

MODULE 2

Preprocessing messy and dynamic data

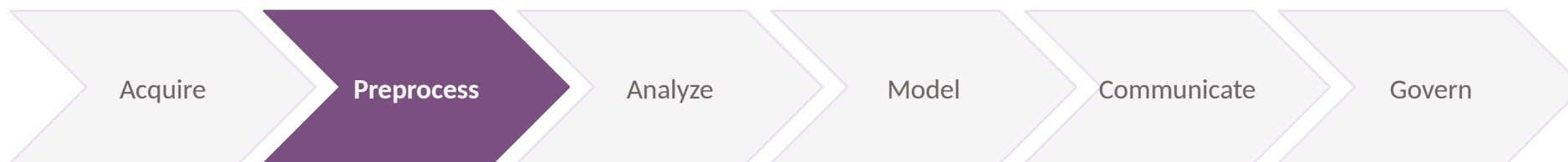
Real data arrives broken. Cleaning is a chain of defensible, reversible, documented decisions, not a chore.

Watch a little, then do a little



Each recorded segment is short. You pause, practice the concept hands-on, then return for the next piece.

Where this module sits in the lifecycle



This module builds the preprocessing layer. Every downstream model inherits the assumptions you make about missing values, outliers, units, and duplicates here, so each fix must be defensible, reversible, and documented.

PART 1 OF 2

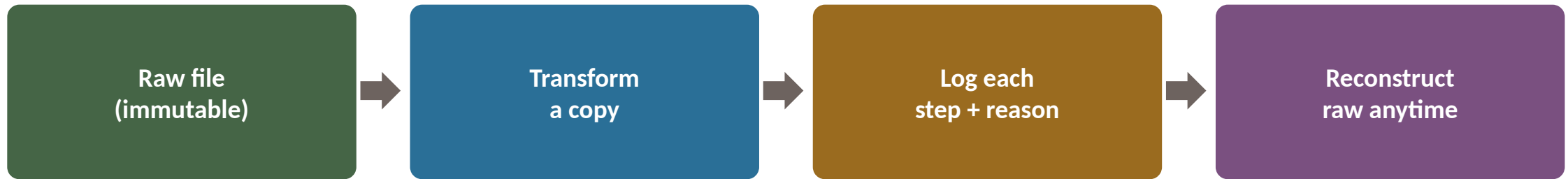
Cleaning as defensible decisions

About 12 minutes on camera

“I dropped the bad rows” is never a methods sentence

- Name the mechanism you believe produced a defect before you choose a fix. The mechanism, not the symptom, dictates the honest remedy.
- Keep the raw data immutable, transform only copies, and log every step so the raw state can always be reconstructed.
- A defensible sentence states the action, the assumption behind it, and how it can be reversed.

Reversible by construction



Every transformation records its rationale, its mechanism assumption, and how to undo it. That log is the deliverable a reviewer audits.

Outliers are not all the same

Data-entry error

A physically impossible value, a negative age, a mis-keyed unit. Correct or remove it, and record why.

Genuine extreme

A rare but real observation, a true high earner. Keep it; deleting it biases the result.

The test only flags

Grubbs or the generalized ESD marks a candidate for investigation, not an automatic deletion.

State a rule

Decide the treatment by a stated, reproducible criterion, not by eyeballing and reflex.

NOW TRY IT

Kaggle Learn: Data Cleaning + Feature Engineering

- Complete the Data Cleaning micro-course: missing values, scaling, dates, encodings, inconsistent entries.
- Complete the Feature Engineering micro-course, watching for features that leak the target.

About 45 minutes total, auto-graded, then come back for Part 2.

PART 2 OF 2

Missing-data mechanisms and leakage

About 13 minutes on camera

Diagnose the mechanism first

MCAR. missing completely at random, the hole is unrelated to any value, so deletion is unbiased but wasteful

MAR. missing at random given observed variables, so a model using those variables can impute honestly

MNAR. missing depends on the withheld value itself, so naive deletion or mean-fill biases your estimates

The rule. the mechanism, not the appearance of the hole, dictates which fix is honest

Leakage-safe feature engineering



The classic mistake is imputing or standardizing before the split, which leaks held-out information into training and inflates every score.

NOW TRY IT

missingno: see the shape of what is absent

- Load a small messy dataset and generate the matrix, bar, and heatmap or dendrogram views.
- Read the correlation of the nulls: columns whose missingness co-occurs point away from MCAR, toward MAR or MNAR.

About 15 minutes, then produce the formative cleaning log.

Formative deliverable (completion points)

- Submit a cleaning log for a small messy dataset of your choosing.
- For each decision record three things: the action, the mechanism assumption behind it (which of MCAR / MAR / MNAR and what evidence), and how the step is reversible.
- State how you checked each decision, for example comparing distributions before and after, or confirming row counts reconcile after a join.

How it counts. Submit the work with a short reflection or evidence you ran it. Credited on completion, not scored for correctness. This is the practice that makes the graded simulation go well.

GRADED ASSESSMENT

Recover the Signal

- You receive a per-student-seeded messy dataset with MNAR missingness, unlabeled unit inconsistencies, duplicates, and outliers of a known distribution.
- Recover the underlying signal; your recovered statistics are scored against a hidden truth only the instructor holds.
- The truth is private and unique to your seed, so a borrowed solution cannot match it. The grade rewards a principled, defended process.