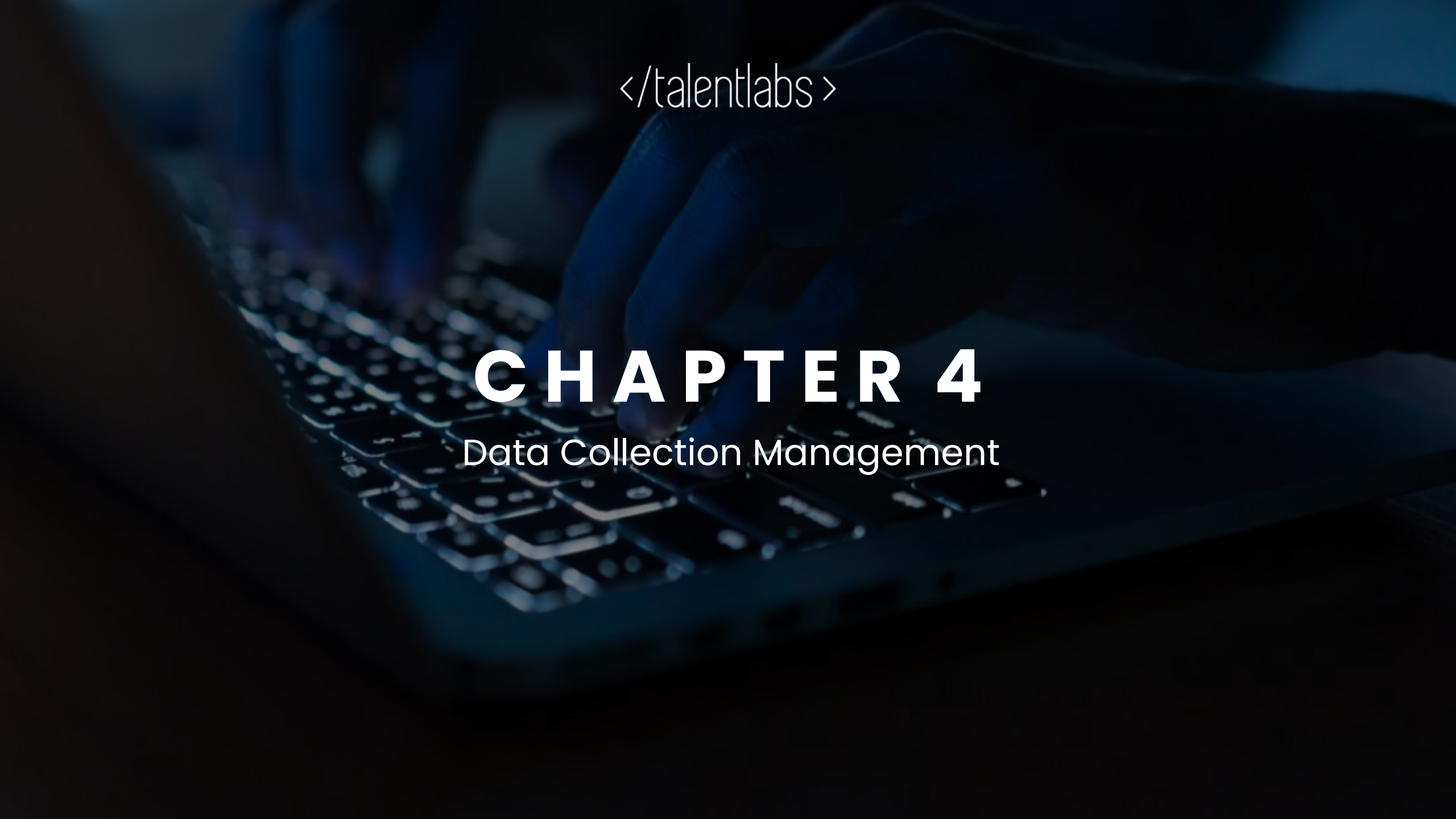


CHAPTER 4

Data Collection Management

Learning Objectives

- Describe what to include in a data collection plan
- Understand the challenges of data collection
- Consider the issues of sampling, bias, privacy and security

Agenda

- Data collection plan
- Sample size
- The issue of data bias
- Privacy and Security
- Quality assurance & best practice

</talentlabs>

Data collection plan

What to include in the data collection plan

Challenges of data collection

Data is like garbage. You'd better know what you are going to do with it before you collect it.

– ***Mark Twain***

</talentlabs>

Data Collection Plan

Know your problem

Define the data you need

Determine what data is available

Identify the source of the data

Explain how privacy will be maintained

Explain how the data will be collected

State how data will be stored and secured

Consider timescales

Explain how you accounted for sampling & bias

</talentlabs>

Challenges of data collection

Issues with data **quality**

Scope of the data – where do you draw the line?

Lack of data

Security

</talentlabs>

Data Collection Plan Summary

1. State the problem you are solving and the questions you are answering.
2. Identify the data you need to solve the problem and whether that data is already available.
3. Clearly state the scope of the data required.
4. Clearly state the timescale for the data collection process.
5. Describe where the data will be sourced from and how.
6. Identify who will collect the data.
7. Describe how the data will be stored and kept secure.
8. Explain any privacy concerns and how they will be mitigated.
9. Explain how bias and sampling have been accounted for.
10. Describe anything else that might be important to the data collection process.
11. State limitations and potential challenges and mitigation procedures (Plan B)

</talentlabs>

Sample size

What is a sample?

How to sample data?

What is a sample?

A sample is a **subset** of the whole **population**.

</talentlabs>

Why is sampling important?

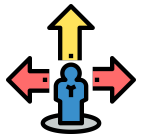
Consider the US presidential elections example

- A magazine sent out a poll to predict the winner of the US election in 1936 (the same magazine was successful in predicting the president winner in the previous election).
- **2.4 million** respondents replied and Franklin Roosevelt was favoured by **43%** of the people.
- The magazine announced that he was going to lose by a big margin.
- In reality he won by gaining **62%** of the vote. How did this happen?
- The poll included the readers of the magazine, car owners and people listed in the phone book. The magazine then went out of business (losing the subscribers to Time magazine).

</talentlabs>

How to sample data?

Ideally gather data on whole population! However this isn't always possible.



So, decide on

Sample size

Accessibility of the sample

Time frame

Keep in mind, the sample needs to be

Diverse

Consistent

Transparent

</talentlabs>

Determining the ideal sample size

Sample too large?

Increases cost, time and complexity.

Sample too small?

Outliers and lack of diversity.

Calculate the ideal sample size

1. **Population**
   Total number of people
2. **Margin of error**
   The accuracy range from whole population (%)
3. **Confidence level**
   How confident you are of the result (%)

If **20%** of your sample like to play basketball with a margin of error of **2%** and a confidence level of **95%** you are confident that in **19 out of 20 cases** the result would be between **18-22%.**

Now use a sample calculator online!

For example:

For a population of **1,000,000** people, with a margin of error of **2%** and a confidence level of **95%**

The ideal sample size is:
2,396 people

</talentlabs>

The issue of data bias

What is bias and fairness

Common types of bias

Understanding bias

Bias is either subconscious or conscious preference in favor of or against one group or person.

Bias can affect data in a negative way.

Fairness – ensuring data collection doesn't create bias.

Need a fair and inclusive data collection system.

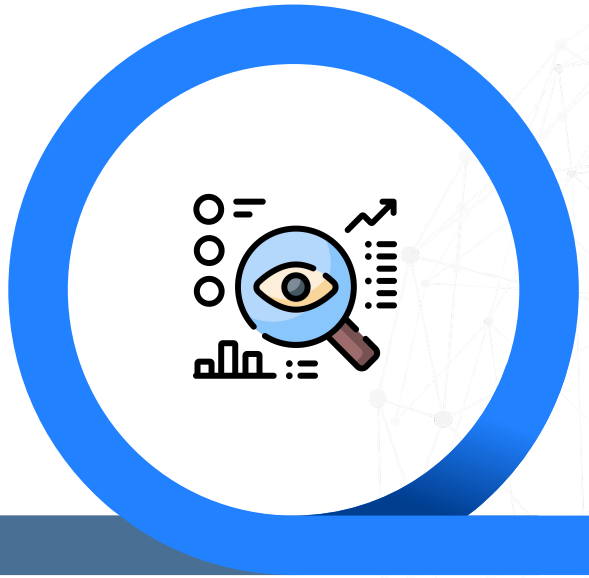
Not that simple!

</talentlabs>

Types of bias and what to avoid!

Sampling bias

Unrepresentative sample



Observer bias

Different people observing the same information differently

Confirmation bias

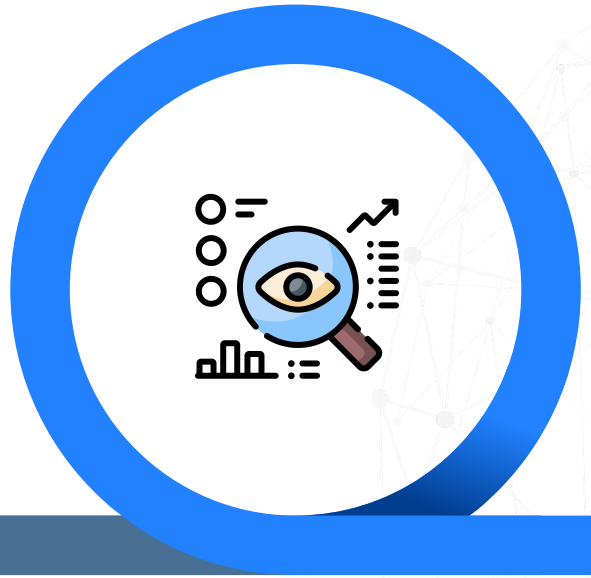
Analysing data in a way to confirm a previous hypothesis or conclusion

</talentlabs>

Types of bias examples

Sampling bias

A survey of university students to measure young adult use of illegal drugs – doesn't account for non-university adults



Observer bias

Scouts in sport – observing the same people with different perception

Confirmation bias

Researcher looking for data to confirm his experimental conclusions – not being open minded to faults in original hypotheses

</talentlabs>

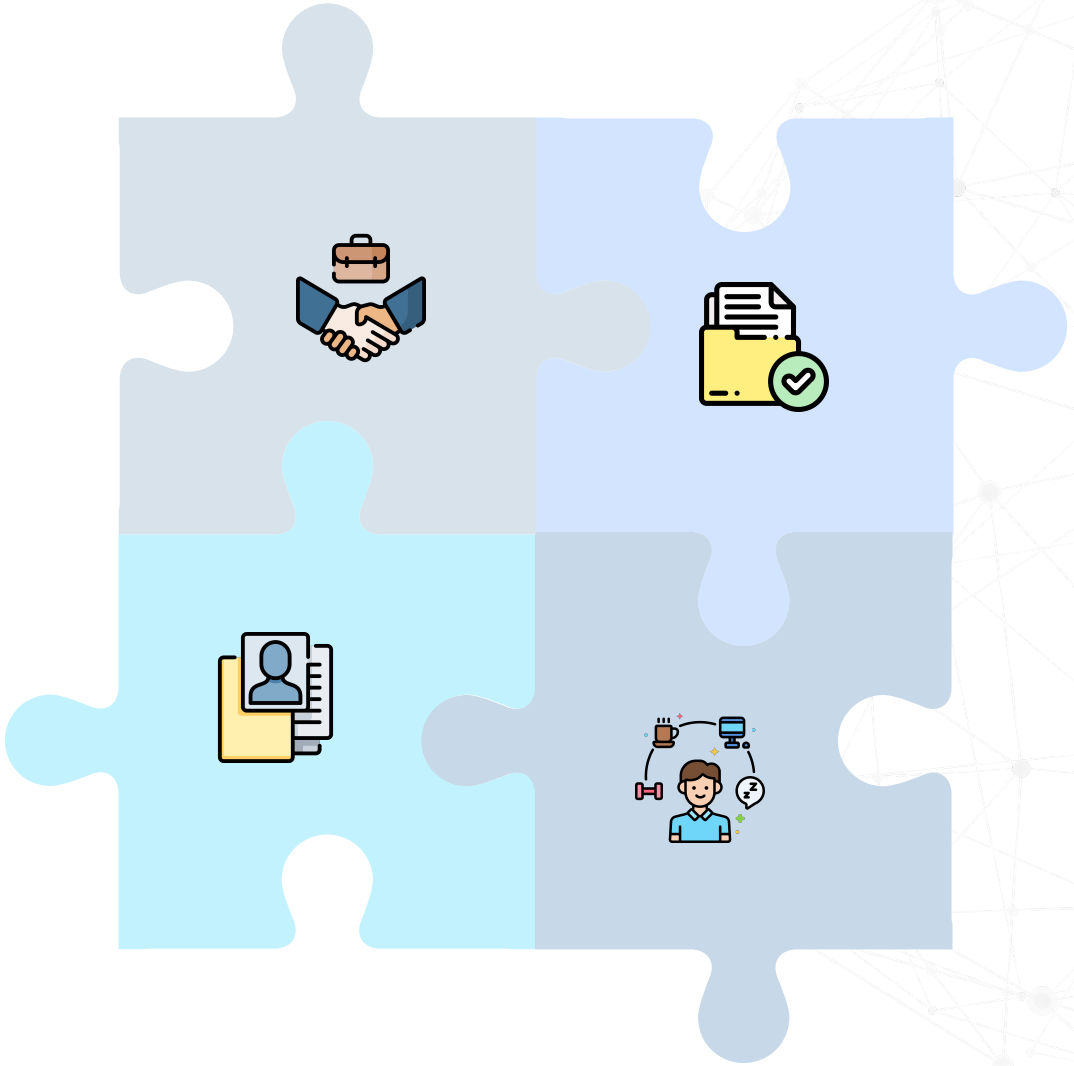
Privacy and security

- Data privacy
- Privacy risks
- Privacy strategies

Data Evolution

Businesses and governments unlocking more and more actionable insights from data.

Even without revealing personal information!



A result of:

- Big Data: velocity, volume, variety
- Advancement in analytics: machine learning, computational power, unstructured data processing

Data can now infer:

- Habits
- Lifestyle
- Social network

</talentlabs>

Privacy

Data protection:

Users have legal right to their data

Organisations must put measures in place to protect individual data.

</talentlabs>

Privacy Risks

Unauthorised access

Data linkage

- Linking unconnected data to profile an individual
- IP address to link with social media
- Postal code to give away location

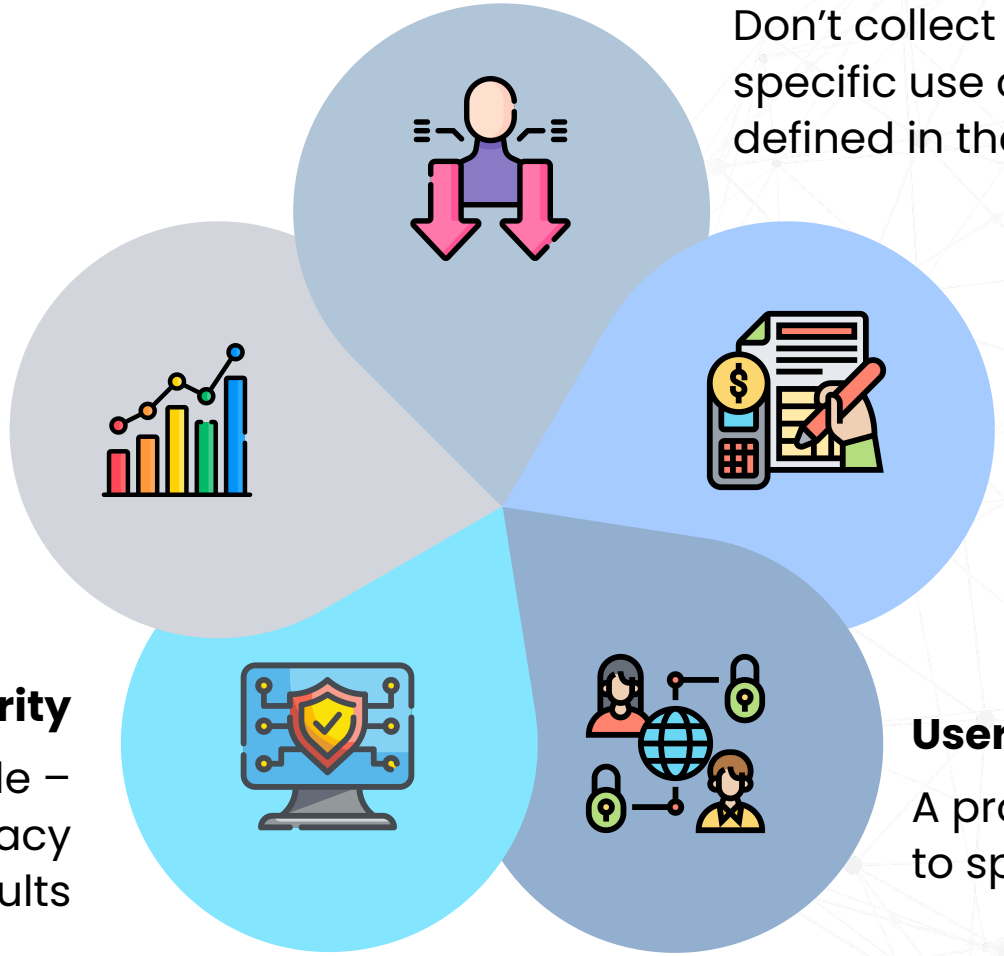
Secondary data use

- Individual needs to give explicit consent

Privacy Strategies

Minimising data

Don’t collect private data unless there is a specific use case, the purpose of which is well defined in the data collection plan



Synthetic data

With large enough datasets, can afford to create fictitious individuals

Data anonymisation

Make personally identifiable information anonymous. Such as: financial data, medical records and IP addresses.

Security

Applied throughout data lifecycle – ideally offering strong privacy defaults

User access

A process which allows users access to specific information

Quality assurance & best practice

Data quality

Data collection best practice

Data quality

Make sure data collection is accurate and consistent!

Common errors

- Date formats – dd/mm/yyyy **vs** mm/dd/yyyy **vs** dd/mm/yy
- Units
- Duplicated data collection
- Description discrepancies (white shark **vs** shark **vs** great white)

</talentlabs>

Data quality

Neighborhood	Block	Address	Gross square size	Year built	Sale price	Sale date
Jamaica	9755	89-21 153rd Street	51000 feet	1936	11000000USD	08-31-2017
Flatiron	851	45 EAST 22ND STREET	-	0	10496435USD	08-31-2017
FLATIRON	851	45 EAST 22ND STREET	-	0	10220998USD	08-31-2017
JAMAICA	9762	88-22 PARSONS BOULEVARD	8500 metres	1928	9000000USD	31-08-17
MIDTOWN EAST	1344	305 EAST 51ST STREET	-	2007	$5400000	31/08/2017

</talentlabs>

Data collection best practice

Make data easily accessible and easy to understand!

Revisit your plan and constantly evaluate it!

Don’t forget about privacy, bias and sampling!

</talentlabs>

Summary and Assignment

Data Collection Plan Summary

1. State the problem you are solving and the questions you are answering.
2. Identify the data you need to solve the problem and whether that data is already available.
3. Clearly state the scope of the data required.
4. Clearly state the timescale for the data collection process.
5. Describe where the data will be sourced from and how.
6. Identify who will collect the data.
7. Describe how the data will be stored and kept secure.
8. Explain any privacy concerns and how they will be mitigated.
9. Explain how bias and sampling have been accounted for.
10. Describe anything else that might be important to the data collection process.
11. State limitations and potential challenges and mitigation procedures (Plan B)

</talentlabs>

Why is sampling important?

Consider the US presidential elections example

- A magazine sent out a poll to predict the winner of the US election in 1936 (the same magazine was successful in predicting the president winner in the previous election).
- **2.4 million** respondents replied and Franklin Roosevelt was favoured by **43%** of the people.
- The magazine announced that he was going to lose by a big margin.
- In reality he won by gaining **62%** of the vote. How did this happen?
- The poll included the readers of the magazine, car owners and people listed in the phone book. The magazine then went out of business (losing the subscribers to Time magazine).

</talentlabs>

Determining the ideal sample size

Sample too large?

Increase cost, time and complexity.

Sample too small?

Outliers and lack of diversity.

Calculate the ideal sample size

1. **Population**
   Total number of people
2. **Margin of error**
   The accuracy range from whole population (%)
3. **Confidence level**
   How confident you are of the result (%)

If **20%** of your sample like to play basketball with a margin of error of **2%** and a confidence level of **95%** you are confident that in **19 out of 20 cases** the result would be between **18-22%.**

Now use a sample calculator online!

For example:

For a population of **1,000,000** people, with a margin of error of **2%** and a confidence level of **95%**

The ideal sample size is:
2,396 people

</talentlabs>

What is bias?

Bias is either subconscious or conscious preference in favor of or against one group or person.

Bias can affect data in a negative way.

Fairness – ensuring data collection doesn't create bias.

Need a fair and inclusive data collection system.

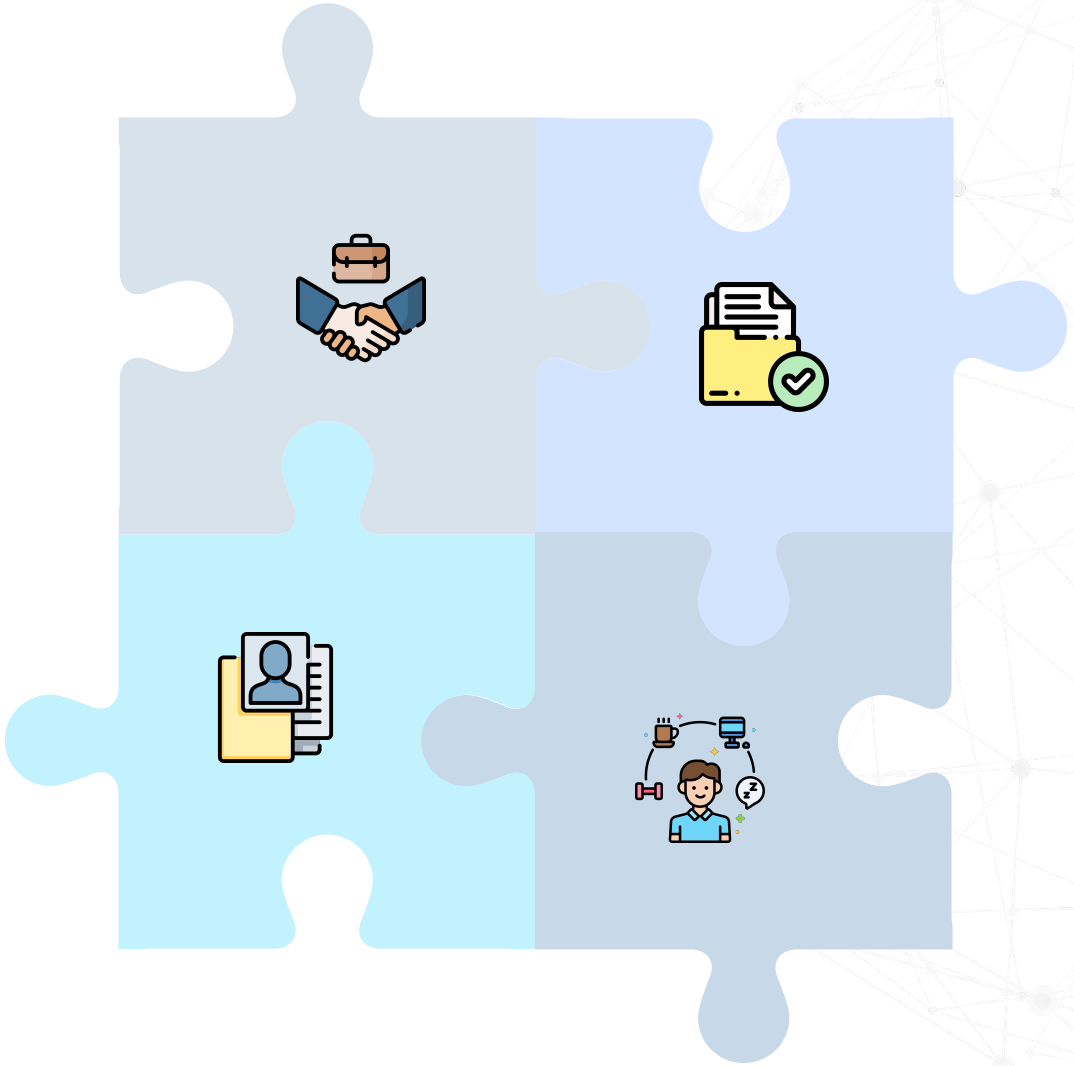
Not that simple!

</talentlabs>

Data Evolution

Businesses and governments unlocking more and more actionable insights from data.

Even without revealing personal information!



A result of:

- Big Data: velocity, volume, variety
- Advancement in analytics: machine learning, computational power, unstructured data processing

Data can now infer:

- Habits
- Lifestyle
- Social network

</talentlabs>

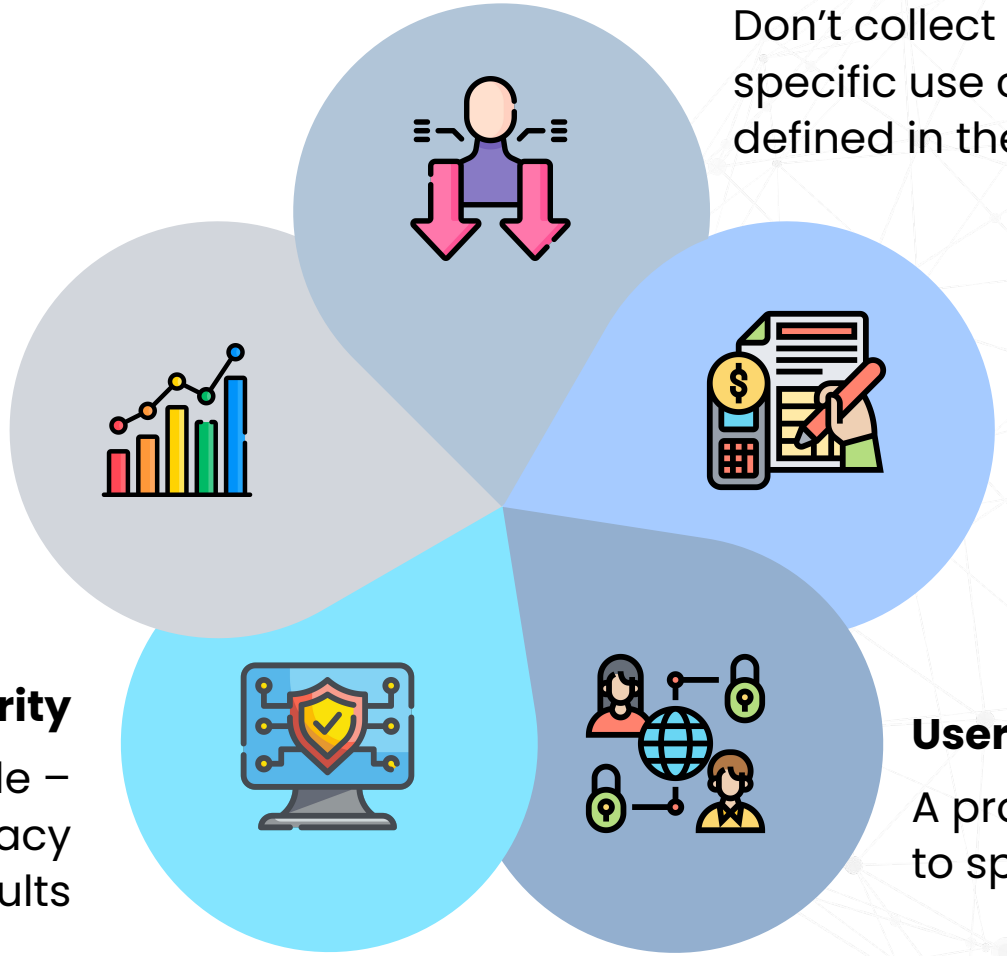
Privacy Strategies

Minimising data

Don't collect private data unless there is a specific use case, the purpose of which is well defined in the data collection plan

Synthetic data

With large enough datasets, can afford to create fictitious individuals



Data anonymisation

Make personally identifiable information anonymous. Such as: financial data, medical records and IP addresses.

Security

Applied throughout data lifecycle – ideally offering strong privacy defaults

User access

A process which allows users access to specific information

Data quality

Make sure data collection is accurate and consistent!

Common errors

- Date formats – dd/mm/yyyy **vs** mm/dd/yyyy **vs** dd/mm/yy
- Units
- Duplicated data collection
- Description discrepancies (white shark **vs** shark **vs** great white)

</talentlabs>

Data collection best practice

Make data easily accessible and easy to understand!

Revisit your plan and constantly evaluate it!

Don’t forget about privacy, bias and sampling!

</talentlabs>

Assignment Information

Multiple choice questions

Data Collection Plan

</talentlabs>