Extending the Versatile Workhorse of Blackbox Neural Architecture Search

Frank Hutter

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

@FrankRHutter
@AutoML_org

These slides are available at www.automl.org/talks -- all references are hyperlinks

Big Picture Motivation: Why Deep Learning Succeeded

Traditional ML practice before Deep Learning

Data

→ Domain expert engineers features → Traditional ML (XGBoost etc)

trial & error

Deep Learning

Data

✔ saves manual labour

✔ better results by end-to-end joint optimization

→ Deep Learning (end-to-end)

From Deep Learning 1.0 to Deep Learning 2.0

Traditional ML practice before Deep Learning

Data → Domain expert engineers features → Traditional ML (XGBoost etc)

trial & error

Deep Learning

Data → DL expert chooses architecture & hyperparameters → Deep Learning (end-to-end)

trial & error

From Deep Learning 1.0 to Deep Learning 2.0

Deep Learning 2.0

Data

- saves human labour
- better results by end-to-end joint optimization

→ Deep Learning 2.0 (end-to-end)

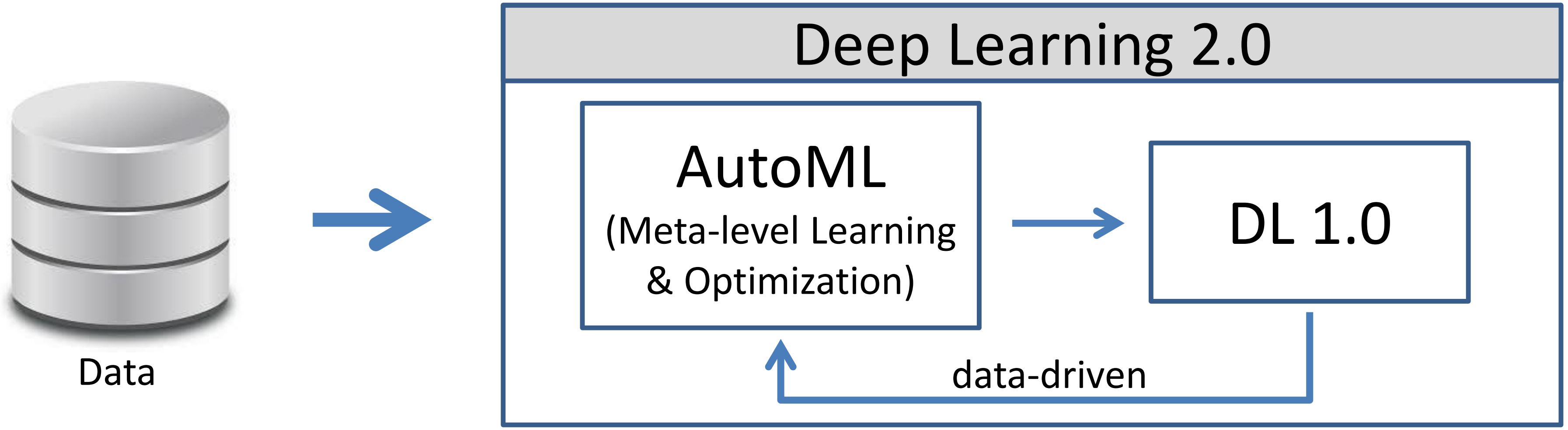
Deep Learning 1.0

Data

→ DL expert chooses architecture & hyperparameters → Deep Learning (end-to-end)

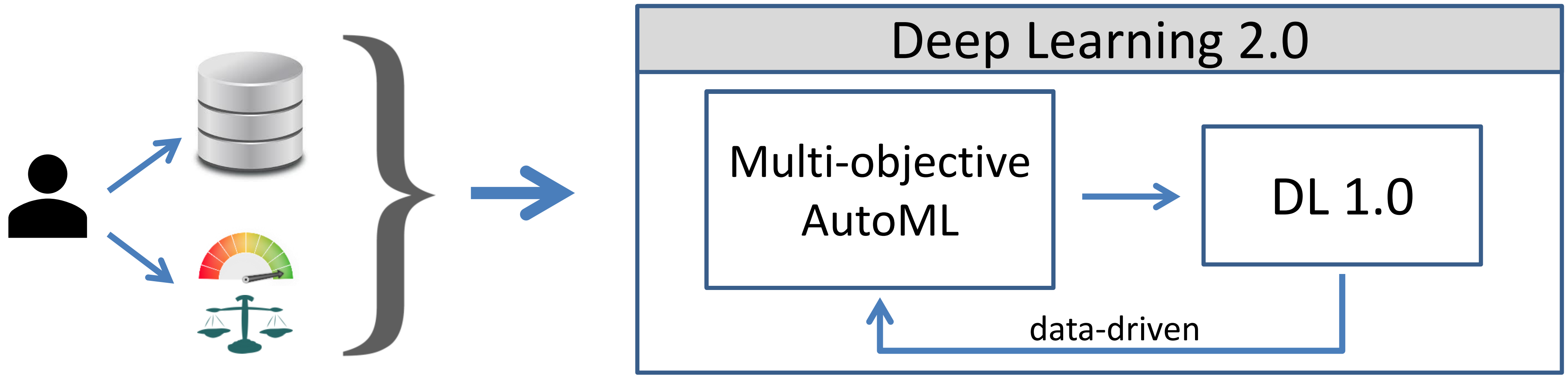
trial & error

From Deep Learning 1.0 to Deep Learning 2.0



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

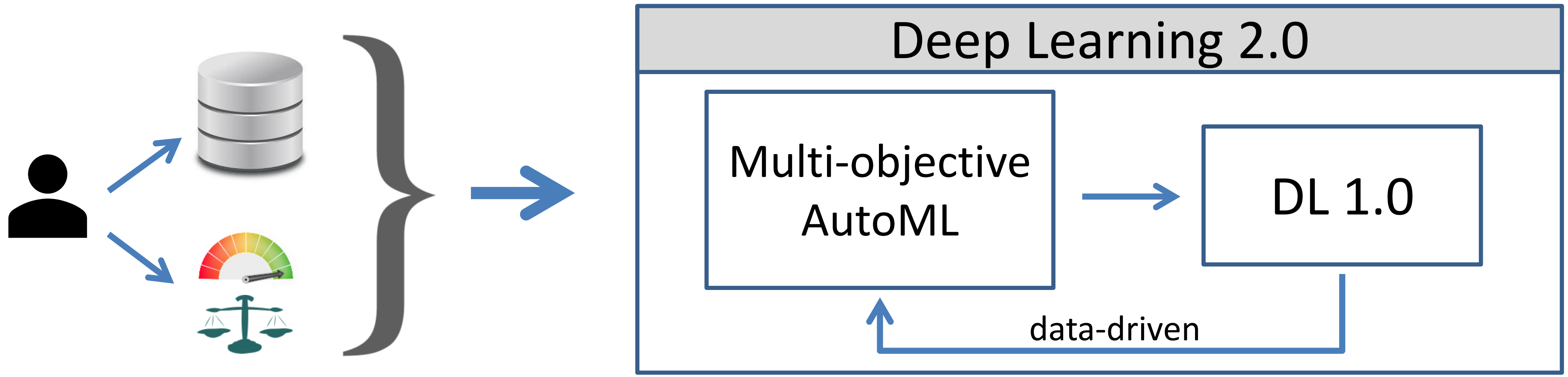
From Deep Learning 1.0 to Deep Learning 2.0



- domain expert can specify objectives

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

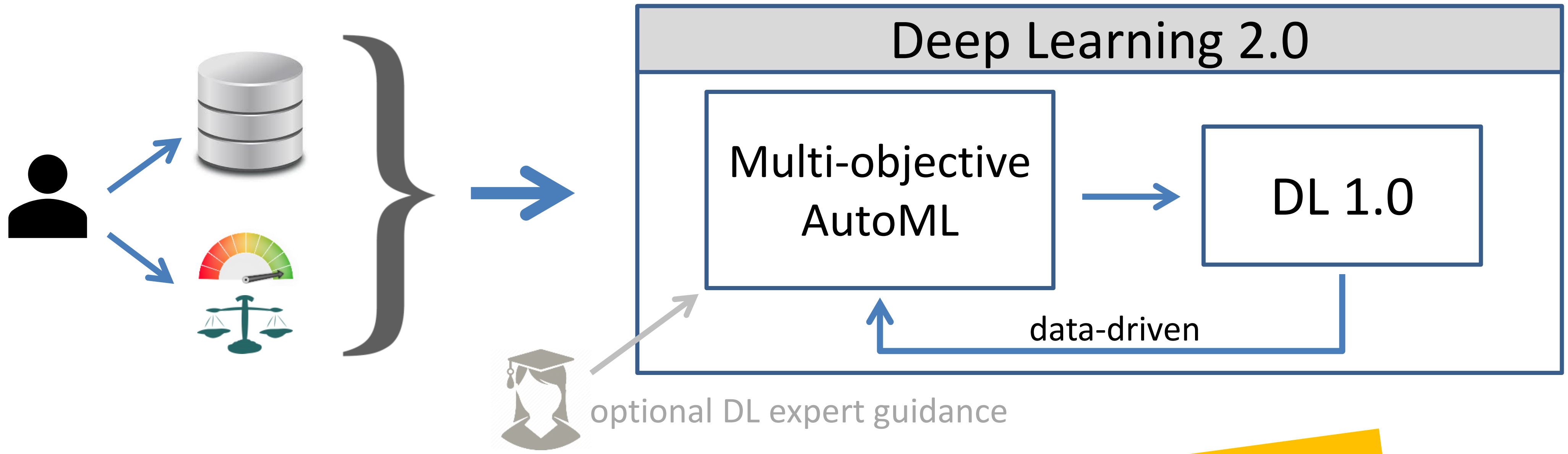
Expected Impact of Deep Learning 2.0



- **Paradigm-changing: democratizing Deep Learning**
  – DL 2.0 projects possible without a DL expert
  – DL 2.0 directly optimizes for user‘s objectives
  → Trustworthy AI by design

DL 2.0 will be even more pervasive than DL 1.0, with huge impact on the billion-dollar DL market

Integrating Human Expert Knowledge



NAS is very core to Deep Learning 2.0

- **Paradigm-changing: de... ...ing**
  - DL 2.0 proj... ...without a DL expert
  - DL 2.0 directly optimizes for user's objectives
    → Trustworthy AI by design

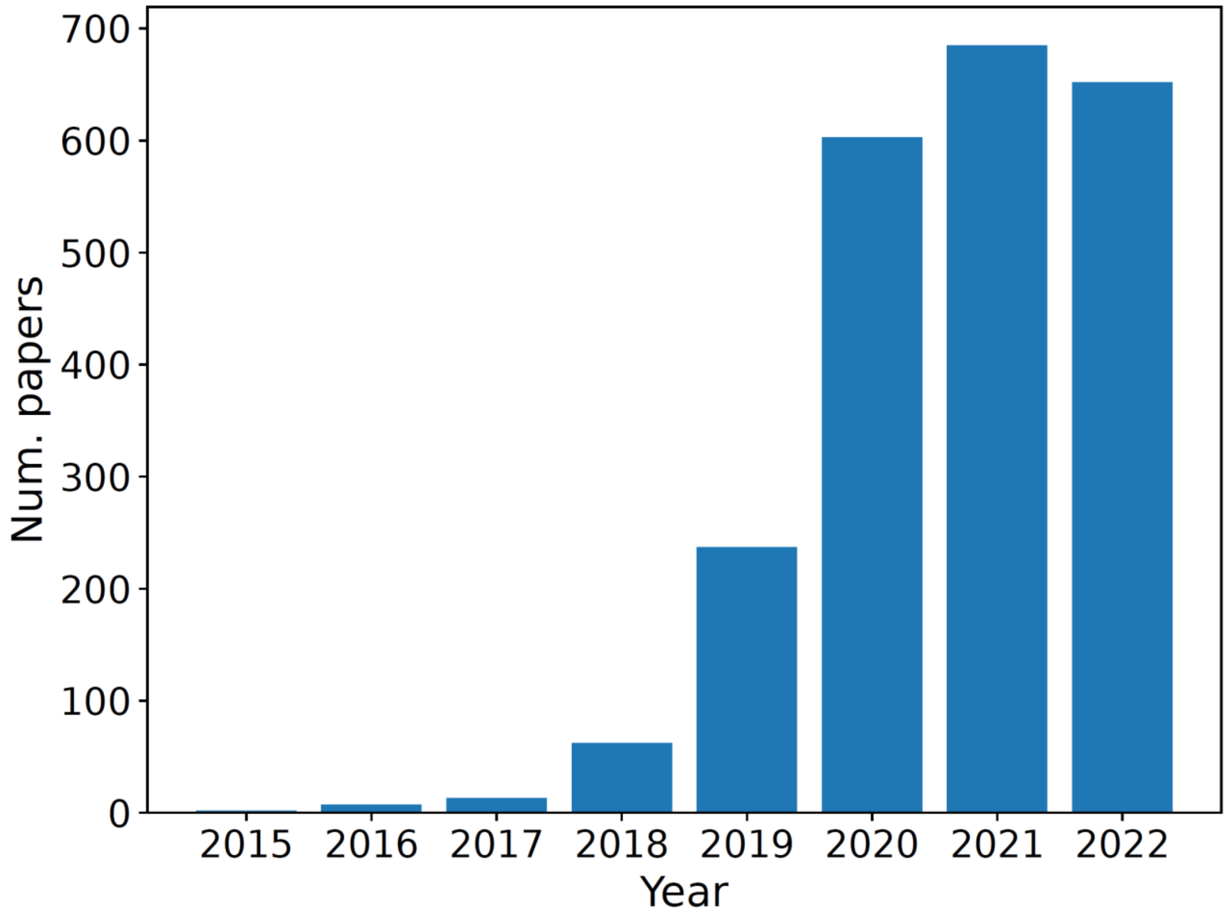
DL 2.0 will be even more pervasive than DL 1.0, with huge impact on the billion-dollar DL market

A Critical Look at the Field of NAS (by someone invested in the field)

- **Over 2000 NAS papers in the last 4 years**
  - see NAS literature list by Deng & Lindauer https://www.automl.org/automl/literature-on-neural-architecture-search/
  - see the survey by White et al, 2023

- **BUT:**
  - Transformers, ViT and ConvNext were **discovered manually, *not* by NAS**
  - Not a focus of the NAS community to create **robust & efficient AutoML systems**

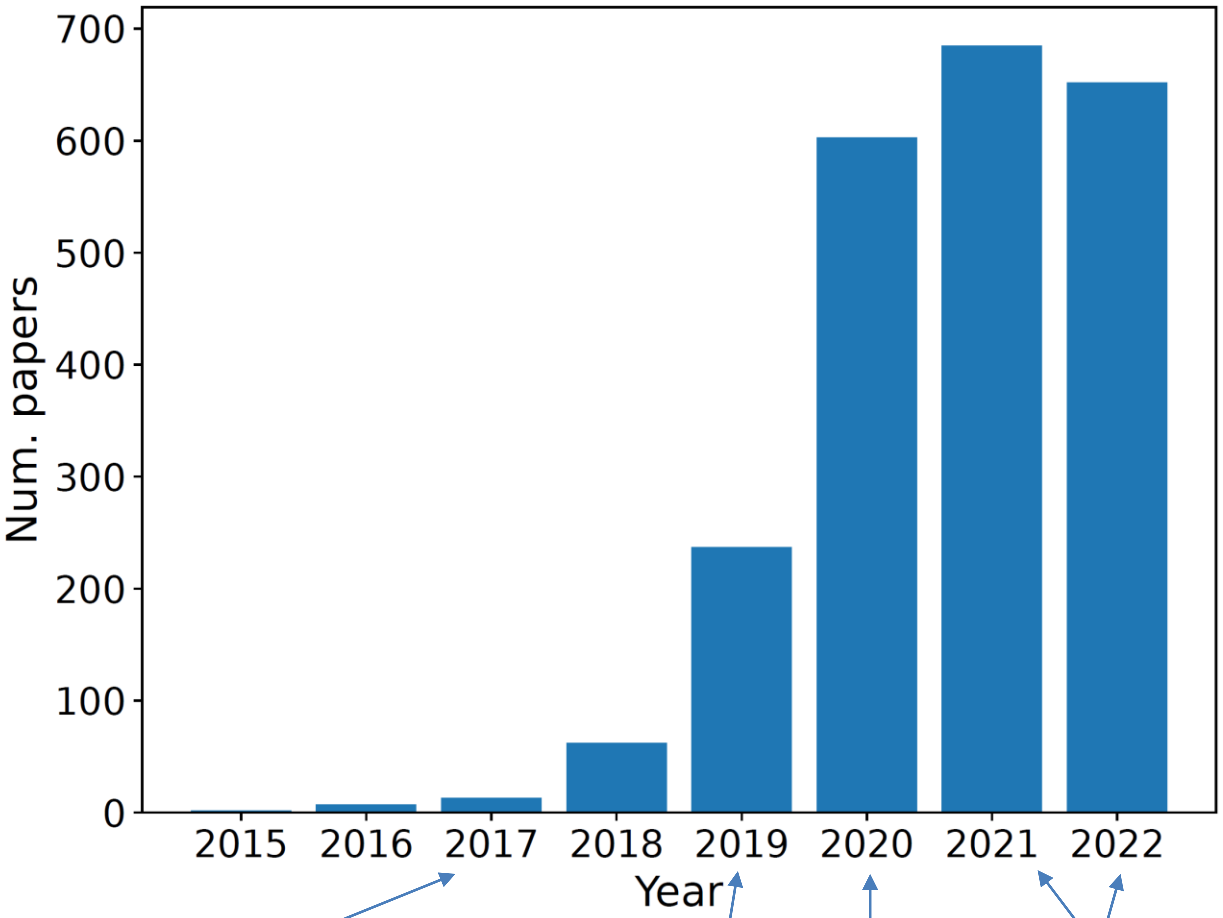
#new NAS papers per year (including arXiv, etc)



A Critical Look at the Field of NAS (by someone invested in the field)

- Early blackbox methods were **extremely expensive**
- **DARTS** is **fast** but has **catastrophic failure modes**
  - E.g., all skip connections
- **ZC proxies** very popular but not a silver bullet: very varied performance across search spaces

#new NAS papers per year (including arXiv, etc)



NAS with RL [Zoph & Le, ICLR 2017] 800 GPUs for 2 weeks for CIFAR-10

DARTS [Liu et al, ICLR 2019] 4 GPU days for CIFAR-10

[Zela et al, ICLR 2020] showing failure modes

ZC proxies e.g., [Abdelfattah et al, 2021]

Three Possible Use Cases of NAS

- Improvements of an existing architecture family
  - **Small search spaces**, e.g., layer-wise hyperparameters: #kernels, kernel sizes
  - **Reduce latency**: distillation, pruning, etc
  - **Already widely used** for **hardware-aware NAS**

Focus of this talk

- AutoML
  - Given a new dataset, robustly & efficiently make predictions
  - **Not a focus of the NAS community** (but of the lightweight NAS competition)

- **Discovering novel architectures**
  - Overcome restrictive search spaces to allow the discovery of novel architectures

Outline

- Bayesian optimization and how to speed it up
  - Bayesian optimization
  - Multi-fidelity optimization
  - Meta-learning

- Extensions of blackbox NAS
  - Transfer-NAS [Shala et al, ICML 2023 top 5%]
  - Hierarchical spaces [Schrodi et al, NeurIPS 2022 WS on meta-learning]
  - Include hyperparameters: JAHS [Bansal, NeurIPS 2022 D&B oral]
  - Multi-objective JAHS for fair face recognition [Dooley et al, NeurIPS 2022 WS on meta-learning]

Bayesian Optimization in a Nutshell

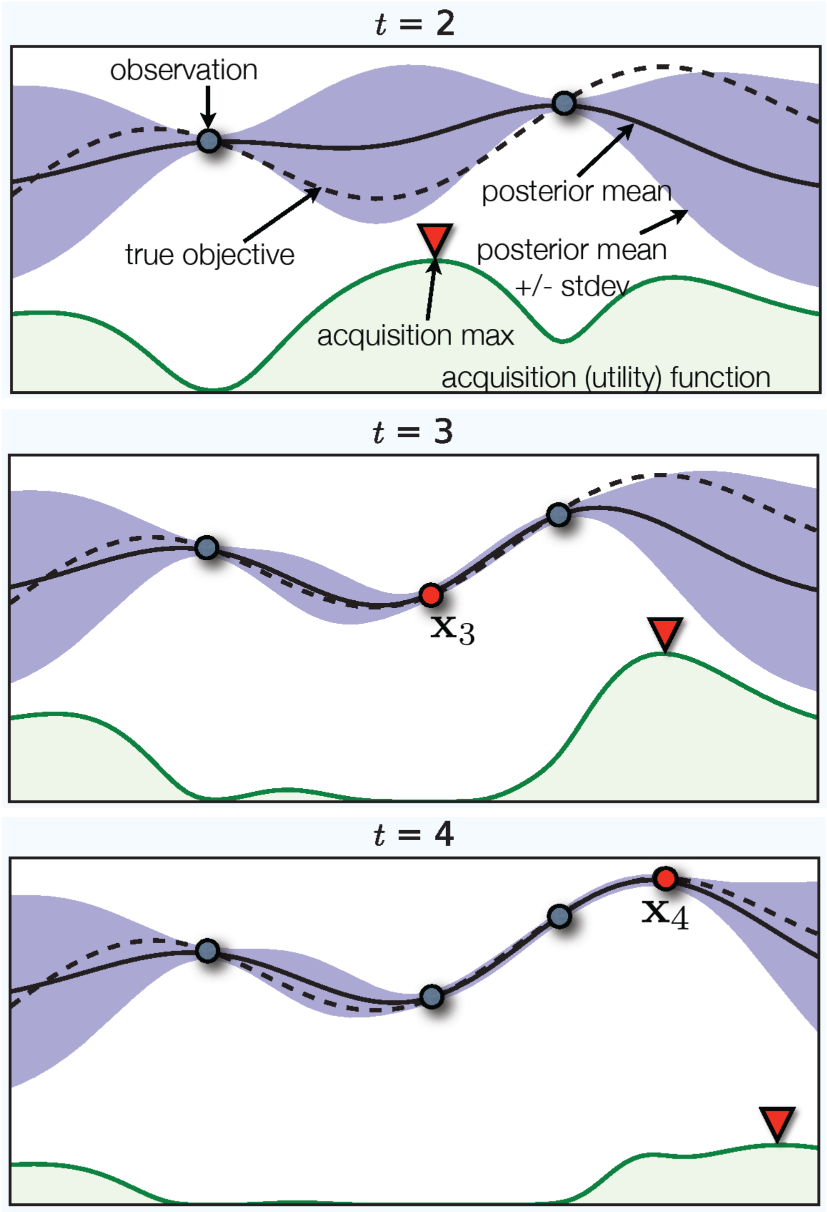
- Prominent approach to optimize **expensive blackbox** functions [Mockus et al., '78]

$$\max_{x \in X} f(x) \qquad x \rightarrow \blacksquare \rightarrow f(x)$$

- Efficient in the number of function evaluations

- Still works when objective is nonconvex, noisy, has unknown derivatives, etc

- Recent convergence results
  [Srinivas et al, '10; Bull '11;
  de Freitas, Smola, Zoghi, '12]



An Important Component of Bayes Opt: the Surrogate Model

- The Standard Model: a **Gaussian process**

  + strong calibration

  + mathematical convenience

  +/- depends crucially on the used kernel

- **Bayesian neural networks**

  + flexibility

  - not good for few data points (unless in-context learned, see [Müller et al, ICML 2023])

- **Random forests**

  + flexibility

  + strong off-the-shelf usability

Outline

- Bayesian optimization and how to speed it up
  - Bayesian optimization
  - Multi-fidelity optimization
  - Meta-learning

- Extensions of blackbox NAS
  - Transfer-NAS [Shala et al, ICML 2023 top 5%]
  - Hierarchical spaces [Schrodi et al, NeurIPS 2022 WS on meta-learning]
  - Include hyperparameters: JAHS [Bansal, NeurIPS 2022 D&B oral]
  - Multi-objective JAHS for fair face recognition [Dooley et al, NeurIPS 2022 WS on meta-learning]

Multi-Fidelity Optimization

- **Key Idea: use cheap approximations of expensive blackbox**

- **Cheap approximations exist in many applications**
  - Fewer epochs of iterative training algorithms (e.g., SGD)
  - Subset of data
  - Downsampled images in object recognition
  - Shallower/slimmer neural networks
  - Shorter MCMC chains in Bayesian deep learning
  - Fewer trials in deep reinforcement learning
  - …

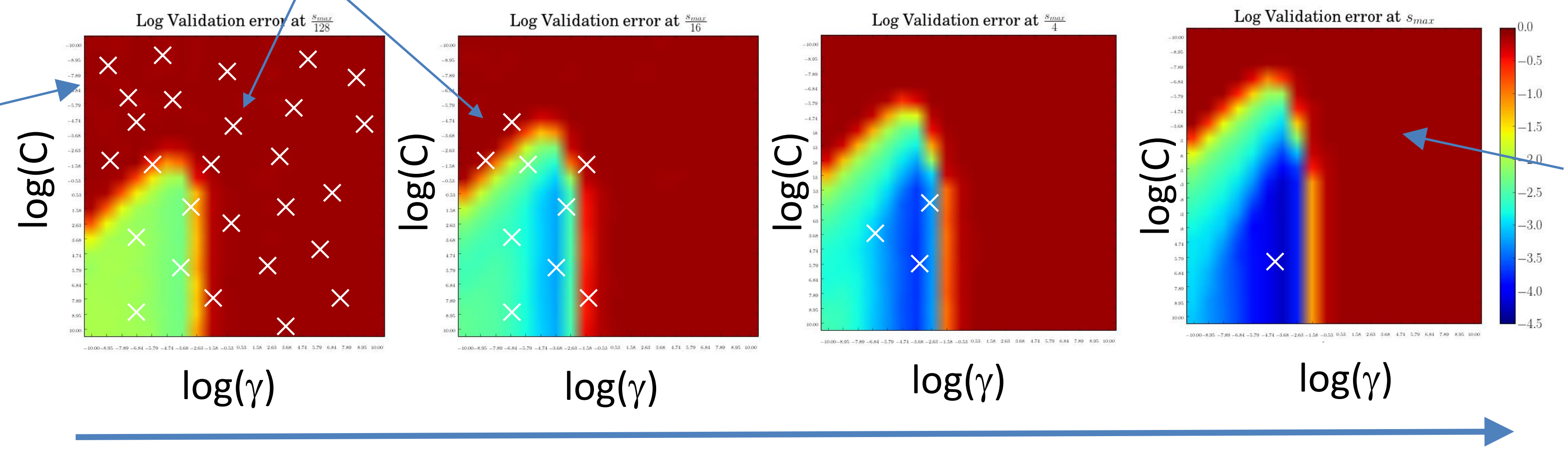
Example for Multi-Fidelity Optimization

- **One possible approximation: use a subset of the data**
  – Many cheap evaluations on small subsets
  – Few expensive evaluations on the full data

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

× denotes evaluated hyperparameter settings

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



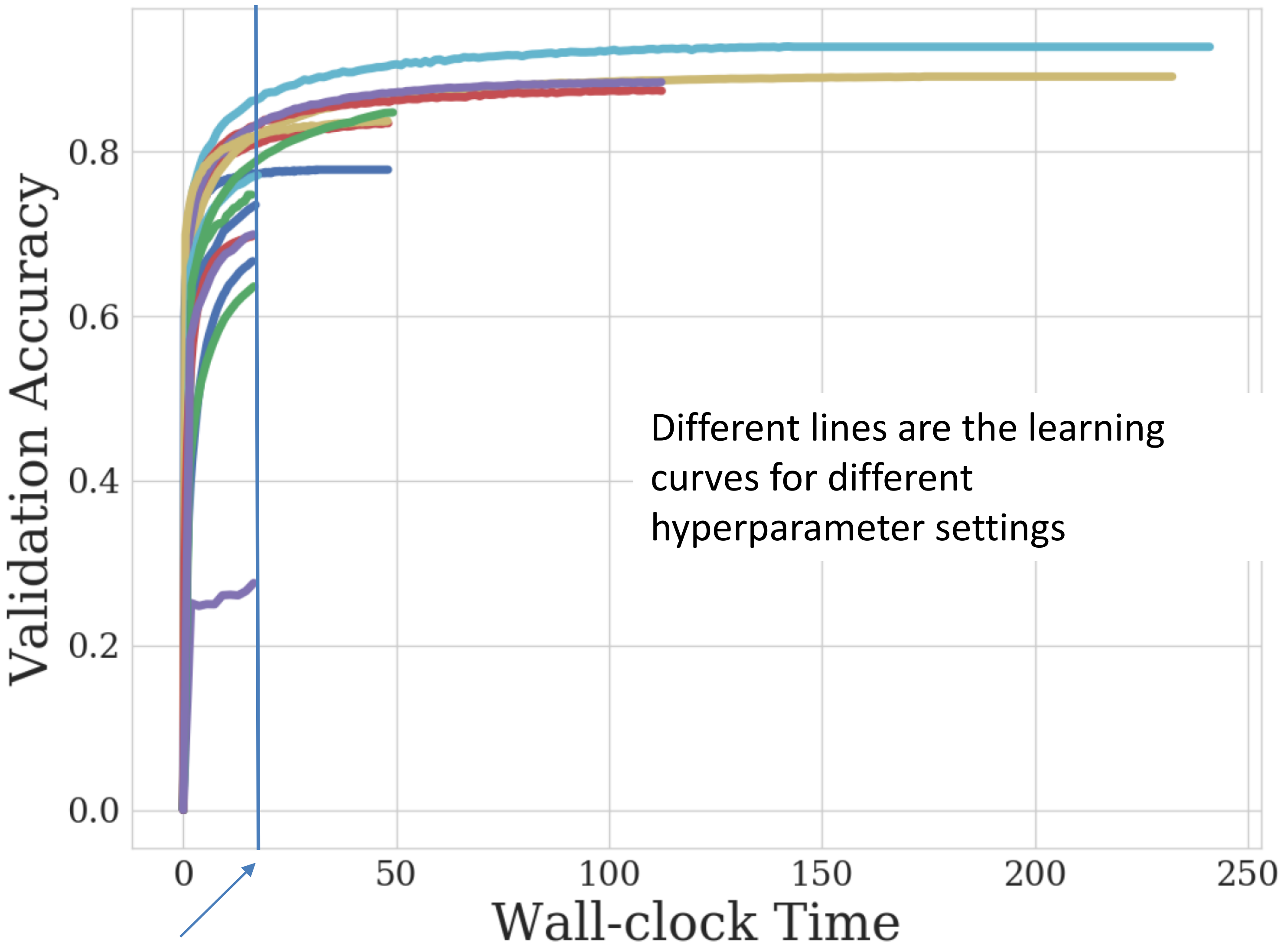
Log validation error when trained on full dataset of 50.000 data points (takes several hours to compute)

Size of subset (of MNIST)

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

Successive Halving (SH) Algorithm When the Fidelity is Runtime

[Jamieson & Talwalkar, AISTATS 2016]



Different lines are the learning curves for different hyperparameter settings

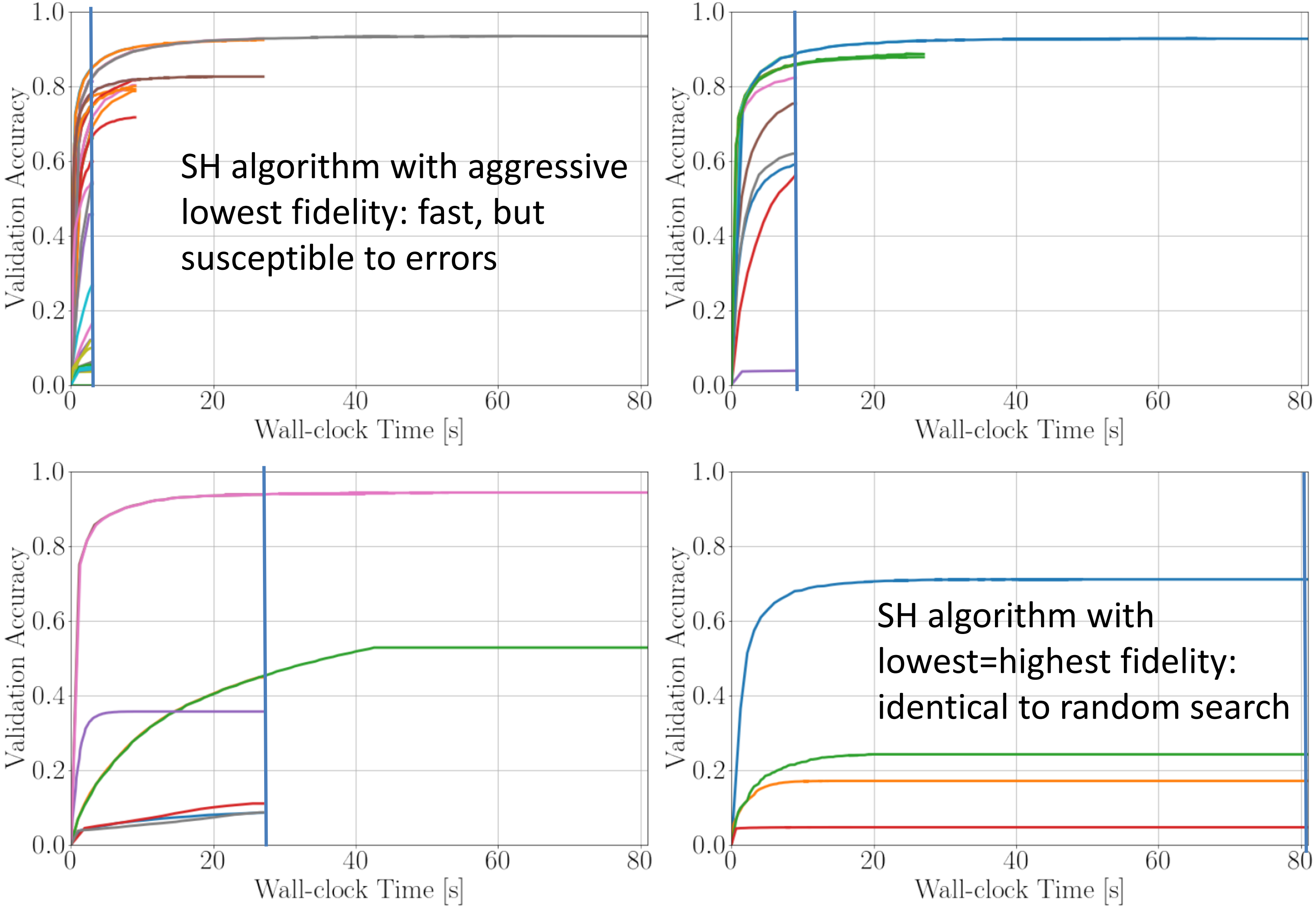
Lowest fidelity

An Extension of SH with Theoretical Guarantees: Hyperband

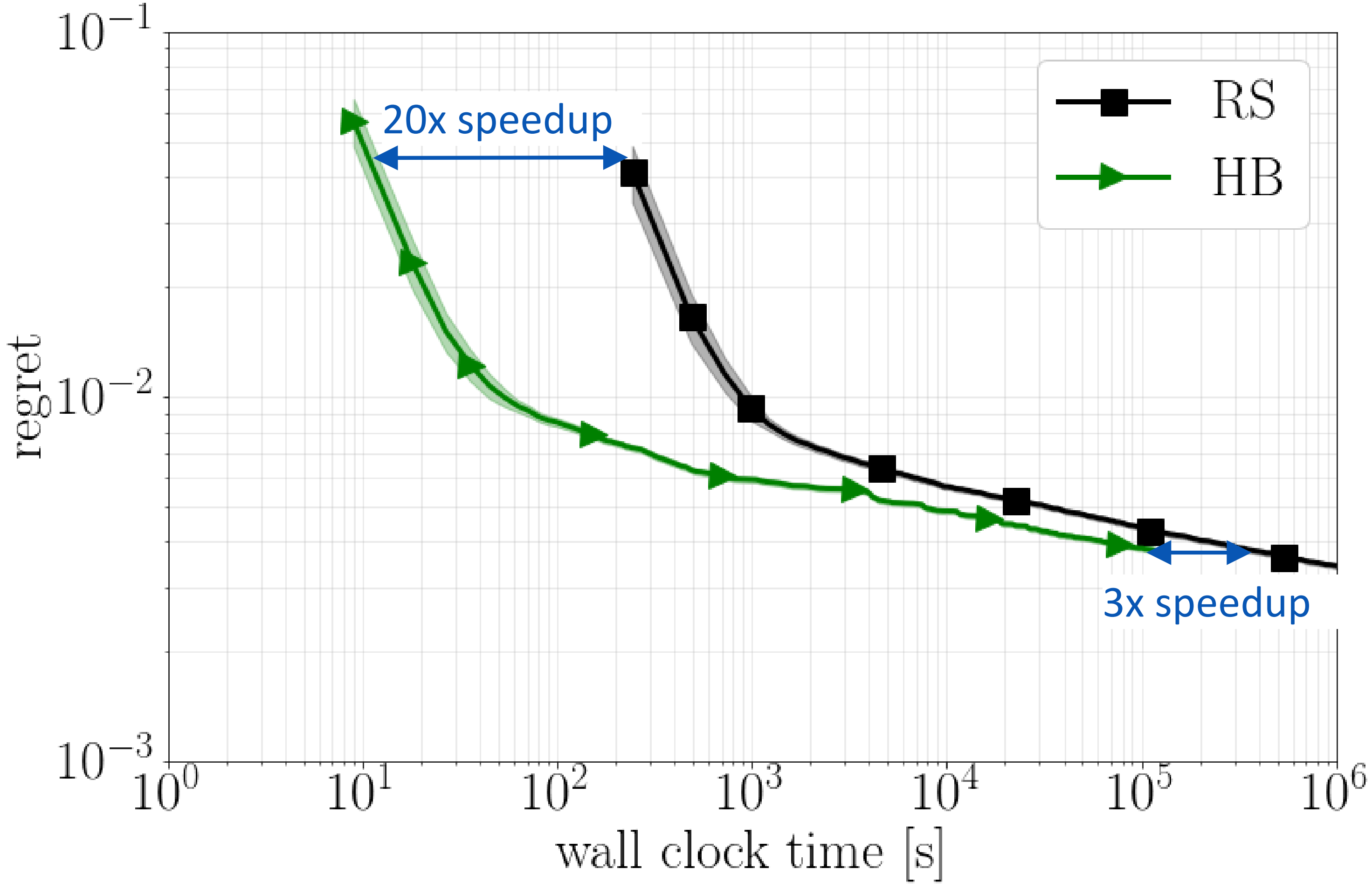
[Li et al, ICLR 2017]

- **Main idea:** hedge against errors in cheap approximations

- **Algorithm:** run multiple copies of SH in parallel, starting at different cheapest fidelities



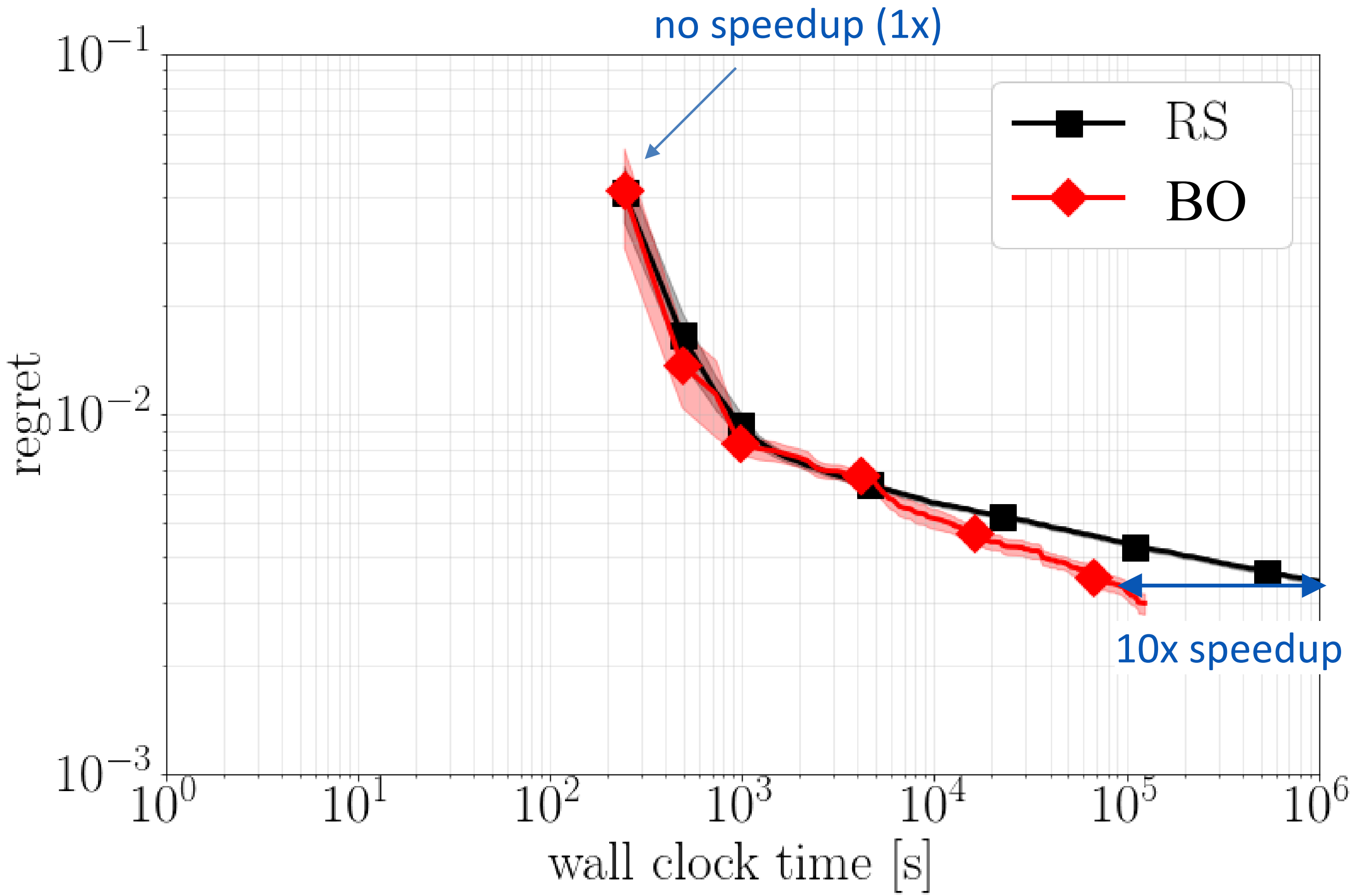
Hyperband vs. Random Search



Biggest advantage: much improved **anytime performance**

Auto-Net on dataset adult

Bayesian Optimization vs Random Search

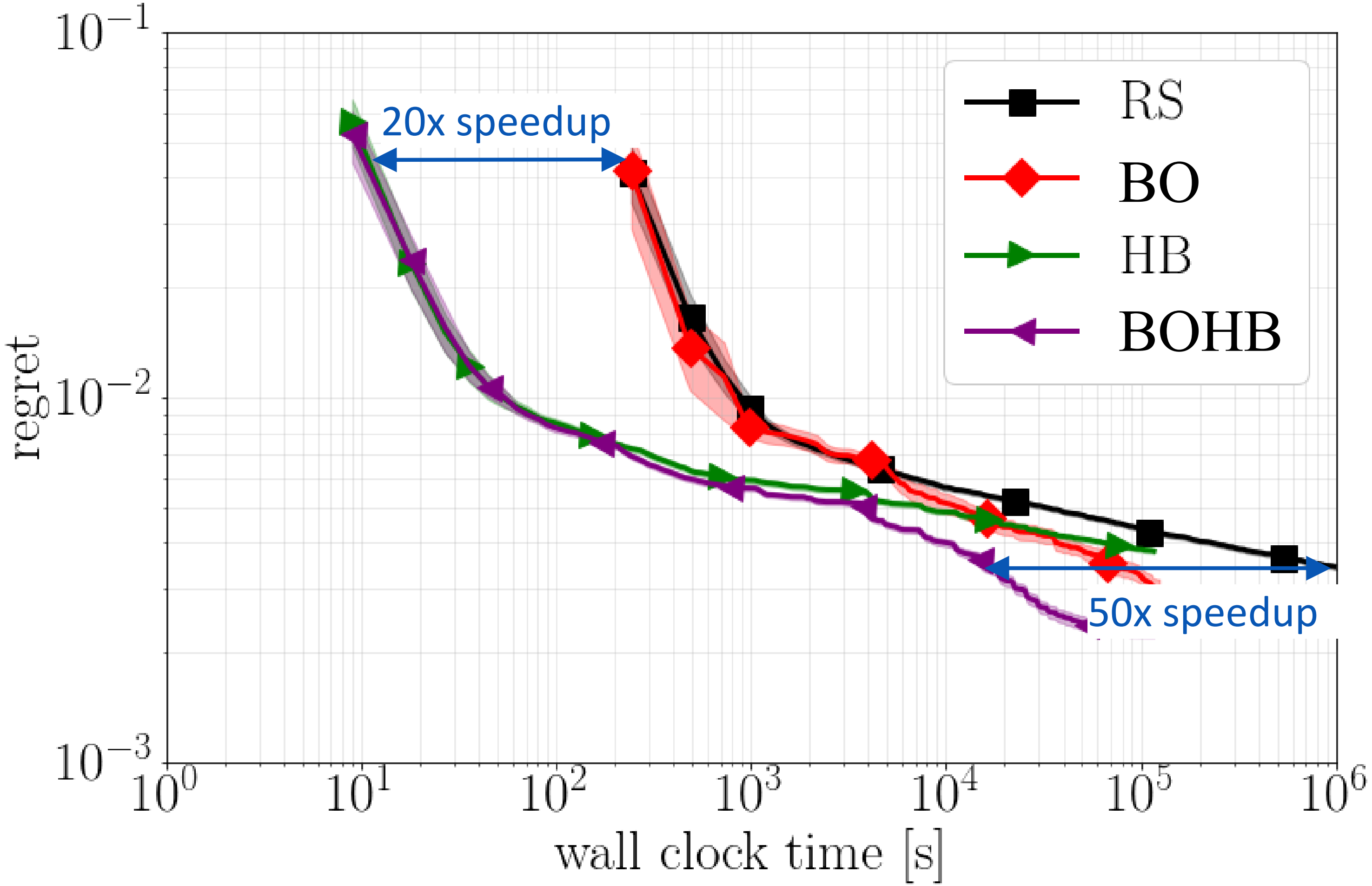


Biggest advantage: much improved **final performance**

Auto-Net on dataset adult

BOHB: Combining Bayesian Optimization & Hyperband

[Falkner, Klein & Hutter, ICML 2018]



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

Auto-Net on dataset adult

Outline

- Bayesian optimization and how to speed it up
  – Bayesian optimization
  – Multi-fidelity optimization
  ⇨ Meta-learning

- Extensions of blackbox NAS
  – Transfer-NAS [Shala et al, ICML 2023 top 5%]
  – Hierarchical spaces [Schrodi et al, NeurIPS 2022 WS on meta-learning]
  – Include hyperparameters: JAHS [Bansal, NeurIPS 2022 D&B oral]
  – Multi-objective JAHS for fair face recognition [Dooley et al, NeurIPS 2022 WS on meta-learning]

Meta-learning Across Datasets

- There are many different ways to meta-learn across datasets

- Relevant here
  – Pre-train Bayesian optimization‘s surrogate model across datasets

Outline

- Bayesian optimization and how to speed it up
  - Bayesian optimization
  - Multi-fidelity optimization
  - Meta-learning

- Extensions of blackbox NAS
  - Transfer-NAS [Shala et al, ICML 2023 top 5%]
  - Hierarchical spaces [Schrodi et al, NeurIPS 2022 WS on meta-learning]
  - Include hyperparameters: JAHS [Bansal, NeurIPS 2022 D&B oral]
  - Multi-objective JAHS for fair face recognition [Dooley et al, NeurIPS 2022 WS on meta-learning]

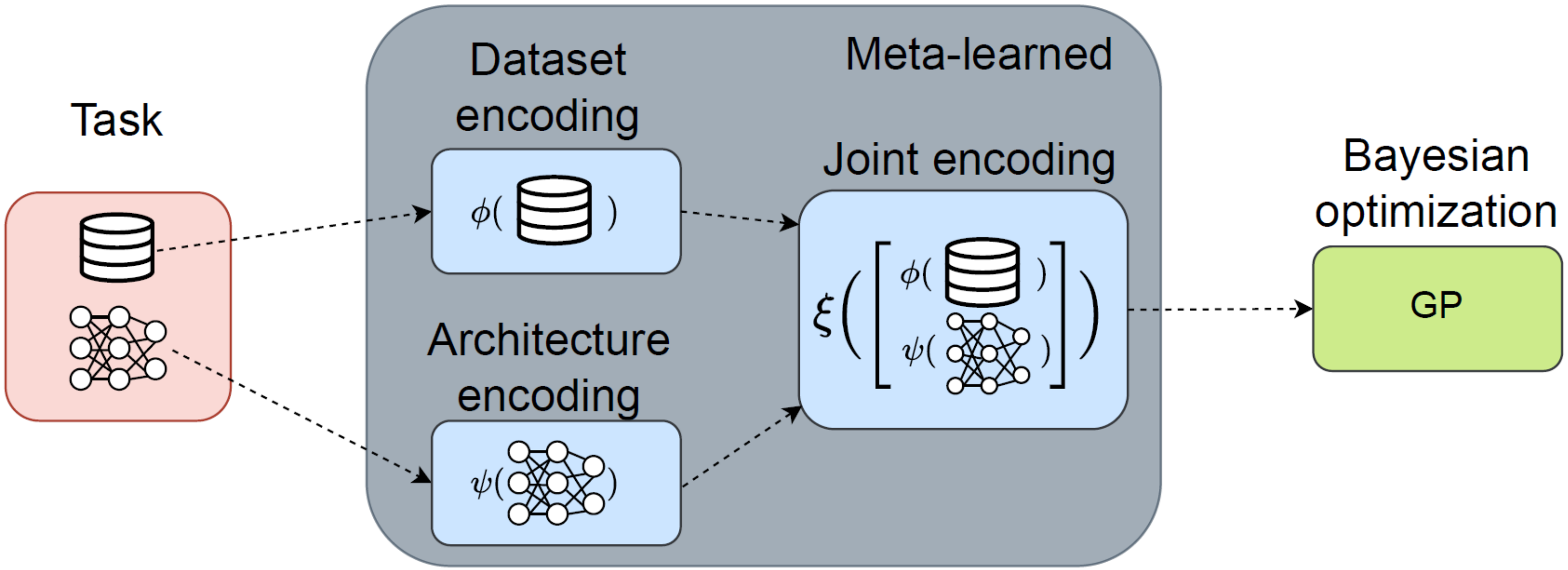
Surrogate Model Used: GPs with Deep Kernels

- Combine the benefits of Gaussian processes (GPs) and neural networks (NNs)
  - GPs:
    - Calibrated uncertainty quantification
    - Strong performance with few samples
  - NNs
    - Flexibility to learn meaningful features
    - Scalability

- High-level overview
  - Use neural networks to learn embeddings for inputs
  - Use Gaussian processes with a kernel function on top of the embeddings

Transfer NAS with Meta-learned Bayesian Surrogates

- Surrogate model
  - Learn a dataset embedding $\phi : \mathcal{D} \to \mathbb{R}^L$ (with a transformer)
  - Learn an architecture embedding $\psi : \chi \to \mathbb{R}^K$ (with a graph neural network)
  - Fit a kernel $\xi : \mathbb{R}^{K+L} \to \mathbb{R}^M$ on the concatenation of these embeddings
  - Meta-learn these embeddings across datasets and architectures



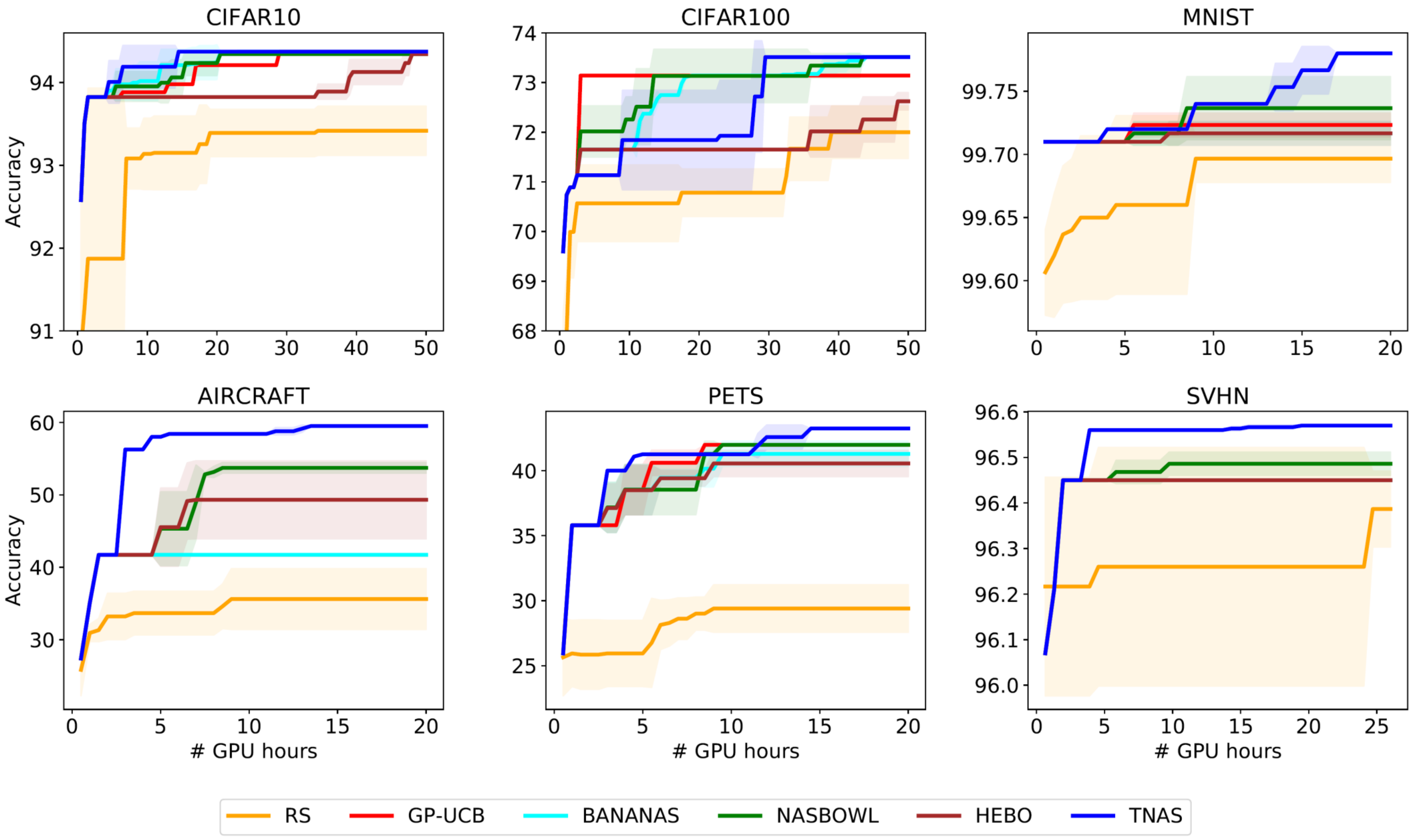
The Rest of the Pipeline

- Experimental setup taken from RapidNAS [Lee et al, ICLR 2021]
  - Meta-train data: 4230 subtasks of downsampled ImageNet with 20 classes each
  - Test data: MNIST, SVHN, CIFAR-10, CIFAR-100, Aircraft, and Oxford-IIIT Pets

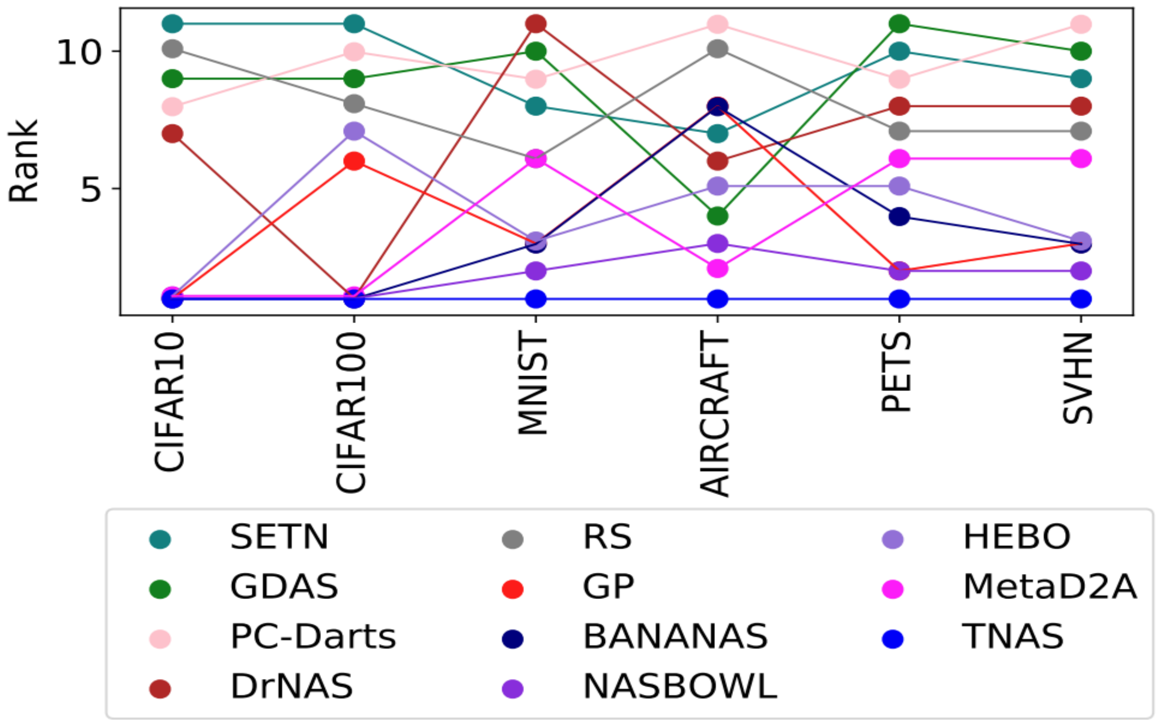
- Search space
  - NB201 (open question: performance with general search spaces)

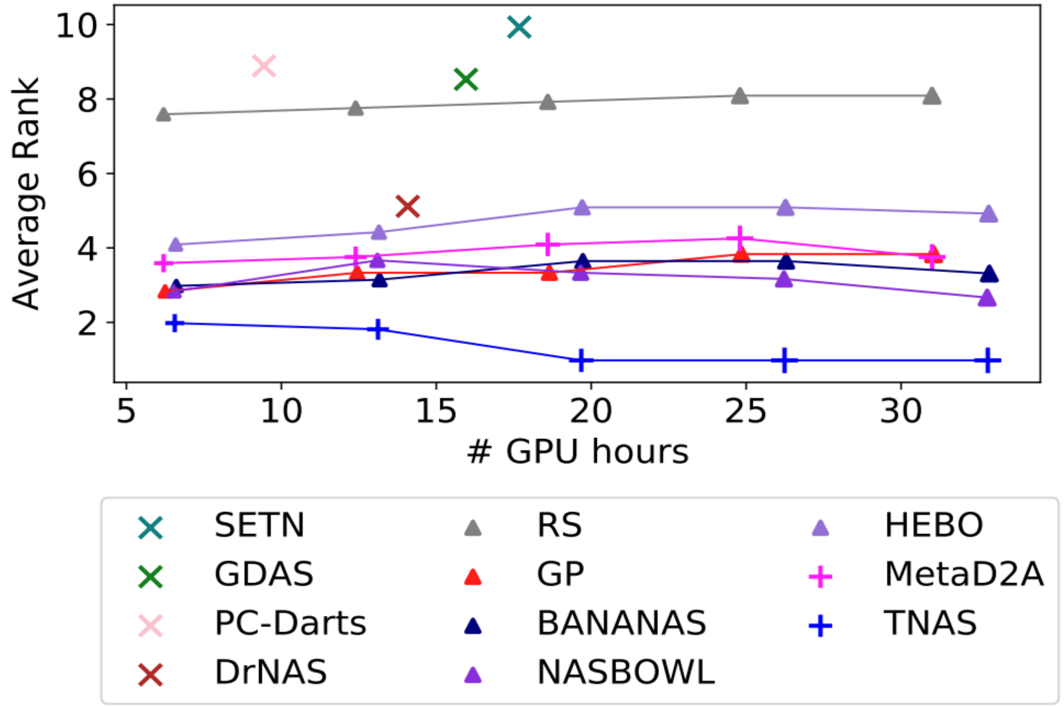
- Standard Bayesian optimization, for up to 30 evaluations
  - With deep GPs based on the meta-learned combined kernel
  - Fine-tune the embeddings after each function evaluation

Results Over Time



Results: Robustness & Efficiency





- Transfer-NAS yields the best results on six different image classification tasks
- It is also the most robust approach, even for very short runtimes
  - In particular, it **performs favourably against one-shot methods**

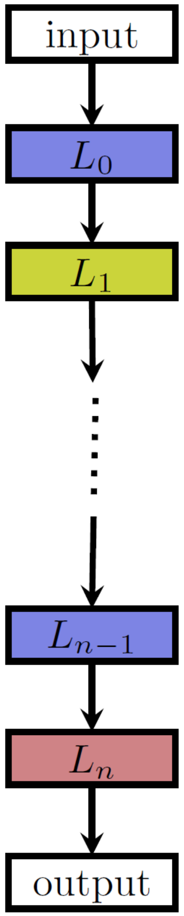
Outline

- Bayesian optimization and how to speed it up
  - Bayesian optimization
  - Multi-fidelity optimization
  - Meta-learning

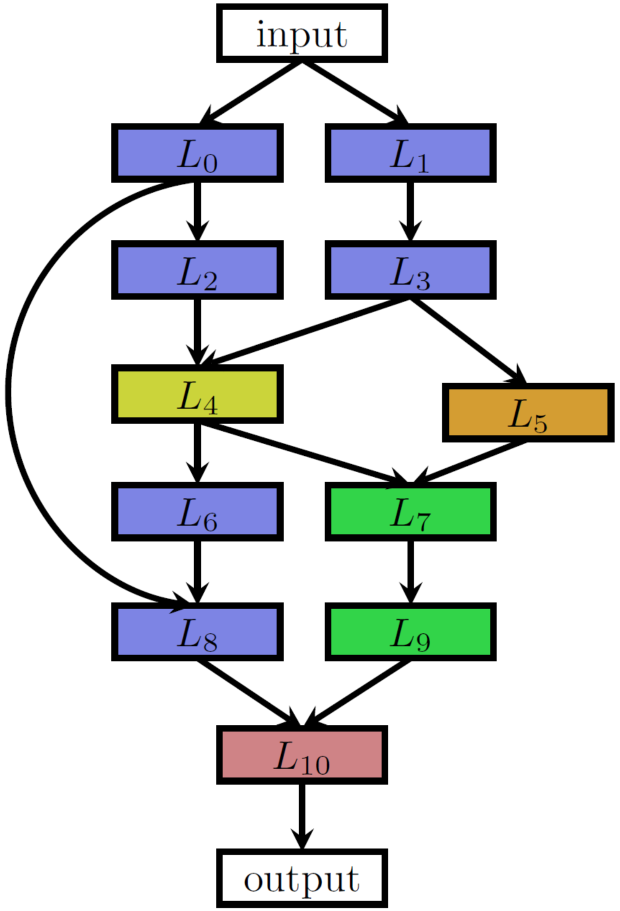
- Extensions of blackbox NAS
  - Transfer-NAS [Shala et al, ICML 2023 top 5%]
  - Hierarchical spaces [Schrodi et al, NeurIPS 2022 WS on meta-learning]
  - Include hyperparameters: JAHS [Bansal, NeurIPS 2022 D&B oral]
  - Multi-objective JAHS for fair face recognition [Dooley et al, NeurIPS 2022 WS on meta-learning]

Basic Neural Architecture Search Spaces

[Elsken et al., JMLR 2019]



Chain-structured space
(different colours:
different layer types)



More complex space
with multiple branches
and skip connections

Cell Search Spaces

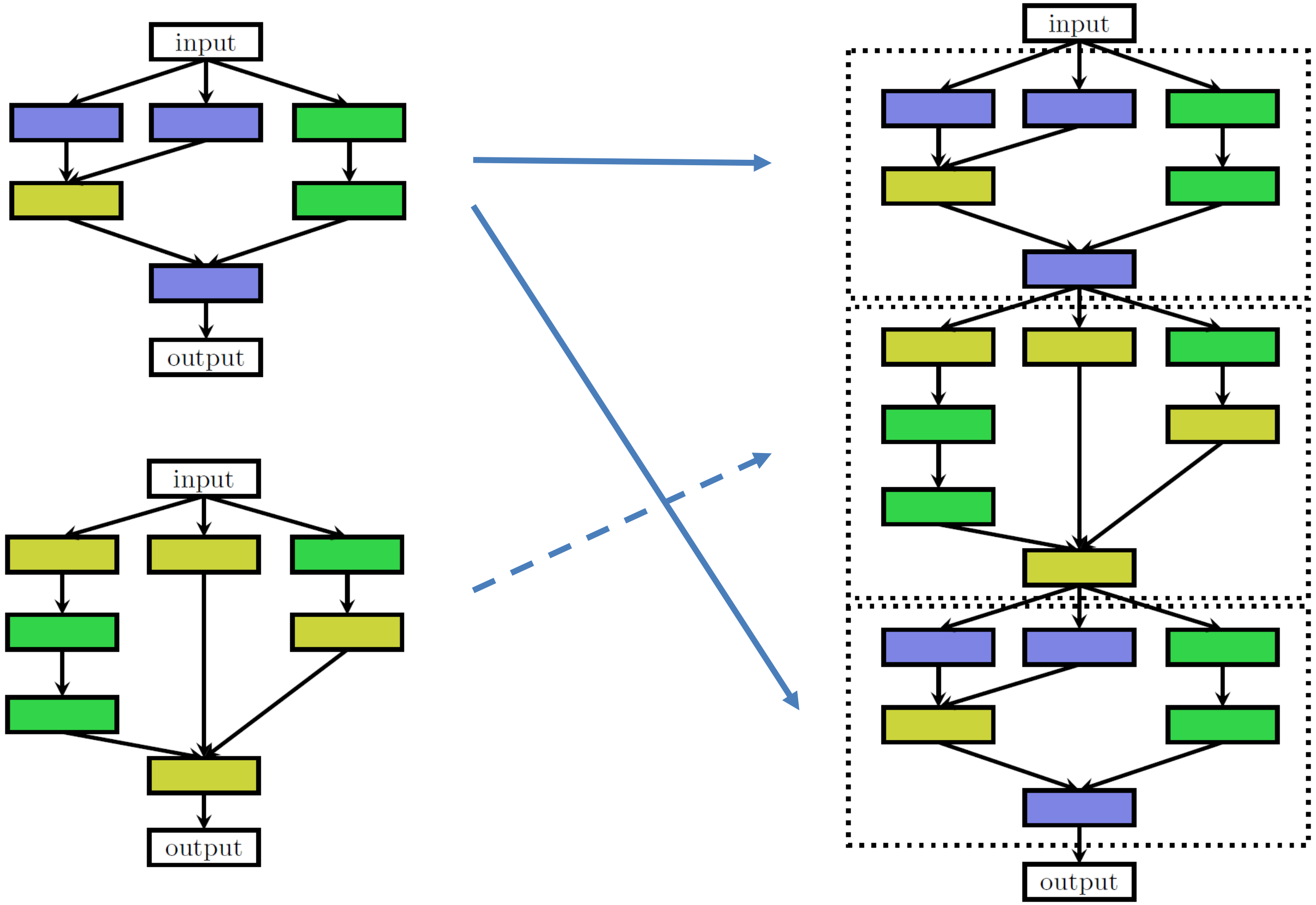
Introduced by Zoph et al. [CVPR 2018]

Two possible cells

Architecture composed of stacking together individual cells

normal cell: preserves spatial resolution

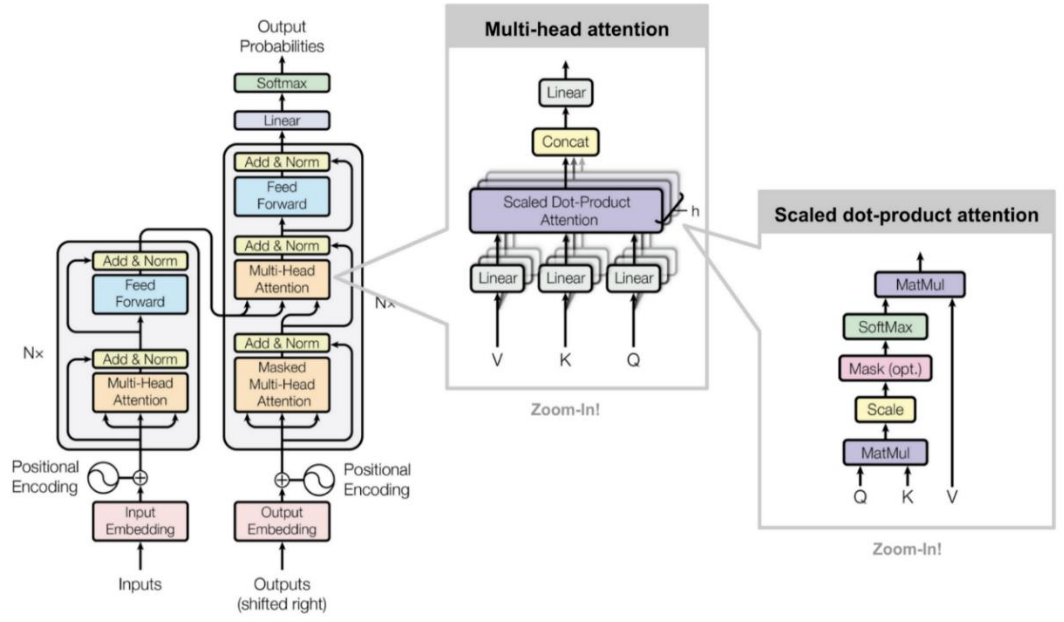
reduction cell: reduces spatial resolution



Hierarchical Cell Spaces

- Reuse of substructures, like in cell search spaces
- But choices on many different levels
- Some examples in the literature (e.g., [Liu et al, ICLR 2018]), but understudied

- Potential example for an element of a hierarchical space: transformers



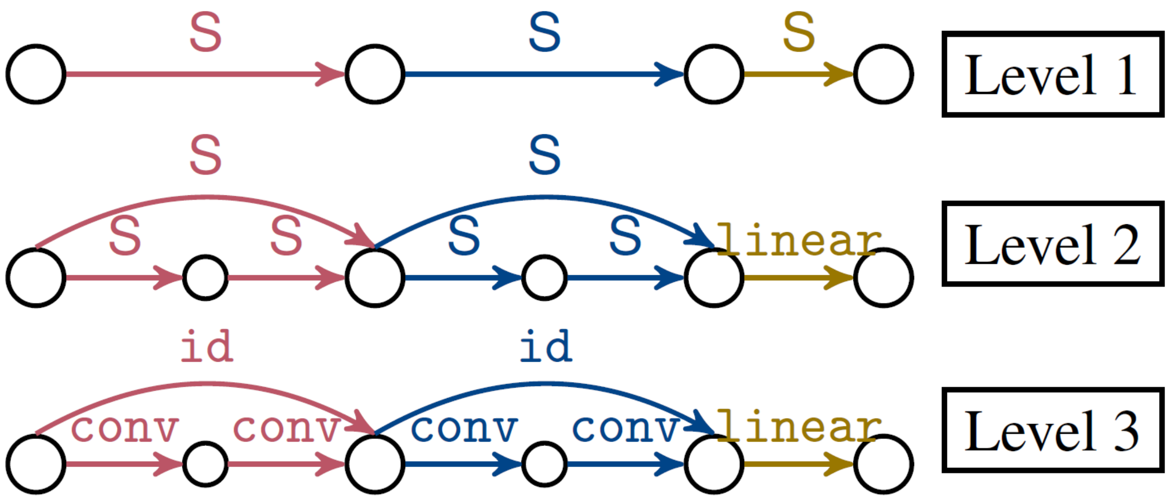
Formulation of Search Spaces by Context-free Grammars

- Choices on multiple levels of the architecture
- Can be described by context-free grammars
- At each level, we apply another production rule to add more detail

$$\mathrm{S} \dashrightarrow \texttt{Linear}(\mathrm{S},\ \mathrm{S},\ \mathrm{S})$$

$$\dashrightarrow \texttt{Linear}(\texttt{Residual}(\mathrm{S},\ \mathrm{S},\ \mathrm{S}),\ \texttt{Residual}(\mathrm{S},\ \mathrm{S},\ \mathrm{S}),\ \texttt{linear})$$

$$\dashrightarrow \texttt{Linear}(\texttt{Residual}(\texttt{conv},\ \texttt{id},\ \texttt{conv}),\ \texttt{Residual}(\texttt{conv},\ \texttt{id},\ \texttt{conv}),\ \texttt{linear})$$



Example Grammar: Hierarchical Extension of NB201

$$
\begin{aligned}
\mathrm{D2} &::= \texttt{Sequential3}(\mathrm{D1},\ \mathrm{D1},\ \mathrm{D0}) \mid \texttt{Sequential3}(\mathrm{D0},\ \mathrm{D1},\ \mathrm{D1}) \mid \texttt{Sequential4}(\mathrm{D1},\ \mathrm{D1},\ \mathrm{D0},\ \mathrm{D0}) \\
\mathrm{D1} &::= \texttt{Sequential3}(\mathrm{C},\ \mathrm{C},\ \mathrm{D}) \mid \texttt{Sequential4}(\mathrm{C},\ \mathrm{C},\ \mathrm{C},\ \mathrm{D}) \mid \texttt{Residual3}(\mathrm{C},\ \mathrm{C},\ \mathrm{D},\ \mathrm{D}) \\
\mathrm{D0} &::= \texttt{Sequential3}(\mathrm{C},\ \mathrm{C},\ \mathrm{CL}) \mid \texttt{Sequential4}(\mathrm{C},\ \mathrm{C},\ \mathrm{C},\ \mathrm{CL}) \mid \texttt{Residual3}(\mathrm{C},\ \mathrm{C},\ \mathrm{CL},\ \mathrm{CL}) \\
\mathrm{D} &::= \texttt{Sequential2}(\mathrm{CL},\ \texttt{down}) \mid \texttt{Sequential3}(\mathrm{CL},\ \mathrm{CL},\ \texttt{down}) \mid \texttt{Residual2}(\mathrm{CL},\ \texttt{down},\ \texttt{down}) \\
\mathrm{C} &::= \texttt{Sequential2}(\mathrm{CL},\ \mathrm{CL}) \mid \texttt{Sequential3}(\mathrm{CL},\ \mathrm{CL},\ \mathrm{CL}) \mid \texttt{Residual2}(\mathrm{CL},\ \mathrm{CL},\ \mathrm{CL})
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{CL} &::= \texttt{Cell}(\mathrm{OP},\ \mathrm{OP},\ \mathrm{OP},\ \mathrm{OP},\ \mathrm{OP},\ \mathrm{OP}) \\
\mathrm{OP} &::= \texttt{zero} \mid \texttt{id} \mid \mathrm{CONVBLOCK} \mid \texttt{avg\_pool} \\
\mathrm{CONVBLOCK} &::= \texttt{Sequential3}(\mathrm{ACT},\ \mathrm{CONV},\ \mathrm{NORM}) \\
\mathrm{ACT} &::= \texttt{relu} \mid \texttt{hardswish} \mid \texttt{mish} \\
\mathrm{CONV} &::= \texttt{conv1x1} \mid \texttt{conv3x3} \mid \texttt{dconv3x3} \\
\mathrm{NORM} &::= \texttt{batch} \mid \texttt{instance} \mid \texttt{layer} \quad .
\end{aligned}
$$

Blue: additional macro-level choices

Red: original NB201

Brown: additional low-level choices

Size of combined search space: $\approx 10^{446}$ architectures

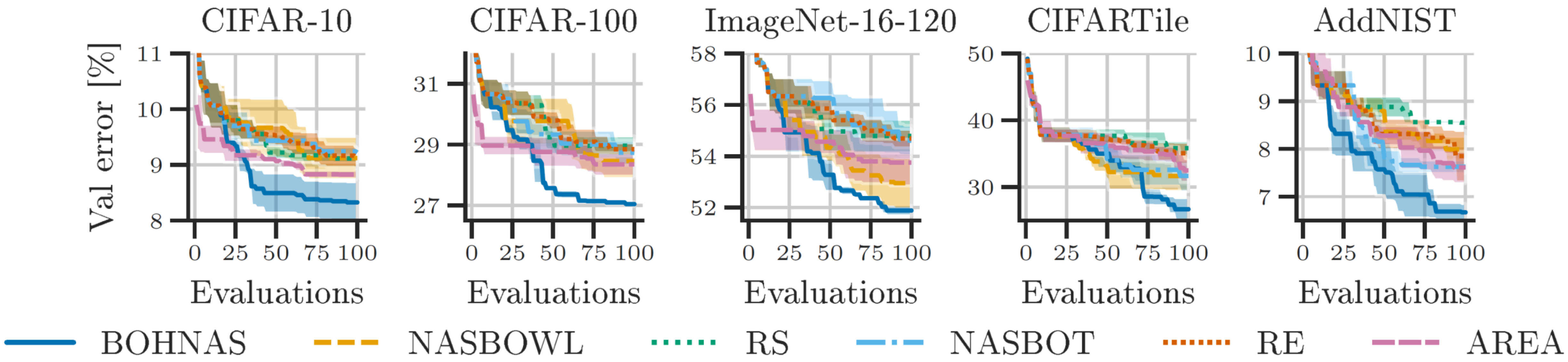
Search Strategy

- Bayesian optimization with a special kernel
- Hierarchical Weisfeiler-Lehman kernel (hWL)
  - WL kernel as in NAS-BOWL [Ru et al, ICLR 2021]
  - Apply WL kernel for each level of abstraction:

$$
\begin{aligned}
F_3(\omega) &= \texttt{Sequential}(\texttt{Residual}(\texttt{conv},\ \texttt{id},\ \texttt{conv}),\ \texttt{Residual}(\texttt{conv},\ \texttt{id},\ \texttt{conv}),\ \texttt{linear})\ , \\
F_2(\omega) &= \texttt{Sequential}(\texttt{Residual},\ \texttt{Residual},\ \texttt{linear})\quad , \\
F_1(\omega) &= \texttt{Sequential}\quad .
\end{aligned}
$$

- Fostering regularity through substitution
  - Reusing partial architectures at multiple places in the architecture
  - The reused architecture can itself contain a reused part

Result 1: The Search Strategy is Much More Efficient



Our approach: BOHNAS

Result 2: The Found Results are Competitive

Dataset Training prot.	C10 NB201†	C100 NB201†	IM16-120 NB201†	CTile NB201†	AddNIST NB201†	C10 Act. func.	C10 DARTS	IM DARTS (transfer)
Previous best	5.63	26.51	53.15	35.75	7.4	**8.32**	**2.65***	24.63*
BOHNAS	**5.02**	**25.41**	**48.16**	**30.33**	**4.57**	8.31	2.68	**24.48**
Difference	+0.51	+1.1	+4.99	+5.42	+2.83	+0.01	-0.03	+0.15

† includes hierarchical search space variants. * reproduced results using the reported genotype.

- Often better performance than any previous method
  - Also better performance than the *best* architecture in the NB201 space
  - With 100 function evaluations
- Activation function search: 1000 function evaluations;
  better than ReLU (8.93) and Swish (8.61) based on Swish search space

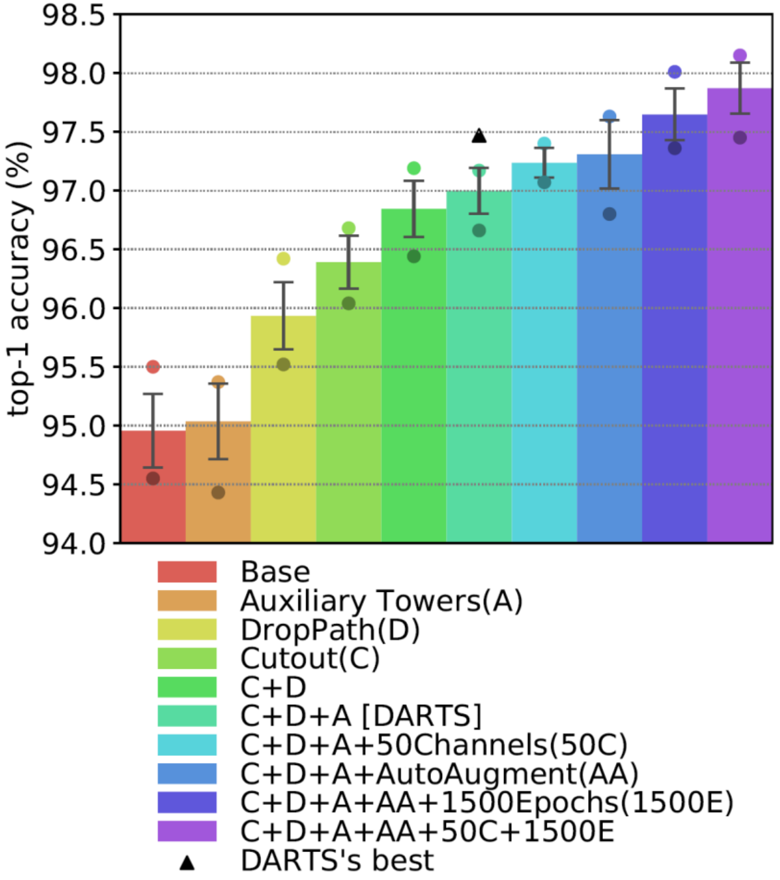
Outline

- Bayesian optimization and how to speed it up
  - Bayesian optimization
  - Multi-fidelity optimization
  - Meta-learning

- Extensions of blackbox NAS
  - Transfer-NAS [Shala et al, ICML 2023 top 5%]
  - Hierarchical spaces [Schrodi et al, NeurIPS 2022 WS on meta-learning]
  - Include hyperparameters: JAHS [Bansal, NeurIPS 2022 D&B oral]
  - Multi-objective JAHS for fair face recognition [Dooley et al, NeurIPS 2022 WS on meta-learning]

Hyperparameters are Important

- Hyperparameters can be more important than architectures
- E.g., [Yang et al, ICLR 2020: "NAS Evaluation is Frustratingly Hard"]
- There are interaction effects between architectures and hyperparameters



JAHS-Bench-201: a Large Scale Benchmark for Joint NAS+HPO

[Bansal et al, NeurIPS Datasets & Benchmarks Track 2022, oral]

- **Joint Architecture and Hyperparameter Search (JAHS)**
- JAHS-Bench-201 extends the prominent NAS-Bench-201 with
  - 4 different hyperparameters
  - 4 different fidelities

Space	Property	Description	# Values
Architecture	Cell Space	NAS-Bench-201	15,625
Hyperparameter	Activation	ReLU/Hardswish/Mish	3
	Learning Rate	$[10^{-3}, 10^{0}]$	Continuous
	Weight Decay	$[10^{-5}, 10^{-2}]$	Continuous
	Trivial Augment	On/Off	2
Fidelity	N	Depth Multiplier	3
	W	Width Multiplier	3
	R	Resolution Multiplier	3
	Epoch	# Training epochs	200

- Evaluations on 3 data sets
- 140 million performance data points
- The largest database of neural network performance to date

Adding Hyperparameters into NAS

- Impossible for some NAS approaches
  - One-shot NAS
  - ZC proxies

- Trivial for blackbox NAS
  - Simply extend the search space by the hyperparameters
  - Bayesian optimization then over a joint space of architectures and hyperparameters

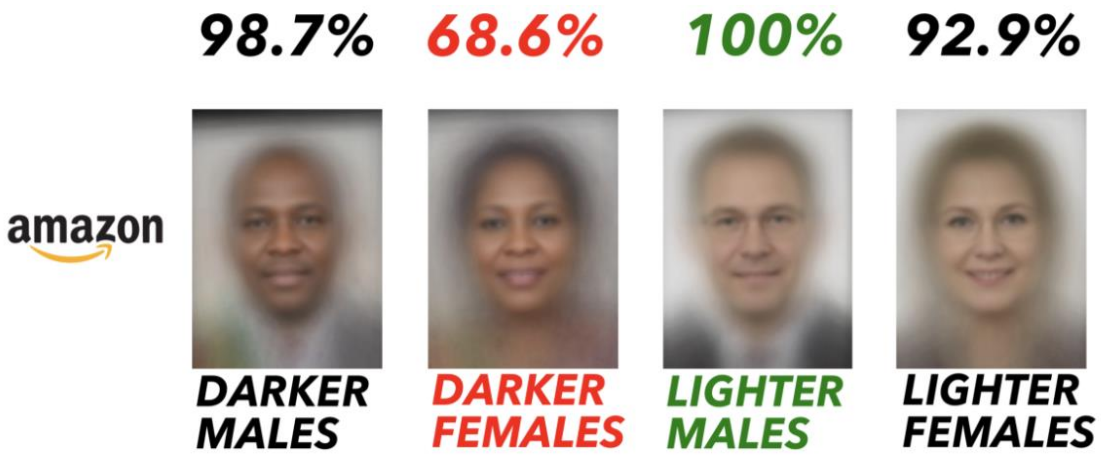
Outline

- Bayesian optimization and how to speed it up
  - Bayesian optimization
  - Multi-fidelity optimization
  - Meta-learning

- Extensions of blackbox NAS
  - Transfer-NAS [Shala et al, ICML 2023 top 5%]
  - Hierarchical spaces [Schrodi et al, NeurIPS 2022 WS on meta-learning]
  - Include hyperparameters: JAHS [Bansal, NeurIPS 2022 D&B oral]
  - Multi-objective JAHS for fair face recognition [Dooley et al, NeurIPS 2022 WS on meta-learning]

Can DL 2.0 Help with Fairness? A Case Study in Face Recognition

- Facial recognition (FR) systems are known to exhibit bias
  - sociodemographic dimensions, like gender and race



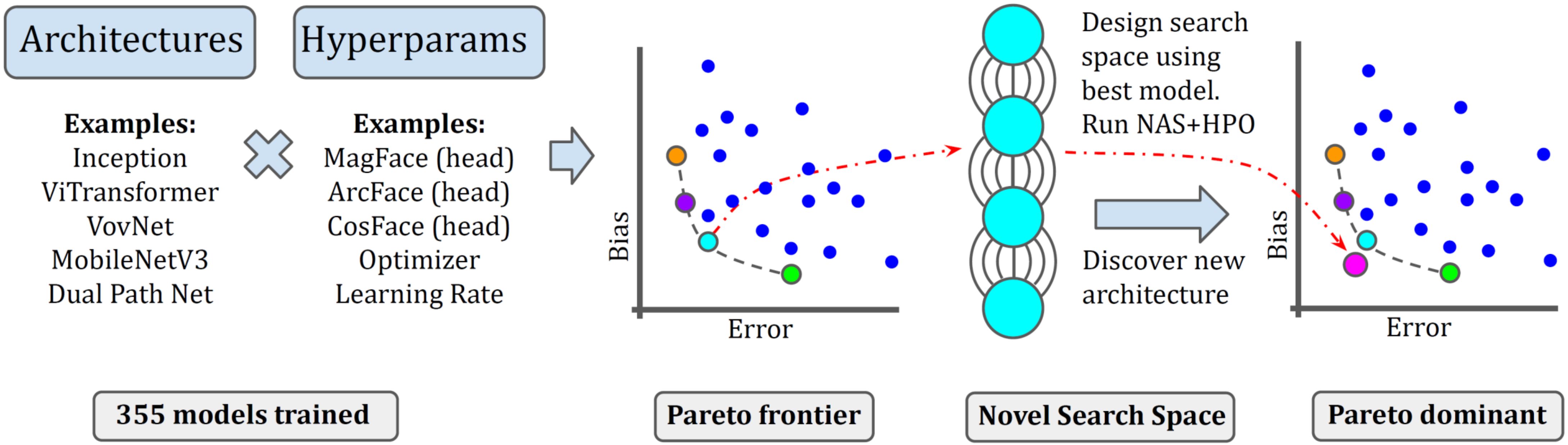
	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
Amazon	98.7%	68.6%	100.0%	92.9%	31.4%
Kairos	98.7%	77.5%	100.0%	93.6%	22.5%
IBM	99.4%	83.0%	99.7%	97.6%	16.7%
Face++	98.7%	95.9%	99.5%	99.0%	3.6%
Microsoft	99.7%	98.5%	100.0%	99.7%	1.5%

- Face recognition is used by law enforcement agencies for sensitive applications
  - Identifying suspects; tracking down missing persons; biometric security

- How can we improve this?
  - Pre-processing, training, and post-processing methods have failed to close the gap
  - Can Deep Learning 2.0 help?

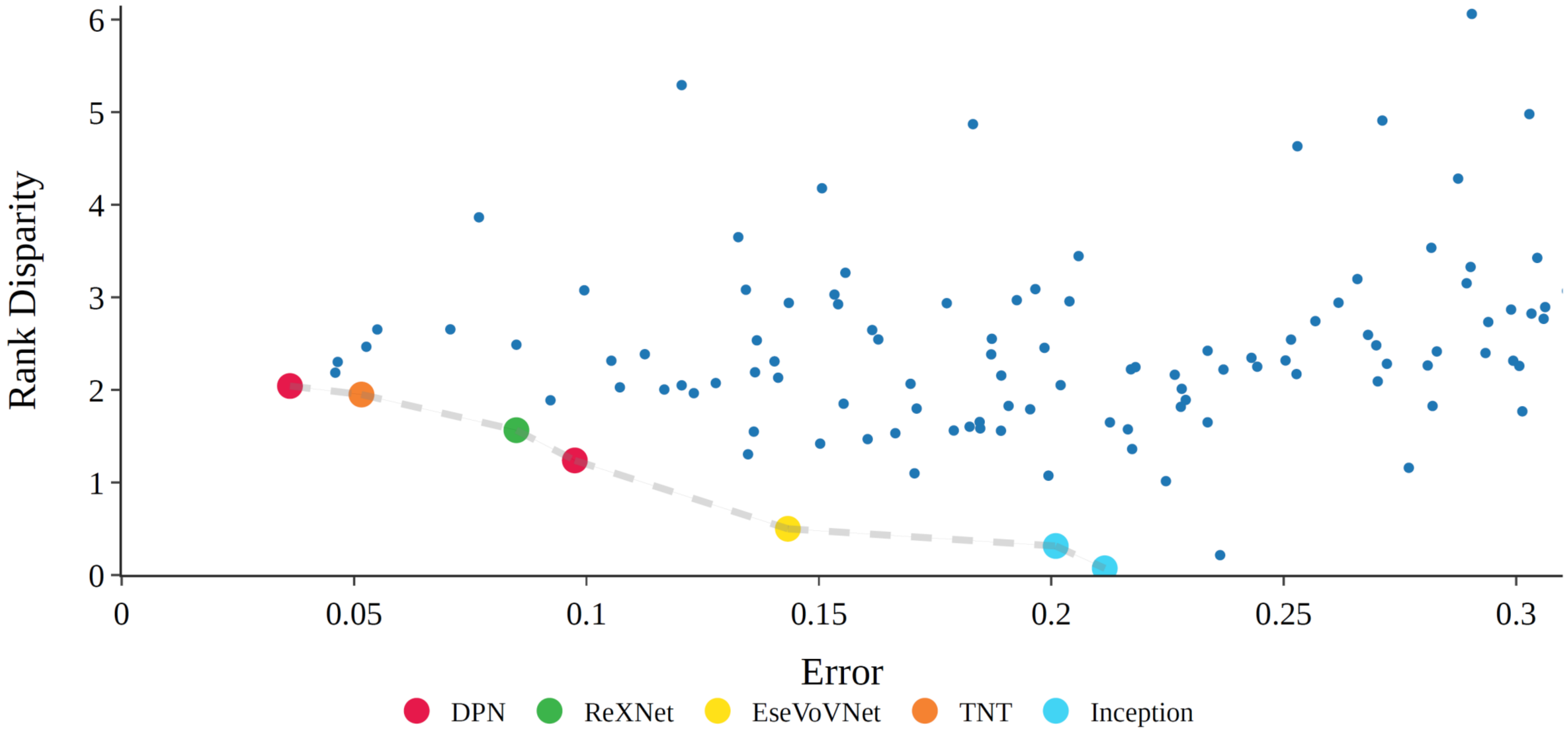
Deep Learning 2.0 to Find Better & Fairer Models

- Dataset: CelebA face recognition
- Protected attribute: Gender
- Fairness metric: difference in quality of classification („Rank disparity”)



- **Result: fairer and more accurate than traditional fairness mitigation algorithms**

Phase 1: Assessment of Fairness & Error of Many Models



Phase 2: Multi-objective Optimization for Fairness & Error

- Architecture space:
  - DPN block of dual path networks
- Hyperparameter space:
  - Optimizer
  - Learning rate
  - Type of head/loss
- Search method used:
  - Bayesian optimization package SMAC3 [Lindauer et al, JMLR 2022]
    - Natively supports multi-fidelity, multi-objective optimization

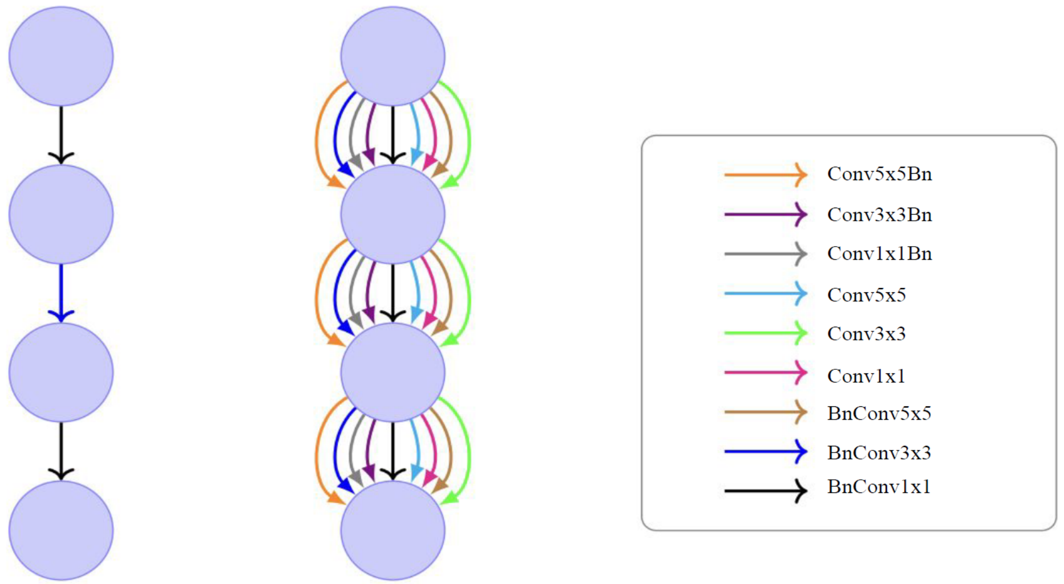
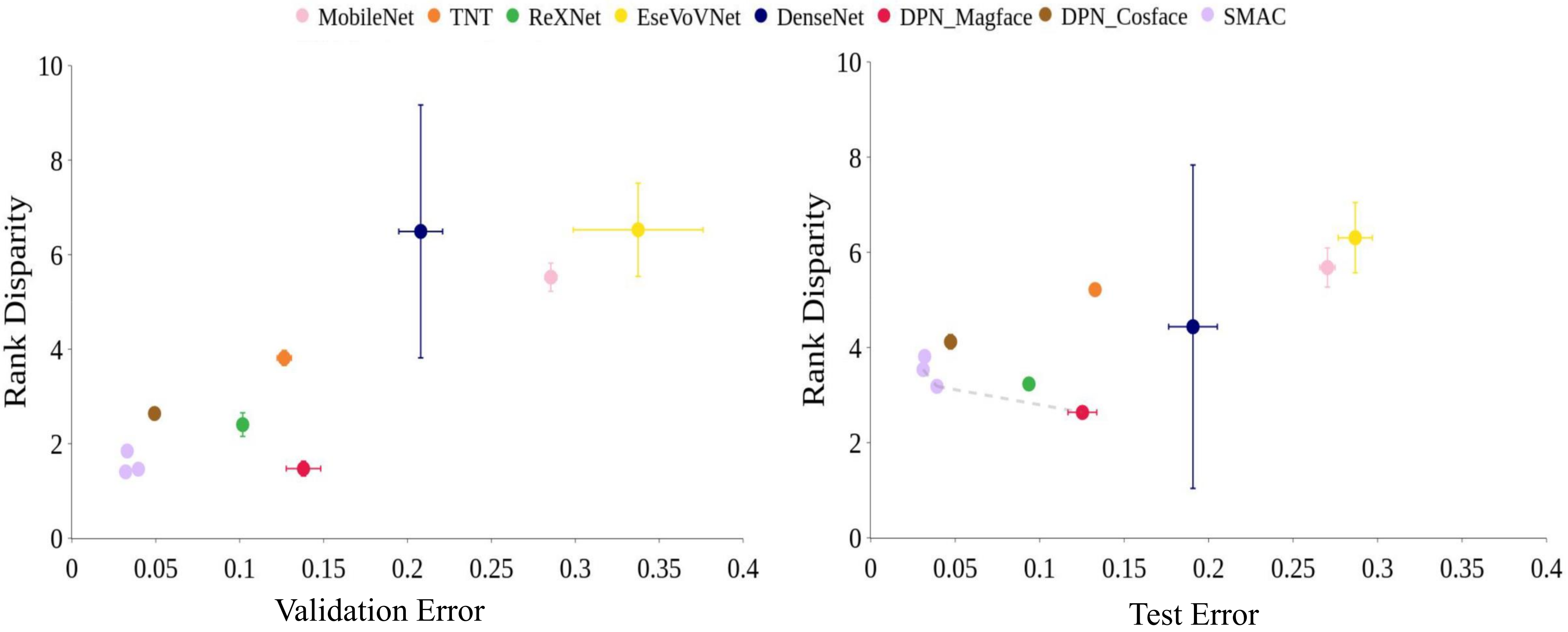


Figure 4: DPN block (left) vs. our searchable block (right).

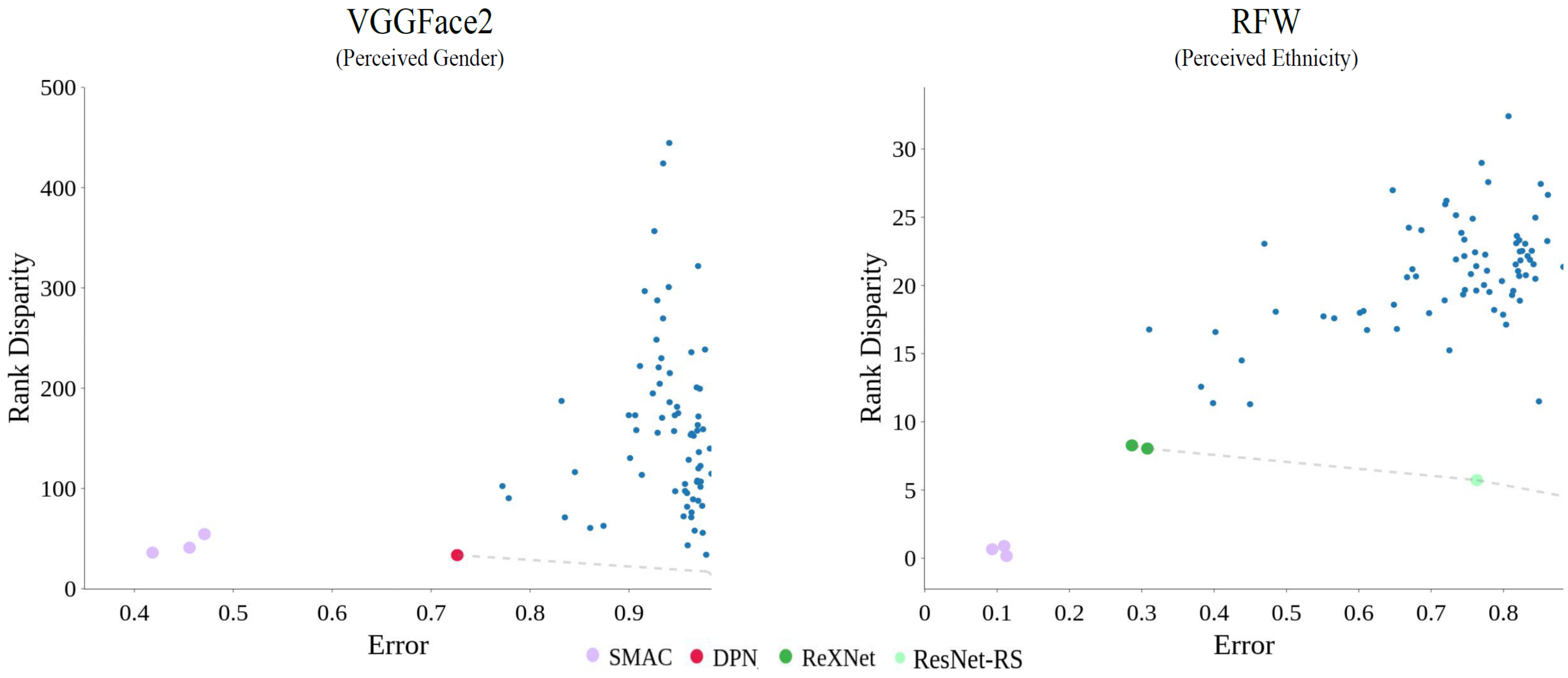
Table 1: Searchable hyperparameter choices.

Hyperparameter	**Choices**
Architecture Head/Loss	MagFace, ArcFace, CosFace
Optimizer Type	Adam, AdamW, SGD
Learning rate (conditional)	Adam/AdamW $\rightarrow [1e-4, 1e-2]$, SGD $\rightarrow [0.09, 0.8]$

Phase 2: Multi-objective Optimization for Fairness & Error



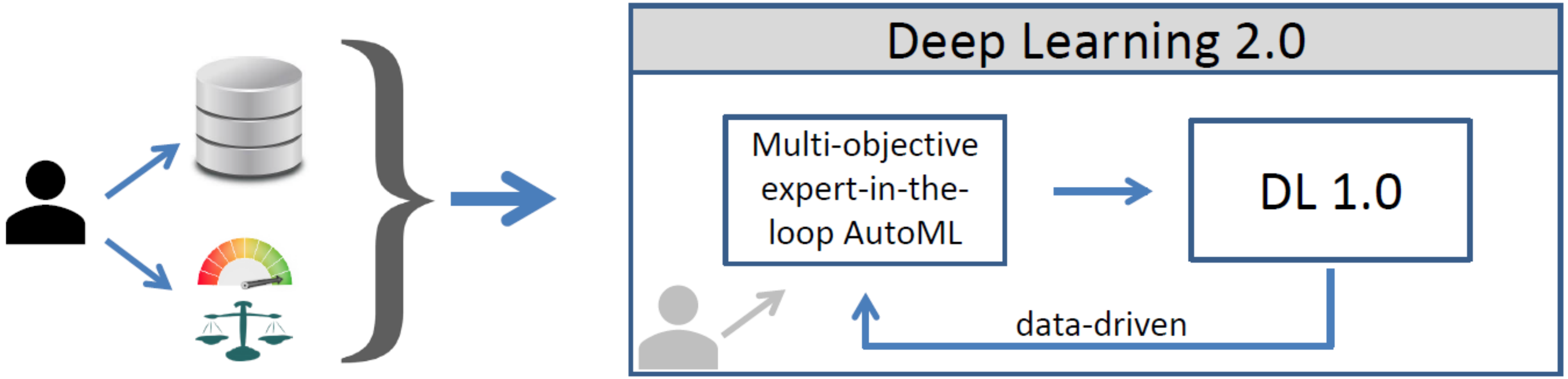
Generalization to Other Fairness Tasks



Our found architecture & hyperparameters appear to still be strong w.r.t. other metrics

Take-aways

Deep Learning 2.0: expert-guided Auto-DL for the objectives at hand

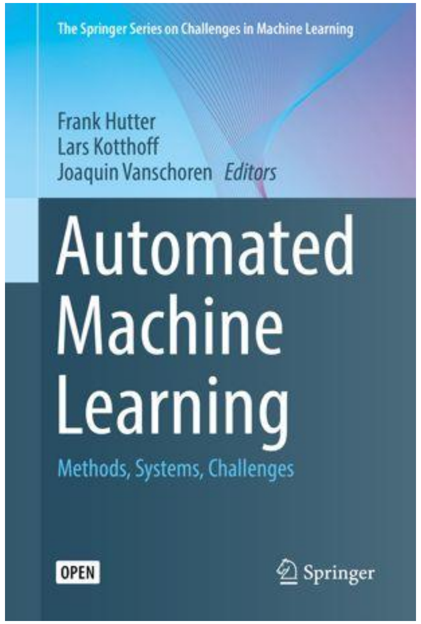


Blackbox NAS to power DL 2.0

- Its speed can rival one-shot methods
- Flexibility
  - Hierarchical search spaces
  - JAHS: including hyperparameters
- Fair face recognition by JAHS

all our code
is open-source:
github.com/automl

get involved:
AutoML conference series
automl.cc



Thank you for your attention!

Funding sources

BOSCH

erc European Research Council

Emmy Noether-Programm
DFG Deutsche Forschungsgemeinschaft

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

DFG
Priority Programme 1527

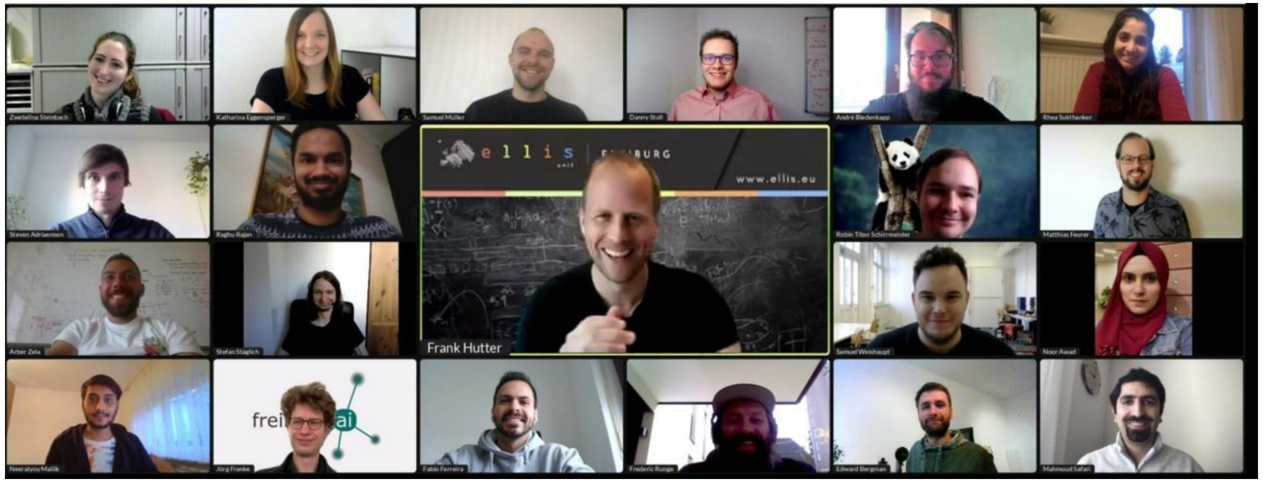
TAILOR
https://tailor-network.eu/

DFG Deutsche Forschungsgemeinschaft

European Commission

BrainLinks BrainTools
acting.thoughts

My fantastic team



@FrankRHutter
@AutoML_org

AutoML.org