

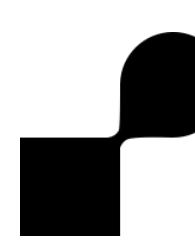


Explanation Method

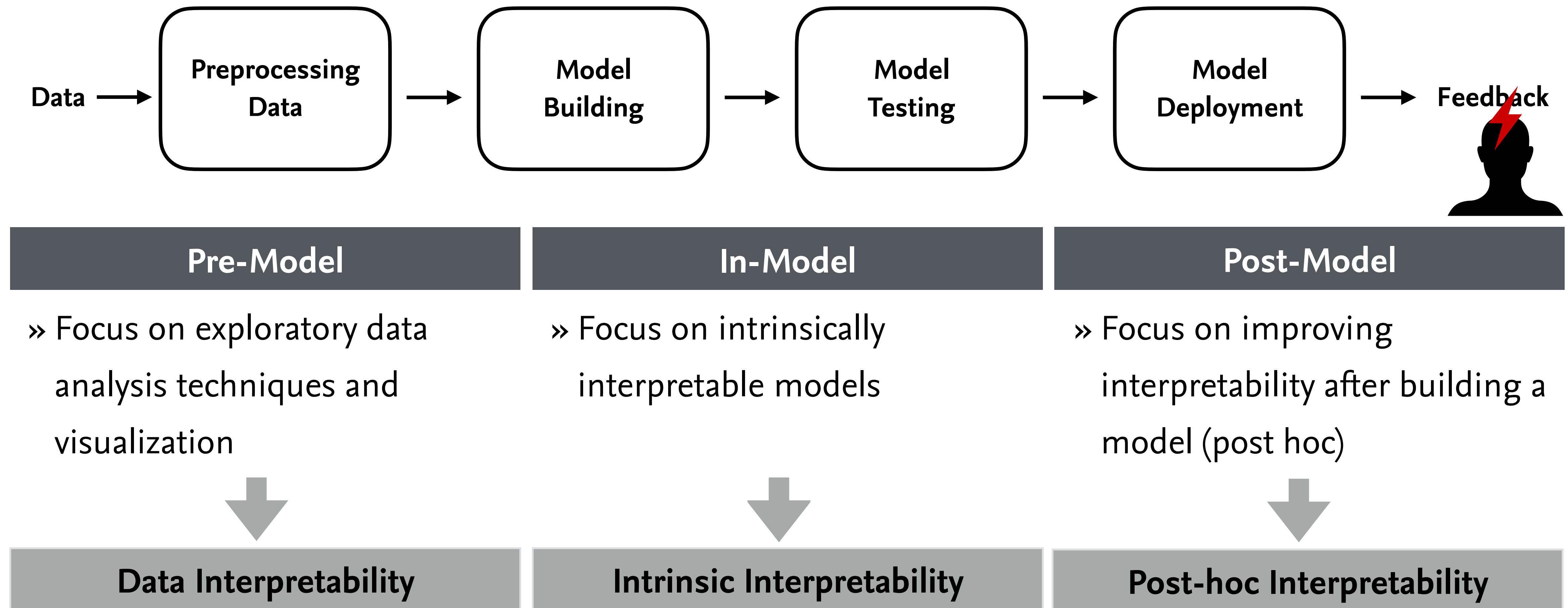
Human-Centered Computing (HCC) Research Group
Institute for Computer Science, Freie Universität Berlin

Lars Sipos

11.02.2022, Introduction to Focus Area

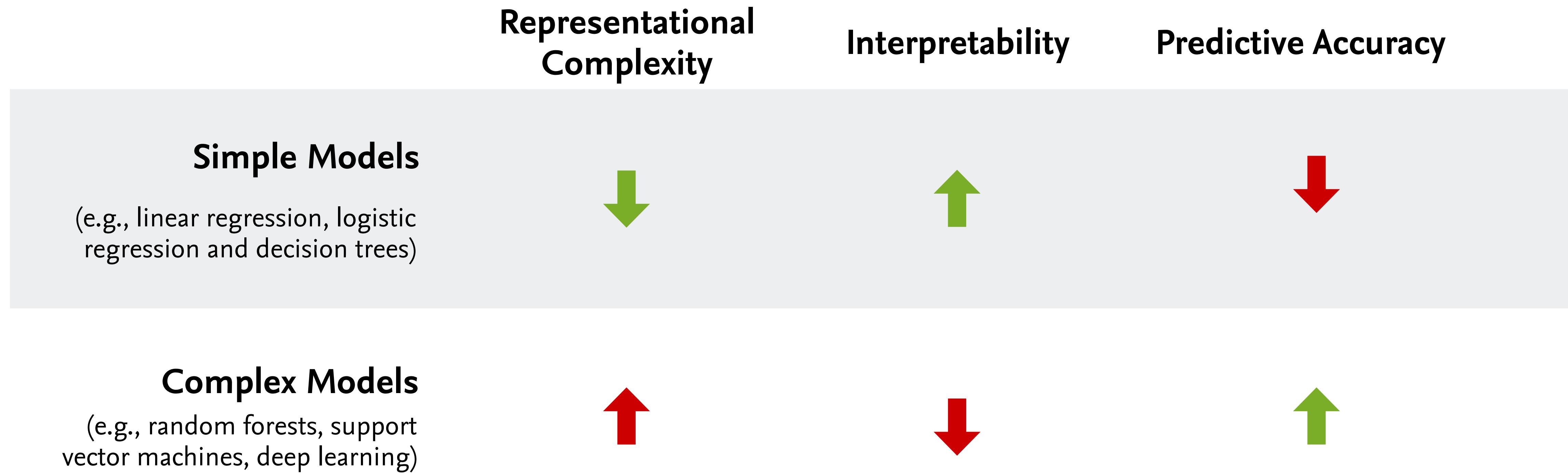


Directions of Interpretability



Carvalho, Diogo V., Eduardo M. Pereira, and Jaime S. Cardoso. "Machine learning interpretability: A survey on methods and metrics." *Electronics* 8.8 (2019): 832.

Challenge of Interpretability vs. Accuracy/Complexity

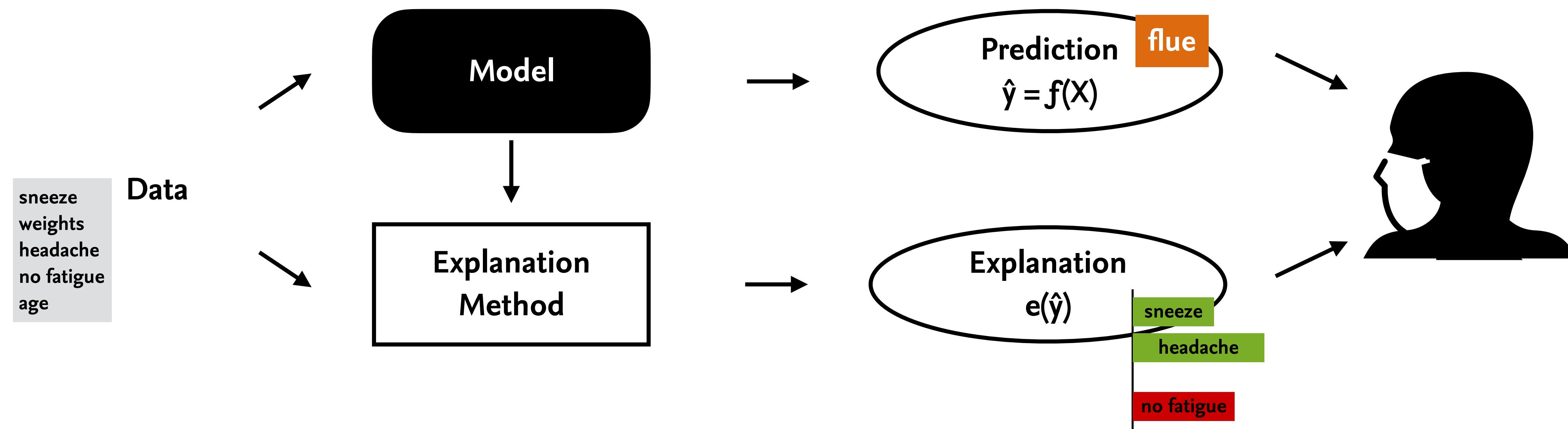


Bamman, D. (2016). Interpretability in human-centered data science. In CSCW Workshop on Human-Centered Data Science. https://cscw2016hcds.files.wordpress.com/2015/10/bamman_hcds.pdf

Martens, D., Vanthienen, J., Verbeke, W., & Baesens, B. (2011). Performance of classification models from a user perspective. Decision Support Systems, 51(4), 782-793.

Explanation Method

An explanation method is a pattern or a mechanisms that generates explanations to establish post-hoc interpretability. An explanation method is based on an explainable AI (XAI) algorithm.



Liao, Q. V., Gruen, D., & Miller, S. (2020, January 8). Questioning the AI: Informing Design Practices for Explainable AI User Experiences. <http://doi.org/10.1145/3313831.3376590>
Image from Carvalho, Diogo V., Eduardo M. Pereira, and Jaime S. Cardoso. "Machine learning interpretability: A survey on methods and metrics." *Electronics* 8.8 (2019): 832.

Local Interpretable Model-Agnostic Explanations (LIME)

Introduced by Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin in 2016.

LIME aims to help users to build trust in a prediction by explaining individual predictions.

The explanation method works on images as well as textual data.

Fast computation but based on heuristics.

Key Ideas:

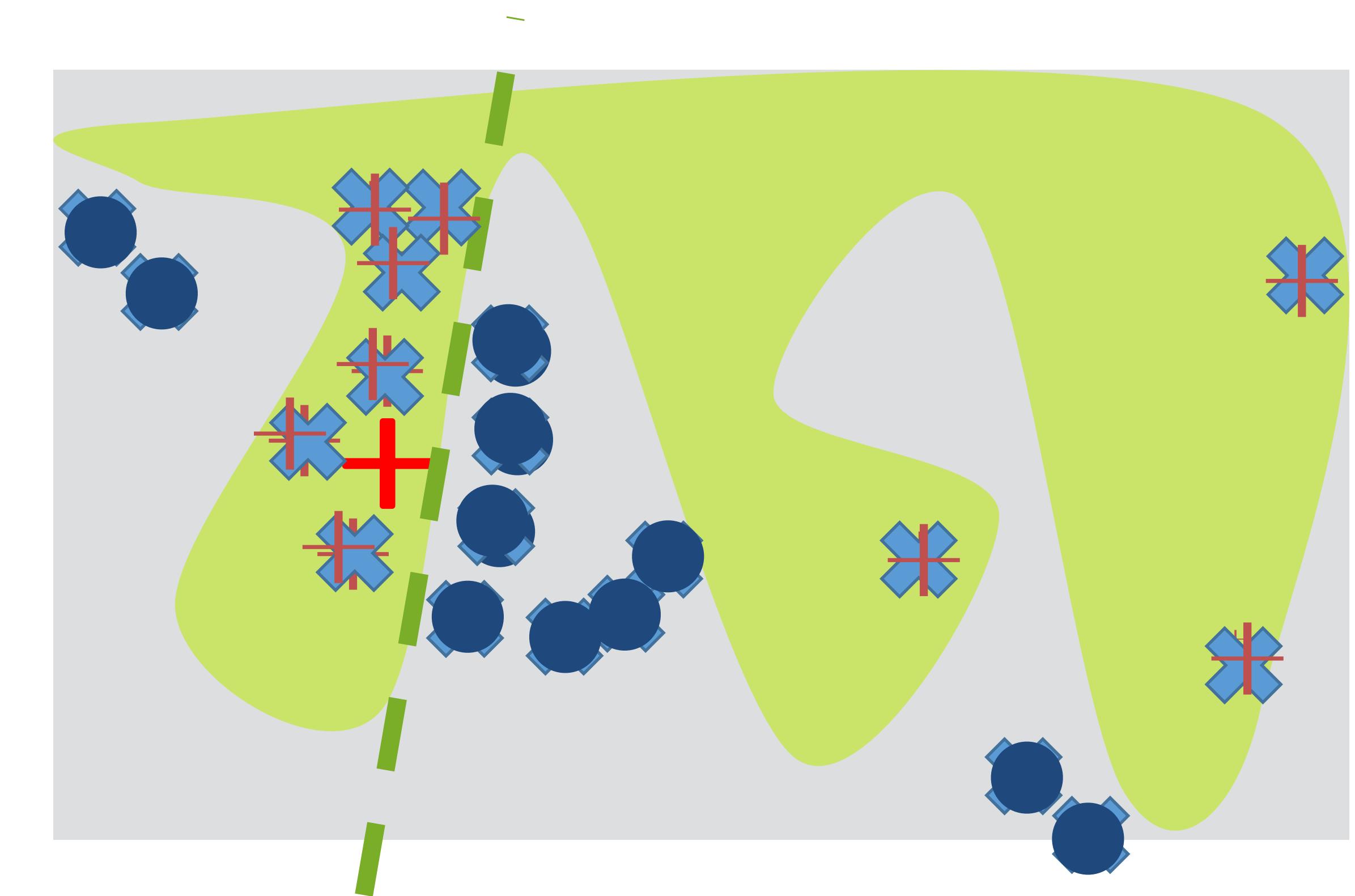
- » Pick a model class interpretable by humans
- » Locally approximate global (blackbox) model



General Approach of LIME

Explain the prediction for input x_i :

1. Sample points around x_i .
2. Use complex model to predict labels for each sample.
3. Weigh samples according to distance to x_i .
4. Learn new simple model on weighted samples.
5. Use simple model to explain.



Slides adapted from Ribeiro, M.T.; Singh, S.; Guestrin, C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier.

Use LIME for Understanding Model Predictions



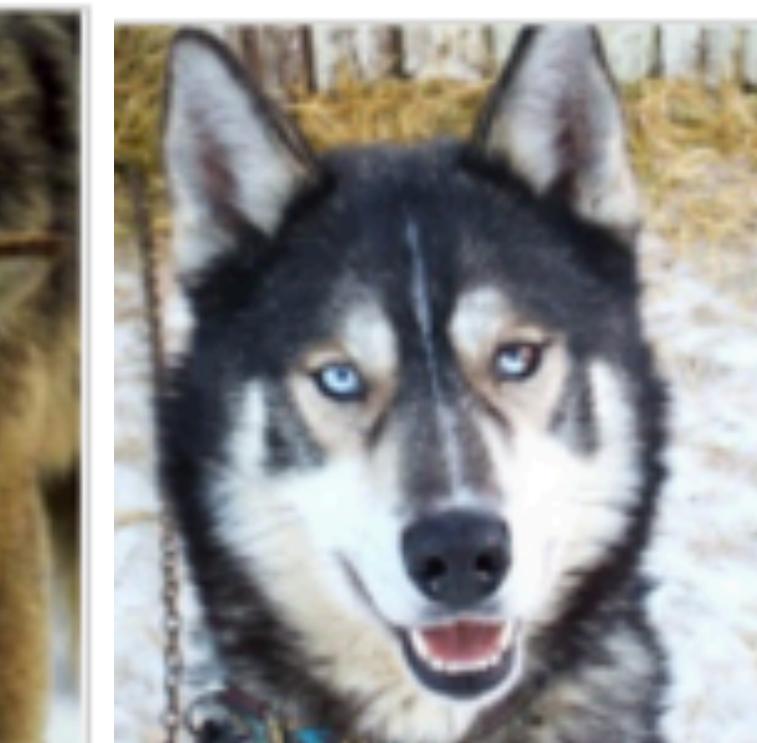
Predicted: **wolf**
True: **wolf**



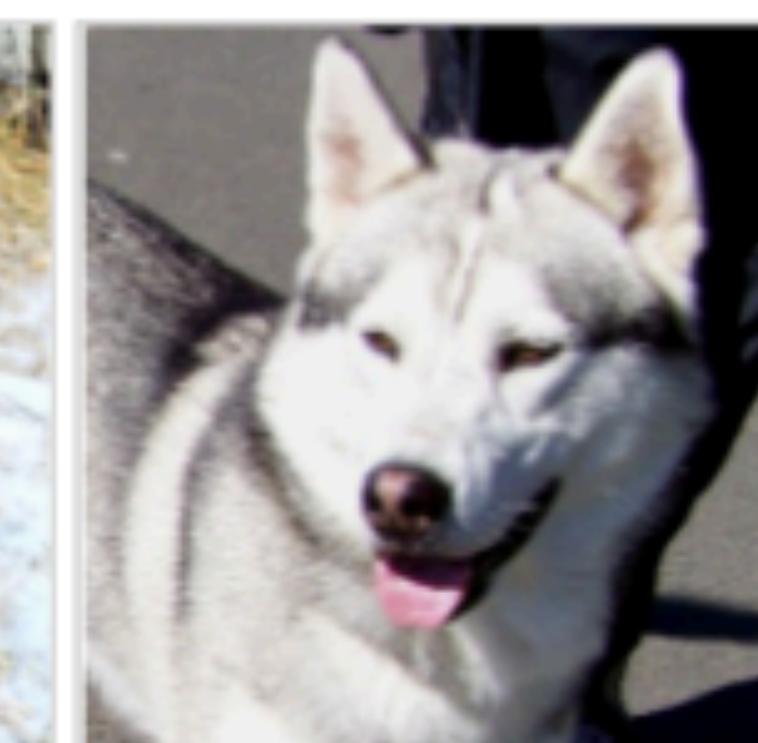
Predicted: **husky**
True: **husky**



Predicted: **wolf**
True: **wolf**



Predicted: **wolf**
True: **husky**



Predicted: **husky**
True: **husky**



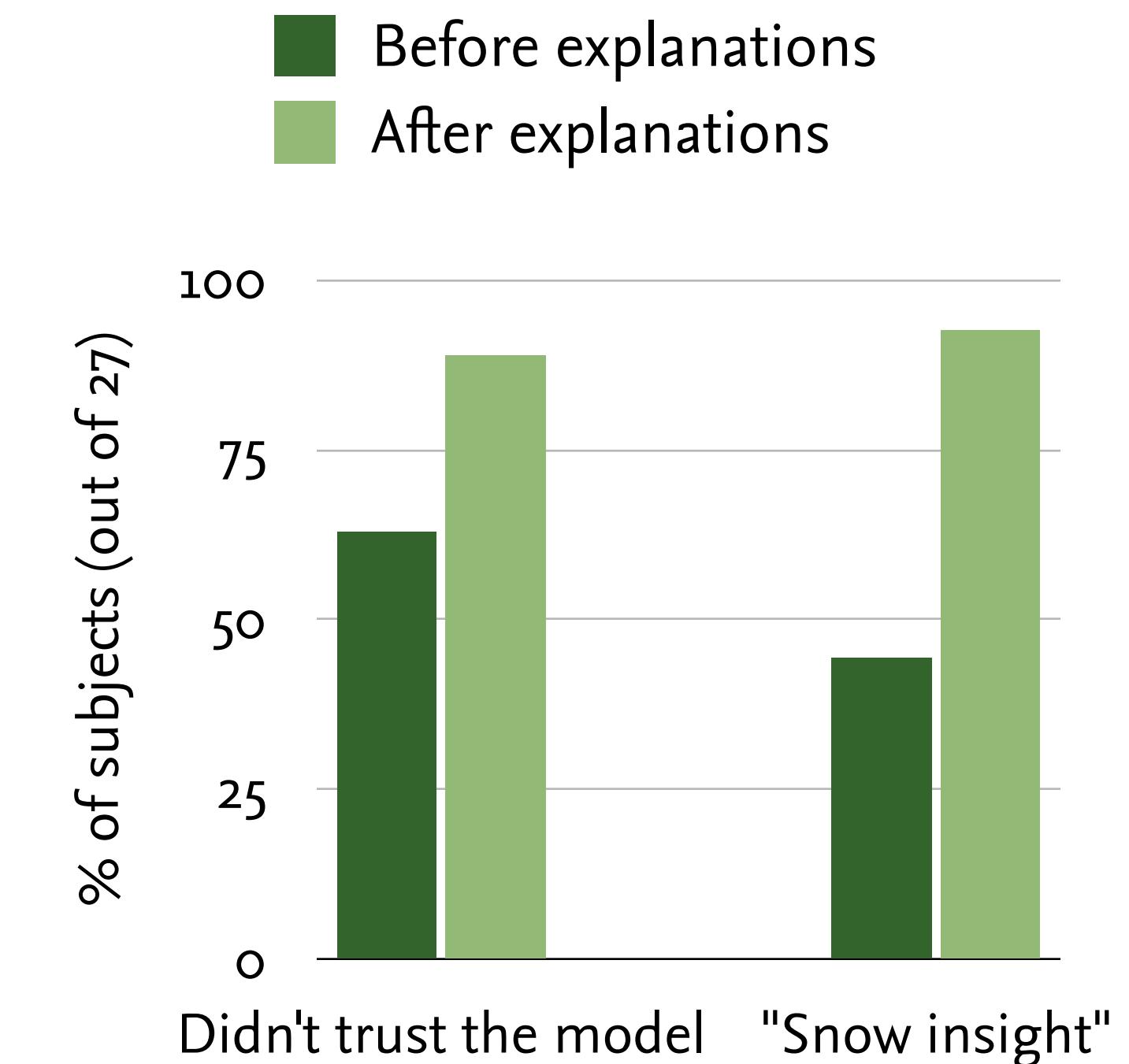
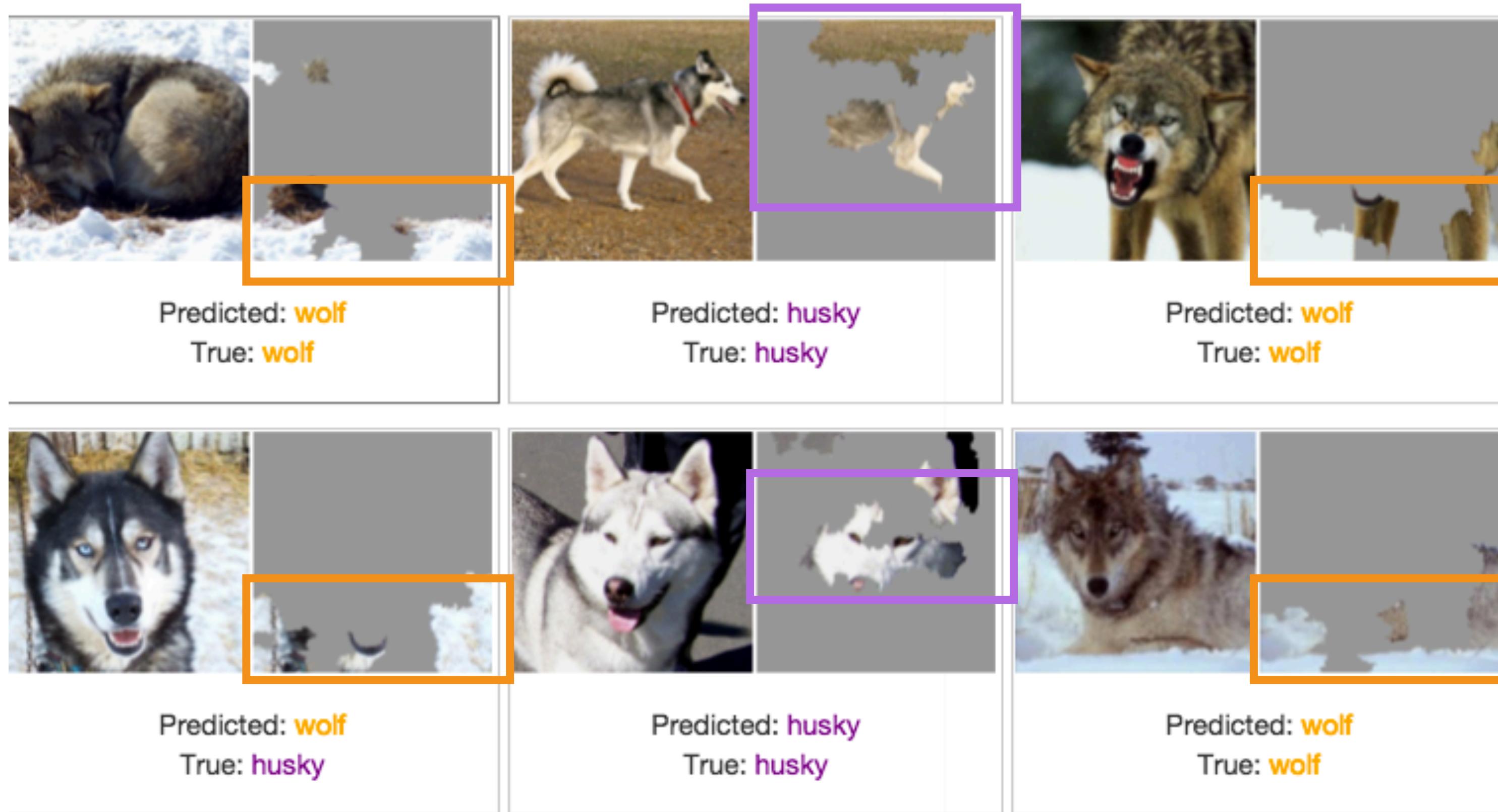
Predicted: **wolf**
True: **wolf**



Only one mistake - do you trust this model?

Slides adapted from Ribeiro, M.T.; Singh, S.; Guestrin, C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier.

Use LIME for understanding Model Predictions (cont.)



It is a snow detector...

Slides adapted from Ribeiro, M.T.; Singh, S.; Guestrin, C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier.

Exercise Task

In this exercise task you will get familiar with LIME.

The exercise can be found on Github:

<https://github.com/FUB-HCC/hcds-intro-to-profile-area-2022>