Team AstraZeneca

Company mentors: Isabel Smith Nino (main point of contact), Antonia Langfelder, Anna Gogleva

**Introduction:**

AstraZeneca is a large research-based BioPharmaceutical company which focuses on many disease areas such as oncology and respiratory. The company has around 60,000 employees and three main hubs in Cambridge, UK; Gaithersburg, Maryland, USA; and Gothenburg, Sweden.

Gene Expression Omnibus (GEO) is a public genomics data repository containing experimental data submitted by scientists. Through the website, users can query and download gene expression profiles for humans and other species. More information can be found at: <https://www.ncbi.nlm.nih.gov/geo/info/overview.html>. The focus of this project will be experiments performed on humans or human cell lines to study diseases.

Experiments generally consist of a “healthy” sample being compared to a “diseased” sample, and then observing which genes change their expression. This information can inform drug development, helping to prioritize compounds that can target genes that drive disease progression.

**Objective:**

The goal of this project is to focus on the studies conducted on humans and use the text in the description to understand what disease areas are being studied, how this has changed over time, and what can be said about the future. You can also think about interesting ways to visualise these results. The questions are open-ended to give you a chance to explore. If other interesting questions occur to you, you are free to use the data to answer them.

Extra credit: Within each study description, there are a list of samples. However, it is not always clear (to a computer) whether the sample name references a healthy or diseased sample. The question is then: can you use the summary text to infer which samples are healthy and which are diseased?

Main parts:

1. Extract the text data from GEO
   1. Series homepage: <https://www.ncbi.nlm.nih.gov/geo/browse/?view=series>
   2. Example experiment: <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE117746>
   3. Information on programmatic access: <https://www.ncbi.nlm.nih.gov/geo/info/geo_paccess.html>
2. Perform NLP analysis on the extracted text data