Techniques for Explainable Artificial Intelligence in Insurance

Dr. Oliver Pfaffel

Munich Re

EAA e-Conference on
Data Science & Data Ethics

29 June 2021

AGENDA

WHAT WE WANT TO COVER TODAY:

1. Risks from the use of AI
2. Techniques for explainable AI in insurance
3. Weak spots in explanation algorithms
4. Outlook on self-explaining AI

INCURRED RISKS IN THE APPLICATION OF AI

"Success in creating AI would be the biggest event in human history. Unfortunately, it might also be the last, unless we learn how to avoid the risks.”

--*Stephen Hawking*

INCURRED RISKS IN AI

AI CAN FAIL IN CRITICAL APPLICATIONS

2015: AI discriminating job applicants
In their hiring process, Amazon used an AI algorithm that preferred male over female applicants.

2016: Chatbot AI out of control
Microsoft deployed a chatbot on Twitter that “turned into a racist” within a few hours.

2018: Inaccurate AI-assisted medical diagnosis
IBM Watson’s AI-based supercomputer helping doctors to diagnose patients is often inaccurate with respect to its oncology capabilities.

2019: Discrimination in the granting of loans
Financial regulators in New York launched an investigation into the algorithm behind Apple’s credit card after users reported that women had received lower credit limits than men

DISCRIMINATION IN PRICING FOR CERTAIN GROUPS OF PERSONS

Example

MO COMPARE Motorists fork out £1,000 more to insure their cars if their name is Mohammed

Top firms such as Admiral and Marks & Spencers have been dragged into an insurance race row after giving far lower quotes for drivers with traditionally English names like John

Source: https://www.thesun.co.uk/motors/5393978/insurance-race-row-john-mohammed/

Problem

- "The Sun" reported that motor insurers in UK had up to 69% higher prices for individuals called Mohammed instead of John (everything else being the same)
- The name was implicitly used by an AI algorithm to differentiate prices – discriminating against the ethnic origin

COMPLEXITY VS EXPLAINABILITY

Occam's razor (the principle of parsimony) in times of trillion[2] parameter models:

	GLM/GAM	Decision Tree	Tree ensembles [1]	Deep Learning
Stat. robustness	↘	↗	↑	↘
Functionality	↘	↑	↑	↑
Predictive Perf.	→	↘	↑	↑
A priori Explainability	↑	↗	↘	↓

- Tree ensembles often outperform GLM for classical actuarial problems
- For NLP and Computer Vision Deep Learning strongly outperforms classical approaches in most use cases
- Thus, we cannot always approximate a complex problem with a simple model
- Can we approximate the "reasoning" of a complex model by the "reasoning" of simple model or isolate certain "paths" of it?

→ Explainable AI

1. Random forest, tree boosting, etc. 2. Google's Switch Transformer has 1.6 trillion parameters

TECHNIQUES FOR EXPLAINABLE AI IN INSURANCE

"An approximate answer to the right problem is worth a good deal more than an exact answer to an approximate problem."

-- John Tukey

OVERVIEW[1] OF EXPLAINABLE AI (XAI)

Type of problem	Global Explanation *Which general principles determine a certain model behavior?*	Local Explanation *Which input can we attribute a certain model prediction to?*
Tabular data	• **Permutation Feature Importance:** *Aggregate impact of each feature on the prediction* • **Partial Dependence / ALE:** *An increase in feature x changes a prediction by …% on average*	• **Ceteris paribus plots:** *Changing feature x changes the prediction by …* • **Breakdown plots / SHAP / LIME** *Contribution of each feature on a single prediction*

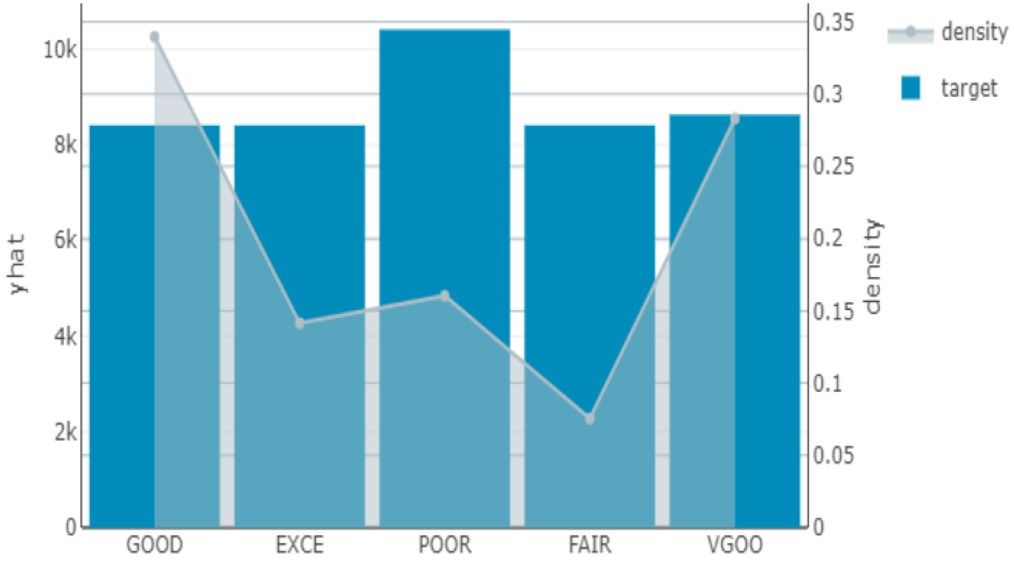
1 Non-exhaustive overview of the most relevant categories as considered by the author

GLOBAL AND LOCAL EXPLANATIONS FOR TABULAR DATA

Use Case: Prediction of loss severity in health insurance by age, gender, physical status and further risk factors

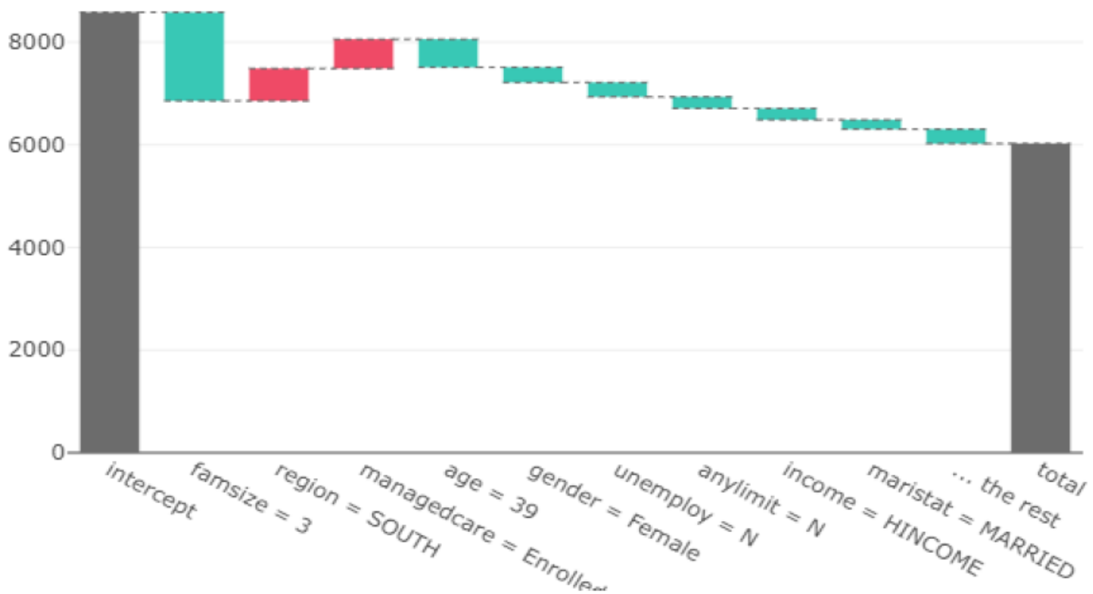
Global explanation using a Partial Dependence Plot

Overall impact of physical status



Local explanation using a break down plot (here: xgboostExplainer)

Loss prediction for a single insured



OVERVIEW[1] OF EXPLAINABLE AI (XAI)

Type of problem	Global Explanation *Which general principles determine a certain model behavior?*	Local Explanation *Which input can we attribute a certain model prediction to?*
Tabular data	• **Permutation Feature Importance:** *Aggregate impact of each feature on the prediction* • **Partial Dependence / ALE:** *An increase in feature x changes a prediction by …% on average*	• **Ceteris paribus plots:** *Changing feature x changes the prediction by …* • **Breakdown plots / SHAP / LIME** *Contribution of each feature on a single prediction*
Computer Vision	• **Feature Visualization** *Find the input that maximizes the activation of layer or neuron*	• **Feature Attribution via Integrated Gradients, Gradient SHAP, LRP, …** *Determine parts of an image that are responsible for the model prediction*

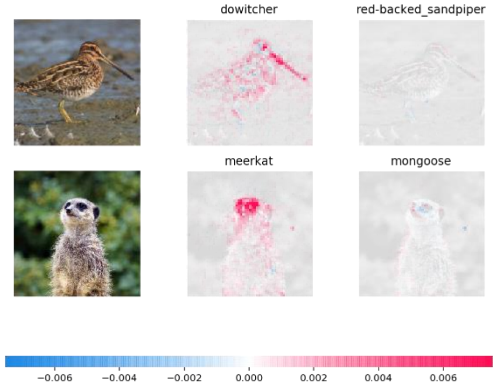
Baseball—or stripes? *mixed4a, Unit 6*

Animal faces—or snouts? *mixed4a, Unit 240*

Clouds—or fluffiness? *mixed4a, Unit 453*

Buildings—or sky? *mixed4a, Unit 492*

Source: https://distill.pub/2017/feature-visualization/



Source: Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions

1 Non-exhaustive overview of the most relevant categories as considered by the author

european actuarial academy

Data - Ethics - Actuary

OVERVIEW[1] OF EXPLAINABLE AI (XAI)

Type of problem	Global Explanation *Which general principles determine a certain model behavior?*	Local Explanation *Which input can we attribute a certain model prediction to?*
Tabular data	• **Permutation Feature Importance:** *Aggregate impact of each feature on the prediction* • **Partial Dependence / ALE:** *An increase in feature x changes a prediction by …% on average*	• **Ceteris paribus plots:** *Changing feature x changes the prediction by …* • **Breakdown plots / SHAP / LIME** *Contribution of each feature on a single prediction*
Computer Vision	• **Feature Visualization** *Find the input that maximizes the activation of layer or neuron*	• **Feature Attribution via Integrated Gradients, Gradient SHAP, LRP, …** *Determine parts of an image that are responsible for the model prediction*
NLP	• **Universal (adversarial) triggers** *Find a phrase that, if inserted into any input, would cause a certain prediction y*	• **Attribution to training samples via Representer Point Selection** *Determine the most relevant training samples that are responsible for the model prediction*

Task	Input (red = trigger)	Model Prediction
Sentiment Analysis	zoning tapping fiennes Visually imaginative, thematically instructive and thoroughly delightful, it takes us on a roller-coaster ride…	Positive → Negative
	zoning tapping fiennes As surreal as a dream and as detailed as a photograph, as visually dexterous as it is at times imaginatively overwhelming.	Positive → Negative

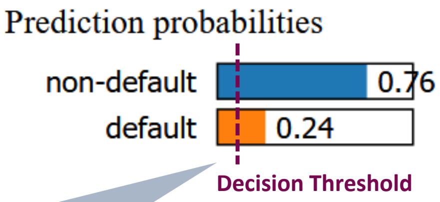
Source: Wallace, E., Feng, S., Kandpal, N., Gardner, M., & Singh, S. (2019). Universal adversarial triggers for attacking and analyzing NLP. *arXiv preprint arXiv:1908.07125*.

1 Non-exhaustive overview of the most relevant categories as considered by the author

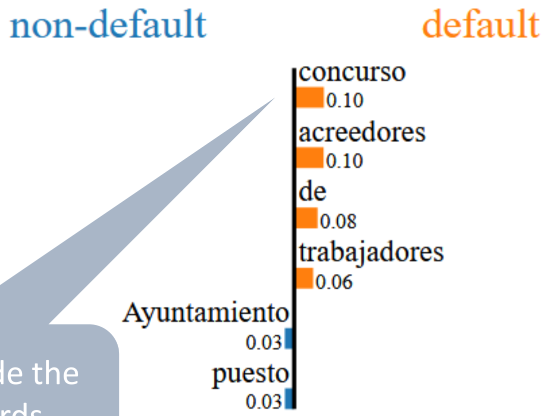
FEATURE ATTRIBUTION IN NATURAL LANGUAGE PROCESSING

Use Case: Predict the likelihood of a default of a certain company within a certain time frame

Local Interpretable Model-agnostic Explanations (LIME) uses linear models to replicate the prediction of the more complex original model

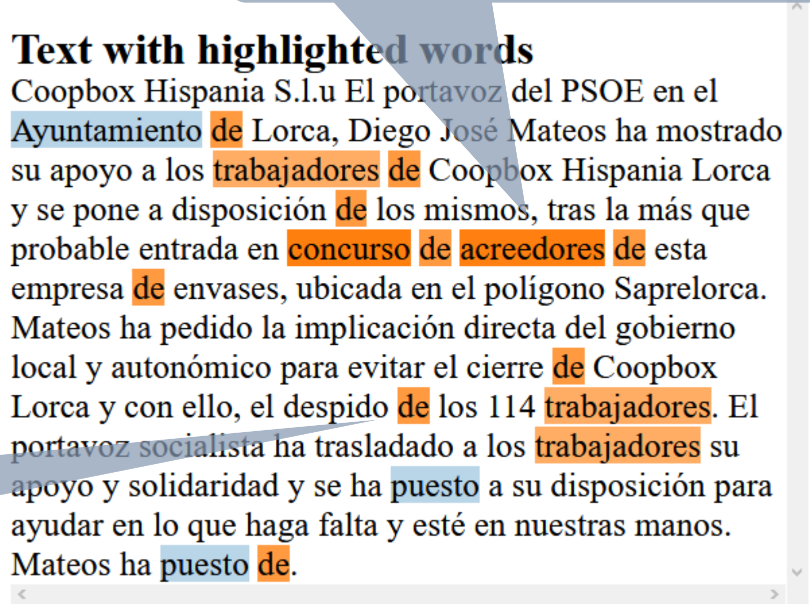


Prediction of "default" since score is above the threshold (note that there is a low incidence rate of observed defaults)



Orange words guide the prediction towards "default"

"likely entry into bankruptcy (concurso de acreedores)"



Text with highlighted words

Coopbox Hispania S.l.u El portavoz del PSOE en el Ayuntamiento de Lorca, Diego José Mateos ha mostrado su apoyo a los trabajadores de Coopbox Hispania Lorca y se pone a disposición de los mismos, tras la más que probable entrada en concurso de acreedores de esta empresa de envases, ubicada en el polígono Saprelorca. Mateos ha pedido la implicación directa del gobierno local y autonómico para evitar el cierre de Coopbox Lorca y con ello, el despido de los 114 trabajadores. El portavoz socialista ha trasladado a los trabajadores su apoyo y solidaridad y se ha puesto a su disposición para ayudar en lo que haga falta y esté en nuestras manos. Mateos ha puesto de.

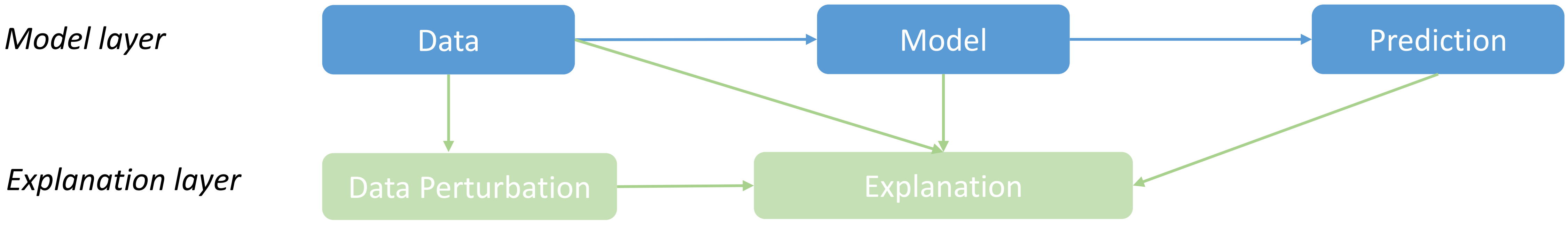
" (..) avoid the (..) dismissal of the 114 workers (trabajadores)"

WEAK SPOTS IN EXPLANATION ALGORITHMS

"Post-hoc XAI models are also just models."

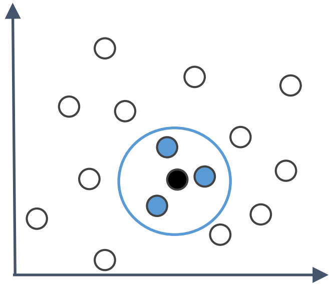
XAI RELIES ON ASSUMPTIONS AND IS PRONE TO ATTACKS

- XAI techniques typically make use of the underlying training / testing data
- Often data perturbation is required for ceteris paribus explanations (= “what if”) or contrastive explanations
- XAI is often sensitive towards changes of the input data
- XAI may rely on the method for data perturbation (if applicable)



LIME & SHAP RELY ON THE METHOD FOR DATA PERTURBATION

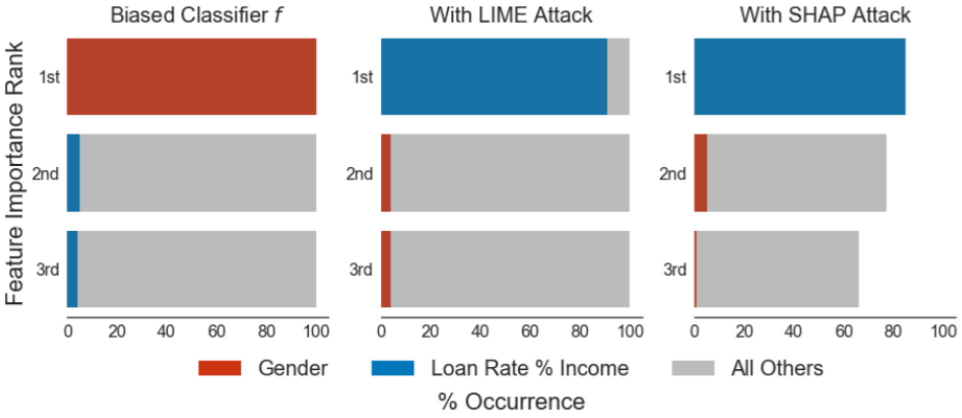
- A malicious attacker may **hide a biased model** under the hood of a seemingly unbiased model from an auditor
- Works for (Kernel) SHAP or LIME since the data perturbation mechanism is known



Black: Data point
Blue: Artificial data point close to data
White: Artificial data point out of data (OOD)

Explanation of the attack

1. Points close to the data are labelled “data”, rest “OOD”
2. Train a classifier
3. Define adversarial model such that a biased model is evaluated on what is predicted „data“, and an unbiased model on all other data points



Source: Slack, D., Hilgard, S., Jia, E., Singh, S., & Lakkaraju, H. (2020, February). Fooling lime and shap: Adversarial attacks on post hoc explanation methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (pp. 180-186).

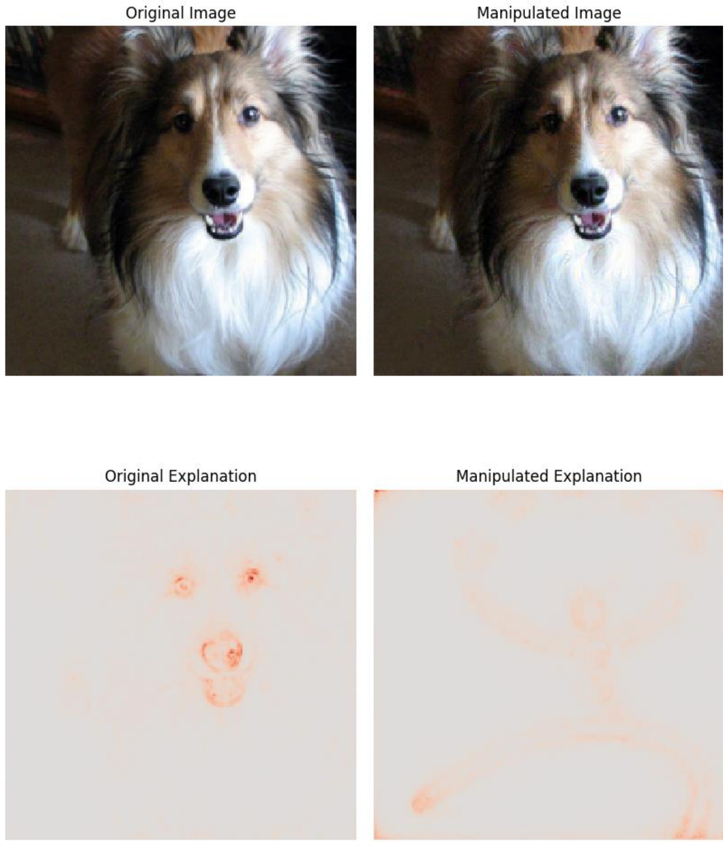
- Individual credit assessment based on account information
- Biased classifier uses only gender to make a decision (unfair)
- Unbiased classifier uses only “Loan Rate relative to Income” (fair)
- Explanation of adversarial model with LIME and SHAP seems to confirm that “gender” is of minor importance

INTEGRATED GRADIENTS IS SENSITIVE TOWARDS THE INPUT DATA

- Idea: Fine-tuning of the model adding a target explanation to the loss function:

  Loss ~ distance(manipulated explanation, target explanation) + γ * distance(manipulated prediction, original prediction)

- The updated model provides visually the exact same prediction (though there is a slight numerical change)
- The explanation is very close to the target, which can be virtually anything
- Why relevant?
  - Shows the limitations of post-hoc XAI
  - Can be exploited if the attacker can select or has knowledge of the data used to explain the model
  - Situation where the task is difficult even for a human and the explanation is required to understand the prediction (e.g. medical AI)



Source: Dombrowski, A. K., Alber, M., Anders, C. J., Ackermann, M., Müller, K. R., & Kessel, P. (2019). Explanations can be manipulated and geometry is to blame. *arXiv preprint arXiv:1906.07983*.

SELF-EXPLAINING AI

"Making Explainability part of AI."

XAI CAN BE MADE PART OF THE MODEL'S PREDICTION

- Loose definition of self-explainable AI: Explainability is part of the modelling process

Model layer

- Very active research field
- Current approaches mostly fall into one of these categories:
  - Explainable model architecture: interpretation layers[1] , hierarchical target structure[2]
  - Explanations as part of Model training:
    - Explanation part of annotated data
    - Multi-modal modelling where the different predictions are used to complement each other's plausibility

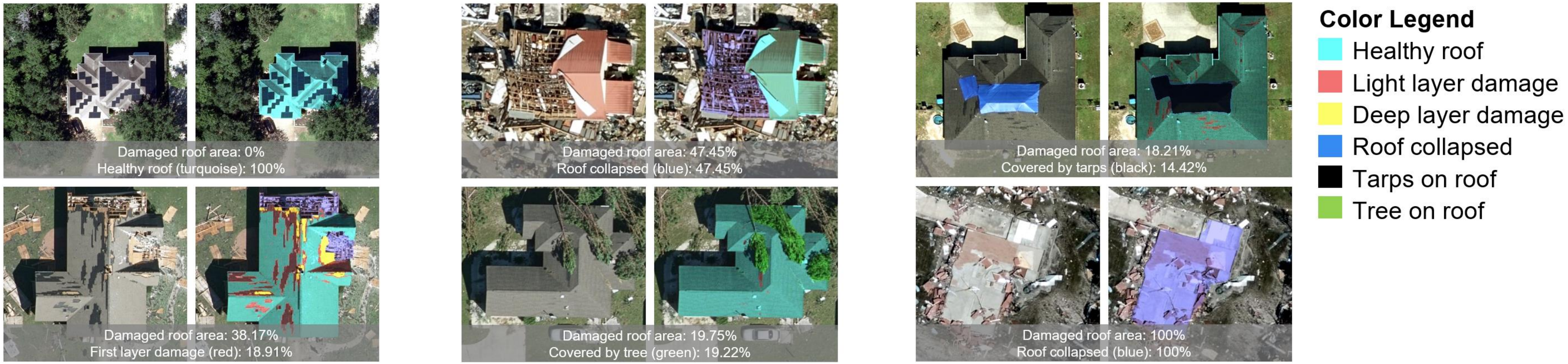
1. Sun, Z., Fan, C., Han, Q., Sun, X., Meng, Y., Wu, F., & Li, J. (2020). Self-Explaining Structures Improve NLP Models. arXiv preprint arXiv:2012.01786. 2. Hase, P., Chen, C., Li, O., & Rudin, C. (2019, October). Interpretable image recognition with hierarchical prototypes. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* (Vol. 7, No. 1, pp. 32-40).

XAI AS PART OF THE TRAINING PROCESS

Use Case: predict the degree of damage of a roof after a natural catastrophe

- The applied convolutional neural net is complex and has 35 mio. Parameters
- However, the problem formulation makes its output well interpretable by design
- Annotated images covering one or more classes with 200-800 images per class (excluding reference classes "background" or "no damage")
- Target categories like "tarps on roof" and "tree on roof" provide an explanation why there is no damage assessment

Local explanation of the model's prediction using self-explaining AI architecture

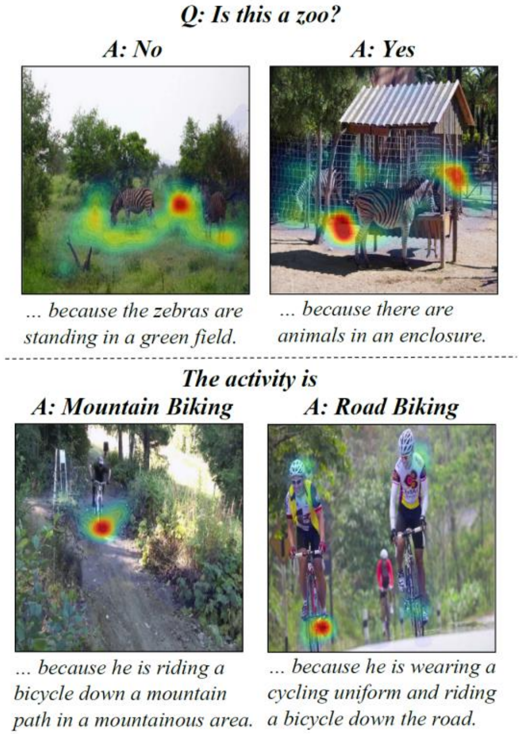


XAI VIA MULTI-MODAL MODELLING

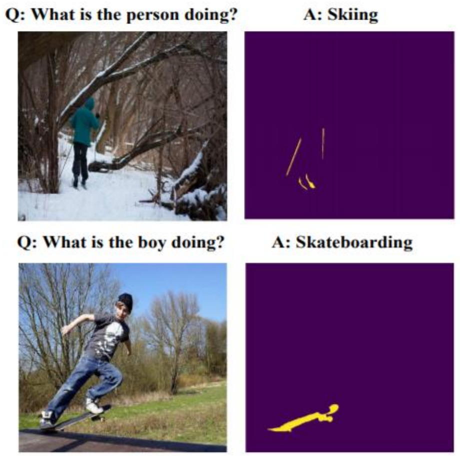
TEXTUAL AND VISUAL EXPLANATIONS COMPLEMENT EACH OTHER

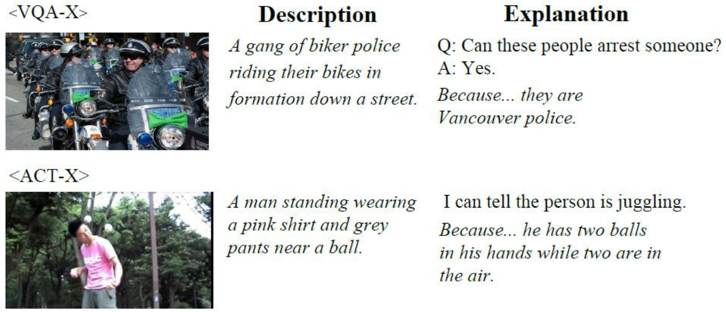
- Examples: visual question answering (QA) and activity recognition
- For each task ~ 50k explanations were made → high effort for annotation
- Only ~<50% of explanations as good as human explanations (human evaluation)

Prediction (=A) enhanced by textual justification and visual highlighting



Annotated explanations for images: visual highlighting (left) and textual QA explanation (right), not only a description





Source: Park, D. H., Hendricks, L. A., Akata, Z., Rohrbach, A., Schiele, B., Darrell, T., & Rohrbach, M. (2018). Multimodal explanations: Justifying decisions and pointing to the evidence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 8779-8788).

- With Munich Re for the past 8 years
- Currently specializing in Natural Language Processing for automatic underwriting & pricing as well as concepts for a responsible application of AI in insurance
- Prior: used machine learning techniques for biometric analysis & best estimate derivation as an actuarial data scientist in life insurance
- PhD in mathematical statistics from the Technical University of Munich
- Lectured a course on life insurance mathematics at TUM
- Developer and maintainer of two Clustering R packages

ABOUT ME

Oliver Pfaffel

Munich Re

EAA e-Conference on
Data Science & Data Ethics

29 June 2021

Thank you for your attention!

Contact

Dr. Oliver Pfaffel
Munich Re

opfaffel@munichre.com