AWS re:Invent

NOV. 28 – DEC. 2, 2022 | LAS VEGAS, NV

BOA401

Explainable attention-based NLP using perturbation methods

Cyrus Vahid
Principal AI Specialist Developer Advocate
AWS

Saousan Kaddami
Research Scientist
AWS

Introduction

Session content

Why explainable AI?

How does XAI work?

Perturbation-based models

Perturbation strategies

Project report

Lessons learned

XAI in Amazon SageMaker

Why explainable AI?

What is explainable AI?

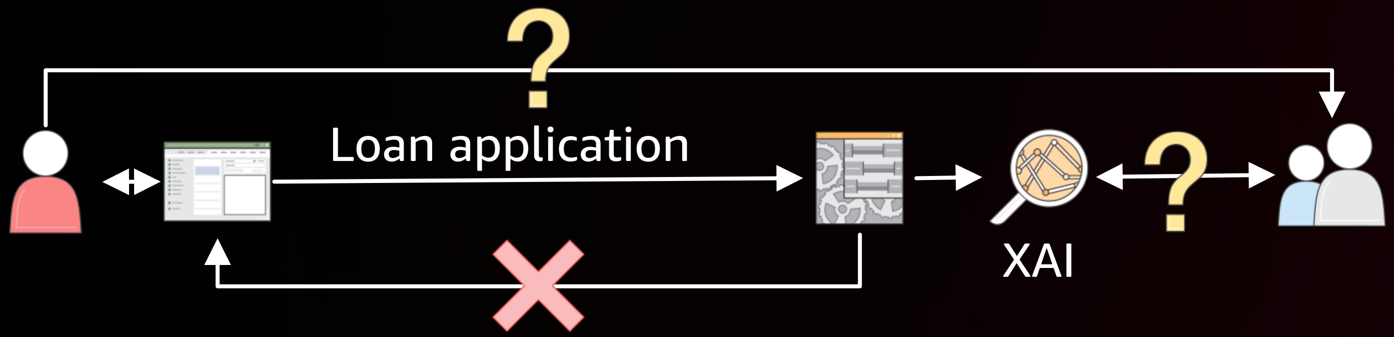
- Explainable AI is the process of explaining the behavior of a model in terms that are intuitive and understandable to human common sense
  - Local explanation: Explaining why a single decision was made
  - Global explanation: Explaining how a model arrives at its decisions

Local explanation

Explaining why a single decision was made

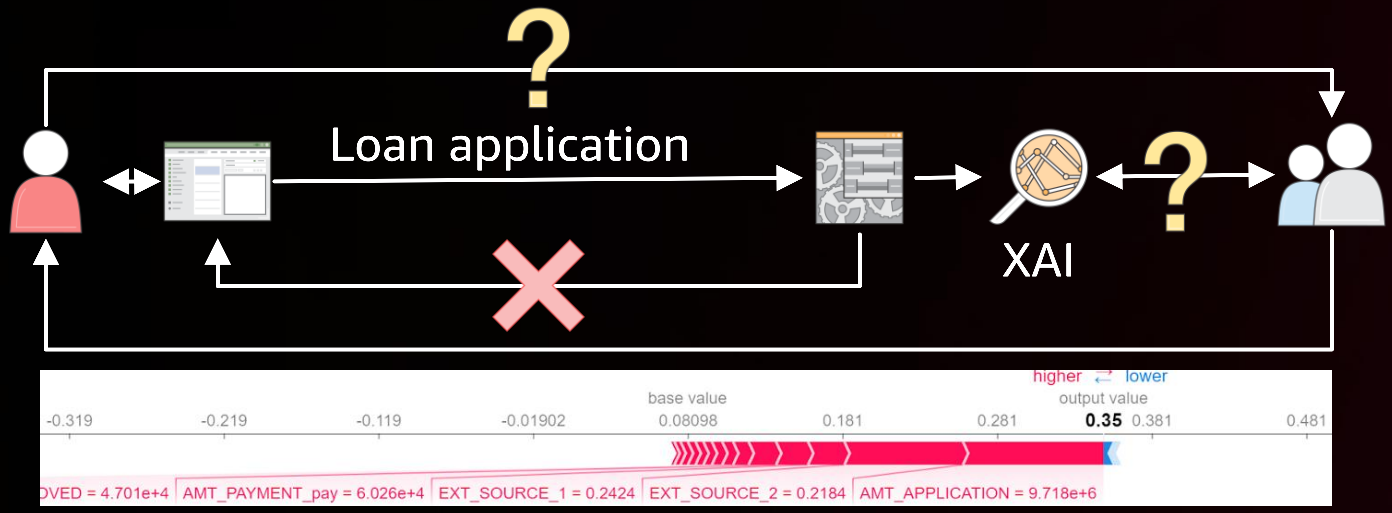
Local explanation

Explaining why a single decision was made



Local explanation

Explaining why a single decision was made



Why we should care

Trust

Transparency

Fairness

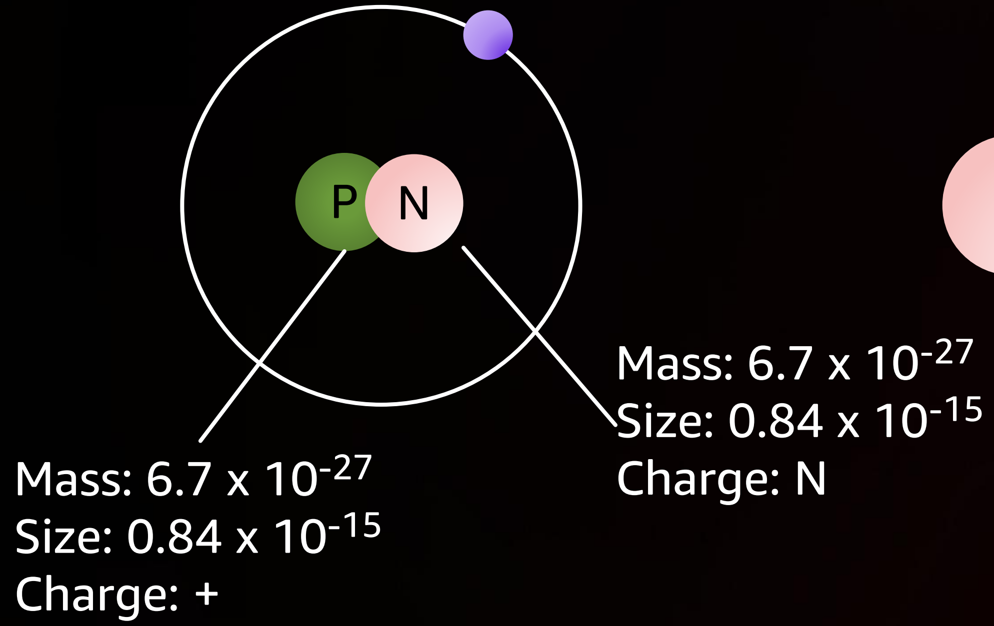
Accountability

Troubleshooting

Training

Information constituency

- The fundamental idea behind constituency is to represent class abstraction
- Example: The study of molecules is at the right level of detail to understand matter, whereas the study of subatomic particles is too detailed

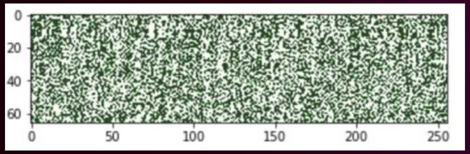


Complex models don’t have information constituency

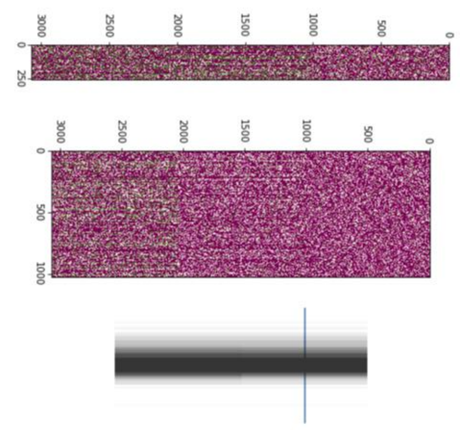
Three-layer text generator

```
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 embedding (Embedding)       multiple                  16896

 gru (GRU)                   multiple                  3938304

 dense (Dense)               multiple                  67650

=================================================================
Total params: 4,022,850
Trainable params: 4,022,850
Non-trainable params: 0
_________________________________________________________________
```

Weight visualization



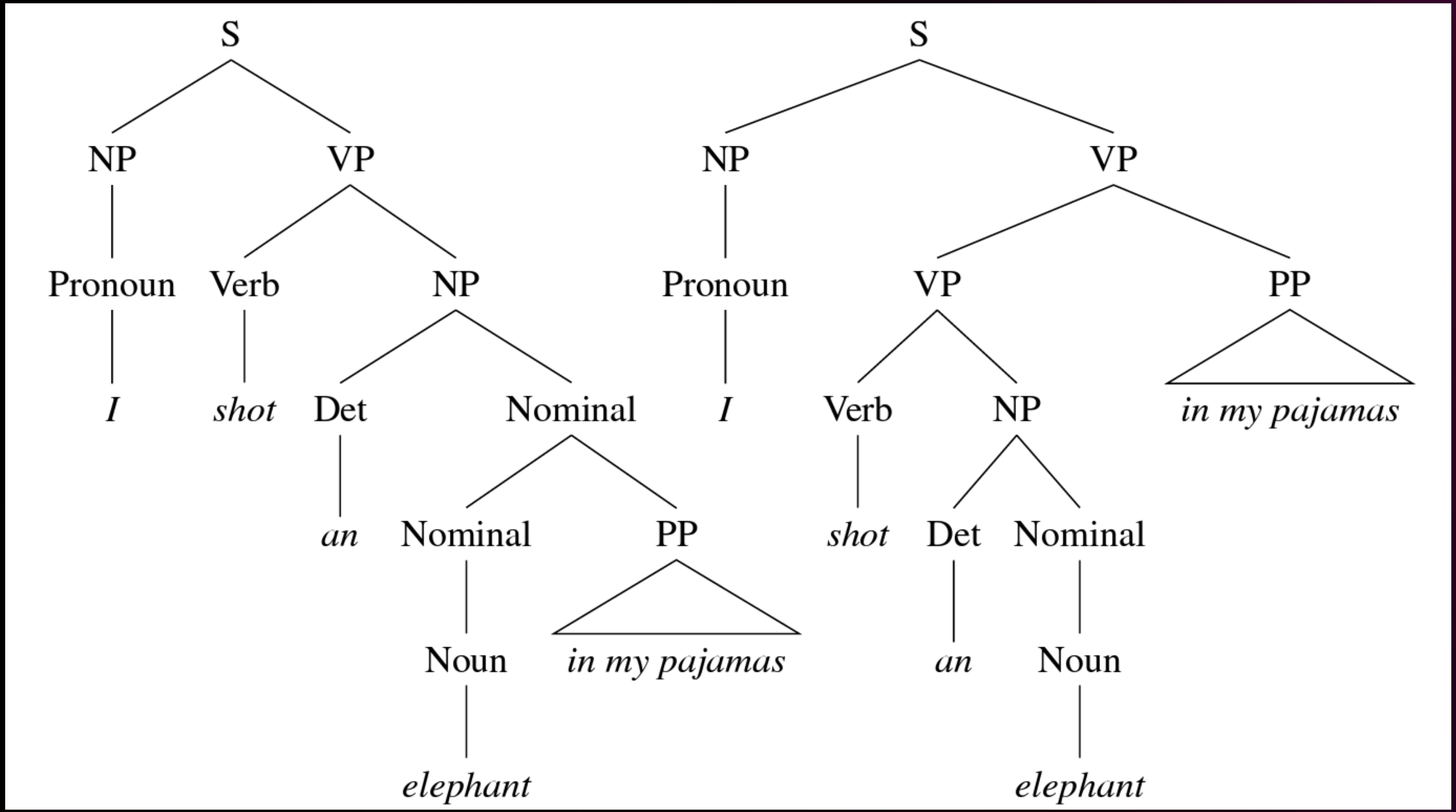
Embedding layer



GRU layer; T: gate weights, M: recurrent weight, B: bias

Dense layer; L: kernel matrix; R: bias vector

This is what we actually understand



Left: Humorous reading,
elephant is in the pyjamas

Right: Person shot the
elephant wearing pyjamas

https://web.stanford.edu/~jurafsky/slp3/13.pdf

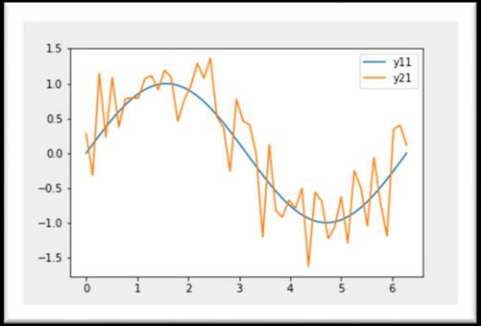
How does XAI work?

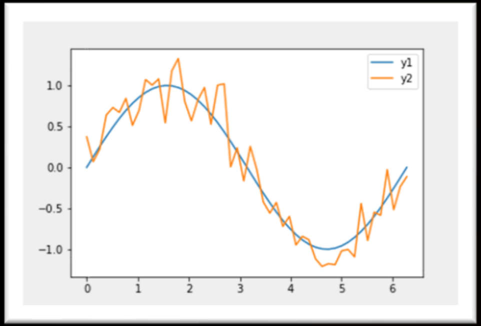
Our context

- Goal: Finding a simple (surrogate) model that explains the original model
- Perturbation-based: Partitioning the input to super-pixels; turning the super-pixels on and off randomly and measuring the effect on the output
- Feature attribution: Discovering how important each super-pixel is to the output
- Local explainability: Explaining individual decisions
- Blackbox approach: We only know what goes in the model and what comes out
- Post-hoc: Explainability is applied to the model after the model is built

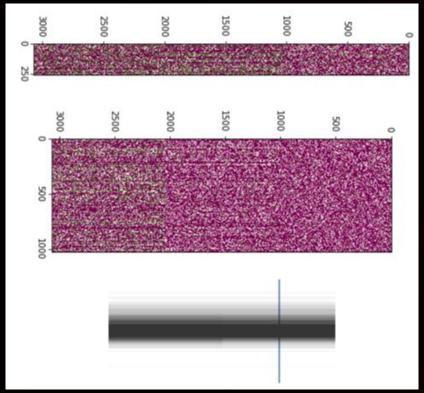
Characteristics of a good explanation

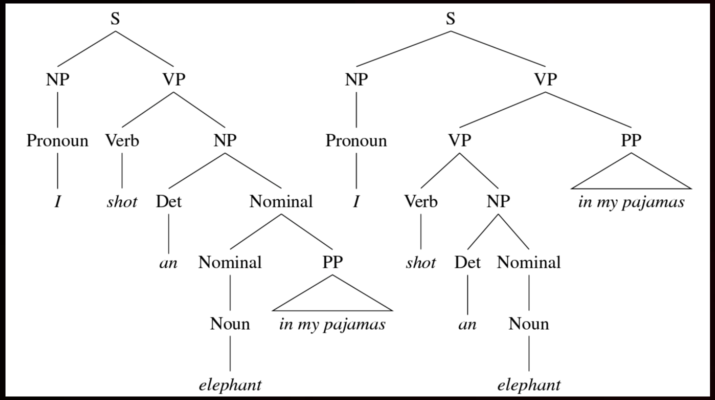
High local fidelity



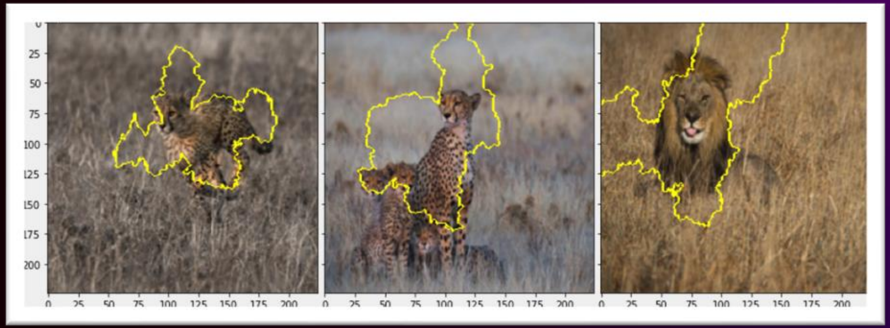


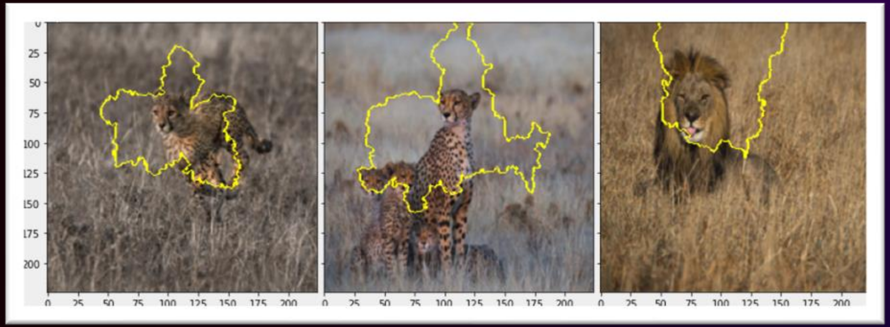
Intuitive interpretability





Model-agnostic



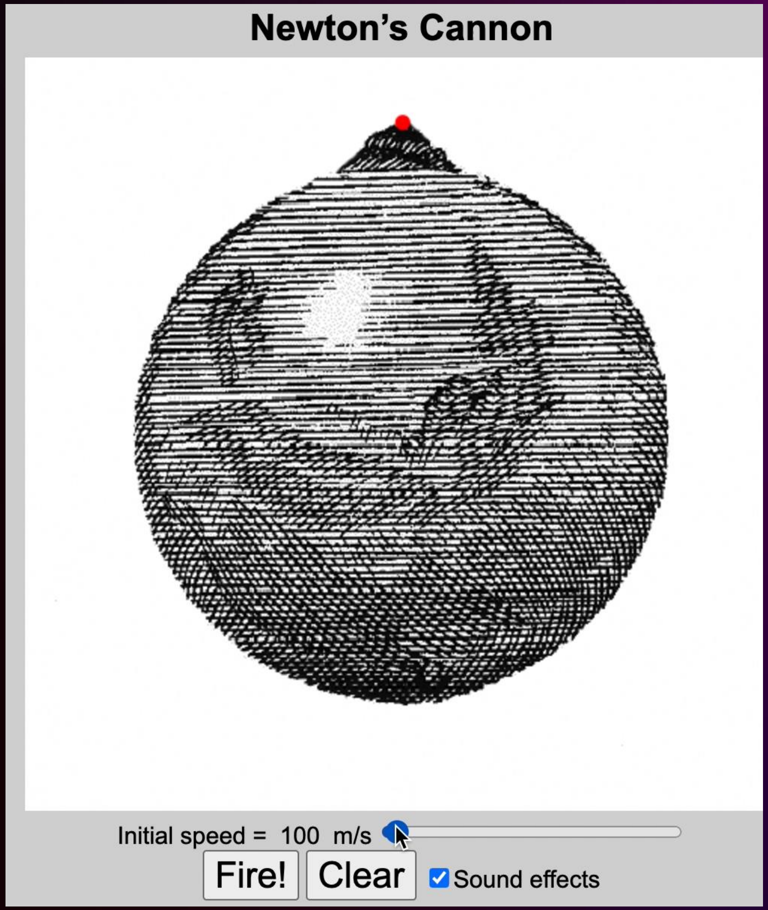


LIME

Local Interpretable Model-agnostic Explanation

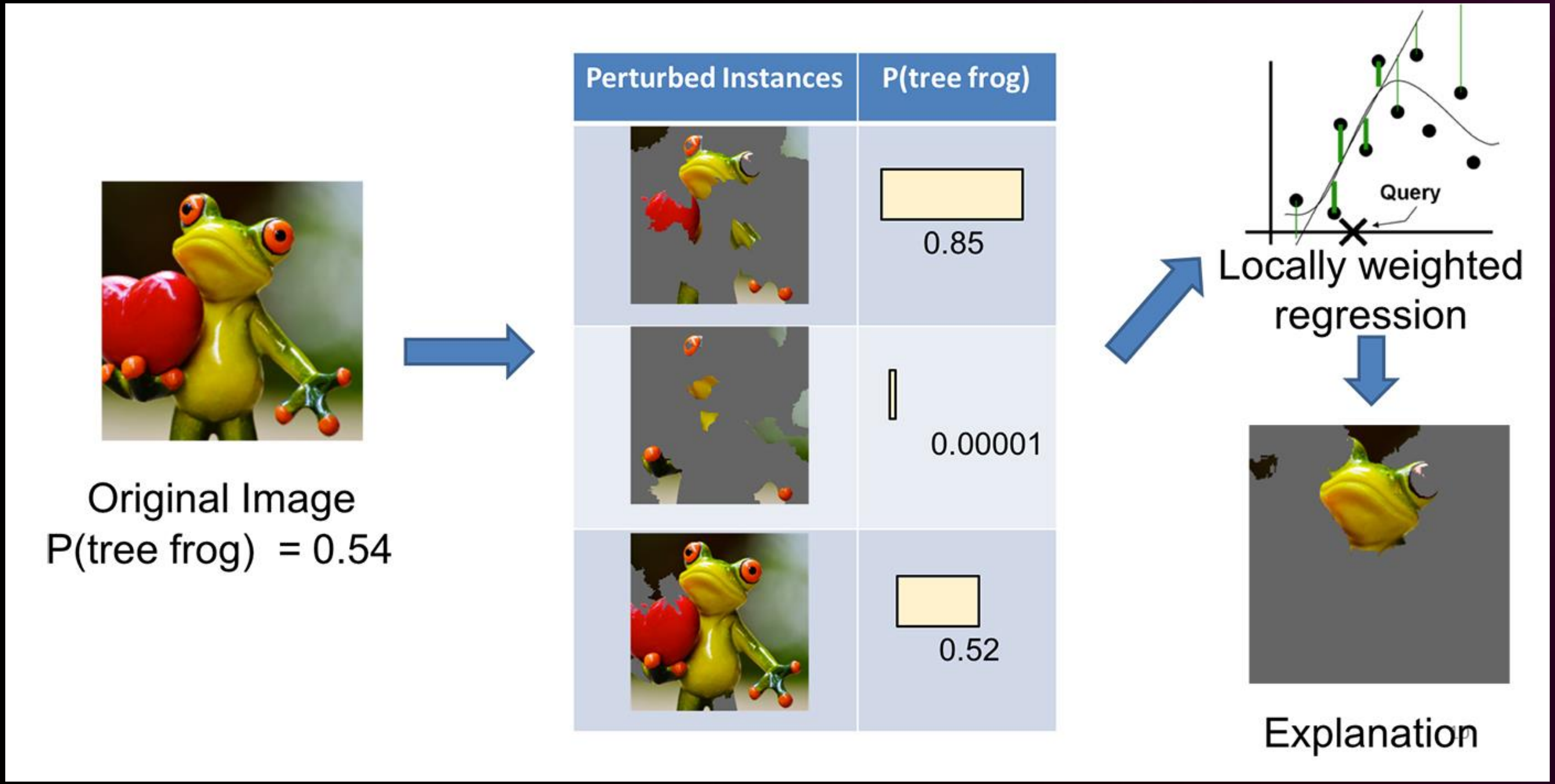
LIME

- The idea behind LIME is to perturb the input and see how it affects predictions
- LIME provides a local explanation by approximating local decisions to a simpler linear model
- Example: Newtonian physics ignores the curvature of the Earth when calculating the trajectory of a ball for short distances



https://physics.weber.edu/schroeder/software/NewtonsCannon.html

LIME – An example approach



1 – The original data is transformed to the dataset with perturbations
2 – Some of the regions of super pixels are turned off
3 – A regression model learns the effect of omitting turned-off regions on the original prediction
4 – Highest impact regions are presented as explanations
(Source: Marco Tulio Ribeiro, Pixabay)

SHAP

Cooperative n-person games

- Cooperative games measure how forming coalitions can help players in a competitive situation
- The goal is to calculate each person's contribution to the success of the collective
- Question: How much does each player contribute to a goal scored by a football team?

Definitions – Coalition function

- A sales team with three members – an account manager, a solutions architect, and a technical account manager (A, B, and C respectively) – can have the following payout table for closing a 250K deal:

Coalition	Payout
ϕ	0
A	100K
B	80K
C	65K
A,B	160K
A, C	140K
B,C	100K
A,B,C = Ω	250K

- $v(A) > v(B) > v(C)$
- $v(A+B) < v(A) + v(B)$
- $v(A+C) < v(A) + v(B)$
- $v(B+C) < v(B) + v(C)$
- $v(A+B+C) > v(A) + v(B) + v(C)$

Suggested solutions

- An obvious solution is equal payouts, but this is unfair to the high performing players
- Another solution could be: $\lambda = \left\{\frac{100}{245}, \frac{80}{245}, \frac{65}{245}\right\}$
  - This also is not a good solution as it discourages teamplay

Shapley values were proposed in 1952 by 2012 Nobel Prize laureate Lloyd Shapley in his work on game theory in the context of cooperative n-person games

Unique solution – Shapley values

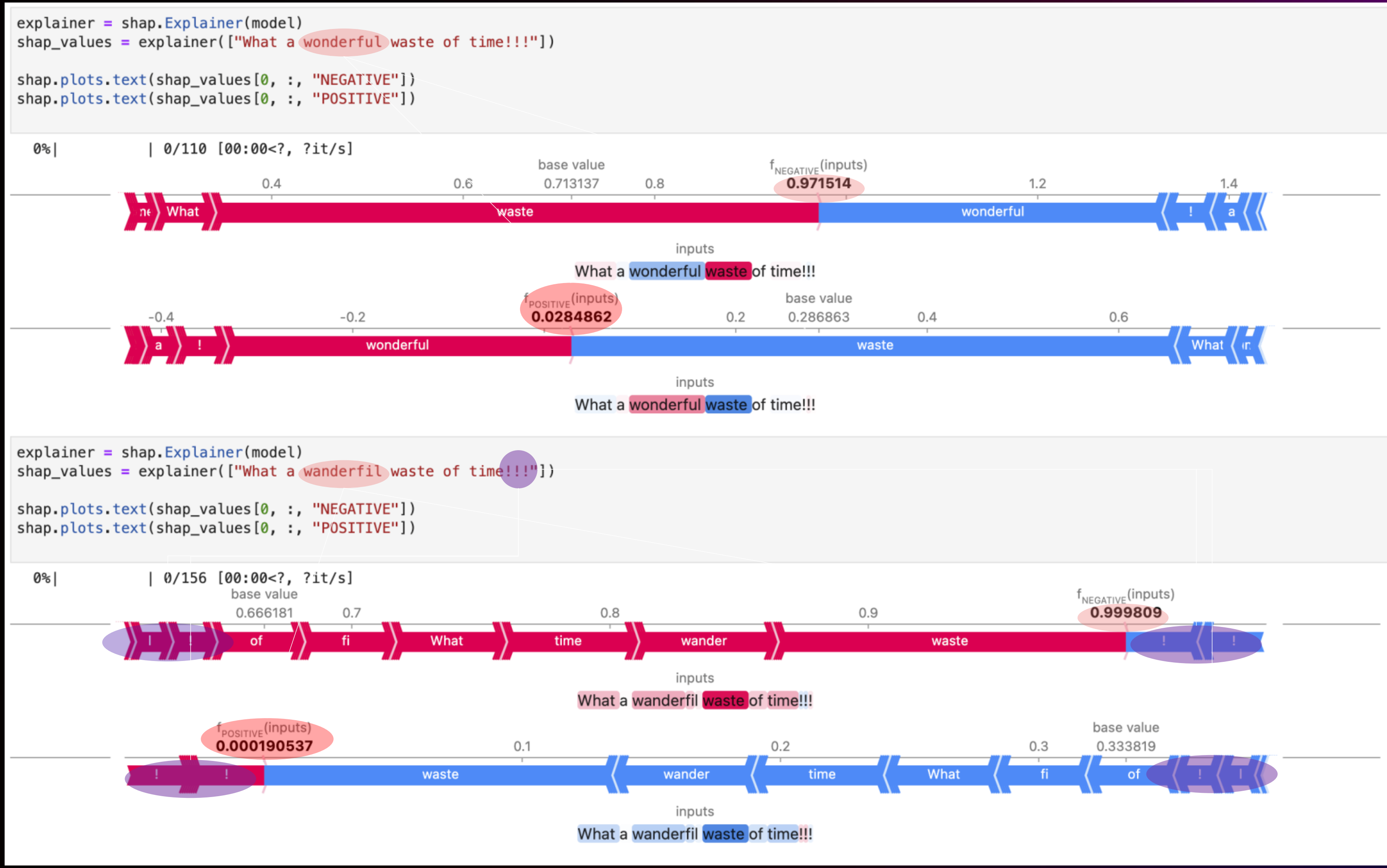
- Provided certain conditions are met, there is a unique solution to the fair division of the payout function among players

Coalition	Payout
A	100K
B	40K
A,B	160K

$$\lambda(N=\{A,B\},v)=\{110,50\}$$

SHAP for sentiment analysis

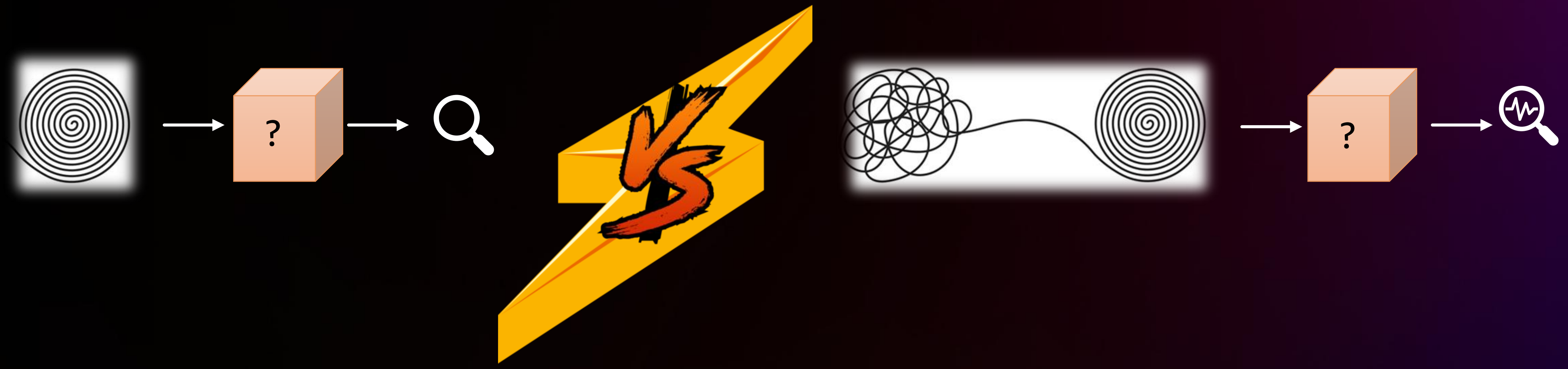
- Tripped by typos
- In this instance, aware of irony
- Unaware of how to deal with punctuation in context detection



Perturbation strategies

Perturbation methods for explainability

- Start: Seminal study by Zeiler and Fergus for images
- Idea: Perturbate input and observe changes in the output



Examples of perturbation approaches on different data

Data	Potential Method	Process
Images	Occlusion	Make perturbations **patch** by patch or **pixel** by pixel
Tabular	LIME	Approximate black box models with a **local interpretable** one
Videos	Temporal masks	Make normalized **freeze** and **reverse** perturbations
Reinforcement learning entities	Rewards corruption	Disturb reinforcement learning agents with **corrupted rewards**

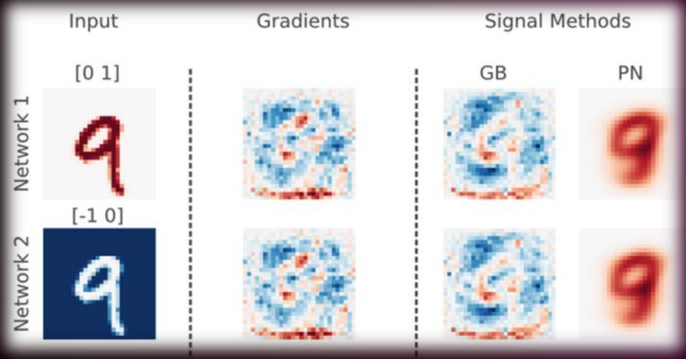
Saliency methods for perturbations

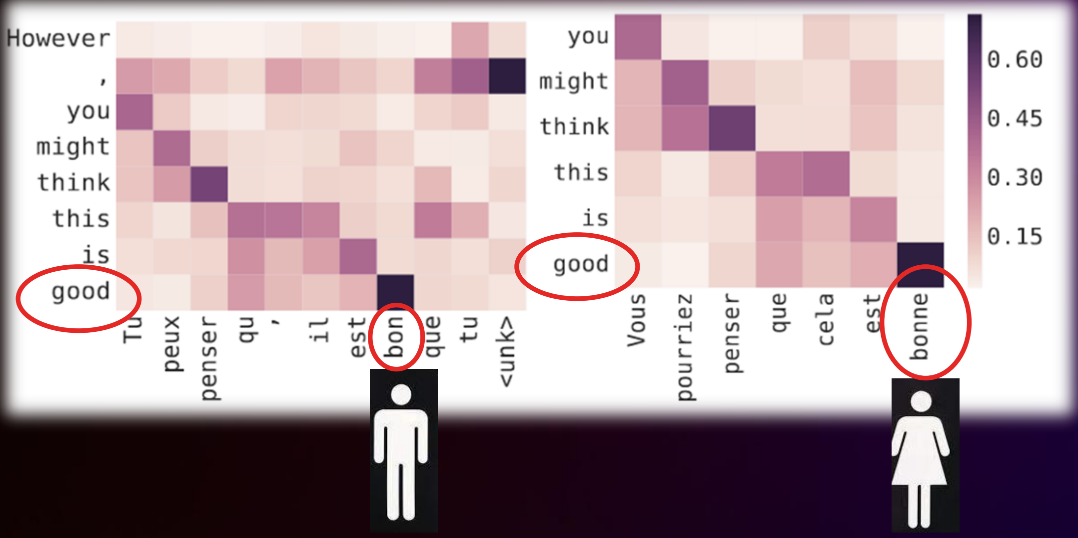
Saliency methods:

- Measure the **consistency** between the rationales **before** and **after** perturbations

Attention-based models:

- Assign a **distribution** of **importance** score over the input tokens to represent their impacts on model predictions
- Use attention **weights** as **token** importance scores





Drastic adversarial perturbation

(1) Naïve approach

```
text_1= 'When life gives you lemons, make lemonade.'
text = 'If we are to teach real peace in this world, and if we are to carry on a real war against war, we shall have to begin with the children.'
text_augmentation(text_1,'MASK' , 1)
```

(2) Attacks by character substitution

```
import nlpaug.augmenter.char as nac
aug_char_random=nac.random.RandomCharAug(action='substitute', name='RandomChar_Aug')
aug_char_keyboard = nac.keyboard.KeyboardAug(name='Keyboard_Aug')
text = 'You must be the change you wish to see in the world.'

text_augmentation(text,aug_char_keyboard , 1)
```

```
'You musy be the change you wich to see in the wkrld.'
```

References :
Adversial attacks to textual data: https://arxiv.org/pdf/2108.04990v2.pdf

Gradual adversarial perturbation

(3) Attacks by synonym substitution

```
import nlpaug.augmenter.word as naw
aug_syn = naw.SynonymAug(aug_src='wordnet', model_path=None, name='Synonym_Aug')

text = 'An apple a day keeps the doctor away.'

text_augmentation(text,aug_syn, 1)
```

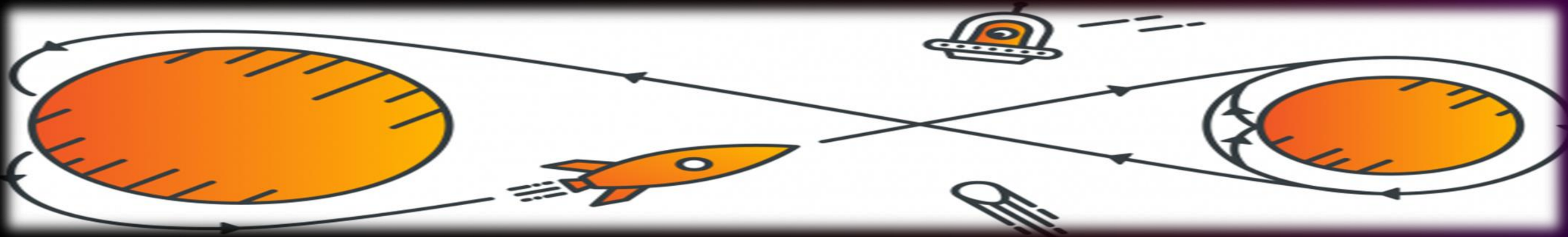
(4) Attacks by paraphrase

```
import nlpaug.augmenter.word as naw
aug_bert = naw.ContextualWordEmbsAug(model_path='distilbert-base-uncased', action=ACT, top_k=TOPK)

text = 'You must be the change you wish to see in the world.'

text_augmentation(text,aug_char_keyboard, 1)
```

References :
Adversarial attacks to textual data: https://arxiv.org/pdf/2108.04990v2.pdf

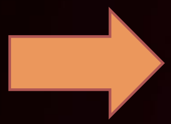
Pitfalls of perturbation methods

- Risk of OOD (Out Of Distribution) generation
- Interpretability vs robustness



Original Sentence:

"This article includes answers what options have for **software intel based unix system**"

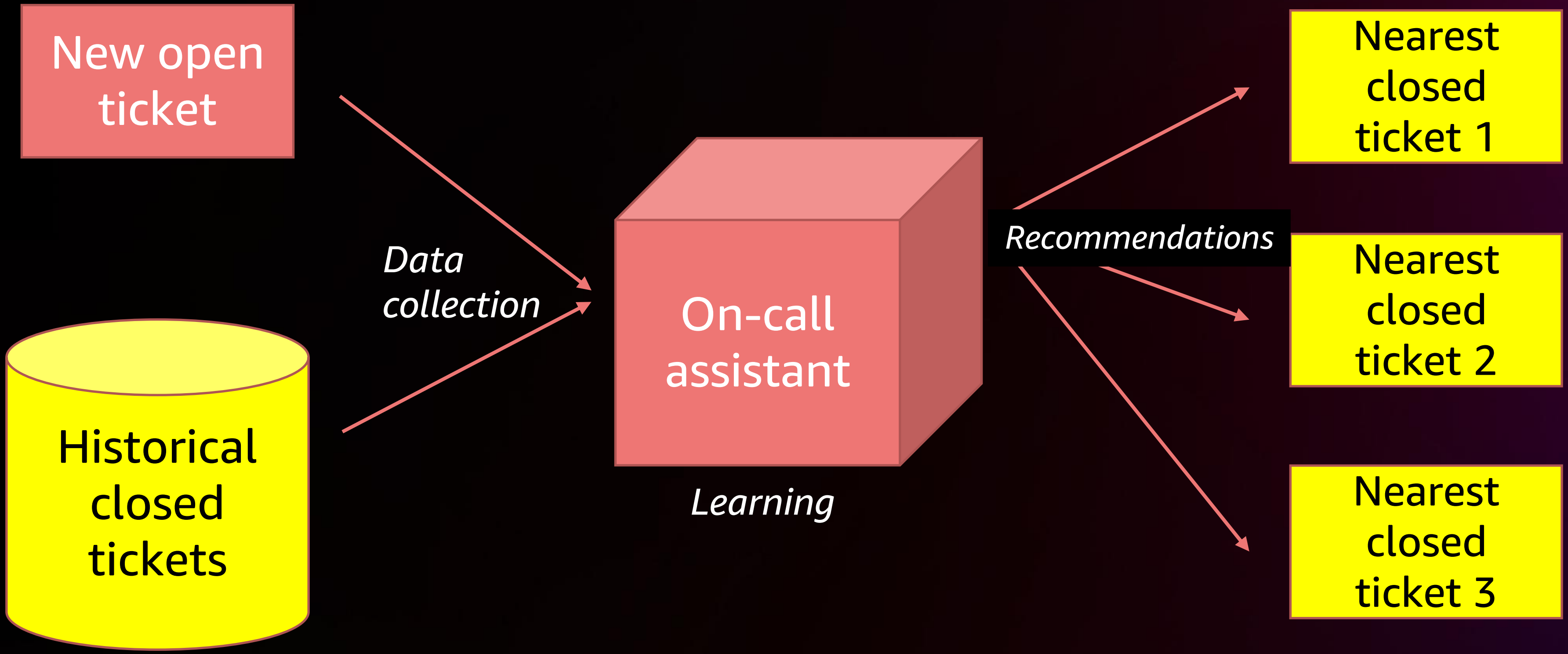


Output Sentence:

"This article explores a recent study on a large scale of **global climate system climate change** and **climate warming"**

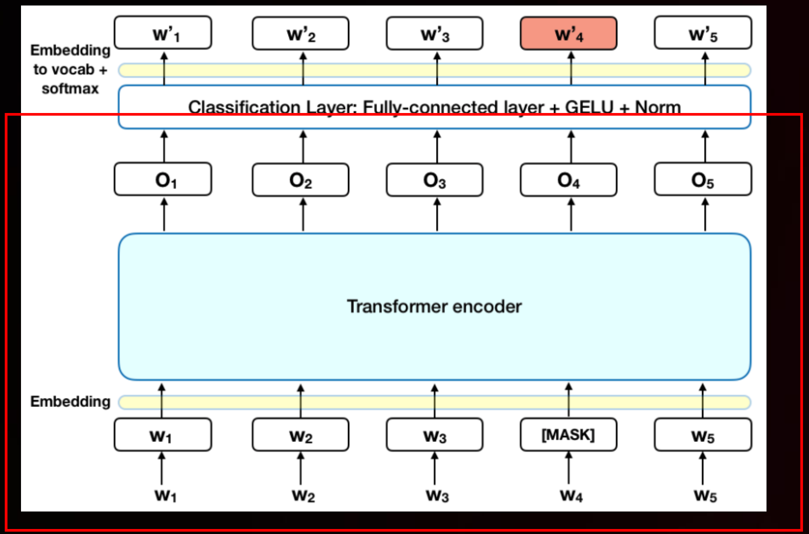
Project report

The on-call assistant model



Natural language DL attention-based encoder

1. Encoding of ticket descriptions into vectors after text cleaning:
   - Removal of common expressions (for example, ‘Describe Current Behavior’)
   - Cleaning from digits + stemming
   - BERT encoding:

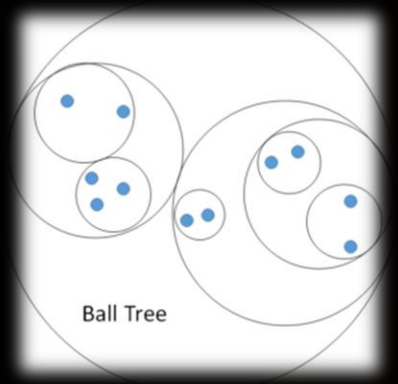


$$w_i \Rightarrow 12\text{x}768$$
$$v_i \Rightarrow 1\text{x}768$$
$$\mathrm{T} = \frac{1}{N} * \sum_{i=1}^{N} v_i$$

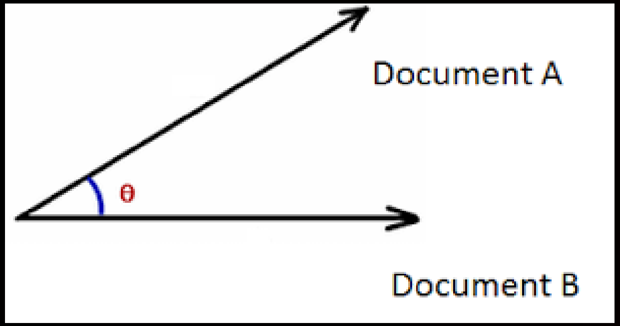
Unsupervised model for nearest neighbors

2. Looking for n (here, 4) nearest neighbors from the closed tickets:

- Ball-Tree approach:



- Brute Cosine Similarity approach:



Challenges and goals

Challenges:

- Lack of quantitative metrics to evaluate, troube shoot and imporve the model
- Consistent requests from the business to provide a text-based explanation of the outputs

Goal from perturbation-based explainability:

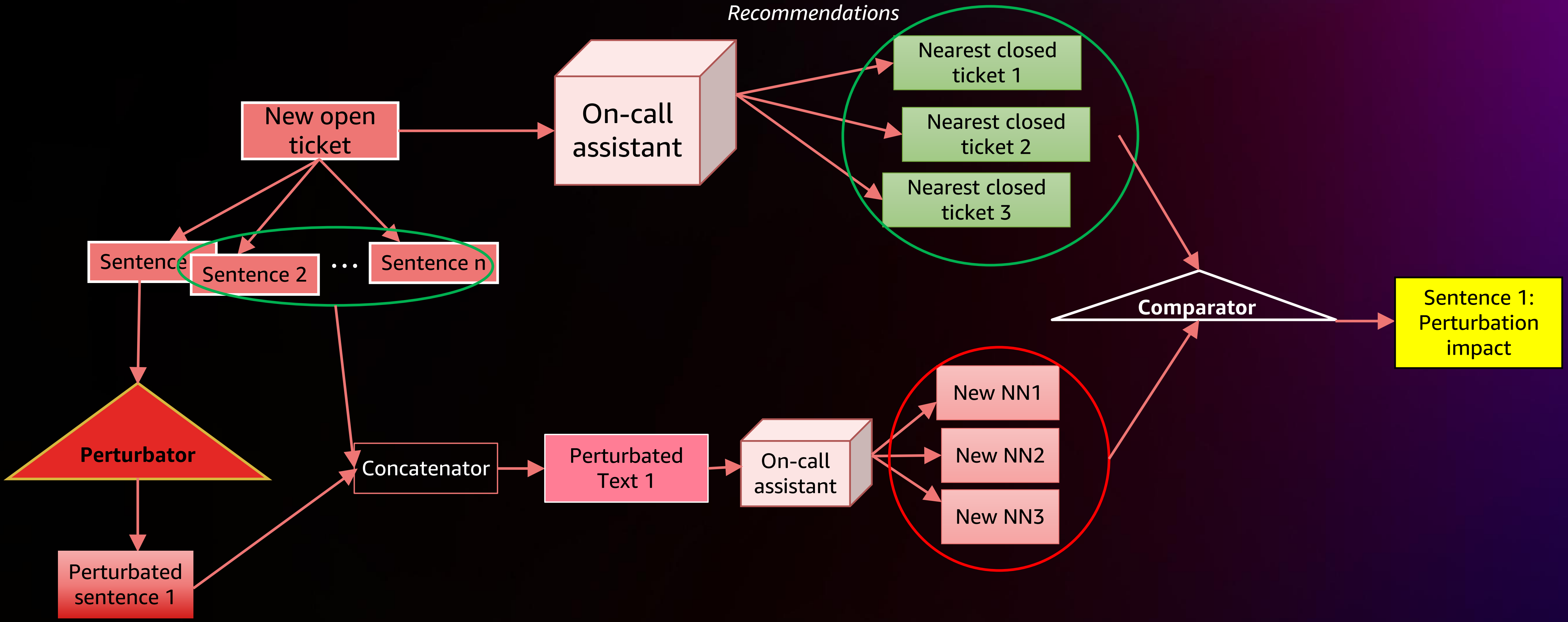
✓ Improve the model

✓ Troubleshoot

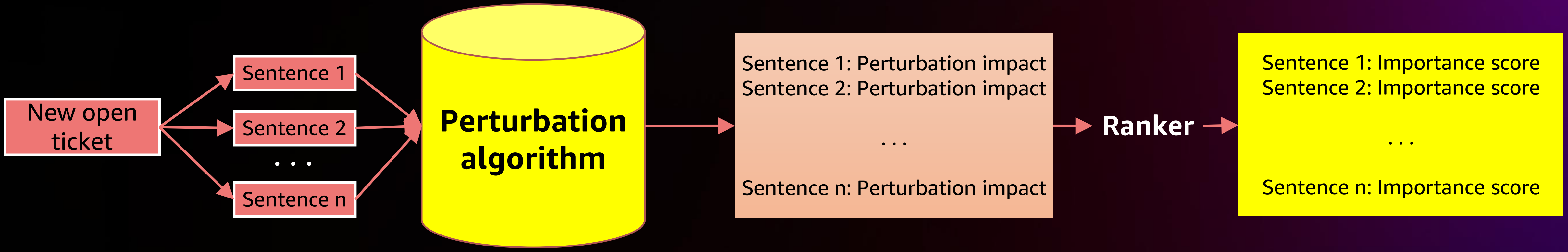
✓ Provide interpetability to the business

Perturbation-based explainability

Our perturbation algorithm



Sentence importance algorithm



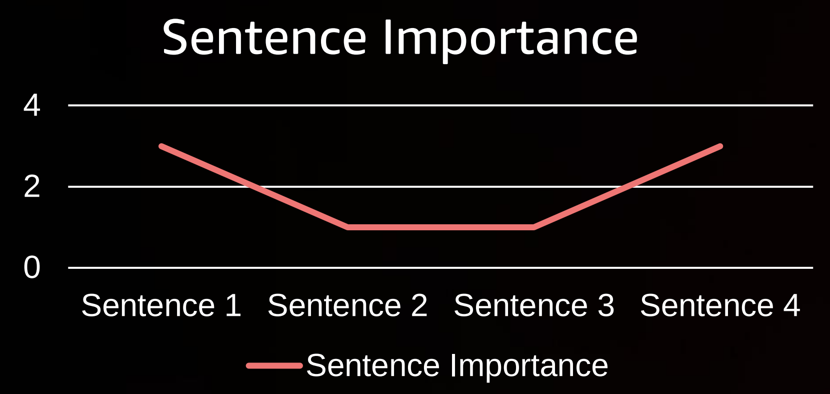
Our results

Importance curve for NN changes

What is happening?

The tool Pandax is facing bugs when trying to refresh the interface. Could you please have a look at the attached screenshot?

Expect an answer from Engineering team within 2 business days.



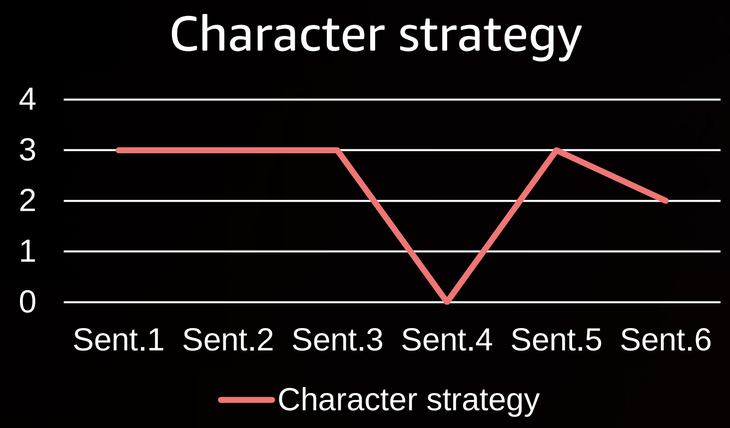
What is happening? We can't get refreshed schedules in the forecasting model Adapt. There should be an issue with the source data. Please have a look at the following link :www.xxxxx.com. Expect an answer from Engineering team within 2 business days.

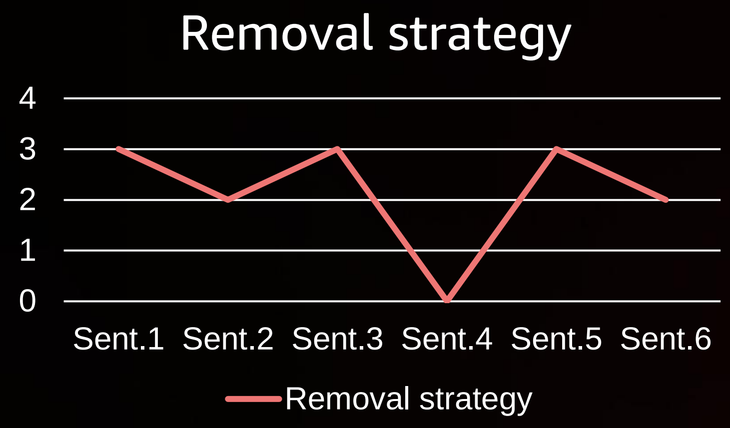
What is happening? I can't open Safari on the vpn. What is the issue? Expect an answer from Engineering team within 2 business days.

What is happening? I can't login to Tokyo API, could you check my credentials? Expect an answer from Engineering team within 2 business days.

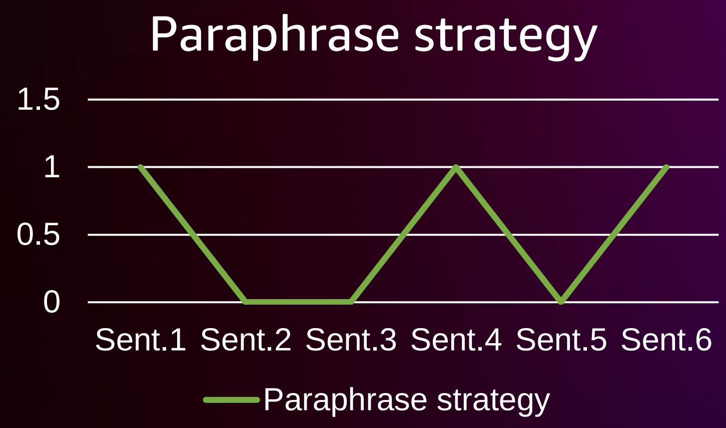
Efficiency of drastic changes for explainability

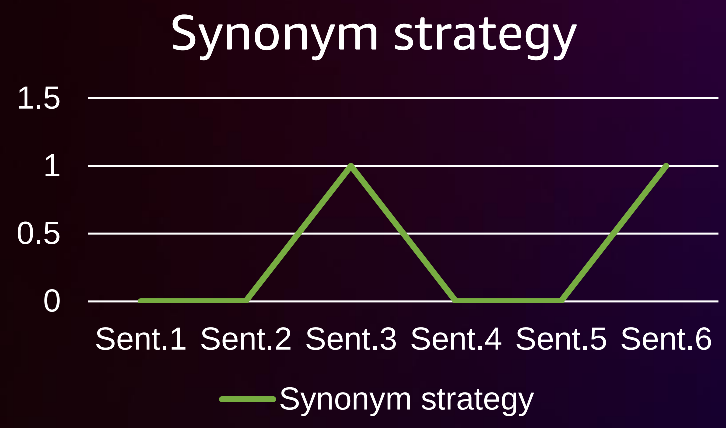
Drastic





Gradual





Outcomes and practical tips

Outcomes and practical tips

Outcomes:

- ✓ Succeeded to troubleshoot the On-Call Assistant
  - ✓ Discovered unimportant repeated sentences with high interpretability scores
- ✓ Provided explainability to the business stakeholders

Technical tips:

- Perturbation methods are generalizable to all DL models
- Explainability - drastic changes; Robustness – gradual changes
- Combine with embedded layers’ analysis ➔ understand neurons’ learning

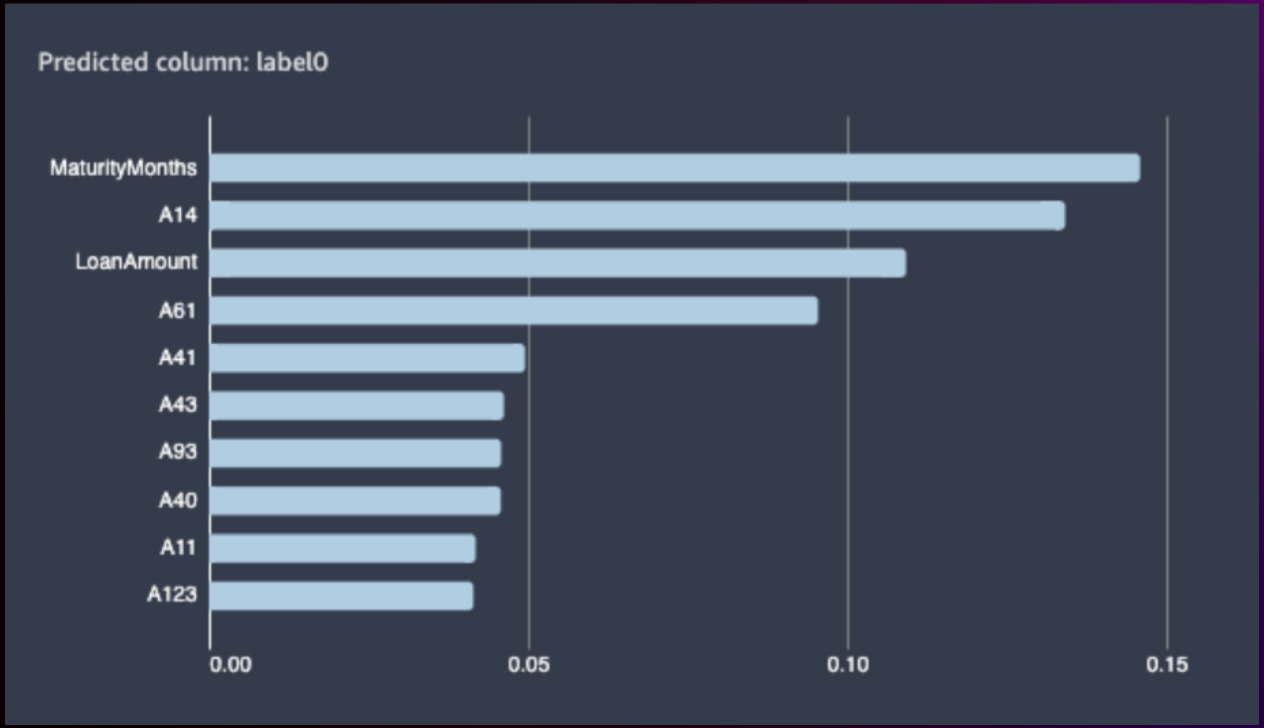
Amazon SageMaker and explainable AI

Amazon SageMaker Clarify

- Provides both local and global explainability
- Detects data biases in pre-processing phase
- Detects biases in trained model
- Monitors your models for emerging biases caused by changes in the real world

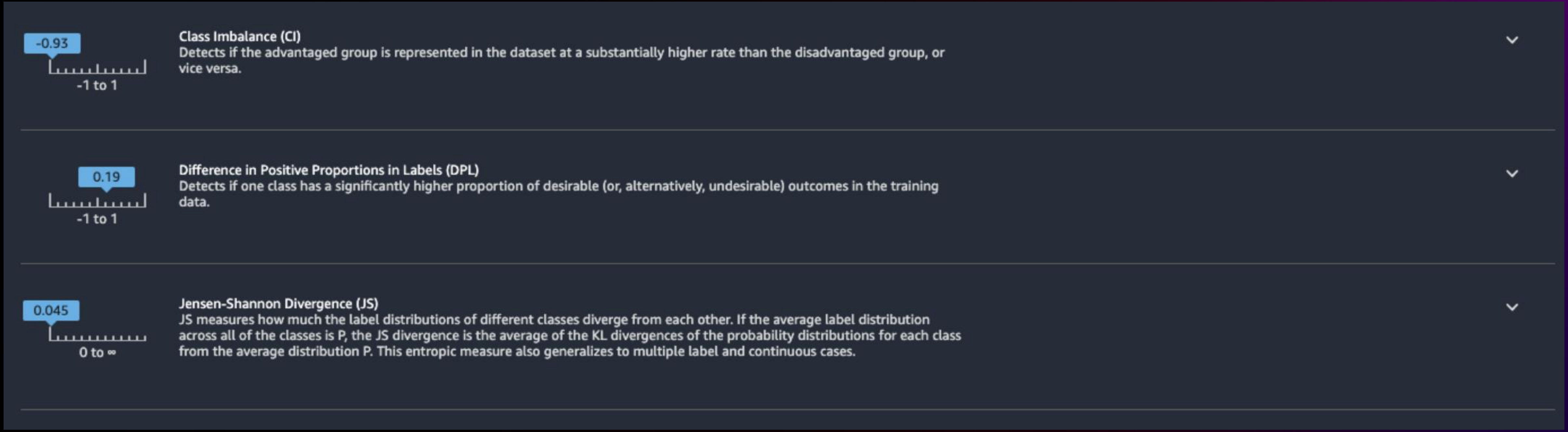
Global explainability

- Providing global explainability by detecting feature attributions and generating reports using integration with SageMaker Experiments



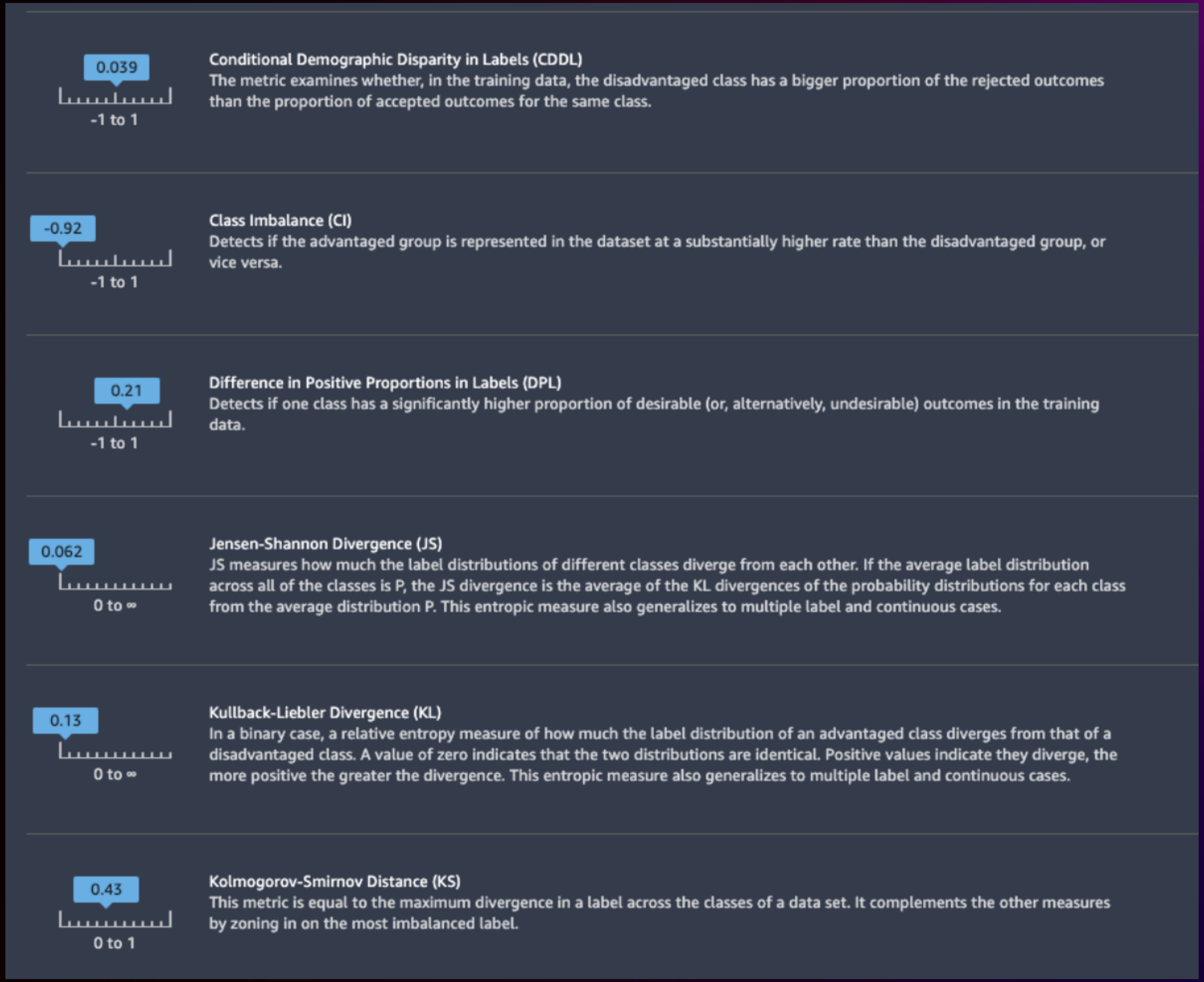
Data bias

- Checks for bias in the data itself during pre-processing phase through integration with Data Wrangler



Model bias

- Checks your trained model for biases, such as predictions that produce a negative result more frequently for one group than they do for another



Monitoring your model for bias

- Although your initial data or model may not have been biased, changes in the world – for example, demographical changes – may introduce biases to a model that has already been trained
- Clarify can monitor the emergence of such biases through integration with Model Monitor

Screenshot of bias monitoring in SageMaker Model Monitor

Monitoring model behaviour

- Changes in the real world can have an impact on feature importance
- A drop in property process might reduce importance of income
- Integration with Model Monitor helps Clarify to provide users with such reports

Thank you!

Cyrus Vahid
cyrusmv@amazon.com

Saousan Kaddami
kaddams@amazon.lu

Please complete the session survey in the **mobile app**