A Decade of AutoML: Reflections and the Road Ahead

KDD 2023 Test of Time Award Presentation

Chris Thornton **Frank Hutter** **Holger Hoos** **Kevin Leyton-Brown**

University of Freiburg
fh@cs.uni-freiburg.de

@FrankRHutter
@AutoML_org

The authors of the KDD 2013 paper: then and now

Auto-WEKA: Combined Selection and Hyperparameter Optimization of Classification Algorithms

Chris Thornton Frank Hutter Holger H. Hoos Kevin Leyton-Brown

Department of Computer Science, University of British Columbia

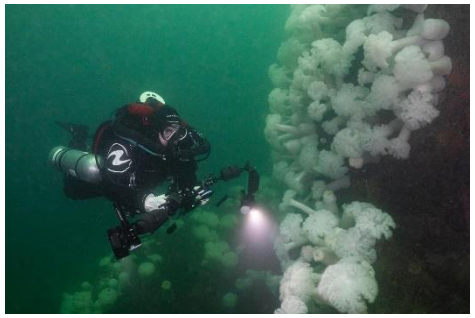
Vancouver, Canada

Chris Thornton

Frank Hutter

Holger Hoos

Kevin Leyton-Brown



University of Freiburg

RWTH Aachen

UBC

The problem we saw in 2013

- Excellent OSS ML software: **WEKA** [Witten et al, since 1999]
  - GUI-based, particularly popular with **novice users**
  - ML growing quickly, already 18k citations for WEKA

- But novice users have a problem: **which method to use?**
  - 39 different classification methods
  - With up to 8 hyperparameters each

Classifier	Categorical	Numeric
Bayes Net	2	0
Naive Bayes	2	0
Naive Bayes Multinomial	0	0
Gaussian Process	3	6
Linear Regression	2	1
Logistic Regression	0	1
Single-Layer Perceptron	5	2
Stochastic Gradient Descent	3	2
SVM	4	6
Simple Linear Regression	0	0
Simple Logistic Regression	2	1
Voted Perceptron	1	2
KNN	4	1
K-Star	2	1
Decision Table	4	0
RIPPER	3	1
M5 Rules	3	1
1-R	0	1
PART	2	2
0-R	0	0
Decision Stump	0	0
C4.5 Decision Tree	6	2
Logistic Model Tree	5	2
M5 Tree	3	1
Random Forest	2	3
Random Tree	4	4
REP Tree	2	3
Locally Weighted Learning*	3	0
AdaBoost M1*	2	2
Additive Regression*	1	2
Attribute Selected*	2	0
Bagging*	1	2
Classification via Regression*	0	0
LogitBoost*	4	4
MultiClass Classifier*	3	0
Random Committee*	0	1
Random Subspace*	0	2
Voting+	1	0
Stacking+	0	0

- **Goals of AutoML**
  - **Democratize ML**: allow anyone to achieve SOTA ML
  - **Productivity tool** for experts

Our background on automated algorithm configuration in 2013

- **Automated algorithm configuration**: finding an algorithm's best setting
  - Topic of my PhD thesis [2009]

- **Many successful fielded applications** in combinatorial optimization
  - E.g., SAT: **500x speedup for software verification** by tuning **26 parameters** [FMCAD, 2017]
  - E.g., MIP: **50x speedups of CPLEX** solver by tuning **61 parameters** [CPAIOR, 2020]

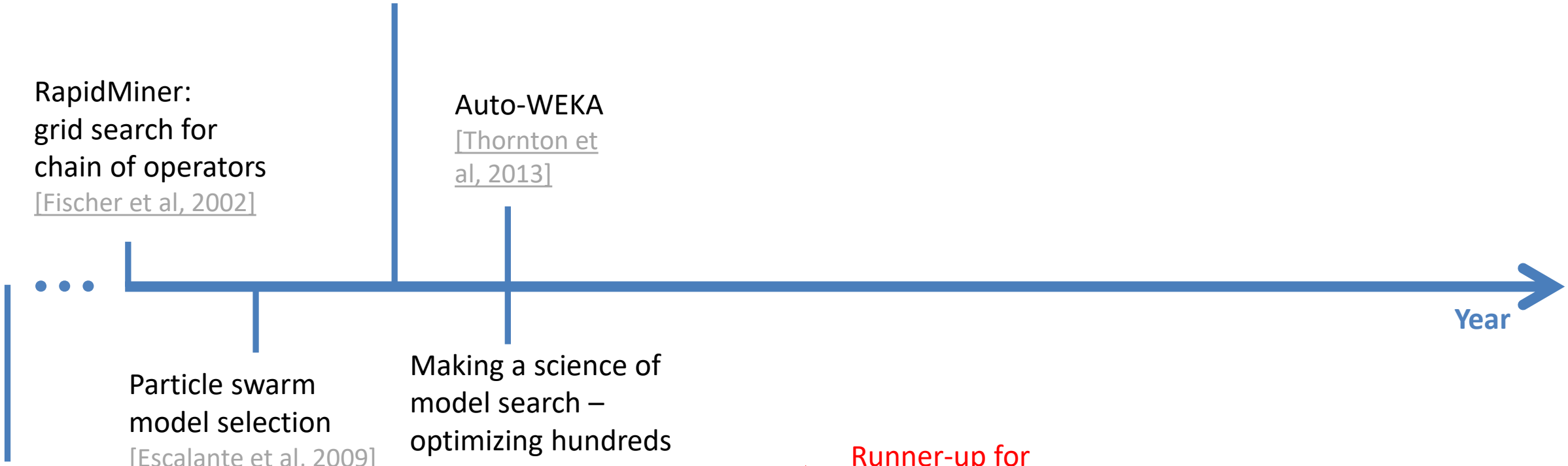
- Had just developed **SMAC: high-dimensional Bayesian optimization**
  - Already ML on the meta-level; obvious to also apply to ML on the base level

Related work on AutoML systems in 2013

SMAC [Hutter et al, 2011]
Spearmint [Snoek et al, 2011]
TPE [Bergstra et al, 2012]

RapidMiner:
grid search for
chain of operators
[Fischer et al, 2002]

Auto-WEKA
[Thornton et
al, 2013]



Particle swarm
model selection
[Escalante et al, 2009]

Making a science of
model search –
optimizing hundreds
of dimensions of
vision architectures
[Bergstra et al, 2013]

Runner-up for
ICML Test of Time Award

1990s: many works
on neuro-evolution

Model selection & hyperparameter optimization

- **Model selection** (with cross-validation)

$$A^* \in \operatorname*{argmin}_{A \in \mathcal{A}} \frac{1}{k} \sum_{i=1}^{k} \underbrace{\mathcal{L}(A, \mathcal{D}_{\text{train}}^{(i)}, \mathcal{D}_{\text{valid}}^{(i)})}$$

Loss of algorithm A on
i-th cross-validation fold

- **Hyperparameter optimization**

$$\boldsymbol{\lambda}^* \in \operatorname*{argmin}_{\boldsymbol{\lambda} \in \boldsymbol{\Lambda}} \frac{1}{k} \sum_{i=1}^{k} \underbrace{\mathcal{L}(A_{\boldsymbol{\lambda}}, \mathcal{D}_{\text{train}}^{(i)}, \mathcal{D}_{\text{valid}}^{(i)})}$$

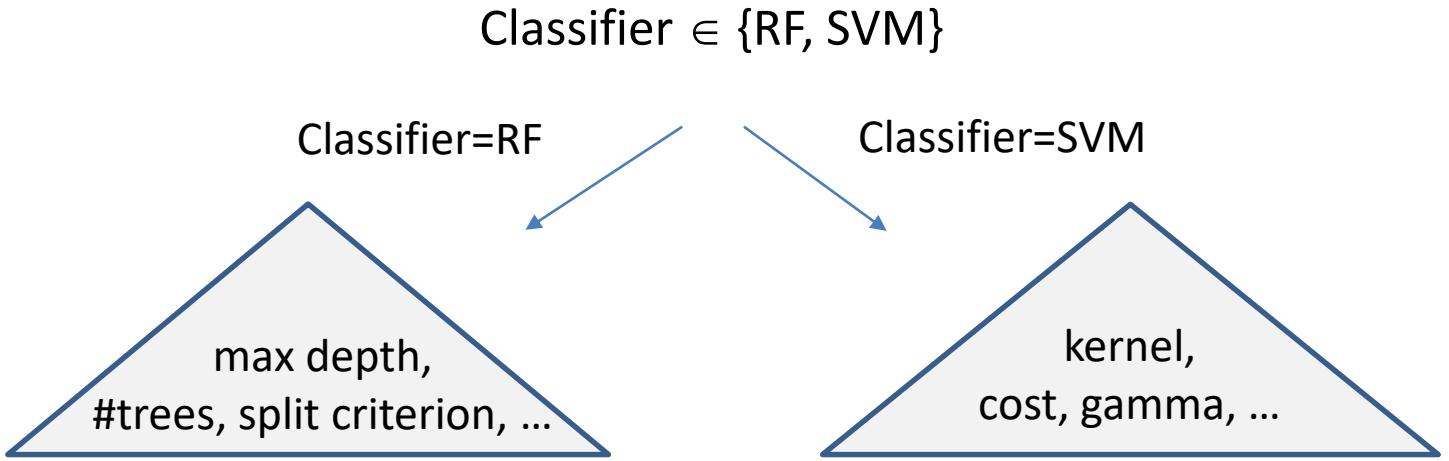
Loss of algorithm A with
hyperparameters λ on
i-th cross-validation fold

Technical innovation: tackling these problems jointly

- **Combined Algorithm Selection and Hyperparameter optimization (CASH)**

$$A^*_{\boldsymbol{\lambda}^*} \in \underset{A^{(j)} \in \mathcal{A}, \boldsymbol{\lambda} \in \boldsymbol{\Lambda}^{(j)}}{\operatorname{argmin}} \frac{1}{k} \sum_{i=1}^{k} \underbrace{\mathcal{L}(A^{(j)}_{\boldsymbol{\lambda}}, \mathcal{D}^{(i)}_{\text{train}}, \mathcal{D}^{(i)}_{\text{valid}})}_{\text{Loss of algorithm } A^{(j)} \text{ with hyperparameters } \lambda \text{ on i-th cross-validation fold}}$$

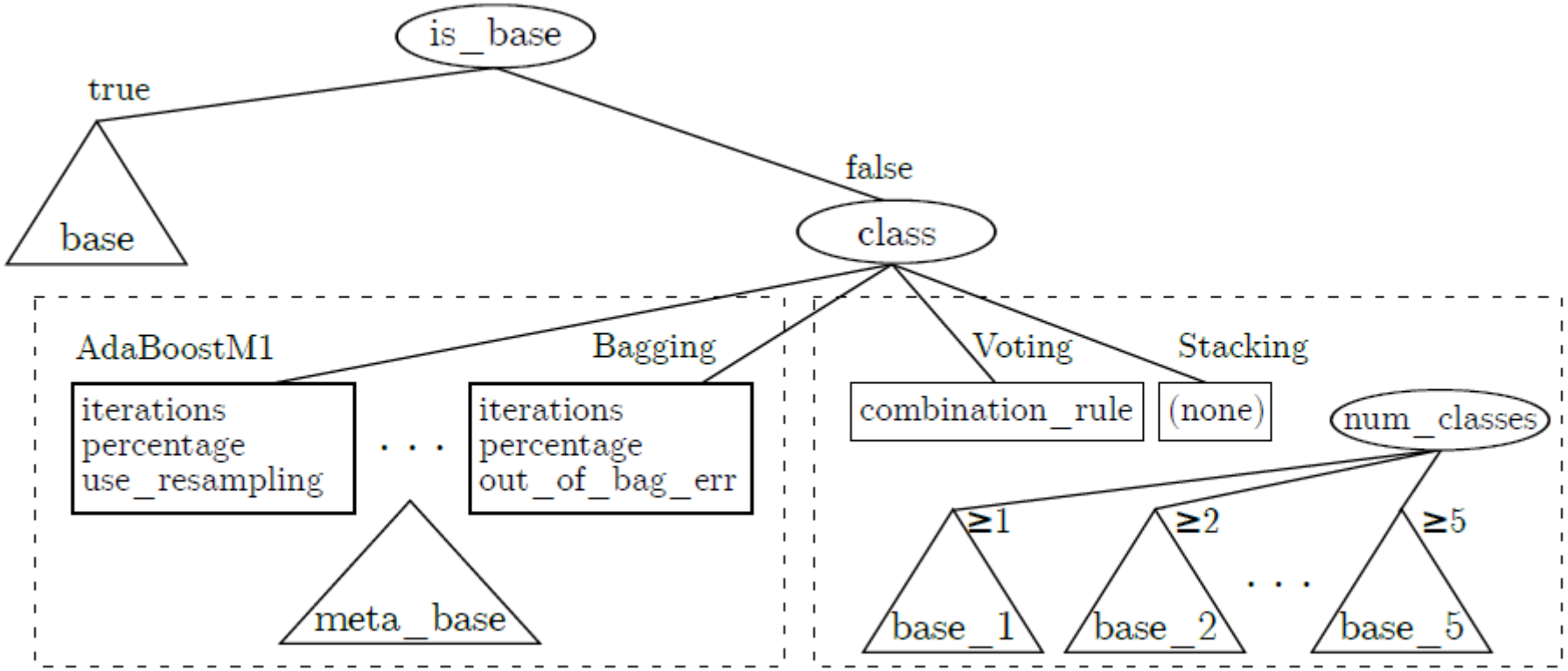
- This can be written as a **single HPO problem with hierarchical dependencies**



Auto-WEKA's configuration space

- **Base classifier**
  - 27 choices, each with their own hyperparameters

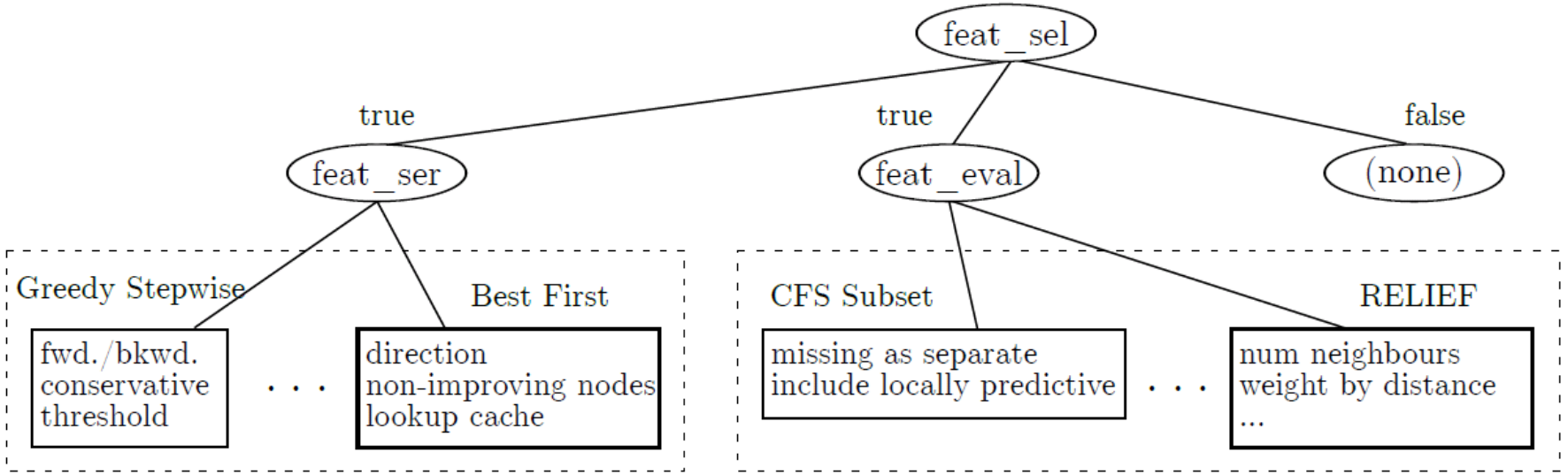
- **Hierachical structure on top of base classifiers**



Auto-WEKA's configuration space

- **Feature selection**
  - **Search** method: which feature subset
  - **Evaluation** method: how to evaluate feature subsets in that search
  - All of these methods have further hyperparameters

- In total: **768 hyperparameters, 10^{47} configurations**

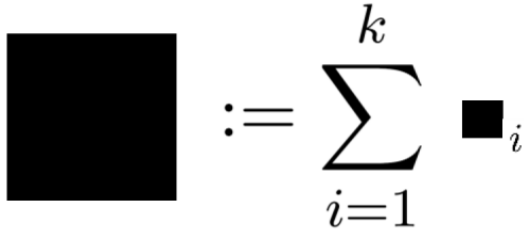


Methods for CASH

- **Baselines**
  - Model selection only (default hyperparameters)
  - Random search: union over grids of 27 base classifiers

- **Auto-WEKA: Bayesian optimization with tree-based models**
  - Variant 1: BO with tree-structured kernel density estimator, **TPE** [Bergstra et al, '11]
  - Variant 2: BO with random forest model, **SMAC** [Hutter et al, '11]

- **Speedup technique for optimizing CV in SMAC**
  - Skip configuration if results too poor on first folds



Experimental Setup

- **21 prominent datasets** from the ML literature
  - Many datasets from UCI
  - CIFAR-10
  - MNIST
  - KDD-Cup
- Data split into **training & test**
- Running Auto-WEKA
  - Training time: **30 hours on 4 cores**
  - Optimize cross-validation on the training portion
  - Evaluate on the test portion

Empirical results

- **Auto-WEKA outperforms best base classifier**
  - Even if best base classifier is chosen on test set
  - In 6/21 cases: **over 10% relative improvements**
- **Auto-WEKA vs random grid search**
  - **100 times faster**
  - CV: **Better in 20/21 cases**
- Within Auto-WEKA, **SMAC vs TPE**
  - CV: **SMAC better in 19/21 cases**, 1 tie
  - Difference usually small,
    only substantial in few cases (SMAC better)

	Rand. Grid	Auto-WEKA	
		TPE	SMAC
Dexter	7.48	9.90	**5.48**
GermanCredit	22.45	21.43	**19.59**
Dorothea	6.03	6.93	**5.52**
Yeast	38.87	**35.03**	36.27
Amazon	**43.94**	48.43	48.30
Secom	6.12	6.25	**5.34**
Semeion	6.52	6.91	**4.86**
Car	1.54	0.94	**0.71**
Madelon	24.26	24.26	**20.87**
KR-vs-KP	0.70	0.45	**0.32**
Abalone	72.45	72.20	**71.76**
Wine Quality	37.28	35.94	**34.74**
Waveform	12.73	12.57	**11.71**
Gisette	3.27	3.70	**2.42**
Convex	28.50	29.04	**24.70**
CIFAR-10-Small	65.11	57.97	**57.76**
MNIST Basic	4.00	13.64	**3.64**
Rot. MNIST + BI	59.75	73.04	**59.61**
Shuttle	0.0263	**0.0230**	**0.0230**
KDD09-Appentency	1.88	1.88	**1.75**
CIFAR-10	65.54	66.68	**63.21**

Follow-up: Auto-WEKA 2.0 JMLR MLOSS 2016 tool paper

Auto-WEKA 2.0: Automatic model selection and hyperparameter optimization in WEKA

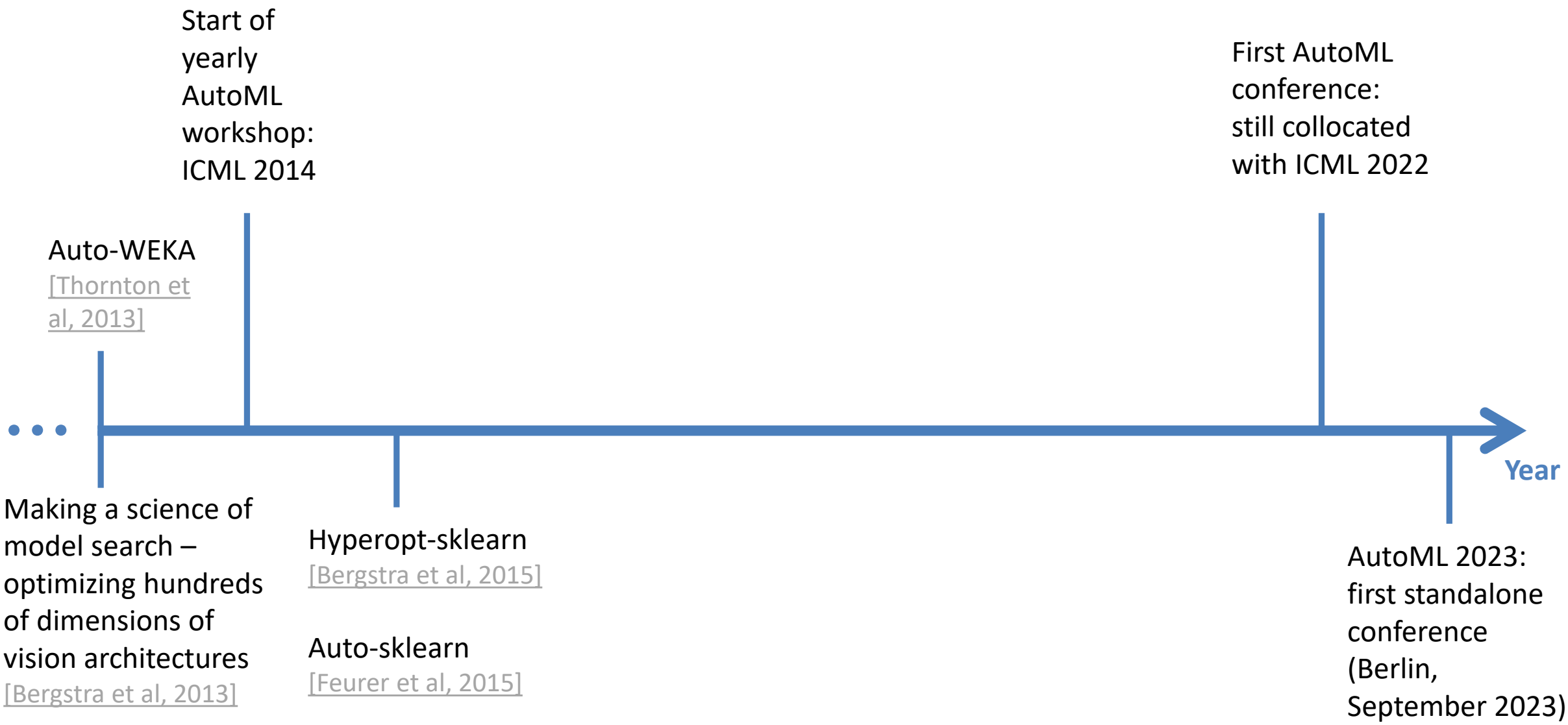
Lars Kotthoff LARSKO@CS.UBC.CA
Chris Thornton CWTHORNT@CS.UBC.CA
Holger H. Hoos HOOS@CS.UBC.CA
Frank Hutter FH@CS.UNI-FREIBURG.DE
Kevin Leyton-Brown KEVINLB@CS.UBC.CA

Lars Kotthoff

- Regression
- Parallelization
- **Fully integrated with WEKA**: available through WEKA's package manager
  - Over 1000 downloads per month

What happened since?

AutoML: A Timeline



Auto-sklearn (NeurIPS 2015)

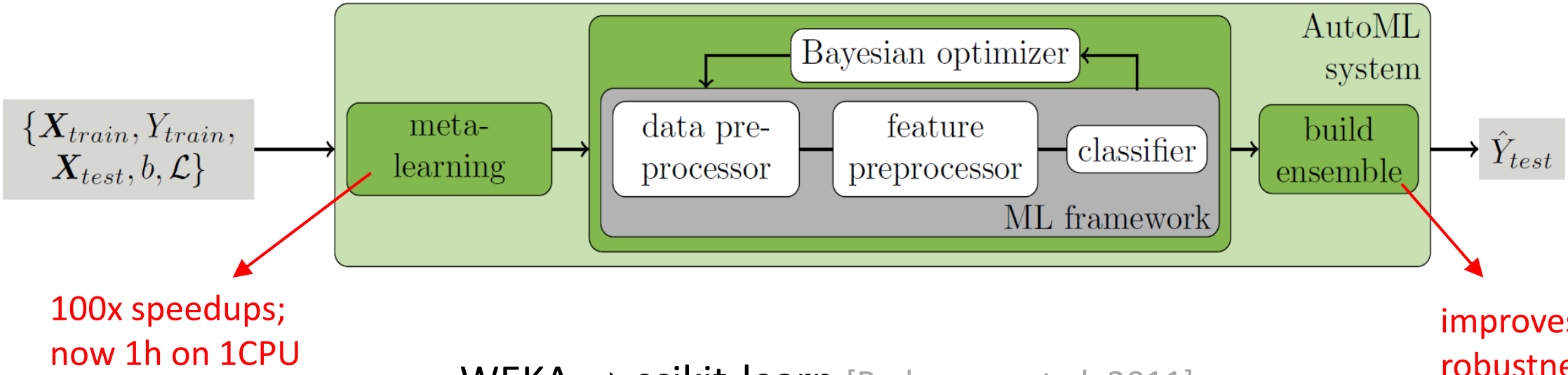
Efficient and Robust Automated Machine Learning

Matthias Feurer **Aaron Klein** **Katharina Eggensperger**
Jost Tobias Springenberg **Manuel Blum** **Frank Hutter**
Department of Computer Science
University of Freiburg, Germany
{feurerm,kleinaa,eggenspk,springj,mblum,fh}@cs.uni-freiburg.de

Matthias Feurer

Aaron Klein

Katharina Eggensperger

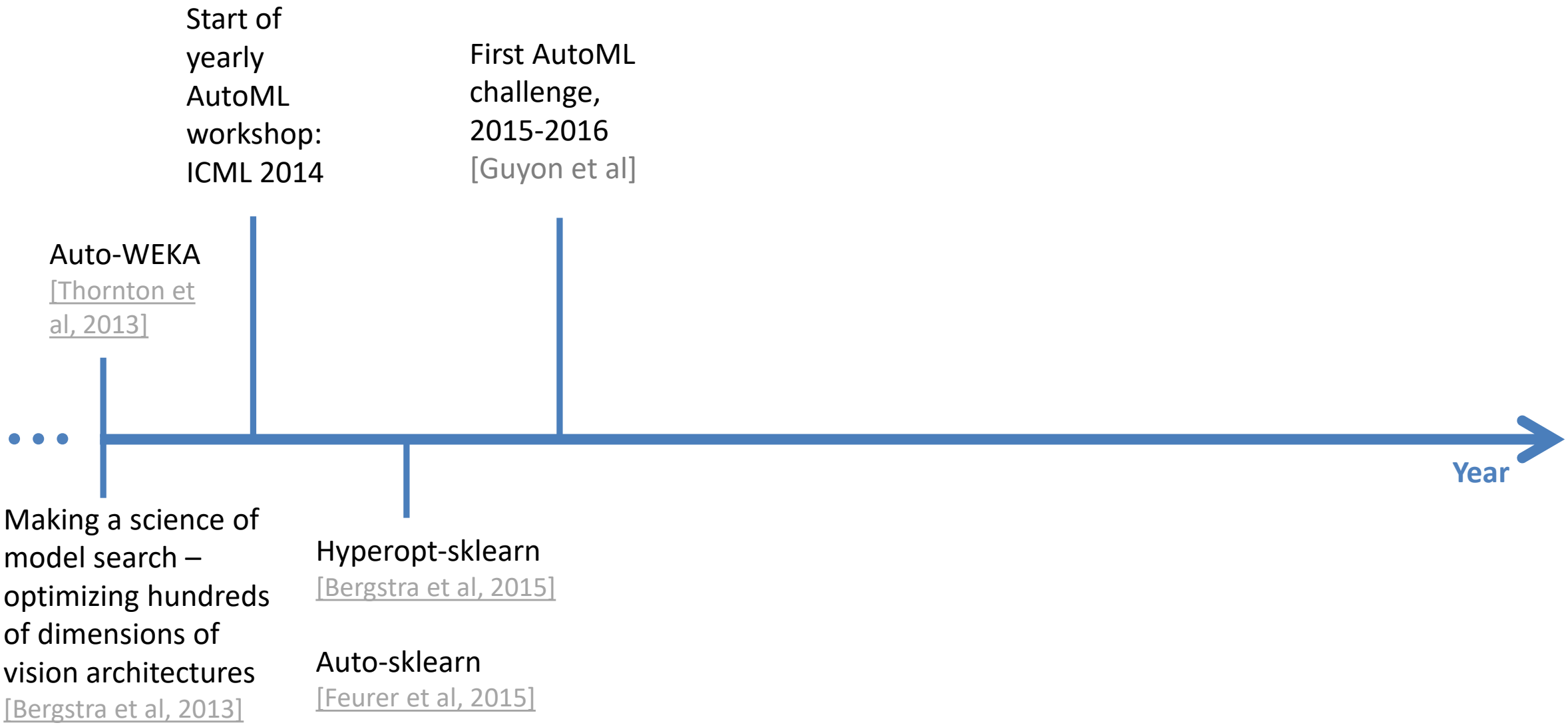


100x speedups;
now 1h on 1CPU

improves robustness

WEKA → scikit-learn [Pedregosa et al, 2011]

AutoML: A Timeline



The AutoML challenge

- A 14-month long competition that allowed us fully build Auto-sklearn

- 2 tracks with 5 phases each
  - Tweakathon (Kaggle-like) with up to 130 human teams
  - AutoML system track
  - Ever more challenging datasets

- In the last 2 phases, **Auto-sklearn won both tracks**
  - https://github.com/automl/auto-sklearn, 7k stars
  - Now **> 25k downloads per month** on PyPI

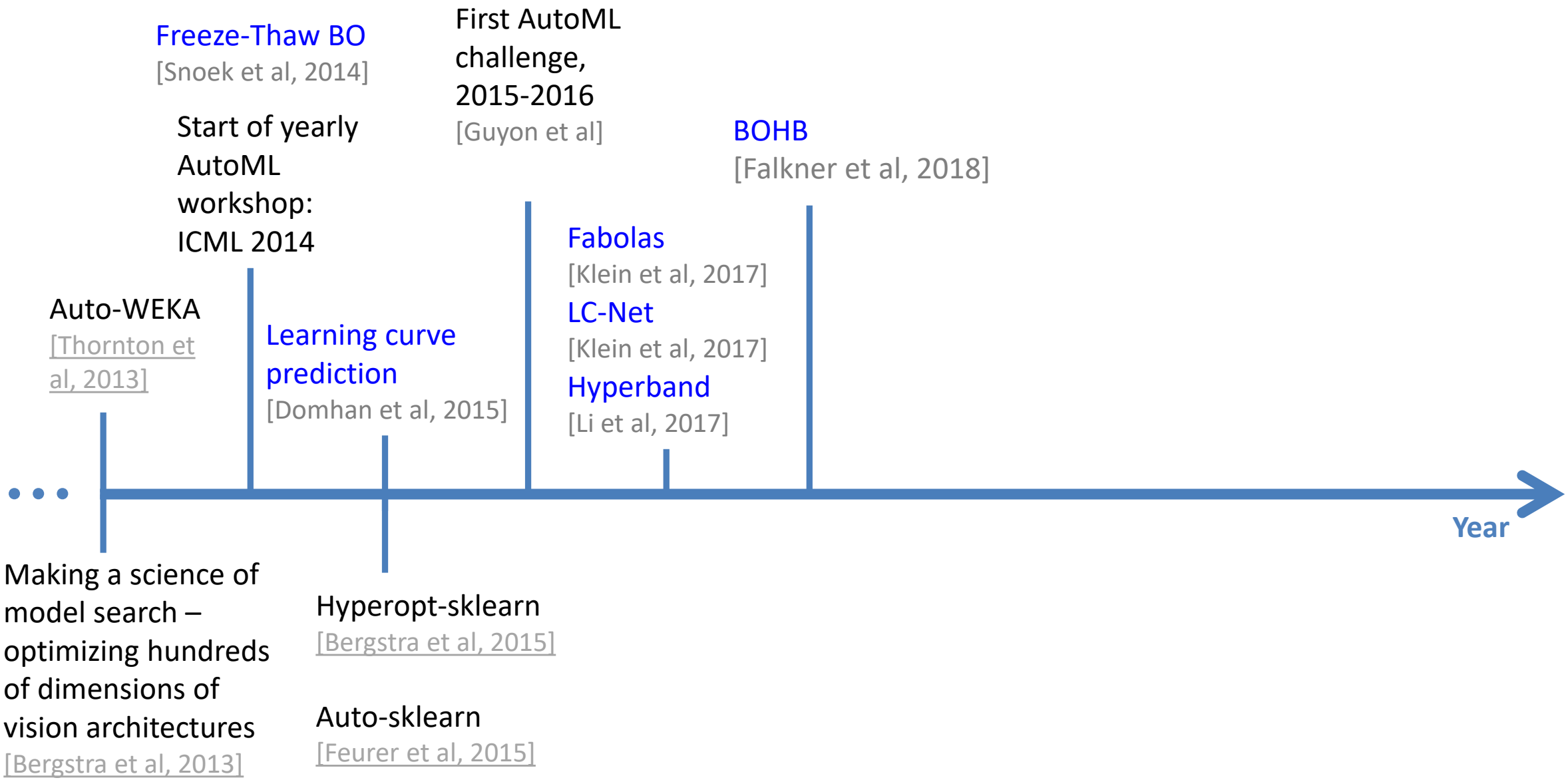
Huge thanks to Isabelle for running this pivotal challenge!

Isabelle Guyon

Dinner with the winners of the competition at ICML 2015

Towards AutoDL:
Multi-Fidelity Optimization for Greater Efficiency

AutoML: A Timeline

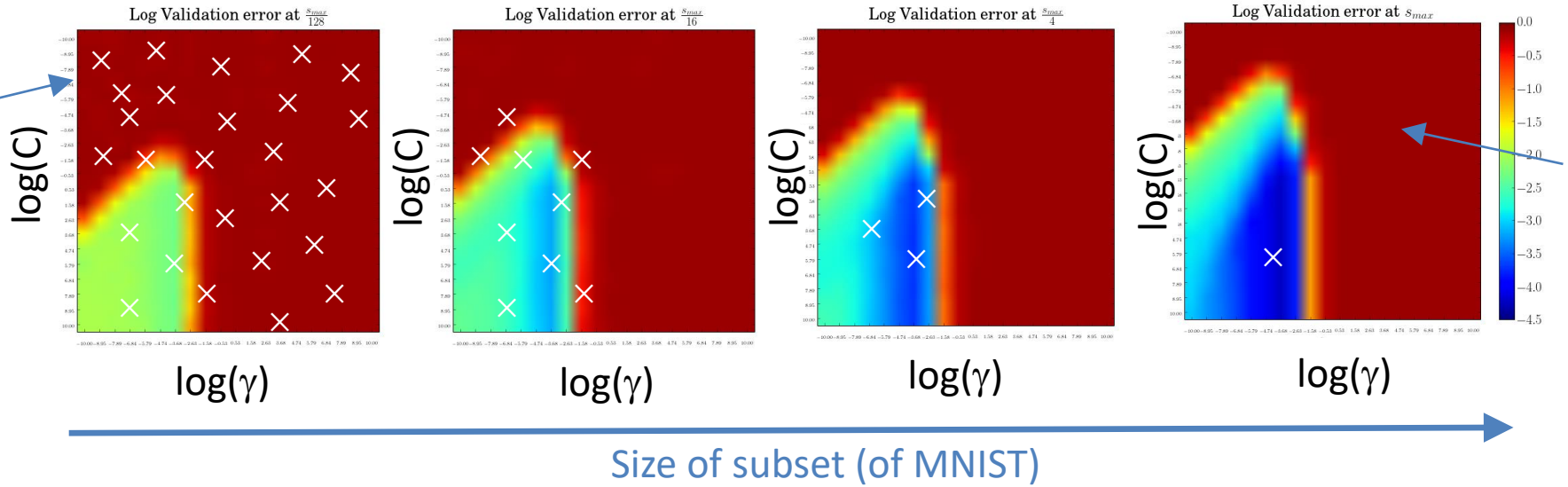


Multi-Fidelity Optimization

- **Use cheap proxies, e.g. subset of the data**
  - Do most of the work on these cheap proxies

- E.g.: Support Vector Machine (SVM) on MNIST dataset (hyperparameters: C, γ)

Log validation error (red:bad; green:OK; blue:good) when trained on subset of 390 data points (takes few seconds to compute)



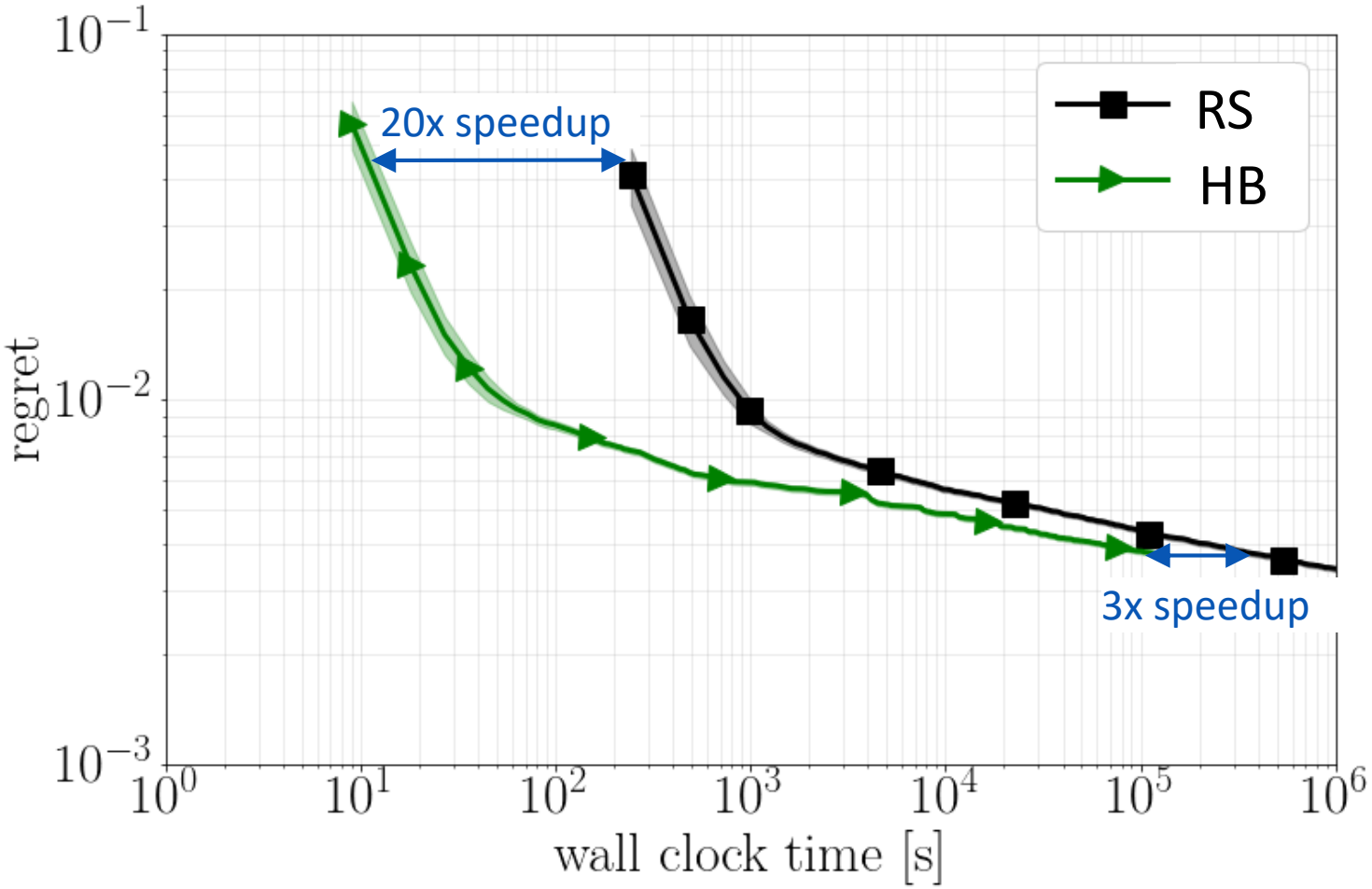
Log validation error when trained on full dataset of 50.000 data points (takes several hours to compute)

→ **up to 1000x speedups over blackbox optimization on full data** [Klein et al, AISTATS 2017]

→ **similar ideas: Successive Halving & Hyperband** [Li et al, ICLR 2017]

Empirical Evaluation: Hyperband vs. Random Search

[Falkner, Klein & Hutter, ICML 2018]

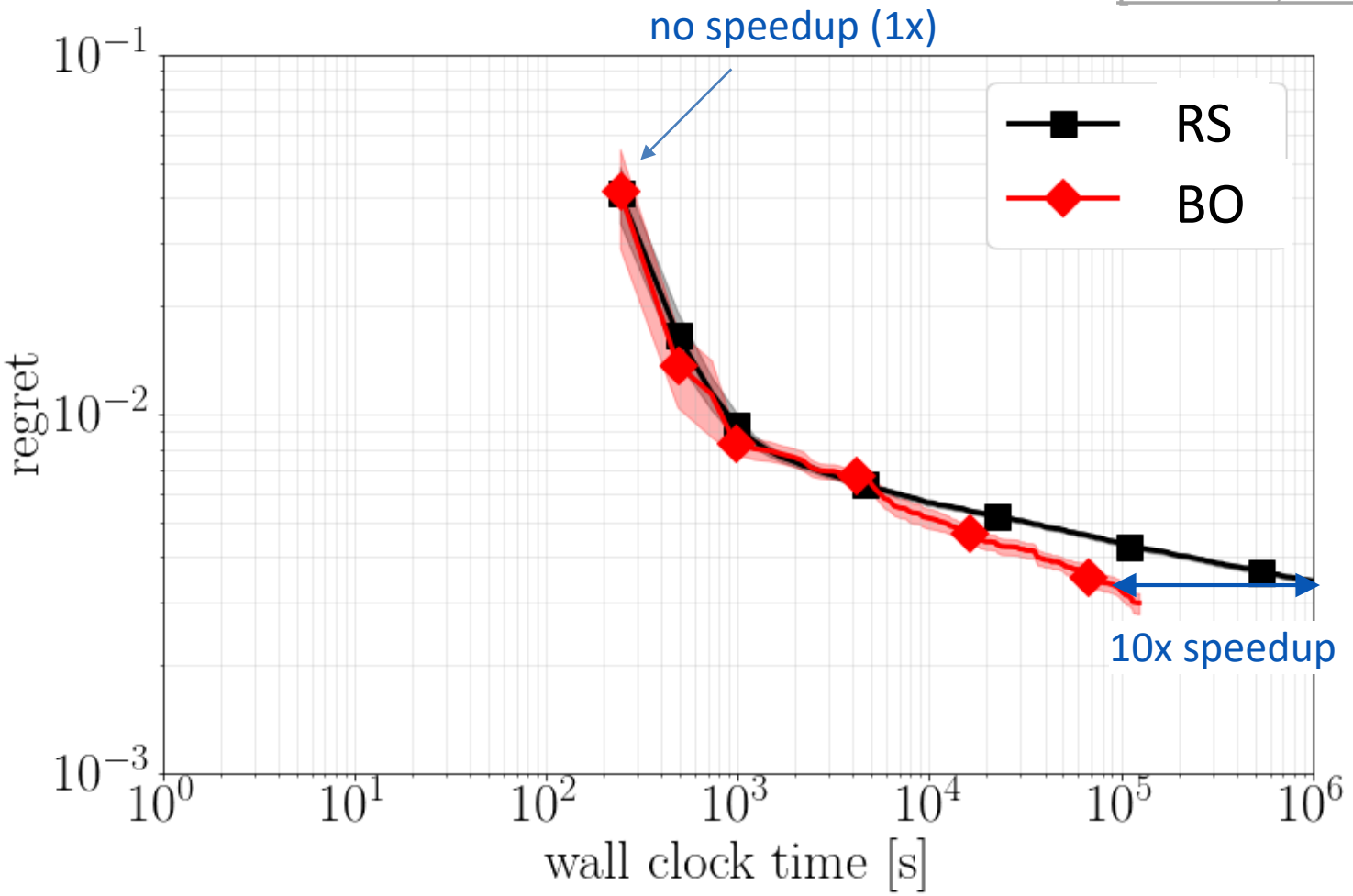


Biggest advantage: much improved **anytime performance**

Auto-Net on dataset adult

Empirical Evaluation: Bayesian Optimization vs. Random Search

[Falkner, Klein & Hutter, ICML 2018]

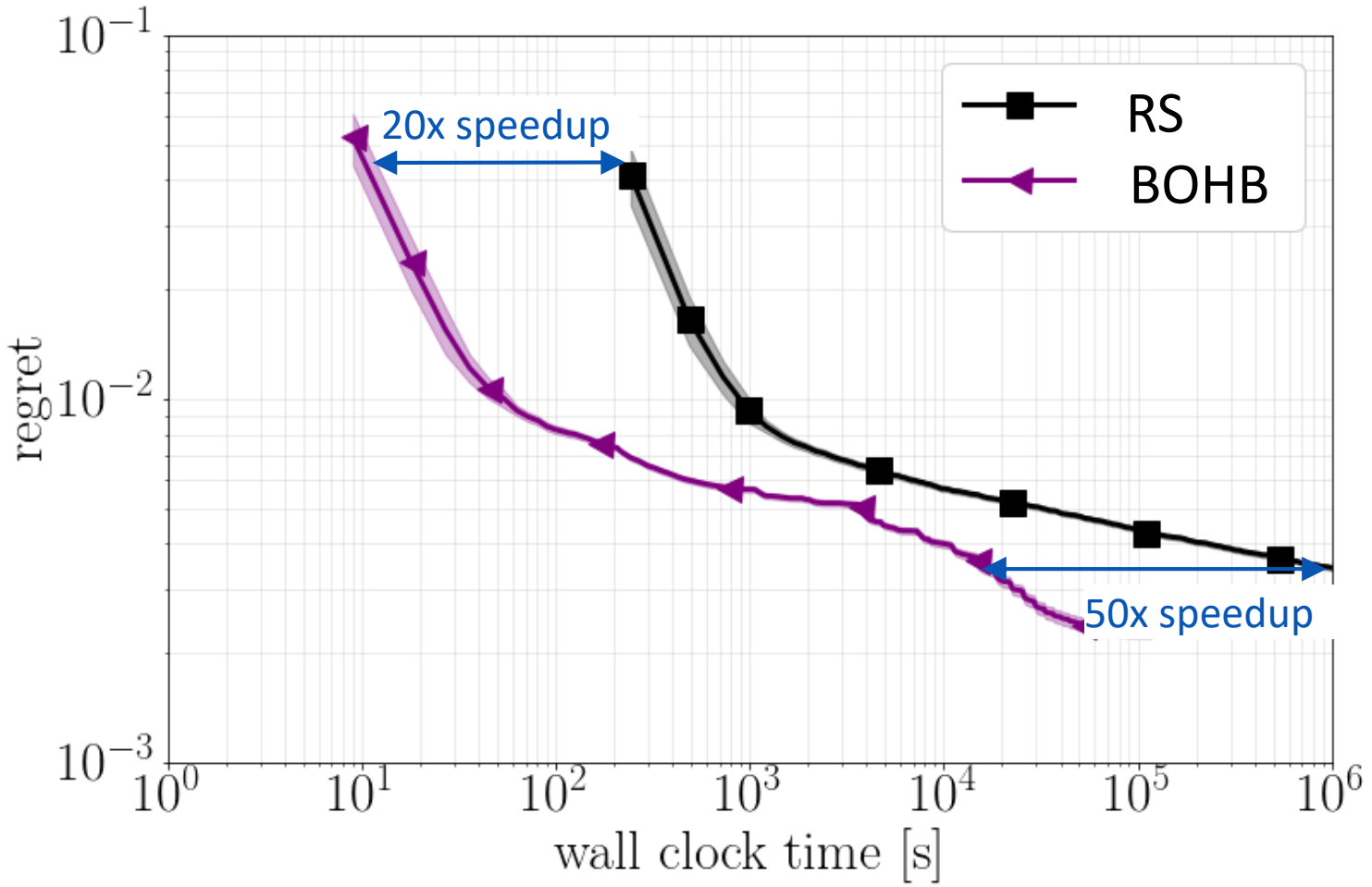


Biggest advantage: much improved **final performance**

Auto-Net on dataset adult

BOHB: Combining Bayesian Optimization & Hyperband

[Falkner, Klein & Hutter, ICML 2018]



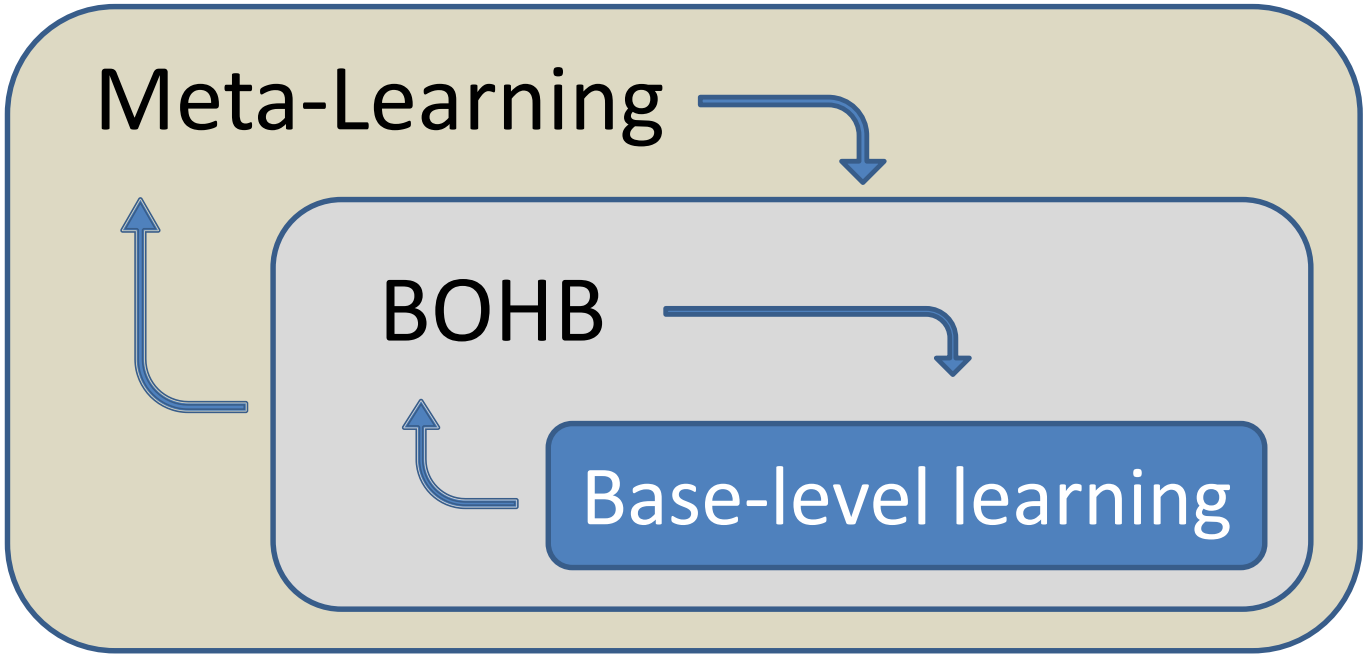
Best of both worlds: strong **anytime and final performance**

Auto-Net on dataset adult

Combination with Meta-Learning

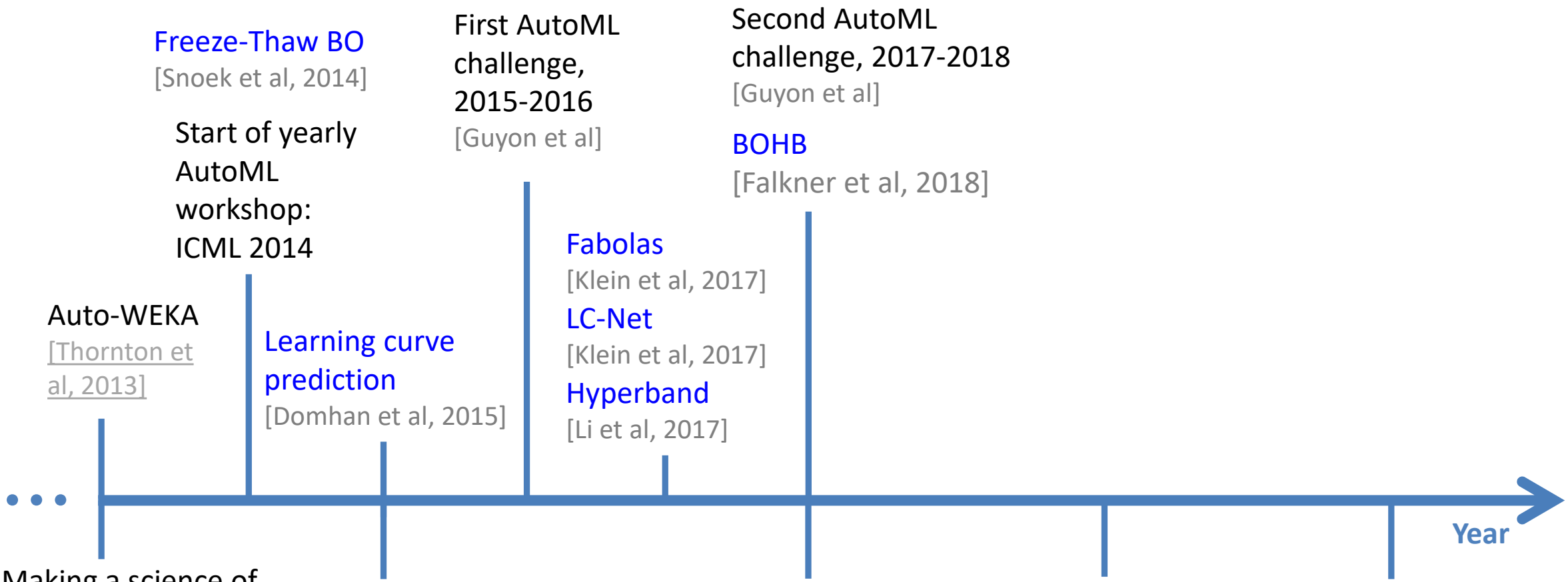
[Feurer et al, 2018]

- Meta-learn the set of configurations to start BOHB with



- About **100x speedups due to meta-learning** across datasets
- Used in Auto-sklearn 2.0 to win the second AutoML challenge

AutoML: A Timeline



Neural architecture search

AutoML: A Timeline

Making a science of model search – optimizing hundreds of dimensions of vision architectures [Bergstra et al, 2013]

Auto-WEKA [Thornton et al, 2013]

Start of yearly AutoML workshop: ICML 2014

Learning curve prediction [Domhan et al, 2015]

Hyperopt-sklearn [Bergstra et al, 2015]

Auto-sklearn [Feurer et al, 2015]

Freeze-Thaw BO [Snoek et al, 2014]

First AutoML challenge, 2015-2016 [Guyon et al]

Fabolas [Klein et al, 2017]

LC-Net [Klein et al, 2017]

Hyperband [Li et al, 2017]

NAS with RL [Zoph & Le, 2017]

BOHB [Falkner et al, 2017

Second AutoML challenge, 2017-20 [Guyon et al]

PoSH-Auto-sklearn [Feurer et al, 2018]

Auto-Keras [Jin et al, 2018]

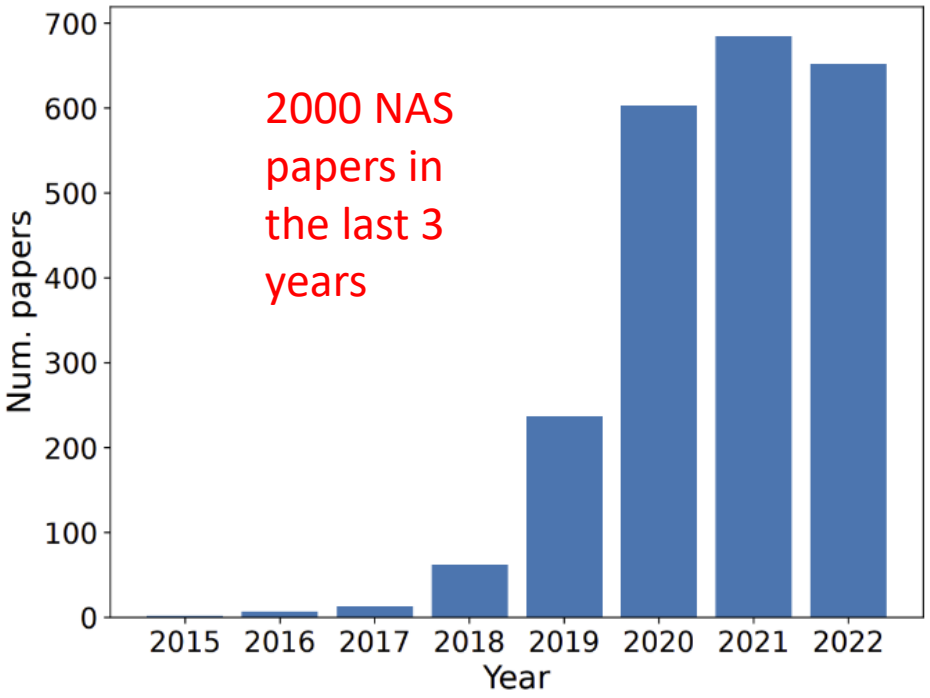
Tabular NAS benchmarks [Ying et al, 2019]

Auto-PyTorch [Zimmer et al, 2020]

AutoGluon [Erickson et al, 2020]

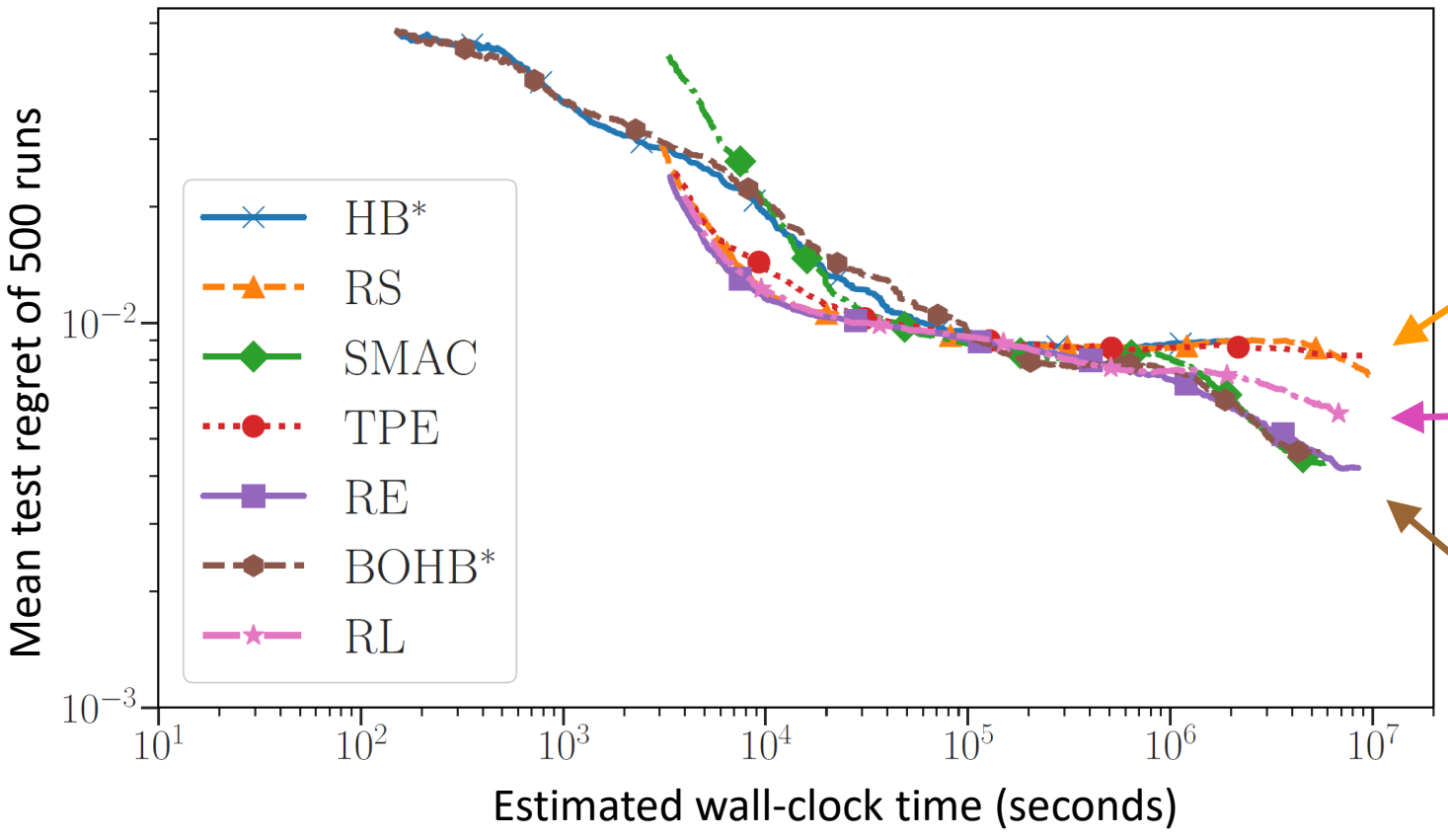
Auto-sklearn 2.0 [Feurer et al, 2022]

Year



2019 result: Bayesian optimization outperforms RL for NAS

[Ying et al., ICML 2019]



Random search

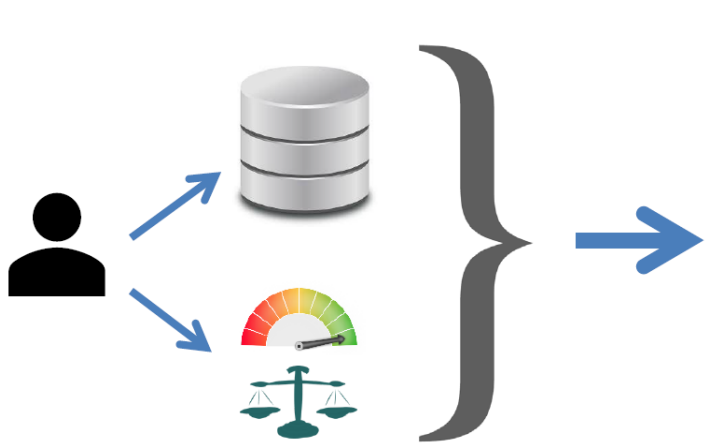
Reinforcement Learning

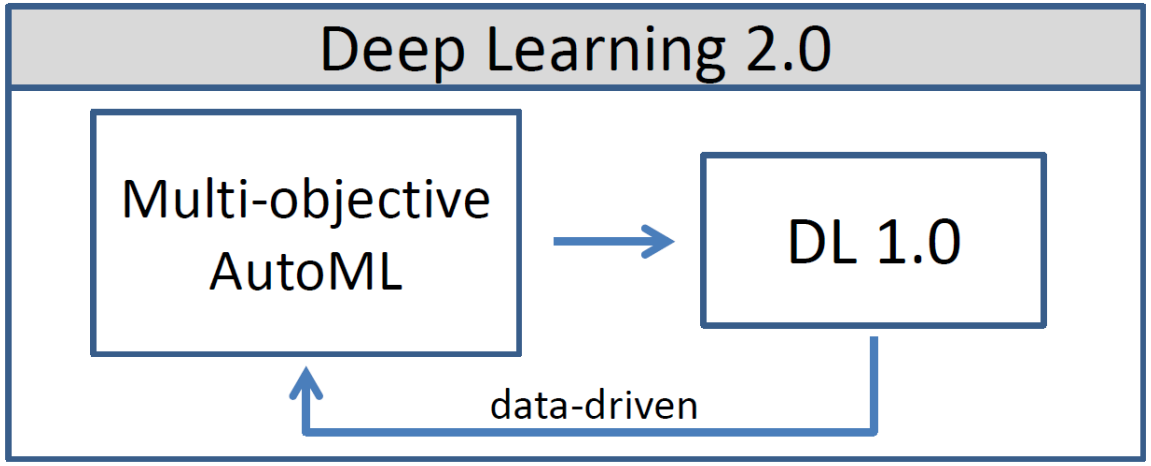
Regularized Evolution & Bayesian Optimization (SMAC / BOHB)

- **SMAC** (published 2011) **outperforms RL** (published 2017)
- Bergstra [2013] optimized 238 architecture choices, Zoph & Le [2017] only 50
- **Google did not invent NAS or AutoML** (and it wasn‘t invented in 2017)

Trustworthy AutoML

Multi-objective AutoML: trustworthiness as the system's objective





- domain expert can specify objectives

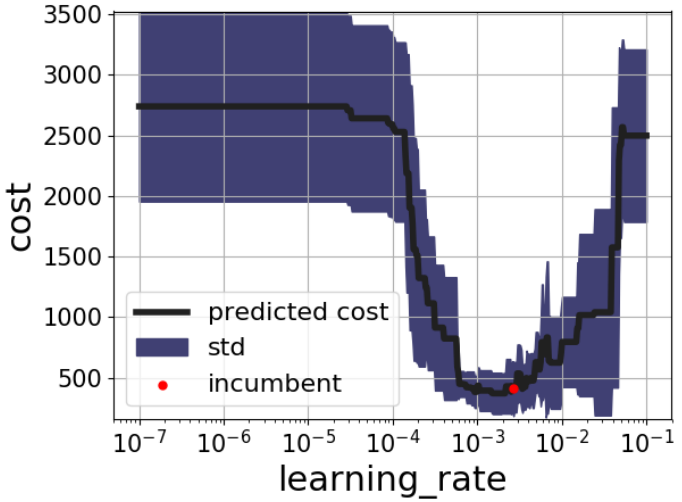
- fairness
- robustness
- model calibration
- interpretability
- latency of predictions
- size(memory) of the model

European Research Council

Example: **Fairer and more accurate than traditional fairness mitigation algorithms** [Dooley et al, 2022]

Additional Trustworthy Dimensions

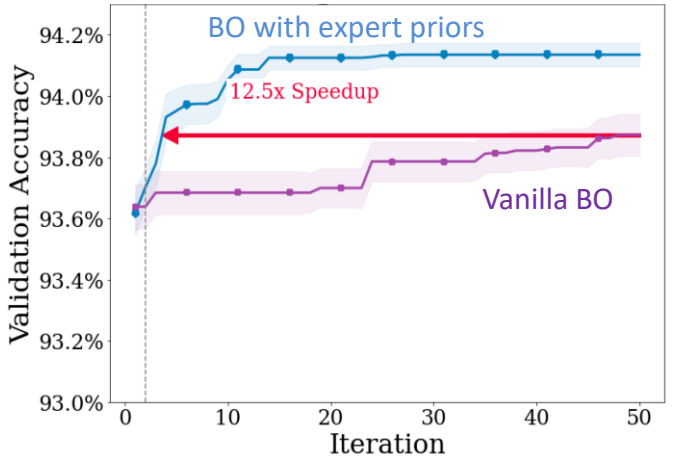
- Explainable AutoML
  - Hyperparameter importance analysis
  - Automated report generation



Marius Lindauer

erc European Research Council

- Human-in-the-loop
  - Guiding the HPO system
  - User priors on good configurations



The Age of Large Pretrained Models

AutoML in the age of large pretrained models

- **AutoML for pretraining**
  - Efficiency is key (e.g., $US 4.6M for training GPT-3)
- **AutoML for fine-tuning**
  - Another CASH problem: choice of pre-trained model and how to fine-tune it
  - QuickTune system for few-shot image classification
  - Perfect match for AutoML: cheap & many different objectives for fine-tuning
- **AutoML for prompting**
- **Meta-learning the entire classification algorithm: TabPFN**

TabPFN, a GPT-like model for tabular data

[Hollmann et al, ICLR 2023]

- TabPFN is a **transformer pretrained to do tabular classification**
- Framed as next-word prediction: $x_1, y_1, \ldots, x_n, y_n, x_{n+1}, ?$

- To be more precise:

$\{(x_1, y_1), \ldots, (x_n, y_n)\}, x_{n+1} \longrightarrow$ TabPFN $\longrightarrow \hat{y}_{n+1}$

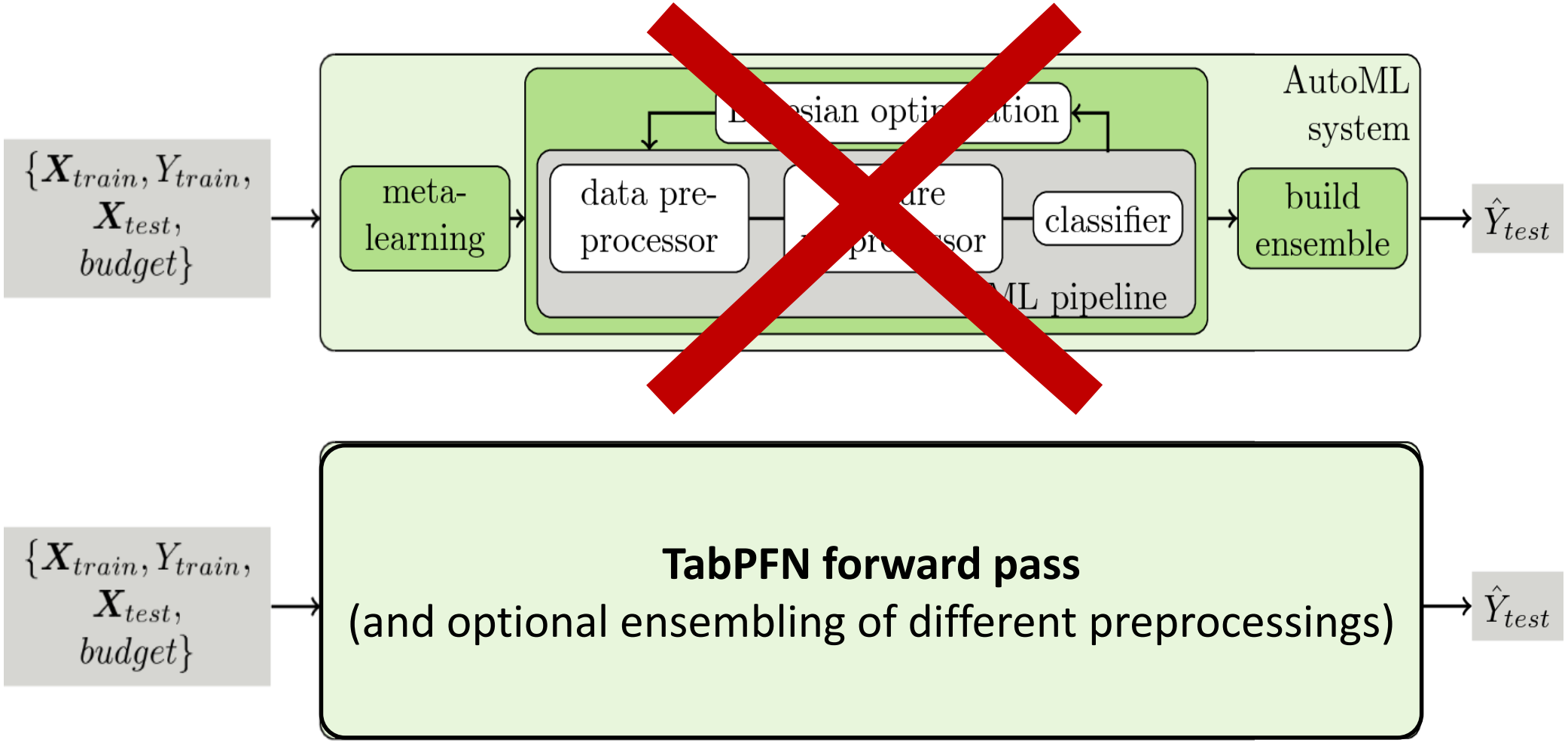
- To be even more precise:

$\{(x_1, y_1), \ldots, (x_n, y_n)\}, x_{n+1} \longrightarrow$ TabPFN $\longrightarrow p(y_{n+1} \mid x_{n+1}, \{(x_1, y_1), \ldots, (x_n, y_n)\})$

TabPFN, a GPT-like model for tabular data

[Hollmann et al, ICLR 2023]

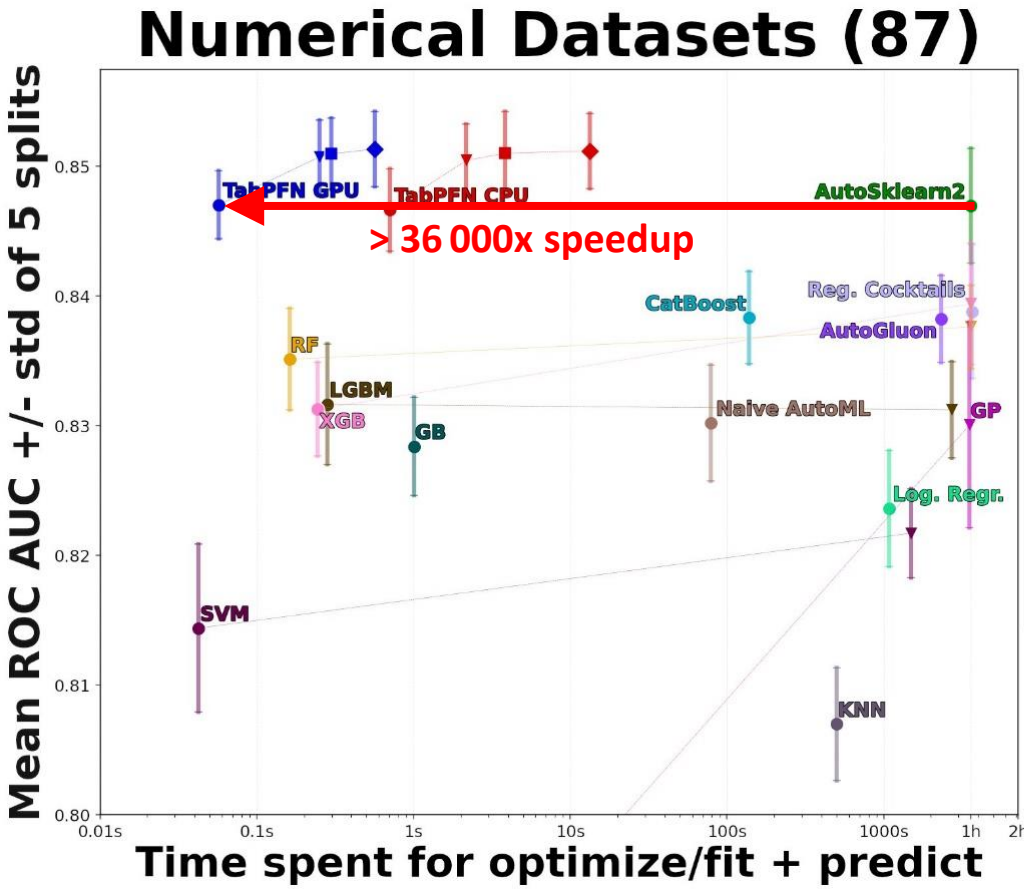
On a new dataset, TabPFN only executes a single forward pass



Quantitative results (87 numerical datasets without missing values)

- **Better performance in 1s than than any other ML / AutoML method in 1h**
  - Disclaimer: these are average results; TabPFN is not the best on every single dataset

- **Limitations we're working on**
  - Size: up to 1000 data points, 100 features, 10 classes
  - Not (yet) designed for: categorical features, missing values, uninformative features
  - High inference time



Wrap-up

AutoML: AI that Builds and Improves AI

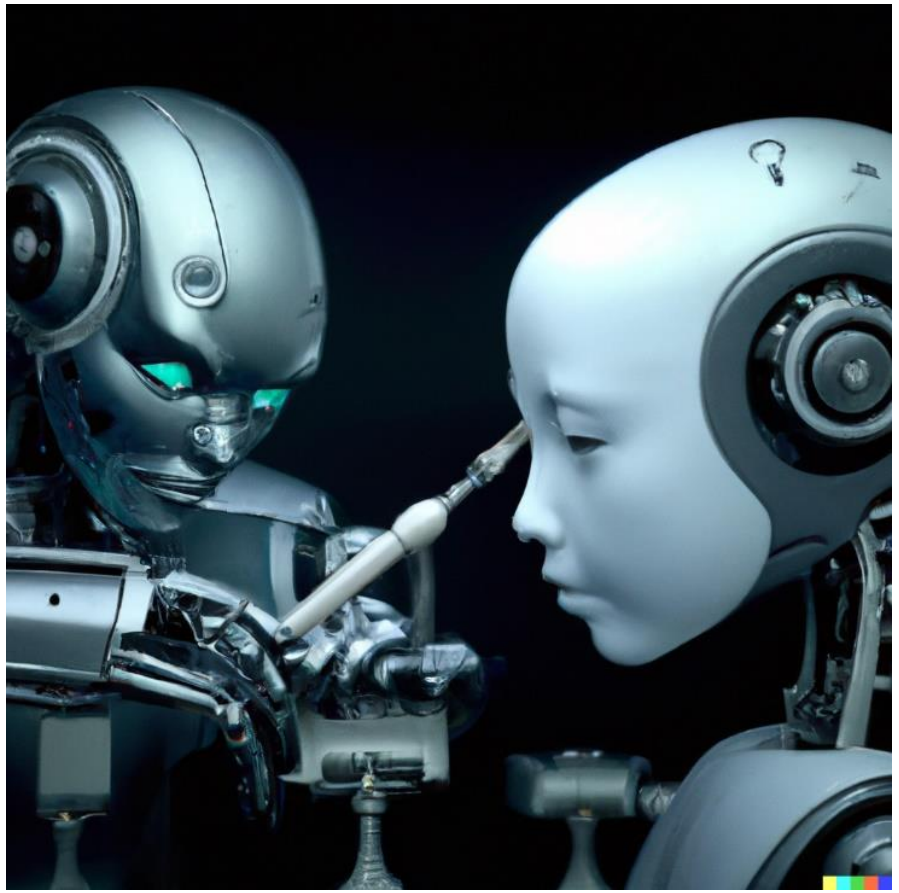


Image credit: DALLE-2

… to be more

- performant
- fair
- calibrated
- energy-efficient
- robust
- aligned
- …

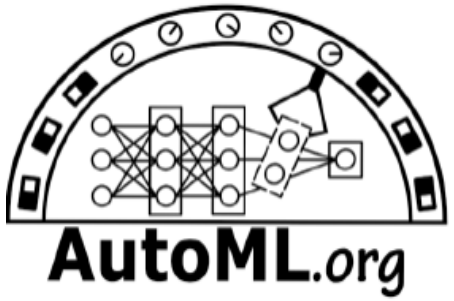
… for the good of humanity

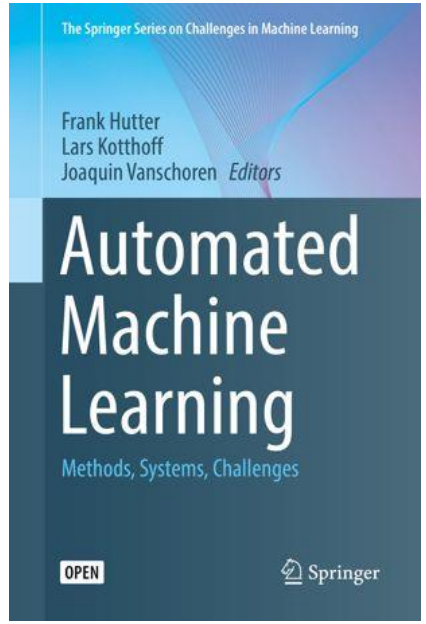
Take-aways

- **AutoML is here to stay**
  - **Democratization** of ML
  - **Productivity** tool

all our code
is open-source:

/automl





- **AutoML does not have to be slow**
  - Meta-learning
  - Multi-fidelity optimization
  - Guidance by human experts

- **Trustworthy AutoML is key**
  - For the good of humanity

AutoML Conference 2023

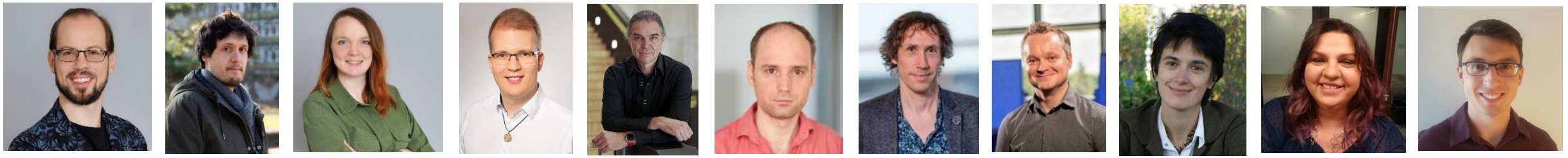
Potsdam/Berlin, Germany

September 12th – 15th 2023

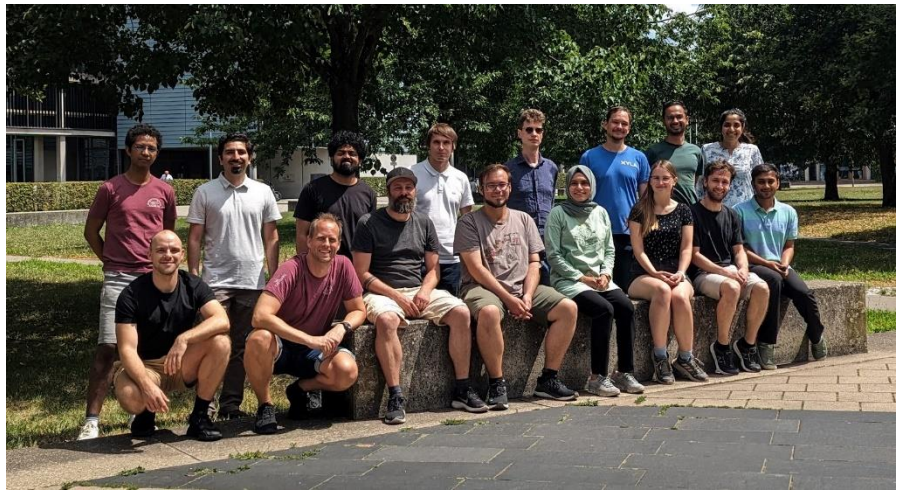
www.automl.cc

Thanks to the amazing AutoML community and those who've helped build it

Lead organizers of AutoML conference, workshops, fall schools and seminar series



AutoML Fall School 2022



My fantastic group