ASSIGNMENT 3 – BIG DATA ANALYSIS

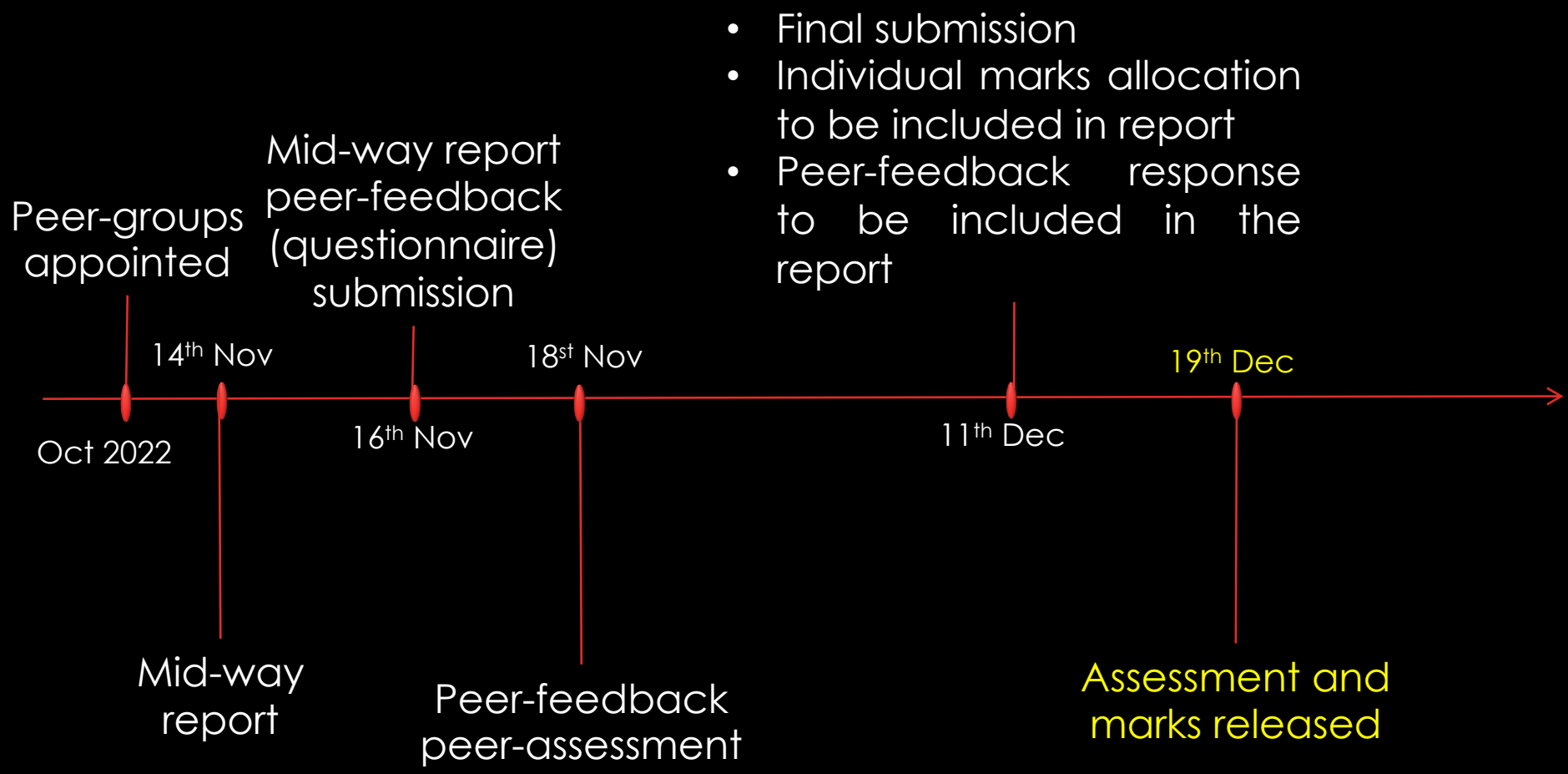
Alessandra Mileo
Alessandra.mileo@dcu.ie

DELIVERABLES

- Mid-way Report (max 500 words, approx. 2 pages)
  - What dataset
  - What technology (and why)
  - What analytics (e.g what you want to gain from the analytics on the data)
  - The plan for the team roles and tasks
- Final Report (max 3 pages) – details on content later

  Suggestion: living document for the team (use googledoc or GitLab Wiki)
- Application/code –
  - Source code (gitlab)
  - Max 5 minutes screencast demo of the different components running

TIMELINE


Peer-groups appointed
Mid-way report peer-feedback (questionnaire) submission
Final submission
Individual marks allocation to be included in report
Peer-feedback response to be included in the report
14th Nov
18st Nov
19th Dec
Oct 2022
16th Nov
11th Dec
Mid-way report
Peer-feedback peer-assessment
Assessment and marks released

TIMING

- Groups finalised (done)
- Intro to Assignment 3 (today, week 7)
- Mid-Way Report due (Mon 14th Nov – Week 9)
- Peer-Feedback due (Wed 17th Nov– Week 9)
- Peer-Feedback assessed (Fri 19th Nov – Week 9)
- Final Delivery (Mon 12th Dec – Week 12+)

REQUIREMENTS

- Uses Big Data Processing (batch or stream)
- Uses a large public dataset, example sources:
  - https://www.kaggle.com/datasets
  - https://data.london.gov.uk/
  - http://hadoopilluminated.com/hadoop_illuminated/Public_Bigdata_Sets.html
  - https://aws.amazon.com/public-datasets/
  - https://github.com/caesar0301/awesome-public-datasets

MIDWAY REPORT FEEDBACK GUIDELINES

Criteria	Feedback 1	Feedback 2	Feedback 3
General Idea	What I find bad/can be improved in this idea is…	What I find good/has potential in this idea is…	One problem/challenge you might encounter is…
Technology fit	Identify 2 technologies and say if they fit or not	For technology that fits, suggest one way the app could be expanded using such technology	For technology that does not fit, say why
Task allocation and roles	Is the task allocation fair/reasonable?	If yes, identify who you think is a good leader candidate and why	If no, say why, identifying at least 1 way to improve it

PEER FEEDBACK

- Peer Feedback: each group will be assigned a peer-group to provide feedback to (Wed 17th Nov)
- Feedback has the form of a questionnaire I provide and will be filled in and submitted as a .pdf file (on Loop)

Feedback Questionnaire Form

Your group ID:
Group ID of the report being reviewed:

Section 1 – General Idea
What I find bad/can be improved in this idea is

What I find good/has potential in this idea is

One problem/challenge you might encounter with this idea is

Section 2 – Technology

Technology 1
One key technology used in this project is
Does this technology fit the general idea?
☐ Yes ☐ No
If Yes, one way the app can be expanded/improved using such technology is
If No, why?

Technology 2
Another key technology used in this project is
Does this technology fit the general idea?
☐ Yes ☐ No
If Yes, one way the app can be expanded/improved using such technology is
If No, why?

Section 3 – Task allocation and user roles
Is the task allocation fair/reasonable?
☐Yes ☐No
If Yes, who do you think is a good leader candidate for the group and why?

If No, say why and indicate at least one advice for improving fairness in task allocation.

MIDWAY REPORT PEER-ASSESSMENT

- Each group will receive a score for the quality of the feedback provided (max 7 marks) by their peer group through Loop (due Fri 19th Nov)

Criteria	Poor (0-2)	Fair (3-5)	Good (6-7)
Quality of the peer feedback provided	Incomplete and of little use (explain why)	Complete but only partially helpful (identify what was not helpful and what was)	Complete and very helpful (identify what part of the feedback you will take action on

Peer Group

(CONTD) MIDWAY REPORT PEER-ASSESSMENT

- Each group will provide a response to the feedback as a section of their final report, and will receive a score for the quality of the response to feedback (max 7 marks) as part of the final group mark.

Criteria	Poor (0-2)	Fair (3-5)	Good (6-7)	
Quality of the peer feedback provided	Incomplete and of little use (explain why)	Complete but only partially helpful (identify what was not helpful and what was)	Complete and very helpful (identify what part of the feedback you will take action on	Peer Group
Peer feedback response	Comments have not been addressed, neither by taking them on board nor by explaining why not	Comments have been partially addressed	Completely answered to all comments either by taking them on board or by explaining why not	Lecturer

FINAL REPORT DOCUMENT

1. Introduction and motivation for the analysis and the choice of data and technology
2. What Data – identified souce, description of data
3. What Analytics – preparation, processing, querying, storing
4. Analytics task and results – insights you got from your analysis, appropriately visualised and explained
5. Related Work – any similar analysis used in any known application?
6. Challenges and lessons learned from working on the selected dataset with the selected technology
7. Responsibility statement (which group member did what?)
   - Include specific tasks and general roles
   - You can refer to http://www.belbin.com/about/belbin-team-roles to identify roles
   - Include individual contribution from each member
8. Response to peer feedback

GROUP PROJECT ASSESSMENT GUIDELINES

Criteria	0-3	4-8	9-14	15-19 (max 76)
Design	Design not appropriate for the problem the app aims to solve	Design addresses the problem but could be improved in data preparation and processing	Design could be improved either in data preparation or in processing	Design is well thought, data and processing properly identified and used
Technology fit (including interface)	Selected technology not appropriate or missing	Technology is appropriate but not always correctly/fully used	Technology is mostly appropriate and correctly used but lacks some functionalities	Technology is fully appropriate and correctly used, all functionalities are present and accessible
Report	Incomplete and poorly structured	Structured but with missing key elements and/or errors	Mostly Complete and structured but with imprecisions	Complete, structured, professional
Video Walkthrough	Poor and unstructured	Unstructured but mostly complete	Mostly complete and structured	Complete, well structured and of professional quality

INDIVIDUAL MARKING

- Each group receives **100 percentage points (%)** to be divided among group members (the sum of all percentages should be 100)

- Examples (group scoring highest marks in all categories):

	Group mark	Individual mark	Total Mark
Student 1	63	33.3% = 10	73
Student 2	63	33.3% = 10	73
Student 3	63	33.3% = 10	73

	Group mark	Individual mark	Total Mark
Student 1	63	50% = 10	73
Student 2	63	25% = 5	68
Student 3	63	25% = 5	68

(CONTD) INDIVIDUAL MARKING

- In genera:
  - Team member(s) with the max contribution m% will get 10 marks
  - Any other team member with contribution p% will get x marks where

$$x = p*10/m$$

- Allocation of % should be **UNANIMOUS** and documented in the section on "Responsibility statement" of the final report
- Special case: a student who did not participate at all in the project, will not get any mark for any of the category and will loose 50% of the marks

ASSIGNMENT - EXEMPLARS

Description	Big data technologies
Investigating DCU collaborations through citation graph analysis: Scraping DORAS and Google scholar to build a novel dataset	PySpark for cleaning, Spark GraphX in Scala for graph analysis
Book Recommender System: UCSD Book dataset (2.3M books and +800K users)	Hive for cleaning and reducing the data, PySpark ML libraries for a hybrid recommender system
STEM Salary expectations, variations and Covid-19: Keggle dataset	Pig for cleaning, PySpark/SparkSQL for simple analysis, SparkML deployed on google cloud dataproc for complex anlaysis