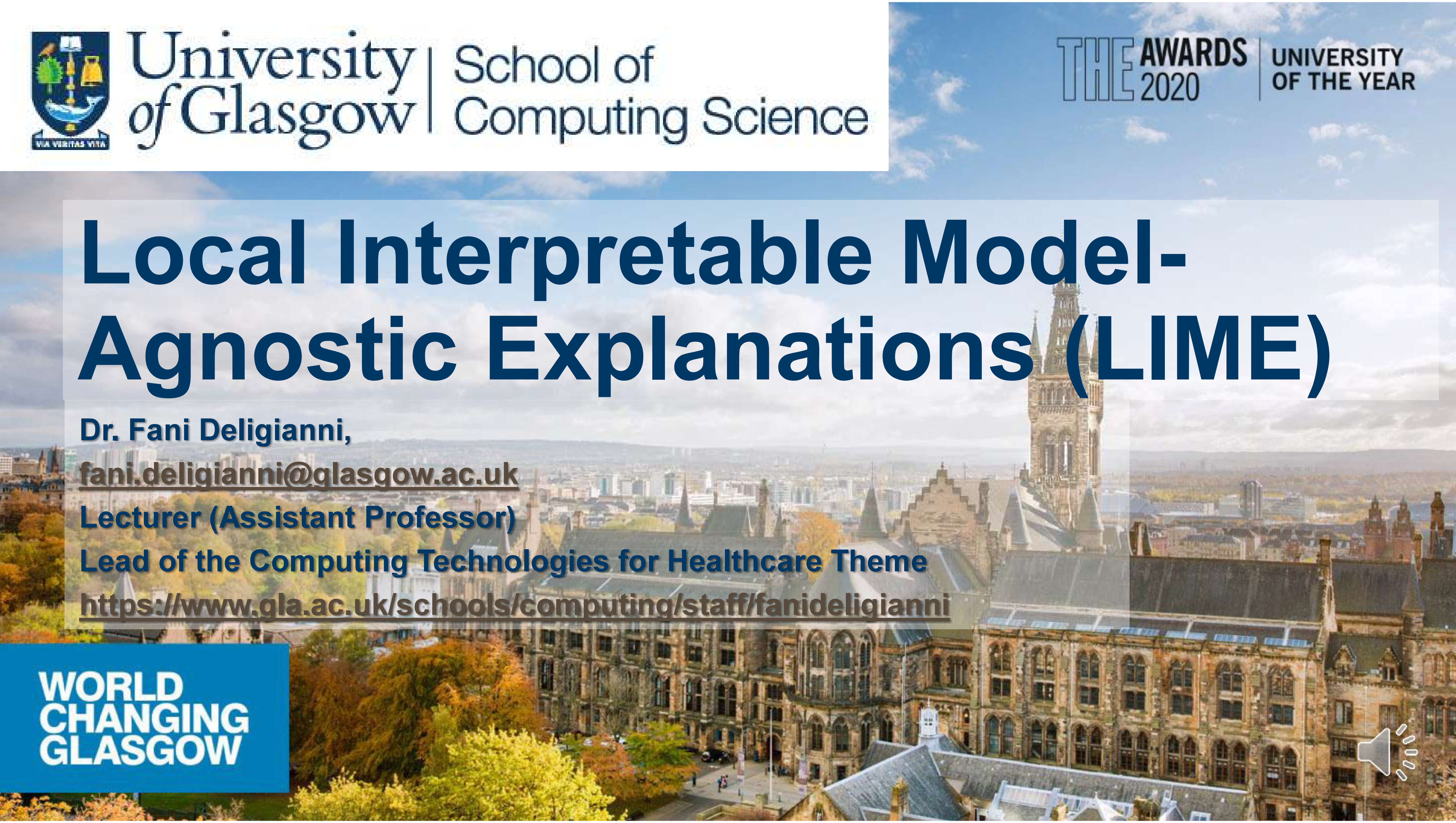
University of Glasgow | School of Computing Science

THE AWARDS 2020 | UNIVERSITY OF THE YEAR



Local Interpretable Model-Agnostic Explanations (LIME)

Dr. Fani Deligianni,

fani.deligianni@glasgow.ac.uk

Lecturer (Assistant Professor)

Lead of the Computing Technologies for Healthcare Theme

https://www.gla.ac.uk/schools/computing/staff/fanideligianni

WORLD CHANGING GLASGOW

Model Agnostic Approaches

- Permutation Feature Importance
- **Local Interpretable Model-agnostic Explanations**
- Shapley Additive Explanations

LIME

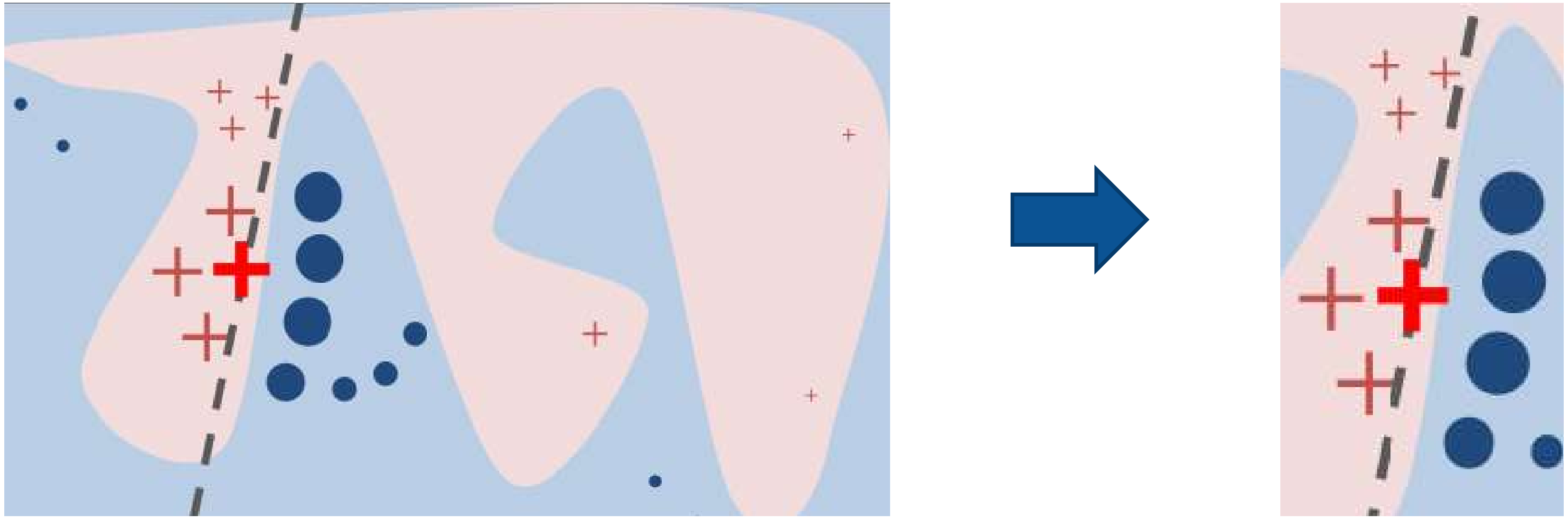
Local Interpretable Model-agnostic Explanations:

- Locally faithful explanations
- Based on a surrogate (locally linear) model

LIME

Local Interpretable Model-agnostic Explanations:

- Locally faithful explanations
- Based on a surrogate (ie. locally linear) model

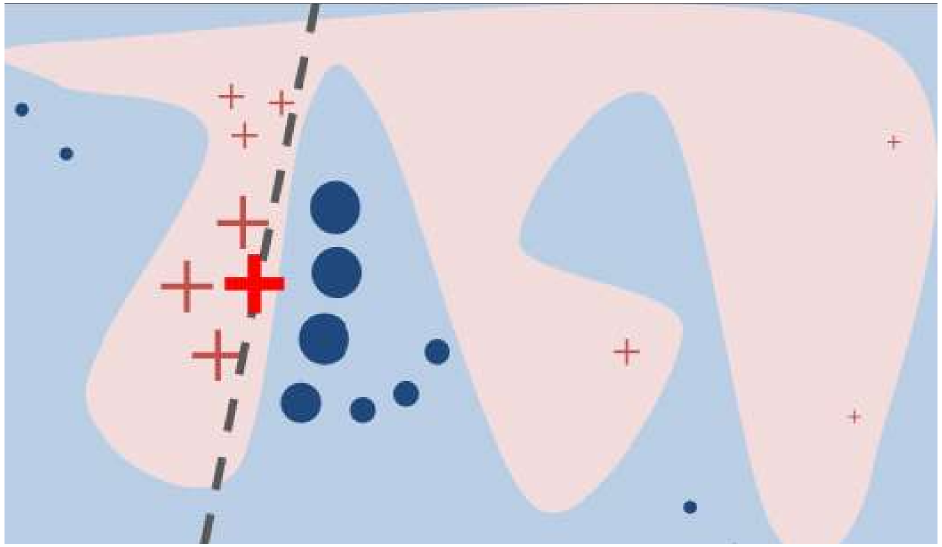


Ribeiro et al. ‘Model-Agnostic Interpretability of Machine Learning’, 2016

LIME - Formulation

Local Interpretable Model-agnostic Explanations:

- Locally faithful explanations
- Based on a surrogate (ie. locally linear) model



Model to explain

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

$$\xi(x) = \underset{g \in G}{\operatorname{argmin}} \; \mathcal{L}(f, g, \Pi_x) + \Omega(g)$$

Explanation

Proximity Measure

Ribeiro et al. ‘Model-Agnostic Interpretability of Machine Learning’, 2016

LIME – Explanations

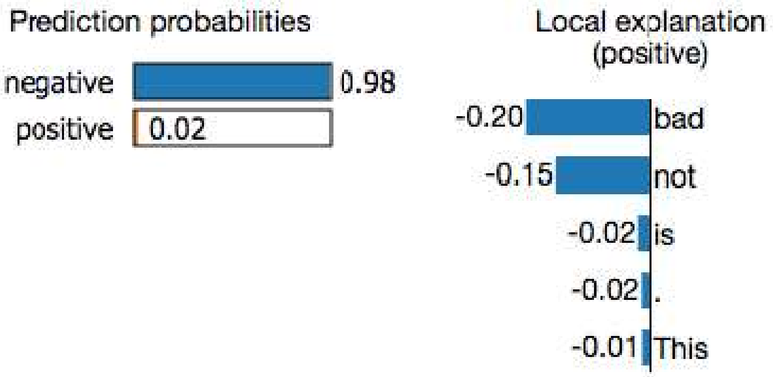
Local Interpretable Model-agnostic Explanations:

- Allow accurate explanations while it retains model flexibility
- The explanation should be accessible even to the non experts
- Small switching costs with relation to changes to the model

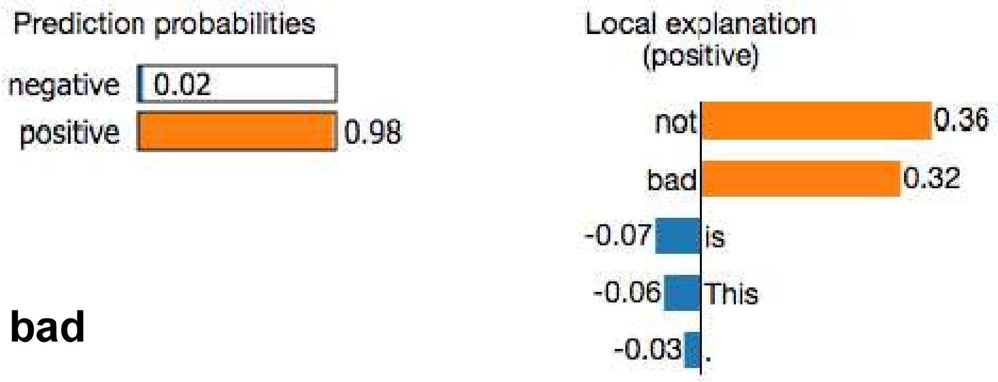
LIME – Explanations

Local Interpretable Model-agnostic Explanations:

- Allow accurate explanations while it retains model flexibility
- The explanation should be accessible even to the non experts
- Small switching costs with relation to changes to the model



This is not too bad



LIME Explanation of a logistic regression

LIME Explanation of an LSTM Model

Summary

- LIME provides a qualitative understanding between input variables and response
- LIME explanations are easily understood by non-experts
- LIME explanations are locally faithful to the model’s behavior
- LIME is a model agnostic approach and thus model switching costs are small

References

- Ribeiro et al. ‘Model-Agnostic Interpretability of Machine Learning’, ICML Workshop on Human Interpretability in Machine Learning, 2016.
- Ribeiro et al. ‘"Why Should I Trust You?": Explaining the Predictions of Any Classifier’, Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016.