OpenML

NETWORKED MACHINE LEARNING

#OpenML

JOAQUIN VANSCHOREN (TU/E),2015

Research different.

NETWORKED SCIENCE
A REALTIME, WORLDWIDE LAB

Research different.

Collaborate large-scale, realtime

Automated sharing (via tools, reproducible)

Linked data: all data, code, results linked online

Large, interdisciplinary, trusted teams

Gain time, productivity, visibility, control

Research different.

Polymaths: Solve math problems through massive collaboration

Broadcast question, organise all ideas on online platform (wiki, blog)

Solved hard problems in weeks, **intense** collaboration, everyone playing to their strengths

Many (joint) publications

If you organise interesting ideas, many minds will build on them

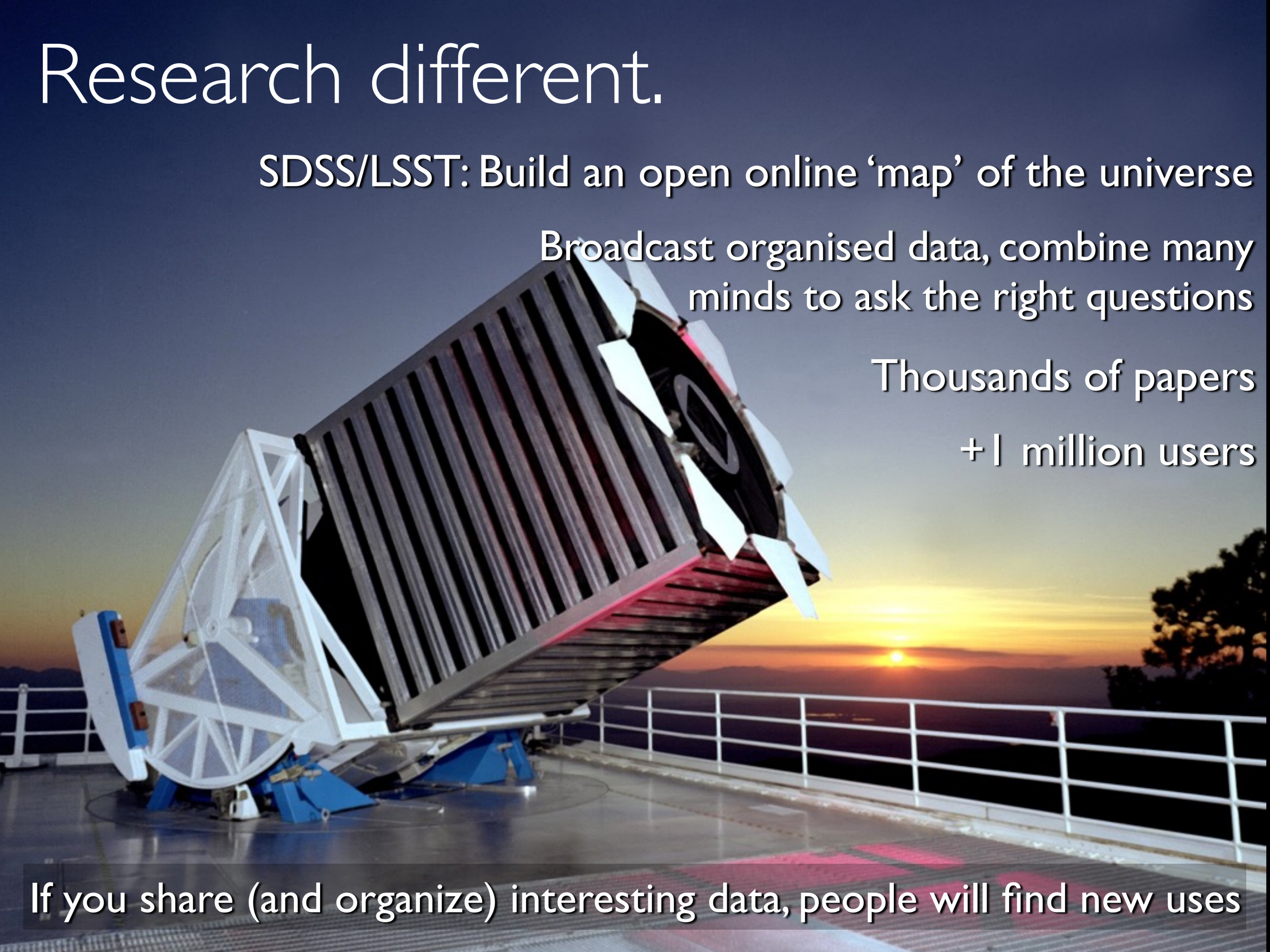
Research different.

SDSS/LSST: Build an open online ‘map’ of the universe

Broadcast organised data, combine many minds to ask the right questions

Thousands of papers

+1 million users



If you share (and organize) interesting data, people will find new uses

Research different.

- Zoo Universe, Apple ResearchKit,…
- Offer right tools so that anybody can contribute, instantly
- Many novel discoveries by scientists and citizens

If you allow collaboration to scale, you can solve any problem

Designed Serendipity

Designed Serendipity

Be the right person, on the right place, at the right time

Or: organise all ideas and data in one place, so that anyone can reuse and build on it, any time, any place

What's hard for one person, is easy for another.
What's surprising to some makes perfect sense to others.
If you have one half an idea, someone else may have the other half.

Tackle hard problems by 'connecting brains'

Allow to scale:
add transparency,
remove friction

Effortless (micro)contributions, of any kind

Track contributions, show impact openly

Organised body of compatible data and tools

Scientists keep control of what is shared when and with whom

Why machine learning (data science)

Complex code, large-scale data, experiments (impossible to print)

Experiments not shared online: impossible to build on prior work: inhibits deeper analysis (e.g. meta-learning)

Low reproducibility, generalisability (studies contradict)

Virtual walls between communities, domains, even continents. Scientists don't collaborate or understand each other, even if they work on the same topic.

What if we could all connect with each other, and with other scientists, to explore and apply machine learning?

Not just machine learning

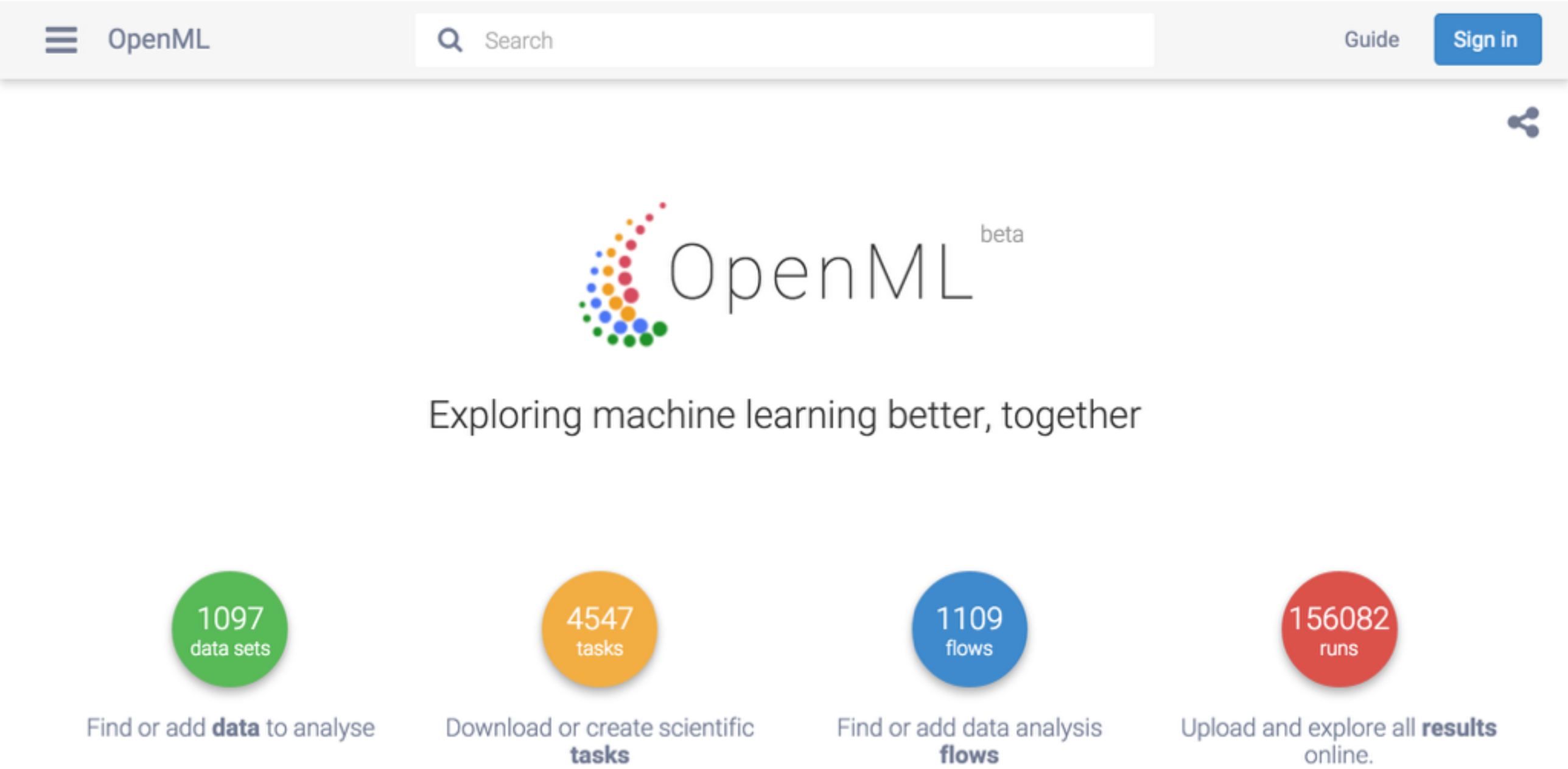
J. Ioannidis (2014)

85% research resources are wasted: many new associations/effects are false, exaggerated, translation into applications is inefficient

High false positive rate: underpowered (small scale) studies, small effect sizes, flexibility in design, biases, irreproducibility, lack of collaboration.

Some sciences have increased credibility: large-scale collaboration, replication culture, registration/sharing of data, reproducibility, better statistical methods and thresholds, better study design, training

New incentives: extra ‘points’ for replicated publication, successful translation, sharing data, training, reviewing. Lose points for refuted publications



Demo

(if wifi allows)

Data from
various sources
analysed and
organised online
for easy access

Scientists can **broadcast data,** explaining the challenge that needs to be addressed. OpenML will (for known data formats) **automatically analyze the data**, compute data characteristics, **annotate and index it for easy search**

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

Tasks are **realtime (collaborative) data mining challenges**, allowing anyone to build on previous results. OpenML creates **machine-readable descriptions** so that tools can **automatically download data**, use the correct procedures, and **upload all results online**.

Realtime challenges

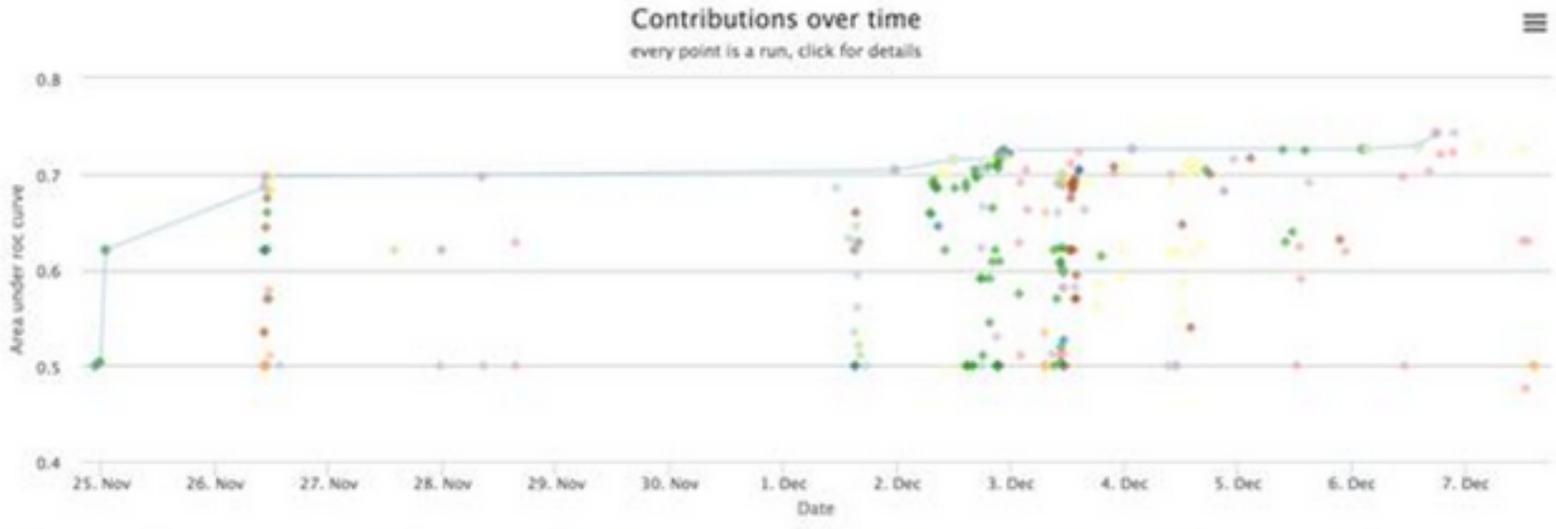
Rogier Beckers
@RogierBeckers

Follow

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

View translation

Lauradorp, Landgraaf



RETWEETS
2

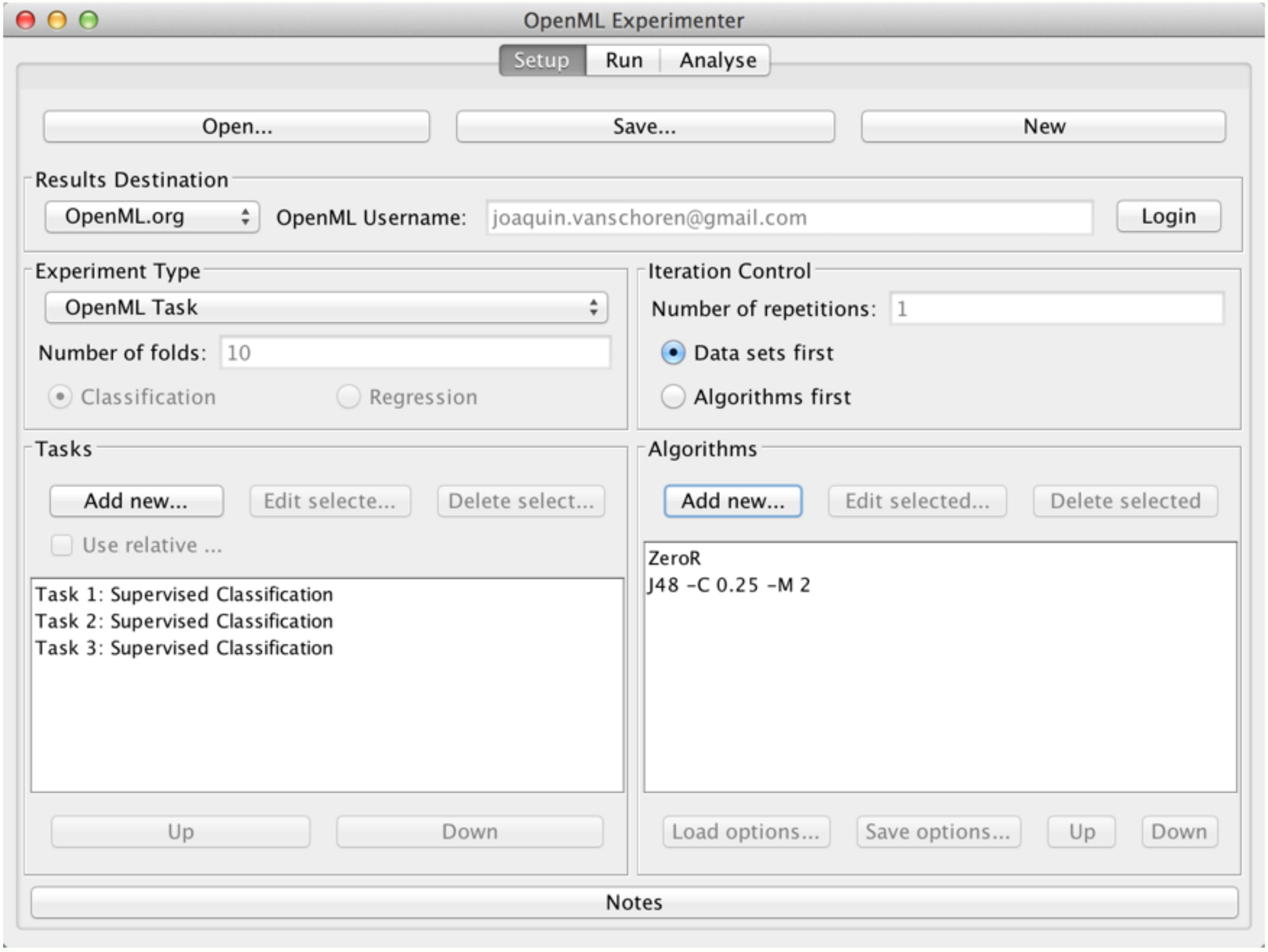
FAVORITES
2

9:48 PM - 7 Dec 2014

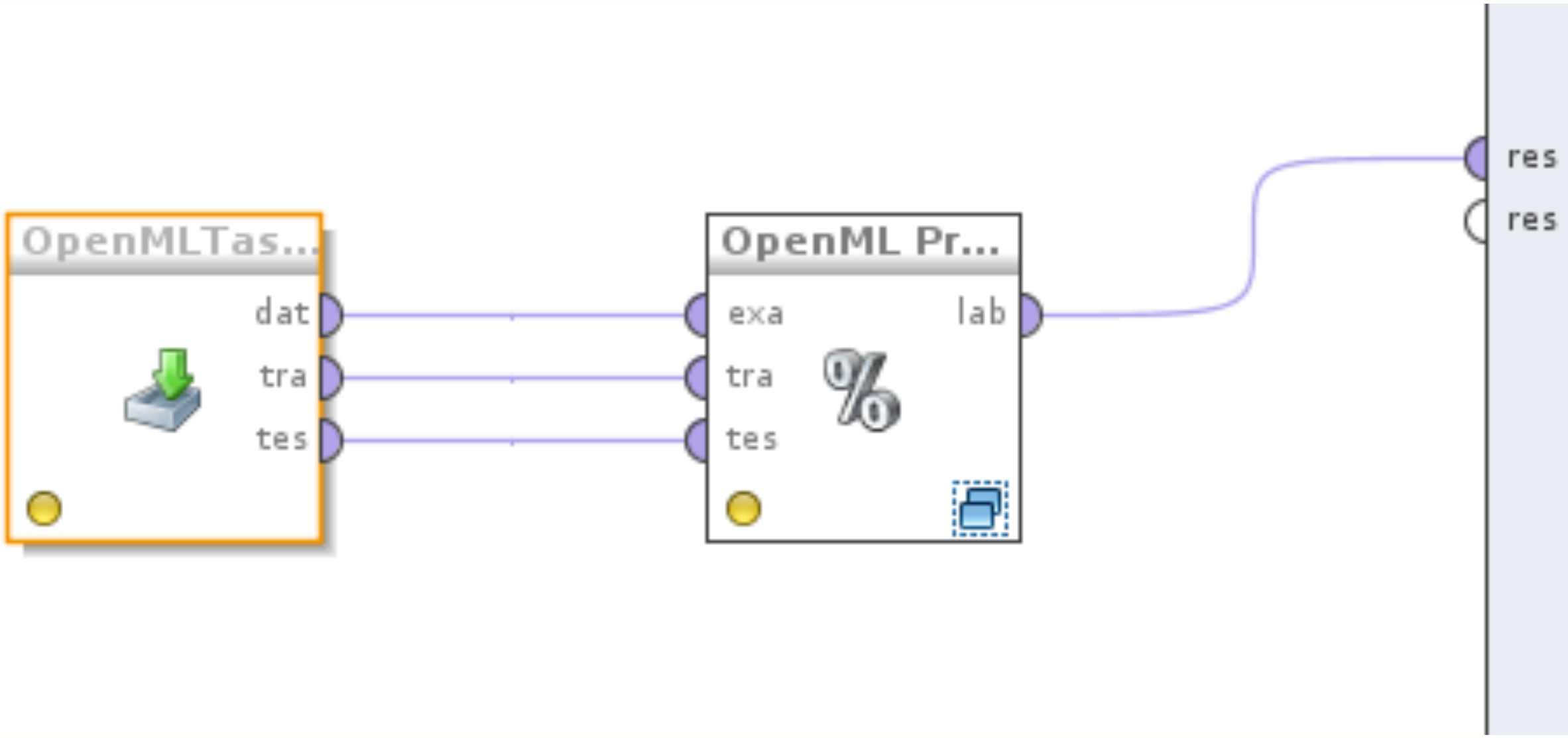
Tool plugins
for automated
data download,
workflow upload and
experiment logging
and sharing

Flows are implementations of algorithms, workflows, or scripts **solving OpenML tasks**. OpenML keeps track of **flow details and versioning**, **organizes all their results** for easy comparison, even across tools.

WEKA plugin


OpenML Experimenter
Setup
Run
Analyse
Open...
Save...
New
Results Destination
OpenML.org
OpenML Username:
joaquin.vanschoren@gmail.com
Login
Experiment Type
OpenML Task
Number of folds:
10
Classification
Regression
Iteration Control
Number of repetitions:
1
Data sets first
Algorithms first
Tasks
Add new...
Edit selecte...
Delete select...
Use relative ...
Task 1: Supervised Classification
Task 2: Supervised Classification
Task 3: Supervised Classification
Up
Down
Algorithms
Add new...
Edit selected...
Delete selected
ZeroR
J48 -C 0.25 -M 2
Load options...
Save options...
Up
Down
Notes

RapidMiner plugin



1. OPERATOR TO DOWNLOAD TASK (TASK TYPE SPECIFIC)
2. SUBWORKFLOW THAT SOLVES THE TASK, GENERATES RESULTS
3. OPERATOR FOR UPLOADING RESULTS

```
<oml:task xmlns:oml="http://openml.org/openml">
  <oml:task_id>59</oml:task_id>
  <oml:task_type>Supervised Classification</oml:task_type>
  <oml:input name="source_data">
    <oml:data_set>
      <oml:data_set_id>61</oml:data_set_id>
      <oml:target_feature>class</oml:target_feature>
    </oml:data_set>
  </oml:input>
  <oml:input name="estimation_procedure">
    <oml:estimation_procedure>
      <oml:type>crossvalidation</oml:type>
      <oml:data_splits_url>
        http://openml.org/api_splits/get/59/Task_59_splits.arff
      </oml:data_splits_url>
      <oml:parameter name="number_repeats">1</oml:parameter>
      <oml:parameter name="number_folds">10</oml:parameter>
      <oml:parameter name="percentage"/>
      <oml:parameter name="stratified_sampling">true</oml:parameter>
    </oml:estimation_procedure>
  </oml:input>
  <oml:input name="evaluation_measures">
    <oml:evaluation_measures>
      <oml:evaluation_measure/>
    </oml:evaluation_measures>
  </oml:input>
  <oml:output name="predictions">
    <oml:predictions>
      <oml:format>ARFF</oml:format>
      <oml:feature name="repeat" type="integer"/>
      <oml:feature name="fold" type="integer"/>
      <oml:feature name="row_id" type="integer"/>
      <oml:feature name="confidence.classname" type="numeric"/>
      <oml:feature name="prediction" type="string"/>
    </oml:predictions>
  </oml:output>
</oml:task>
```

Machine-interpretable tasks

```
@relation openml_task_77_predictions

@attribute repeat numeric
@attribute fold numeric
@attribute sample numeric
@attribute row_id numeric
@attribute confidence.1 numeric
@attribute confidence.2 numeric
@attribute confidence.3 numeric
@attribute confidence.4 numeric
@attribute confidence.5 numeric
@attribute confidence.6 numeric
@attribute confidence.7 numeric
@attribute confidence.8 numeric
@attribute confidence.9 numeric
@attribute confidence.10 numeric
@attribute prediction {1,2,3,4,5,6,7,8,9,10}
@attribute correct {1,2,3,4,5,6,7,8,9,10}

@data
0,0,0,9,1,0,0,0,0,0,0,0,0,0,1,1
0,0,0,388,0,1,0,0,0,0,0,0,0,0,2,2
0,0,0,474,0,0,1,0,0,0,0,0,0,0,3,3
0,0,0,750,0,0,0,1,0,0,0,0,0,0,4,4
0,0,0,903,0,0,0,0,0.75,0,0,0.25,0,0,5,5
0,0,0,1023,0,0,1,0,0,0,0,0,0,0,3,6
0,0,0,1252,0,0,0.111111,0,0,0,0.222222,0,0,0.666667,10,7
0,0,0,1483,0,0,0,0,0,0,0,1,0,0,8,8
0,0,0,1682,0,0,0,0,0,0,0,0,1,0,9,9
0,0,0,1870,0,0,0.111111,0,0,0,0.222222,0,0,0.666667,10,10
0,0,0,3,1,0,0,0,0,0,0,0,0,0,1,1
0,0,0,398,0,1,0,0,0,0,0,0,0,0,2,2
0,0,0,468,0,0,0.333333,0,0,0.666667,0,0,0,0,6,3
0,0,0,705,0,0,0.333333,0,0,0.666667,0,0,0,0,6,4
0,0,0,990,0,0,0,0,0.75,0,0,0.25,0,0,5,5
0,0,0,1131,0,0,0,0,0,1,0,0,0,0,6,6
0,0,0,1355,0,0,0.111111,0,0,0,0.222222,0,0,0.666667,10,7
0,0,0,1443,0,0,0,0,0,0,0,1,0,0,8,8
0,0,0,1663,0,0,0,0,0,0,0,0,1,0,9,9
0,0,0,1996,0,0,0.111111,0,0,0,0.222222,0,0,0.666667,10,10
0,0,0,27,1,0,0,0,0,0,0,0,0,0,1,1
0,0,0,380,0,1,0,0,0,0,0,0,0,0,2,2
0,0,0,487,0,0,1,0,0,0,0,0,0,0,3,3
0,0,0,714,0,0,0,1,0,0,0,0,0,0,4,4
0,0,0,912,0,0,0,0,0.75,0,0,0.25,0,0,5,5
0,0,0,1192,0,0,0,1,0,0,0,0,0,0,4,6
0,0,0,1375,0,0,0,0,0,0,1,0,0,0,7,7
0,0,0,1477,0,1,0,0,0,0,0,0,0,0,2,8
0,0,0,1648,0,0,0,0,0,0,0,0,1,0,9,9
0,0,0,1903,0,0,0.111111,0,0,0,0.222222,0,0,0.666667,10,10
0,0,0,164,0,0,1,0,0,0,0,0,0,0,3,1
```

```
J48 pruned tree
------------------

att1 <= 0
|   att5 <= 1.548576: 2 (202.0/15.0)
|   att5 > 1.548576
|   |   att6 <= 8574.682952
|   |   |   att6 <= 5804.513685
|   |   |   |   att5 <= 1.609052
|   |   |   |   |   att2 <= 2
|   |   |   |   |   |   att4 <= 144.472861
|   |   |   |   |   |   |   att6 <= 4528.578862: 8 (3.0/1.0)
|   |   |   |   |   |   |   att6 > 4528.578862: 2 (3.0/1.0)
|   |   |   |   |   |   att4 > 144.472861
|   |   |   |   |   |   |   att6 <= 5120.638273: 8 (7.0)
|   |   |   |   |   |   |   att6 > 5120.638273: 5 (3.0/1.0)
|   |   |   |   |   att2 > 2
|   |   |   |   |   |   att5 <= 1.589619: 5 (3.0)
|   |   |   |   |   |   att5 > 1.589619: 2 (2.0)
|   |   |   |   att5 > 1.609052: 8 (104.0/6.0)
|   |   |   att6 > 5804.513685
|   |   |   |   att2 <= 2
|   |   |   |   |   att4 <= 163.296861
|   |   |   |   |   |   att5 <= 1.773592
|   |   |   |   |   |   |   att6 <= 6493.591741
|   |   |   |   |   |   |   |   att4 <= 155.794861: 5 (22.0/8.0)
|   |   |   |   |   |   |   |   att4 > 155.794861
|   |   |   |   |   |   |   |   |   att5 <= 1.643385: 4 (4.0/1.0)
|   |   |   |   |   |   |   |   |   att5 > 1.643385
|   |   |   |   |   |   |   |   |   |   att5 <= 1.743657: 8 (7.0/1.0)
|   |   |   |   |   |   |   |   |   |   att5 > 1.743657: 4 (2.0/1.0)
|   |   |   |   |   |   |   att6 > 6493.591741
|   |   |   |   |   |   |   |   att4 <= 161.344861: 5 (16.0/2.0)
|   |   |   |   |   |   |   |   att4 > 161.344861
|   |   |   |   |   |   |   |   |   att6 <= 7329.973134: 3 (6.0/2.0)
|   |   |   |   |   |   |   |   |   att6 > 7329.973134: 5 (2.0)
|   |   |   |   |   |   att5 > 1.773592
|   |   |   |   |   |   |   att6 <= 7329.973134: 8 (21.0/5.0)
|   |   |   |   |   |   |   att6 > 7329.973134: 5 (5.0/2.0)
|   |   |   |   |   att4 > 163.296861
|   |   |   |   |   |   att6 <= 7908.003397
|   |   |   |   |   |   |   att5 <= 1.703938
|   |   |   |   |   |   |   |   att5 <= 1.639492: 4 (2.0/1.0)
|   |   |   |   |   |   |   |   att5 > 1.639492: 3 (2.0)
|   |   |   |   |   |   |   att5 > 1.703938: 8 (50.0/6.0)
|   |   |   |   |   |   att6 > 7908.003397
|   |   |   |   |   |   |   att4 <= 172.932861
|   |   |   |   |   |   |   |   att6 <= 8306.160447
|   |   |   |   |   |   |   |   |   att6 <= 8105.99096: 6 (3.0/1.0)
|   |   |   |   |   |   |   |   |   att6 > 8105.99096: 4 (5.0)
|   |   |   |   |   |   |   |   att6 > 8306.160447
```

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

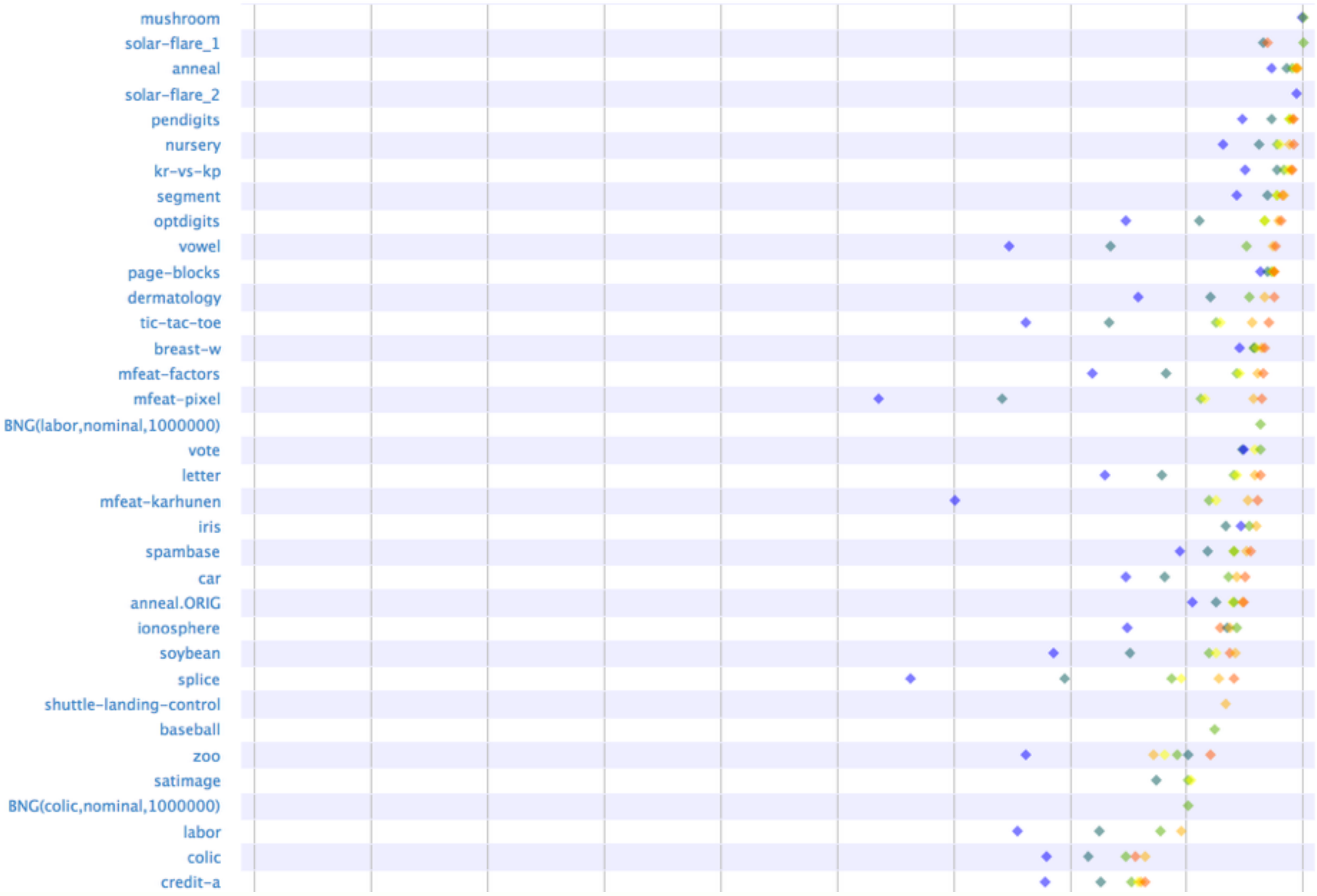
Runs contain the results that **flows** obtained on specific tasks. Runs are **fully reproducible**, linked to the underlying data, tasks, flows and authors. OpenML **organizes all results online** for **discovery, comparison and reuse**

Performance evaluation

Overview of results per flow

Evaluation measure: predictive accuracy

Parameter: I



OpenML.org

Search, add/ delete, share, studies,…

Data

Tasks

Flows

Runs

Java API

R API

Python API

Web API

moa

Open for Innovation KNIME

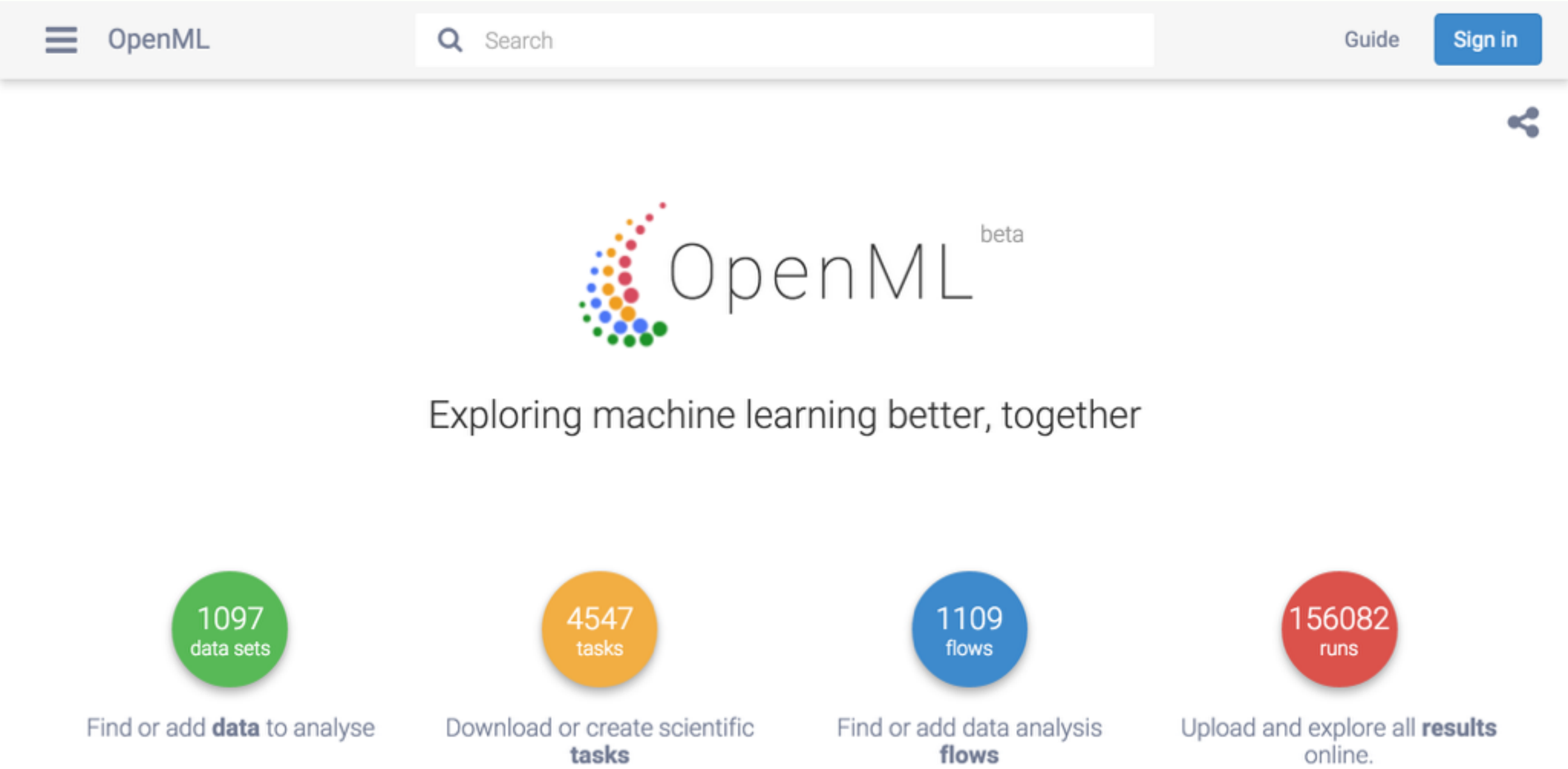
rapidminer

mlr

caret*, ipred*

SciKit*

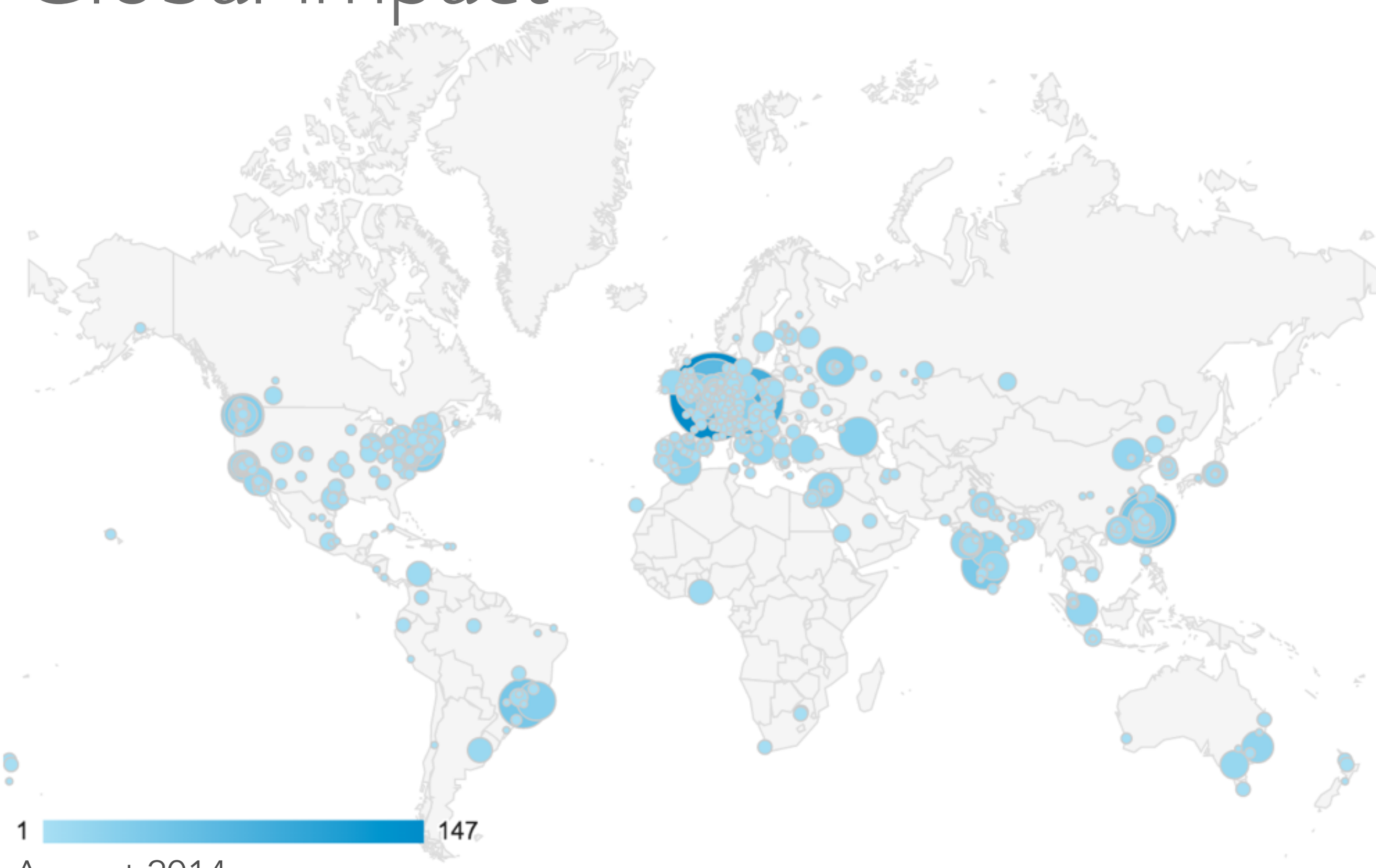
…



Demo

(if wifi allows)

Global impact



August 2014

Soon…

OpenML studies (notebooks)

- online counterpart of paper (url backlink)
- linked to paper to promote citation
- collection of datasets, flows, runs, results + description
- easily include (build on) data of others

Teams

- Add scientists in teams (circles)
- Share resources, results within team only
- Make public at any time (e.g. after publication)

Impact tracking

- Profile page: statistics of activity and impact on OpenML
- Altmetrics: e.g. datasets shared and reused

OpenML for algorithm selection

Meta-data:
- Growing collection of datasets
- Wide array of meta-features (more coming)
- Wide range of integrated machine learning algorithms

Algorithm selection
- Use existing experiments to train algorithm selection techniques
- Upload dataset, ask system to propose ML algorithms
- People can start, track, visualise AS process from website

Meta-learning support:
- View run status
- Generate datasets from meta-data
- Meta-learning studies

Current work

Active testing with subsampling (Brazdil, van Rijn):

- Iteratively select most promising algorithm based on previous performance on 'similar' datasets (same algorithms win or lose)

Stream algorithm selection (Pfahringer, Holmes, van Rijn):

- Build meta-models to predict most promising technique based on current properties of data (moving window)

Dataset similarity on latent meta-features (Sebag):

- Matrix factorization to extract latent meta-features

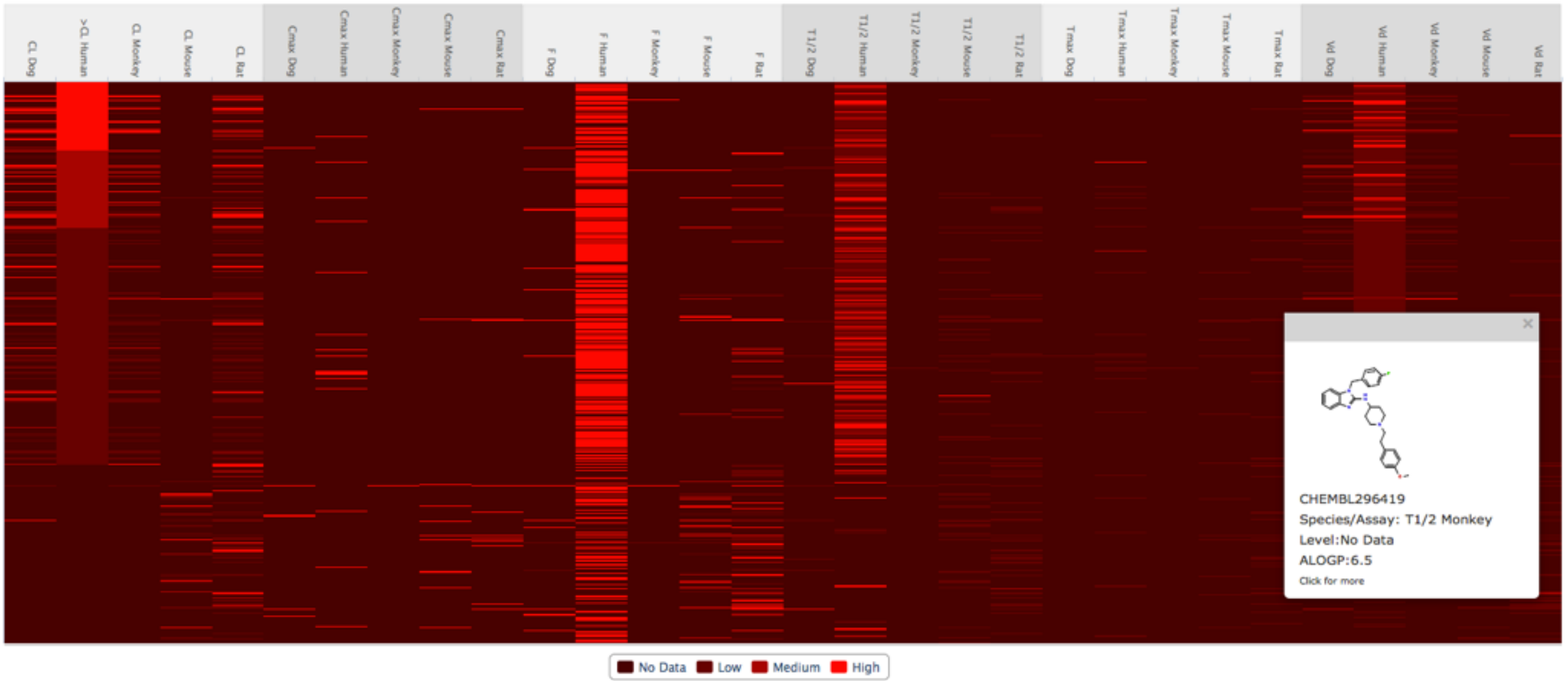
Meta-QSAR (King, Soldatova, Sadawi):

- Meta-learning to predict best approaches for mining QSAR data

Meta-QSAR

Drug discovery for rare tropical diseases (malaria)

- Data from ChEMBL processed into 10.000s datasets
- Meta-data on molecules (6000+ features), target enzymes
- We learn which features and algorithms are most useful



OpenML

JOIN THE CLUB!

Joaquin Vanschoren
University of Eindhoven

Hendrik Blockeel
University of Leuven

Geoffrey Holmes
University of Waikato

Jan van Rijn
University of Leiden

Luis Torgo
University of Porto

Bernd Bischl
Universty of Dortmund

Simon Fischer
RapidMiner

Patrick Winter
KNIME.com

Bo Gao
University of Leuven

Milan Vukicevic
University of Belgrade

Sandro Radovanović
University of Belgrade

You?
Join now!