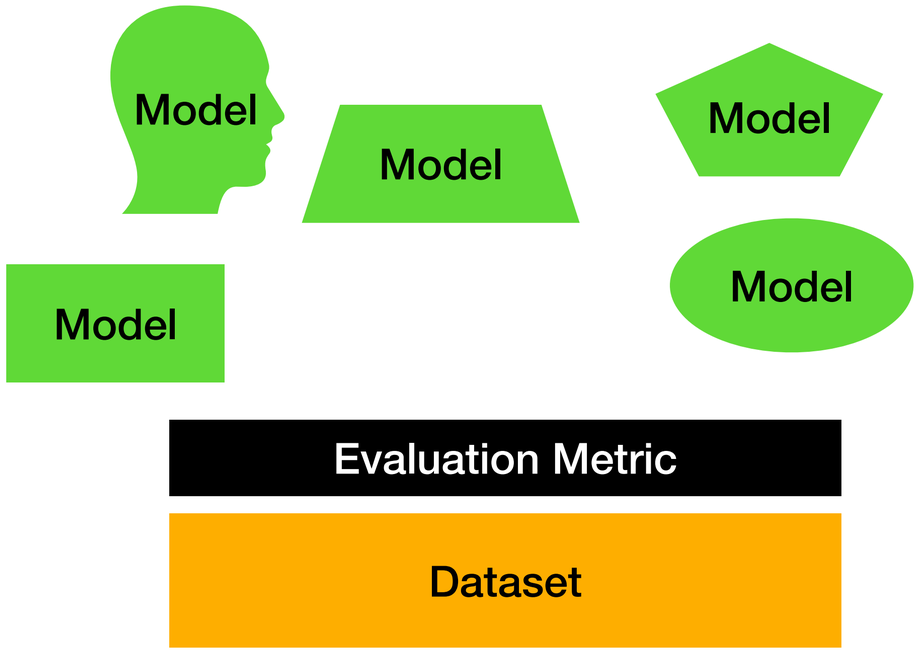
Context for Interpreting Benchmark Performances

ACL-2021 Workshop on Benchmarking: Past, Present and Future (BPPF)

Eunsol Choi
The University of Texas at Austin

Putting benchmark results into context

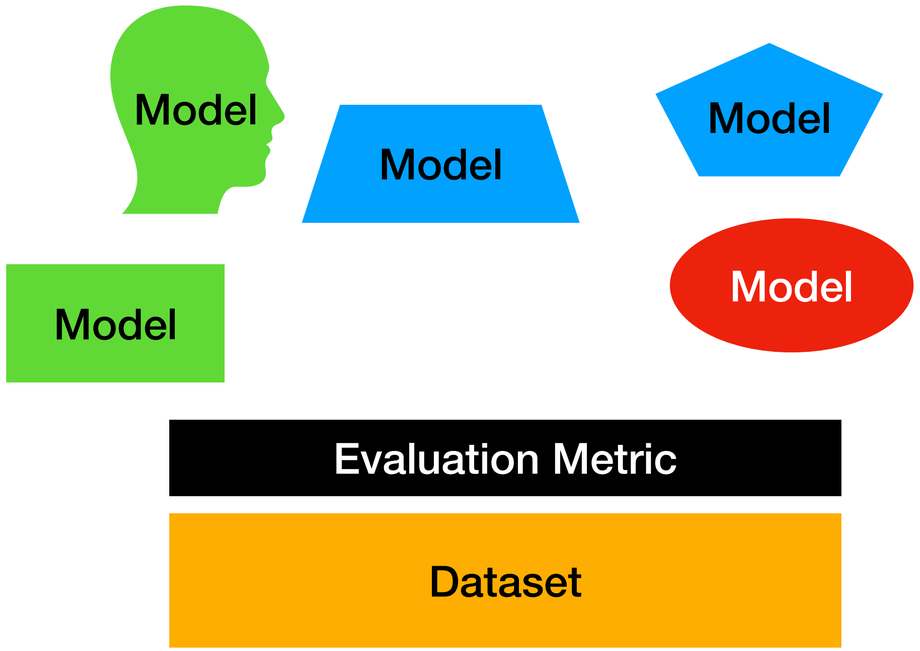


$f : x \rightarrow y$

$\text{score}(y, y^{pred})$

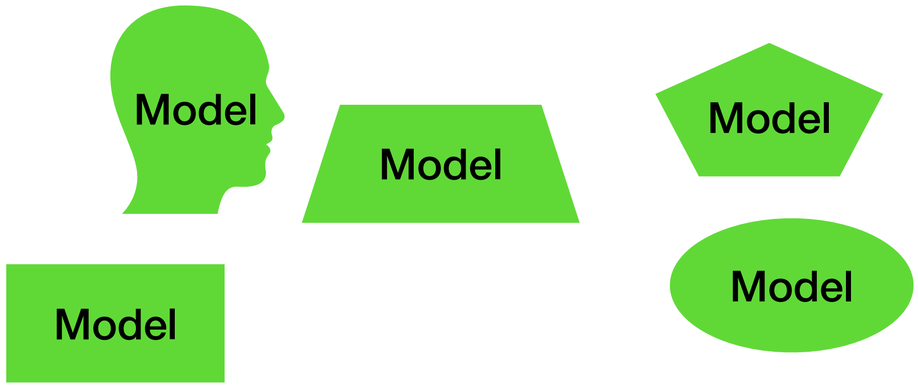
$\{(x, y)\}$

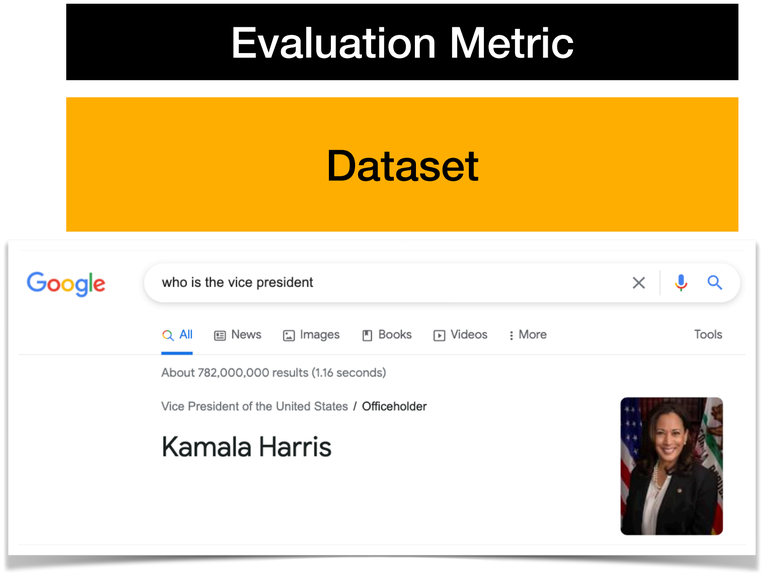
Putting benchmark results into context



- Contexts for **models**
  - How to make a **fair and useful** comparison across different models?

Putting benchmark results into context





- Contexts for **models**
  - How to make a **fair and useful** comparison across different models?

- Contexts for benchmark **data**
  - What are the **implicit** extra-linguistic contexts (temporal, geographical, cultural) of our datasets?

NLP Benchmarks in 2021

General Benchmarks

GLUE — Language Understanding

XTREME — Multilingual Language Understanding

.
.
.

Specialized Benchmarks

WMT — Machine Translation

Multihop reasoning

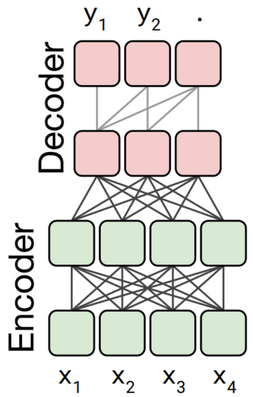
Winograde — Commonsense Reasoning

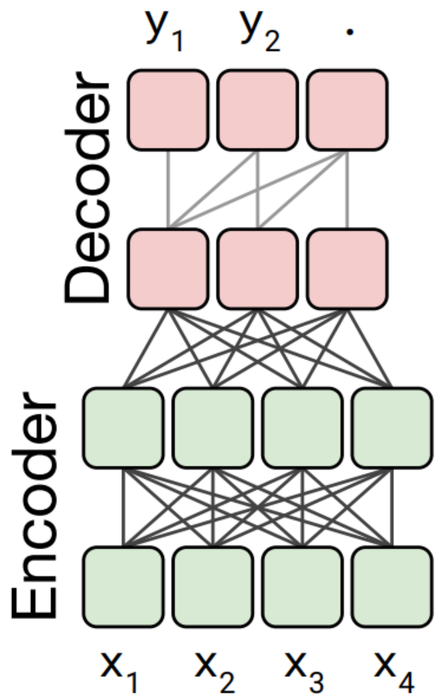
DROP — Numeric reasoning

Natural Questions — Information seeking questions

.
.
.

Which models are successful in leaderboard?





Larger pre-trained models, trained longer on larger amount of data, perform consistently better on multiple benchmarks

[Raffel et al, 2020, Liu et al, 2019]

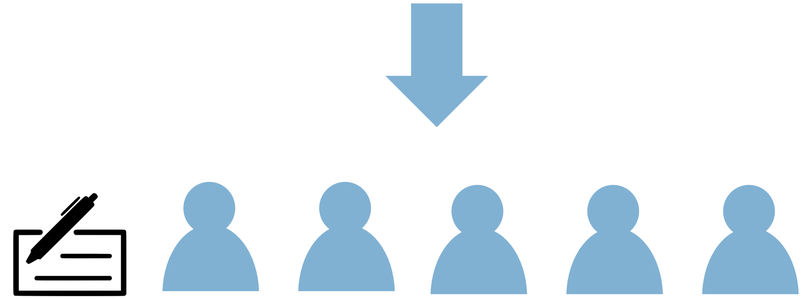
Goal: Compare models using similar amount of resources

- Under the fixed resource constraints, what would be the best model achieving highest accuracy on QA task?
- Efficient QA NeurIPS competition 2020

Sewon Min, Adam Roberts, Chris Alberti, Colin Raffel, Danqi Chen, Eunsol Choi, Hannaneh Hajishirzi, Jennimaria Palomaki, Jordan Boyd-Graber, Kelvin Guu, Kenton Lee, Michael Collins, Tom Kwiatkowski
NeurIPS 2020 competition https://efficientqa.github.io/

Google AI

Task: Open-domain question answering



Q: which country is located north of the Belgium?
A: [“The Netherlands”, “Netherlands”]

How should we measure efficiency?

- Speed
  - Training time
  - Inference time

- Memory
  - How much storage is needed to store your model?

How should we measure efficiency?

- Speed
  - Training time
  - Inference time

- Memory
  - How much storage is needed to store your model?

- None of these measures is straightforward…

Efficient QA competition: memory constraints

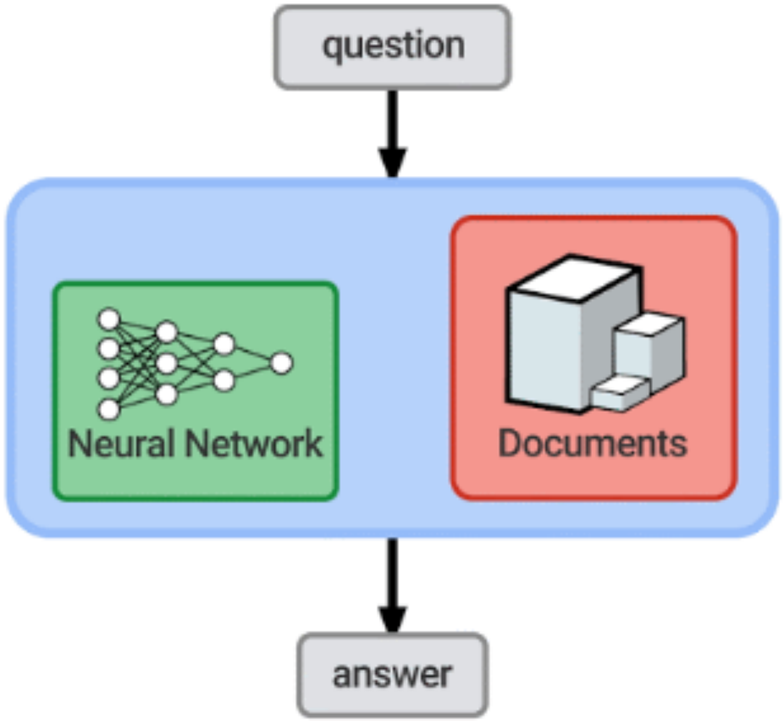
The number of bytes required to store a Docker image that contains the complete, self-contained QA system. This Docker image must include all code, libraries, parameters, and data required by your system.

4 tracks:

- Unrestricted
- Under 6GB
- Under 500MB
- Smallest system getting 25% accuracy

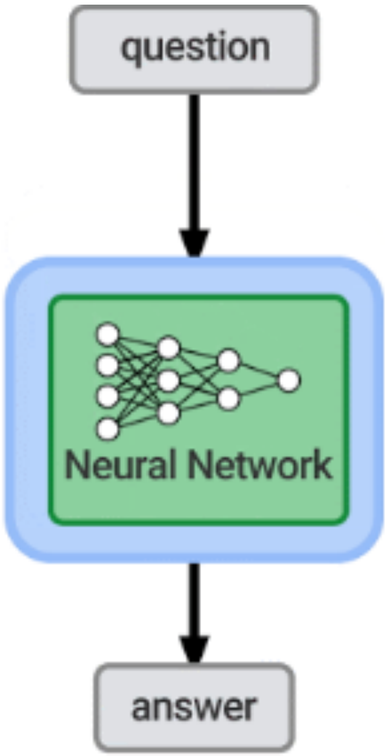
Two architectures for open domain QA

Retrieval based model



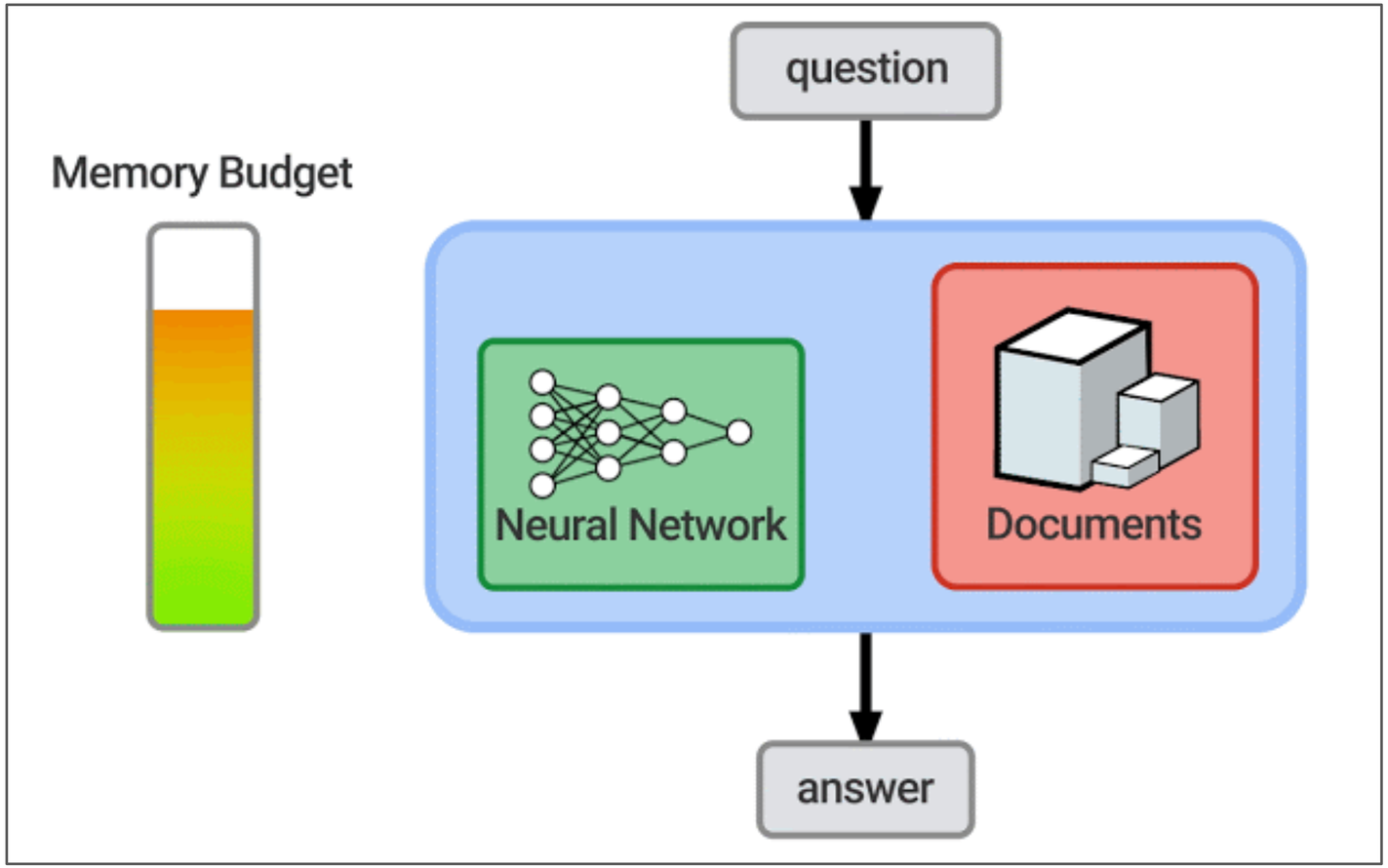
[Chen et al 2017, Lee et al 2019]

Closed book model



[Roberts et al 2021, Fevry et al 2020]

Comparing different architecture under the same memory budget

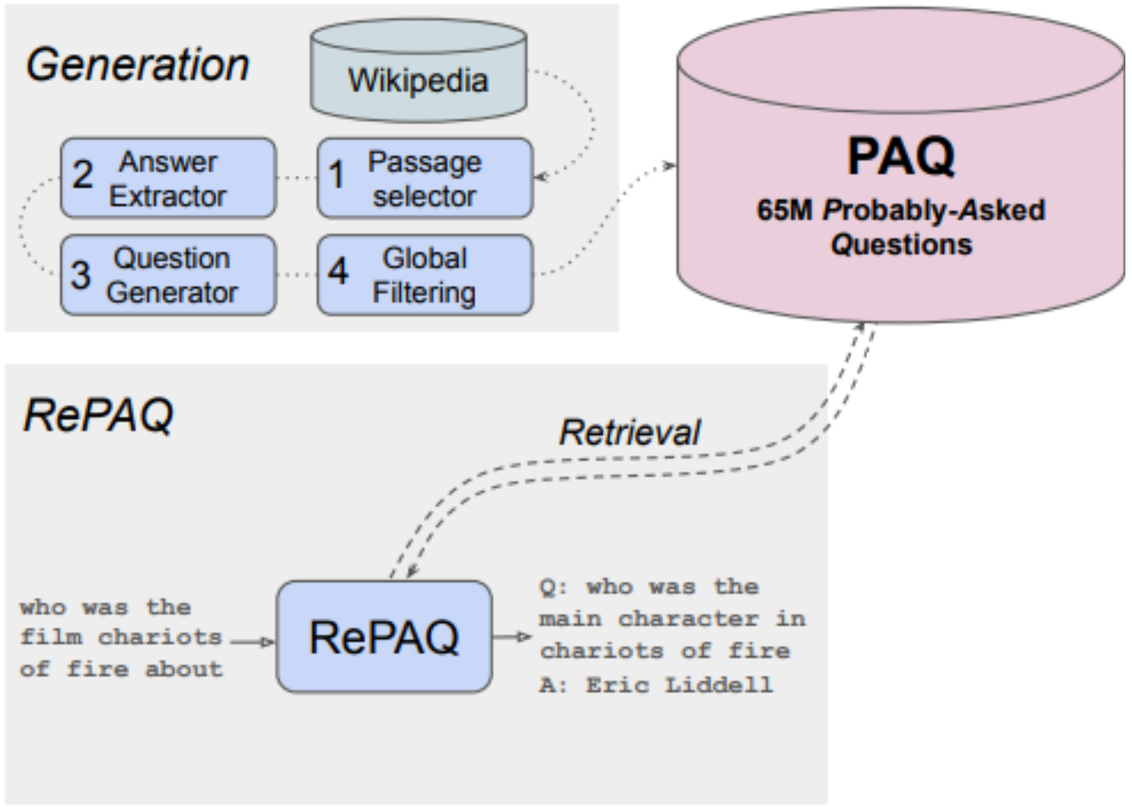


Results of the Efficient QA competition

- 39 submissions from 18 unique teams.
- Many interesting approaches to make models more memory efficient.
- Please take a look at our report!

Track	Team	Video	Accuracy (auto)	Accuracy (human)
Unrestricted	Microsoft Research & Dynamics 365	link	54.00	65.80
Unrestricted	Facebook AI	link	53.89	67.38
6GB	FAIR-Paris&London	link	53.33	65.18
6GB	Studio Ousia & Tohoku University	link	50.17	62.01
6GB	Brno University of Technology	link	47.28	58.96
500MB	UCL+Facebook AI	link	33.44	39.40
500MB	Naver Clova	link	32.06	42.23
25% smallest	UCL+Facebook AI	link	26.78	32.45

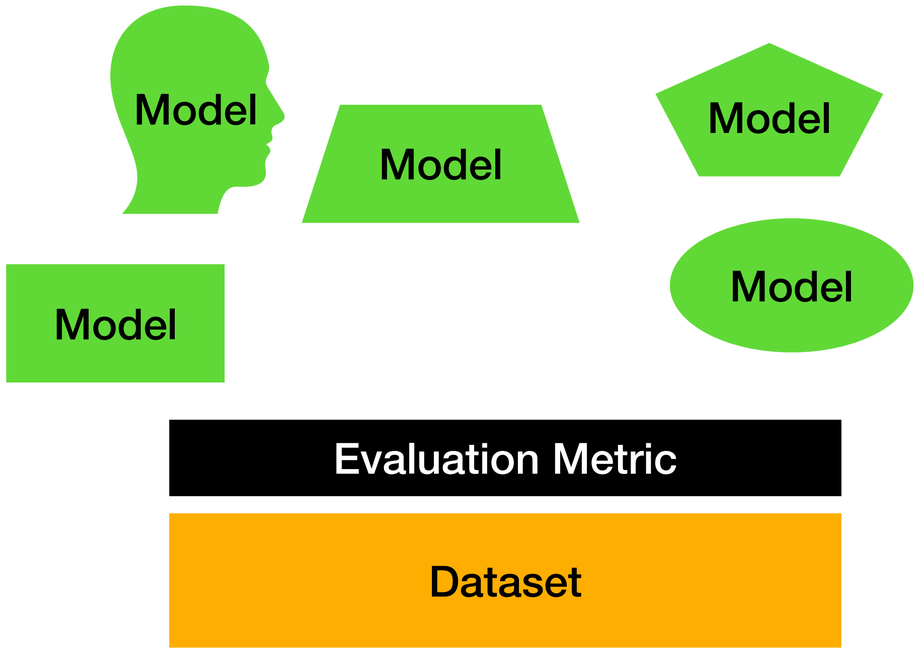
Memory constraint motivates new modeling approaches



Instead of using document corpus, they use generated [question, answer] pairs as a knowledge source

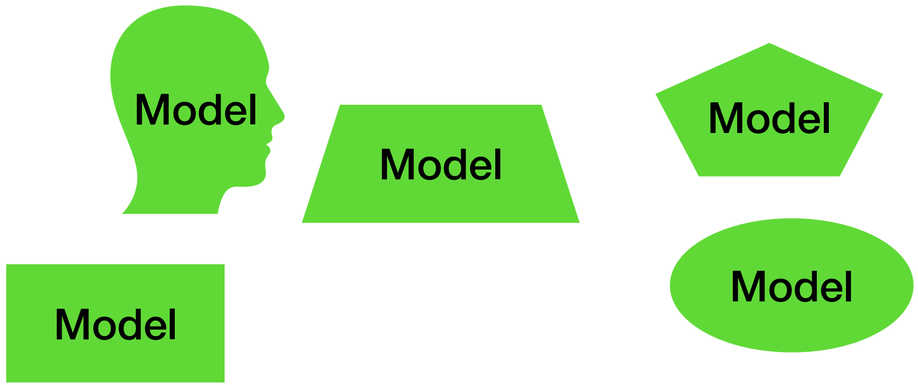
[Lewis et al, ArXiv 2021] https://arxiv.org/pdf/2102.07033.pdf

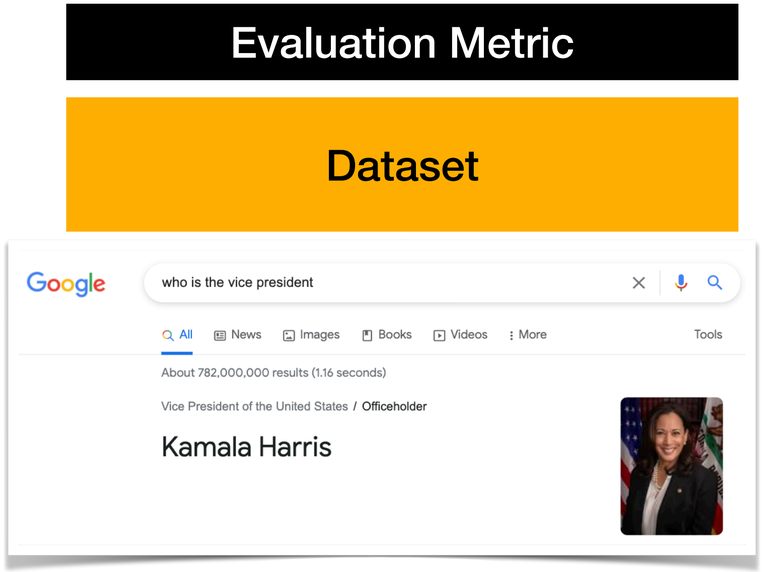
Putting benchmark results into context



- Contexts for **models**
  - How to make a **fair and useful** comparison across different models?
- Contexts for benchmark **data**
  - What are the ***implicit*** extra-linguistic contexts (temporal, geographical, cultural) of our datasets?

Putting benchmark results into context

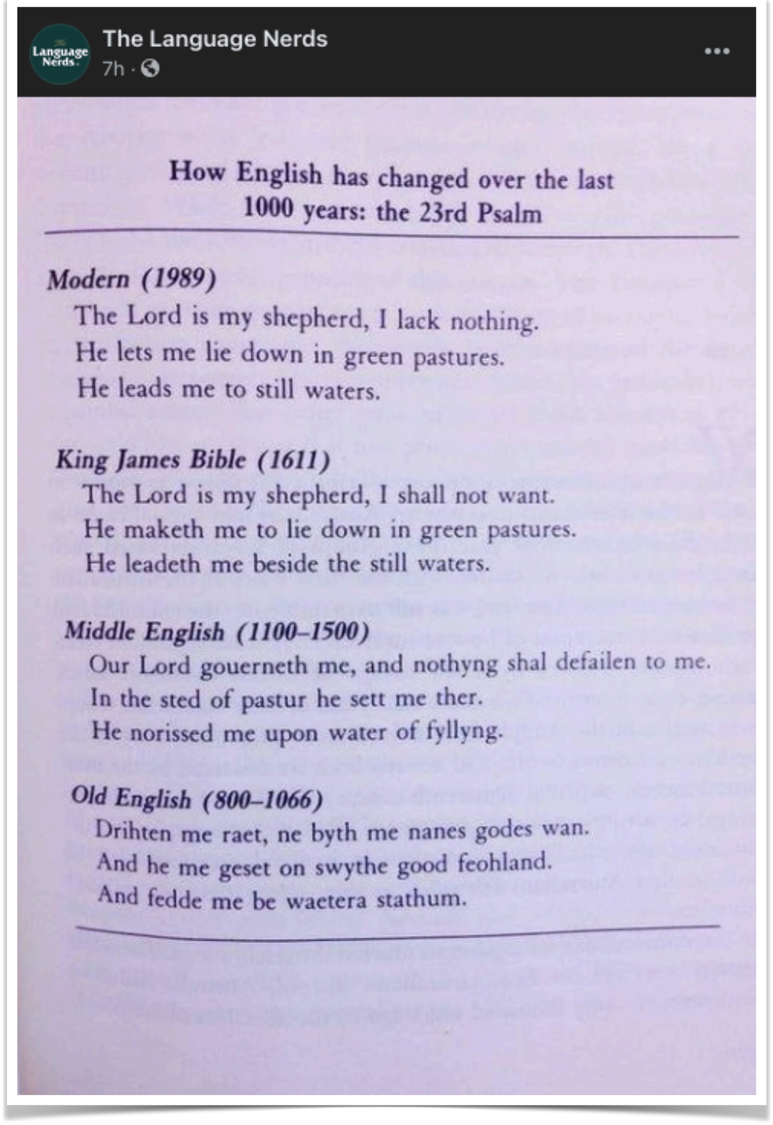




- Contexts for **models**
  - How to make a **fair and useful** comparison across different models?

- Contexts for benchmark **data**
  - What are the ***implicit*** extra-linguistic contexts (temporal, geographical, cultural) of our datasets?

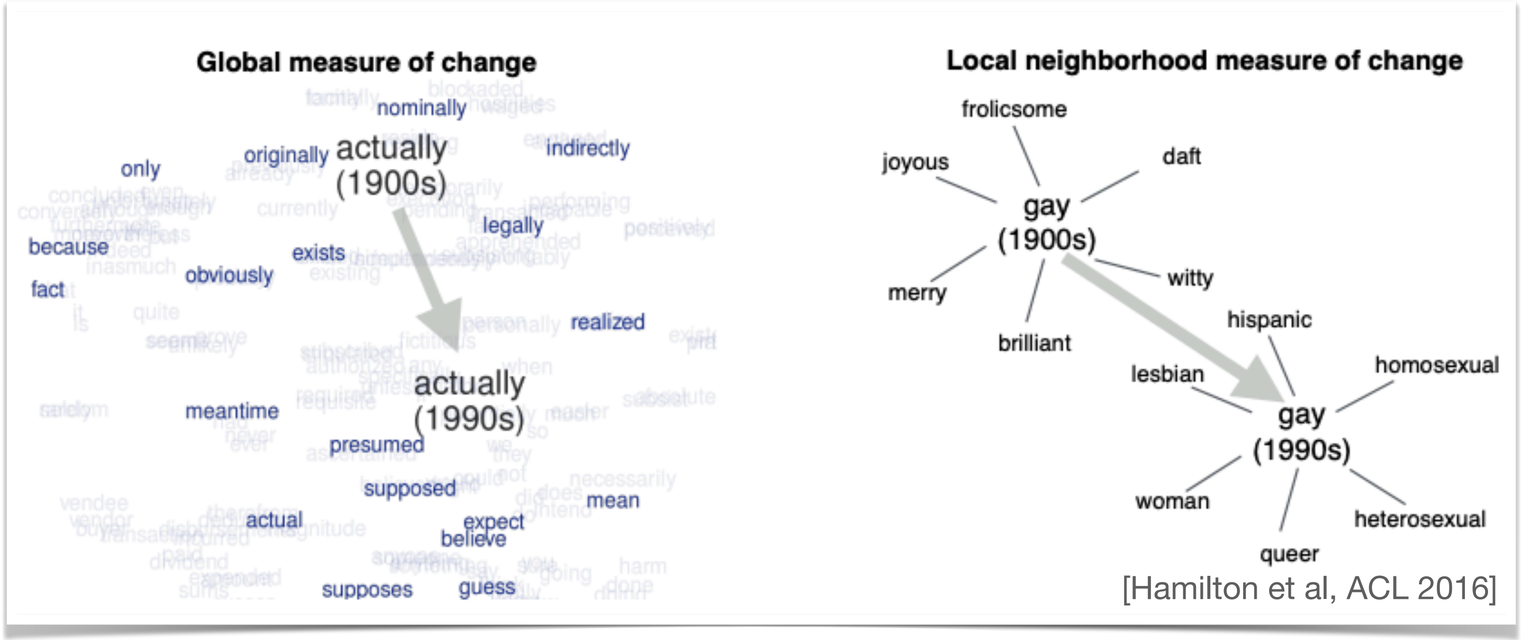
Language evolves



When was the first time Google was used as a verb?

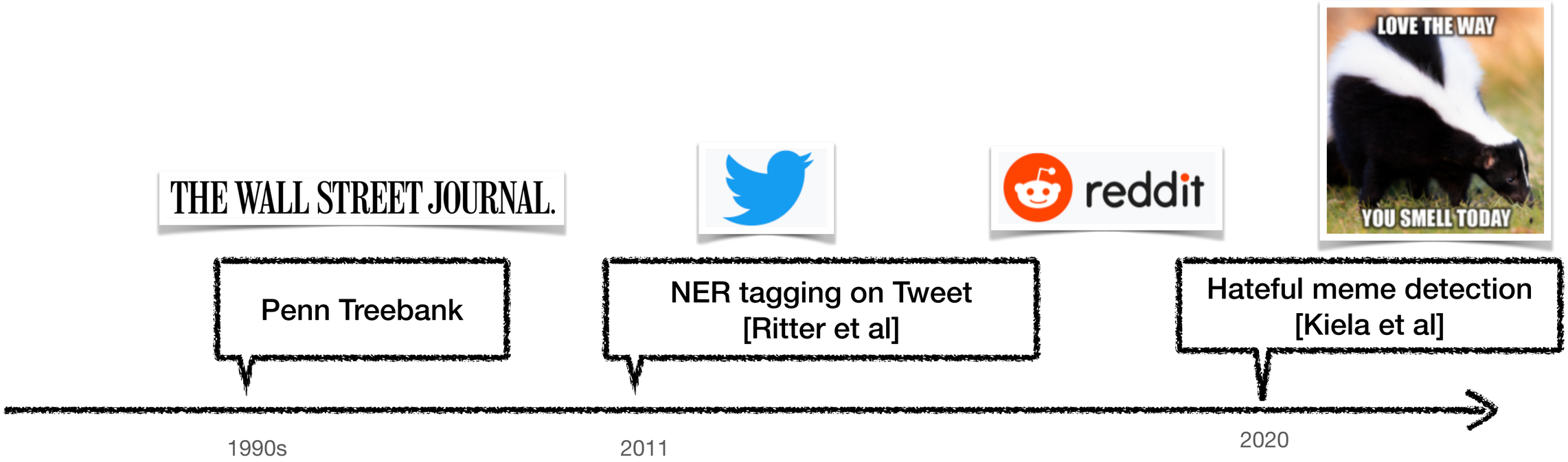
Is Google officially a verb?

Is Google a noun or verb?



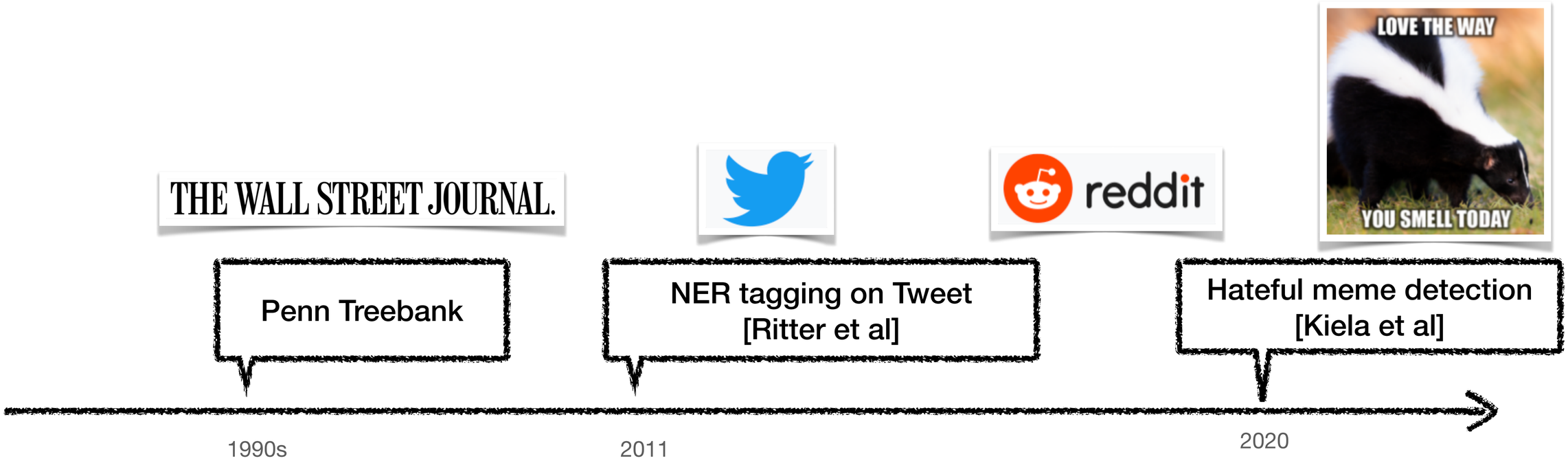
[Hamilton et al, ACL 2016]

New benchmarks reflect the changes in language usages



Datasets are supposed to be a representation of the world [Torralba and Efros, CVPR 2011]

New benchmarks reflect the changes in language usages

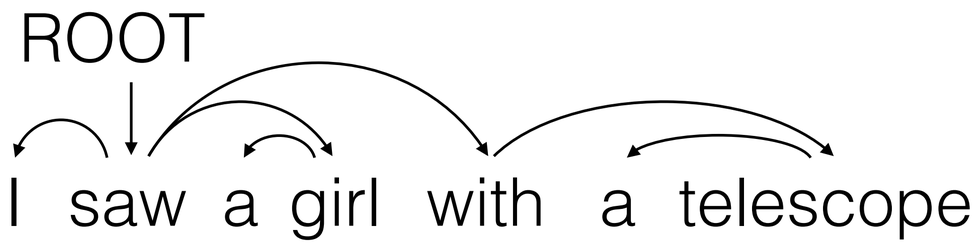


Datasets are supposed to be a representation of the world [Torralba and Efros, CVPR 2011]

- Each dataset makes implicit assumptions about the ***world (extra-linguistic contexts)*** it represents.

How benchmark is changing

Linguistic Knowledge



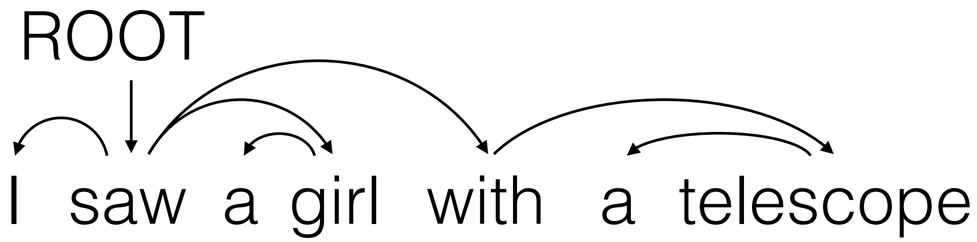
a framework for consistent annotation of ***grammar*** *(parts of speech, morphological features, and syntactic dependencies)*

Past **Today**

[Universal Dependency, de Marneffe et al, 2014]

How benchmark is changing

Linguistic Knowledge



*a framework for consistent annotation of **grammar** (parts of speech, morphological features, and syntactic dependencies)*

World Knowledge

The theory of relativity was developed by [**MASK**].

*LAMA is a probe for analyzing the factual and commonsense **knowledge** contained in pretrained language models.*

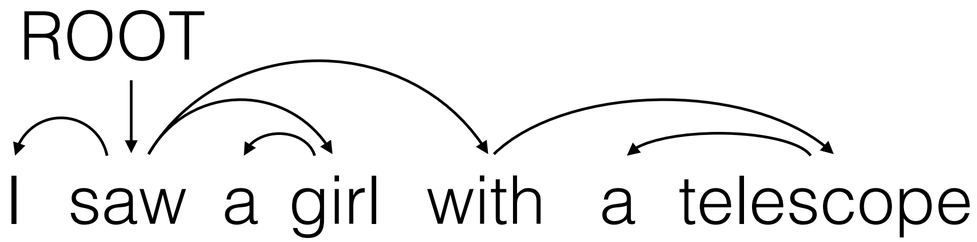
Past → **Today**

[Universal Dependency, de Marneffe et al, 2014]

[Petroni et al, EMNLP 2019]

How benchmark is changing

Linguistic Knowledge



*a framework for consistent annotation of **grammar** (parts of speech, morphological features, and syntactic dependencies)*

World Knowledge

The theory of relativity was developed by [**Einstein**].

*LAMA is a probe for analyzing the factual and commonsense **knowledge** contained in pretrained language models.*

Past → **Today**

[Universal Dependency, de Marneffe et al, 2014]

[Petroni et al, EMNLP 2019]

How QA benchmark is changing

Question : What shift happened in animal regulation in 1963 in U.S?

Document Context :

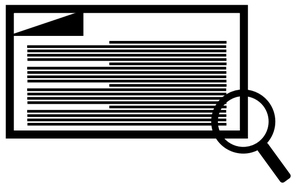
The Lacey Act of 1900 was the first federal law that regulated commercial animal markets. It prohibited interstate commerce of animals killed in violation of state game laws, and covered all wildlife. Whereas the Lacey Act dealt with game animal management and market commerce species, a major shift in focus occurred by 1963 to habitat preservation instead of take regulations. A provision was added by Congress in the Land and Water Conservation Fund Act of…

Answer is span in the provided document

Question : When did Joe Biden graduate from college?

WIKIPEDIA
The Free Encyclopedia

1. **Article retrieval**
Find top Wikipedia page in the same language by Google Search



2. **Answer annotation**
Crowdworker finds answer and evidence paragraph

Answer: 1965 (string)

Past → **Today**

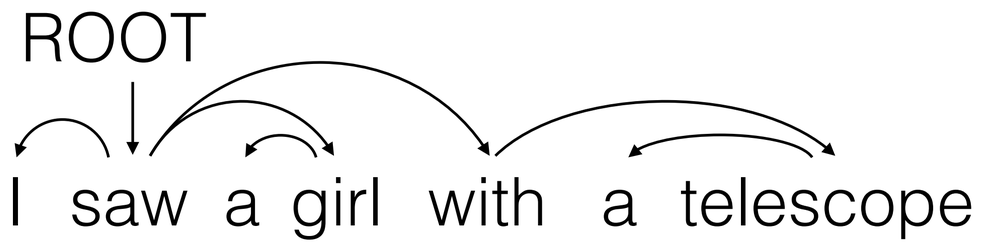
[Rajpurkar et al 2016]

[Asai et al 2021]

Temporal dependence of benchmarks

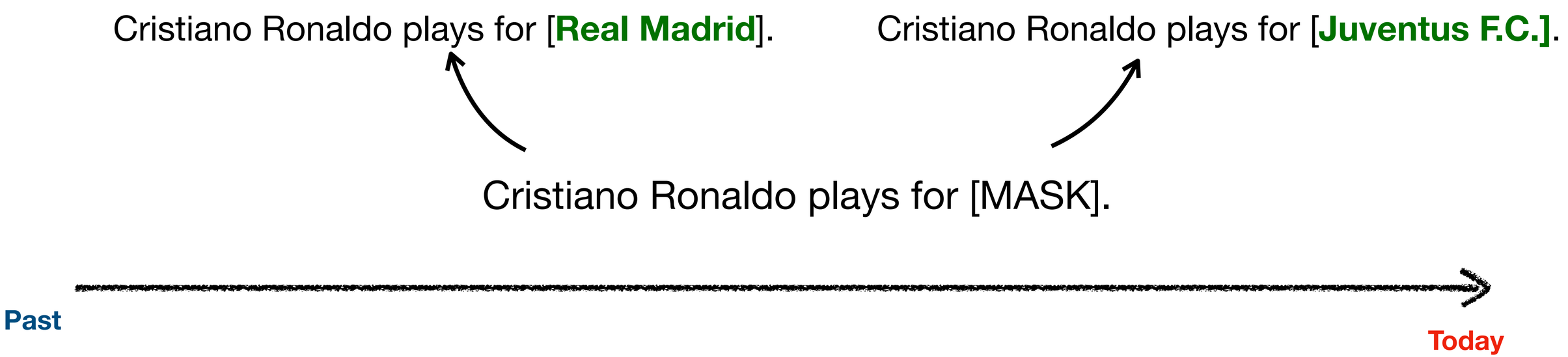
Sentiment Analysis

Linguistic Structure Analysis



Past → **Today**

Temporal dependence of benchmarks



Time-Aware Language Models as Temporal Knowledge Bases
[Dhingra et al, ArXiv 2021] https://arxiv.org/pdf/2106.15110.pdf

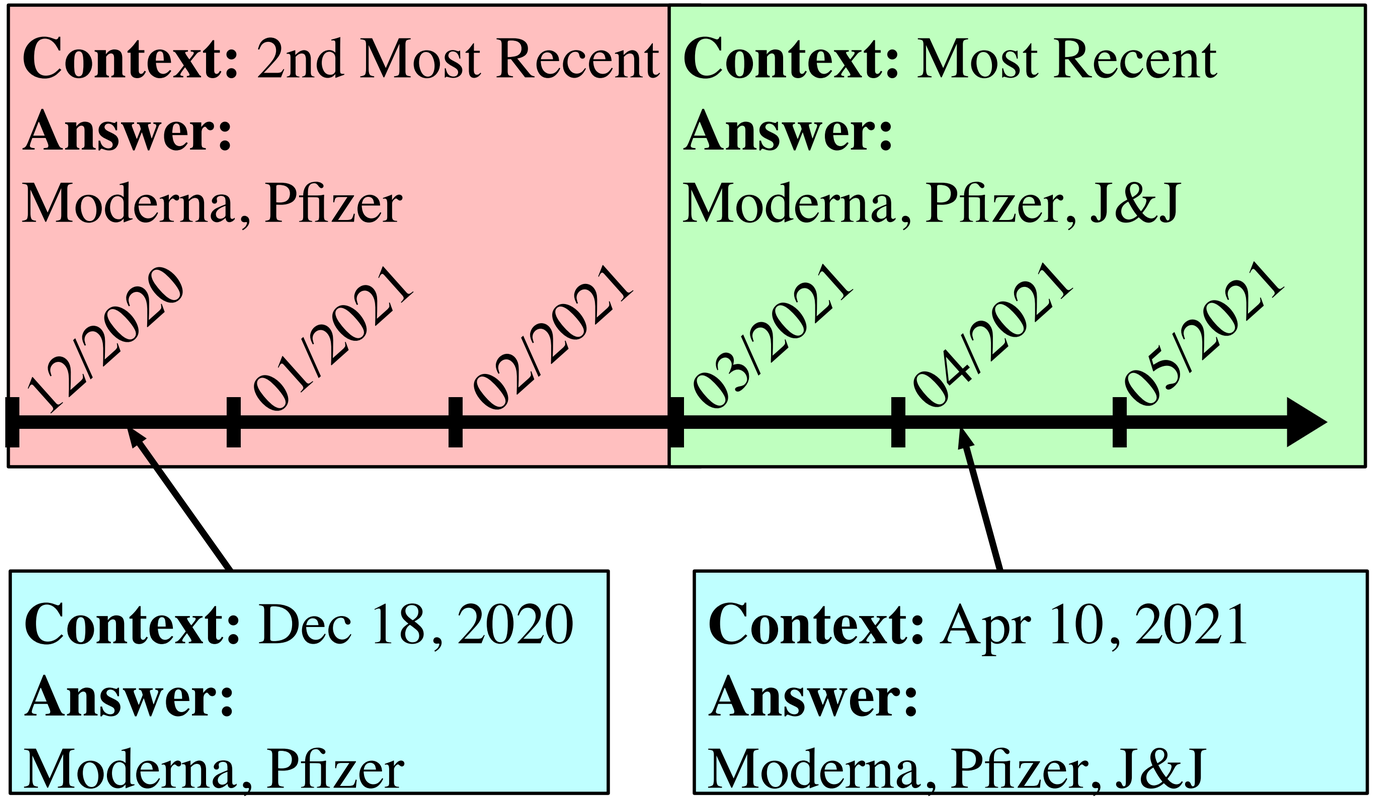
Temporal dependence of model

Pre-trained language model trained on a corpus collected up to a fixed timeframe, and perform worse in the realistic setup of predicting future utterances!

Pitfalls of Static Language Modelling [Lazaridou et al, ArXiv 2021]

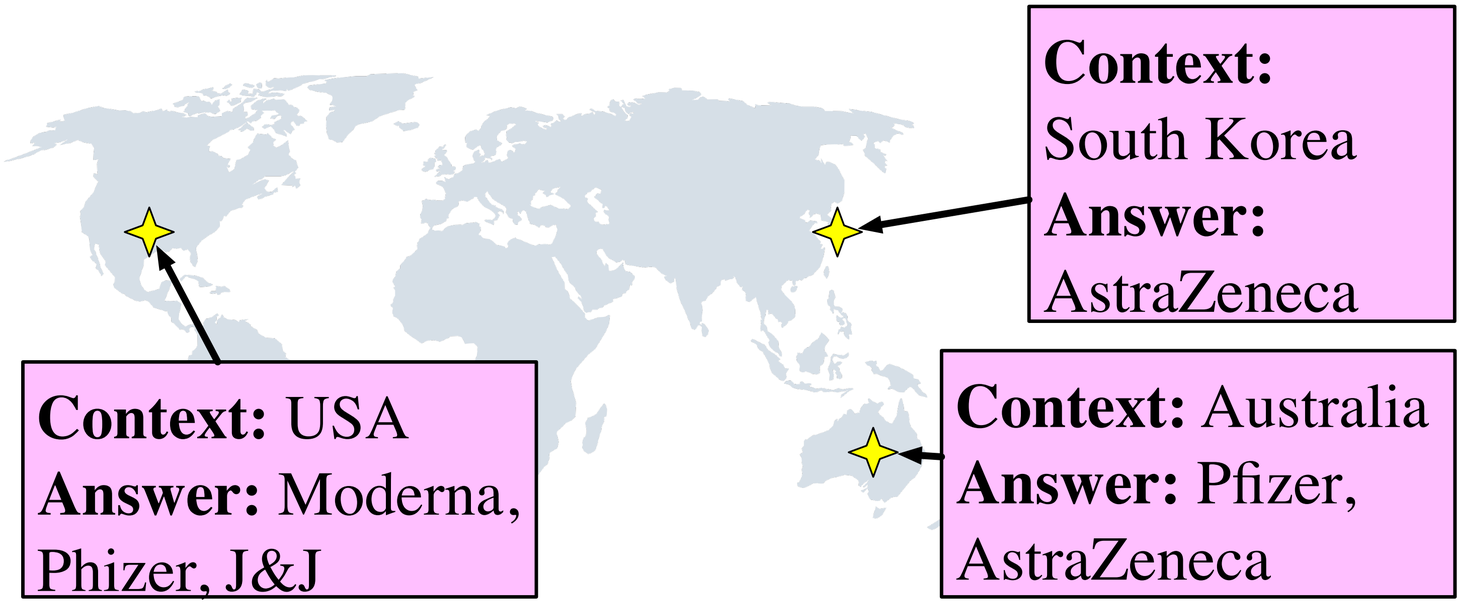
Temporal dependence of QA

Question: Which COVID-19 vaccines have been authorized for adults in the US?

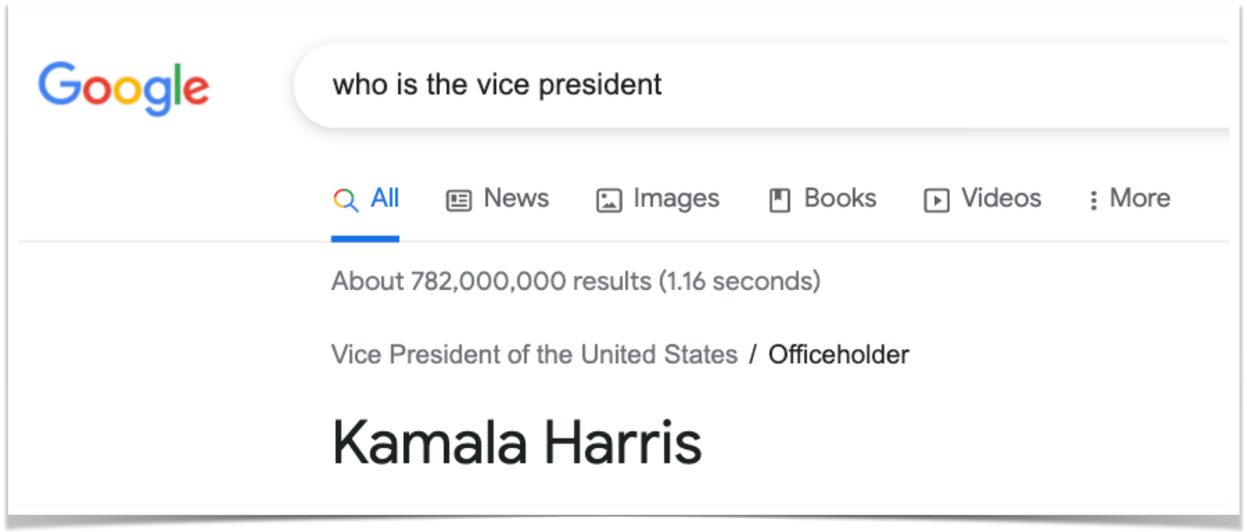


Geographical dependence of QA

Question: Which COVID-19 vaccines have been authorized by our government?



Extra-linguistic contexts for benchmark



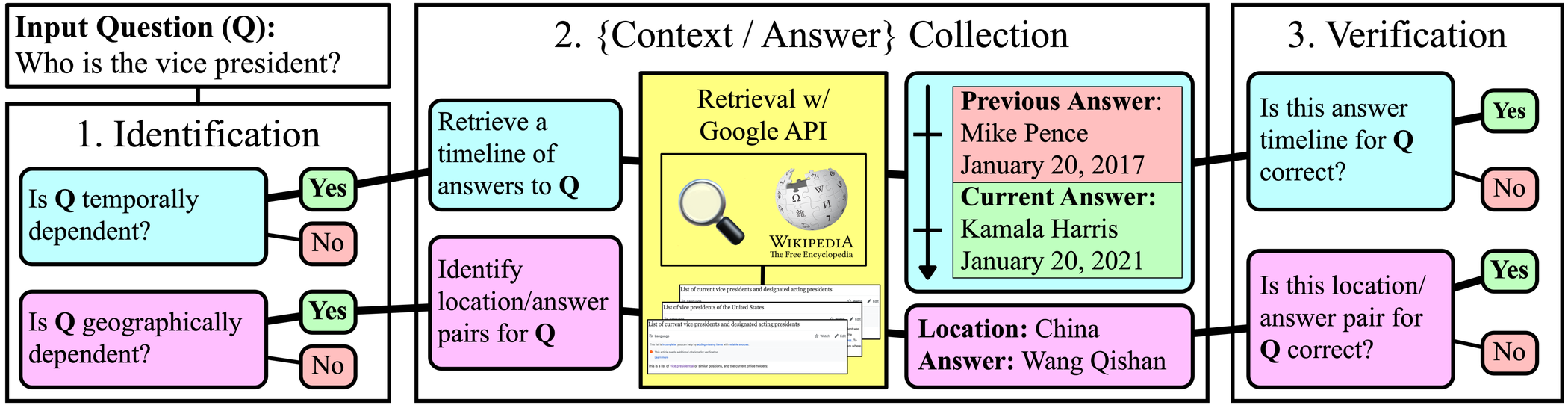
- Temporally, often assumed to be at the time stamp of dataset creation
- Geographically, often assumed to be populated areas with more NLP researchers

SituatedQA: Integrating Extra-linguistic Contexts into QA

- We study how answer to the same question changes based on ***where*** the question was asked, and ***when*** the question was posed
- Integrating temporal⌛ and geographical🌎 contexts into question answering task
- Given a question and its corresponding extra-linguistic context, find the answer

[Zhang and Choi, In Submission, 2021]

Adding extra-linguistic context into QA dataset



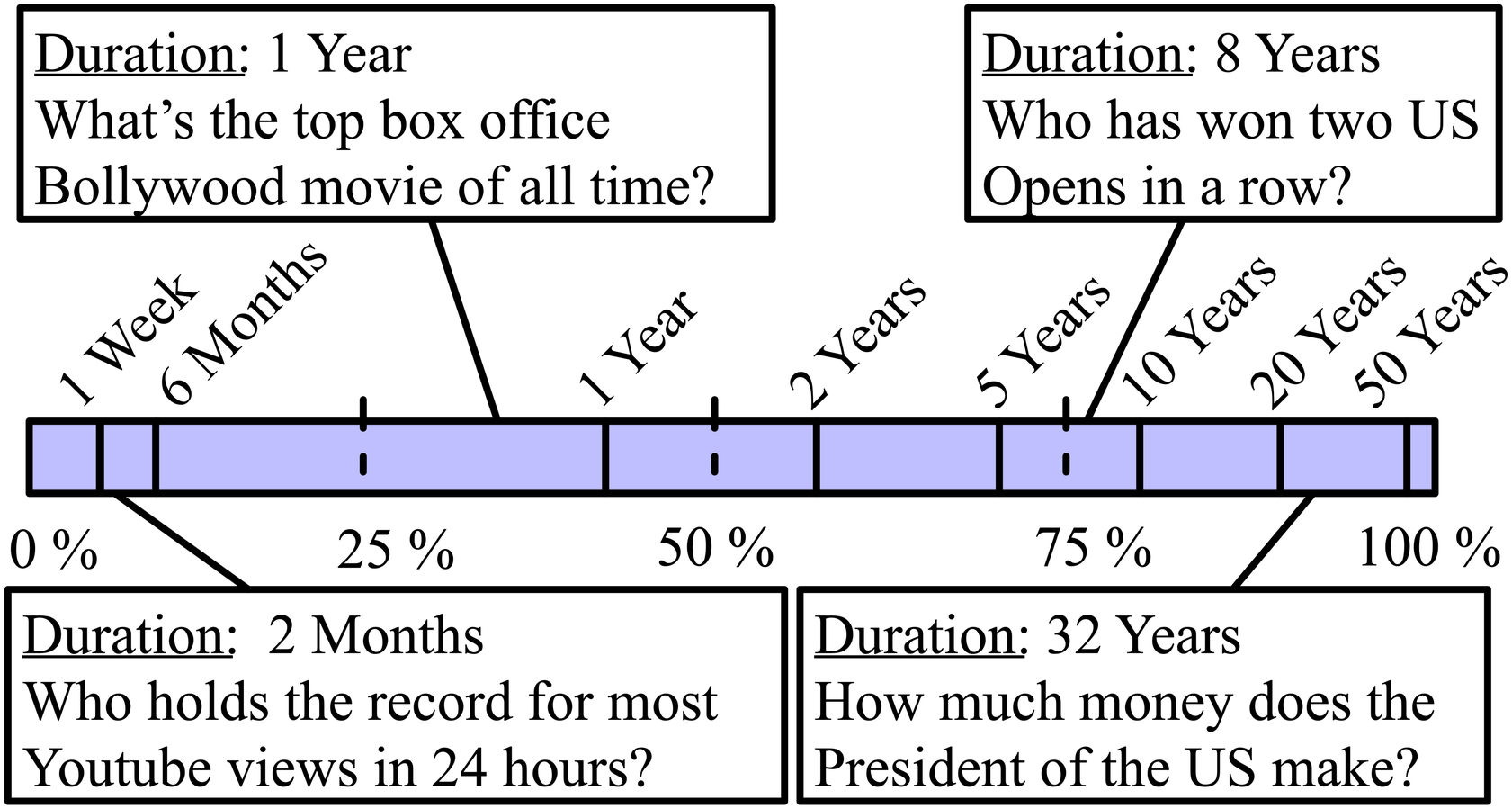
- On 5 open retrieval QA (Natural Questions, WebQuestions, TyDiQA (English only), MS-MARCO) datasets, at least 10% of them were temporally dependent, some as high as 40%

Situated QA: Task

- Given a question, a context type and a context value, find the appropriate answer

Question	Context Type	Context Value	Answer
Who went number 1 in the WNBA draft?	Temporal Relative	Current	Sabrina Ionescu
		Previous	Jackie Young
How many seasons are there for American Horror Story?	Temporal Absolute	Sep 13, 2017	9
		Sep 18, 2019	10
When was the last time states were created?	Geographical	Nigeria	1996
		US	1959

How Frequently Facts Change?



Temporal dependence of QA model

Natural Questions

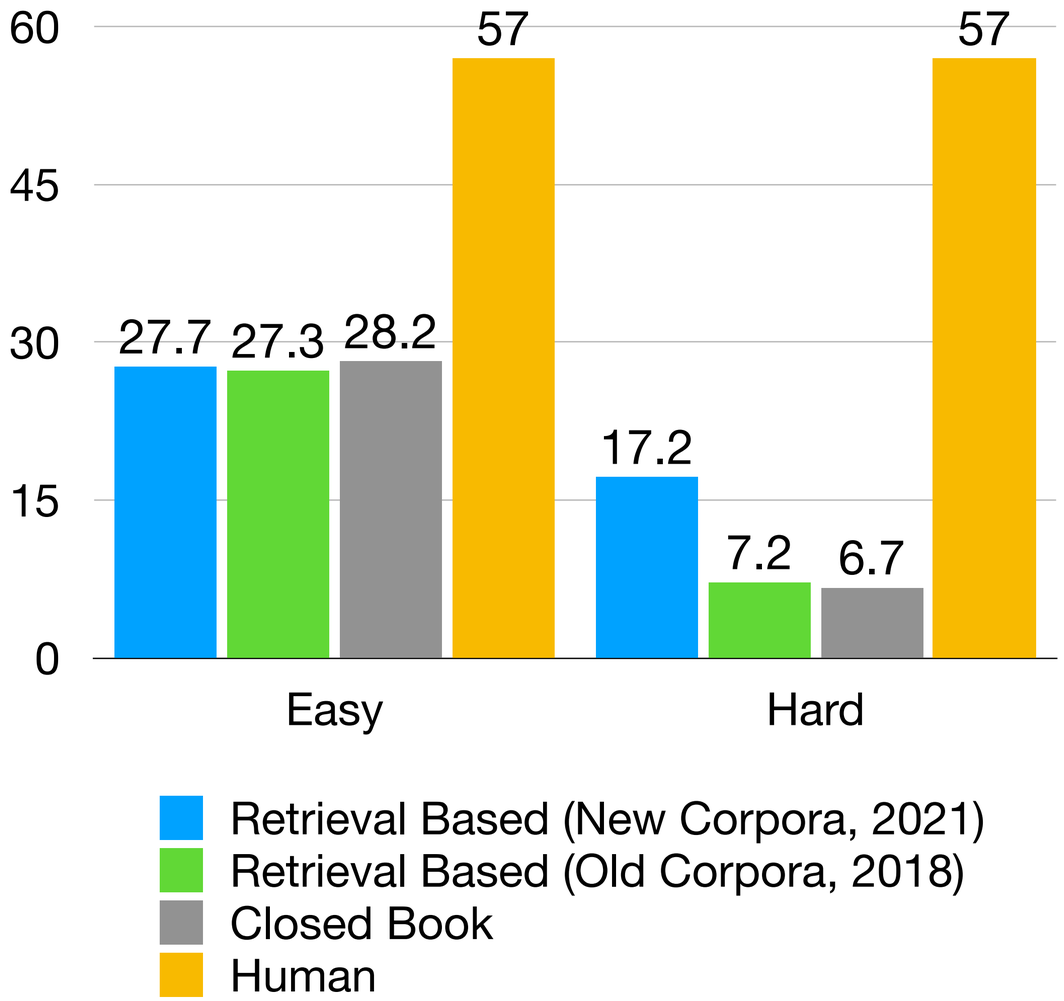
Both pre-trained language model and large-scale training data are from 2018!

Would that impact model performances?

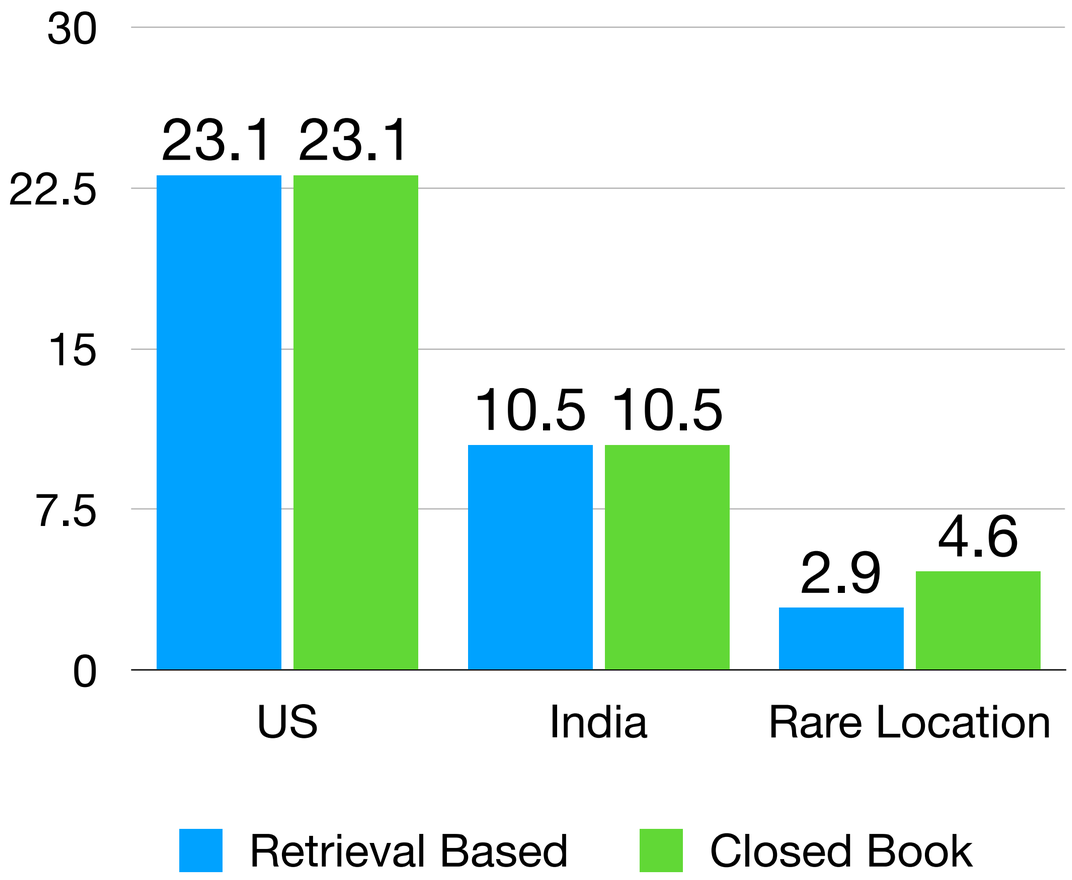
- Keep the original open retrieval QA task set up:
  - Given a question, provide an answer

    Q: Who went number 1 in the WNBA draft? A: Sabrina Ionescu
  - Assuming present (early 2021) as the temporal context.
  - Model pretrained with original NQ data (2018), fine-tuned further with our data (2021)
- Split the evaluation data into two subsets:
  - Easy: Questions where answer has not changed since 2018
  - Hard: Questions where answer is updated since 2018

Results on SituatedQA: Temporal Context



Results on Situated QA: Geographical Context



- Ambiguous question with missing geographical context:

> When was the last time states were created?

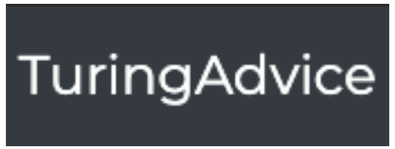
- Compare the predicted answer against answers from specific geographical context as gold reference
  - Not surprisingly, the context is often assumed to be in US!

How should we keep benchmarks relevant for the future?

- We can fix the world knowledge that we will query from:

[Petroni et al, NAACL 2021]

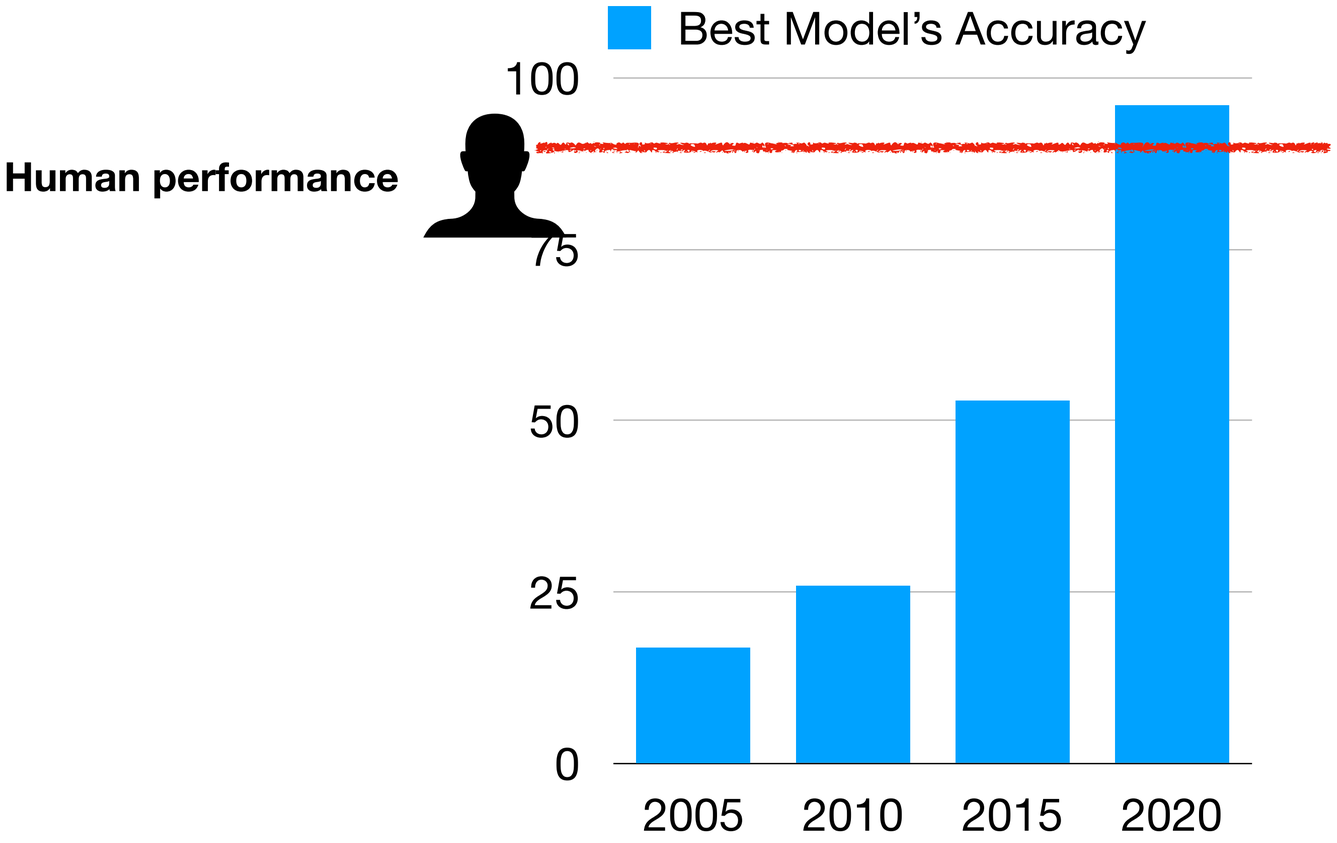
- We will keep updating the test set with newly acquired examples:



[Zeller et al, NAACL 2021]

[Kiela et al, ArXiv 2021]

When does benchmark stop being relevant?



- When model performance exceeds human performance?
- When language or knowledge in the dataset no longer reflect our world?

Summary / Concluding Thoughts

- We should consider context (e.g., computational resources) when evaluating different models in benchmarks
- We should consider extra-linguistic contexts that are assumed in our benchmark datasets
- How can we keep our benchmark ***relevant*** for the future, when language usage changes dynamically?

Thank you !

Email:
eunsol@utexas.edu

Twitter:
@eunsolc