CS 11-747 Neural Networks for NLP

Model Interpretation

Danish Pruthi

March 11*, 2021

*or whenever you watch this

Why interpretability?

Example from Caruana et al.

Why interpretability?

- **Task:** predict probability of death for patients with pneumonia

- **Why**: so that high-risk patients can be admitted, low risk patients can be treated as outpatients

Example from Caruana et al.

Why interpretability?

- **Task:** predict probability of death for patients with pneumonia
- **Why**: so that high-risk patients can be admitted, low risk patients can be treated as outpatients
- $\text{AUC}_{\text{Neural networks}} > \text{AUC}_{\text{Logistic Regression}}$

Example from Caruana et al.

Why interpretability?

- **Task:** predict probability of death for patients with pneumonia

- **Why**: so that high-risk patients can be admitted, low risk patients can be treated as outpatients

- $AUC_{\text{Neural networks}} > AUC_{\text{Logistic Regression}}$

- Rule based classifier

  HasAsthma(X) —> LowerRisk(X) 😕

Example from Caruana et al.

Why interpretability?

- **Task:** predict probability of death for patients with pneumonia

- **Why**: so that high-risk patients can be admitted, low risk patients can be treated as outpatients

- $\text{AUC}_{\text{Neural networks}} > \text{AUC}_{\text{Logistic Regression}}$

- Rule based classifier

HasAsthma(X) —> LowerRisk(X)

more intensive care

Example from Caruana et al.

Why interpretability?

Why interpretability?

- **Debug models** to uncover (and subsequently fix) issues, biases

Why interpretability?

- **Debug models** to uncover (and subsequently fix) issues, biases
- **Engender trust** among stakeholders

Why interpretability?

- **Debug models** to uncover (and subsequently fix) issues, biases
- **Engender trust** among stakeholders
- **Comply with regulations** which mandate "right to explanations"

Why interpretability?

- **Debug models** to uncover (and subsequently fix) issues, biases
- **Engender trust** among stakeholders
- **Comply with regulations** which mandate "right to explanations"
- **Assess robustness**: how will the model perform *in the wild*

Why interpretability?

- **Debug models** to uncover (and subsequently fix) issues, biases
- **Engender trust** among stakeholders
- **Comply with regulations** which mandate "right to explanations"
- **Assess robustness**: how will the model perform *in the wild*
- **Provide recourse**

Why interpretability?

- **Debug models** to uncover (and subsequently fix) issues, biases
- **Engender trust** among stakeholders
- **Comply with regulations** which mandate "right to explanations"
- **Assess robustness**: how will the model perform *in the wild*
- **Provide recourse**
- **Identify causal factors** behind the predictions
- and more….

The why question

Why did a model make a certain prediction for a given example?

The why question

Why did a model make a certain prediction for a given example?

$$f(\mathbf{x}) = w_1\mathbf{x}_1 + w_2\mathbf{x}_2$$

The why question

Why did a model make a certain prediction for a given example?

$$f(\mathbf{x}) = w_1\mathbf{x}_1 + w_2\mathbf{x}_2$$

Estimates Rent

House

Area

Population Density

The why question

Why did a model make a certain prediction for a given example?

$$f(\mathbf{x}) = \overset{2.0}{w_1}\mathbf{x}_1 + \overset{0.01}{w_2}\mathbf{x}_2$$

Estimates Rent

House

Area

Population Density

The why question

Why did a model make a certain prediction for a given example?

$$\underset{\text{Estimates Rent}}{f}(\underset{\text{House}}{\mathbf{x}}) = \overset{2.0}{w_1}\,\underset{\text{Area}}{\mathbf{x}_1} + \overset{0.01}{w_2}\,\underset{\text{Population Density}}{\mathbf{x}_2}$$

- How the answer is computed? (mechanistic details)

The why question

Why did a model make a certain prediction for a given example?

$$\underbrace{f(\underset{\uparrow\ \text{House}}{\mathbf{x}})}_{\text{Estimates Rent}} = \overset{2.0}{w_1}\underset{\uparrow\ \text{Area}}{\mathbf{x}_1} + \overset{0.01}{w_2}\underset{\uparrow\ \text{Population Density}}{\mathbf{x}_2}$$

- How the answer is computed? (mechanistic details)
- Relative importance of each feature?

The why question

Why did a model make a certain prediction for a given example?

$$\underset{\text{Estimates Rent}}{f}(\underset{\text{House}}{\mathbf{x}}) = \overset{2.0}{w_1}\underset{\text{Area}}{\mathbf{x}_1} + \overset{0.01}{w_2}\underset{\text{Population Density}}{\mathbf{x}_2}$$

- How the answer is computed? (mechanistic details)
- Relative importance of each feature?
- How did we end up with these parameters?

The why question

Why did a model make a certain prediction for a given example?

$$\underset{\text{Estimates Rent}}{f}(\underset{\text{House}}{\mathbf{x}}) = \overset{2.0}{w_1}\underset{\text{Area}}{\mathbf{x}_1} + \overset{0.01}{w_2}\underset{\text{Population Density}}{\mathbf{x}_2}$$

- How the answer is computed? (mechanistic details)
- Relative importance of each feature?
- How did we end up with these parameters?
  - What was the training objective?

The why question

Why did a model make a certain prediction for a given example?

$$f(\mathbf{x}) = \overset{2.0}{w_1}\mathbf{x}_1 + \overset{0.01}{w_2}\mathbf{x}_2$$

Estimates Rent

House

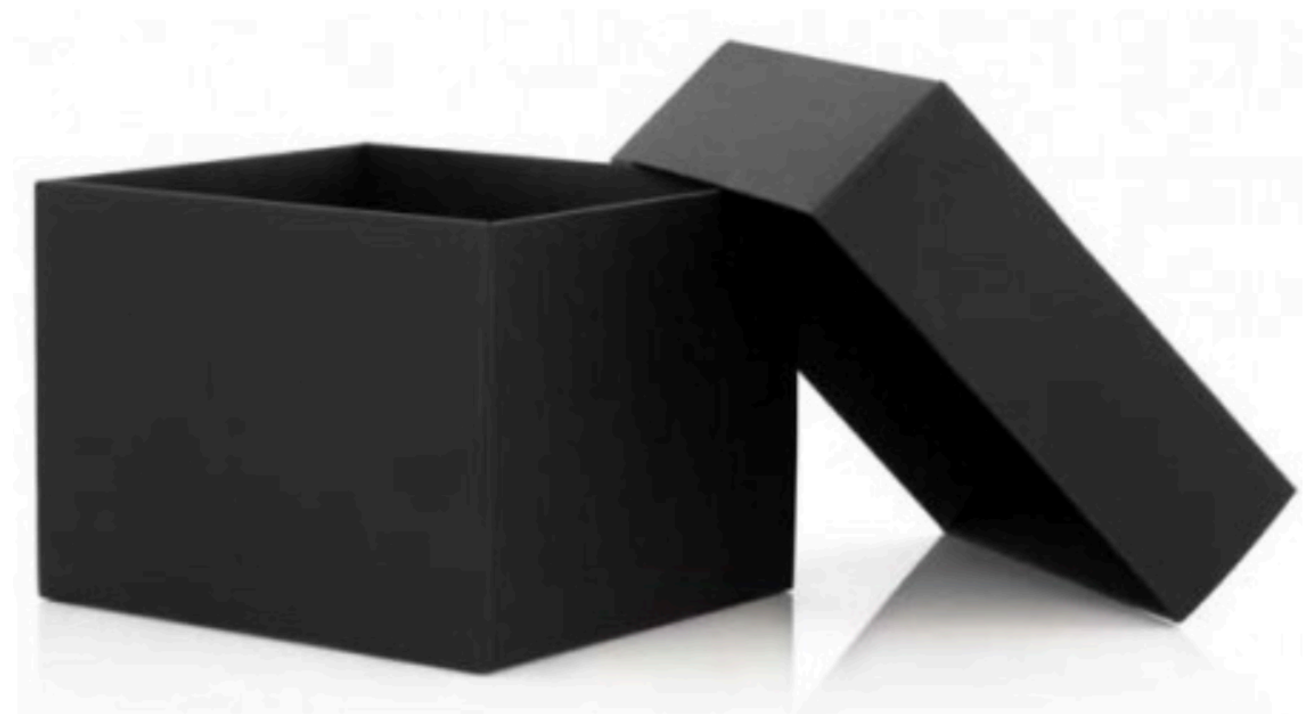
Area

Population Density

- How the answer is computed? (mechanistic details)
- Relative importance of each feature?
- How did we end up with these parameters?
    - What was the training objective?
    - What was the data? Which city? Is it representative?

Two broad themes

- What is the model learning?
- Can we explain the prediction?



Two broad themes

global interpretation

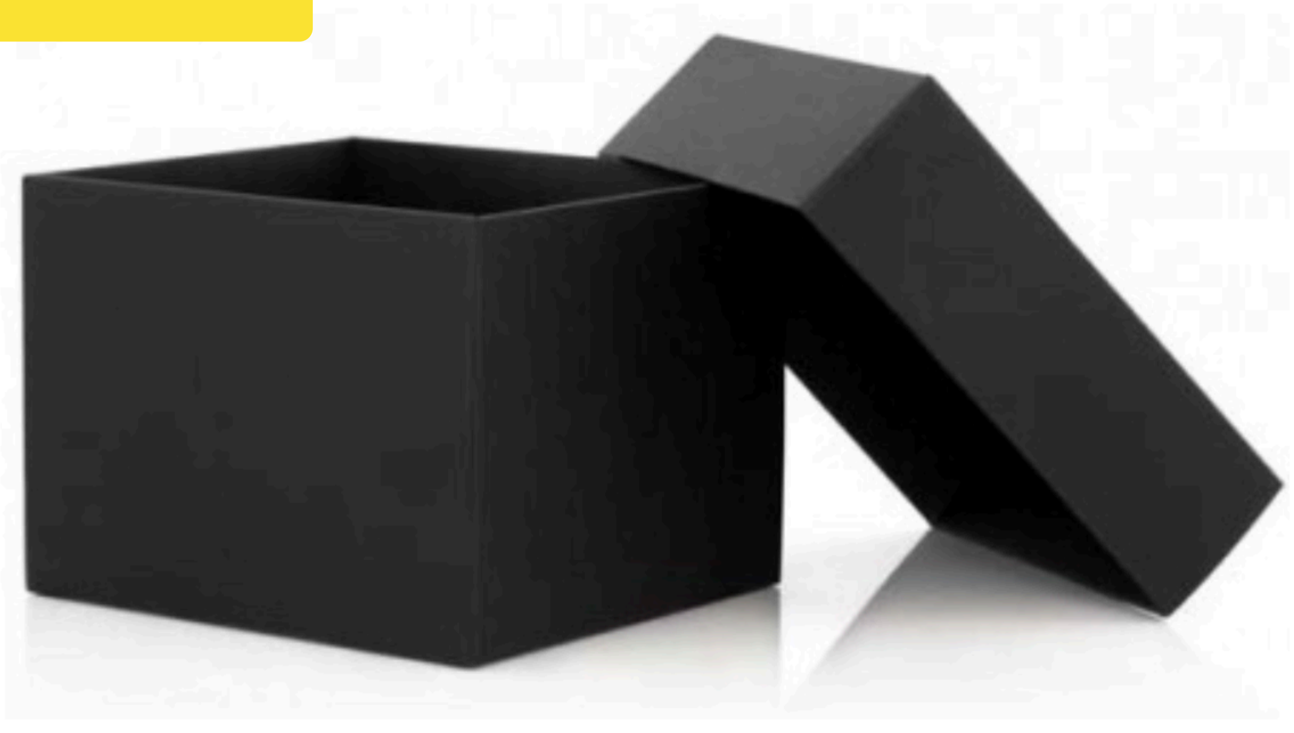
- What is the model learning?

- Can we explain the prediction?

Two broad themes

global interpretation

- What is the model learning?



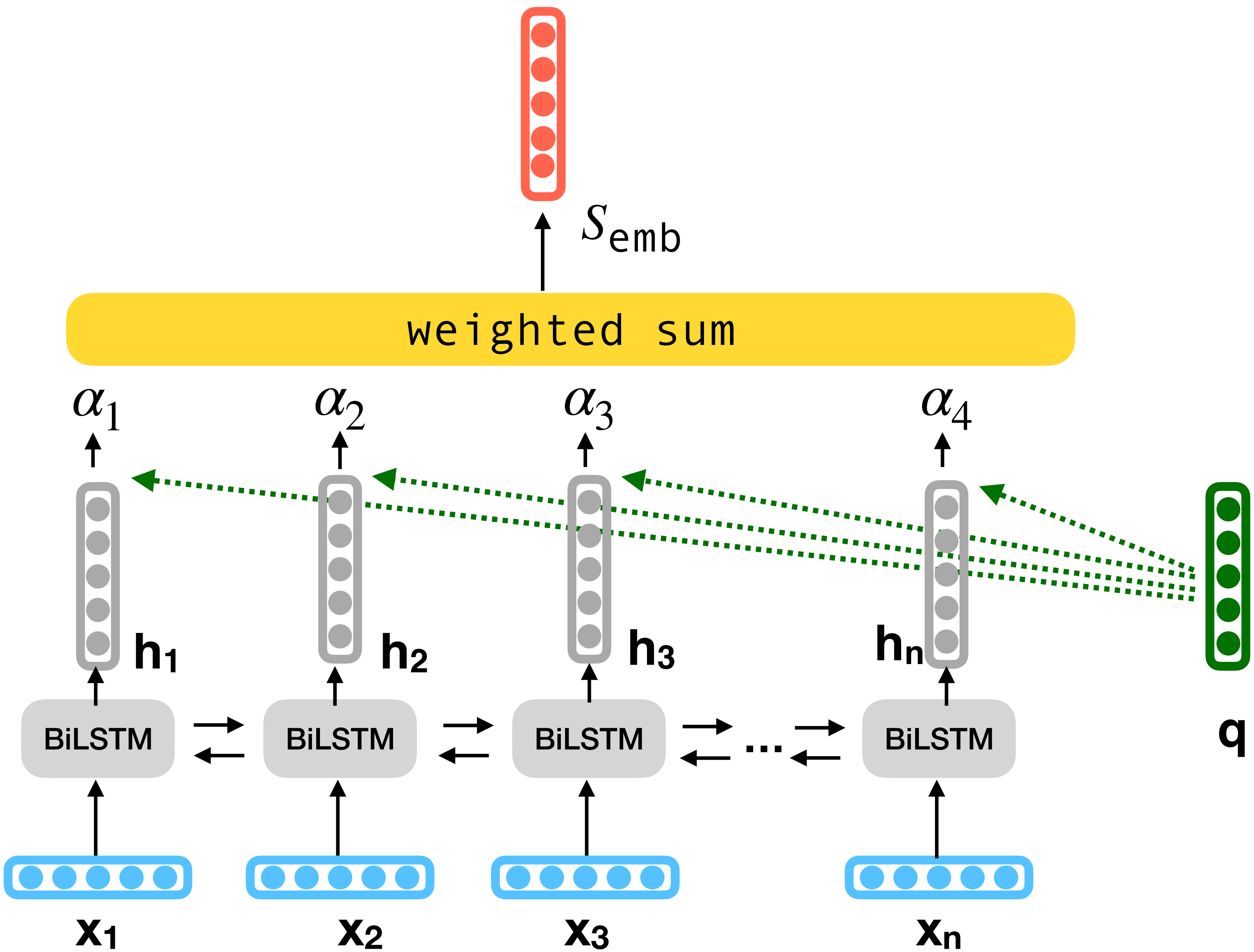
- Can we explain the prediction?

local interpretation

What is the model learning?

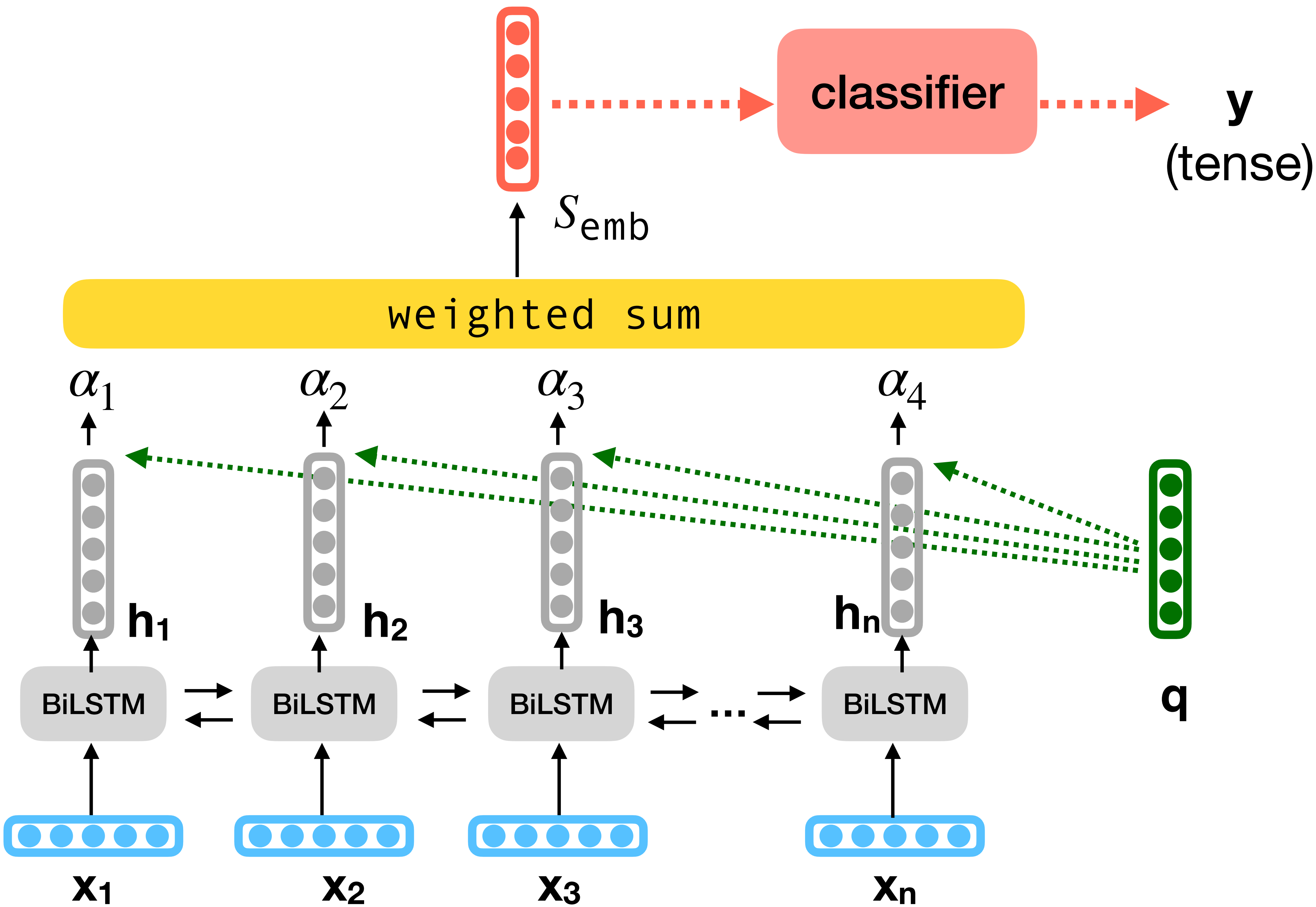
Probing

Probing



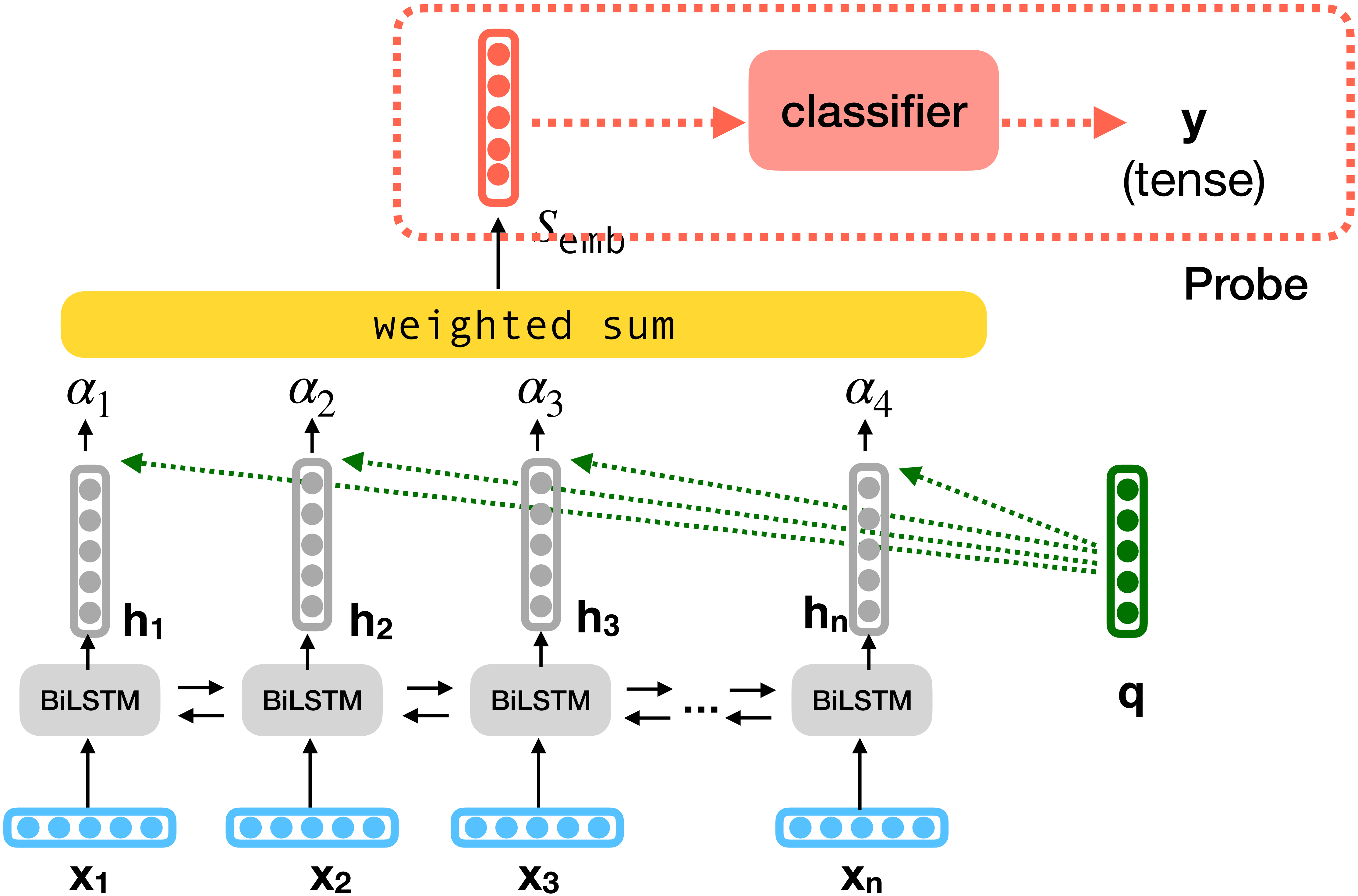
$s_{\mathtt{emb}}$
weighted sum
α_1
α_2
α_3
α_4
$\mathbf{h_1}$
$\mathbf{h_2}$
$\mathbf{h_3}$
$\mathbf{h_n}$
$\mathbf{q}$
BiLSTM
BiLSTM
BiLSTM
...
BiLSTM
$\mathbf{x_1}$
$\mathbf{x_2}$
$\mathbf{x_3}$
$\mathbf{x_n}$

Probing



classifier
y
(tense)
$s_{\mathtt{emb}}$
weighted sum
α_1
α_2
α_3
α_4
h1
h2
h3
hn
BiLSTM
BiLSTM
BiLSTM
BiLSTM
...
q
x1
x2
x3
xn

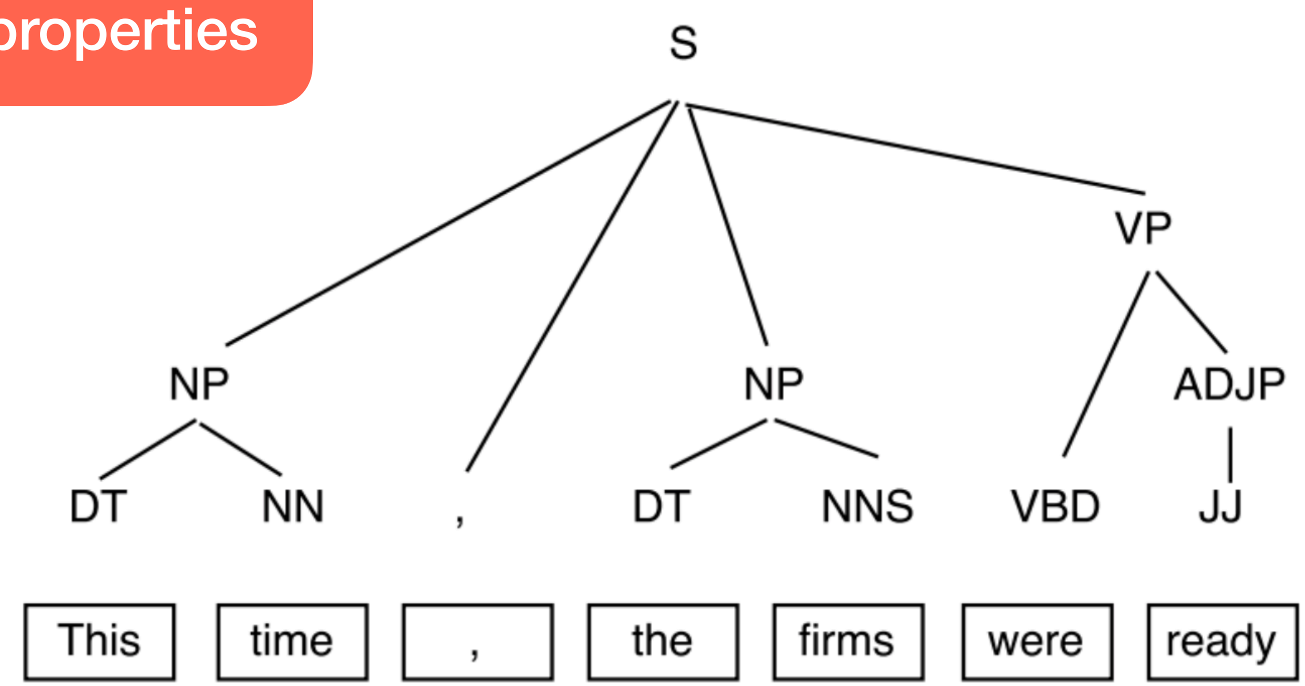
Probing



classifier
y
(tense)
s_{emb}
Probe
weighted sum
α_1
α_2
α_3
α_4
h1
h2
h3
hn
q
BiLSTM
BiLSTM
BiLSTM
...
BiLSTM
x1
x2
x3
xn

Source Syntax in NMT

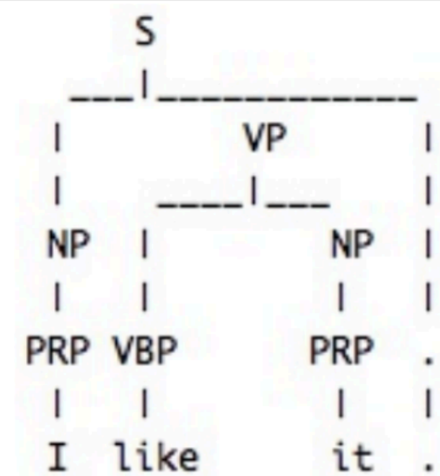
5 syntactic properties



POS	DT	NN	,	DT	NNS	VBD	JJ
Smallest Phrase Constituent	NP	NP	-	NP	NP	VP	ADJP
Top-level Syntactic Sequence	(NP, NP, VP)						
Voice	Active						
Tense	Past						

Does String-Based Neural MT Learn Source Syntax? Shi et al. EMNLP 2016

Source Syntax in NMT

<table>
<tr><th>Model</th><th>Source</th><th>Target</th></tr>
<tr><td>E2E</td><td>I like it .</td><td>I like it .</td></tr>
<tr><td>PE2PE</td><td>it I . like</td><td>it I . like</td></tr>
<tr><td>E2F</td><td>I like it .</td><td>J'aime ça.</td></tr>
<tr><td>E2G</td><td>I like it .</td><td>Ich mag das.</td></tr>
<tr><td>E2P</td><td>I like it .</td><td>

(S (NP PRP)$_{NP}$ (VP VBP (NP PRP)$_{NP}$)$_{VP}$.)$_{S}$</td></tr>
</table>

Figure 1: Sample inputs and outputs of the E2E, PE2PE, E2F, E2G, and E2P models.

Does String-Based Neural MT Learn Source Syntax? Shi et al. EMNLP 2016

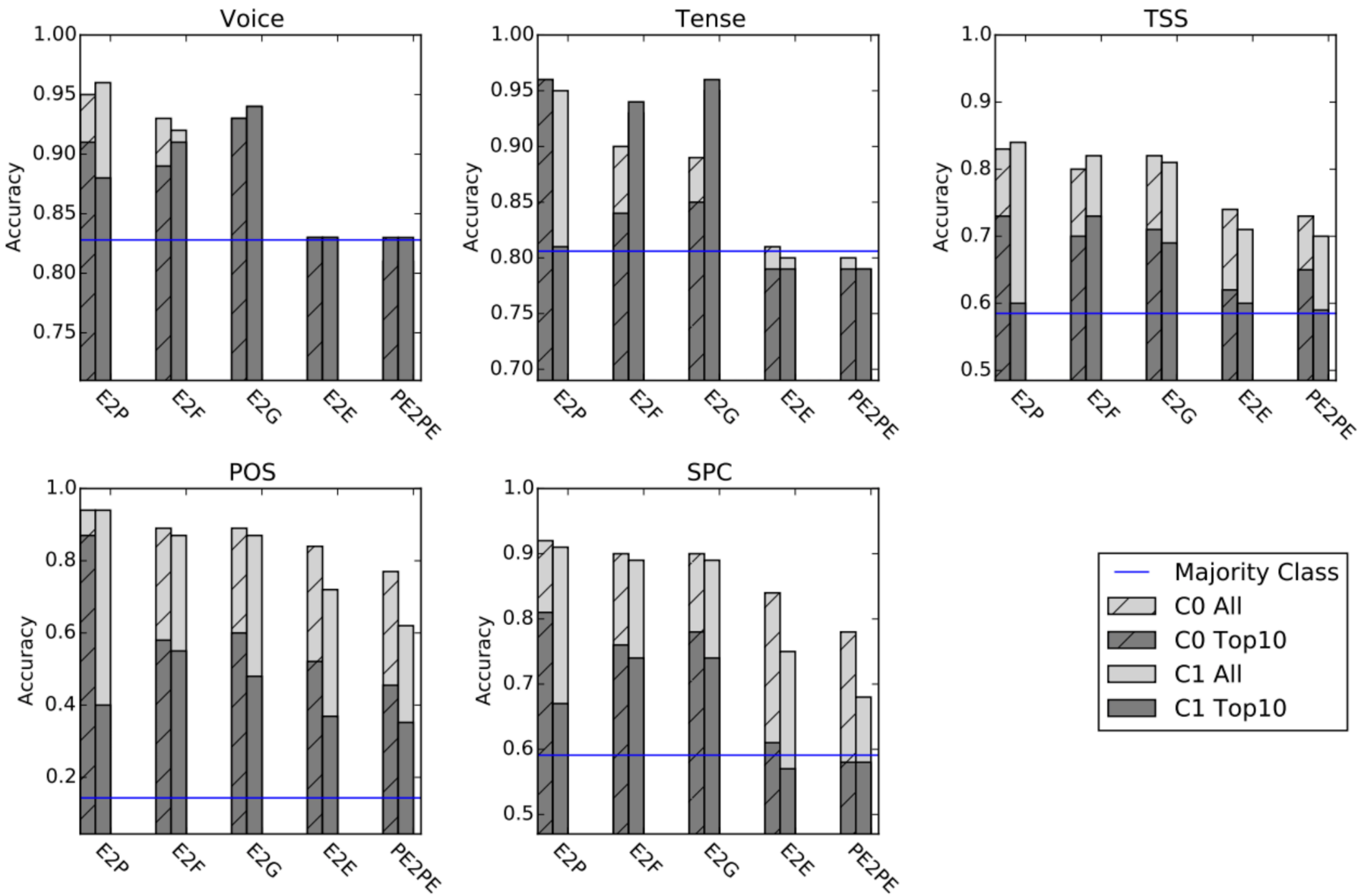
Source Syntax in NMT

Model	Accuracy
Majority Class	82.8
English to French (E2F)	92.8
English to English (E2E)	82.7

Table 1: Voice (active/passive) prediction accuracy using the encoding vector of an NMT system. The majority class baseline always chooses active.

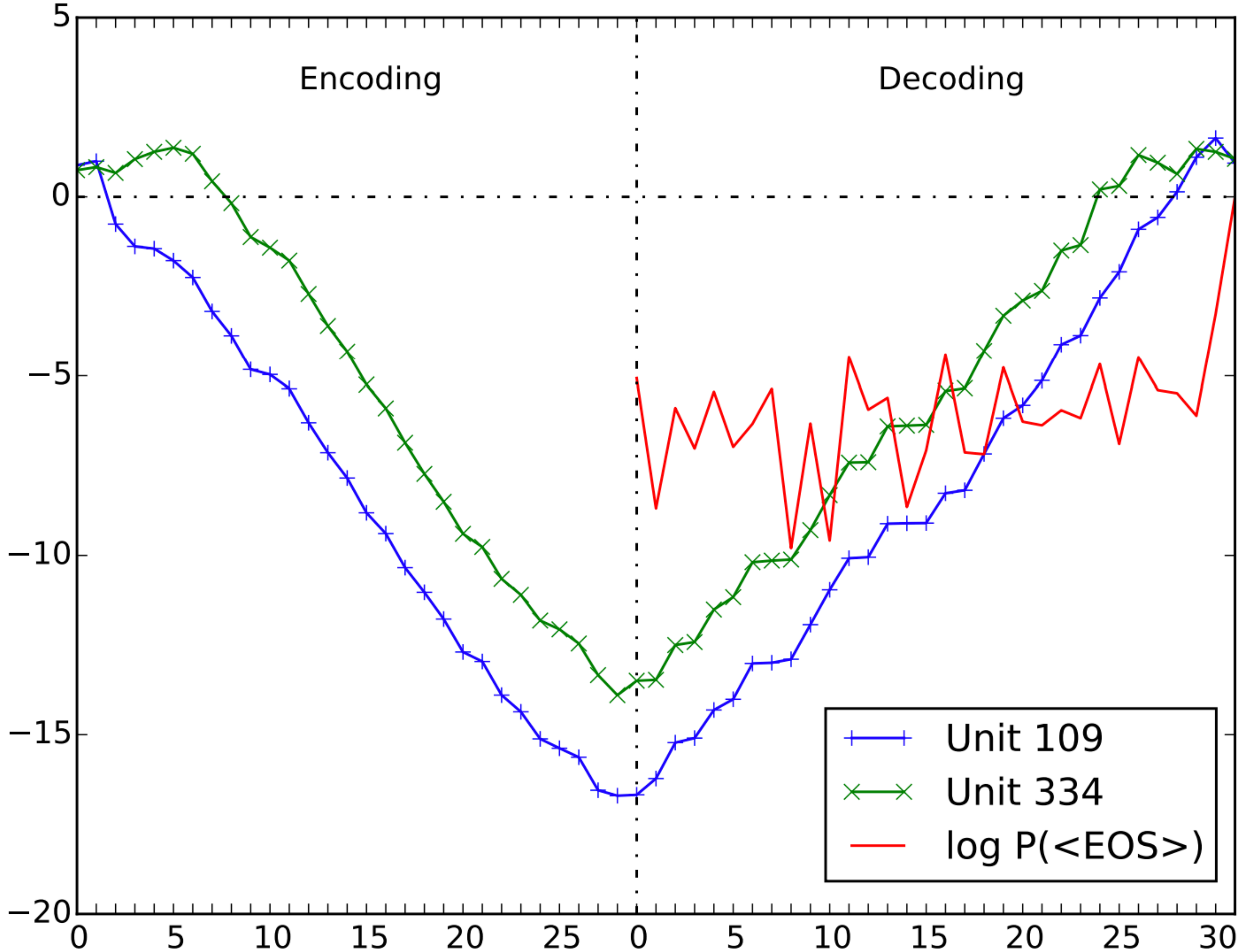
Does String-Based Neural MT Learn Source Syntax? Shi et al. EMNLP 2016

Source Syntax in NMT



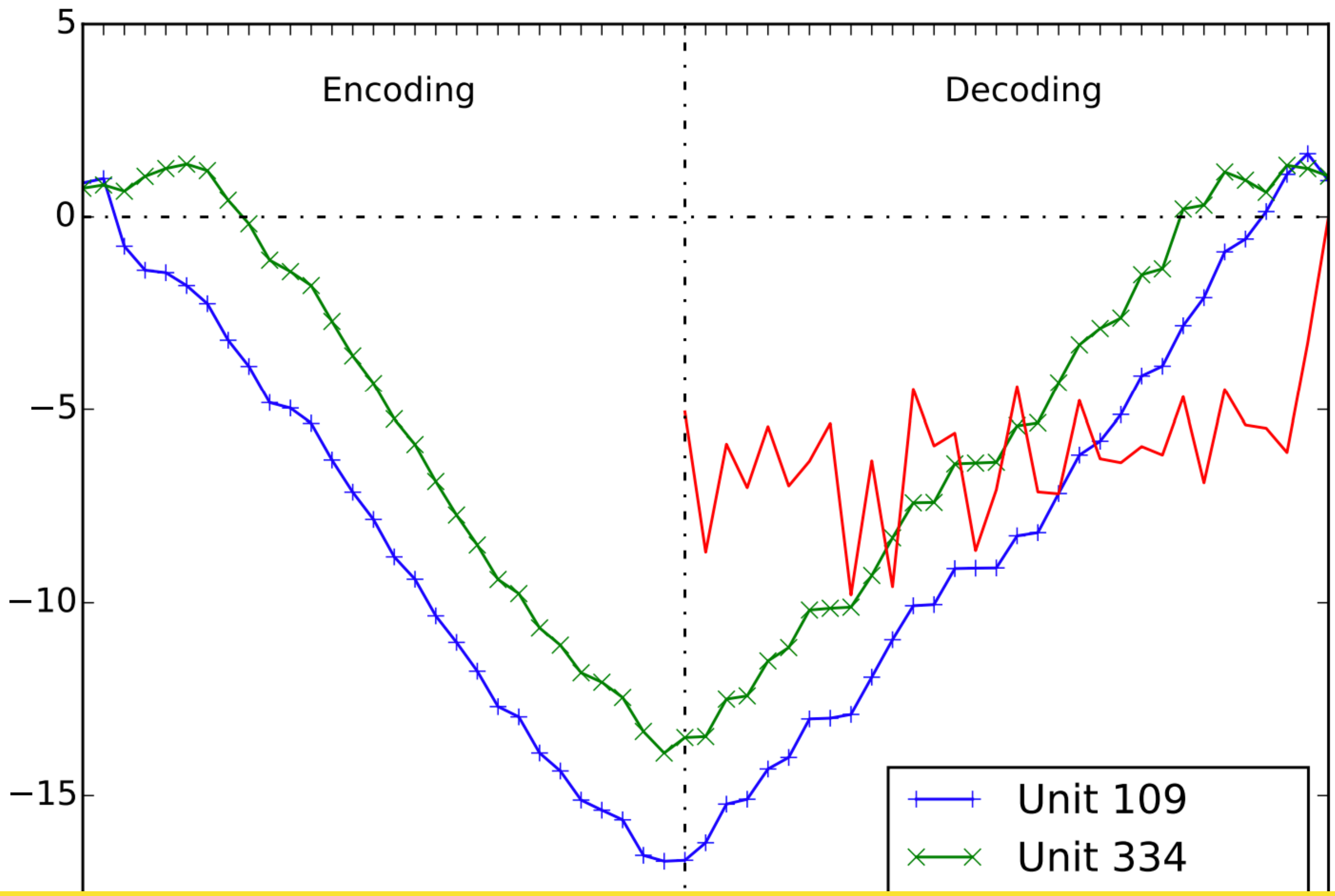
Does String-Based Neural MT Learn Source Syntax? Shi et al. EMNLP 2016

Why neural translations are the right length?



Shi et al. EMNLP 2016

Why neural translations are the right length?



Note: LSTMs can learn to count, whereas GRUs can not do unbounded counting (Weiss et al. ACL 2018)

Shi et al. EMNLP 2016

Issues with probing

- Did I interpret the representation or my probing classifier learn the task itself (Hewitt et al. 2019)
- Can only probe for properties you have supervision for
- Correlation doesn't imply causation
- and more…

[Probing Classifiers: Promises, Shortcomings, and Alternatives](#)
by Yonatan Belinkov

Summary: What is the model learning?

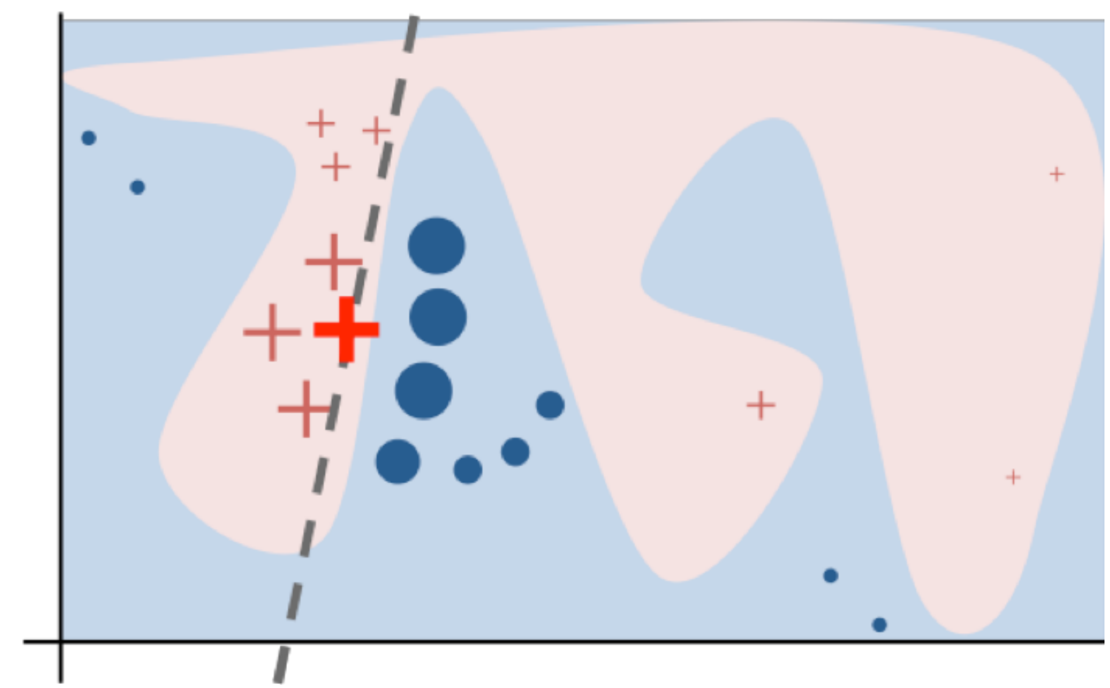
https://boknilev.github.io/nlp-analysis-methods/table1.html

Explain the prediction

Explanation Technique: LIME

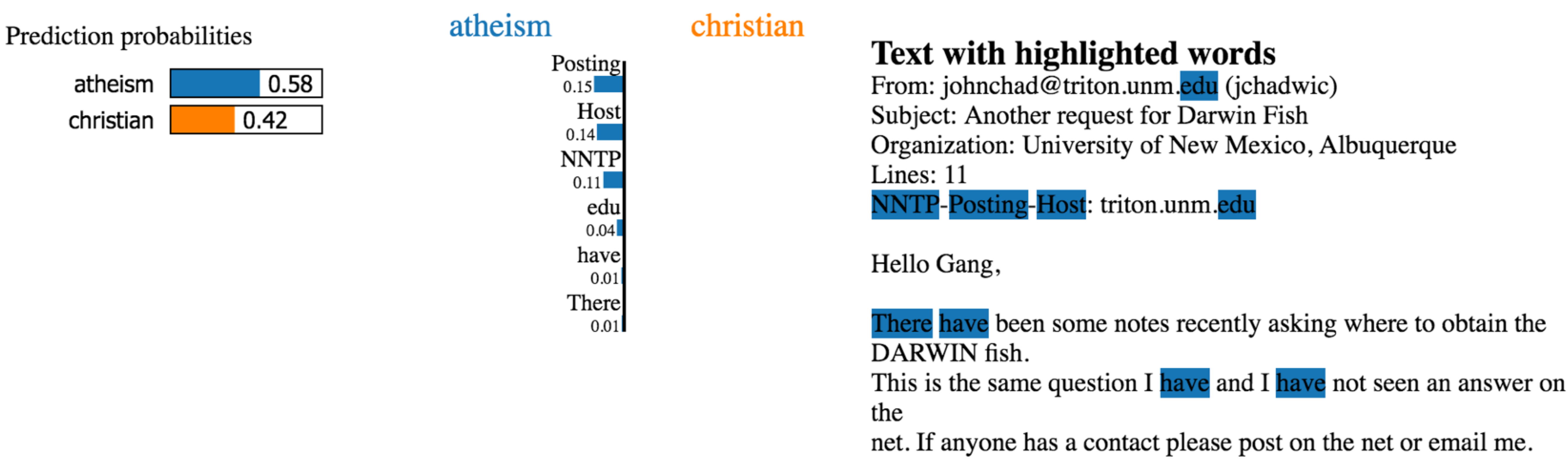
Ribeiro et al, KDD 2016

Explanation Technique: LIME



Ribeiro et al, KDD 2016

Explanation Technique: LIME



Ribeiro et al, KDD 2016

Explanation Techniques: gradient based importance scores

Method	Attribution $R_i^c(x)$
Gradient * Input	$x_i \cdot \frac{\partial S_c(x)}{\partial x_i}$
Integrated Gradient	$(x_i - \bar{x_i}) \cdot \int_{\alpha=0}^{1} \frac{\partial S_c(\tilde{x})}{\partial (\tilde{x_i})}\Big\|_{\tilde{x}=\bar{x}+\alpha(x-\bar{x})} d\alpha$
ϵ-LRP	$x_i \cdot \frac{\partial^g S_c(x)}{\partial x_i}, \quad g = \frac{f(z)}{z}$
DeepLIFT	$(x_i - \bar{x_i}) \cdot \frac{\partial^g S_c(x)}{\partial x_i}, \quad g = \frac{f(z) - f(\bar{z})}{z - \bar{z}}$

Example of attributions on MNIST

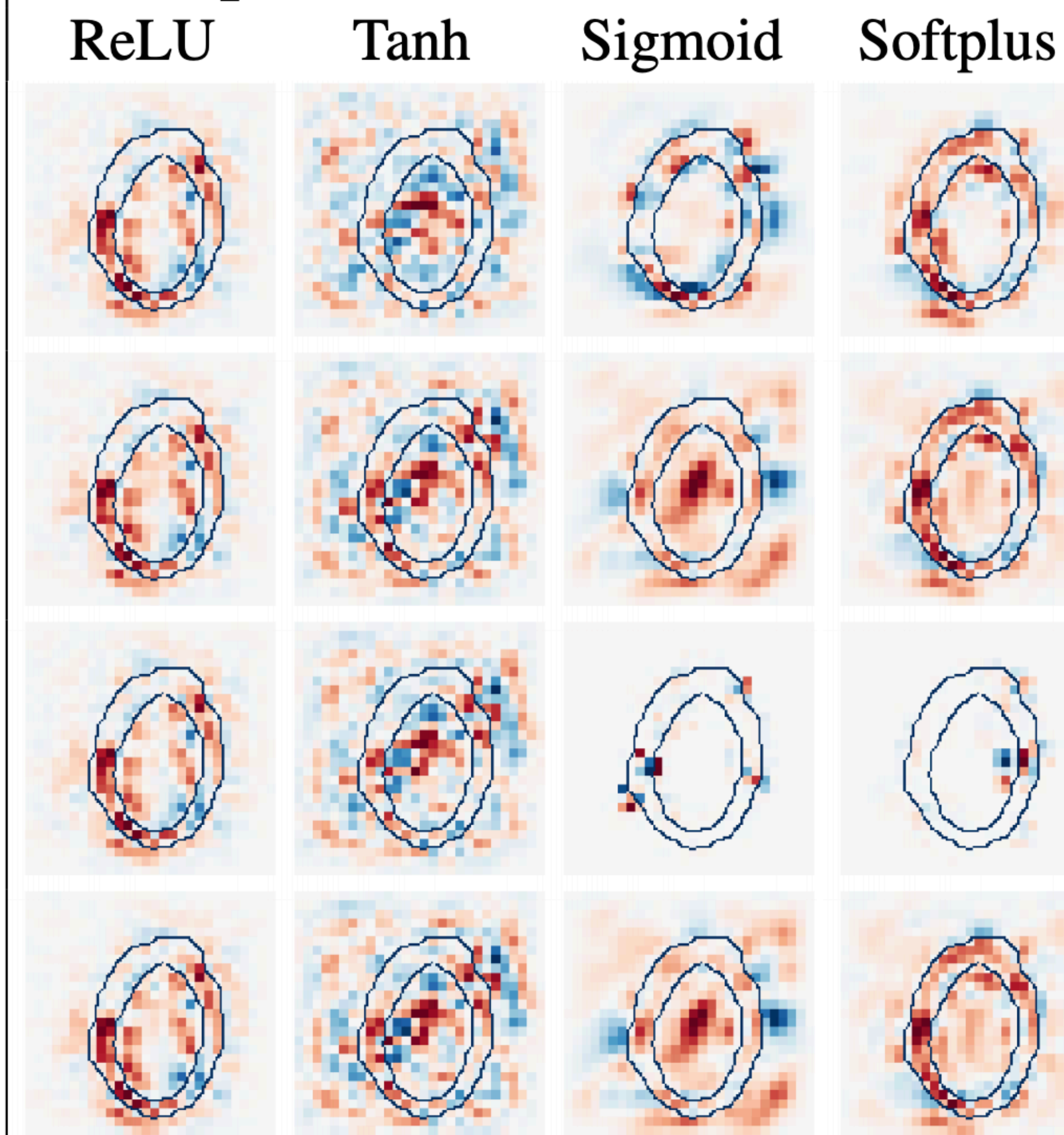
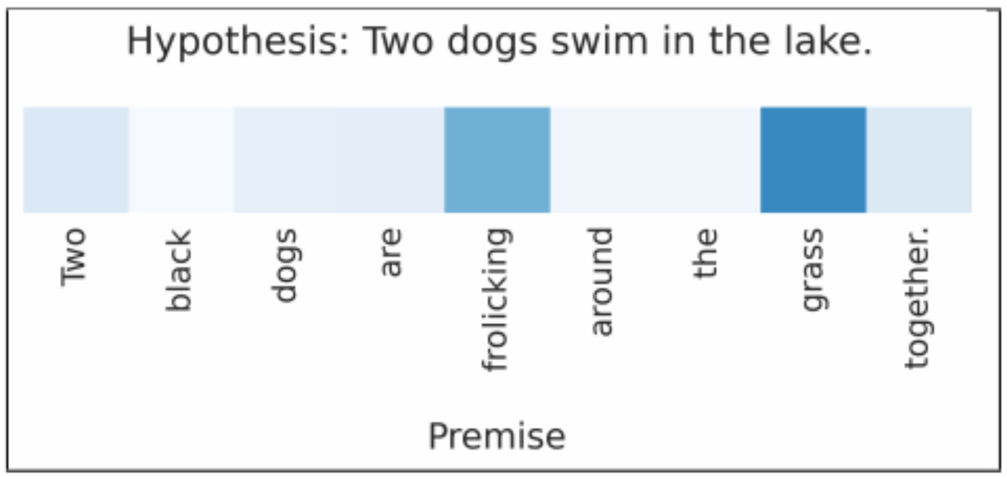


Figure from Ancona et al, ICLR 2018

Explanation Technique: Attention

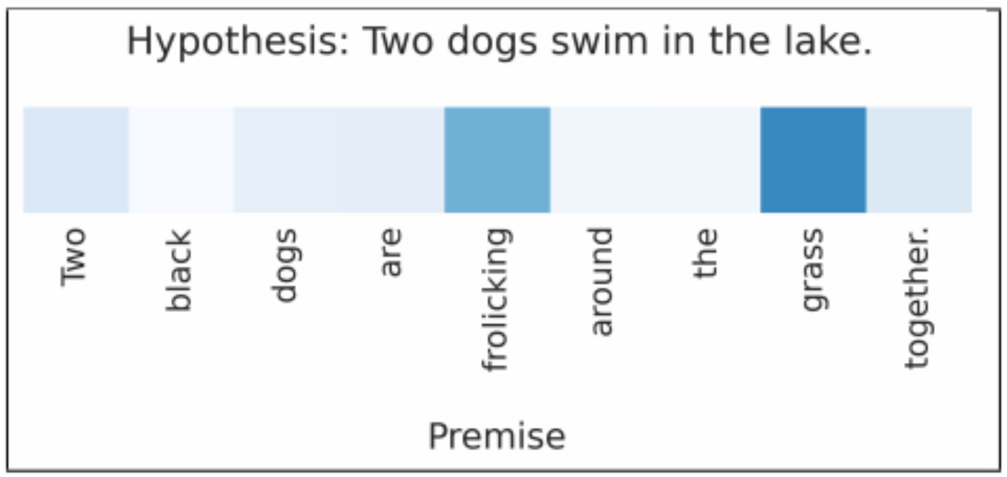
Explanation Technique: Attention



Entailment

Rocktäschel et al, 2015

Explanation Technique: Attention

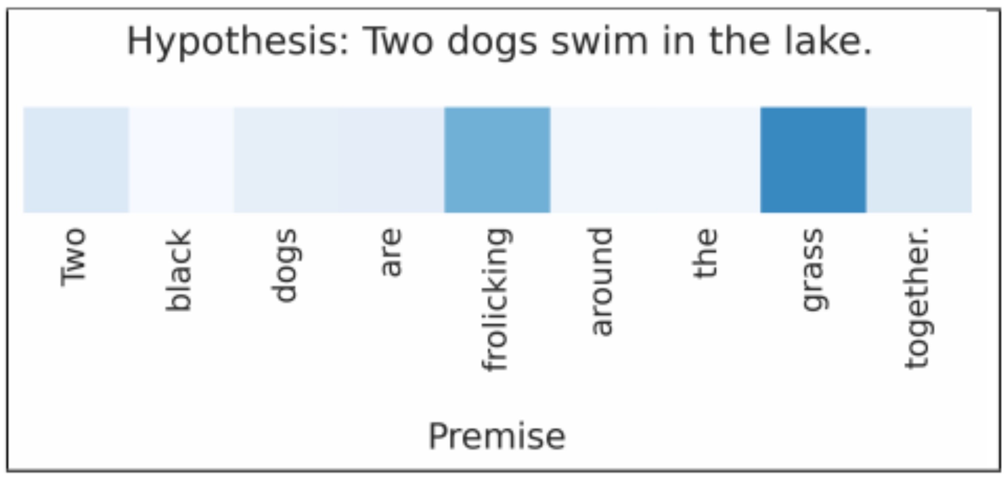


Entailment
Rocktäschel et al, 2015

A stop sign is on a road with a mountain in the background.

Image captioning
Xu et al, 2015

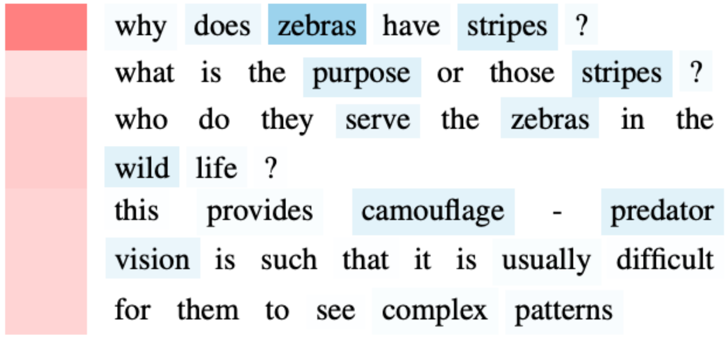
Explanation Technique: Attention



Entailment
Rocktäschel et al, 2015

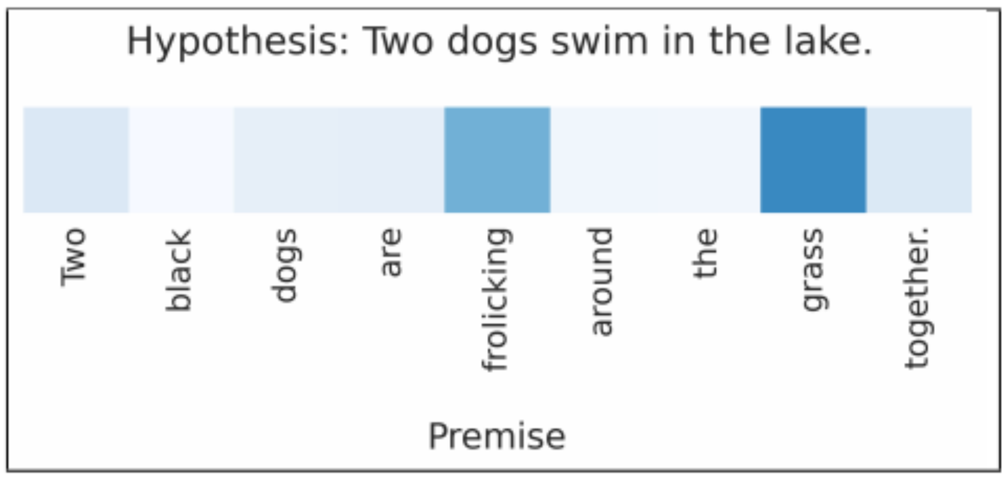
A stop sign is on a road with a mountain in the background.

Image captioning
Xu et al, 2015

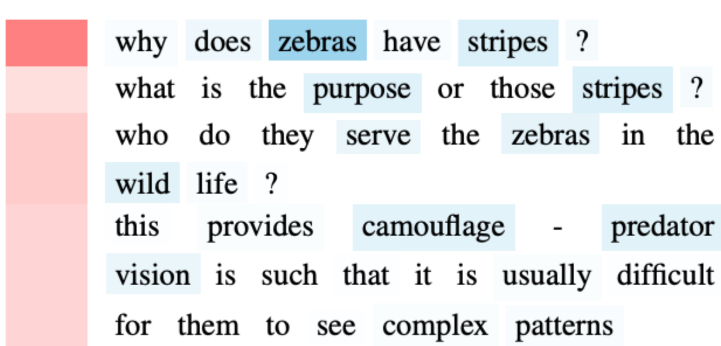


Document classification
Yang et al, 2016

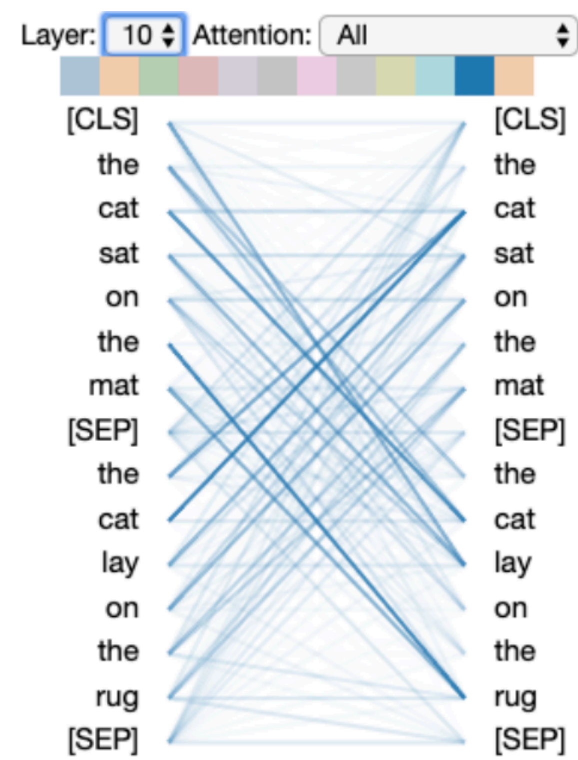
Explanation Technique: Attention



Entailment
Rocktäschel et al, 2015



Document classification
Yang et al, 2016



BERTViz
Vig et al, 2019

Explanation Technique: Attention

Explanation Technique: Attention

Attention is not Explanation

Sarthak Jain
Northeastern University
jain.sar@husky.neu.edu

Byron C. Wallace
Northeastern University
b.wallace@northeastern.edu

Explanation Technique: Attention

Attention is not Explanation

Sarthak Jain
Northeastern University
jain.sar@husky.neu.edu

Byron C. Wallace
Northeastern University
b.wallace@northeastern.edu

1. Attention is only mildly correlated with other importance score techniques

2. Counterfactual attention weights should yield different predictions, but they do not

Attention is not not Explanation

Sarah Wiegreffe*
School of Interactive Computing
Georgia Institute of Technology
saw@gatech.edu

Yuval Pinter*
School of Interactive Computing
Georgia Institute of Technology
uvp@gatech.edu

Attention is not not Explanation

Sarah Wiegreffe*
School of Interactive Computing
Georgia Institute of Technology
saw@gatech.edu

Yuval Pinter*
School of Interactive Computing
Georgia Institute of Technology
uvp@gatech.edu

"Attention *might* be an explanation."

Attention is not not Explanation

Sarah Wiegreffe*
School of Interactive Computing
Georgia Institute of Technology
saw@gatech.edu

Yuval Pinter*
School of Interactive Computing
Georgia Institute of Technology
uvp@gatech.edu

"Attention *might* be an explanation."

- Attention scores can provide a (plausible) explanation not **the explanation.**
- Attention is not explanation if you don't need it
- Agree that attention is indeed manipulable,

"this should provide pause to researchers who are looking to attention distributions for one true, faithful interpretation of the link their model has established between inputs and outputs."

Learning to Deceive with Attention-Based Explanations

Danish Pruthi[†], Mansi Gupta[‡], Bhuwan Dhingra[†], Graham Neubig[†], Zachary C. Lipton[†]
[†]Carnegie Mellon University, Pittsburgh, USA
[‡]Twitter, New York, USA
ddanish@cs.cmu.edu, mansig@twitter.com,
{bdhingra, gneubig, zlipton}@cs.cmu.edu

Attention	**Biography**	**Label**
Original	Ms. X practices medicine in Memphis, TN and is affiliated ... Ms. X speaks English and Spanish.	Physician
Ours	Ms. X practices medicine in Memphis , TN and is affiliated ... Ms. X speaks English and Spanish.	Physician

Learning to Deceive with Attention-Based Explanations

Danish Pruthi[†], Mansi Gupta[‡], Bhuwan Dhingra[†], Graham Neubig[†], Zachary C. Lipton[†]
[†]Carnegie Mellon University, Pittsburgh, USA
[‡]Twitter, New York, USA
ddanish@cs.cmu.edu, mansig@twitter.com,
{bdhingra, gneubig, zlipton}@cs.cmu.edu

Attention	Biography	Label
Original	Ms. X practices medicine in Memphis, TN and is affiliated ... Ms. X speaks English and Spanish.	Physician
Ours	Ms. X practices medicine in Memphis , TN and is affiliated ... Ms. X speaks English and Spanish.	Physician

- Manipulated models perform better than no-attention models
- Elucidate some workarounds (what happens behind the scenes)

Explanation Technique: Extractive Rationale Generation

Key idea: find minimal span(s) of text that can (by themselves) explain the prediction

- Generator (x) outputs a probability distribution of each word being the rational
- Encoder (x) predicts the output using the snippet of text x
- Regularization to support contiguous and minimal spans

Review

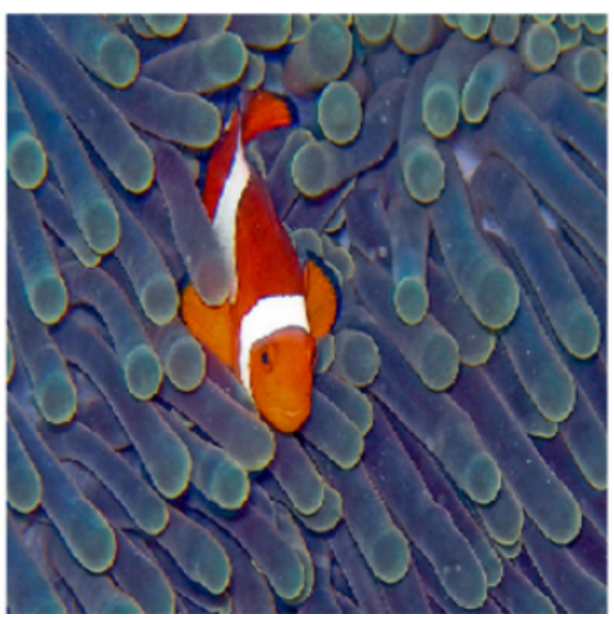
the beer was n't what i expected, and i'm not sure it's "true to style", but i thought it was delicious. **a very pleasant ruby red-amber color** with a relatively brilliant finish, but a limited amount of carbonation, from the look of it. aroma is what i think an amber ale should be - a nice blend of caramel and happiness bound together.

Ratings *Look:* 5 stars *Smell:* 4 stars

Figure 1: An example of a review with ranking in two categories. The rationale for Look prediction is shown in bold.

Explanation Technique: Influence Functions

- What would happen if a given training point didn't exist?
- Retraining the network is prohibitively slow, hence approximate the effect using influence functions.



Most influential train images

Koh & Liang, ICML 2017

How to evaluate?

Agreement among explanations

$\boldsymbol{e}_T^{(i)}(\boldsymbol{x}) \setminus \boldsymbol{e}_T^{(j)}(\boldsymbol{x})$	Random	Grad Norm	Grad $\times$ Input	LIME	Integrated Gradients	Attention
Random	1.00	0.10	0.10	0.10	0.10	0.10
Grad Norm	0.10	1.00	0.27	0.13	0.22	0.30
Grad $\times$ Input	0.10	0.27	1.00	0.11	0.16	0.17
LIME	0.10	0.13	0.11	1.00	0.16	0.15
Integrated Gradients	0.10	0.22	0.16	0.16	1.00	0.24
Attention	0.10	0.30	0.17	0.15	0.24	1.00

Overlap among the top-10% tokens selected by different explanation methods for sentiment analysis

How to evaluate?

How to evaluate?

how many townships have a population above 50 ? [prediction: NUMERIC]
what is the difference in population between fora and masilo [prediction: NUMERIC]
how many athletes are not ranked ? [prediction: NUMERIC]
what is the total number of points scored ? [prediction: NUMERIC]
which film was before the audacity of democracy ? [prediction: STRING]
which year did she work on the most films ? [prediction: DATETIME]
what year was the last school established ? [prediction: DATETIME]
when did ed sheeran get his first number one of the year ? [prediction: DATETIME]
did charles oakley play more minutes than robert parish ? [prediction: YESNO]

Figure 4. **Attributions from question classification model.** Term color indicates attribution strength—Red is positive, Blue is negative, and Gray is neutral (zero). The predicted class is specified in square brackets.

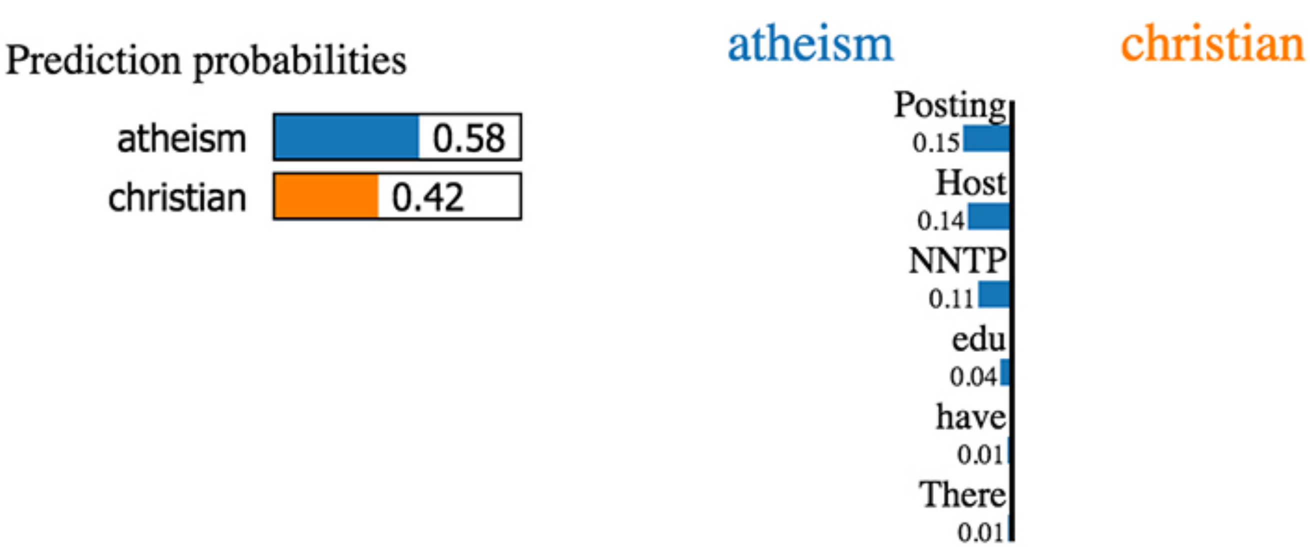
Integrated Gradients

How to evaluate?

how many townships have a population above 50 ? [prediction: NUMERIC]
what is the difference in population between fora and masilo [prediction: NUMERIC]
how many athletes are not ranked ? [prediction: NUMERIC]
what is the total number of points scored ? [prediction: NUMERIC]
which film was before the audacity of democracy ? [prediction: STRING]
which year did she work on the most films ? [prediction: DATETIME]
what year was the last school established ? [prediction: DATETIME]
when did ed sheeran get his first number one of the year ? [prediction: DATETIME]
did charles oakley play more minutes than robert parish ? [prediction: YESNO]

Figure 4. **Attributions from question classification model.** Term color indicates attribution strength—Red is positive, Blue is negative, and Gray is neutral (zero). The predicted class is specified in square brackets.

Integrated Gradients



Text with highlighted words
From: johnchad@triton.unm.edu (jchadwic)
Subject: Another request for Darwin Fish
Organization: University of New Mexico, Albuquerque
Lines: 11
NNTP-Posting-Host: triton.unm.edu

Hello Gang,

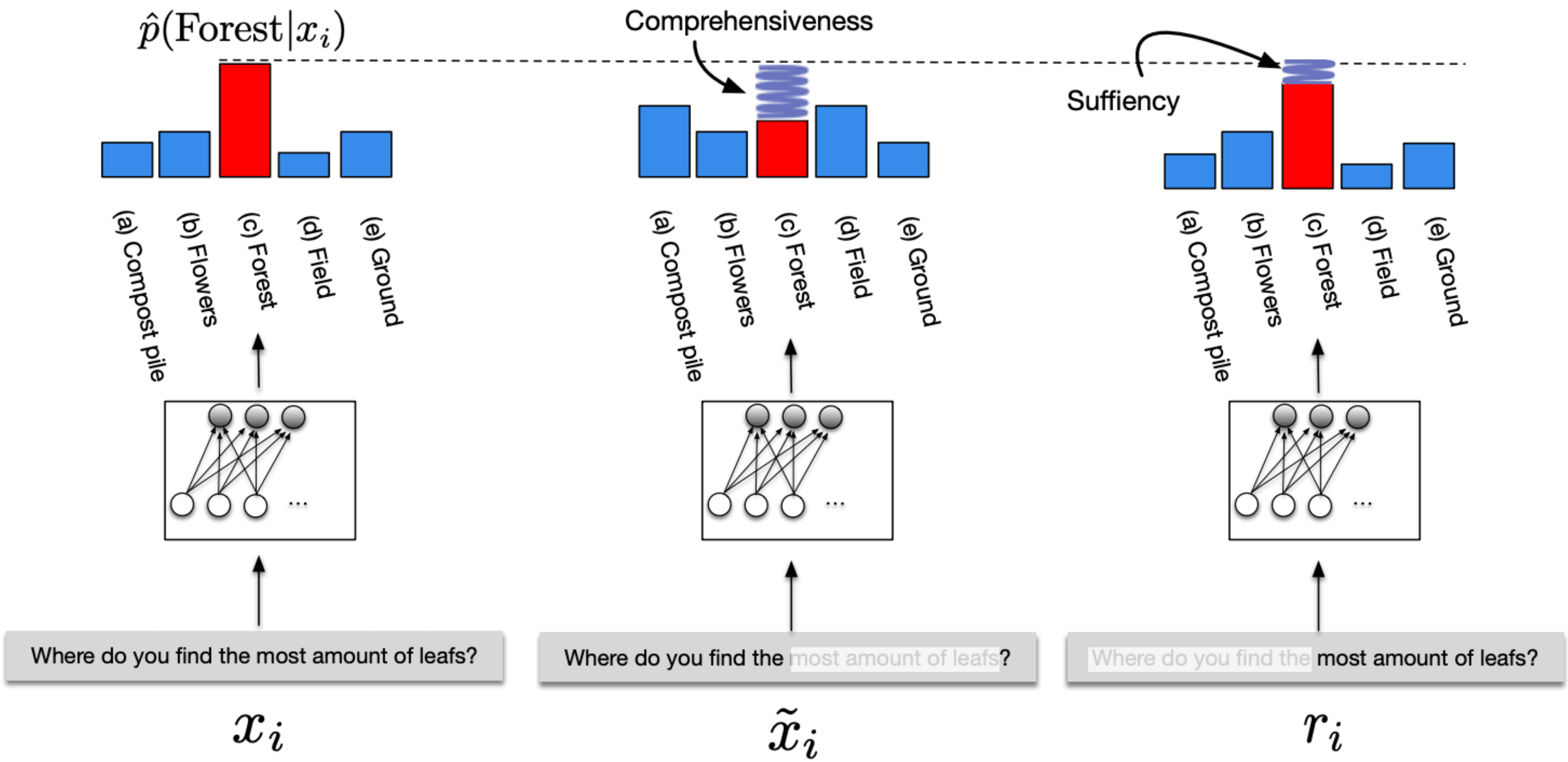
There have been some notes recently asking where to obtain the DARWIN fish.
This is the same question I have and I have not seen an answer on the
net. If anyone has a contact please post on the net or email me.

LIME

How to evaluate?

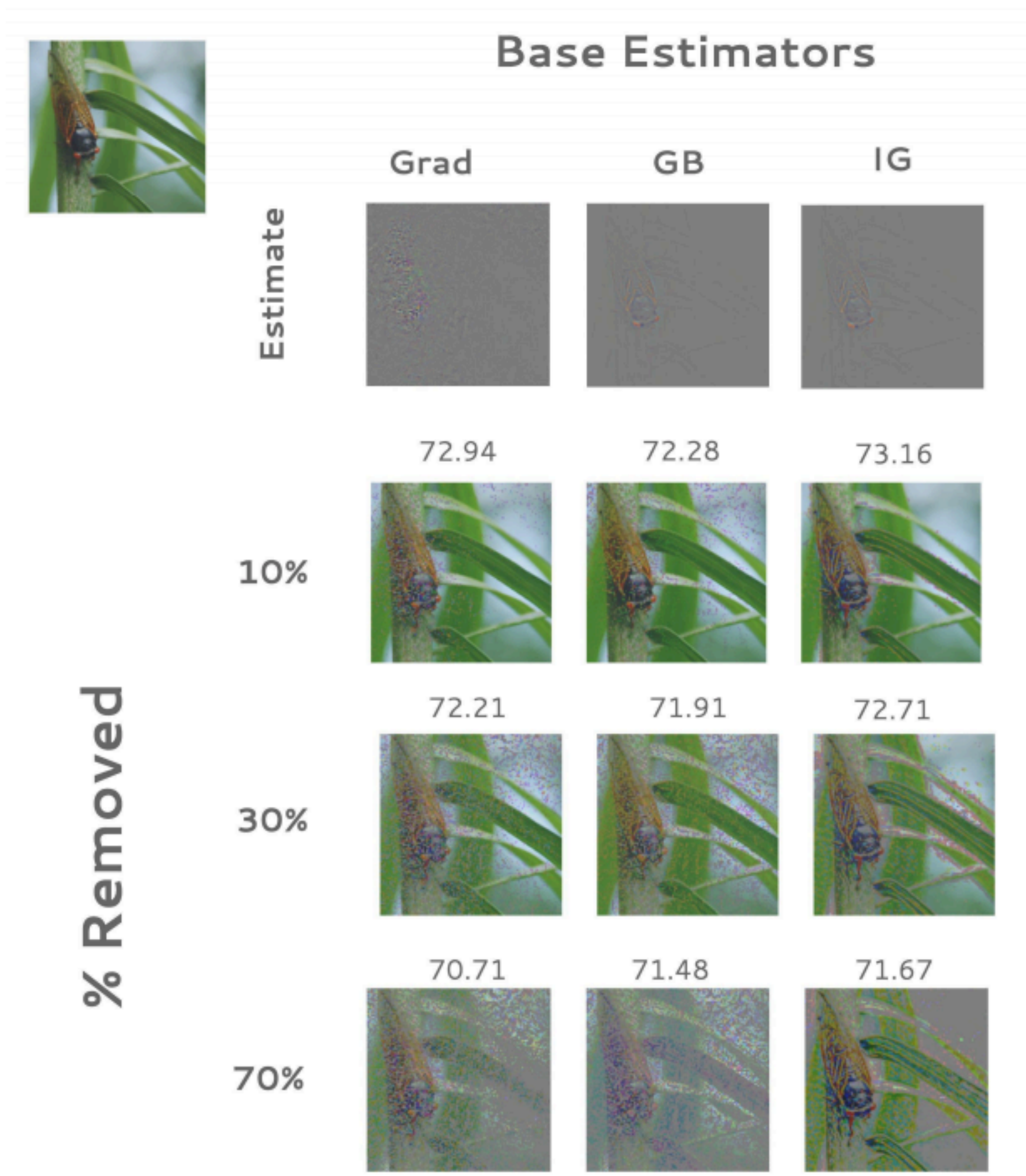
ERASER benchmark (DeYoung et al. 2020)

How to evaluate?



ERASER benchmark (DeYoung et al. 2020)

How to evaluate?



RemOve And Retrain (ROAR) benchmark
(Hooker et al. 2019)

How to evaluate?

Morphosyntactic Agreement

The link provided by the editor above **encourages ….**

Poerner et al, ACL 2018

How to evaluate?

Morphosyntactic Agreement

The link provided by the editor above **encourages ….**

Hybrid documents

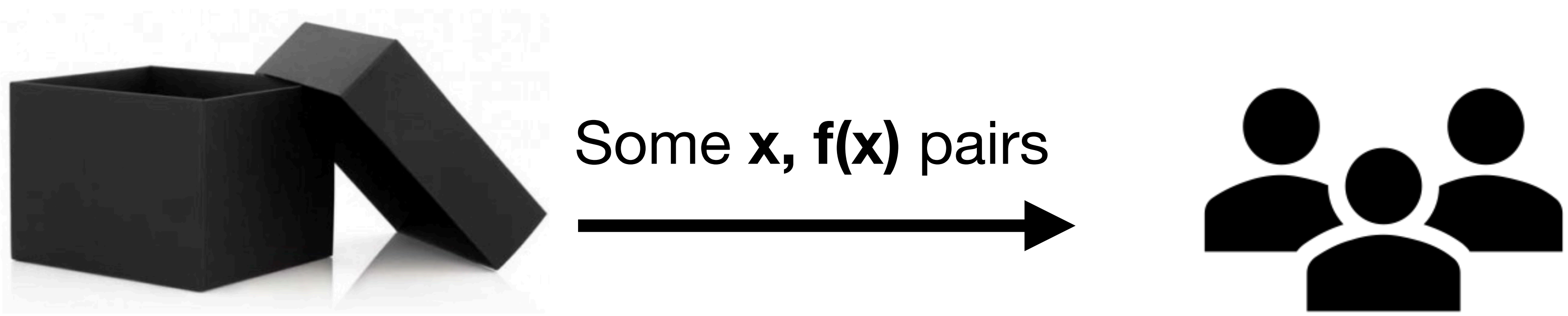
This is collected from Document 1. This text comes from Document 2. …. This text is taken from Document n.

Poerner et al, ACL 2018

Our proposal

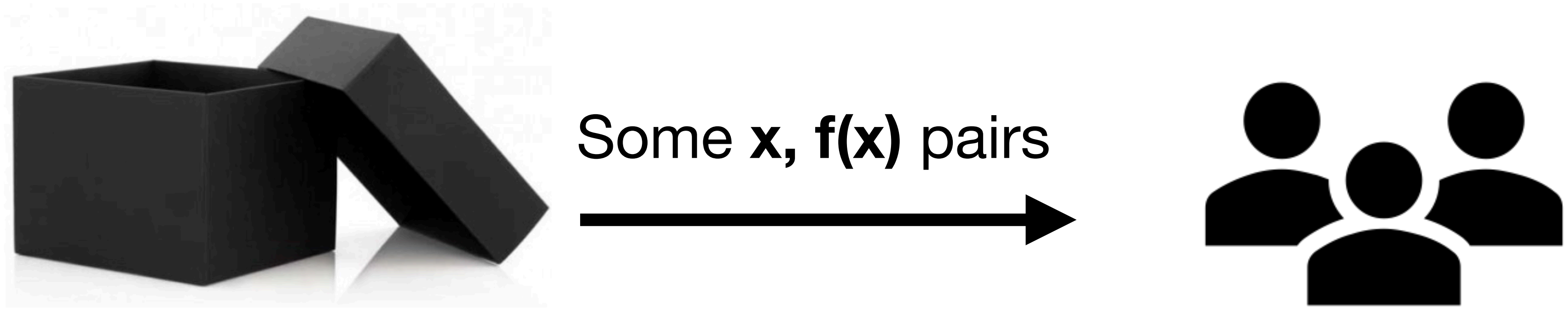
Pruthi et al. 2020: https://arxiv.org/pdf/2012.00893.pdf

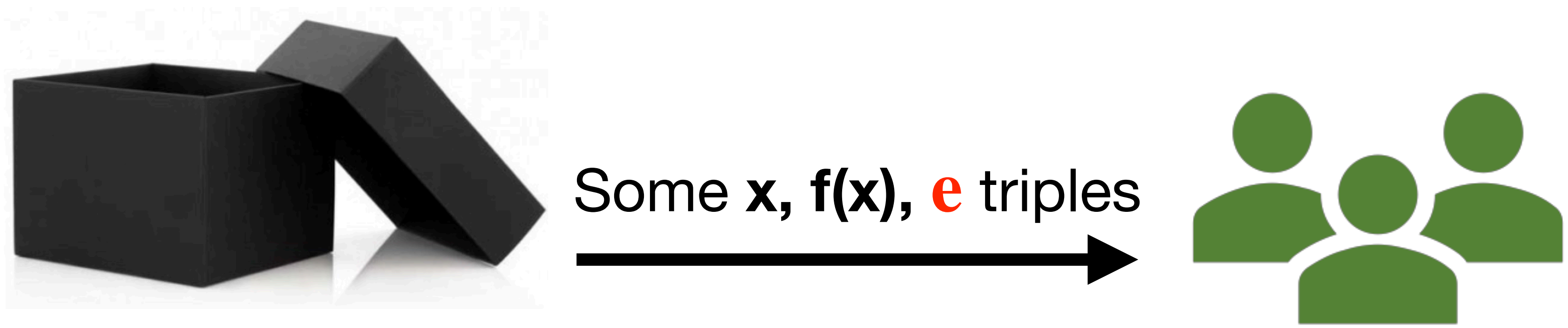
Our proposal



Pruthi et al. 2020: https://arxiv.org/pdf/2012.00893.pdf

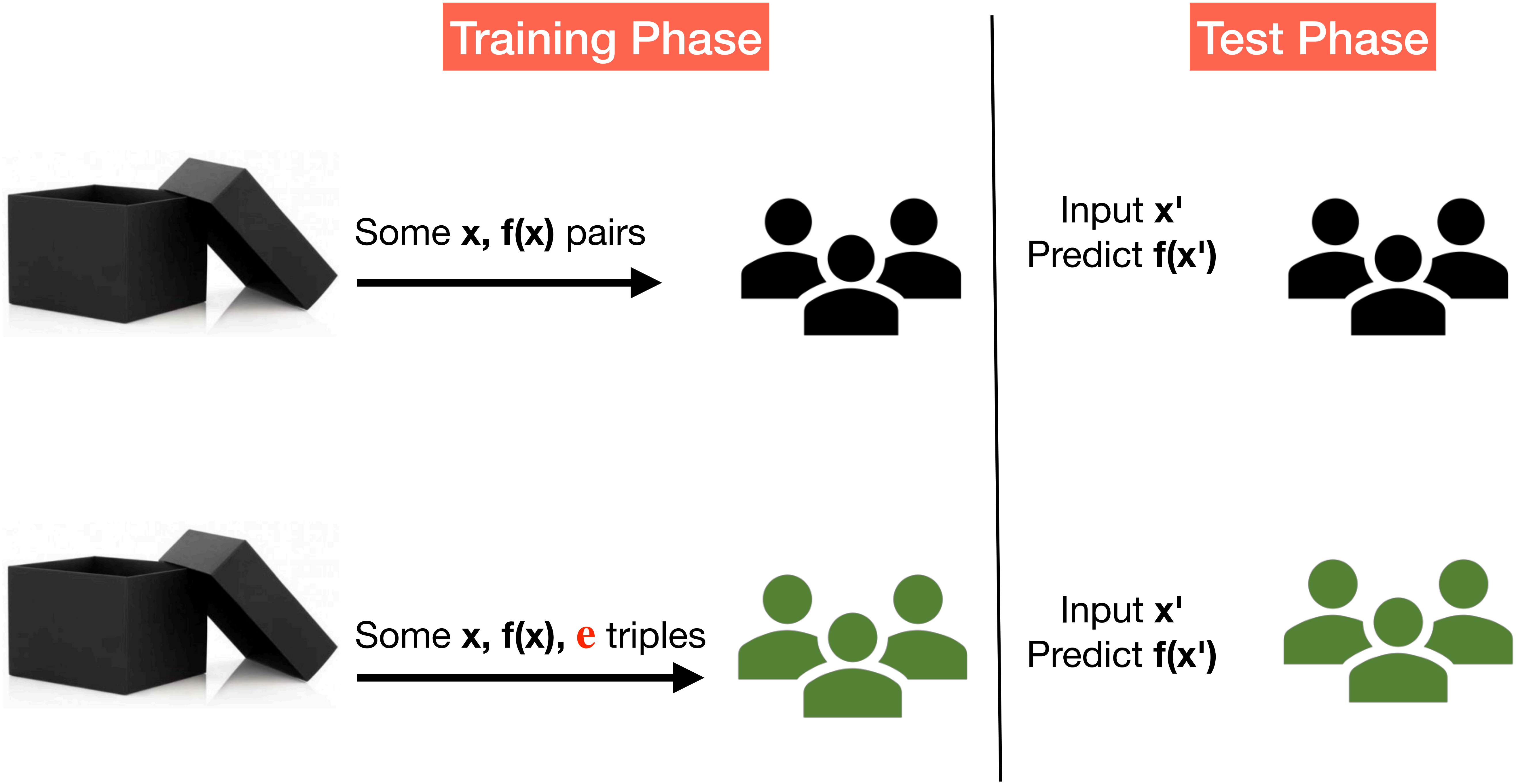
Our proposal





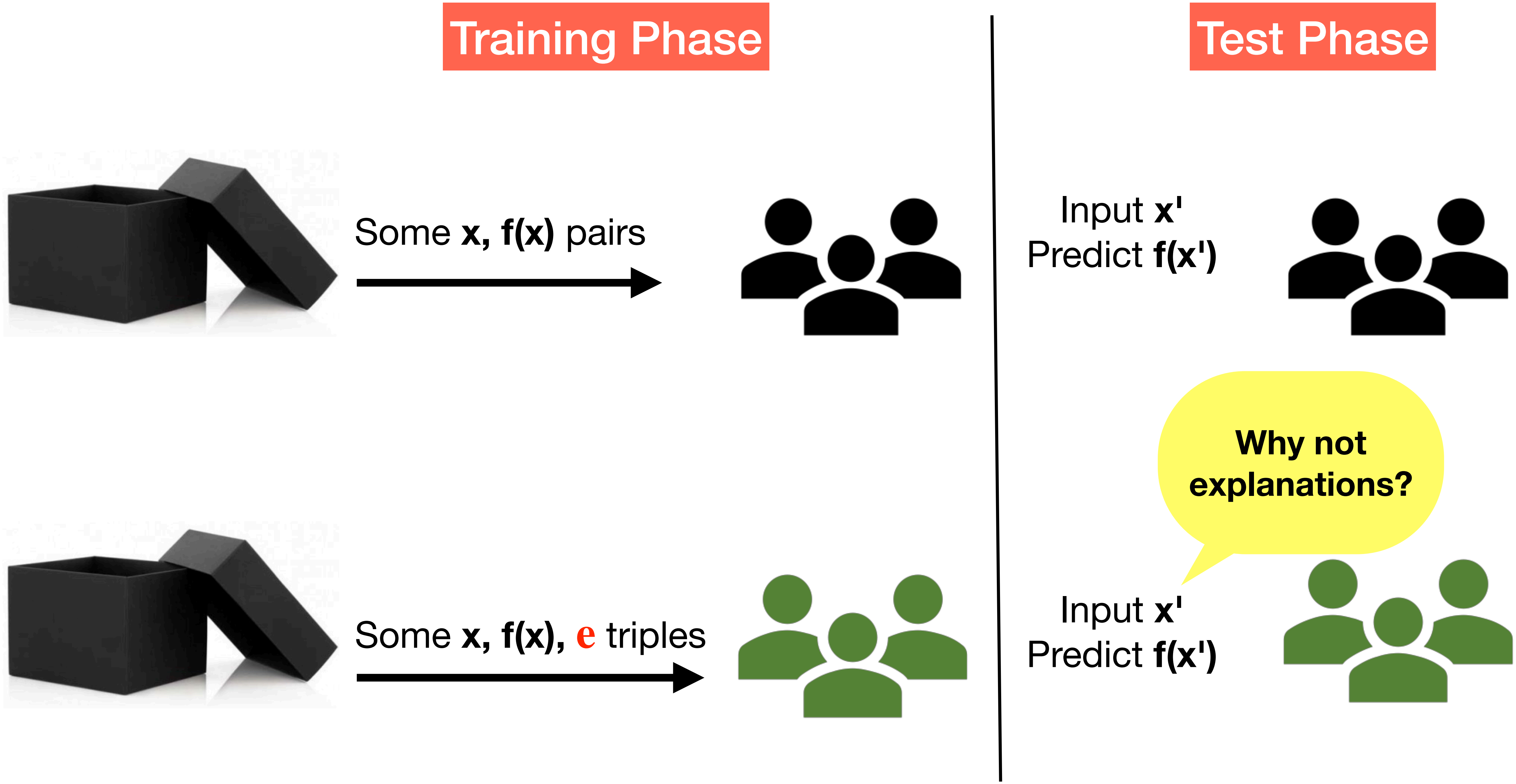
Pruthi et al. 2020: https://arxiv.org/pdf/2012.00893.pdf

Our proposal



Pruthi et al. 2020: https://arxiv.org/pdf/2012.00893.pdf

Our proposal



Pruthi et al. 2020: https://arxiv.org/pdf/2012.00893.pdf

Summarizing two directions

What is the model learning?

- Input: a model M, **a (linguistic) property P**
- Output: extent to which M captures P
- Techniques: classification, regression
- Evaluation: implicit

Explain the prediction

- Input: a model M, **a test example X**
- Output: an explanation E
- Techniques: varied …
- Evaluation: complicated

Discussion

What are you interpretability needs?

Thank You!