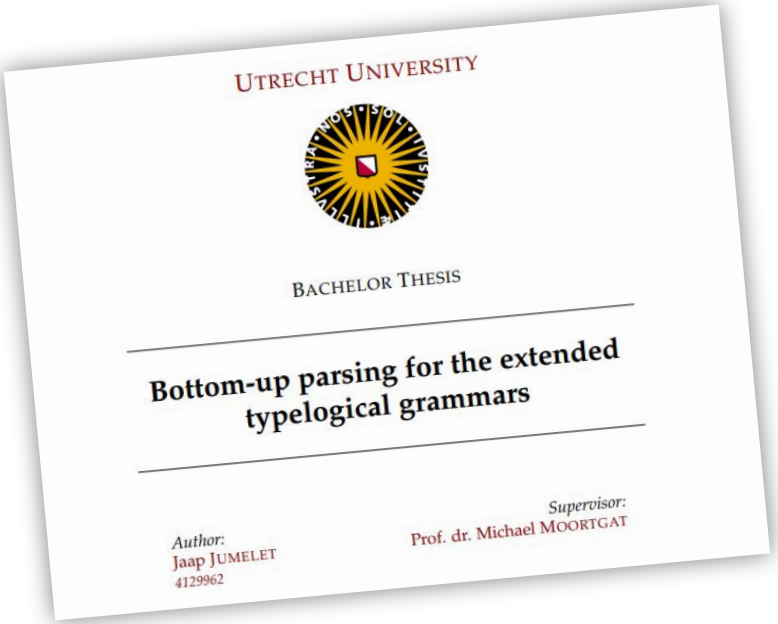
NLP1

Evaluating and Interpreting Language Models

Jaap Jumelet | ILLC, University of Amsterdam

Who Am I?

- BSc Artificial Intelligence at *Universiteit Utrecht (2013-2017)*
  - *Logic*
  - *(Computational) Linguistics*
  - *Theoretical Computer Science*


UTRECHT UNIVERSITY

BACHELOR THESIS

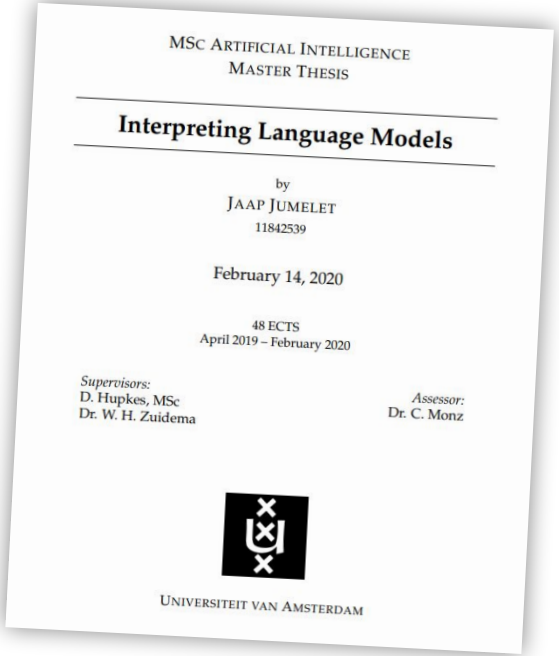
Bottom-up parsing for the extended typelogical grammars

Author:
Jaap JUMELET
4129962

Supervisor:
Prof. dr. Michael MOORTGAT

Who Am I?

- BSc Artificial Intelligence at *Universiteit Utrecht (2013-2017)*
- MSc Artificial Intelligence at the *UvA (2017 - 2020)*
  - *Natural Language Processing*
  - *Machine/Deep Learning*
  - *Explainable AI*



MSc Artificial Intelligence
Master Thesis

Interpreting Language Models

by
Jaap Jumelet
11842539

February 14, 2020

48 ECTS
April 2019 – February 2020

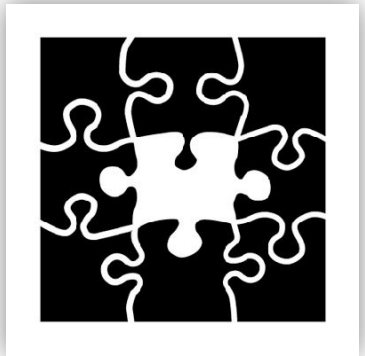
Supervisors:
D. Hupkes, MSc
Dr. W. H. Zuidema

Assessor:
Dr. C. Monz

Universiteit van Amsterdam

Who Am I?

- BSc Artificial Intelligence at *Universiteit Utrecht (2013-2017)*
- MSc Artificial Intelligence at the *UvA (2017 - 2020)*
- PhD candidate at the *Institute for Logic Language, and Computation (ILLC)* at the University of Amsterdam with **Jelle Zuidema**
- Interested in:
  - Language models *(but who isn't, nowadays...)*
  - Interpretability
  - (Psycho-)linguistics & NLP
  - Grammar / Hierarchical Structure



Plan for today

- Interpretability
  - *Why* do we need interpretability?
  - What is an **explanation**?
  - Explanation **faithfulness**
- Interpretability Methods
  - Behavioural studies
  - Probing
  - Feature Attributions

Why do we need interpretability?

Let's take a step back to **2001**

Statistical Science
2001, Vol. 16, No. 3, 199–231

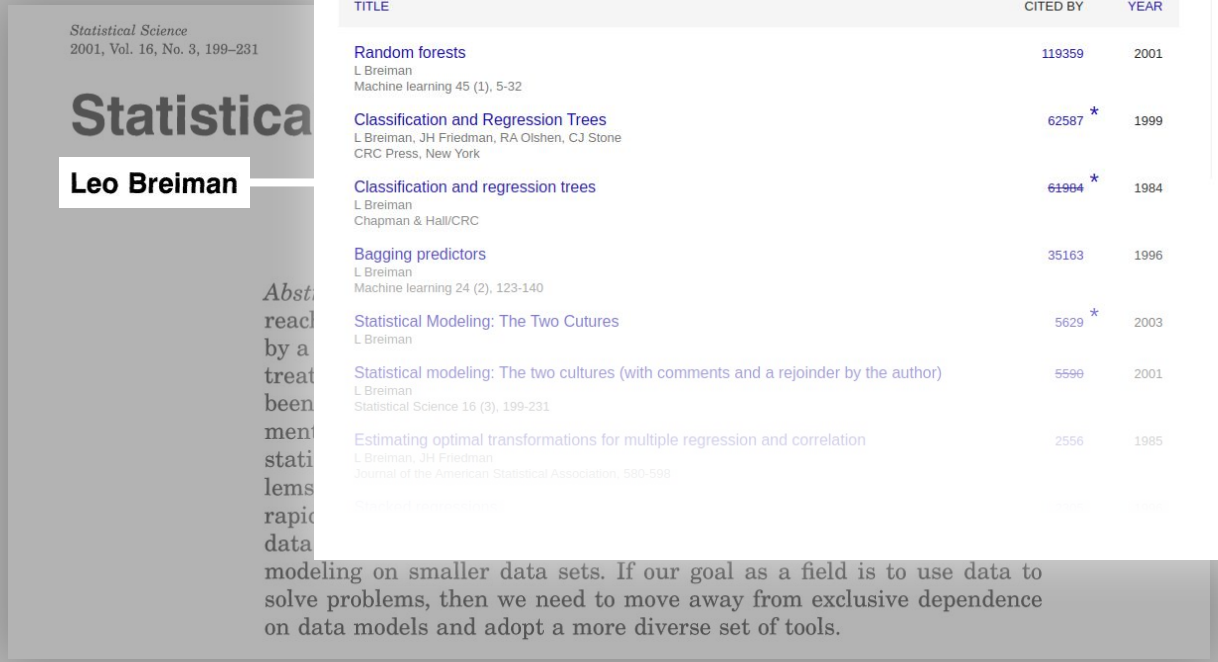
Statistical Modeling: The Two Cultures

Leo Breiman

Abstract. There are two cultures in the use of statistical modeling to reach conclusions from data. One assumes that the data are generated by a given stochastic data model. The other uses algorithmic models and treats the data mechanism as unknown. The statistical community has been committed to the almost exclusive use of data models. This commitment has led to irrelevant theory, questionable conclusions, and has kept statisticians from working on a large range of interesting current problems. Algorithmic modeling, both in theory and practice, has developed rapidly in fields outside statistics. It can be used both on large complex data sets and as a more accurate and informative alternative to data modeling on smaller data sets. If our goal as a field is to use data to solve problems, then we need to move away from exclusive dependence on data models and adopt a more diverse set of tools.

Why do we need interpretability?

Let's take a step back



Statistical Science
2001, Vol. 16, No. 3, 199–231

Statistica

Leo Breiman

Abst
reach
by a
treat
been
ment
stati
lems
rapi
data
modeling on smaller data sets. If our goal as a field is to use data to solve problems, then we need to move away from exclusive dependence on data models and adopt a more diverse set of tools.

Leo Breiman 1928-2005

FOLLOW

Professor of Statistics, UC Berkeley
Verified email at stat.berkeley.edu - Homepage
Data Analysis Statistics Machine Learning

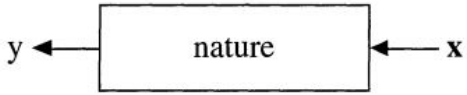
TITLE	CITED BY	YEAR
Random forests L Breiman Machine learning 45 (1), 5-32	119359	2001
Classification and Regression Trees L Breiman, JH Friedman, RA Olshen, CJ Stone CRC Press, New York	62587 *	1999
Classification and regression trees L Breiman Chapman & Hall/CRC	~~61984~~ *	1984
Bagging predictors L Breiman Machine learning 24 (2), 123-140	35163	1996
Statistical Modeling: The Two Cutures L Breiman	5629 *	2003
Statistical modeling: The two cultures (with comments and a rejoinder by the author) L Breiman Statistical Science 16 (3), 199-231	~~5590~~	2001
Estimating optimal transformations for multiple regression and correlation L Breiman, JH Friedman Journal of the American Statistical Association, 580-598	2556	1985

Cited by VIEW ALL

	All	Since 2018
Citations	249340	133584
h-index	53	40
i10-index	85	45

Why do we need interpretability?

Statistics starts with data. Think of the data as being generated by a black box in which a vector of input variables **x** (independent variables) go in one side, and on the other side the response variables **y** come out. Inside the black box, nature functions to associate the predictor variables with the response variables, so the picture is like this:



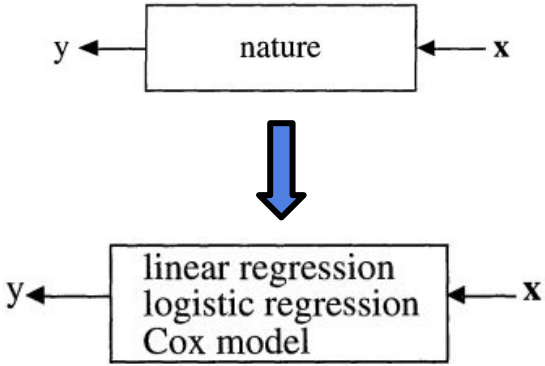
There are two goals in analyzing the data:

Prediction. To be able to predict what the responses are going to be to future input variables;
Information. To extract some information about how nature is associating the response variables to the input variables.

Breiman (2001) - *Statistical Modeling: The Two Cultures*

Why do we need interpretability?

The **Data Modelling** culture:



Assumes an **explicit** and ***interpretable*** relationship between input *x* and output *y*

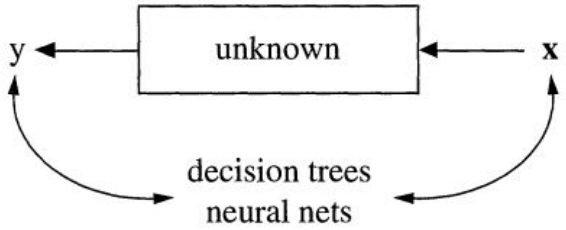
Model validation. Yes–no using goodness-of-fit tests and residual examination.
Estimated culture population. 98% of all statisticians.

Breiman (2001) - *Statistical Modeling: The Two Cultures*

Why do we need interpretability?

The **Algorithmic Modelling** culture:

The analysis in this culture considers the inside of the box complex and unknown. Their approach is to find a function $f(\mathbf{x})$—an algorithm that operates on $\mathbf{x}$ to predict the responses $\mathbf{y}$. Their black box looks like this:



Model validation. Measured by predictive accuracy.
Estimated culture population. 2% of statisticians, many in other fields.

Breiman (2001) - *Statistical Modeling: The Two Cultures*

Why do we need interpretability?

Why do we need interpretability?

The **desiderata** of algorithmic models:

1. Fairness
   - *What biases does it contain? Does it discriminate against particular groups?*
2. Trustworthiness
   - *Models that are deployed carry a degree of responsibility, can we trust them?*
3. Robustness
   - *Does our model generalise robustly to unseen data?*
4. Faithfulness
   - *How faithful are model explanations to its actual reasoning?*

Lipton (2018) - *The Mythos of Model Interpretability*

NOS Nieuws • Zondag 18 juni 2023, 06:29 •
Aangepast maandag 19 juni 2023, 12:34

Nauwelijks zicht op 'zwarte zoemende dozen' van overheid: 'Algoritmeregister wassen neus'

Hatixhe Raba
redacteur Binnenland

Het algoritmeregister, dat de overheid transparanter moet maken, wordt een half jaar na de lancering nauwelijks ingevuld. En de informatie die er wel in staat, is niet erg toegankelijk.

In het register maken overheidsinstellingen, bijvoorbeeld gemeentes en ministeries, openbaar hoe ze tot (semi-)geautomatiseerde besluiten komen. Het algoritmegebruik door de overheid ligt onder een vergrootglas door de toeslagenaffaire, waarin dubbele nationaliteit werd meegewogen in een algoritme voor risicoanalyse van toeslagaanvragen.

Bij de lancering in december 2022 stonden er 109 algoritmes op de website. Die

Why do we need

The **desiderata** of algo

1. Fairness
   - *What biases does* *st particular groups?*
2. Trustworthiness
   - *Models that are de* *y, can we trust them?*
3. Robustness
   - *Does our model ge*
4. Faithfulness
   - *How faithful are m* *ing?*

) - The Mythos of Model Interpretability

BBC

NEWS Menu

Business | Market Data | New Tech Economy | Technology of Bu

Apple's 'sexist' credit card investigated by US regulator

11 November 2019

A US financial regulator has opened an investigation into claims Apple's credit card offered different credit limits for men and women.

It follows complaints - including from Apple's co-founder Steve Wozniak - that algorithms used to set limits might be inherently biased against women.

New York's Department of Financial Services (DFS) has

Why do we need

The **desiderata** of algori

1. Fairness
   - *What biases does it* *particular groups?*
2. Trustworthiness
   - *Models that are depr* *can we trust them?*
3. Robustness
   - *Does our model gen*
4. Faithfulness
   - *How faithful are moc* *g?*

BBC

NEWS

Menu

Tech

Facebook apology as AI labels black men 'primates'

6 September 2021

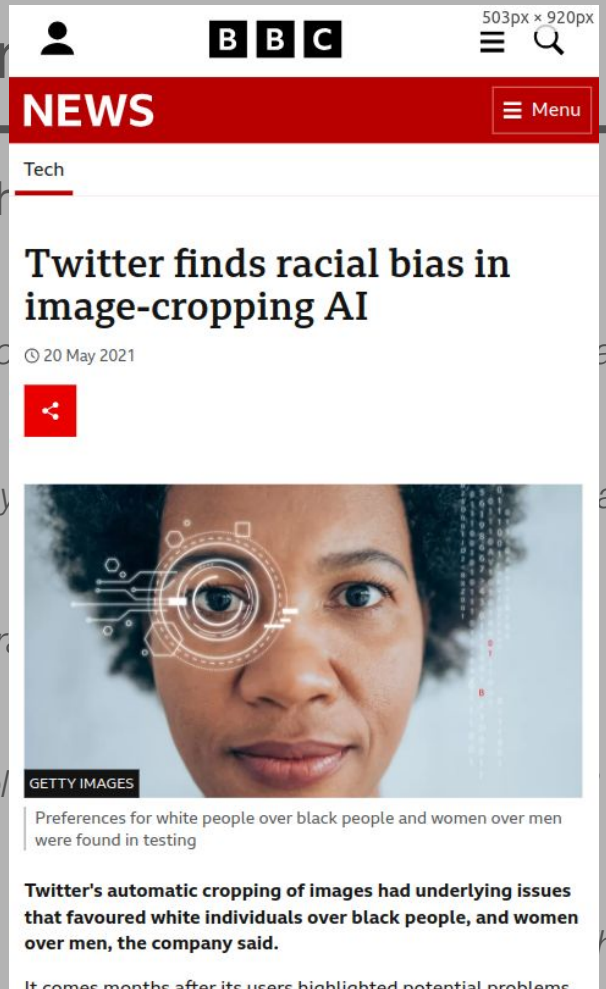
Facebook users who watched a newspaper video featuring black men were asked if they wanted to "keep seeing videos about primates" by an artificial-intelligence recommendation system.

Facebook told BBC News it "was clearly an unacceptable error", disabled the system and launched an investigation.

Why do we need in

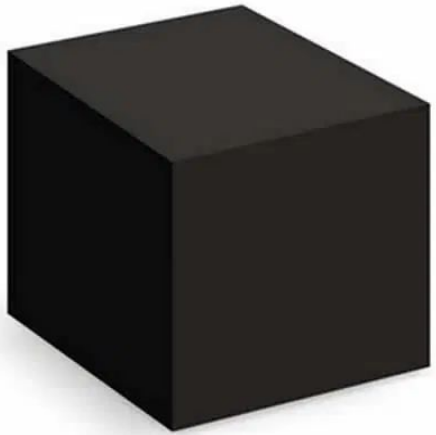
The **desiderata** of algorith

1. Fairness
   - *What biases does it co articular groups?*
2. Trustworthiness
   - *Models that are deploy an we trust them?*
3. Robustness
   - *Does our model gener*
4. Faithfulness
   - *How faithful are model*



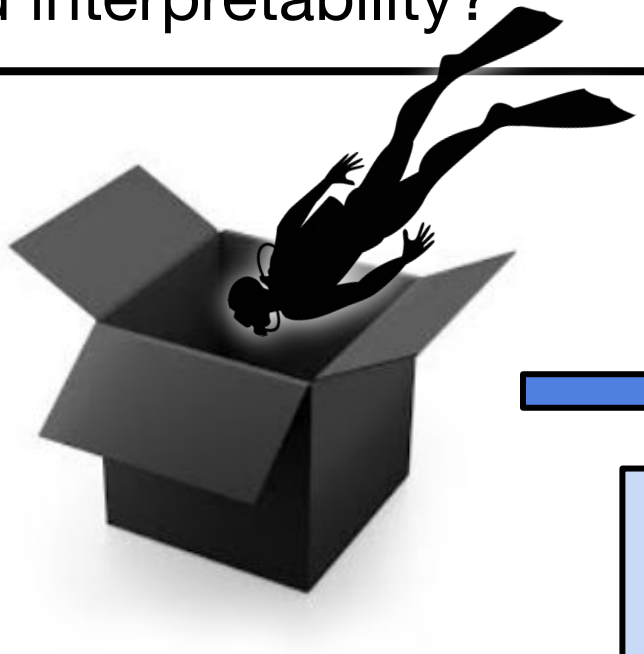
he Mythos of Model Interpretability

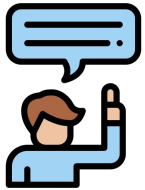
Why do we need interpretability?



?

Why do we need interpretability?





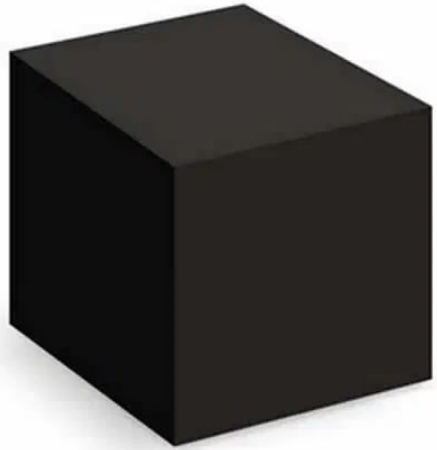
Post-hoc
Explanation

How do we explain a model?

Henk is a 36 year old male lawyer from Amsterdam

How do we explain a model?

Henk is a 36 year old male lawyer from Amsterdam



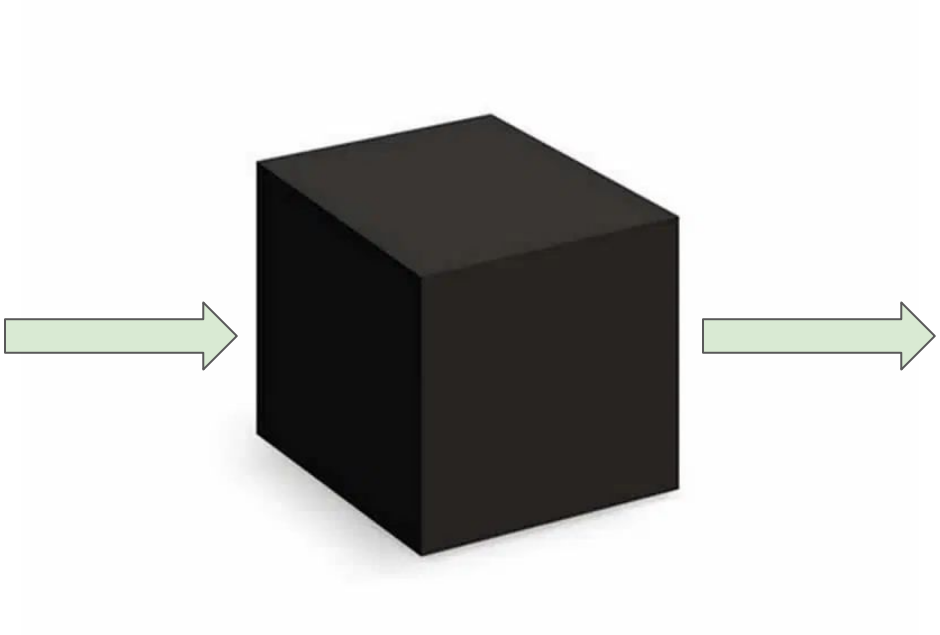
€5000 credit

How do we explain a model?

Suzan is a 32 year old female doctor from Utrecht

How do we explain a model?

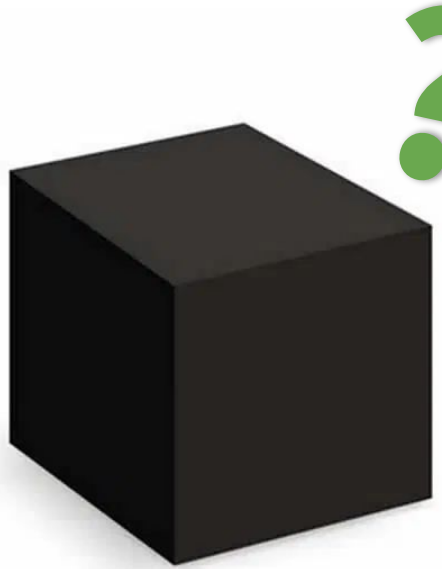
Suzan is a 32 year old female doctor from Utrecht



€1000 credit

How do we explain a model?

Suzan is a 32 year old female doctor from Utrecht

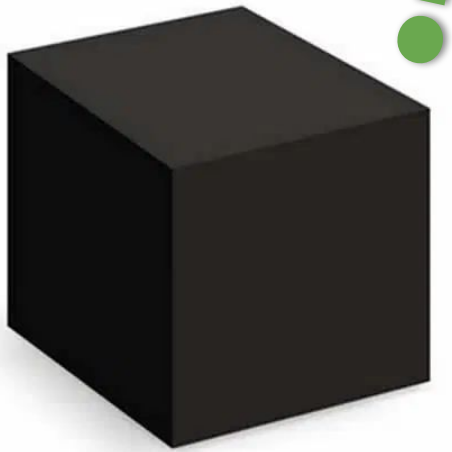


€1000 credit

Why does Suzan get less than Henk? Because of her **age**? **Gender**? **Occupation**? **Location**?

How do we explain a model?

Suzan is a 32 year old female doctor from Utrecht



?

female : **0.90**
32 : 0.05
doctor : 0.03
Utrecht : 0.02

€1000 credit

How do we explain a model?

- We often explain events by pointing out the most **important** factors
- This is often done in **contrast** to a neutral **baseline**

Input: *Can you stop the dog from*
Output: barking

1. Why did the model predict "barking"?
Can you stop the dog from

2. Why did the model predict "barking" *instead of* "crying"?
Can you stop the dog from

3. Why did the model predict "barking" *instead of* "walking"?
Can you stop the dog from

Yin & Neubig (2022) - *Interpreting Language Models with Contrastive Explanations*

Explanation Faithfulness

How do we ensure that a model explanation actually represents a model's reasoning?

Explanation Faithfulness

How do we ensure that a model explanation actually represents a model's reasoning?

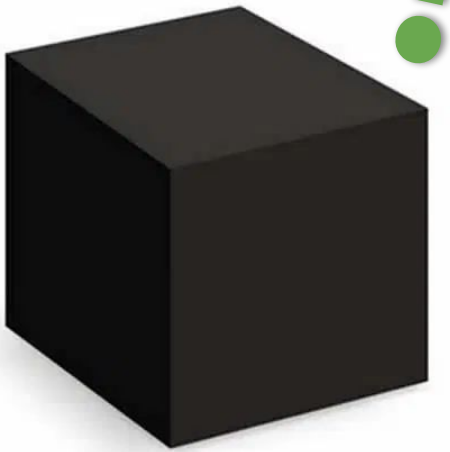
Plausibility does **not** imply faithfulness!

Models can be *right for the wrong reasons*!

But how do we ever know our explanation is truly faithful to the model?

Explanation Faithfulness

Suzan is a 32 year old female doctor from Utrecht



female	: 0.02
32	: **0.93**
doctor	: 0.03
Utrecht	: 0.02

€1000 credit

Explanation Methods

Levels of explanation *granularity*:

1. **Behavioural**
   - Model remains a black-box
   - Predictions of model are the main object of interest

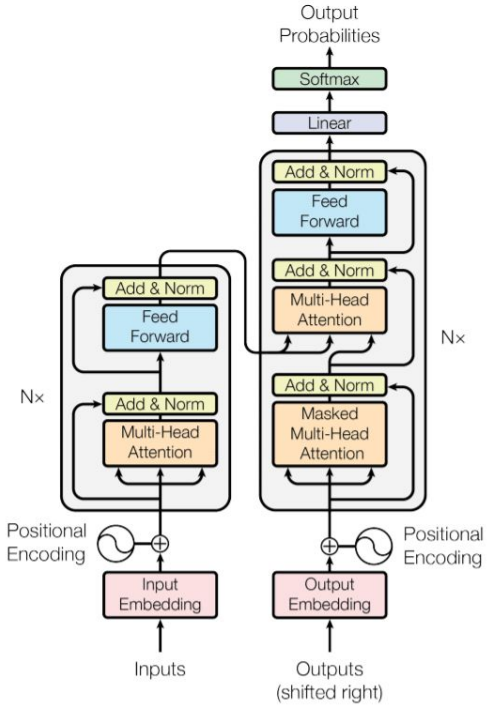


Figure 1: The Transformer - model architecture.

Explanation Methods

Levels of explanation *granularity*:

1. **Behavioural**
   - Model remains a black-box
   - Predictions of model are the main object of interest
2. **Attributional**
   - Which input features were most *important* for a prediction?

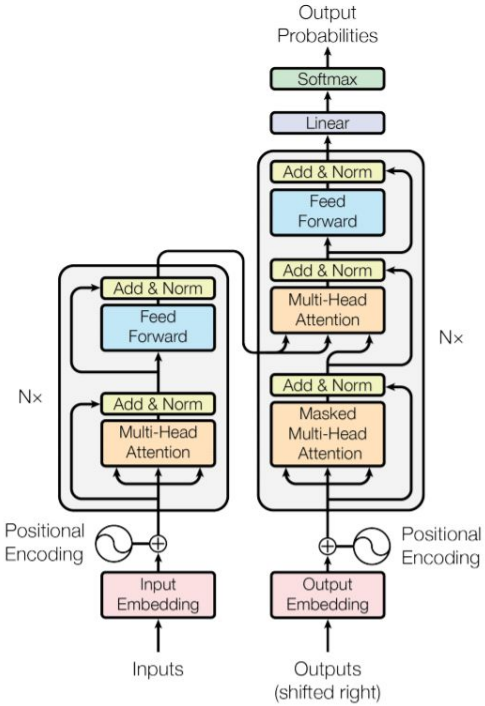


Figure 1: The Transformer - model architecture.

Explanation Methods

Levels of explanation *granularity*:

1. **Behavioural**
   - Model remains a black-box
   - Predictions of model are the main object of interest
2. **Attributional**
   - Which input features were most *important* for a prediction?
3. **Probing**
   - What abstract features are encoded by the model?
   - Performed layer-wise

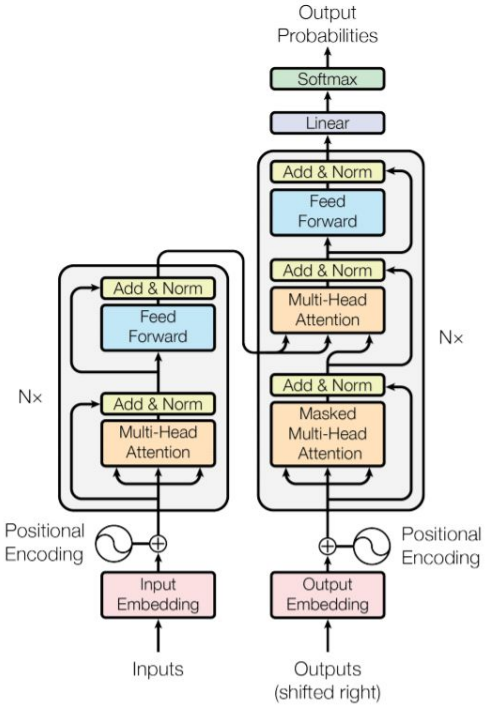


Figure 1: The Transformer - model architecture.

Explanation Methods

Levels of explanation *granularity*:

1. **Behavioural**
   - Model remains a black-box
   - Predictions of model are the main object of interest
2. **Attributional**
   - Which input features were most *important* for a prediction?
3. **Probing**
   - What abstract features are encoded by the model?
   - Performed layer-wise
4. **Mechanistic**
   - Can we identify specific *circuits* responsible for a particular behaviour?

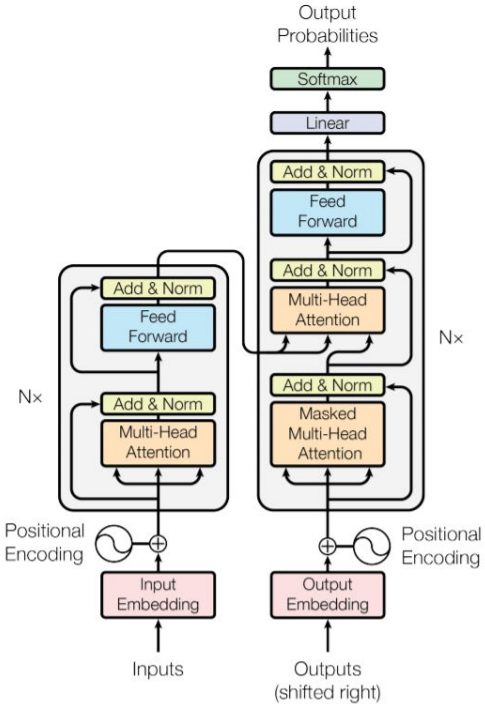


Figure 1: The Transformer - model architecture.

Behavioural Interpretability

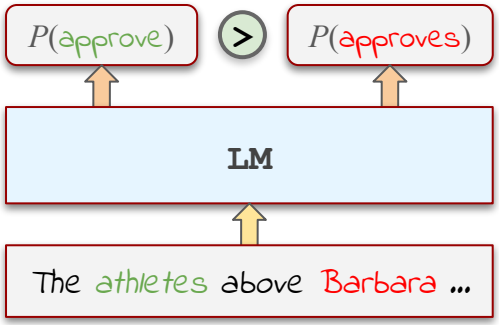
How can we understand a model better, without ‘opening the black box’?

- Using carefully crafted **minimal pairs** we can investigate a model’s performance on a specific phenomenon.

Behavioural Interpretability

How can we understand a model better, without ‘opening the black box’?

- Using carefully crafted **minimal pairs** we can investigate a model’s performance on a specific phenomenon.
- This type of experiment only requires access to the **output probabilities** of the model.



BLiMP

- The **Benchmark of Linguistic Minimal Pairs for English**
- Tests the capacity of language models for a wide range of *linguistic phenomena*
- Allows us to test and compare language model performance regardless of size
- Comparison done based on *sentence probability:*

$$\mathrm{P}(\text{grammatical sentence}) > \mathrm{P}(\text{ungrammatical sentence})$$

Warstadt et al. (2020) - BLiMP: The Benchmark of Linguistic Minimal Pairs for English

BLiMP

Phenomenon	N	Acceptable Example	Unacceptable Example
ANAPHOR AGR.	2	*Many girls insulted themselves.*	*Many girls insulted herself.*

Warstadt et al. (2020) - BLiMP: The Benchmark of Linguistic Minimal Pairs for English

BLiMP

Phenomenon	N	Acceptable Example	Unacceptable Example
ANAPHOR AGR.	2	*Many girls insulted themselves.*	*Many girls insulted herself.*
ARG. STRUCTURE	9	*Rose wasn't disturbing Mark.*	*Rose wasn't boasting Mark.*
BINDING	7	*Carlos said that Lori helped him.*	*Carlos said that Lori helped himself.*

Warstadt et al. (2020) - BLiMP: The Benchmark of Linguistic Minimal Pairs for English

BLiMP

Phenomenon	N	Acceptable Example	Unacceptable Example
ANAPHOR AGR.	2	*Many girls insulted themselves.*	*Many girls insulted herself.*
ARG. STRUCTURE	9	*Rose wasn't disturbing Mark.*	*Rose wasn't boasting Mark.*
BINDING	7	*Carlos said that Lori helped him.*	*Carlos said that Lori helped himself.*
CONTROL/RAISING	5	*There was bound to be a fish escaping.*	*There was unable to be a fish escaping.*
DET.-NOUN AGR.	8	*Rachelle had bought that chair.*	*Rachelle had bought that chairs.*

Warstadt et al. (2020) - BLiMP: The Benchmark of Linguistic Minimal Pairs for English

BLiMP

Phenomenon	N	Acceptable Example	Unacceptable Example
ANAPHOR AGR.	2	*Many girls insulted themselves.*	*Many girls insulted herself.*
ARG. STRUCTURE	9	*Rose wasn't disturbing Mark.*	*Rose wasn't boasting Mark.*
BINDING	7	*Carlos said that Lori helped him.*	*Carlos said that Lori helped himself.*
CONTROL/RAISING	5	*There was bound to be a fish escaping.*	*There was unable to be a fish escaping.*
DET.-NOUN AGR.	8	*Rachelle had bought that chair.*	*Rachelle had bought that chairs.*
ELLIPSIS	2	*Anne's doctor cleans one important book and Stacey cleans a few.*	*Anne's doctor cleans one book and Stacey cleans a few important.*
FILLER-GAP	7	*Brett knew what many waiters find.*	*Brett knew that many waiters find.*
IRREGULAR FORMS	2	*Aaron broke the unicycle.*	*Aaron broken the unicycle.*
ISLAND EFFECTS	8	*Which bikes is John fixing?*	*Which is John fixing bikes?*
NPI LICENSING	7	*The truck has clearly tipped over.*	*The truck has ever tipped over.*
QUANTIFIERS	4	*No boy knew fewer than six guys.*	*No boy knew at most six guys.*
SUBJECT-VERB AGR.	6	*These casseroles disgust Kayla.*	*These casseroles disgusts Kayla.*

Table 1: Minimal pairs from each of the twelve linguistic phenomenon categories covered by BLiMP. Differences are underlined. *N* is the number of 1,000-example minimal pair paradigms within each broad category.

Warstadt et al. (2020) - BLiMP: The Benchmark of Linguistic Minimal Pairs for English

BLiMP

Model	Overall	Ana. agr	Arg. str	Binding	Ctrl. rais.	D-n agr	Ellipsis	Filler. gap	Irregular	Island	NPI	Quantifiers	S-v agr
5-gram	61.2	47.9	71.9	64.4	68.5	70.0	36.9	60.2	79.5	57.2	45.5	53.5	60.3
LSTM	69.8	91.7	73.2	73.5	67.0	85.4	67.6	73.9	89.1	46.6	51.7	64.5	80.1
TXL	69.6	94.1	72.2	74.7	71.5	83.0	77.2	66.6	78.2	48.4	55.2	69.3	76.0
GPT-2	83.0	99.3	81.8	80.9	81.9	95.8	89.3	81.3	91.9	72.7	76.8	79.0	86.4
Human	88.6	97.5	90.0	87.3	83.9	92.2	85.0	86.9	97.0	84.9	88.1	86.6	90.9

Table 3: Percentage accuracy of four baseline models and raw human performance on BLiMP using a forced-choice task. A random guessing baseline would achieve an accuracy of 50%.

Warstadt et al. (2020) - BLiMP: The Benchmark of Linguistic Minimal Pairs for English

BLiMP

Figure 1: Heatmap showing the correlation between models' accuracies in each of the 67 paradigms.

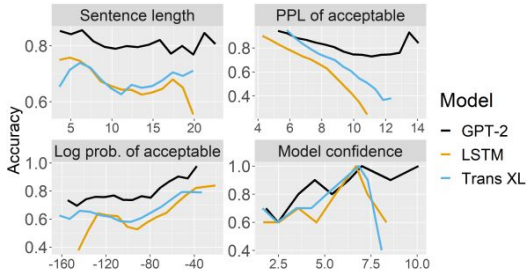


Figure 2: Models' performance on BLiMP as a function of sentence length, perplexity, log probability of the acceptable sentence, and model confidence (calculated as $|\log P(S_1) - \log P(S_2)|$).

Warstadt et al. (2020) - BLiMP: The Benchmark of Linguistic Minimal Pairs for English

BLiMP

Figure 1: Heatmap showing the correlation between models' accuracies in each of the 67 paradigms.

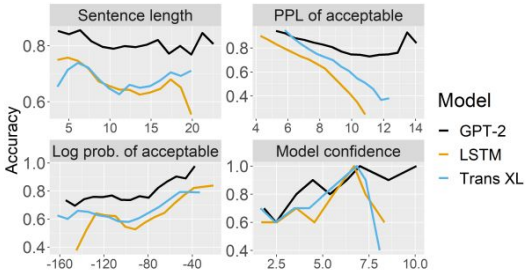
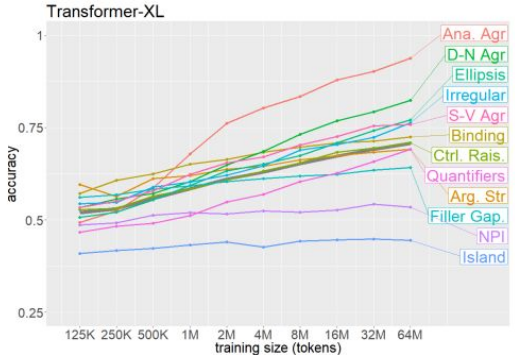


Figure 2: Models' performance on BLiMP as a function of sentence length, perplexity, log probability of the acceptable sentence, and model confidence (calculated as $|\log P(S_1) - \log P(S_2)|$).



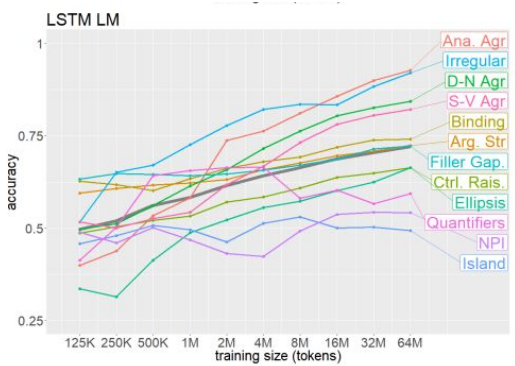


Figure 5: Transformer-XL (top) and LSTM LM (bottom) performance as a function of training size and phenomena in BLiMP. The gray line shows the average across all phenomena.

Warstadt et al. (2020) - BLiMP: The Benchmark of Linguistic Minimal Pairs for English

Behavioural Tests for Uncovering Biases

We can use behavioural tests to investigate how a model acquires behaviour during training.

Back in 2021 we ran this experiment:

- LSTM LM trained on 100M Wikipedia tokens
- Evaluated on BLiMP *during* training
- In particular on *anaphora agreement:*

  E.g. *Katherine can't help* herself / *himself

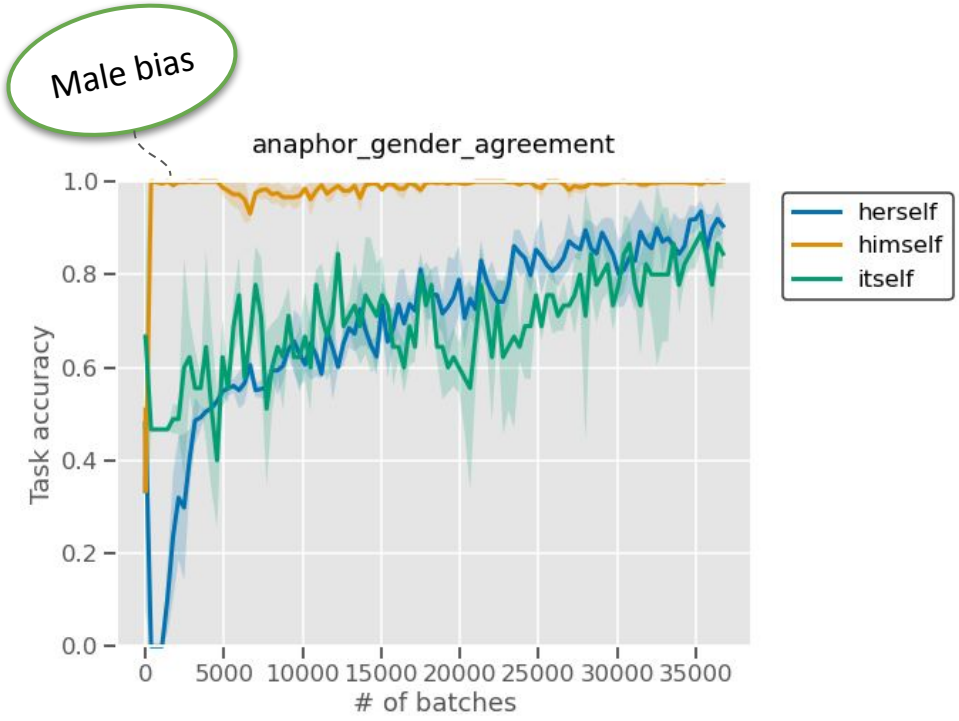
Behavioural Tests for Uncovering Biases

We can use behavioural tests to investigate how a model acquires behaviour during training.

Back in 2021 we ran this experiment:

- LSTM LM trained on 100M Wikipedia tokens
- Evaluated on BLiMP *during* training
- In particular on *anaphora agreement:*

  E.g. *Katherine can't help* herself / *himself



Van der Wal, Jumelet, Schulz & Zuidema (2021) - The Birth of Bias

Limitations of Behavioural Tests

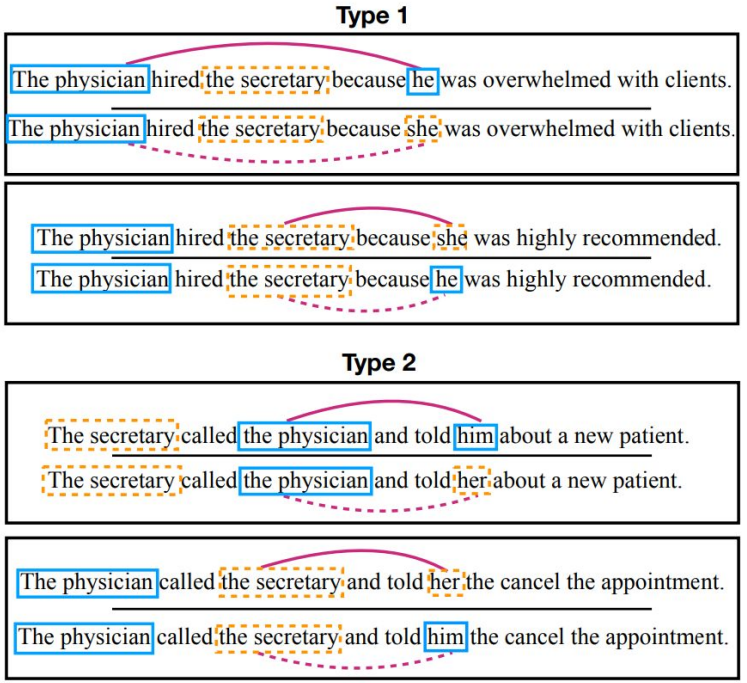
Behavioural tests show us a model's response to a particular input

- We now know roughly ***what*** a model can do.
- ***Why*** a model gave a particular response is not clear though!

Limitations of Behavioural Tests

Behavioural tests show us a model's response to a particular input

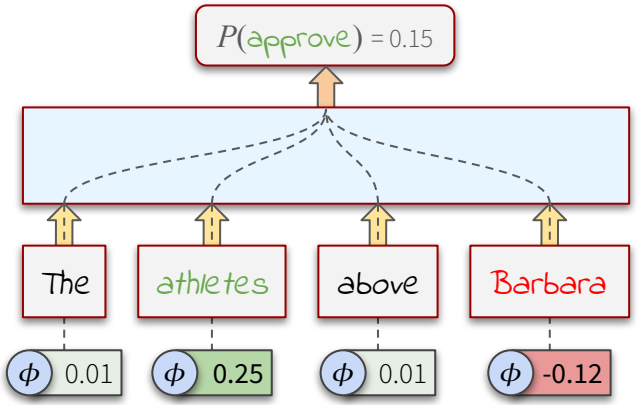
- We now know roughly ***what*** a model can do.
- ***Why*** a model gave a particular response is not clear though!
- Complex phenomena require more complex explanations
- E.g. coreference resolution:



Zhao et al. (2018) - *Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods*

Feature Attribution Methods

- **Feature attribution methods** explain model predictions in terms of the strongest *contributing* features.
- By normalizing such scores we get an insight into the relative importance of each feature.
- Shows us the *rationale* of a model behind a prediction → useful for uncovering biases!



Pronoun Resolution

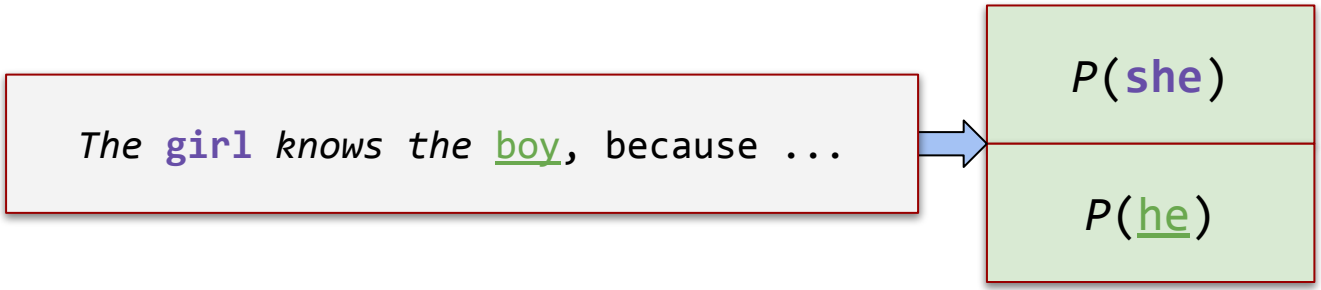
The **girl** knows the boy, because **she** had spoken to him earlier.

The **girl** knows the boy, because he lives next-door to **her**.

Jumelet et al. (2019) - *Analysing Neural Language Models: Contextual Decomposition Reveals Default Reasoning in Number and Gender Assignment*

Pronoun Resolution

Pronoun resolution:

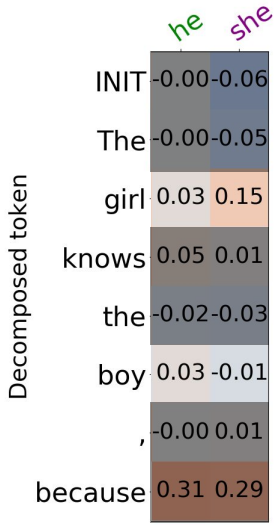


Pronoun Resolution

The girl *knows the* boy, because he/she

Pronoun Resolution

The girl knows the boy, because he/she



Pronoun Resolution

The girl knows the boy, because he/she

$C(\text{he}) - C(\text{she})$

- Female preference
+ Male preference

Pronoun Resolution

The girl knows the boy, because he/she

C(he) - C(she)

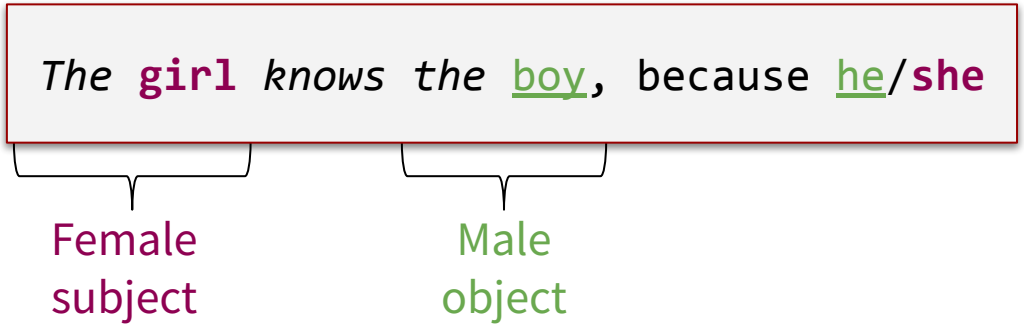
-0.12

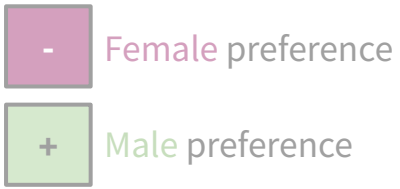
0.04

\- Female preference

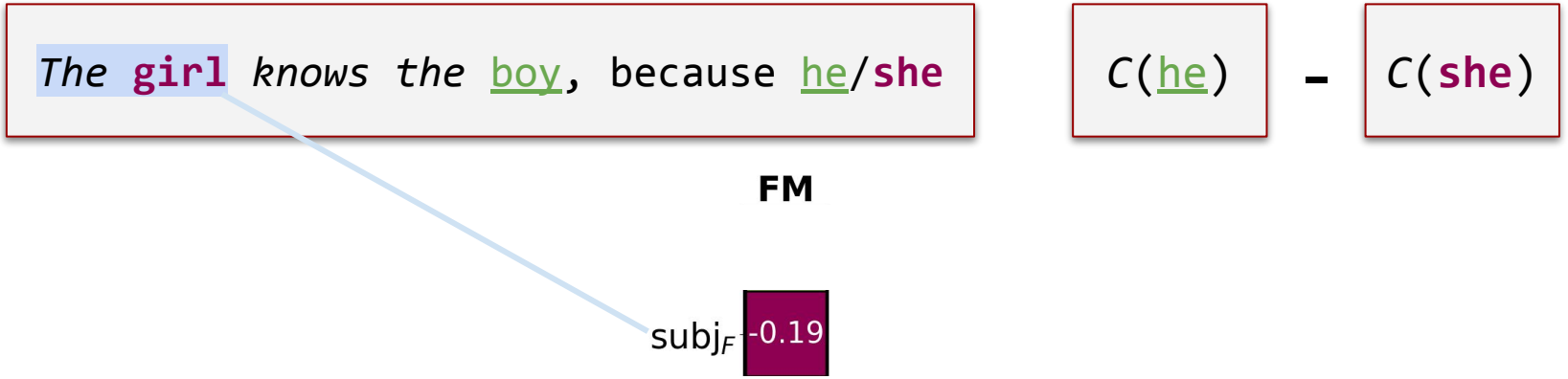
\+ Male preference

Average contributions

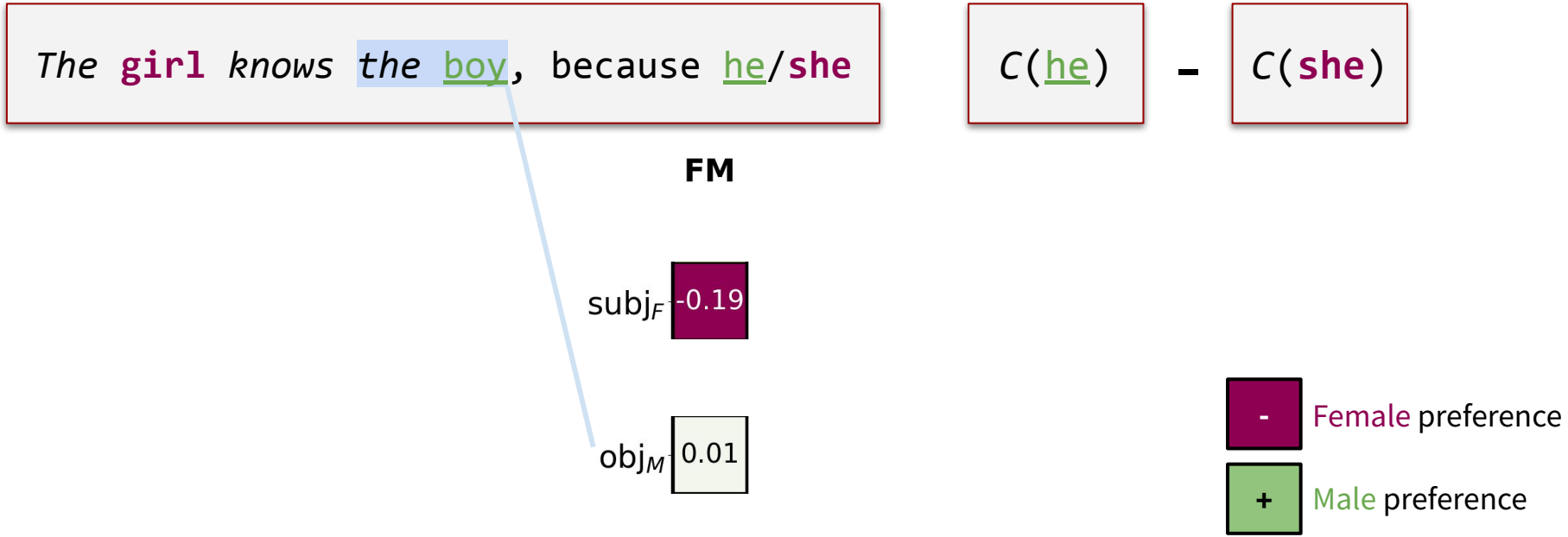




Average contributions



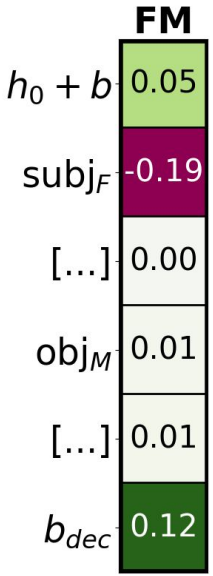
Average contributions



Average contributions

The **girl** *knows the* boy, because he/**she**

C(he) - C(**she**)



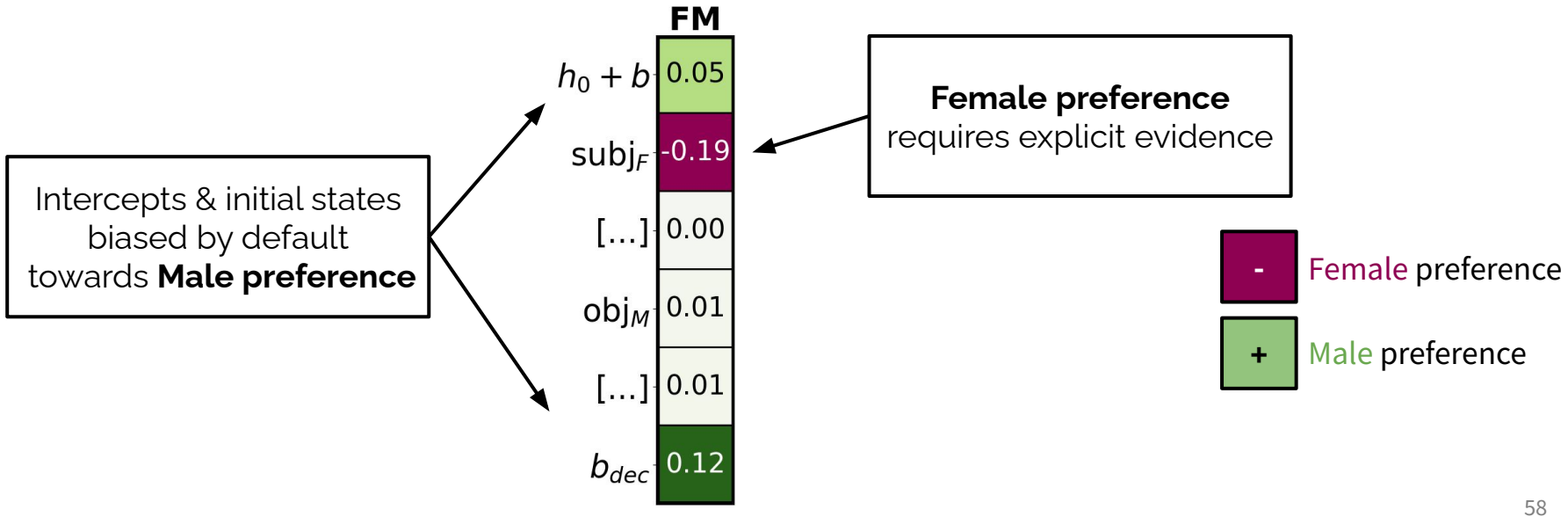
- Female preference

+ Male preference

Default Reasoning?

The **girl** *knows the* boy, because he/**she**

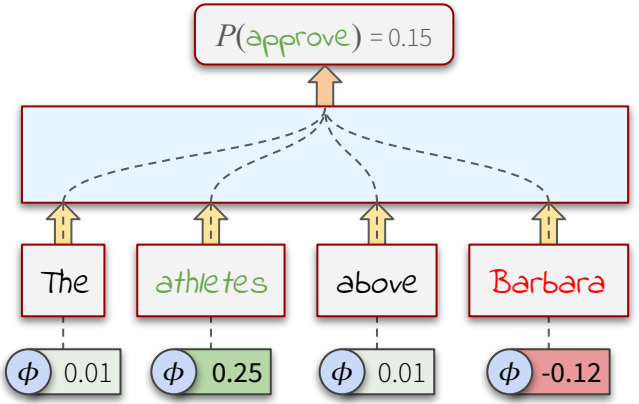
C(he) - C(**she**)



Feature Attribution Methods

How do we compute the relative importance of a feature?

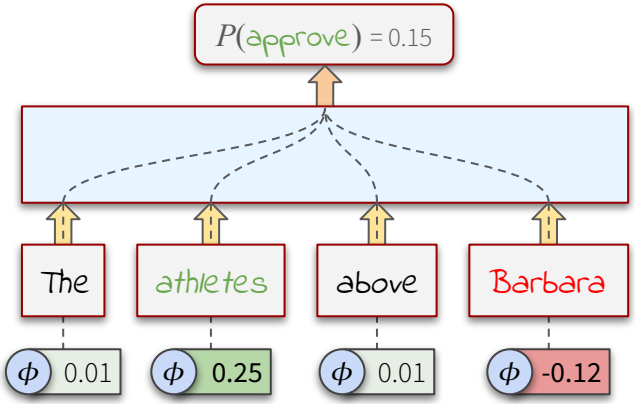
- Often this is done by **perturbing** parts of the input, and measuring the *change* in model output.



Feature Attribution Methods

How do we compute the relative importance of a feature?

- Often this is done by **perturbing** parts of the input, and measuring the *change* in model output.
- How should we perturb?
- How can we represent the *missingness* of a feature?
- How should we measure the change?



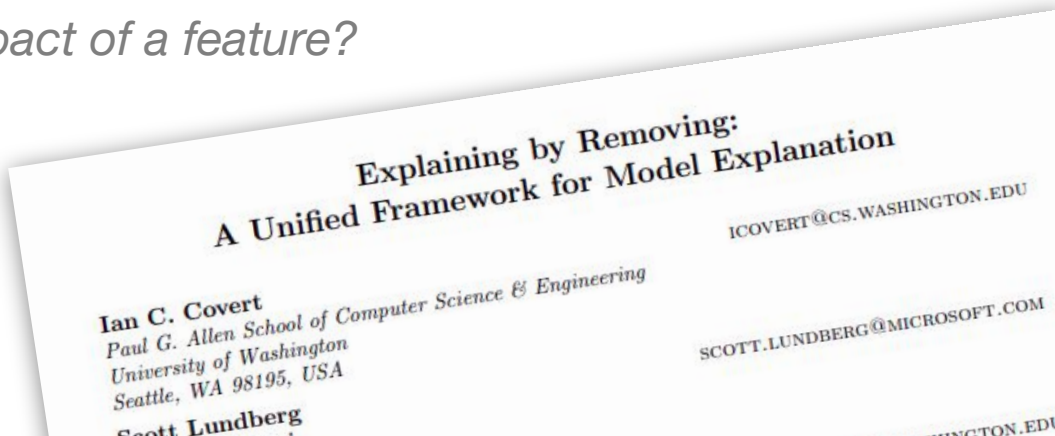
Attribution Dimensions

1. **Feature Removal**

   How do we deal with removed features?

2. **Feature Influence**

   How do we quantify the impact of a feature?



Explaining by Removing:
A Unified Framework for Model Explanation

Ian C. Covert
Paul G. Allen School of Computer Science & Engineering
University of Washington
Seattle, WA 98195, USA
ICOVERT@CS.WASHINGTON.EDU

SCOTT.LUNDBERG@MICROSOFT.COM

Feature Removal

1 **Static** Baseline

$$v(\mathbf{x}_S) = f(\mathbf{x}_S \cup \mathbf{x}'_{\setminus S})$$

Feature Removal

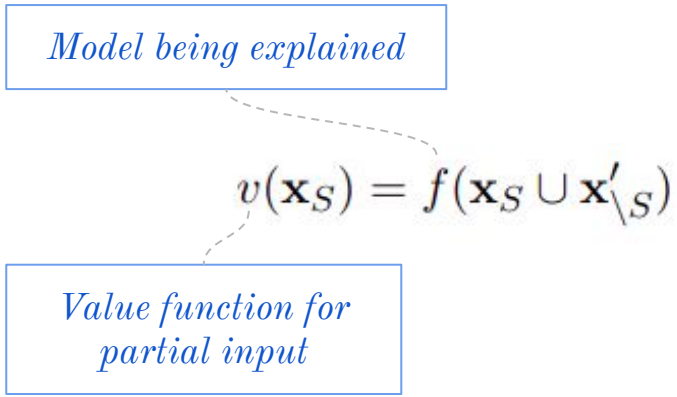
1 **Static** Baseline

$$v(\mathbf{x}_S) = f(\mathbf{x}_S \cup \mathbf{x}'_{\setminus S})$$

Value function for partial input

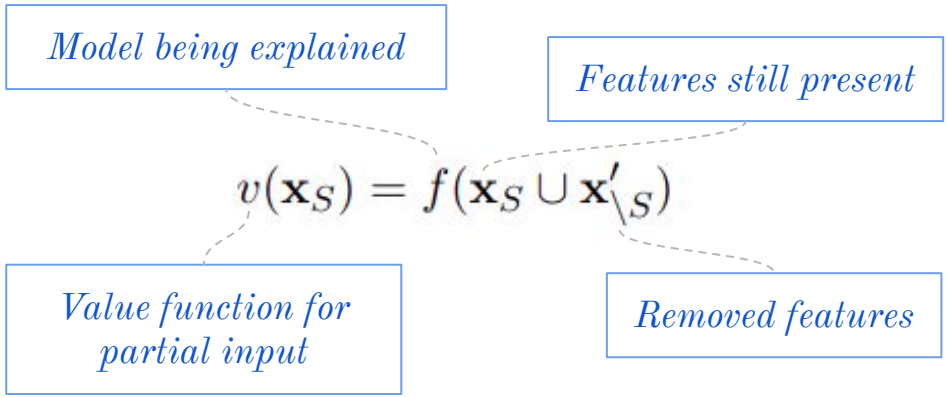
Feature Removal

1 **Static** Baseline



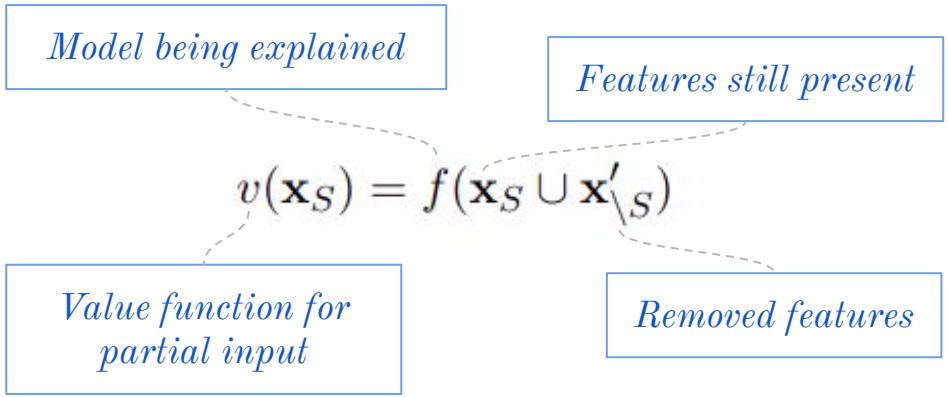
Feature Removal

1 **Static** Baseline



Feature Removal

1 **Static** Baseline



x	= “This movie is not bad”
x’	= “<pad> <pad> <pad> <pad> <pad>”
S	= {1, 2, 3, 5}
$x_S \cup x'_{\setminus s}$	= “This movie is <pad> bad”

Feature Removal

2 **Interventional** background distribution

$$v(\mathbf{x}_S) = \mathbb{E}_{\mathbf{x}'_{\setminus S}}\left[f(\mathbf{x}_S \cup \mathbf{x}'_{\setminus S})\right]$$

Expectation over removed features

Feature Removal

2 **Interventional** background distribution

$$v(\mathbf{x}_S) = \mathbb{E}_{\mathbf{x}'_{\setminus S}}\left[f(\mathbf{x}_S \cup \mathbf{x}'_{\setminus S})\right]$$

Expectation over removed features

x = “This movie is not bad”

S = {1, 2, 3, 5}

$x_S \cup x'_{\setminus s}$ = “This movie is *the* *is* *walk* ... bad”

Feature Removal

3 **Observational** background distribution

Conditioned on present features

$$v(\mathbf{x}_S) = \mathbb{E}_{\mathbf{x}'_{\setminus S}}\left[f(\mathbf{x}_S \cup \mathbf{x}'_{\setminus S}) \mid \mathbf{x}_S\right]$$

Expectation over removed features

Feature Removal

3 **Observational** background distribution

Conditioned on present features

$$v(\mathbf{x}_S) = \mathbb{E}_{\mathbf{x}'_{\setminus S}}\left[f(\mathbf{x}_S \cup \mathbf{x}'_{\setminus S}) \mid \mathbf{x}_S\right]$$

Expectation over removed features

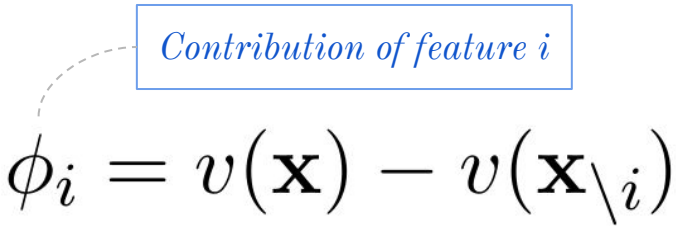
x = “This movie is not bad”

S = {1, 2, 3, 5}

$x_S \cup x'_{\setminus s}$ = “This movie is *very* / *quite* / *pretty* / ... bad”

Feature Influence

1 **Ablation**



x = “*This movie is not bad*”

x’ = <pad>

ϕ_{not} = f(“*This movie is not bad*”) - f(“*This movie is* <pad> *bad*”)

Feature Influence

2 Shapley Value

$$\phi_i = \overbrace{\sum_{S \subseteq \mathbf{x} \setminus \{i\}}}^{\text{coalitions without } i} p(S) \cdot \overbrace{\left(v(\mathbf{x}_{S \cup i}) - v(\mathbf{x}_S)\right)}^{\text{marginal contribution of } i \text{ to coalition}}$$

$$p(S) = \underbrace{\frac{|S|!(|\mathbf{x}| - 1 - |S|)!}{|\mathbf{x}|!}}_{\text{relative number of coalitions of size } |S|}$$

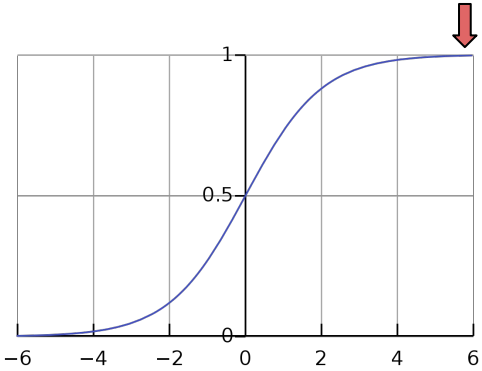
Completeness *axiom*:

$$f(x) = \sum_i \phi_i$$

Feature Influence

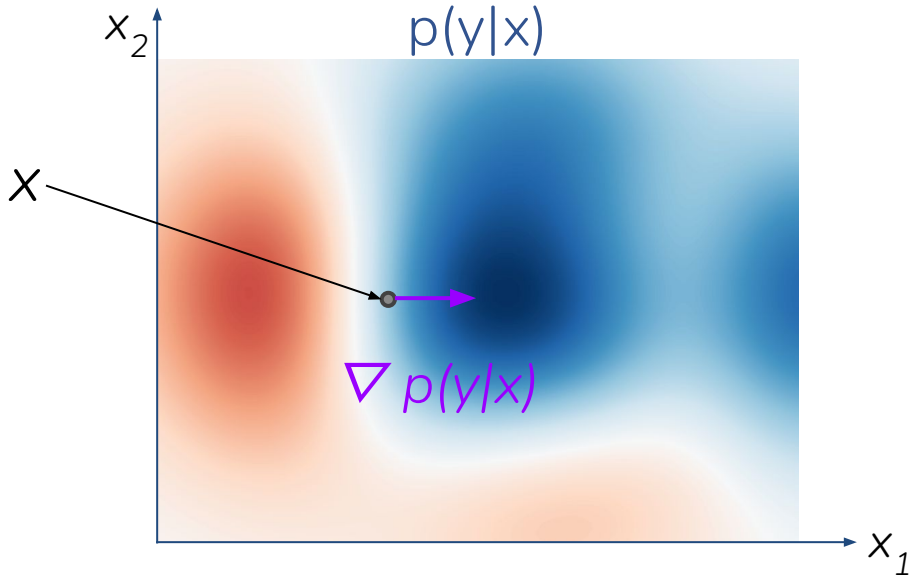
3 **Gradients**

Plain Gradients: $\phi_i = \frac{f(x)}{\partial x_i}$



Highlighting via Input Gradients

- Estimate importance of a feature using derivative of output w.r.t that feature
- i.e., with a “tiny change” to the feature, what happens to the prediction?

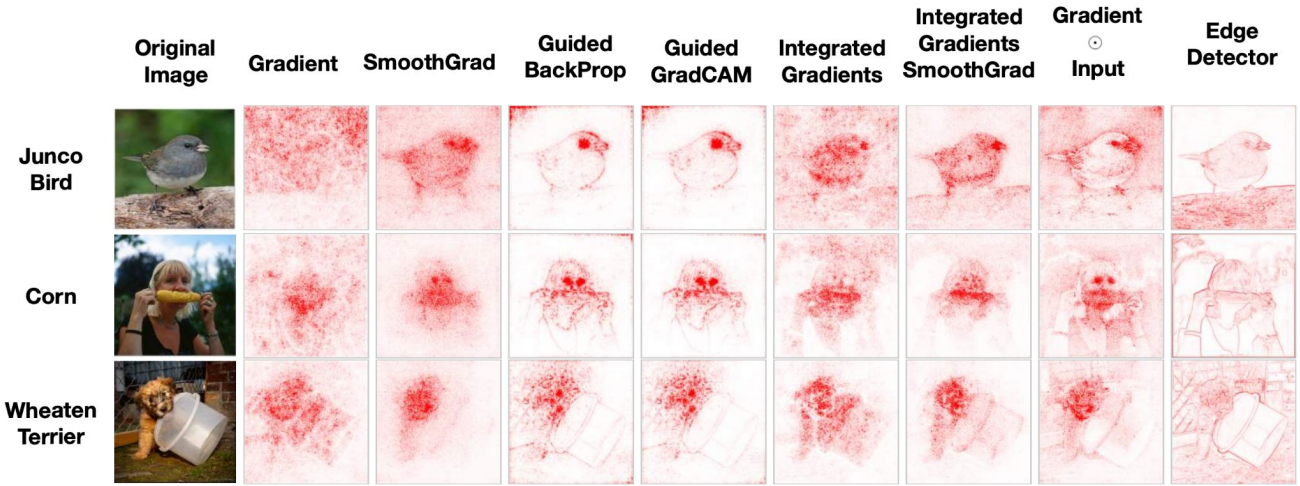


- We then visualize the importance values of each feature in a heatmap

[Simonyan et al. 2014]

Slide credit: Ana Marasovic

Example of highlighting: Image classification



[Adebayo et al., 2018]

Slide credit: Ana Marasovic

Gradient-based Highlightings for NLP

For NLP, derivative of output w.r.t a feature

=

derivative of **output** w.r.t an **input token**

What to use as the output?

- Top prediction probability
- Top prediction logits
- Loss (with the top prediction as the ground-truth class)

Token is actually an embedding. How to turn gradient w.r.t embedding into a scalar score?

- Sum it?
- Take an L_p norm?
- Dot product with embedding itself?

Do we normalize values across sentence?

direction lead to a decrease in the loss

$$-\nabla_{e(t)}\mathcal{L}_{\hat{y}} \cdot e(t)$$

L1-normalized across all tokens

Eqn from [Han et al. 2020]

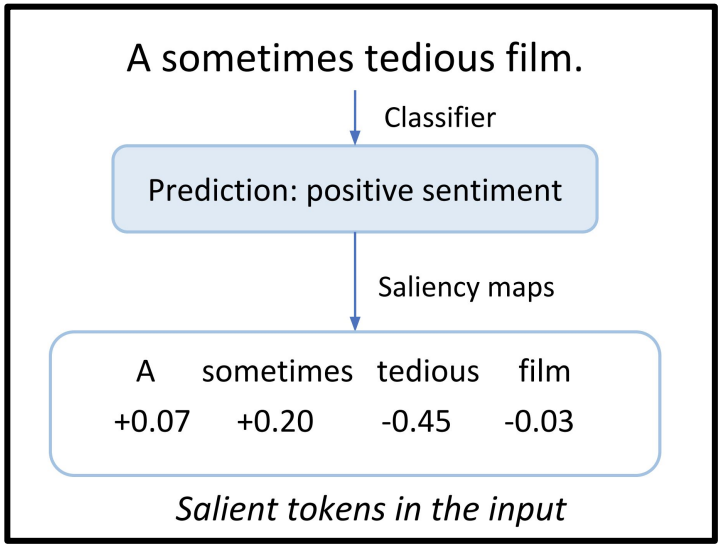
Slide credit: Ana Marasovic

Gradient-based Highlightings for NLP

For NLP, derivative of output w.r.t a feature

=

derivative of **output** w.r.t an **input token**



direction lead to a decrease in the loss

$$-\nabla_{e(t)}\mathcal{L}_{\hat{y}} \cdot e(t)$$

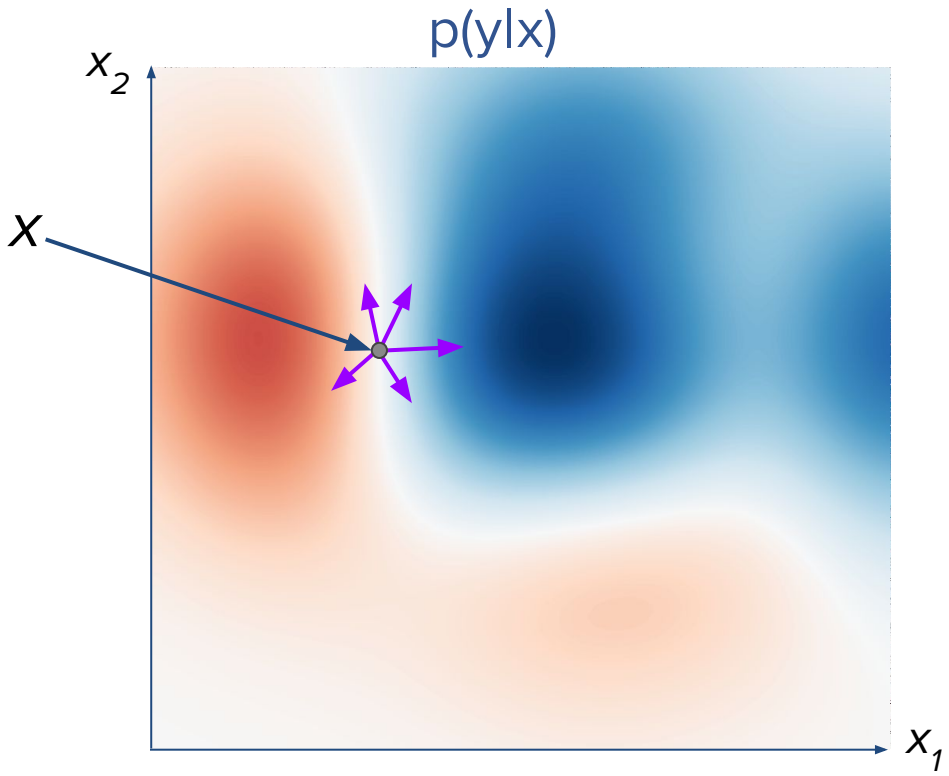
L1-normalized across all tokens

Eqn from [Han et al. 2020]

Slide credit: Ana Marasovic

Problems with Using Gradient for Highlighting

- Too “local” and thus sensitive to slight perturbations

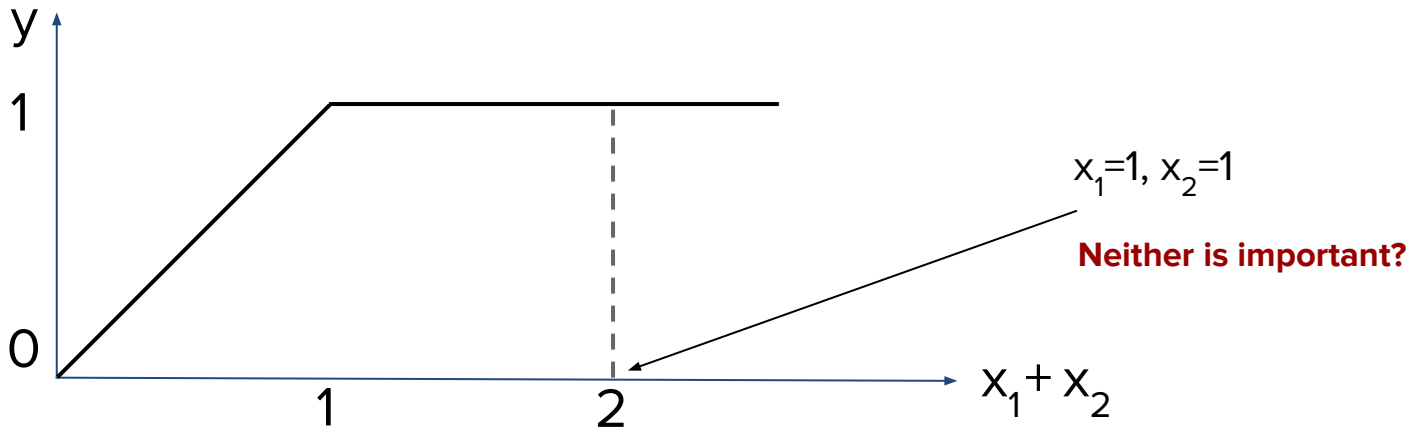


Slide credit: Ana Marasovic

Problems with Using Gradient for Highlighting

- too "local" and thus sensitive to slight perturbations
- "saturated outputs" lead to unintuitive gradients

$$y = x_1 + x_2 \quad \text{when } (x_1 + x_2) < 1$$
$$1 \quad \text{when } (x_1 + x_2) >= 1$$

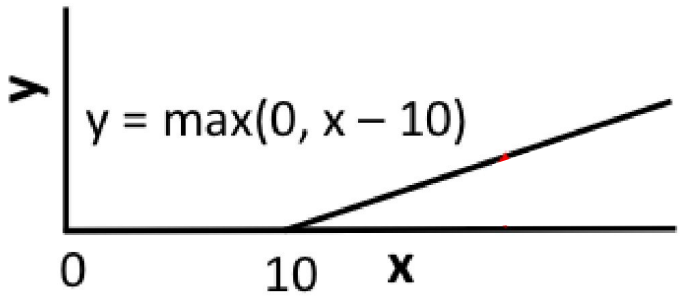


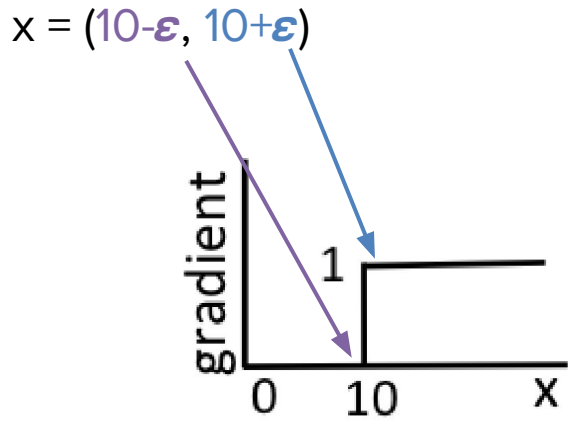
[Shrikumar et al. 2017]

Slide credit: Ana Marasovic

Problems with Using Gradient for Highlighting

- too "local" and thus sensitive to slight perturbations
- "saturated outputs" lead to unintuitive gradients
- discontinuous gradients (e.g., thresholding) are problematic





[Shrikumar et al. 2017]

Slide credit: Ana Marasovic

Extensions of Vanilla Gradient

- too “local” and thus sensitive to slight perturbations
- “saturated outputs” lead to unintuitive gradients
- discontinuous gradients (e.g., thresholding) are problematic

How to mitigate these issues? Don’t rely on a single gradient calculation:

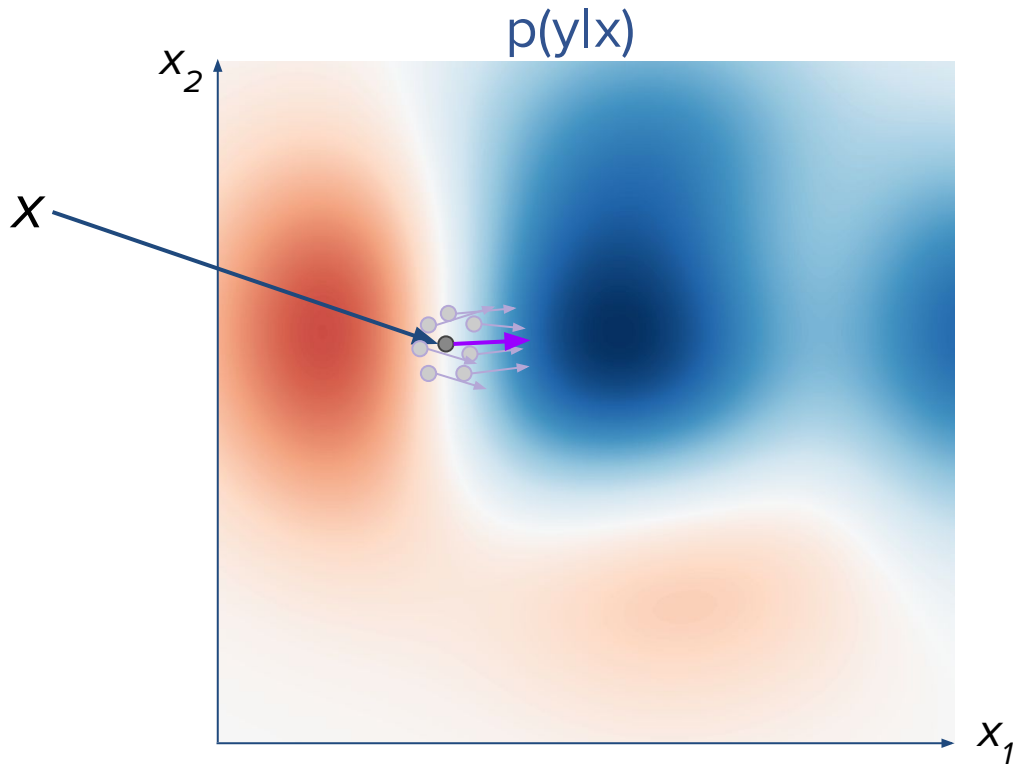
- SmoothGrad
- Integrated Gradients

Other approaches, e.g., LRP, DeepLIFT, GradCAM. Not covered here.

Slide credit: Ana Marasovic

Extensions of Vanilla Gradient

SmoothGrad: add gaussian noise to input and average the gradient

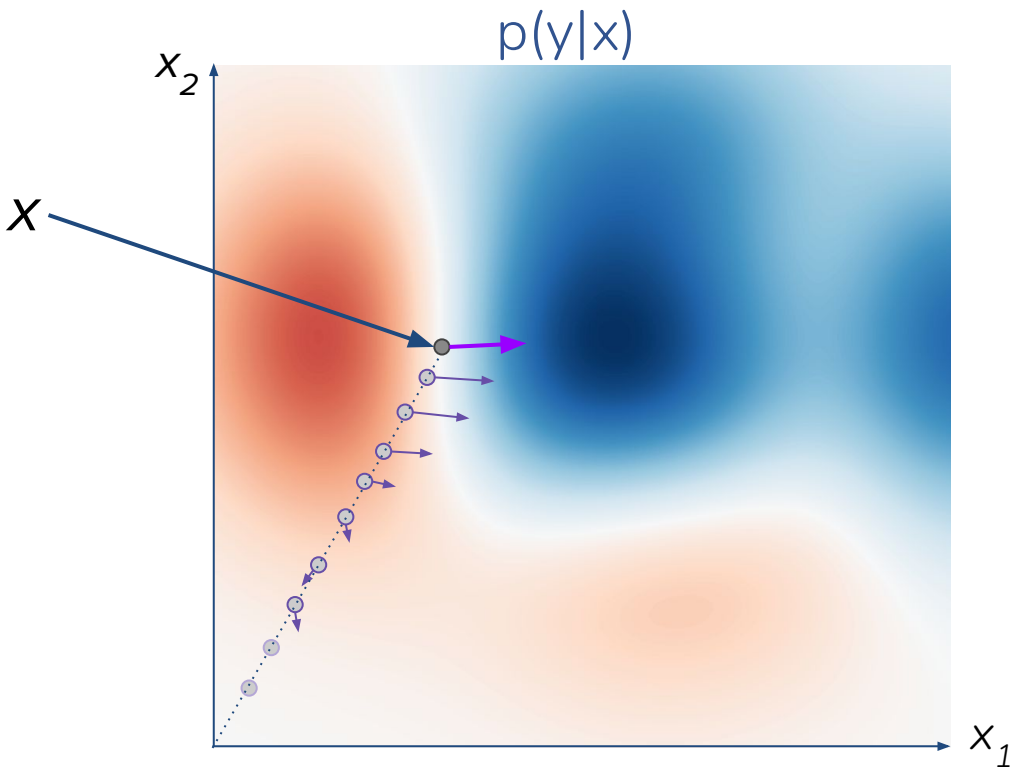


[Smilkov et al. 2017]

Slide credit: Ana Marasovic

Extensions of Vanilla Gradient

Integrated Gradients: average gradients along path from zero to input



[Sundararajan et al. 2017]

Slide credit: Ana Marasovic

Summary of Gradient-based Highlighting

Positives:

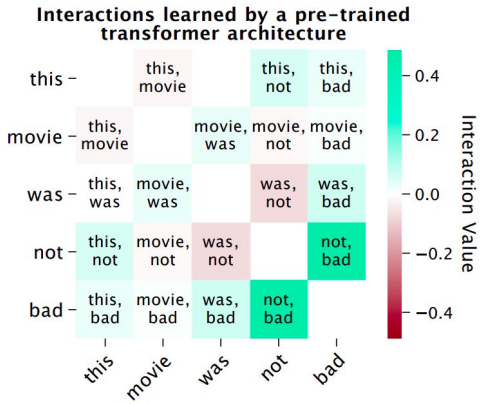
- Fast to compute: single (or a few) calls to backward()
- Visually appealing: spectrum of importance values

Negatives:

- Needs white-box (gradient) access to the model
- Gradients can be unintuitive with saturated or thresholded values
- Difficult to apply to non-classification tasks
- Highlighting cannot do anything if a model uses knowledge (such as commonse) that is not explicitly mentioned in the input
- Ignore the interactions between words/pixels (e.g., “not good”)

Slide credit: Ana Marasovic

Summary of Gradient-based Highlighting



- Ignore the interactions between words/pixels (e.g., “not good”)

Slide credit: Ana Marasovic

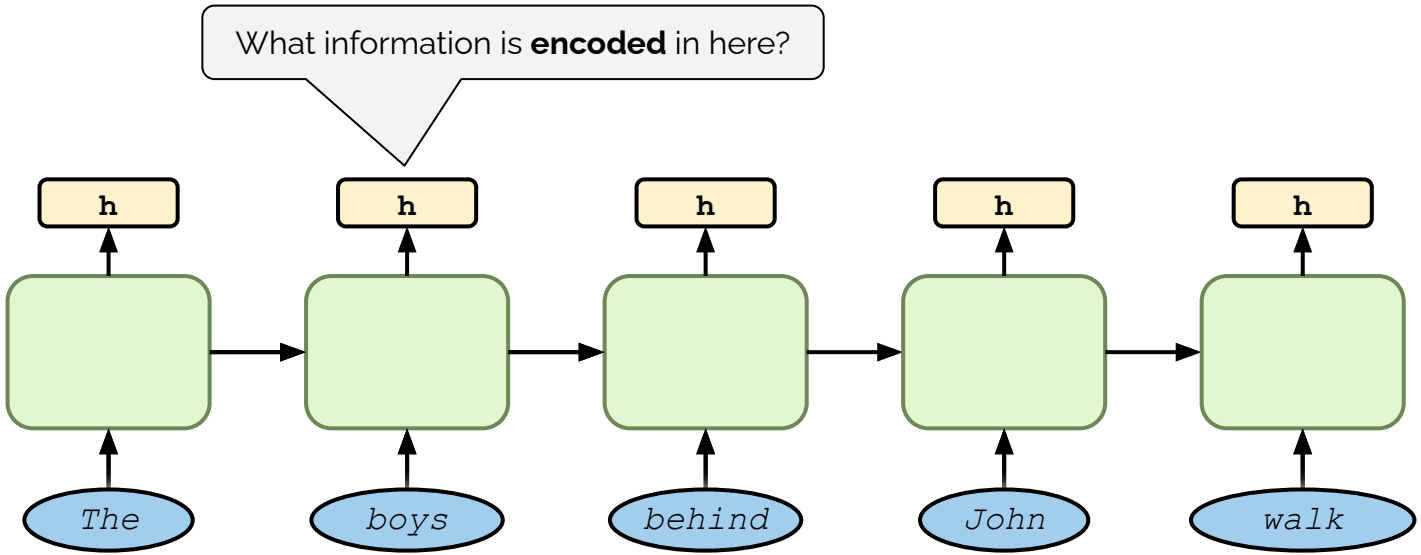
Probing

Feature attribution methods showed us which input features were important for a prediction.

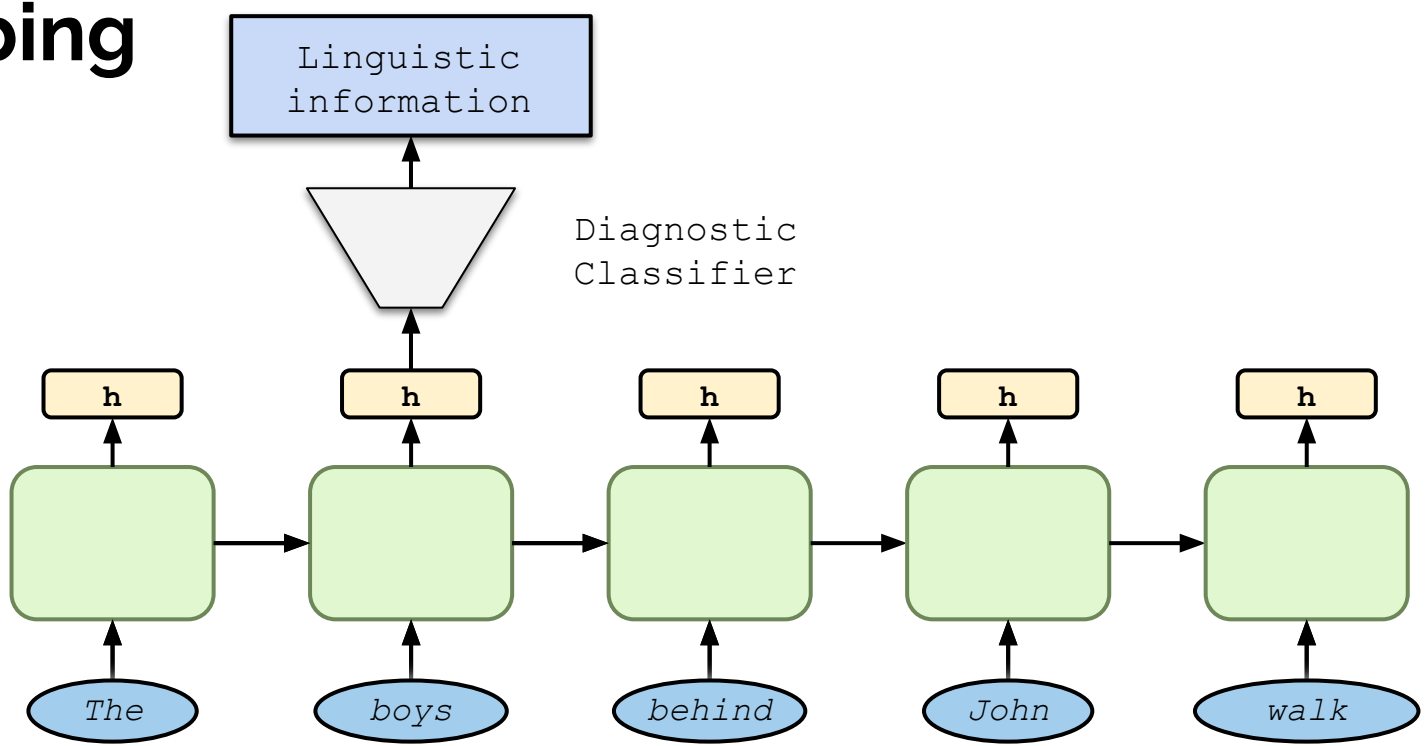
- ✗ They do not show *where* in the model predictions are formed
- ✗ They give no insight into **higher-level** concepts such as ‘gender’, ‘number’, or ‘part-of-speech’ class.

Instead, we can turn to **probing**, in which we train classifiers on top of model representations!

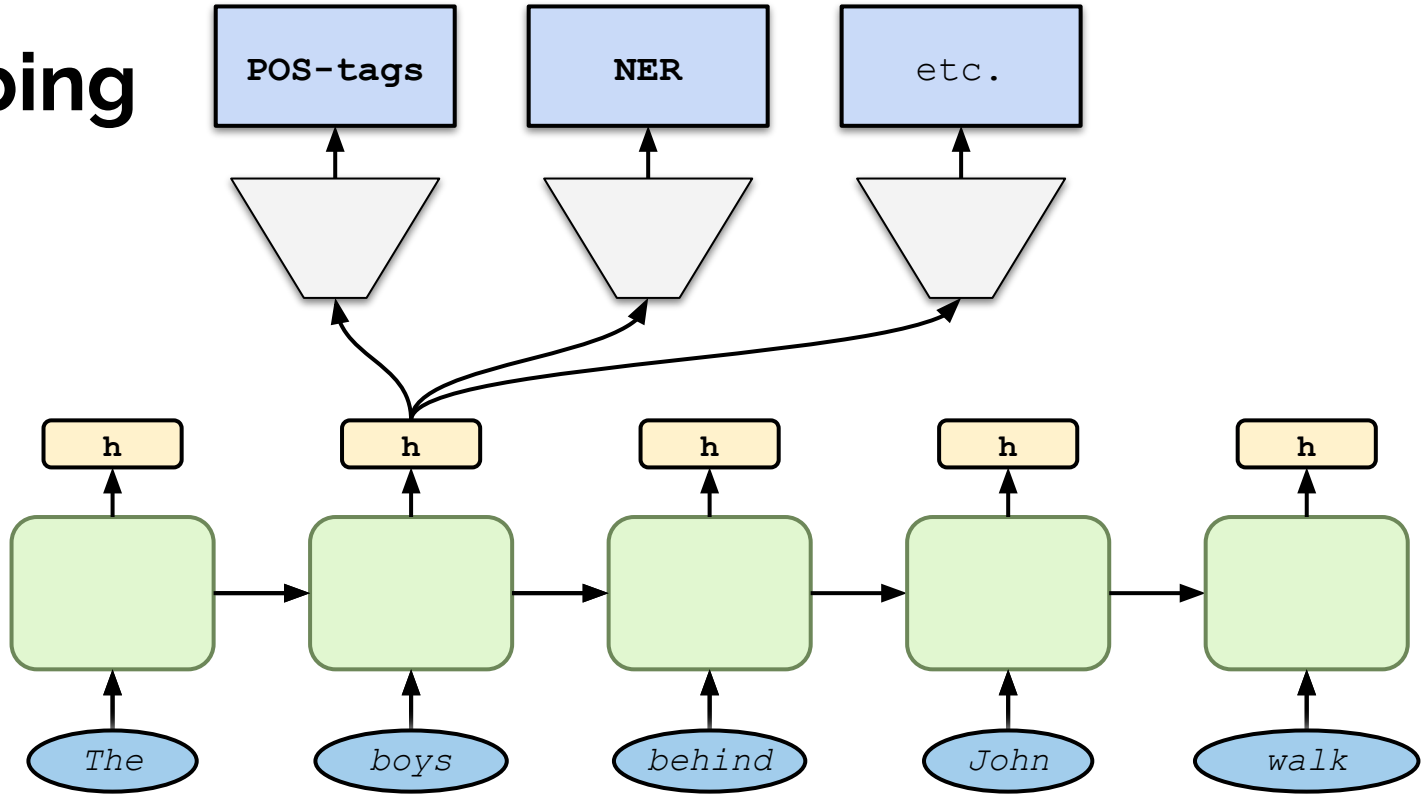
Probing


What information is **encoded** in here?
h
h
h
h
h
The
boys
behind
John
walk

Probing


Linguistic information
Diagnostic Classifier
h
h
h
h
h
The
boys
behind
John
walk

Probing



Representations

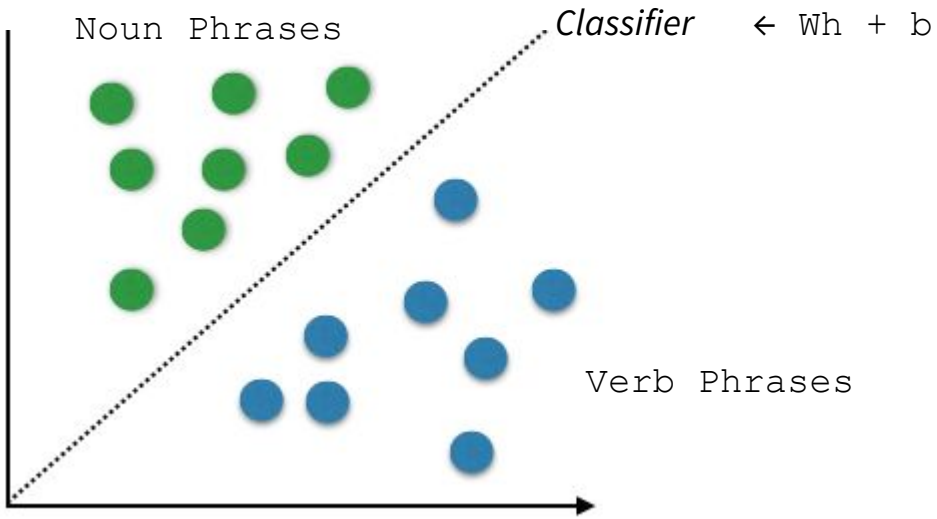
- Representations are just a point in a vector space
- But, it is likely that the representation of “cat” is somewhat similar to “dog”

$$h_{cat} \approx h_{dog}$$

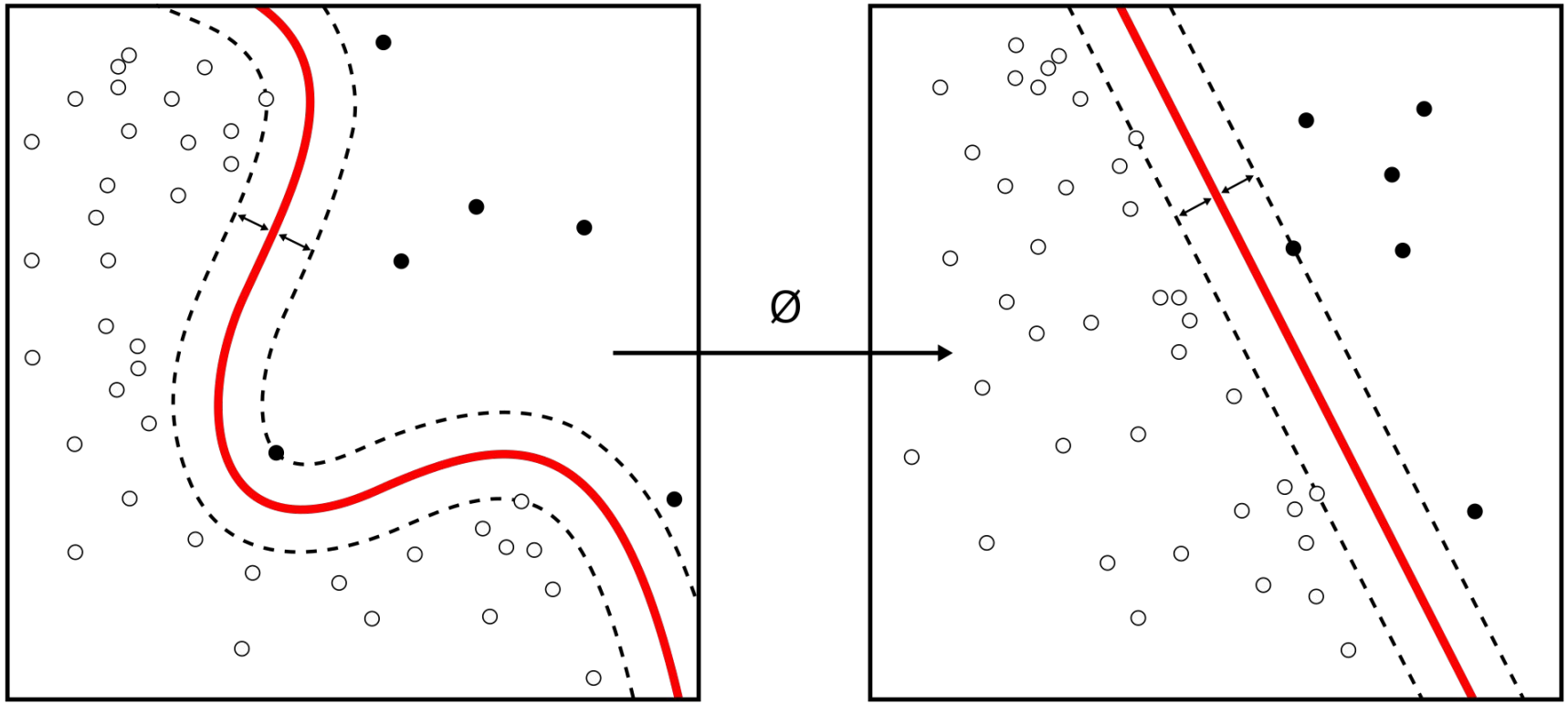
- More generally, the representation of **nouns** are likely to be similar, and <u>distinct</u> from **verbs**, **determiners**, **adverbs**, etc.

$$h_{NOUN} \not\approx h_{VERB}$$

What does *probed info* imply?



Why linear?



Probing

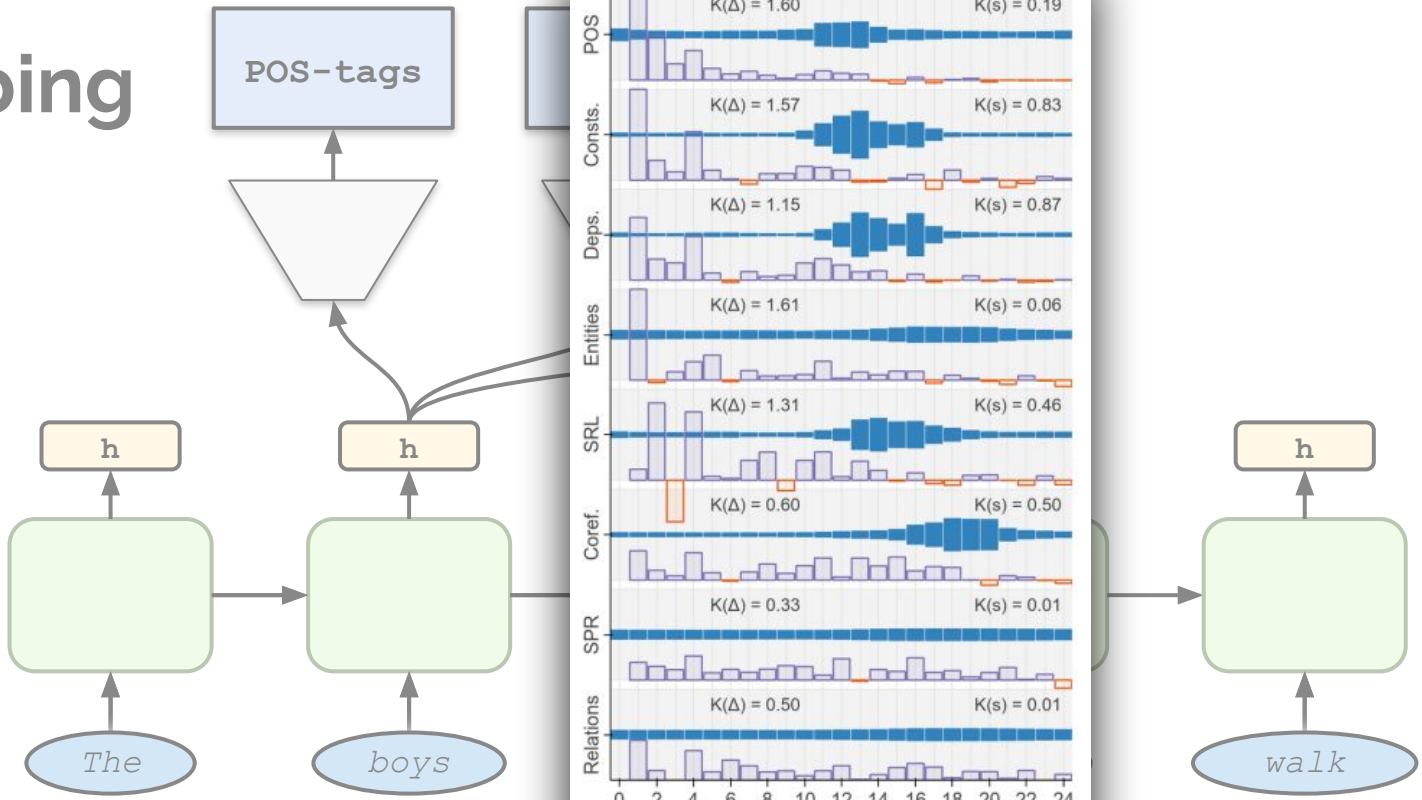
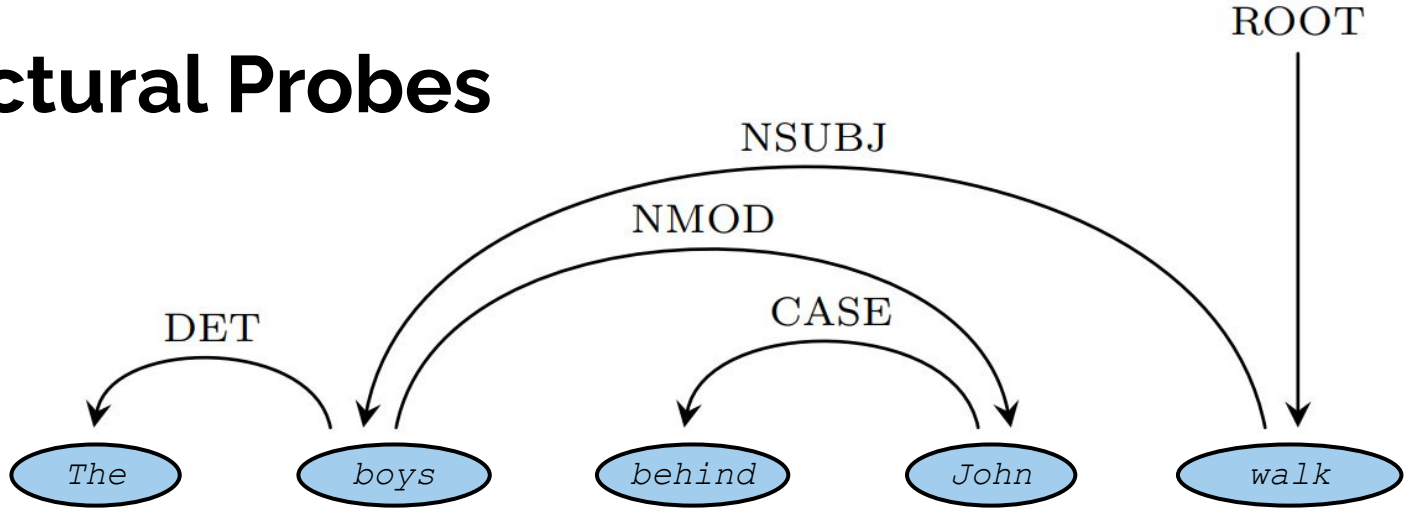


Figure 2: Layer-wise metrics on BERT-large. Solid (blue) are mixing weights $s_\tau^{(\ell)}$ (§3.1); outlined (purple) are differential scores $\Delta_\tau^{(\ell)}$ (§3.2), normalized for each task. Horizontal axis is encoder layer.

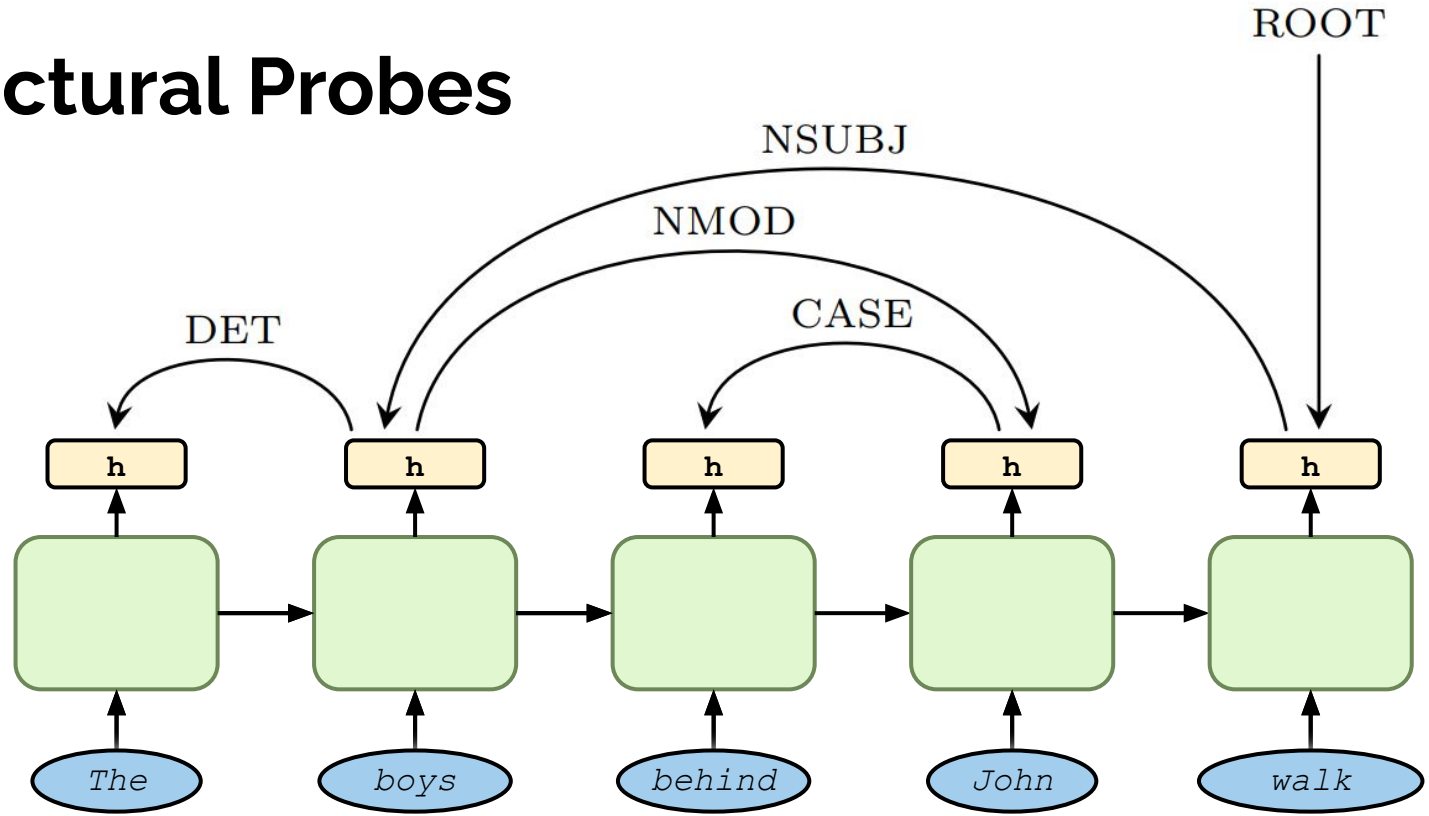
BERT Rediscovers the Classical NLP Pipeline

Tenney et al. (2019)

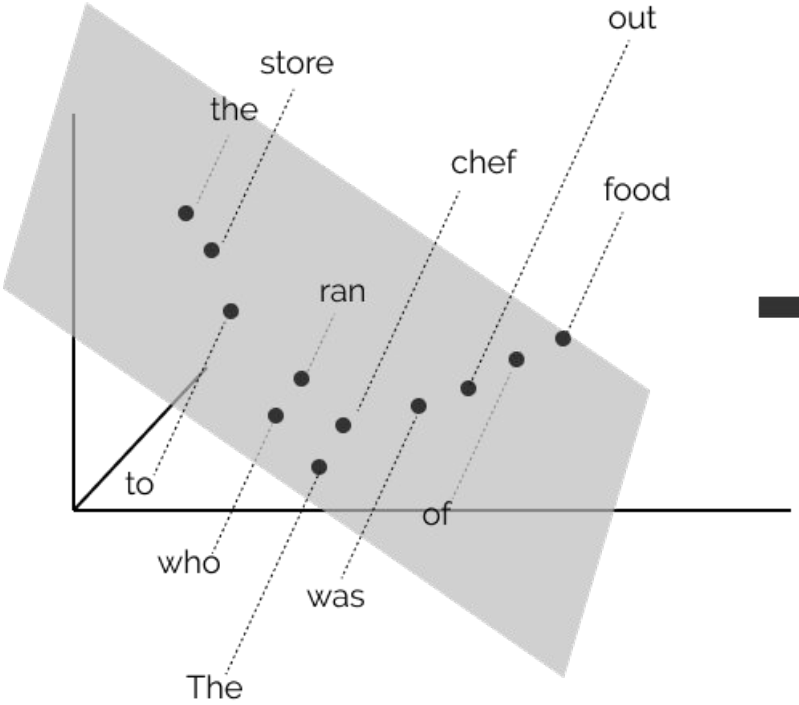
Structural Probes

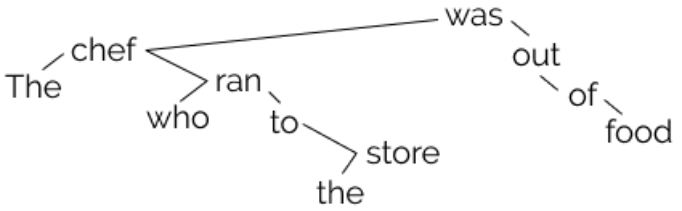


Structural Probes



Structural Probes





A Structural Probe for Finding Syntax in Word Representations

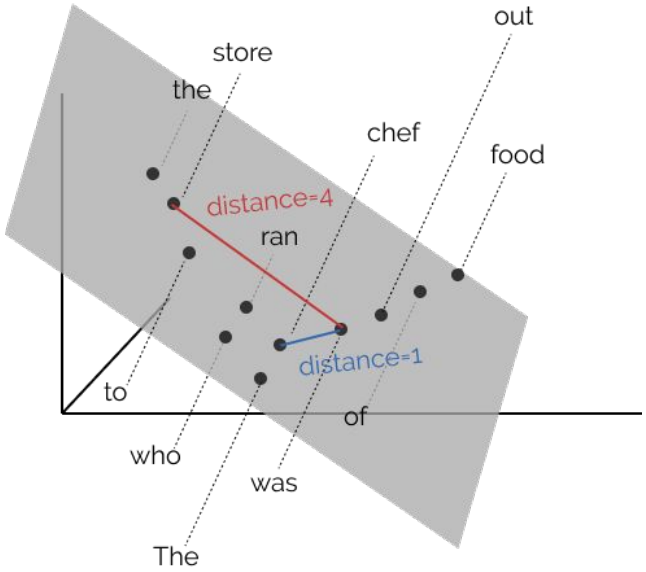
John Hewitt
Stanford University
johnhew@stanford.edu

Christopher D. Manning
Stanford University
manning@stanford.edu

Distance Metric

$d(x, y) = \| x - y \|$ *Norm induced metric*

$d(x, y) = \| x - y \|_B$ *We calculate the distance in the subspace transformed by the linear map B*



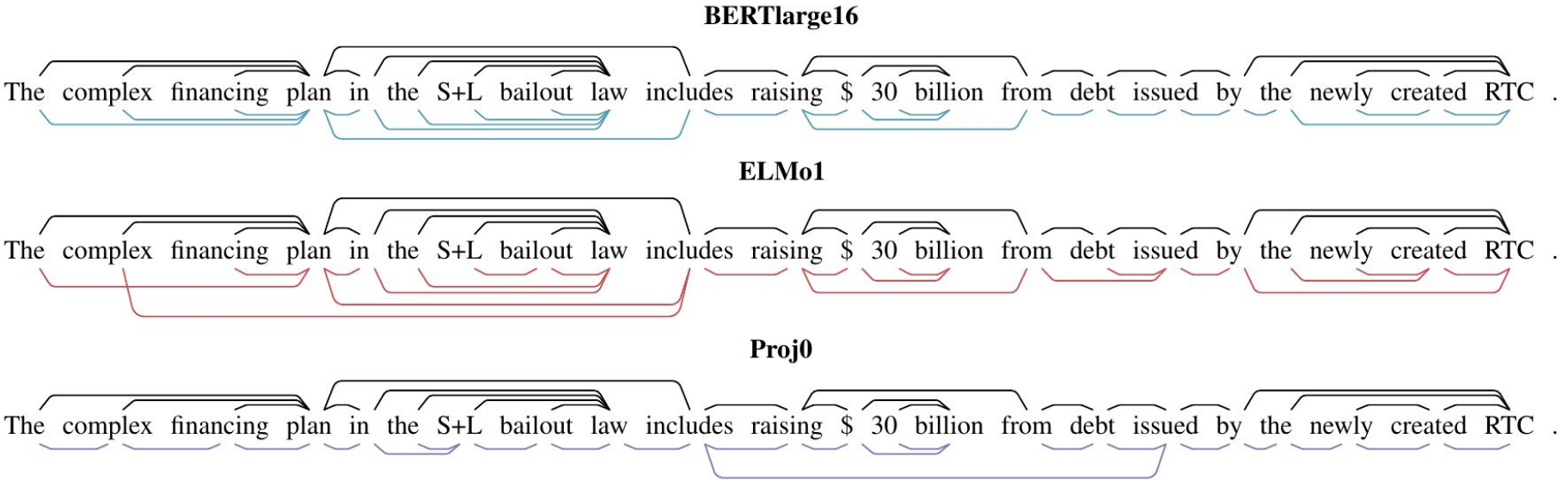
Objective

True distance (*gold parse*)
We recover this from our treebank

$$\min_{B} \sum_{\ell} \frac{1}{|s_\ell|^2} \sum_{i,j} \left(d(w_i, w_j) - \|B(h_i - h_j)\|^2 \right)$$

Predicted distance

Parse trees


BERTlarge16
The complex financing plan in the S+L bailout law includes raising $ 30 billion from debt issued by the newly created RTC .
ELMo1
The complex financing plan in the S+L bailout law includes raising $ 30 billion from debt issued by the newly created RTC .
Proj0
The complex financing plan in the S+L bailout law includes raising $ 30 billion from debt issued by the newly created RTC .

Recap

- The huge size of current NLP models has made us lose **transparency**
- Interpretability is **vital** for gaining trust in black-box models
- Interpretability is also vital for understanding the **linguistic capacities** of NLP models
- We can explain a model at increasing levels of granularity
  - Behavioural tests
  - Feature attributions
  - Probing
  - *(Not covered today)* Mechanistic Interpretability
    - ➢ *Check out Interpretability & Explainability in AI, Block 6!*
- Thanks for listening!

References

1. Breiman (2001) - *Statistical Modeling: The Two Cultures*
2. Lipton (2018) - *The Mythos of Model Interpretability*
3. Yin & Neubig (2022) - *Interpreting Language Models with Contrastive Explanations*
4. McCoy et al. (2021) - *Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference*
5. Warstadt et al. (2020) - *BLiMP: The Benchmark of Linguistic Minimal Pairs for English*
6. Jumelet et al. (2019) - *Analysing Neural Language Models: Contextual Decomposition Reveals Default Reasoning in Number and Gender Assignment*
7. Covert et al. (2021) - *Explaining by removing: a unified framework for model explanation*
8. Tenney et al. (2019) - *BERT Rediscovers the Classical NLP Pipeline*
9. Hewitt et al. (2019) - *A Structural Probe for Finding Syntax in Word Representations*