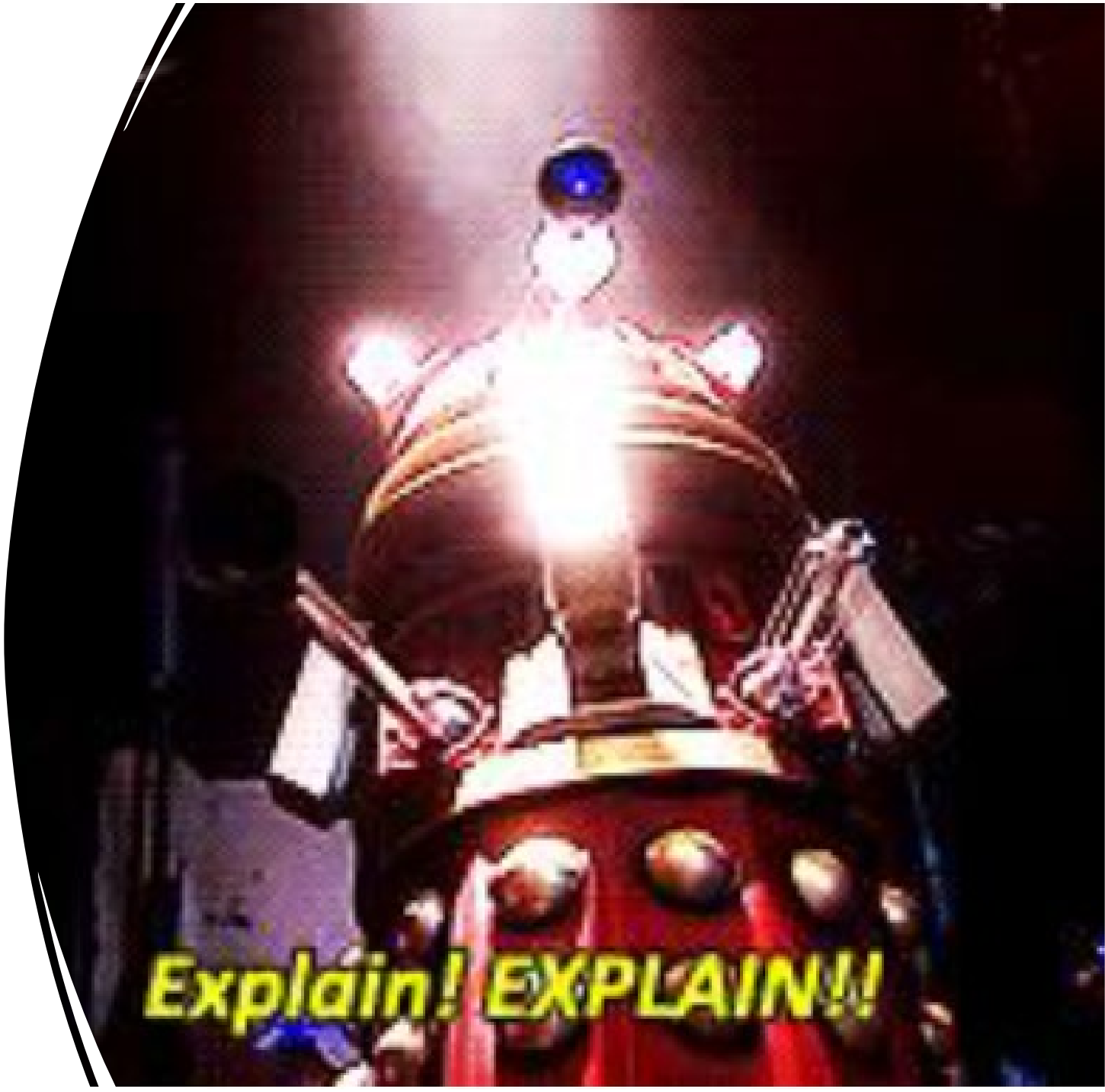
X-AI

Explainable Artificial Intelligence

Qasim Zia

Interpretability

- Why Interpret ?



The current state of machine learning

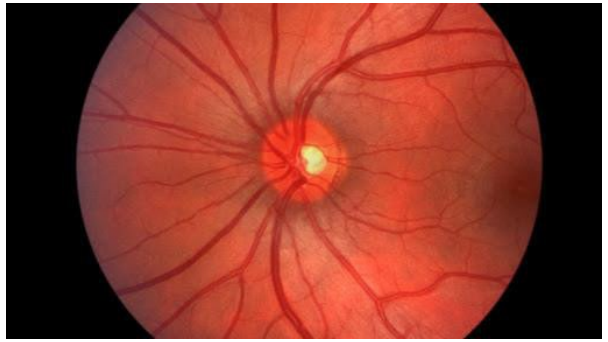
And its uses

...

https://www.tesla.com/videos/autopilot-self-driving-hardware-neighborhood-long

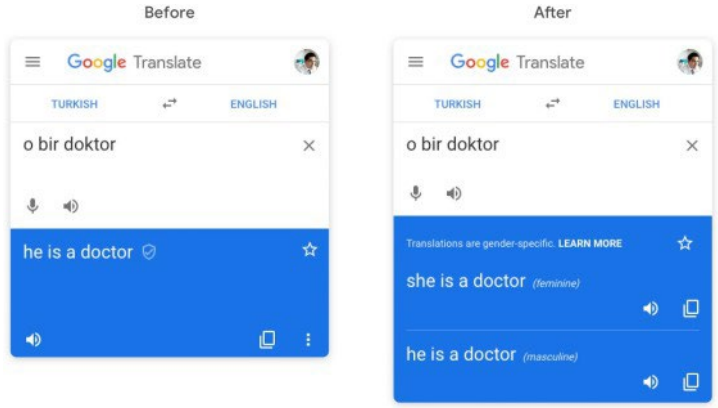
NYPost

MIT Technology Review



DeepMind

DeepMind



So are we in the golden age of AI ?

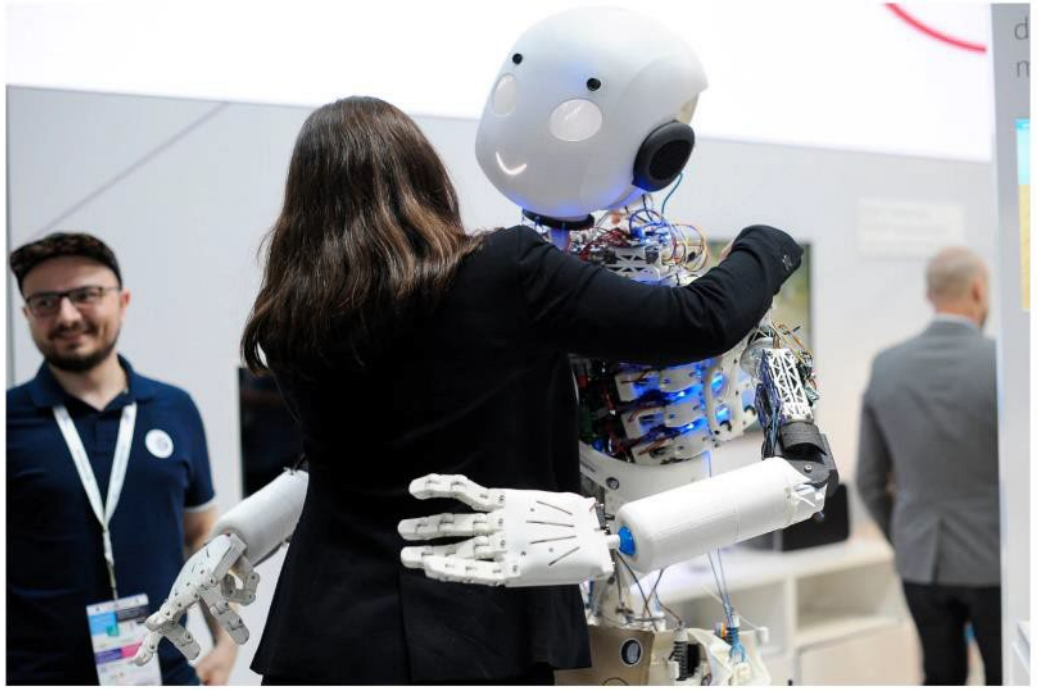
Safety and well being

Tesla hit parked police car 'while using Autopilot'

30 May 2018 Share

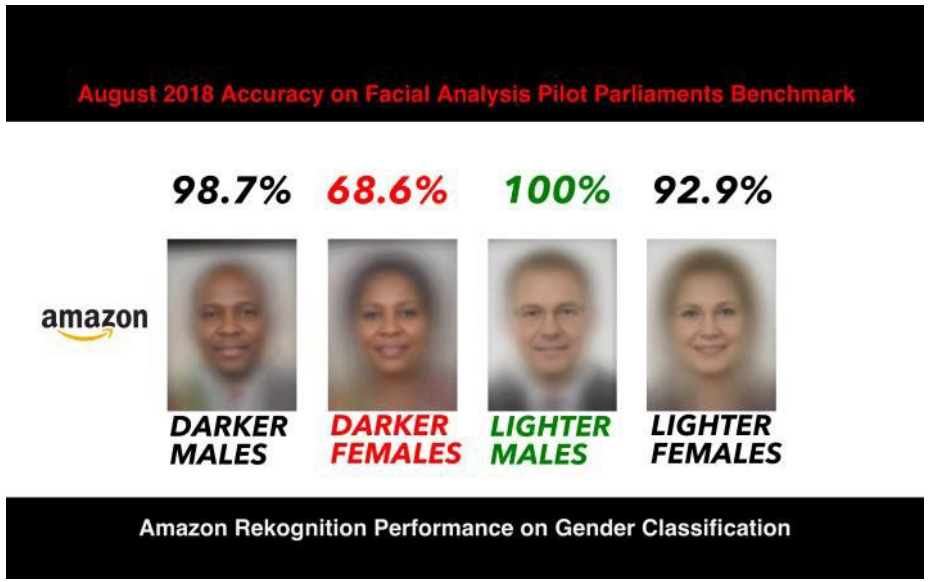
A number of Tesla vehicles have been involved in crashes.

Warnings of a Dark Side to A.I. in Health Care

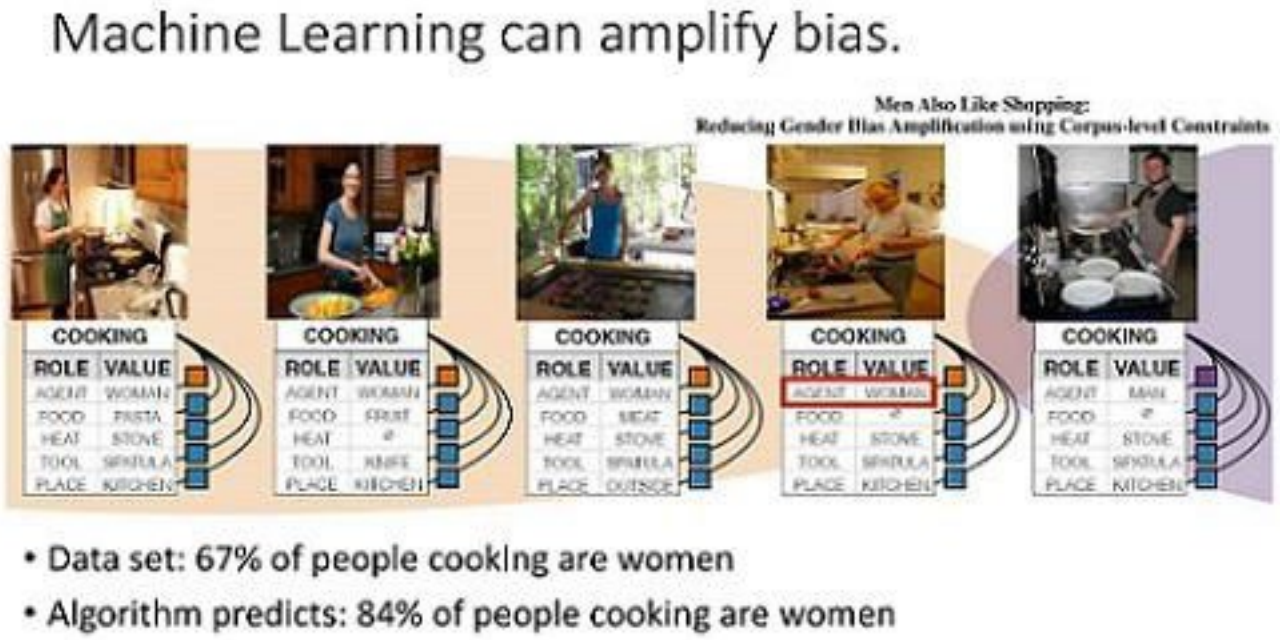


Scientists worry that with just tiny tweaks to data, neural networks can be fooled into committing "adversarial attacks" that mislead rather than help. Joan Cros/NurPhoto, via Getty Images

Bias in algorithms



https://medium.com/@Joy.Buolamwini/response-racial-and-gender-bias-in-amazon-rekognition-commercial-ai-system-for-analyzing-faces-a289222eeced

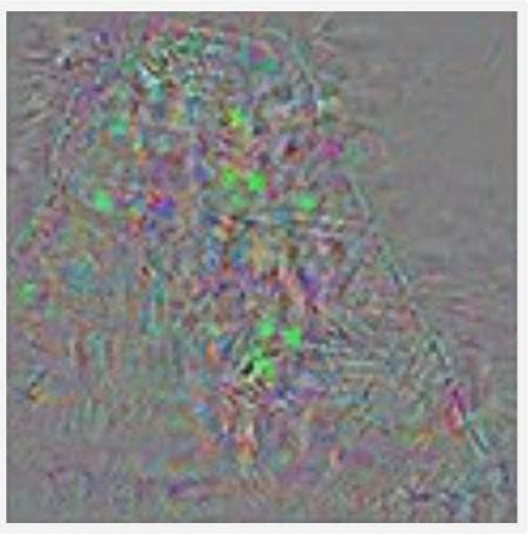


https://www.infoq.com/presentations/unconscious-bias-machine-learning/

Adversarial Examples

Original image

Temple (97%)



Perturbations

Adversarial example

Ostrich (98%)

Legal Issues - GDPR

And more …

- Interactive feedback - can model learn from human actions in online setting ? (Can you tell a model to not repeat a specific mistake ?)

- Recourse – Can a model tell us what actions we can take to change its output ? (For example, what can you do to improve your credit score?)

In general, it seems like there are few fundamental problems –

- We don't trust the models
- We don't know what happens in extreme cases
- Mistakes can be expensive / harmful
- Does the model makes similar mistakes as humans ?
- How to change model when things go wrong ?

Interpretability is one way we try to deal with these problems

What is interpretability ?

There is no standard definition –

Most agree it is something different from performance.

Ability to explain or to present a model in understandable terms to humans (Doshi-Velez 2017)

Cynical view – It is what makes you feel good about the model.

It really depends on target audience.

Model based vs Model Agnostic

- **Can it explain only few classes of models ?**

Example –

Rationales
LR / Decision Trees
Attention
Gradients (Differentiable Models only)

- **Can it explain any model ?**

Example –

LIME – Locally Interpretable Model Agnostic Explanations

SHAP – Shapley Values

What is LIME?

Local

Interpretable

Model-agnostic

Explanations

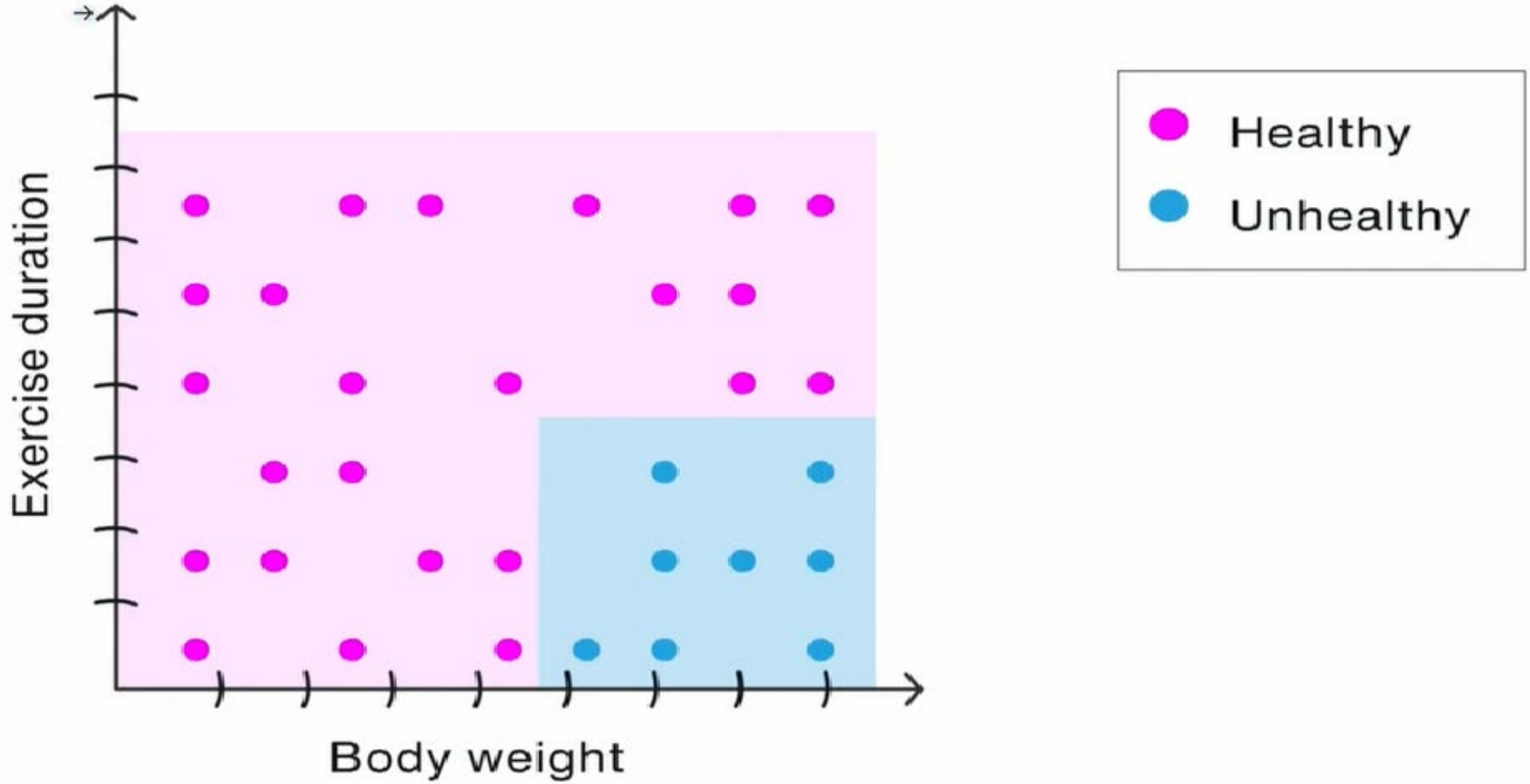
Local neighborhood of the instance

A human should be able to interpret

Applicable for all models

Explanation that helps the interpretation

Global and Local Interpretability

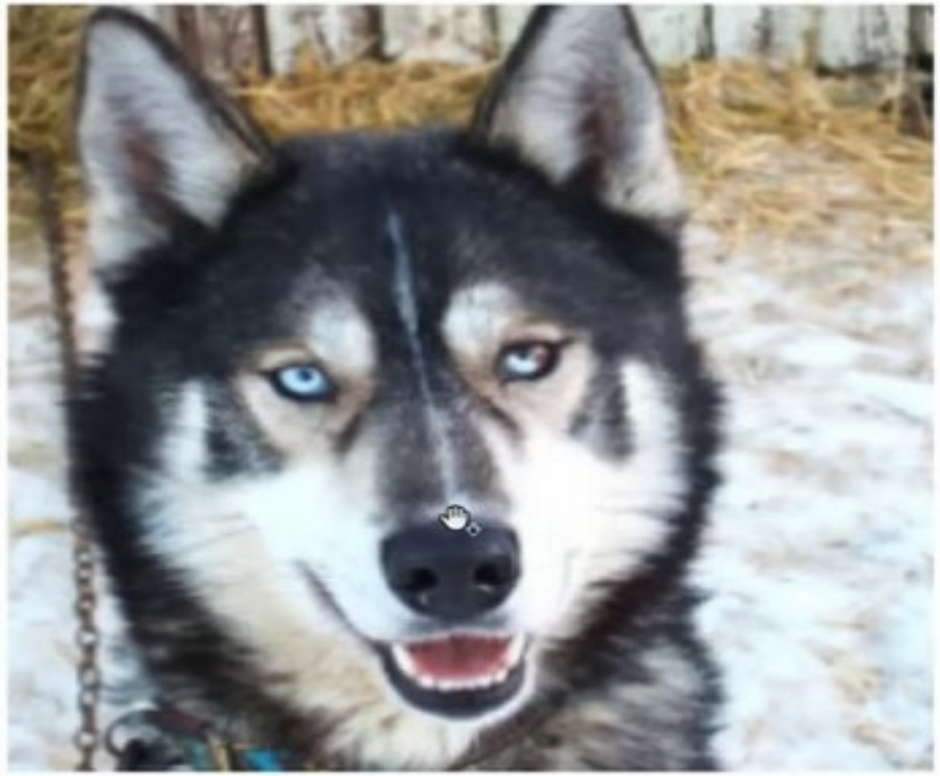


Text Classification

I have a medical emergency. Hence won't be able to attend the meeting today.	Important
Hi may I get the information about your service?	Not important

I have a medical emergency. Hence won't be able to attend the meeting today.	Important
Hi may I get the information about your service?	Not important

Husky or Wolf?

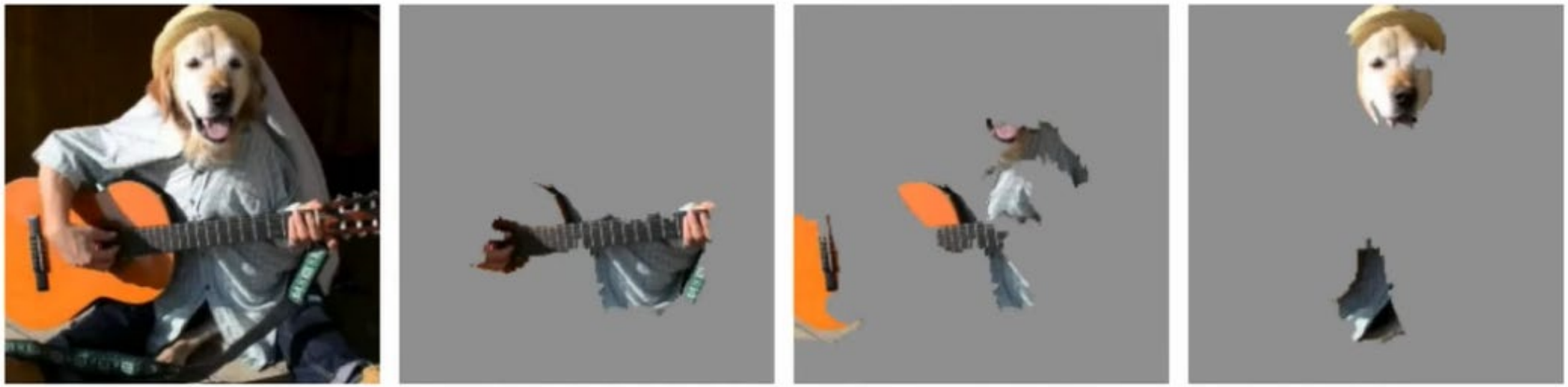


(a) Husky classified as wolf

(b) Explanation

The model detects snow instead of wolf. It cannot be trusted!

Explaining an Image Classification Prediction



(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

Explaining an image classification prediction made by Google's Inception neural network. The top 3 classes predicted are "Electric Guitar" (p = 0.32), "Acoustic guitar" (p = 0.24) "Labrador" (p = 0.21).

The model detects features correctly. It can be trusted.

Intuition Behind LIME

Linear models are readily interpretable.

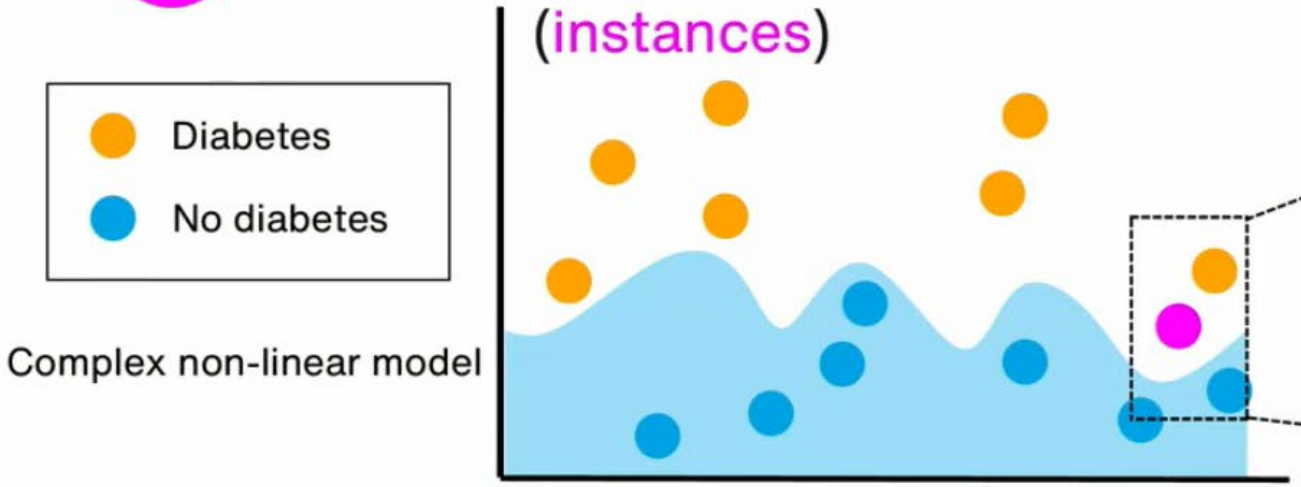
For example if we have

$$y = w_1x_1 + w_2x_2 + w_3x_3$$

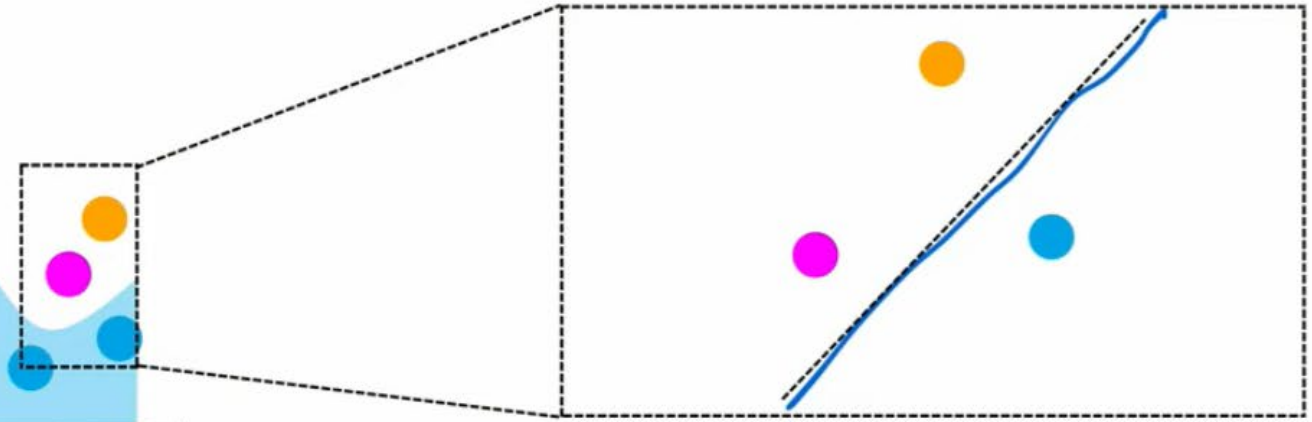
w1x1 is the contribution to the prediction of the feature x1 for the specific data instance and a high value means a high contribution.

LIME Implementation in 4 Steps

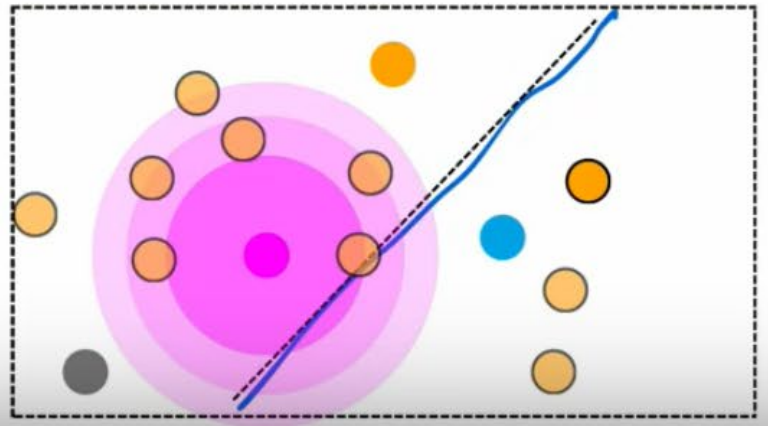
1 Global model, local data point (instances)



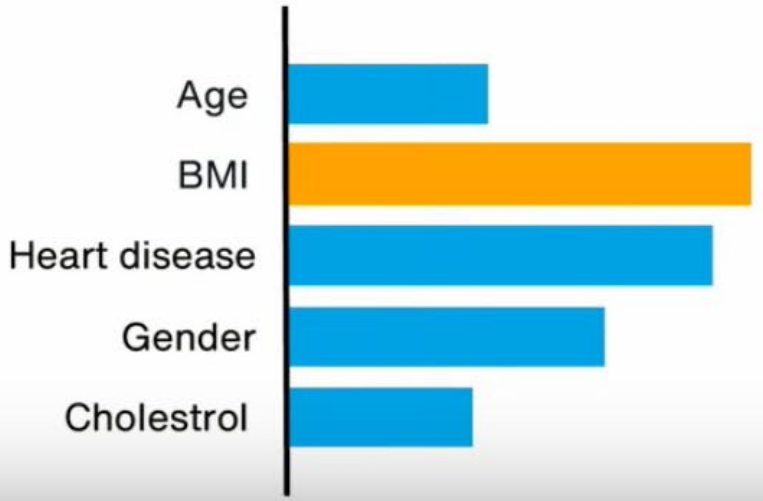
2 Local linear model near the point



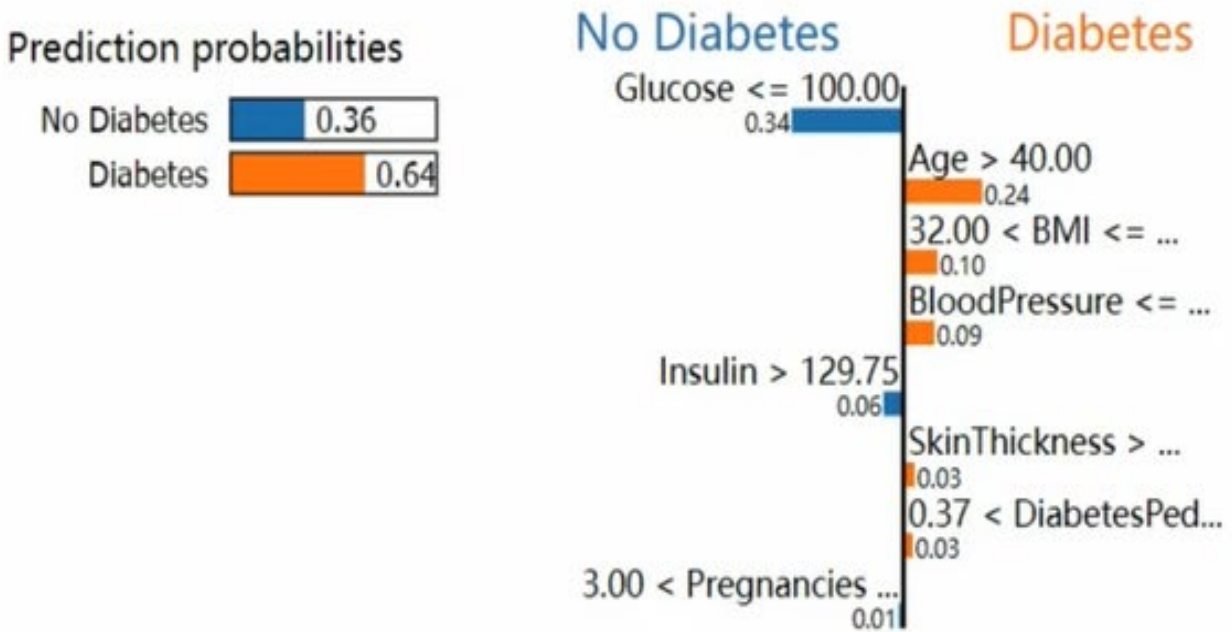
3 Perturbed data weighted according to the distance to the local point



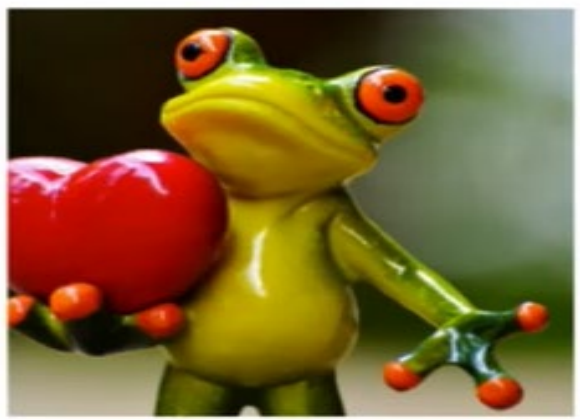
4 Features relevant to the data point



LIME Prediction



LIME Applied to Images



Original Image

Interpretable Components

Subdivide the image to interpretable components

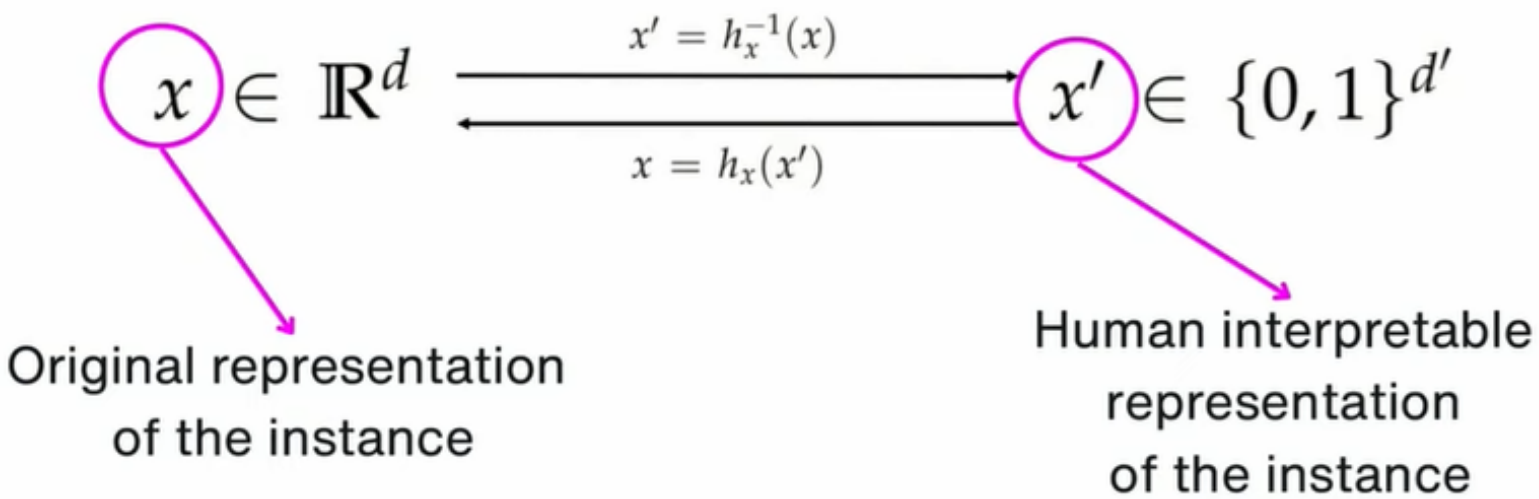
What Does The Neural Network See?

P() = 0.54

P() = 0.07

P() = 0.05

LIME Math: Data Representation



Since tabular data is already interpretable by humans, there is no need for transformation

$$x' \in \mathbb{R}^{d'} \longrightarrow x = x' \quad (d = d')$$

LIME Math: Optimization Problem

$$\xi(x) = \underset{g \in G}{\arg\min}\, \mathcal{L}(f, g, \pi_x) + \Omega(g)$$

zeta

Predicted instance we want to explain

Family of interpretable models

Complex model being explained

Simple/surrogate model

Neighbourhood of x

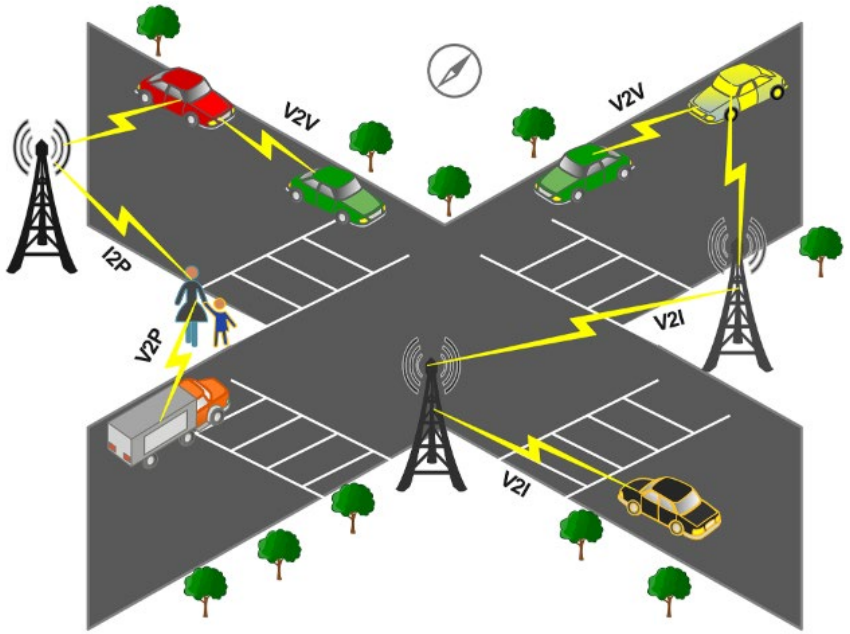
Complexity measure of g

x be a row in a dataframe

Gender	Age	Exercise minutes	BMI	Weight
Male	37	125	25	76
Female	66	20	21	53
Female	71	12	28	77

Lime in DT-VANET

- Improved Explain ability of AI Models
- Real-time Diagnostics
- Model Validation and Debugging
- Enhanced Safety Protocols



References

- https://vizuara.ai/
- Ribeiro, M.T., Singh, S. and 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).
- **Interpretability in Machine Learning** Northeastern University Course