Advantages, properties and limitations of model agnostic explainability techniques in Machine Learning

Workshop ‘Verfahren zur Interpretierbarkeit von neuronalen Netzen’
Invited Talk | June 21, 2022

ZERTIFIZIERTE KI
Qualität sichern. Fortschritt gestalten.
QUANTPI

Outline of today's talk on model agnostic explainability

Today's agenda

1. Introduction to model agnostic XAI algorithms
2. Heatmaps in traffic road sign recognition
3. Properties of explanations
4. Feature importance in anomalous transaction detection
5. Numerical convergence of explanations
6. Conclusion

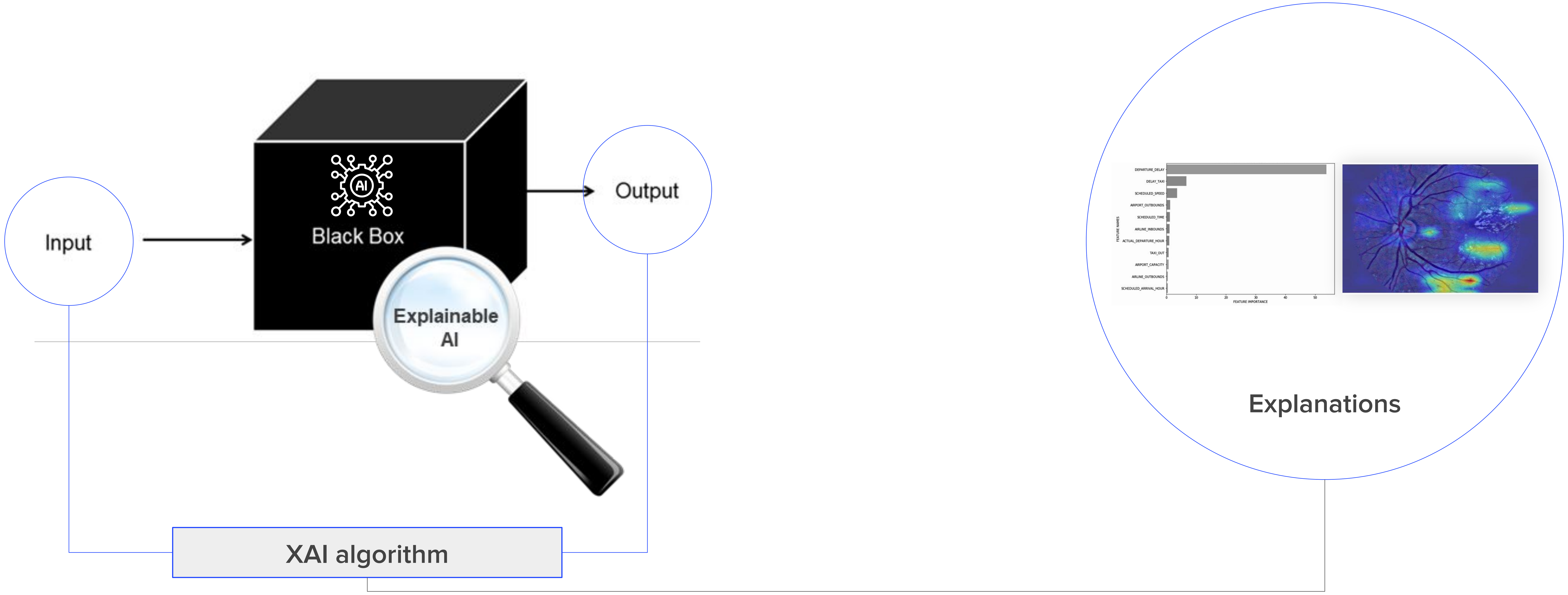
Goals for this talk

1. Introduce you to the properties and limitations of model agnostic XAI algorithms.
2. Discuss their application on tabular and image data.
3. Motivate the importance of calibrating XAI algorithms to generate stable explanations.

Dr. Antoine Gautier | CRO
antoine.gautier@quantpi.com

Principle of model agnostic XAI algorithms

Model agnostic XAI algorithms
Explanations are produced by perturbing inputs and comparing outputs.

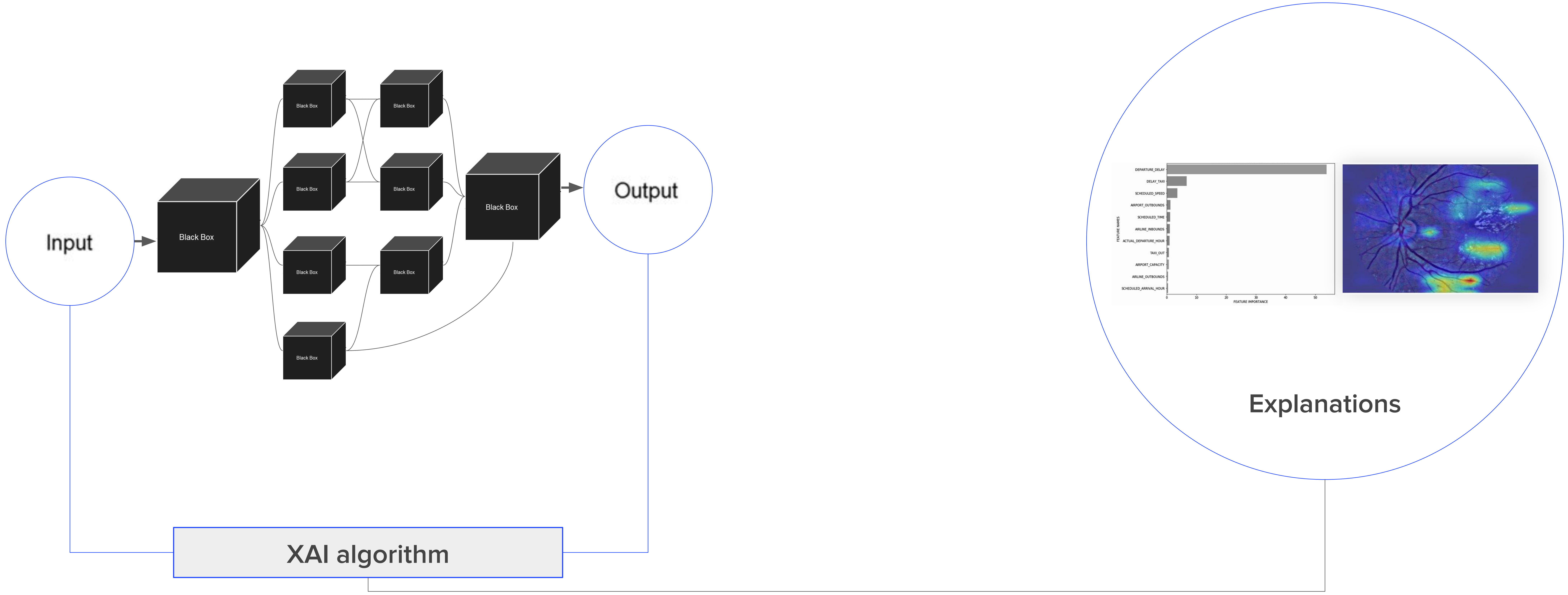


Dr. Antoine Gautier | CRO
antoine.gautier@quantpi.com

Model agnostic XAI algorithms can be applied to any callable black-box

Model agnostic XAI algorithms

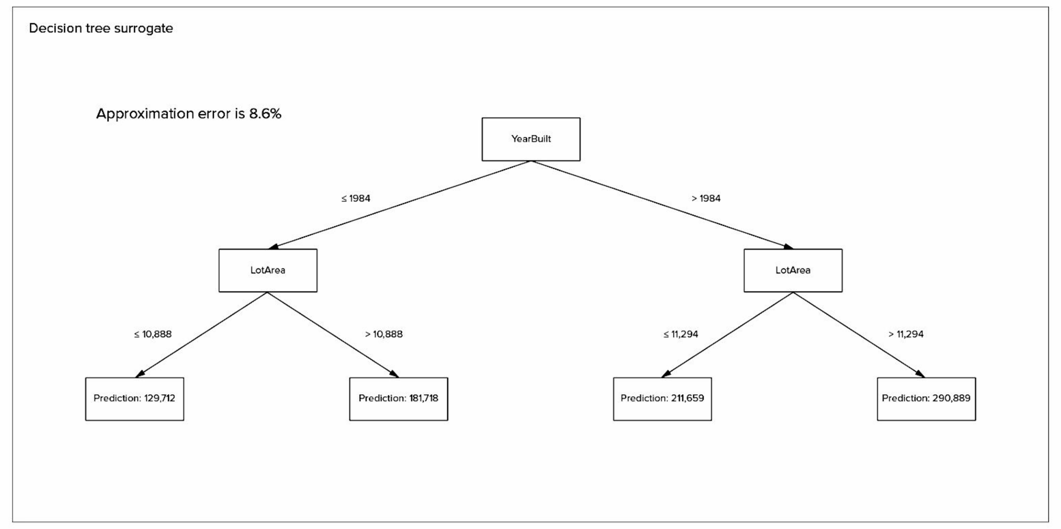
Explanations are produced by perturbing inputs and comparing outputs.



Four big families of model agnostic XAI approaches

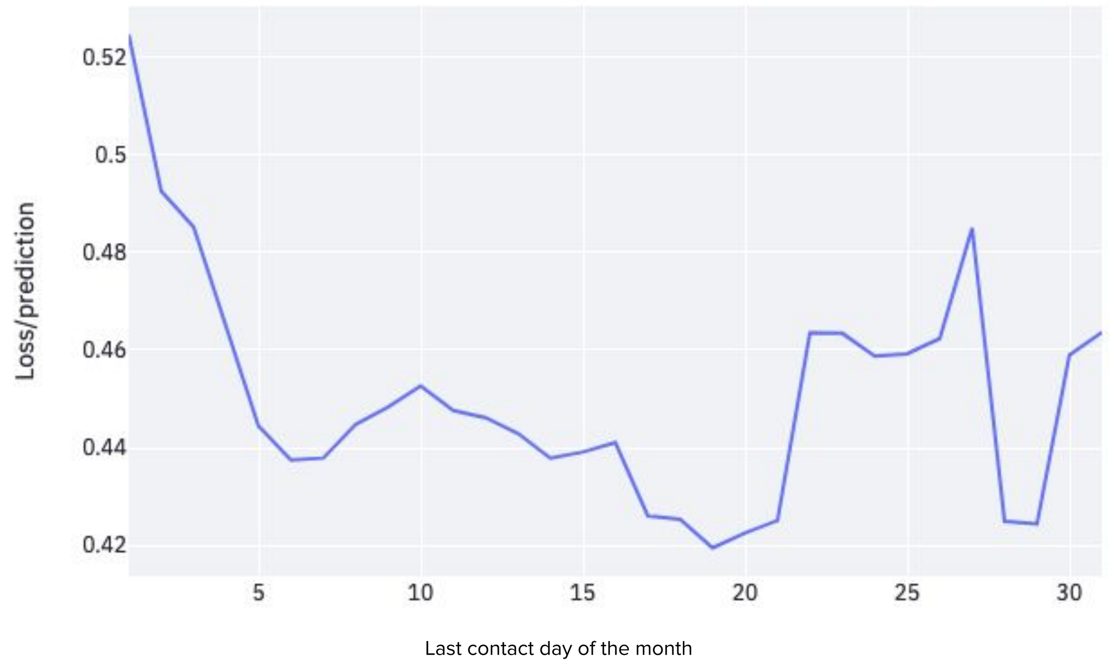
Transparent surrogates

- LIME
- Decision rules
- …



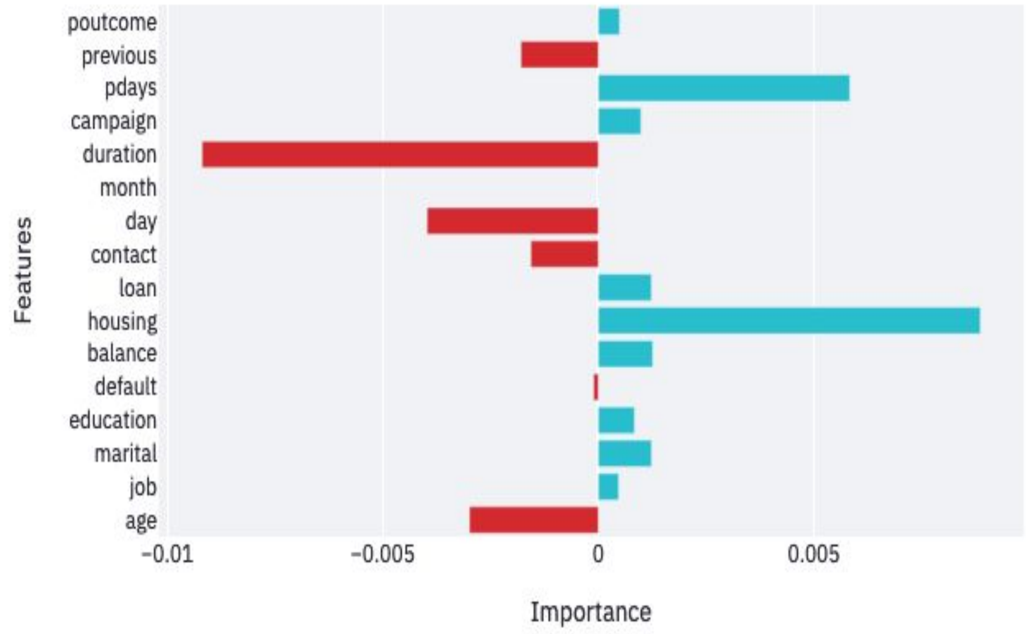
Generalized sensitivity

- PD-plots and M-plots
- Sensitivity analysis
- …



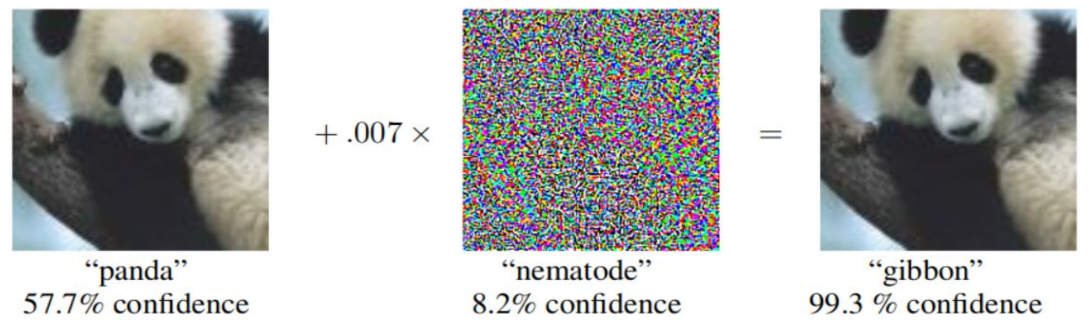
Feature importance

- Shapley values
- Model reliance
- …



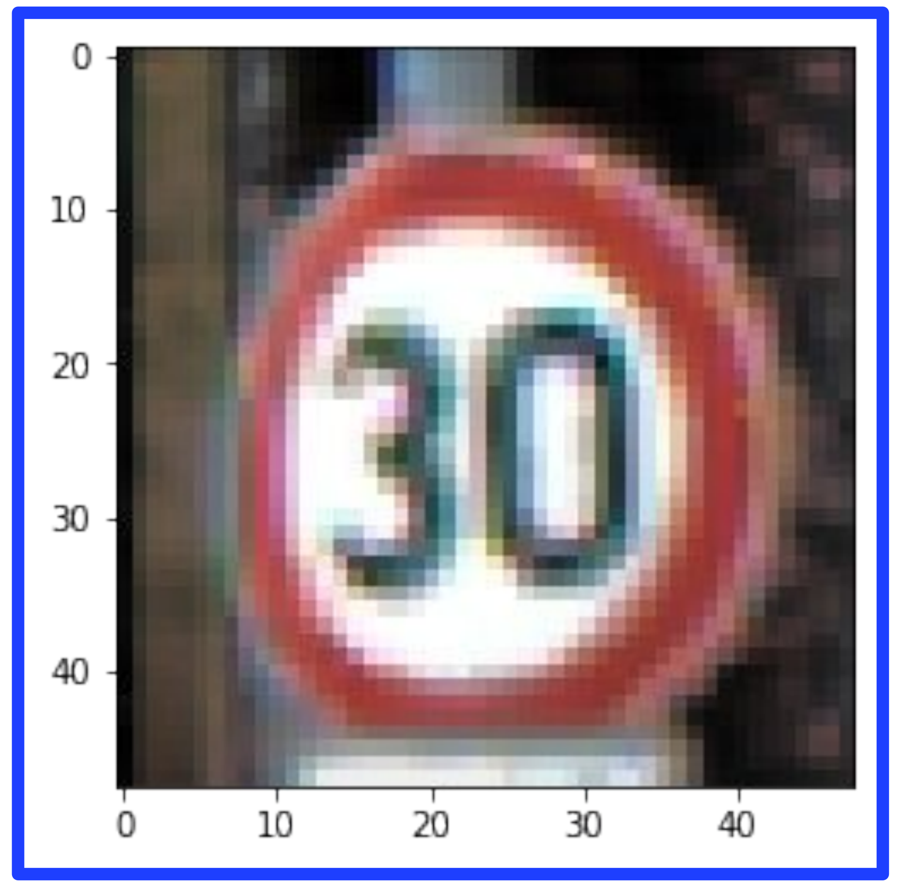
Perturbations

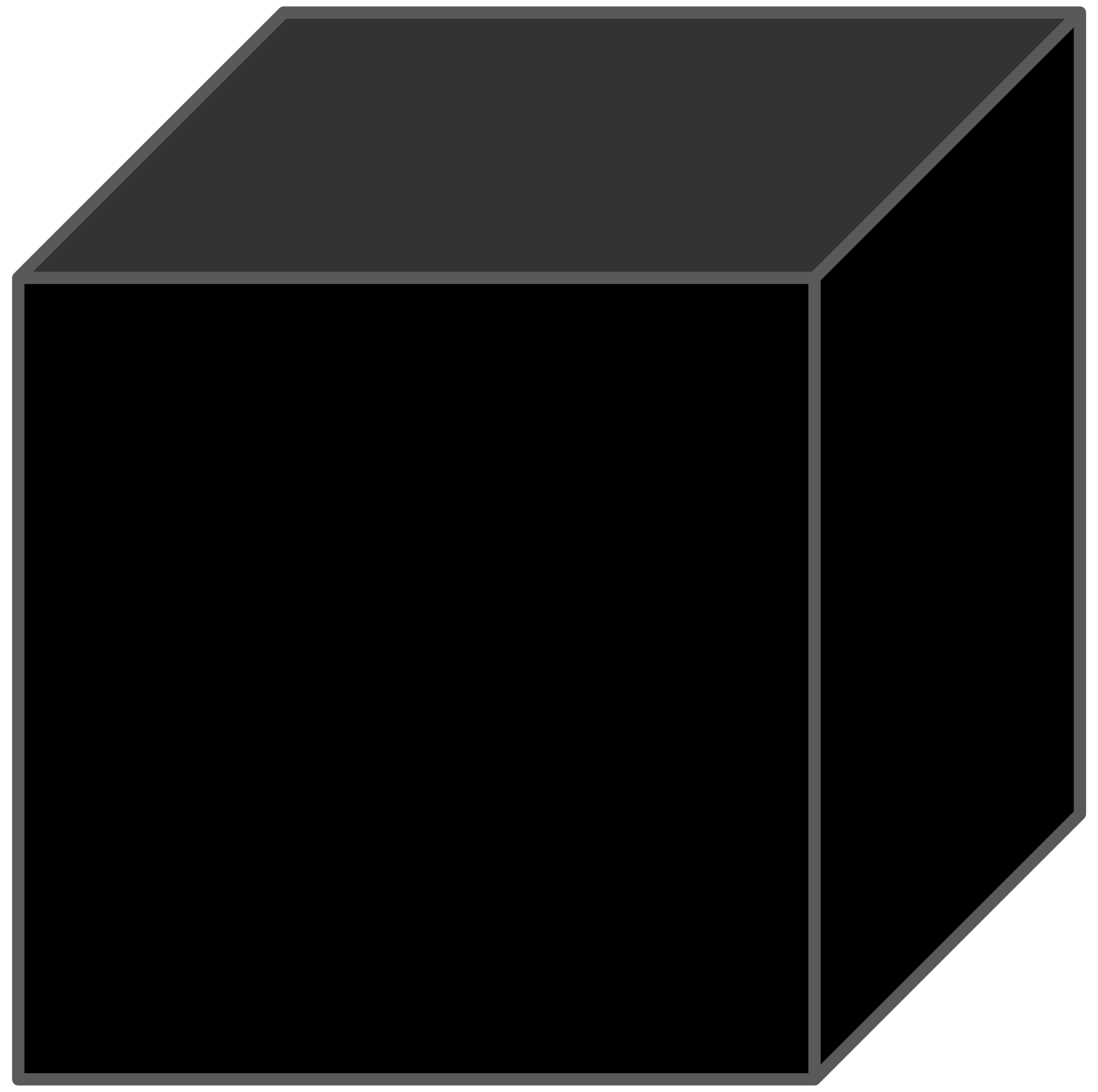
- Counter-factual
- Adversarial perturbations
- …



Example I: Convolutional neural network for road sign recognition

Traffic road sign recognition





Class	Value
Speed limit 30	0.99
Speed limit 80	0.00 ... 1
...	...

Input
Image of traffic road sign

Model
Black box

Output
Probability score for 43 road signs

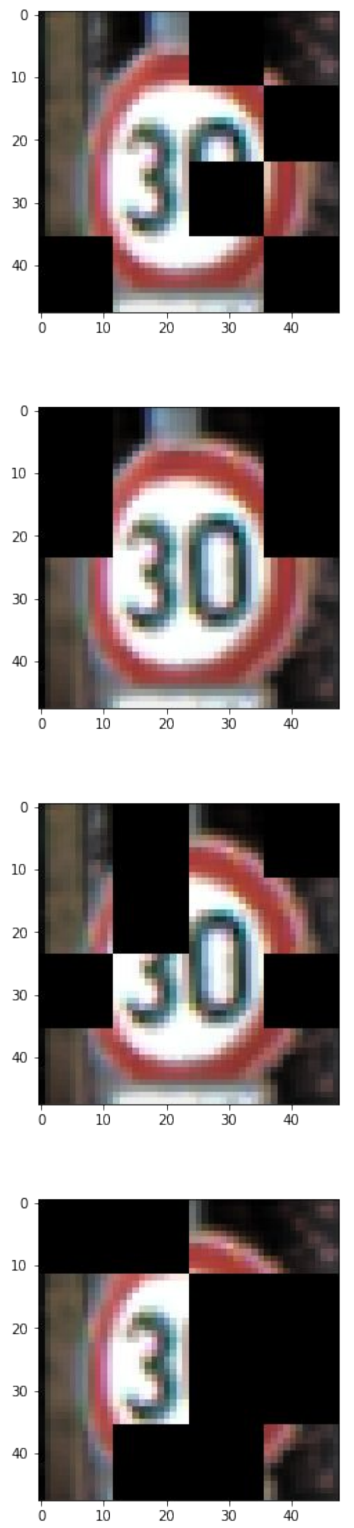
Dataset
German traffic road signs (resized)

Tested model
AlexNet (5 conv. layers)

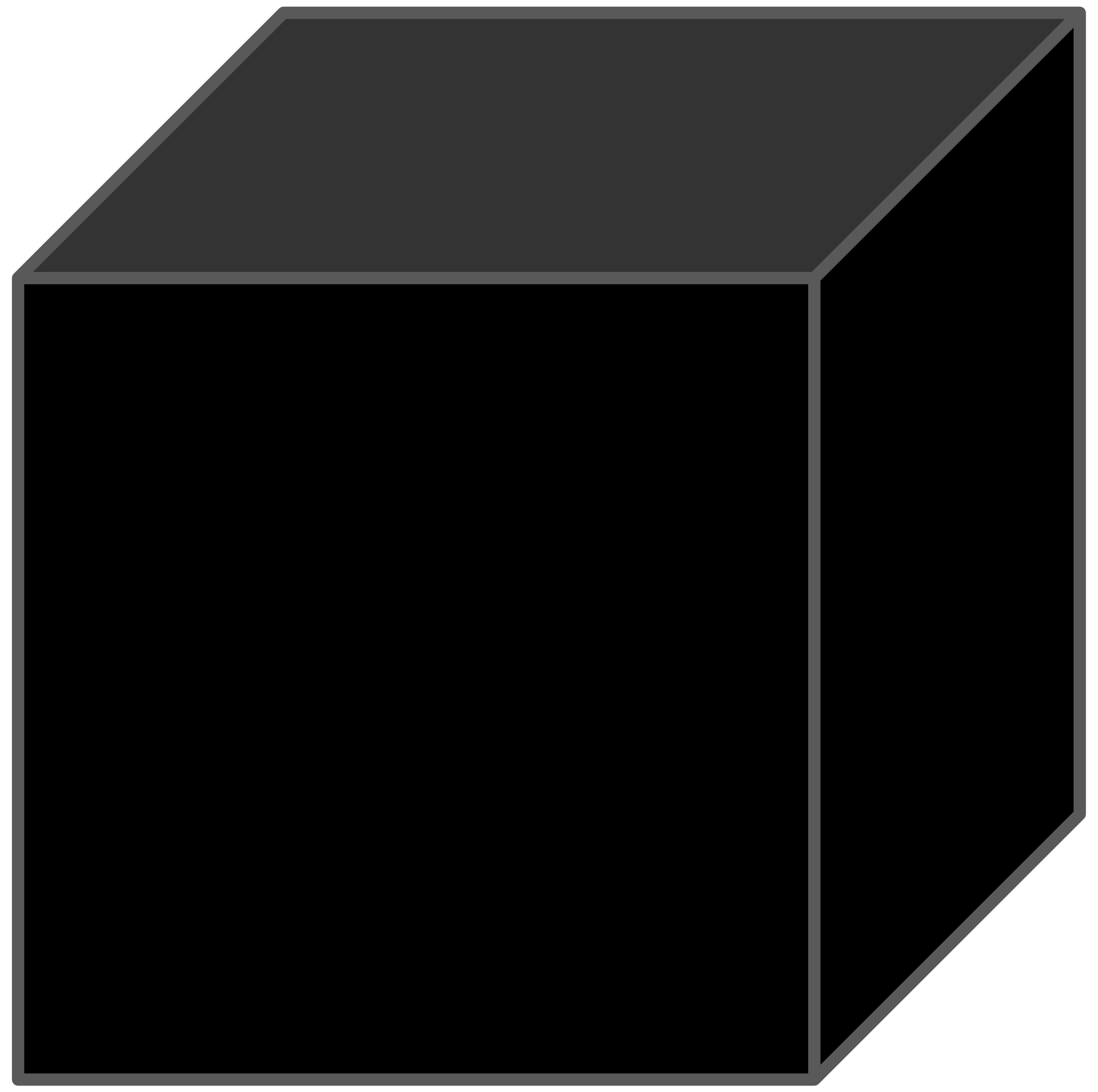
Model access
Confidence scores with probability for each class

Process of a model agnostic XAI technique in the road sign recognition example

1. Perturb input | **2. Query model** | **3. Compare outputs**



...



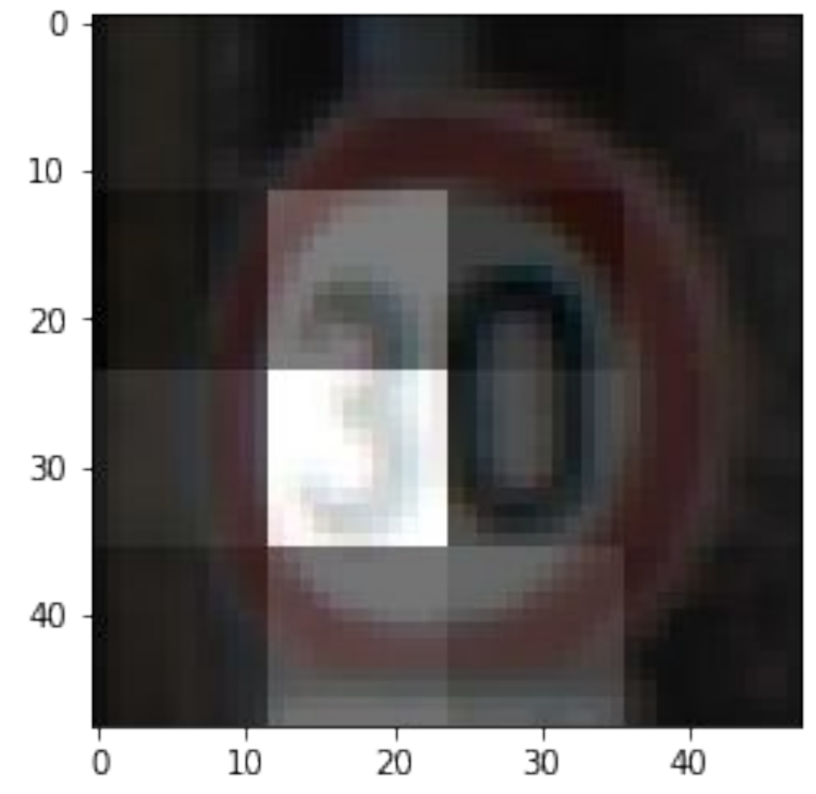
Black box model

Predicted values
0.31
0.14
0.15
0.92
0.65
0.35
...

Explanation

0,07	0,06	0,06	0,08
0,02	0,3	0,01	0,09
0,14	1,0	0,08	0,09
0,08	0,27	0,2	0,06

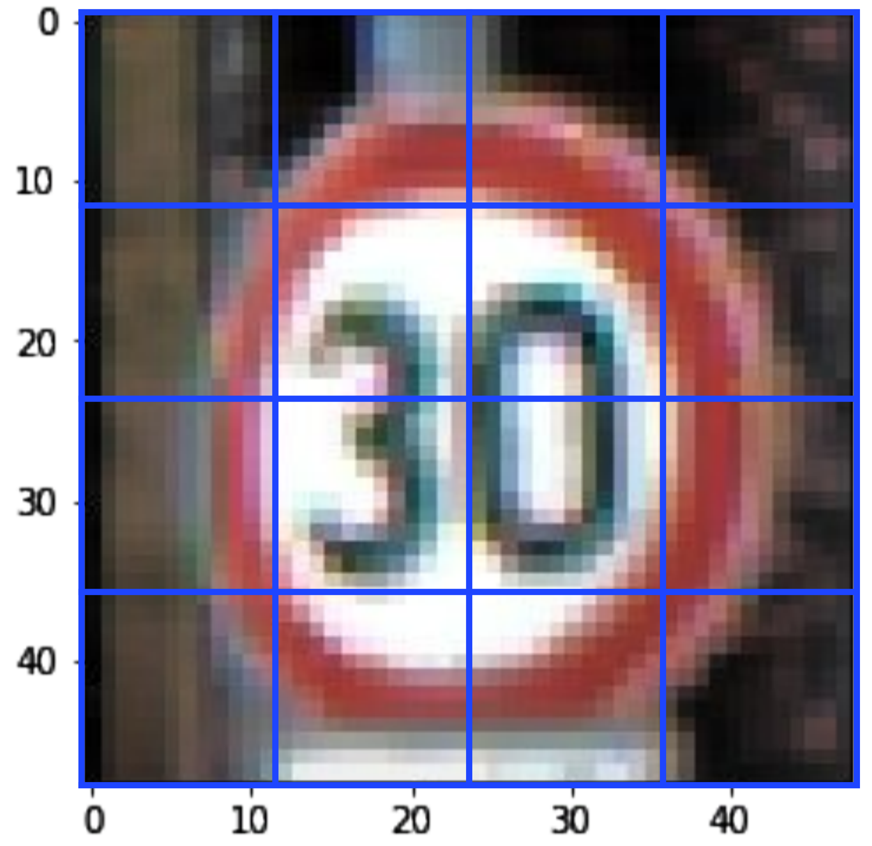
Visualisation



Explanations are useful to verify absence of biases and increase trust

Explanation on images can identify the most important regions for prediction

Explanation for “Speed limit 30” with a 4x4 grid



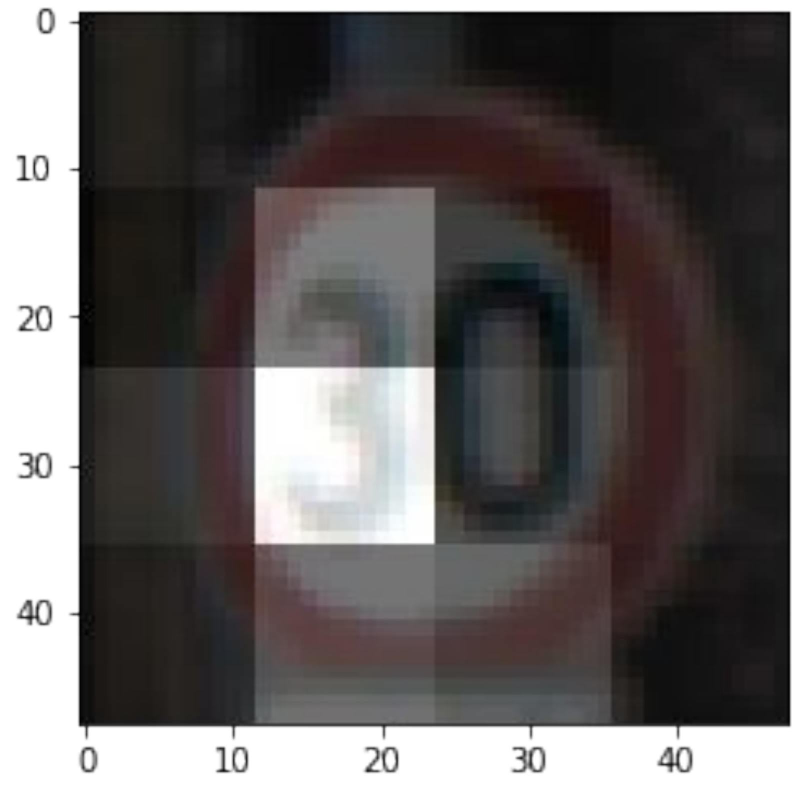
Segmentation

Split the image to be explained into regions

0,07	0,06	0,06	0,08
0,02	0,3	0,01	0,09
0,14	1,0	0,08	0,09
0,08	0,27	0,2	0,06

Explanation

- Important regions have large score
- Not important regions have low score



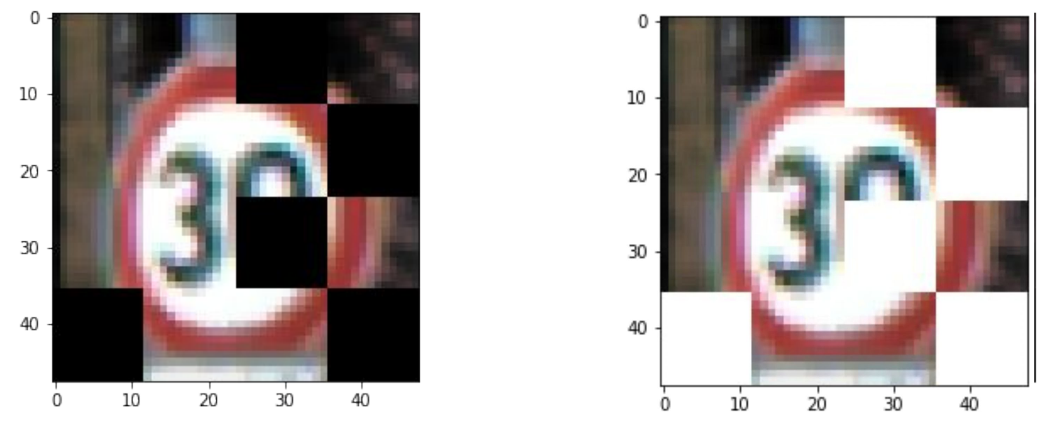
Visualising of the explanation

- White regions are important for the prediction
- Black regions are not important for the prediction

Dr. Antoine Gautier | CRO
antoine.gautier@quantpi.com

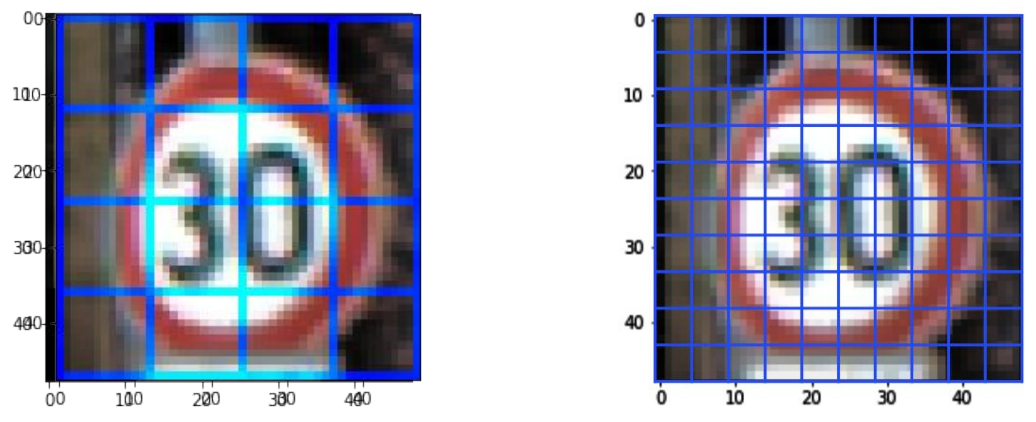
Explanations have four main hyper-paremeters that need to be carefuly selected

Choice of baseline



Different ways to mask the image produce different explanations

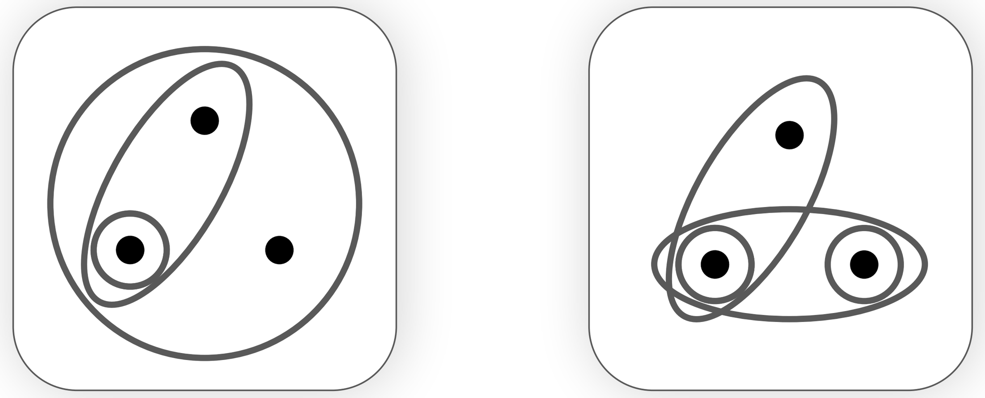
Choice of regions



Different segmentations of the image produce different explanations

Parametrize the perturbation of inputs

Choice of explainer



Different ways to combine outputs produce different explanations

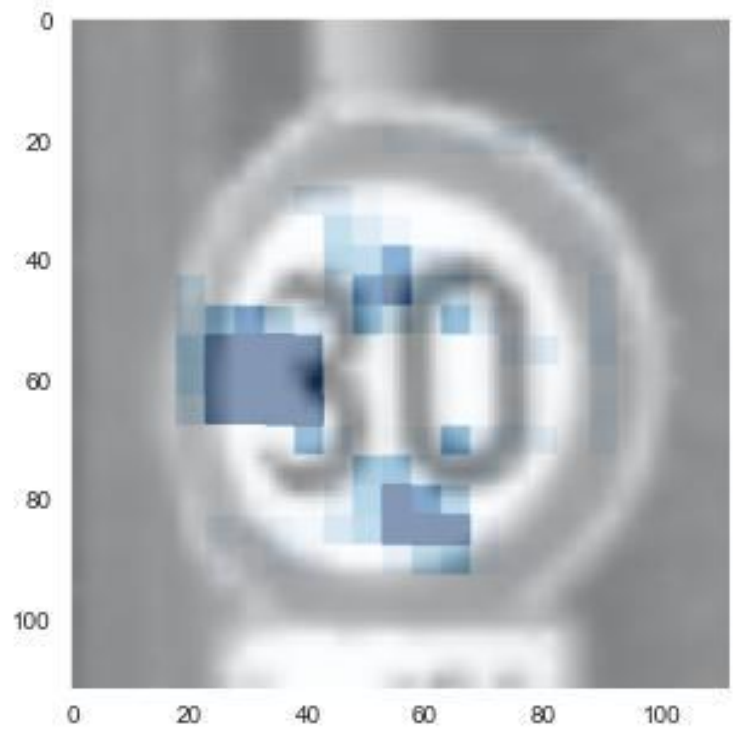
Choice of metric

Output
Prob. of Speed limit 30
Prob. of Speed limit 80
Cross-entropy error
...

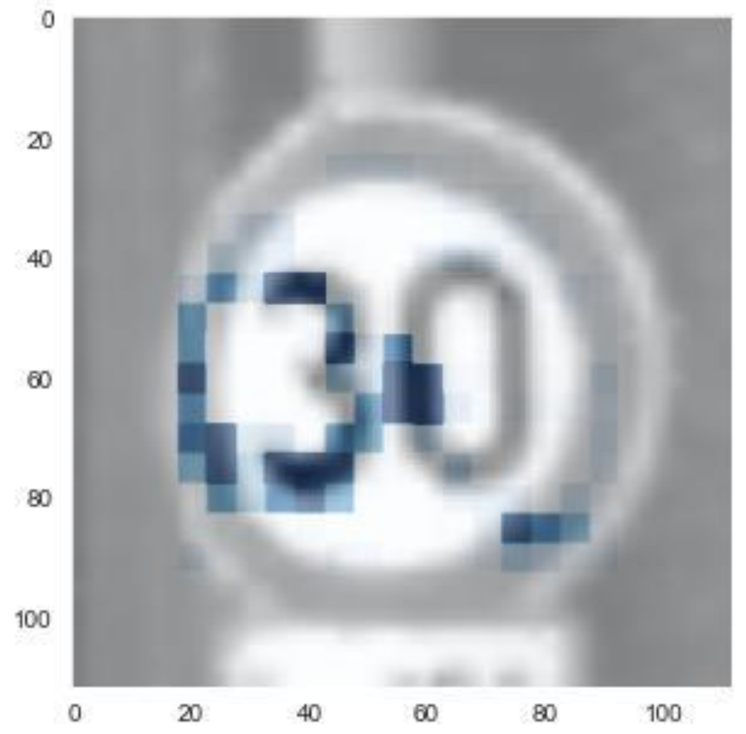
Different output types produce different explanations

Parametrize the comparison of outputs

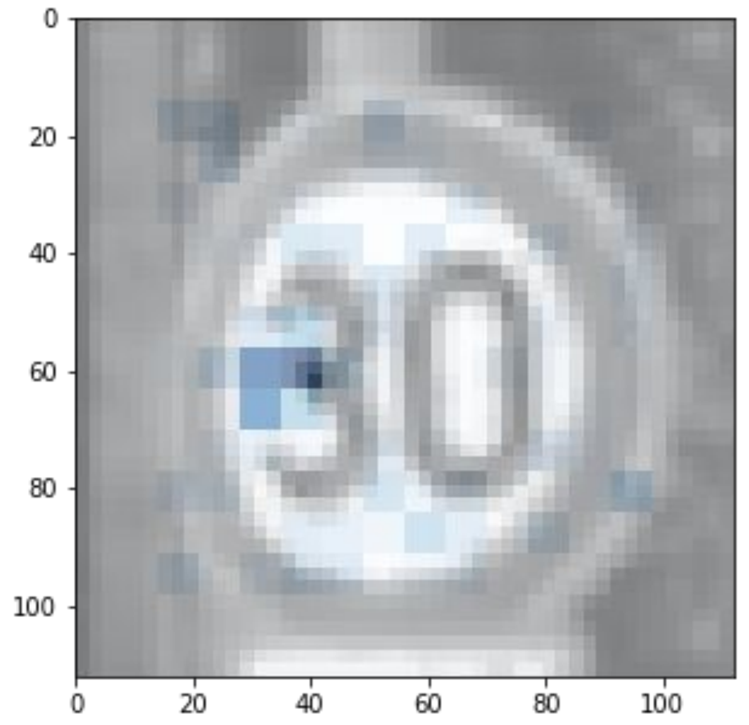
Examples of explanations for different choices of hyper-parameters



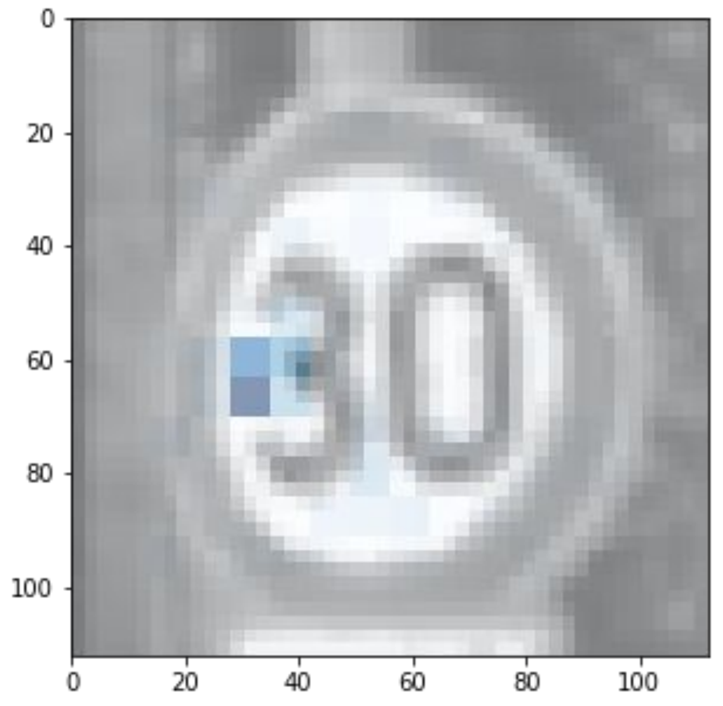
"30" sign with black baseline



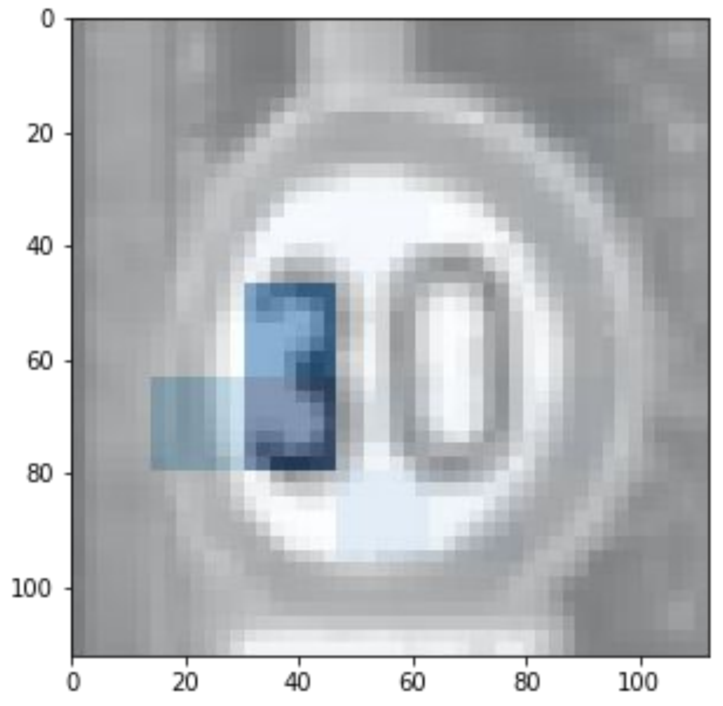
Change the baseline to white



Change the XAI method

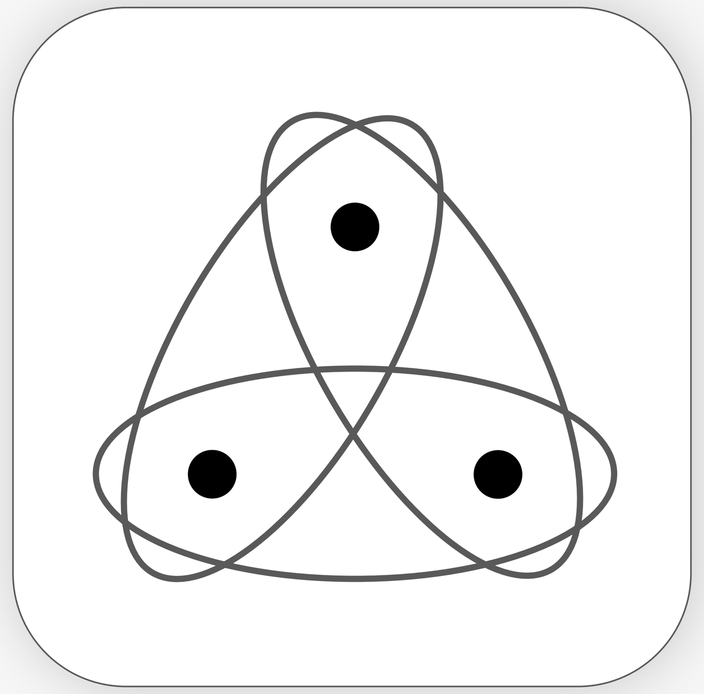


Change the metric



Change the grid size

Explanation generated by “removing” a single region

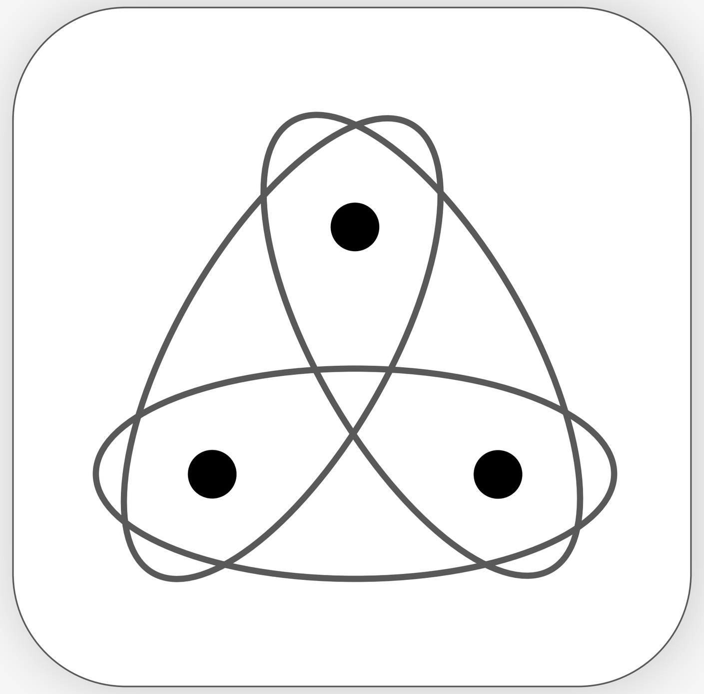


Key concept:
The importance of a region is given by the difference in prediction when it is hidden.

Example of execution

Dr. Antoine Gautier | CRO
antoine.gautier@quantpi.com

Properties of explanations generated by “removing” a single region



Key concept:
The importance of a region is given by the difference in prediction when it is hidden.

Advantages and properties

<u>Advantages:</u>

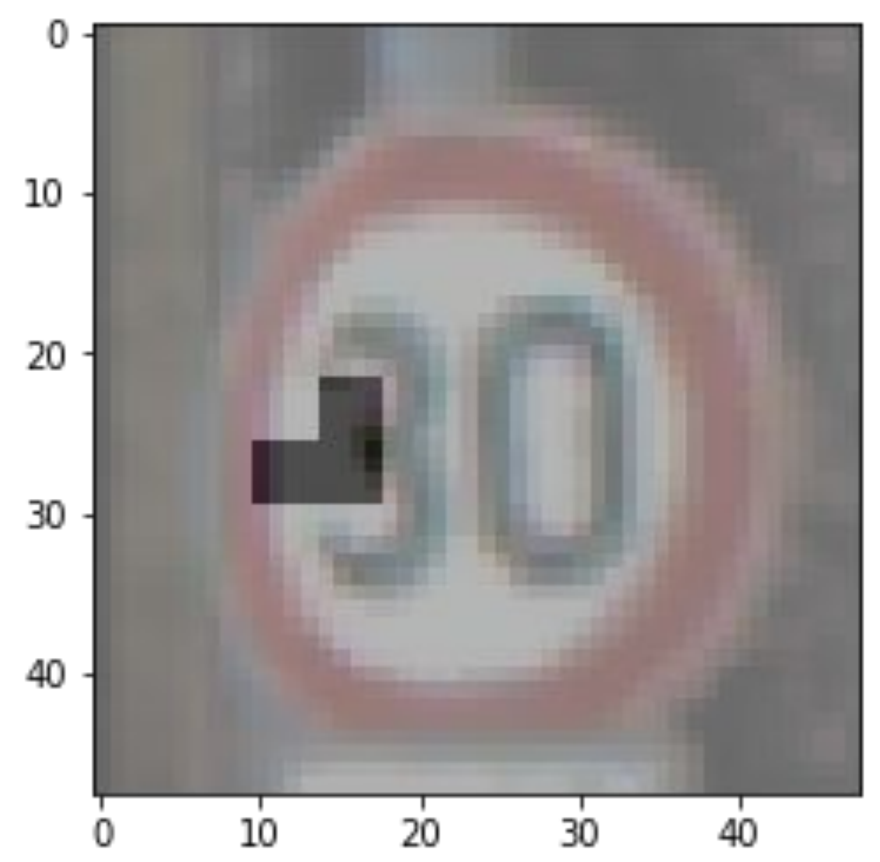
- Easy to interpret
- Low computational cost per explanation
  # model queries = 1 + # cells in the grid

<u>Properties:</u>

- Dummy: Regions unused for predictions have score 0
- Symmetry: Regions equally impacting features have the same score
- Addivity: Explanations of sum of black equal the sum of the explanations

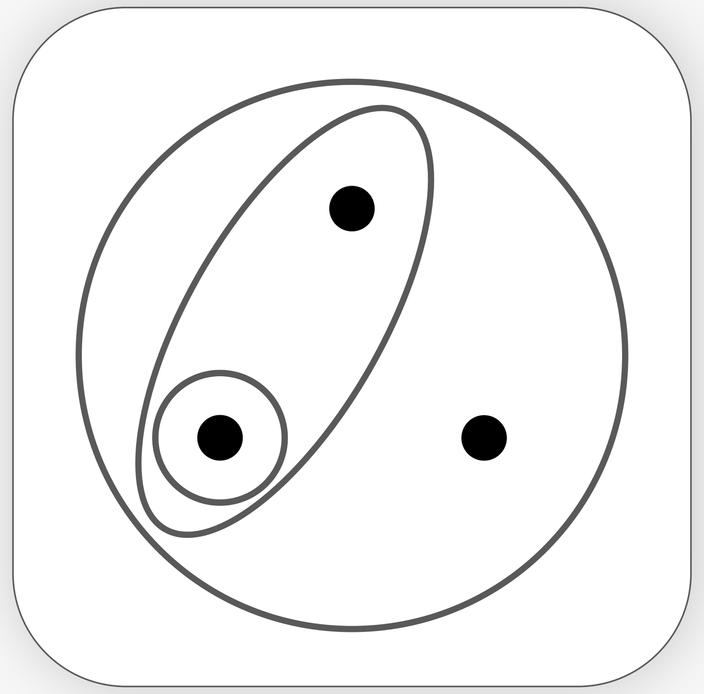
Limitations

- Only considers one type of synergies between the regions
- May be inappropriate for grids with small regions



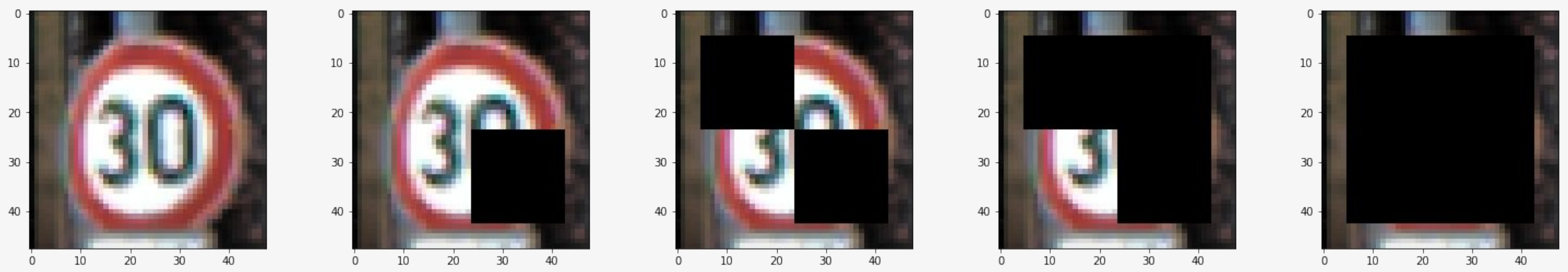
Dr. Antoine Gautier | CRO
antoine.gautier@quantpi.com

Scoring based on orderings of regions

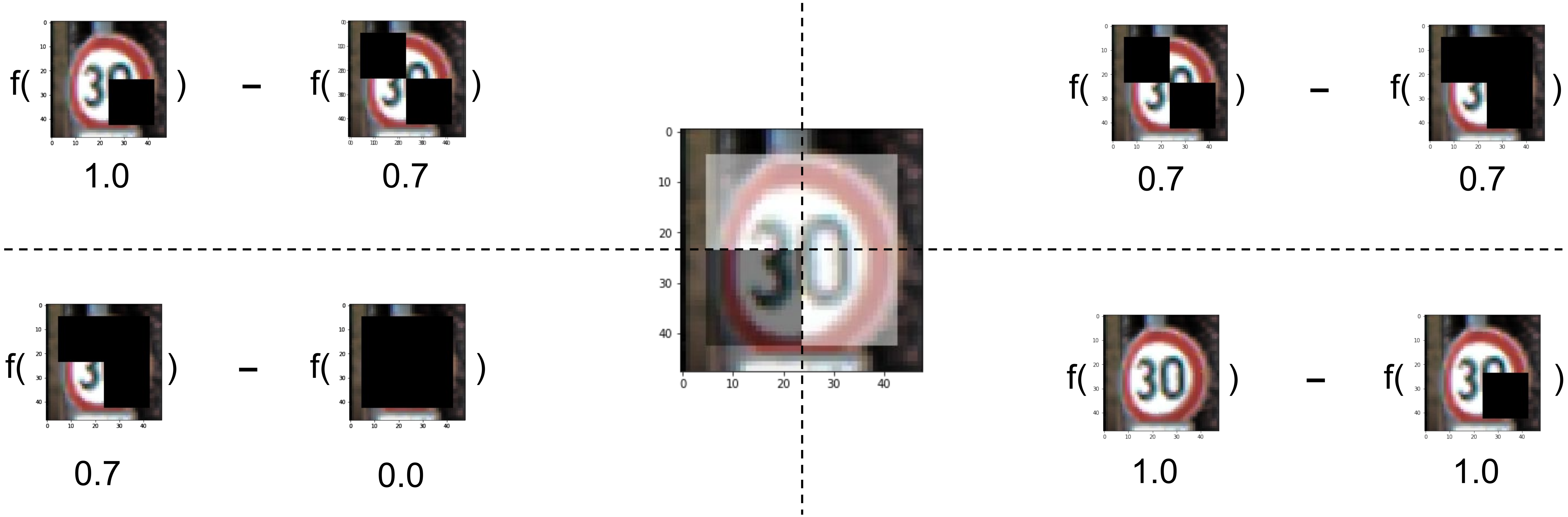


Key concept:
The importance score of a region is computed by hiding regions one after the other in a particular order.

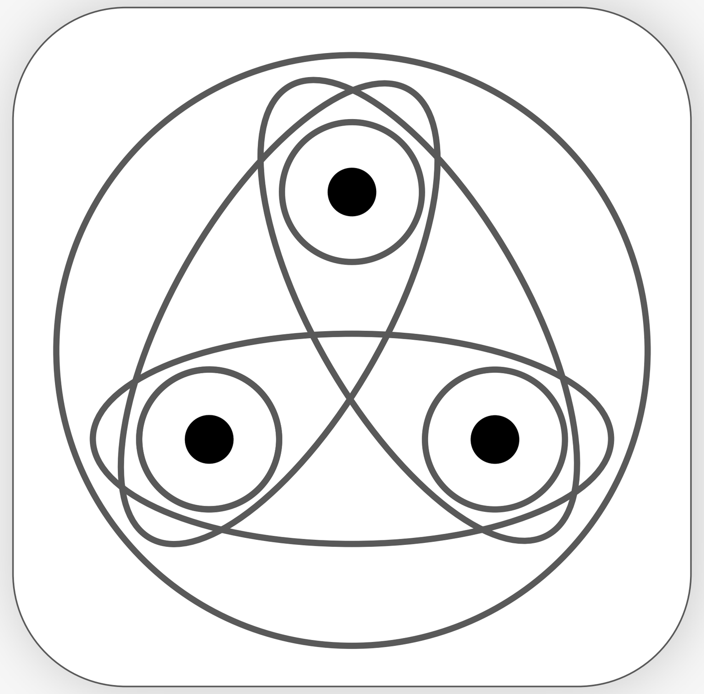
Example of hiding procedure given an ordering



Example of execution



Shapley values explanation



Key concept:
The importance of a region is computed by averaging the scores on every possible order of the regions.

Average marginal contribution for all permutations

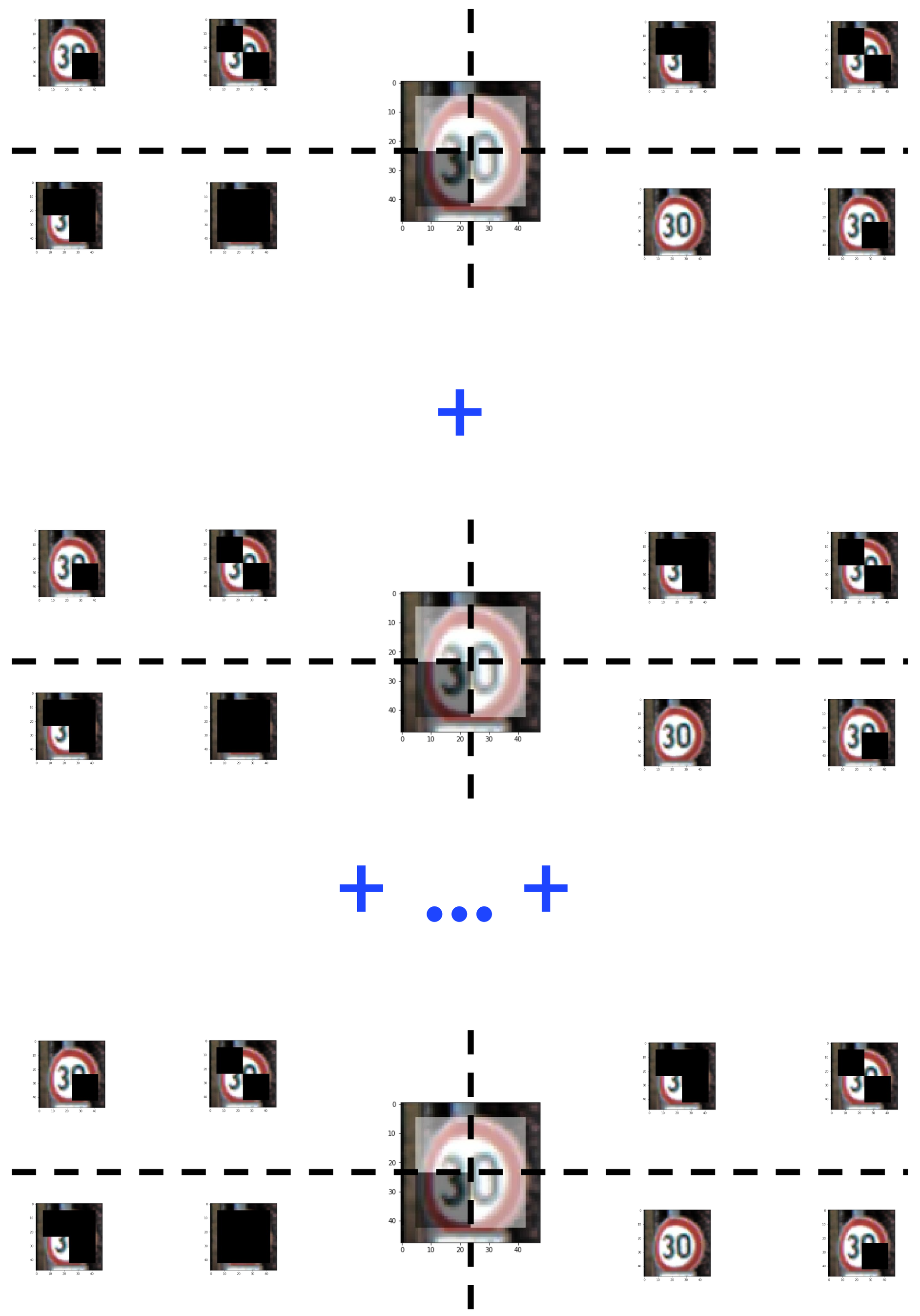
Permutation 1

+

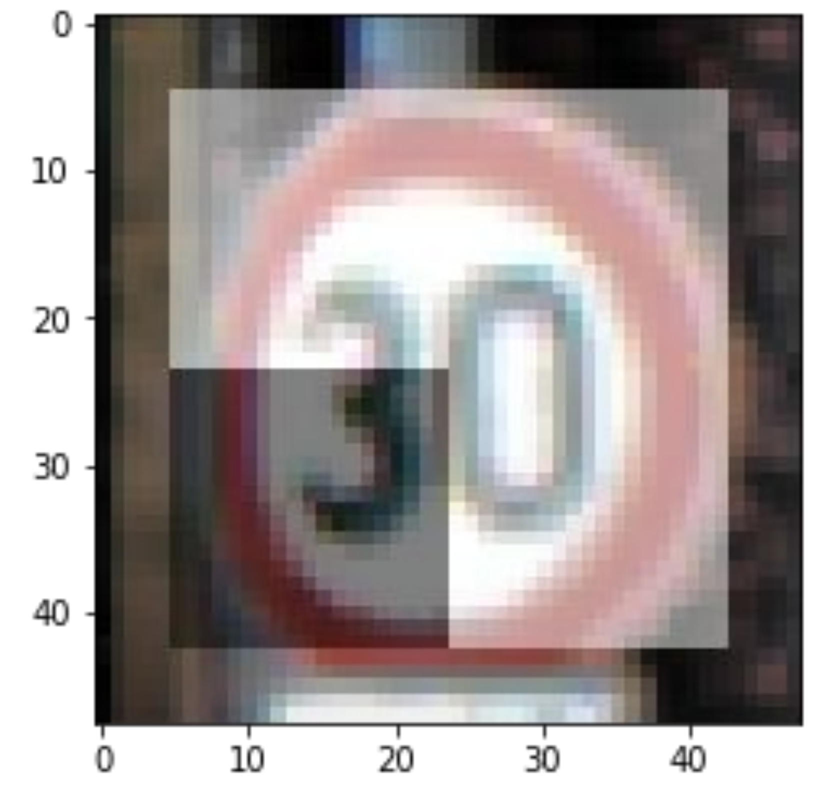
Permutation 2

+ ... +

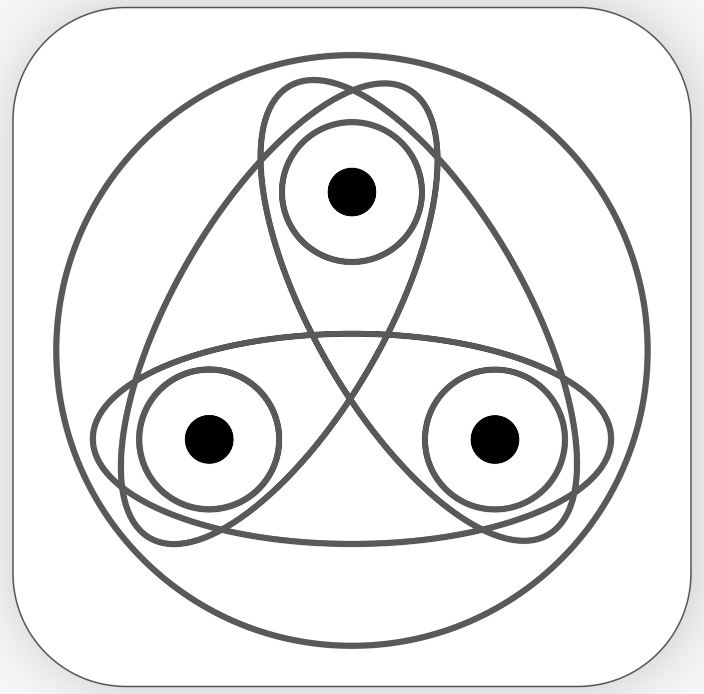
Permutation N!



=



Properties of Shapley values explanations



Key concept:
The importance of a region is given by the difference in prediction when it is hidden.

Advantages and properties

Advantages:

- Consider all possible synergies between regions

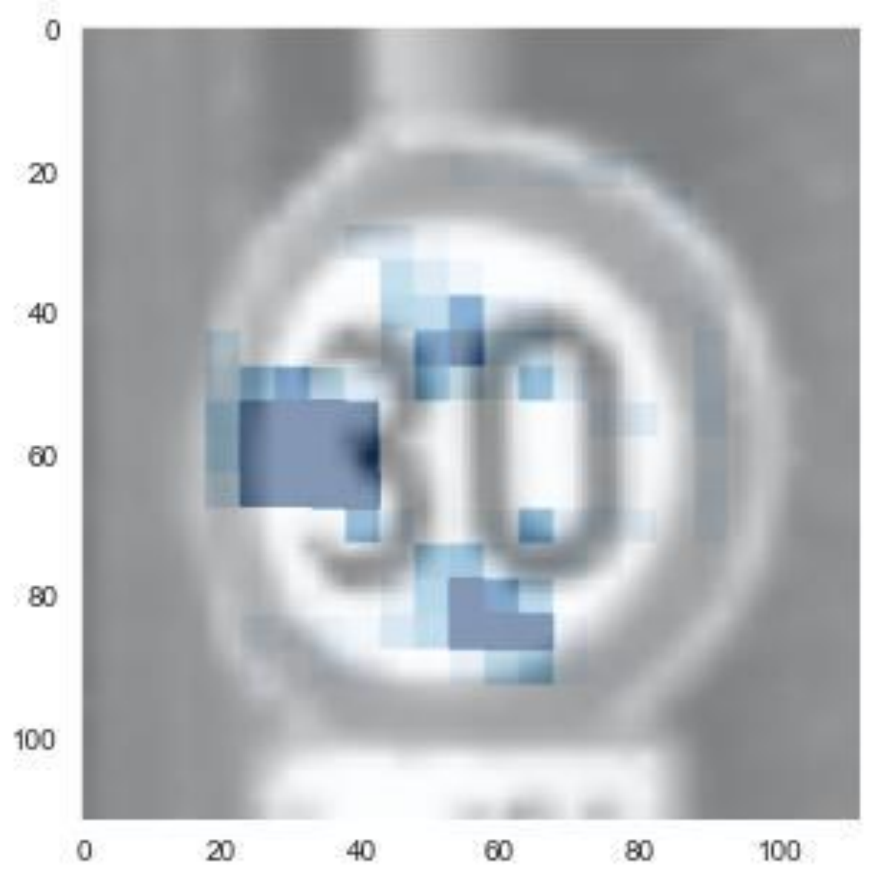
Properties:

- Dummy: Regions unused for predictions have score 0
- Symmetry: Regions equally impacting features have the same score
- Addivity: Explanation of sum of black boxes equal the sum of the explanations of each
- Efficiency: The score of all regions sum to the value of the prediction

Limitations

- Require more technical understanding for interpretation
- High computational cost per explanation
  # model queries = (# cells in the grid)!

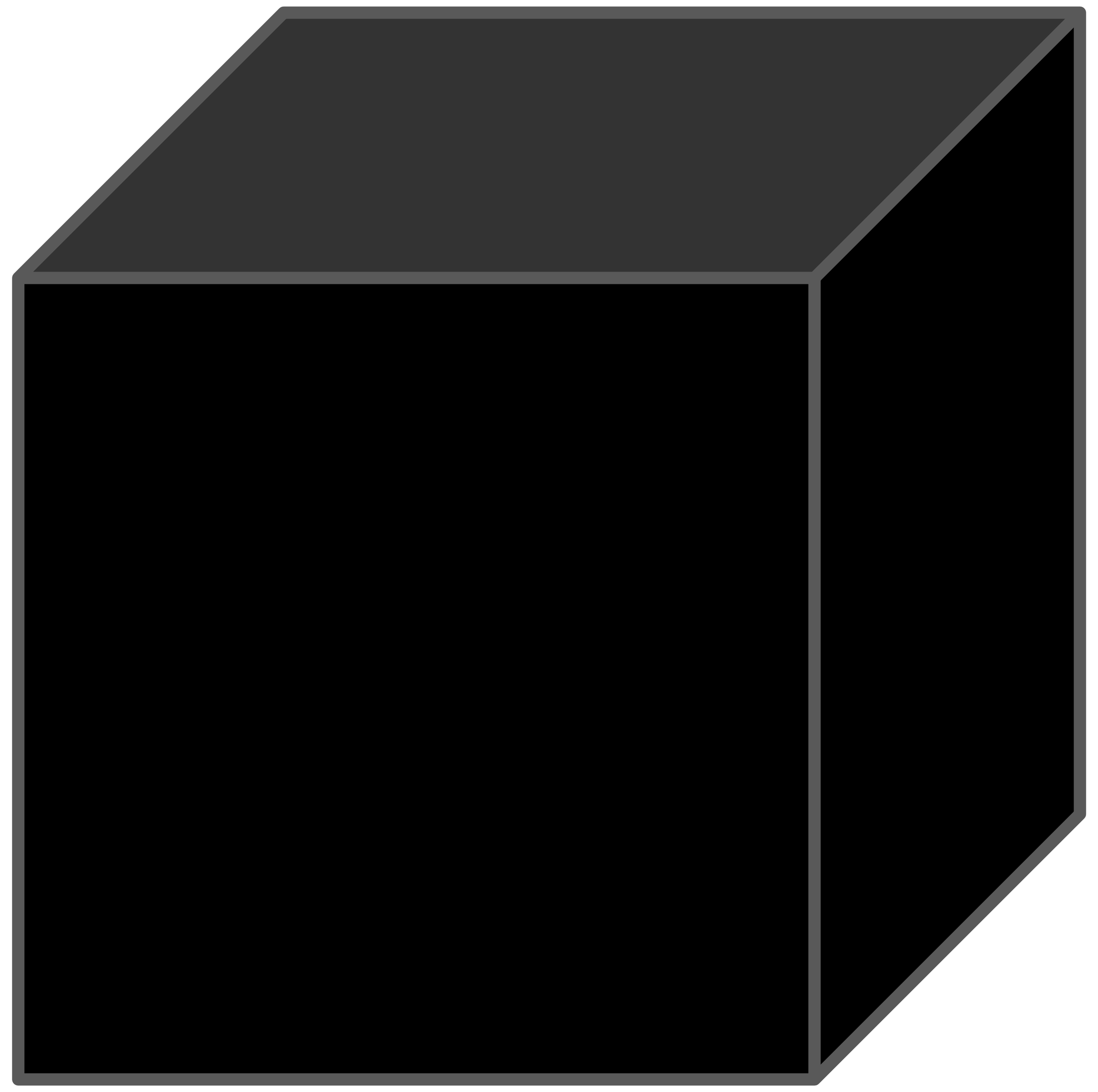
  => In practice, only compute approximations



Example 2: Deep auto-encoder on transactions in ERP data

Fraud detection in enterprise resource planning (ERP) systems

Feature	Value
Currency:	€
Amount:	500
...	...
Type:	office supply



Anomaly score
0.314

Input
Transaction

Model
Black box

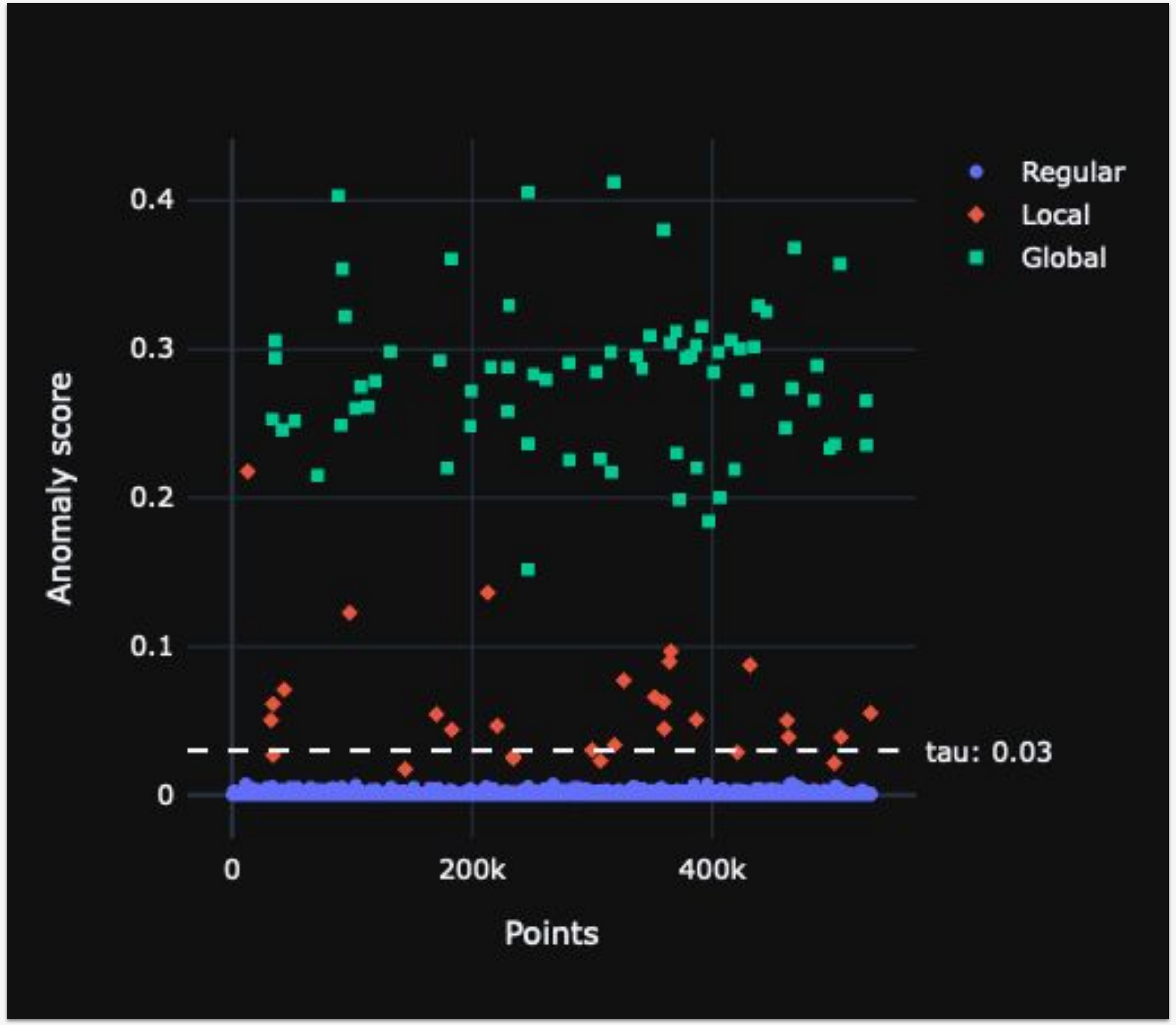
Output: Anomaly score
- High score implies transaction is anomalous
- Low score implies transaction is regular

Dataset
German traffic road signs (resized)

Model
Deep auto-encoder (18 hidden layers)

Model output
Anomaly score = Reconstruction error

Anomaly detection in ERP system



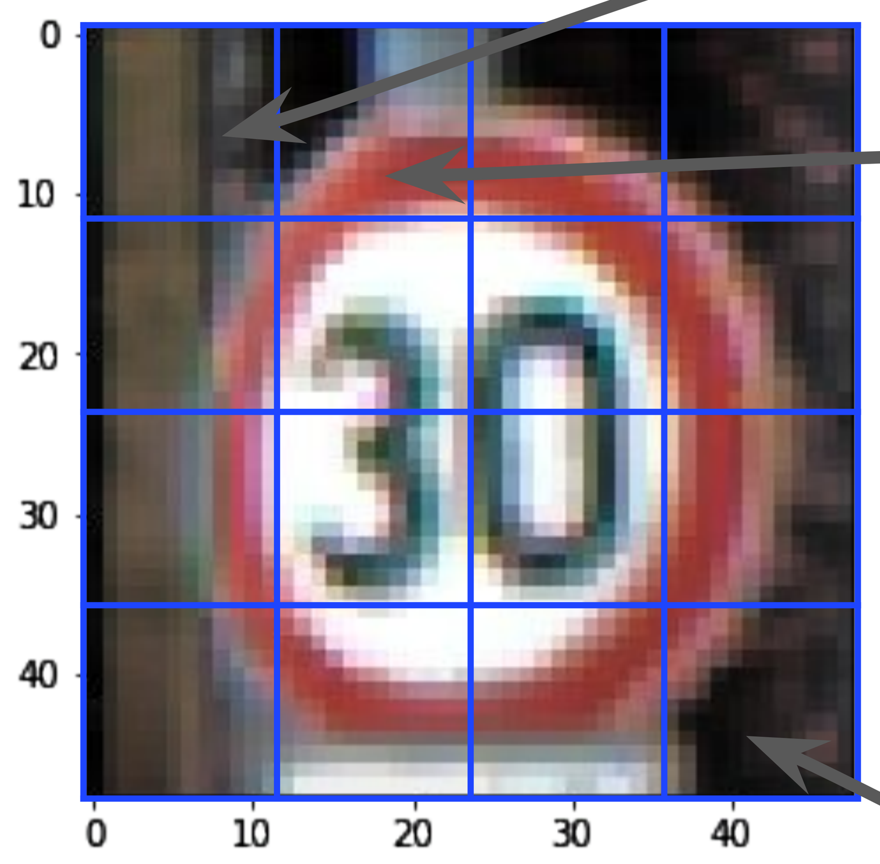
Anomaly detection

Anomaly score computed on 500k points

- **Regular transaction.**
  Normal transaction without anomaly.
- **Global anomalies.**
  Typically large difference in few features
  E.g. typing mistake, measurement failure, etc.
- **Local anomalies.**
  Deviates from joint feature distribution.
  Suspicious case requiring deeper investigation.

XAI modelling of image and tabular inputs is similar

The cells in the grid on the image can be identified with features in tabular data



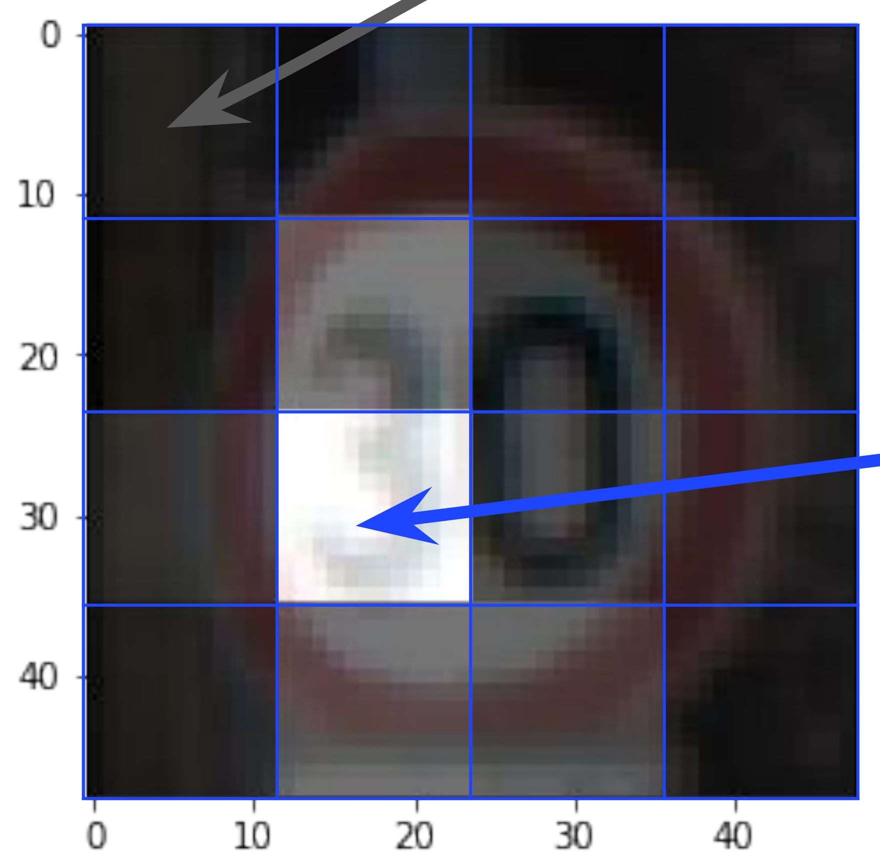
...

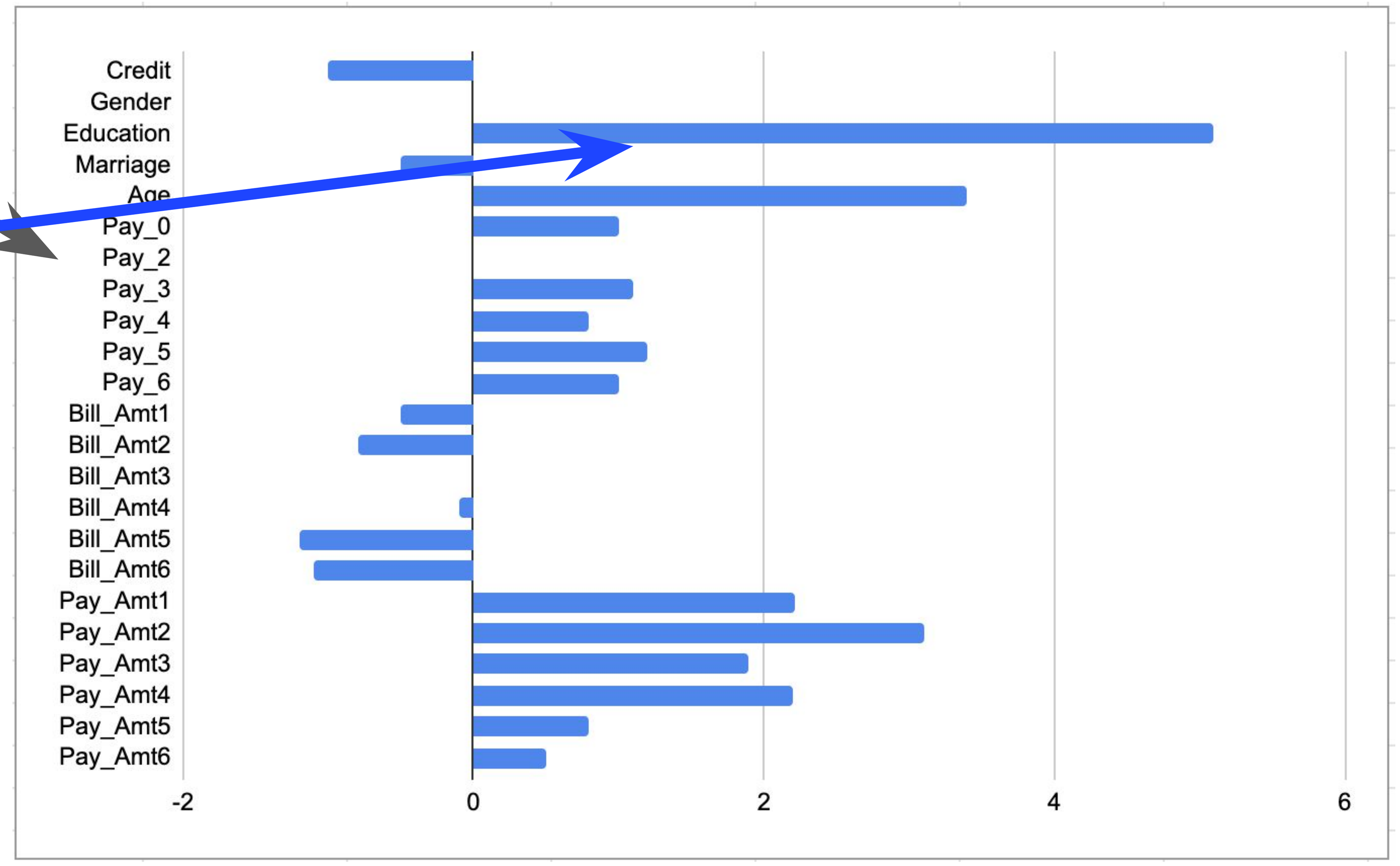
Feature	Value
Currency:	€
Amount:	500
...	...
Type:	office supply

Dr. Antoine Gautier | CRO
antoine.gautier@quantpi.com

XAI representations of image and tabular explanations can be interpreted similarly

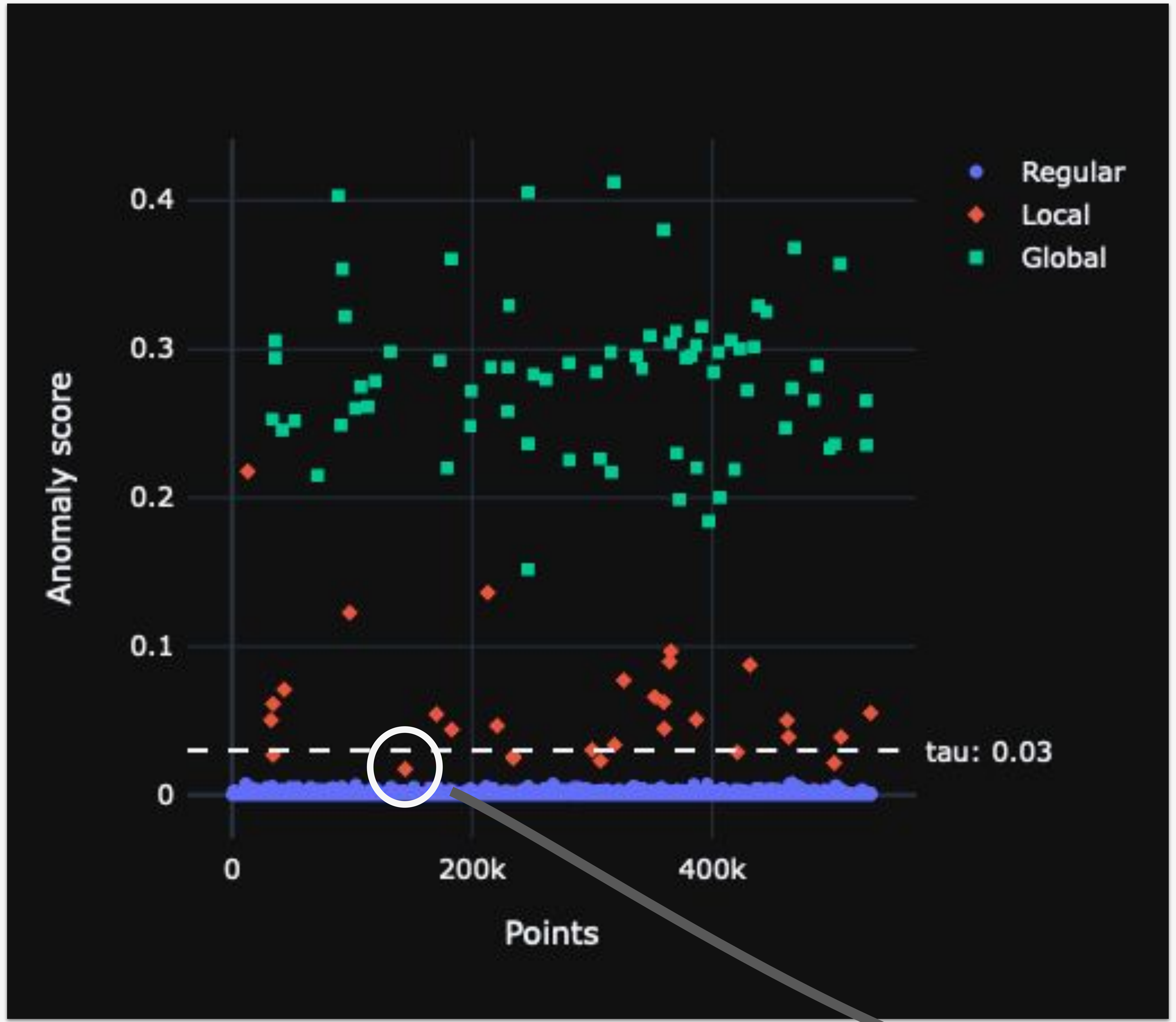
An important region (bright cell) is similar to an important feature (long bar)





An unimportant region (dark cell) is similar to an unimportant feature (small bar)

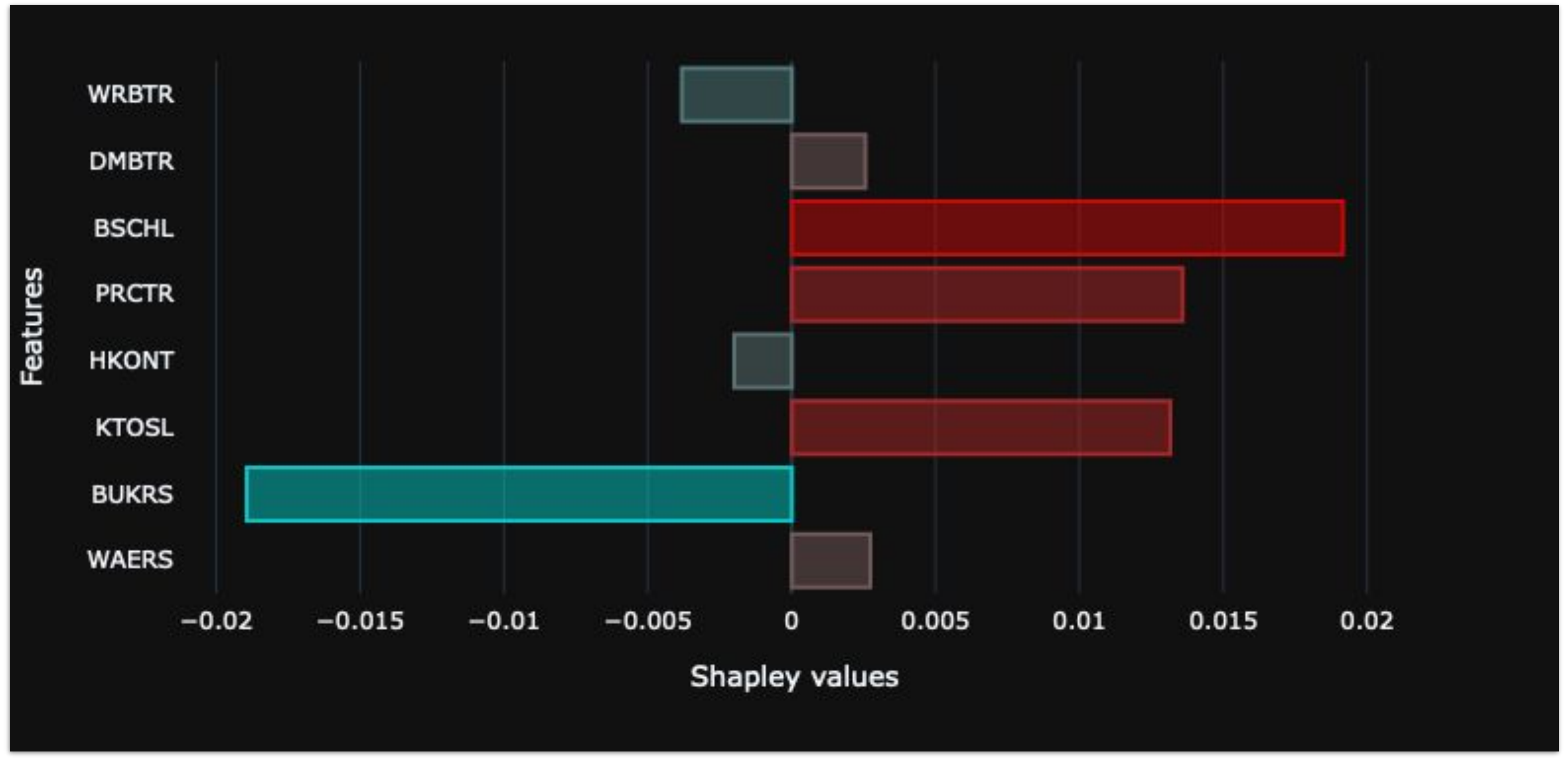
Explanation of a single transaction



Explanation of the anomaly score of a single transaction

Explanation for the prediction of one particular transaction:

- Red bars indicate that the feature increases the anomaly score
- Blue bars indicate that the feature decreases the anomaly score
- Small bars indicate that the feature has low impact
- Large bars indicate that the feature has high impact



Explanations are useful to reduce time on identifying false positive and investigating true positive

Note: Single feature removal is good for global anomalies and Shapley value is better for local anomalies

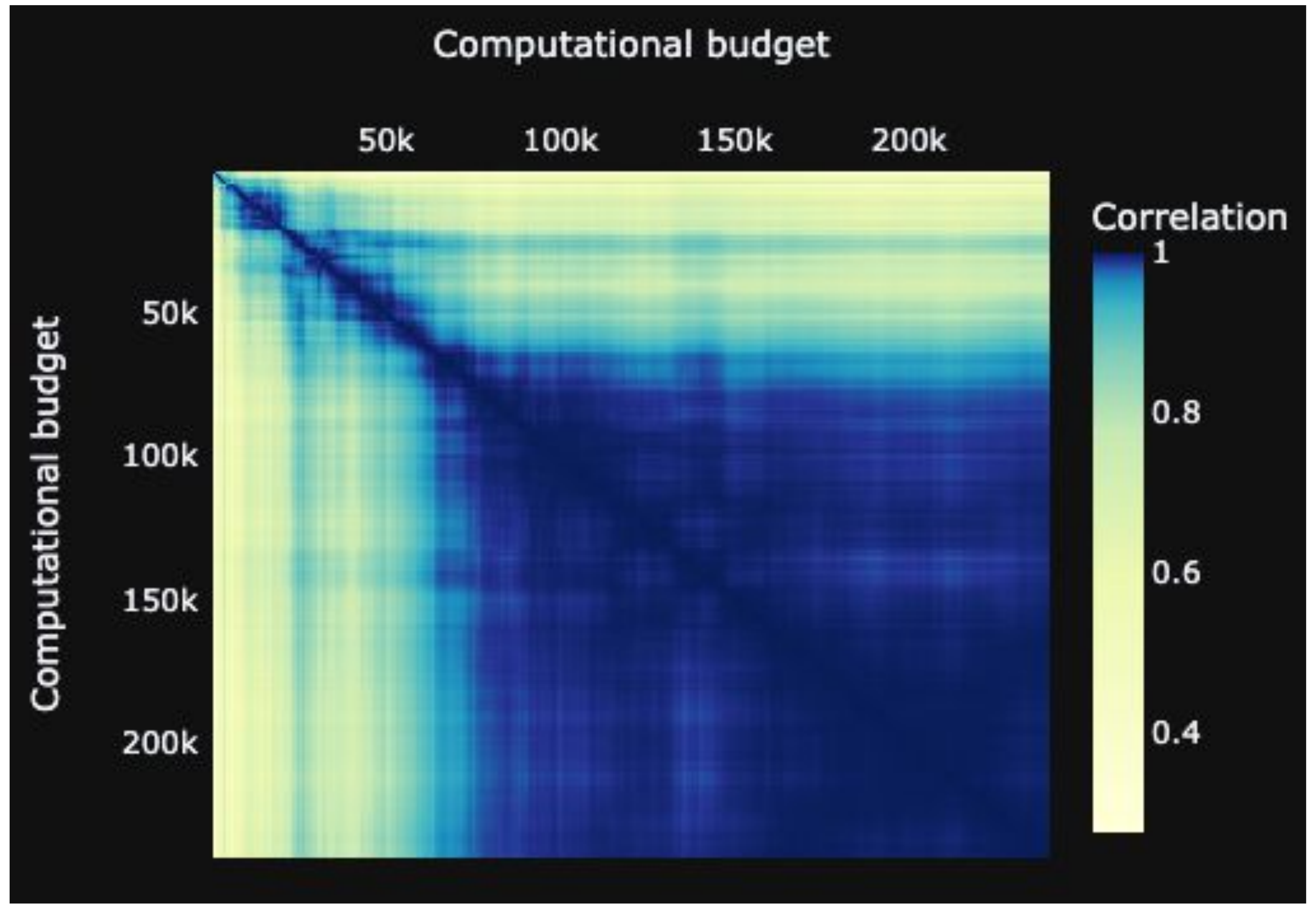
Dr. Antoine Gautier | CRO
antoine.gautier@quantpi.com

Analysing the convergence and stability of explanations is essential for reliable intrepretation

Convergence of scores

Average pearson correlation between feature scores of identical points with different computational budgets.

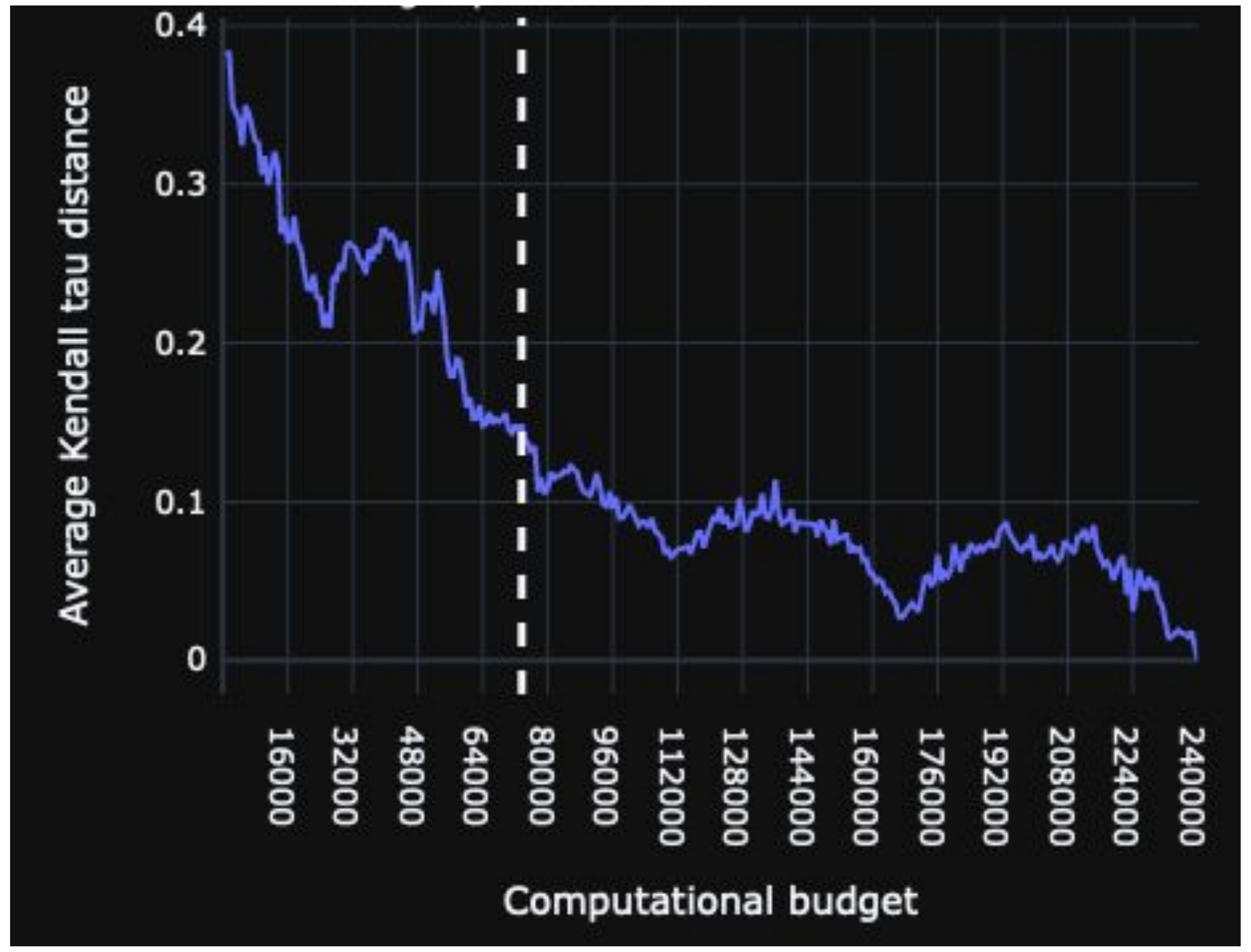
Yellow means the feature scores are very different.
Blue means the feature scores are very correlated.



Convergence of rankings

Average change in Kendall-tau distance between ranking of features induced by explanations differing by one unit of computational budhet.

Low value is good means the ranking has converged.
High value indicates that the ranking is still oscillating a lot with respect to the computational budget.



Dr. Antoine Gautier | CRO
antoine.gautier@quantpi.com

Let’s unlock the full potential of AI

Dr. Antoine Gautier
Co-Founder & Chief Research Officer

Mail : antoine.gautier@quantpi.com

QuantPi GmbH
Halbergstraße 4
66121 Saarbrücken

www.quantpi.com
contact@quantpi.com
+49 681 309 828 99