Chapter 1

Data processing and information

LEARNING INTENTIONS

By the end of this chapter, you will be able to:

- explain the difference between data and information
- explain the use of direct and indirect sources of data
- evaluate the advantages and disadvantages of direct and indirect sources of data
- describe factors that affect the quality of information
- understand the need for encryption
- describe different methods of encryption
- describe encryption protocols
- explain how encryption is used
- evaluate the advantages and disadvantages of different protocols and methods of encryption
- describe the use of validation methods
- explain the need for both validation and verification
- describe batch, online and real-time processing methods

CONTINUED

- give examples of when batch, online and real-time processing methods are used
- write an algorithm
- evaluate the advantages and disadvantages of different processing methods.

BEFORE YOU START

- Do you know the difference between input and output?
- Do you understand that when input data is processed it can be stored or sent to output?
- Do you understand the term encryption?
- Do you understand the term hacking?
- Do you know what a protocol is?
- Do you understand the structure of a table used in a database?
- Are you able to use a spell checker?

Introduction

We live in a world where we rely on data and information. It is important that data is accurate. Digital technology helps us to manage the input and transfer of data so that it is fit for purpose for its intended audience.

1.1 Data and information

Data

Data is raw numbers, letters, symbols, sounds or images with no meaning.

Some examples of data are:

P952BR
@bbcclick
359
23557.99

KEY WORD

data: raw numbers, letters, symbols, sounds or images without meaning

A picture without context is a further example of raw data.

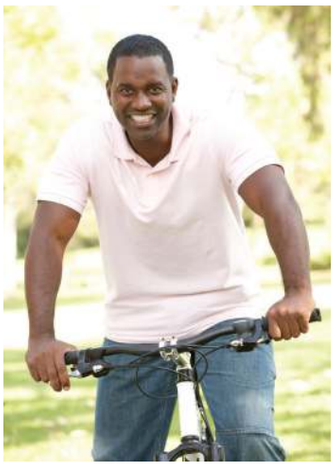

Figure 1.1: Example of raw data.

The data P952BR could have several meanings. It could be:

- a product code
- a postal/ZIP code
- a car registration number.

Because we do not know what the data means, it is meaningless.

REFLECTION

When answering a question such as 'Give one item of data', do not try to explain what the data means because it then becomes **information**. Just give the raw numbers, letters, symbols or image.

KEY WORD

information: data with context and meaning

Information

When data items are given context and meaning, they become information. A person reading the information will then know what it means.

Data is given context by identifying what sort of data it is. This still does not make it information, but it is a step on the way to it becoming information.

Data	Context	Comment
P952BR	A product code	This is a product code, but it is still not known what it is a product code for, so it is still data.
@bbcclick	A Twitter handle	This is an address used for Twitter, but it is not information unless it is known to be a Twitter handle or used within Twitter software. It's also not known whose address it is.
359	Price in Pakistani Rupees	This is a currency value, but it is not known what the price is for, so it is still data.

Table 1.1: Examples of data being given context.

For the data to become information, it needs to be given meaning. Information is useful because it means something.

Data	Context	Meaning
P952BR	A product code	A product code for a can of noodles.
@bbcclick	A Twitter handle	The Twitter address for the BBC's weekly technology show, Click, which is worth watching on BBC World News and BBC2 to keep up to date with technology.
359	Price in Pakistani rupees	The price of a mobile phone cover.

Table 1.2: Examples of data being given context and meaning to become information.

Questions

A company creates websites using style sheets.

1 Identify one item of data that will be used by the company.

2 Describe how this item of data can become information.

Data sources

Direct data source

Data collected from a **direct data source** (primary source) must be used for the same purpose for which it was collected.

KEY WORDS

direct data source: data that is collected for the purpose for which it will be used

The data will often have been collected or requested by the person who intends to use the data. The data must not already exist for another purpose though. When collecting the data, the person collecting should know for what purpose they intend to use the data.

Figure 1.2: Direct data source.

For example, a sports shop wants to find out what other shops are charging for trainers. There are various direct sources from which this data can be collected. These could include:

- visiting the other shops and noting down the prices
- visiting the other shops' websites and noting down the prices
- carrying out a survey of other shop owners to ask their prices (although the shop owners are unlikely to want to give these).

Questionnaires can be used to gather specific data, such as opinions about an event that has taken place. Questionnaires are particularly useful when there are a large number of respondents and statistical analysis will be carried out on the results. Questions on a questionnaire need to be structured carefully to:

- elicit the information required
- enable analysis of the data effectively
- gather enough information without putting people off from completing the questionnaire.

Online questionnaires enable quicker analysis of data because the users fill in the data online and then the data is entered directly into a database. Online questionnaires save time because no further data entry by a third party is necessary.

Interviews are another direct source of information. Questions are asked directly to respondents and the interviewer can ask the respondent to elaborate on their answers.

Indirect data source

Figure 1.3: Indirect data source.

Data collected from an **indirect data source** (secondary source) already existed for another purpose. Although it may have been collected by the person who intends to use it, it was often collected by a different person or organisation.

KEY WORDS

indirect data source: data that was collected for a different purpose (secondary source)

For example, a sports shop could use various indirect sources to find out what other shops are charging for trainers, including:

- carrying out a survey of customers who have purchased trainers from the other shops (in this case, the price will be the one paid by the customer, which may have been a different price to that charged now, or it may have been discounted at the time)
- looking at till receipts from the shop (the price is printed on the till receipt for the purpose of providing proof of purchase, not for identifying prices).

PRACTICAL ACTIVITY 1.01

Which of the following are direct data sources and which are indirect data sources?

Data	Reason collected	Reason used
Names and email addresses of members of a political party	To record their membership and to be able to contact them.	To contact members by email to see if they will donate some money.
Employee attendance dates and times	To identify when employees attended work and to calculate their wages.	To allow a police officer to check an employee's alibi if a crime has been committed.
Flight times and prices from airline websites	To compare the prices and times for a trip to Florida.	To decide the best flight to use for a trip to Florida.
Names, ages and addresses of people	For a national census.	To allow a marketing company to find out which areas have the highest population of children.
Weather measurements from a weather station	To record the current weather.	To show the current temperature and rainfall on a website.

REFLECTION

Direct data is usually used by the person that collected it and for the purpose they collected it. However, it's also possible for a person to collect data from an indirect (secondary) source. For example, if a journalist is writing a news article and bases his story on existing news articles, then he has used indirect sources rather than interviewing the people involved in the original story.

One indirect source of information that is commonly used is an electoral register. Governments keep a register of people who are registered to vote in each household. This register includes names and addresses. Its main purpose is to enable those people to vote in elections. However, it can also be used by credit reference agencies to check whether a person lives at the address they say they do, or by marketing agencies to send direct marketing to the people listed on the register. There is an option for individual entries on an electoral register to be hidden from public view.

Businesses that want to send marketing letters will often purchase a list of email addresses or telephone numbers or addresses of people. Selling data is a big business, especially if the data enables a company to direct their marketing at their target market. For example, a company selling IT textbooks to schools would benefit greatly from a list of email addresses of IT teachers. Different countries have different laws about how personal data can be used in this way, but most developed nations have data protection laws that require companies to get consent from customers before the customers' data can be shared with a third party.

Advantages and disadvantages of gathering data from direct and indirect data sources

The general rule is that data collected directly for the purpose for which it is intended is more likely to be accurate and relevant than data that is obtained from existing data (indirect source).

Direct data source	Indirect data source
The data will be relevant because what is needed has been collected.	Additional data that is not required will exist that may take time to sort through and some data that is required may not exist.
The original source is known and so can be trusted.	The original source may not be known and so it can't be assumed that it is reliable.
It can take a long time to gather original data rather than use data that already exists.	The data is immediately available.
A large sample of statistical data can be difficult to collect for one-off purposes.	If statistical analysis is required, then there are more likely to be large samples available.

Direct data source	Indirect data source
The data is likely to be up to date because it has been collected recently.	Data may be out of date because it was collected at a different time.
Bias can be eliminated by asking specific questions.	Original data may be biased due to its source.
The data can be collected and presented in the format required.	The data is unlikely to be in the format required, which may make extracting the data difficult.

Table 1.3: Comparing direct and indirect sources.

Questions

This spreadsheet is used to calculate the area of a driveway.

	A	B	C
1	**Area calculator**		
2	Length =	3	m
3	Width =	5	m
4	Area =	15	m^2

Figure 1.4: Part of a spreadsheet.

The builder using the spreadsheet needs to know the length and width of a driveway for a customer.

3 Identify one direct source the builder could use to find the length and width.

4 Identify one indirect source the builder could use to find the length and width.

5 Give one advantage of using the direct source instead of the indirect source to find the length and width.

Note: static and dynamic data is extension content, and is not part of the syllabus.

1.2 Quality of information

The quality of information is determined by a number of attributes.

Accuracy

Information that is inaccurate is clearly not good enough. Data must be accurate in order to be considered of good quality. Imagine being told that you need to check in at the airport 45 minutes before the flight leaves, so you turn up at 18:10 for a 19:05 flight, only to find that you were actually supposed to check in one hour before the flight leaves.

Examples of inaccurate information include:

- decimal point in the wrong place, for example $90.30 instead of $903.00 could suggest a product is much cheaper than it really is
- misspelling such as 'stair' instead of 'stare', where words have completely different meanings
- misplaced characters, such as a licence plate of BW9EP3T instead of BW93PET.

Relevance

Information must be relevant to its purpose. Having additional information that is not required means that the user must search through the data to find what is actually required.

Examples of irrelevant information include:

- being given a bus timetable when you want to catch a train
- being told the rental price of a car when you want to buy the car
- a user guide for a mobile phone that includes instructions on how to assemble a plug.

Age

Information must be up to date in order to be useful. Old information is likely to be out of date and therefore no longer useful. When using indirect data sources, always check when the information was produced.

Examples of out of date information include:

- the number of residents in a town based on a census from 2011, but 500 new homes have been built in the town since then
- a rugby score that has not been updated for 5 minutes, during which time a player scored.

Level of detail

There needs to be the right amount of information for it to be good quality. It's possible to have either too little or too much information provided. If there is too much information, then it can be difficult to find the exact information required. If there is not enough information, then it is not possible to use it correctly.

For example, a person orders a pizza. They ask for a large pepperoni to be delivered. They forgot to say what type of base they wanted and where it should be delivered to. The pizza company does not have enough information to fulfil the order.

Another example could be a traveller who needs to catch a train from Bhopal to Kacheguda. The traveller phones the rail company to find out the time of departure and arrival for trains, but they have to listen to all the times of the stations in between before they get the arrival time at Kacheguda.

Completeness

All information that is required must be provided in order for it to be of good quality. Not having all the information required means it cannot be used properly.

For example, a person has booked their car in for a service over the phone. The mechanic at the garage tells them the name of the street but doesn't give the building number.

PRACTICAL ACTIVITY 1.02

Look at this invitation.

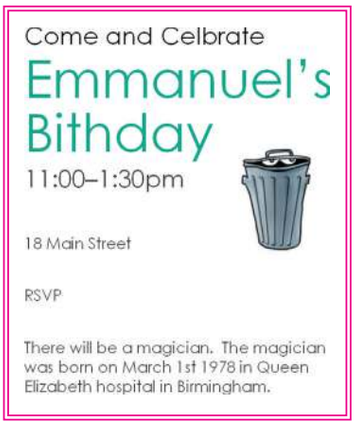

Figure 1.5: Birthday party invitation.

CONTINUED

Describe how accuracy, relevance, level of detail and completeness affect the quality of information in the invitation.

Questions

6 Identify three factors that could affect the quality of information.

7 Describe how the age of information could affect the quality of information within a user guide for a mobile phone.

1.3 Encryption

One specific type of encoding is **encryption**. Encryption is when data is scrambled so that the data cannot be understood. Data can be encrypted when it is stored on disks or other storage media, or it can be encrypted when it is sent across a network, such as a local area network or the internet. Encryption is important when sending or storing sensitive data such as personal data or a company's sales figures. Data being sent across a network or the internet can be intercepted by hackers. Data stored on storage media could be stolen or lost. The purpose of encryption is to make the data difficult or impossible to read if it is accessed by an unauthorised user. Accessing encrypted data legitimately is known as decryption.

KEY WORD

encryption: scrambling data so it cannot be understood without a decryption key to make it unreadable if intercepted

Caesar cipher

A cipher is a secret way of writing. In other words it is a code. Ciphers are used to convert a message into an encrypted message. It is a special type of algorithm which defines the set of rules to follow to encrypt a message. Roman Emperor Julius Caesar created the Caesar cipher so that he could communicate in secret with his generals.

The Caesar cipher is sometimes known as a shift cipher because it selects replacement letters by shifting along the alphabet.

WORKED EXAMPLE 1.01

In this example the alphabet is to be shifted by three (+3) letters so that A = D, B = E and so on.

Original	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
Encrypted	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C

So to encrypt the word 'Hello', you would use:

H = K, E = H, L = O, O = R

which gives KHOOR.

While the Caesar cipher is very easy to use, it's also very easy to crack.

PRACTICAL ACTIVITY 1.03

1. Using the Caesar cipher +3 example above, write an encrypted message to a friend. Ask your friend to decipher it.
2. Choose how many letters you are going to shift by and write another encrypted message to a friend. Don't tell your friend how many letters you shifted by. Your friend should try to decipher the code by working out which letters appear most commonly.
3. Look online for how to 'create a cipher wheel ' and use it to encrypt and decrypt messages.

Symmetric encryption

Symmetic encryption is the oldest method of encryption. It requires both the sender and recipient to possess a secret encryption and decryption key known as a private key. With symmetric encryption, the secret key needs to be sent to the recipient. This could be done at a separate time, but it still has to be transmitted whether by post or over the internet and it could be intercepted.

Figure 1.6: Symmetric encryption.

Asymmetric encryption

Asymmetric encryption is also known as public-key cryptography. Asymmetric encryption overcomes the problem of symmetric encryption keys being intercepted by using a pair of keys. This will include a public key which is available to anybody wanting to send data, and a private key that is known only to the recipient. The key is the algorithm required to encrypt and decrypt the data.

The process works like this:

Figure 1.7: Asymmetric encryption.

Here is an example. Tomasz sends a message to Helene. Tomasz encrypts the message using Helene's public key. Helene receives the encrypted message and decrypts it using her private key.

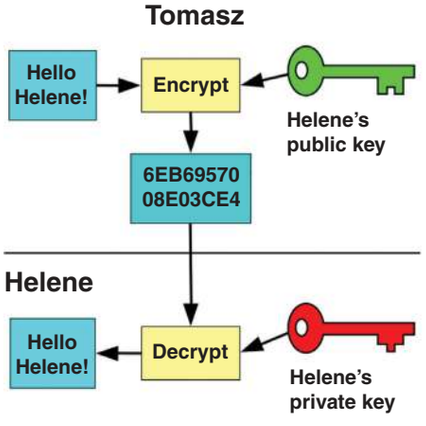

Figure 1.8: Example of asymmetric encryption.

Asymmetric encryption requires a lot more processing than symmetric encryption and so it takes longer to decrypt the data. However, as the decryption key does not have to be transmitted, it is more secure than symmetric encryption.

In order to find a public key, digital certificates are required which identify the user or server and provide the public key. A digital certificate is unique to each user or server. A digital certificate usually includes:

- organisation name
- organisation that issued the certificate
- user's email address
- user's country
- user's public key.

When encrypted data is required by a recipient, the computer will request the digital certificate from the sender. The public key can be found within the digital certificate.

Asymmetric encryption is used for Secure Sockets Layer (**SSL**) which is the security method used for secure websites. Transport Layer Security (**TLS**) has superseded SSL but they are both often referred to as SSL. Once SSL has established an authenticated session, the client and server will create symmetric keys for faster secure communication.

KEY WORDS

SSL: Secure Socket Layer

TLS: Transport Layer Security

PRACTICAL ACTIVITY 1.04

Find and watch a video about SSL.

Applications of encryption

Hard disk

Disk encryption will encrypt every single bit of data stored on a disk. This is different to encrypting single files. In order to access any file on the disk, the encryption key will be required. This type of encryption is not limited to disks and can be used on other storage media such as backup tapes and Universal Serial Bus (USB) flash memory. It is particularly important that USB flash memory and backup tapes are encrypted because these are portable storage media and so are susceptible to being lost or stolen. If the whole medium is encrypted, then anybody trying to access the data will not be able to understand it. The data is usually accessed by entering a password or using a fingerprint to unlock the encryption.

HTTPS

Normal web pages that are not encrypted are fetched and transmitted using Hypertext Transfer Protocol (HTTP). Anybody who intercepts web pages or data being sent over HTTP would be able to read the contents of the web page or the data. This is particularly a problem when sending sensitive data, such as credit card information or usernames and passwords.

Hypertext Transfer Protocol Secure (HTTPS) is the encryption standard used for secure web pages. It uses Secure Socket Layer (SSL) or Transport Layer Security (TLS) to encrypt and decrypt pages and information sent and received by web users. SSL was first used in 1996 and was replaced by TLS in 1999. SSL can still be used, but it has vulnerabilities so it's not recommended. TLS is the protocol that is used by banks when a user logs onto online banking. A secure web page can be spotted by its address beginning with https:// and in addition some browsers display a small padlock.

KEY WORD

HTTPS: Hypertext Transfer Protocol Secure

Figure 1.9: The 's' after 'http' and the padlock indicate that this is a secure website.

When a browser requests a secure page, it will check the digital certificate to ensure that it is trusted, valid and that the certificate is related to the site from which it is coming. The browser then uses the public key to encrypt a new symmetric key that is sent to the web server. The browser and web server can then communicate using a symmetric encryption key, which is much faster than asymmetric encryption.

WORKED EXAMPLE 1.02

The web browser requests the certificate from the web server.

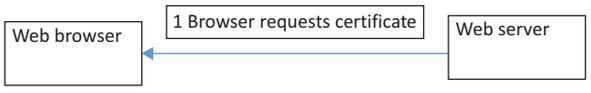

Figure 1.10: Asymmetric cryptography.

The web browser then uses the web server's public key to encrypt a new symmetric key and sends that encrypted symmetric key to the web server. The web server uses its own private key to decrypt the new symmetric key.

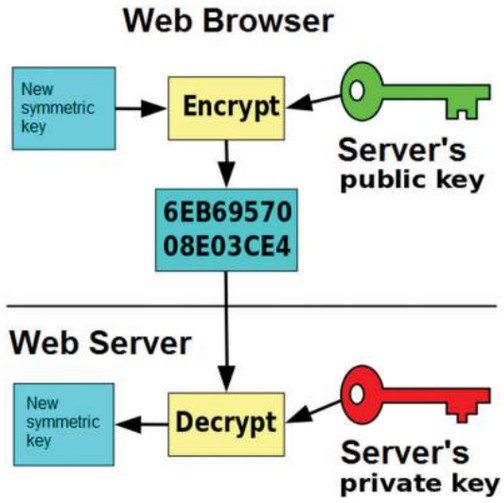

Figure 1.11: Secure website identification.

The browser and web server now communicate using the same symmetric key.

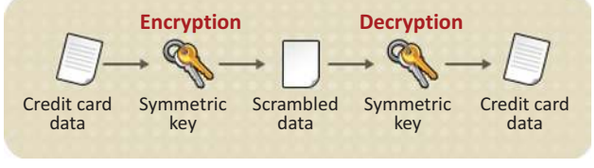

Figure 1.12: Symmetric encryption.

Email

Email encryption uses asymmetric encryption. This means that recipients of emails must have the private key that matches the public key used to encrypt the original email. In order for this to work, both the sender and recipient need to send each other a digitally-signed message that will add the person's digital certificate to the contact for that person. Encrypting an email will also encrypt any attachments.

How encryption protects data

Encryption only scrambles the data so that if it is found, it cannot be understood. It does not stop the data from being intercepted, stolen or lost. However, with strong 256-bit AES encryption, it is virtually impossible for somebody to decrypt the data and so it is effectively protected from prying eyes.

REFLECTION

Most Wi-Fi access points and Wi-Fi routers use encryption. This serves two purposes. The first is to only allow people who know the 'key' (usually a password) to access the network, so that any unauthorised users cannot gain access. The second is to encrypt the data, so that it cannot be understood by somebody 'snooping' on the Wi-Fi network.

Wi-Fi networks are particularly susceptible to 'snooping' because no wires are required to connect to the network. It is possible to sit in a car outside somebody's house and see the Wi-Fi network. The 'key' stops that person from accessing the network and also stops that person from understanding the data that is moving around the network.

Did you know that, if you access a public Wi-Fi hotspot that is 'open' and therefore not encrypted, anybody with the right software can see what you are sending over the network, including your emails? This applies to laptops, tablets and mobile phones or any other device using public Wi-Fi.