In this paper, we will approach the problem of generating counterfactuals for processes. The literature has provided a multitude of techniques to generate counterfactuals for AI models, that are derived from static data[1]. However, little research has focussed on counterfactuals for dynamic data[2].

For process data, the literature often uses terms like structured and semi-structured, as they are related to the staticity and dynamicity. Both, structuredness and semi-structuredness, often relate to the data model, in which we structure the information at hand. As static data neither changes over time nor changes its structure, we can use structured data-formats such as tables to capture the information and each data point is an independent entity. We can take the MNIST dataset[3] or Iris dataset[1, 4] as examples for structured and static data. In both datasets, all data points are independent and have the same amount of attributes. In contrast, semi-structured data does not have to follow these stric characteristics. Here, data points often belong to a group of data points which constitutes the full entity. Furthermore, the attributes of each data point may vary. The grouping mechanism could take the form of associative links, class associations or temporal cause-effect relationships. Examples of these are Part-of-Speech datasets like Penn Treebank set[6]. Here, we often associate each data point with a sentence and each of the words are temporally related. The temporal relationship between data points makes turns the data into *dynamic* data. However, not all semi-structured data sets are dynamic and vice versa. Ontologies, like DBPedia[2] can be considered semi-structured but not dynamic. However, structured data will almost always be static. Lastly, there is also unstructured data, which does not incorporate any specific data model. Corpora like the Brown dataset[5], for instance, are collections of text heavy unstructured information. In Figure 1, we show various examples of data.
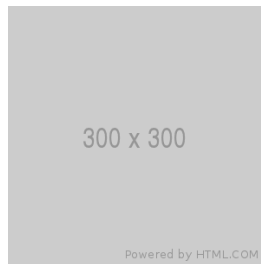

Figure 1: Figure shows two schematic examples of static structured and dynamic semistructured data.

---

[1]With static data, we refer to data that does not change over a time dimension.

[2]With dynamic data, we refer to data that has a temporal relationship as a major component, which is also inherently sequential

A major reason, why there has not been much research on counterfactuals for dynamic semi-structured data, emerges from a multitude of challenges, when dealing with counterfactuals and sequences. Three of these challenges are particularly important.

First, counterfactuals within AI attempt to explain outcomes which never occured. A *what-if* questions often refer to hypothetical scenarios. Therefore, there is no evidential data, from which we can infer predictions. Subsequently, this lack of evidence further complicates the evaluation of generated counterfactuals. In other words, you cannot validate the correctness of a theoretical outcome that has never occured.

Second, sequential data is highly variable in length, but process steps have complicated factors, too. The sequential nature of the data impedes the tractability of many problems due to the combinatorial explosion of possible sequences. Furthermore, the data generated is seldomly one-dimensional or discrete. Henceforth, each dimension's contribution can vary in dependance of its context, the time and magnitude.

Third, process data often requires knowledge of the causal structures that produced the data in the first place. However, these structures are often hidden and it is a NP-hard problem to elicit them[7].

These challenges make the field, in which we can contribute a vast endeavor.