OpenML

TAKING MACHINE LEARNING RESEARCH ONLINE

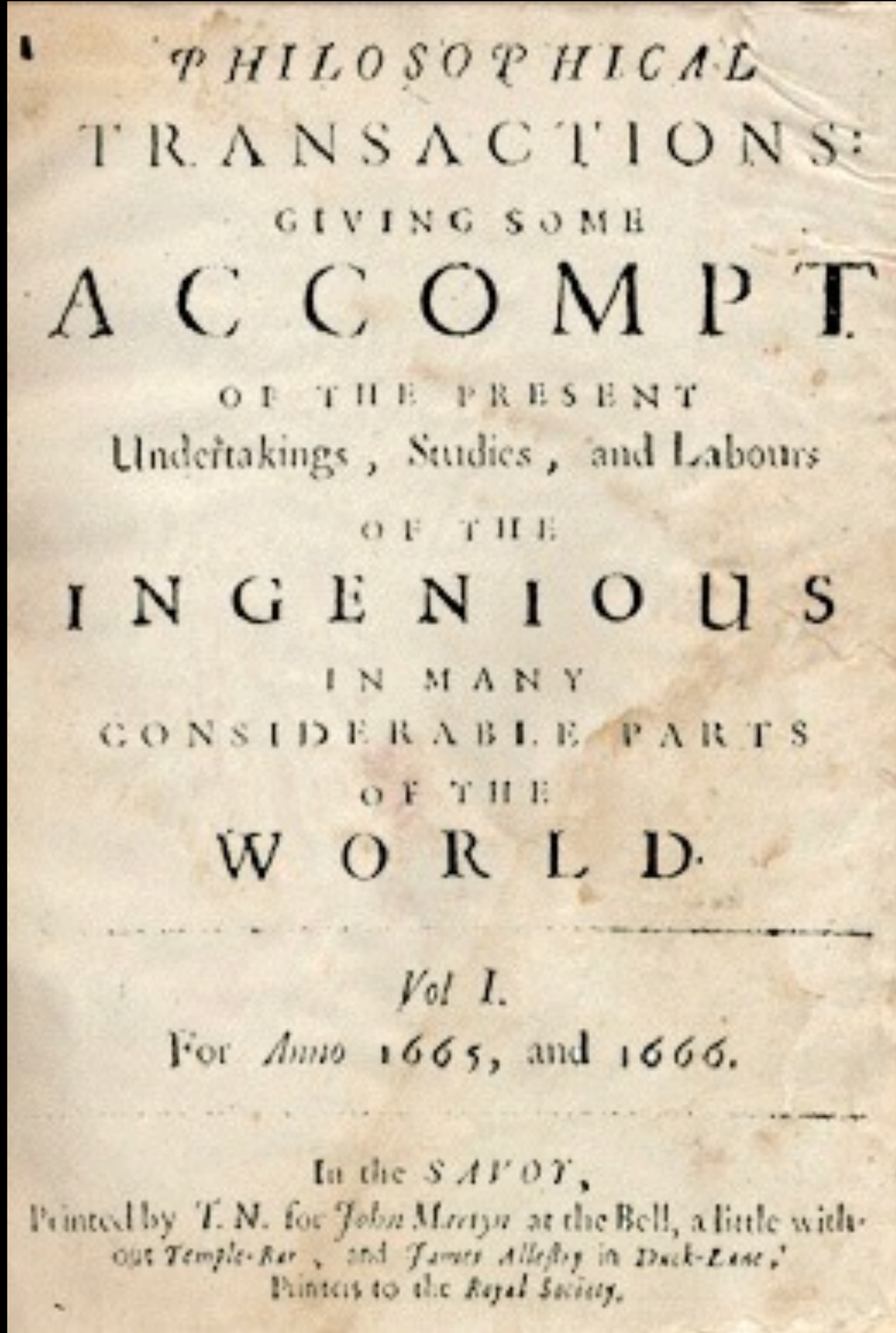
Joaquin Vanschoren (TU/e) 2015

AFTER 300 YEARS

IS PRINTING PRESS STILL THE BEST MEDIUM?

FOR MACHINE LEARNING?

- Code, data too complex (published separately)
- Experiment details scant
- Results unactionable, hard to reproduce, reuse
- Papers not updatable
- Slow, limited impact tracking
- Publication bias



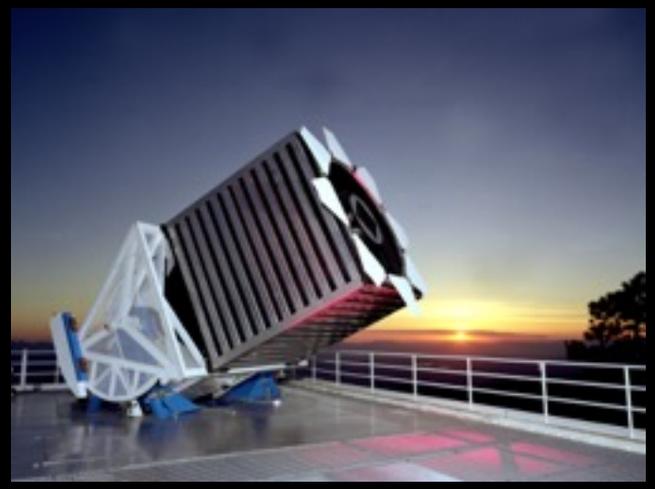
PHILOSOPHICAL
TRANSACTIONS:
GIVING SOME
ACCOMPT
OF THE PRESENT
Undertakings, Studies, and Labours
OF THE
INGENIOUS
IN MANY
CONSIDERABLE PARTS
OF THE
WORLD.

Vol I.
For Anno 1665, and 1666.

In the SAVOY,
Printed by T. N. for John Martyn at the Bell, a little without Temple-Bar, and James Allestry in Duck-Lane, Printers to the Royal Society.

NETWORKED SCIENCE

Polymaths: Mathematicians solved centuries-old problems within weeks by collaborating openly online



SDSS: Thousands of astronomical papers published on organised, online data from a single telescope

Galaxy Zoo: Amateur astronomers make new discoveries by looking through thousands of images

Why? Designed serendipity

Broadcasting data fosters spontaneous, unexpected discoveries

What's hard for one scientist is easy for another: connect minds

How? Remove friction

Organized body of compatible scientific data (and tools) online

Micro-contributions: seconds, not days

Easy, organised communication

Track who did what, give credit

FRICTION-LESS ENVIRONMENT FOR
MACHINE LEARNING RESEARCH

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

Easy to use: Automated download/upload within your ML environment

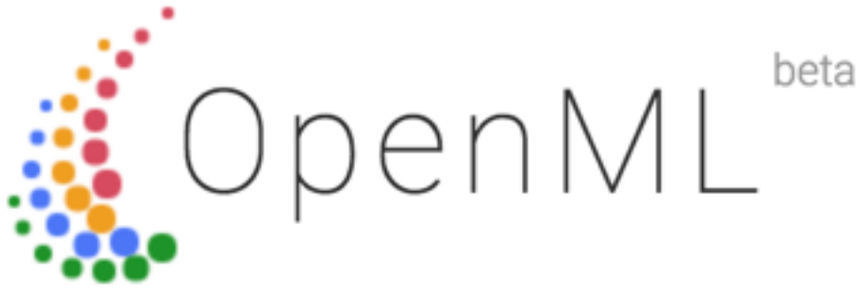
Micro-contributions: Upload single dataset, algorithm, experiment

Easy communication: Online discussions per dataset, algorithm, experiment

Reputation: Auto-tracking of downloads, reuse, likes.

Real time: Share and reuse instantly, openly or in circles of trusted people

OpenML Search



Exploring machine learning better, together

1345 data sets

Find or add **data** to analyse

4830 tasks

Download or create scientific **tasks**

1400 flows

Find or add data analysis **flows**

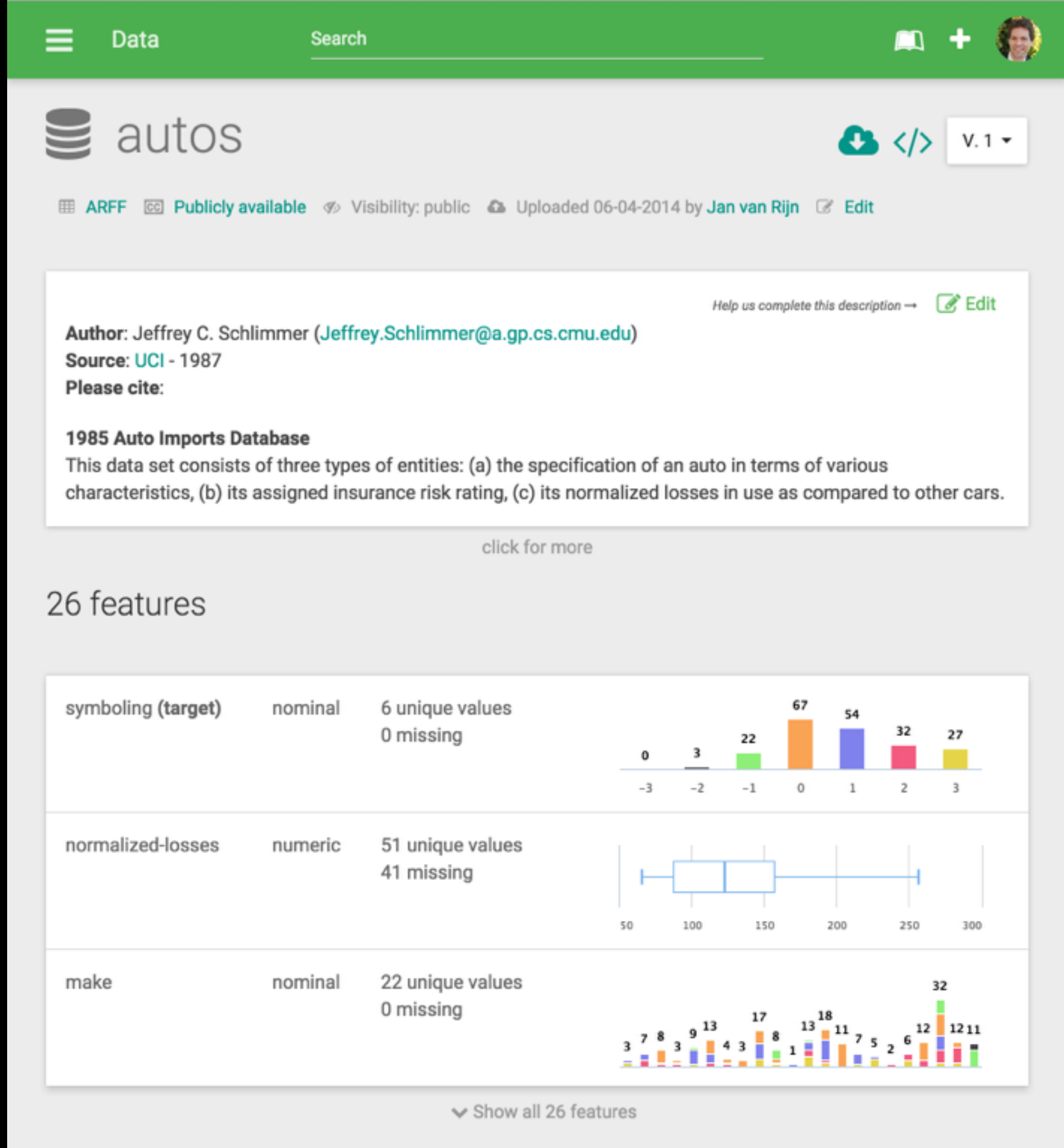
452379 runs

Upload and explore all **results** online.

Data from
various sources
analysed and
organised online
for easy access

Scientists **broadcast data** by uploading or linking from existing repos. OpenML will **automatically check and analyze the data**, compute characteristics, **annotate, version and index it for easy search**

- Search on keywords or properties
- Wiki-like descriptions
- Analysis and visualisation of features
- Auto-calculation of large range of meta-features



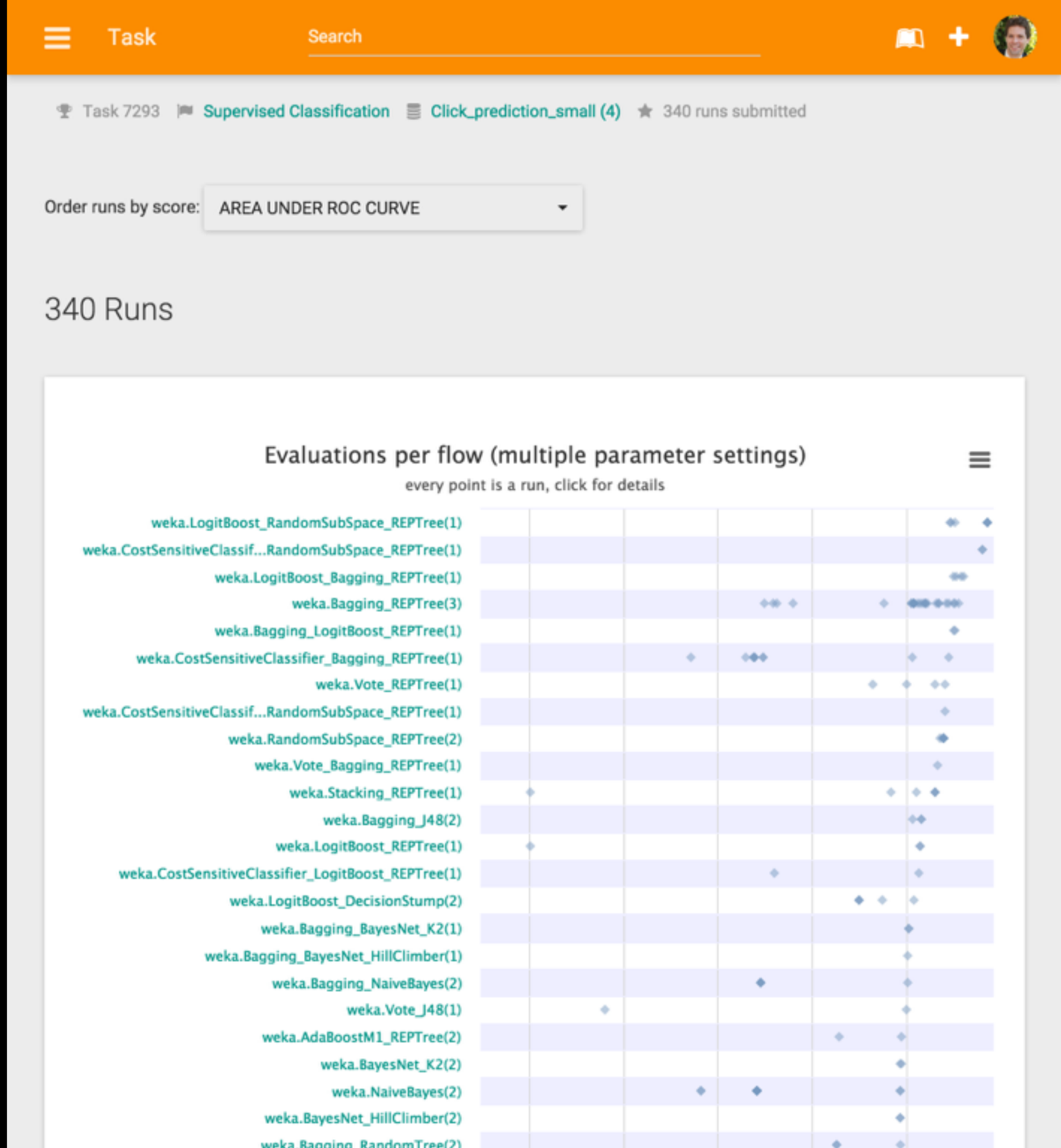
Scientific tasks that can be **interpreted by tools** and **solved collaboratively**

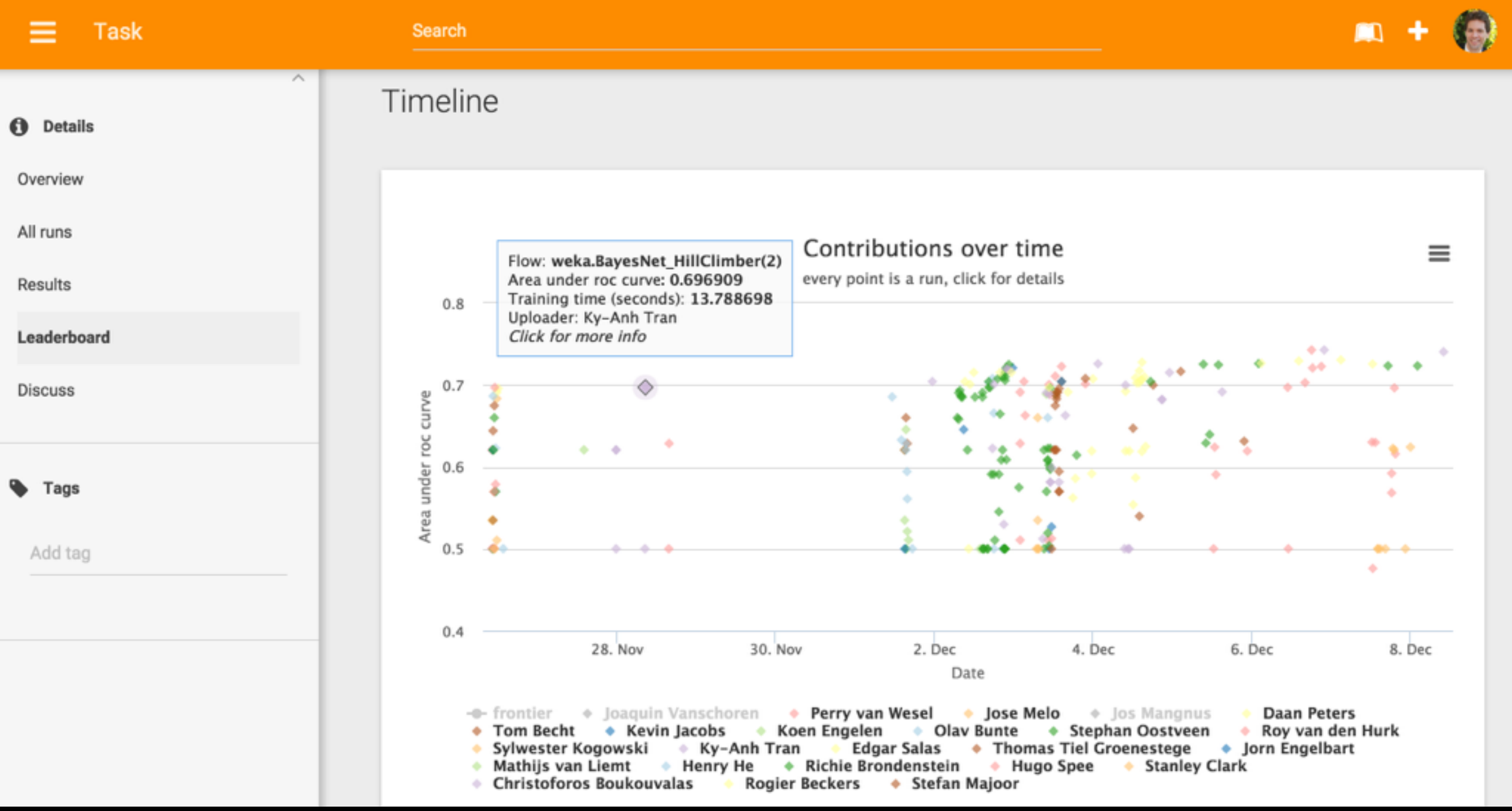
Tasks: containers with all data, goals, procedures.

Machine-readable: tools can **automatically download data,** use correct procedures, and **upload results**.

Creates **realtime, collaborative data mining challenges.**

- *Example: Classification on click prediction dataset, using 10-fold CV and AUC*
- People submit results (e.g. predictions)
- Server-side evaluation (many measures)
- All results organized online, per algorithm, parameter setting
- Online visualizations: every dot is a run plotted by score





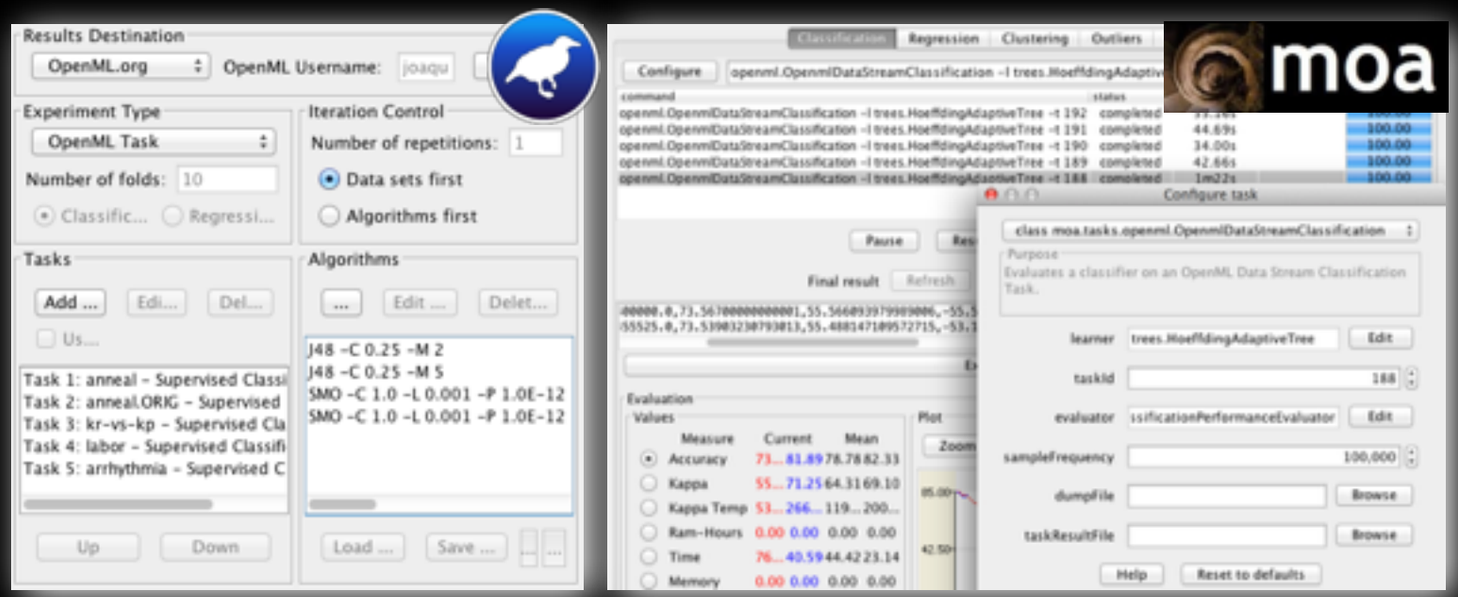
- Leaderboards visualize progress over time: who delivered breakthroughs when, who built on top of previous solutions
- Collaborative: all code and data available, learn from others, form teams
- Real-time: who submits first gets credit, others can improve immediately

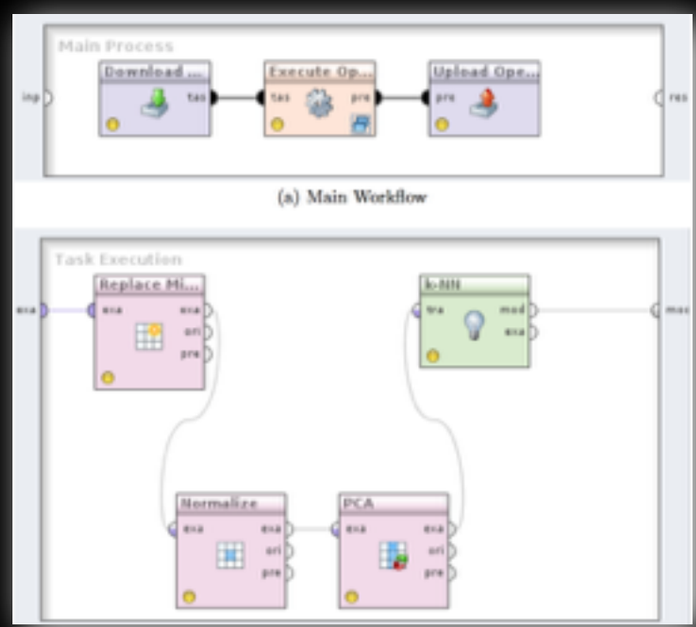
Machine learning flows (code) that can **solve tasks** and **report results.**

Flows: wrappers that read **tasks**, return required **results**.
Scientists upload code or link from existing repositories/libraries.
Tool integrations allow automated **data download, flow upload** and **experiment logging and sharing.**

REST API + Java, R, Python APIs

- WEKA/MOA plugins: automatically load tasks, export results



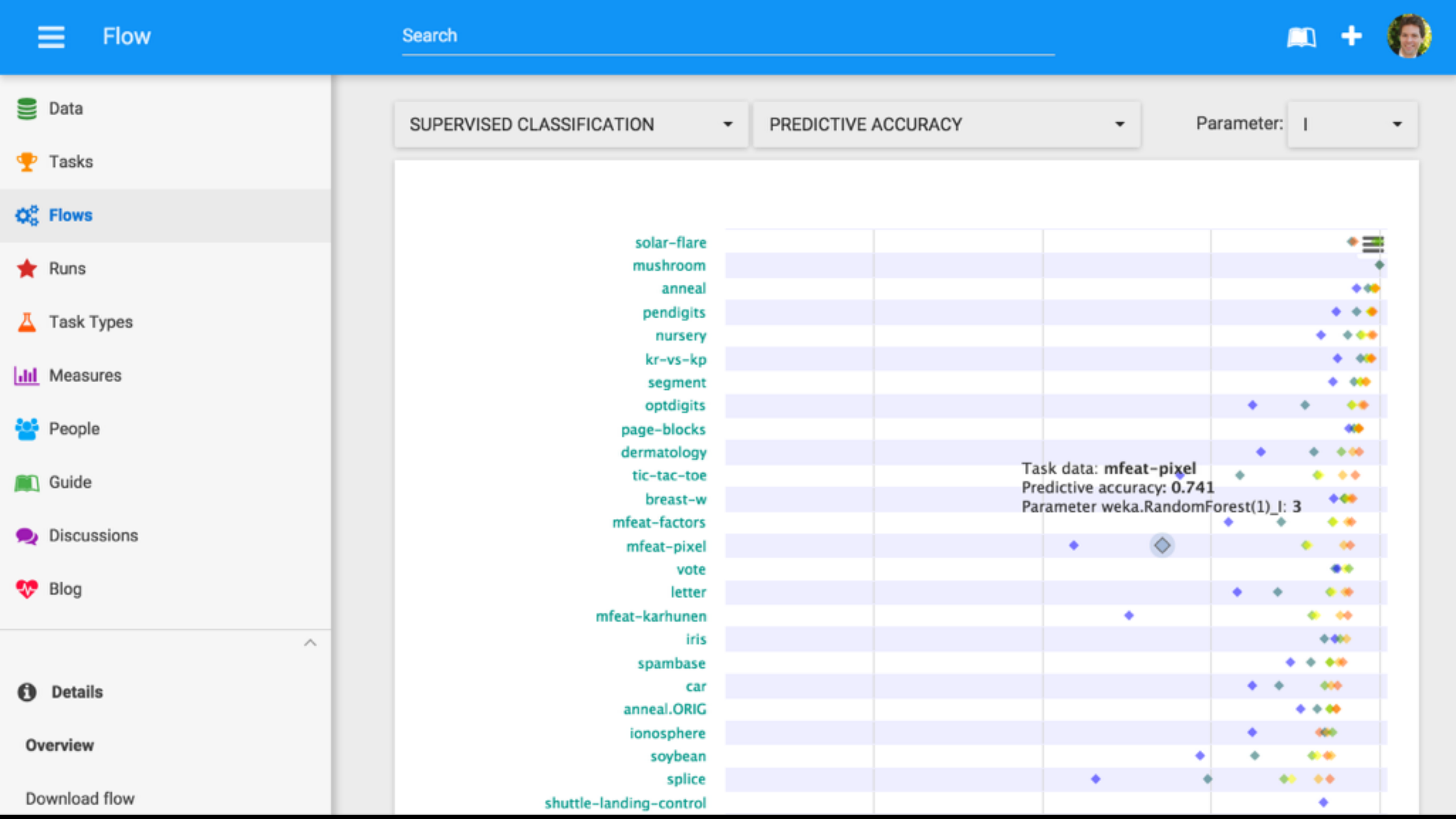


- RapidMiner plugin: new operators to load tasks, export results and subworkflow

- R/Python interfaces: functions to down/upload data, code, results in few lines of code

```
from openml.apiconnector import APIConnector
from sklearn import preprocessing, ensemble
connector = APIConnector(username=username, password=passw
dataset = connector.download_dataset(31)
X, y, categorical = dataset.get_pandas()
enc = preprocessing.OneHotEncoder(categorical_features=cat
X = enc.transform(X).todense()
clf = ensemble.RandomForestClassifier()
clf.fit(X, y)
```

```
library(OpenML); library(mlr)
authenticateUser(username = "user", passwo
task = getOMLTask(task.id = 1L)
lrn = makeLearner("classif.randomForest")
run.mlr = runTaskMlr(task, lrn)
run.id = uploadOMLRun(run.mlr)
```



- All results obtained with same flow organised online
- Results linked to data sets, parameter settings -> trends/comparisons
- Visualisations (dots are models, ranked by score, colored by parameters)

Experiments auto-uploaded, linked to **data, flows** and **authors**, and organised for easy reuse

Runs uploaded by **flows**, contain fully reproducible results for all **tasks**. OpenML **evaluates and organizes all results online** for **discovery, comparison and reuse**

- Detailed run info
- Author, data, flow, parameter settings, result files, …
- Evaluation details (e.g., results per sample)

Run

Run 84087

JSON XML

Task 7293 (Supervised Classification) Click_prediction_small Uploaded 01-01-2015 by Ky-Anh Tran

Flow

weka.Bagging_BayesNet_K2(1)	Leo Breiman (1996). Bagging predictors. Machir
weka.Bagging_BayesNet_K2(1)_P	100
weka.Bagging_BayesNet_K2(1)_S	1
weka.Bagging_BayesNet_K2(1)_num-slots	8

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

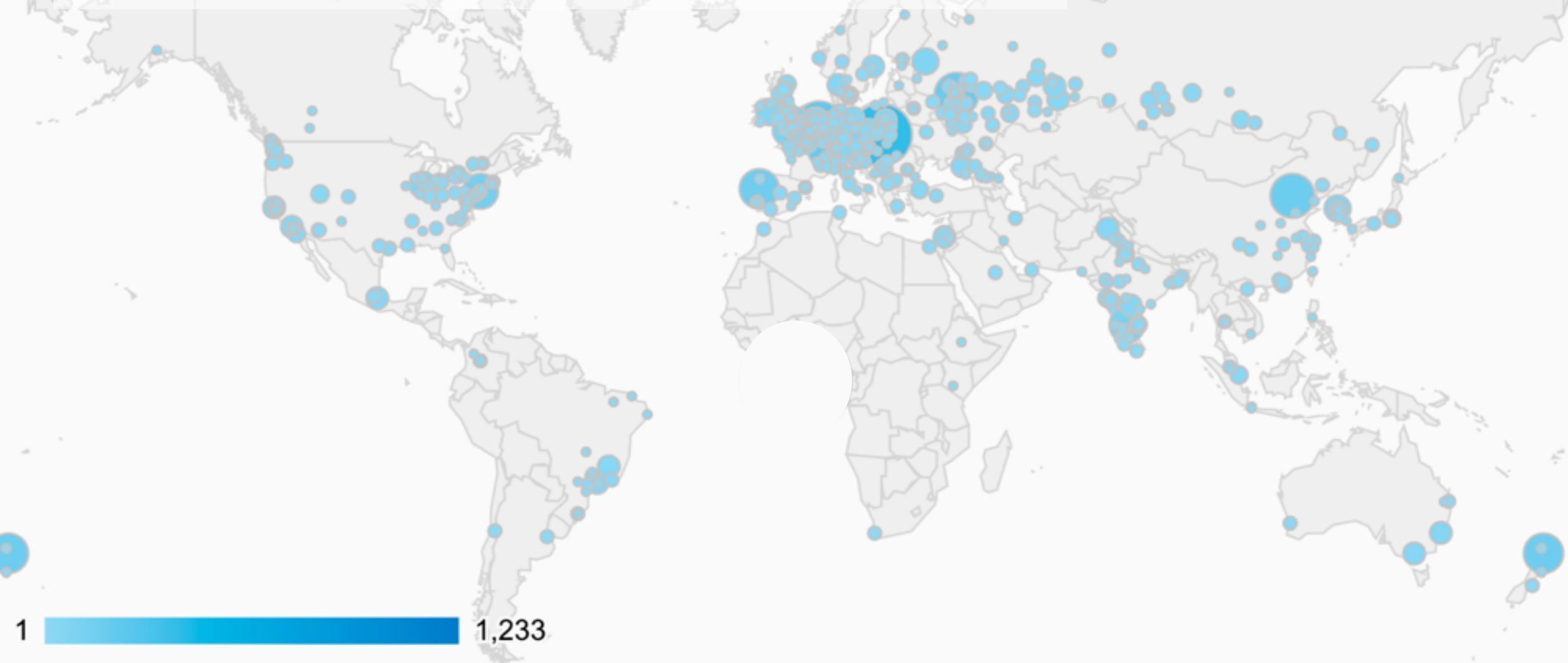
Per class

0	1
0.7007	0.7007

Cross-validation details (10-fold Crossvalidation)

OpenML Community

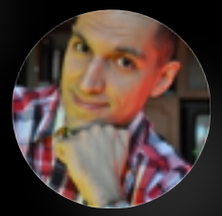
Used all over the world (and still in beta)
Great open source community of GitHub
450+ active users, many more passive ones
1000s of datasets, flows, 450000+ runs



Jan-Jun 2015

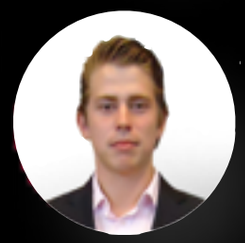
THANK YOU

#OpenML

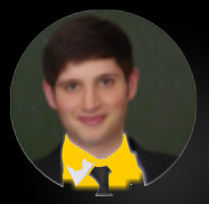


Nenad Tomašev

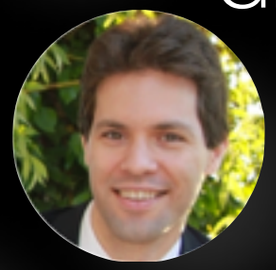
Luis Torgo



Jan van Rijn

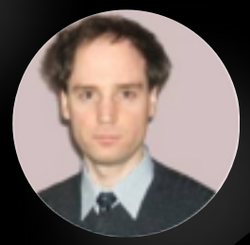


Giuseppe Casalicchio



Joaquin Vanschoren

Michel Lang



Bernd Bischl

Matthias Feurer

You? Please join us :)

Things we're working on

Circles

Create collaborations with trusted researchers

Share results within team prior to publication

Projects (e-papers)

- Online counterpart of a paper, linkable
- Merge data, code, experiments (new or old)
- Public or shared within circle

Altmetrics

- Measure real impact of your work
- Reuse, downloads, likes of data, code, projects,…
- Online reputation (more sharing)

Things we're working on (please join)

Distributed computing

- Create jobs online, run anywhere you want
- Locally, clusters, clouds

Algorithm selection, hyperparameter tuning

- Upload dataset, system recommends techniques
- Model-based optimisation techniques
- Continuous improvement (learns from past)

Things we're working on (please join)

Data repository connections

- Wonderful open data repo's (e.g. rOpenSci)
- More data formats, data set analysis

Algorithm/code connections

- Improved API's (R,Java,Python,CLI,…)
- Your favourite tool integrated

Statistical analysis

- Proper significance testing in comparisons
- Recommend evaluation techniques (e.g. CV)

Online task creation

- Definition of scientific tasks
- Freeform tasks or server-side support