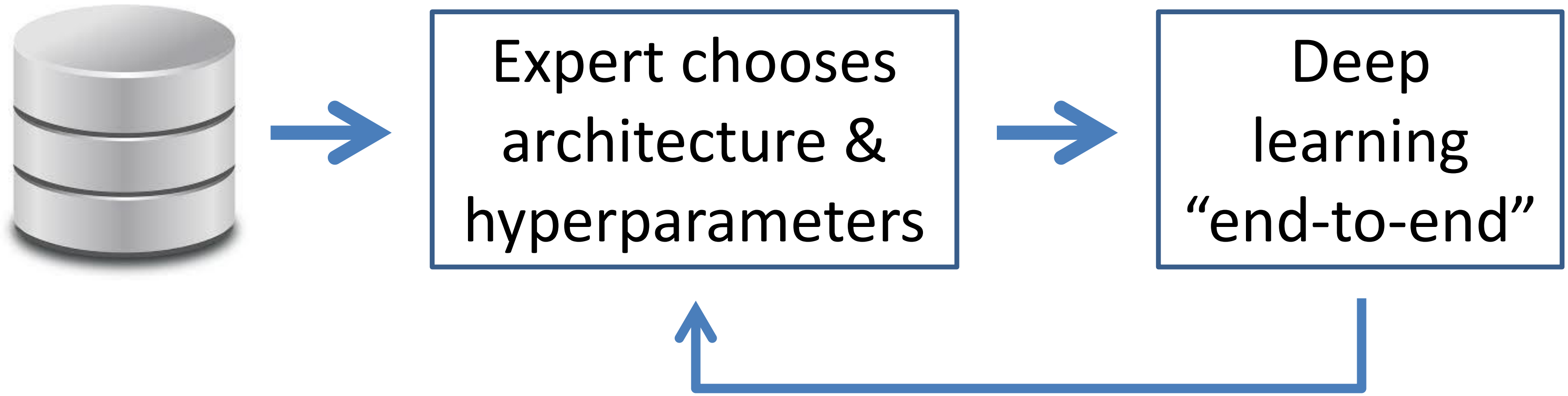
Automatic Machine Learning (AutoML) and How To Speed It Up

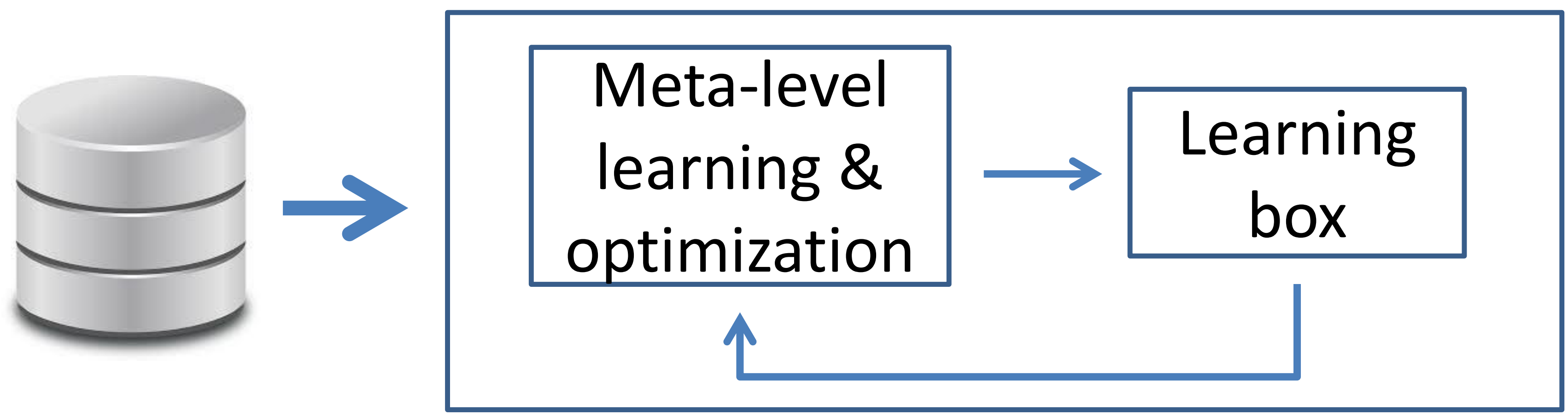
Frank Hutter
Department of Computer Science
University of Freiburg, Germany
fh@cs.uni-freiburg.de

AutoML and Meta-Learning

Current deep learning practice

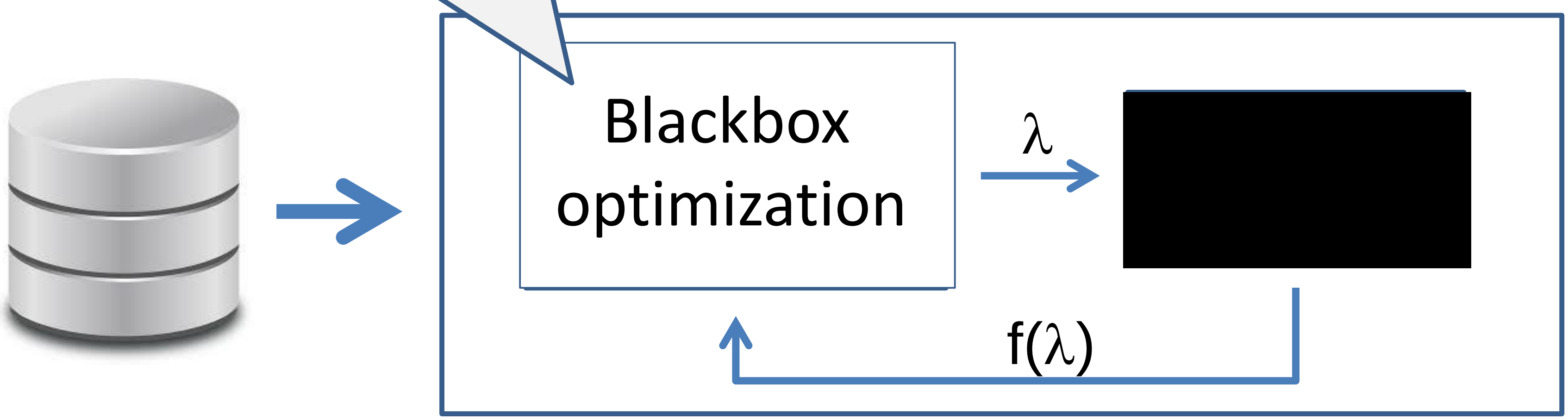


AutoML: true end-to-end learning



AutoML as Blackbox Optimization

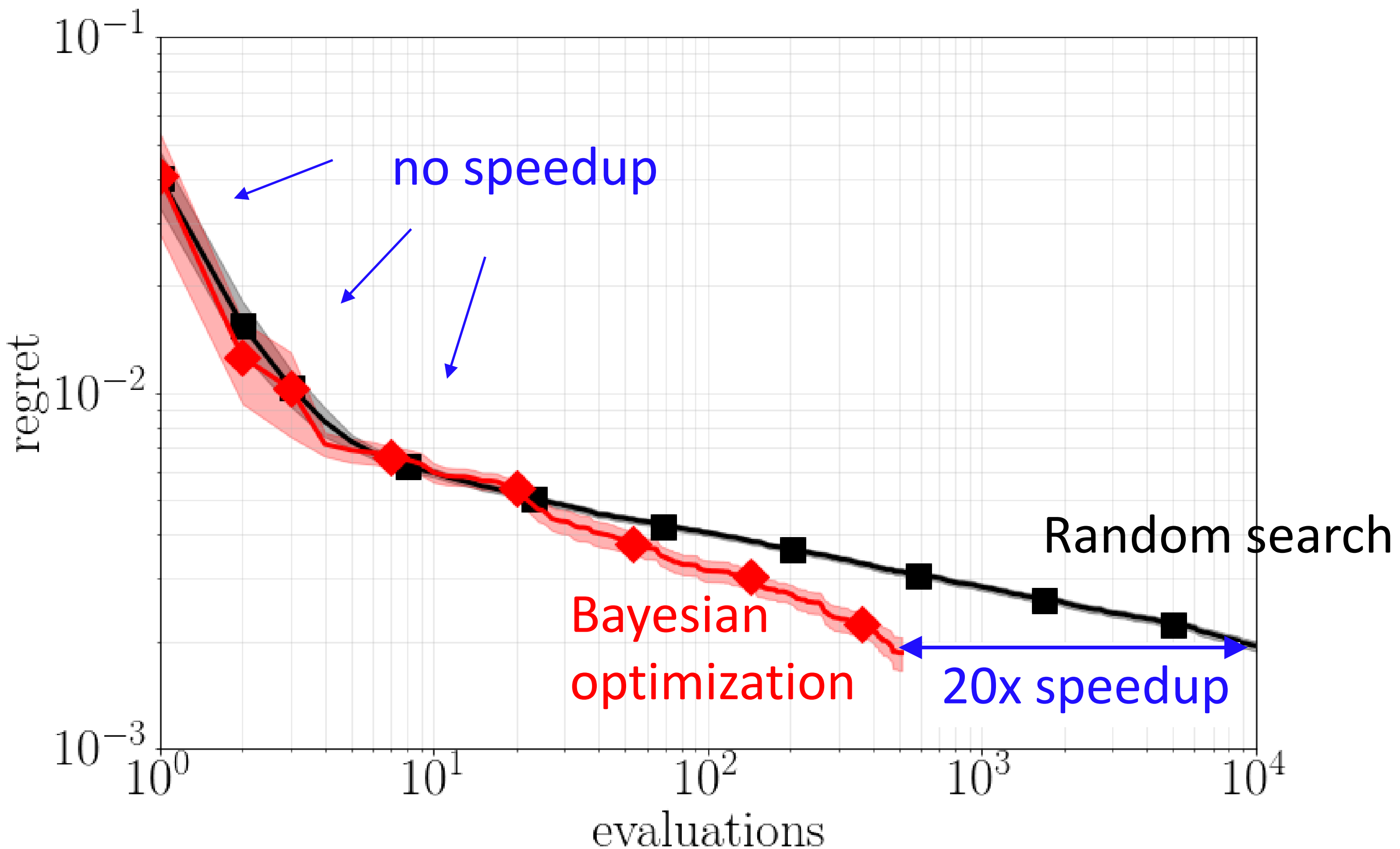
Random search,
evolutionary methods,
reinforcement learning,
…
Bayesian optimization



Effectiveness of Bayesian Optimization

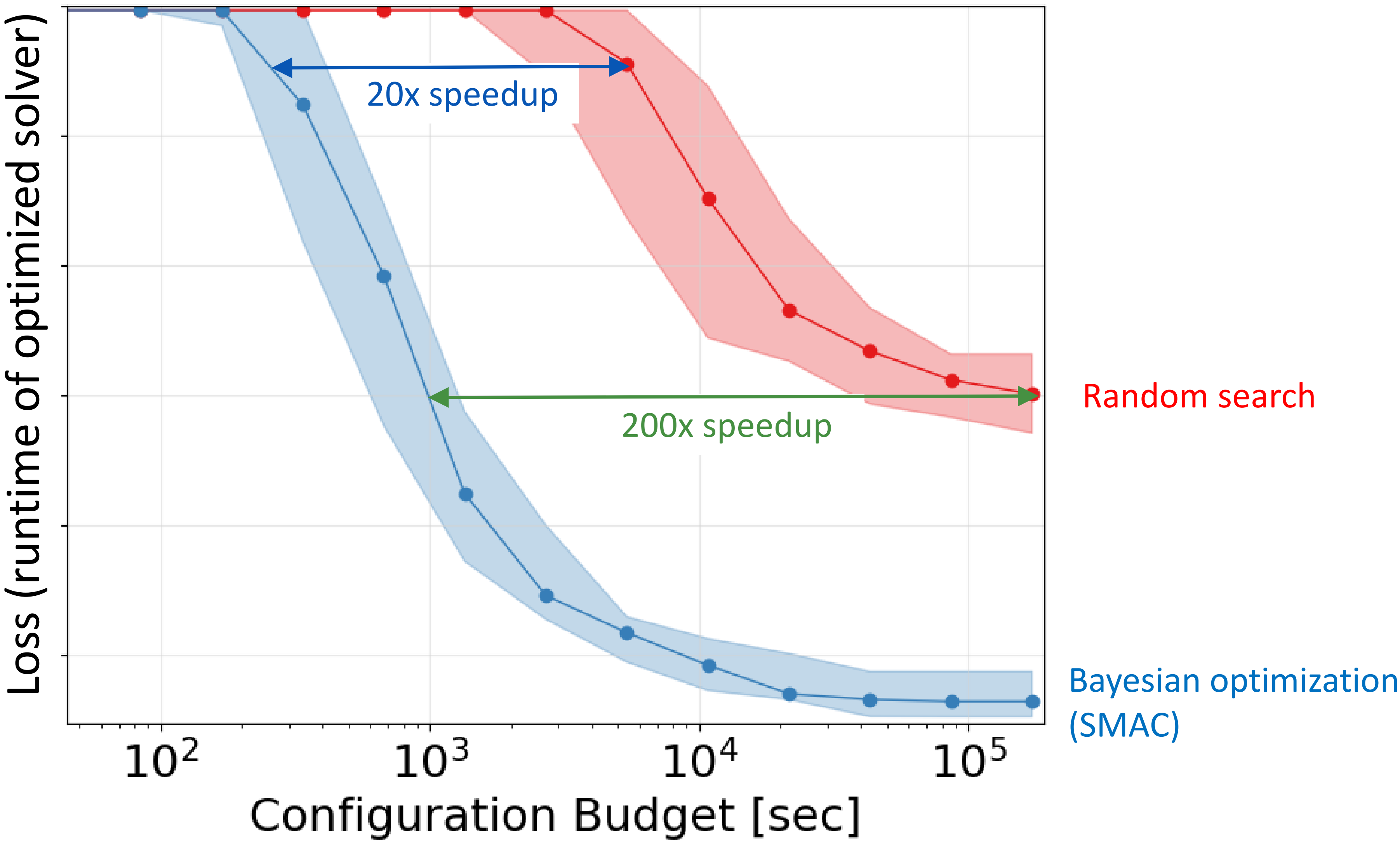
"Sometimes, BayesOpt is only twice as fast as Random Search"

- But sometimes it is dramatically faster



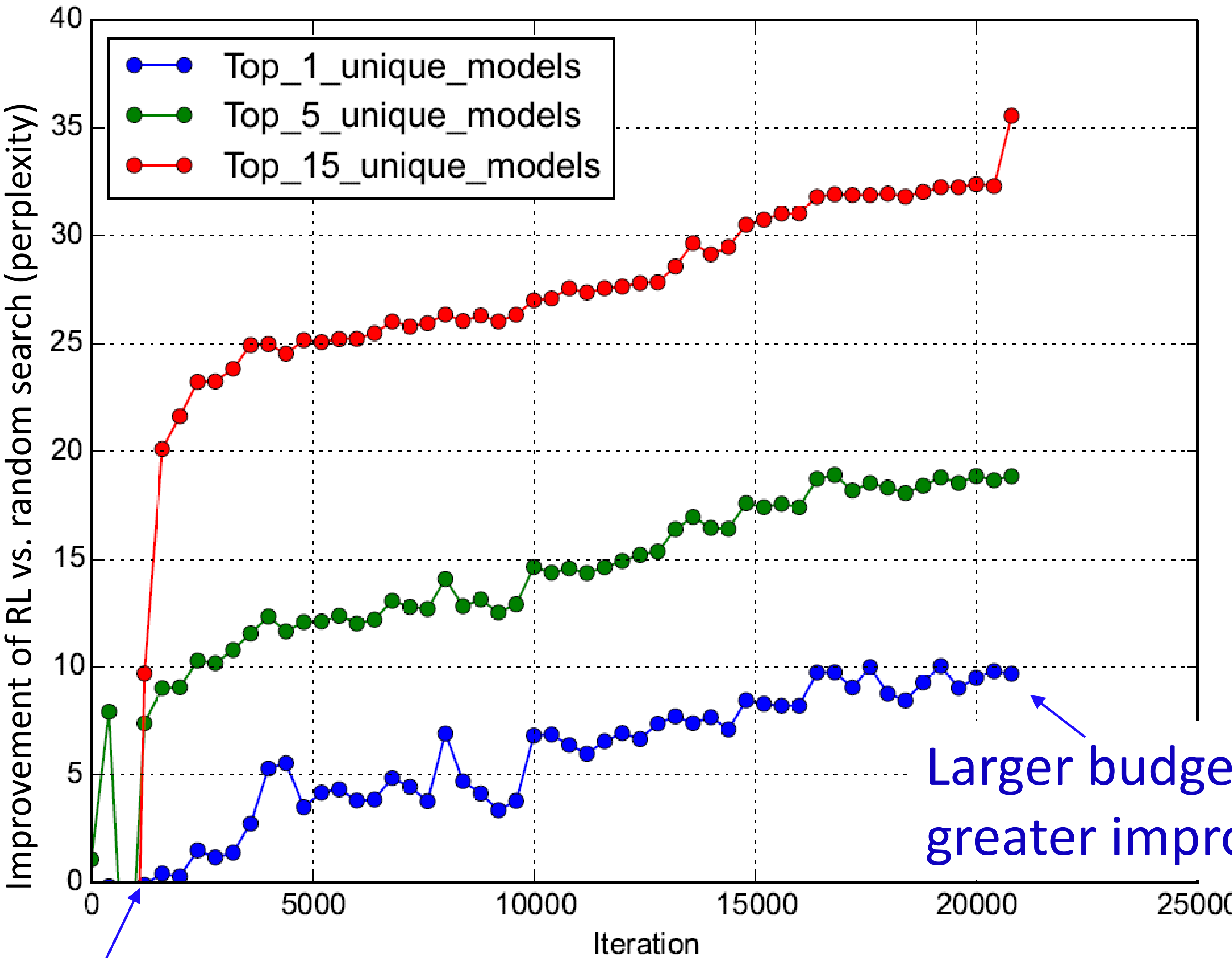
Example: Optimizing a deep feedforward net on dataset adult, 7 hyperparameters

Effectiveness of Bayesian Optimization



Example: Optimizing CPLEX on combinatorial auctions (Regions 100), 76 hyperparameters

Same Pattern Occurs in RL vs. Random Search

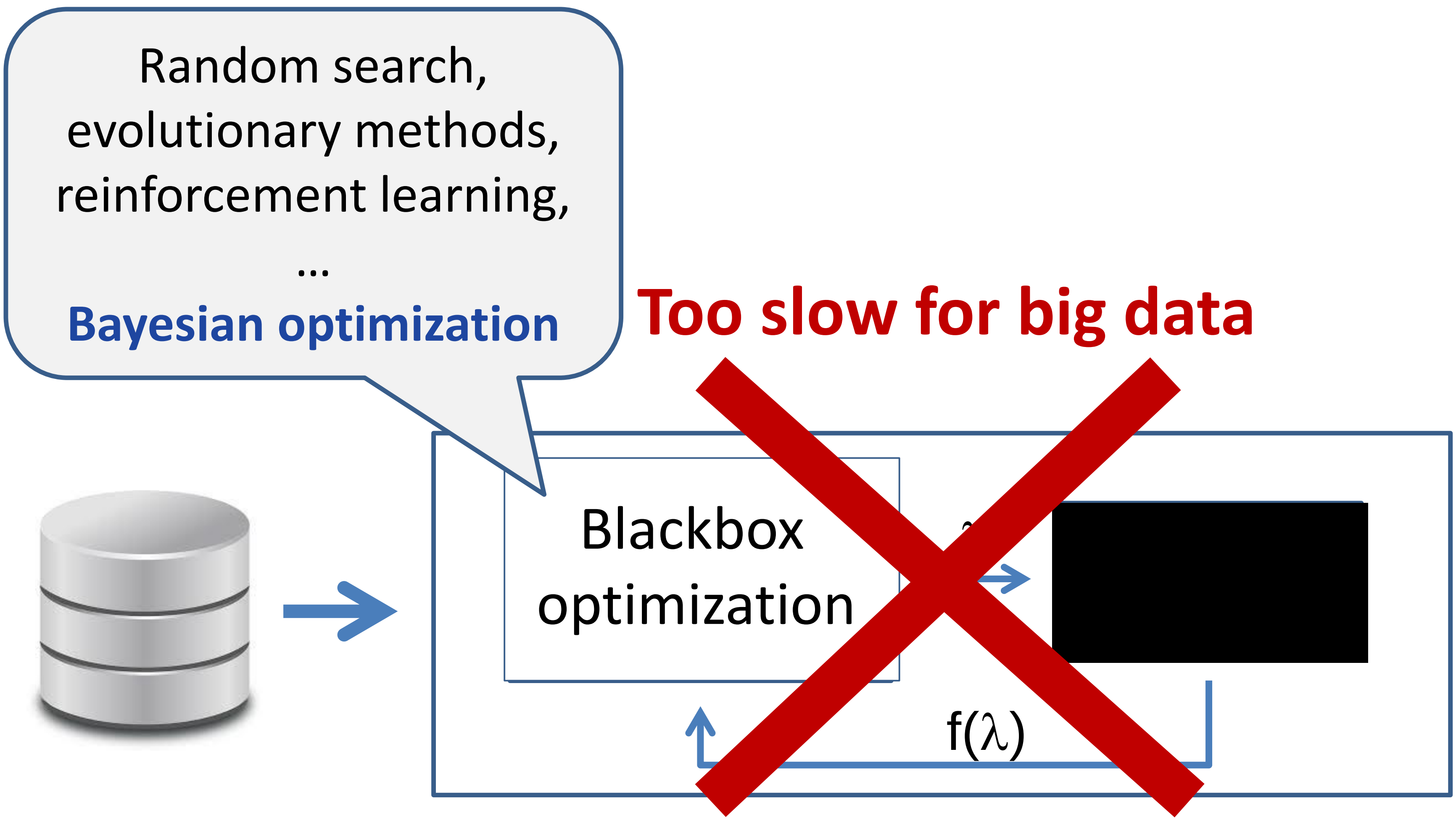


Larger budgets: greater improvements

Up to 1200 function evaluations: RL not better than Random Search

Figure taken from „Neural Architecture Search by Reinforcement Learning“, Zoph & Le

AutoML as Blackbox Optimization



3 AutoML systems

6 ways to go beyond blackbox optimization

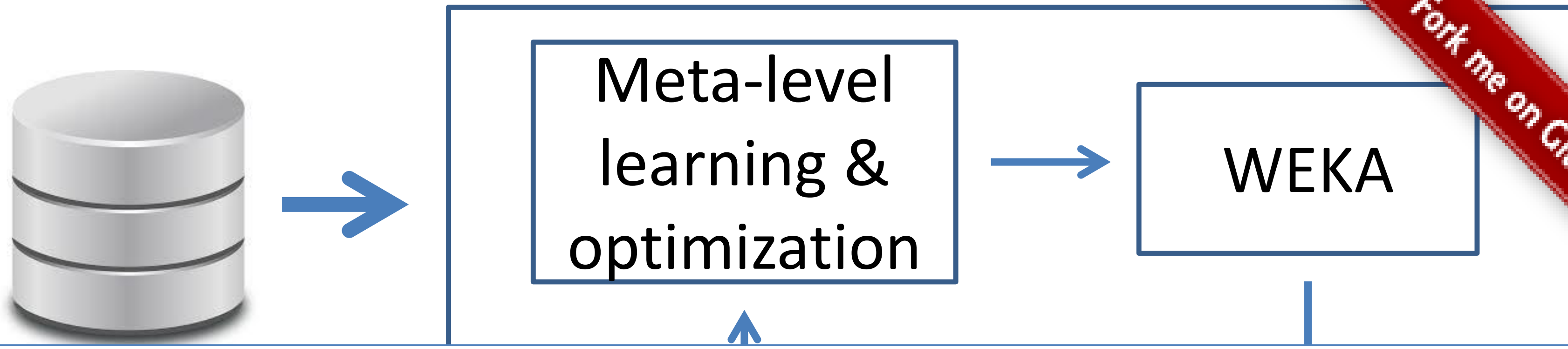
Benchmark: AutoML Challenge

- **Large-scale challenge run by ChaLearn & CodaLab**
  - 17 months, 5 phases with 5 new datasets each (2015-2016)
  - 2 tracks: code submissions / Kaggle-like human track
- **Code submissions: true end-to-end learning necessary**
  - Get training data, learn model, make predictions for test data
  - 1 hour end-to-end
- **25 datasets from wide range of application areas**
  - Already featurized
  - Inputs: features X, targets y

AutoML System 1: Auto-WEKA

[Thornton, Hutter, Hoos, Leyton-Brown, KDD 2013; Kotthoff e… …MLR 2016]

Fork me on GitHub



Available in WEKA package manager; ≈400 downloads/week

- **Parameterize ML framework: WEKA** [Witten et al, 1999-current]
  - 27 base classifiers (with up to 10 hyperparameters each)
  - 2 ensemble methods; in total: 786 hyperparameters

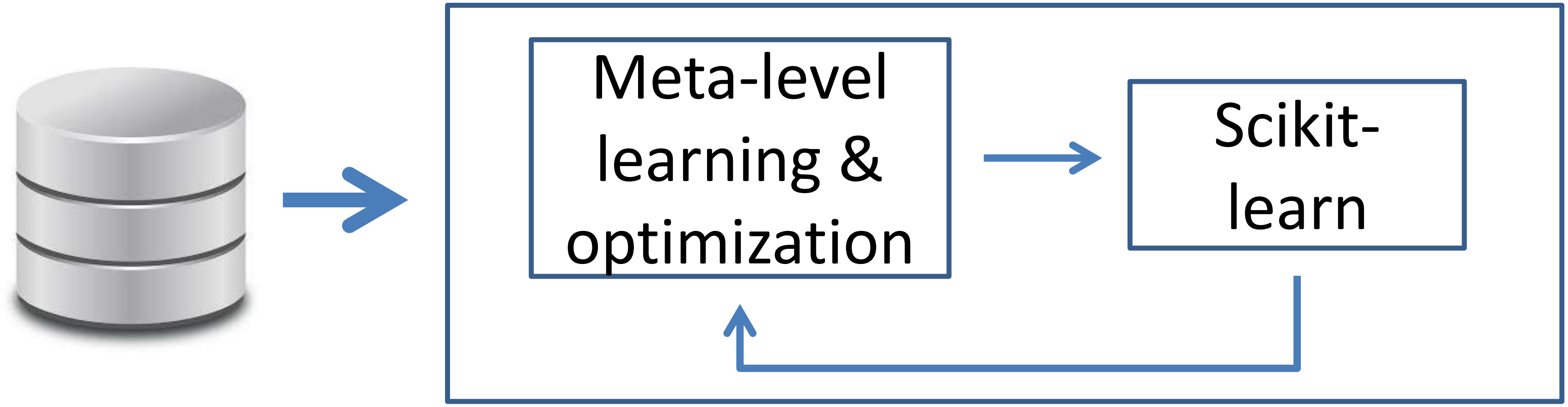
- Optimize **CV performance** by Bayesian optimization (SMAC)

Only evaluate more folds for good configurations
  - 5x speedups for 10-fold CV

$$\blacksquare := \sum_{i=1}^{k} \blacksquare_i$$

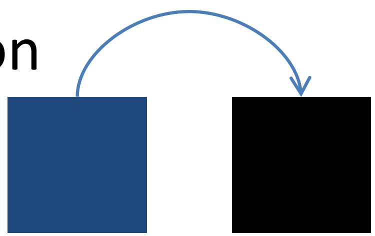
AutoML System 2: Auto-sklearn

[Feurer, Klein, Eggensperger, Springenberg, Blum, Hutter; NIPS 2015]



- Optimize CV performance by SMAC $\blacksquare := \sum_{i=1}^{k} \blacksquare_i$

2 **Meta-learning** to warmstart Bayesian optimization



- Reasoning over different datasets
- Dramatically speeds up the search (2 days $\rightarrow$ 1 hour)

3 Automated **posthoc ensemble construction** to combine the models Bayesian optimization evaluated

- Efficiently re-uses its data; improves robustness

Auto-sklearn: Ready for Prime Time

- Winning approach in the AutoML challenge
  - **Auto-track: overall winner, 1st place in 3 phases**, 2nd pl n 1
    - Close competitor: variant of automatic statistician [Lloyd et al]
  - **Human track: always in top-3 vs. 150 teams of human exper**
  - **Final two rounds: won both tracks**

Fork me on GitHub

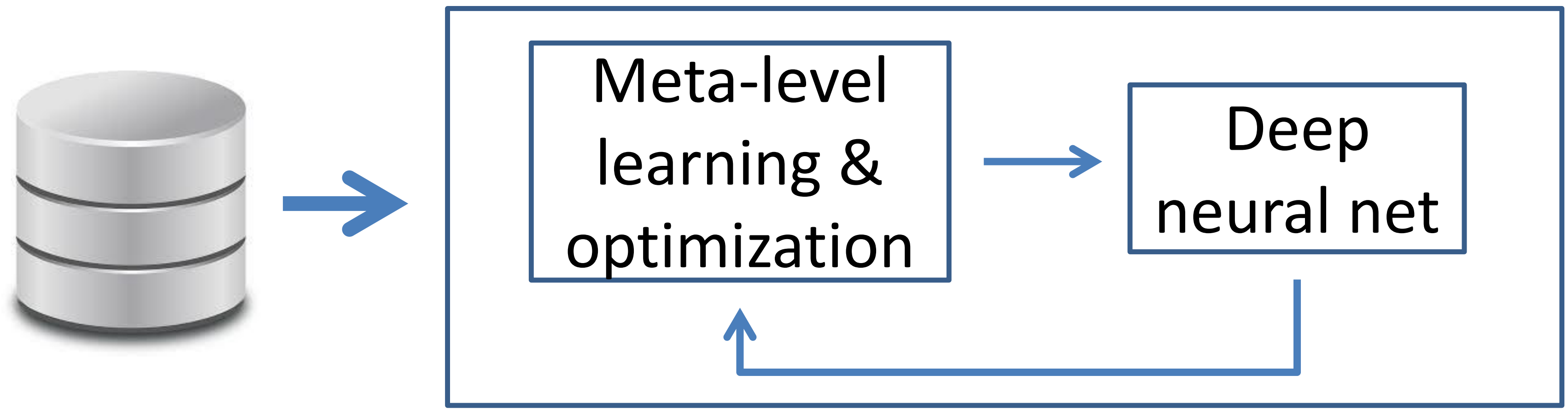
https://github.com/automl/auto-sklearn

Watch 121 | Star 1,638 | Fork 298

- Trivial to use:

```
import autosklearn.classification as cls
automl = cls.AutoSklearnClassifier()
automl.fit(X_train, y_train)
y_hat = automl.predict(X_test)
```

AutoML System 3: Auto-Net



- CV performance optimized by SMAC $\blacksquare := \sum_{i=1}^{k} \blacksquare_i$

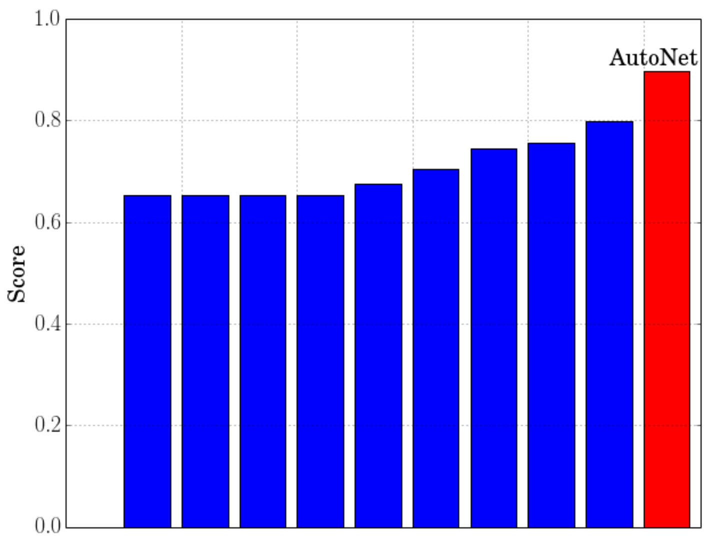
- Joint optimization of:
  - Network architecture
  - Hyperparameters

Auto-Net in AutoML Challenge

[Mendoza, Klein, Feurer, Springenberg & Hutter, AutoML 2016]

- Featurized data $\rightarrow$ fully-connected network
  - Up to 5 layers (with 3 layer hyperparameters each)
  - 14 network hyperparameters, in total **29 hyperparameters**
  - Optimized for 18h on 5GPUs

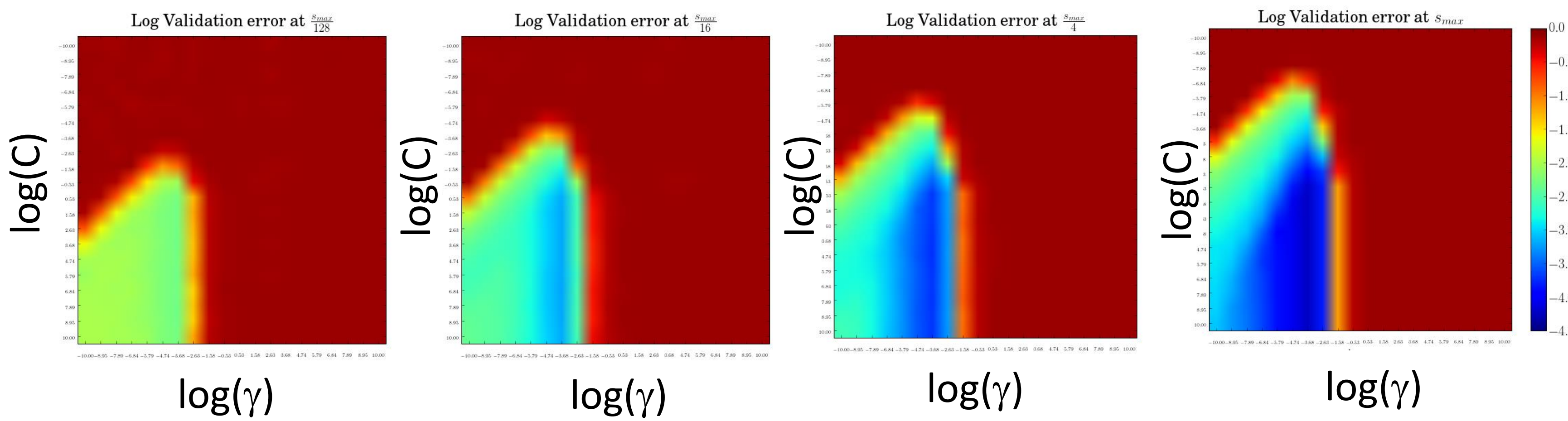
- Auto-Net won several datasets against human experts
  - E.g., Alexis data set:
    - 54491 data points,
      5000 features, 18 classes
  - **First automated deep learning system to win a ML competition data set** against human experts



Using Cheap Approximations of the Blackbox

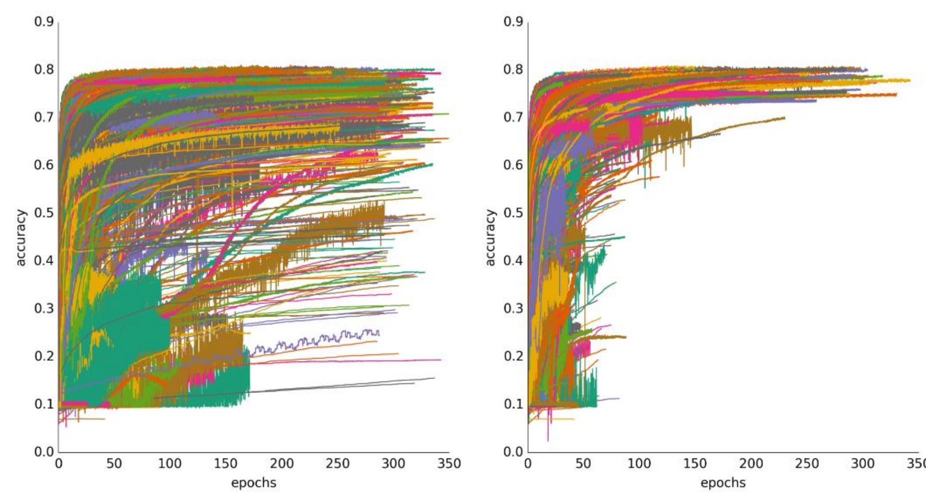
4 Reasoning across **subsets of the data**

– Up to 1000x speedups [Klein et al, AISTATS 2017]



5 Reasoning across **training epochs**

[Swersky et al, arXiv 2014]
[Domahn et al, IJCAI 2015]

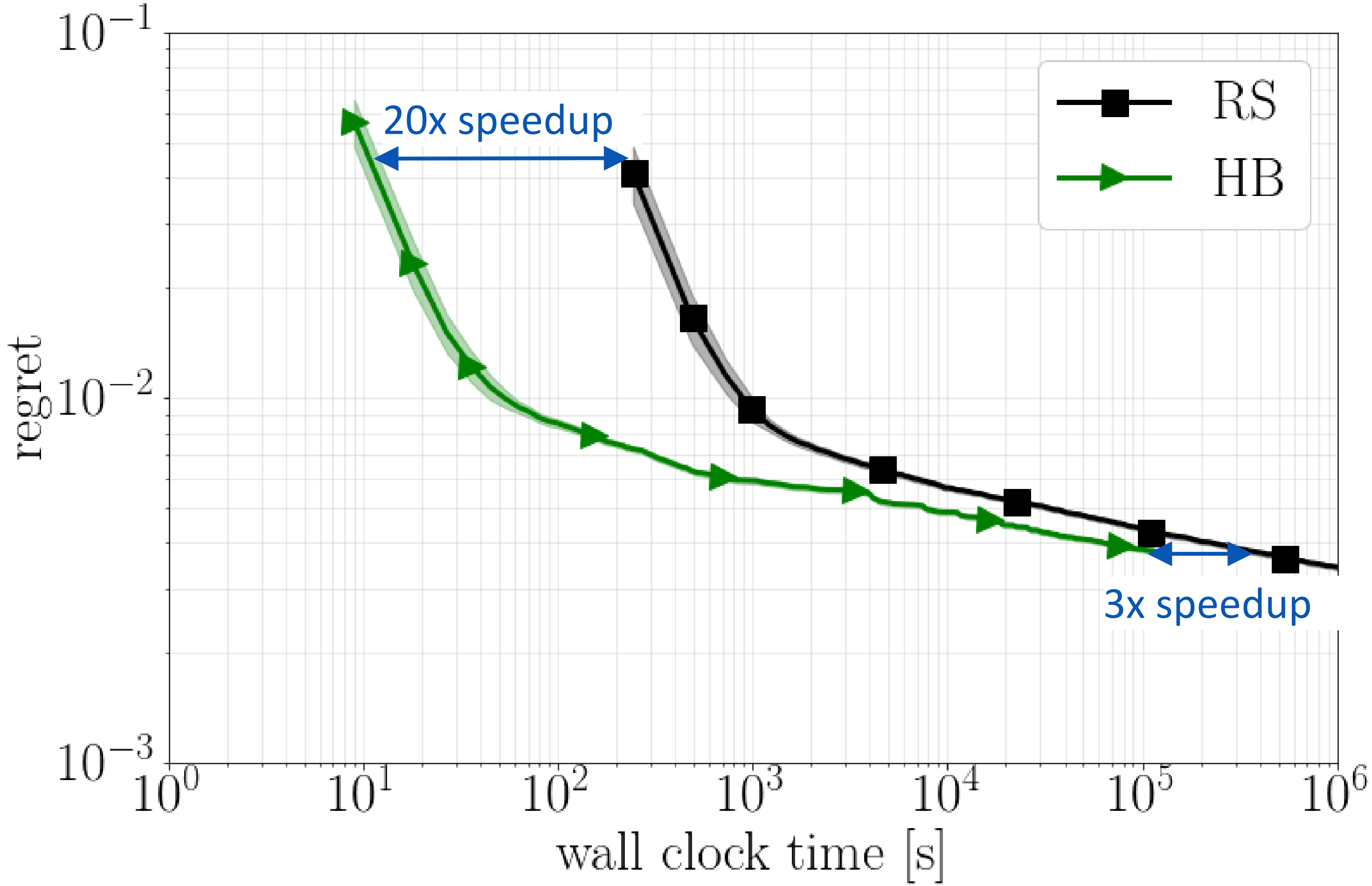


Hyperband & Successive Halving

- Successive Halving [Jamieson & Talwalkar, AISTATS 2015]
  - Run N (=many) configurations for a small budget B
  - Iteratively:
    Select best half of configurations and double their budget

- Hyperband [Li et al, ICLR 2017]
  - Calls Successive Halving iteratively with
    different tradeoffs of N and B

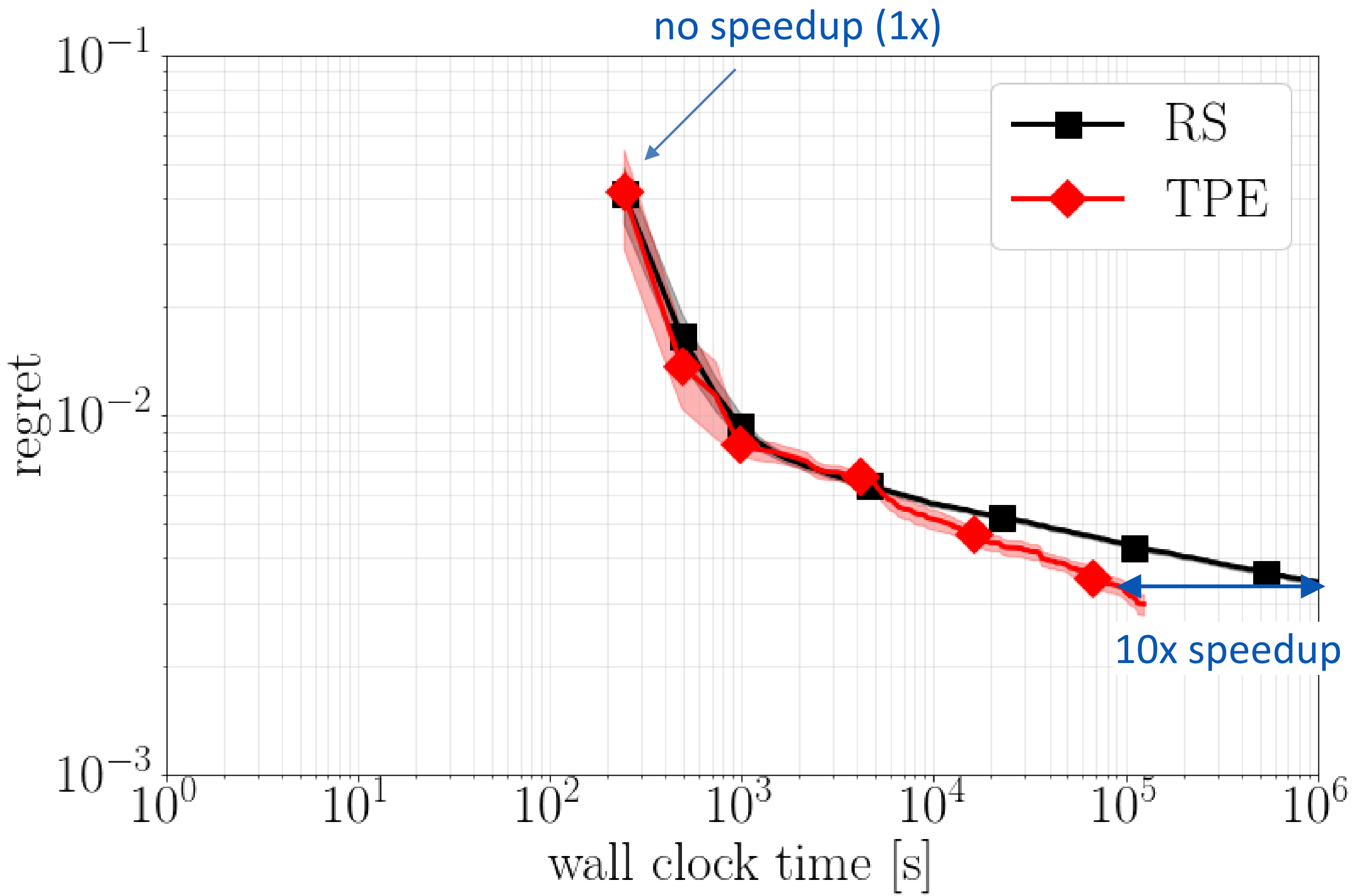
Hyperband vs. Random Search



Biggest advantage: much improved **anytime performance**

Auto-Net on dataset adult

Bayesian Optimization vs. Random Search

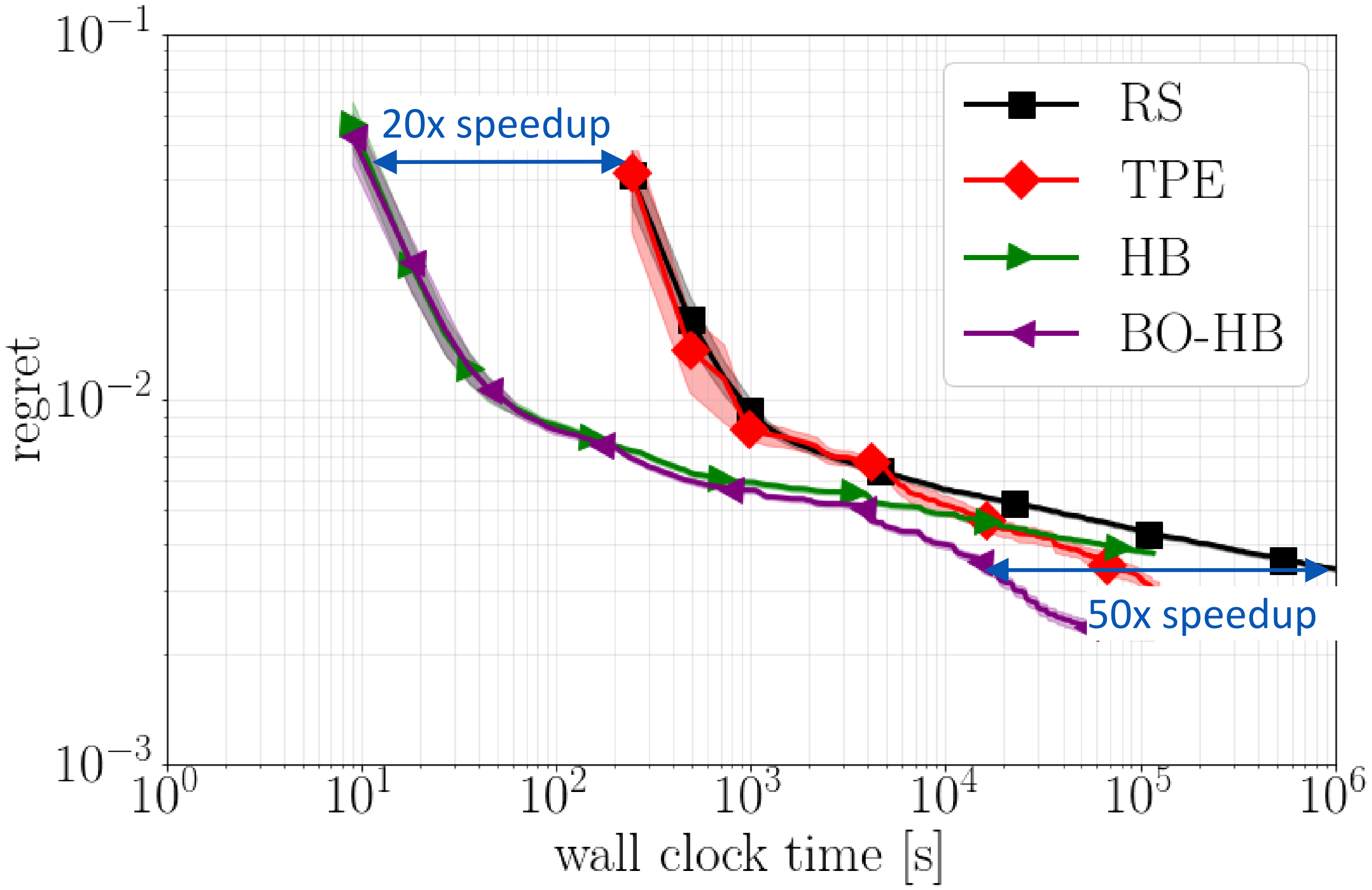


Biggest advantage: much improved **final performance**

Auto-Net on dataset adult

Combining Bayesian Optimization & Hyperband

[Falkner, Klein & Hutter, BayesOpt 2017]

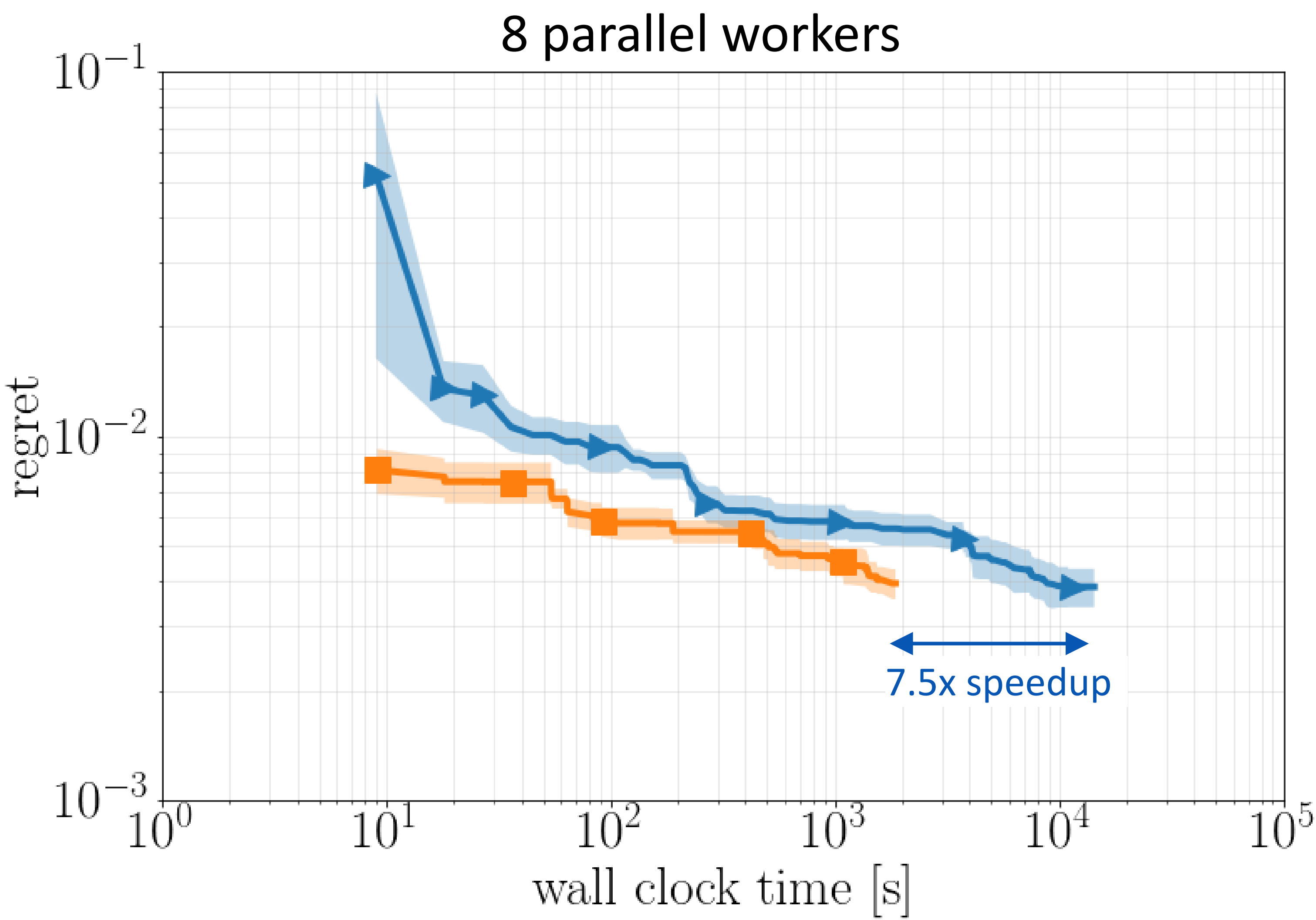


Best of both worlds: strong **anytime and final performance**

Auto-Net on dataset adult

Almost Linear Speedups By Parallelization

[Falkner, Klein & Hutter, BayesOpt 2017]



Auto-Net on dataset adult

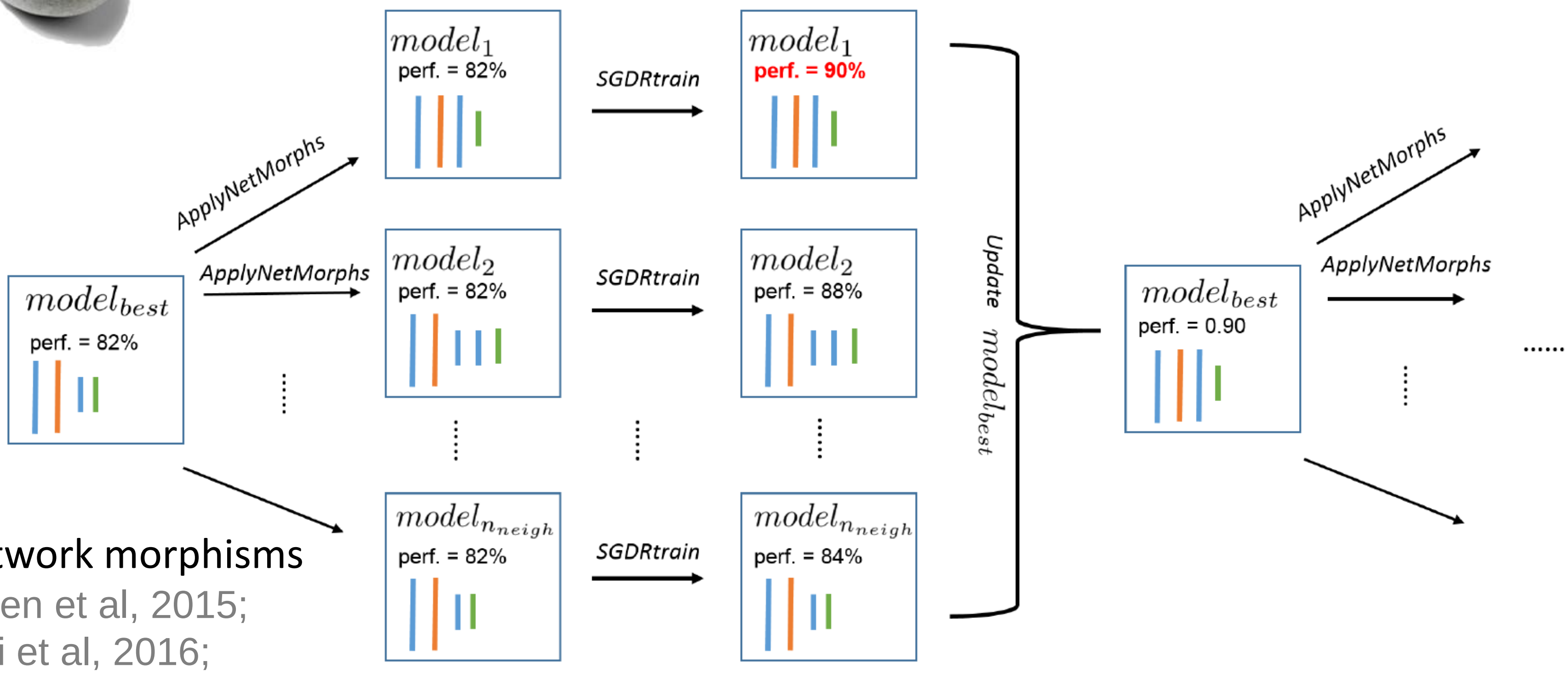
Tuning CNNs on a Budget: CIFAR-10

[Falkner, Klein & Hutter, BayesOpt 2017]

- Six design decisions
  - Depth, widening factor
  - Learning rate, batch size, weight decay, momentum
- Maximum budget per CNN run: **2 hours on a Titan X**
  - Ran BO-HB for 12 hours on 10 GPUs
  - Result: **4% test error**
- Maximum budget per CNN run: 3 hours on a Titan X
  - Ran BO-HB for **12 hours on 10 GPUs**
  - Result: **3.5% test error**

[Elsken, Metzen & Hutter, MetaLearn 2017]

6 Online Adaptation of Architecture & Hyperparams



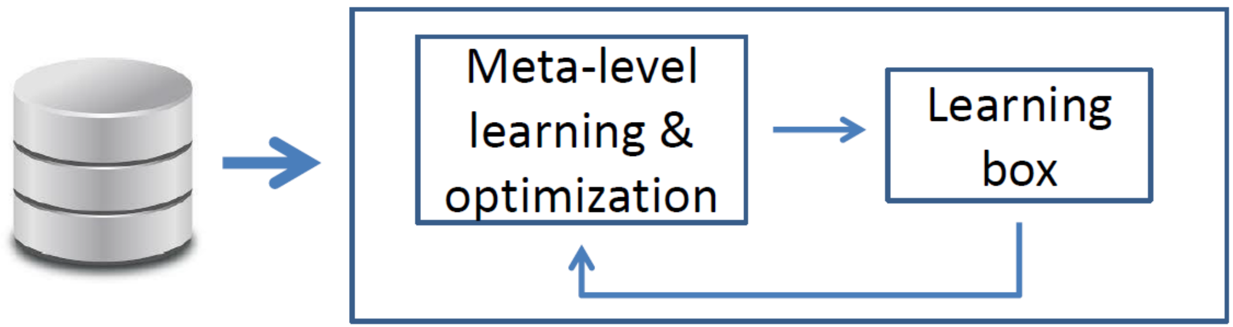
Network morphisms
[Chen et al, 2015; Wei et al, 2016; Cai et al, 2017]

Cosine annealing
[Loshchilov & Hutter, 2017]

Result: **architecture search in 12 hours on 1 GPU**: 5.7% on CIFAR-10

Conclusion

- Bayesian optimization enables **true end-to-end learning**
  - Auto-WEKA, Auto-sklearn & Auto-Net



- Large speedups by going beyond blackbox optimization
  - Learning across datasets
  - Learning across data subsets & epochs
  - Combination of Hyperband and Bayesian optimization
  - Online adaptation of architectures & hyperparameters

- Links to code: http://automl.org

Thanks!

Funding sources

European Research Council

Emmy Noether-Programm

DFG Deutsche Forschungsgemeinschaft

BrainLinks BrainTools acting.thoughts

DFG Priority Programme 1527

European Commission

EU project RobDREAM

My fantastic team

Other collaborators

UBC: Chris Thornton, **Holger Hoos**, **Kevin Leyton-Brown**, Kevin Murphy

DeepMind: Ziyu Wang, Nando de Freitas

Bosch: Thomas Elsken, Jan Hendrik Metzen

MPI Tübingen: Philipp Hennig

Uni Freiburg: Tobias Springenberg, Robin Schirrmeister, Tonio Ball, Thomas Brox, Wolfram Burgard

I'm looking for more great postdocs!