SAMPLE ANSWERS & NOTES

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

Q 1(a) **[7 Marks]**

Given the following brief to design a system for a data collection task, list three (3) important questions you would ask your client and suggest a type of database system to use, giving a reason for your choice.

> "We are collecting data to use for a marketing campaign by DCU to increase public transport use when travelling to campus. Data sources include public surveys, records from Transport Ireland and information from DCU estates."

Q 1(b) **[7 Marks]**

(i) Why is it useful to categorise data?
(ii) You have data from the last 5 track and field meets run by Athletics Ireland. Identify some different category descriptions for the following pieces of data:

A. Athlete profile
B. List of races
C. Gold, Silver and Bronze winners
D. Times of the winners from the sprint races

Q 1(c) **[6 Marks]**

Give two (2) advantages & two (2) disadvantages of non-relational databases and give an example of a when a non-relational database would be useful.

Q 1(d) **[5 Marks]**

Given a generic data analytics pipeline – Gathering, Processing, Analysing, Presenting, Preserving – describe in 1 or 2 sentences the activities that can occur at each stage.

Q 1(a): Designing a System for Data Collection

To design a system for the data collection task described in the brief, you would want to ask your client some key questions. Here are three important questions along with a suggested type of database system and the reason for the choice:

1. Question: What specific data points are critical for the marketing campaign?

- Suggested Database System: Relational Database
- Reason: Relational databases are well-suited for structured data, and since the data sources include surveys, records, and information, a relational database can efficiently organize and link these structured datasets.

2. Question: How frequently will the data be updated, and is real-time access necessary?

- Suggested Database System: NoSQL Database (e.g., MongoDB)

- Reason: NoSQL databases are suitable for handling large volumes of unstructured data and provide flexibility for frequent updates. If real-time access is crucial, a NoSQL solution might be preferable.

3. Question: What are the security and privacy requirements for handling this data?

- Suggested Database System: Relational Database with strong security features
- Reason: Relational databases often have robust security features and can be configured to enforce data integrity and access controls, crucial for handling sensitive information.

Q 1(b): Categorizing Data

(i) Why is it useful to categorize data?

- Categorizing data is useful for organizing information, improving searchability, and facilitating analysis. It helps in creating a structured framework for better understanding and utilization of the data.

(ii) Category Descriptions for the Data:

- A. Athlete Profile: Individual Athlete Information
- B. List of Races: Event Information
- C. Gold, Silver, and Bronze Winners: Medalist Information
- D. Times of the Winners from Sprint Races: Performance Metrics

Q 1(c): Non-relational Databases

- **Advantages:**

1. Scalability: Non-relational databases can easily scale horizontally to handle large amounts of data.
2. Flexibility: They are suitable for handling unstructured or semi-structured data.

- **Disadvantages:**

1. Consistency: May lack the strong consistency guarantees of relational databases.
2. Complexity: Non-relational databases can be more complex to query and manage.

- Example Scenario for Non-relational Database:
- When dealing with a large-scale, dynamic system like a social media platform where user-generated content varies widely and needs to be rapidly scaled.

Q 1(d): Data Analytics Pipeline

- Gathering: Collecting raw data from various sources.
- Processing: Cleaning and transforming the raw data into a usable format.
- Analyzing: Extracting insights and patterns from the processed data.
- Presenting: Visualizing the analyzed data for interpretation.
- Preserving: Storing the results and maintaining the data for future reference or analysis.

I hope this helps! Good luck with your exam. If you have any further questions or need clarification on any point, feel free to ask.

2.

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

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

(i) What is metadata?
(ii) Name three (3) things that metadata is used for.
(iii) Give examples of different types of metadata.

Q 2(b) **[7 Marks]**

Give simple example metadata describing the module CA682. Why would a standard be useful for this type of metadata and identify one (1) problem with enforcing a standard.

Q 2(c) **[6 Marks]**

Big data is characterised by its Volume, Velocity, Variety and recently Veracity. Veracity refers to how reliable the data is.
(i) Why does big data have a particular issue with data uncertainty?
(ii) Give one (1) method for checking the veracity of a big data source.

Q 2(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.

Q 2(a): Metadata

(i) What is metadata?

Answer: Metadata is data that provides information about other data. It offers context, meaning, and structure to help understand, manage, and use the actual data.

(ii) Name three (3) things that metadata is used for.

Answer:

Data Description: Describing the content, format, and structure of the data.

Data Administration: Managing and organizing data, including access controls.

Data Discovery: Facilitating the search and retrieval of specific data.

(iii) Examples of different types of metadata.

Answer:

Descriptive Metadata: Title, author, and keywords.

Structural Metadata: Relationships between data elements.

Administrative Metadata: Access rights, data ownership, and creation date.

Q 2(b): Metadata for Module CA682

Example Metadata for CA682:
Module Code: CA682
Module Title: Database Management Systems
Instructor: Prof. Smith
Credit Hours: 5
Prerequisites: CA681
Why a Standard is Useful:

A standard ensures consistency and interoperability, making it easier for systems to understand and process the metadata consistently. For example, a standard could define a specific format for dates or a standardized way of expressing prerequisites.

Identify One Problem with Enforcing a Standard:

One problem is that enforcing a standard can be challenging when dealing with legacy systems or when there is a need for flexibility due to evolving requirements. Strict adherence might hinder adaptation to new circumstances.

Q 2(c): Big Data and Veracity

(i) Why does big data have a particular issue with data uncertainty?

Answer: Big data often incorporates data from diverse and unstructured sources, leading to uncertainty in terms of accuracy, reliability, and consistency.

(ii) Method for Checking Veracity of a Big Data Source:

Answer: Cross-referencing with multiple sources or employing data validation algorithms can help verify the veracity of big data. Comparing information from different sources can enhance confidence in the reliability of the data.

Q 2(d): Scraping Data from a Website

Process of Scraping Data from a Website:

Identify Target Data: Determine the specific data on the website to be extracted.

Use Web Scraping Tools: Employ tools or libraries (e.g., BeautifulSoup, Scrapy in Python) to extract data from HTML.

Parse and Clean Data: Process the extracted data, removing irrelevant information and ensuring data quality.

Store or Analyze Data: Save the scraped data for analysis or use it directly in the desired application.

Two Rules to Remember When Using Web Scraping as a Data Source:

Respect Terms of Service: Ensure compliance with the website's terms of service to avoid legal issues.

Rate Limiting: Implement rate limiting to avoid overwhelming the website's servers and to be considerate of their resources.

3.

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

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

You are collecting data to for a marketing campaign to increase public transport use to travel to DCU. You have the following data sources:

A. Survey of current transport patterns of staff and students.
B. Access logs from Transport Ireland app filtered by GPS location for the Glasnevin Campus.
C. Map of transport options for the DCU campuses.
D. Medical research data showing improved heart health from regular walking based on user's wearing fitbit sensors.

For **each** of these sources, identify one (1) possible cause and consequence of poor quality data.

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

Pick one of the data sources listed in Q3(a).

(i) Give an example of an approach to cleaning data that you could use.
(ii) Give an example of how you could enforce better data quality.

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

(i) What are **constraints** with respect to Data Quality? Ensure you define and distinguish between static and dynamic constraints.

(ii) What are two (2) potential problems with enforcing constraints to improve data quality?

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

Open datasets are made freely available for all people to access. Identify and explain two (2) potential problems that may arise in making data open or using open data.

Q 3(a): Data Sources - Causes and Consequences of Poor Quality Data

Survey of current transport patterns of staff and students (Source A):

Possible Cause of Poor Quality Data: Respondents may provide inaccurate information due to recall bias or intentional misreporting.

Consequence: Misleading insights and ineffective campaign strategies based on unreliable data.

Access logs from Transport Ireland app (Source B):

Possible Cause of Poor Quality Data: Inaccurate GPS signals or technical glitches in the app could result in incorrect location data.

Consequence: Incorrect understanding of popular transport routes, leading to misguided efforts to improve certain routes.

Map of transport options for the DCU campuses (Source C):

Possible Cause of Poor Quality Data: Outdated or incomplete information about available transport options.

Consequence: Users may not be aware of new transport options, leading to an incomplete understanding of available choices.

Medical research data showing improved heart health from regular walking (Source D):

Possible Cause of Poor Quality Data: Users may not consistently wear or use Fitbit sensors, leading to incomplete or biased health data.

Consequence: Overestimation of the health benefits of regular walking, potentially leading to misguided marketing messages.

Q 3(b): Cleaning and Enforcing Data Quality for Chosen Data Source

Chosen Data Source: Access logs from Transport Ireland app (Source B)

(i) Example of an Approach to Cleaning Data:

Implement outlier detection algorithms to identify and remove anomalous GPS coordinates that might result from technical errors or deliberate manipulation.

(ii) Example of Enforcing Better Data Quality:

Implement data validation rules within the app to check the plausibility of GPS coordinates and prompt users for confirmation if data seems inconsistent. Additionally, regular updates to the app can address technical issues affecting data quality.

Q 3(c): Constraints and Problems with Enforcing Constraints

(i) Constraints with Respect to Data Quality:

Static Constraints: Fundamental rules defined during database design (e.g., data types, uniqueness constraints).

Dynamic Constraints: Rules that evolve with time and may be context-dependent (e.g., allowable range of values for a changing parameter).

(ii) Potential Problems with Enforcing Constraints:

Rigidity: Static constraints can become a problem when data requirements change, leading to the need for frequent modifications to the database structure.

Performance Impact: Enforcing constraints, especially dynamic ones, may introduce computational overhead, impacting system performance.

Q 3(d): Problems with Open Datasets

Potential Problem 1: Lack of Control over Data Quality:

Explanation: Open datasets may lack the quality control mechanisms present in proprietary datasets, leading to inconsistencies, errors, and biases that can affect downstream analyses.

Potential Problem 2: Privacy and Security Concerns:

Explanation: Open datasets, especially when dealing with personal or sensitive information, can pose privacy risks if not adequately anonymized. Security concerns may arise if the data contains vulnerabilities that could be exploited.

4.

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

Q 4(a) **[8 Marks]**

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

A. Summary of voter choices in the Irish Election 2016
B. Annual income for Computer Science students grouped by university
C. Population trend for Ireland over the last decade
D. Average rental prices for each Irish County in 2016

Q 4(b) **[7 Marks]**

In the appendix, Figure 1 shows a graph. Identify three (3) problems with the design and suggest a better method for showing the information, giving a specific chart type that could be used. You do not need to view the appendix in colour.

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

Explain what D3.js is and give 2 examples of things it doesn't do. Explain the difference between rules and selectors (CSS) and how they are used in D3.js.

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

How can design rules help to make better data visualisations? Give an example.

Q 4(a): Visualization Tasks and Appropriate Graph Types

A. Summary of voter choices in the Irish Election 2016:

- **Appropriate Graph Type: Stacked Bar Chart**

Justification: A stacked bar chart can effectively show the distribution of votes for each party in the election, providing a clear visual representation of the total votes and the contribution of each party.

B. Annual income for Computer Science students grouped by university:

- **Appropriate Graph Type: Grouped Bar Chart**

Justification: A grouped bar chart allows for a comparison of annual incomes across different universities for Computer Science students, making it easy to identify income differences between universities.

C. Population trend for Ireland over the last decade:

- **Appropriate Graph Type: Line Chart**

Justification: A line chart is suitable for displaying trends over time. It can effectively show the population changes in Ireland over the last decade, allowing for a clear understanding of the trend.

D. Average rental prices for each Irish County in 2016:

- **Appropriate Graph Type: Choropleth Map**

Justification: A choropleth map can provide a geographic representation of average rental prices in each Irish County in 2016, allowing for easy identification of regional variations.

Q 4(b): Figure 1 in the Appendix - Problems and Suggestions

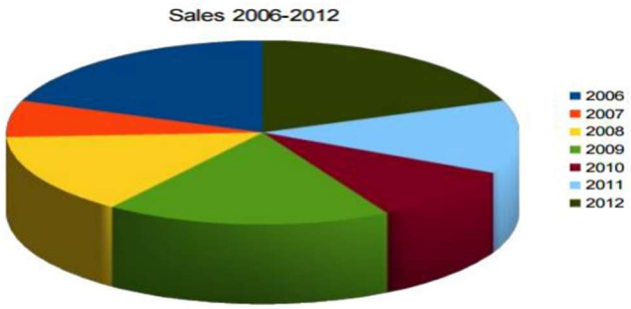


Figure 1 – a 3d pie-chart

Problems with Design:

- **Inadequate Axis Labels:** The graph lacks clear and descriptive axis labels, making it challenging to interpret the data.
- **Unclear Legend:** The legend is not well-placed or clearly labeled, causing confusion about the meaning of different colors.
- **Poor Color Choice:** If color is used, it may not be accessible for individuals with color vision deficiencies.

Better Method and Specific Chart Type:

- **Suggested Chart Type:** Grouped Bar Chart with Descriptive Axis Labels and a Clear Legend

Q 4(c): D3.js - Explanation and Differences

D3.js Explanation:

- D3.js (Data-Driven Documents): A JavaScript library for creating dynamic, interactive data visualizations in the web browser. It binds data to the Document Object Model (DOM) and applies data-driven transformations to the document.

Two Things D3.js Doesn't Do:

- High-Level Chart Abstractions: D3.js is low-level, allowing great flexibility but requiring more effort for creating high-level chart abstractions.

Built-in Interactivity: While D3.js provides tools for building interactive visualizations, it doesn't offer pre-built interactive components like tooltips or zoom controls.

Difference Between Rules and Selectors in CSS and Their Use in D3.js:

- Rules: Style definitions applied to selected elements.
- Selectors: Patterns that match sets of elements.
- Use in D3.js: D3.js uses CSS selectors to target and modify elements in the DOM, and rules define the styles applied to those elements.

Q 4(d): Design Rules for Better Data Visualizations

How Design Rules Help:

Explanation: Design rules provide guidelines for creating effective and meaningful data visualizations, ensuring clarity, accuracy, and interpretability. For example, using appropriate color schemes, emphasizing key data points, and providing clear labels can enhance the overall understanding of the information being presented.

Example:

Use of Consistent Color Palette: By following a consistent color palette for different categories in a chart, viewers can easily associate colors with specific data elements, improving comprehension and reducing potential confusion.

5.

Q 5(c) **[6 Marks]**

In visualisation it is important to understand pre-attentive processing. Explain what a pre-attentive feature is and describe an experiment to determine if a feature is pre-attentive or not.

Q 5(d) **[6 Marks]**

Correctly match the following depth cues for human vision.

Occlusion	A. moving the head slightly to create differences in the sensed images
Convergence	B. difference in direction of our eyes when looking at closer objects
Accommodation	C. images sensed by our two eyes are slightly different and this difference is used to determine depth
Aerial Haze	D. blocking of more distant objects by closer objects
Binocular Disparity	E. objects on the far horizon look hazy due to particles in the air
Motion Parallax	F. muscle tension from re-focussing the eye

Which cue is (mostly) used to create the 3D effect in movies?

Q 5(c): Pre-Attentive Processing

Explanation of Pre-Attentive Feature:

A pre-attentive feature is a visual attribute that is rapidly and effortlessly processed by the human visual system without the need for focused attention. These features can be quickly perceived and recognized, facilitating the initial processing of visual information.

Experiment to Determine if a Feature is Pre-Attentive:

One experiment to determine if a feature is pre-attentive involves the use of a visual search task. In this task, participants are presented with a set of visual stimuli, and their task is to quickly identify a target with a specific feature among distractors.

Pre-Attentive Feature Test:

If participants can easily and quickly identify the target based on a specific feature (e.g., color, orientation, size) without focused attention, it suggests that the feature is pre-attentive.

Control Condition:

To establish a baseline, a control condition with distractors that do not share the specific feature can be included. If the search time is significantly faster in the presence of the distinctive feature, it supports the pre-attentive nature of that feature.

Q 5(d): Matching Depth Cues for Human Vision

- Occlusion (D): Blocking of more distant objects by closer objects.
- Convergence (B): Difference in direction of our eyes when looking at closer objects.
- Accommodation (F): Muscle tension from re-focusing the eye.
- Aerial Haze (E): Objects on the far horizon look hazy due to particles in the air.
- Binocular Disparity (C): Images sensed by our two eyes are slightly different, and this difference is used to determine depth.
- Motion Parallax (A): Moving the head slightly to create differences in the sensed images.

3D Effect in Movies:

Binocular Disparity (C): Binocular disparity, or the slight difference in the images sensed by our two eyes, is mostly used to create the 3D effect in movies. By presenting slightly different images to each eye, filmmakers can simulate the way our eyes perceive depth in the real world, creating a stereoscopic effect for a more immersive viewing experience.

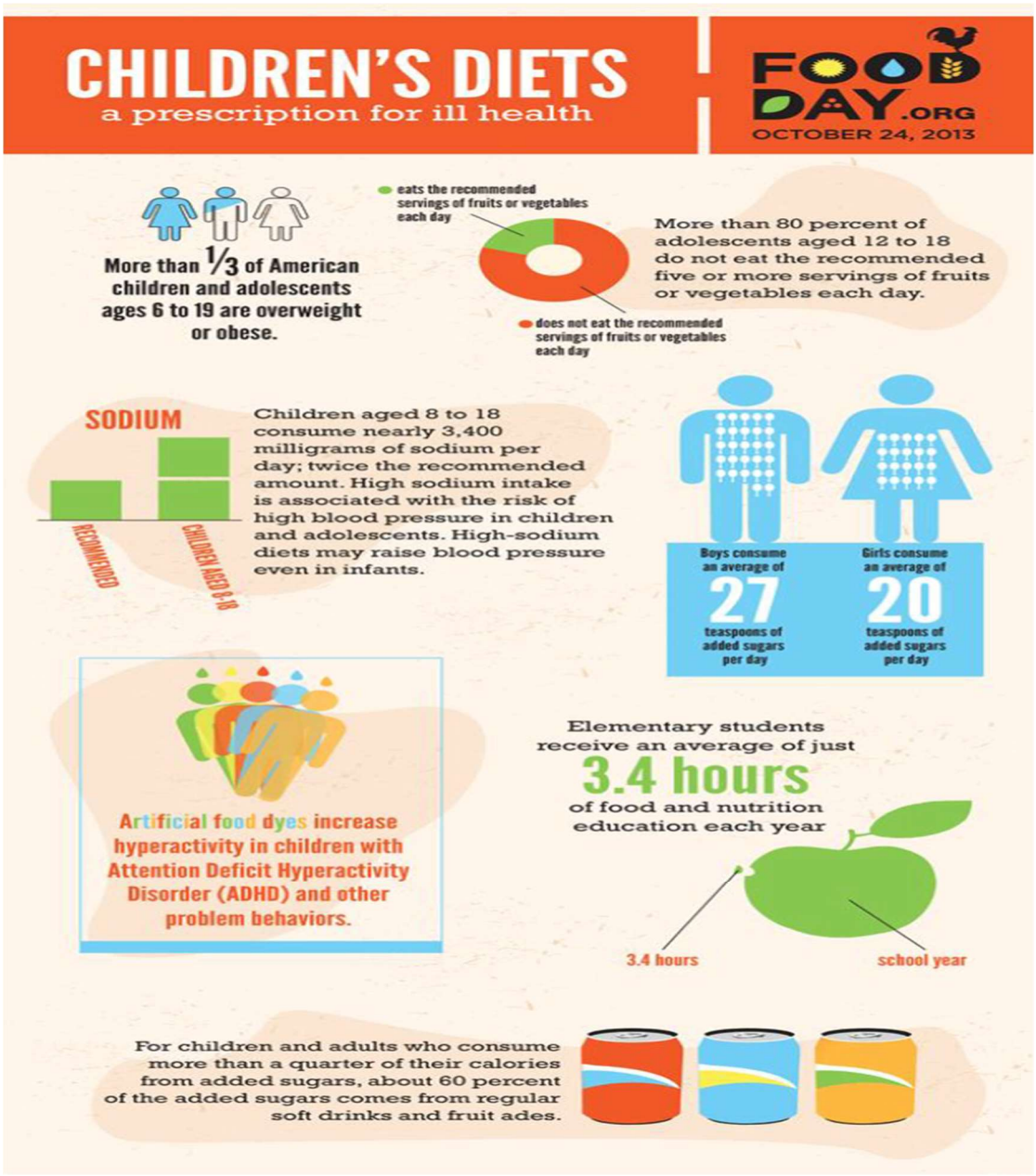
QUESTION 5 ***[TOTAL MARKS: 25]***

Q 5(a) **[7 Marks]**

In the appendix, Figure 2 shows a graphic.
What visual communication goals are evident?
Identify two (2) design principles and explain how the graphic applies them to fulfil the communication goals.

Q 5(b) **[6 Marks]**

Using Figure 2, explain and give examples of two (2) gestalt principles and how they are used.



Q 5(a) Visual communication goals

The visual communication goals of Figure 2 are to:

- Inform the reader about the poor dietary habits of children and adolescents.
- Raise awareness of the health risks associated with these poor dietary habits.
- Encourage readers to take action to improve their children's diets.

Design principles

Two design principles that the graphic applies to fulfill these communication goals are:

- Contrast: The graphic uses contrast to highlight key information and make it visually appealing. For example, the text "Children's Diets: A Prescription for Ill Health" is in a large, bold font and is set against a dark background. This makes it stand out from the rest of the graphic and immediately grabs the reader's attention.
- Proximity: The graphic uses proximity to group related information together. For example, the statistics about children's dietary habits and the health risks associated with those habits are grouped together. This makes it easy for the reader to understand the relationship between the two.

How the graphic applies the design principles

The graphic uses the principle of contrast to highlight the following key information:

- The title of the graphic: "Children's Diets: A Prescription for Ill Health"
- The statistics about children's dietary habits: "More than 80% of adolescents aged 12 to 18 do not eat the recommended five or more servings of fruits or vegetables each day." and "More than 1/3 of American children and adolescents ages 6 to 19 are overweight or obese."
- The health risks associated with poor dietary habits: "High sodium intake is associated with the risk of high blood pressure in children and adolescents." and "Artificial food dyes increase hyperactivity in children with Attention Deficit Hyperactivity Disorder (ADHD) and other problem behaviors."

- The graphic uses the principle of proximity to group the following related information together:

- The statistics about children's dietary habits and the health risks associated with those habits
- The recommendations for improving children's diets

Q 5(b) Gestalt principles

Two gestalt principles that are used in Figure 2 are:

- Closure: The principle of closure states that the human mind tends to fill in gaps and create a complete image, even if all of the information is not present. The graphic uses this principle to create a sense of urgency and alarm. For example, the image of the child's plate with the unhealthy food is incomplete. The child's hand is reaching for the plate, but the food itself is not shown. This leaves the reader to imagine what the food is, and the fact that it is not shown suggests that it is something unhealthy.
- Similarity: The principle of similarity states that the human mind tends to group similar objects together. The graphic uses this principle to highlight the fact that many different types of unhealthy food are popular among children and adolescents. For example, the image of the child's plate includes a variety of unhealthy foods, such as pizza, soda, and chips. This grouping of similar foods makes it clear that these foods are all part of the problem of poor dietary habits in children.

How the graphic uses the gestalt principles

The graphic uses the principle of closure to create a sense of urgency and alarm. The incomplete image of the child's plate with the unhealthy food leaves the reader to imagine what the food is, and the fact that it is not shown suggests that it is something unhealthy. This creates a sense of mystery and suspense, which encourages the reader to learn more about the health risks associated with poor dietary habits in children.

The graphic uses the principle of similarity to highlight the fact that many different types of unhealthy food are popular among children and adolescents. The grouping of similar foods makes it clear that these foods are all part of the problem of poor dietary habits in children. This helps to raise awareness of the problem and encourages readers to take action to improve their children's diets.