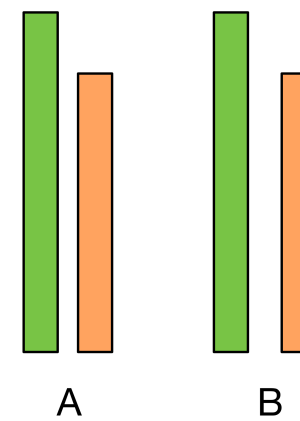
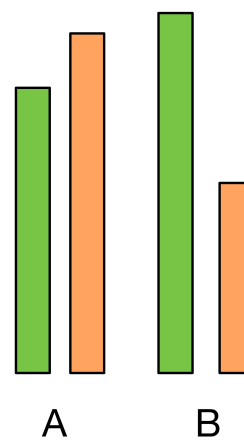
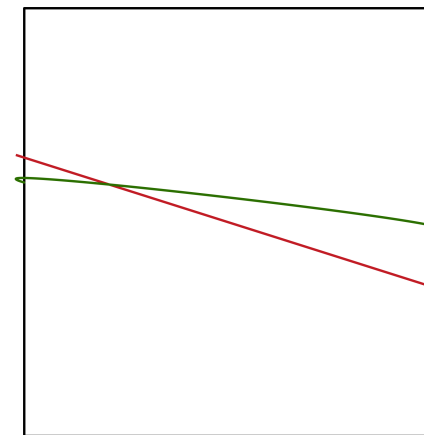
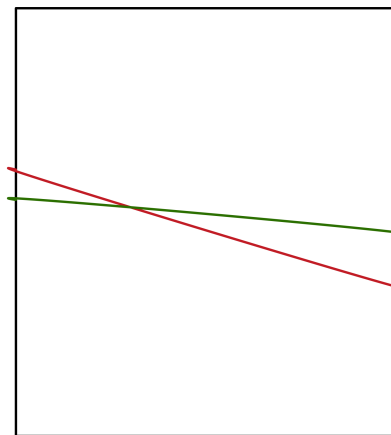
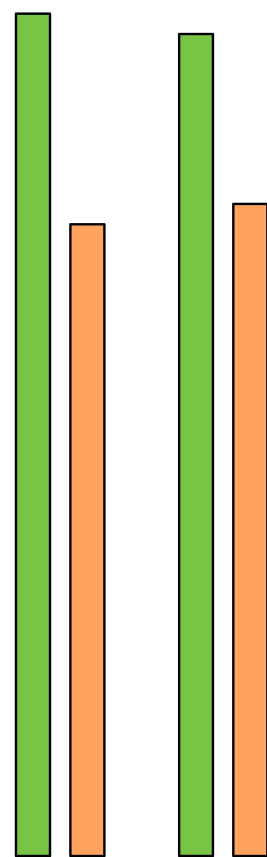


# Analyzing NN & Discovering Bias

Antske Fokkens & Pia Sommerauer  
2 July 2019

# Bias (sentiment mining)



# Bias in Embeddings

Analogy	Reported	Index	1st answer	2nd answer
<a href="#">Mikolov et al. (2013a)</a>				
man king woman	queen	2	king	queen
Paris France Tokyo	Japan	1	Japan	Tokyo
brother sister grandson	granddaughter	1	granddaughter	niece
big bigger cold	colder	2	cold	colder
Einstein scientist Picasso	painter	1	painter	scientist
<a href="#">Bolukbasi et al. (2016)</a>				
man computer_programmer woman	homemaker	2	computer_programmer	homemaker
he doctor she	nurse	2	doctor	nurse
she interior_designer he	architect	2	interior_designer	architect
she feminism he	conservatism	4	feminism	liberalism
she lovely he	brilliant	10	lovely	magnificent
she sewing he	carpentry	4	sewing	woodworking
<a href="#">Manzini et al. (2019b)</a>				
black criminal caucasian	lawful	13	legal	statutory
caucasian lawful black	criminal	2	lawful	criminal
caucasian hillbilly asian	yuppie	3	hillbilly	hippy
asian yuppie caucasian	hillbilly	2	yuppie	hillbilly
asian engineer black	killer	39	operator	jockey
black killer asian	engineer	7	killer	impostor
christian conservative jew	liberal	4	centrist	democrat
jew liberal christian	conservative	2	liberal	conservative
muslim terrorist jew	journalist	4	hacker	protestor
jew journalist muslim	terrorist	2	purportedly	terrorist
christian conservative muslim	regressive	53	moderate	conservative
muslim regressive christian	conservative	13	regressive	progressive

# First Answer

- Rich and varied enough evaluation data
- Error Analysis
- Coming soon:

Tongshuang Wu, Marco Tulio Ribeiro, Jeffrey Heer and Daniel Weld (2019) Errudite: Scalable, Reproducible, and Testable Error Analysis. *Proceedings of ACL*

# Analyzing your model

- Typical problems:
  - preprocessing problems
  - errors in gold
  - topic bias

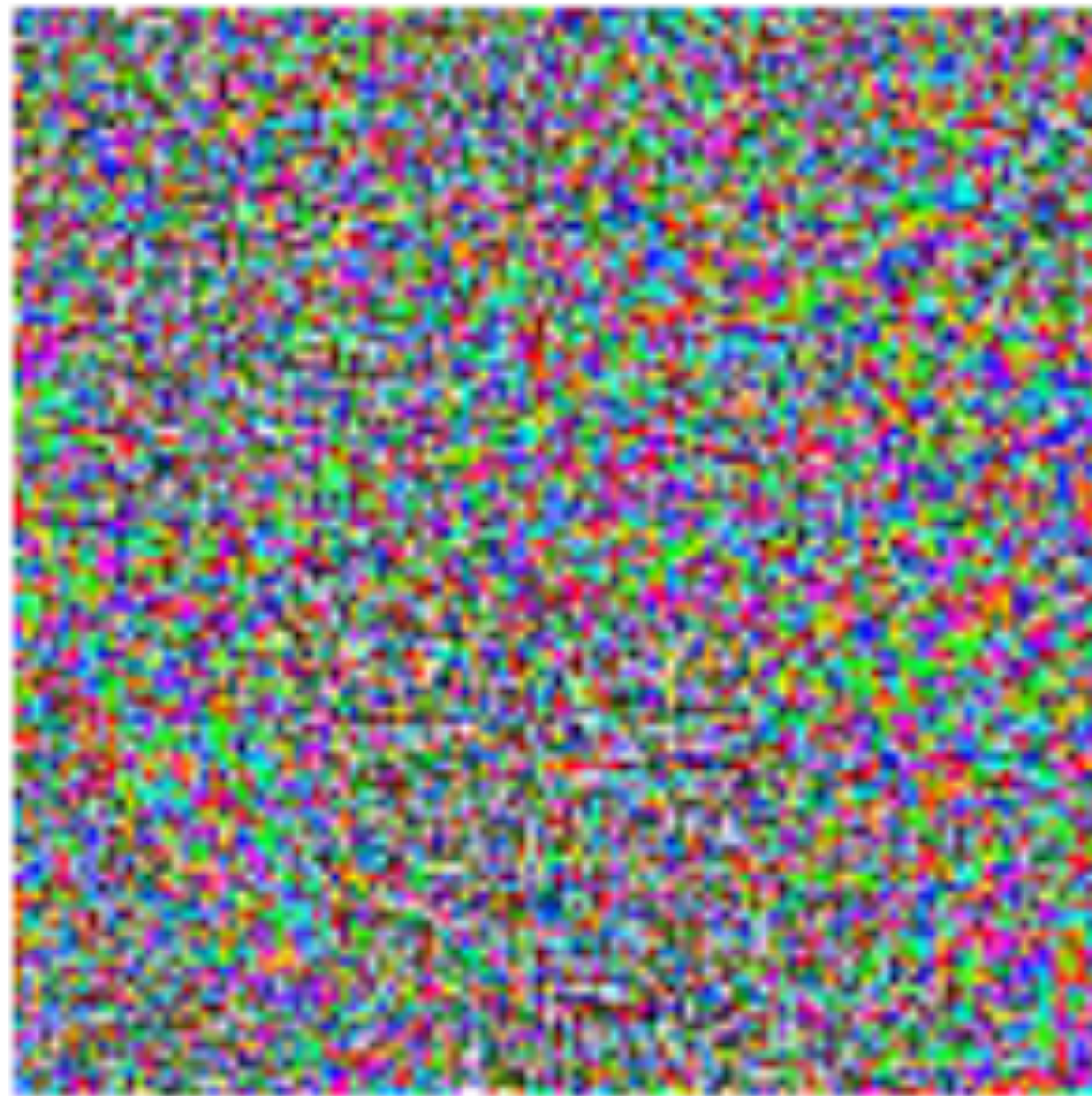
# Good old days...

- Naïve Bayes/Maximum Entropy Classifiers:
  - directly study correlating feature values & outcome
- Logistic Regression:
  - check the weights of specific features & extract most influential ones
- Often: possibilities of creating confidence scores

# Nowadays...

- Neural networks are proverbial blackbox structures:
  - most input representations are opaque high-density embeddings
  - so are all internal representations
  - they are highly connected: which components provide actual input?

# Neural Networks



**don't know when they don't know...**



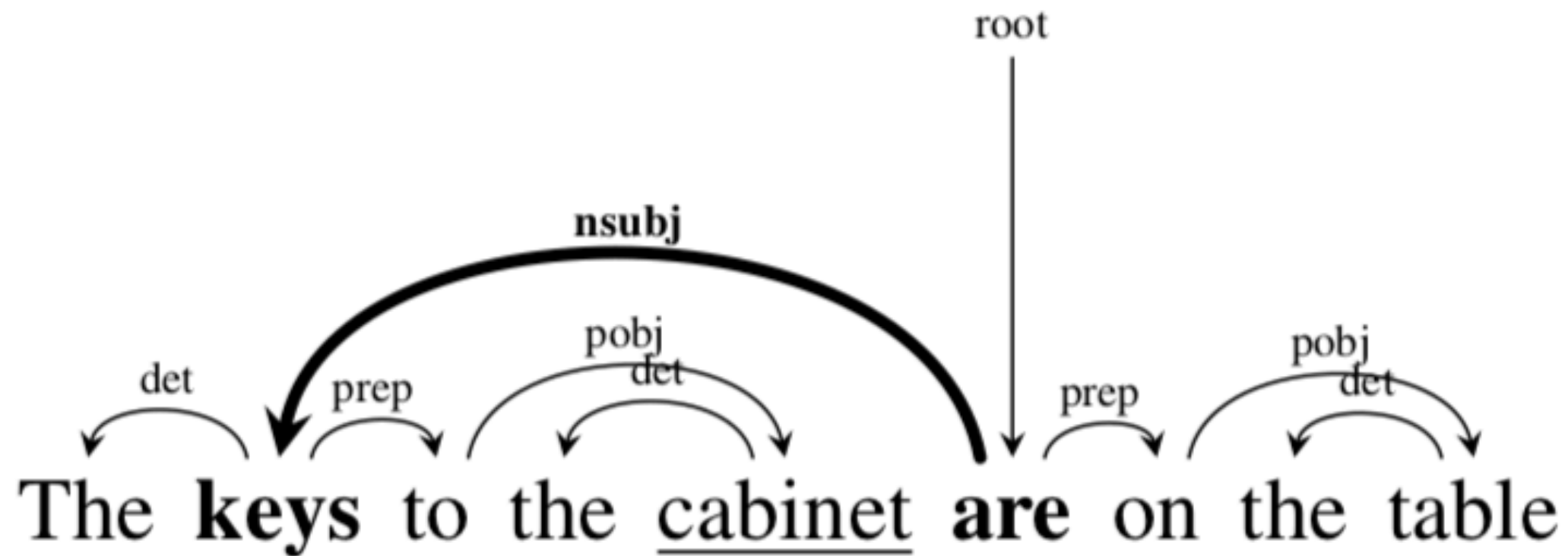
# analyzing neural networks

- Questions:
  - what input has most impact on the prediction?
  - what information is used by the system?  
=> identify what is represented
  - can we explain specific predictions?
  - what (minimal) changes to some examples lead to different predictions?

# What input has most influence?

- Main strategies:
  1. Select specific input data
  2. track the signal:
    - which dimensions are lighting up
    - which parts of the input provided these strong signals

# Select specific input



Linzen et al. (2016)

# Heatmaps

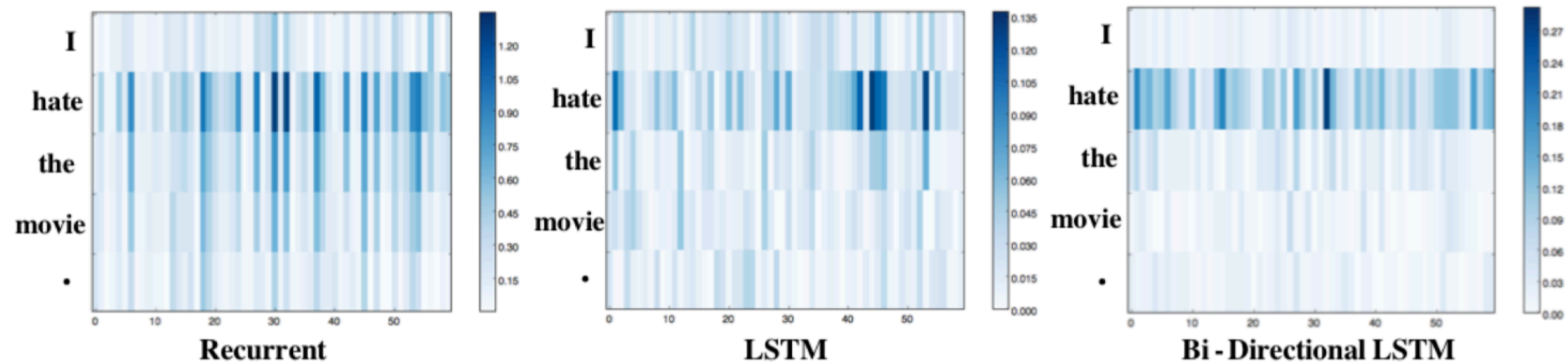
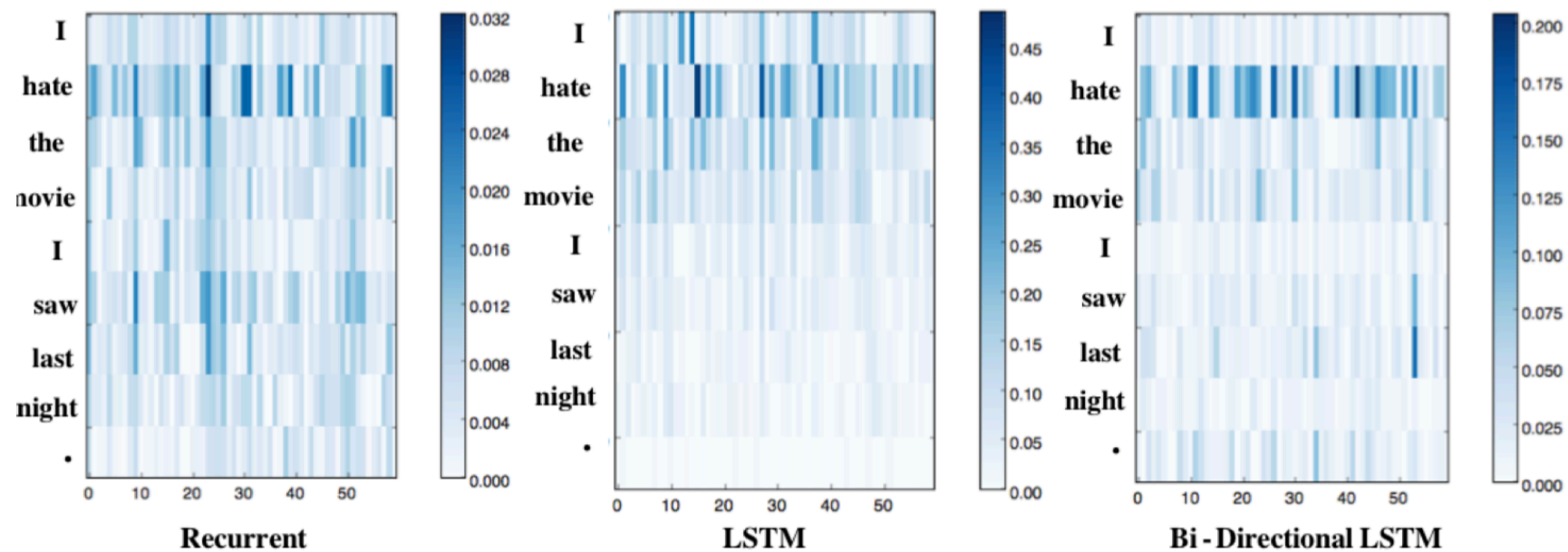
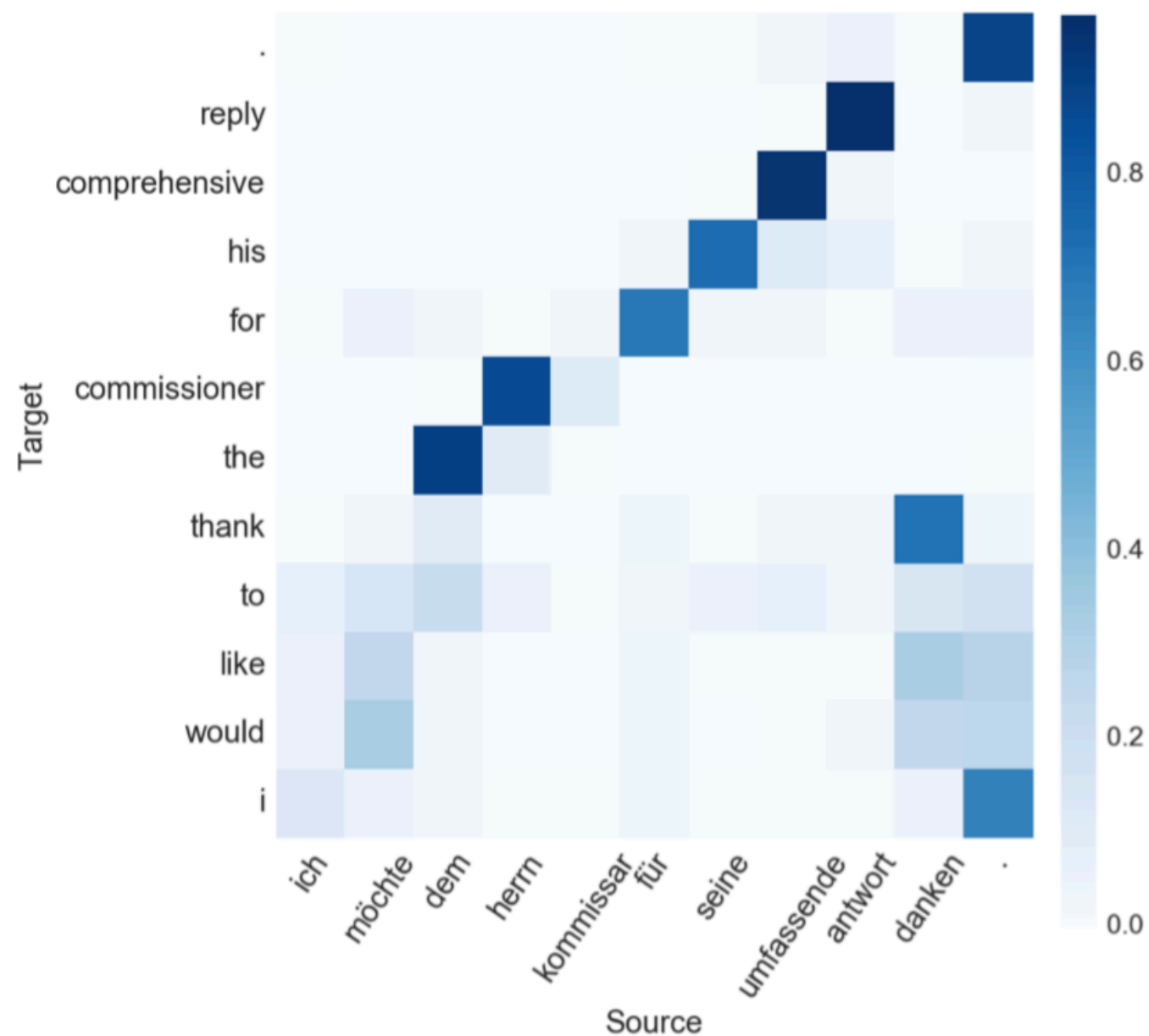


Figure 5: Saliency heatmap for for "I hate the movie ." Each row corresponds to saliency scores for the correspondent word representation with each grid representing each dimension.



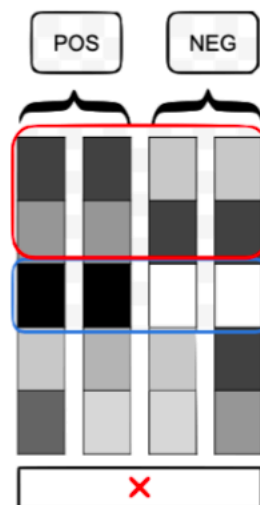
# Investigating Attention



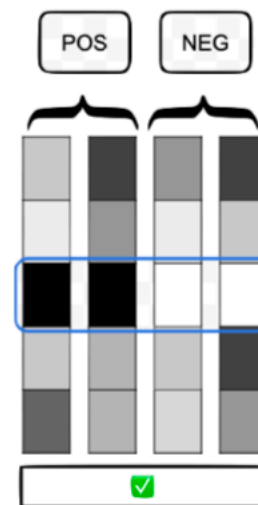
# What information is represented?

1. Diagnostic classifiers: can a basic machine learning algorithm predict specific information?
2. What patterns are found in a specific layer and how do they relate to embeddings trained for a specific task?

# Diagnostic Classifiers



(a) Target property (present in black, absent in white) and other dimensions correlates with positive and negative classes.



(b) The only dimension correlating with the positive and negative classes is the target dimension of the target property.



(c) Highly similar positive and negative examples that can only be distinguished by the dimension of the target property

# SVCCA

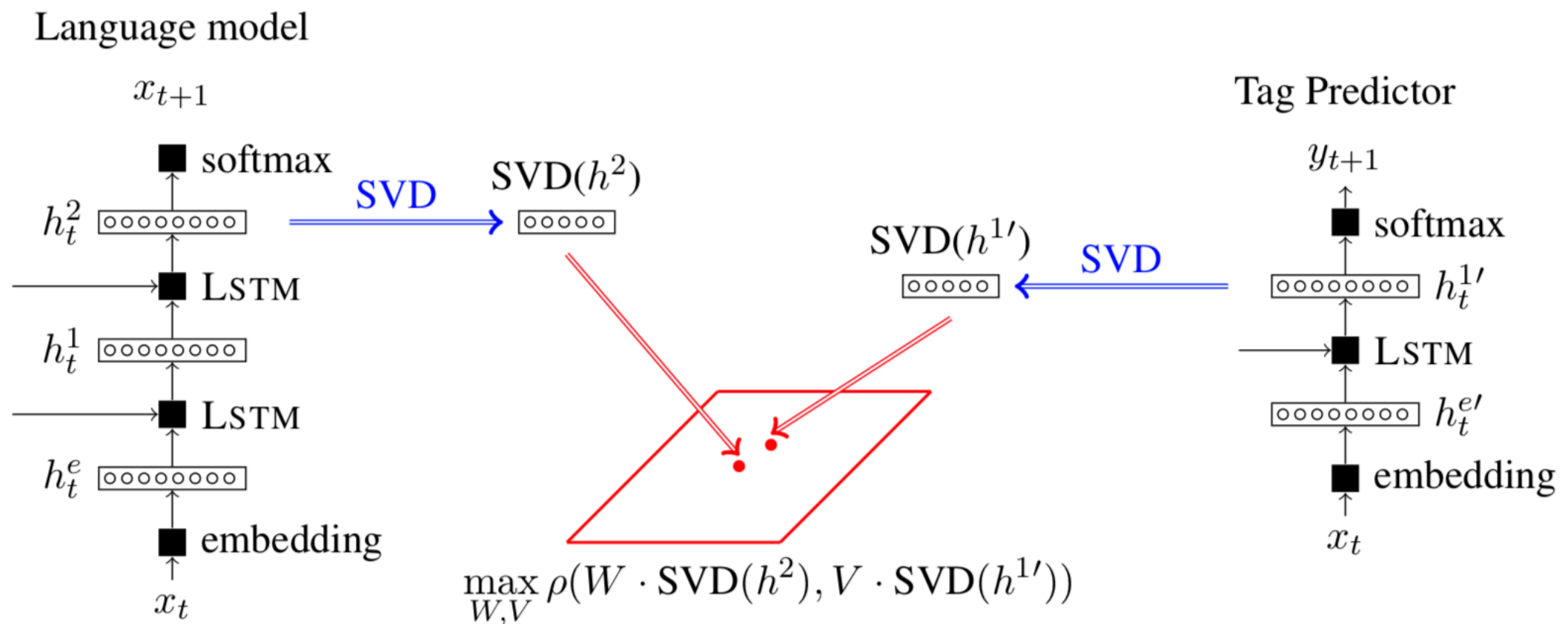


Figure 1: SVCCA used to compare the layer  $h^2$  of a language model and layer  $h^{1'}$  of a tagger.



# Explaining Predictions

- Identify landmarks and link those to representative examples in training data (Croce et al. 2018)

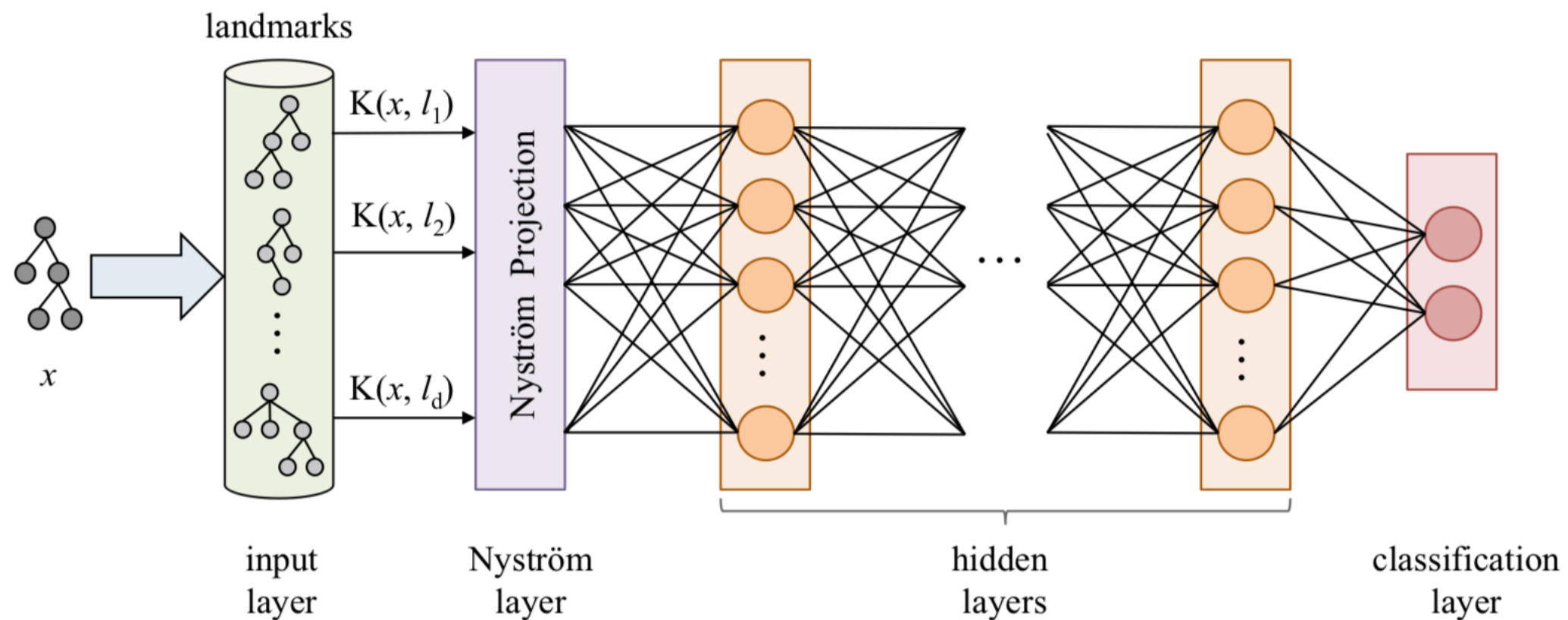


Figure 1: Kernel-based Deep Architecture.

# Explaining Predictions

- Identifying input - output pairs as explanations

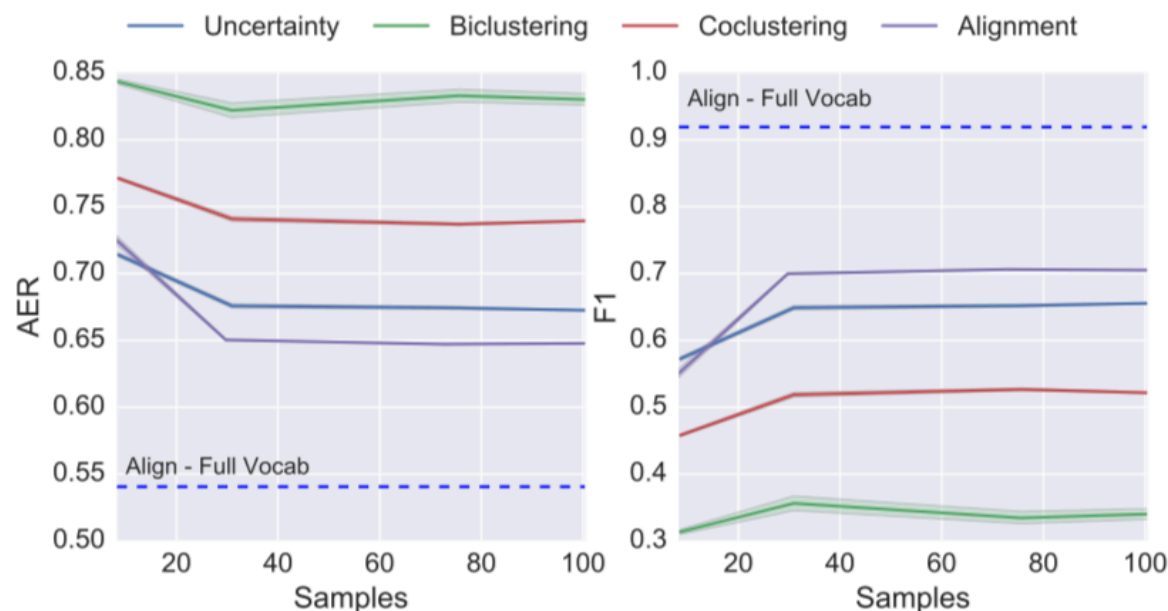


Figure 2: Arpabet test results as a function of number of perturbations used. Shown are mean plus confidence bounds over 5 repetitions. **Left:** Alignment Error Rate, **Right:** F1 over edge prediction.

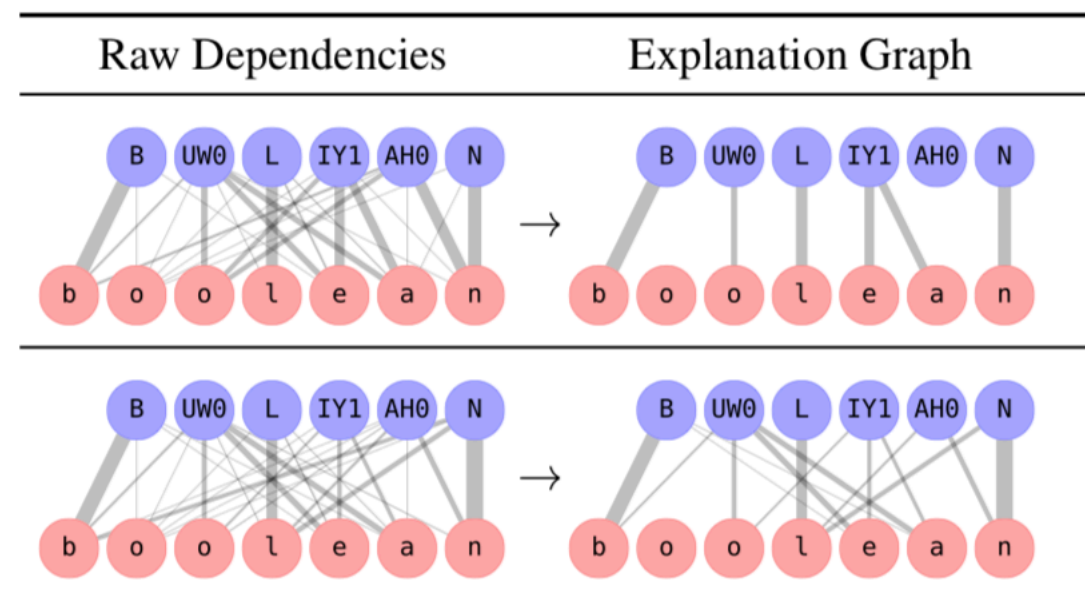
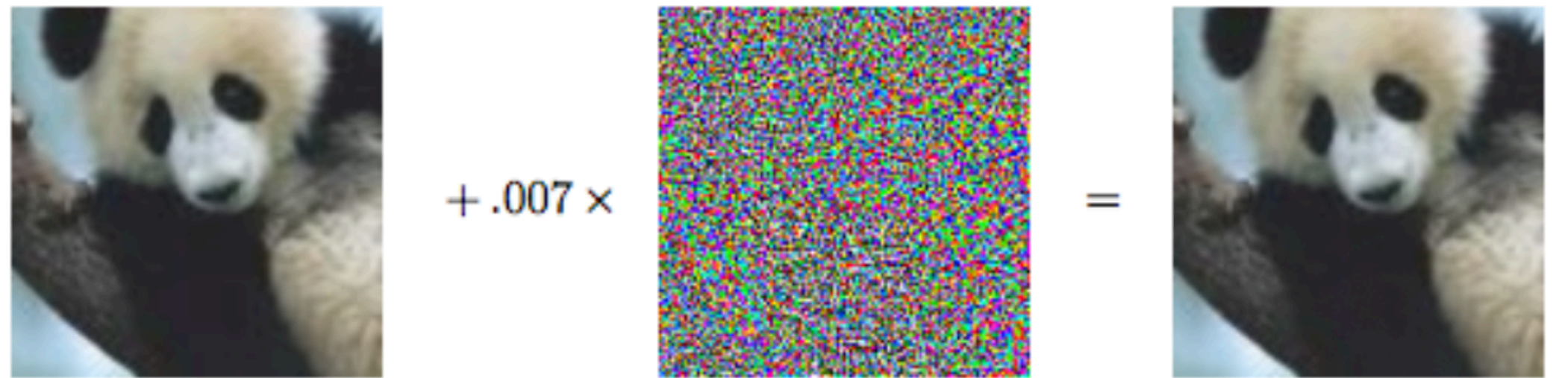


Table 1: Inferred dependency graphs before (left) and after (right) explanation selection for the prediction: *boolean*  $\mapsto$  B UW0 L IY1 AH0 N, in independent runs with large (top) and small (bottom) clustering parameter  $k$ .

# Adversarial Examples



$x$   
“panda”  
57.7% confidence

$+ .007 \times$

$\text{sign}(\nabla_x J(\theta, x, y))$   
“nematode”  
8.2% confidence

$=$

$x + \epsilon \text{sign}(\nabla_x J(\theta, x, y))$   
“gibbon”  
99.3 % confidence

- Idea: identify the minimal modification that leads to a different prediction

# Adversarial Examples (NLP)

Transformation Rules	#Flips		
$(WP\ is \rightarrow WP's)$	70 (1%)	<b>Original:</b> What is the oncorhynchus also called? <b>A:</b> chum salmon	<b>Original:</b> How long is the Rhine? <b>A:</b> 1,230 km
$(? \rightarrow ??)$	202(3%)	<b>Changed:</b> <b>What's</b> the oncorhynchus also called? <b>A:</b> keta	<b>Changed:</b> How long is the Rhine? <b>?</b> <b>A:</b> more than 1,050,000

(a) Example Rules

(b) Example for  $(WP\ is \rightarrow WP's)$

(c) Example for  $(? \rightarrow ??)$

Figure 2: **Semantically Equivalent Adversarial Rules:** For the task of question answering, the proposed approach identifies transformation rules for questions in (a) that result in paraphrases of the queries, but lead to incorrect answers (#Flips is the number of times this happens in the validation data). We show examples of rephrased questions that result in incorrect answers for the two rules in (b) and (c).

Ribeiro et al. (2018)

<https://www.aclweb.org/anthology/P18-1079>

# Adversarial Examples (NLP)

In the United States especially, several high-profile cases such as Debra LaFave, Pamela Rogers, and Mary Kay Letourneau have caused increased scrutiny on teacher misconduct.

(a) Input Paragraph

Q: What has been the result of this publicity?  
A: increased scrutiny on teacher misconduct

(b) Original Question and Answer

Q: What **haL** been the result of this publicity?  
A: **teacher misconduct**

(c) Adversarial Q & A (Ebrahimi et al., 2018)

Q: **What's** been the result of this publicity?  
A: **teacher misconduct**

(d) **Semantically Equivalent Adversary**

Figure 1: Adversarial examples for question answering, where the model predicts the correct answer for the question and input paragraph (1a and 1b). It is possible to fool the model by adversarially changing a single character (1c), but at the cost of making the question nonsensical. A **Semantically Equivalent Adversary** (1d) results in an incorrect answer while preserving semantics.



# Towards CLEARBOX?

- Can we also do proofs on:
  - what neural networks represent
  - how they learn?
- Future work: collaboration with Bettina Speckmann's group (TU Eindhoven)
  - expertise on geometric algebra
  - expertise on visual analytics

# Learning more

- The following paper provides a good starting point to find various lines of research on analyzing neural networks:

Belinkov, Yonatan, and James Glass. "Analysis methods in neural language processing: A survey." *Transactions of the Association for Computational Linguistics* 7 (2019): 49-72.

- This paper provides an overview of the first international workshop on this topic:

Alishahi, Afra, Grzegorz Chrupala, and Tal Linzen. "Analyzing and Interpreting Neural Networks for NLP: A Report on the First BlackboxNLP Workshop." *Natural Language Engineering* (2019).

# References

- Alvarez-Melis, David, and Tommi Jaakkola. "A causal framework for explaining the predictions of black-box sequence-to-sequence models." In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 412-421. 2017.
- Croce, Danilo, Daniele Rossini, and Roberto Basili. "Explaining non-linear classifier decisions within kernel-based deep architectures." In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pp. 16-24. 2018.
- Ghader, Hamidreza, and Christof Monz. "What does Attention in Neural Machine Translation Pay Attention to?." In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 30-39. 2017.
- Li, Jiwei, Xinlei Chen, Eduard Hovy, and Dan Jurafsky. "Visualizing and understanding neural models in nlp." *arXiv preprint arXiv:1506.01066* (2015).
- Linzen, Tal, Emmanuel Dupoux, and Yoav Goldberg. "Assessing the ability of LSTMs to learn syntax-sensitive dependencies." *Transactions of the Association for Computational Linguistics* 4 (2016): 521-535.
- Nissim, Malvina, Rik van Noord, and Rob van der Goot. "Fair is Better than Sensational: Man is to Doctor as Woman is to Doctor." *arXiv preprint arXiv:1905.09866* (2019).
- Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "Semantically equivalent adversarial rules for debugging nlp models." In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 856-865. 2018.
- Saphra, Naomi, and Adam Lopez. "Understanding Learning Dynamics Of Language Models with SVCCA." *arXiv preprint arXiv:1811.00225* (2018).
- Sommerauer, Pia, and Antske Fokkens. "Firearms and Tigers are Dangerous, Kitchen Knives and Zebras are Not: Testing whether Word Embeddings Can Tell." In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pp. 276-286. 2018.
- Tongshuang Wu, Marco Tulio Ribeiro, Jeffrey Heer and Daniel Weld (2019) Errudite: Scalable, Reproducible, and Testable Error Analysis. *Proceedings of ACL*