



NLP1

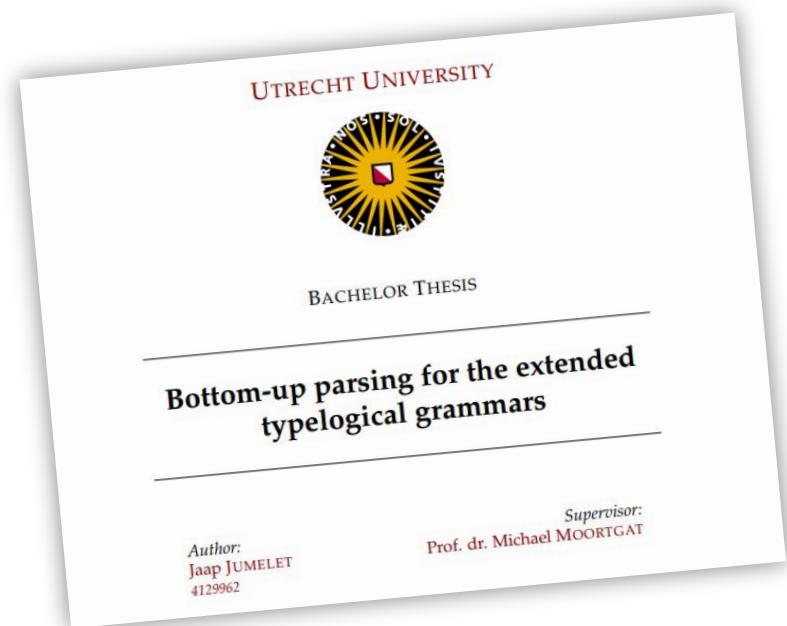
Evaluating and Interpreting Language Models

Jaap Jumelet | ILLC, University of Amsterdam



Who Am I?

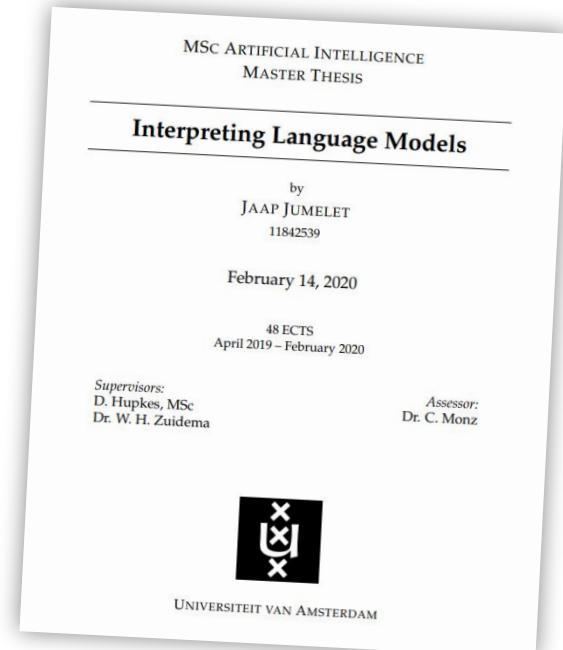
- BSc Artificial Intelligence at *Universiteit Utrecht* (2013-2017)
 - *Logic*
 - *(Computational) Linguistics*
 - *Theoretical Computer Science*





Who Am I?

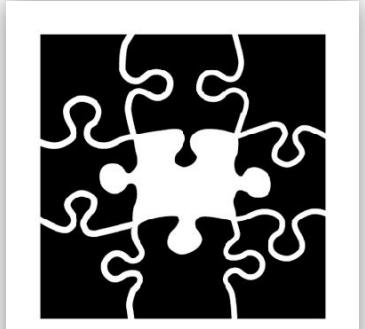
- BSc Artificial Intelligence at *Universiteit Utrecht* (2013-2017)
- MSc Artificial Intelligence at the *UvA* (2017 - 2020)
 - *Natural Language Processing*
 - *Machine/Deep Learning*
 - *Explainable AI*





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

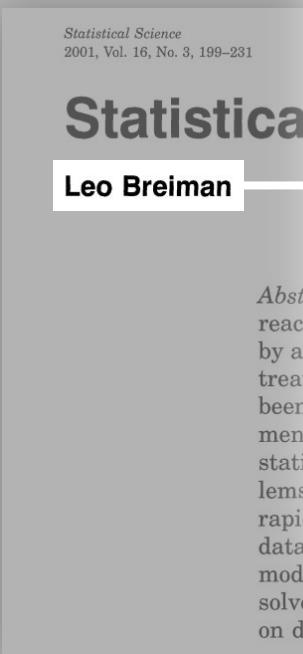
Leo Breiman

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



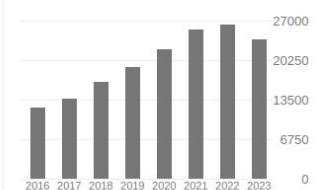
Leo Breiman 1928-2005

Professor of Statistics, [UC Berkeley](#)
Verified email at stat.berkeley.edu - Homepage

Data Analysis Statistics Machine Learning



Cited by	VIEW ALL	
	All	Since 2018
Citations	249340	133584
h-index	53	40
i10-Index	85	45



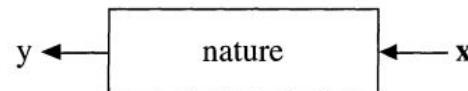
TITLE	CITED BY	YEAR
Random forests L Breiman Machine learning 45 (1), 5-32	119359	2001
Classification and Regression Trees L Breiman, JH Friedman, RA Olshen, CJ Stone CRC Press, New York	62587 *	1999
Classification and regression trees L Breiman Chapman & Hall/CRC	61904 *	1984
Bagging predictors L Breiman 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	5599	2001
Estimating optimal transformations for multiple regression and correlation L Breiman, JH Friedman Journal of the American Statistical Association, 580-598	2556	1985

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?

Statistics starts with data. Think of the data as being generated by a black box in which a vector of input variables \mathbf{x} (independent variables) go in one side, and on the other side the response variables \mathbf{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:

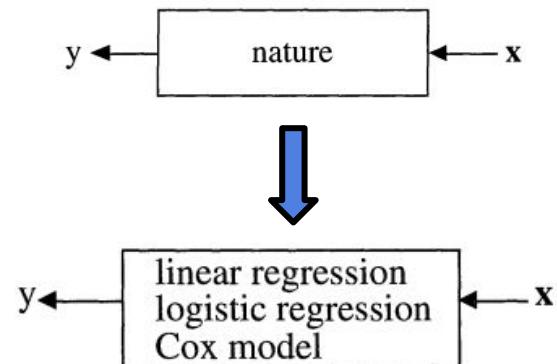
Prediction. To be able to predict what the responses are going to be to future input variables;

Information. To extract some information about how nature is associating the response variables to the input variables.



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.

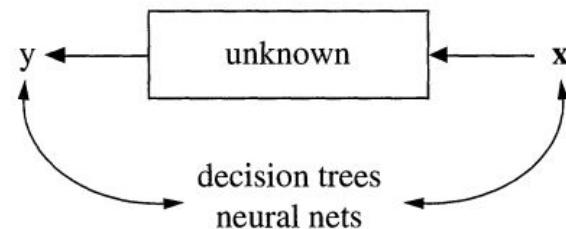
Estimated culture population. 98% of all statisticians.



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:

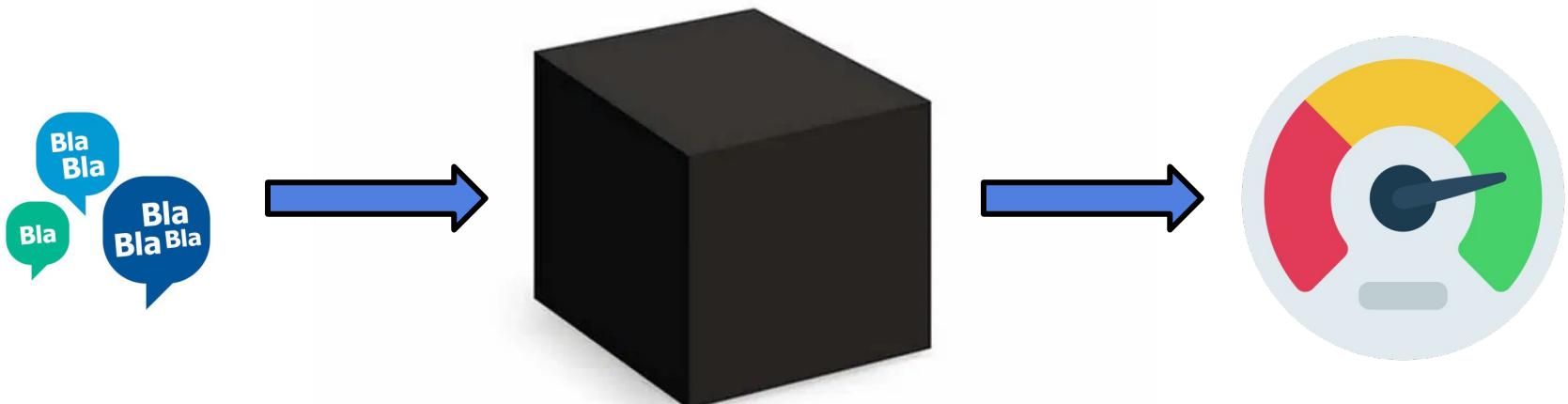


Model validation. Measured by predictive accuracy.

Estimated culture population. 2% of statisticians,
many in other fields.



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?*



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?*



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?*



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?*



Why do we need the desired outcome?

1. Fairness

- Wants to be fair

2. Trustworthiness

- Must be trustworthy

3. Robustness

- Doesn't break

4. Faithfulness

- Has the right answer

The screenshot shows a search result for 'Vervanging lichtmasten' (Replacing streetlights). The result is titled 'Risico-indicatoren & voorspelmodel jonge aanwas in de drugscriminaliteit' (Risk indicators & predictive model for young drug-related crime). It includes details such as 'Organisatie: Gemeente Utrecht', 'Type algoritme: Vervanging lichtmasten', 'Afdeling: Veiligheid', and 'Beleidsterrein: Veiligheid'. Below this, there is a summary of the algorithm: 'Het verkrijgt weke indicatoren jongeren levensstijl om beslissingen te nemen in de drugscriminaliteit en het voorstellen in welke buurten veel leefbare jongeren wonen. Deze toepassing vindt plaats binnen de City Deal "Zicht op Ondermijning". Meer specifieke informatie niet dit algoritme volgt.' The main title of the page is 'Het Algoritmeregister van de Nederlandse overheid'.

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'



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

groups?

ist them?

of Model Interpretability



Why do we need interpretability?

BBC NEWS

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

The desiderata of algorithmic interpretability

1. Fairness

- *What biases does our model have?*

2. Trustworthiness

- *Models that are deterministic*

3. Robustness

- *Does our model generalize well?*

4. Faithfulness

- *How faithful are models to the real world?*

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

st particular groups?

ty, can we trust them?

ing?

) - The Mythos of Model Interpretability



Why do we need algorithms?

The desiderata of algorithms

1. Fairness

- *What biases does it have?*

2. Trustworthiness

- *Models that are deployed*

3. Robustness

- *Does our model generalize well?*

4. Faithfulness

- *How faithful are models to reality?*

The BBC News mobile interface is shown, featuring the BBC logo, a user icon, a search bar, and a red 'NEWS' banner. Below the banner, the word 'Tech' is underlined, indicating the category of the news article. The main headline reads 'Facebook apology as AI labels black men 'primates''. The date '6 September 2021' is listed below the headline. A small red square icon with a white arrow is visible to the left of the headline.

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.

particular groups?

can we trust them?

g?

The Mythos of Model Interpretability



Why do we need interpretable models?

The desiderata of algorithms

1. Fairness

- *What biases does it contain?*

2. Trustworthiness

- *Models that are deployed*

3. Robustness

- *Does our model generalize well?*

4. Faithfulness

- *How faithful are models to reality?*

Tech

Twitter finds racial bias in image-cropping AI

© 20 May 2021

GETTY IMAGES

Preferences for white people over black people and women over men were found in testing

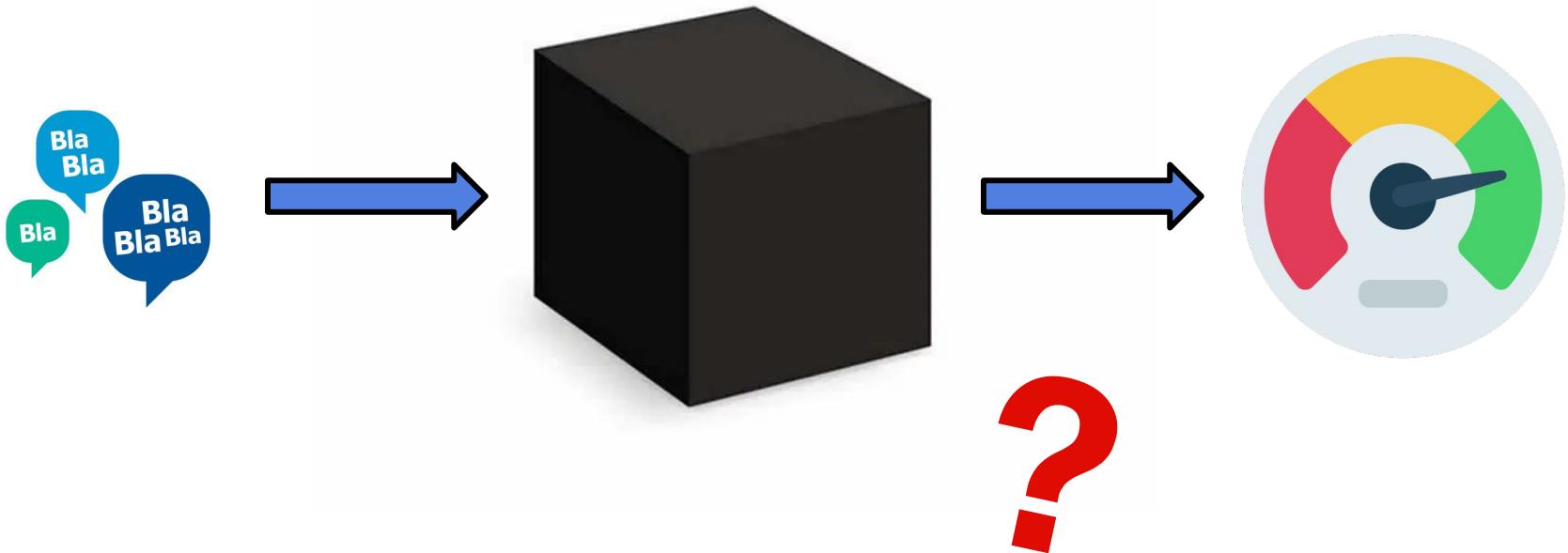
Twitter's automatic cropping of images had underlying issues that favoured white individuals over black people, and women over men, the company said.

It comes months after its users highlighted potential problems

he Mythos of Model Interpretability

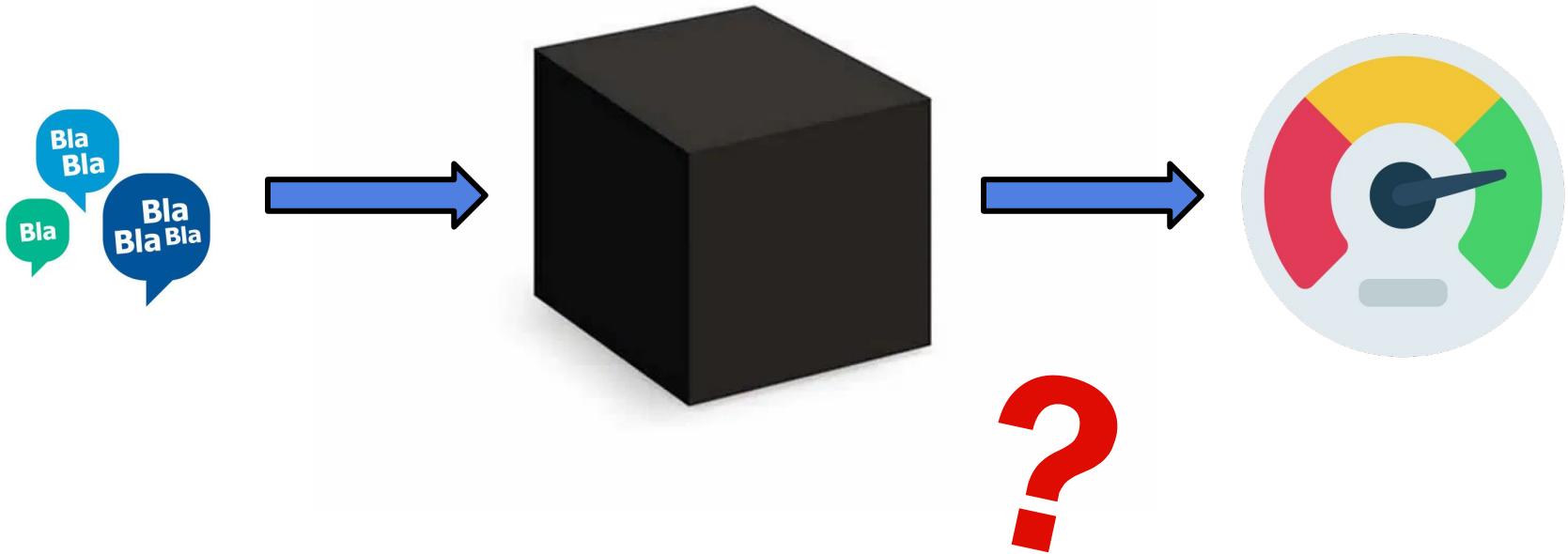


Why do we need interpretability?





Why do we need interpretability?



Personal reason: “scientific curiosity”, *how is it so good?*



Why



You

Can we ever truly understand a large-scale AI model's internal reasoning?



ChatGPT

Understanding the internal reasoning of a large-scale AI model, such as GPT-3.5, can be a challenging task. While we can gain insights into the functioning of AI models through various interpretability techniques, achieving a complete and comprehensive understanding of their internal reasoning remains a significant challenge.

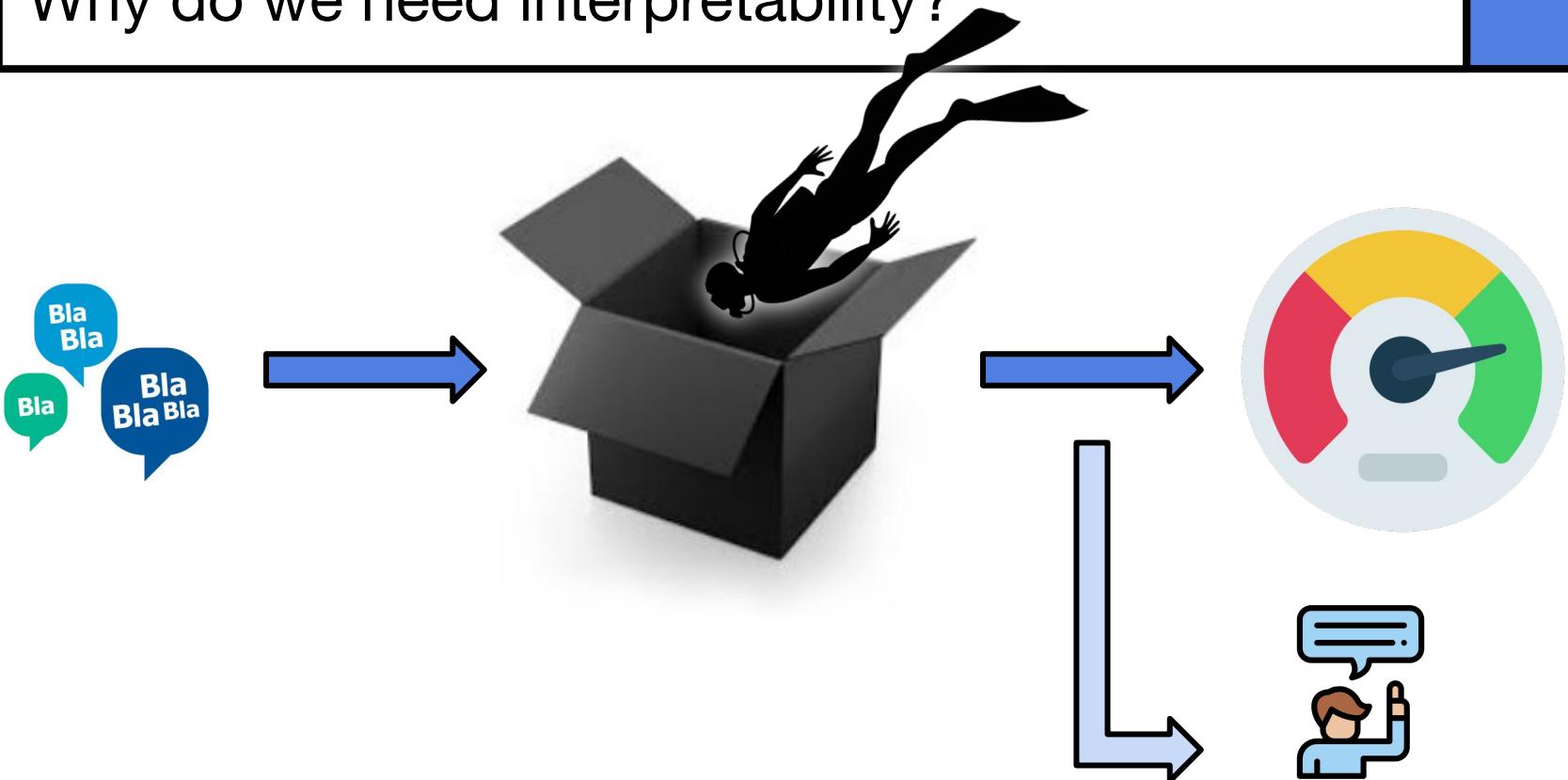
Large-scale AI models, like GPT-3.5, operate using complex architectures with millions or even billions of parameters. These models are trained on vast amounts of data, which allows them to learn patterns and make predictions. However, the internal workings of these models are highly intricate and often referred to as "black boxes" because it can be difficult to trace how a specific input leads to a particular output.

Several techniques have been developed to interpret and explain the decisions made by AI models. These include attention visualization, feature importance analysis, and saliency mapping, among others. These methods attempt to highlight the areas of input data that the model focuses on or deems important for making predictions. While these techniques can provide some insights, they may not capture the full extent of the model's reasoning.





Why do we need interpretability?



*Post-hoc
Explanation*



How do we explain a model?



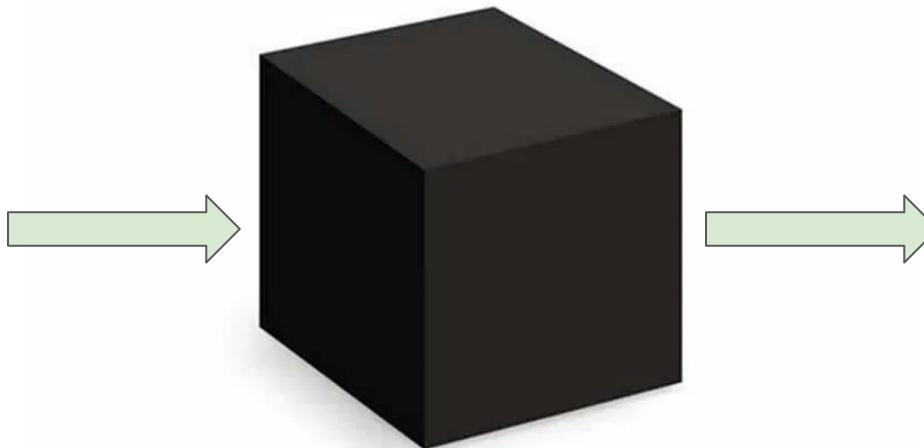
Henk is a 36 year
old male lawyer
from Amsterdam



How do we explain a model?



Henk is a 36 year old male lawyer from Amsterdam



€5000
credit



How do we explain a model?



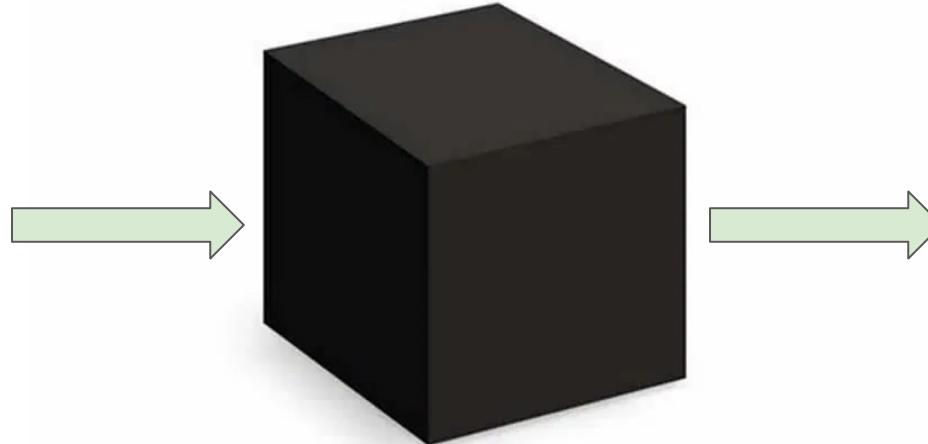
Suzan is a 32 year
old female doctor
from Utrecht



How do we explain a model?



Suzan is a 32 year old female doctor from Utrecht



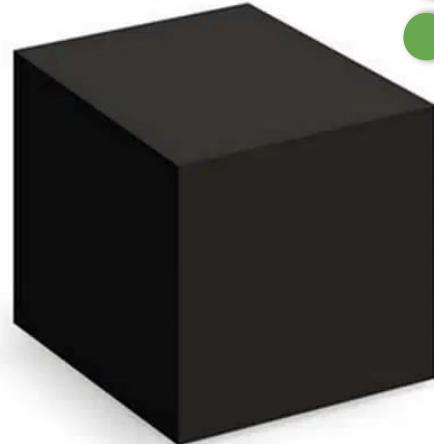
€1000
credit



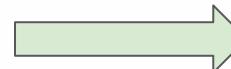
How do we explain a model?



Suzan is a 32 year old female doctor from Utrecht



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



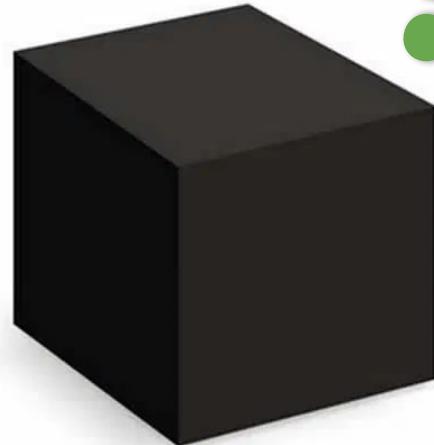
€1000 credit



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



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?

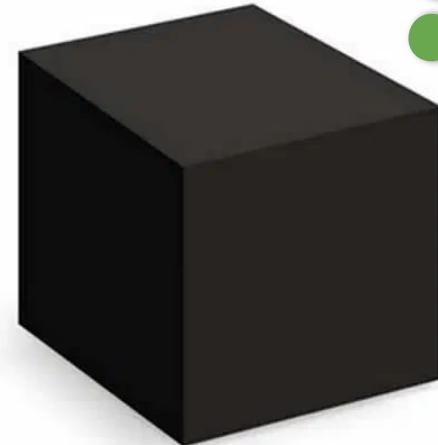




Explanation Faithfulness



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

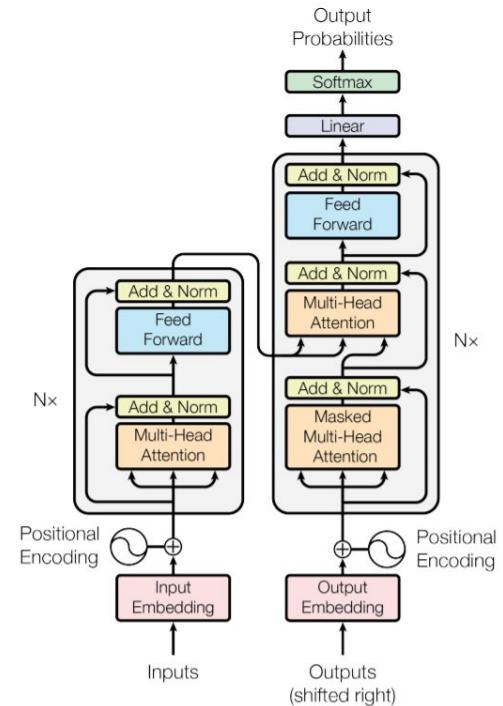


Figure 1: The Transformer - model architecture.



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?

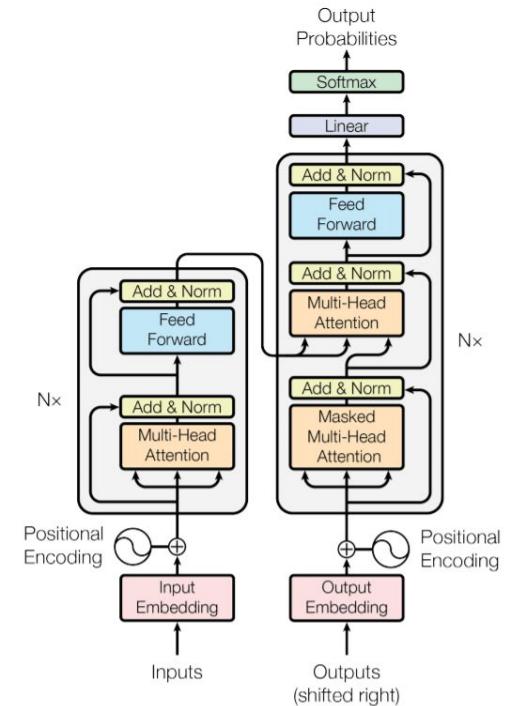


Figure 1: The Transformer - model architecture.



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

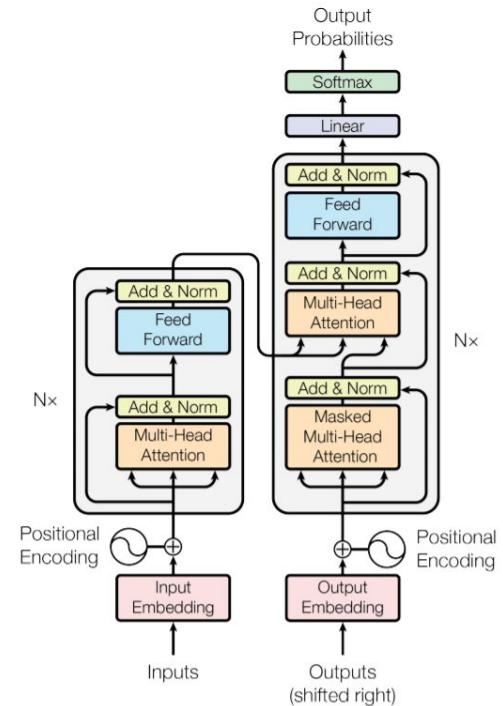


Figure 1: The Transformer - model architecture.



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?

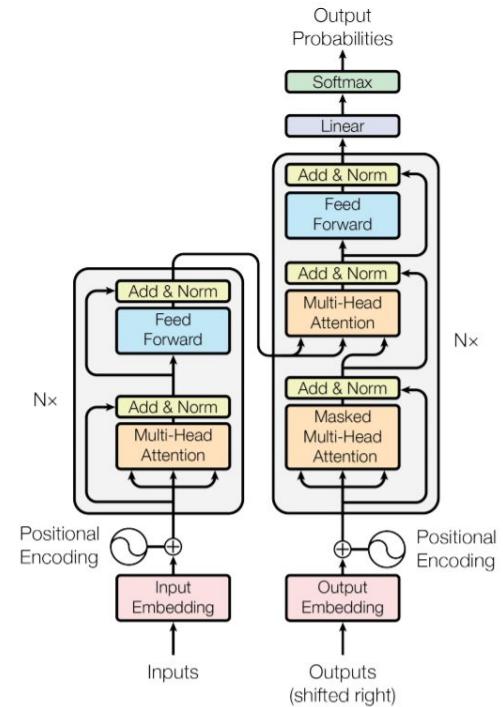


Figure 1: The Transformer - model architecture.



Behavioural Interpretability

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

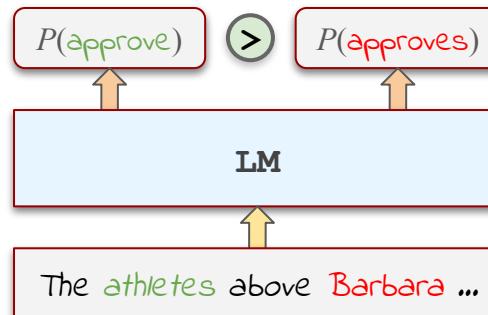
- Using carefully crafted **minimal pairs** we can investigate a model’s performance on a specific phenomenon.



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.





- 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*:

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



Phenomenon	N	Acceptable Example	Unacceptable Example
ANAPHOR AGR.	2	<i>Many girls insulted <u>themselves</u>.</i>	<i>Many girls insulted <u>herself</u>.</i>



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



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



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

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.



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%.



BLiMP

	5-gram	0.34	0.39	0.58	0.59	1
5-gram		0.34	0.39	0.58	0.59	1
LSTM	0.49		0.63	0.9	1	0.59
TXL	0.48	0.68		1	0.9	0.58
GPT-2	0.54	1	0.68		0.63	0.39
human	1	0.54	0.48	0.49		0.34
	human	GPT-2	TXL	LSTM	5-gram	

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

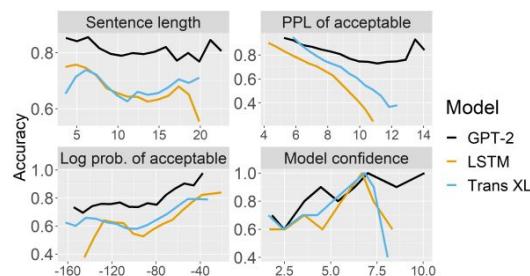


Figure 2: Models' performance on BLiMP as a function of sentence length, perplexity, log probability of the acceptable sentence, and model confidence (calculated as $|\log P(S_1) - \log P(S_2)|$).



BLiMP

	human	GPT-2	TXL	LSTM	5-gram
5-gram	0.34	0.39	0.58	0.59	1
LSTM	0.49	0.63	0.9	1	0.59
TXL	0.48	0.68	1	0.9	0.58
GPT-2	0.54	1	0.68	0.63	0.39
human	1	0.54	0.48	0.49	0.34

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

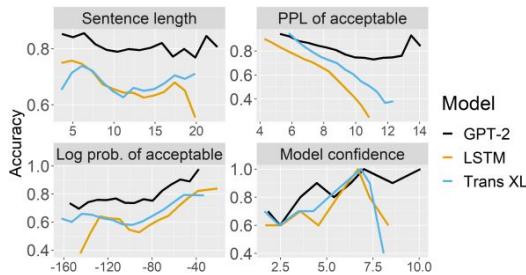


Figure 2: Models' performance on BLiMP as a function of sentence length, perplexity, log probability of the acceptable sentence, and model confidence (calculated as $|\log P(S_1) - \log P(S_2)|$).

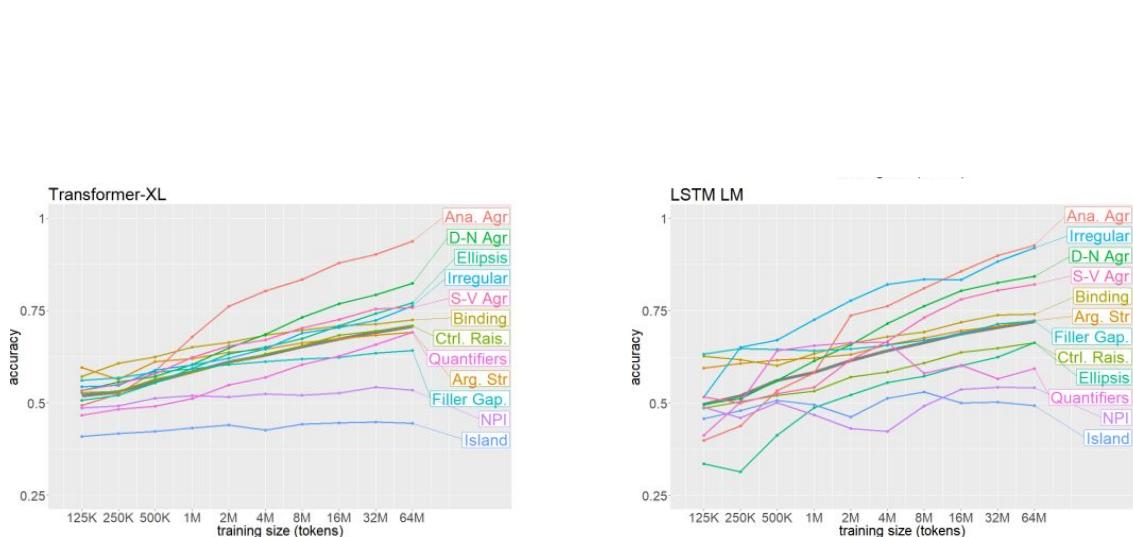


Figure 5: Transformer-XL (top) and LSTM LM (bottom) performance as a function of training size and phenomena in BLiMP. The gray line shows the average across all phenomena.



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*

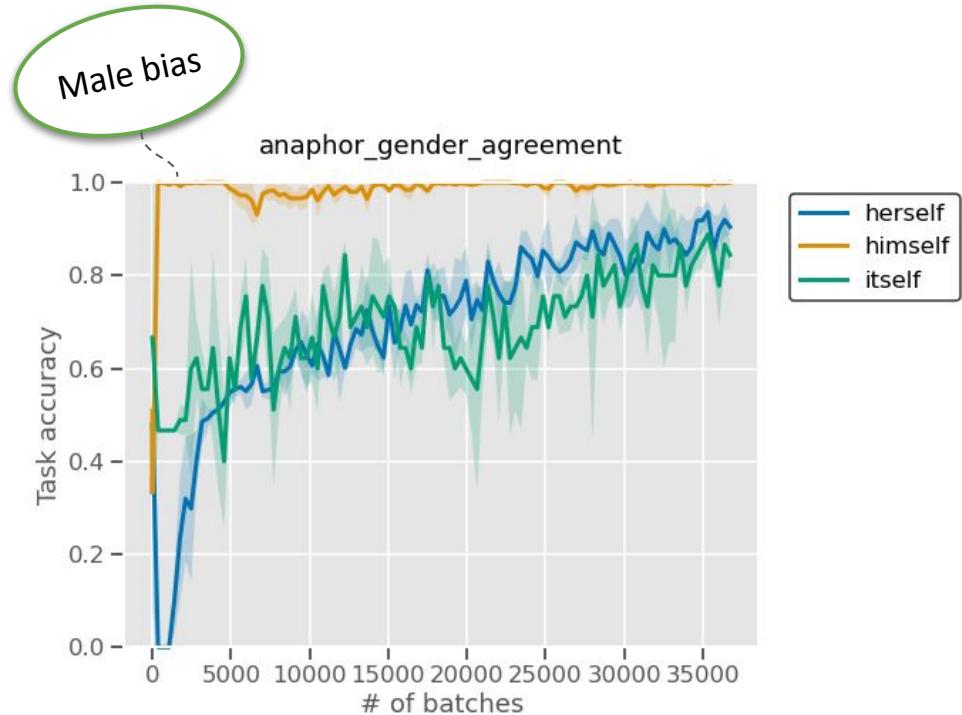


Behavioural Tests for Uncovering Biases

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*





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:

Type 1

The physician hired the secretary because he was overwhelmed with clients.

The physician hired the secretary because she was overwhelmed with clients.

The physician hired the secretary because she was highly recommended.

The physician hired the secretary because he was highly recommended.

Type 2

The secretary called the physician and told him about a new patient.

The secretary called the physician and told her about a new patient.

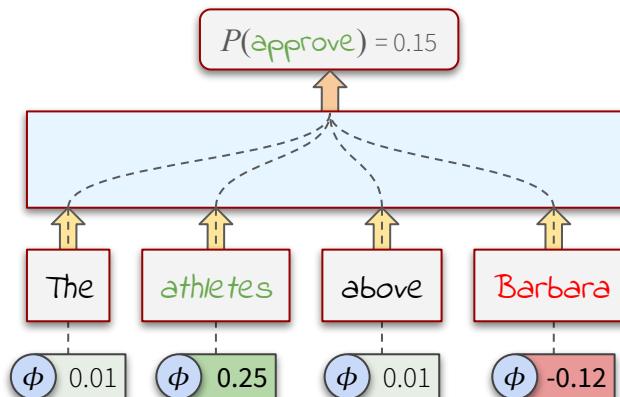
The physician called the secretary and told her to cancel the appointment.

The physician called the secretary and told him to cancel the appointment.



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**.

Pronoun Resolution

Pronoun resolution:

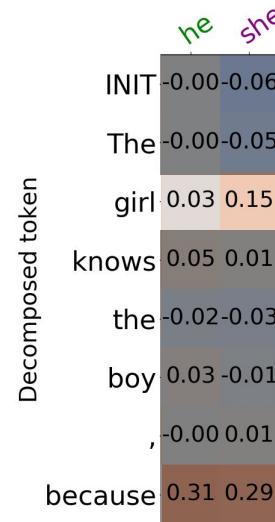
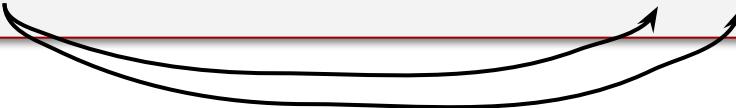
The girl knows the boy, because ...

$P(\text{she})$

$P(\text{he})$

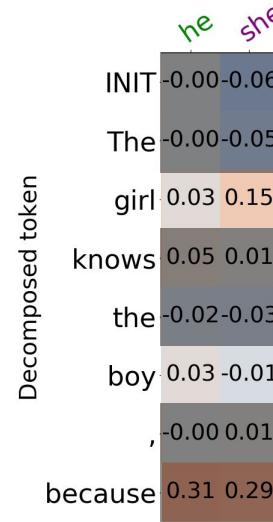
Pronoun Resolution

The girl knows the boy, because he/she



Pronoun Resolution

The girl knows the boy, because he/she

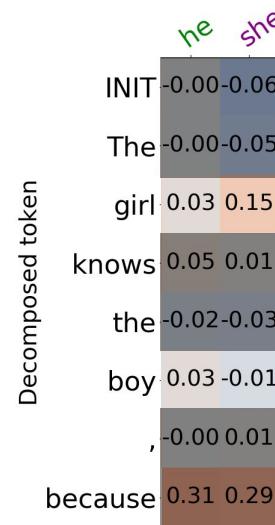


Pronoun Resolution

The girl knows the boy, because he/she

$C(\text{he})$

$- C(\text{she})$

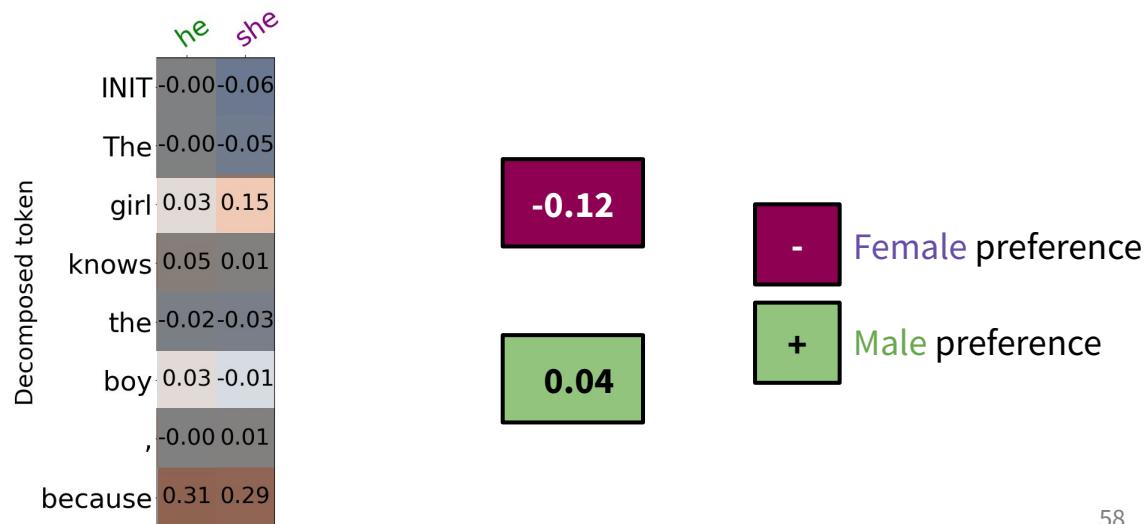


- Female preference
- + Male preference

Pronoun Resolution

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

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



Average contributions

The girl knows the boy, because he/she

Female
subject

Male
object

$C(\text{he})$

-

$C(\text{she})$

- Female preference
- + Male preference

Average contributions

The girl knows the boy, because he/she

$C(\text{he})$

$- C(\text{she})$

FM

subj_F -0.19

- Female preference
- + Male preference

Average contributions

The girl knows the boy, because he/she

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

FM

subj_F -0.19

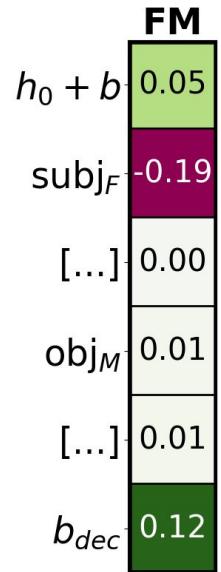
obj_M 0.01

- Female preference
- + Male preference

Average contributions

The girl knows the boy, because he/she

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



- Female preference
- + Male preference

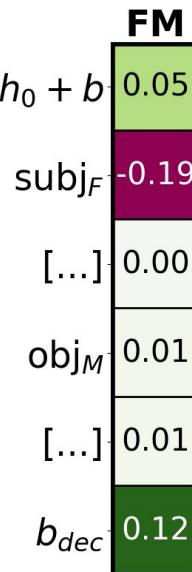
Default Reasoning?

The girl knows the boy, because he/she

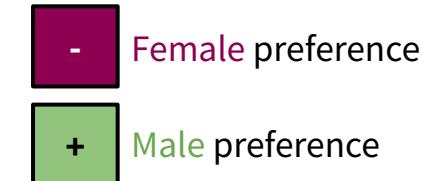
$C(\text{he})$

$- C(\text{she})$

Intercepts & initial states
biased by default
towards **Male preference**



Female preference
requires explicit evidence

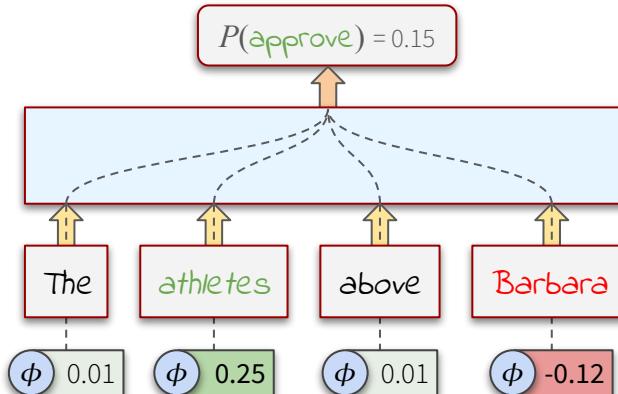




Feature Attribution Methods

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.

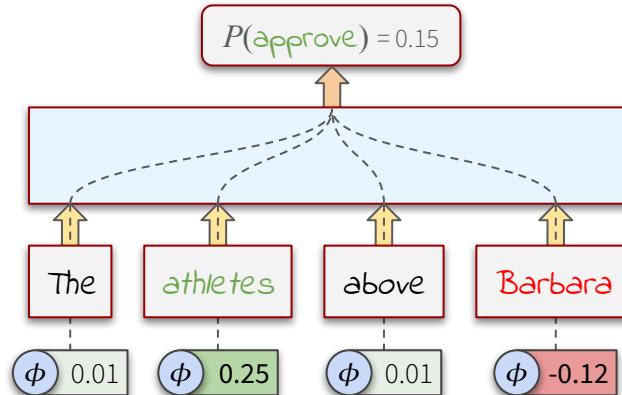




Feature Attribution Methods

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

ICOVERT@CS.WASHINGTON.EDU

SCOTT.LUNDBERG@MICROSOFT.COM

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

Scott Lundberg

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

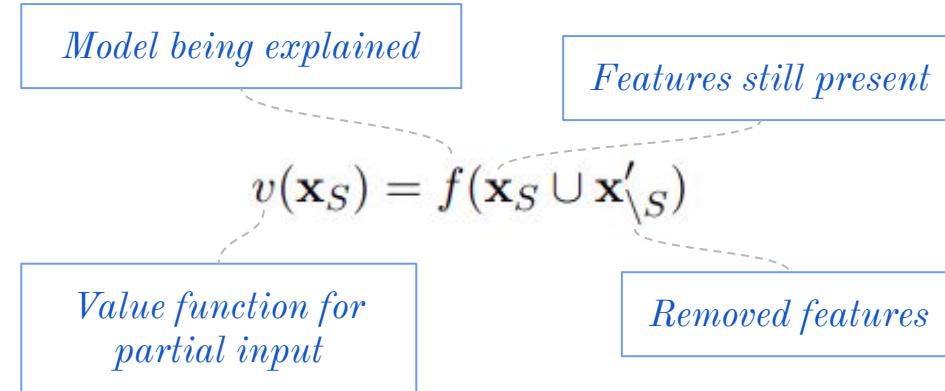
Model being explained

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

*Value function for
partial input*

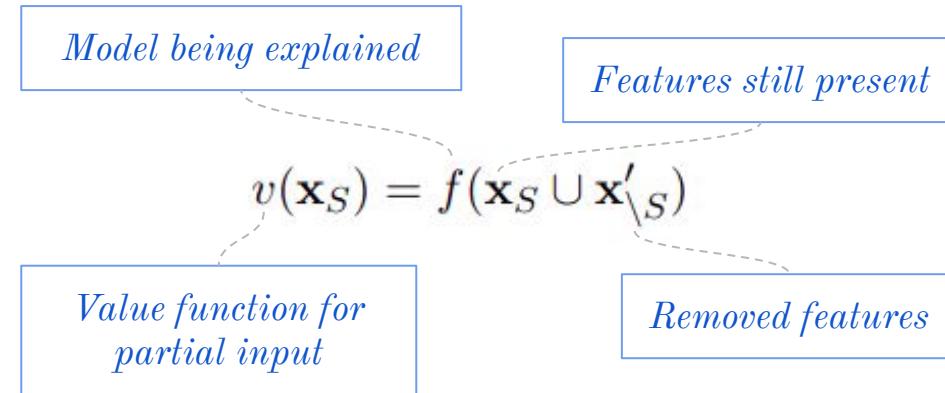
Feature Removal

1 Static Baseline



Feature Removal

1 Static Baseline



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

Feature Removal

2 Interventional background distribution

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

Expectation over removed features

Feature Removal

2 Interventional background distribution

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

Expectation over removed features

\mathbf{x} = “This movie is not bad”

S = {1, 2, 3, 5}

$\mathbf{x}_S \cup \mathbf{x}'_{\setminus S}$ = “This movie is *the* *bad*”

is
walk

...

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

Conditioned on present features

3 Observational background distribution

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

Expectation over removed features

\mathbf{x} = “This movie is not bad”

S = {1, 2, 3, 5}

$\mathbf{x}_S \cup \mathbf{x}'_{\setminus S}$ = “This movie is very
quite
pretty
... bad”

Feature Influence

1 Ablation

Contribution of feature i

$$\phi_i = v(\mathbf{x}) - v(\mathbf{x}_{\setminus i})$$

\mathbf{x} = “*This movie is not bad*”

\mathbf{x}' = <pad>

$$\Phi_{not} = f(\text{“This movie is not bad”}) - f(\text{“This movie is <pad> bad”})$$

Feature Influence

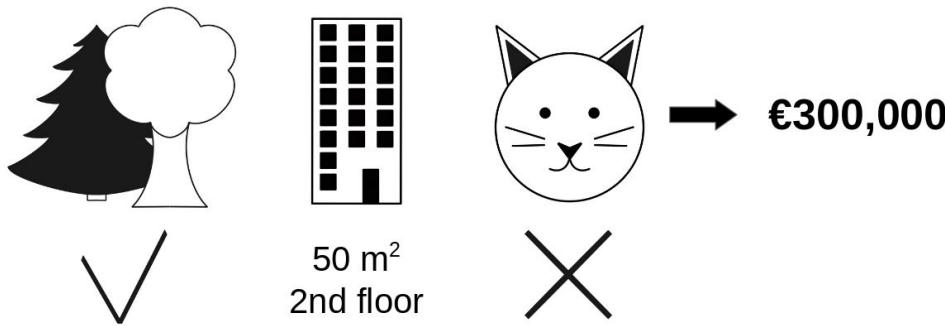
2 Shapley Value

$$\phi_i = \sum_{\substack{\text{coalitions} \\ S \subseteq \mathbf{x} \setminus \{i\}}} p(S) \cdot \overbrace{(v(\mathbf{x}_{S \cup i}) - v(\mathbf{x}_S))}^{\text{marginal contribution of } i \text{ to coalition}}$$
$$p(S) = \underbrace{\frac{|S|!(|\mathbf{x}| - 1 - |S|)!}{|\mathbf{x}|!}}_{\text{relative number of coalitions of size } |S|}$$

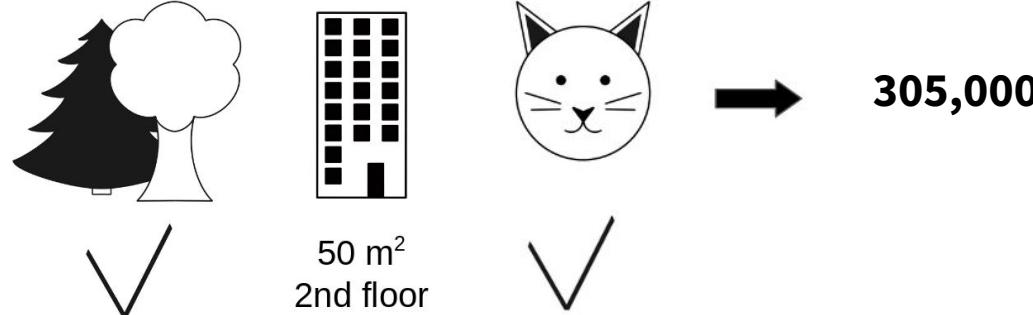
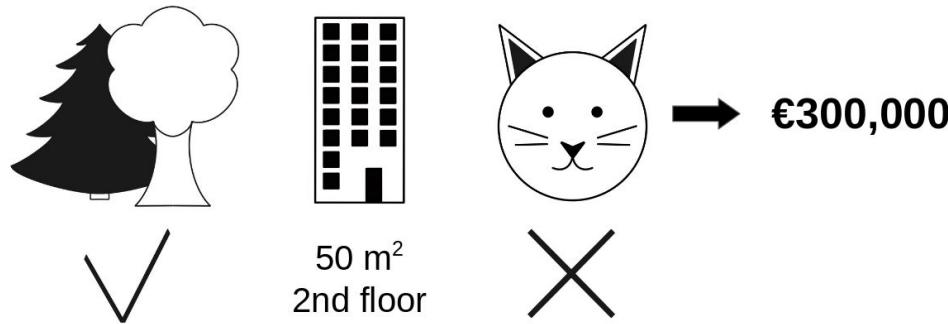
Completeness axiom:

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

Shapley Values



Shapley Values



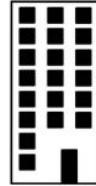
Shapley Values



€310,000



50 m²
1st floor



€320,000



50 m²
1st floor



Shapley Values



50m²



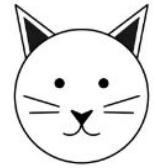
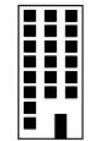
50m²



50m²



50m²



Feature Influence

2 Shapley Value

$$\phi_i = \sum_{\substack{\text{coalitions} \\ S \subseteq \mathbf{x} \setminus \{i\}}} p(S) \cdot \overbrace{(v(\mathbf{x}_{S \cup i}) - v(\mathbf{x}_S))}^{\text{marginal contribution of } i \text{ to coalition}}$$
$$p(S) = \underbrace{\frac{|S|!(|\mathbf{x}| - 1 - |S|)!}{|\mathbf{x}|!}}_{\text{relative number of coalitions of size } |S|}$$

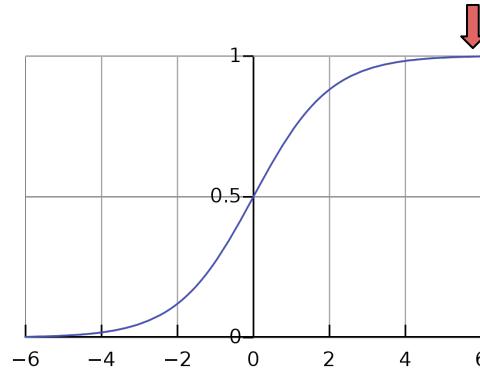
Completeness axiom:

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

Feature Influence

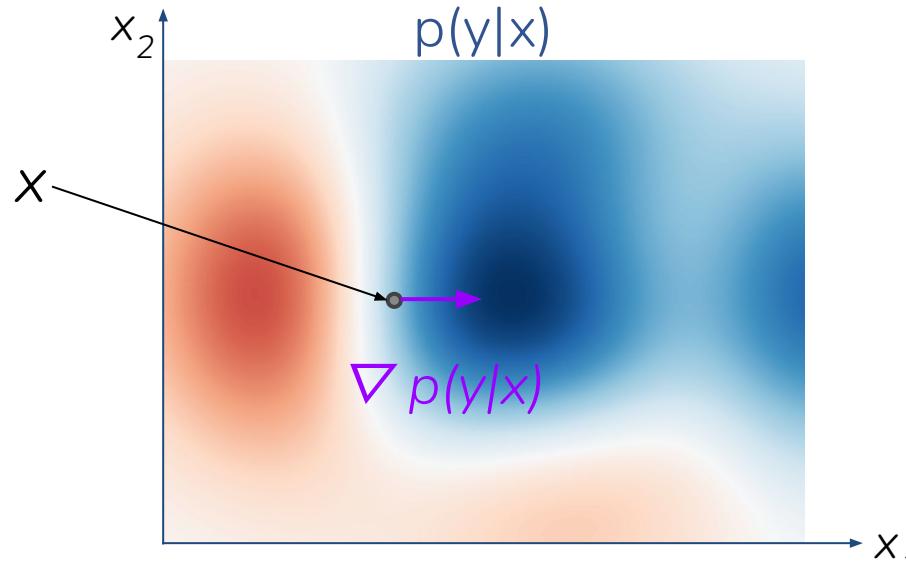
3 Gradients

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



Highlighting via Input Gradients

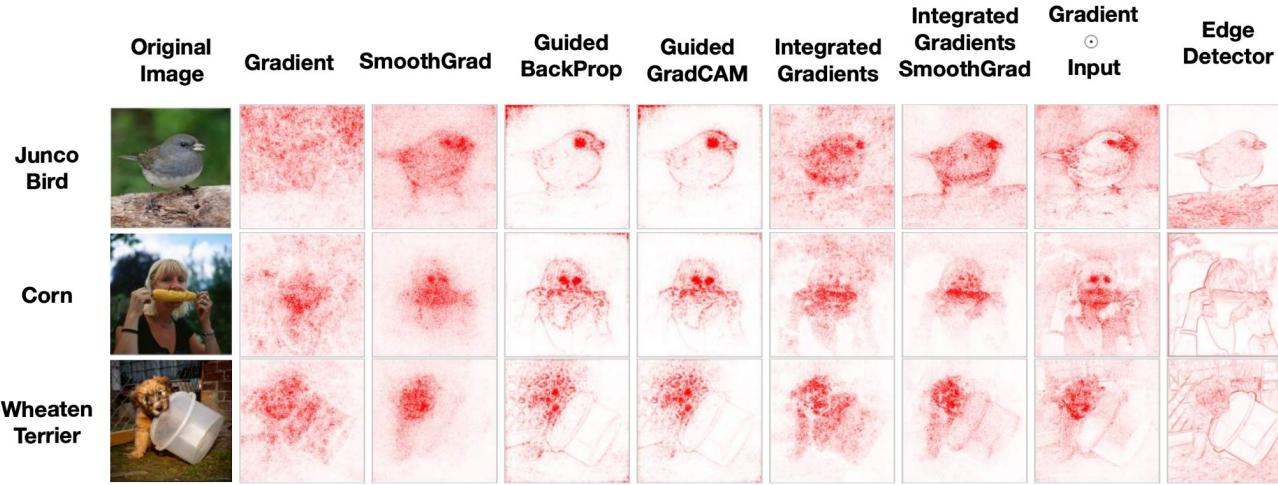
- Estimate importance of a feature using derivative of output w.r.t that feature
- i.e., with a “tiny change” to the feature, what happens to the prediction?



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

[[Simonyan et al. 2014](#)]

Example of highlighting: Image classification



[Adebayo et al., 2018]

85

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?

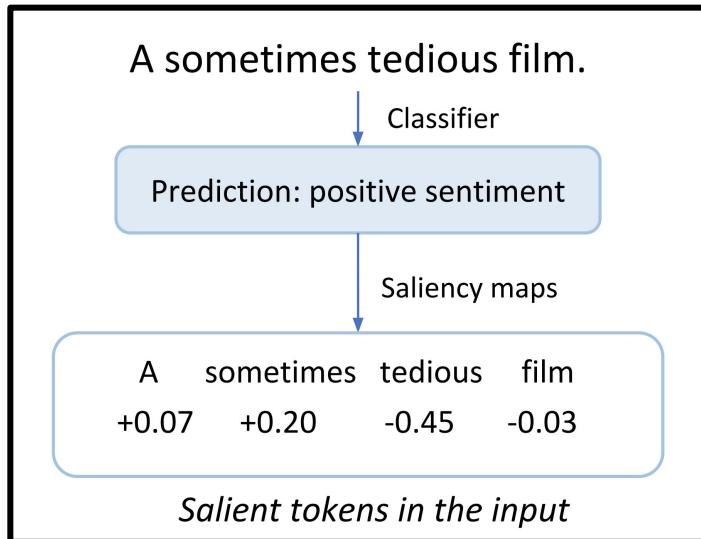
direction lead to
a decrease in
the loss

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

L1-normalized across all tokens

Gradient-based Highlightings for NLP

For NLP, derivative of output w.r.t a feature
=
derivative of **output** w.r.t an **input token**



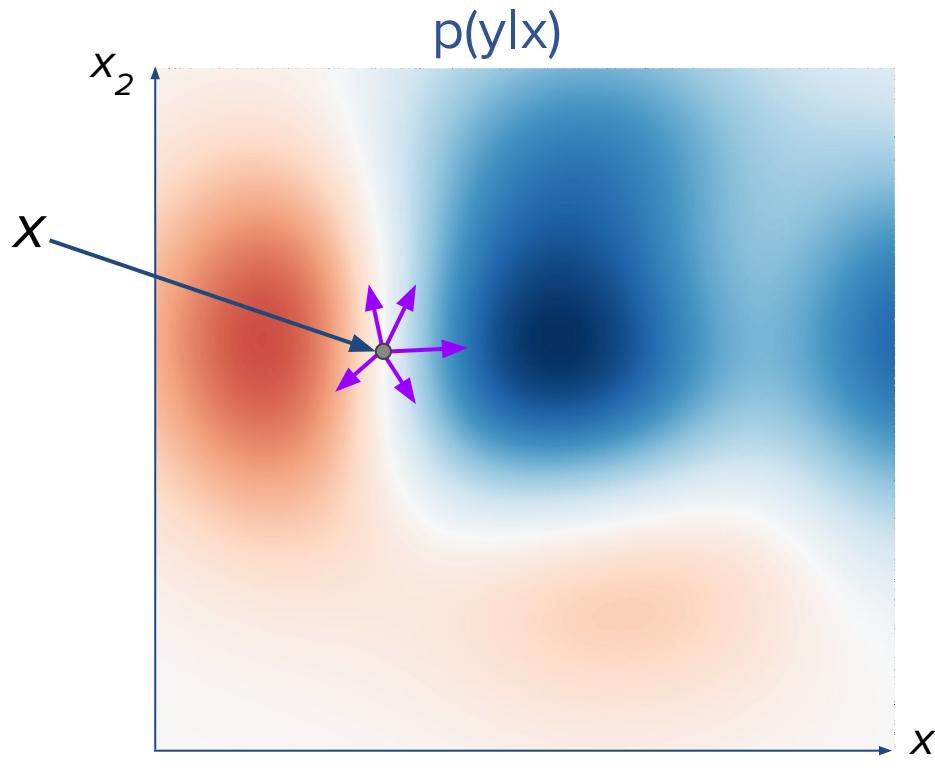
direction lead to
a decrease in
the loss

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

L1-normalized across all tokens

Problems with Using Gradient for Highlighting

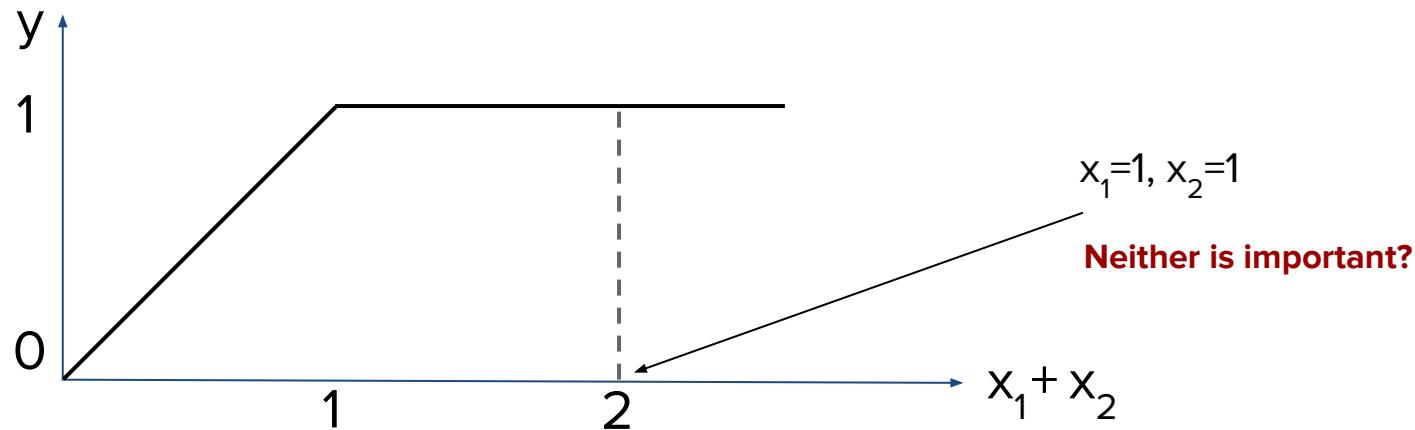
- Too “local” and thus sensitive to slight perturbations



Problems with Using Gradient for Highlighting

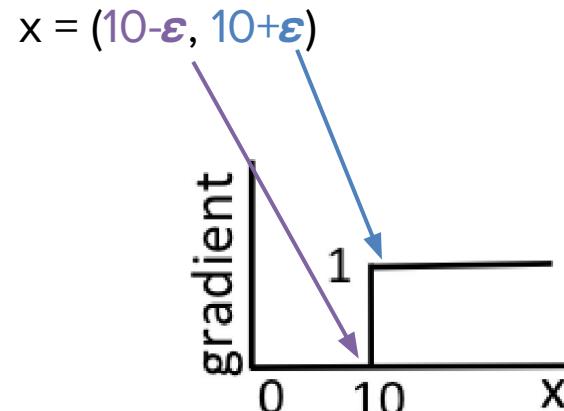
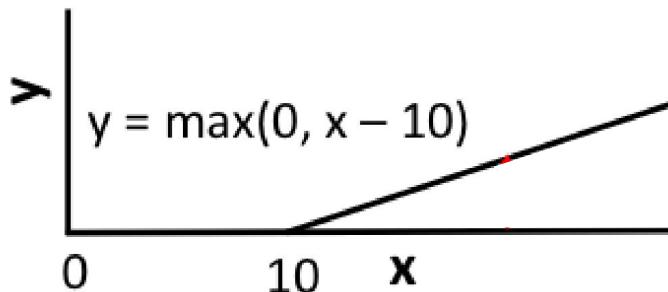
- too “local” and thus sensitive to slight perturbations
- “saturated outputs” lead to unintuitive gradients

$$y = \begin{cases} x_1 + x_2 & \text{when } (x_1 + x_2) < 1 \\ 1 & \text{when } (x_1 + x_2) \geq 1 \end{cases}$$



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



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

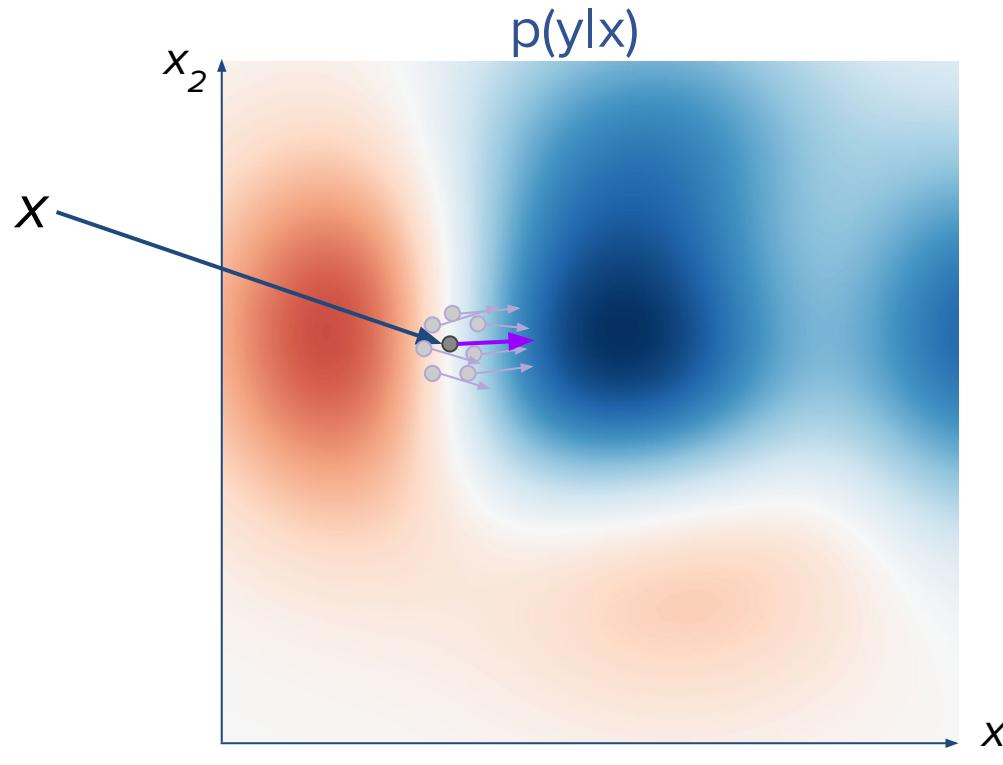
How to mitigate these issues? Don’t rely on a single gradient calculation:

- SmoothGrad
- Integrated Gradients

Other approaches, e.g., [LRP](#), [DeepLIFT](#), [GradCAM](#). Not covered here.

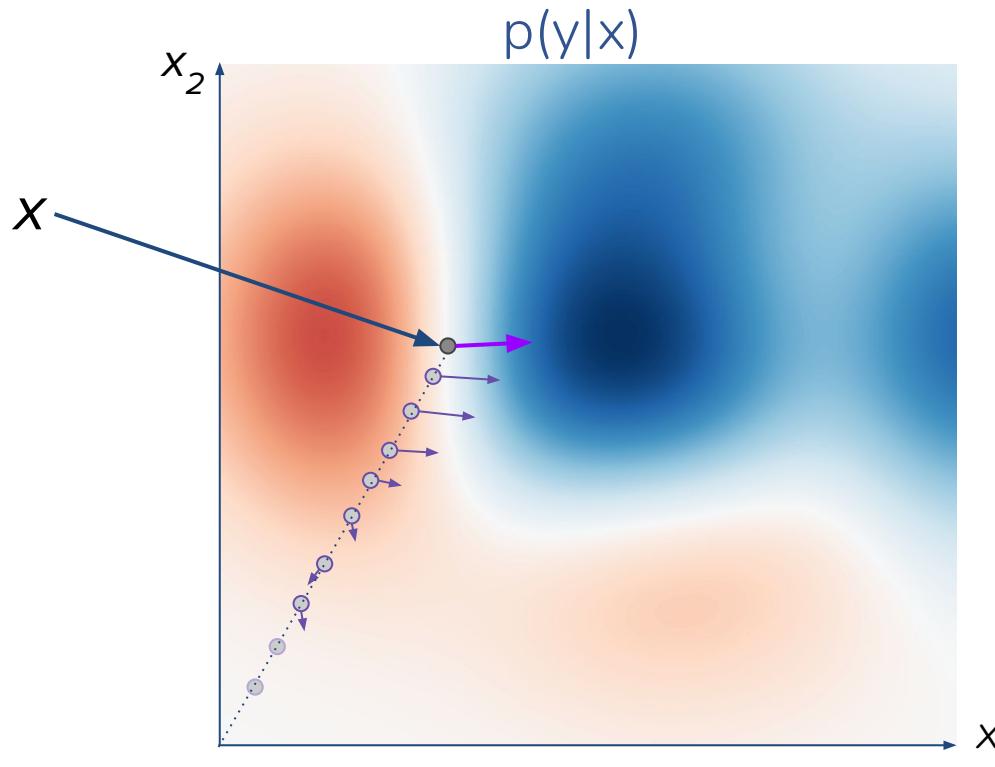
Extensions of Vanilla Gradient

SmoothGrad: add gaussian noise to input and average the gradient



Extensions of Vanilla Gradient

Integrated Gradients: average gradients along path from zero to input



Summary of Gradient-based Highlighting

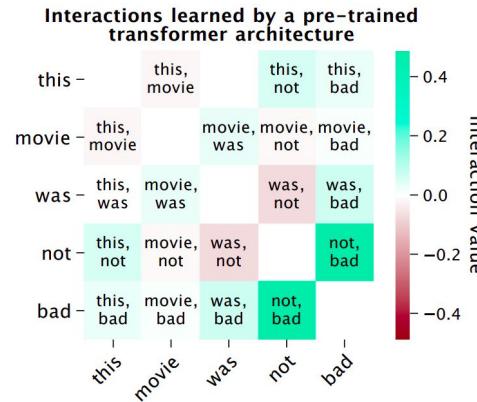
Positives:

- 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 common sense) that is not explicitly mentioned in the input
- Ignore the interactions between words/pixels (e.g., “not good”)

Summary of Gradient-based Highlighting



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



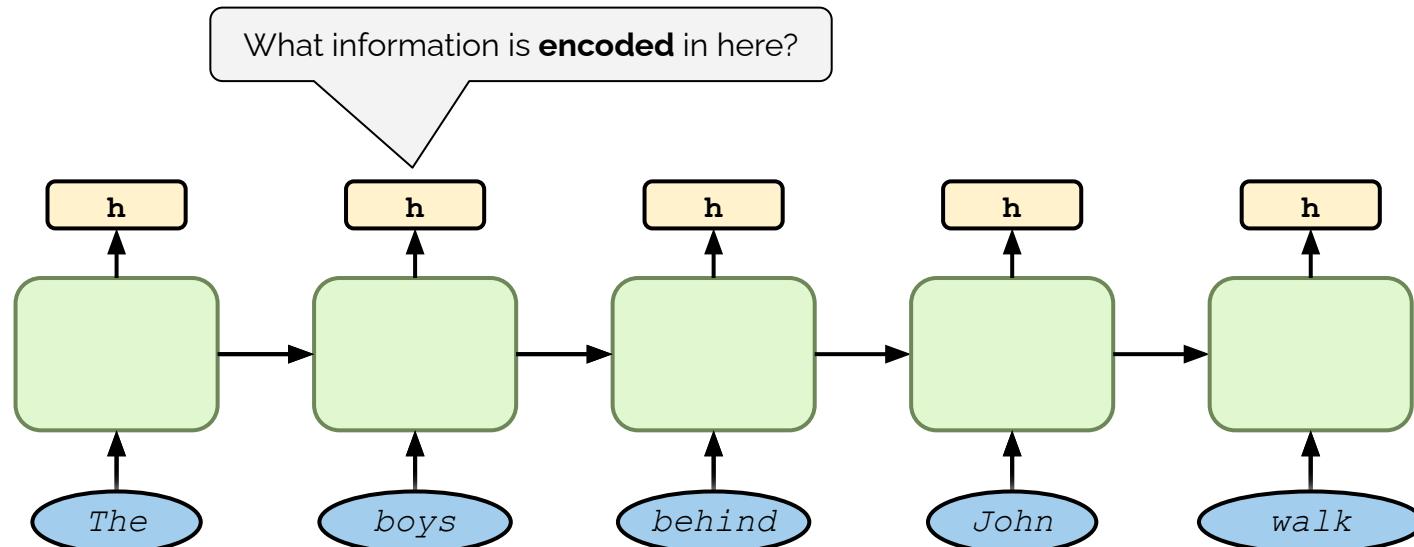
Probing

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

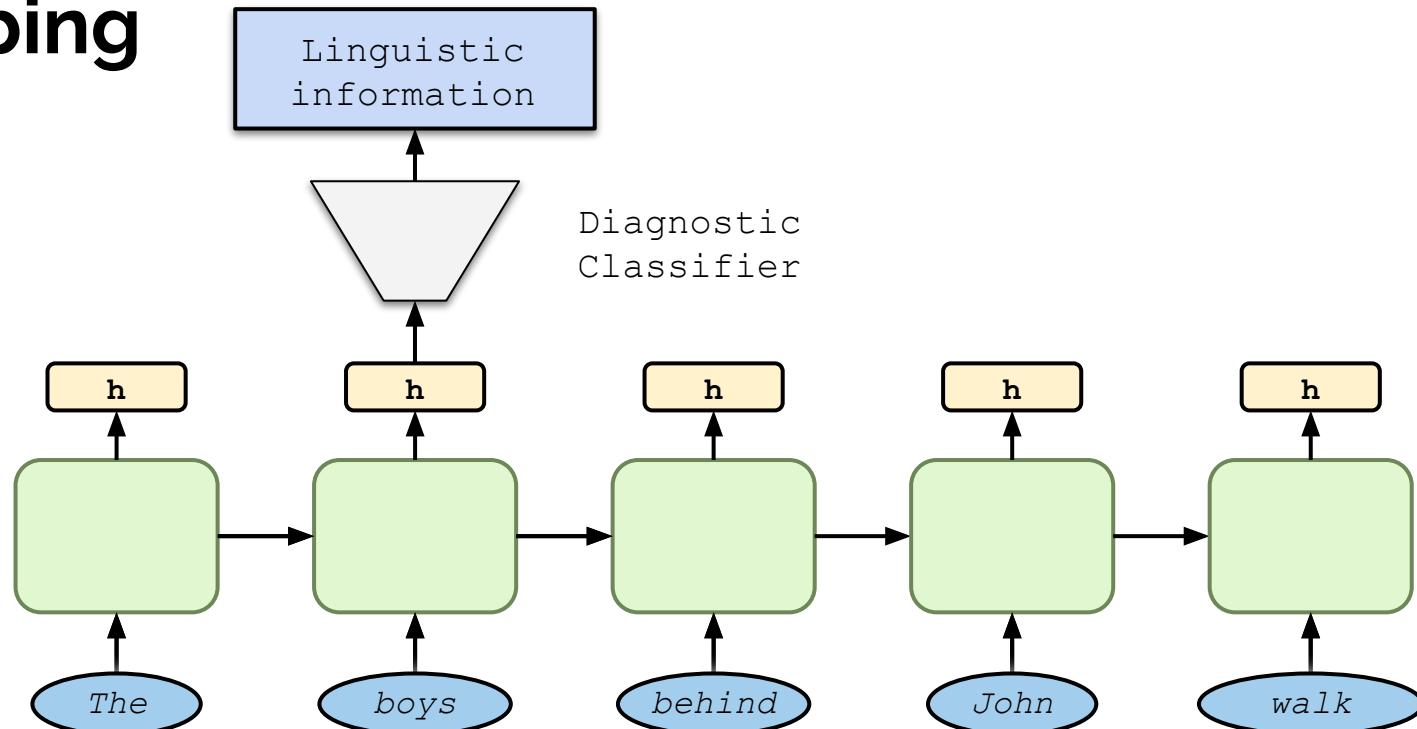
- ✗ They do not show *where* in the model predictions are formed
- ✗ They give no insight into **higher-level** concepts such as ‘gender’, ‘number’, or ‘part-of-speech’ class.

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

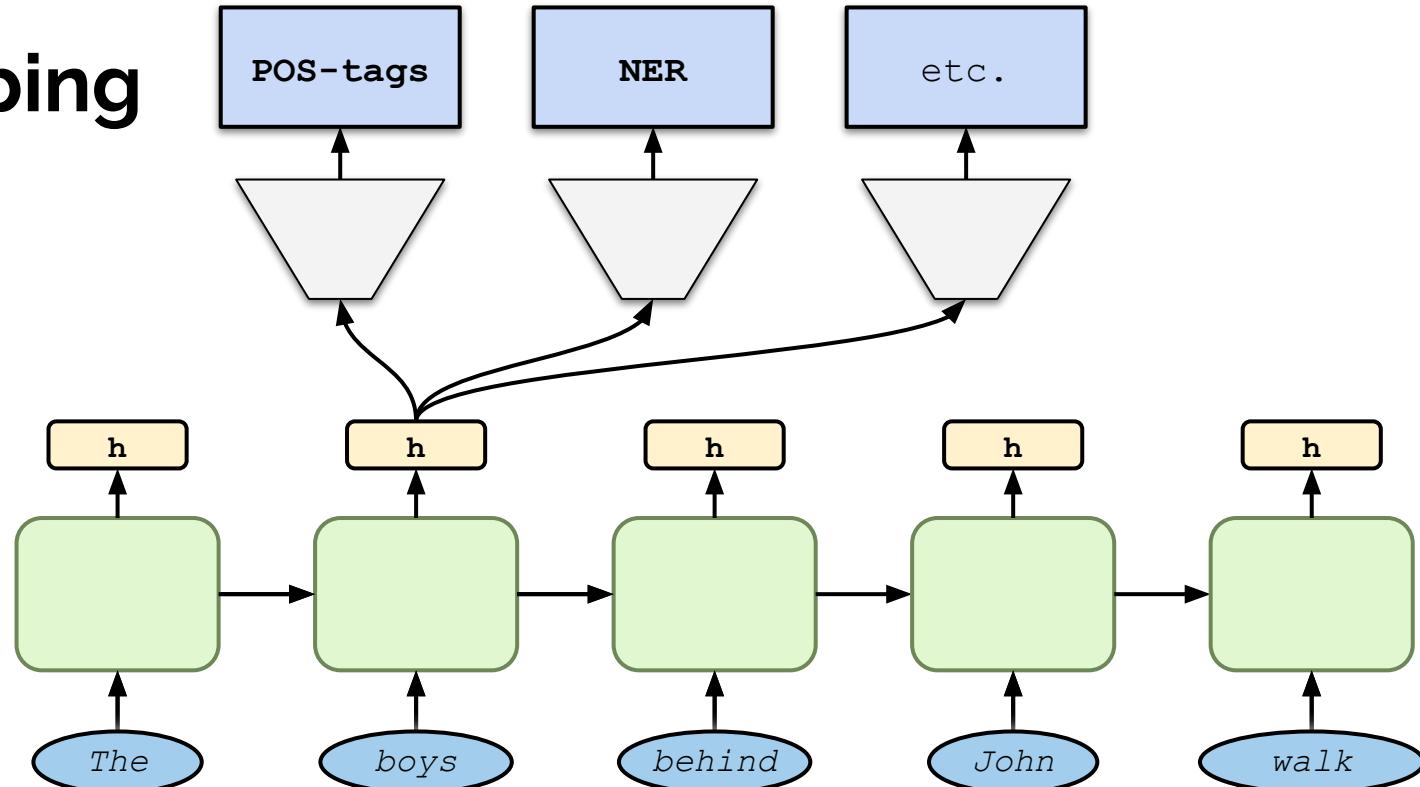
Probing



Probing



Probing



Representations

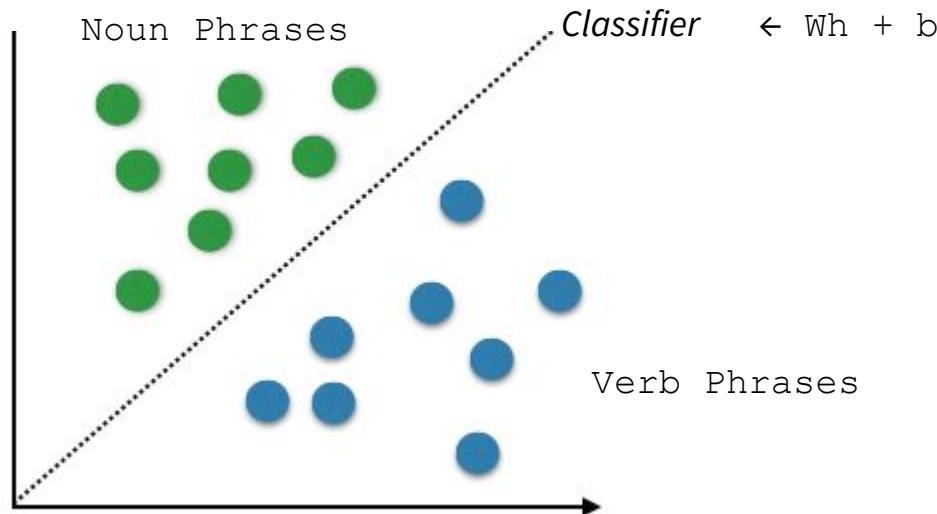
- Representations are just a point in a vector space
- But, it is likely that the representation of “cat” is somewhat similar to “dog”

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

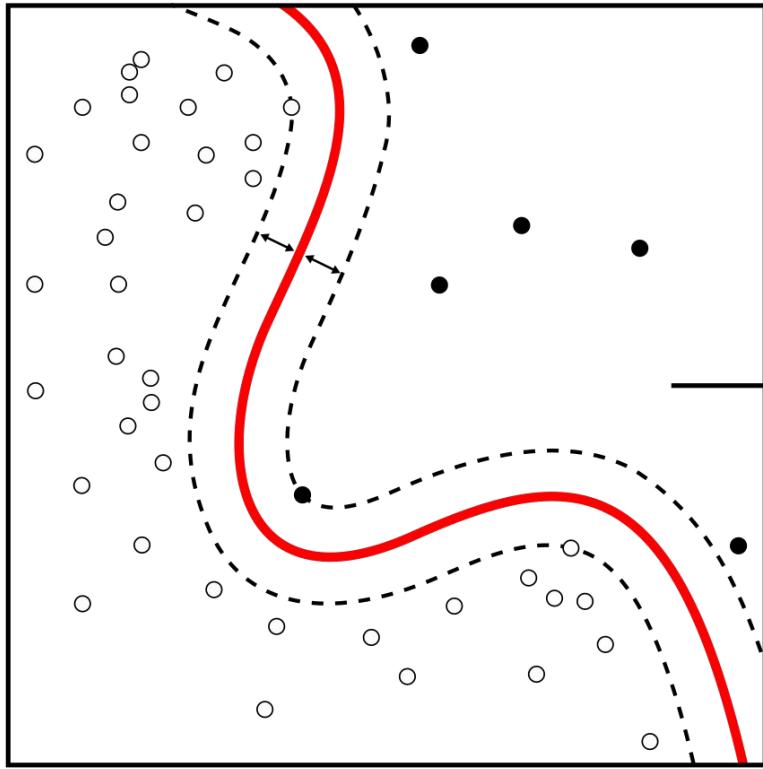
- More generally, the representation of **nouns** are likely to be similar, and distinct from **verbs, determiners, adverbs**, etc.

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

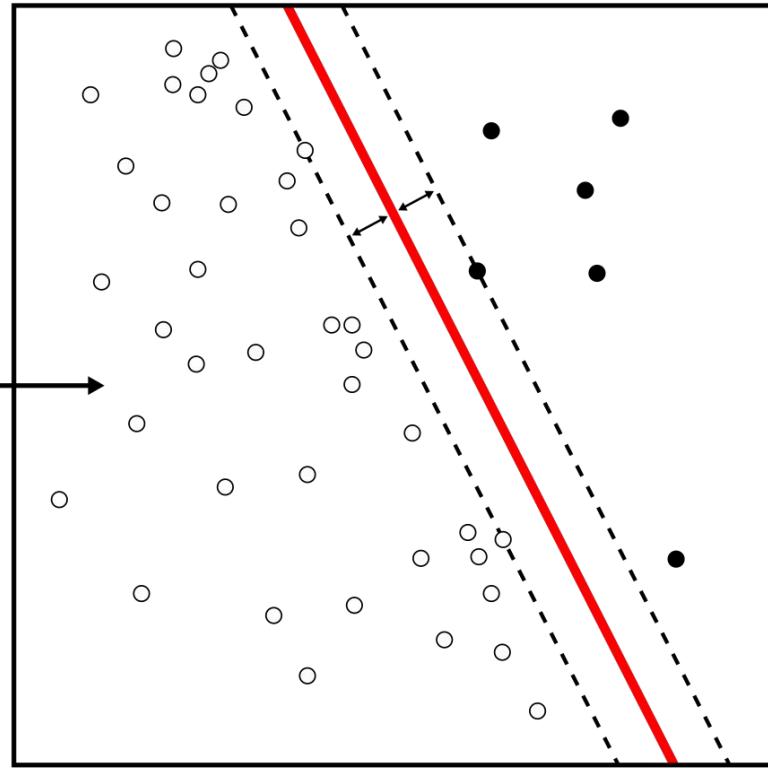
What does *probed info* imply?



Why linear?



\emptyset



Probing

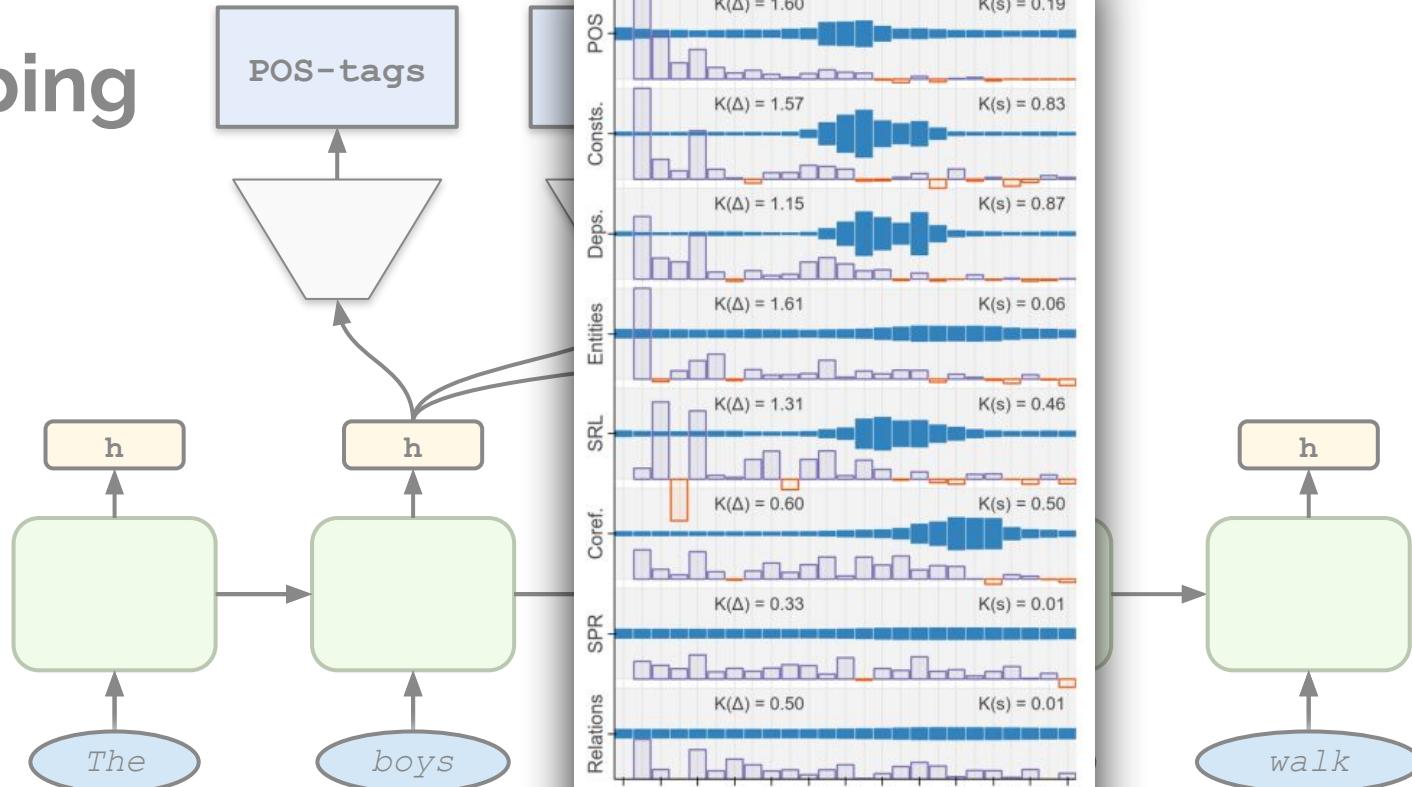


Figure 2: Layer-wise metrics on BERT-large. Solid (blue) are mixing weights $s_\tau^{(\ell)}$ (§3.1); outlined (purple) are differential scores $\Delta_\tau^{(\ell)}$ (§3.2), normalized for each task. Horizontal axis is encoder layer.



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. Gonen et al. (2019) - *How Does Grammatical Gender Affect Noun Representations in Gender-Marking Languages?*
6. Warstadt et al. (2020) - *BLiMP: The Benchmark of Linguistic Minimal Pairs for English*
7. Jumelet et al. (2019) - *Analysing Neural Language Models: Contextual Decomposition Reveals Default Reasoning in Number and Gender Assignment*
8. Covert et al. (2021) - *Explaining by removing: a unified framework for model explanation*
9. Tenney et al. (2019) - *BERT RedisCOVERS the Classical NLP Pipeline*
10. Hewitt et al. (2019) - *A Structural Probe for Finding Syntax in Word Representations*