# Student enterprise project outline

| | |
|---|---|
| **PROJECT TITLE:** | **Cluster Interpretation via Dimensionality Reduction** |
| **REFERENCE:** | FocusWare-03-cluster-interpretation |

**BACKGROUND:**

FocusWare (https://chriswalshaw.co.uk/focusware/) is University-adjacent organisation created by Prof. Chris Walshaw to help market the NetWorks mobile optimisation software and to exploit collaborations with Nokia Siemens and the University of Malaga:

- https://patentscope.wipo.int/search/en/detail.jsf?docId=WO2011144922
- https://link.springer.com/article/10.1007/s11277-010-9963-1
- https://link.springer.com/article/10.1007/s10732-010-9148-9

Although the primary software product is focused on graph & network optimisation, a number of related side topics are currently of interest.

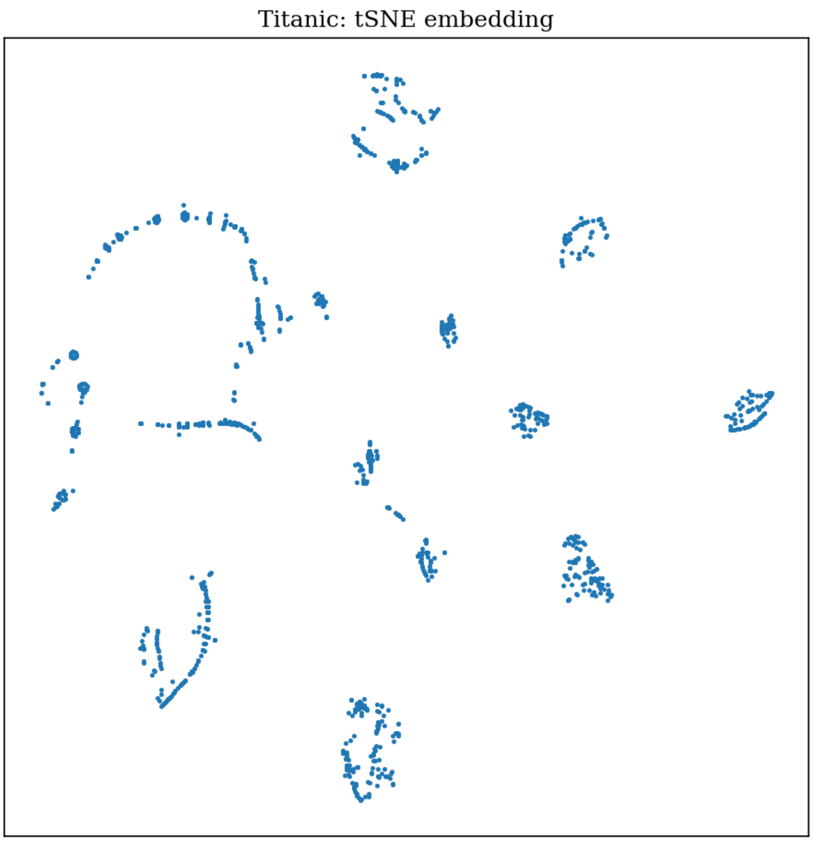
**PROJECT OVERVIEW:**

As data collection grows, the data analysis becomes increasingly more complex. Although we have simple techniques (such as segmentation, sampling, filtering) to help deal with data that has thousand or even millions of records (rows/datapoints/etc), it is much harder to deal with data that has many dimensions (columns/features/variables/etc). Even data with just 10-20 dimensions can be difficult to analyse and some datasets can have thousands of dimensions.

In recent years, **dimensionality reduction** has become a popular tool for visualising and understanding the vast amount of high dimensional data that is now available. The most well-known are PCA and tSNE but more recently other popular techniques, e.g. UMAP, have emerged. These can be coupled with **clustering algorithms** such as k-Means or DBSCAN to segment the data into smaller, closely related clusters. However, it is not always easy to interpret what each cluster represents.

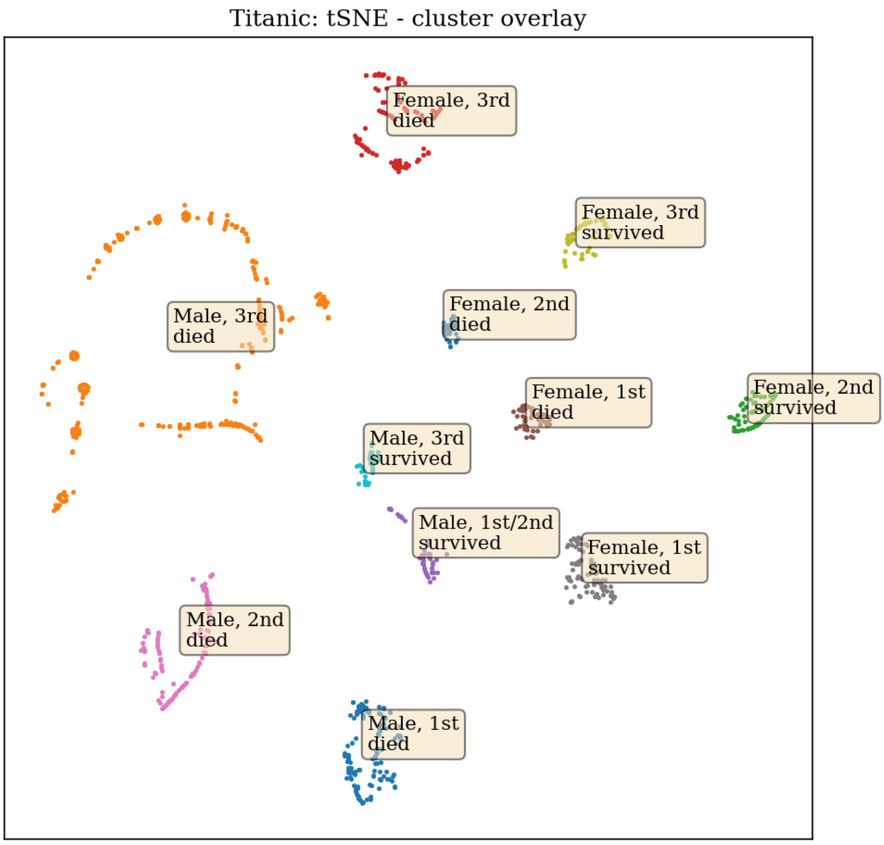
The aim of this project is to devise and implement Python-based methods to assist in (semi-)automatically interpreting cluster composition, both for standard datasets and on data provided by FocusWare. An **example** is shown on the following page.

| | |
|---|---|
| **TECHNICAL SKILLS REQUIRED** | Python / Machine Learning |
| **OTHER SKILLS REQUIRED** | Project management, time management, interest in data science & data visualisation |
| **OTHER IMPORTANT INFORMATION** | Please email Chris Walshaw (c.walshaw@gre.ac.uk) to register interest in and/or apply for this project. |



Titanic: tSNE embedding

The well-known Titanic passenger dataset, embedded into 2 dimensions by tSNE.



Titanic: tSNE - cluster overlay

The same dataset, clustered by k-Means and with the clusters interpreted manually.