# Exercise set 5: Explainable AI

## Bonus submission deadline: 21.04.2021 23:59

The exercises marked with **BONUS** should be returned in Moodle by the submission deadline. Successfully solving it will give you bonus points to boost your final grade. During the exercise session, the solution of the exercises will be discussed.

## Exercise 1: LIME (BONUS)

The Locally Interpretable Model-agnostic Explanations (LIME) algorithm offers a framework to give a human-readable interpretation of deep-learning classification methods. The idea was originally presented in the paper *Why Should I Trust You?': Explaining the Predictions of Any Classifier*, whose arxiv preprint is available here. The pdf of the article is also included in the exercise folder. Your task is read the paper and write a short summary (1-2 pages) about the ideas of the paper.You can use the following questions as a guide:

- What is the problem that the authors try to solve?
- What is the main idea behind LIME? How does the algorithm work?
- How did the authors evaluate their proposed method? What were the results?
- What are the main advantages/strengths of the proposed method?
- What are the main disadvantages/limitations of the proposed method?

## Exercise 2: LIME in practice

Nothing to do here, the TA will go through an example of how to apply LIME to image classification. You can open the example by running the following command in your matlab prompt.

```
openExample('nnet/UnderstandNetworkPredictionsUsingLIMEExample')
```

## Exercise 3: GRAD-CAM

Nothing to do here, the TA will go through an example of explainable image classification with GRAD-CAM. You can open the example by running the following command in your matlab prompt.

```
openExample('nnet/GradCAMRevealsTheWhyBehindDeepLearningDecisionsExample')
```