
The Data Lifecycle

A data science perspective

Richard Djarbeng

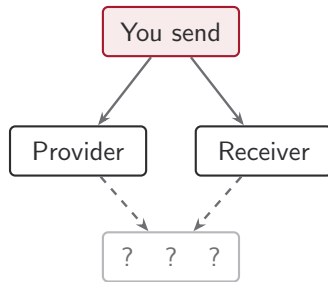
Where is your data right now?

This morning you sent GH¢50 by **Mobile Money**. That transfer became a record: amount, time, sender, receiver.

Q1. Who now holds a copy of that record?

Q2. Will it still exist in 5 years? Who decides?

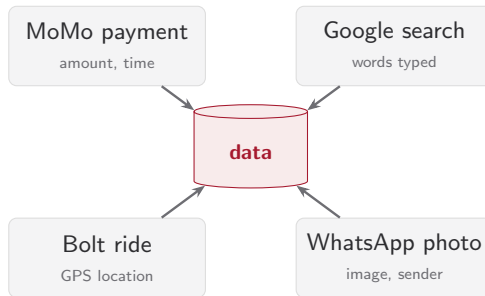
Answer in the chat.



First: what is data?

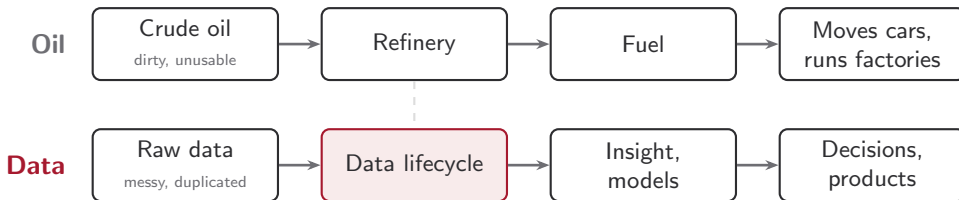
Data is any recorded fact about the world: a number, a word, a time, a location, an image.

Today almost everything we do leaves a record. Most of it is created by our phones, without us typing anything.



Data is the new oil

Valuable, but only after it is refined



The data lifecycle is the **refinery**.

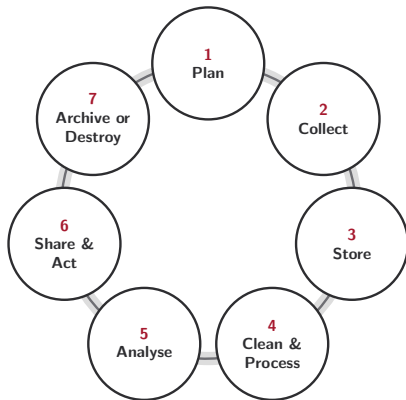
Phrase attributed to mathematician Clive Humby, 2006 · *The Economist*, "The world's most valuable resource is no longer oil, but data", May 2017.

What is the data lifecycle?

From a data science perspective

In data science, the **data lifecycle** is the path data takes from a **question** we want answered, to a **decision**, and finally to deletion.

Each stage has its own question:
what to record, where to keep it, whether it is correct, who may use it, and when to delete it.



Why study it?

Errors travel

A wrong value recorded at stage 2 is still wrong when a manager uses it at stage 6.

The law applies

Ghana's Data Protection Act (Act 843) covers personal data at every stage.

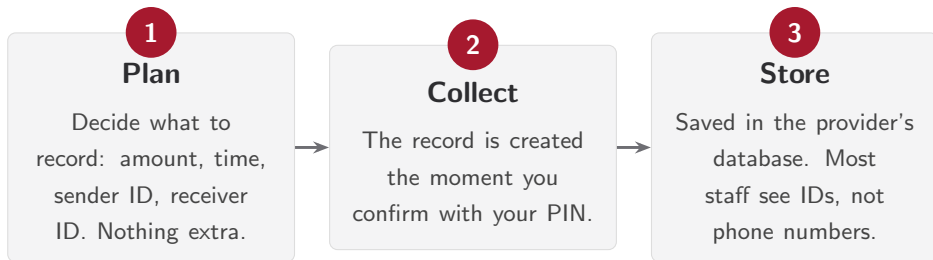
Models inherit it

A model learns from whatever data it is given, including its mistakes.

Data Protection Act, 2012 (Act 843), Republic of Ghana. Full text: ghalii.org/akn/gh/act/2012/843

Stages 1 to 3: Plan, Collect, Store

Following your GH¢ 50 transfer



Stage 4: Clean & Process

Can we trust the data? A real export, before cleaning

momo_transfers.csv					
	A	B	C	D	E
1	transaction_id	time	sender_id	amount_ghs	status
2	TX-10481	14:02:11	U-3371	200.00	success
3	TX-10481	14:02:11	U-3371	200.00	success
4	TX-10482	14:10:45	U-1204		success
5	TX-10483	14:31:02	U-9012	GHS 50	success
6	TX-10484	14:33:20	U-4410	75.50	SUCCESS

← duplicate row

← amount missing

← text, not a number

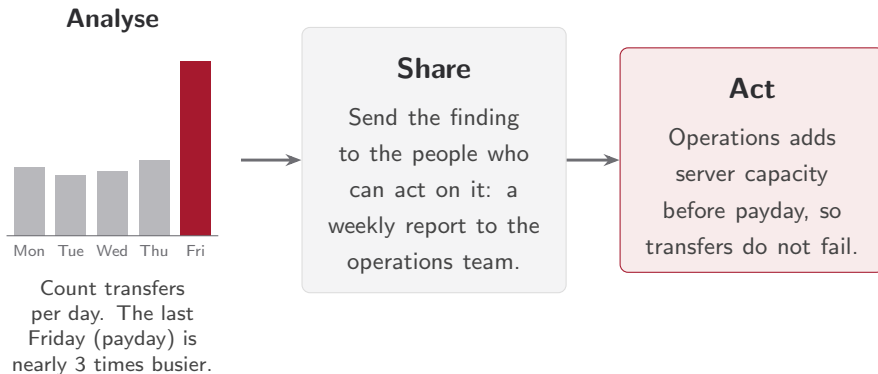
← inconsistent label

Clean: delete row 3 · flag row 4 · GHS 50 → 50.00 · SUCCESS → success

If row 3 stays in, what breaks?

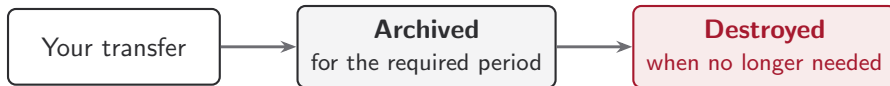
Stages 5 and 6: Analyse, Share & Act

From clean data to a decision



Stage 7: Archive or Destroy

[Back to Q2](#)

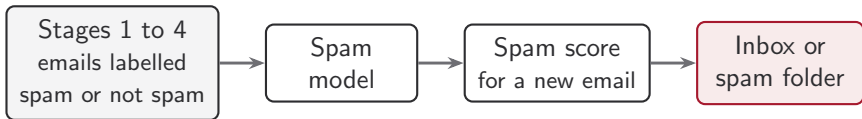


Q2: it exists as long as the provider's policy says.

Data Protection Act, 2012 (Act 843); Anti-Money Laundering Act, 2020 (Act 1044), Republic of Ghana. Both at ghalii.org

One place it is applied: machine learning

Example: the spam filter in your email



The model learns only from the emails it was given.

Wrong labels or missing types of scam mean real emails land in spam and scams reach your inbox.

Same idea for MoMo: a fraud model learns from past transfers, then decides whether to block yours.

Your turn: which stage went wrong?

A MoMo provider built a fraud model. Two things went wrong.

1

Problem 1

The model learned only from January to June transfers.

In December it blocks thousands of normal Christmas payments as fraud.

2

Problem 2

One input was the column `account_frozen`. Accounts are frozen only *after* fraud is confirmed.

In testing it was 99% correct.
Live, it misses most fraud.

In pairs: in which of the 7 stages did each problem start?

Plan · Collect · Store · Clean & Process · Analyse · Share & Act · Archive or Destroy

Your turn: which stage went wrong?

Answers

1

Collect

December was never collected.

The model has never seen holiday spending, so it treats it as unusual, and unusual looks like fraud.

2

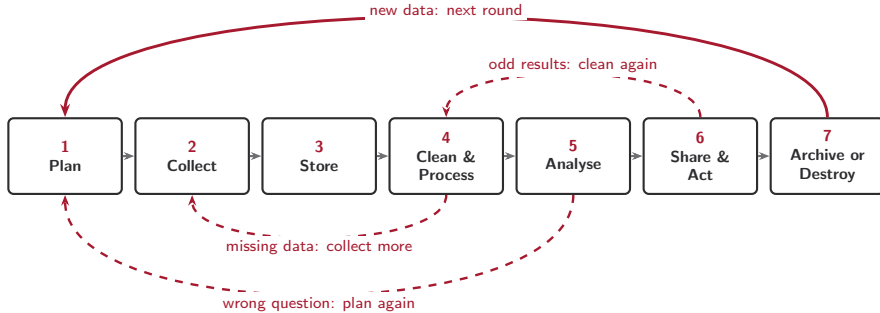
Clean & Process

Data leakage. When a new transfer arrives, no account has been frozen yet. The model relied on information it will never have.

Both started in the data, not the model.

Data leakage: S. Kaufman, S. Rosset, C. Perlich, O. Stitelman, *Leakage in Data Mining*, ACM TKDD, 2012.

Summary

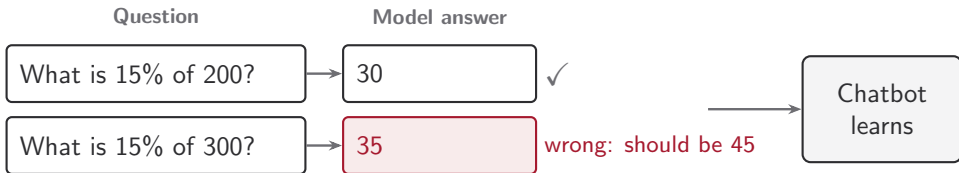


You can go back to an earlier stage from **any** stage, not only after stage 7.

Exit question: one app you used today. Where could its data go wrong?

Next time: the same lifecycle, for chatbot data

To teach a chatbot a new job, we give it thousands of examples, like **past exam questions with model answers**.



Collect: who writes thousands of these?
Often another AI.

Clean: who catches the wrong answers
before the chatbot learns them?

L. Ouyang et al., NeurIPS 2022 · Y. Wang et al., "Self-Instruct", ACL 2023.

Bonus: further reading

On data and the data lifecycle

- Data Protection Act, 2012 (Act 843), Ghana. ghalii.org/akn/gh/act/2012/843
- H. Wickham, “Tidy Data”, *Journal of Statistical Software*, 2014. how to organise a clean table
- *10 minutes to pandas*, pandas documentation. pandas.pydata.org/docs/user_guide/10min.html
- P. Chapman et al., *CRISP-DM 1.0: Step-by-step data mining guide*, 2000.
- DAMA International, *Data Management Body of Knowledge (DMBOK)*, 2nd ed., 2017.

On machine learning

- Google, *Machine Learning Crash Course*. developers.google.com/machine-learning/crash-course
- M. Zinkevich, *Rules of Machine Learning*, Google. developers.google.com/machine-learning/guides/rules-of-ml
- S. Kaufman et al., “Leakage in Data Mining”, *ACM TKDD*, 2012.
- D. Sculley et al., “Hidden Technical Debt in Machine Learning Systems”, *NeurIPS* 2015.
- Y. Wang et al., “Self-Instruct”, *ACL* 2023. for next session

Thank You

Questions?

Richard Djarbeng

MSc Electrical and Computer Engineering, Carnegie Mellon University

Bonus: Theory or practice?

In industry, every stage has a person and a tool

Stages	Who	Example tool	What the tool does
Collect, Store	Data engineer	PostgreSQL	open-source database that stores the records
Clean, Analyse	Data scientist	pandas (Python)	library for cleaning and summarising tables
Share & Act	ML engineer	MLflow	open-source tool, from Databricks, that records which data and settings produced each model
Archive or Destroy	Compliance officer	retention policy	written rule: how long each type of data is kept

PostgreSQL: [postgresql.org](https://www.postgresql.org) pandas: pandas.pydata.org MLflow: mlflow.org

Bonus: Use cases beyond machine learning

The lifecycle applies to any data an organisation keeps

Use	Data	Key stage	Why
National census	household answers	Clean	missing or double-counted homes distort population figures
Voter register	names, IDs, photos	Clean	duplicate registrations must be removed before an election
University records	grades, transcripts	Archive	graduates request transcripts decades later
Hospital records	patient visits	Store	only staff treating the patient should see them
Monthly sales report	shop receipts	Share & Act	managers restock based on what sold

Bonus: Use cases in machine learning

A use case is a real problem solved by running data through the lifecycle

Problem	Data used	Decision made	Stage that matters most
Stop card and MoMo fraud	past transactions	block or allow a payment	Archive: law sets how long
Predict hospital readmission	patient visit records	which patients get follow-up	Store: who may see records
Recommend the next video	what you watched	what to show you next	Collect: new views feed the loop
Price a ride (Uber, Bolt)	GPS trips, demand	surge price, driver placement	Clean: GPS gaps and errors

Bonus: Why data scientists care

Most time goes to data

In a 2020 survey, data scientists said about 45% of their time goes to loading and cleaning data.

Most failures are data failures

When a model goes wrong, the cause is usually leakage, duplicates or missing periods.

Reproducibility

You must be able to say exactly which data trained a model, or you cannot fix or explain it.

Legal responsibility

Keeping personal data “just in case” can break Act 843.

Anaconda, *State of Data Science 2020* (reported by Datanami, July 2020) · Kaufman et al., *ACM TKDD*, 2012 · Act 843.

Bonus: Who cares?

Regulators

Data Protection
Commission,
Bank of Ghana

Companies

Bad data leads
to wasted
money and
wrong decisions

Customers

It is their
payments, health
and location

Data scientists

When a model
fails, they must
find where the
data went wrong

Data Protection Commission, Ghana: dataprotection.org.gh Bank of Ghana: bog.gov.gh