OpenML

DEMOCRATIZING AND AUTOMATING MACHINE LEARNING

JOAQUIN VANSCHOREN, TU EINDHOVEN, 2017

@joavanschoren

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

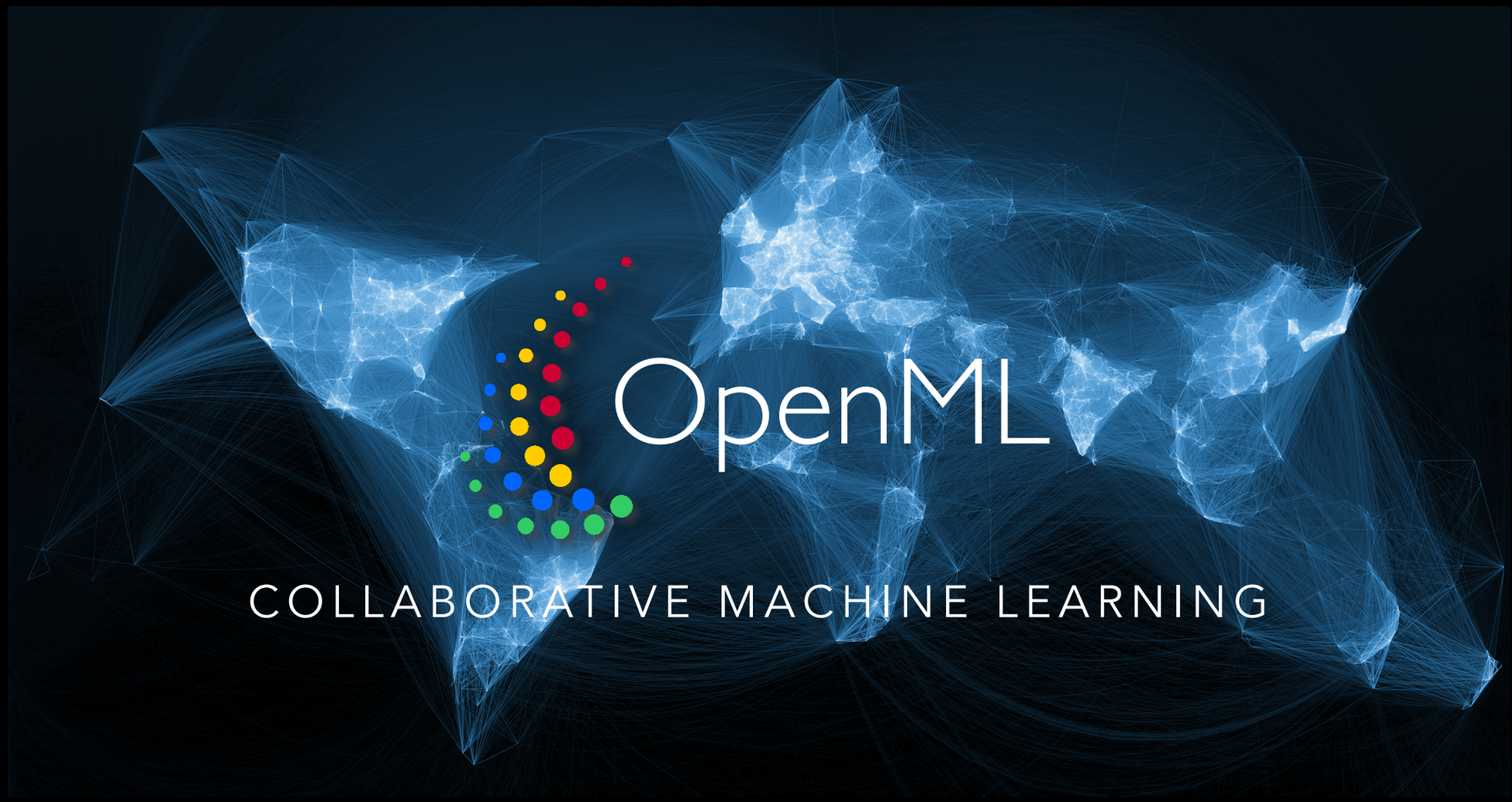
COLLABORATIVELY

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

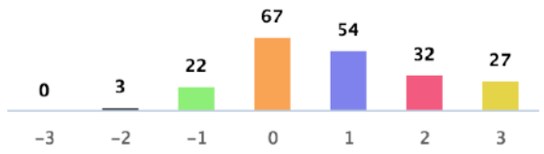
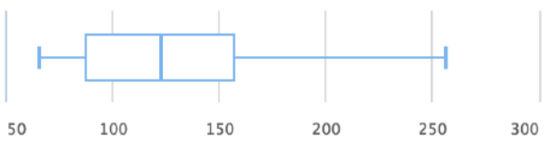
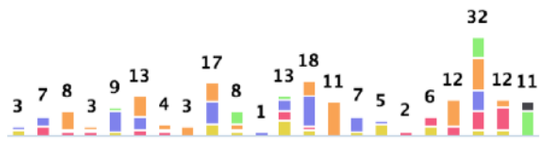
It starts with data

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

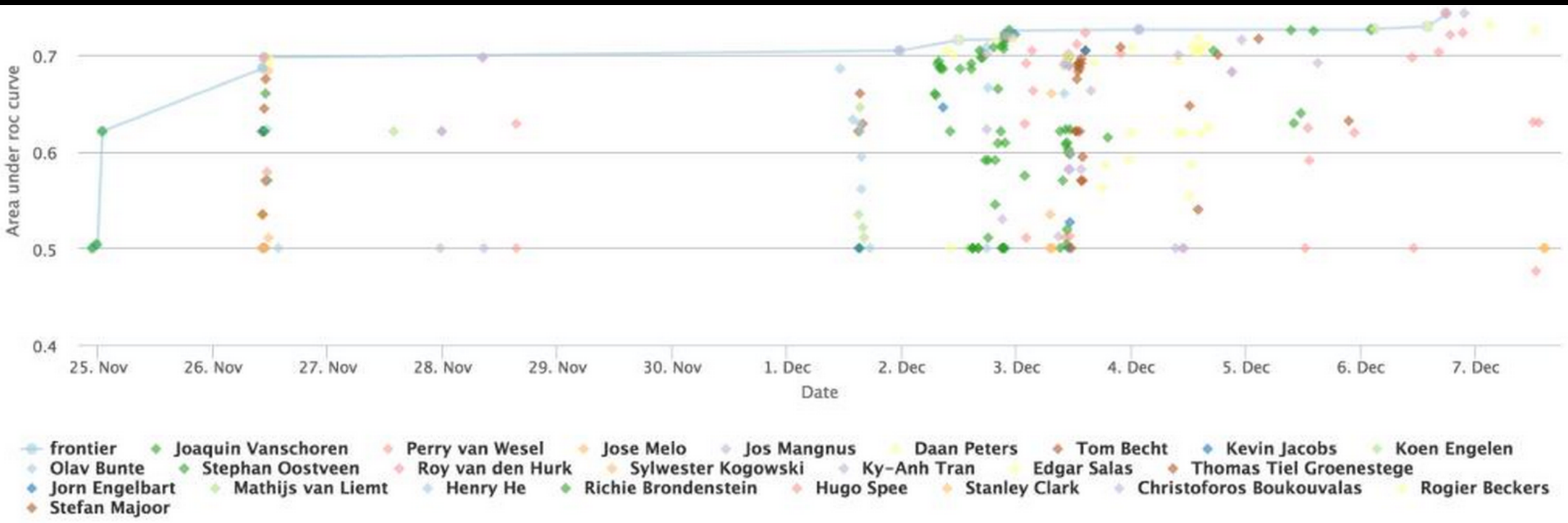
Set your goals, find help

Tasks contain data, goals, procedures.

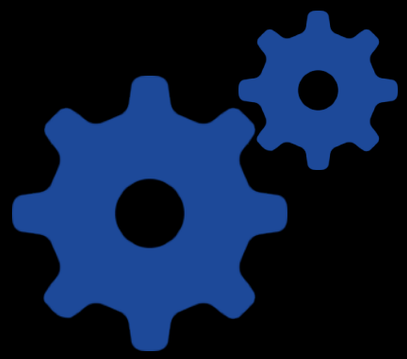
Readable by tools, automates experimentation

All results organized online: **realtime overview**

All results organized online: **realtime overview**

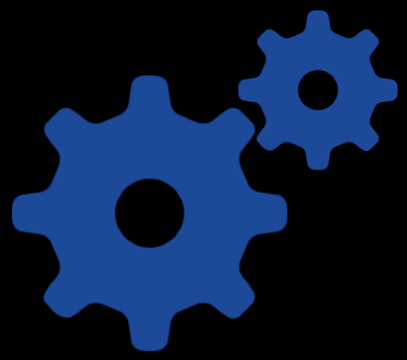


Explore all possible solutions

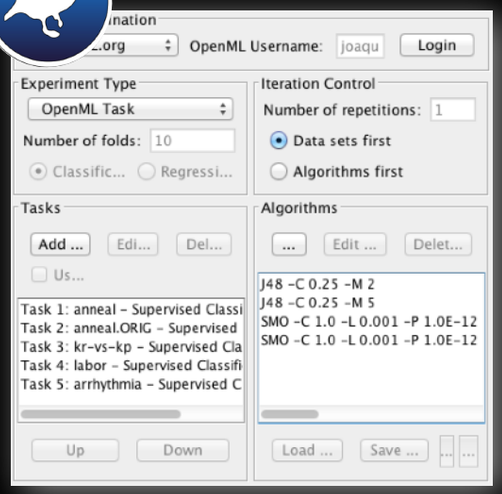


Flows (workflows, scripts) can run anywhere (locally)

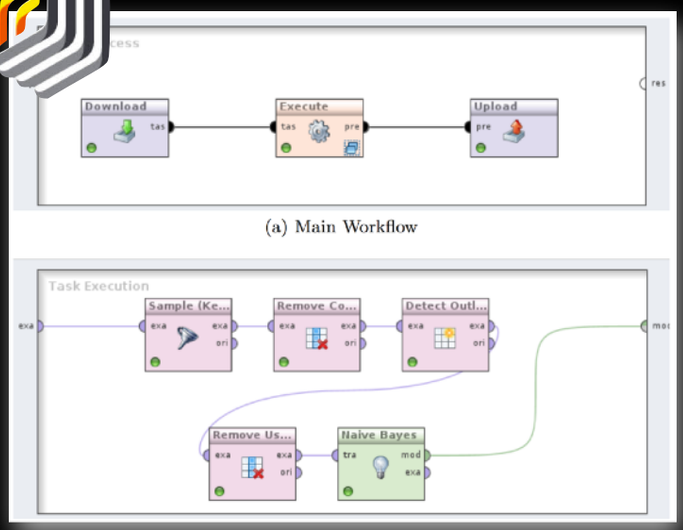
Tool integrations + APIs (REST, R, Python, Java,…)



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

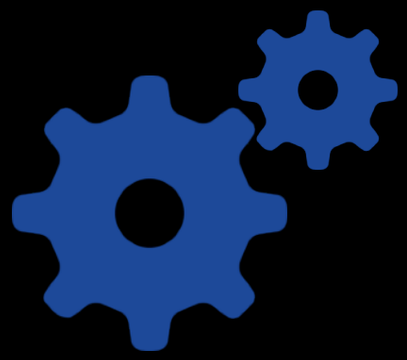


```python
from sklearn import tree
from openml import tasks,runs
task = tasks.get_task(14951)
clf = tree.DecisionTreeClassifier()
run = runs.run_task(task, clf)
return_code, response = run.publish()
```



(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)

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

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.

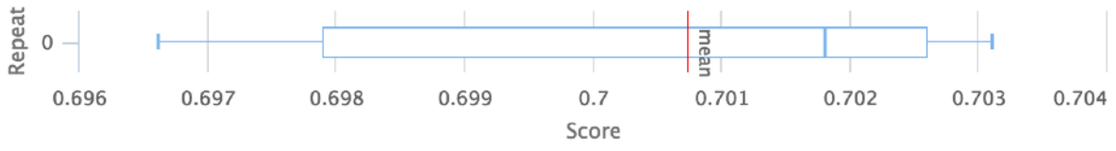
Area under ROC curve

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

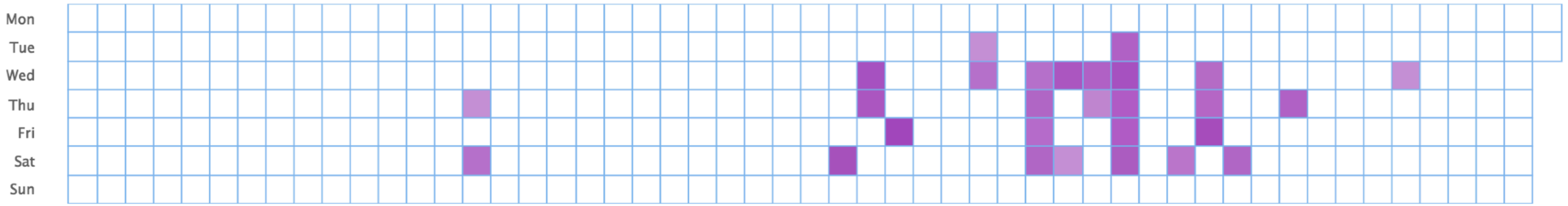
Activity	Reach	Impact	Uploads			
1649.5	12	195	1	35	0	1605

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**

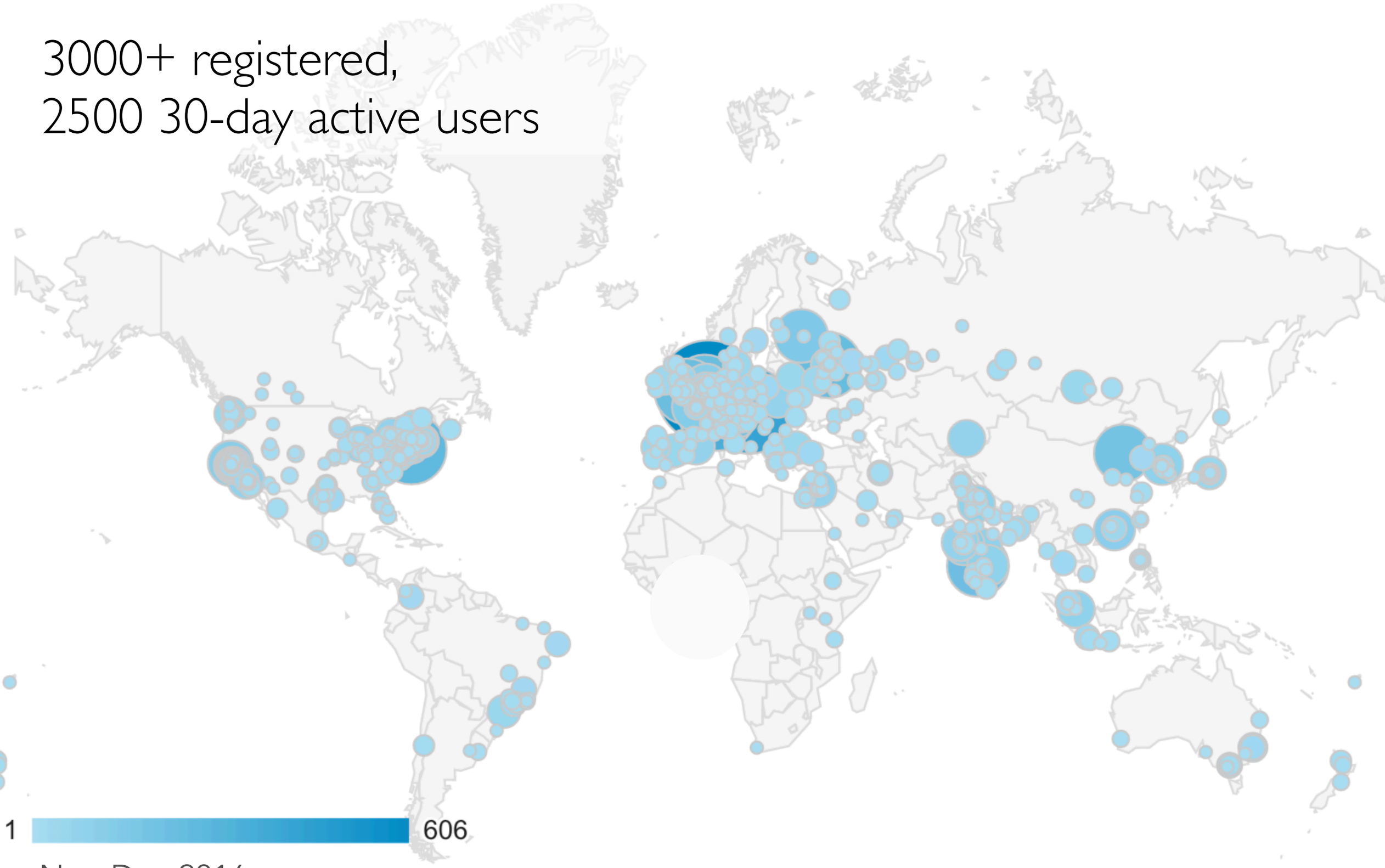
1698629 runs

Upload and explore all **results** online.

Download and share data, flows and runs through:

OpenML Community

3000+ registered,
2500 30-day active users



Nov-Dec 2016

Collaboration tools (in progress)

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

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

(Classroom) challenges
Open/closed, people can learn from each other

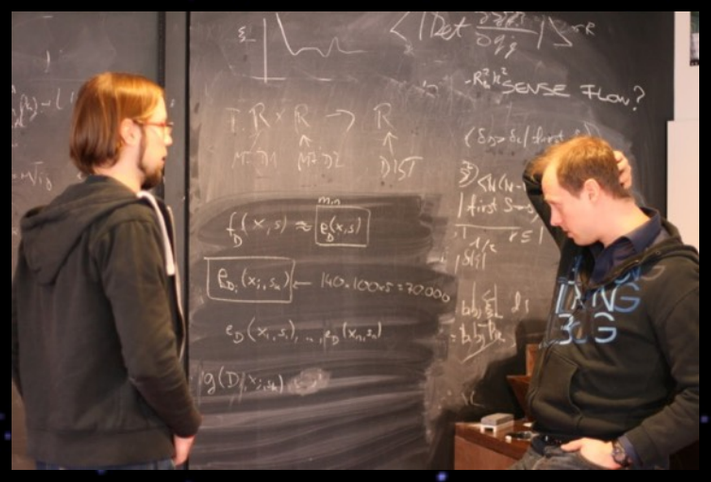
Code submissions
Sharing versioned code, docker images, archiving

Join Us!

www.openml.org

Next hackathon:

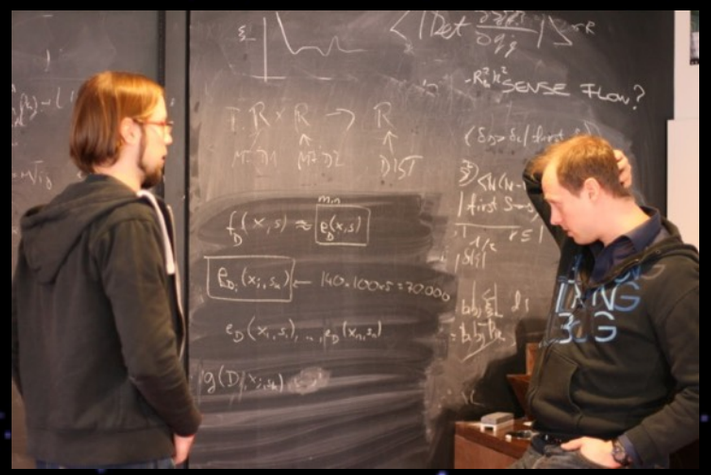
- October 9-13
- Lorentz Center, Leiden



Help us :)

We are always looking for:

- Code contributions (open source)
  - New tool/platform integration
  - New bots
  - Your own ideas
- Interesting datasets
- Computing resources
- Funding ideas



@open_ml

OpenML

Thank You

Now try it yourself :)