OpenML

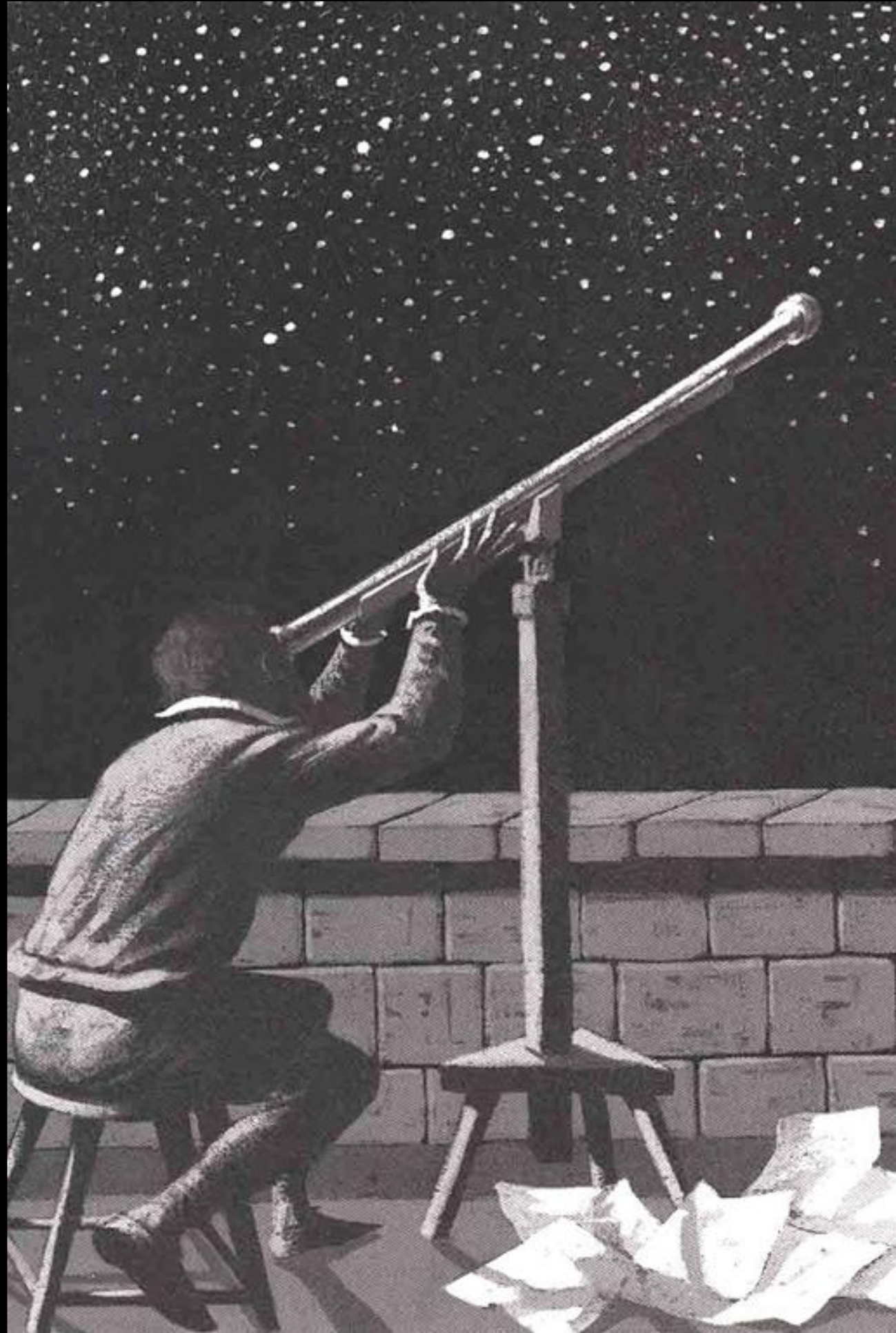
TOWARDS NETWORKED, AUTOMATED MACHINE LEARNING

Joaquin Vanschoren (TU/e) 2015

GALILEO GALILEI
DISCOVERS SATURN'S RINGS

1610

'SMAISMRMILMEPOETAL
EUMIBUNENUGTTAUIRAS'

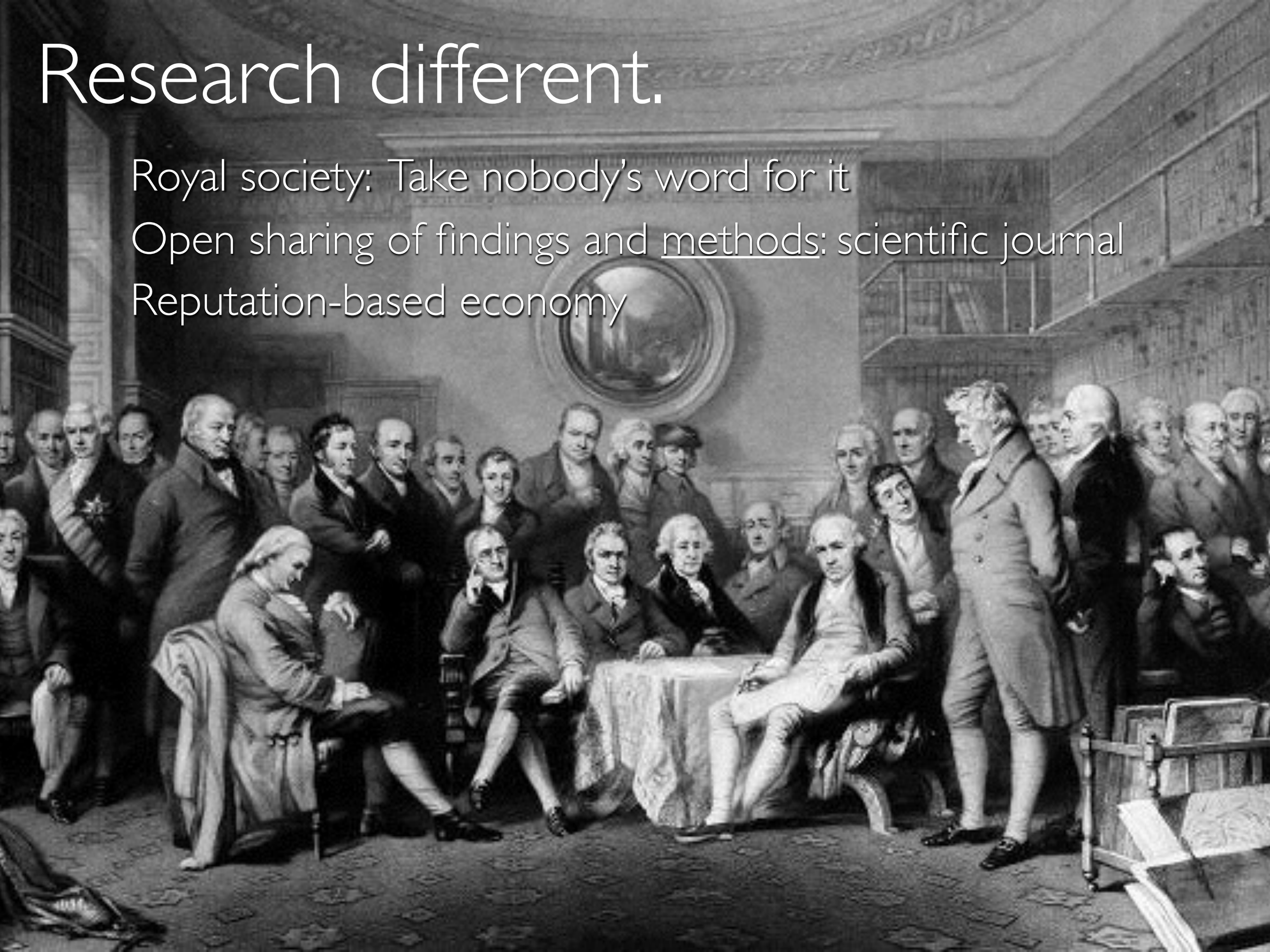


Research different.

Royal society: Take nobody's word for it

Open sharing of findings and methods: scientific journal

Reputation-based economy

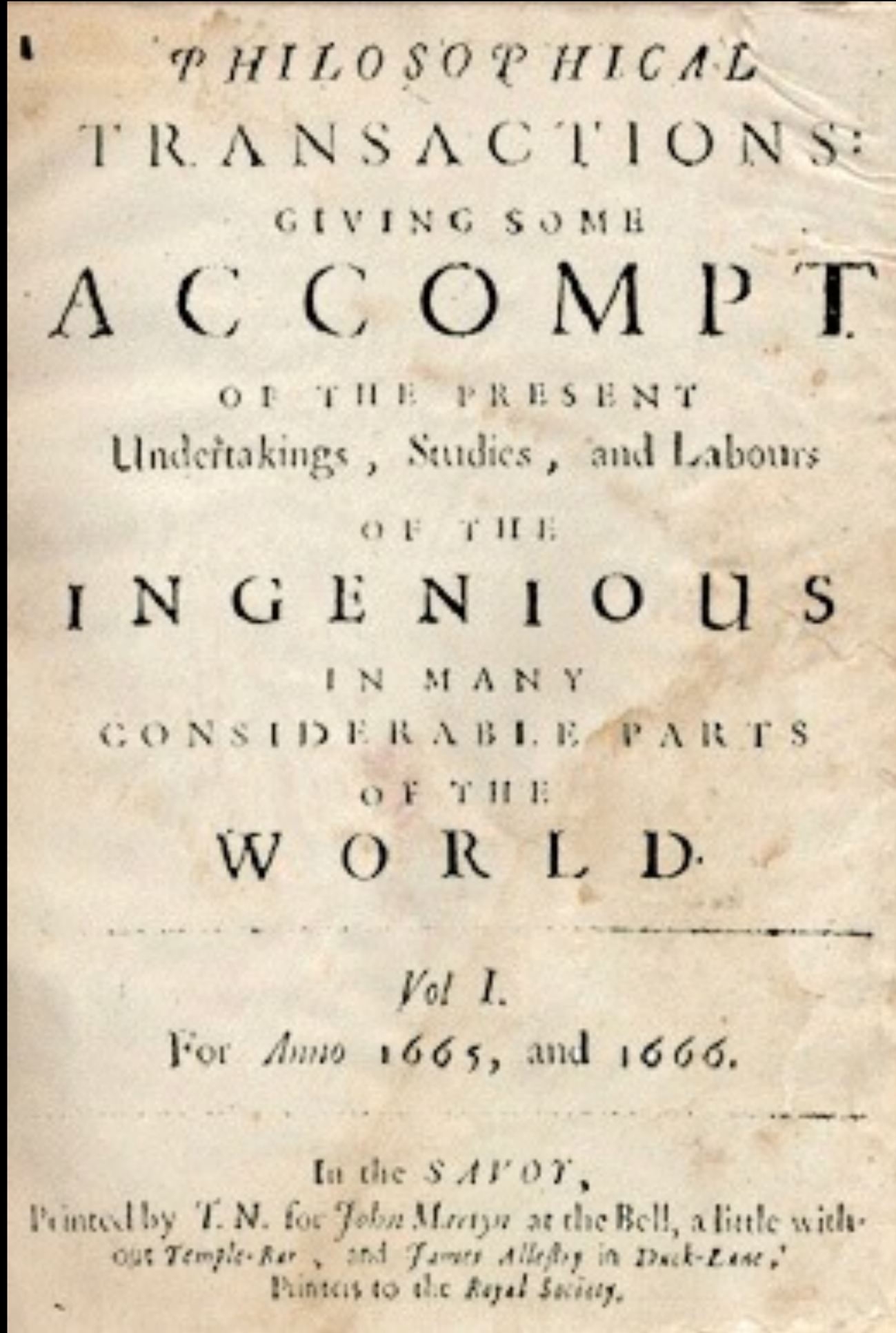


AFTER 300 YEARS

IS THE PRINTING PRESS STILL THE BEST MEDIUM?

FOR MACHINE LEARNING?

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

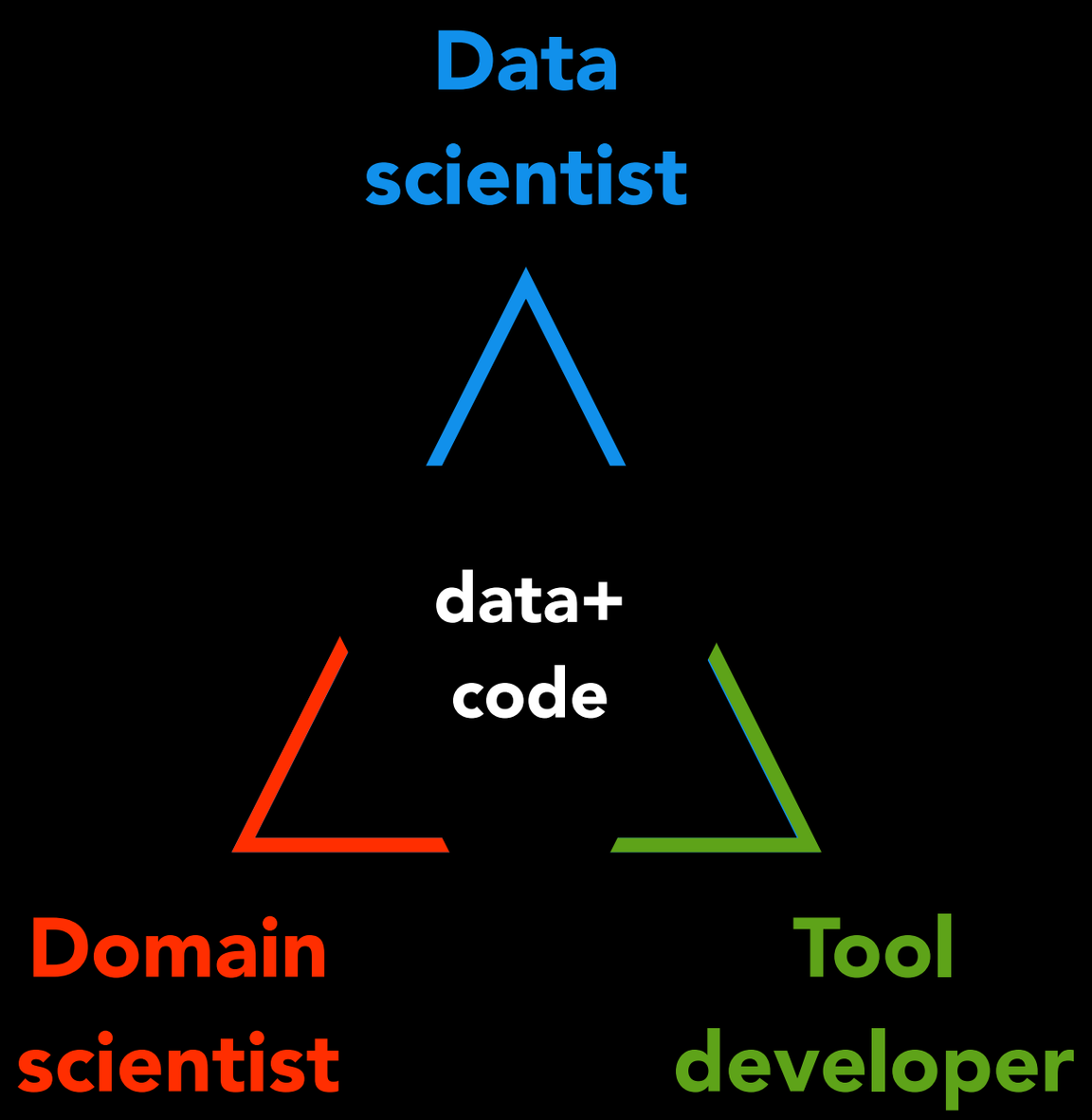


PHILOSOPHICAL
TRANSACTIONS:
GIVING SOME
ACCOMPT
OF THE PRESENT
Undertakings, Studies, and Labours
OF THE
INGENIOUS
IN MANY
CONSIDERABLE PARTS
OF THE
WORLD.

Vol I.
For Anno 1665, and 1666.

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

Gaps in data-driven science



Domain scientists: doubts about latest/best techniques, run simple models on complex data, small (biased) collaborations

Data scientists: don't speak the language, don't know how to access scientific databases, run complex models on simple data

Industry: small companies delimitated by lack of access to data/expertise

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.Ds 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

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

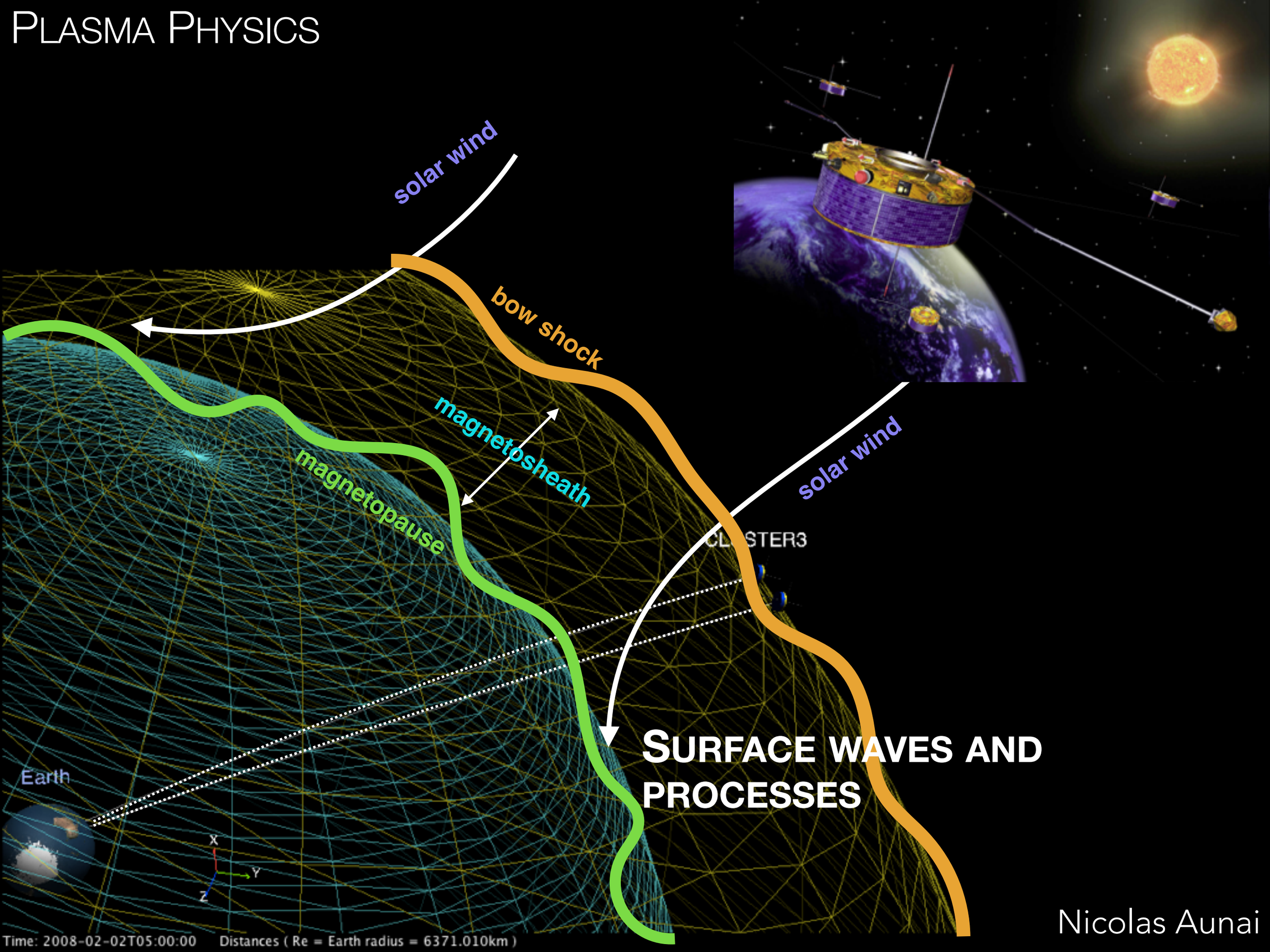
Research findings less likely to be true if:

- Studies are smaller, few collaborators
- Effect sizes are smaller
- Flexibility in designs, definitions, outcomes, analytical modes
- Teams are chasing statistical significance

Increase credibility:

- Large-scale interdisciplinary collaboration
- Replication culture, reproducibility
- Registration/sharing of data
- Better statistical methods, models
- Better study design, training

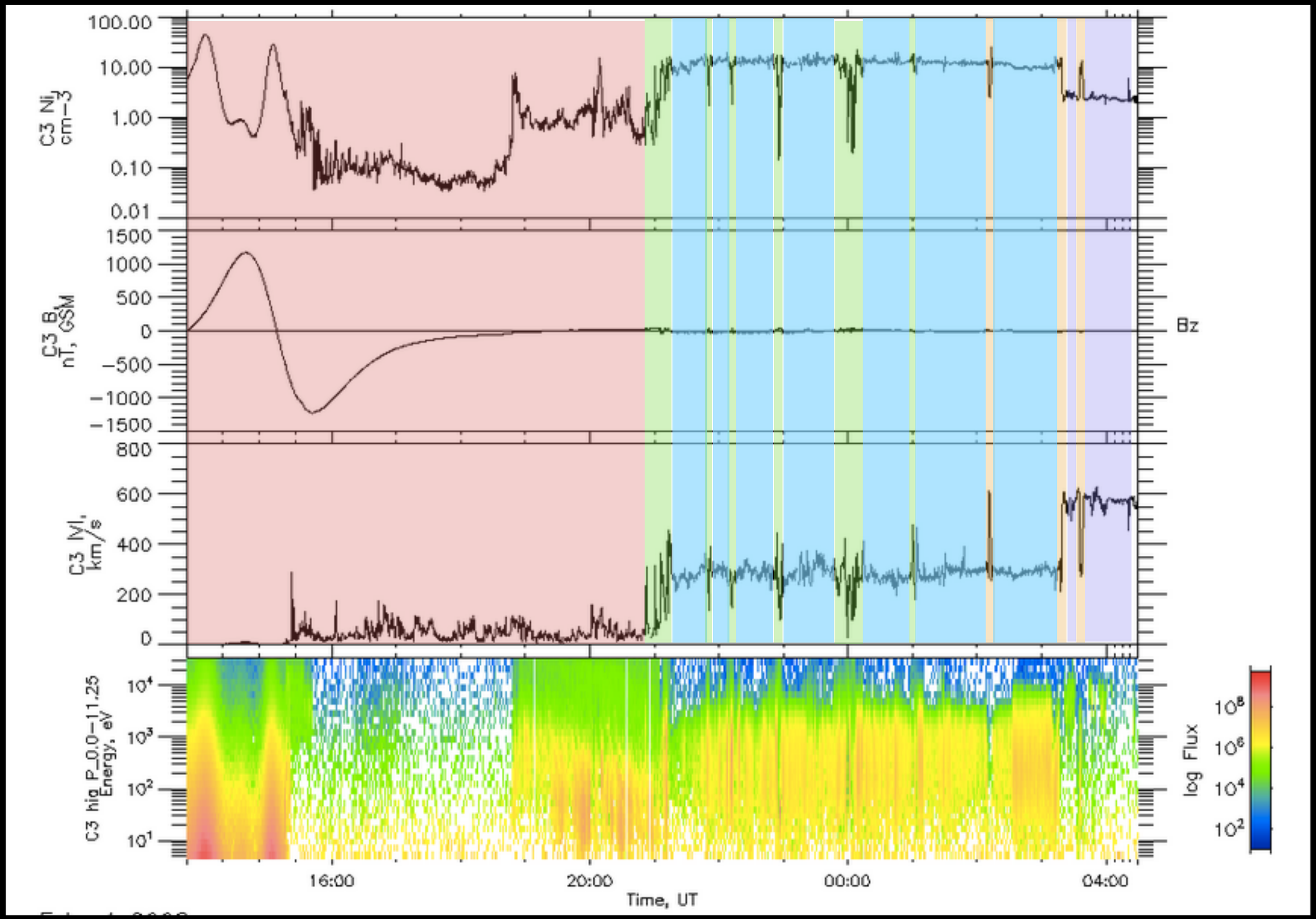
Plasma Physics



Surface waves and processes

Nicolas Aunai

LABELLING (CATALOGUING) OF EVENTS STILL DONE MANUALLY



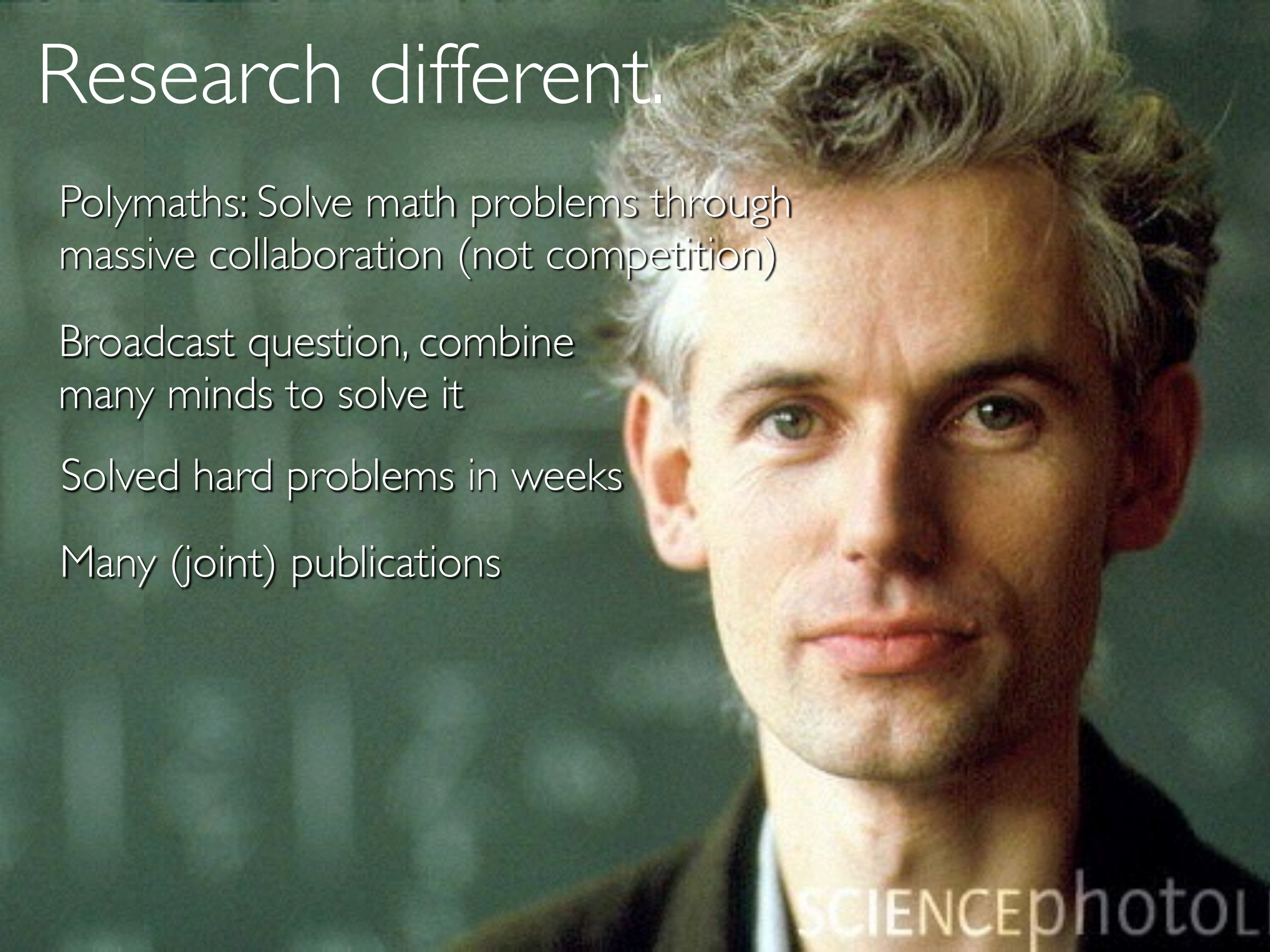
Research different.

Polymaths: Solve math problems through massive collaboration (not competition)

Broadcast question, combine many minds to solve it

Solved hard problems in weeks

Many (joint) publications



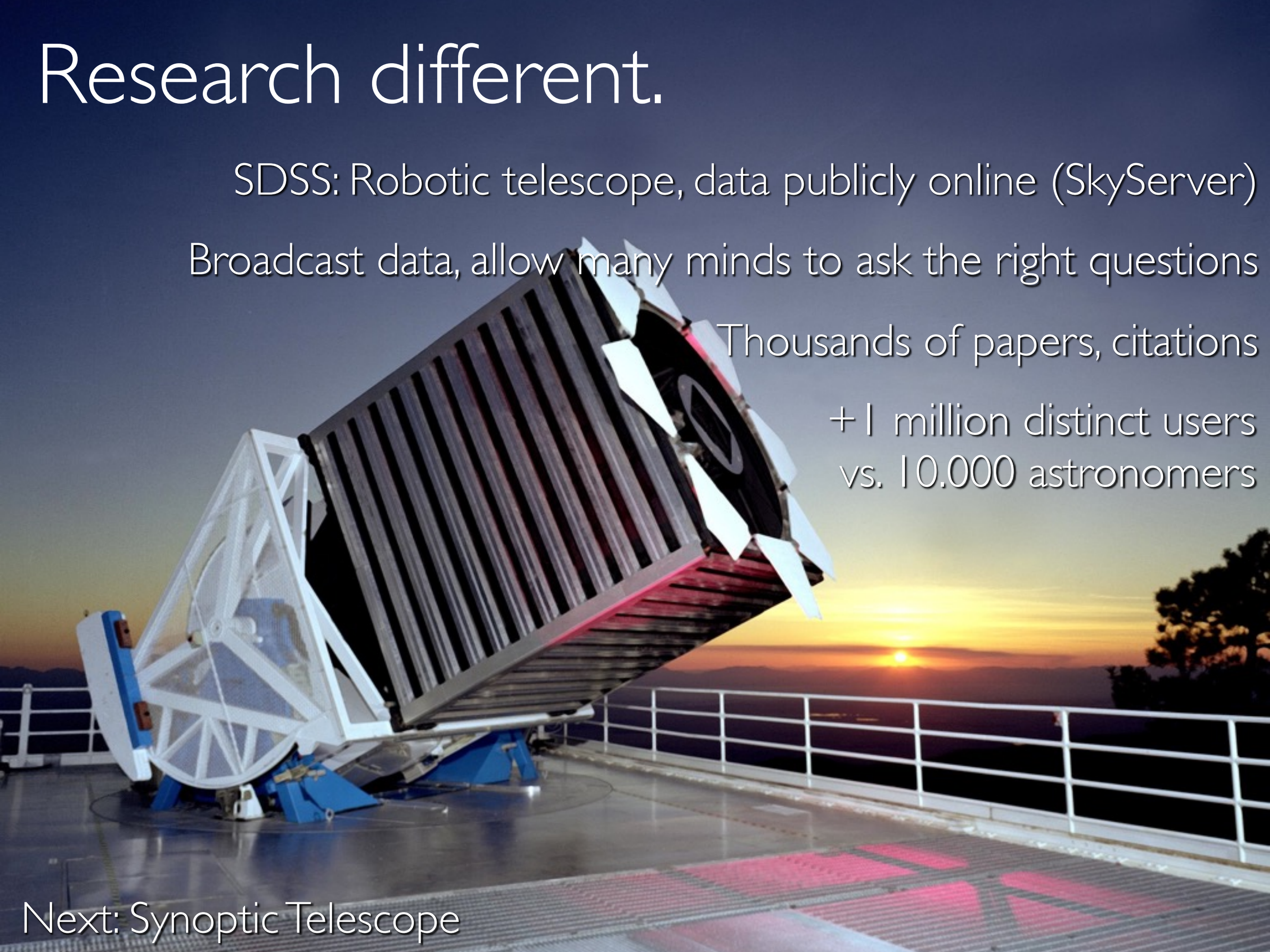
Research different.

SDSS: Robotic telescope, data publicly online (SkyServer)

Broadcast data, allow many minds to ask the right questions

Thousands of papers, citations

+1 million distinct users
vs. 10.000 astronomers



Next: Synoptic Telescope

Research different.

How do you label a million galaxies?

Offer right tools so that anybody can contribute, instantly

Many novel discoveries by scientists <u>and citizens</u>

More? Read ‘Networked science’ by M. Nielsen

Why? Designed serendipity

Broadcasting data fosters spontaneous, unexpected discoveries

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

How? Remove friction

Organized body of compatible scientific data (and tools) online

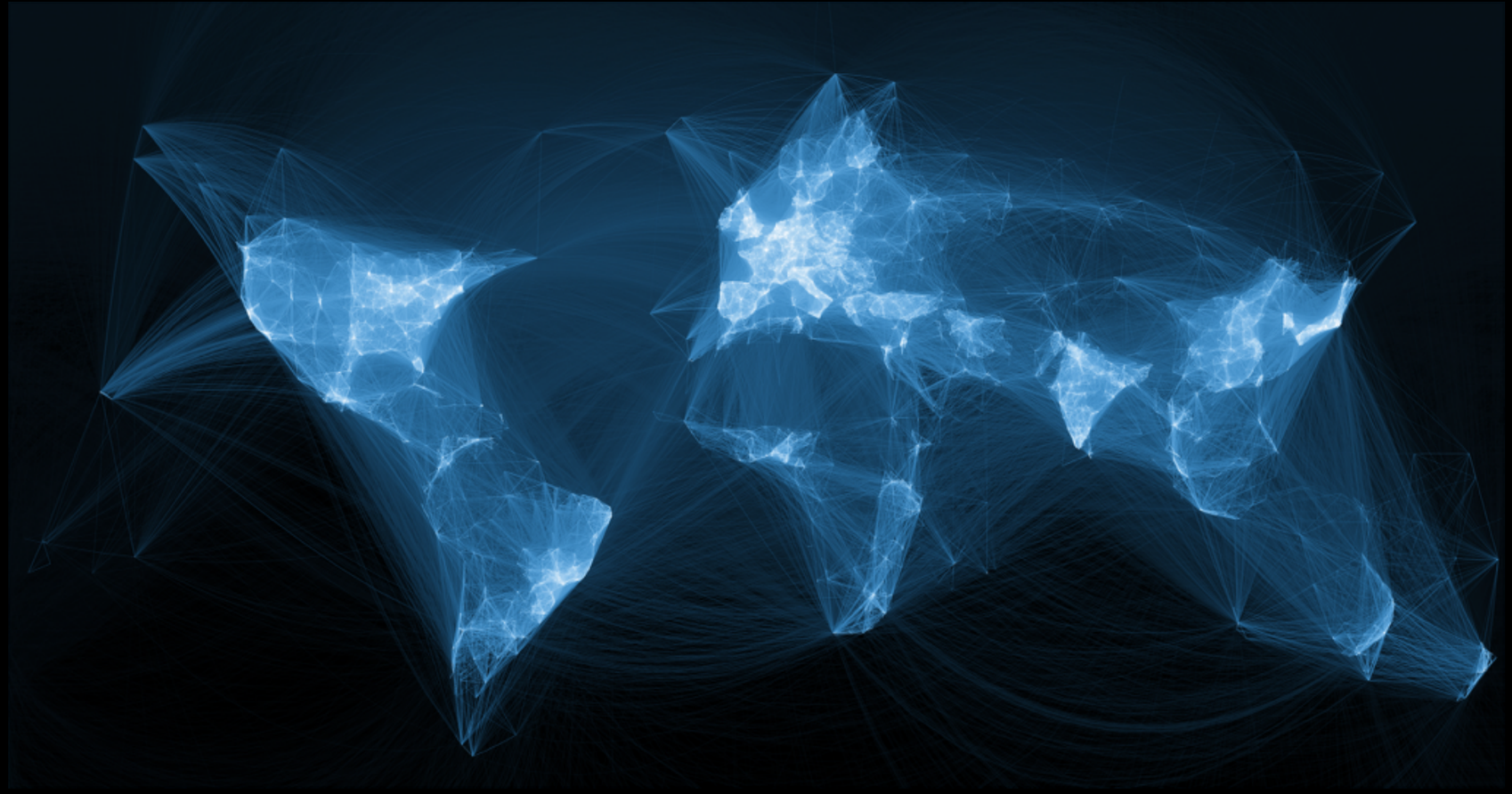
Micro-contributions: seconds, not days

Easy, organised communication

Track who did what, give credit

OpenML

A REALTIME, WORLDWIDE LAB



FRICTION-LESS ENVIRONMENT FOR MACHINE LEARNING RESEARCH

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

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

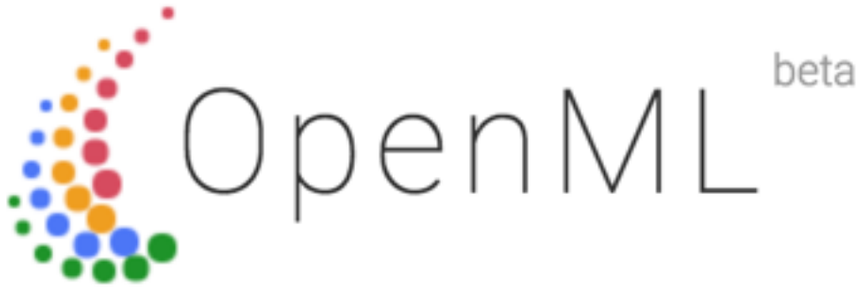
Micro-contributions: Upload single dataset, algorithm, experiment

Easy communication: Online discussions per dataset, algorithm, experiment

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

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

OpenML Search



Exploring machine learning better, together

1345 data sets

Find or add **data** to analyse

4830 tasks

Download or create scientific **tasks**

1400 flows

Find or add data analysis **flows**

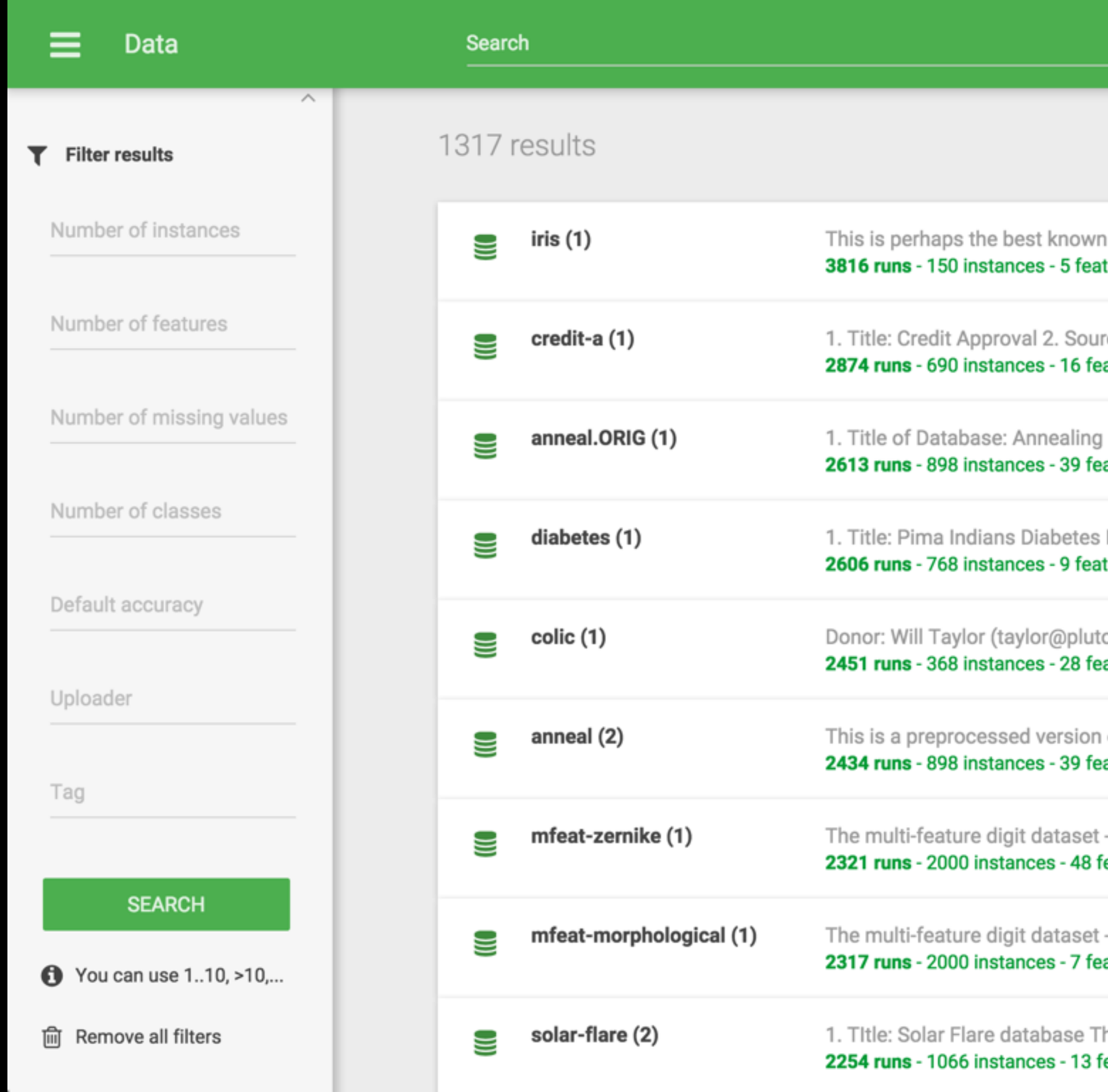
452379 runs

Upload and explore all **results** online.

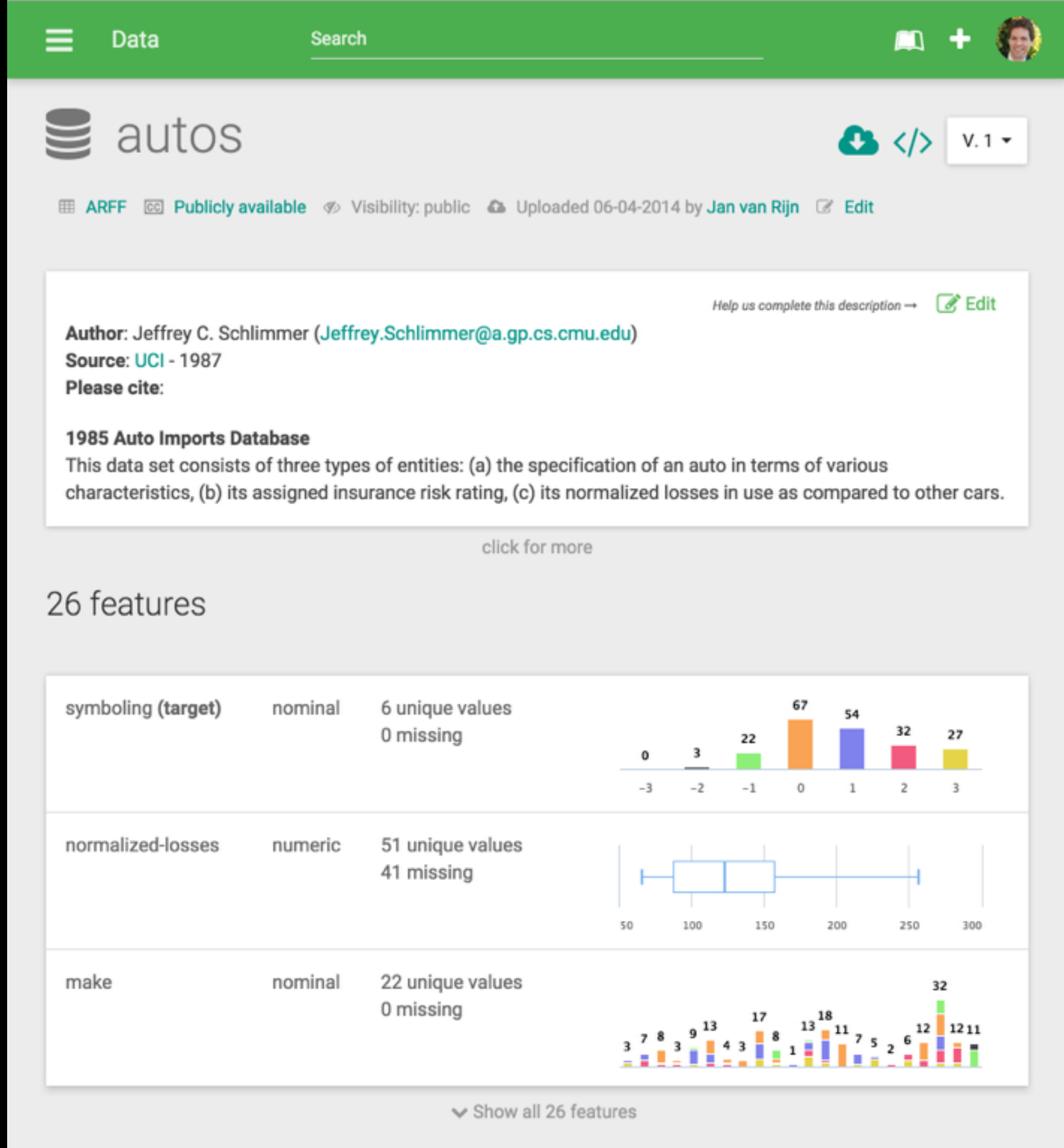
Data from
various sources
analysed and
organised online
for easy access

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

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



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



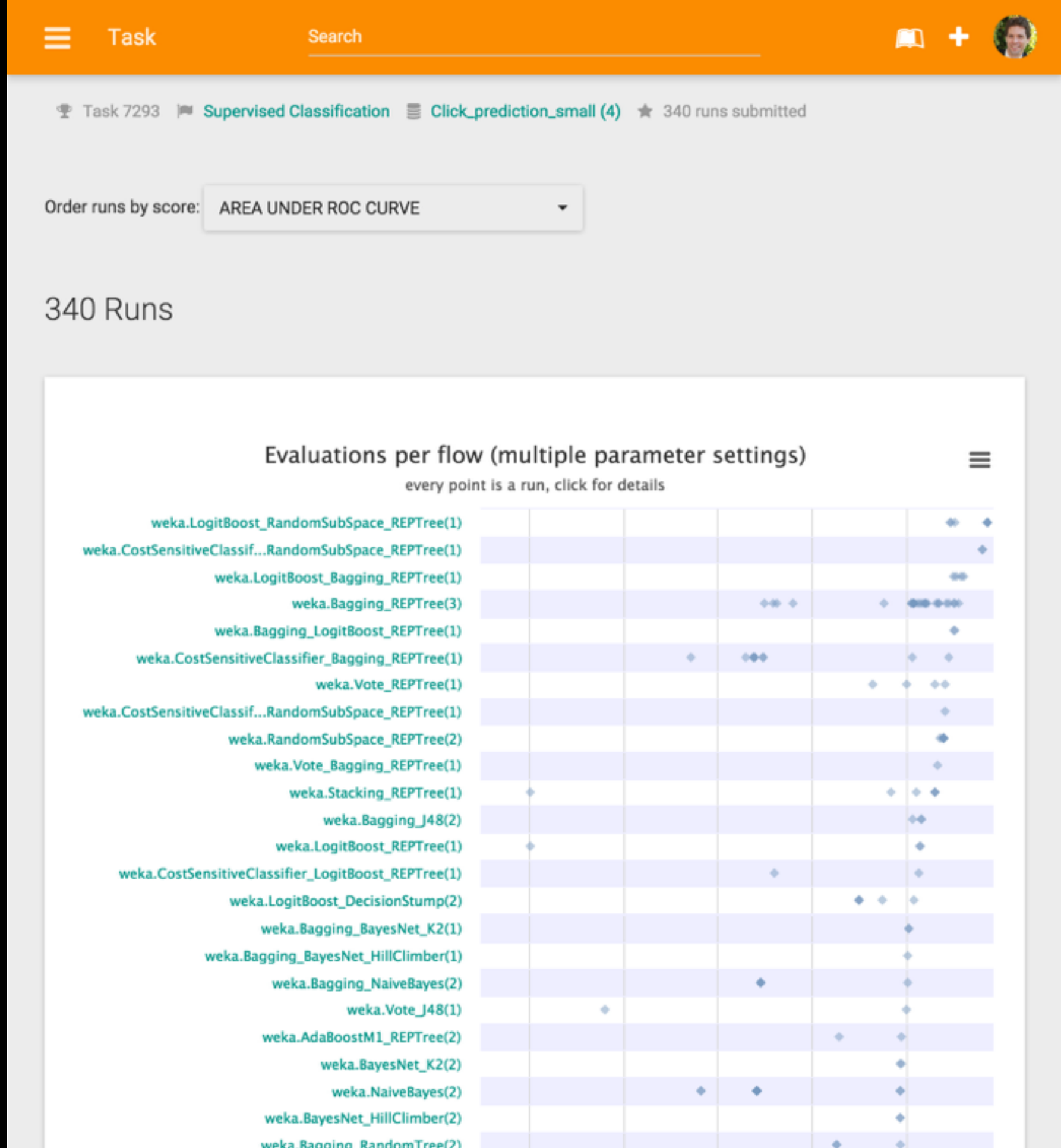
Scientific tasks that can be **interpreted by tools** and **solved collaboratively**

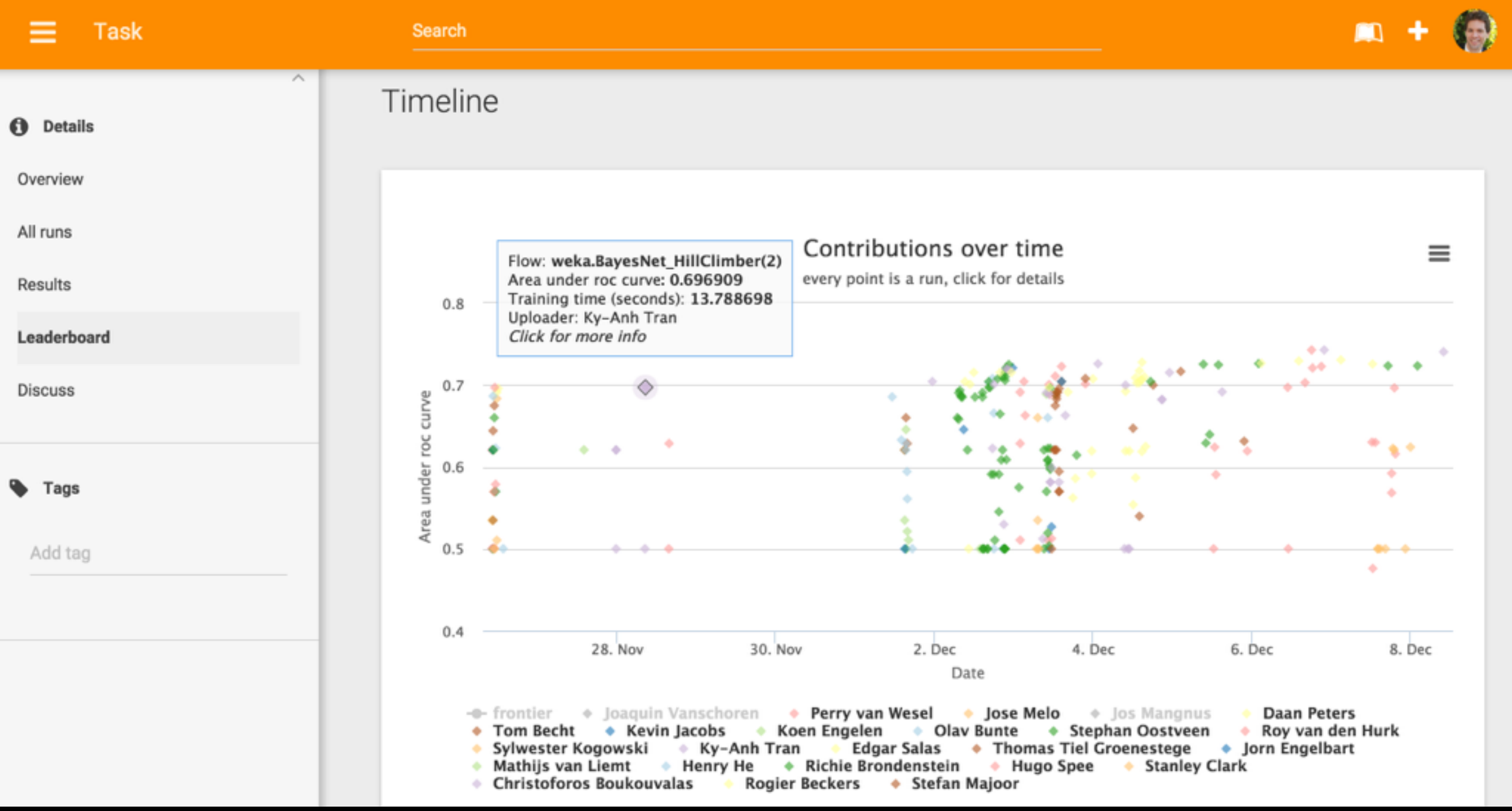
Tasks: containers with all data, goals, procedures.

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

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

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





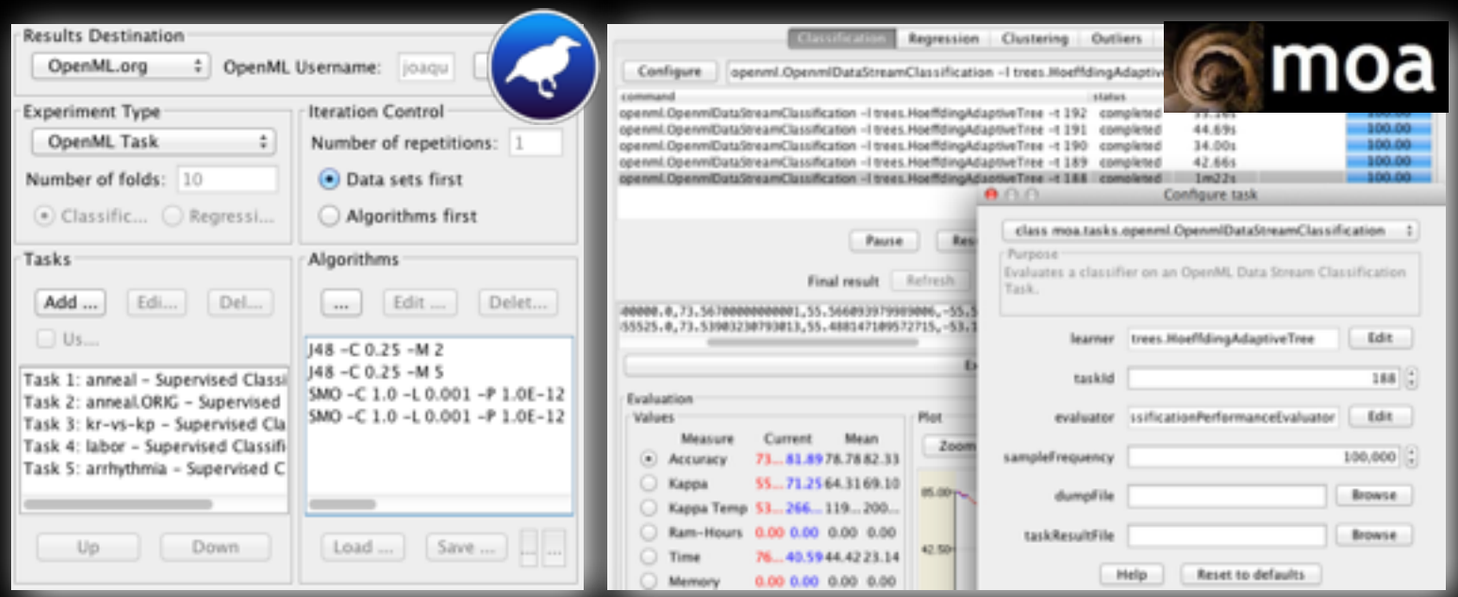
- Leaderboards visualize progress over time: who delivered breakthroughs when, who built on top of previous solutions
- Collaborative: all code and data available, learn from others, form teams
- Real-time: who submits first gets credit, others can improve immediately

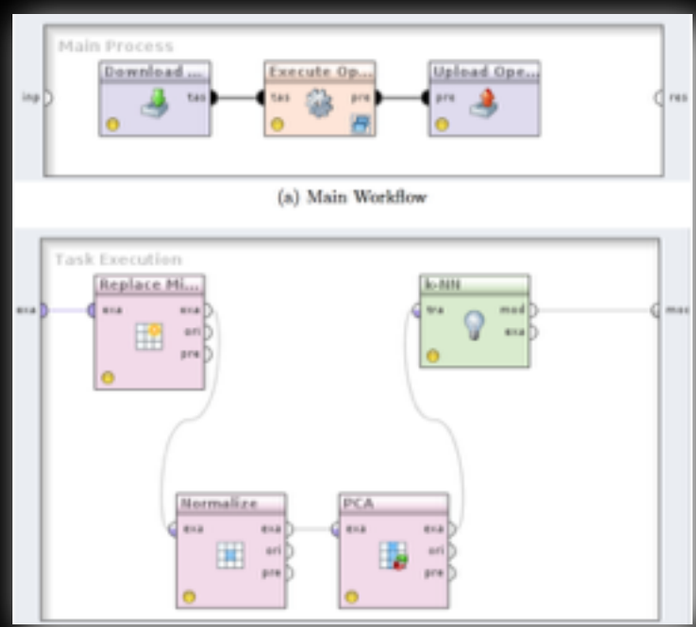
Machine learning flows (code) that can **solve tasks** and **report results.**

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

REST API + Java, R, Python APIs

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



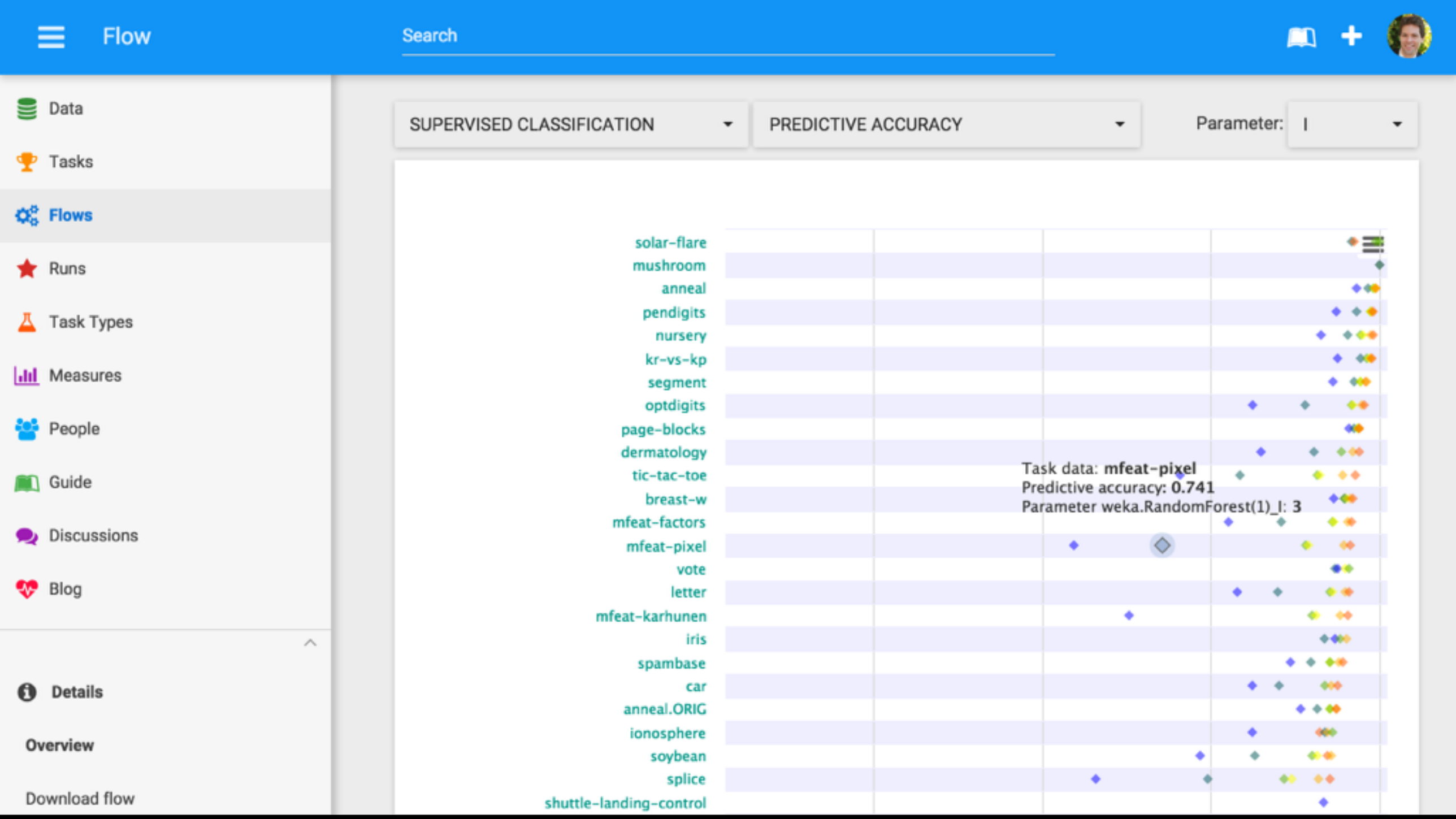


- RapidMiner plugin: new operators to load tasks, export results and subworkflow

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

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

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



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

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

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

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

Run

Run 84087

JSON XML

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

Flow

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

Result files

Description
XML file describing the run, including user-defined evaluation measures.

Model readable
A human-readable description of the model that was built.

Model serialized
A serialized description of the model that can be read by the tool that generated it

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

Area under ROC curve

0.7007 ± 0.0023

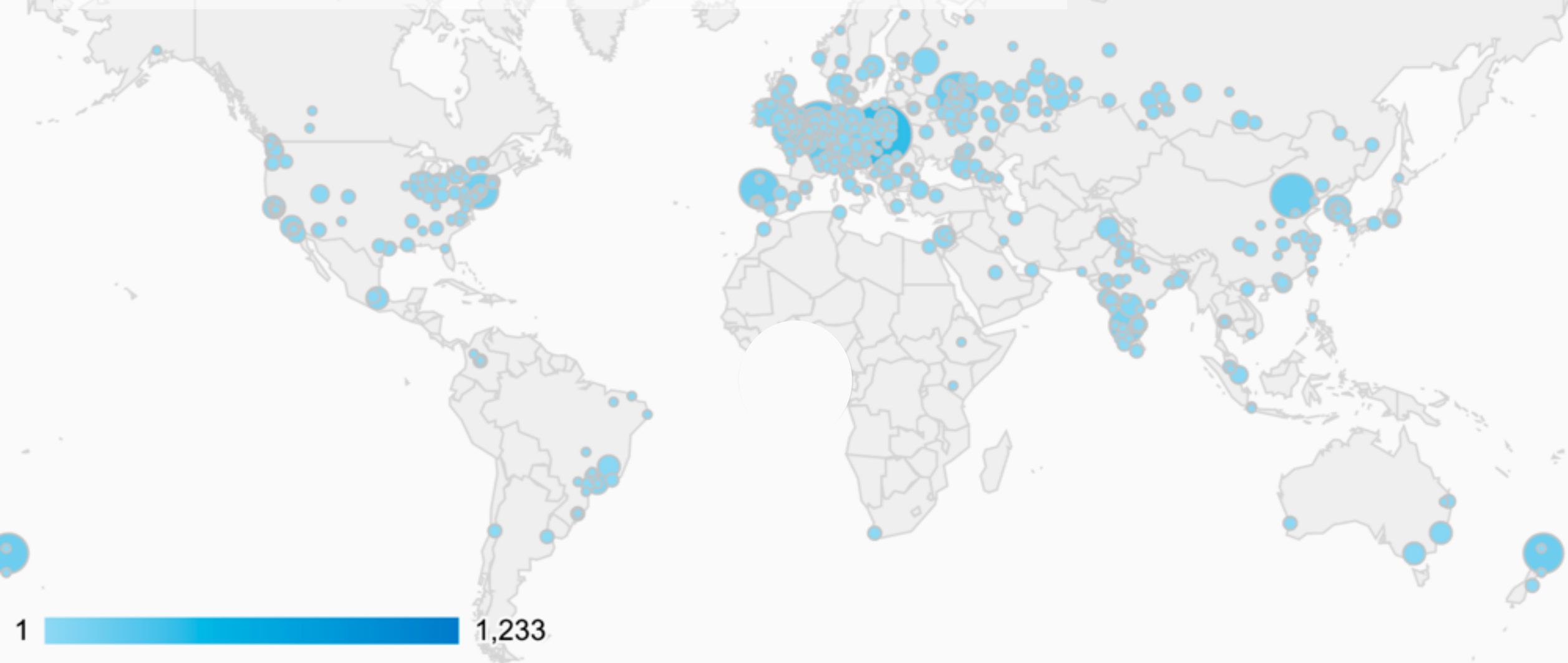
Per class

0	1
0.7007	0.7007

Cross-validation details (10-fold Crossvalidation)

OpenML Community

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



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

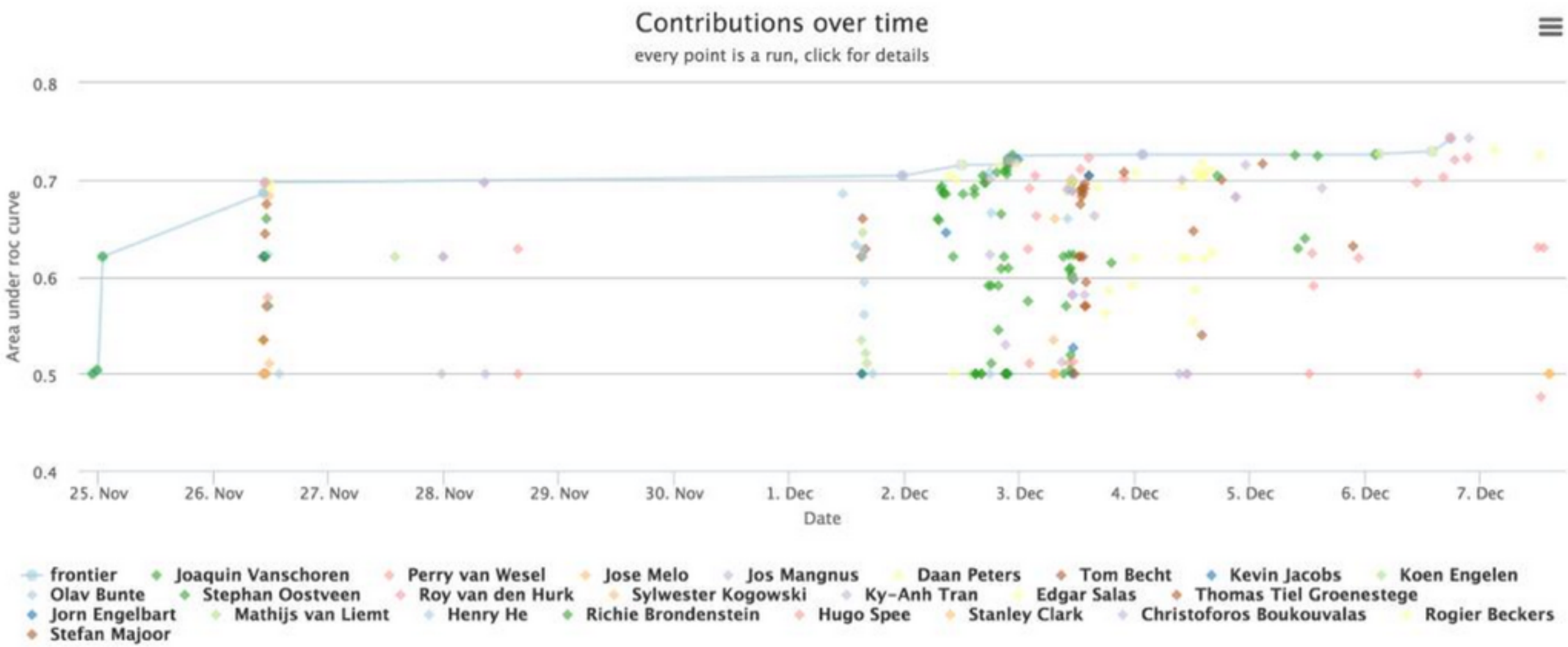
Statistical analysis

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

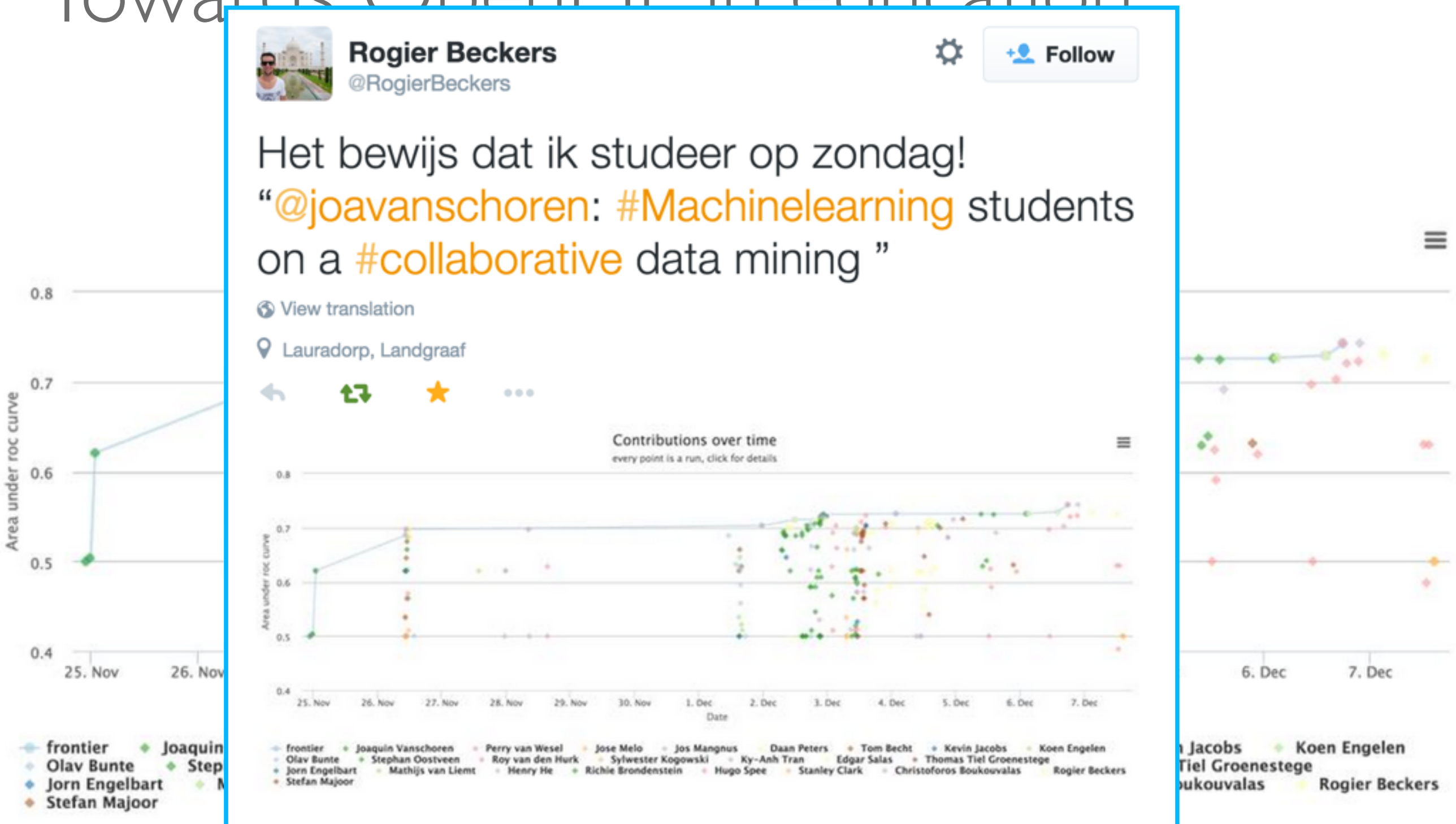
Online task creation

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

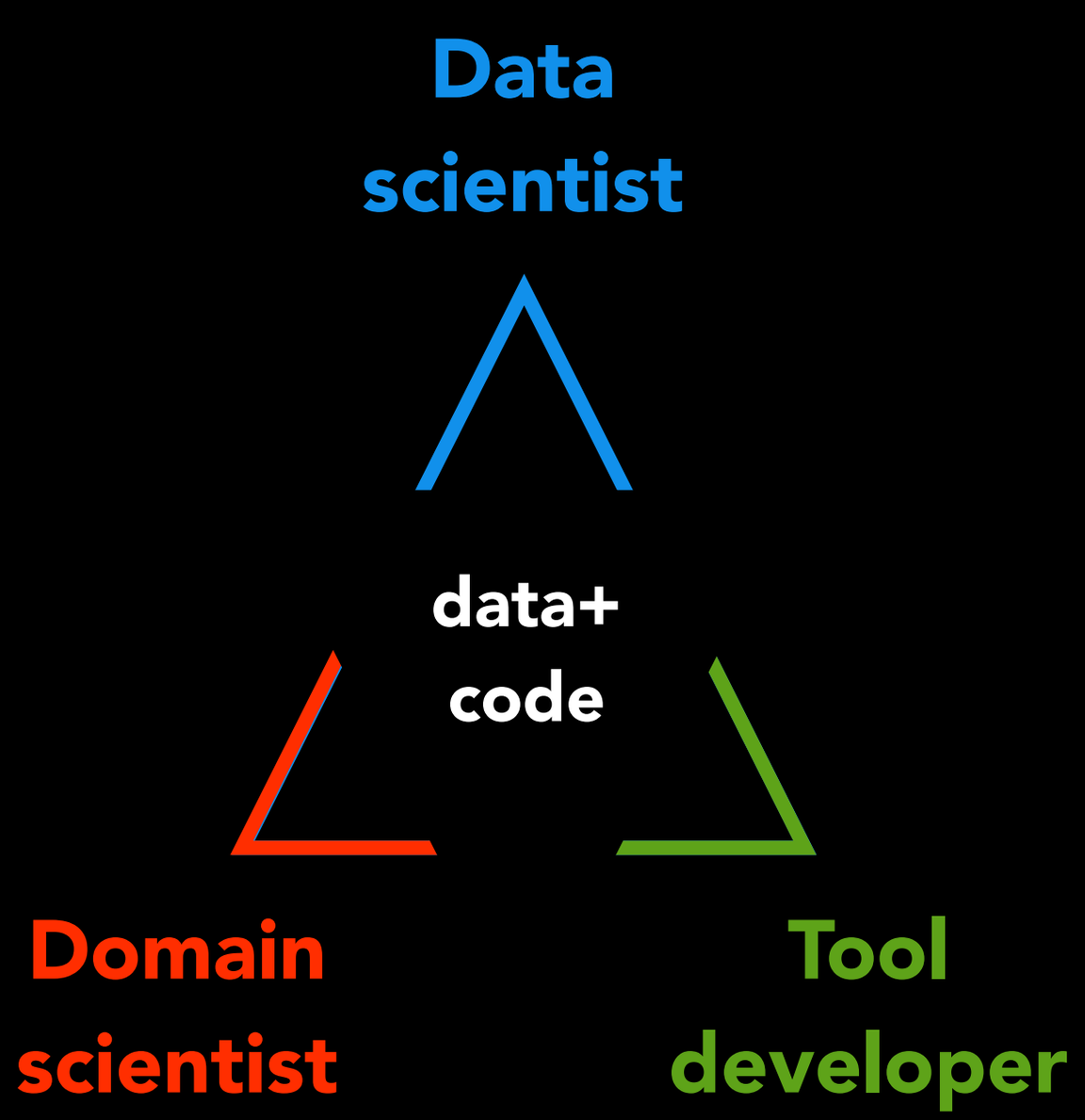
Towards OpenML in education



Towards OpenML in education



Bridging gaps in data-driven science

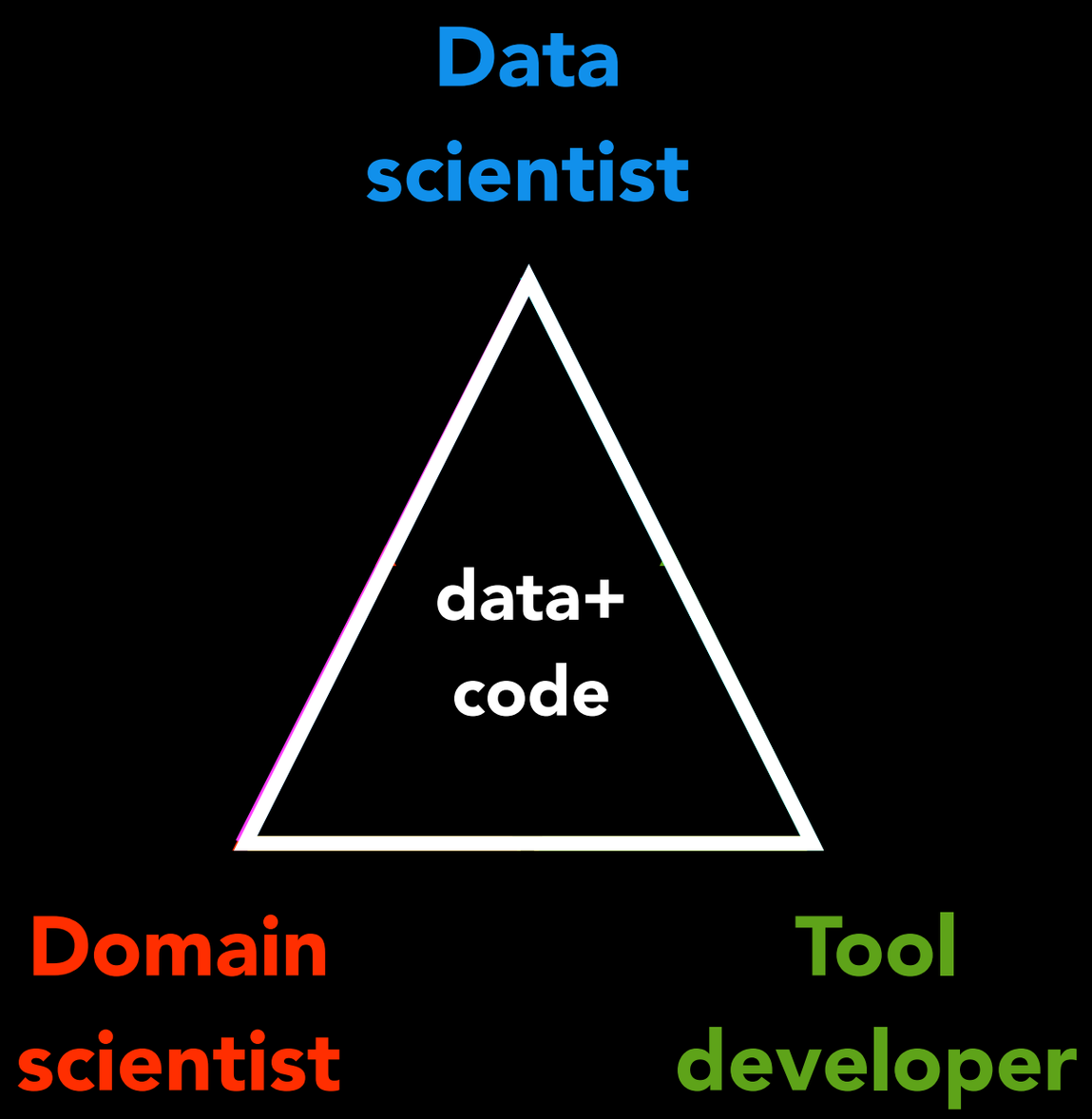


Collaboratory: bring data scientists and domain scientists together online (and their data and tools)

Easy, large-scale collab: Extract actionable datasets, key tools. Scientists shares data and get help, DS can test technique on many current datasets.

Real-time collab: share experiments automatically, discuss online. Automate experimentation.

Bridging gaps in data-driven science

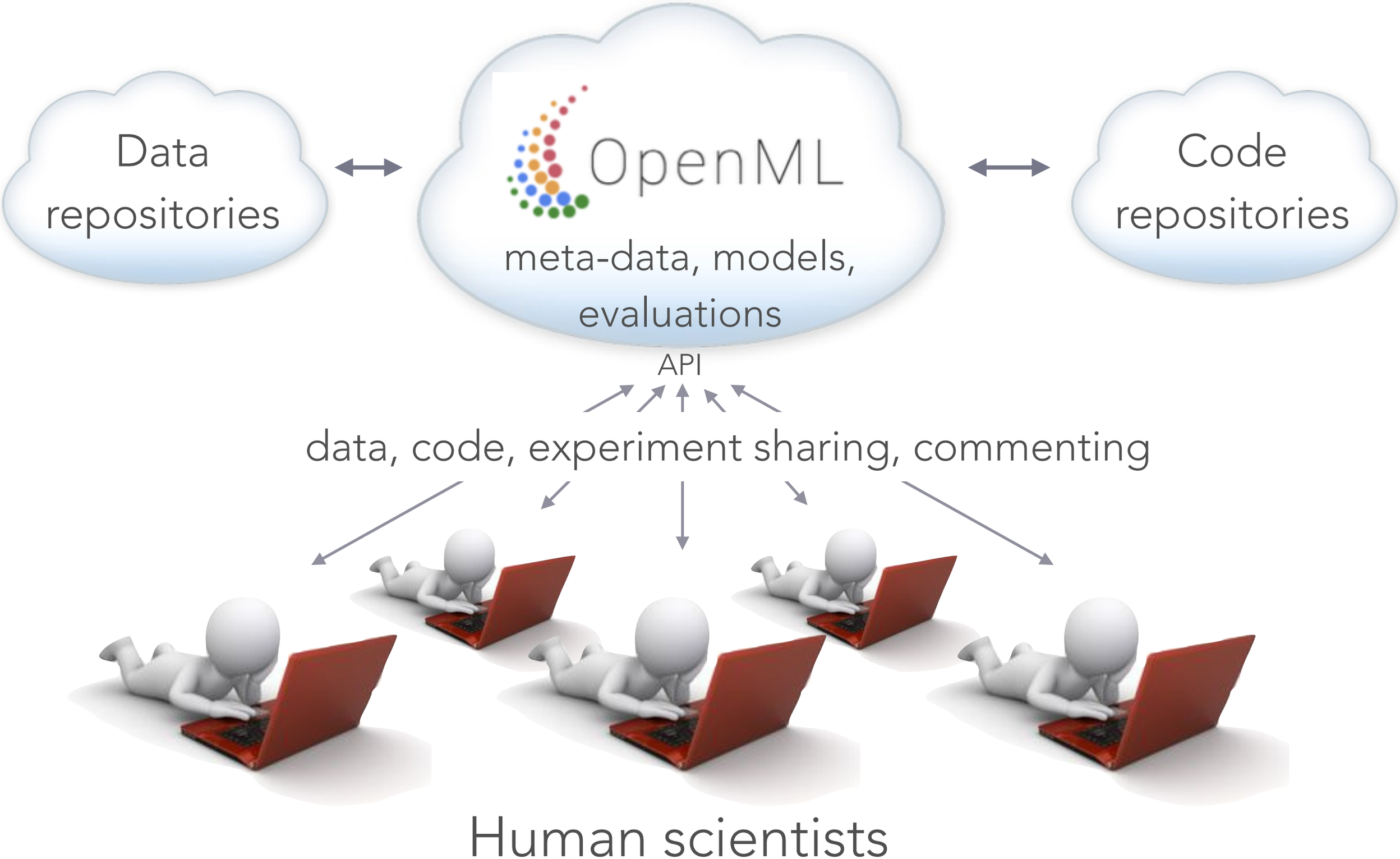


Collaboratory: bring data scientists and domain scientists together online (and their data and tools)

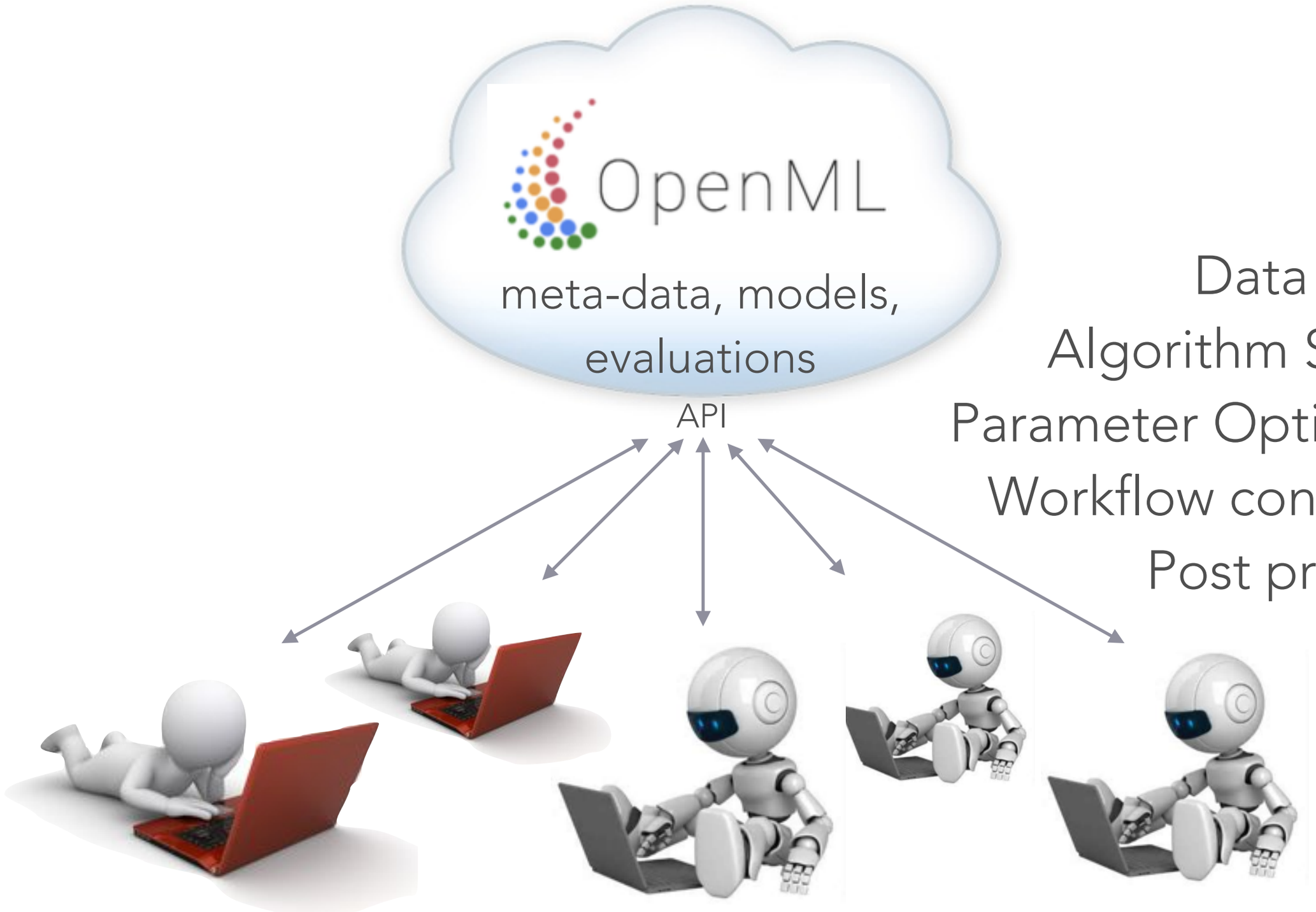
Easy, large-scale collab: Extract actionable datasets, key tools. Scientists shares data and get help, DS can test technique on many current datasets.

Real-time collab: share experiments automatically, discuss online. Automate experimentation.

Networked data(-driven) science



Automating machine learning



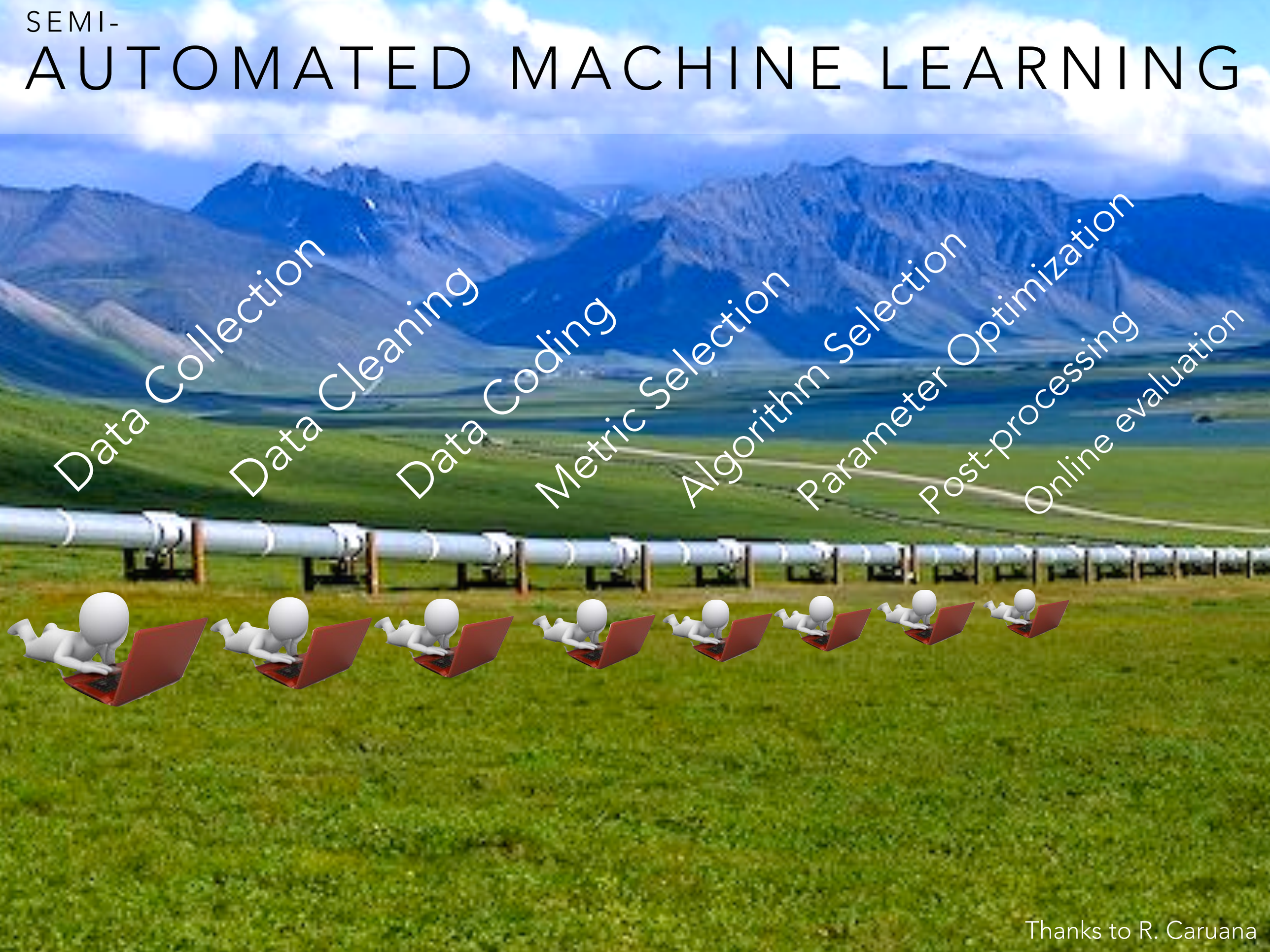
Data cleaning
Algorithm Selection
Parameter Optimization
Workflow construction
Post processing

Human scientists

Automated processes

SEMI-

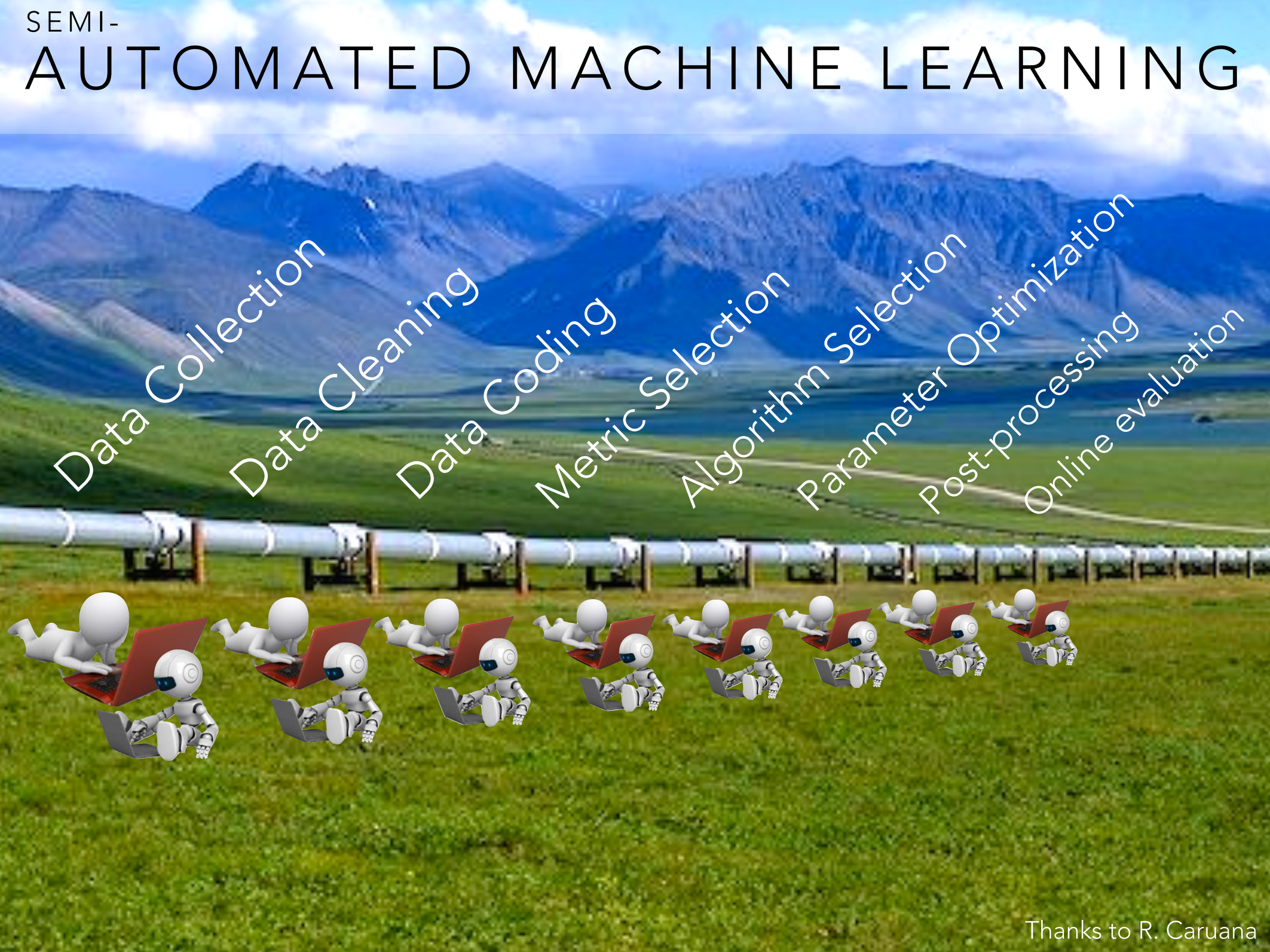
AUTOMATED MACHINE LEARNING



Thanks to R. Caruana

SEMI-

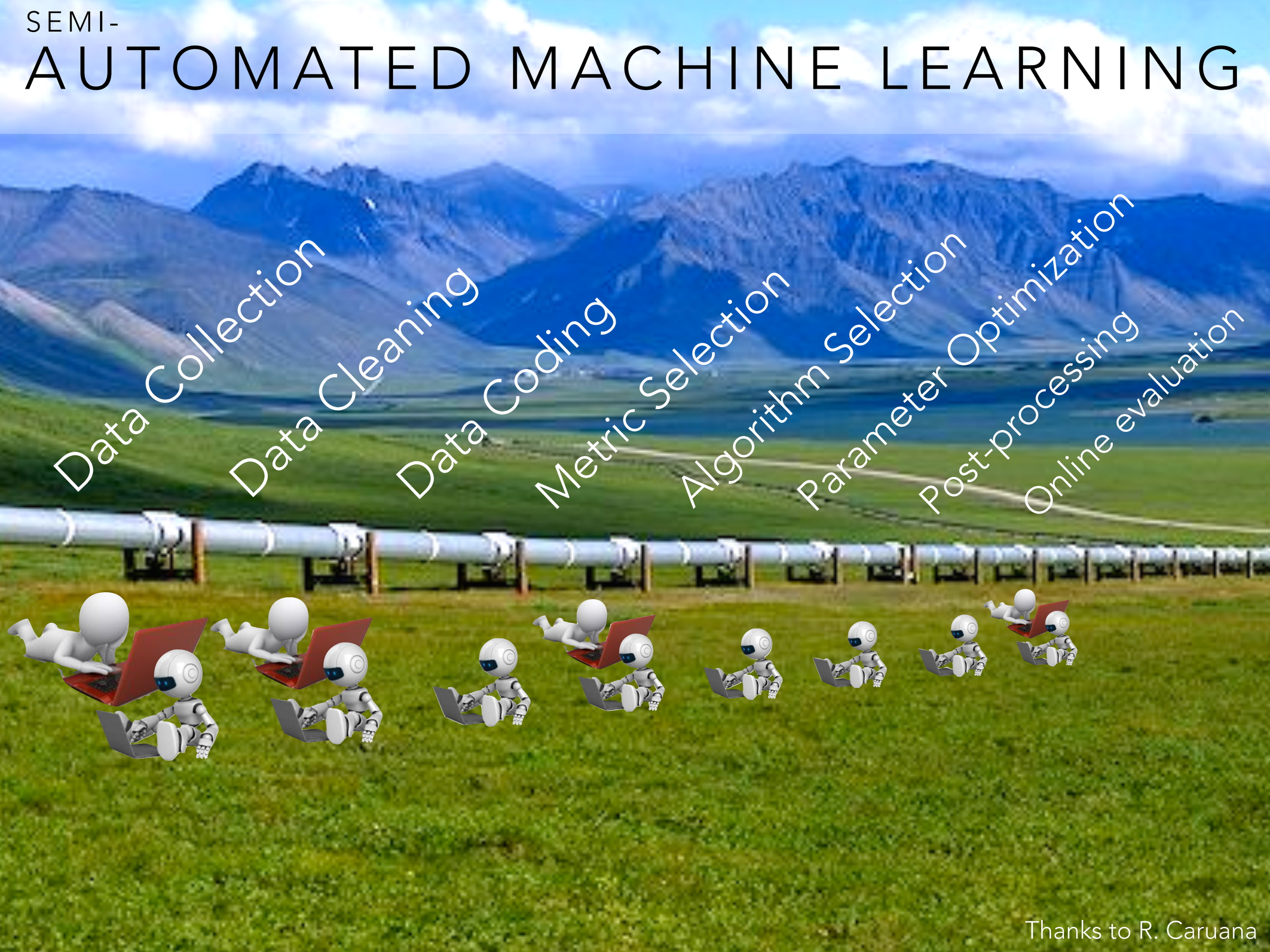
AUTOMATED MACHINE LEARNING



Thanks to R. Caruana

SEMI-

AUTOMATED MACHINE LEARNING



Thanks to R. Caruana

MACHINE LEARNING PIPELINE

Data Collection
Data Cleaning
Data Coding
Metric Selection
Algorithm Selection
Parameter Optimization
Post-processing
Online evaluation

Manual heuristic search: surprisingly suboptimal
Grid search: only effective with very small number of parameters
Random search: better with larger number of parameters
Bayesian Optimization: better with very large number of parameters

Critical with modern (low bias) algorithms: boosting, deep learning,…

Thanks to R. Caruana

Bayesian Optimization

1. Initial sample

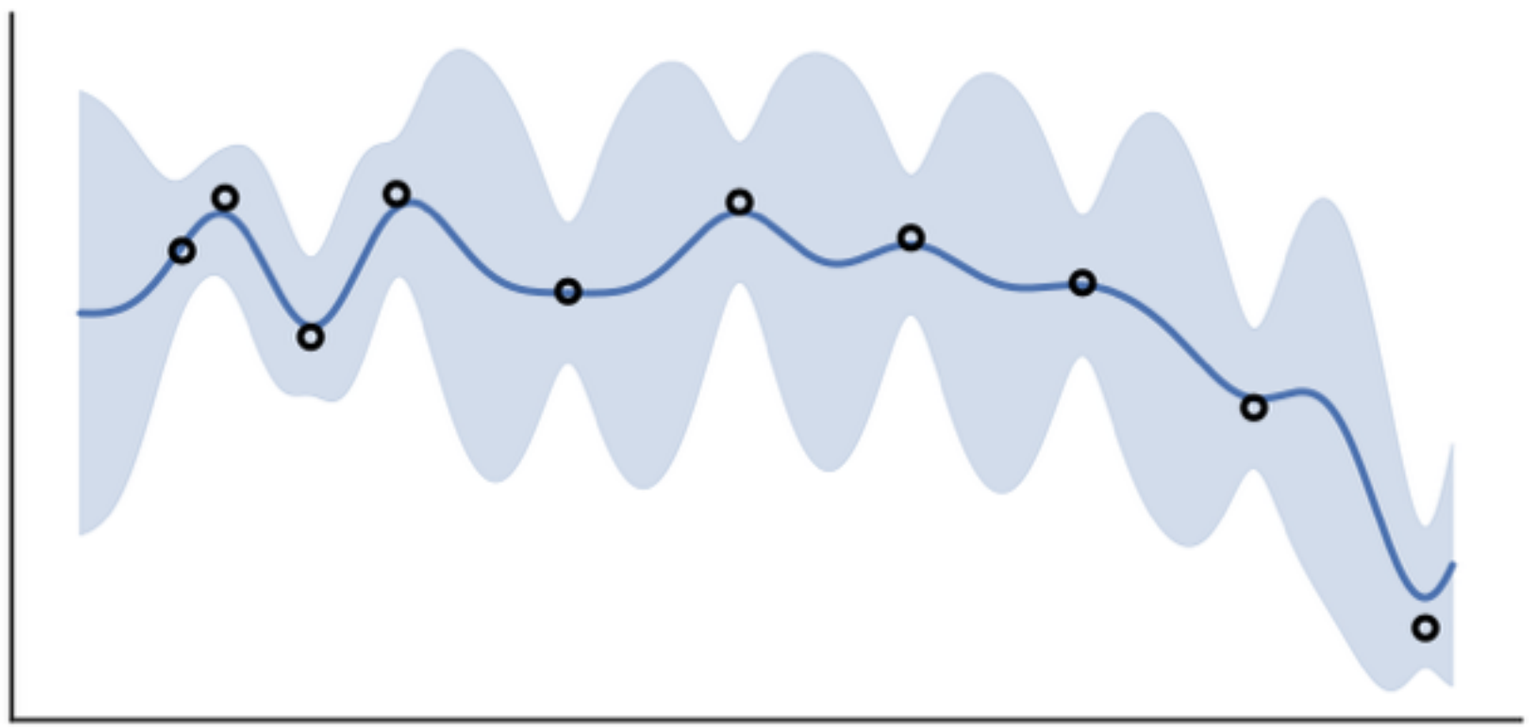
Thanks to Matthew Hoffmann

Bayesian Optimization

1. Initial sample

Thanks to Matthew Hoffmann

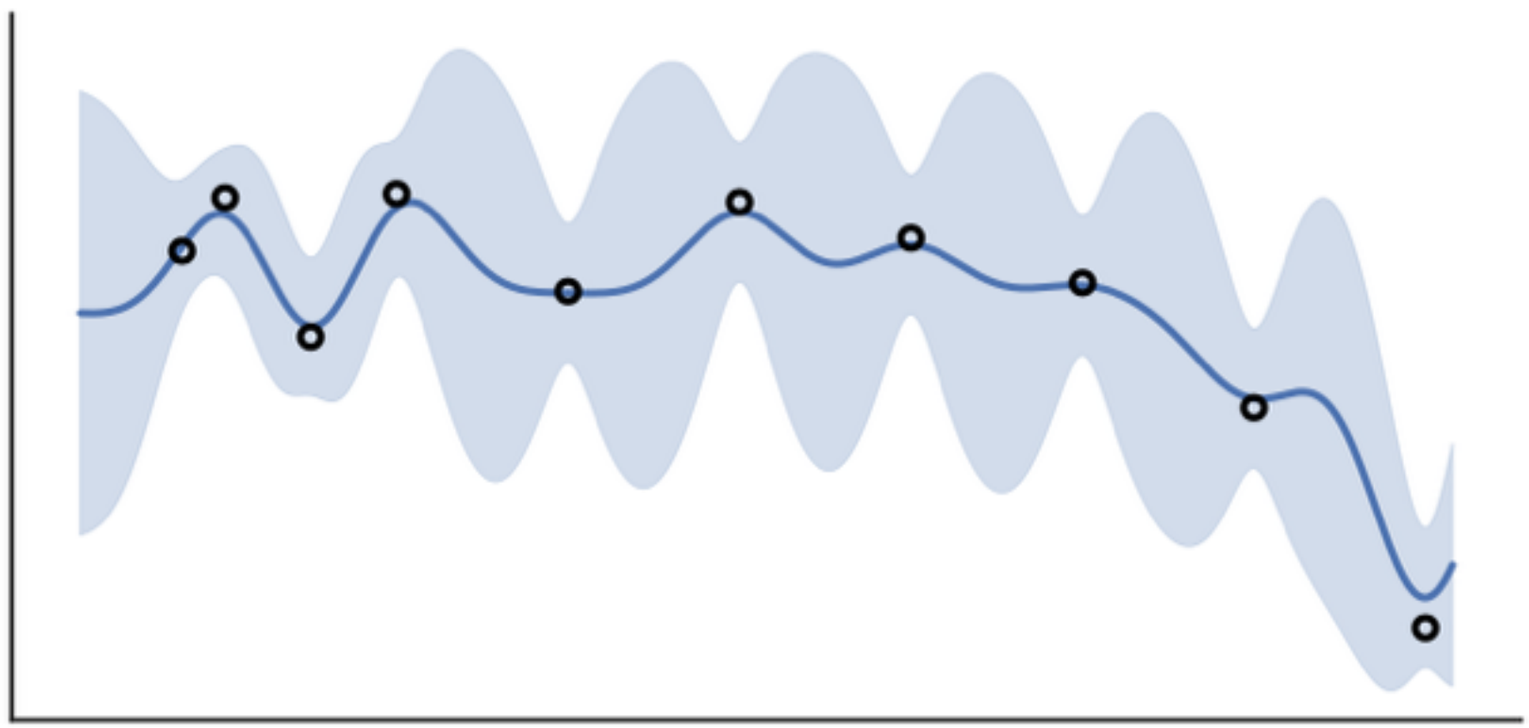
Bayesian Optimization

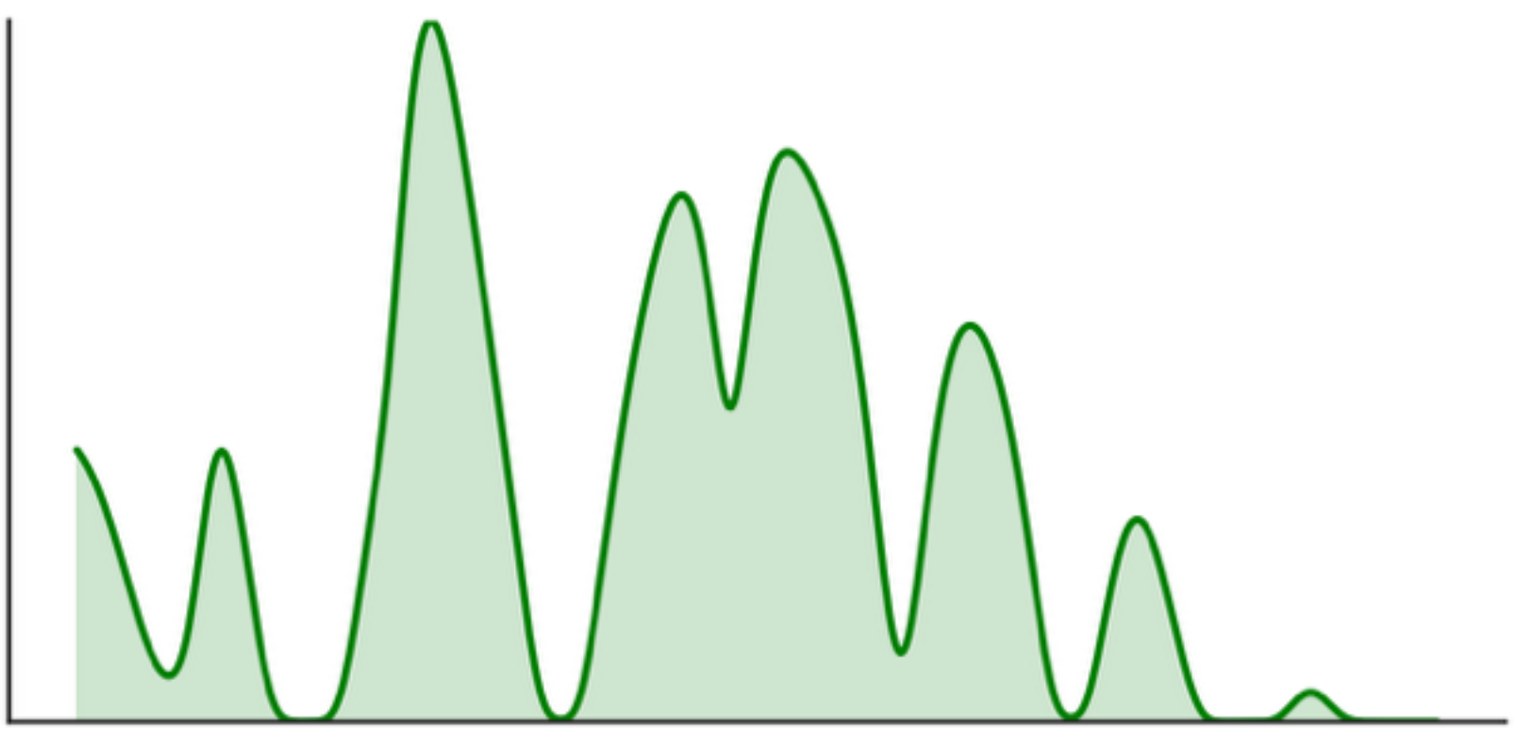


1. Initial sample
2. Posterior model

Thanks to Matthew Hoffmann

Bayesian Optimization

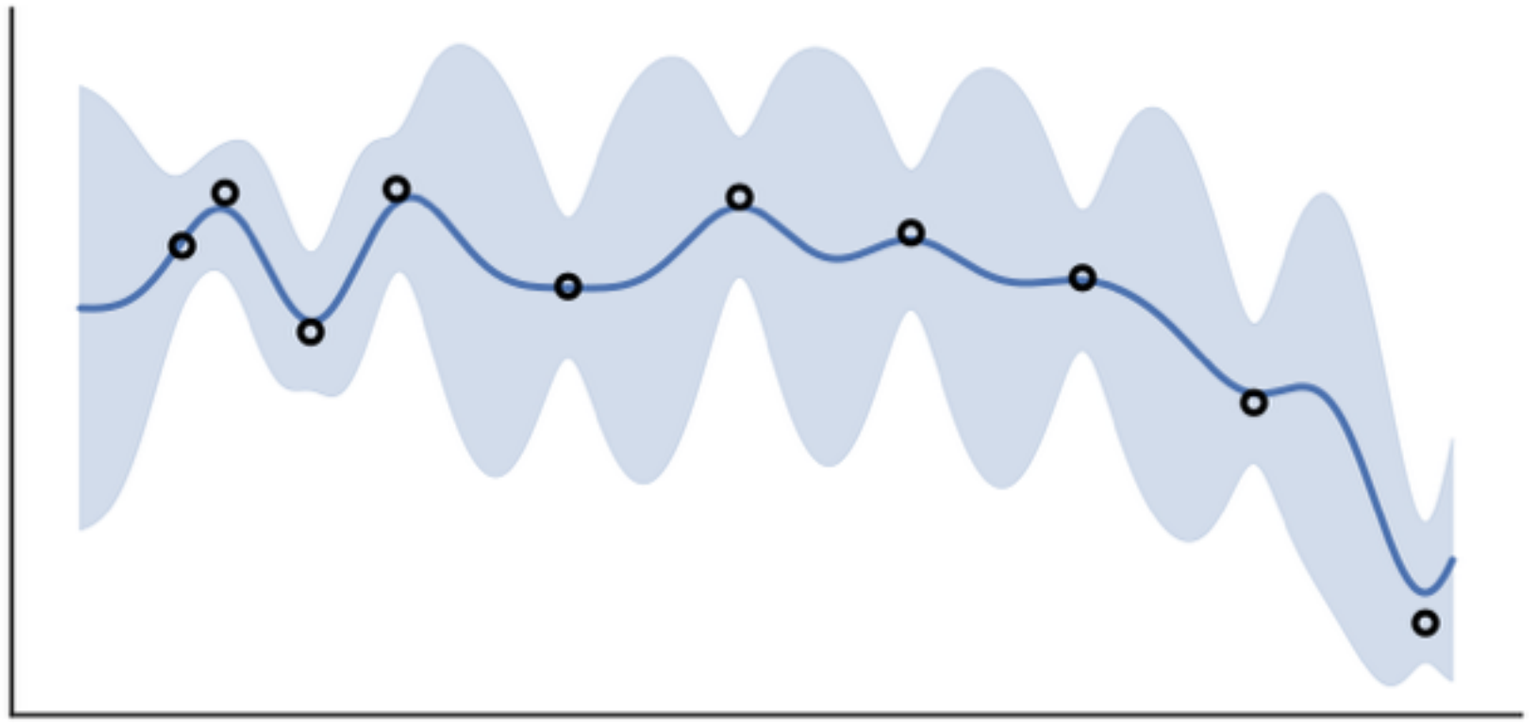


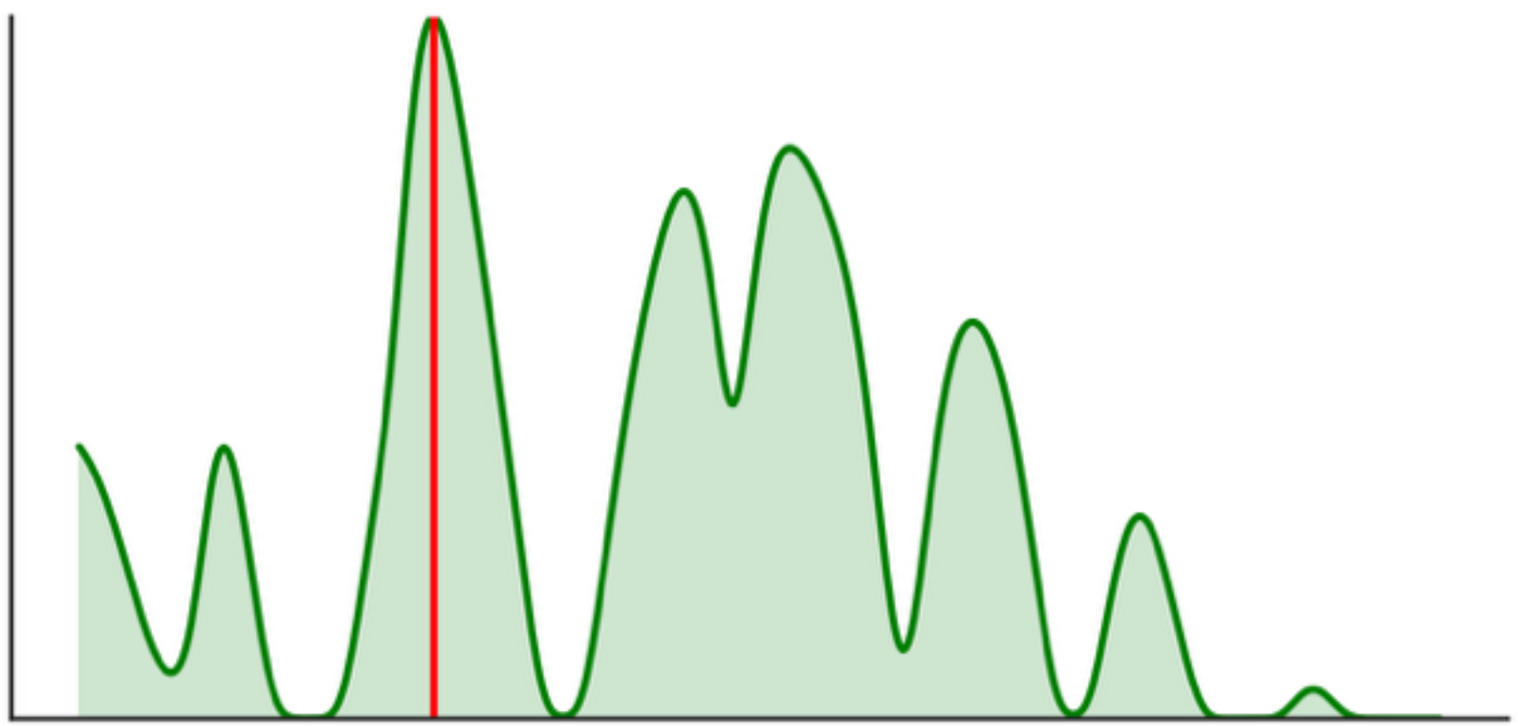


1. Initial sample
2. Posterior model
3. Exploration strategy (acquisition function)

Thanks to Matthew Hoffmann

Bayesian Optimization

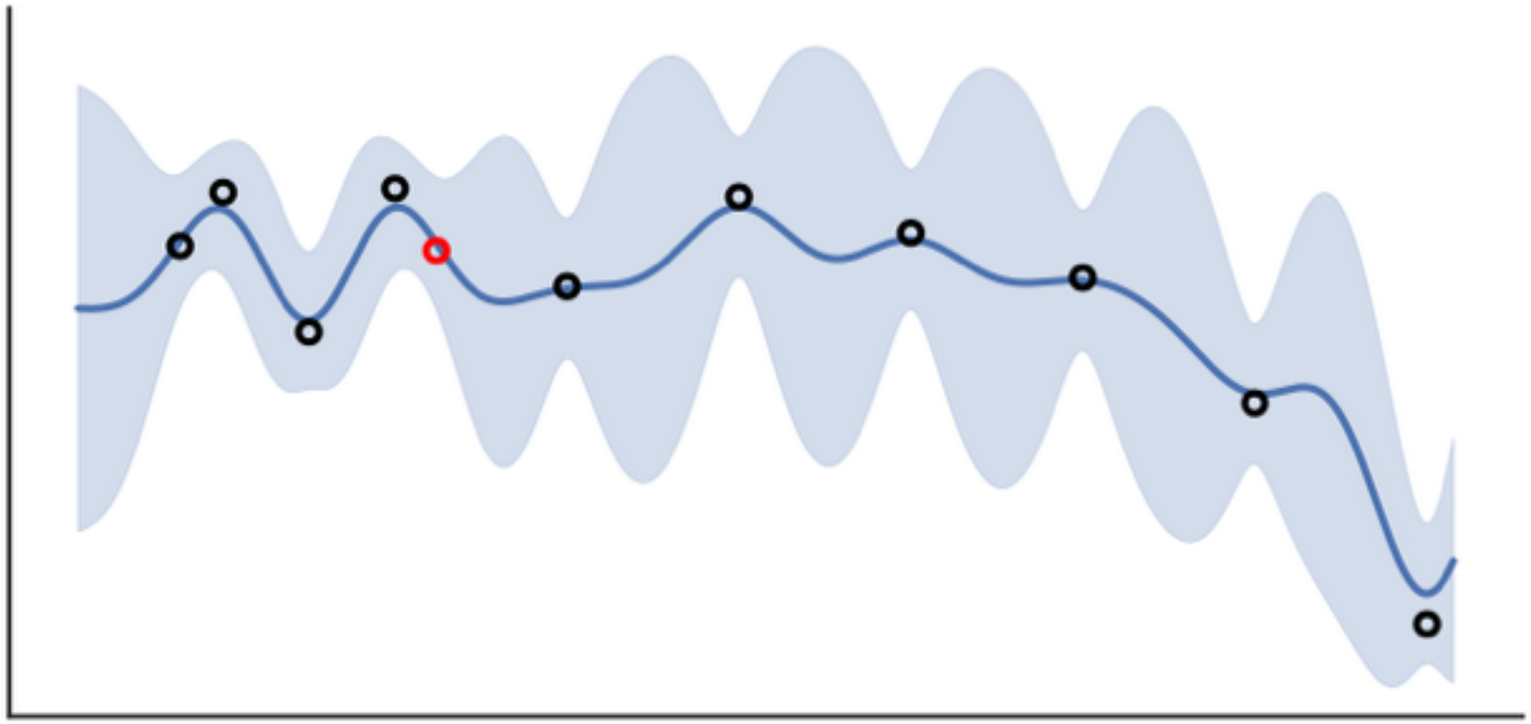


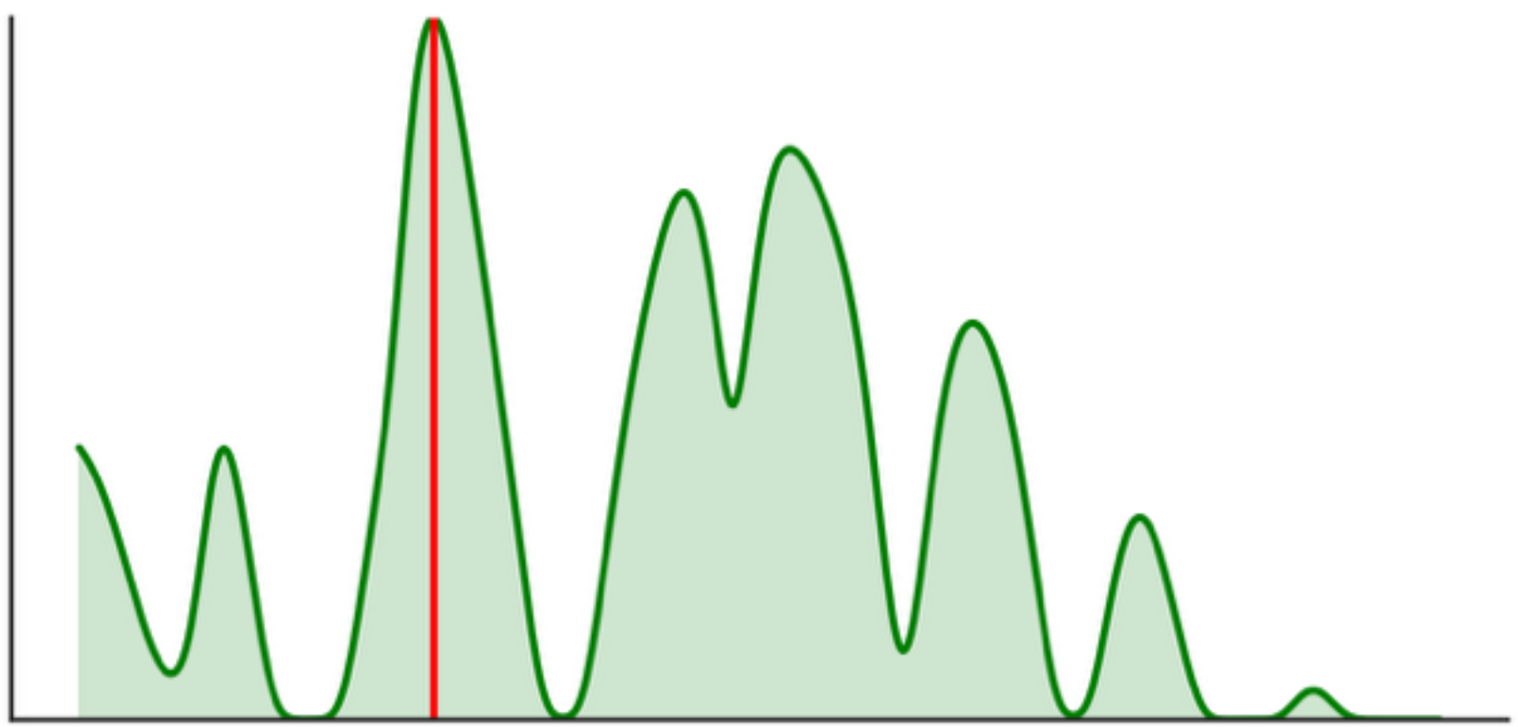


1. Initial sample
2. Posterior model
3. Exploration strategy (acquisition function)
4. Optimize it

Thanks to Matthew Hoffmann

Bayesian Optimization

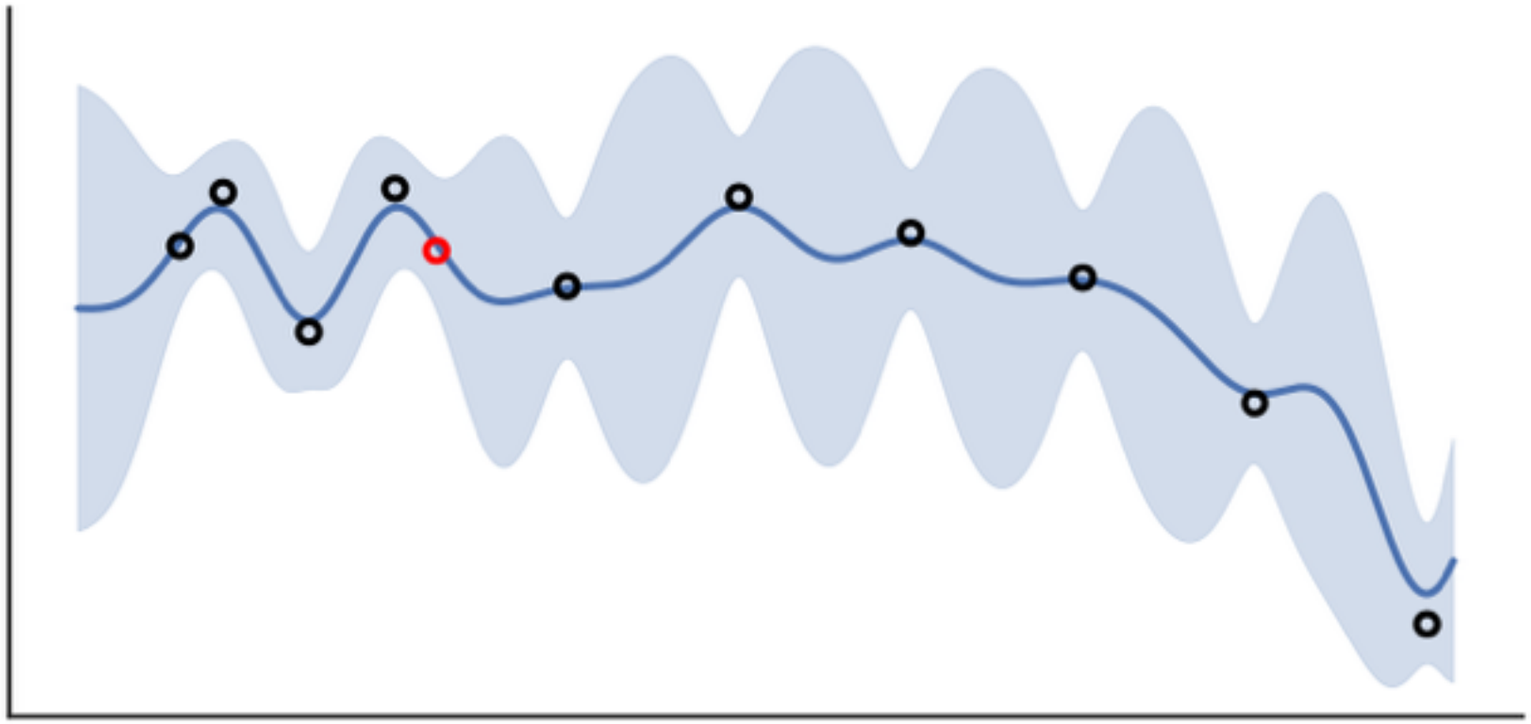


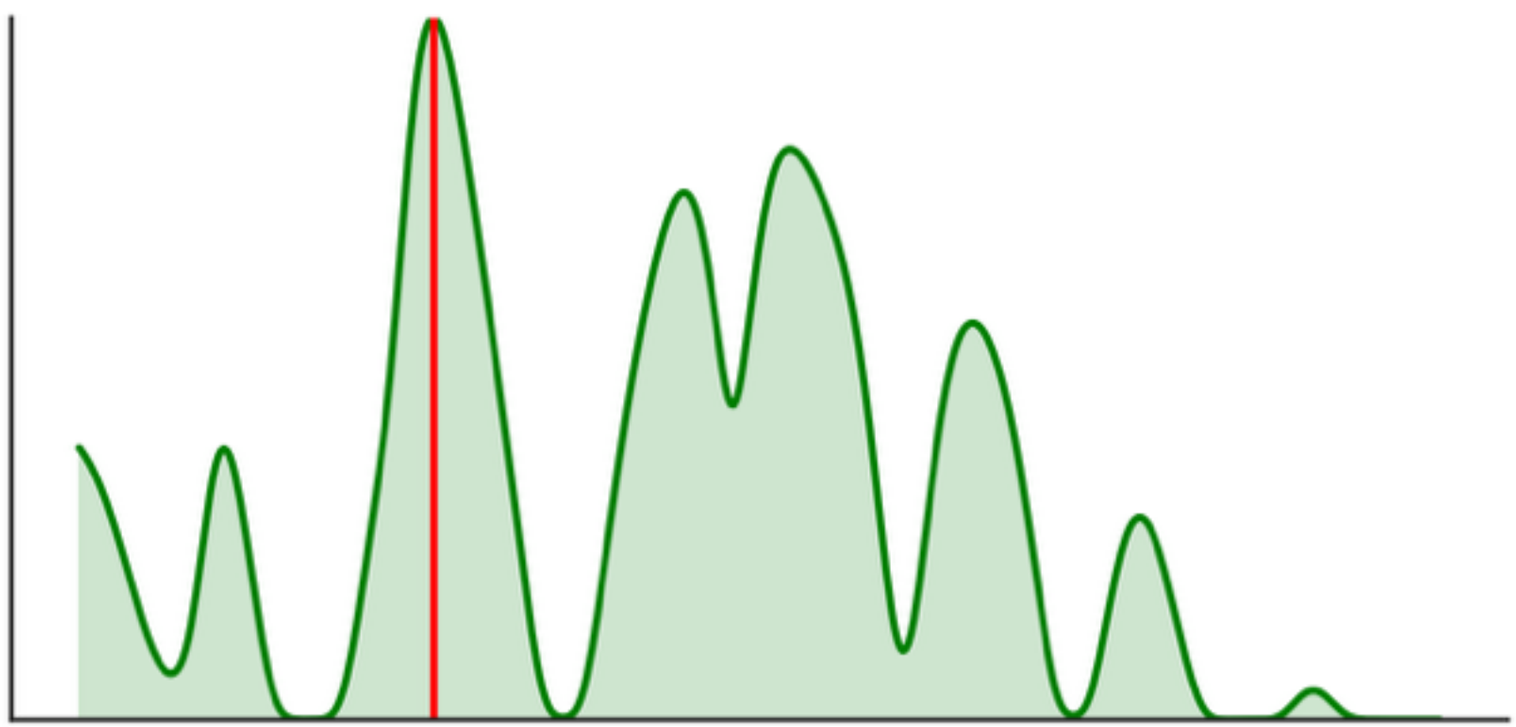


1. Initial sample
2. Posterior model
3. Exploration strategy (acquisition function)
4. Optimize it
5. Sample new data, update model

Thanks to Matthew Hoffmann

Bayesian Optimization





1. Initial sample
2. Posterior model
3. Exploration strategy (acquisition function)
4. Optimize it
5. Sample new data, update model
6. Repeat

Thanks to Matthew Hoffmann

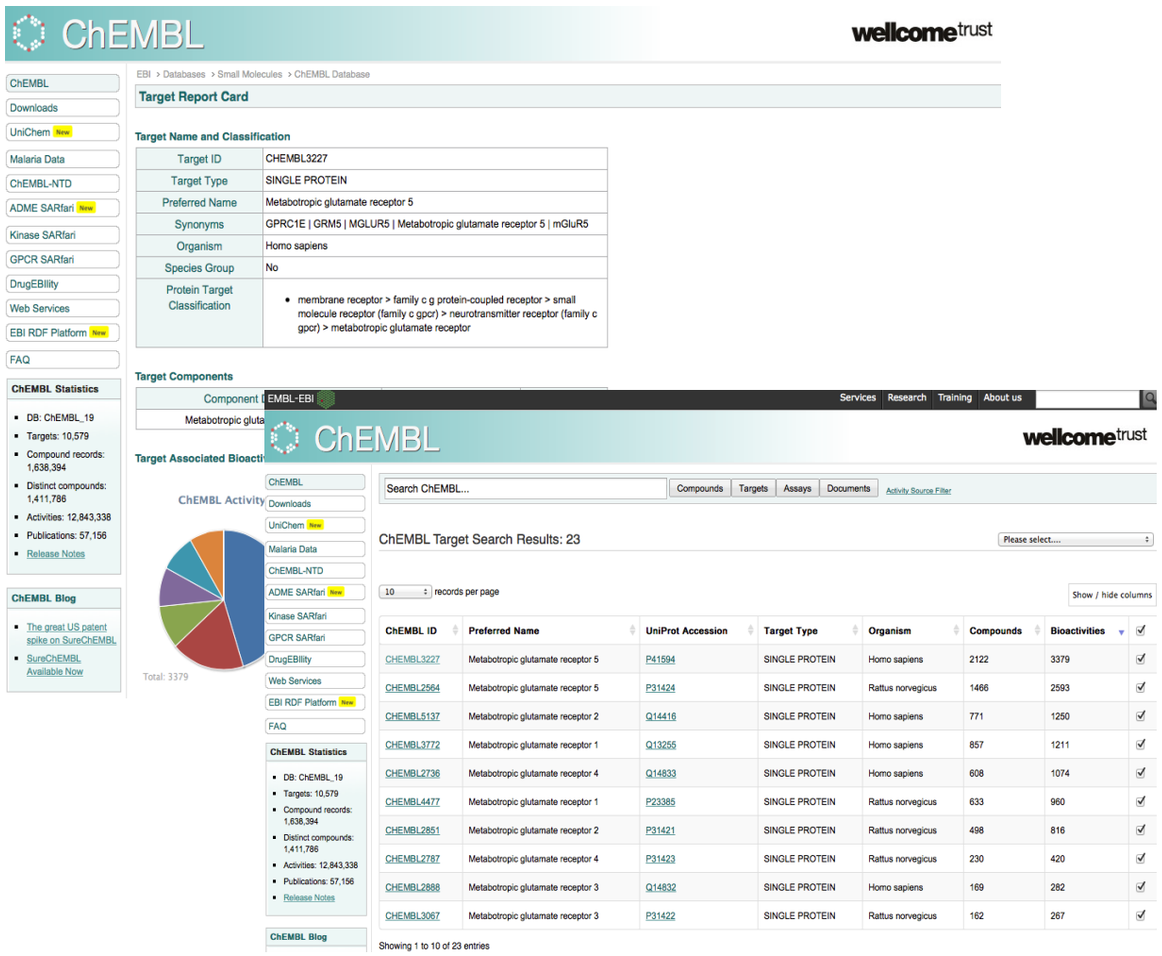
MACHINE LEARNING PIPELINE

Data Collection
Data Cleaning
Data Coding
Metric Selection
Algorithm Selection
Parameter Optimization
Post-processing
Online evaluation

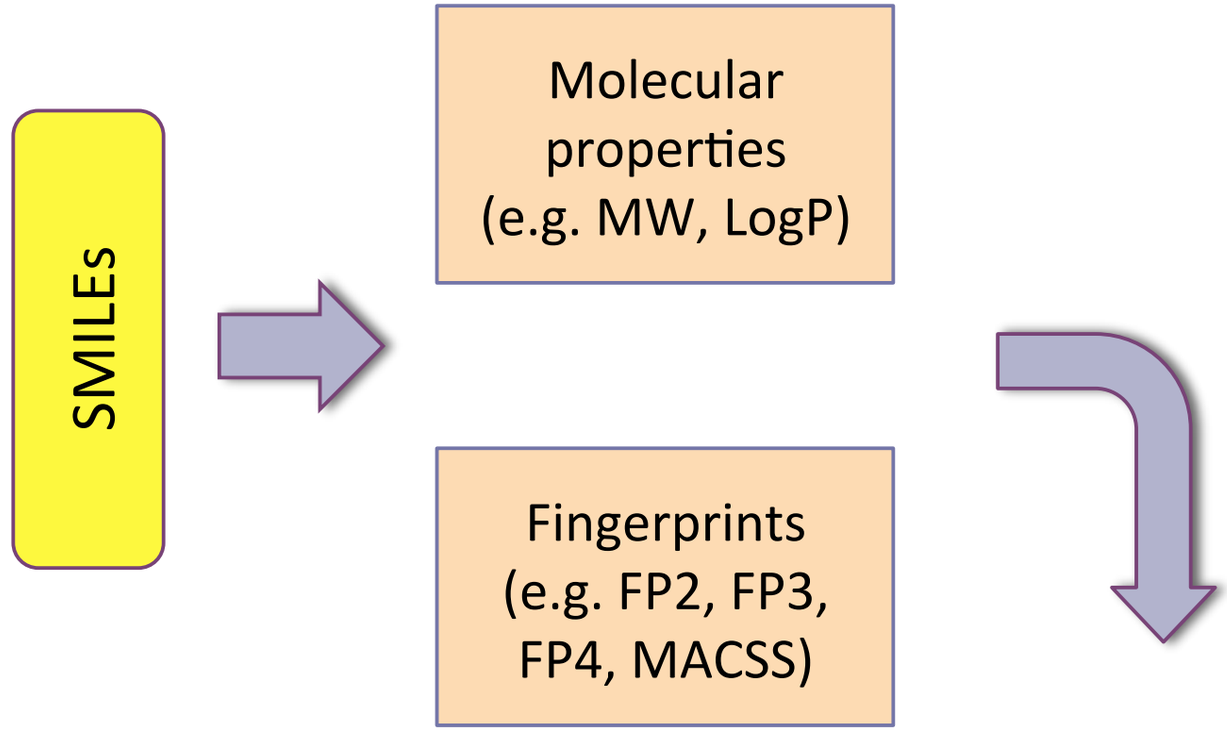
Meta-learning: Learn link between data properties & algorithm performance
Active testing: Sequentially predict algorithm, learn from evaluation score
Recommender systems: Evaluations are 'ratings' of datasets by algorithms

Data properties are crucial, depend on data domain (streams, graphs,…)
Best combined with parameter optimisation

OpenML in drug discovery



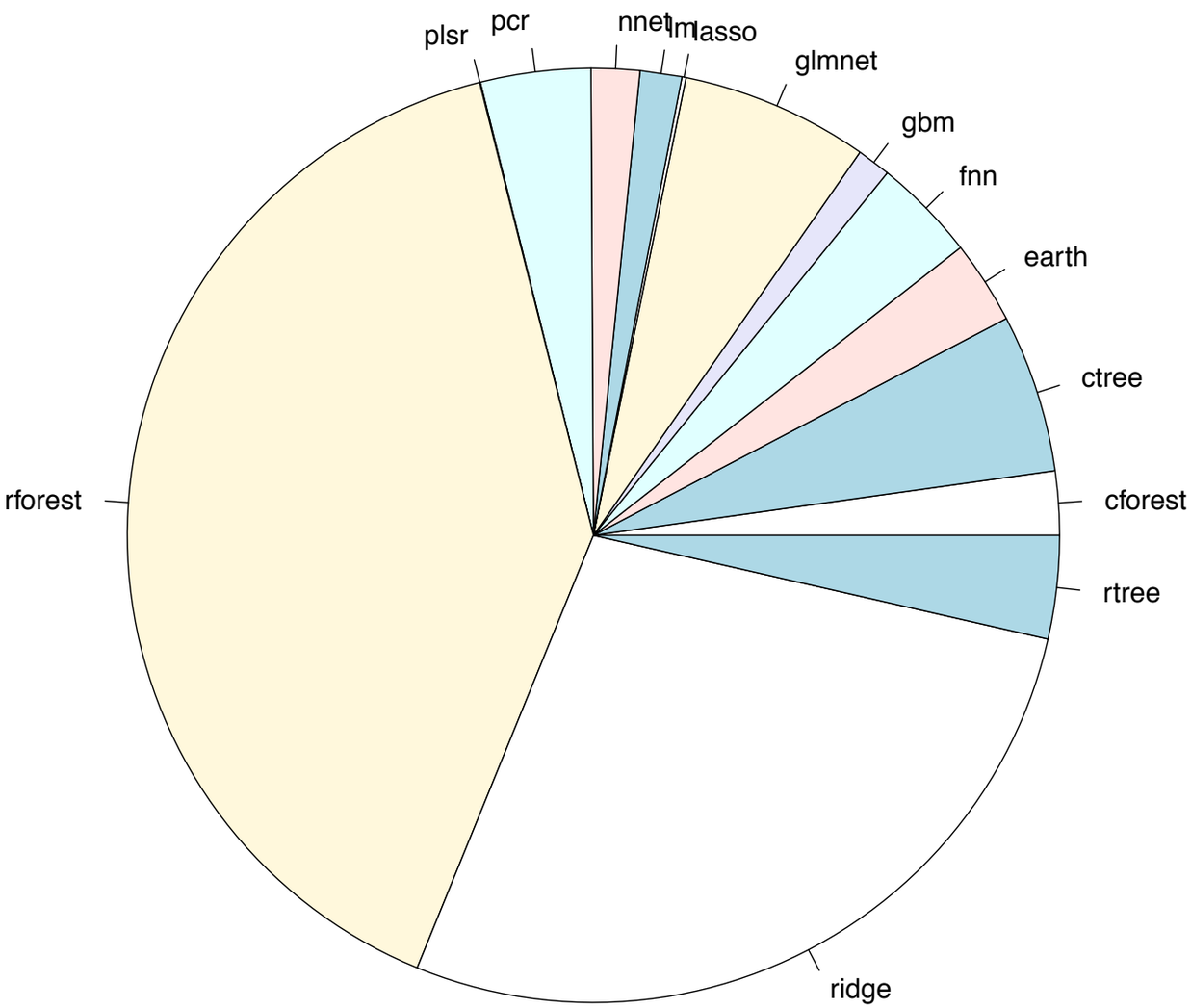
ChEMBL database

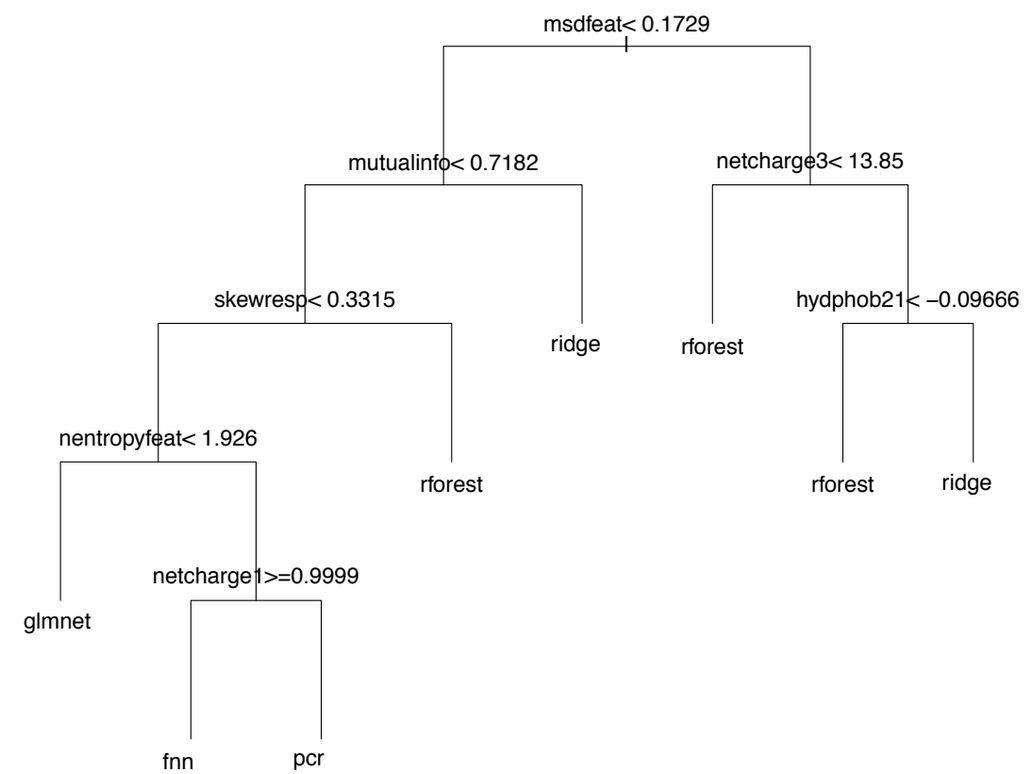


MW	LogP	TPSA	b1	b2	b3	b4	b5	b6	b7	b8	b9	
377.435	3.883	77.85	1	1	0	0	0	0	0	0	0	
341.361	3.411	74.73	1	1	0	1	0	0	0	0	0	...
197.188	-2.089	103.78	1	1	0	1	0	0	0	1	0	
346.813	4.705	50.70	1	0	0	1	0	0	0	0	0	
							⋮					

10.000+ regression datasets

OpenML in drug discovery





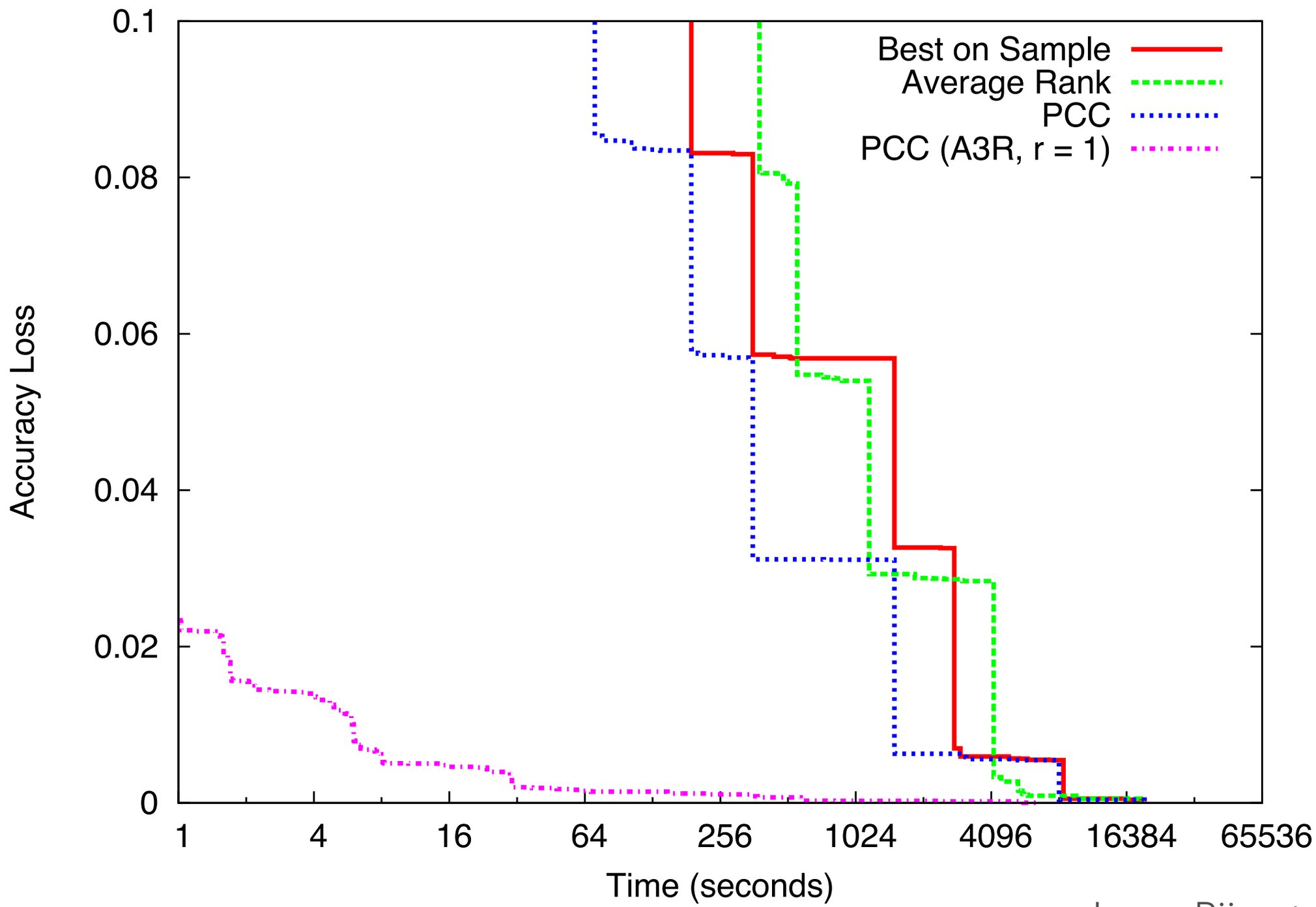
ECFP4_1024	FCFP4_1024	**ECFP6_1024**	FCFP6_1024
0.697	0.427	**0.725**	0.627

Predict best algorithm with meta-models

I. Olier et al. MetaSel@ECMLPKDD 2015

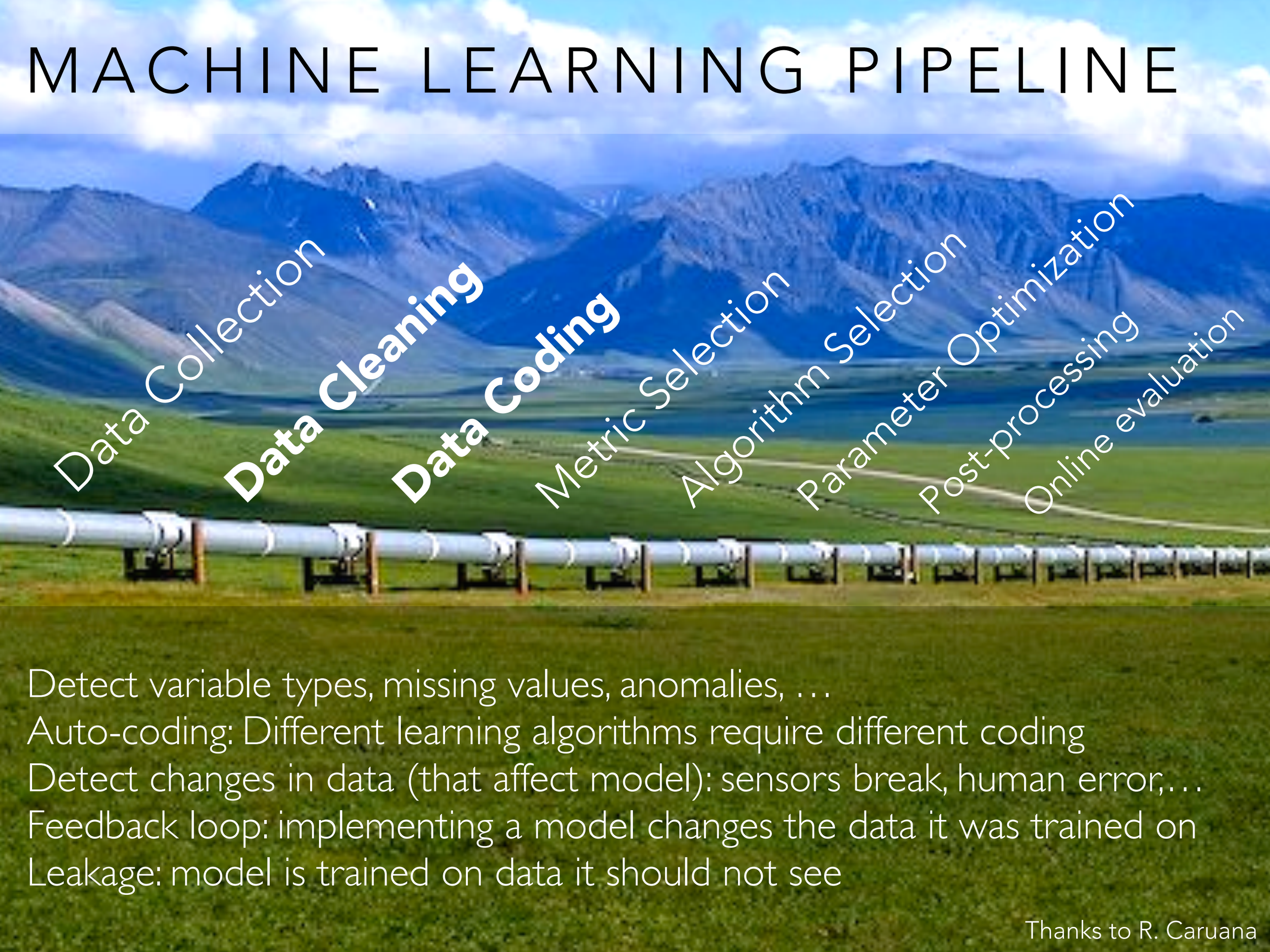
Fast algorithm selection

By evaluating algorithm on smaller data samples, multi-objective evaluation



J. van Rijn et al. IDA 2015

MACHINE LEARNING PIPELINE



Detect variable types, missing values, anomalies, …
Auto-coding: Different learning algorithms require different coding
Detect changes in data (that affect model): sensors break, human error,…
Feedback loop: implementing a model changes the data it was trained on
Leakage: model is trained on data it should not see

Thanks to R. Caruana

Towards an AI for data-driven research

Symbiotic AI: learns from thousands of human data scientists how to analyse data. Learns which workflows work well on which data.

Collaborates with scientists: take care of tedious, error-prone tasks: clean data, find similar data, try state-of-the-art, algorithm selection, configuration,…

Couple human expertise and machine learning: offer the right information at the right time to the right person, so that she can make informed decisions based on clear evidence.

Towards an AI for data-driven research

AI (novice) - what?
Upload dataset, give 1 day to return best possible model

Man-Computer Symbiosis - why?
Answer specific questions by scientist.
Combine machine learning with talents of human mind (hunches, expertise).
Explains *why* recommendation made.

Expert level (expert) - how?
Expert asks system *how* it did something, system explains all details, show code, data

Towards an AI for data-driven research

Look at my new data set :)

I analysed it, here's a full report.

Can you remove the outliers?

I removed them, does this look ok?

Yes, now I want to predict X

OK, let me build you some optimized classification models.

Are there similar data sets?

Here's a ranked list, and some papers.

What's the state-of-the-art?

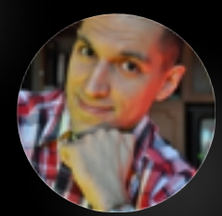
I'm running all state-of-the-art techniques, I'll send a report soon.

What's the difference with algo A and B?

Here's a comparison. B looks most promising on high-dimensional data

THANK YOU

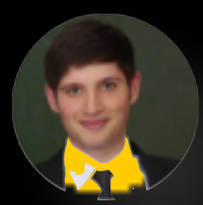
#OpenML



Nenad Tomašev

Luis Torgo

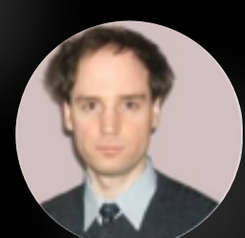
Jan van Rijn



Giuseppe Casalicchio

Joaquin Vanschoren

Michel Lang



Bernd Bischl

Matthias Feurer

You?