EXPLORE AI ACADEMY

Data sources and access

Sources of data

There is no analysis without data

DATA is at the heart of **data analysis**.

DATA COLLECTION is the process of gathering relevant **data** from a variety of useful sources.

DATA ANALYSIS is all about uncovering useful insights and trends from **data**.

There is no business value without data analysis

We create...

...when suitable data is **COLLECTED**,

...the **DATA** is managed, and

... **DATA ANALYSIS** is used to solve business problems.

Where to source your data

Readily available data

Company data

Data that a company **collects** and **stores** in a storage location with limited access.

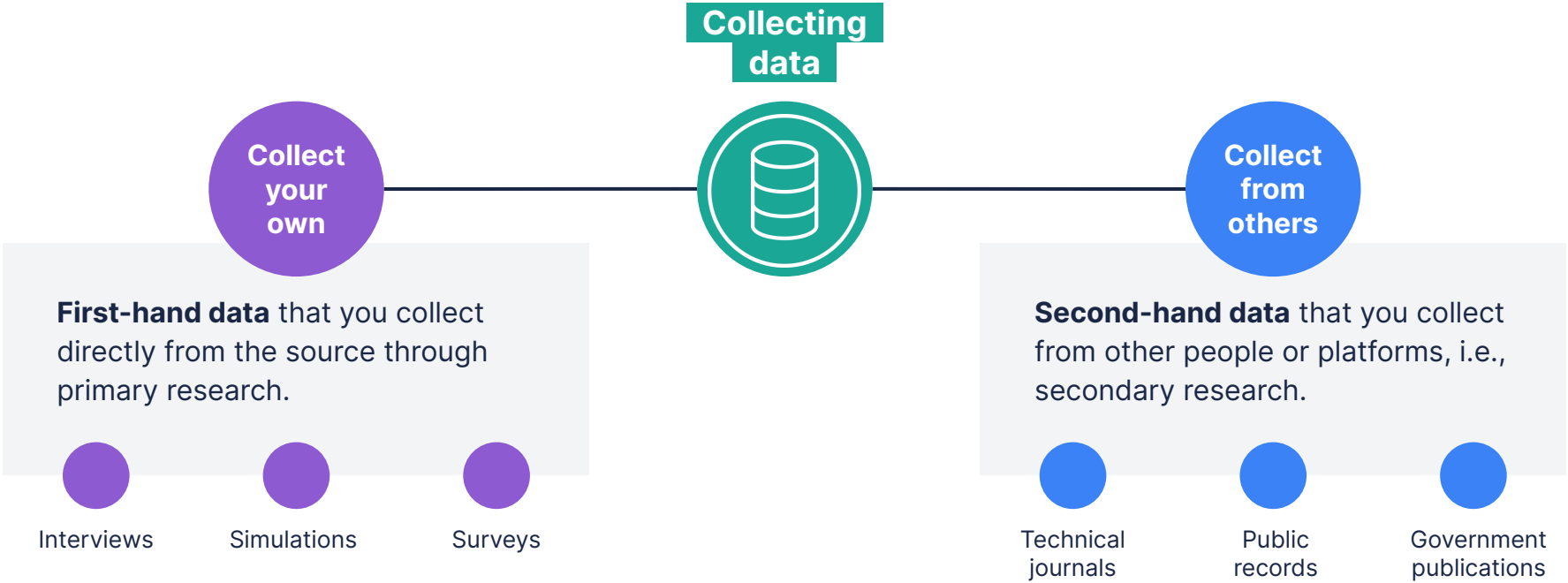
- Sales data
- Financial transactions
- Customer data

Open source data

Public records by government and non-government organisations.

- Medical data
- Census data
- Weather data

Where to source your data



Evaluate your data sources

> Despite the vast amount of available data, not all of it can be trusted. We need to assess the **strengths**, **limitations**, or any **biases** associated with a **data source**.

Use the **C.R.A.A.P test** to evaluate our data sources:

Currency

How timely is the data? Is it current and with regular updates, or is it outdated?

Relevance

How applicable is the data to our needs?

Authority

Who is the source of the data, and what **credentials** do they hold?

Accuracy

How truthful or **correct** is the data? Can we verify the data from other sources or our knowledge?

Purpose

What is the data **intended** for? Are there **possible biases**?

Understand the data you need

> Data is the link between the problem you are trying to solve and the intended solution. Ask the right questions using the 5W2H (what, why, where, when, who, how, how much) method to define the problem and understand the data that you need.

What data are you looking for?

- **Relevant** – data that pertain to the problem at hand and are in the required format.
- **Dependable** – data that are well structured, correct, and consistent.

NB: It **can't be 100% perfect**. It is all about finding a balance between the ideal and what is available.

What do you want to avoid?

- **Wrong data** – your analysis will be pointless (GIGO: garbage in, garbage out).
- **Too little data** – not enough to give you desired results.
- **Too much data** – cumbersome to work with and can cause time wastage.

Data in the real world

Even if we understand what we need from our data and that we can trust it, data in the real world is...

messy

+

difficult to work with

There are several ways to validate and convert real-world data into more **usable** and **useful** forms.

Real-world data is messy

> We collect raw data in different ways and from a variety of sources; it is likely to be **messy** and **unclean**, which makes it pretty useless.

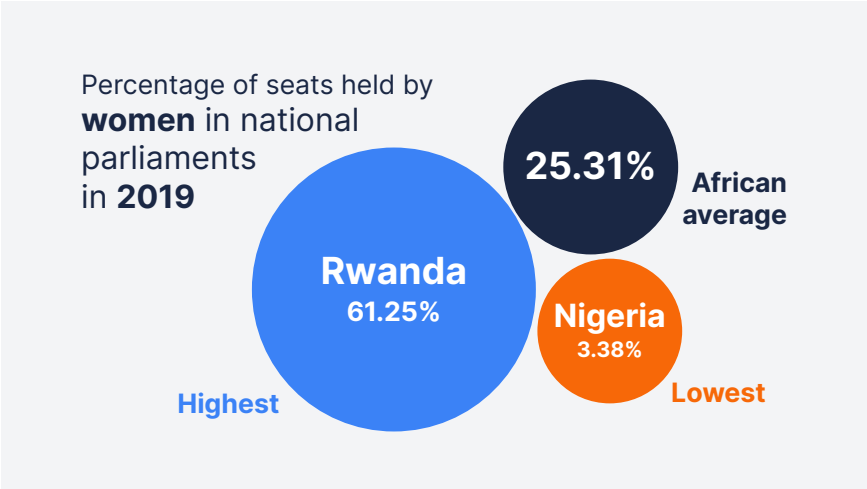
Unstructured data

Country KEN Year 2019 Proportion of seats 0.21776504298
Country South Africa Year 2019 Proportion of seats 0.46347607053
Country Seychelles Year 2019 Proportion of seats 0.21212121212
Country Sierra Leone Year 2019 Proportion of seats 0.12328767123
Country Nigeria Year 2019 Proportion of seats 0.03380281691

	Country	Year	Proportion of seats held by women in national parliaments
Inconsistencies	KEN	2019	0.21776504298
	South Africa	2019	0.46347607053
	Seychelles	2019	0.21212121212
Missing values	Sierra Leone		0.12328767123
	Nigeria	2019	0.03380281691
Erroneous values	Rwanda	2019	o.6125
Duplicates	Nigeria	2019	0.03380281691

Real-world data is messy

> We clean this messy data to convert it into **useful, structured data** that can be used for **analysis**.



Country	Year	Proportion of seats held by women in national parliaments
Kenya	2019	0.21776504298
South Africa	2019	0.46347607053
Seychelles	2019	0.21212121212
Sierra Leone	2019	0.12328767123
Nigeria	2019	0.03380281691
Rwanda	2019	0.6125

Data validation

> Data validation is the process of checking the accuracy and quality of data before use. We can put several checks in place that define the rules and constraints against which ingested data are compared.

Data type

It's of the **correct data type**, for example, numeric.

Format

It follows a **predefined format**, for example, a date format.

Length

It's of the **appropriate length**, for example, a phone number.

Consistent

It follows a consistent **logical order**, for example, check-in and check-out dates.

Range

It falls within a **specified range**, for example, dates within a month.

Uniqueness

Unique entries have **not** been **duplicated**, for example, IDs.

Presence

No entries have been left **blank**, especially for mandatory fields.

Look up

It conforms to a **set of acceptable values**, for example, days of the week.

Get started with what you have!

Data gathering is an **iterative process**.