



OpenML

COLLABORATIVE MACHINE LEARNING

JOAQUIN VANSCHOREN, TU EINDHOVEN, 2015



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

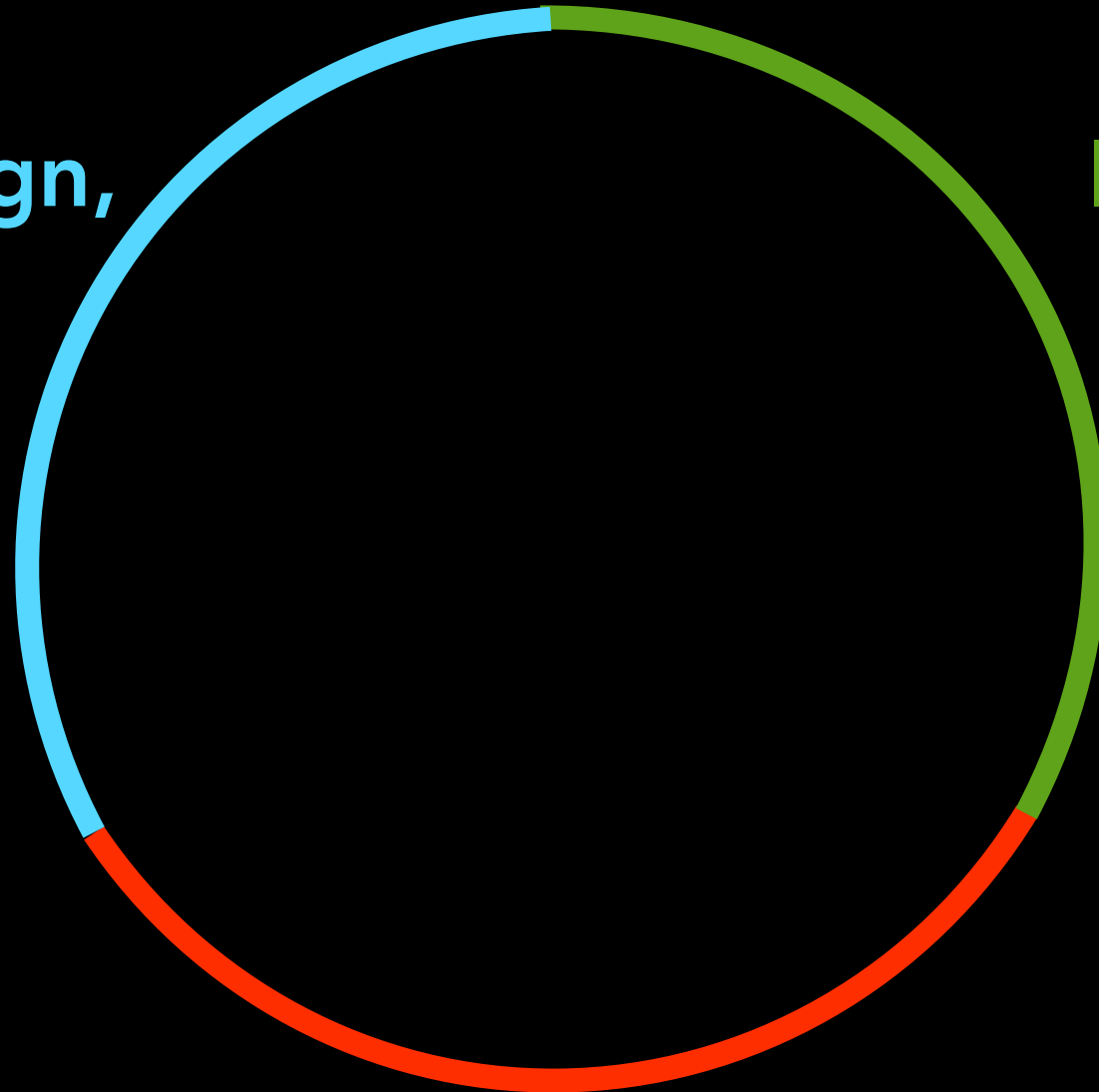
Data(driven) science ecosystem

Few of us are experts in all crafts at once (we collaborate)

**Algorithm design,
selection**

**Implementing,
running code**

**Problem definition,
data collection**



Gaps in the ecosystem

Algorithm design,
selection

Domain experts: learning and trying latest/best data science techniques **takes lots of time**

Algorithm experts: learning domain language, finding latest/relevant data **takes lots of time**

Unnecessary friction: time lost on tasks that others do in a fraction, automate altogether

Problem definition,
data collection

MUCH OF WHAT MEDICAL RESEARCHERS conclude in their studies is misleading, exaggerated, or flat-out wrong. So why are doctors—to a striking extent—still drawing upon misinformation in their everyday practice? Dr. John Ioannidis has spent his career challenging his peers by exposing their bad science.

LIES, DAMNED LIES, AND MEDICAL SCIENCE

By DAVID H. FREEDMAN

IN 2001, RUMORS were circulating in Greek hospitals that surgery residents, eager to rack up scalpel time, were falsely diagnosing hapless Albanian immigrants with appendicitis. At the University of Ioannina medical school's teaching hospital, a newly minted doctor named Athina Tzioumi was discussing the rumors with colleagues when a professor who had overheard asked her if she'd like to try to prove whether they were true—he seemed to be almost daring her. She accepted the challenge and, with the professor's and other colleagues' help, eventually produced a formal study showing that, for whatever reason, the appendices removed from patients with Albanian names in six Greek hospitals were more than three times as likely to be perfectly healthy as those removed from patients with Greek names. "It was hard to find a journal willing to publish it, but we did," recalls Tzioumi. "I also discovered

that I really liked research." Good thing, because the study had actually been a sort of audition. The professor, it turned out, had been putting together a team of exceptionally brash and curious young clinicians and Ph.D.s to join him in tackling an unusual and controversial agenda.

Last spring, I sat in on one of the team's weekly meetings on the medical school's campus, which is plunked crazily across a series of sharp hills. The building in which we met, like most at the school, had the look of a barracks and was festooned with political graffiti. But the group convened in a spacious conference room that would have been at home at a Silicon Valley start-up. Sprawled around a large table were Tzioumi and eight other youngish Greek researchers and physicians who, in contrast to the posty younger staff frequently seen in U.S. hospitals, looked like the casually glamorous cast of a television medical drama. The professor,



Dr. John Ioannidis, photographed in August at Stanford University's Cecil H. Green Library

In subfields, up to 85% medical research resources are wasted

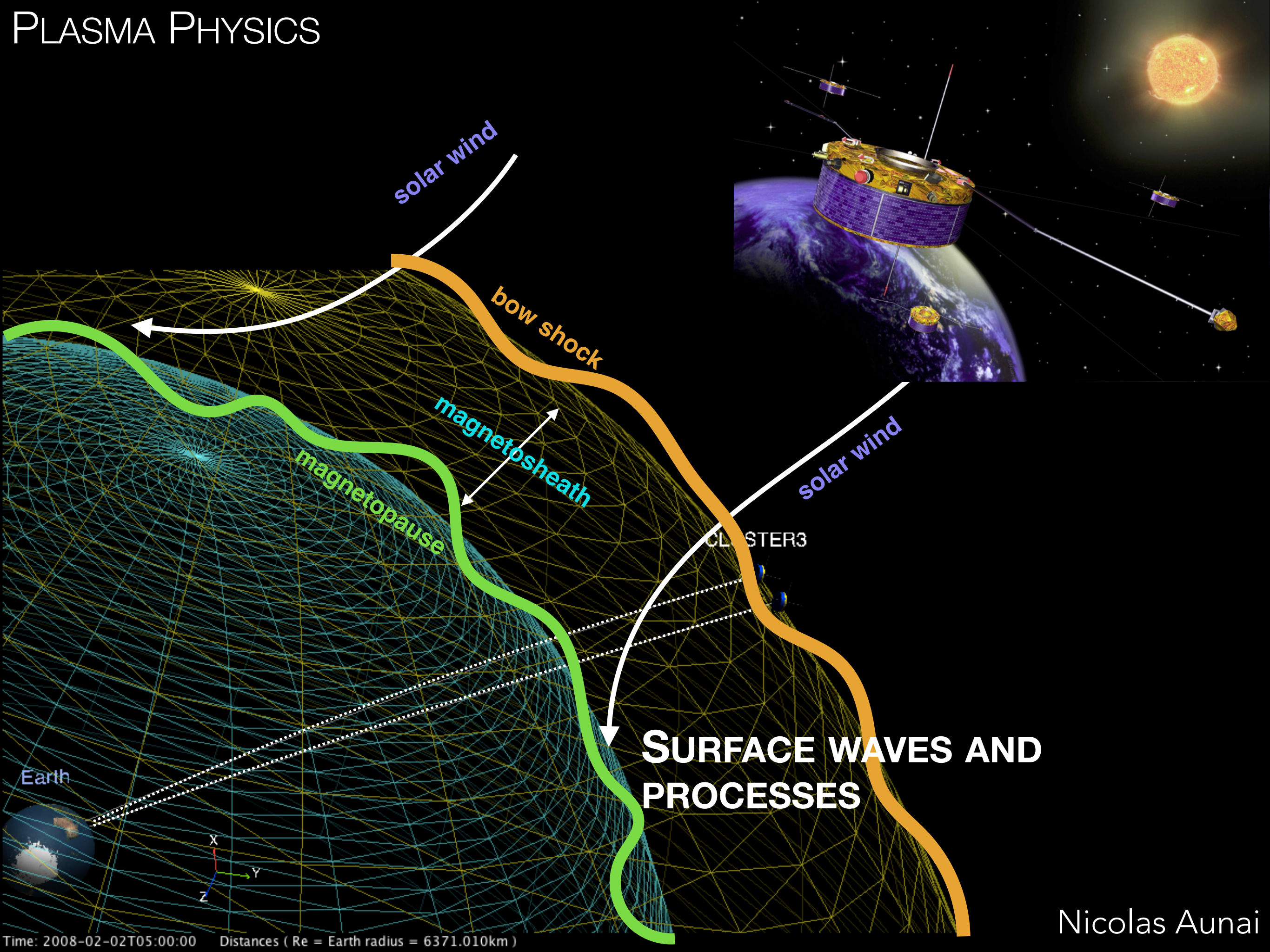
- underpowered, irreproducible research
- improper analysis: false associations
- flexibility in design, analysis
- not translatable to applications

Research better if:

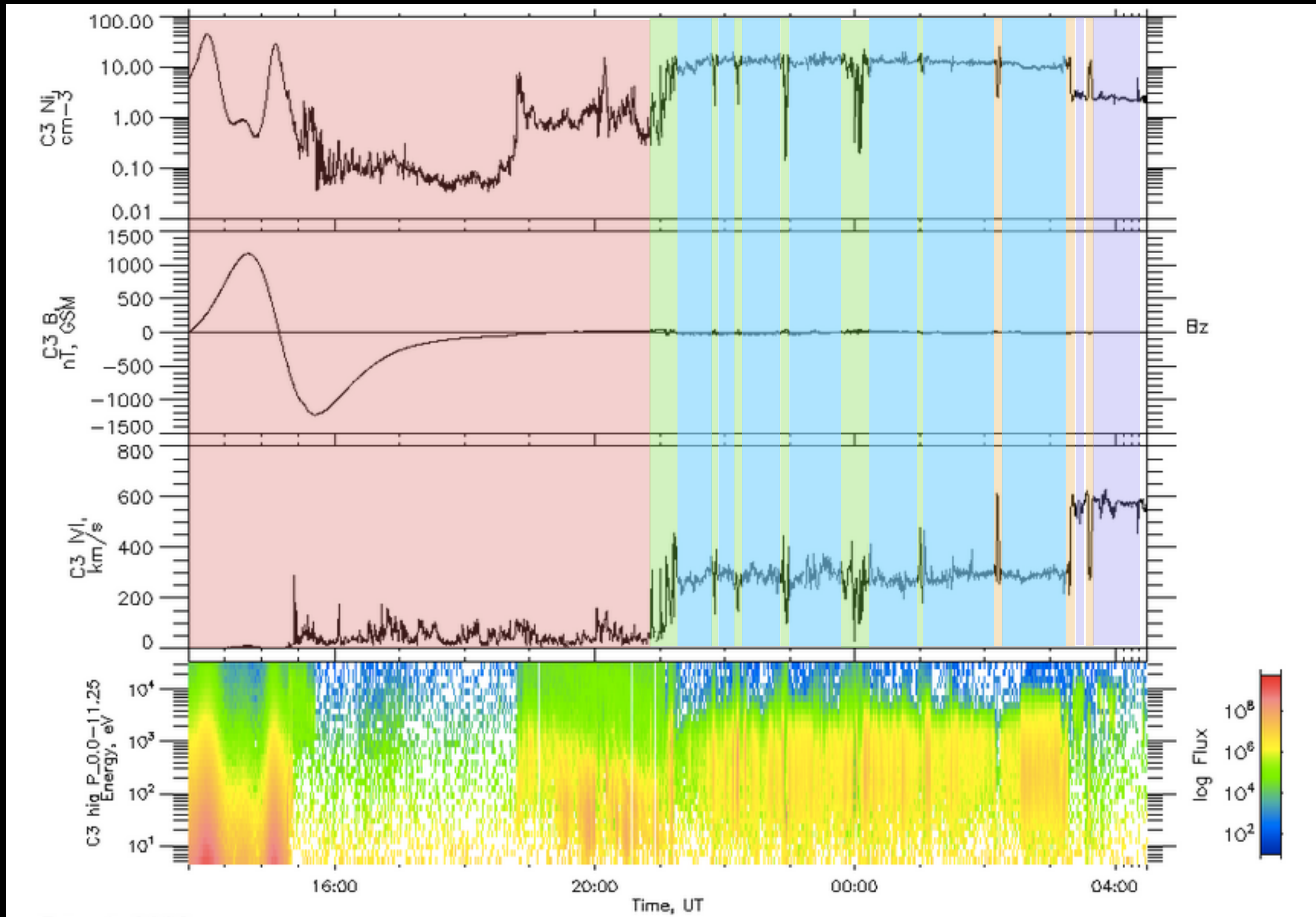
- Large-scale, interdisciplinary
- Easy replication
- Open sharing of data, protocols, software
- Better methods, tests



PLASMA PHYSICS



LABELLING (CATALOGUING) OF EVENTS STILL DONE MANUALLY



Gaps in the ecosystem

Algorithm design,
selection

Slows down, underpowers research

Ioannidis: 85% medical research ineffective

Similar findings in other fields

Enterprises lack access to expertise, data

Duplicate work, suboptimal solutions

Value from data could be faster, cheaper

Shortage of data science expertise

McKinsey: 190k data scientists needed by 2018

Data science needs to scale: frictionless collaboration across fields and labs, open data, democratisation, automate drudge work

Problem definition,
data collection

Gaps in the ecosystem

Algorithm design,
selection

Small-scale collaboration

Problem definition,
data collection

Necessary, but:

People's attention **doesn't scale.**

Time is limited

Likely biased: asking N experts gives you
 N different answers

Gaps in the ecosystem

Algorithm design,
selection

Literature

Problem definition,
data collection

Yes, but:

Slow: *too many* papers, domain-specific jargon, little cross-domain relevance. Faster to just try things yourself

Mining the literature? OK, but information in papers is often imprecise, aggregated, hard to reproduce

Paper is a 300 year old medium, internet is a much better one

Gaps in the ecosystem

Algorithm design,
selection

Networked Science

Problem definition,
data collection

Broadcast data so that many minds analyse the data in different ways

Broadcast code so that many minds can apply it on their own data

Organize everything on a collaboration platform

Research different.

Polymaths: Solve math problems through massive collaboration

Broadcast question, combine many minds to solve it

Solved hard problems in weeks

Many (joint) publications

Research different.

SDSS: Robotic telescope, data publicly online (SkyServer)

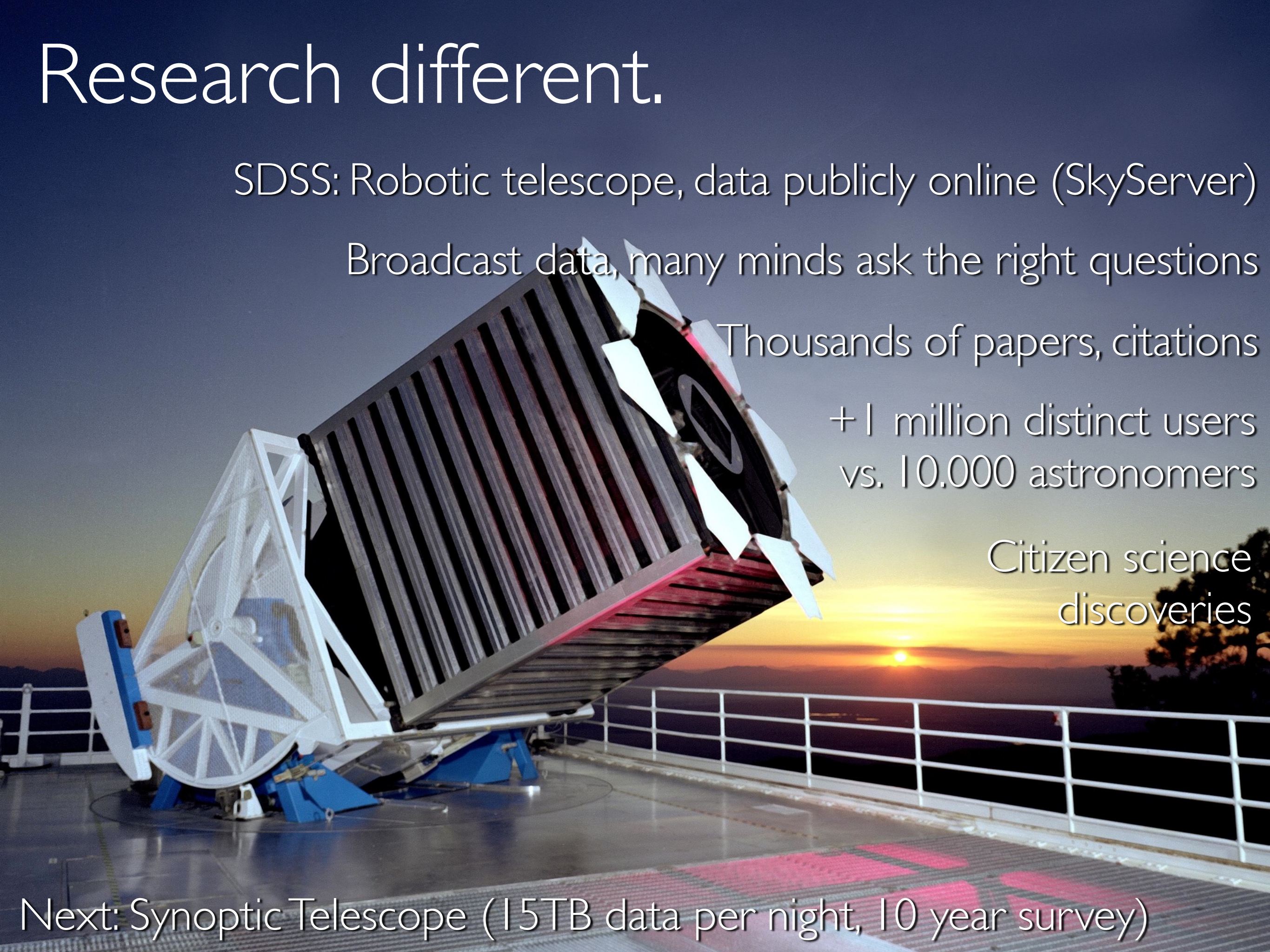
Broadcast data, many minds ask the right questions

Thousands of papers, citations

+ 1 million distinct users
vs. 10,000 astronomers

Citizen science
discoveries

Next: Synoptic Telescope (15TB data per night, 10 year survey)



Research different.

How do you label a million galaxies?

Offer simple tools so that anybody can contribute, instantly

Many novel discoveries by scientists and citizens

Designed serendipity

What's hard/surprising for one scientist is easy for another

If data/code is accessed easily, it will be used in unexpected ways

Data is the new soil. Algorithms are like seeds that we sow on them

Create a sandbox for trying out the latest algorithms on the latest data, and analyze the results together



Remove friction

Contribute in seconds, not days
Use infrastructure, tools that
scientists already use

Organized body of compatible
and actionable scientific data
and tools

Easy, organised communication

Track who did what, give credit



Millions of real, open datasets are generated

- *Drug activity, gene expressions, astronomical observations, text,...*

Extensive toolboxes exist to analyse data

- *MLR, SKLearn, RapidMiner, KNIME, WEKA, AmazonML, AzureML,...*

Massive amounts of experiments are run, most of this information is lost forever (in people's heads)

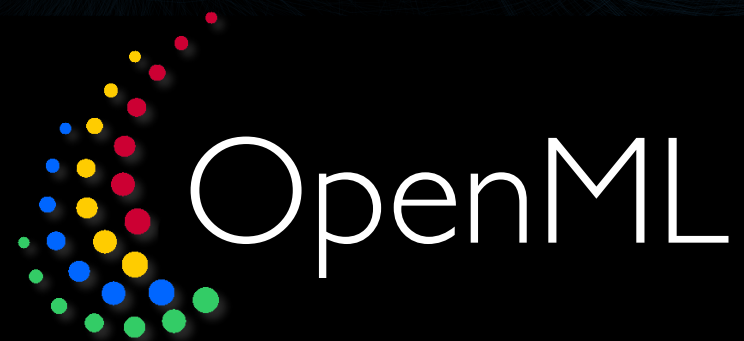
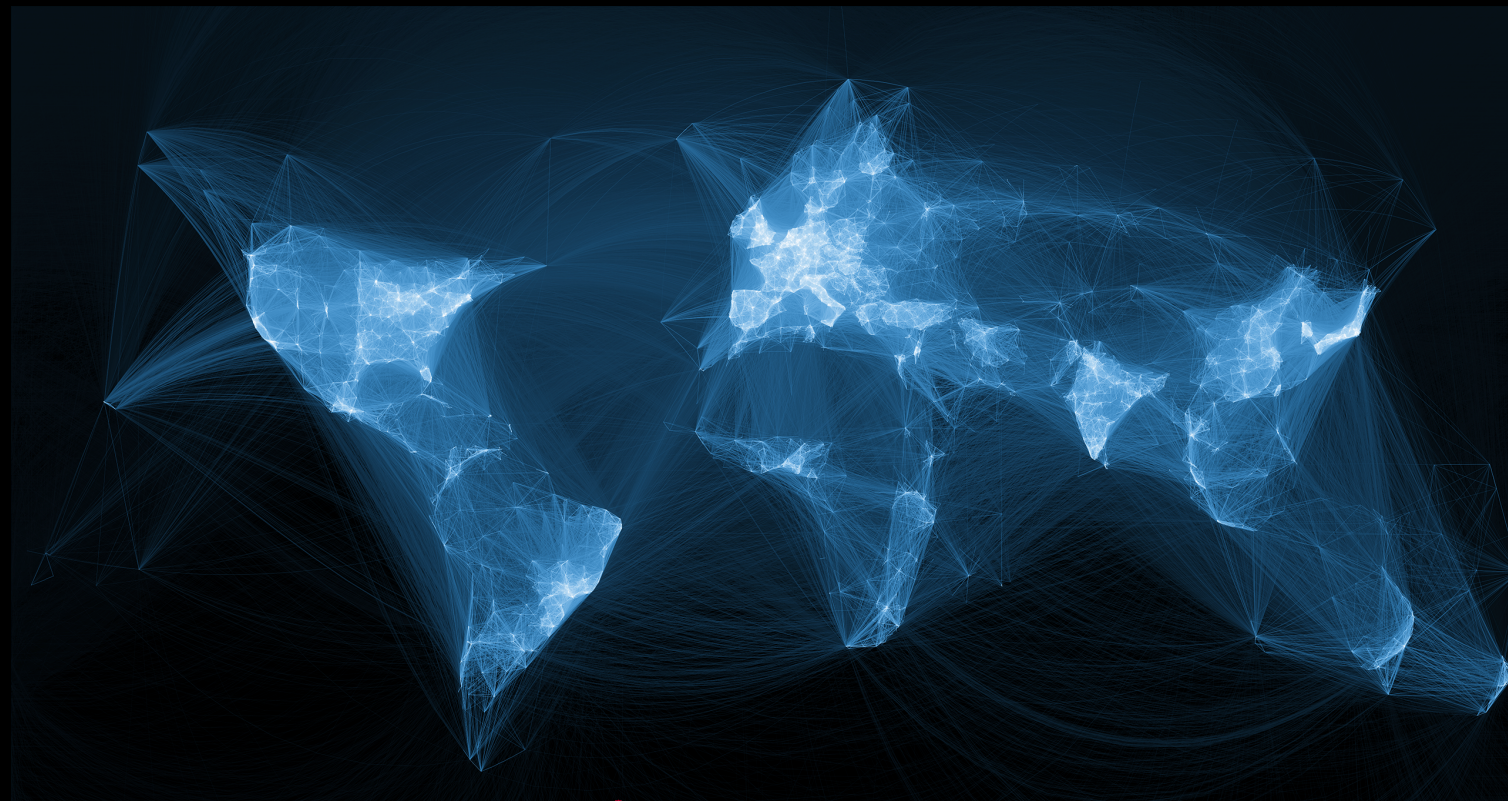
- *No learning across people, labs, fields*
- *Start from scratch, slows research and innovation*

Let's connect machine learning environments to network, so that we can organize, learn from experience

Connect minds, collaborate globally in real time

Train algorithms to automate data science





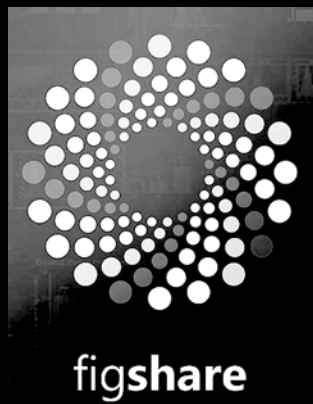
FRICITIONLESS, NETWORKED MACHINE LEARNING

Easy to use: Integrated in ML environments. Automated, reproducible sharing

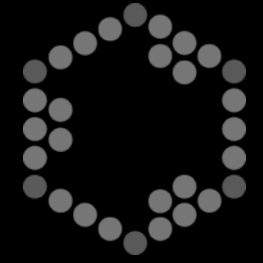
Organized data: Experiments connected to data, code, people anywhere

Easy to contribute: Post single dataset, algorithm, experiment, comment

Reward structure*: Build reputation and trust (e-citations, social interaction)



zenodo



mldata.org
machine learning data set repository

kaggle

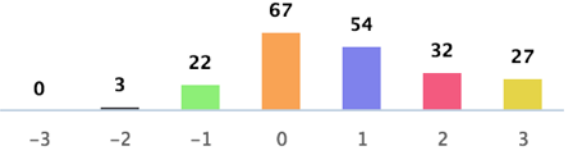
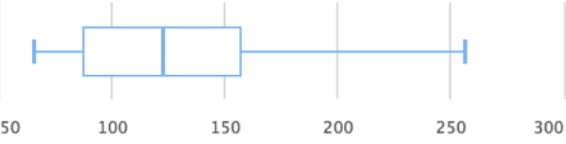
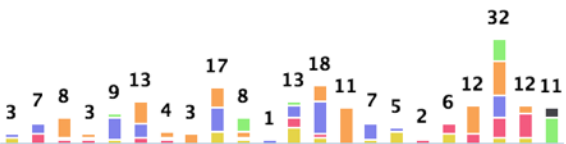


Data (ARFF) uploaded or referenced, versioned
analysed, characterized, organised online








analysed, characterized, 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

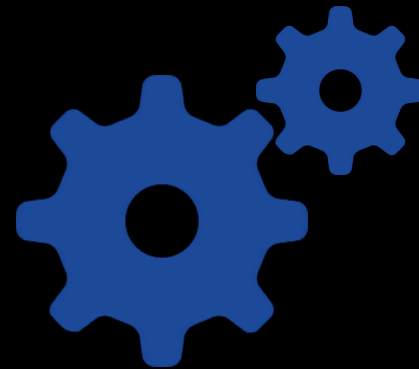
 DefaultAccuracy	0.33	The predictive a
 NumberOfClasses	7	The number of c
 NumberOfFeatures	26	The number of f
 NumberOfInstances	205	The number of i
 NumberOfMissingVal...	59	Counts the total



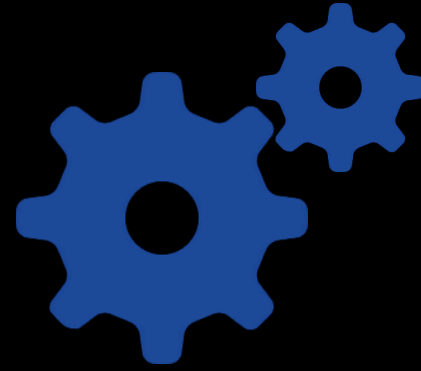
Tasks contain data, goals, procedures.

Readable by tools, automates experimentation

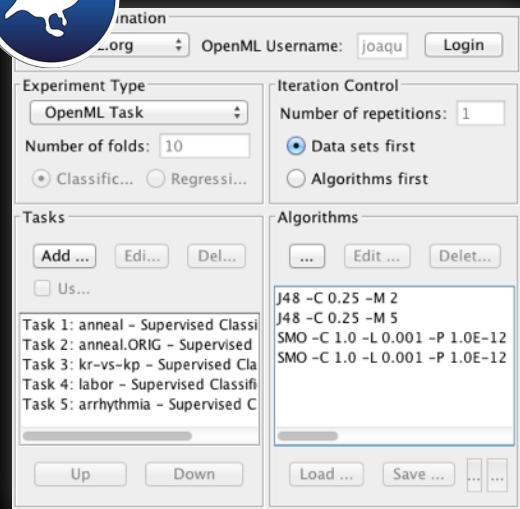
Results organized online: **realtime overview**



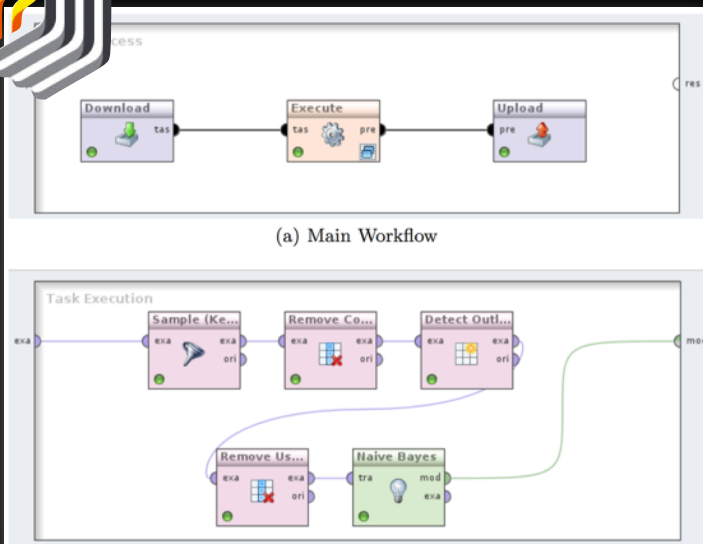
Flows (code) run locally, registered + versioned.
Integrations + APIs (REST, Java, R, Python,...)



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



```
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(target=dataset.default_target)
clf = ensemble.RandomForestClassifier()
clf.fit(X, y)
```



```
library(OpenML); library(mlr)
```

```
task = getOMLTask(task.id = 1L)
lrn = makeLearner("classif.randomForest")
run.mlr = runTaskMlr(task, lrn)
run.id = uploadOMLRun(run.mlr)
```




Experiments auto-uploaded, evaluated online
reproducible, linked to **data**, **flows** and **authors**



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.

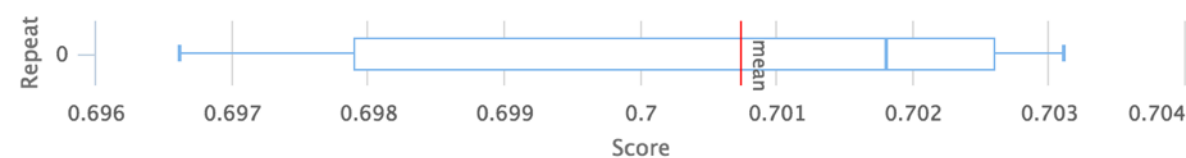
Area under ROC curve

0.7007 \pm 0.0023

Per class

0	1
0.7007	0.7007

Cross-validation details (10-fold Crossvalidation)



Explore

Reuse

Share





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




- Search by keywords or properties
- Filters
- Tagging




The screenshot shows a data search interface with a green header bar containing a menu icon, the word "Data", and a search bar. On the left, a "Filter results" sidebar is open, listing various filter categories: "Number of instances", "Number of features", "Number of missing values", "Number of classes", "Default accuracy", "Uploader", and "Tag". A green "SEARCH" button is at the bottom of the sidebar, along with an information icon and the text "You can use 1..10, >10,..." and a trash icon with the text "Remove all filters". The main area displays "1317 results" and a list of datasets. Each dataset entry includes a database icon, the dataset name with a count in parentheses, and a brief description followed by statistics in green text: "runs", "instances", and "features".






Dataset Name	Count	Statistics
iris	(1)	3816 runs - 150 instances - 5 features
credit-a	(1)	2874 runs - 690 instances - 16 features
anneal.ORIG	(1)	2613 runs - 898 instances - 39 features
diabetes	(1)	2606 runs - 768 instances - 9 features
colic	(1)	2451 runs - 368 instances - 28 features
anneal	(2)	2434 runs - 898 instances - 39 features
mfeat-zernike	(1)	2321 runs - 2000 instances - 48 features
mfeat-morphological	(1)	2317 runs - 2000 instances - 7 features
solar-flare	(2)	2254 runs - 1066 instances - 13 features


Data

- Wiki-like descriptions
- Analysis and visualisation of features

☰ Data Search   

 autos   V. 1 ▾

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

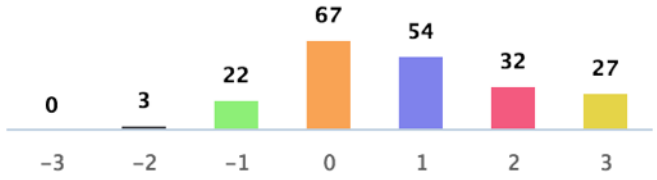
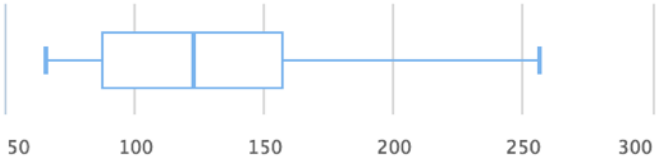
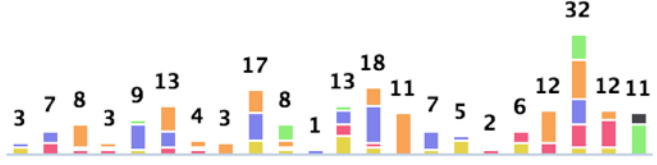
Help us complete this description →  Edit

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](#)

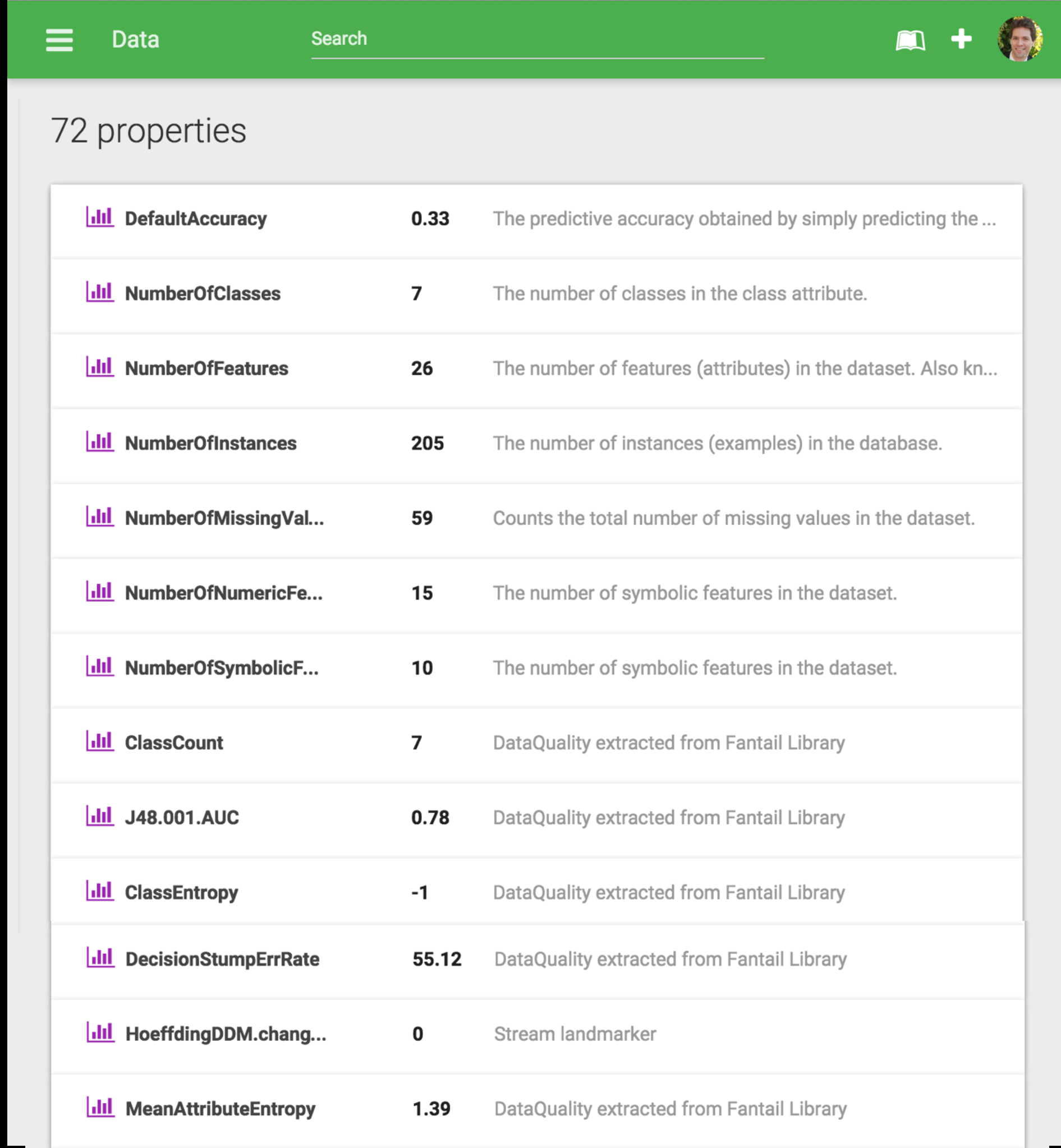
26 features

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](#)

Data

- Wiki-like descriptions
- Analysis and visualisation of features
- Auto-calculation of large range of meta-features
 - discover similar datasets
 - learn across datasets

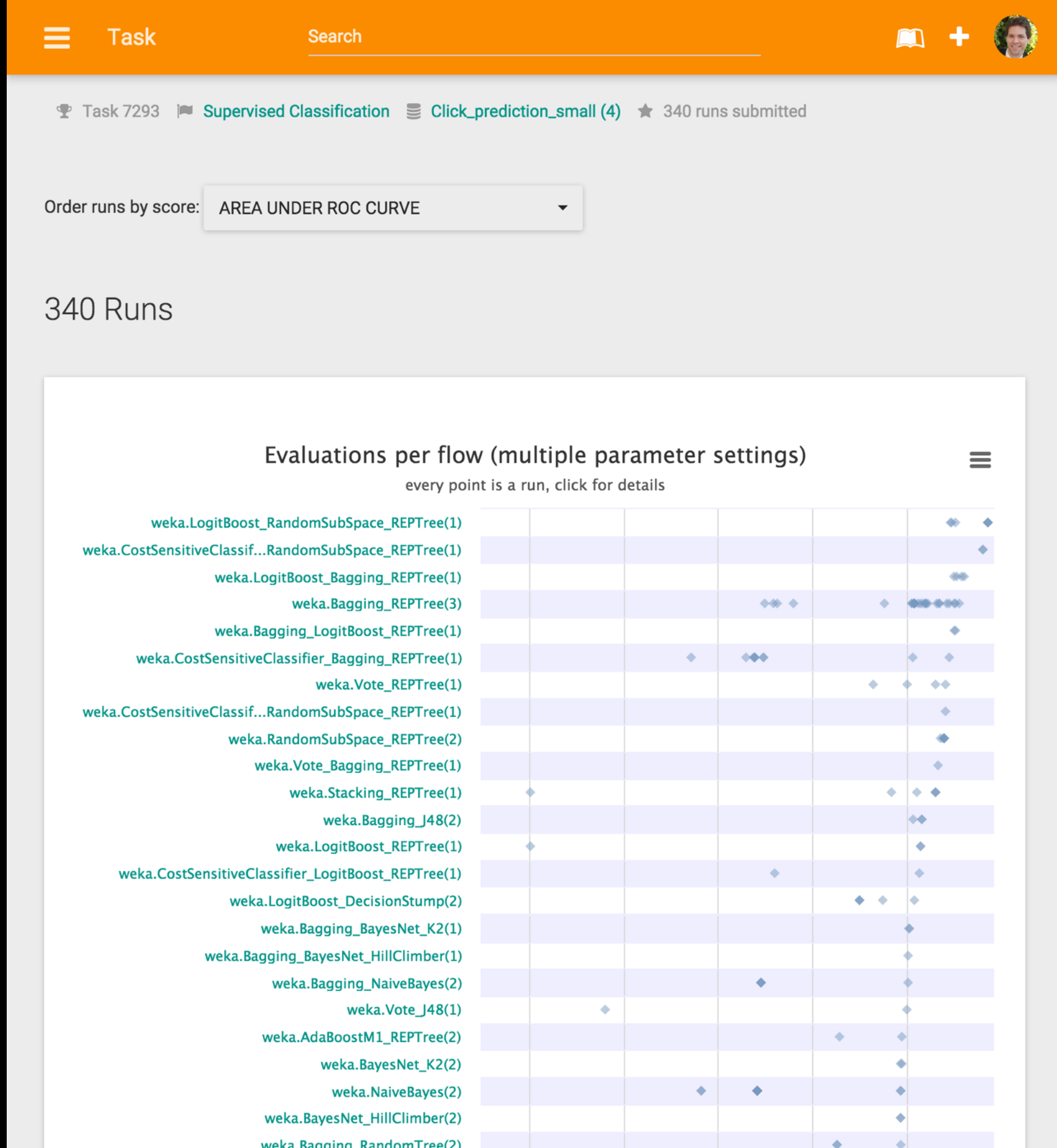


72 properties

 DefaultAccuracy	0.33	The predictive accuracy obtained by simply predicting the ...
 NumberOfClasses	7	The number of classes in the class attribute.
 NumberOfFeatures	26	The number of features (attributes) in the dataset. Also kn...
 NumberOfInstances	205	The number of instances (examples) in the database.
 NumberOfMissingVal...	59	Counts the total number of missing values in the dataset.
 NumberOfNumericFe...	15	The number of symbolic features in the dataset.
 NumberOfSymbolicF...	10	The number of symbolic features in the dataset.
 ClassCount	7	DataQuality extracted from Fantail Library
 J48.001.AUC	0.78	DataQuality extracted from Fantail Library
 ClassEntropy	-1	DataQuality extracted from Fantail Library
 DecisionStumpErrRate	55.12	DataQuality extracted from Fantail Library
 HoeffdingDDM.chang...	0	Stream landmarker
 MeanAttributeEntropy	1.39	DataQuality extracted from Fantail Library

Tasks

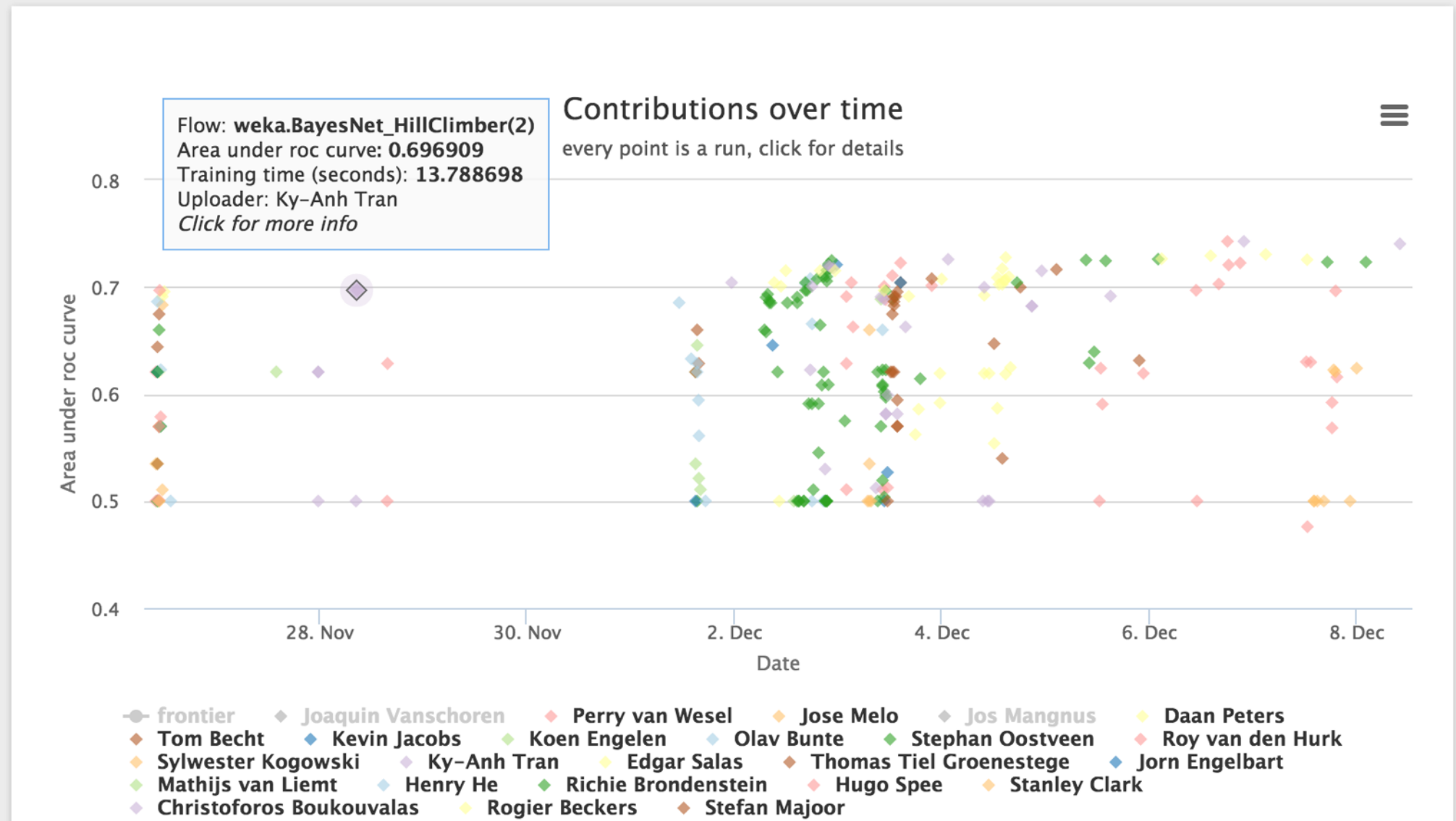
- Example: Classification on click prediction dataset, using 10-fold CV and AUC
- People submit results (e.g. predictions)
- Server-side evaluation (many measures)
- All results organized online, per algorithm, parameter setting
- Online visualizations: every dot is a run plotted by score





Timeline

- Details**
- Overview
- All runs
- Results
- Leaderboard**
- Discuss
- Tags
- Add tag



- 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
- Real-time: clear who submitted first, others can improve immediately

Classroom challenges



Rogier Beckers

@RogierBeckers



[Follow](#)

Het bewijs dat ik studeer op zondag!

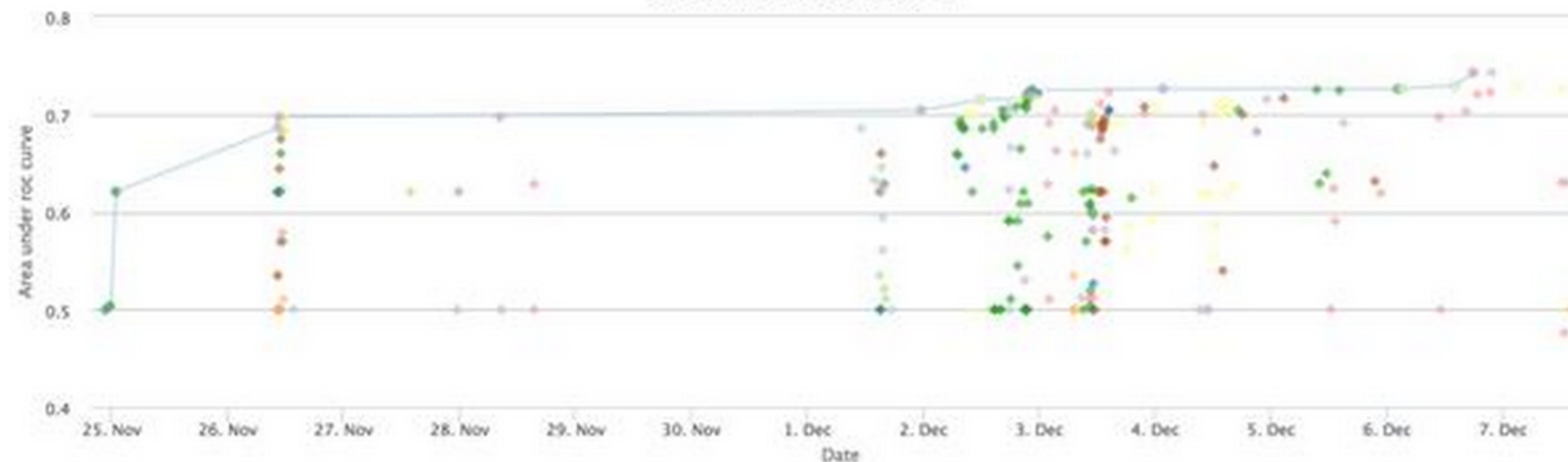
“@joavanschoren: #Machinelearning students on a #collaborative data mining”

[View translation](#)

[Lauradorp, Landgraaf](#)



Contributions over time
every point is a run, click for details



- frontier
- Olav Bunte
- Jorn Engelbart
- Stefan Majoor
- Joaquin Vanschoren
- Stephan Oostveen
- Mathijs van Liemt
- Perry van Wesel
- Roy van den Hurk
- Henry He
- Jose Melo
- Sylwester Kogowski
- Richie Brondenstein
- Jos Mangnus
- Ky-Anh Tran
- Hugo Spee
- Daan Peters
- Edgar Salas
- Stanley Clark
- Tom Becht
- Kevin Jacobs
- Christoforos Boukouvalas
- Koen Engelen
- Thomas Tiel Groenestege
- Rogier Beckers

RETWEETS

2

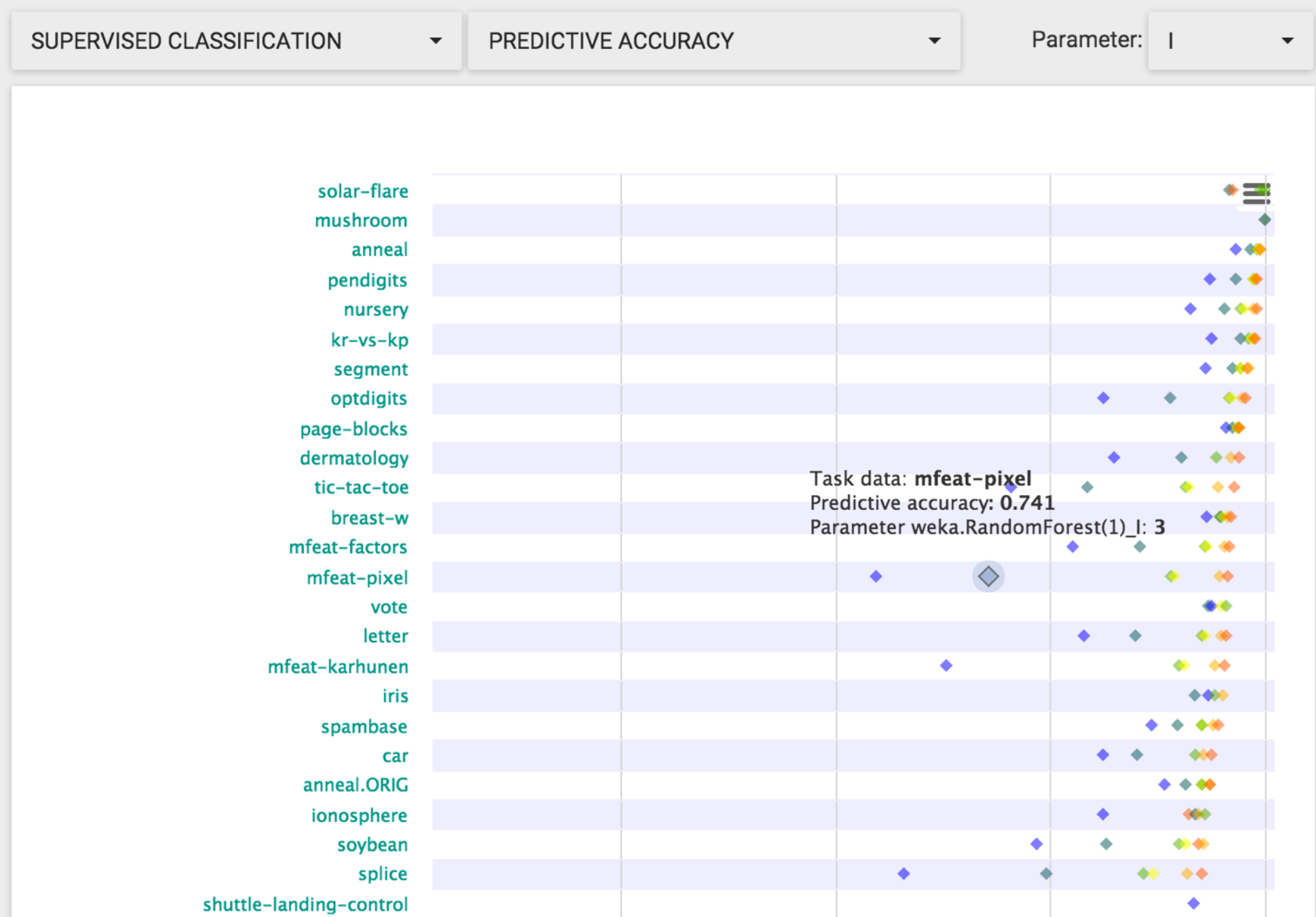
FAVORITES

2



9:48 PM - 7 Dec 2014

- 📄 Data
- 🏆 Tasks
- ⚙️ Flows
- ★ Runs
- 🧪 Task Types
- 📊 Measures
- 👥 People
- 📖 Guide
- 💬 Discussions
- ❤️ Blog
- ⌆
- 📄 Details
- Overview
- Download flow



- 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)

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

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

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)

Explore

Reuse

Share



R API

Idem for Java, Python

Tutorial: <http://www.openml.org/guide>

List datasets

```
datasets = listOMLDataSets() # returns active data sets
datasets[1:3, 3:6]
```

##	name	NumberOfClasses	NumberOfFeatures	NumberOfInstances
## 1	anneal	6	39	898
## 2	anneal	6	39	898
## 3	kr-vs-kp	2	37	3196

List flows

```
flows = listOMLFlows()
flows[1:7, 1:2]
```

##	implementation.id	full.name
## 1	1	openml.evaluation.EuclideanDistance(1.0)
## 2	2	openml.evaluation.PolynomialKernel(1.0)
## 3	3	openml.evaluation.RBFKernel(1.0)
## 4	4	openml.evaluation.area_under_roc_curve(1.0)
## 5	5	openml.evaluation.average_cost(1.0)
## 6	6	openml.evaluation.build_cpu_time(1.0)
## 7	7	openml.evaluation.build_memory(1.0)

List tasks

```
tasks = listOMLTasks()
tasks[1:6, 1:5]
```

##	task_id	task_type	did	status	name
## 1	1	Supervised Classification	1	active	anneal
## 2	2	Supervised Classification	2	active	anneal
## 3	3	Supervised Classification	3	active	kr-vs-kp
## 4	4	Supervised Classification	4	active	labor
## 5	5	Supervised Classification	5	active	arrhythmia
## 6	6	Supervised Classification	6	active	letter

List runs and results

```
runs = listOMLRuns(task.id = 59L) # must be re
head(runs)
```

```
runresults = listOMLRunResults(task.id = 59L)
colnames(runresults)
```

R API

Idem for Java, Python

Tutorial: <http://www.openml.org/guide>

Download datasets

```
iris.data2 = getOMLDataSet(did = 61L) # the iris data set has
iris.data2
```

```
##
## Data Set "iris" :: (Version = 1, OpenML ID = 61)
##   Collection Date      : 1936
##   Creator(s)          : R.A. Fisher
##   Default Target Attribute: class
```

Download flows

```
flow = getOMLFlow(flow.id = 1248L)
flow
```

```
##
## Flow "classif.randomForest" :: (Version = 1, Flow ID = 124
##   External Version      : 4.6-10
##   Dependencies         : mlr_2.3, randomForest_4.6.10
##   Number of Flow Parameters: 12
##   Number of Flow Components: 0
```

Download tasks

```
task = getOMLTask(task.id = 59L)
task
```

```
##
## OpenML Task 59 :: (Data ID = 61)
##   Task Type           : Supervised Classification
##   Data Set            : iris :: (Version = 1, OpenML ID = 61)
##   Target Feature(s)  : class
##   Estimation Procedure : Stratified crossvalidation (1 x 10 f
```

```
iris.data = task$input$data.set$data
head(iris.data)
```

```
##   sepallength sepalwidth petallength petalwidth
## 0           5.1         3.5         1.4         0.2 Iri
## 1           4.9         3.0         1.4         0.2 Iri
## 2           4.7         3.2         1.3         0.2 Iri
## 3           4.6         3.1         1.5         0.2 Iri
## 4           5.0         3.6         1.4         0.2 Iri
## 5           5.4         3.9         1.7         0.4 Iri
```


R API

Idem for Java, Python

Tutorial: <http://www.openml.org/guide>

Download run

```
run = getOMLRun(run.id = 234L)
```

Download run with predictions

```
run.pred = getOMLRun(run.id = 234L, get.predictions = TRUE)  
all.equal(run.pred$predictions, getOMLPredictions(run))
```


Explore

Reuse

Share



WEKA

OpenML extension
in plugin manager

The screenshot shows the configuration window for the OpenML extension in WEKA. It is divided into several sections:

- Results Destination:** A dropdown menu set to "OpenML.org", an "OpenML Username" field containing "joaqu", and a "Login" button.
- Experiment Type:** A dropdown menu set to "OpenML Task", a "Number of folds" field set to "10", and two radio buttons: "Classific..." (selected) and "Regressi...".
- Iteration Control:** A "Number of repetitions" field set to "1", and two radio buttons: "Data sets first" (selected) and "Algorithms first".
- Tasks:** Three buttons: "Add ...", "Edi...", and "Del...". A checkbox "Us..." is below them. A list box contains five tasks:
 - Task 1: anneal - Supervised Classi
 - Task 2: anneal.ORIG - Supervised
 - Task 3: kr-vs-kp - Supervised Cla
 - Task 4: labor - Supervised Classifi
 - Task 5: arrhythmia - Supervised CBelow the list are "Up" and "Down" buttons.
- Algorithms:** Three buttons: "...", "Edit ...", and "Delet...". A list box contains four algorithm configurations:
 - J48 -C 0.25 -M 2
 - J48 -C 0.25 -M 5
 - SMO -C 1.0 -L 0.001 -P 1.0E-12
 - SMO -C 1.0 -L 0.001 -P 1.0E-12Below the list are "Load ...", "Save ...", and two empty buttons.

MOA

Classification | Regression | Clustering | Outliers | Concept Drift

Configure Run

command	status	time elapsed	current activi...	% complete
openml.OpenmlDataStreamClassification -l trees.HoeffdingAdaptiveTree -t 192	completed	55.16s		100.00
openml.OpenmlDataStreamClassification -l trees.HoeffdingAdaptiveTree -t 191	completed	44.69s		100.00
openml.OpenmlDataStreamClassification -l trees.HoeffdingAdaptiveTree -t 190	completed	34.00s		100.00
openml.OpenmlDataStreamClassification -l trees.HoeffdingAdaptiveTree -t 189	completed	42.66s		100.00
openml.OpenmlDataStreamClassification -l trees.HoeffdingAdaptiveTree -t 188	completed	1m22s		100.00

Pause | Resume

Final result Refresh

100000.0, 73.56700000000001, 55.566093979989006, -55.5
155525.0, 73.53903230793013, 55.488147109572715, -53.1

Ex

Evaluation

Values

Measure	Current	Mean
Accuracy	73... 81.89	78.78 82.33
Kappa	55... 71.25	64.31 69.10
Kappa Temp	53... 266...	119... 200...
Ram-Hours	0.00 0.00	0.00 0.00
Time	76... 40.59	44.42 23.14

Plot

Zoom

85.00
42.50

Configure task

class moa.tasks.openml.OpenmlDataStreamClassification

Purpose
Evaluates a classifier on an OpenML Data Stream Classification Task.

learner trees.HoeffdingAdaptiveTree Edit

taskId 188

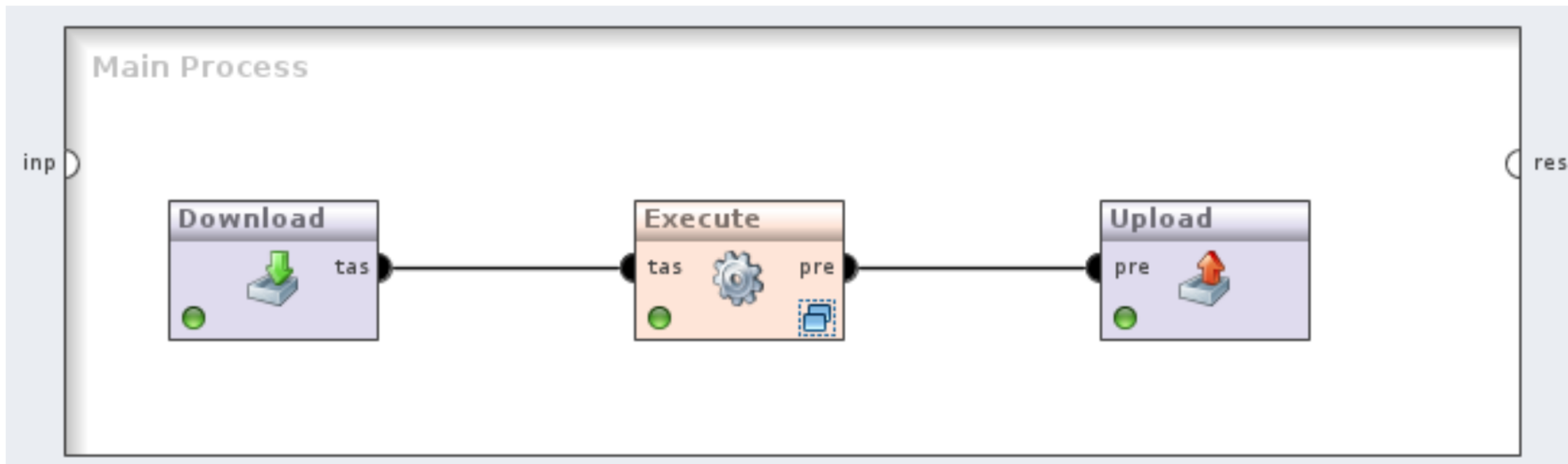
evaluator ssificationPerformanceEvaluator Edit

sampleFrequency 100,000

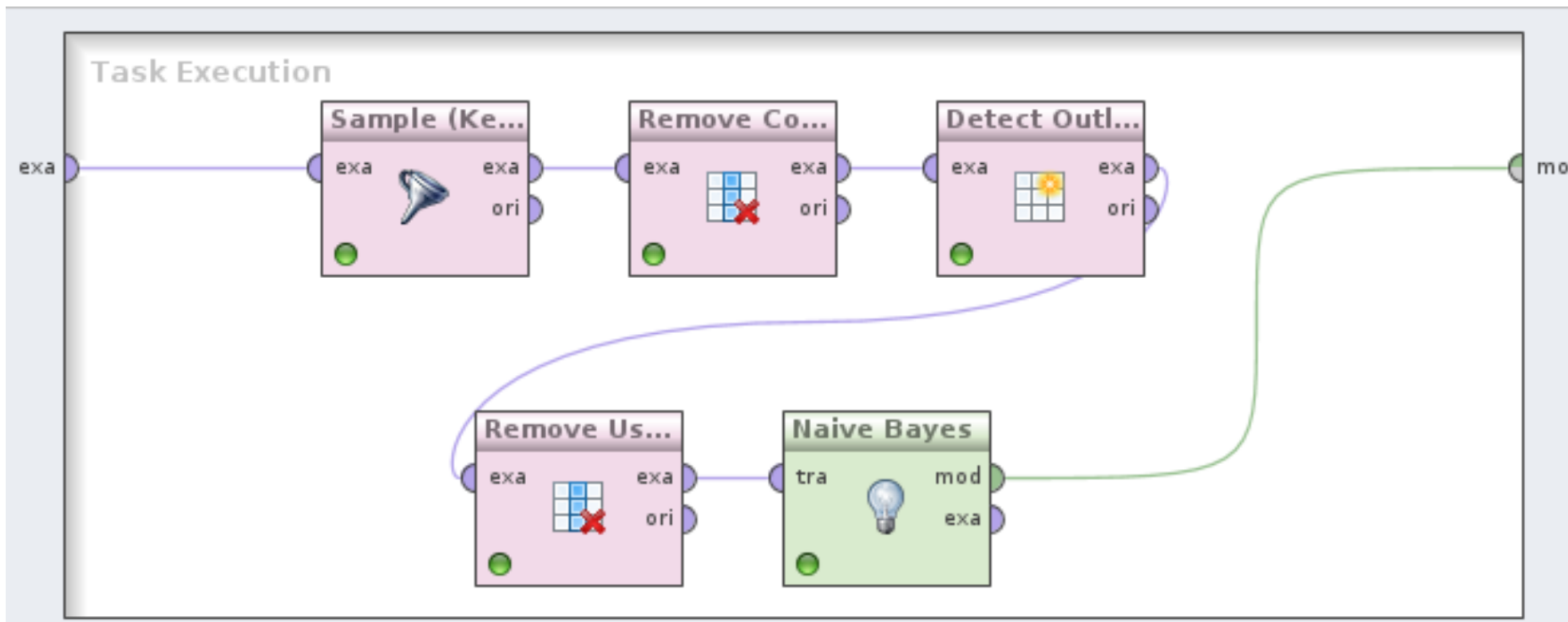
dumpFile Browse

taskResultFile Browse

RapidMiner: 3 new operators



Main Workflow



R API

Run a task
(with mlr):

```
task = getOMLTask(task.id = 59L)  
task
```

```
library(mlr)  
lrn = makeLearner("classif.rpart")  
run.mlr = runTaskMlr(task, lrn)
```

```
run.mlr
```

```
##  
## OpenML Run NA :: (Task ID = 59, Flow ID = NA)  
##  
## Resample Result  
## Task: data  
## Learner: classif.rpart  
## acc.aggr: 0.94  
## acc.mean: 0.94  
## acc.sd: 0.05  
## Runtime: 0.152566
```


R API

Run a task
(with mlr):

```
library(mlr)  
lrn = makeLearner("classif.rpart")  
run.mlr = runTaskMlr(task, lrn)
```

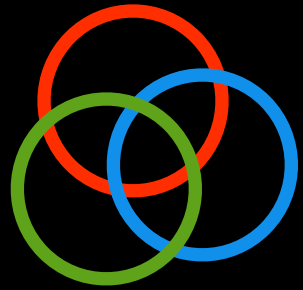
```
run.mlr
```

```
##  
## OpenML Run NA :: (Task ID = 59, Flow ID = NA)  
##  
## Resample Result  
## Task: data  
## Learner: classif.rpart  
## acc.aggr: 0.94  
## acc.mean: 0.94  
## acc.sd: 0.05  
## Runtime: 0.152566
```

And upload:

```
run.id = uploadOMLRun(run.mlr)
```

Collaboration tools (in progress)



Circles

Create collaborations with trusted researchers
Share results within team prior to publication



Studies (e-papers)

- Start a question, invite others (or everyone)
- Online counterpart of a paper, linkable



Reputation

- Auto-track reuse of shared resources (data, code)
- Prove your activity, reach, impact



Notebooks

- Easy sharing, collaboration on scripts

Data science, differently



Change scale

- Invite anyone / everyone to work with your data
- Global organization: state-of-the-art online
- Discover interesting people, data, code,...



Change speed

- Easy data/code reuse, automated sharing
- Real-time collaboration (seconds, not days)



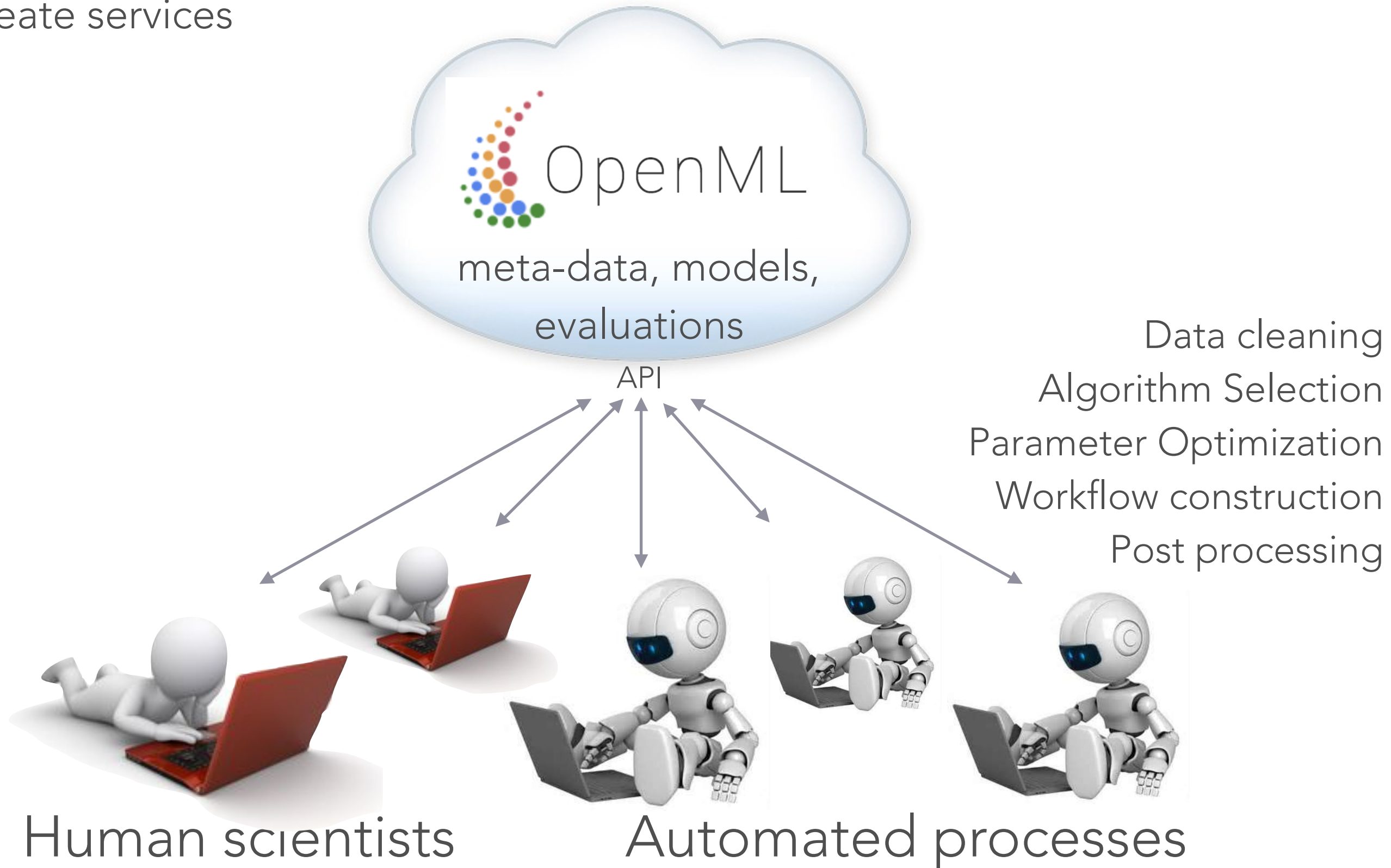
Automation

- Bots that automatically run algorithms on data
- Discover similar datasets, do basic analysis
- Learn from lots of data: select/optimize algorithms

AutoML: automating machine learning

Learn from many datasets and experiments how to do data analysis

Create services

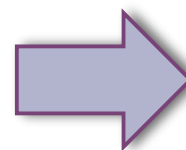


OpenML in drug discovery

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

The screenshot displays the ChEMBL website interface. At the top, the ChEMBL logo and Wellcome Trust branding are visible. The main content area is titled 'Target Report Card' for target CHEMBL3227. It lists target details such as 'Target ID', 'Target Type', 'Preferred Name', 'Synonyms', 'Organism', and 'Species Group'. Below this, there is a 'Target Components' section and a 'Target Associated Bioactivities' section. A search bar is present, and below it, a table shows 'ChEMBL Target Search Results: 23'. The table columns include ChEMBL ID, Preferred Name, UniProt Accession, Target Type, Organism, Compounds, and Bioactivities. A pie chart on the left side of the search results section shows 'ChEMBL Activity' with a total of 3379.

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
							⋮				

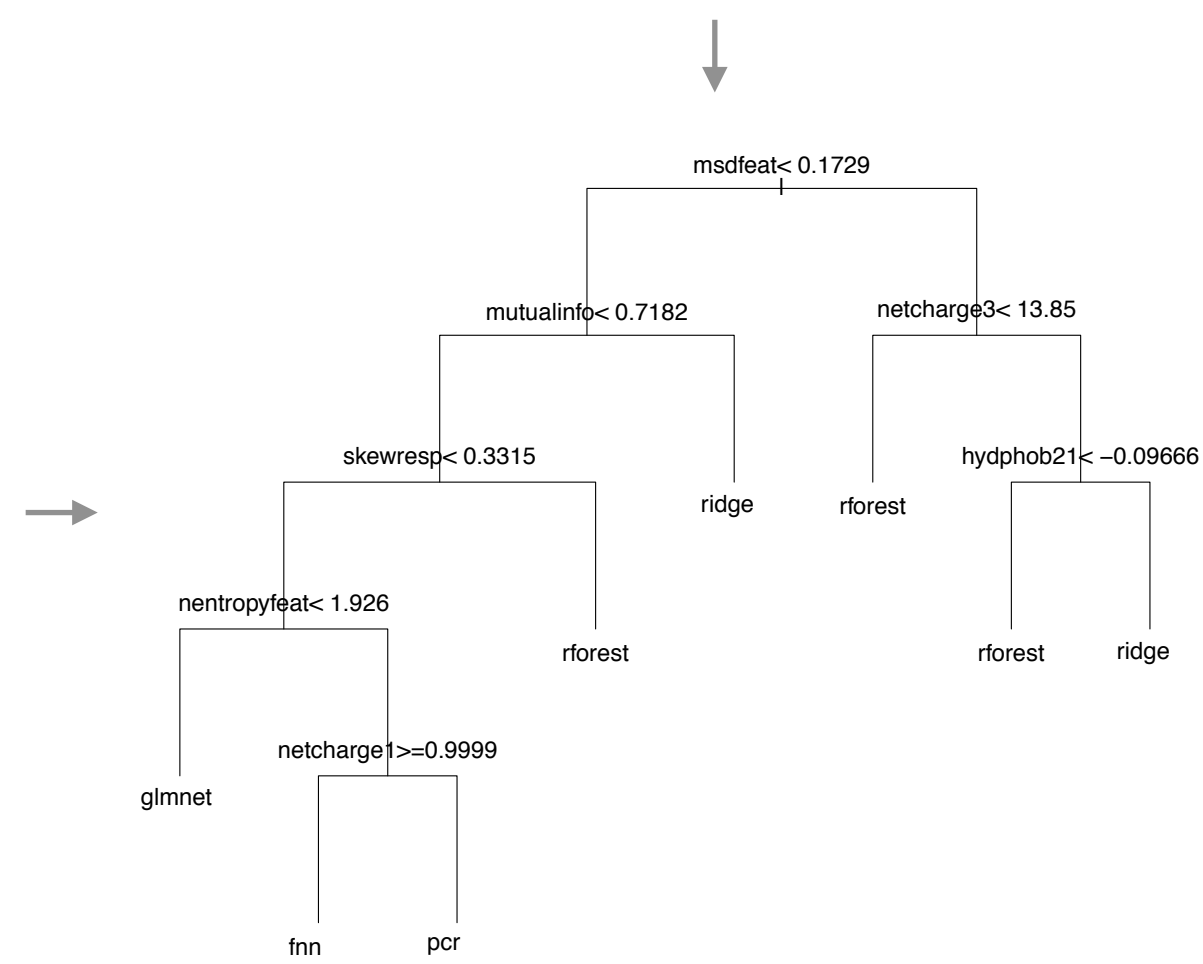
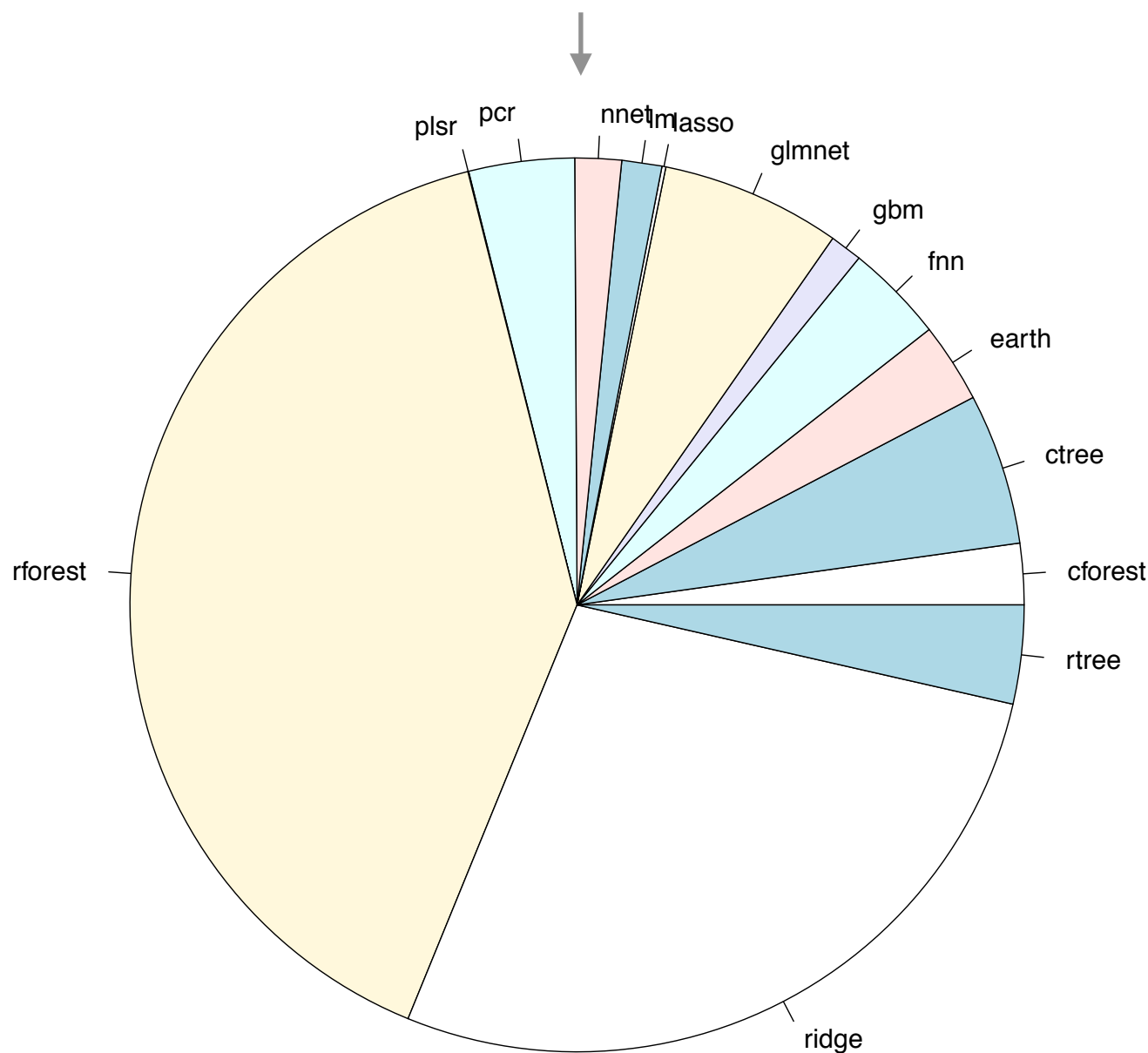
10.000+ regression datasets
2750 targets have >10 compounds, x 4 fingerprints

OpenML in drug discovery

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
							⋮				

Metafeatures:

- simple, statistical, info-theoretic, landmarks
- target: aliphatic index, hydrophobicity, net charge, mol. weight, sequence length, ...



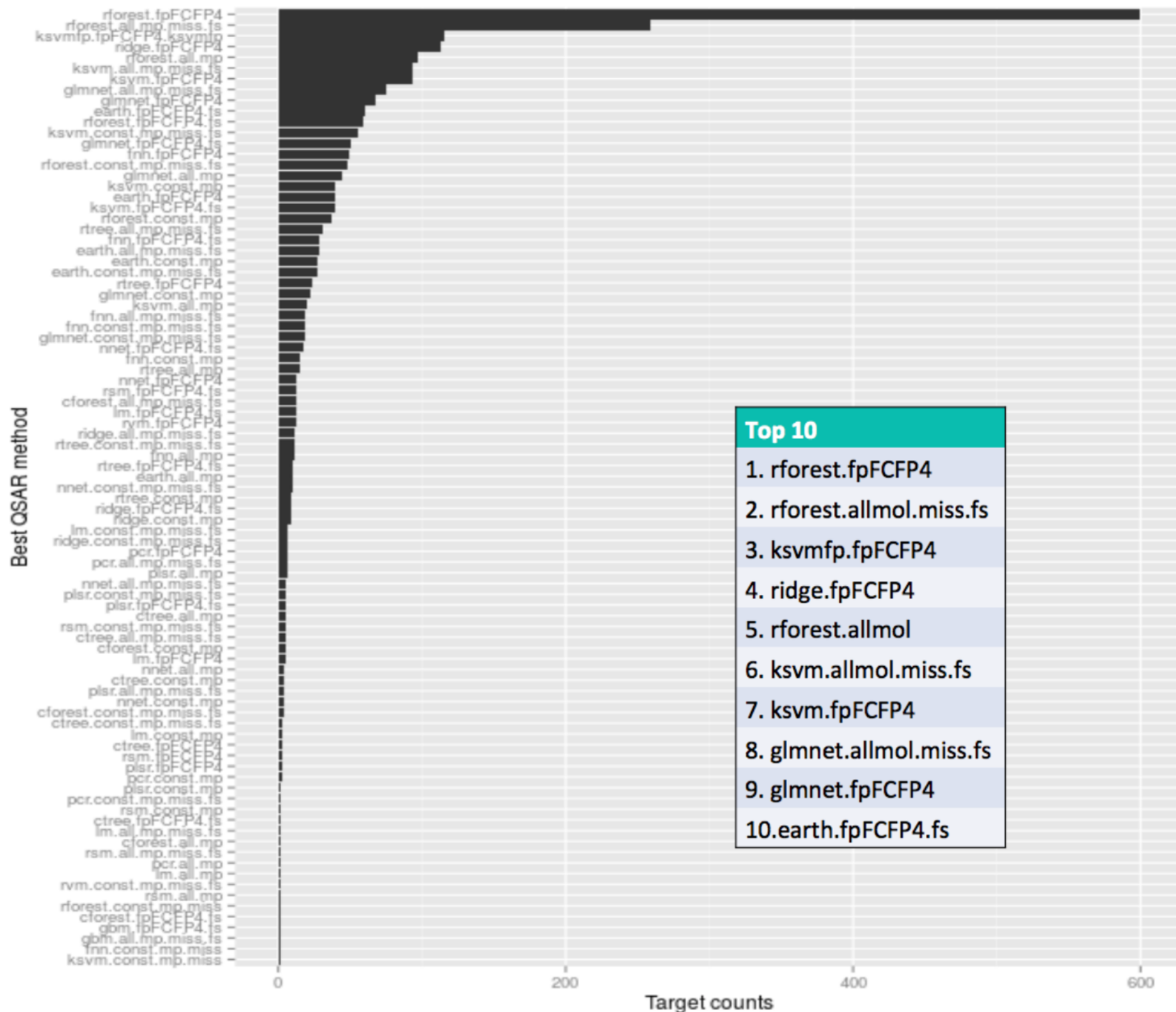
ECFP4_1024	FCFP4_1024	ECFP6_1024	FCFP6_1024
0.697	0.427	0.725	0.627

Predict best algorithm with meta-models

OpenML in drug discovery

Best algorithms?

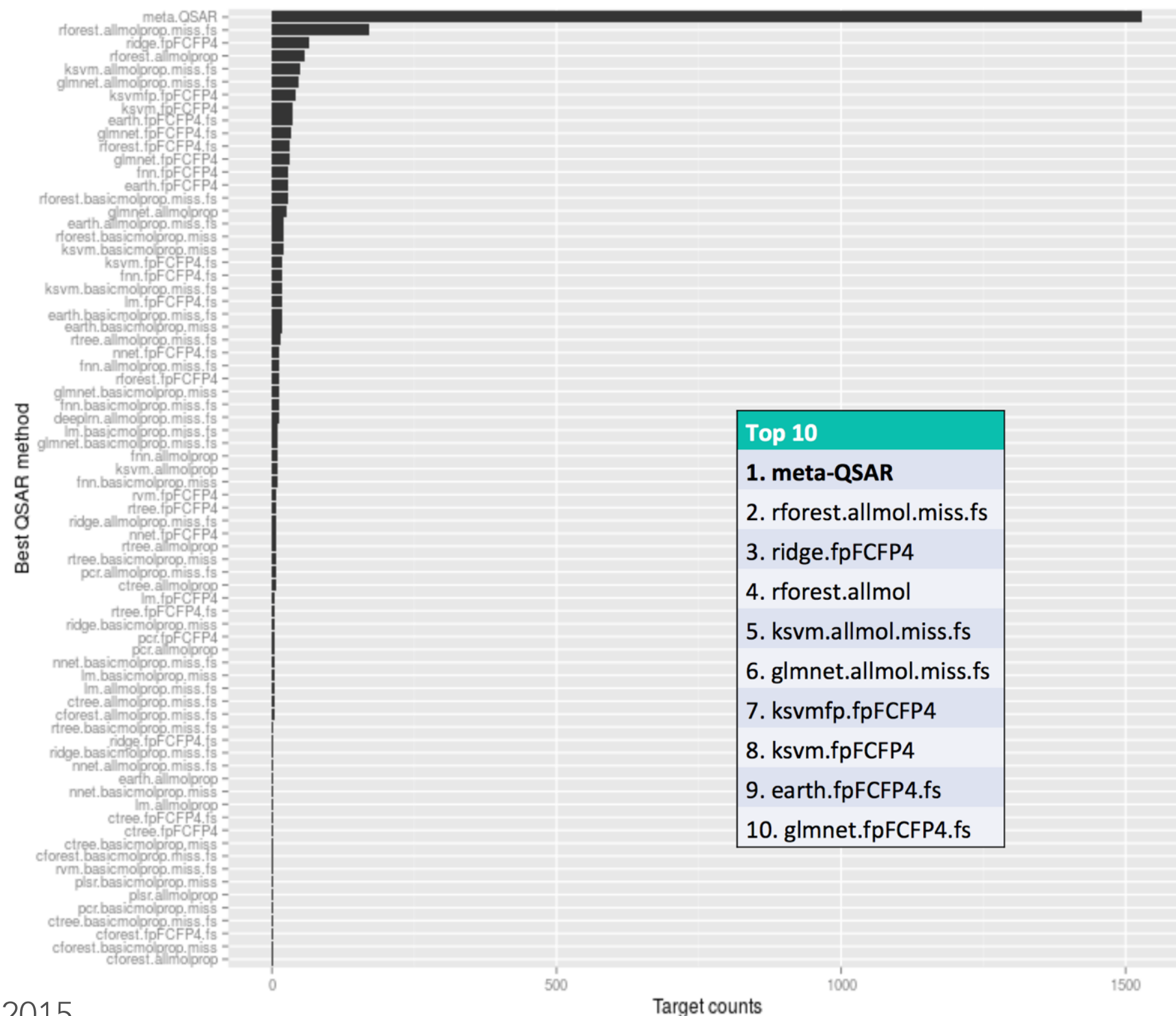
Best features?



OpenML in drug discovery

New technique: stacked generalisation

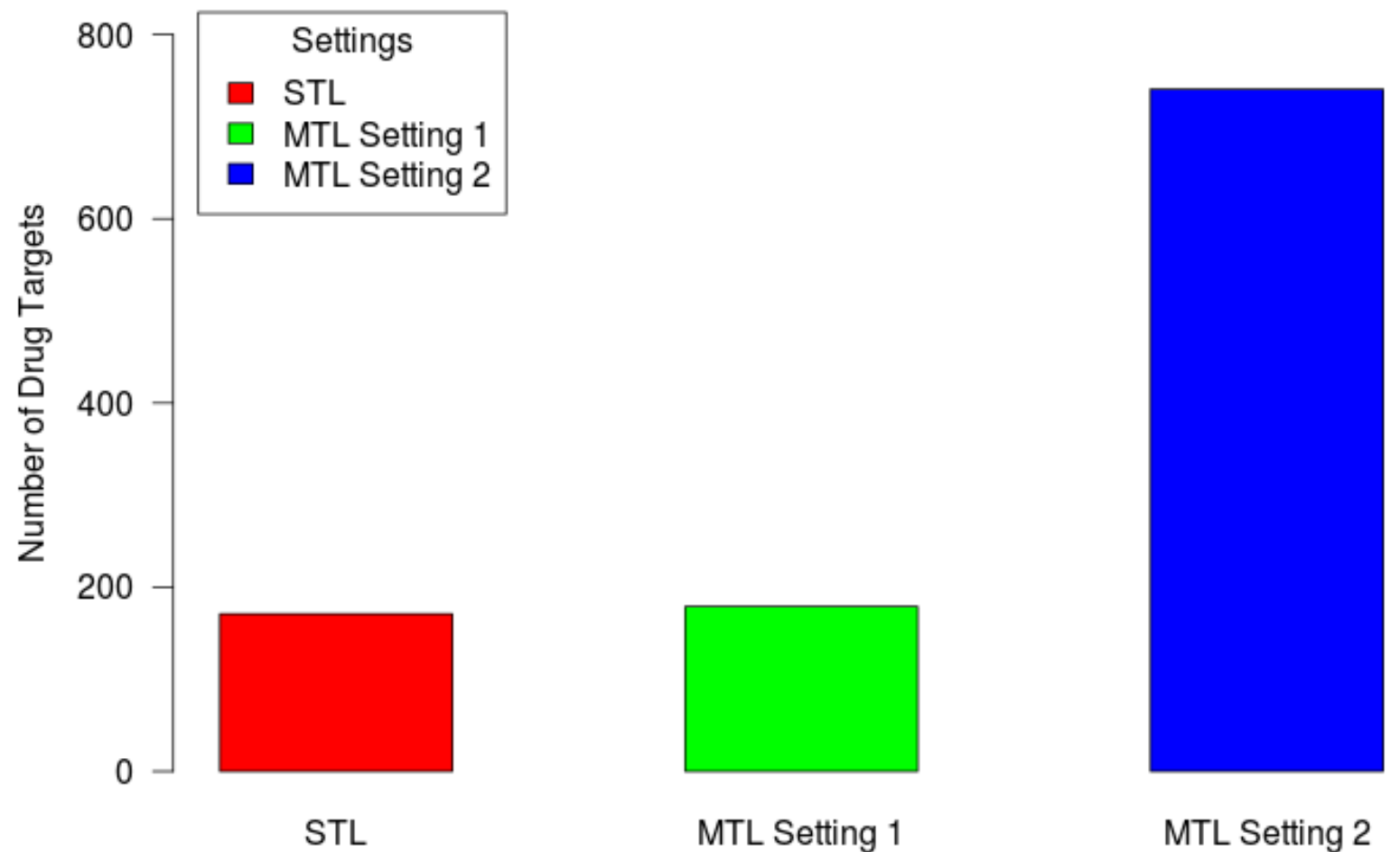
Best model by far



OpenML in drug discovery

New technique: multi-target learning

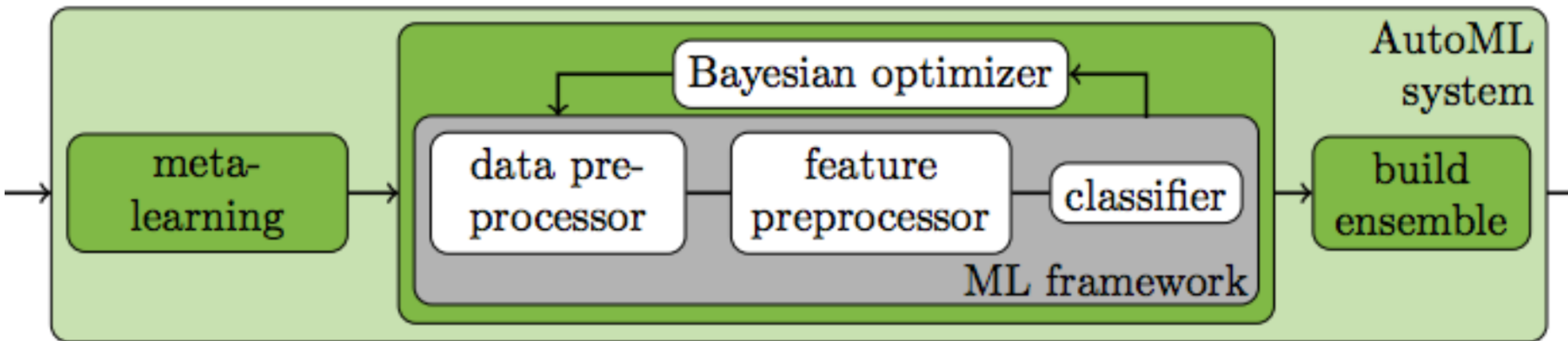
When few drugs are tested on a given target, combine with the data on 'related' targets (same family, similar gene sequence)



We just scratched the surface. Data will be available on OpenML for many more studies and ideas, to test new algorithms,...

Automating ML

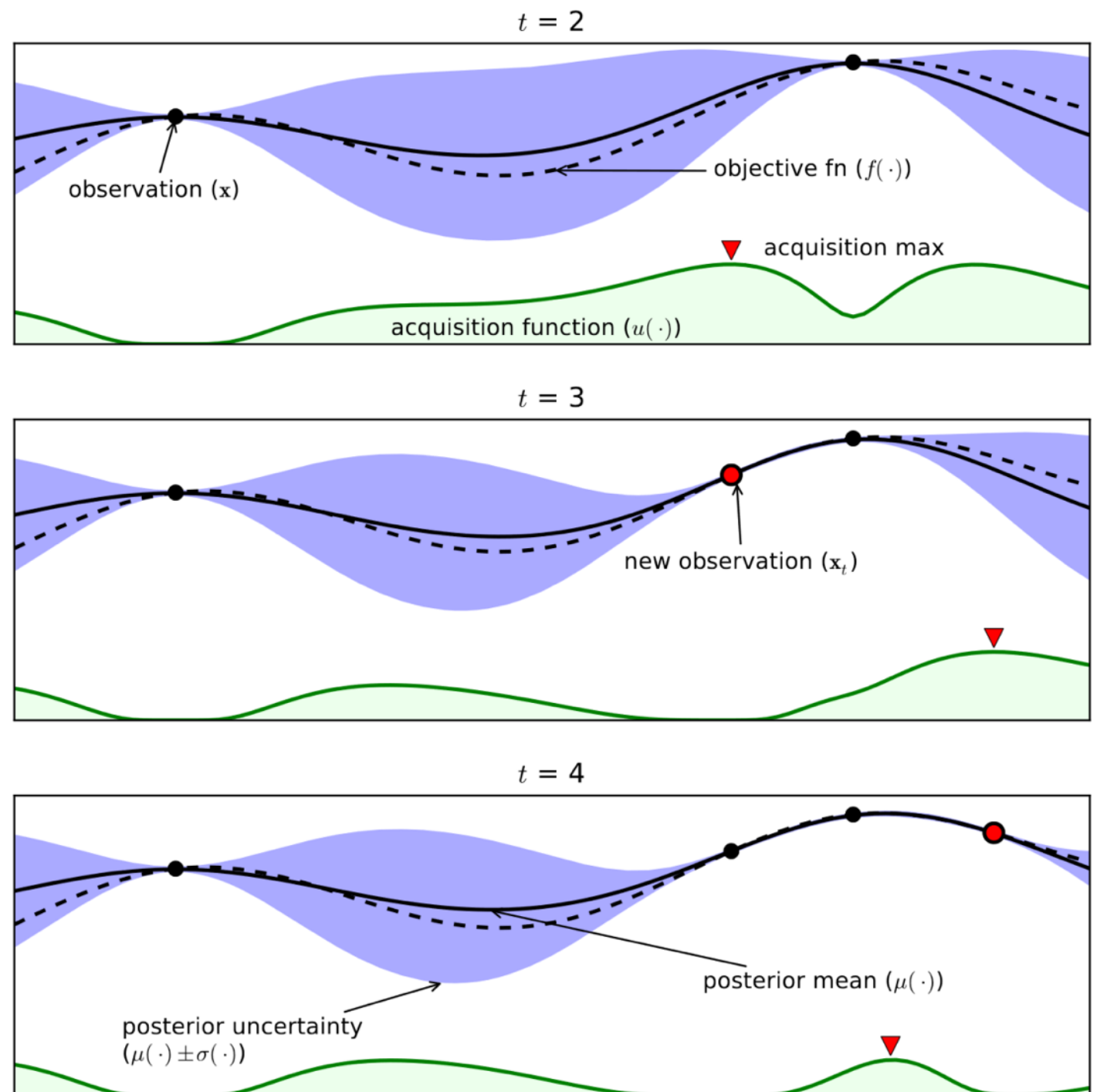
AutoML challenge: winning solution used OpenML and meta-learning



We just scratched the surface. Data will available on OpenML for many more studies and ideas, to test new algorithms,...

Automating ML

Learn parameter space of algorithms over many datasets, include them in acquisition functions.

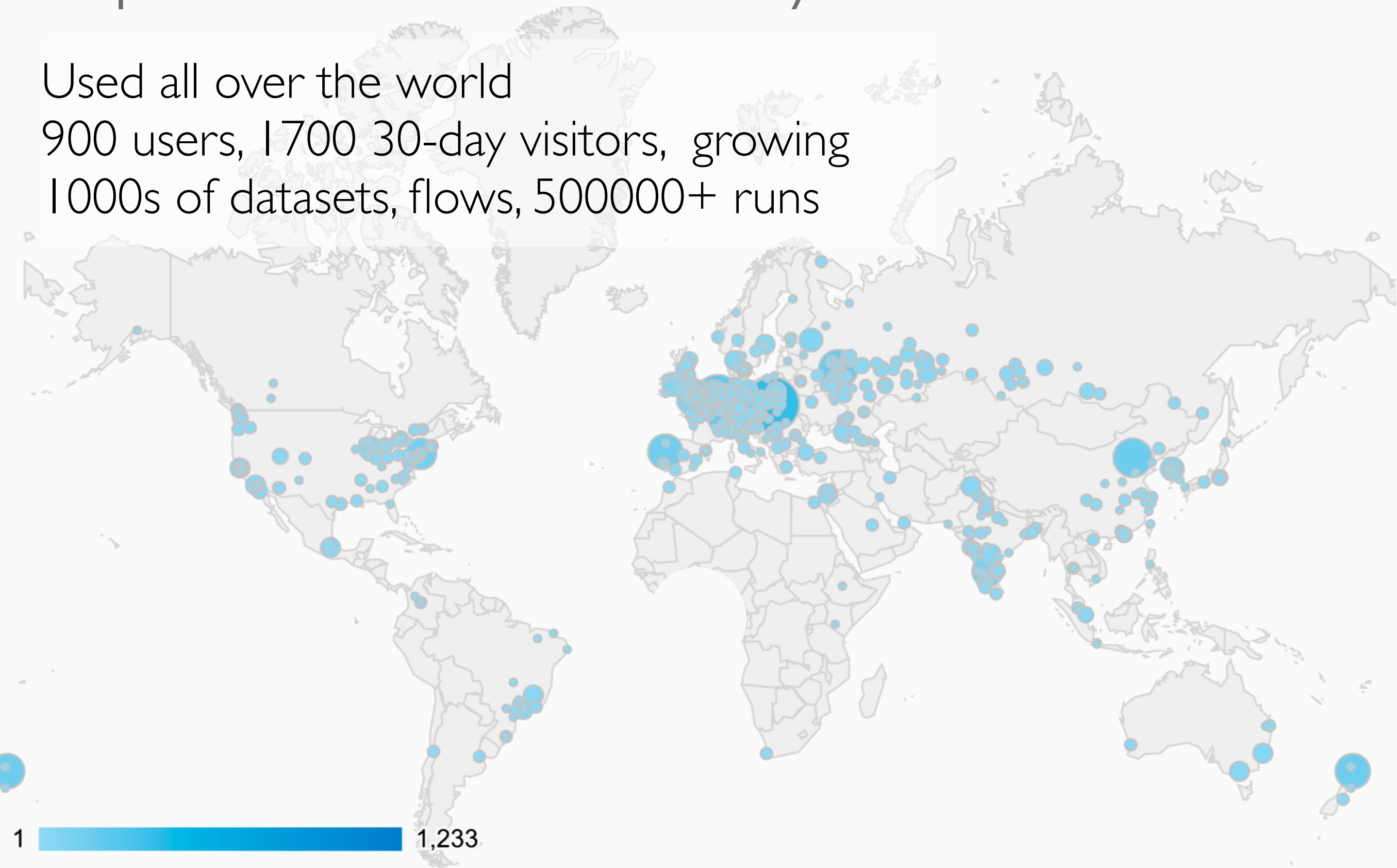


OpenML Community

Used all over the world

900 users, 1700 30-day visitors, growing

1000s of datasets, flows, 500000+ runs



1

1,233

Jan-Jun 2015

Join OpenML

- Open Source, on GitHub
- Regular workshops, hackathons



Next workshop:

- Lorentz Center (Leiden),
14-18 March 2016

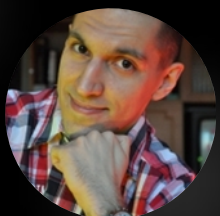
THANK YOU

 #OpenML



Jakob Bossek

Farzan Majdani



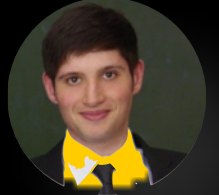
Nenad Tomašev



Luis Torgo



Jan van Rijn



Giuseppe Casalicchio



Joaquin Vanschoren



Michel Lang



Bernd Bischl



Matthias Feurer

You?