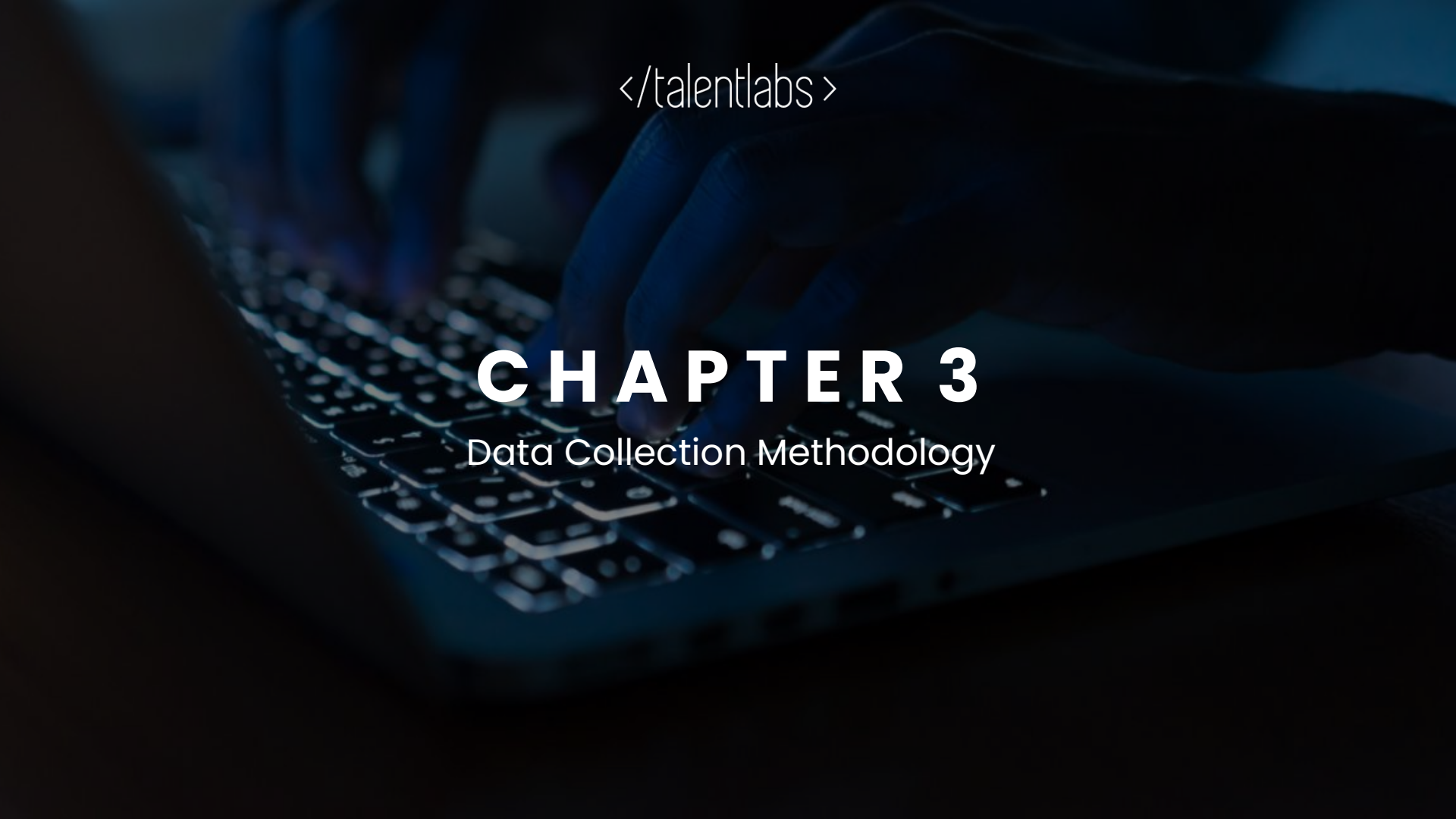

CHAPTER 3
Data Collection Methodology

Learning Objectives

- Describe the data collection workflow
- Understand how to define the problem and identify stakeholders
- Explain the tools and methods of data collection
- Appreciate the importance of documenting the process

Agenda

- What is data collection
- Understanding the problem and stakeholders
- Methods of data collection
- Data collection in real-time
- Documenting the process

</talentlabs>

What is data collection?

- Data collection definition
- Data collection within the analytics workflow
- Key steps of data collection

Data collection is the process of **gathering** data.

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

</talentlabs>

Data collection workflow

01 Define the problem (ask questions!)

02 Make a data collection plan

03 Collect the data

04 Document the process

</talentlabs>

Data collection workflow

What **type of data** do you require to achieve the aim?

What is the **aim** of the project?

What **methods** will be required to collect and store the data?

</talentlabs>

Understanding the problem and stakeholders

- Why it's important to understand the problem
- Stakeholder examples
- Asking good questions

Understanding the problem

What are you trying to solve?

Why?

Data the right tool?

Is data available?

What's in and out of scope?

Who are the stakeholders and

What are their expectations?

</talentlabs>

Stakeholders

Stakeholders are people who have invested **resources** and **time** into the project.

Stakeholder groups within the company

Analytics team

Customer-facing team

Executives team

</talentlabs>

Asking the right questions

- Asking questions is part of the process.
- Questions are the **foundation** for the rest of the project.
- When asking questions consider:
  - Relationships between variables
  - Domain
  - Scope

</talentlabs>

Examples of bad questions

Leading questions

Prompts the respondent towards giving a predetermined answer

Vague questions

Not precise or don't give context

Confusing questions

Difficult to understand

</talentlabs>

Examples of bad questions

Leading questions

This is our best product, isn't it? (instead – use neutral language)

Vague questions

Are we growing?

(instead – be precise)

Confusing questions

If we had to open a store in either America, Australia, Hong Kong or Germany where would you choose? Consider profit, marketing, historic sales data and opportunities.

(instead – split, simplify and reduce the content)

</talentlabs>

How to ask good questions - SMART

Specific – Focused

Measurable – Quantifiable and can be assessed

Action-focused – Inspire change

Relevant – Significant to the problem and important

Time-bound – Has a time range

</talentlabs>

How to ask good questions - SMART

Specific **M**easurable **A**ction-focused **R**elevant **T**ime-bound

Can we increase sales in cars?

Can we increase our **Ford F-Series sales** by **25%** by the **end of the year** **via Google Ads** and meet our **growth expectations**?

</talentlabs>

Methods of data collection

- Tools for data collection
- Primary and secondary data
- Methods of data collection
- Where to find data online

Tools for data collection

01

Google Sheets

Spreadsheets

02

Query Languages

</talentlabs>

Primary and Secondary Data Collection

Primary

Gathered by the researcher first-hand

- Quality, trustworthiness, understanding
- Time, cost

Secondary

Gathered by other researchers

- Time, cost
- Quality, trustworthiness, understanding

</talentlabs>

Data Collection Method Examples

Web scraping

Experiments

Simulations

Automated (websites)

Interviews

Government records

Reports and statements

Sensor (industrial machine)

Tracking (cookies)

Questionnaires

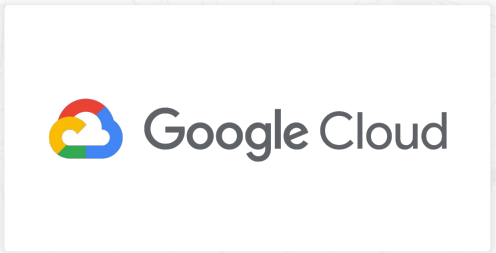
</talentlabs>

Data Collection Method Examples

Web scraping	Selenium and BeautifulSoup in Python
Automated	Google Analytics
Experiments	Laboratory + industrial experiments or research
Reports	Company financial records
Simulations	CFD analysis
Government records	Crime rates
Questionnaires	Online survey
Interviews	Speaking to experts
Sensors	Installing sensors on a car, smart watch

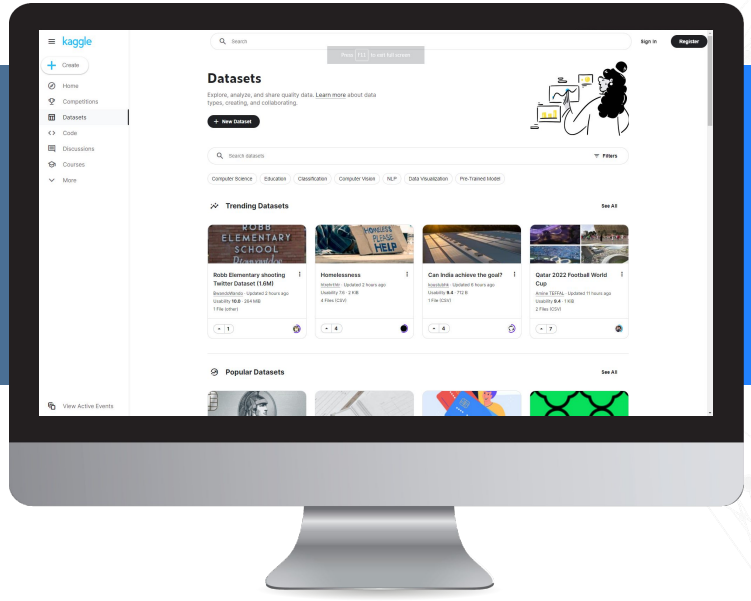
</talentlabs>

Where to find data online



</talentlabs>

Hands-on activity



Go and explore some public data repositories!
For example at:
https://www.kaggle.com/datasets

kaggle

</talentlabs>

Data collection in real time

- What is real-time data?
- Examples of real-time data
- Applications of real-time data

What is real-time data?

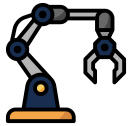
Real-time data is data that is **accessible as soon as its created** and collected.

Accessible for **live** decision making.

</talentlabs>

What is real-time data?

E-commerce purchases



Equipment sensor data

Weather data

Live streaming data

</talentlabs>

Applications of real-time data

- Retail
- Predictive Maintenance
- Healthcare

</talentlabs>

Documenting the process

Documentation

Why document data collection?

What does documenting involve?

Providing the information regarding the decisions made and the data collected during the process.

</talentlabs>

What does documenting involve?

Metadata

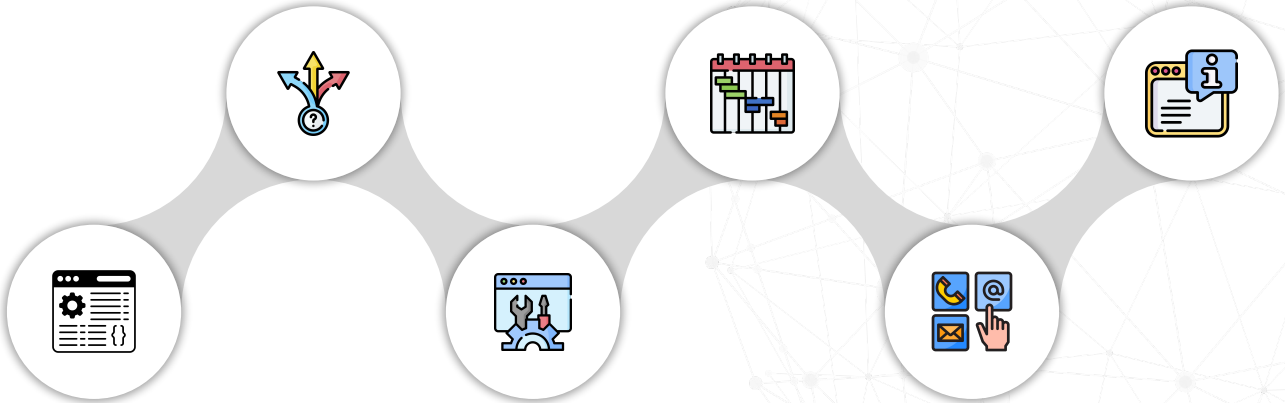
Explain decision making steps

Explain tools used throughout the process

Add dates and timelines

Add contact information for data collectors and project lead

Anything else you think might be useful to future data users



</talentlabs>

What does documenting involve?

- Context
- Usability
- Trust

Summary and assignment

Data collection workflow

01 Define the problem (ask questions!)

02 Make a data collection plan

03 Collect the data

04 Document the process

</talentlabs>

Data collection workflow

What **type of data** do you require to achieve the aim?

What is the **aim** of the project?

What **methods** will be required to collect and store the data?

</talentlabs>

Data Collection Method Examples

Web scraping	Selenium and BeautifulSoup in Python
Automated	Google Analytics
Experiments	Laboratory + industrial experiments or research
Reports	Company financial records
Simulations	CFD analysis
Government records	Crime rates
Questionnaires	Online survey
Interviews	Speaking to experts
Sensors	installing sensors on a car, smart watch

</talentlabs>

Stakeholders

Stakeholders are people who have invested **resources** and **time** into the project.

Understanding the problem

What are you trying to solve?

What + why?

Data the right tool?

Is data available?

What's in and out of scope?

Who are the stakeholders and

What are their expectations?

</talentlabs>

How to ask good questions - SMART

Specific **M**easurable **A**ction-focused **R**elevant **T**ime-bound

Can we increase sales in cars?

Can we increase our **Ford F-Series sales** by **25%** by the **end of the year** **via Google** Ads and meet our **growth expectations**?

</talentlabs>

Primary and Secondary Data Collection

Primary

Gathered by the researcher first-hand

- Quality, trustworthiness, understanding
- Time, cost

Secondary

Gathered by other researchers

- Time, cost
- Quality, trustworthiness, understanding

</talentlabs>

Applications of real-time data

- Retail
- Predictive Maintenance
- Healthcare

</talentlabs>

What does documenting involve?

Metadata

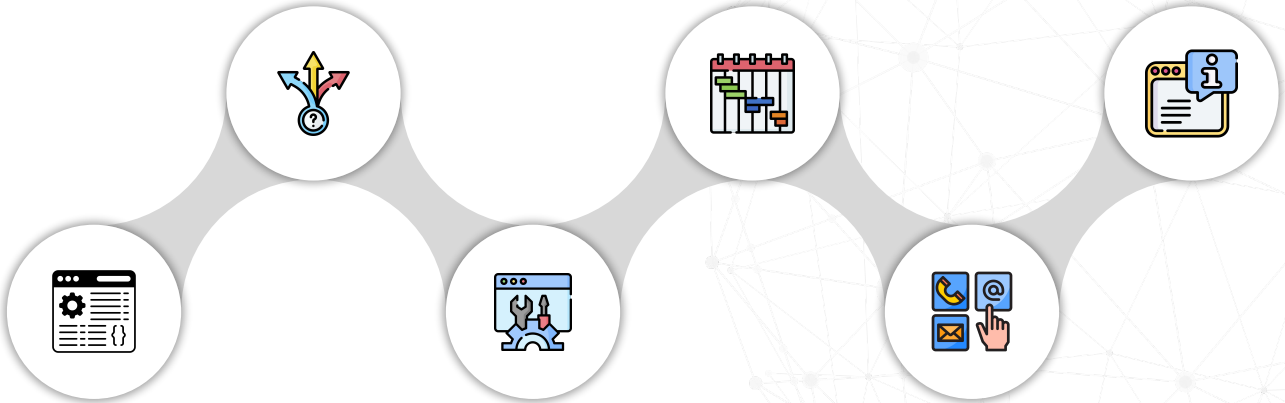
Explain decision making steps

Explain tools used throughout the process

Add dates and timelines

Add contact information for data collectors and project lead

Anything else you think might be useful to future data users



</talentlabs>

Assignment Information

Short-questions

Mini-project

</talentlabs>