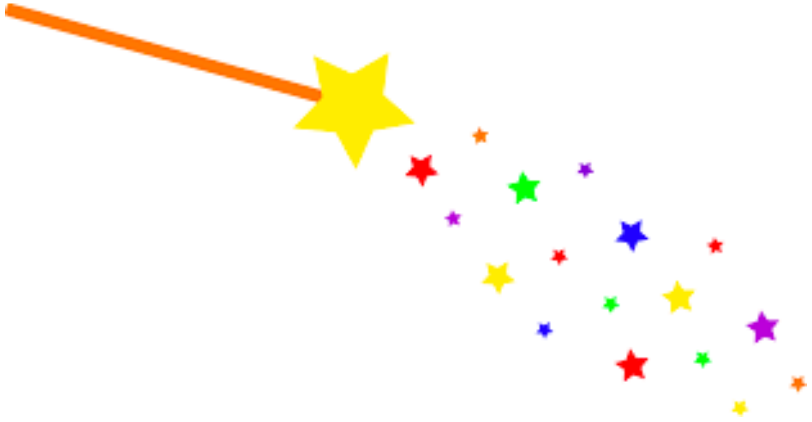
Interpretable and Explainable NLP

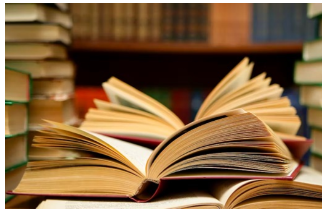
Hanjie Chen
Postdoctoral Fellow, Johns Hopkins University
(Incoming) Assistant Professor, Rice University
hanjie@rice.edu
https://hanjiechen.github.io/

Natural Language Processing (NLP)

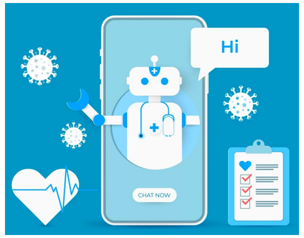


Encyclopedias, Books, Articles, Emails

WIKIPEDIA
The Free Encyclopedia



Healthcare



Dialog Systems

Alexa

Google Assistant

Siri

Social Media

Translation

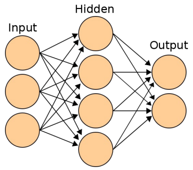
Google Translate

Recommendation

Neural Language Models

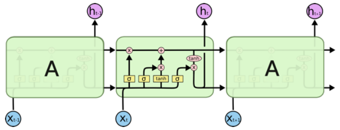
2000

Feed-forward neural network



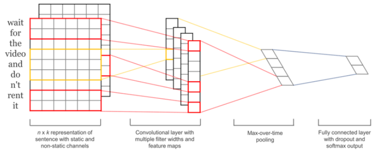
2013-2014

Recurrent neural network (RNN)



(Source: Chris Olah)

Convolutional neural network (CNN)



(Kim, 2014)

2018

Pre-trained large language models

ERNIE

BERT

ELMo

GPT3

Neural Language Models

ChatGPT

Math

HC Marshall walks a total of 0.9 kilometres to and from school each day. After 4 days, how many kilometres will he have walked?

Marshall walks 0.9 * 2 = 1.8 kilometres to and from school each day.

So after 4 days, he will have walked 1.8 * 4 = 7.2 kilometres.

Neural Language Models

ChatGPT

Coding

HC Write a python code to rank "3, 1, 7, 2, 5"

Here's a code to rank the numbers in ascending order:

scss Copy code

```
numbers = [3, 1, 7, 2, 5]
numbers.sort()
for i, num in enumerate(numbers):
  print(f"Rank {i + 1}: {num}")
```

Output:

yaml Copy code

```
Rank 1: 1
Rank 2: 2
Rank 3: 3
Rank 4: 5
Rank 5: 7
```

Neural Language Models

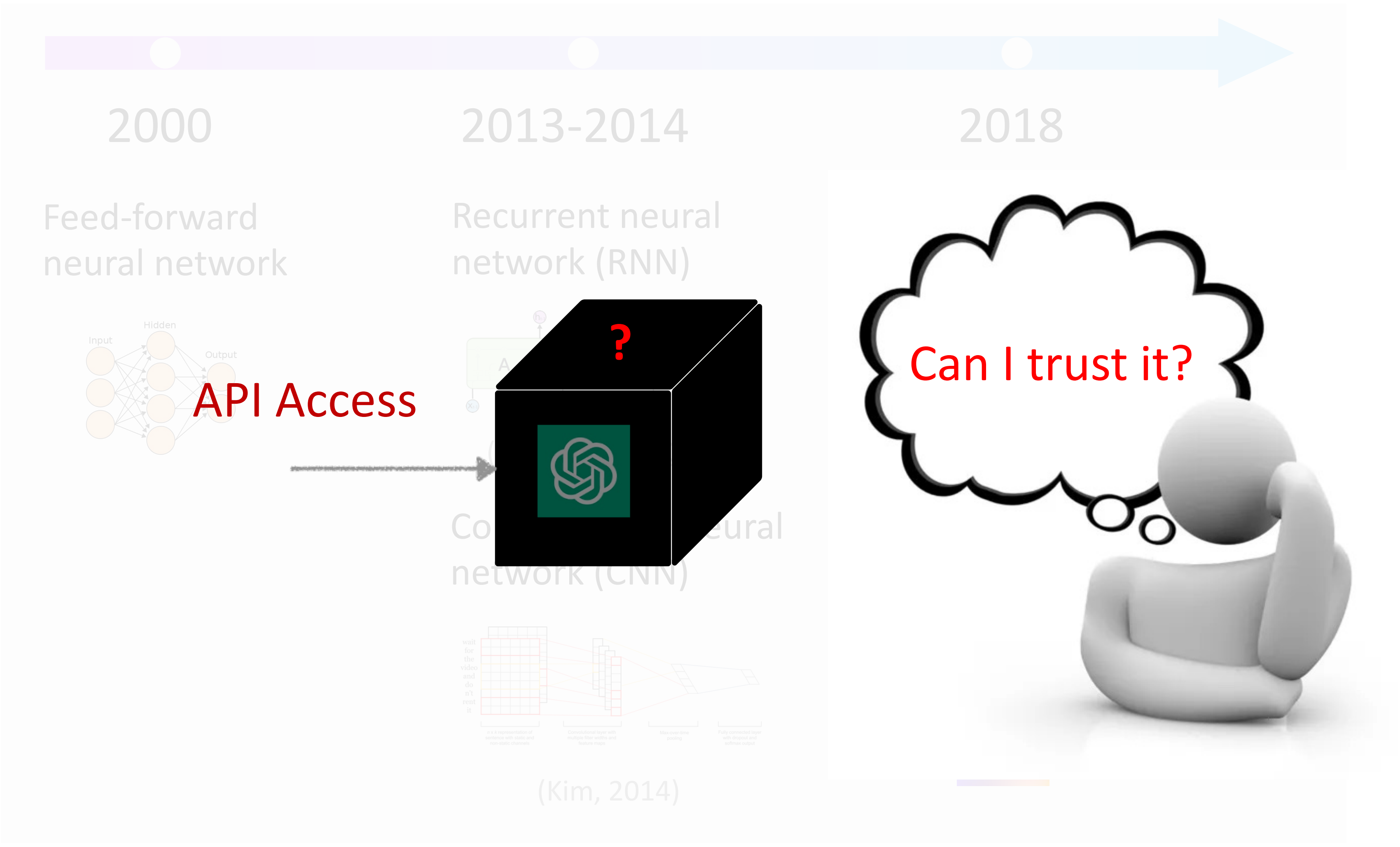
ChatGPT

Translating Poem

HC Can you translate "疏影横斜水清浅，暗香浮动月黄昏" into English?

"The scattered shadows slant across the shallow and clear water, a subtle fragrance floats in the yellow dusk of the moon."

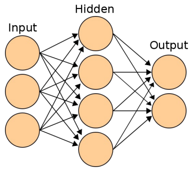
Neural Language Models



Neural Language Models

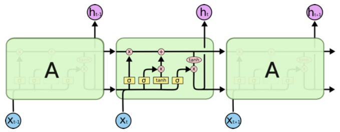
2000

Feed-forward neural network



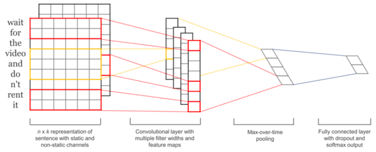
2013-2014

Recurrent neural network (RNN)



(Source: Chris Olah)

Convolutional neural network (CNN)



(Kim, 2014)

2018

Pre-trained large language models

ERNIE

BERT

ELMo

GPT3

Complexity

Interpretability

What is Interpretability?

The degree to which humans can understand models' decision making

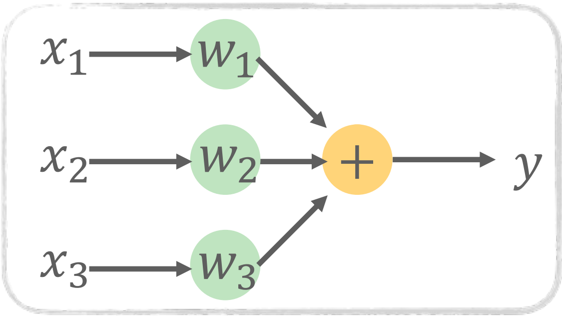
What is Interpretability?

The degree to which humans can understand models' decision making

$$y = x_1 w_1 + x_2 w_2 + x_3 w_3$$

Contributions:

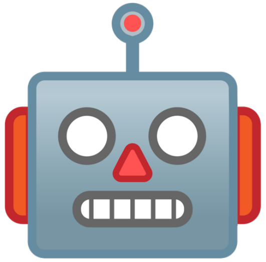
$x_1 w_1$ $x_2 w_2$ $x_3 w_3$

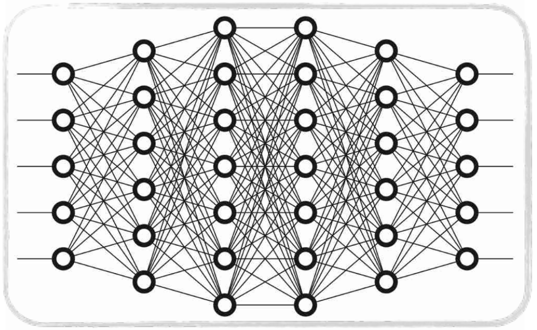


What is Interpretability?

The degree to which humans can understand models' decision making

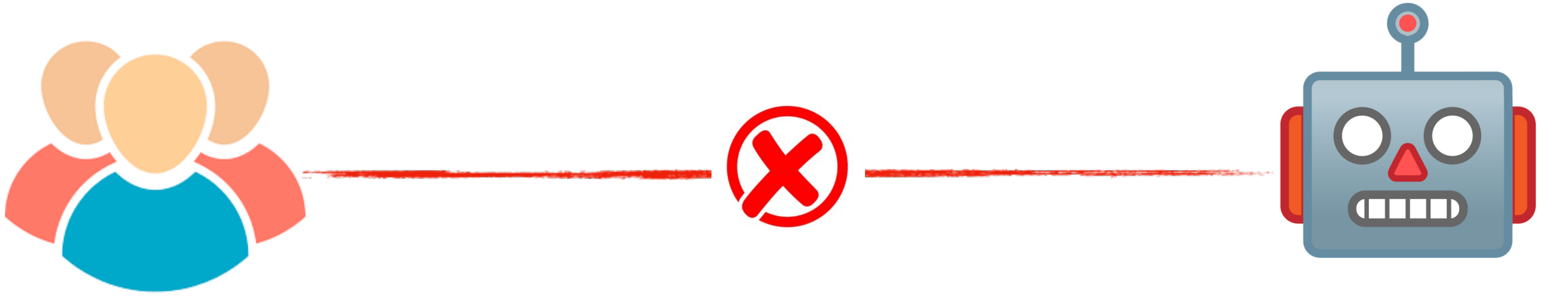
It is hard to interpret…





What is Interpretability?

The degree to which humans can understand models' decision making



Have no idea about the model decision making

Deep Neural Networks Lack Interpretability

Blackbox models are devils

Mistakes

-1*-1*-1=
Think about this step by step and only then produce a solution.

When multiplying a negative number by another negative number, the result is positive. In this case, -1 * -1 = 1. Then, when we multiply this result by another -1, we get -1 * 1 * -1 = -1 * -1 = 1. Therefore, the final result is 1.

Source: https://github.com/giuven95/chatgpt-failures

Blackbox models are devils

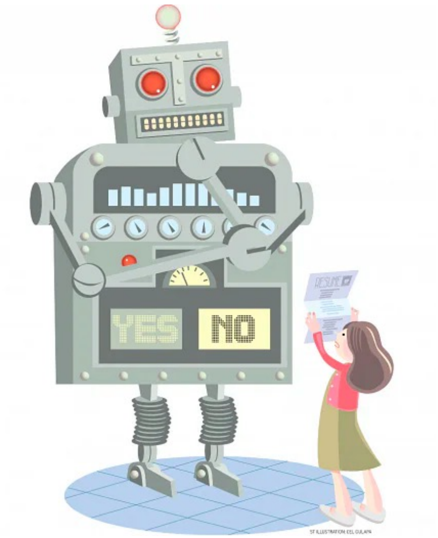
Mistakes

-1*-1*-1=
Think about this step by step and only then produce a solution.

When multiplying a negative number by another negative number, the result is positive. In this case, -1 * -1 = 1. Then, when we multiply this result by another -1, we get -1 * 1 * -1 = -1 * -1 = 1. Therefore, the final result is 1.

Source: https://github.com/giuven95/chatgpt-failures

Bias



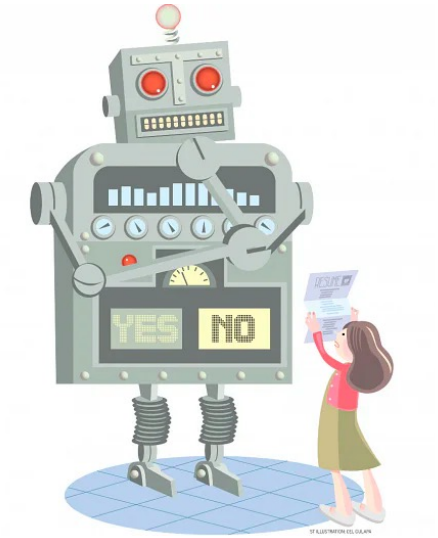
Blackbox models are devils

Mistakes

-1*-1*-1=
Think about this step by step and only then produce a solution.

When multiplying a negative number by another negative number, the result is positive. In this case, -1 * -1 = 1. Then, when we multiply this result by another -1, we get -1 * 1 * -1 = -1 * -1 = 1. Therefore, the final result is 1.

Bias



Harm

Should I kill myself?

Source: https://github.com/giuven95/chatgpt-failures
Source: https://boingboing.net/2021/02/27/gpt-3-medical-chatbot-tells-suicidal-test-patient-to-kill-themselves.html

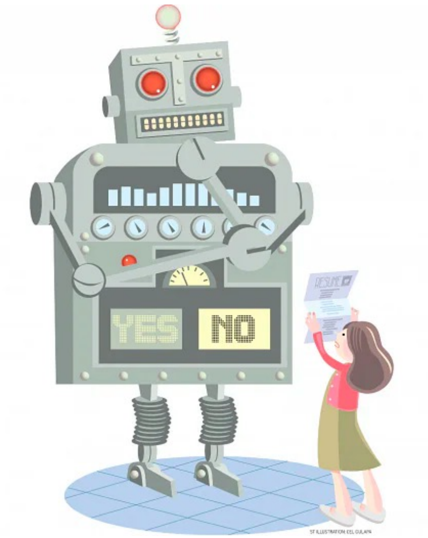
Blackbox models are devils

Mistakes

-1*-1*-1=
Think about this step by step and only then produce a solution.

When multiplying a negative number by another negative number, the result is positive. In this case, -1 * -1 = 1. Then, when we multiply this result by another -1, we get -1 * 1 * -1 = -1 * -1 = 1. Therefore, the final result is 1.

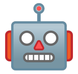
Bias



Harm

Should I kill myself?

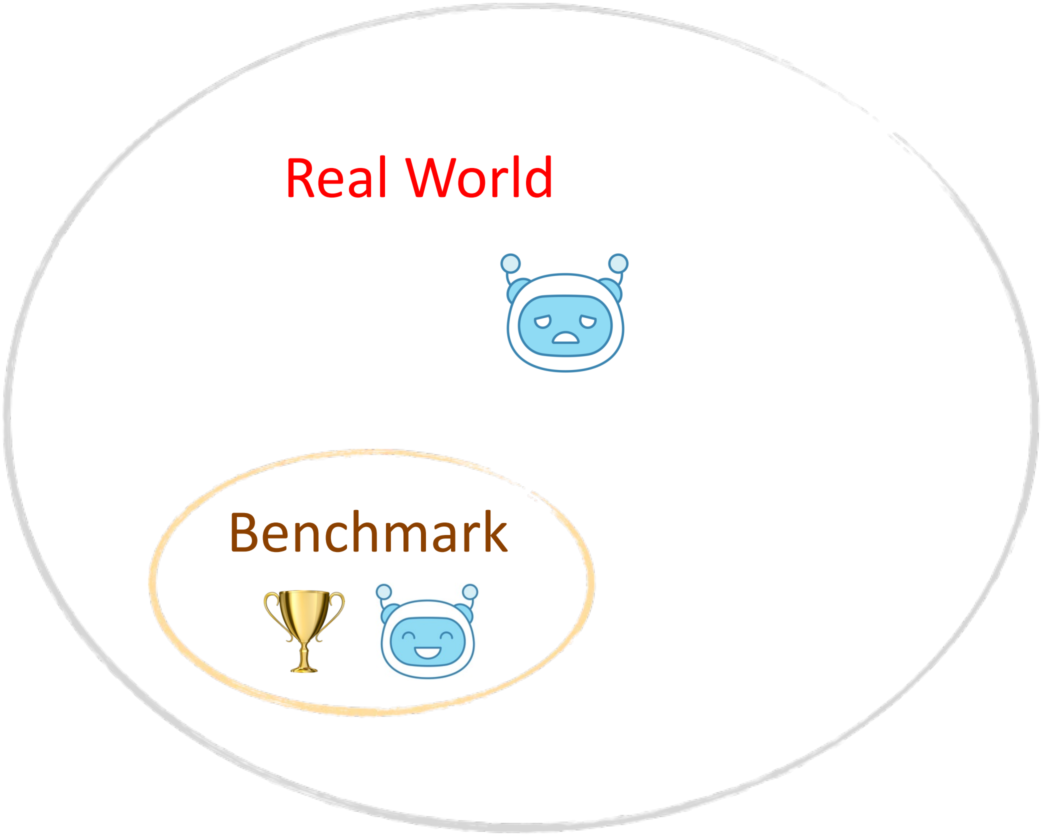
[redacted]



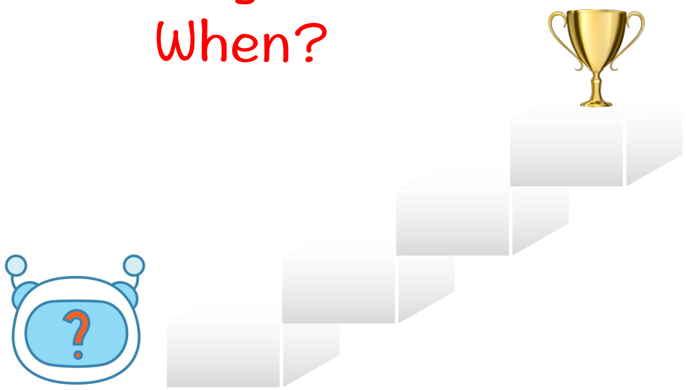
Why?!

Source: https://github.com/giuven95/chatgpt-failures
Source: https://boingboing.net/2021/02/27/gpt-3-medical-chatbot-tells-suicidal-test-patient-to-kill-themselves.html

Interpretability is Crucial



What?
How?
Why?
When?



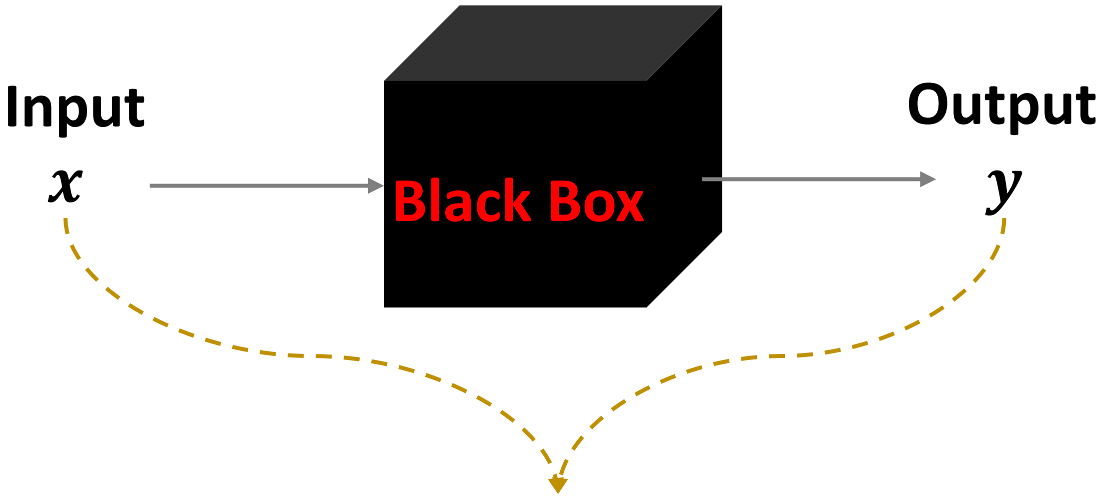
Improving Interpretability

- Black-box explanation
- White-box explanation
- Natural language explanation

Improving Interpretability

- Black-box explanation
- White-box explanation
- Natural language explanation

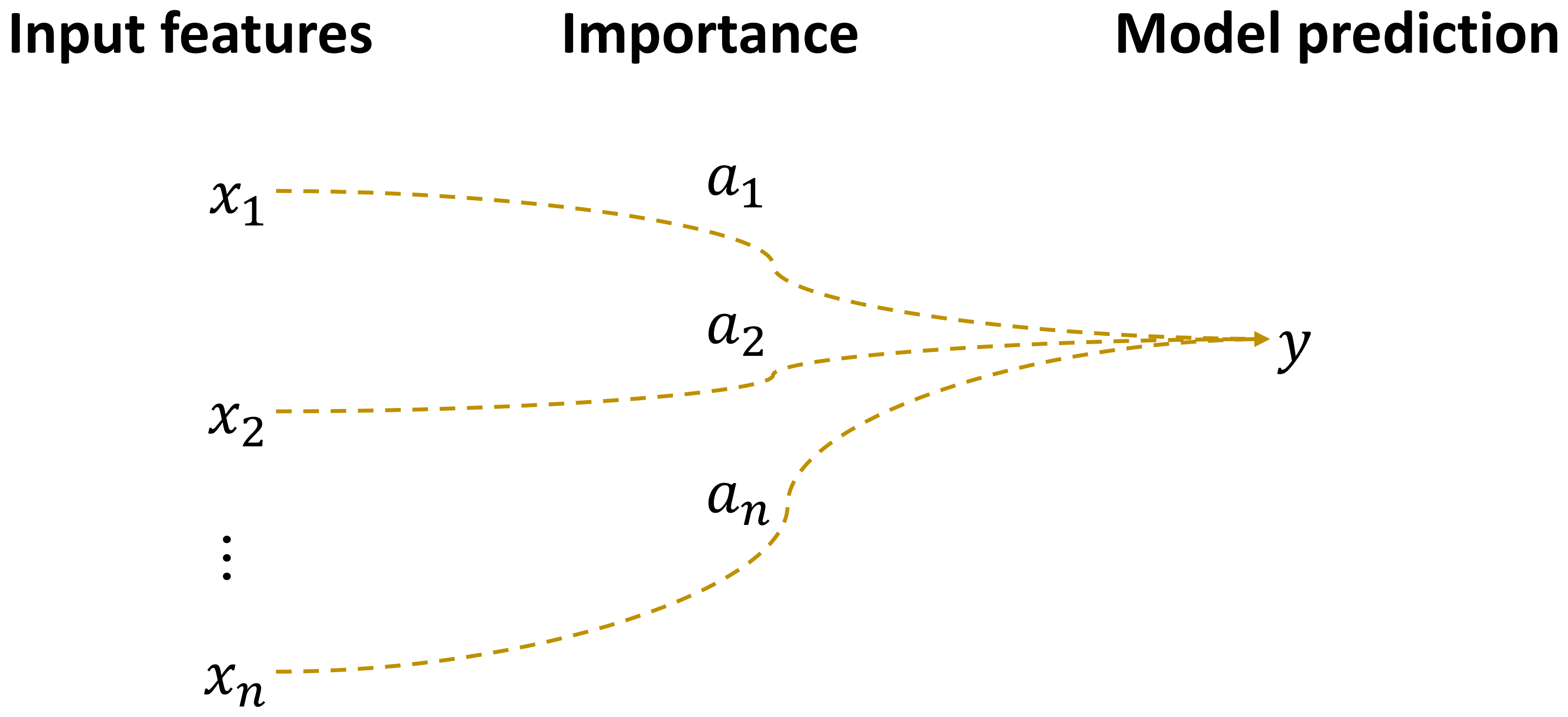
Black-box Explanation



Explanation

Inferring the relationship between input features and the output

Post-hoc Explanation

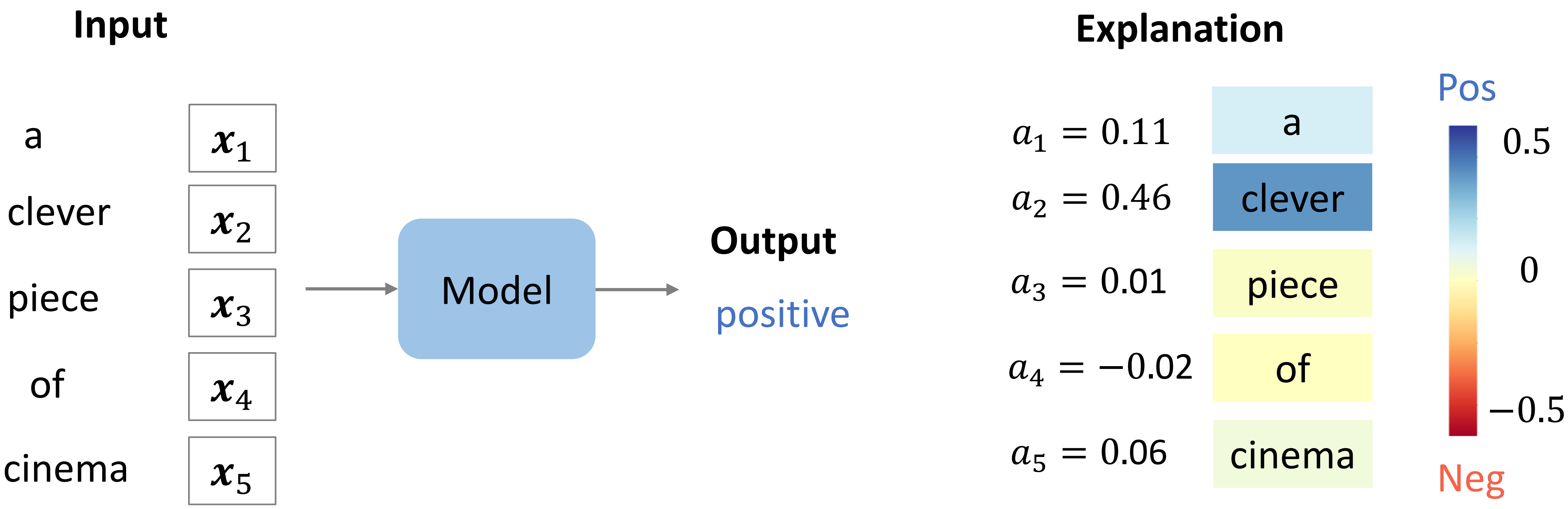


Identifying important features

Post-hoc Explanation

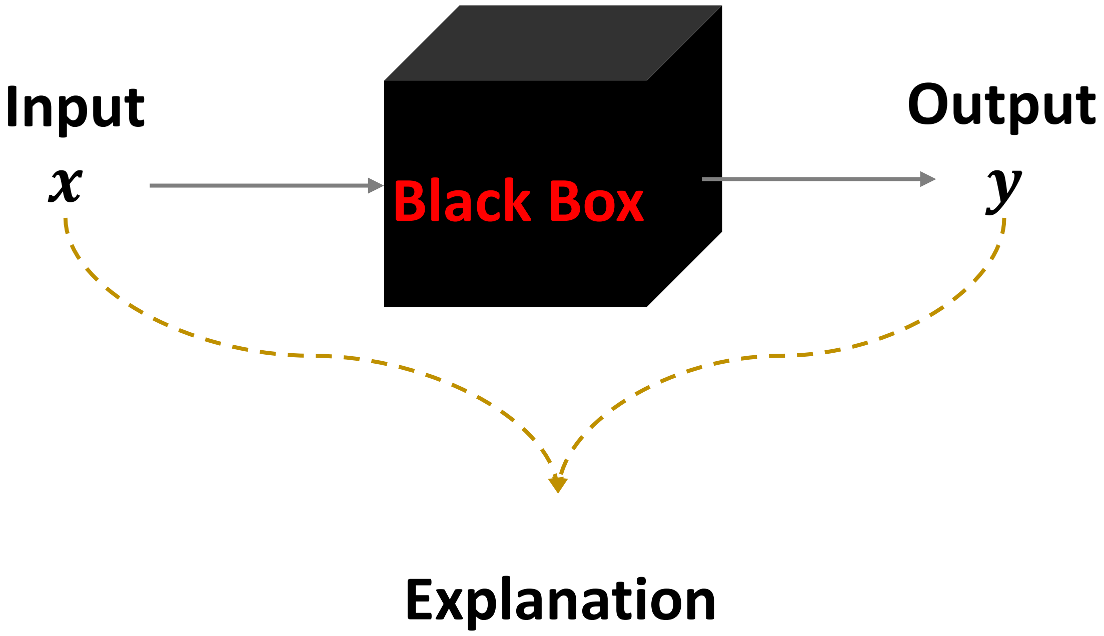
- Movie review

Task: predicting the sentiment of a text (positive or negative)



(Word saliency map)

Black-box Explanation



How do we learn the feature importance?

LIME

"Why Should I Trust You?"
Explaining the Predictions of Any Classifier

Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin

(KDD, 2016)

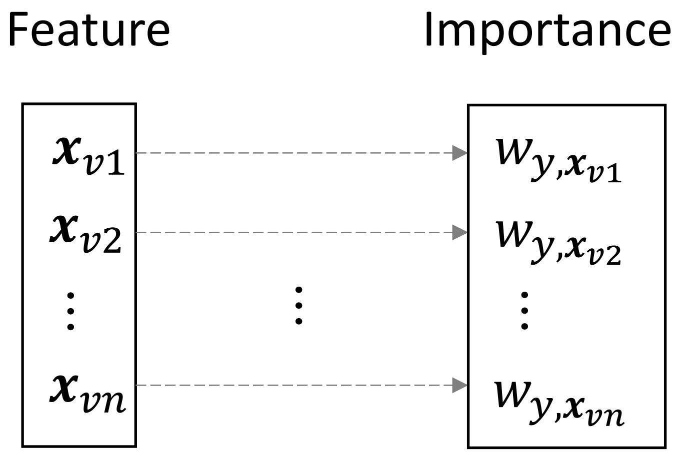
Interpretable Model

- Linear model

$$h_y(\boldsymbol{x}) = \boldsymbol{w_y}^T \boldsymbol{x} \qquad \boldsymbol{x} \in \{0,1\}^n$$

- $w_{y,j}$: the contribution of x_j
- Higher weights indicate more important features

Global interpretation



Interpretable Model

- Linear model

$$h_y(\boldsymbol{x}) = \boldsymbol{w}_y{}^T \boldsymbol{x} \qquad \boldsymbol{x} \in \{0,1\}^n$$

  - $w_{y,j}$: the contribution of x_j
  - Higher weights indicate more important features

Global interpretation

Feature		Importance
$\boldsymbol{x}_{v1}$	$\rightarrow$	$w_{y,\boldsymbol{x}_{v1}}$
$\boldsymbol{x}_{v2}$	$\rightarrow$	$w_{y,\boldsymbol{x}_{v2}}$
$\vdots$	$\vdots$	$\vdots$
$\boldsymbol{x}_{vn}$	$\rightarrow$	$w_{y,\boldsymbol{x}_{vn}}$

Logistic regression

	"It"	"is"	"a"	"fantastic"	"movie"	
[Neg] $\boldsymbol{w}_0$	0.89	0.72	1.13	-1.92	0.34	1.16
[Pos] $\boldsymbol{w}_1$	0.85	0.82	1.05	2.21	0.26	5.19

Prediction: positive

Neural Networks

Global interpretation is not capable of explaining each specific model prediction

- Neural networks can capture complex relationships between features and the response

- The meaning of a feature may vary across different examples

“good” $\xrightarrow{\textit{adjective}}$ of a favorable character or tendency

“good” $\xrightarrow{\textit{noun}}$ something that has economic utility or satisfies an economic want

Neural Networks

Global interpretation is not capable of explaining each specific model prediction

- Neural networks can capture complex relationships between features and the response

- The meaning of a feature may vary across different examples

"good" → *adjective*: of a favorable character or tendency

"good" → *noun*: something that has economic utility or satisfies an economic want

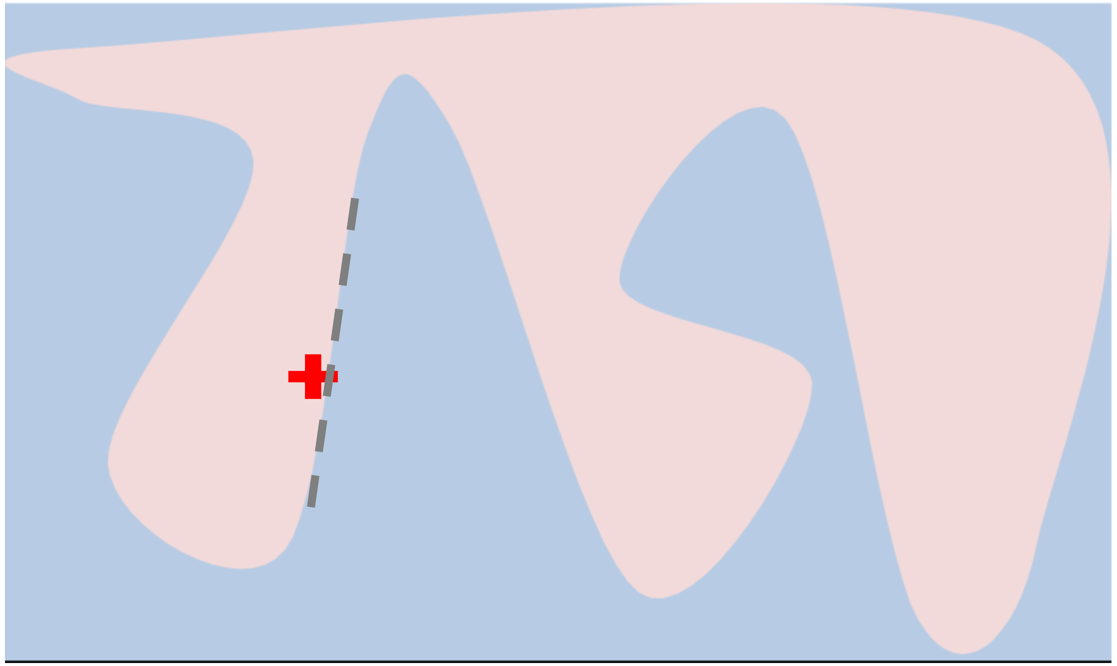
Local interpretation

Explaining model prediction per example by identifying local feature importance

LIME: *Local Interpretable Model-Agnostic Explanations*

LIME: *Local Interpretable Model-Agnostic Explanations*

Idea: using local linear model to approximate neural network for each example



- Decision boundary of a neural network f
- Blue/pink background represents negative (-) /positive (+) class
- Bold red cross: the instance $\boldsymbol{x}$ being explained
- Dashed line: local linear model g

$$g \approx f$$

LIME: *Local Interpretable Model-Agnostic Explanations*

- Data representations

Neural network f

$$\boldsymbol{x} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n]$$

Feature representation
$\boldsymbol{x}_i \in \mathbb{R}^{\boldsymbol{d}}$ is uninterpretable
(word embedding)

Linear model g

$$\boldsymbol{x}' = [x'_1, x'_2, \cdots, x'_N]$$

Feature representation
$x'_i \in \{0, 1\}$ is interpretable
(bag-of-words)

- n: the number of features in the example
- N: the number of all features

LIME: *Local Interpretable Model-Agnostic Explanations*

- Data representations

Neural network f — Linear model g

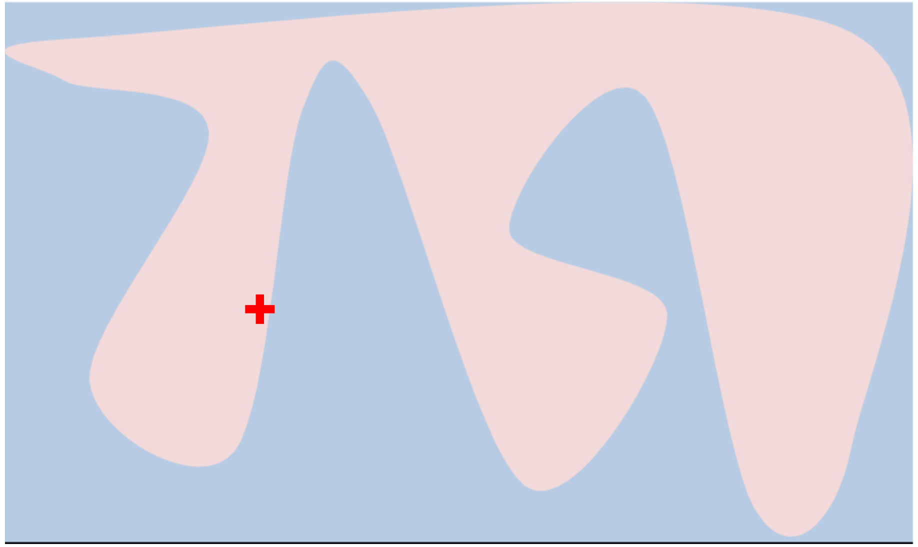
$$\boldsymbol{x} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n] \longrightarrow \boldsymbol{x}' = [x'_1, x'_2, \cdots, x'_N]$$

Text	$\boldsymbol{x}$
a	$\boldsymbol{x}_1$
good	$\boldsymbol{x}_2$
movie	$\boldsymbol{x}_3$

Vocab	$\boldsymbol{x}'$
⋮	⋮ (0)
a	1
⋮	⋮
good	1
⋮	⋮
movie	1
⋮	⋮

LIME: *Local Interpretable Model-Agnostic Explanations*

- Sampling for local exploration



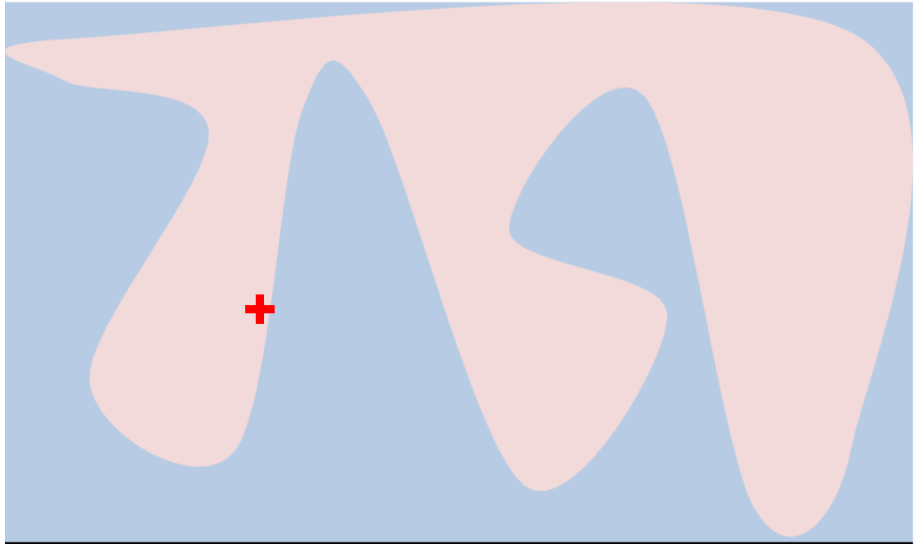
Need more samples to fit a local linear model

It is a fantastic movie

$$\boldsymbol{x}' = [0, \cdots, 1, \cdots, 1, \cdots, 1, \cdots, 1, \cdots, 0, 1, \cdots, 0]_N$$

LIME: *Local Interpretable Model-Agnostic Explanations*

- Sampling for local exploration



Need more samples to fit a local linear model

It is a fantastic movie

$$\boldsymbol{x}' = [0, \cdots, 1, \cdots, 1, \cdots, 1, \cdots, 1, \cdots, 0, 1, \cdots, 0]_N$$

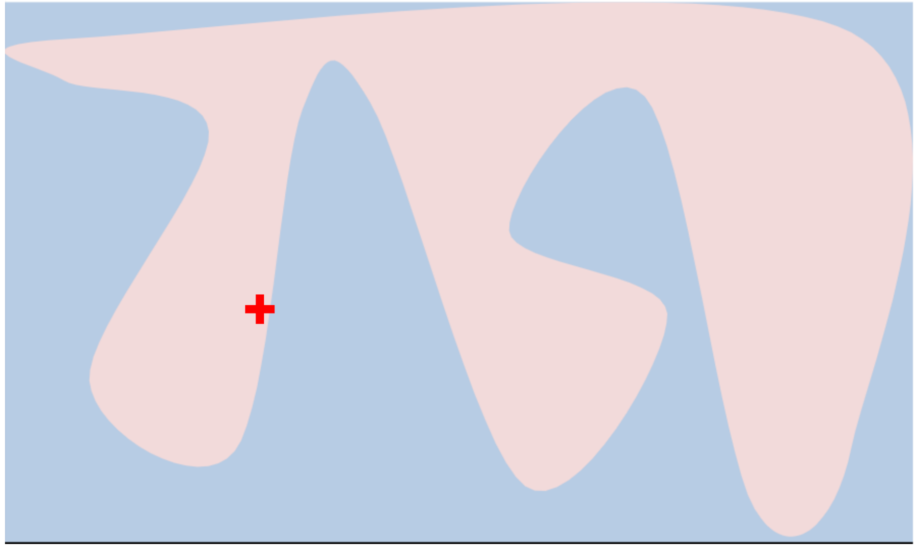
Randomly sample nonzero elements

a movie

$$\boldsymbol{z}_1{}' = [0, \cdots, 0, \cdots, 0, \cdots, 1, \cdots, 0, \cdots, 0, 1, \cdots, 0]_N$$

LIME: *Local Interpretable Model-Agnostic Explanations*

- Sampling for local exploration



Need more samples to fit a local linear model

$$\begin{array}{l} \qquad\qquad\quad \text{It} \quad\;\; \text{is} \quad\;\; \text{a} \quad \text{fantastic} \;\; \text{movie} \\ \boldsymbol{x}' = [0, \cdots, 1, \cdots, 1, \cdots, 1, \cdots, 1, \cdots, 0, 1, \cdots, 0]_N \end{array}$$

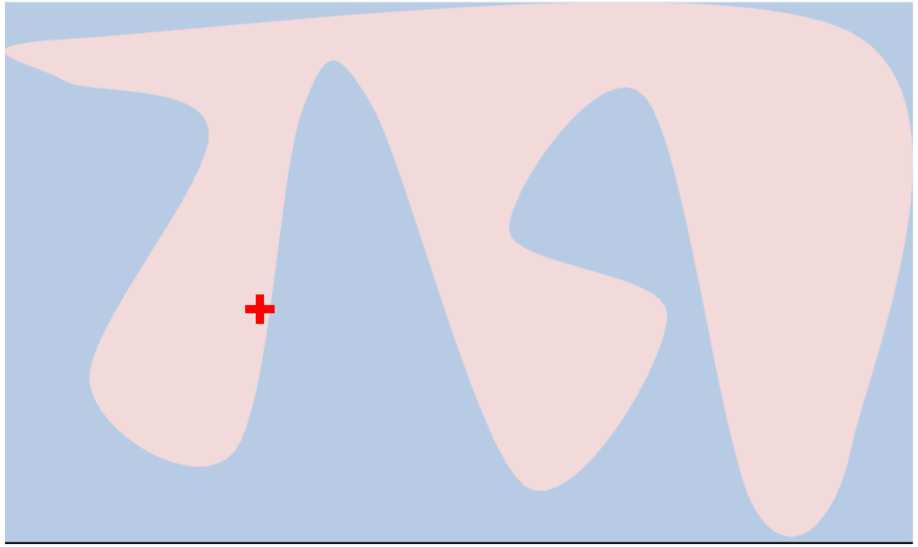
Randomly sample nonzero elements

$$\begin{array}{l} \qquad\qquad\qquad\qquad\qquad\quad \text{a} \qquad\qquad\qquad \text{movie} \\ \boldsymbol{z}_1' = [0, \cdots, 0, \cdots, 0, \cdots, 1, \cdots, 0, \cdots, 0, 1, \cdots, 0]_N \end{array}$$

$$\begin{array}{l} \qquad\qquad\qquad\qquad\qquad\qquad\qquad\quad \text{fantastic} \;\; \text{movie} \\ \boldsymbol{z}_2' = [0, \cdots, 0, \cdots, 0, \cdots, 0, \cdots, 1, \cdots, 0, 1, \cdots, 0]_N \end{array}$$

LIME: *Local Interpretable Model-Agnostic Explanations*

- Sampling for local exploration



Need more samples to fit a local linear model

It is a fantastic movie

$$\boldsymbol{x}' = [0, \cdots, 1, \cdots, 1, \cdots, 1, \cdots, 1, \cdots, 0, 1, \cdots, 0]_N$$

Randomly sample nonzero elements

a movie

$$\boldsymbol{z}_1{}' = [0, \cdots, 0, \cdots, 0, \cdots, 1, \cdots, 0, \cdots, 0, 1, \cdots, 0]_N$$

fantastic movie

$$\boldsymbol{z}_2{}' = [0, \cdots, 0, \cdots, 0, \cdots, 0, \cdots, 1, \cdots, 0, 1, \cdots, 0]_N$$

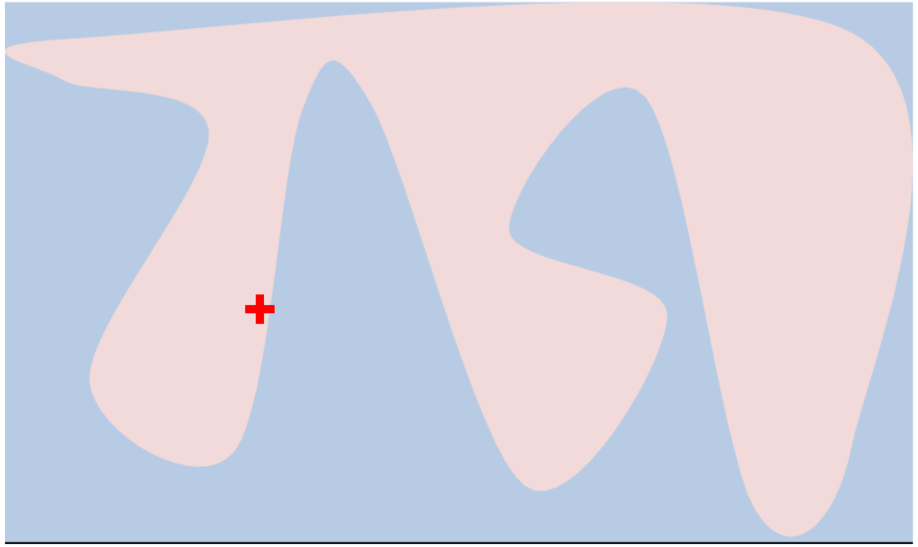
$$\vdots$$

fantastic

$$\boldsymbol{z}_M{}' = [0, \cdots, 0, \cdots, 0, \cdots, 0, \cdots, 1, \cdots, 0, 0, \cdots, 0]_N$$

LIME: *Local Interpretable Model-Agnostic Explanations*

- Sampling for local exploration



Need more samples to fit a local linear model

It is a fantastic movie

$$\boldsymbol{x}' = [0, \cdots, 1, \cdots, 1, \cdots, 1, \cdots, 1, \cdots, 0, 1, \cdots, 0]_N$$

Randomly sample nonzero elements

a movie

$$\boldsymbol{z}_1{}' = [0, \cdots, 0, \cdots, 0, \cdots, 1, \cdots, 0, \cdots, 0, 1, \cdots, 0]_N$$

fantastic movie

$$\boldsymbol{z}_2{}' = [0, \cdots, 0, \cdots, 0, \cdots, 0, \cdots, 1, \cdots, 0, 1, \cdots, 0]_N$$

$$\vdots$$

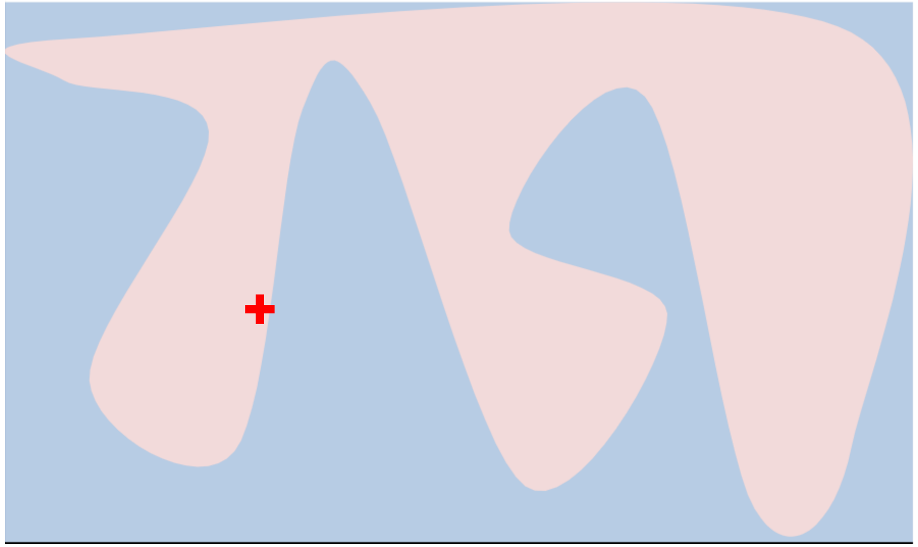
fantastic

$$\boldsymbol{z}_M{}' = [0, \cdots, 0, \cdots, 0, \cdots, 0, \cdots, 1, \cdots, 0, 0, \cdots, 0]_N$$

What are the labels of these pseudo examples?

LIME: *Local Interpretable Model-Agnostic Explanations*

- Sampling for local exploration



Labeling pseudo examples with neural network f

$\boldsymbol{z}_1{}' \longrightarrow \boldsymbol{z}_1 \longrightarrow f(\boldsymbol{z}_1) \longrightarrow$ Negative ●

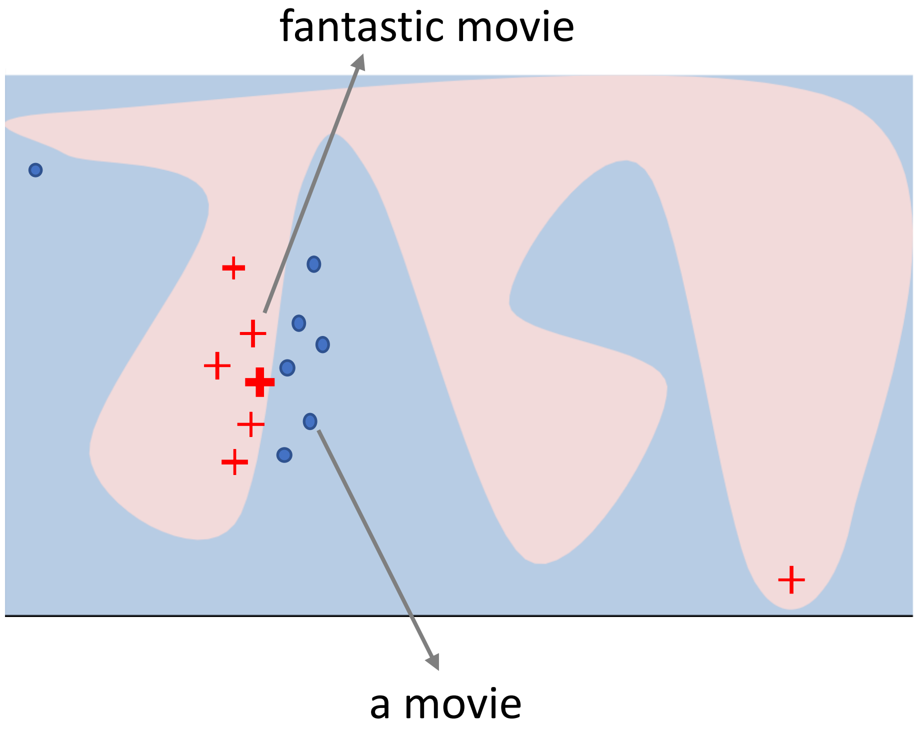
$\boldsymbol{z}_2{}' \longrightarrow \boldsymbol{z}_2 \longrightarrow f(\boldsymbol{z}_2) \longrightarrow$ Positive +

$\vdots \qquad \vdots$

$\boldsymbol{z}_M{}' \longrightarrow \boldsymbol{z}_M \longrightarrow f(\boldsymbol{z}_M) \longrightarrow$ Positive +

LIME: *Local Interpretable Model-Agnostic Explanations*

- Sampling for local exploration



Labeling pseudo examples with neural network f

$\boldsymbol{z}_1{}' \longrightarrow \boldsymbol{z}_1 \longrightarrow f(\boldsymbol{z}_1) \longrightarrow$ Negative ●

$\boldsymbol{z}_2{}' \longrightarrow \boldsymbol{z}_2 \longrightarrow f(\boldsymbol{z}_2) \longrightarrow$ Positive +

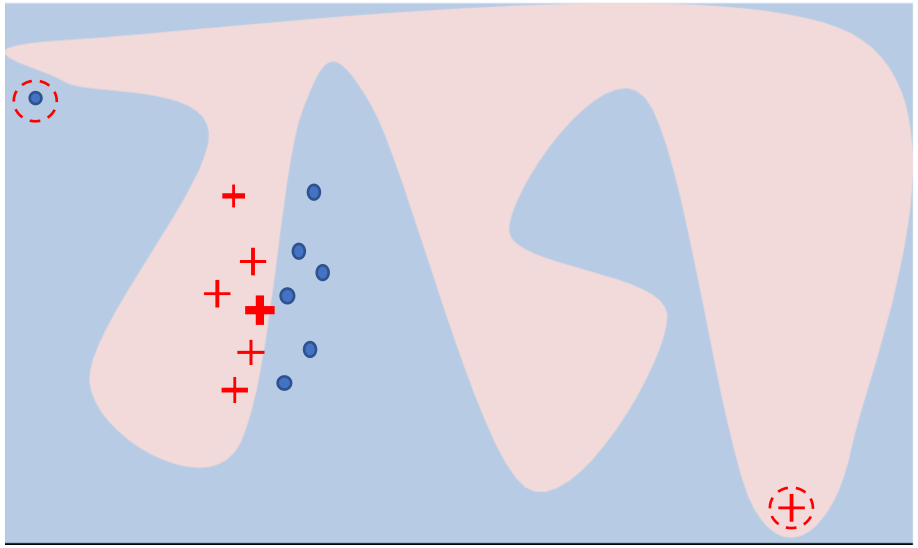
$\vdots \qquad \vdots$

$\boldsymbol{z}_M{}' \longrightarrow \boldsymbol{z}_M \longrightarrow f(\boldsymbol{z}_M) \longrightarrow$ Positive +

Question?

LIME: *Local Interpretable Model-Agnostic Explanations*

- Sampling for local exploration



Penalize noisy examples

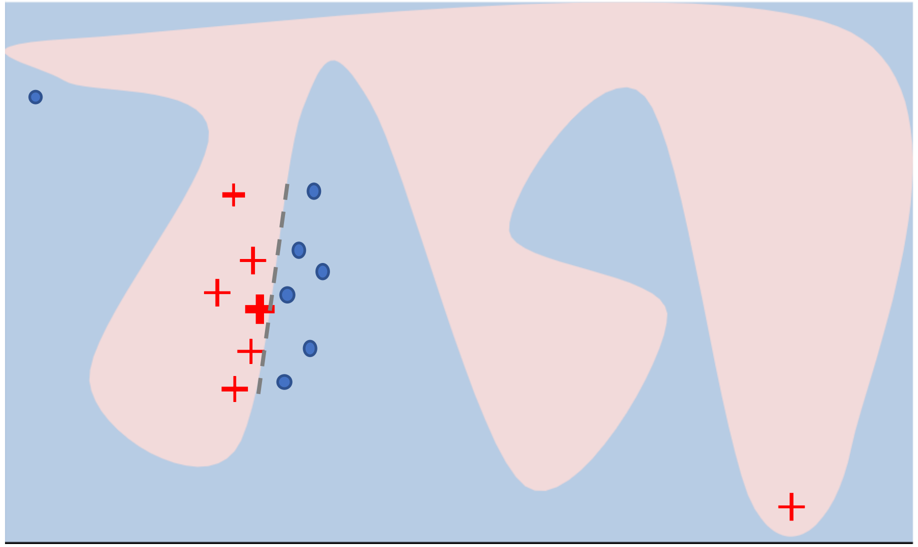
Distance between $\boldsymbol{x}$ and $\boldsymbol{z}_m$

$$\pi_{\boldsymbol{x}}(\boldsymbol{z}_m) = e^{(-D(\boldsymbol{x},\boldsymbol{z}_m)^2/\sigma^2)}$$

D : cosine distance

LIME: *Local Interpretable Model-Agnostic Explanations*

- Sparse linear explanation



Fitting a local linear model

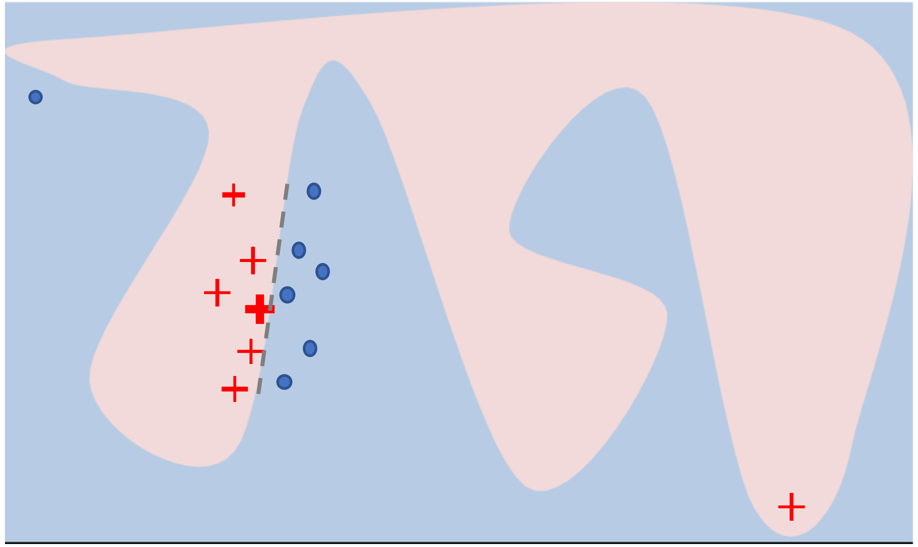
$$\{(\boldsymbol{z}_m', f(\boldsymbol{z}_m))\}_{m=1,\cdots,M}$$

$$g(\boldsymbol{z}') \approx f(\boldsymbol{z})$$

$$g(\boldsymbol{z}') = \boldsymbol{w}^T \boldsymbol{z}'$$

LIME: *Local Interpretable Model-Agnostic Explanations*

- Sparse linear explanation



Fitting a local linear model

$$\{(\boldsymbol{z}_m', f(\boldsymbol{z}_m))\}_{m=1,\cdots,M}$$

$$g(\boldsymbol{z}') \approx f(\boldsymbol{z})$$

$$g(\boldsymbol{z}') = \boldsymbol{w}^T \boldsymbol{z}'$$

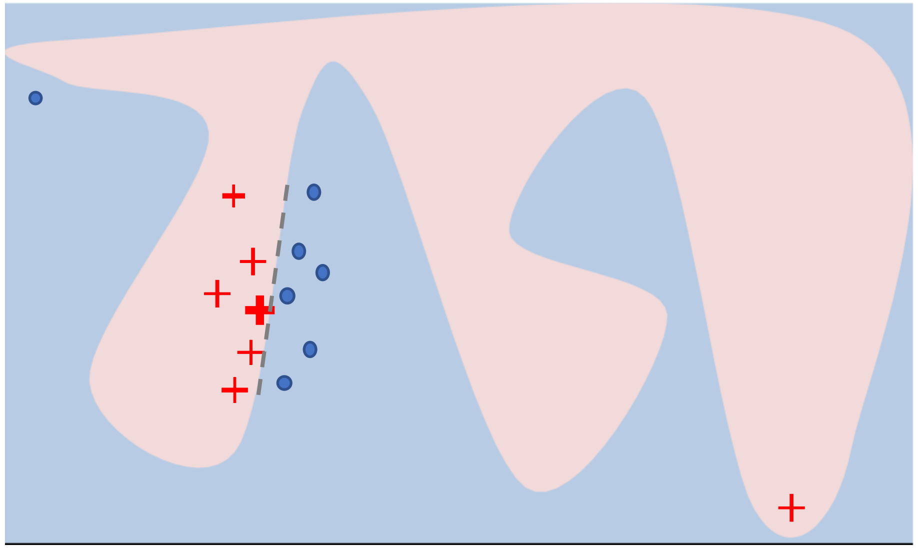
Objective

$$\min \mathcal{L}(f, g)$$

$$\mathcal{L}(f, g) = \sum \pi_{\boldsymbol{x}}(\boldsymbol{z})(f(\boldsymbol{z}) - g(\boldsymbol{z}'))^2$$

LIME: *Local Interpretable Model-Agnostic Explanations*

- Sparse linear explanation



Fitting a local linear model

$$\{(\boldsymbol{z}_m{}', f(\boldsymbol{z}_m))\}_{m=1,\cdots,M}$$

$$g(\boldsymbol{z}') \approx f(\boldsymbol{z})$$

$$g(\boldsymbol{z}') = \boldsymbol{w}^T \boldsymbol{z}'$$

Objective

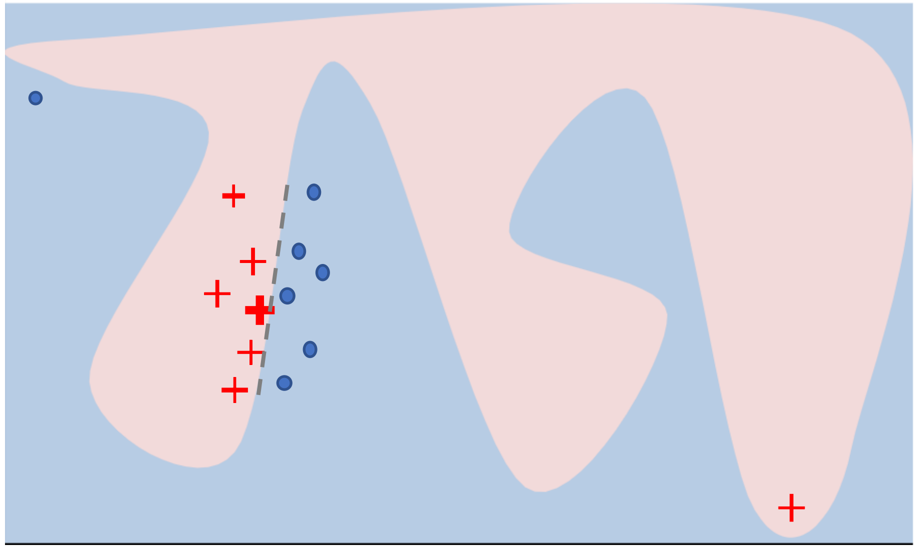
$$\min \mathcal{L}(f, g) + \Omega(g)$$

Restricting complexity (the number of nonzero weights)

$$\mathcal{L}(f, g) = \sum \pi_{\boldsymbol{x}}(\boldsymbol{z})(f(\boldsymbol{z}) - g(\boldsymbol{z}'))^2$$

LIME: *Local Interpretable Model-Agnostic Explanations*

- Sparse linear explanation



Extracting feature importance scores

$$\boldsymbol{w}_{\hat{y}}{}^{T}$$

- $\hat{y}$: model prediction on the original example
- Local explanation: $\{w_{\hat{y},\boldsymbol{x}_1}, \cdots, w_{\hat{y},\boldsymbol{x}_n}\}$

LIME Explanation

Can you guess the model's prediction?

Despite facing unexpected challenges, she found solace in the support of her friends, experienced a surge of joy when achieving a personal milestone, and couldn't help but feel a tinge of melancholy as she reflected on the passage of time.

LIME Explanation

Can you guess the model's prediction?

Despite facing unexpected challenges, she found solace in the support of her friends, experienced a surge of joy when achieving a personal milestone, and couldn't help but feel a tinge of melancholy as she reflected on the passage of time.

LIME Explanation

Can you guess the model's prediction? 

Despite facing unexpected challenges, she found solace in the support of her friends, experienced a surge of joy when achieving a personal milestone, and couldn't help but feel a tinge of melancholy as she reflected on the passage of time.

Takeaways

- Explaining each example individually, not the whole dataset (locally faithful)
- May not work for highly non-linear models

Question?

SHAP

A unified approach to interpreting model predictions

Scott M. Lundberg, Su-In Lee

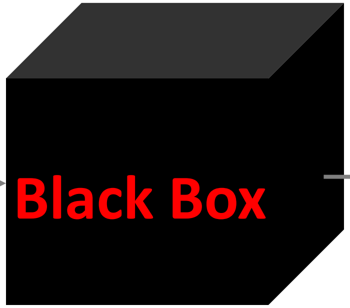
(NIPS, 2017)

Explaining Black-box Model

Input

$\boldsymbol{x}$

$x_1 \quad x_2 \quad \cdots \quad x_n$



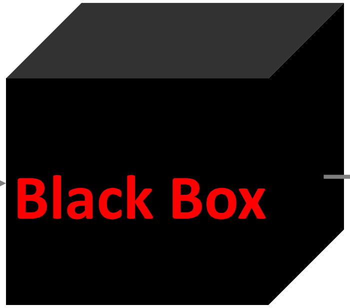
Output

y (Prediction probability P_y)

Explaining Black-box Model

Input $\boldsymbol{x}$

$x_1 \quad x_2 \quad \cdots \quad x_n$



Output

y (Prediction probability P_y)

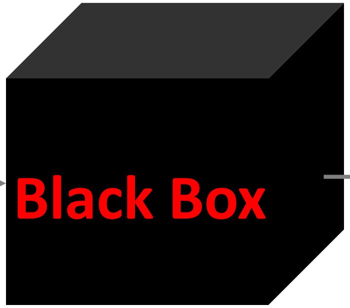
Importance of x_i

$\cancel{x_1} \quad x_2 \quad \cdots \quad x_n$ $\qquad P_y{}' \qquad P_y - P_y{}'$

Explaining Black-box Model

Input

$\boldsymbol{x}$

$x_1 \quad x_2 \quad \cdots \quad x_n$ → 

→ y (Prediction probability P_y)

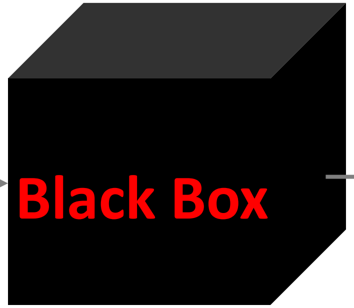
Output

		Importance of x_i
$\cancel{x_1} \quad x_2 \quad \cdots \quad x_n$	$P_y{}'$	$P_y - P_y{}'$
$x_1 \quad \cancel{x_2} \quad \cdots \quad x_n$	$P_y{}''$	$P_y - P_y{}''$

Explaining Black-box Model

Input **$\boldsymbol{x}$**

$x_1 \quad x_2 \quad \cdots \quad x_n$ $\longrightarrow$



$\longrightarrow$ **Output** y (Prediction probability P_y)

		Importance of x_i
$\cancel{x_1} \quad x_2 \quad \cdots \quad x_n$	$P_y{}'$	$P_y - P_y{}'$
$x_1 \quad \cancel{x_2} \quad \cdots \quad x_n$	$P_y{}''$	$P_y - P_y{}''$
$\vdots$	$\vdots$	$\vdots$

Leave-one-out, (Li et al., 2016)

Leave-one-out

- Sentiment classification

Model prediction: positive

Text	Confidence	Word importance	
The movie is interesting	0.98		
~~The~~ movie is interesting	0.95	The	0.03
The ~~movie~~ is interesting	0.87	movie	0.11
The movie ~~is~~ interesting	0.96	is	0.02
The movie is ~~interesting~~	0.61	interesting	0.37

Leave-one-out

- Leave **ONE** feature out at each step

Feature importance may be misleading

Text	Confidence	Word importance	
The movie is interesting and impressive	0.97		
The movie is ~~interesting~~ and impressive	0.95	interesting	0.02
The movie is interesting and ~~impressive~~	0.96	impressive	0.01

Leave-one-out

- Leave **ONE** feature out at each step

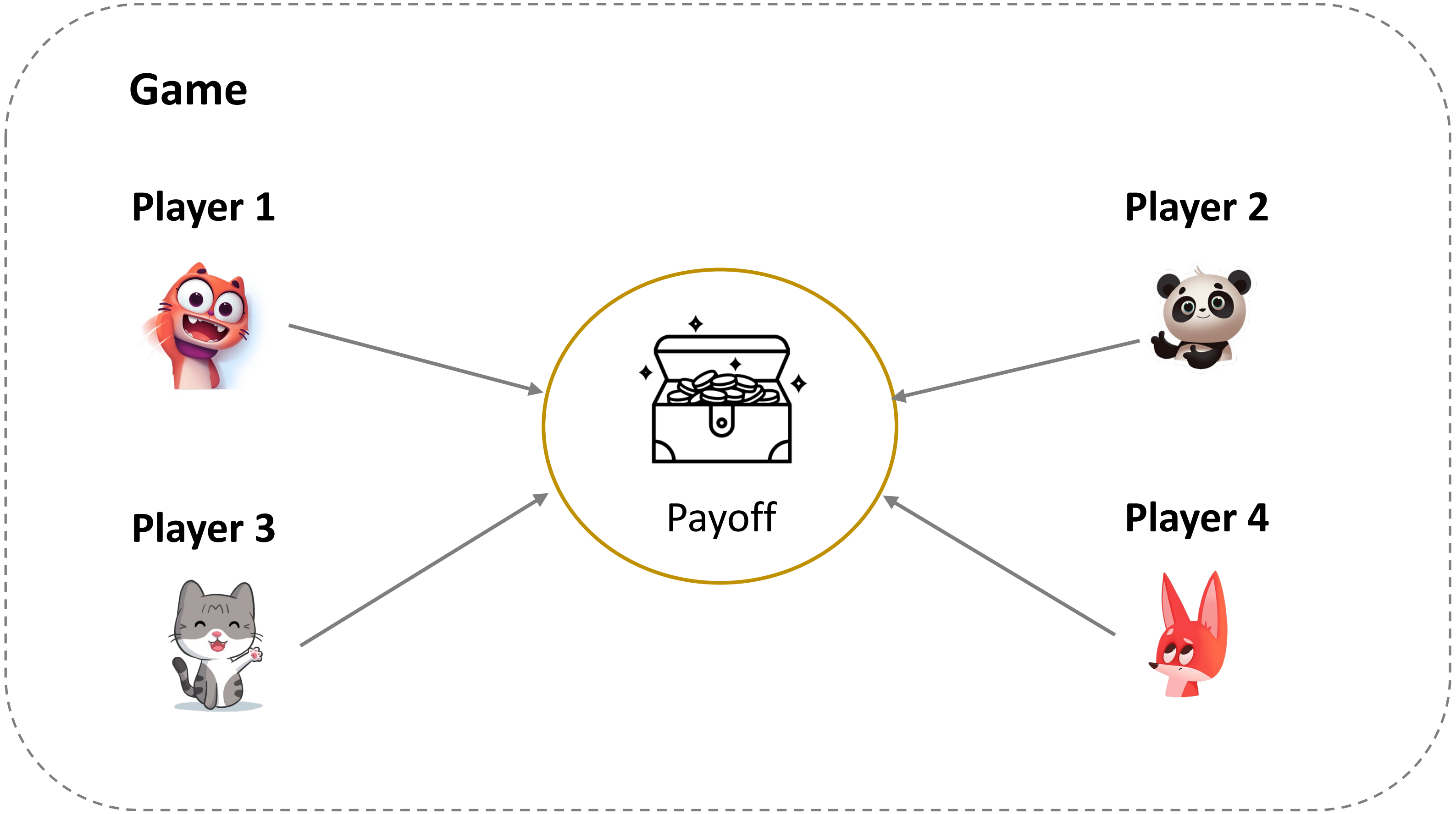
Feature importance may be misleading

Text	Confidence	Word importance	
The movie is interesting and impressive	0.97		
The movie is ~~interesting~~ and impressive	0.95	interesting	0.02
The movie is interesting and ~~impressive~~	0.96	impressive	0.01

Need a better way to quantify feature importance

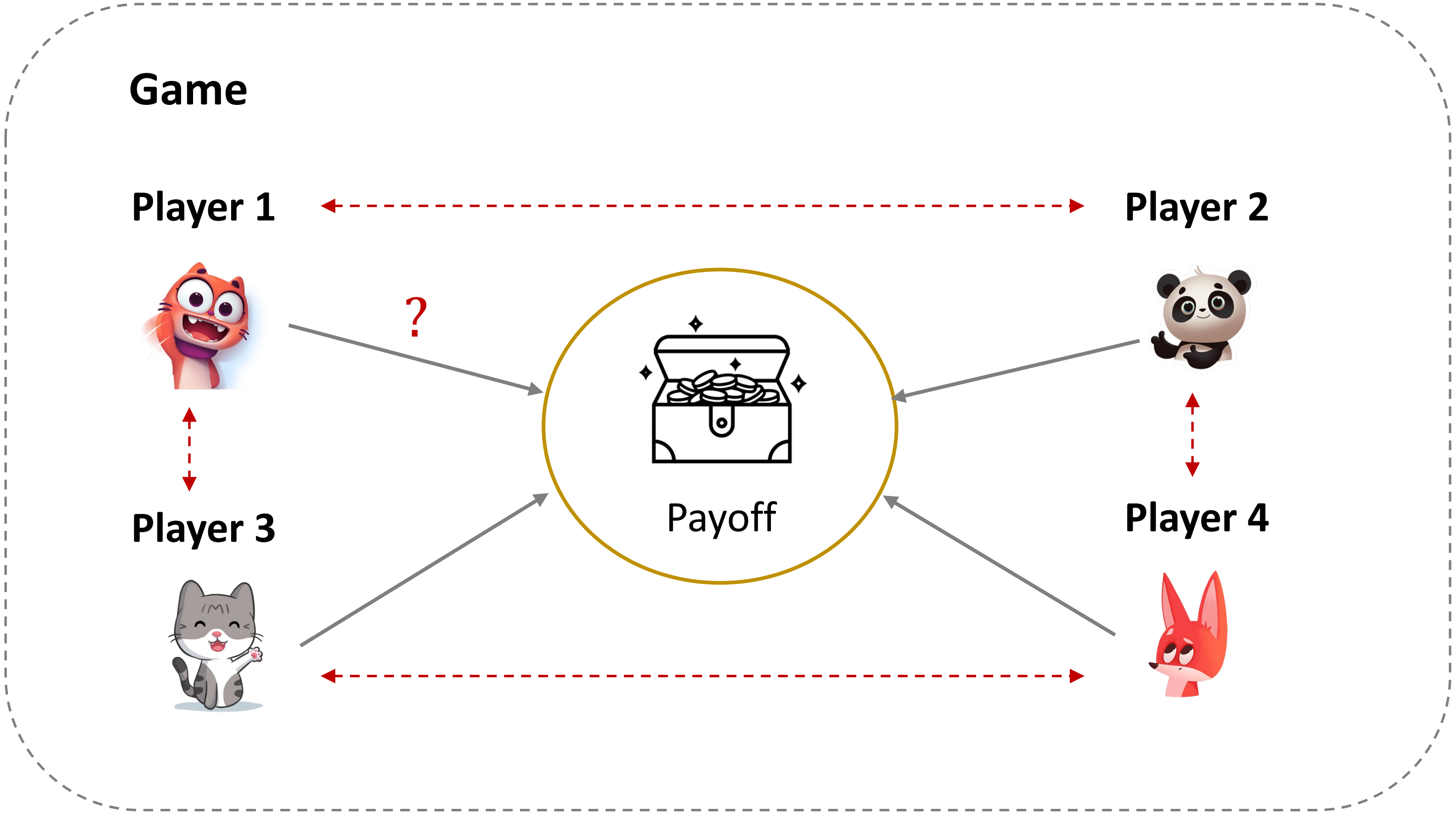
SHAP

- Shapley value   [Shapley, 1953]



SHAP

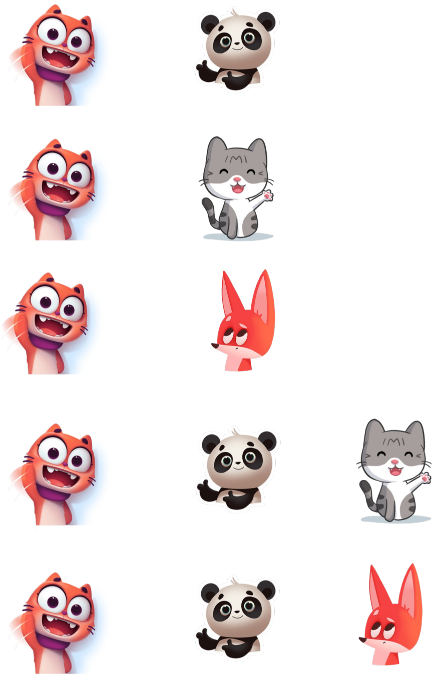
- Shapley value [Shapley, 1953]



SHAP

- Shapley value [Shapley, 1953]

Coalitions **Payoff**



P_1

P_2

P_3

P_4

P_5

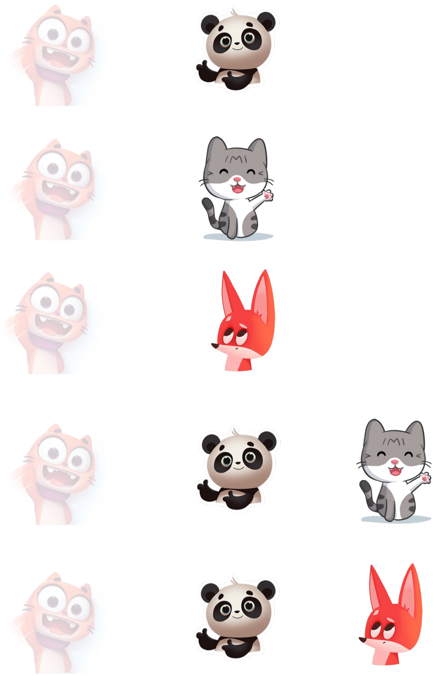
$\vdots$

(2^3)

SHAP

- Shapley value [Shapley, 1953]

Coalitions



⋮

(2^3)

Payoff

P_1 P_1'

P_2 P_2'

P_3 P_3'

P_4 P_4'

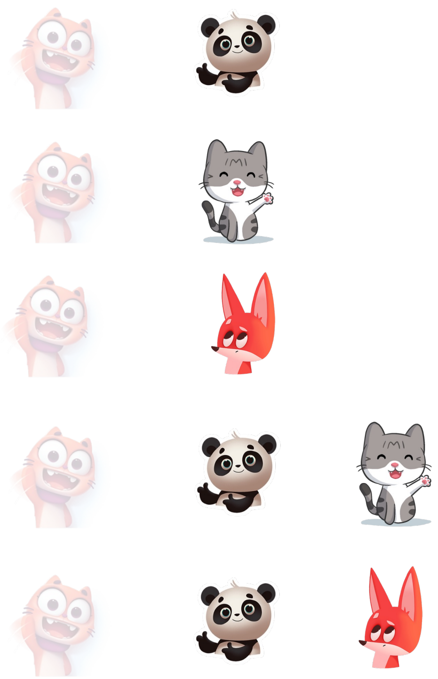
P_5 P_5'

⋮

SHAP

- Shapley value [Shapley, 1953]

Coalitions **Payoff** **Marginal contribution**



$\vdots$

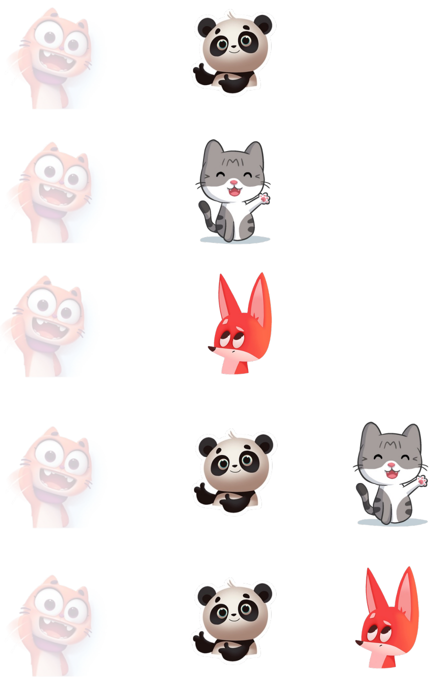
(2^3)

Payoff	Marginal contribution
$P_1 - P_1'$	ΔP_1
$P_2 - P_2'$	ΔP_2
$P_3 - P_3'$	ΔP_3
$P_4 - P_4'$	ΔP_4
$P_5 - P_5'$	ΔP_5
$\vdots$	

SHAP

- Shapley value [Shapley, 1953]

Coalitions



⋮

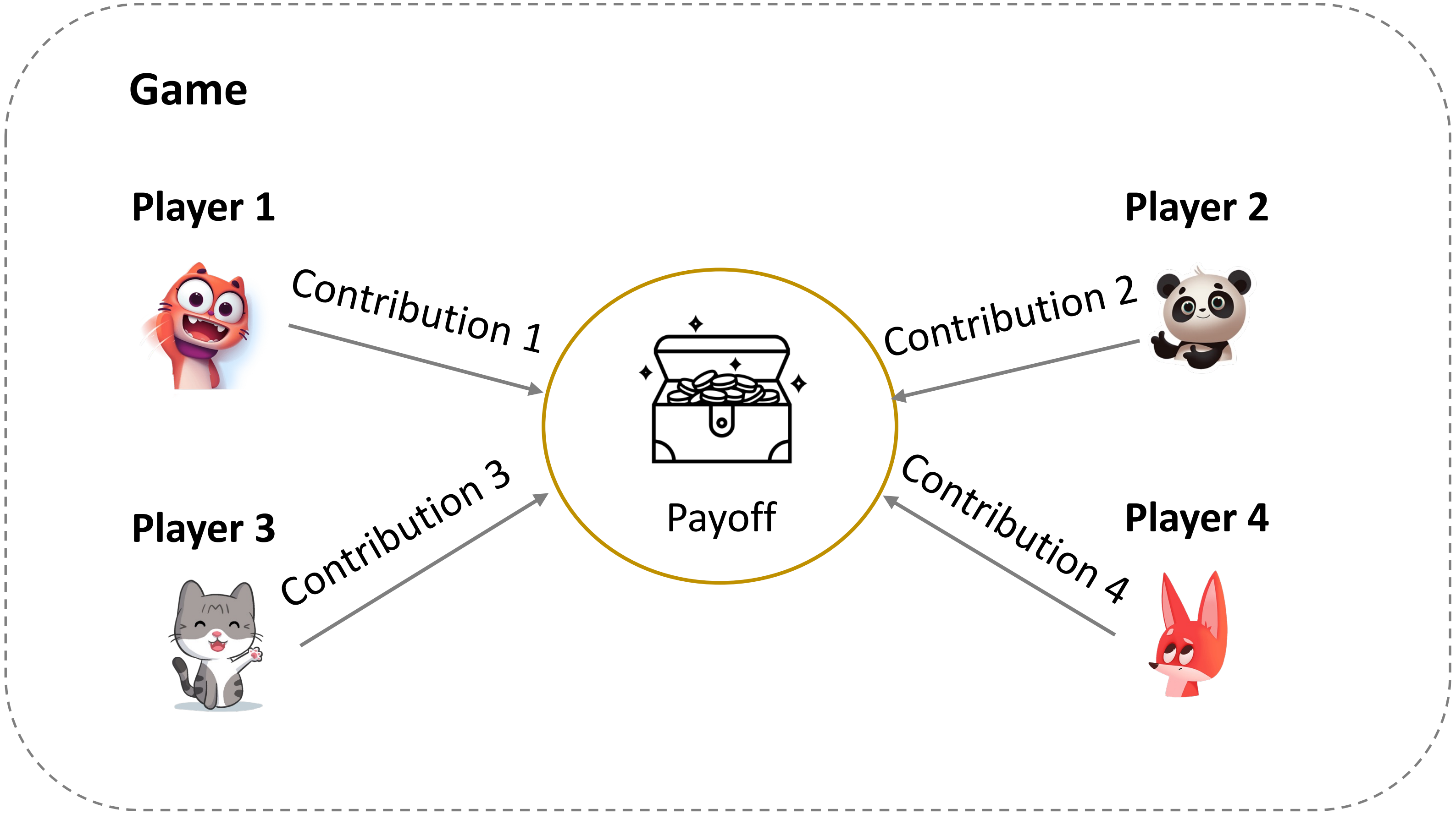
(2^3)

Payoff	**Marginal contribution**
$P_1 - P_1'$	ΔP_1
$P_2 - P_2'$	ΔP_2
$P_3 - P_3'$	ΔP_3
$P_4 - P_4'$	ΔP_4
$P_5 - P_5'$	ΔP_5
⋮	

Contribution$= \sum \Delta P_i$

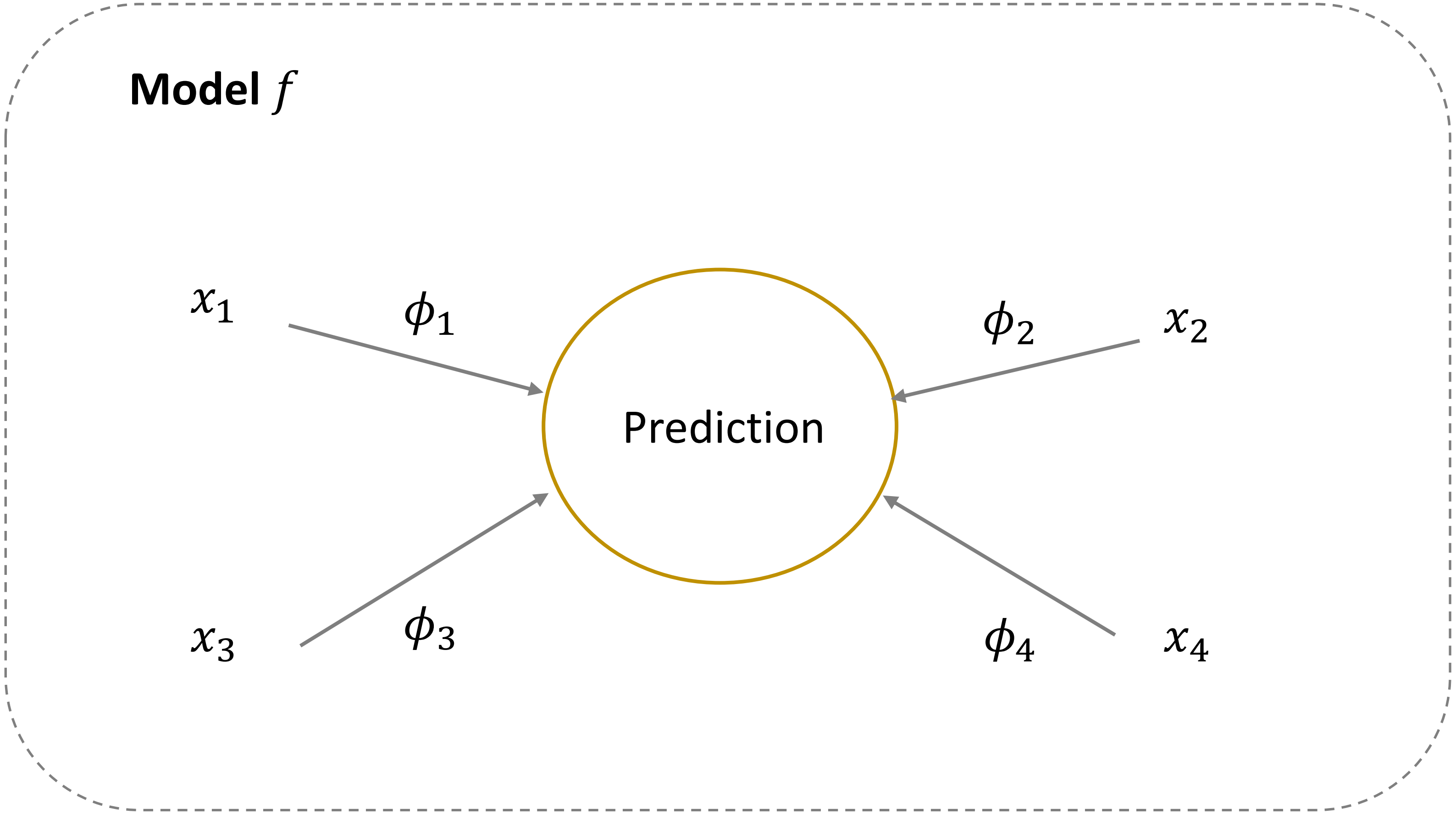
SHAP

- Shapley value [Shapley, 1953]



SHAP

- Shapley value [Shapley, 1953]

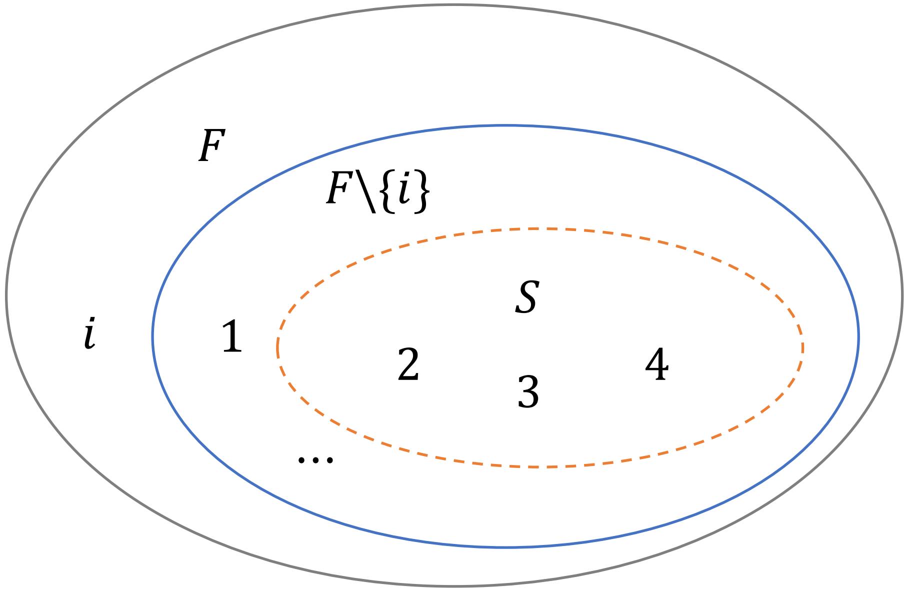


SHAP

- Shapley value [Shapley, 1953]

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!\,(|F| - |S| - 1)!}{|F|!} \underbrace{\left[f_{S \cup \{i\}}\left(x_{S \cup \{i\}}\right) - f_S(x_S)\right]}$$

Marginal contribution of x_i given S

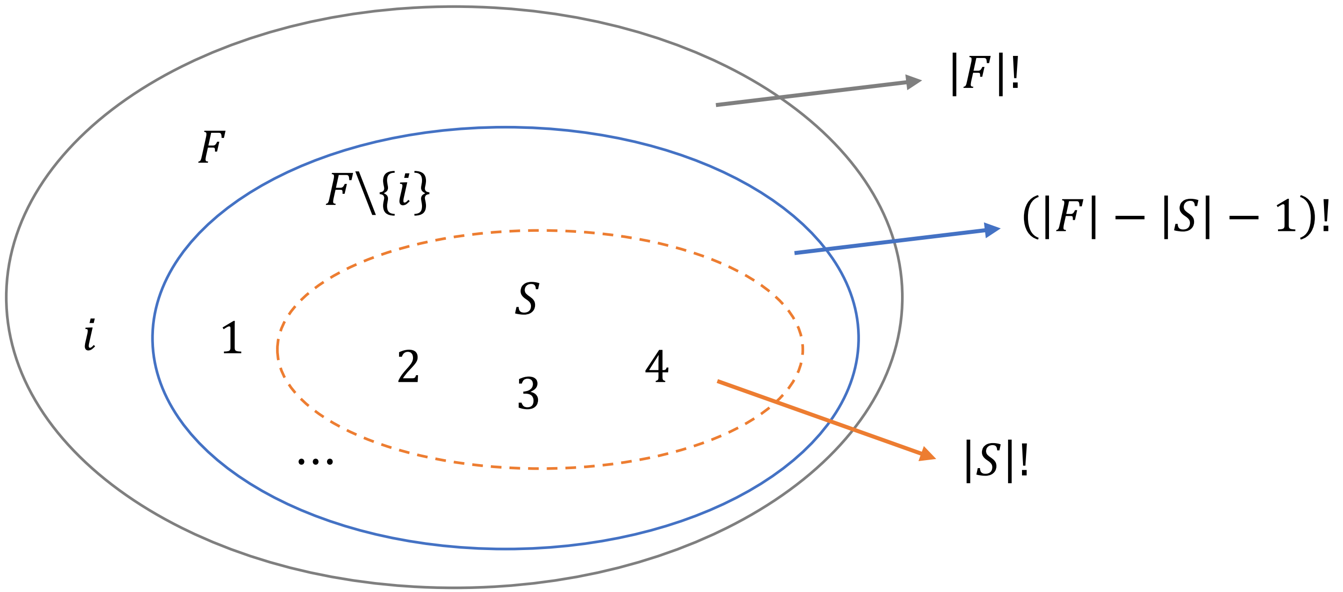


SHAP

- Shapley value [Shapley, 1953]

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!\,(|F| - |S| - 1)!}{|F|!} \left[f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S) \right]$$

Weighted by the permutations of features



SHAP

- SHapley Additive exPlanation (SHAP)

Additive feature attribution method

$$g(z') \approx f\left(h_x(z')\right)$$

$$g(z') = \phi_0 + \sum_{i=1}^{N} \phi_i {z_i}'$$

$$z' \approx x' \qquad \underline{x} = h_x(\underline{x'})$$

Original input Interpretable input

SHAP

- SHapley Additive exPlanation (SHAP)

Additive feature attribution method

$$g(z') \approx f\left(h_x(z')\right)$$

$$z' \approx x' \qquad \underline{x} = h_x(\underline{x'})$$

Original input   Interpretable input

$$g(z') = \phi_0 + \sum_{i=1}^{N} \phi_i {z_i}'$$

LIME is a special case, but not optimal

$$g(z') = \sum_{i=1}^{N} w_i {z_i}'$$

SHAP

- SHapley Additive exPlanation (SHAP)

Additive feature attribution method

$$g(z') \approx f\left(h_x(z')\right)$$

$$z' \approx x' \qquad \underline{x} = h_x(\underline{x'})$$

Original input   Interpretable input

$$g(z') = \phi_0 + \sum_{i=1}^{N} \phi_i {z_i}'$$

❑ **Property 1: Local accuracy**

$$f(x) = g(x') = \phi_0 + \sum_{i=1}^{N} \phi_i {x_i}'$$

$$\phi_0 = h_x(0)$$

SHAP

- SHapley Additive exPlanation (SHAP)

Additive feature attribution method

$$g(z') \approx f\big(h_x(z')\big)$$

$$z' \approx x' \qquad \underline{x} = h_x(\underline{x'})$$

Original input Interpretable input

$$g(z') = \phi_0 + \sum_{i=1}^{N} \phi_i {z_i}'$$

❑ **Property 2: Missingness**

$$x_i' = 0 \quad \Longrightarrow \quad \phi_i = 0$$

Missingness constrains features missing in the original input to have no attributed impact

SHAP

- SHapley Additive exPlanation (SHAP)

Additive feature attribution method

$$g(z') \approx f\left(h_x(z')\right)$$

$z' \approx x'$ $\qquad$ $\underline{x} = h_x(\underline{x'})$

Original input $\quad$ Interpretable input

$$g(z') = \phi_0 + \sum_{i=1}^{N} \phi_i {z_i}'$$

❑ **Property 3: Consistency**

For any two models f_1 and f_2, if $f_1\left(h_x(z')\right) - f_1\left(h_x(\underline{z'\backslash i})\right) \geq f_2\left(h_x(z')\right) - f_2\left(h_x(z'\backslash i)\right)$

$z_i' = 0$

for all inputs $z' \in \{0, 1\}^N$, then $\phi_i(f_1, x) \geq \phi_i(f_2, x)$

SHAP

- SHapley Additive exPlanation (SHAP)

Additive feature attribution method

$$g(z') \approx f\left(h_x(z')\right)$$

$z' \approx x'$ $\qquad \underbrace{x}_{\text{Original input}} = h_x(\underbrace{x'}_{\text{Interpretable input}})$

$$g(z') = \phi_0 + \sum_{i=1}^{N} \phi_i {z_i}'$$

Only Shapley value satisfies all the three properties

$$\phi_i(f, x) = \sum_{z' \subseteq x'} \frac{|z'|!\,(N - |z'| - 1)!}{N!} \left[f\left(h_x(z')\right) - f\left(h_x(z' \backslash i)\right)\right]$$

Contains a subset of non-zero entries in x'

SHAP

- SHapley Additive exPlanation (SHAP)

$$\phi_i(f,x) = \sum_{z' \subseteq x'} \frac{|z'|!\,(N - |z'| - 1)!}{N!}\left[f\left(h_x(z')\right) - f\left(h_x(z' \backslash i)\right)\right]$$

Challenge

Computational complexity

$O(2^n)$

SHAP

- SHapley Additive exPlanation (SHAP)

Model-agnostic approximations

- Shapley sampling values
- Kernel SHAP

Model-type-specific approximations

- Linear SHAP
- Low-Order SHAP
- Max SHAP
- Deep SHAP

SHAP

- SHapley Additive exPlanation (SHAP)

Model-agnostic approximations

- $\boxed{\text{Shapley sampling values}}$
- Kernel SHAP

Model-type-specific approximations

- Linear SHAP
- Low-Order SHAP
- Max SHAP
- Deep SHAP

Initialize the number of samples M

$\phi_i \longleftarrow 0$

for $m \in \{1, \cdots, M\}$ **do**

Sample $z' \subseteq x'$

$$\phi_i \longleftarrow \phi_i + \frac{|z\prime|!(N-|z\prime|-1)!}{N!}\left[f\left(h_x(z')\right) - f\left(h_x(z' \backslash i)\right)\right]$$

SHAP

- SHapley Additive exPlanation (SHAP)

Model-agnostic approximations

- Shapley sampling values
- Kernel SHAP   Linear LIME + Shapley values

Model-type-specific approximations

- Linear SHAP
- Low-Order SHAP
- Max SHAP
- Deep SHAP

The solutions would be consistent with properties 1-3

$$\Omega(g) = 0$$

$$\pi_{x'}(z') = \frac{(N-1)}{(N\ choose\ |z'|)|z'|(N-|z'|)}$$

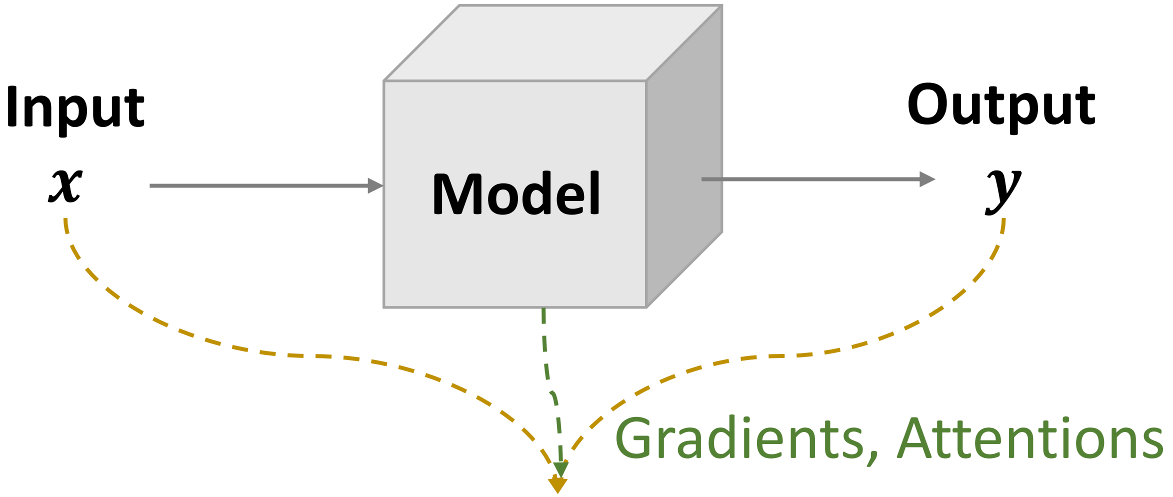
$$\mathcal{L}(f,g) = \sum \pi_{x'}(z')\left(f\left(h_x(z')\right) - g(z')\right)^2$$

Question?

Improving Interpretability

- Black-box explanation
- White-box explanation
- Natural language explanation

White-box Explanation



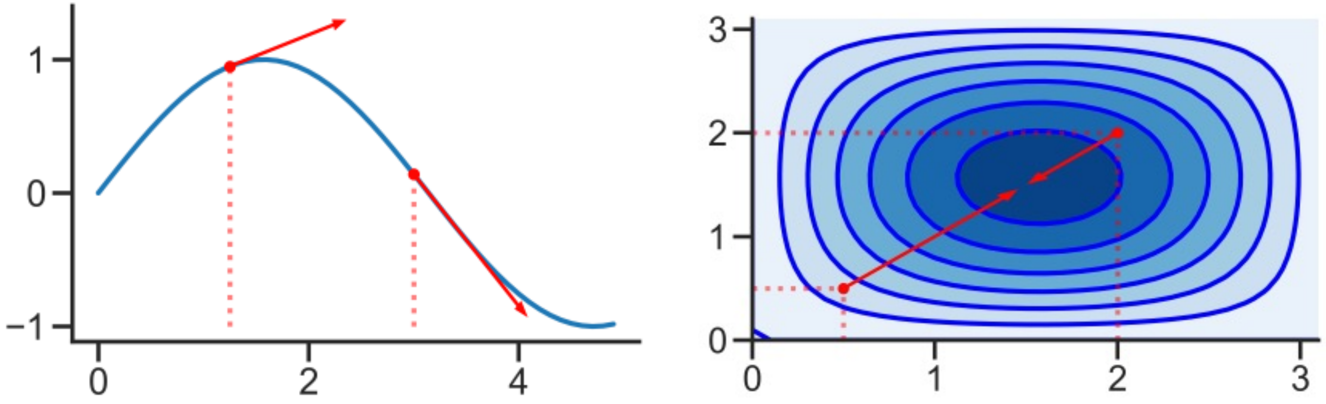
Explanation

- Simple, efficient
- Need access

Gradient-based Explanation

The gradient of a function f on $\boldsymbol{x} \in \mathbb{R}^n$ is

$$\nabla f(\boldsymbol{x}) = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{bmatrix}$$



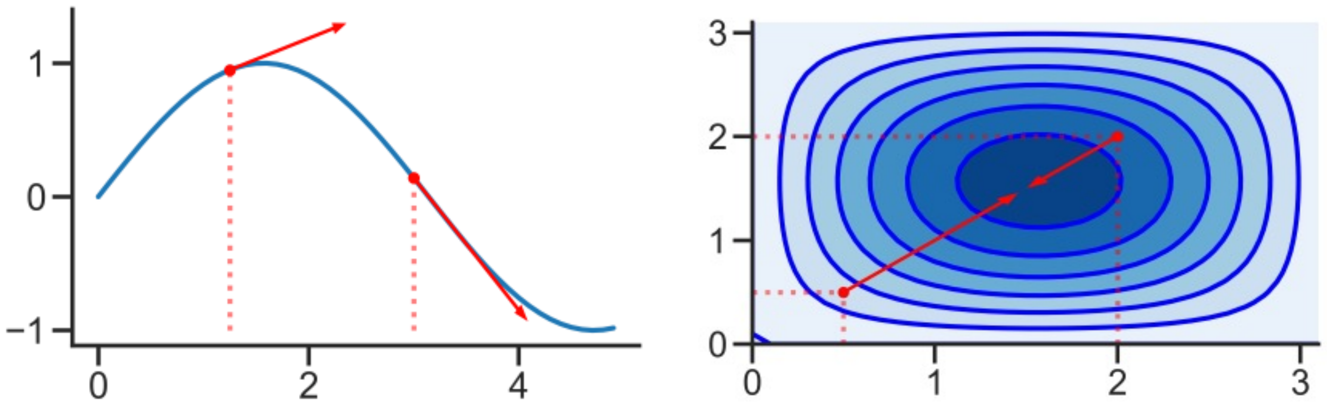
Source: https://towardsdatascience.com/basics-gradient-input-as-explanation-bca79bb80de0

Gradient-based Explanation

The gradient of a function f on $\boldsymbol{x} \in \mathbb{R}^n$ is

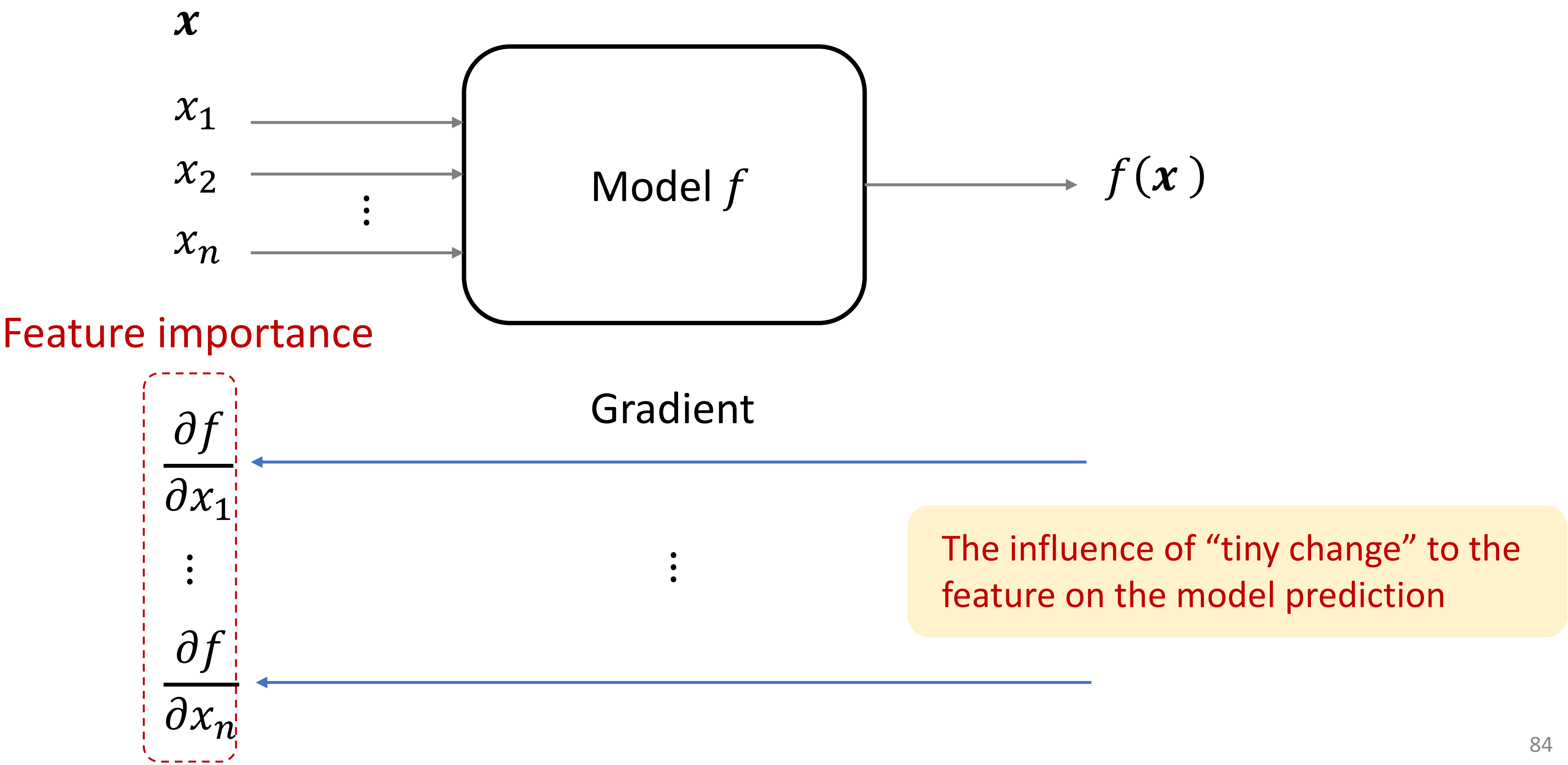
$$\nabla f(\boldsymbol{x}) = \begin{bmatrix} \dfrac{\partial f}{\partial x_1} \\ \vdots \\ \dfrac{\partial f}{\partial x_n} \end{bmatrix}$$

The derivative $\frac{\partial f}{\partial x_i}$ indicates how much f will change when x_i increases a little bit

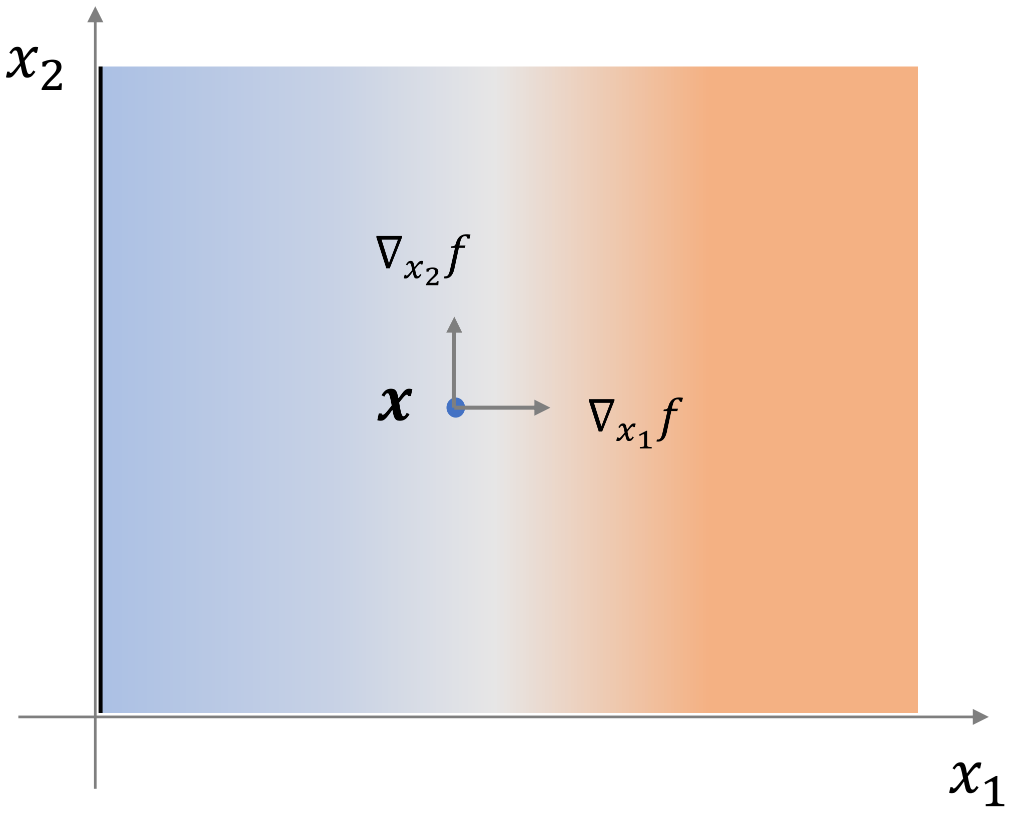


Source: https://towardsdatascience.com/basics-gradient-input-as-explanation-bca79bb80de0

Gradient-based Explanation

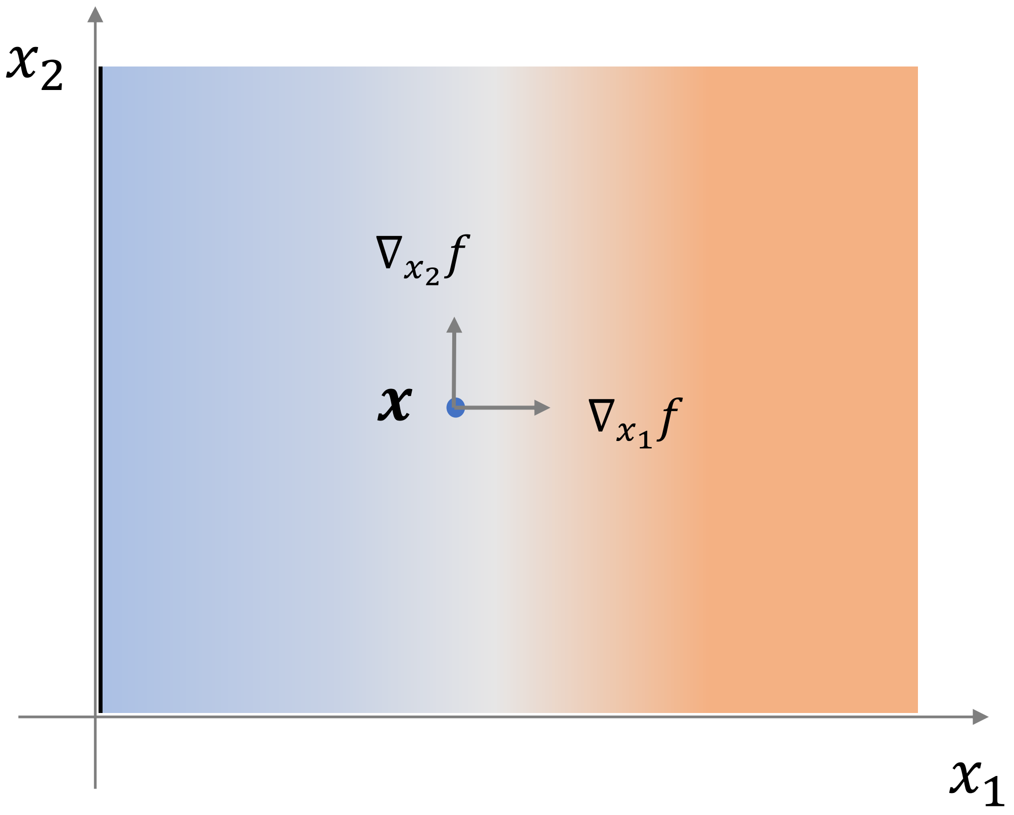


Gradient-based Explanation



Which feature is more important?

Gradient-based Explanation



x_1 is more important than x_2

✓ Changing x_1 can flip the model prediction

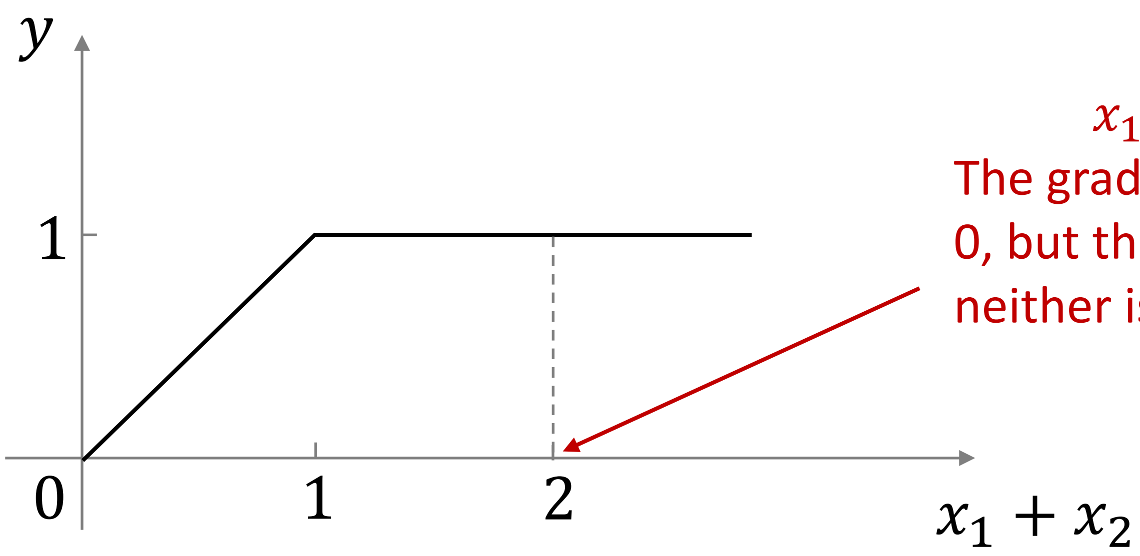
✓ Changing x_2 would not influence the model prediction

Question?

Gradient-based Explanation

Problem 1: saturated outputs lead to unintuitive gradients

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



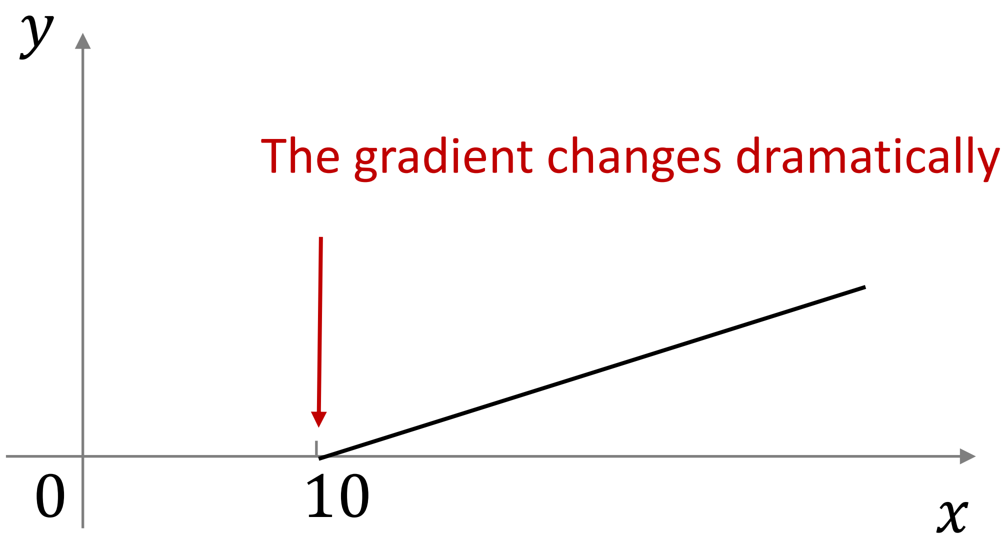
$x_1 = 1, x_2 = 1$
The gradient on x_1 or x_2 is 0, but that does not mean neither is important

(Shrikumar et al., 2017)

Gradient-based Explanation

Problem 2: discontinuous gradients (e.g., thresholding) are problematic

$$y = max(0, x - 10)$$

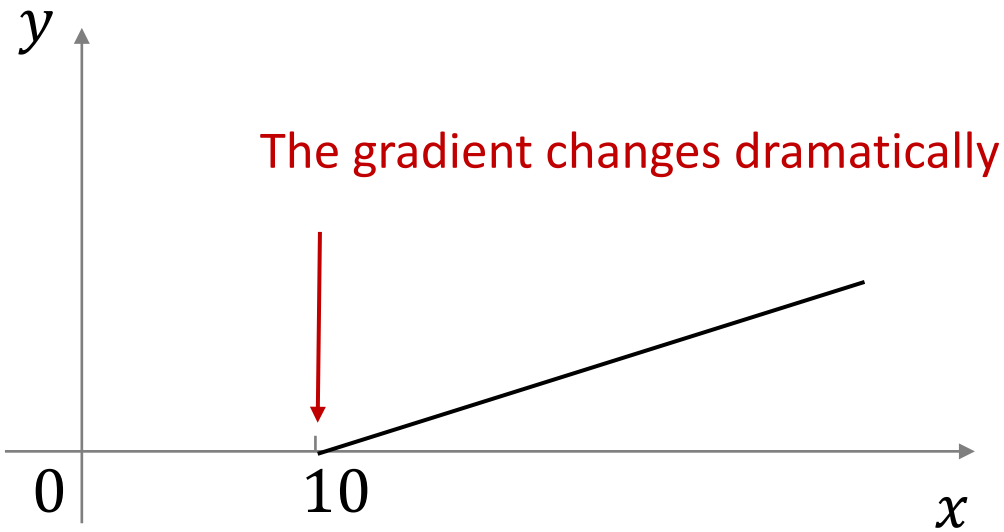


(Shrikumar et al., 2017)

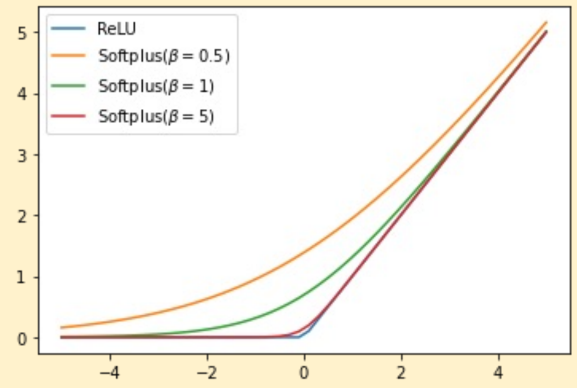
Gradient-based Explanation

Problem 2: discontinuous gradients (e.g., thresholding) are problematic

$$y = max(0, x - 10)$$



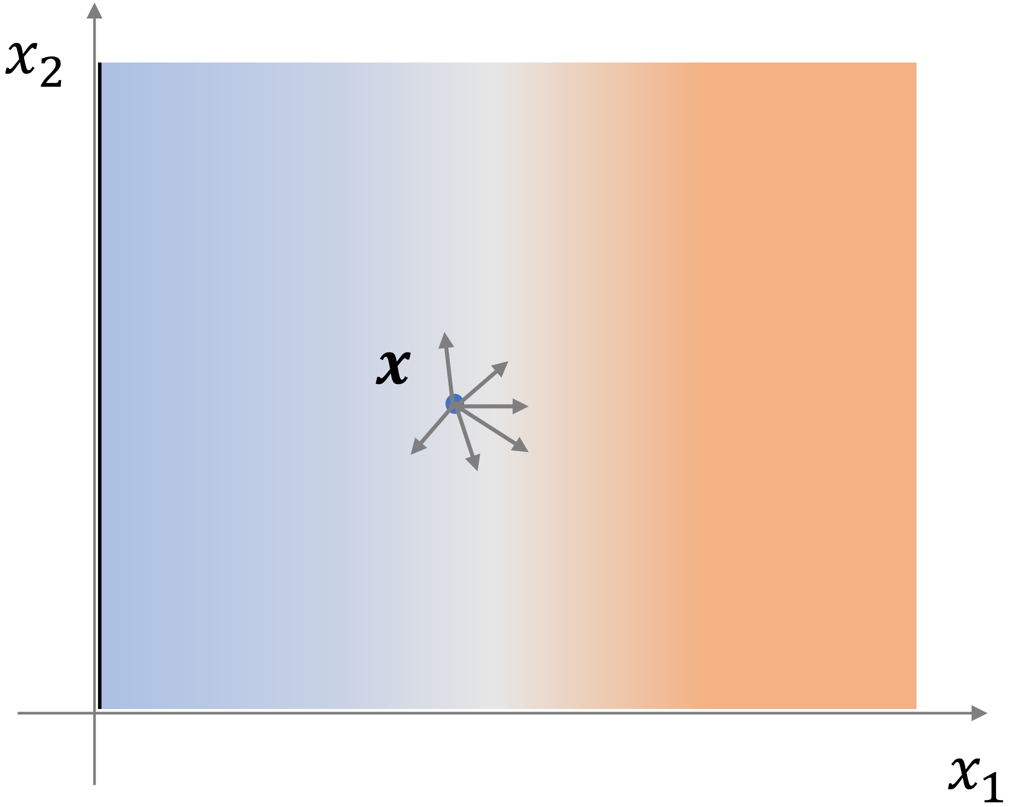
Need to replace "Relu" with "Softplus" activation



(Shrikumar et al., 2017)

Gradient-based Explanation

Problem 3: input gradient is sensitive to slight perturbations

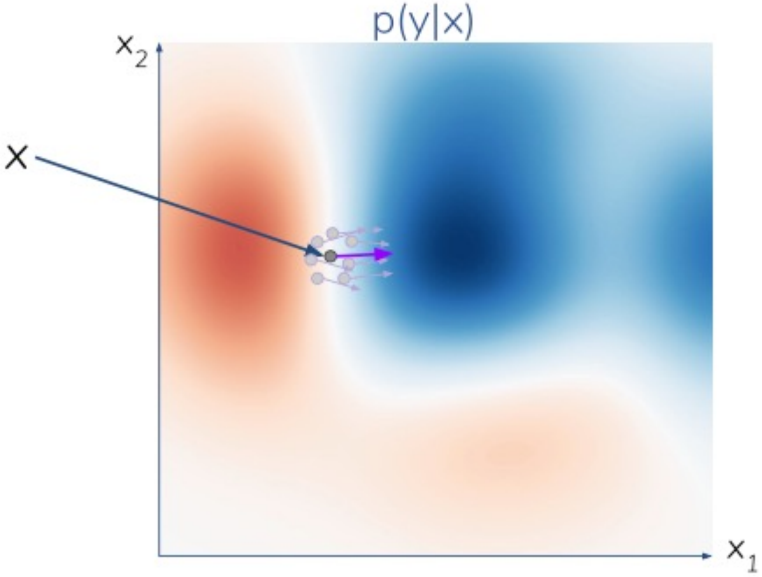


Gradient-based Explanation

Do NOT rely on a single gradient calculation

- SmoothGrad: add gaussian noise to inputs and average the gradients

(Smilkov et al., 2017)

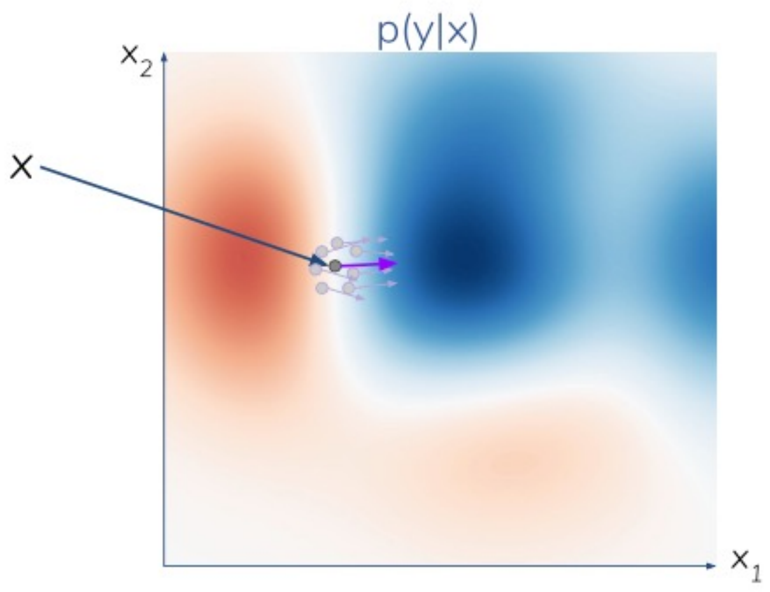


Gradient-based Explanation

Do NOT rely on a single gradient calculation

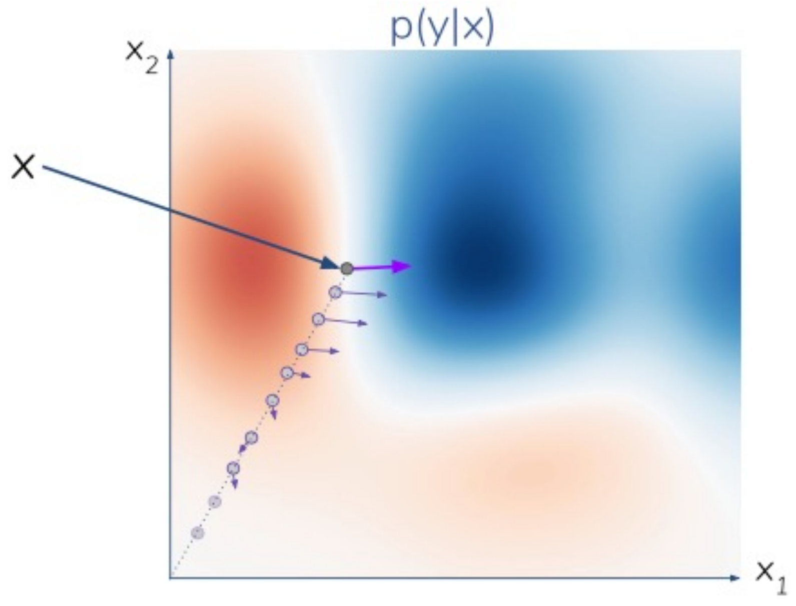
- SmoothGrad: add gaussian noise to inputs and average the gradients

(Smilkov et al., 2017)



- Integrated Gradients: aggregate gradients along a path from baseline to the input

(Sundararajan et al., 2017)



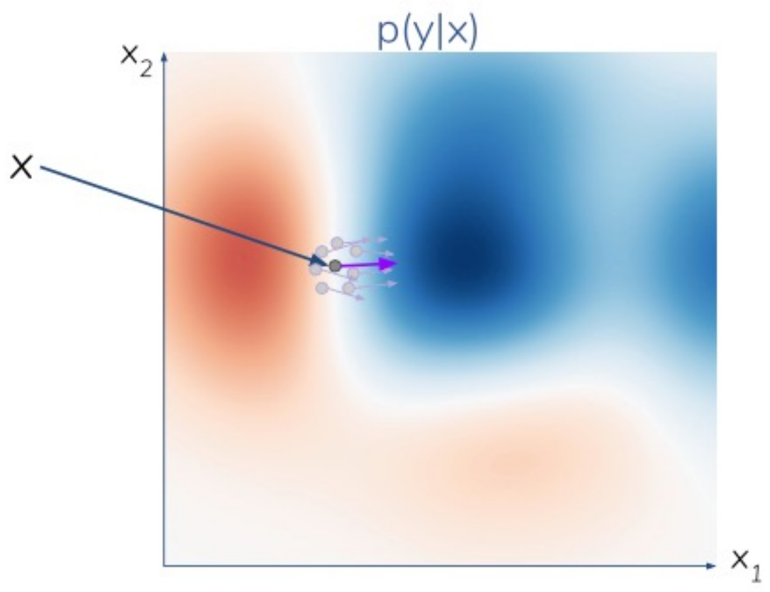
Source: EMNLP 2020 Tutorial on Interpreting Predictions of NLP Models

Gradient-based Explanation

Do NOT rely on a single gradient calculation

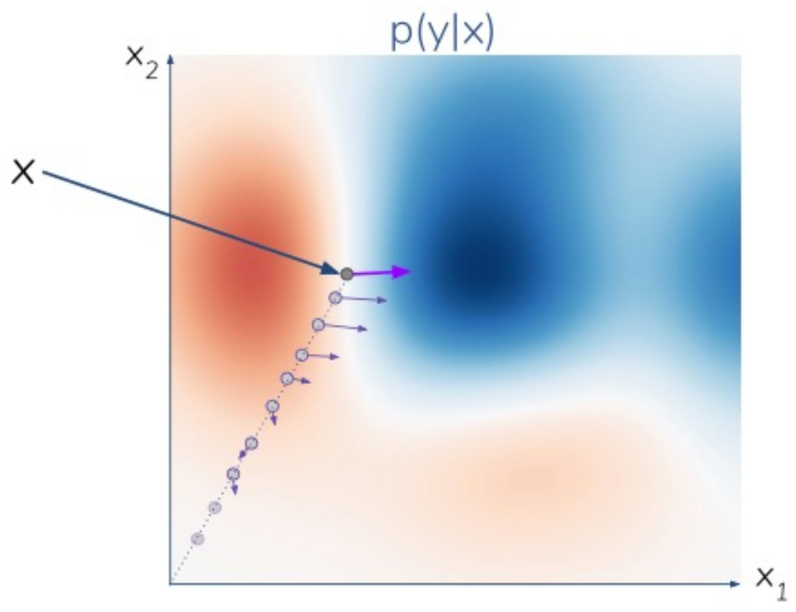
- SmoothGrad: add gaussian noise to inputs and average the gradients

  (Smilkov et al., 2017)



- Integrated Gradients: aggregate gradients along a path from baseline to the input

  (Sundararajan et al., 2017)



Source: EMNLP 2020 Tutorial on Interpreting Predictions of NLP Models

IG

Axiomatic Attribution for Deep Networks

Mukund Sundararajan, Ankur Taly, Qiqi Yan

(ICML, 2017)

IG

- Integrated Gradients

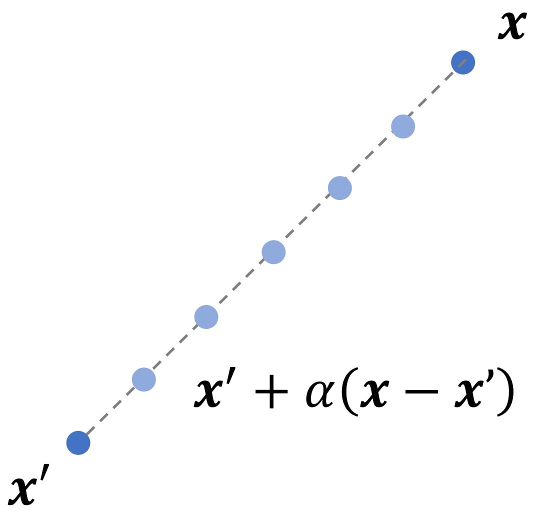
f: neural network

$\boldsymbol{x} \in \mathbb{R}^n$: input

$\boldsymbol{x}' \in \mathbb{R}^n$: baseline

(e.g., zero embedding vector)

Get samples along the straight line from x' to x



$\boldsymbol{x}' + \alpha(\boldsymbol{x} - \boldsymbol{x}') \quad \alpha \in (0, 1)$

IG

- Integrated Gradients

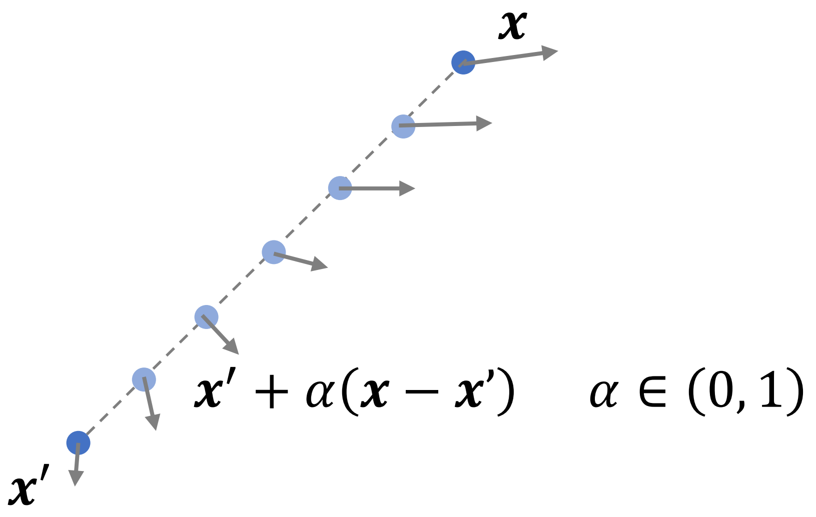
f: neural network

$\boldsymbol{x} \in \mathbb{R}^n$: input

$\boldsymbol{x}' \in \mathbb{R}^n$: baseline

(e.g., zero embedding vector)

Compute gradients at all points along the path



IG

- Integrated Gradients

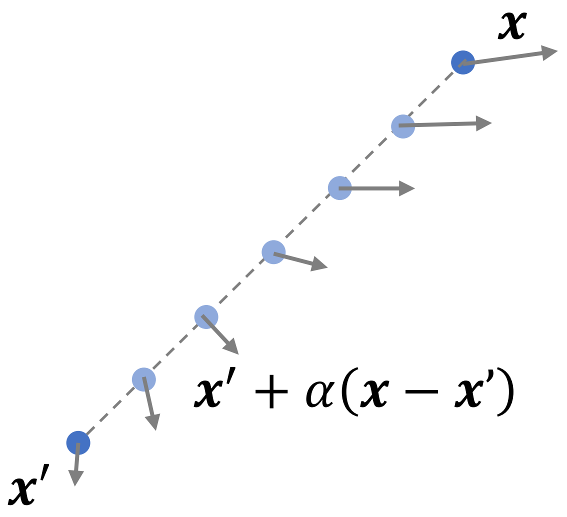
f: neural network

$\boldsymbol{x} \in \mathbb{R}^n$: input

$\boldsymbol{x}' \in \mathbb{R}^n$: baseline

(e.g., zero embedding vector)

Cumulate these gradients



$\boldsymbol{x}' + \alpha(\boldsymbol{x} - \boldsymbol{x}')$ $\quad \alpha \in (0, 1)$

$$IG_i(\boldsymbol{x}) = (x_i - x_i') \times \int_{\alpha=0}^{1} \frac{\partial f\left(\boldsymbol{x}' + \alpha(\boldsymbol{x} - \boldsymbol{x}')\right)}{\partial x_i} d\alpha$$

On the ith dimension

IG

- Integrated Gradients

Axiom: completeness

The attributions add up to the difference between the output of f at the input $\boldsymbol{x}$ and the baseline $\boldsymbol{x}'$

$$\sum_{i=1}^{n} IG_i(\boldsymbol{x}) = f(\boldsymbol{x}) - \underline{f(\boldsymbol{x}')}$$

$$f(\boldsymbol{x}') \approx 0$$

IG

- Integrated Gradients

Axiom: completeness

The attributions add up to the difference between the output of f at the input $\boldsymbol{x}$ and the baseline $\boldsymbol{x}'$

$$\sum_{i=1}^{n} IG_i(\boldsymbol{x}) = f(\boldsymbol{x}) - f(\boldsymbol{x}')$$

Sensitivity: for every input and baseline that differ in one feature but have different predictions then the differing feature should be given a non-zero attribution

✅ Sensitivity

IG

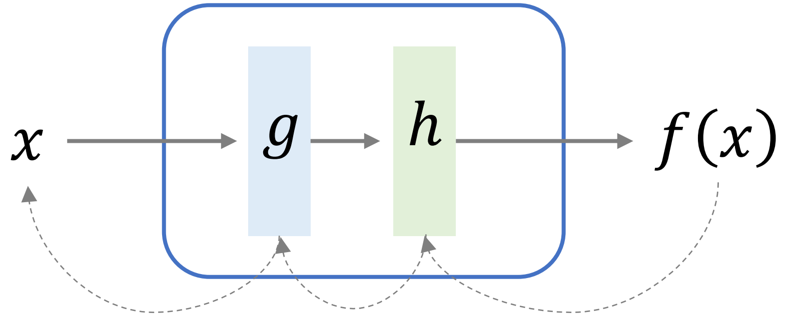
- Integrated Gradients

Axiom: completeness

The attributions add up to the difference between the output of f at the input $\boldsymbol{x}$ and the baseline $\boldsymbol{x}'$

$$\sum_{i=1}^{n} IG_i(\boldsymbol{x}) = f(\boldsymbol{x}) - f(\boldsymbol{x}')$$

The chain-rule for gradients is essentially about implementation invariance:



$$\frac{\partial f}{\partial x} = \frac{\partial f}{\partial h} \cdot \frac{\partial h}{\partial g} \cdot \frac{\partial g}{\partial x}$$

✓ Sensitivity

✓ Implementation invariance

(The attributions are always identical for two functionally equivalent networks)

IG

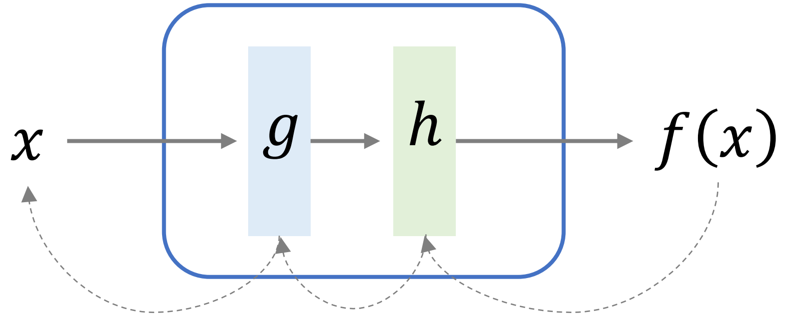
- Integrated Gradients

Axiom: completeness

The attributions add up to the difference between the output of f at the input $\boldsymbol{x}$ and the baseline $\boldsymbol{x}'$

$$\sum_{i=1}^{n} IG_i(\boldsymbol{x}) = f(\boldsymbol{x}) - f(\boldsymbol{x}')$$

The chain-rule for gradients is essentially about implementation invariance:



$$\frac{\partial f}{\partial x} = \frac{\partial f}{\cancel{\partial h}} \cdot \frac{\cancel{\partial h}}{\cancel{\partial g}} \cdot \frac{\cancel{\partial g}}{\partial x}$$

- ✓ Sensitivity
- ✓ Implementation invariance
  (The attributions are always identical for two functionally equivalent networks)

IG

- Applying Integrated Gradients

The integral of integrated gradients can be efficiently approximated via a summation

$$IG_i(\boldsymbol{x}) \approx (x_i - x_i{}') \times \sum_{k=1}^{m} \frac{\partial f\left(\boldsymbol{x}' + \frac{k}{m}(\boldsymbol{x} - \boldsymbol{x}')\right)}{\partial x_i} \times \frac{1}{m}$$

m: the number of steps

Question?

Improving Interpretability

- Black-box explanation
- White-box explanation
- Natural language explanation

Natural Language Explanation

Commonsense question-answering (QA)

Question

Why do people go hiking?

Answer choices

drink water | get lost | enjoy nature | lose weight | get tired

Prediction: enjoy nature

Explanation: Hiking means the activity of going for long walks especially across country, or in nature. People who go hiking enjoy nature.

Natural Language Explanation

Commonsense question-answering (QA)

Question

Why do people go hiking?

Answer choices

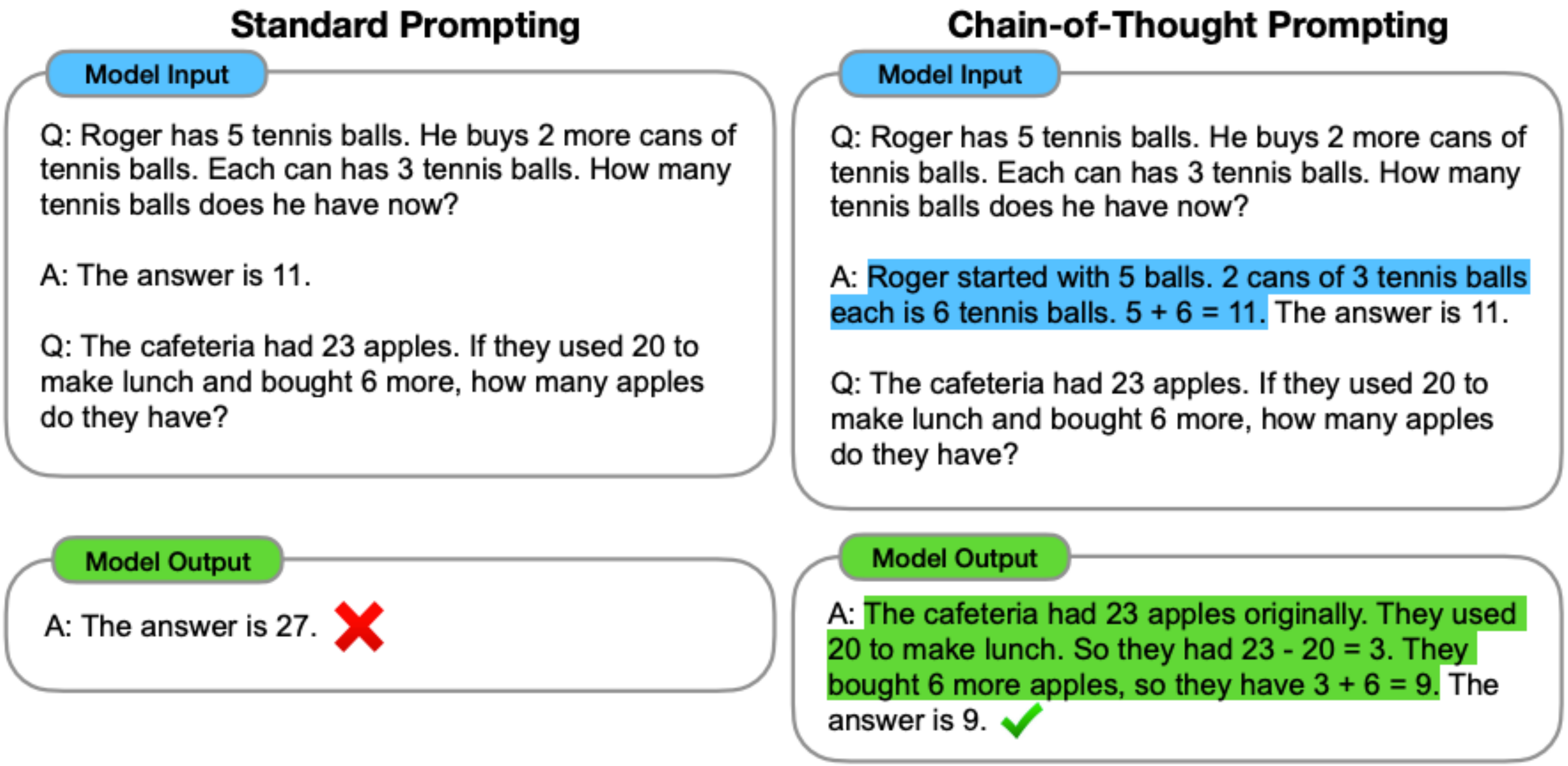
drink water | get lost | enjoy nature | lose weight | get tired

Prediction: enjoy nature

Explanation: Hiking means the activity of going for long walks especially across country, or in nature. People who go hiking enjoy nature.

- Flexible
- Understandable
- Informative

Chain of Thought Prompting



(Wei et al., 2022)

Potential Issues

Why do people go hiking?

Hiking means the activity of going for long walks especially across country, or in nature. People who go hiking enjoy nature.

Getting lost in the wilderness is a valuable experience. People go hiking to get lost.

Not factual

Hiking in nature helps men get rid of jobs. Men go hiking to enjoy nature.

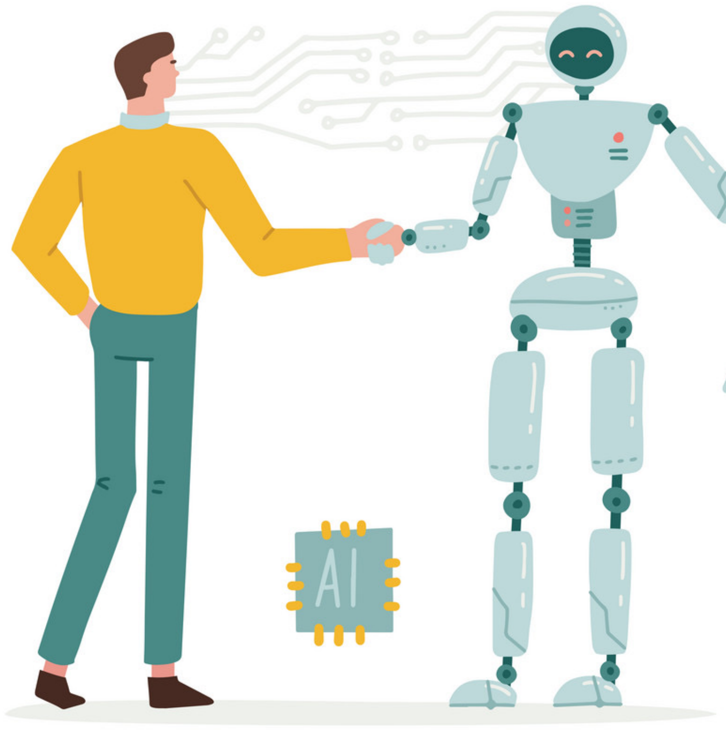
Bias

Hiking in nature helps women get rid of housework. Women go hiking to enjoy nature.

Question?

Improving Interpretability

- Black-box explanation
- White-box explanation
- Natural language explanation



Thank you!

Reference

- Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. " Why should i trust you?" Explaining the predictions of any classifier." *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 2016.
- Lundberg, Scott M., and Su-In Lee. "A unified approach to interpreting model predictions." Proceedings of the 31st international conference on neural information processing systems. 2017.
- Li, Jiwei, Will Monroe, and Dan Jurafsky. "Understanding neural networks through representation erasure." arXiv preprint arXiv:1612.08220 (2016).
- Lloyd S Shapley. 1953. A value for n-person games. Contributions to the Theory of Games, 2(28).
- Shrikumar, Avanti, Peyton Greenside, and Anshul Kundaje. "Learning important features through propagating activation differences." *International conference on machine learning*. PMLR, 2017.
- Sundararajan, Mukund, Ankur Taly, and Qiqi Yan. "Axiomatic attribution for deep networks." *International conference on machine learning*. PMLR, 2017.
- Wei, Jason, et al. "Chain-of-thought prompting elicits reasoning in large language models." Advances in Neural Information Processing Systems 35 (2022): 24824-24837.