EXPLORE AI ACADEMY

Ways of work

EGAD: An introduction to GATHER

Ways of work

The EGAD framework

> The EGAD framework provides us with a guide on how to **leverage data to solve problems**. It includes everything from understanding the problem to maintaining the solution.

	Draft	Do	Deliver	Decompress
EXPLAIN	Problem statement	Storytelling	Communication	Feedback
GATHER	Problem landscape	Databases	Data engineering	Insights
ANALYSE	Equation of value	Programming	Solution governance	Performance metrics
DEPLOY	Project management	Version control	Production	Maintenance

Ways of work

GATHER

> GATHER in the EGAD framework focuses on how we **collect**, **transform**, and **store data**, and how we **translate** findings **into insights**.

	DRAFT	DO	DELIVER	DECOMPRESS
GATHER	**Problem landscape** If we **understand** the **landscape** of the **problem** we are trying to solve,	**Databases** we can **collect** and **store** it in appropriate **formats** and **storage mediums**,	**Data engineering** that will allow us to **clean** and **wrangle** the data,	**Insights** to transform it into **information** that gives data **context**, **meaning**, and **purpose**.

Ways of work

GATHER, DRAFT

> The first phase of **GATHER**, **DRAFT**, ensures that we have everything we need before trying to solve a problem.

Once we have an appropriate problem statement, we need to GATHER the data, information, and knowledge we need:

- We have to have access to and **GATHER data** that will ensure success.
- **GATHER information** about the project and its past. *(What was tried in the past and how successful was it.)*
- Make sure we are able to **GATHER knowledge** that will help in the execution. *(The right people and the right tools.)*

What else can we GATHER on the problem landscape?

Data: Who owns and signs it off? Where is it? Is it structured or unstructured?

Information: What does the data dictionary say? How is the data gathered and updated?

Knowledge: Have we solved a similar problem before? Do we need a subject matter expert?

Ways of work

Gathering data in the real world

There's no such thing as a **perfect** dataset. When gathering data, the process often involves **checking** its **accuracy**, **cleaning** anomalies, and **formatting** it appropriately.

The challenge lies in **bridging** the gap between the **available data** and the **specific requirements** we have in mind.

What do we do with imperfect data?

The Pareto principle states that roughly **80% of the effects come from 20% of the causes**. For data professionals, this means quickly understanding the **20% of data** that accounts for **80% of the results**.

The process:

1. Use what data is **available** and **get started**.
2. Do **descriptive analytics** and build a **model**.
3. Analyse our **results** to determine what **more** data we might need to continue.
4. **Repeat** the process with additional data until we have our desired result.

What is involved in gathering data?

Processes involved

- **Finding data**, e.g. using web scraping to extract data.
- **Creating data**, e.g. collecting or transforming data.
- **Storing data**, e.g. using AWS to maintain databases.
- **Managing data**, e.g. backing up and granting access to data.

A continual process

- We need data to **solve problems**, so we gather it after we have specified the problem.
- A continual feedback loop exists between **E-G-A-D** (and we never stop gathering data).

Relevance

- It's impossible to do data science without **good-quality data** – garbage in, garbage out!
- Data needs to be in the correct **format** in order to **analyse** and **visualise** it.

Where do we get data?

> Getting data is a critical part of data science. Sometimes we get lucky and the data are **already available**, and other times we need to **collect our own** data.

Using other people's data

Open data sources, for example, Stats SA, UCT's Data Portal, City of Cape Town, The World Bank.

Proprietary data sources, for example, industry datasets or company-specific datasets.

Note: You should not share any proprietary data without written consent from the source. You also need to be aware of regulations like the POPI Act.

Collecting your own data

Primary research, including surveys, interviews, and simulating data.

Collect other people's data, for example, use web scraping to pull data off websites, APIs to pull data off systems and specific applications, or capture data electronically that used to be on paper.

Ways of work

GATHER, DO

> The **DO** phase of **GATHER** focuses on understanding what **data is**, how to **store it**, and **how to query** and **transform** it.

There are **multiple mediums** for **storing** data and these are constantly changing and improving with time.

As data professionals, the fundamentals of **databases** are imperative to understand.

Access to the **cloud** allows data professionals to execute **larger data transformations** almost in **real time**.

Old school: From “data” written on rock and clay tables we moved to writing and typing on paper and storing it in filing cabinets.

Local storage: Data stored physically on a local computer, external drive, or on a server in a database or in a file system.

Cloud storage: We are now starting to store and transform data in the cloud, e.g. Amazon Web Services, Microsoft Azure, or Google Cloud.

Ways of work

GATHER, DELIVER to DECOMPRESS

> The **DELIVER** phase of **GATHER** focuses on the **systems** and **infrastructure** we need to **transform massive amounts of data into valuable insights**.

Data engineering is not only a profession but a practice of designing, constructing, and maintaining **systems**, **architectures**, and **pipelines** that allow for the **collection**, **storage**, and **processing of data**.

Data engineering lays the **foundation** for **effective data analysis**, **machine learning**, and **business intelligence** by ensuring that data are accessible, accurate, and ready for exploration.

Ways of work

The importance of GATHER

Every dataset, irrespective of its source, will have **imperfections** and will require thorough **cleaning** and **preprocessing**.

We need data to solve problems, so we gather it **after** we have specified **the problem**.

It is a **continual process** that encompasses sourcing, curating, storing, and managing data to maintain **quality** and **readiness for analysis**.

Throughout this course, you'll **learn** various **skills** and **tools** you can apply in the DRAFT, DO, DELIVER, DECOMPRESS phases of many different **projects** to solve real **problems**.