Searching for **architectures** and **BERT moments** in specialized AI applications

Misha Khodak

CS 839

3 March 2026

Outline of the lecture

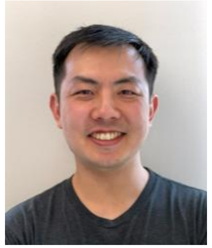
1. specialized foundation models (**FMs**)
   a. the large-scale pretraining paradigm
   b. our investigation comparing the latest specialized FMs to **classical supervised methods**

2. neural architecture search (**NAS**)
   a. using training data to specify a model class
   b. developing a NAS method that works for **data beyond vision and text**

Zongzhe Xu

Ritvik Gupta

Wenduo Cheng



Alex Shen

Junhong Shen

Ameet Talwalkar

Paradigm shifts in the last decade of AI

ImageNet moment: train deep nets on GPUs

2012

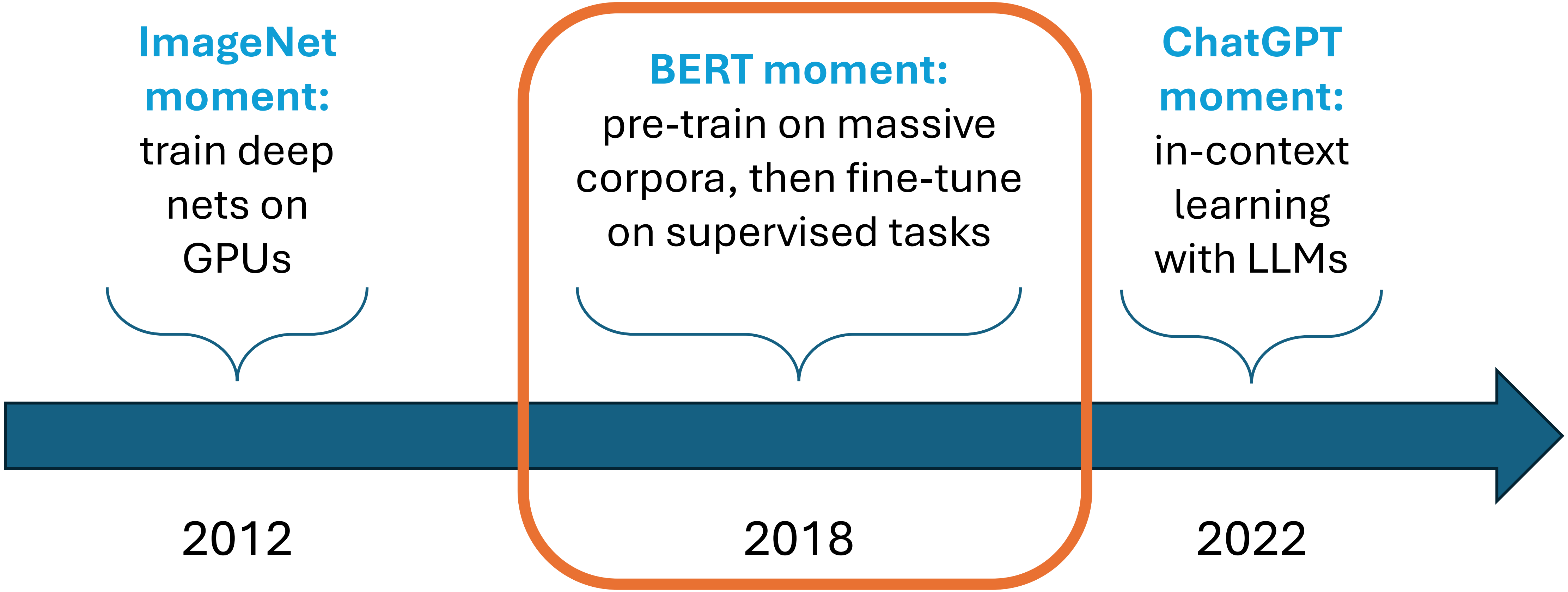
BERT moment: pre-train on massive corpora, then fine-tune on supervised tasks

2018

ChatGPT moment: in-context learning with LLMs

2022

Paradigm shifts in the last decade of AI



Paradigm shifts in the last decade of AI

Before: fit neural networks directly to **supervised** data (labeled pairs (x, y))

BERT moment:
pre-train on massive corpora, then fine-tune on supervised tasks

After: fine-tune **foundation models** pretrained on massive datasets

2018

In detail: What is **supervised learning** ?

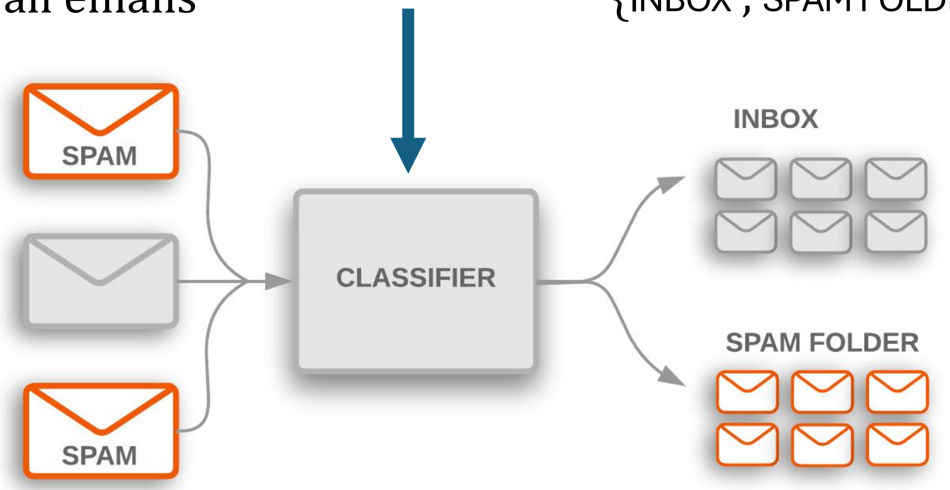
Given a task with some labeled data pairs $(x, y) \in X \times Y$

1. choose parameterized model class of functions $f_\theta : X \mapsto Y$
2. set parameters θ by optimizing performance on your data
3. tune as needed

$X =$ all emails

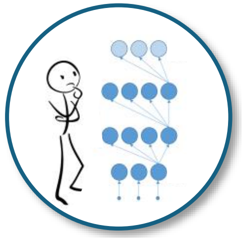
f_θ

$Y = \{\text{INBOX , SPAM FOLDER}\}$

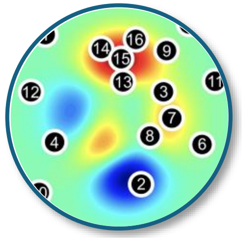


$$\max_\theta \sum_{(x,y)} [f_\theta(x) == y]$$

The traditional supervised workflow



model
development



hyper-
parameter
optimization

train model
on target task

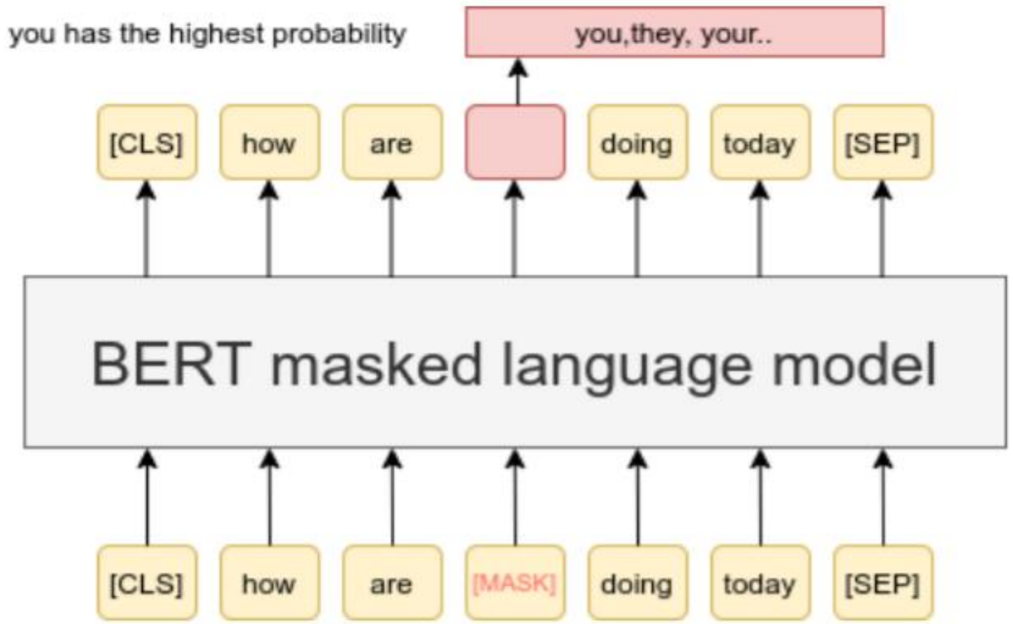
choose a
parameterized
model

tune as needed

optimize
performance on
your data

What happened in 2018?

Google trained a large model (**BERT**) to fill in masked words, training it on all of Wikipedia and a large set of public domain books.



Devlin et al. *BERT: Pre-training of deep bidirectional transformers for language understanding*. NAACL 2019.

What happened in 2018?

Google trained a large model (**BERT**) to fill in masked words, training it on all of Wikipedia and a large set of public domain books.

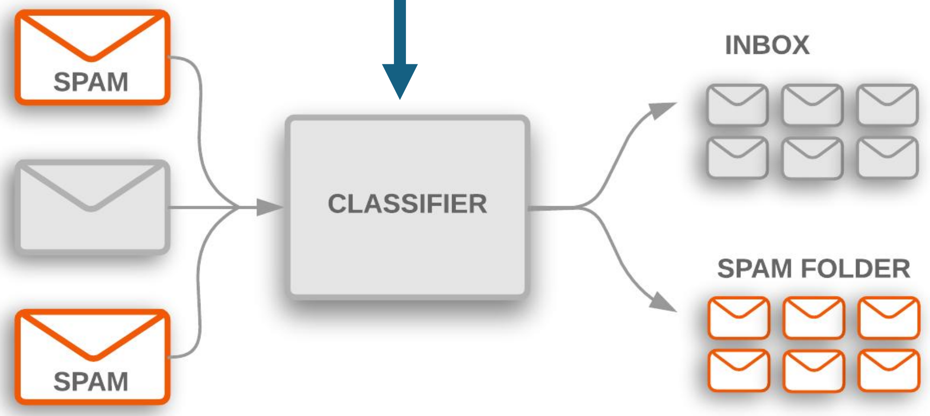
When adapted (**fine-tuned**) on supervised tasks, BERT did **dramatically better** across-the-board.

Such models were later termed **foundation models**

$X =$ all emails

f_θ

$Y =$ {INBOX , SPAM FOLDER}



$$\boxed{\max_\theta} \sum_{(x,y)} [f_\theta(x) == y]$$

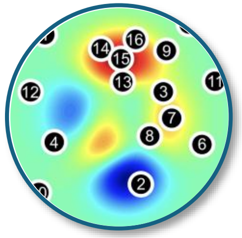
initialization

BERT masked language model

Bommasani et al. *On the opportunities and risks of foundation models*. 2021.

The foundation model workflow

pretrain FM
on massive
dataset



hyper-
parameter
optimization

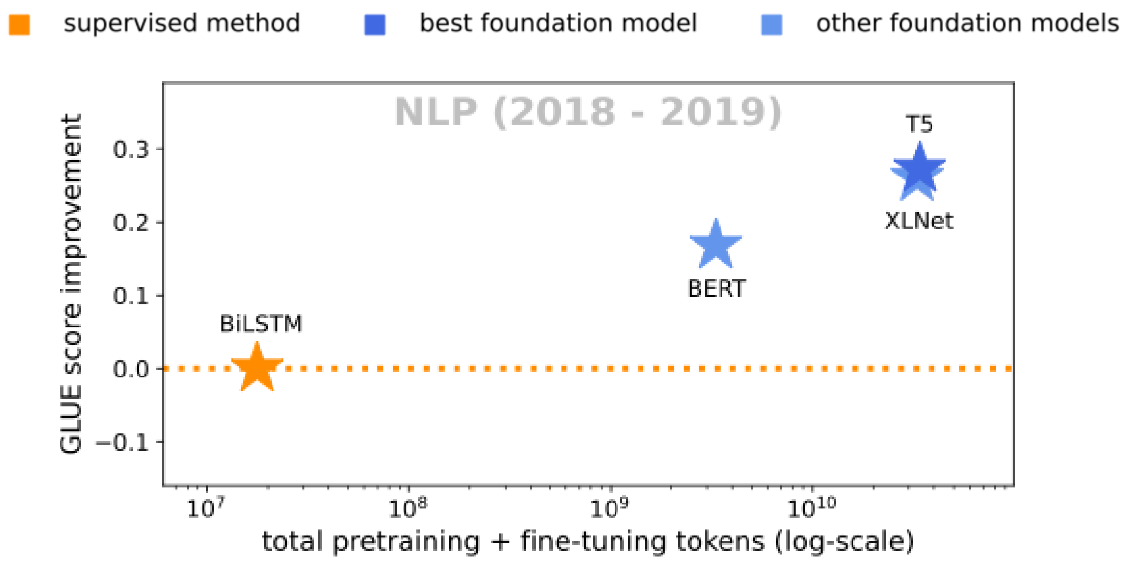
fine-tune FM
on target task

(or download
one since you're
not Google)

tune as needed

optimize
performance on
your data

A paradigm shift in natural language processing (NLP)



Before: well-tuned supervised models are competitive and cheaper

BERT moment: pre-train on massive corpora, then fine-tune on supervised tasks

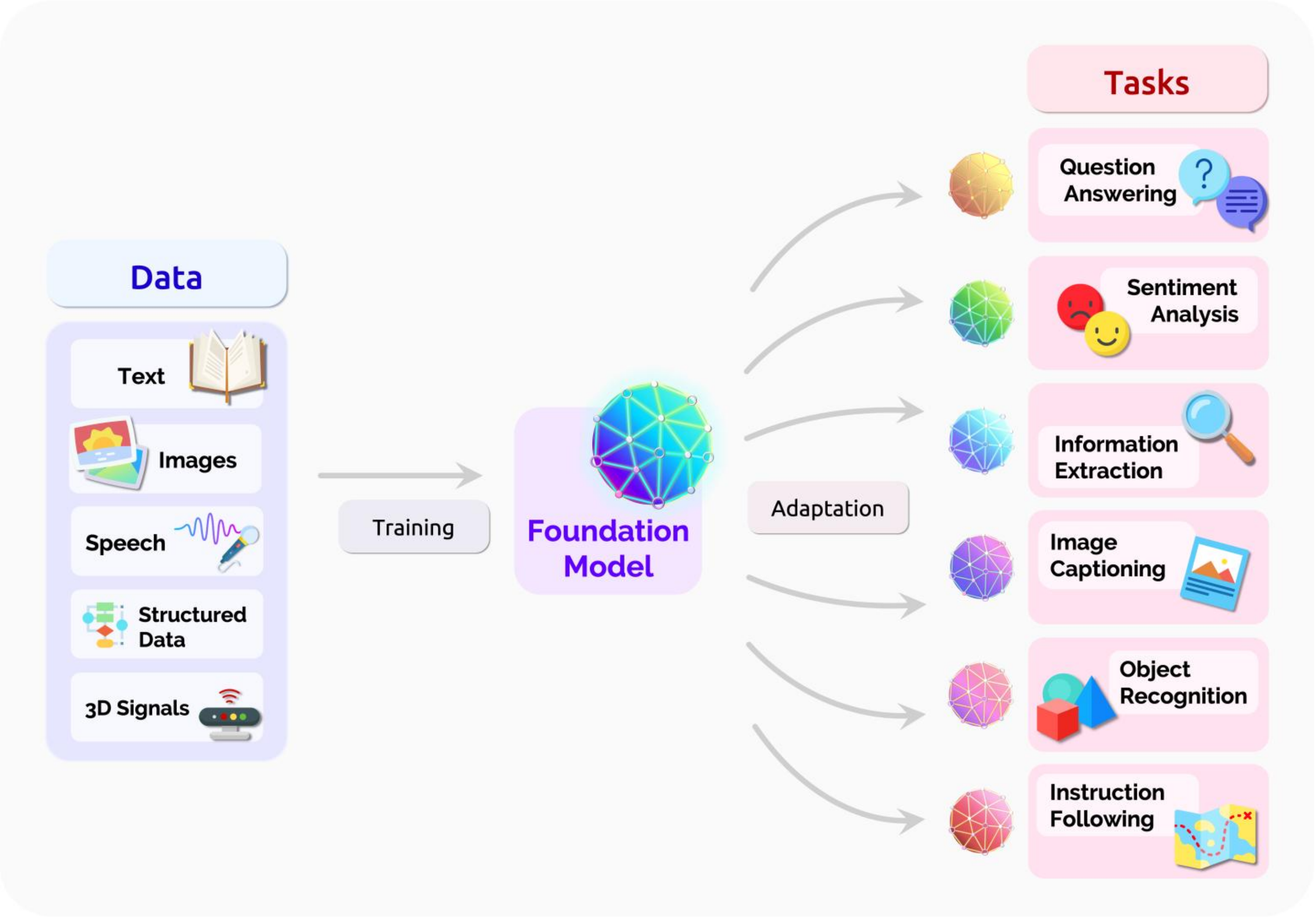
After: nobody trains from scratch anymore, everybody fine-tunes FMs

2018

More recently: An explosion of interest in **specialized FMs**

Why?

1. Many specialized domains have massive unlabeled datasets
2. Obvious success in text / vision / audio
3. AI hype

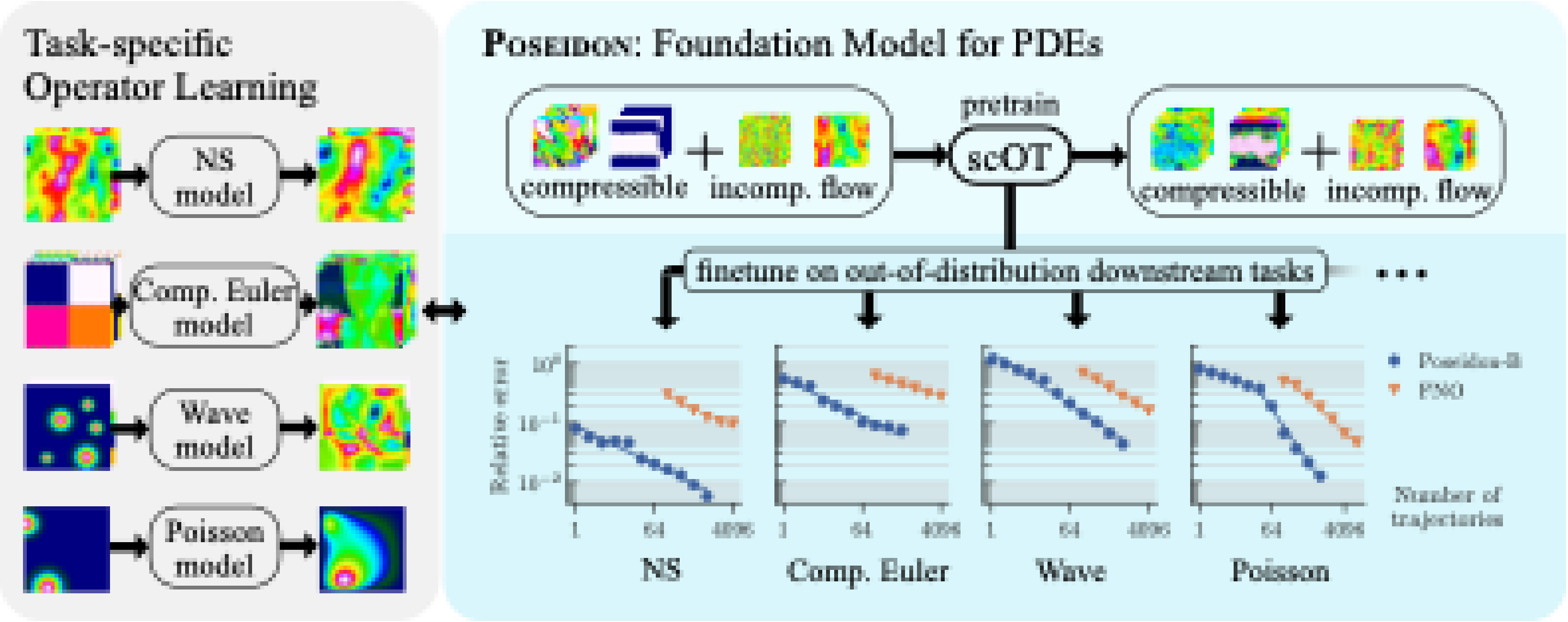


Bommasani et al. *On the opportunities and risks of foundation models*. 2021.

More recently: An explosion of interest in **specialized FMs**

Example: the Poseidon FM for neural PDE solving

1. pretrained on 78K fluid simulations
2. model size: 629M



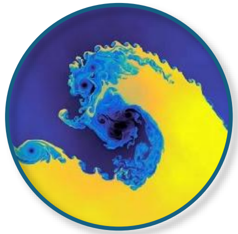
Herde et al. *Poseidon: Efficient foundation models for PDEs*. NeurIPS 2024.

More recently: An explosion of interest in **specialized FMs**

Computer vision

- data: billions of natural images
- success of “FMs” pre-dates BERT

Fluid dynamics



- data: thousands of PDE simulations (small)
- FMs: few well-known FMs (other than Poseidon)

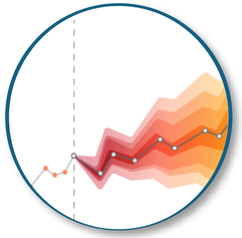
Genomic sequences

data: trillions of human genome base-pairs
FMs: ≥10, including
- Enformer (252M)
- NT-Multispecies (2.5B)

Satellite imaging

data: millions of pictures of Earth
FMs: ≥ 11, including
- DOFA (337M)
- ScaleMAE (323M)

Time series



data: millions of unique series
FMs: ≥9, including
- MOMENT (385M)
- TEMPO (345M)

Have there been **BERT moments** in specialized domains?

[Xu*-Gupta*-Cheng-Shen-Shen-Talwalkar-**K**, ICLR 2025]

Computer vision

- data: billions of natural images
- success of “FMs” pre-dates BERT

Fluid dynamics

- data: thousands of PDE simulations (small)
- FMs: few well-known FMs (other than Poseidon)

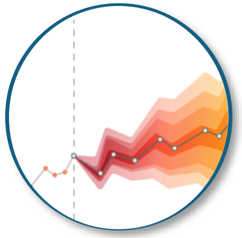
Genomic sequences

data: trillions of human genome base-pairs
FMs: ≥10, including
- Enformer (252M)
- NT-Multispecies (2.5B)

Satellite imaging

data: millions of pictures of Earth
FMs: ≥ 11, including
- DOFA (337M)
- ScaleMAE (323M)

Time series



data: millions of unique series
FMs: ≥9, including
- MOMENT (385M)
- TEMPO (345M)

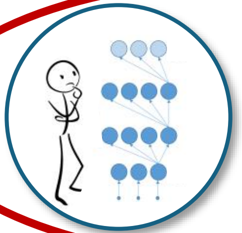
Have there been **BERT moments** in specialized domains?

[Xu*-Gupta*-Cheng-Shen-Shen-Talwalkar-**K**, ICLR 2025]

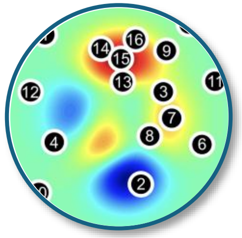
Investigation outline

1. pick **three specialized domains** with many large foundation models (genomics, satellite, time series)
2. identify **benchmark tasks** used to evaluate them
3. question: does fine-tuning those FMs on those tasks **beat traditional supervised learning** given the same compute budget?

Traditional supervised workflow



model development

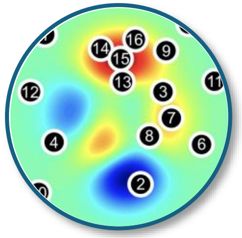


hyper-parameter optimization

train model on target task

Too many tasks to do by hand:

1. 18 genomics tasks
2. 9 satellite tasks
3. 28 time series tasks



hyper-parameter optimization

fine-tune FM on target task

Have there been **BERT moments** in specialized domains?

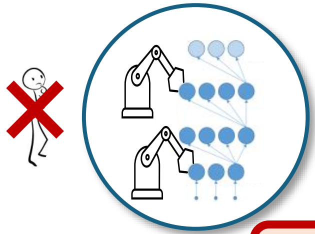
[Xu*-Gupta*-Cheng-Shen-Shen-Talwalkar-**K**, ICLR 2025]

Investigation outline

1. pick **three specialized domains** with many large foundation models (genomics, satellite, time series)
2. identify **benchmark tasks** used to evaluate them
3. question: does fine-tuning those FMs on those tasks **beat traditional supervised learning** given the same compute budget?

~~Traditional~~

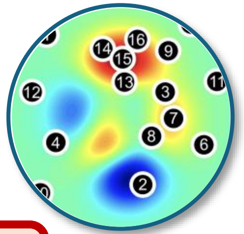
Simulated supervised workflow



~~model development~~

neural architecture search

second part of the talk

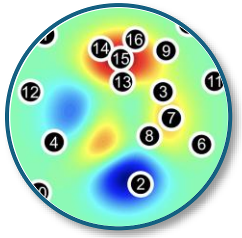


hyper-parameter optimization

train model on target task

Foundation model workflow

pretrain FM on massive dataset

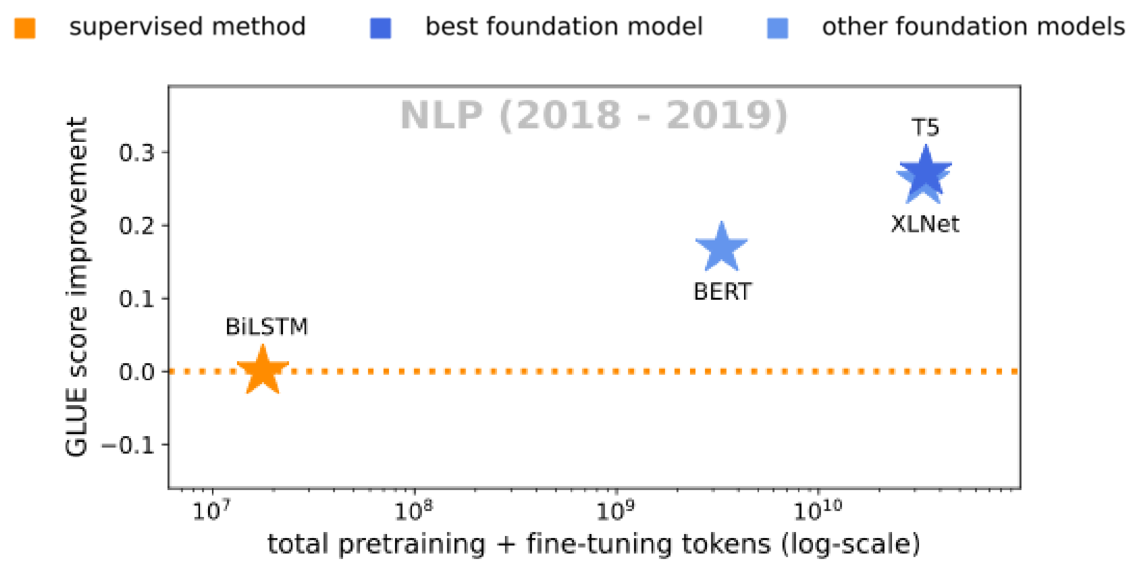


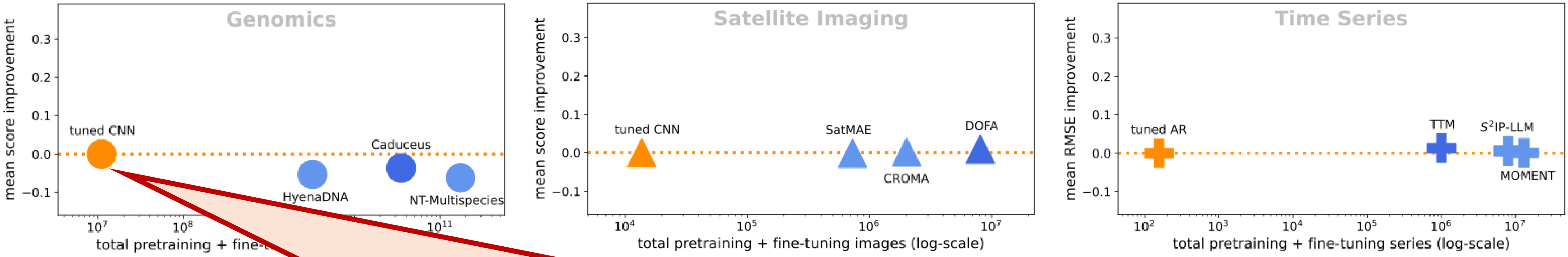
hyper-parameter optimization

fine-tune FM on target task

Have there been **BERT moments** in specialized domains?

[Xu*-Gupta*-Cheng-Shen-Shen-Talwalkar-**K**, ICLR 2025]



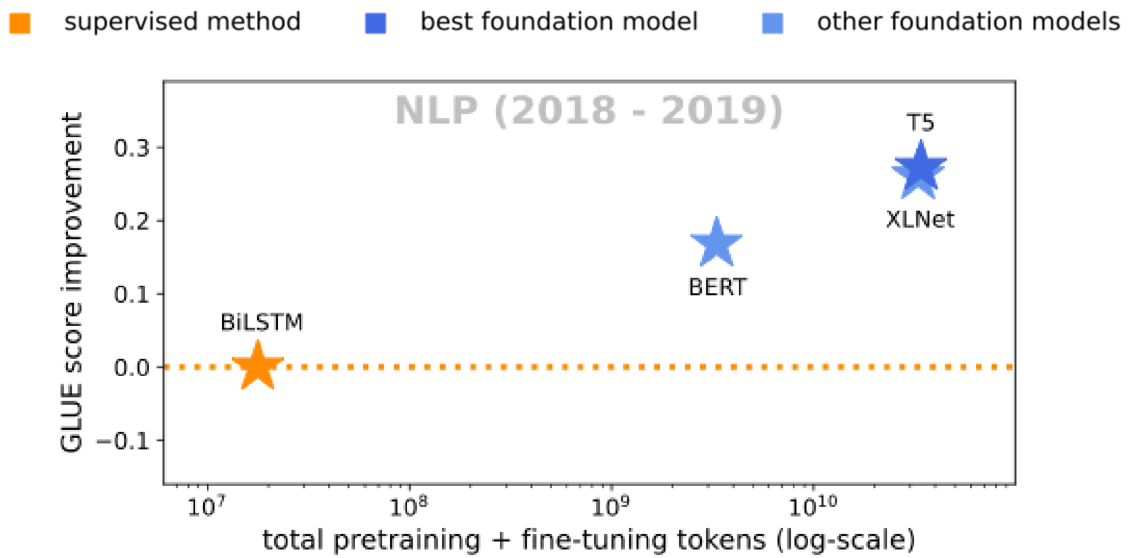


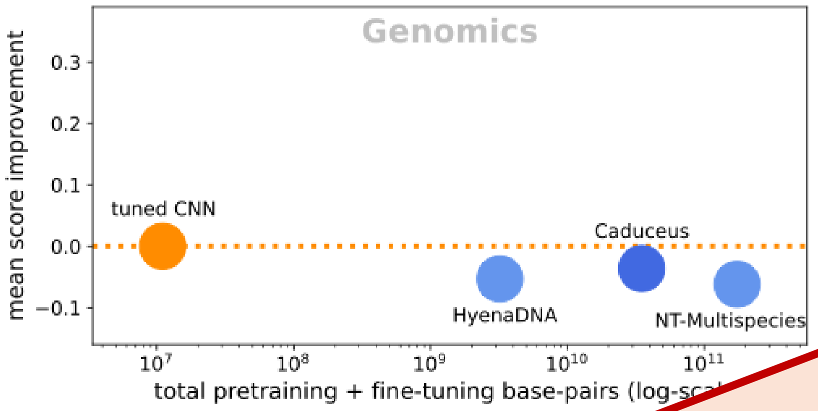
Genomics

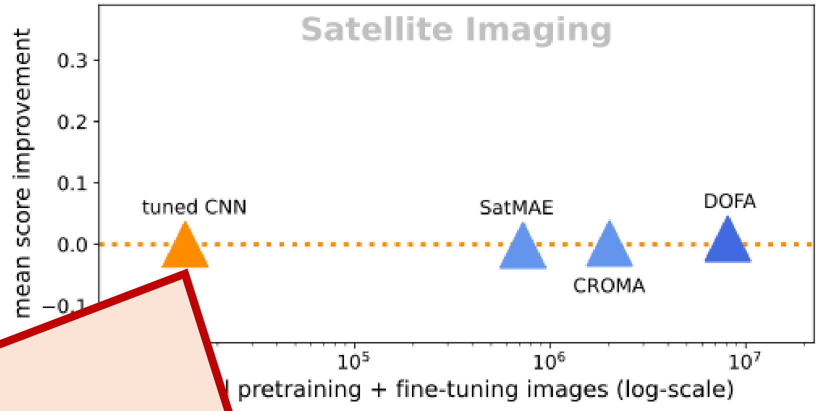
- neural architecture search consistently discovered models that beat **200x**-bigger FMs
- we set the new state-of-the-art on a leading benchmark (Nucleotide Transformer)

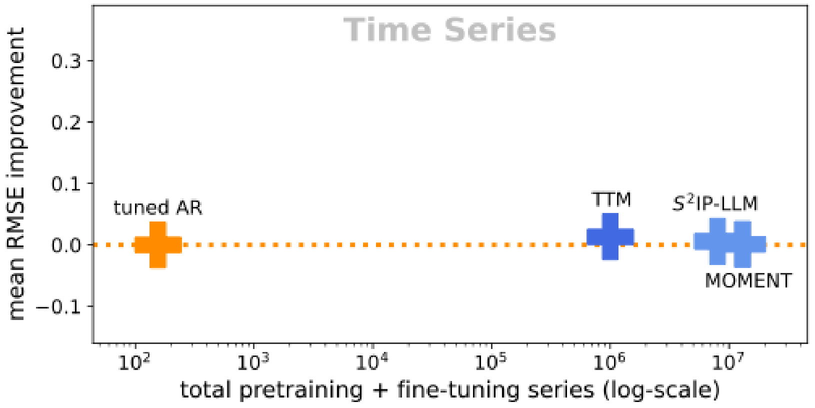
Have there been **BERT moments** in specialized domains?

[Xu*-Gupta*-Cheng-Shen-Shen-Talwalkar-**K**, ICLR 2025]







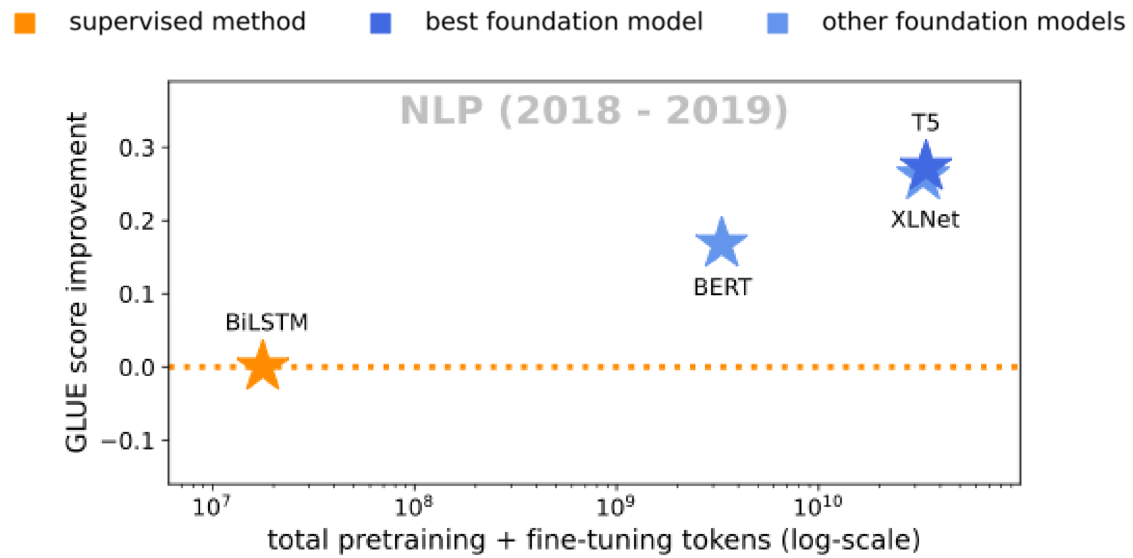


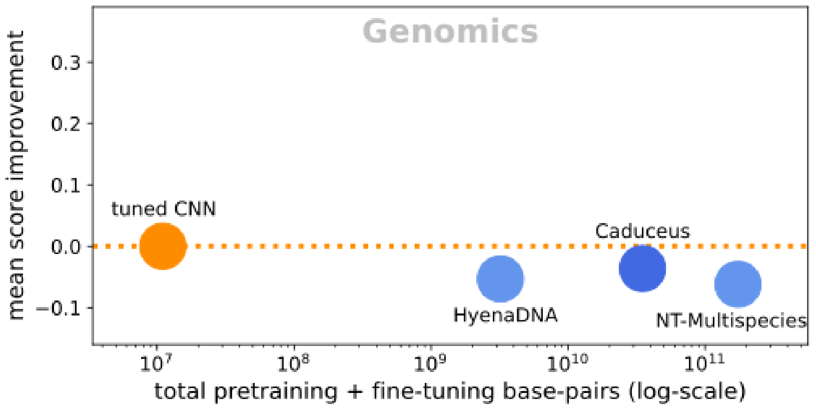
Satellite imaging

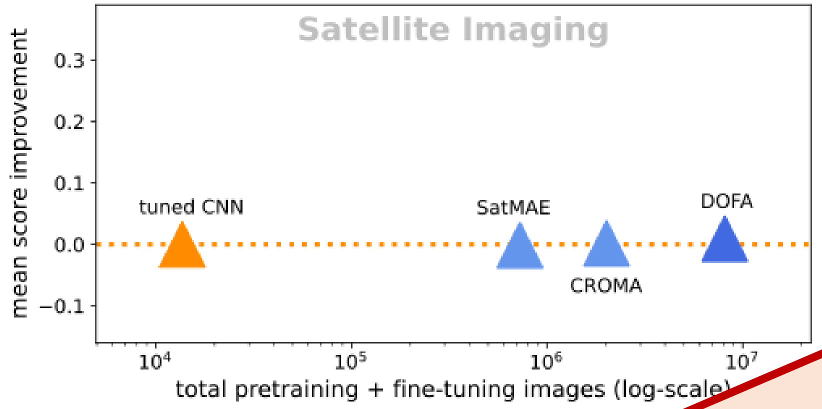
- neural architecture search consistently discovered models matching **10x**-bigger FMs
- we achieved performance competitive with the latest FMs on GeoBench

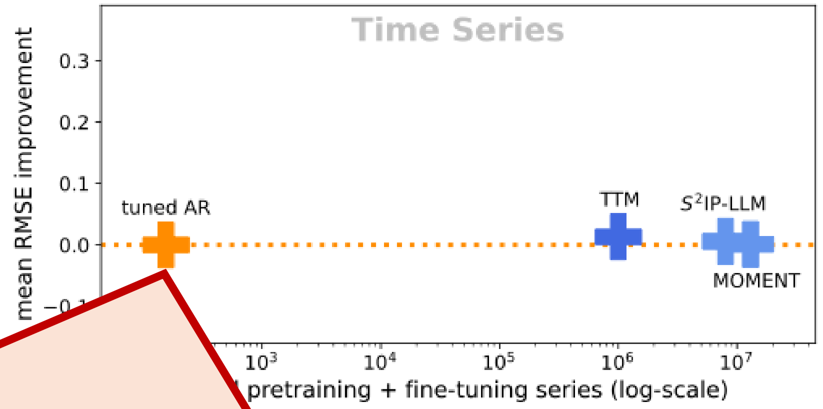
Have there been **BERT moments** in specialized domains?

[Xu*-Gupta*-Cheng-Shen-Shen-Talwalkar-**K**, ICLR 2025]









Time series

- tuning a linear (AR) model is competitive with **750,000x**-bigger (!) time series FMs
- past work only tested linear models with **5 steps** of history while giving FMs **512 steps**

Aside: **what happened with time series?**

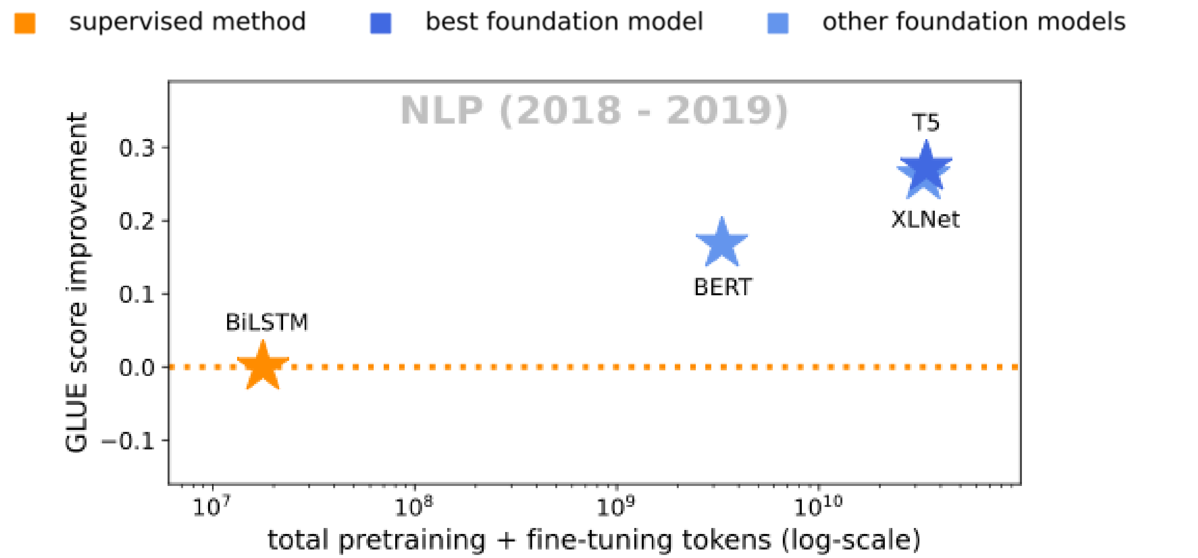
A tragedy in three acts:

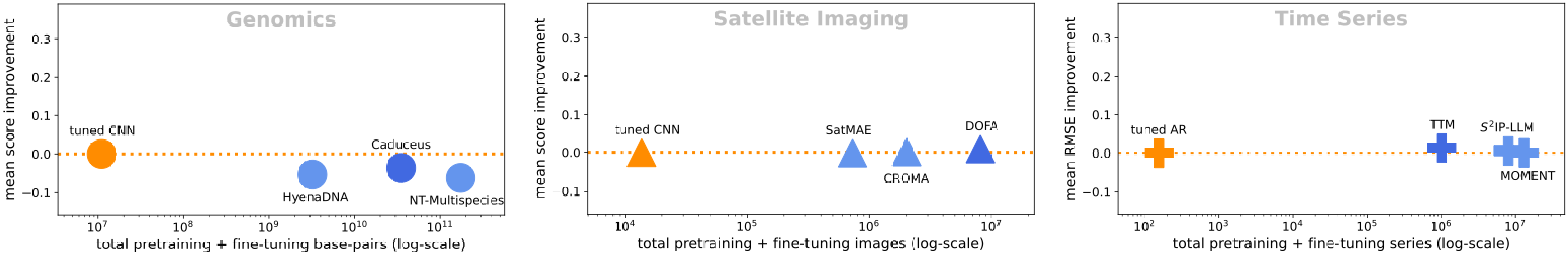
1. Early supervised work compared against Auto-ARIMA, which
   [Hyndman & Khandakar, JSS 2008]
   - strictly generalizes basic linear autoregression ($\mathrm{AR}(p)$) by adding a moving average ($\mathrm{MA}(q)$) term
   - has context length parameters (p and q) that determine model expressivity / how much of the past is used to predict
   - is **too expensive to fit on a context length > 5**
2. Time series FMs took Auto-ARIMA numbers directly from those papers **while giving their FMs 512 steps of context**
3. We found **$\mathrm{AR}(p)$ is very cheap to fit with $p = 512$** and performs as well as those time series FMs (github.com/Zongzhe-Xu/AutoAR)

Challu et al. *N-HiTS: Neural hierarchical interpolation for time series forecasting.* AAAI 2022.

Have there been **BERT moments** in specialized domains?

[Xu*-Gupta*-Cheng-Shen-Shen-Talwalkar-**K**, ICLR 2025]



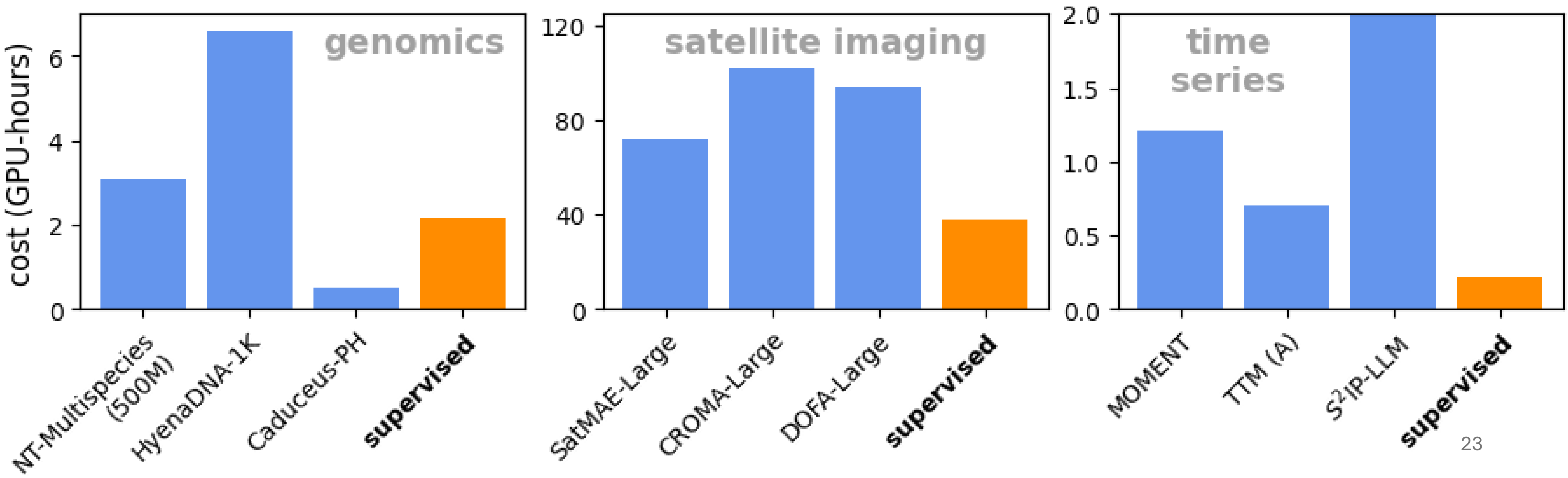


Answer:

not in the three domains that have seen some of the most foundation modeling work outside text / images / audio

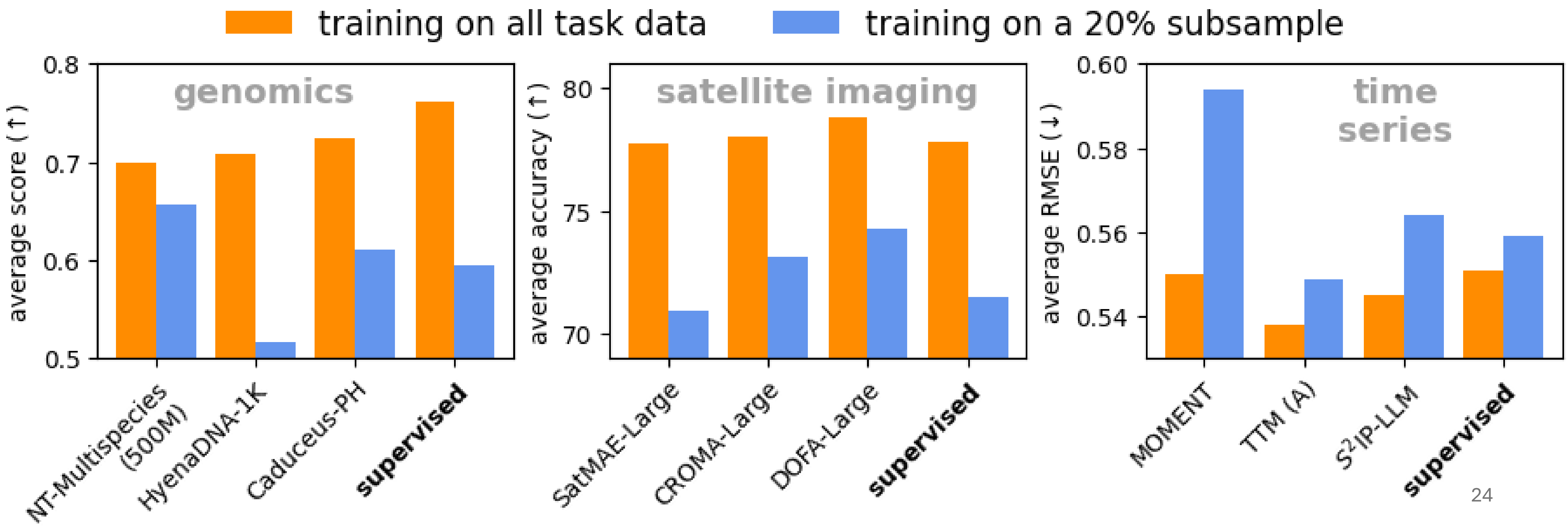
Maybe FMs are more efficient after pre-training?

- supervised models are usually smaller and thus cheaper than FMs
- supervised training is usually faster, **even accounting for the cost of neural architecture search**



Maybe FMs require less data to do well?

- need to design **better benchmarks** that assess this
- supervised learning isn't dramatically worse **even with 5x less data**



Summary:
Evaluating specialized FMs

lots of resources have been spent to train specialized FMs, and yet they are **beaten on their own benchmarks** by **cheaper supervised models**

to accurately measure progress, FM benchmarks should

1. be **difficult**, i.e. not solved by supervised learning
2. reflect **use-cases** targeted by FMs
3. cover **diverse** applications

before using an FM, check if a **well-tuned supervised method** suffices

Outline of the lecture

1. specialized foundation models (**FMs**)
    a. the large-scale pretraining paradigm
    b. our investigation comparing the latest specialized FMs to **classical supervised methods**

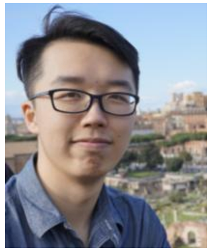
2. neural architecture search (**NAS**)
    a. using training data to specify a model class
    b. developing a NAS method that works for **data beyond vision and text**


Nick Roberts


Junhong Shen


Tri Dao


Liam Li


Renbo Tu


Chris Ré


Fred Sala

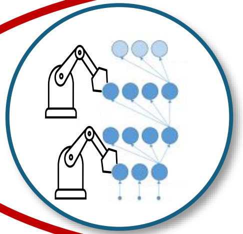

Ameet Talwalkar

Simulating the development of supervised models using **neural architecture search**

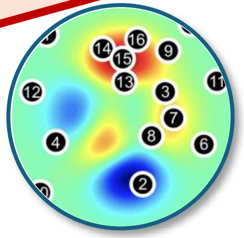
We need a search method that

1. efficiently returns performant supervised models
2. works on multiple data modalities and dimensionalities

Simulated supervised workflow



neural architecture search

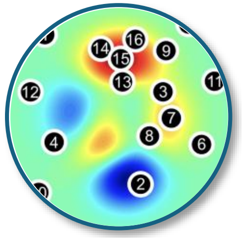


hyper-parameter optimization

train model on target task

Foundation model workflow

pretrain FM on massive dataset



hyper-parameter optimization

fine-tune FM on target task

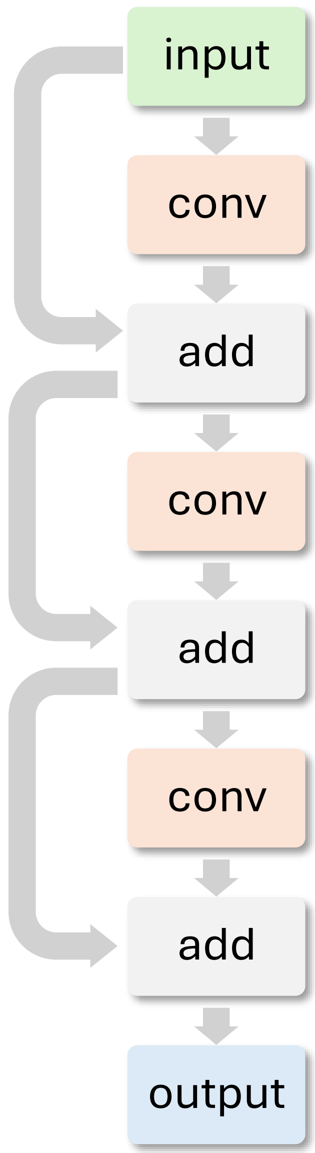
What is neural architecture search (NAS) ?

What is a **neural architecture** ?

- labeled directed graph that defines the model class
- specifies what **operations** (additions, concatenations, multiplies, convolutions, ...) to apply to the input to output a prediction, and in what order
- does **not** specify the parameters of those operations (those are set by training on data)

NAS also uses training data to define the architecture
(by setting the graph's edges and operation labels)

ResNet architecture
[He et al., CVPR 2015]



Why use neural architecture search (NAS) ?

- state-of-the-art architectures can be **very complex**

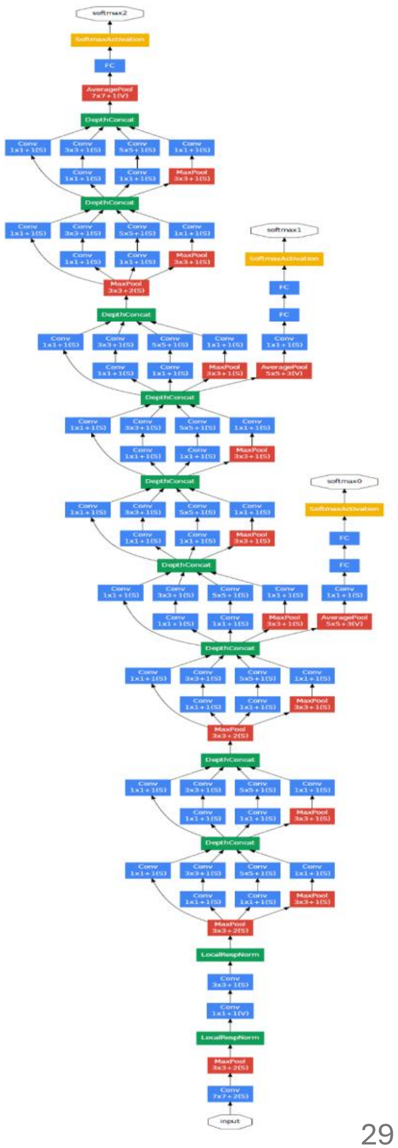
- **too many tasks to hand-select** a customized architecture each

Q: Isn't this just hyperparameter optimization?

A: Yes, but

- many **more** hyperparameters
- the hyperparameters are **architectural**

Inception v3 architecture
[Szegedy et al., CVPR 2016]



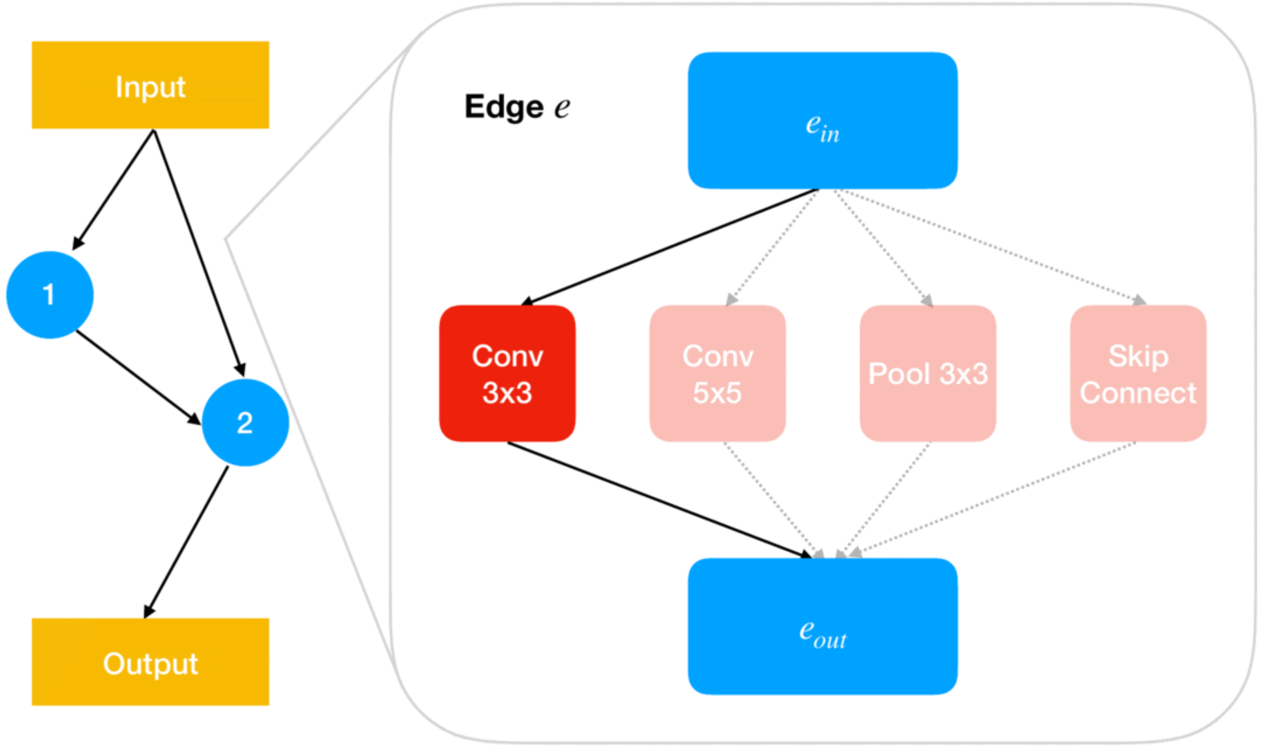
Isn’t NAS super inefficient?

Architectural hyperparameters enable **good heuristics** such as **weight-sharing**

- instead of training many discrete architectures, train one architecture that is a weighted combination of them, then round the weights

Main takeaway:

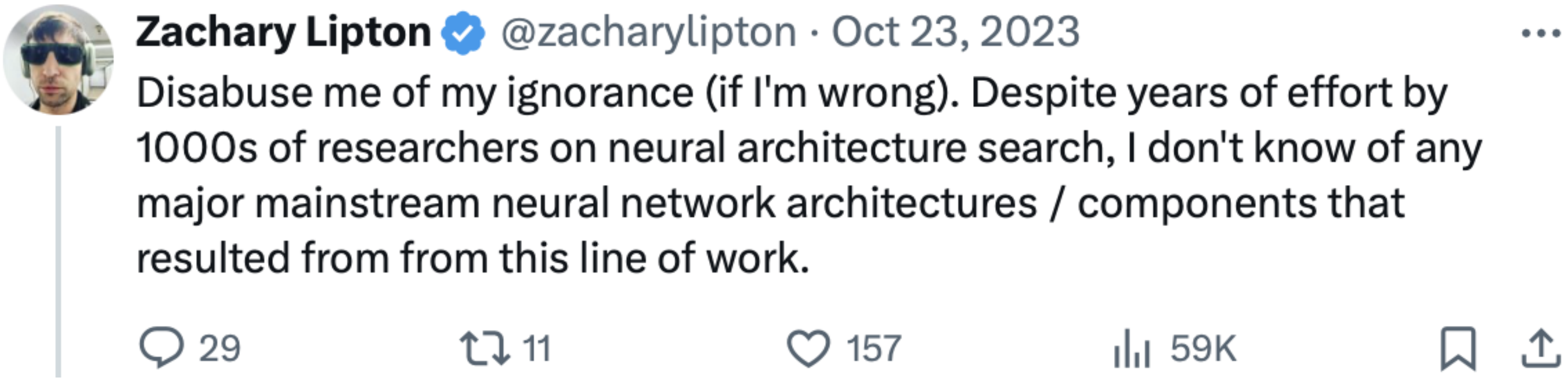
- with the right heuristic, **NAS has less overhead than training a single network** from scratch



Pham et al. *Efficient neural architecture search via parameter sharing*. ICML 2018.

But what is NAS actually useful for?

Nothing?


Zachary Lipton @zacharylipton · Oct 23, 2023

Disabuse me of my ignorance (if I'm wrong). Despite years of effort by 1000s of researchers on neural architecture search, I don't know of any major mainstream neural network architectures / components that resulted from from this line of work.

29 11 157 59K

But what is NAS actually useful for?

~~Nothing?~~

Getting SOTA on ImageNet?

Zachary Lipton @zacharylipton · Oct 23, 2023

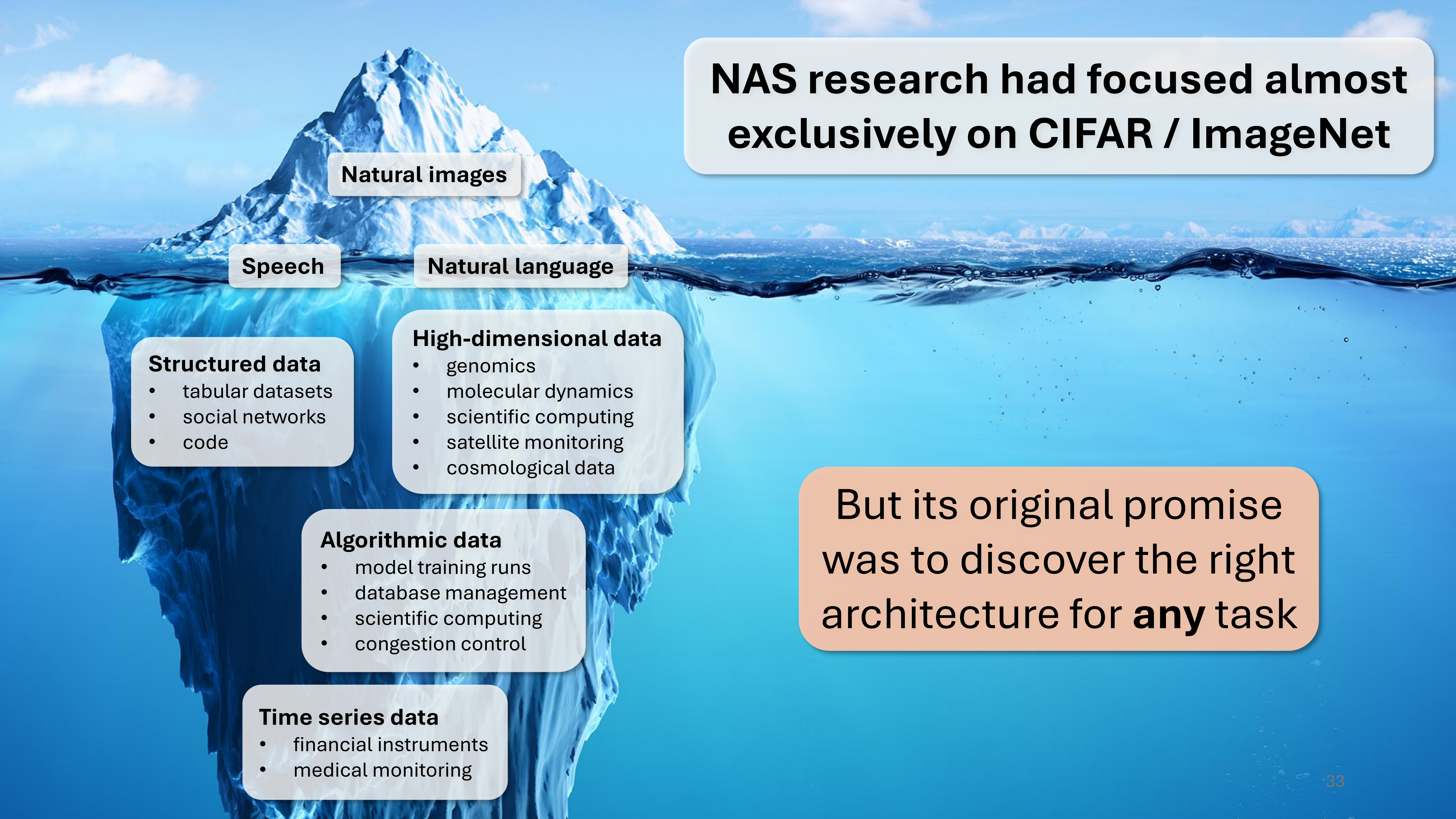
Disabuse me of my ignorance (if I'm wrong). Despite years of effort by 1000s of researchers on neural architecture search, I don't know of any major mainstream neural network architectures / components that resulted from from this line of work.

29 11 157 59K

Colin White @crwhite_ml

Since 2017, SotA on ImageNet has been a NAS architecture 8 times. Most notably from EfficientNet and follow-ups.
NAS hasn't found "brand new" architectures, but it's had great success in making existing architectures/components better and more efficient

NAS research had focused almost exclusively on CIFAR / ImageNet



But its original promise was to discover the right architecture for **any** task

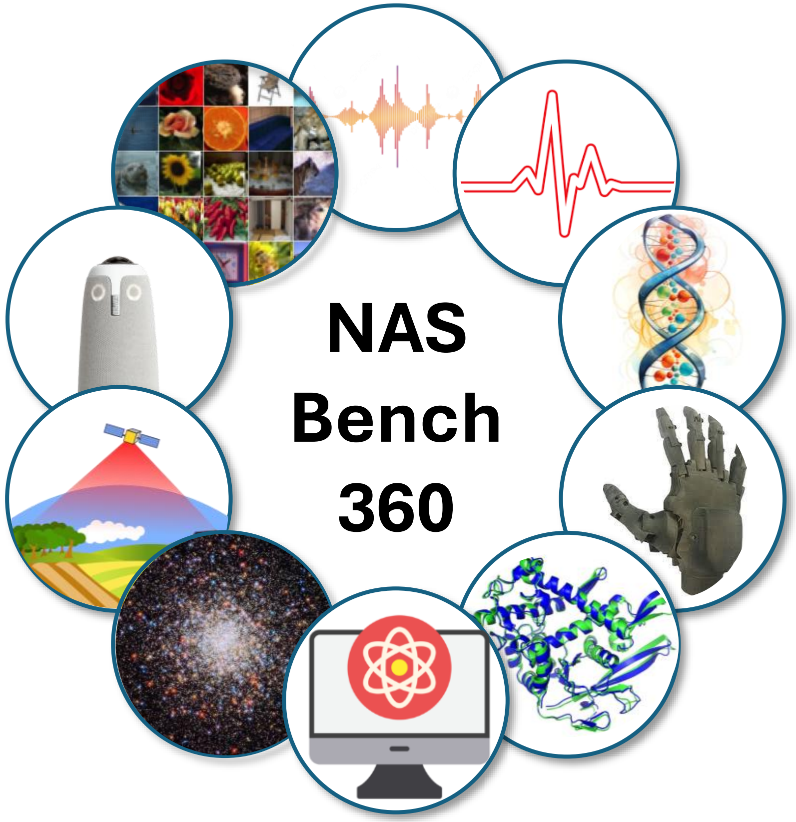
Does NAS work on tasks **outside of computer vision** ?

[Tu*-Roberts*-**K**-Shen-Sala-Talwalkar, NeurIPS 2022 Datasets & Benchmarks]

we built a ten-task benchmark called **NAS-Bench-360** to find out

unfortunately, state-of-the-art NAS methods only beat expert-designed architectures on **3 / 10** tasks ☹

- these are arguably the most interesting use-cases for architecture search
- they are also types of tasks we care about when **evaluating specialized FMs**

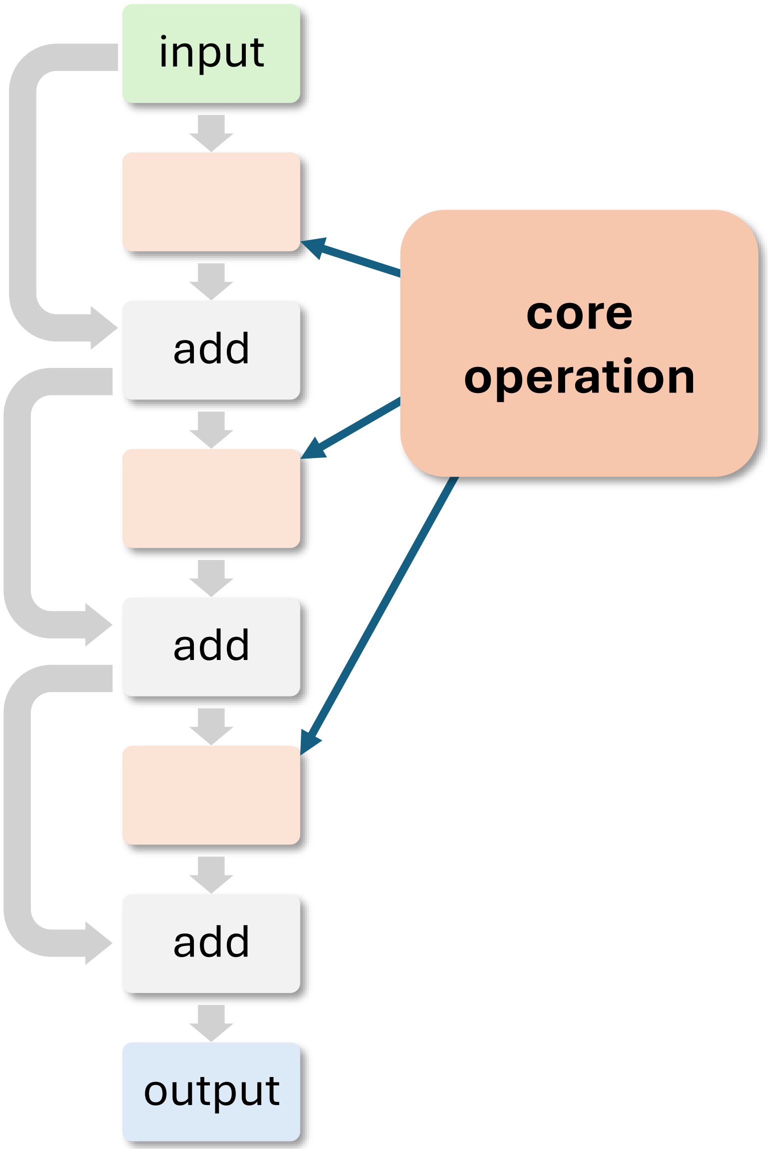


Transferring **neural architectures** to diverse data modalities

[Roberts*-**K***-Dao-Li-Ré-Talwalkar, NeurIPS 2021; Shen*-**K***-Talwalkar, NeurIPS 2022]

Why did the best NAS methods fail?

- NAS has classically aimed for a good neural network **graph topology**
- but most breakthrough architectures have similar topologies built around **one core operation**



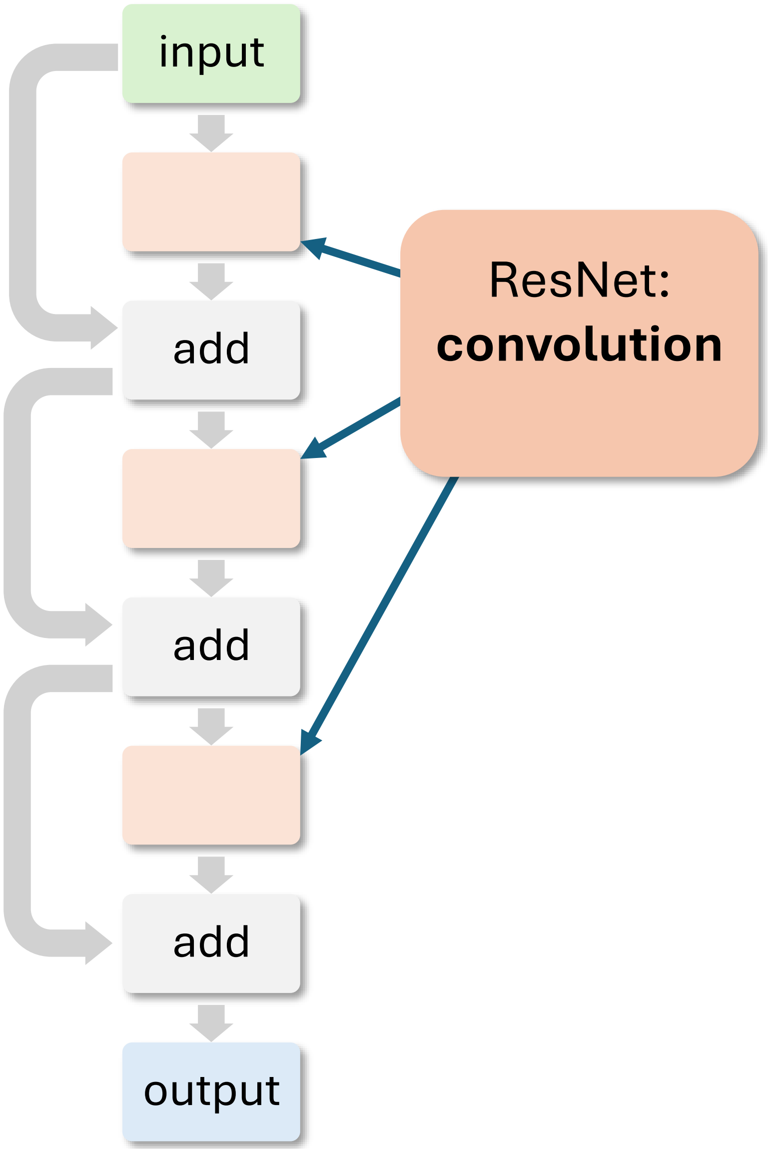
Zoph & Le. *Neural architecture search with reinforcement learning*. ICLR 2017.

Transferring **neural architectures** to diverse data modalities

[Roberts*-**K***-Dao-Li-Ré-Talwalkar, NeurIPS 2021; Shen*-**K***-Talwalkar, NeurIPS 2022]

Why did the best NAS methods fail?

- NAS has classically aimed for a good neural network **graph topology**
- but most breakthrough architectures have similar topologies built around **one core operation**



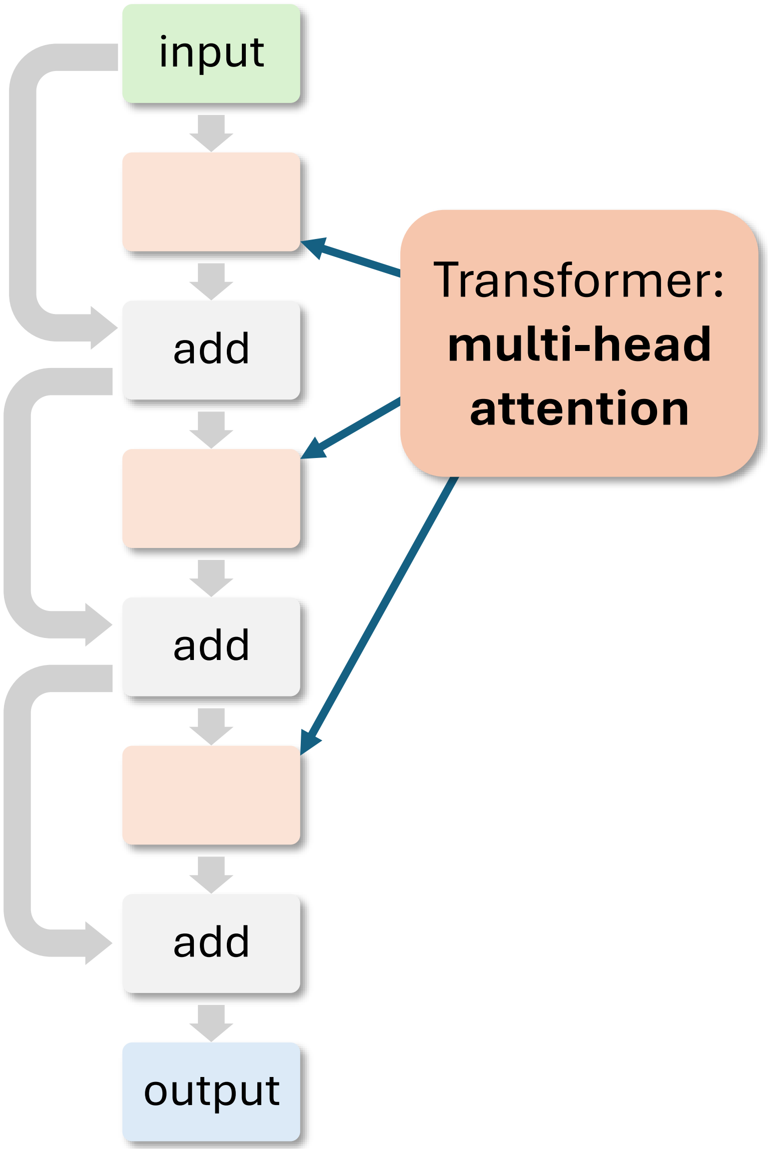
He et al. *Deep residual learning for image recognition*. CVPR 2016.

Transferring **neural architectures** to diverse data modalities

[Roberts*-**K***-Dao-Li-Ré-Talwalkar, NeurIPS 2021; Shen*-**K***-Talwalkar, NeurIPS 2022]

Why did the best NAS methods fail?

- NAS has classically aimed for a good neural network **graph topology**
- but most breakthrough architectures have similar topologies built around **one core operation**



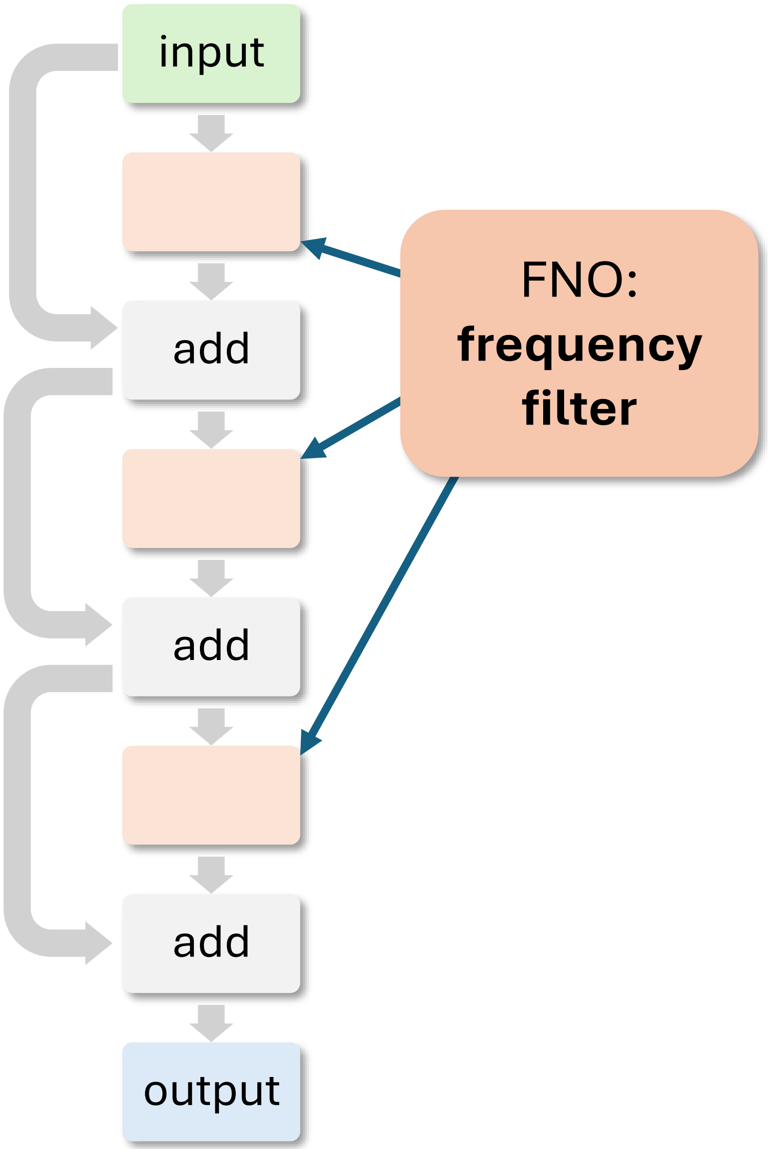
Vaswani et al. *Attention is all you need.* NeurIPS 2017.

Transferring **neural architectures** to diverse data modalities

[Roberts*-**K***-Dao-Li-Ré-Talwalkar, NeurIPS 2021; Shen*-**K***-Talwalkar, NeurIPS 2022]

Why did the best NAS methods fail?

- NAS has classically aimed for a good neural network **graph topology**
- but most breakthrough architectures have similar topologies built around **one core operation**



Li et al. *Fourier neural operator for parametric partial differential equations*. ICLR 2021.

Transferring **neural architectures** to diverse data modalities

[Roberts*-**K***-Dao-Li-Ré-Talwalkar, NeurIPS 2021; Shen*-**K***-Talwalkar, NeurIPS 2022]

Why did the best NAS methods fail?

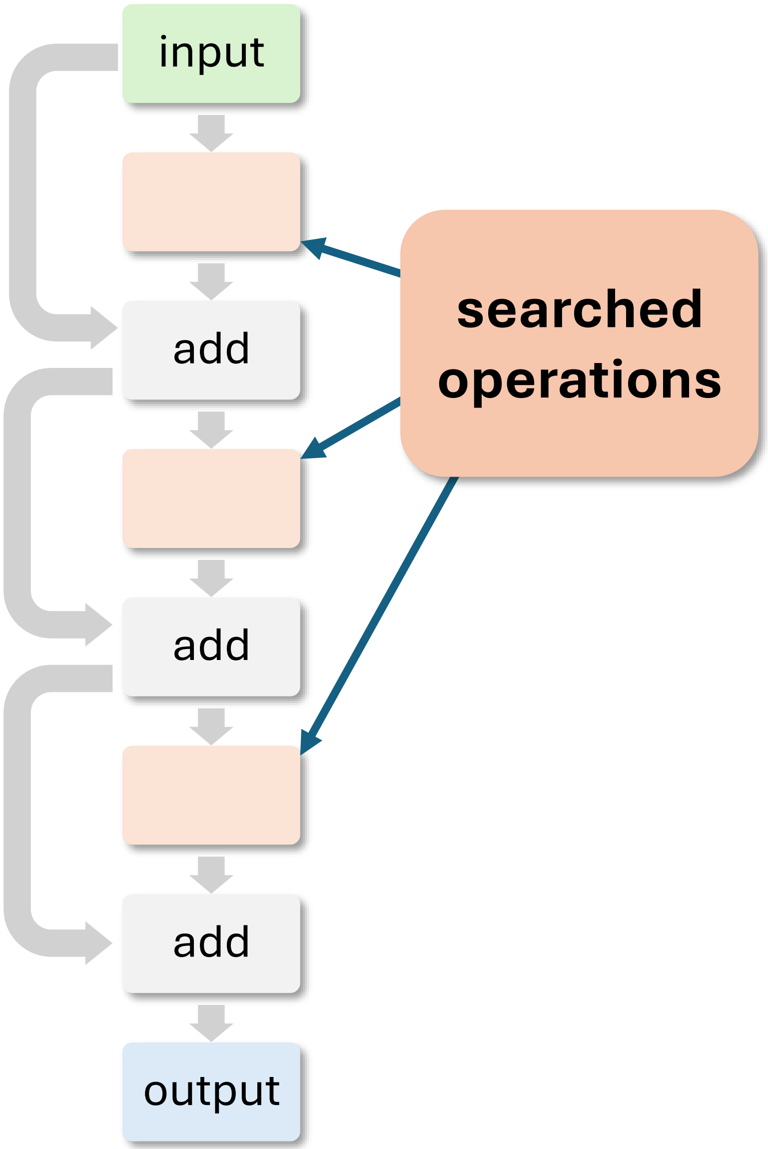
- NAS has classically aimed for a good neural network **graph topology**
- but most breakthrough architectures have similar topologies built around **one core operation**

Idea: fix a topology and search for good **operations** over an expressive search space

[Roberts*-**K***-Dao-Li-Ré-Talwalkar, NeurIPS 2021]

too expensive ☹

XD



Transferring **neural architectures** to diverse data modalities

[Roberts*-**K***-Dao-Li-Ré-Talwalkar, NeurIPS 2021; Shen*-**K***-Talwalkar, NeurIPS 2022]

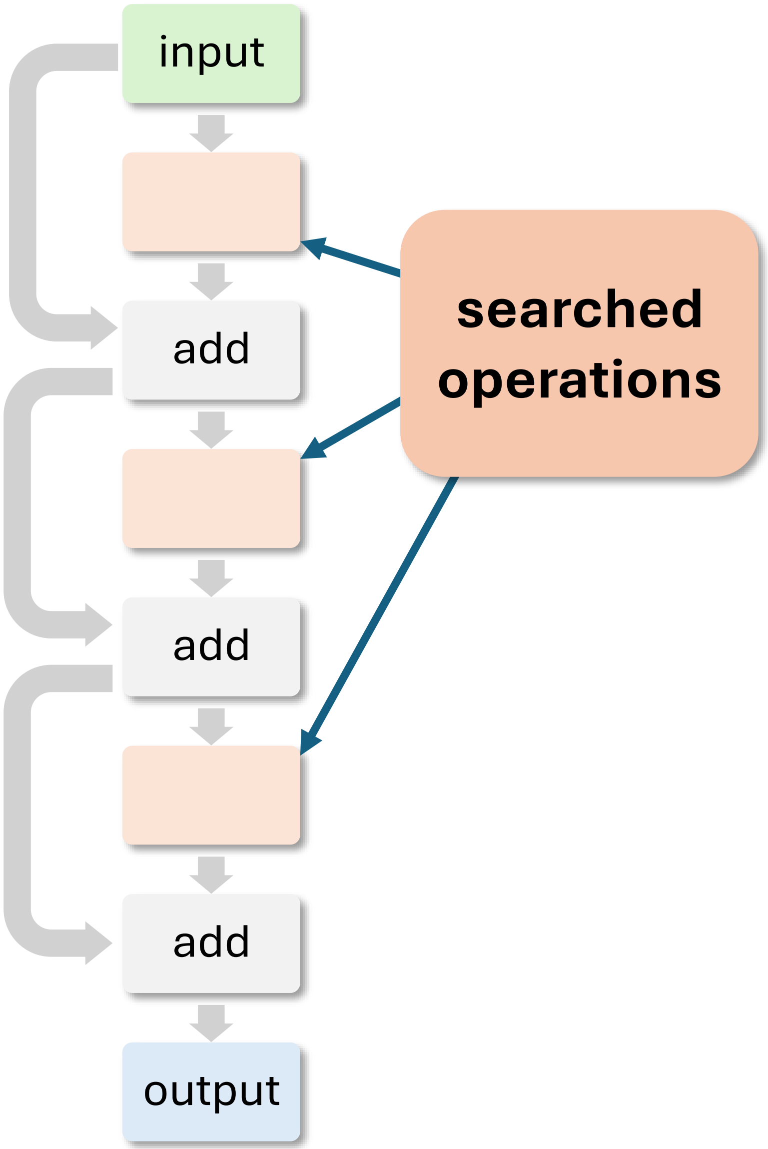
Why did the best NAS methods fail?

- NAS has classically aimed for a good neural network **graph topology**
- but most breakthrough architectures have similar topologies built around **one core operation**

Idea: fix a topology and search for good ~~**operations** over an expressive search space~~ convolutional **kernel sizes / dilation rates** to use at each layer

[Shen*-**K***-Talwalkar, NeurIPS 2022]

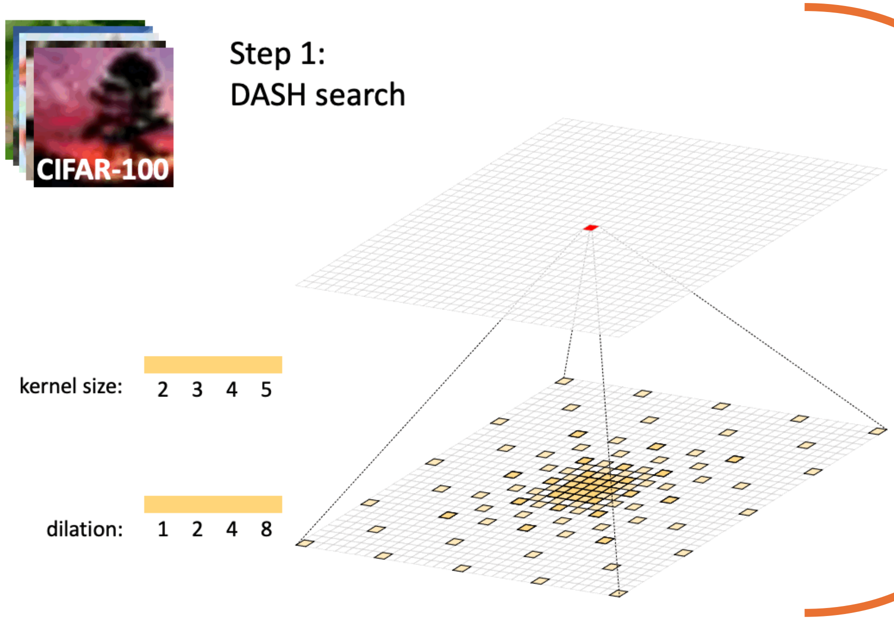
DASH

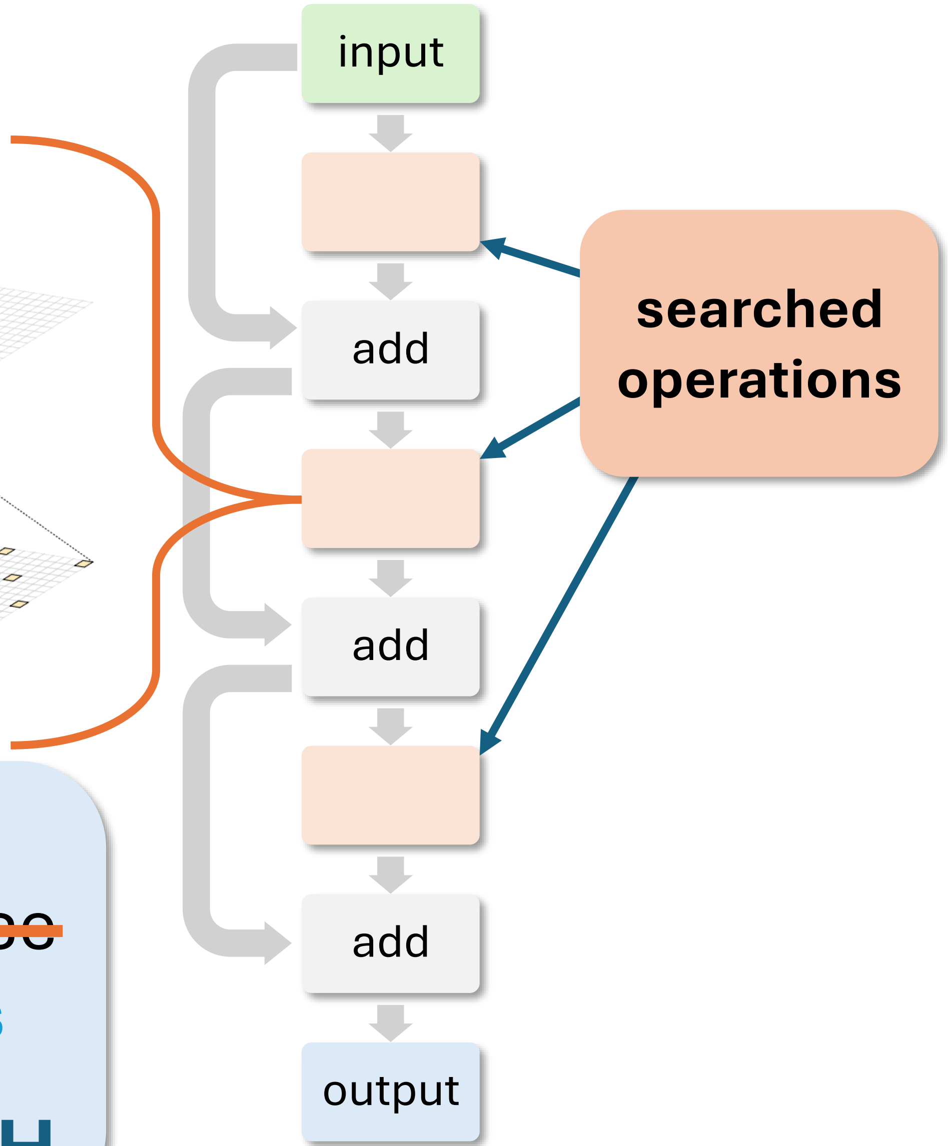


Transferring **neural architectures** to diverse data modalities

[Roberts*-**K***-Dao-Li-Ré-Talwalkar, NeurIPS 2021; Shen*-**K***-Talwalkar, NeurIPS 2022]

Step 1:
DASH search





Idea: fix a topology and search for good ~~**operations** over an expressive search space~~ convolutional **kernel sizes / dilation rates** to use at each layer

[Shen*-**K***-Talwalkar, NeurIPS 2022]

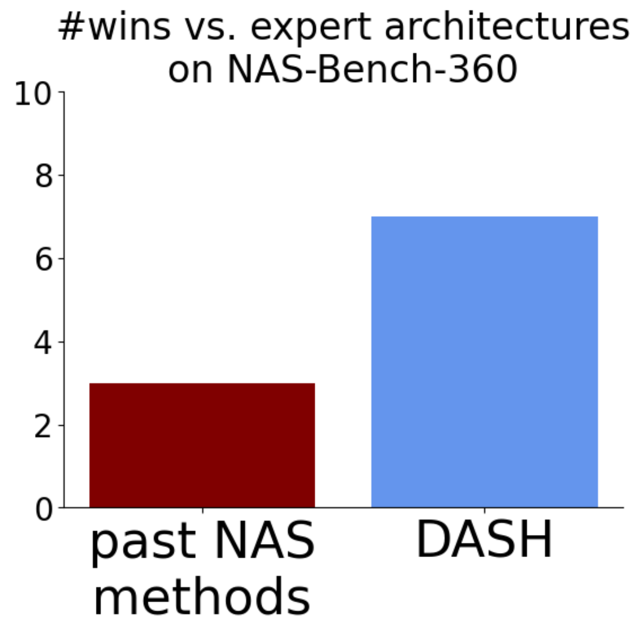
DASH

The DASH algorithm for NAS

[Shen*-**K***-Talwalkar, NeurIPS 2022]

- fixes the topology and searches for good convolutional **kernel sizes / dilation rates** to use at each layer
- finds good architectures **faster than they take to train**
- beats **7 of 10 expert architectures** on NAS-Bench-360
- adapted by the **2nd-place team** in the AutoML Decathlon 2022 competition
- adapted to show that supervised learning is competitive with specialized FMs

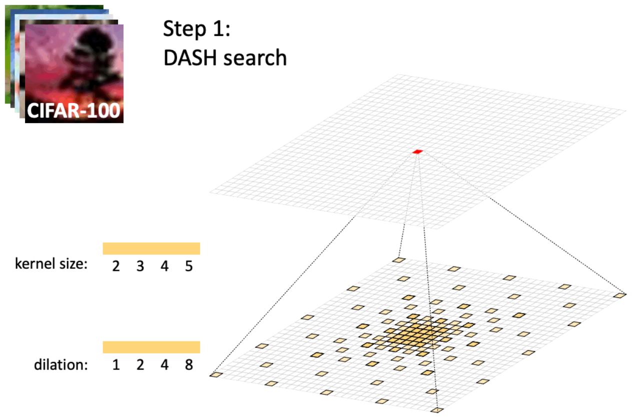
  [Xu*-Gupta*-Cheng-Shen-Shen-Talwalkar-**K**, ICLR 2025]



How do we use DASH to **investigate specialized FMs**?

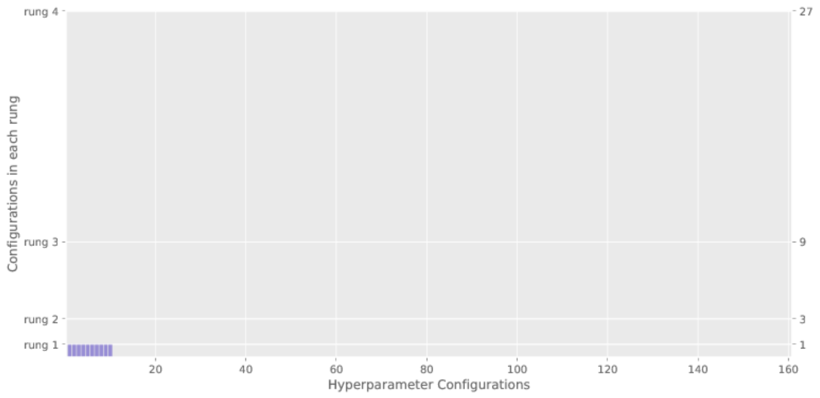
After finding an architecture, we set non-architectural (training) hyperparameters using the standard tuner **ASHA** in an inner loop

DASHA:



DASH

+



ASHA

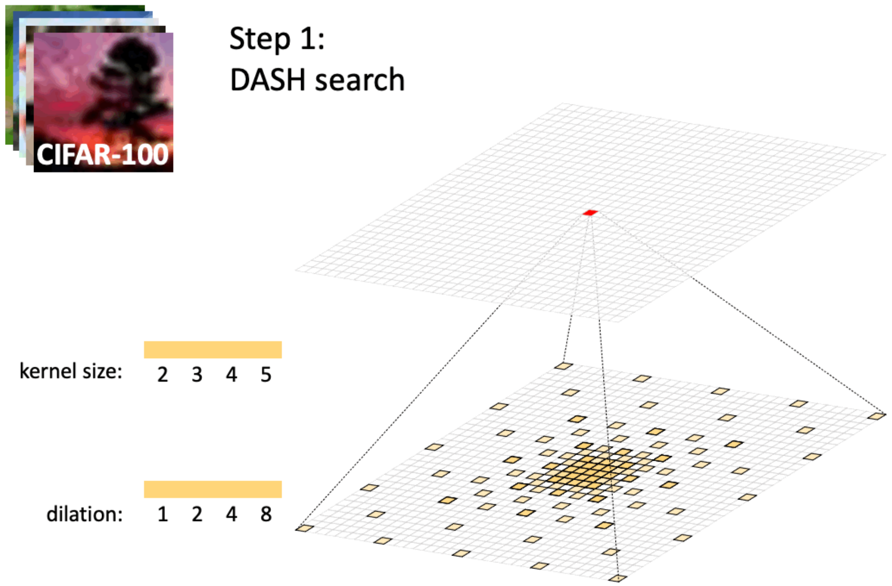
github.com/ritvikgupta199/DASHA

Li et al. *A system for massively parallel hyperparameter tuning.* MLSys 2020.

Summary:
Broadly effective architecture search

try DASH if your convolutional neural net is not doing as well as you'd like

1. works in different input dimensions
2. often very efficient
3. extensively evaluated on a diverse array of tasks
4. you want a strong baseline for your new specialized foundation model

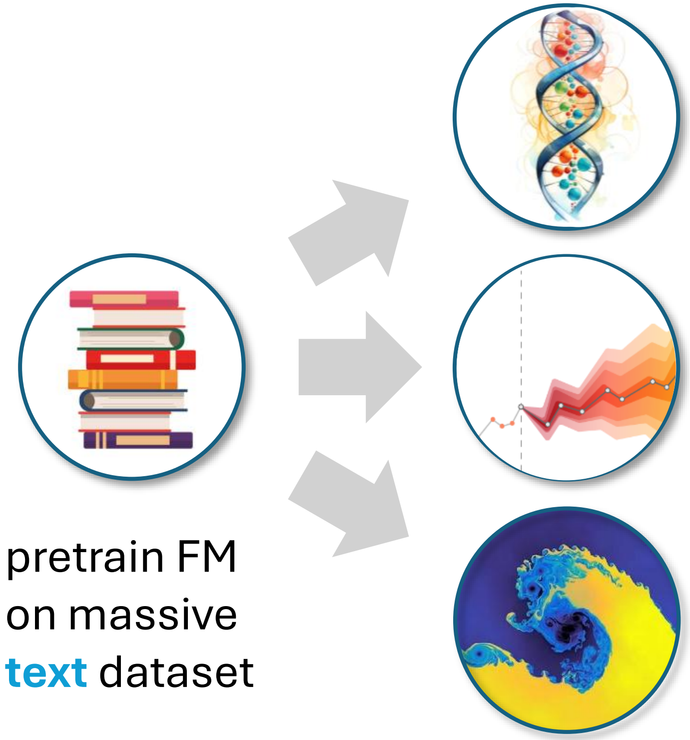


Ongoing direction: Transferring FMs across modalities

recent work shows you can fine-tune text and vision FMs on **completely different** data modalities and get results competitive with specialized FMs
[Shen-Li-Dery-Staten-**K**-Neubig-Talwalkar, ICML 2023]

Challenges:

1. scaling to frontier models (many-billion parameters)
2. will this still work with better benchmarks not solved by supervised learning?



pretrain FM on massive **text** dataset

fine-tune FM on **genomics** tasks
[Chen-Shen-**K**-Ma-Talwalkar, 2025]

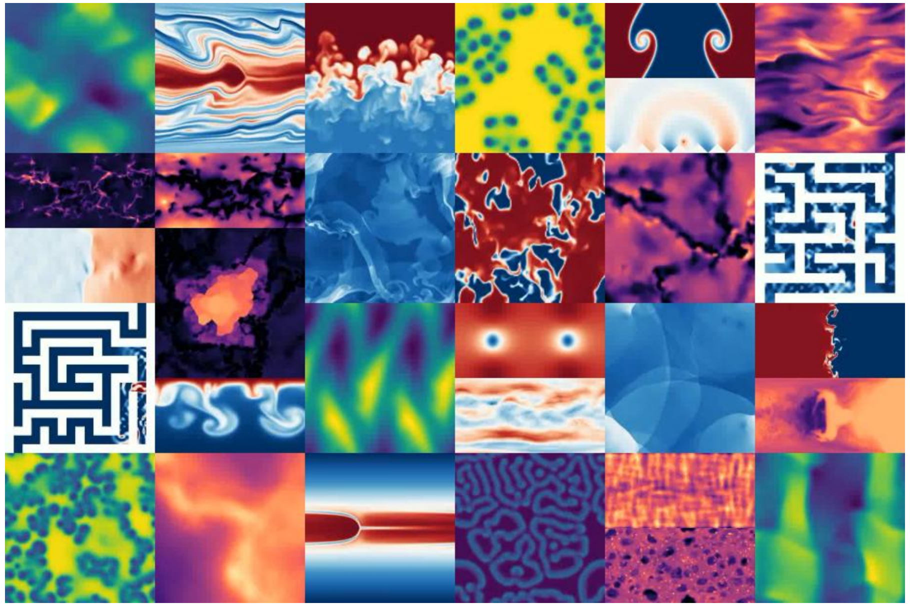
fine-tune FM on **time series** tasks
[Zhou et al., NeurIPS 2023; Jin et al., ICLR 2024]

fine-tune FM on **PDE-solving** tasks
[Shen et al., ICML 2023; Shen et al., TMLR 2024]

Ongoing direction:
Developing **better specialized foundation models**

specialized domains may yet have their BERT moments with

1. better, **domain-specific** modeling
2. better, **use-case-aware** benchmarks
   - e.g. CASP for protein-folding
     [Kryshtafovych et al., 2021]
3. better, larger **pretraining data**
   - e.g. the Well for PDE solving
     [Ohana et al., 2024]



TheWell

Thank you!

khodak@wisc.edu
pages.cs.wisc.edu/~khodak


Zongzhe Xu


Ritvik Gupta


Wenduo Cheng

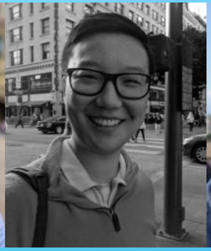

Alex Shen


Junhong Shen


Ameet Talwalkar


Nick Roberts


Tri Dao


Liam Li


Renbo Tu


Chris Ré


Fred Sala


Lucio Dery


Jian Ma


Graham Neubig


Corey Staten