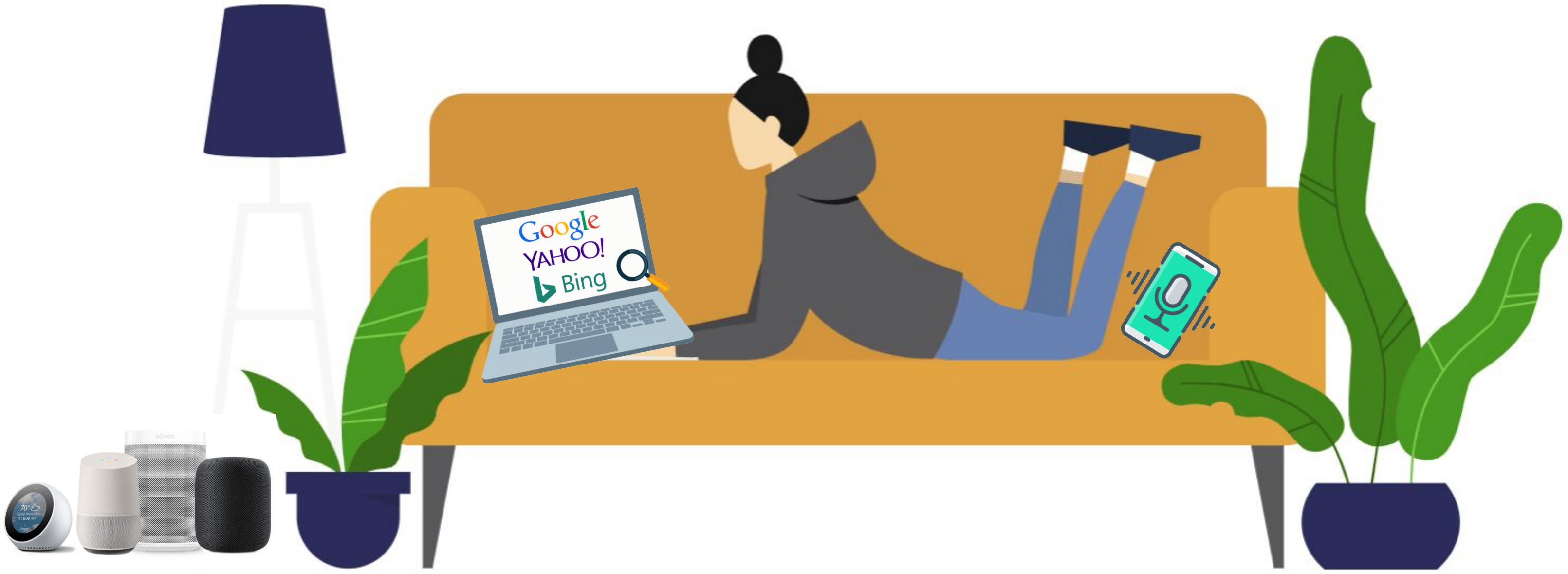
A Quick Tour of NLP Explainability

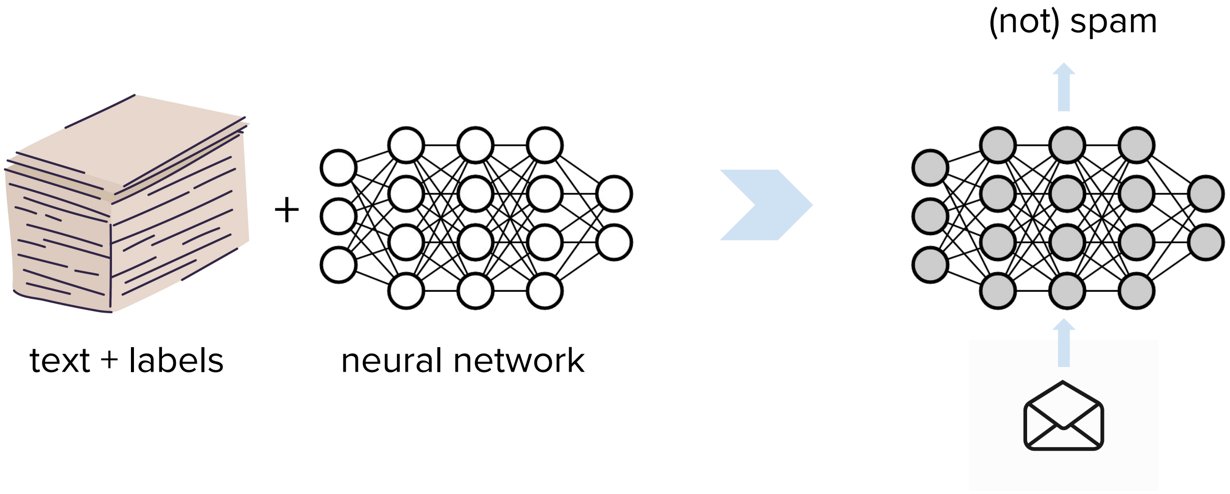
Ana Marasović

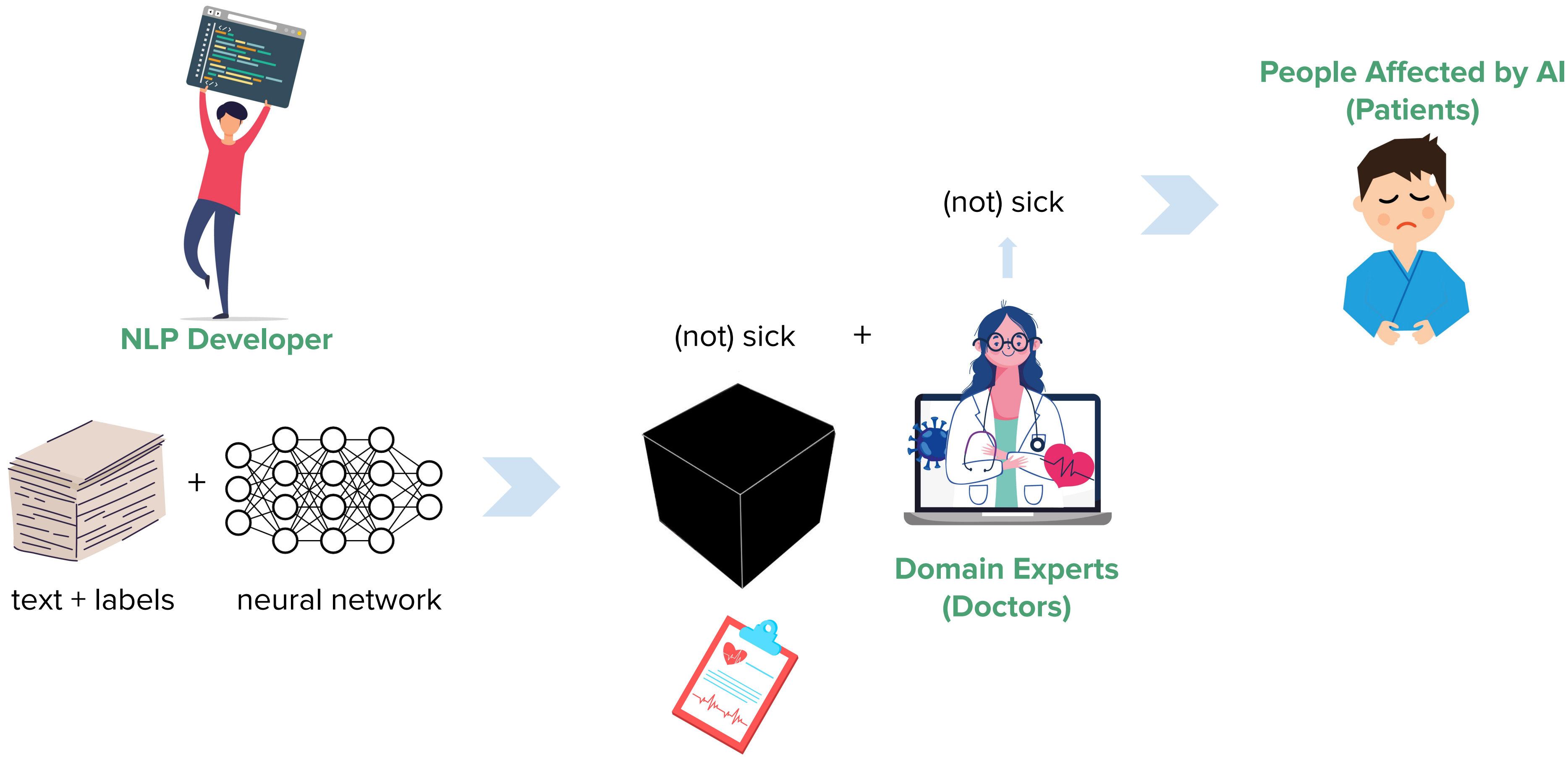
Allen Institute for AI (AI2) × AllenNLP × University of Washington

NLP technology has become an integral part of most people's daily lives




SPAM


(not) spam
+
text + labels
neural network


NLP Developer
text + labels
+
neural network
(not) sick
+
Domain Experts
(Doctors)
(not) sick
People Affected by AI
(Patients)

Increasingly harder to opt out

	Promised Benefits	Risks
Doctors	➔ Faster diagnosis ➔ Better treatment ➔ Less burnout & stress	➔ Hurt their patients ➔ Bad performance review ➔ Getting fired ➔ Lawsuits
Patients	➔ Faster diagnosis ➔ Better treatment	➔ Delayed care ➔ Wrong treatment ➔ Death

Why is this input assigned this answer?

How to change the answer?

Lecture Outline

- Why did my model make this prediction?
  - Part I: Gradient-Based Highlighting
  - Part II: Free-Text Explanations
  - Part III: Influential Train Examples

- Why did my model predict P rather than Q?
  - Contrastive Editing

- How and who explanations help?

Why did my model make this prediction?

Part I: Gradient-Based Highlighting

Slides for this part are copied & slightly modified from
the EMNLP tutorial “Interpreting Predictions of NLP Models”:
https://github.com/Eric-Wallace/interpretability-tutorial-emnlp2020

Thanks to the tutorial creators **Eric Wallace**, **Matt Gardner**, & **Sameer Singh!**

Why did my model make this prediction?

$\Big\downarrow$ *highlighting*

Which parts of the input are responsible for this prediction?

Highlighting methods highlight input features (pixels, words, ect.) that were important for a model prediction

Input highlights are also known as:

1. Saliency maps (for images)
2. Sensitivity maps (for images)
3. Input (feature) attribution
4. Input feature importance
5. Input feature relevance
6. Input feature contribution
7. Extractive rationales

Highlighting Techniques in General

- ➔ Compute the relative "*importance*" of each token in the input
- ➔ "Importance" is, loosely:
  if you change or remove the token, how much is the prediction affected?

Examples of Highlights:

Sentiment an intelligent fiction about learning through cultural clash.

MLM [CLS] The [MASK] ran to the emergency room to see her patient . [SEP]

[Ribeiro et al. 2016, Murdoch et al. 2018, Wallace et al. 2019]

- ➔ Compute the relative “*importance*” of each token in the input
- ➔ “Importance” is, loosely:
  if you change or remove the token, how much is the prediction affected?

“Importance” is measured with:

1. Gradients magnitudes
2. Attention scores
3. Input perturbations

…

- ➔ Compute the relative “*importance*” of each token in the input
- ➔ “Importance” is, loosely:
  if you change or remove the token, how much is the prediction affected?

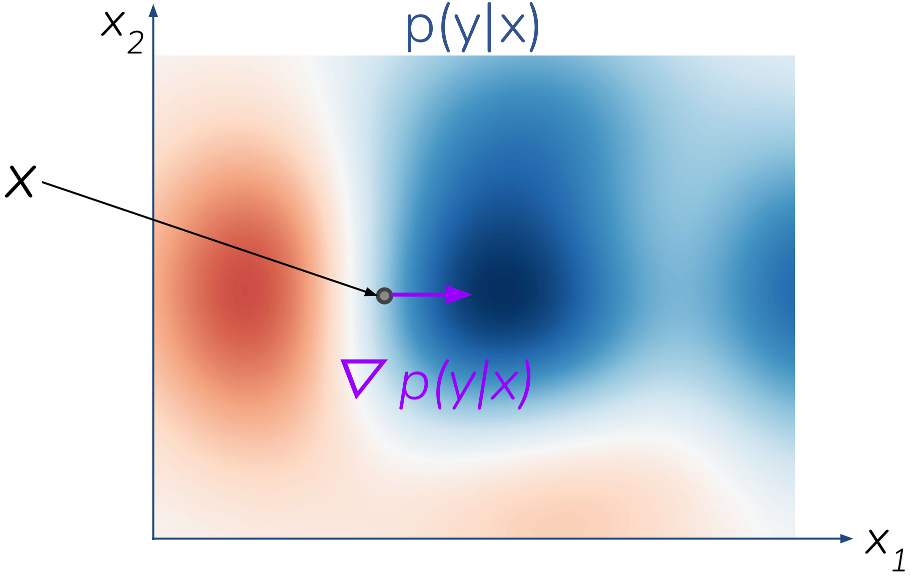
“Importance” is measured with:

1. **Gradient magnitudes**
2. Attention scores
3. Input perturbations

   …

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]

Gradient-based Highlights 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)

Word 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?

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

Eqn from [Han et al. 2020]

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
- Not “customizable”
    - small changes in a individual “token” are not necessarily meaningful
    - distance is implicitly Euclidean (L_2)
- Gradients can be unintuitive with saturated or thresholded values
- Difficult to apply to non-classification tasks

Lecture Outline

- Why did my model make this prediction?
  - Part I: Gradient-Based Highlighting
  - Part II: Free-Text Explanations
  - Part III: Influential Train Examples

- Why did my model predict P rather than Q?
  - Contrastive Editing

- How and who explanations help?

Why did my model make this prediction?

Part II: Free-Text Explanations

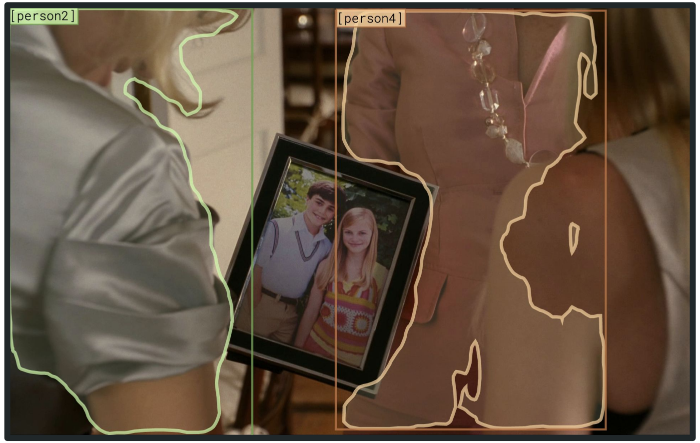
Why did my model make this prediction?

Free-text explanations

Answer in plain English that immediately gives the gist of the reasoning

Answering “why” by highlighting…

…doesn’t work when the reason is not explicitly stated in the input



[Zellers et al., 2019]

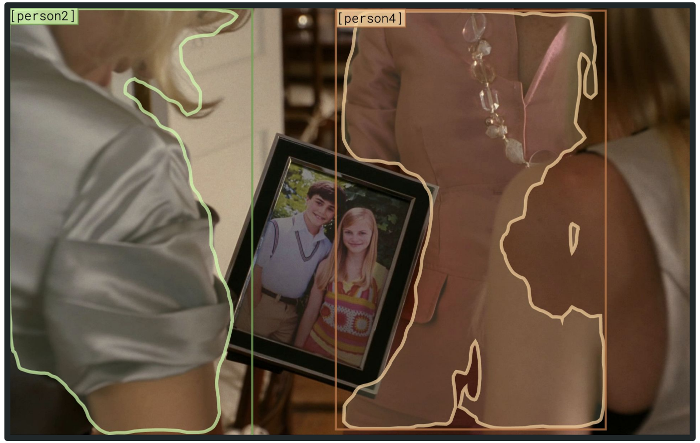
Question: What is going to happen next?

Answer: [person2] holding the photo will tell [person4] how cute their children are.

Free-text explanation: It looks like [person4] is showing the photo to [person2], and they will want to be polite.

Answering “why” by highlighting…

…doesn’t work when the reason is not explicitly stated in the input



[Zellers et al., 2019]

Free-text explanation:

- [person4] is showing the photo to [person2]
- [person2] will want to be polite

We cannot highlight this in the input!

Answering “why” by highlighting...

...doesn’t work when the reason is not explicitly stated in the input

Question: Where is a frisbee in play likely to be?

Answer choices: outside, park, roof, tree, <u>air</u>

Free-text explanation: A frisbee is a concave plastic disc designed for skimming through the air as an outdoor game so while in play it is most likely to be in the air.

[Aggarwal et al., 2021]

How to generate free-text explanations?

Step 1:

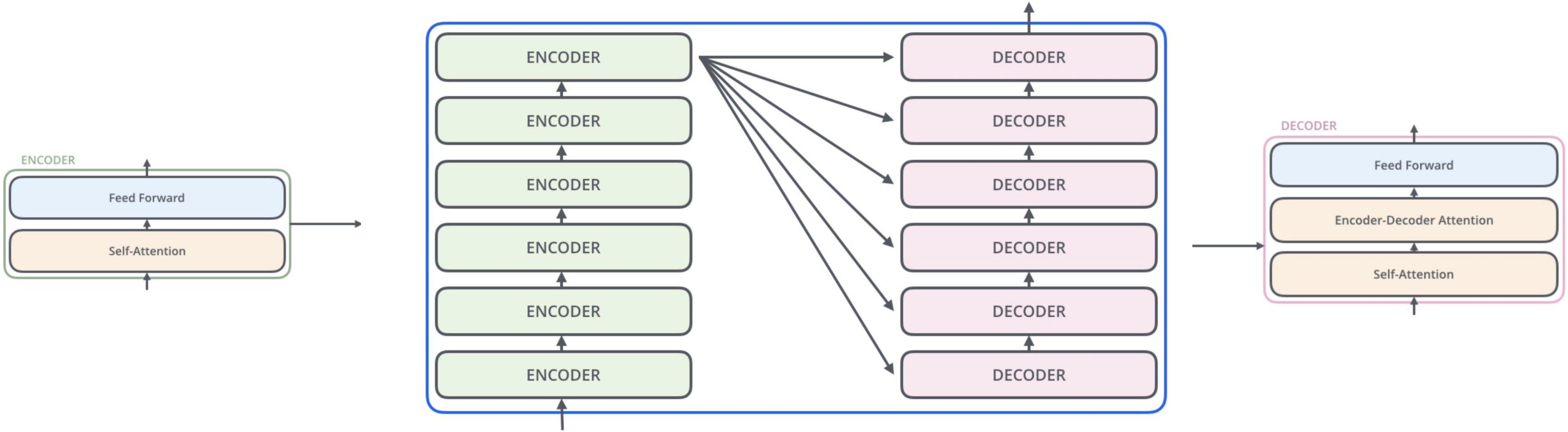
Find some human-written explanations[◊]

Step 2:

Finetune a pretrained transformer-based generation models (GPT-2)

[◊] [Wiegreffe* and Marasović*, NeurIPS 2021]

Generating Explanations

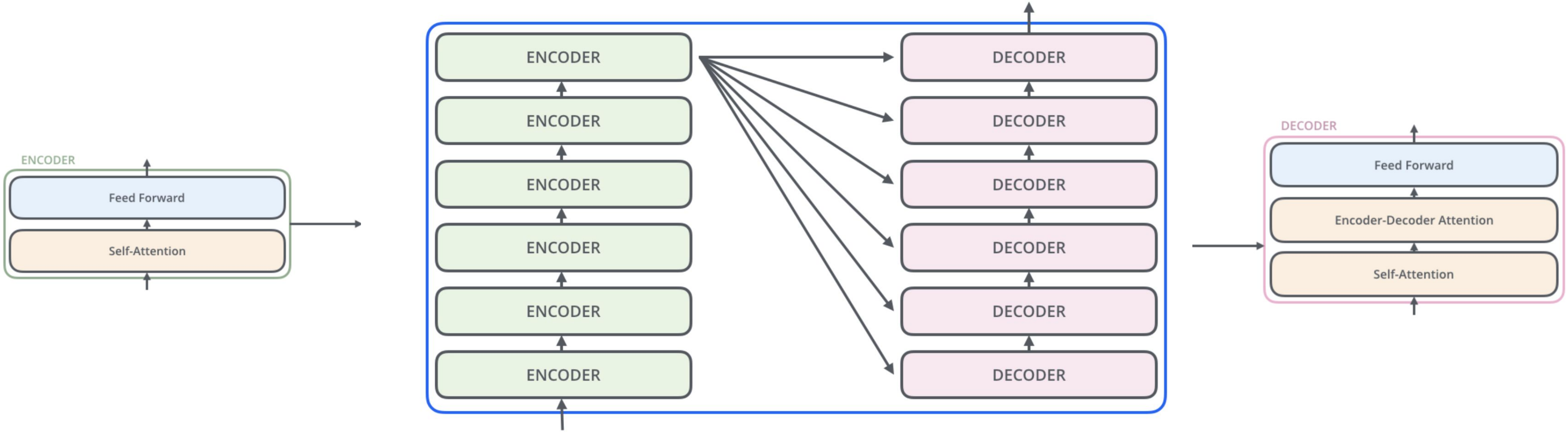



```
question: where is a frisbee in play likely
to be? choice: outside choice: park choice:
roof choice: tree choice: air
```

[Marasović et al., Findings of EMNLP 2020]

Generating Explanations

`Air because a frisbee is a concave plastic disc designed for skimming through the air as an outdoor game so while in play it is most likely to be in the air.`



`question: where is a frisbee in play likely to be? choice: outside choice: park choice: roof choice: tree choice: air`

[Marasović et al., Findings of EMNLP 2020]

Summary of Free-Text Explanations

Positives:

- Easy to comprehend, cognitive load of understanding is low
- Can explain instances of reasoning tasks

Negatives:

- Standard approach requires human-written explanations for supervision
- Can be used to deceive users

Lecture Outline

- Why did my model make this prediction?
  - Part I: Gradient-Based Highlighting
  - Part II: Free-Text Explanations
  - Part III: Influential Train Examples
- Why did my model predict P rather than Q?
  - Contrastive Editing
- How and who explanations help?

Why did my model make this prediction?

Part III: Influential Train Examples

Slides for this part are copied & slightly modified from
the EMNLP tutorial "Interpreting Predictions of NLP Models":
https://github.com/Eric-Wallace/interpretability-tutorial-emnlp2020

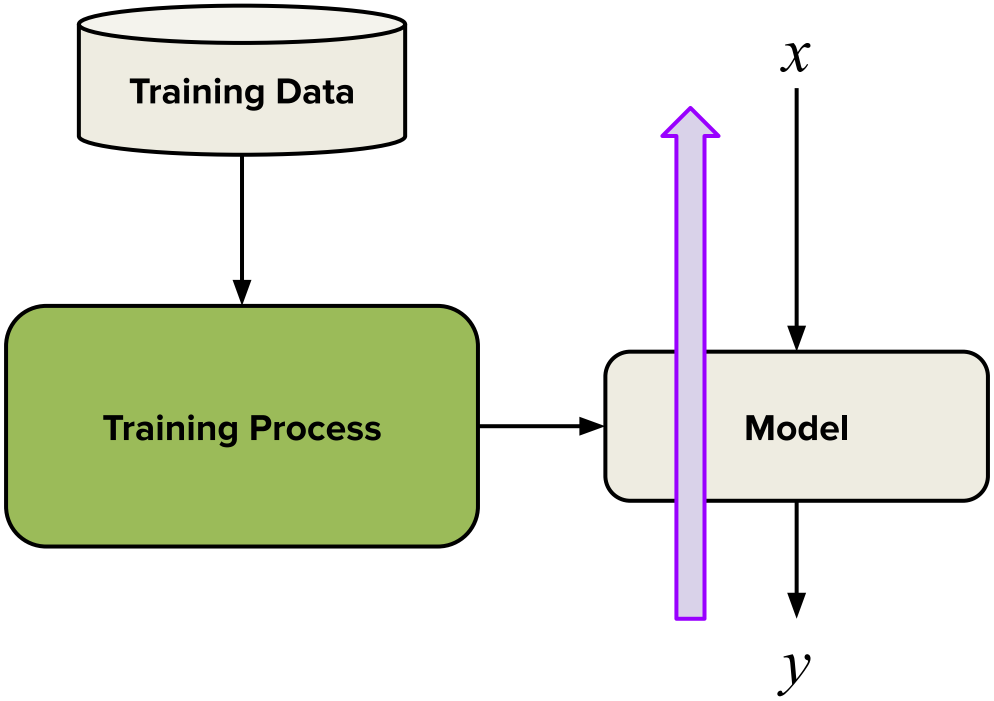
Thanks to the tutorial creators **Eric Wallace**, **Matt Gardner**, & **Sameer Singh!**

Why did my model make this prediction?

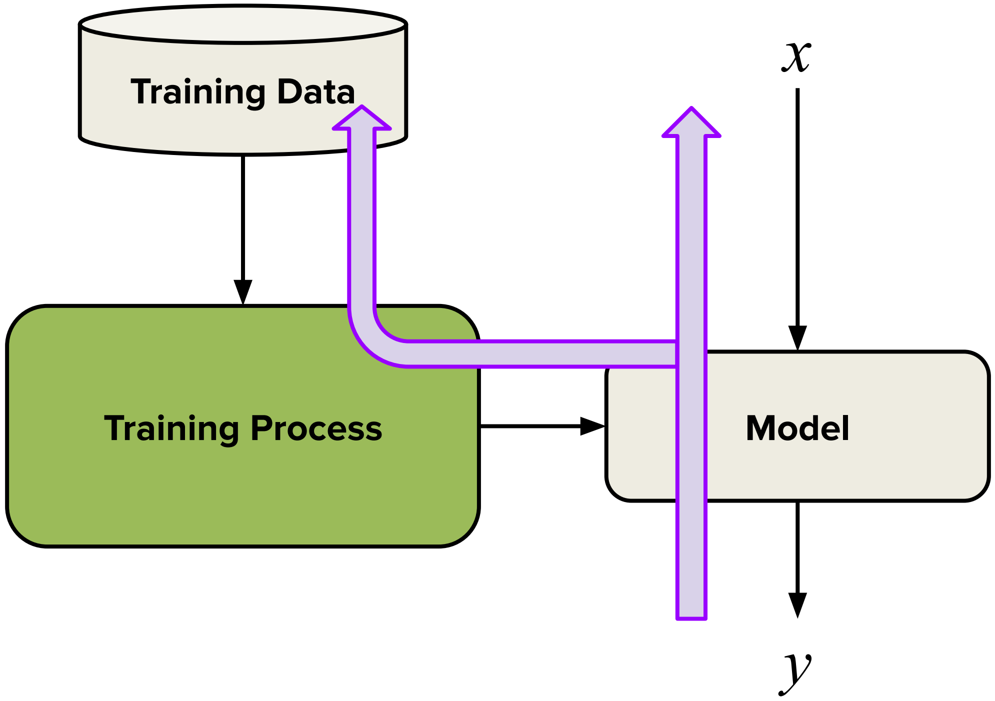
↓

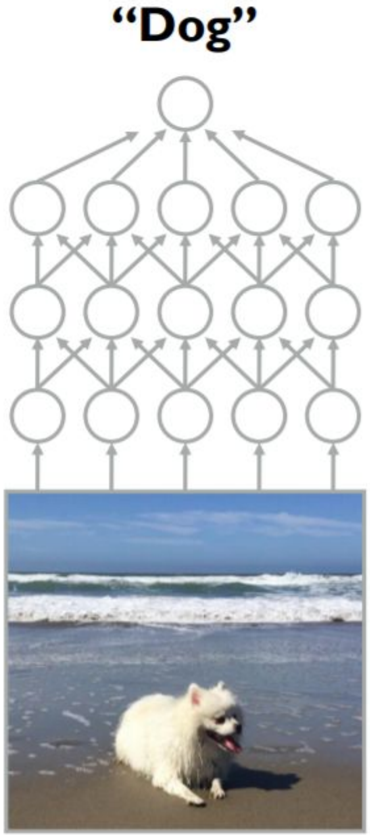
Which training examples were responsible for this prediction?

So far...

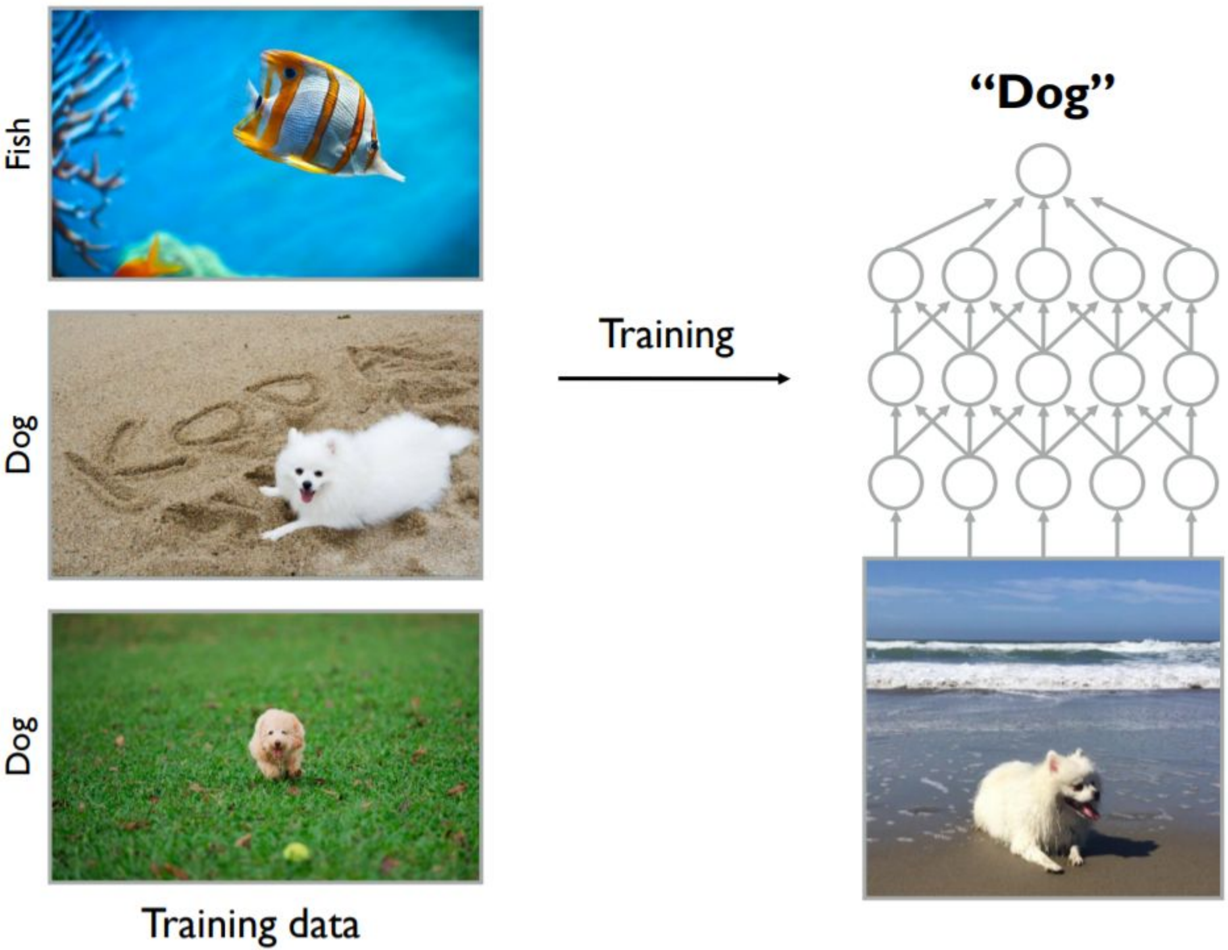


Data Influence



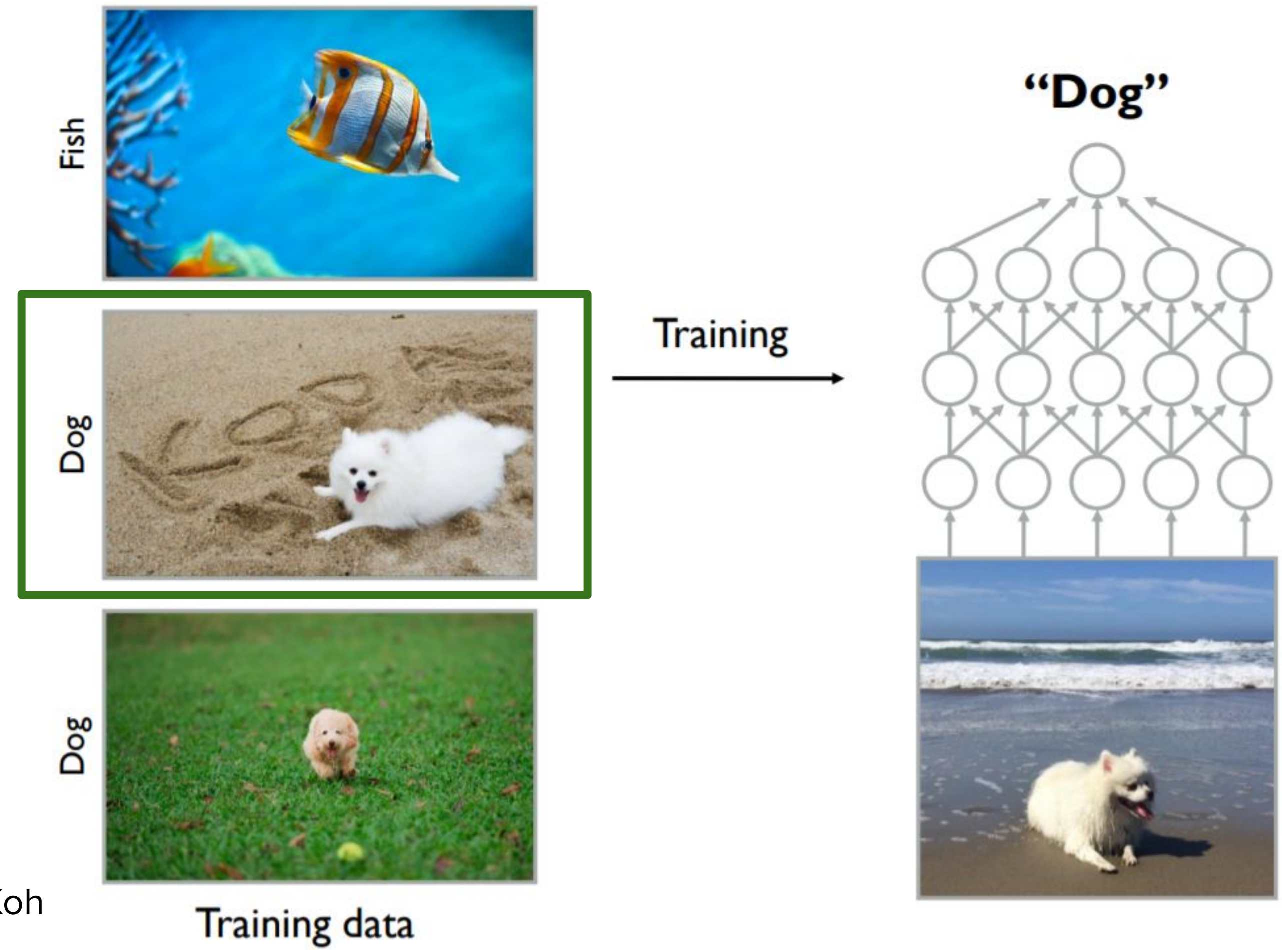


Slide: Pang Wei Koh



Slide: Pang Wei Koh

Most Important



Slide: Pang Wei Koh

Data Influence: Example Use Cases [Yeh et al. 2018]

Test Example

Polar Bear ✗

Data Influence: Example Use Cases [Yeh et al. 2018]

Test Example

Polar Bear ✘

Influential Training Examples

Polar Bear ✘

Beaver

Pig

Influence Functions

Why did my model make this prediction?

↓

Which training examples were responsible for this prediction?

Influence Functions

Why did my model make this prediction?

↓

Which training examples were responsible for this prediction?

↓

Which examples, if removed, would change the loss a lot?

[Koh and Liang 2017]

Pick $\hat{\theta}$ to minimize $\frac{1}{n}\sum_{i=1}^{n} L(z_i, \theta)$

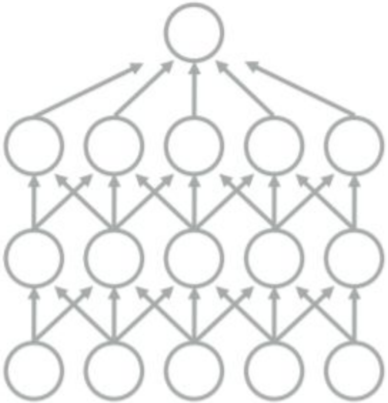
Fish

Dog

Dog

Training data $z_1, z_2, \ldots, z_n$

"Dog"



$\hat{\theta}$

Pick $\hat{\theta}$ to minimize $\frac{1}{n}\sum_{i=1}^{n} L(z_i, \theta)$

Fish

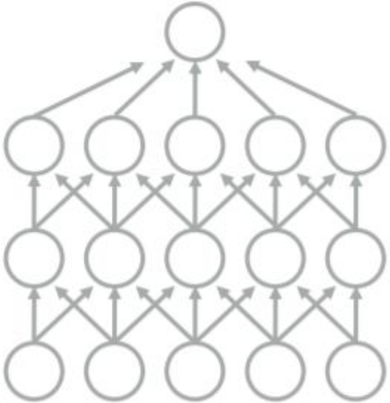
Dog

z_{train}

Dog

Training data $z_1, z_2, \ldots, z_n$

"Dog"



$\hat{\theta}$

Fish

Dog

z_{train}

Dog

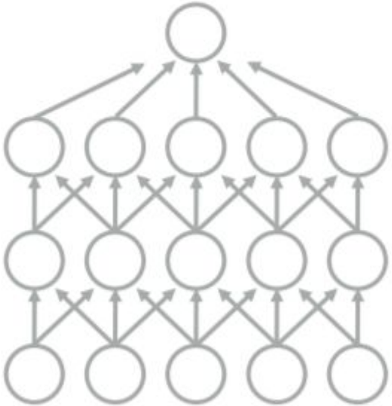
Training data $z_1, z_2, \ldots, z_n$

Pick $\hat{\theta}$ to minimize $\frac{1}{n}\sum_{i=1}^{n} L(z_i, \theta)$

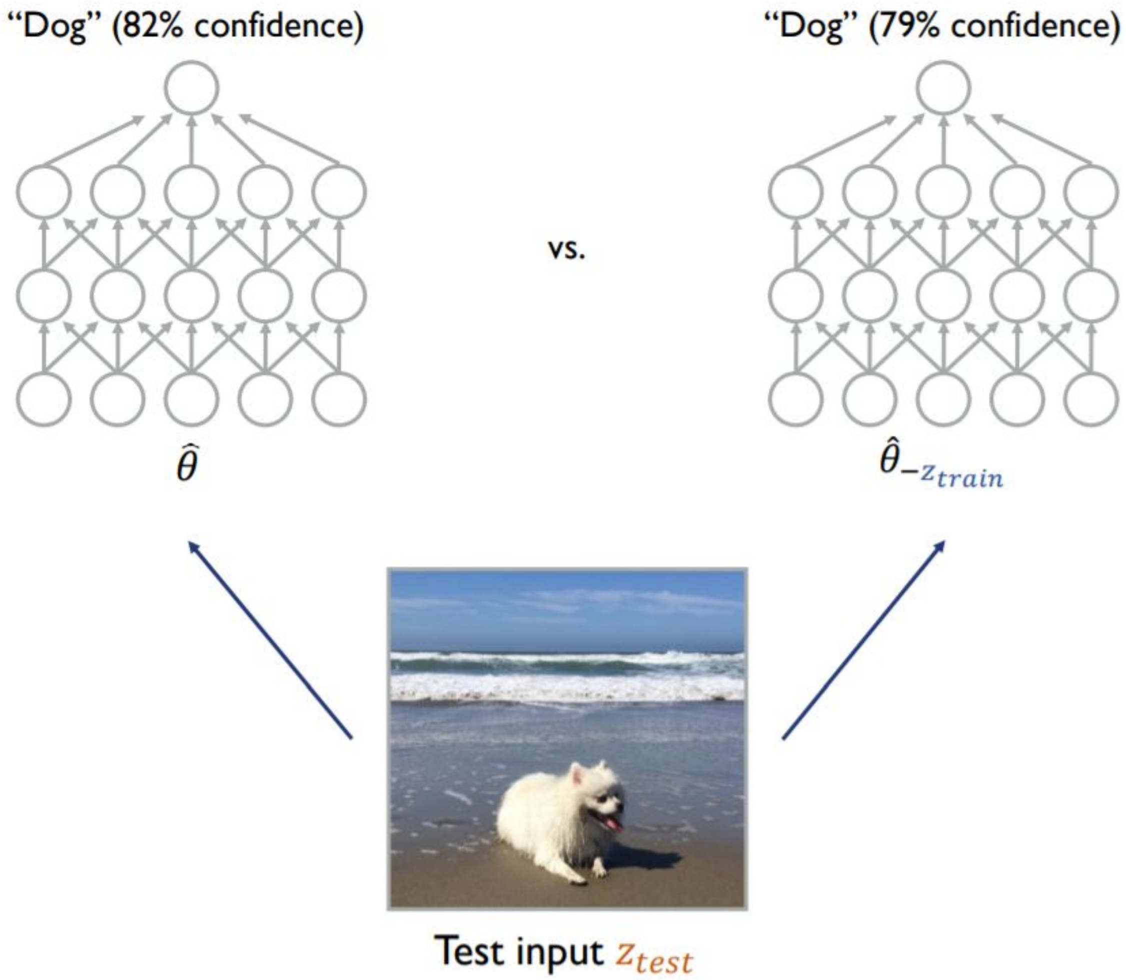
Pick $\hat{\theta}_{-z_{train}}$ to minimize

$$\frac{1}{n}\sum_{i=1}^{n} L(z_i, \theta) - \frac{1}{n} L(z_{train}, \theta)$$

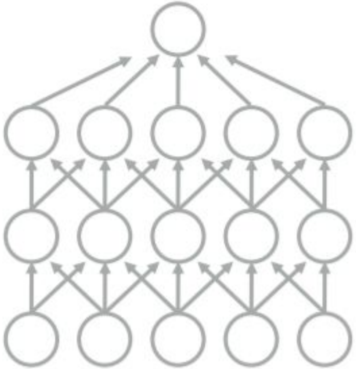
"Dog"



$\hat{\theta}_{-z_{train}}$


"Dog" (82% confidence)
"Dog" (79% confidence)
vs.
$\hat{\theta}$
$\hat{\theta}_{-z_{train}}$
Test input z_{test}

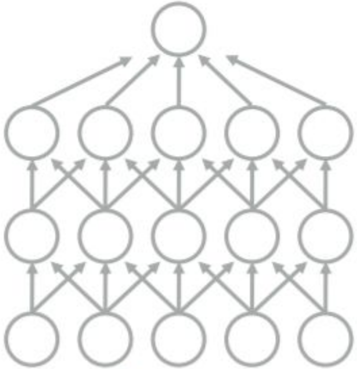
"Dog" (82% confidence)



$\hat{\theta}$

vs.

"Dog" (79% confidence)



$\hat{\theta}_{-z_{train}}$

What is $L(z_{test}, \hat{\theta}_{-z_{train}}) - L(z_{test}, \hat{\theta})$?

Use Case of Data Influence: Text Classification (NLI)

Test input	
P: The manager was encouraged by the secretary. *H:* The secretary encouraged the manager.	{entail}

Most supporting training examples	
P: Because you're having fun. *H:* Because you're having fun.	[entail]
P: I don't know if I was in heaven or hell, said Lillian Carter, the president's mother, after a visit. *H:* The president's mother visited.	[entail]
P: Inverse price caps. *H:* Inward caps on price.	[entail]
P: Do it now, think 'bout it later. *H:* Don't think about it now, just do it.	[entail]

[Han et al. 2020]

Influence Functions Summary

Pros:

- Principled approach (in the convex setting) for estimating influence of individual training points
- Works empirically for many models

Influence Functions Summary

Cons:

- Influential points can be uninterpretable
  - What influence did it actually have?
- Computationally expensive [Garima et al. 2020]
  - Especially with large training data!
- Often requires approximations that may be invalid [Basu et al. 2020]
  - Would prediction really change if training example wasn't there?
- How does it interact with pretrained models?
  - Are the influential points too specific to choice of pretrained models?

Need more work in this area!

Lecture Outline

- ➔ Why did my model make this prediction?
  - Part I: Gradient-Based Highlighting
  - Part II: Free-Text Explanations
  - Part III: Influential Train Examples
- ➔ Why did my model predict P rather than Q?
  - Contrastive Editing
- ➔ How and who explanations help?

Why did my model predict P rather than Q?

So far:

Why did my model make this prediction?

Insights from Social Science

Explanations are **contrastive** = responses to:

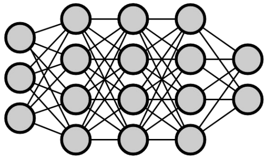
“Why P rather than Q?”

“What changes to the input would hypothetically change the answer from P to Q?”

where **P** is an observed event **(fact)**, and **Q** an imagined, counterfactual event that did not occur **(foil)**

[Miller, 2019]

do not admit to hospital



The patient reports to have a strong headache… They have dementia in Alzheimer's disease… They have not previously been treated for cardiovascular problems…

Why am I sent home?

- ➔ Headache
- ➔ Dementia in Alzheimer’s disease
- ➔ No cardiovascular problems

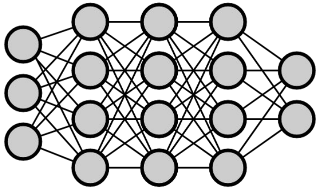
…

Why am I sent home [**rather than** admitted to a hospital]?

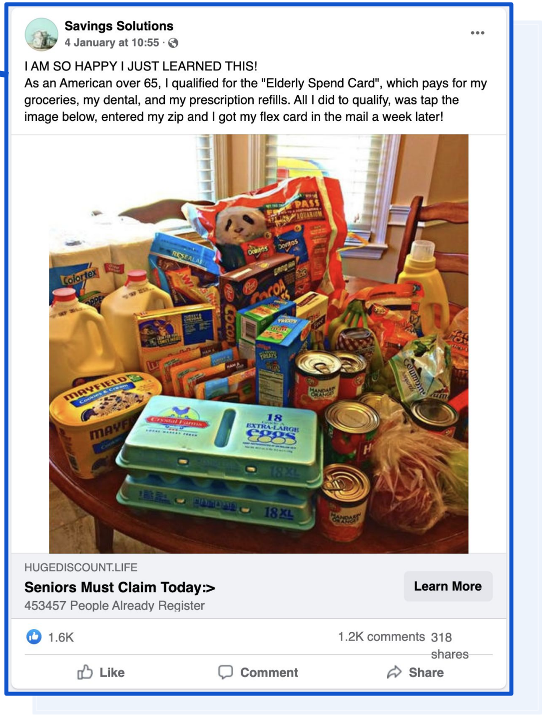
- ➔ No cardiovascular problems

The patient reports to have a strong headache… They have dementia in Alzheimer's disease… They have ~~not~~ previously been treated for cardiovascular problems…

misleading



\+ documents from the Web



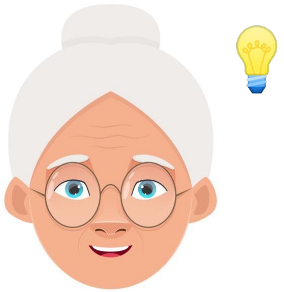
Why is my post misleading?
How can I change it to make it clear/correct?

Why is my post misleading?
How can I change it to make it clear/correct?

misleading

I AM SO HAPPY I JUST LEARNED THIS! As an American over 65, I qualified for the "Elderly Spend Card", which pays for my groceries, my dental, and my prescription refills. All I did to qualify, was tap the image below, entered my zip and I got my flex card in the mail a week later!

Contrastive explanations: explain **how to minimally modify the input to change the prediction to something else**



misleading

I AM SO HAPPY I JUST LEARNED THIS! As an American over 65, I qualified for the “Elderly Spend Card”, which pays for my groceries, my dental, and my prescription refills. All I did to qualify, was tap the image below, entered my zip and I got my flex card in the mail a week later!

correct

I AM SO HAPPY I JUST LEARNED THIS! As ~~an American over 65~~ **someone who has private health insurance with the Medicare Advantage plan, lives in X, and is chronically ill**, I qualified for the “Elderly Spend Card”, which pays for my groceries, my dental, and my prescription refills. All I did ~~to qualify~~, was tap the image below, entered my zip and I got my flex card in the mail a week later!

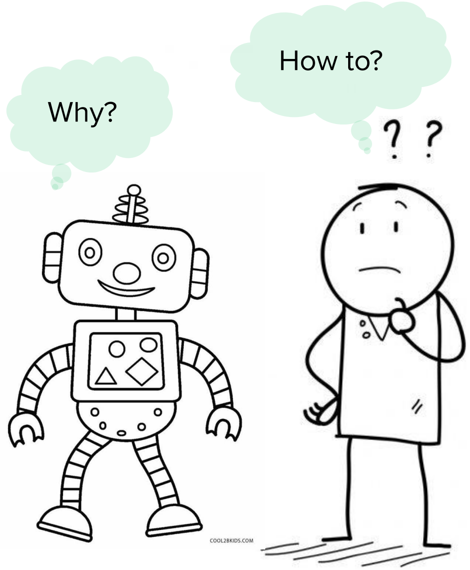
Contrastive explanations: explain **how to minimally modify the input to change the prediction to something else**

"Understanding how people define, generate, select, evaluate, and present explanations seems almost essential"

People assign human-like traits to AI models (**anthropomorphic bias**)

⇒ People expect explanations of models' behavior to follow the same conceptual framework used to explain human behavior

⇒ No users' agency otherwise



[Miller, 2019]

Contrastive Explanations of NLP Models

Contrastive input editing:
Minimal edits to the input that change model output to the contrast case

Yang et al. COLING 2020.

Jacovi and Goldberg. TACL 2021.

Ross et al. Findings of ACL 2021.

Wu et al. ACL 2021.

Collect **free-text** human **contrastive explanations**, ...

...and **generate them** left-to-right Chen et al. ACL 2021.

...abstract them into templates, automatically fill in the templates **(template-based infilling)**

Paranjape et al. Findings of ACL 2021.

Contrastive vector representation:
A dense representation of the input that captures latent features that differentiate two classes

Jacovi et al. EMNLP 2021.

Contrastive Explanations of NLP Models

Contrastive input editing:
Minimal edits to the input that change model output to the contrast case

Yang et al. COLING 2020.

Jacovi and Goldberg. TACL 2020.

Ross et al. Findings of ACL 2021.

Wu et al. ACL 2020.

Collect **free-text** human **contrastive explanations**, …

…and **generate them** left-to-right Chen et al. ACL 2021.

…abstract them into templates, automatically fill in the templates **(template-based infilling)**

Paranjape et al. Findings of ACL 2021.

Contrastive vector representation:
A dense representation of the input that captures latent features that differentiate two classes

Jacovi et al. EMNLP 2021.

Contrastive Explanations via **Contrastive Editing**

Question:
Ann and her children are going to Linda's home _____.

(a) by bus (b) by car (c) on foot (d) by train

Why **"by train"** (d) and not **"on foot"** (c)?
How to change the answer from **"by train"** (d) to **"on foot"** (c)?

Context:
...Dear Ann, I hope that you and your children will be here in two weeks. My husband and I will go to meet you at the train station. Our town is small...

MiCE-Edited Context:
...Dear Ann, I hope that you and your children will be here in two weeks. My husband and I will go to meet you at ~~the train station~~ **your home on foot**. Our ~~town~~ **house** is small...

[Ross, **Marasović**, Peters, Findings of ACL 2021]

Goal:

Explain a **Predictor** model by *automatically* finding a **minimal edit** to the input that **causes Predictor's output to change to the contrast case**

[Ross, **Marasović**, Peters, Findings of ACL 2021]

Goal:

Explain a **Predictor** model by *automatically* finding a **minimal edit** to the input that **causes Predictor's output to change to the contrast case**

A very high-level idea of 🐭:

➔ Use an **Editor** model to edit the input by **masking input words & filling masked positions** until we find cause Predictor's output to change to the contrast case

➔ Simultaneously, minimize the masking percentage ∼ the edit size

[Ross, **Marasović**, Peters, Findings of ACL 2021]

input: `Sylvester Stallone has made some crap films in his lifetime, but this has got to be one of the worst. A totally dull story...`

the contrast label (foil)

```
label: positive input: Sylvester Stallone has made some crap films in his lifetime,
but this has got to be one of the worst. A totally dull story...
```

[Ross, **Marasović**, Peters, Findings of ACL 2021]

the contrast label (foil)

`label: positive input: Sylvester Stallone has made some crap films in his lifetime, but this has got to be one of the worst. A totally dull story...`

mask *n%* of input tokens

`label: positive input: Sylvester Stallone has made some` **<mask>** `films in his lifetime, but this has got to be one of the` **<mask>**`. A totally` **<mask>** `story...`

[Ross, **Marasović**, Peters, Findings of ACL 2021]

the contrast label (foil)

label: positive input: Sylvester Stallone has made some crap films in his lifetime, but this has got to be one of the worst. A totally dull story...

mask *n%* of input tokens

label: positive input: Sylvester Stallone has made some **<mask>** films in his lifetime, but this has got to be one of the **<mask>**. A totally **<mask>** story...

sample 15 spans at each masked position

1. label: positive input: Sylvester Stallone has made some good films in his lifetime, but this has got to be one of the worst. A totally novel story...

2. label: positive input: Sylvester Stallone has made some great films in his lifetime, but this has got to be one of the greatest of all time. A totally boring story...

...

15. label: positive input: Sylvester Stallone has made some wonderful films in his lifetime, but this has got to be one of the greatest. A totally tedious story...

the contrast label (foil)

label: positive input: Sylvester Stallone has made some crap films in his lifetime, but this has got to be one of the worst. A totally dull story...

mask *n%* of input tokens

label: positive input: Sylvester Stallone has made some **<mask>** films in his lifetime, but this has got to be one of the **<mask>**. A totally **<mask>** story...

sample 15 spans at each masked position

get the logit of the contrast label

1. label: positive input: Sylvester Stallone has made some good films in his lifetime, but this has got to be one of the worst. A totally novel story...

$$l(pos) = 0.2$$

2. label: positive input: Sylvester Stallone has made some great films in his lifetime, but this has got to be one of the greatest of all time. A totally boring story...

$$l(pos) = 0.6$$

...

15. label: positive input: Sylvester Stallone has made some wonderful films in his lifetime, but this has got to be one of the greatest. A totally tedious story...

$$l(pos) = 0.65$$

1. Prepend the contrast label to the input
2. Mask *n%* of the input tokens
3. Sample 15 spans at masked positions

× **4 different values of *n* to minimize the edit**

4*15=60 samples

[Ross, **Marasović**, Peters, Findings of ACL 2021]

1. Prepend the contrast label to the input
2. Mask *n*% of the input tokens
3. Sample 15 spans at masked positions

× **4 different values of *n* to minimize the edit**

How to pick which values for *n*?

Binary search on [0,55]

[Ross, **Marasović**, Peters, Findings of ACL 2021]

1. Prepend the contrast label to the input
2. Mask *n*% of the input tokens
3. Sample 15 spans at masked positions

× **4 different values of *n* to minimize the edit**

How to pick which values for *n*?

Binary search on [0,55]

Start: $n^{(1)}=27.5\%$

[Ross, **Marasović**, Peters, Findings of ACL 2021]

1. Prepend the contrast label to the input
2. Mask *n*% of the input tokens
3. Sample 15 spans at masked positions

$\times$

4 different values of *n* to minimize the edit

How to pick which values for *n*?

Binary search on [0,55]

Start: $n^{(1)}=27.5\%$

➔ If a contrastive edit found: $n^{(2)}=13.75\%$

[Ross, Marasović, Peters, Findings of ACL 2021]

1. Prepend the contrast label to the input
2. Mask *n*% of the input tokens
3. Sample 15 spans at masked positions

$\times$

4 different values of *n* to minimize the edit

How to pick which values for *n*?

Binary search on [0,55]

Start: $n^{(1)}=27.5\%$

➔ If a contrastive edit found: $n^{(2)}=13.75\%$

➔ If a contrastive edit **not** found: $n^{(2)}=41.25\%$

[Ross, Marasović, Peters, Findings of ACL 2021]

1. Prepend the contrast label to the input
2. Mask *n%* of the input tokens
3. Sample 15 spans at masked positions

× **4 different values of *n* to minimize the edit**

How to pick which values for *n*?

Binary search on [0,55]

Start: $n^{(1)}$=27.5%

- ➔ If a contrastive edit found: $n^{(2)}$=13.75%
  - ◆ If a contrastive edit found: $n^{(3)}$=6.875%
- ➔ If a contrastive edit **not** found: $n^{(2)}$=41.25%
  - ◆ If a contrastive edit found: $n^{(3)}$=20.625%

[Ross, Marasović, Peters, Findings of ACL 2021]

1. Prepend the contrast label to the input
2. Mask *n%* of the input tokens
3. Sample 15 spans at masked positions

× **4 different values of *n* to minimize the edit**

How to pick which values for *n*?

Binary search on [0,55]

Start: $n^{(1)}=27.5\%$

- ➔ If a contrastive edit found: $n^{(2)}=13.75\%$
  - ◆ If a contrastive edit found: $n^{(3)}=6.875\%$
  - ◆ If a contrastive edit **not** found: $n^{(3)}=20.625\%$
- ➔ If a contrastive edit **not** found: $n^{(2)}=41.25\%$
  - ◆ If a contrastive edit found: $n^{(3)}=20.625\%$
  - ◆ If a contrastive edit **not** found: $n^{(3)}=48.125\%$

[Ross, Marasović, Peters, Findings of ACL 2021]

1. Prepend the contrast label to the input
2. Mask *n%* of the input tokens
3. Sample 15 spans at masked positions

× **4 different values of *n* to minimize the edit**

How to pick masking positions?

Based on token importance for the original prediction

Rank input tokens based on the gradient magnitude of the model we're explaining

Mask top-*n*% of **ranked** tokens

[Ross, Marasović, Peters, Findings of ACL 2021]

1. Prepend the contrast label to the input
2. Mask *n%* of the input tokens
3. Sample 15 spans at masked positions

$\times$

4 different values of *n* to minimize the edit

4*15=60 samples

[Ross, Marasović, Peters, Findings of ACL 2021]

1. Prepend the contrast label to the input
2. Mask *n*% of the input tokens
3. Sample 15 spans at masked positions

×

4 different values of *n* to minimize the edit

4*15=60 samples

rank 60 samples w.r.t. the logit of the contrast label

[Ross, Marasović, Peters, Findings of ACL 2021]

1. Prepend the contrast label to the input
2. Mask *n%* of the input tokens
3. Sample 15 spans at masked positions

× **4 different values of *n* to minimize the edit**

4*15=60 samples

rank 60 samples w.r.t. the logit of the contrast label

keep top-3 samples **beam**

[Ross, Marasović, Peters, Findings of ACL 2021]

1. Prepend the contrast label to the input
2. Mask *n%* of the input tokens
3. Sample 15 spans at masked positions

× **4 different values of *n* to minimize the edit**

4*15=60 samples

rank 60 samples w.r.t. the logit of the contrast label

keep top-3 samples **beam**

if a contrastive edit is found

[Ross, Marasović, Peters, Findings of ACL 2021]

Repeat for every instance in the beam at most **2 more rounds**

1. Prepend the contrast label to the input
2. Mask *n*% of the input tokens
3. Sample 15 spans at masked positions

× **4 different values of *n* to minimize the edit**

4*15=60 samples

rank 60 samples w.r.t. the logit of the contrast label

keep top-3 samples **beam**

Can a pretrained model without any additional tweaks fill in the spans?

[Ross, Marasović, Peters, Findings of ACL 2021]

Can a pretrained model without any additional tweaks fill in the spans?

We find it's important to **prepare the editor** by finetuning it to infill masked spans given masked text and **a target end-task label**

`(standard masking)` `Sylvester Stallone has made some` **`<mask>`** `films in his lifetime, but this has got to be one of the` **`<mask>`**`. A totally` **`<mask>`** `story...`

`(targeted masking)` `label: negative` `input: Sylvester Stallone has made some` **`<mask>`** `films in his lifetime, but this has got to be one of the` **`<mask>`**`. A totally` **`<mask>`** `story...`

[Ross, Marasović, Peters, Findings of ACL 2021]

Can a pretrained model without any additional tweaks fill in the spans?

We find it's important to **prepare the editor** by finetuning it to infill masked spans given masked text and **a target end-task label**

We find that **labels predicted by the model** we're explaining **can be used** in this step without a big loss in performance

[Ross, Marasović, Peters, Findings of ACL 2021]

Can a pretrained model without any additional tweaks fill in the spans?

We find it's important to **prepare the editor** by finetuning it to infill masked spans given masked text and **a target end-task label**

We find that **labels predicted by the model** we're explaining **can be used** in this step without a big loss in performance

Gradient-based masking in this step gives better performance

[Ross, Marasović, Peters, Findings of ACL 2021]

MiCE is a two-stage approach to generating contrastive edits

- ➔ Stage 1: Prepare an editor
- ➔ Stage 2: Make edits guided with gradients & logits of the model we're explaining

[Ross, Marasović, Peters, Findings of ACL 2021]

The maximum number of iterations for a single instance:

$$\overbrace{\text{\# binary search levels } \mathtt{s} \times \text{\# samples at each maskin position } \mathtt{m}}^{\text{first round}} +$$

$$\underbrace{\text{beam size } \mathtt{b} \times \text{\# binary search levels } \mathtt{s} \times \text{\# samples at each masking position } \mathtt{m} \times \text{\# of rounds}}_{\text{other rounds}} =$$

$4 \times 15 + 3 \times 4 \times 15 \times 2 = 420$

That's a lot, and also there is no guarantee that a smaller contrastive edit does not exist

[Ross, Marasović, Peters, Findings of ACL 2021]

Methodology for Detecting Artifacts with Local Explanations

1. **Construct a validation set:** use a standard split, or intentionally construct a small set of potentially challenging samples

2. **Produce local explanations** for examples in Step 1

3. **Identify candidate artifacts:**
   a. **Granular:** aggregate the important granular features from local explanations in Step 2 & identify features that appear disproportionately
   b. **Abstract:** inspect local explanations from Step 2 manually

4. **Verify candidate artifacts** by manipulating examples in Step 1, e.g., observing the effect of removing/replacing identified artifacts on the model prediction

Modified from the methodology in [Pezeshkpour et al., 2021]

How MiCE Edits Can Be Used?

MiCE's edits can offer hypotheses about model "bugs"

Original prediction: positive

> An interesting pairing of stories, this little flick manages to bring together seemingly different characters and story lines all in the backdrop of WWII and succeeds in tying them together without losing the audience. I was impressed by the depth portrayed by the different characters and also by how much I really felt I understood them and their motivations, even though the time spent on the development of each character was very limited. The outstanding acting abilities of the individuals involved with this picture are easily noted. A fun, stylized movie with a slew of comic moments and a bunch more head shaking events. 7/10

[Ross, Marasović, Peters, Findings of ACL 2021]

How MiCE Edits Can Be Used?

MiCE's edits can offer hypotheses about model "bugs"

`MiCE's edit × contrast prediction (negative)`

An interesting pairing of stories, this little flick manages to bring together seemingly different characters and story lines all in the backdrop of WWII and succeeds in tying them together without losing the audience. I was impressed by the depth portrayed by the different characters and also by how much I really felt I understood them and their motivations, even though the time spent on the development of each character was very limited. The outstanding acting abilities of the individuals involved with this picture are easily noted. A fun, stylized movie with a slew of comic moments and a bunch more head shaking events. ~~7/10~~ 4/10

[Ross, Marasović, Peters, Findings of ACL 2021]

How MiCE Edits Can Be Used?

MiCE's edits can offer hypotheses about model "bugs"

`MiCE's edit` × **`contrast prediction (negative)`**

An interesting pairing of stories, this little flick manages to bring together seemingly different characters and story lines all in the backdrop of WWII and succeeds in tying them together without losing the audience. I was impressed by the depth portrayed by the different characters and also by how much I really felt I understood them and their motivations, even though the time spent on the development of each character was very limited. The outstanding acting abilities of the individuals involved with this picture are easily noted. A fun, stylized movie with a slew of comic moments and a bunch more head shaking events. ~~7/10~~ 4/10

[Ross, Marasović, Peters, Findings of ACL 2021]

How MiCE Edits Can Be Used?

MiCE's edits can offer hypotheses about model "bugs"

Hypothesis:
Model learned to rely heavily on numerical ratings ⭐

Test the hypothesis using MiCE's edits:

1. Filter instances with edits smaller than ≤ 0.05

2. Select tokens that are removed/inserted more than expected given their frequency in the original IMDB inputs

$y_c =$ ***positive***		$y_c =$ ***negative***	
Removed	**Inserted**	**Removed**	**Inserted**
4/10	excellent	10/10	awful
ridiculous	enjoy	8/10	disappointed
horrible	amazing	7/10	1
4	entertaining	9	4
predictable	10	enjoyable	annoying

[Ross, Marasović, Peters, Findings of ACL 2021]

Who? What are *expectations*, *background*, & *needs* of a person for who explanations are introduced?

Debugging

Why? What are the goals of producing explanations?

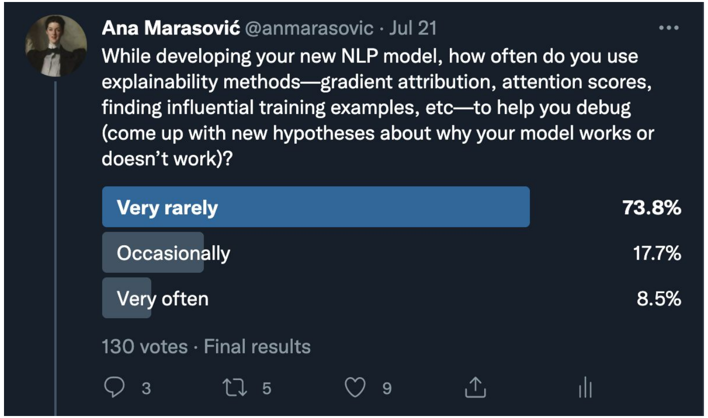
What is the content we should to include in the explanation?

How? What type of explanation is the most appropriate?

“Who? Why? What? How?” framework introduced in [Ribera and Lapedriza, IUI Workshops 2019]

How and who explanations help?

Although local explanations are specifically motivated for people to use, there is no convincing evidence yet that local explanations help people who are using language technology



An AI model is **trustworthy** to a given contract if it is capable of maintaining the contract.

If a human perceives that an AI model is trustworthy to a contract, and therefore accepts vulnerability to AI's actions, then the human **trusts** AI contractually. Otherwise, human **distrusts** AI contractually.

Trust does not exist if the human does not perceive risk.

Human's contractual trust in AI is **warranted** if it is caused by trustworthiness in AI. Otherwise, human's trust in AI is **unwarranted**.

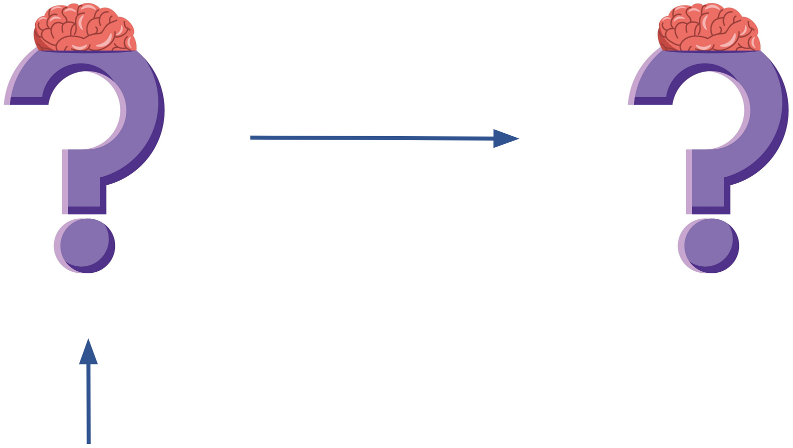
[Jacovi, **Marasović**, Miller, Goldberg; FAccT 2021]

Trust does not exist if the human does not perceive risk, but…

Researchers focus on grand AI challenges that people are good at (e.g., commonsense QA, *“Where is a frisbee in play likely to be?”*)

Researchers focus use simple tasks that people don’t need help with (e.g., claim verification against a very short text)

Who? What are *expectations*, *background*, & *needs* of a person for who explanations are introduced?



Why? What are the goals of producing explanations?

What is the content we should to include in the explanation?

How? What type of explanation is the most appropriate?

"Who? Why? What? How?" framework introduced in [Ribera and Lapedriza, IUI Workshops 2019]

Thank you!

Questions?

References that are not links

- Ribera and Lapedriza. Can we do better explanations? A proposal of user-centered explainable AI. IUI Workshops 2019.
- Miller. Explanation in artificial intelligence: Insights from the social sciences. AIJ 2019.
- Yang et al. Generating Plausible Counterfactual Explanations for Deep Transformers in Financial Text Classification. COLING 2020.
- Jacovi and Goldberg. Aligning Faithful Interpretations with their Social Attribution. TACL 2021.
- Chen et al. KACE: Generating Knowledge-Aware Contrastive Explanations for NLI. ACL 2021.
- Ross et al. Explaining NLP Models via Minimal Contrastive Editing (MiCE). Findings of ACL 2021.
- Paranjape et al. Prompting Contrastive Explanations for Commonsense Reasoning Tasks. Findings of ACL 2021.
- Wu et al. Polyjuice: Generating Counterfactuals for Explaining, Evaluating, and Improving Models. ACL 2021.
- Jacovi et al. Contrastive Explanations for Model Interpretability. EMNLP 2021.