

|  |  |
| --- | --- |
| **Company / Organisation Name:** | Ofcom |
| **Team / Department**: | Research & Intelligence, Strategy & Research Group |
| **Address:** | Riverside House, 2a Southwark Bridge Road, London SE1 9HA |

**Provisional title for project:**

Clustering online services used by UK users to inform Ofcom’s online safety work

**Short description of the problem that would be addressed by the project:**

Ofcom has regulatory duties under the Online Safety Act. While the largest online services are easily identifiable and well targeted, there are nearly 700,000 services with fewer than 100,000 UK visits per month (sometimes known as ‘longtail services’). Many of these have responsibilities under the Act, posing an important challenge: how can Ofcom focus attention efficiently on the most important areas?

Current methods allow for filtering based on size (visits metric) and type (industry metric, e.g. social media, search engines) but is more possible? Grouping those 700,000 services into manageable categories relevant to online safety areas, in a way that supports programmatic monitoring, could be a major benefit in guiding Ofcom’s ability to allocate resources effectively and make informed decisions.

**Short description of the data sources that would be used in the project, and how they would be used**

The primary dataset for this project comes from SimilarWeb and includes detailed information on 700,000 websites and their characteristics. These characteristics cover audience metrics, site metadata, technologies employed, marketing strategies, and more. The dataset is managed using Azure/Databricks, which provides the capability to integrate additional data sources if needed. To prepare the SimilarWeb data for clustering algorithms, some exploration and pre-processing will likely be needed. This includes establishing an analytical approach to data transformation, such as handling categorical variables and implementing dimensionality reduction techniques. An assessment of clustering techniques will also be needed to determine the most appropriate method, as different data distributions result in different shaped clusters, and different algorithms are better suited to specific scenarios.

An initial step could involve applying clustering to a subset of the data, to allow for the use of semi-supervised learning techniques, where clusters are first identified and human-labelled, followed by the incorporation of additional data to refine results. This approach could improve the accuracy of the clustering process. Even if the results are negative, understanding the reasons behind different accuracy levels between clustering and classification methods will provide valuable insights.

The primary research question is whether the available data can enable the grouping of online services in a way that is meaningful to Ofcom’s duties under the Online Safety Act. Any relevant findings will inform Ofcom’s ongoing monitoring of the Online Safety sector. If specific categories of services are identified as risky or particularly relevant, these categories could be integrated into Ofcom’s regular reporting and used to guide its policy development and supervisory activities.

**Would any work by the student need to be carried out on site at the Company (with the exception of supervisory**

**Meetings)?**

No, online interaction is the normal way of working but we have an office in London where the student can work if helpful.

**Any issues of data confidentiality and IPR that would need to be resolved**

Reports or publications will need sign off from Ofcom before being used externally.

**Essential skills**

Data analytics, including knowledge of unsupervised learning techniques. Python, SQL or similar.

**Desirable skills**

Segmentation clustering/market analysis skills.

**Preferred degree programmes (if any)**

Not specified but preference would be Data Science or Geographic Data Science

**Preferred selection method**

CV sift and Online Interview

**Support and training offered by the company**

Weekly online meetings. Initial training on dataset familiarisation and available analytical tools and resources will be provided.  The student will have support from a Senior Associate who is familiar with the datasets used.

**Financial assistance offered by the company**

£500 stipend upon completion

**Any other comments**

If there are any questions about the 2025 programme, please contact Richard Arnold at [richard.arnold@ucl.ac.uk](mailto:richard.arnold@ucl.ac.uk). The completed form should also be returned to this address.