

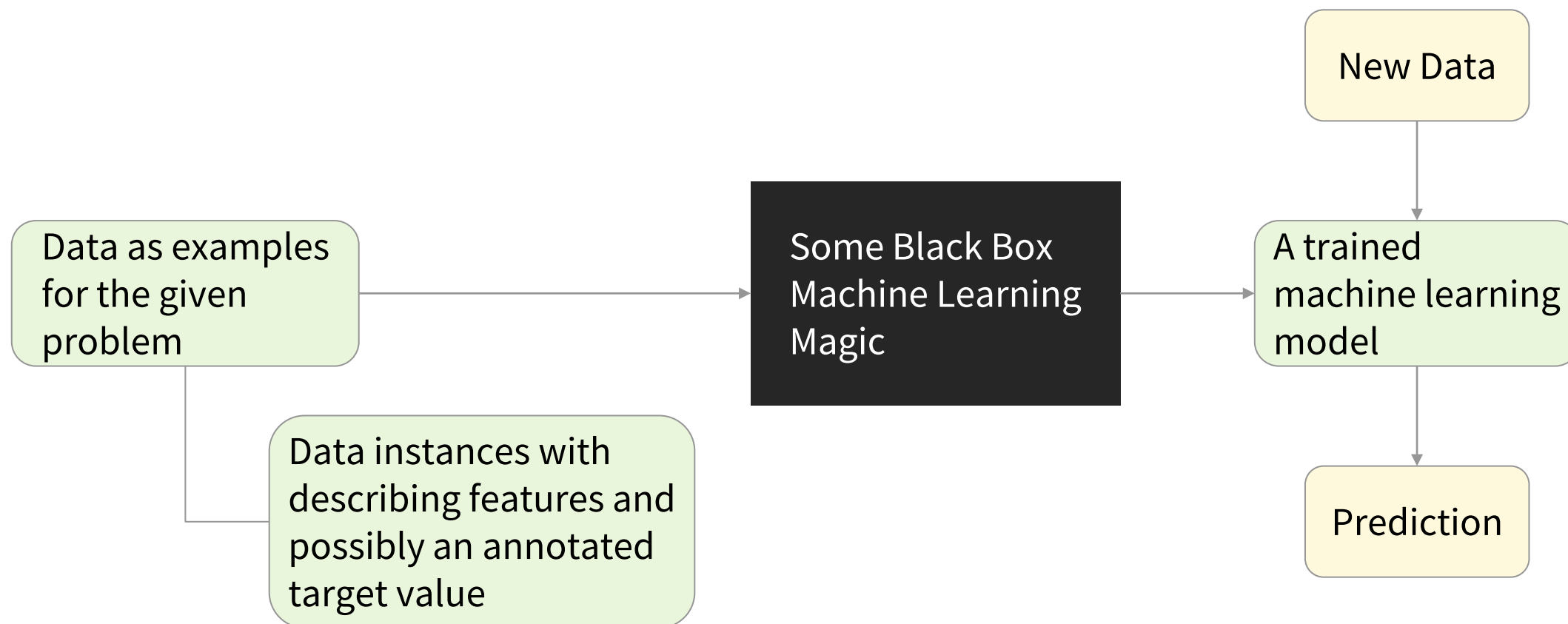
OUT OF THE BLACK BOX! EXPLAINABLE MACHINE LEARNING

Prof. Dr. Korinna Bade

Hochschule Anhalt – Fachbereich Informatik und Sprachen

E-Mail: korinna.bade@hs-anhalt.de

MACHINE LEARNING






INTERPRETABILITY AND EXPLAINABILITY

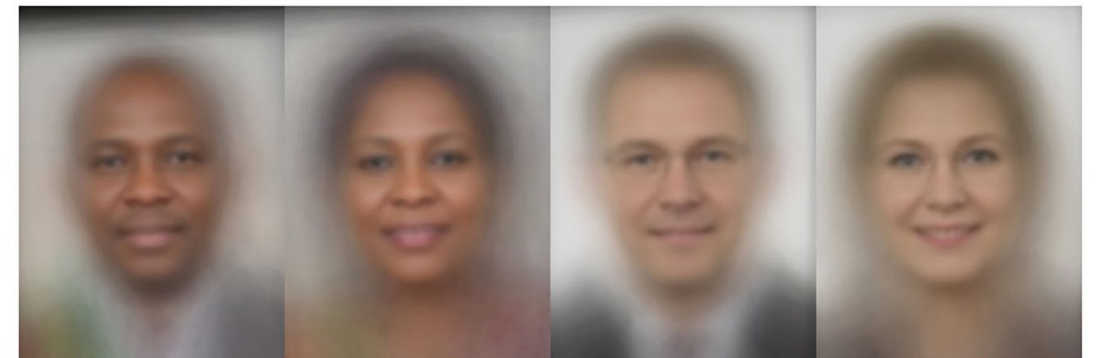
- Tim Miller (2017)
 - »...the degree to which an **observer** can understand the cause of a decision.«
- Been Kim et al. (2016)
 - »...a method is interpretable if a **user** can correctly and efficiently predict the method's results.«
- From the users perspective!
- Interpretability: Understanding the inner workings of the models
- Explainability: Explaining the decisions made
- Necessity
 - Are there significant consequences for unacceptable results?



EXPLAINABILITY

- Intellectual and social motivations
 - Understanding model decisions
 - Gaining trust in model decisions
 - Free of errors
 - Fair, non-discriminatory
- Commercial and legal motivations
 - Improvement of established, analytical processes
 - Compliance with legal regulations

Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
 Microsoft	94.0%	79.2%	100%	98.3%	20.8%
 FACE++	99.3%	65.5%	99.2%	94.0%	33.8%
 IBM	88.0%	65.3%	99.7%	92.9%	34.4%



Gender Shades project, MIT Media Lab, CC BY-NC-ND 2017 - 2020



METHODS FOR EXPLAINABILITY

- Models, which are directly interpretable
 - E.g. decision trees
- Model specific methods
 - E.g. methods specific for neural networks
- Model agnostic methods
 - Independent of the model -> used only as black box



MODEL AGNOSTIC METHODS

- Model is a black-box
 - > Analysis based on input data and observed output
- What we can do:
 - Analyze the given data
 - Visualize dependencies
 - Modify input data and observe the changes in the output
 - > draw conclusions about the model



MODEL AGNOSTIC METHODS - EXAMPLES

- Permutation feature importance
 - Which features have the main impact on the prediction?
 - How?
 - Observe impact on model performance when permutating feature values
- Partial dependence plot
 - What is the average correlation between feature and prediction?
- Example based explanations
 - Counterfactual explanations
 - Which minimal change in the feature values of an instance is necessary to gain the desired prediction?
 - Prototypes and critiques
 - Prototype: representative data instance
 - Critique: data instance not well represented through a prototype



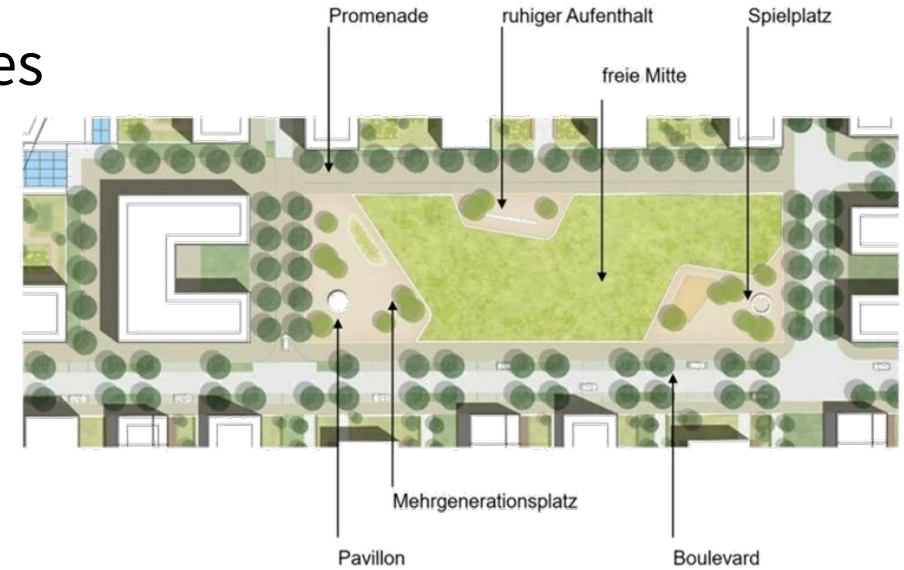
EXPLAINABILITY IN PRACTICE

- Examples of two current projects
 - Explaining clustering in the context of citizen participation
 - Explaining deep neural networks in the context of object detection for autonomous driving



EXAMPLE 1: EXPLAINING CLUSTERING

- Project in the domain of participatory processes
- Process considered
 - Citizens hand in submissions concerning a development plan of a legal institution
 - E.g. a city wants to plan the development of a property unused so far
 - Institution has to consider all submissions and has to treat them equally
- Clustering of submissions to bring together similar submission / submissions addressing the same issue
- Questions of explainability
 - What submissions are in one cluster / in different clusters and why?
 - How does the clustering changes when new submissions arrive or parameters of the algorithm are altered?

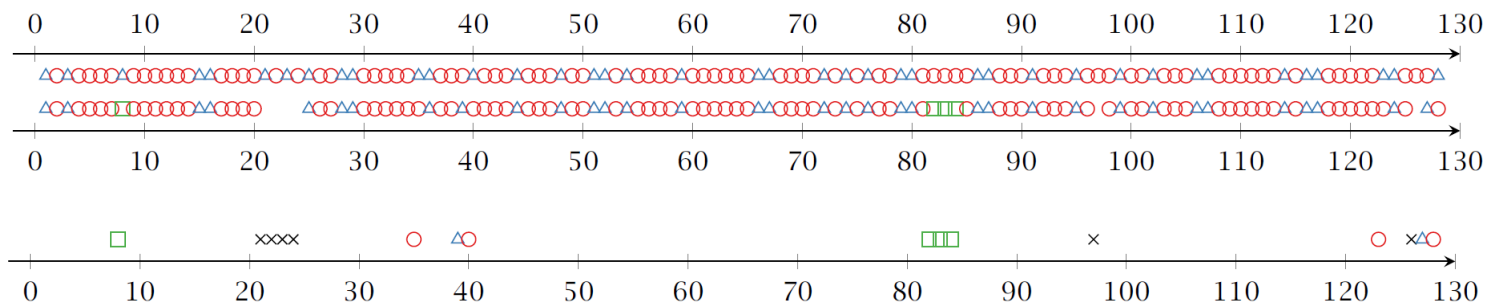


Stadt Unterschleißheim 2022

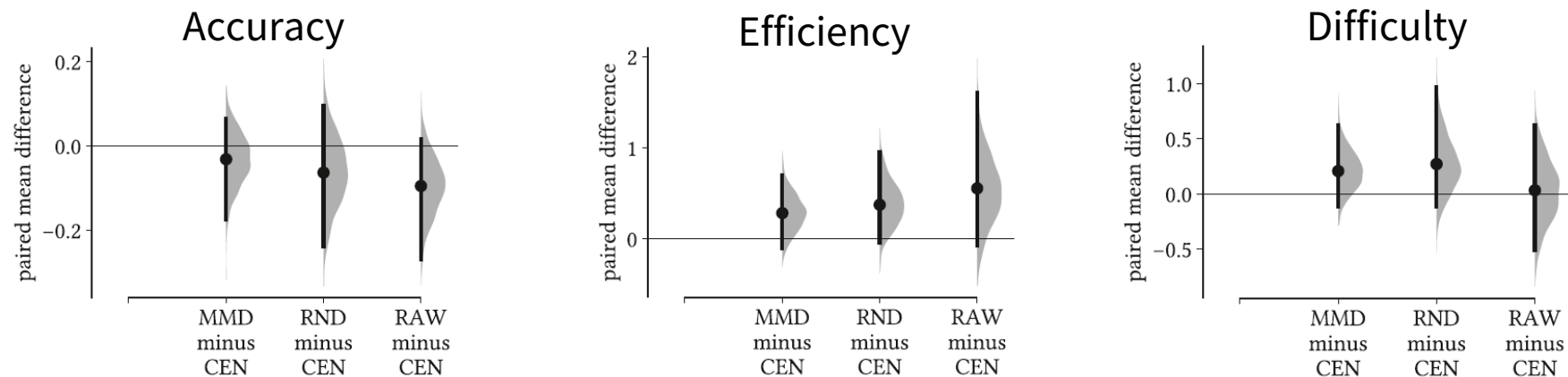


EXAMPLE 1: EXPLAINING CLUSTERING

- Cluster difference model, e.g. for visualization

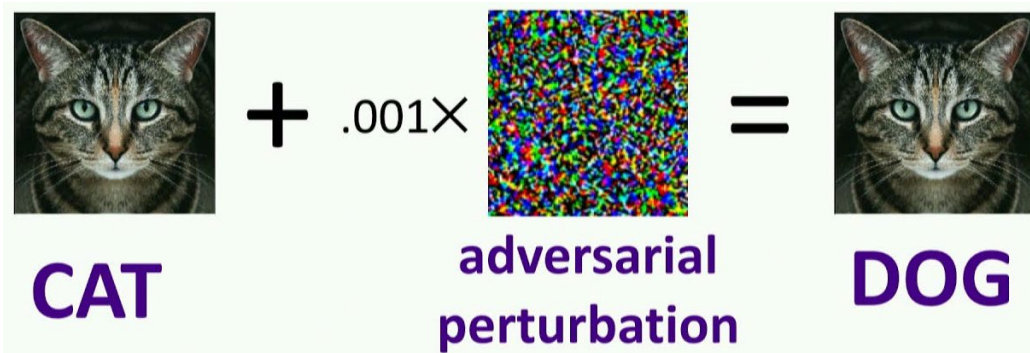
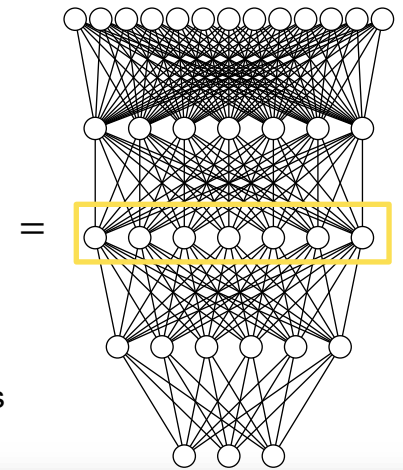
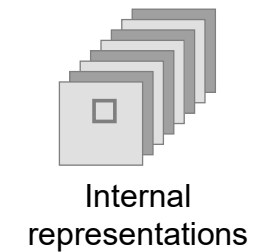


- Studying example based explanations with a user study

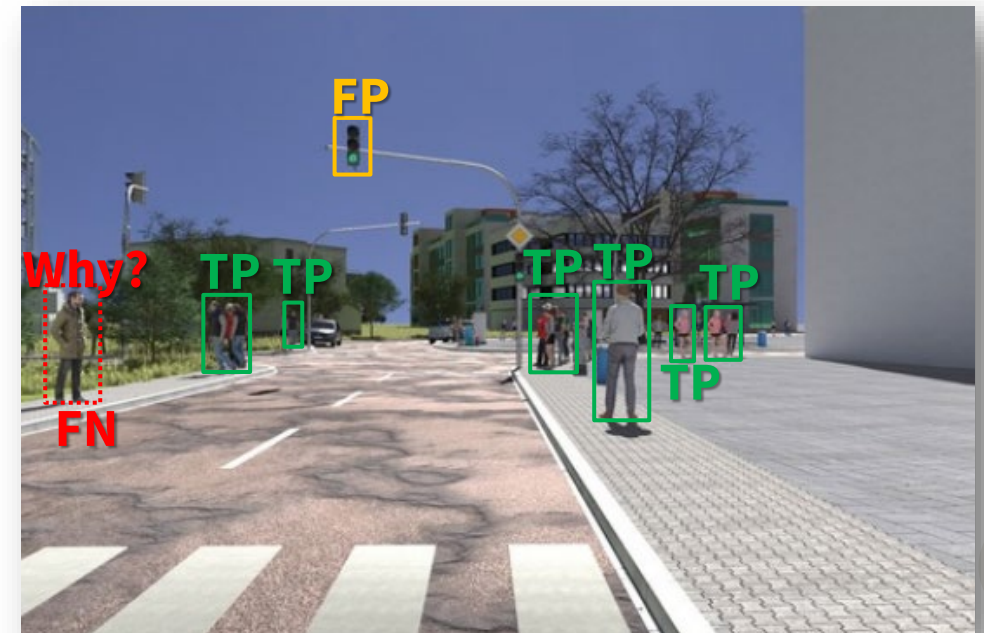


EXAMPLE 2: EXPLAINING DEEP NEURAL NETWORKS

- Project in the domain of autonomous driving
- Detecting pedestrians in images through deep neural networks
- Goal: identify and understand errors of the detection network
- Challenge: networks are very complex and mathematical models
- Sources of errors different to humans

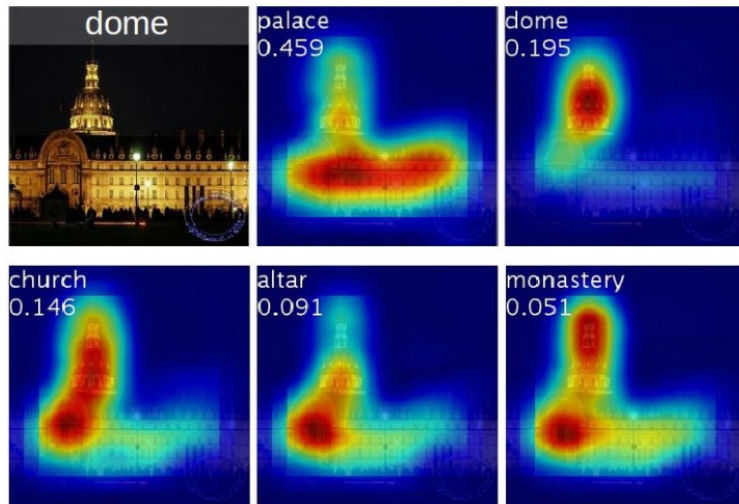


I.J. Goodfellow, J. Shlens, C. Szegedy: Explaining and harnessing adversarial examples, 2015

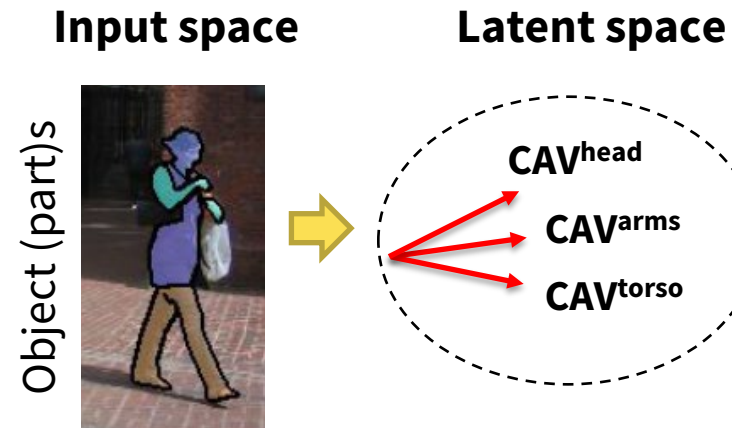


EXAMPLE 2: EXPLAINING DEEP NEURAL NETWORKS

- Saliency maps
 - Simple visualization of layer activation
- Semantic concepts
 - Vector representation of concepts



Zhou et al.: Learning deep features for discriminative localization, 2016



- Understanding concepts in different layers and networks, e.g. stability and similarity





Bildquelle: <https://xkcd.com/1838>

Thank you!

Any questions?

