CONTEXTUALLY

EN**H**ANCED

INTEPRETABLE

LOCAL

EX **L**AINABLE

A**I**

A DATA CONTEXT-AWARE PERTURBATION METHOD FOR XAI

SAIF ANWAR, NATHAN GRIFFITHS, ABHIR BHALERAO, THOMAS POPHAM, MARK BELL

ICML
International Conference
On Machine Learning

WARWICK
THE UNIVERSITY OF WARWICK

KEY DEFINITIONS:

Explainability: Providing evidence or reasoning for all model outputs via an explanation

Interpretability: All explanations must be understandable to end users

Faithfulness: A measure of how accurately an explanation reflects the behaviour of an AI sytem

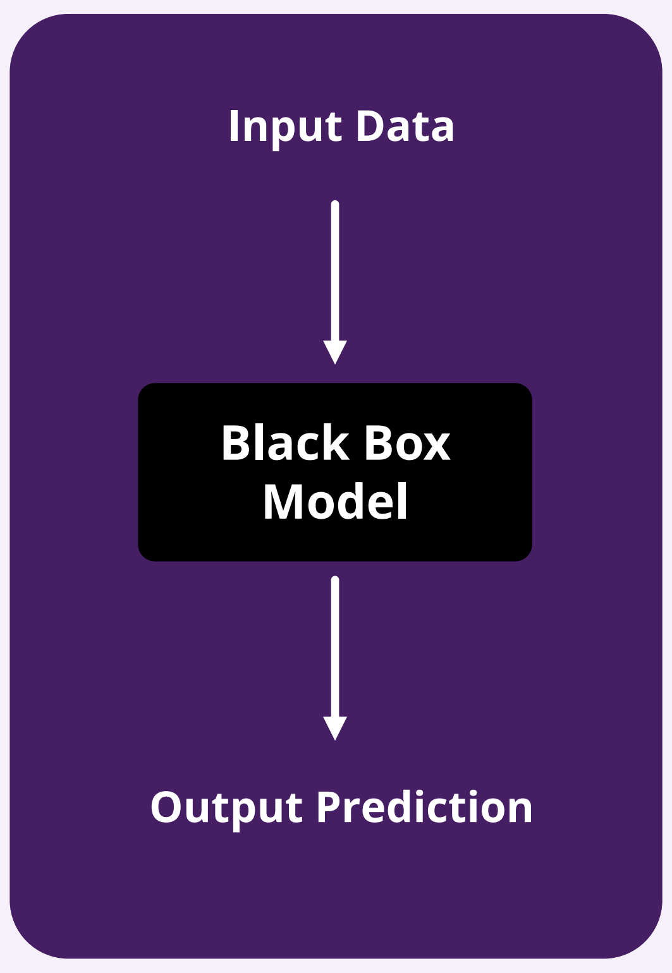
1. WHY DO WE NEED EXPLAINABLE AI?

Some fields have high associated risk

Need to be able to trust decisions in these scenarios

Many ML methods are very complex

We don't understand their inner workings



XAI methods aim to increase confidence in AI

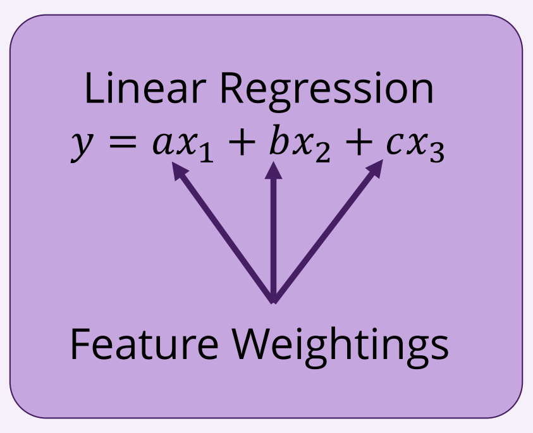
Understanding model reasoning for a prediction

Assess vulnerabilities of a model

Ensure fulfilment of societal and regulatory standards

2. TYPES OF EXPLANATIONS

INHERENTLY INTERPRETABLE MODELS



Some ML models are inherently interpretable:

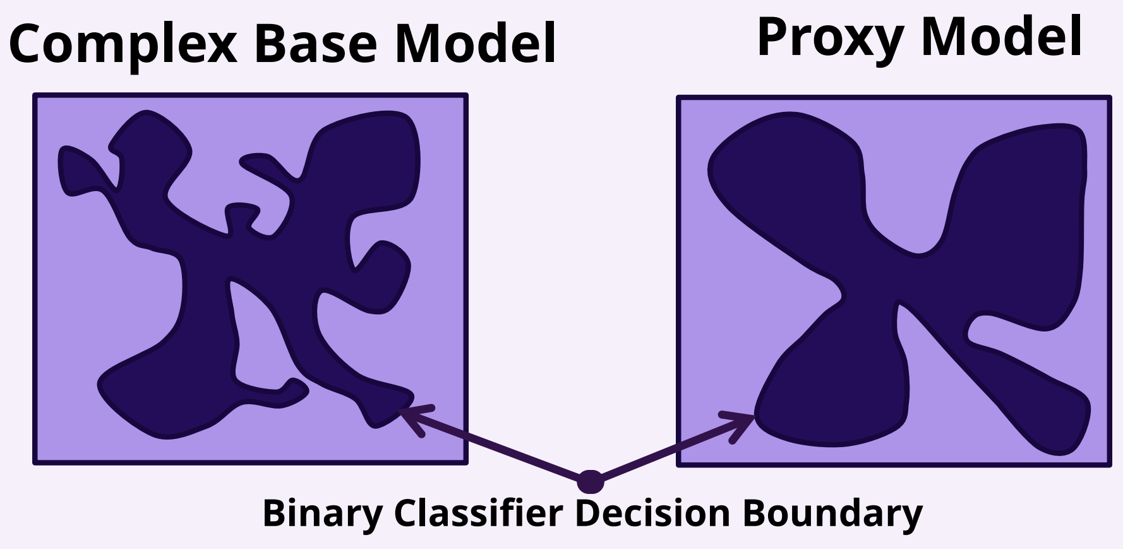
- Linear regression where feature coefficients can be observed
- Decision trees where the decision path can be traced

In these cases, an explanation is the base model itself, which is completely faithful to its own behaviour.

MODEL AGNOSTIC METHODS

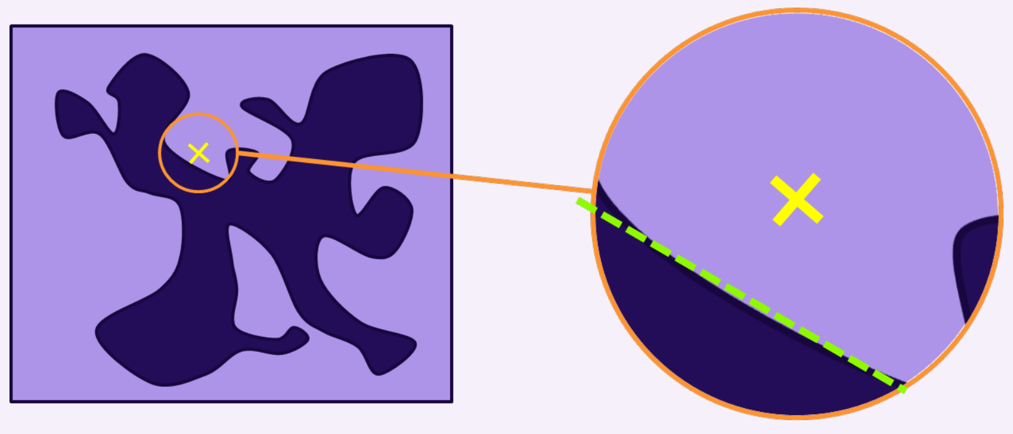
Model-agnostic methods only use input and output data to generate explanations for any type of model.

Proxy models attempt to simplify the behaviour of a complex base model using an inherently interpretable model which is used as an explanation.



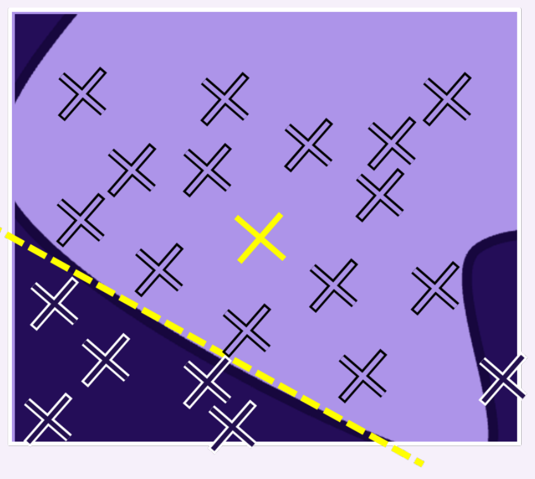
LOCAL PROXY MODELS

- Global proxy models approximate the entire model using the original training data which often leads to oversimplification of the complex model's behaviour
- Local proxy models approximate the behaviour around a specific instance for which the prediction is being explained



There is often not enough training data in a given locality to fit a suitable proxy model.

- Proxy models are instead fit to perturbations of the instance being explained
- The target labels are the base models predictions for these perturbations



The error of the proxy model quantifies the faithfulness towards the base model behaviour.

3. MOTIVATION - LIME

Local Interpretable Model-agnostic Explanations (LIME) is a well-regarded perturbation-based XAI method which fits a linear regression model in the locality of an instance being explained.

PROXIMITY MEASURES

When calculating the faithfulness of an explanation, the contribution of each perturbation, z, towards the error is weighted by its proximity to the instance being explained, $\pi_x(z)$, which is calculated as shown.

$$\pi_x(z) = \exp(\frac{-D^2(x,z)}{\sigma^2})$$

D is a predefined distance function and σ is a locality hyperparameter. In LIME D is Euclidean distance. This is not appropriate for all features, such as:

- Time (Cyclic), e.g. 23:00 -> 01:00 = 2 hours
- Decibels: Logarithmic scale

PERTURBATION GENERATION

- Perturbations are generated by sampling the training data using a Normal distribution and are then scaled around the instance being explained.
- Features are perturbed independently therefore feature dependencies are not considered. For example, assume that as the population of a city grows, traffic congestion increases. Although these features are correlated, ignoring feature dependencies may result in a perturbation combining lowered congestion with increased population.

A lack of context regarding data features can lead to unrealistic perturbations and proximity measures which leads to unreliable explanations.

4. METHOD - CHILLI

We propose CHILLI, a data-context aware XAI framework for generating locally faithful explanations.

PROXIMITY MEASURES

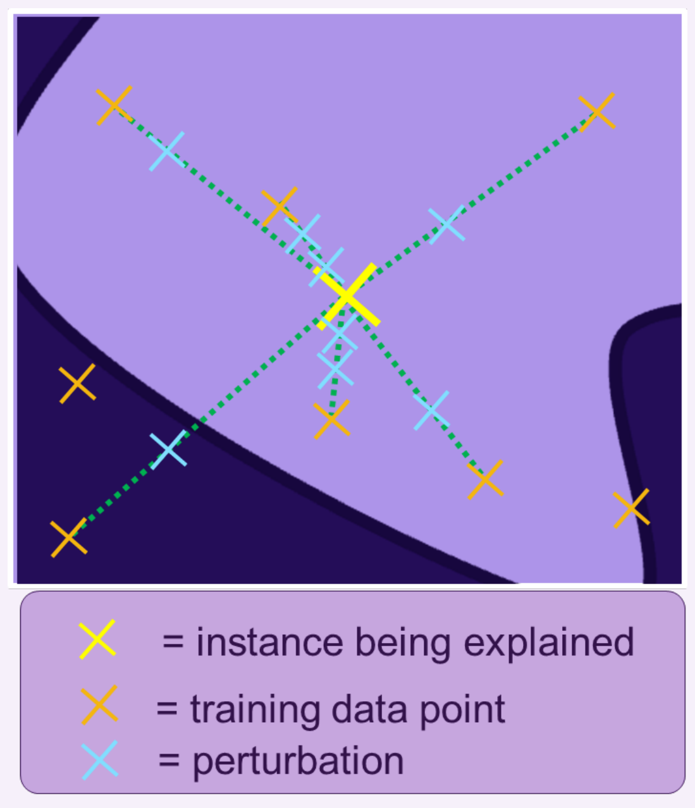
Each feature, i, is given a contextually appropriate distance function, D_i. When calculating the proximity between two points, the distance is calculated independently in each feature dimension and then aggregated as shown below.

$$\pi_x(z) = \exp(\frac{-\frac{1}{d}(\sum_{i \in d} D_i(x_i, z_i))^{\,2}}{\sigma^2})$$

PERTURBATION GENERATION

We generate perturbations using a novel method inspired by Synthetic Minority Oversampling Technique (SMOTE). The algorithm is outlined below.

1. Select a datapoint, x', from the training data where the probability of, x', being selected is directly proportional to its proximity from the instance, x, being explained
2. Select a random value, I, between 0 and 1
3. Linearly interpolate between x and x' by a factor of I in all feature dimensions to generate a single perturbation
4. Repeat 1-3 for the desired number of perturbations.



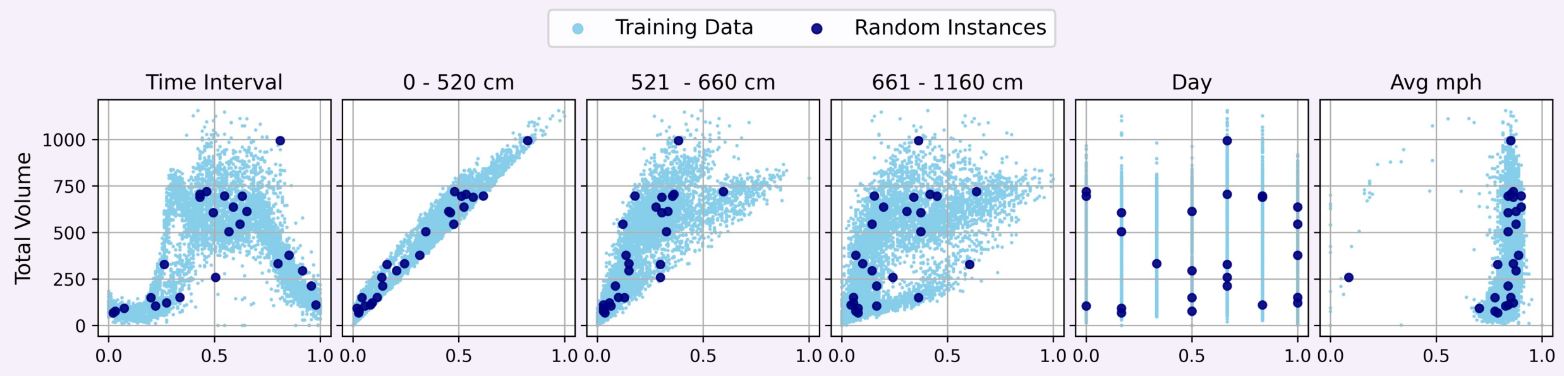
× = instance being explained
× = training data point
× = perturbation

Such a perturbation generation method ensures feature dependency within the training data is maintained and all perturbations remain within appropriate contextual bounds.

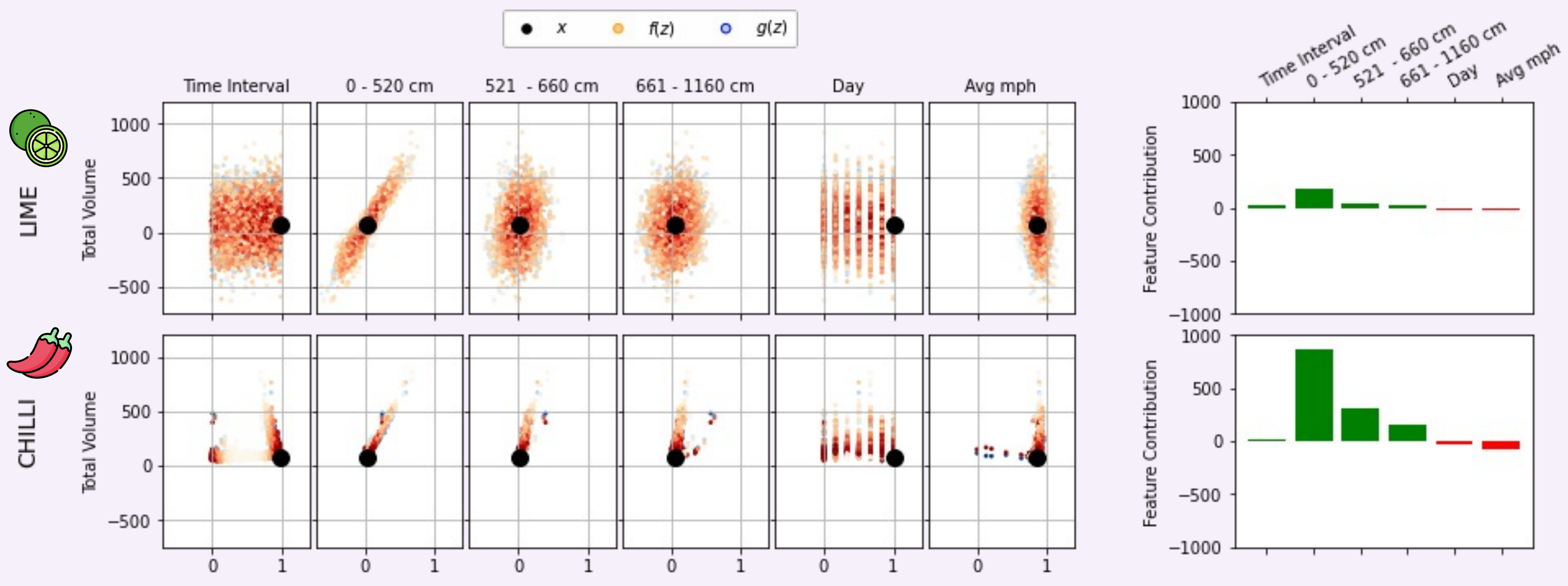
5. EXPERIMENTS & RESULTS

We compare the performance of CHILLI and LIME on the following datasets:

- **MIDAS:** Predicting air temperature at a given location using hourly weather observations from itself and neighbouring locations
- **WebTRIS:** Predicting volume of traffic flow in a given 15 minute interval using time, day, average speed and numbers of different size vehicles



The WebTRIS data (left) is shown in its individual feature dimensions against the target variable



The above figure shows the perturbations generated by LIME and CHILLI, as well as the respective explanations produced by fitting a linear proxy model to the perturbations.

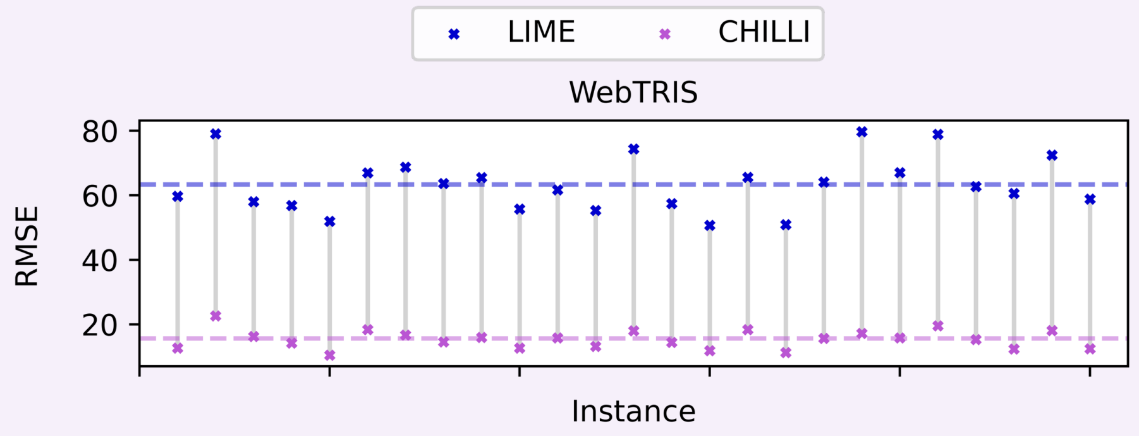
PERTURBATIONS

LIME:

- Fall outside the appropriate bounds and imply impossible values, e.g. negative speed
- Do not represent the training data used for the base model
- Are not local to the instance being explained

CHILLI:

- Are all within the appropriate feature bounds
- Resemble the distribution of the base model's training data for all features
- Are local to the instance being explained



EXPLANATION

- The explanation generated by LIME has much smaller coefficients since the explanation is generalised over the entire feature space
- CHILLI produces an explanation with much stronger coefficients and is also based on sound intuition

FAITHFULNESS

- Comparing the explanations produced by LIME and CHILLI over 25 random instances (left), explanations produced by CHILLI achieve a lower error than those produced by LIME in every instance
- An average **75% error reduction** is achieved by CHILLI compared to LIME
- A lower error implies a more faithful and trustworthy explanation

6. CONCLUSION

We explored the effect of incorporating contextual domain knowledge into a model-agnostic perturbation based XAI approach, namely LIME. We proposed novel methods for generating contextually appropriate perturbations and proximity measures to accurately consider feature dependencies and constraints. It was found that a lack of domain knowledge resulted in unrepresentative, generalised and unreliable explanations.

Our novel method, CHILLI, outperformed LIME in all tested instances with intuitively sound explanations resulting in a 75% increase in explanation faithfulness.