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

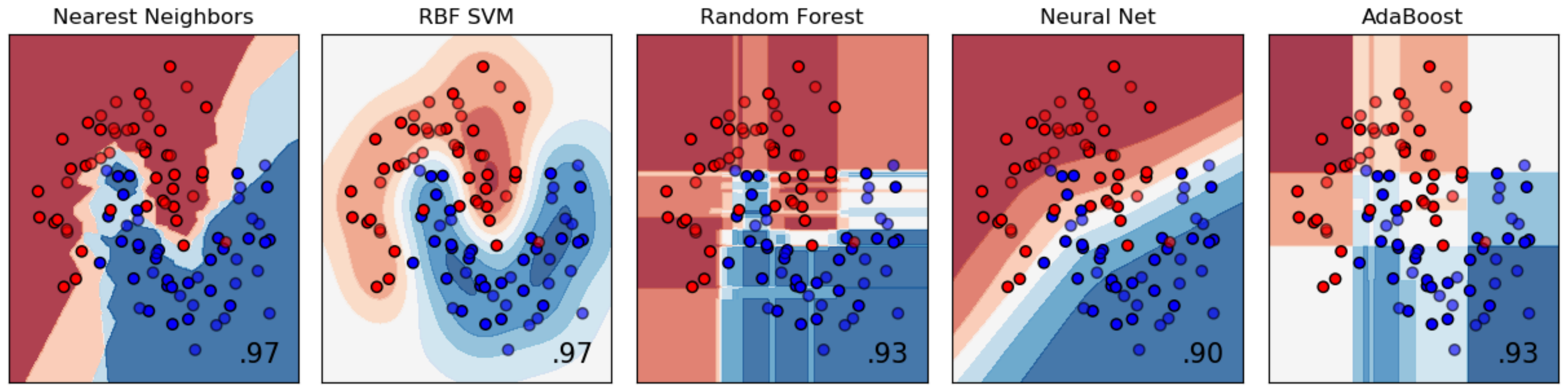
collection,
cleaning,
preprocessing,
featurization,
selection,…

Data

Machine Learning is labor-intensive

Infinite range of possibilities, tacit experience

Model selection



collection, cleaning, preprocessing, featurization, selection,…

Data

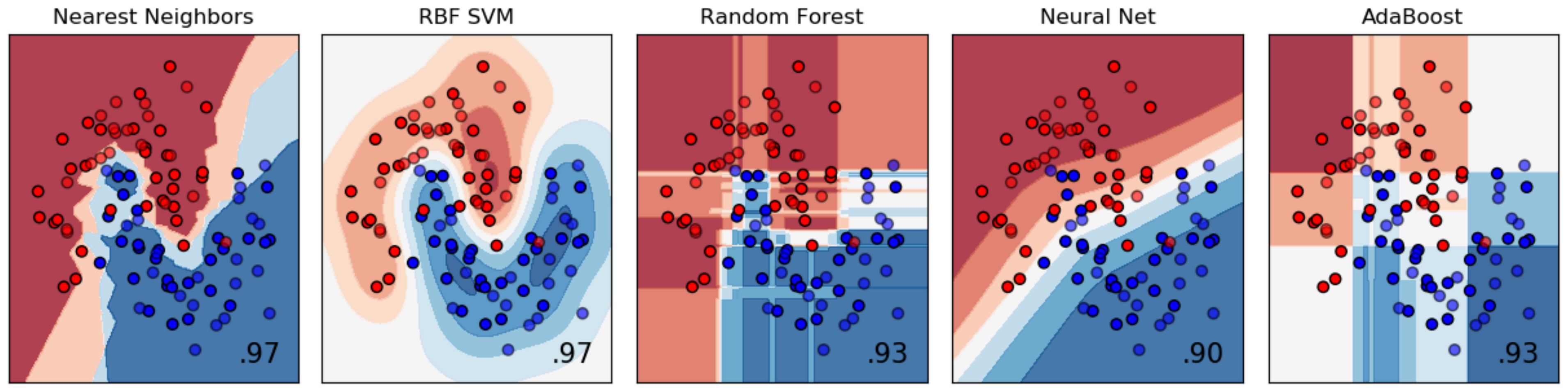
Machine Learning is labor-intensive

Infinite range of possibilities, tacit experience

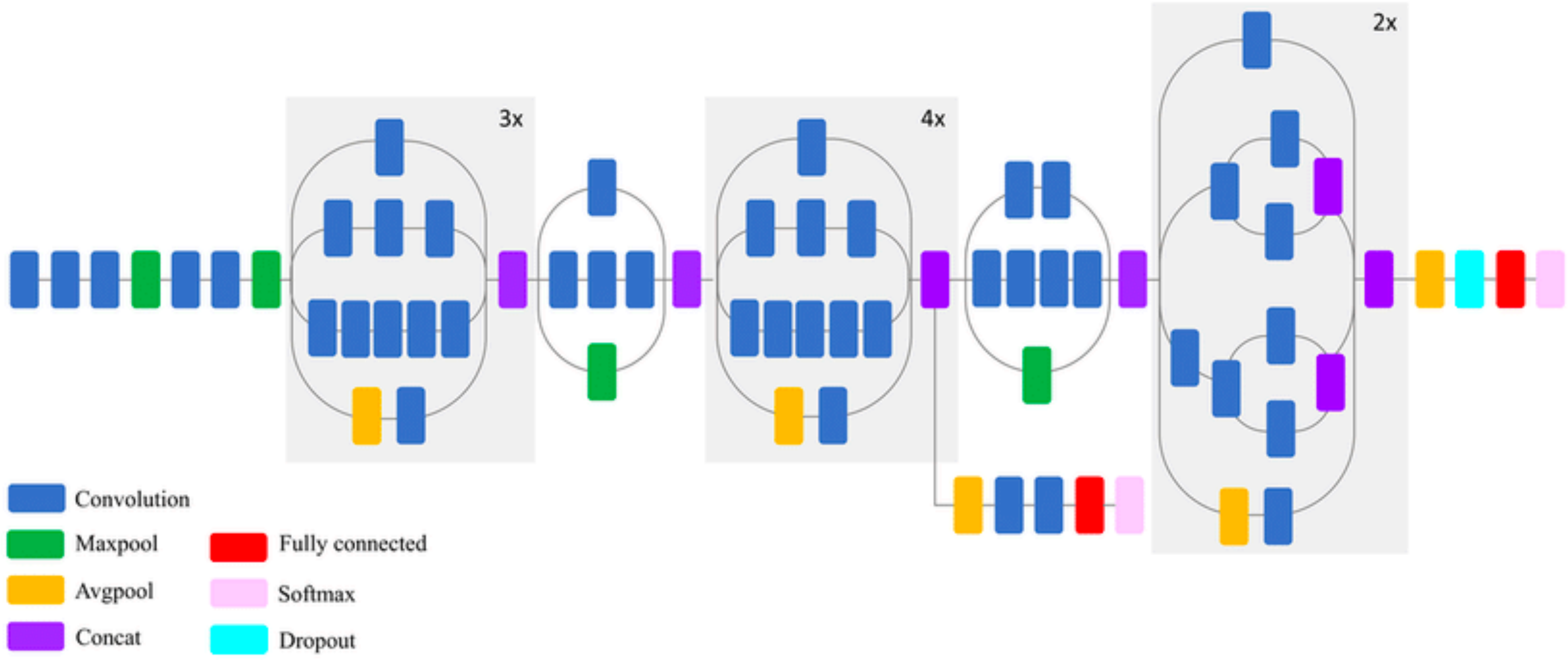
collection,
cleaning,
preprocessing,
featurization,
selection,…

Data

Model selection



Neural architecture search



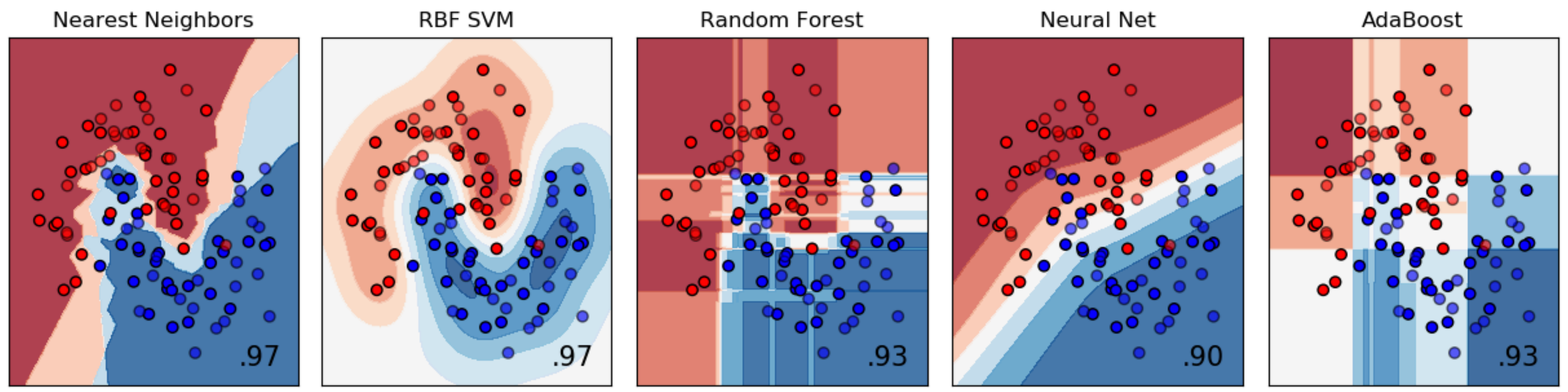
Machine Learning is labor-intensive

Infinite range of possibilities, tacit experience

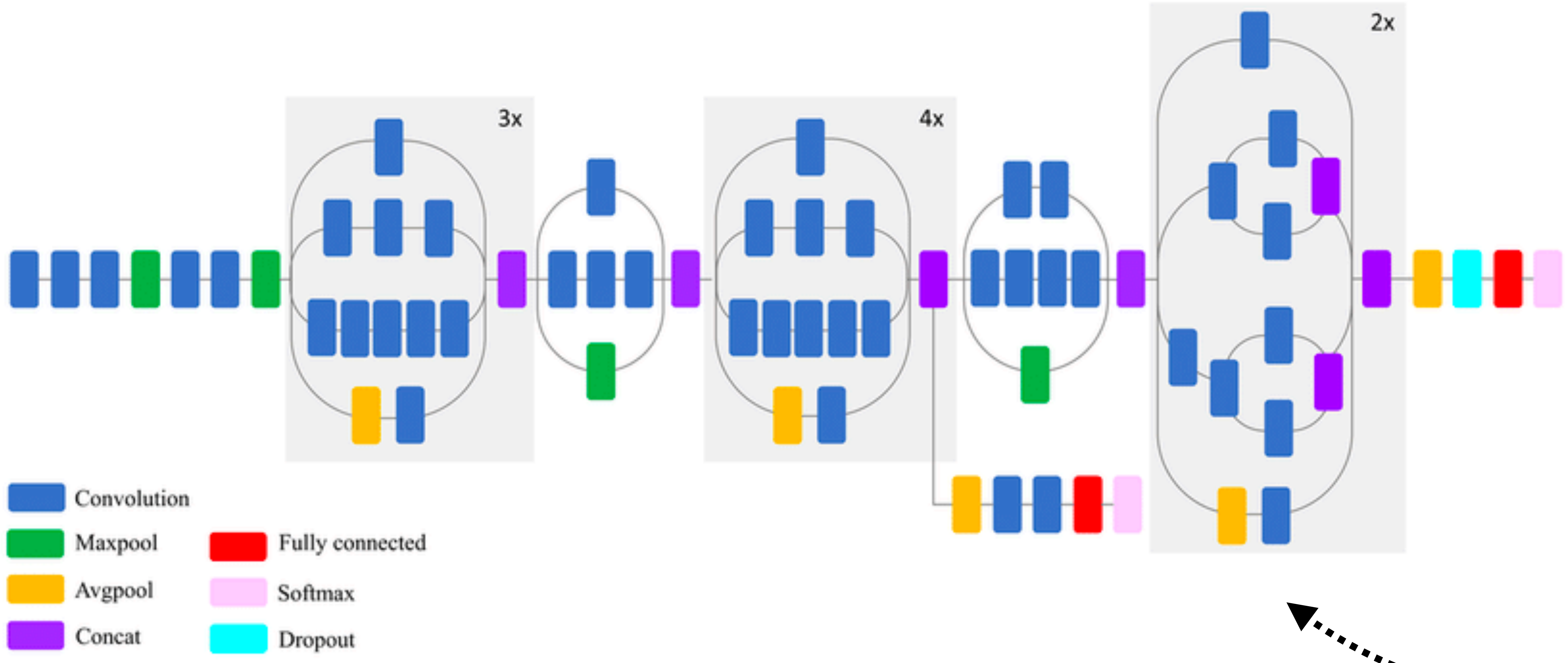
collection,
cleaning,
preprocessing,
featurization,
selection,…

Data

Model selection



Neural architecture search



Transfer / continual / meta learning

(pre-trained) weights, optimisers, …

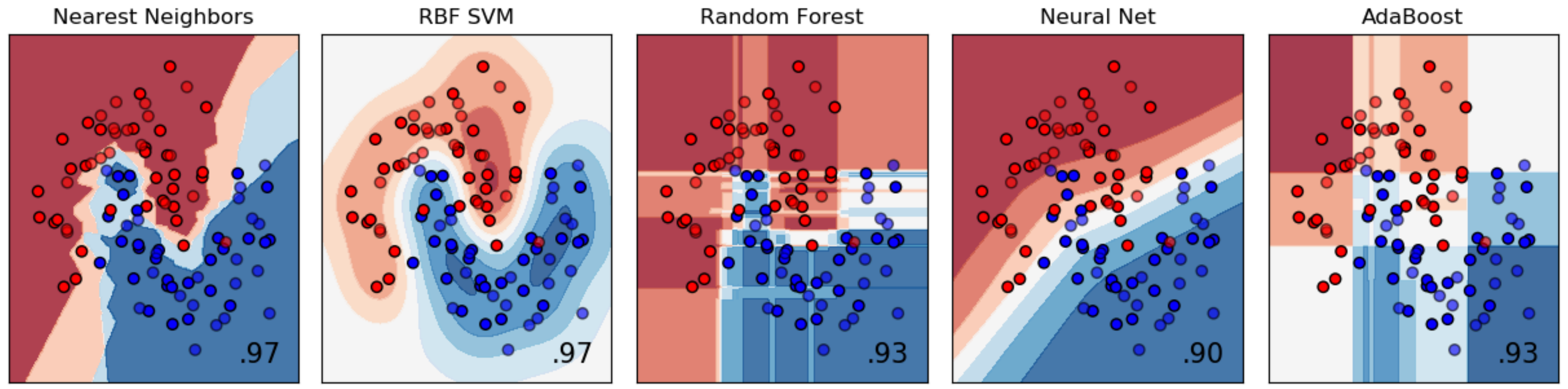
Machine Learning is labor-intensive

Infinite range of possibilities, tacit experience

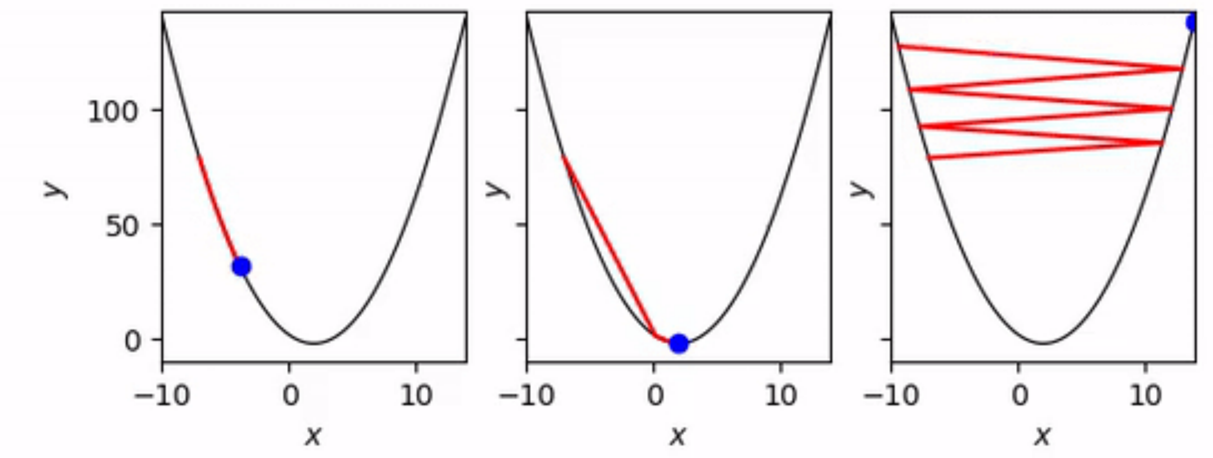
collection, cleaning, preprocessing, featurization, selection,…

Data

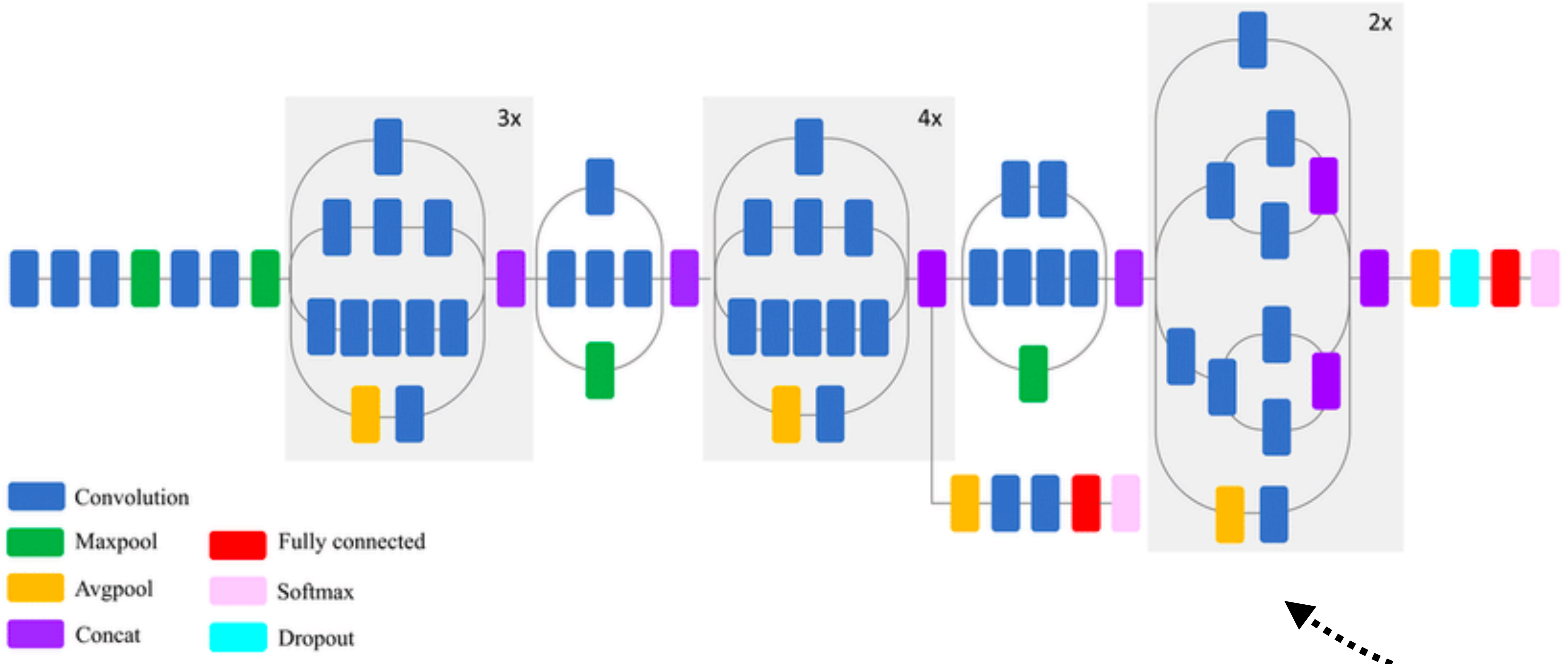
Model selection



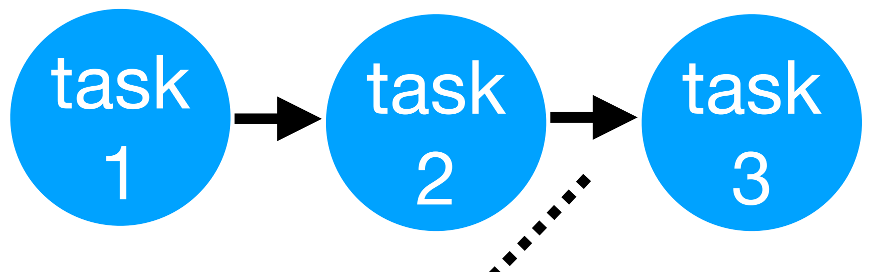
Hyperparameter tuning



Neural architecture search



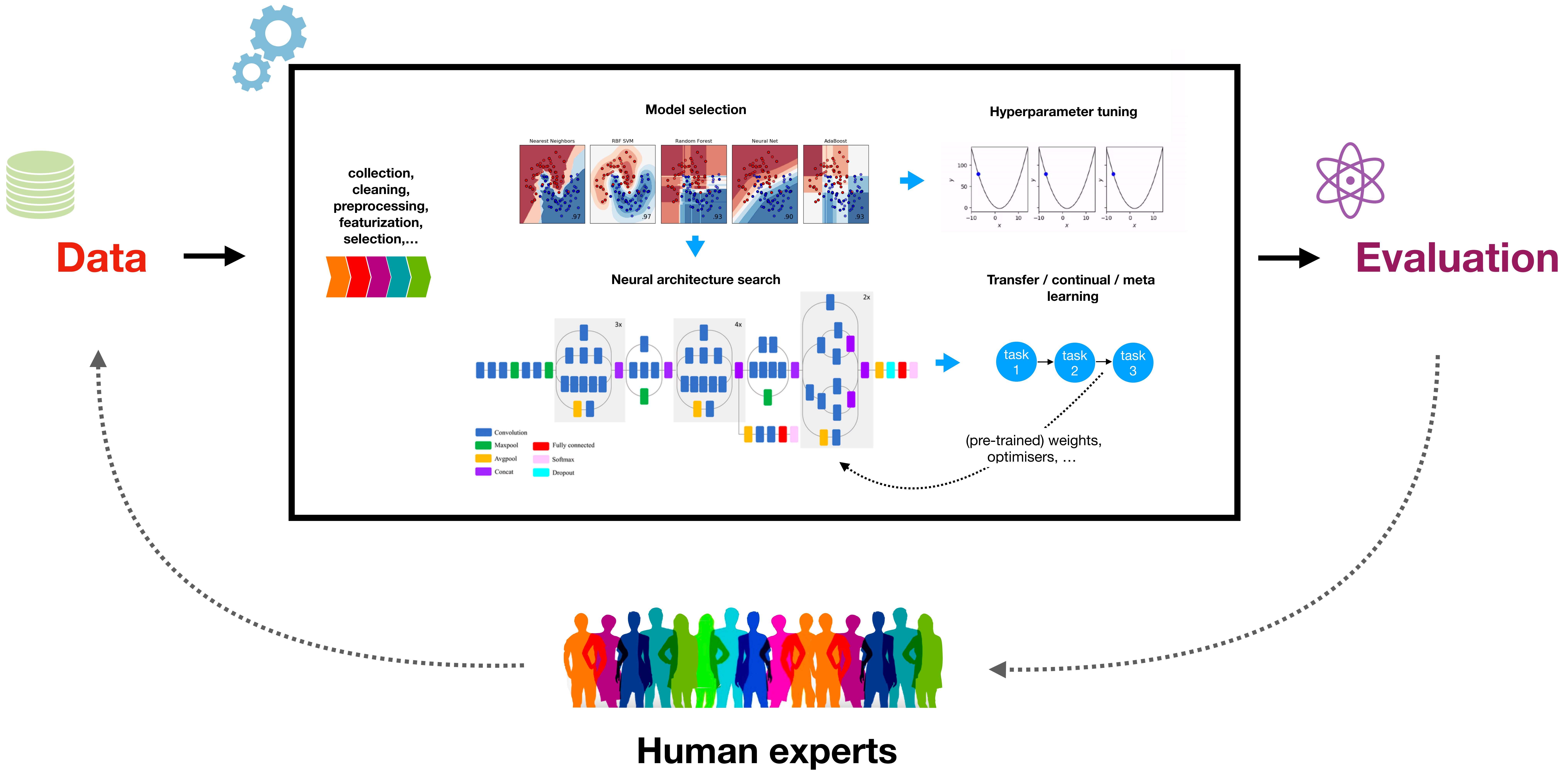
Transfer / continual / meta learning



(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

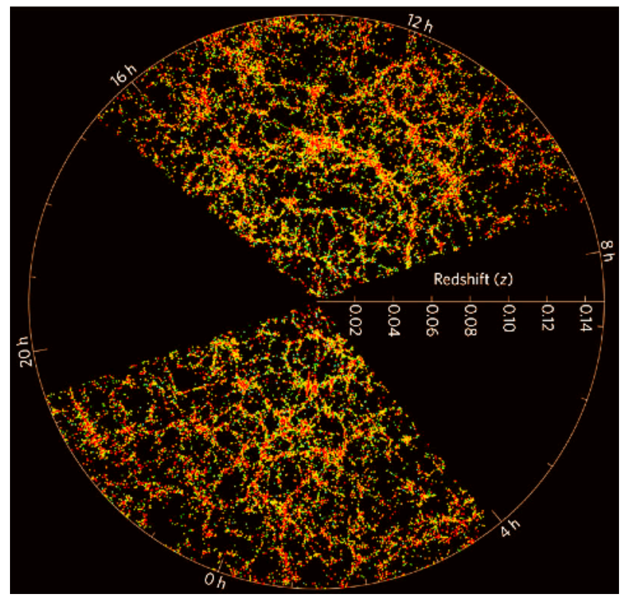
and make it universally accessible and useful?

Democratizing data

mapping the universe

Democratizing data

map of the universe



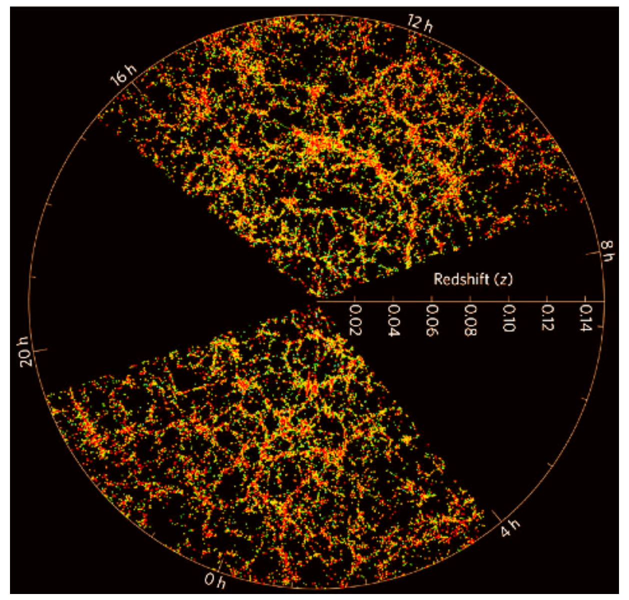
Well-organized data, easily accessible, uniform meta-data

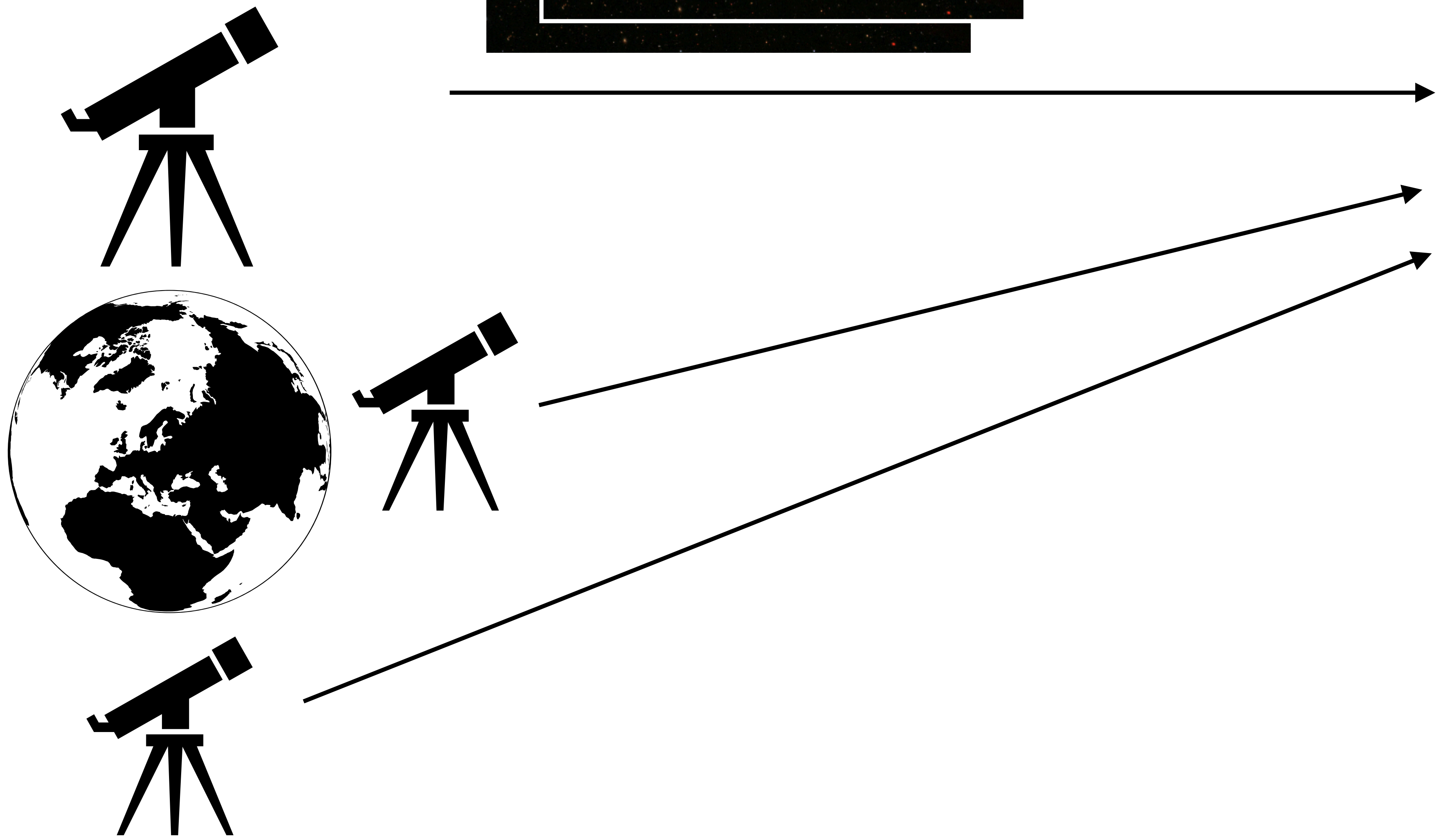
Machine learning models

New discoveries, 1000s of papers

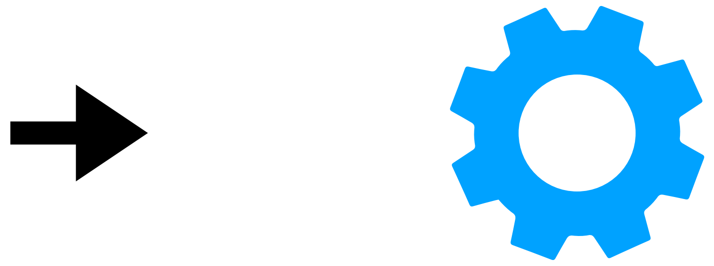
Democratizing data

map of the universe

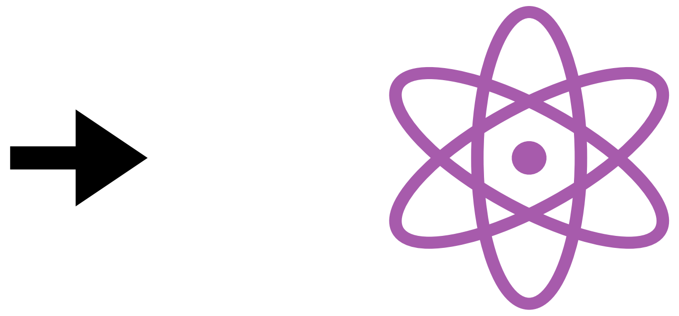




Well-organized data, easily accessible, uniform meta-data



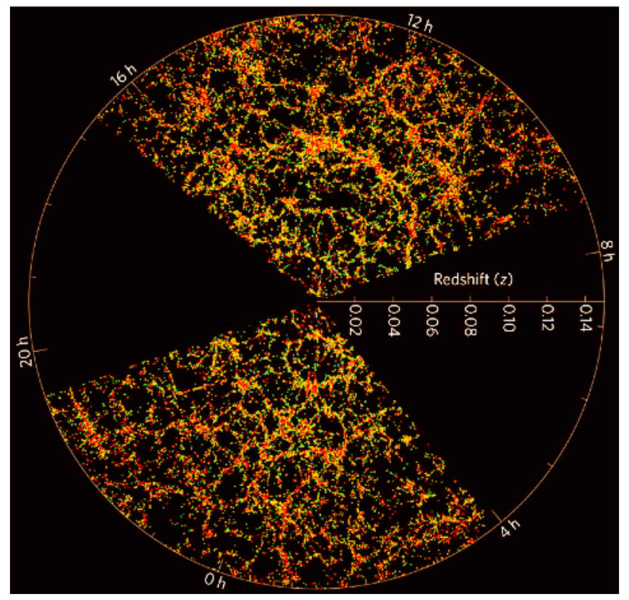
Machine learning models



New discoveries, 1000s of papers

Democratizing data

map of the universe



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

Machine learning models

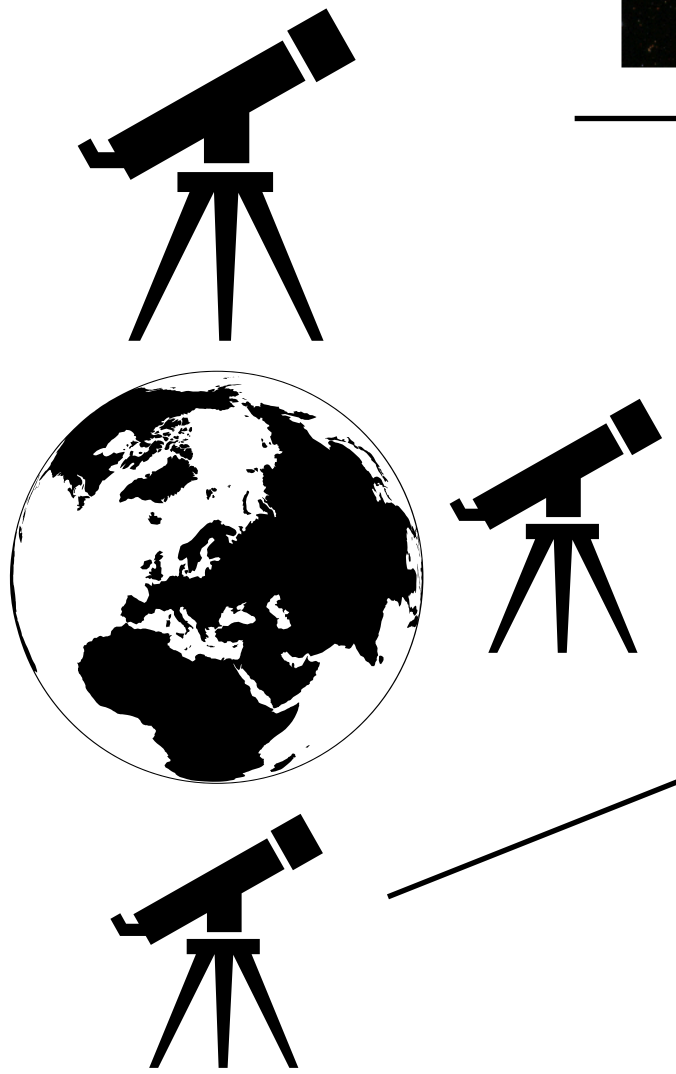
New discoveries, 1000s of papers

Citizen Science projects

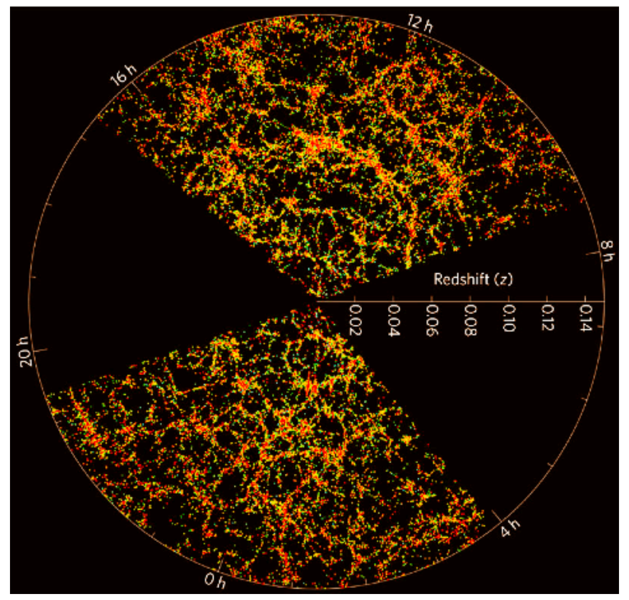
GALAXY ZOO

Democratizing data

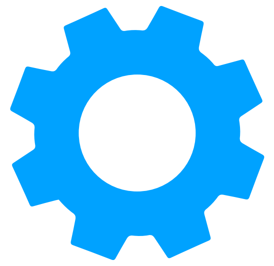
How can we generalize this idea?



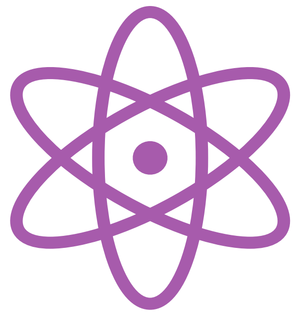
map of the universe



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



Machine learning models

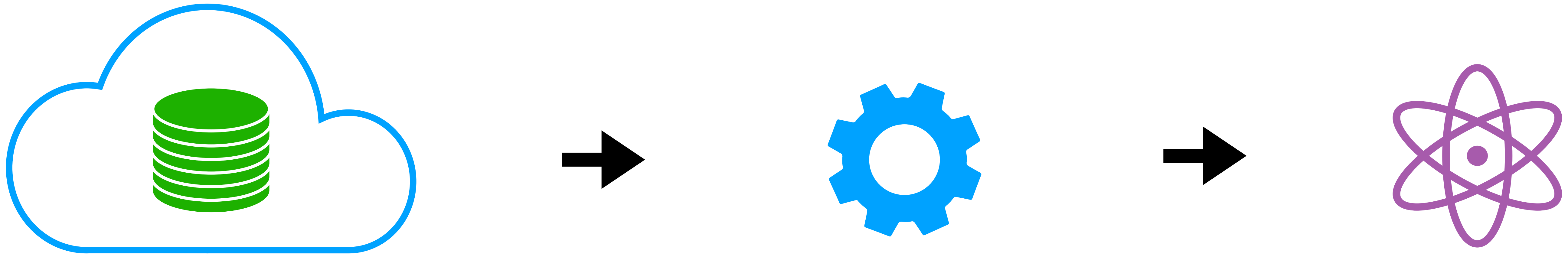


New discoveries, 1000s of papers

Citizen Science projects

GALAXY ZOO

Democratizing ML data



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

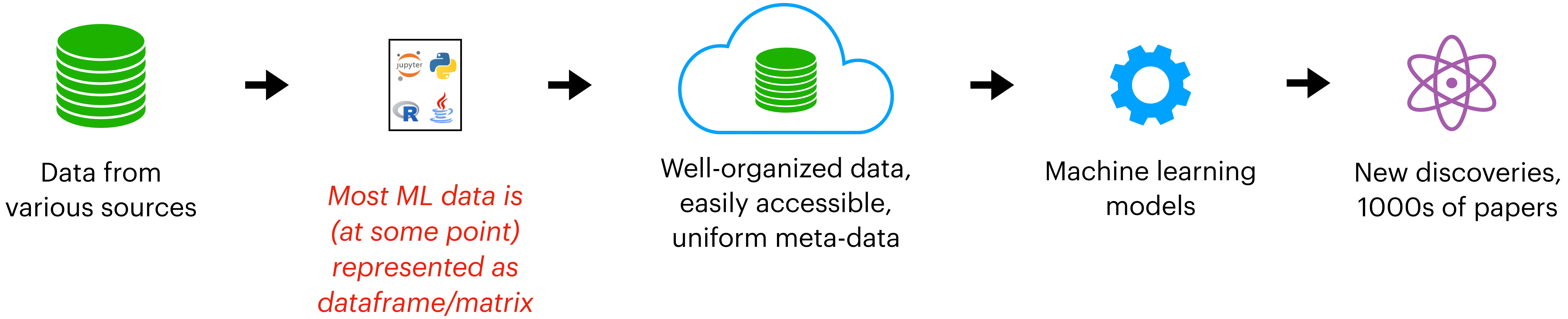
Machine learning models

(Assuming we have these)

New discoveries, 1000s of papers

Democratizing ML data

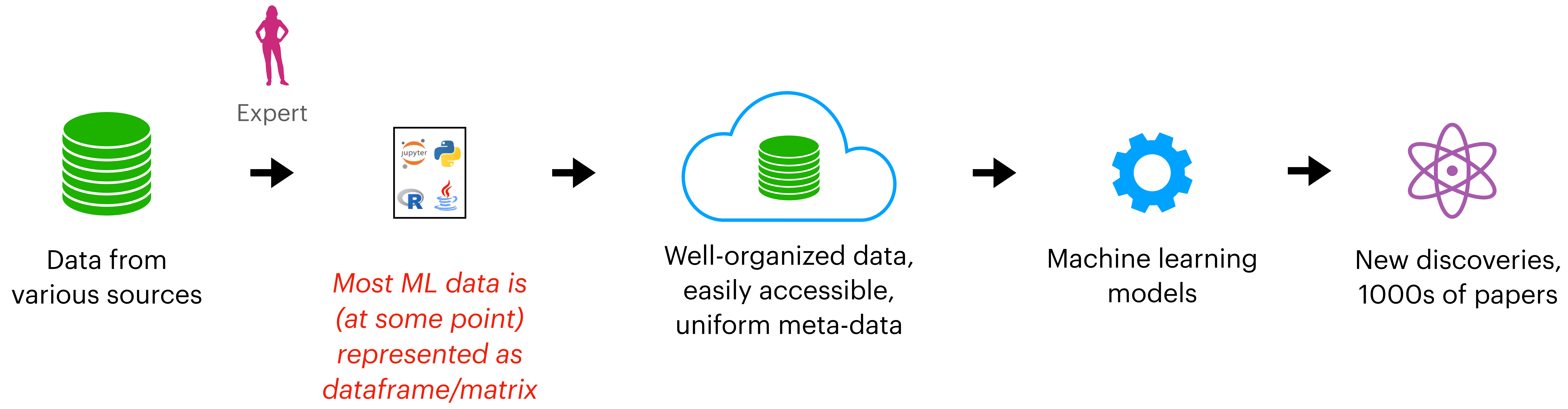
Democratizing ML data



(Required anyway for many ML models)

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

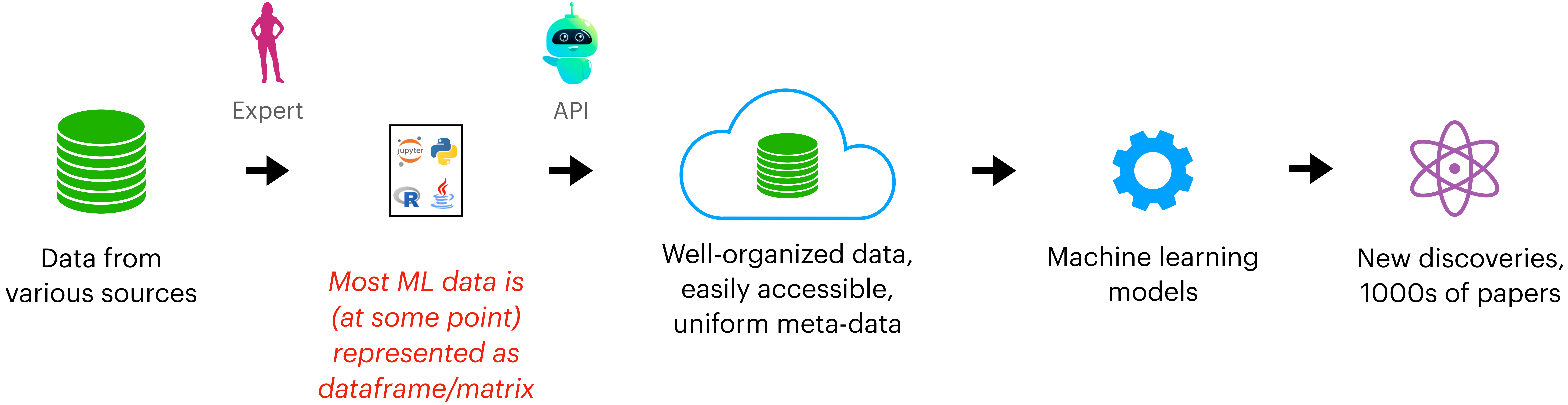
Democratizing ML data



(Required anyway for many ML models)

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

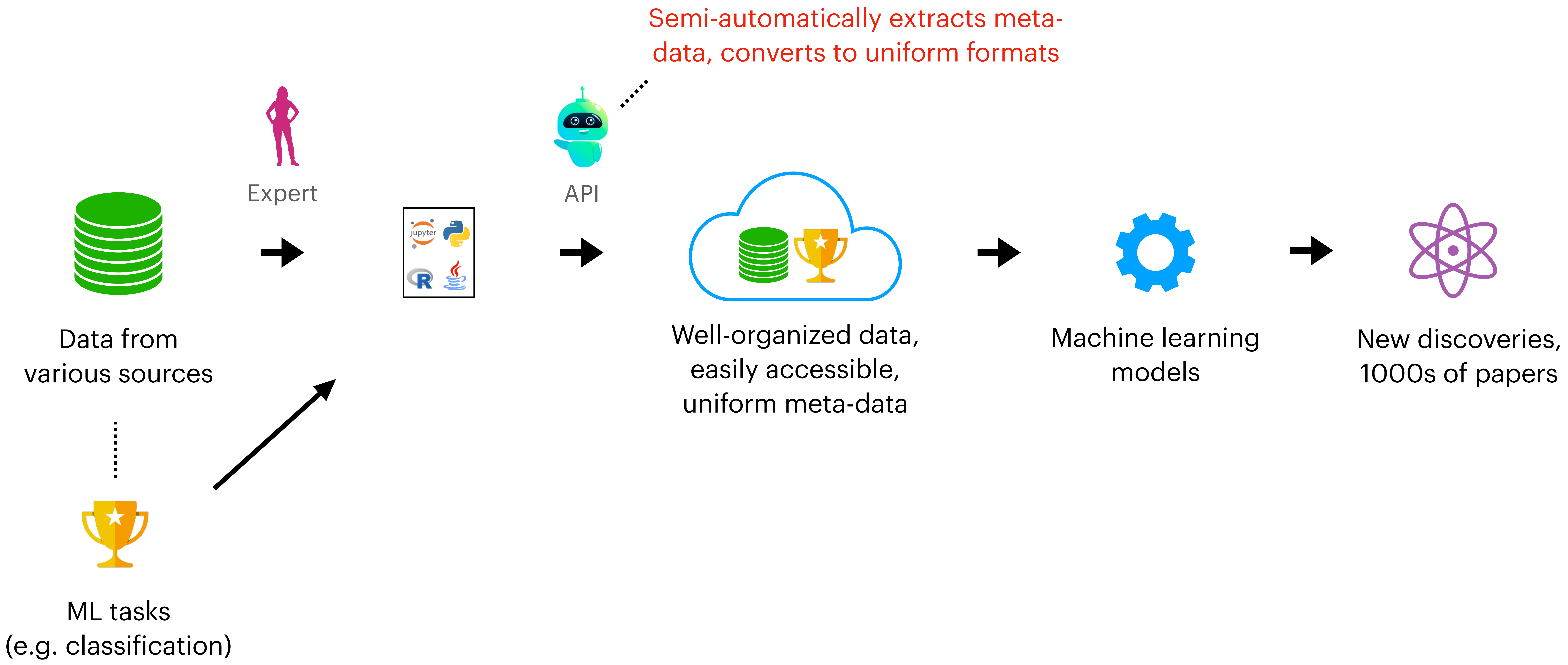
Democratizing ML data



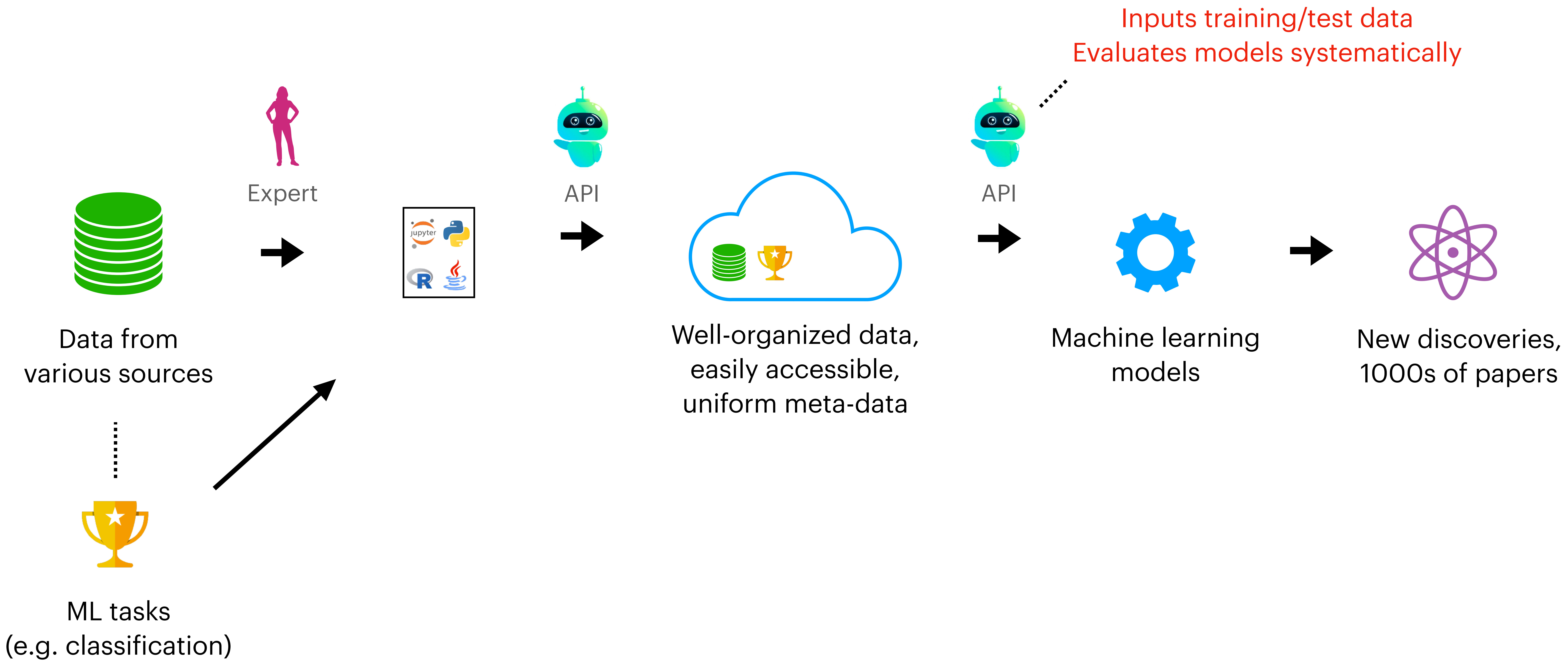
(Required anyway for many ML models)

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

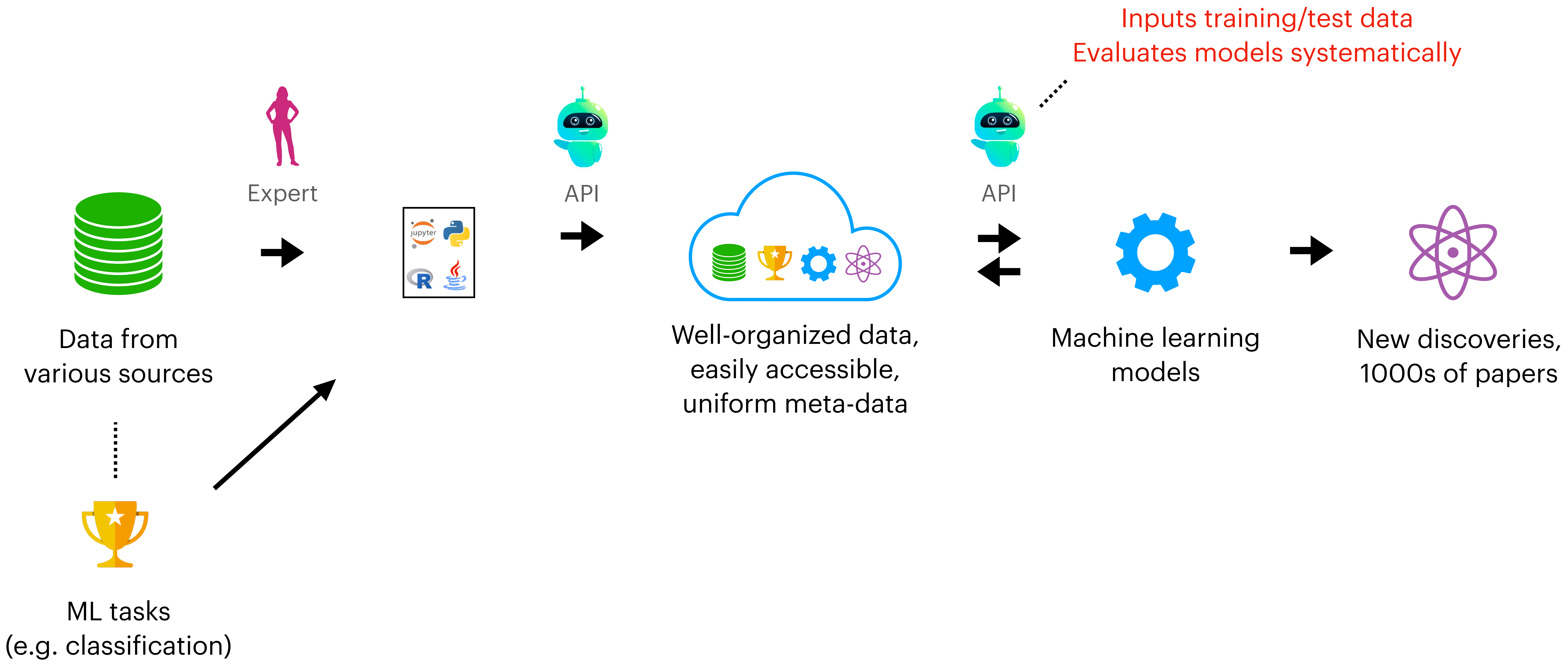
Democratizing ML data



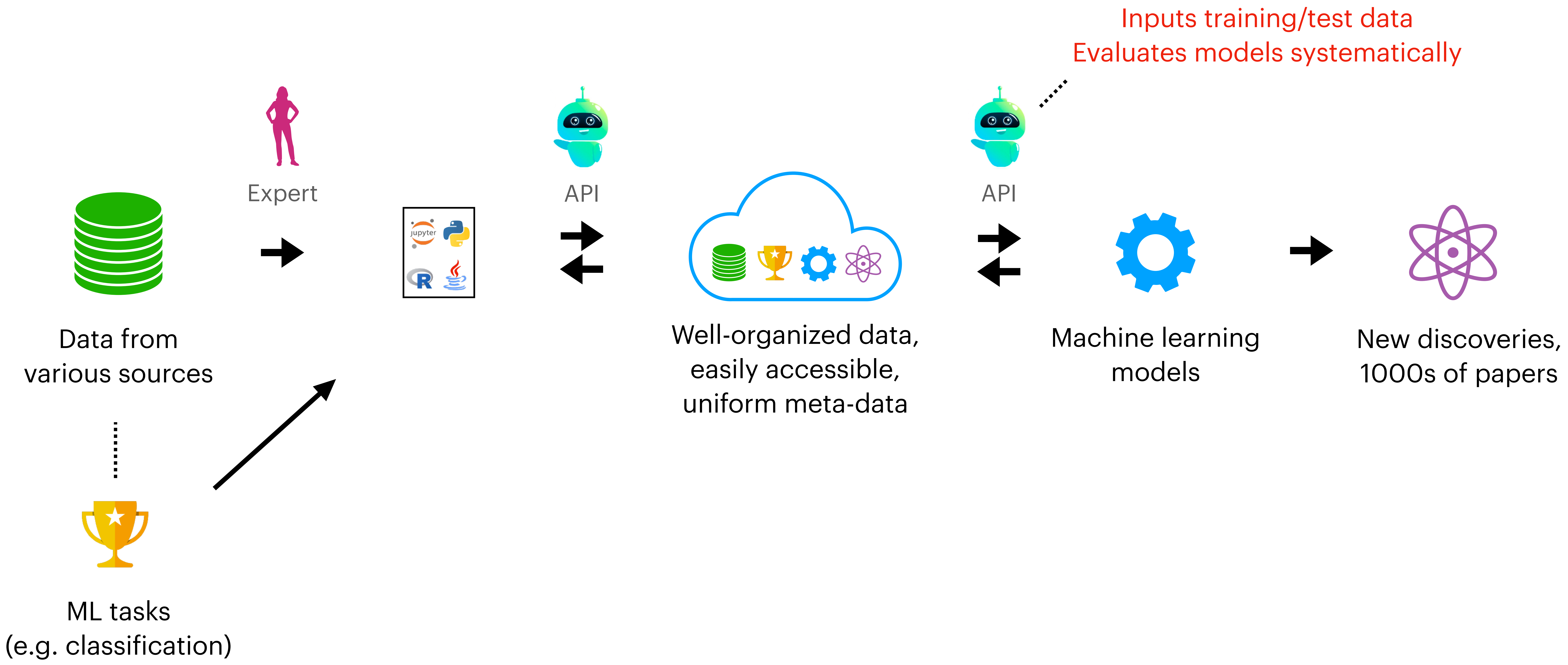
Democratizing ML data



Democratizing ML data

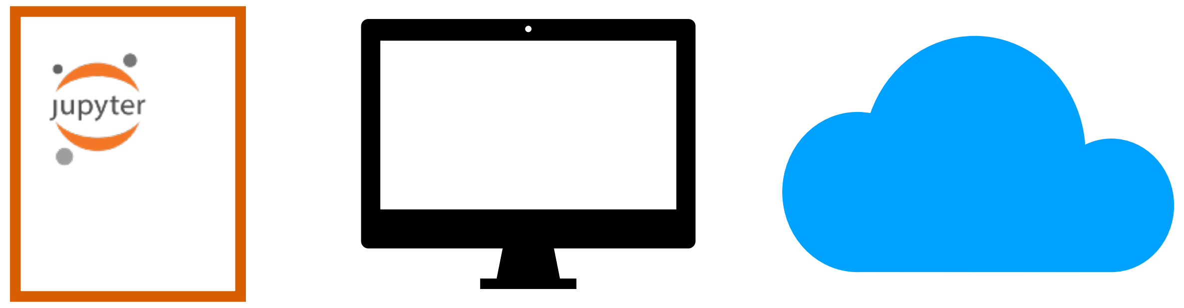


Democratizing ML data



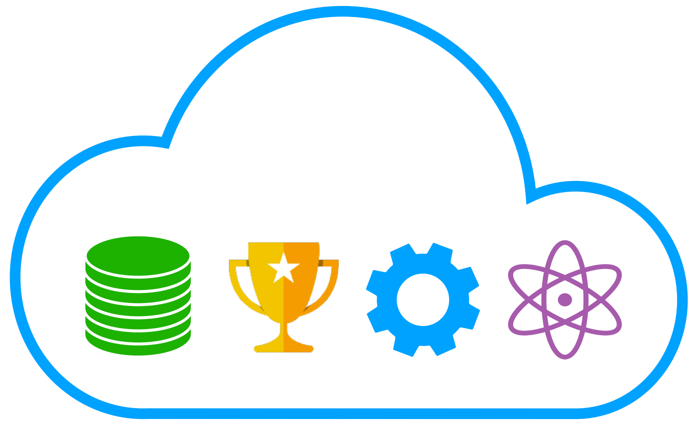
OpenML

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



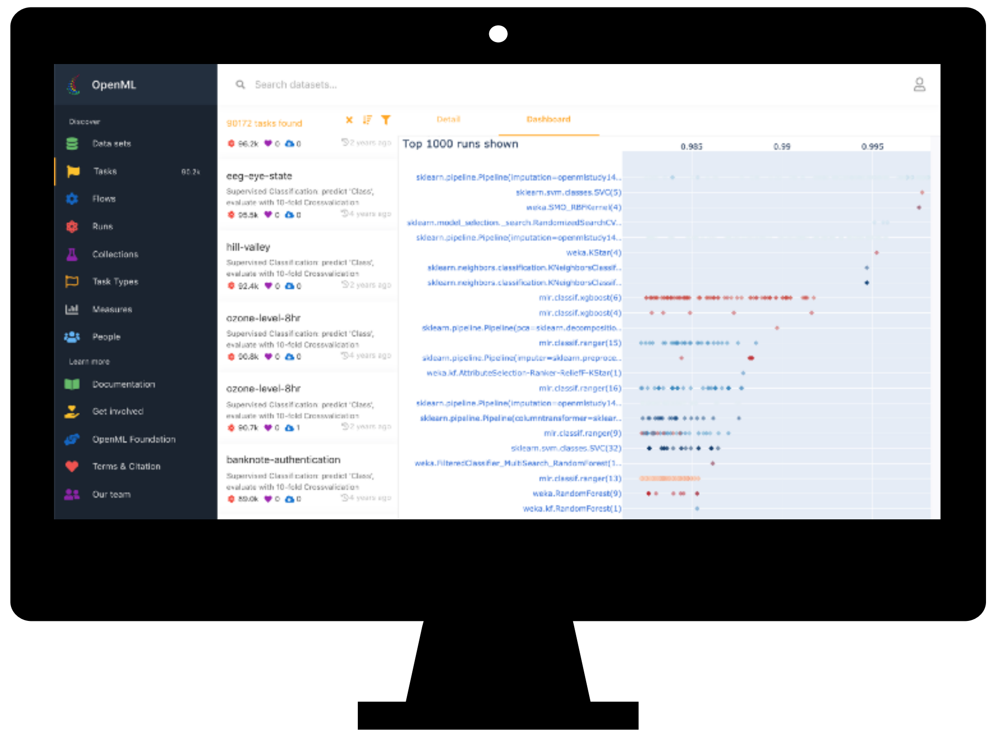
Accessible from anywhere, anytime
(scripts, notebooks, apps, cloud jobs)

API



OpenML

API



Website
(new.openml.org)

OpenML web interface

Search

Datasets

Dataset analysis

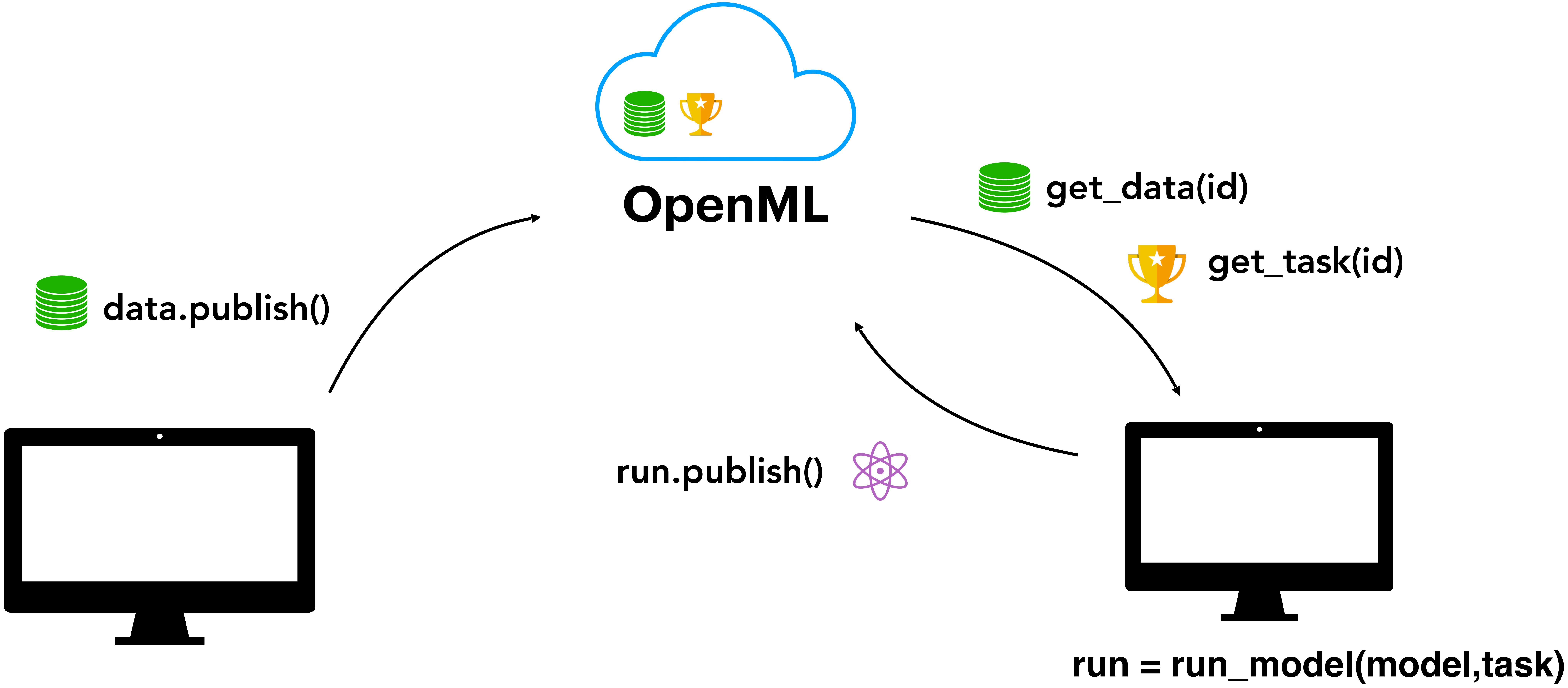
OpenML web interface

Tasks

Algorithms

Evaluations (every dot is a model)

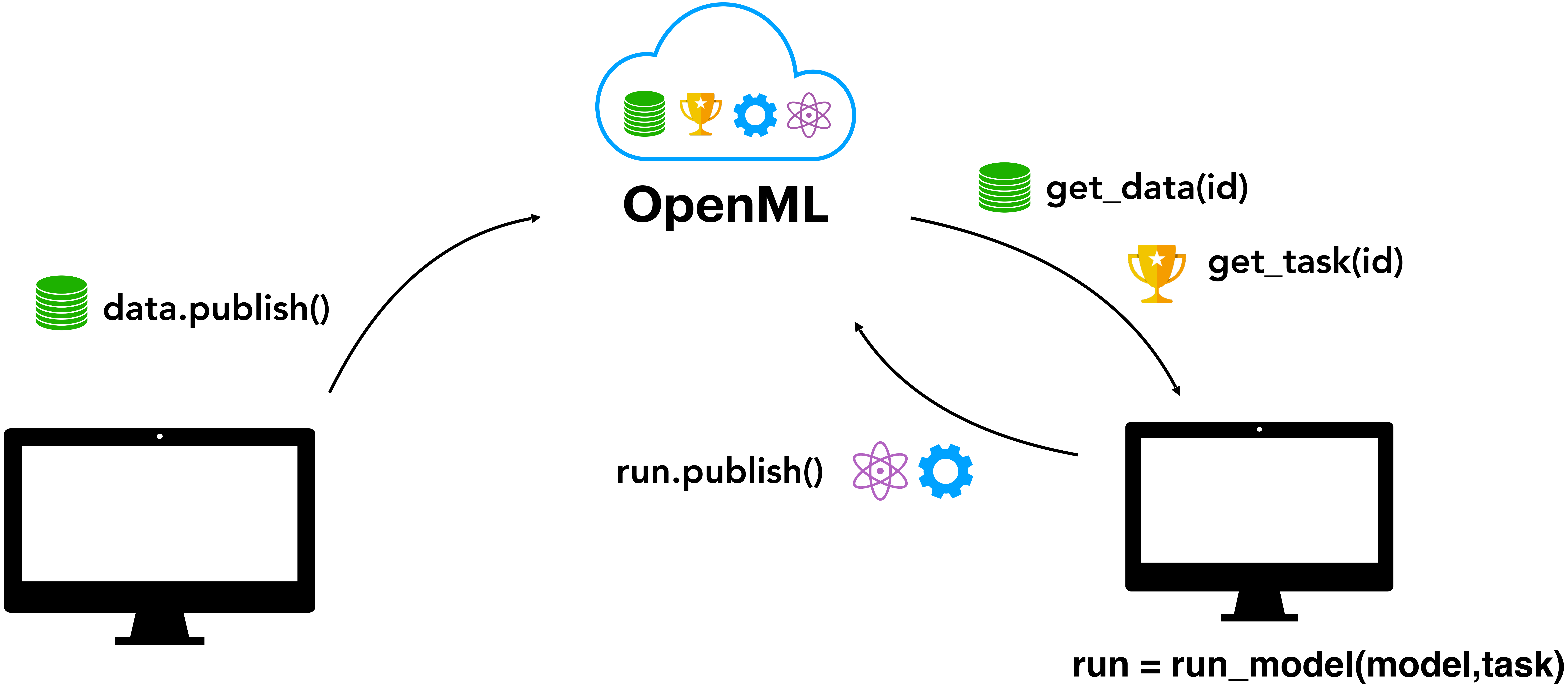
Frictionless machine learning



APIs:

Integrations:

Frictionless machine learning



APIs:

Integrations:

Frictionless machine learning

```python
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

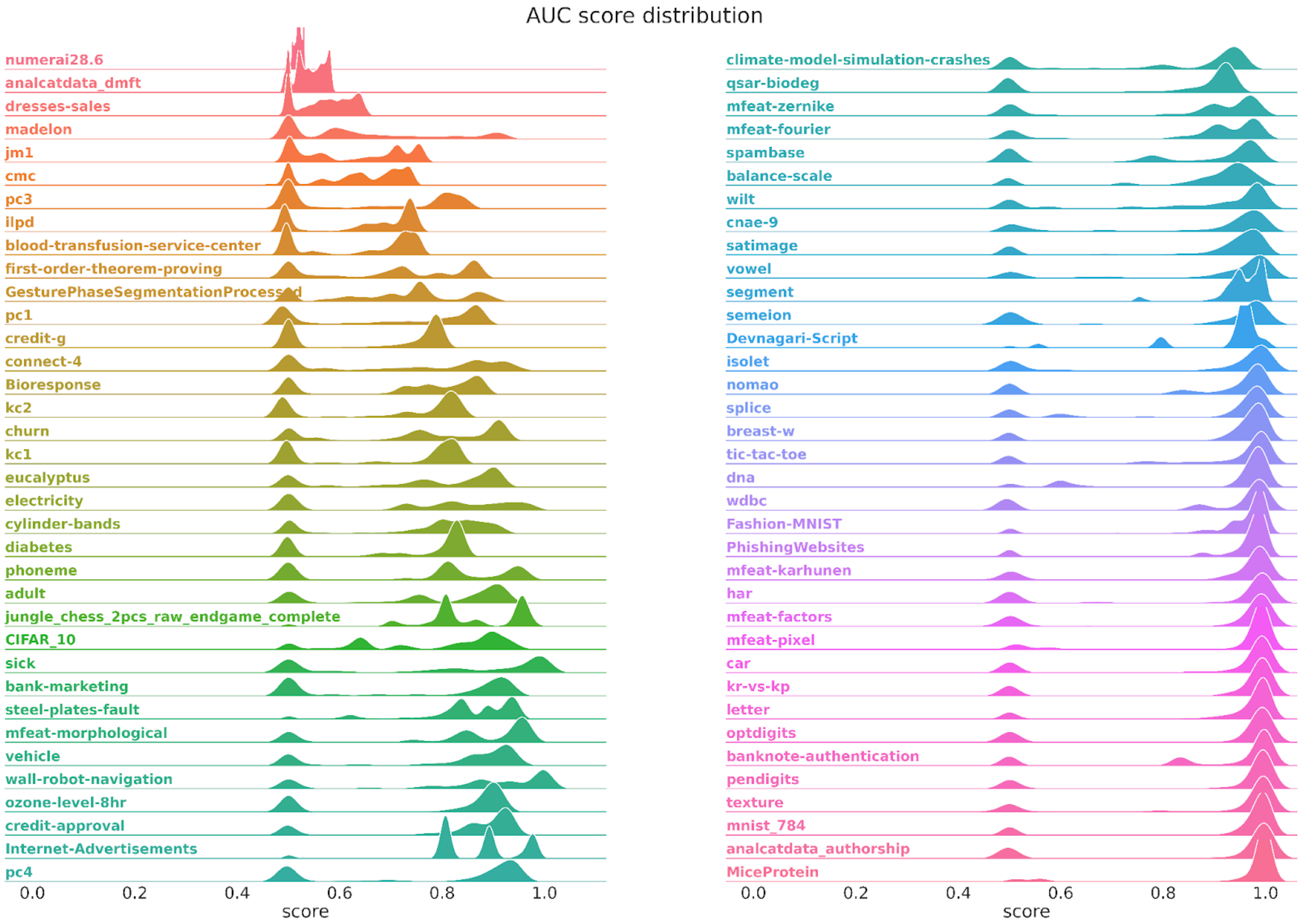
Benchmarking suites

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

Benchmarking suites

- Example: OpenML-CC18
  - 3.8 million results

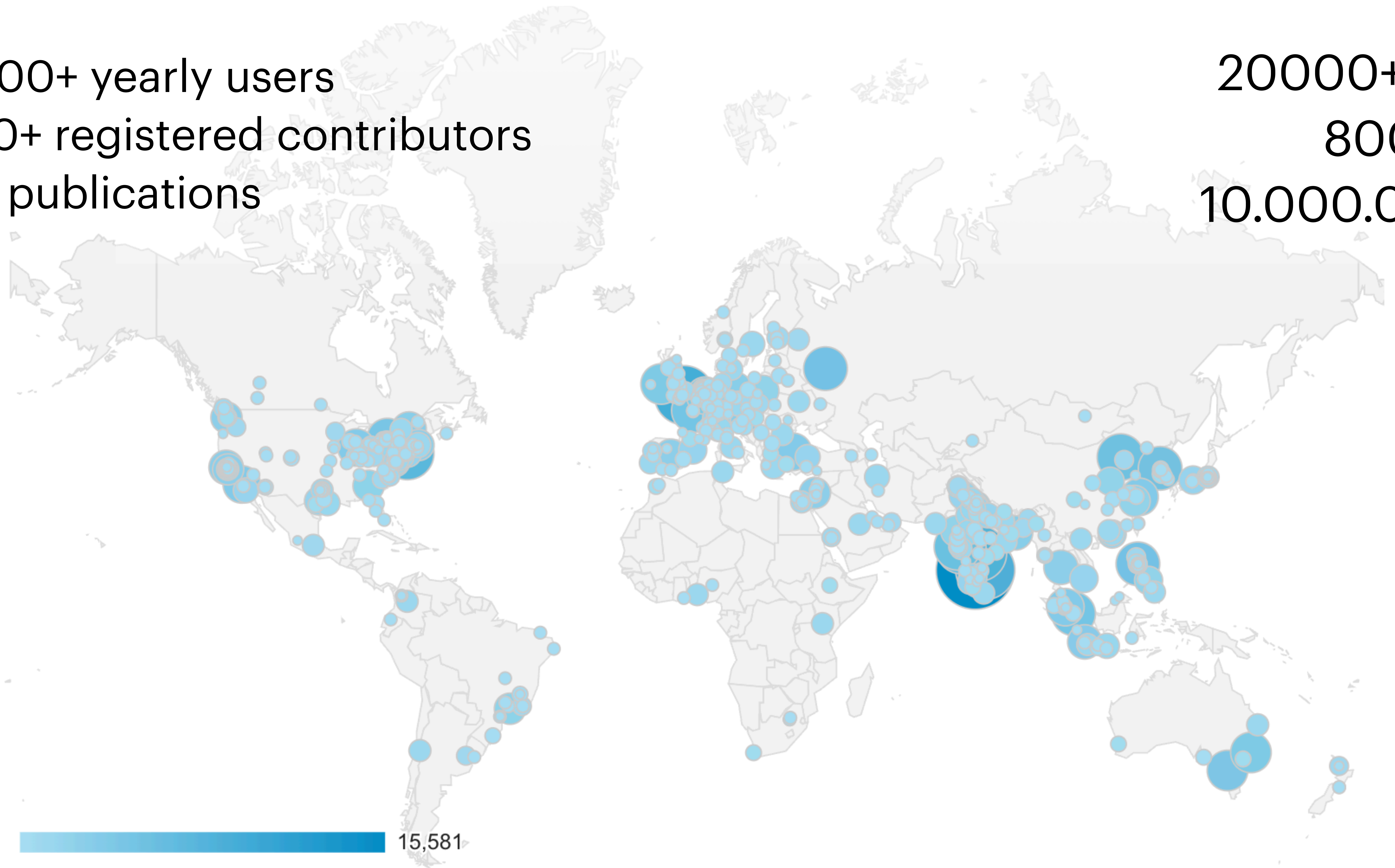
- Classification only
- 72 datasets
- Contain missing values and categorical features
- Medium-sized (500-100000 observations, <5000 features after one-hot-encoding)
- Not unbalanced
- No groups/block/time dependencies
- No sparse data
- Some more subjective criteria (see paper)



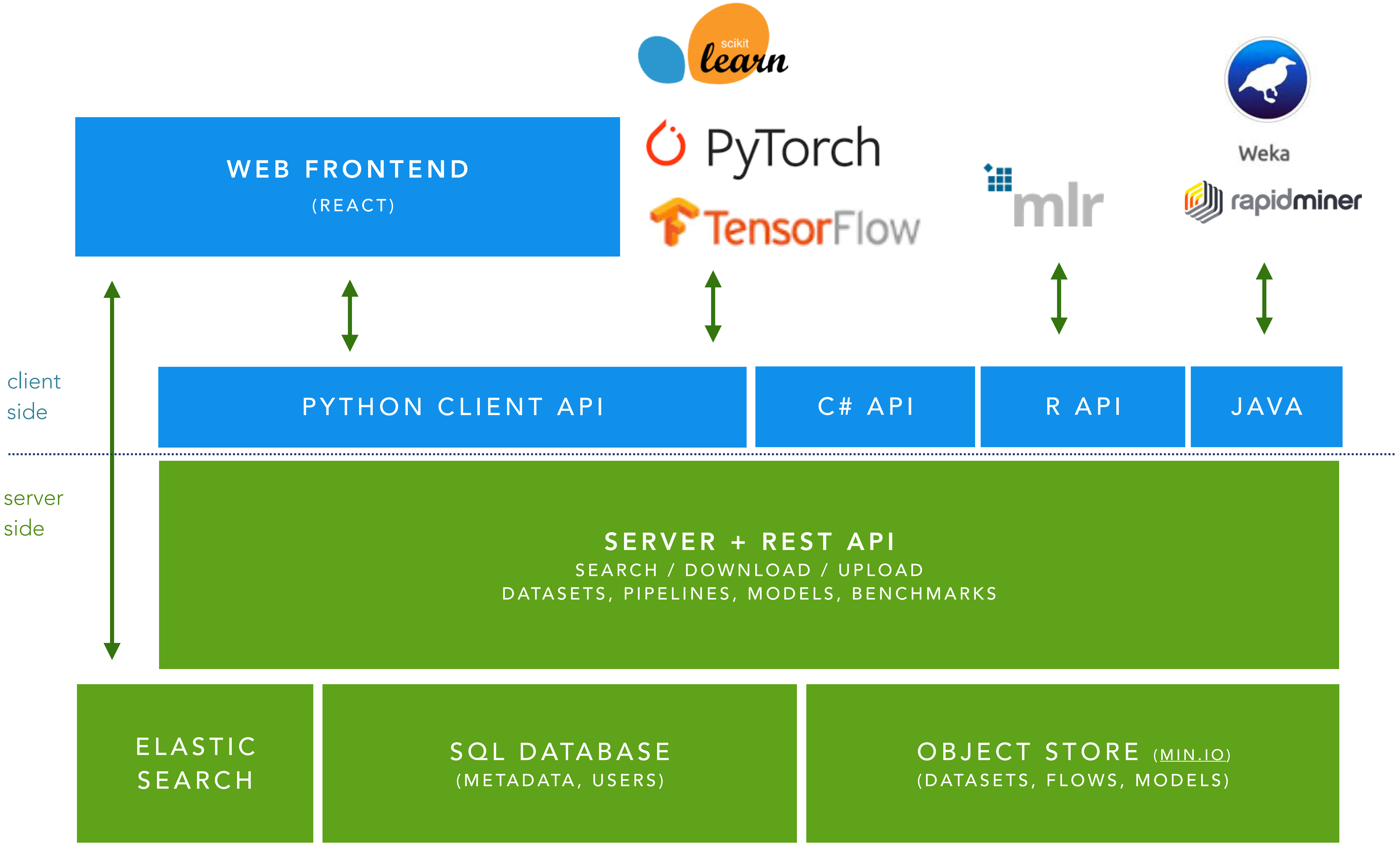
OpenML Community

250000+ yearly users
13000+ registered contributors
900+ publications

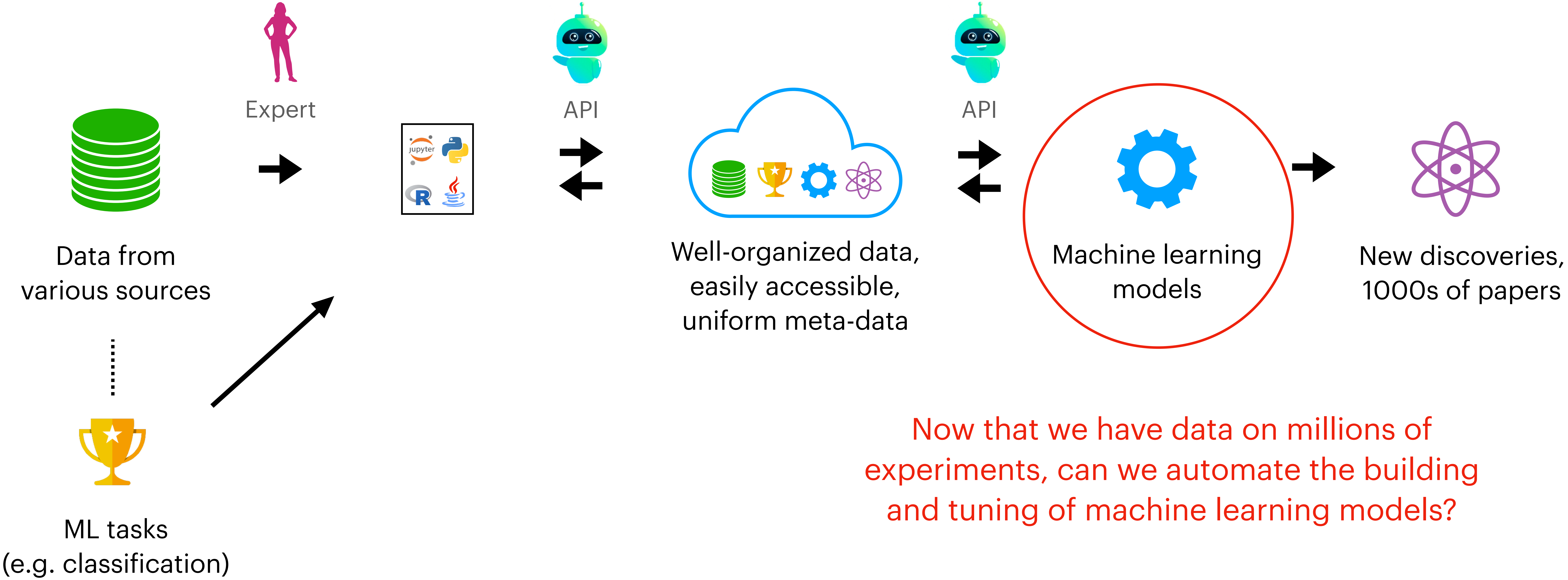
20000+ datasets
8000+ flows
10.000.000+ runs



OpenML Architecture



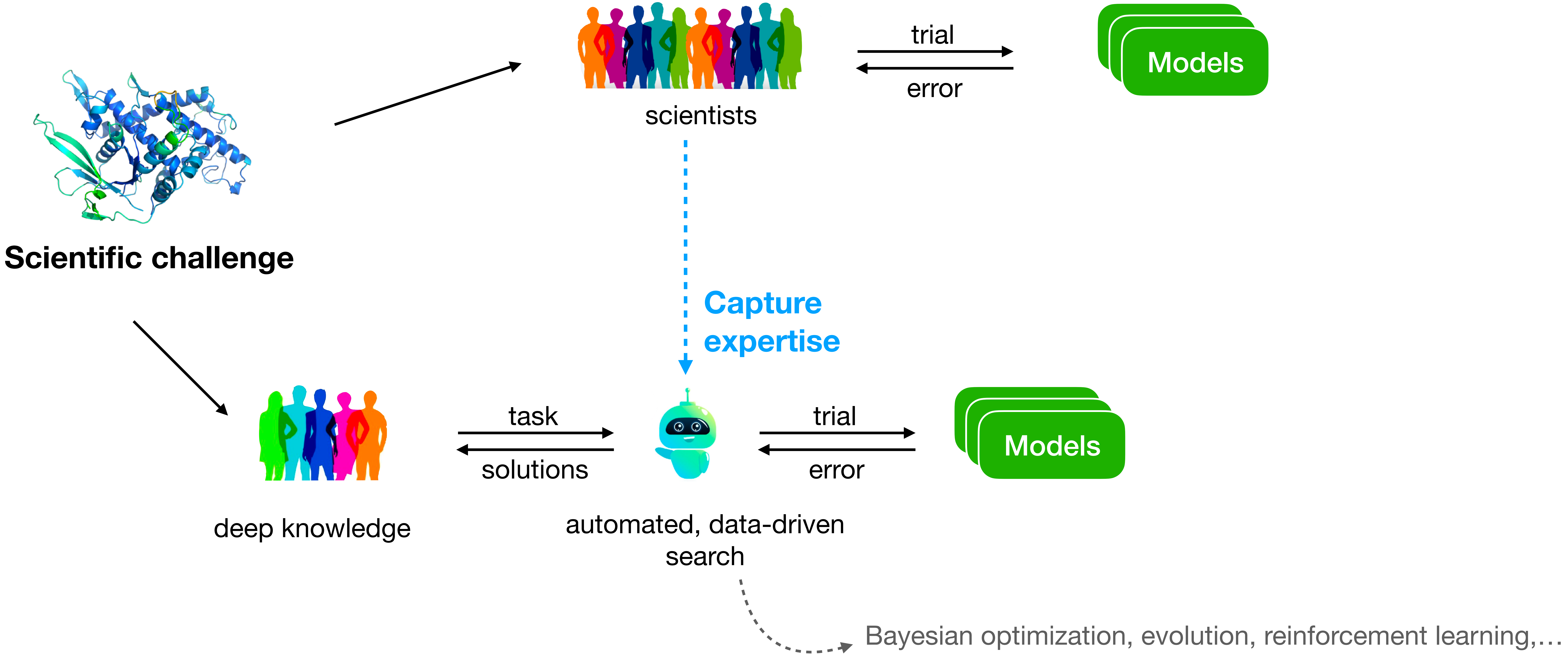
Democratizing machine learning itself



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

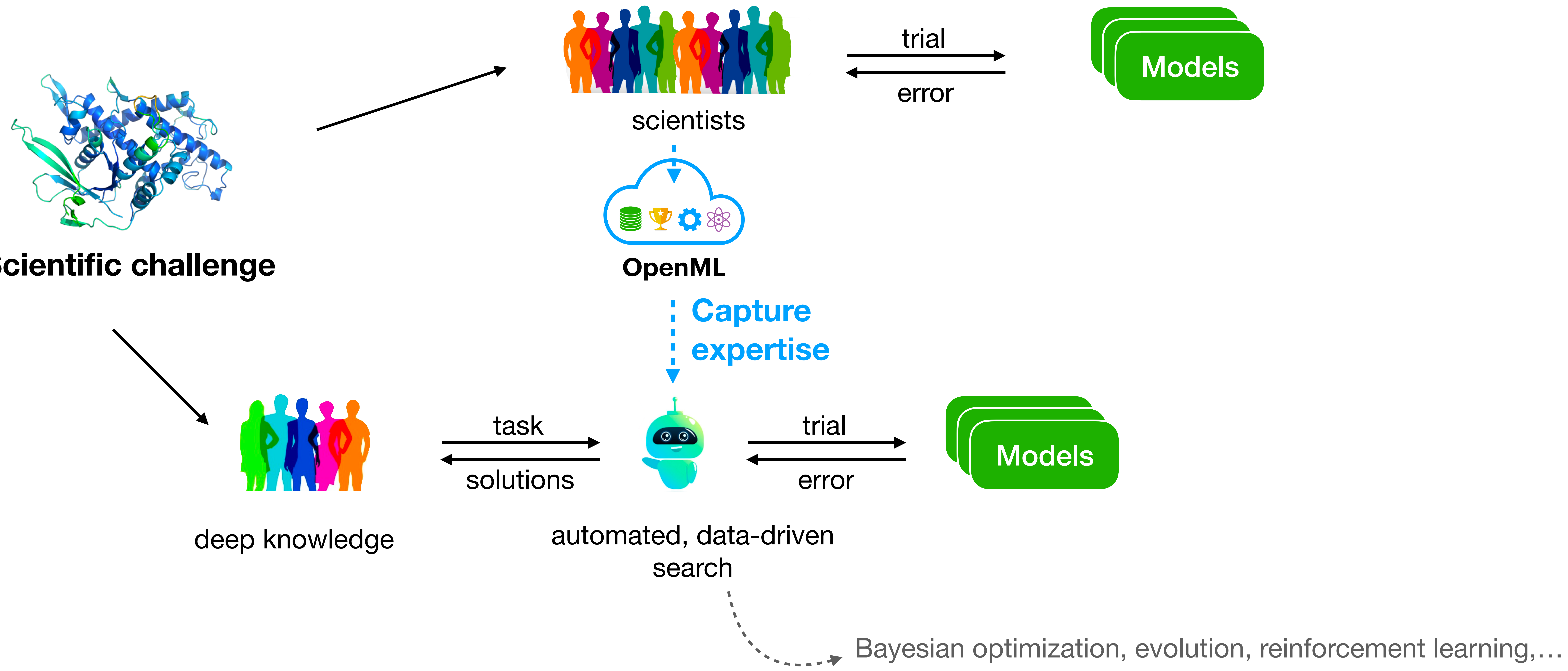
Automatic Machine Learning (AutoML)

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



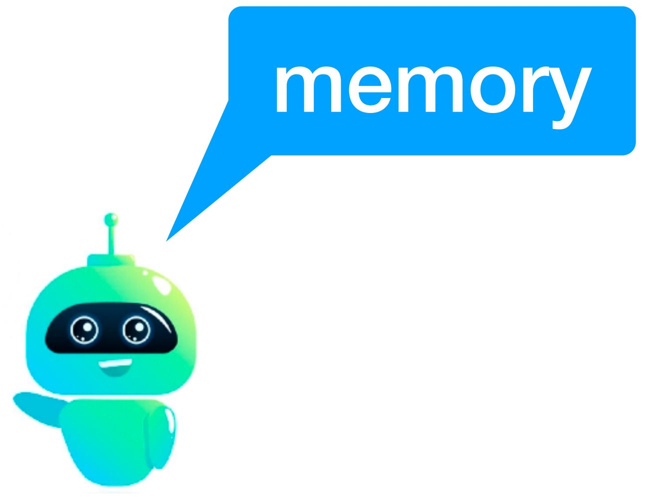
Automatic Machine Learning (AutoML)

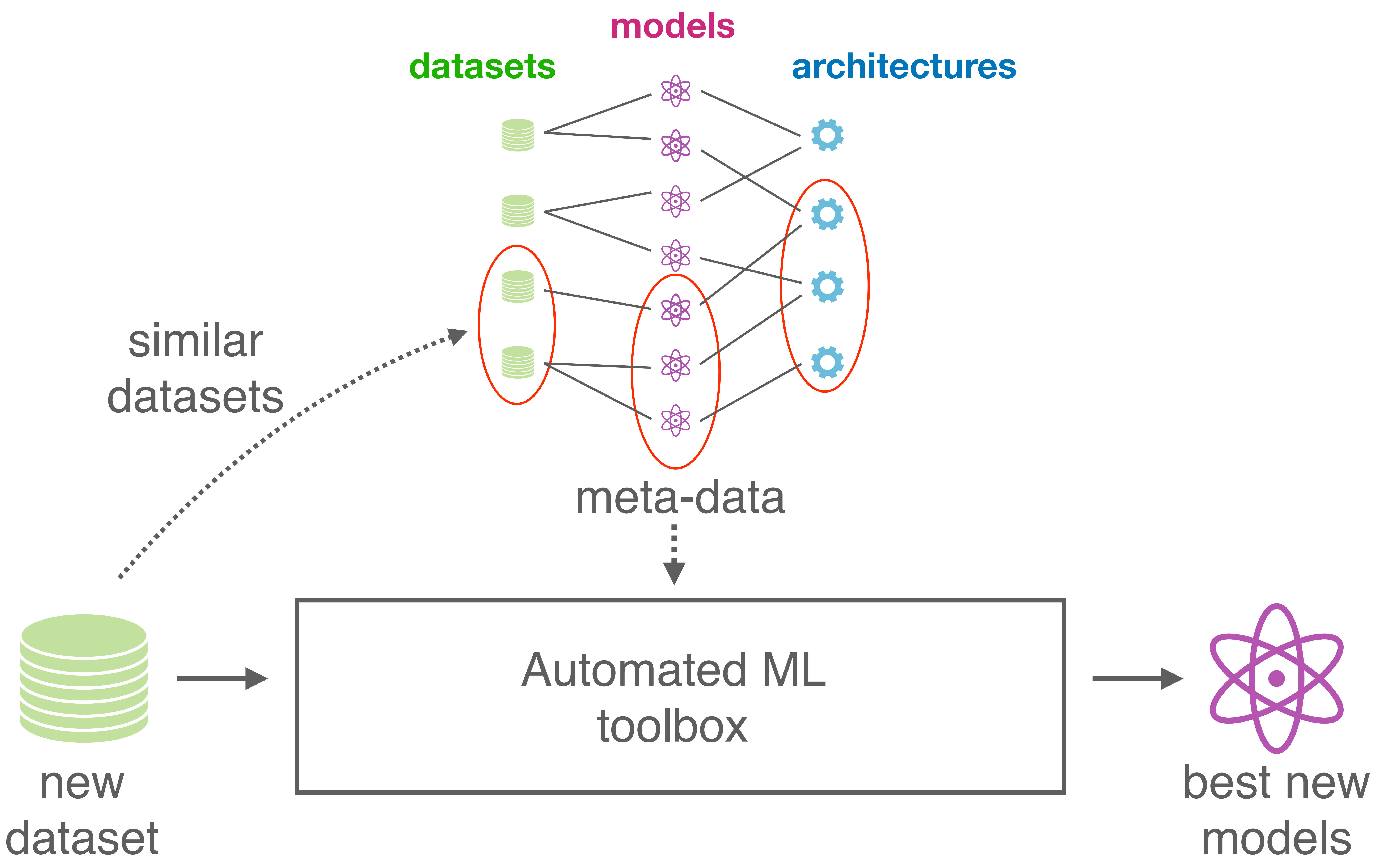
Replace manual trial and error with automated search (based on prior experience)



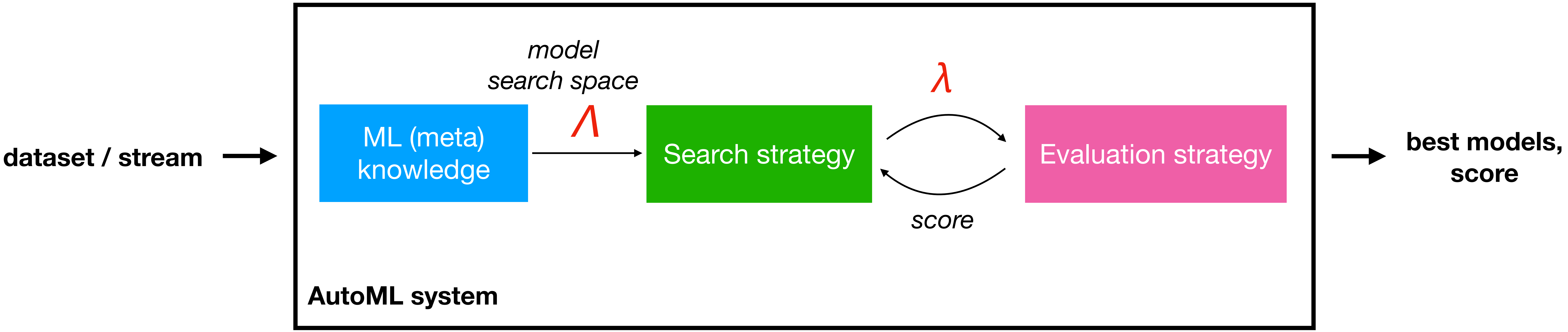
OpenML as a global memory

Machine-readable repository of machine learning results

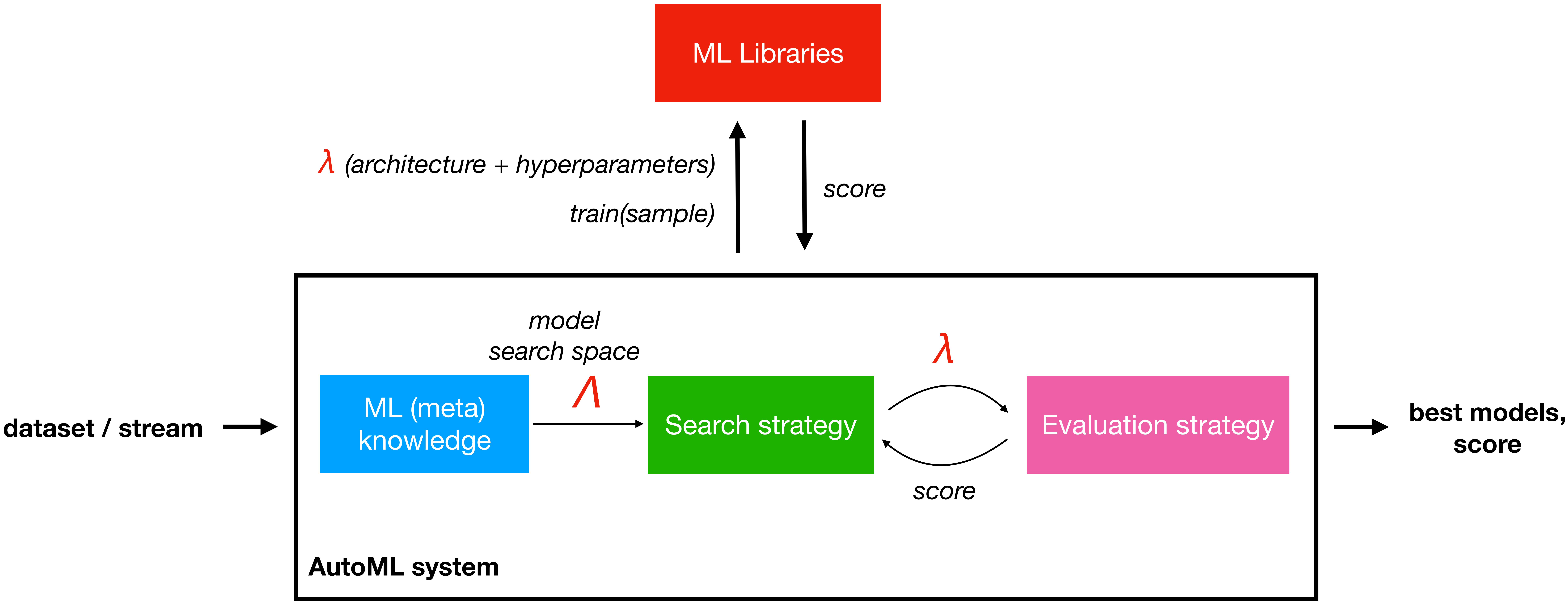




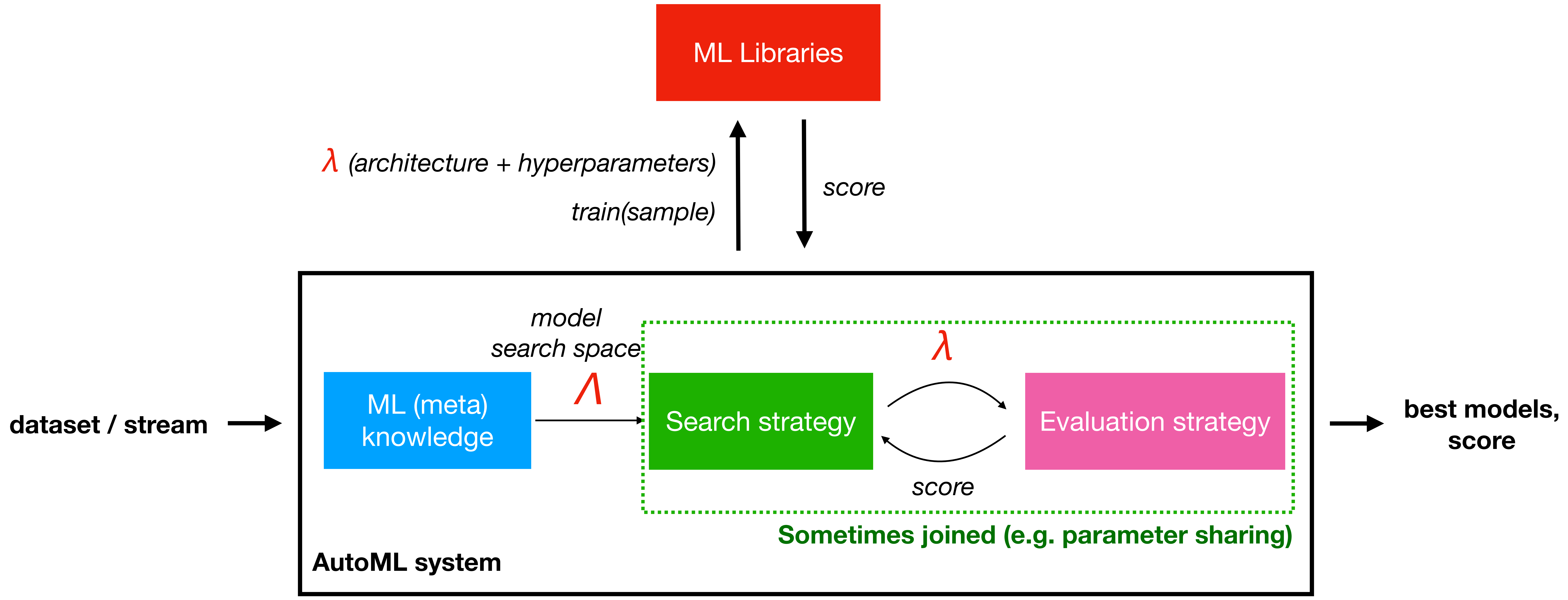
Structure of AutoML systems



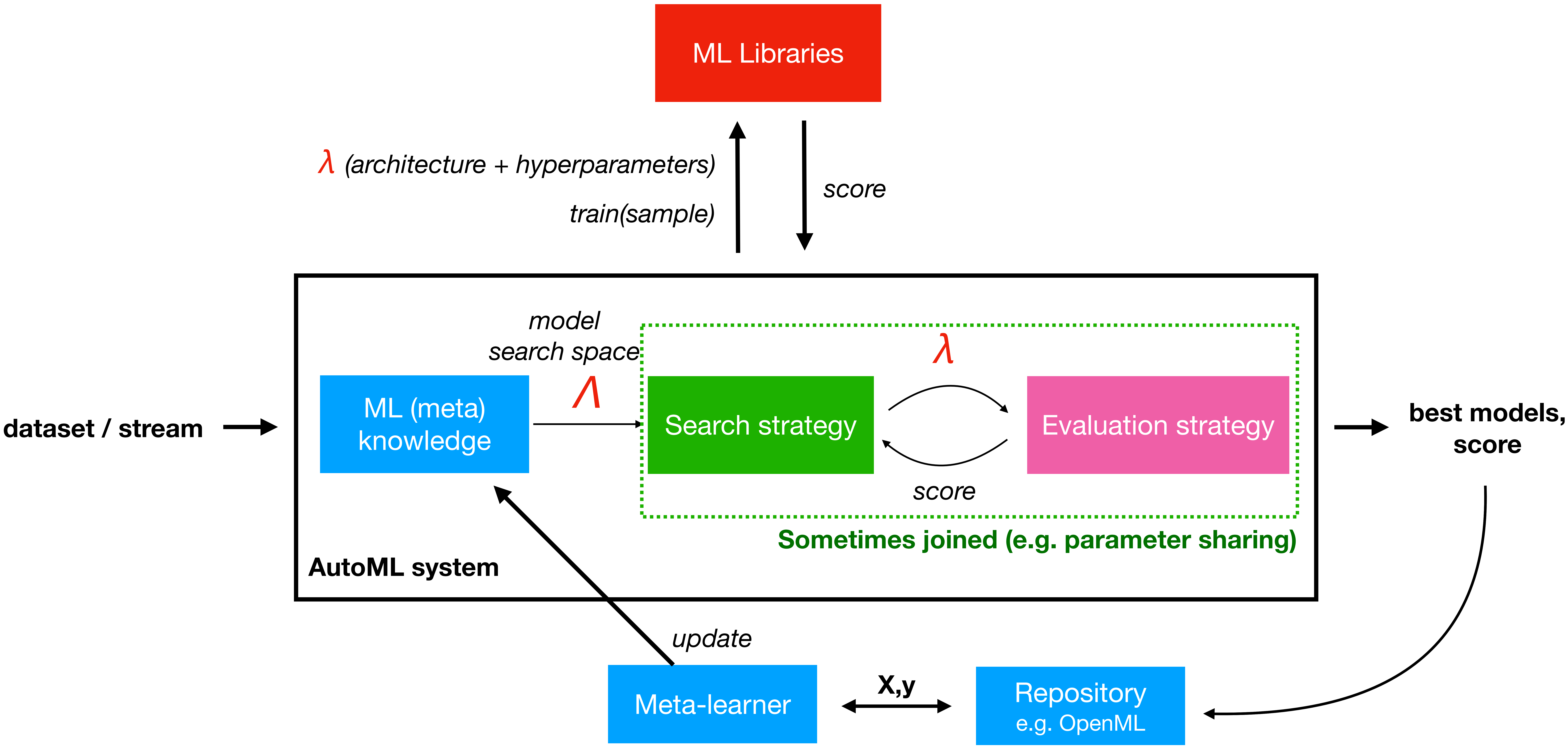
Structure of AutoML systems



Structure of AutoML systems

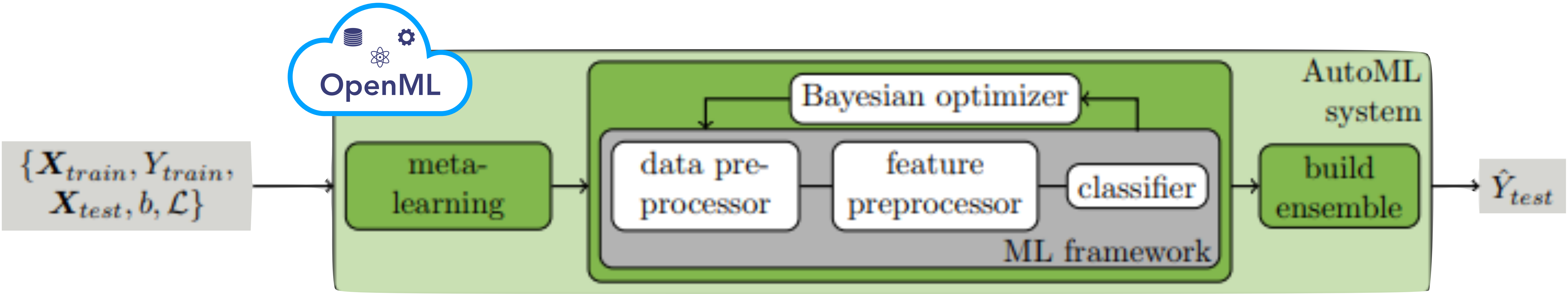


Structure of learning AutoML systems



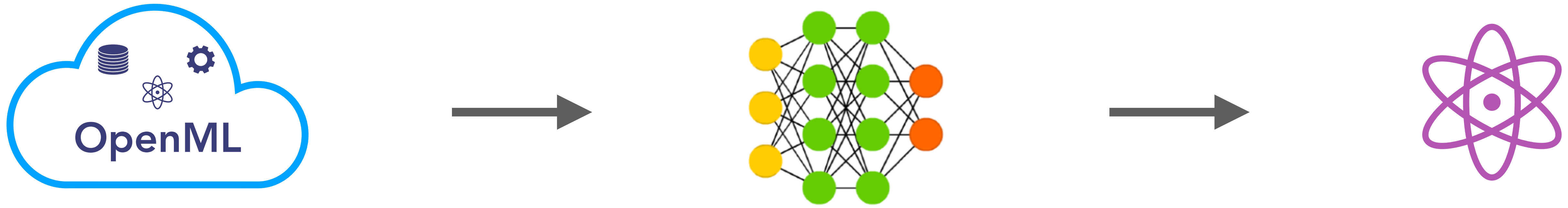
Automating machine learning

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

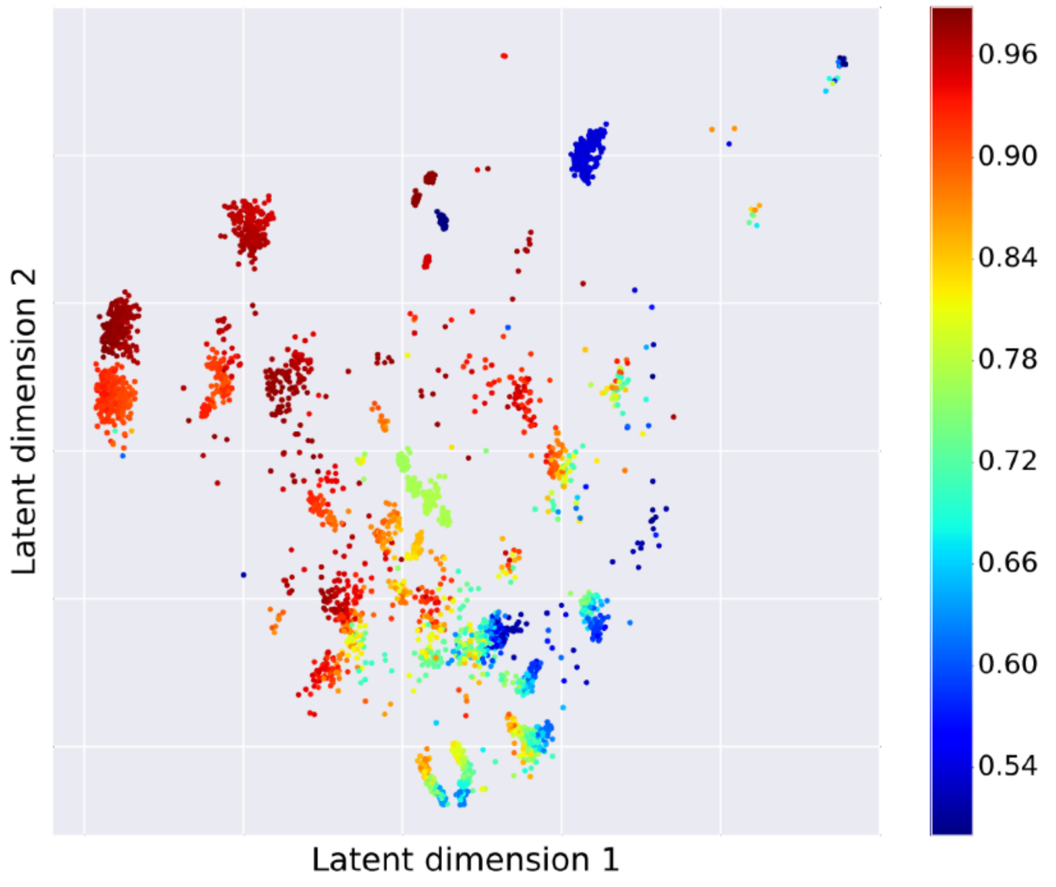


likely best models

Perrone et al. 2018

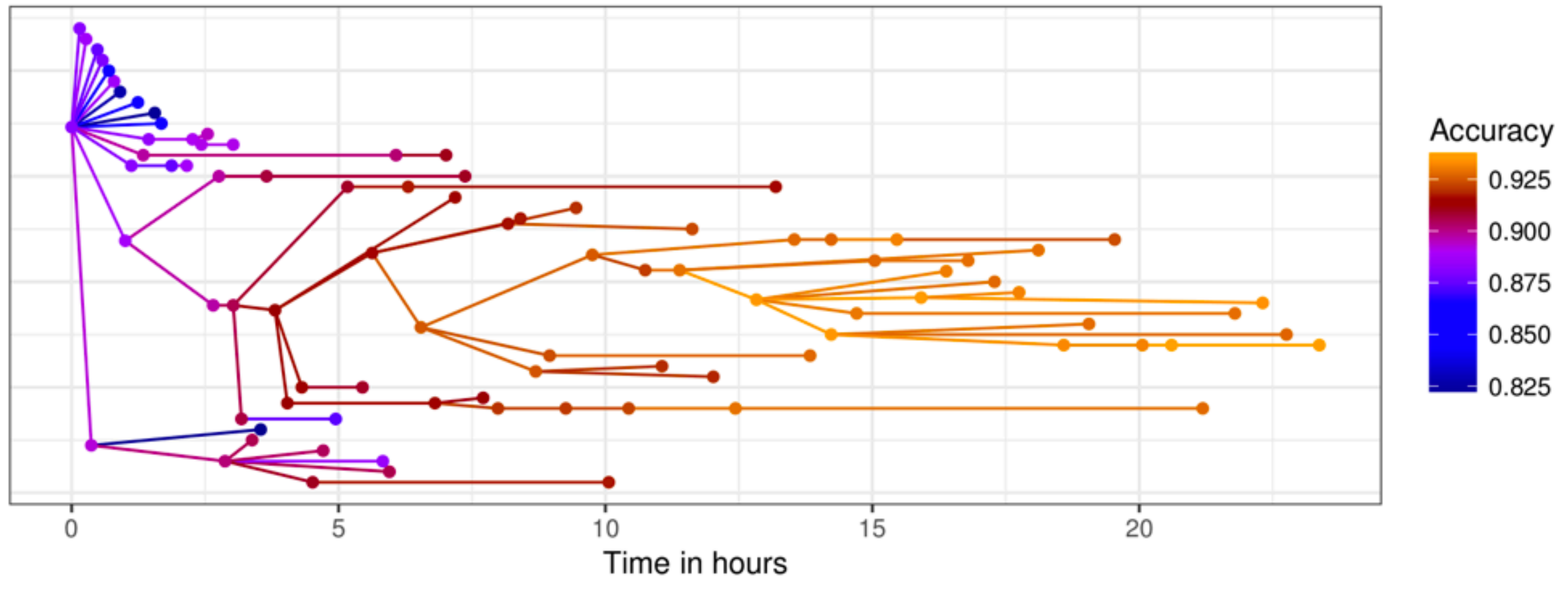
Automating machine learning

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



Fusi et al. 2018

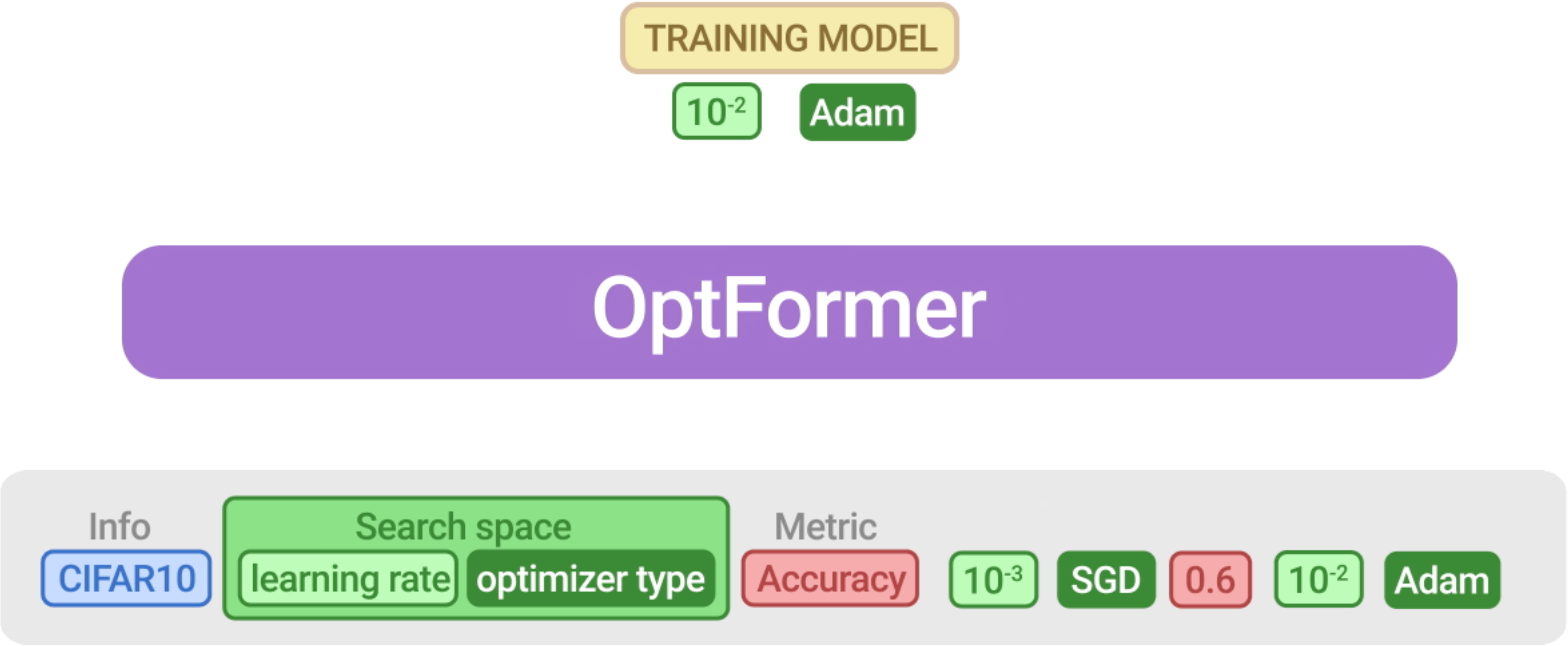
GAMA (TU/e): modular AutoML system, handles wide range of tasks



Gijsbers et al. 2018 - 2022

Automating machine learning

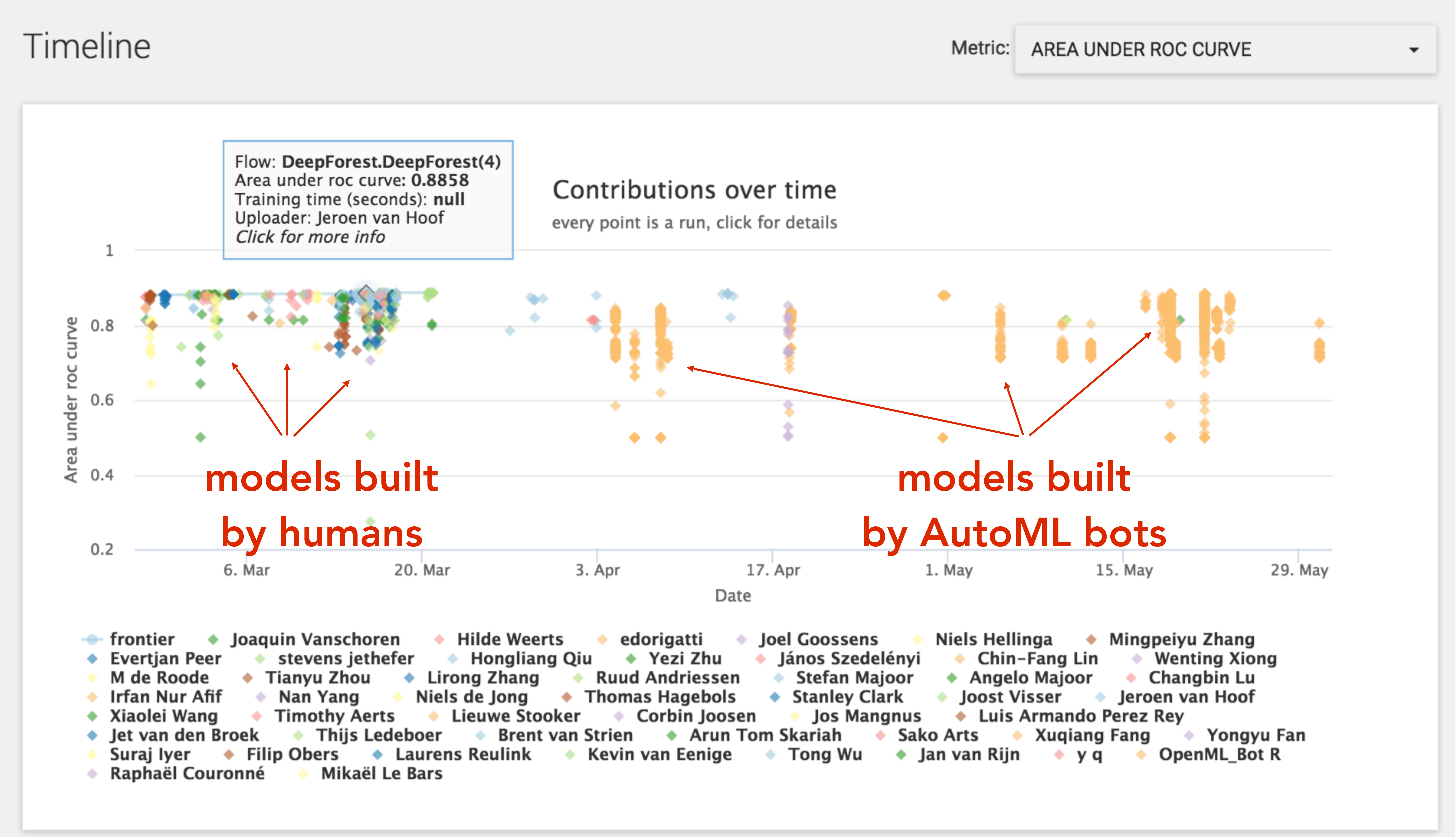
OptFormer (DeepMind): uses OpenML to train a transformer model, predict the next models to try



Chen et al. 2022

Human-AI interaction

Algorithms learn from models shared by humans
Humans learn from models built by bots



Join us! (and change the world)

Active open source community

- Hackathons 2-3x a year

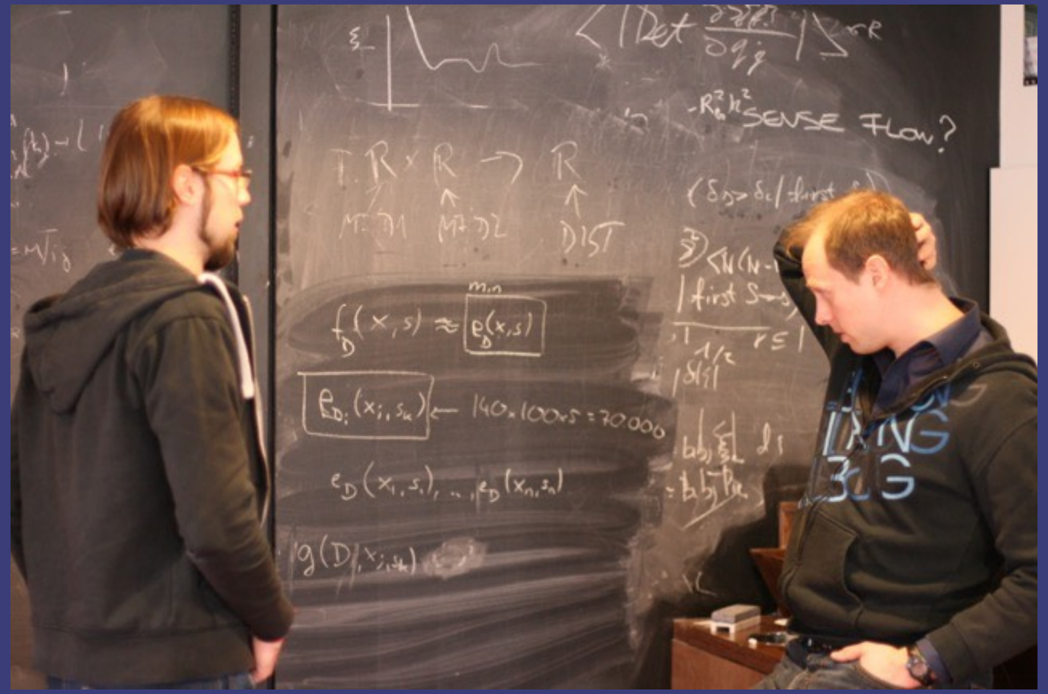
OpenML Foundation

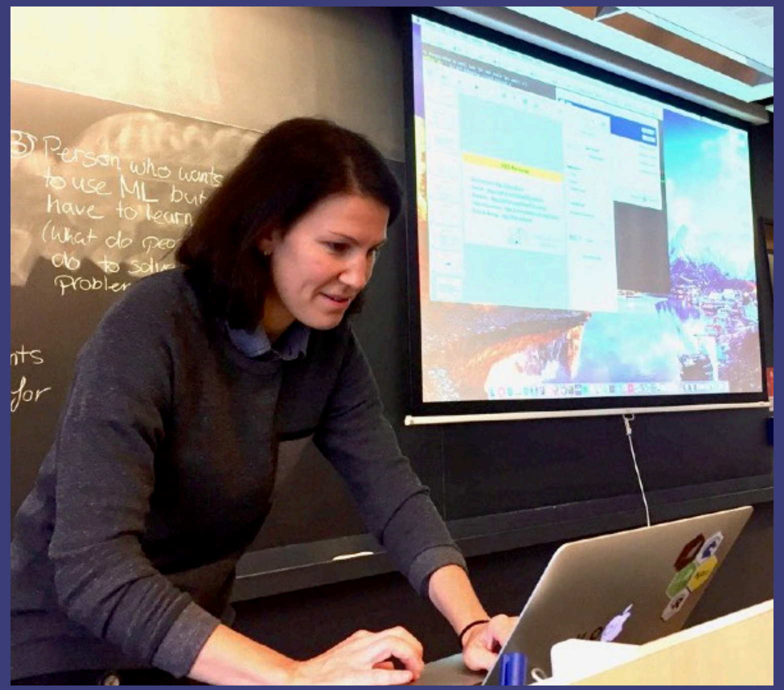
- Sponsorship, science

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


Guiseppe Casalicchio


Michel Lang


Jan van Rijn


Bilge Celik


Janek Thomas


Erin Ledell


Prabhant Singh


Marcel Wever


Sebastian Fischer


Neil Lawrence

and many more!

Thank you!

谢谢

@open_ml

OpenML

www.openml.org