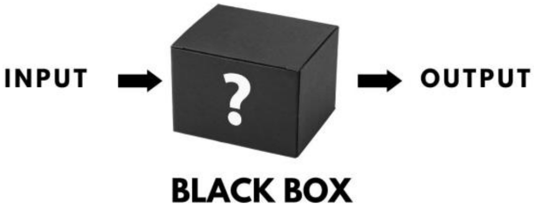
Sardar Vallabhbhai National Institute of Technology, Surat, India

Recent Advancement in Artificial Intelligence and Robotics (RA-AIR-24), 08-13, July-2024



Explainable AI (XAI): Making AI Understandable to Humans

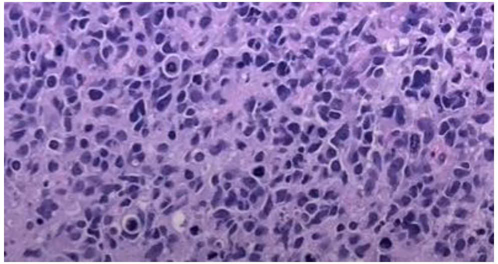
Praveen Kumar Chandaliya
Department of Artificial Intelligence
pkc@aid.svnit.ac.in
SVNIT, Surat

AI Successes

- Science (Astronomy, neuroscience, medical imaging, bio-informatics)
- Environment (Energy, climate, weather, resources)
- Retail (Intelligent stock control, demographic store placement)
- Manufacturing (Intelligent control, automated monitoring, detection methods)
- Security (Intelligent smoke alarms, fraud detection)
- Marketing (Promotions, ...)
- Management (Scheduling, timetabling)
- Finance (Credit scoring, risk analysis...)
- Web data (Information retrieval, information extraction, ...)

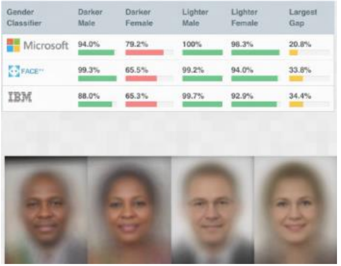
Where AI is yet to fulfill its promise

Arizona police released photographs from the pedestrian death involving an Uber self-driving car.



Disease misclassification

safety-critical systems: glass cockpit of a C-141, Space Shuttle and control room of a nuclear power plant.



Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
Microsoft	94.0%	79.2%	100%	98.3%	20.8%
FACE++	99.3%	65.5%	99.2%	94.0%	33.8%
IBM	88.0%	65.3%	99.7%	92.9%	34.4%

COTS gender classification

Characterizing these applications

- **Wrong decisions can be costly and dangerous**
- **Accuracy is not the only objective**
- **Need for multi-dimensional perspective**

Why Explainable AI ?

- Human-understandable rationale in decision-making
- Trust/confidence in system
- Compliance with ethical principles
- Enhanced control and robustness
- Openness of discovery and scientific research

European Union's General Data Protection Regulation (GDPR)

"A business using personal data for automated processing-making must be able to explain how the system makes decisions, See Article 15 (1) (h) and Recital of GDPR"

Why Explainable AI ?

India's Digital Personal Data Protection Act (DPDP) , 2023, India

An Act to provide for the processing of digital personal data in a manner that recognises both the right of **individuals to protect their personal data** and the need to process such personal data for lawful purposes and for matters connected therewith or incidental thereto.

Right to an explanation is a right to be given an explanation for an output of the algorithm

Goal of XPI: FATE in AI

Fairness

Accountability

Transparency

Ethics

Transparency: Ensuring that AI systems are transparent and their decision-making processes are understandable by users and stakeholders.

Fairness: Mitigating bias and discrimination in AI-driven decision-making to promote equitable outcomes.

Accountability: Establishing clear lines of responsibility and liability for the actions and decisions of AI systems.

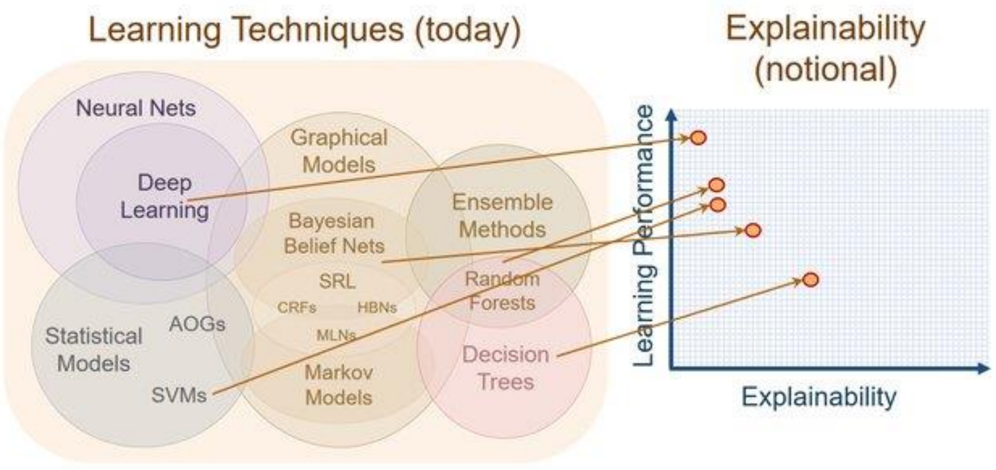
Privacy: Protecting sensitive data and personal information used in the development and deployment of XAI systems.

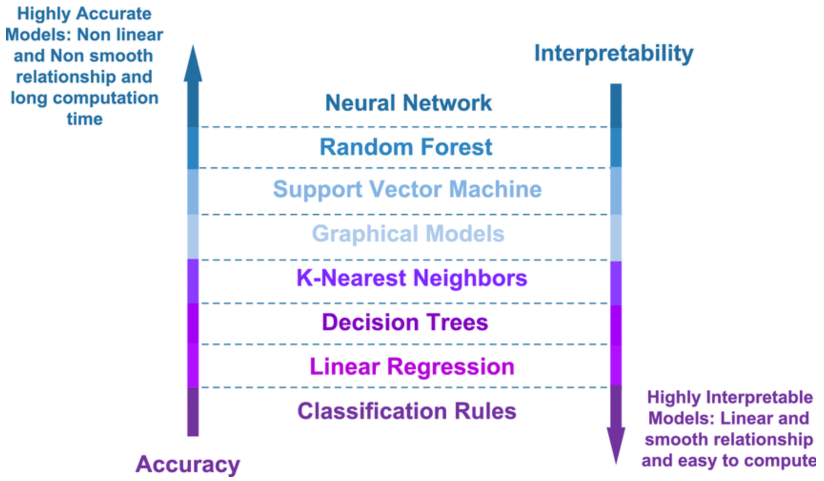
Artificial Intelligence (AI) is at the forefront of modern technology, and its effects are felt in many areas of society. To prevent algorithmic disparities, fairness, accountability, transparency, and ethics (FATE) in AI are being implemented

Credit: https://www.microsoft.com/en-us/research/theme/fate/

Credit: Lecue et al., Tutorial on XAI. AAAI 2020. https://xaitutorial2020.github.io/

Trade-off between accuracy vs. explainability



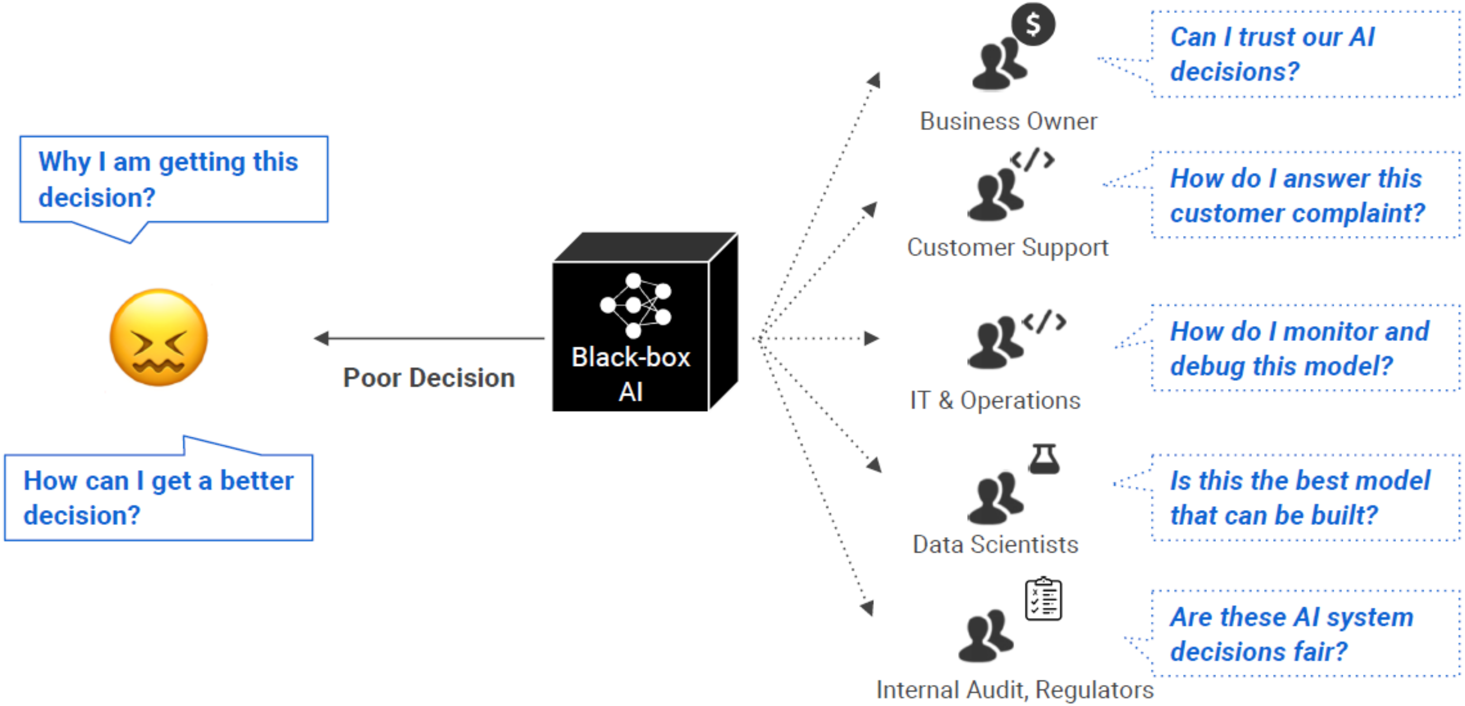


A trade-off between accuracy and explainability

67% of the businesses leaders taking part in PwC's 2017 Global CEO Survey believe that **AI and automation will impact negatively on stakeholder** trust levels in their industry in next five years.

Source: Taymouri et al., Business Process Variant Analysis: Survey and Classification

Whom is XAI for ?



Explainability vs. Interpretability

Interpretability

- Understand what a model did or might have done.

Explainability

- Summarizing the reason for neural network behaviour, gaining trust of users, producing insights or causes of decisions

XAI Taxonomy

Agnosticity

Model-**agnostic**	Applicable to all model types
Model-**specific**	Only applicable to a specific model type

Scope

Global explanation	Explaining the whole model
Local explanation	Explaining individual predictions

Data Type

- Graph
- Image
- Text
- Tabular

Explanation Type

Visual

Data visualisation techniques may be used to understand the prediction or choice made over the input data.

Feature importance

After all possible combinations, we get the feature importance based on its average predicted marginal contribution to the model's decision.

Data points

This category includes all methods that return data points (already existent or newly created) to make a model interpretable.

Surrogate models

We can explain our complex model’s prediction by using a simplified model (surrogate model) to approximate it around the prediction.

LIME: Local Interpretable Model agnostic Explanation

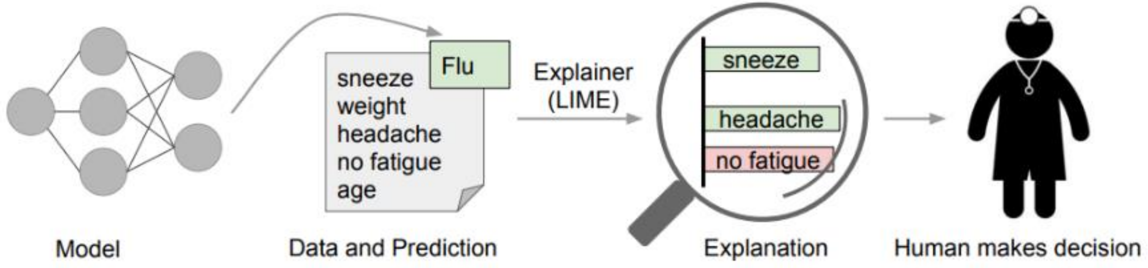
Local**:** Instead of explaining the predictions globally by building a global surrogate model, LIME focuses on training local surrogate models to explain individual prediction.

Interpretable**:** The Idea behind LIME is to make easy to interpretable local model such as linear regression which can explain your black box model locally.

Model Agnostic: LIME can be applied to any black box model irrespective of the technique as long as it can predict probability .

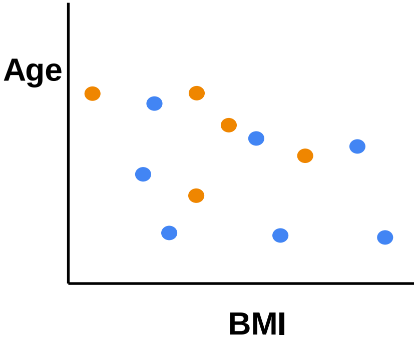
Explanations**:** LIME will be able to explain how the model behaves, which features it picks up and what kinds of interactions between them takes place to drive the predictions

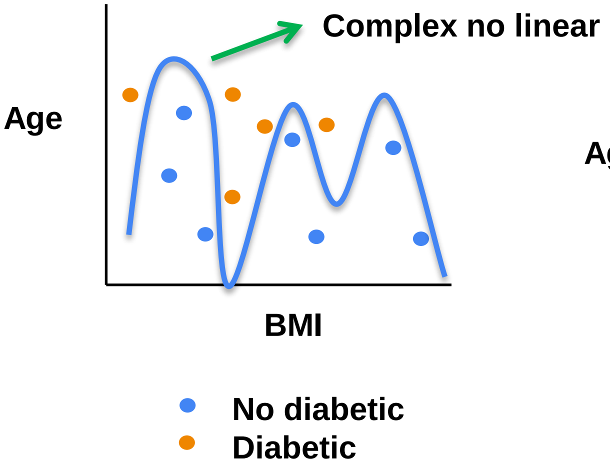
- Lime is a post hoc technique, which means that this is applied **after the event i.e. after model training.**
- Model internals are “hidden”, it works on tabular, image, graph, and text data.
- Explanations are **locally** faithful, but not necessarily globally

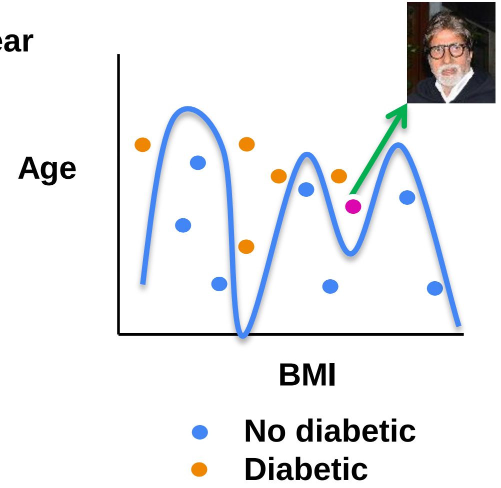


LIME: Diabetic Database

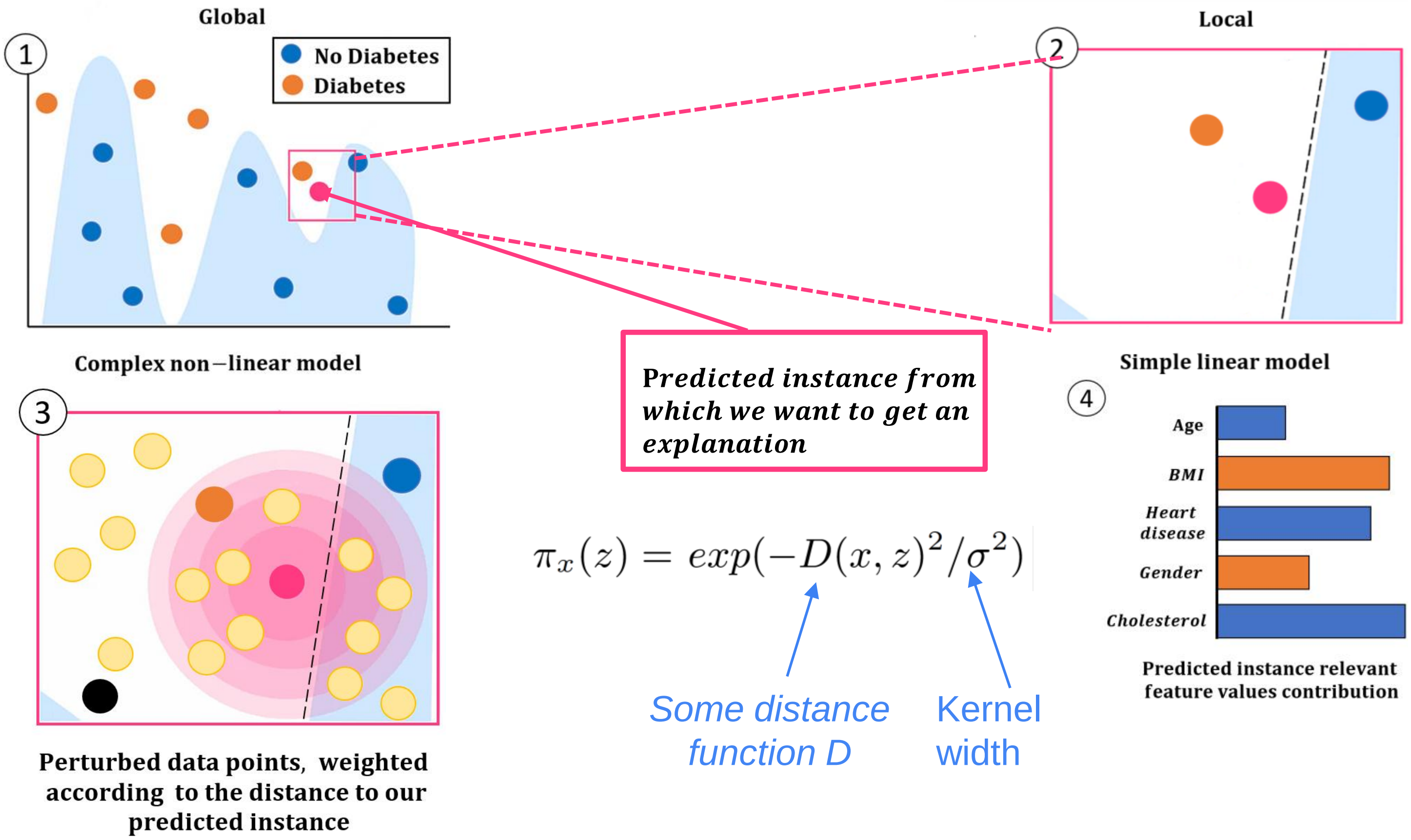
Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	Diabetes Pedigree Function	Age	Outcome
6	148	72	35	0	33.6	0.627	50	**1**
1	85	66	29	0	26.6	0.351	31	**0**
8	183	64	0	0	23.3	0.672	32	**1**
1	89	66	23	94	28.1	0.167	21	**0**







LIME step by step



$$\pi_x(z) = exp(-D(x,z)^2/\sigma^2)$$

The math in LIME

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

Family of interpretable models (Linear regression)

Complex model

$f : \mathbb{R}^d \rightarrow \mathbb{R}$

Simple interpretable model

Local neighbourhood of x (Proximity)

Regularizar

$x \in \mathbb{R}^d \longrightarrow$ number of features

$Age = 56$	$Gender = F$	$BMI = 30$	……	$diabetic = yes$

Loss terms

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

Train a weighted, interpretable model on the dataset with the perturbed instances

(1) $$\mathcal{L}(f, g, \pi_x) = \sum_{z \in \mathcal{Z}} \pi_x(z)\left(f(z) - g(z)\right)^2$$

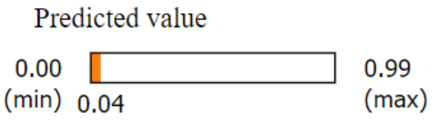
Complex model prediction

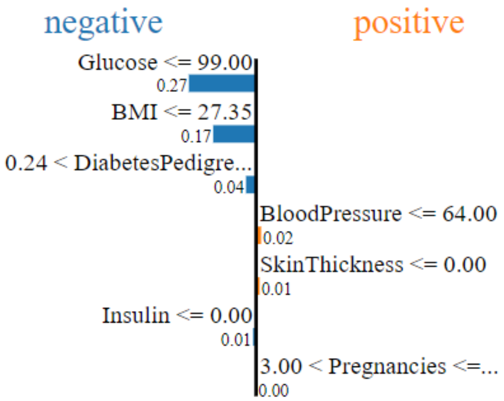
Simple model prediction

(2) $\Omega(g)$ LIME uses sparse linear models (K – LASSO)

LIME Result on Diabetic Database

```
Intercept 0.5191701535536987
Prediction_local [0.0678266]
Right: 0.04
```





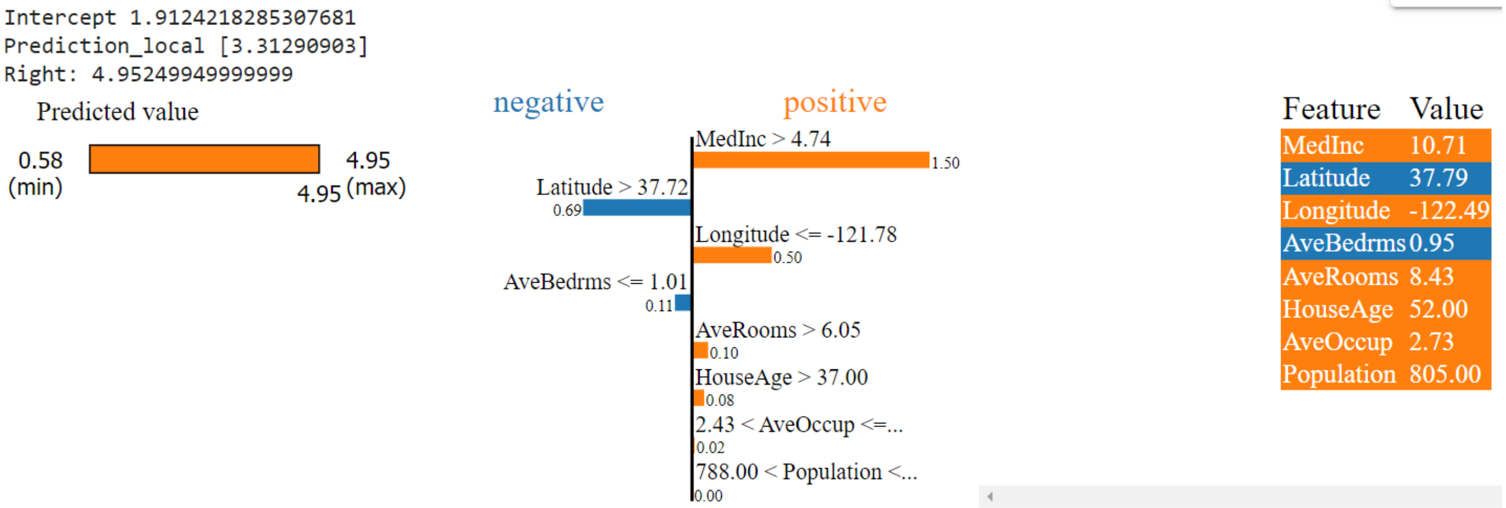
Feature	Value
Glucose	73.00
BMI	26.80
DiabetesPedigreeFunction	0.27
BloodPressure	60.00
SkinThickness	0.00
Insulin	0.00
Pregnancies	5.00

LIME application on tabular data

- *Housing public tabular database*
- *Random forest fit with this database*
- *LIME explanations*

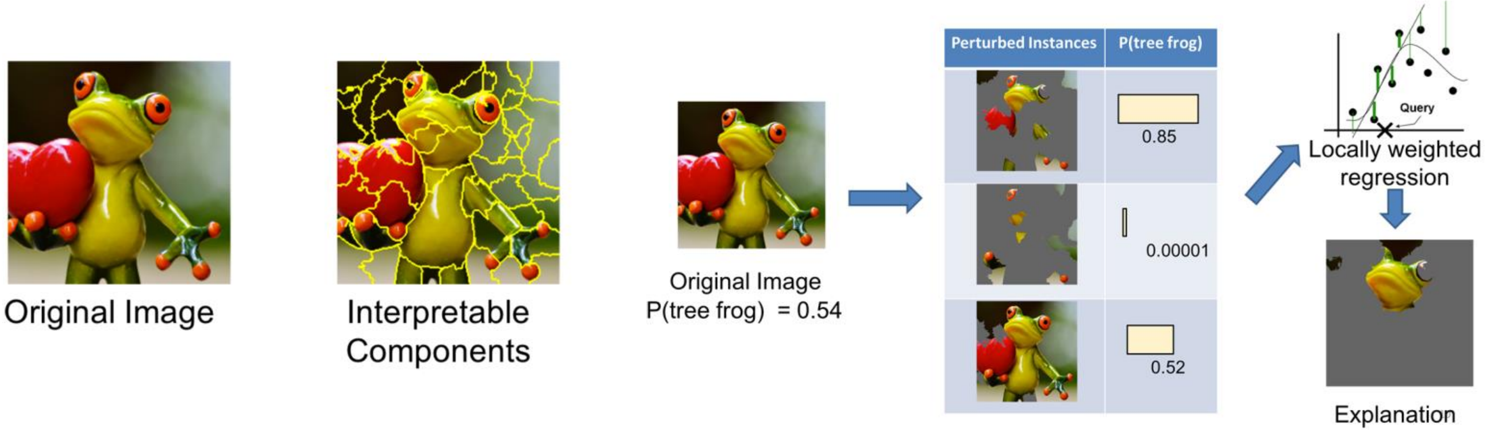
	MedInc	HouseAge	AveRooms	AveBedrms	Population	AveOccup	Latitude	Longitude	label
0	8.3252	41.0	6.984127	1.023810	322.0	2.555556	37.88	-122.23	4.526
1	8.3014	21.0	6.238137	0.971880	2401.0	2.109842	37.86	-122.22	3.585
2	7.2574	52.0	8.288136	1.073446	496.0	2.802260	37.85	-122.24	3.521
3	5.6431	52.0	5.817352	1.073059	558.0	2.547945	37.85	-122.25	3.413
4	3.8462	52.0	6.281853	1.081081	565.0	2.181467	37.85	-122.25	3.422

LIME Result on Housing dataset



- **Right**: prediction given by trees regressor prediction model.
- **Prediction_local:** This denotes the value outputted by a linear model trained on the perturbed samples.

LIME works for image classification

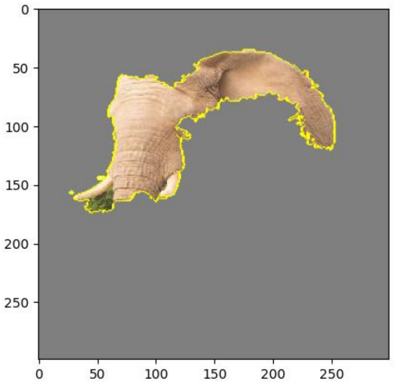


Popular code
https://github.com/marcotcr/lime

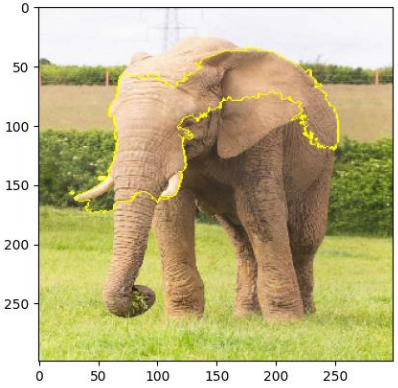
1. Generate a data set of perturbed instances by turning some of the interpretable components “off” (making them gray)
2. For each perturbed instance, we get the probability that a tree frog is in the image according to the model.
3. Learn a simple (linear) model on this data set, which is locally weighted—that is, we care more about making mistakes in perturbed instances that are more similar to the original image.
4. Present the superpixels with highest positive weights as an explanation, graying out everything else.

LIME application on image

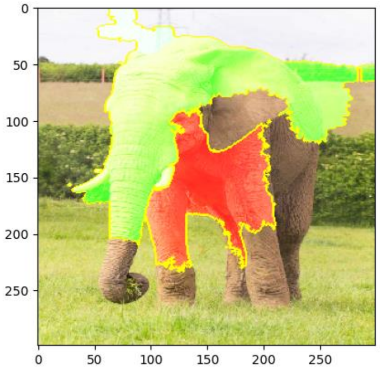
Input image



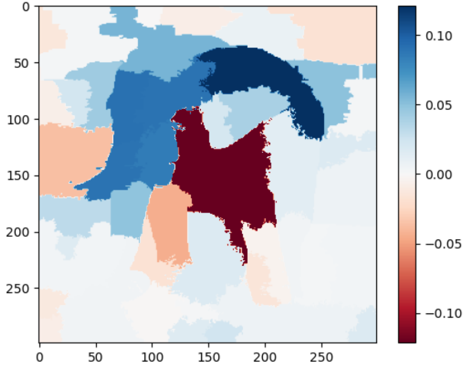
Superpixel for the top most Prediction



Superpixel for the top most Prediction



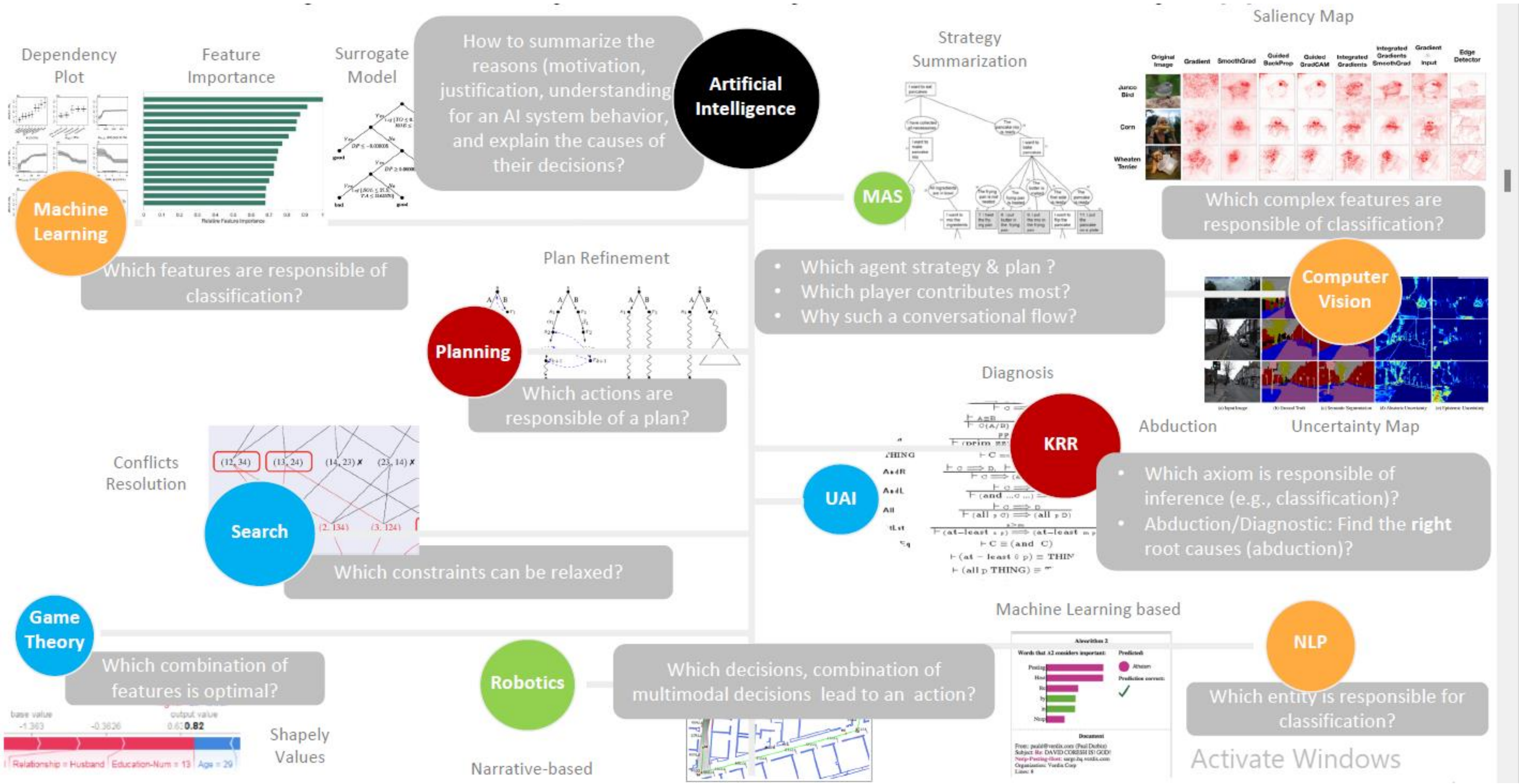
Positive and negative



Heat map

https://github.com/prodramp/DeepWorks/tree/main/MLI-XAI

XAI: One Objective, Many 'AI's, Many Definitions, Many Approaches



"I'M A LITTLE SURPRISED. WITH SUCH EXTENSIVE EXPERIENCE IN PREDICTIVE ANALYSIS, YOU SHOULD'VE KNOWN WE WOULDN'T HIRE YOU."

Dr. Praveen Kumar Chandaliya
Department of Artificial Intelligence
pkc@aid.svnit.ac.in
SVNIT, Surat, India