Alan Turing Institute Data Study Group – Syngenta Problem Statement

Syngenta is a business that employs world class science and innovative crop solutions to transform how crops are grown, enabling millions of growers to make better use of available resources.

In collaboration with academic and industrial partners and with UK government funding, Syngenta is working on a project to support farming decisions by simplifying farmers' access to knowledge.

Benchmarking is the relative comparison of efficiency and productivity between growers and is a longstanding practice in the agricultural industry. This project intends to research a data-driven approach to benchmarking, that can extend the benefit beyond national boundaries, for both developed and developing countries. This involves understanding whether the necessary data can be captured, transferred, integrated and interrogated robustly and cost-effectively.

The project anticipates that around 50 variables will be required to form a representative 'scenario'. This in itself is a data challenge as it requires sourcing and processing data from a large set of data sources. Once the data set has been assembled, there are further mathematical and data science challenges.

To begin with is the substantial double-sided question of the resolution at which the variables should be measured and aggregated over, which is related to and dependent on the size or area that a 'scenario' has. We have inter-dependent choices to make and would like help exploring these.

Once the scenario parameters have been defined, we wish to understand which methods would best allow the comparison of individual scenarios with each other, given that they will be comprised of high dimensional, multivariate data. An initial decision would be whether to opt for supervised or unsupervised machine learning. Is it OK to largely discretise the variables or adopt a kernel method, to allow scenarios to be categorised but at the cost of potentially losing information? Alternatively, unsupervised methods involving clustering and proximity analysis may be preferable but have the potential for untenably high computation cost.

We would like to investigate this across the breadth of data interrogation approaches and welcome fresh perspectives and untypical statistical approaches. For example, one niche but very deep area is that of soil data, potentially including the rhizosphere and soil metagenomics. If we had consistent access to such data, how could we sensibly include it in our data set and match scenarios to that?