Data Science Toolbox Assessed Coursework 3: Data at Scale

Deadline: Wednesday noon, Week 24

Group Project description

You will **choose an application domain** that your group will work with for Assessment 3. Your challenge is to apply **Massively parallel Data Science technology** to that data.

You should:

a) choose an appropriate **scientific/analysis question**;
b) use an appropriate strategy to **learn about the computational performance** of the model(s);
c) apply the model(s) to achieve your question.

Appropriate Methods

- Classifiers
- Neural Networks
  - Auto-encoders
  - Choice of Deep Learning platform
  - Choice of architecture
  - Interpreting Deep Learning decisions
- Exploratory Data Analysis
  - Graph visualisation and algorithms on graphs

Appropriate Technologies

Models and scientific questions include but are not limited to:

- Algorithmic approaches
- Parallelism via GPUs
- Map/Reduce
- (Py)Spark
- Pregel/GraphX for graph-based computation
- Distributed data vs Distributed computation paradigms
- Comparison between parallel and single-machine problems
- Scaling performance in terms of data volume and resource allocated

Advice on assessment

You will be assessed on:

a) the implementation of the model, that is, you can be awarded credit for:
   - additional implementation if an off-the-shelf implementation falls short,

- exploring multiple implementations,
- examining the mathematical details of choices.

b) the application of the model to **your chosen domain**, that is you can be awarded credit for:
   - identifying an appropriate dataset;
   - using your understanding of the structure of datasets to make arguments comparing the dataset you chose to one that you might encounter in a "real" cyber-security setting;
   - plotting or otherwise describing various inputs, outputs, or parameters.

c) the correctness of the methods used to achieve their stated goals.

d) the robustness of the results in supporting the conclusions.

In order to be attributed credit for your efforts to choose appropriate data, ensure that you document the data exploration process. You should aim to demonstrate diligence that there is no more appropriate data source in your chosen category.

You do not need to excel in all areas in order to get a high mark. Instead, you need to perform robustly in all areas and additionally demonstrate insight somewhere to score highly.

You do not need to work in a genuine high-volume environment, but you should demonstrate how you expect your method would perform at scale.

Individual reflection description

- Discuss the rationale behind the inference goal that you selected;
- Discuss the parallelisation paradigm that you explored;
- Discuss what changes you might have to make were the volume of data to be increased by a factor of 1000;
- Relate your data source to those you might encounter in a real-world setting;
- Discuss a mathematical issue raised in the project, different to those of your group.