Data-Centric Engineering Training

Code

Dates TBC

Introduction to engineering data analysis and reproducible workflows

Course materials

Slides will be presented along with an accompanying computational document that includes chunks of `Python`, `R`, and `Julia` code necessary to analyse the data, and solve the problems in the various examples. To achieve this, various libraries/packages have been used, and these will need to be installed and loaded for the code to run.

Load Python Packages | Load R Packages | Load Julia Packages

▼ Code

```
import cmdstanpy, multiprocessing, math
import numpy as np, pandas as pd
from scipy import stats
```

Tip: Loading packages

In `R`, `Python` and `Julia` packages first need to be installed. Guidance on installing packages can be found here (for `R`), here (for `Python`), and here for `Julia`.

Packages only need to be installed once (unless they are uninstalled), but then need to be loaded each time you want to make direct use of the functions or data they contain. This document will not detail the workings of each package, but such information can be found online, for example here is the website for the `Python` 'pandas' package.

Course objectives

- Attendees will learn how to perform calculations in a reproduceable way using modern programming languages and libraries (including probabilistic programming). The course will introduce multiple challenges associated with analysing engineering data.
- We will introduce the concept of (multi-variate) uncertainty: where it comes from, how to quantify it, and how to propagate it through engineering models in a transparent and reproducible way, to justify our decisions and recommendations.
- Atendees will learn new tools and techniques that will help them in their day-to-day work, developing skills that will help them introduce methods of "data-centric engineering" into their organisations.

A note on spreadsheets...

Engineering calculations are often performed using spreadsheet software. This course, however, will demonstrate how to perform engineering analysis using programming languages (and accompanying libraries). The primary reasons for this are as follows:

- When running challenging simulations, or working with large amounts of data, spreadsheets may become unmanageably slow.
- Spreadsheet software has reproducibility challenges. Data can be difficult to distinguish from calculated quantities, and the sequence of calculations can also be difficult to identify. Reproducable workflows are important for sharing work with colleagues, re-running calculations, and documenting that calculations have been performed correctly. The Alan Turing Institute has developed The Turing Way, which provides guidance on best practice for reproducible workflows.
- The software used in this course is all freely available and benefit from large user communities. Modern analysis tools are primarily developed (and maintained) for the programming languages rather than spreadsheets.
- Calculations using spreadsheets are often unreliable. Errors *arise* in spreadsheets for many reasons, such as obscuring (or deleting) data, incorrect assumptions regarding data types, uninterpretable functions, and automatic filling. Errors often *remain* in spreadsheets due to the challenges associated with testing, documenting and version control.

Some notable examples of high consequence spreadsheet errors are recorded by the European Spreadsheet Risks Interest Group (EuSpRiG) and further discussion on the incompatibility of spreadsheets and good data management (in the context of research) can be found in a presentation from Monash University, here.

Instructors

Andrew Duncan, PhD

Andrew is a senior lecturer in statistics and data-centric engineering at Imperial college London and previously led the data-centric engineering programme at the Alan Turing Institute. He is an expert in uncertainty quantification, big data environments, DevOps, MLOps and web services. He has previously supported many large industrial projects across engineering sectors, including energy, aerospace, and maritime.

Professor Adam Sobey, PhD

Adam Sobey is an Associate Professor in the Maritime Engineering group at the University of Southampton, Group Lead for Marine and Maritime in the Data-Centric Engineering Programme of The Alan Turing Institute. His work in maritime engineering was incorpoertated into Lloyd's Register's design guidance.

Domenic Di Francesco, PhD, CEng(MIMechE)

Domenic worked in the energy industry becoming a chartered mechanical engineer before completing his PhD in computational statistics. He is now a research fellow at the Alan Turing Institute and a visiting researcher at Cambridge Univeristy, with expertise in decision making under uncertainty and data-centric engineering.

Peter Yatsyshin, PhD

Peter is a cross-disciplinary researcher and lecturer with publications in machine learning, statistical physics and applied mathematics. He is an Alan Turing Research Fellow and an honorary research fellow at Imperial College London. Peter is an expert in physics-informed models, uncertainty quantification, and differential equaions.

Example: analysis of material test data

Material testing has been completed for a new constructrion project. There are a limited number of measurements of material toughness. These values meet the contractual requirements, and so have been accepted.

Challenges

How can the data be analysed to answer the following questions:

- What value of toughness should be used in calculations?
  - What are the industry standard approaches to find this value?
  - What are some alternative approaches?
- How can we analyse the data to obtain a probabilistic estimate of the toughness?
  - How can we fit a probabilistic model in Excel?
  - How can we fit a probabilistic model in `R`, `Python` and `Julia`, and why should we consider doing this?
  - How can we also account for statistical (epistemic) uncertainty by using probabilistic programming?
  - How can we give the model a helpful starting point (prior) to further improve predictions?
  - The vendor has provided some measurement uncertainty - how can we incorporate this into our calculation?
- A new use case has been proposed, requiring a new minimum toughness. The engineering team have been asked whether it is safe to use the material. How can we answer this question?
  - How can we decide whether more testing is required?

This example will discuss the merits and challenges of different tools and methods, including:

- maximum likelihood estimation
- Bayesian inference
- value of information analysis.