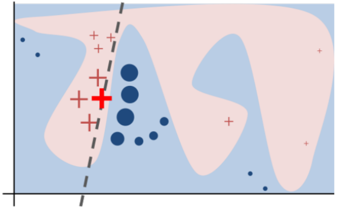
Interpretable Machine Learning

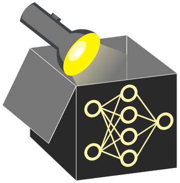
Local Explanations
Introduction



Learning goals

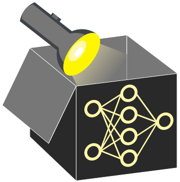
- Understand motivation for local explanations
- Develop an intuition for possible use-cases
- Know characteristics of local explanation methods

METHODOLOGICAL MOTIVATION



- Purpose of local explanations:
  - Insight into the driving factors for a **particular prediction/decision**
  - Understand ML model decisions in a **local neighborhood** of a given input (e.g., feature vector)

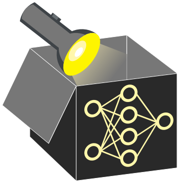
METHODOLOGICAL MOTIVATION



- Purpose of local explanations:
  - Insight into the driving factors for a **particular prediction/decision**
  - Understand ML model decisions in a **local neighborhood** of a given input (e.g., feature vector)

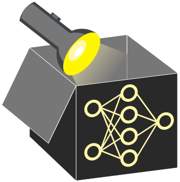
- Local Methods can address questions such as:
  - **Why** did the model decide to predict $\hat{y}$ for input $\mathbf{x}$?
  - **How** does the model behave for observations similar to $\mathbf{x}$?
  - **What if** some features of $\mathbf{x}$ had different values?
  - **Where** (in which regions in $\mathcal{X}$) does the model fail?

SOCIAL MOTIVATION



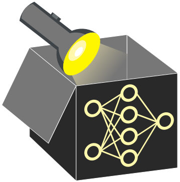
- Explanations for laypersons should be tailored to the **explainee**
  $\rightsquigarrow$ **case specific**, **human-intelligible**, **faithful** to explained mechanism

SOCIAL MOTIVATION



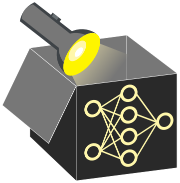
- Explanations for laypersons should be tailored to the **explainee**
  $\rightsquigarrow$ **case specific**, **human-intelligible**, **faithful** to explained mechanism
- If algorithms make decisions in **socially/safety critical domains**, end users have a justified interest in receiving explanations

SOCIAL MOTIVATION



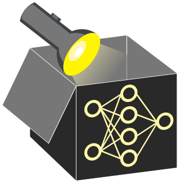
- Explanations for laypersons should be tailored to the **explainee**
  $\rightsquigarrow$ **case specific**, **human-intelligible**, **faithful** to explained mechanism
- If algorithms make decisions in **socially/safety critical domains**, end users have a justified interest in receiving explanations
- Local explanations cannot only increase **user trust**, but also help to detect **critical local biases** in algorithmic decision making

SOCIAL MOTIVATION



- Explanations for laypersons should be tailored to the **explainee**
  $\rightsquigarrow$ **case specific**, **human-intelligible**, **faithful** to explained mechanism
- If algorithms make decisions in **socially/safety critical domains**, end users have a justified interest in receiving explanations
- Local explanations cannot only increase **user trust**, but also help to detect **critical local biases** in algorithmic decision making
- European citizens have the legally binding **right to explanation** as given in the General Data Protection Regulation (GDPR) and the AI Act
  - $\rightsquigarrow$ Instead of explaining the entire (complex) model (with potential market secrets), explanations in a case-by-case usage are more reasonable

GDPR & AI ACT: THE RIGHT TO EXPLANATION



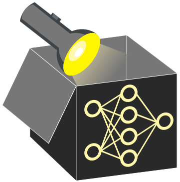
"The data subject should have the right not to be subject to a decision [...] based solely on automated processing [...], such as automatic refusal of an online credit application or e-recruiting practices without any human intervention.
[. . .]
In any case, such processing should be subject to suitable safeguards, which should include [...] the **right to obtain** [...] **an explanation of the decision reached after such assessment and to challenge the decision**."

▸ "Recital 71, GDPR" 2016

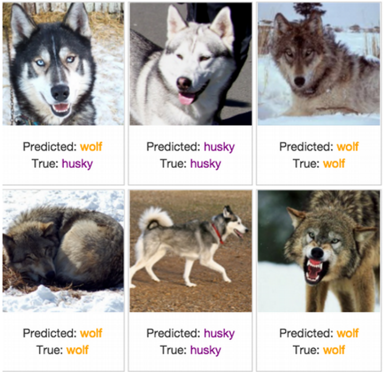
"Any affected person [...] shall have the right to obtain from the deployer clear and meaningful explanations of the role of the AI system in the decision-making procedure and the main elements of the decision taken."

▸ "Art. 86, AI Act" 2021

EXAMPLE: HUSKY OR WOLF?



- We trained a model to predict if an image shows a wolf or a husky
- Below predictions on six test images are given
- Do you trust our predictor?

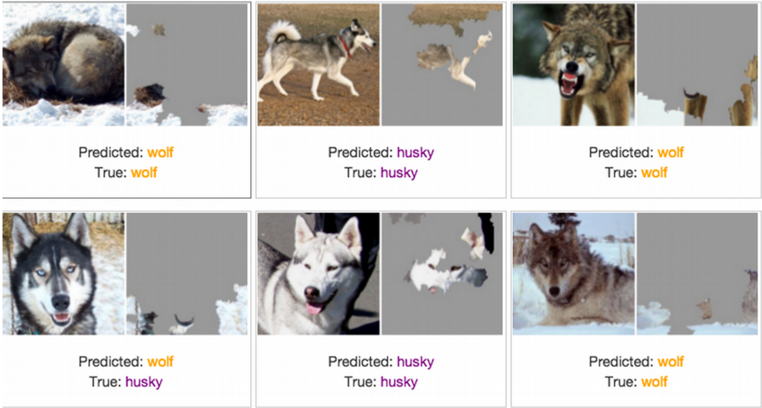


Source: [Sameer Singh 2018]

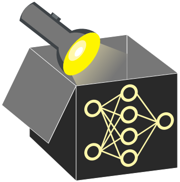
- Sometimes the ML model is wrong
- Can you guess the pattern the ML model learned to identify a wolf?

EXAMPLE: HUSKY OR WOLF? USING LIME

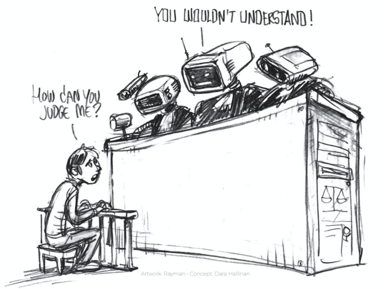
- Local explanations highlight parts of image which led to the prediction
  $\rightsquigarrow$ our predictor is actually a snow detector



Source: [Sameer Singh 2018]



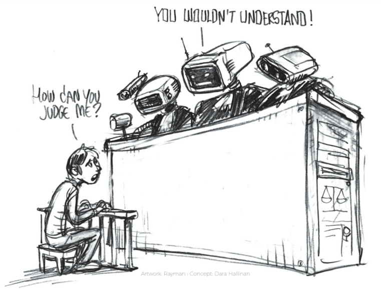
EXAMPLE: LOAN APPLICATION



Source: [https://www.elte.hu]

- Imagine: You apply for a loan at an online bank and are immediately rejected without reasons

EXAMPLE: LOAN APPLICATION

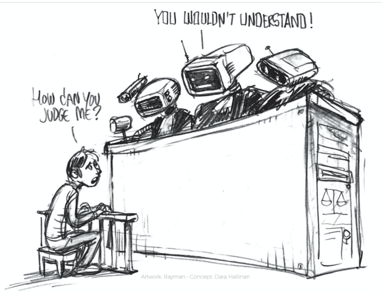


Source: [https://www.elte.hu]

- Imagine: You apply for a loan at an online bank and are immediately rejected without reasons
- Bank could e.g. provide a counterfactual explanation using local explanation methods:

  > “If you were older than 21, your loan application would have been accepted."

EXAMPLE: LOAN APPLICATION

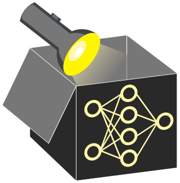


Source: [https://www.elte.hu]

- Imagine: You apply for a loan at an online bank and are immediately rejected without reasons
- Bank could e.g. provide a counterfactual explanation using local explanation methods:

  > “If you were older than 21, your loan application would have been accepted."

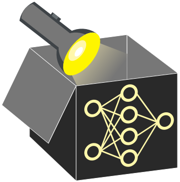
⇝ helps to understand the decision and to take actions for recourse (if req.)



EXAMPLE: STOP OR RIGHT-OF-WAY?

- Imagine:
  - You work at a car company that develops image classifiers for autonomous driving
  - You show your model the following image (an adversarial example)

Source: [Eykholt et. al 2018]

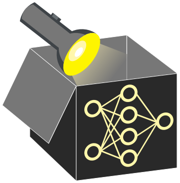


EXAMPLE: STOP OR RIGHT-OF-WAY?

- Imagine:
  - You work at a car company that develops image classifiers for autonomous driving
  - You show your model the following image (an adversarial example)
  - Classifier is 99% sure it describes a right-of-way sign
- Would you entrust other people's lives into the hands of this software?

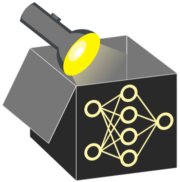
Source: [Eykholt et. al 2018]

CHARACTERISTICS OF LOCAL EXPLANATIONS



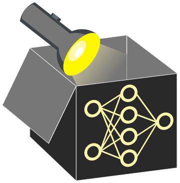
- **Explanation scope:** Specific to one prediction, valid only in local environment

CHARACTERISTICS OF LOCAL EXPLANATIONS



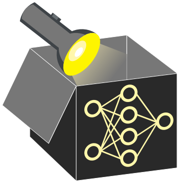
- **Explanation scope:** Specific to one prediction, valid only in local environment
- **Applicable model classes:**
  - Model-agnostic (by design)
  - Model-specific variants (exploit internal structure for speed/accuracy)

CHARACTERISTICS OF LOCAL EXPLANATIONS



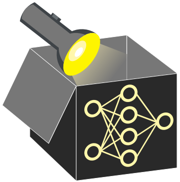
- **Explanation scope:** Specific to one prediction, valid only in local environment
- **Applicable model classes:**
  - Model-agnostic (by design)
  - Model-specific variants (exploit internal structure for speed/accuracy)
- **Audience:** ML engineers, laypersons, and domain experts

CHARACTERISTICS OF LOCAL EXPLANATIONS



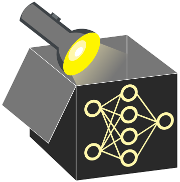
- **Explanation scope:** Specific to one prediction, valid only in local environment
- **Applicable model classes:**
  - Model-agnostic (by design)
  - Model-specific variants (exploit internal structure for speed/accuracy)
- **Audience:** ML engineers, laypersons, and domain experts
- **Supported data types:** Broad applicability across modalities (tabular, image, text, audio), but method-specific adaptations often required

CHARACTERISTICS OF LOCAL EXPLANATIONS



- **Explanation scope:** Specific to one prediction, valid only in local environment
- **Applicable model classes:**
  - Model-agnostic (by design)
  - Model-specific variants (exploit internal structure for speed/accuracy)
- **Audience:** ML engineers, laypersons, and domain experts
- **Supported data types:** Broad applicability across modalities (tabular, image, text, audio), but method-specific adaptations often required
- **Main method families**
  - Single ICE curves
  - Shapley / SHAP values
  - LIME / Anchors
  - Counterfactual explanations
  - Adversarial examples

CREDIT DATASET



- We illustrate local explanation methods on the German credit data
  ▸ "see Kaggle" n.d.
- 522 observations, 9 features containing credit and customer information
- Binary target "risk" indicates if a customer has a 'good' or 'bad' credit risk
- We merged categories with few observations

name	type	range
age	numeric	[19, 75]
sex	factor	{male, female}
job	factor	{0, 1, 2, 3}
housing	factor	{free, own, rent}
saving.accounts	factor	{little, moderate, rich}
checking.accounts	factor	{little, moderate, rich}
credit.amount	numeric	[276, 18424]
duration	numeric	[6, 72]
purpose	numeric	{others, car, furniture, radio/TV}
risk	factor	{good, bad}