
multiverse

Data Analysis in Industry

Session 2

Session Outline

Principles of Data Classification

Types of Data

Data Structures

Data Sources

Data Storage

Data Quality

Data Usage

The Portfolio

Recap

Learning Objectives

- Identify **Business Specific Rules** related to datasets and data characteristics that will influence project design and analysis
- Describe the key characteristics of the different **Data Formats** and how to work with them

multiverse

Principles of Data Classification

Things to consider

Things to consider

What are the types of data?

Things to consider

What are the types of data?

Where is the data stored?

Things to consider

What are the types of data?

Where is the data stored?

Where did the data come from?

Things to consider

What are the types of data?

Where is the data stored?

Where did the data come from?

Is any data sensitive?

Things to consider

What are the types of data?

Where is the data stored?

Where did the data come from?

Is any data sensitive?

Who has access to the data?

Things to consider

What are the types of data?

Where is the data stored?

Where did the data come from?

Is any data sensitive?

Who has access to the data?

How is the data protected?

Things to consider

What are the types of data?

Where is the data stored?

Where did the data come from?

Is any data sensitive?

Who has access to the data?

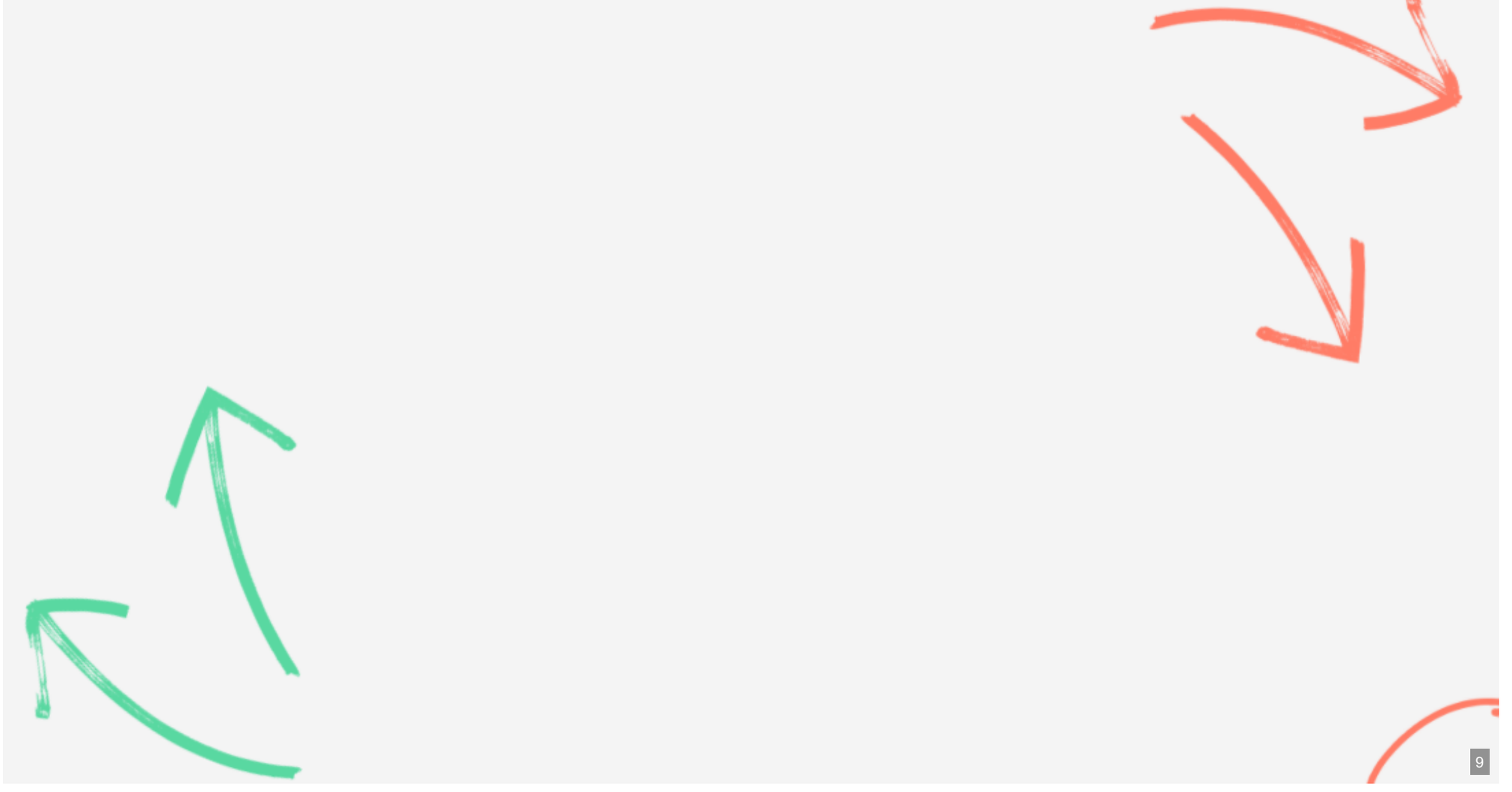
How is the data protected?

How are you going to comply with GDPR?

Content	Context	User

multiverse

Types of Data



Quantitative

Quantitative

Discrete

Quantitative

Discrete

Continuous


Quantitative
Discrete
Continuous
Qualitative

Quantitative

Discrete

Continuous

Qualitative

Binomial

Quantitative

Discrete

Continuous

Qualitative

Binomial

Nominal


Quantitative
Discrete
Continuous
Qualitative
Binomial
Nominal
Ordinal

Quantitative

Discrete

Continuous

Numerical data that can be 'counted'

e.g. number of marbles, siblings, customers, etc

Quantitative

Discrete

Continuous

Numerical data that can be 'measured'

e.g. temperature, weight, height

Categorical data that has two options

e.g. true or false, heads or tails, yes or no

Qualitative

Binomial

Nominal

Ordinal

Categorical data that has multiple options but no implied order

e.g. colour, job title, error type, etc

Qualitative

Binomial

Nominal

Ordinal

Categorical data that multiple options and an implied order

e.g. likert scale, coffee cup size, salary band, etc

Qualitative

Binomial

Nominal

Ordinal

Identify the Qualitative Data

Weight of a baby	Emotional state	Colour of a bottled drink
Political opinion	Your height	Number of shoes you own
Car type	Holiday destination	Distance to your nearest shop
Number of classes on a timetable	Movie rating	IQ score

Identify the Qualitative Data

Weight of a baby	Emotional state	Colour of a bottled drink
Political opinion	Your height	Number of shoes you own
Car type	Holiday destination	Distance to your nearest shop
Number of classes on a timetable	Movie rating	IQ score

Activity

In groups discuss data you use regularly and whether it is quantitative or qualitative

- What subdivision does it fall under?
- How do you visualise it?
- How do you use it?

multiverse

Data Structures

	STRUCTURED DATA	UNSTRUCTURED DATA
CHARACTERISTICS:	▪ Pre-defined data models ▪ Usually text only ▪ Easy to search	▪ No pred-defined data model ▪ May be text, images, audio, video or other formats ▪ Difficult to search
STORED IN:	▪ Relational databases ▪ Data warehouses	▪ Applications ▪ NoSQL databases ▪ Data lakes
GENERATED BY:	▪ Humans or machines	▪ Humans or machines

	STRUCTURED DATA	UNSTRUCTURED DATA
APPLICATION EXAMPLES:	▪ Online reservation system ▪ Inventory control ▪ CRM systems ▪ ERP systems	▪ Word processing ▪ Presentation software ▪ Email clients ▪ Media editing tools
DATA EXAMPLES:	▪ Dates ▪ Product names and numbers ▪ Customer name ▪ Error code ▪ Transaction information	▪ Text files ▪ Audio files ▪ Video files ▪ Images ▪ Emails and reports

Structured

Unstructured

Highly organised

Easily read by machines

YEAR	SITES	PARTICIPATION	MEALS SERVED
1968	0.9	56	0.2
1969	1.2	99	0.3
1970	1.9	227	1.8
1971	3.2	569	8.2
1972	6.5	1080	21.9
1973	11.2	1437	26.6
1974	10.6	1403	33.6
1975	12.0	1785	50.3
1976	16.0	2453	73.4

YEAR	SITES	PARTICIPATION	MEALS SERVED
1968	0.9	56	0.2
1969	1.2	99	0.3
1970	1.9	227	1.8
1971	3.2	569	8.2
1972	6.5	1080	21.9
1973	11.2	1437	26.6
1974	10.6	1403	33.6
1975	12.0	1785	50.3
1976	16.0	2453	73.4

YEAR	SITES	PARTICIPATION	MEALS SERVED
1968	0.9	56	0.2
1969	1.2	99	0.3
1970	1.9	227	1.8
1971	3.2	569	8.2
1972	6.5	1080	21.9
1973	11.2	1437	26.6
1974	10.6	1403	33.6
1975	12.0	1785	50.3
1976	16.0	2453	73.4

Cannot be processed using conventional tools

Be careful!

Sometimes data looks structured but isn't. For example, Excel spreadsheets have no rules around usage, so you can have multiple tables or different data types in one column.

STRUCTURE	FEATURES
File	▪ Used to store information ▪ Used by computers to read and write information that needs to be processed ▪ Organised into record
List	▪ Contains elements of different data types ▪ E.g. ('John', 10, 7.2, True)
Array	▪ Data can be identified by their index position ▪ Similar to a list but can have multiple dimensions ▪ A 2 dimensional array is a matrix
Table	▪ Typical data files with labelled columns (fields) and rows (records)
Tree	▪ Hierarchical collection of data with parent and child nodes

Activity

Discuss whether the data you use regularly is structured on unstructured

Further reading:

Data Lake vs Data Warehouse
SQL vs NoSQL
Structured vs Unstructured Data

multiverse

Data Sources

Public Data

Client Data

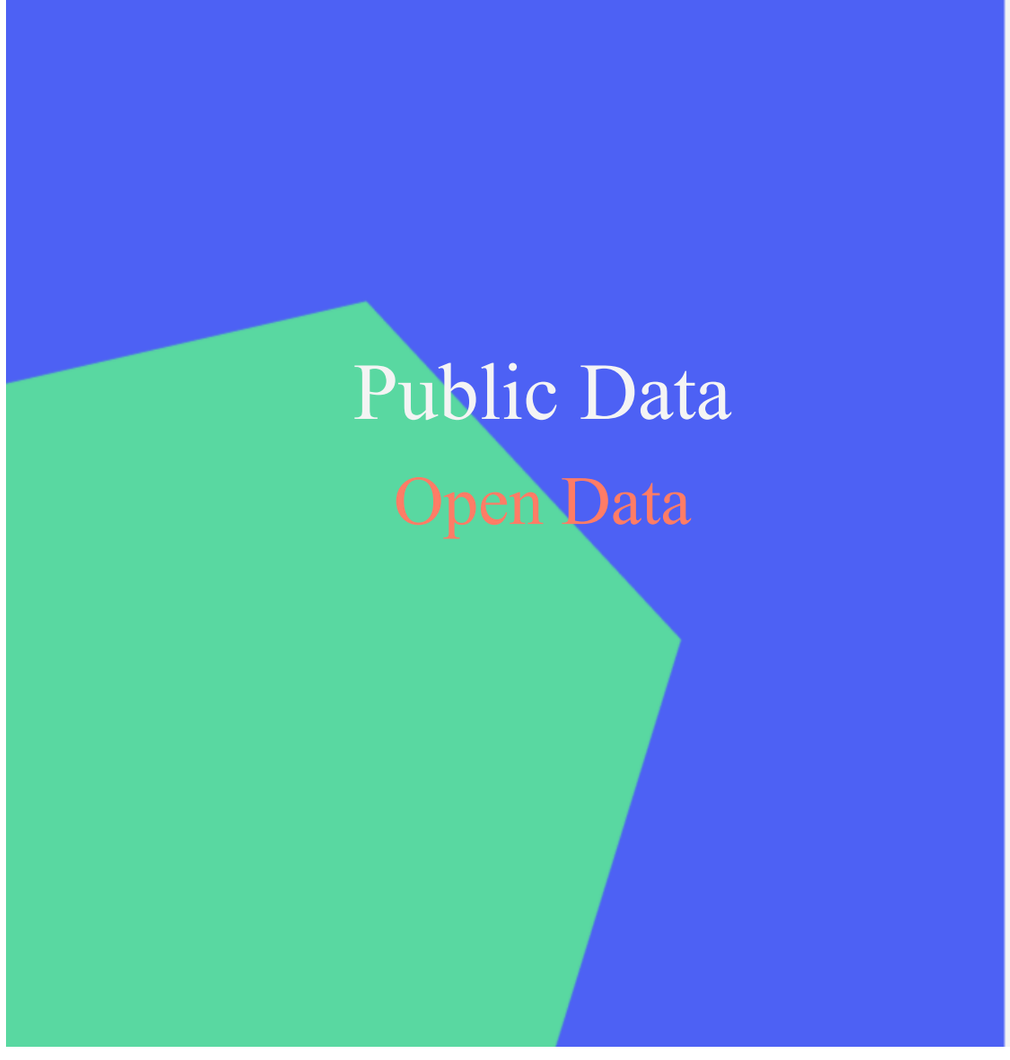
Proprietary Data

Research Data

Public Data

Open Data

Data that can be moved freely, reused and redistributed, although hard to change or modify

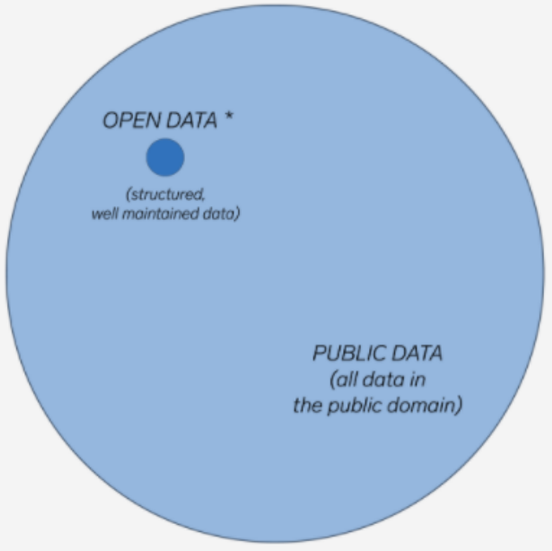


A subset of public data but:

- Smaller in volume
- More likely to be structured
- More likely to be open licensed
- Better maintained and more reliable through sanctioned portals
- May require a nominal fee to be used

According to the Open Knowledge Foundation:

"Open data and content can be freely used, modified, and shared by anyone and for any purpose."



** According to the Open Data Barometer's Global Report 2017, only **7%** of key datasets across 115 countries were considered open. The open data circle size is **7%** of data otherwise considered public.*

Proprietary

Operational

Administrative

Data that is owned and stored within an organisation. Proprietary data may be protected by patents, copyrights/trademarks or trade laws.

Proprietary

Operational

Administrative

Proprietary data that is produced by your organisations day to day operations.

E.g. customer, inventory or purchase data

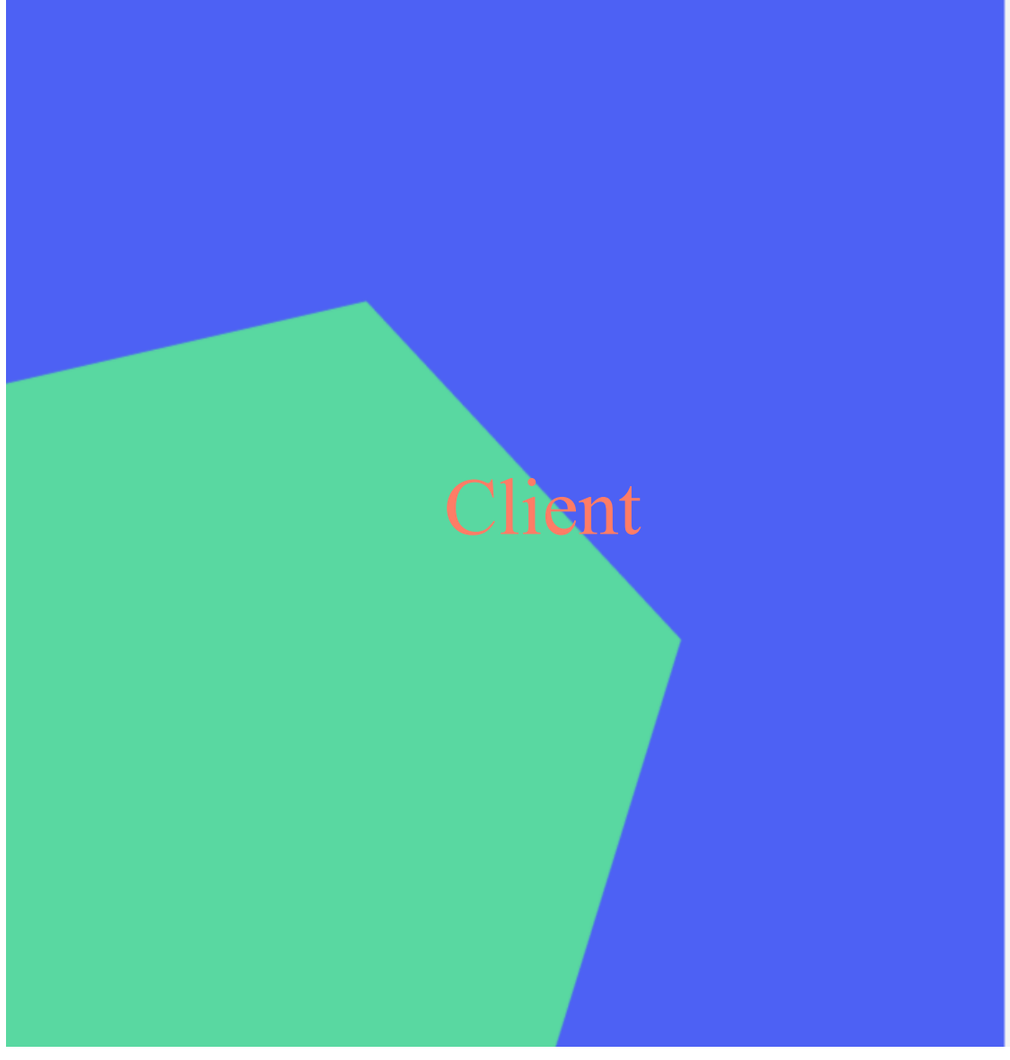
Proprietary

Operational

Administrative

Required to run an organisations day to day operations

E.g. HR, payroll, admin



Proprietary data provided by a client

E.g. data provided by a consultancy firm

Research

Observational

Simulation

Derived

Data from a third party that is made available to you under a licence agreement or has been collected, generated or created to validate original research findings.

Research

Observational

Simulation

Derived

Data gathered from observing trends in the population or from experiments

For example, are shoppers more likely to buy items at eye level?

Research

Observational

Simulation

Derived

Data gathered from a theoretical experiment based on past information

For example, simulating what will happen to the housing market if interest rates rise.

Research

Observational

Simulation

Derived

Data that has been created from other sources

For example, a data warehouse created with ETL

Identifiability	Sensitivity	Availability
Can someone be identified?	What damage can be done?	How readily available is the data?

Things to consider:

Things to consider:

Data Accuracy - Can we trust this data? Is it up to date? Is it relevant?

Things to consider:

Data Accuracy - Can we trust this data? Is it up to date? Is it relevant?

Limitations of Data - Are things excluded?

Things to consider:

Data Accuracy - Can we trust this data? Is it up to date? Is it relevant?

Limitations of Data - Are things excluded?

Compatibility with other data sources - Can we join this to our data?

Things to consider:

Data Accuracy - Can we trust this data? Is it up to date? Is it relevant?

Limitations of Data - Are things excluded?

Compatibility with other data sources - Can we join this to our data?

Legal & regulatory rights to data - Are we allowed to use this data?

Things to consider:

Data Accuracy - Can we trust this data? Is it up to date? Is it relevant?

Limitations of Data - Are things excluded?

Compatibility with other data sources - Can we join this to our data?

Legal & regulatory rights to data - Are we allowed to use this data?

Business Context - Do we understand the quirks of this data?

Activity

Open each of the files and discuss what the defining features are of each

- What do you think the benefits are?
- What about limitations?
- Do you think they are easier for a human or computer to read?
- Which tools/software can you use with each?

Extensible Markup Language
Comma Separated Values
Text File
Rich Text Format
Excel
JavaScript Object Notation

FILE FORMAT	PROPERTIES	BENEFITS	LIMITATIONS
.xml (e**X**tensible **M**arkup **L**anguage)	A hierarchy based markup language that uses user defined keywords to tag data	▪ Easily read by machines ▪ Portable to many different systems	▪ Hard for humans to read ▪ Large size due to repeated markups
.csv (**C**omma **S**eparated **V**alues)	Tabular data separated by commas. Is a raw text value	▪ Lightweight ▪ Easily read by many applications	▪ If there are commas within the data they need to be 'text qualified' so interpreter knows they are not delimiters
.rtf (**R**ich **T**ext **F**ormat)	A file that is stored as Raw text but has a markup language to denote basic formatting such as bold, underline etc.	▪ Fairly lightweight ▪ Suitable for holding documents, not actual data	▪ Rarely used ▪ Hard to read due to markups ▪ Used only for wordpad

FILE FORMAT	PROPERTIES	BENEFITS	LIMITATIONS
.txt (Text)	Text-based with no formatting or tags. Can be delimited by anything.	▪ Flexible ▪ Lightweight ▪ Easily read	▪ Can easily break ▪ Needs text qualification
.xlsx (Excel File)	Proprietary spreadsheet file format created by Microsoft Excel	▪ Many users are comfortable with this format ▪ Widely used	▪ Large file size ▪ Specialist software needed to view or edit ▪ Hard for applications to read
.json (**J**ava**S**cript **O**bject **N**otation	Text-based open standard designed for human-readable data interchange.	▪ Structure easily read by applications ▪ Lightweight	▪ No error handling ▪ Can leave your machine vulnerable to attacks if taken from an untrusted source

multiverse

Data Storage

Common Types of Database

Common Types of Database

Relational Database Management System (RDBMS)

Common Types of Database

Relational Database Management System (RDBMS)

Not Only SQL (NoSQL)

Access Security

Access

Security

Activity

Discuss what types of database you may have access to in your role.

Who else has access?

What security steps does your organisation have in place?

multiverse

Data Quality




Accuracy	Complete	Consistency	Uniqueness	Timeliness

If you find an error...

You should either...

If you find an error...

You should either...

→ Correct it

If you find an error...

You should either...

→ Correct it

→ Impute a new value

If you find an error...

You should either...

→ Correct it

→ Impute a new value

→ Remove it

If you find an error...

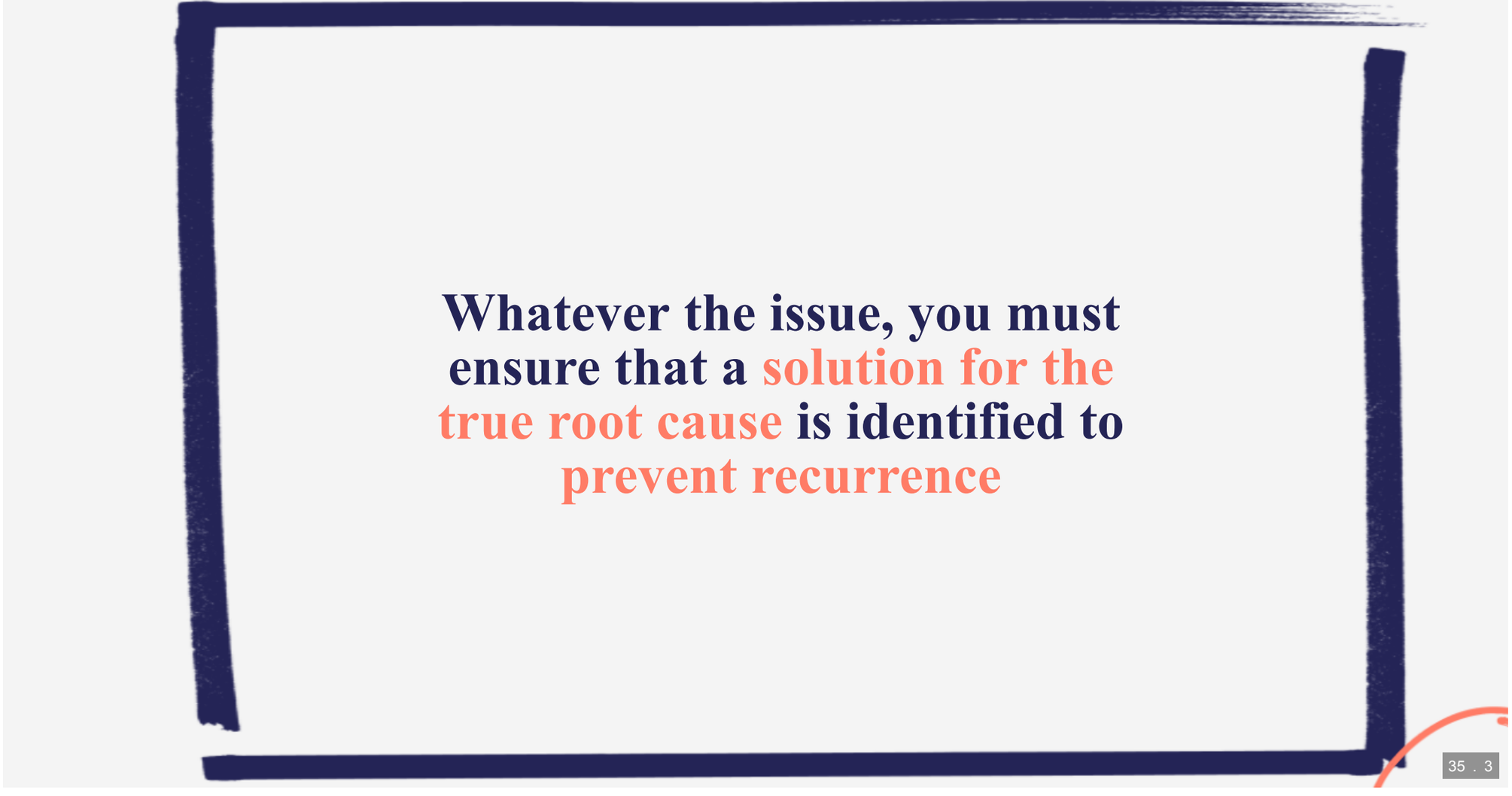
You should either...

→ Correct it

→ Impute a new value

→ Remove it

→ Ignore it


Whatever the issue, you must ensure that a solution for the true root cause is identified to prevent recurrence
35 . 3

What are the consequences of poor data quality?

What are the consequences of poor data quality?

→ Bad Business Decisions

What are the consequences of poor data quality?

→ Bad Business Decisions

→ Inefficient Business Practices

What are the consequences of poor data quality?

→ Bad Business Decisions

→ Inefficient Business Practices

→ Lost Market Reputation

What are the consequences of poor data quality?

→ Bad Business Decisions

→ Inefficient Business Practices

→ Lost Market Reputation

→ Missed Opportunities

What are the consequences of poor data quality?

→ Bad Business Decisions

→ Inefficient Business Practices

→ Lost Market Reputation

→ Missed Opportunities

→ Lost Revenue

What are the consequences of poor data quality?

→ Bad Business Decisions

→ Inefficient Business Practices

→ Lost Market Reputation

→ Missed Opportunities

→ Lost Revenue

→ Breach of Data Protection Laws

Activity

Examine this file and write down any problems you find with the data

multiverse

Data Usage

General Data Protection Regulation (GDPR 2018)

1	Data must be processed **lawfully, fairly and transparently**
2	Data must be collected for **specified, explicit and legitimate purposes**
3	Data must be **adequate, relevant and limited to what is necessary** for processing
4	Data must be **accurate and kept up to date**
5	Data must be **kept only for as long as is necessary** for processing
6	Data must be processed in a manner that **ensures its security**

Why?

What?

When?

Where?

How?

Everyone has a right to be informed...

Why?

What?

When?

Where?

How?

Everyone has a right to be informed...

- How their data will be used

Why?

What?

When?

Where?

How?

Everyone has a right to be informed...

- How long their data will be kept

Why?

What?

When?

Where?

How?

Everyone has a right to be informed...

- Where it will be processed

Why?

What?

When?

Where?

How?

Everyone has a right to be informed...

- Who else will have access

Why?

What?

When?

Where?

How?

Everyone has a right to be informed...

- How they can access their data

Why?

What?

When?

Where?

How?

Everyone has a right to be informed...

- How they can correct any information

Why?

What?

When?

Where?

How?

Everyone has a right to be informed...

- How they can delete their information

Why?

What?

When?

Where?

How?

Everyone has a right to be informed...

- How they can prevent their data being processed any further

What counts as PII?

What counts as PII?

- Name

What counts as PII?

- Name
- Address

What counts as PII?

- Name
- Address
- Contact Details

What counts as PII?

- Name
- Address
- Contact Details
- Bank Details

What counts as PII?

- Name
- Address
- Contact Details
- Bank Details
- Driving License

What counts as PII?

- Name
- Address
- Contact Details
- Bank Details
- Driving License
- Passport Number

What counts as PII?

- Name
- Address
- Contact Details
- Bank Details
- Driving License
- Passport Number
- IP address

Why?

What?

When?

Where?

How?

Processing data includes...

Why?

What?

When?

Where?

How?

Processing data includes...

- Collecting, recording, storing, organising and deleting

Why?

What?

When?

Where?

How?

Processing data includes...

- Collecting, recording, storing, organising and deleting
- Only using it as agreed by the owner

Why?

What?

When?

Where?

How?

Processing data includes...

- Collecting, recording, storing, organising and deleting
- Only using it as agreed by the owner
- Keeping it correct, up to date and relevant

What About Consent?

Consent is ***"any freely given, specific, informed and unambiguous indication of the data subject's wishes by which he or she, by a statement or by a clear affirmative action, signifies agreement to the processing of personal data relating to him or her".***

Article 4(11) GDPR

Why?

What?

When?

Where?

How?

- Data should only be kept as long as necessary

Why?

What?

When?

Where?

How?

- Data should only be kept as long as necessary

- Your organisation will have a policy which defines 'reasonable use'

Why?

What?

When?

Where?

How?

- Data should only be kept as long as necessary
- Your organisation will have a policy which defines 'reasonable use'
- If the data is no longer necessary it must be deleted

Why?

What?

When?

Where?

How?

- Data should only be kept as long as necessary
- Your organisation will have a policy which defines 'reasonable use'
- If the data is no longer necessary it must be deleted
- This includes all digital, hard copies and backups

Why?

What?

When?

Where?

How?

- Where data is stored must be stated

Why?

What?

When?

Where?

How?

- Where data is stored must be stated
- As does everyone who will have access

Why?

What?

When?

Where?

How?

- Where data is stored must be stated
- As does everyone who will have access
- Organisations will have strict policies on who can access data

Why?

What?

When?

Where?

How?

- Where data is stored must be stated
- As does everyone who will have access
- Organisations will have strict policies on who can access data
- They will also set guidelines on how access can be granted

Why?

What?

When?

Where?

How?

- Organisations should keep a regularly reviewed record of who has access

Why?

What?

When?

Where?

How?

- Organisations should keep a regularly reviewed record of who has access

- Data should be kept secure with protections such as special logins

Why?

What?

When?

Where?

How?

- Organisations should keep a regularly reviewed record of who has access

- Data should be kept secure with protections such as special logins

- Where possible, data should be processed on company machines or secure VPNs

Why?

What?

When?

Where?

How?

- Organisations should keep a regularly reviewed record of who has access

- Data should be kept secure with protections such as special logins

- Where possible, data should be processed on company machines or secure VPNs

- Employees should be given training in how to keep data secure

Contributing to a Safe Environment

Everybody who uses personal data in their role must comply with GDPR

Documenting

- How was the data obtained?
- How did you ensure the data was accurate and up to date?
- What did you intend to do with the data?
- How long did you intend to use it for?
- What will you do with the obsolete data?

Activity

Think about some of the data you use in your role

Try and answer the questions on the previous slide and discuss in your group

Are there any other data policies your organisation has put in place?

When things go wrong...

Facebook Data Breach
July 2017-September 2018

29 Million people affected

British Airways Hack
August 2018 - September 2018

380,000 people affected

multiverse

Privacy By Design

Privacy must become integral to organisational priorities, project objectives, design processes, and planning operations. Privacy must be embedded into every standard, protocol and process that touches our lives.

1. *Proactive* not *Reactive*

Anticipate data breaches before they happen by putting into place appropriate security measures

2. Privacy as the *Default*

Personally Identifiable Information is automatically made secure to ensure personal data is kept safe

3. Privacy *Embedded* into *Design*

Privacy measures are introduced as the system is being developed, not added on later

4. Full *Functionality*

User experience vs security should not be a debate. Users must expect their data to be safe *and* be able to enjoy full use of the system

5. End-to-End *Security*

Personal data must be kept secure at all points within a system

6. *Visibility* and *Transparency*

Data owners must be allowed to see how information moves through your system

7. *Respect* for User *Privacy*

Keeping PII safe must be your main priority

multiverse

The Portfolio

When delivering a project write up you will need to evidence and include your data classification process.

This includes:

When delivering a project write up you will need to evidence and include your data classification process.

This includes:

- Describing data types, structures and sources

When delivering a project write up you will need to evidence and include your data classification process.

This includes:

- Describing data types, structures and sources
- Where it is stored and how it is accessed

When delivering a project write up you will need to evidence and include your data classification process.

This includes:

- Describing data types, structures and sources
- Where it is stored and how it is accessed
- What company policies you followed to ensure it was safe, clean and useable in accordance with GDPR

Your Portfolio

Your Portfolio

→ Serves as evidence of applying your skills

Your Portfolio

→ Serves as evidence of applying your skills

→ Will contain project write ups and reflective journals

Your Portfolio

→ Serves as evidence of applying your skills

→ Will contain project write ups and reflective journals

→ You can start building your portfolio once you start applying acquired skills

Your Portfolio

→ Serves as evidence of applying your skills

→ Will contain project write ups and reflective journals

→ You can start building your portfolio once you start applying acquired skills

→ Check sample portfolios and a project write up template

multiverse

Recap

Learning Objectives

- Identify **business specific rules** related to datasets and data characteristics that will influence project design and analysis
- Describe the key characteristics of the different **Data Formats** and how to work with them

ASSIGNMENT	
PART 1- DATA ANALYTICS LIFE CYCLE	
Use a work-related example to identify the stages of the Data Analytics Lifecycle. Describe what happened in each stage and highlight what was your role in the process. In the end, add a summary of the project/analysis including the main findings, what went well and what could have been improved.	
Word Count	Max 1500 words
Deadline	3 weeks
Deliverables	Word Document or PowerPoint presentation

ASSIGNMENT	
PART 2- PROJECT BRIEF	
Use a work-related example to create a project brief. This could be related to a project you are about to start or something new. Your brief should contain a business problem, the wider context of the analysis and a plan of action to solve the problem.	
Word Count	Max 1500 words
Deadline	4 weeks
Deliverables	Word Document

multiverse

Complete Session Attendance Log and Update Your OTJ