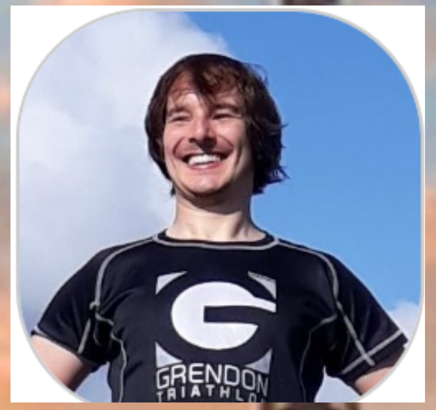
ICCV23
PARIS

🍔 BigMAC: Big Model Adaptation for Computer vision

Speakers:

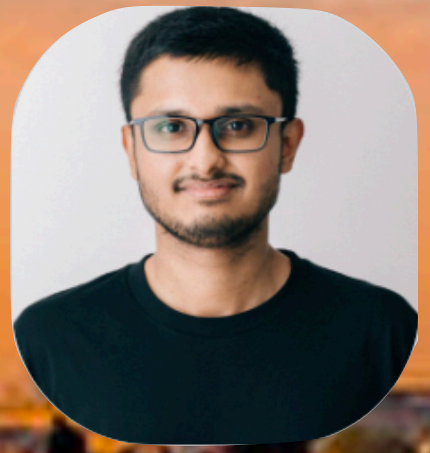

Neil Houlsby
Google Brain

Maria Attarian
Google Brain, U. of Toronto

Ludwig Schmidt
U. of Washington

Ishan Misra
Meta AI

Aditi Raghunathan
Carnegie Mellon University


Sayak Paul
HuggingFace

Carl Vondrick
Columbia University

Organised by:

Yuki M. Asano,

Tengda Han

Mathilde Caron

Phillip Isola

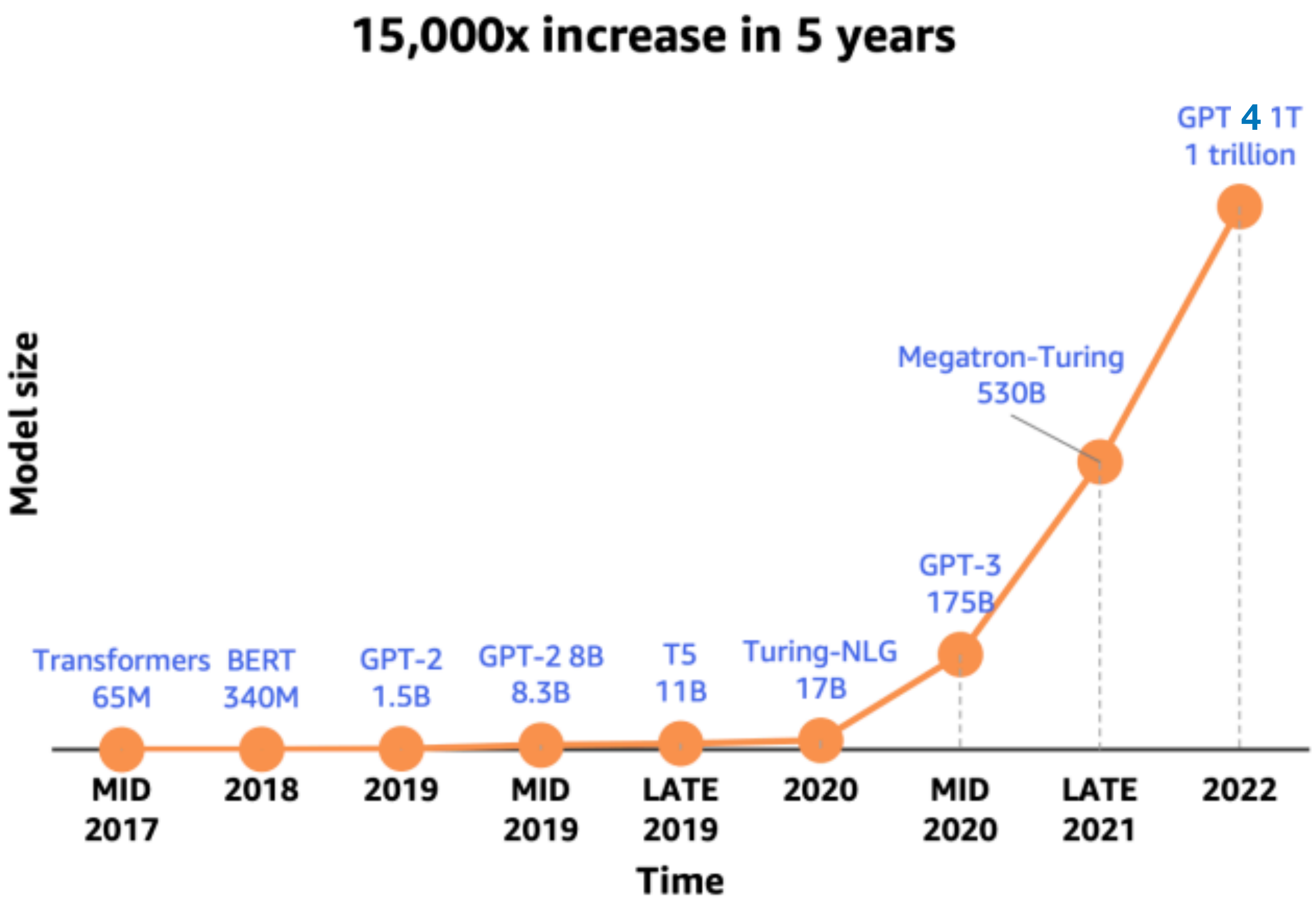
Serge Belongie

BigMAC Schedule

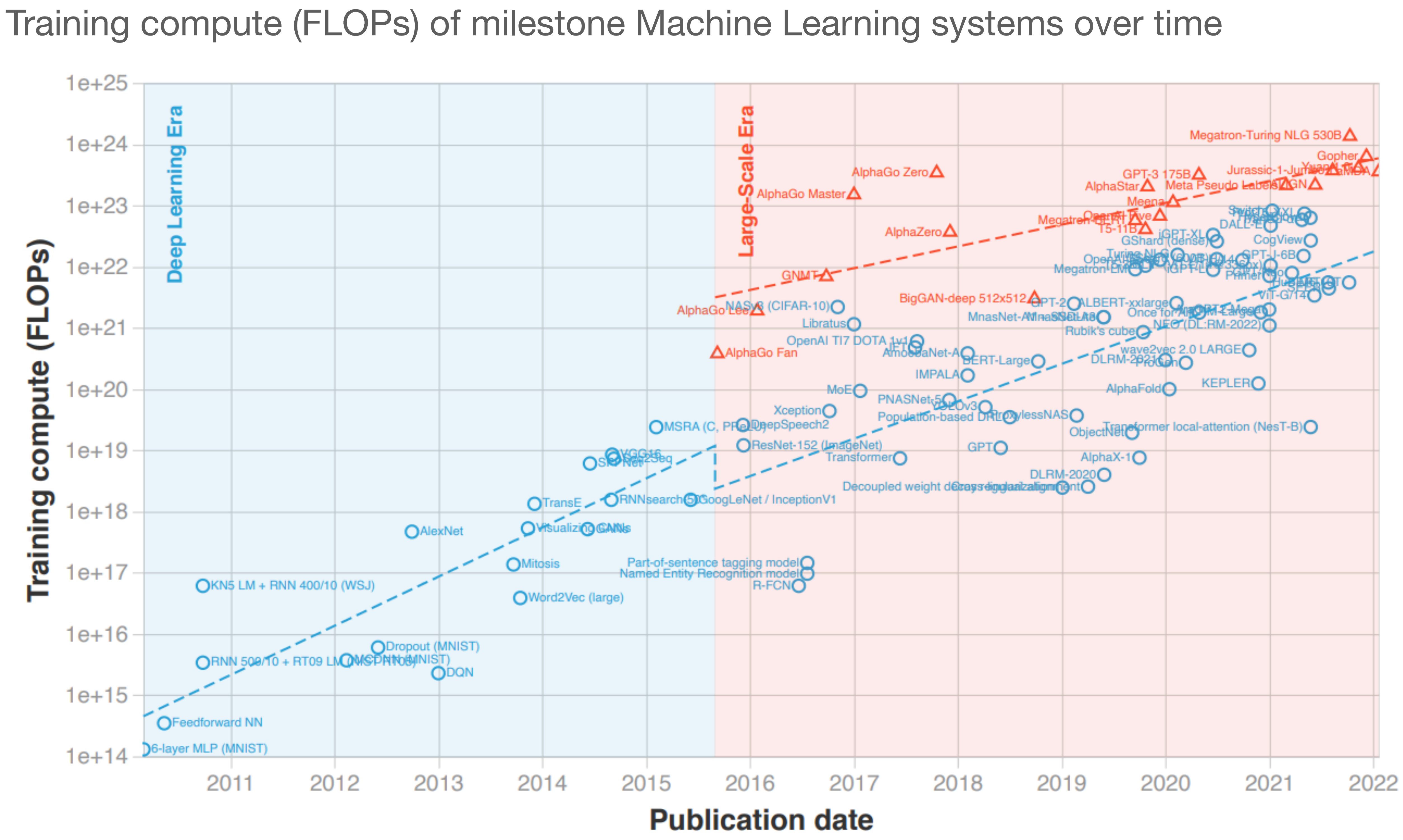
Time	Speaker	Affiliation
9:00 am - 9:15 am	Welcome and Introduction	
9:15 am - 9:45 am	Neil Houlsby	Google Brain
9:45 am - 10:15 am	Maria Attarian	Google Brain, University of Toronto
10:15 am - 10:45 am	Ludwig Schmidt	University of Washington
10:45 am - 11:00 am	Coffee Break	
10:00 am - 11:30 am	Ishan Misra	Meta AI
11:30 am - 12:00 pm	Aditi Raghunathan	Carnegie Mellon University
12:00 pm - 12:30 pm	Sayak Paul	HuggingFace
12:30 pm - 1:00 pm	Carl Vondrick	Columbia University
1:00 pm	Closing remarks	

Why now?

Big Models have arrived. So we need to figure out how to use them.



15,000x increase in 5 years

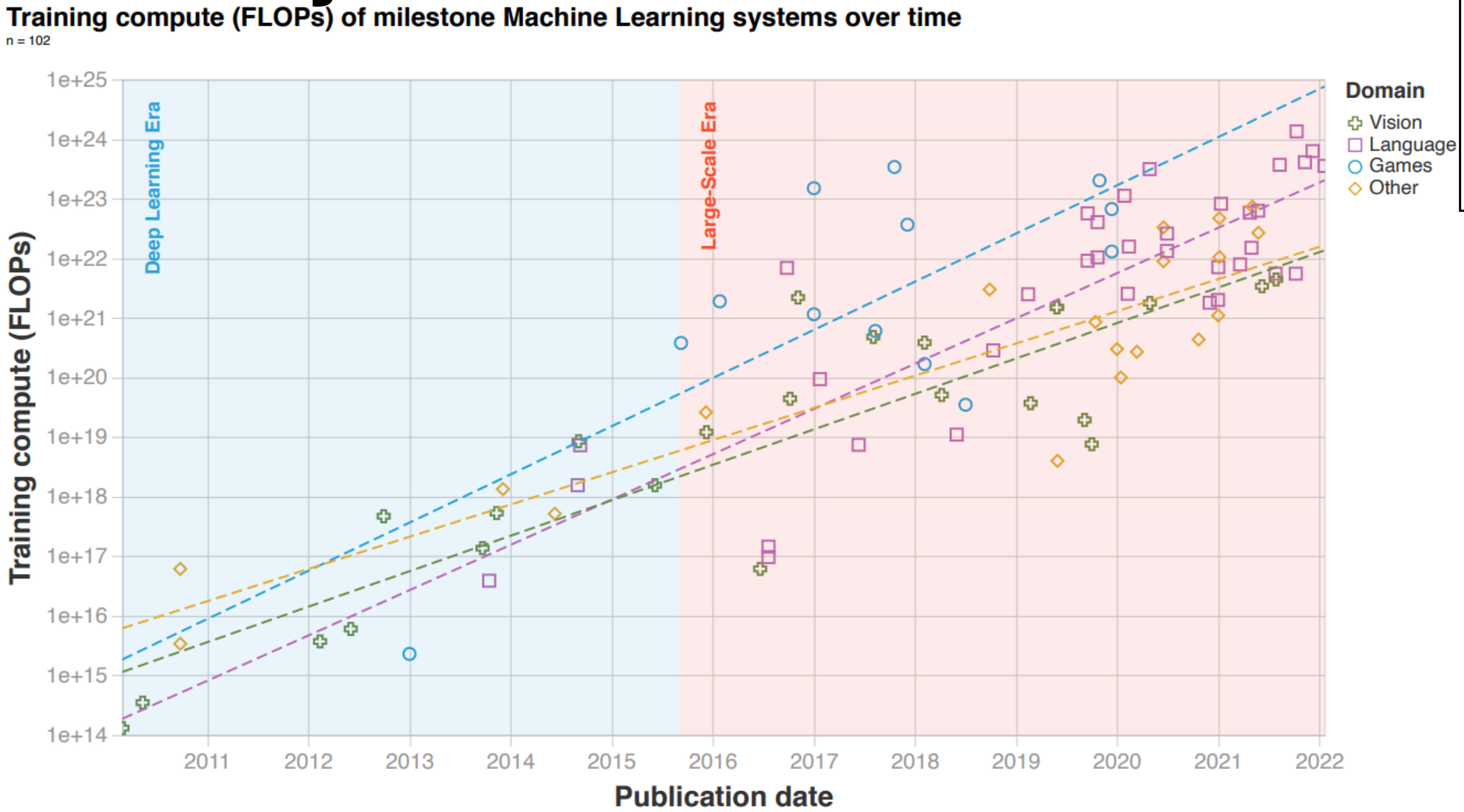


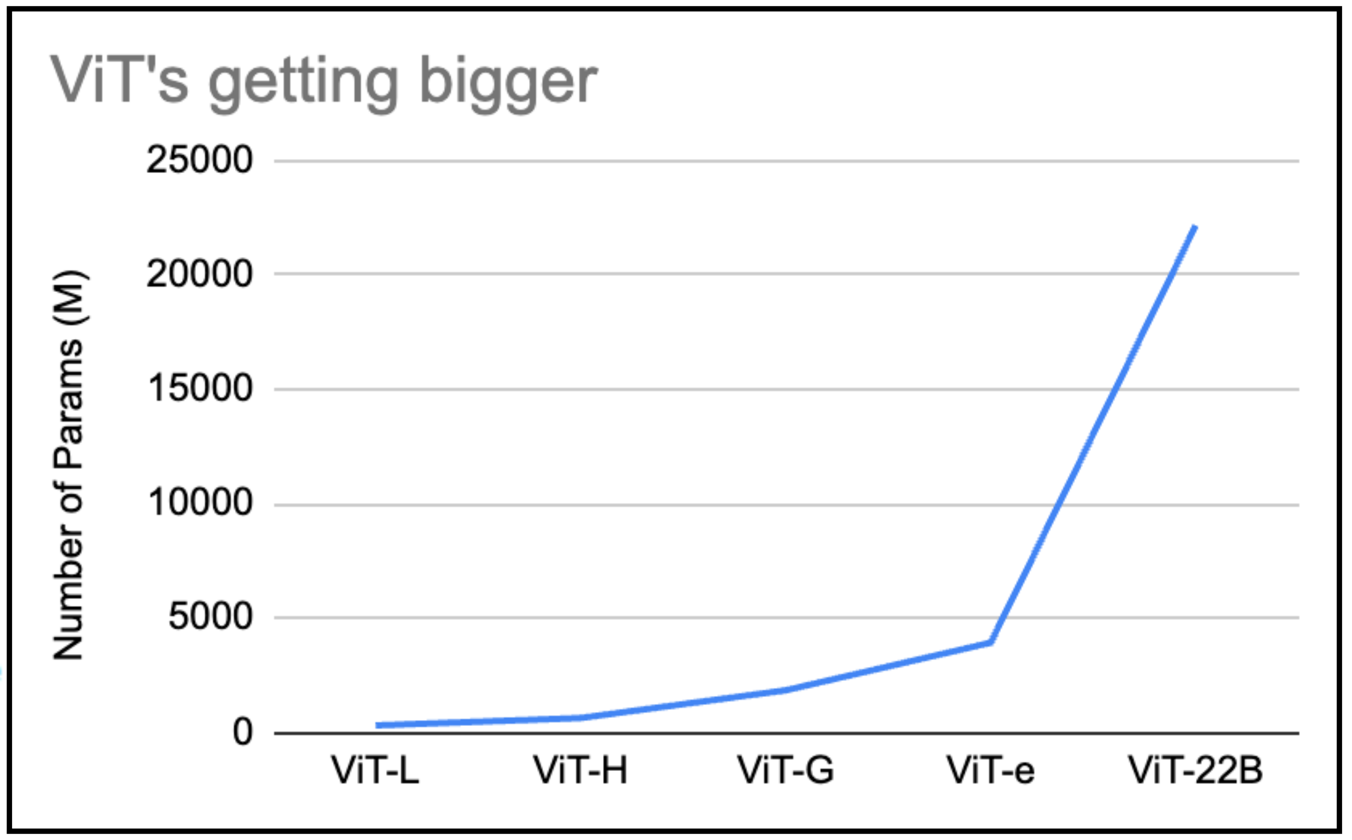
Training compute (FLOPs) of milestone Machine Learning systems over time

- tremendous growth in LLM sizes
- Multiple growth trends

Sevilla et al. Compute Trends Across Three Eras of Machine Learning. 2022

That's just NLP... or is it? No.





"Compute Requirements: ViT-22B was trained on 1024 TPU V4 chips [..]"

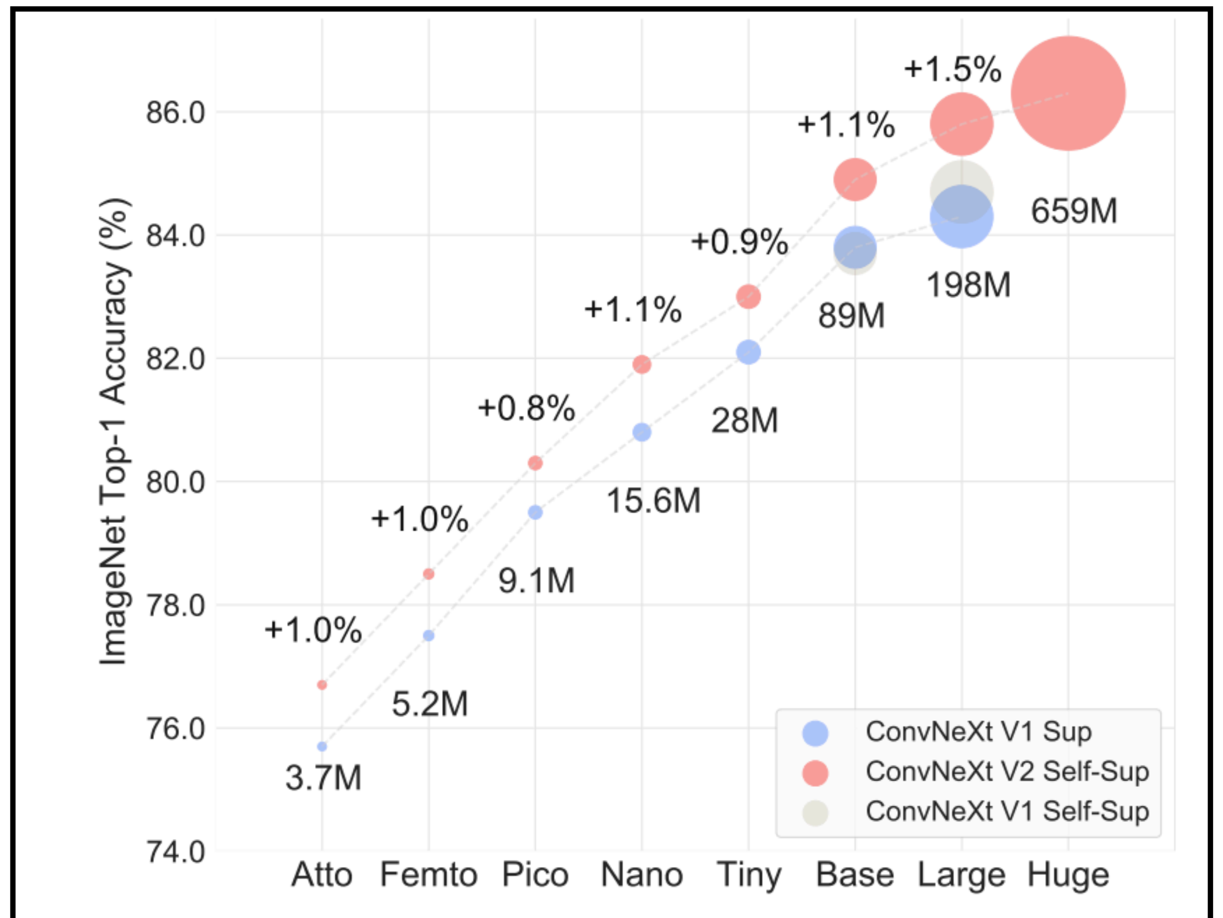
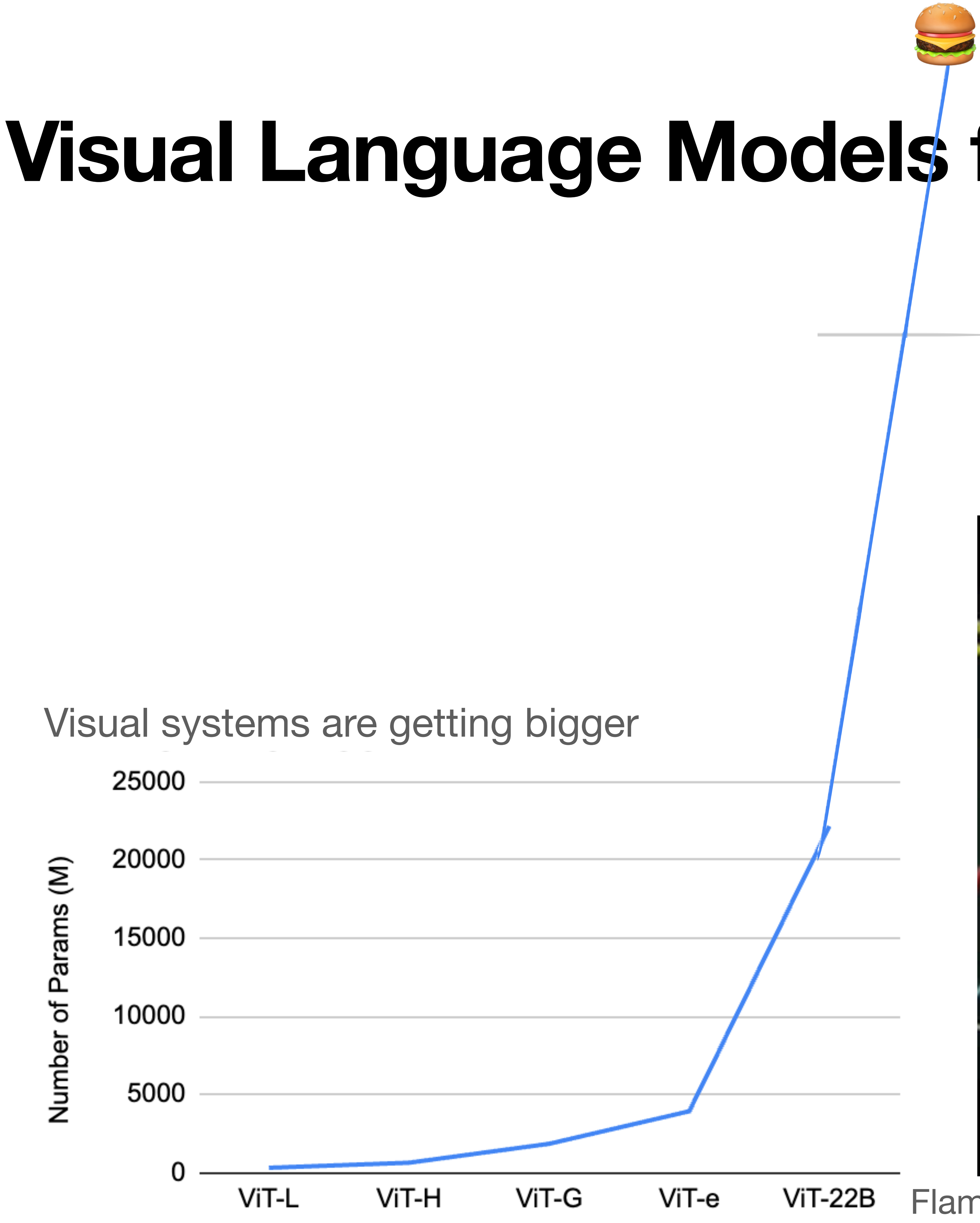


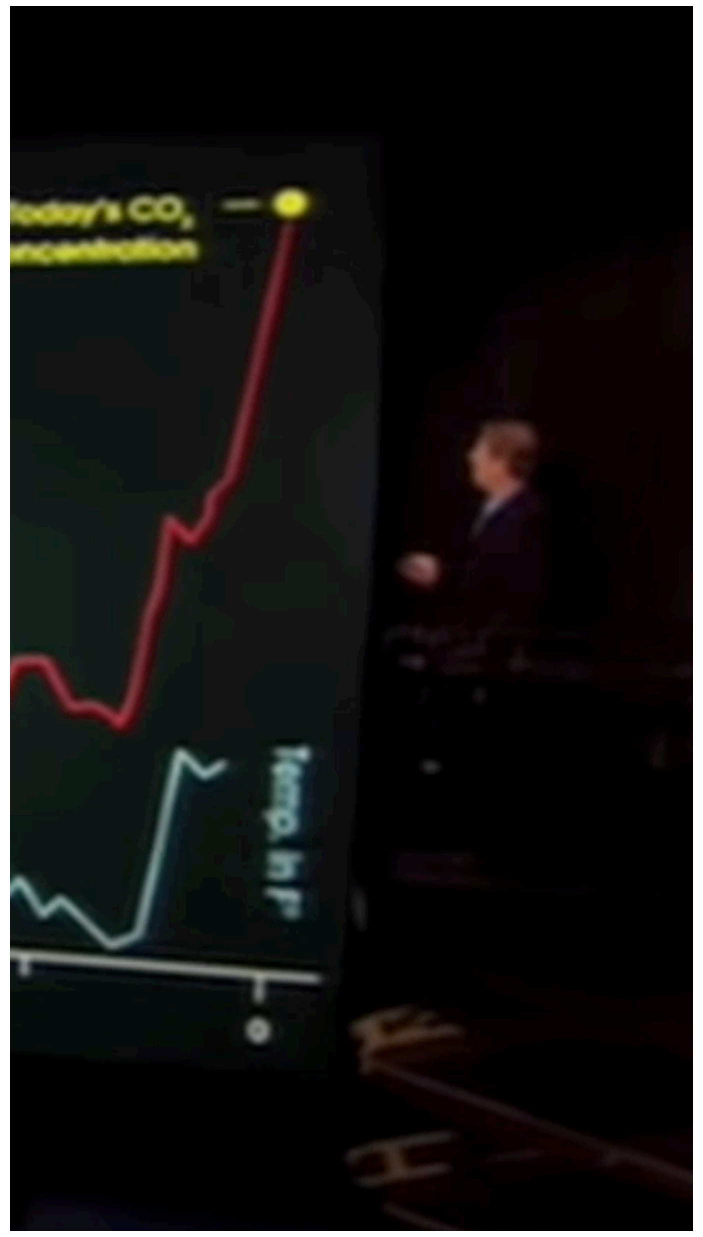
Figure 1. **ConvNeXt V2 model scaling**. The ConvNeXt V2 model, which has been pre-trained using our fully convolutional masked autoencoder framework, performs significantly better than the previous version across a wide range of model sizes.

- With Transformers in vision, scale has arrived here too.

Sevilla et al. Compute Trends Across Three Eras of Machine Learning. 2022
Dehghani et al. Scaling Vision Transformers to 22 Billion Parameters. ICML 2023
Woo et al. ConvNeXt V2: Co-designing and Scaling ConvNets with Masked Autoencoders. CVPR 2023

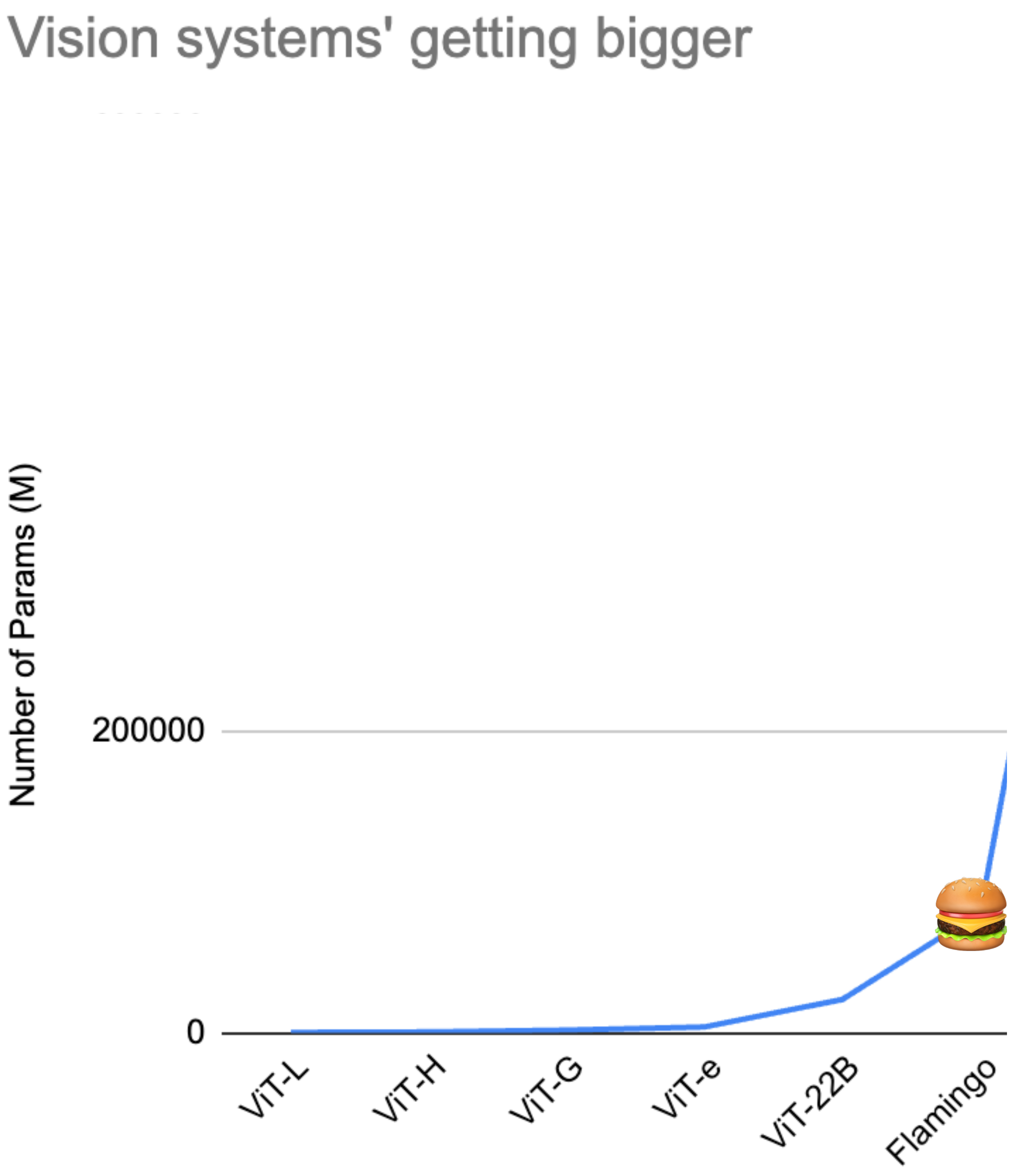
Visual Language Models further increases the #params, by a lot





Alayrac et al. Flamingo: a Visual Language Model for Few-Shot Learning. NeurIPS 2022

Visual Language Models further increases the #params, by a lot



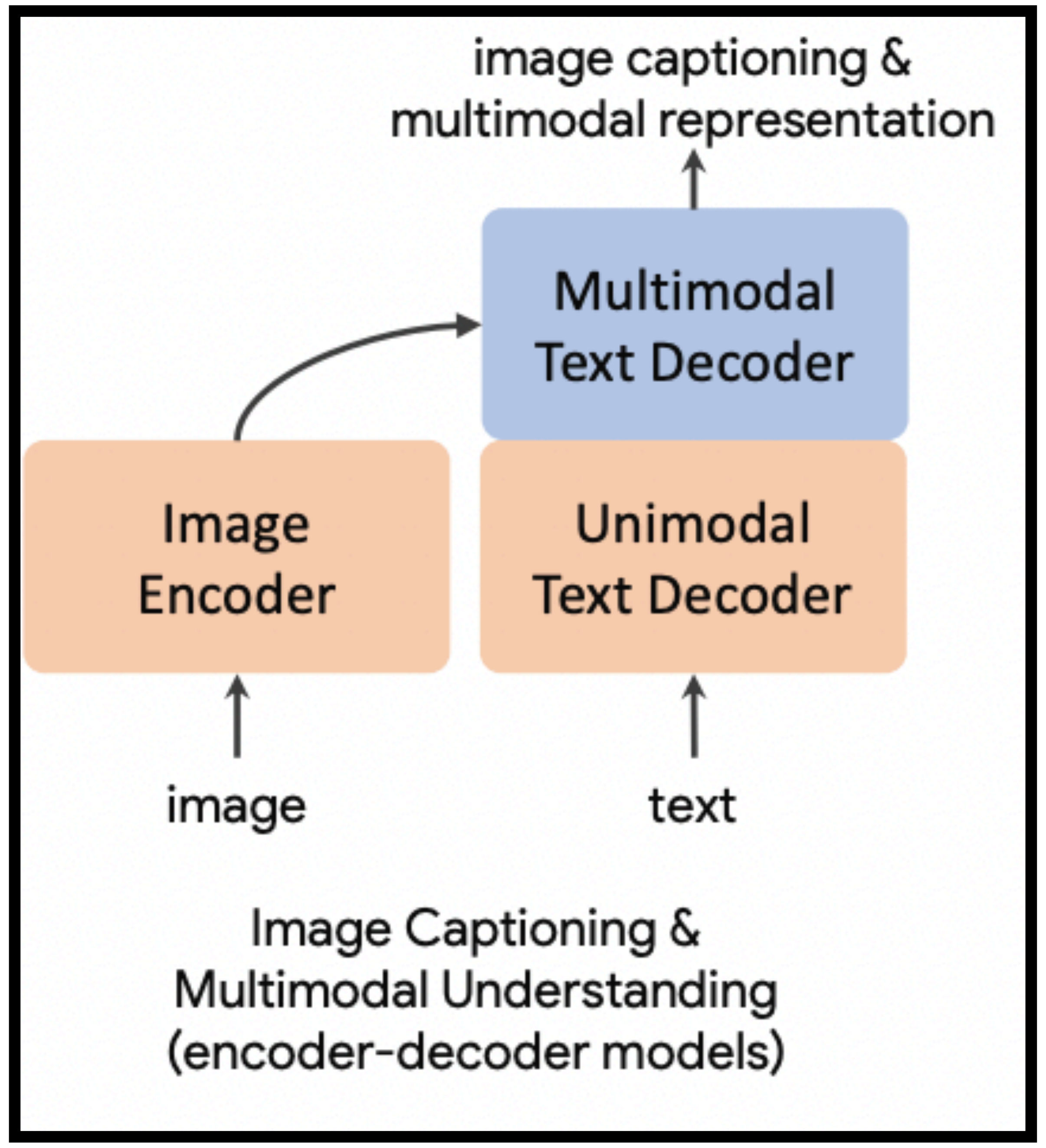


image captioning & multimodal representation
Multimodal Text Decoder
Image Encoder
Unimodal Text Decoder
image
text
Image Captioning & Multimodal Understanding (encoder-decoder models)

General design of VLMs

- Flamingo, BLIP, CM3, Frozen, CoCa, ALIGN, Fromage, VisualLLM, ...

Yu et al. Contrastive Captioners are Image-Text Foundation Models. TMLR 2022
Alayrac et al. Flamingo: a Visual Language Model for Few-Shot Learning. NeurIPS 2022
Tsimpoukelli et al. Multimodal Few-Shot Learning with Frozen Language Models. NeurIPS 2021.
Koh et al. Grounding Language Models to Images for Multimodal Generation. 2023
Li et al. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. 2023
Aghajanyan et al. CM3: A Causal Masked Multimodal Model of the Internet. 2022

The base/foundation vs the adaptation.

(Too a large extent) a well known recipe.

@tipsybartender)

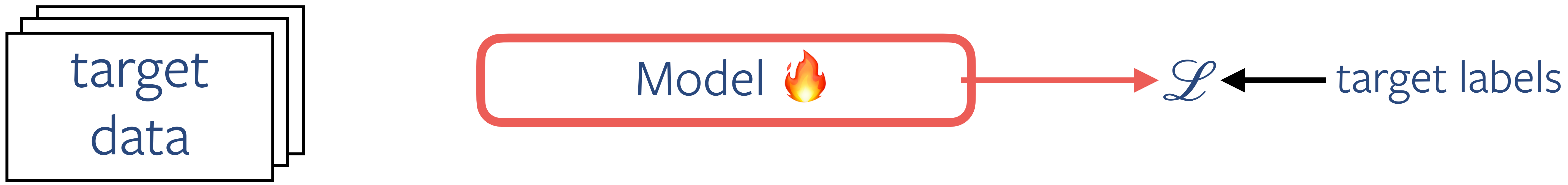
Adaptation strategies require **small work/parameters/GPU,** but have a **large effect.**

What to do with those big models?

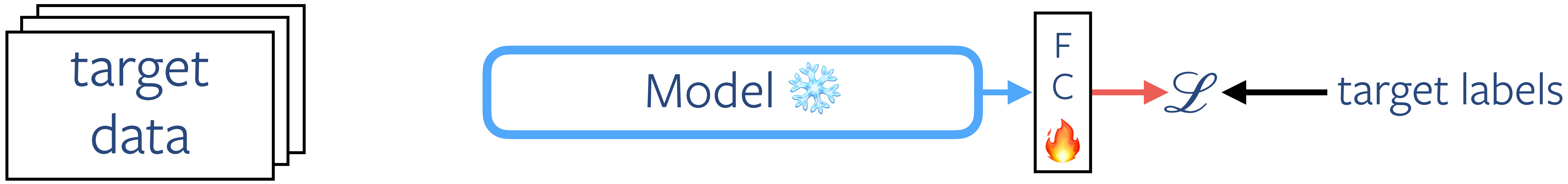
From my Self-supervised and vision-language learning lectures 2023

Main ways of adapting models (1/2)

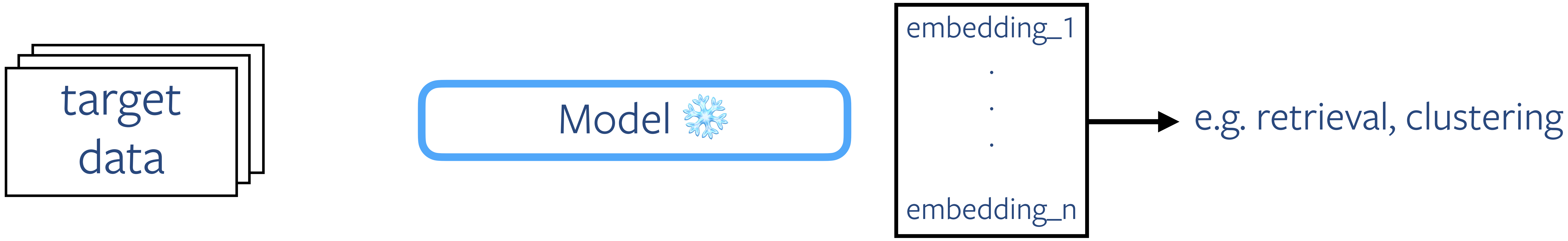
Full-finetuning



Limited-finetuning (e.g. linear probing)

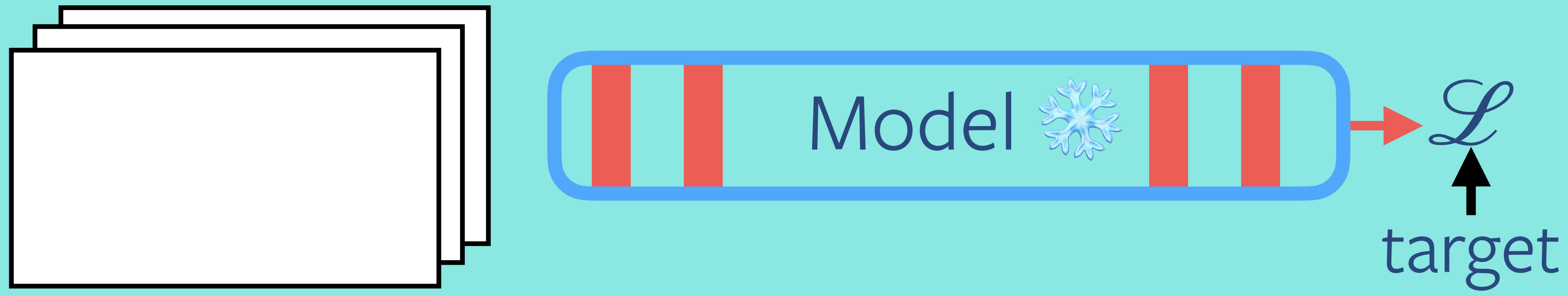


No-finetuning (e.g. used for retrieving similar instances)



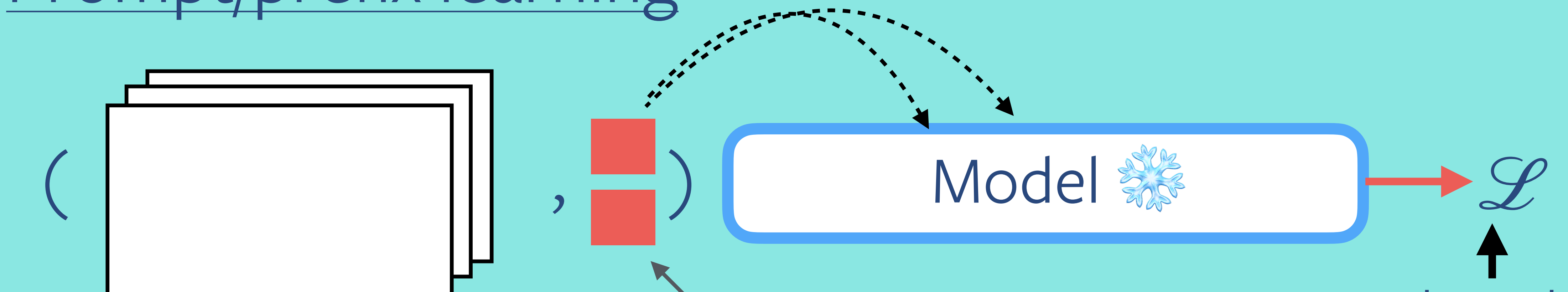
Parameter-efficient Finetuning (PEFT) ideas

More params inside model: Adapters, LoRA etc.



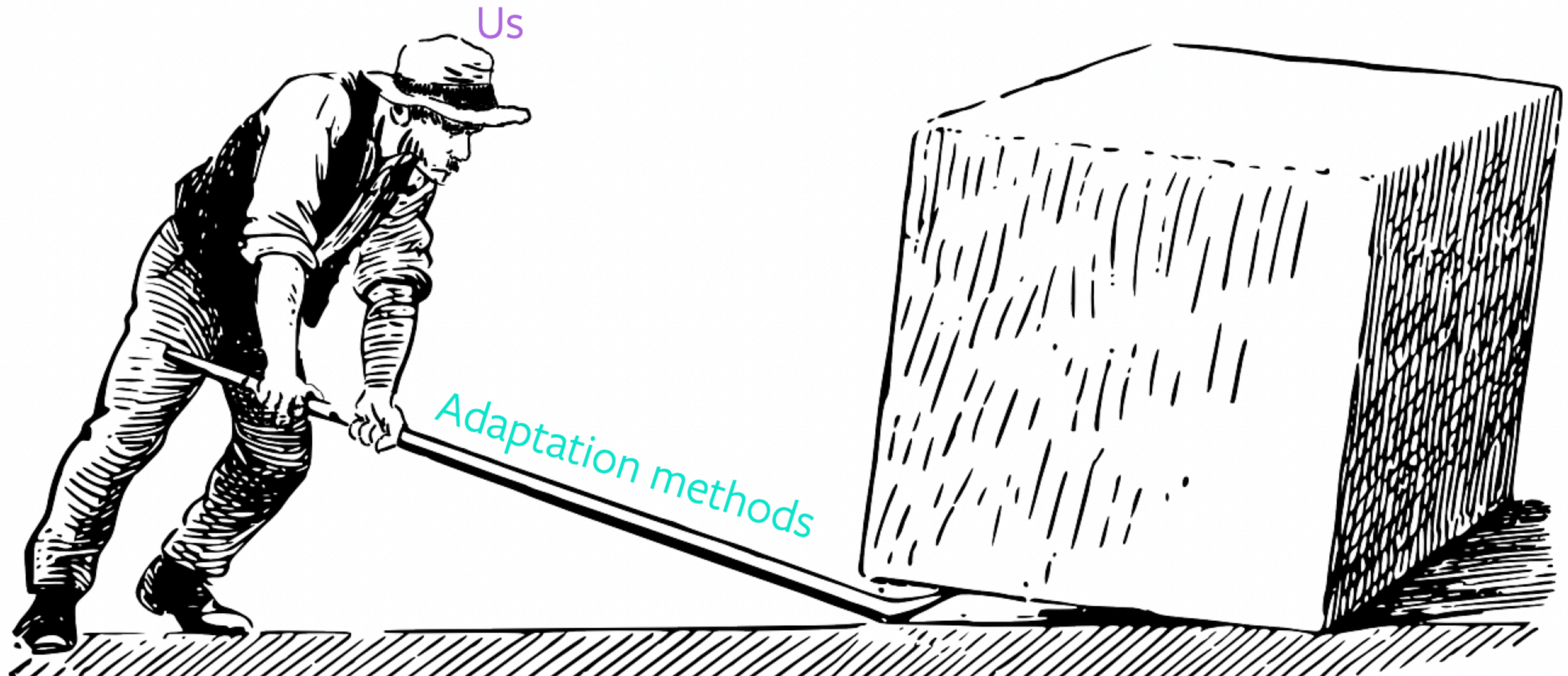
- e.g.: 1x1 convs, Residual-MLPs, only BN or bias params, binary masks, low-rank adaptation of matmuls

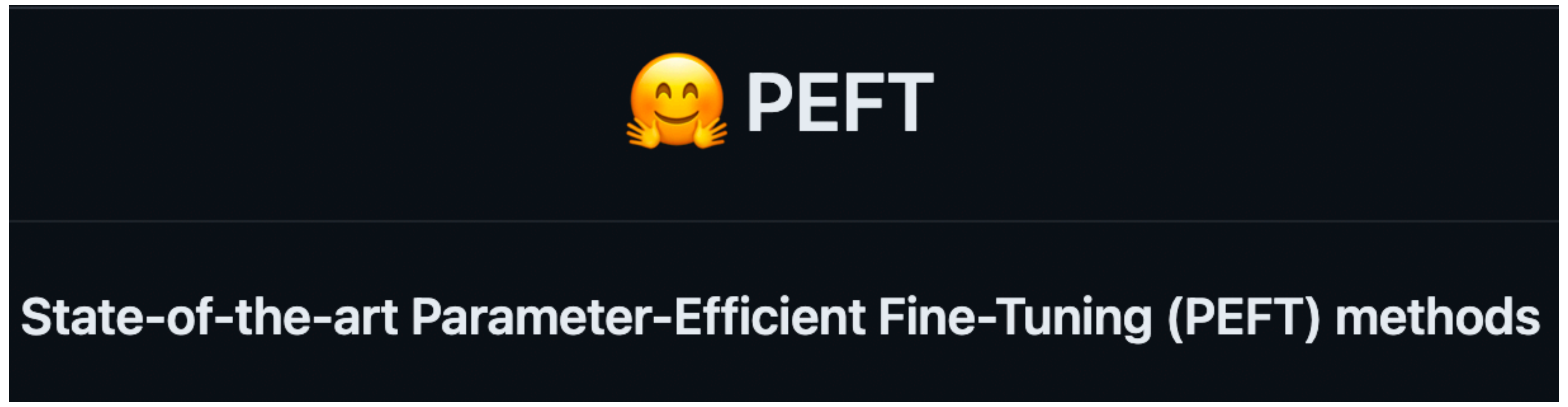
Prompt/prefix learning



learnable, additional inputs

- similar to prompt manual engineering [like "step-by-step" or "trending on artstation"]





🤗 PEFT

State-of-the-art Parameter-Efficient Fine-Tuning (PEFT) methods

https://github.com/adapter-hub/adapter-transforme

https://github.com/huggingface/peft

Schedule

Time	Speaker	Affiliation
9:00 am - 9:15 am	Welcome and Introduction	
9:15 am - 9:45 am	Neil Houlsby	Google Brain
9:45 am - 10:15 am	Maria Attarian	Google Brain, University of Toronto
10:15 am - 10:45 am	Ludwig Schmidt	University of Washington
10:45 am - 11:00 am	Coffee Break	
10:00 am - 11:30 am	Ishan Misra	Meta AI
11:30 am - 12:00 pm	Aditi Raghunathan	Carnegie Mellon University
12:00 pm - 12:30 pm	Sayak Paul	HuggingFace
12:30 pm - 1:00 pm	Carl Vondrick	Columbia University
1:00 pm	Closing remarks	