Exercises

1	2	3	4	5

Surname, First name

____

5ARB0 Data science: data acquisition and analysis
Trial exam

1	1	1	1	1	1	1
2	2	2	2	2	2	2
3	3	3	3	3	3	3
4	4	4	4	4	4	4
5	5	5	5	5	5	5
6	6	6	6	6	6	6
7	7	7	7	7	7	7
8	8	8	8	8	8	8
9	9	9	9	9	9	9
0	0	0	0	0	0	0

Instructions

- Write down your name and your student ID at the appropriate places above.
- Make sure that you enter your student ID by coloring the appropriate boxes and also fill it in in the top row.
- Use a black or blue pen and color the full box black or blue.
- Provide your answers in the answer box underneath a question. **If you need more space, you can use the extra pages provided at the end of the exam form.**
- If you use extra pages, make sure that you clearly mark which (sub)question you are answering on the extra pages.
- Hand in all pages.
- Do not remove the staple. If it is detached, check that you hand in all pages.

Permitted examination aids:

- calculator;
- dictionary.

Important notes

- This exam consists of 4 questions, each with sub-questions, and has 16 pages.
- The last two pages are empty and meant as extra space if you need this for answering the questions.
- If you want to make use of the extra space, clearly refer to these pages in your answer in the box underneath the question. Mark clearly on the extra pages which (sub)question you are answering.
- The time allotted for the exam is **two** hours (120 minutes).
- It is not permitted to leave the examination room within 15 minutes of the start and within the final 15 minutes of the examination, unless stated otherwise.
- **Examination scripts (fully completed examination paper, stating name, student number, etc.) must always be handed in, as well as any scrap papers you might use.**
- Students are not permitted to share examination aids or lend them to each other.

Regulations:

- Students are only permitted to visit the toilets under supervision.
- The house rules must be observed during the examination.
- The instructions of examiners and invigilators must be followed.
- No pencil cases are permitted on desks.

During written examinations, the following actions will in any case be deemed to constitute fraud or attempted fraud:

- using another person’s proof of identity/campus card (student identity card);
- having a mobile telephone or any other type of media-carrying device on your desk or in your clothes;
- using, or attempting to use, unauthorized resources and aids, such as the internet, a mobile telephone, etc.;
- having any paper at hand other than that provided by TU/e, unless stated otherwise;
- visiting the toilet (or going outside) without permission or supervision.

Question 1 - Data Acquisition

4p **1a** In data collection, one often distinguishes between *primary data collection* and *secondary data collection*. Explain what is meant by each type and give one example of each.

6p **1b** Analog data (*e.g.*, signals) acquired from sensors usually have noise. Such an issue can be mitigated in a pre-processing step by applying a filter to the signal. Explain how each of the following types of filters work.

- Low-pass filter.
- Rolling median filter.
- Band-pass filter.

5p **1c** Consider the following time series: [45, 30, 12, 55, 14, 18, 23, 50, 37, 22, 25, 25, 25, 30, 34, 42, 18, 21, 20, 35]. What would be the output if you filter this time series with a 4-period moving average filter? Explain also how you deal with the boundary effects.

5p **1d** In a data acquisition process, events are stored and tied to the moment they happen. These events can be of different types, *i.e.*, they can be composed of data types that differ for each event recorded (*e.g.*, heart rate measurement, exercise, meal, mood). An AI&ES student wants to store such events data using a CSV (comma separated value) file. Discuss whether this is a good choice for storing such data, motivating your choice. What are the advantages and/or disadvantages that you recognize?

Let us now consider a dataset composed of multiple time series. One of these time series has recorded heart rate, measured every 5 minutes. Another time series of the same dataset contains measurements of calories burned, again every 5 minutes. Answer the following questions.

3p **1e** Which steps could be applied to allow the use of these two time series together as a new stream of events?

2p **1f** Which type of information can be retrieved from these new events?

Question 2 - Data analysis

Suppose you were given the following data set of 4 samples in 2 dimensions:

$$x_1 = \begin{bmatrix} 2 \\ 1 \end{bmatrix}, \quad x_2 = \begin{bmatrix} 2 \\ 3 \end{bmatrix}, \quad x_3 = \begin{bmatrix} 4 \\ 1 \end{bmatrix}, \quad x_4 = \begin{bmatrix} 0 \\ 3 \end{bmatrix}.$$

You want to extract meaningful features from these points.

3p **2a** What does ”Principal Component Analysis” entail? Describe what is extracted and how the data is processed.

10p **2b** Compute the covariance matrix of this data set.

8p **2c** Suppose that the eigenvectors of the covariance matrix are:

$$Q = \begin{bmatrix} -2 & -3 \\ -3 & 2 \end{bmatrix}.$$

Map the data onto the principal components.

4p **2d** Suppose that the eigenvalues of the covariance matrix are λ_1, and λ_2 with $\lambda_1 >> \lambda_2 > 0$. What does this imply for the distribution of the data points?

Question 3 - Non-technical aspects

6p **3a** Name three laws, codes of conduct or guidelines that are relevant for research with humans and describe briefly what this law, code, or guideline is about.

9p **3b** Select one of the laws, codes, or guidelines from Question 3a. Give an example of a research study that this law applies to. Explain how this law is relevant for this type of research.

Consider the following study information on test persons, recruitment, and research methods.

Research question: Which conflicts occur on a company work floor?

Test persons: People working at company X, or the stakeholders at Company X will participate in the study. In total 20 people work at company X. Most of them can be described as

- Gender: men.
- Age range: 25-35.
- Educational level: HBO, WO (higher education, university),
- All participants are at least 18 years old.

Recruitment: Company X is a small company, and we have a complete list of software developers working at company X. We will send every one of them a personalised invitation, explaining the importance of the research, its benefits and the potential risks the study participation might entail.

Research methods:

- Observation of the participants during work, and in particular scrum events, such as sprint retrospectives and sprint planning. While observing, we will take notes of the situations, which will together form a diary.
- Survey to collect data from the participants on whether they are aware of any conflicts. Participants fill in several surveys during the study.
- Interview to gain insights into the participants' views
- Data collection and analysis of discussions from the previous year on the internal social platforms like TEAMS.
- Analysis of the notes from past meetings provided to us by company X.
- Participants will be observed during certain events, such as sprint planning and notes will be taken during those events.
- Participants will be interviewed. During the interview, a recording will take place and notes will be made. Then the results will be compared with the previously mentioned data.

10p **3c** Given the above description, identify three risks in one or more of the fields of ethics, privacy or security and their potential consequences. Discuss how you would mitigate these risks.

Question 4 - Data management

8p **4a** What are the four elements of FAIR data principles? Explain what is meant by each of them.

4p **4b** Explain what is meant by a glossary in a knowledge organization system.

The AHIMA Data Quality Management Model distinguishes amongst different characteristics of data quality. For each of the following characteristics, explain briefly what is meant by it.

2p **4c** Accessibility

2p **4d** Currency

2p **4e** Granularity

2p **4f** Timeliness

5p **4g** *A Dutch hospital invests large amounts of manpower and money for implementing an electronic health record (EHR) to digitize its information systems. The expectation is that the new digital system will help the hospital operate more efficiently and more effectively by having the right information at the right place at the correct time. After introducing the EHR, complaints start coming from different departments that obtaining the answers to queries takes a long time, and oftentimes the answers are incorrect. For example, it takes a full week to determine the number of patients who have undergone a particular procedure, and the answer returned may not be correct. Upon an examination of the problems, it is observed that although the medical staff have entered a lot of data into the system, most of the information is entered in the free-text fields of the EHR as physicians' or nurses' notes.*

Write a short essay discussing the data quality issues that you determine in this example, by using the data quality characteristics of the AHIMA data quality model as a reference.

Extra space for answering questions

5 Please refer clearly in your answer that you will make use of this extra space.

↷

You have reached the end of this exam.