National College of Ireland

BSc (Hons) / HCert in Data Science – Year 1 – Full-time (BSHDS1, HCDS1)
BSc (Hons) / HCert in Computing – Year 1 – Full-time (BSHC1, HCCOMP1)
BSc (Hons) in Computing (Advanced Entry) – Year 2 – Full-time (BSHC2_AE)

Semester 1, 2025/26

Release Date: Monday 17th November 2025
Due Date: Tuesday 16th December 2025 @ 23:55

Lecturer:
Arghir-Nicolae Moldovan, Charlyn Rosales

Introduction to Data Science
CA Type: Individual Project

WEIGHT: 70% of overall marks for this module. The assignment will be marked out of 100.

INTRUCTIONS:

- Read the below description carefully and make sure you **follow the requirements**.
- Also read the **additional information provided on Moodle**.
- **Ask the lecturer** if you need additional clarifications.
- **Attend the classes** as the lecturer may provide additional formative feedback.
- This is an individual assessment, so you are **not allowed** to collaborate or share your work with other colleagues.

SUBMISSION DETAILS:

Check on Moodle for instructions and submission links. The following deliverables must be submitted:

1) The report in **MS Word** document format
2) Additional files **as separate PDF or Word** documents:
   a) Declaration of Ethics Consideration
   b) Ethical Review Application Form
   c) AI Acknowledgement Form
   d) Appendix with evidence of work

Late submissions without an approved extension through NCI360 will be subject to the standard NCI penalties (i.e., 10% for up to 1 week, plus 5% for any additional week, after 6 weeks you get ZERO). But note that there are other strict deadlines to provide the results to the Exams Office.

TURNITIN: All report submissions will be electronically screened for evidence of academic misconduct (e.g., plagiarism and collusion). The use of AI tools for any part of the work must be summarized and referenced in the project report and detailed in the AI declaration (see Moodle for additional details).

Duration of CA: As per the above dates.
Attachments: None

Task

In this project, you are tasked to identify and document appropriate sources of your own **personal data footprint** (without providing the data), their characteristics and structure (if online), and prepare ethical review documentation for a data science study on the personal data footprint. This study should be mapped to the key stages of either **KDD or CRISP-DM methodology**.

Emphasis should be placed on how you ensure adherence to relevant privacy, ethical policies and, if appropriate, legal frameworks for the proposed study. You should also note how participant risks are mitigated. You should report all these aspects by completing and submitting (to the lecturer not the NCI Ethics Committee) a Declaration of Ethics Consideration and an Ethical Review Application Form.

You should not reveal personal information as part of this project but only describe the data structure and how it could be used to formulate a small-scale data science project. You should, however, note and discuss the following:

- How the data would be collected;
- The high-level approach to analysing it;
- The expected outcome(s) of the project and the implications of the findings;
- Note aspects of the study that could cause privacy, legal and/or ethical concern.

Note that you **are not required to do in-depth practical data analysis** beyond exploring, understanding and summarising the datasets.

Criteria for Datasets

The following criteria should be considered when searching for personal datasets:

- The data should be related to your **personal activity** on various online websites / apps.
- **Create a longer list** of websites / apps where you are more active.
- Search online to find out if they allow you to **access and download** your personal data.
- **Select a total of 3 personal datasets** that you will use for the project.
    - The **3 datasets** should be complementary and/or from different companies.
    - You can also select **4 datasets** if 2 datasets are from websites/apps belonging to the same large company (e.g., Google, Meta).
- Ideally, the above selection should **include 1 AI-related dataset** that comprises personal data from your interactions with an AI tool.
    - Create a list of AI tools that you use and check if any of them allow you to export any personal data.
    - If you cannot include any AI-related personal dataset, explain why in the data section.
- Note that some datasets may be large and complex and it would not be feasible to say that your study will analyse all datasets and all of their contents. So, in the data section of your project report you should include separate summary tables, for example:
    - 1st table to summarise the complete personal datasets.
    - 2nd table to summarise some parts / files that will be the focus of your data science study (e.g., files that are relevant for 3-4 analyses, that potentially would lead to more interesting findings, that allow you to make connections between the datasets and lead to combined findings obtained from multiple datasets, etc.).

Reference Management

The following guidelines must be followed for referencing:

- All students must use a **Reference Management System (RMS)** such as Zotero or Mendeley.
- All **in-text citations and references** must be inserted in the MS Word report with the RMS.
- IEEE referencing style must be used.
- Make sure you carefully check each reference in your RMS library and fill any important details that are missing as sometimes the RMS cannot automatically get them correctly.
- Consult this guide on which details are important (but make sure to also add location and publisher for conference papers / book chapters, and DOI for any reference that has one, add URLs for webpages, tools, etc): https://libguides.ncirl.ie/referencingandavoidingplagiarism/ieee
- Note that you are not required to write a literature review section.
- However, you must reference all the relevant resources that you have used or consulted for the project (e.g., additional references used in the introduction to set the context of the project, webpages describing how to download personal datasets, generic or data science tools used for exploring the datasets, AI tools used for any part of the work, research articles that inspired the methodology, references consulted when writing the ethics section, etc.).

Report Length, Template, OneDrive

The following guidelines must be followed when writing the CA report:

- The report must be **concise**.
- The expected page length of the project report is **3 pages in total**. This limit includes all the text and figures / tables from title to references (including them).
- The report can be shorter / longer but you may be penalised (i.e., for not including sufficient details or for not making good use of the space / template.
- Write the report in a **MS Word** document**.**
- Write the report in your **NCI OneDrive** account **gradually** as you work (do not just copy everything at the end).
- The report must be formatted in the **IEEE double column template**. The IEEE template is available on Moodle, and can also be found at the URL: https://www.ieee.org/conferences/publishing/templates.html
- Add your student name and other details, using the IEEE authors style at the top of page 1.

Appendix

Please follow the below requirements:

- Each student must create and submit a separate appendix document using a **generic 1 column** template in MS Word.
- The appendix must include screenshots of the **report version history** from OneDrive (i.e., showing the your name and all the dates when you edited the file).
- Additionally, the appendix should include screenshots of the **RMS project collection** showing the complete list / table of references with main columns (i.e., authors, title, publication year, and **date when each reference was first added** to the collection).
- The appendix must also include a few screenshots as proof that you **downloaded and explored** the datasets to understand their contents (e.g., properties of each dataset main folder showing size and number of files, a text editor search result to find number of instances in a semi-structured file such as JSON/ XML/ HTML). Note that students must **be careful** not to include any personal data of other users in the screenshots (e.g., chat messages with other users).

Report Structure

The project report should be structured as follows:

- **Abstract**: A roughly 150-word executive summary of the study.
- **Introduction**: Briefly set the context, motivate the work, and summarise the objectives (for example what are you trying to find out).
- **Data**: Summarise the process/strategy you followed to identify suitable personal datasets that meet the criteria specified above. Summarise the characteristics of the datasets in large tables spanning over 2 columns (e.g., size in MB, number of files, files format, structured/semi-structured/unstructured, number of attributes and their type, number of instances, etc.). Summarise both the complete datasets and separately the specific parts / files that you will select from the datasets and will be the focus of your data science study. Make sure to include references to the sources of your datasets for example the webpages detailing how to access and download the datasets.
- **Methodology**: Essentially, provide a step-by-step description of how you would apply KDD or CRISP-DM to a study on the personal footprint data. This section should be broken into subsections corresponding to the different steps of the chosen methodology. Each subsection should provide a discussion of the methods and/or technologies to be applied at that step of the methodology. The methodology steps should include some citations to past research papers that inspired your methodology, for example papers that analysed data from the same websites / apps that you selected (i.e., this is to provide evidence that your specified methods or technologies were used before, not that you just wrote based on your own thinking with little or no understanding).
- **Ethical Considerations:** Briefly summarise the aspects of the study that could cause privacy, legal and/or ethical concern, and how did you mitigate the risks (these should be detailed in the separate forms). Clearly indicate if the ethical risks apply only to you or if personal datasets also include data generated by other people (e.g., communication messages, collaborative data generation, etc.). This section should also include some citations to past research papers or other sources (e.g., check how past research papers using human participants summarised the ethical / privacy aspects, also identify and reference some papers / sources that discussed ethical / privacy issues and mitigation solutions relevant in your study context or to data science methods / technologies that you specified in the methodology section).
- **Findings:** Discuss 3-4 interesting findings that you observed from the high-level data exploration you conducted to understand and summarise the datasets. Discuss how selecting and considering files from different personal datasets led to more interesting findings (than analysing each dataset independently). It would be great if you can summarise some findings in summary tables (e.g., descriptive stats / numeric summaries), but the short page length will not allow to include many figures. You are not expected to do in-depth analysis of the datasets, so the findings can be broader and inspired by past research papers.
- **Conclusions and Future Work**: Briefly discuss the main conclusions of the project, what would you do differently or how would you extend the work if you had more time and in-depth knowledge of DS methods. Discuss if you consider changing your interaction with any of the online websites / apps (e.g., to reduce their data collection, improve your privacy).
- **References**: A complete bibliography of all the resources used or consulted for the project (e.g., additional references used in the introduction to set the context of the project, webpages describing how to download personal datasets, generic or data science tools used for exploring the datasets, AI tools used for any part of the work, research articles that inspired the methodology, references consulted when writing the ethics section, etc.).

Marking Rubric

Criteria	H1 (> 70%)	H2.1 (60 – 69%)	H2.2 (50 – 59%)	Pass (40 – 49%)	Fail (< 40%)
Objectives (10%)	Challenging objectives are well presented, met, and thoroughly discussed.	Reasonable objectives are well presented, met, and discussed.	Reasonable objectives are clear and are mostly met.	There are clear objectives, which are at least partially met.	Cannot discern objectives and/or if objectives were met.
Data (20%)	Detailed description of data search process, excellent variety and use of tools. Excellent selection of 3-4 suitable, relevant and complementary datasets. Excellent description of the data and use of summary table.	Good description of data search process, good variety & use of tools. Good selection of 3 suitable, relevant and complementary datasets (maybe 2 from 1 company) Good description of the data and use of summary tables.	Appropriate description of data search process, appropriate use of some tools. Appropriate selection of 2 suitable relevant, and complementary datasets. Appropriate description of the data and use of tables.	Basic description of data search process, and basic use of tools. Suitable, but somewhat trivial selection of 1-2 datasets. The data was described to some extent, basic tables.	Poor description of data search process, and poor use of tools. Poor selection, perhaps not suitable or not personal datasets. The data was poorly described, no summary table. Insufficient evidence that the datasets were explored.
Methodology (30%)	All stages of KDD or CRISP-DM have been thoroughly and accurately applied and detailed. Deep discussion of methods or technologies to be applied.	All stages of KDD or CRISP-DM have been appropriately applied and detailed. Good discussion of methods or technologies to be applied.	Most stages of KDD or CRISP-DM have been appropriately applied and described. Adequate discussion of methods or technologies to be applied.	Basic attempt to apply KDD or CRISP-DM. Some appropriate methods or technologies to be applied, are discussed, but the description lacks depth.	Deviates significantly from KDD or CRISP-DM, the methodology steps were poorly understood and described. It is unclear what methods or technologies will be applied.
Ethics (15%)	Excellent understanding of potential ethical issues and how they may be mitigated. Detailed forms, the proposal would pass ethical review.	Good understanding of potential ethical issues and how they may be mitigated. Complete forms, the proposal would pass ethical review.	Acceptable understanding of the potential ethical issues and how they may be mitigated. The proposal would probably pass ethical review, but with minor revisions.	Some understanding of the potential ethical issues and how they may be mitigated. The proposal would probably pass ethical review, but with major revisions.	Limited to no understanding of the potential ethical issues and how they may be mitigated. The proposal would not pass ethical review.
Findings & Conclusions (15%)	Excellent presentation and interpretation of 3-4 findings, insightful conclusions & future work.	Good presentation and interpretation of 3 findings, detailed conclusions & future work.	Appropriate presentation and interpretation of 2 findings, adequate conclusions & future work.	Basic presentation and interpretation of 1-2 findings, some conclusions & future work.	Little to no findings, poor or limited conclusions & future work.
Report Quality, Referencing, Appendix (10%)	Well-structured and written, with no (large) language errors. The IEEE double-column template and page length is well adhered to. References are complete, appropriately and correctly used. Exemplary use of RMS features. Very comprehensive appendix.	Report has a few language and/or style errors. IEEE template and length limit are adhered to. References are complete and correctly used. Good use of RMS features. Detailed appendix with sufficient evidence.	Report is readable with some language and/or style errors. IEEE template is largely adhered to. References are mostly complete and correctly used. Adequate use of RMS. Appropriate appendix but could be more detailed.	Report is readable with many language and/or style errors. IEEE template may have been broken. References are few and/or mostly incomplete. Basic or limited use of RMS. Appendix has some but incomplete evidence.	Littered with typos, and/or poor use of English. IEEE template was not used or broken significantly. References (if any) are probably incomplete and poorly used. No use of RMS.