EXPLAINABLE ARTIFICIAL INTELLIGENCE

L193 – Student Presentations 1 – Lent 2025

UNIVERSITY OF CAMBRIDGE

PAPER TIME!

Papers! Papers! Papers! Papers! Papers! Papers!

"*A PhD student drowning under a pile of papers*" – Quite literal interpretation by ChatGPT-4o

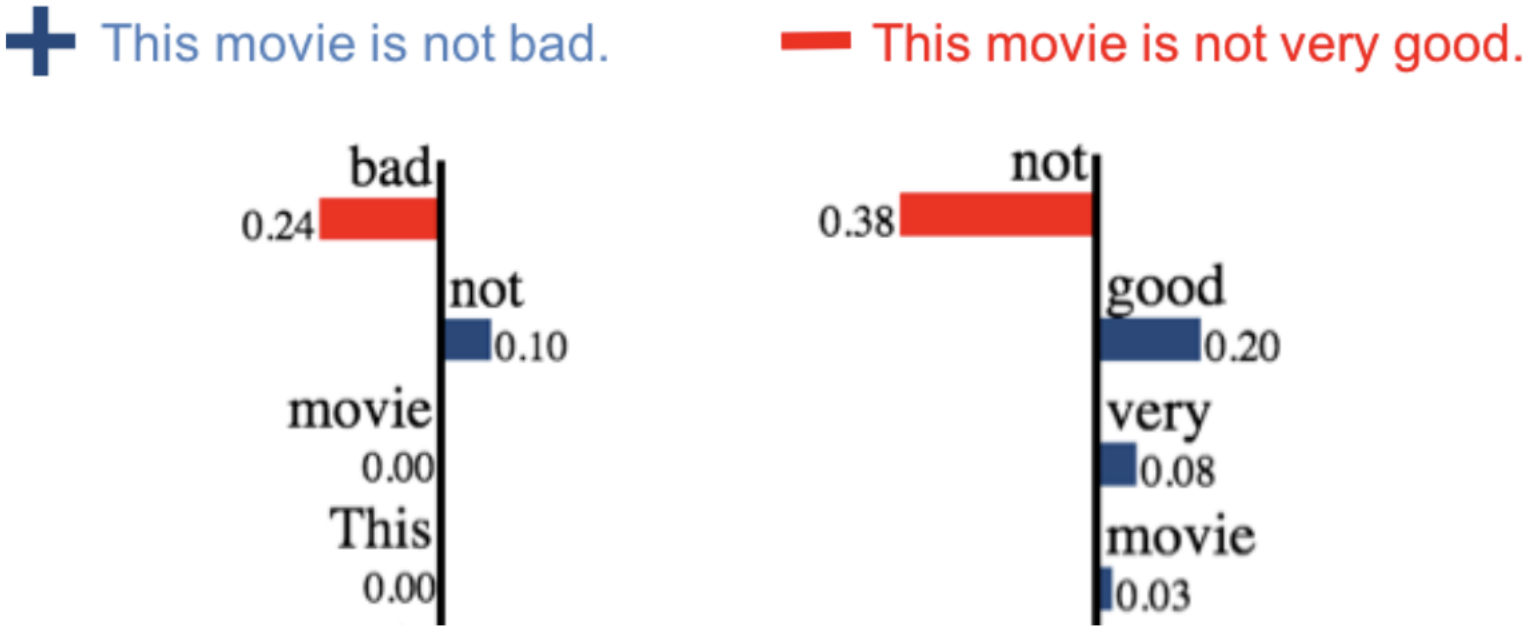
PAPER RECAP

ANCHORS (RIBEIRO ET AL., 2018)

Ribeiro et al. "**Anchors: High-precision model-agnostic explanations**." AAAI 2018.

ANCHORS: TAKEAWAYS

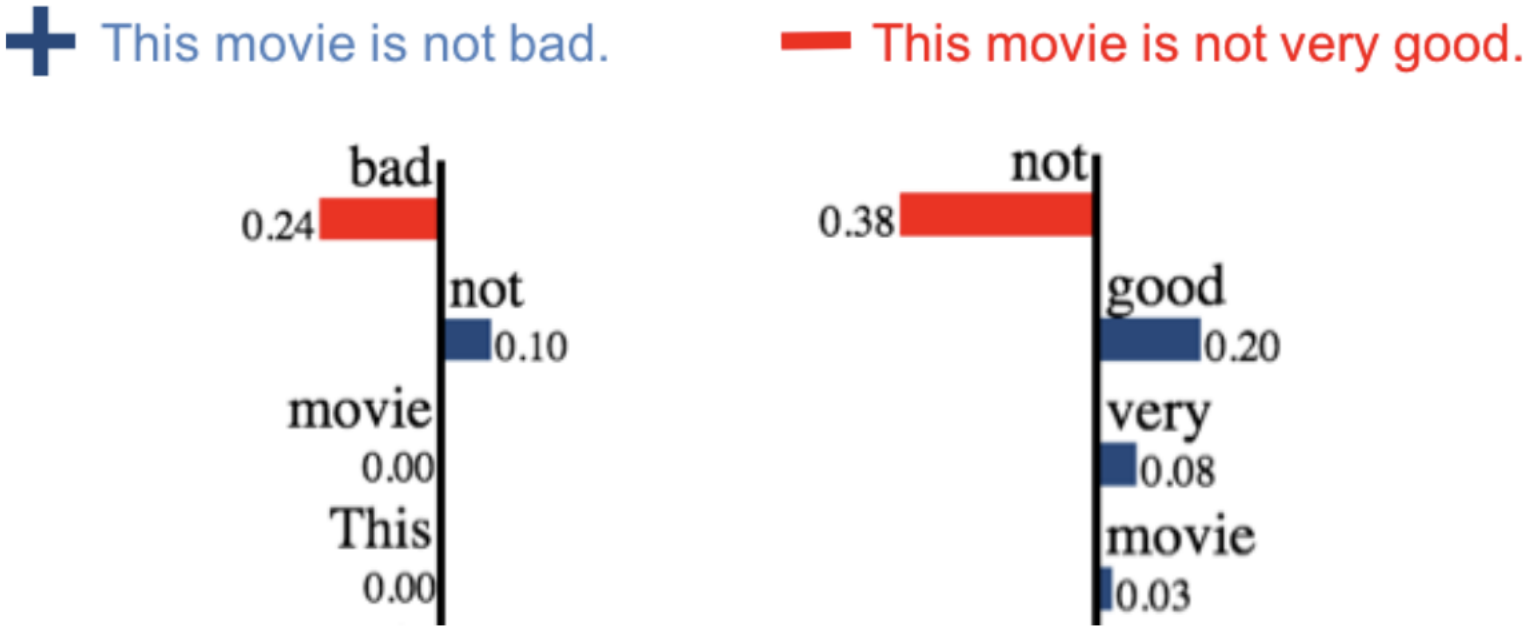
1. Feature interactions are context-dependent, but linear local explanations (e.g., LIME) are not



Looking at the LHS explanation, we may think that "not" is a positive word for predicting a good review!

ANCHORS: TAKEAWAYS

1. Feature interactions are context-dependent, but linear local explanations (e.g., LIME) are not



A local linear explanation's coverage is unclear!

ANCHORS: TAKEAWAYS

2. Instead, we could look at sets of features that "anchor" a sample's prediction to its current value

Given

Sample $\boldsymbol{x}$

Black-box model $\mathbf{f}(\boldsymbol{x})$

Data distribution $D(z|A)$
(distribution over perturbed samples near x)

ANCHORS: TAKEAWAYS

2. Instead, we could look at sets of features that "anchor" a sample's prediction to its current value

Given

Sample $\boldsymbol{x}$

Black-box model $\mathbf{f}(\boldsymbol{x})$

Data distribution $D(z|A)$
(distribution over perturbed samples near x)

We Want

An anchor (i.e., a feature rule) A such that:

1. A applies to at least τ samples in D

$$P\left(\mathrm{prec}(A) \geq \tau\right) \geq 1 - \delta$$

$$\mathrm{prec}(A) = \mathbb{E}_{\mathcal{D}(z|A)}\left[\mathbb{1}_{f(x)=f(z)}\right]$$

ANCHORS: TAKEAWAYS

2. Instead, we could look at sets of features that "anchor" a sample's prediction to its current value

Given

Sample $\boldsymbol{x}$

Black-box model $\mathrm{f}(\boldsymbol{x})$

Data distribution $D(z|A)$
(distribution over perturbed samples near x)

We Want

An anchor (i.e., a feature rule) A such that:

1. A applies to at least τ samples in D

2. A applies to the largest number of samples

$$\max_{A \text{ s.t. } P(\text{prec}(A) \geq \tau) \geq 1-\delta} \text{cov}(A).$$

ANCHORS: TAKEAWAYS

3. We construct this by iteratively building anchors, one feature at a time

Can be framed as a multi-armed bandit problem:

1. Each new feature x_i is an arm
2. The true precision of adding x_i to the current rule A is the reward, and
3. An evaluation of $f(\cdot)$ on a sample in $D(z \mid A \in \{x_i\})$ is an arm pull

This means we can solve this task via methods for bandit problems!

ANCHORS: TAKEAWAYS

3. We construct this by iteratively building anchors, one feature at a time

Can be framed as a multi-armed bandit problem:

1. Each new feature x_i is an arm
2. The true precision of adding x_i to the current rule A is the reward, and
3. An evaluation of $f(\cdot)$ on a sample in $D(z \mid A \in \{x_i\})$ is an arm pull

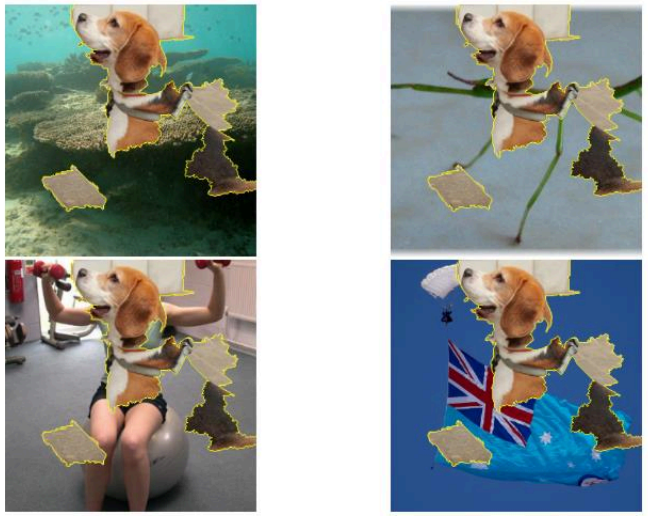
This means we can solve this task via methods for bandit problems!

We can extend this to a set of potential rule candidates via beam-search

ANCHORS: TAKEAWAYS

4. We extract anchors for a sample using domain-specific perturbation distributions $D(z \mid A)$

A



$D(z \mid A)$

INTEGRATED GRADIENTS (SUNDARARAJAN ET AL., 2017)

Sundararajan et al. "Axiomatic attribution for deep networks." ICML 2017

INTEGRATED GRADIENTS: TAKEAWAYS

1. Feature attribution methods should be:

   a) Sensitive: if a change in a feature changes the output, that feature must have non-zero attribution

Vanilla Gradient fails to achieve this!

(consider a ReLU nonlinearity with a sharp transition)

INTEGRATED GRADIENTS: TAKEAWAYS

1. Feature attribution methods should be:
   a) Sensitive: if a change in a feature changes the output, that feature must have non-zero attribution
   b) Implementation Invariance: functionally equivalent networks should produce identical attributions.

INTEGRATED GRADIENTS: TAKEAWAYS

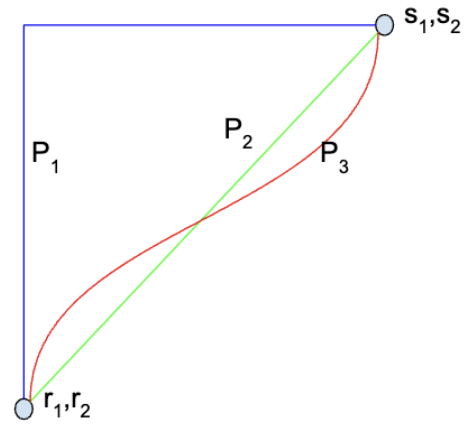
2. Integrated Gradients achieve both of these axioms (and also completeness) by integrating the gradients on a path from a baseline input to the sample of interest.

$$\text{IntegratedGrads}_i(x) ::= (x_i - x'_i) \times \int_{\alpha=0}^{1} \frac{\partial F(x' + \alpha \times (x - x'))}{\partial x_i} \, d\alpha$$

INTEGRATED GRADIENTS: TAKEAWAYS

2. Integrated Gradients achieve both of these axioms (and also completeness) by integrating the gradients on a path from a baseline input to the sample of interest.

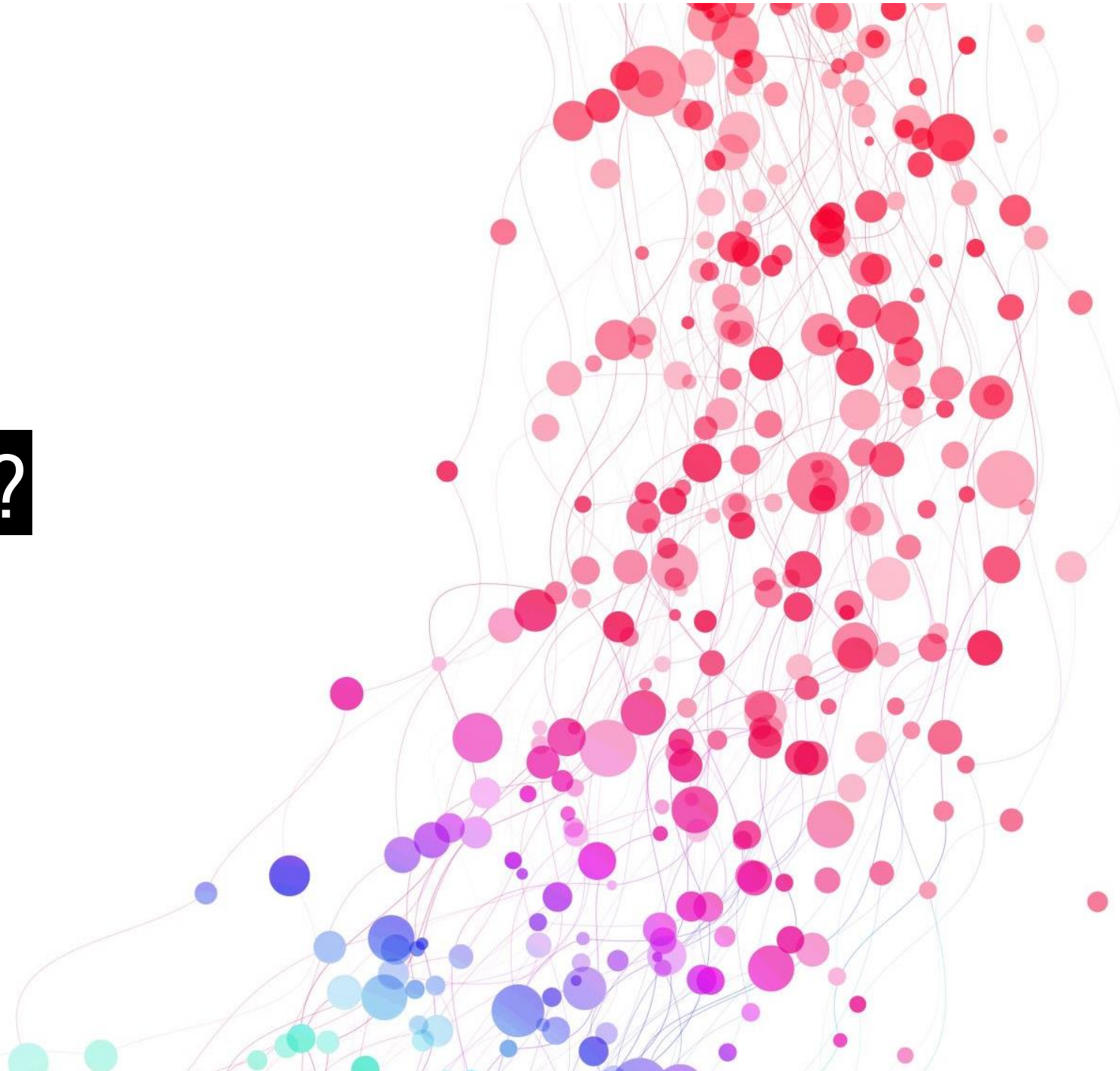
$$\text{IntegratedGrads}_i(x) ::= (x_i - x_i') \times \int_{\alpha=0}^{1} \frac{\partial F(x' + \alpha \times (x - x'))}{\partial x_i} \, d\alpha$$



INTEGRATED GRADIENTS: TAKEAWAYS

3. "Path Methods" are the only methods that satisfy sensitivity, implementation invariance, and completeness

QUESTIONS?

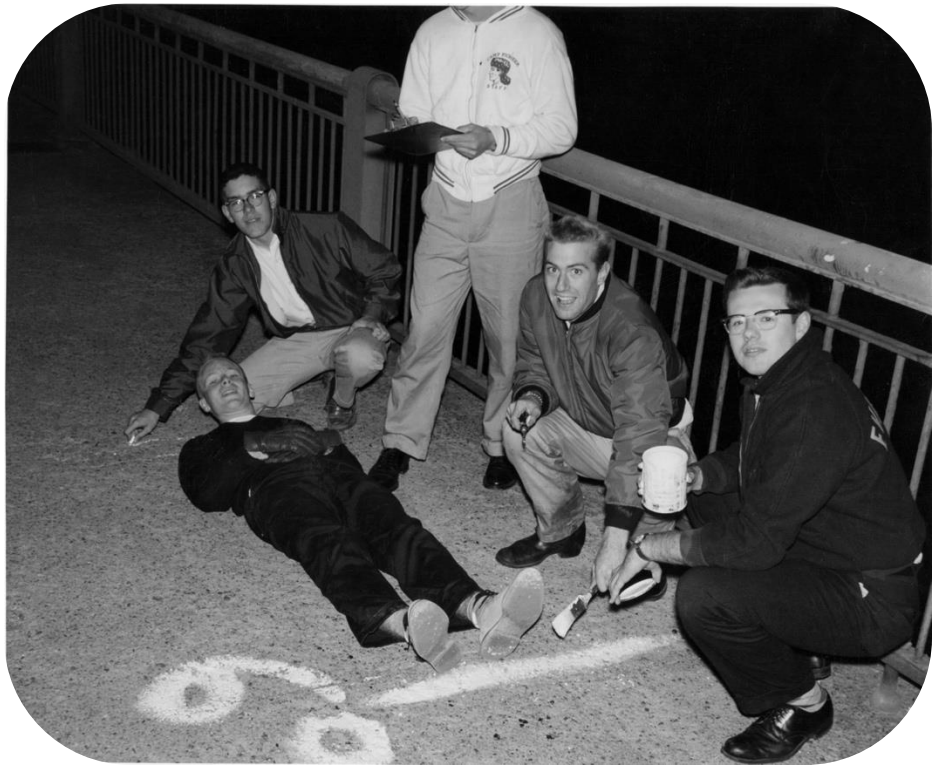


PAPER DISTRIBUTION TIME!

This is Oliver Smoot, he is about 1.702 m high (1 smoot)

PAPER DISTRIBUTION TIME!

This is Oliver Smoot, he is about 1.702 m high (1 smoot)



PAPER DISTRIBUTION TIME!

How many smoots are there in the Mathematical Bridge?

(Hint: one *smoot* is 1.702 m, roughly 0.89 *mateos*)

PAPER DISTRIBUTION TIME!

How many smoots are there in the Mathematical Bridge?

(Hint: one *smoot* is 1.702 m, roughly 0.89 *mateos*)

Answer: 9.072 smoots (approx. 15.44 m)

PAPER DISTRIBUTION TIME!

Paper 1: Adebayo, Julius, et al. "Sanity checks for saliency maps." NeurIPS (2018).

Paper 2: Bau, David, et al. "Network dissection: Quantifying interpretability of deep visual representations." CVPR (2017)