Exercises

1	2	3	4	5

Surname, First name

Model solutions to the Trial Exam

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.

Data acquired directly from the original sources with a specific purpose in mind is called primary data collection. An example is data obtained from surveys regarding people's satisfaction with a product or service, for the purpose of determining customer satisfaction.

Access to data that was previously collected for a different (primary) purpose and is now being re-used for another purpose than the primary purpose is called secondary data collection. Hence, the primary data becomes secondary when its role switches to the use of people/project different than the original one that created it. An example is re-using data that was collected to treat patients for developing a machine learning model for early detection of a disease.

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.

A low-pass filter attenuates signals above a particular cut-off frequency. Hence, low frequency signals are allowed to pass through, and the high frequencies are filtered out.

Consider a set of consecutive data points (series). A rolling median filter slides a moving window of size *n* over the data points and replaces the value that it is on with the median value in the sliding window. This generates a new series containing all median values for every instance of the sliding window, which results in the filtered series.

A bandpass filter allows signals to pass within a particular frequency range (its bandwidth) and attenuates signals outside this range. Hence, lower and higher frequencies outside the bandwidth are filtered out.

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.

*Calculate the mean of all possible windows of size 4. The gives as output: [**45**, **41.25**, **33**, 35.5, 27.75, 24.75, 27.5, 26.25, 32.0, 33.0, 33.5, 27.25, 24.25, 26.25, 28.5, 32.75, 31.0, 28.75, 25.25, 23.5].*

The boundary effects occur start of the filter, when there are not four values in the signal known (see the red, bold, italic first three values above). In the example above, the series is prepended with the first value (45) to calculate the average of the values in the sliding window. Other alternatives might have been to prepend with 0 (zero padding) or not calculating a filter output for the first three values.

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?

CSV files are very simple to read and understand. They are compact, interchangeable, and useful when storing entries containing the same types of values. In other words, all entries should respect the same columns, i.e. a column always contains the same type of data. Therefore, if the data set is structured and every column in a table has similar entries, a CSV file is a good choice.

However, the data types are quite different in the example given in the question. A heart rate measurement has different characteristics than an exercise or a meal event. Then, it is not convenient to use a csv file for this type of data. For example, the csv file would have multiple columns that would be used only by specific events (e.g., mood does not have a “carbohydrates” value, but meal does). This increases the file size, and also reduce its readability. Hence, the choice of the student of using a csv file is not optimal.

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?

There is no guarantee that such events will happen in a synchronized manner, and so, in cases like this, a preprocessing step of data synchronization is needed.

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

Different types of information could be derived from this data set. For instance, by using both heart rate and calories burned together, one could infer different levels of cardio related activities, the periods of the day where someone performs such activities, cardio peaks and lows not tied to any relevant number of calories burned (anomalies), etc.

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.

It means finding an orthogonal basis corresponding to the directions in which the data varies maximally and projecting the data onto that basis.

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

The student has to determine the mean first:

$$\hat{m} = \frac{1}{4} \begin{bmatrix} 2+2+4+0 \\ 1+3+1+3 \end{bmatrix} = \begin{bmatrix} 2 \\ 2 \end{bmatrix}.$$

The mean is subtracted from the data:

$$x_1 - \hat{m} = \begin{bmatrix} 0 \\ -1 \end{bmatrix}, \quad x_2 - \hat{m} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad x_3 - \hat{m} = \begin{bmatrix} 2 \\ -1 \end{bmatrix}, \quad x_4 - \hat{m} = \begin{bmatrix} -2 \\ 1 \end{bmatrix}.$$

Compute the outer product of each sample with itself and sum:

$$\sum_{i=1}^{N} (x_i - \hat{m})(x_i - \hat{m})^\top = \left(\begin{bmatrix} 0 \\ -1 \end{bmatrix} \begin{bmatrix} 0 & -1 \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} \begin{bmatrix} 0 & 1 \end{bmatrix} + \begin{bmatrix} 2 \\ -1 \end{bmatrix} \begin{bmatrix} 2 & -1 \end{bmatrix} + \begin{bmatrix} -2 \\ 1 \end{bmatrix} \begin{bmatrix} -2 & 1 \end{bmatrix} \right)$$
$$= \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} + \begin{bmatrix} 4 & -2 \\ -2 & 1 \end{bmatrix} + \begin{bmatrix} 4 & -2 \\ -2 & 1 \end{bmatrix}$$
$$= \begin{bmatrix} 8 & -4 \\ -4 & 4 \end{bmatrix}$$

Normalize the outer product:

$$S = \frac{1}{4} \begin{bmatrix} 8 & -4 \\ -4 & 4 \end{bmatrix} = \begin{bmatrix} 2 & -1 \\ -1 & 1 \end{bmatrix}$$

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.

The formula for mapping the data onto the principal components is $z_i = Q(x_i - \hat{m})$ for each i.

First, subtract the mean from the data

$$x_1 - \hat{m} = \begin{bmatrix} 0 \\ -1 \end{bmatrix}, \quad x_2 - \hat{m} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad x_3 - \hat{m} = \begin{bmatrix} 2 \\ -1 \end{bmatrix}, \quad x_4 - \hat{m} = \begin{bmatrix} -2 \\ 1 \end{bmatrix}.$$

$$z_1 = \begin{bmatrix} -2 * 0 + -3 * -1 \\ -3 * 0 + 2 * -1 \end{bmatrix} = \begin{bmatrix} 3 \\ -2 \end{bmatrix}$$

$$z_2 = \begin{bmatrix} -2 * 0 + -3 * 1 \\ -3 * 0 + 2 * 1 \end{bmatrix} = \begin{bmatrix} -3 \\ 2 \end{bmatrix}$$

$$z_3 = \begin{bmatrix} -2 * 2 + -3 * -1 \\ -3 * 2 + 2 * -1 \end{bmatrix} = \begin{bmatrix} -1 \\ -8 \end{bmatrix}$$

$$z_4 = \begin{bmatrix} -2 * -2 + -3 * 1 \\ -3 * -2 + 2 * 1 \end{bmatrix} = \begin{bmatrix} 1 \\ 8 \end{bmatrix}$$

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?

It means the distribution of the data is much longer in one direction than in the other direction, i.e., it is a cigar-shaped ellipsoid.

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.

Declaration of Helsinki: statement of ethical principles in clinical research

Good Clinical Practice: international harmonized ethical and scientific standard for designing, recording and reporting clinical trials with human subjects

General Data Protection Regulation (GDPR): EU regulation on data protection and privacy

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.

For example, a research study in which participants are being asked to provide their home and work addresses to investigate the relation between travel distances and the use of public transport. Working with home addresses is processing personal data and therefore subject to the GDPR.

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:

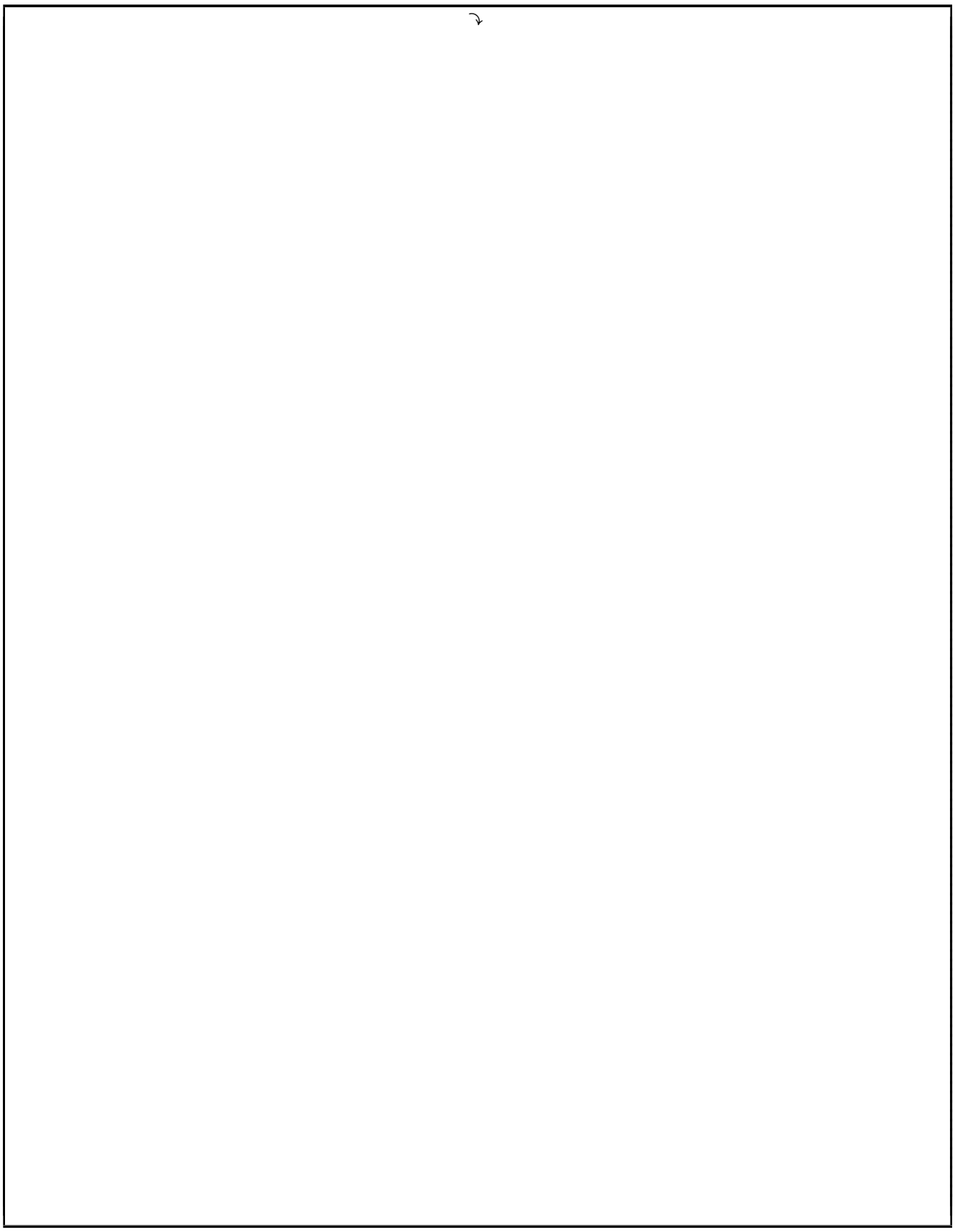
- Observation of the participants during work, and in particular scrum events, such as sprint retrospectives and sprint planning. While observing, we will take notes of the situations, which will together form a diary.
- Survey to collect data from the participants on whether they are aware of any conflicts. 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.

There is no clear informed consent procedure, people might participate in research without them knowing it. A mitigation strategy is to set up a proper IC procedure and explicitly ask for consent.

The company is small (20 people). Therefore, it is difficult to exclude people who do not want to participate in the real-life observations. A mitigation strategy is to organize separate meetings with the test persons.

Data collection from social internal platforms like TEAMS might contain sensitive and private information. Even people who consented might not recall if they wrote sensitive information on social platform. A mitigation strategy is, ensuring privacy by not reporting any information that can be traced to persons.



Question 4 - Data management

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

The data should be findable: the metadata and the data should be easy to find for *both humans and computers*.

The data should be accessible: data can be reached by authorized and authenticated parties (computers or humans).

The data should be interoperable: different parties can exchange and share data reliably, using a shared semantic and syntactic base.

The data should be re-usable: data can be used by parties other than those who did the primary data collection in an accurate and relevant manner.

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

A glossary is an alphabetical list of terms with associated definitions.

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

The level of ease and efficiency at which data are legally obtainable, within a well-protected and controlled environment.

2p **4d** Currency

The extent to which data are up to date; a datum value is up to date if it is current for a specific point in time, and it is outdated if it was current at a preceding time but incorrect at a later time.

2p **4e** Granularity

The level of detail at which the attributes and characteristics of data quality in healthcare data are defined.

2p **4f** Timeliness

The availability of up - to - date data within the useful, operative, or indicated time.

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.

The case describes that the system allows recording of structured data, but apparently, the physicians enter most of the data in the free text fields, as a narrative. This relates to the **data collection processes** in the AHIMA data quality (DQ) functions. Furthermore, there are aspects related to **application functions**, since the data quality does not seem to be appropriate for the purpose of use (e.g. automated reporting). Presumably, there are no issues related to the **warehousing function**, since the system is able to store the relevant information digitally, but the **analysis function** is hampered because of the difficulty of processing and analyzing unstructured text.

Further, one can observe the following regarding data quality characteristics.

Data accuracy could be good, but it should be noted that it is easy to make typos and mistakes in free text fields. Hence, this should be investigated further.

Data accessibility is problematic, since the information that is required is not available easily for the purpose in mind (e.g. reporting), even though the data is available digitally.

Data consistency is at risk, since different care personnel might be (unconsciously) using different definitions for concepts and information.

Data currency is not guaranteed, since most of the information seems to be entered manually by humans, and it is not known if the information entered is the most current (or up to date) one.

Data definition is also not guaranteed, since the text entered might entail different definitions across people and departments.

Since the data is entered by professionals who know what they do, it can be expected that the risks for the **relevance**, **comprehensiveness** and the **granularity** of data are not large.

It is not possible to make a statement from the available information for the **data precision** and **data timeliness** characteristics. It can be assumed that these are in order, however, since the complaints seem to be more about the queries and reporting, and not so much about the lack of precision or timeliness of the data.

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.