**Problem Statement:** Structured data extraction from lengthy agreement documents aimed at saving time and cost.

**Parallel application:** Law analytics

**Project Overview:**

* We fine tune these using a [token classification down-stream task](https://huggingface.co/transformers/custom_datasets.html#tok-ner) for the purposes of identifying the information we wish to extract and attempt to determine the best models which can be then used for inference and well-structured data extraction.
* We use the ‘[F1 Score](https://scikit-learn.org/stable/modules/model_evaluation.html#precision-recall-f-measure-metrics)’ as the metric we are trying to optimise. The F1 score allows us to balance recall and precision, minimising false positives and false negatives in our predictions.

**Tools Used:**

Python; PowerPoint; Online open-source data

**Data Source:** We will be using [database available online](https://drive.google.com/drive/u/0/folders/1Yu-JnZj1LbVBfTdPiHfMDnaKZj4eqks8) on a open source public interest project started in 2021.

**Roadmap:**

1. Data exploration
2. Corpus creation & Pre-processing
3. Labelling creation for data extraction
4. Validation of model
5. Insight collation