



`</talentlabs>`

CHAPTER 2

Data Preparation

Learning Objectives



Explain how data is prepared for analysis



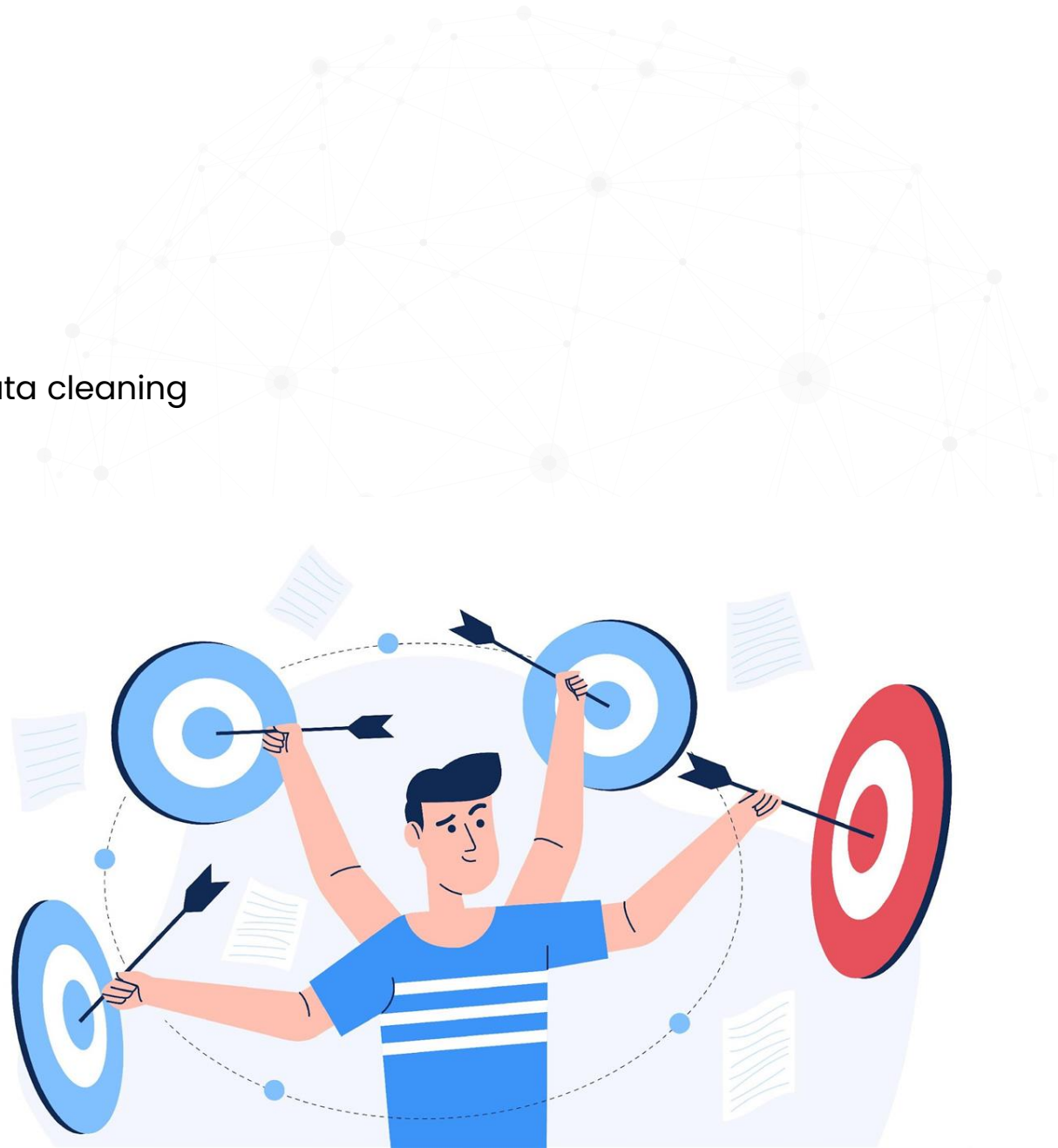
Describe the difference between data wrangling and data cleaning



Understand how to handle a new dataset



Appreciate the importance of verification and documentation when processing data



Agenda

- Data wrangling
- Cleaning and enriching data
- How to handle a new dataset
- Verification and documentation
- Chapter summary & assignment



Data Wrangling



The data wrangling workflow



Data wrangling tools



Typical Analytics Workflow

Analysts use data to **generate insights**.

01



Understand the problem
and the desired outcome

02



Data collection

03



Data wrangling

04



Data analysis and
visualisation

05



Documenting the process


06



Effectively communi-
cating the final report and
insights to stakeholders

Data wrangling is the process of transforming data from a 'raw' to a more usable form. During data wrangling, collected data is prepared for analysis.

Raw data



Name	Company Salary
Dr John Smith	"\$50,000"
Ms Jane Doe	45000
Mr Mike Bloggs	37000
Miss Claire Smith	42000

Data Wrangling Workflow



Inspect



Structure



Clean



Enrich



Validate



Document & Publish



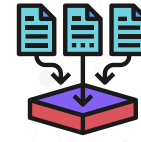
Data Wrangling Definitions



Data Cleaning vs Data Wrangling

Data cleaning: removing flawed data

Data wrangling: transforming data into a more usable form



Enriching data/Data Wrangling

Translating clean data into a more usable form.

Data Wrangling Workflow



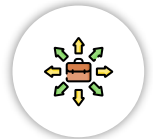
Inspect



Structure



Clean



Enrich (Wrangle)



Validate



Document & Publish



Data Wrangling Tools



Google Sheets

Spreadsheets

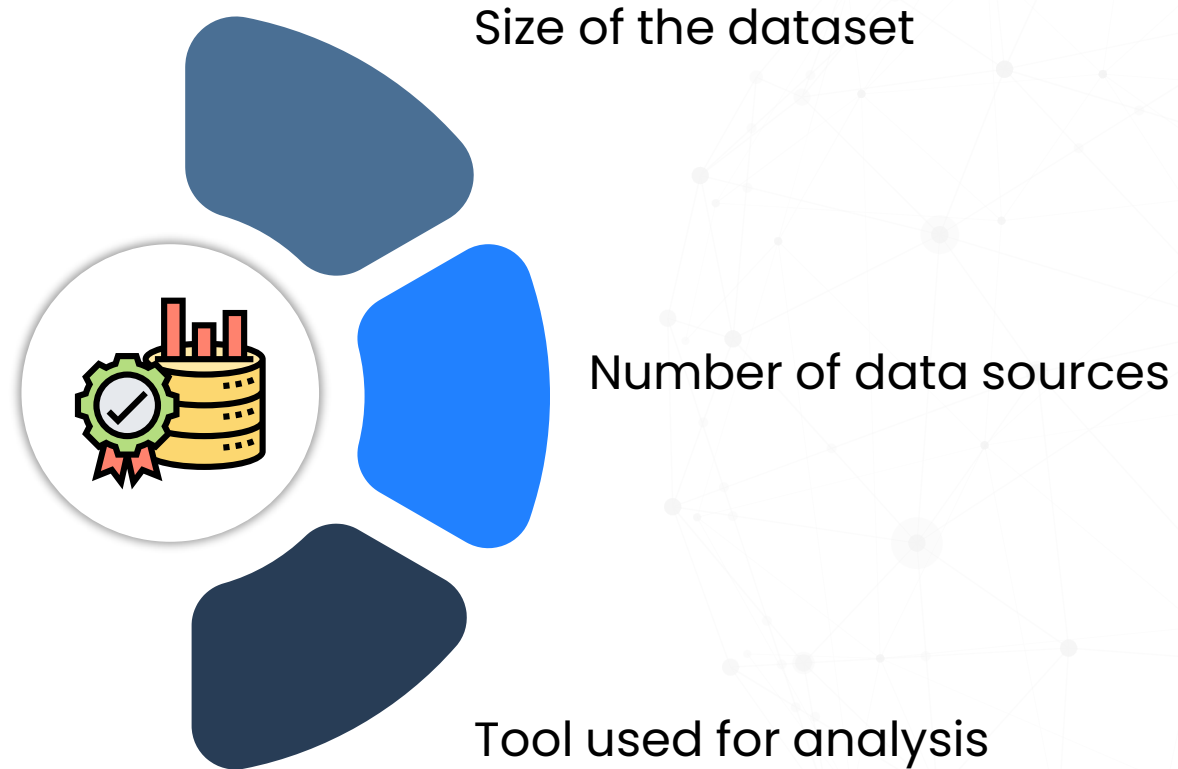


SQL



Python

Data Wrangling Tools



Cleaning and Enriching data



Dirty and clean data



Data cleaning workflow



Handling missing values



Data enrichment



Data Cleaning vs Data Wrangling



Data Cleaning

Removing flawed data



Data Wrangling

Transforming clean data into a more usable form

Dirty vs Clean Data

member_number	date	item_description
01808	21-07-2015	Whole milk
"2552"	05/01/2015	Tropical Fruit
2300	2015.12.12	Vegetables
3037	01/02/2015	tropical fruit

Member number	Date	Item Description
1808	21/07/2015	Whole milk
2552	05/01/2015	Tropical Fruit
2300	12/12/2015	Vegetables
3037	01/02/2015	Tropical fruit

Dirty data – the four Is

01

Incomplete

Member number	Date
1808	21/07/2015
	05/01/2015

02

Incorrect

Member number	Date	Item Description
1808	21/07/2015	Wholemilk
1808	21/07/2015	Wholemilk

03

Inconsistent

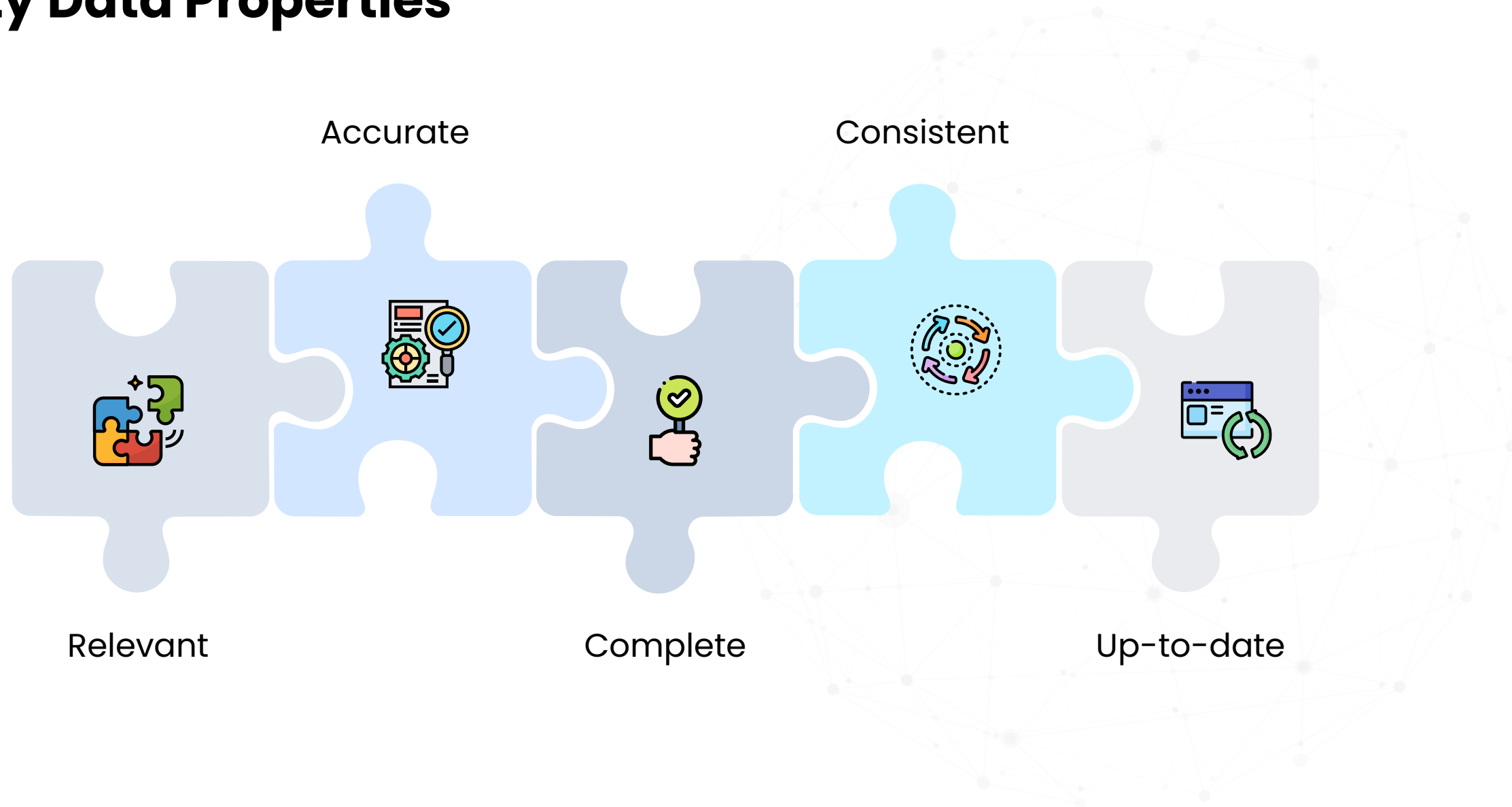
Member number	Date	Item Description
1808	07-21-2015	Whole milk
2552	05/01/2015	Whole milk

04

Irrelevant

Member number	Date	Item Description	Temperature (oC)
1808	21/07/2015	Whole milk	32
2552	05/01/2015	Tropical Fruit	23

Quality Data Properties



Benefits of Cleaning Data



Increases productivity



Improves insight reliability



Saves time and money



Allows for better decision making

Data Cleaning Workflow



Audit



Outline workflow



Implement workflow



Validate the quality



Report

Common issues

- Not fixing source of error
- Not backing up raw data
- Not allowing sufficient time for data cleaning

How to handle missing data?



Drop



Impute (fill in)



Flag

Depends on:

- Dataset size
- Percentage of missing values
- Type of missing data
- Reason for missing

Data Wrangling Workflow



Inspect



Structure



Clean



Enrich (Wrangle)



Validate



Document & Publish



Data Wrangling/Enrichment for Structured Data

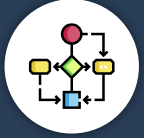
- Sorting columns
- Filtering columns
- Merging data from multiple sources
- Removing outliers
- Concatenating and splitting string columns
- Combining information into new columns

Name	Salary (\$)
Mr John Smith	45000
Ms Jane Doe	50000



Name	Salary (\$)	Gender	Salary Deviation from Mean (\$)
Mr John Smith	45000	Male	-2500
Ms Jane Doe	50000	Female	2500

How to handle a new dataset



Data wrangling workflow revisited



Organising the data



Data cleaning checklist



Handling a new dataset



Backup the raw data file



Organize the data



Inspect the data



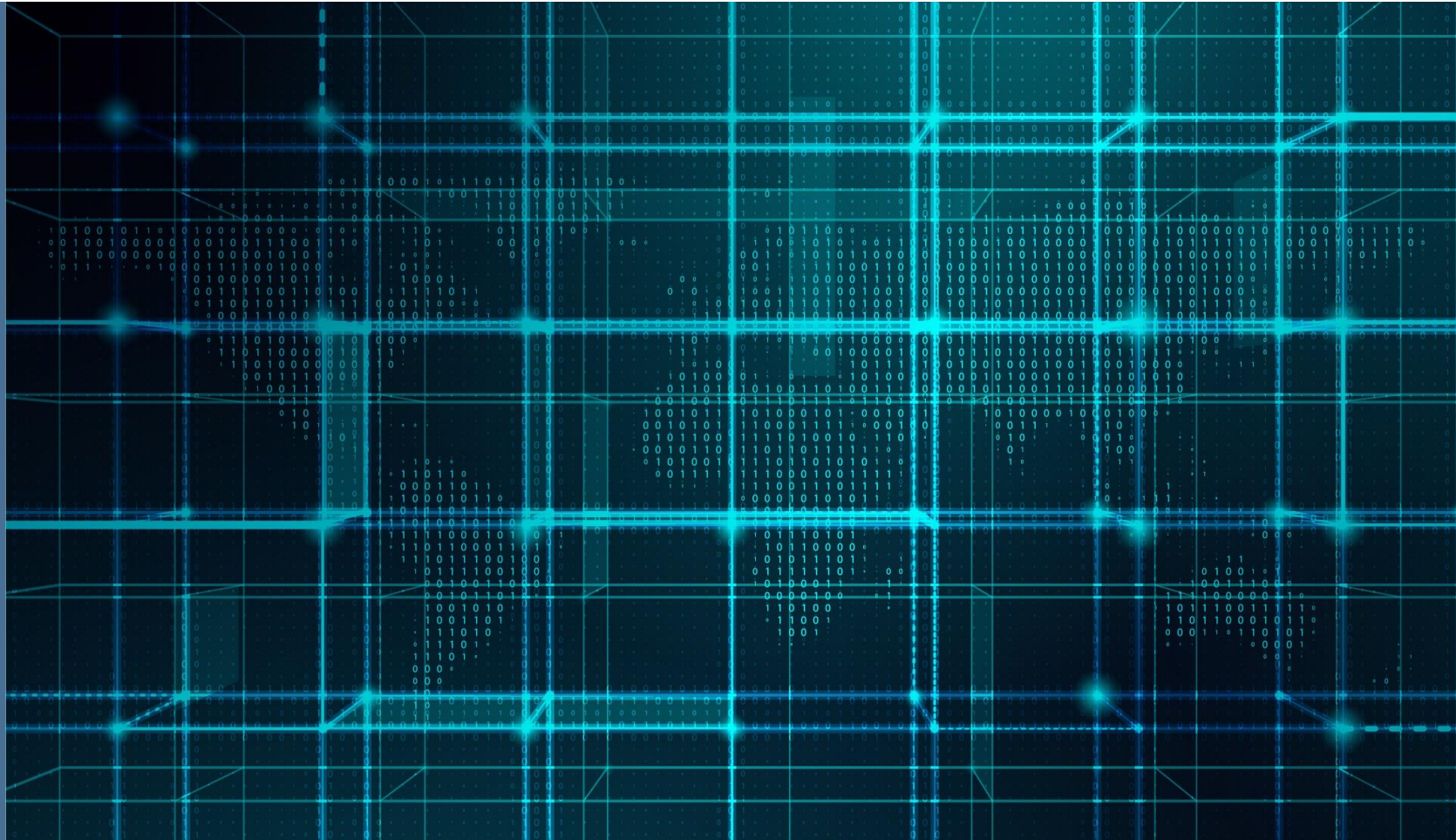
Clean the data



Enrich data as necessary



Verify the data



Organising the data – file structure

- Make a folder structure that is easy to navigate
- Give files and folders clear names
- Align naming and storage practices with your team/organization
- Separate ongoing and completed work










Organising the data – before inspection

- Remove unused and unnecessary data
- Merge data if needed
- Prepare general view (if in spreadsheets)



Data Cleaning Checklist

01		Irrelevant columns
02		Duplicates
03		Data types
04		Missing data
05		Inconsistent data entry
06		Errors/misspellings
07		Format



Verification and Documentation



Data validation










Documentation





Data validation involves making sure most common problems were identified and corrected.

Data Cleaning Checklist

01		Irrelevant columns
02		Duplicates
03		Data types
04		Missing data
05		Inconsistent data entry
06		Errors/misspellings
07		Format





Documentation involves creating a record of the changes made to the raw data during the data cleaning effort.

Documentation



Saves time



Reference for future users



Improves credibility

Make a record of:

- Changes made
- Reasoning behind decisions
- Date
- Person + approver
- Version number

Chapter Summary & Assignment



Data wrangling is the process of transforming data from a 'raw' to a more usable form. During data wrangling, collected data is prepared for analysis.

Data Wrangling Definitions



Data Cleaning vs Data Wrangling

Data cleaning: removing flawed data

Data wrangling: transforming data into a more usable form



Enriching data/Data Wrangling

Translating clean data into a more usable form.

Data Wrangling Workflow



Inspect



Structure



Clean



Enrich (Wrangle)



Validate



Document & Publish



Data Wrangling Tools



Google Sheets

Spreadsheets



SQL



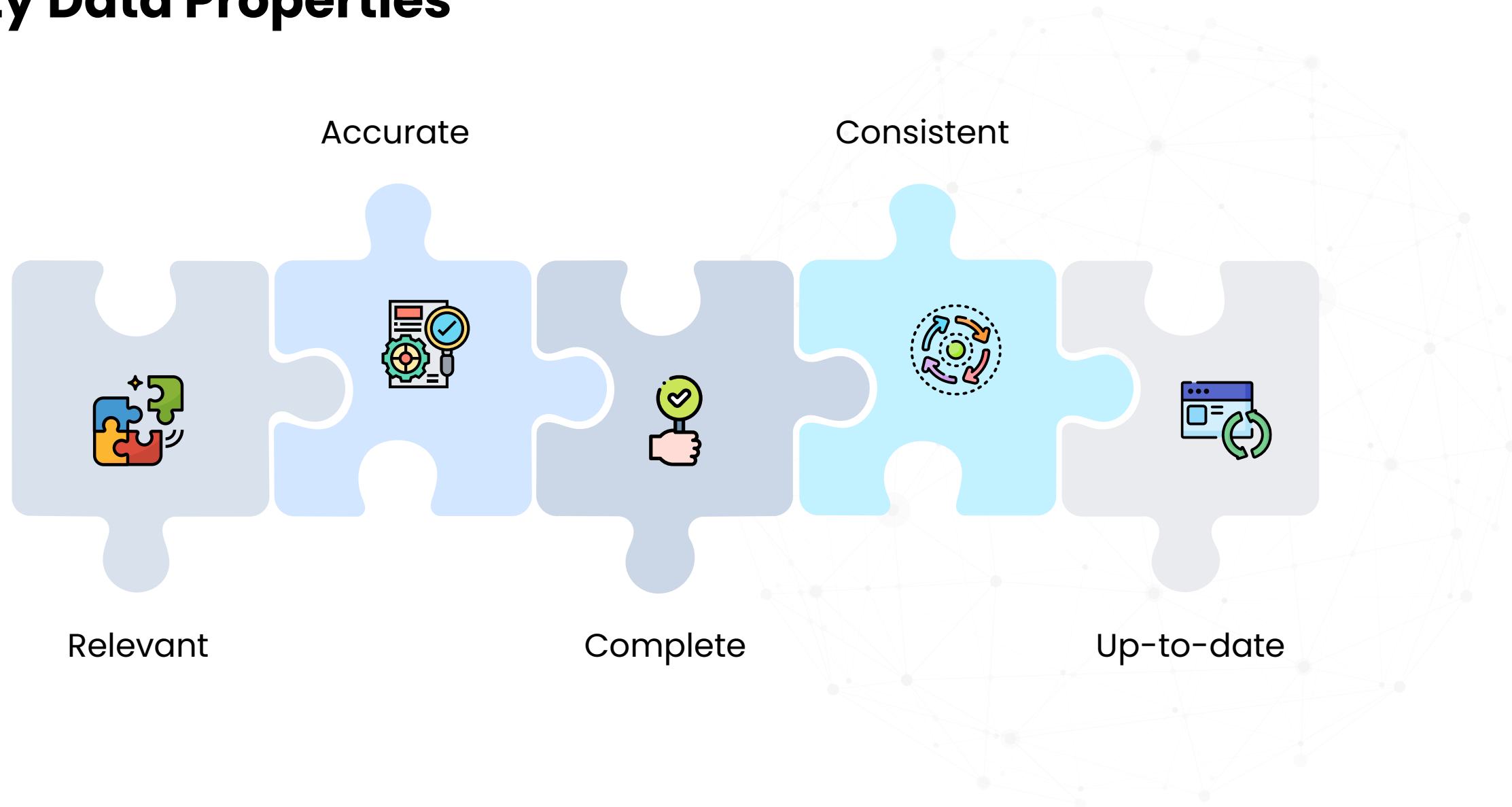
Python

Dirty vs Clean Data

Member number	Date	Item Description
01808	21-07-2015	Whole milk
"2552"	05/01/2015	Tropical Fruit
2300	2015.12.12	Vegetables
3037	01/02/2015	tropical fruit

Member number	Date	Item Description
1808	21/07/2015	Whole milk
2552	05/01/2015	Tropical Fruit
2300	12/12/2015	Vegetables
3037	01/02/2015	Tropical fruit

Quality Data Properties



Data Cleaning Workflow



Audit



Outline workflow



Implement workflow



Validate the quality



Report

Common issues

- Not fixing source of error
- Not backing up raw data
- Not allowing sufficient time for data cleaning

How to handle missing data?



Drop



Impute (fill in)



Flag

Depends on:

- Dataset size
- Percentage of missing values
- Type of missing data
- Reason for missing

Handling a new dataset



Backup the raw data file



Organize the data



Inspect the data



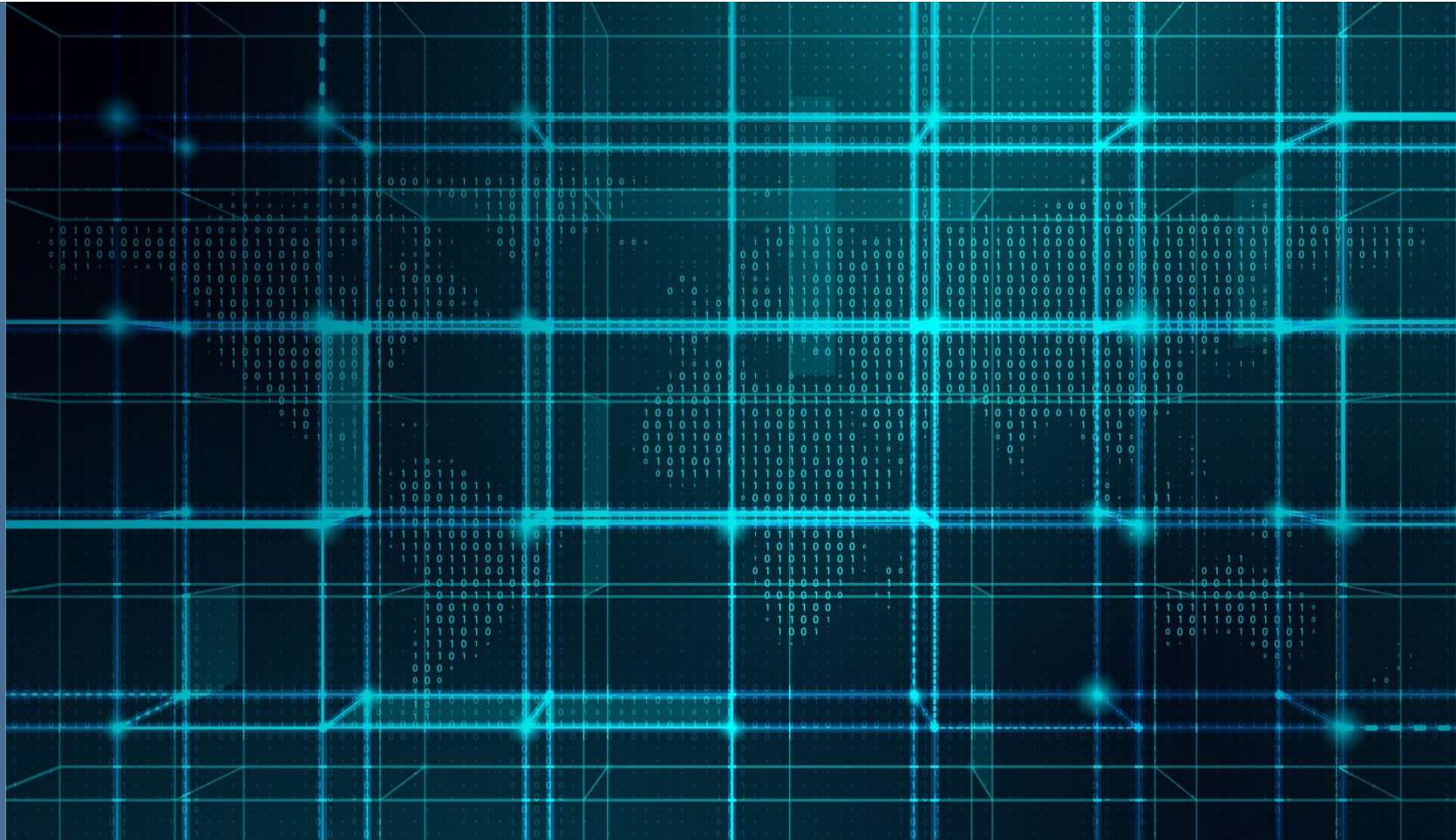
Clean the data










Enrich data as necessary



Verify the data



Data Cleaning Checklist

01		Irrelevant columns
02		Duplicates
03		Data types
04		Missing data
05		Inconsistent data entry
06		Errors/misspellings
07		Format





Data validation involves making sure most common problems were identified and corrected.



Documentation involves creating a record of the changes made to the raw data during the data cleaning effort.

Assignment Information

Multiple choice questions

