

Interpreting NLP Models

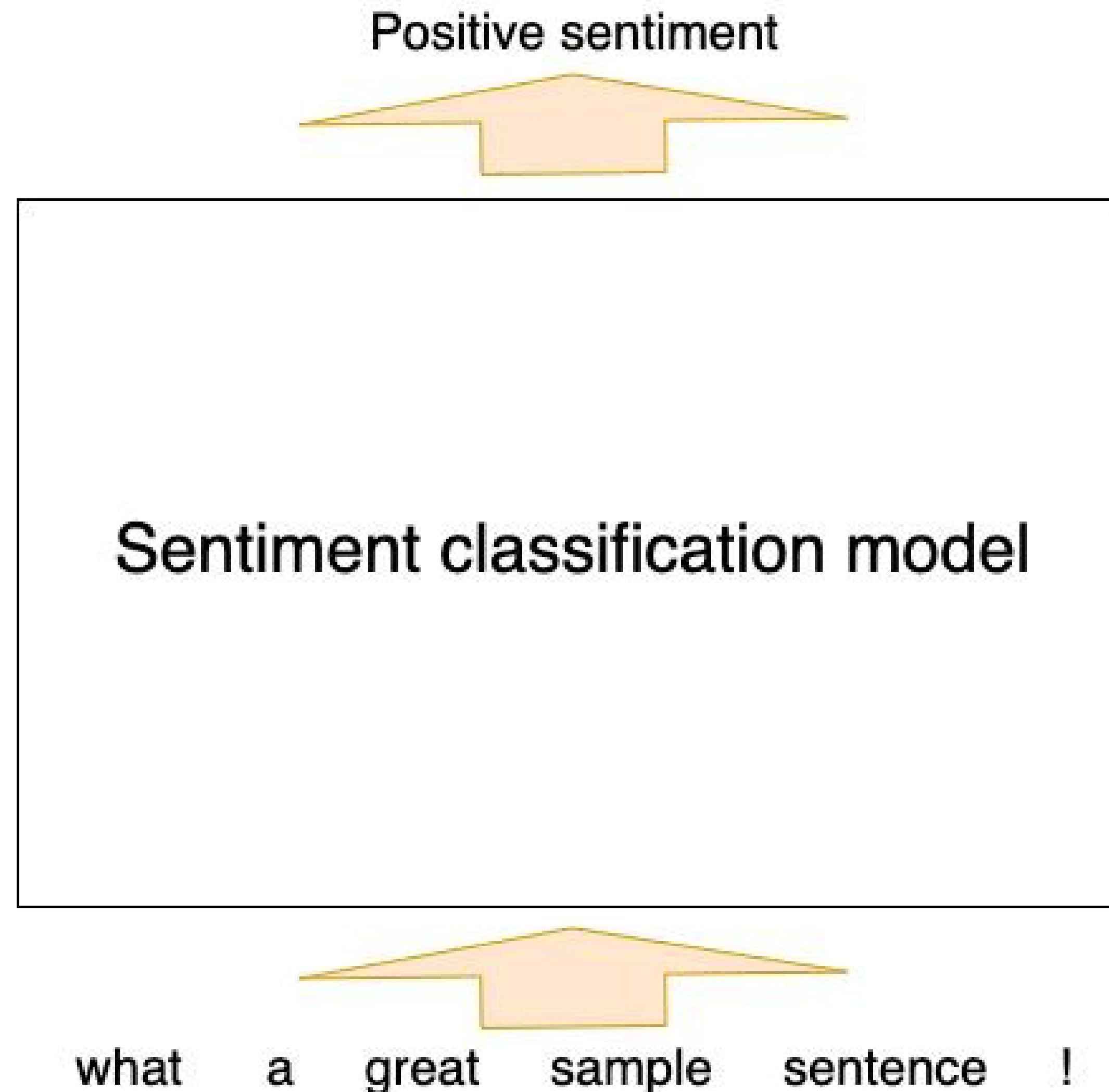
CSE 5525: Foundations of Speech and Natural Language
Processing

<https://shocheen.github.io/courses/cse-5525-fall-2025>

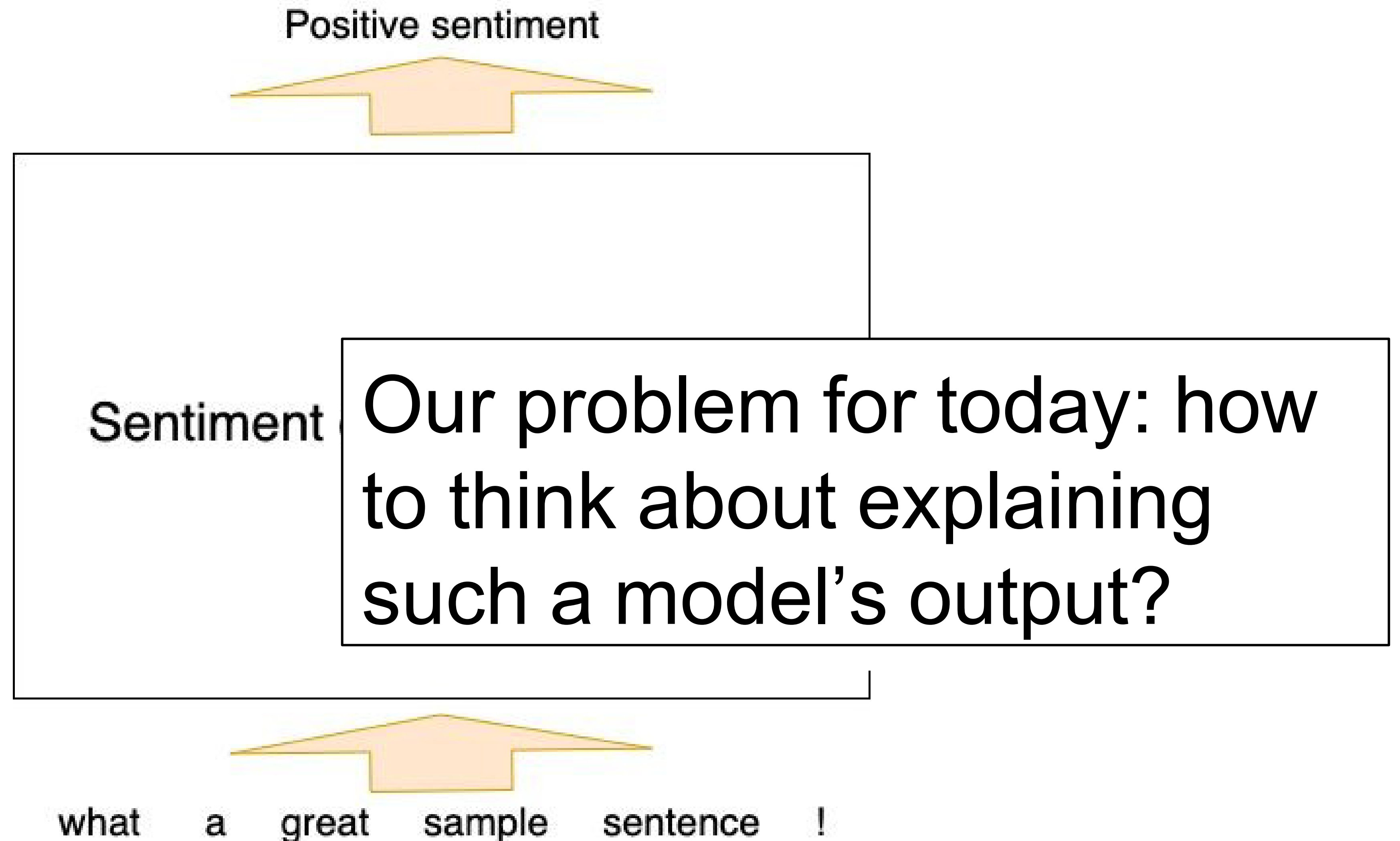


THE OHIO STATE UNIVERSITY

A motivating scenario: text classification

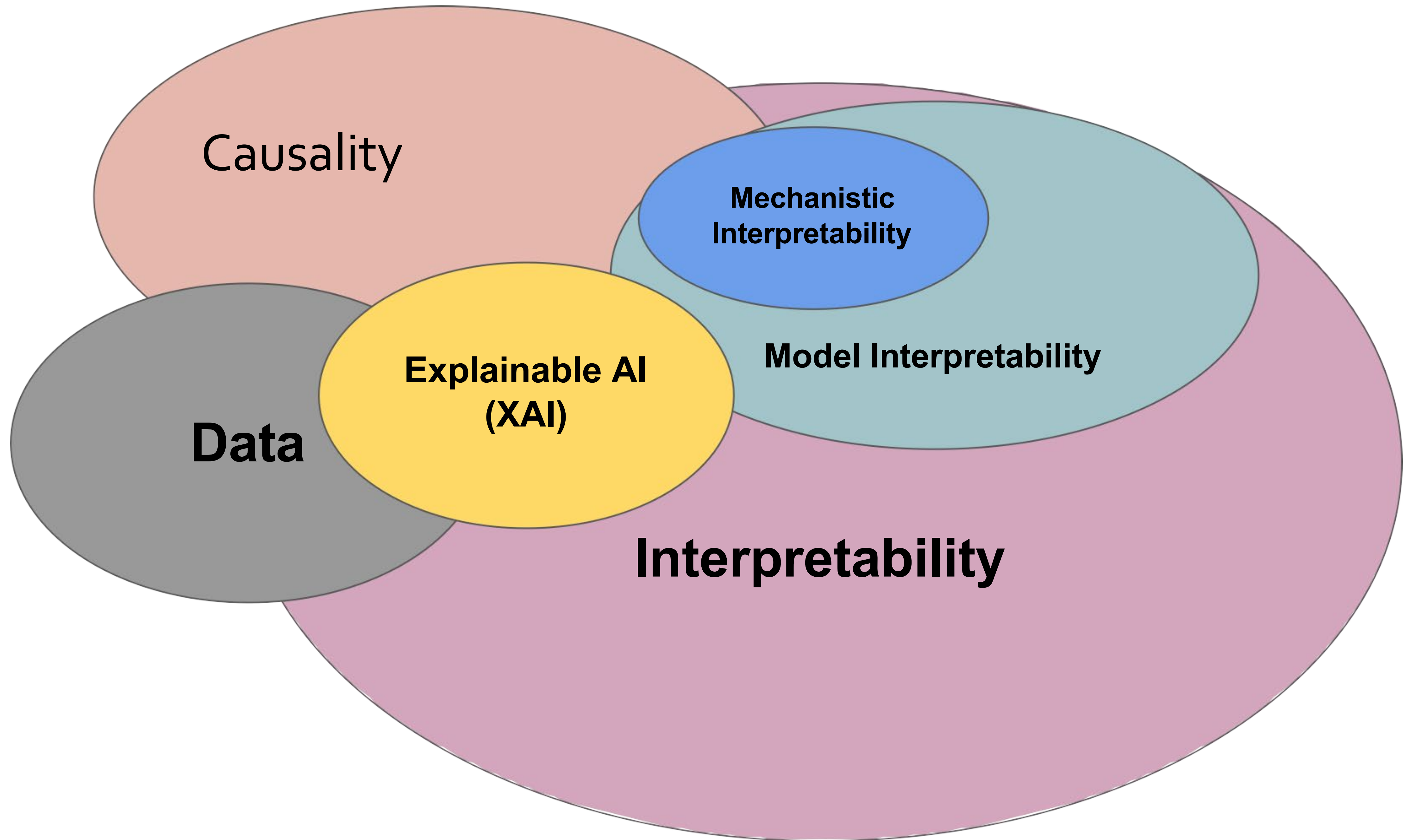


A motivating scenario: text classification



Why might we care about interpreting the reasons for a model's predictions?

- To debug a model
- To help us gain insight into the training data
- To increase confidence in a model by making it easier to flag poor reasons for making a decision
- Helpful to people in human-in-the-loop scenarios for deciding when to take a model's advice into account
- For ethical reasons in cases where people affected by a model's decision are owed an explanation

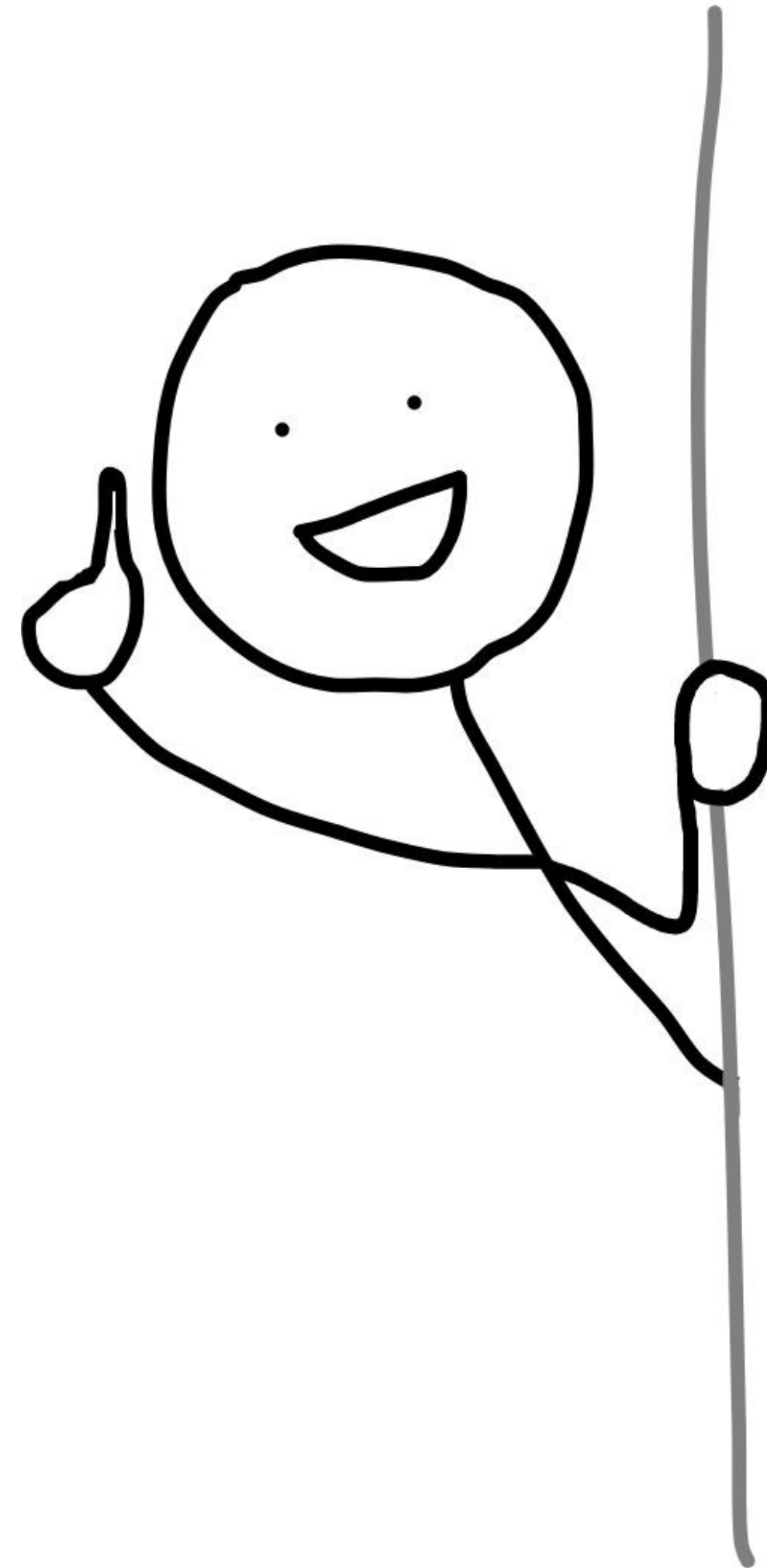


An outline for what we'll talk about

- What do we mean when we talk about an "explanation"? What about an "interpretable model"?
- Global vs local interpretations
- Walking through some post-hoc methods for interpreting a model (Probing, LIME, integrated gradients, attention), plus some discussion of evaluation
- (If time) Basics of mechanistic interpretability

A quick aside about scope

- Most of what we'll be talking about is not exclusively applicable to NLP, but to interpreting (some) machine learning models more broadly
- (but it's work that comes up a lot in interpretability discussions on the NLP side of things too)



Defining terms

What qualities do we look for in an **explanation**?

- Faithfulness
 - Is the explanation true to what the model did?
- Utility to humans
 - Is the explanation helpful to end users?

(see Doshi-Velez and Kim 2017, Madsen et al. 2022)

What qualities do we look for in a (**globally**) interpretable **model**?

- Algorithmic transparency e.g., guarantees about convergence or the shape of the error surface
- Decomposability: Are the different pieces of the model understandable on their own?
- Simulatability: Can a person hold the whole model in their head at once?

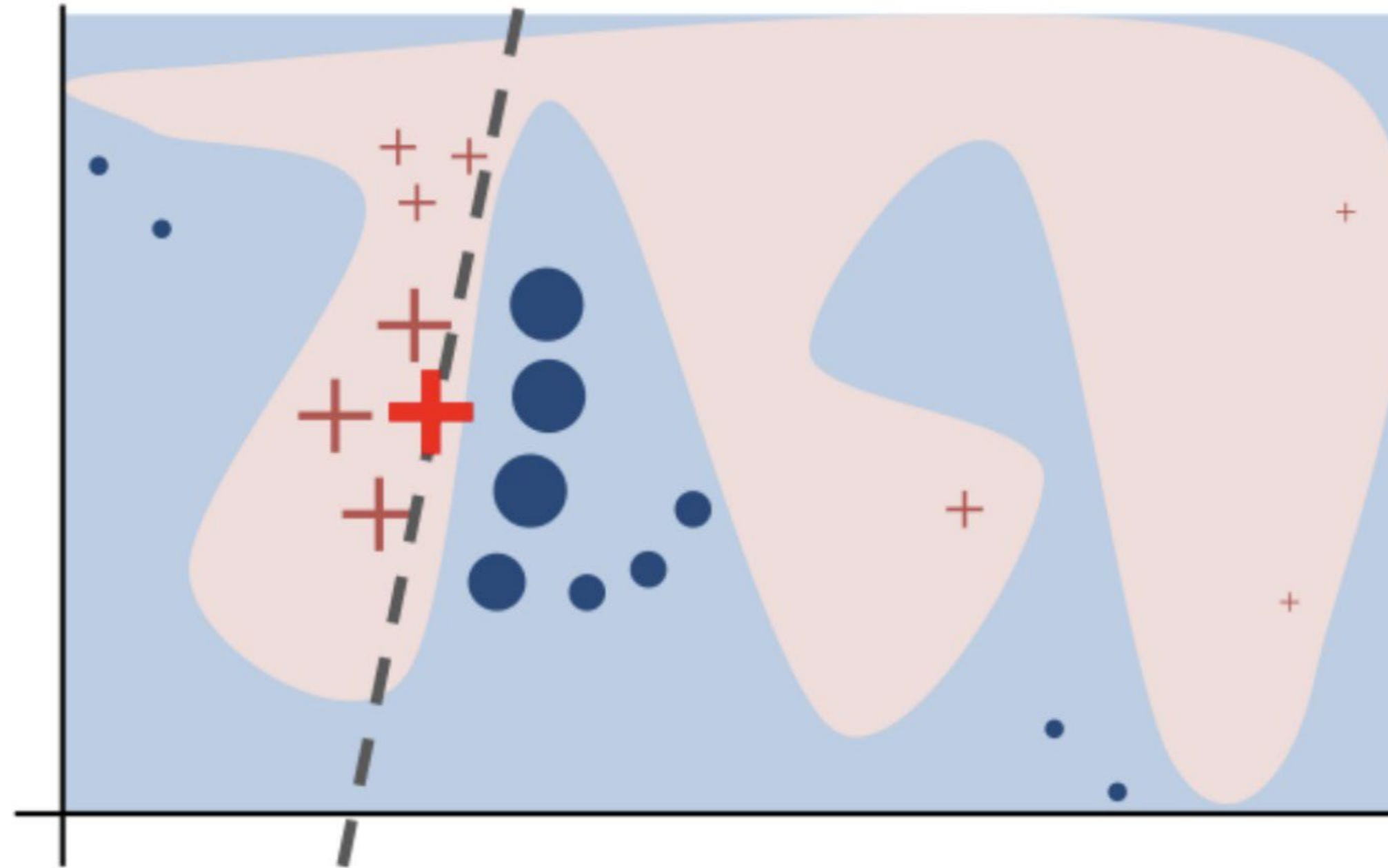
What qualities do we look for in a (**globally**) interpretable **model**?

- Algorithmic transparency e.g., guarantees about convergence or the shape of the error surface
- Decomposability: Are the different pieces of the model understandable on their own?
- Simulatability: Can a person hold

These concepts are a very tall order for current NLP models! So we'll focus on **local** explanations.

Global versus local explanations

To borrow a figure from Ribeiro et al. 2016:



- A global explanation describes the entire model across all its possible inputs.
- A local explanation describes only the parts of the model relevant for a particular instance's decision
 - E.g. which parts of the input were responsible for the model's prediction on this particular data point?
 - Or which training example(s) lead the model to make this decision.

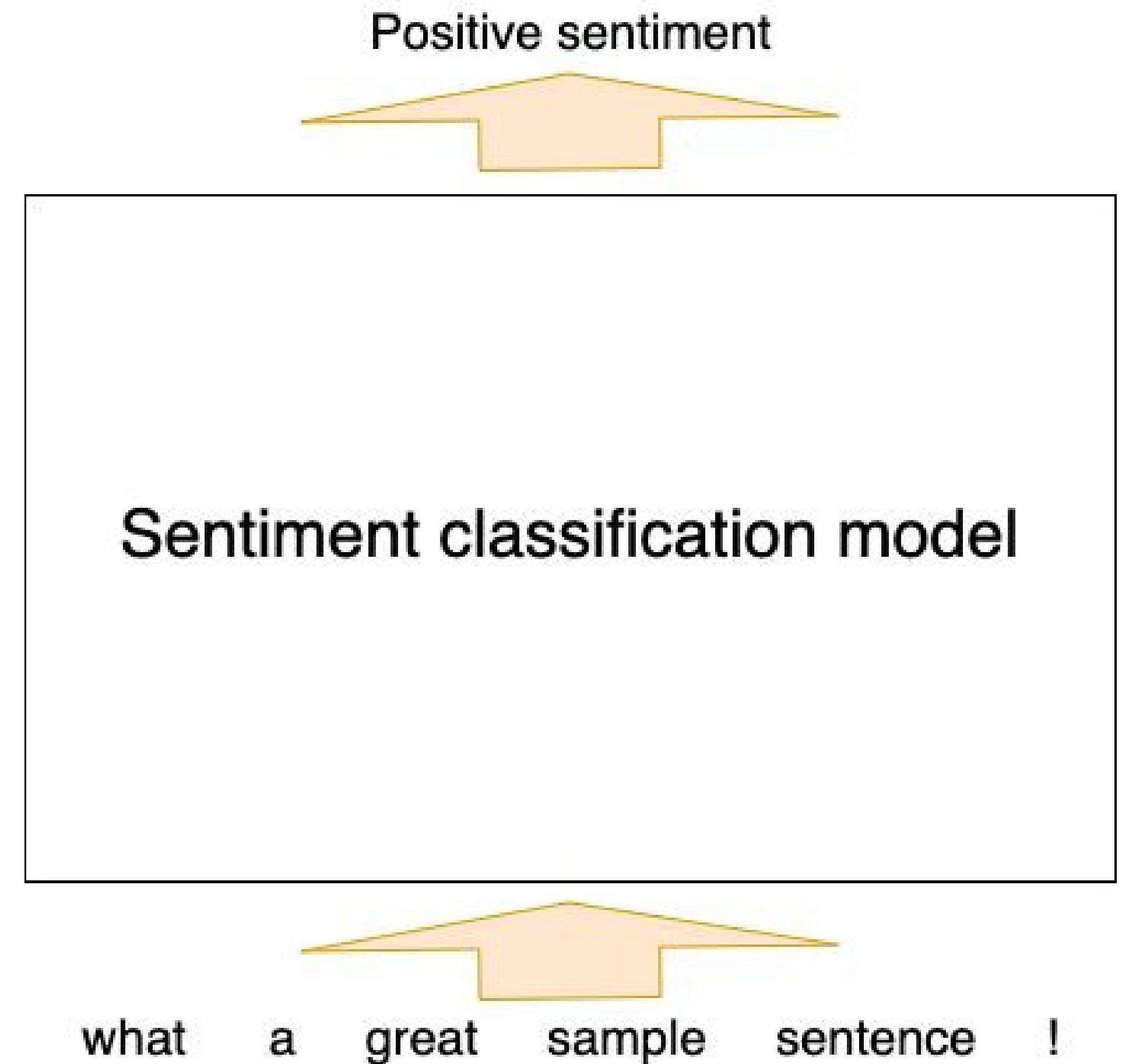
What choices are available to someone who hopes to interpret their eventual model?

- Restrict yourself to a class of model that's more readily interpretable (See: enduring popularity of linear models in applied-NLP settings)
- Apply a post-hoc, model-agnostic method for producing explanations
- Figure out how to interpret your model of choice

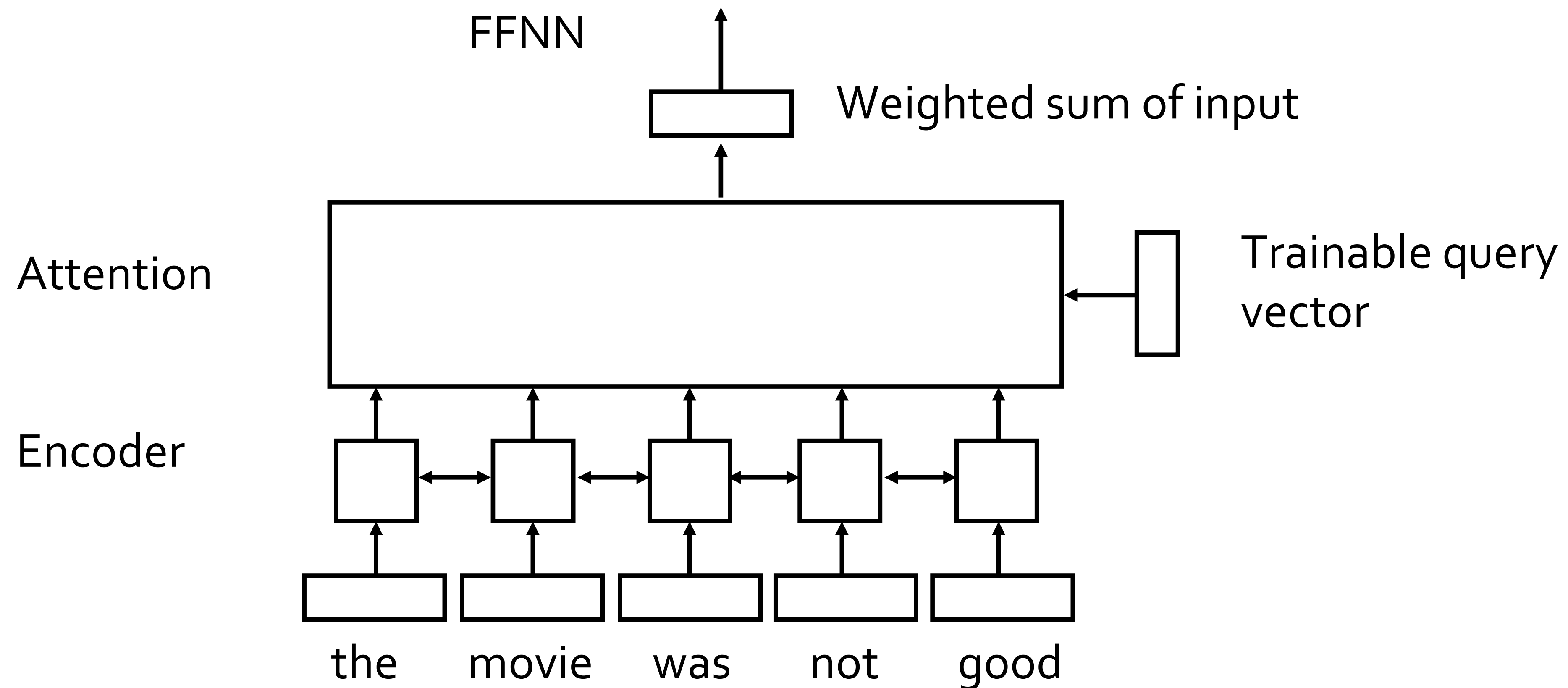
Post-hoc methods for interpreting models

What do we mean by “post-hoc”?

- Any method for getting us an explanation that doesn't make assumptions about the structure of the model.
- Key challenge for these methods: how do we get information about what caused the model to make its decision without access to the model's intermediate calculations?

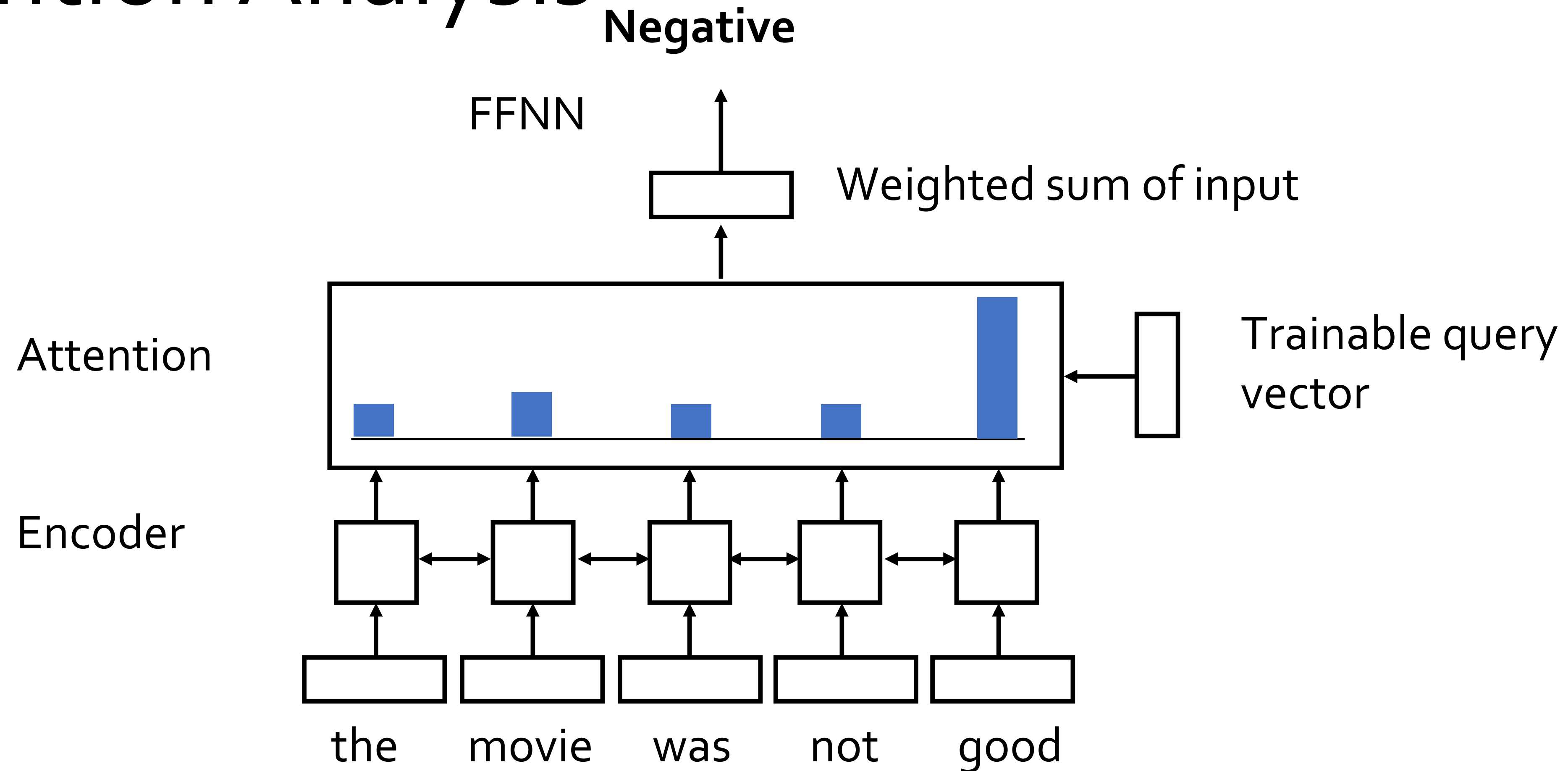


Sentiment Analysis with Attention



- ▶ Similar to a deep averaging (DAN) model, but (1) extra RNN layer; (2) attention layer instead of just a sum

Attention Analysis



- ▶ Attention places most mass on *good* — did the model ignore *not*?
- ▶ What if we removed *not* from the input?

Attention Analysis

after 15 minutes watching the movie i was asking myself what to do leave the theater sleep or try to keep watching the movie to see if there was anything worth i finally watched the movie what a waste of time maybe i am not a 5 years old kid anymore

original α

$$f(x|\alpha, \theta) = 0.01$$

after 15 minutes watching the movie i was asking myself what to do leave the theater sleep or try to keep watching the movie to see if there was anything worth i finally watched the movie what a waste of time maybe i am not a 5 years old kid anymore

adversarial $\tilde{\alpha}$

$$f(x|\tilde{\alpha}, \theta) = 0.01$$

- ▶ They show it is possible to modify attention while preserving the prediction probabilities
- ▶ Does this convince you that explanation is not helpful?

Is attention = explanation?

Is "Attention = Explanation"? Past, Present, and Future

Attention mechanisms have become a core component of neural models in Natural Language Processing over the past decade. These mechanisms not only deliver substantial performance improvements but also claim to offer insights into the models' inner workings. In this talk, we will highlight a series of contributions we have made that provided a critical perspective on the role of attention as a faithful explanation for model predictions, and sparked a larger conversation on the overarching goals of interpretability methods in NLP. We'll contrast our methodological approaches and findings to highlight that there is no one-size-fits-all answer to the question "Is attention explanation?". Finally, we'll explore the role of attention as an explanation mechanism in today's NLP landscape.

Relevant papers: [Jain & Wallace \(2019\)](#), [Wiegrefe & Pinter \(2019\)](#)



Sarah Wiegrefe

Postdoc at
AI2



Sarthak Jain

Applied Scientist at
AWS

[Big Picture Workshop](#)

Local Explanations

- ▶ An explanation could help us answer counterfactual questions: if the input were \mathbf{x}' instead of \mathbf{x} , what would the output be?

| | Model |
|---|-------|
| <i>that movie was not great , in fact it was terrible !</i> | — |
| <i>that movie was not _____ , in fact it was terrible !</i> | — |
| <i>that movie was _____ great , in fact it was _____ !</i> | + |

- ▶ Attention can't necessarily help us answer this!

Erasure Method

- ▶ Delete each word one by one and see how prediction prob changes

that movie was not great , in fact it was terrible !

— prob = 0.97

___ movie was not great , in fact it was terrible !

— prob = 0.97

that ___ was not great , in fact it was terrible !

— prob = 0.98

that movie ___not great, in fact it was terrible !

— prob = 0.97

that movie was ___great, in fact it was terrible !

— prob = 0.8

that movie was not ___, in fact it was terrible !

— prob = 0.99

Erasure Method

- ▶ Output: highlights of the input based on how strongly each word affects the output

that movie was not great , in fact it was terrible !

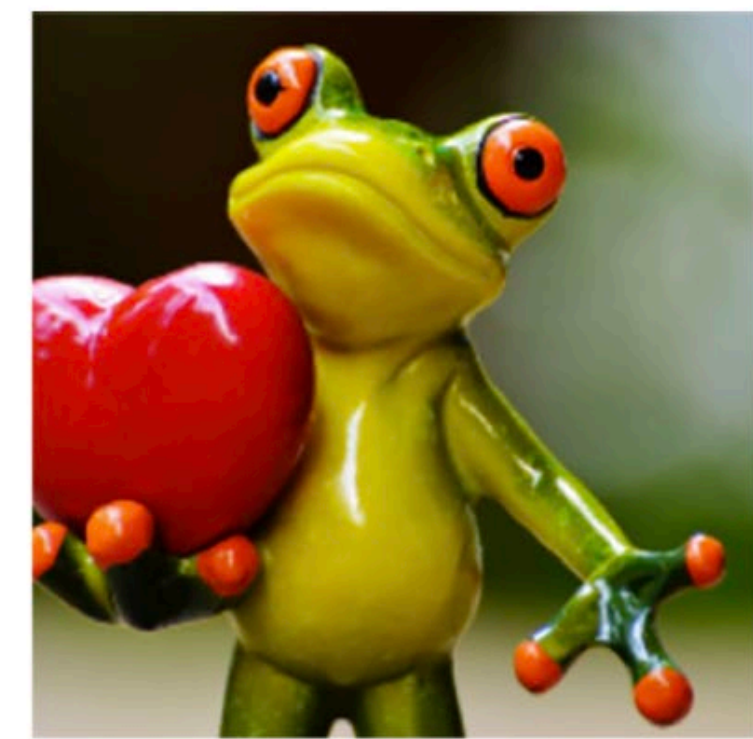
- ▶ **not** contributed to predicting the negative class (removing it made it less negative), **great** contributed to predicting the positive class (removing it made it more negative)
- ▶ Will this work well?
 - ▶ Inputs are now unnatural, model may behave in “weird” ways
 - ▶ Saturation: if there are two features that each contribute to negative predictions, removing each one individually may not do much

Demo Time!

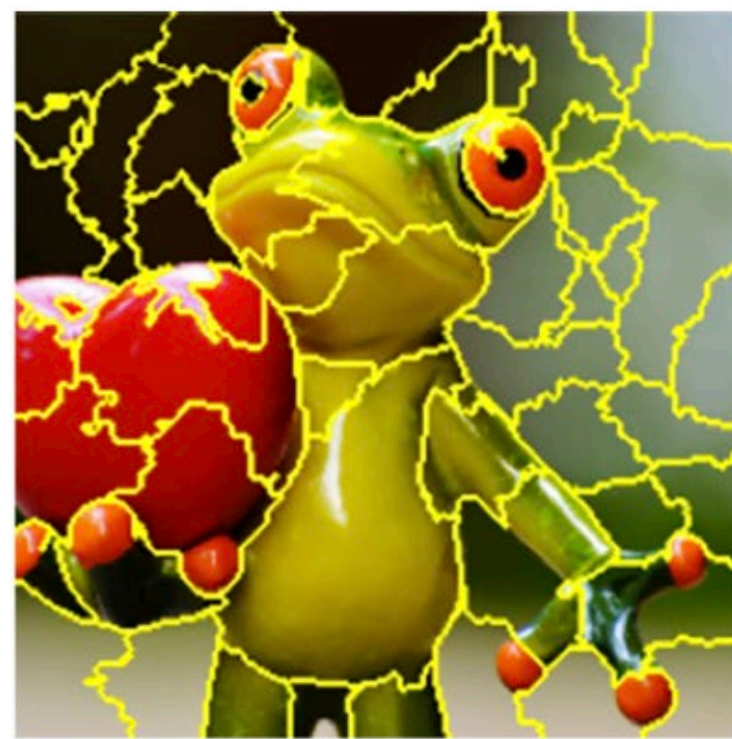
LIME

- ▶ Locally-interpretable, model-agnostic explanations (LIME)
- ▶ Similar to erasure method, but we're going to delete collections of things at once
- ▶ Can lead to more realistic input (although people often just delete words with it)
- ▶ More scalable to complex settings

LIME



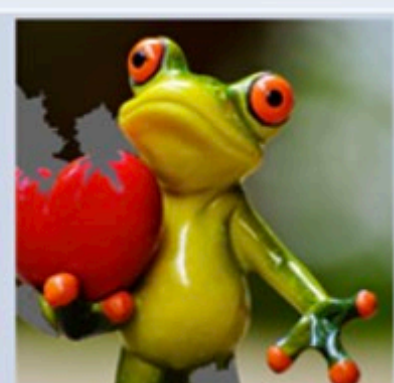


Original Image

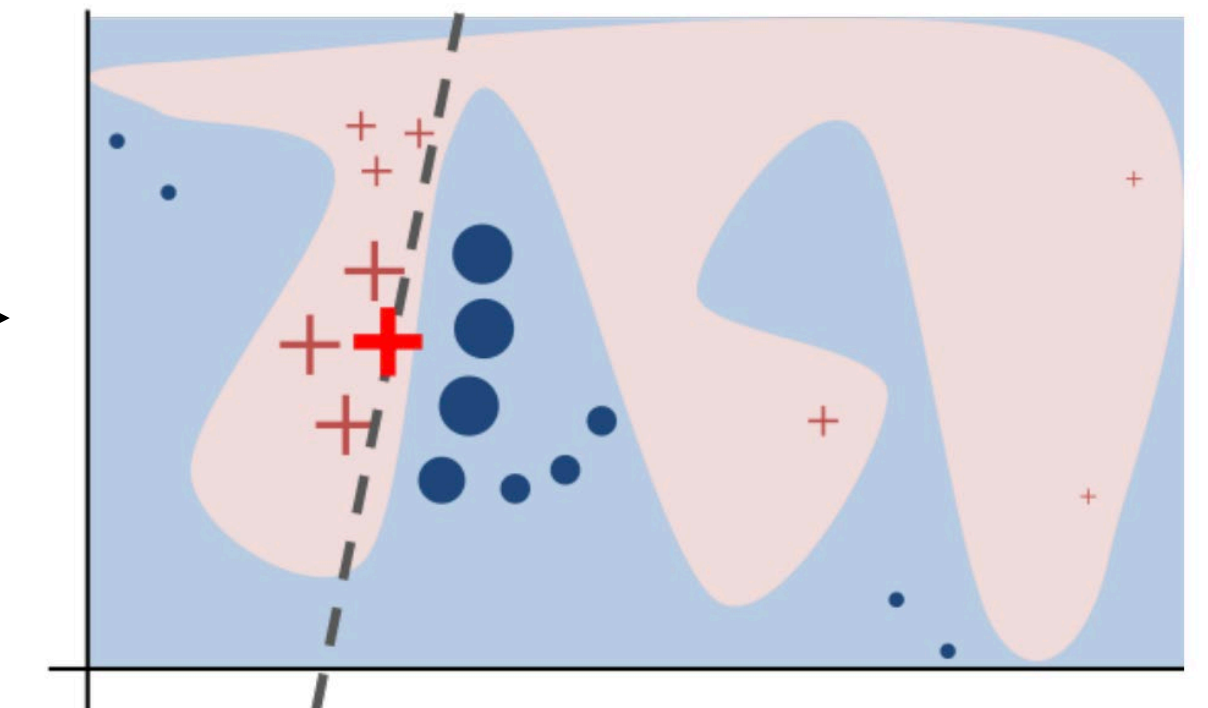


Interpretable Components

- ▶ Break input into components (for text: could use words, phrases, sentences, ...)

| Perturbed Instances | P(tree frog) |
|---|--------------------------------|
|  | <div><div></div></div> 0.85 |
|  | <div><div></div></div> 0.00001 |
|  | <div><div></div></div> 0.52 |

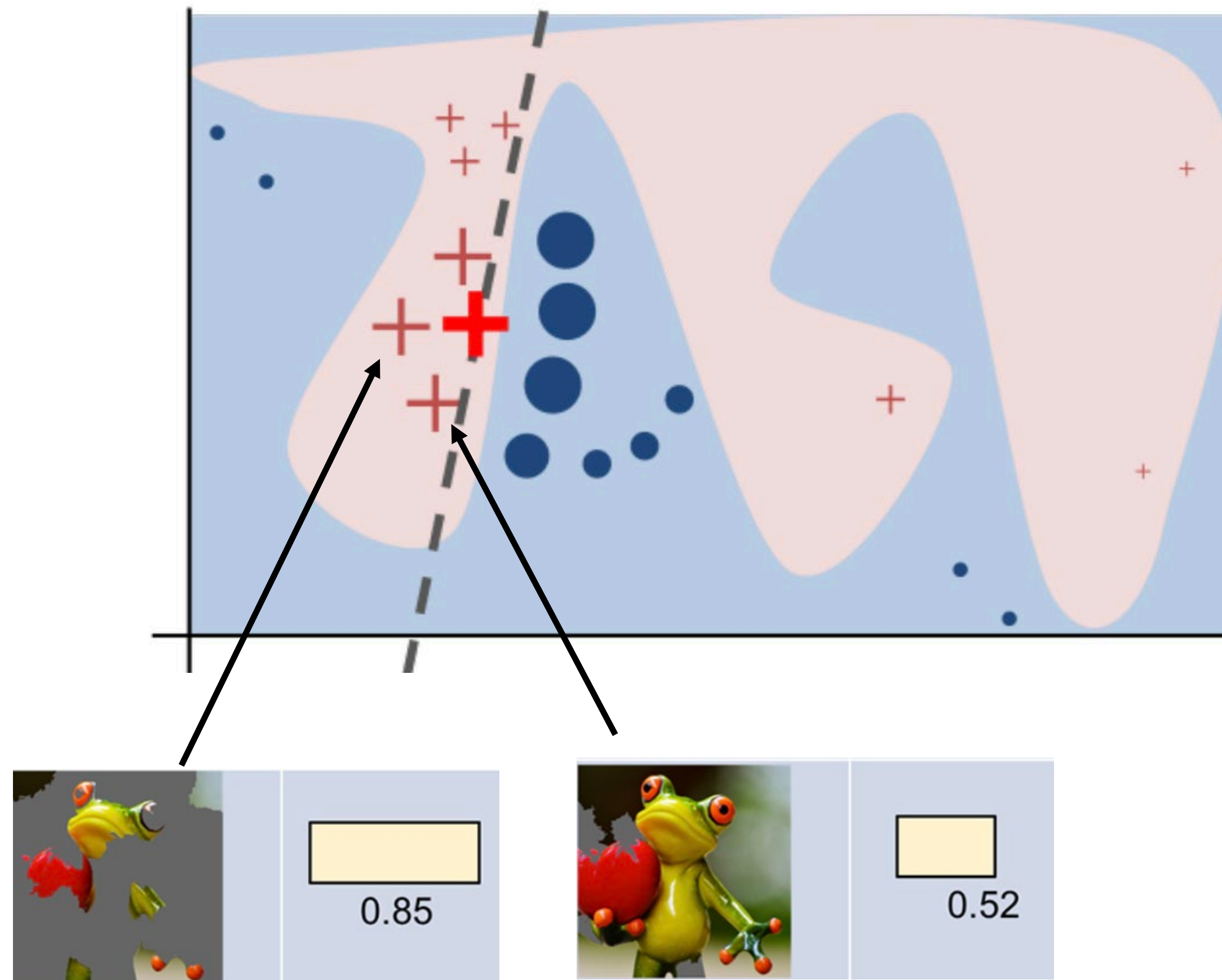
- ▶ Check predictions on subsets of those



- ▶ Now we have model predictions on perturbed examples

<https://www.oreilly.com/learning/introduction-to-local-interpretable-model-agnostic-explanations-lime>

LIME



- ▶ This is what the model is doing on perturbed examples of the input
- ▶ Now we train a classifier to predict **the model's behavior** based on **what subset of the input it sees**
- ▶ The weights of that classifier tell us which parts of the input are important

LIME

- ▶ This secondary classifier's **weights** now give us highlights on the input

The movie is mediocre, maybe even bad.

Negative 99.8%

The movie is mediocre, maybe even ~~bad~~.

Negative 98.0%

The movie is ~~mediocre~~, maybe even bad.

Negative 98.7%

The movie is ~~mediocre~~, maybe even ~~bad~~.

Positive 63.4%

The movie is ~~mediocre~~, ~~maybe~~ even ~~bad~~.

Positive 74.5%

The ~~movie~~ is mediocre, maybe even ~~bad~~.

Negative 97.9%

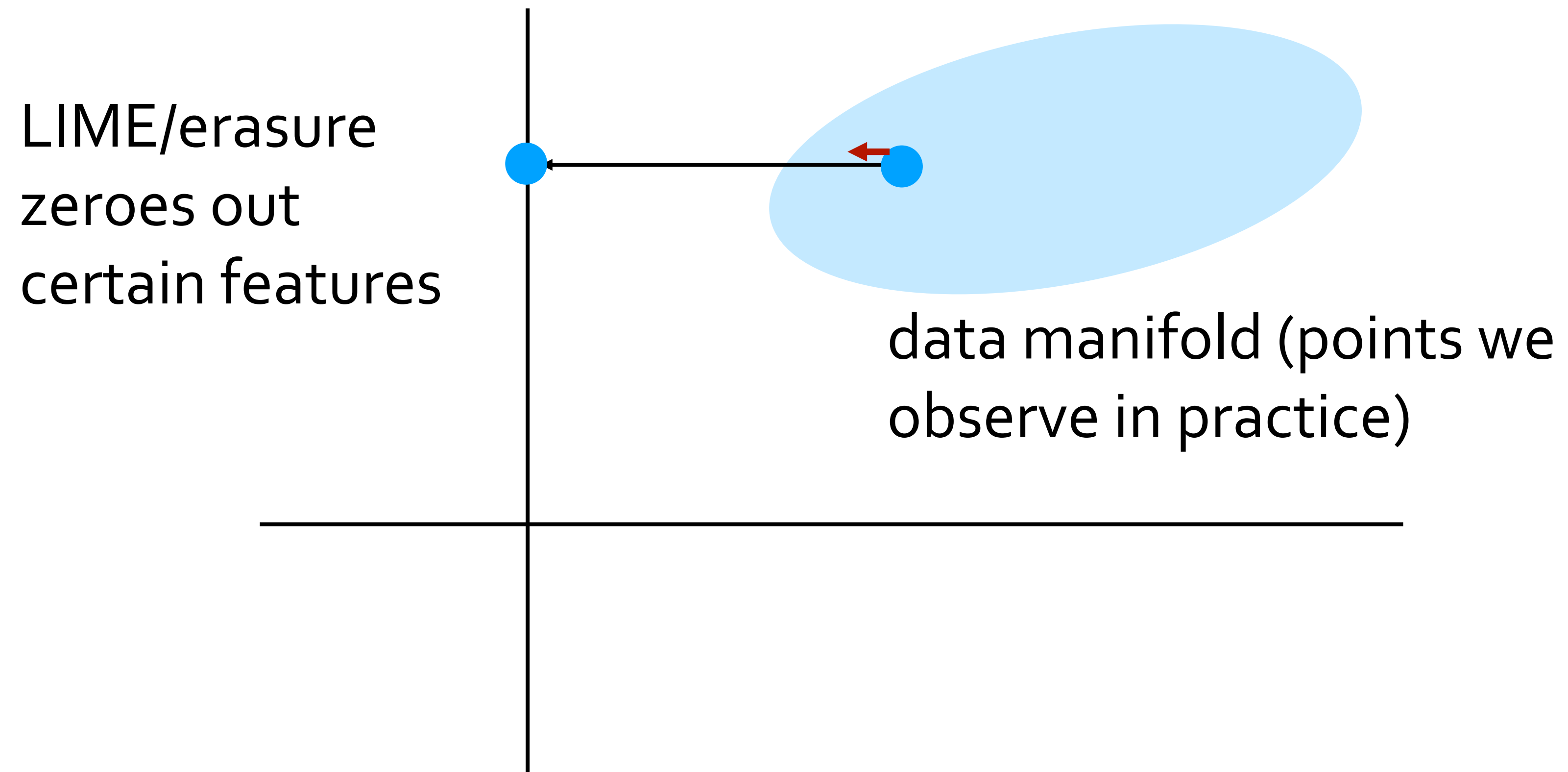
The movie is **mediocre**, maybe even **bad**.

Problems with LIME

- ▶ Lots of moving parts here: what perturbations to use? what model to train? etc.
- ▶ Expensive to call the model all these times
- ▶ Linear assumption about interactions may not be reliable

Problems with LIME

- ▶ Problem: fully removing pieces of the input may cause it to be very unnatural



- ▶ Alternative approach: look at what this perturbation does locally right around the data point using **gradients**

Demo Time!

Gradient-based Methods

Gradient-based Methods

score = weights * features
(or an NN, or whatever)

Learning a model

Compute derivative of score
with respect to weights: how
can changing weights improve
score of correct class?

Gradient-based Explanations

Compute derivative of score
with respect to ***features***: how
can changing ***features***
improve score of correct class?

Gradient-based Methods

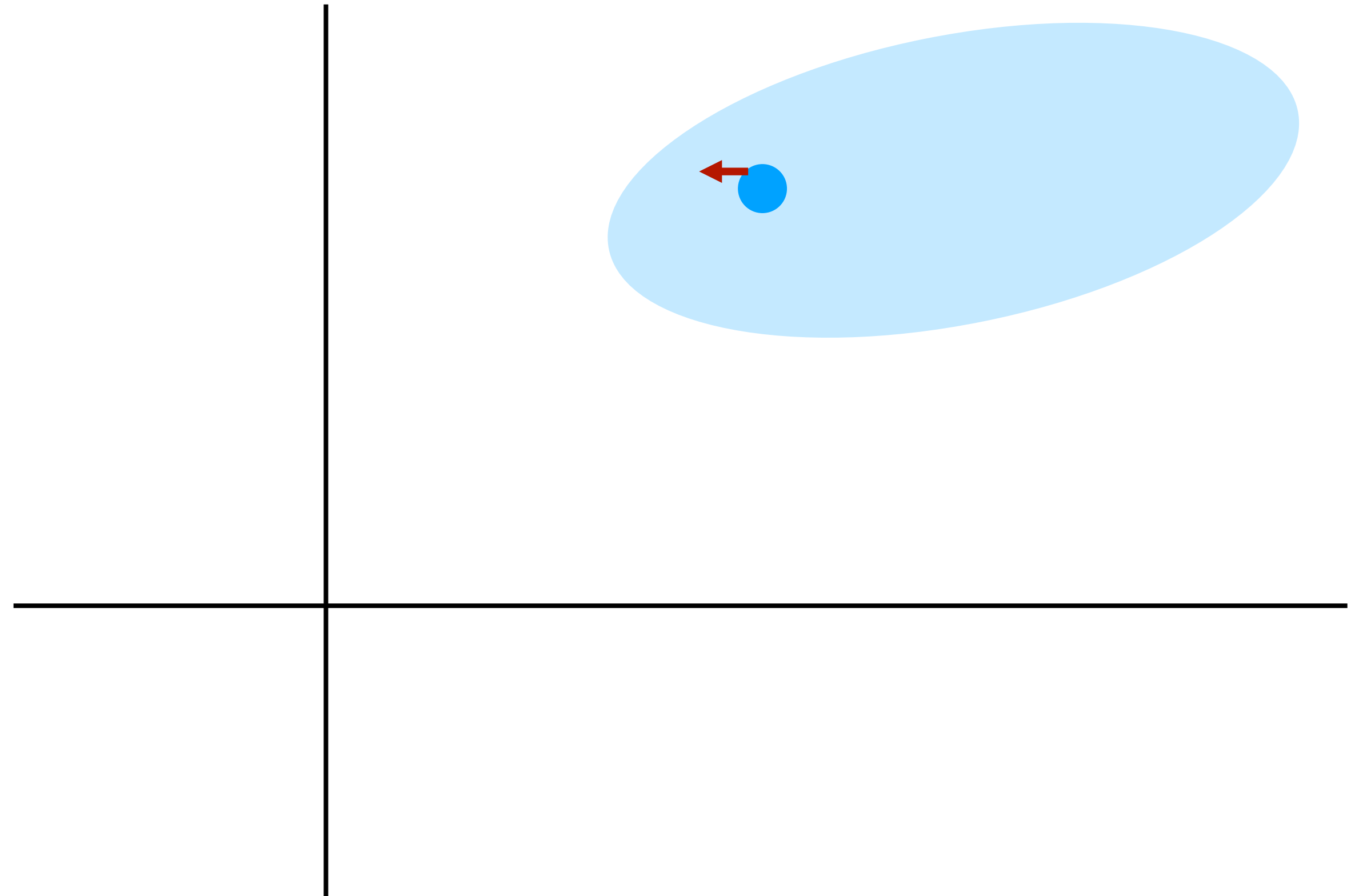
- ▶ Originally used for images

S_c = score of class c

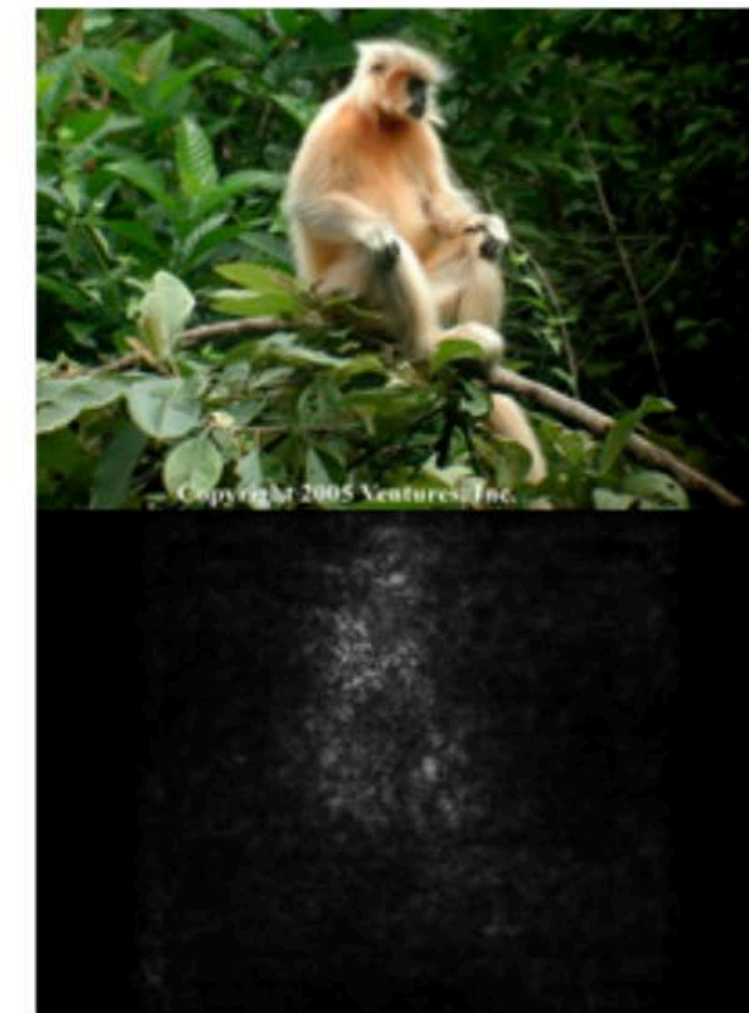
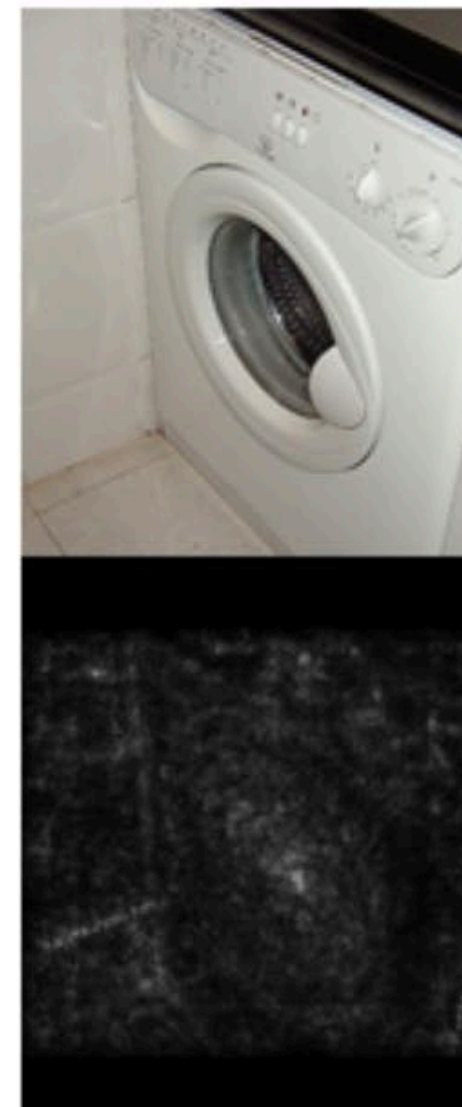
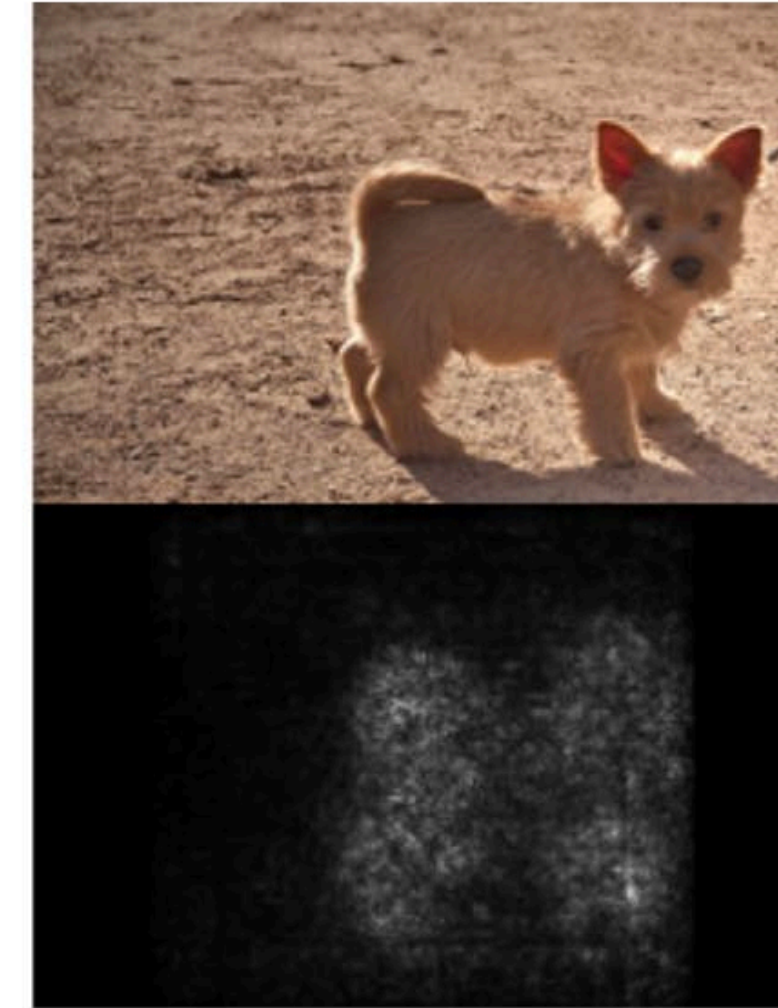
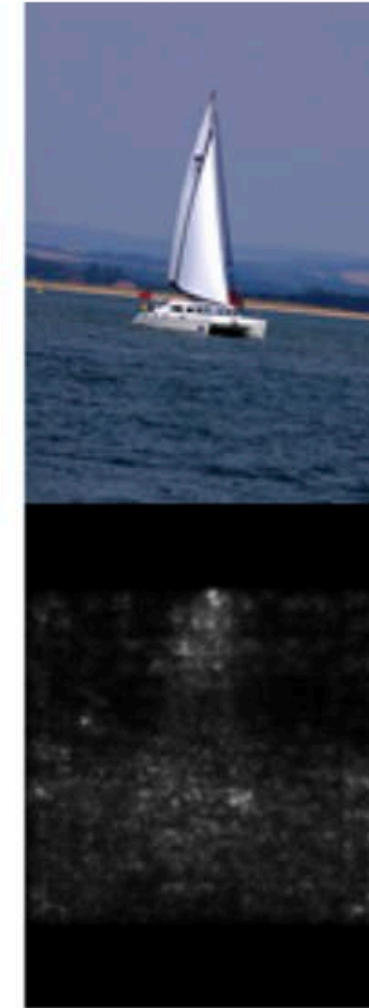
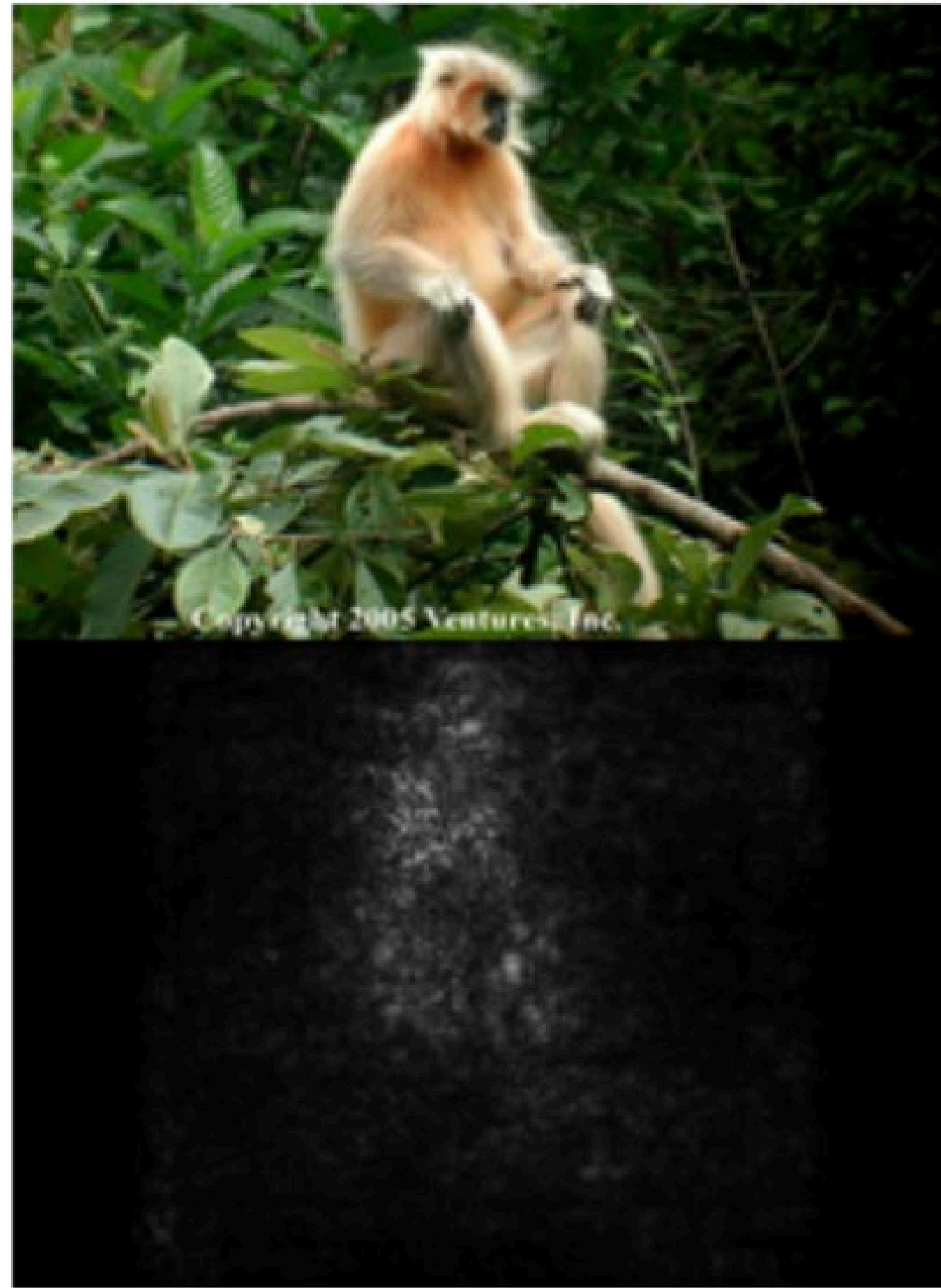
I_o = current image

$$w = \left. \frac{\partial S_c}{\partial I} \right|_{I_o}$$

- ▶ Higher gradient magnitude = small change in pixels leads to large change in prediction



Gradient-based Methods



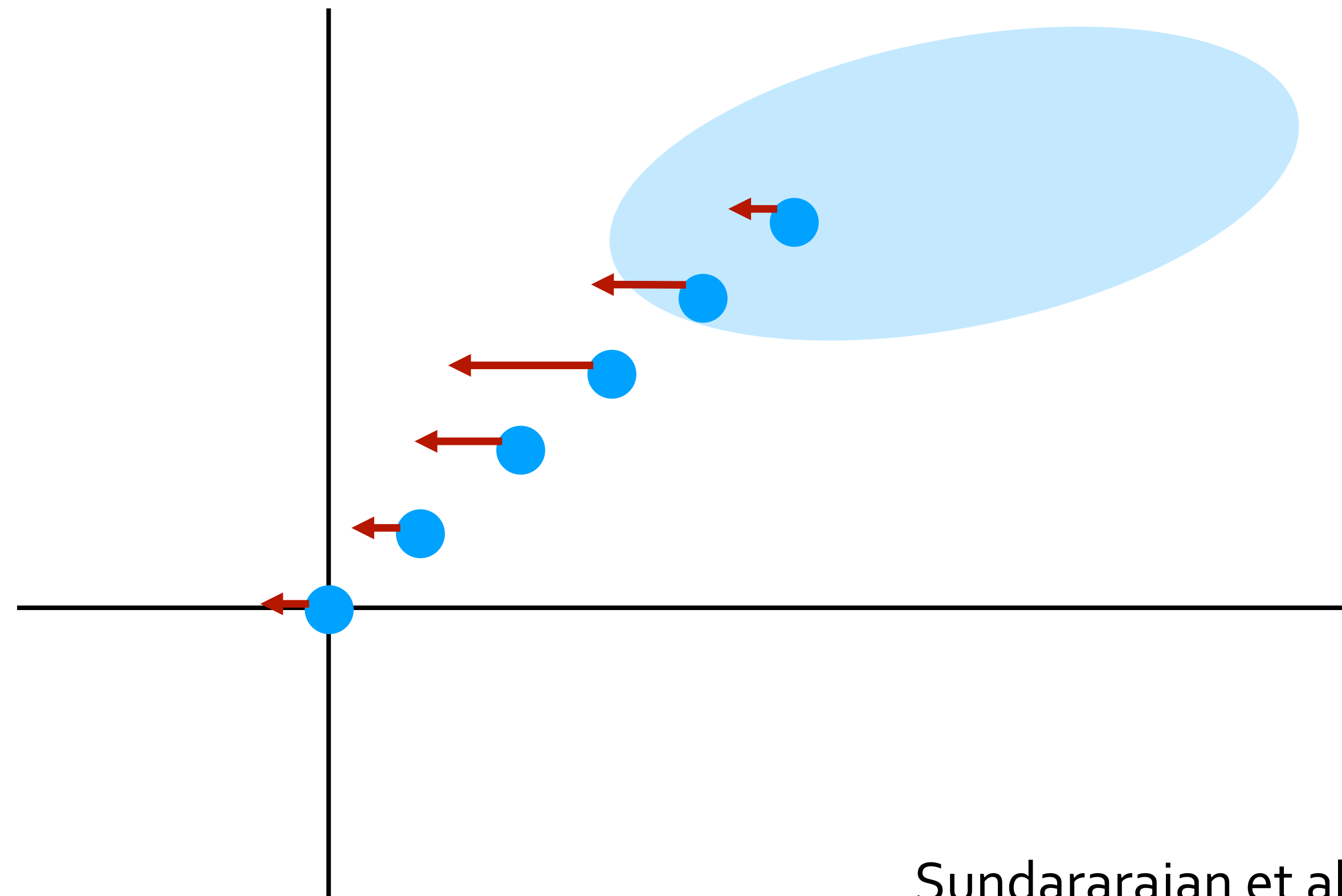
Simonyan et al. (2013)

Integrated Gradients

- Suppose you have prediction = A OR B for features A and B. Changing either feature doesn't change the prediction, but changing both would. Gradient-based method says neither is important

- Integrated gradients: compute gradients along a path from the origin to the current data point, aggregate these to learn feature importance

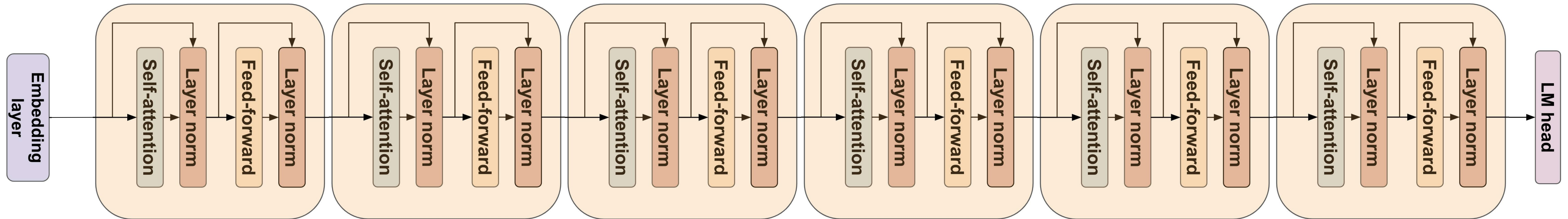
- Intermediate points can reveal new info about features



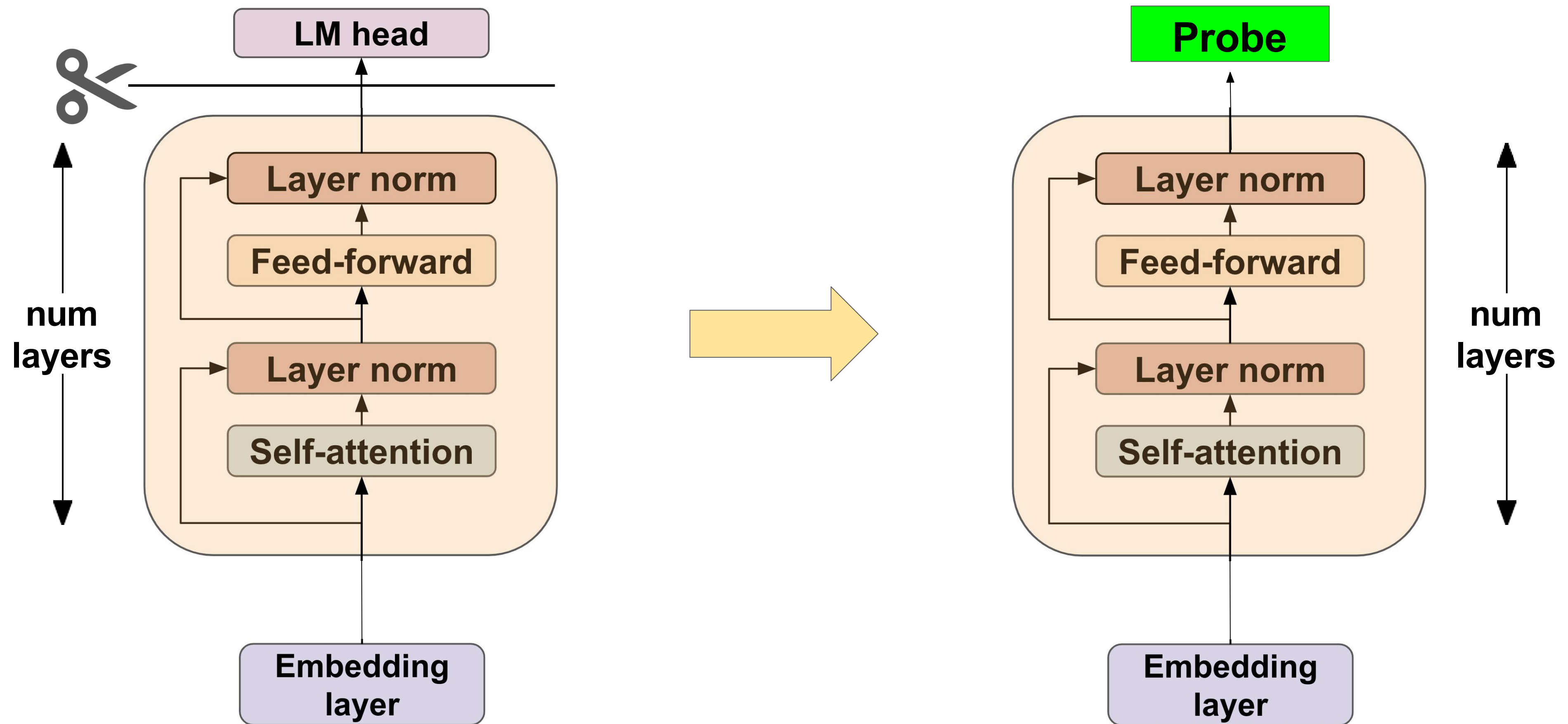
Demo Time!

PROBING

Transformer models are HUGE



How do we make sense of this huge model?

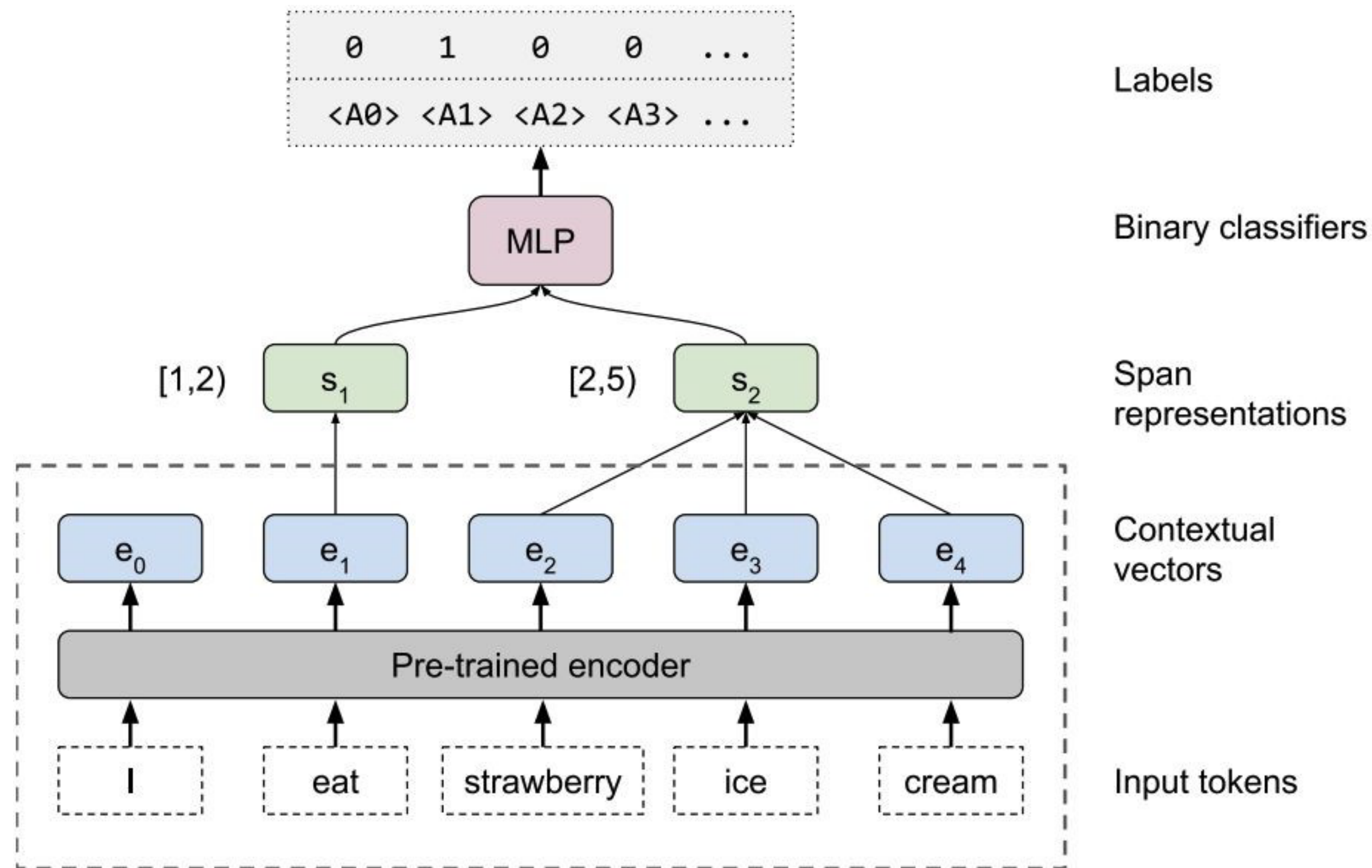


What is a Probe?

Definition: *A classifier that is specifically trained to predict some property from a pretrained model's representations.*

Edge Probing (Tenney et al. 2019)

- General method that works to probe different types of information



BERT rediscovers the NLP pipeline (Tenney et al. 2019)

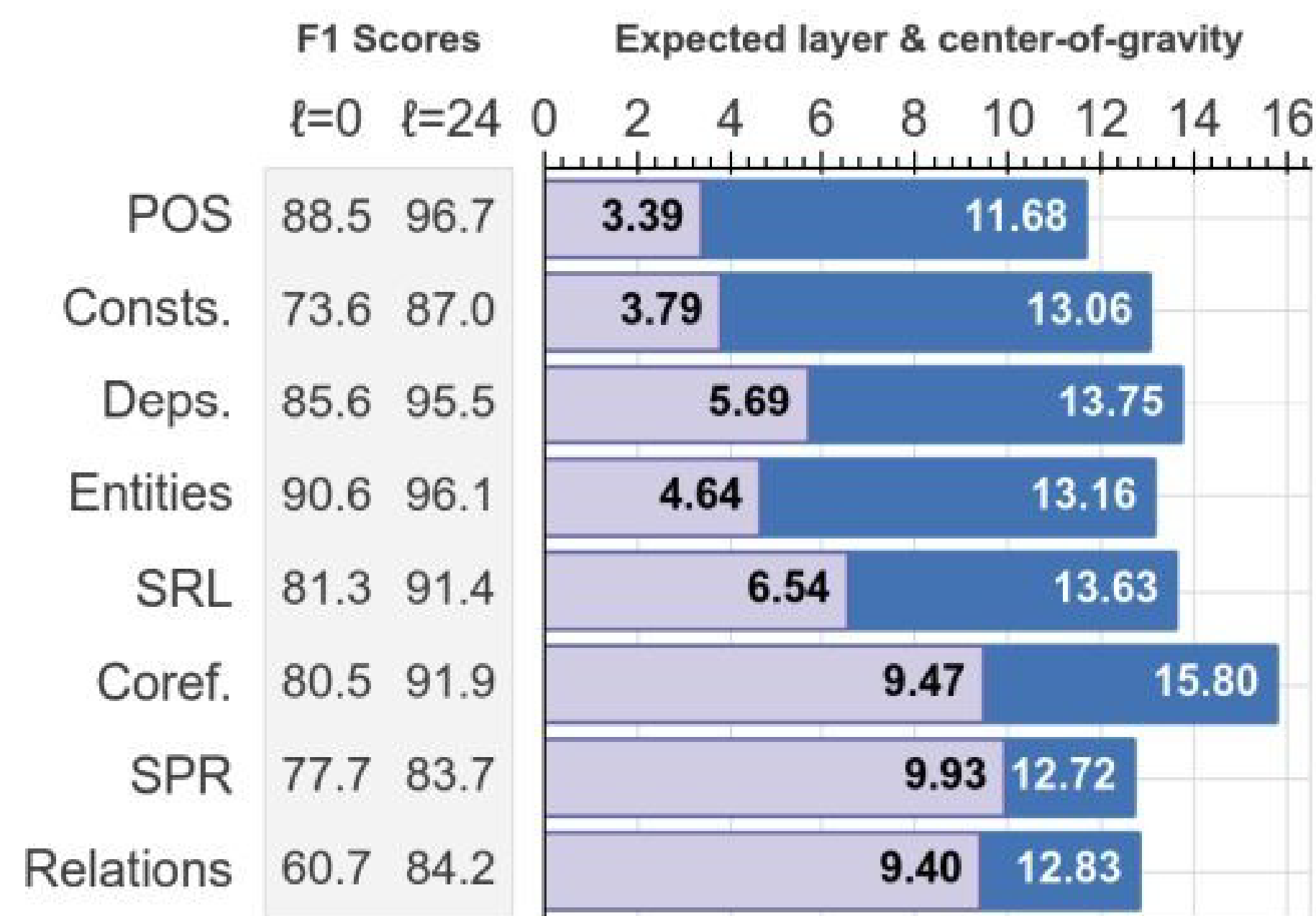


Figure 1: Summary statistics on BERT-large. Columns on left show F1 dev-set scores for the baseline ($P_{\tau}^{(0)}$) and full-model ($P_{\tau}^{(L)}$) probes. Dark (blue) are the mixing weight center of gravity (Eq. 2); light (purple) are the expected layer from the cumulative scores (Eq. 4).

POS - part of speech tagging (e.g. this word is a noun)

consts - constituent labeling (e.g. is this span a noun phrase)

deps - dependency labeling (e.g. is span_one the subject and span_two the object)

entities - named entity labeling (e.g. this word is a person)

SRL - semantic role labeling (what roles are the spans playing with each other: “Mary (pusher) pushed John (pushee)”

coref - coreference (do span_one and span_two refer to the same entity or event)

SPR - semantic proto-role (identifying attributes like awareness so is Mary aware that they are doing the pushing)

relations - relation classification (predicting the real-world relation between two spans given a set of these)

Issues with Probing (Belinkov et al. 2021)

- Probe works
 - Representation encodes information
 - Probe solved task by itself
- Probe doesn't work
 - Representation lacks the information
 - Representation encodes information, but probe is not the right function class
- We want to probe *tasks*, but require supervised data, so instead we probe *datasets*
- Probes designed this way are *correlative* not *causative*

Other Probing Works

Information-Theoretic Probing with Minimum Description Length

Elena Voita^{1,2}

Ivan Titov^{1,2}

Amnesic Probing: Behavioral Explanation with Amnesic Counterfactuals

Yanai Elazar^{1,2} Shauli Ravfogel^{1,2} Alon Jacovi¹ Yoav Goldberg^{1,2}


Low-Complexity Probing via Finding Subnetworks

Steven Cao^{1,2}

Victor Sanh²

Alexander M. Rush²

Pareto Probing: Trading Off Accuracy for Complexity

Tiago Pimentel^{*},  Naomi Saphra^{*},  Adina Williams  Ryan Cotterell  

Evaluating Explanations

Faithfulness vs. Plausibility

- ▶ Suppose our model is a bag-of-words model with the following:

the = -1, movie = -1, good = +3, bad = 0

the movie was good prediction score=+1

the movie was bad prediction score=-2

- ▶ Suppose explanation returned by LIME is:

the movie was good

the movie was bad

- ▶ Is this a "correct" explanation?

Faithfulness vs. Plausibility

- ▶ *Plausible* explanation: matches what a human would do

the movie was **good**

the movie was **bad**

- ▶ Maybe useful to explain a task to a human, but it's not what the model is really doing!

- ▶ *Faithful* explanation: actually reflects the behavior of the model

the movie was **good**

the movie was **bad**

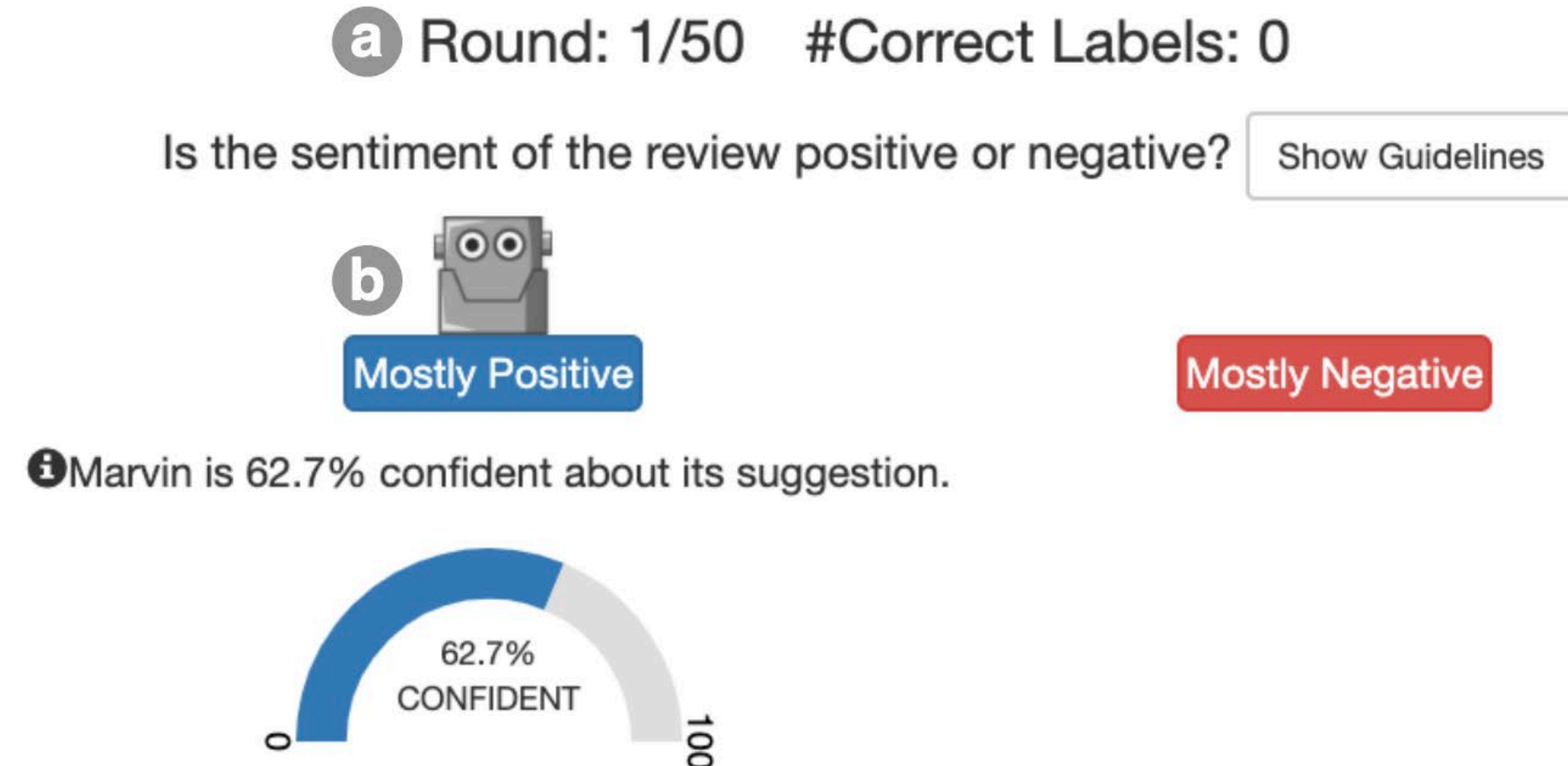
- ▶ We usually prefer faithful explanations; non-faithful explanations are actually deceiving us about what our models are doing!
- ▶ Rudin: *Stop Explaining Black Box Models for High-Stakes Decisions and Use Interpretable Models Instead*

Evaluating Explanations

- ▶ Nguyen (2018): delete words from the input and see how quickly the model flips its prediction?
 - ▶ Downside: not a “real” use case
- ▶ Hase and Bansal (2020): counterfactual simulatability: user should be able to predict what the model would do in another situation
 - ▶ Hard to evaluate

Evaluating Explanations

I, like others **was very excited to read this book.** I thought it would show another side to how the Tate family dealt with the murder of their daughter Sharon. I didn't have to read much to realize however that the book is was not going to be what I expected. It is full of added dialog and assumptions. It makes it hard to tell where the truth ends and the embellishments begin. It reads more like fan fiction than a true account of this family's tragedy. I did enjoy looking at the early pictures of Sharon that I had never seen before but they were **hardly worth the price of the book.**



- ▶ Human is trying to label the sentiment. The AI provides its prediction to try to help. Does the human-AI team beat human/AI on their own?
- ▶ AI provides both an explanation for its prediction (blue) and also a possible counterargument (red)

What to Expect from Explanations?

Ye et al. (2021)

- ▶ What do we really want from explanations?
 - ▶ Explanations should describe model behavior with respect to counterfactuals (Miller, 2019; Jacovi and Goldberg, 2021)

The movie is not that bad.

The movie is not ____ ____.



- ▶ What about **realistic counterfactuals**? Since dropping tokens isn't always meaningful

The movie is not actually bad.

- ▶ We are going to evaluate explanations based on whether they can tell us useful things about model behavior

A Multi-hop QA Example

Ye et al. (2021)

- ▶ We formulate a hypothesis about the model's behavior, and test it using counterfactuals

Base Example

Are Super High Me and All in This Tea both documentaries?

Super High Me is a 2008 **documentary** film about smoking.
All in This Tea is a 2007 **documentary** film.

YES

Token-Level Explanation

<s> Are Super High Me and All in This Tea both documentaries ?
</s> Super High Me is a 2008 **documentary** film about
smoking . All in This Tea is a 2007 **documentary** film . </s>

Expected Behavior

The hypothesis is true.

Hypothesis



The QA model is looking at the
two **documentary** tokens

Realistic Counterfactuals

Super High Me is a 2008 **romance** film about smoking.
All in This Tea is a 2007 **documentary** film.

YES

Super High Me is a 2008 **documentary** film about smoking.
All in This Tea is a 2007 **romance** film.

YES

Super High Me is a 2008 **romance** film about smoking.
All in This Tea is a 2007 **romance** film.

YES

Actual Behavior

Mismatch

The hypothesis is not true.
Model always predict YES.

Takeaways

- ▶ Lots of ongoing research:
 - ▶ How do we interpret explanations?
 - ▶ How do *users* interpret our explanations?
 - ▶ How should *automated systems* make use of explanations?
- ▶ Emerging consensus: there is no one-size-fits-all solution. There are many formats of explanation that all have their uses — choice may be application specific
- ▶ This research has taken a bit of a back seat during the current era of LLMs.

Takeaways

- ▶ Many other ways to do explanation:
 - ▶ Diagnostic test sets (“unit tests” for models)
 - ▶ Building models that are explicitly interpretable (there is some work on interpretable architectures)
Interpretable Architectures and Algorithms for Natural Language Processing? (Yadav, Rohan Kumar 2022)

What is Mechanistic Interpretability?

Definition: *The study of reverse engineering parametric models (often neural networks) from their learned weights into more human-interpretable algorithmic units.*

Notable Work

- Analysis of 1 and 2-layer MLPs and Transformers to find circuits (Olah et al. 2021)
- Induction Heads (Olsson et al. 2022)
- Neuron Polysemanticity (Elhage et al. 2021; 2022)

Anthropic

A Mathematical Framework for Transformer Circuits

Introduced a formalism to
the analysis of
Transformers

AUTHORS

Nelson Elhage^{*†} Neel Nanda^{*} Catherine Olsson^{*} Tom Henighan[‡] Nicholas Joseph[‡]
Ben Mann[‡] Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma,
Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones,
Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark,
Jared Kaplan, Sam McCandlish, Chris Olah[‡]

AFFILIATION

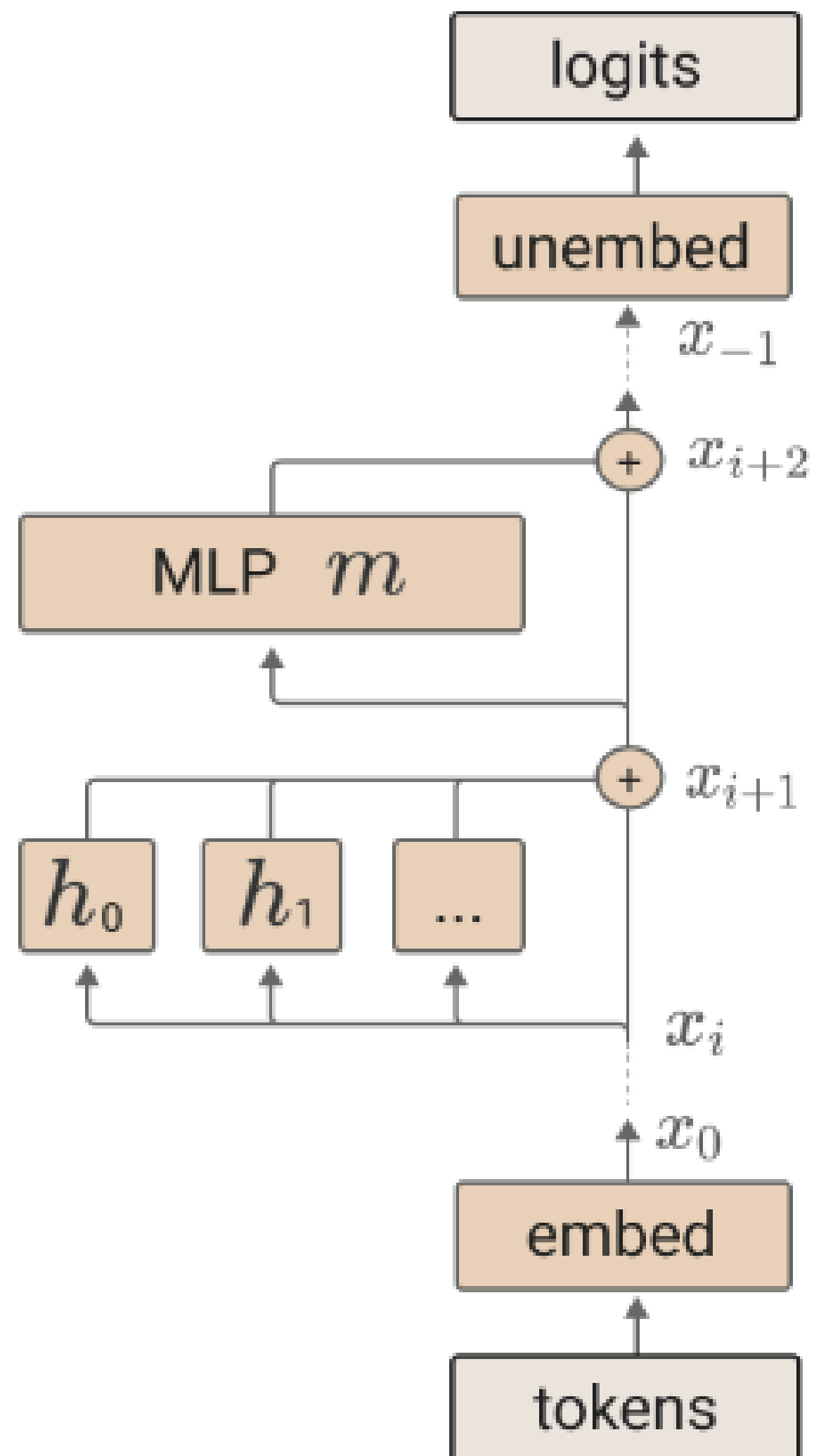
Anthropic

PUBLISHED

Dec 22, 2021

^{*} Core Research Contributor; [†] Core Infrastructure Contributor; [‡] Correspondence to colah@anthropic.com;
Author contributions statement below.

Residual Streams



The final logits are produced by applying the unembedding.

$$T(t) = W_U x_{-1}$$

An MLP layer, m , is run and added to the residual stream.

$$x_{i+2} = x_{i+1} + m(x_{i+1})$$

Each attention head, h , is run and added to the residual stream.

$$x_{i+1} = x_i + \sum_{h \in H_i} h(x_i)$$

Token embedding.

$$x_0 = W_E t$$



One
residual
block

Model Editing

Target: *A concept or specific fact needs to be changed in the model*

Approach: *Changing the weights of the model to edit the model's belief of that fact/concept?*

ROME (Meng et al. 2022)

- Use causal tracing to isolate the causal effect of individual hidden states when processing a fact
- Introduce rank-one model editing (ROME) to edit the model

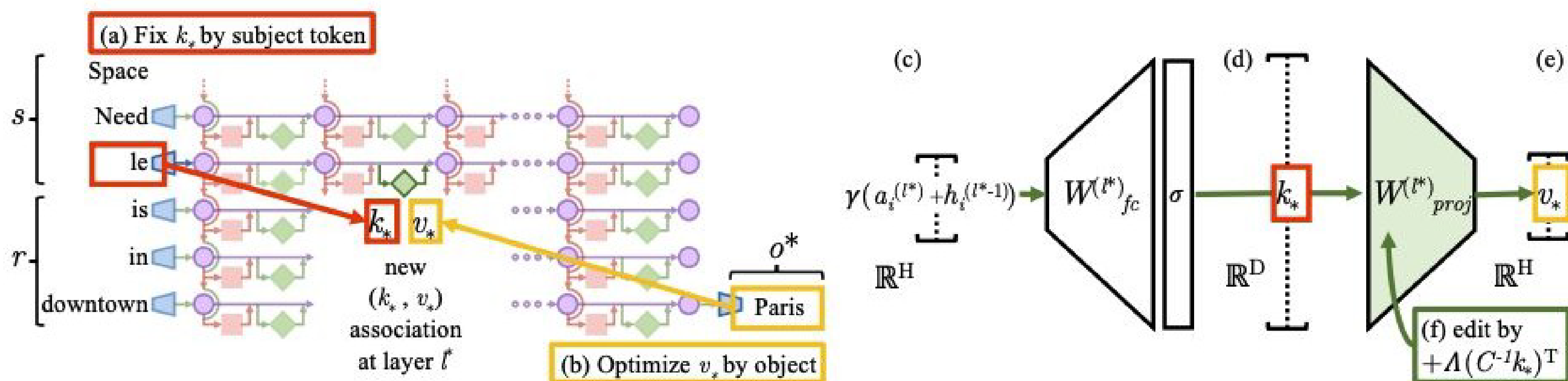


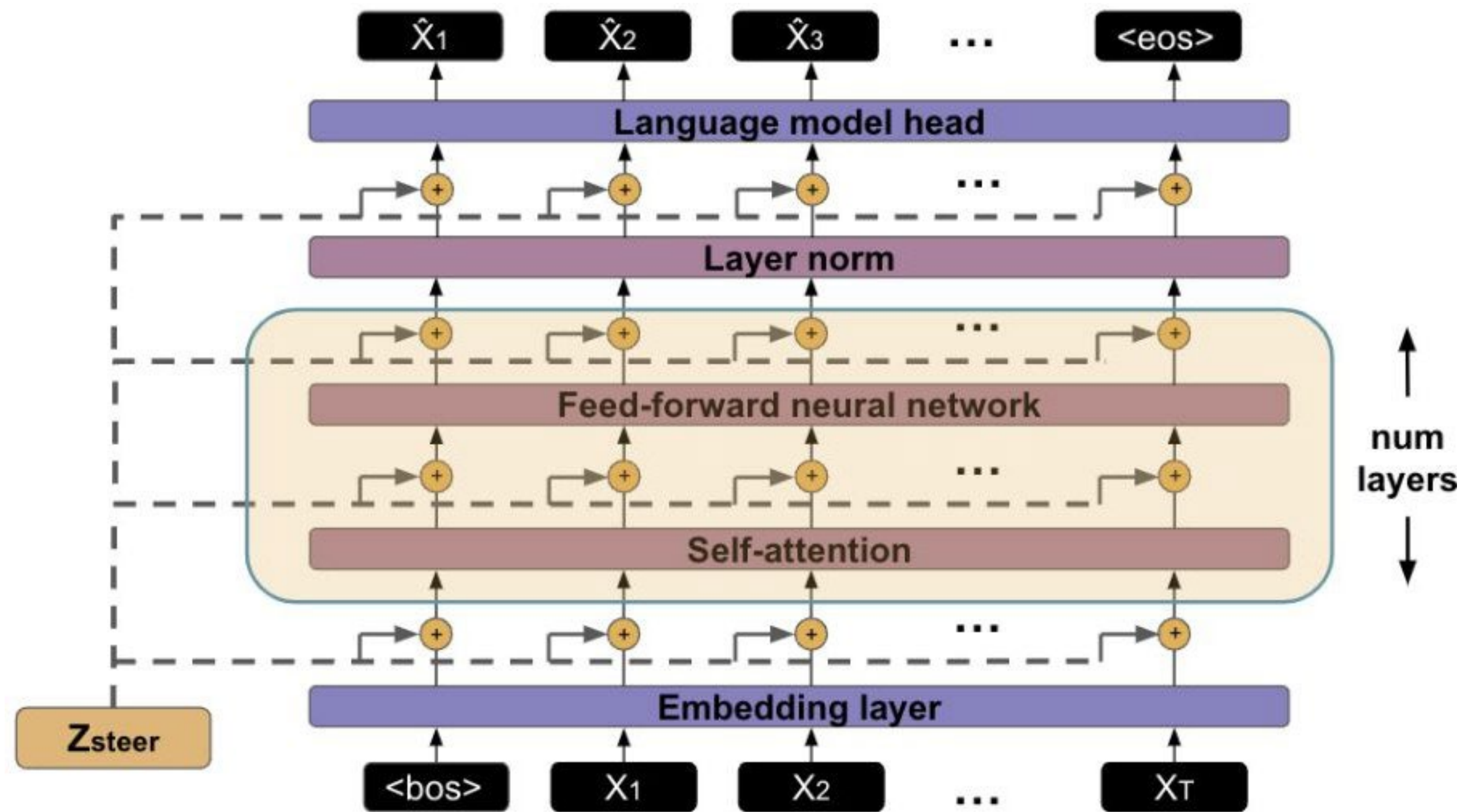
Figure 4: **Editing one MLP layer with ROME.** To associate *Space Needle* with *Paris*, the ROME method inserts a new (k_*, v_*) association into layer l^* , where (a) key k_* is determined by the subject and (b) value v_* is optimized to select the object. (c) Hidden state at layer l^* and token i is expanded to produce (d) the key vector k_* for the subject. (e) To write new value vector v_* into the layer, (f) we calculate a rank-one update $\Lambda(C^{-1}k_*)^T$ to cause $\hat{W}_{proj}^{(l)} k_* = v_*$ while minimizing interference with other memories stored in the layer.

Steering Vectors (Subramani et al. 2019; 2020; 2022)

Steering Vectors: a fixed-length vector that steers a language model to generate a specific sequence exactly when added to the hidden states of a model at a specific location.

This is our stick that we're poking a language model with.

Extracting steering vectors



ALGORITHM 1: Extracting z_{steer} for a sentence

Input : x – target sentence

M – pretrained language model

θ – pretrained language model weights

I_L – injection location

I_T – injection timestep

d – dimension of z_{steer}

Output : z_{steer} – extracted candidate steering vector

```

1  $z_{\text{steer}} \sim \text{xavier\_normal}(d)$ 
2 for  $i \leftarrow [1, 2, \dots, N]$  do
3    $\text{logits} = M_{\theta}.\text{forward}(x, z_{\text{steer}}, I_L, I_T)$ 
4    $\mathcal{L} = \text{XENT}(\text{logits}, x)$ 
5    $\mathcal{L}.\text{backward}()$ 
6    $z_{\text{steer}} = z_{\text{steer}} + lr * \frac{\partial \mathcal{L}}{\partial z_{\text{steer}}}$ 
7 end
8 return  $z_{\text{steer}}$ 

```

Steering vector results

| Steering vectors | |
|-------------------------|--------------------------------|
| Positive Input | the taste is excellent! |
| +1.0 * $z_{tonegative}$ | the taste is excellent! |
| +2.0 * $z_{tonegative}$ | the taste is unpleasant. |
| Negative Input | the desserts were very bland. |
| +1.0 * $z_{topositive}$ | the desserts were very bland . |
| +2.0 * $z_{topositive}$ | the desserts were very tasty. |

Inference-time Interventions (Li et al. 2023)

- Use linear probes to find attention heads that correspond to the desired attribute
- Shift attention head activations during inference along directions determined by these probes

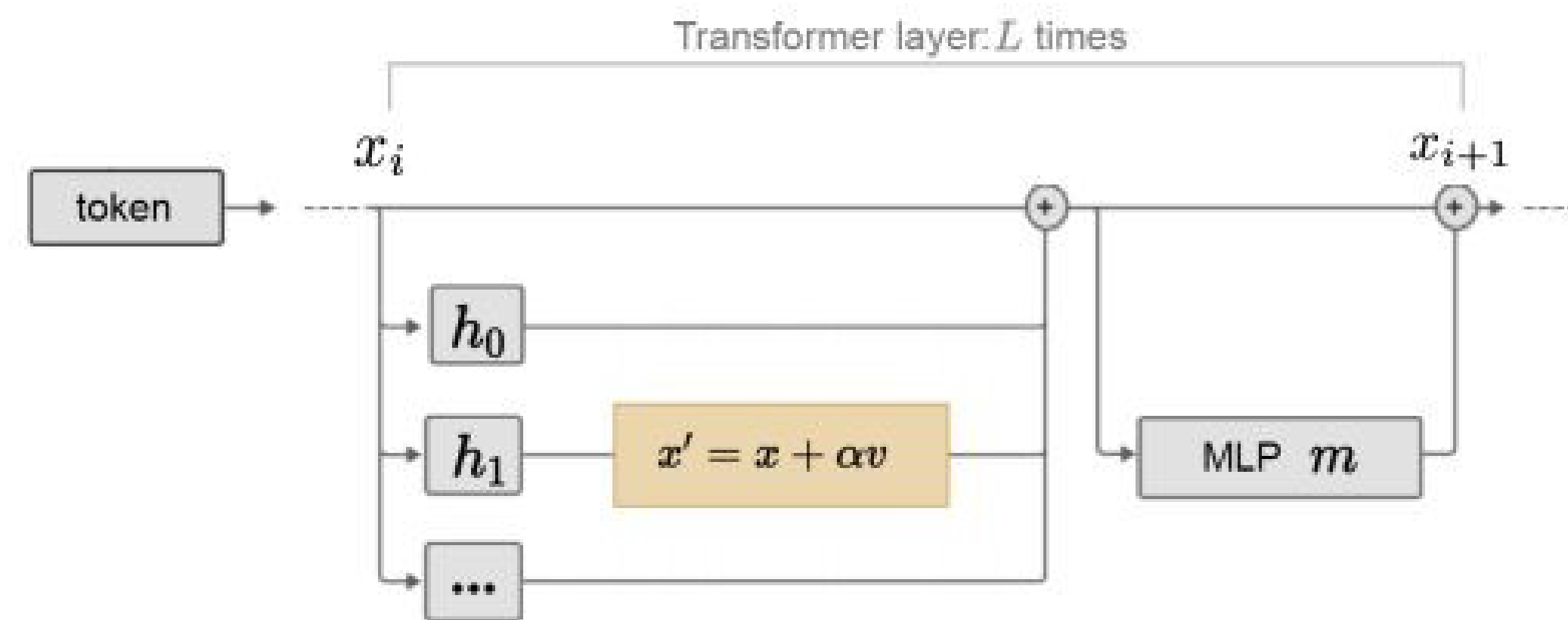
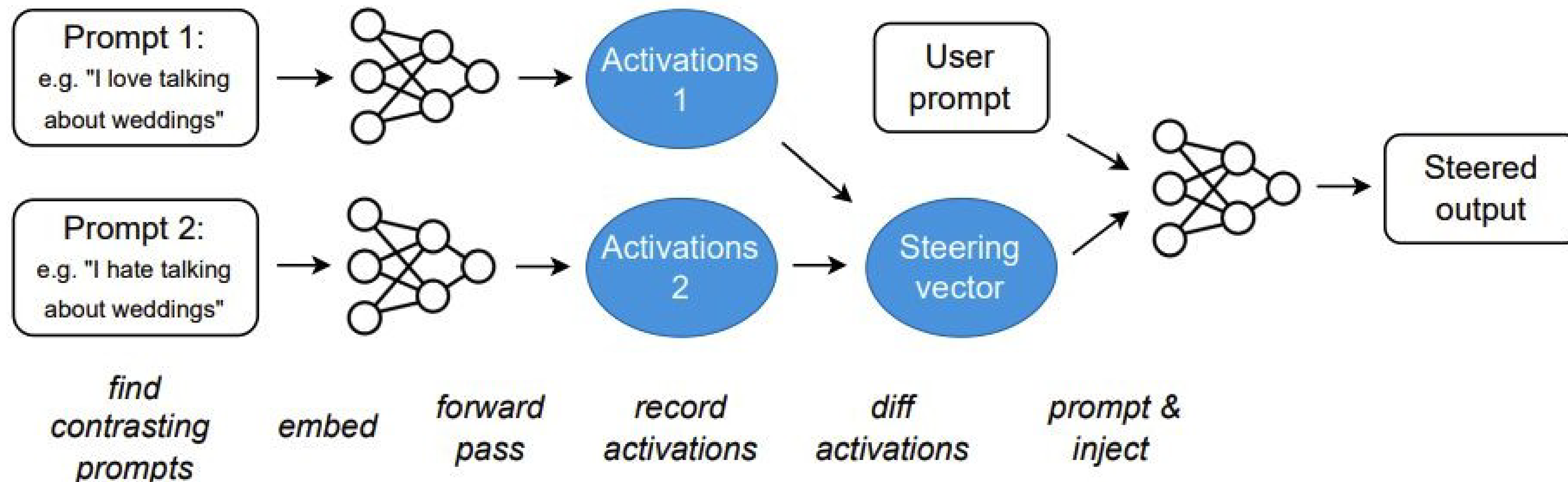


Figure 3: A sketch of the computation on the last token of a transformer with inference-time intervention (ITI) highlighted.

More activation manipulation

- Contrastive steering vectors (Turner et al. 2023; Rinsky et al. 2023)

Figure 1: Schematic of the Activation Addition (**ActAdd**) method. \bigcirc = natural language text; \bullet = vectors of activations just before a specified layer. In this example, the output is heavily biased towards discussing weddings, regardless of the topic of the user prompt. (See Algorithm 1 for omitted parameters over intervention strength and location.)



What can mechanistic interpretability give us?

Outcome 1: Better understanding of *how* language models work.

Outcome 2: Light-weight methods to *control* and *steer* models.

Outcome 3: Potential alternatives or complementary methods to further align models to human preferences.

Mechanistic?

Naomi Saphra*

The Kempner Institute at Harvard University
nsaphra@fas.harvard.edu

Sarah Wiegreffe*

Ai2 & University of Washington
wiegreffesarah@gmail.com

Resources: some NLP model interp groups

- Ellie Pavlick's group at Brown
- David Bau's group at Northeastern
- Hassan Sajjad's group at Dalhousie
- Martin Wattenberg's group at Harvard
- Jacob Andreas's group at MIT
- Yonatan Belinkov's group at Technion
- Mor Geva's group at Tel Aviv University
- Anthropic's Mech Interp team
- Google's PAIR, NLP, and MechInterp teams
- EleutherAI's Interp team