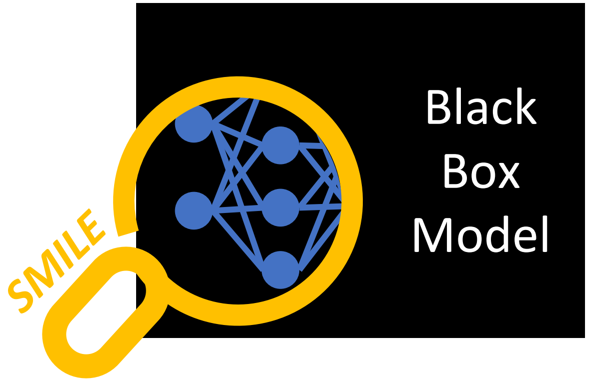


From Machine Learning to Generative AI: Advancing Statistical Model-Agnostic Interpretability with Local Explanations (SMILE)

Aslansefat, K., Hashemian, M., Walker, M., Akram, M. N., Sorokos, I., & Papadopoulos, Y. (2023). Explaining black boxes with a SMILE: Statistical Model-agnostic Interpretability with Local Explanations. ***IEEE Software***, *41*(1), 87-97.

Table of Content

What I am going to discuss

Highlights on eXplainable Artificial Intelligence (XAI)

Some basic information.

LIME Explainability Procedure

Discussing how LIME works on a simple example.

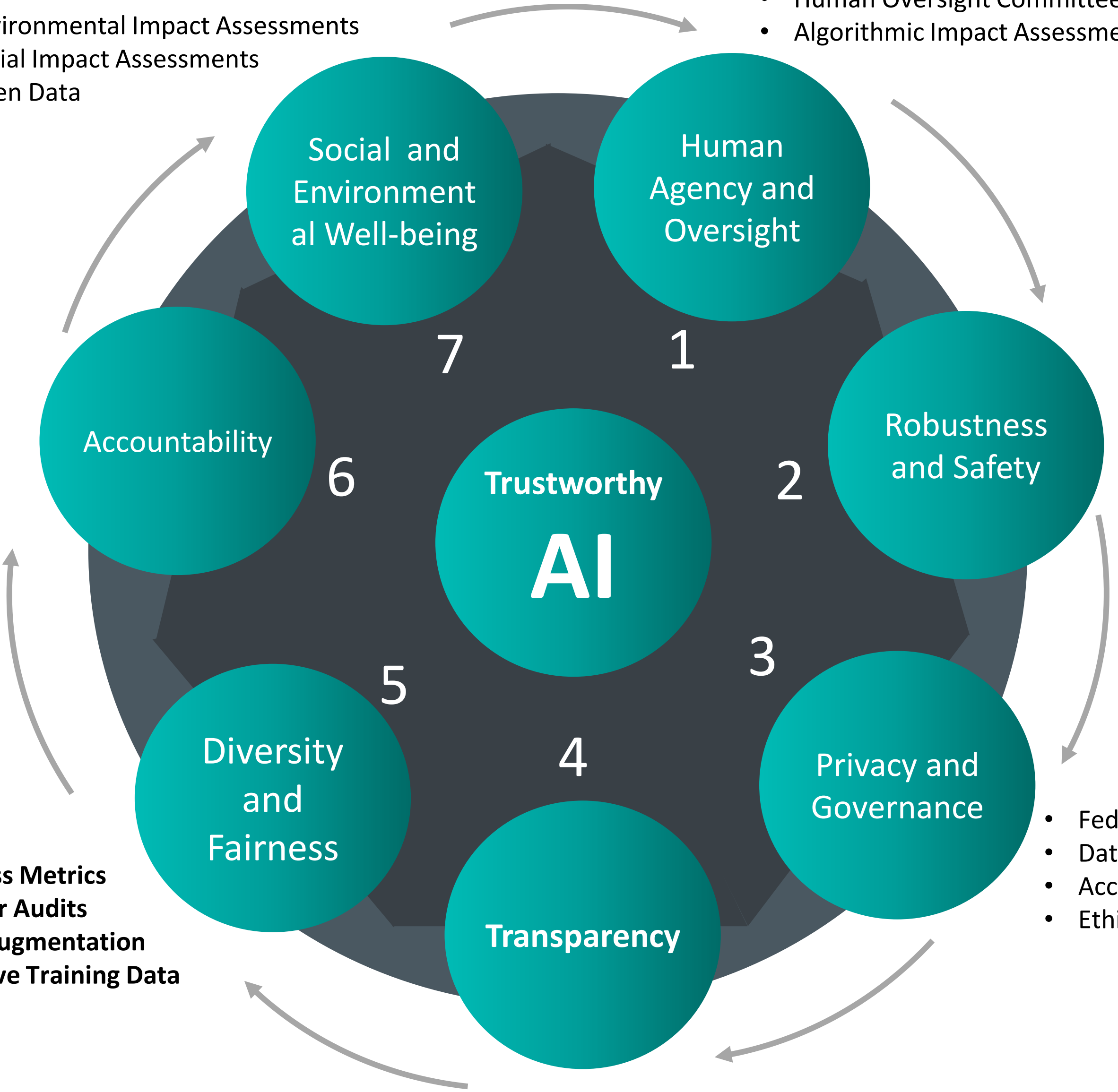
SMILE IDEA

Discussing the idea and providing some experiments.

Expanding SMILE for GenAI

Discussing how SMILE has been expanded for GenAI (LLMs, MLLMs, etc).

EU AI ACT



Trustworthy

AI

1 Human Agency and Oversight

- Human-in-the-loop
- Human Oversight Committee
- Algorithmic Impact Assessments

2 Robustness and Safety

- Robust Data Pre-processing
- Robust Feature Engineering
- Ensemble Methods
- Error Handling and Fail-Safes

3 Privacy and Governance

- Federated Learning
- Data Anonymisation
- Access Control
- Ethical Framework

4 Transparency

- **Explainability**
- **Interpretability**

5 Diversity and Fairness

- **Fairness Metrics**
- **Regular Audits**
- **Data Augmentation**
- **Inclusive Training Data**

6 Accountability

- Data Logging
- Model Versioning
- Auditing
- Risk Management

7 Social and Environmental Well-being

- Environmental Impact Assessments
- Social Impact Assessments
- Open Data

Key Questions

What explanation methods are applied, and are they appropriate for the model and user needs?

Who are the target users of explanations, and how are their needs and understanding levels considered?

Where in the system lifecycle is explainability integrated (e.g., model design, deployment)?

Which model components are explained, and what type of explanations are provided

When is explainability assessed or updated (e.g., continuously, post-deployment)?

Key Stakeholders

Data Scientist

- Technical report
- XAI metrics

- Mitigation strategies
- Improvement recommendations

Business Owner

- Business implications
- Potential risks
- Opportunities
- Technical overview

Regulators

- Compliance shown
- XAI evidence
- Mitigation details

Consumer

Examples:

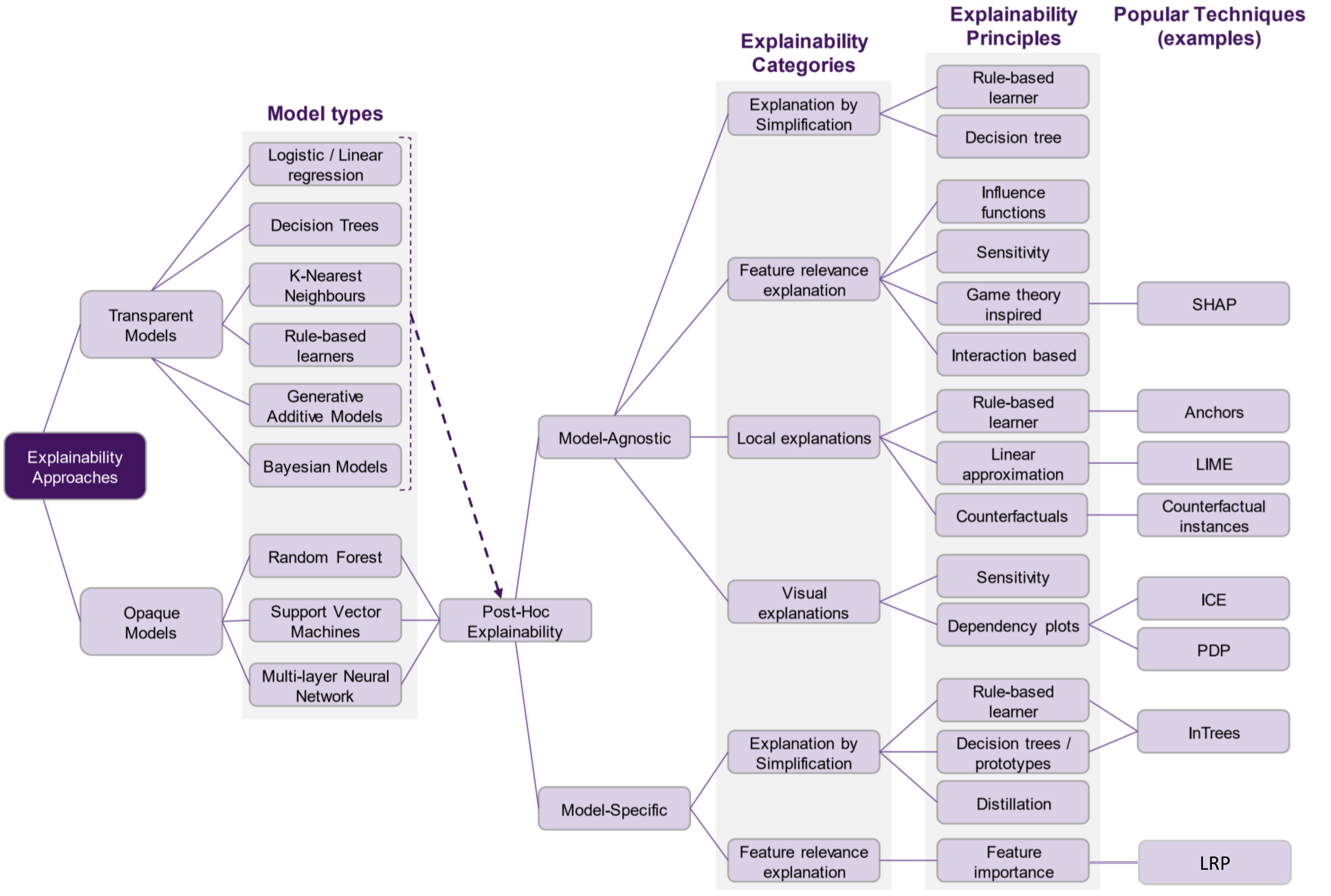
- User-friendly infographics.
- Online user feedback forms.
- Demographic-based testing.
- Personalized & non-biased product recommendations.

Fairness Report:

- Non-technical focus
- Model effects
- XAI steps
- Feedback avenues

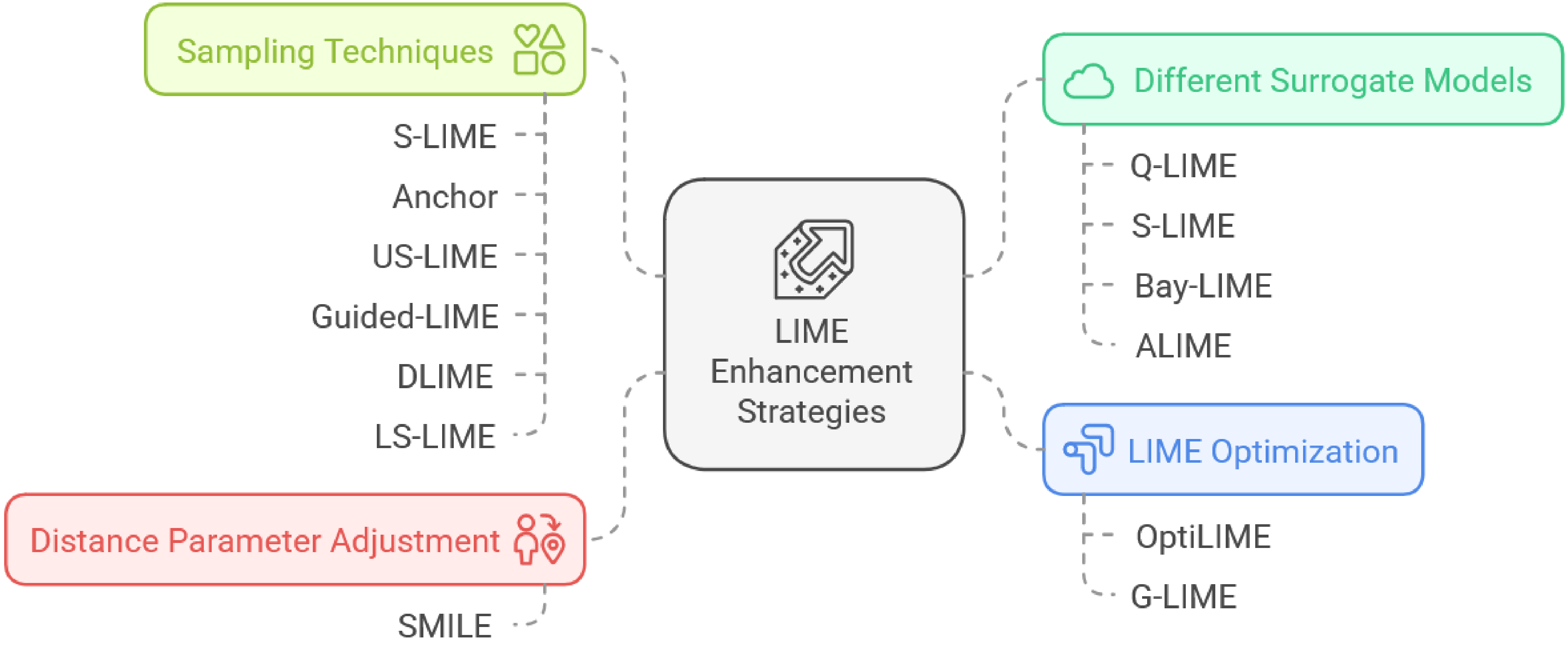
XAI Advocacy Groups

- Bias transparency
- Societal impacts
- Addressing steps
- Real-world evidence



Belle, V., & Papantonis, I. (2021). Principles and practice of explainable machine learning. *Frontiers in big Data*, 39.

LIME-based Approaches



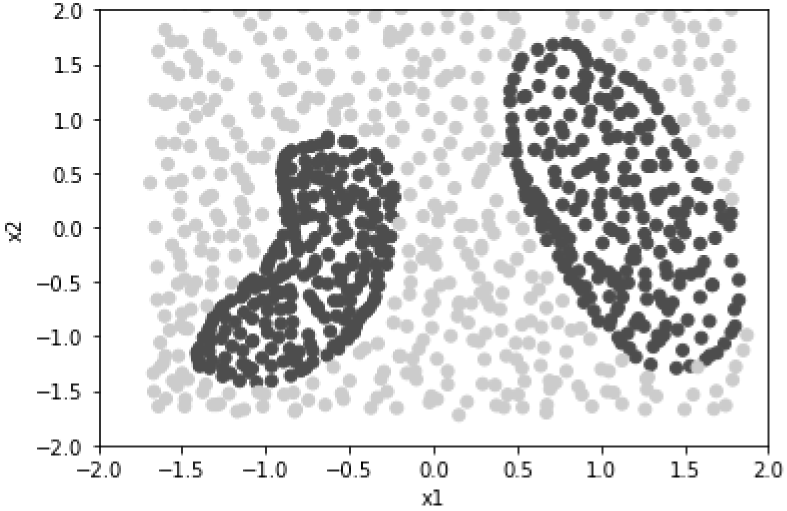
LIME

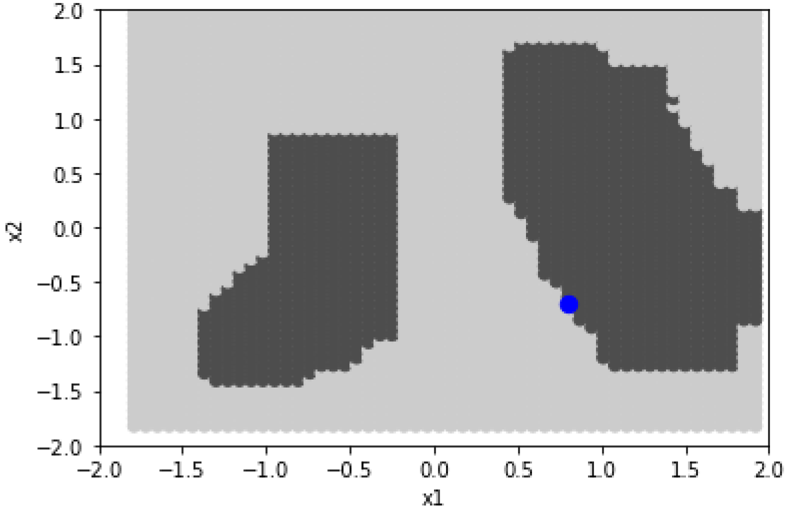
Discussing LIME Algorithm in a simple tabular example

GIF from: storyset.com

LIME: Local Interpretable Model-agnostic Explanations

Let's discuss a simple example of LIME on a tabular hypothetical dataset

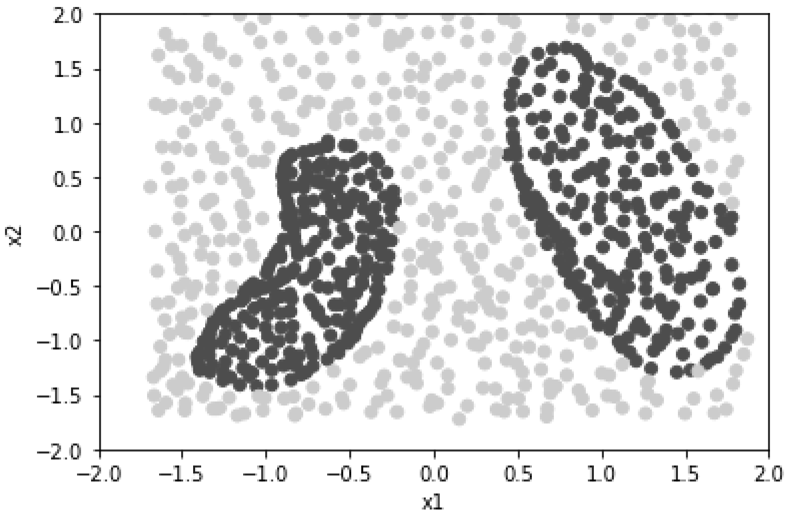


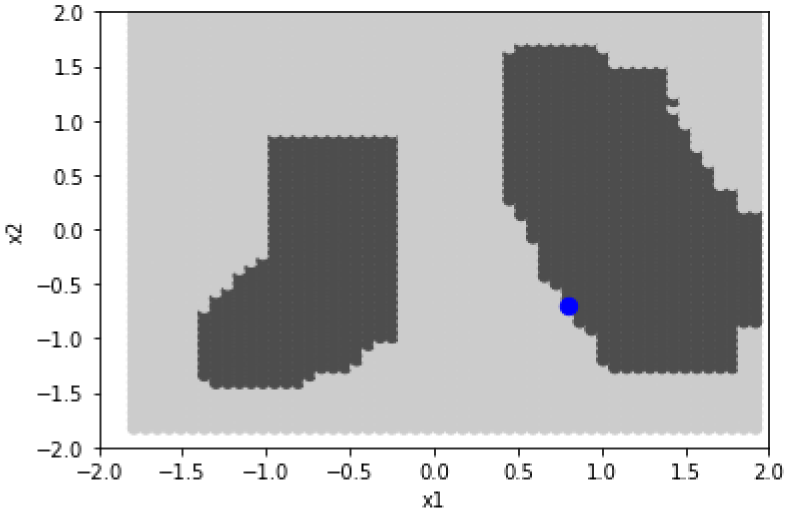


Source: https://arteagac.github.io/blog

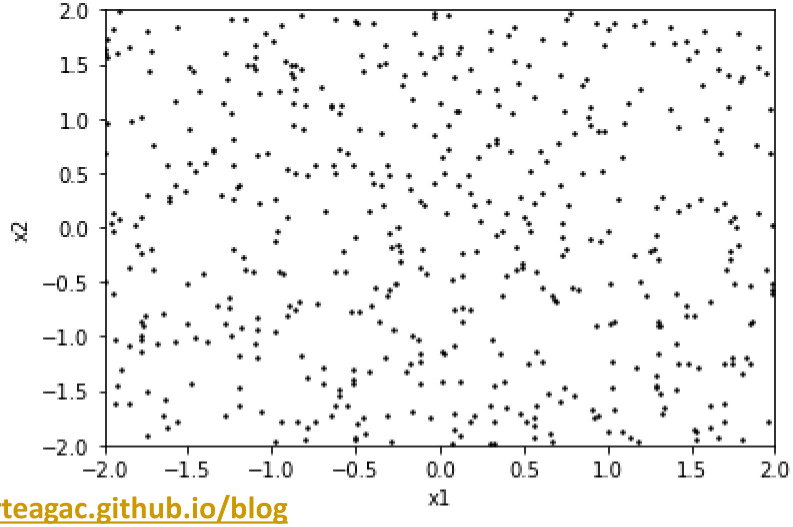
LIME: Local Interpretable Model-agnostic Explanations

Let’s discuss a simple example of LIME on a tabular hypothetical dataset

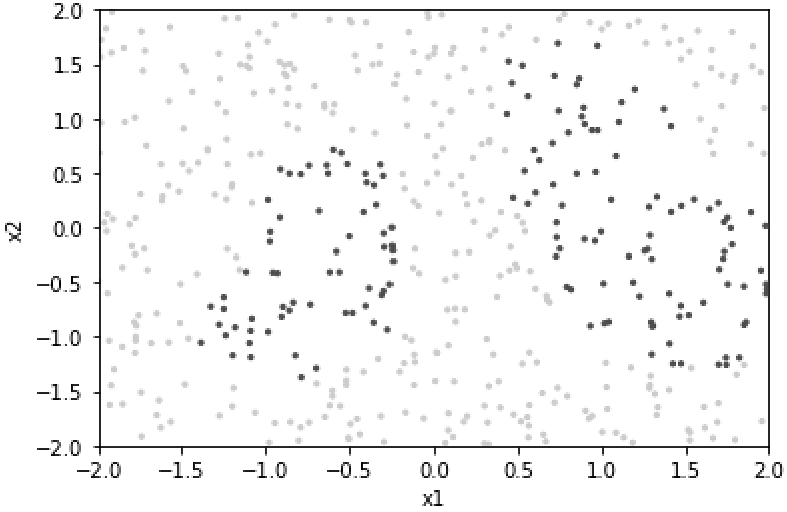




Generating uniform random numbers



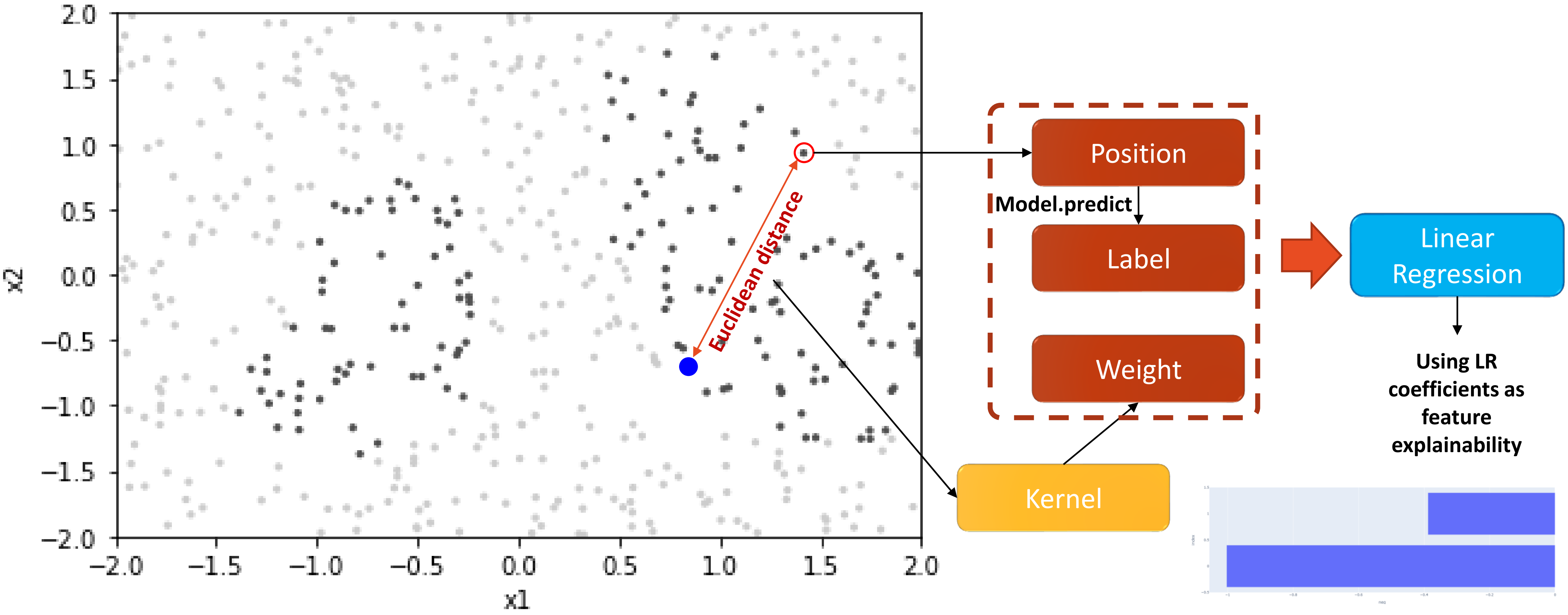
Using model.predict(X_perturb)



Source: https://arteagac.github.io/blog

LIME: Local Interpretable Model-agnostic Explanations

Let’s discuss a simple example of LIME on a tabular hypothetical dataset



Position

Model.predict

Label

Weight

Kernel

Linear Regression

Using LR coefficients as feature explainability

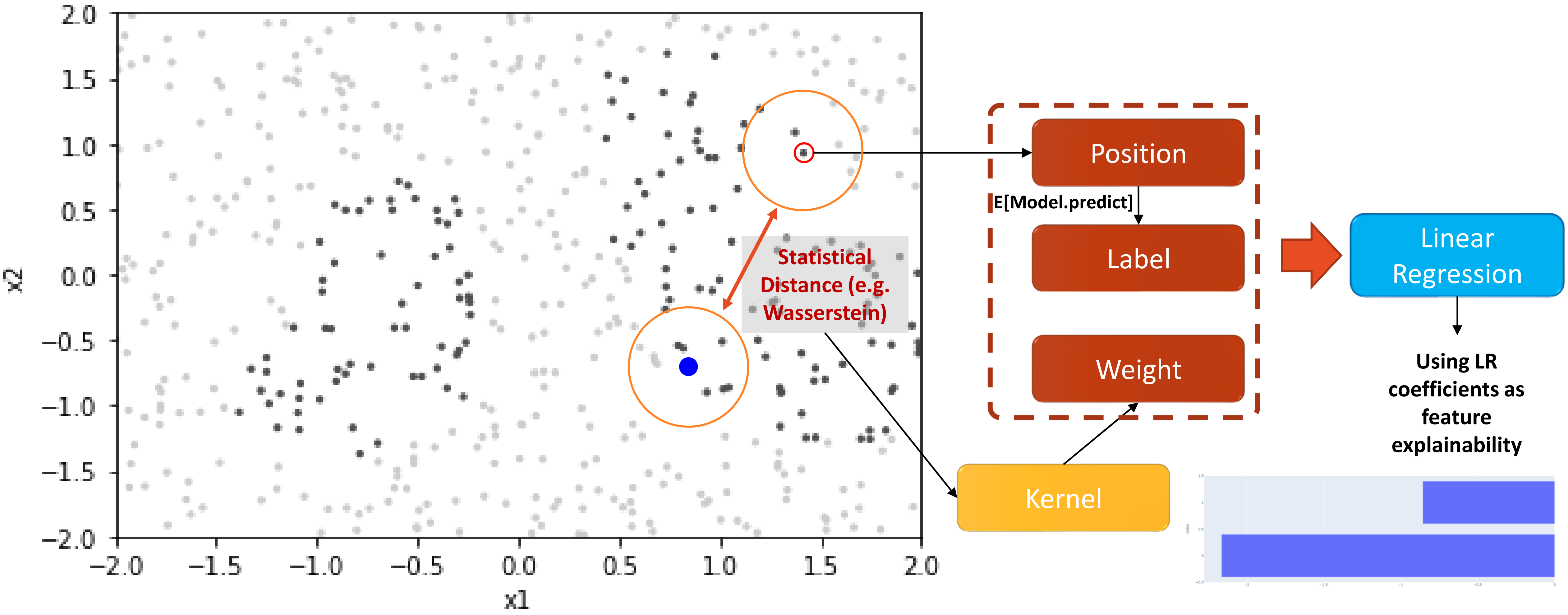
IDEA!

SMILE: Statistical Model-agnostic Interpretability with Local Explanations

GIF from: storyset.com

SMILE: Statistical Model-agnostic Interpretability with Local Explanations

Let’s discuss the same simple example with SMILE



Statistical Distance (e.g. Wasserstein)

Position

E[Model.predict]

Label

Weight

Kernel

Linear Regression

Using LR coefficients as feature explainability

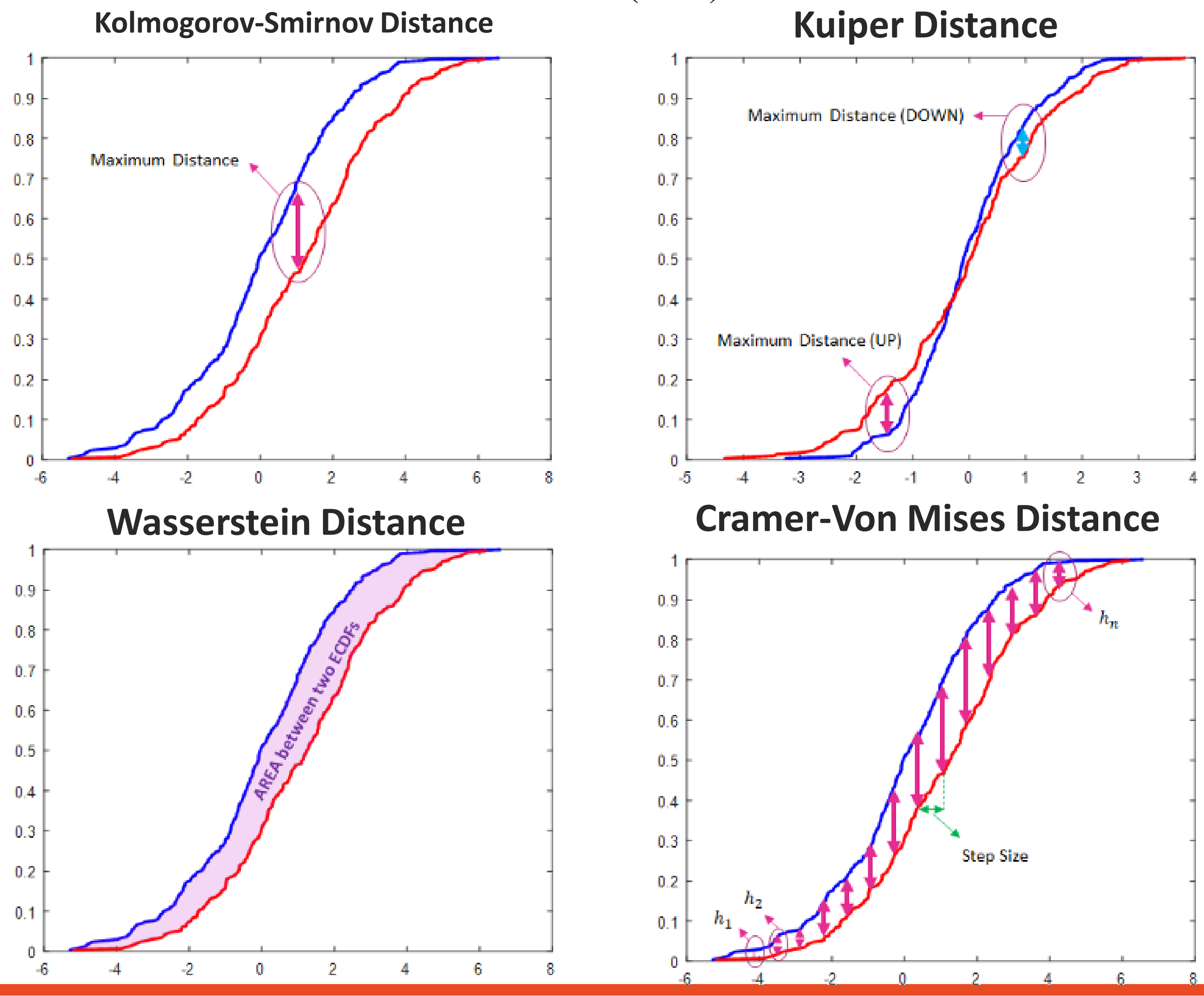
SMILE: Statistical Model-agnostic Interpretability with Local Explanations

Cumulative Distribution Function (CDF) Distance Measures

- Wasserstein
- Kolmogorov-Smirnov
- Kuiper
- Anderson-Darling
- Cramer-Von Mises
- Wasserstein-Anderson-Darling (WAD)

SMILE: Statistical Model-agnostic Interpretability with Local Explanations

Cumulative Distribution Function (CDF) Distance Measures



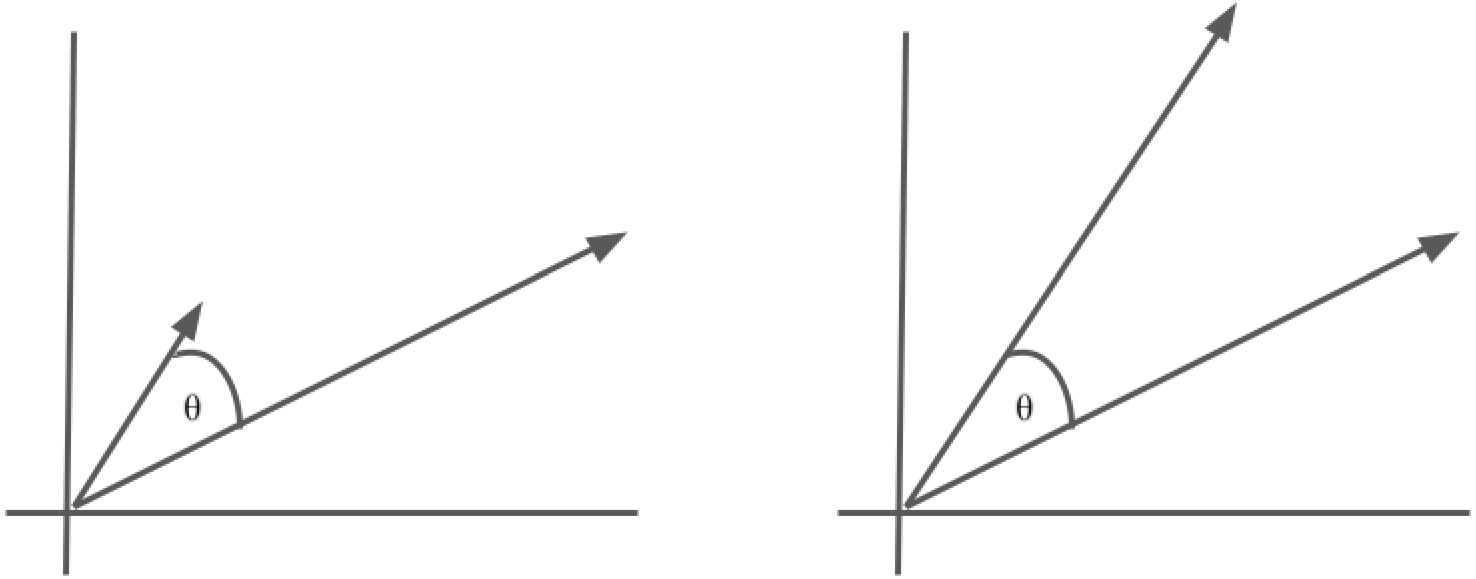
Kolmogorov-Smirnov Distance

Kuiper Distance

Wasserstein Distance

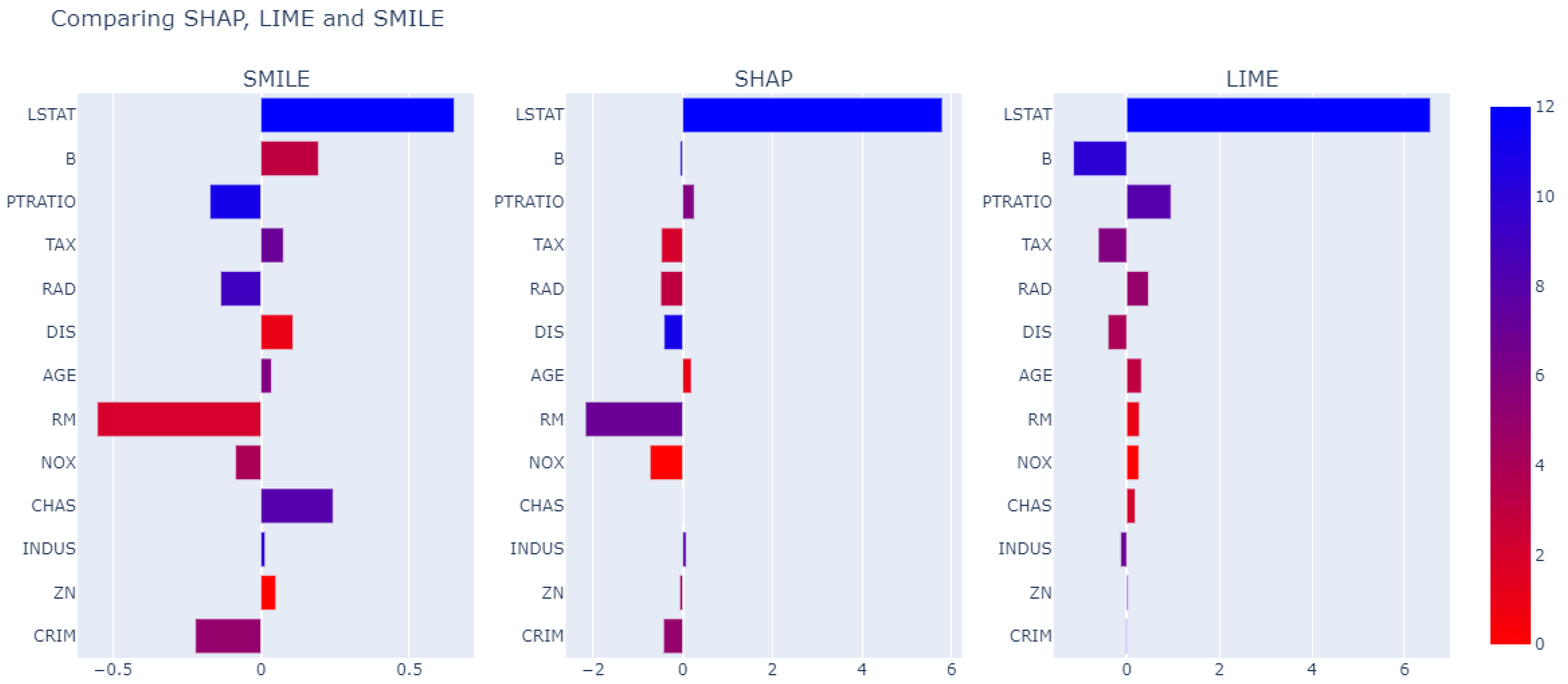
Cramer-Von Mises Distance

Wasserstein vs. Cosine



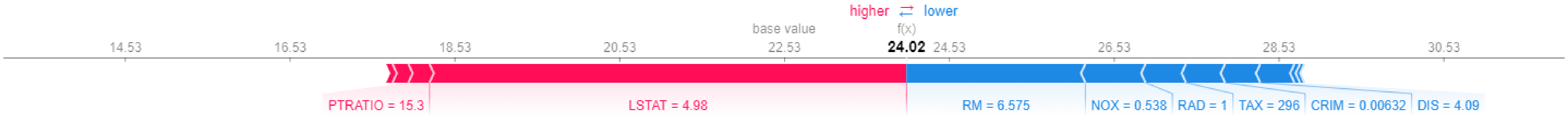
Source : Horan, C. (2021) **Wasserstein Distance and Textual Similarity,** https://neptune.ai/blog/wasserstein-distance-and-textual-similarity

Comparing SMILE with LIME and SHAP for Boston House Pricing Prediction

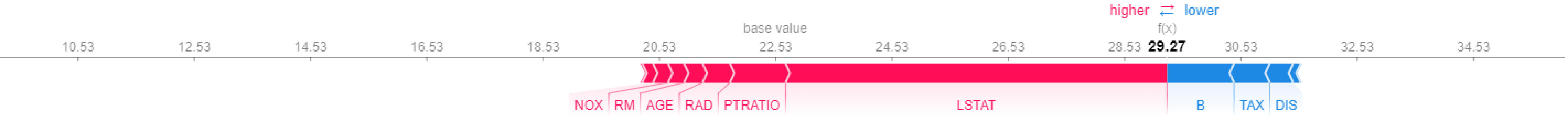


Comparing SMILE with LIME and SHAP with Force Plot

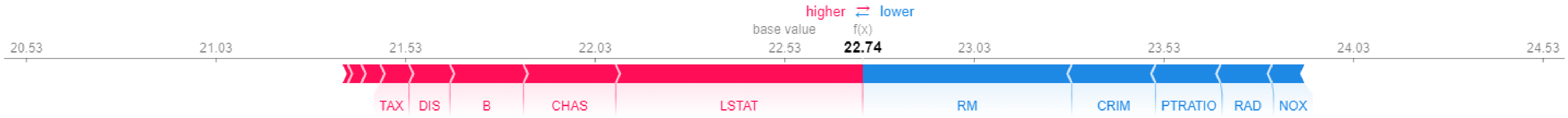
SHAP



LIME

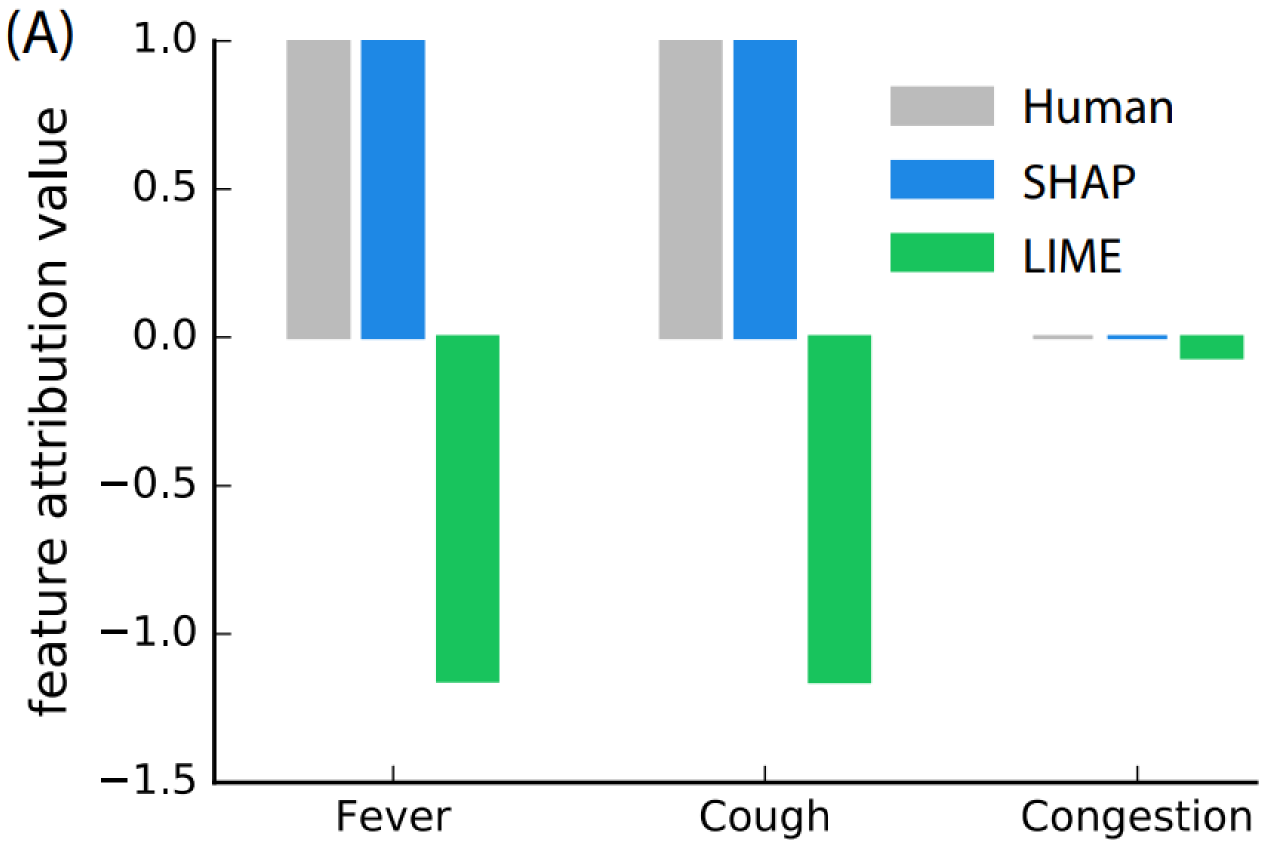


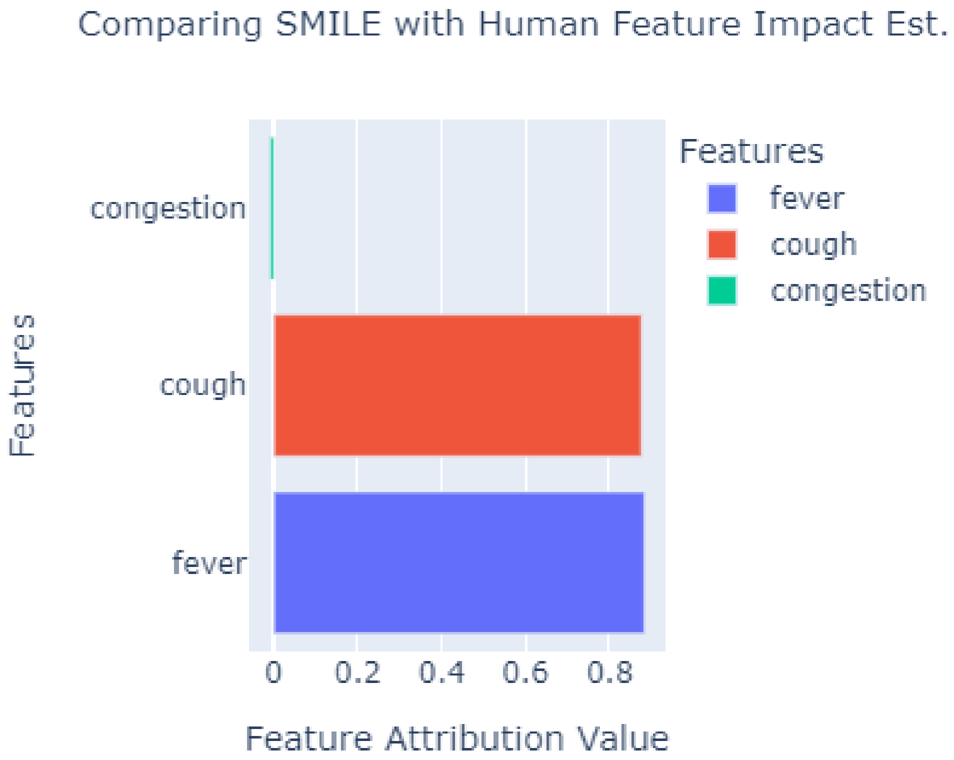
SMILE



Comparing with Human Feature Impact Estimates

Let's discuss a simple example of LIME on a tabular hypothetical dataset





Lundberg, S. M., & Lee, S. I. (2017, December). A unified approach to interpreting model predictions. In *Proceedings of the 31st international conference on neural information processing systems* (pp. 4768-4777).

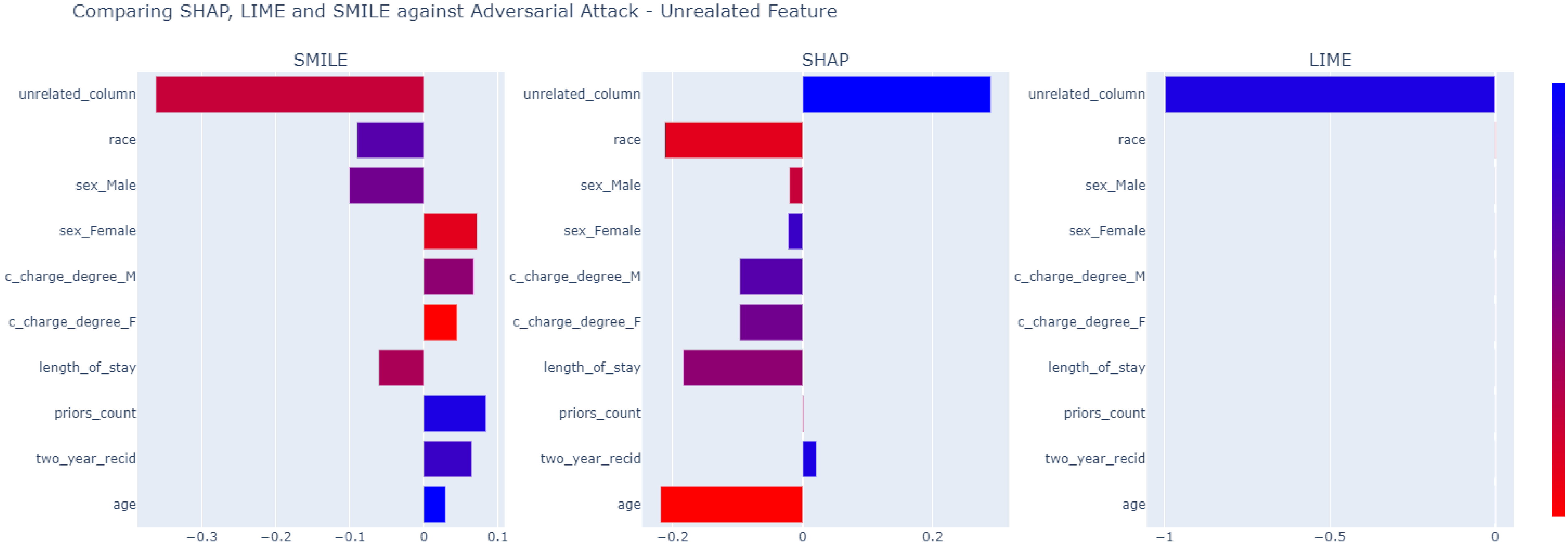
Adversarial Attacks

Fooling LIME and SHAP: Adversarial Attacks on Post hoc Explanation Methods by Dylan Slack et al. (2020)

GIF from: storyset.com

Adversarial Attacks

Fooling LIME and SHAP: Adversarial Attacks on Post-hoc Explanation Methods



* COMPAS dataset for recidivism risk prediction

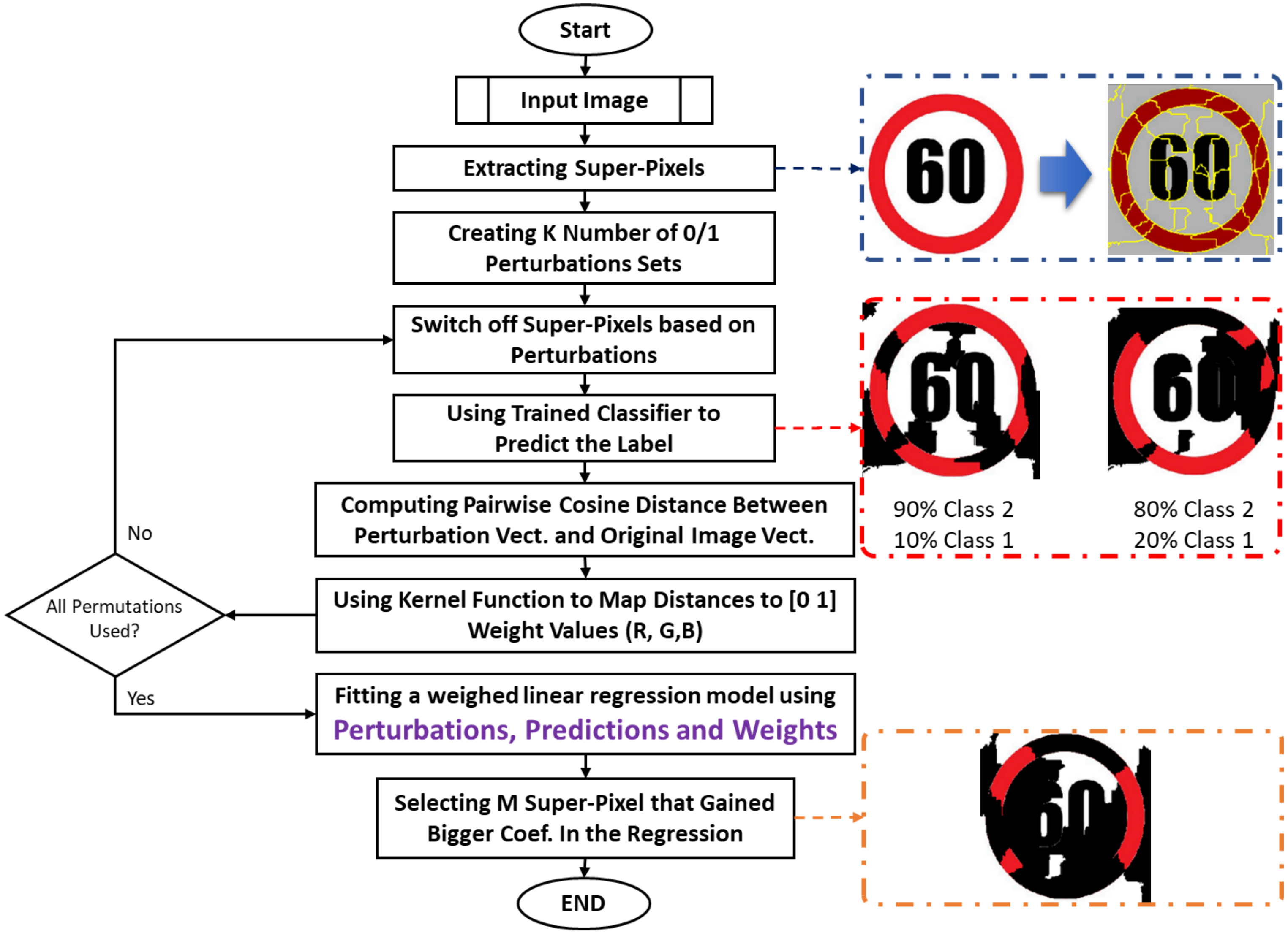
SMILE for Images

Explaining the SMILE Procedure for local explainability in Image classifiers.

GIF from: storyset.com

LIME for Image Classifiers

How LIME provides explainability for image classifiers?



Start

Input Image

Extracting Super-Pixels

Creating K Number of 0/1 Perturbations Sets

Switch off Super-Pixels based on Perturbations

Using Trained Classifier to Predict the Label

Computing Pairwise Cosine Distance Between Perturbation Vect. and Original Image Vect.

Using Kernel Function to Map Distances to [0 1] Weight Values (R, G,B)

All Permutations Used?

No

Yes

Fitting a weighed linear regression model using **Perturbations, Predictions and Weights**

Selecting M Super-Pixel that Gained Bigger Coef. In the Regression

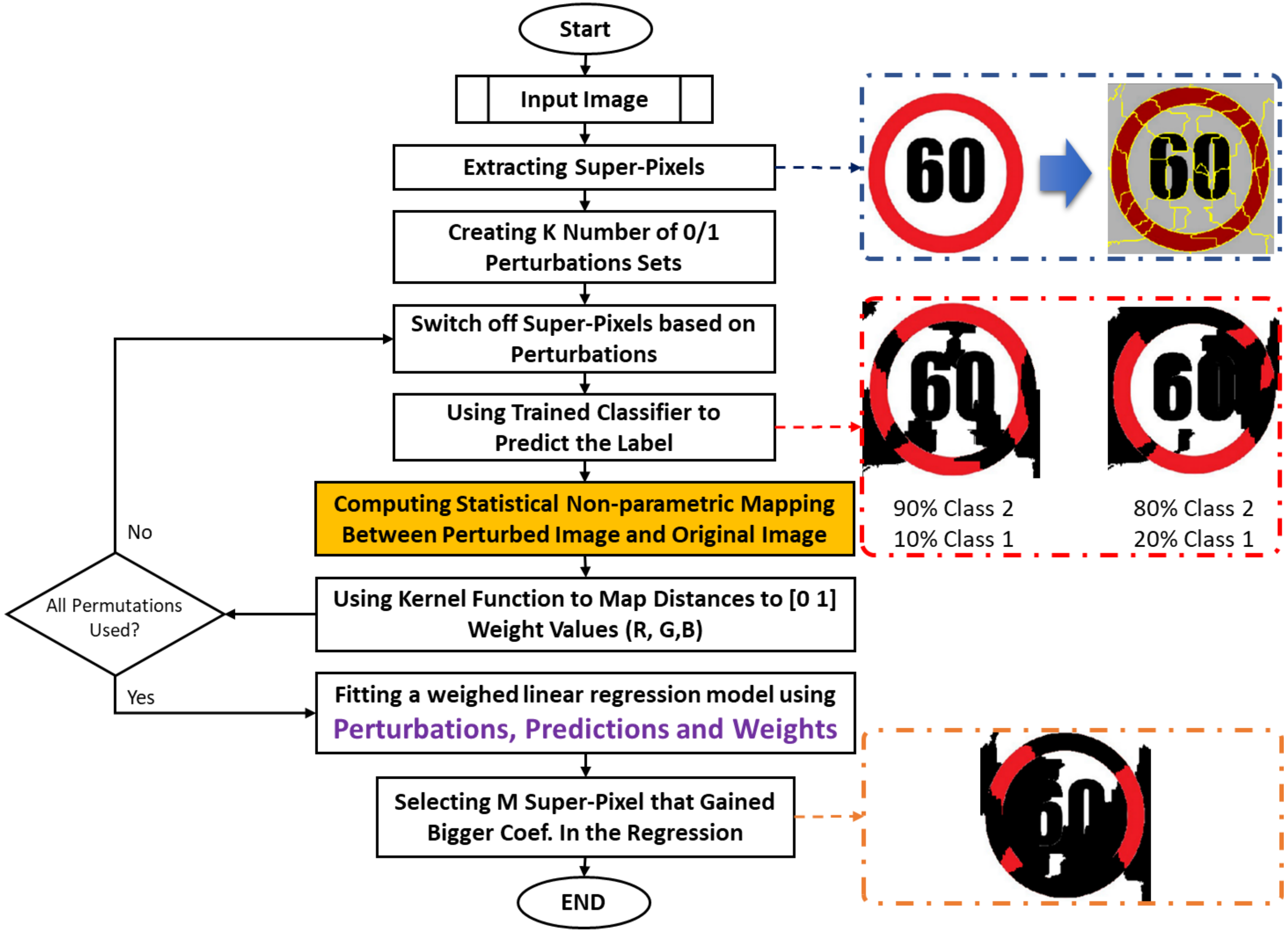
END

90% Class 2
10% Class 1

80% Class 2
20% Class 1

SMILE for Image Classifiers

How SMILE provides explainability for image classifiers?



Start

Input Image

Extracting Super-Pixels

Creating K Number of 0/1 Perturbations Sets

Switch off Super-Pixels based on Perturbations

Using Trained Classifier to Predict the Label

Computing Statistical Non-parametric Mapping Between Perturbed Image and Original Image

Using Kernel Function to Map Distances to [0 1] Weight Values (R, G,B)

All Permutations Used?

No

Yes

Fitting a weighed linear regression model using Perturbations, Predictions and Weights

Selecting M Super-Pixel that Gained Bigger Coef. In the Regression

END

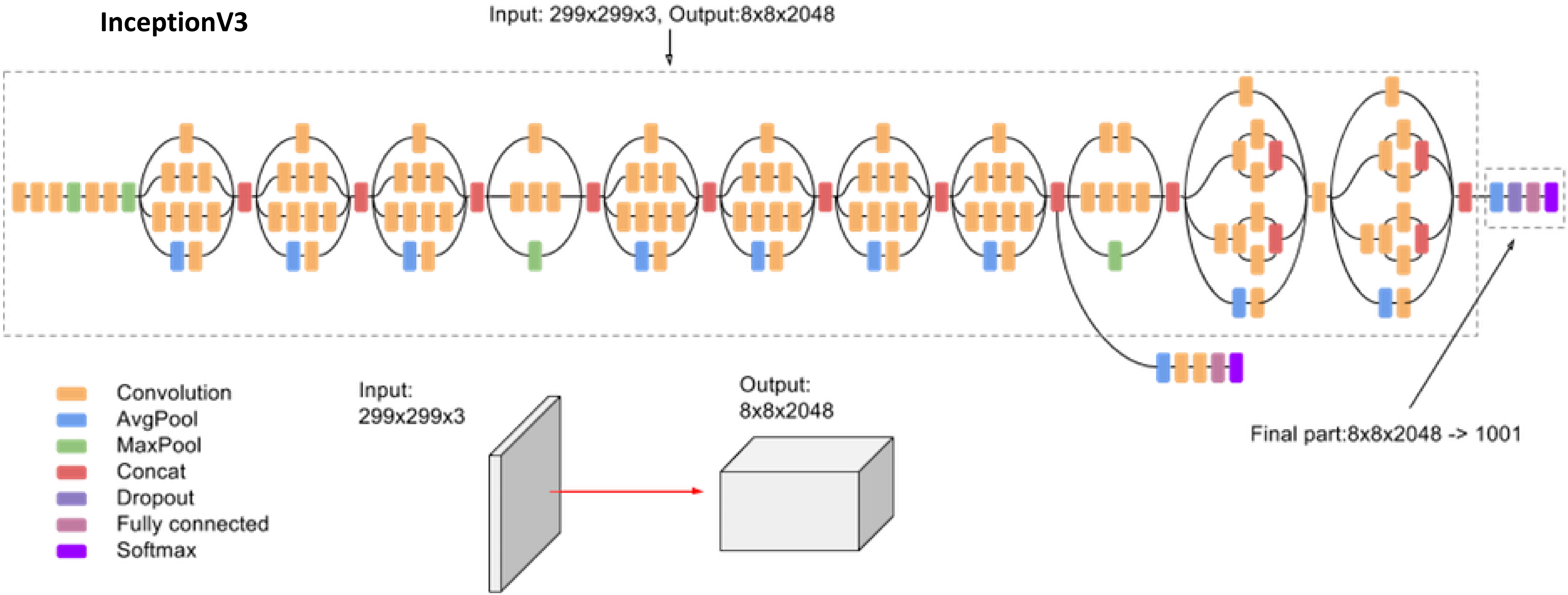
90% Class 2
10% Class 1

80% Class 2
20% Class 1

SMILE for Image Classifiers

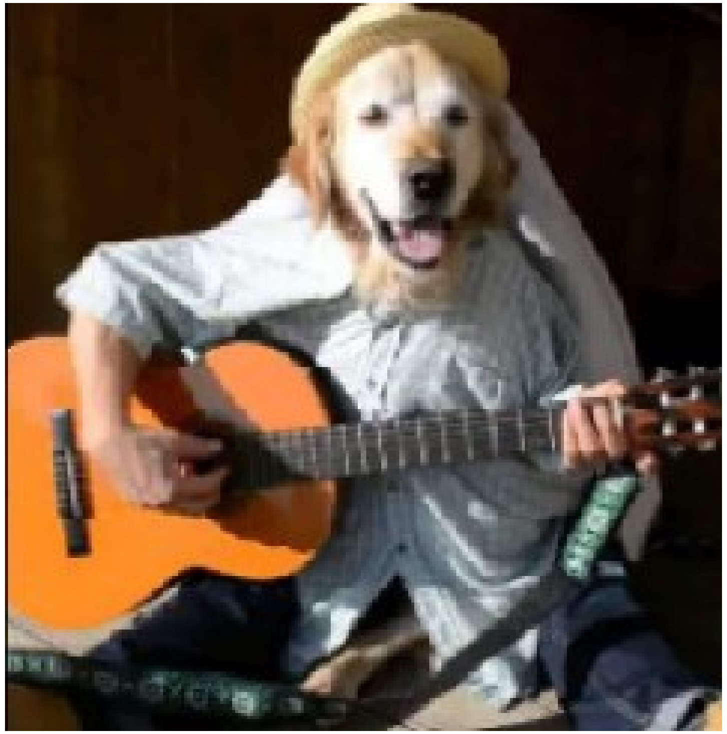
How SMILE provides explainability for image classifiers?

InceptionV3

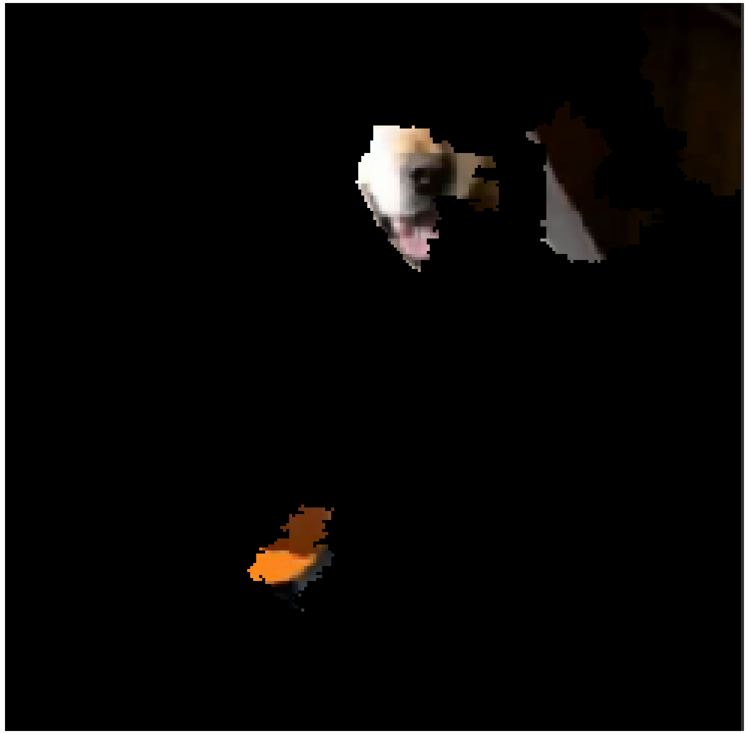


Comparison between LIME and SMILE for Image Classification

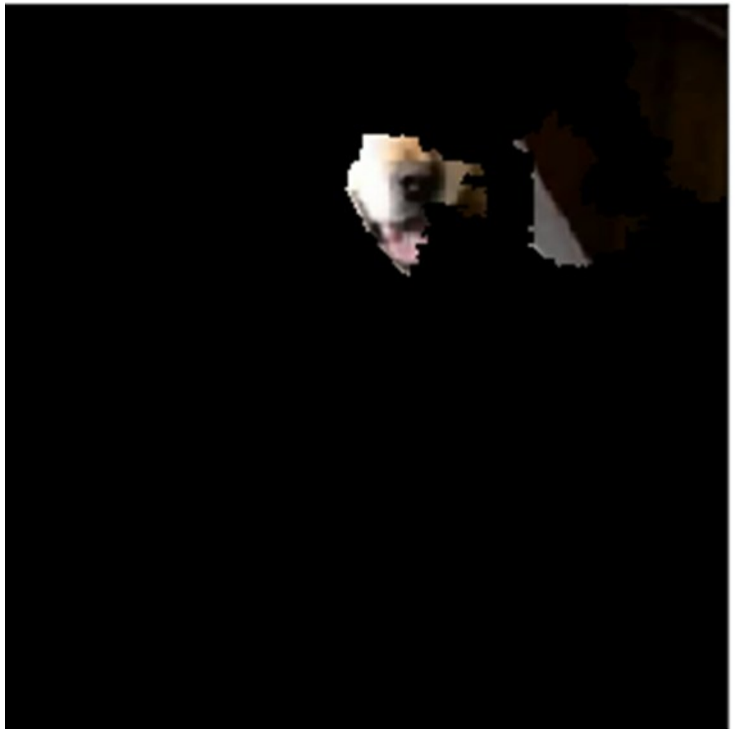
Original



LIME



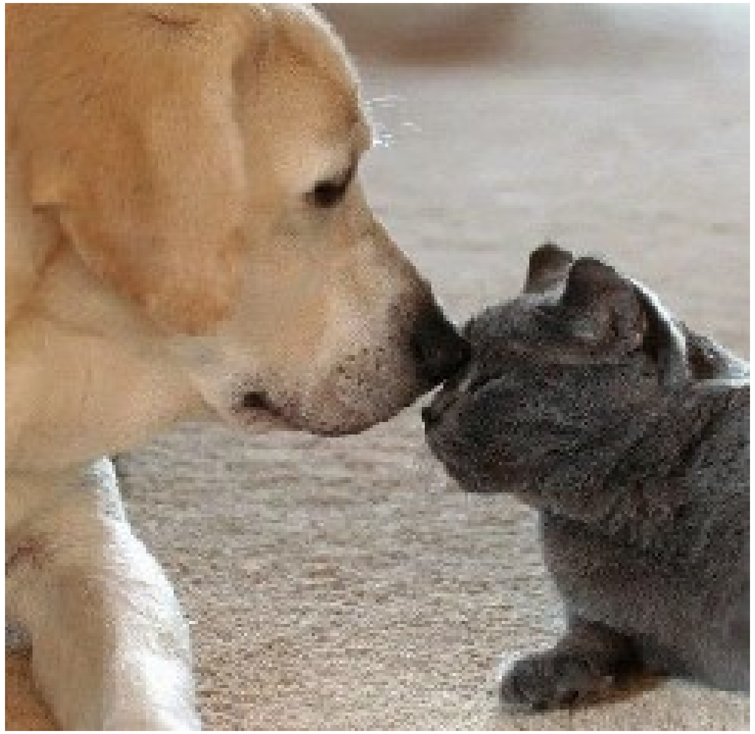
SMILE



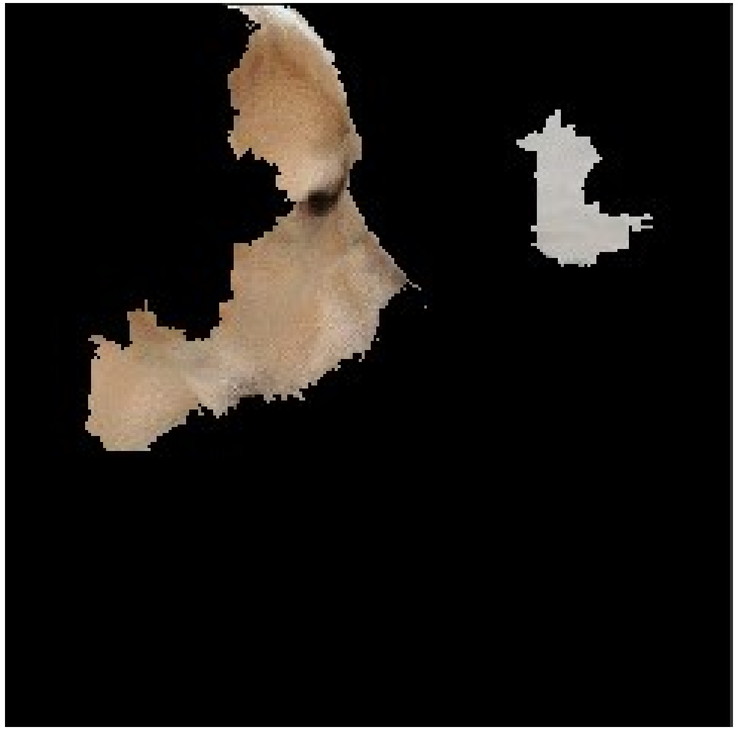
[1] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August). " Why should i trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 1135-1144

Comparison between LIME and SMILE for Image Classification

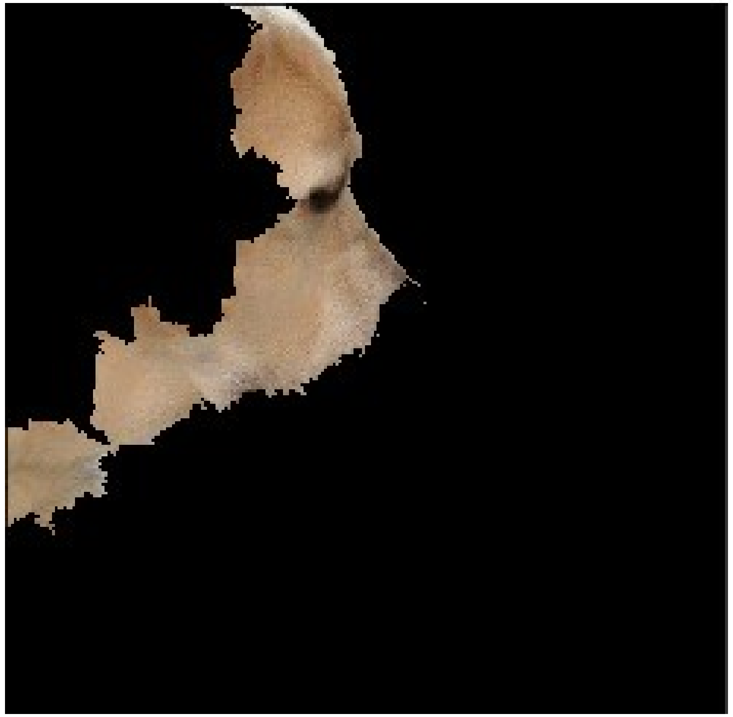
Original



LIME



SMILE



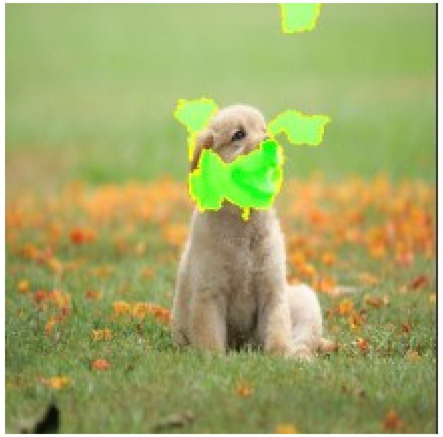
In SMILE, as we process more information (images instead of perturbation vectors), we expect more accurate results.

Source: https://arteagac.github.io/blog

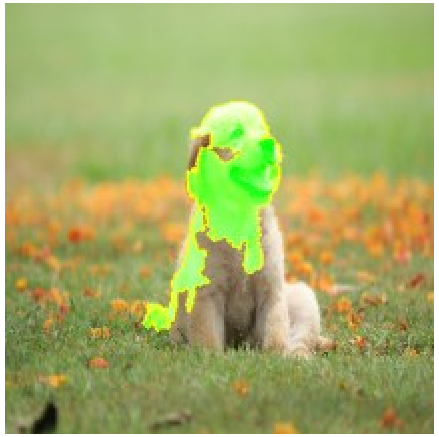
Comparison between BayLIME and SMILE for Image Classification

Original

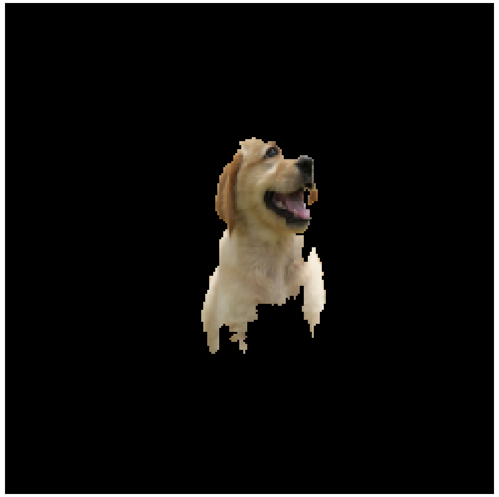
BayLIME – “non_Bay”



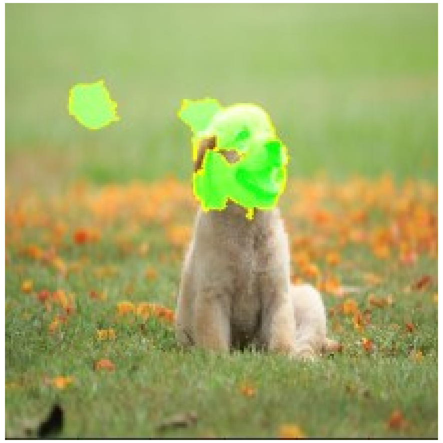
BayLIME – “Bay_non_info_prior”



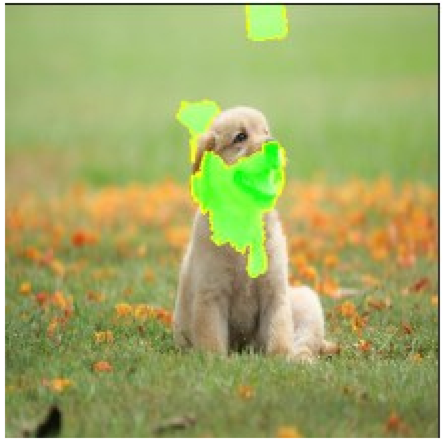
SMILE



BayLIME – “Bay_info_prior”

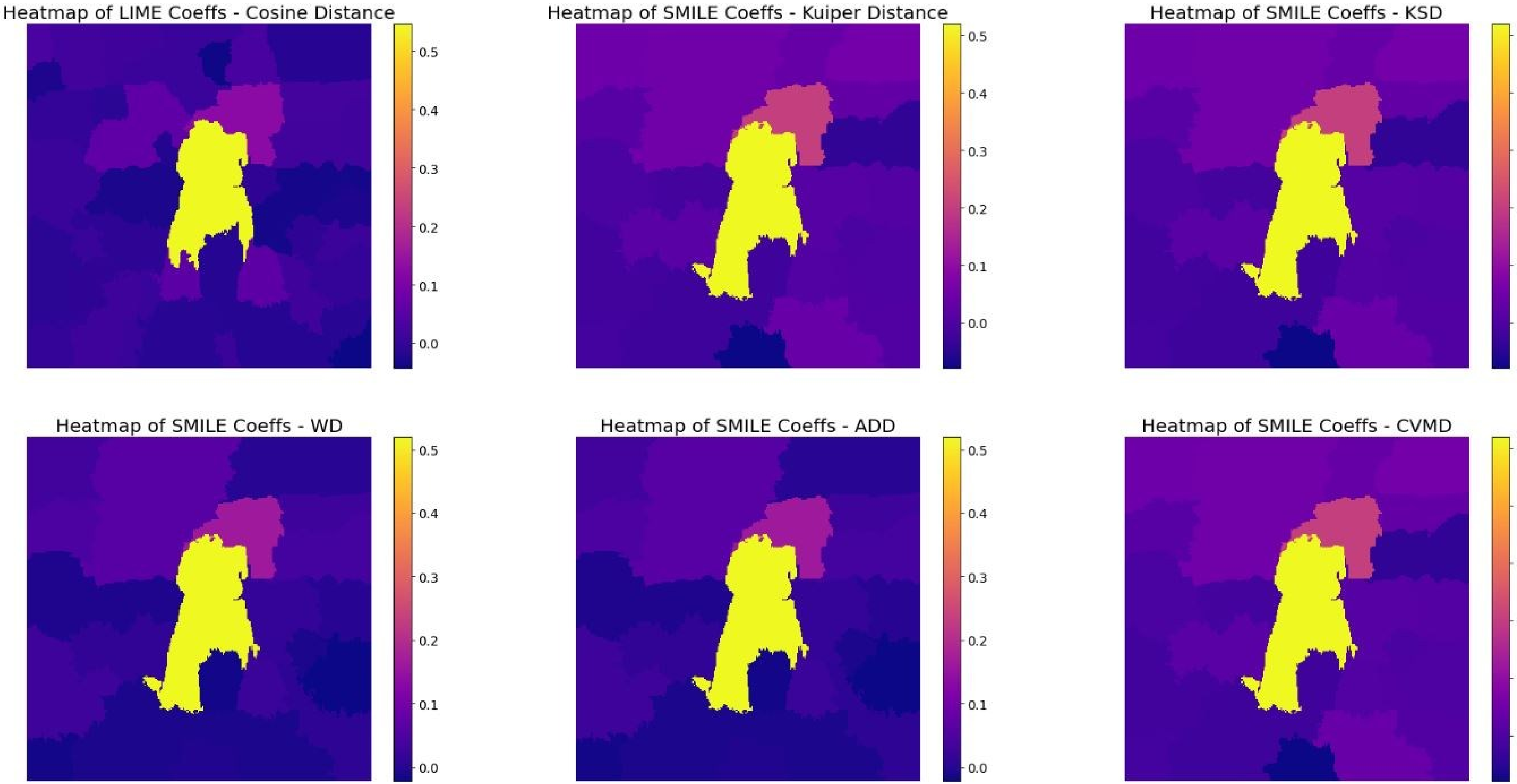


BayLIME – “BayesianRidge_inf_prior_fit_alpha”



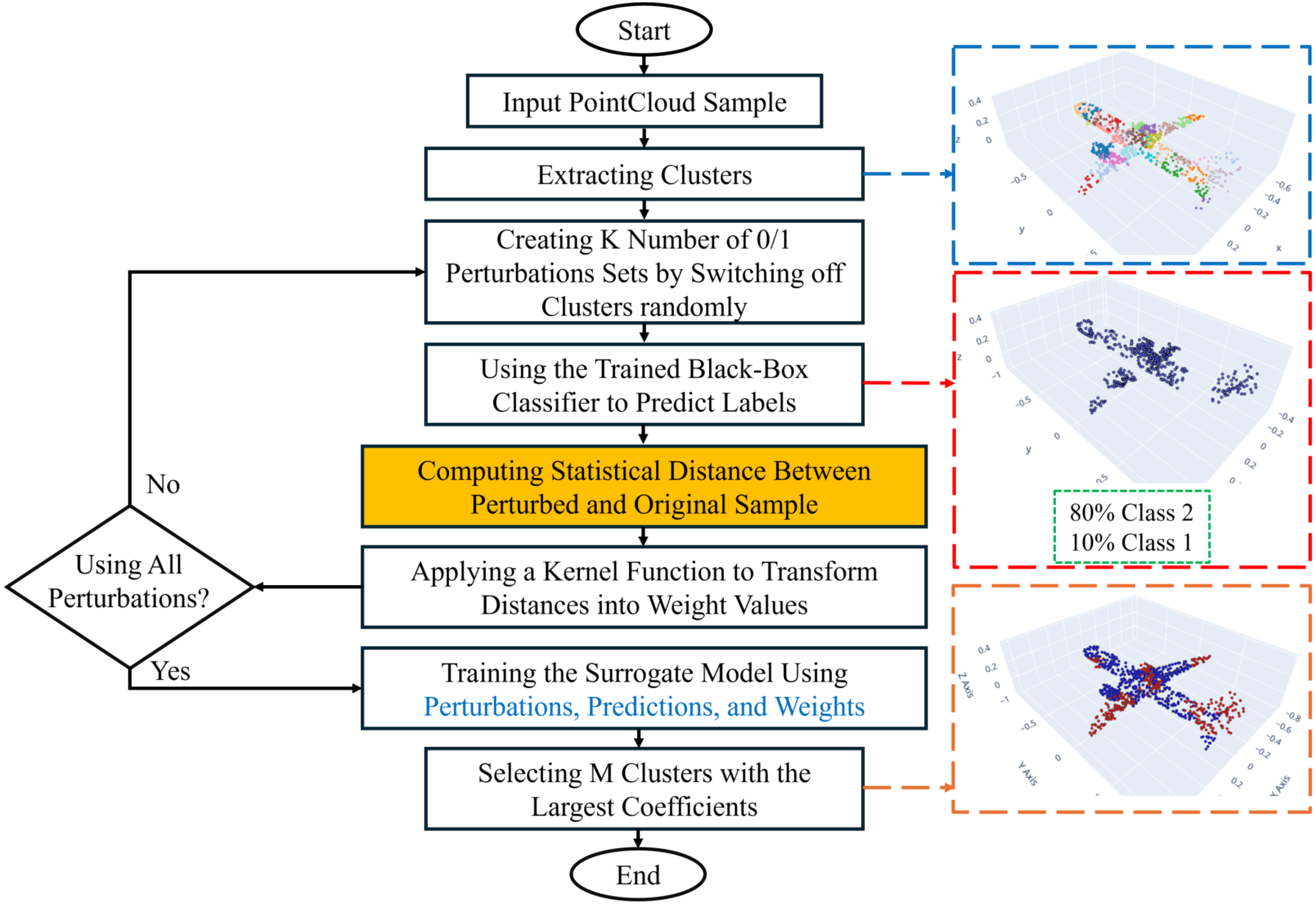
Source: https://github.com/x-y-zhao/BayLime/blob/master/BayLIME_tutorial_images.ipynb

Heatmap Comparison between SMILE Distance Measure for Image Classification



SMILE for Point Cloud

Explaining the SMILE Procedure for local explainability in Point Cloud classifiers.



Start

Input PointCloud Sample

Extracting Clusters

Creating K Number of 0/1 Perturbations Sets by Switching off Clusters randomly

Using the Trained Black-Box Classifier to Predict Labels

Computing Statistical Distance Between Perturbed and Original Sample

Applying a Kernel Function to Transform Distances into Weight Values

Using All Perturbations?

No

Yes

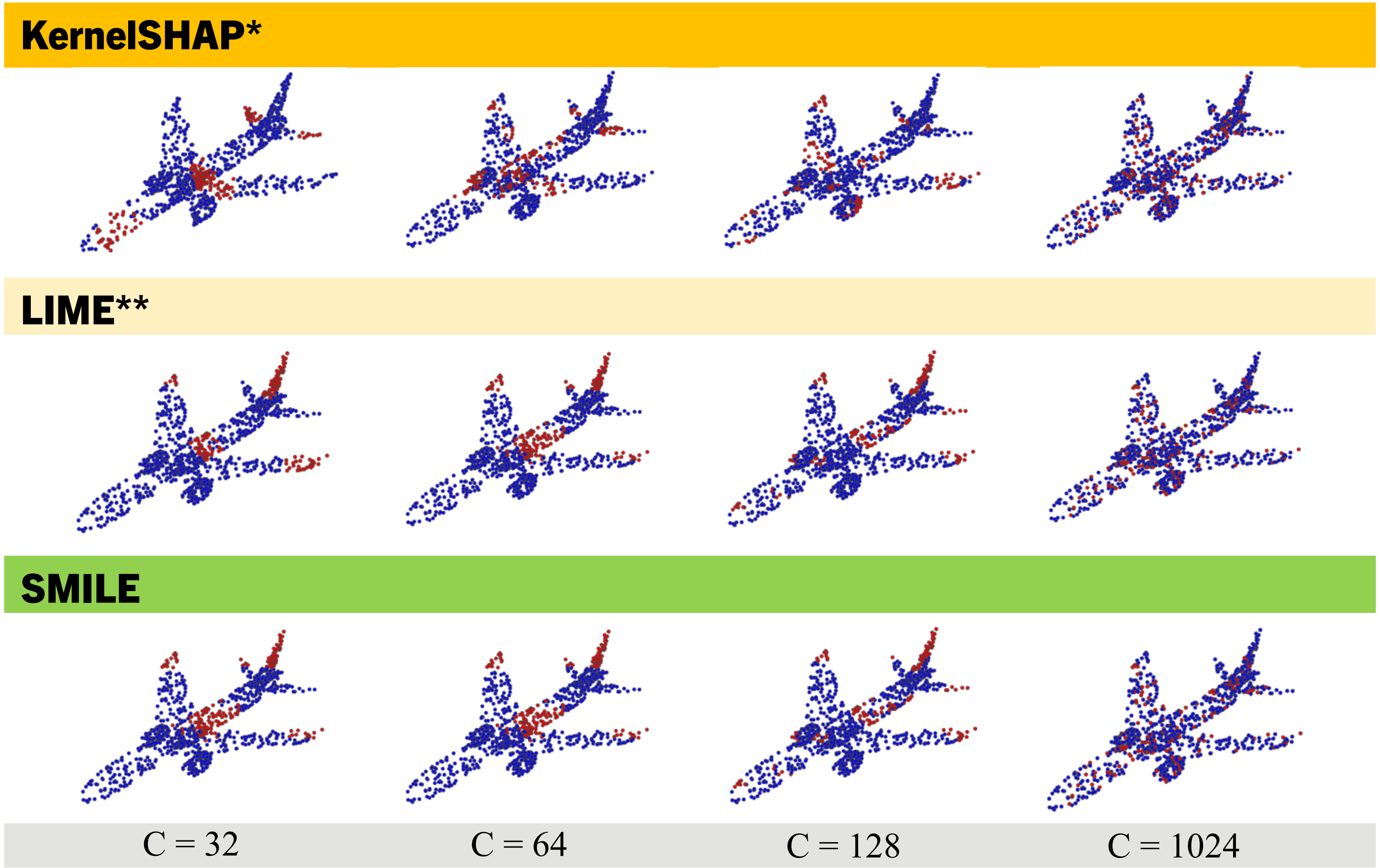
Training the Surrogate Model Using Perturbations, Predictions, and Weights

Selecting M Clusters with the Largest Coefficients

End

80% Class 2
10% Class 1

Ahmadi, S. M., Aslansefat, K., Valcarce-Dineiro, R., & Barnfather, J. (2024). Explainability of Point Cloud Neural Networks Using SMILE: Statistical Model-Agnostic Interpretability with Local Explanations. *arXiv preprint arXiv:2410.15374*.



KernelSHAP*

LIME**

SMILE

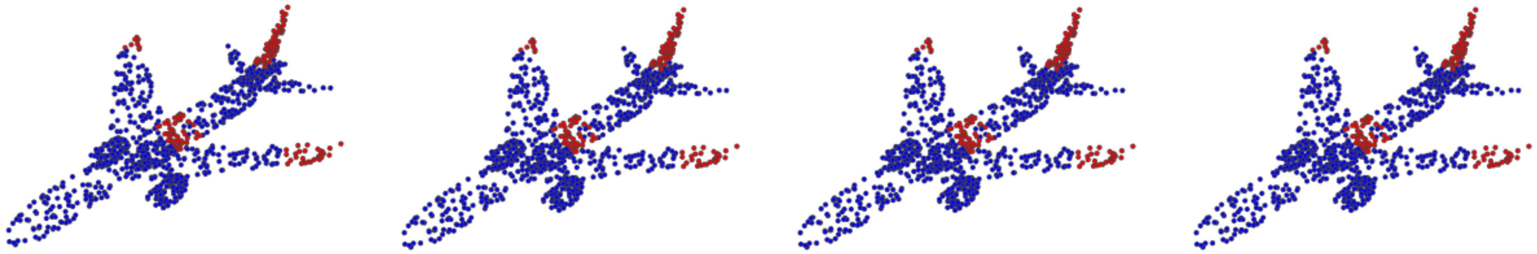
C = 32 C = 64 C = 128 C = 1024

References:

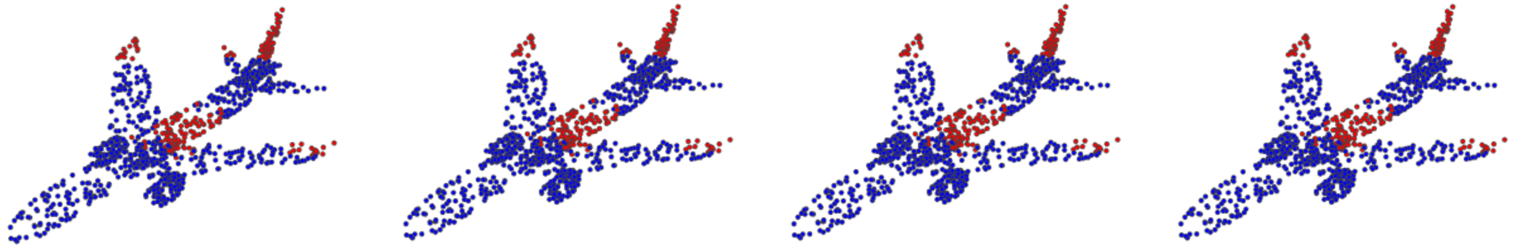
* Inspired by Lundberg, S.M. and S.-I. Lee, *A unified approach to interpreting model predictions.* Advances in neural information processing systems, 2017. **30**.

** Tan, H. and H. Kotthaus. *Surrogate model-based explainability methods for point cloud nns*. in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2022.

LIME*



SMILE (all versions)



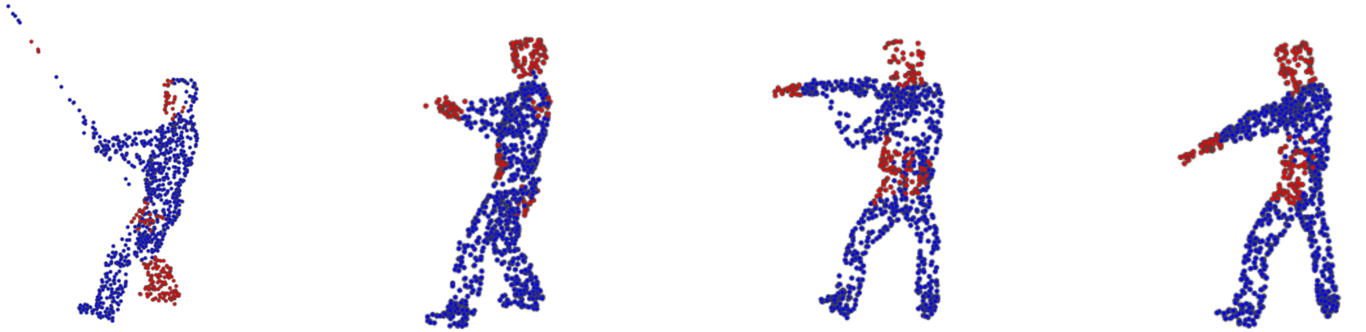
$\sigma = 0.1$ $\sigma = 0.2$ $\sigma = 0.3$ $\sigma = 0.7$

References:

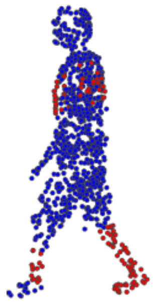
[*] Tan, H. and H. Kotthaus. *Surrogate model-based explainability methods for point cloud nns*. in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2022.

ModelNet40:

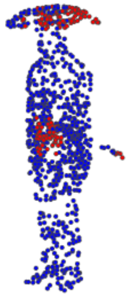
Correctly Classified:



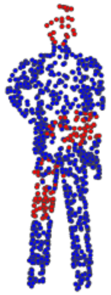
Misclassified:



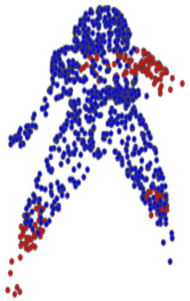
Predicted as
‘Lamp’



Predicted as
‘Plant’



Predicted as
‘Bottle’



Predicted as
‘Lamp’

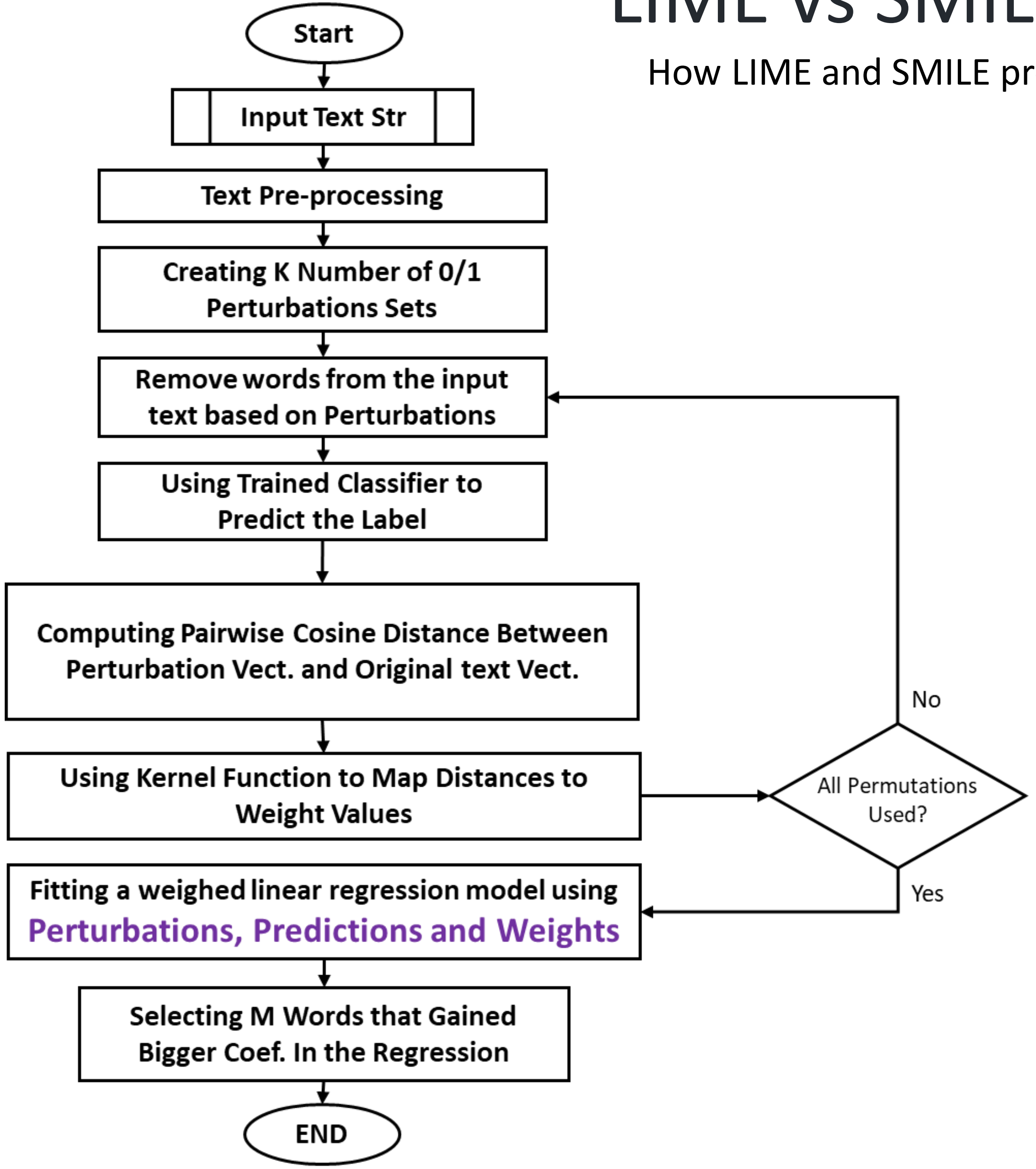
SMILE for Texts

Explaining the SMILE Procedure for local explainability in Text classifiers.

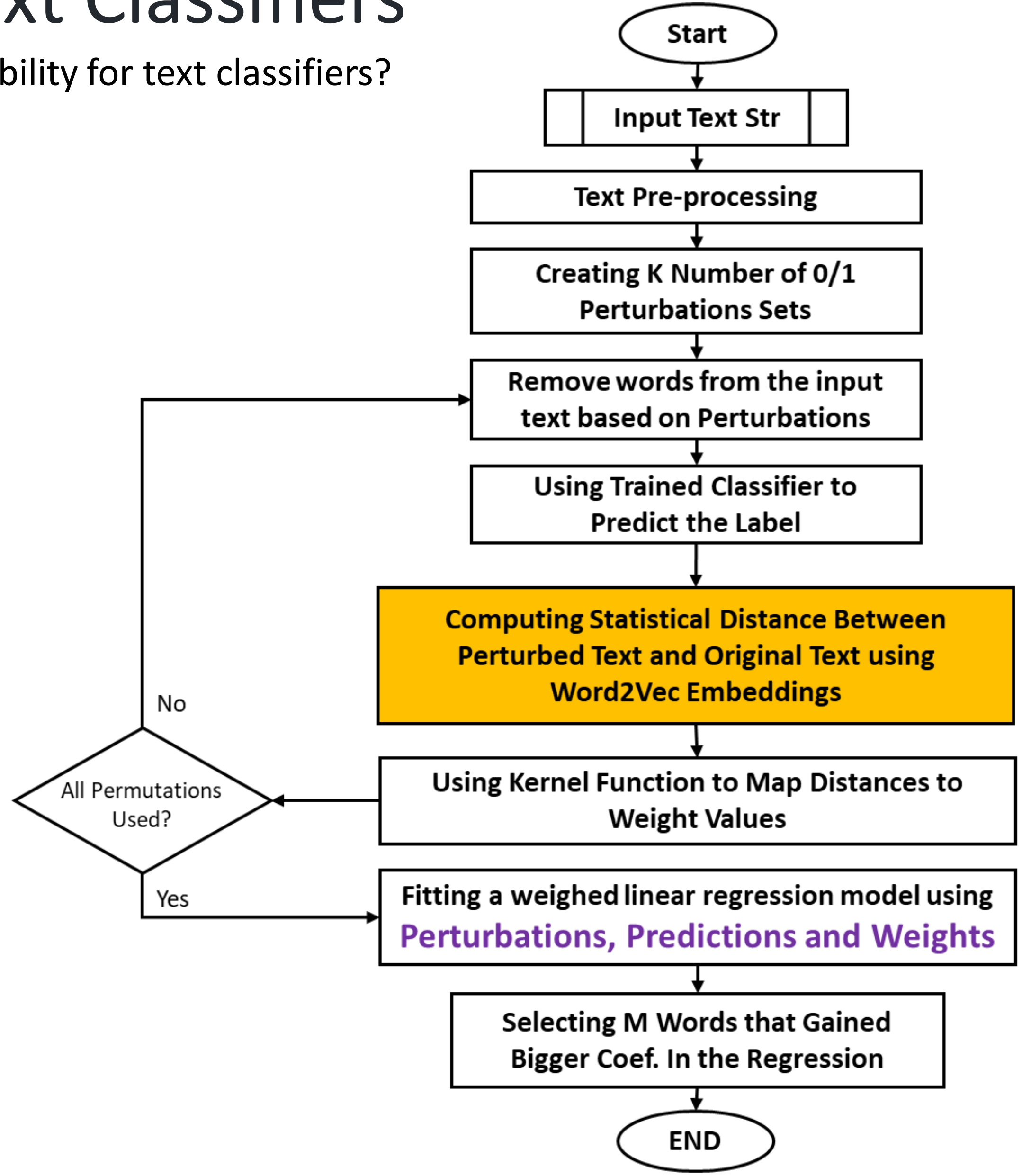
GIF from: storyset.com

LIME vs SMILE for Text Classifiers

How LIME and SMILE provide explainability for text classifiers?



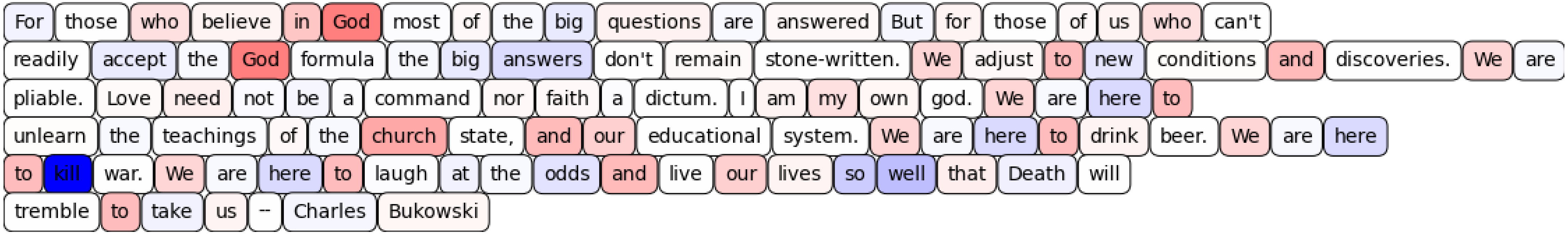
Start
Input Text Str
Text Pre-processing
Creating K Number of 0/1 Perturbations Sets
Remove words from the input text based on Perturbations
Using Trained Classifier to Predict the Label
Computing Pairwise Cosine Distance Between Perturbation Vect. and Original text Vect.
Using Kernel Function to Map Distances to Weight Values
All Permutations Used?
No
Yes
Fitting a weighed linear regression model using Perturbations, Predictions and Weights
Selecting M Words that Gained Bigger Coef. In the Regression
END



Start
Input Text Str
Text Pre-processing
Creating K Number of 0/1 Perturbations Sets
Remove words from the input text based on Perturbations
Using Trained Classifier to Predict the Label
Computing Statistical Distance Between Perturbed Text and Original Text using Word2Vec Embeddings
Using Kernel Function to Map Distances to Weight Values
All Permutations Used?
No
Yes
Fitting a weighed linear regression model using Perturbations, Predictions and Weights
Selecting M Words that Gained Bigger Coef. In the Regression
END

SMILE for Text Classifiers - Visualization

“For those who believe in God most of the big questions are answered But for those of us who can't readily accept the God formula the big answers don't remain stone-written. We adjust to new conditions and discoveries. We are pliable. Love need not be a command nor faith a dictum. I am my own god. We are here to unlearn the teachings of the church state, and our educational system. We are here to drink beer. We are here to kill war. We are here to laugh at the odds and live our lives so well that Death will tremble to take us - Charles Bukowski"

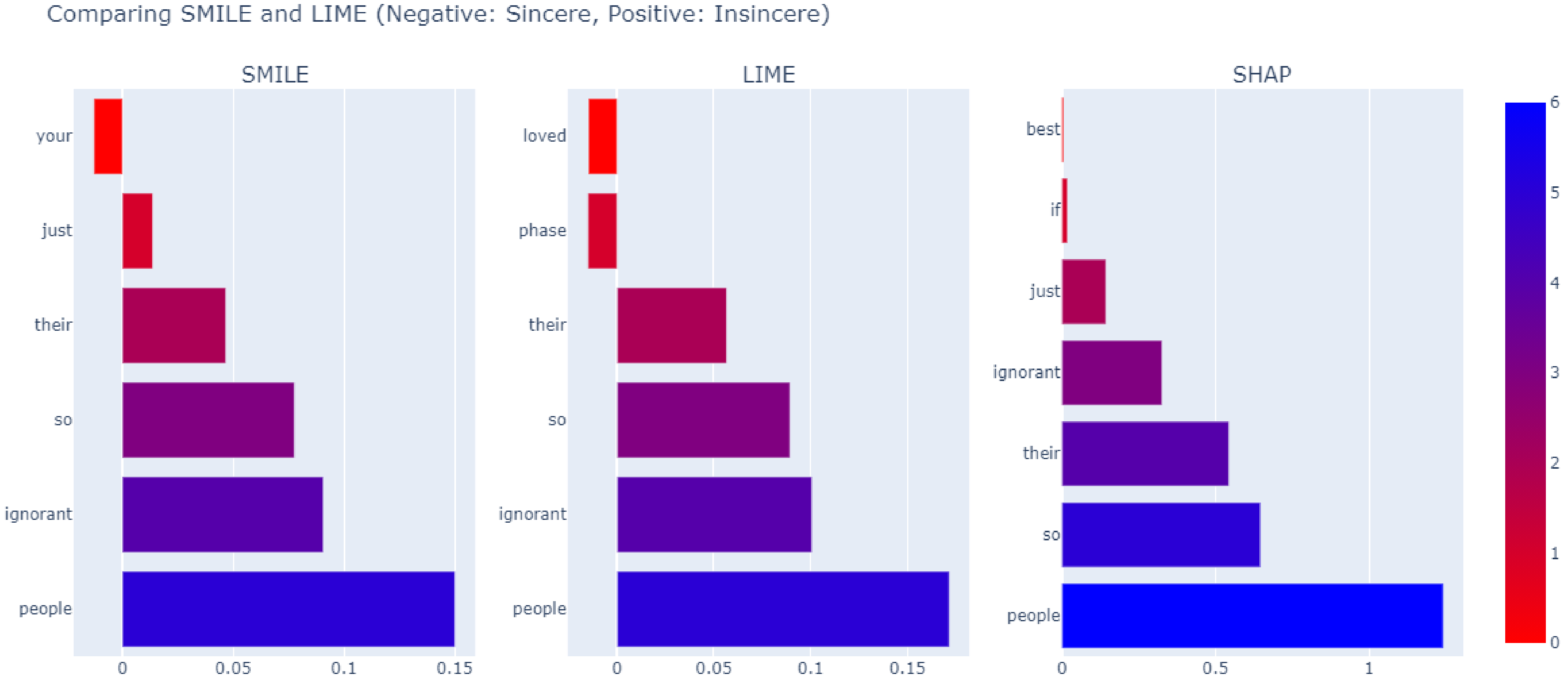


For those who believe in God most of the big questions are answered But for those of us who can't
readily accept the God formula the big answers don't remain stone-written. We adjust to new conditions and discoveries. We are
pliable. Love need not be a command nor faith a dictum. I am my own god. We are here to
unlearn the teachings of the church state, and our educational system. We are here to drink beer. We are here
to kill war. We are here to laugh at the odds and live our lives so well that Death will
tremble to take us -- Charles Bukowski

SMILE vs LIME and SHAP for Text Classifiers:

Quora Insincere Questions Classification

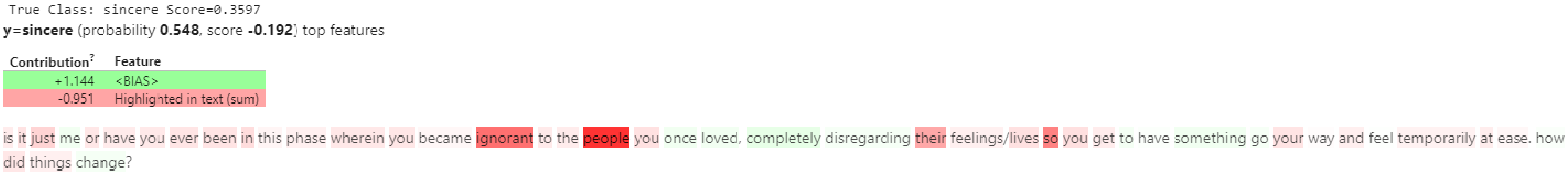
“Is it just me or have you ever been in this phase wherein you became ignorant to the people you once loved, completely disregarding their feelings/lives so you get to have something go your way and feel temporarily at ease. How did things change?" `True Class: sincere Score=0.3597`



SMILE vs ELi5 Visualisation for Text Classifiers:

Quora Insincere Questions Classification

“Is it just me or have you ever been in this phase wherein you became ignorant to the people you once loved, completely disregarding their feelings/lives so you get to have something go your way and feel temporarily at ease. How did things change?"

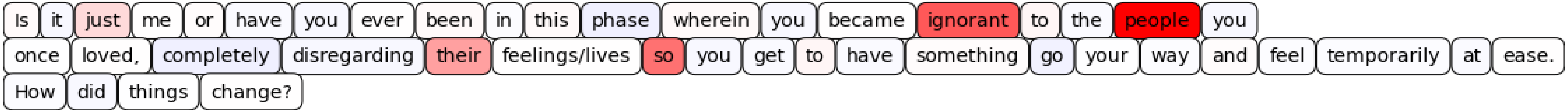


```
True Class: sincere Score=0.3597
```

y=sincere (probability **0.548**, score **-0.192**) top features

Contribution?	Feature
+1.144	<BIAS>
-0.951	Highlighted in text (sum)

is it just me or have you ever been in this phase wherein you became ignorant to the people you once loved, completely disregarding their feelings/lives so you get to have something go your way and feel temporarily at ease. how did things change?



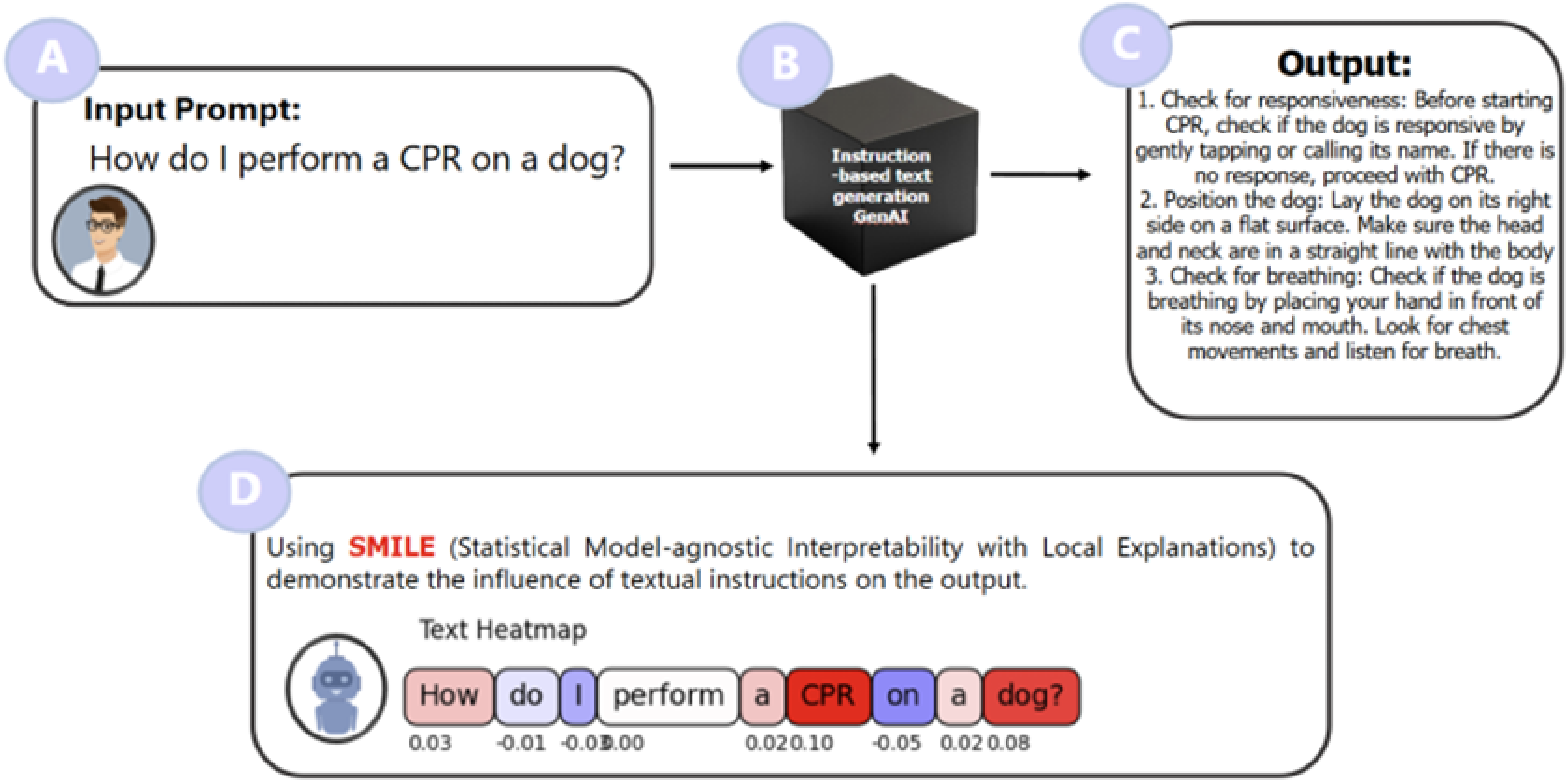
Is it just me or have you ever been in this phase wherein you became ignorant to the people you once loved, completely disregarding their feelings/lives so you get to have something go your way and feel temporarily at ease. How did things change?

SMILE for Generative AI

Explaining the SMILE Procedure for local explainability in Text classifiers.

GIF from: storyset.com

gSMILE for Generative Models (LLMs, VLMs, MLLMs)



Input Prompt:

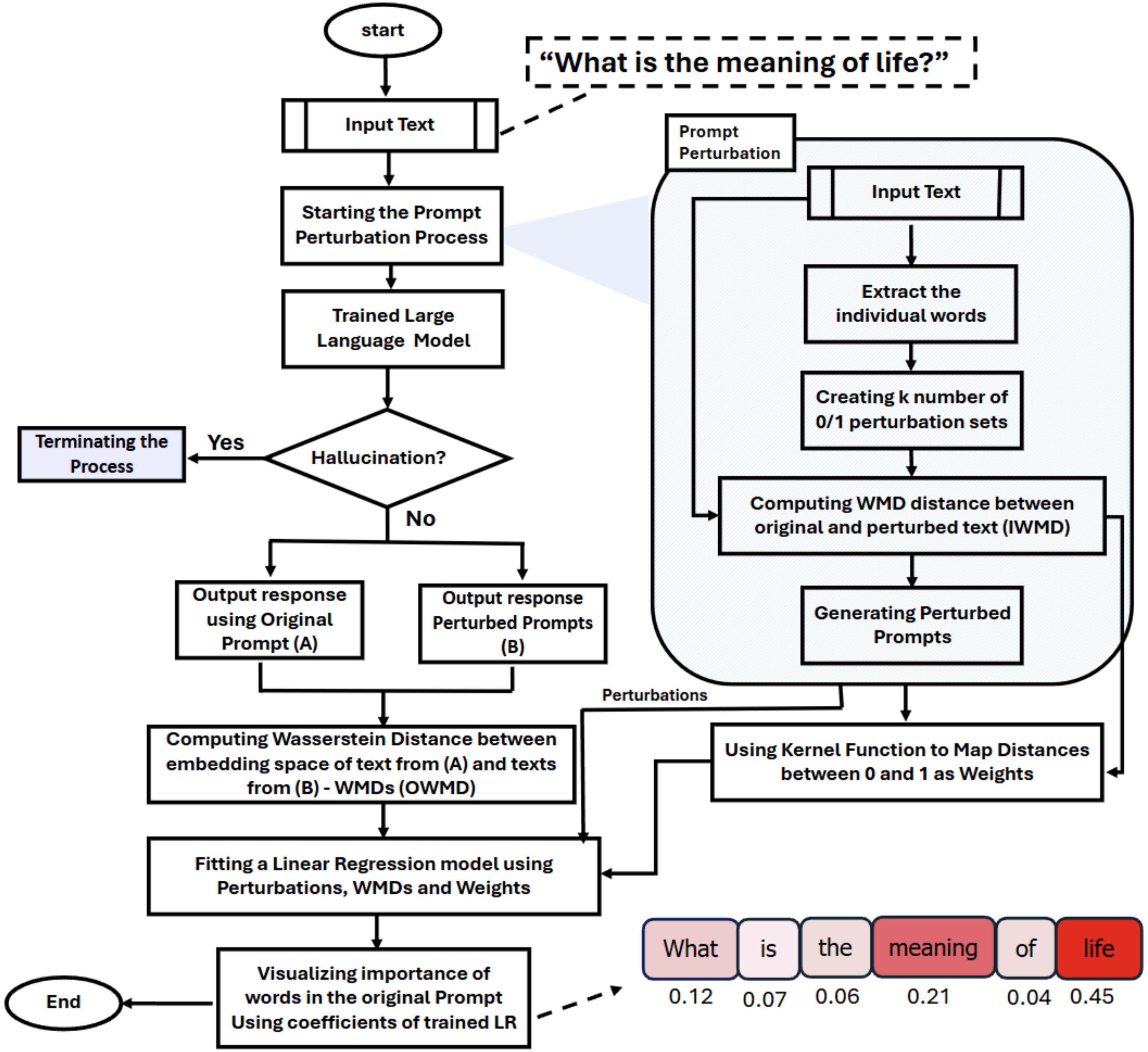
How do I perform a CPR on a dog?

Output:

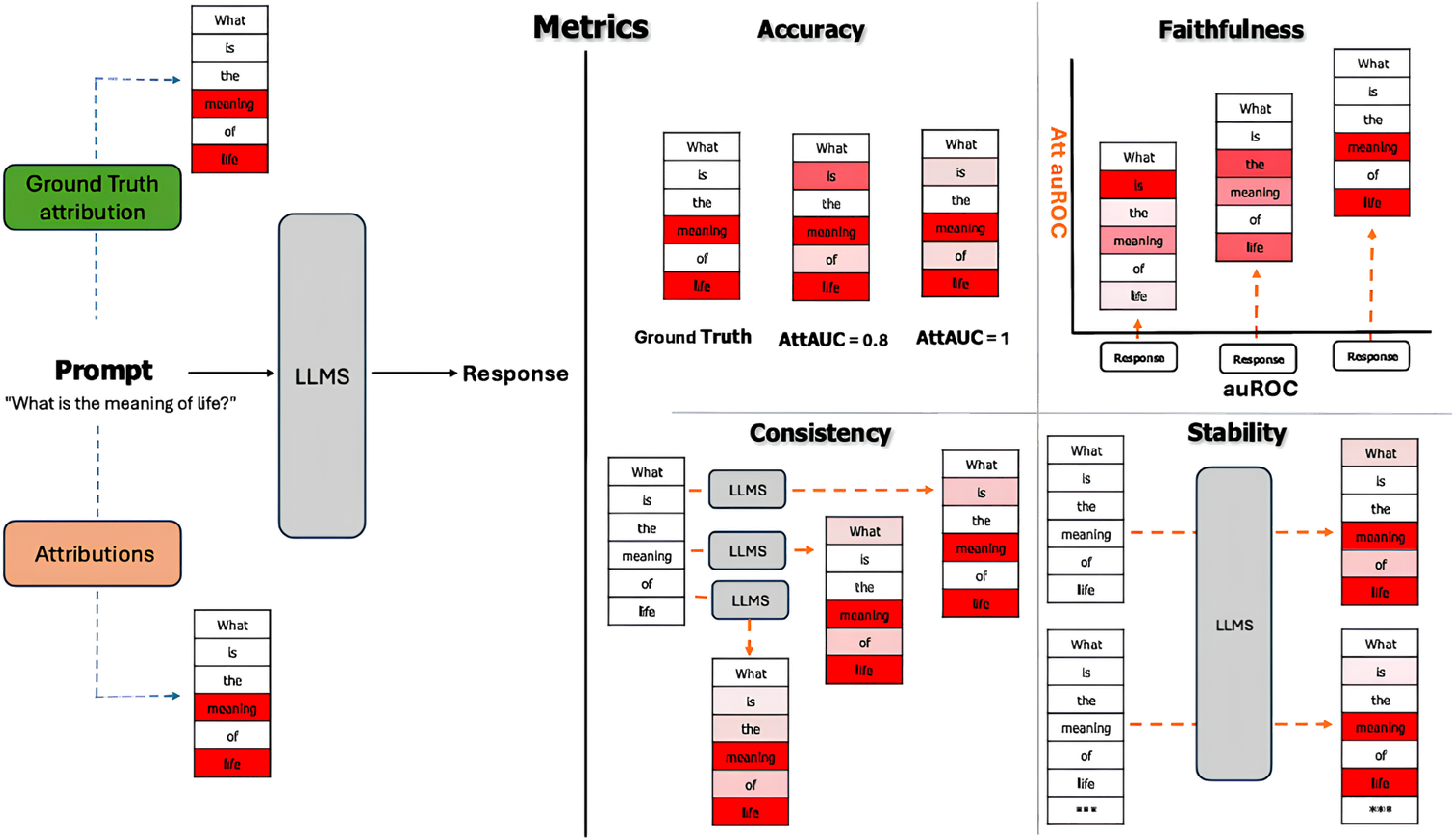
1. Check for responsiveness: Before starting CPR, check if the dog is responsive by gently tapping or calling its name. If there is no response, proceed with CPR.
2. Position the dog: Lay the dog on its right side on a flat surface. Make sure the head and neck are in a straight line with the body
3. Check for breathing: Check if the dog is breathing by placing your hand in front of its nose and mouth. Look for chest movements and listen for breath.

Using **SMILE** (Statistical Model-agnostic Interpretability with Local Explanations) to demonstrate the influence of textual instructions on the output.

Dehghani, Z., Aslansefat, K., Khan, A., & Akram, M. N. (2025). Explainability of Large Language Models using SMILE: Statistical Model-agnostic Interpretability with Local Explanations. *arXiv preprint arXiv:2505.21657*. *(Submitted to IEEE Transactions on Artificial Intelligence)*.

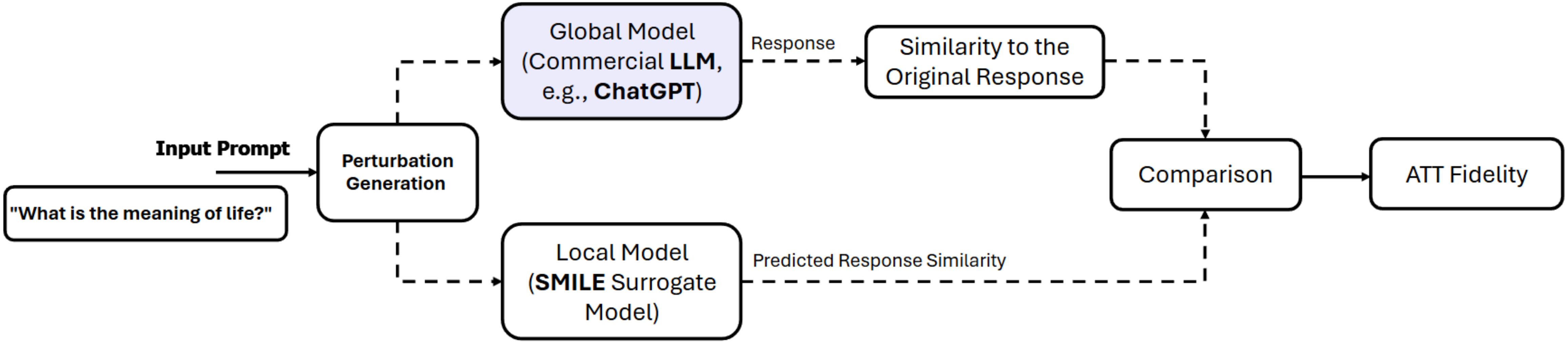

start
Input Text
"What is the meaning of life?"
Starting the Prompt Perturbation Process
Trained Large Language Model
Hallucination?
Yes
Terminating the Process
No
Output response using Original Prompt (A)
Output response Perturbed Prompts (B)
Prompt Perturbation
Input Text
Extract the individual words
Creating k number of 0/1 perturbation sets
Computing WMD distance between original and perturbed text (IWMD)
Generating Perturbed Prompts
Perturbations
Computing Wasserstein Distance between embedding space of text from (A) and texts from (B) - WMDs (OWMD)
Using Kernel Function to Map Distances between 0 and 1 as Weights
Fitting a Linear Regression model using Perturbations, WMDs and Weights
Visualizing importance of words in the original Prompt Using coefficients of trained LR
End
What
is
the
meaning
of
life
0.12
0.07
0.06
0.21
0.04
0.45

Dehghani, Z., Aslansefat, K., Khan, A., & Akram, M. N. (2025). Explainability of Large Language Models using SMILE: Statistical Model-agnostic Interpretability with Local Explanations. *arXiv preprint arXiv:2505.21657. (Submitted to IEEE Transactions on Artificial Intelligence).*



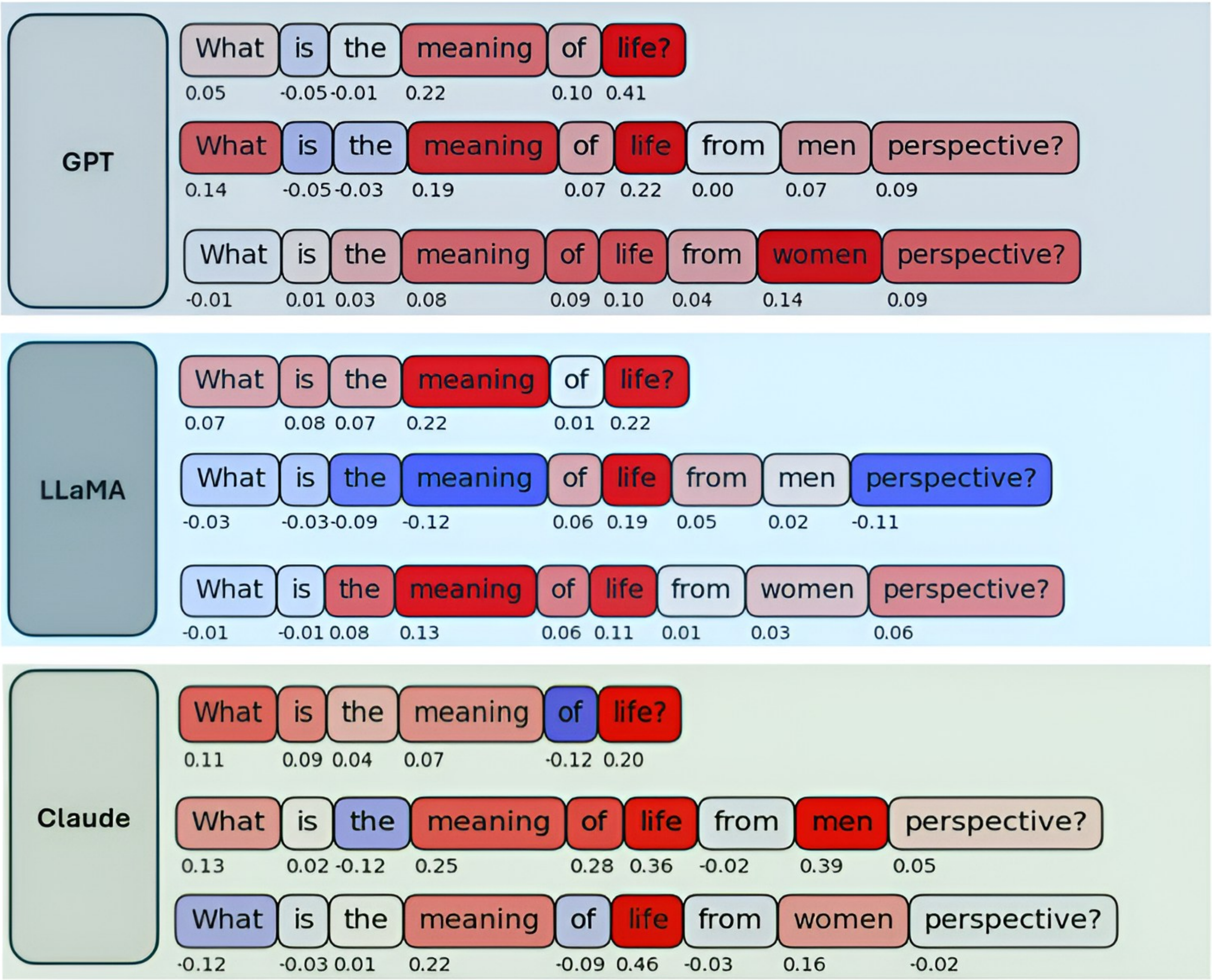
Dehghani, Z., Aslansefat, K., Khan, A., & Akram, M. N. (2025). Explainability of Large Language Models using SMILE: Statistical Model-agnostic Interpretability with Local Explanations. *arXiv preprint arXiv:2505.21657*. *(Submitted to IEEE Transactions on Artificial Intelligence)*.

A Change in Fidelity Assessment for LLMs


Input Prompt
"What is the meaning of life?"
Perturbation Generation
Global Model (Commercial **LLM**, e.g., **ChatGPT**)
Response
Similarity to the Original Response
Local Model (**SMILE** Surrogate Model)
Predicted Response Similarity
Comparison
ATT Fidelity

Dehghani, Z., Aslansefat, K., Khan, A., & Akram, M. N. (2025). Explainability of Large Language Models using SMILE: Statistical Model-agnostic Interpretability with Local Explanations. *arXiv preprint arXiv:2505.21657. (Submitted to IEEE Transactions on Artificial Intelligence)*.

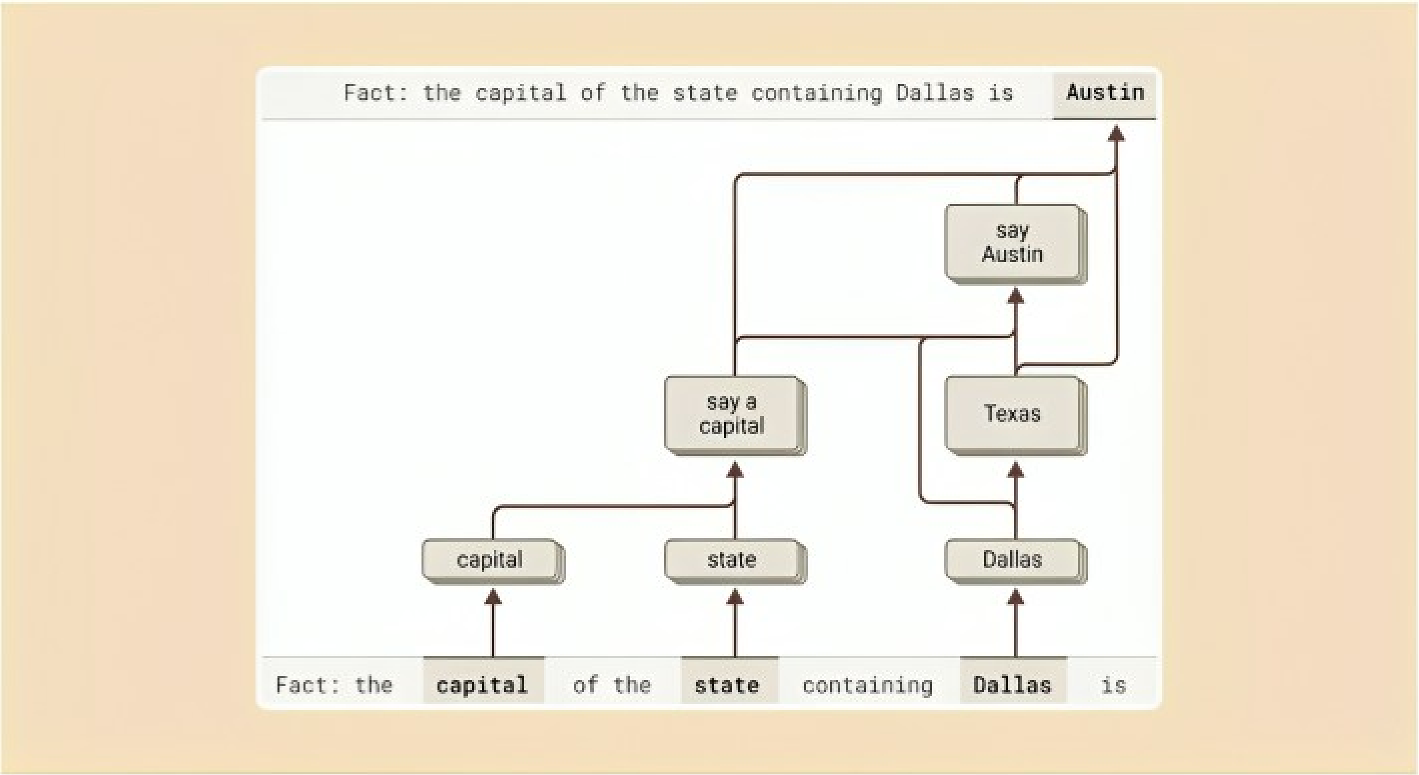
The use of gSMILE for Bias and Fairness Evaluation



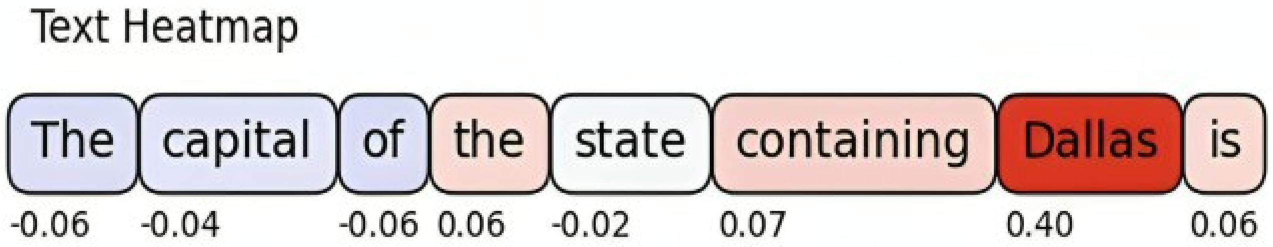
Dehghani, Z., Aslansefat, K., Khan, A., & Akram, M. N. (2025). Explainability of Large Language Models using SMILE: Statistical Model-agnostic Interpretability with Local Explanations. *arXiv preprint arXiv:2505.21657*. *(Submitted to IEEE Transactions on Artificial Intelligence)*.

Model-agnostics vs. Model-specific

Anthropic Research

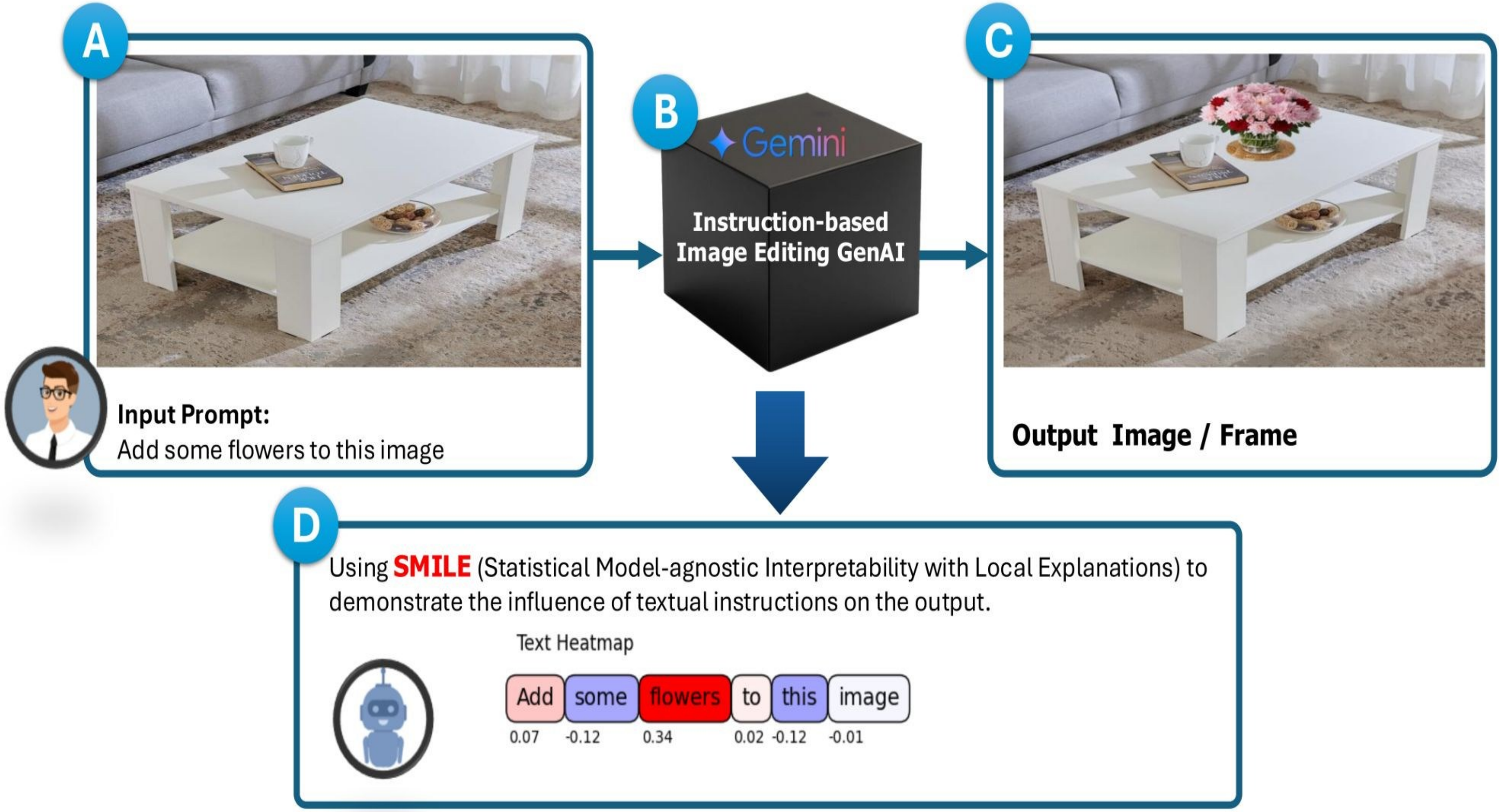

Fact: the capital of the state containing Dallas is Austin
say Austin
say a capital
Texas
capital
state
Dallas
Fact: the capital of the state containing Dallas is

SMILE Explainability



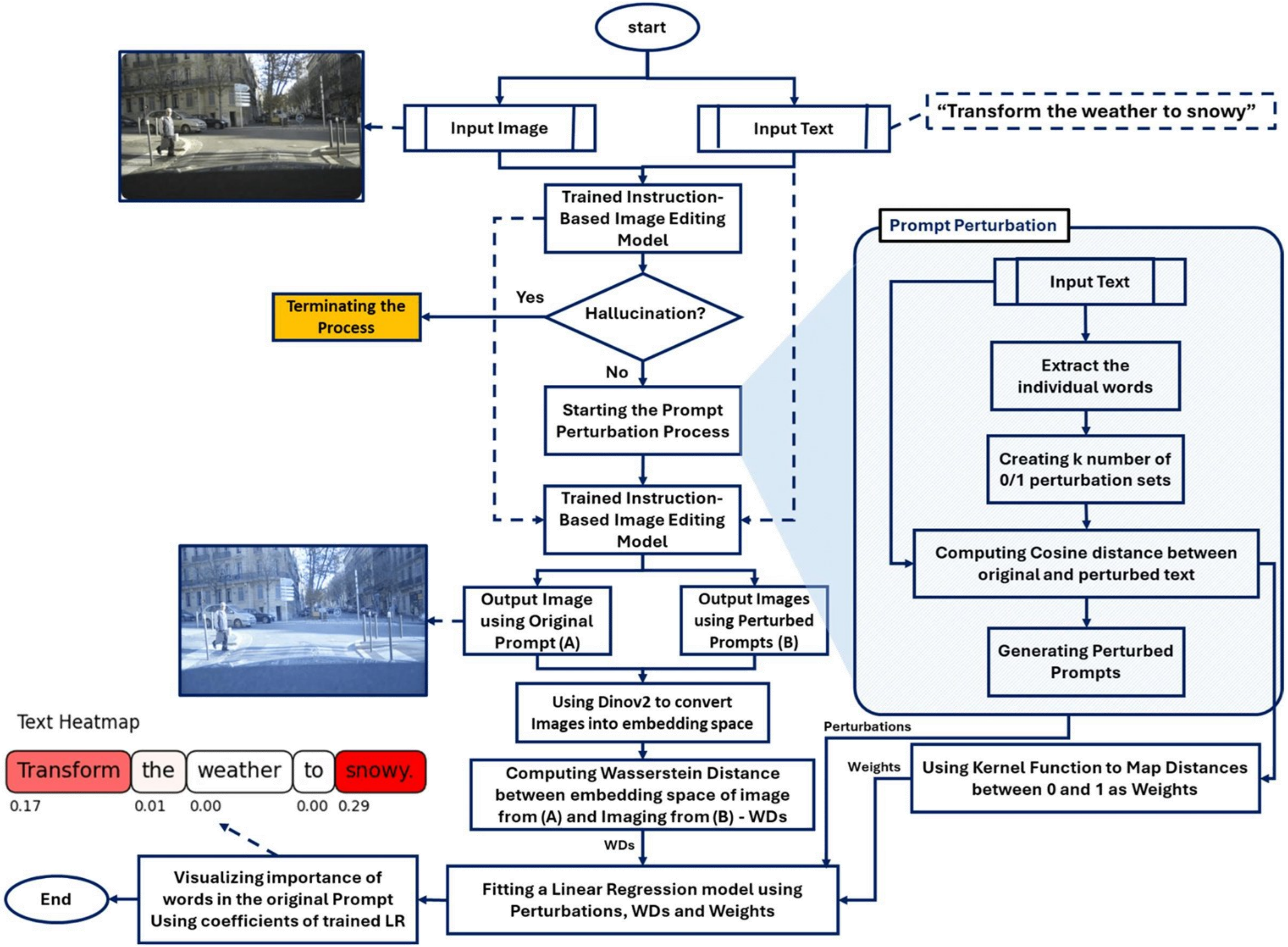
Dehghani, Z., Aslansefat, K., Khan, A., & Akram, M. N. (2025). Explainability of Large Language Models using SMILE: Statistical Model-agnostic Interpretability with Local Explanations. *arXiv preprint arXiv:2505.21657*. *(Submitted to IEEE Transactions on Artificial Intelligence)*.

gSMILE for Multimodal Models


A
Input Prompt:
Add some flowers to this image
B
Gemini
Instruction-based Image Editing GenAI
C
Output Image / Frame
D
Using SMILE (Statistical Model-agnostic Interpretability with Local Explanations) to demonstrate the influence of textual instructions on the output.
Text Heatmap
Add 0.07
some -0.12
flowers 0.34
to 0.02
this -0.12
image -0.01

https://www.kaggle.com/code/zeinabdehghani/explaining-gemini-image-editing-with-smile/notebook

Dehghani, Z., Aslansefat, K., Khan, A., Rivera, A. R., George, F., & Khalid, M. (2024). Mapping the Mind of an Instruction-based Image Editing using SMILE. *arXiv preprint arXiv:2412.16277*. *(Submitted to Nature npj Artificial Intelligence)*.



start

Input Image

Input Text

"Transform the weather to snowy"

Trained Instruction-Based Image Editing Model

Hallucination?

Yes

Terminating the Process

No

Starting the Prompt Perturbation Process

Trained Instruction-Based Image Editing Model

Output Image using Original Prompt (A)

Output Images using Perturbed Prompts (B)

Using Dinov2 to convert Images into embedding space

Computing Wasserstein Distance between embedding space of image from (A) and Imaging from (B) - WDs

WDs

Fitting a Linear Regression model using Perturbations, WDs and Weights

Visualizing importance of words in the original Prompt Using coefficients of trained LR

End

Prompt Perturbation

Input Text

Extract the individual words

Creating k number of 0/1 perturbation sets

Computing Cosine distance between original and perturbed text

Generating Perturbed Prompts

Perturbations

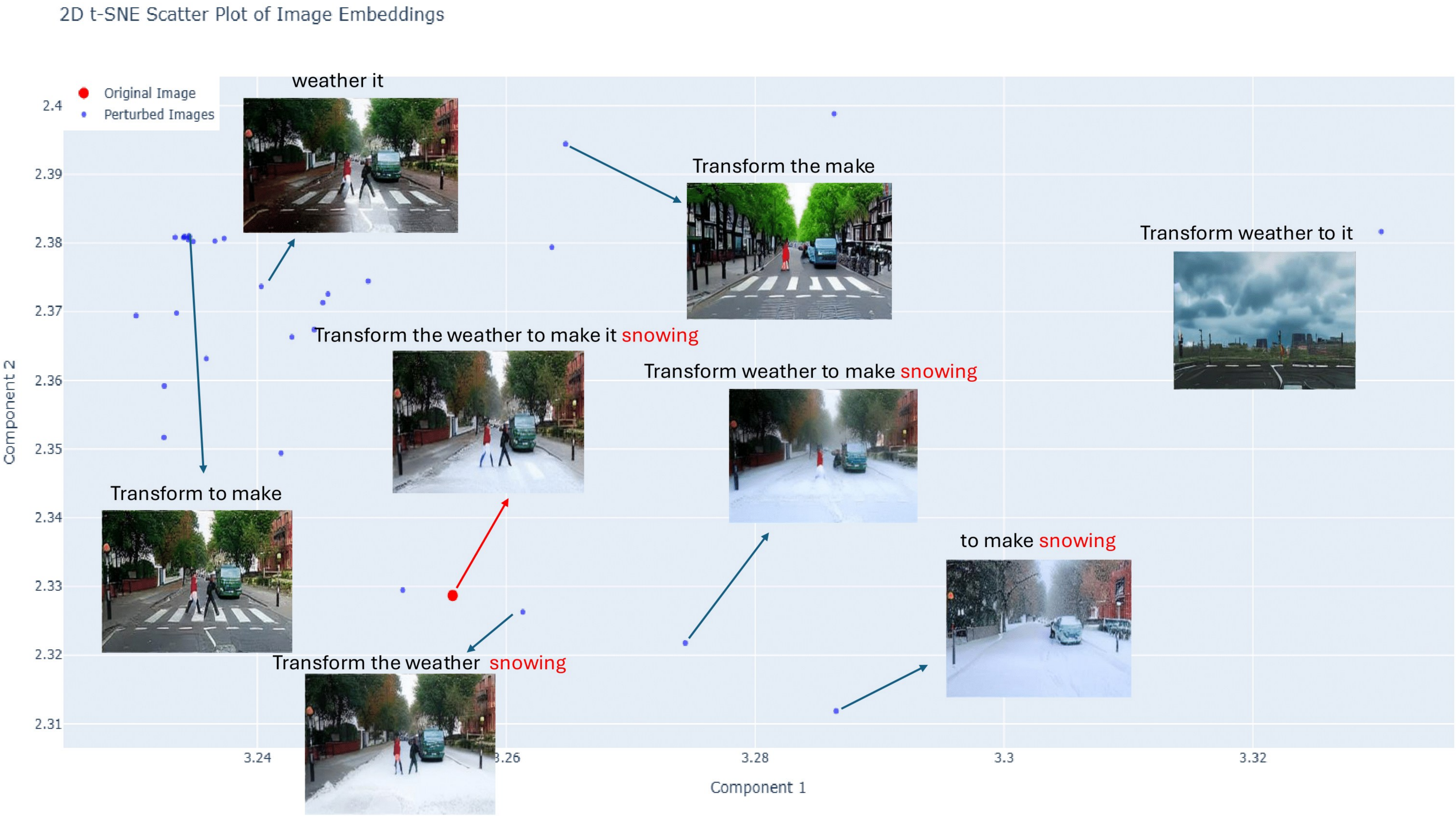
Weights

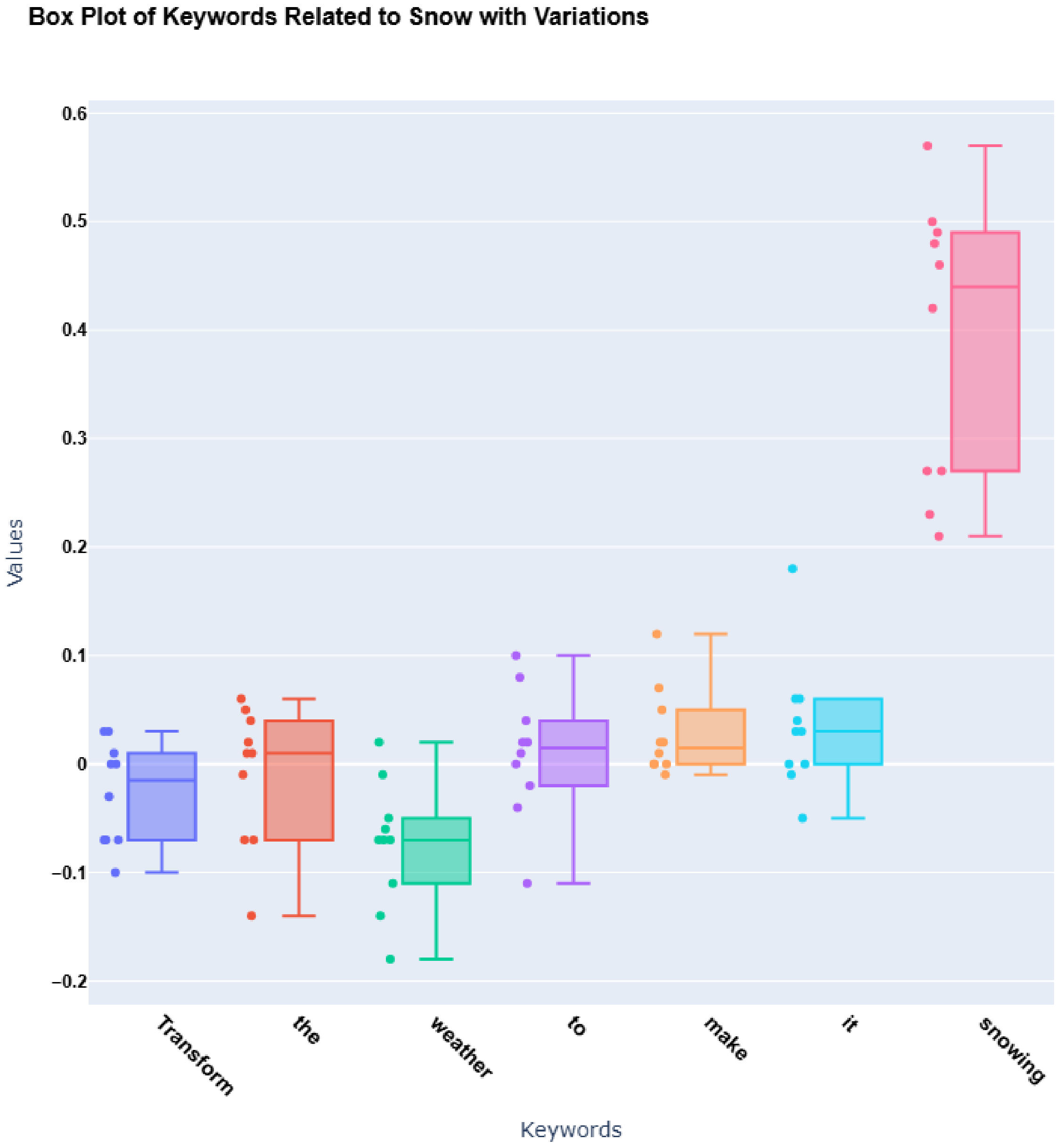
Using Kernel Function to Map Distances between 0 and 1 as Weights

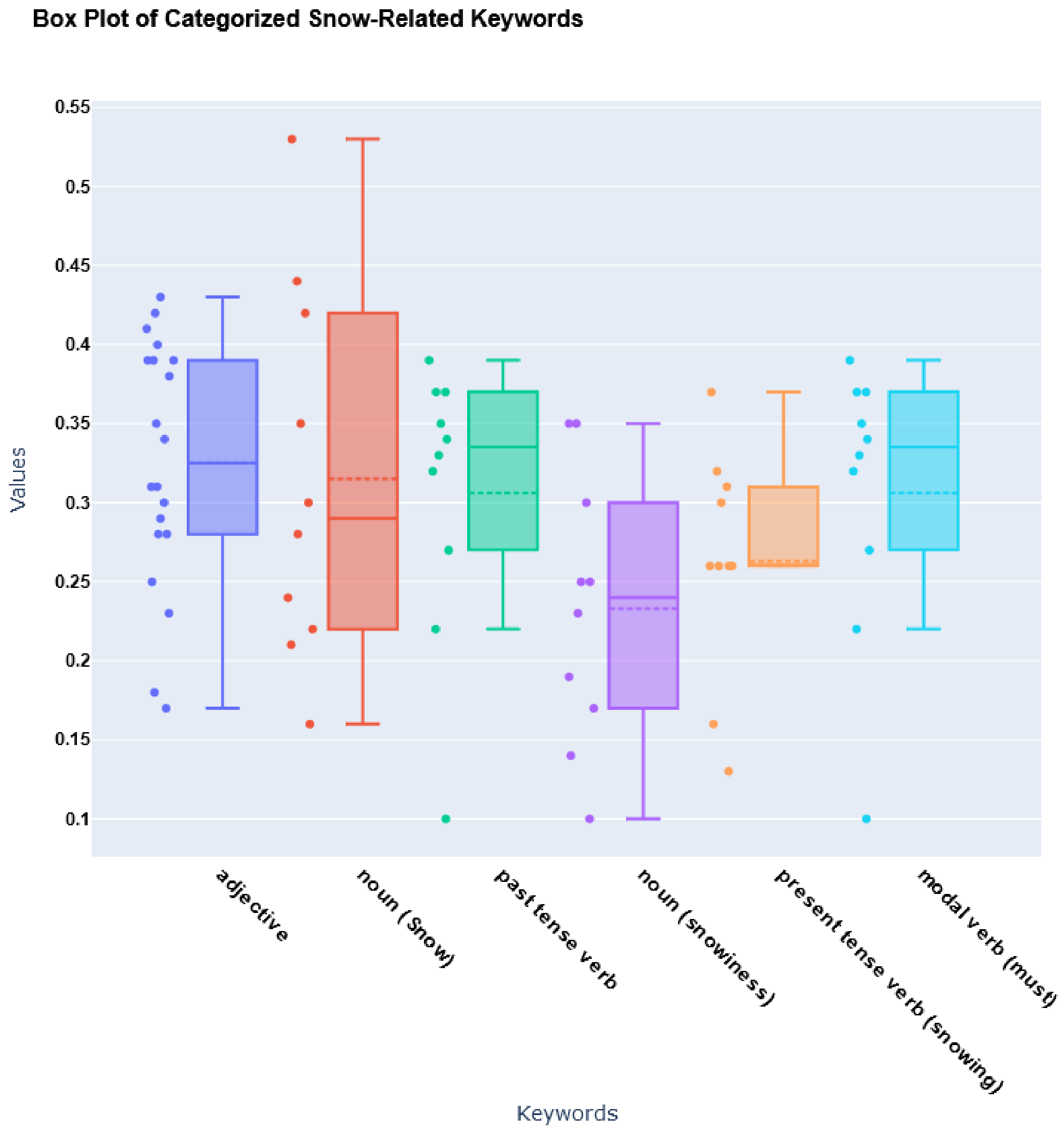
Text Heatmap

Transform	the	weather	to	snowy.
0.17	0.01	0.00	0.00	0.29

Dehghani, Z., Aslansefat, K., Khan, A., Rivera, A. R., George, F., & Khalid, M. (2024). Mapping the Mind of an Instruction-based Image Editing using SMILE. *arXiv preprint arXiv:2412.16277*. *(Submitted to Nature npj Artificial Intelligence)*.

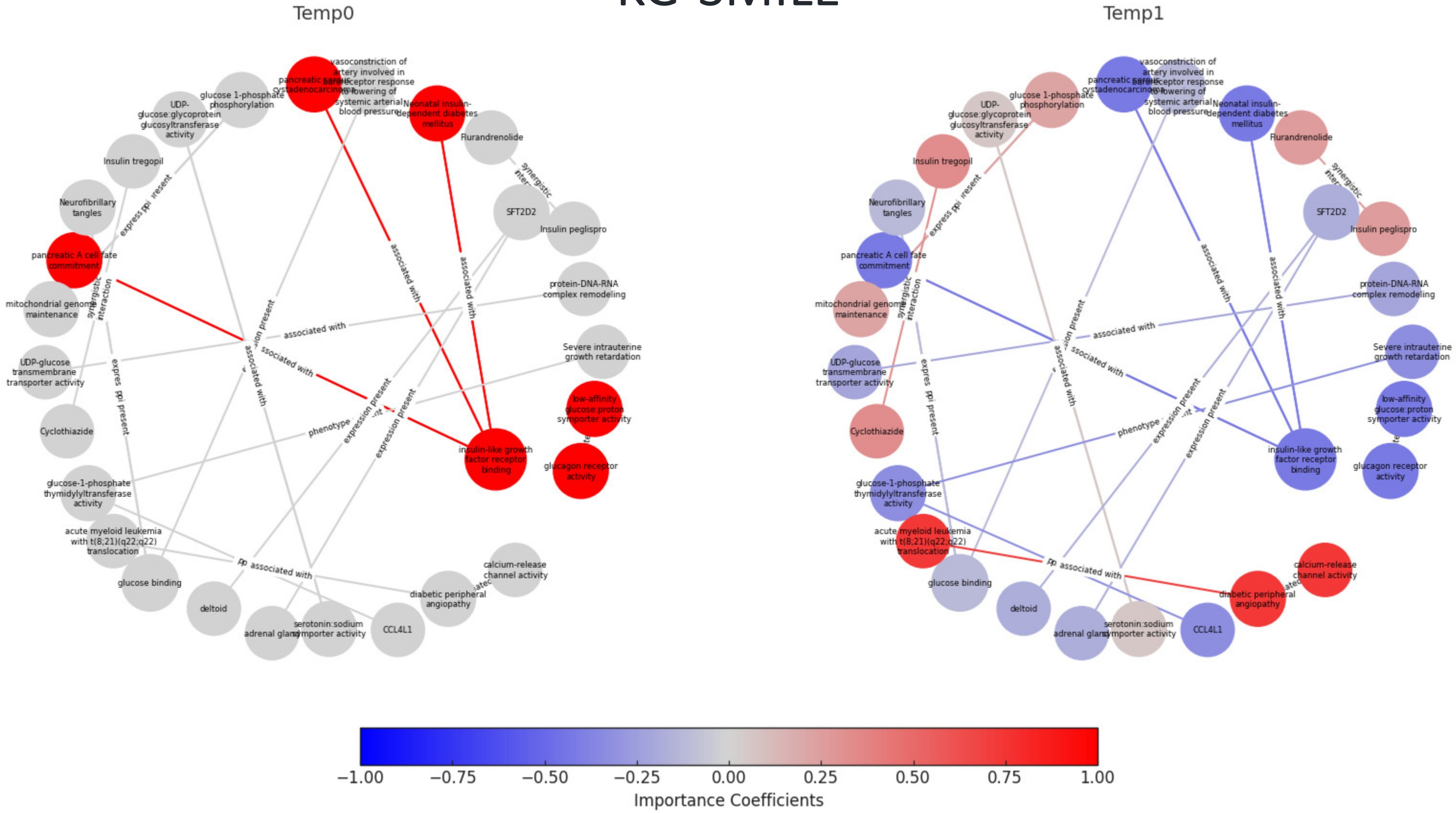

2D t-SNE Scatter Plot of Image Embeddings
● Original Image
● Perturbed Images
weather it
Transform the make
Transform weather to it
Transform the weather to make it snowing
Transform weather to make snowing
Transform to make
to make snowing
Transform the weather snowing
2.4
2.39
2.38
2.37
2.36
2.35
2.34
2.33
2.32
2.31
Component 2
3.24
3.26
3.28
3.3
3.32
Component 1





Dehghani, Z., Aslansefat, K., Khan, A., Rivera, A. R., George, F., & Khalid, M. (2024). Mapping the Mind of an Instruction-based Image Editing using SMILE. *arXiv preprint arXiv:2412.16277*. *(Submitted to Nature npj Artificial Intelligence)*.

KG-SMILE


Temp0
Temp1
vasoconstriction of artery involved in baroreceptor response to lowering of systemic arterial blood pressure
pancreatic serous cystadenocarcinoma
glucose 1-phosphate phosphorylation
UDP-glucose:glycoprotein glucosyltransferase activity
Neonatal insulin-dependent diabetes mellitus
Flurandrenolide
Insulin tregopil
synergistic inter
SFT2D2
Neurofibrillary tangles
express ppi present
Insulin peglispro
pancreatic A cell fate commitment
synergistic interaction
protein-DNA-RNA complex remodeling
mitochondrial genome maintenance
associated with
Severe intrauterine growth retardation
UDP-glucose transmembrane transporter activity
expres ppi present
expression present
low-affinity glucose:proton symporter activity
phenotype
Cyclothiazide
insulin-like growth factor receptor binding
glucagon receptor activity
glucose-1-phosphate thymidylyltransferase activity
acute myeloid leukemia with t(8;21)(q22;q22) translocation
ppi associated with
calcium-release channel activity
glucose binding
diabetic peripheral angiopathy
deltoid
adrenal gland
serotonin:sodium symporter activity
CCL4L1
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00
Importance Coefficients

Moghaddam, Z. Z. S., Dehghani, Z., Rani, M., Aslansefat, K., Mishra, B. K., Kureshi, R. R., & Thakker, D. (2025). Explainable Knowledge Graph Retrieval-Augmented Generation (KG-RAG) with KG-SMILE. *arXiv preprint arXiv:2509.03626*. (Submitted to Information Systems Frontiers).

Conclusion

- ✓ SMILE extends LIME by integrating statistical distance measures (e.g., Wasserstein, Cramér–von Mises, Kuiper) to produce more stable and reliable local explanations.
- ✓ The method is completely model-agnostic, making it suitable for explaining classical ML, deep learning, and Generative AI (gSMILE) models.
- ✓ SMILE inherits the visual interpretability of LIME and the quantitative depth of SHAP, combining their advantages while improving robustness.
- ✓ Across images, text, and 3D point cloud modalities, SMILE consistently delivers more coherent and less noisy feature attributions.
- ✓ Experimental results demonstrate that SMILE is more resilient to adversarial manipulation than LIME and SHAP, though not entirely immune.
- ✓ The framework now supports LLMs, VLMs, and multimodal models through gSMILE, enabling explainability, fidelity analysis, and bias/fairness evaluation for generative tasks.
- ✓ Future directions include expanding SMILE for various directions, including Image Captioning, QML, and Geo-SMILE.

SMILE Reproducibility

GitHub https://github.com/Dependable-Intelligent-Systems-Lab/xwhy

GitHub https://github.com/koo-ec/KG-SMILE

GitHub https://github.com/Sara068/Mapping-the-Mind-of-an-Instruction-based-Image-Editing-using-SMILE

GitHub https://github.com/Sara068/LLM-SMILE

GitHub https://github.com/koo-ec/Geo-SMILE (Not Public Yet)

Publications

[1] Aslansefat, K., Hashemian, M., Walker, M., Akram, M. N., Sorokos, I., & Papadopoulos, Y. (2023). Explaining black boxes with a SMILE: Statistical Model-agnostic Interpretability with Local Explanations. ***IEEE Software***, *41*(1), 87-97.

[2] Ahmadi, S. M., Aslansefat, K., Valcarce-Dineiro, R., & Barnfather, J. (2024). Explainability of Point Cloud Neural Networks Using SMILE: Statistical Model-Agnostic Interpretability with Local Explanations. *arXiv preprint arXiv:2410.15374 (Submitted to Engineering Applications of Artificial Intelligence)*.

[3] Dehghani, Z., Aslansefat, K., Khan, A., Rivera, A. R., George, F., & Khalid, M. (2024). Mapping the Mind of an Instruction-based Image Editing using SMILE. *arXiv preprint arXiv:2412.16277*. *(Submitted to Nature npj Artificial Intelligence)*.

[4] Dehghani, Z., Aslansefat, K., Khan, A., & Akram, M. N. (2025). Explainability of Large Language Models using SMILE: Statistical Model-agnostic Interpretability with Local Explanations. *arXiv preprint arXiv:2505.21657*. *(Submitted to IEEE Transactions on Artificial Intelligence)*.

[5] Moghaddam, Z. Z. S., Dehghani, Z., Rani, M., Aslansefat, K., Mishra, B. K., Kureshi, R. R., & Thakker, D. (2025). Explainable Knowledge Graph Retrieval-Augmented Generation (KG-RAG) with KG-SMILE. *arXiv preprint arXiv:2509.03626*. (Submitted to Information Systems Frontiers).

References

Here you can find the references that has been used in this presentation

[1] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August). " Why should i trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 1135-1144).

[2] Lundberg, S. M., & Lee, S. I. (2017, December). A unified approach to interpreting model predictions. In Proceedings of the 31st international conference on neural information processing systems (pp. 4768-4777).

[3] Slack, D., Hilgard, S., Jia, E., Singh, S., & Lakkaraju, H. (2020, February). Fooling lime and shap: Adversarial attacks on post hoc explanation methods. In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (pp. 180-186).

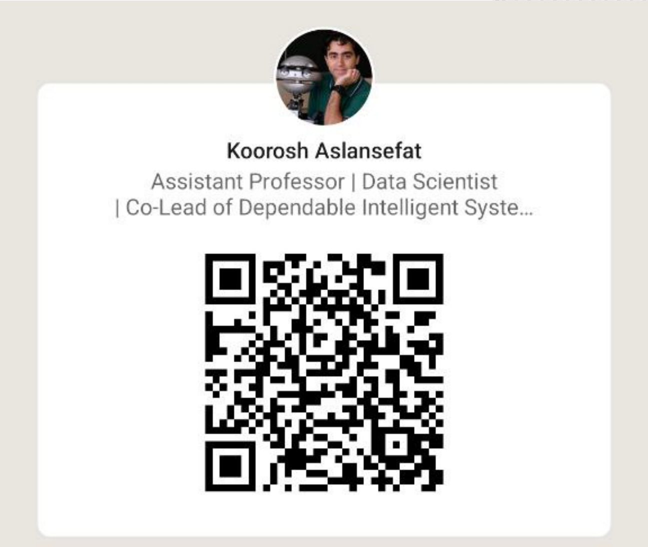
[4] Ghorbani, A., Abid, A., & Zou, J. (2019, July). Interpretation of neural networks is fragile. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 33, No. 01, pp. 3681-3688).

[5] Belle, V., & Papantonis, I. (2021). Principles and practice of explainable machine learning. Frontiers in big Data, 39.

[6] Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58, 82-115.

Thank You

If you have any question, please feel free to ask



Koorosh Aslansefat

Assistant Professor | Data Scientist
| Co-Lead of Dependable Intelligent Syste...

k.aslansefat@hull.ac.uk