

# Data Science Toolbox Portfolio Questions

## 03 Latent Structures, PCA, and Clustering

Daniel Lawson — University of Bristol

### Block 3

## Portfolio 03

Choose **one question** and write up to **one page** about it. You are free to conduct further experiments to add weight to your results, and any additional material you generate can be submitted as an appendix. See [The Assessment Page](#) for advice.

These questions may make reference to the content from the current block.

**Question R03.1:** Read the (python) documentation for [Sparse SVD](#). How is this making computational efficiencies over a standard SVD and when might this fail? The documentation discusses the sense in which the results are accurate in terms of “subspace\_angles”. Discuss what this means, and how it relates to the use of the SVD in PCA.

**Question R03.2:** Read the [documentation about how HDBSCAN works](#). Reflect on the importance of dimension in this for the construction of the nearest-neighbour step. You might want to refer to results in the literature such as “[When is Nearest Neighbor meaningful?](#)”.

**Question R03.3:** Imagine that you are trying to understand the Cyber Security data from [Workshop 3.3](#) for the purposes of predicting whether traffic is “normal”. Consider the advantages and disadvantages of enriching the feature set via a) dimensionality reduction, and b) clustering, for the purpose of passing to a classifier. You may wish to perform experiments (and cite results placed in your appendix) for this task.