## Slide 1

Hello everyone, welcome to the 6th topic counterfactual explanation. Before we dive into the technical details, let`s see an intuitive example of counterfactual thinking. So what does counterfactual mean exactly? Actually counterfactual thinking is quite common in our daily life. Have you ever notice that, when a sport game completes, those who won a bronze medal is much happier than those who won a silver one? One explanation could be, the silver winner has lost the first place to the gold medal winner, he/she would think, `I could have won the gold medal if I run a little bit faster`. On the contrary, the bronze winner has gained the third place by winning against the 4th player. So he/she would think `I could lose the medal if I run a little bit slower`. Both of them think of a slightly different world, either a better world or a worse world. Such different worlds are called counterfactual examples.

## Slide 2

How does those counterfactual examples, those closest world help us in interpreting an automated decision? To understand a model, we could either reveal the internal logic of the model (called explain ability), or see the influence of external variables on the output (called interpretability). Counterfactual explanation deals with the latter one. The idea is simple. By generating a slightly different world, input it into the model, and see if the output would be different than the original output.

Counterfactual explanation is a model-agnostic, local, example based method. It does not assume the structure of the model, it only explains a single concrete example, and for this example, it only explains its output rather than the whole model.

## Slide 3

With all these constraints, counterfactual explanation sounds not very informative and its application could be limited. What is the point explaining only the outputs of one single example? There is one scenario that we are not interested in the overall explanation: the lay audience. Lay audience means ordinary people, just like you and me, who do not have expert knowledge about machine learning. These lay audience expect an explanation phased in natural language rather than diagrams. For example, `you could obtain the loan if your credit score is 5 points higher`, or `you could earn more than 50 thousand dollars annually if you had a doctor degree`.

Are these explanations valid? Do they give us the information about the model? Do they help us to understand the model? From the perspective of a lay audience, the reason to understand an automated decision actually one of these three: 1. To understand the decision 2. To obtain guidance for future actions 3. To contest a decision.

## Slide 4

What kind of features should the generated example have?

First of all, the generated example should have a different prediction result. For binary classifications, this means the generated example lies in the opposite side of the decision boundary.

As we have mentioned before, the new possible world should be slightly different to the original world, as for the data, the distance between them should be close, the generated example should be in the vicinity of the original input.

What`s more, for better human readability, the change to the original input should be sparse. The fewer entries are changed the better. The reason for this is twofold: Only by this way human is able to understand which entry has the significant influence on the final decision, besides, counterfactual explanation not only helps to understand but also help to act. Sparsity is also the prerequisite for an actionable explanation.

## Slide 5 (gradient based)

Now we take a few minutes to dive deeper into the techniques. I have to declaim that what I listed here are just two methods to generate a counterfactual, they are the most common methods among all literature.

The first kind method is gradient based. We capture the demands in loss terms and use an energy function for optimization.

Predict an output

Calculate the loss/error of your output

Update your prediction system

## Slide 6 (data evolution)

Question: what are non-differentiable models?

The gradient based method

## Slide 7 (adversarial example)

Have a look at the generated counterfactual examples, do they satisfy our demand?

Sparsity? Validity? Proximity?

Data distribution – Plausibility?

Without out plausibility, we are gaming the system. We are deceiving, cheating the model. Even though we success due to the ill designed model, this result is still not wanted.

Adversarial example (AE) is another highly-related topic beside counterfactual explanations. An adversarial example is an input that is almost identical to the original input, but when we feed it into the model, surprisingly the model will predict totally wrong. It is somehow common knowledge that, automated decision models nowadays are quite vulnerable to adversarial attacks. You can find tons of paper categorized into attacker and defender, and this army compete seem endless.

So how do we distinguish between adversarial examples and counterfactual examples? When we generate a counterfactual example, how do we avoid this pitfall?

## Slide 8 (Solution to adversarial)

One way to avoid generating an adversarial example is to respect the data distribution. See this diagram for example…

## Slide 9 (Actionability)

Let`s recap the aim of counterfactual explanations. Counterfactual explanation aims to help a lay audience to understand the decision, give guidance for future action and help the user protest when necessary.