The Problem
The current high-quality air pollution monitoring station network in the UK is spatially sparse with heterogeneous placement and commonly suffers from missing data temporally. In England, only 103 Automatic Urban and Rural Network (AURN) monitoring stations were online for NO\textsubscript{2} between 2014-2018, resulting in large areas of England missing localised measurements.

The Solution
We propose a data-driven supervised machine learning model to fill the missing air pollution concentration data temporally and spatially. We do this by training the model to understand the relationship between the environment around the monitoring station and the air pollution concentrations measured.

Model Framework

How Decision Trees Work
The type of machine learning model used is called a decision tree. Figure 1 shows a notional example decision tree for air pollution.

The first stage, training, is where the model analyses the historical data to learn patterns, in this case, the relationship between environmental conditions and air pollution levels. For example, the scenario of low air pollution concentrations at a time of high wind, shown as the green route in Figure 1. The data from existing monitoring stations is used as historical training data.

The model allows us to predict air pollution levels from the known environmental conditions based on the relationship learnt during training. This is known as the prediction stage.

Understanding Our Past

Filling Missing Data Temporally
Since we have data describing the environmental conditions for each AURN station, we can predict the air pollution concentrations at all timestamps. Missing data, such as those in the monitoring station in the figure, can now be filled in, providing a complete dataset of hourly concentrations.

Filling Missing Data Spatially
Further to filling in data temporally, we can fill data spatially. Environmental data is available throughout England; therefore, we can predict concentrations at all locations. The model predictions equate to 355,827 stations at 1km intervals, creating a dataset valued at £70 Billion.

Exploring Hypothetical Situations
The model can also predict future air pollution concentrations for hypothetical situations. For example, the model can answer the question: How would NO\textsubscript{2} air pollution concentrations change if a motorway was built in the East of England? Any environmental condition used to train the model can be changed to explore hypothetical situations.