

CS378: Natural Language Processing

Lecture 22: Dataset Bias and Spurious Correlations

Greg Durrett



TEXAS

The University of Texas at Austin



Announcements

- ▶ A5 due today
- ▶ Final project released today (more details at the end of today's lecture)



Recap

- ▶ **Pretraining (BERT):**
 - ▶ Train a big model to fill in masked-out words, then adapt it to other tasks. Led to big gains in **question answering** and **NLI** performance. BART/T5, GPT-3, etc. push this further.
- ▶ **Question answering (QA):**
 - ▶ “What was Marie Curie the first female recipient of?”
-> “The Nobel Prize” (find this span in a document containing the answer)
- ▶ **Natural language inference (NLI):**
 - ▶ "But I thought you'd sworn off coffee."
contradicts "I thought that you vowed to drink more coffee."



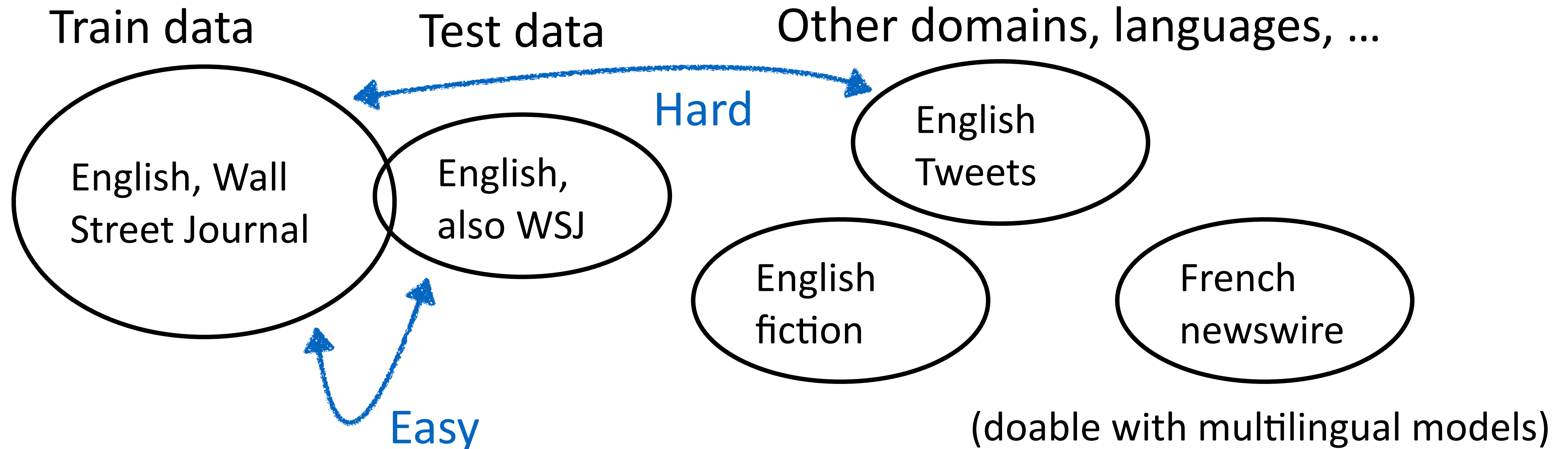
Model Performance

- ▶ If models can be fine-tuned on large datasets and perform very well on the held-out test dataset, is the problem solved?
- ▶ What can go wrong?



Generalization

- ▶ If a model does well on train but poorly on test data, it *doesn't generalize*
- ▶ A model can do well on its test data and still fail to generalize *out of distribution* — arguably an even more important notion
- ▶ Many notions of generalization. Example: POS tagging





Generalization: QA

Train data

SQuAD: factoid questions with answers on Wikipedia

Test data

SQuAD

Other *domains*

Science questions

Unanswerable questions

French questions

Other types of reasoning, such as *multi-hop questions*

Who won the Nobel in Chemistry the year Marie Curie won the Nobel in Physics?



Generalization

- ▶ Just doing well on a single test set **is not that useful**
- ▶ We want POS taggers, QA systems, and more that can generalize to new settings so we can deploy them in practice
- ▶ Sometimes, you can get **very good test performance** but the model **generalizes very poorly**. How does this happen?

Annotation Artifacts, Reasoning Shortcuts: QA



Annotation Artifacts

- ▶ Some datasets might be easy because of how they're constructed, especially in QA and NLI

What becomes of Macbeth?

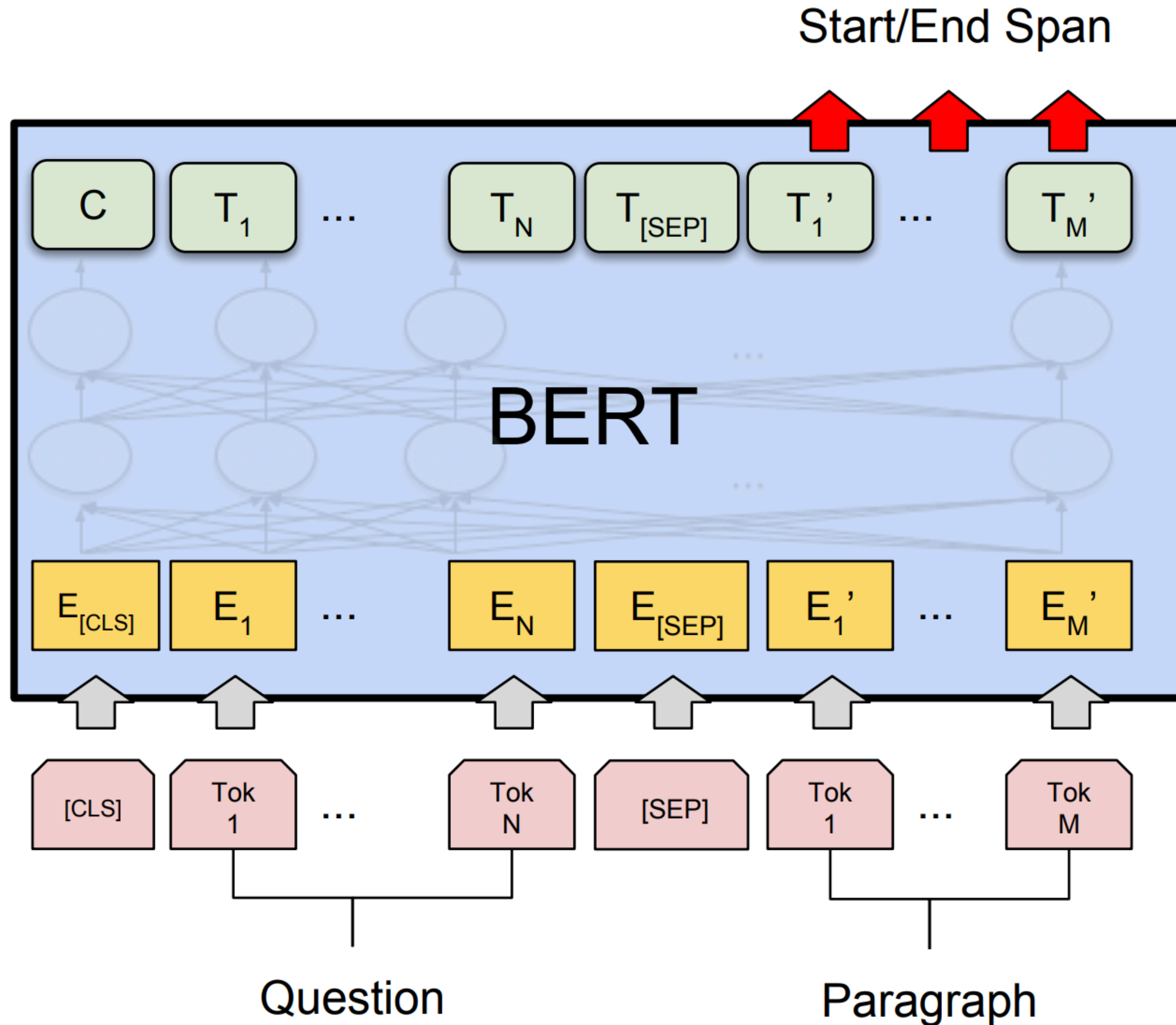
What does Macduff do to Macbeth?

What violent act does Macduff perform upon Macbeth?

- ▶ All questions have the same answer. But some are more easily guessable



Reminder: QA with BERT





QA: Answer Type Heuristics

What degree did Martin Luther receive on October 19, 1512?

On October 19, 1512, Luther was awarded his doctorate of theology and, on October 21, 1512, was received into the senate of the theological faculty of the University of Wittenberg. He spent the rest of his career in this position at the University of Wittenberg.

- ▶ What should the model be doing? Corresponding Martin Luther with Luther, matching October 19, 1512 between question and passage



QA: Answer Type Heuristics

What degree did Martin Luther receive?

What degree ____?

On October 19, 1512, Luther was awarded his doctorate of theology and, on October 21, 1512, was received into the senate of the theological faculty of the University of Wittenberg. He spent the rest of his career in this position at the University of Wittenberg.

- ▶ Only one possible degree here! Model only needs to see “what degree” and will not learn to use the rest of the context!



QA: Answer Type Heuristics

- ▶ Question type is powerful indicator. Only a couple of locations in this context!

Where _____?

On October 19, 1512, Luther was awarded his doctorate of theology and, on October 21, 1512, was received into the senate of the theological faculty of the University of Wittenberg. He spent the rest of his career in this position at the University of Wittenberg.

Who _____?

When _____?



QA: Answer Type Heuristics

- ▶ Question type is powerful indicator. Only a couple of locations in this context!

Where _____? Who _____? When _____?

On October 19, 1512, Luther was awarded his doctorate of theology and, on October 21, 1512, was received into the senate of the theological faculty of the University of Wittenberg. He spent the rest of his career in this position at the University of Wittenberg.

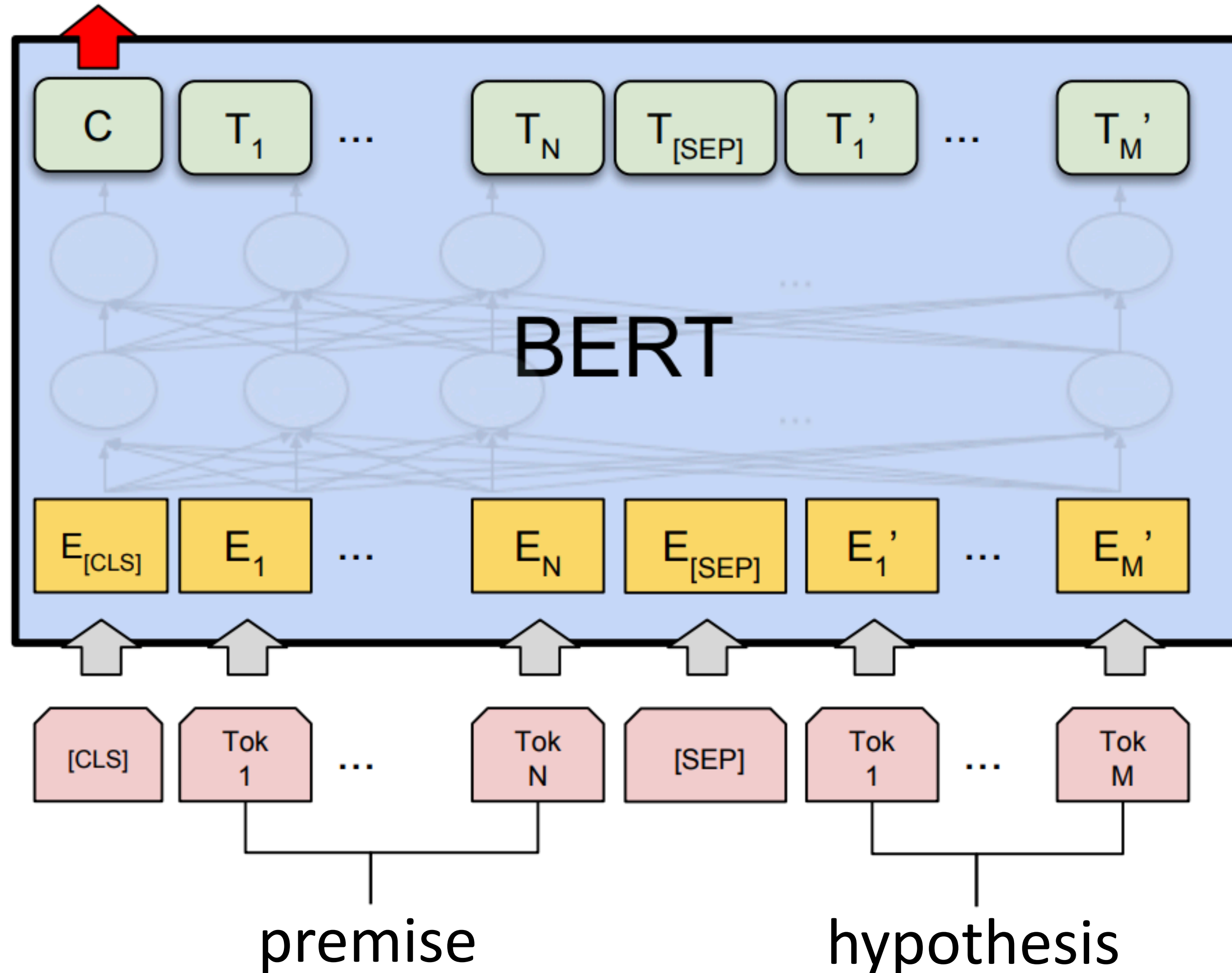
- ▶ What will happen if we train on this data?
 - ▶ Will loss decrease?
 - ▶ How will the model learn to “behave”?

Annotation Artifacts, Reasoning Shortcuts: NLI



Reminder: NLI with BERT

entailed/neutral/contradiction





NLI: Hypothesis-only Baselines

Premise: *A woman on a deck is selling bamboo sticks.*

Label?

Hypothesis: *A man is selling bamboo sticks*

Hypothesis: *A man is juggling flaming chainsaws*

Hypothesis: *Eighteen flying monkeys are in outer space*

- ▶ Not all of these things have the same likelihood of being true a priori
- ▶ What might the model learn to do in this case?



NLI: Hypothesis-only Baselines

Premise	A woman selling bamboo sticks talking to two men on a loading dock.
Entailment	There are at least three people on a loading dock.
Neutral	A woman is selling bamboo sticks to help provide for her family.
Contradiction	A woman is not taking money for any of her sticks.

- ▶ What's different about this neutral sentence?
 - ▶ To create neutral sentences: annotators *add information*
- ▶ What's different about this contradictory sentence?
 - ▶ To create contradictions: annotators *add negation*
- ▶ These are not broadly representative of what can happen in other settings. There is no “natural” distribution of NLI, but this is still very restrictive



NLI: Hypothesis-only Baselines

Premise	A woman selling bamboo sticks talking to two men on a loading dock.
Entailment	There are at least three people on a loading dock.
Neutral	A woman is selling bamboo sticks to help provide for her family .
Contradiction	A woman is not taking money for any of her sticks.

- ▶ Models can detect new information or negation easily
- ▶ Models can do very well *without looking at the premise*

Performance of models that only look at the hypothesis:
~70% on 3-class SNLI dataset

	Hyp-only model	Majority class	
SNLI	69.17	33.82	+35.35
MNLI-1	55.52	35.45	+20.07
MNLI-2	55.18	35.22	+19.96



NLI: Heuristics (HANS)

Heuristic	Definition	Example
Lexical overlap	Assume that a premise entails all hypotheses constructed from words in the premise	The doctor was paid by the actor. ————→ The doctor paid the actor. WRONG
Subsequence	Assume that a premise entails all of its contiguous subsequences.	The doctor near the actor danced. ————→ The actor danced. WRONG
Constituent	Assume that a premise entails all complete subtrees in its parse tree.	If the artist slept , the actor ran. ————→ The artist slept. WRONG

- ▶ Word overlap supersedes actual reasoning in these cases
- ▶ They create a test set (HANS) consisting of cases where heuristics like word overlap are misleading. Very low performance



Evidence of Spurious Correlations: Contrast Sets

- ▶ How do we control for annotation artifacts? Things like “premises and hypotheses overlap too much” aren’t easy to see!
- ▶ For any particular effect like lexical overlap, we could try to annotate data that “breaks” that effect
- ▶ Issue: breaking one correlation may just result in another one surfacing. How do we “break” them all at the same time?
- ▶ Solution: construct new examples through *minimal edits that change the label*.



Evidence of Spurious Correlations: Contrast Sets

Hardly one to be faulted for his ambition or his vision, it is genuinely unexpected, then, to see all Park's effort add up to so very little. ... The premise is promising, gags are copious and offbeat humour abounds but it all fails miserably to create any meaningful connection with the audience.

(Label: Negative)

Hardly one to be faulted for his ambition or his vision, **here we see all Park's effort come to fruition.** ... The premise is **perfect**, gags are **hilarious** and offbeat humour abounds, **and it creates a deep** connection with the audience.

(Label: Positive)

- ▶ By minimally editing an example, we control for pretty much all of the possible shortcuts that apply to the original.
- ▶ E.g., [summary starts with "*Hardly*" -> negative] is a pattern that could not hold anymore



Evidence of Spurious Correlations: Contrast Sets

Dataset	# Examples	# Sets	Model	Original Test		Contrast
NLVR2	994	479	LXMERT	76.4	61.1	(-15.3)
IMDb	488	488	BERT	93.8	84.2	(-9.6)
MATRES	401	239	CogCompTime2.0	73.2	63.3	(-9.9)
UD English	150	150	Biaffine + ELMo	64.7	46.0	(-18.7)
PERSPECTRUM	217	217	RoBERTa	90.3	85.7	(-4.6)
DROP	947	623	MTMSN	79.9	54.2	(-25.7)

Solutions



Broad Solutions

- ▶ Most solutions involve changing what data is trained on
 - ▶ Subset of data
 - ▶ Soft subset (i.e., reweight the existing examples)
 - ▶ Superset: add adversarially-constructed data, contrast sets, etc.
- ▶ For subsets: what do we train on?
 - ▶ Don't train on stuff that allows you to cheat
 - ▶ Train on examples that teach the real task rather than shortcuts



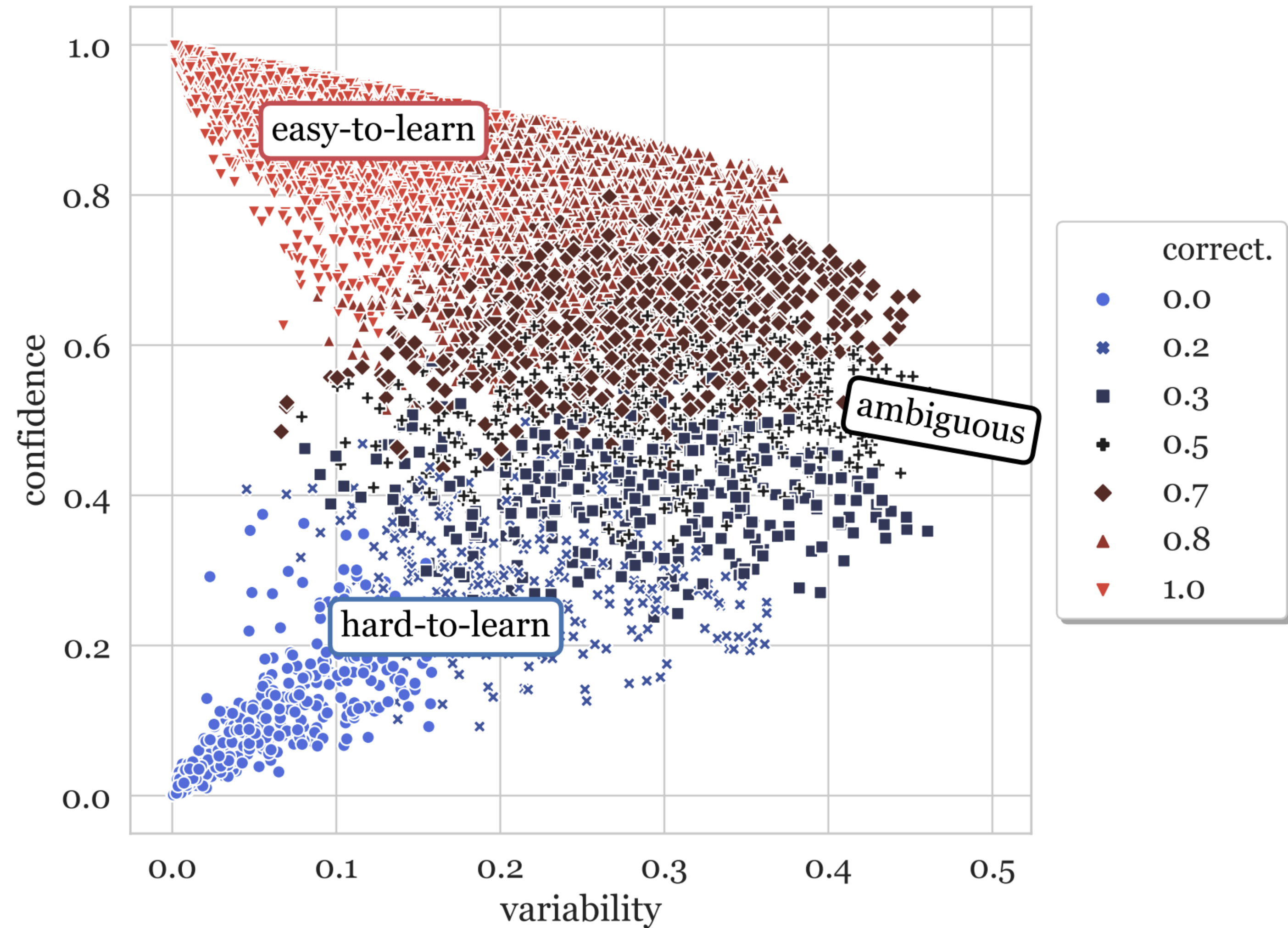
Dataset Cartography

- ▶ What happens with each particular example during training?
- ▶ Spurious correlations are *easy to learn*: a model should learn these early and always get them right
- ▶ Imagine a very challenging example
 - ▶ Model prediction may change a lot as it learns this example, may be variable in its predictions
- ▶ Imagine a mislabeled example
 - ▶ Probably just always wrong unless it gets overfit



Data Maps

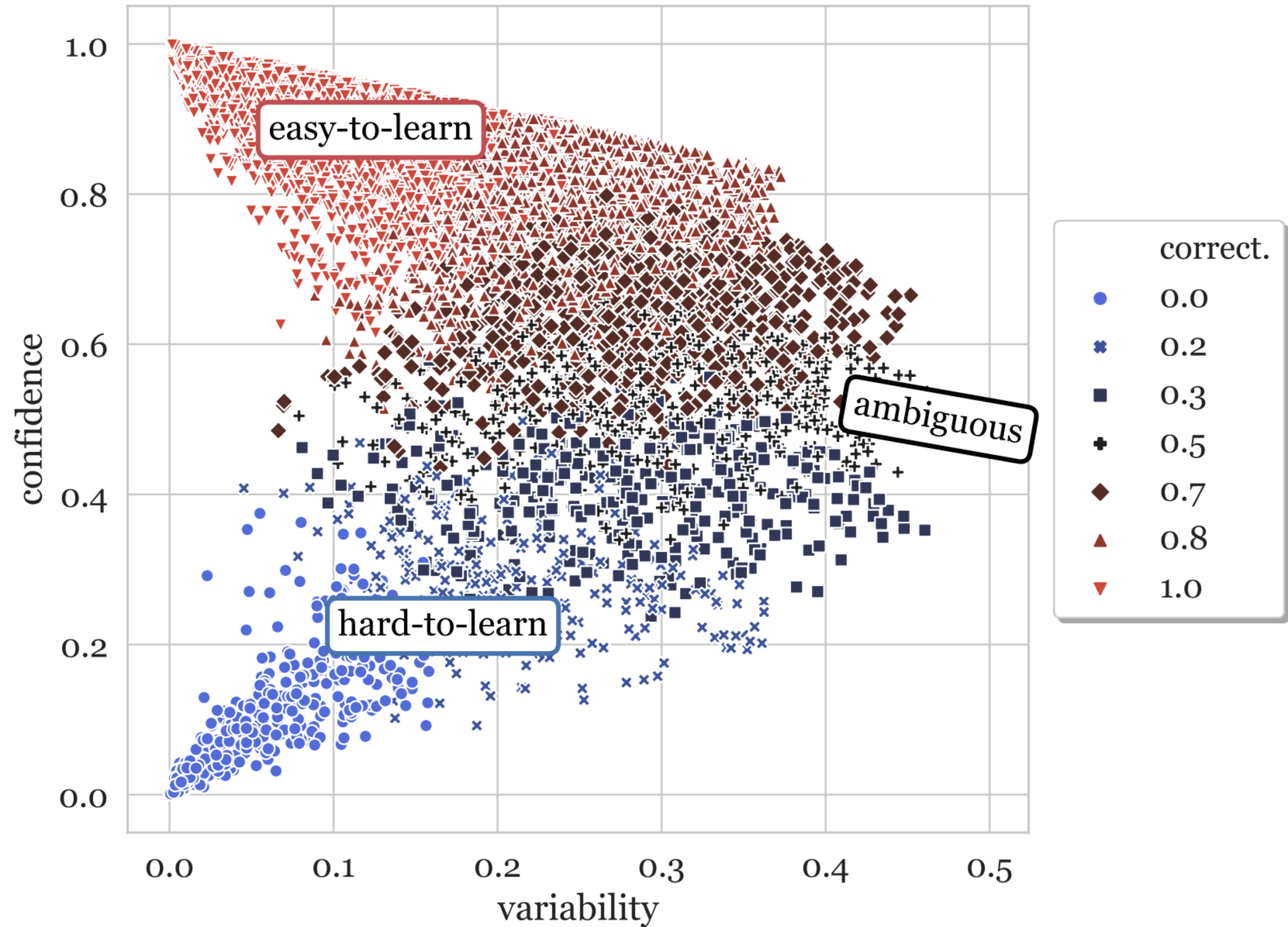
- ▶ Confidence: mean probability of correct label
- ▶ Variability: standard deviation in probability of the correct label
- ▶ Ambiguous examples: possible learnable (model knows it sometimes but not other times), but hard!





Data Maps

- ▶ What to do with them?
- ▶ Training on hard-to-learn or ambiguous examples leads to better performance out-of-domain





Debiasing

- ▶ Other ways to identify easy examples other than data maps
- ▶ Train some kind of a weak model and discount examples that it fits easily

$$\mathcal{L}(\theta_d) = - \left(1 - p_b^{(i,c)}\right) y^{(i)} \cdot \log p_d$$

one-hot label vector

log probability of each label

probability under a copy of the model trained for a few epochs on a small subset of data (bad model)



Debiasing

Method	MNLI (Acc.)		
	dev	<i>HANS</i>	Δ
BERT-base	84.5	61.5	-
Reweighting <small>known-bias</small>	83.5 [‡]	69.2 [‡]	+7.7
Reweighting <small>self-debias</small>	81.4	68.6	+7.1
Reweighting <small>♠ self-debias</small>	82.3	69.7	+8.2

- ▶ On the challenging HANS test set for NLI, this debiasing improves performance substantially
- ▶ In-domain MNLI performance goes down



Debiasing

- ▶ Other work has explored similar approaches using a known bias model

$$\hat{p}_i = \textit{softmax}(\log(p_i) + \log(b_i))$$

probabilities from learned bias model — like the weak model from Utama et al. (prev. slides), but you define its structure

- ▶ *Ensembles* the weak model with the model you actually learn.
- ▶ Your actual model learns the *residuals* of the weak model: the difference between the weak model's output distribution and the target distribution.
- ▶ This lets it avoid learning the weak model's biases!



Core Principles

- ▶ By reweighting data or changing the training paradigm, you can learn a model that generalizes better
- ▶ Most gains will show up **out-of-domain**. Very hard to get substantial improvements on the same dataset, unless you consider small subsets of examples (e.g., the toughest 1% of examples by some measure)

Final Project
(see spec and GitHub)