

**SEMESTER 1 EXAMINATIONS 2019/2020**

**MODULE:** CA682, CA682D - Data Management and Visualisation

**PROGRAMME(S):**

MCM M.Sc. in Computing

CAPT PhD-track

ECSA Study Abroad (Engineering & Computing)

CAPM MSc

MDPP MA in Data Protection and Privacy Law

GCAI Grad Cert in Artificial Intelligence

ECSAO Study Abroad (Engineering & Computing) **YEAR OF STUDY:** 1,2,3,O,X

|  |  |  |
| --- | --- | --- |
| **EXAMINER(S):** |  |  |
| Dr. Suzanne Little | (Internal) | (Ext:6052) |
| Dr. Brian MacNamee | (External) | External |
| Dr. Dan Emanoil Grigoras | (External) | External |

**TIME ALLOWED:** 3 Hours

**INSTRUCTIONS:** Answer **four (4)** questions from the five available.

**PLEASE DO NOT TURN OVER THIS PAGE UNTIL YOU ARE INSTRUCTED TO DO SO.**

The use of programmable or text storing calculators is expressly forbidden.

Please note that where a candidate answers more than the required number of questions, the examiner will mark all questions attempted and then select the highest scoring ones.

***There are no additional requirements for this paper.***

CA682, CA682D - Data Management and Visualisation

Semester 1 Examinations 2019/2020 Page 1 of 7

***QUESTION 1 [TOTAL MARKS: 25]***

**Q 1(a) [9 Marks]**

Given the following brief to design a system for a data collection task:

(i) List three (3) important questions you would ask your client.

(ii) Describe the data and/or file formats that you are likely to use in collecting the data.

(iii) Suggest a type of database system to use for this project, giving a reason for your choice.

“ Local councils in Ireland want to understand water usage patterns and customer sentiment in response to proposed water charges. They have historical records of water usage (Litres per day per household), data from water processing plants (e.g., volume processed per day, record of faults, duration of outages, cost per litre processed), maps of the existing water pipelines, access to Central Statistics Office data with population information (e.g., density, age, rate of change) and have commissioned a survey of Irish adults (e.g., opinion on water charges, how they use water). They want to identify where future investment in infrastructure should occur. ”

Here is a comprehensive response to your query, utilizing the information provided in the sources and our previous conversation:

### Understanding Water Usage Patterns and Customer Sentiment

#### (i) Important Questions for the Client

Before designing a system for this data collection task, it's crucial to gain a deeper understanding of the client's needs and objectives. Based on the provided brief, here are three important questions to ask the local councils:

* **What specific insights are you looking for regarding water usage patterns?** Understanding the council's specific goals will shape the data collection and analysis process. For example, are they interested in identifying households with unusually high water consumption, or are they focused on analyzing usage trends over time and correlating those trends with factors like population density or age?
* **How will the customer sentiment data be used to inform investment decisions?** It's essential to clarify how the survey data on water charges and water usage habits will be integrated with the other data sources to guide infrastructure investments. Will negative sentiment in specific areas lead to prioritizing upgrades even if those areas have relatively low water usage?
* **What are the desired outputs and reporting requirements for this project?** Clarifying the format and frequency of reports, as well as the level of detail required, will ensure that the data collection and analysis efforts align with the council's expectations. Do they require interactive dashboards to visualize water usage patterns in real-time, or are they primarily interested in summary reports with key findings and recommendations?

#### (ii) Data and File Formats

The brief outlines a variety of data sources already available to the councils. The data and file formats likely to be used in collecting the data include:

* **Historical Water Usage Records:** These records are likely stored in a database or spreadsheet format, such as **CSV (Comma-Separated Values)** or **XLSX (Microsoft Excel)**. The data will include fields like household ID, date, and water consumption in litres per day.
* **Water Processing Plant Data:** Data from water processing plants could be in various formats, including **sensor readings, log files, and maintenance records**. These may be in proprietary formats specific to the plant's systems. **It would be important to assess the compatibility of these formats with the chosen database system.**
* **Maps of Water Pipelines:** These maps might be in **GIS (Geographic Information System)** formats like **Shapefiles** or **GeoJSON**, containing data on pipeline location, capacity, and age.
* **Central Statistics Office (CSO) Data:** CSO data is typically available in **CSV or Excel formats**. This data will provide information on population density, age distribution, and other relevant demographic factors.
* **Survey of Irish Adults:** The survey data will likely be collected using **online survey tools or paper-based questionnaires**. The resulting data will likely be in a spreadsheet format, such as CSV or XLSX, and will include demographic information, opinions on water charges, and details about water usage habits.

#### (iii) Database System

A **relational database management system (RDBMS)** would be a suitable choice for this project. RDBMS, like Oracle, PostgreSQL, MySQL/MariaDB, SQLite, IBM DB2, Microsoft SQL Server, store data in tables with relationships between them. This type of database offers several advantages for managing the council's data:

* **Structured Data:** RDBMS are well-suited for handling structured data like the water usage records, CSO data, and survey results. The data can be organized into tables with defined columns and data types, ensuring consistency and integrity.
* **Relationships:** RDBMS allow for establishing relationships between different tables, which is crucial for this project. For instance, you can link household water usage data with their location on the pipeline map and with demographic information from the CSO data. This relational structure facilitates complex queries and analysis.
* **Querying and Analysis:** RDBMS utilize **SQL (Structured Query Language)** for querying and data manipulation. SQL is a powerful and widely used language, enabling the councils to perform sophisticated analyses to identify water usage patterns and correlate them with other factors. They could, for instance, identify areas with high water usage, aging infrastructure, and negative sentiment towards water charges, which could be prioritized for investment.
* **Data Integrity and Consistency:** RDBMS enforce data integrity through constraints and relationships, ensuring that the data remains accurate and consistent over time. This is particularly important for maintaining the quality of the data used for decision-making.

While RDBMS are generally considered "adequate for all tasks but not excellent at any of them", they are easy to use, have low resource requirements, are well-supported by various software, are familiar to many data professionals, and are appropriate for the volume of data in this project. For these reasons, an RDBMS would be a practical and effective choice for this data collection and analysis project.

It is important to note that other database systems, such as **NoSQL databases**, might be considered if the councils need to handle unstructured data, require greater flexibility in data modeling, or are dealing with massive datasets. However, based on the information provided in the brief, a relational database system appears to be the most appropriate choice for this particular application.

**Q 1(b) [8 Marks]**

Categorise the following data attributes as either Qualitative or Quantitative; Discrete or Continuous (if appropriate) and Nominal or Ordinal or Interval or Ratio.

(i) Type of pet (e.g., cat, dog, bird, fish) (ii) Number of pets currently owned

(iii) Weight of the pets

(iv) Happiness of pet owners (self-rated from 1 to 5).

Let's categorize the given data attributes using the definitions from the sources:

* **Qualitative data** describes qualities or characteristics. It is **categorical** and generally **non-numerical**. Examples include gender, country of origin, favorite color. **Limited mathematical functions** can be performed on qualitative data.
* **Quantitative data** represents quantities or measurements. It is **numerical**, and **"all the maths"** can be performed on it. Shoe size, temperature, and bank balance are examples of quantitative data.
* **Discrete data** is quantitative data that can only take **certain values** with "gaps" between them. This data type often results from **counting**, like the number of times someone attended an event or the number of crimes reported.
* **Continuous data**, also quantitative, can theoretically take **any value** within a given range. The **precision of the measurement** depends on the measuring device's capability. Continuous data usually results from **measurements**, like cholesterol levels or the time required to complete a task.

The **four levels of data measurement**, NOIR (Nominal, Ordinal, Interval, Ratio), are:

* **Nominal:** These are labels or categories with **no inherent order**. Examples are gender, department, and language. Nominal data can only be compared for equality or set membership.
* **Ordinal:** Ordinal data has labels **with order**. Examples are temperature (very hot, hot, warm, mild), medals (gold, silver, bronze), and Likert scales. While they can be arranged in order, they cannot be added or subtracted.
* **Interval:** Interval data uses numbers with **proportionate spaces between values**. This means we can talk about the "difference" between values. Income, shoe size, and temperature in Celsius or Fahrenheit are examples of interval data. Interval data **lacks a true zero point**.
* **Ratio:** Ratio data also uses numbers but **includes a true zero point**. This means we can multiply and divide these values. Age, amount of rainfall, book sales, and temperature in Kelvin are examples of ratio data.

### Categorizing the Pet Attributes

Now let's categorize the provided pet attributes based on the above definitions:

#### (i) Type of Pet

* **Qualitative:** It describes a characteristic (the species of the pet).
* **Nominal:** There's no inherent order among "cat," "dog," "bird," and "fish."

#### (ii) Number of Pets Currently Owned

* **Quantitative:** It represents a countable quantity.
* **Discrete:** You can only own whole numbers of pets.
* **Ratio:** There's a true zero point (meaning "no pets").

#### (iii) Weight of the Pets

* **Quantitative:** It's a measurable value.
* **Continuous:** A pet's weight can theoretically take any value within a range.
* **Ratio:** There's a true zero point (meaning "no weight").

#### (iv) Happiness of Pet Owners (Self-Rated from 1 to 5)

* **Qualitative:** It represents a subjective perception of happiness.
* **Ordinal:** The scale from 1 to 5 implies an order (1 being the least happy and 5 the most happy), but the differences between each level are not necessarily equal.

This categorization provides a clear understanding of the nature of each attribute and how it can be analyzed.

**Q 1(c) [4 Marks]**

Which of the following descriptions of data ([A], [B] or [C]) are mostly likely to be

classified as “big data”? Briefly explain your reasoning, including any assumptions, referring to the standard “V’s” for defining “big data” .

[A] The “Titanic” dataset showing passenger details from the final voyage of the ship. [B] Records from Spotify of the tracks listened to by each user (est. 232M users).

[C] Sales records from the DCU merchandise store.

### "Big" Data: Spotify vs. Titanic vs. DCU Merchandise

"Big data" is often characterized by the **four Vs**: **Volume, Variety, Velocity, and Veracity**. Here's how these characteristics apply to each of the data descriptions you provided:

#### [A] Titanic Dataset

* **Volume:** This dataset is relatively small. There were approximately 2,200 passengers and crew aboard the Titanic. Even with detailed information for each individual, the overall data volume would be manageable with traditional data processing techniques.
* **Variety:** The data likely includes passenger demographics, ticket class, cabin information, and survival status. While somewhat varied, it's still structured data that can be easily stored in a relational database.
* **Velocity:** The data was collected at a single point in time and remains static. There's no ongoing data generation.
* **Veracity:** The data is likely to be fairly accurate, given the historical significance of the event and efforts to document the passenger manifest.

#### [B] Spotify User Listening Records

* **Volume:** With an estimated 232 million users, the volume of data generated by Spotify's listening records would be immense. Each user's listening history, potentially spanning years, would include details on tracks, artists, albums, playlists, listening duration, and listening time. Storing and processing this amount of data would pose significant challenges.
* **Variety:** The data is varied, including structured data like track metadata and user demographics, as well as unstructured data like audio features and potentially social media interactions related to listening habits.
* **Velocity:** Spotify users generate data continuously as they stream music. The rate of data generation is high, requiring systems that can handle this constant influx of information.
* **Veracity:** While Spotify's data collection mechanisms are likely robust, user data can be inherently messy. For instance, inaccurate user profiles, skipped tracks, or incomplete listening sessions could affect data quality.

#### [C] DCU Merchandise Store Sales Records

* **Volume:** The volume of sales records for the DCU merchandise store would depend on factors like the store's size, customer base, and transaction frequency. It's likely to be much smaller than Spotify's data, especially given that DCU is a single institution.
* **Variety:** This data would include structured sales transactions with details like product IDs, quantities, prices, and customer information. The variety would be relatively low compared to Spotify's data.
* **Velocity:** Sales data is generated with each transaction, but the velocity would depend on the store's sales volume. It's unlikely to be as high as Spotify's streaming data.
* **Veracity:** Sales data is generally considered to be accurate, especially with modern point-of-sale systems. However, human error in data entry or system glitches could impact data quality.

### Conclusion

Based on the four Vs, **the records from Spotify ([B]) are the most likely to be classified as "big data"**. The sheer volume of data from millions of users, the variety of data types, the high velocity of data generation, and the potential challenges in ensuring data veracity all point towards the characteristics commonly associated with big data.

The Titanic dataset ([A]) is too small in volume, and the DCU merchandise store sales records ([C]) are likely to be manageable with traditional data processing techniques, assuming DCU's merchandise store is of typical scale for a single university.

**Q 1(d) [4 Marks]**

Describe the process of scraping data from a website. Give two (2) rules that you should remember when using this as data source.

Answer:

Web scraping is the process of automatically extracting data from websites. It involves using software tools to retrieve the underlying HTML code of a webpage and then parsing this code to identify and extract the desired information. This information can then be stored in a structured format for further analysis or use. The sources explain that this process typically involves three key steps:

1.

Parse HTML: The first step is to retrieve the HTML code of the target website. This code contains the structure and content of the webpage, including text, images, tables, and other elements. Tools known as "crawlers" or "spiders" are often used to fetch this HTML data.

2.

Match Patterns: Once the HTML code is obtained, the next step is to identify the specific data elements to be extracted. This often involves using regular expressions or other pattern-matching techniques to locate the relevant information within the HTML structure. The crawler needs to be able to recognize and distinguish between different elements like headings, paragraphs, lists, and tables.

3.

Identify Links (Repeat): Many websites organize information across multiple pages. To capture a complete dataset, the crawler must identify and follow links to other relevant pages. This can be done recursively, with the crawler extracting data from each linked page until all relevant content is gathered.

Rule 1

Rule 2

●

Data Quality: The sources emphasize the importance of data quality throughout the data lifecycle, particularly during the "Gathering" phase. Website structures can be inconsistent, and data might be presented in various formats, making it prone to errors or inconsistencies. Scraping tools may not always accurately extract data, leading to missing values, incorrect data types, or formatting problems. It's essential to thoroughly clean and validate the scraped data before using it for any analysis or decision-making. This might involve checking for missing values, correcting data types, standardizing formats, and removing duplicates.●

Respect Website Terms and Conditions: Websites often have terms of service or robots.txt files that specify how their data can be accessed and used. Violating these terms could have legal consequences. It's crucial to always review a website's terms of service and robots.txt file before scraping to ensure compliance and respect for the website owner's policies. Excessive scraping can also burden a website's server, potentially disrupting its normal operation. Be mindful of the frequency and volume of your scraping requests to avoid causing unintended problems.

***[End of Question 1]***

CA682, CA682D – Data Management and Visualisation

Semester 1 Examinations 2019/2020 Page 2 of 7

***QUESTION 2 [TOTAL MARKS: 25]***

**Q 2(a) [10 Marks]**

“ Local councils in Ireland want to understand water usage patterns and customer sentiment in response to proposed water charges. They have historical records of water usage (Litres per day per household), data from water processing plants (e.g., volume processed per day, record of faults, duration of outages, cost per litre processed), maps of the existing water pipelines (location, capacity), access to Central Statistics Office data with population information (e.g., density, age, rate of change) and have commissioned a survey of Irish adults (e.g., opinion on water charges, how they use water). They want to identify where future investment in infrastructure should occur. ”

Using the UK data archive Data Management Lifecycle (1. Creating Data; 2.

Processing Data; 3. Analysing Data; 4. Preserving Data; 5. Giving Access to Data; 6. Re-Using data), explain how you could go about this task and give examples of data analytics tasks, methods and tools that could be used at each stage where relevant.

**Q 2(b) [8 Marks]**

Metadata is used to describe and define the content of data so it can be found and used more easily. List and explain, with examples, four (4) potential issues with

using metadata created by human users.

### Identifying Areas for Infrastructure Investment Using the Data Management Lifecycle

To help the local councils in Ireland identify areas for future infrastructure investment based on water usage patterns and customer sentiment, the UK data archive Data Management Lifecycle can be used as a roadmap:

**1. Creating Data:**

* **Commissioning Surveys:** The councils have already commissioned a survey to gather data on public opinion regarding water charges and how people use water. This is a primary data source specific to their needs. The survey data should be collected in a structured format suitable for analysis, potentially using spreadsheets or survey tools. [no source]
* **Gathering Data from Various Sources:** The councils have access to various data sources:
  + Historical water usage records (Litres per day per household).
  + Data from water processing plants (volume processed, faults, outages, cost per litre).
  + Maps of existing water pipelines (location and capacity).
  + Central Statistics Office data with population information (density, age, rate of change).

**2. Processing Data:**

* **Cleaning Data:** The diverse data sources will likely have inconsistencies and errors. Data cleaning is essential to address issues like:
  + **Missing values:** For example, some households may have incomplete water usage records.
  + **Inconsistent formats:** Data from different sources might use varying units or conventions.
  + **Duplicate entries:** The survey might have duplicate responses that need to be removed. Duplicate entries can also occur during data integration, for example, if the same household is listed with a slightly different address in the water usage records and the CSO data.
* **Aligning Data:** This step involves transforming data into a common format for analysis.
  + Units need to be standardized: For example, ensuring all volume measurements are in liters.
  + Geographic references: Maps, water pipeline data, and household locations should be aligned using a consistent coordinate system.
* **Integrating Data:** This involves combining data from multiple sources into a unified dataset for analysis.
  + **Relational databases:** This approach can link data from different sources based on common identifiers, such as household IDs or geographic locations.
  + **Data Warehousing:** A data warehouse designed for analytical processing can be used to store the integrated data.

**3. Analysing Data:**

* **Data Exploration:** Tools like Jupyter notebooks, Tableau, or even spreadsheets can be used to initially explore the data, identify patterns, and formulate hypotheses. This exploration might involve:
  + Calculating descriptive statistics for water usage across different demographics and geographic areas.
  + Visualizing water usage trends over time and correlating them with population growth, seasonal changes, or events like droughts.
  + Analyzing survey responses to identify areas with strong positive or negative sentiment towards water charges. Word clouds or sentiment analysis tools could be useful here. [no source]
* **Statistical Analysis:**
  + Regression analysis: This can be used to model the relationship between water usage and factors like household size, age of residents, income, or property value. [no source]
  + Time series analysis: This can help forecast future water demand based on historical usage patterns and population projections. [no source]
* **Spatial Analysis:**
  + Geographic Information System (GIS): This software can be used to overlay water usage data, pipeline maps, and population density on a map to visualize areas with high demand, limited capacity, or potential growth. [no source]
  + Network analysis: This can be used to model the water distribution network, identifying bottlenecks, vulnerabilities, and areas requiring expansion or upgrades. [no source]
* **Sentiment Analysis:**
  + Machine learning algorithms can be applied to analyze survey responses and social media data to gauge public sentiment towards water charges and proposed infrastructure investments. [no source]

**4. Preserving Data:**

* **Database Management:** Choosing an appropriate database management system is crucial for long-term data storage, access, and integrity. Options include relational databases, NoSQL databases (document, column, graph), or a combination (polyglot persistence).
* **Data Backup and Recovery:** Regular backups should be implemented to prevent data loss in case of system failures or disasters.
* **Data Security:** Measures must be taken to protect sensitive personal data collected through surveys, adhering to data protection regulations like GDPR.

**5. Giving Access to Data:**

* **Data Visualization and Communication:** Tools like Python's Matplotlib, Bokeh, or Seaborn can be used to create visualizations (charts, graphs, maps) to communicate findings to stakeholders, including the public, council members, and policymakers. These visualizations should be clear, concise, and easy to understand, avoiding common pitfalls like misleading scales or inappropriate chart types.
* **Data Reporting:** Prepare comprehensive reports summarizing findings from the analysis, including recommendations for infrastructure investments.
* **Open Data Platforms:** Depending on data sensitivity and privacy considerations, the councils could consider publishing anonymized and aggregated data on open data platforms, encouraging transparency and enabling further research and analysis by external parties.

**6. Re-using Data:**

* **Continuous Monitoring and Evaluation:** The integrated dataset and analytical models can be used for ongoing monitoring of water usage patterns, evaluating the impact of water charges, and assessing the effectiveness of infrastructure investments. [no source]
* **Predictive Modeling:** Machine learning techniques can be applied to build models that predict future water demand, allowing for proactive planning and infrastructure development. [no source]
* **Scenario Planning:** The data can be used to simulate the impact of different policy decisions or infrastructure investments, supporting informed decision-making. [no source]
* **Data Archiving:** Ensure that data is archived appropriately for long-term preservation and potential future use.

This approach, grounded in the UK data archive Data Management Lifecycle, provides a structured framework for leveraging the available data to identify areas where future investment in water infrastructure is most needed. The specific methods and tools used at each stage can be tailored to the specific data characteristics and the councils' analytical goals.

This response incorporates information from the sources you provided, as well as general knowledge and best practices in data analysis and infrastructure planning. You may want to independently verify the information that is not explicitly sourced.

**Q2(b) Issues with Human-Created Metadata**

Metadata created by human users, while valuable, can present challenges. Here are four potential issues:

**1. Inconsistency:** Different individuals might use varying terms or descriptions for the same concept, leading to difficulty in searching and retrieving data.

●

**Example:** One person might describe a garment's color as "teal," while another uses "turquoise."1

**2. Subjectivity:** Human interpretation can introduce bias, making it difficult to compare or combine data from different sources.

●

**Example:** Describing the condition of a piece of clothing as "good" is subjective and might differ from another person's assessment.1

**3. Incompleteness:** Users might omit crucial information or provide insufficient detail, hindering the usefulness of the metadata.

●

**Example:** Failing to mention the brand of a clothing item can limit the search capabilities.1

**4. Lack of Standardization:** Without adherence to established standards or controlled vocabularies, metadata can become chaotic and difficult to manage, especially when dealing with large volumes of data.1

Using a metadata standard helps improve data quality by promoting consistency, reducing ambiguity, and enabling better data discovery and integration.1 However, enforcing a metadata standard can be challenging due to resistance to change, varying levels of user understanding, and the need for ongoing training and maintenance.

**Q 2(c) [4 Marks]**

Open datasets are made freely available for all people to access. Identify and

explain two (2) potential problems that may arise either in making data open or using open data.

## Potential Problems with Open Datasets

Open datasets, while beneficial, can present challenges in both their creation and utilization:

**Data Privacy and Protection:** Open datasets, even if intended for public use, can inadvertently contain sensitive personal information.

* + **GDPR regulations** highlight the importance of protecting personal data. Sensitive data like medical conditions, religious beliefs, and criminal activity requires explicit consent from the data subject for processing. Even seemingly innocuous data, when combined with other datasets, could potentially be used to identify individuals. For example, a dataset with location data and timestamps, if correlated with social media posts or other publicly available information, might reveal an individual's movements and activities.
  + **Data breaches** pose a significant risk. The Optus data breach in Australia in 2022 exposed email addresses, home addresses, dates of birth, and even passport or license numbers of millions of customers. This incident emphasizes the vulnerability of data, even for organizations with established data protection measures.
  + In the context of the Irish councils' water usage project, ensuring the privacy of individuals is paramount. Survey responses, linked to household water usage data, could potentially reveal sensitive information about individuals' habits and lifestyles. Anonymization techniques, such as aggregating data to a regional level rather than individual households, can help mitigate privacy risks.

**Data Quality and Reliability:** Open datasets may lack the rigorous quality control measures found in more controlled environments.

* + **Data cleaning** is often necessary to address issues like missing values, inconsistent formats, and duplicate entries. Manual data entry errors, typos, and artifacts introduced during data gathering can compromise data quality. For example, inconsistencies in units of measurement (centimeters vs. inches) or variations in the format of dates can introduce errors. In the case of the Irish councils, ensuring data quality is crucial for accurate analysis. For example, verifying the completeness and accuracy of historical water usage records and standardizing data formats from different processing plants would be essential.
  + **Metadata**, which describes and defines the content of data, plays a crucial role in ensuring findability and usability. However, **human-generated metadata can be problematic**. People may lie, be lazy, or simply make mistakes. Additionally, there may be semantic overlap between metadata elements, requiring judgment from the person assigning the metadata. These issues can lead to difficulties in searching, selecting, and interpreting open datasets. For the Irish councils, clear and accurate metadata would be essential for understanding and utilizing the various datasets, especially given their diverse sources.

Addressing these potential problems requires a multifaceted approach.

* For data privacy and protection, anonymization techniques, data encryption, and adherence to data protection regulations are crucial.
* For data quality and reliability, robust data cleaning processes, metadata standards, and community-driven validation efforts can help ensure the usability and trustworthiness of open datasets.

These considerations highlight the importance of responsible data management practices, both for those making data open and those utilizing it.

**Q 2(d) [3 Marks]**

HDFS enables Map-Reduce processing in Hadoop by replicating large files into

blocks across DataNodes, managed by NameNodes. How does HDFS prevent or limit data corruption errors?

### Data Replication and Checksums

HDFS employs several mechanisms to prevent and limit data corruption:

* **Replication:** HDFS replicates data blocks across multiple DataNodes. The default replication factor is three, meaning each block is stored on three different nodes. If one DataNode fails or experiences corruption, the data can be retrieved from the replicas on other nodes. This ensures data availability and redundancy.
* **Checksums:** HDFS uses checksums to verify data integrity. When a data block is written to a DataNode, a checksum is calculated and stored along with the block. When the block is read, the checksum is recalculated and compared to the stored checksum. If the checksums do not match, it indicates that the data block has been corrupted. [no source]
* **NameNode:** The NameNode is responsible for managing the file system metadata, including the location of data blocks and their replicas. It periodically communicates with DataNodes to ensure they are alive and healthy. If a DataNode fails, the NameNode will schedule the replication of the missing blocks to other DataNodes. This ensures data consistency and prevents data loss due to node failures.

These mechanisms work together to provide a highly reliable and fault-tolerant storage system for Hadoop. By replicating data and using checksums, HDFS minimizes the risk of data corruption and ensures that data remains accessible even in the event of hardware failures. This is essential for supporting MapReduce processing, which relies on the availability of data across the cluster.

This response incorporates information from the sources you provided, as well as general knowledge about HDFS. You may want to independently verify the information that is not explicitly sourced.

***[End of Question 2]***

CA682, CA682D – Data Management and Visualisation

Semester 1 Examinations 2019/2020 Page 3 of 7

***QUESTION 3 [TOTAL MARKS: 25]***

**Q 3(a) [8 Marks]**

(i) Give simple example metadata (3-4 attributes) describing your pen (ie, what you are using to write this exam paper).

(ii) For each example you’ve identified, is it Descriptive, Administrative or Structural metadata?

(iii) How could a standard be used for this type of metadata? (iv) Identify one (1) problem with enforcing a standard.

## Simple Metadata for a Pen

**(i) Example metadata:**

* **Brand:** BIC
* **Colour:** Blue
* **Type:** Ballpoint

**(ii) Type of metadata:**

* **Brand:** Descriptive (describes the pen)
* **Colour:** Descriptive (describes the pen)
* **Type:** Descriptive (describes the pen)

**(iii) Using a standard:**

A metadata standard could ensure consistency in describing pens. For example, a standard could define:

* Controlled vocabulary for "Brand" to ensure consistent naming (e.g., "BIC" not "BiC")
* Specific colour names or codes (e.g., "Pantone 293 C" for blue)
* Clear definitions for pen types (e.g., distinguishing "ballpoint" from "rollerball")

Standards like **Dublin Core**, while designed for broader information objects, could provide a framework. **Dublin Core** proposes a standard set of metadata elements, intended to be simple enough for a document's author to supply. However, adapting it for specific objects like pens would require additional specifications.

**(iv) Problem with enforcing a standard:**

* One problem is **ensuring adherence**. People may not understand the standard or see the benefit of using it. For example, someone might describe a pen as "dark blue" instead of using the standard colour code. This can lead to inconsistencies and difficulties in finding and comparing information about pens. The sources mention that enforcing a standard requires trade-offs between organization (adding, duplicate detection, storage) and retrieval (query, search).
* **People may find ways to work around the standard.** This is particularly true if the standard is perceived as too restrictive or if it does not adequately cover all possible variations. The sources do not provide specific examples of this, but it is a general challenge with enforcing any standard.

This response is primarily based on the sources you provided, particularly those related to metadata and standards like Dublin Core. Some of the specific examples provided are not directly from the sources and are based on general knowledge about pens and metadata. You may want to verify this information independently.

**Q 3(b) [6 Marks]**

Having successfully gathered data for local councils to understand water usage (Q1 or Q2), give two (2) examples of possible *data glitches* and explain how they might result in poor decision making.

## Data Glitches in the Water Usage Project

**Example 1:**

**Missing values in historical water usage records:** Some households might have missing data for certain periods, perhaps due to meter malfunctions or data entry errors during data gathering.

**Consequence:** If these missing values are not properly handled, they could lead to an underestimation of water usage in specific areas. This could result in inadequate investment in those areas, leading to water shortages or low water pressure during peak demand.

**Example 2:**

**Inconsistent units of measurement in data from water processing plants:** Some plants might report volume in cubic meters, while others report in liters.

**Consequence:** If these inconsistencies are not addressed during data integration and cleaning, they will lead to inaccurate calculations of total water processed and the cost per liter. This could result in poor decisions about the efficiency of different plants and where to allocate resources for upgrades or maintenance.

**Q 3(c) [6 Marks]**

Identify three (3) possible data errors in the sample view of a simple expenses table below. What methods for data cleaning would you recommend to clean this dataset?

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Jan** | **Feb** | **Apr** | **May** | **Jun** |
| **Jane** | €5 | €5 | €10 | €9 | €8 |
| **John** | 6.2 | €7 | €6 | €5 | €6 |
| **Sally** | €10 | €7 | €10 | €9 | €7 |
| **Simon** | €6 | €10 | €5 | €9 |  |
| **Total Expenses** | €27 | €29 | €21 | €32 | €21 |

## 

## Possible Data Errors and Data Cleaning Methods

### Possible Data Errors

* **Inconsistent currency symbol:** John's expenses for January are missing the Euro symbol (€). This inconsistency could cause problems in calculations or analysis if the software used to process the data is sensitive to formatting.
* **Missing data:** Simon's expenses for June are missing. This could be a data entry error or it could be that Simon did not have any expenses in June. Either way, it is important to identify and handle missing data appropriately.
* **Incorrect calculation for Total Expenses in January:** The total expenses for January are listed as €27. However, the sum of the individual expenses for January (€5 + 6.2 + €10 + €6) is **€27.2**. This could be a simple calculation error or it could indicate a more serious problem with the data entry process.

### Data Cleaning Methods

* **Data validation rules:** Implementing data validation rules in the data entry process can prevent many of these errors from occurring in the first place. For example, a validation rule could require that all currency values include the Euro symbol or that all cells for a particular month are filled in. This is a form of "implementing process mandates" to fix the human element of data entry problems.
* **Standardization:** Standardizing data formats can help prevent inconsistencies. In this case, ensuring that all currency values include the Euro symbol and using a consistent format for decimal values would improve data quality.
* **Missing data imputation:** There are various methods for handling missing data, such as replacing the missing value with the average value for that person or using a more sophisticated statistical model. The choice of method will depend on the nature of the data and the goals of the analysis.
* **Data exploration and verification:** Carefully examining the data for errors and inconsistencies is an essential step in data cleaning. Tools such as spreadsheets or purpose-built data cleaning tools like OpenRefine can assist in this process. Visualizing the data can also help to identify outliers or patterns that might indicate errors.
* **Documentation:** Documenting the data cleaning process is important for ensuring reproducibility and transparency. This documentation should include the steps taken to clean the data, the tools used, and any assumptions made.

These methods can be applied using a variety of tools, including:

* **Spreadsheets:** Spreadsheets like Google Sheets and Microsoft Excel provide basic functionality for data cleaning, such as sorting, filtering, and applying formulas.
* **Purpose-built tools:** Tools like OpenRefine, RapidMiner, and Tableau Prep offer more advanced features for data cleaning, such as data profiling, transformation, and validation.
* **General Purpose Languages:** Programming languages like Python and R provide a high degree of flexibility and control over the data cleaning process.

It's important to note that the choice of data cleaning methods will depend on the specific dataset and the goals of the analysis. A combination of different techniques is often necessary to ensure data quality.

The sources provide general guidance on data cleaning, but they don't address this specific example directly. The suggestions made here are based on common data cleaning practices and general knowledge about data quality. It would be a good idea to explore these methods in more detail to determine the most appropriate approach for this specific dataset.

**Q 3(d) [5 Marks]**

Give an example of data that could be considered sensitive or personal data under GDPR regulations. You are working for DCU, someone breaks into your office and steals a laptop with this data stored on it! What actions should you take?

## Stolen Laptop Containing Sensitive Personal Data

**Example of Sensitive Data Under GDPR:** A spreadsheet containing the names, student numbers, home addresses, dates of birth, and exam results of students enrolled in a particular course at DCU would be considered sensitive personal data under GDPR regulations. This is because this information can directly or indirectly identify an individual and is not data of public domain. Additionally, under GDPR, this type of data requires **unambiguous consent** from the data subject (the students) to be processed.

**Actions to Take After Laptop Theft:** If someone breaks into your office and steals a laptop containing this data, you should take the following actions:

1. **Immediately report the incident to the DCU Data Protection Officer (DPO).** GDPR regulations stipulate that any unauthorised access to personal data, such as through theft, is considered a data breach. The DPO is responsible for managing data breaches and ensuring compliance with GDPR.
2. **Document the details of the incident.** This should include the date and time of the theft, a description of the stolen laptop, the data stored on the laptop, and any other relevant information. This documentation will be crucial for any subsequent investigation or reporting requirements.
3. **Assess the potential risks to the individuals whose data was stored on the laptop.** Consider the sensitivity of the data and the potential harm that could result from its unauthorized disclosure. For instance, if the data includes financial information or health records, the risks are significantly higher.
4. **Take steps to mitigate the risks.** This may involve:
   1. **Notifying the affected individuals.** Depending on the severity of the breach and the potential risks, it may be necessary to inform the students whose data was stolen.
   2. **Providing credit monitoring or identity theft protection services.**
   3. **Implementing additional security measures to prevent future breaches.**
5. **Cooperate with any investigations by the DPO or relevant authorities.** Be prepared to provide information and documentation related to the incident.
6. **Review DCU's data protection policies and procedures.** Identify any weaknesses that may have contributed to the breach and take steps to improve security measures. Consider implementing strong passwords, encryption, and regular data backups to minimize risks in the future.

The GDPR's purpose is to protect the misuse of sensitive and personal data. It's essential to follow the regulations and take appropriate action in case of a data breach to minimize potential harm to individuals and maintain trust.

***[End of Question 3]***

CA682, CA682D – Data Management and Visualisation

Semester 1 Examinations 2019/2020 Page 4 of 7

***QUESTION 4 [TOTAL MARKS: 25]***

**Q 4(a) [12 Marks]**

In the appendix, Figure 1 shows a graphic.

1. What visual communication goals are evident? [2 marks]

Visual communication goals can be categorised into four types: **Information, Persuasion, Education and Entertainment**. **Information** goals involve structured data, where structure is key. The level of detail (macro vs. micro), layout, and colour contribute to the credibility of the information presented. It is important to ensure the source data is of high quality. **Persuasion** goals are focused on eliciting a specific response from the receiver. For instance, advertising utilises information to convey a particular message. **Education** goals involve the transfer of knowledge and skills. **Entertainment** aims to provide pleasure, diversion, and amusement.

The sources provide examples of each goal:

* **Information:** A graphic illustrating the efficacy of various snake oil supplements exemplifies the goal of providing information.
* **Persuasion:** A vintage Guinness advertisement demonstrates the goal of persuasion by associating the beverage with strength.
* **Education:** Textbooks, online learning resources, brochures, and movies are typical examples of educational visual communication.
* **Entertainment:** Artistic expressions, video games, film, television, and ebooks serve the purpose of entertainment.

Furthermore, a good visualisation design process considers various aspects such as the **story (concept), goal (function), information (data), and visual form (metaphor)**. The effectiveness of the visualisation depends on how well these elements are integrated. For instance, a successful visualisation could involve a research document, data, and outline for information; a proof of concept or prototype for the goal; a schematic or wireframe for the story, and sketches or artwork for the visual form. The ultimate aim is to achieve a successful visualisation that is neither boring nor useless.

1. Identify two (2) design principles and explain how the graphic applies them to fulfil the communication goals. [4 marks]

The sources mention several design principles, but don't explicitly apply them to a specific graphic. Therefore, without a provided graphic, a complete analysis cannot be given. Here's a summary of design principles from the sources and how they can be applied in general:

### Design Principles for Visual Communication

* **Structure:** A clear structure is crucial for presenting information effectively. This involves organizing data logically, using visual hierarchies (headings, subheadings, bullet points, etc.), and ensuring a clear flow of information.
  + **Example:** A diagram explaining a complex process might use a flow chart format with clear visual cues to guide the viewer's understanding.
* **Level of Detail:** Balancing macro and micro levels of detail ensures viewers can grasp the overall message while also having access to specific data points.
  + **Example:** A graph might use a zoomed-out view to show overall trends, but also offer interactive elements to allow users to zoom in and explore individual data points.
* **Layout:** The arrangement of elements (text, images, charts) greatly impacts readability and comprehension. Consider visual weight, white space, and alignment.
  + **Example:** Using a grid system can create a visually appealing and organized layout that makes it easy for viewers to follow the information presented.
* **Colour:** Colour choices affect readability, aesthetics, and can evoke emotions. Consider colour palettes, contrast, and cultural associations.
  + **Example:** Using a contrasting colour for important data points can draw attention to key information, while maintaining a consistent colour scheme can ensure a cohesive and professional look.
* **Credibility:** Design choices influence how trustworthy the information appears. Using reputable sources, clear labeling, and avoiding unnecessary embellishments enhances credibility.
  + **Example:** A data visualization might include citations to sources or provide a clear methodology to demonstrate transparency and increase trust in the presented information.
* **Research Audience:** Understanding the target audience's demographics, knowledge level, and interests helps tailor the design for maximum impact.
  + **Example:** A visualisation for experts in a specific field might use technical terminology, while one for a general audience would need simpler explanations.
* **Illustrations, Themes, Grouping:** Visual elements should attract the viewer's attention and guide them through the information.
  + **Example:** Using relevant icons or images can make data more engaging, and grouping related data points can highlight patterns and insights.
* **Divide Information Into Chunks:** Breaking down information into manageable sections makes it easier to understand and process.
  + **Example:** A complex topic might be presented in a series of slides or chapters, each focusing on a specific aspect.
* **Legibility:** The information must be easy to read. This includes choosing appropriate fonts, font sizes, and ensuring sufficient contrast between text and background.
  + **Example:** Using a sans-serif font in a large enough size for body text can improve readability.
* **Progressive Disclosure:** Gradually revealing information can prevent cognitive overload and help viewers absorb information step by step.
  + **Example:** An interactive data visualization might start with a basic overview and then allow users to explore more details as they interact with the graphic.
* **Focus On Narrative:** For entertainment purposes, a strong narrative can engage the viewer and create an immersive experience.
  + **Example:** A movie trailer might use editing techniques and music to build suspense and draw the audience into the story.
* **Lighting, Layout, Multimodality:** The construction of visual content influences how it is perceived.
  + **Example:** In a film, the use of lighting, camera angles, and sound design all contribute to creating a specific mood and atmosphere.
* **Style:** Developing a consistent and recognisable visual style can enhance brand identity and make content more memorable.
  + **Example:** A company might use a consistent color palette and font in all its marketing materials to create a unified brand identity.

To apply these principles to a specific graphic, you would need to consider:

1. **What are the communication goals of the graphic?** (e.g., inform, persuade, educate, entertain)
2. **How does the graphic use visual elements to achieve those goals?**
3. **How effective are the design choices?**

Once you provide the image, it will be possible to apply these principles and answer your query comprehensively.

1. The figure has been converted to greyscale. What colours would you recommend to use to highlight important points and why? [4 marks]

This response draws on information from the sources and previous turns in the conversation to recommend colours and explain their use in highlighting information.

### Colour Recommendations to Highlight Information

The sources emphasise the importance of using colour effectively in data visualisations to guide the viewer's attention and enhance comprehension. Colour choices should be guided by the principles of trustworthiness, accessibility, and elegance.

**Recommended Colours:**

* **Red:** Red is often associated with importance, urgency, or warning. Using red sparingly can effectively highlight critical data points or areas that require immediate attention.
* **Green:** Green is commonly associated with positive connotations, such as growth, success, or correctness. It can be used to highlight positive trends or favorable outcomes.
* **Blue:** Blue often represents calmness, trust, and stability. It is a versatile colour that can be used for various purposes, such as highlighting background information or emphasizing data points without creating a sense of urgency.

**Considerations for Colour Choice:**

* **Trustworthiness:** Avoid using colour palettes that are misleading or distort the data.
* **Accessibility:** Ensure that colour choices are accessible to individuals with colour vision deficiencies. Use sufficient contrast between colours and consider providing alternative ways to represent information, such as patterns or labels.
* **Elegance:** Aim for a visually appealing and cohesive colour scheme. Avoid using too many colours, as this can create a cluttered and distracting visual.
* **Cultural Significance:** Be mindful of the cultural significance of colours. Red and green have particular meanings in some cultures. If your audience is diverse, it's essential to research colour associations to avoid misinterpretations. This information is not from the sources.

**Example:** In a bar chart showing sales figures, you could use red to highlight the highest sales month, green to indicate months where sales exceeded targets, and blue for the remaining months.

By using colour strategically, you can guide your audience's attention to the most important information and create a more engaging and informative visual presentation.

1. Identify two (2) attributes that the graphic uses to encode data. [2 marks]

Attributes

Now that we have the raw material to encode our data -- points, lines and area (and

occasionally forms), let’s look at what attributes we can change to capture the data values.

These can be divided into three classes: quantitative, categorical and relational. You should

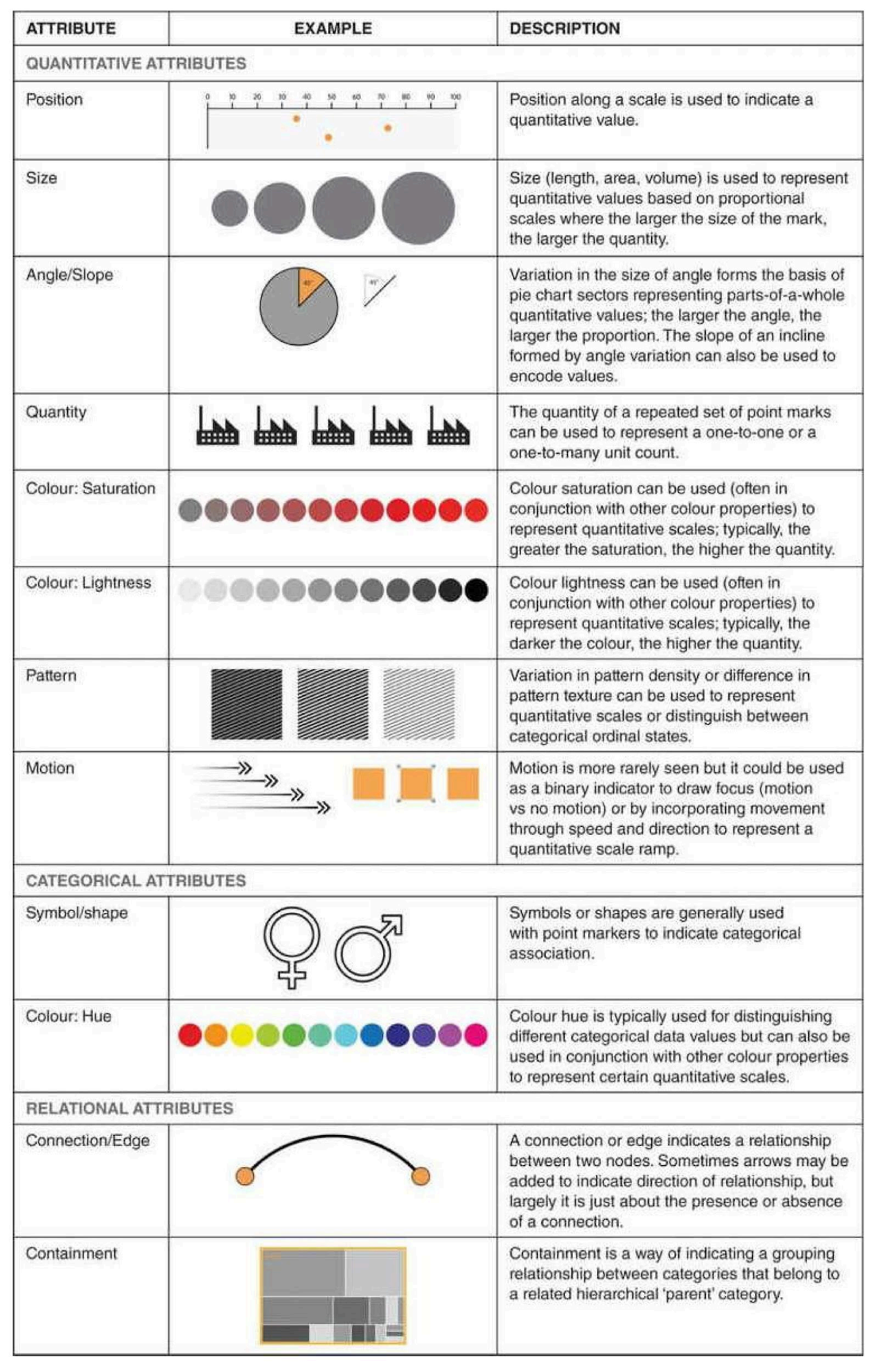
recognise these names from looking at data types. There is overlap between the attributes

used in these different classes and it is possible to encode a categorical value using

something like position or size.

The table below (from Kirk, 2016, p161) shows the attributes for each class with an example

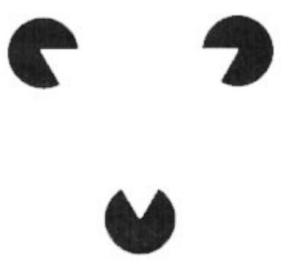
of how it might appear and a description.

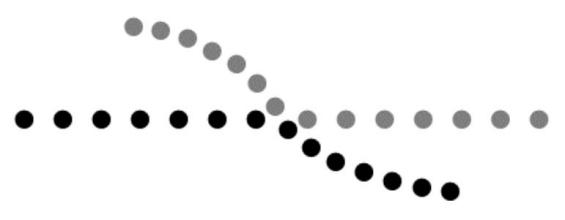


**Q 4(b) [6 Marks]**

Identify the gestalt principles of visualisation present in each of the 3 images ([A], [B] & [C]) below.



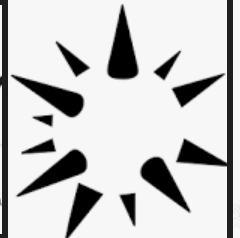




[B] [C]

[A]

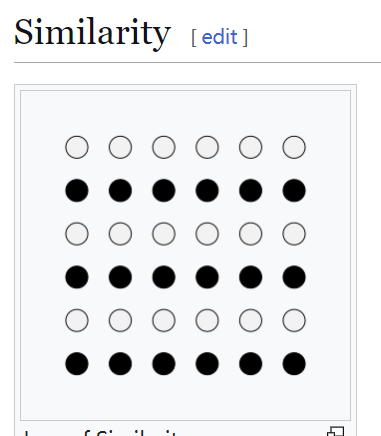
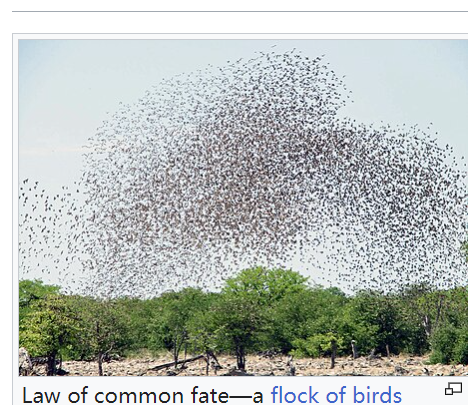
Here is a breakdown of the Gestalt principles present in the provided images:

* **A Closure**: We tend to perceive incomplete shapes as complete by filling in the missing gaps. This principle can be applied to design logos or icons where a complete shape is suggested through strategically placed elements.
* 
* **[B] Symmetry**: This principle posits that symmetrical elements are perceived as belonging together, forming a coherent shape. In Image [B], the two curved shapes, despite being separate entities, are perceived as parts of a single, symmetrical object (a goblet or vase).
* **[C] Continuity**: The **principle of continuity** suggests that we tend to perceive elements arranged along a line or curve as belonging together, even when there are breaks in the line. The dots in Image [C] are perceived as forming a single, continuous line due to their arrangement.

**Note**: The specific images referred to as [A], [B], and [C] are not provided in the text. Therefore, the analysis relies on a general understanding of these Gestalt principles. To provide a more precise response, please share the images.

The Gestalt Principles of Visualisation

The sources provide a list of Gestalt principles. These principles, stemming from Gestalt psychology, describe how humans perceive and group visual elements.

* **Proximity**: Elements close to each other are perceived as a group. For example, placing related data points close together in a graph helps viewers understand their connection.
* **Similarity**: We group elements with similar visual attributes, like shape, color, or size. Imagine a scatterplot where different data groups are represented by different colored dots.
* 
* **Continuity**: We perceive elements along a line or curve as connected, even if there are breaks. For instance, a line graph effectively uses continuity to show trends over time.
* **Symmetry**: Symmetrical elements are seen as a unified whole. A visualisation might employ symmetry to create a balanced and harmonious design.
* **Closure**: We tend to perceive incomplete shapes as complete by filling in the missing gaps. This principle can be applied to design logos or icons where a complete shape is suggested through strategically placed elements.
* **Figure and Ground**: We distinguish between the object of focus (figure) and its background (ground). A clear separation of figure and ground in a visualisation helps ensure the data stands out and is easily comprehensible.
* 

These principles guide designers in creating effective visualisations that are easily understood by viewers.

**Q 4(c) [4 Marks]**

In some visualisations, size of objects represents a quantity. If using a 2d shape

such as a circle to represent a quantity, what does the designer need to be careful of? You can draw a sketch to illustrate the problem.

Using Area to Represent Quantity

When using circles to represent quantities in a visualisation, designers need to be careful about how the **area** of the circle relates to the quantity. It's easy to mistakenly use the **diameter** or **radius** of the circle to represent the quantity, which can lead to misleading visualisations.

**The problem:** Humans naturally perceive the size of a circle based on its diameter or radius, not its area. If the diameter of a circle is doubled, its area actually increases by a factor of four. This can result in the larger circle appearing much larger than it should be relative to the smaller circle, creating a distorted representation of the data.

**Example**:

Imagine two circles representing sales figures for different products.

* Product A has sales of €100.
* Product B has sales of €200.

If the designer uses the diameter of the circle to represent the sales, the circle for Product B will have twice the diameter of Product A. However, this will make the circle for Product B look **four times larger**, not twice as large, which misrepresents the actual difference in sales.

**Solution**: To accurately represent quantities using circles, the designer should ensure that the **area** of the circle is proportional to the quantity.

* For instance, if Product A's sales are represented by a circle with an area of 10 square units, then Product B's sales should be represented by a circle with an area of 20 square units. This ensures that the visual representation accurately reflects the 2:1 ratio in sales between the two products.

**Sketches:**

The sources do not contain any sketches. Here are some sketches illustrating the problem:

**Incorrect**: Circles with diameters proportional to the quantities:

**Correct**: Circles with areas proportional to the quantities:

By being mindful of the relationship between area and perceived size, designers can create more accurate and effective visualisations using circles.

**Q 4(d) [3 Marks]**

Which is of greater importance in a visualisation - luminance (brightness/contrast) or colour (hue)? Justify your answer.

## Luminance versus Colour in Visualisations

**Luminance (brightness/contrast) is generally of greater importance in a visualisation than colour (hue)**. This is because luminance plays a crucial role in our ability to perceive shapes, patterns, and depth, which are fundamental to understanding any visual representation.

* **Luminance and Human Vision:** The human visual system is highly sensitive to differences in luminance. Our eyes have specialized cells called rods and cones that detect light and colour. Rods are primarily responsible for vision in low-light conditions and are more sensitive to luminance changes, while cones are responsible for colour vision and work best in brighter light.
* **Pre-Attentive Processing:** Luminance contrast is a key element in pre-attentive processing – the rapid, subconscious processing of visual information that helps us quickly identify important elements in a scene. High luminance contrast makes elements stand out, grabbing our attention before we even consciously process the image. For instance, a bright red data point on a dark background will instantly draw the eye.
* **Accessibility and Clarity:** Visualisations with strong luminance contrast are generally more accessible to viewers with colour vision deficiencies. They also tend to be clearer and easier to understand for all viewers, as the information is encoded in a way that our visual system is naturally attuned to.

While colour can be a valuable tool in visualisations, it should be used strategically and with careful consideration.

* **Colour's Role:** Colour is most effective for encoding categorical data, differentiating distinct groups or categories. For example, different colours can represent different countries on a map or different product categories in a sales chart.
* **Colour Challenges:** Overusing colour can create visual clutter, making it difficult to focus on the important information. Additionally, cultural associations with colours can vary widely, making it important to choose colours carefully to avoid unintended meanings.

In summary, luminance is the foundation of visual perception. It should be prioritised to ensure clarity and accessibility in visualisations. Colour should be used sparingly and purposefully to enhance the visualisation without overwhelming the message.

***[End of Question 4]***

CA682, CA682D – Data Management and Visualisation

Semester 1 Examinations 2019/2020 Page 5 of 7

***QUESTION 5 [TOTAL MARKS: 25]***

**Q 5(a) [12 Marks]**

(i) Identify two (2) specific problems with the design of the graph below. [2 marks]

(ii) Using the data from the graph, **sketch** an alternative graph for displaying the information to highlight why CA682 (Data Management and

Visualisation) is the most important module in the Data Analytics course. If you want to use colour then you can write in the colour name. [6 marks]

(iii) Label (ie, identify) on your sketch the following graph components: x-axis, y-axis, title and the marks used. [4 marks]

|  |
| --- |
|  |

**Q 5(b) [8 Marks]**

Given the following visualisation tasks, suggest an appropriate graph type (specific chart type not just a category) for each to display the information and give a brief

justification.

[A] Understand the relationship between maximum daily temperature (⁰C) and average daily personal water consumption (Litres) in Ireland.

[B] Show the improvement in sales (total profit in €) over the past 5 years for your product compared to your competitors.

[C] The most popular method of travel to DCU during 2019. [D] Distribution of grades in CA682 over the past 5 years.

**Q 5(c) [5 Marks]**

Describe the four stages of understanding that happen when you look at a graphic or

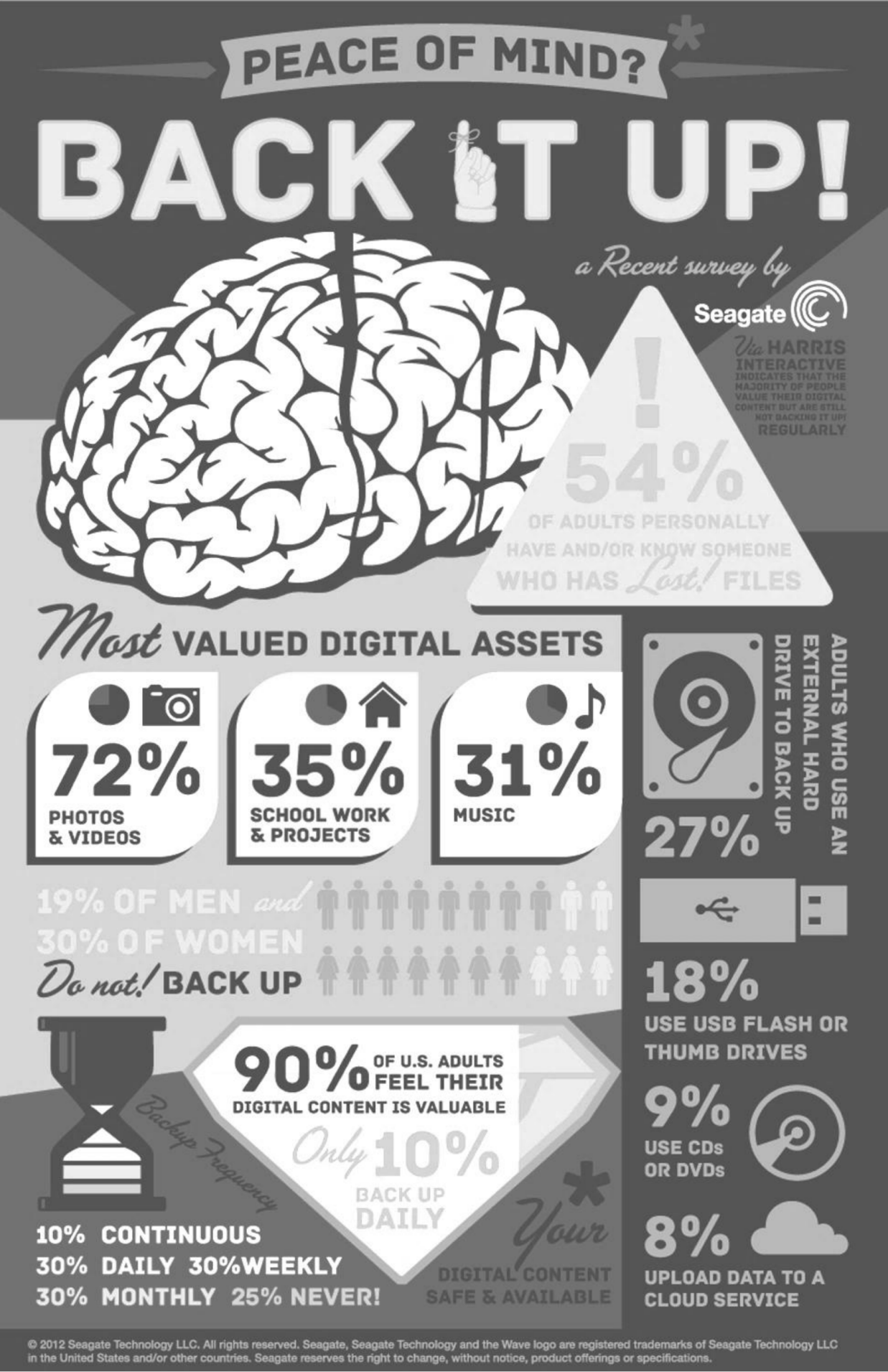
chart. Why does this mean that 3D effects in 2D graphs make understanding difficult?

***[End of Question 5]***

CA682, CA682D – Data Management and Visualisation

Semester 1 Examinations 2019/2020 Page 6 of 7

***APPENDICES***



***[END OF APPENDICES] [END OF EXAM]***

CA682, CA682D – Data Management and Visualisation

Semester 1 Examinations 2019/2020 Page 7 of 7