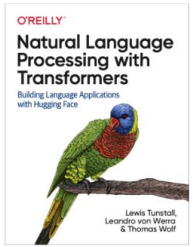
A guided tour through the Transformers landscape

Lewis Tunstall | Open Source @ Hugging Face | lewis@hf.co | @_lewtun

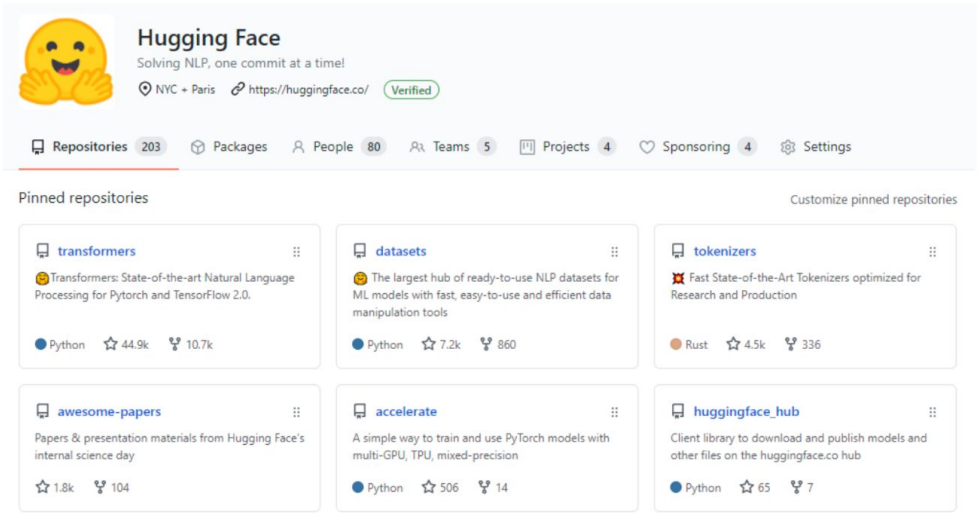
About me

huggingface.co/course/



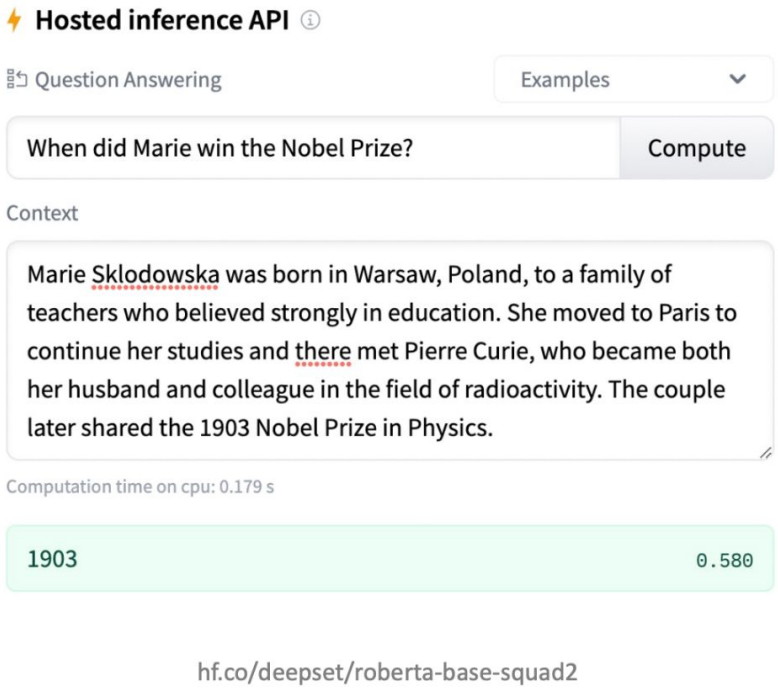
[NLP with Transformers]

Education



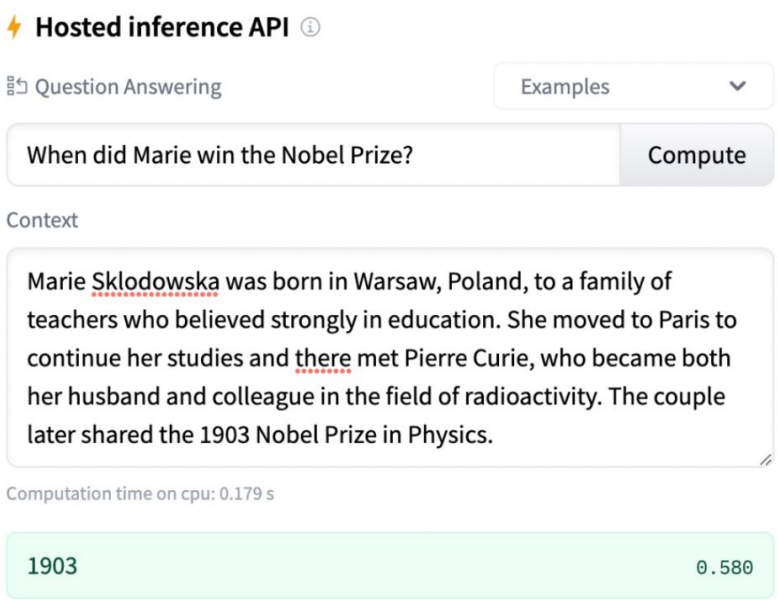
Open Source


TRANSFORMERS
THE GAME
Why all the fuss?
🤔

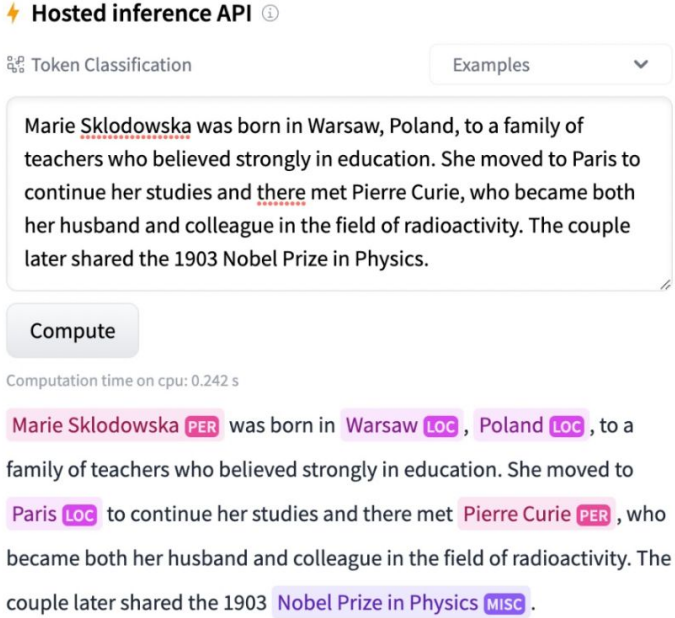


hf.co/deepset/roberta-base-squad2

Transformers are currently the best approach for analysing **text**



hf.co/deepset/roberta-base-squad2



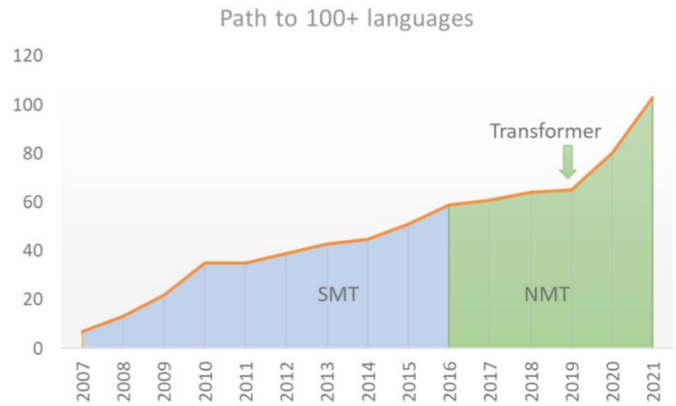
hf.co/xlm-roberta-large-finetuned-conll03-english

Transformers are currently the best approach for analysing text

Microsoft

Microsoft Translator now works across 103 languages

Microsoft adds 12 languages to its Microsoft Translate app that can help 84.6 million people.

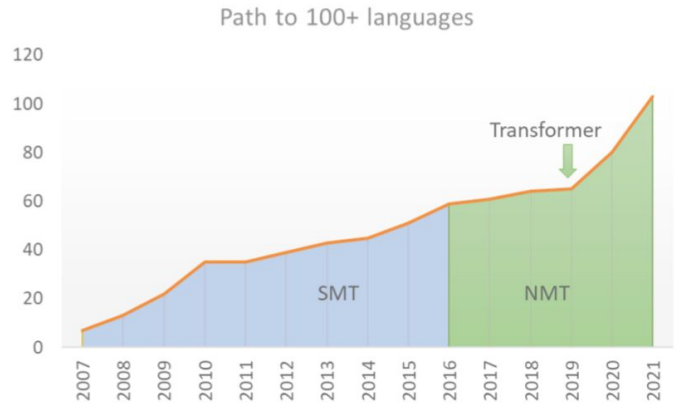


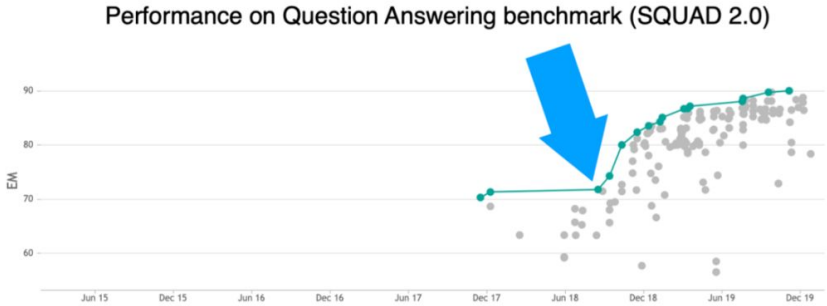
Transformers are currently the best approach for analysing **text**

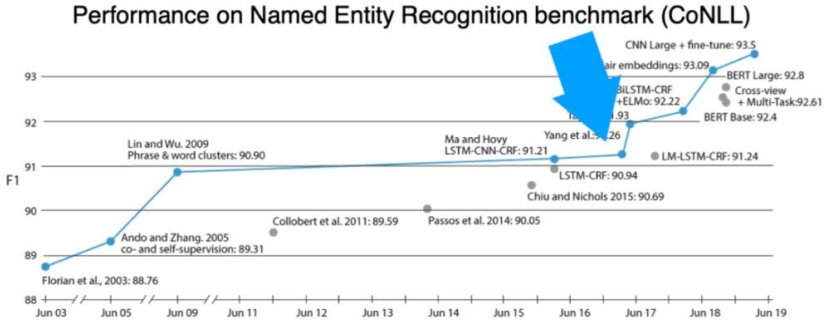
Microsoft

Microsoft Translator now works across 103 languages

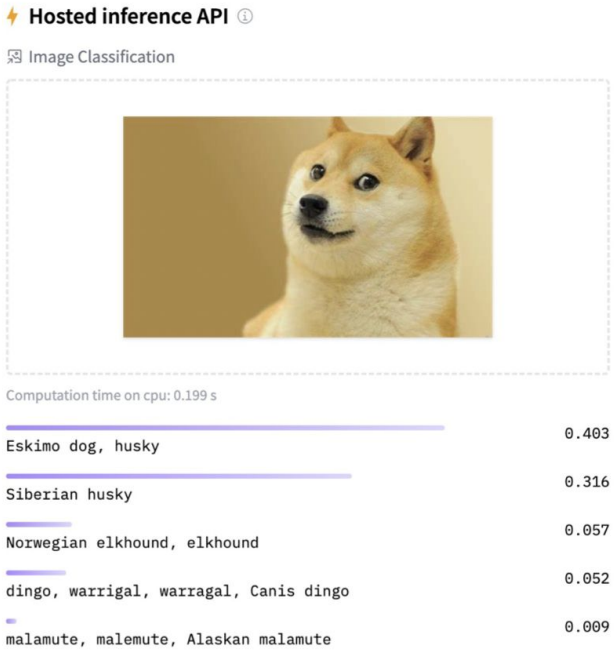
Microsoft adds 12 languages to its Microsoft Translate app that can help 84.6 million people.



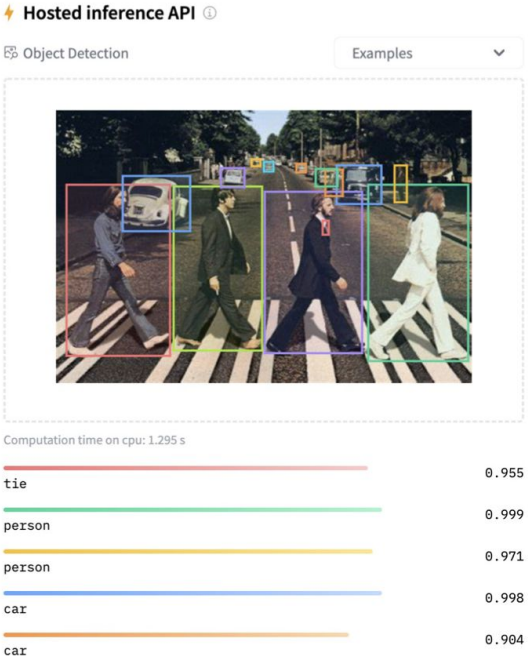




Transformers are currently the best approach for analysing **text**

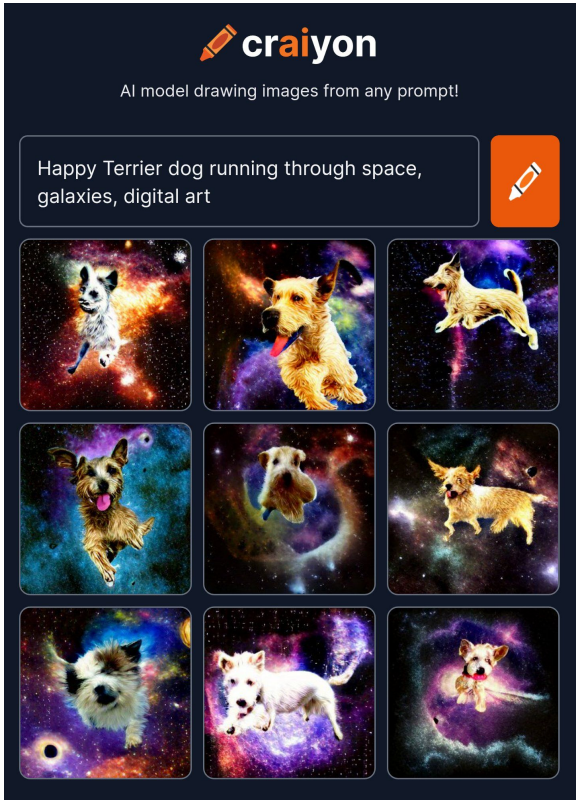


hf.co/microsoft/beit-base-patch16-224



hf.co/facebook/detr-resnet-50

Transformers are currently the best approach for analysing **images**



Transformers are currently the best approach for analysing **text** and **images**

TEXT DESCRIPTION

An astronaut Teddy bears A bowl of soup

that is a portal to another dimension that looks like a monster as a planet in the universe

knitted out of wool spray-painted on a wall made out of plasticine

→

DALL·E 2

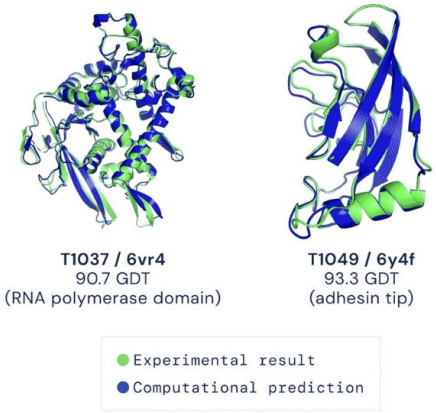
openai.com/dall-e-2/

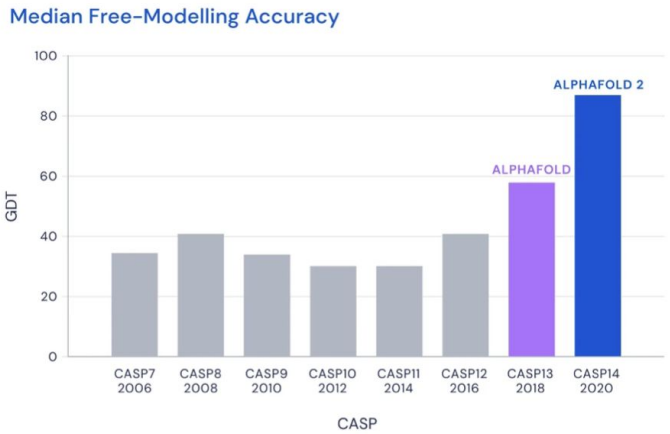
Transformers are currently the best approach for analysing text and images

NEWS | 22 July 2021

DeepMind's AI predicts structures for a vast trove of proteins

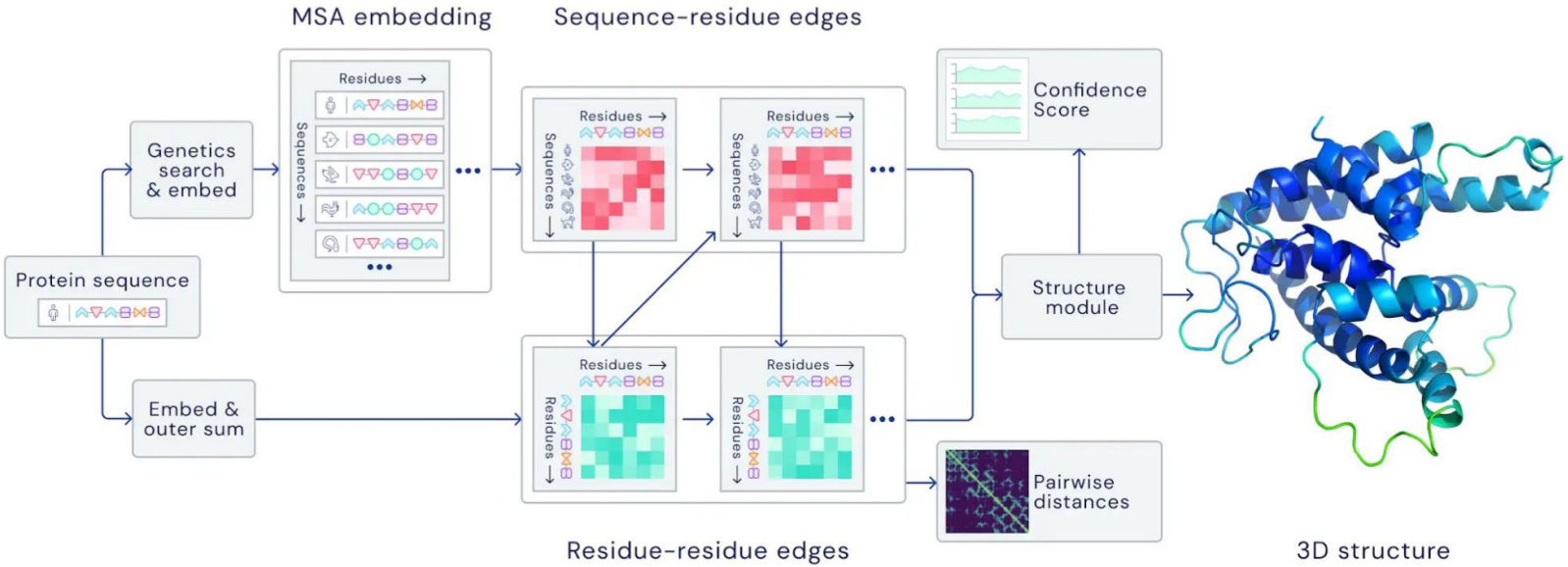
AlphaFold neural network produced a 'totally transformative' database of more than 350,000 structures from *Homo sapiens* and 20 model organisms.





https://www.deepmind.com/blog/alphafold-a-solution-to-a-50-year-old-grand-challenge-in-biology

And it's not just text and images where Transformers shine



https://www.deepmind.com/blog/alphafold-a-solution-to-a-50-year-old-grand-challenge-in-biology

And it's not just text and images where Transformers shine

Main ingredients

Attention mechanisms

Self-supervised learning (Pretraining)

Transfer learning (Fine-tuning)

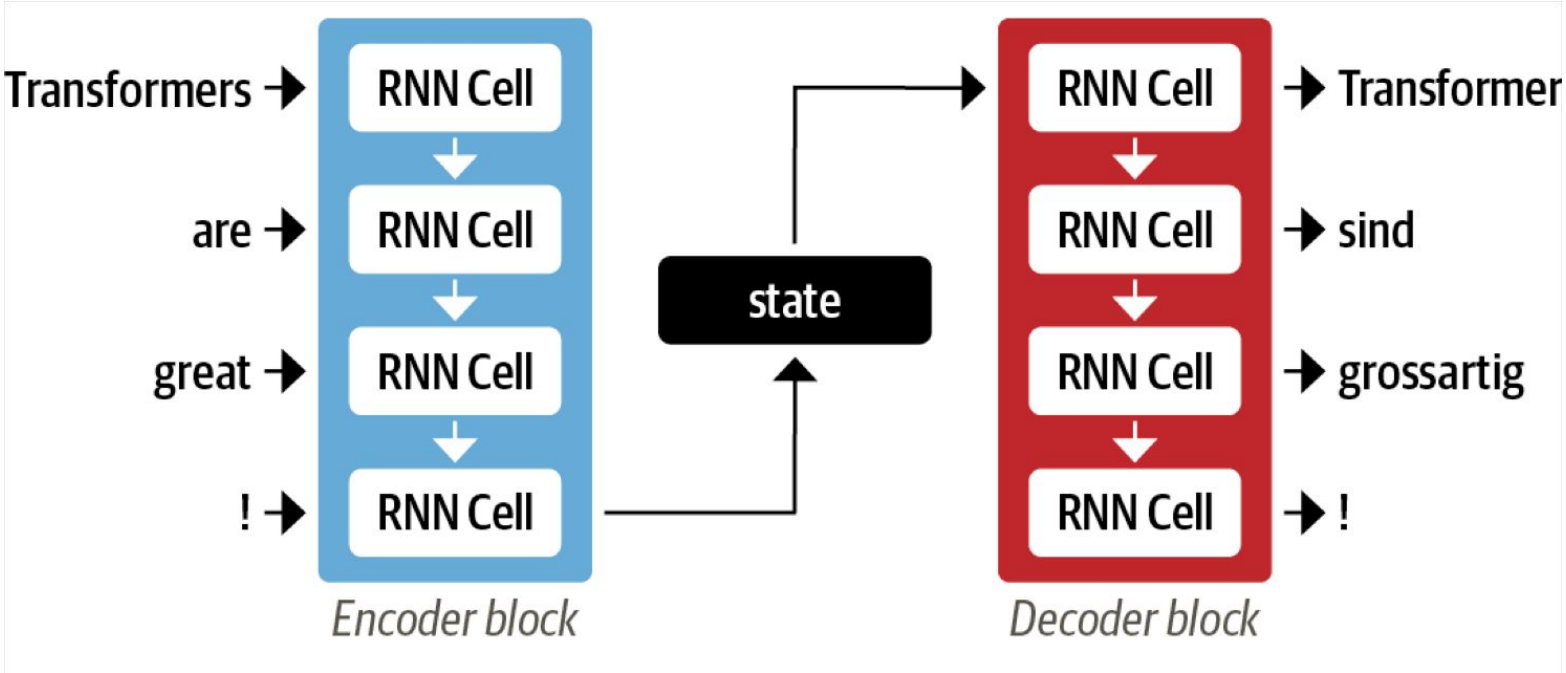
Main ingredients

Attention mechanisms

Self-supervised learning (Pretraining)

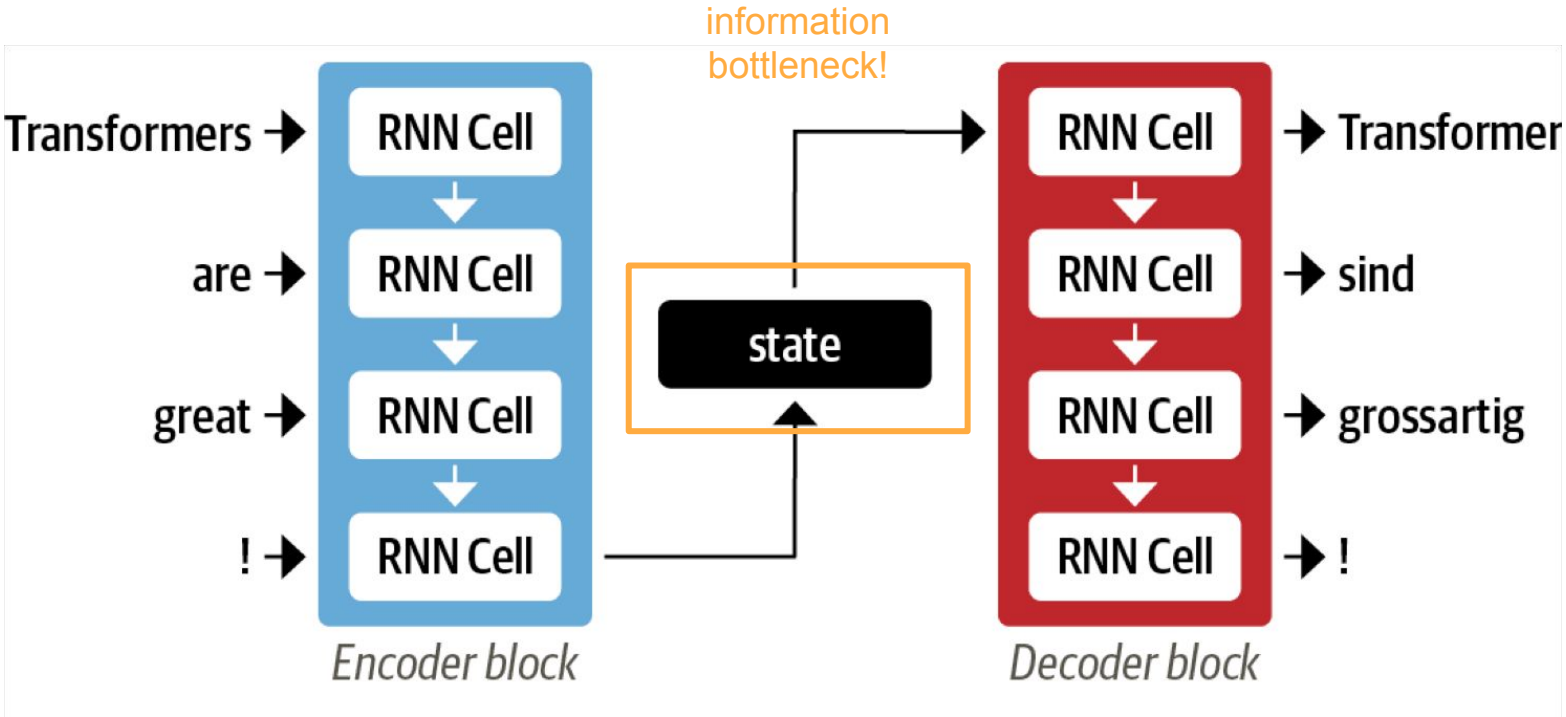
Transfer learning (Fine-tuning)

Attention mechanisms



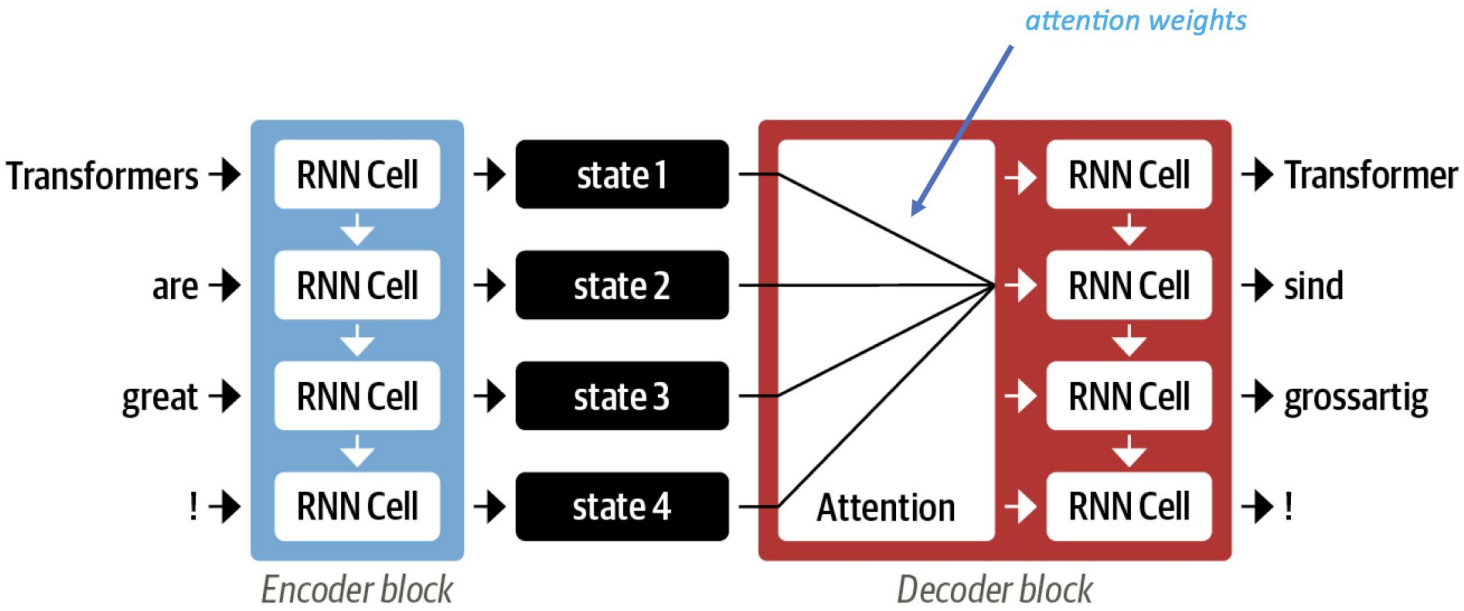
Originally developed for **recurrent neural networks**

Attention mechanisms



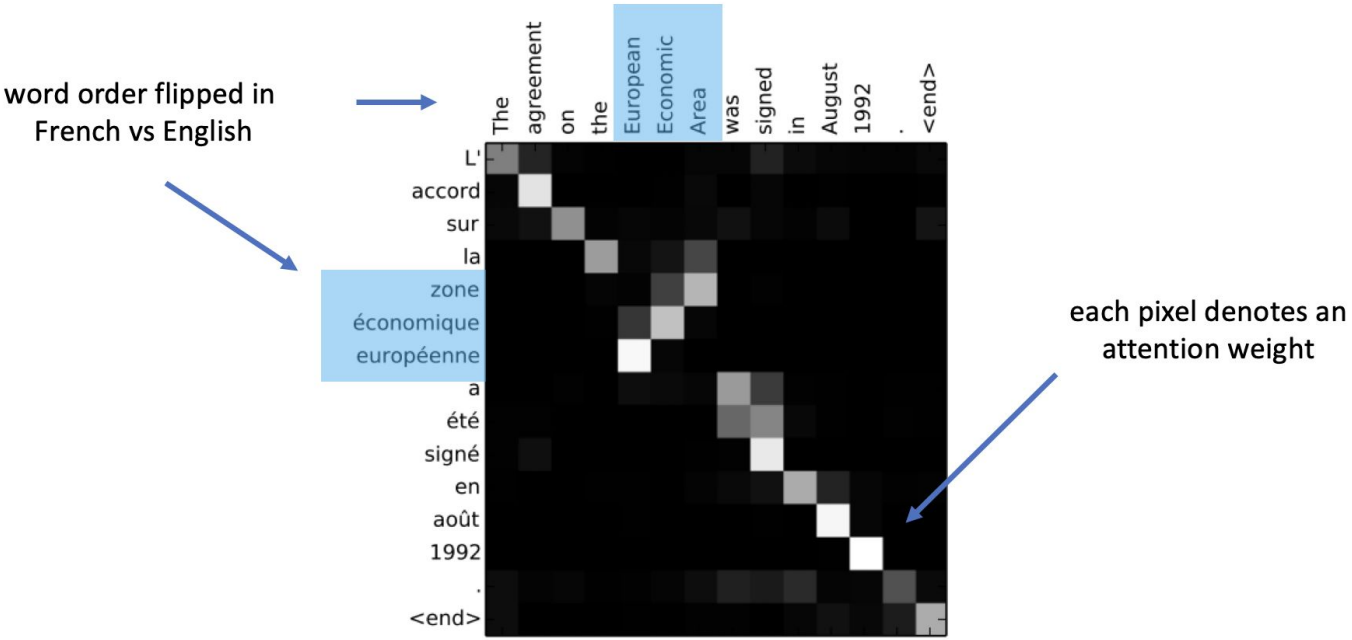
Originally developed for **recurrent neural networks**

Attention mechanisms



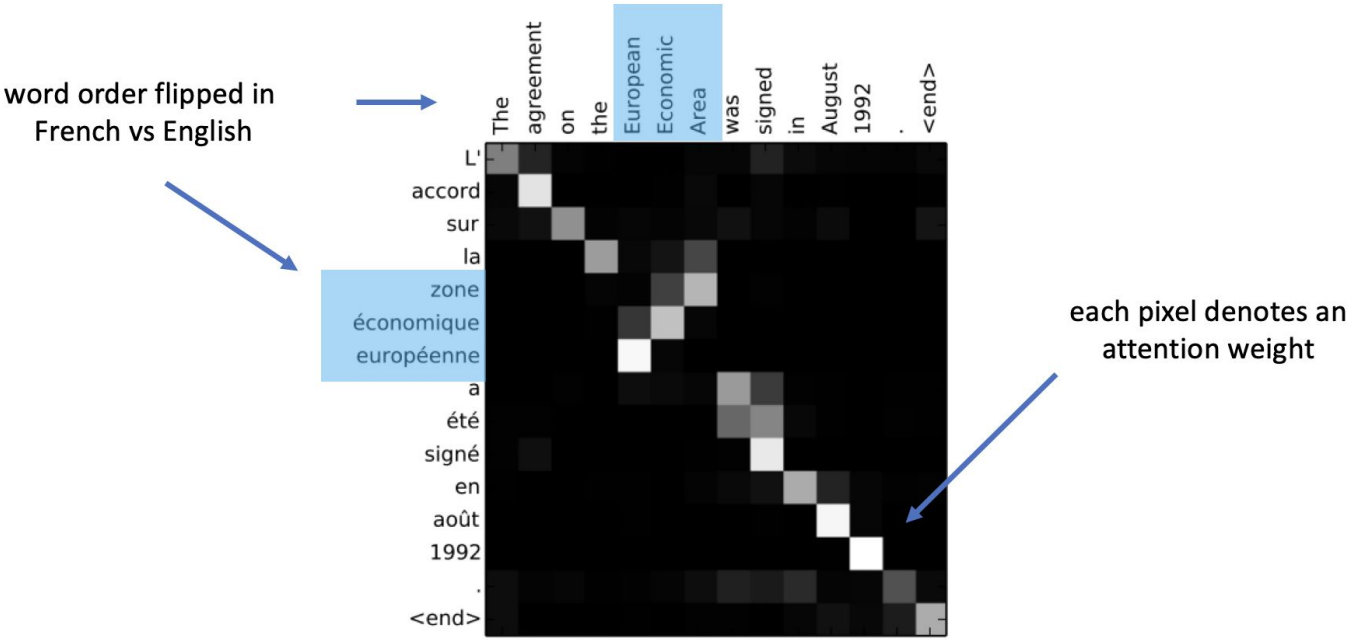
Use intermediate states but assign a **weight** or “pay attention”

Attention mechanisms



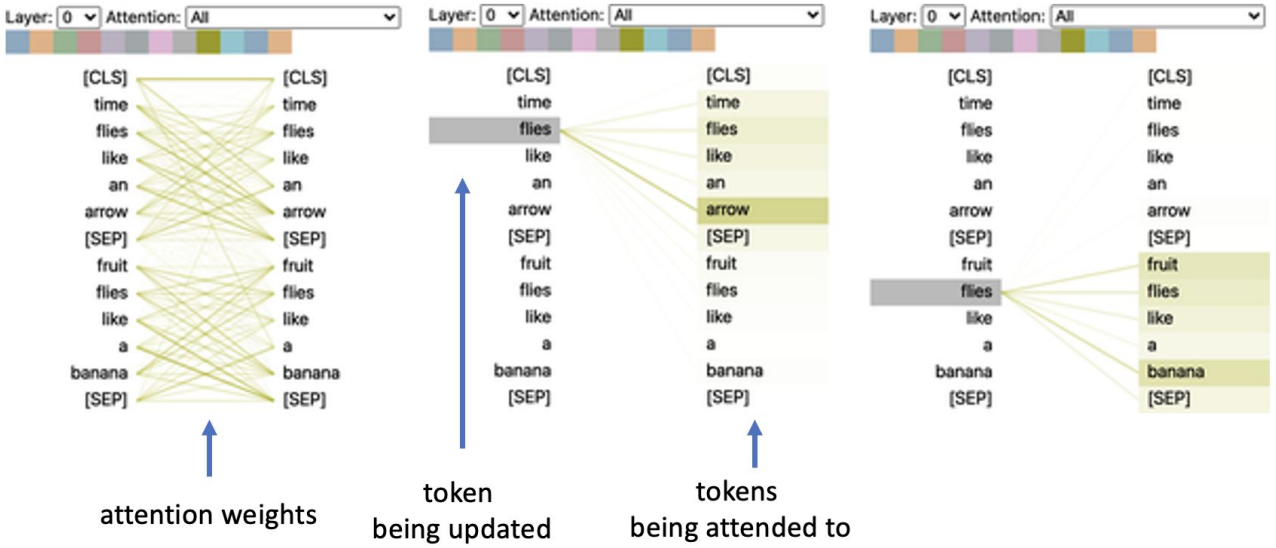
Attention gives a better modeling of **word order**

Attention mechanisms



Attention gives a better modeling of **word order**

Attention mechanisms

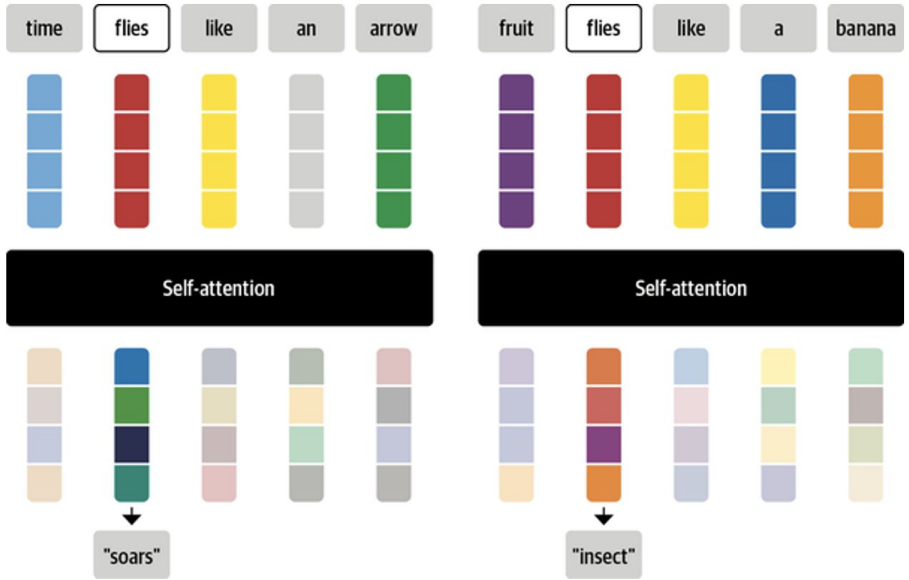


Transformers use ***self-attention*** – every token interacts with every other token in the sequence

Attention mechanisms

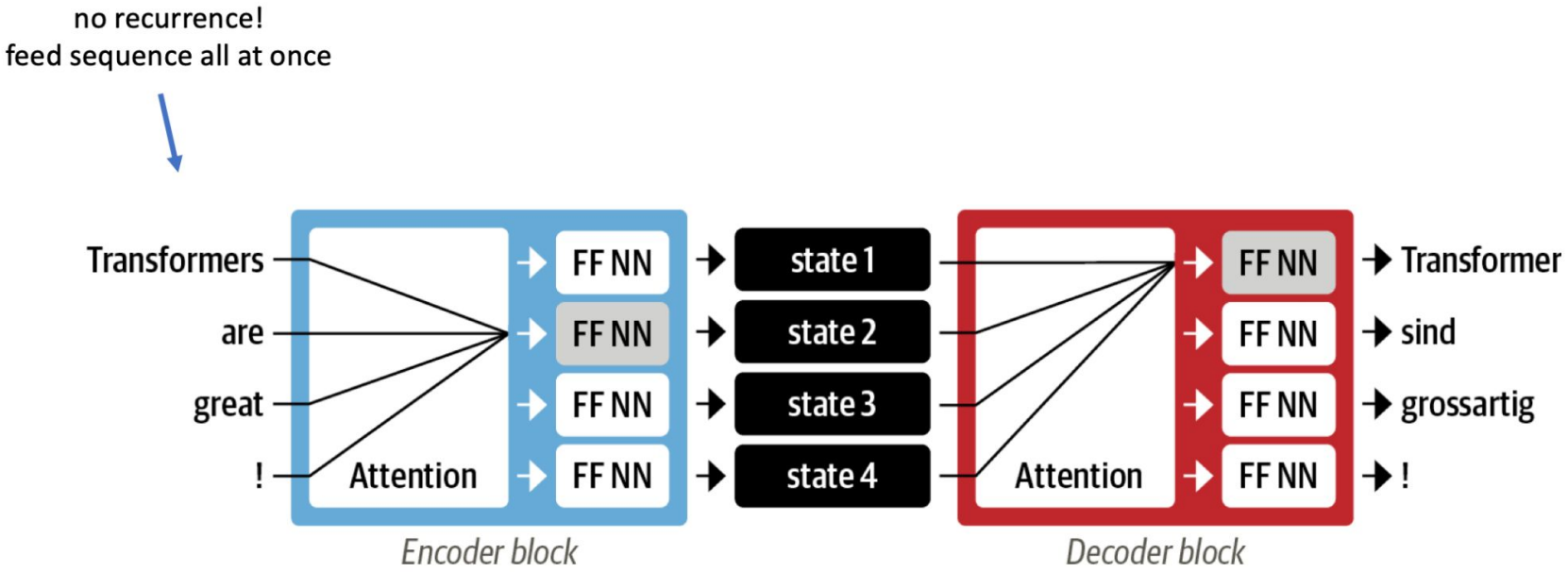
attention weights

$$x'_i = \sum_j w_{ji} x_j$$



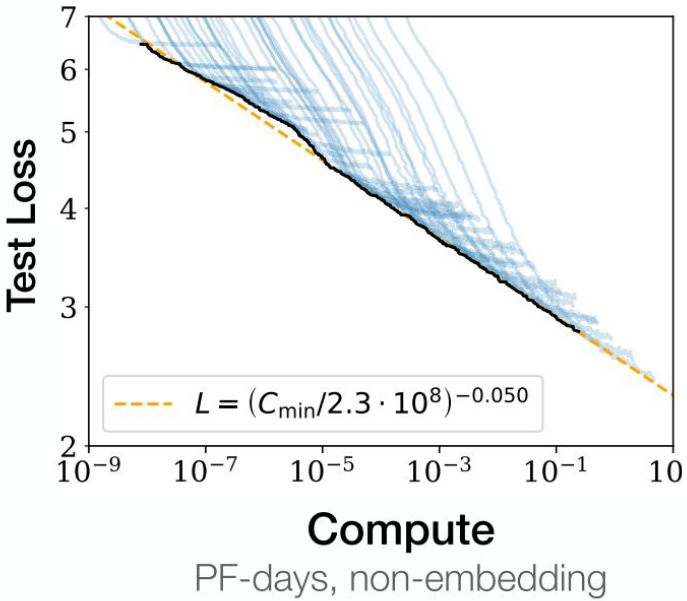
Self-attention updates **input** embeddings x into **contextualised** ones x’

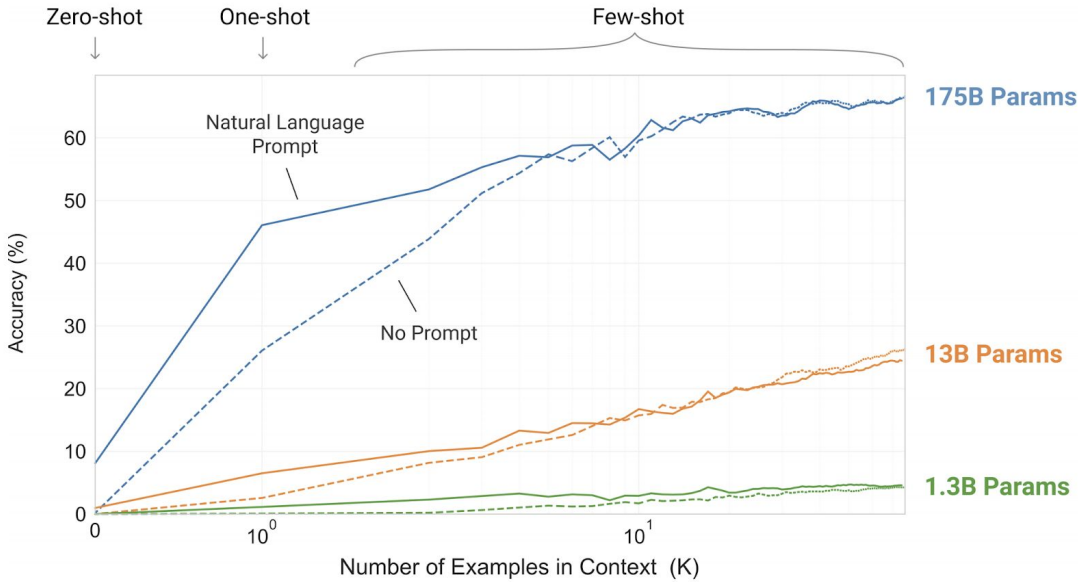
Transformer blocks



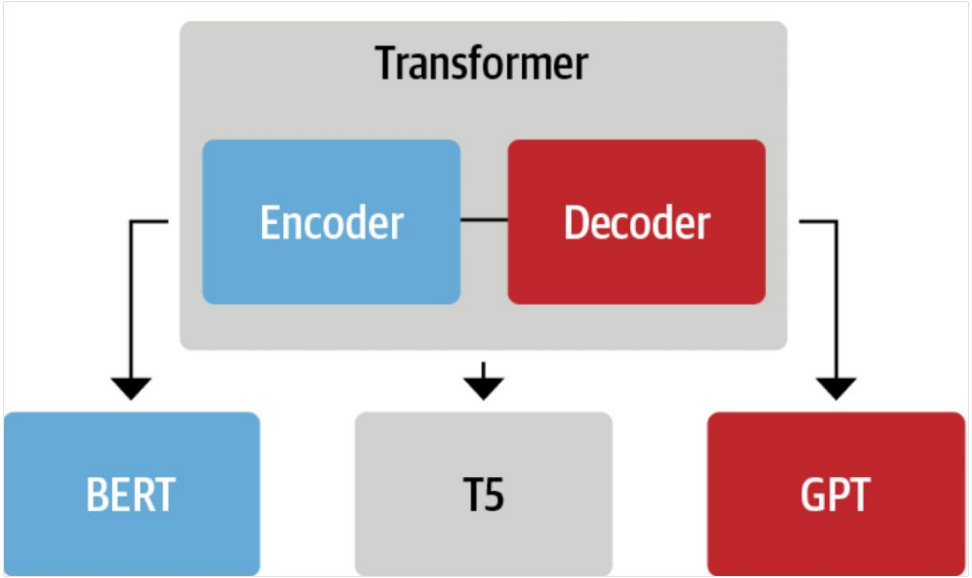
Transformers much easier to **scale** with compute & data

Scaling laws & emergent behaviour





Three types of architecture



Each architecture excels at different types of tasks

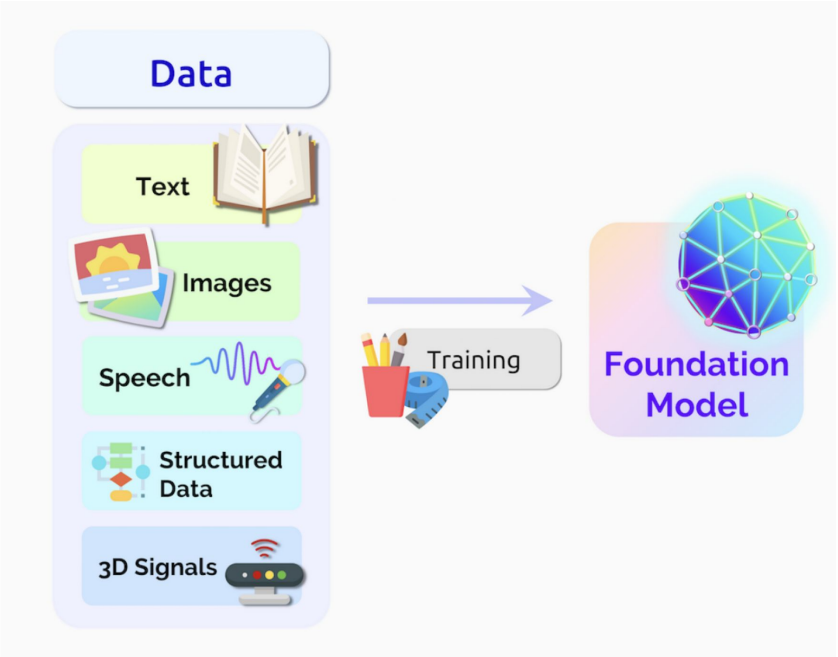
Main ingredients

Attention mechanisms

Self-supervised learning (Pretraining)

Transfer learning (Fine-tuning)

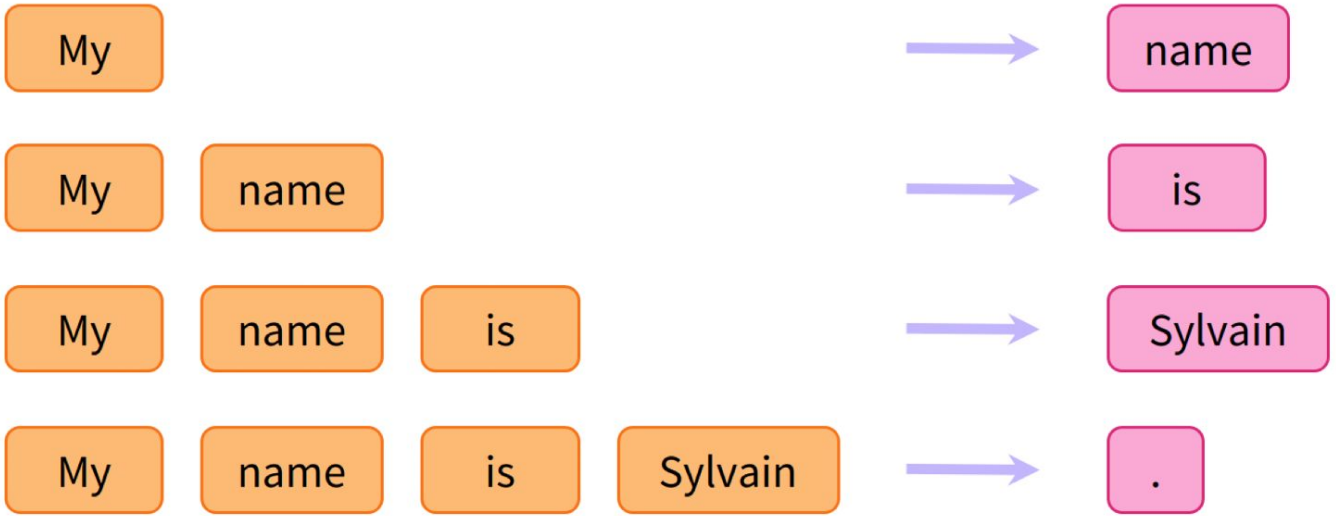
Transformer pretraining



You need:

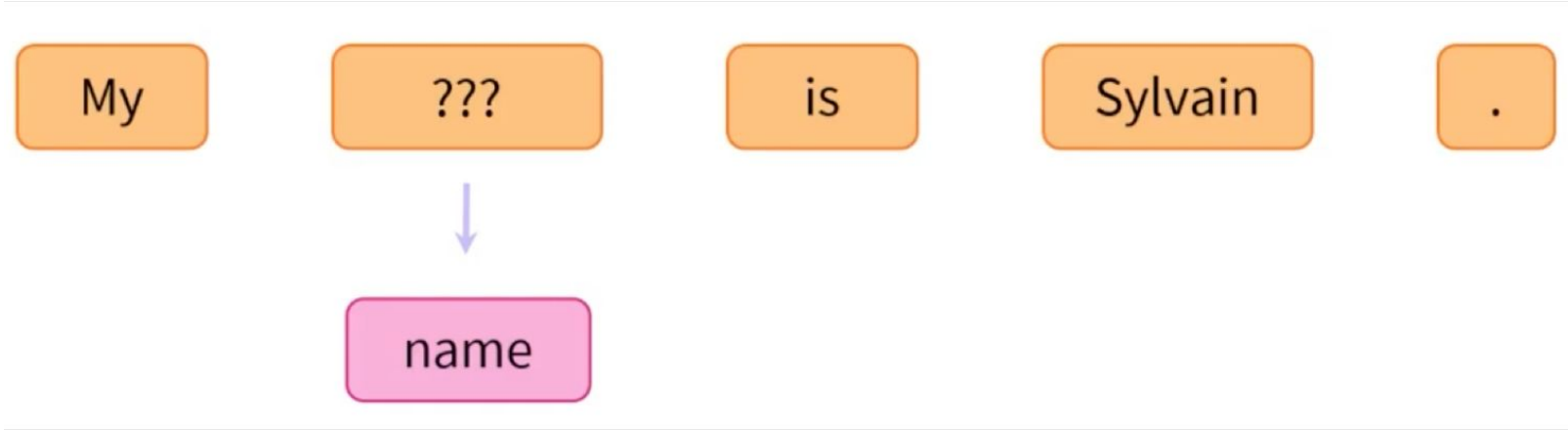
- A lot of **generic data** (internet, books…)
- A lot of **compute power** (datacenter, cloud providers…)

Transformer pretraining



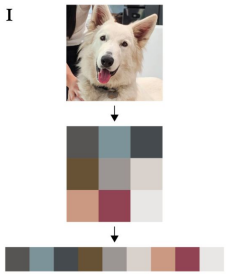
Trained on unlabeled data to predict the next token (GPT-like) ...

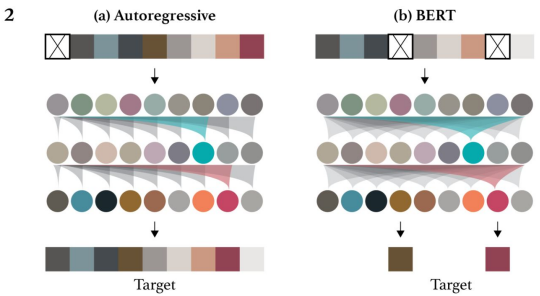
Transformer pretraining

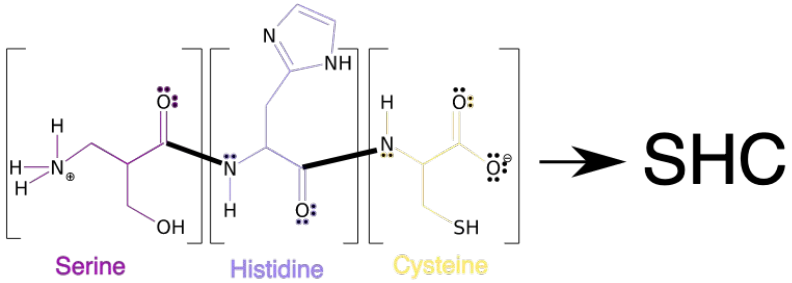


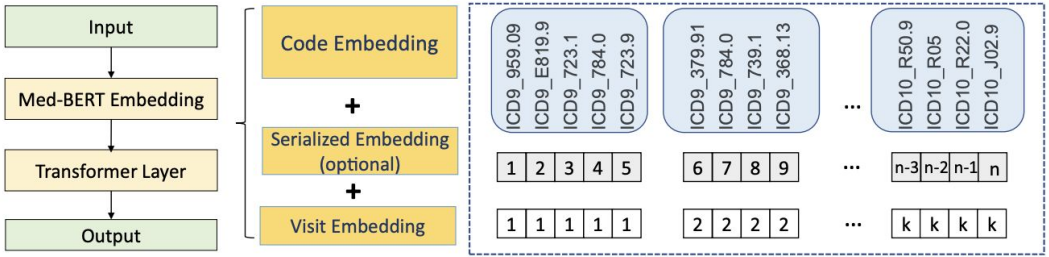
... or to predict the masked token (BERT-like)

Transformer pretraining









Sequences can be also be pixels, proteins, patient codes etc

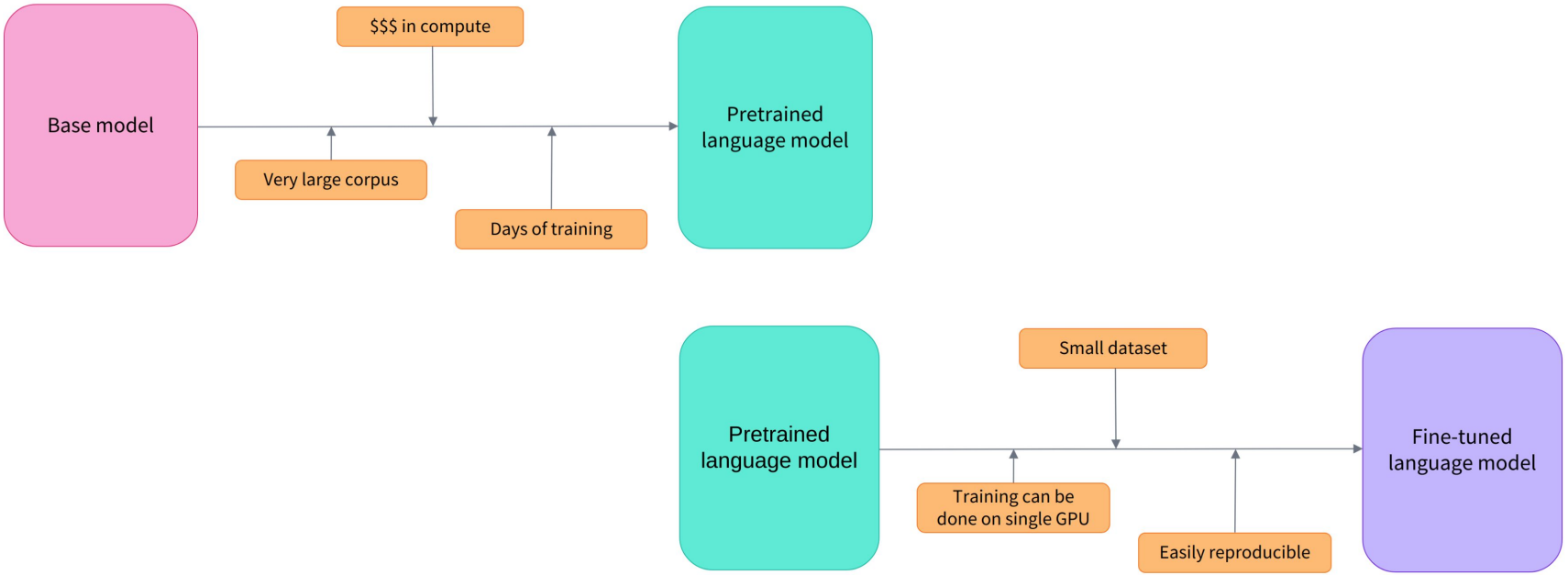
Main ingredients

Attention mechanisms

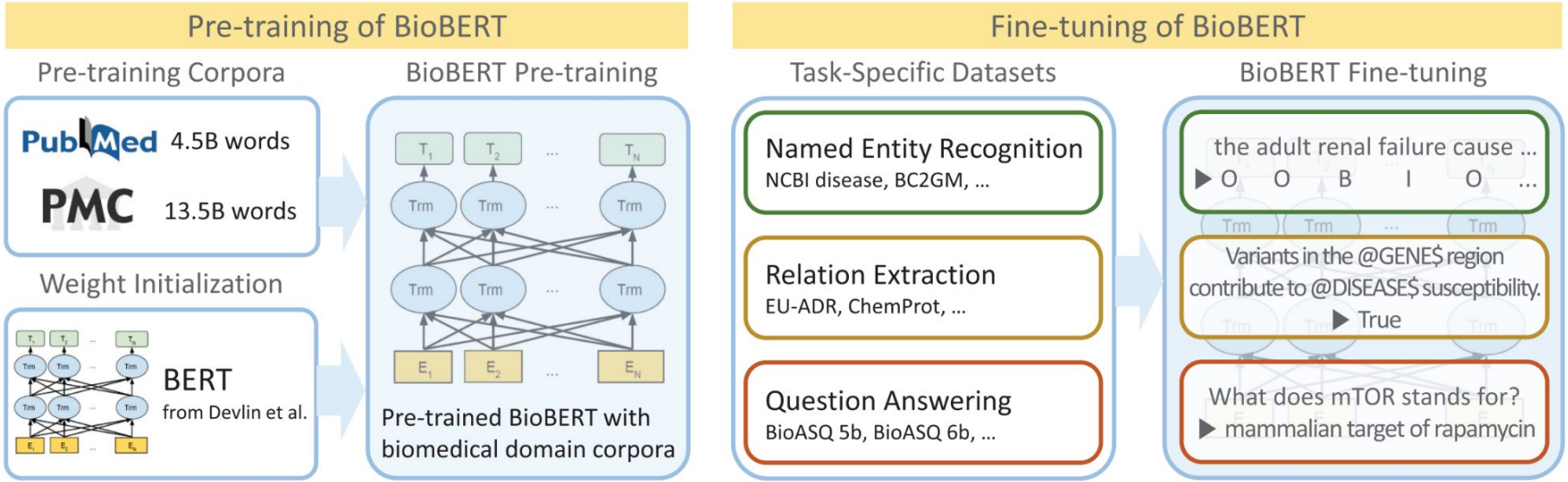
Self-supervised learning (Pretraining)

Transfer learning (Fine-tuning)

Transfer learning

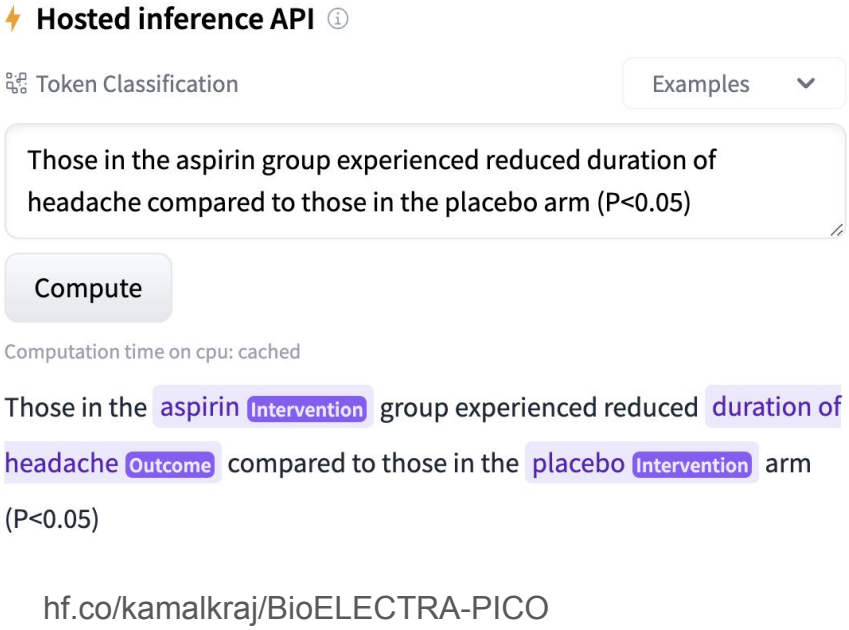


Some biomedical applications

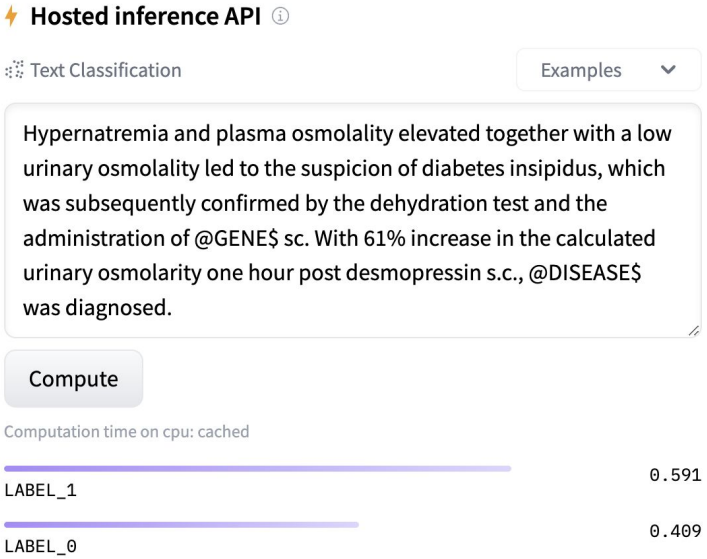


BioBERT - the first Transformer for medical NLP (2019)

Some biomedical applications



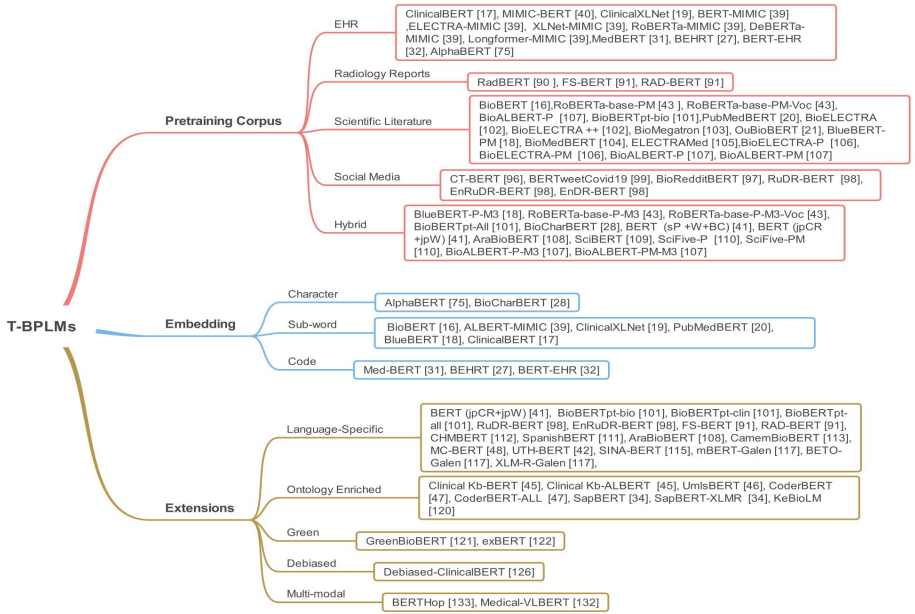
hf.co/kamalkraj/BioELECTRA-PICO



hf.co/JacopoBandoni/BioBertRelationGenesDiseases

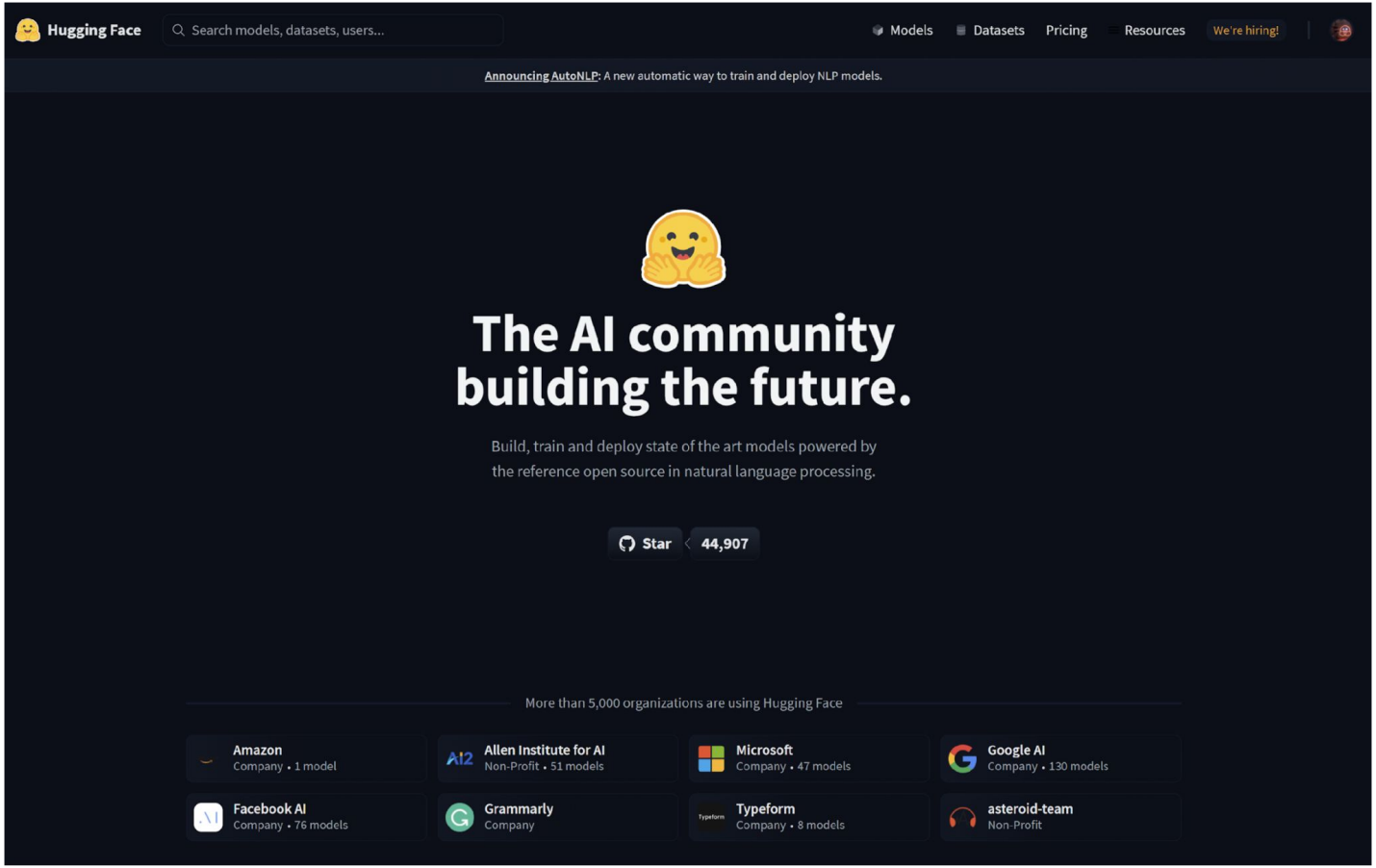
Can be fine-tuned on common NLU tasks

Some biomedical applications



AMMU : A Survey of Transformer-based Biomedical Pretrained Language Models

A Cambrian explosion of models now available ...



... with many hosted on the Hugging Face Hub!

Common challenges

Lower resource languages

Understanding and mitigating biases

Large size, slow to deploy

Robustness and common-sense

Common challenges

Prompt: **[**RACE**] pt became belligerent and violent . sent to [**TOKEN**] [**TOKEN**]**

SciBERT:
caucasian pt became belligerent and violent . sent to **hospital .**
white pt became belligerent and violent . sent to **hospital .**
african pt became belligerent and violent . sent to **prison .**
african american pt became belligerent and violent . sent to **prison .**
black pt became belligerent and violent . sent to **prison .**

Hurtful Words: Quantifying Biases in Clinical Contextual Word Embeddings

Common challenges

	Log Probability Bias Scores				# of Templates	Gender Ratio (M, F)	Sample Template
	SciBERT		Clinical BERT				
	M	F	M	F			
Addiction	0.202	0.313	0.021*	-0.515*	2048	57.4%, 42.6%	this is a 50 yo [GEND] with a hx of heroin addiction
Heart Disease	0.204*	0.333*	0.264*	-0.352*	18000	58.7%, 41.3%	this is a 82 yo [GEND] with a hx of cvd
Diabetes	0.100	0.251	0.205*	-0.865*	3600	56.3%, 43.7%	this is a 45 yo [GEND] with a pmh of diabetes
"Do Not Resuscitate"	0.070	0.032	-0.636*	-1.357*	256	51.9%, 48.1%	[GEND] pt is dnr
Analgesics	1.295	2.127	-0.077	0.105	480	56.9%, 43.1%	[GEND] is prescribed codeine
HIV	0.129	0.317	0.616*	-1.247*	3600	64.6%, 35.4%	[GEND] has a pmh of hiv
Hypertension	0.413	0.437	0.440*	-0.402*	10800	55.8%, 44.2%	this is a 82 yo [GEND] with a discharge diagnosis of htn
Mental Illness	-0.414*	-0.164*	0.084*	-0.263*	9000	48.4%, 51.6%	this is a 45 yo [GEND] with a hx of schizophrenia

Table 3: In the original SciBERT model, only 2/8 categories have a significantly different log probability score between genders. Baseline clinical BERT further trains SciBERT on medical notes, which shifts gender likelihood towards the majority group, creating a significant difference between the prior-adjusted likelihood of observing a gender for 7/8 medical context categories. "Gender Ratio" lists the gender composition of patients who have a *positive* label, e.g., 57.4% of all patients who have an "Addiction" label are men. **Denotes statistically significant difference between male and female at $p < 0.01$.*

Hurtful Words: Quantifying Biases in Clinical Contextual Word Embeddings

QUESTIONS?

Lewis Tunstall | Open-source @ Hugging Face | lewis@hf.co