#OpenML

OpenML

COLLABORATIVE MACHINE LEARNING

JOAQUIN VANSCHOREN (@JOAVANSCHOREN), TU EINDHOVEN, 2016

GALILEO 1610

smaismrmilmepoetaleumibunenugttauiras

A TALE OF MANY MINDS

BRAHE 1600

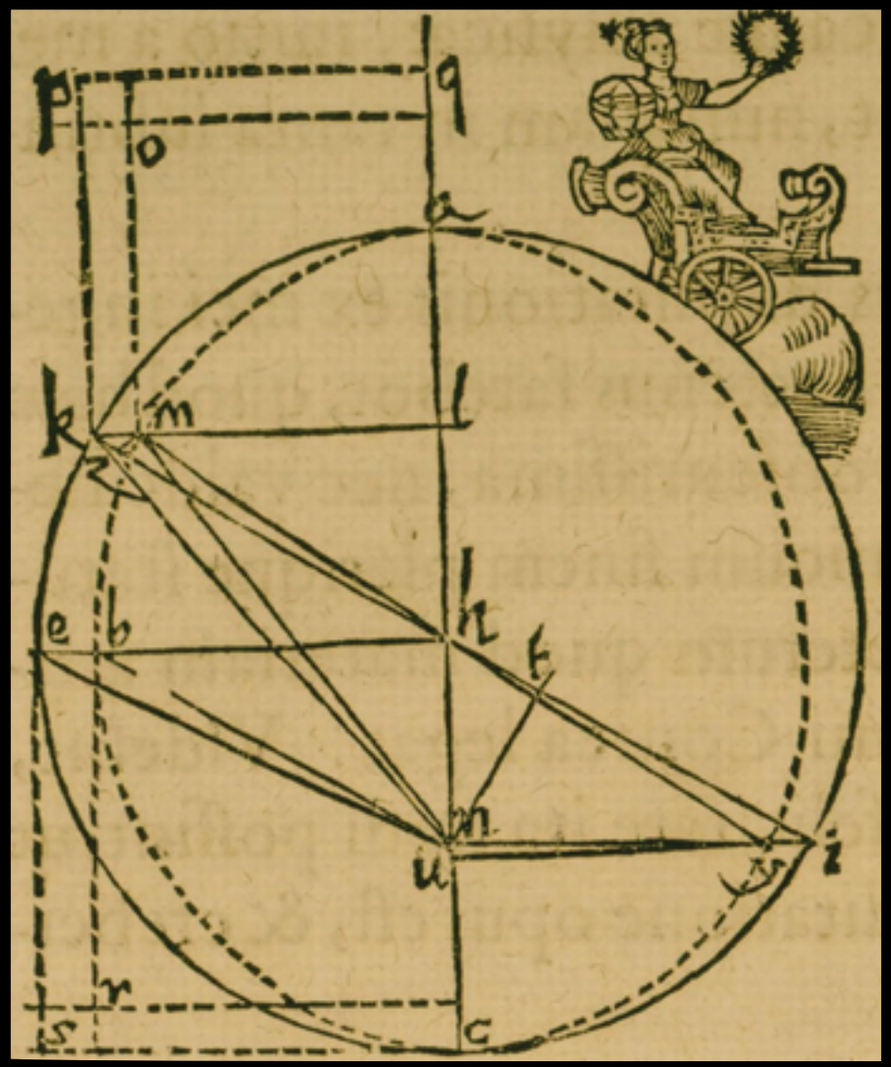
A

PARS QVARTA.

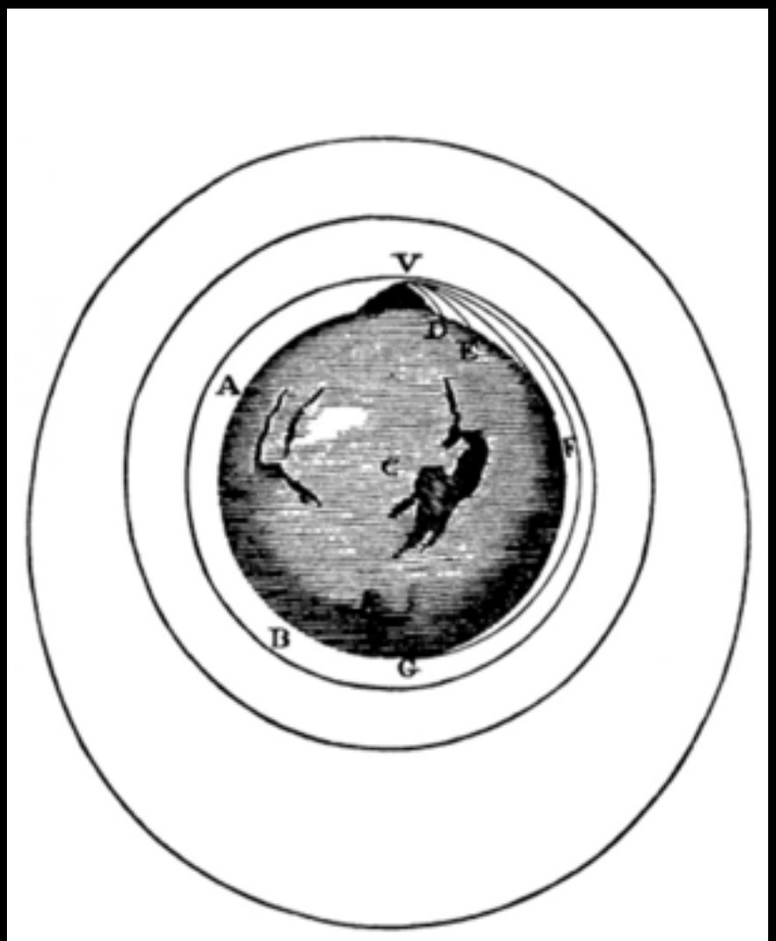
	Tempus	Locus ☉	Solis a Terra distantia	Martis a Sole distantia	
1582.	23 Nove. H. 16. 0	11.41 ♐	98345	158852	
	26 Dece. H. 8.30	15. 4 ♑	98226	162104	1
	30 Dece. H. 8.10	19. 9 ♑	98252	162443	1
1583.	26 Janua. H. 6.15	16.33 ♒	98624	164421	
1584.	21 Dece. H. 14. 0	10.16 ♑	98207	164907	
1585.	24 Janua. H. 9. 0	14.53 ♒	98595	166210	1
	4 Febr. H. 6.40	26.10 ♒	98830	166400	2
	12 Mart. H. 10.30	2.16 ♈	99858	166170	
1587.	25 Janua. H. 17. 0	16. 1 ♒	98611	166232	
	4 Mart. H. 13.24	24. 0 ♓	99595	164737	2
	10 Mart. H. 11.30	29.52 ♓	99780	164382	2
	21 April. H. 9.30	10.48 ♉	101010	161027	1
1589.	8 Mart. H. 16.24	28.36 ♓	99736	161000	1
	13 April. H. 11.15	3.38 ♉	100810	157141	
	15 April. H. 12. 5	5.36 ♉	100866	156900	
	6 Maji. H. 11.20	25.49 ♉	101366	154326	1
1591.	13 Maji. H. 14. 0	2.10 ♊	101467	147891	1
	6 Junii H. 12.20	24.59 ♊	101769	144981	2
	10 Junii H. 11.50	28.47 ♊	101789	144526	2
	28 Junii H. 10.24	15.51 ♋	101770	142608	
1593.	21 Julii H. 14. 0	8.26 ♌	101498	138376	2
	22 Aug. H. 12.20	9.11 ♍	100761	138463	1
	29 Aug. H. 10.20	11.54 ♍	100562	138682	1
	3 Octo. H. 8. 0	20.15 ♎	99500	140697	
1595.	17 Sept. H. 16.45	4.18 ♎	99990	143222	2
	27 Octo. H. 12.20	13.59 ♏	98851	147890	1
	3 Nove. H. 12. 0	21. 2 ♏	98694	148773	1
	18 Dece. H. 8. 0	6.43 ♑	98200	154539	1

KEPLER 1610-1620

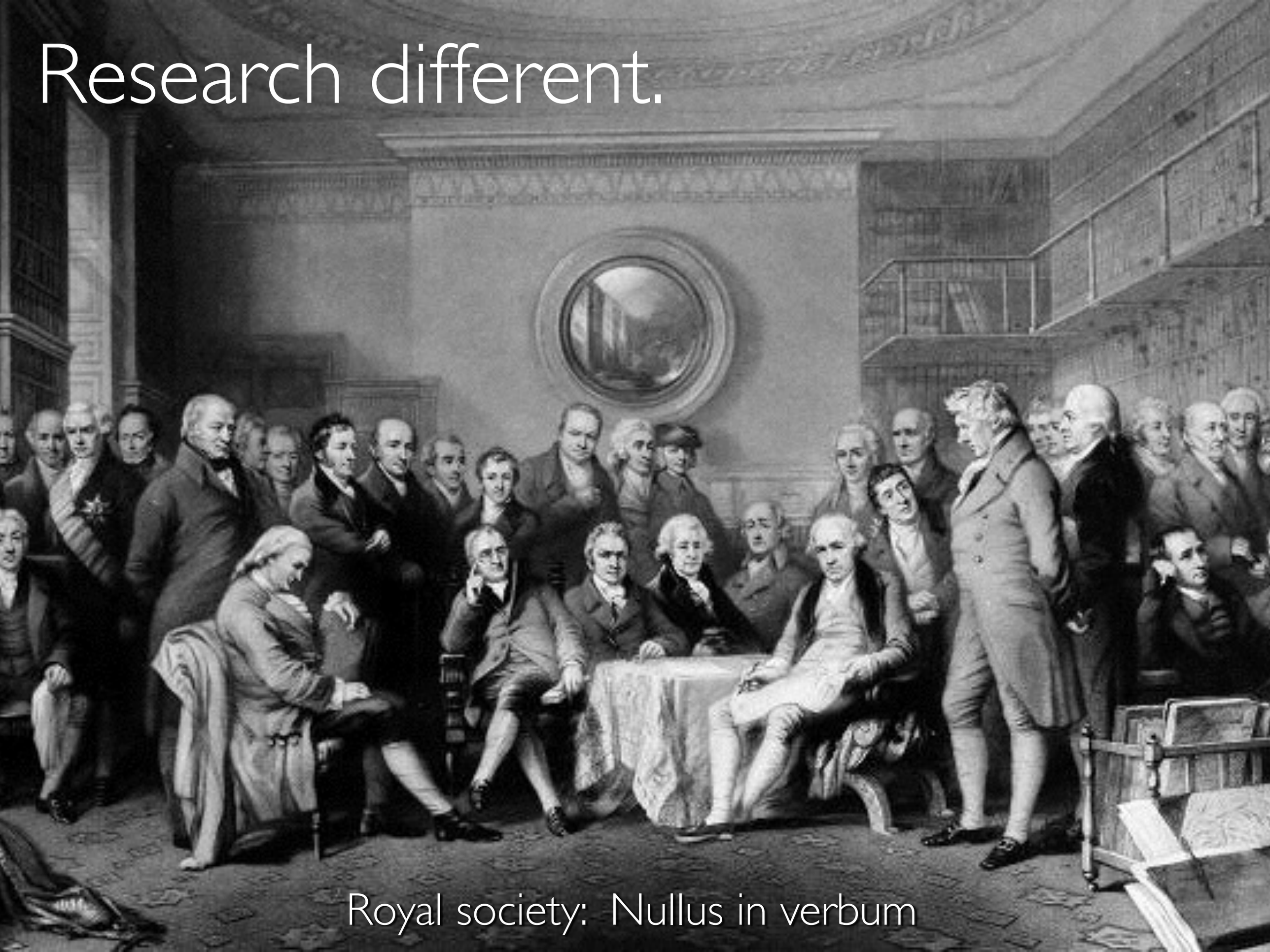
B



NEWTON 1686



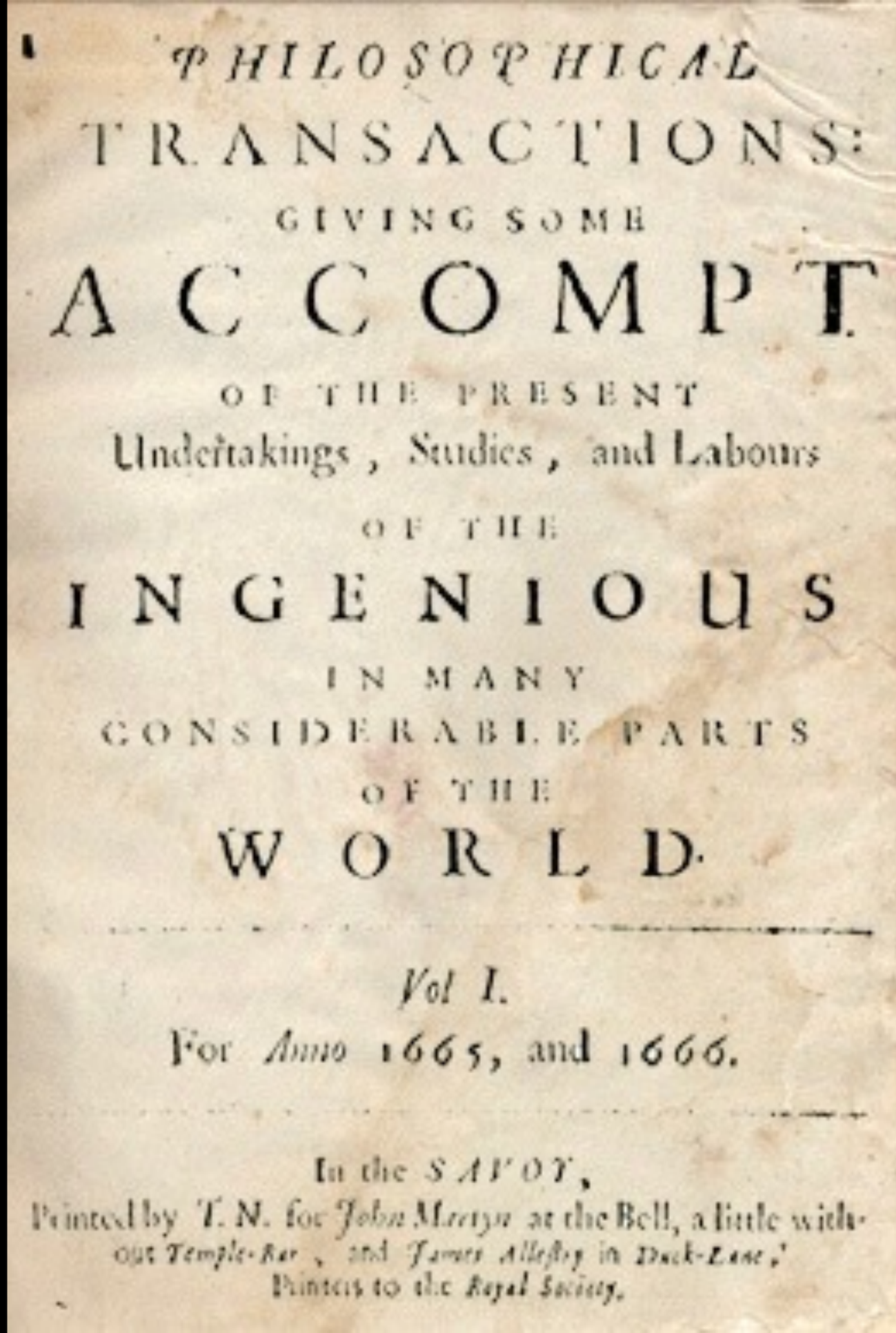
Research different.



Royal society: Nullus in verbum

IS PAPER THE BEST MEDIUM? FOR MACHINE LEARNING?

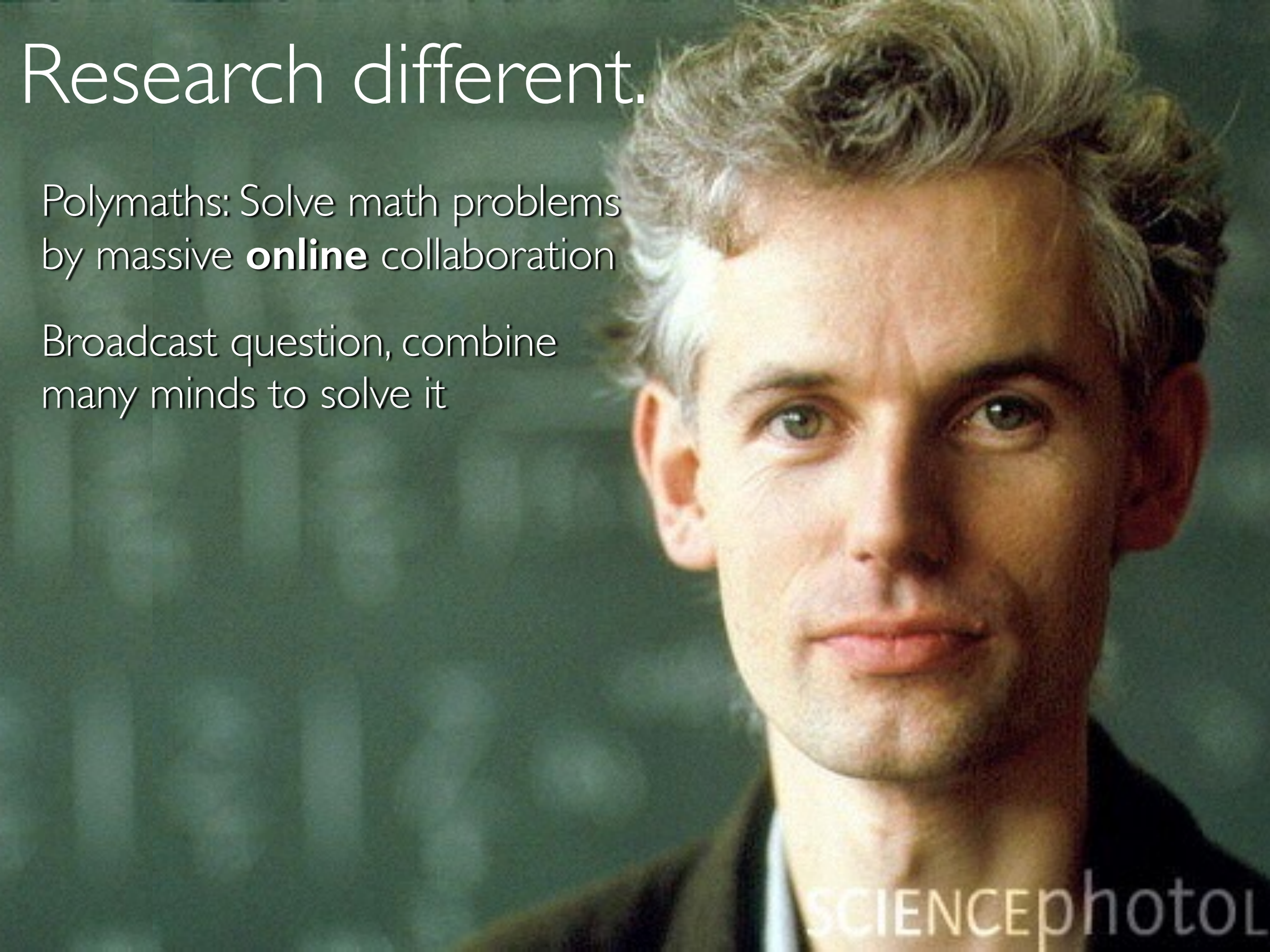
- Data hard to find, reuse
- Code hard to find, reuse
- Results hard to reproduce, compare, reuse
- Publication bias
- Mostly solitary, offline work
- Lot of time wasted on tedious jobs (cleaning, tuning,…)


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.

Research different.

Polymaths: Solve math problems
by massive **online** collaboration

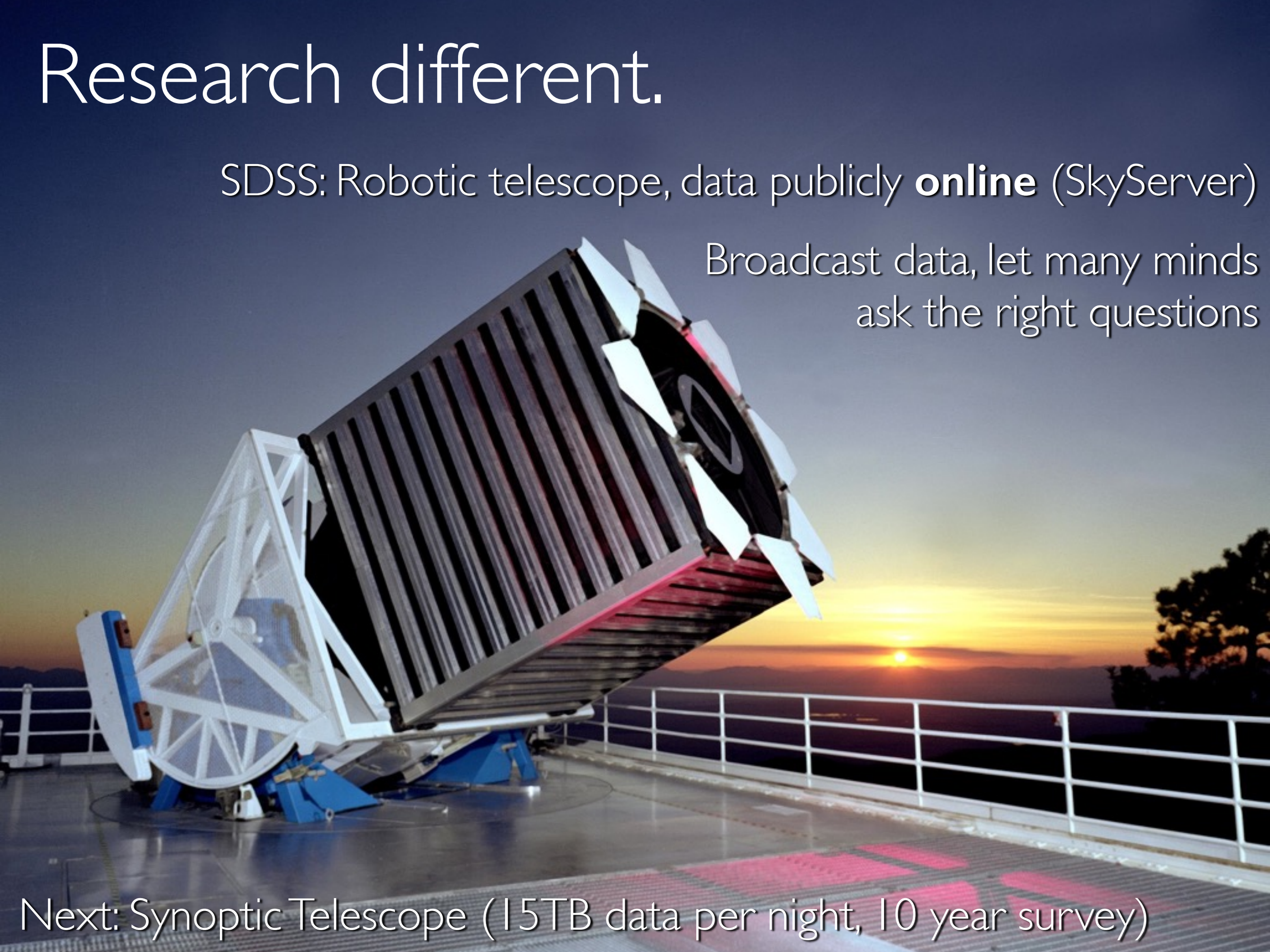
Broadcast question, combine
many minds to solve it



Research different.

SDSS: Robotic telescope, data publicly **online** (SkyServer)

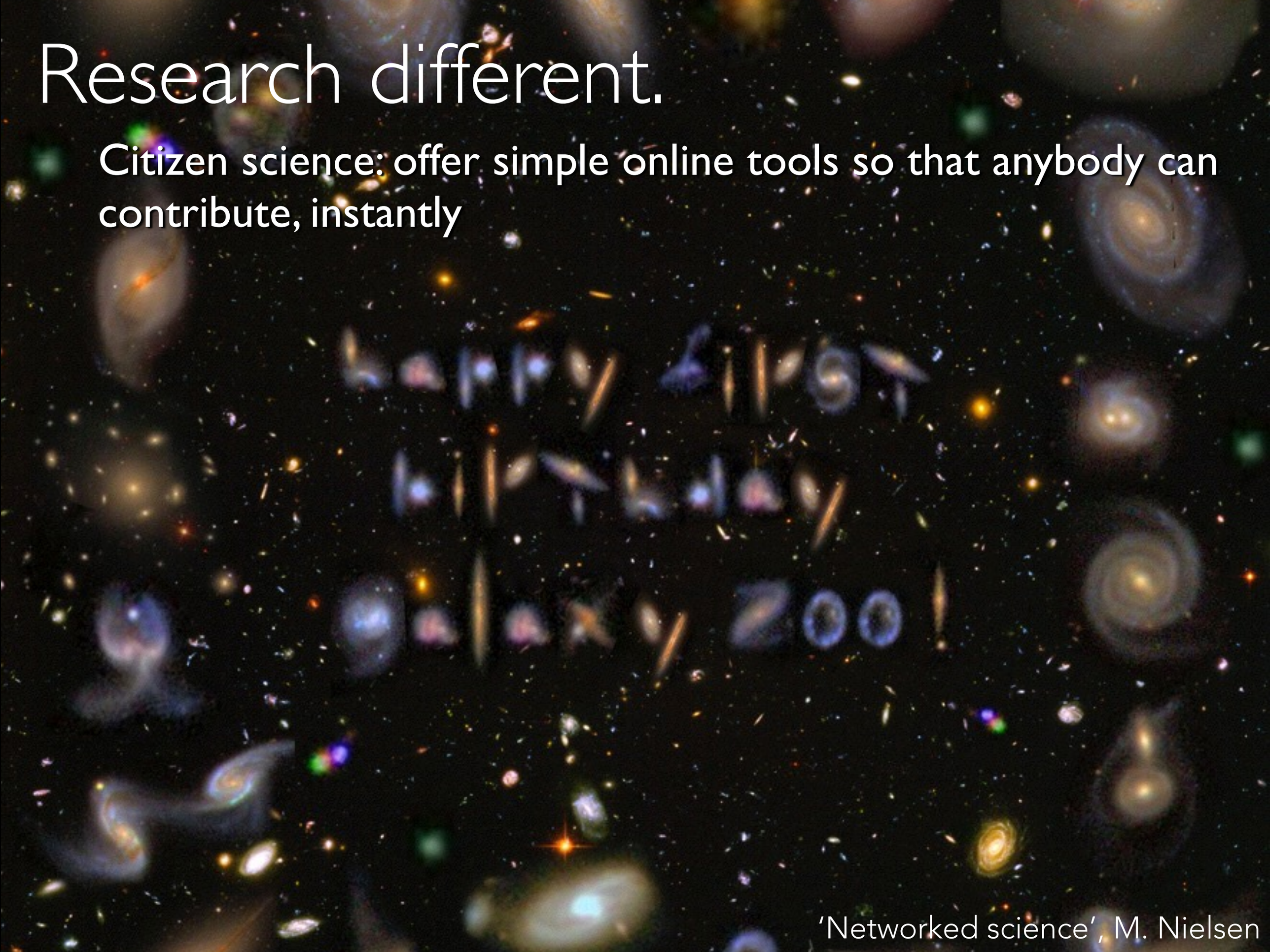
Broadcast data, let many minds
ask the right questions



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

Research different.

Citizen science: offer simple online tools so that anybody can contribute, instantly



‘Networked science’, M. Nielsen

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

Remove friction

Contribute easily

Easy access to organized data and tools

Give credit

Data is the new soil

Algorithms are like seeds that we sow on them

Let many people analyse data in many different ways, make new discoveries, share them easily

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

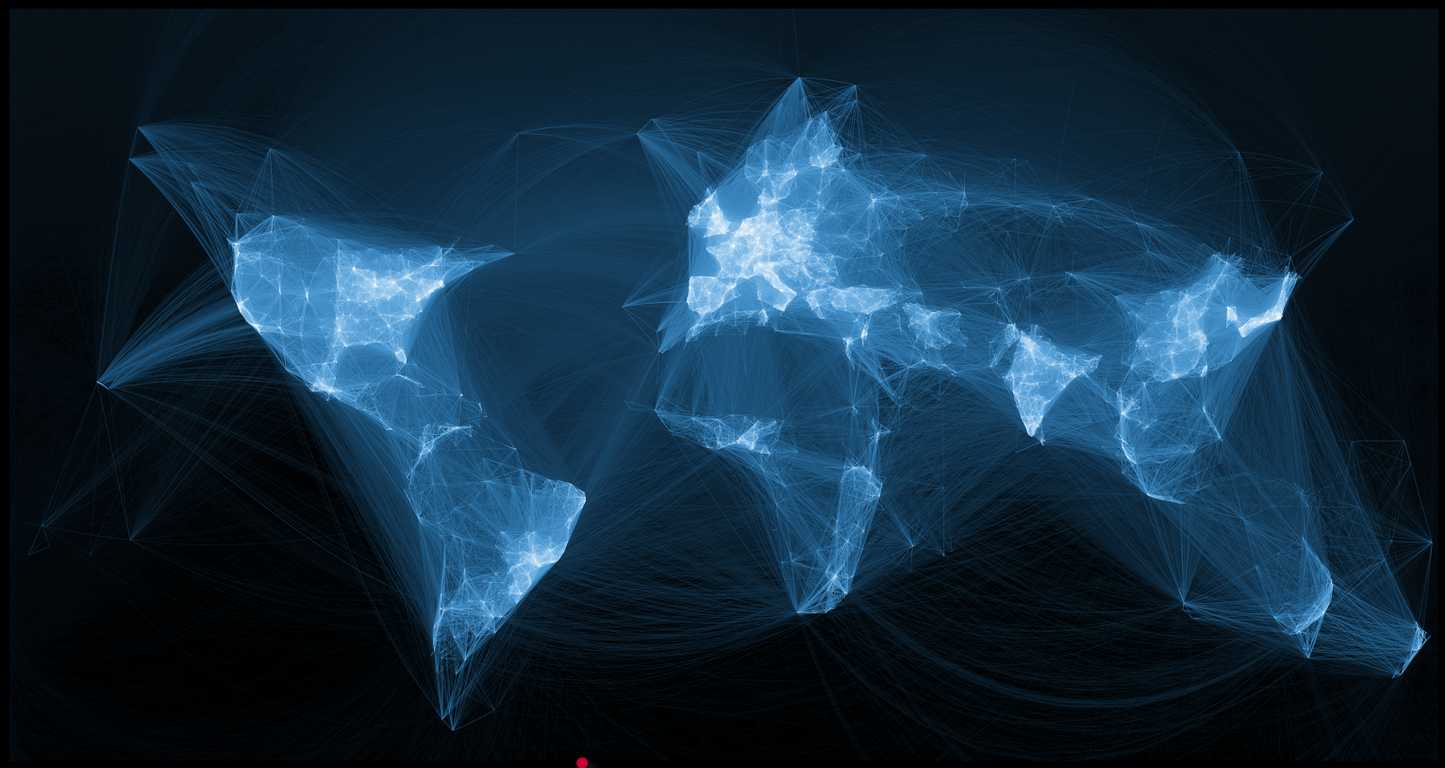
Millions of real, open datasets are generated

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

Extensive toolboxes exist to analyse data

- *SKLearn, MLR, RapidMiner, KNIME, WEKA, Watson, TensorFlow,…*

Missing link: extend toolboxes to share data and experiments on common collaboration platform



OpenML

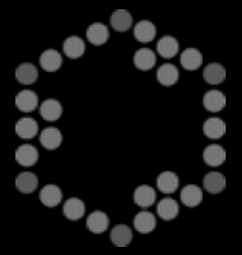
COLLABORATIVE MACHINE LEARNING

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)

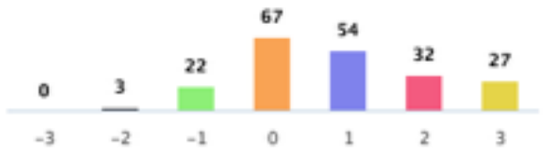
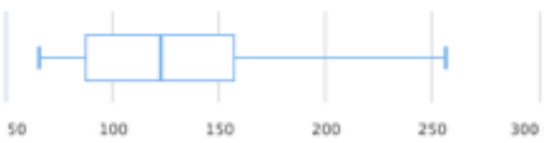
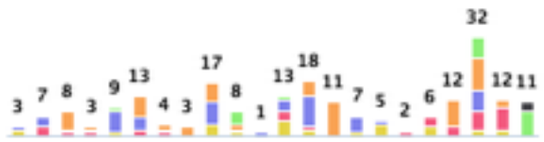


Data (ARFF) uploaded or referenced, versioned

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

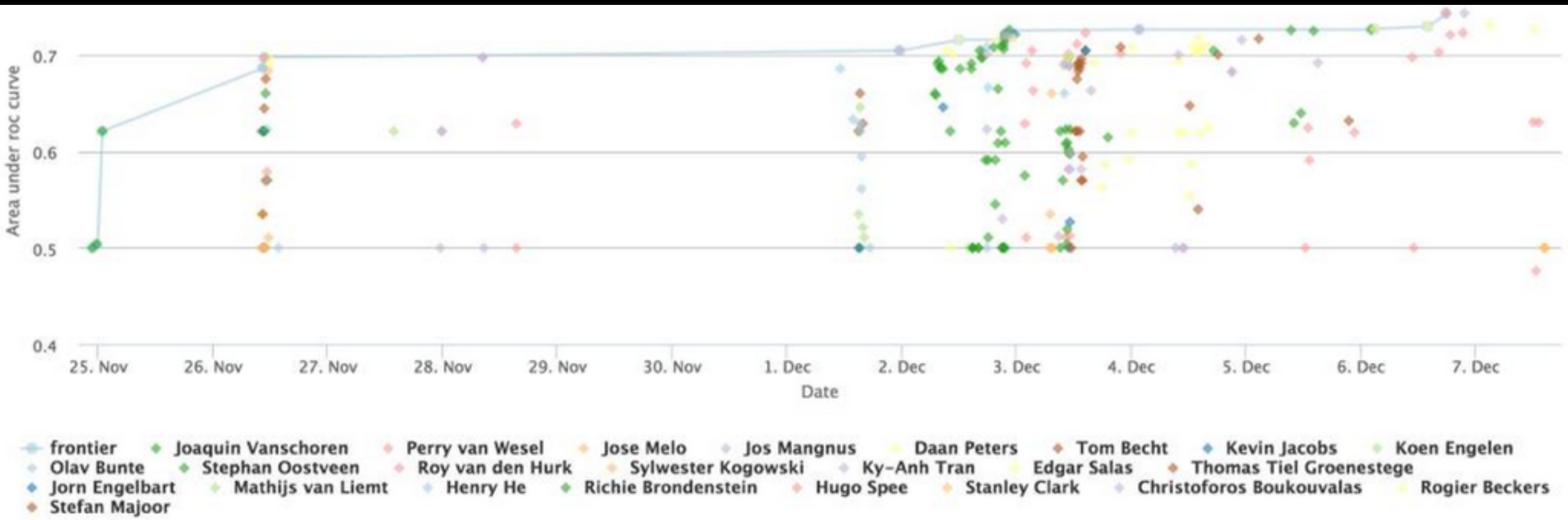
DefaultAccuracy	**0.33**	The predictive a
NumberOfClasses	**7**	The number of c
NumberOfFeatures	**26**	The number of f
NumberOfInstances	**205**	The number of i
NumberOfMissingVal...	**59**	Counts the total

Tasks contain data, goals, procedures.

Readable by tools, automates experimentation

Results organized online: **realtime overview**

Results organized online: **realtime overview**



Flows (code) run locally, auto-registered by tools

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

Integrations + APIs (REST, Java, R, 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

```
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.

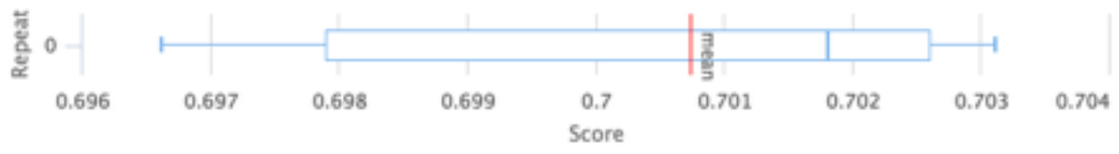
Area under ROC curve

0.7007 ± 0.0023

Per class

0	1
0.7007	0.7007

Cross-validation details (10-fold Crossvalidation)



OpenML beta

Exploring machine learning better, together

19590 data sets

Find or add **data** to analyse

46701 tasks

Download or create scientific **tasks**

2691 flows

Find or add data analysis **flows**

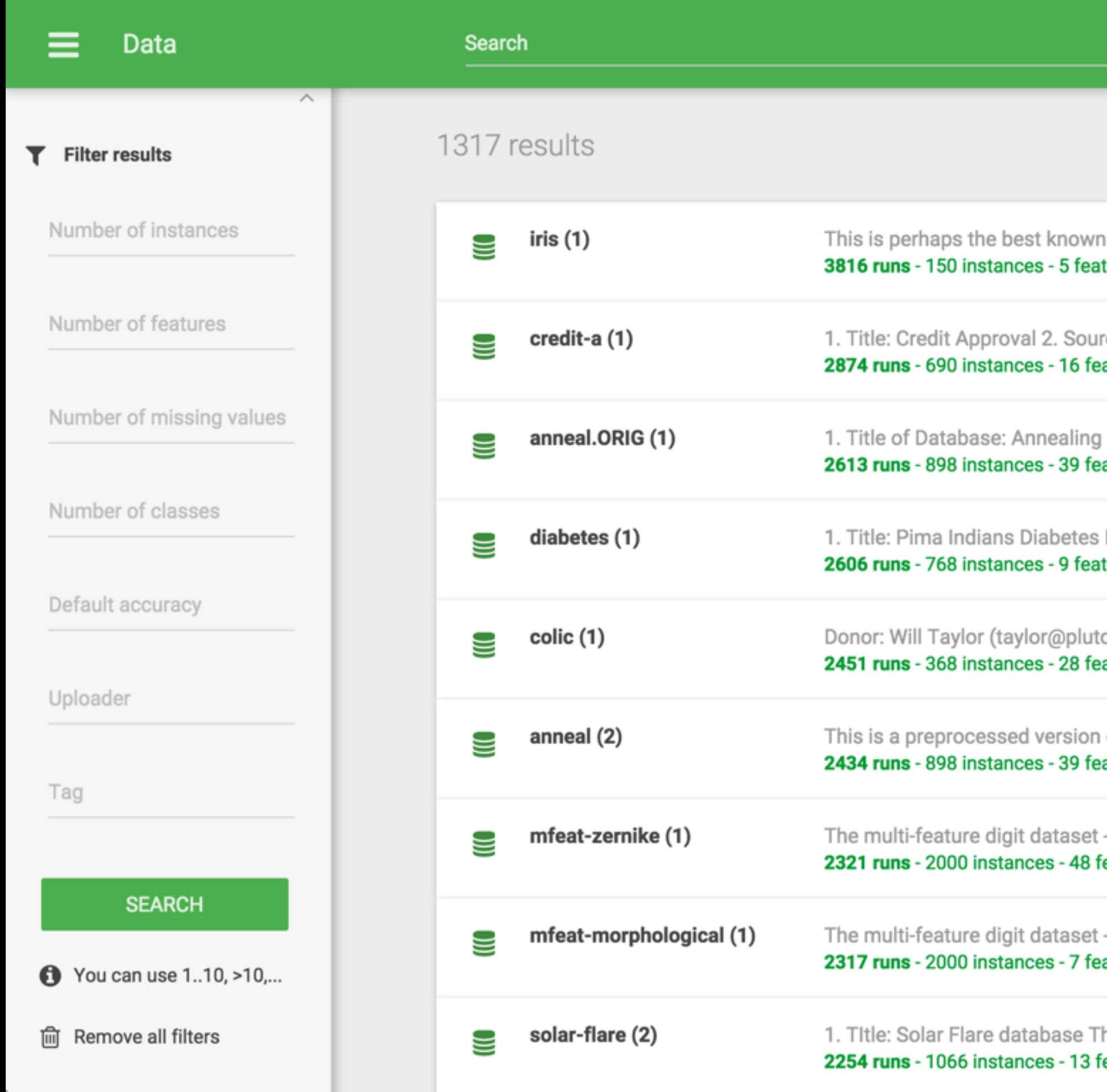
454945 runs

Upload and explore all **results** online.

Download and share data, flows and runs through:

Data

- Search by keywords or properties
- Filters
- Tagging



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

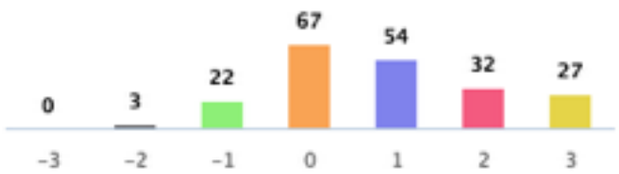
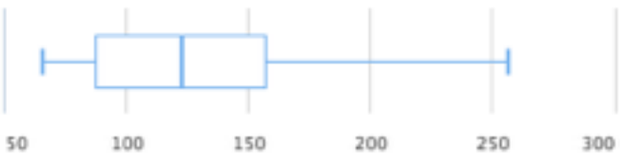
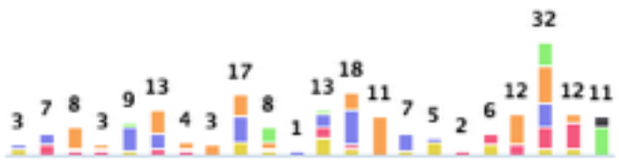
Help us complete this description → Edit

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

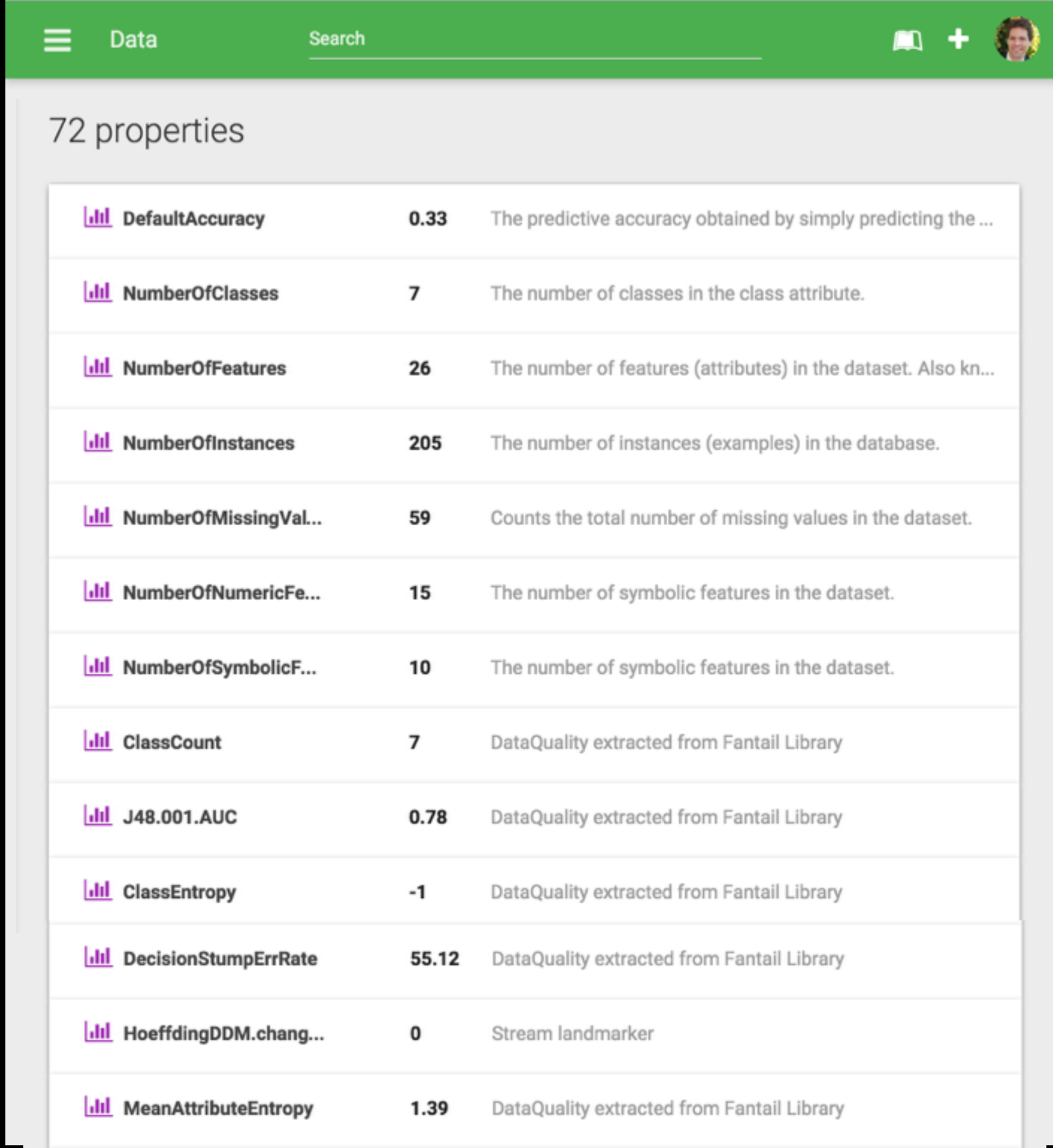
26 features

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

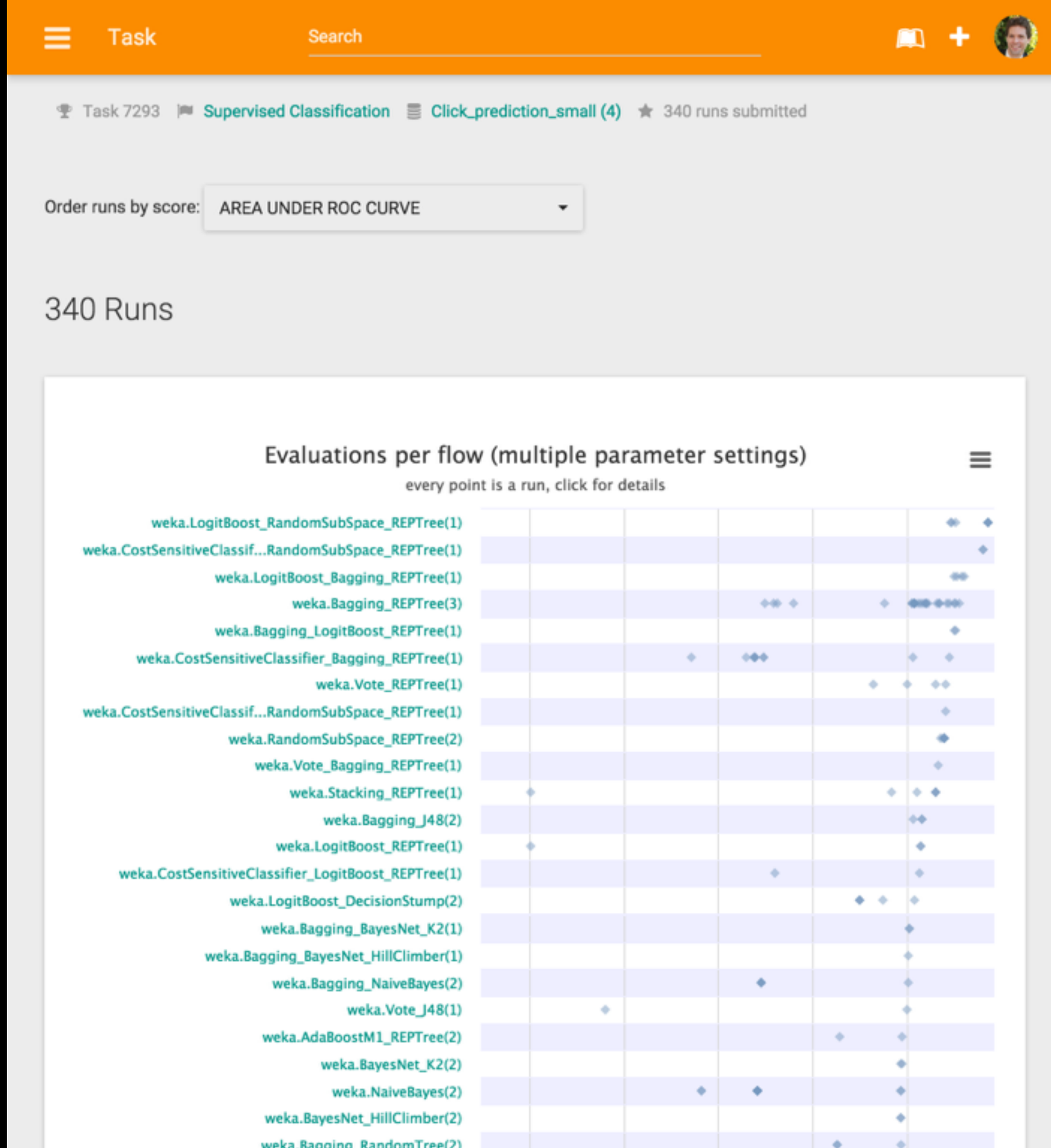
Data

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



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



Task

Search

Details

Overview

All runs

Results

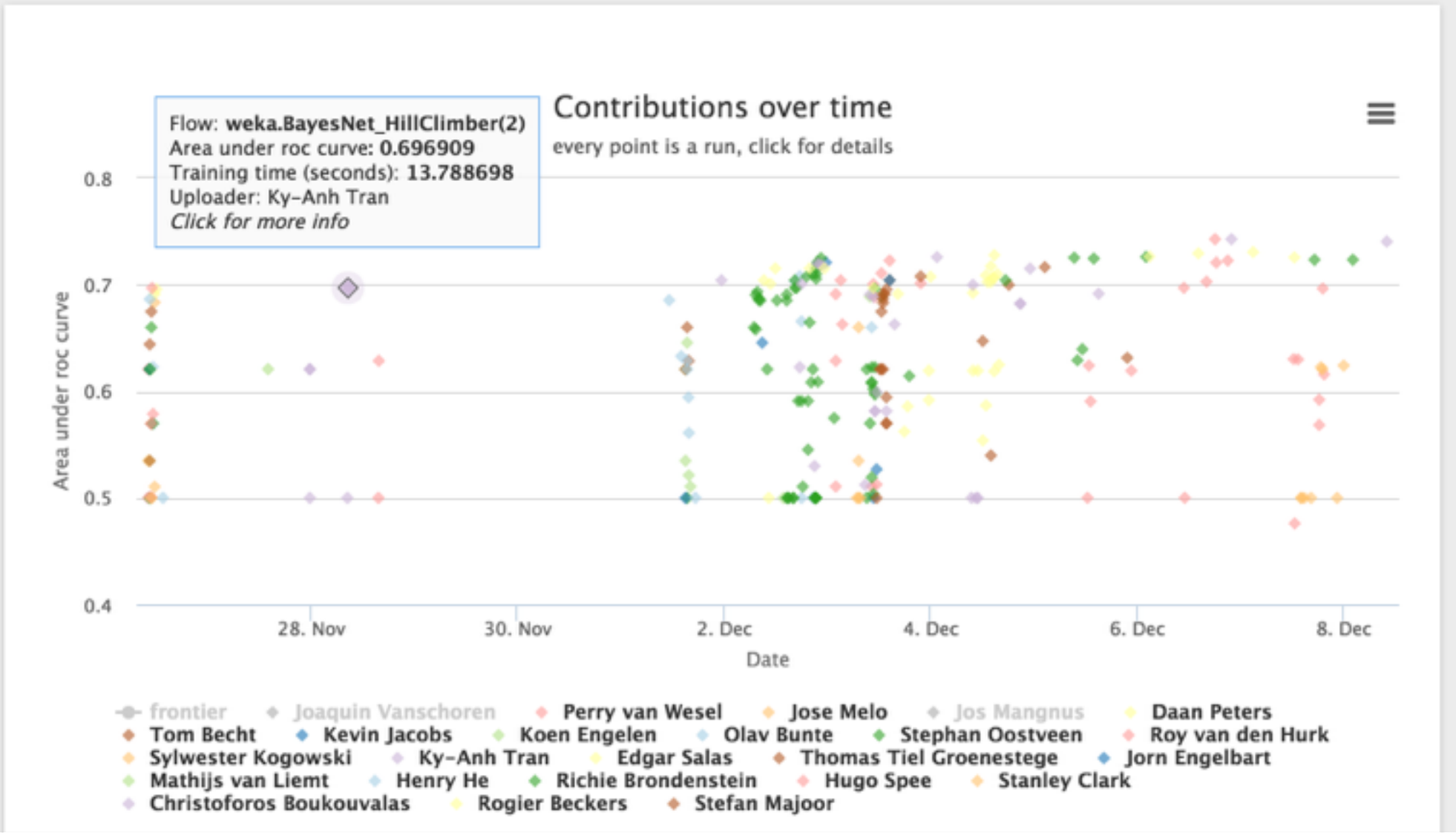
Leaderboard

Discuss

Tags

Add tag

Timeline



- 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

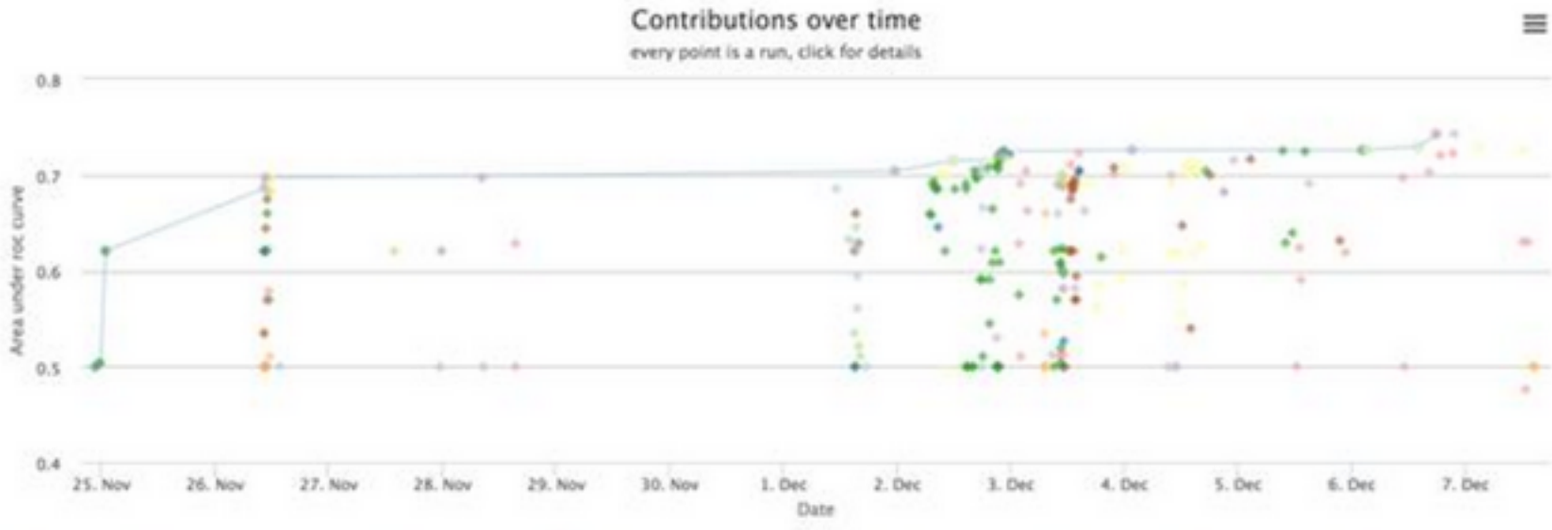
Rogier Beckers
@RogierBeckers

Follow

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

View translation

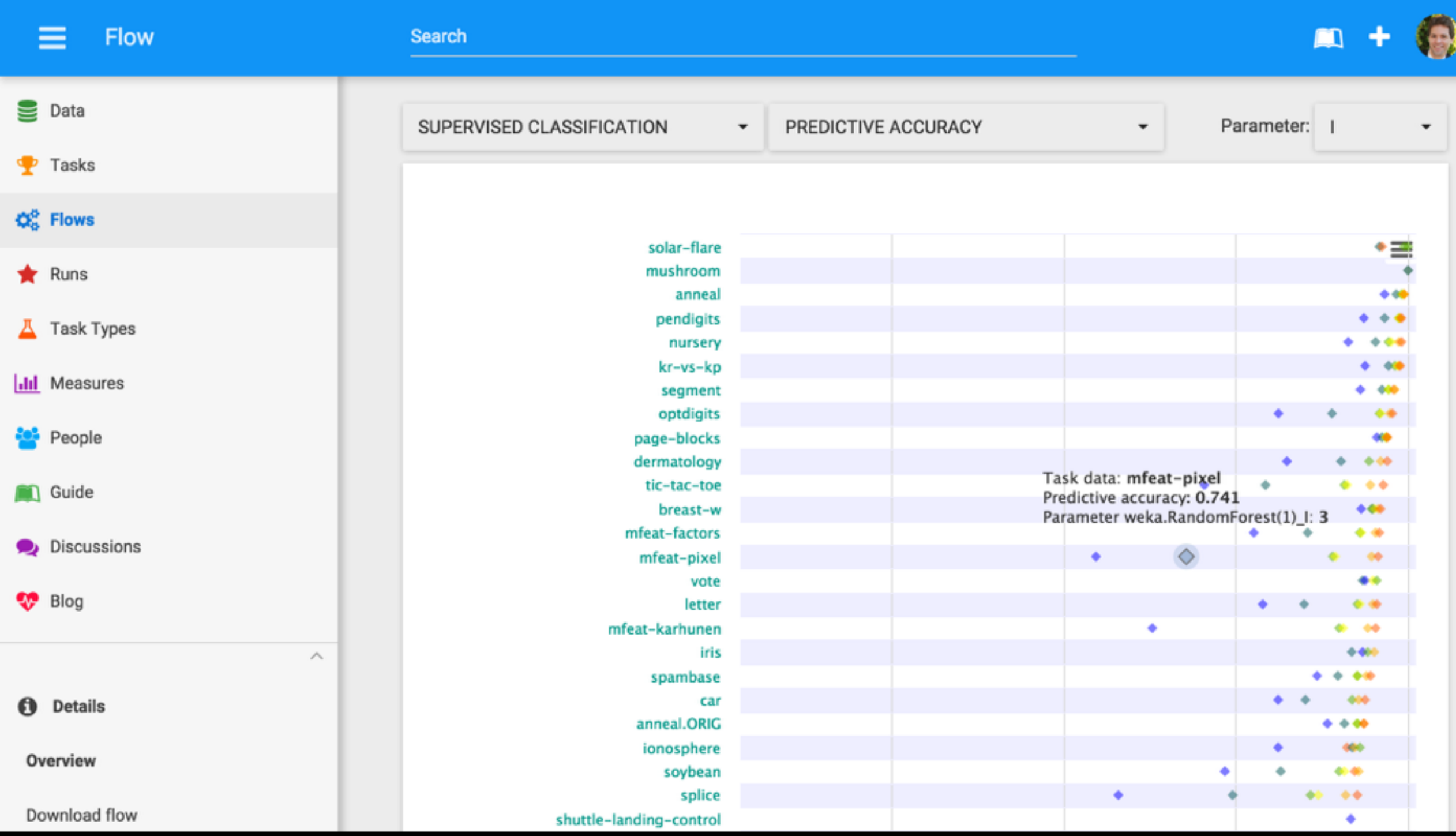
Lauradorp, Landgraaf



RETWEETS
2

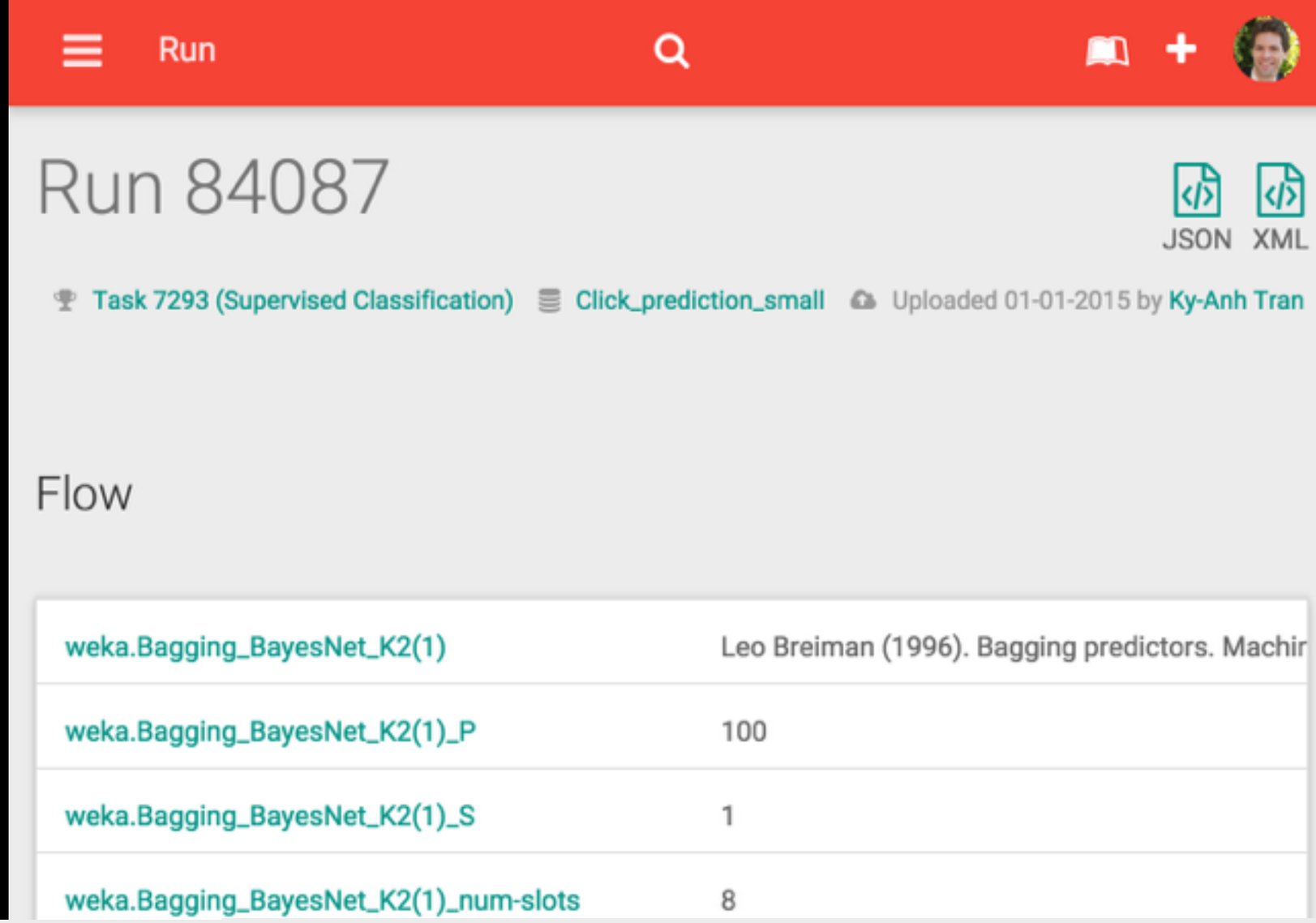
FAVORITES
2

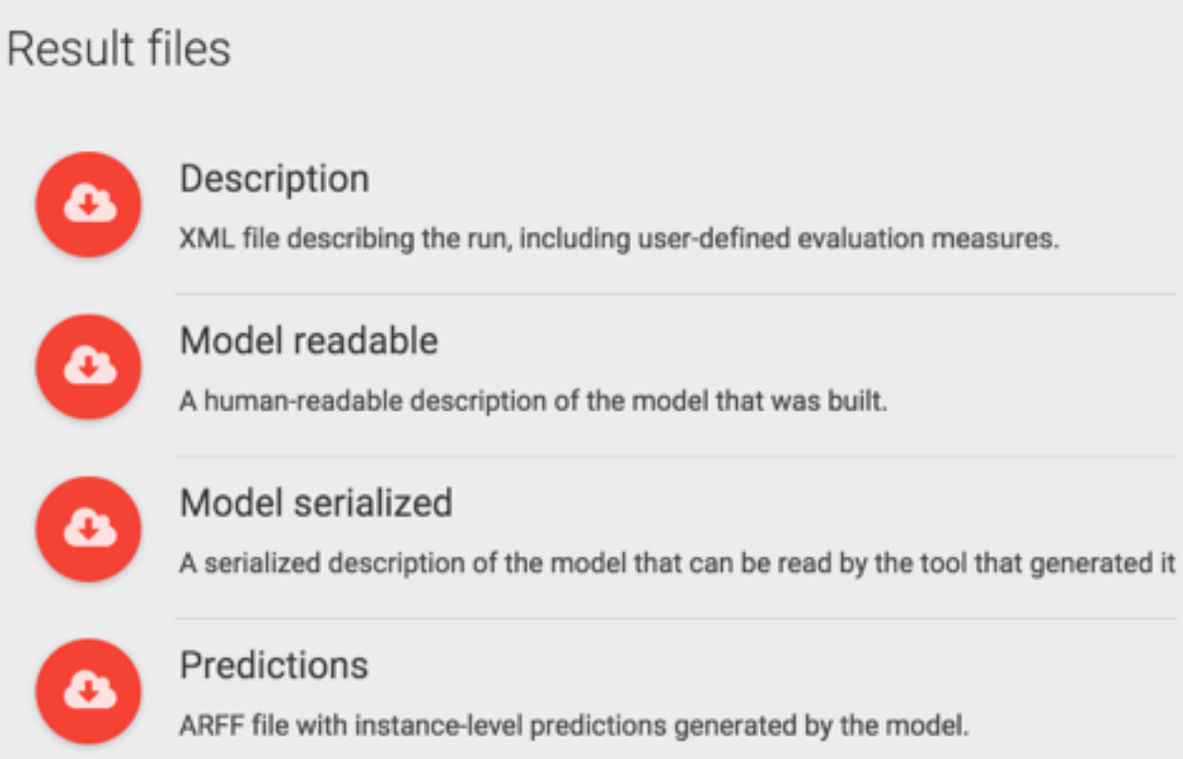
9:48 PM - 7 Dec 2014

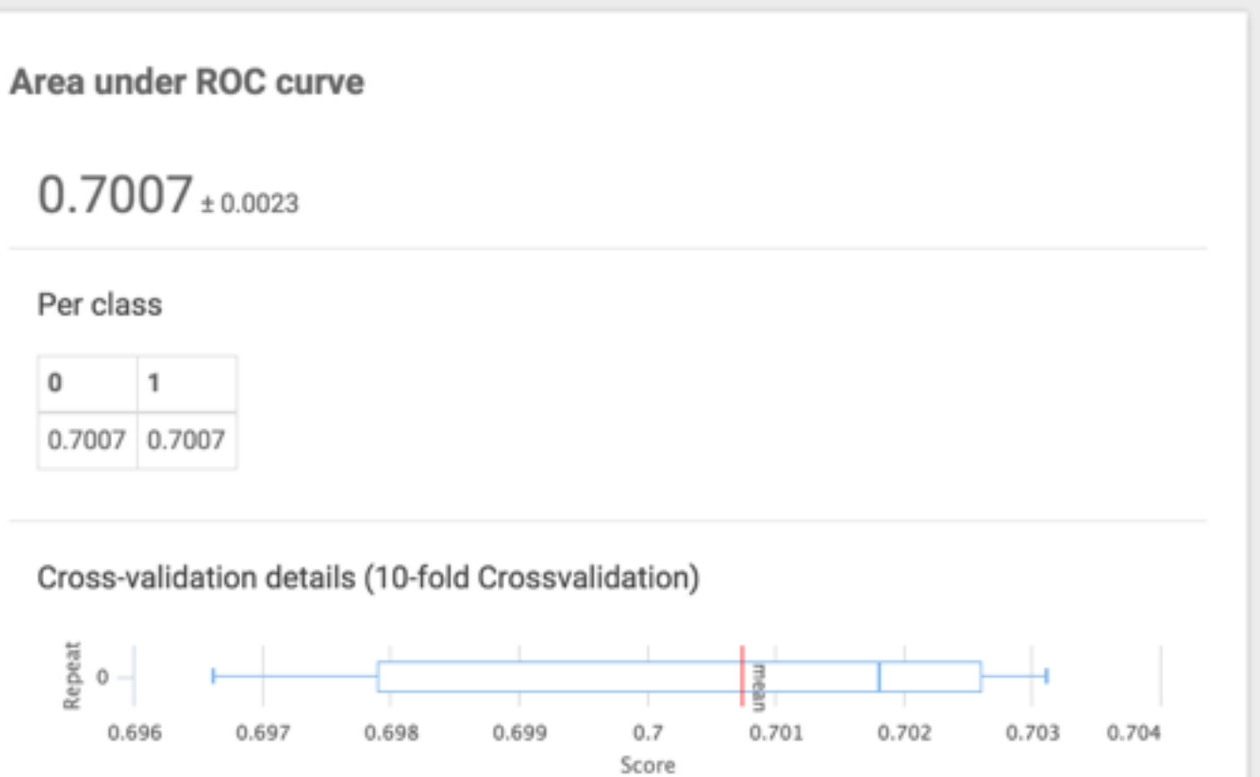


- All results obtained with same flow organised online
- Results linked to data sets, parameter settings -> trends/comparisons
- Visualisations (dots are models, ranked by score, colored by parameters)

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







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 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 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 runs and results

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

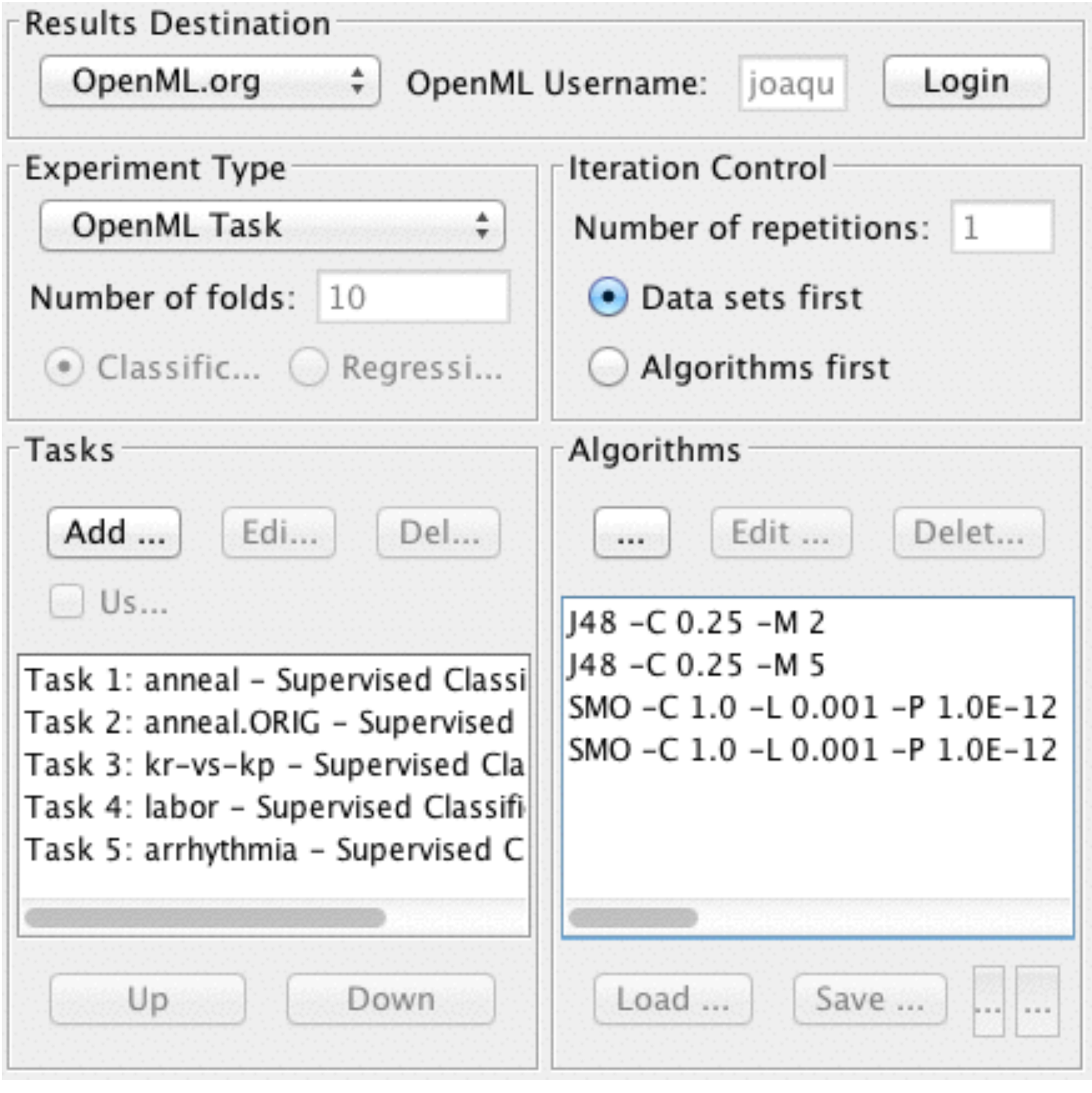
Explore

Reuse

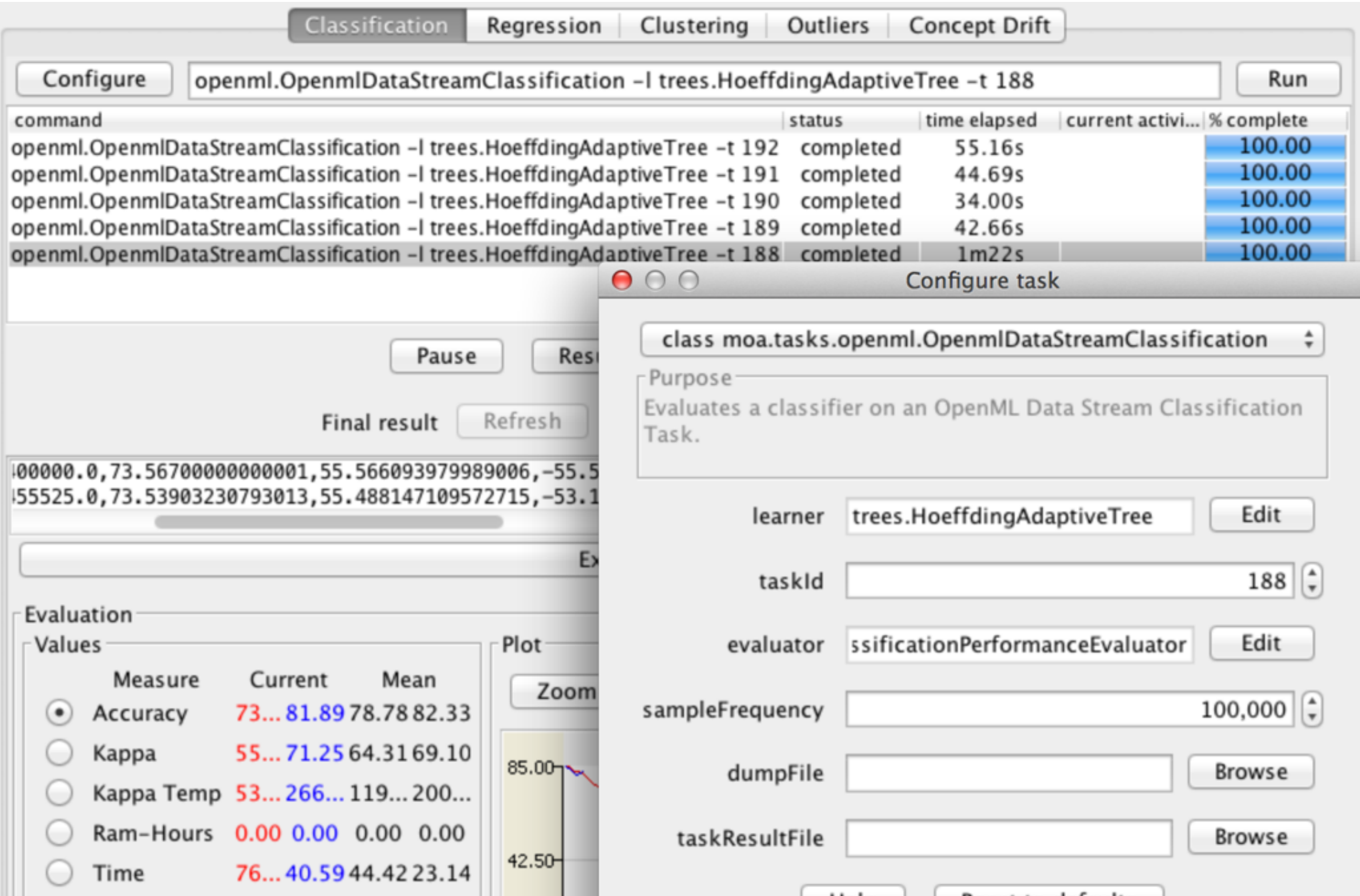
Share

WEKA

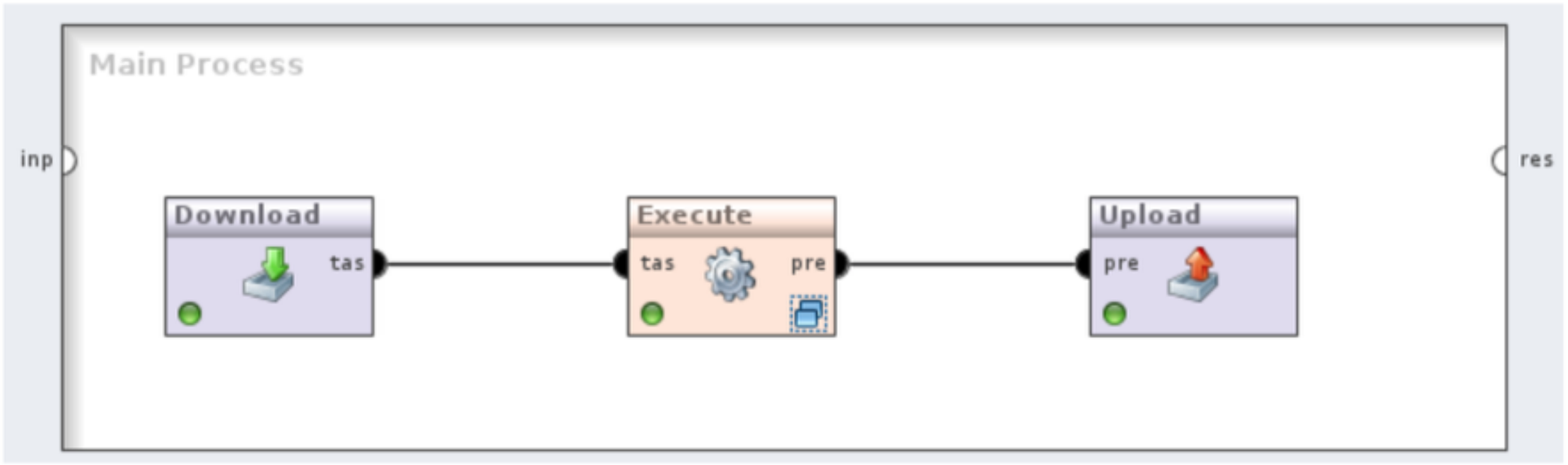
OpenML extension in plugin manager



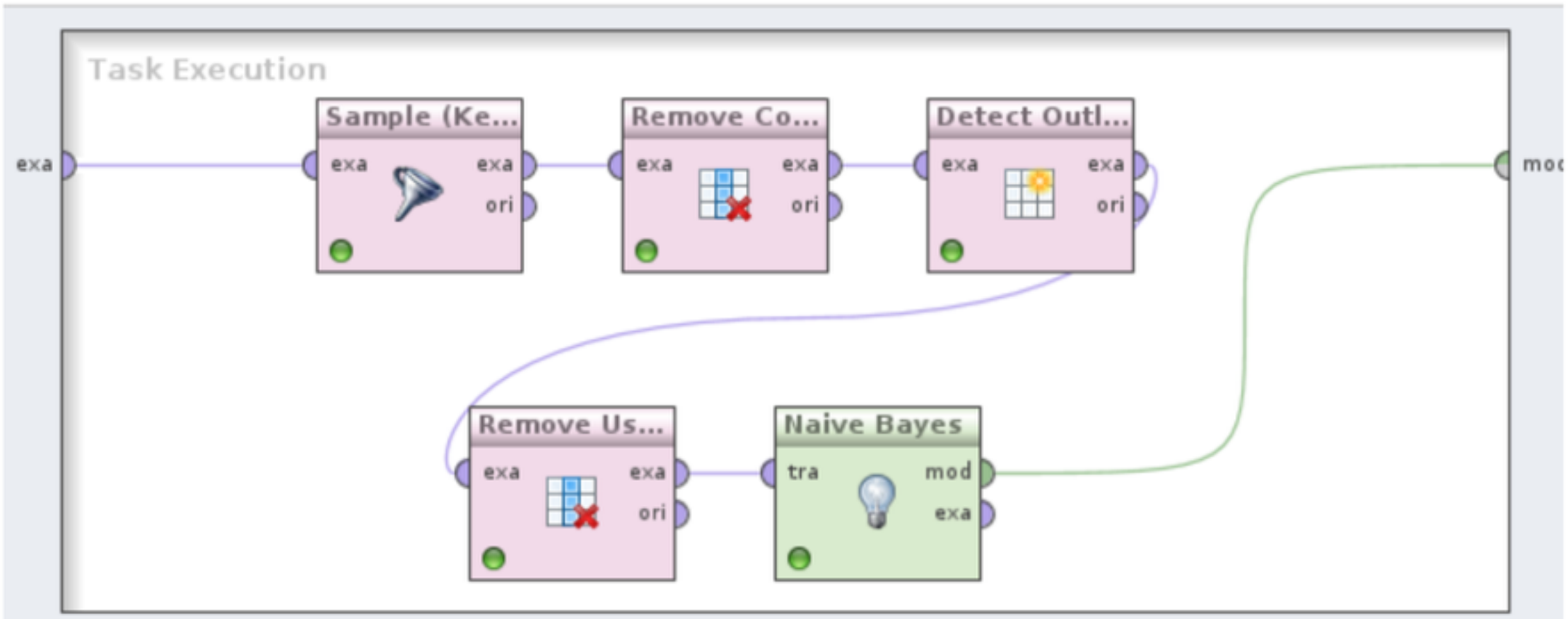
MOA


Classification
Regression
Clustering
Outliers
Concept Drift
Configure
openml.OpenmlDataStreamClassification -l trees.HoeffdingAdaptiveTree -t 188
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
Final result
Refresh
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

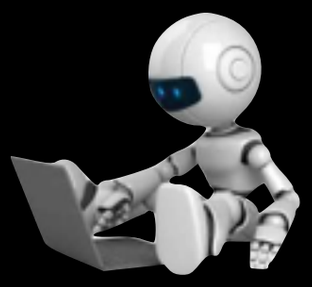
Data science, differently

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)



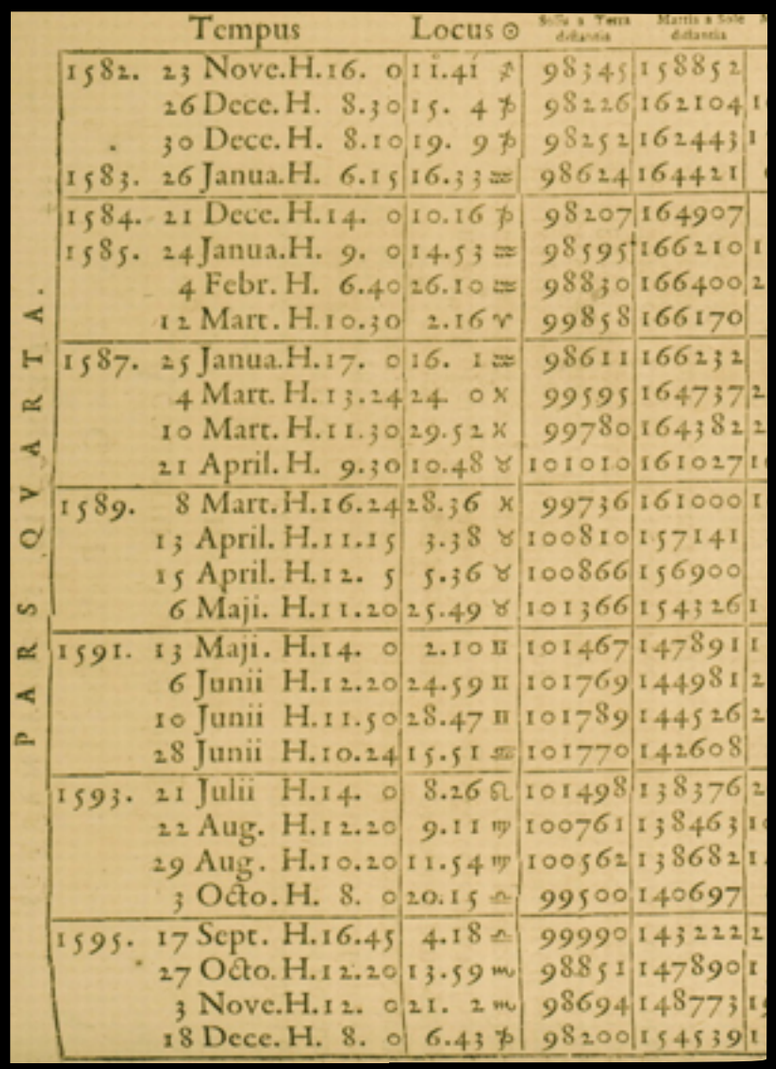
Automation

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

A NEW TALE OF MANY MINDS

‘BIG’ DATA

A



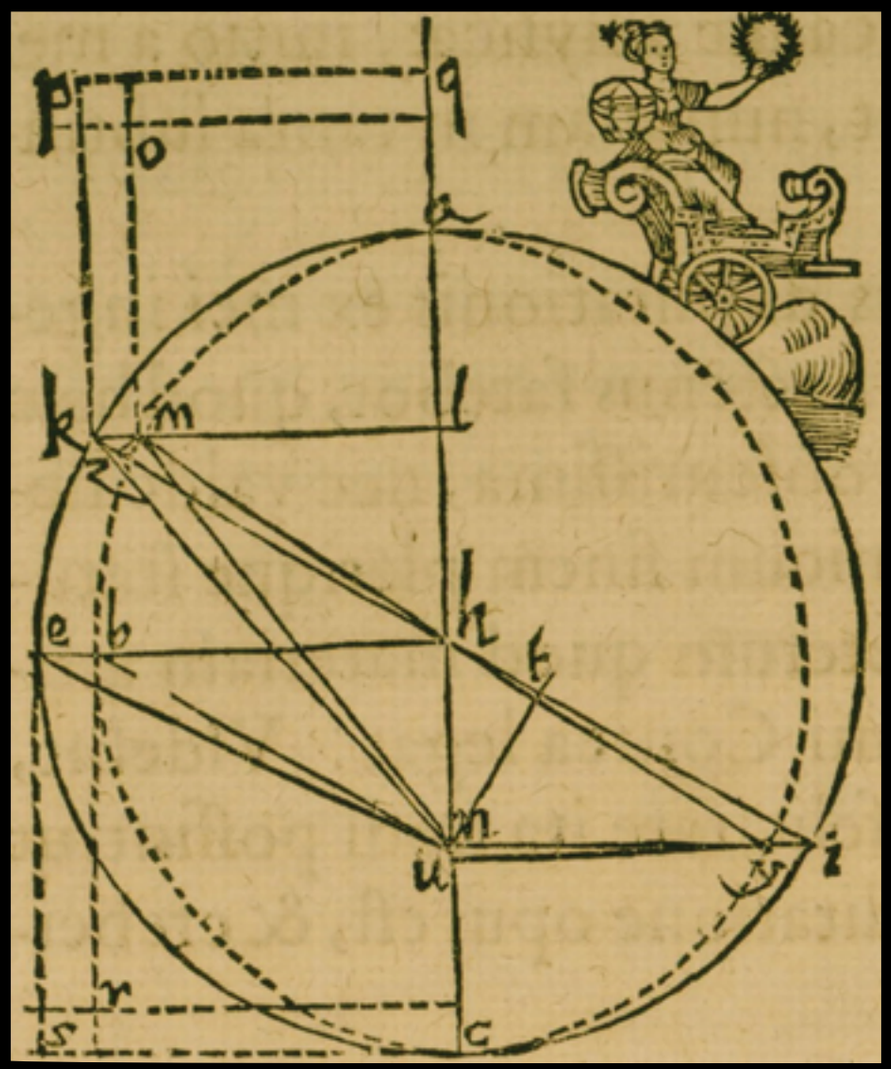
	Tempus	Locus ☉	Solis a Venere distantia	Martis a Sole distantia	
1582.	23 Nove. H. 16. 0	11.41 ♐	98345	158852	
	26 Dece. H. 8.30	15. 4 ♑	98226	162104	1
	30 Dece. H. 8.10	19. 9 ♑	98252	162443	1
1583.	26 Janua. H. 6.15	16.33 ♒	98624	164421	
1584.	21 Dece. H. 14. 0	10.16 ♑	98207	164907	
1585.	24 Janua. H. 9. 0	14.53 ♒	98595	166210	1
	4 Febr. H. 6.40	26.10 ♒	98830	166400	2
	12 Mart. H. 10.30	2.16 ♈	99858	166170	
1587.	25 Janua. H. 17. 0	16. 1 ♒	98611	166232	
	4 Mart. H. 13.24	24. 0 ♓	99595	164737	2
	10 Mart. H. 11.30	29.52 ♓	99780	164382	2
	21 April. H. 9.30	10.48 ♉	101010	161027	1
1589.	8 Mart. H. 16.24	28.36 ♓	99736	161000	1
	13 April. H. 11.15	3.38 ♉	100810	157141	
	15 April. H. 12. 5	5.36 ♉	100866	156900	
	6 Maji. H. 11.20	25.49 ♉	101366	154326	1
1591.	13 Maji. H. 14. 0	2.10 ♊	101467	147891	1
	6 Junii H. 12.20	24.59 ♊	101769	144981	2
	10 Junii H. 11.50	28.47 ♊	101789	144526	2
	28 Junii H. 10.24	15.51 ♋	101770	142608	
1593.	21 Julii H. 14. 0	8.26 ♌	101498	138376	2
	22 Aug. H. 12.20	9.11 ♍	100761	138463	1
	29 Aug. H. 10.20	11.54 ♍	100562	138682	1
	3 Octo. H. 8. 0	20.15 ♎	99500	140697	
1595.	17 Sept. H. 16.45	4.18 ♎	99990	143222	2
	27 Octo. H. 12.20	13.59 ♏	98851	147890	1
	3 Nove. H. 12. 0	21. 2 ♏	98694	148773	1
	18 Dece. H. 8. 0	6.43 ♑	98200	154539	1

PARS QVARTA.

OPEN DATA

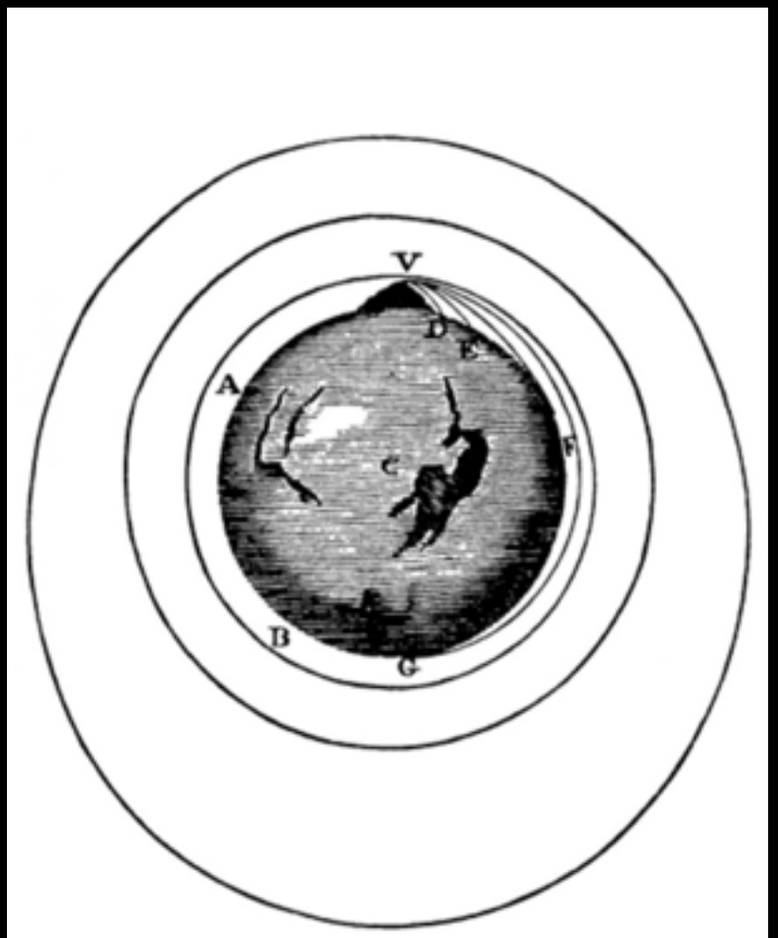
MACHINE LEARNING

B



OPEN SOURCE

NEW THEORY

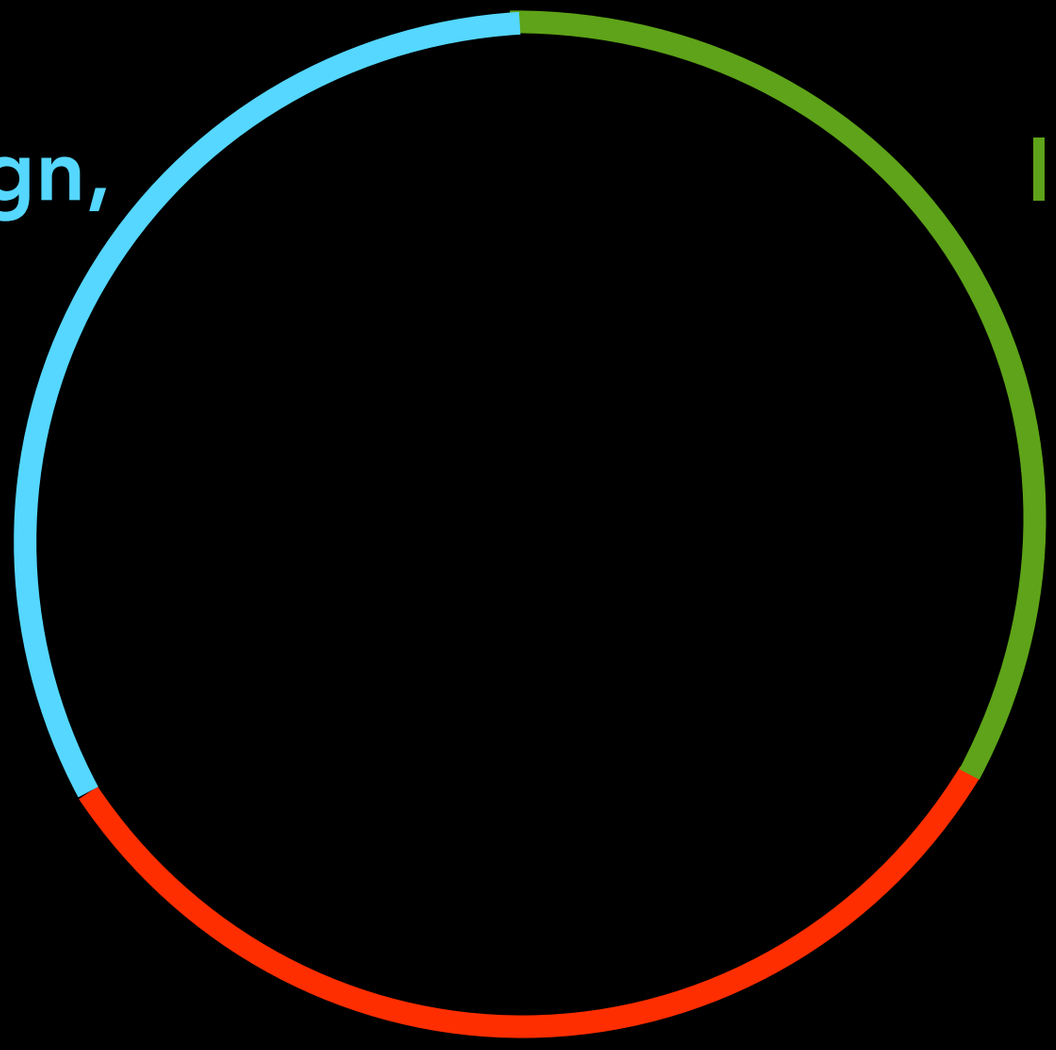


OPEN SCIENCE

Towards a data science collaboratory

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

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

Algorithm design, selection

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 Tatsioni 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 Tatsioni. "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 planked 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 Tatsioni and eight other youngish Greek researchers and physicians who, in contrast to the pasty 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

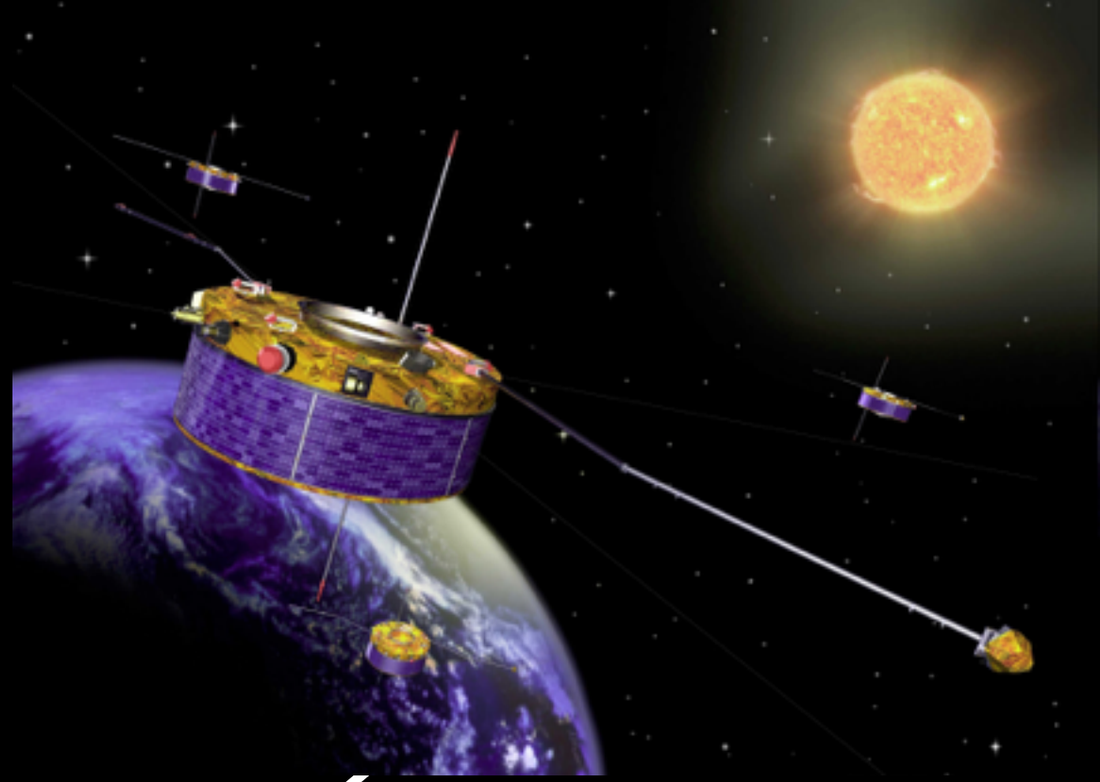
In subfields, up to 85% medical research resources are wasted

- 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



solar wind

bow shock

magnetosheath

magnetopause

solar wind

CLUSTER3

Earth

X

Y

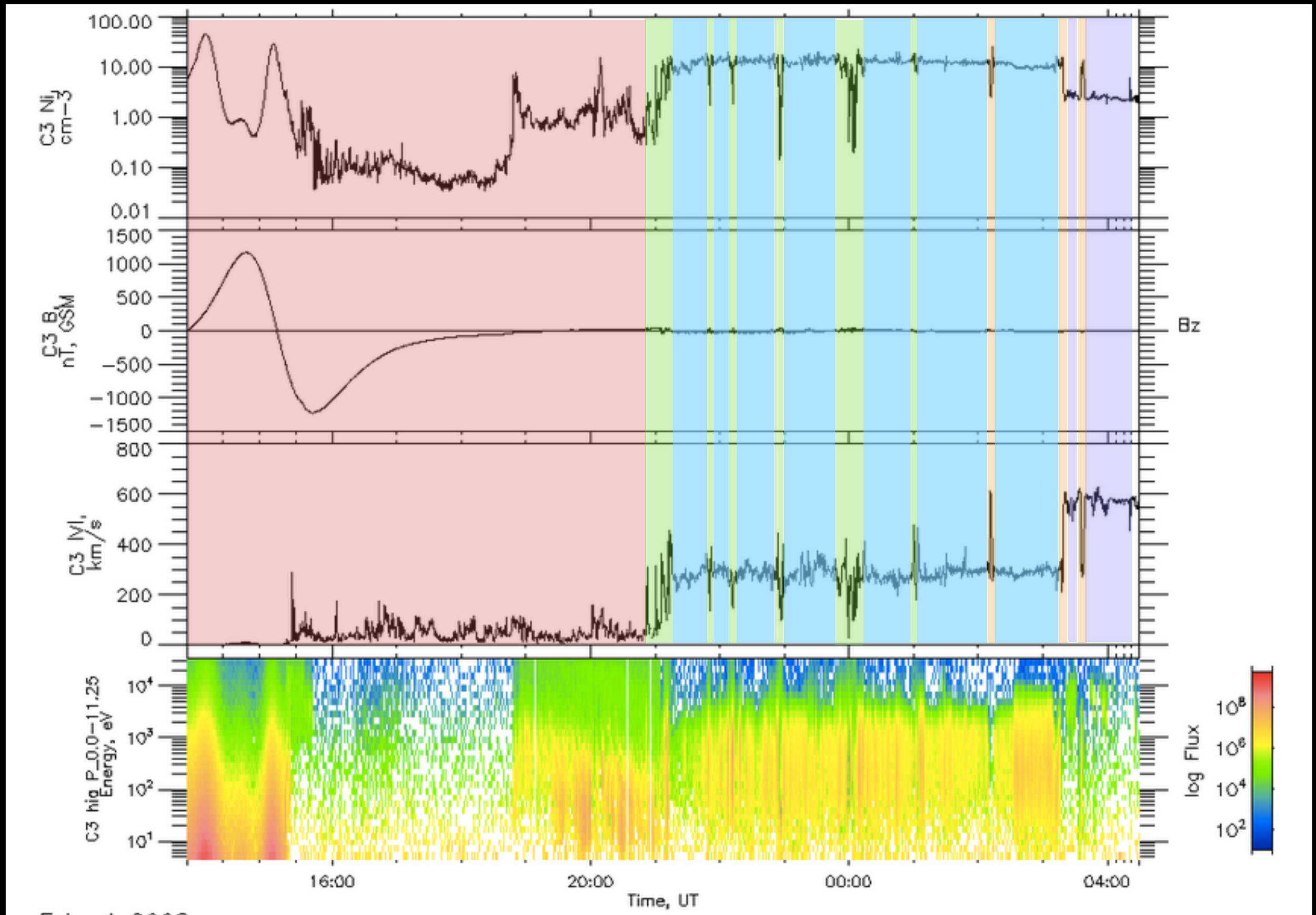
Z

Time: 2008-02-02T05:00:00 Distances (Re = Earth radius = 6371.010km)

SURFACE WAVES AND PROCESSES

Nicolas Aunai

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

Problem definition, data collection

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

Gaps in the ecosystem

Algorithm design, selection

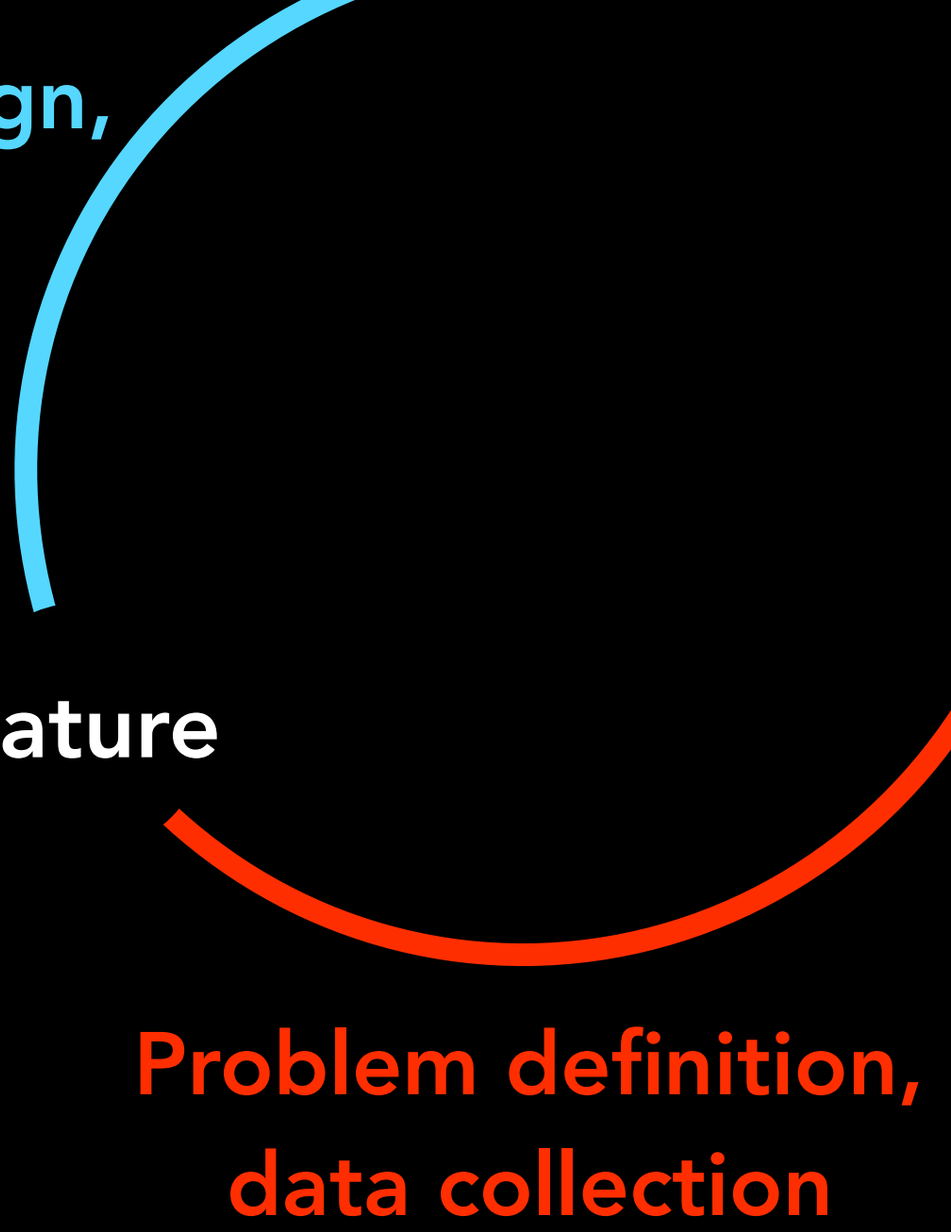
Small-scale collaboration

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



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

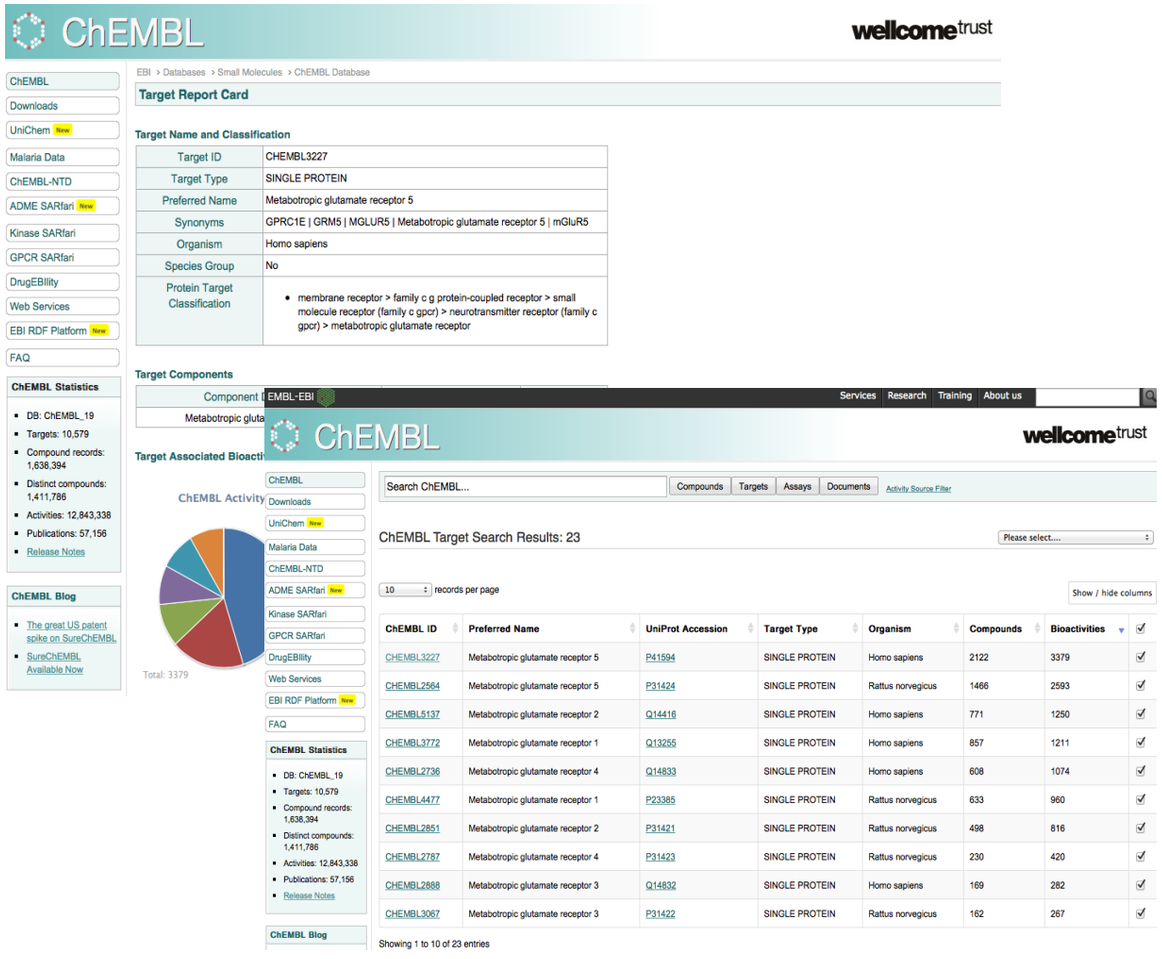
Broadcast data so that many minds analyse the data in different ways

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

Organize everything on a collaboration platform

OpenML in drug discovery

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



SMILES

Molecular properties (e.g. MW, LogP)

Fingerprints (e.g. FP2, FP3, FP4, MACSS)

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	
							⋮					

ChEMBL database

1.4M compounds,10k proteins,
12,8M activities

16.000+ regression datasets

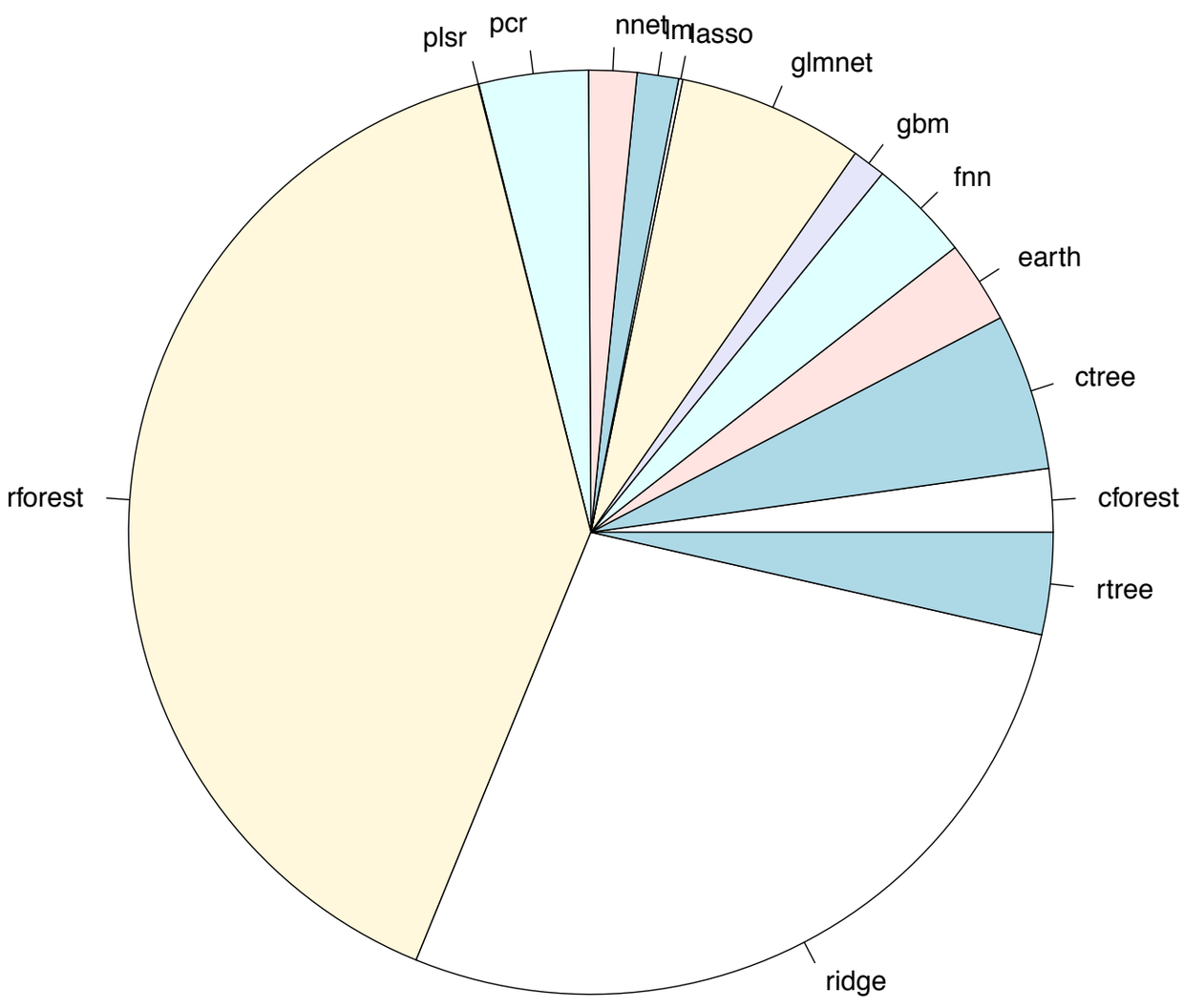
2750 targets have >10 compounds, x 4 fingerprints

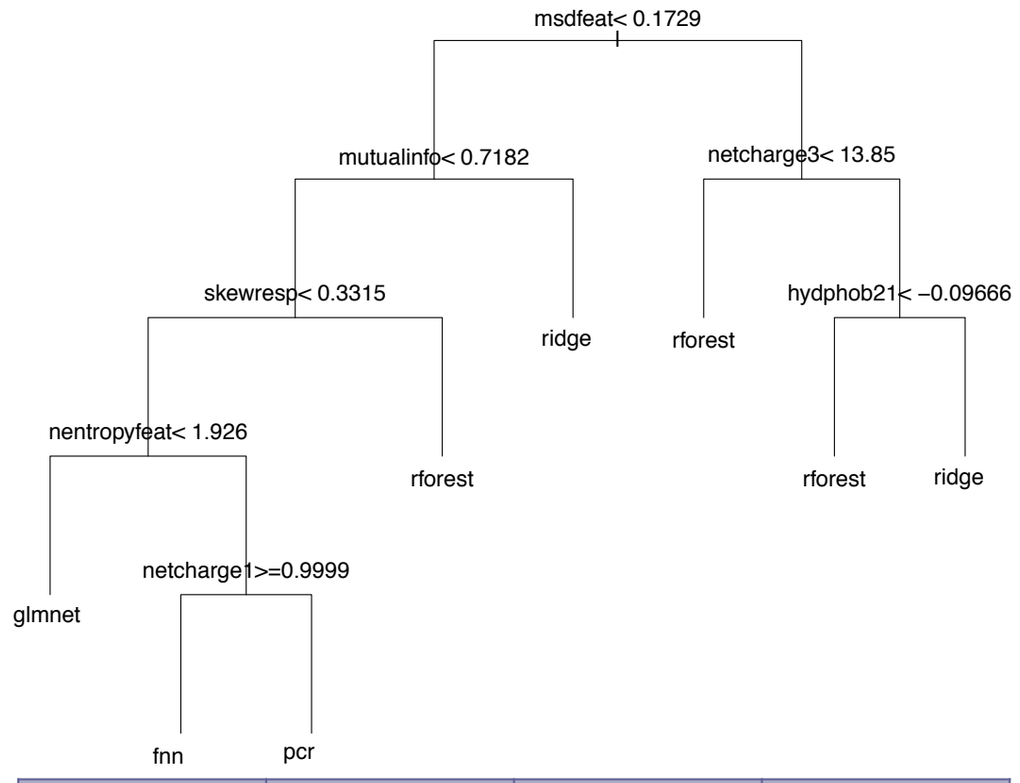
OpenML in drug discovery

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, landmarkers
- target: aliphatic index, hydrophobicity, net charge, mol. weight, sequence length, …





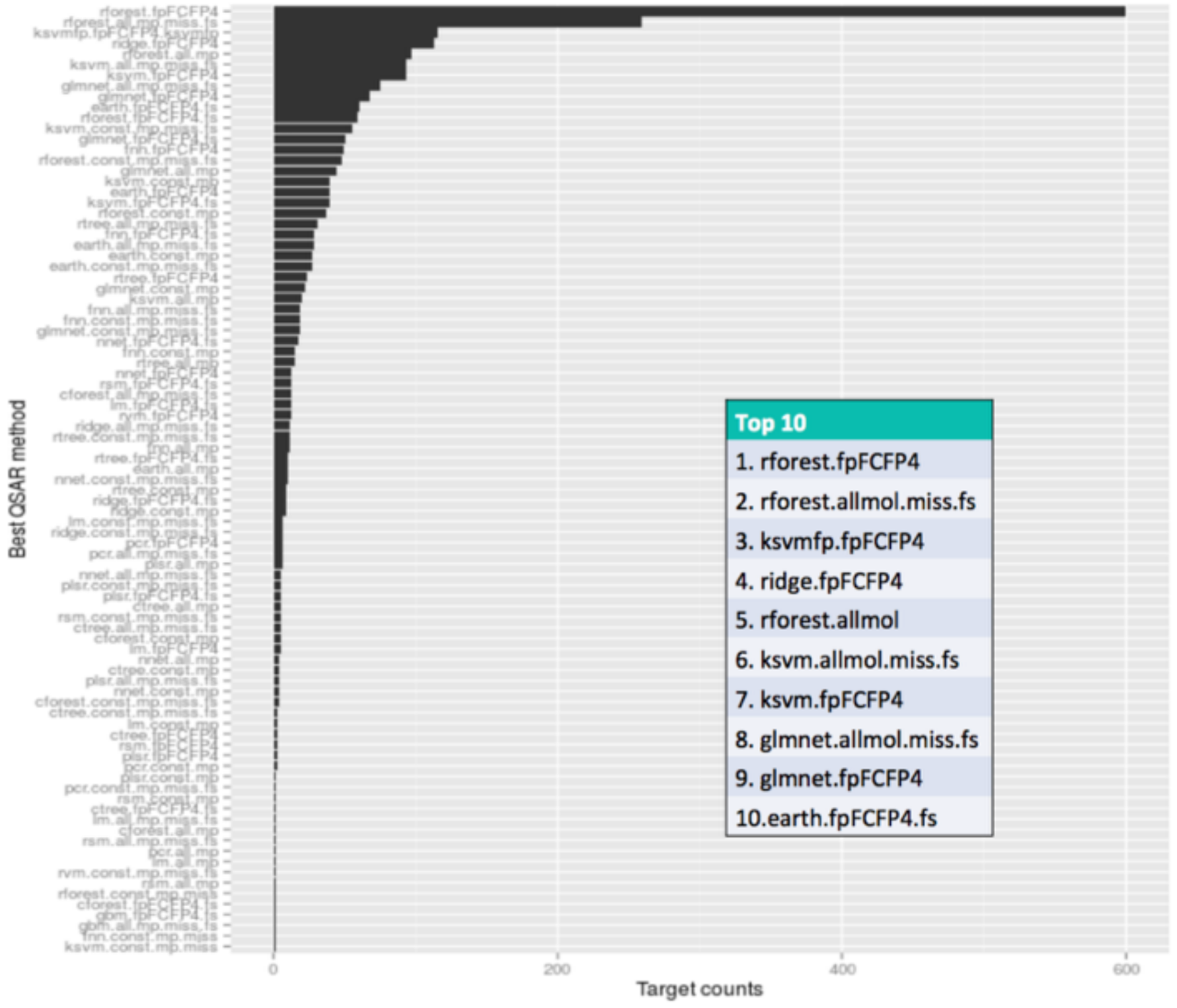
ECFP4_1024	FCFP4_1024	**ECFP6_1024**	FCFP6_1024
0.697	0.427	**0.725**	0.627

Predict best algorithm with meta-models

OpenML in drug discovery

Best algorithms?

Best features?

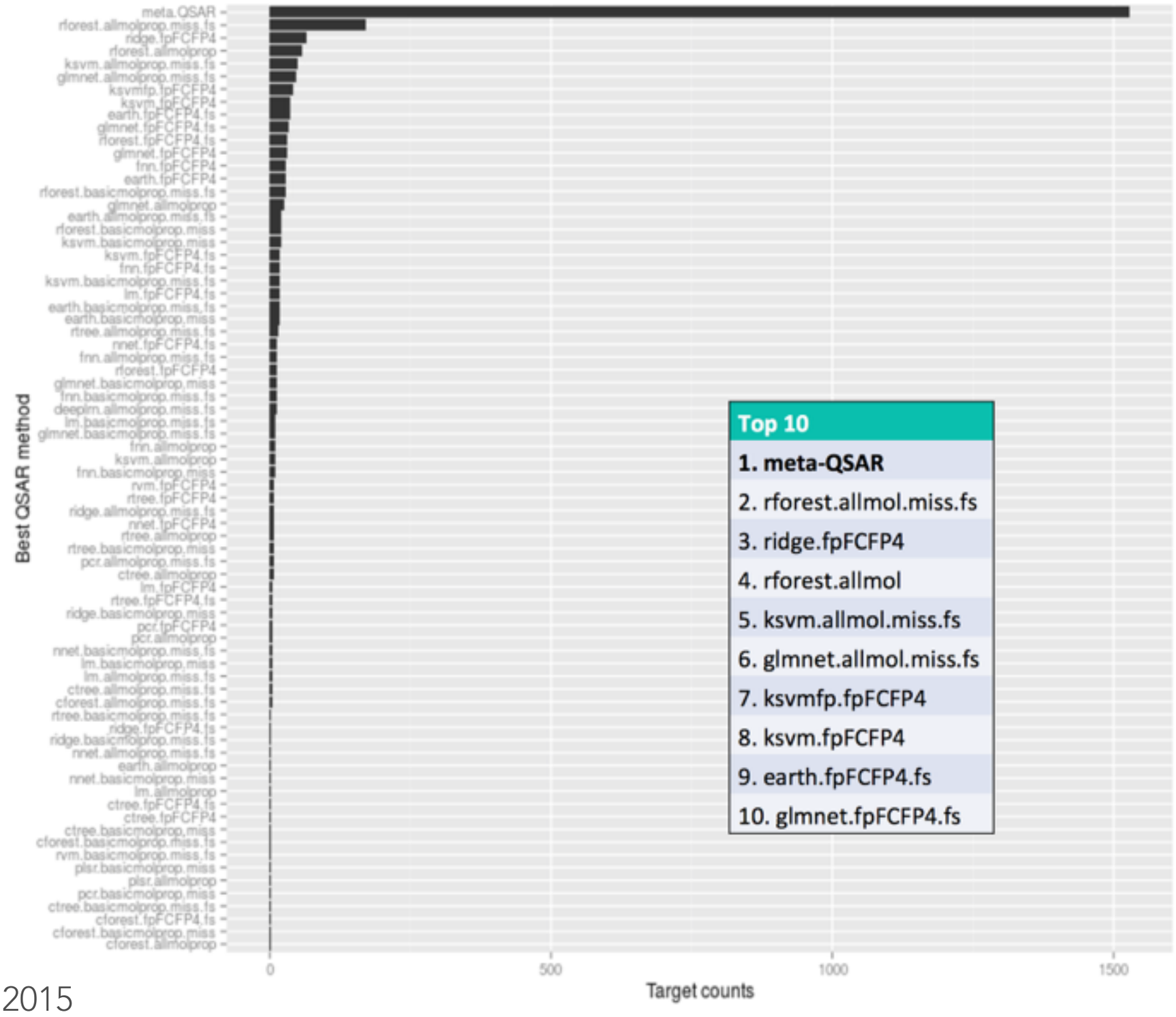


I. Olier et al. MetaSel@ECMLPKDD 2015

OpenML in drug discovery

New technique: stacked generalisation

Best model by far

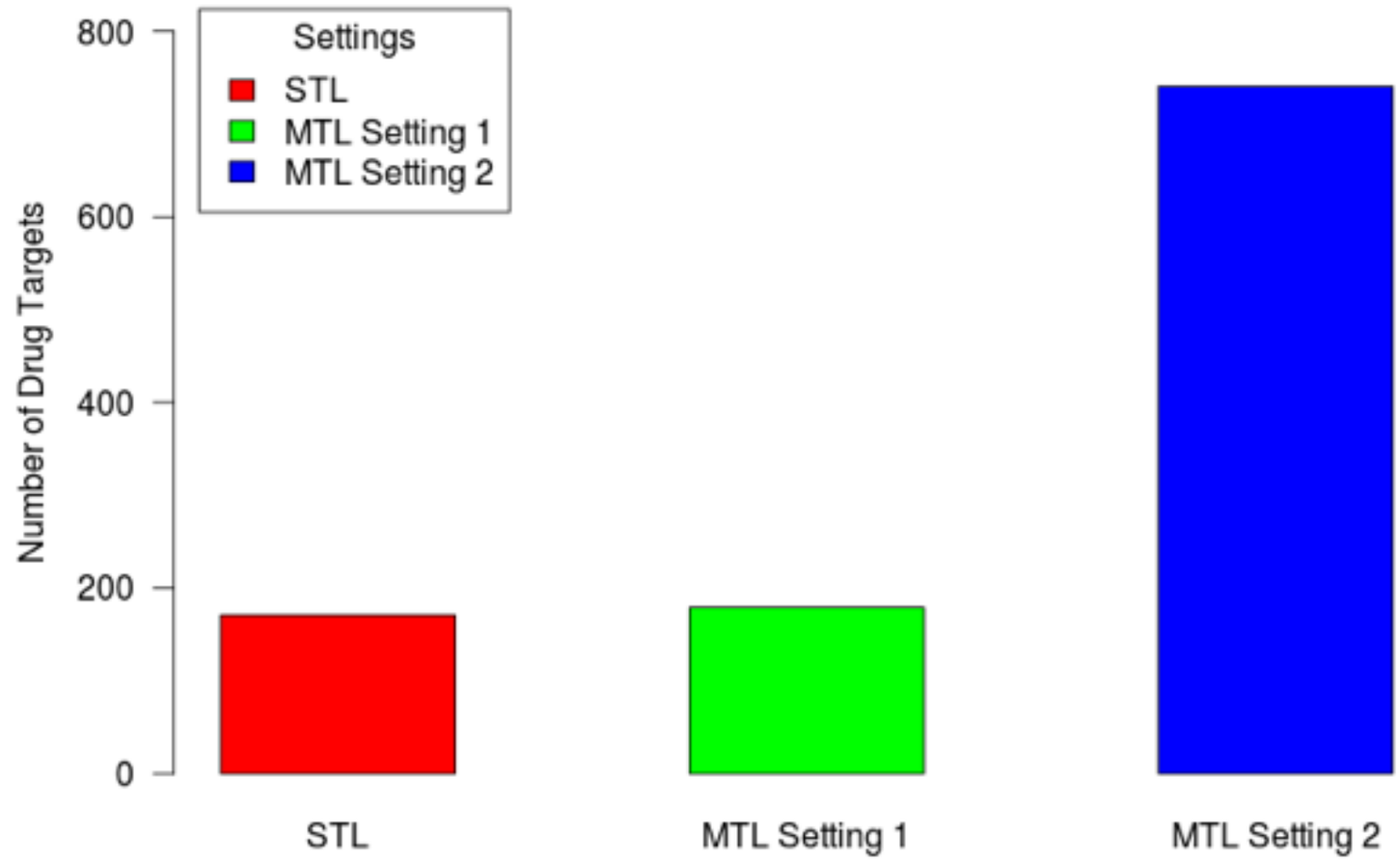


I. Olier et al. MetaSel@ECMLPKDD 2015

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)

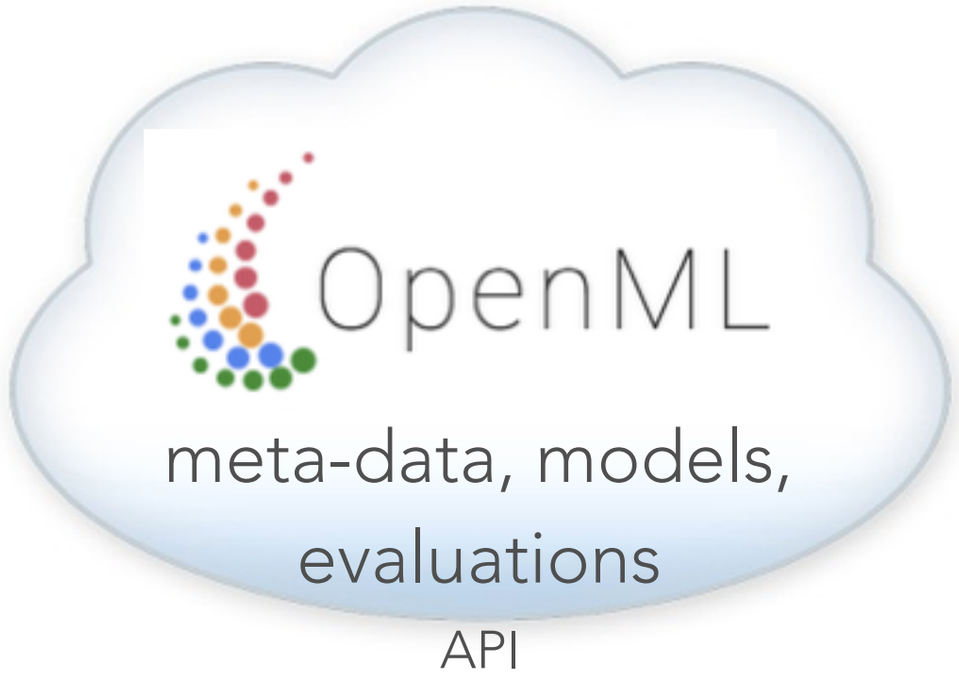


We just scratched the surface. Data will available on OpenML for many more studies and ideas, to test new algorithms,…

AutoML: automating machine learning

Learn from many datasets and experiments how to do data analysis

Create services



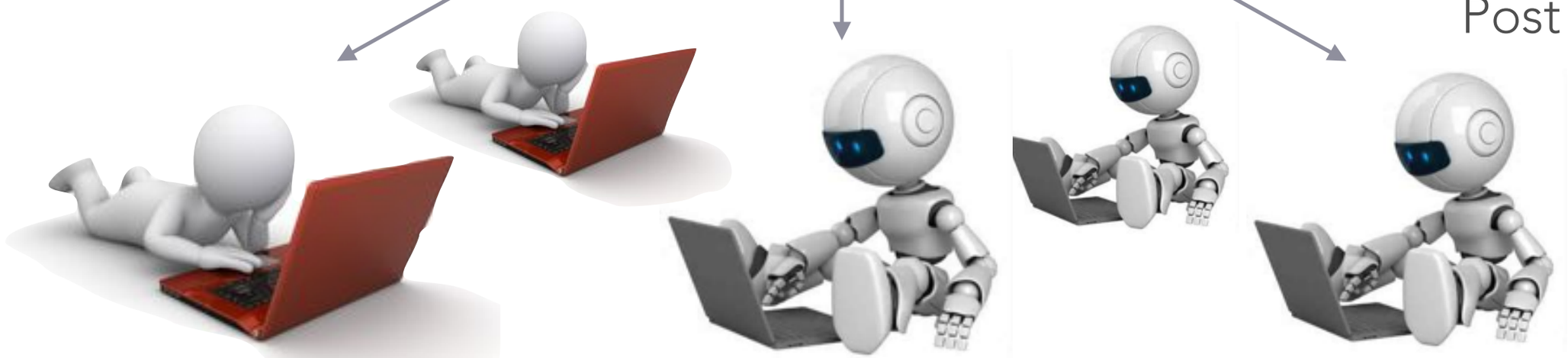
Data cleaning

Algorithm Selection

Parameter Optimization

Workflow construction

Post processing

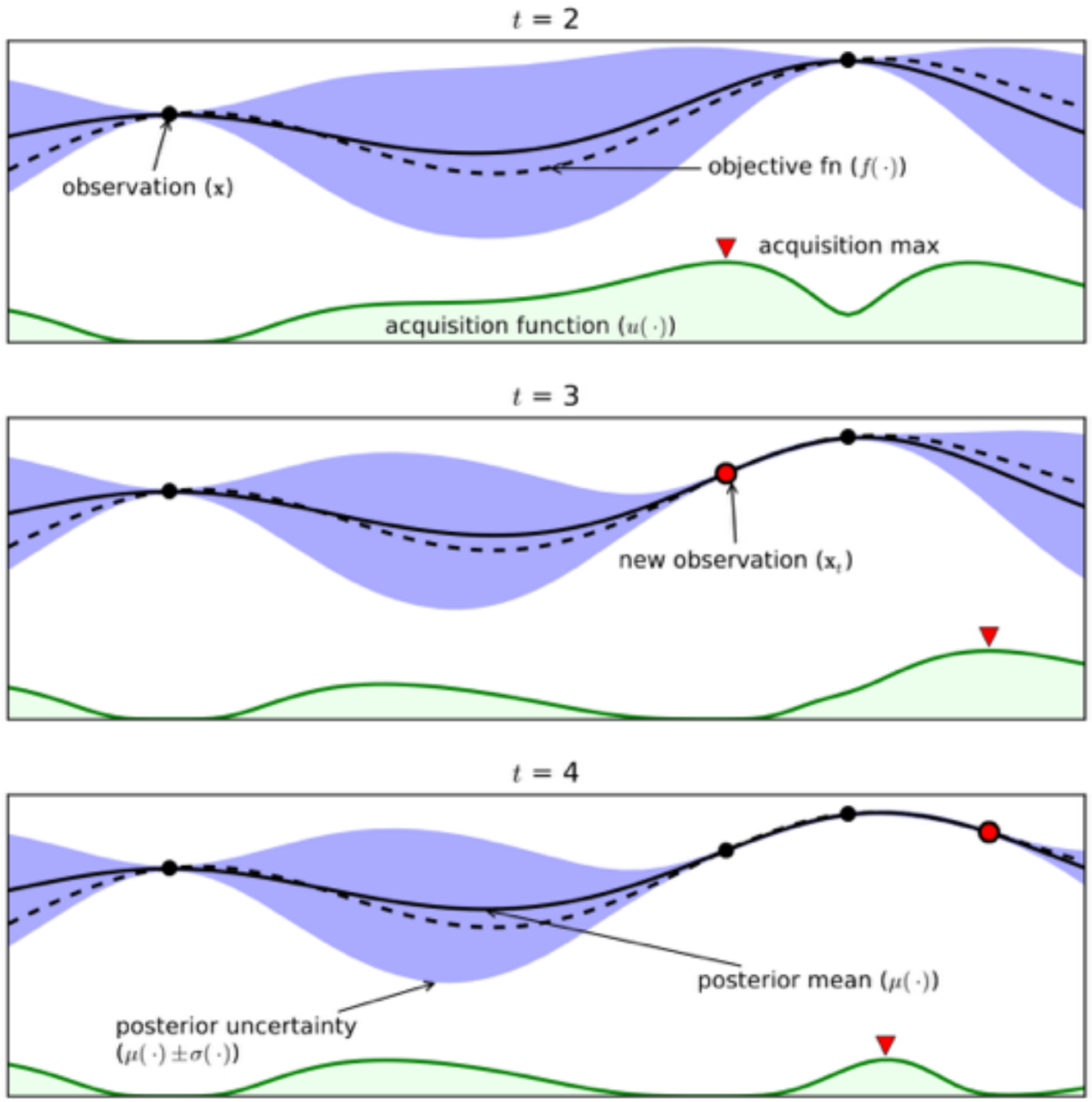


Human scientists

Automated processes

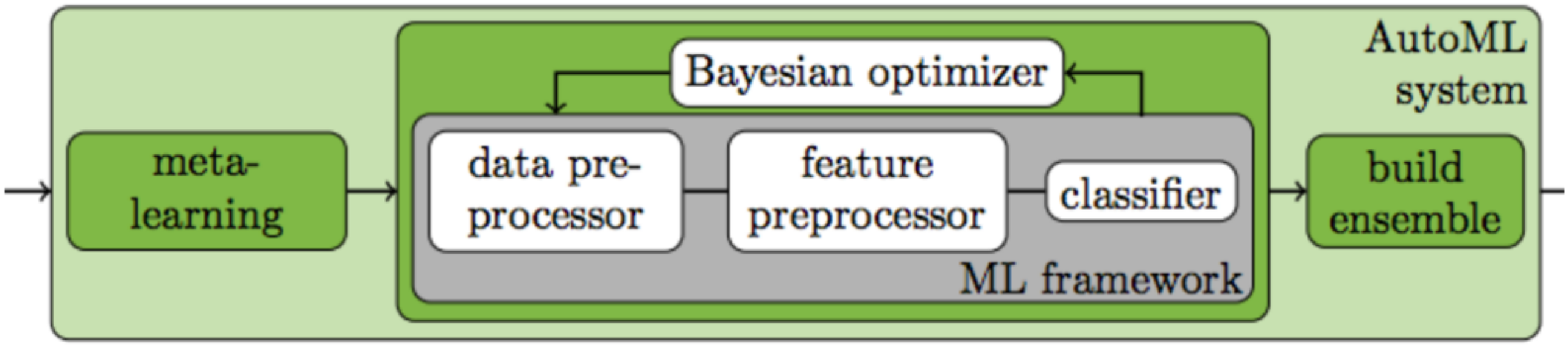
Automating ML

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



Automating ML

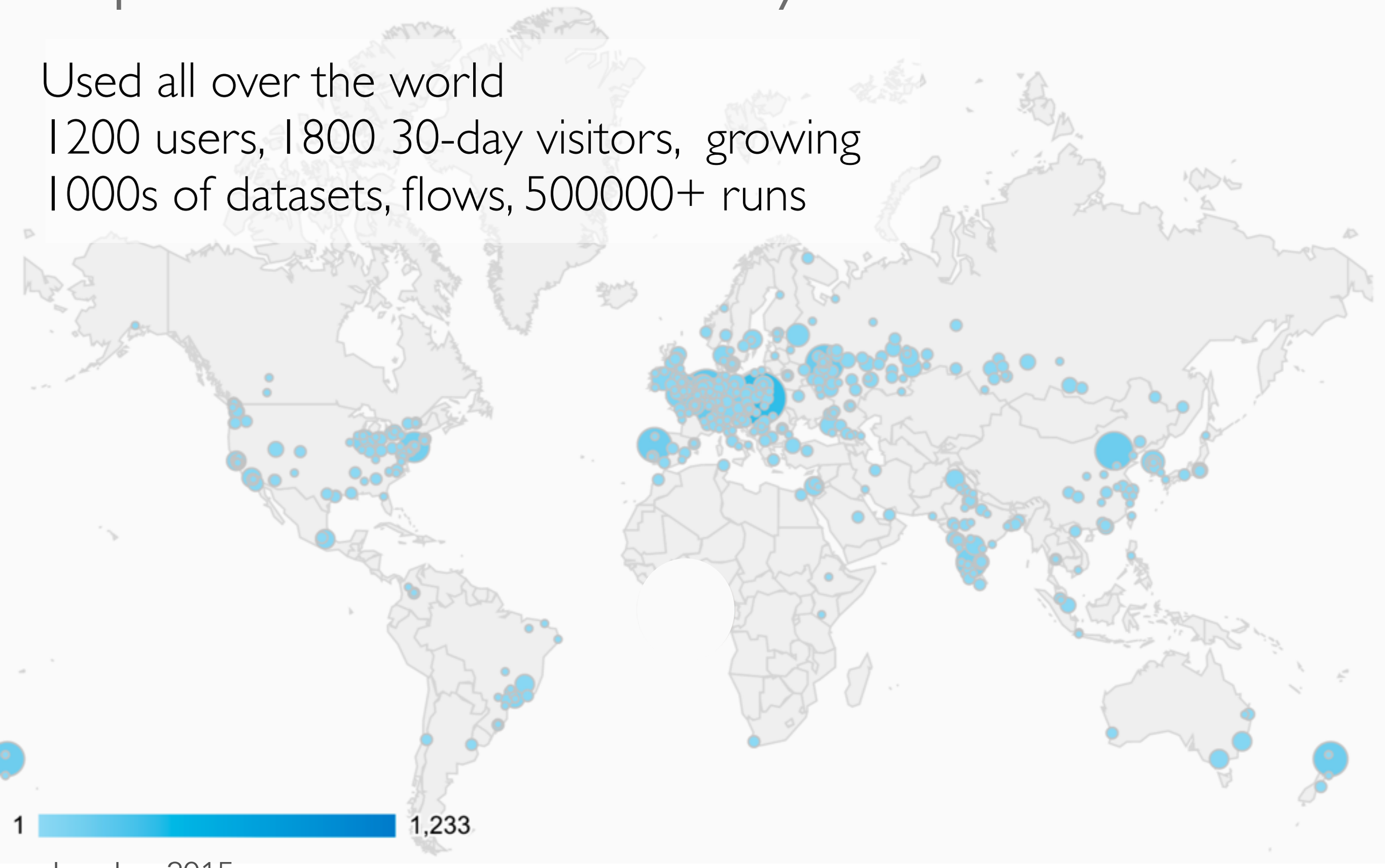
AutoML challenge: winning solution used OpenML and meta-learning
In Python: Auto-SKLearn



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

OpenML Community

Used all over the world
1200 users, 1800 30-day visitors, growing
1000s of datasets, flows, 500000+ runs



Jan-Jun 2015

Join OpenML

- Open Source, on GitHub
- Regular workshops, hackathons

Next workshop:

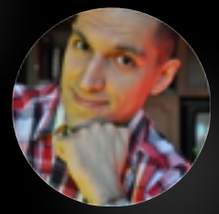
- Eindhoven (NL),
- 5-9 September 2016

THANK YOU

#OpenML

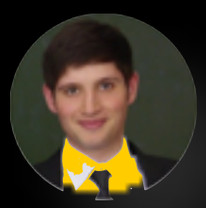

Jakob Bossek

Farzan Majdani


Nenad Tomašev


Luis Torgo


Jan van Rijn


Giuseppe Casalicchio


Joaquin Vanschoren


Michel Lang


Bernd Bischl


Matthias Feurer

You?