The Wikispeedia dataset is an example of observational data since the data was gathered without the researcher interfering with the choice of the player of the links to play the game.

Therefore, one can find one or more confounders acting on the causal diagram defined as:
Link at the top position -> Higher CTR.

So far, one can distinguish two confounders:

1. Articles with only links at the top section:

There may exist articles with only links at the top section. This means that the player does not have the choice to go through the whole article and to click on a link based on its position. In fact, all links share the same position relatively. In this case, a link could have a higher CTR than another, not because of its positioning at the top section, but because the player has no other choice: there are no links to look at outside the top section.

In this project, we will investigate how to deal with this confounder. As a proposed method, we would be interested in exploring the fundamental concept of matching. For this, articles with the same links distributions over their sections are paired up.

2. Unrelated topics between two or more finished paths:

Let's consider two paths **A** and **B** starting from a different first article to reach a different article (the goal). In this case, players who are playing the path **A** choose to click on a link that they estimate it could lead them to the goal. On the other side, players who are playing the path **B** choose to click on a different link that they think it will be the right way towards the goal (which is different from the goal of path **A** by definition).

A link **b** existing in articles from path **B** may have a higher CTR than a link **a** coming from path **A**, not because of its positioning at the top section of the corresponding article, but because link **a** did not appear in all articles of path **B**. In other terms, players from path **B** did not even have the choice between the two links since link **a** won't lead them to final goal.

As for the first confounder, we will investigate the elimination of this confounder based on the concept of matching. The causality analysis should be based on articles sharing the same first article and final article (the goal) in the game. In this way, one could ensure that a difference of CTR between two links is not caused by the choice of a player helping him in reaching a certain goal but since the paths are matching with respect to the first article and the final article, the positioning of two links may be the only cause acting directly on the CTR.

Until now, we only investigated observed covariates in the scope of the naive model. However, this latter may be easily attacked by the presence of unobserved covariates. Further analysis could be explored for the next steps of the project.