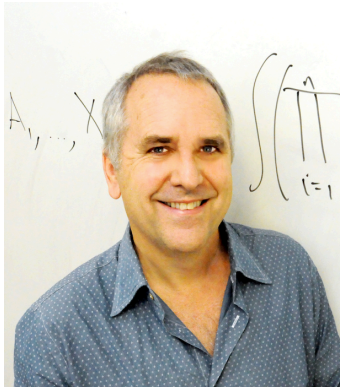
LS-Tree: Model Interpretation When the Data Are Linguistic

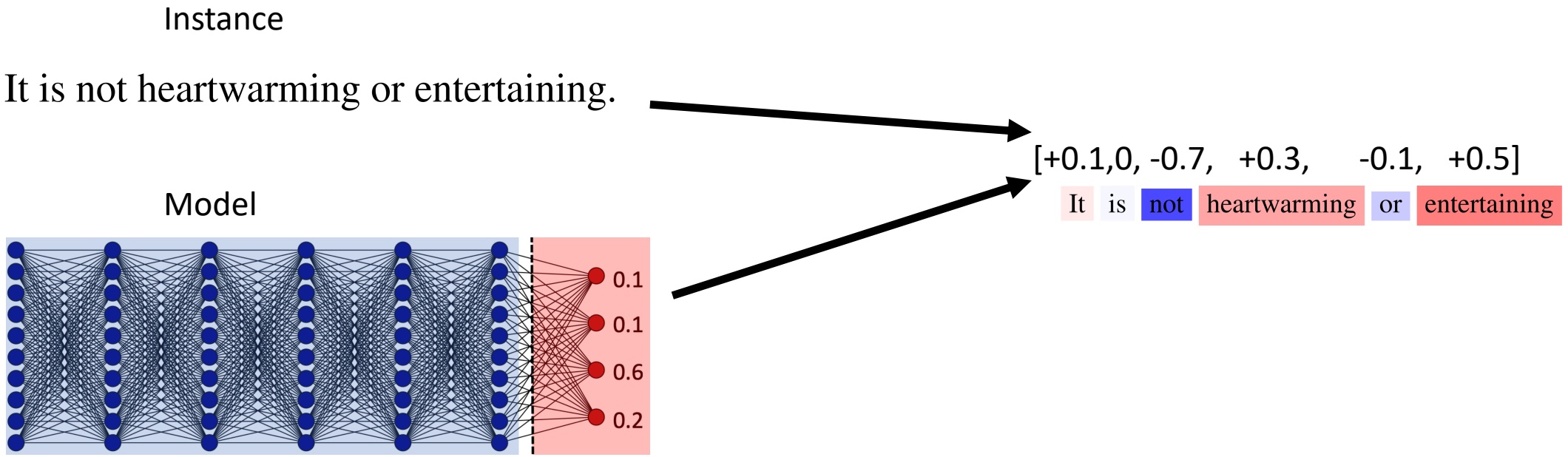
Jianbo Chen and Michael I. Jordan

University of California, Berkeley



Instancewise feature attribution

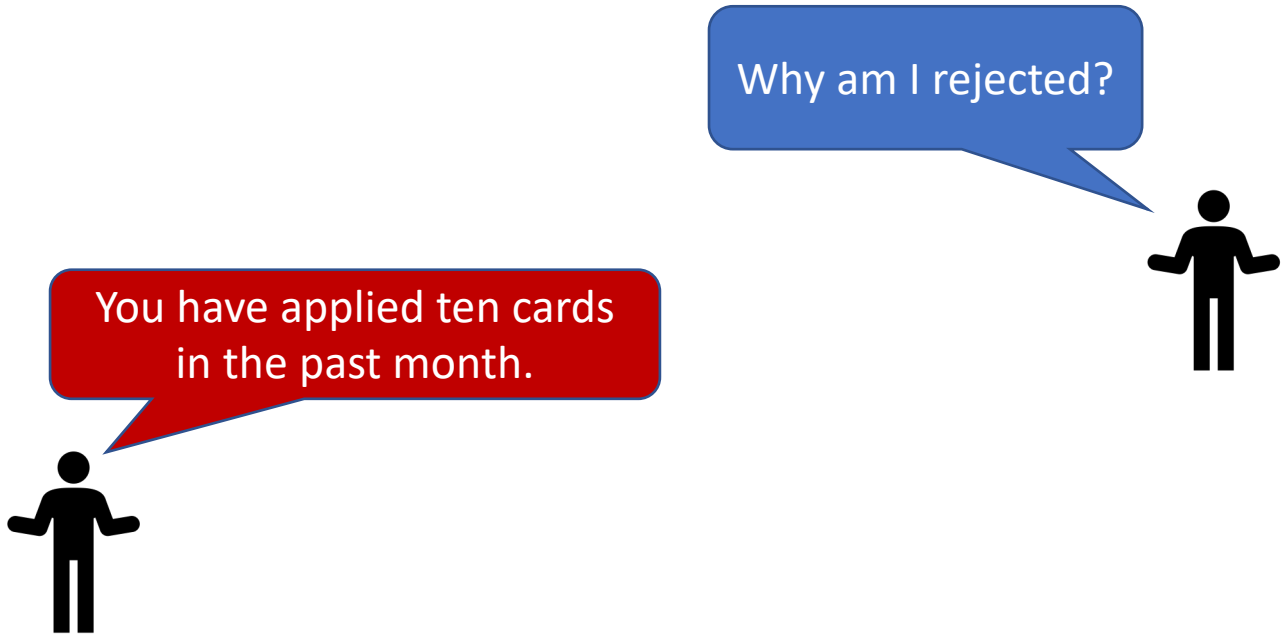
- For a given instance, assign a vector of importance scores for each feature.



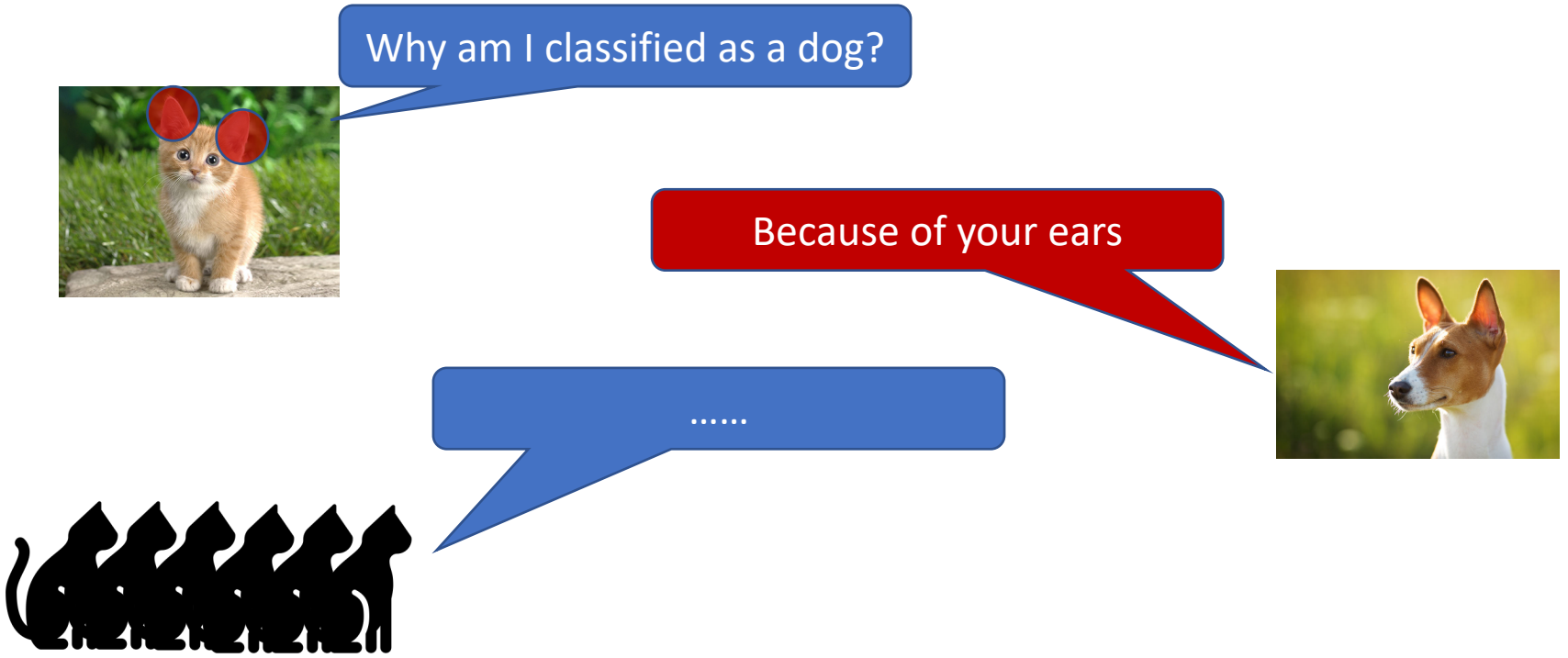

Why model interpretation

Transparency in critical decision making

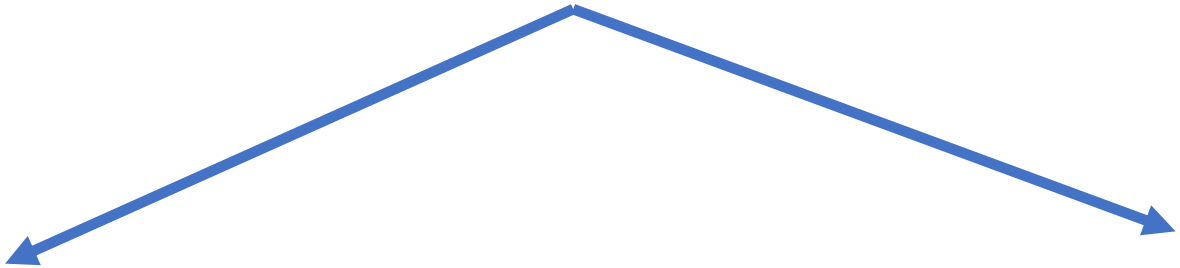
- credit card rejection, fraud detection …



Debugging tools



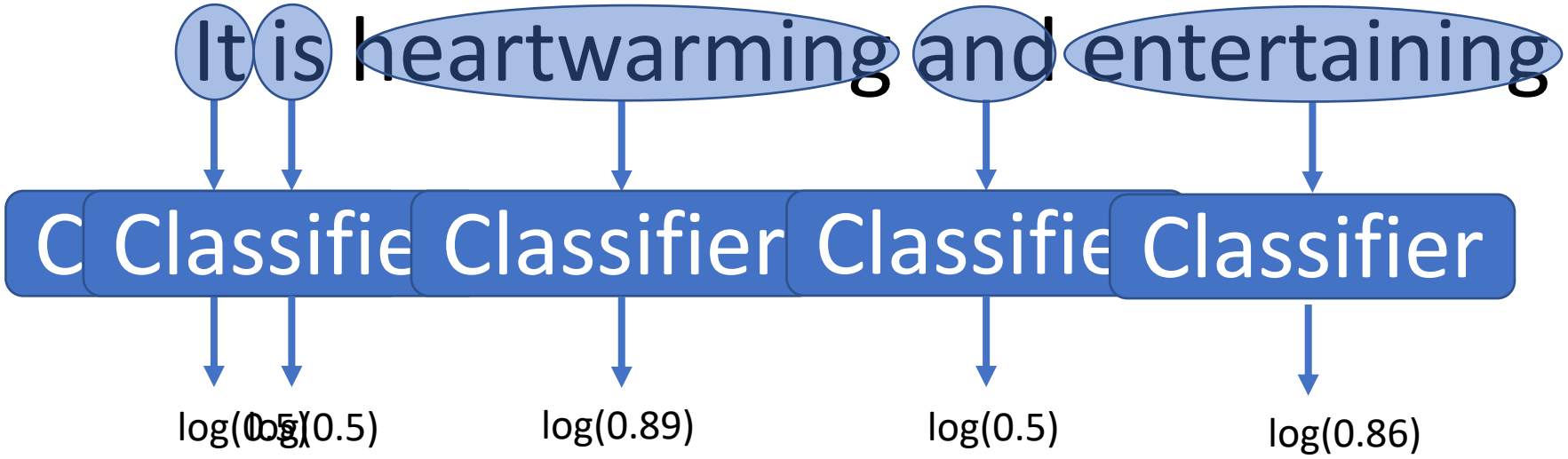
Black-box feature attribution



Model privacy

Convenience

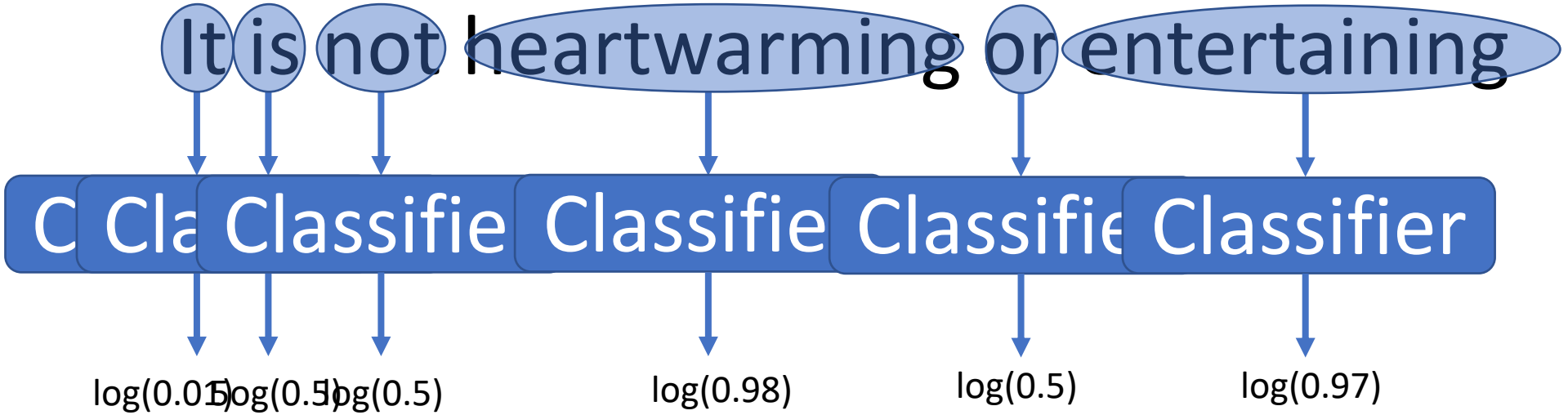
A simple method



Evaluating one feature at a time:

$$\phi_{f,x}(i) = f(\{i\})$$

When it sucks…




How to incorporate interaction?

Existing methods

- LIME (Ribeiro, Singh, and Guestrin 2016)
- Representation Erasure (Li, Monroe, and Jurafsky 2016)
- Quantitative Input Influence (QII) (Datta, Sen, and Zick 2016)
- SHapley Additive exPlanations (SHAP) (Lundberg and Lee 2017)
- L-Shapley and C-Shapley (Chen, Song, Wainwright, and Jordan 2018)
- ……

Procedures of existing methods

- Step 1: Sample word subsets with a certain scheme
- Step 2: Evaluate target model f on each sampled word subset
- Step 3: Combine model evaluations into attribution scores

An illustration – the Shapley value (Shapley 1953)

Shapley value – Step 1 and Step 2

It is not heartwarming or entertaining $f(\text{“not heartwarming”}) - f(\text{“heartwarming”})$

It is not heartwarming or entertaining $f(\text{“It is not”}) - f(\text{“It is”})$

It is not heartwarming or entertaining $f(\text{“It ... not”}) - f(\text{“It”})$

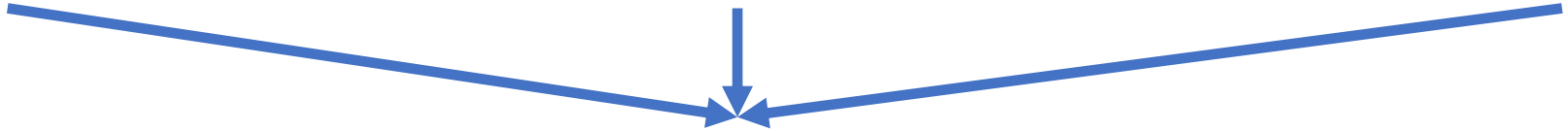
……

Marginal contribution of i to S: $f(S \cup \{i\}) - f(S)$

where $f(S) := f(x_S)$

Shapley Value – Step 3

[Additivity] [Equal Contributions] [Monotonicity]



$$\phi_{f,x}(i) = \frac{1}{d} \sum_{S \subset [d]} \frac{1}{\binom{d-1}{|S|-1}} (f(S \cup \{i\}) - f(S))$$

Example: It is not heartwarming or entertaining

- Quantitative Input Influence (QII) (Datta, Sen, and Zick 2016)
- SHapley Additive exPlanations (SHAP) (Lundberg and Lee 2017)
- L-Shapley and C-Shapley (Chen, Song, Wainwright, and Jordan, 2018)

Limitations of existing methods

Step 1: Sample word subsets with a certain scheme
Step 2: Evaluate target model f on each sampled word subset

It is not heartwarming or entertaining $f(\text{“It ... not”}) - f(\text{“It”})$

‘It … not’ is not natural language.

Human interpretable word combinations

Limitations of existing methods

Step 3: Combine model evaluations into attribution scores for each word.

? Is ‘not’ important as a single word, or because of its interaction with ‘heartwarming’

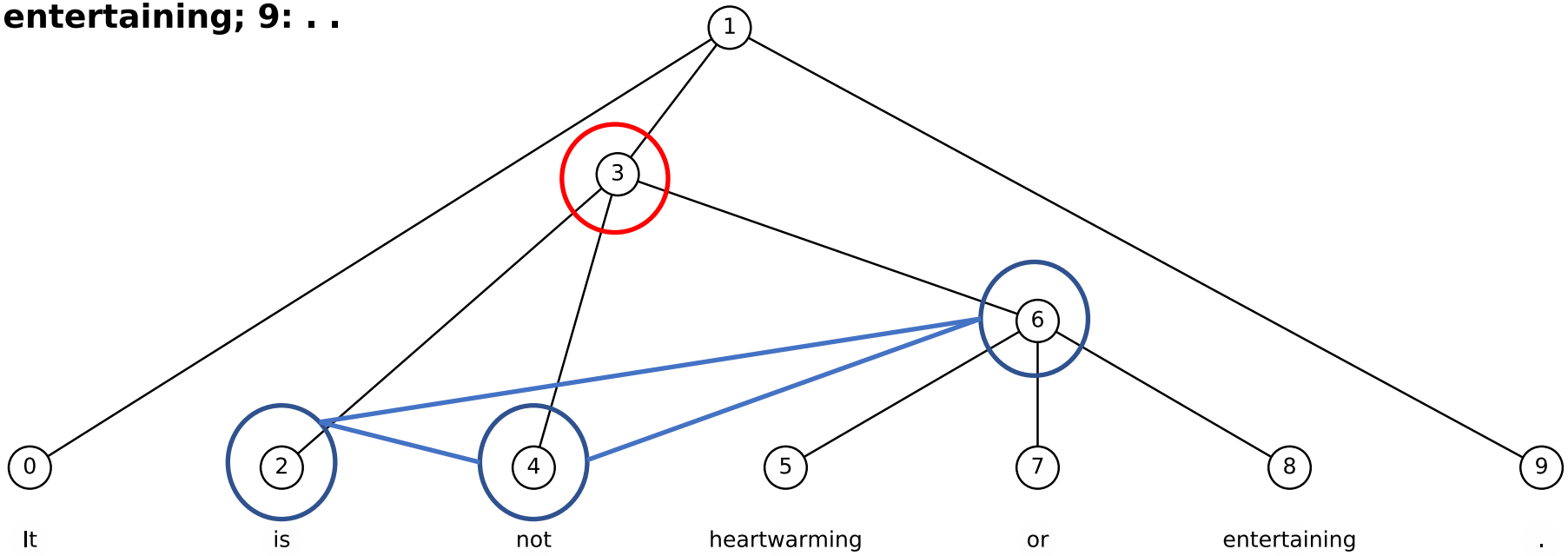
~~It is not heartwarming or entertaining~~

? What expressions are valid to human?

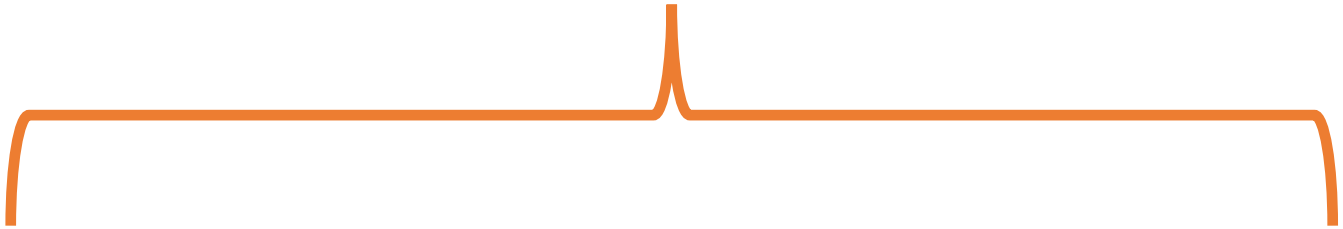
What interactions are we interested in?

Constituency parsing for linguistic data

0: it; 2: is; 4: not; 5: heartwarming;
1: it is not heartwarming or entertaining. ;
3: is not; 6: heartwarming or entertaining;
7: or; 8: entertaining; 9: . .



Our approach: LS-Tree



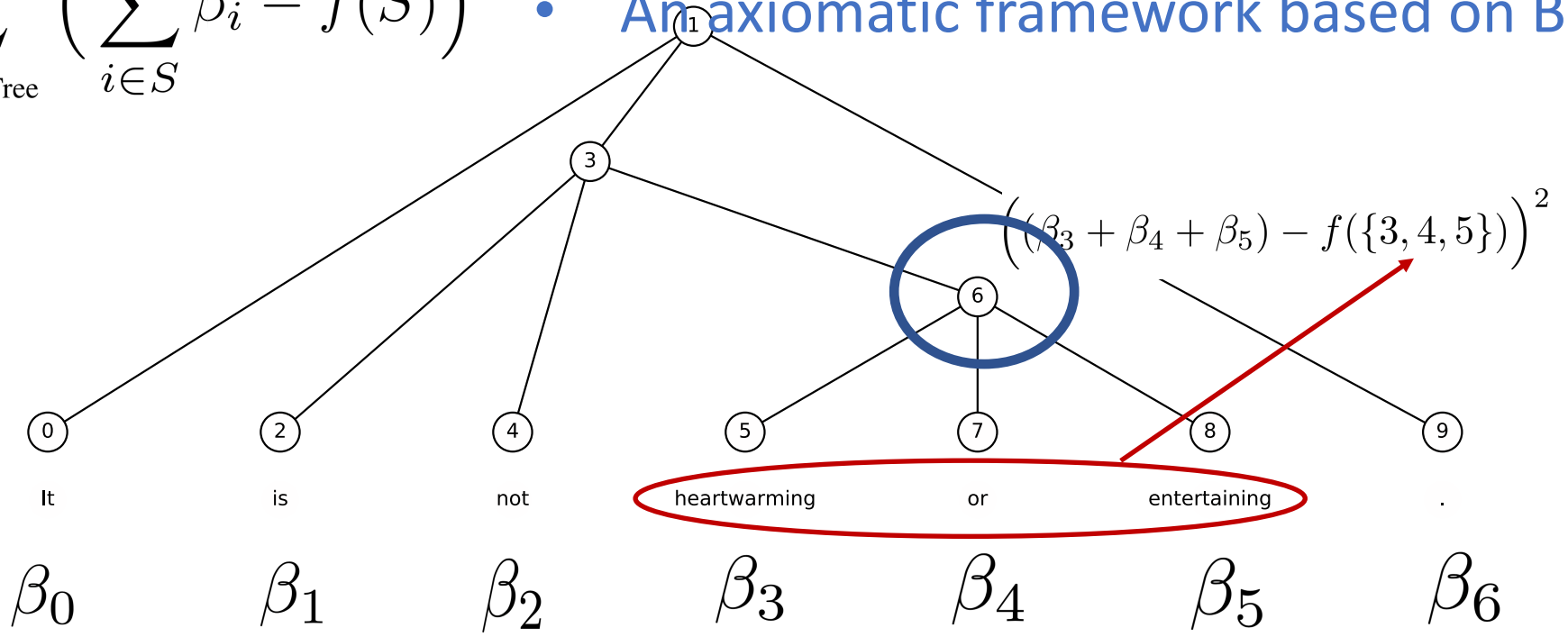
Least squares

Cook’s interaction score

Step 1: Least-squares

$$\min_{\beta} \sum_{S \in V_{\text{Tree}}} \left(\sum_{i \in S} \beta_i - f(S) \right)^2$$

- A linear approximation at nodes of the tree.
- An axiomatic framework based on Banzhaf value.

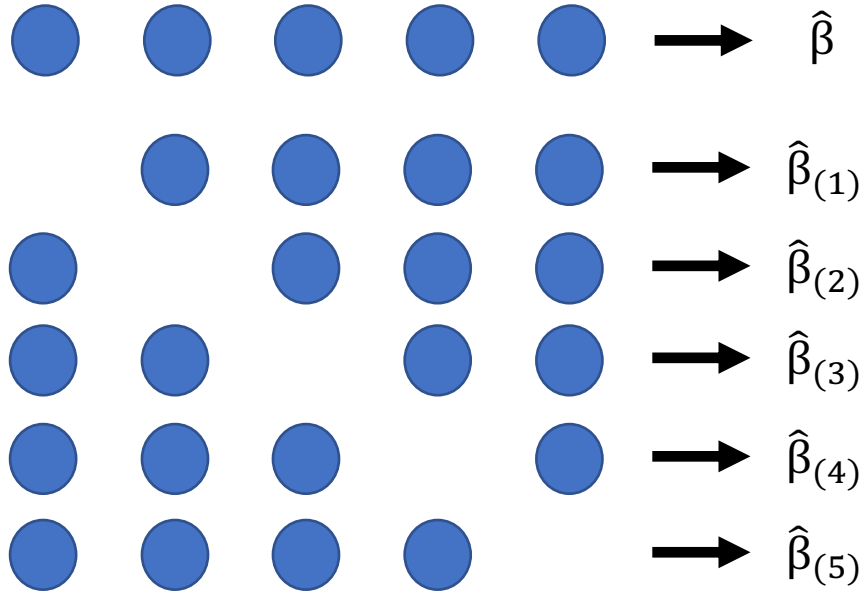


Cook's distance (Cook 1977)

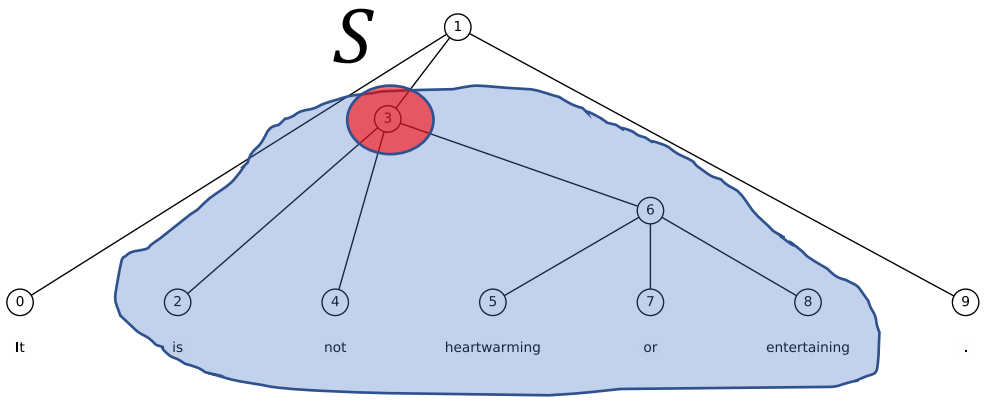
Capture the influence of instance i:

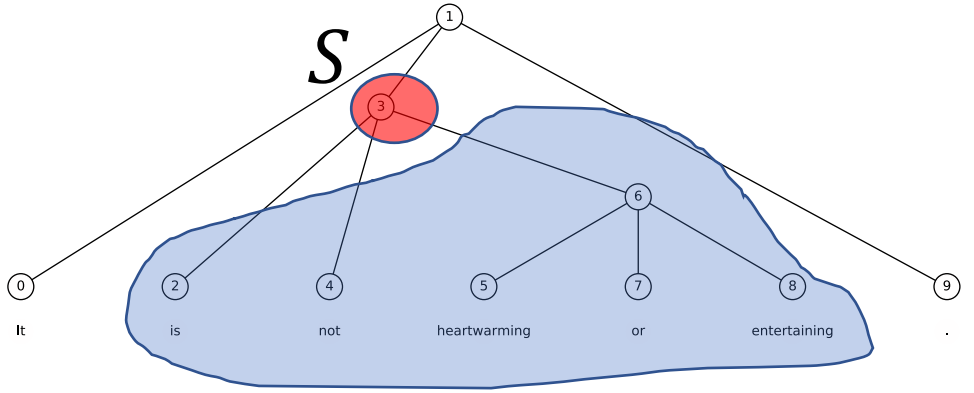
$$D_i = \text{Const.} \cdot (\hat{\beta}_{(i)} - \hat{\beta})^T X^T X (\hat{\beta}_{(i)} - \hat{\beta})$$

$\hat{\beta}_{(i)}$: Fit a linear model without data point i.



Step 2: Influence of the interaction at node S





$\beta_{\geq S}$ ← Distance → $\beta_{>S}$

All nodes: An $\Theta(d^3)$ algorithm using the Sherman-Morrison formula.

Properties of LS-Tree

- Constituency parsing: Incorporate prior knowledge
- Cook's distance: Attribute interactions
- Complexity
  - Linear query complexity
  - Sherman-Morrison: Cubic computational complexity

Adversative relations

- not, but, yet, though, although, even though, whereas, except, despite, in spite of

Size of data set

10K

100K

600K

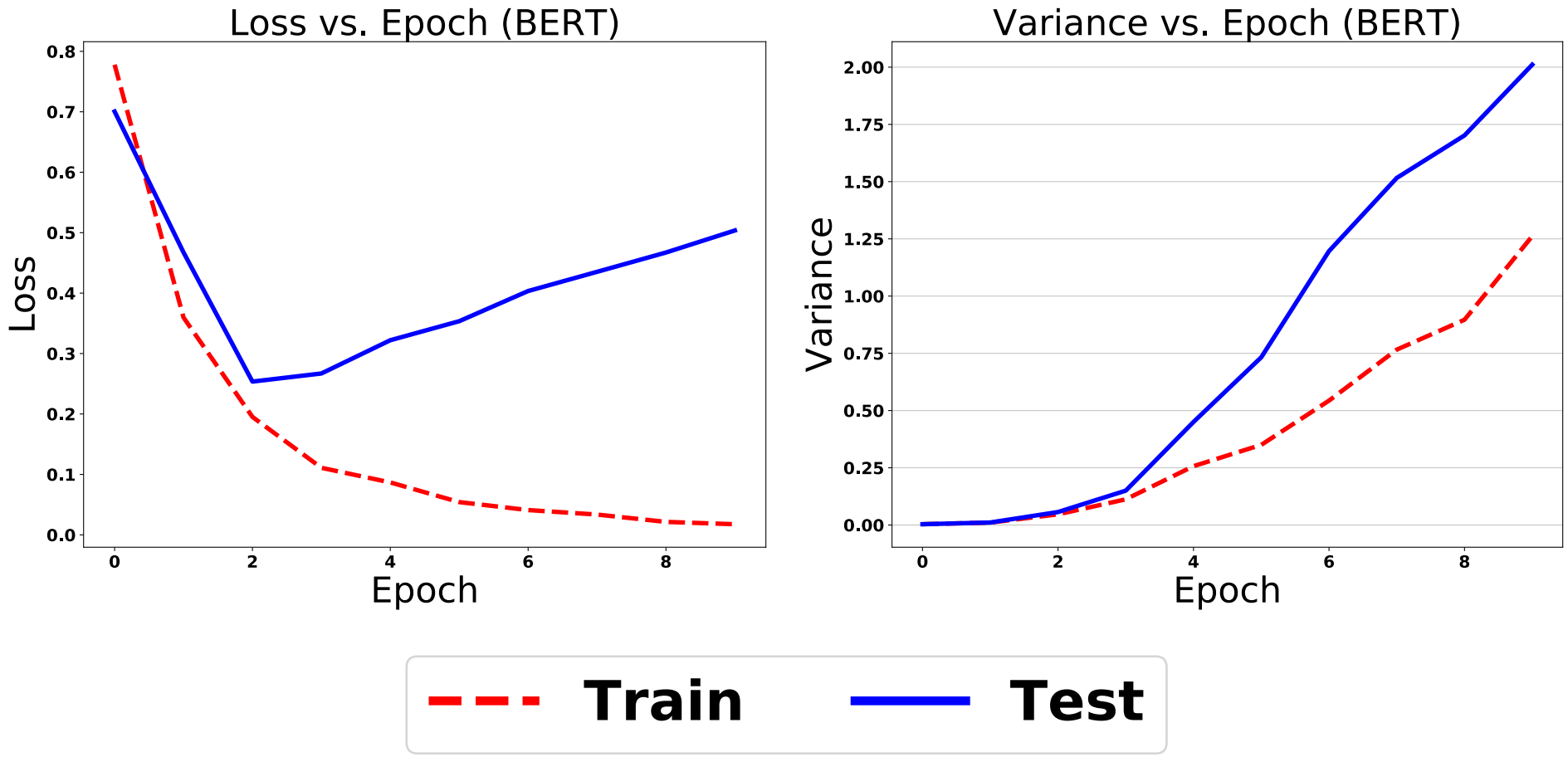
Dataset	Model	Avg. Score	not	but	yet	though	although	even though	whereas	except	despite	in spite of
SST	BoW	0.153	0.000(6.318)	0.000(0.079)	0.000(2.005)	0.000(0.865)	0.000(2.222)	0.000(0.000)	-(-)	0.000(4.280)	0.000(3.519)	0.000(0.000)
	CNN	0.634	1.673(4.592)	1.694(1.444)	0.568(0.959)	0.213(0.735)	0.915(0.462)	0.626(0.407)	-(-)	0.948(1.175)	**1.452**(4.270)	**2.119**(1.943)
	LSTM	0.79	**1.746**(2.580)	1.502(0.453)	1.449(2.368)	1.153(1.094)	0.338(0.197)	1.794(0.998)	-(-)	**2.353**(3.835)	1.256(1.818)	0.590(0.624)
	BERT	1.238	1.714(4.383)	**2.148**(1.760)	**1.669**(3.120)	**1.525**(3.268)	**1.741**(3.256)	**1.885**(2.092)	-(-)	1.156(3.331)	1.160(2.998)	0.864(2.352)
IMDB	BoW	0.038	0.000(2.683)	0.000(0.263)	0.000(2.210)	0.000(1.473)	0.000(1.710)	0.000(0.000)	0.000(3.604)	0.000(1.342)	0.000(0.132)	-(-)
	CNN	0.424	1.050(0.819)	**3.442**(0.021)	**1.689**(0.295)	0.922(0.085)	1.036(0.071)	1.175(0.467)	0.469(1.064)	**1.590**(4.067)	0.363(0.434)	-(-)
	LSTM	0.126	0.960(3.087)	2.222(0.524)	1.500(0.238)	0.611(0.087)	0.492(1.270)	0.944(0.683)	**1.222**(3.865)	1.294(4.008)	0.286(0.508)	-(-)
	BERT	1.159	**1.616**(2.057)	3.390(1.800)	1.644(1.152)	**1.371**(2.061)	**1.735**(2.123)	**1.457**(1.557)	0.285(0.430)	1.421(2.060)	**1.518**(2.241)	-(-)
Yelp	BoW	0.035	0.000(8.488)	0.000(1.015)	0.000(3.553)	0.000(1.664)	0.000(1.128)	0.000(0.000)	0.000(0.536)	0.000(0.367)	0.000(1.213)	-(-)
	CNN	0.161	**2.287**(3.467)	**2.454**(0.932)	0.516(0.043)	0.988(0.435)	**0.889**(0.075)	0.789(0.621)	0.286(0.671)	**0.522**(2.529)	0.423(0.889)	-(-)
	LSTM	0.224	2.173(5.950)	1.712(1.676)	**0.988**(2.065)	0.984(1.310)	0.706(1.194)	0.559(0.483)	**1.395**(1.793)	0.344(1.408)	0.514(1.153)	-(-)
	BERT	0.746	1.384(2.106)	2.448(0.658)	0.781(0.184)	**1.336**(0.953)	0.596(0.615)	**1.019**(0.880)	0.095(0.162)	0.331(0.074)	**1.041**(0.414)	-(-)

Is “while” indicating a contrast?

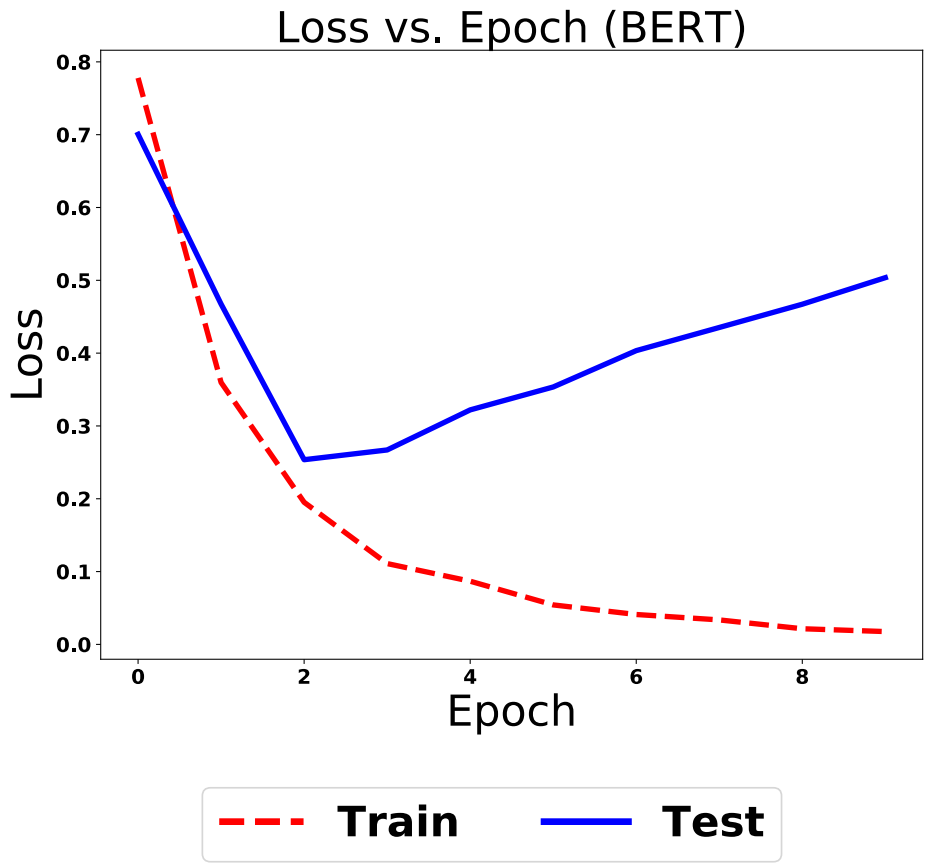
Interaction scores of the parent node of “while”.

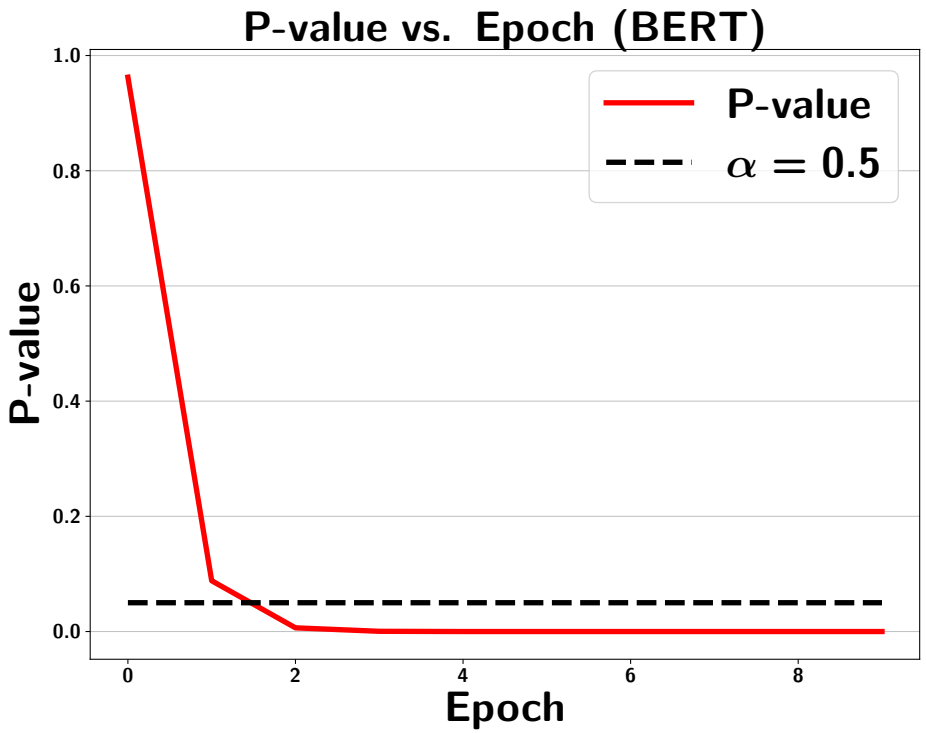
Sentence	Meaning	BoW	CNN	LSTM	BERT
... He said he couldn’t help. We had to walk **while** the snow blew in our faces. When we were almost there, we saw the shuttle pull out with the smoking shuttle driver in it, driving in the opposite direction, away from us. I can not believe how rude they were.	during the time that	0.000(0.338) ✓	0.781(0.300) ✗	1.761(0.839) ✗	0.062(0.092) ✓
... I ordered a cappuccino. It tasted like milk and no coffee. I was exceptionally disappointed. So **while** the place has a great reputation, even they can screw it up if they don t pay attention to detail, and at this level they should never screw it up. I had a better cup at Martys Market for crying out loud!	whereas (indicating a contrast)	0.000(0.338) ✗	1.142(0.300) ✓	2.155(0.839) ✓	2.167(0.092) ✓
Usually asking the server what is her favorite dish gets you a pretty good recommendation, but in this case, it was crap! The smoked brisket had that discoloration that happens to meat when it’s been sitting out for a **while**. And it wasn’t even tender!! Am I asking for too much?	a period of time	0.000(0.338) ✓	0.206(0.300) ✓	0.465(0.839) ✓	0.082 (0.092) ✓

Overfitting



Overfitting – A Permutation Test





Questions

Or email to jianbochen@berkeley.edu

Title: LS-Tree: Model Interpretation When the Data Are Linguistic

Code: To appear at https://github.com/Jianbo-Lab/LS-Tree