

TAKING MACHINE LEARNING RESEARCH ONLINE

Joaquin Vanschoren (TU/e) 2015

AFTER 300 YEARS IS PRINTING PRESS STILL THE BEST MEDIUM? FOR MACHINE LEARNING?

- Code too complex (online)
- Data sets too large (online)
- Experiment details scant
- Results unactionable, hard to reproduce, reuse
- Papers not updatable
- Slow, limited impact tracking
- Publication bias
- No online public discussion

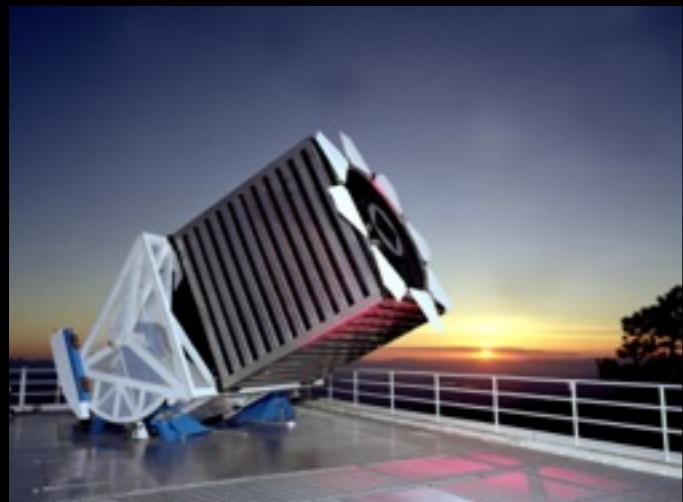
PHILOSOPHICAL
TRANSACTIONS:
GIVING SOME
ACCOMP'T
OF THE PRESENT
Undertakings, Studies, and Labours
OF THE
INGENIOUS
IN MANY
CONSIDERABLE PARTS
OF THE
WORLD.

Vol I.
For Anno 1665, and 1666.

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

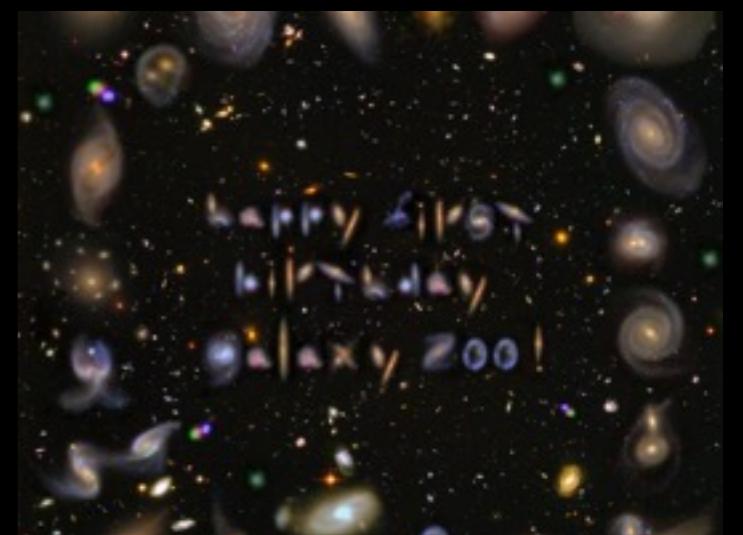
NETWORKED SCIENCE

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

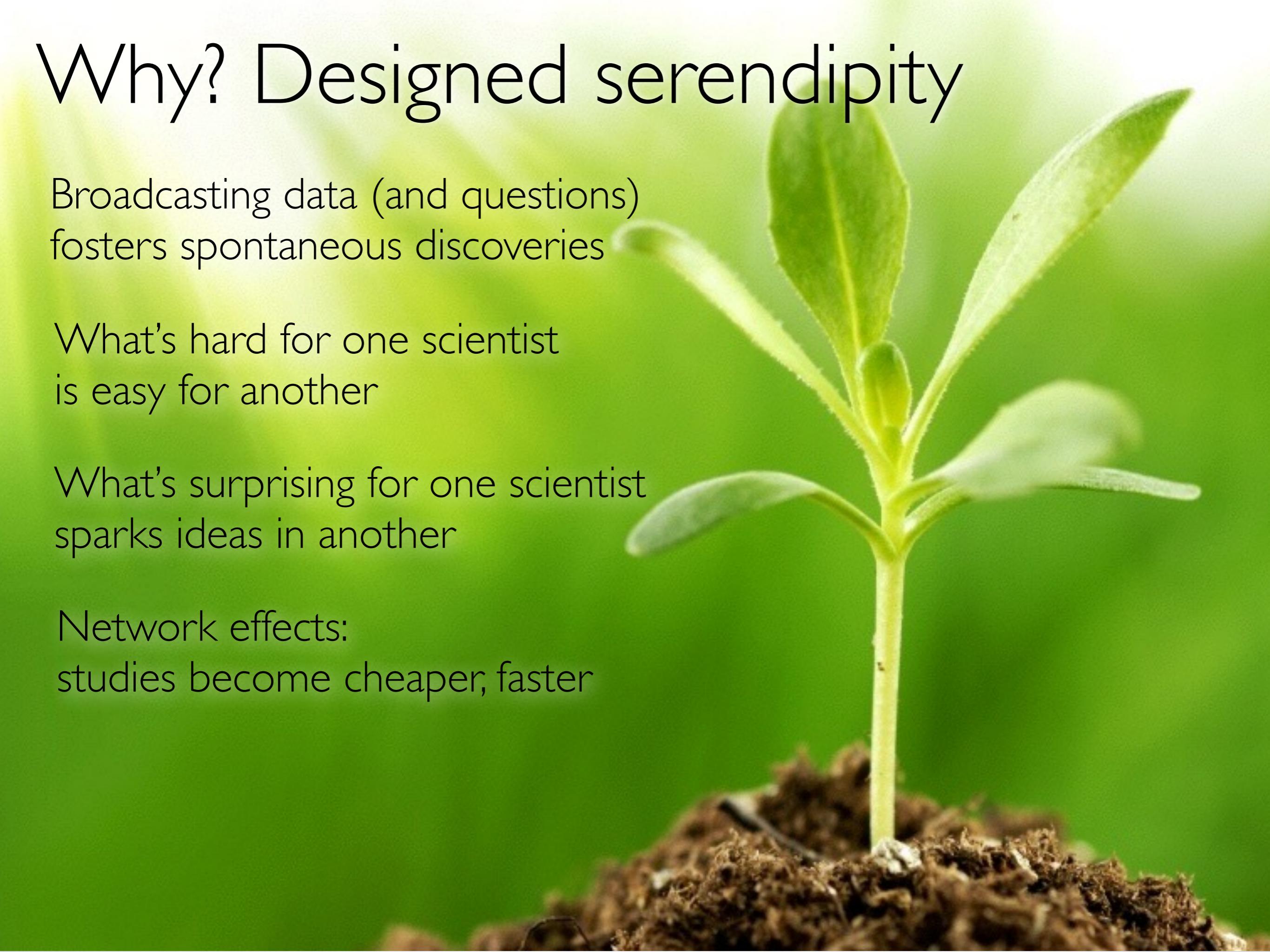


SDSS: Thousands of astronomical papers published on organised, online data from one single telescope

Galaxy Zoo: Amateur astronomers made new discoveries by simply looking through thousands of images of galaxies



Why? Designed serendipity

A close-up photograph of a young plant seedling with several green leaves sprouting from a thin stem. The plant is growing out of a mound of dark brown, textured soil. The background is a solid, vibrant green color.

Broadcasting data (and questions)
fosters spontaneous discoveries

What's hard for one scientist
is easy for another

What's surprising for one scientist
sparks ideas in another

Network effects:
studies become cheaper, faster

Only works of you remove friction

Organised body of compatible
scientific data (and tools)

Micro-contributions: seconds, not days

Easy, organised communication

Measure impact (altmetrics)
Give credit when credit is due





OpenML

A REALTIME, WORLDWIDE LAB



FRICTION-LESS ENVIRONMENT FOR MACHINE LEARNING RESEARCH

Organized: Easy to find datasets, code, experiments. All connected, uniform.
Linked to people. Reproducible.

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

Micro-contributions: Upload single dataset, algorithm, experiment

Easy communication: Online discussions per resource (+github)

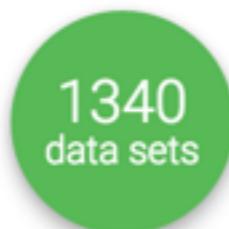
Reputation: Auto-tracking of downloads, reuse, likes.

Autonomy: Work openly or in circles of trusted people (preliminary work)



OpenML^{beta}

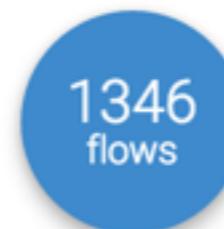
Exploring machine learning better, together



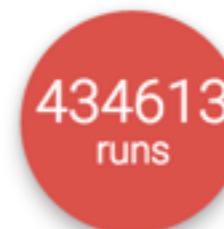
Find or add **data** to analyse



Download or create scientific **tasks**



Find or add data analysis **flows**



Upload and explore all **results** online.



Data from
various sources
**analysed and
organised online**
for easy access

Scientists can **broadcast data** by uploading or linking from existing repositories. OpenML will (for known data formats) **automatically analyze the data**, compute data characteristics, **annotate, version and index it for easy search**

- Search datasets by keywords or properties
- Filters
- Tagging
- Through website or API

Data

Search

Filter results

Number of instances

Number of features

Number of missing values

Number of classes

Default accuracy

Uploader

Tag

SEARCH

i You can use 1..10, >10, ...

Remove all filters

1317 results

 iris (1)	This is perhaps the best known 3816 runs - 150 instances - 5 features
 credit-a (1)	1. Title: Credit Approval 2. Source: 2874 runs - 690 instances - 16 features
 anneal.ORIG (1)	1. Title of Database: Annealing 2613 runs - 898 instances - 39 features
 diabetes (1)	1. Title: Pima Indians Diabetes 2606 runs - 768 instances - 9 features
 colic (1)	Donor: Will Taylor (taylor@pluto) 2451 runs - 368 instances - 28 features
 anneal (2)	This is a preprocessed version of 2434 runs - 898 instances - 39 features
 mfeat-zernike (1)	The multi-feature digit dataset - 2321 runs - 2000 instances - 48 features
 mfeat-morphological (1)	The multi-feature digit dataset - 2317 runs - 2000 instances - 7 features
 solar-flare (2)	1. Title: Solar Flare database The 2254 runs - 1066 instances - 13 features

- Wiki-like descriptions (add what you know about it)
- Analysis and visualisation of features to spot problems
- Auto-calculation of large range of meta-features

Data Search ☰ + 

autos

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

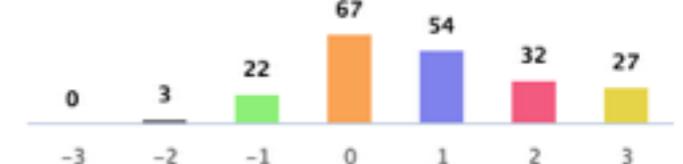
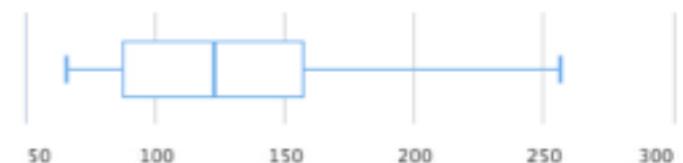
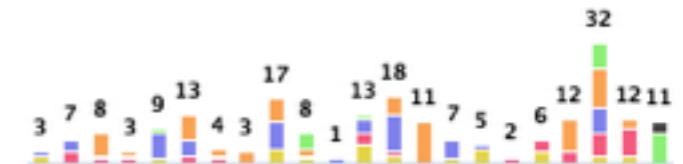
Help us complete this description → Edit

Author: Jeffrey C. Schlimmer (Jeffrey.Schlimmer@a.gp.cs.cmu.edu)
Source: UCI - 1987
Please cite:

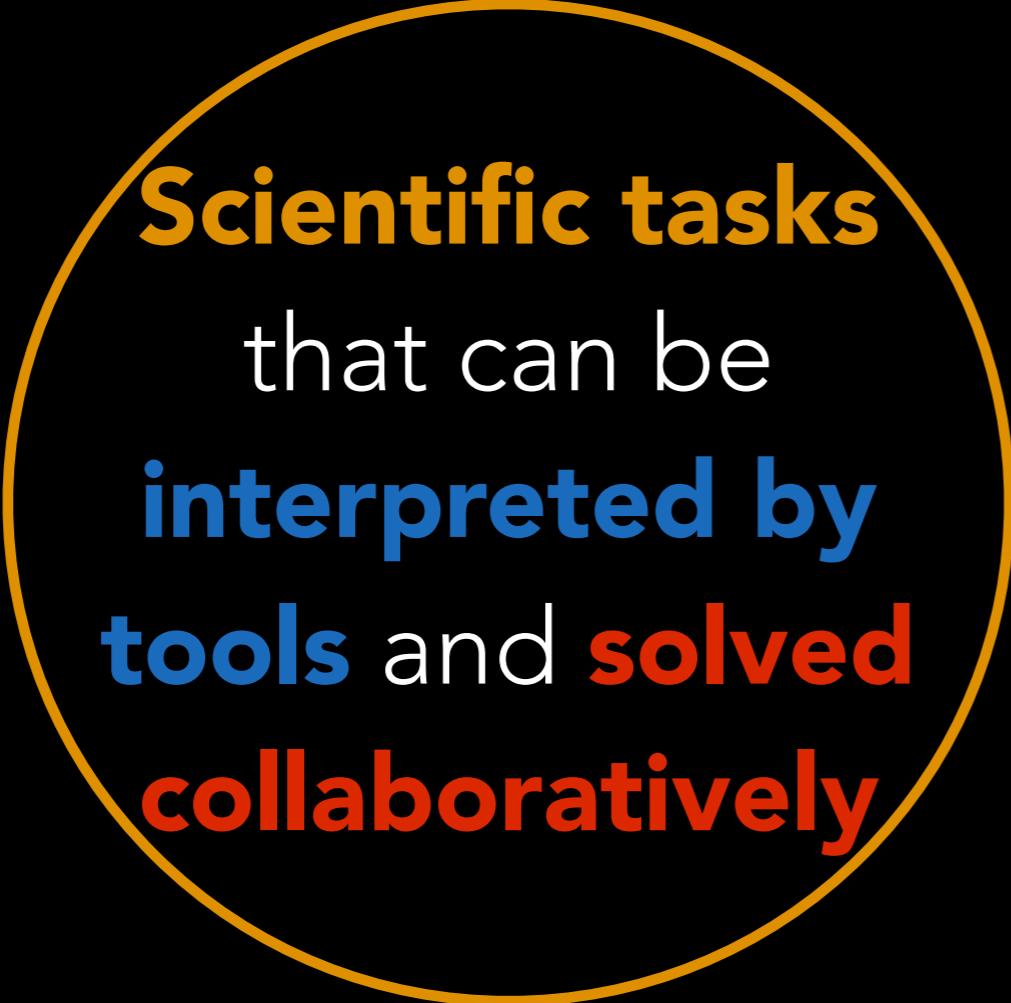
1985 Auto Imports Database
This data set consists of three types of entities: (a) the specification of an auto in terms of various characteristics, (b) its assigned insurance risk rating, (c) its normalized losses in use as compared to other cars.

[click for more](#)

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](#)

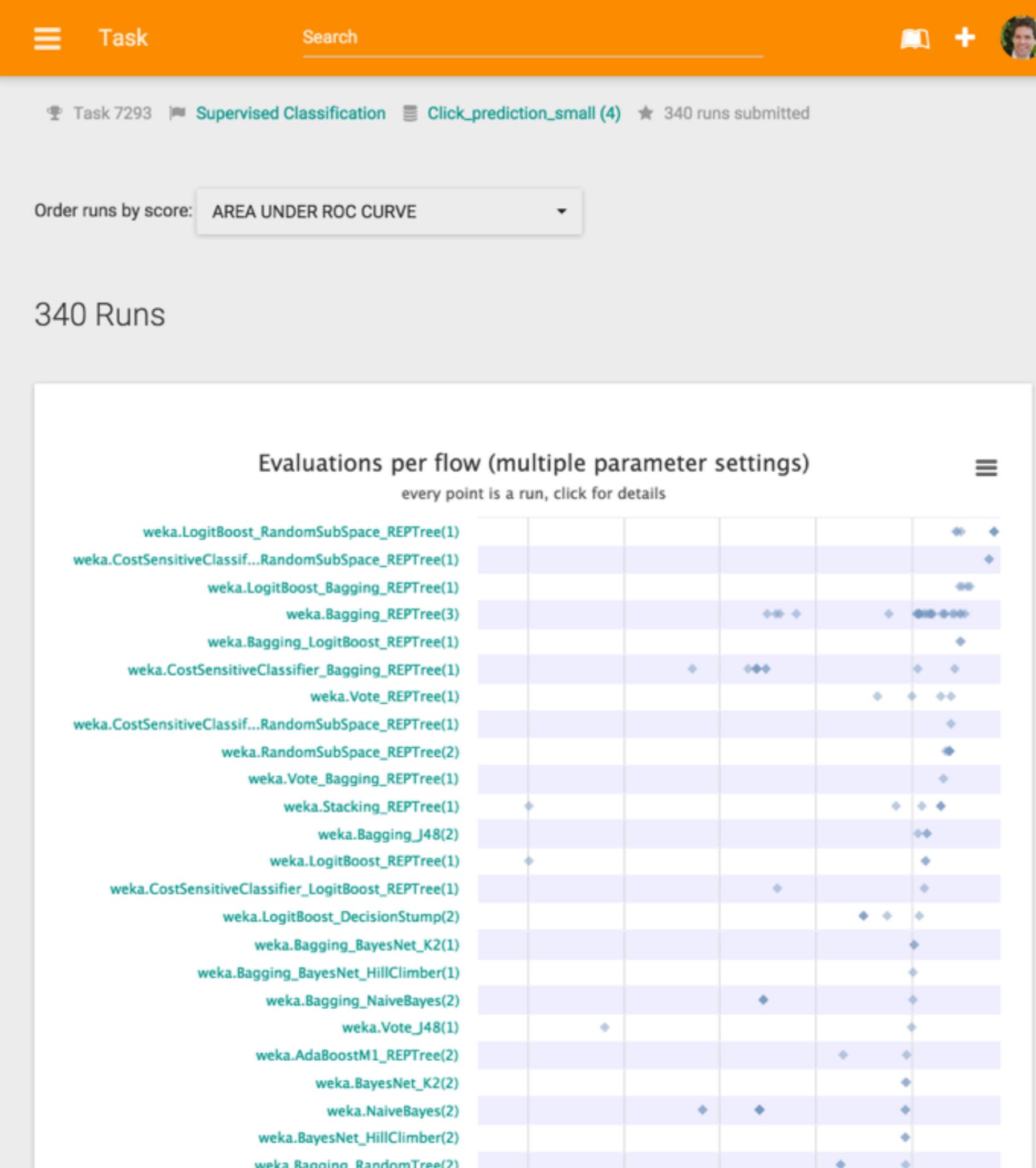


Scientific tasks
that can be
interpreted by
tools and **solved**
collaboratively

Scientists can create *tasks* for certain problems (e.g. classification): containers with all data and **machine-readable descriptions** so that tools can **automatically download data**, use the correct procedures, and **upload all results**.

All results are organized online, creating **realtime (collaborative) data mining challenges** where anyone can build on previous results.

- People submit results (e.g. predictions, models), obtained on their machines
 - OpenML computes large range of performance metrics, depending on task type
 - All results organized online
 - Online visualizations: every dot is a model obtained by specific solution and parameter variations.



Timeline

Details

Overview

All runs

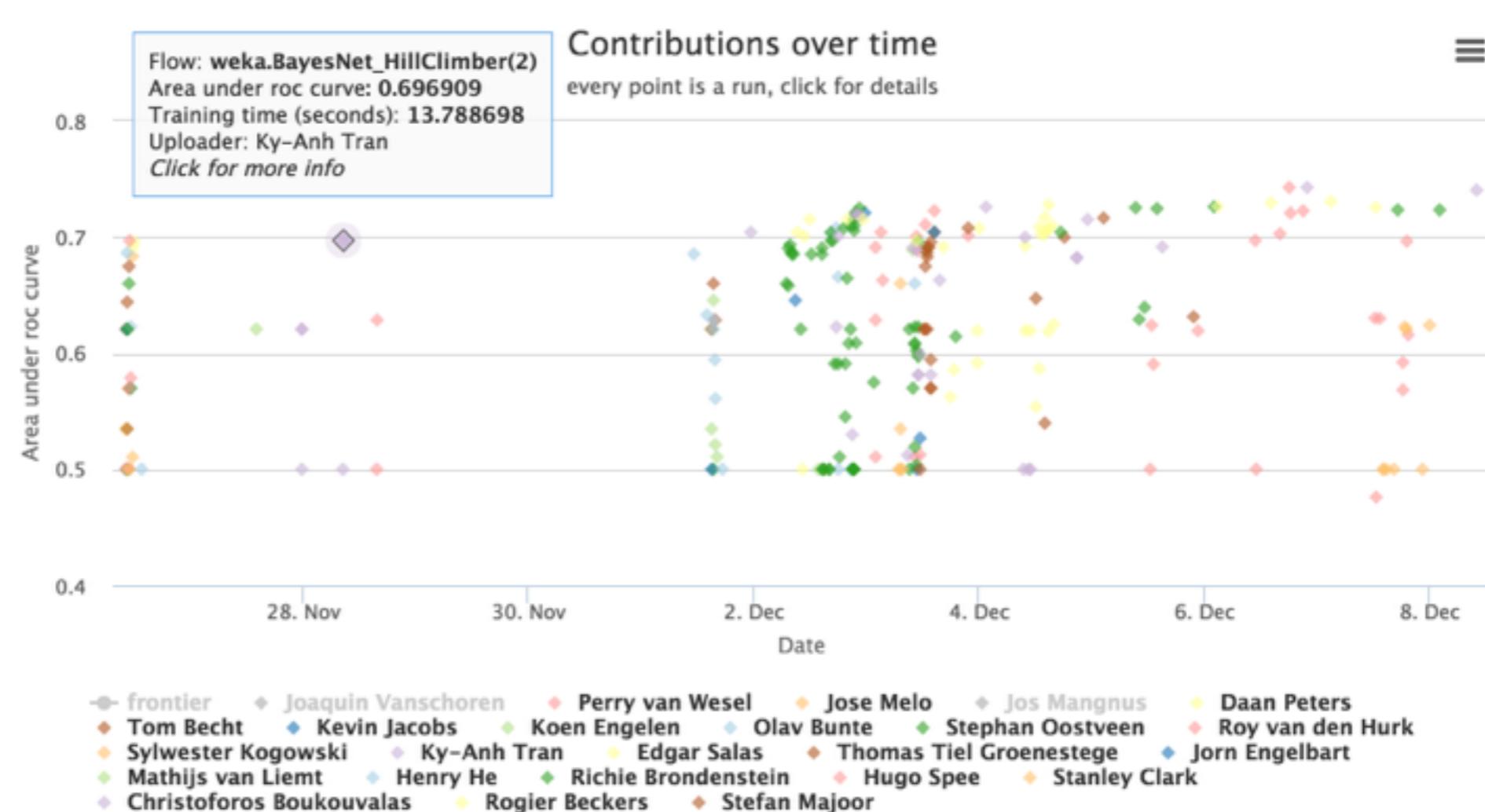
Results

Leaderboard

Discuss

Tags

Add tag



- Leaderboards with visualisation of progress over time: who delivered breakthrough results, who build on top of previous solutions
- Collaborative setting: you can see details of other solutions, rerun, learn from others
- Real-time: who submits first gets credit, others can improve immediately

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

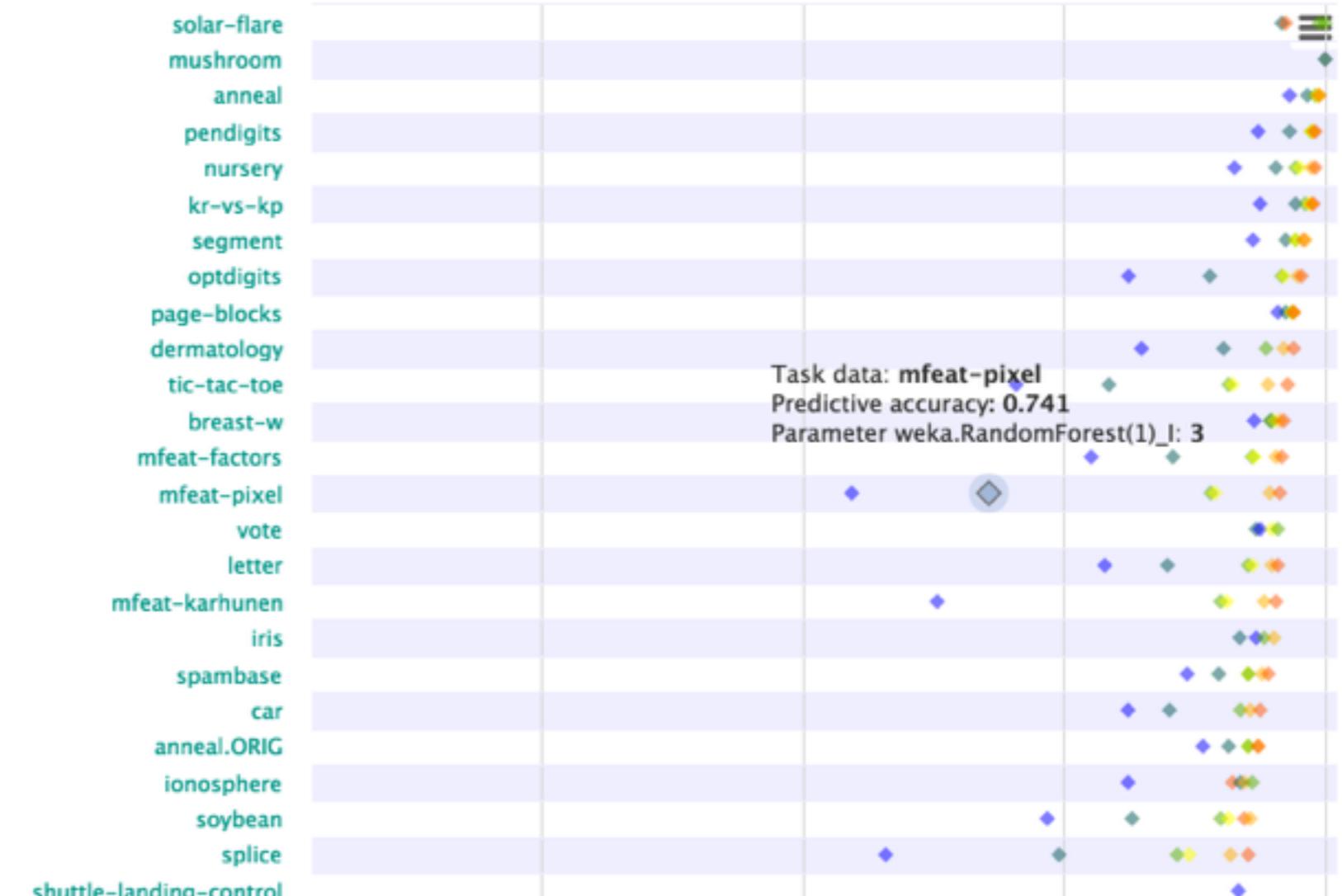
Scientists upload code used in experiments, or link from existing repositories. OpenML keeps track of **flow details and versioning**, **organizes all results** for easy comparison, even across tools. Tool integrations allow automated **data download**, **flow upload** and **experiment logging and sharing**.

 Data Tasks Flows Runs Task Types Measures People Guide Discussions Blog Details Overview Download flow

SUPERVISED CLASSIFICATION

PREDICTIVE ACCURACY

Parameter: I



- All results obtained with the same flow organised online
- Results linked to data sets, parameter settings, other details
- Also visualised online (dots are models, color-coded by parameter setting)

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

Runs contain the results that **flows** obtained on specific tasks. Runs are **fully reproducible**, linked to the underlying data, tasks, flows and authors. OpenML **organizes all results online** for **discovery, comparison and reuse**



Run



- All experiment (run) details stored
- Author, date, flow, parameter settings, result files, links to underlying data
- Computed evaluations, also for individual folds, samples

Run 84087

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

Area under ROC curve

0.7007 \pm 0.0023

Per class

0	1
0.7007	0.7007

Cross-validation details (10-fold Crossvalidation)



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.



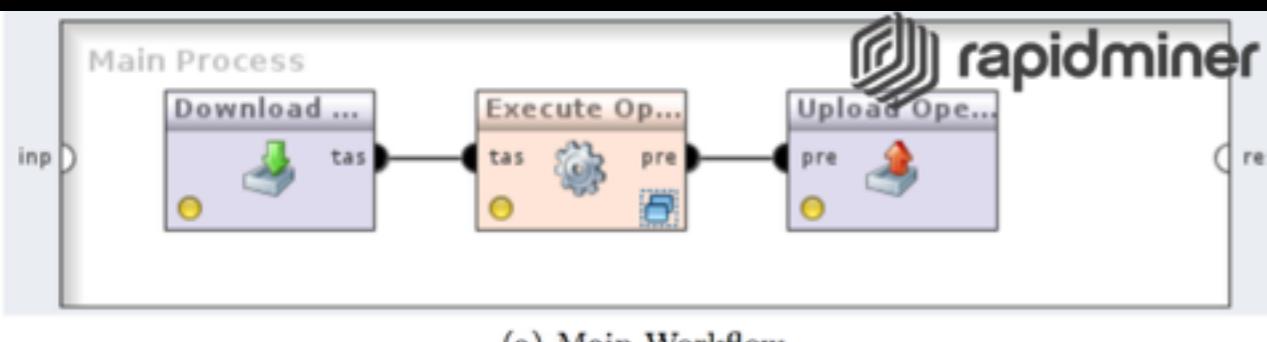
Predictions

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

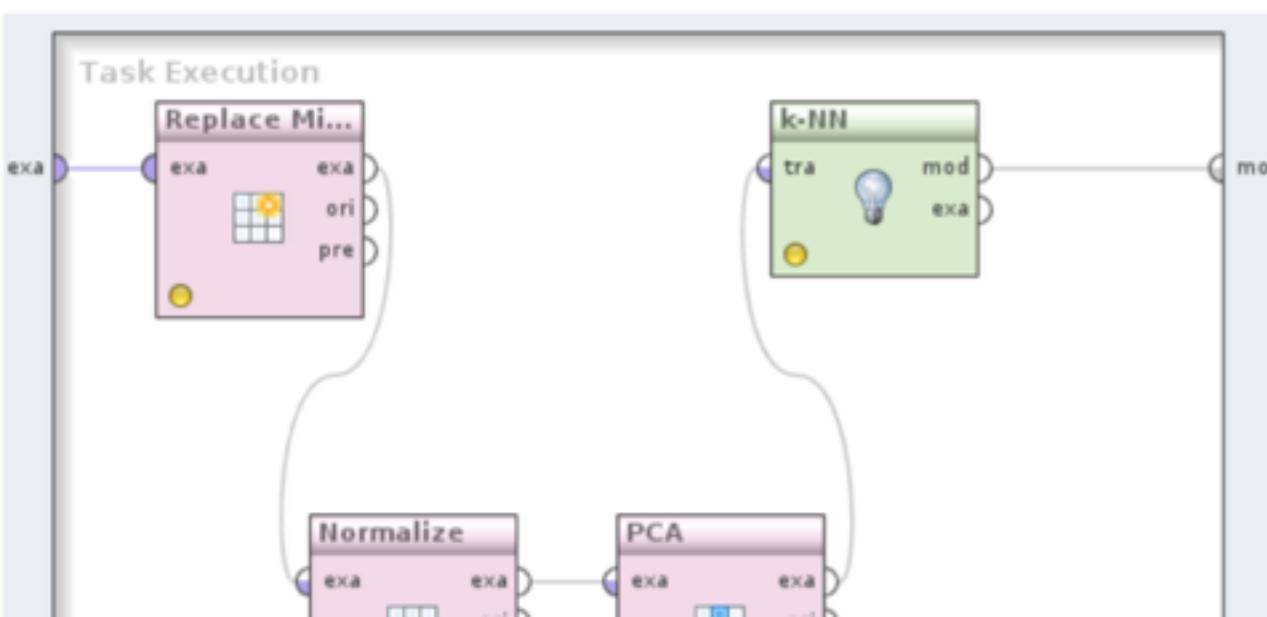
The screenshot shows the OpenML interface. At the top, there's a 'Results Destination' section with dropdowns for 'OpenML.org' and 'OpenML Username: joaqu'. To the right is a blue bird icon. Below this, the 'Experiment Type' section has 'OpenML Task' selected. The 'Iteration Control' section shows 'Number of repetitions: 1' with 'Data sets first' selected. Under 'Tasks', there are buttons for 'Add ...', 'Edi...', and 'Del...'. A list of tasks includes: 'Task 1: anneal - Supervised Classi', 'Task 2: anneal.ORIG - Supervised', 'Task 3: kr-vs-kp - Supervised Cla', 'Task 4: labor - Supervised Classifi', and 'Task 5: arrhythmia - Supervised C'. On the right, there's a 'Classification' tab selected in a toolbar. Below it is a table of command-line logs for a 'trees.HoeffdingAdaptiveTree' task, showing status, time, and command details. A 'Configure task' button is at the bottom. Another window on the right shows 'moa' software configuration for 'openml.OpenmlDataStreamClassification' with various parameters like learner, taskID, evaluator, etc.

- REST API + language specific APIs in Java, R, Python
- Download all data, flows, previous results in your favorite ML environment, automatically upload all your results
- WEKA/MOA: Available as plugin

- Also support for workflow-based tools (RapidMiner, ADAMS)
- R/Python: download data, run machine learning algorithms and upload results in just a few lines of code.



(a) Main Workflow



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



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

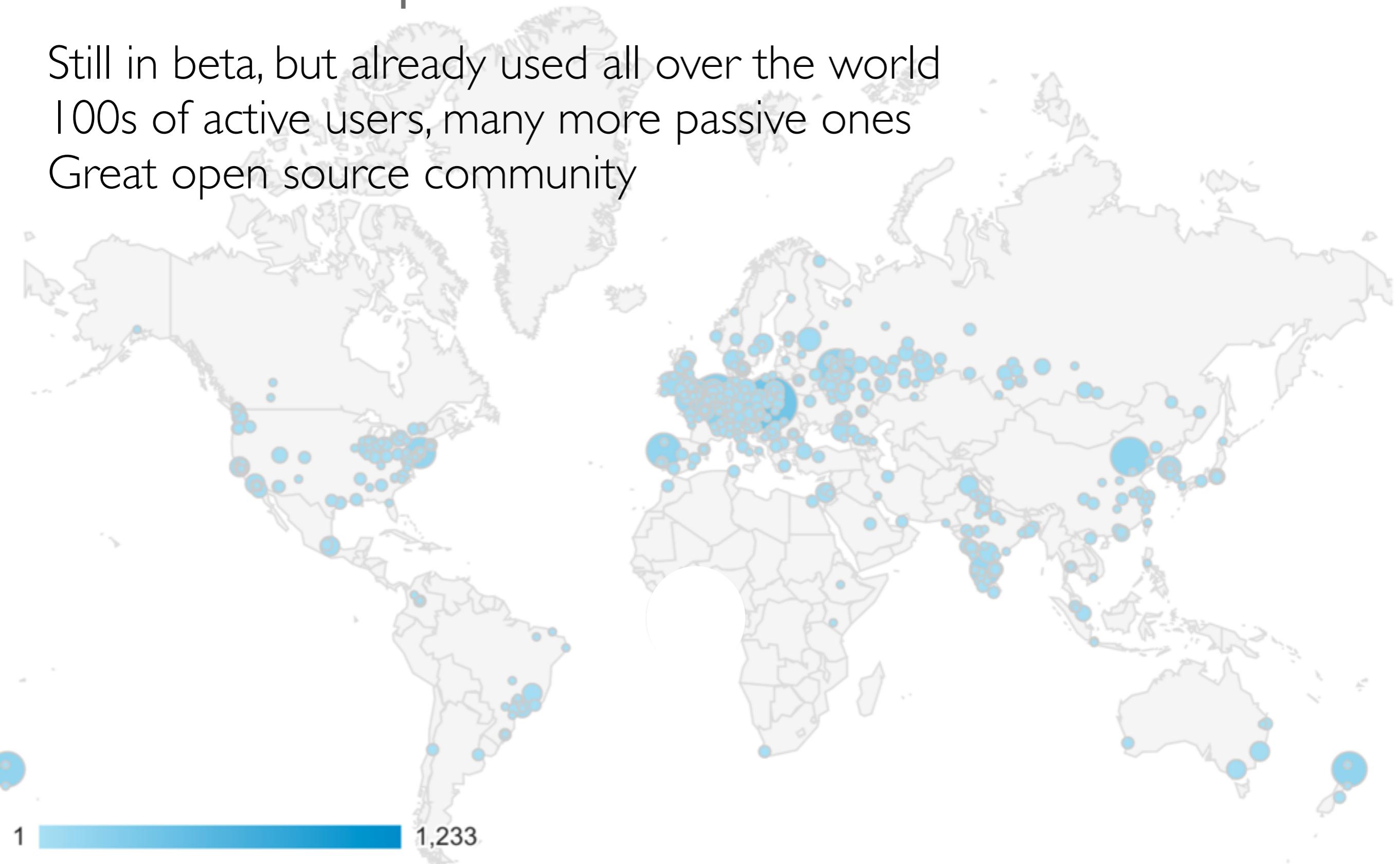


Global impact

Still in beta, but already used all over the world

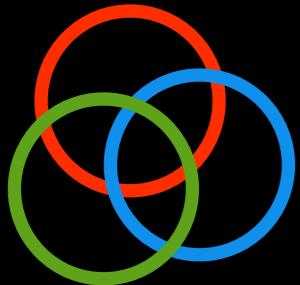
100s of active users, many more passive ones

Great open source community



Jan-Jun 2015

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

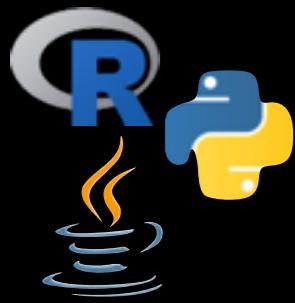
- Upload dataset, system recommends techniques
- Model-based optimisation techniques
- Continuous improvement (learns from past)

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

THANK YOU



#OpenML

Nenad Tomašev

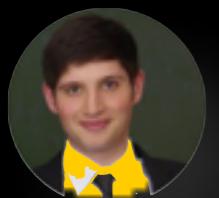


Luis Torgo



Jan van Rijn

Giuseppe Casalicchio



Joaquin Vanschoren



Michel Lang



Bernd Bischl



Matthias Feurer

You?