Chapter 1

Background

1.1 Plan for this section

1. Background (Context, Definitions)
2. Previous work (Related works, Particularly timely or relevant work)

1.1.1 Explainable AI

Some of the reasons driving research in explainable AI (XAI hereafter) are summarized in **zhangSurvey2021**:

- Scientists that use AI to achieve their performance goals hope to recover human-understandable information pertaining to their problem of interest;
- Regulators hope to make AI-based decisions accountable, especially those that affect data subjects directly (e.g. in insurance, banking, or even in criminal justice);
- AI users hope to limit unexpected failure scenarios in their work, especially in cases where failure is unacceptable.

The point has been made **rudinWhy2019** that using inherently interpretable models should be encouraged, rather than relying on new techniques to "explain" black-box methods.

Because of how ubiquitous AI is becoming, XAI research is also blossoming and the research field is vast. Some taxonomies have emerged recently that help differentiate methods **zhangSurvey2021**; **bellePrinciples2021**. One key characteristic of an XAI method is its *scope*: a *global* XAI method applies to a *whole* fully trained ML model, while a *local* method gives an explanation of *one prediction* by the model. Global methods report general information that does not necessarily reflect how the model acts in all situations, while local methods produce results with a more limited range, which can therefore be misleading. Yet, assuming that aggregating many local explanations can help us grasp the model's general behaviour, it can be tempting to only focus on those.

1.1.2 Local explainability

Having discarded all global methods, we can still subdivide local methods into several strategies. Some of them are outlined in **bellePrinciples2021**:

- Rule-based approaches (Anchors)

- Local approximation approaches (LIME, LFA)
- Counterfactual approaches (Wachter)

Rule-based

CF is dual to rule-based

Counterfactual explanations (CFX hereafter) consist in producing *virtual points* that are close to the explained input, but for which the model prediction is different. They aim to answer the question: “What would have to change in the input for the model to change its prediction?” This is reminiscent of adversarial perturbations, in which a small perturbation can suffice to "trick" the model into making a prediction nonsensical. In CFX the goal is to keep the virtual points as realistic as possible.

CFX provide value in several ways:

- They indirectly show which features are being used by the model in its decisions;
- They provide insight on the model’s understanding of what makes an input point realistic;
- They constitute a way to provide *recourse* when the model is used in a setting where recourse is required or desirable.