# Downforce Technologies Data Science Interview Test – Soil Organic Carbon Prediction

## Background

Soil Organic Carbon (SOC) is a key indicator of soil health and plays an important role in agriculture, climate mitigation, and ecosystem functioning. Traditionally, SOC is measured through in-field soil sampling followed by laboratory analysis, which is accurate but costly and spatially sparse. In this task, the goal is to explore how satellite-derived representations can be used to estimate SOC remotely, calibrated using ground-based soil samples.

The ground-truth samples used here are derived from the **LUCAS Soil Survey** (Land Use/Cover Area frame Survey), a harmonised topsoil dataset collected across the European Union using standardised field and laboratory protocols.

## Data Provided

You are given the following inputs:

1. **Training data**
    - File: `training_data.csv`
    - Contains UK soil samples from the LUCAS dataset.
    - Columns include:
        - Coordinates (EPSG:27700)
        - `TargetSOC` (target variable to predict)
        - Sampling date
        - 64 AlphaEarth foundation model embeddings for **2022**
        - 64 AlphaEarth foundation model embeddings for **2024**
2. **Raster data**
    - Two large GeoTIFF files containing Google AlphaEarth foundation model embeddings (stored as integers) for a UK tile:
        - 2022: `s3://us-west-2.opendata.source.coop/tge-labs/aef/v1/annual/2022/30N/xkou4w6uhyogespuy-0000008192-0000000000.tiff`
        - 2024: `s3://us-west-2.opendata.source.coop/tge-labs/aef/v1/annual/2024/30N/xks7764uu0jo1h8jh-0000008192-0000000000.tiff`
3. **Utility script**
    - A Python script that dequantises the integer-valued embeddings in the GeoTIFFs into floating-point values, consistent with the embeddings in the CSV file.

More information about the AlphaEarth foundation model can be found here:
https://source.coop/tge-labs/aef

## Tasks

1. Train **any machine learning model** to predict `TargetSOC` using `training_data.csv`.
2. Provide all your work in a **single notebook**.
3. Clearly explain:
    - Feature selection and preprocessing
    - Model choice
    - Training and validation strategy
4. Use plots or visualisations where helpful to describe:
    - The dataset
    - Model behaviour or results
5. This exercise is **not about maximising accuracy**. We are primarily interested in how you approach and reason about the problem.

### Optional Bonus

- Apply your trained model to the provided GeoTIFF embeddings to generate a **dense SOC prediction map** for the tile.

## Deliverables

- A runnable notebook containing:
    - Code
    - Explanations and reasoning
    - Visualisations (where relevant)
- A `requirements.txt` file listing **all Python libraries** required to run the notebook.