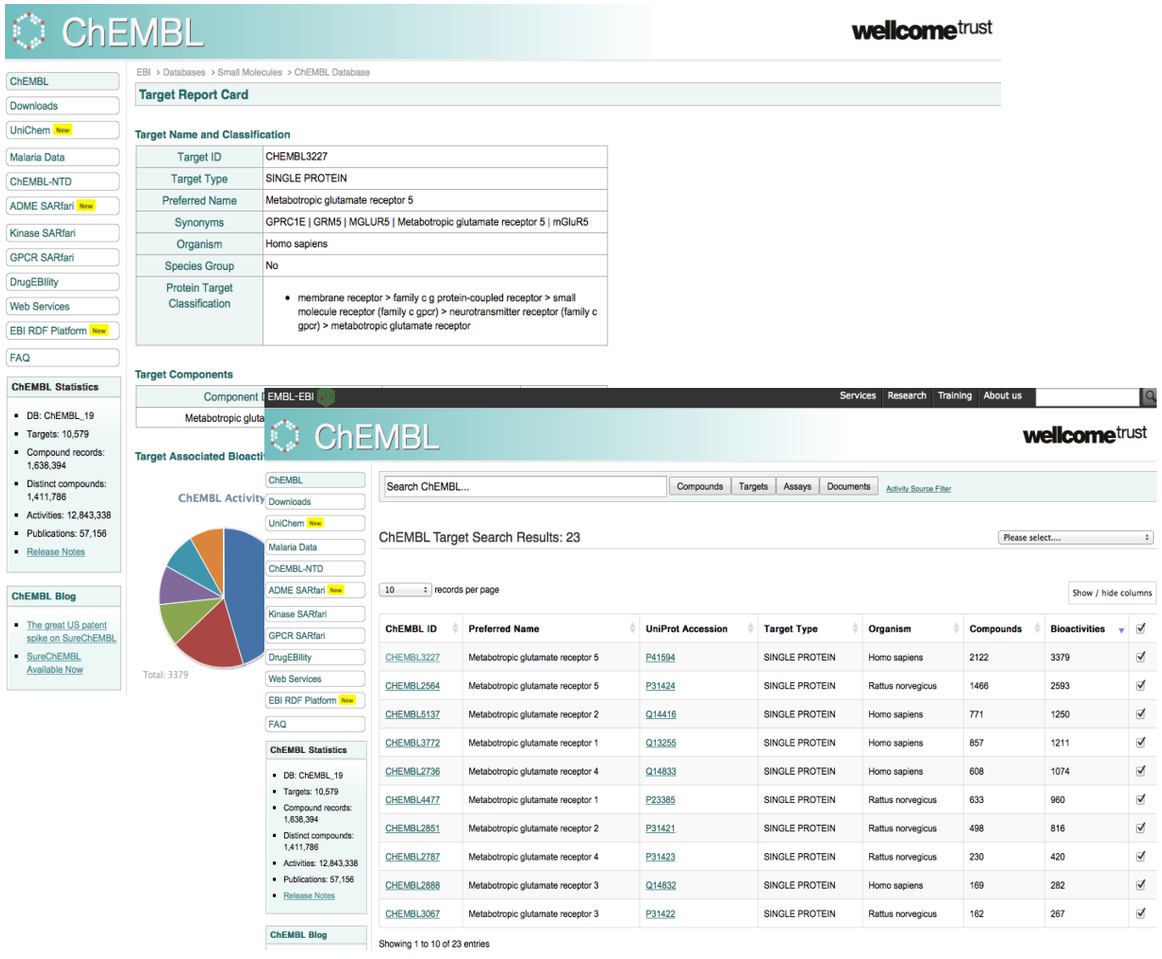


Hyperparameter optimization challenge
Every iteration can be uploaded, prior ones downloaded

Predict energy usage in smart energy
API streams data, predictions uploaded hourly/daily

OpenML in drug discovery

Predict which drugs will inhibit certain proteins (and hence viruses, parasites,…)



SMILES

Molecular properties (e.g. MW, LogP)

Fingerprints (e.g. FP2, FP3, FP4, MACSS)

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	
							⋮					

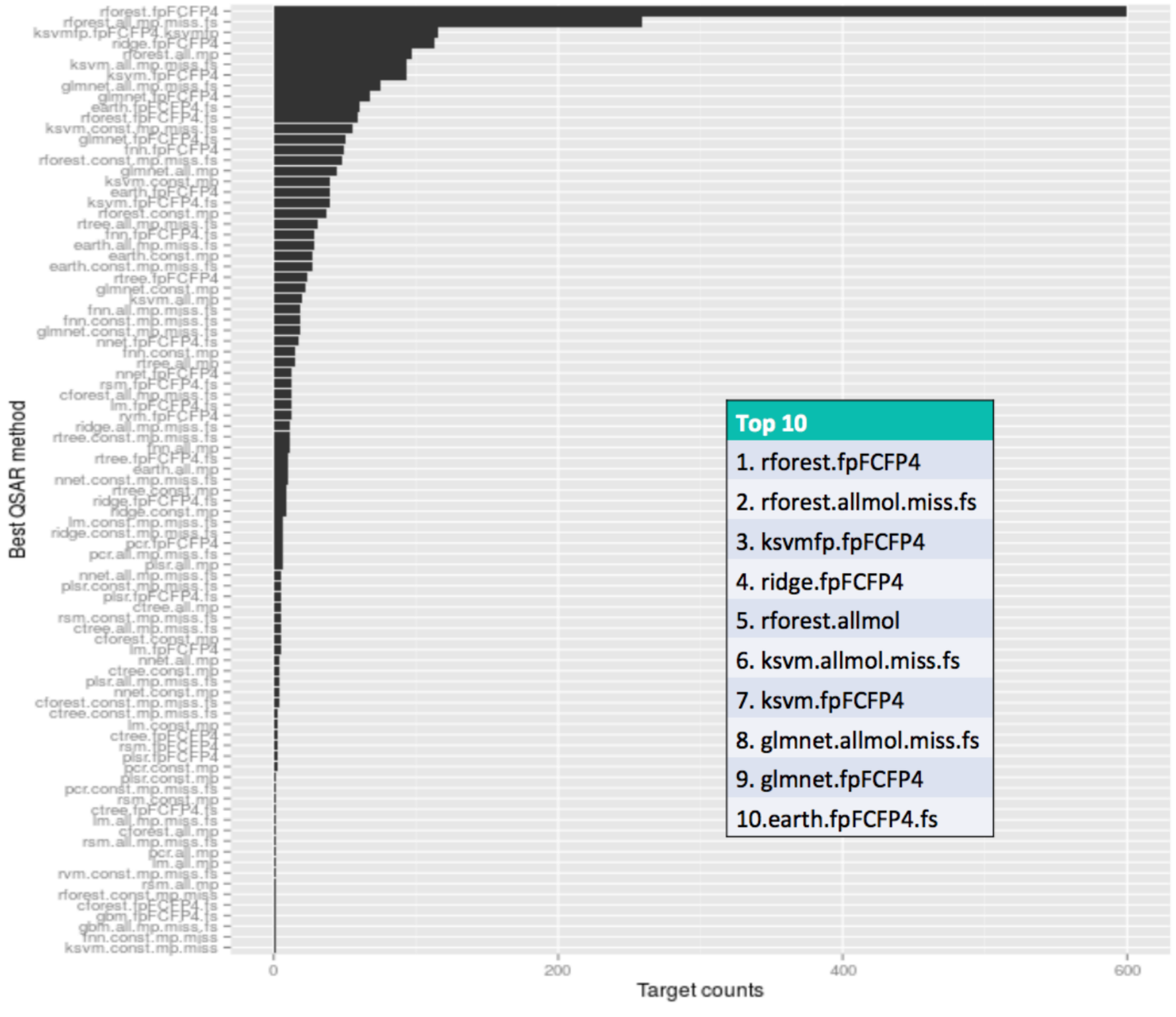
16.000+ QSAR datasets

2750 targets (proteins), x 6 feature representations

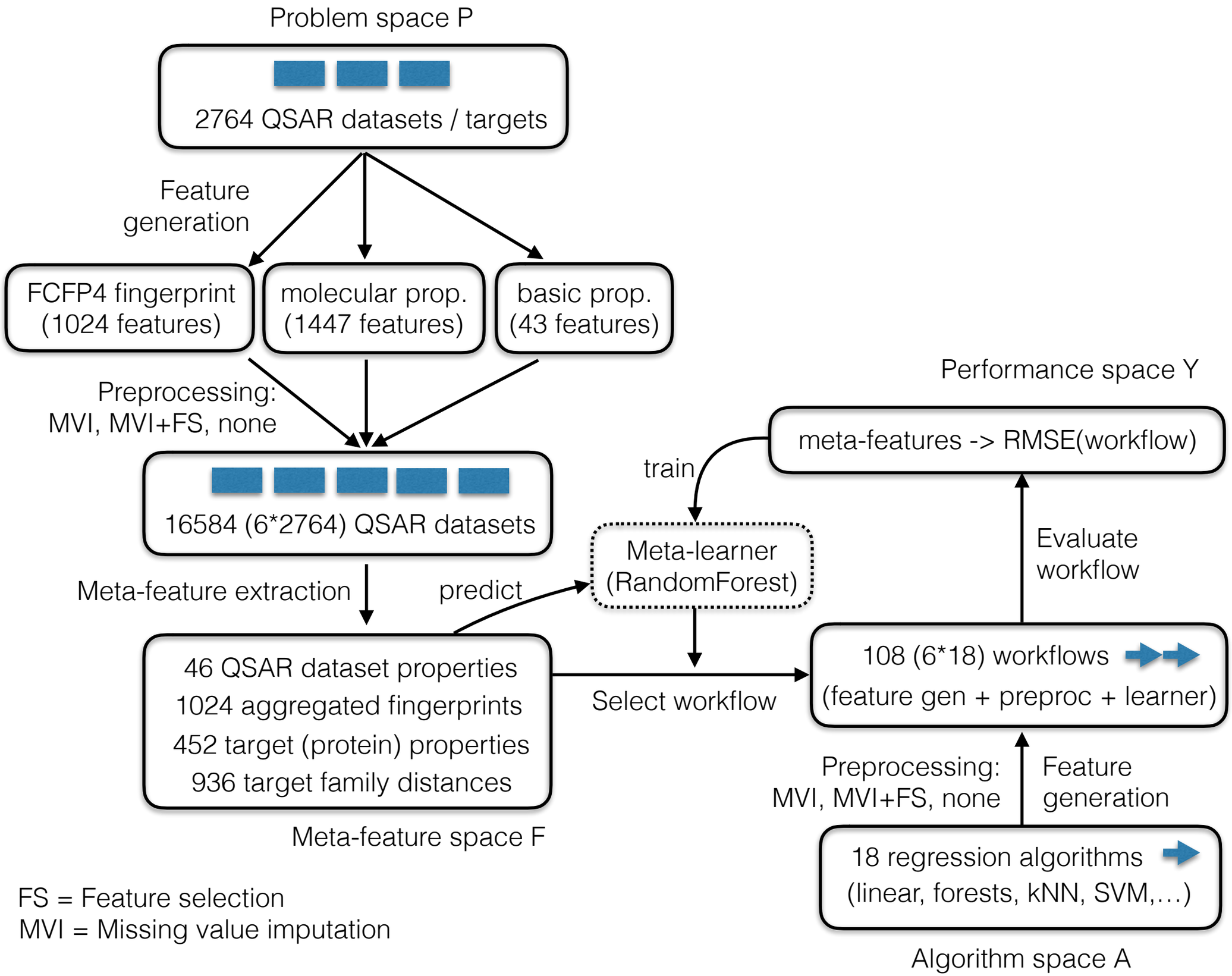
OpenML in drug discovery

Best algorithms?

Best features?



Predicting workflows with MetaLearning



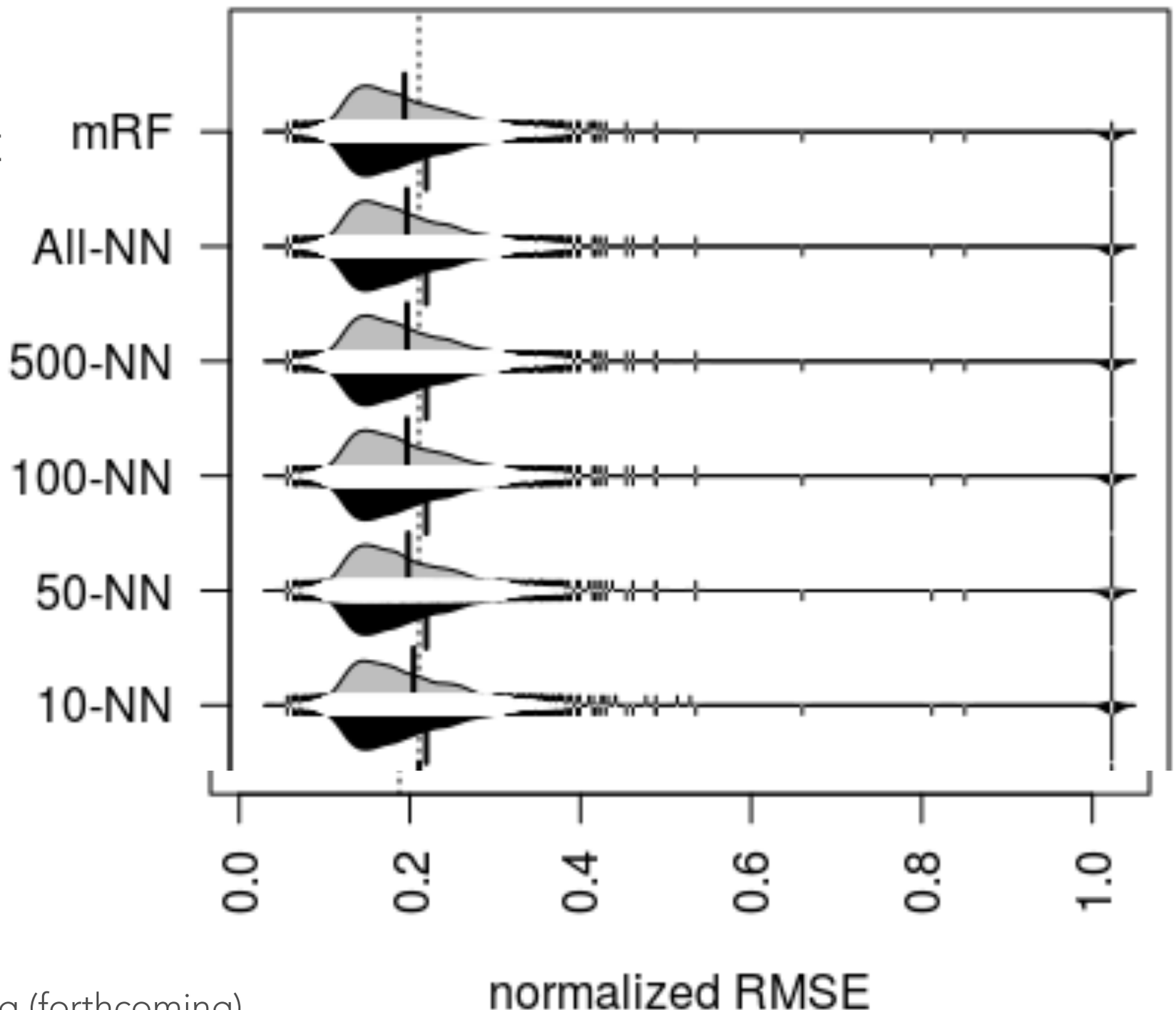
OpenML in drug discovery

Random forest (black) vs meta-learner (grey)

Meta-learner:

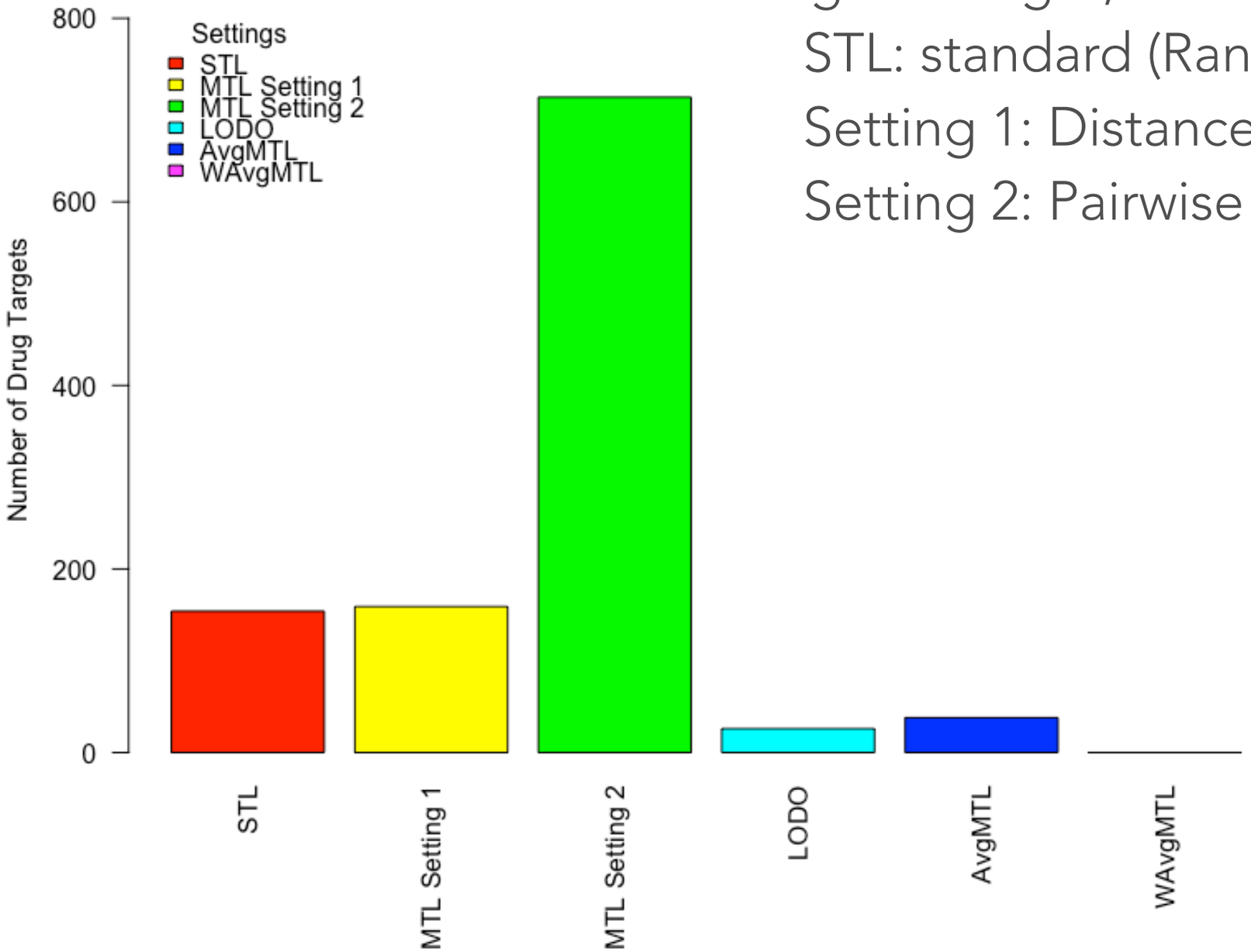
mRF: Random forest

k-NN: kNN



I. Olier et al. (2016) Machine Learning (forthcoming)

OpenML in drug discovery



Multi-target learning: if few drugs are tested on a given target, include data on 'related' targets:

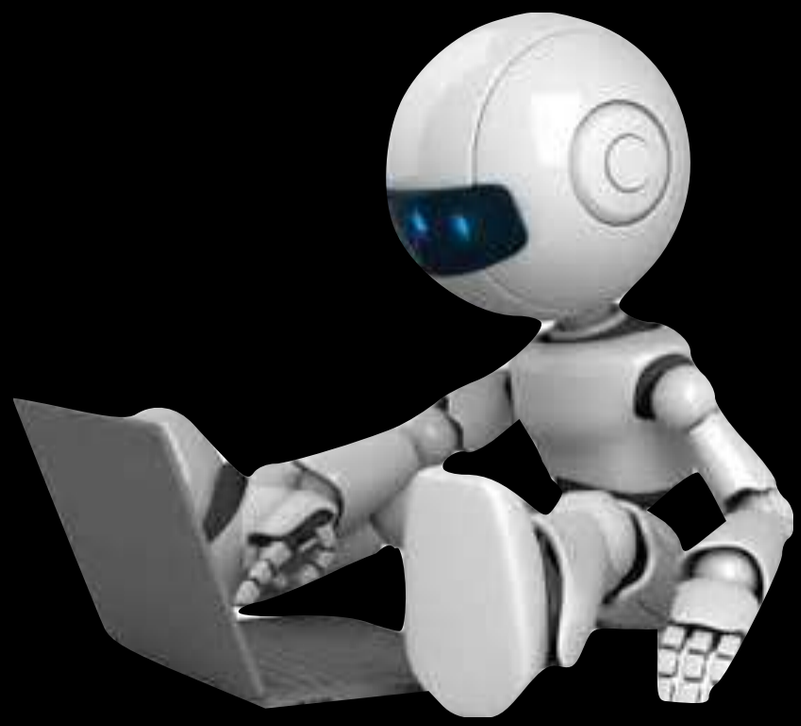
STL: standard (Random Forests)

Setting 1: Distance is taxonomy (ChEMBL)

Setting 2: Pairwise sequence alignment

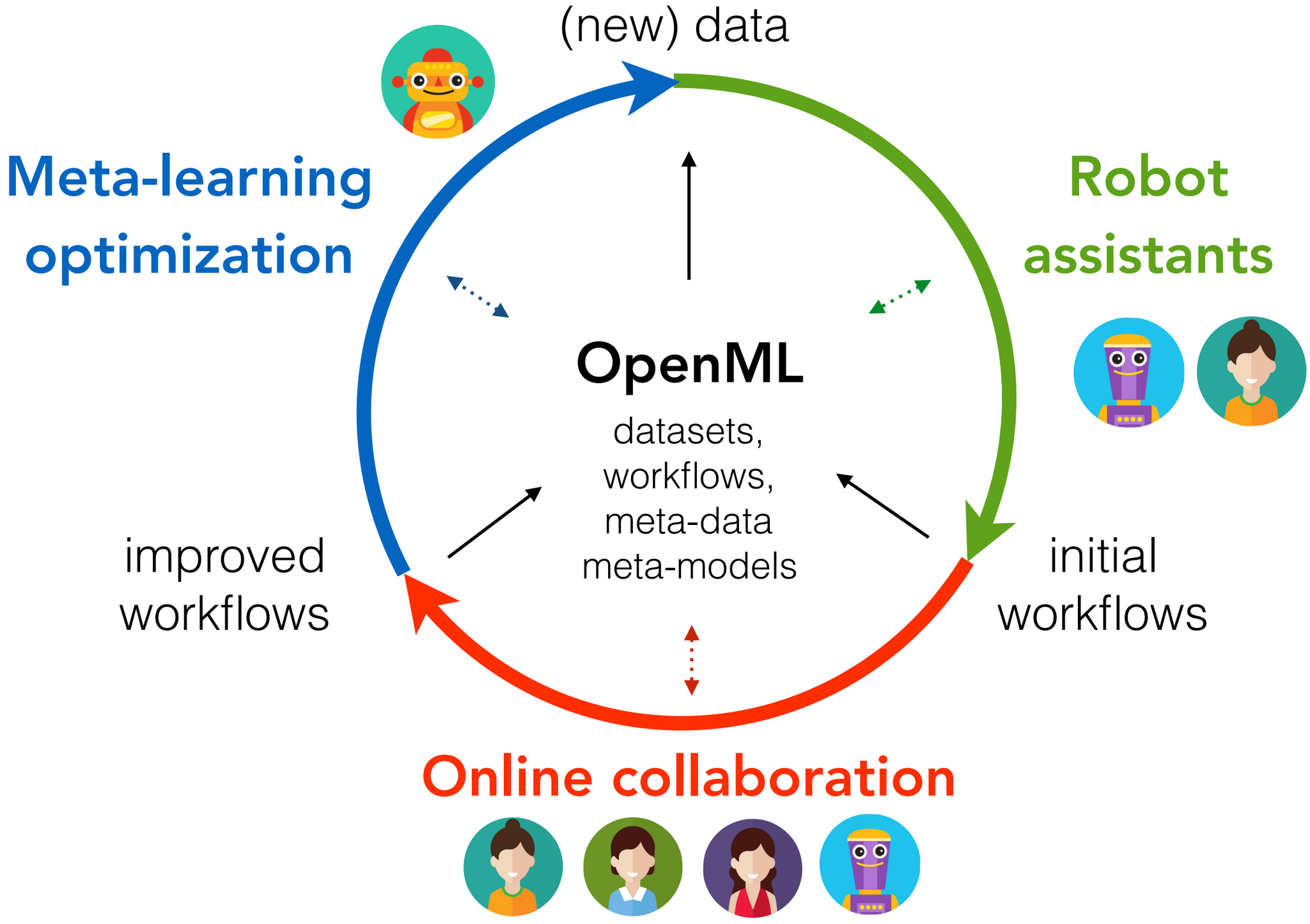
We just scratched the surface. All data is available on OpenML.

Automating machine learning

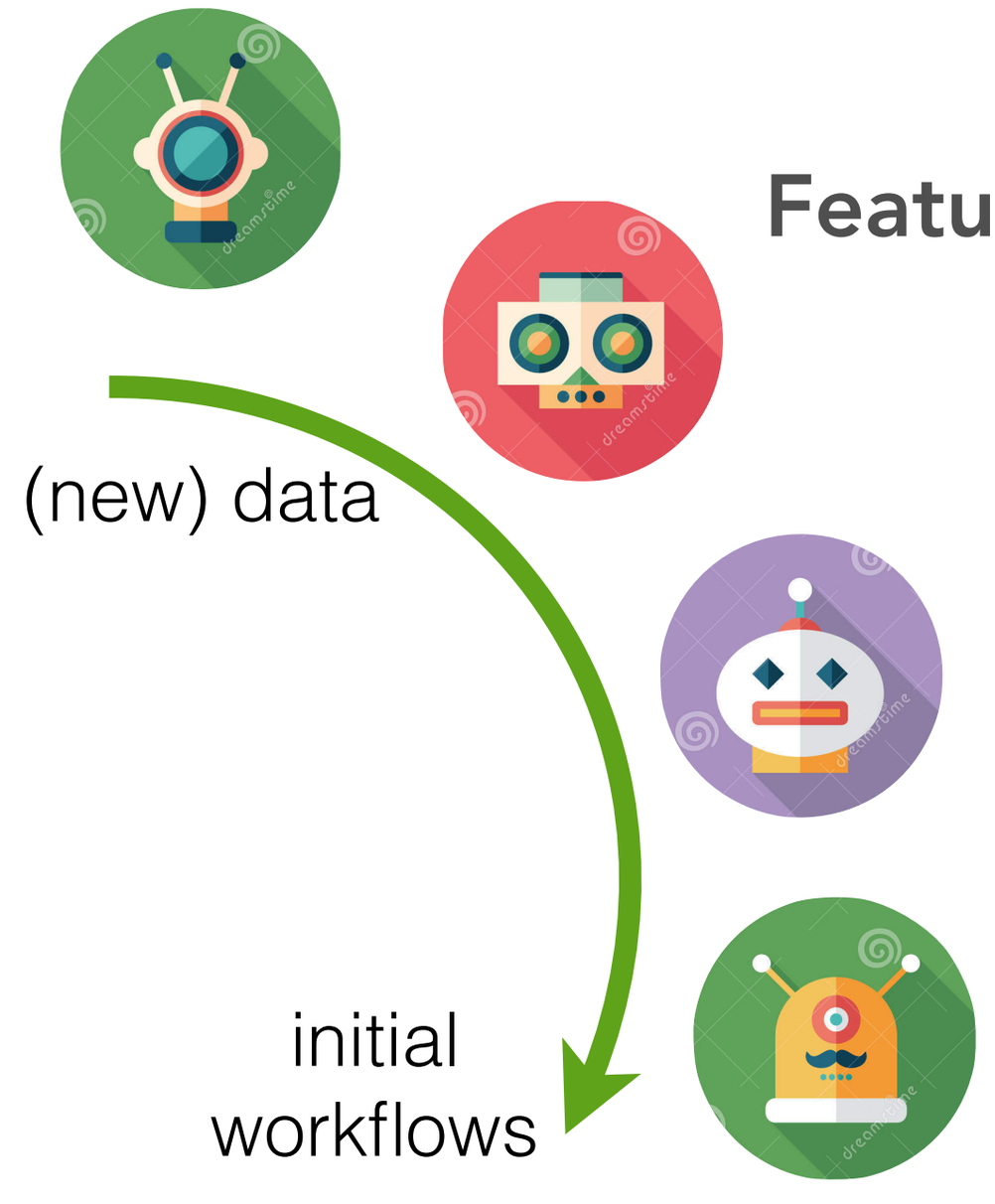


- Lots of experimental data: learn to learn across datasets
- Bots that automatically run algorithms on data
- Data-driven modeling: learn how to build models based on many prior experiments

Automating machine learning: a human-robot symbiosis (future)



Robot assistants



Runtime prediction bot: predicts how long an ML algorithm will run on your data

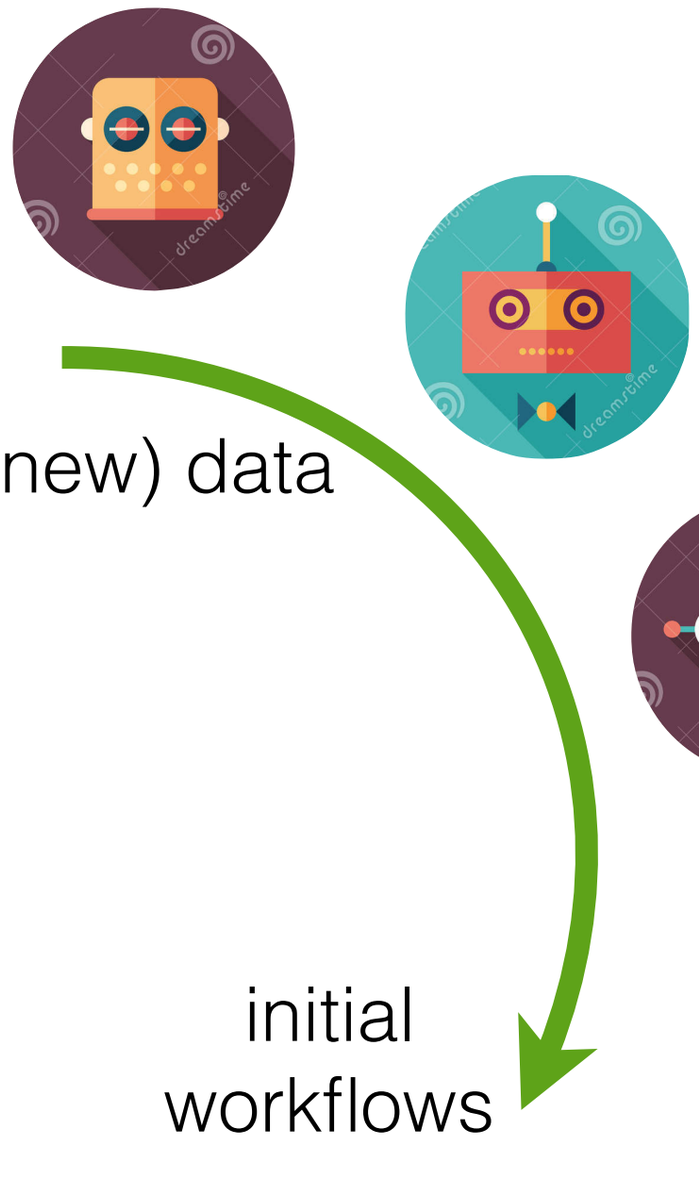
Feature selection bot: recommends/runs feature selection techniques

Imputation bot: recommends/runs missing value imputation techniques

Outlier detection bot: recommends/runs outlier detection techniques

Robot assistants

Random Bot: runs random search given a hyperparameter space



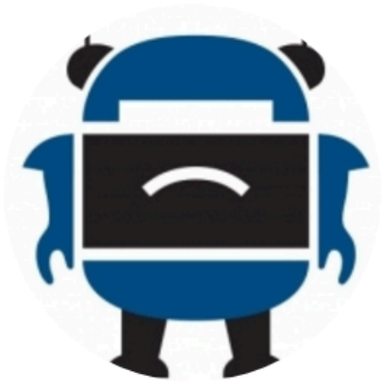
Greedy Bot: learns key algorithms, hyper-parameters, ranges. Tries those first.

Optimization bots: runs advanced hyperparameter optimization

Workflow bot: build ML workflows, in collaboration with other bots

Random Bot running on OpenML

People Search



Random Bot

Joined 2016-03-16

Activity	Reach	Impact	Uploads			
5716	17	0	0	17	0	1740

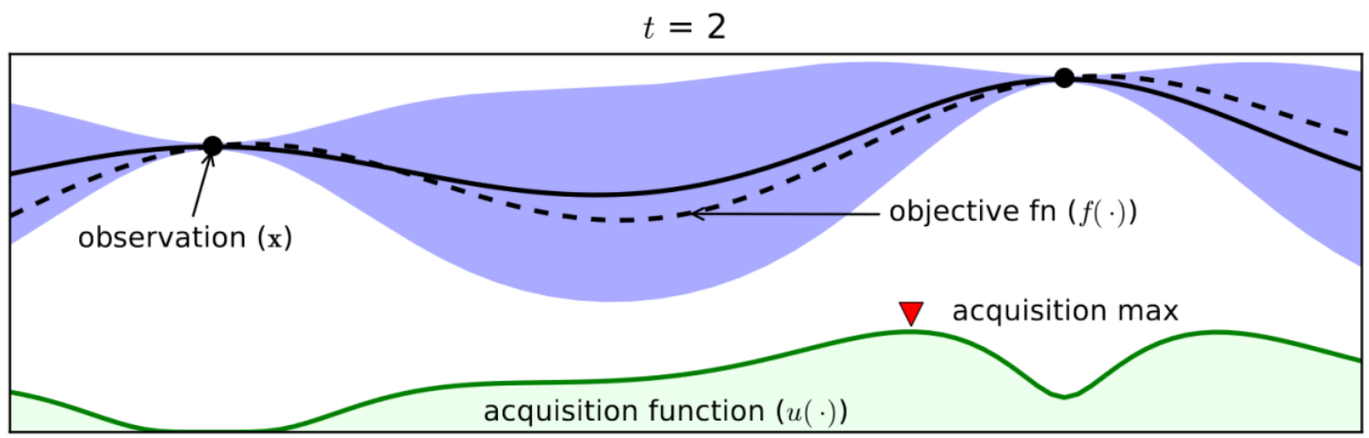
EDIT PROFILE

	Activity	Reach		Impact
Data Sets	0	0	0	0
Flows	17	0	0	0
Tasks	0	0	0	
Runs	1740	0	0	

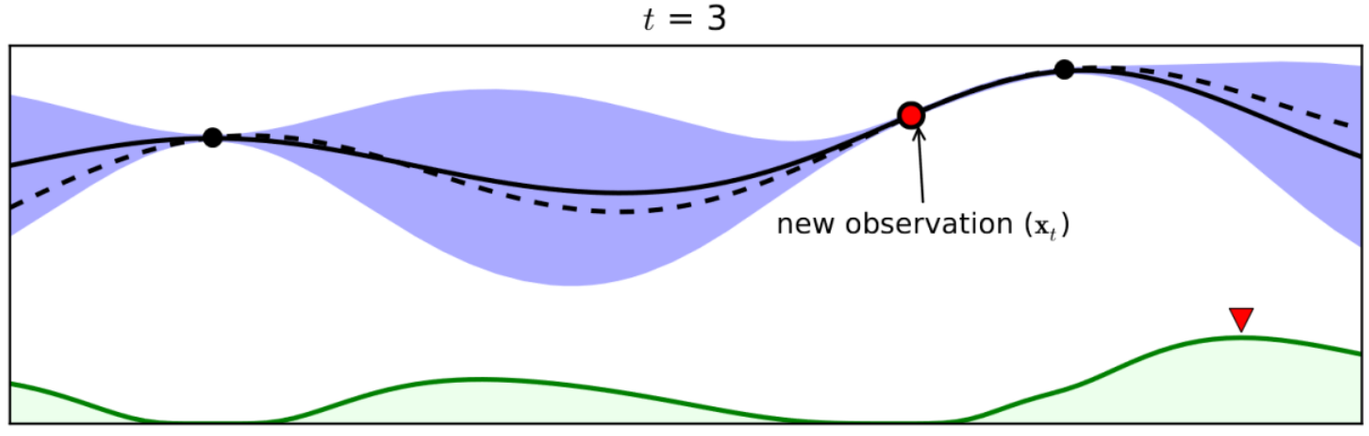
Bayesian optimization + metalearning

Include meta-data from prior datasets

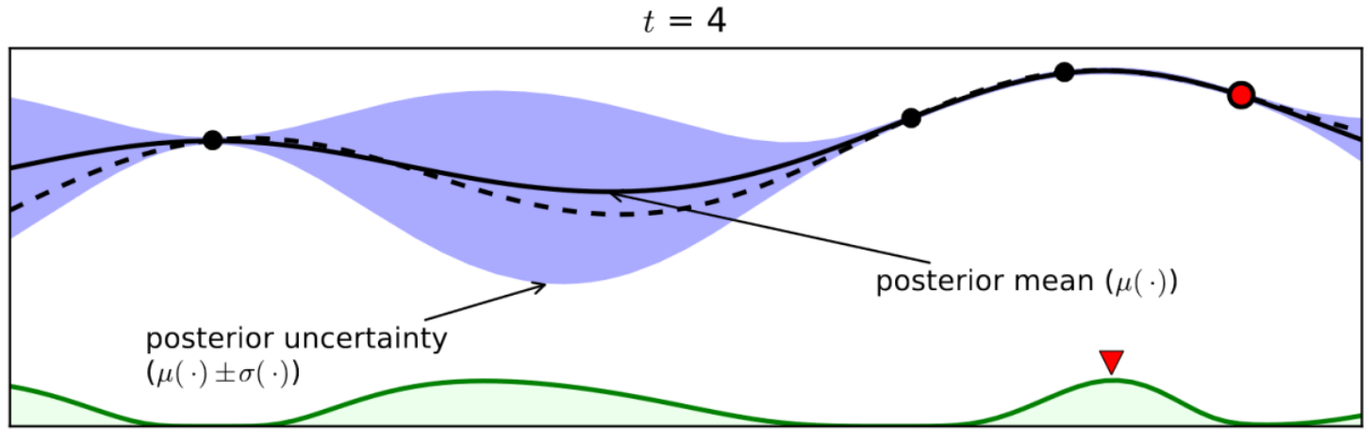
- Warm start: initialize search with promising configurations (*AutoML challenge winner*)



- Surrogate models with prior (focus on best parameters, ranges)

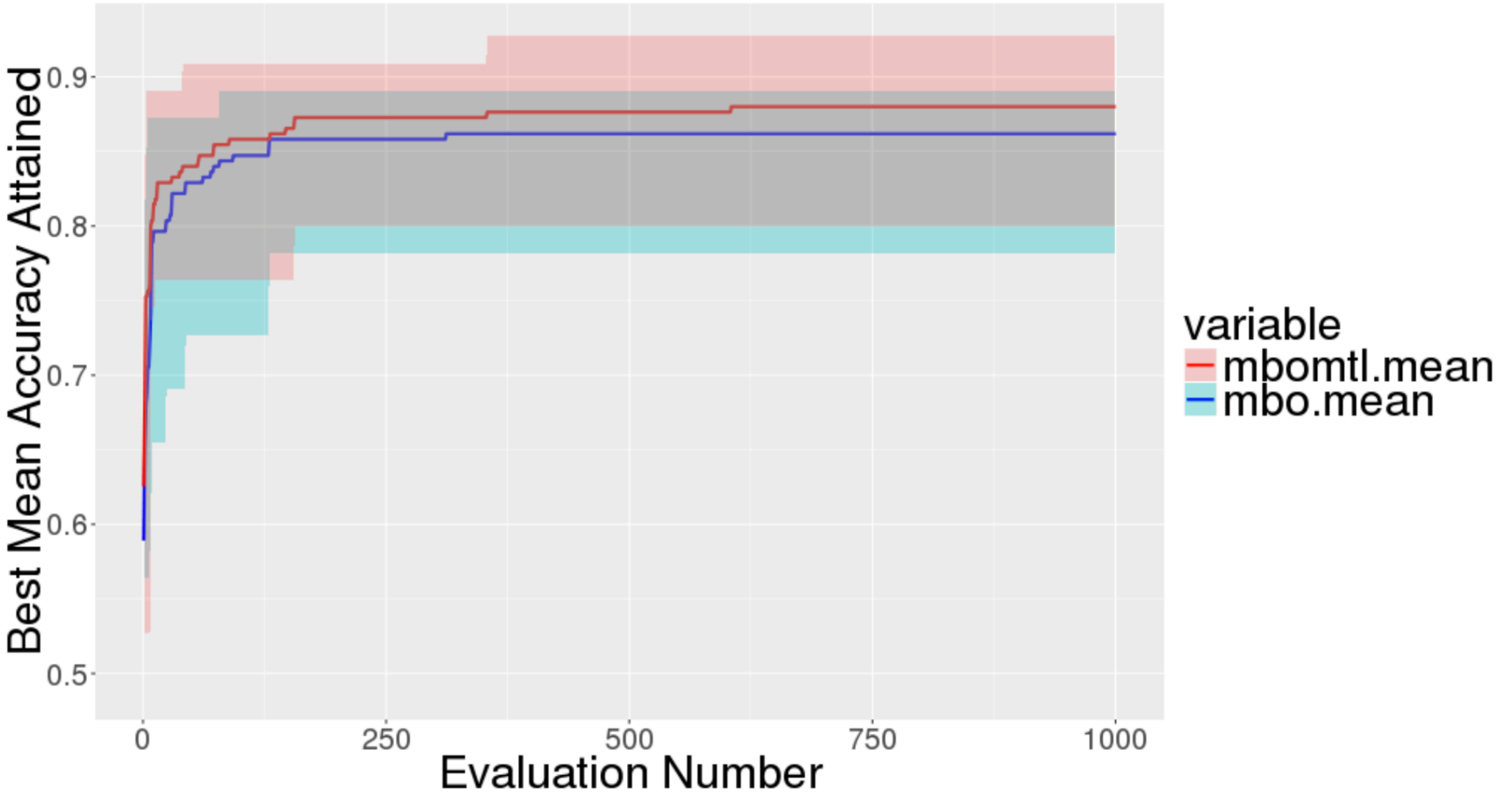


- Acquisition functions based on meta-models (predict performance, trained on prior datasets)

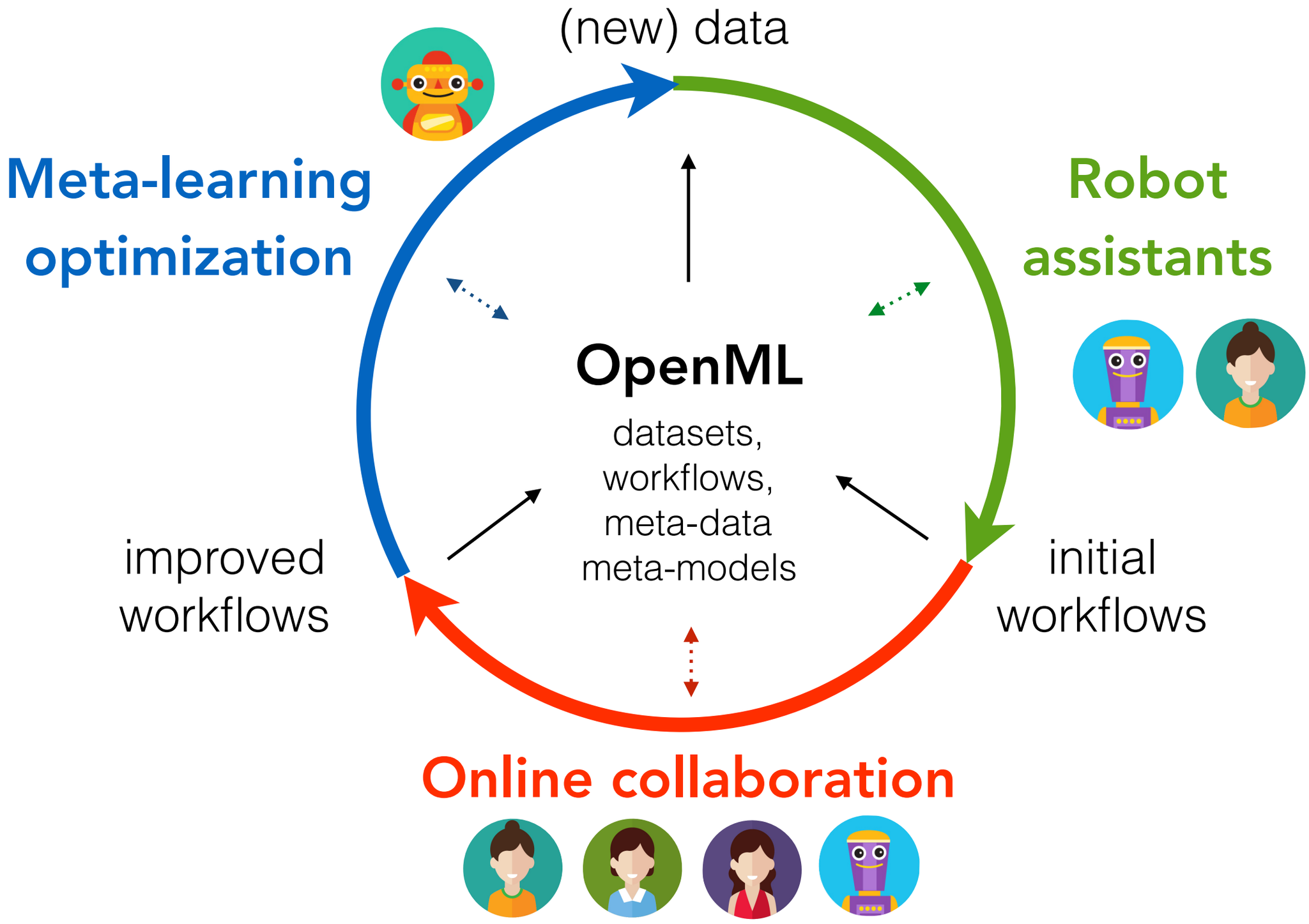


Combine meta-learning and optimization

- Acquisition functions based on meta-models



Learn from results of humans and robots, use that to build better bots to help humans




WHAT IF WE CAN EXPLORE DATA
COLLABORATIVELY
ON WEB SCALE
IN REAL TIME

Join Us!

www.openml.org

Join our hackathons

@open_ml

OpenML

Thank You