**Data Science for Industry Short Course**

***Task for Certificate of Completion***

The US Consumer Financial Protection Bureau provides a public dataset of complaints they have received about financial products and services (<https://catalog.data.gov/dataset/consumer-complaint-database>). This project uses a subset of 20 000 of these complaints, saved as **complaints.RData**. Each record/row captures a single complaint, on the following variables:

* **date\_received:** the date the complaint was received
* **product**: a broad categorisation of the type of financial product involved (5 categories: bank account or service, credit card, credit reporting, debt collection, mortgage)
* **consumer\_complaint\_narrative**: the submitted text explaining the nature of the complaint.
* **consumer\_compensated**: a binary indicator of whether the consumer was compensated (monetarily or non-monetarily) after the complaint.

For the task you can choose to do ANY ONE of the following (or both, if you like, but only one is required for the certificate of completion)

1. Use the bing sentiment dictionary to calculate sentiment scores for the complaints, and plot a histogram of sentiment scores for each product class. What patterns do you notice?
2. Construct a neural network classifier that, given only the text of the complaints, predicts whether the claim will be compensated or not. Use a feedforward neural network based on bag-of-words features, with or without tf-idf weighting. Assess the out-of-sample performance of your classifier.

Your submission should consists of the following items:

1. Your **report**, writtenin whatever word processing software you like (e.g. word, latex). The report should contain a description of the problem, the approach you took, and your results. There is a page limit of 8 pages for the report.
2. Your **code**, either embedded in the document if an Rmd file, or as a separate .R or .Rmd file, or jupyter notebook. The key thing is that I need to be able to run your code and reproduce your results, so there should be clear instructions on how to use your script(s). The code itself should not be displayed in the final typeset document (use “echo = FALSE” for Rmd files) and not pasted as an appendix in your report.

Please submit a single .zip file containing your report and all code, to **ian.durbach@uct.ac.za**

The submission deadline is on or before Friday 6th September 2019. This is just so that we can get the certificates done without undue delays – if you need more time, let me know (within reason).