eXplainable Artificial Intelligence (XAI)

UNIVERSITY OF MANNHEIM
Business School

Lecture Series: Data Science in Action

Prof. Dr. Kevin Bauer

Assistant Professor for E-Business and E-Government

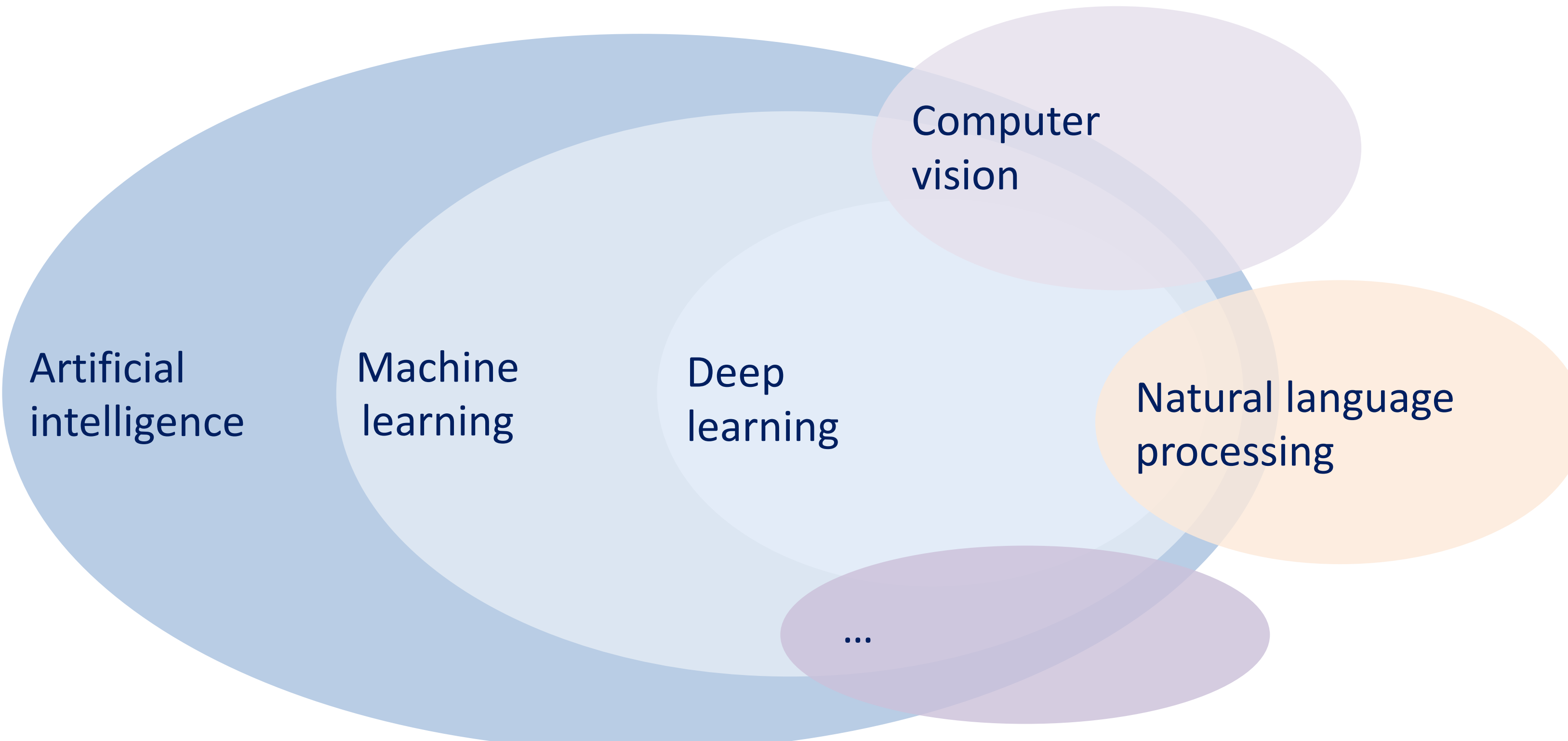
AACSB ACCREDITED EQUIS ACCREDITED ASSOCIATION OF MBAS ACCREDITED

Artificial Intelligence

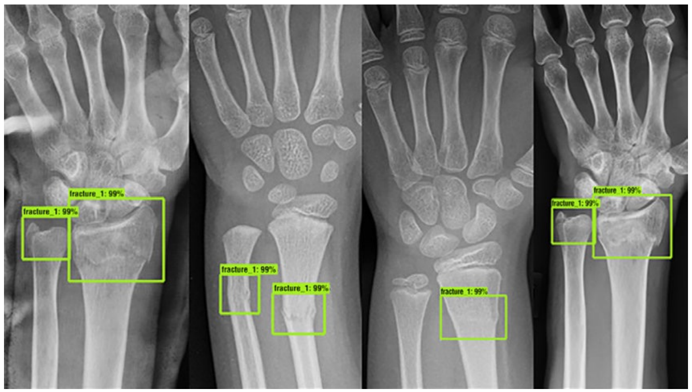
"[…] computer systems able to perform tasks normally requiring human intelligence"

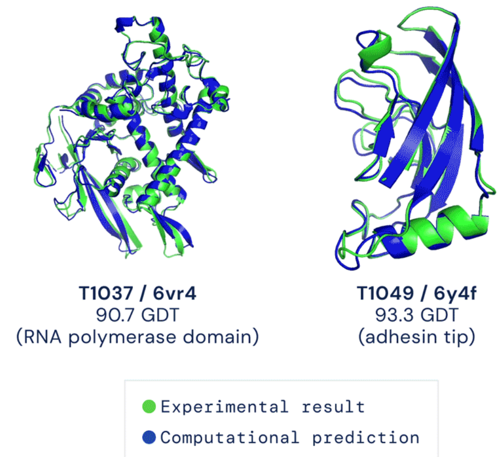
Oxford dictionary

→ AI is the current **frontier** of our efforts to make machines „intelligent“



Predictive AI and decision making





- ML informs decision-making under uncertainty
  - Provides probability that an uncertain state of the world occurs / is true
  - High predictive accuracy fosters AI use in consequential domains
- Other prominent use cases (in business)
  - Sales forecasting
  - Energy consumption forecasting
  - Applicant screening
  - …

Modern AI often black box for humans

Features → Black box → **Prediction of outcome**

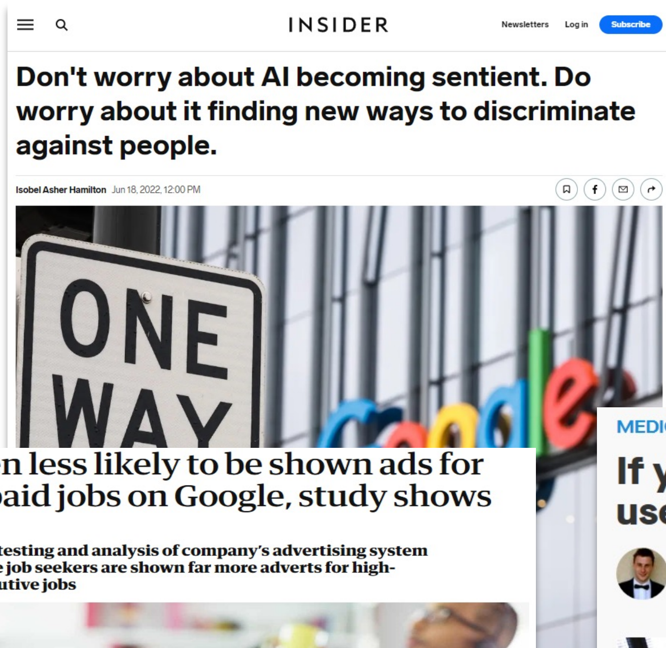
Age: 45
Sex: Female
Income: 70K
Children: 2
…

Fraudulent claim: 76%

Why is this problematic?

AI is prone to errors

AI can be biased


INSIDER

Newsletters Log in Subscribe

Don't worry about AI becoming sentient. Do worry about it finding new ways to discriminate against people.

Isobel Asher Hamilton Jun 18, 2022, 12:00 PM


The New York Times

Apple Card Investigated After Gender Discrimination Complaints

A prominent software developer said on Twitter that the credit card was "sexist" against women applying for credit.


Women less likely to be shown ads for high-paid jobs on Google, study shows

Automated testing and analysis of company's advertising system reveals male job seekers are shown far more adverts for high-paying executive jobs


MEDICAL MALAISE

If you're not a white male, artificial intelligence's use in healthcare could be dangerous

By Robert David Hart • July 10, 2017

Legal and regulatory initiatives

- **General Data Protection Regulation (GDPR)** in the EU demands "*transparent [data] processing*" drawing upon "*appropriate mathematical or statistical procedures*"
- EU's proposal for **AI Act**:
  "AI systems used to evaluate the credit score or creditworthiness of natural persons should be classified as high-risk AI systems, since they determine those persons' access to financial resources or essential services such as housing […]"
- **Algorithmic Accountability Act** in the US goes in similar direction and effectively requires businesses to promote transparency in their AI systems
- …

https://www.digitalsme.eu/artificial-intelligence-act/

https://www.protocol.com/enterprise/revised-algorithmic-accountability-bill-ai

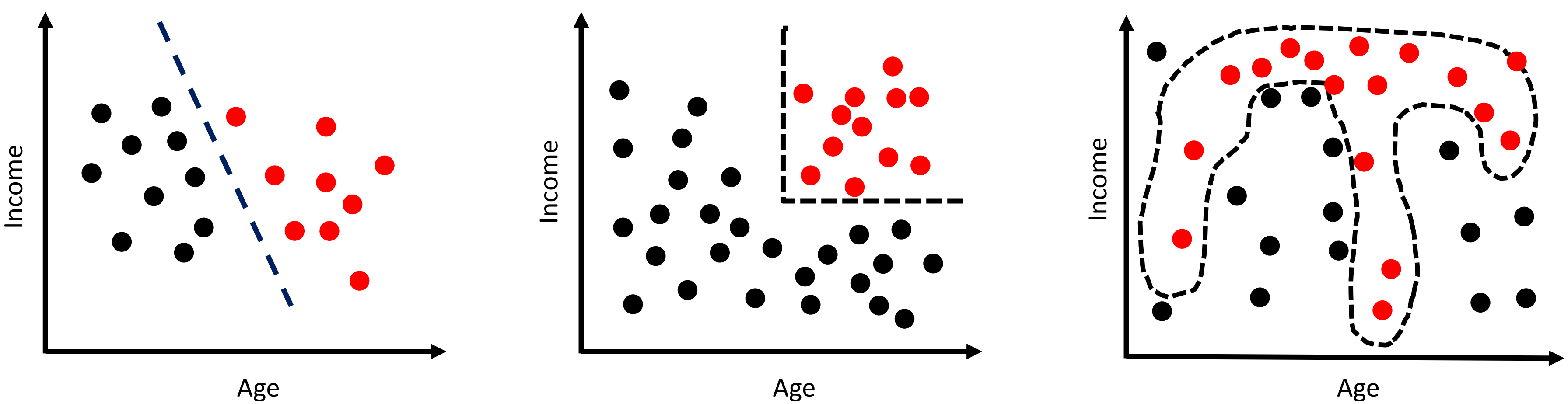
Another curse of dimensionality

UNIVERSITY OF MANNHEIM
Business School

(1) Linear model:
5*Inc+Age-7 > 0
Otherwise

(2) Non-linear model:
Inc & Age > 0.5
Otherwise

(3) High dimensional model:
?



- (1) and (2) understandable for humans
- (3) unclear how decisions occur; we could fit a linear model for (3) but this would lead to many errors

Four arguments pro explaining black boxes

[Adadi & Berrada 2018]

- **Explain to justify:** ensure an auditable and provable way to defend outputs
  - Individual customer **inquiring why** she was assessed as high risk for fraud
- **Explain to control:** identify and override erroneous predictions
  - Insurance agent can better **understand when to overrule** and adjust the premium estimation for individual customers
- **Explain to improve:** enable improvement of ML models
  - Developers can understand what information the model uses and how, enabling them to **correct biased behaviors and build more generalizable mode**ls
- **Explain to discover:** enable to recognize previously unknown patterns
  - Identify new patterns in big data structures allowing users to **learn from the AI**

eXplainable AI (XAI)

- **Explainability**: degree to which humans can understand model predictions to ensure fairness, accountability, and transparency

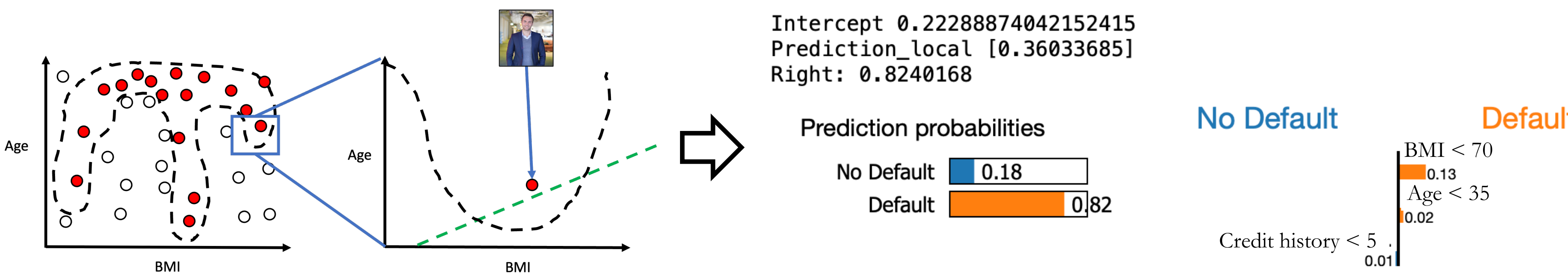
- **What** to explain?
  - Local (single prediction)
  - Global (whole prediction model)
- **How** to explain?
  - Inherently interpretable models → e.g., via logistic regression
  - Post-hoc explanations
    - Model-specific → e.g., based on splits in tree algorithms
    - Model-agnostic → e.g., based on surrogate models

Surrogate explanations: LIME & SHAP

Local Interpretable Model-Agnostic Explanations

[Ribeiro et al. 2016]

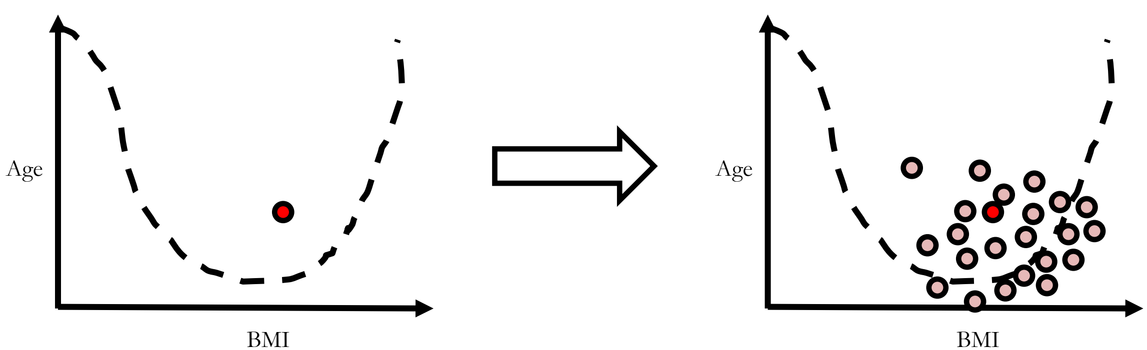
- Explains why the model makes a specific prediction for a specific data point
- Concrete implementation of how to build local surrogate models]
- Provides explanations based on input features. Example: why did I not get a loan?



Intuition behind LIME (tabular data)

Step 1: Create perturbed data around instance to be explained

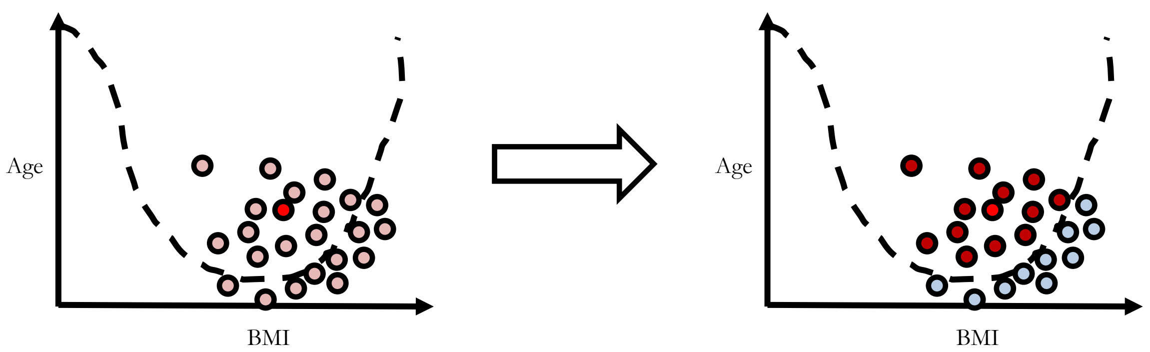
- Perturbation of data based on empirical distributions in training set
  - E.g., slightly increase the BMI and slightly decrease the Age
- Random creation of synthetic observations



Intuition behind LIME

Step 2: Use complex model to make predictions for new data

- Results in a new, artificial data set
- labels (=predictions)
- features (=perturbations)

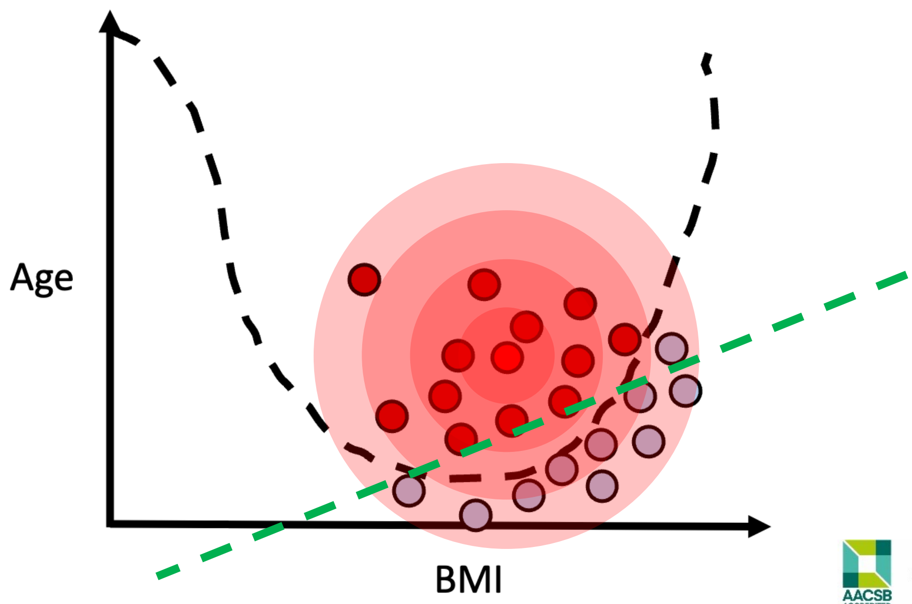


Intuition behind LIME

UNIVERSITY OF MANNHEIM
Business School

Step 3: Train a simple (linear) model on the new data

- Weighting of new data points according to distance to instance we aim to explain
- The further away from original point, the less important (not in local neighborhood)
- Simple model provides insights into local working of complex one, e.g., LASSO coefficients



AACSB ACCREDITED EQUIS ACCREDITED ASSOCIATION OF MBAS ACCREDITED

Intuition behind LIME

Set of interpretable models *G*

Regularization of surrogate to make it simple

$$\xi(x) = argmin_{g \in G}\, \mathcal{L}(f, g, \pi_x) + \Omega(g)$$

Data point *x*

Complex model *f*

Surrogate model *g*

Neighborhood of *x*

- $\mathcal{L}(f, g, \pi_x)$: find a simple model ***g*** that approximates the complex model ***f*** well in the local neighborhood $\boldsymbol{\pi_x}$ of the current data point ***x***
- $\Omega(g)$: penalize surrogate model ***g***'s complexity to ensure interpretability

(Dis)advantages of LIME

+ Advantages

Model agnostic and freedom to choose surrogate model

Short and simple explanations

Applicable for tabular, image, and text data

− Disadvantages

Definition of neighbourhood

Unlikely synthetic data points

Very similar data points may obtain very different explanations (instability)

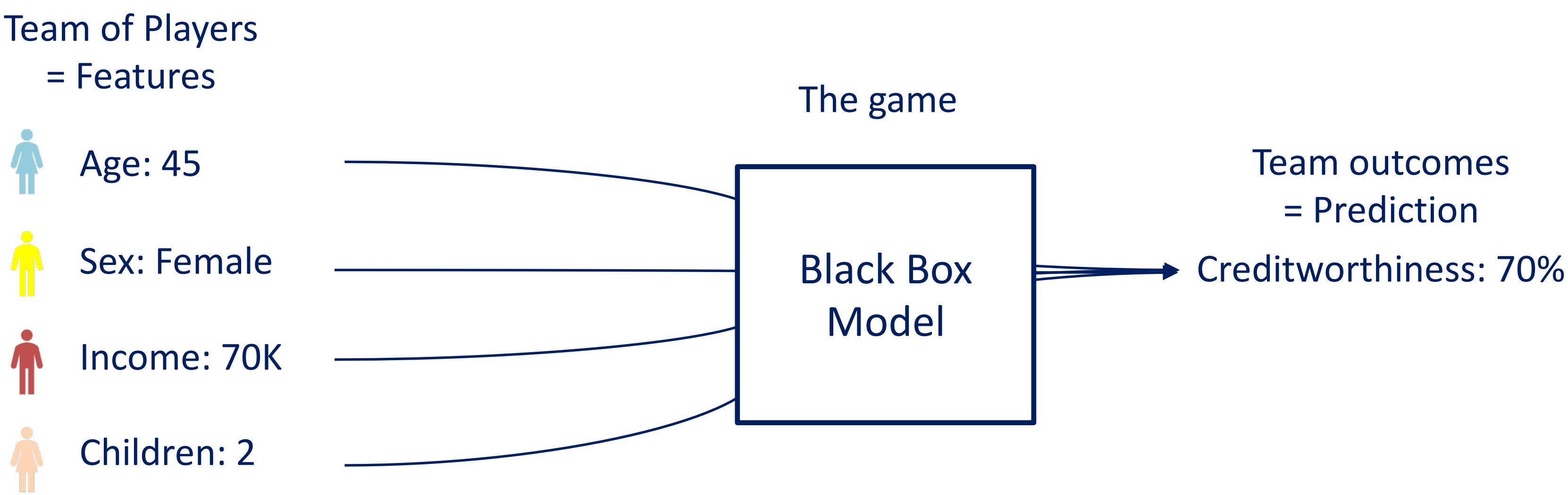
Susceptible to manipulation

SHapley Additive exPlanations

[Lundberg & Lee, 2017]

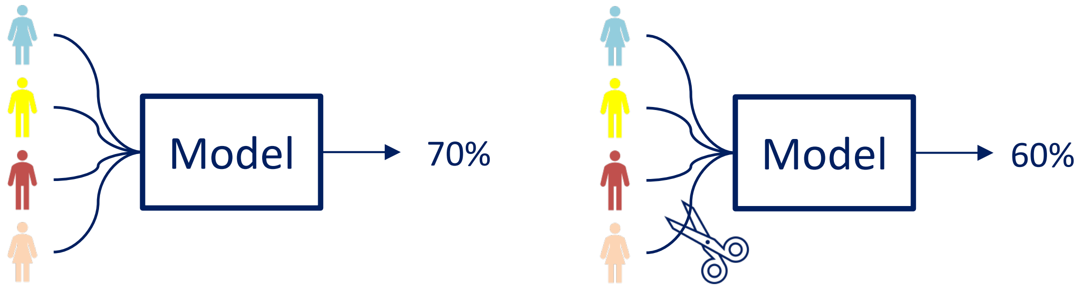
SHAP

- Rooted in Collaborative Game Theory [Shapley 1951]
- Average individual contribution of a player in a team to the outcome of the group
- SHAP asks: What outcome would the group achieve (prediction) if a specific player (feature) would have been excluded?

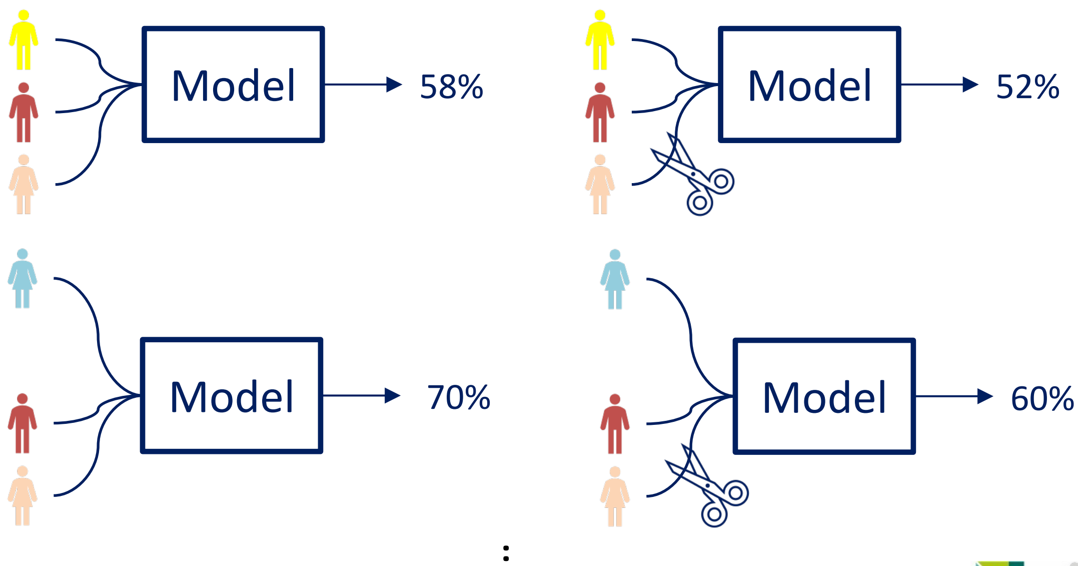


Intuition behind SHAP

- Remove a player to compute marginal change in prediction



- Average over removal from all possible subsets



Intuition behind SHAP

Data point *x*

Shapley value for *i* = age

$$\phi_i(f,x) = \sum_{z' \subseteq x'} \frac{|z'|!\,(M - |z'| - 1)!}{M!}\left(f_x(z') - f_x(z' \backslash i)\right)$$

Model *f*

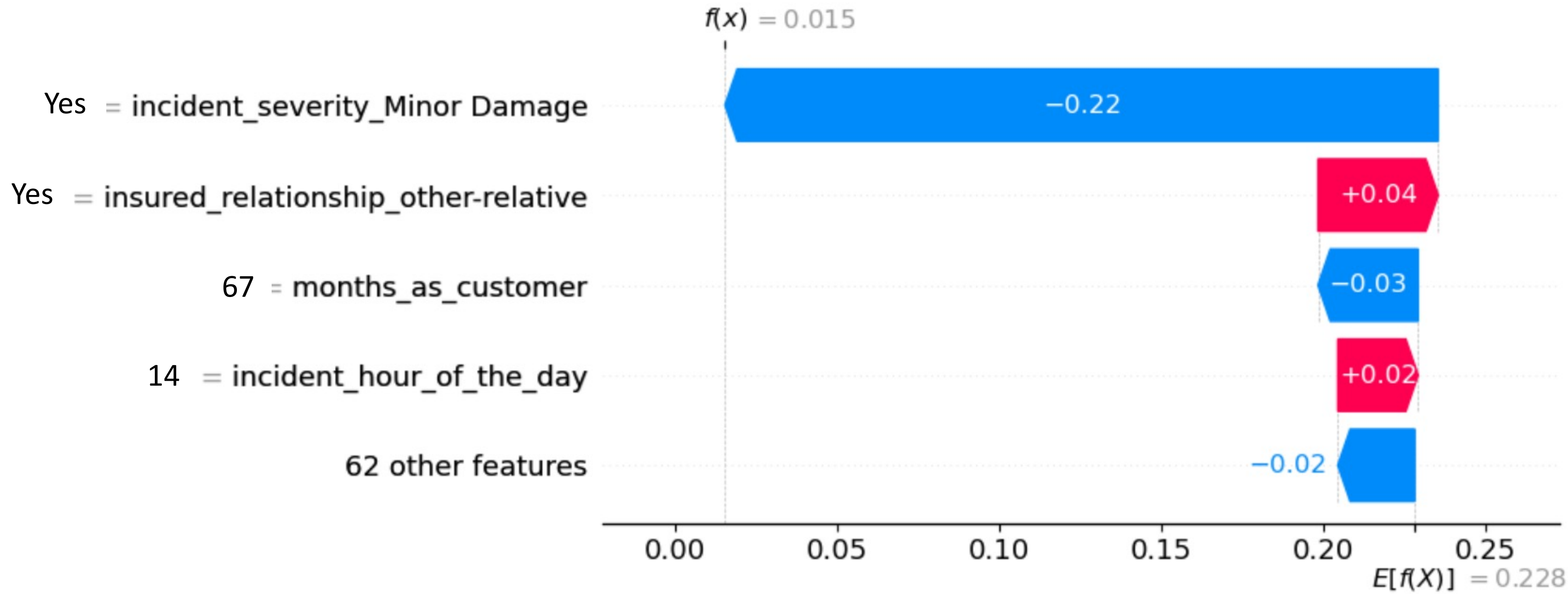
Weighted sum over all
Subsets *z'* of (transformed)
Data point *x;* where *M* is total
number of features in full set

Difference in
Model prediction

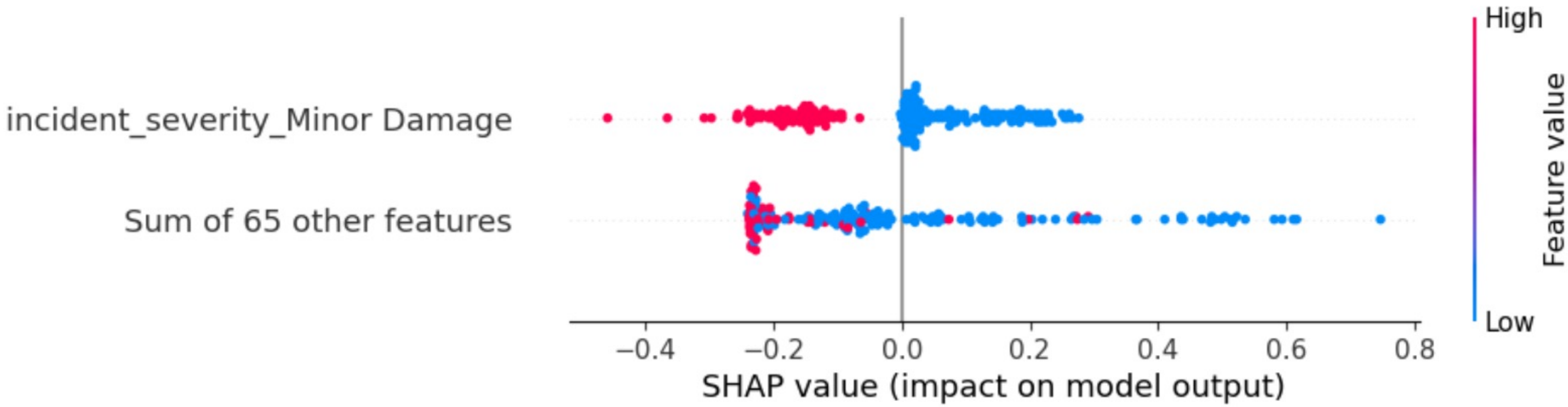
- Removal of information in data set: random draw from background data set (random features has no predictive power)
- Approximate SHAP values due to high complexity (2^N)
- Note: there are model-specific versions of SHAP, e.g., Tree-Shap, Deep-Shap that use model internals

SHAP for XGB model predicting insurance fraud – local explanation

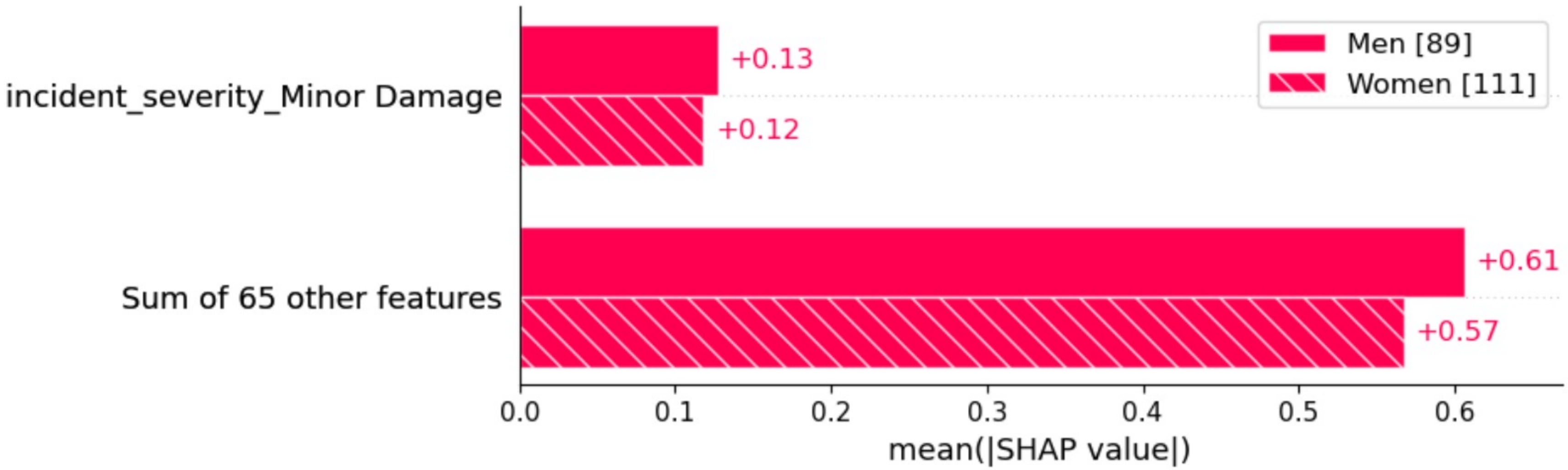
```
Model prediction for instance: 0.015453994274139404
```



SHAP for XGB model predicting insurance fraud – global explanation



SHAP for XGB model predicting insurance fraud – gender bias?



(Dis)advantages of SHAP

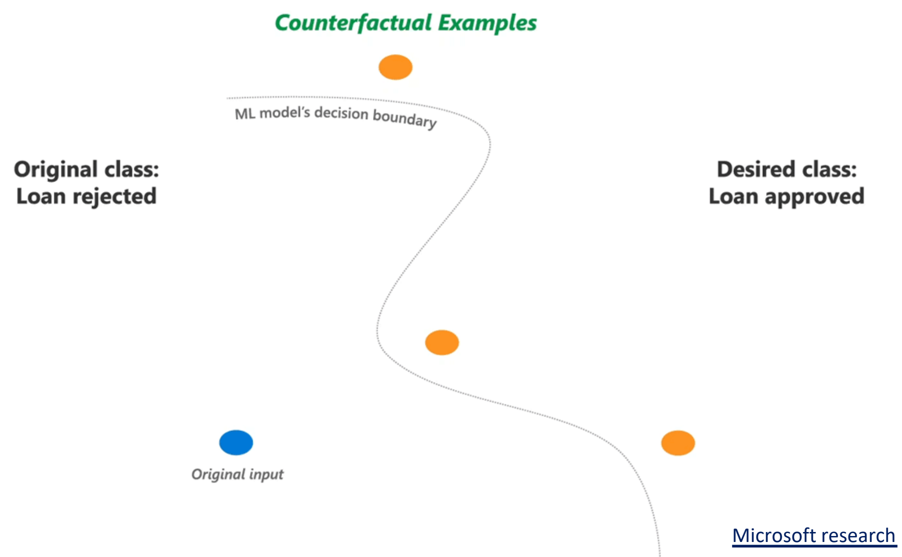
UNIVERSITY OF MANNHEIM
Business School

+	−
Sound theoretical foundation with nice mathematical properties	Typically slow in computation
Consistent global explanations	Depending on type, SHAP can ignore feature dependences
Short and simple explanations	Susceptible to manipulation
Applicable for tabular, image, and text data	

Counterfactual explanations

Counterfactual explanations

UNIVERSITY OF MANNHEIM
Business School



[Microsoft research]

- Counterfactual explanations provide an understanding of model decisions by posing "what if" scenarios.
- Highlight scenarios where small input changes alter model decisions, e.g., reaching a certain threshold
- A counterfactual explanation is the smallest change to feature values changing the prediction to a predefined output.

Implementation of counterfactual explanations

- Maximization problem to find a set of data points that
  - lead to a prediction as close as possible to desired prediction
  - are as similar to original data point as possible
  - change a small set of features
  - represent likely feature combinations
- Different implementations
  - MACE [Karimi et al. (2020)]
  - DiCE (Diverse Counterfactual Explanation) [Mothilal et al., 2020]

DiCE

Query instance (original outcome : 0)

	age	workclass	education	marital_status	occupation	race	gender	hours_per_week	income
0	22.0	Private	HS-grad	Single	Service	White	Female	45.0	0.01904

Diverse Counterfactual set (new outcome : 1)

	age	workclass	education	marital_status	occupation	race	gender	hours_per_week	income
0	70.0	-	Masters	-	White-Collar	-	-	51.0	0.534
1	-	Self-Employed	Doctorate	Married	-	-	-	-	0.861
2	47.0	-	-	Married	-	-	-	-	0.589
3	36.0	-	Prof-school	Married	-	-	-	62.0	0.937

(Dis)advantages of counterfactuals

UNIVERSITY OF MANNHEIM
Business School

+	−
Simple explanation	There are typically multiple counterfactuals
Does not require access to data, only model	For multiple feature value changes it is unclear how changes affected prediction individually
Easy to implement	

A final word of caution

Explanations are no "silver bullet" for AI problems

- Explanations can create **data privacy and intellectual property concerns**
- Explanations may enable people to **game the system**
- **Deliberate manipulation** and hiding of bias [Lakkaraju & Bastani, 2020]
- Explanations may invoke unintended **behavioral side effects** for users
  - **Overreliance** and **blind delegation** to AI [Bauer et al., 2023]
  - **Confirmatory learning** [Bauer et al., 2023]
  - **Informational overload** [Poursabzi-Sangdeh et al., 2021]

Thank you for your attention!

kevin.bauer@uni-mannheim.de

Dr. Kevin Bauer
Human-centric AI | Machine Learning | Economics | Behavioral Economics

Python Live Demo

References

- Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). IEEE access, 6, 52138-52160.
- Bauer, K., Hinz, O., van der Aalst, W., & Weinhardt, C. (2021). Expl (AI) n it to me–explainable AI and information systems research. Business & Information Systems Engineering, 63, 79-82.
- Bauer, K., von Zahn, M., & Hinz, O. (2023). Expl (AI) ned: The impact of explainable artificial intelligence on users' information processing. Information Systems Research.
- Bauer, K., von Zahn, M., & Hinz, O. (2023). Please take over: XAI, delegation of authority, and domain knowledge. SAFE Working Paper No. 394, Available at SSRN: https://ssrn.com/abstract=4512594 or http://dx.doi.org/10.2139/ssrn.4512594
- Deprez, P., Shevchenko, P. V., & Wüthrich, M. V. (2017). Machine learning techniques for mortality modeling. European Actuarial Journal, 7, 337-352.
- Dhieb, N., Ghazzai, H., Besbes, H., & Massoud, Y. (2019, September). Extreme gradient boosting machine learning algorithm for safe auto insurance operations. In 2019 IEEE international conference on vehicular electronics and safety (ICVES) (pp. 1-5). IEEE.
- Guelman, L. (2012). Gradient boosting trees for auto insurance loss cost modeling and prediction. Expert Systems with Applications, 39(3), 3659-3667.
- Lakkaraju, H., & Bastani, O. (2020, February). " How do I fool you?" Manipulating User Trust via Misleading Black Box Explanations. In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (pp. 79-85).
- Liu, Q., Pitt, D., & Wu, X. (2014). On the prediction of claim duration for income protection insurance policyholders. Annals of Actuarial Science, 8(1), 42-62.
- Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in neural information processing systems, 30.
- Poursabzi-Sangdeh F., Goldstein DG., Hofman JM, Wortman Vaughan JW., Wallach H. (2021). Manipulating and measuring model interpretability. Proc. CHI Conf. on Human Factors in Comput. Systems.