

Democratizing machine learning research with OpenML

Joaquin Vanschoren, Eindhoven University of Technology

Machine Learning is labor-intensive

Infinite range of possibilities, tacit experience

😓 Machine Learning is labor-intensive

Infinite range of possibilities, tacit experience

collection,
cleaning,
preprocessing,
featurization,
selection,...

Data



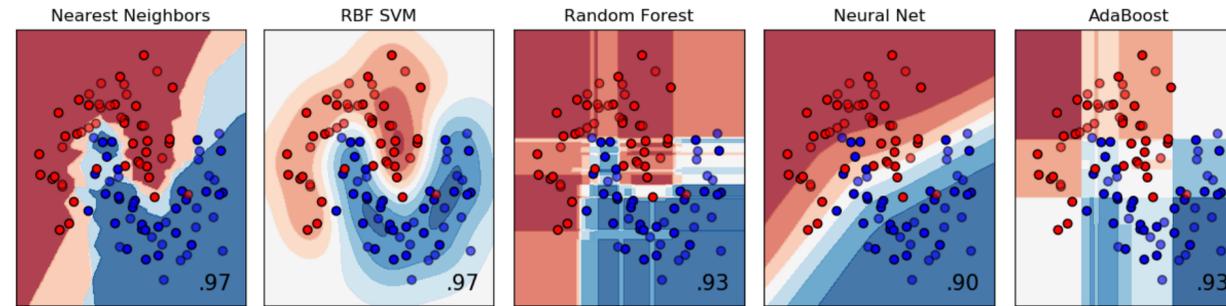
🥵 Machine Learning is labor-intensive

Infinite range of possibilities, tacit experience

Model selection

collection,
cleaning,
preprocessing,
featurization,
selection,...

Data



🥵 Machine Learning is labor-intensive

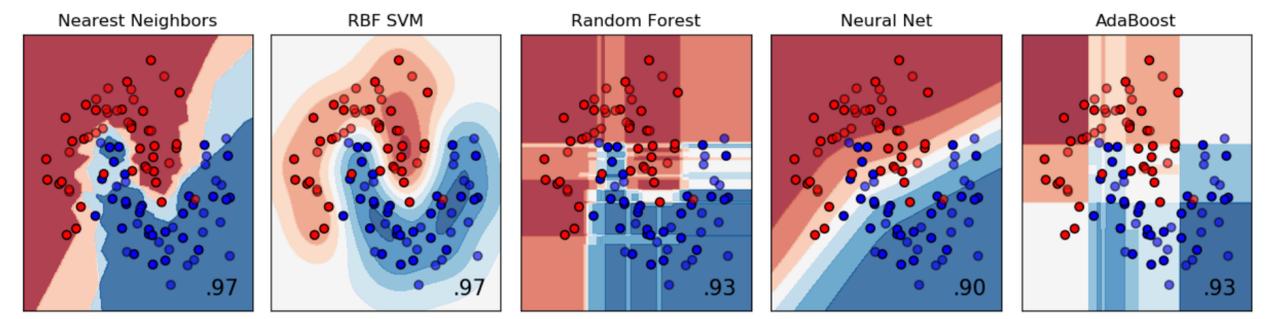
Infinite range of possibilities, tacit experience

collection,
cleaning,
preprocessing,
featurization,
selection,...

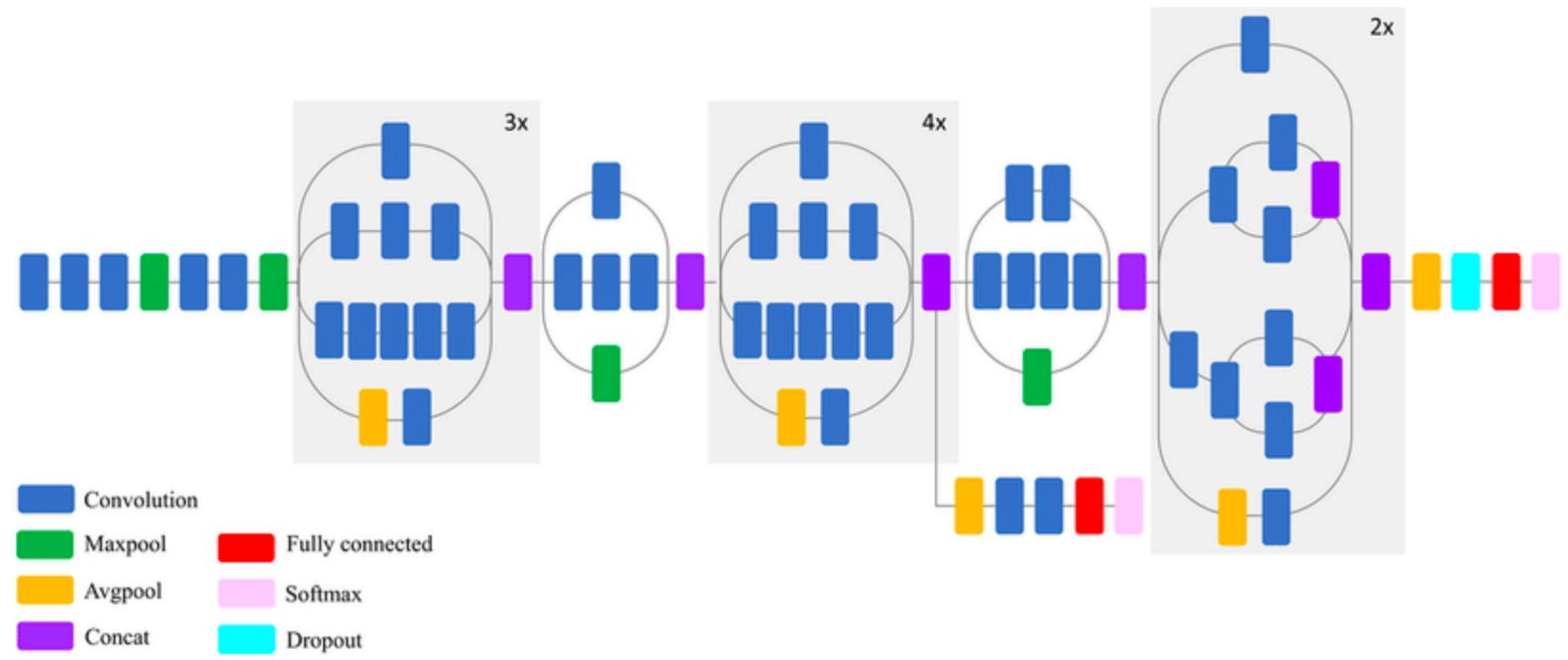
Data



Model selection



Neural architecture search



😓 Machine Learning is labor-intensive

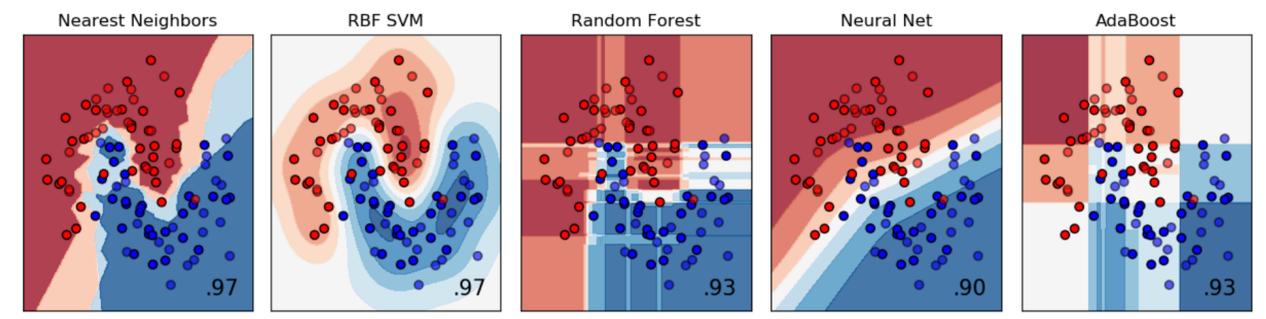
Infinite range of possibilities, tacit experience

collection,
cleaning,
preprocessing,
featurization,
selection,...

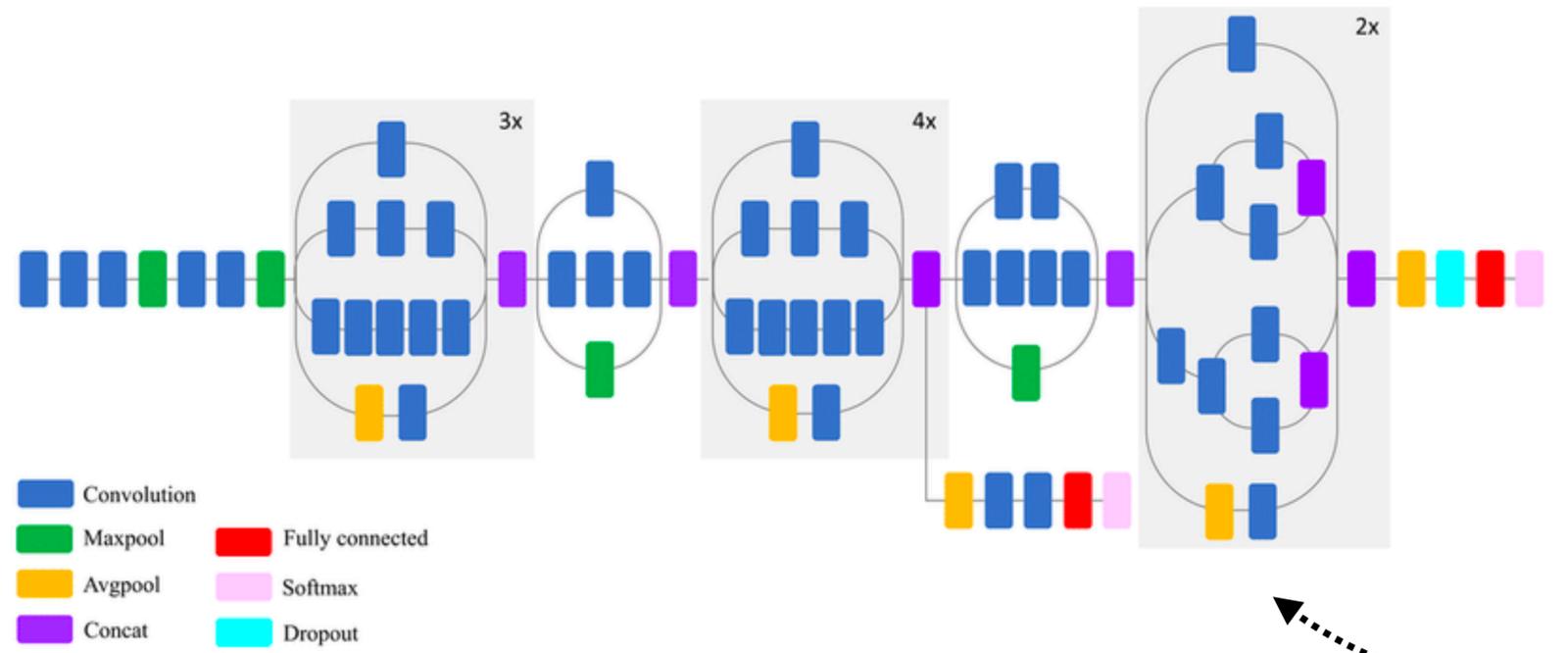
Data



Model selection



Neural architecture search



Transfer / continual / meta learning



(pre-trained) weights,
optimisers, ...



😓 Machine Learning is labor-intensive

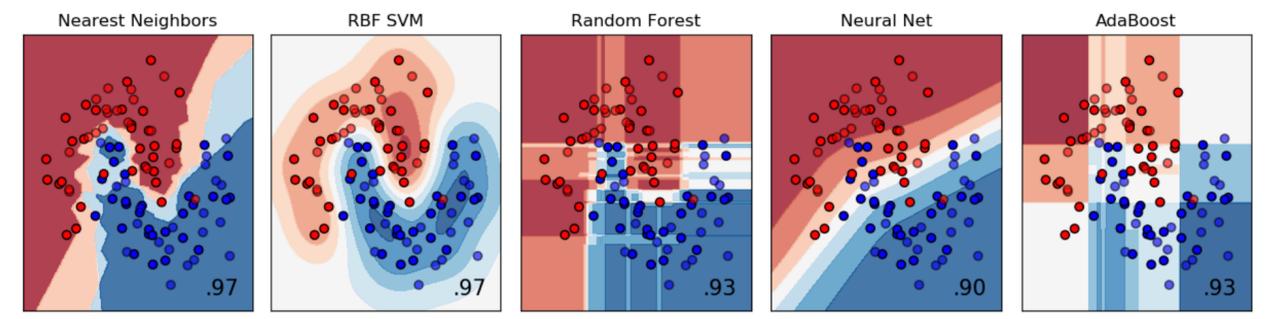
Infinite range of possibilities, tacit experience

collection,
cleaning,
preprocessing,
featurization,
selection,...

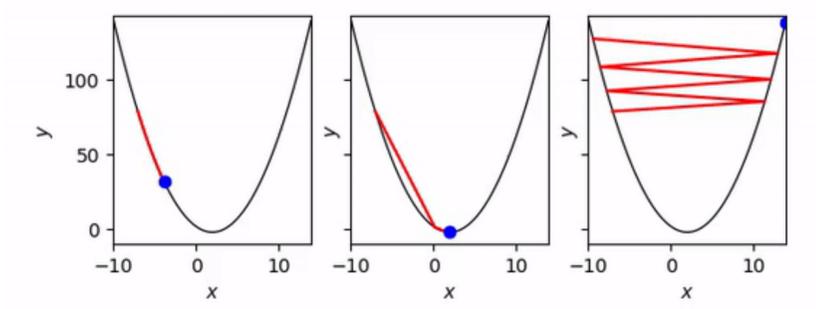
Data



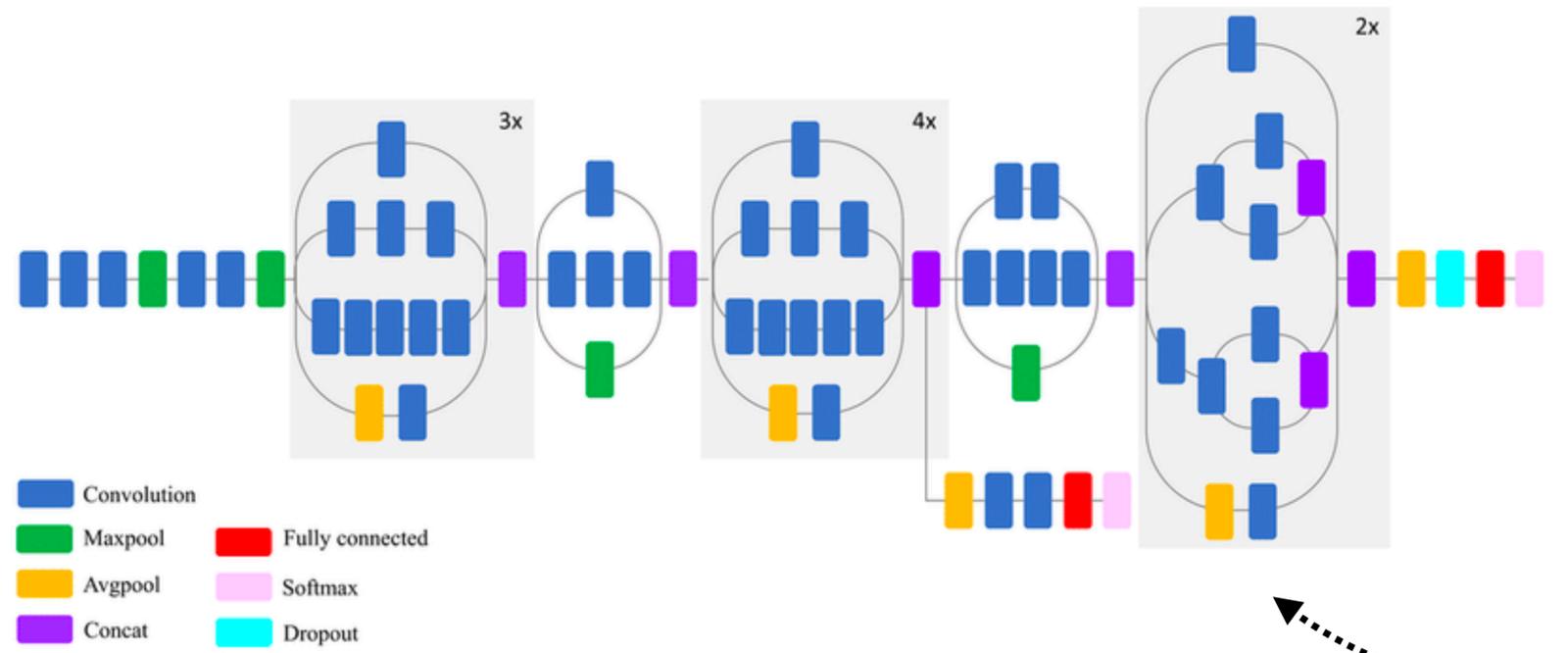
Model selection



Hyperparameter tuning



Neural architecture search



Transfer / continual / meta learning



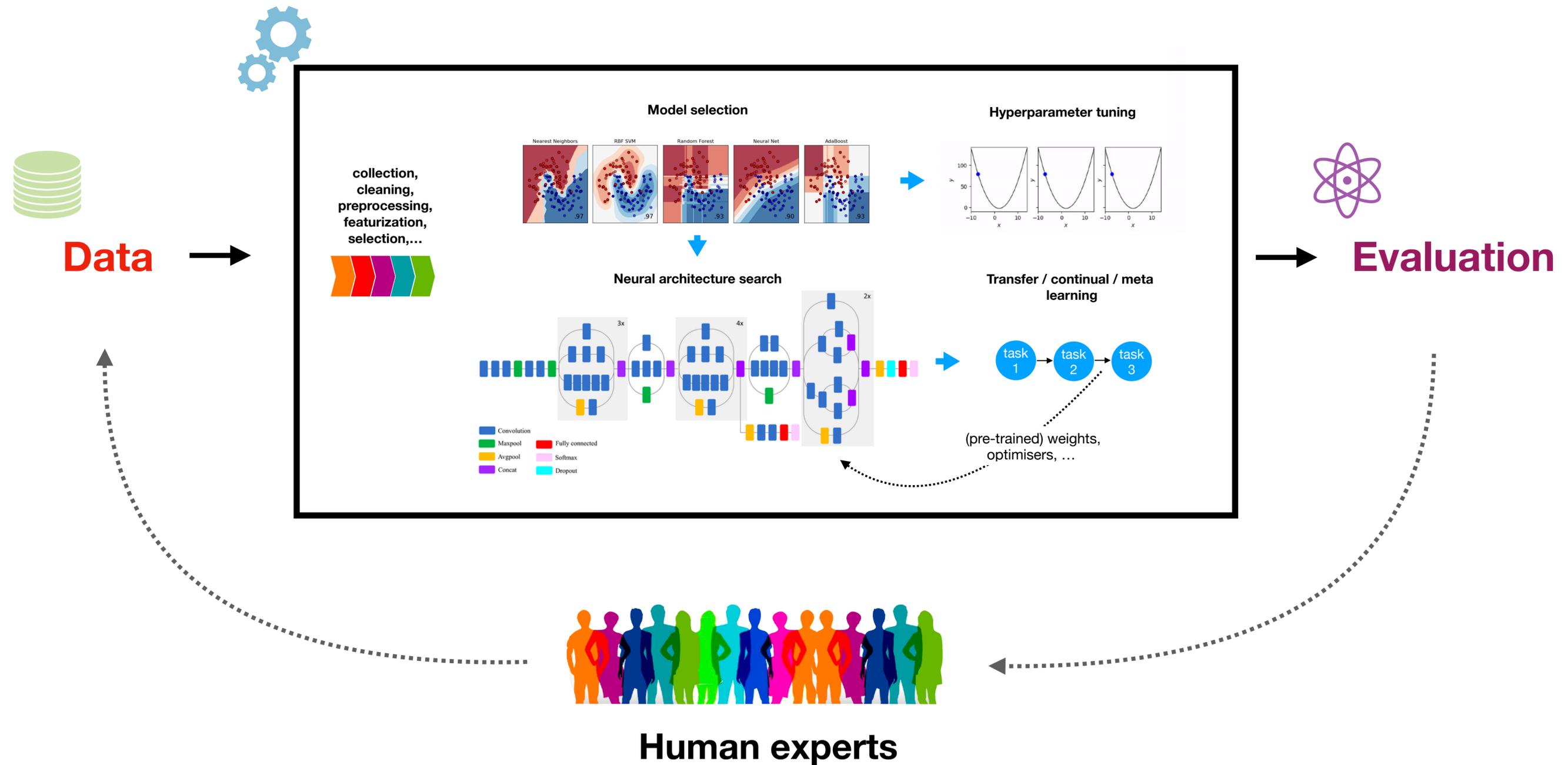
(pre-trained) weights,
optimisers, ...





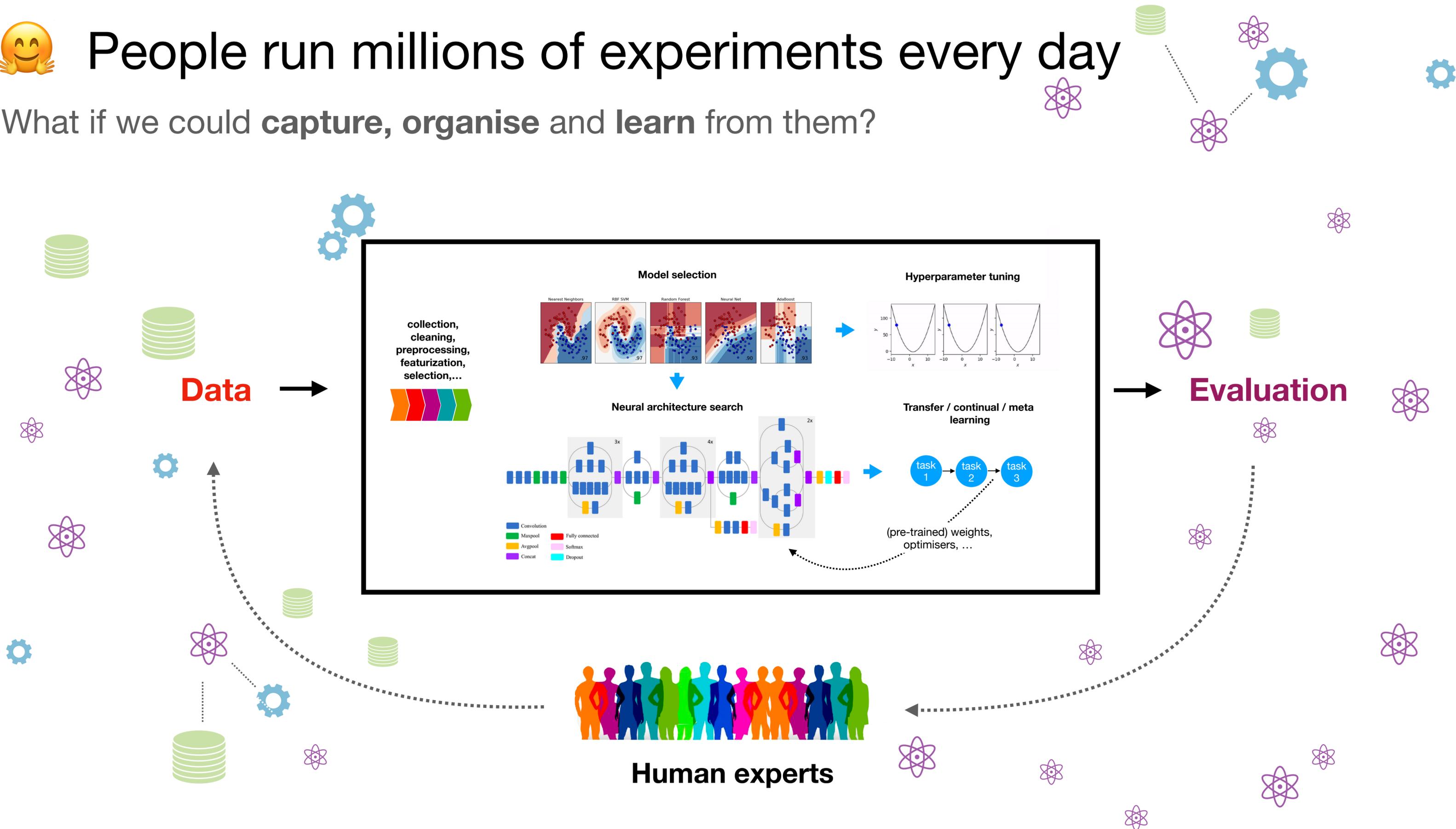
People run millions of experiments every day

What if we could **capture, organise and learn** from them?



👤 People run millions of experiments every day

What if we could **capture, organise and learn** from them?





*What if...
we could organize the
world's **machine learning**
information*

*and make it universally
accessible and useful?*

Democratizing data

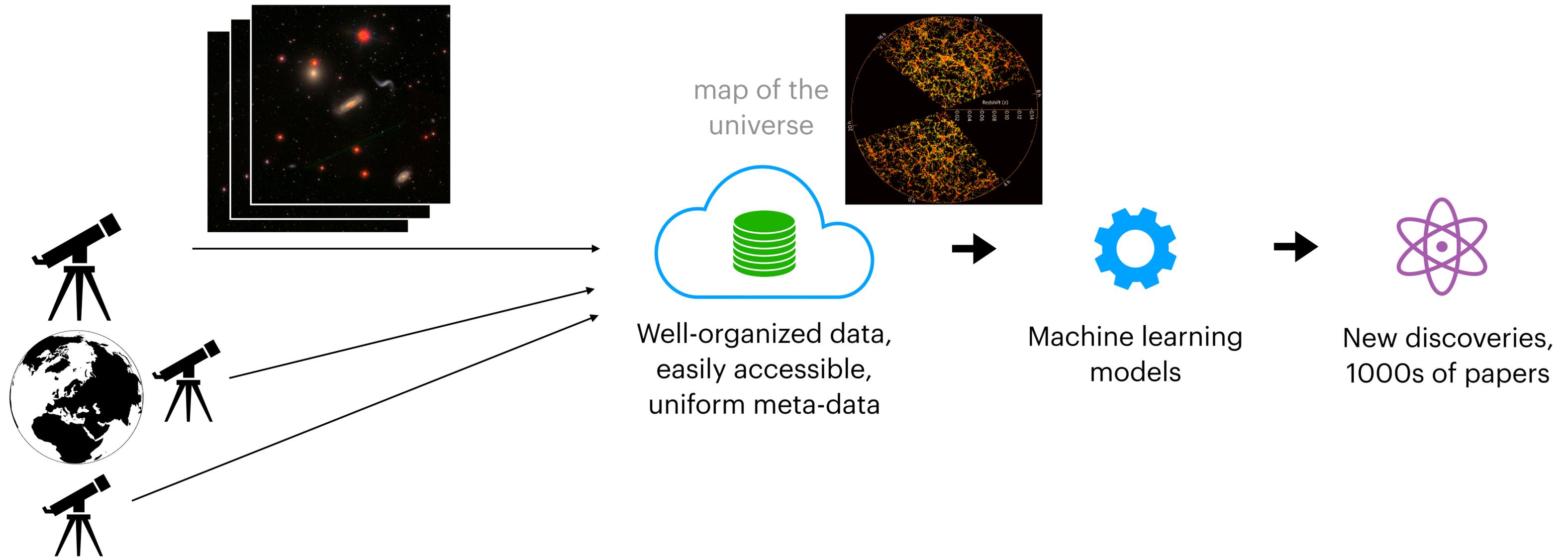
mapping the universe



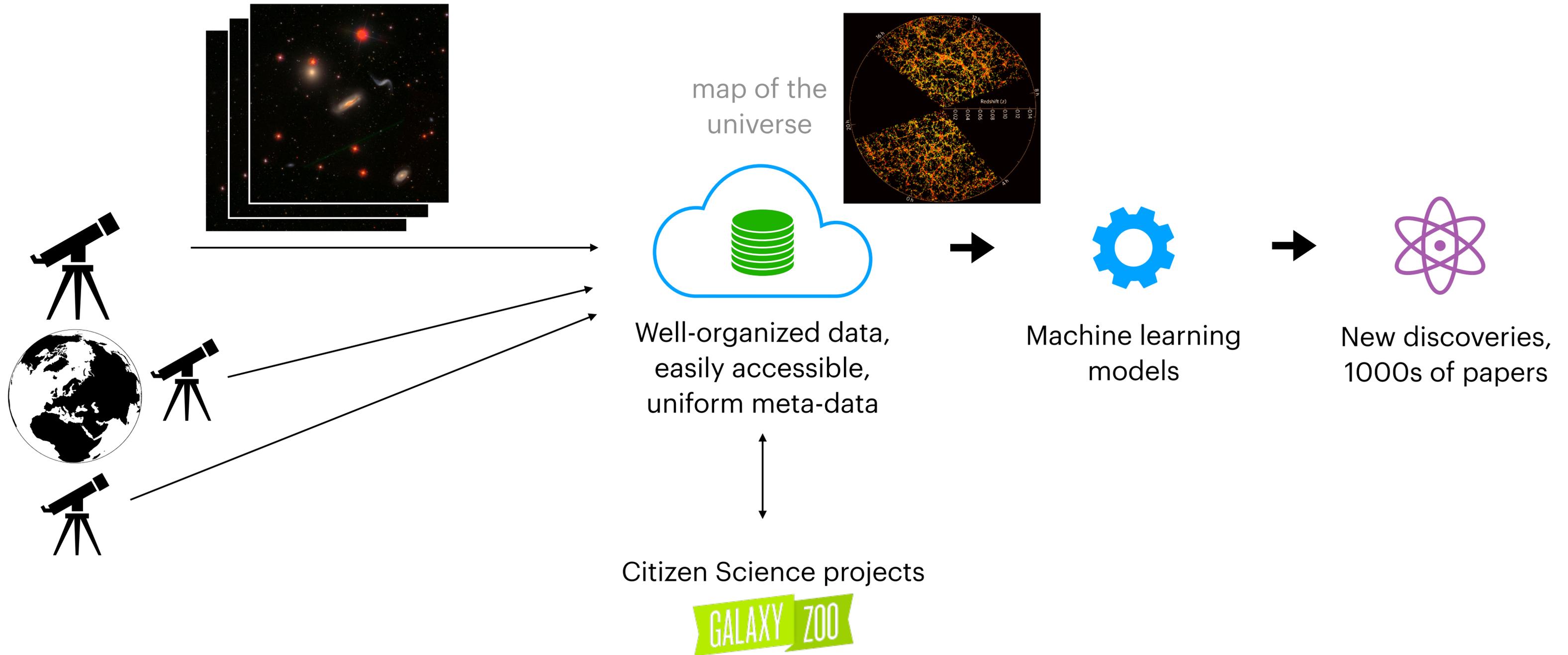
Democratizing data



Democratizing data

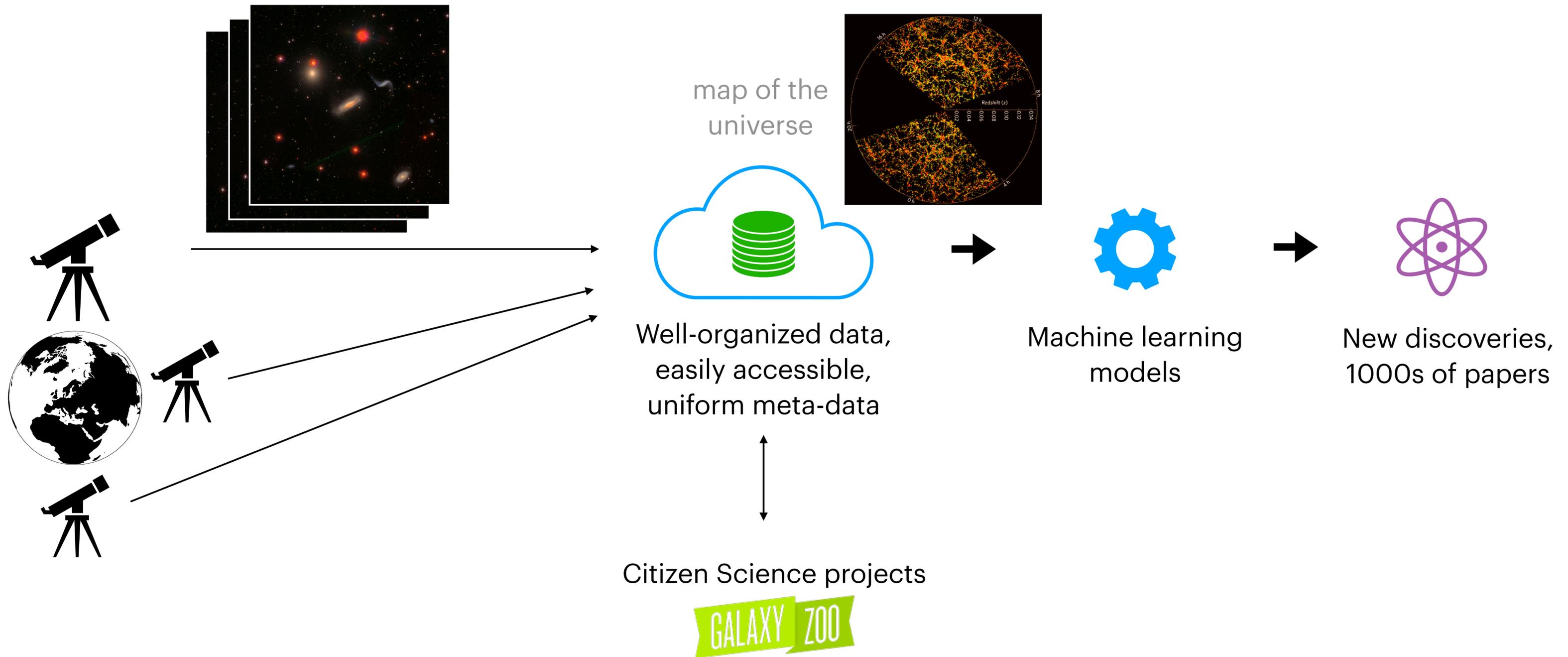


Democratizing data



Democratizing data

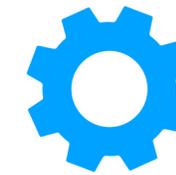
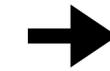
How can we generalize this idea?



Democratizing ML data

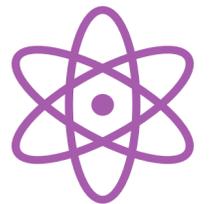
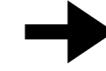


Well-organized data,
easily accessible,
uniform meta-data



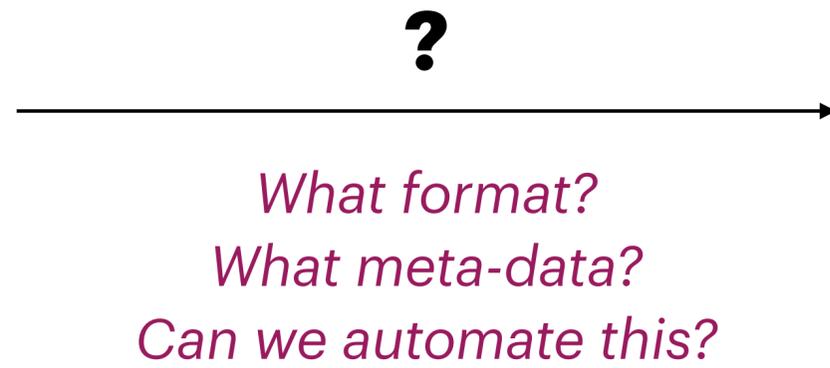
Machine learning
models

(Assuming we have these)

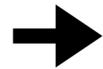


New discoveries,
1000s of papers

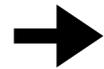
Democratizing ML data



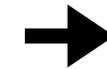
Democratizing ML data



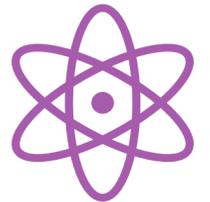
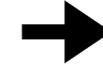
Most ML data is (at some point) represented as dataframe/matrix



Well-organized data, easily accessible, uniform meta-data



Machine learning models

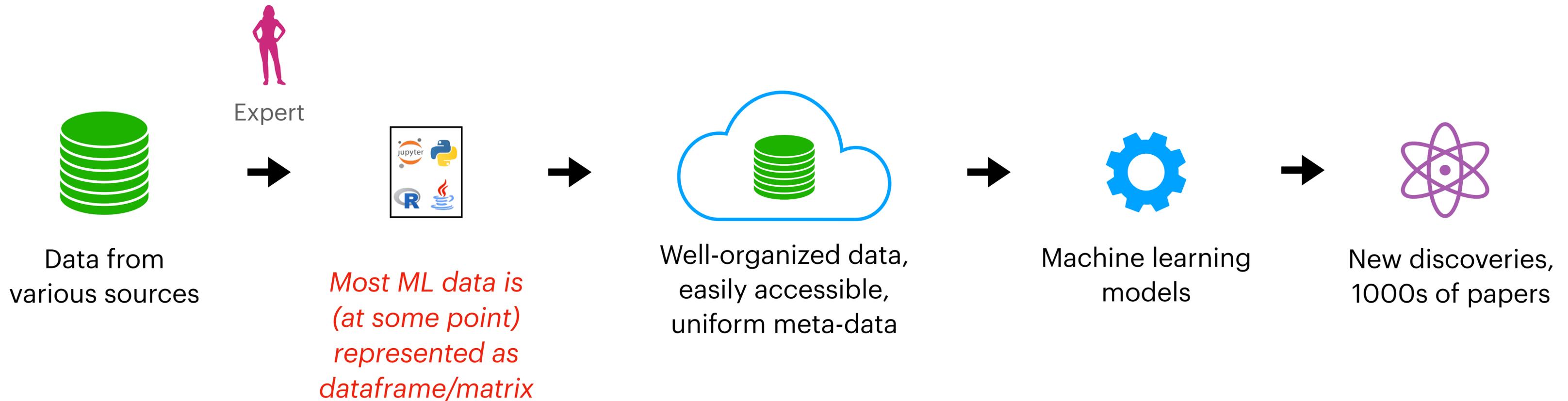


New discoveries, 1000s of papers

(Required anyway for many ML models)

(Can be a data loading script for large and remotely hosted data)

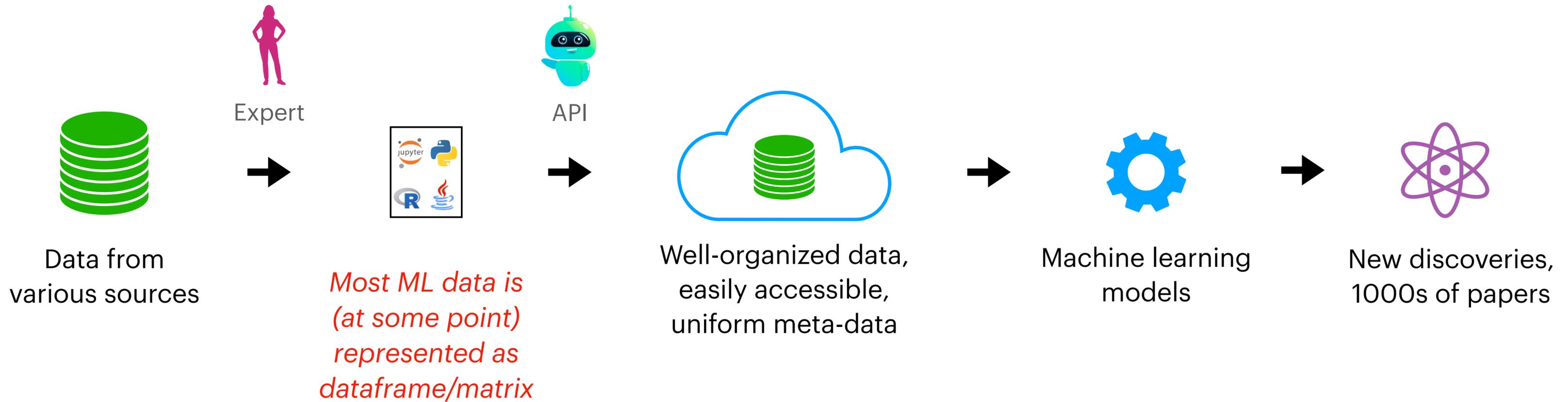
Democratizing ML data



(Required anyway for many ML models)

(Can be a data loading script for large and remotely hosted data)

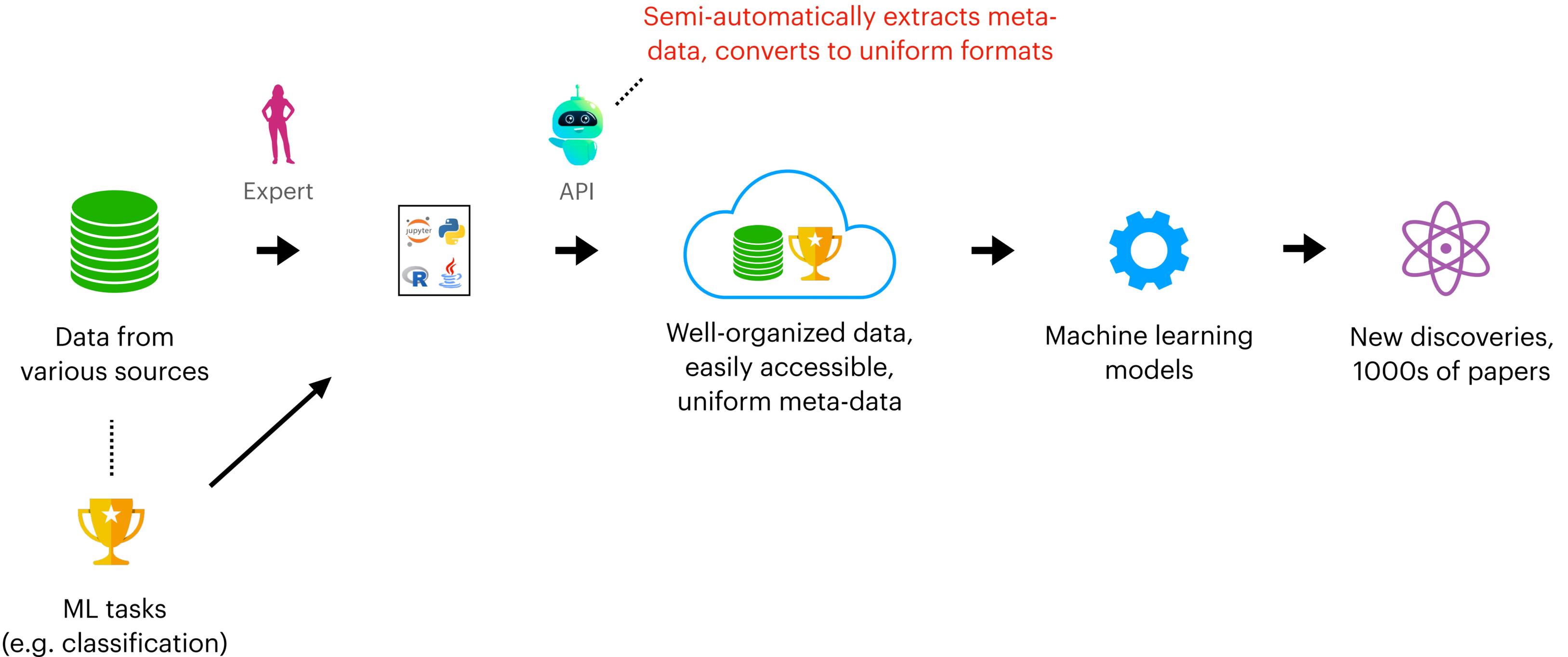
Democratizing ML data



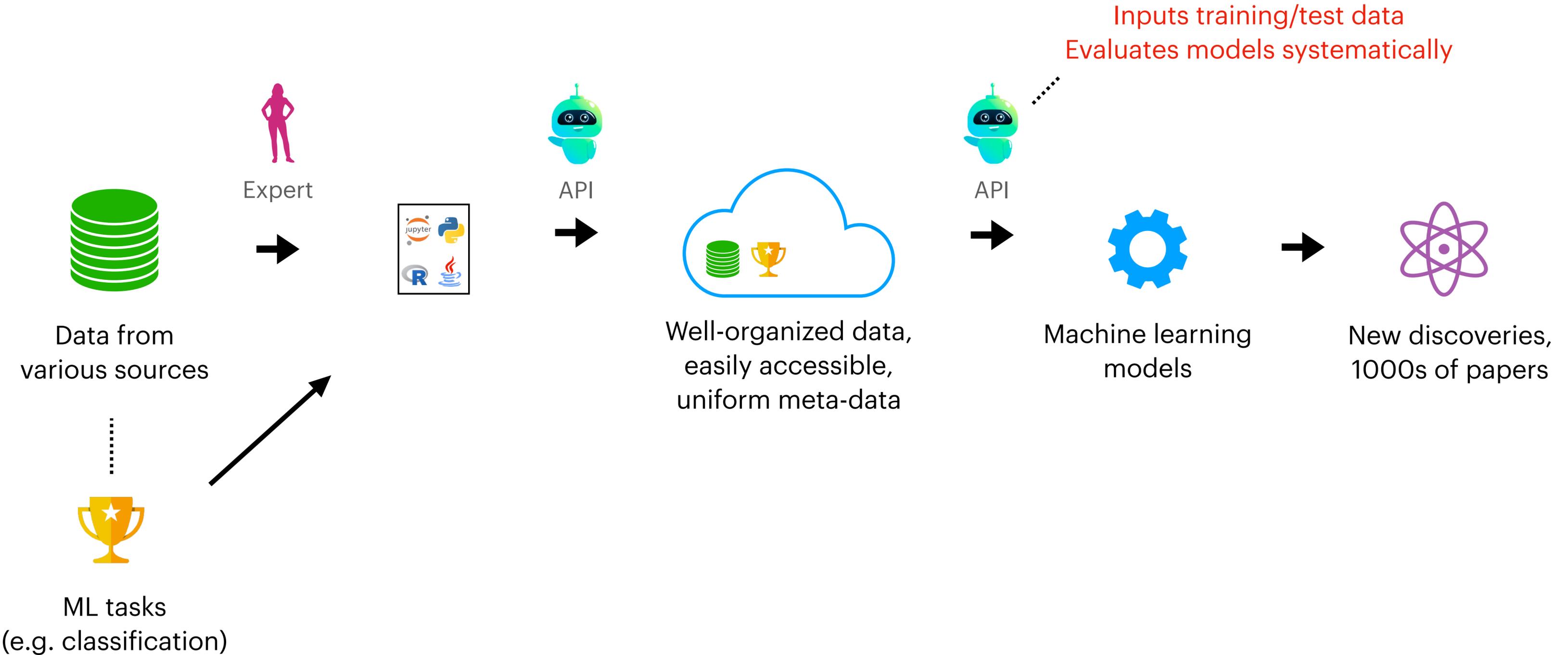
(Required anyway for many ML models)

(Can be a data loading script for large and remotely hosted data)

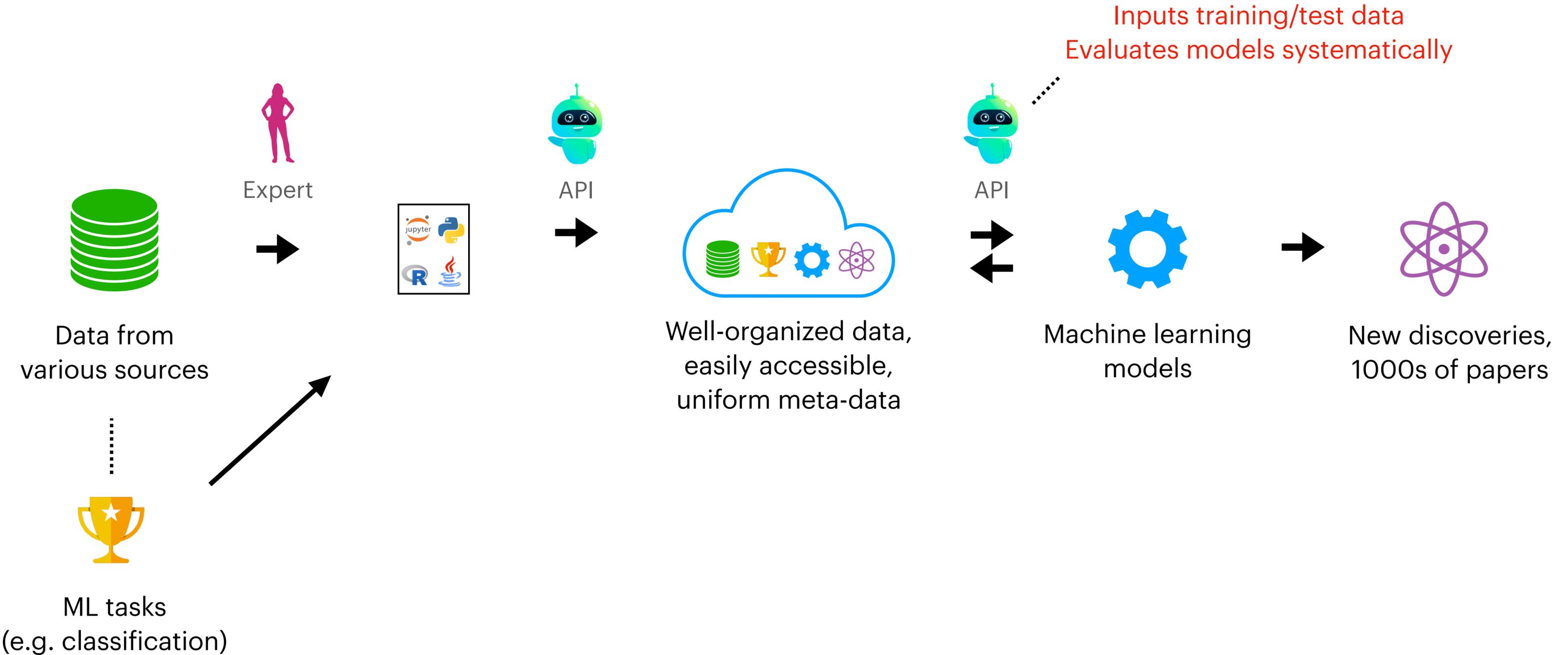
Democratizing ML data



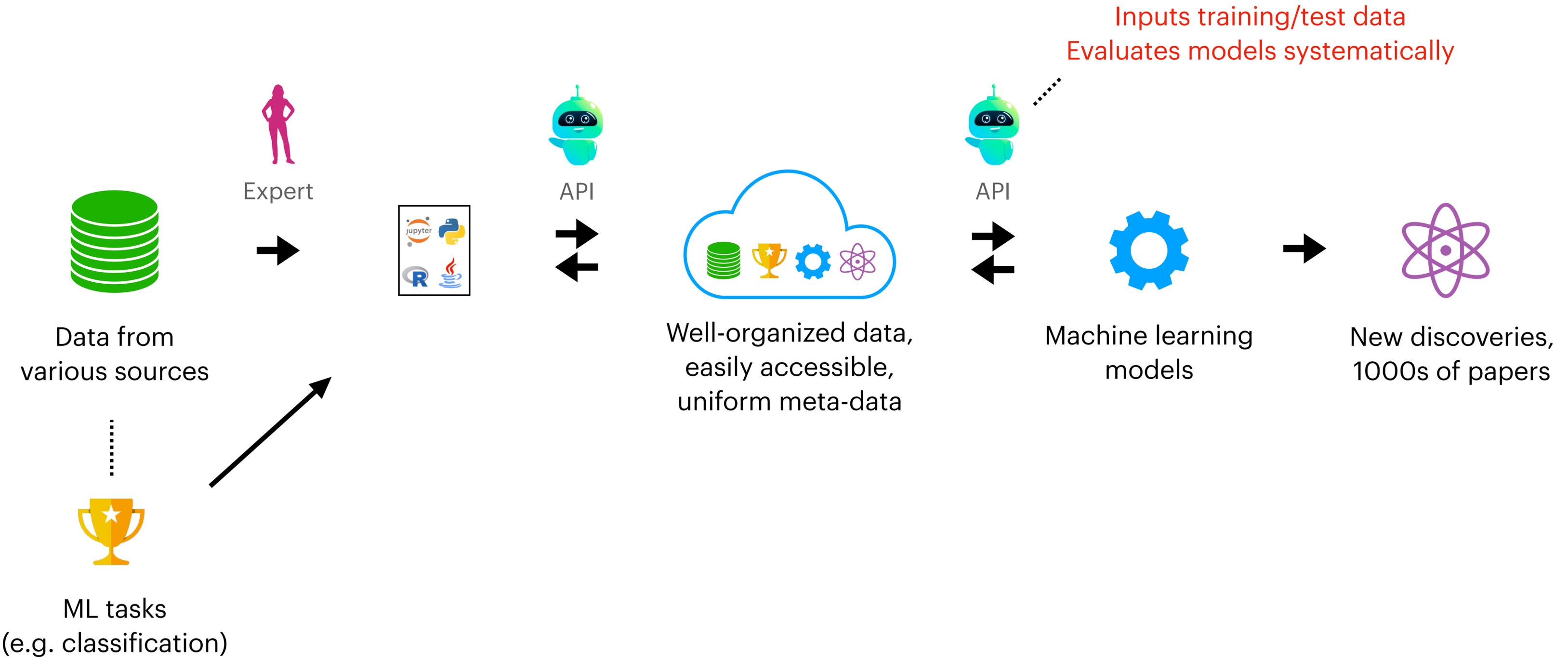
Democratizing ML data



Democratizing ML data



Democratizing ML data



OpenML

An open platform for discovering and sharing ML datasets, algorithms, experiments



API

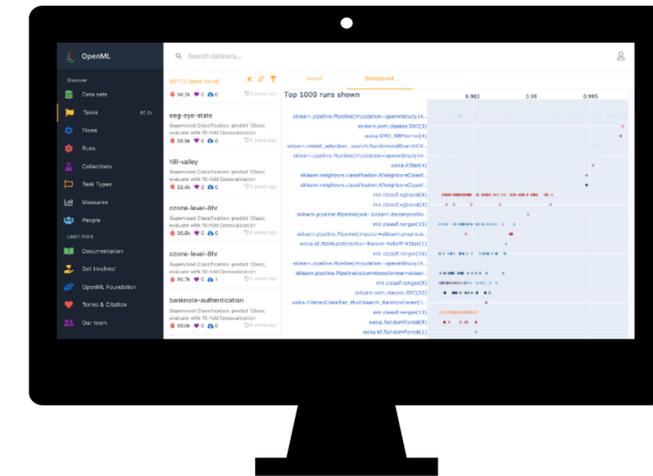


API



Accessible from anywhere, anytime
(scripts, notebooks, apps, cloud jobs)

OpenML



Website
(new.openml.org)

OpenML web interface

Search

Datasets

Dataset analysis

OpenML

Search

- Datasets 7
- Tasks
- Flows
- Runs
- Collections
- Benchmarks
- Task Types
- Measures

Learn

- Documentation
- Blog
- API's
- Contribute
- Meet up
- About us
- Terms & Citation

Minify Dark

covertype

7 datasets found verified

Sign In Sign Up

sylva_prior

Datasets from the Agnostic Learning vs. Prior Knowledge Challenge (<http://www.agnostic.inf.ethz.ch>)

486 14 14.4k x 109 1040 7 years ago v.1

covertype

Normalized version of the Forest Covertypes dataset (see version 1), so that the numerical values are between 0 and 1. Contains the forest cover type for

342 1 40 581k x 55 150 8 years ago v.3

CovPokElec

Dataset created to study concept drift in stream mining. It is constructed by combining the Covertypes, Poker-Hand, and Electricity datasets. More details

332 27 1.46M x 73 149 8 years ago v.1

covertype

Predicting forest cover type from cartographic variables only (no remotely sensed data). The actual forest cover type for a given observation (30 x 30

216 11 110k x 55 180 8 years ago v.1

covertype

This is the famous covertypes dataset in its binary version, retrieved 2013-11-13 from the libSVM site (called covtype.binary there). Additional to the

22 9 581k x 55 293 7 years ago v.2

Data Detail Analysis Tasks

covertype dataset

Choose one or more attributes for distribution plot (first 1k attributes listed)

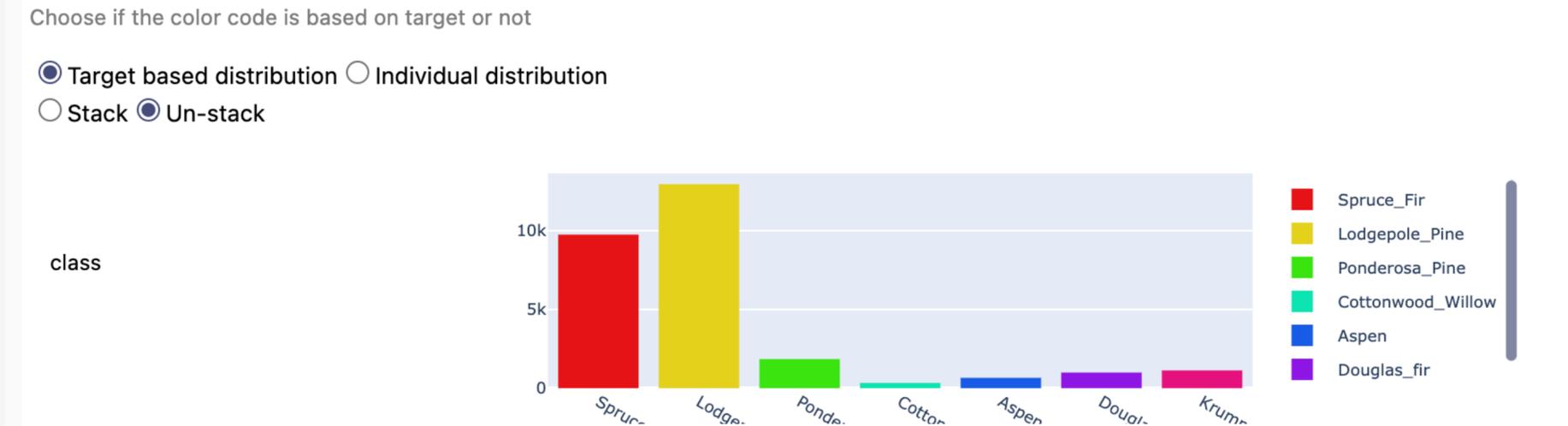
Attribute	DataType	Missing values	# categories	Target	Entropy
<input checked="" type="checkbox"/> class	nominal	0	7	true	1.3
<input checked="" type="checkbox"/> soil_type_28	nominal	0	2		0.01
<input checked="" type="checkbox"/> soil_type_17	nominal	0	2		0.04
<input checked="" type="checkbox"/> soil_type_18	nominal	0	2		0.02
<input checked="" type="checkbox"/> soil_type_19	nominal	0	2		0.04
<input type="checkbox"/> soil_type_20	nominal	0	2		0.08

Distribution plot

Choose if the color code is based on target or not

Target based distribution Individual distribution

Stack Un-stack



OpenML web interface

Tasks

Algorithms

Evaluations (every dot is a model)

OpenML

Discover

- Data sets
- Tasks 90.2k
- Flows
- Runs
- Collections
- Task Types
- Measures
- People

Learn more

- Documentation
- Get involved
- OpenML Foundation
- Terms & Citation
- Our team

Search datasets...

90172 tasks found

96.2k 0 0 2 years ago

eeg-eye-state

Supervised Classification: predict 'Class', evaluate with 10-fold Crossvalidation

95.5k 0 0 4 years ago

hill-valley

Supervised Classification: predict 'Class', evaluate with 10-fold Crossvalidation

92.4k 0 0 2 years ago

ozone-level-8hr

Supervised Classification: predict 'Class', evaluate with 10-fold Crossvalidation

90.8k 0 0 4 years ago

ozone-level-8hr

Supervised Classification: predict 'Class', evaluate with 10-fold Crossvalidation

90.7k 0 1 2 years ago

banknote-authentication

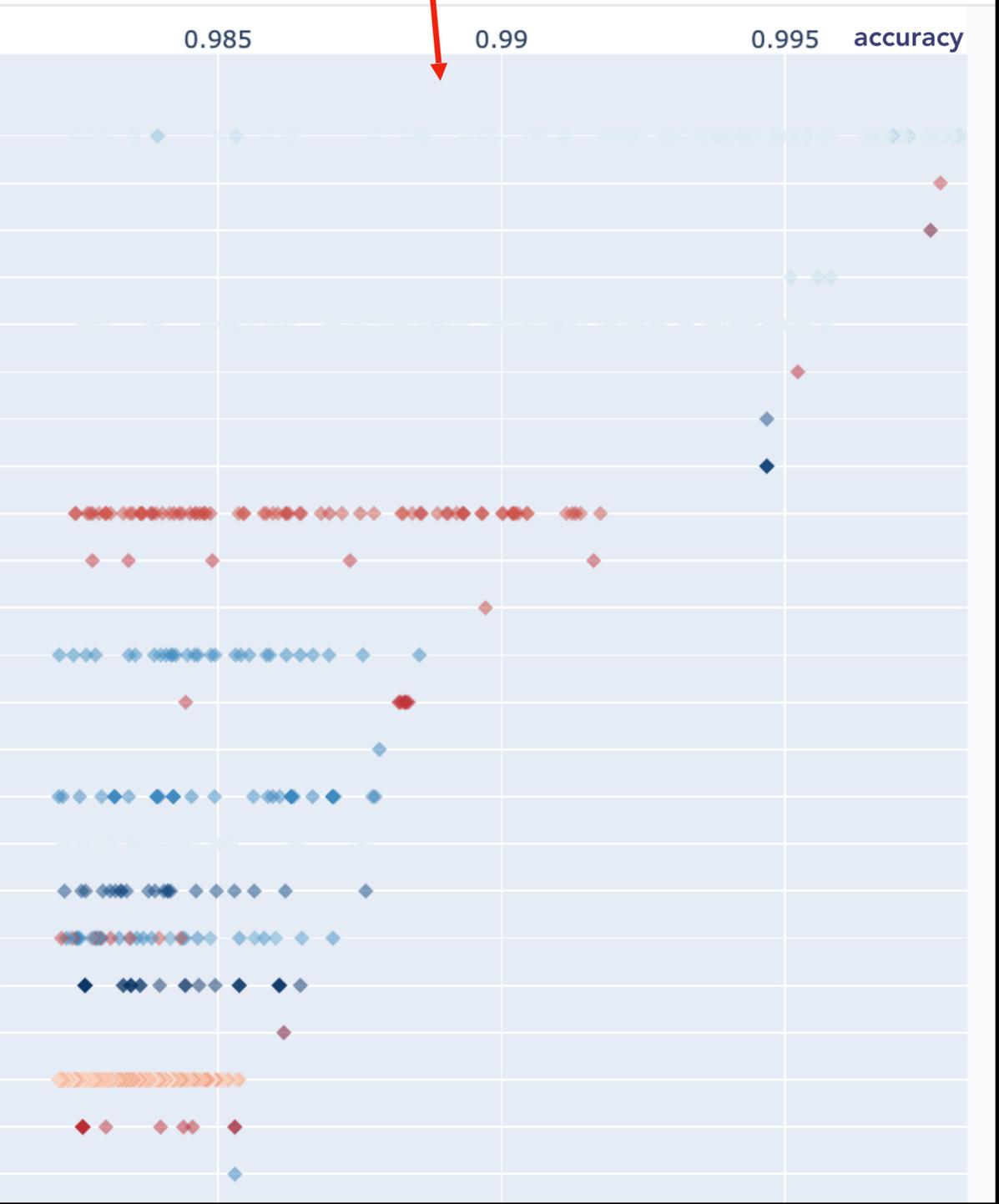
Supervised Classification: predict 'Class', evaluate with 10-fold Crossvalidation

89.0k 0 0 4 years ago

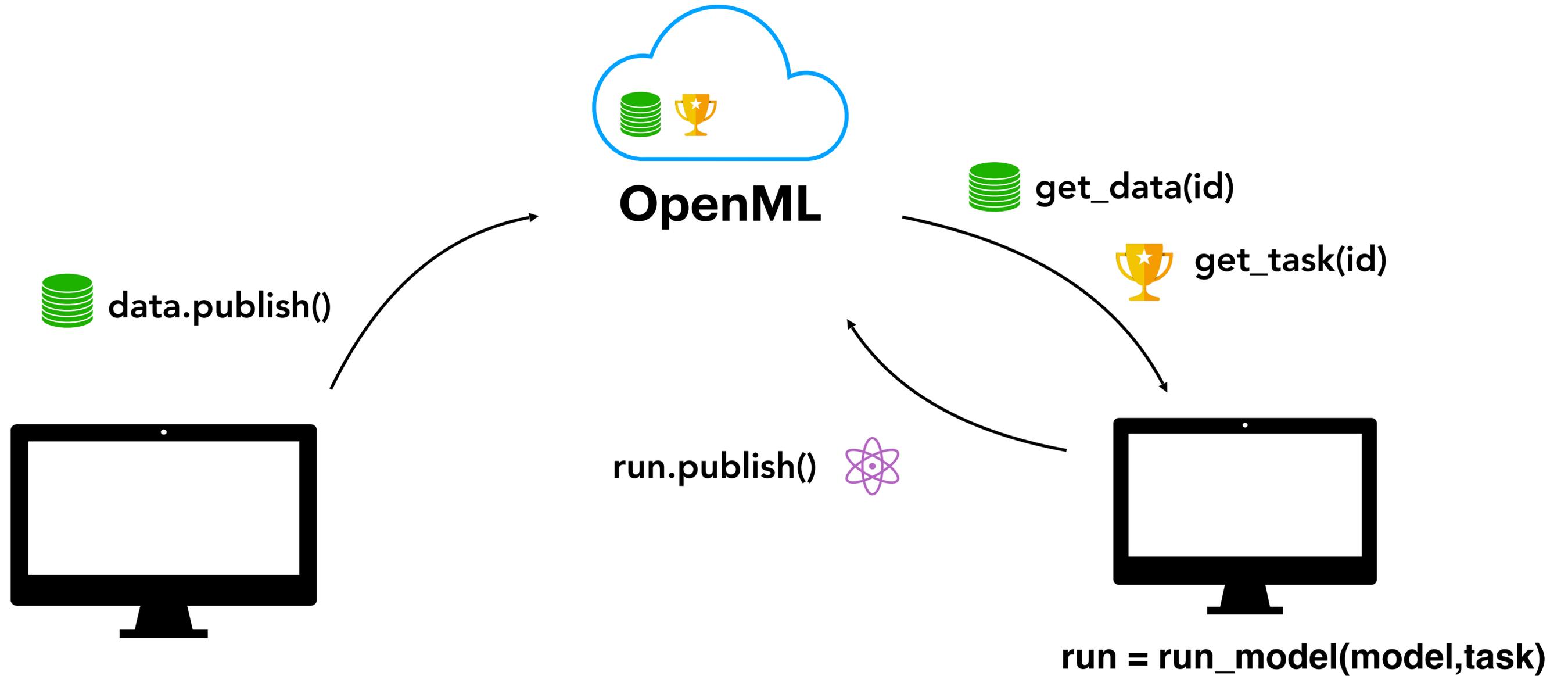
Detail Dashboard

Top 1000 runs shown

- sklearn.pipeline.Pipeline(imputation=openmlstudy14..
- sklearn.svm.classes.SVC(5)
- weka.SMO_RBFKernel(4)
- sklearn.model_selection._search.RandomizedSearchCV..
- sklearn.pipeline.Pipeline(imputation=openmlstudy14..
- weka.KStar(4)
- sklearn.neighbors.classification.KNeighborsClassif..
- sklearn.neighbors.classification.KNeighborsClassif..
- mlr.classif.xgboost(6)
- mlr.classif.xgboost(4)
- sklearn.pipeline.Pipeline(pca=sklearn.decompositio..
- mlr.classif.ranger(15)
- sklearn.pipeline.Pipeline(imputer=sklearn.preproce..
- weka.kf.AttributeSelection-Ranker-ReliefF-KStar(1)
- mlr.classif.ranger(16)
- sklearn.pipeline.Pipeline(imputation=openmlstudy14..
- sklearn.pipeline.Pipeline(columntransformer=sklear..
- mlr.classif.ranger(9)
- sklearn.svm.classes.SVC(32)
- weka.FilteredClassifier_MultiSearch_RandomForest(1..
- mlr.classif.ranger(13)
- weka.RandomForest(9)
- weka.kf.RandomForest(1)



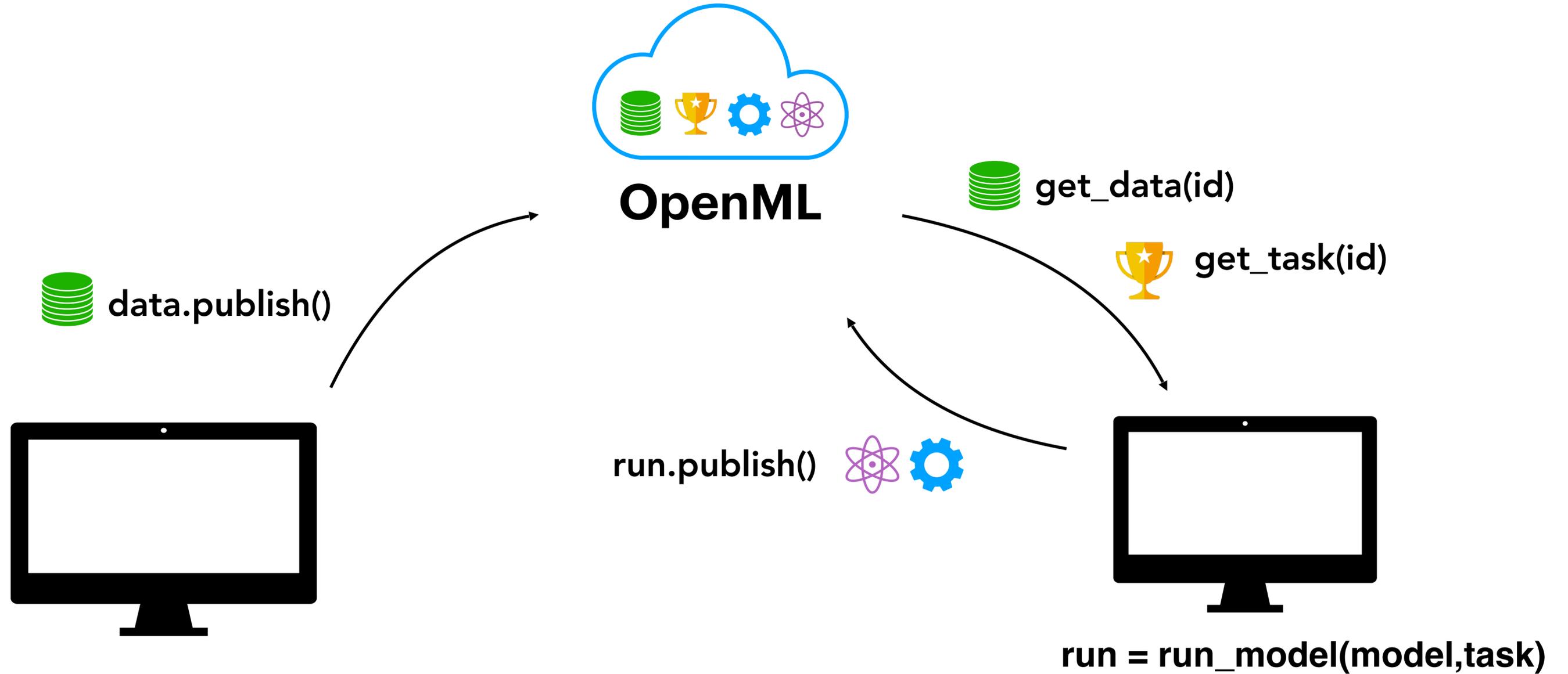
Frictionless machine learning



APIs:    

Integrations:     

Frictionless machine learning



APIs:    

Integrations:     

Frictionless machine learning



```
from sklearn import ensemble
from openml import tasks, runs

model = ensemble.RandomForestClassifier()
task = tasks.get_task(3954)
run = runs.run_model_on_task(model, task)
run.publish()
```



Frictionless machine learning



```
● ● ●  
  
from torch.nn  
from openml import tasks, runs  
  
model = torch.nn.Sequential(processing_net,  
                             features_net, results_net)  
task = tasks.get_task(3954)  
run = runs.run_model_on_task(clf, task)  
run.publish()
```



Benchmarking suites

- How can we build better, more general benchmarks?
 - Start with a large set of datasets (e.g. OpenML)
 - Define strict set of constraints
 - Retrieve and test models on all matching datasets
 - Gather results from different researchers in a central place (e.g. OpenML)
- Offers a way to really use benchmark suites and converge to well-defined accepted suites
- Are meant to be dynamic: evolve with new datasets joining over time

Benchmarking suites

- All results can be streamed, organized, downloaded to/from OpenML

OpenML

Search

- Datasets
- Tasks
- Flows
- Runs
- Collections

Tasks

Runs 130

Task Types

Measures

Learn

- Documentation
- Blog
- API's
- Get involved

Search collections

130 collections found run

Sign In Sign Up

A large-scale comparison of classification algorithms
We investigate the performance of a wide range of classification algorithms on a wide range of datasets to better understand when they perform well and

512 514 63 91.4k 1 6 years ago

Does Feature Selection Improve Classification?
Feature selection can be of value to classification for a variety of reasons. Real world data sets can be rife with irrelevant features, especially if the data was

394 394 24 9.45k 15 3 years ago

AutoML Benchmark Study
Run results of the ongoing AutoML benchmark, see <https://openml.github.io/automlbenchmark/>. The benchmark includes both

19 19 6 117 226 2 years ago

CC18-Example
Description

39 39 1 39 275 23 hours ago

Prefetching for SPARQL endpoints
Data prefetching is a standard technique used to accelerate the access to data

Run Collection Analysis Tasks Runs

Show scatter plot of results

Predictive Accuracy

Show results for each fold (can be slow)

Legend:

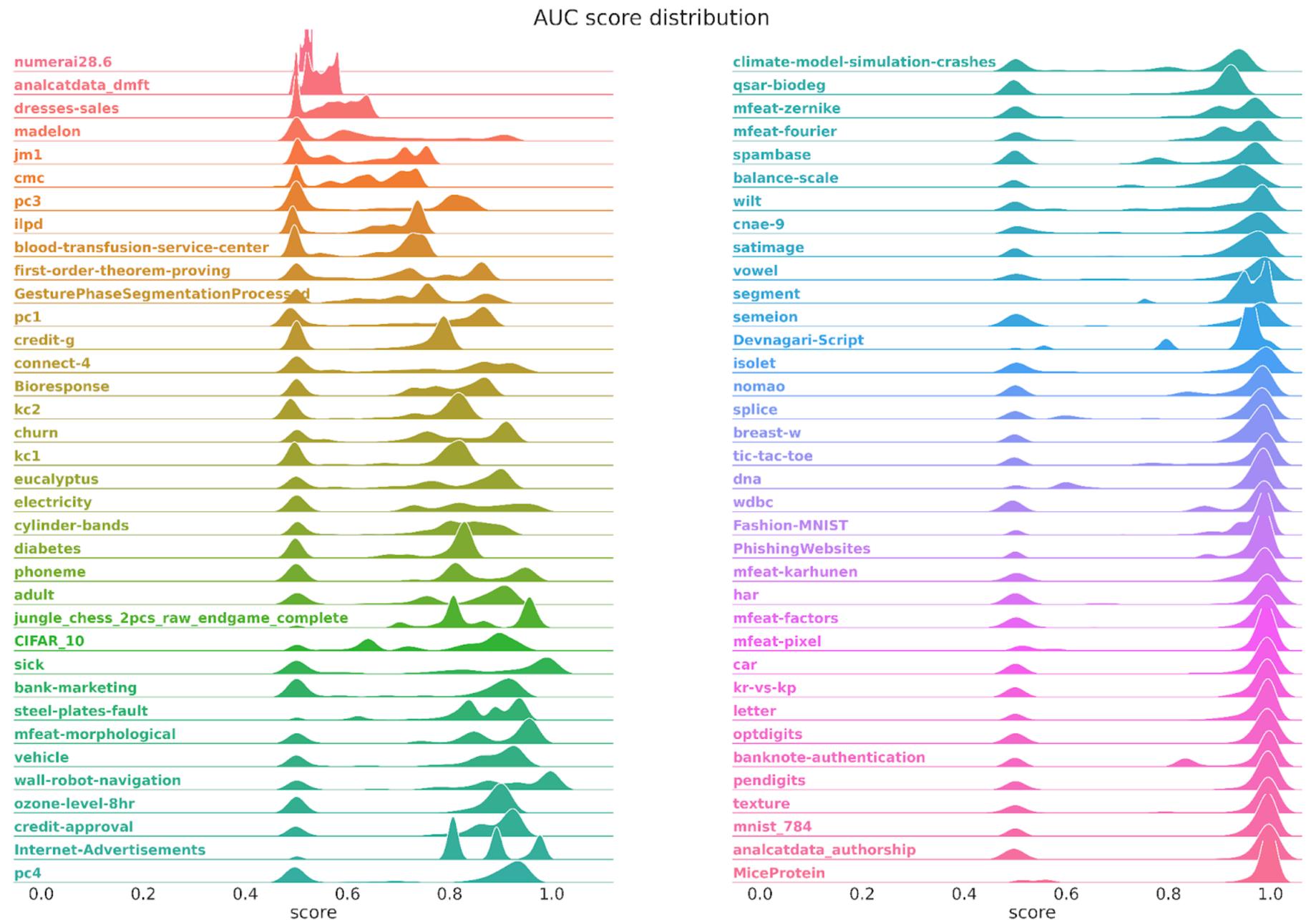
- automlbenchmark_autosklearn(1)
- automlbenchmark_h2oautoml(1)
- automlbenchmark_tpot(1)
- automlbenchmark_autoweka(1)
- automlbenchmark_randomforest(1)
- automlbenchmark_tunedrandomforest(1)

Dataset:

- adult
- Amazon_emplo...
- APSFailure
- bank-marketing
- connect-4
- Fashion-MNIST
- guillermo
- helena
- higgs
- jannis
- jungle_chess...
- KDDCup09_app...
- MiniBooNE
- nomao
- numerai28.6
- riccardo
- robert
- shuttle
- volkert

Benchmarking suites

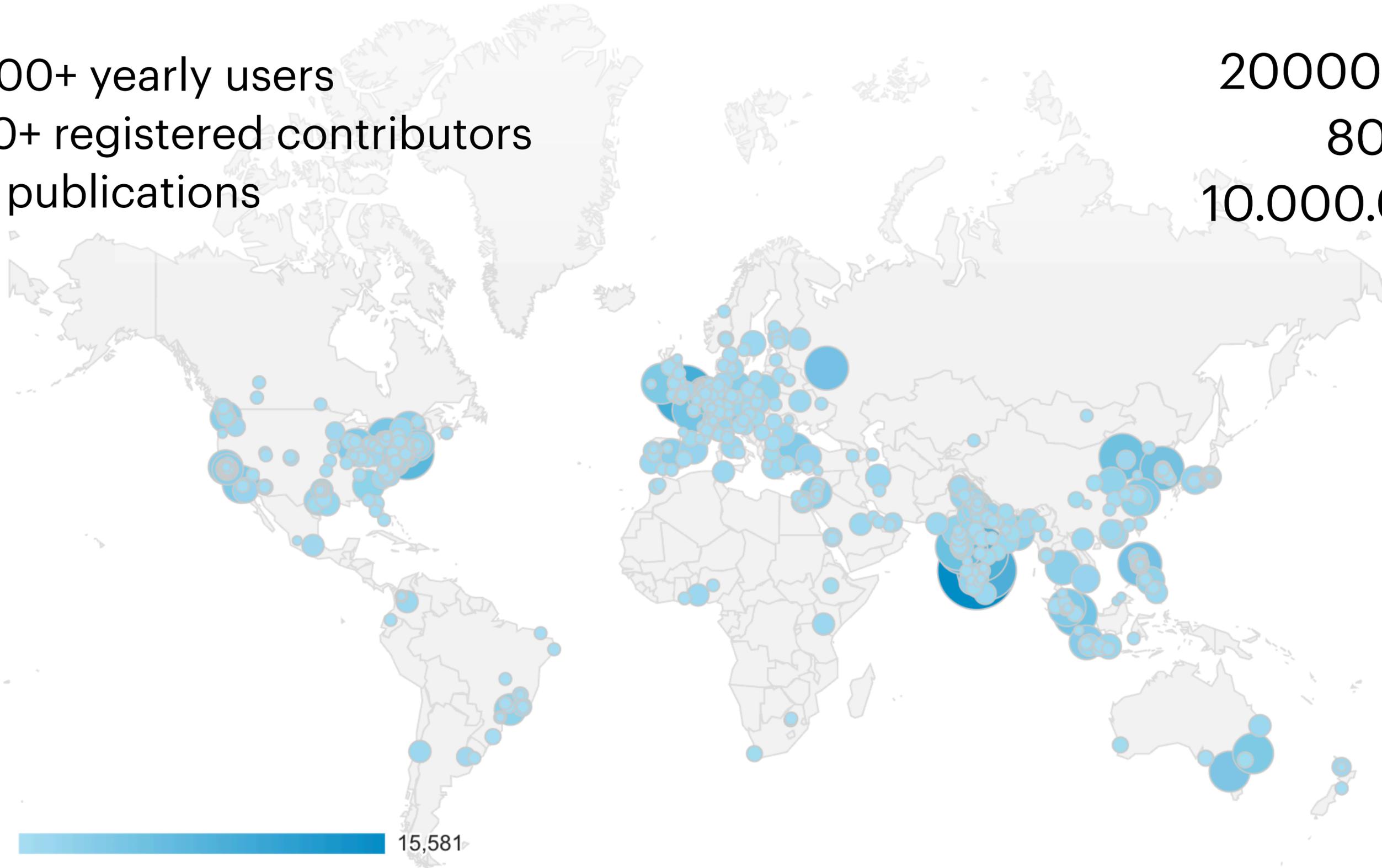
- Example: OpenML-CC18
 - 3.8 million results
- Classification only
- 72 datasets
- Contain missing values and categorical features
- Medium-sized (500-100000 observations, <5000 features after one-hot-encoding)
- Not unbalanced
- No groups/block/time dependencies
- No sparse data
- Some more subjective criteria (see paper)



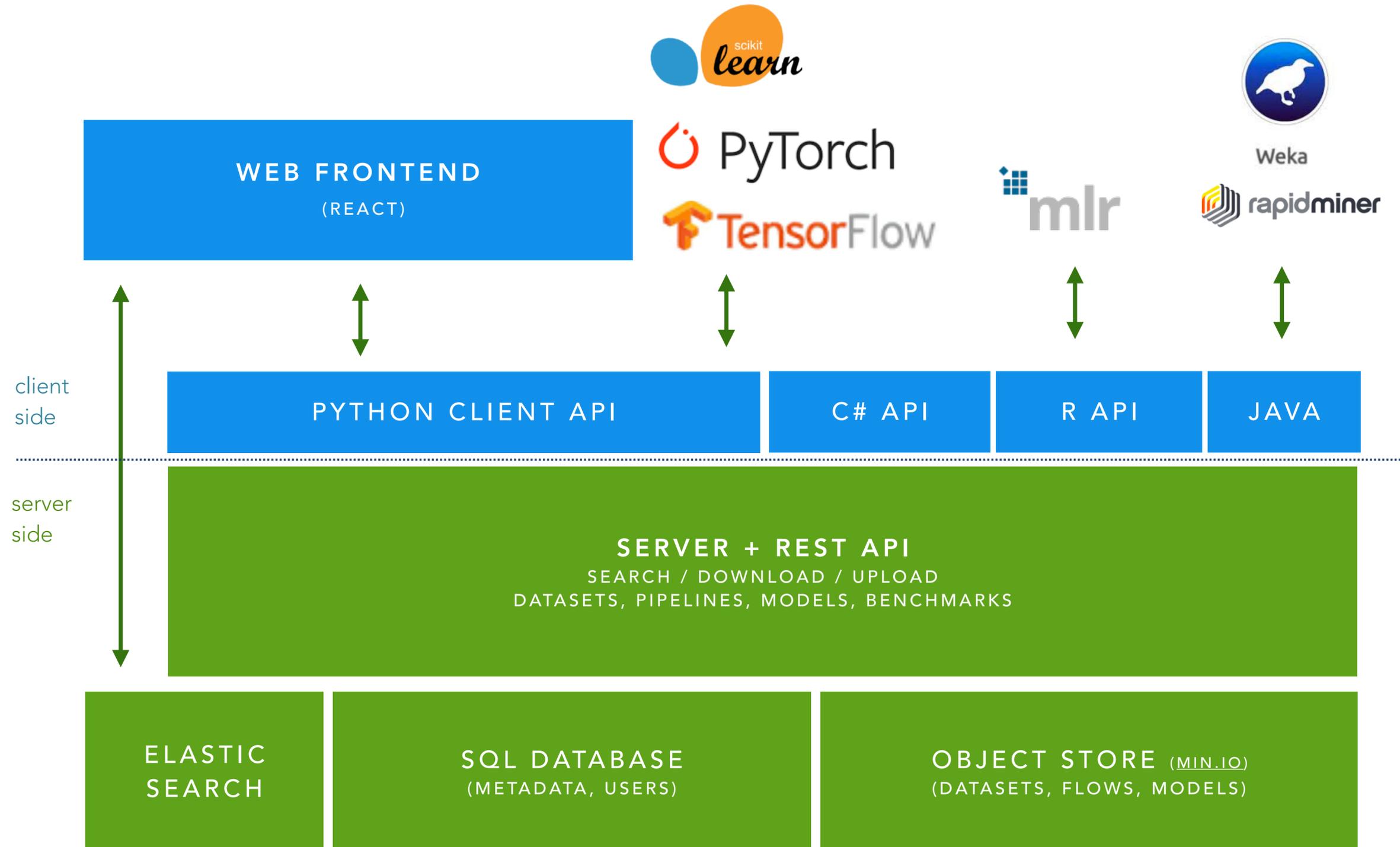
OpenML Community

250000+ yearly users
13000+ registered contributors
900+ publications

20000+ datasets
8000+ flows
10.000.000+ runs



OpenML Architecture



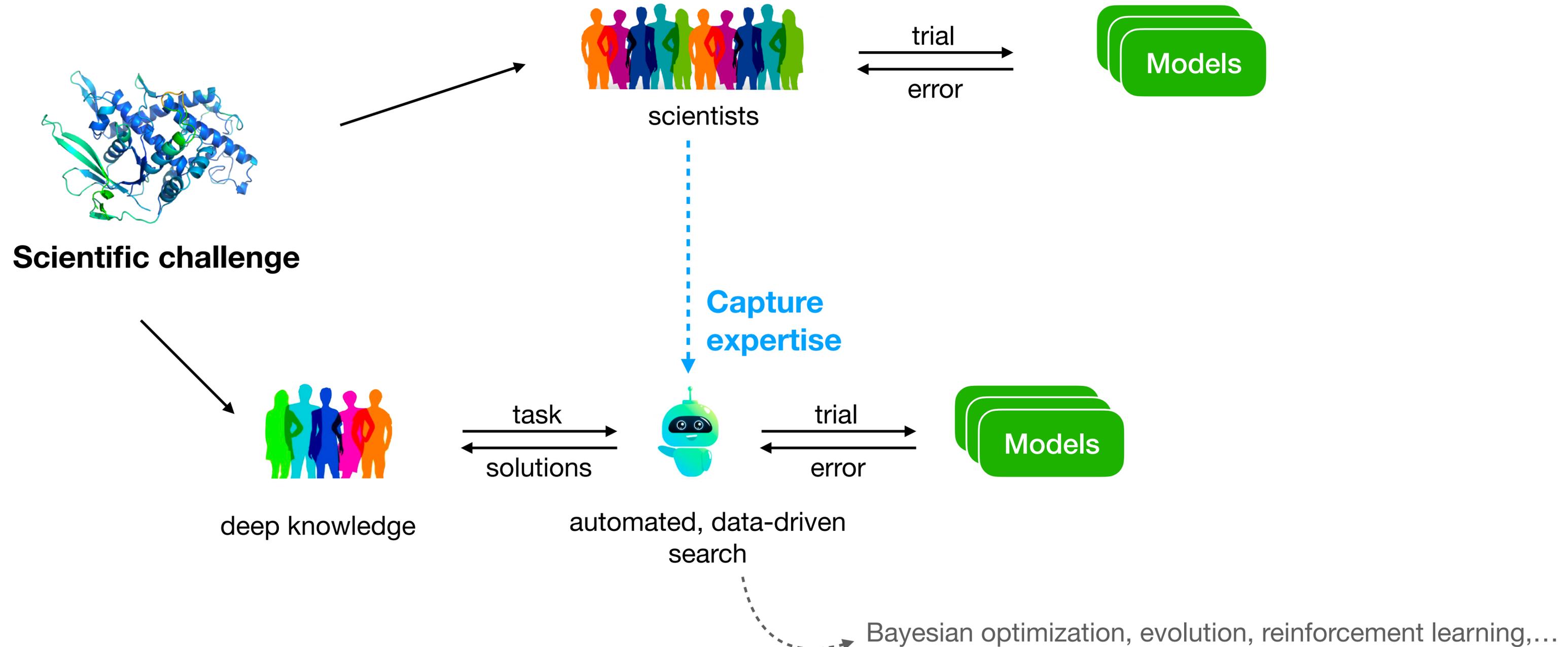
Democratizing machine learning itself



Now that we have data on millions of experiments, can we automate the building and tuning of machine learning models?

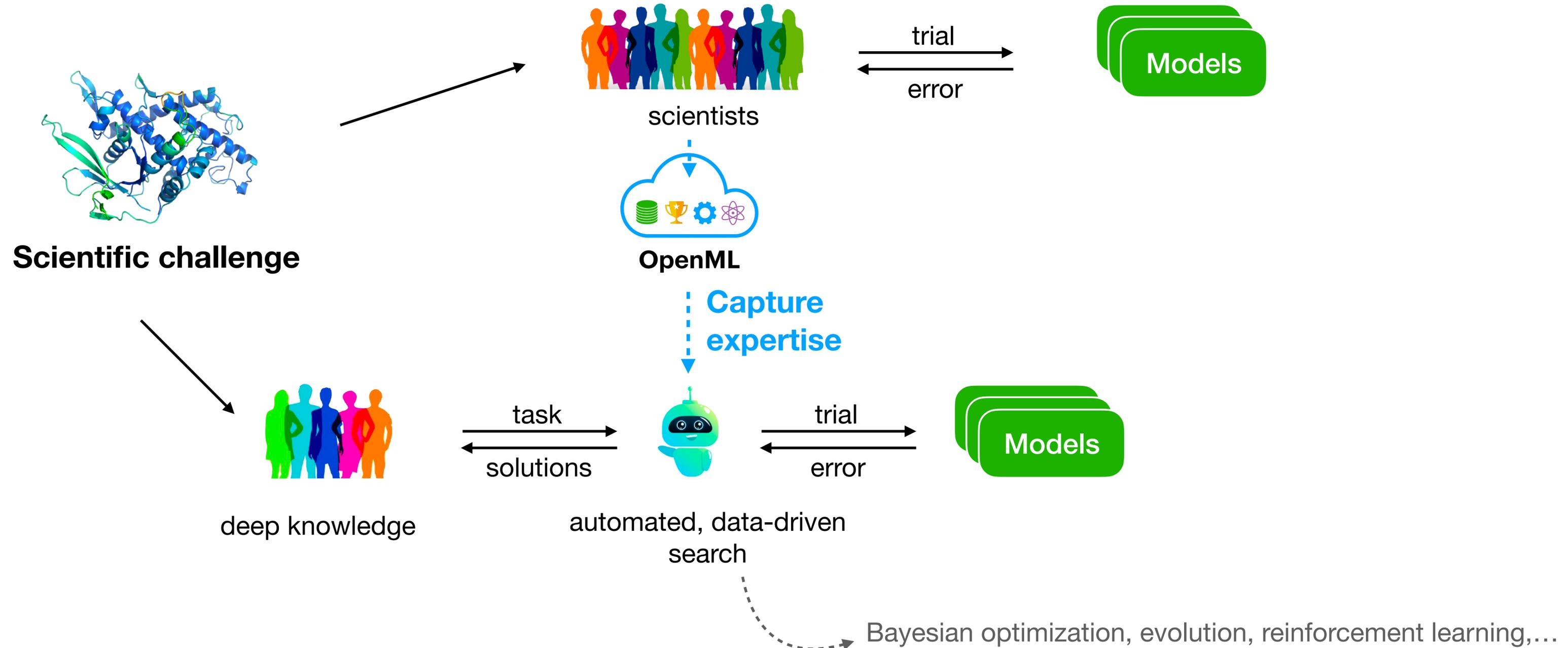
Automatic Machine Learning (AutoML)

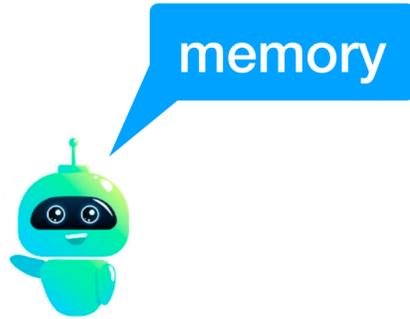
Replace manual trial and error with automated search (based on prior experience)



Automatic Machine Learning (AutoML)

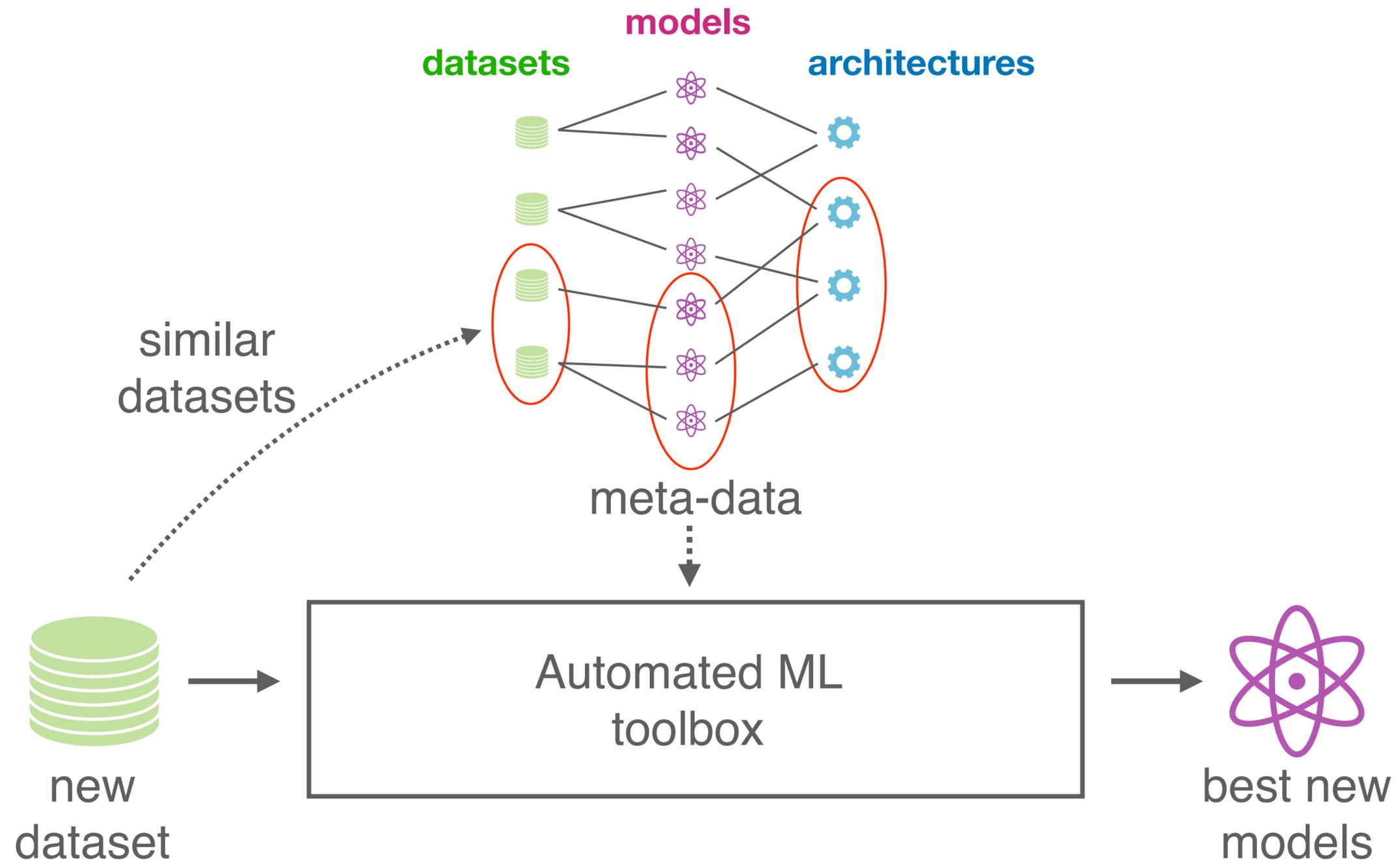
Replace manual trial and error with automated search (based on prior experience)



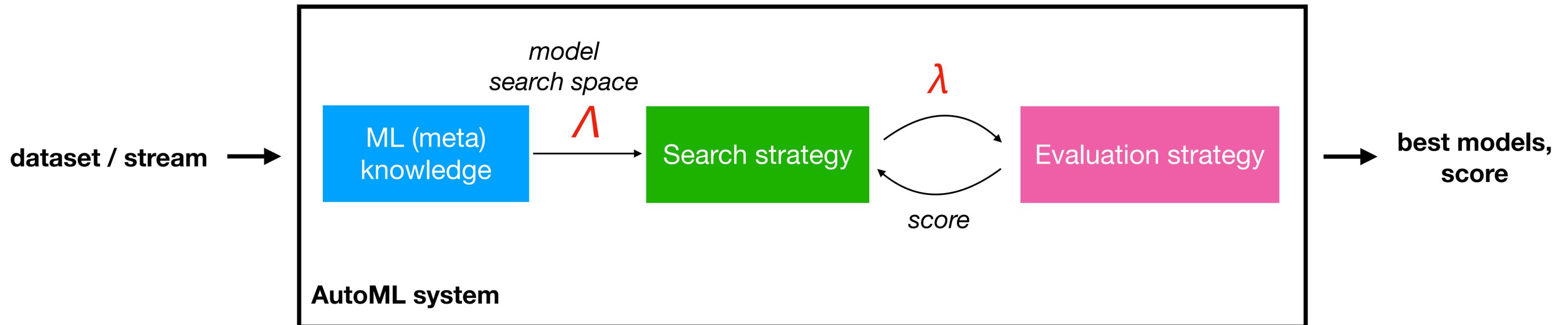


OpenML as a global memory

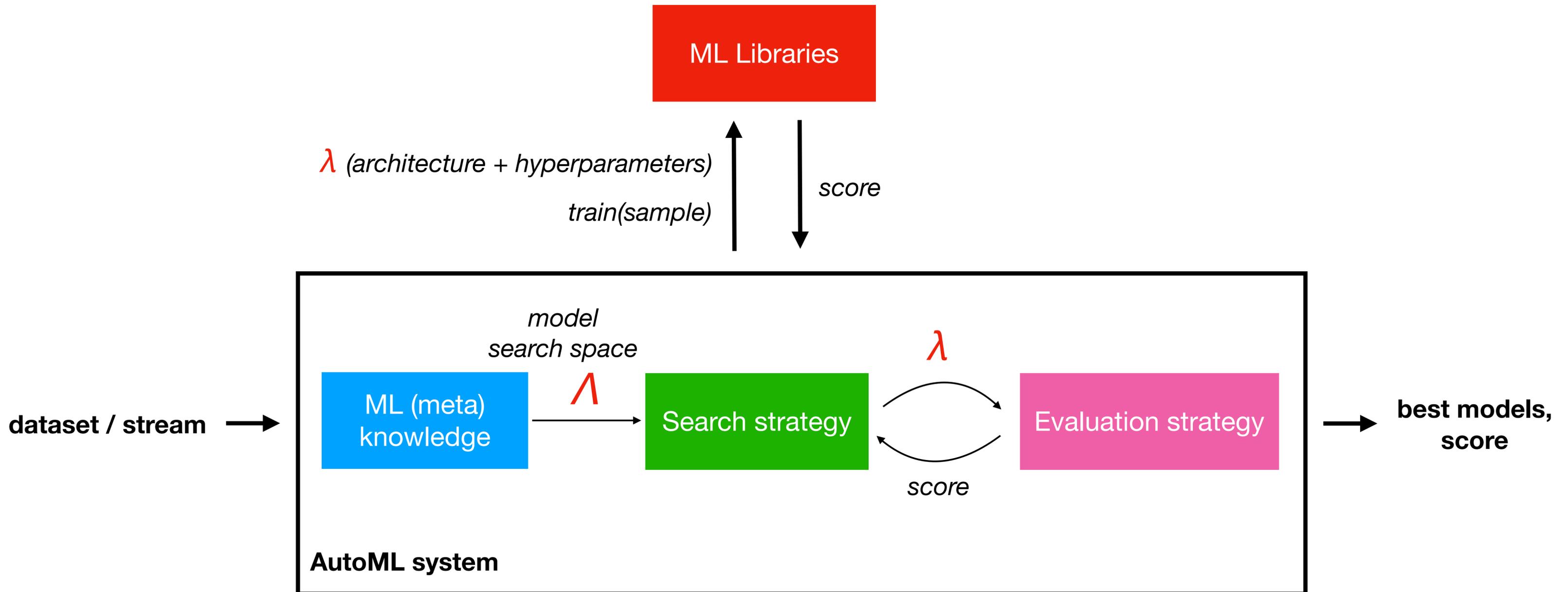
Machine-readable repository of machine learning results



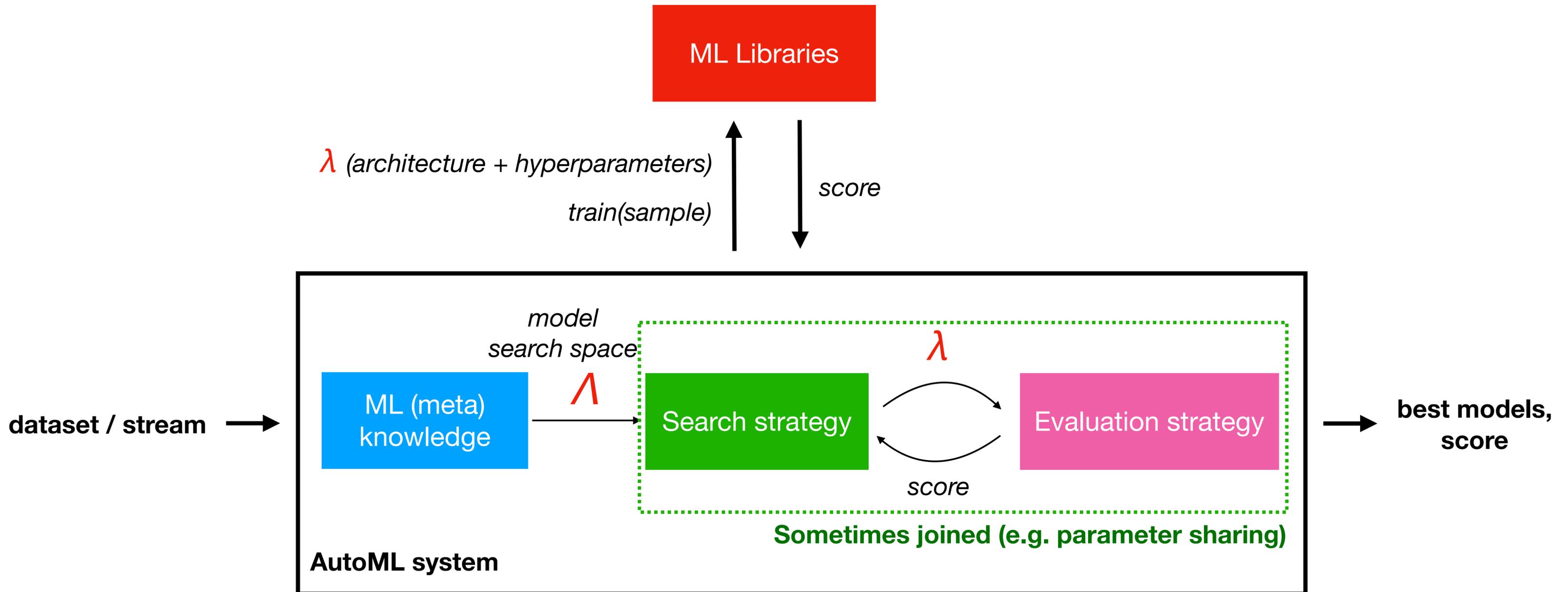
Structure of AutoML systems



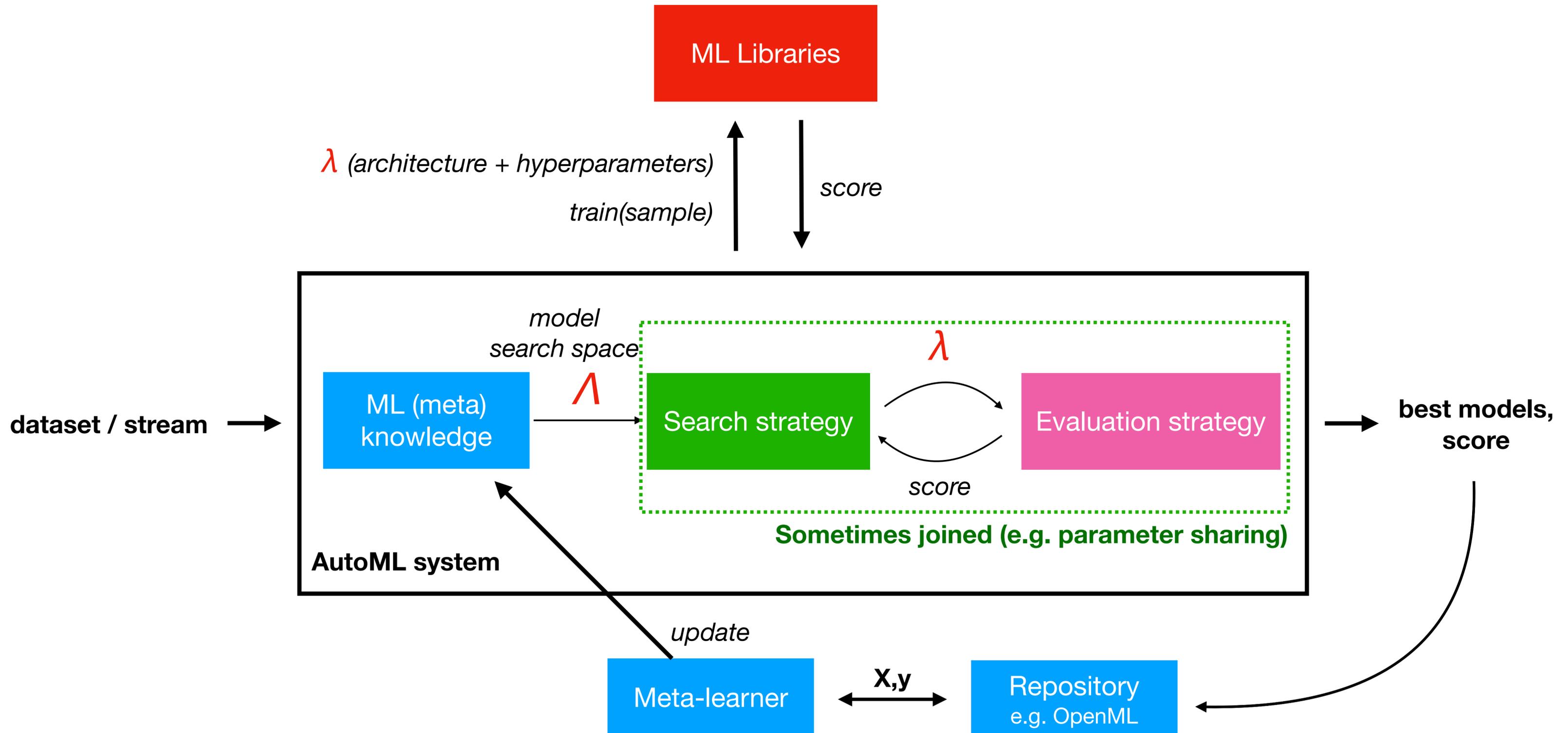
Structure of AutoML systems



Structure of AutoML systems

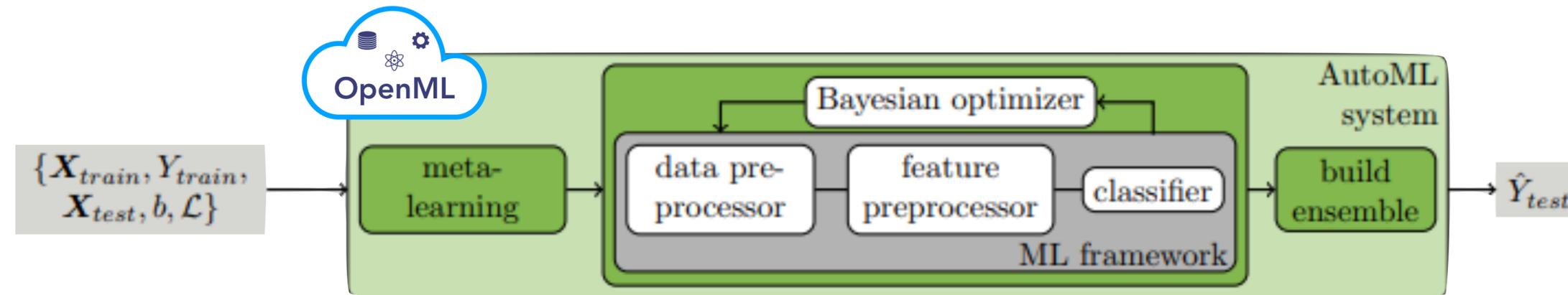


Structure of learning AutoML systems



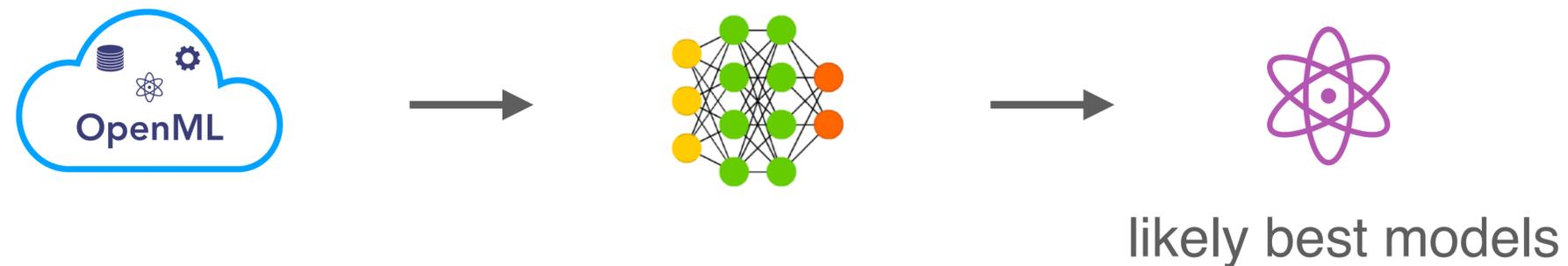
Automating machine learning

auto-sklearn: uses OpenML to *warm-start* the search for the best pipelines



Feurer et al. 2020

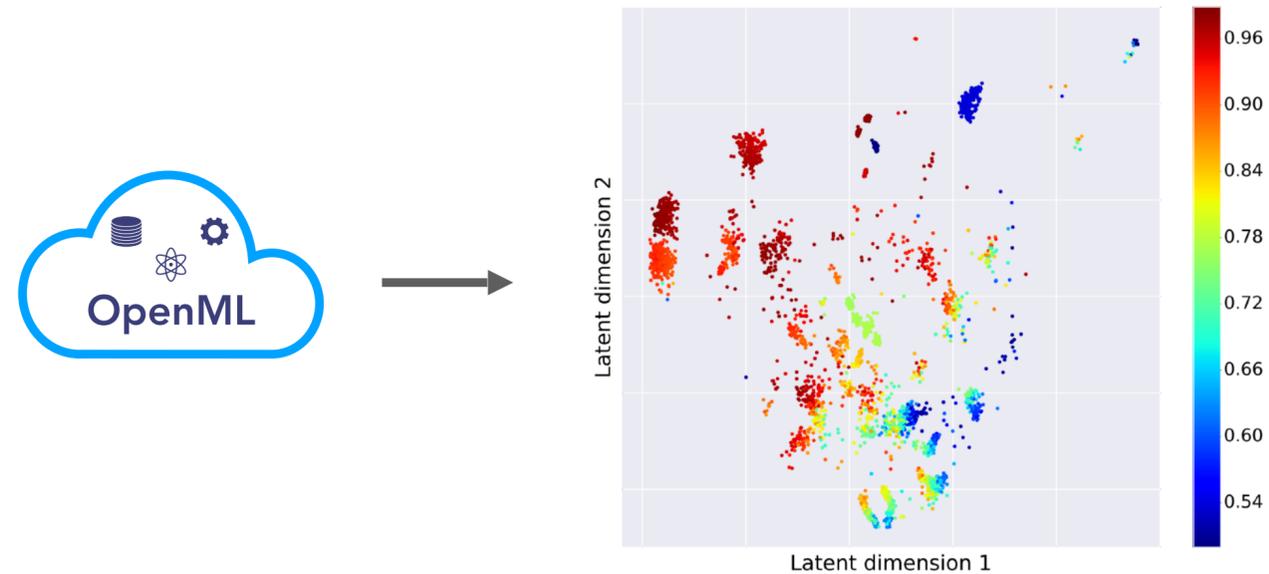
ABLR (Amazon): uses OpenML to learn how to search hyperparameters



Perrone et al. 2018

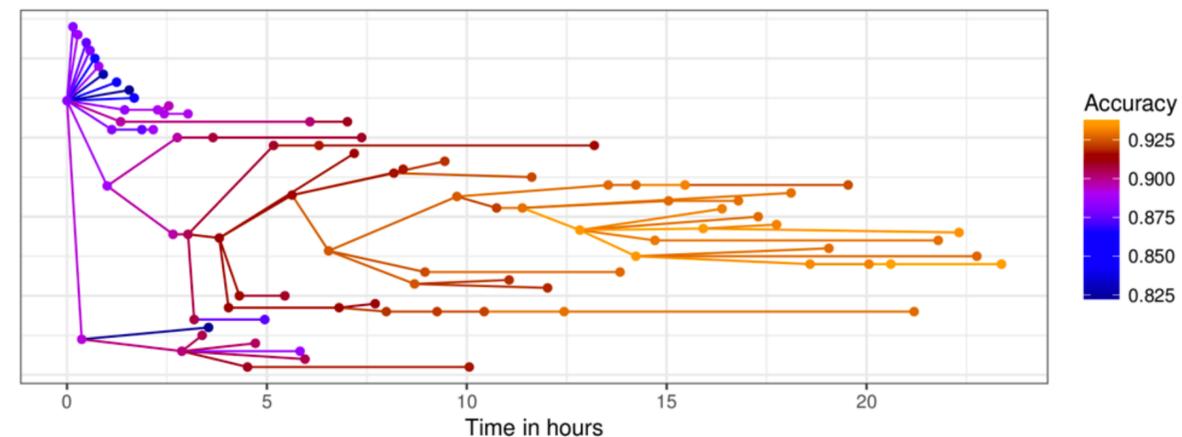
Automating machine learning

ProbMF (Microsoft): uses OpenML to recommend the best algorithms



Fusi et al. 2018

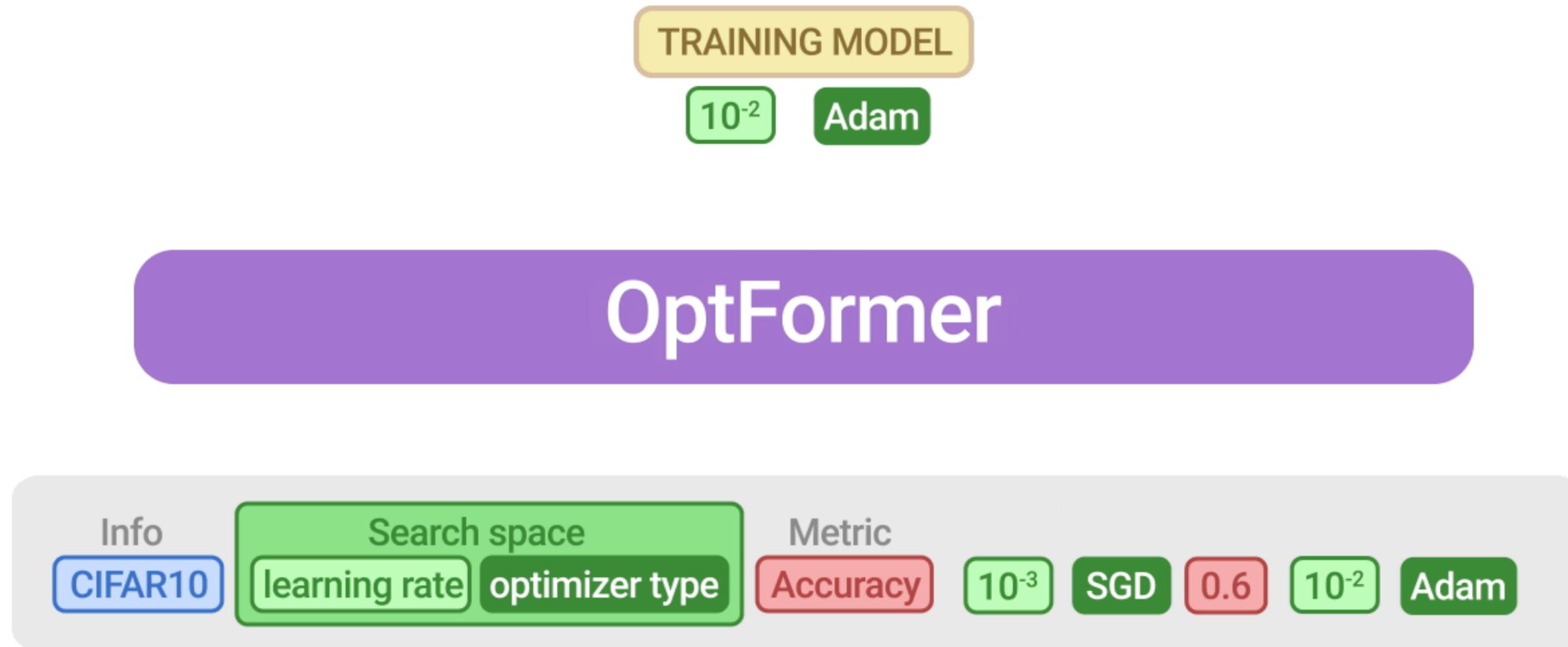
GAMA (TU/e): modular AutoML system, handles wide range of tasks



Gijsbers et al. 2018 - 2022

Automating machine learning

OptFormer (DeepMind): uses OpenML to train a transformer model, predict the next models to try





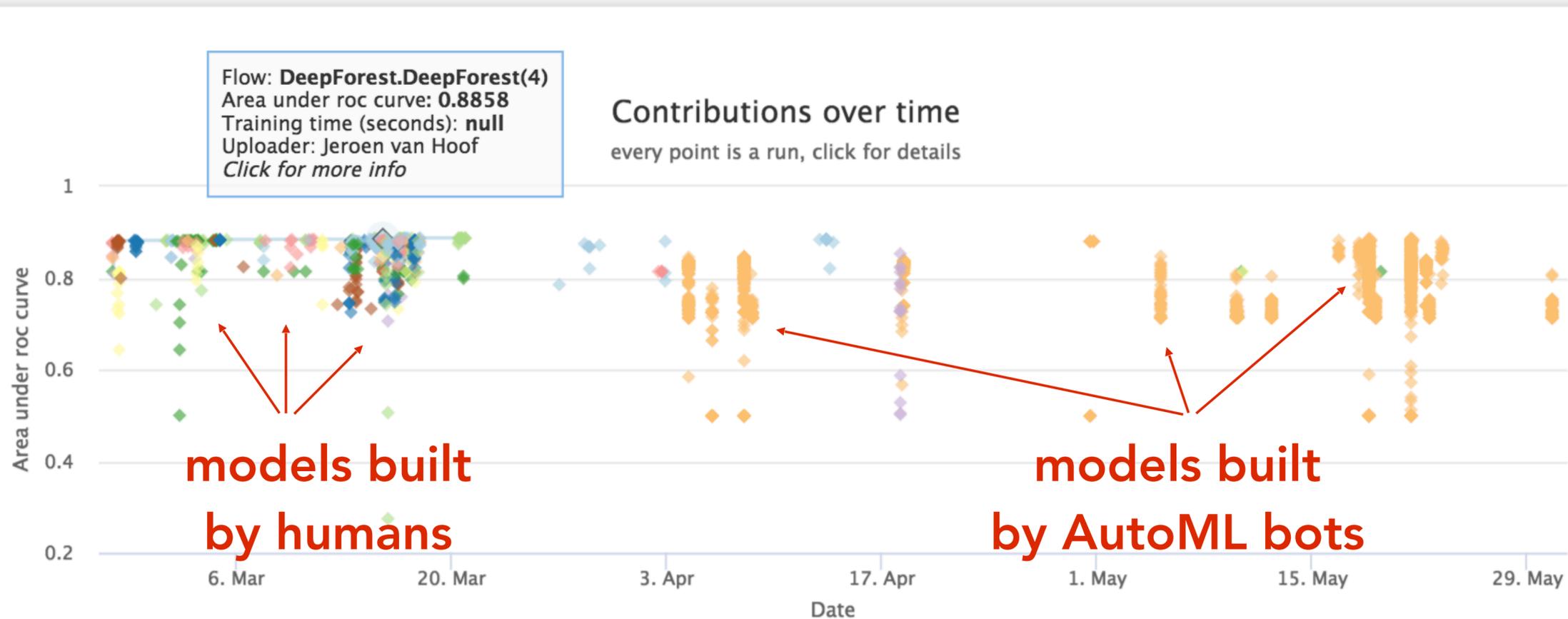
Human-AI interaction

Algorithms learn from models shared by humans

Humans learn from models built by bots

Timeline

Metric: AREA UNDER ROC CURVE



- frontier
- Joaquin Vanschoren
- Hilde Weerts
- edorigatti
- Joel Goossens
- Niels Hellinga
- Mingpeiyu Zhang
- Evertjan Peer
- stevens jethofer
- Hongliang Qiu
- Yezi Zhu
- János Szedelényi
- Chin-Fang Lin
- Wenting Xiong
- M de Roode
- Tianyu Zhou
- Lirong Zhang
- Ruud Andriessen
- Stefan Majoor
- Angelo Majoor
- Changbin Lu
- Irfan Nur Afif
- Nan Yang
- Niels de Jong
- Thomas Hagebols
- Stanley Clark
- Joost Visser
- Jeroen van Hoof
- Xiaolei Wang
- Timothy Aerts
- Lieuwe Stooker
- Corbin Joosen
- Jos Mangnus
- Luis Armando Perez Rey
- Jet van den Broek
- Thijs Ledebouer
- Brent van Strien
- Arun Tom Skariah
- Sako Arts
- Xuqiang Fang
- Yongyu Fan
- Suraj Iyer
- Filip Obers
- Laurens Reulink
- Kevin van Eenige
- Tong Wu
- Jan van Rijn
- y q
- OpenML_Bot R
- Raphaël Couronné
- Mikaël Le Bars

Join us! (and change the world)

Active open source community

- Hackathons 2-3x a year

OpenML Foundation

- Sponsorship, science

OpenML spin-off: PortML

- Services, projects



We're hiring!
2 Research Engineer positions
at TU Eindhoven

Thanks to the entire OpenML star team



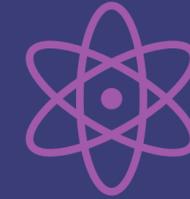
Pieter Gijsbers



Matthias Feurer



Heidi Seibold



Bernd Bischl



Andreas Müller



Sahithya Ravi



Giuseppe Casalicchio



Michel Lang



Jan van Rijn



Prabhant Singh



Marcel Wever



Erin Ledell



Bilge Celik



Janek Thomas



Sebastian Fischer



Neil Lawrence



and many more!

Thank you!

谢谢

 @open_ml

 OpenML

 www.openml.org

