HELMHOLTZ AI

Introduction to Explainable AI

Helmholtz AI @ HIDA Data Science Courses
06. - 07.05.2025

Who are we & what is our mission?

HELMHOLTZ AI Artificial Intelligence Cooperation Unit

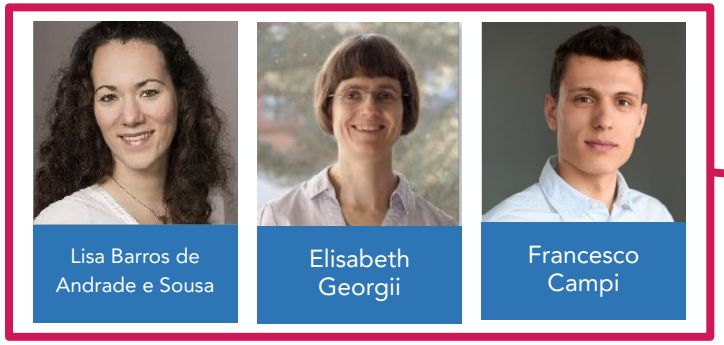
WHAT IS OUR MISSION?

Maximise research impact by democratising access to AI

HOW DO WE DO THAT?

- Short- to mid-term scientific collaboration
- Free of charge
- Easy application

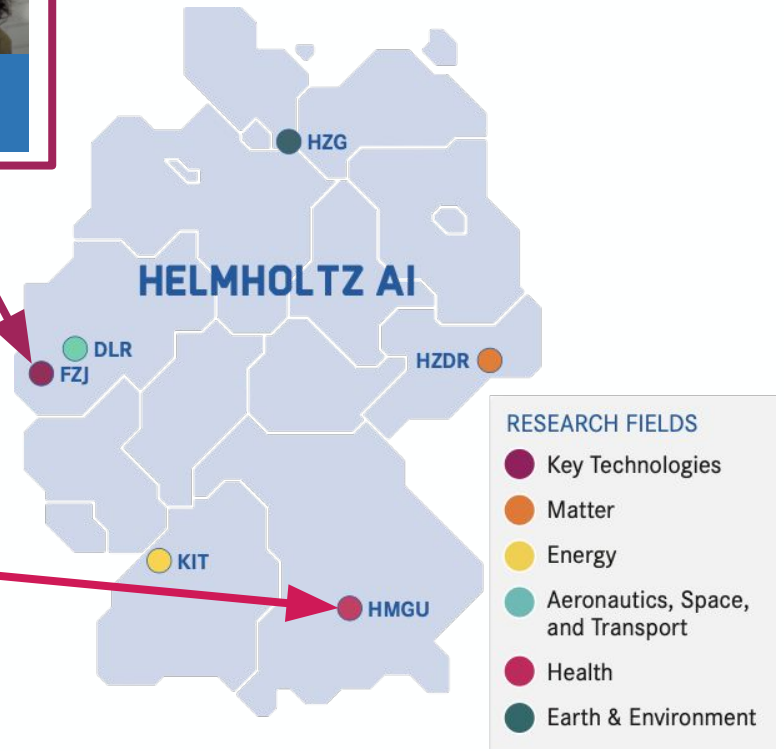
Sabrina Benassou



Lisa Barros de Andrade e Sousa

Elisabeth Georgii

Francesco Campi



Ask your questions in the chat!

Agenda

1. Introduction to eXplainable AI (XAI)
2. Day 1: XAI for Random Forests
   - Model-agnostic methods: Perm. Feature Importance, LIME, SHAP
   - Model-specific methods: Forest-Guided Clustering
   - Comparison of XAI methods for Random Forest models

   Day 2: XAI for CNNs
   - Model-agnostic methods: LIME, SHAP
   - Model-specific methods: Grad-CAM
   - Comparison of XAI methods for CNN models
3. Wrap-Up

GitHub Repository:
https://tinyurl.com/57f25hea

Introduction to eXplainable AI (XAI)

Terminology

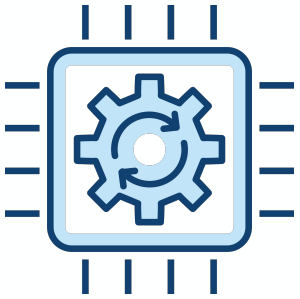
Interpretability or Explainability?

Introduction to eXplainable AI (XAI)

Terminology

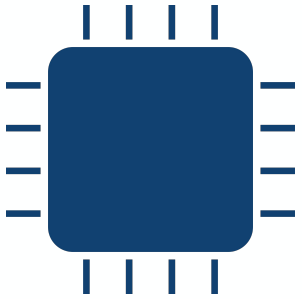
Interpretability

The degree to which a human can understand the internal mechanics of a model without external tools, focussing on the transparency of the model itself.



Explainability

The extent to which the internal mechanics of a machine learning model can be explained in human terms, often using post-hoc methods, focussing on insights into the behavior of black-box models.

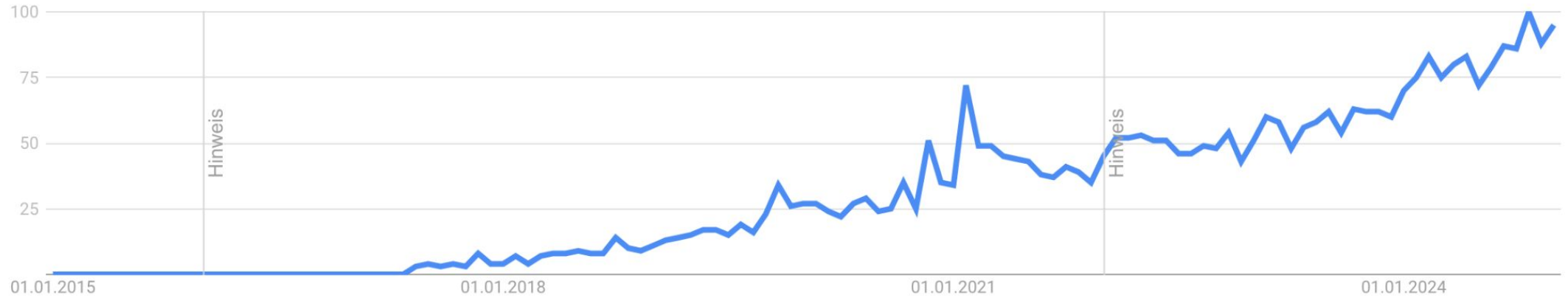


In this course, we will focus only on eXplainable Artificial Intelligence (XAI).

Introduction to eXplainable AI (XAI)

Why is explainability important?

Google Trends Popularity Index of the term "Explainable AI" over the last ten years (2015–2025)



slido

Why is explainability important?

ⓘ Start presenting to display the poll results on this slide.

Introduction to eXplainable AI (XAI)

Why is explainability important?

„The problem is that a single metric, such as classification accuracy, is an incomplete description of most real-world tasks." — (Doshi-Velez et al., 2017)

Introduction to eXplainable AI (XAI)

Why is explainability important? For technological acceptance!

Introduction to eXplainable AI (XAI)

Why is explainability important? To avoid ethical issues!

NEWS | 24 October 2019 | Update 26 October 2019

Millions of black people affected by racial bias in health-care algorithms

Study reveals rampant racism in decision-making software used by US hospitals – and highlights ways to correct it.

Heidi Ledford

Introduction to eXplainable AI (XAI)

Why is explainability important? For knowledge creation!

What Does Deep Learning See? Insights From a Classifier Trained to Predict Contrast Enhancement Phase From CT Images

Kenneth A. Philbrick[1]
Kotaro Yoshida
Dai Inoue
Zeynettin Akkus
Timothy L. Kline
Alexander D. Weston
Panagiotis Korfiatis
Naoki Takahashi
Bradley J. Erickson

OBJECTIVE. Deep learning has shown great promise for improving medical image classification tasks. However, knowing what aspects of an image the deep learning system uses or, in a manner of speaking, sees to make its prediction is difficult.

MATERIALS AND METHODS. Within a radiologic imaging context, we investigated the utility of methods designed to identify features within images on which deep learning activates. In this study, we developed a classifier to identify contrast enhancement phase from whole-slice CT data. We then used this classifier as an easily interpretable system to explore the utility of class activation map (CAMs), gradient-weighted class activation maps (Grad-CAMs), saliency maps, guided backpropagation maps, and the saliency activation map, a novel map reported here, to identify image features the model used when performing prediction.

RESULTS. All techniques identified voxels within imaging that the classifier used. SAMs had greater specificity than did guided backpropagation maps, CAMs, and Grad-CAMs at identifying voxels within imaging that the model used to perform prediction. At shallow network layers, SAMs had greater specificity than Grad-CAMs at identifying input voxels that the layers within the model used to perform prediction.

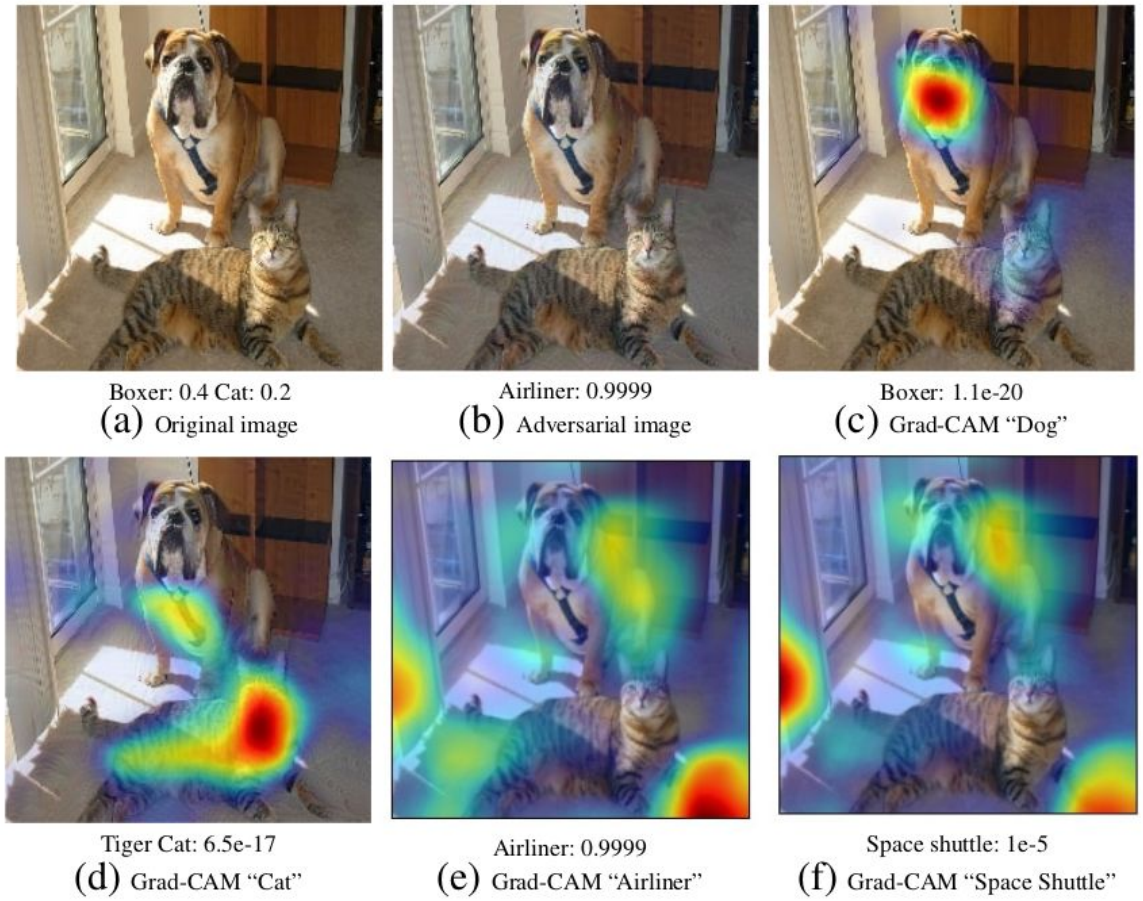
CONCLUSION. As a whole, voxel-level visualizations and visualizations of the imaging features that activate shallow network layers are powerful techniques to identify features that deep learning models use when performing prediction.

Introduction to eXplainable AI (XAI)

Why is explainability important? To meet regulatory requirements!

Introduction to eXplainable AI (XAI)

Why is explainability important? As a defense strategy!



https://campusai.github.io/_papers/Grad-CAM/adversarial.png

Home › Artificial Intelligence and Soft Computing › Conference paper

Explainable AI for Inspecting Adversarial Attacks on Deep Neural Networks

Zuzanna Klawikowska, Agnieszka Mikołajczyk & Michał Grochowski

Conference paper | First Online: 07 October 2020

2252 Accesses | 1 Citations

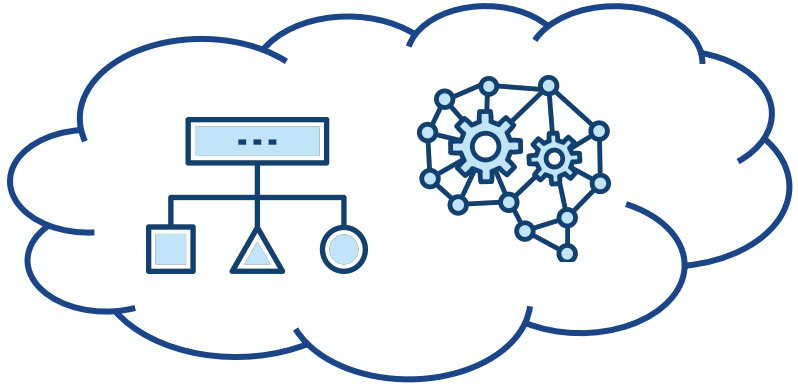
Part of the Lecture Notes in Computer Science book series (LNAI,volume 12415)

Introduction to eXplainable AI (XAI)

Integration of XAI into your Machine Learning workflow

Input Data

ML / DL Model



Output

Explanation

Example: But there were puzzling findings among them, some readers noticed. Although the work identified many human bacteria in cancerous tissues, there was, in addition to the mysterious seaweed bacterium, a marine hydrothermal vent bacterium linked to prostate cancer, and a coral bacterium associated with melanoma. - Science Insider (https://doi.org/10.1126/science.adk1012)

XAI Method

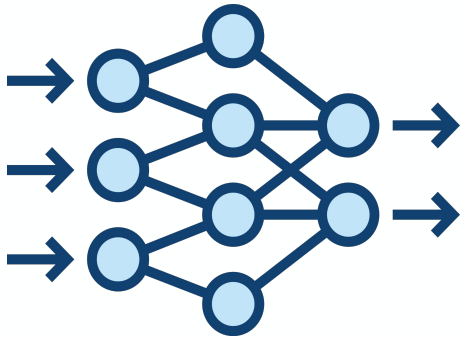
Domai

Introduction to eXplainable AI (XAI)

Integration of XAI into your Machine Learning workflow

input data

ML / DL model



output

Current explanation: This is a cat!

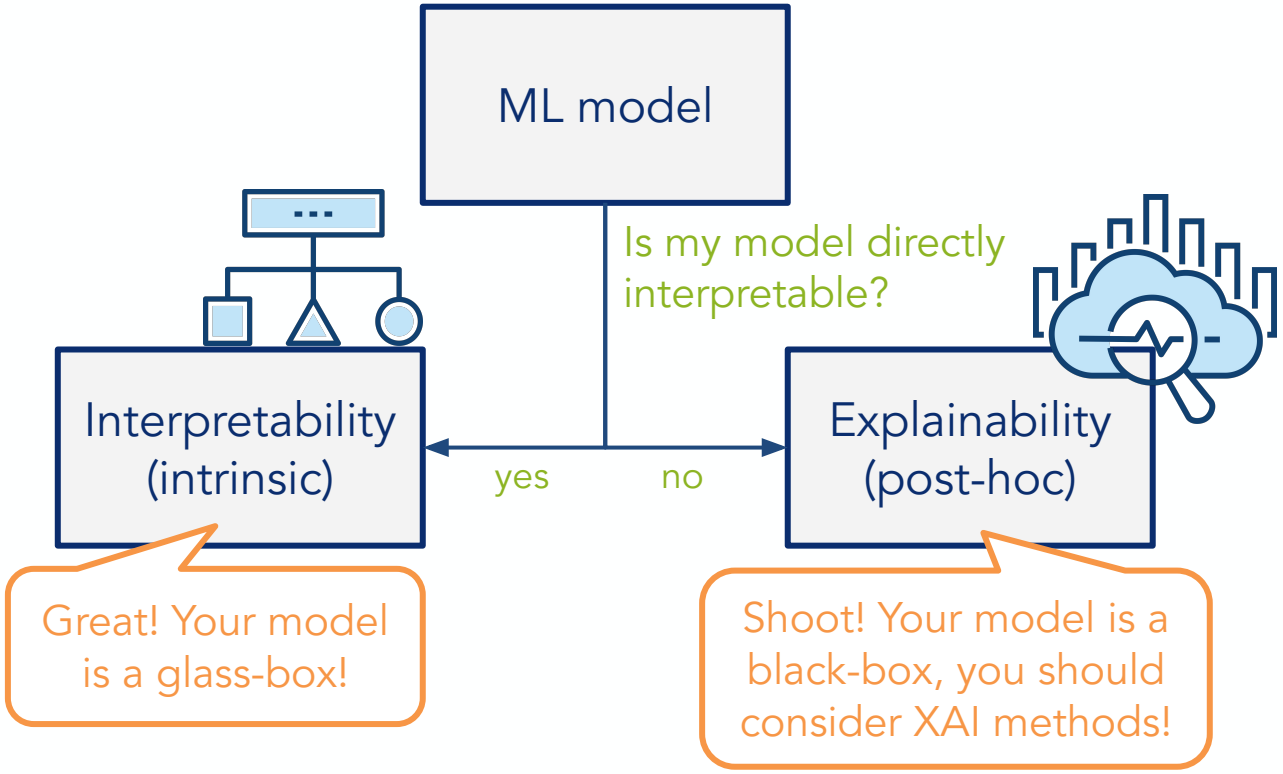
Cat

XAI method + domain knowledge

XAI explanation:
- it has fur, whiskers, and claws
- it has this feature

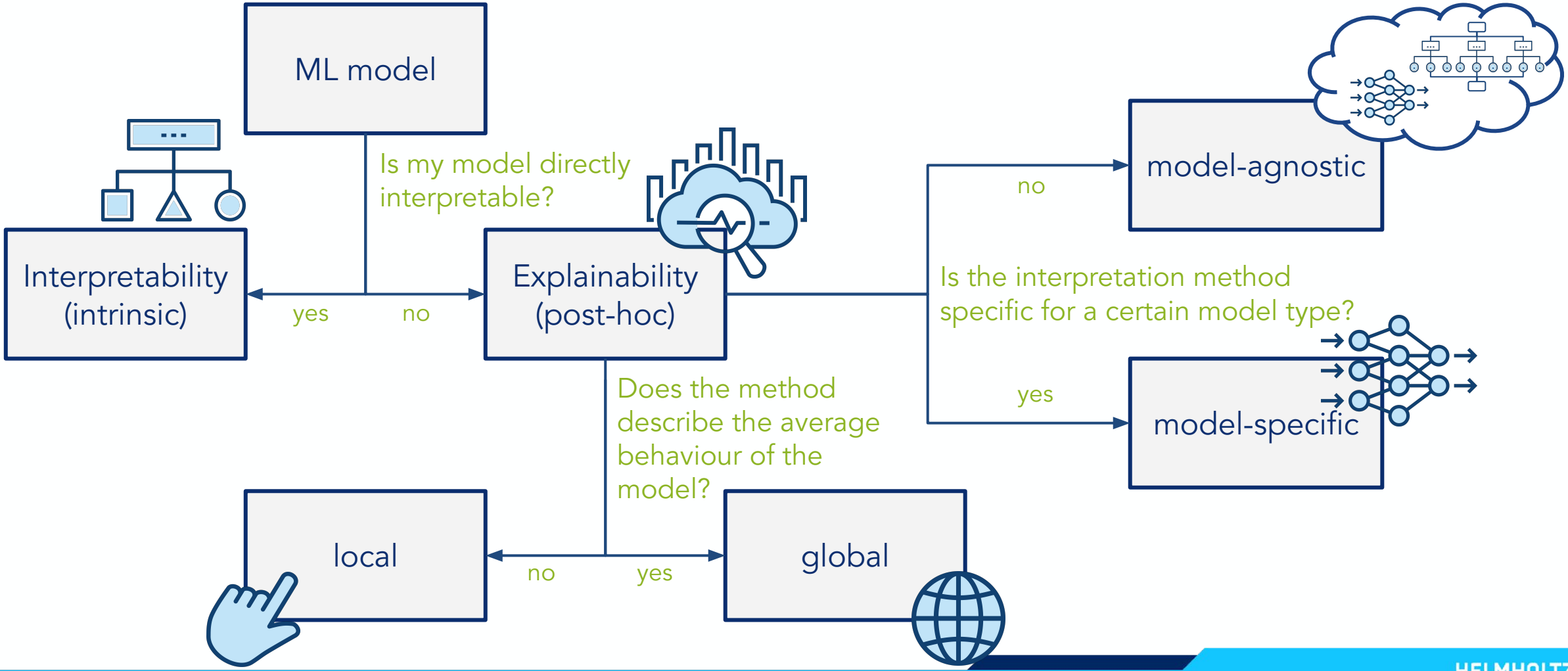
Introduction to eXplainable AI (XAI)

Integration of XAI into your Machine Learning workflow



Introduction to eXplainable AI (XAI)

Integration of XAI into your Machine Learning workflow



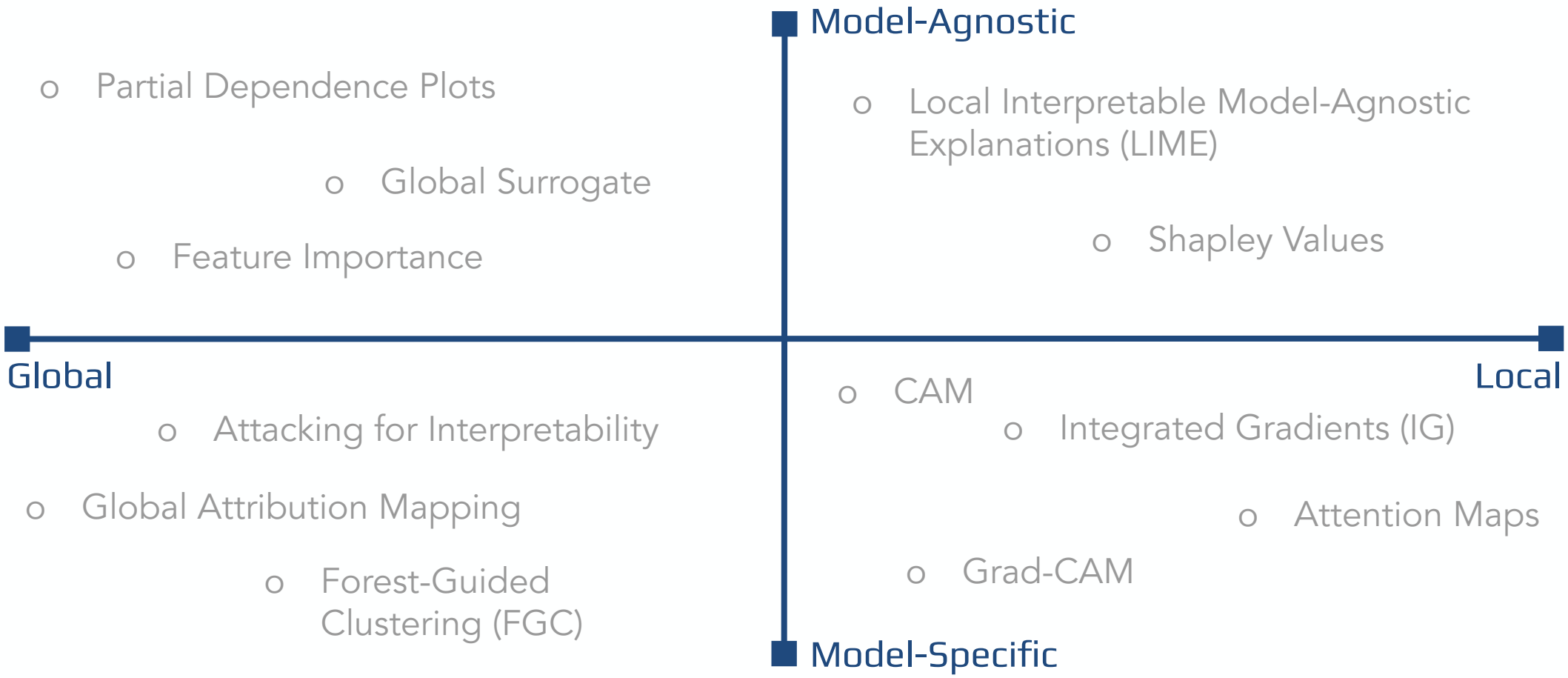
slido

To understand what impact blood pressure has on the survival rate of patient John Doe in a Random Forest model, we need:

ⓘ Start presenting to display the poll results on this slide.

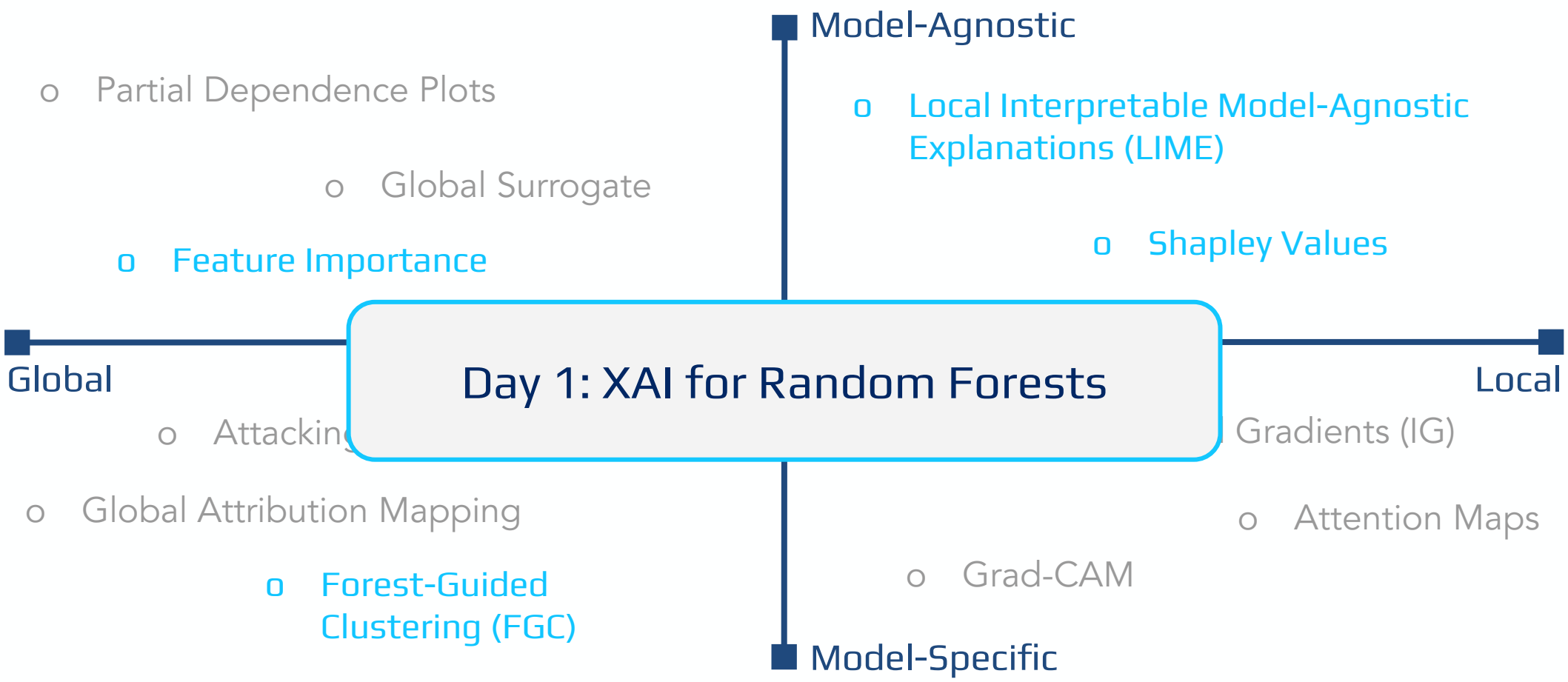
Introduction to eXplainable AI (XAI)

Integration of XAI into your Machine Learning workflow



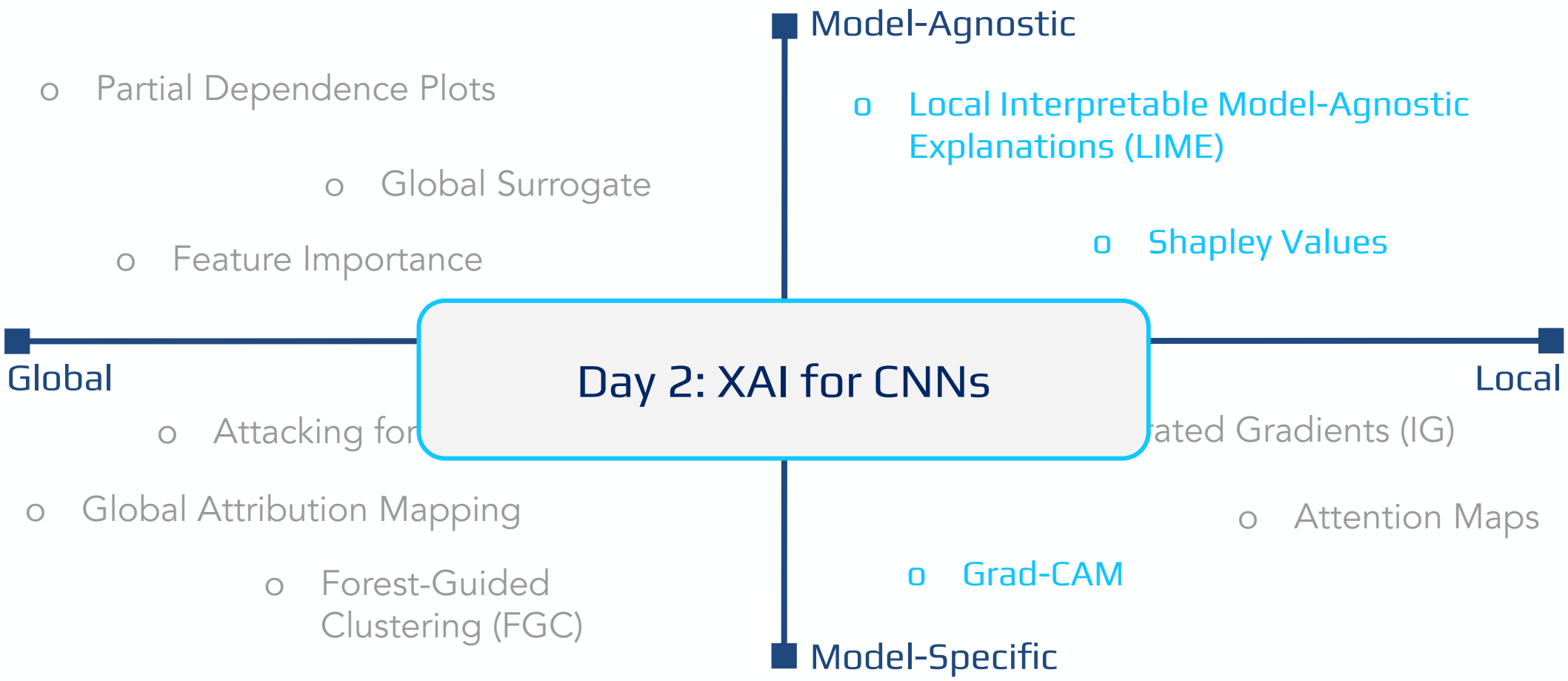
Introduction to eXplainable AI (XAI)

Integration of XAI into your Machine Learning workflow



Introduction to eXplainable AI (XAI)

Integration of XAI into your Machine Learning workflow



You will move now into separate breakout sessions with your tutors!

Agenda

1. Introduction to eXplainable AI (XAI)
2. Day 1: XAI for Random Forests
   - Model-agnostic methods: Perm. Feature Importance, LIME, SHAP
   - Model-specific methods: Forest-Guided Clustering
   - Comparison of XAI methods for Random Forest models

   Day 2: XAI for CNNs
   - Model-agnostic methods: LIME, SHAP
   - Model-specific methods: Grad-CAM
   - Comparison of XAI methods for CNN models
3. Wrap-Up

GitHub Repository:
https://tinyurl.com/57f25hea

Agenda

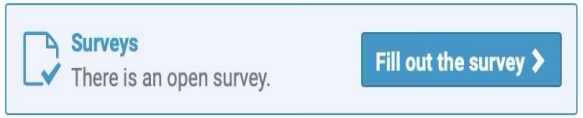
1. Introduction to eXplainable AI (XAI)
2. Track 1: XAI for Random Forests
   - Model-agnostic methods: Perm. Feature Importance, LIME, SHAP
   - Model-specific methods: Forest-Guided Clustering
   - Comparison of XAI methods for Random Forest models

   Track 2: XAI for CNNs
   - Model-agnostic methods: LIME, SHAP
   - Model-specific methods: Grad-CAM
   - Comparison of XAI methods for CNN models
3. Wrap-Up

Wrap-Up

Survey

Survey:



Further Tutorials:

XAI for Transformers

Further Reading