Slide 1:

As you know, for a non-technical person it would be complicated to imagine what is the convolutional networks or how the random forests work. Therefore, for people like that portraying whole AI and Machine learning as black boxes is the easiest way for understanding. Unfortunately, even for technical people is it very complex to comprehend model’s reasoning when we can observe millions and millions parameters. Analyzing them at once seems impossible. That is why we have methods to concentrate on explaining model’s data-driven deciding process.

Slide 2:

There is a chart which shows outline of XAI methods. As you can see there are different ways to interpret our model. It also varies depending on the situation. If we deal with high-risk environment like healthcare it would be suggested to first start from Post-Hoc methods. What is Post-hoc ? It is process of applying methods that analyze model after training. On the other hand, if we have low-risk environment for instance video recommending system then intrinsic methods are advised to use. In this case, Intrinsic methods comprise methods to restrict model’s complexity. It indicates that we intentionally simplify model structure while trying to keep the high-accuracy we had before. As you can see, on graphic we can also notice model-agnostic and model-specific methods. What do they mean simply is that, after we train our model we can either use functions solely meant for our particular model or use functions which are applicable to most models. What is the difference ?

Model-agnostic methods are accessible to use on every model but they do not have access to internal state of model’s structure like weights or bias terms in case of CNN. Therefore, they focus on input-output pairs of models. On the contrary, model-specific methods mean that for our specific model they can access all internal information which allows to present greater explainability in some cases.

Slide 3:

Moreover, we have two different approaches of interpretation methods. Global means that we check overall behavior of the model in all instances. On the other hand, local interpretation methods are meant for explanation of individual predictions

Slide 4:

In this case, we choose intrinsic interpretability method to focus mainly on achieving great explainability and model's decision-making process

. Therefore, applying intrinsic methods which means simplifying model's complexity will result in easier learning algorithm understanding . For instance, if our model would use random forests algorithm a reduced complexity would be a decreased size of decision trees assuming people understand smaller decision trees easier