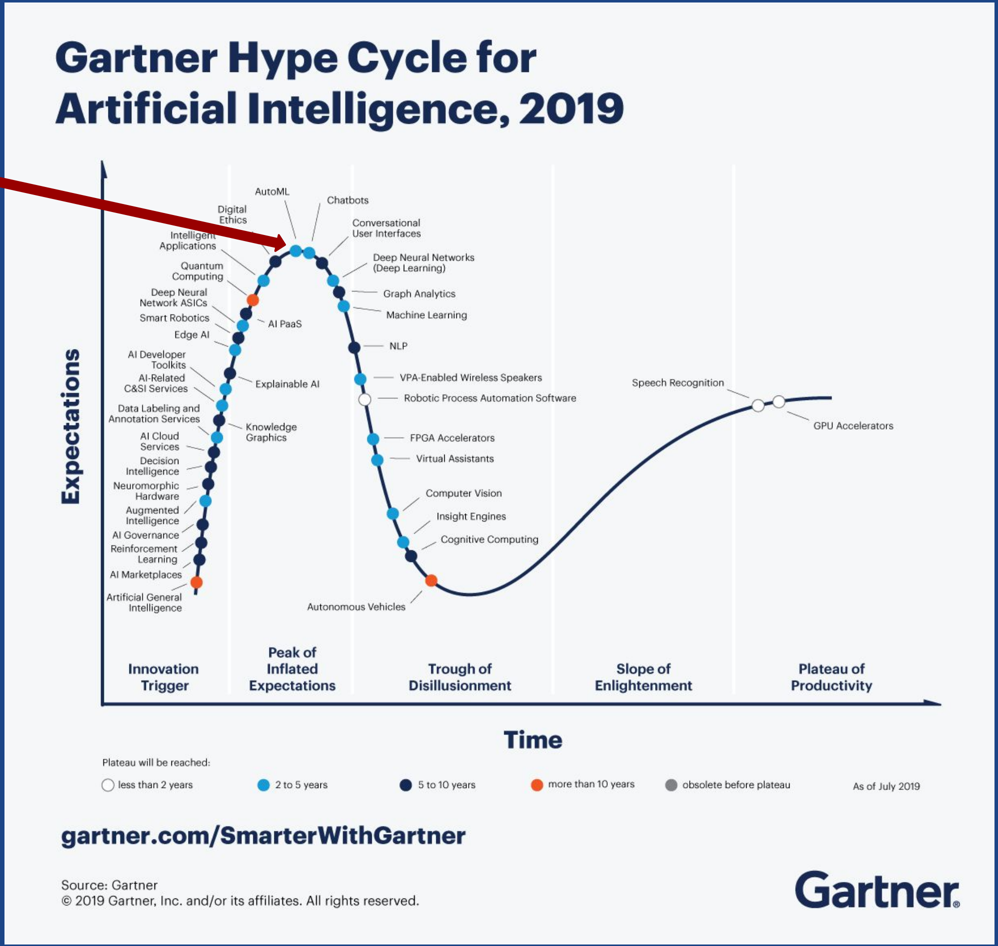
HOMER

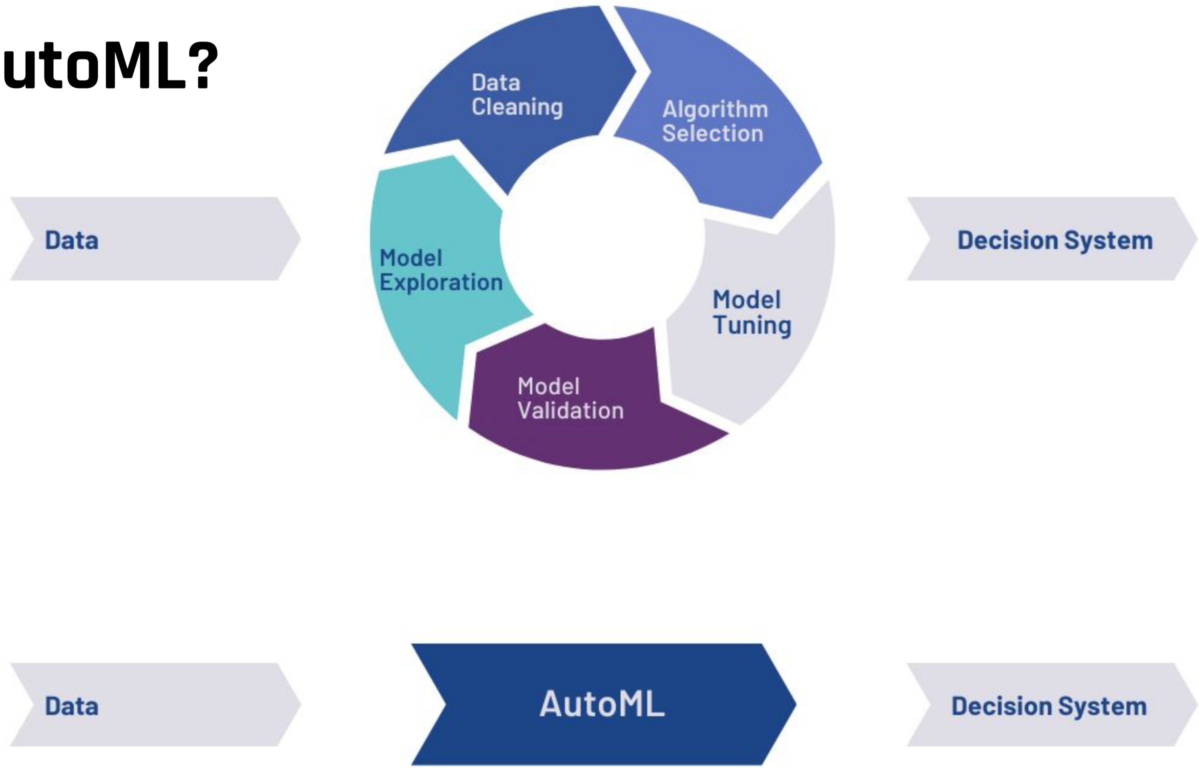
New era of AutoML?

Katarzyna Woźnica, Anna Kozak

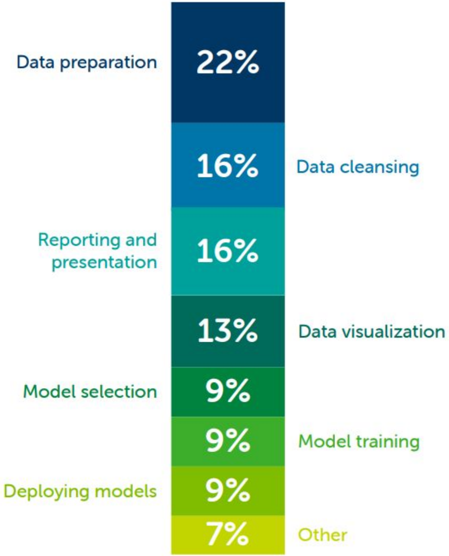

Gartner Hype Cycle for Artificial Intelligence, 2019
AutoML
Chatbots
Digital Ethics
Conversational User Interfaces
Intelligent Applications
Deep Neural Networks (Deep Learning)
Quantum Computing
Graph Analytics
Deep Neural Network ASICs
Machine Learning
Smart Robotics
AI PaaS
Edge AI
NLP
AI Developer Toolkits
AI-Related C&SI Services
Explainable AI
VPA-Enabled Wireless Speakers
Robotic Process Automation Software
Speech Recognition
Data Labeling and Annotation Services
Knowledge Graphics
GPU Accelerators
AI Cloud Services
FPGA Accelerators
Decision Intelligence
Virtual Assistants
Neuromorphic Hardware
Computer Vision
Augmented Intelligence
Insight Engines
AI Governance
Cognitive Computing
Reinforcement Learning
AI Marketplaces
Artificial General Intelligence
Autonomous Vehicles
Expectations
Innovation Trigger
Peak of Inflated Expectations
Trough of Disillusionment
Slope of Enlightenment
Plateau of Productivity
Time
Plateau will be reached:
less than 2 years
2 to 5 years
5 to 10 years
more than 10 years
obsolete before plateau
As of July 2019
gartner.com/SmarterWithGartner
Source: Gartner
© 2019 Gartner, Inc. and/or its affiliates. All rights reserved.
Gartner

What is AutoML?

What is AutoML?

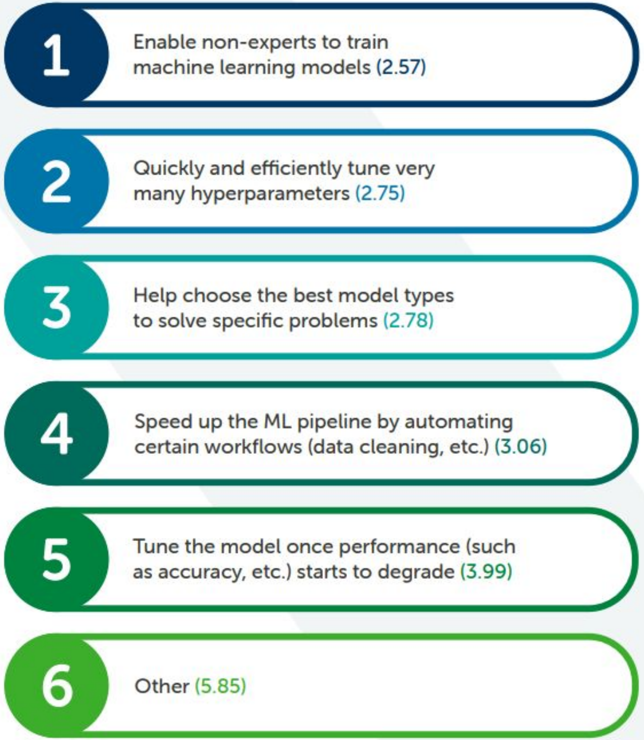


Who is AutoML end user?



n = 1,966

We asked our respondents how much time they spend on the above tasks, and for each item they entered a number reflecting the percentage of time spent relative to the other options. This is the average of the reported percentages.



We asked respondents to drag and rank the options from most to least important, with the first being most important.

n = 2,042

Anaconda, State of Data Science 2022
3,493 individuals from 133 countries)

Who is AutoML end user?

Traditionally, application's developers using statistical and learning methods choose algorithms and tune their parameters empirically, commonly by trial and error; or in the best case, by using prior knowledge of experts on the domain.

[PSMS for Neural Networks, 2007]

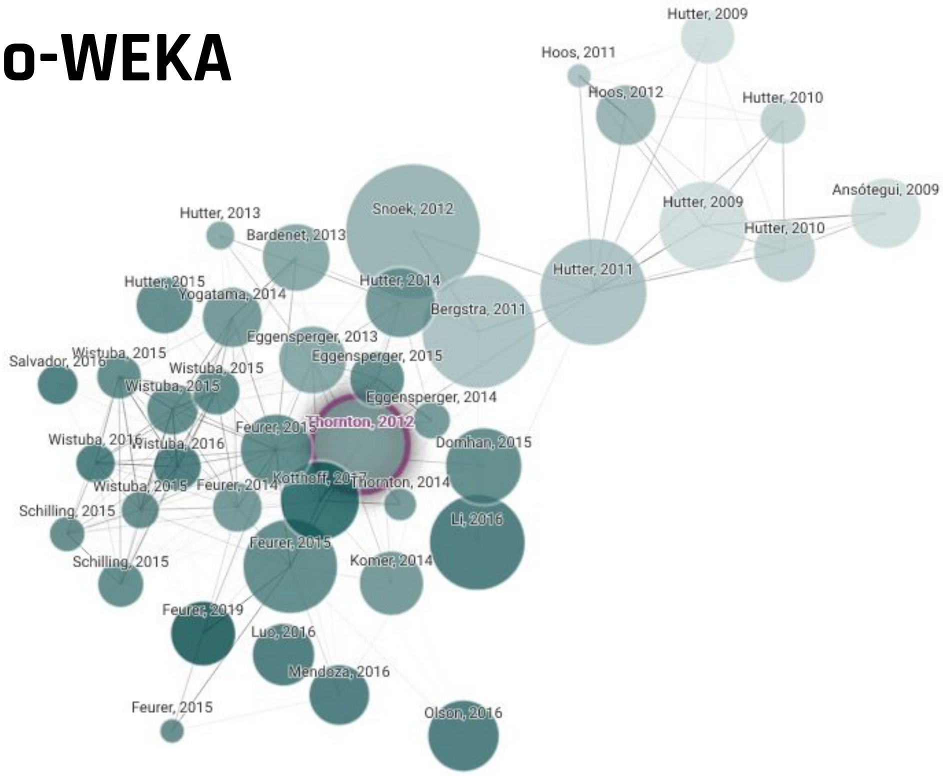
*It can be challenging to make the **right choice** when faced with these degrees of freedom, leaving many users to select algorithms based on reputation or intuitive appeal, and/or to leave hyperparameters set to default values.*

[AutoWEKA, 2013]

*Automated Machine Learning (AutoML) supports **practitioners and researchers** with the tedious task of designing machine learning pipelines and has recently achieved substantial success.*

[Auto-sklearn 2.0, 2022]

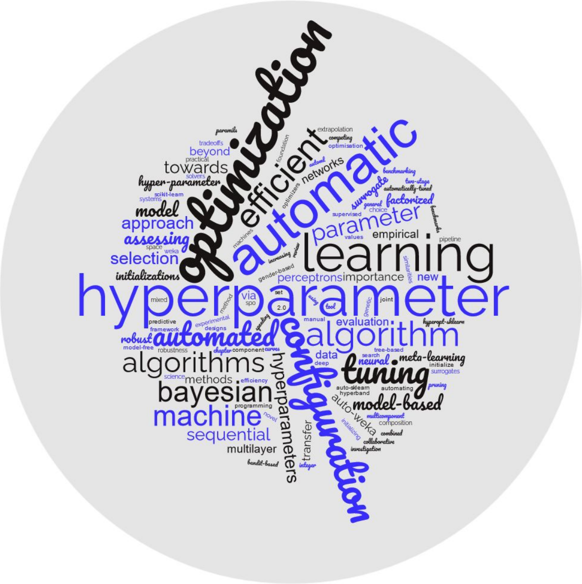
Example of Auto-WEKA



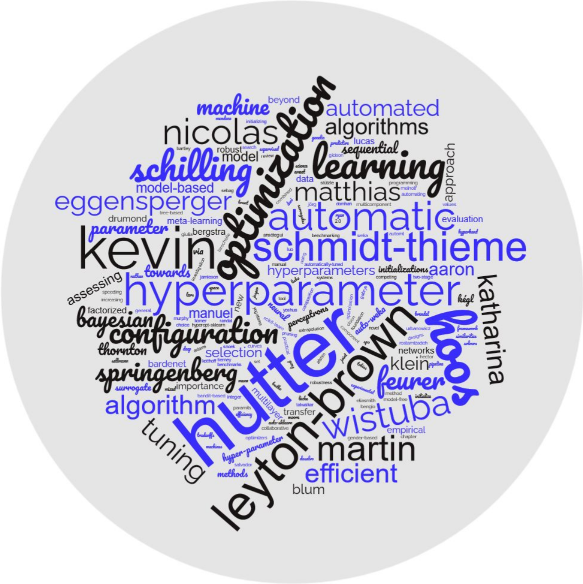
Source: https://www.connectedpapers.com/

Started point: C. Thornton and F. Hutter and H.~H. Hoos and K. Leyton-Brown, *Auto-WEKA: Combined Selection and Hyperparameter Optimization of Classification Algorithms*, 2013

Example of Auto-WEKA



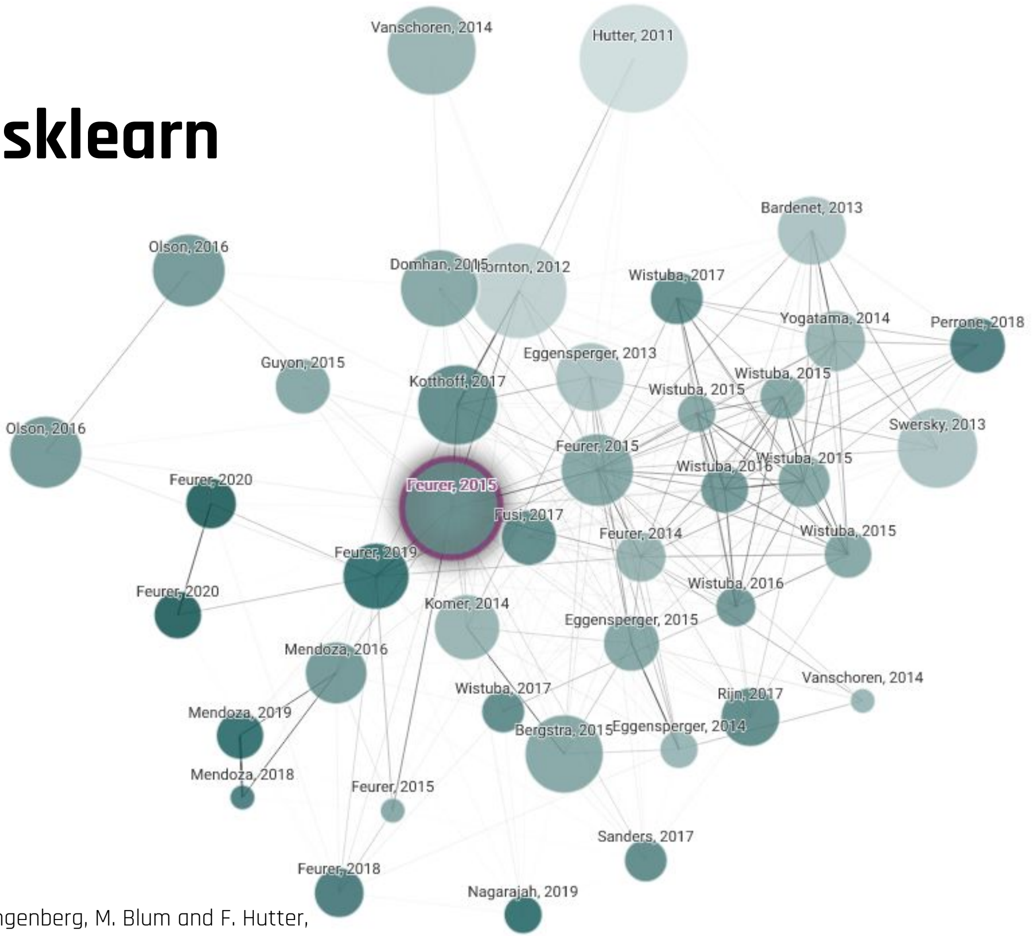
Without authors



With authors

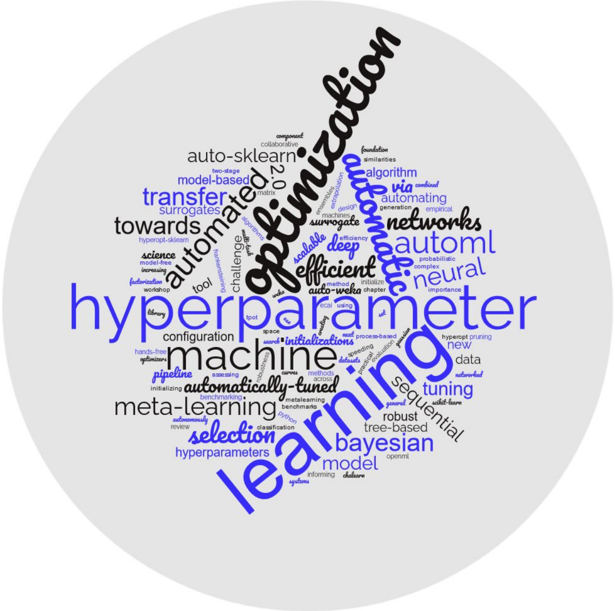
Source: https://www.connectedpapers.com/, https://www.wordclouds.com/
Started point: C. Thornton and F. Hutter and H.~H. Hoos and K. Leyton-Brown, *Auto-WEKA: Combined Selection and Hyperparameter Optimization of Classification Algorithms*, 2013

Example of Auto-sklearn



Source: https://www.connectedpapers.com/
Started point: M, Feurer, A. Klein, K. Eggensperger, J. Springenberg, M. Blum and F. Hutter, *Efficient and Robust Automated Machine Learning*, 2015

Example of Auto-sklearn



Without authors

With authors

Source: https://www.connectedpapers.com/, https://www.wordclouds.com/
Started point: M, Feurer, A. Klein, K. Eggensperger, J. Springenberg, M. Blum and F. Hutter, *Efficient and Robust Automated Machine Learning*, 2015

BUT...

Have we failed to reach out to our target audience?

- Despite various attempts of our community to make other researchers aware of AutoML, it seems that the amount of people we have reached is still rather limited. With new activities, such as our AutoML fall school, we hope to change that in the future.
- The capabilities of our AutoML tools are not well aligned with their needs. In particular,
  - their capabilities are most likely not broad enough by offering only very limited support for data engineering – a task that often requires a significant amount of time and expertise,
  - and the black-box nature of most AutoML processes makes it hard to understand why a certain (ensemble of) model(s) is returned at the end of running AutoML.

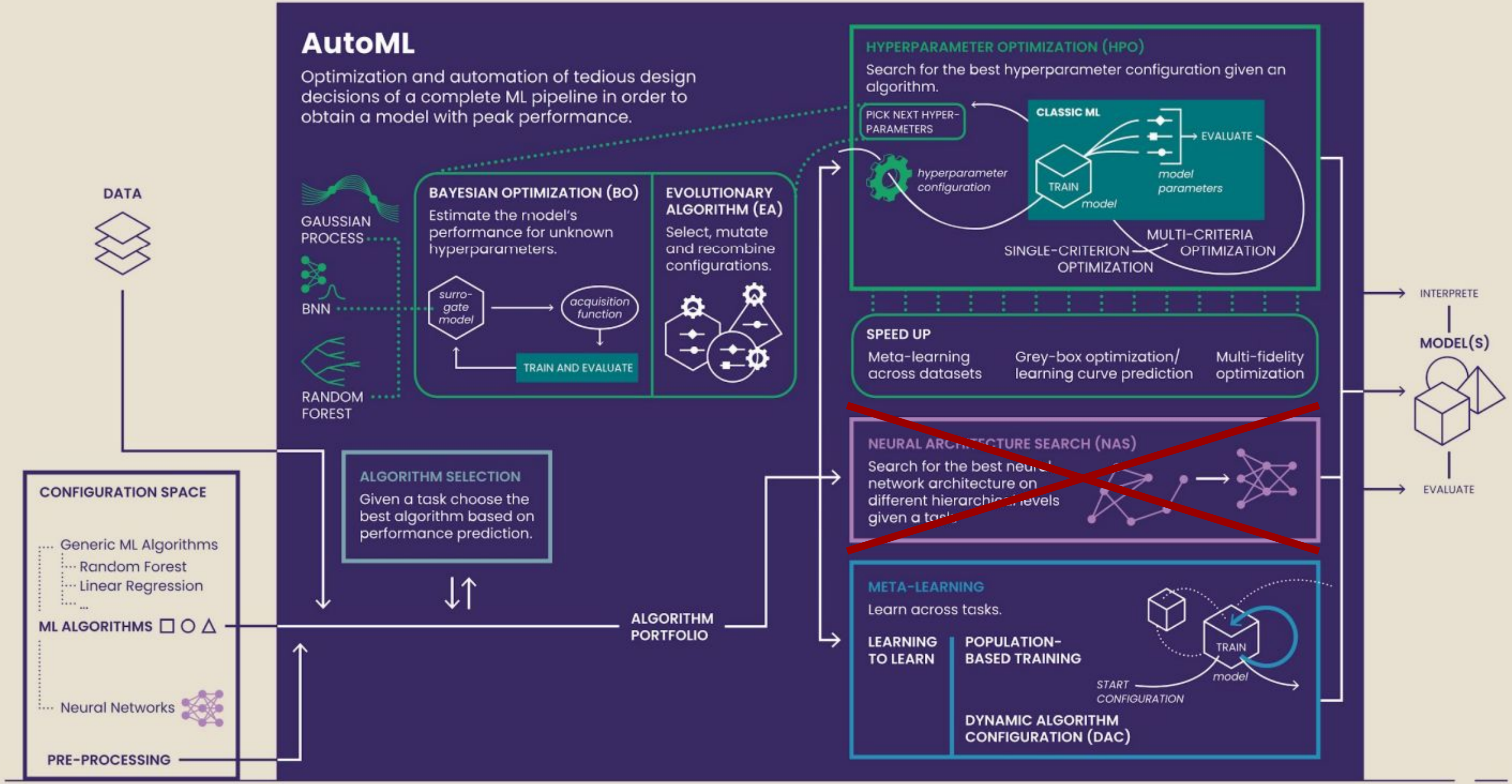
Source: M. Lindauer, A. Tornede, *Rethinking AutoML: Advancing from a Machine-Centered to Human-Centered Paradigm*, *November 30, 2022*
https://www.automl.org/rethinking-automl-advancing-from-a-machine-centered-to-human-centered-paradigm/

Ongoing research

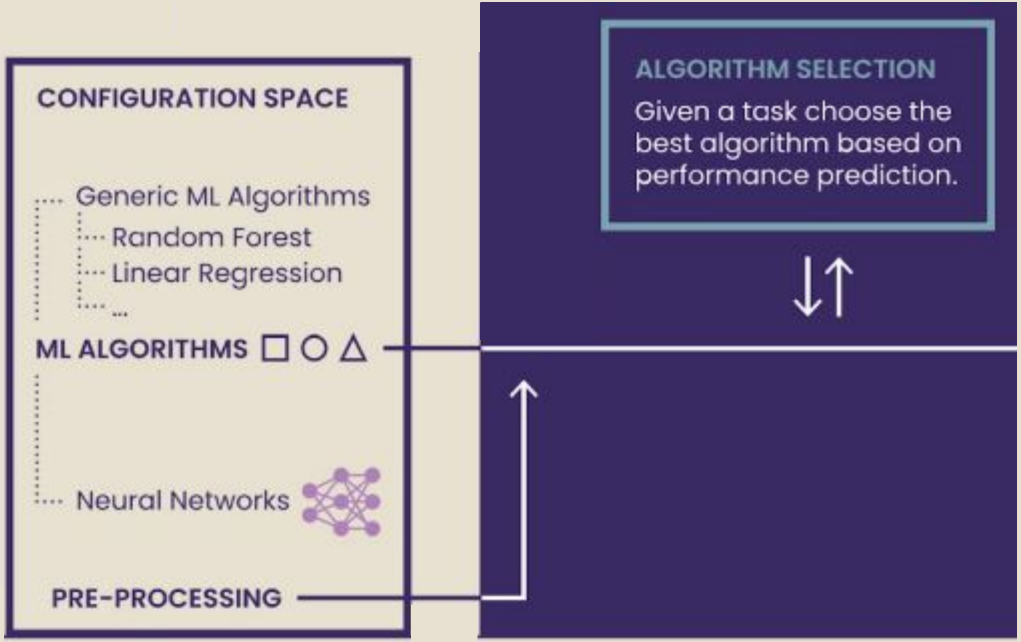
(mostly important for us)

MI

What points have researchers focused on so far?



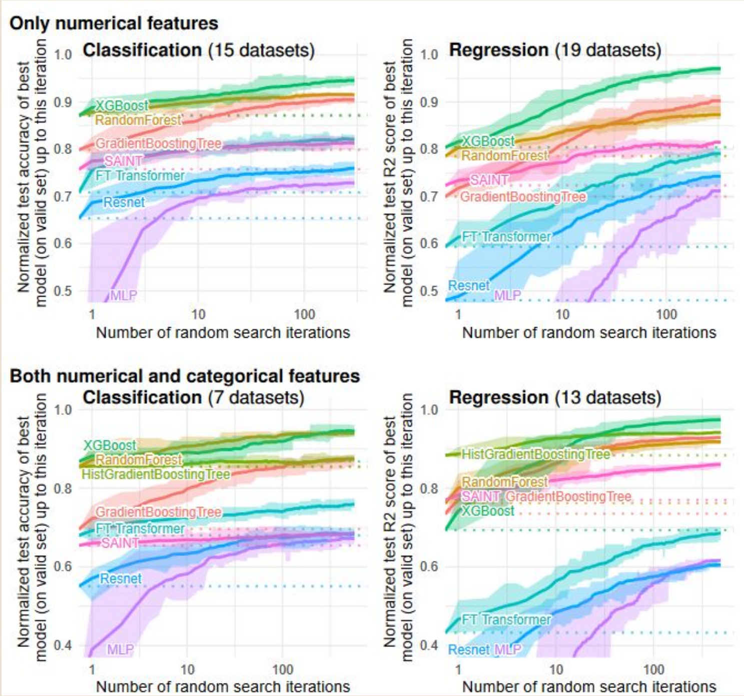
ML Algorithms for tabular data



ML Algorithms for tabular data - latest works

➔ R. Shwartz-Ziv and A. Armon, ***Tabular Data: Deep Learning is Not All You Need***, ICML 2021 Workshop AutoML Program Chairs

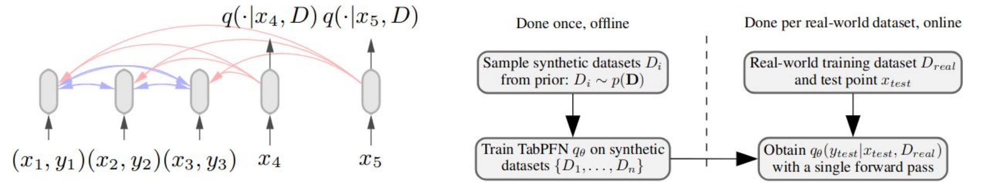
➔ L. Grinsztajn, E. Oyallon, and G. Varoquaux, ***Why do tree-based models still outperform deep learning on typical tabular data?***, NeurIPS 2022 Track Datasets and Benchmarks Program Chairs



L. Grinsztajn, E. Oyallon, and G. Varoquaux, *Why do tree-based models still outperform deep learning on typical tabular data?*, NeurIPS 2022 Track Datasets and Benchmarks Program Chairs

Forester Package @Anna.Kozak

TabPFN



(a) Architecture and attention mechanism (b) Prior-fitting and inference

Figure 1: Left (a): Training samples $\{(x_1, y_1), \ldots, (x_3, y_3)\}$ are transformed to 3 tokens, which attend to each other; test samples x_4 and x_5 attend only to the training samples. Right (b): The PFN learns to approximate the PPD of a given prior in the offline stage to yield predictions on a new dataset in a single forward pass in the online stage. Plots based on [24].

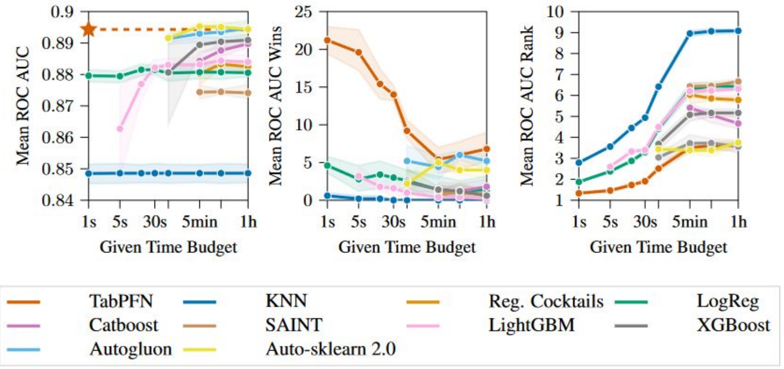


Figure 5: ROC AUC performance over time. We report mean ROC, mean wins and rank along with the 95% confidence interval across 5 repetitions for different time budgets (Unlabelled ticks: 1min, 15min).

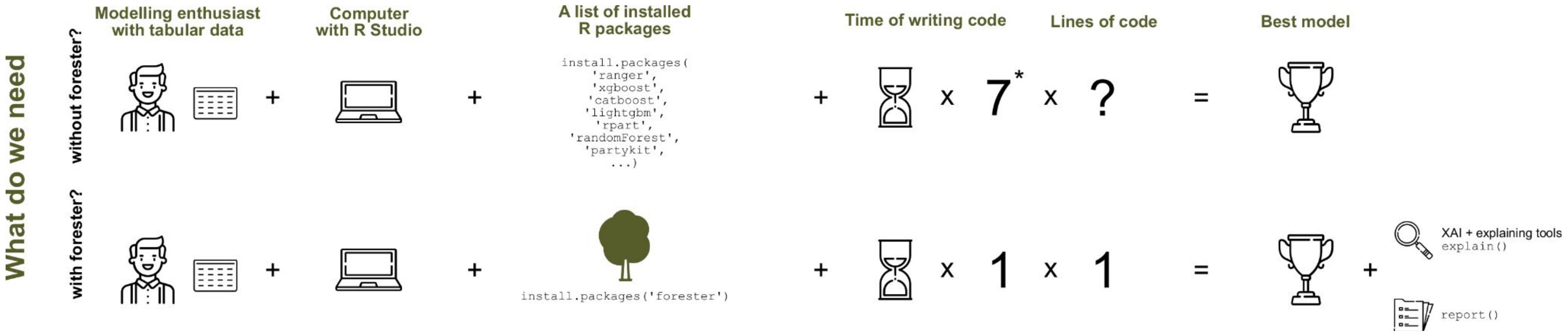
N. Hollmann, S. Müller, K. Eggensperger, and F. Hutter, *TabPFN: A Transformer That Solves Small Tabular Classification Problems in a Second*, 2022

The forester R package

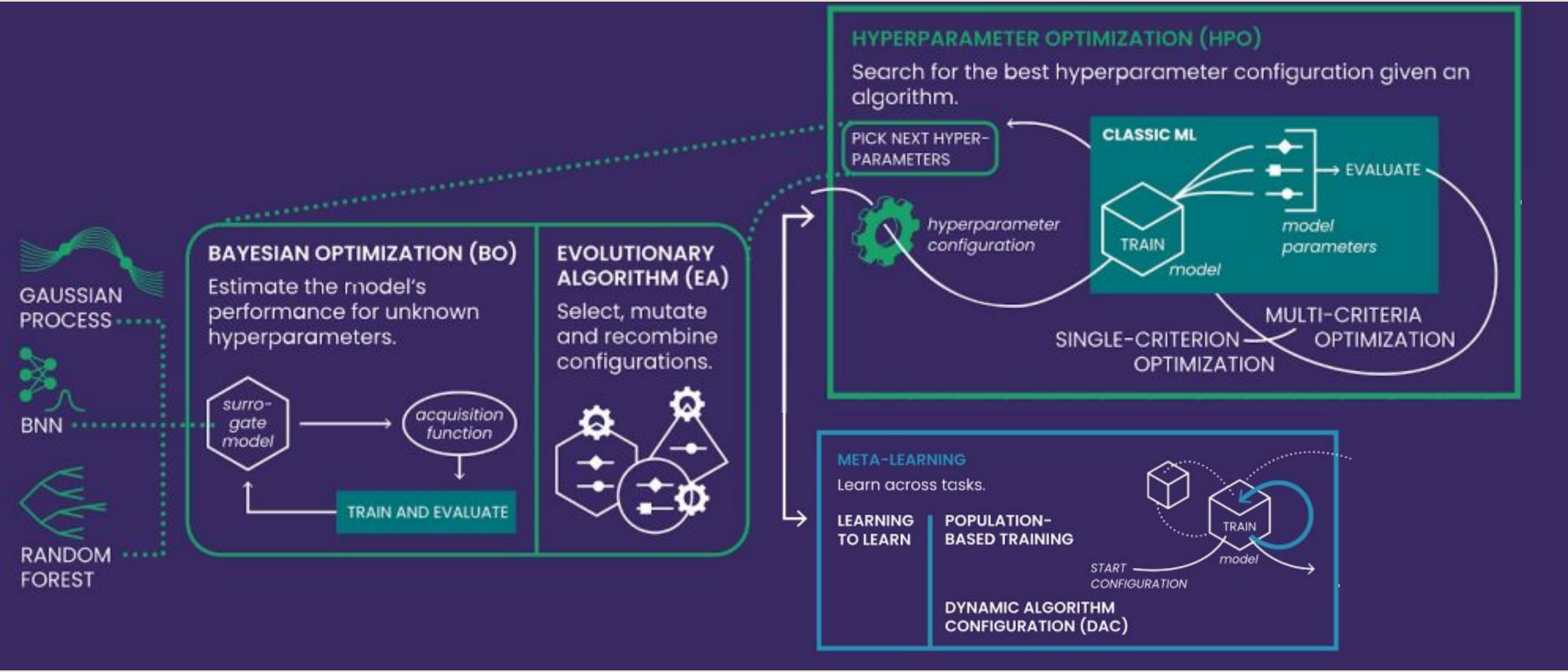
What is **forester**?

- ➔ full automation of the process of training tree-based models
- ➔ no demand for ML expertise
- ➔ powerful tool for making high-quality baseline models for experienced users

! The forester package is **designed for beginners** in data science, but also for more experienced users.



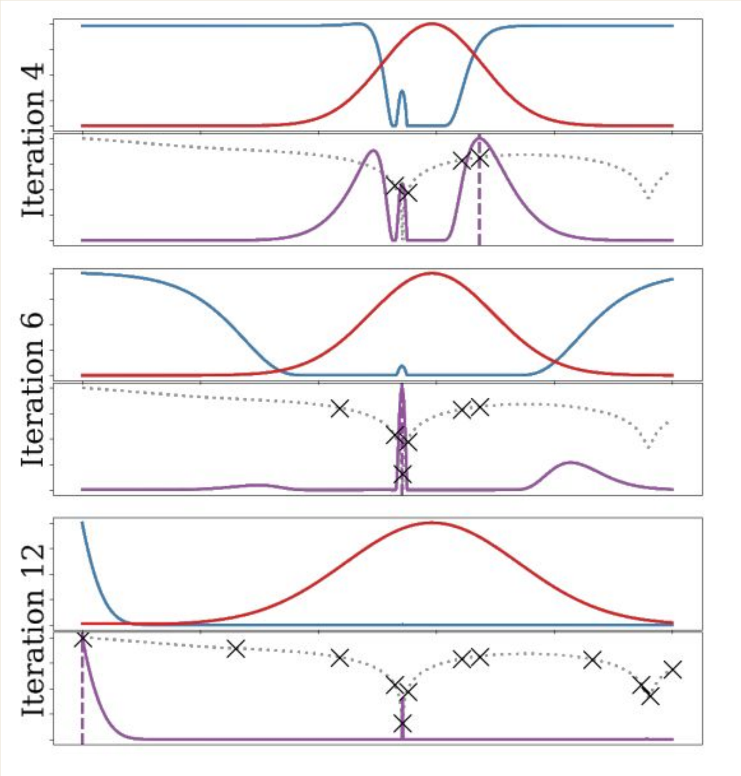
CASH & HPO (for tabular data and classic ML)



CASH & HPO & Meta-learning

Bayesian Optimization with Priors

$$\boldsymbol{x}_n \in \underset{\boldsymbol{x}\in\mathcal{X}}{\arg\max}\, \alpha_\pi(\boldsymbol{x}, \mathcal{D}_n) = \underset{\boldsymbol{x}\in\mathcal{X}}{\arg\max}\, \alpha(\boldsymbol{x}, \mathcal{D}_n)\pi(\boldsymbol{x})^{\beta/n}$$

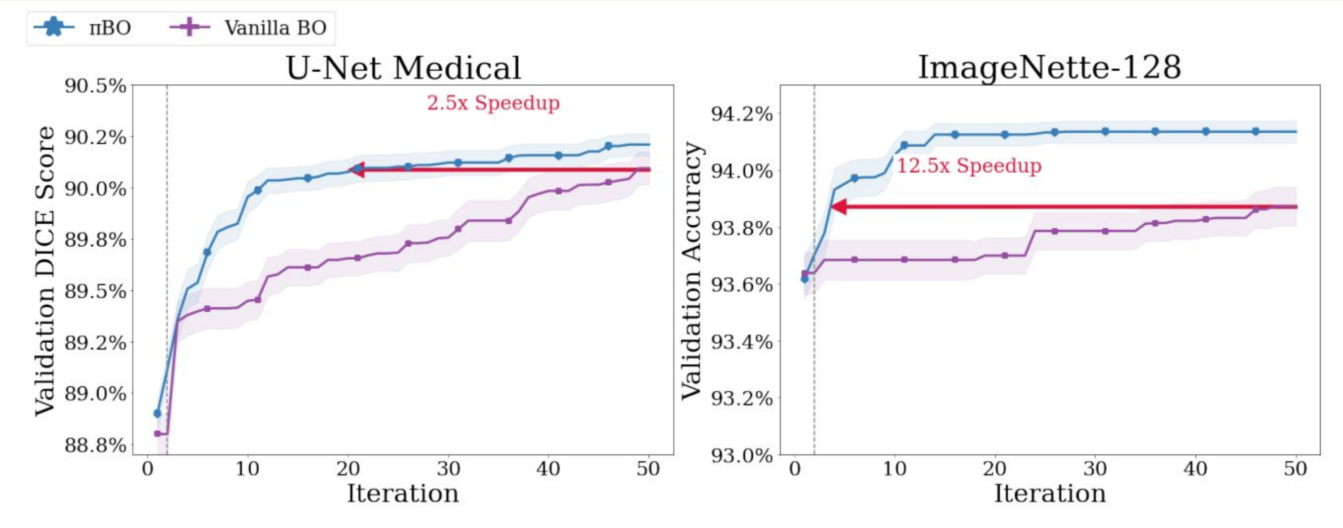


Consolidated learning @Katarzyna.Woźnica

Hvarfner, C., Stoll, D., Souza, A., Lindauer, M., Hutter, F. and Nardi, L. *piBO: Augmenting Acquisition Functions with User Beliefs for Bayesian Optimization.* 2022

CASH & HPO & Meta-learning

Bayesian Optimization with Priors



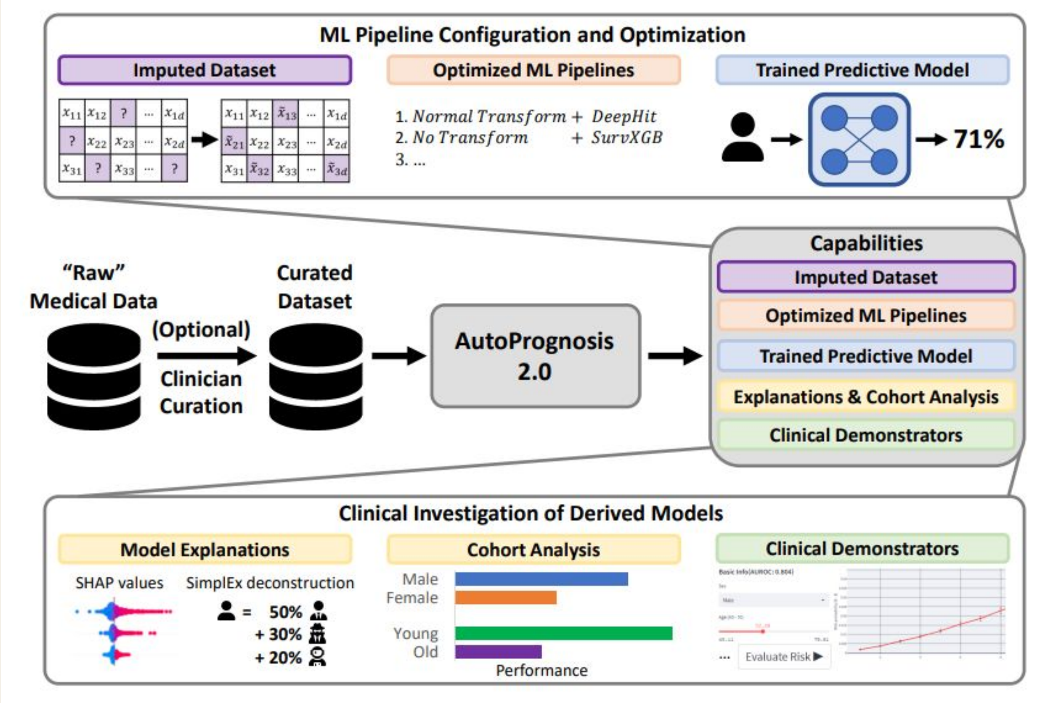
Consolidated learning @Katarzyna.Woźnica

Hvarfner, C., Stoll, D., Souza, A., Lindauer, M., Hutter, F. and Nardi, L. *piBO: Augmenting Acquisition Functions with User Beliefs for Bayesian Optimization.* 2022

CASH & HPO & Meta-learning

AutoPrognosis 1.0 and 2.0

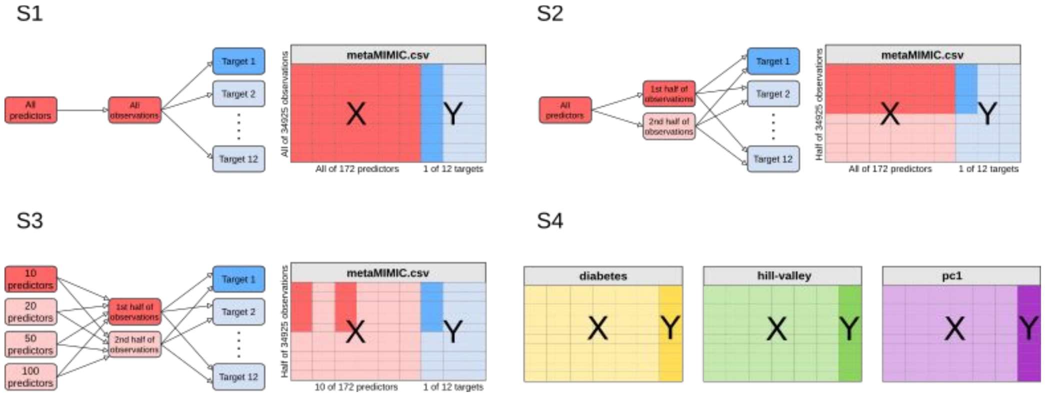
Patient data:
- UK BioBank
- UNOS
- MAGGIC
- SEER



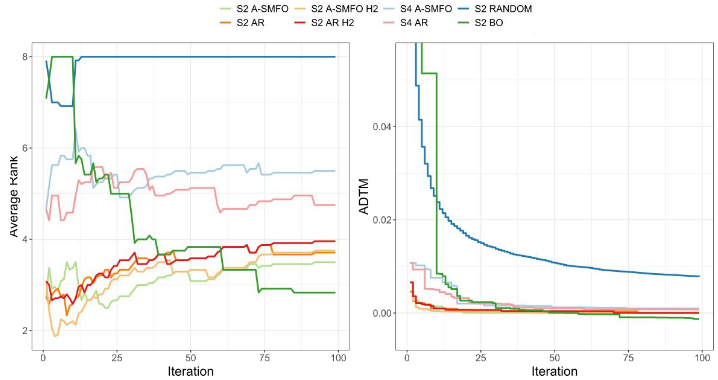
A M. Alaa and M. van der Schaar, *AutoPrognosis: Automated Clinical Prognostic Modeling via Bayesian Optimization with Structured Kernel Learning*, 2018

Consolidated learning

- **Domain-specific meta-train collection regarding prior knowledge**
- **metaMIMIC benchmark**

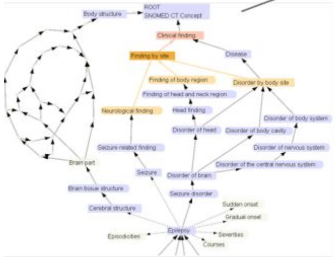


- the definition-based similarity of tasks is positively related to hyperparameters' transferability between them.



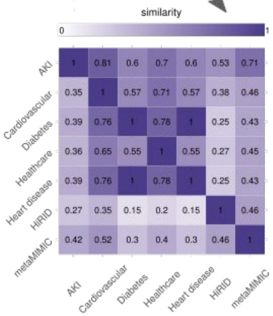
K. Woźnica, M. Grzyb, Z. Trafas, and P. Biecek, *Consolidated learning - a domain-specific model-free optimization strategy with examples for XGBoost and MIMIC-IV*, 2022

Ontology-based semantic meta-features



SnoMED CT hierarchy (visualization)

Lin similarity between terms



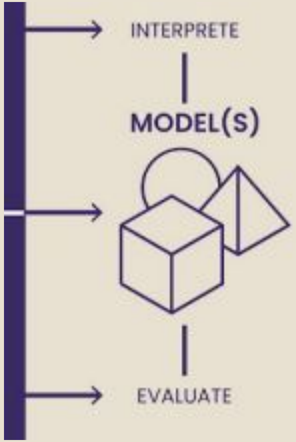
Similarity between datasets

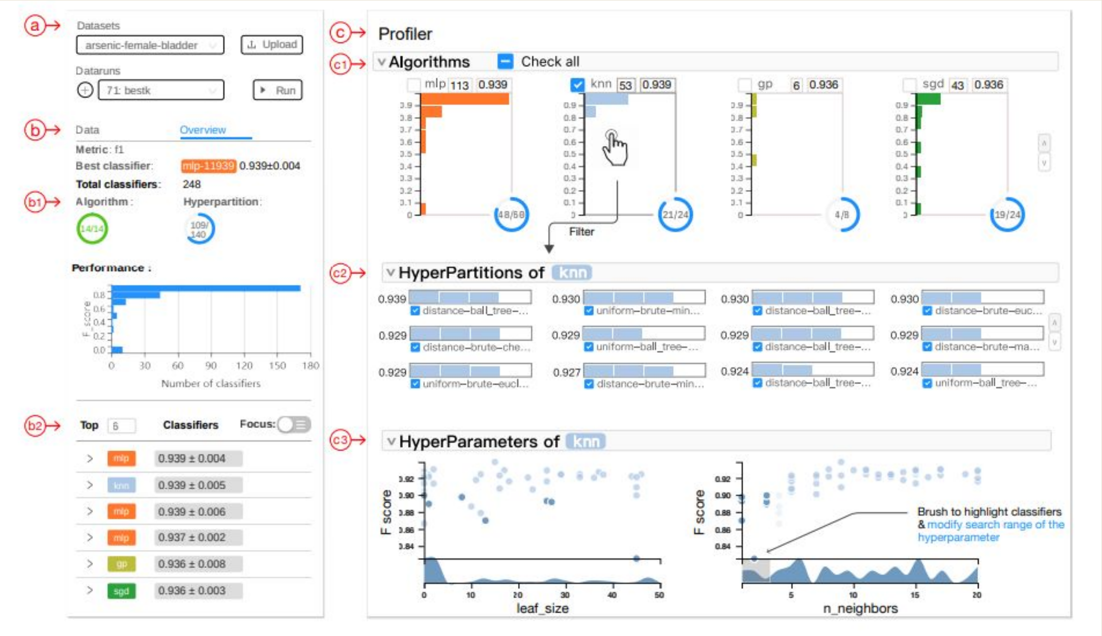
Hypothesis: semantic similarity of variables helps in meta-learning

➔ SnoMED annotated healthcare datasets (metaMIMIC + kaggle)

K. Woźnica, P.Wilczyński, and P. Biecek

ATMSeer

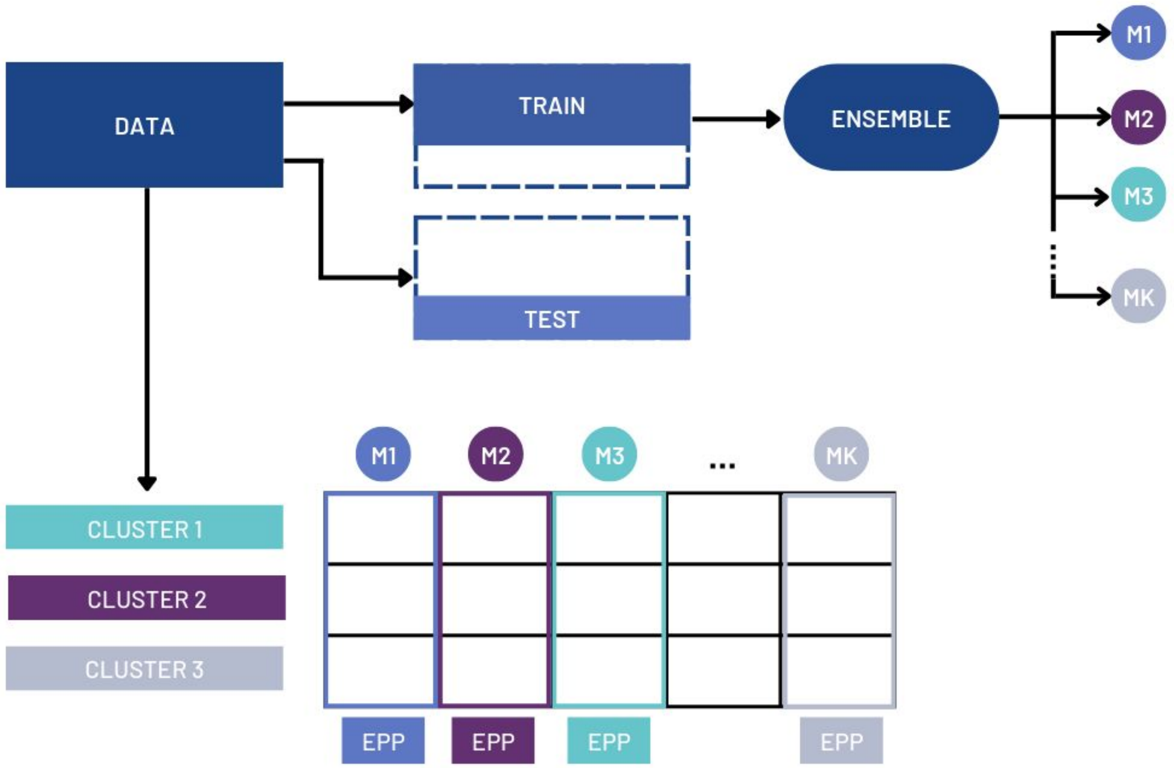




Q. Wang, Y. Ming Z. Jin, Q. Shen, D. Liu, M. J. Smith, K.Veeramachaneni and H. Qu, ATMSeer: Increasing Transparency and Controllability in Automated Machine Learning, 2019

EPP++ @Katarzyna.Woźnica

EPP++


DATA
TRAIN
TEST
ENSEMBLE
M1
M2
M3
MK
CLUSTER 1
CLUSTER 2
CLUSTER 3
M1
M2
M3
...
MK
EPP
EPP
EPP
EPP

Questions?

Challenges

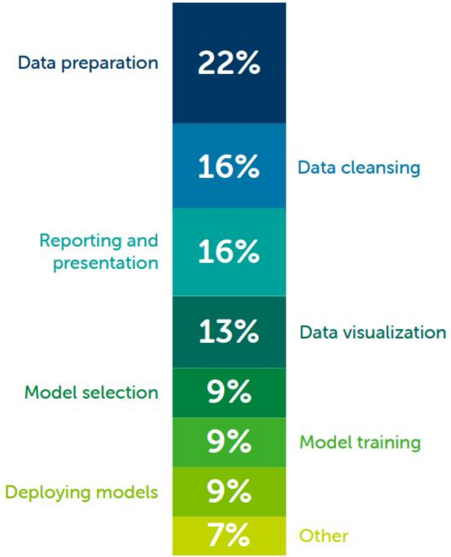
Two perspectives

A. **Automating Data Science** - Tijl De Bie, Luc De Raedt, José Hernández-Orallo, Holger Hoos, Padhraic Smyth, Christopher Williams

B. **Rethinking AutoML: Advancing from a Machine-Centered to Human-Centered Paradigm** - Marius Lindauer & Alexander Tornede

B.

AutoML actually only covers a rather small portion of the data science workflow and thus is only of limited use in practice.

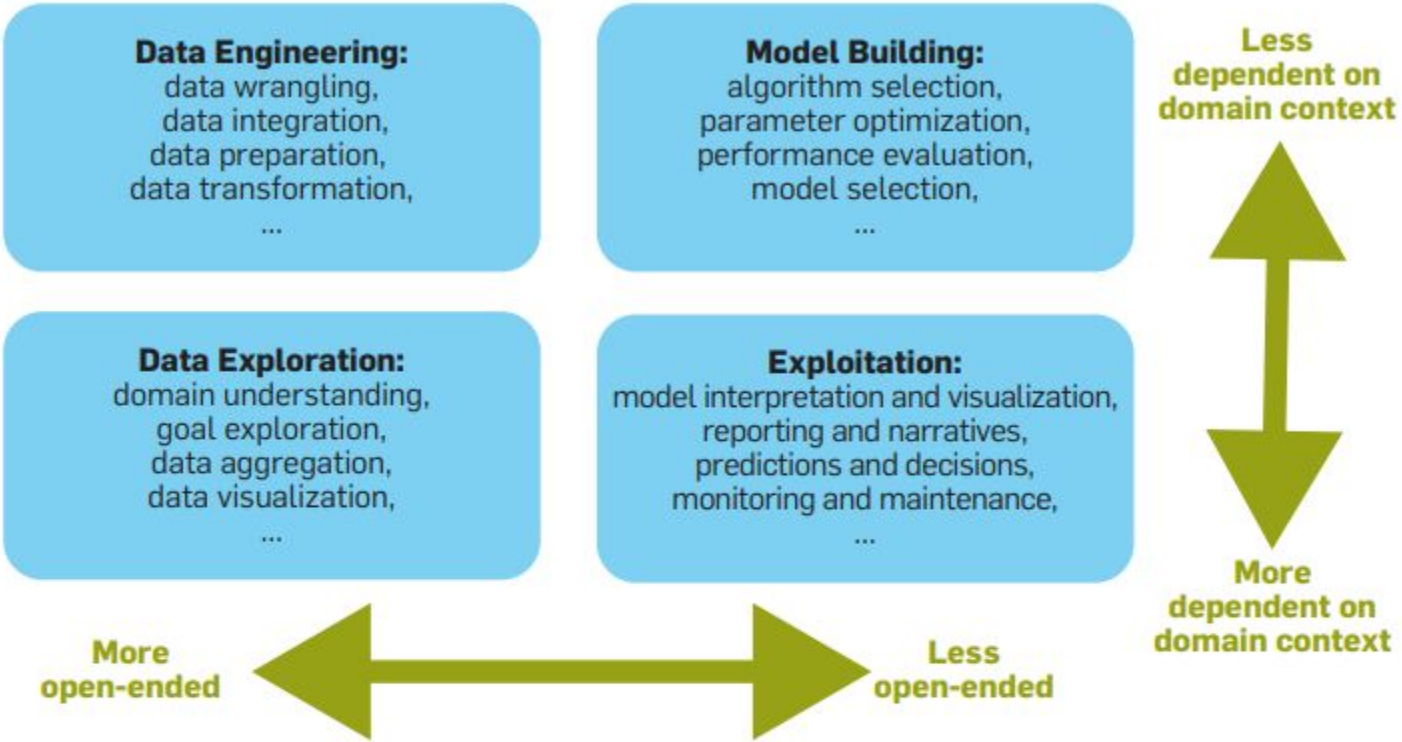


n = 1,966

We asked our respondents how much time they spend on the above tasks, and for each item they entered a number reflecting the percentage of time spent relative to the other options. This is the average of the reported percentages.

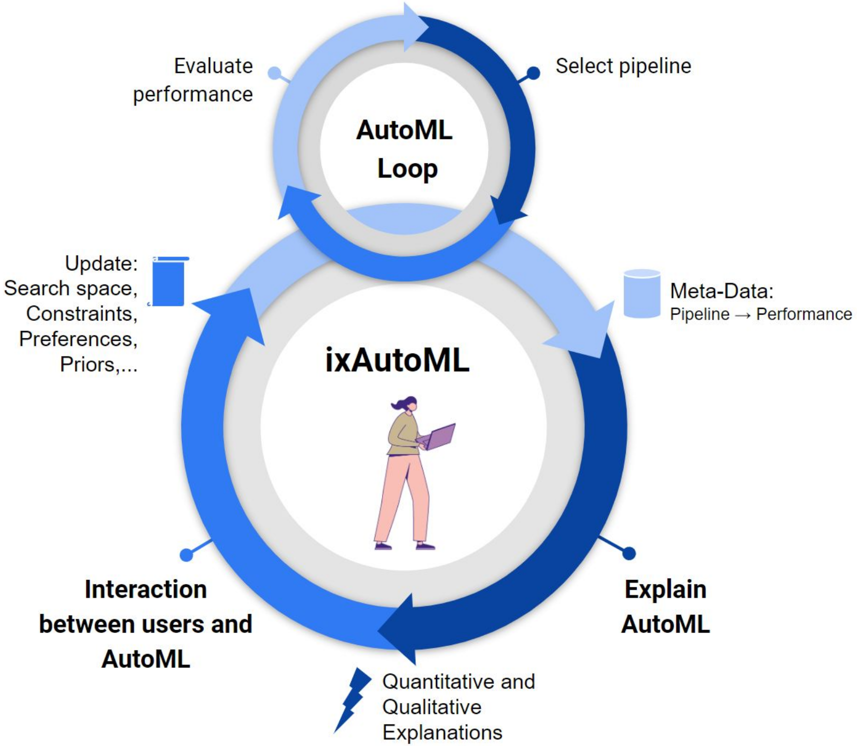
Automated Data Science

A.



Extension of target audience to data scientists

Both the internal process of AutoML tools and how their final result was constructed is often hard to understand, even for AutoML experts, let alone data scientists, leading to a lack of trust in AutoML systems.



EPP++ @Katarzyna.Woźnica

Human-Centered AutoML

- ➔ Increasing the efficiency of AutoML by making use of the best of both worlds: a systematic search of efficient AutoML approaches and human expertise and intuition;
- ➔ A human-in-the-loop AutoML framework that is tailored to the needs of data scientists and thus leading to a more wide spread use of it;
- ➔ Insights into the design of ML applications and thus accelerating research on ML by reproducible and insightful tools;