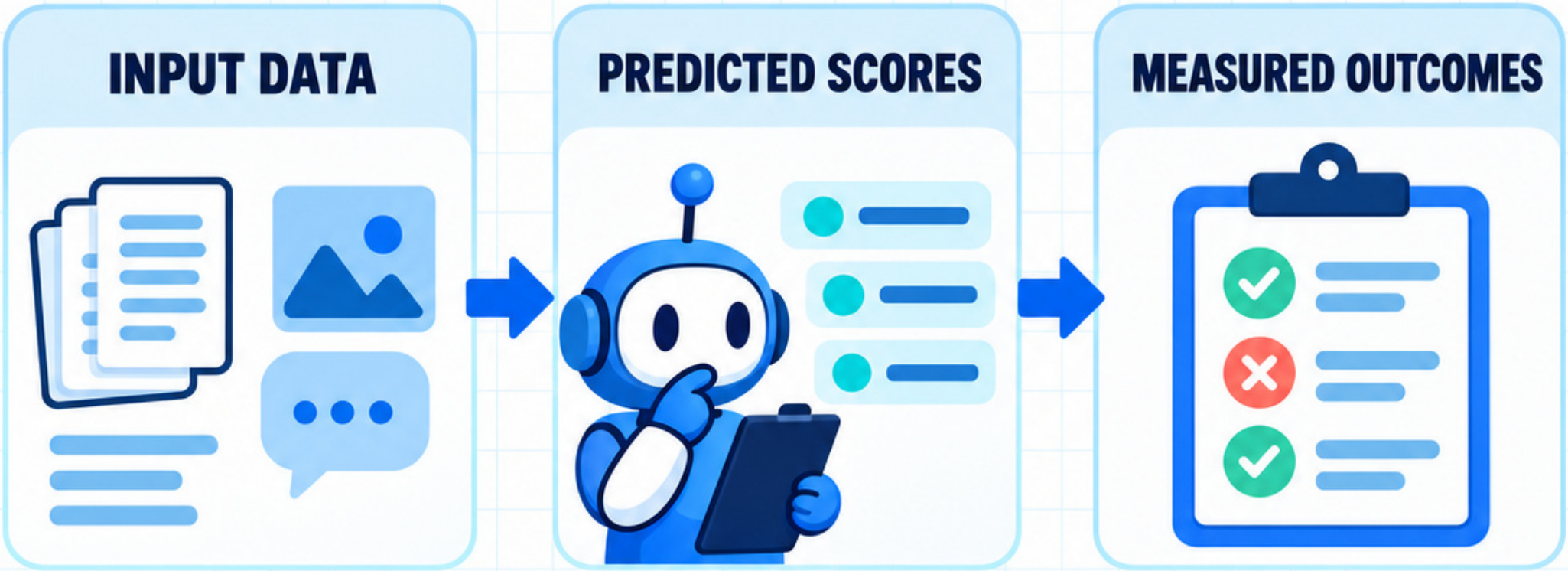
A CHANGING INPUT IS A CLUE

MONITOR DRIFT AND DELAYED OUTCOMES



A model can receive different data and still work.

It can also lose quality without a large input shift.

Today: decide what changed, what you know, and what to do next.

START WITH A MONITORING CONTRACT

MONITORING CONTRACT

- What is measured? Which population?
- What reference and time window?
- What minimum sample size?
- Who investigates, and what action is allowed?

OWNER

Responsible for investigating and taking action.

INVESTIGATION CHECKLIST

An alert without a response rule is unfinished design.

THREE DIFFERENT QUESTIONS

INPUT DRIFT:

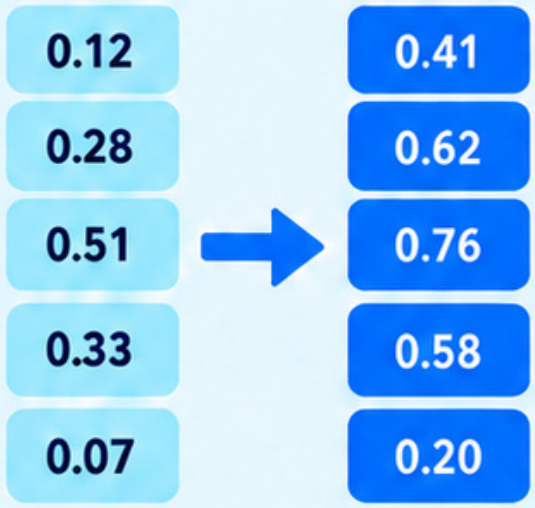
Did feature values change?

Illustrative distributions and scores

PREDICTION DRIFT:

Did model scores change?

Illustrative distributions and scores



QUALITY CHANGE:

Did predictions become less useful against observed outcomes?

The first two can be checked before labels arrive. They do not prove the third.

INPUT DRIFT IN PLAIN WORDS

Illustrative shop example:

Customers now place fewer weekly orders after a price change.

The feature distribution has shifted.

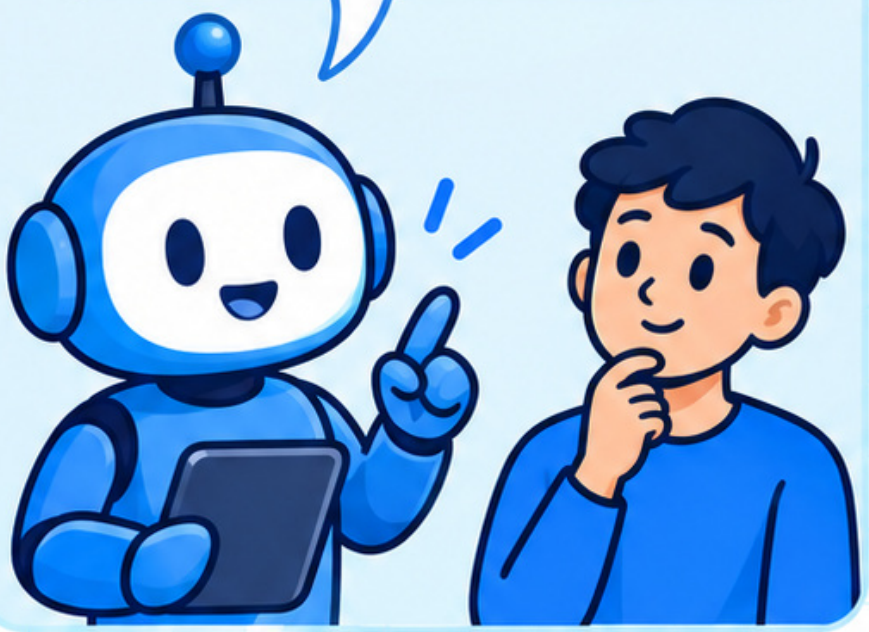
Possible causes include changed behavior, a different customer mix, or a broken event feed.

Investigate the cause before changing the model.

PREDICTION DRIFT IN PLAIN WORDS

Illustrative example:

More customers receive high inactivity scores this week.



Check the model version,

feature versions,

input mix, and missing-value rate.

A score shift can come from the pipeline or the population.

It is not a measured error rate.

QUALITY NEEDS OBSERVED OUTCOMES

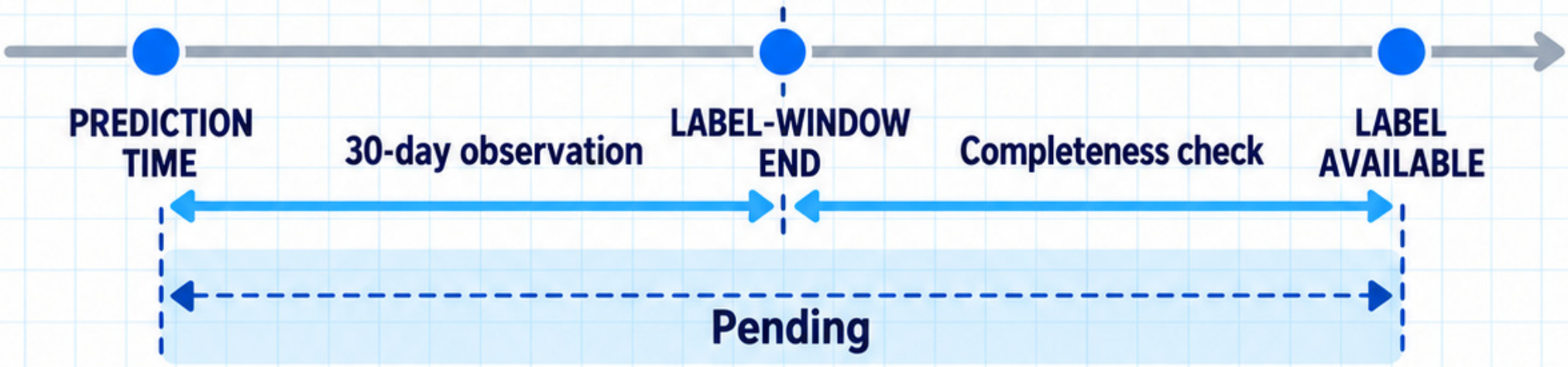
Toy target: no completed purchase during the next 30 days.

At prediction time, the future is unknown.

The 30-day window must finish before "no completed purchase" is final.

Also wait for the declared data-completeness rule.

Pending is not a class label.



KEEP THE CLOCKS SEPARATE

For each prediction, retain:

- Prediction ID and prediction time.
- Model version and score.
- Label-window end.
- Outcome event time and label availability time.

Join the matching label only when it is available to the evaluation run.

Illustrative record IDs

Prediction Table

Prediction ID	Prediction time	Model version	Score	Label-window end
p001	…	…	…	…
p002	…	…	…	…
p003	…	…	…	…
…	…	…	…	…

Join by prediction ID

Illustrative record IDs

Label Table

Prediction ID	Outcome event time	Label availability time
p001	…	…
p002	…	…
p003	…	…
…	…	…

Prediction time

When the prediction was made.

Outcome event time

When the outcome event occurred.

Label-window end

End of the window for the true label.

Label availability time

When the label becomes available.

Keep every clock separate, and join the matching label only when it is available to the evaluation run.

MEASURE LABEL COVERAGE FIRST

Illustrative mature cohort: 100 predictions.

80
have complete labels;

100
total

20
are unresolved.

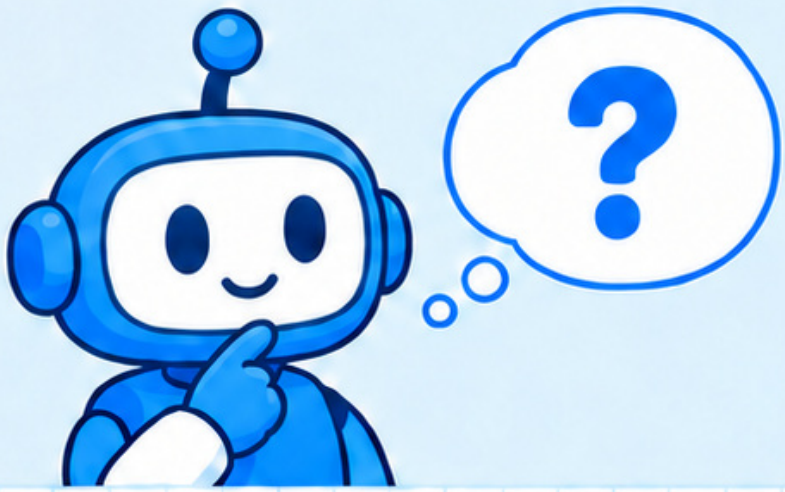
Label coverage = 80 / 100 = 80%.

Quality on those 80 may be biased if missing labels are selective.

Report coverage and exclusion reasons beside quality.

COMPARE LIKE WITH LIKE

Choose a reference that matches the question.

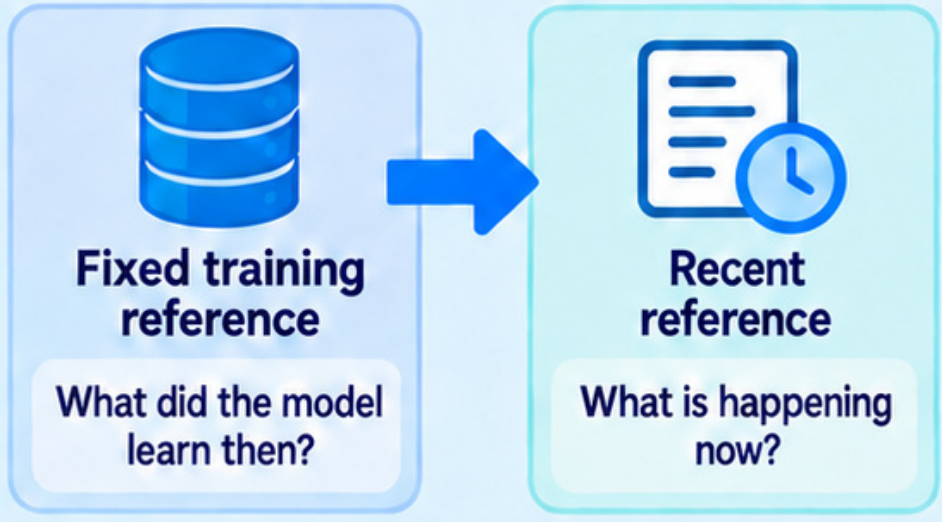


Compare weekends with weekends when seasonality matters.

Keep population, feature definition, model version, and sampling rules explicit.

- Population
- Feature definition
- Model version
- Sampling rules

A fixed training reference and a recent reference answer different questions.



A DRIFT SCORE NEEDS CONTEXT

A statistical test or distance summarizes a difference.

Its behavior depends on sample size, feature type, and the chosen method.

Sample size

Feature type

Method

Configure a meaningful threshold and minimum sample.

Threshold

Minimum sample

Review effect size, source changes, and business context.

Effect size

Source changes

Business context

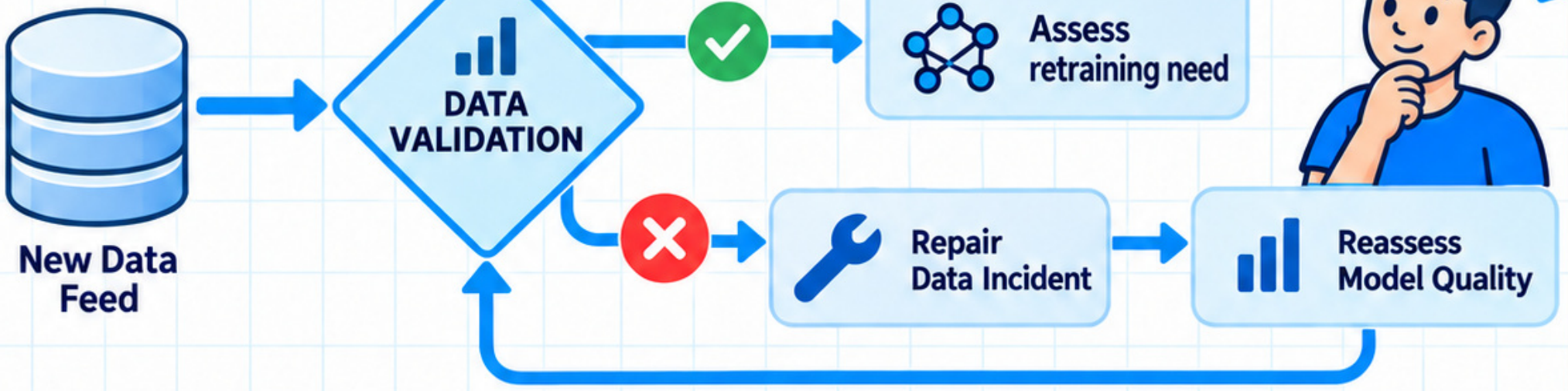
No universal drift number proves failure.

CHECK DATA HEALTH BEFORE RETRAINING

- Missing columns?
- Unexpected nulls?
- Late batches?
- Changed units?
- Duplicate events?
- New category encoding?

A model trained on broken inputs can preserve the problem.

Fix or isolate the data incident, then reassess model quality.



THREE ALERTS WITH THREE RESPONSES

Illustrative policies, not universal thresholds:

SIGNAL	DATA	DRIFT	QUALITY
	source misses its deadline → investigate ingestion.	sustained shift with enough samples → inspect slices and recent changes.	mature-cohort metric breaches its agreed limit → compare a candidate and consider rollback.
EVIDENCE			
RESPONSE	 investigate ingestion.	 inspect slices and recent changes.	 compare a candidate and consider rollback.

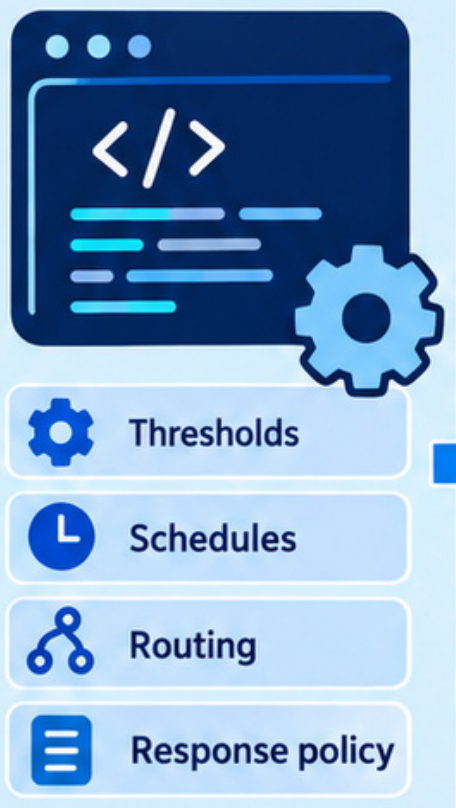
AI SYSTEM DESIGN • DAY 16

13/18

USE TOOLS FOR CLEAR JOBS

SQL joins predictions to available labels and defines cohorts.

Evidently compares reference and current data and reports drift or quality metrics.



Application code owns thresholds, schedules, routing, and response policy.

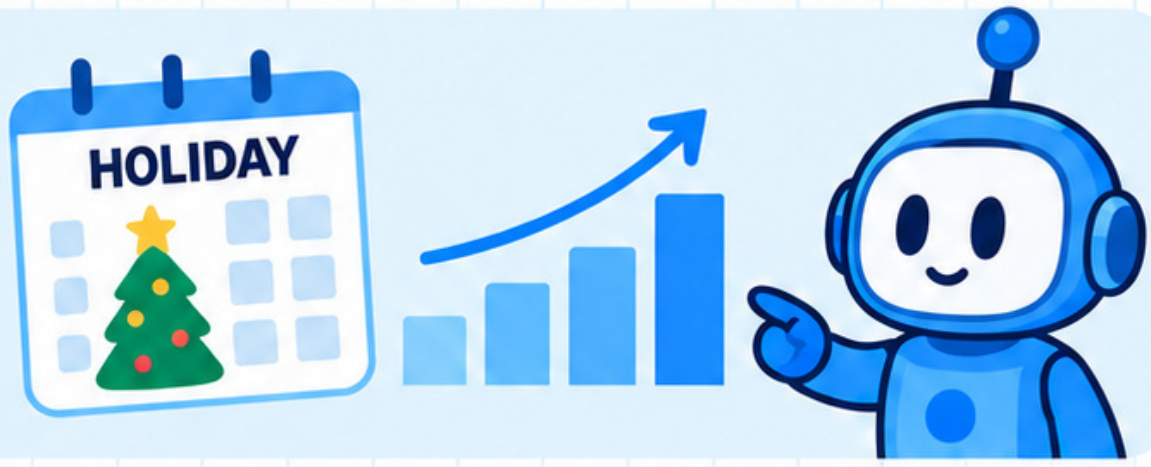
A report does not authorize retraining or deployment.

@learn.machinelearning

REPLAY A DRIFT-ONLY SCENARIO

Illustrative replay:

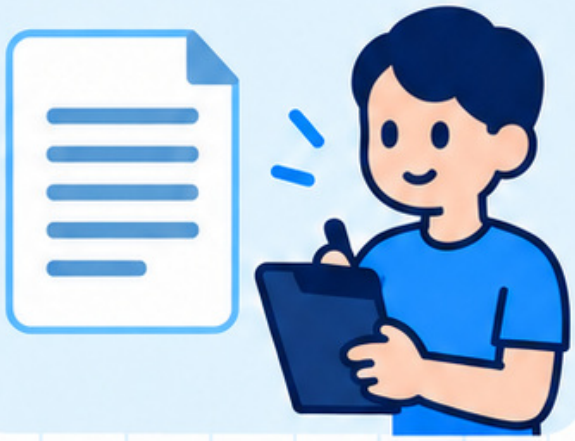
A holiday changes order frequency.



Data checks pass.
Label coverage is complete.

Quality remains inside the agreed range on a comparable mature cohort.

Response: document and monitor the shift.

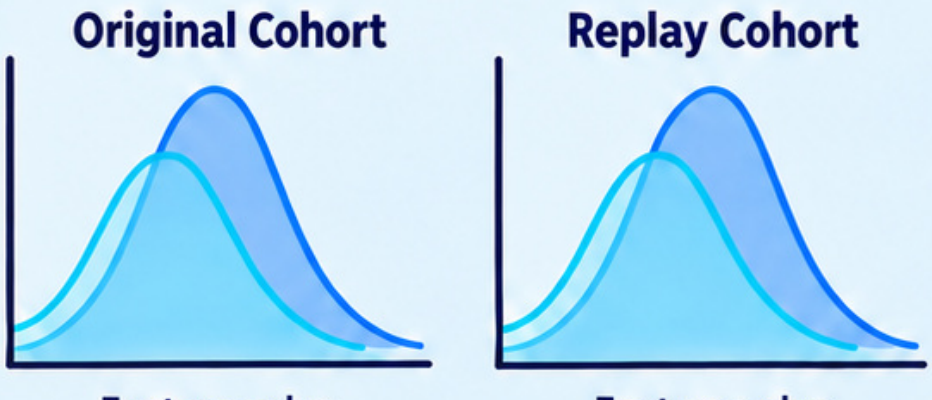


Drift alone does not justify replacing the model.

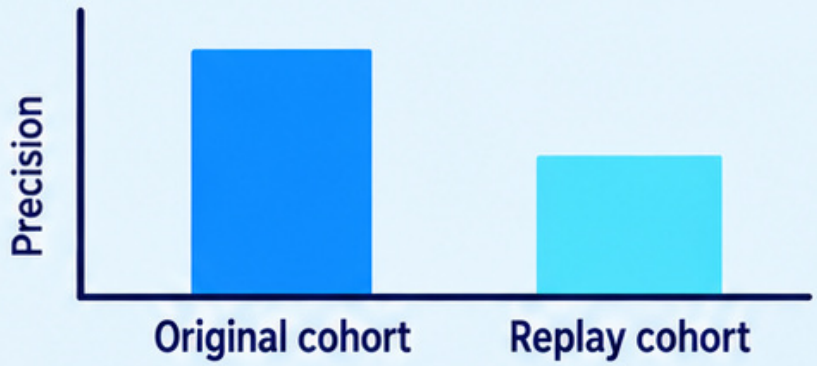
REPLAY A QUALITY REGRESSION

Illustrative replay:

Feature distributions look similar.



A mature, fully labeled cohort shows worse precision at the same review budget.



Check label logic, population slices, score ordering, and interventions.

- Check label logic
- Check population slices
- Check score ordering
- Check interventions

A quiet drift dashboard does not prove a healthy model.

RETRAIN THROUGH THE RELEASE PATH

EVIDENCE

If evidence supports retraining:

CANDIDATE

Create a candidate with versioned inputs.

EVALUATION GATE

Compare against the current model on suitable holdouts.

Check quality, capacity, slices, and operating constraints.

Checks fail

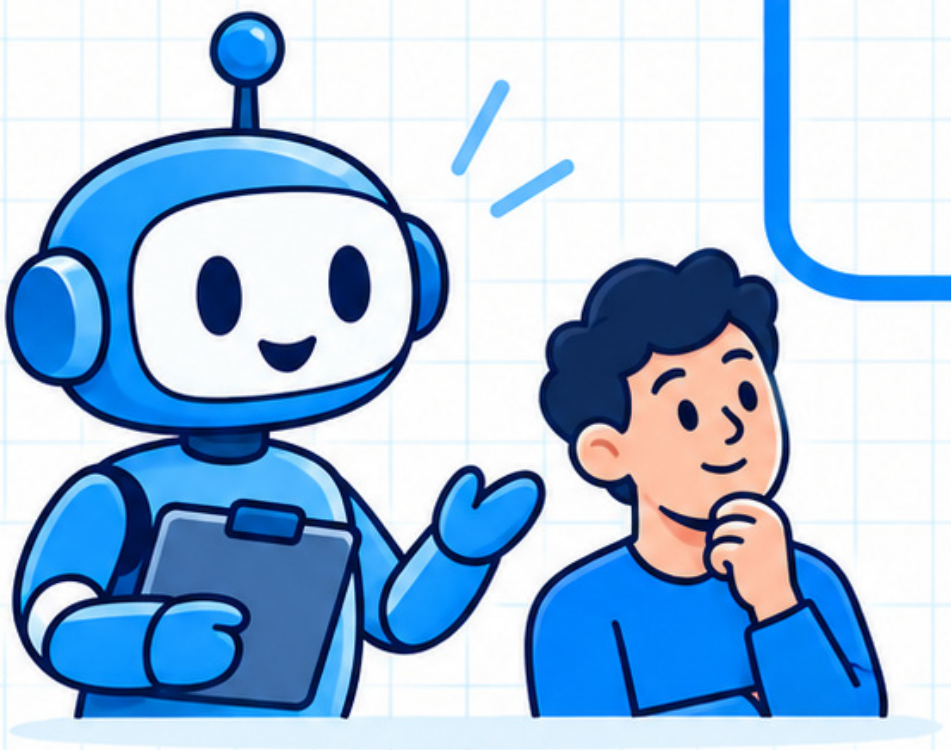
INVESTIGATION

APPROVED RELEASE

Release only after the acceptance checks pass.

ROLLBACK AND MONITORING

Keep rollback and monitoring in the design.



YOUR TURN: CHOOSE THE RESPONSE

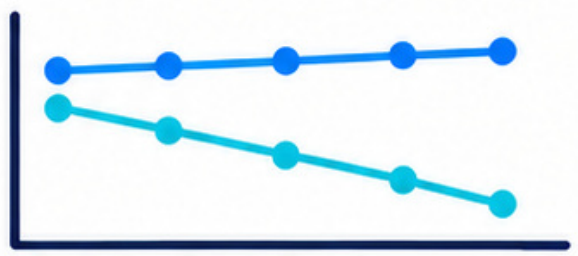
Illustrative cases:

A Order amounts shift during a sale; mature quality is stable.

B A source column suddenly becomes null.

id	source	value
101	A123	42
102	null	17
103	A456	29

C Fully labeled quality falls while inputs look stable.



Choose: monitor context, repair data, or investigate quality.

Monitor context

Repair data

Investigate quality

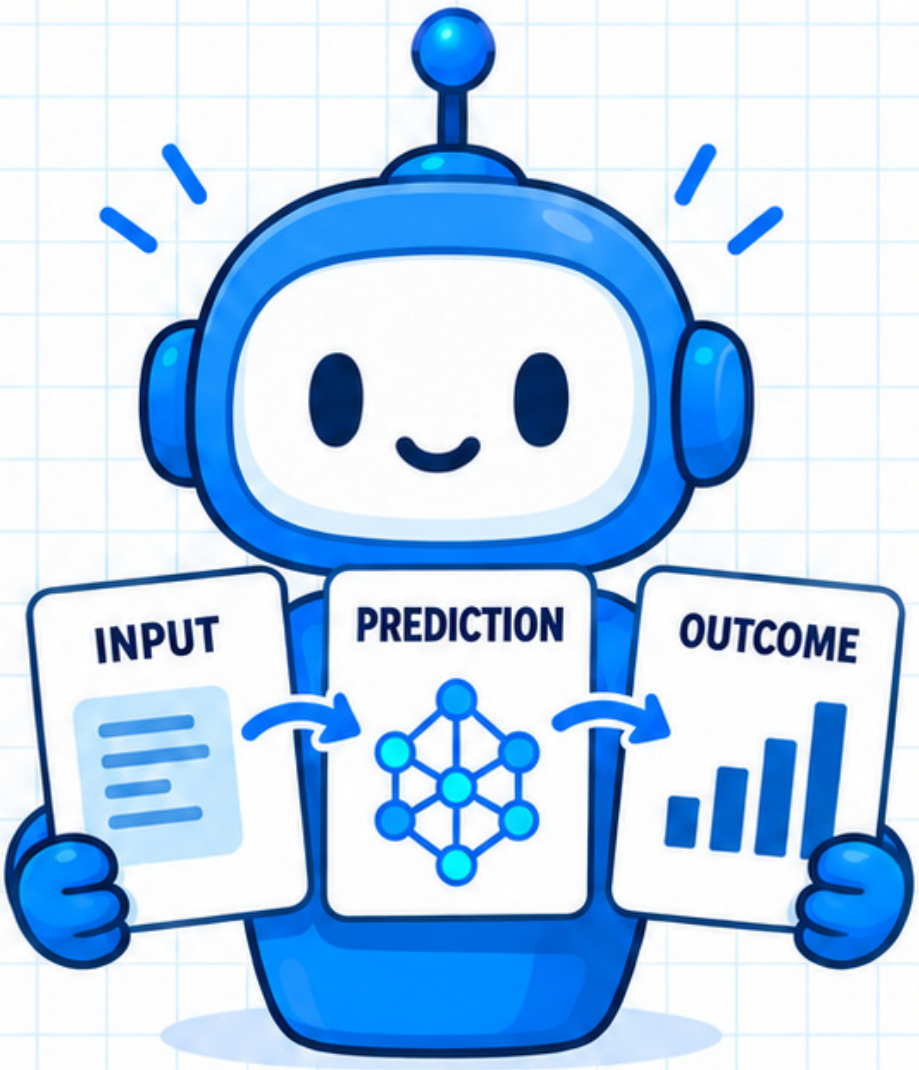
What evidence must you check first?

AI SYSTEM DESIGN • DAY 16

18/18

MONITOR WHAT YOU CAN KNOW

A → monitor the sale context after validating data and comparable quality.



B → investigate and repair the data incident.

C → investigate the quality regression.

Always report label maturity and coverage.

Next: recommendation retrieval and ranking.