Evaluation

October 3rd

Martin Austermuhle
@maustermuhle

Here's some data on how many tickets D.C.'s traffic cameras handed out from Oct. 1, 2021 through March 31, 2022, and how many remain unpaid. As usual, Maryland drivers got the most tickets and had the most outstanding fines. ($31 million worth, of the $59 million in fines.)

f Number of ATE Tickets Issued by T
n-Tri-State Area From 10/01/2021 to (

- ate)	Plate State	Tickets Cou
	DC	132,
	MD	275,
	VA	123,
	Sub-Total	531,
rea	Sub-Total	82,
	Grand Total	**614,**

3: Summary of ATE Outstanding Tickets with No C
rea (DC, MD & VA) vs Non-Tri-State Area As of M

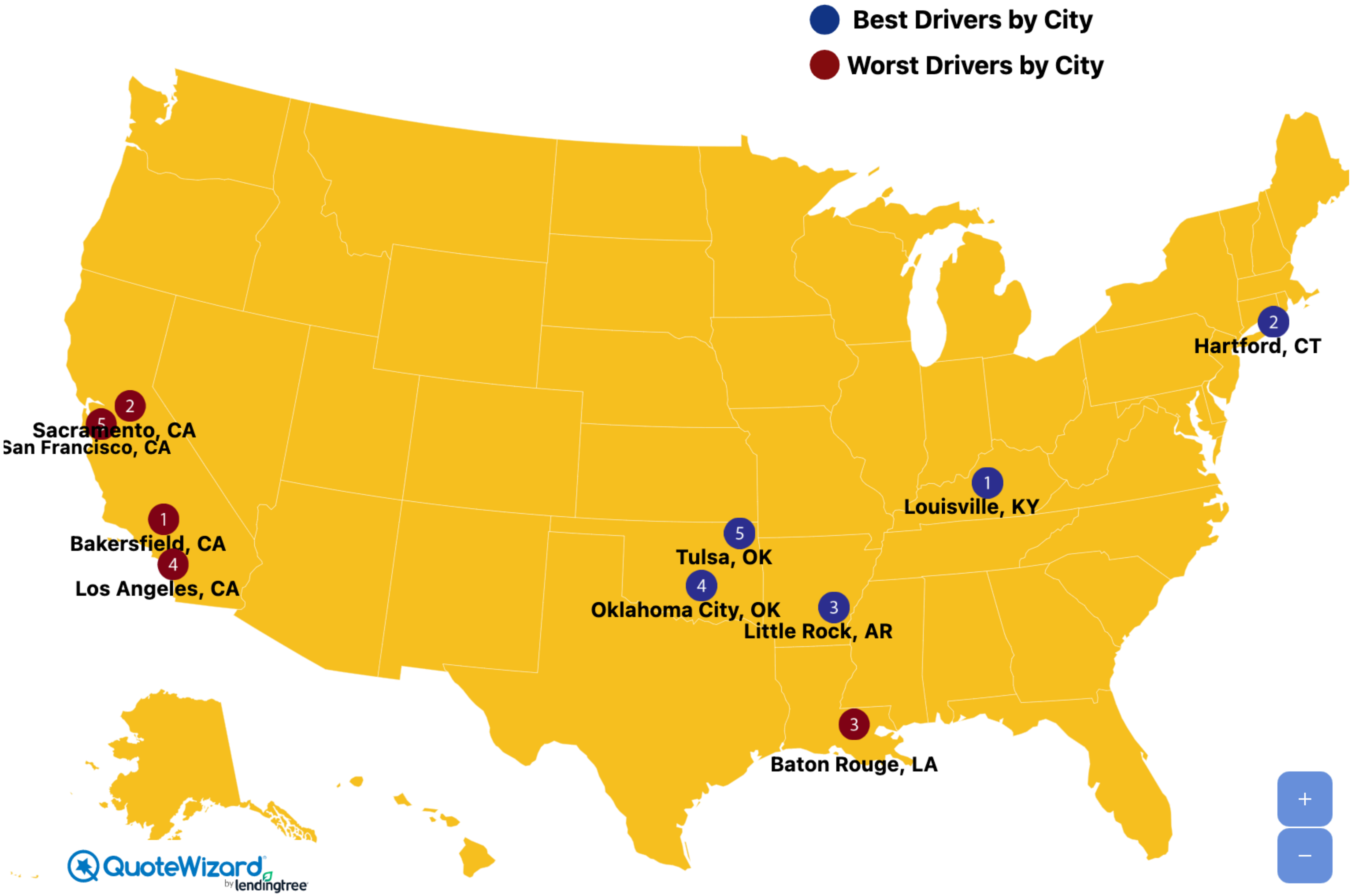
Outstanding Tickets with No Collection - From 10/01/20

Plate State	Number of Tickets without Payments	Dollar Amounts to be Collected
MD	145,618	$31,881,075
VA	71,054	$15,600,367
DC	62,469	$11,954,222
ates	**279,141**	**$59,435,664**
tates	**89.46%**	**89.3**
Plate State	**Number of Tickets without Payments**	**Dollar Amounts to be Collected**
ab Total	**32,880**	**$7,118,525**
nd Total	**312,021**	**$66,554,189**

Table 7: Summary of Number of ATE Tickets Issued by Tri-State (DC, MD, VA) and Non-Tri-State Area From 10/01/2021 to 03/31/2022

Location (Tri-State/Non-Tri-State)	Plate State	Tickets Count	% Of Mailed Tickets
Tri-State Area	DC	132,073	21.50%
	MD	275,625	44.87%
	VA	123,573	20.12%
	Sub-Total	531,271	86.49%
Non-Tri-State Area	Sub-Total	82,956	13.51%
	Grand Total	**614,227**	**100%**

[Source]


Best Drivers by City
Worst Drivers by City
Sacramento, CA
San Francisco, CA
Bakersfield, CA
Los Angeles, CA
Hartford, CT
Louisville, KY
Tulsa, OK
Oklahoma City, OK
Little Rock, AR
Baton Rouge, LA
QuoteWizard
by lendingtree

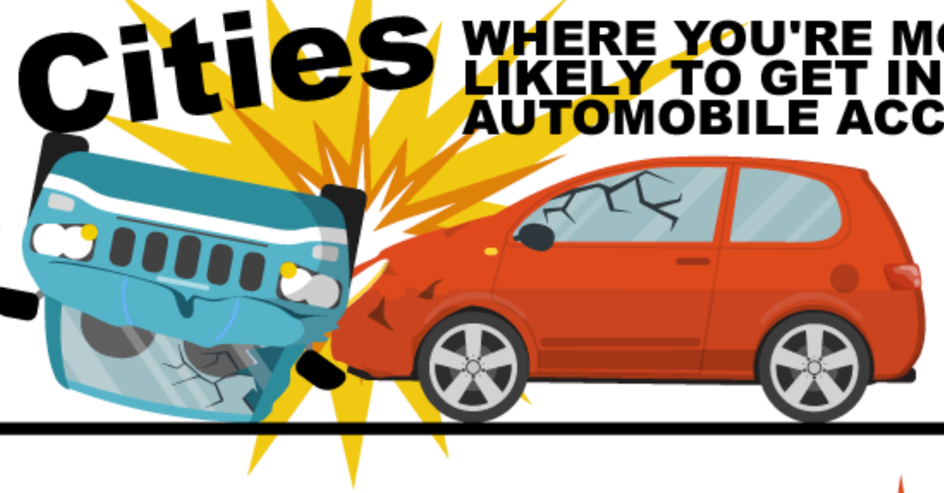
[Source]

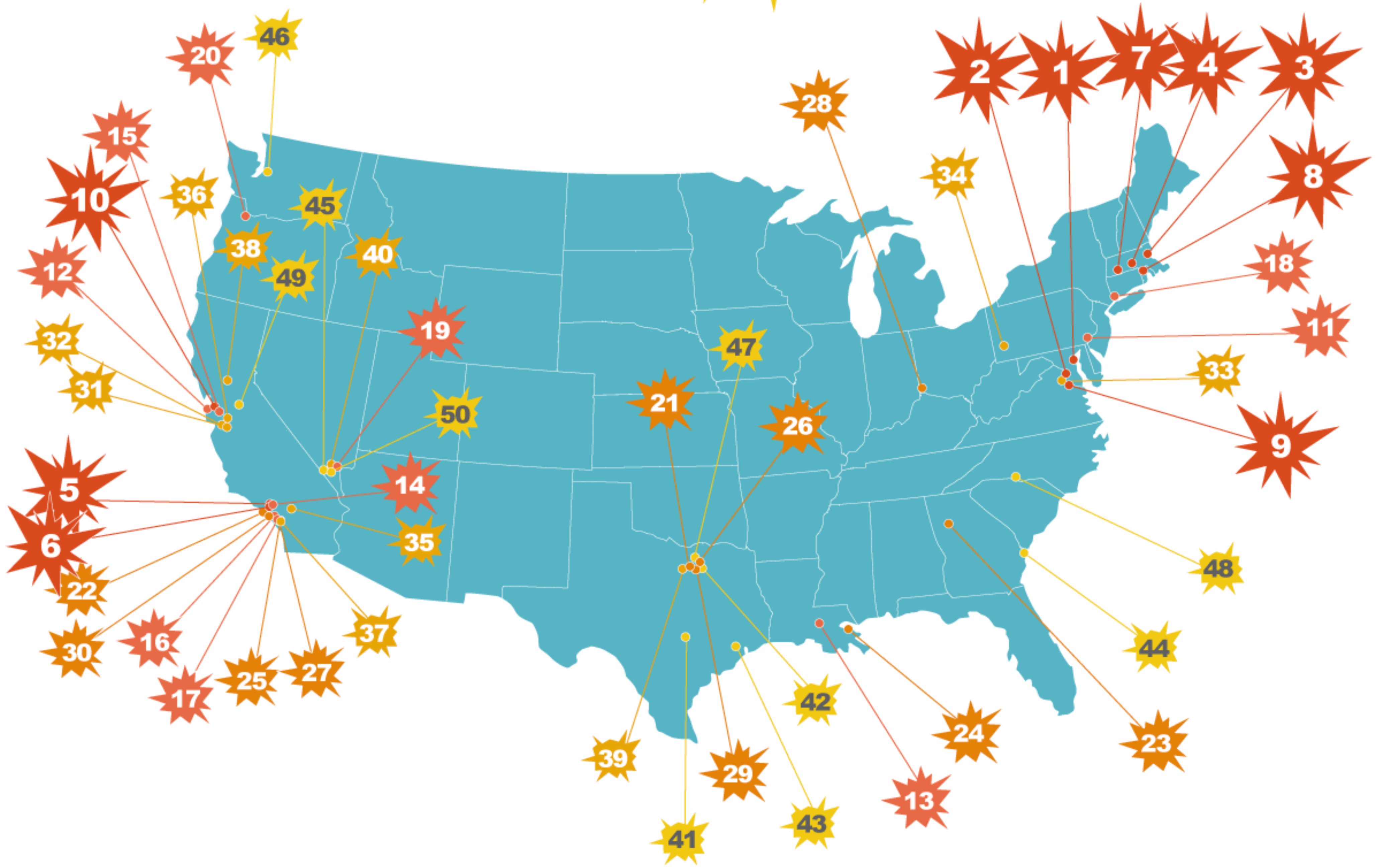
Driving factors include:

- Speeding Tickets
- Citations (running a red light, using a cellphone while driving, etc.)
- Accidents
- DUIs

[Source]

The 50 cities WHERE YOU'RE MOST LIKELY TO GET IN AN AUTOMOBILE ACCIDENT





National average for years between collisions: 10.57
National average for hard-braking events per 1,000 miles: 19

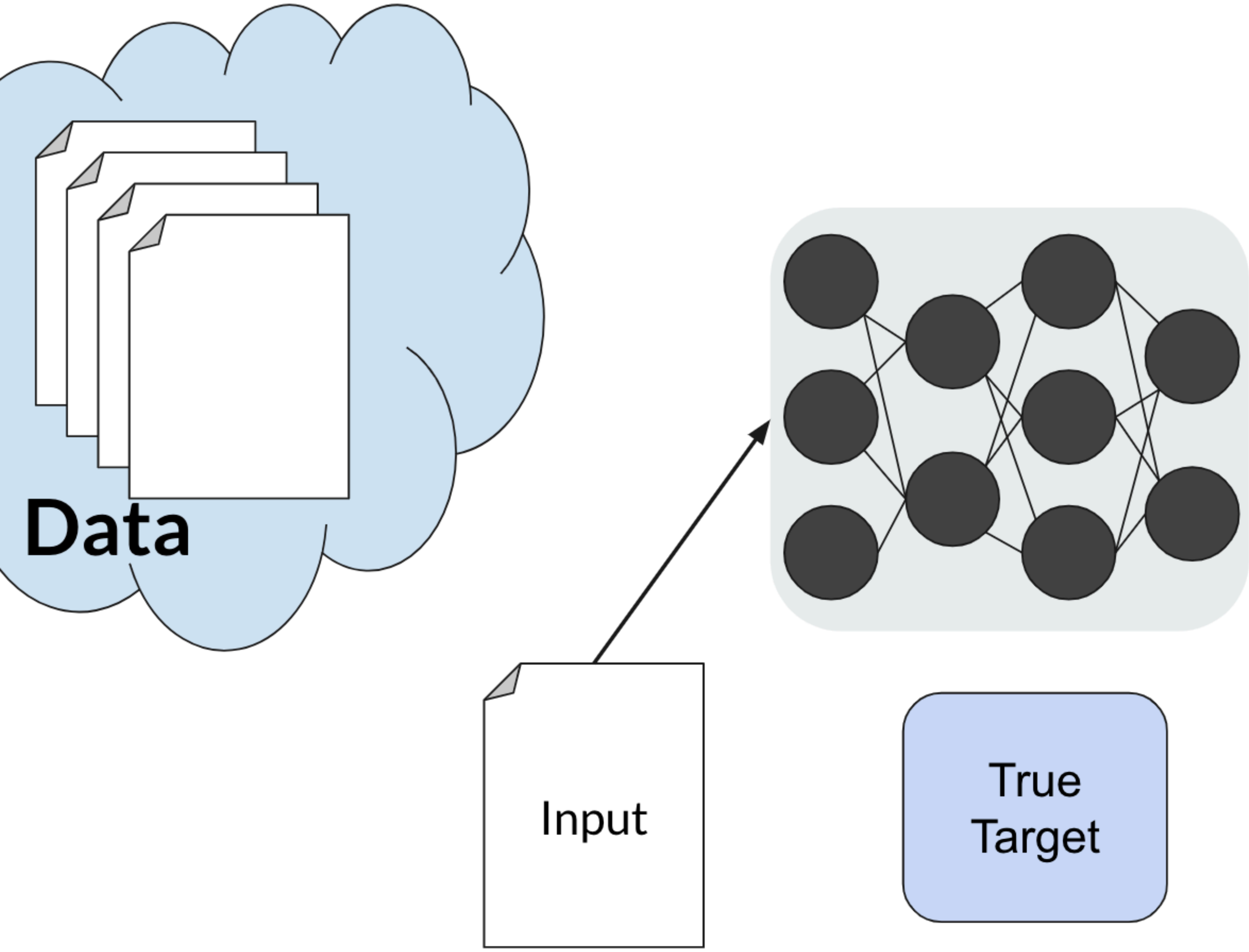
RANK		AVERAGE YEARS BETWEEN COLLISIONS	RELATIVE COLLISION LIKELIHOOD *(Compared to National Average)*	DRIVEWISE® HARD-BRAKING EVENTS PER 1,000 MILES	RANKING *Standardized for Population Density*	RANKING *Standardized for Annual Precipitation*
1	Baltimore, MD	4.19	152.5%	30.58	1	1
2	Washington, DC	4.36	142.3%	27.21	2	2
3	Boston, MA	4.89	116.5%	26.12	4	3
4	Worcester, MA	5.14	105.6%	26.87	3	4
5	Glendale, CA	5.31	99.0%	N/A	5	7
6	Los Angeles, CA	5.81	82.0%	N/A	7	6
7	Springfield, MA	5.82	81.7%	23.98	6	5
8	Providence, RI	6.19	70.8%	26.5	8	9
9	Alexandria, VA	6.22	69.9%	27.48	11	8
10	Oakland, CA	6.31	67.7%	N/A	9	10

[Source]

Which U.S. City Has the Most Car Accidents?

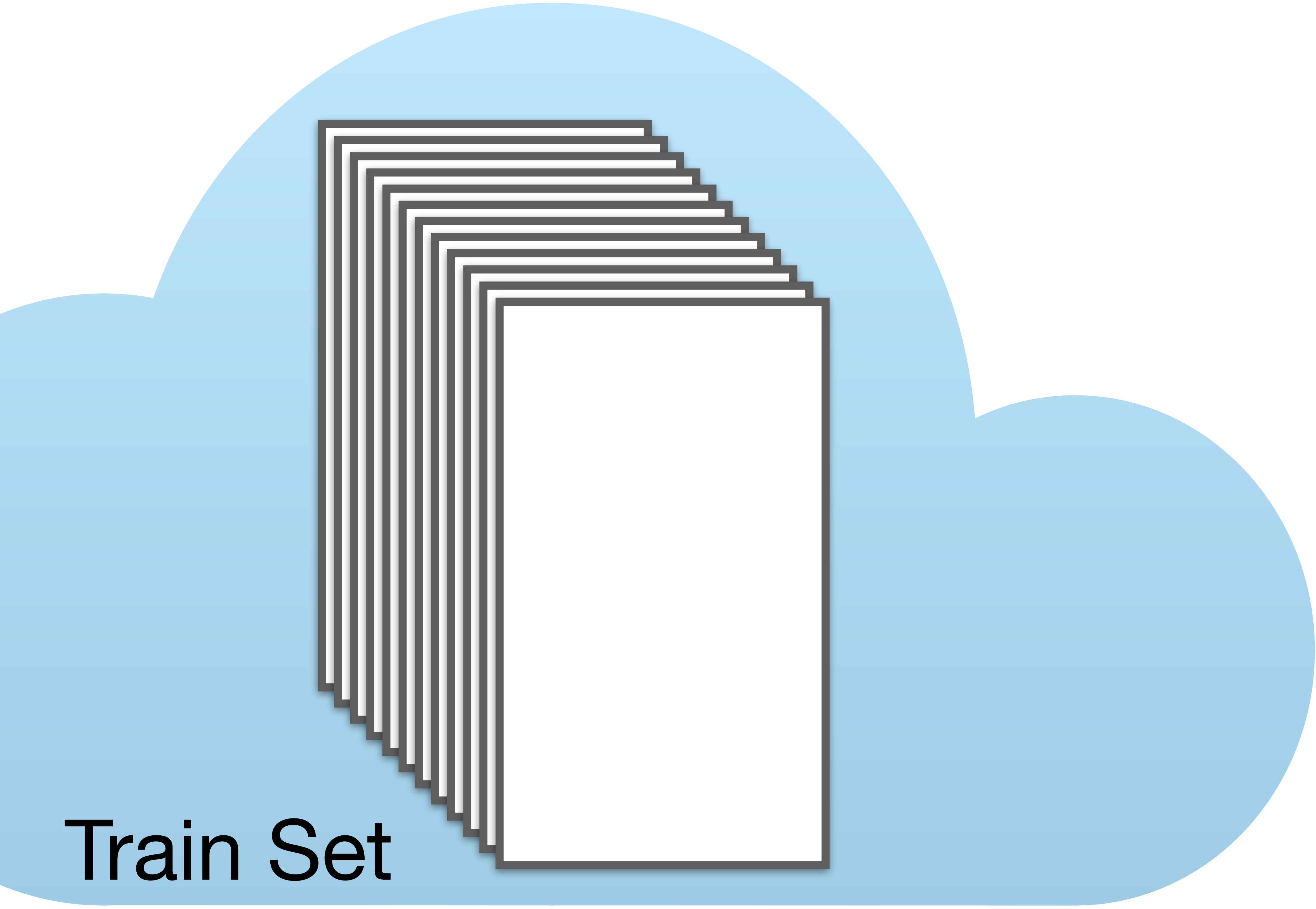
Baltimore, Maryland, ranks as the city with the most automobile accidents in the US. Drivers in Baltimore average one collision every 4.19 years, and there are 38.58 hard-breaking occurrences per 1,000 miles compared to the national average. What’s more, Baltimore scores the lowest in standardizing the ranking for population density and annual precipitation, meaning the drivers themselves are more than likely at fault.

How do we evaluate our models?



- Our model is being optimized to our dataset
- How do we know our model is learning the task and not just memorizing the data?

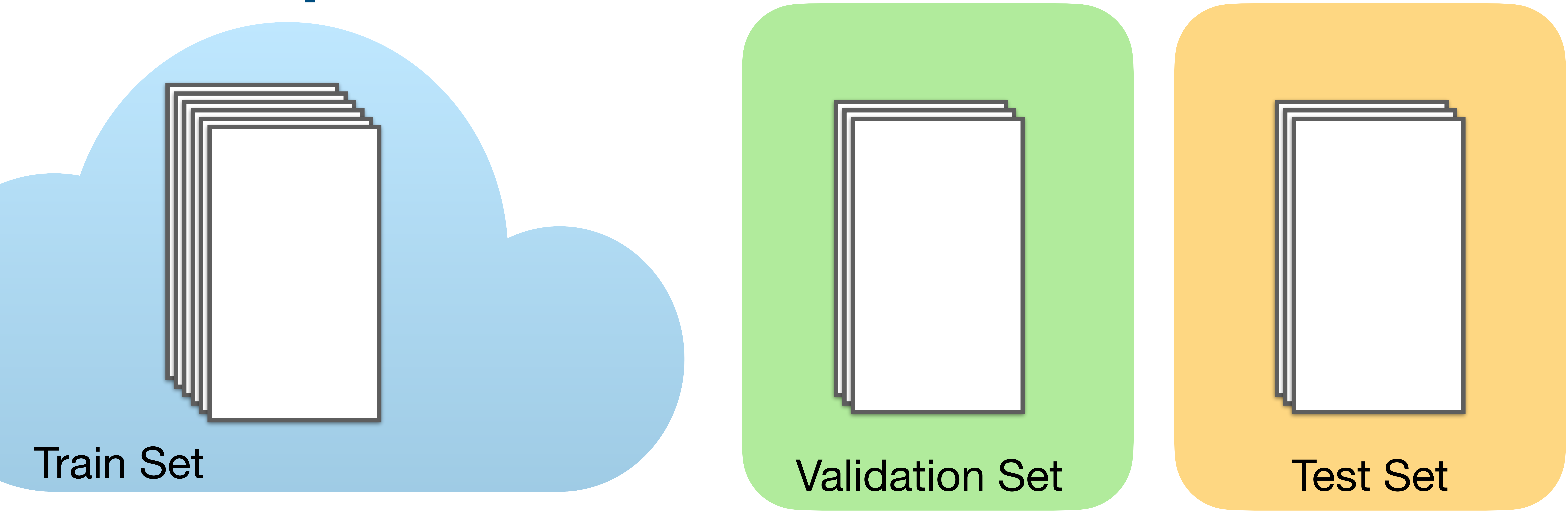
Data Splits



Data Splits

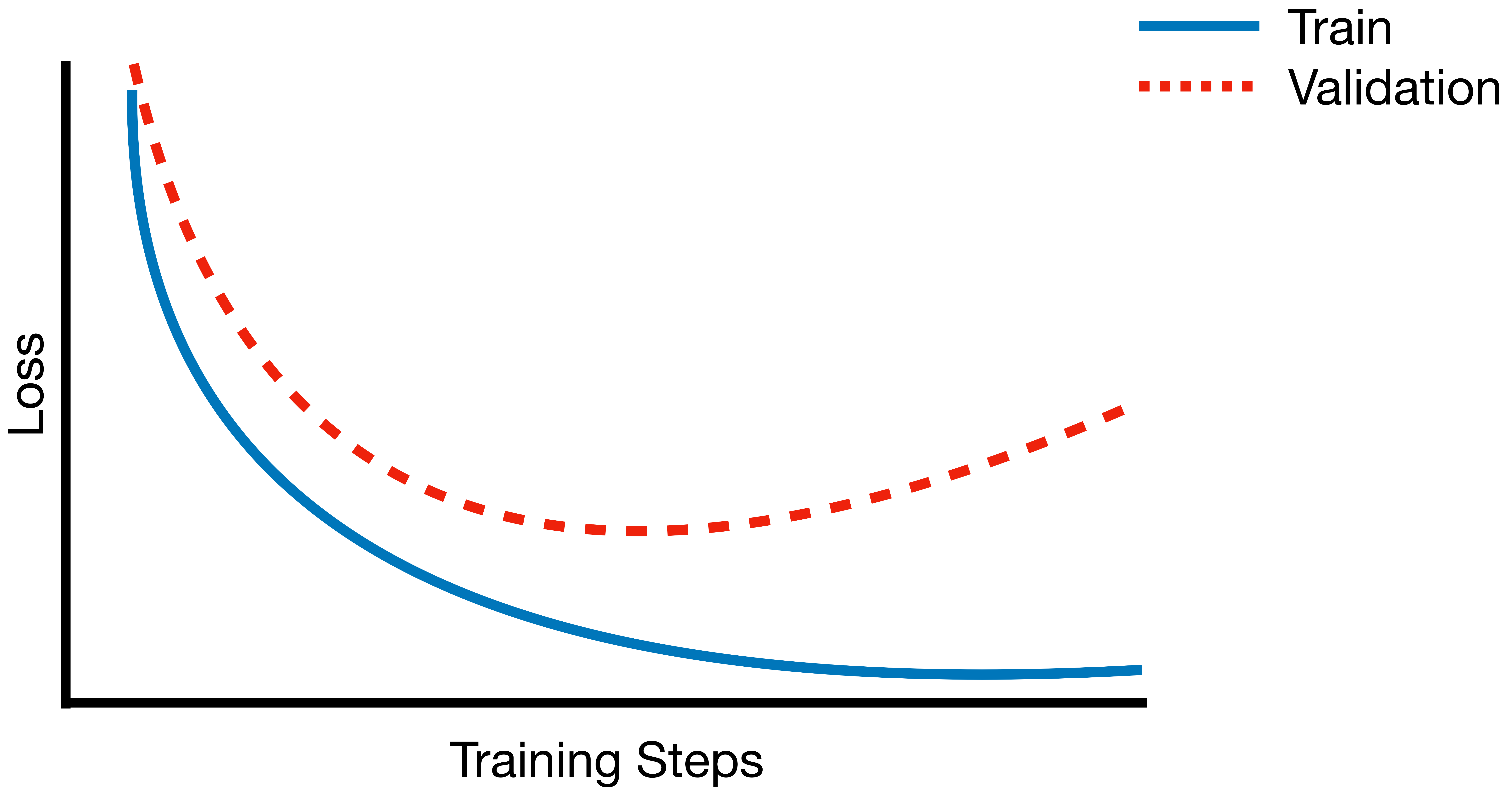
We separate a portion of our data for testing, that is unseen during training

Data Splits

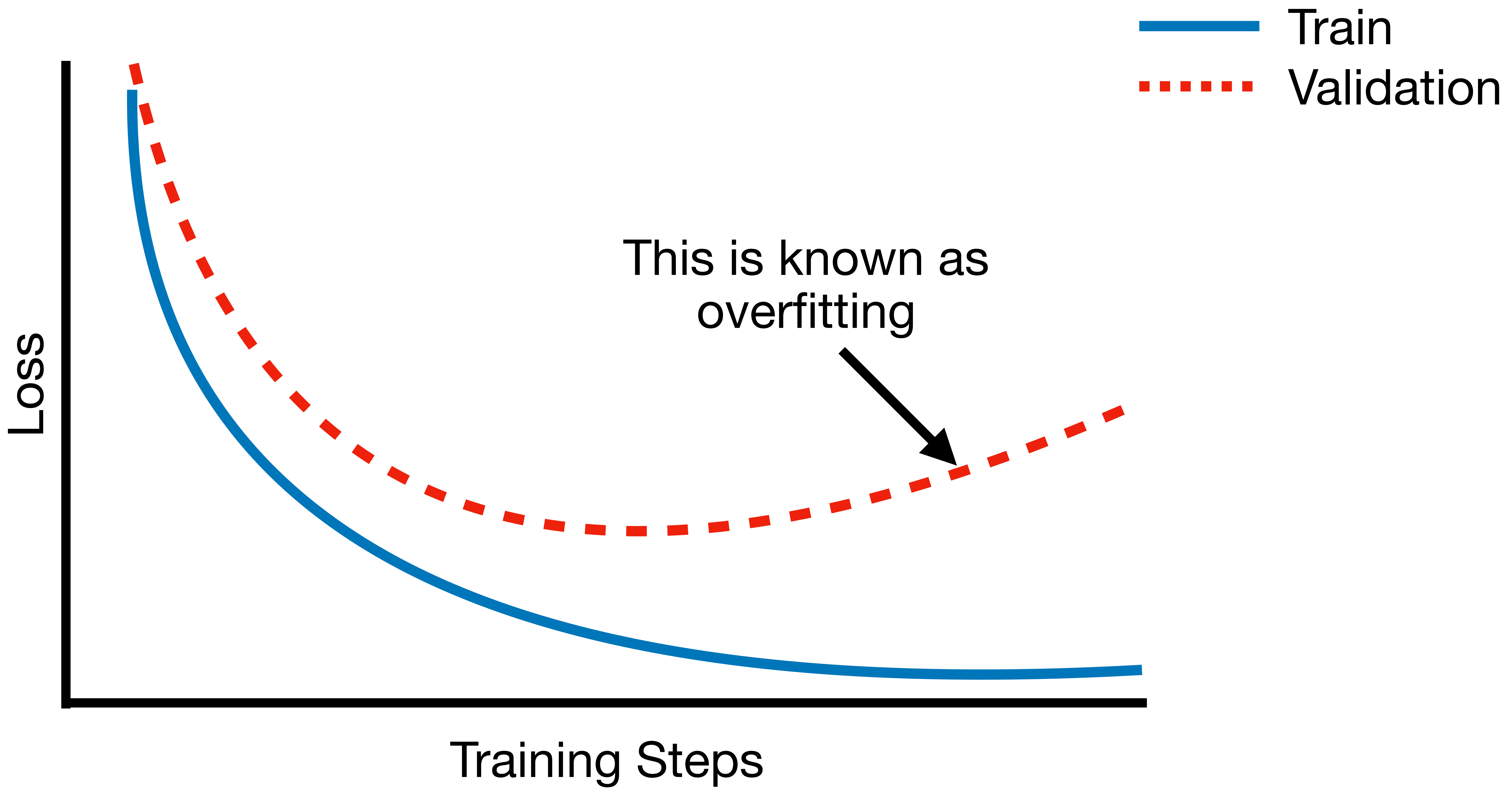


We can also separate a portion of our data for validation, to tune our hyper parameters to

Train Validation Curve



Train Validation Curve



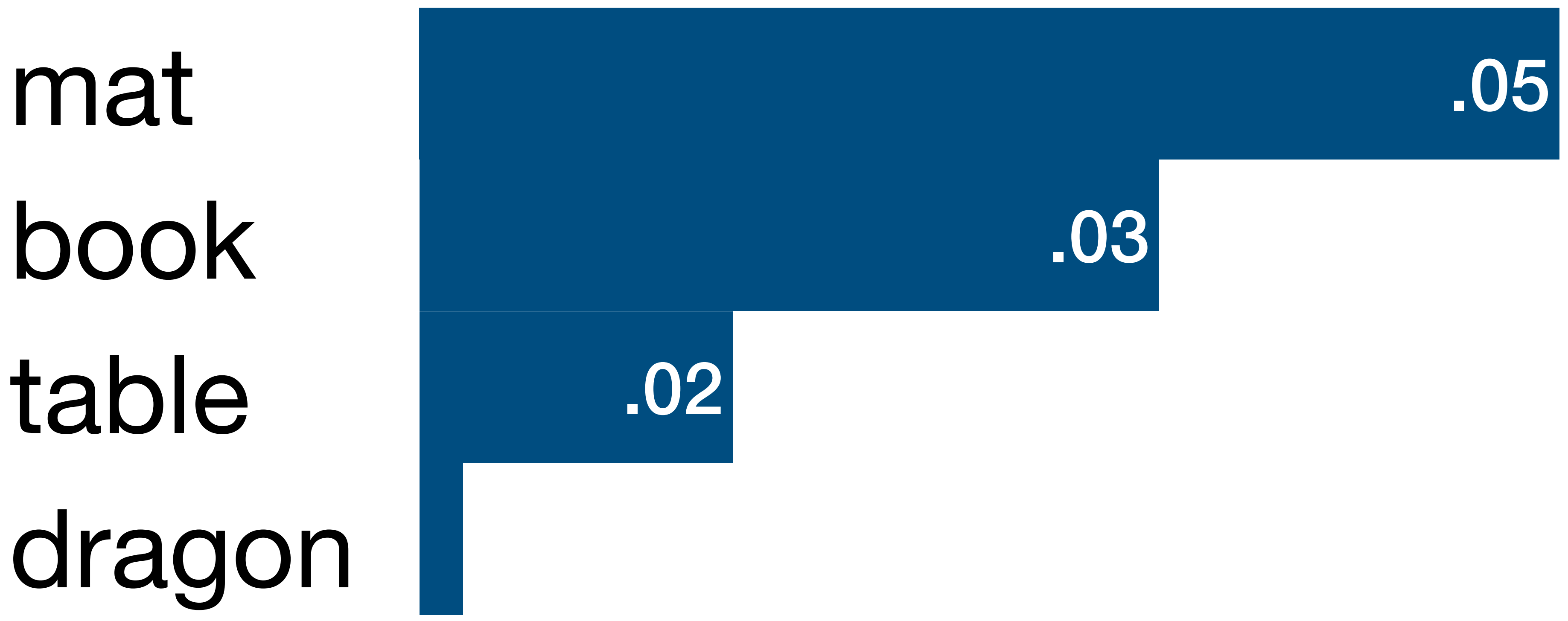
Training Process

1. Train model and evaluate on validation dataset
2. Choose model checkpoint with the best performance on the validation dataset
3. Evaluate on test set

Langauge Modeling

Language Modeling

The cat sat on the



Perplexity

[Adapted from Daniel Khashabi]

A measure if how well a probability distribution predicts a sample

Definition: for a document D with words $w_1, \ldots, w_n$:

$$\mathrm{ppl}(D) = 2^E \text{ where}$$

$$E = -\frac{1}{n} \sum_{i=1}^{n} \log_2 P(w_i \mid w_1, \ldots, w_{i-1})$$

$$E = -\frac{1}{6} \begin{bmatrix} \log_2 \mathbf{P}(\text{mat} \mid \text{the cat sat on the}) + \\ \log_2 \mathbf{P}(\text{the} \mid \text{the cat sat on}) + \\ \log_2 \mathbf{P}(\text{on} \mid \text{the cat sat}) + \\ \log_2 \mathbf{P}(\text{sat} \mid \text{the cat}) + \\ \log_2 \mathbf{P}(\text{cat} \mid \text{the}) + \\ \log_2 \mathbf{P}(\text{the}) \end{bmatrix}$$

Perplexity Base Cases

[Adapted from Daniel Khashabi]

Definition: for a document D with words $w_1, \ldots, w_n$:

$$\text{ppl}(D) = 2^E \text{ where } E = -\frac{1}{n}\sum_{i=1}^{n} \log_2 P(w_i \mid w_1, \ldots, w_{i-1})$$

If P is **uninformative**: $\forall w \in V : P(w_i \mid w_{1:i-1}) = \dfrac{1}{|V|} \Rightarrow \text{ppl}(D) = 2^{-\frac{1}{2}n\log_2\frac{1}{|V|}} = |V|$

If P is **exact**: $P(w_i \mid w_{1:i-1}) = 1 \Rightarrow \text{ppl}2^{-\frac{1}{2}n\log_2 1} = 1$

Perplexity ranges between 1 and $|V|$

Lower perplexity is good!

Perplexity is a measure of a model's **uncertainty** about the next word ("average branching factor")

Perplexity in Different Models

[Source]

Model	Perplexity
Interpolated Kneser-Ney 5-gram (Chelba et al., 2013)	67.6
RNN-1024 + MaxEnt 9-gram (Chelba et al., 2013)	51.3
RNN-2048 + BlackOut sampling (Ji et al., 2015)	68.3
Sparse Non-negative Matrix factorization (Shazeer et al., 2015)	52.9
LSTM-2048 (Jozefowicz et al., 2016)	43.7
2-layer LSTM-8192 (Jozefowicz et al., 2016)	30
Ours small (LSTM-2048)	43.9
Ours large (2-layer LSTM-2048)	39.8

Perplexity in Different Models

[Source]

Model	Layers	Heads	Perplexity
LSTMs (Grave et al., 2017)	-	-	40.8
QRNNs (Merity et al., 2018)	-	-	33.0
Adaptive Transformer (Sukhbaatar et al., 2019)	36	8	20.6
Local Transformer	16	16	19.8
Adaptive Input (Baevski and Auli, 2019)	16	16	18.7
TransformerXL (Dai et al., 2019)	18	16	18.3
Routing Transformer	10	16	**15.8**

Conditional Generative Tasks

Conditional Generation Tasks

- Which metric is most commonly used is often field specific
- Types of metrics
  1. Overlap based
  2. Similarity based
  3. Reference Free

Overlap Based Metrics

- Calculate the overlap between a model’s generation and a gold reference
- Examples:
  - ROUGE
  - BLEU
  - METEOR

ROUGE-N

How much of the reference is captured by the model's output?

$$\text{recall} = \frac{\text{number of n-grams in model and reference}}{\text{number of n-grams in reference}}$$

How much of the model's output is relevant?

$$\text{precision} = \frac{\text{number of n-grams in model and reference}}{\text{number of n-grams in model}}$$

$$\text{F1} = 2 * \frac{\text{precision} * \text{recall}}{\text{precision} + \text{recall}}$$

ROUGE ranges between 0 and 1

Higher ROUGE is good!

ROUGE-1 Example

Reference

the fox jumped over the lazy dog

Model output

the brown fox jumped over the happy dog

$$\text{recall} = \frac{\text{number of n-grams in model and reference}}{\text{number of n-grams in reference}}$$

$$= \frac{6}{7} = 0.85$$

$$\text{precision} = \frac{\text{number of n-grams in model and reference}}{\text{number of n-grams in model}}$$

$$= \frac{6}{8} = 0.75$$

$$F1 = 2 * \frac{\text{precision} * \text{recall}}{\text{precision} + \text{recall}}$$

$$= 2 * \frac{0.85 * 0.75}{0.85 + 0.75} \boxed{= 0.797}$$

Overlap Based Metrics

Pros

- Easy and quick to compute
- Easy to understand (interpretable)
- Not language specific

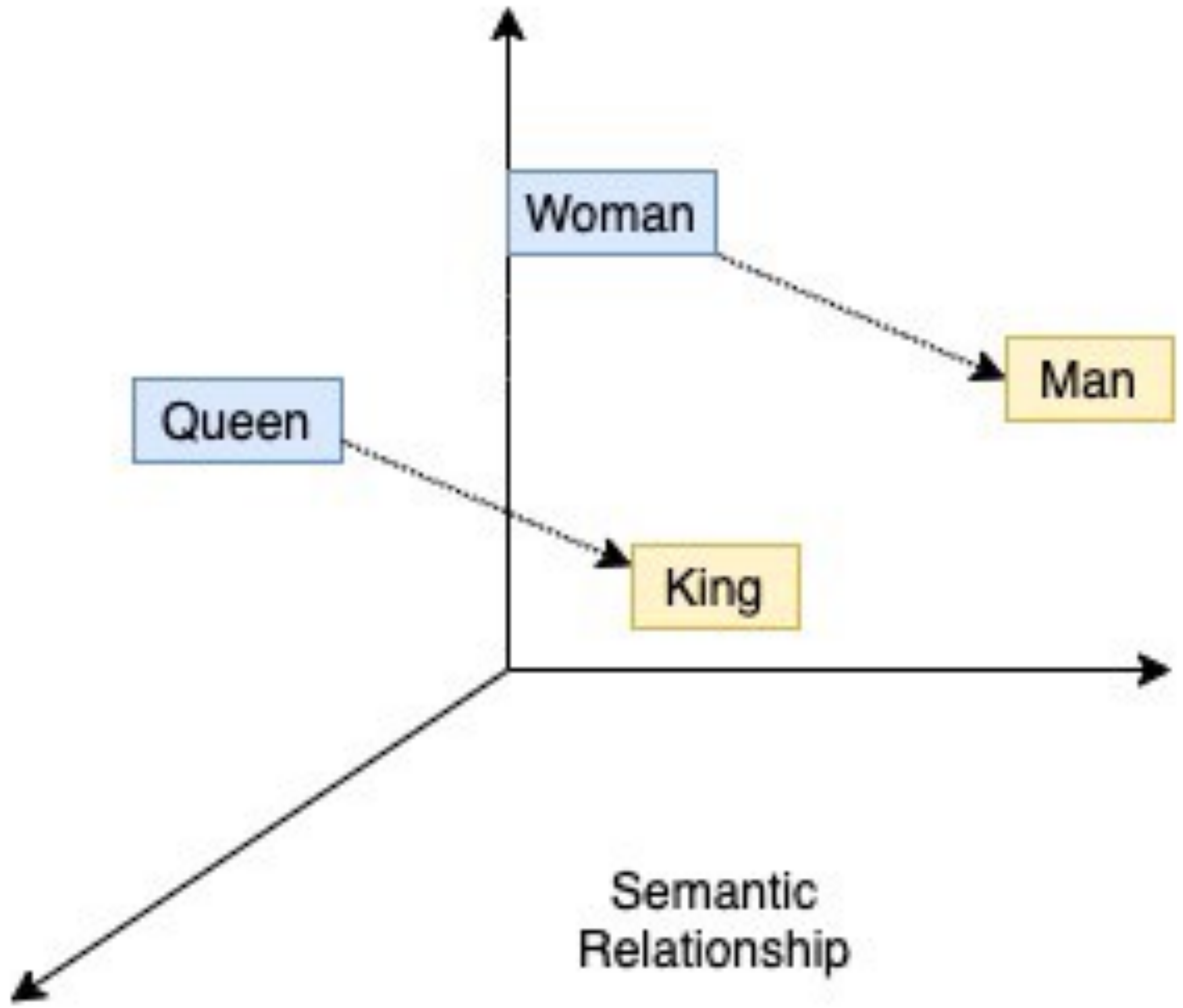
Cons

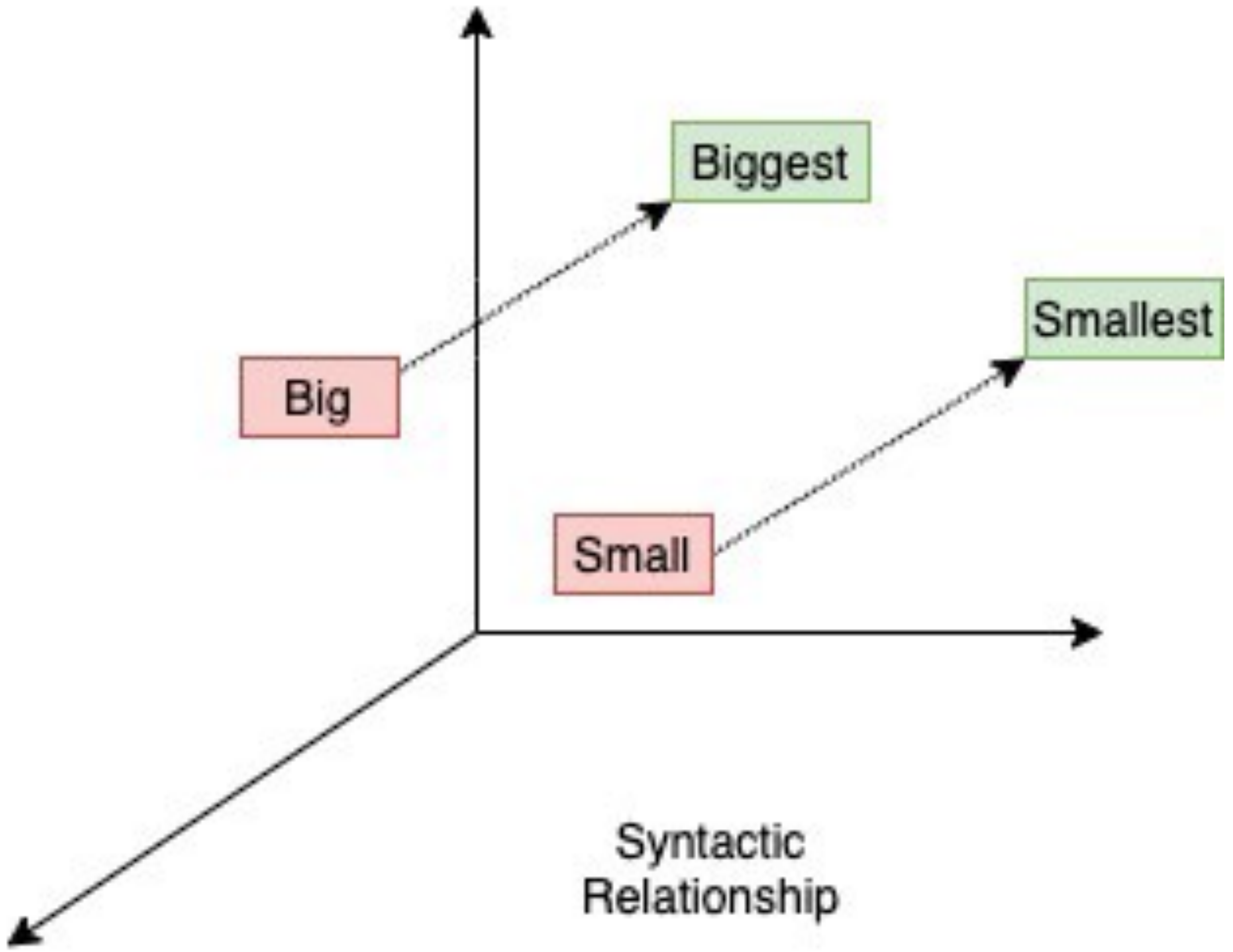
- Can be over-simplistic
- Difficult to capture nuances in language
- Requires annotated data

Similarity Based

- Computes the semantic similarity between the reference and the model output
- Examples:
  - Cosine Similarity
  - BERTScore
  - MoverScore

Recall: Embeddings

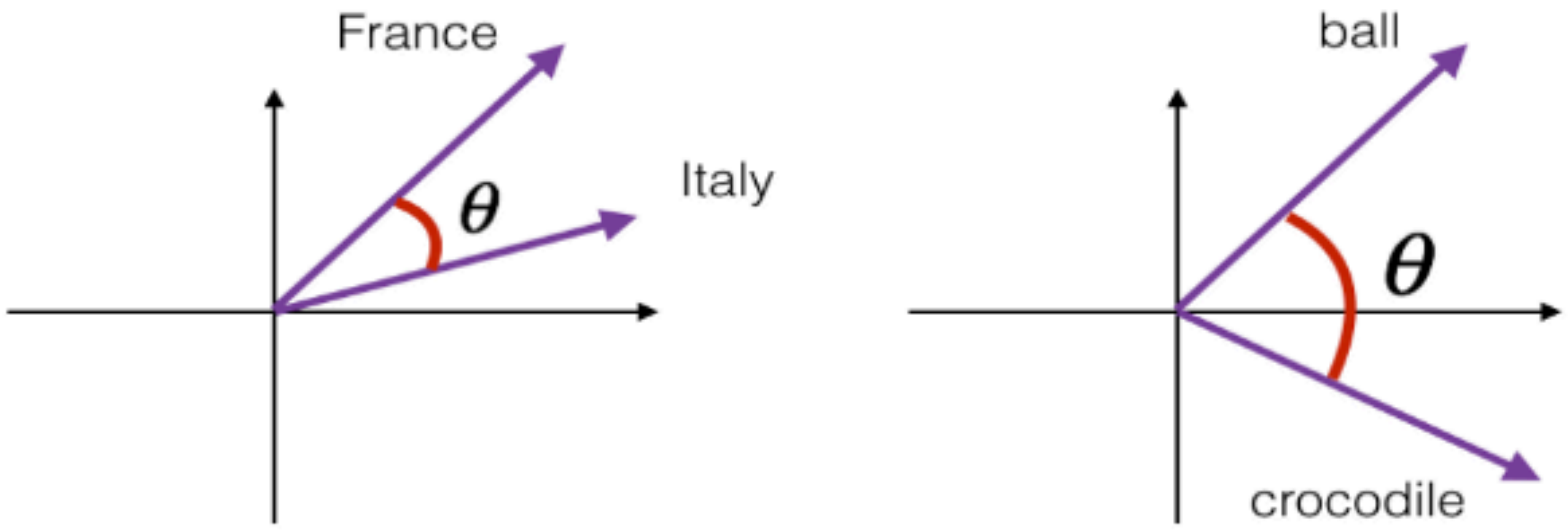




[Image Credit]

Cosine Similarity

[Source]



France and Italy are quite similar

θ is close to 0°

$\cos(\theta) \approx 1$

ball and crocodile are not similar

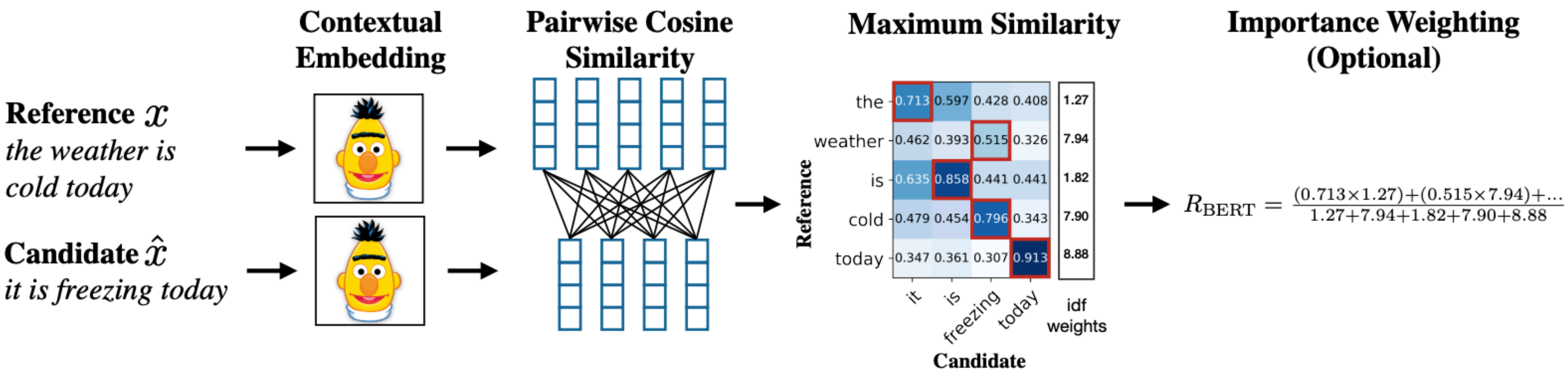
θ is close to 90°

$\cos(\theta) \approx 0$

$$similarity(A,B) = \frac{A \cdot B}{\|A\| \times \|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \times \sqrt{\sum_{i=1}^{n} B_i^2}}$$

BERTScore

[Source]



Similarity Based Metrics

Pros

- Better captures semantic similarities
- Less sensitive to small changes in output
- Generally correlates better with human judgements than overlap metrics

Cons

- Less interpretable
- Relies on trained embeddings, which may be unreliable
- Often not available in all languages

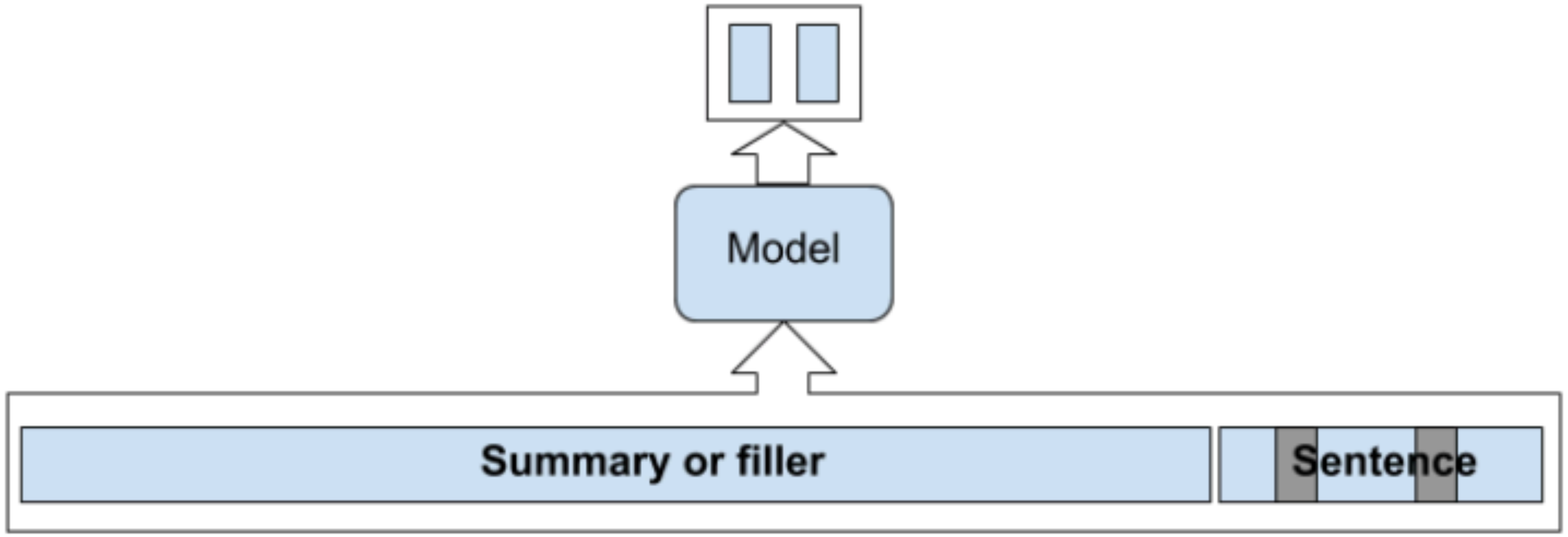
Reference Free Metrics

- Only relies on the input document
- Metrics are more task specific
  - OpenKiwi
  - BLANC
  - SUPERT

BLANC

[Source]

- A reference free summarization metric
- Give a language model a summary and a masked sentence from the original document, test how well the language model can reconstruct the sentence
- Scores range from -1 to 1 and rate the “helpfulness” of a summary



Reference-free metrics

Pros

- Does not require gold references
- Generally correlates higher with human judgements than overlap-based metrics

Cons

- Less interpretable
- Need to design a different metric for each task
- Often not available in all languages

Human Evaluation

Human Evaluation

- Generally better estimation of quality than automatic metrics
- Can can be difficult to design/ expensive to collect
- Considerations when designing a human evaluation schema:
  - How can we avoid biasing the annotators?
  - What do we care to capturing in the evaluation?
  - What background knowledge do our annotators need to have?

Comparative Evaluation

- For each example, sample two outputs (A & B) from different models
- Ask the annotator, do you prefer output A or output B?
- Considerations:
  - Captures relative preference
  - Easy to design
  - Does not easily capture minor differences
  - Does not account for both outputs being bad

Rate the quality of the generations

- Ask the annotator to rate the quality of a generation on a scale
- Difficult to define "quality" ➡️ better to test for specific desired traits
  - Rate grammatically
  - Rate factual correctness

Breaking News!

Posted last week

News Summarization and Evaluation in the Era of GPT-3

Tanya Goyal[1] **Junyi Jessy Li**[2] **Greg Durrett**[1]
[1] Department of Computer Science [2] Department of Linguistics
The University of Texas at Austin
tanyagoyal@utexas.edu

Abstract

The recent success of zero- and few-shot prompting with models like GPT-3 has led to a paradigm shift in NLP research. In this paper, we study its impact on text summarization, focusing on the classic benchmark domain of news summarization. First, we investigate how zero-shot GPT-3 compares against fine-tuned models trained on large summarization datasets. We show that not only do humans overwhelmingly prefer GPT-3 summaries, but these also do not suffer from common dataset-specific issues such as poor factuality. Next, we study what this means for evaluation, particularly the role of gold standard test sets. Our experiments show that both reference-based and reference-free automatic metrics, e.g. recently proposed QA- or entailment-based factuality approaches, cannot reliably evaluate zero-shot summaries. Finally, we discuss future research challenges beyond generic summarization, specifically, keyword- and aspect-based summarization, showing how dominant fine-tuning approaches compare to zero-shot prompting.

CNN article: https://www.cnn.com/2022/09/09/politics/judge-throws-out-trumps-rico-lawsuit-against-hillary-clinton-and-democrats/index.html

Length-constrained prompt

Summarize the above article in 2 sentences.

GPT-3: Donald Trump's lawsuit against Hillary Clinton, the Democratic National Committee, and several other individuals and entities has been dismissed by a federal judge. The judge ruled that Trump's claims are unsupported by any legal authority and that many of his specific factual assertions are implausible or unsupported.

Keyword-constrained prompt

Summarize the above article briefly focusing on Alina Habba.

GPT-3: Alina Habba is an attorney for Donald Trump. She stated that Trump will "immediately move to appeal this decision." Habba also said that the judge's opinion is "rife with erroneous applications of the law" and that it "disregards the numerous independent governmental investigations which substantiate our claim that the defendants conspired to falsely implicate our client and undermine the 2016 Presidential election."

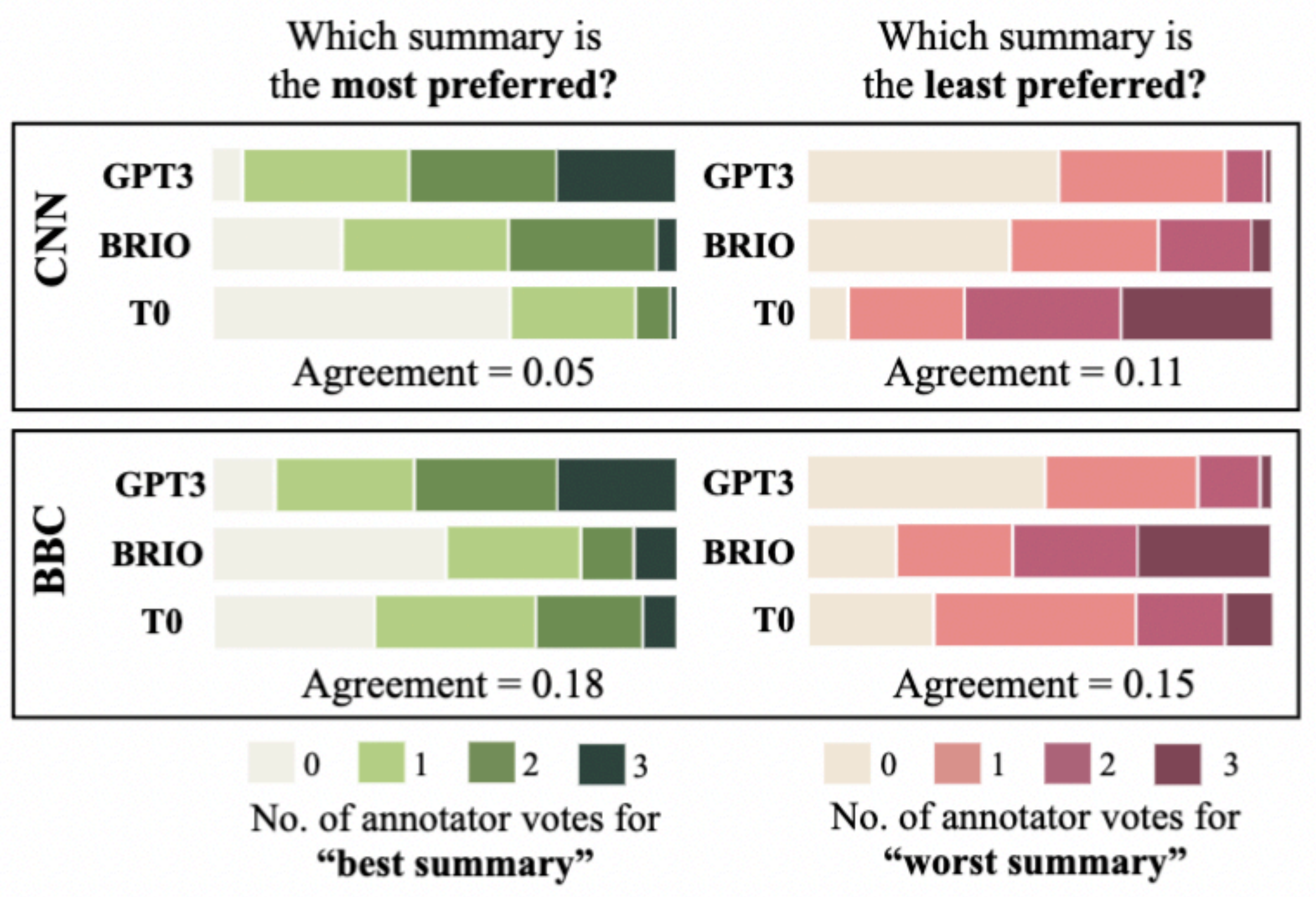
Figure 1: Examples of zero-shot summarization using GPT-3. We can generate summaries following style constraints or queries included in the prompts, allowing us to emulate a range of existing fine-tuned systems.

marizing data from a new source domain or producing a summary in a different style. The success of prompt-based models (GPT-3 (Brown et al., 2020), T0 (Sanh et al., 2022), PaLM (Chowdhery et al., 2022), etc.) provides an alternative approach,

Should we rethink evaluation?

[Source]

- Found that humans prefer GPT-3 summaries to summaries from fine-tuned models
- Automatic metrics did not accurately capture this preference



Karson Elmgren @kelmgren · Sep 27

it turns out evaluation of language models is uh... complicated

> **Tanya Goyal** @tanyaagoyal · Sep 27
>
> ✨New preprint✨
> Zero-shot GPT-3 does *better* at news summarization than any of our fine-tuned models. Humans like these summaries better. But all of our metrics think they're MUCH worse.
>
> Work/ w/ @jessyjli, @gregd_nlp. Check it out here:
> arxiv.org/abs/2209.12356
> [1/6]
>
> Show this thread

Mario Filho
@mariofilhoml

Just like we have a move towards data-centric AI, it might be time to rethink our evaluation metrics and make them more aligned with real-life goals

Good insight. Thanks for sharing @tanyaagoyal

> **Tanya Goyal** @tanyaagoyal · Sep 27
>
> ✨New preprint✨
> Zero-shot GPT-3 does *better* at news summarization than any of our fine-tuned models. Humans like these summaries better. But all of our metrics think they're MUCH worse.
>
> Work/ w/ @jessyjli, @gregd_nlp. Check it out here: arxiv.org/abs/2209.12356
> [1/6]
>
> Show this thread

Xavier Amatriain
@xamat

"We need to rethink automatic evaluation". +100

> **Tanya Goyal** @tanyaagoyal · Sep 27
>
> ✨New preprint✨
> Zero-shot GPT-3 does *better* at news summarization than any of our fine-tuned models. Humans like these summaries better. But all of our metrics think they're MUCH worse.
>
> Work/ w/ @jessyjli, @gregd_nlp. Check it out here: arxiv.org/abs/2209.12356
> [1/6]
>
> Show this thread

Tal Linzen
@tallinzen

Sigh, once again I'm going to have to make a completely new set of slides next time I teach NLP

> **Greg Durrett** @gregd_nlp · Sep 27
>
> Check out Tanya's paper! GPT-3 is a huge paradigm shift for summarization that the community hasn't fully digested yet. You should play around with davinci-002 for your own summ tasks! If there's something you always wanted to do but didn't have data for, it might work zero-shot! twitter.com/tanyaagoyal/st...

9:13 PM · Sep 27, 2022 · Twitter Web App

Next Week

- Ethics
- **Reading:** On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?
  - https://dl.acm.org/doi/pdf/10.1145/3442188.3445922