# Data Science Toolbox Assessed Coursework 2: Data at Scale

**Deadline: Wednesday Noon, Week 11**

## Group Project description

You will **choose an application domain** that your group will work with for Assessment 2. Your challenge is to apply **Neural Networks in a way that would scale** to a deployable data-science solution.

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

When applying Neural Networks consider:

- Will you use it for:
  - Data processing, for example to analyse free text?
  - Classification and/or a prediction task?
  - To analyse complex data types, such as images or graphs?
- What type of Neural Network will you use?
- How can you make it work in the test environment of your laptop/desktop/colab/noteable?
- How can you justify that your approach will scale?
  - Scaling here means that you should consider how your approach would work if the data volume were to increase by a factor of 1000.
  - It would be expected to require access to high-performance computing resources (e.g. GPUs, clusters) to achieve this in practice, but you do not need to actually run at this scale.
  - You should be able to argue that your approach would benefit efficiently from such resources.

### Appropriate Technologies

You may use this opportunity to delve further into:

- Pre-trained models
  - Transfer learning
  - Fine-tuning
  - Few shot learning
- 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" data-science 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 are not expected 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 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.

## Example project ideas

You should know about The Datasets subreddit, Kaggle datasets, and Hugging Face datasets. Here are some example project ideas to get you started:

- How to find datasets? Examine the Reddit Dataset Dataset and train a neural network to classify datasets into categories (e.g., text, image, tabular; large, small). Use human truthing to improve your neural networks performance. Test by using the network to suggest suitable datasets for DST projects.
- You are working for a social media company and want to classify images uploaded by users. Use the Flickr dataset or (a very small part of) the Kaggle Instagram Images dataset with a pre-trained neural network such as CLIP to make tags for images. How do you know whether your approach is working well?
- You are working for an investment bank and want to know how to incorporate news content into stock predictions. Contrast tools such as LLM predictions, and what you can train for a dataset such as Stock News Analysis using e.g. FinBERT linked to historical stock prices, to predict stock price movements based on news articles.
- Consider the Reddit Mental Health Dataset. Use a pre-trained neural network model to identify posts that discuss pharmaceutical products. Is there a sentiment association with some products? How does the discussion of products correlate with public information on prescription rates? Can this be linked to advertising and/or pricing? You may find the Medical Expenditure Panel Survey useful for prescription data in the US.

## Coursework guidance

**This section is the same for every coursework.**

### Submission

**Every group member must submit** something to Blackboard, by the deadline. You must submit an **Individual Reflection** and a **Group Report**.

It is recommended to upload the Reflection directly to blackboard, and in the notes add a link to your Report repository.

You should submit a Report Repository containing:

1. **README.md**: An explanation of:
   - **Project Group**: List who was in the group.
     - It can be helpful to describe briefly what their contribution was, here or in the reading order. Any **Equity** variation should be clearly noted.
   - **Reading order**: the order that your files should be read in, which should explain:

    - Preparation: how to install any packages or software etc that should be installed.
    - Your report content reading order (if there are multiple files), with any additional info about the file you feel appropriate (e.g. if only some of the team were an author; the purpose of the file such as "data downloading", etc)
2. **report/**: a FOLDER, containing all of the files that will be read as part of assessing your project.
    - Label these alphanumerically as "<number>-<name>.<file ending>" for the reading order.
3. **Documentation folders**:
    - Each member of the group should be using the repository to work on the project. Make **one folder** per group member, name and merge your content into the report as possible. This will act as documentation that you have contributed to the project.

There is an Example for you to emulate, with the structure:

- README.md
- report/
    - 01-Data.Rmd
    - 02-R_analysis.Rmd
    - 03-Python_Analysis.ipynb
    - 04-Wrapup.Rmd
- RachelR/test.Rmd
- PeterP/work.ipynb

**Assessment**

- 75% of your mark will be for the group project itself. All students in a project should submit the same project; only one project will be run. The individual marks may be moderated away from the group project mark.
- 25% of your mark will be for an individual reflection, which should be written by you. It should be approx 500-800 words (not strict) which should be individually written.

## Report

All coursework for this unit is based on group work in teams of around 3. Your team will address a single data science challenge. You will have choice about the topic, within the remit of the project description. It is always the intention that you each learn from, and teach, your teammates any skills you can bring to bear on your chosen problem. Your team will submit a single project report, which is a script that can be run to a) obtain data, b) analyse data, and c) produce any figures and tables that you feel are illuminating.

Your project script would **typically** take the form of an **Rstudio markdown** project or a **Jupyter Notebook**. It should be annotated with factual statements

describing what you have done and why in basic terms. Unless otherwise stated, you may choose the programming language but we recommend sticking with python or R since all students are expected to become familiar with these. The results of computations including plots should be displayed and labelled (e.g. with numbers) and if you have not used a seamless method then you must provide a zip file containing both a script, and a pdf or similar document that also contains the output of your script. Your script is expected to run, and if at any stage some manual step is required (for example, to wait for a bluecrystal job submission to finish, or data must be downloaded) this should be carefully noted. You may lose marks if your script needs debugging.

There is no word, page or other limit. Credit will be awarded for making your arguments thoroughly but without repetition or meandering off-topic. Only include material that you feel makes a contribution to the overall project scope. If some research led to a dead end, work it into the results.

Remember to **reference** websites and other resources for content and ideas, in addition to the usual academic referencing. This will assist you in your future projects.

### Report Assessment Criteria

Your project will be assessed against the following criteria:

- Fit and Success.
  - *Choice of project is very important for learning. This section includes finding good datasets and matching them to questions; making progress on hard problems; fulfilling the learning outcomes.*
- Innovation.
  - *Thinking outside of the box and finding resources that are not presented in the course. Innovation can come in the form of data, methods, and mathematical ideas brought in from elsewhere.*
- Citations, Referencing, Literature.
  - *Cite your sources, build up a repertoire of useful content. Link your results to those on analogous problems. Note that many resources are not published papers.*
- Structure and Description.
  - *Your project should be well introduced, and easy to read and undestand. Make good figures and explain them. Structure your project well, stick to the point and note what your results mean.*

### Equity

It is assumed that all group members will contribute equally to the project. However, this is not always possible for a variety of reasons. Your team should discuss this and agree an **equity** or proportional contribution to the group project, accounting for both practical (implementation) and conceptual (theory, methods choice, etc) contributions. If you cannot agree, you should approach

the tutor to try to agree equity before submitting divergent opinions. Try to agree any non-even equity before the project gets underway.

Contributions will be taken into account when assigning individual marks from group reports. Small deviations are unlikely to be given divergent grades.

Individual grades can be moderated up and down based on equity but are unlikely to be increased as much as they are decreased, and the final decision takes into account documentation.

Additional notes:

- It is expected that all group members understand the group submission.
- It is also the intention that they put in equal effort.
- It is not expected that the final script contains content proportional to equity. There are many good reasons that work does not make the final report.
- If you put in lower effort and agree a lower equity, you may receive a proportionally lower group mark.
- If you put in extra effort and agree a higher equity, you may receive a higher mark but the reward is not linear. It is better to have an equal share of a good project, than a high share of a poor project.
- Mathematical contributions and programming contributions can be considered. All contributions should be documented.
- If some people choose lower equity because they could not contribute fully, make this clear in all reflections. The lower manpower may mitigate a low grade.

**Documentation**

All students are expected to contribute to programming. You should each submit your own scripts, session history or similar, that demonstrate that you made some independent effort, even if these did not make it to the final report. If you cannot demonstrate an amount of effort commensurate with your claimed equity, then your mark may be reduced.

Your documentation is likely to take the form of an Rstudio markdown or Jupyter Notebook. It can be long and contain dead ends. It does not need to be documented, nor be able to run from top-to-bottom. It should be unique to you, though is likely to contain content from others' notebooks. You may refer to it in your individual reflection, but if there is excessive material that should have been shared with the group then you will not receive credit for it. You should not try to boost your individual grade by doing extra work here. It may not be carefully read and you may not receive feedback on it. It should be no additional effort to produce this as it should consist of files that you already have.

## Individual Reflection

The purpose of your reflection is to encourage changes in your practice that improve your understanding of Data Science, as well as improve your ability to work in a team on Data Science projects.

You are being assessed on your progression and understanding of the content of the project. It is better to note deficiencies with what you have done, than to try to post-hoc justify something. It is understood that you are under time pressure and may make a poor irreversible decision for the project performance, but that will not strongly affect your mark if the reason for the failure is clear. You must write your writeup independently of the other students, though using the shared understanding gained from working with them.

### Reflection Assessment Criteria

Your reflection should always adress the following areas in addition to what is asked in the specific project description.

- Fit.
  - *Introduce the area and explain the overall goals.*
  - *Justify the decisions made in the project.*
- Depth.
  - *Explain the results and discuss the conclusions. This can focus on your contributions but should also include the project as a whole.*
  - *Briefly explain some aspect of the mathematical model(s) that has been used.*
    - *It is expected that your group will discuss this in detail, and that contribution of understanding is included in the project contributions.*
    - *Each student still must write something in their own words.*
  - *Reflect on the strengths and weaknesses of your approach, and how you might do it differently next time.*
  - *Reflect on any aspects of the project that could be improved, paying particular attention to* **group working**, *technological barriers and solutions.*
- Structure and Description.
  - *Write clearly, use referencing if appropriate. This is a reflection, not a report, so use appropriate language.*
- Evidence.
  - *Your documentation and reflection together are assessed to evidence your individual contribution. Explain this and how it fits into the whole.*

### Learning outcomes

You are reminded that:

- Teamwork is a learning outcome.
- Progress, teaching and sharing is more important than individual technical ability or project success.
- The difficulty of these assessments is beyond what would be expected of an average student alone.
- Most groups will contain a mixture of expertise which should be exploited.
- In the event that your entire group is inexperienced at programming, you still need to meet a minimum standard. However, you can still score well if you focus on a mathematically interesting question.

## Working practice

You should work on this project together. This may mean all group members trying different things and coalescing on a final approach. Trying things that fail is still a contribution. Failure can be included in the report if something meaningful was learnt.

To work physically separately, you should:

a) arrange a suitable **discussion forum** for your group such as a WhatsApp group, slack, etc.
b) arrange a suitable **file sharing location** such as github, OneDrive, Dropbox, or GoogleDrive.
c) Collaboratively decide the final content, merging all versions of the analysis.

You should finalise the project content at least 48 hours before the deadline, so that individual writeups can be written.

## Marking Criteria

The mark ranges and descriptions in normal type below are the University of Bristol Generic Marking criteria that apply to any assessment at the University - these can be found at https://www.bristol.ac.uk/academic-quality. The descriptions in bold type are additional maths-specific criteria introduced primarily to clarify the descriptors in the case of marking maths examinations.

| **0-100 scale** | **Criteria to be satisfied**<br>University generic marking criteria in normal type, **Maths-specific marking criteria in bold** |
|---|---|
| 100<br>94<br>89 | • Work would be worthy of dissemination under appropriate conditions<br>• Mastery of advanced methods and techniques at a level beyond that explicitly taught<br>• Ability to synthesise and employ in an original way ideas from across the subject<br>• In group work, there is evidence of an outstanding individual contribution<br>• Excellent presentation<br>• Outstanding command of critical analysis and judgement **and**<br>• **Work develops concepts not directly presented in course material or uses known concepts to answer hard, unfamiliar questions that require calculations/methods not similar to any course material**<br>• **An elegance of mathematical work beyond that expected for the level of the course**<br>• **Of a quality that could be distributed to fellow students as an example of exceptional work** |
| 83<br>78<br>72 | • Excellent range and depth of attainment of intended learning outcomes<br>• Mastery of a wide range of methods and techniques<br>• Evidence of study and originality clearly beyond the bounds of what has been taught<br>• In group work, there is evidence of an excellent individual contribution<br>• Excellent presentation **and**<br>• **On standard but unfamiliar problems, carrying out calculations with no errors of understanding**<br>• **Demonstrates a high level of technical competence with very few mistakes of any kind**<br>• **Great clarity in mathematical arguments** |
| 68<br>65<br>62 | • Attained all the intended learning outcomes<br>• Able to use well a range of methods and techniques to come to conclusions<br>• Evidence of study, comprehension and synthesis beyond the bounds of what has been explicitly taught<br>• Very good presentation of material<br>• Able to employ critical analysis and judgement<br>• Where group work is involved there is evidence of a productive individual contribution **and**<br>• **Able to make a good attempt at standard but unfamiliar problems, with some minor errors**<br>• **Demonstrates technical competence, perhaps with some shortcomings**<br>• **Clear mathematical arguments** |

| 0-100 scale | Criteria to be satisfied<br>University generic marking criteria in normal type, **Maths-specific marking criteria in bold** |
|---|---|
| 58<br>55<br>52 | • Some limitations in attainment of learning objectives, but has managed to grasp most of them<br>• Able to use most of the methods and techniques taught<br>• Evidence of study and comprehension of what has been taught<br>• Adequate presentation of material<br>• Some grasp of issues and concepts underlying the techniques and material taught<br>• Where group work is involved there is evidence of a positive individual contribution **and**<br>• **Able to start standard but unfamiliar problems but with significant errors**<br>• **Able to complete competently "bookwork" questions that have been seen in the course material** |
| 48<br>45<br>42 | • Limited attainment of intended learning outcomes<br>• Able to use a proportion of the basic methods and techniques taught<br>• Evidence of study and comprehension of what has been taught, but grasp insecure<br>• Poorly presented<br>• Some grasp of the issues and concepts underlying the techniques and material taught, but weak and incomplete **and**<br>• **Able to complete "bookwork" questions that have been seen in course material with few errors**<br>• **Gaps or inconsistencies in the mathematical argument** |
| 35 | • Attainment of only a minority of the learning outcomes<br>• Able to demonstrate a clear but limited use of some of the basic methods and techniques taught<br>• Weak and incomplete grasp of what has been taught<br>• Deficient understanding of the issues and concepts underlying the techniques and material taught **and**<br>• **Able to reproduce work seen in course material, but with some errors** |
| 7-29 | • Attainment of nearly all the intended learning outcomes deficient<br>• Lack of ability to use at all or the right methods and techniques taught<br>• Inadequately and incoherently presented<br>• Wholly deficient grasp of what has been taught<br>• Lack of understanding of the issues and concepts underlying the techniques and material taught **and**<br>• **Unable to reproduce satisfactorily even "bookwork" questions that have been seen in course material** |
| 0 | • No significant assessable material, absent or assessment missing a "must pass" component |