

AutoML Systems and Lookout

Marius Lindauer

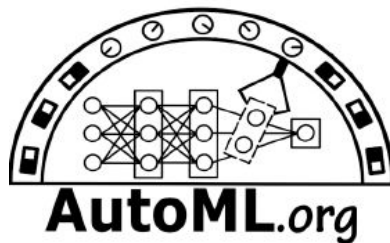
Leibniz Universität Hannover
Germany

 / LindauerMarius
m.lindauer@ai.uni-hannover.de

Katharina Eggensperger

Eberhard Karls Universität Tübingen
Germany

 / KEggensperger
katharina.eggensperger@uni-tuebingen.de





Questions?

Story Line Today

- AutoML Systems
 - Machine Learning Pipelines
 - Auto-Sklearn
 - Demo
- AutoML in the Wild
- Conclusion

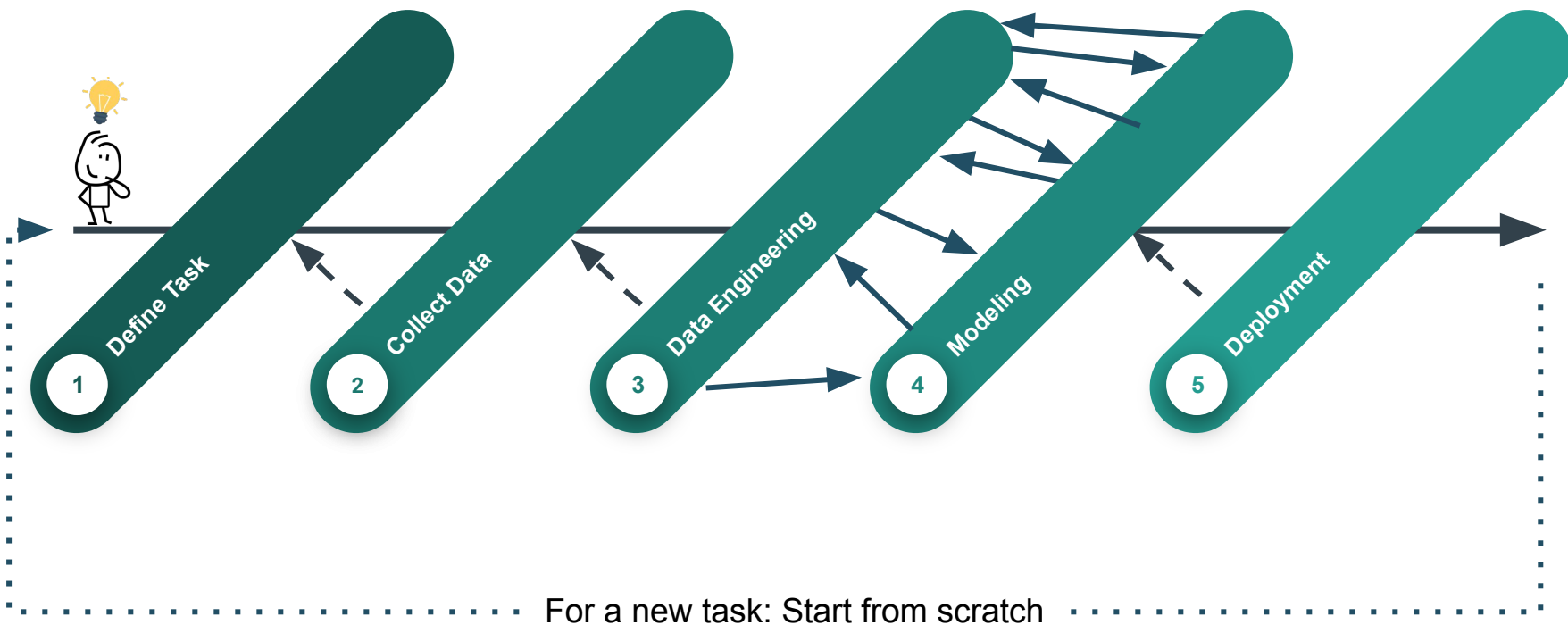
AutoML Systems

>> I need a tool for this!

Machine Learning Pipelines



Why does ML development take a lot of time?



AutoML for Tabular Data. Why?



Malicacid	Feature
Ash	Feature
Alcalinity_of_ash	Feature
Magnesium	Feature
Total_phenols	Feature

1, .28, 2.29, 5.64, 1.04, 3.92, 1065
1, 13.2, 1.78, 2.14, 11.2, 100, 2.65, 2.76, .26, 1.28, 4.38, 1.05, 3.4, 1050
1, 13.16, 2.36, 2.67, 18.6, 101, 2.8, 3.24, .3, 2.81, 5.68, 1.03, 3.17, 1185
1, 14.37, 1.95, 2.5, 16.8, 113, 3.85, 3.49, .24, 2.18, 7.8, .86, 3.45, 1480
1, 13.24, 2.59, 2.87, 21, 118, 2.8, 2.69, .39, 1.82, 4.32, 1.04, 2.93, 735
1, 14.2, 1.76, 2.45, 15.2, 112, 3.27, 3.39, .34, 1.97, 6.75, 1.05, 2.85, 1450
1, 14.39, 1.87, 2.45, 14.6, 96, 2.5, 2.52, .3, 1.98, 5.25, 1.02, 3.58, 1290
1, 14.06, 2.15, 2.61, 17.6, 121, 2.6, 2.51, .31, 1.25, 5.05, 1.06, 3.58, 1295
1, 14.83, 1.64, 2.17, 14, 97, 2.8, 2.98, .29, 1.98, 5.2, 1.08, 2.85, 1045
1, 13.86, 1.35, 2.27, 16, 98, 2.98, 3.15, .22, 1.85, 7.22, 1.01, 3.55, 1045
1, 14.1, 2.16, 2.3, 18, 105, 2.95, 3.32, .22, 2.38, 5.75, 1.25, 3.17, 1510
1, 14.12, 1.48, 2.32, 16.8, 95, 2.2, 2.43, .26, 1.57, 5, 1.17, 2.82, 1280
1, 13.75, 1.73, 2.41, 16, 89, 2.6, 2.76, .29, 1.81, 5.6, 1.15, 2.9, 1320
1, 14.75, 1.73, 2.39, 11.4, 91, 3.1, 3.69, .43, 2.81, 5.4, 1.25, 2.73, 1150
1, 14.38, 1.87, 2.38, 12, 102, 3.3, 3.64, .29, 2.96, 7.5, 1.2, 3, 1547
1, 13.63, 1.81, 2.7, 17.2, 112, 2.85, 2.91, .3, 1.46, 7.3, 1.28, 2.88, 1310
1, 14.3, 1.92, 2.72, 20, 120, 2.8, 3.14, .33, 1.97, 6.2, 1.07, 2.65, 1280
1, 13.83, 1.57, 2.62, 20, 115, 2.95, 3.4, .4, 1.72, 6.6, 1.13, 2.57, 1130
1, 14.19, 1.59, 2.48, 16.5, 108, 3.3, 3.93, .32, 1.86, 8.7, 1.23, 2.82, 1680
1, 13.64, 3.1, 2.56, 15.2, 116, 2.7, 3.03, .17, 1.66, 5.1, 0.6, 3.26, 845

Photo by [Nathan Dumlaio](#) on [Unsplash](#)

Why is this challenging?

- mixed/categorical features
- features on different scales
- missing features
- highly structured data
- feature engineering needed

Why is this relevant?

→ healthcare, biology, social sciences, finance, geoscience, physics, chemistry, mechanics, ...

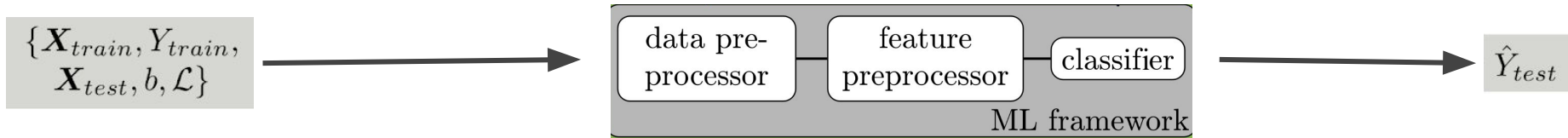
What is awesome about tabular data?

→ There exist hundreds of datasets from different domains with different characteristics

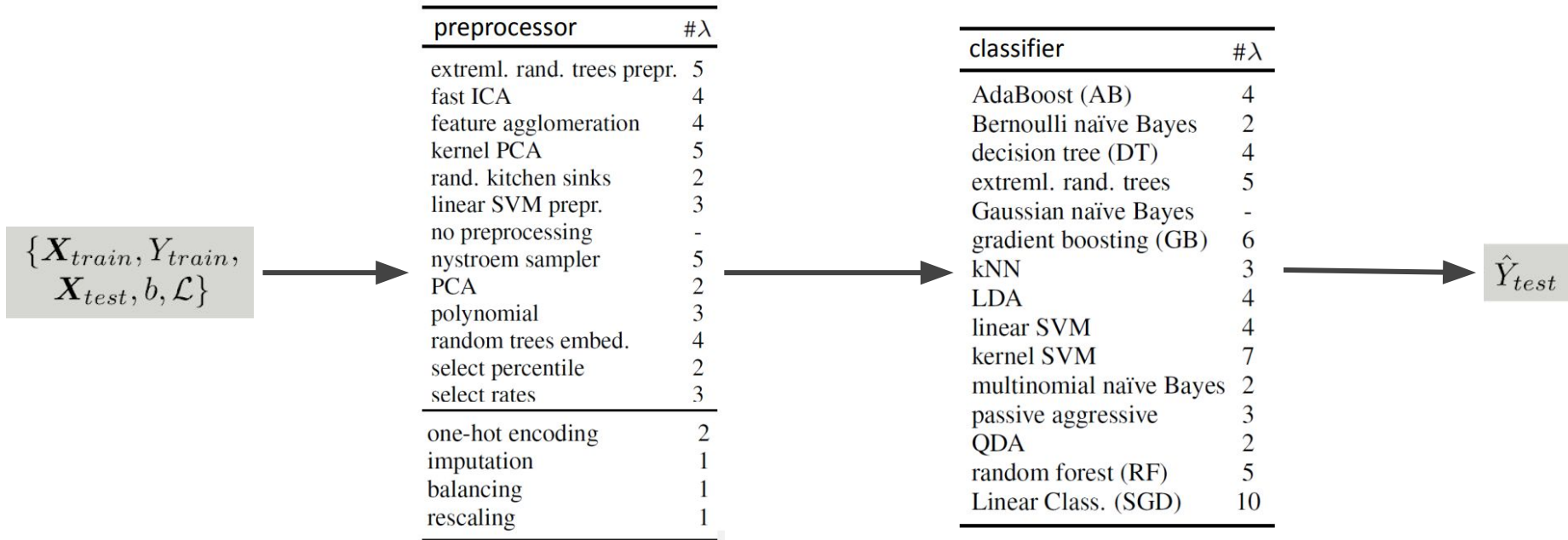
Great for

- **meta-learning**
- **studying algorithms**
- **comparing algorithms**

AutoML Systems (and what they do)



AutoML Systems (and what they do)

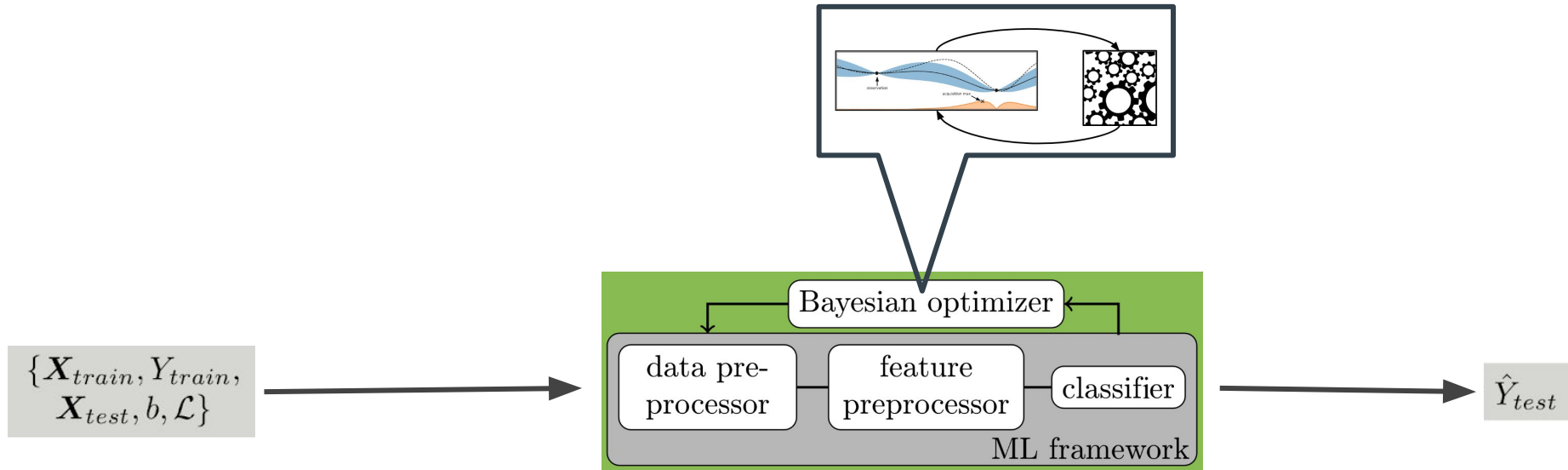


AutoML Systems (and what they want)

*Machine Learning for everyone
in 4 lines of code*

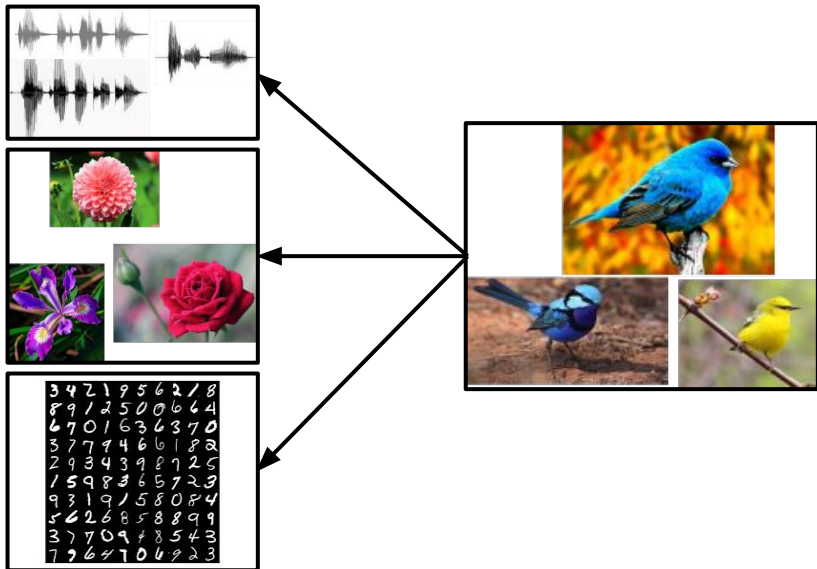
```
import autosklearn.classification
>>> cls = autosklearn.classification.AutoSklearnClassifier()
>>> cls.fit(X_train, y_train)
>>> predictions = cls.predict(X_test)
```

AutoML Systems (and what they do)



More I: Meta-Learning

→ Warmstart Bayesian Optimization



Offline / Before:

- 1) Collect >200 datasets
- 2) Find the best pipeline on each dataset

Online / For a new dataset:

- 1) Compute 38 meta-features, select 25 most similar previous datasets
- 2) Initialize optimization with best pipelines on those datasets

More II: Ensembling

→ Build an ensemble



Image credit: [Photo by Denise Leon](#)

AutoML Systems (and what they do)

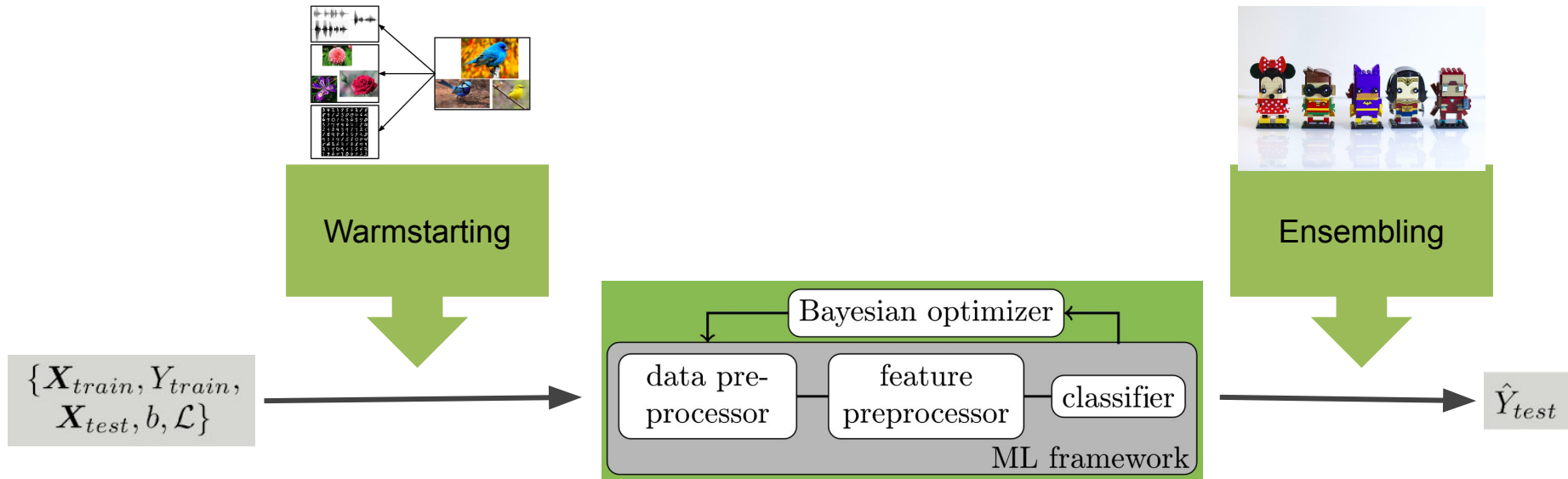


Image credit: [Photo by Denise Leon](#)

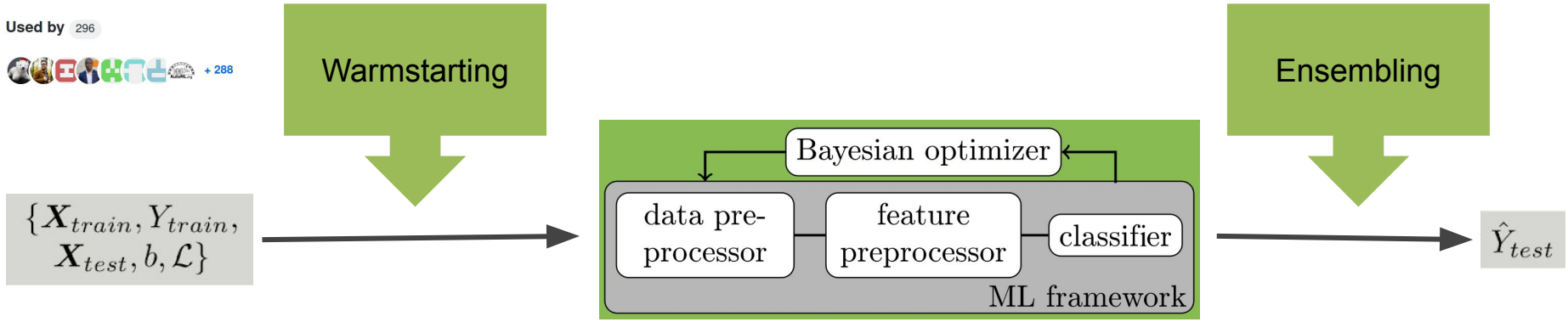
Auto-Sklearn 1.0

automl / auto-sklearn Public

Edit Pins Unwatch 214 Fork 1.2k Starred 6.5k

Code Issues 110 Pull requests 9 Discussions Actions Projects 1 Wiki

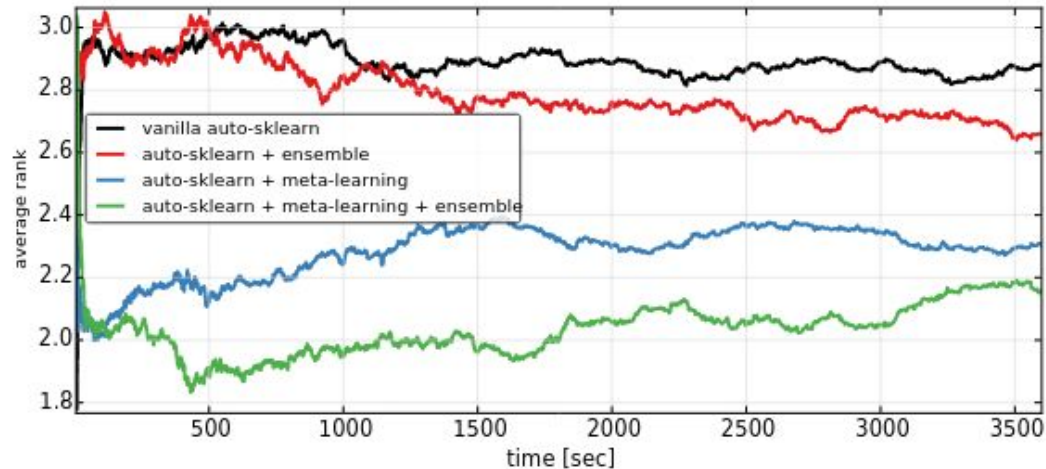
Used by 296



[/automl/auto-sklearn](https://github.com/automl/auto-sklearn)

Contributors 78

Auto-Sklearn 1.0 - Results



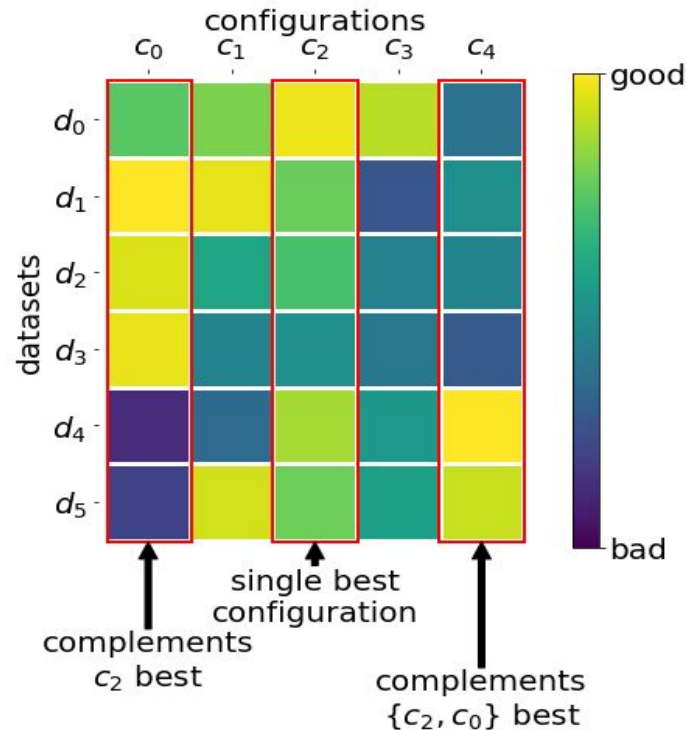
However, some things to be improved

- meta-features can be expensive to compute
- large datasets can be an issue

Even More I: Portfolios

Goal Meta-Learning without meta-features

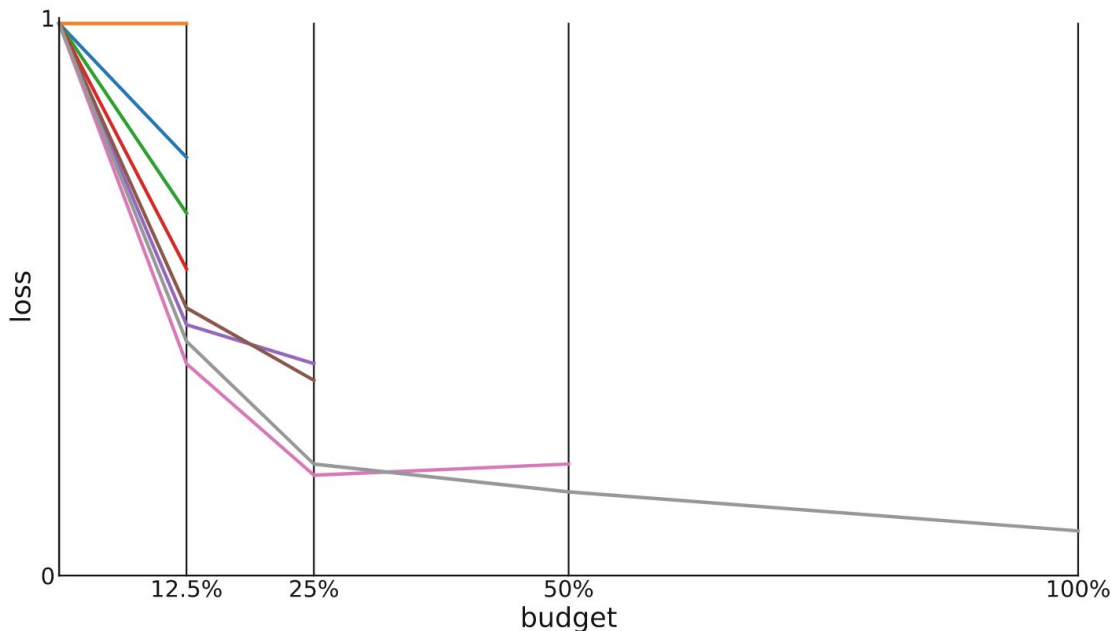
Idea Construct a Portfolio
(a list of diverse pipelines)



Even More II: Successive Halving

Goal Scale to large datasets.

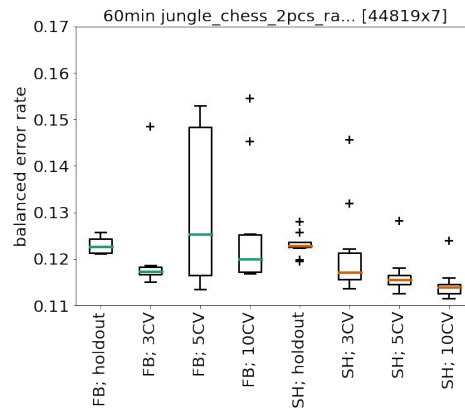
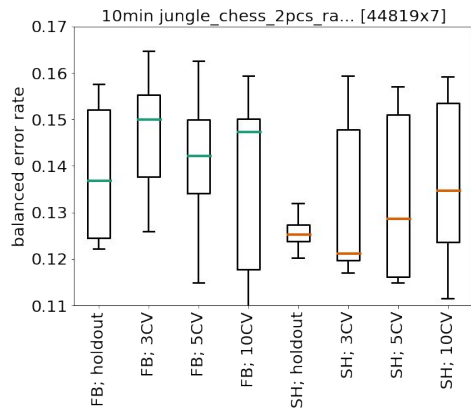
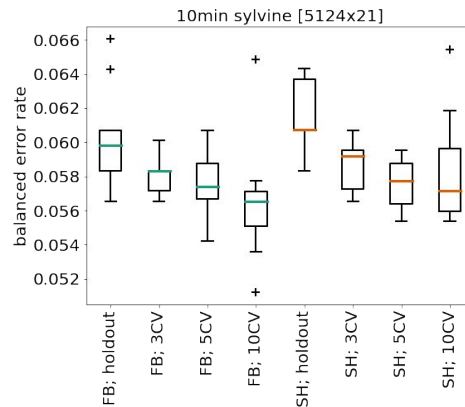
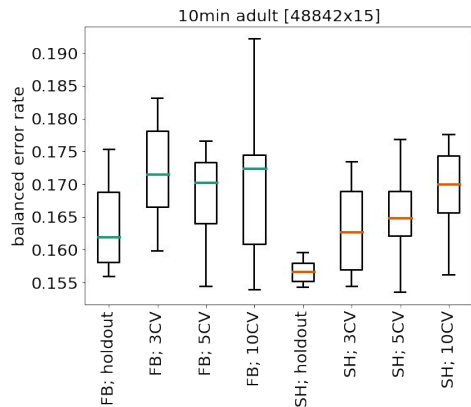
Idea Allocate more resources to promising pipelines



But what about small datasets?

Image Credit - CC-BY
Matthias Feurer and Frank Hutter: *Hyperparameter optimization*
Automated Machine Learning, The Springer Series on Challenges in Machine Learning

Impact of the Optimization Strategy



Wait what? ... Did we make it worse?

Can we automatically select an optimization policy?

→ Yes!

→ We can learn a selector [Feurer et al 2022]



Photo by [John Lockwood](#) on [Unsplash](#)

Learned Selector

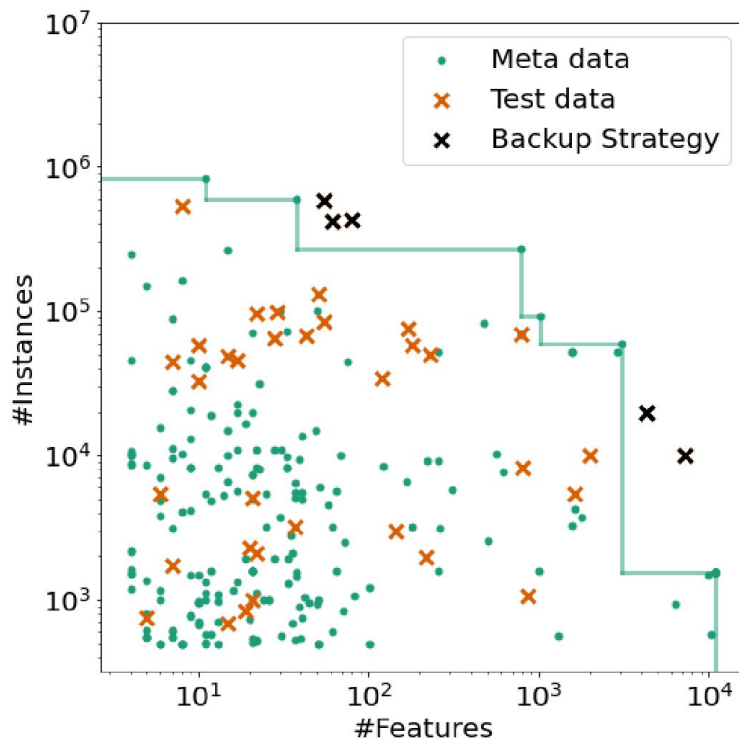
How?

Given a set of meta-datasets,

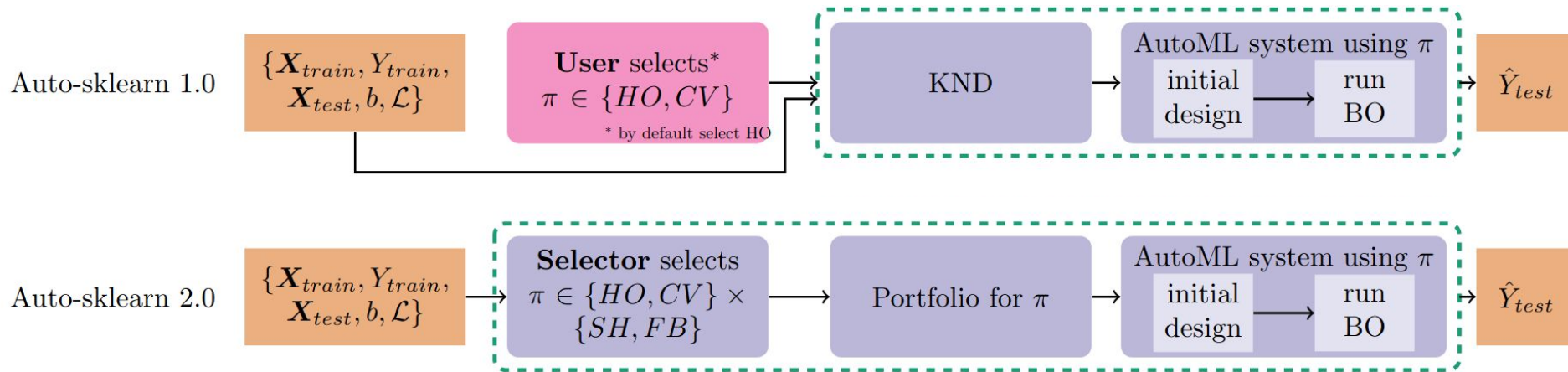
- for each dataset find best policy
- train a meta-selection model

Limitations

- meta-model is trained on a fixed budget
- meta-datasets need to be representative of the new dataset



Autosklearn 1.0 vs Auto-sklearn 2.0



Demo: SMAC / Auto-Sklearn

>> Here's my data. How do I use this?

Other OSS Systems?

 AutoGluon

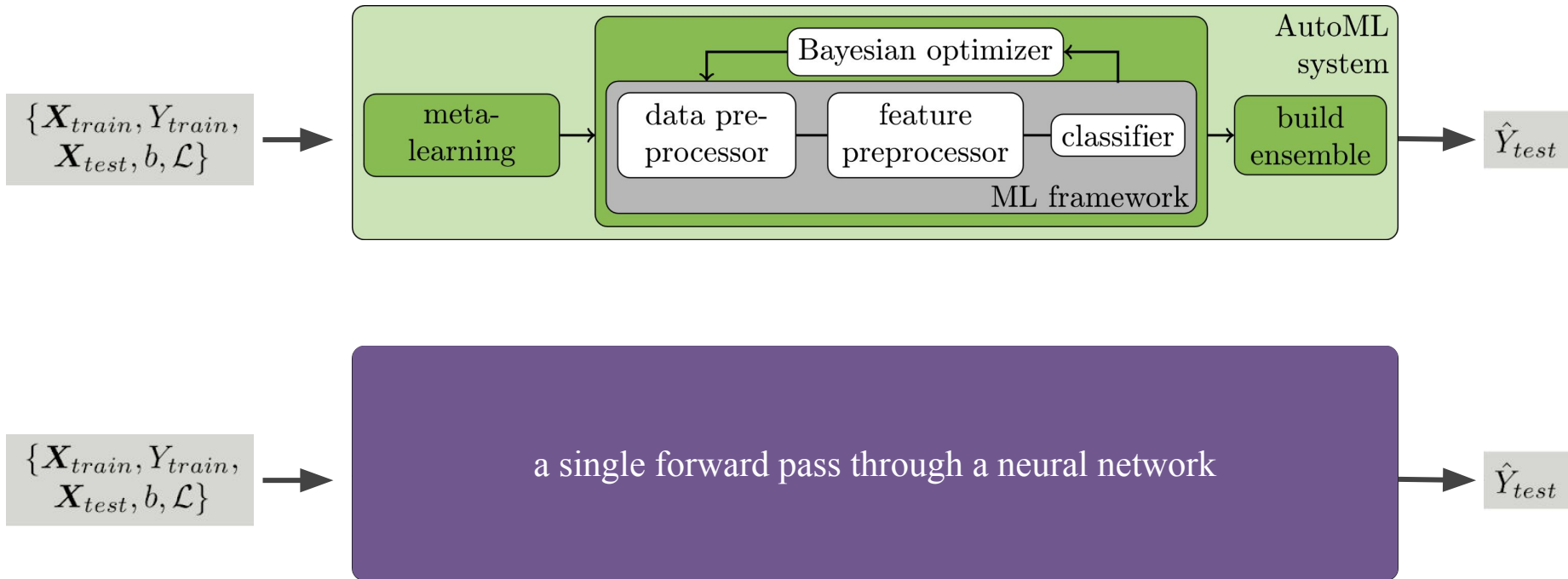
 TPOT

 GAMA

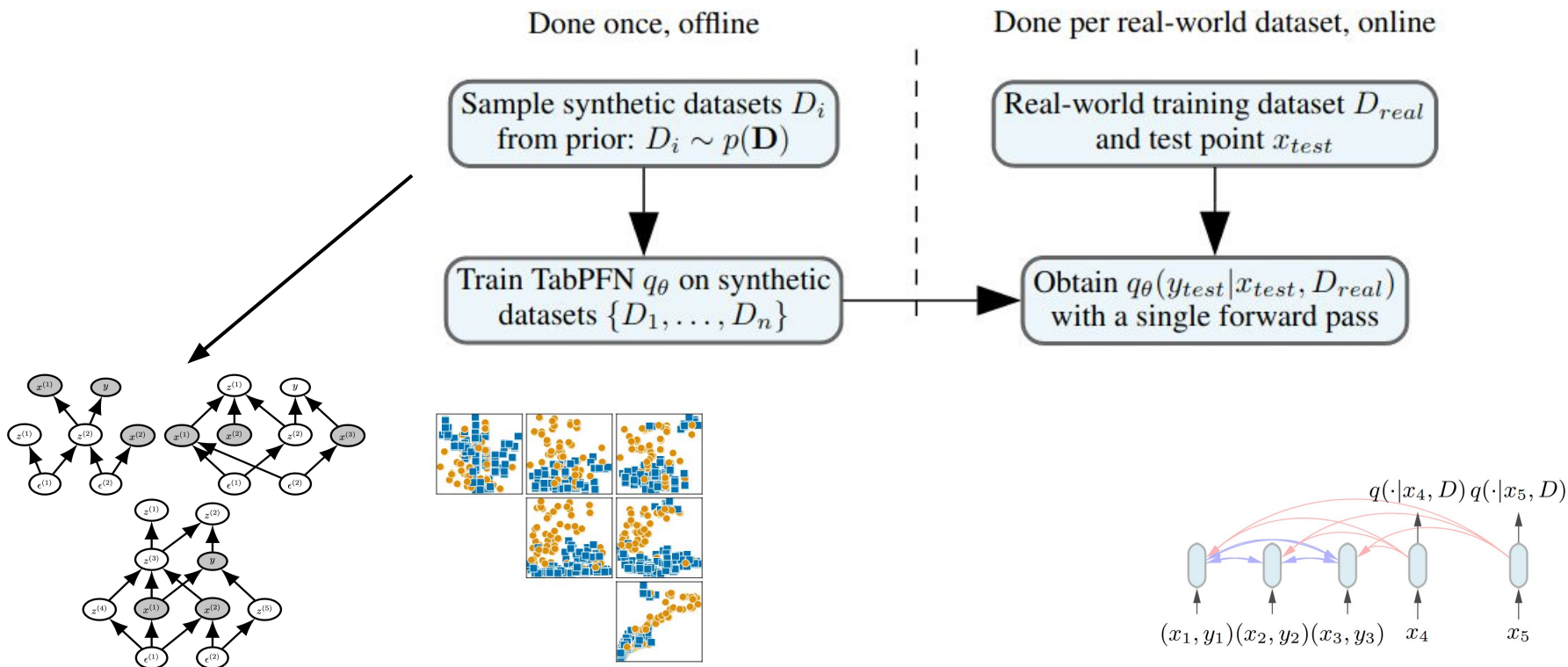
 FLAML

... and many more, see: <https://openml.github.io/automlbenchmark/frameworks.html> [Gijssbers et. al, 2022]

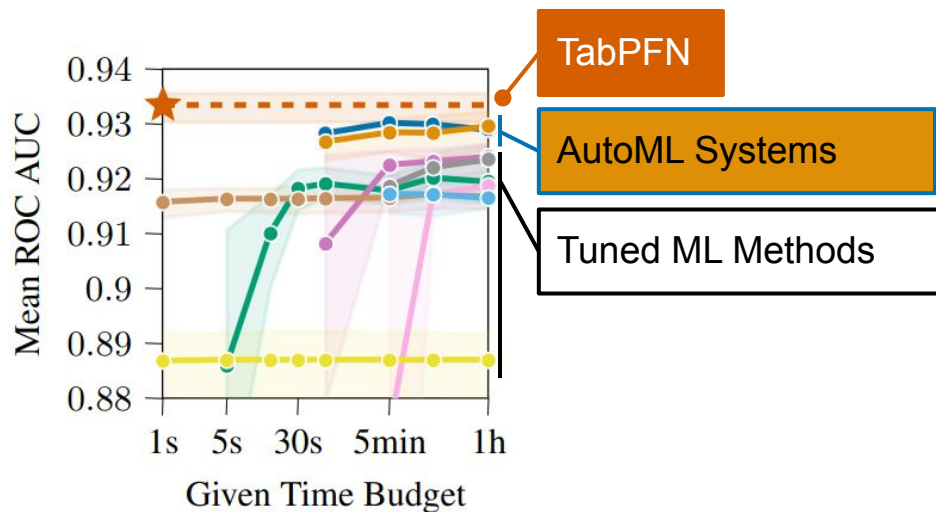
TabPFN: Prior-fitted Networks for Tabular Data



TabPFN: Prior-fitted Networks for Tabular Data



TabPFN: Results



18 small datasets (<1000 samples),
continuous features, no missing
values

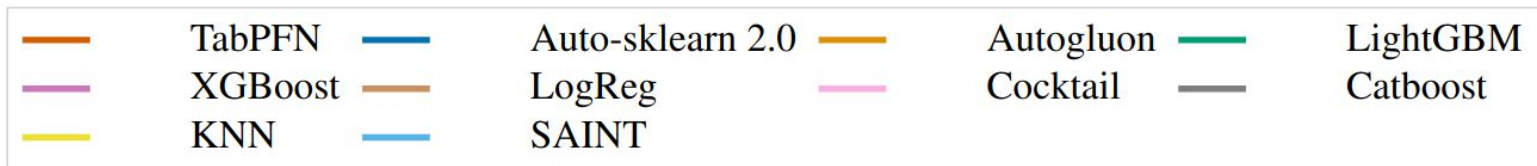


Image source: [\[Hollmann et al. 2023\]](#)

TabPFN: Summary

TL;DR TabPFN, a trained transformer, instantly yields predictions for tabular datasets.

Limitations and Remarks

- Up to 1000 samples
- Up to 100 features
- Up to 10 classes

→ works best on **continuous** datasets **without missing** values

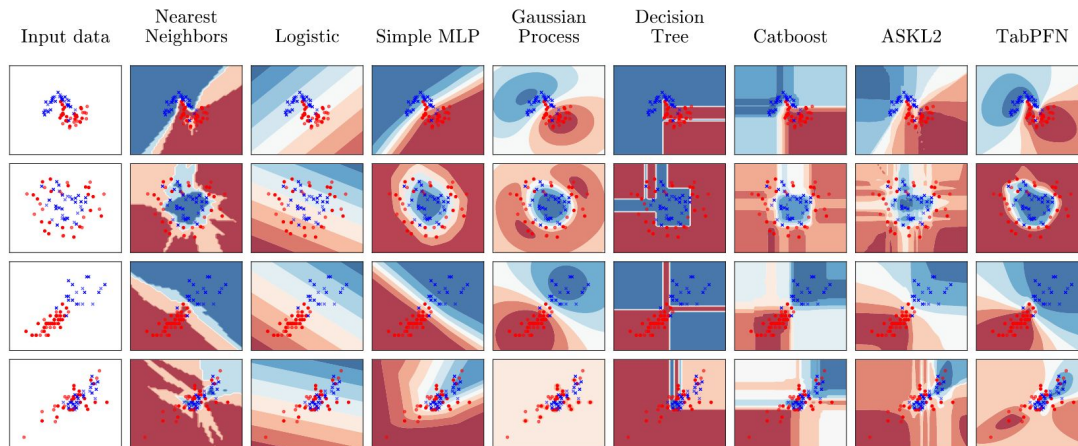


Image source: [\[Hollmann et al. 2023\]](#)



[/automl/TabPFN](#)



Questions?

AutoML in the Wild

>> Anything to consider?

AutoML x Fairness [\[Weerts et al. 2022\]](#)

One of many examples

“During the coronavirus crisis, students had to take exams at home. Universities used anti-cheat software to prevent fraud. Among other things, the software had to recognize the student’s faces. But it couldn’t recognize the student in question, Robin Pocornie. It wasn’t until she pointed an extra light at her face that the surveillance software Proctorio finally recognized her. And in the meantime, she had a lot of extra stress to deal with. She feels discriminated against.” [NL Times, 15.07.2023, [Webcam exam software “discriminatory,” doesn’t recognize darker skin tones, says student](#)]

→ Could’ve AutoML helped here?

→ Can we automate fairness?

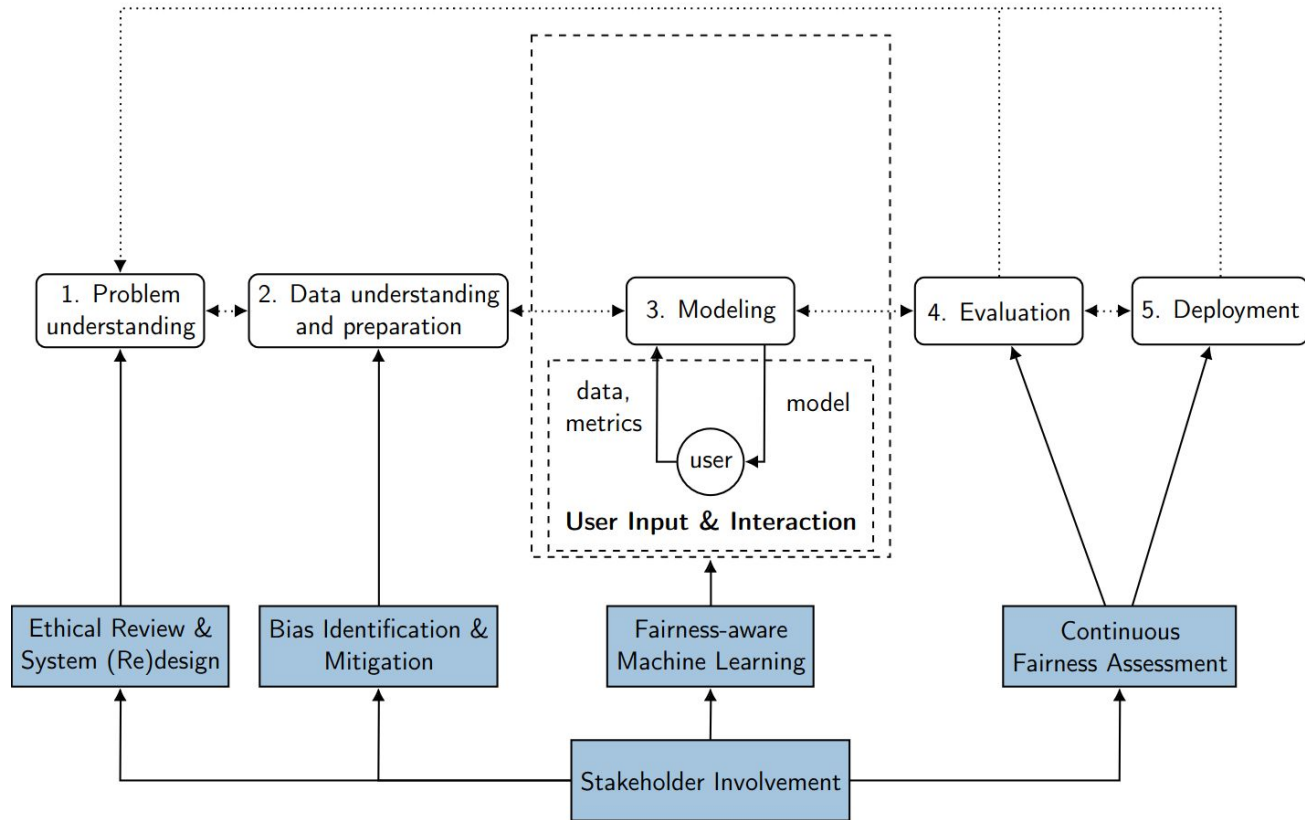


[Photo](#) by cottonbro studio

Based on <https://www.automl.org/can-fairness-be-automated/> and [Weerts et al. 2022]

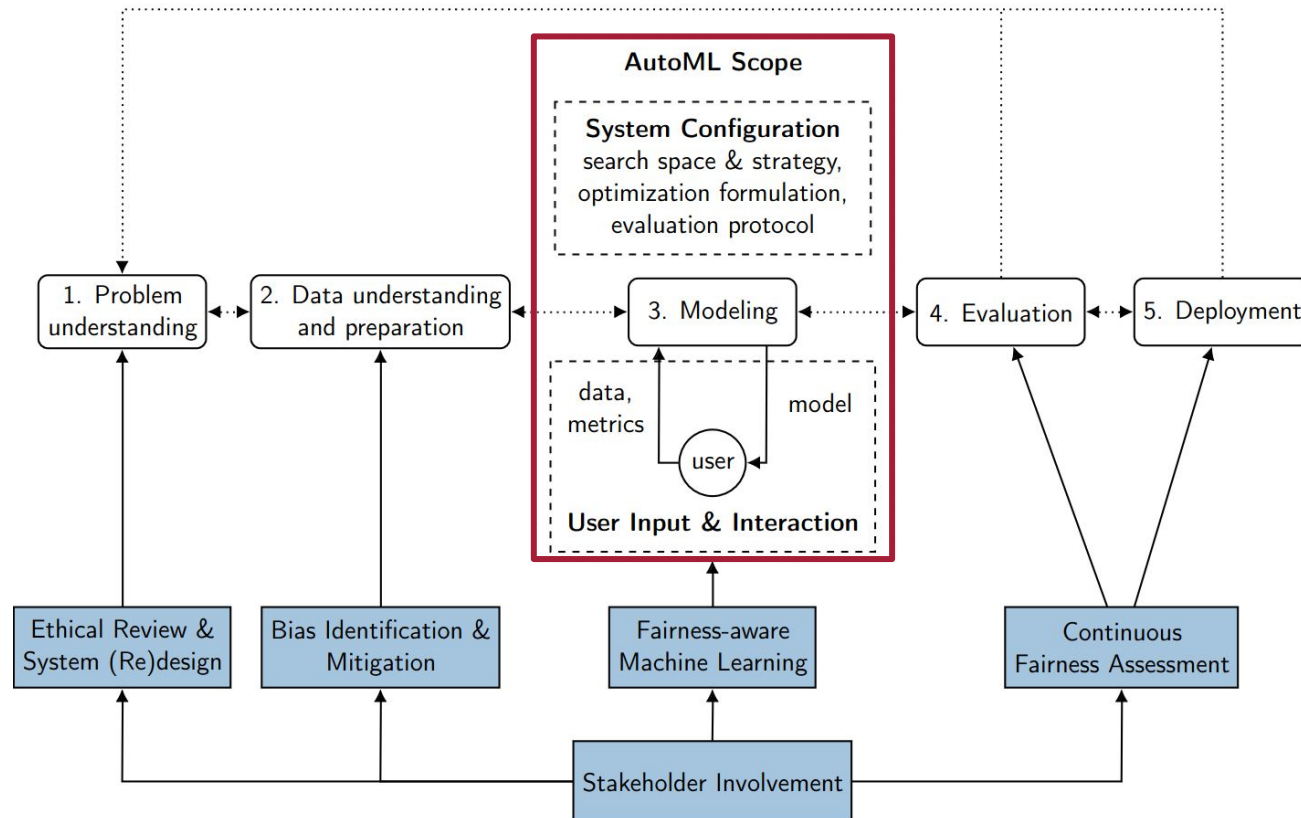
Fairness Considerations in the ML Workflow

[Weerts et al. 2022]



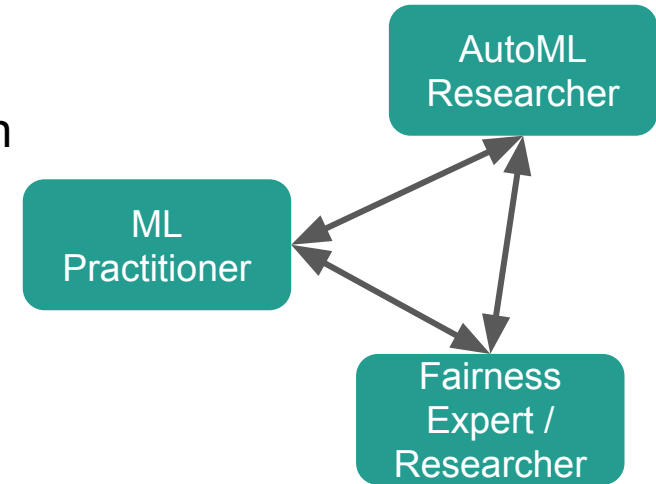
Opportunities for fairness-aware AutoML

[Weerts et al. 2022]



What can we do? Opportunities?

- Codifying best practices
- Better Multi-objective/Constrained optimization
- Better (contextualized) benchmarks
- Better interpretability/explainability
- Better reporting



Technical interventions are **not the sole tool for addressing unfairness!**

→ **No, we can not automate fairness!**

→ But AutoML can allow the user to **spend more time on aspects where a human in the loop is essential**

Green AutoML [\[Tornede et al. 2023\]](#)

Energy-efficient AutoML

Data compression,
Zero-cost AutoML,
multi-fidelity,
intelligent stopping, ...



Searching for Energy-Efficient Models

Model size constraint,
Energy-aware objective functions,
Energy efficient architectures,
Model compression, ...

AutoML for Sustainability

Plastic Litter Detection,
Green Assisted Driving,
Predictive Maintenance, ...

Create Attention

Tracking emissions,
awareness among students,
researchers, industry partners, ...





Questions?

Kahoot Quiz I

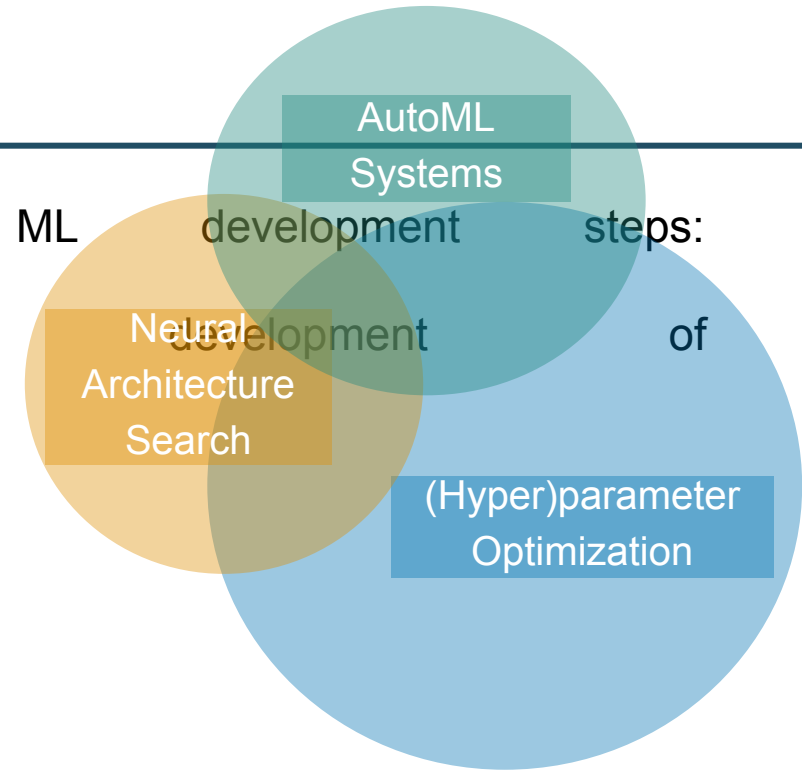
Your feedback

Conclusion

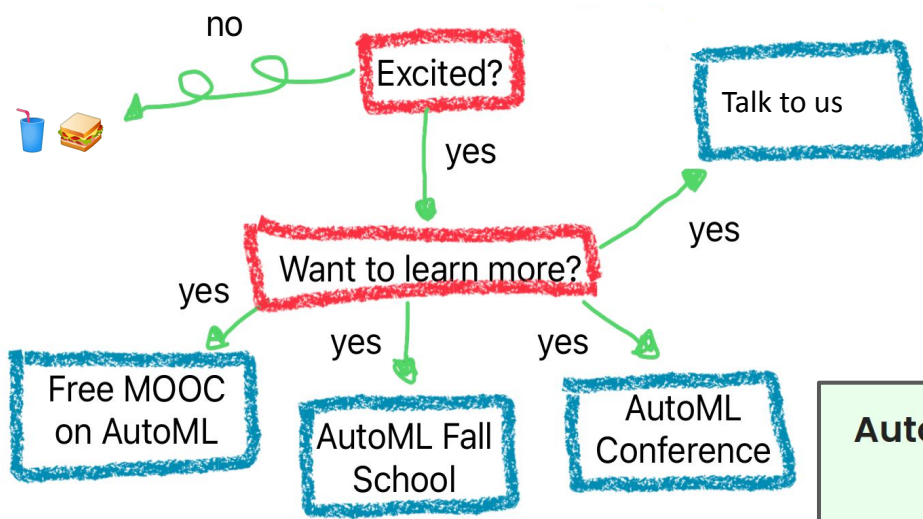
- AutoML helps for many ML development steps:
- HPO, NAS, AutoML systems speeds up of ML applications

Future

- Human-centered and trustworthy AutoML
- Foundation Models X AutoML
- Better Tooling



Advertisement !!!?



AI Campus Original

COURSE
AutoML - Automated Machine Learning

Universität Hannover
Universität Freiburg ...

Prof. Dr. Marius Lindauer
Prof. Dr. Frank Hutter
Prof. Dr. Bernd Bischl

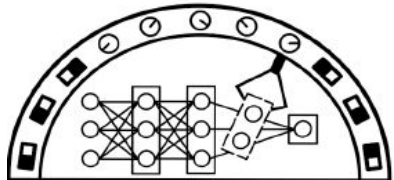
anytime

3rd AutoML Fall School 2023

Date: November 27th - 30th 2023 Place: Munich, Germany

AutoML Conference 2023

Potsdam/Berlin, Germany
September 12th - 15th 2023



For more info visit:

AutoML.org

Thanks.
Have a nice weekend!

