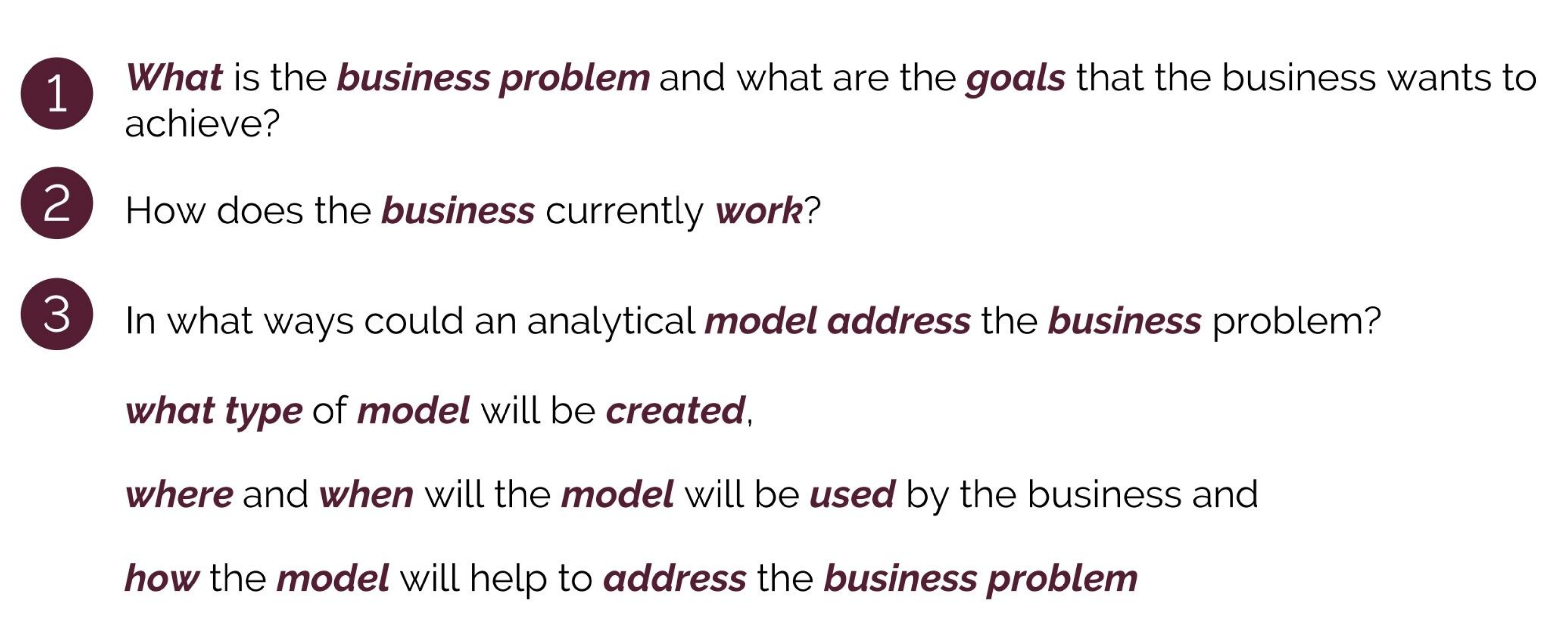
Business understanding phase

→ business problem converted into analytical sol. by answering:



1. **What** is the **business problem** and what are the **goals** that the business wants to achieve?
2. How does the **business** currently **work**?
3. In what ways could an analytical **model address** the **business** problem?

what type of **model** will be **created**,

where and **when** will the **model** will be **used** by the business and

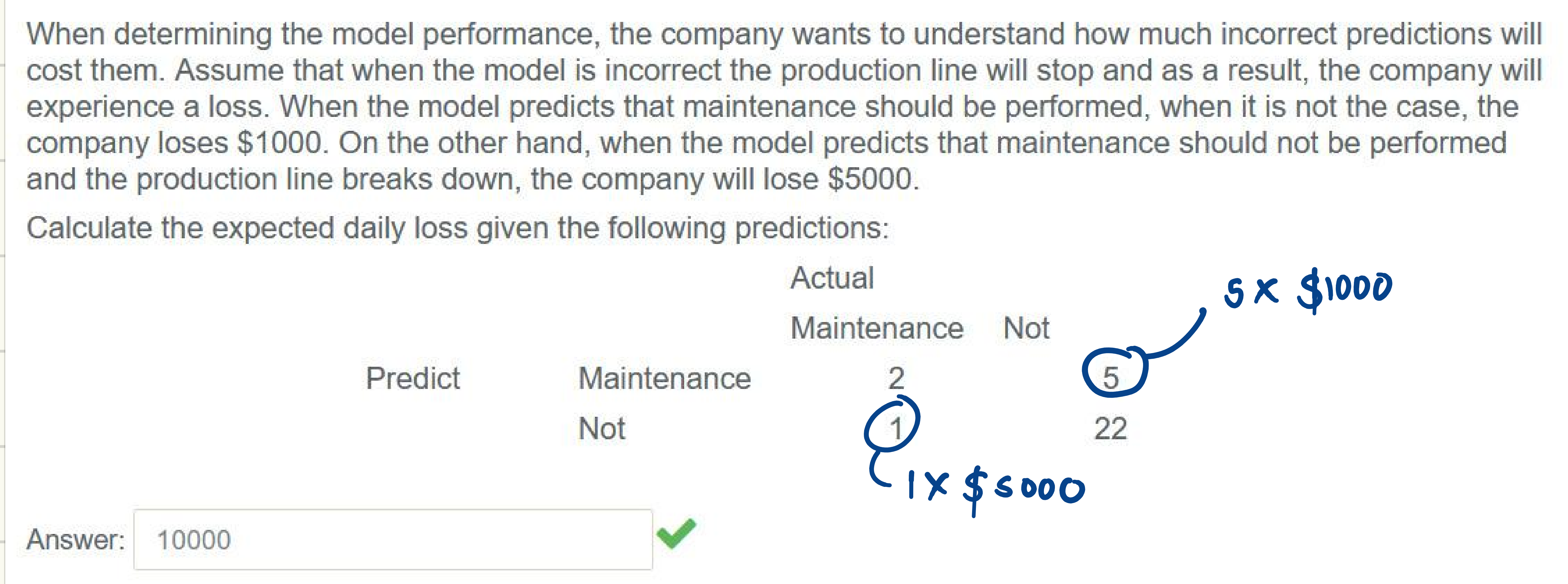
how the **model** will help to **address** the **business problem**

confusion matrix

		Actual	
		true	false
Prediction	true	true positive a	false positive b
	false	false negative c	true negative d

formula for model accuracy: $\frac{(a+d)}{(a+b+c+d)}$

QUESTIONS FROM QUIZ:



When determining the model performance, the company wants to understand how much incorrect predictions will cost them. Assume that when the model is incorrect the production line will stop and as a result, the company will experience a loss. When the model predicts that maintenance should be performed, when it is not the case, the company loses $1000. On the other hand, when the model predicts that maintenance should not be performed and the production line breaks down, the company will lose $5000.

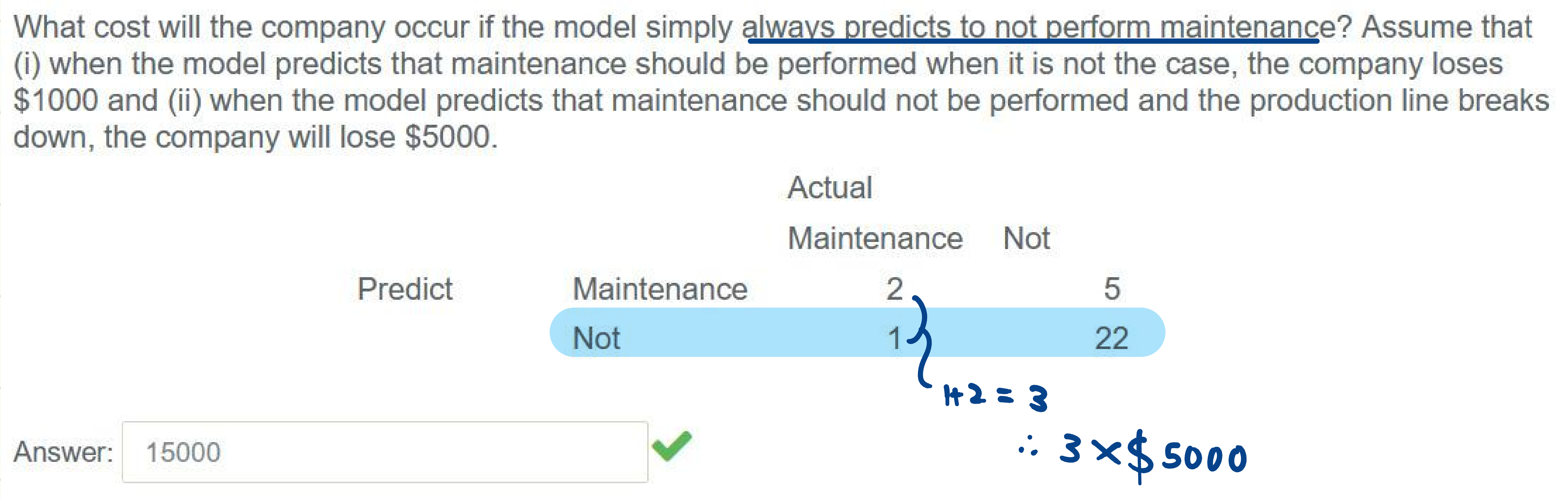
Calculate the expected daily loss given the following predictions:

		Actual	
		Maintenance	Not
Predict	Maintenance	2	5
	Not	1	22

5 x $1000

1x $5000

Answer: 10000 ✓



What cost will the company occur if the model simply always predicts to not perform maintenance? Assume that (i) when the model predicts that maintenance should be performed when it is not the case, the company loses $1000 and (ii) when the model predicts that maintenance should not be performed and the production line breaks down, the company will lose $5000.

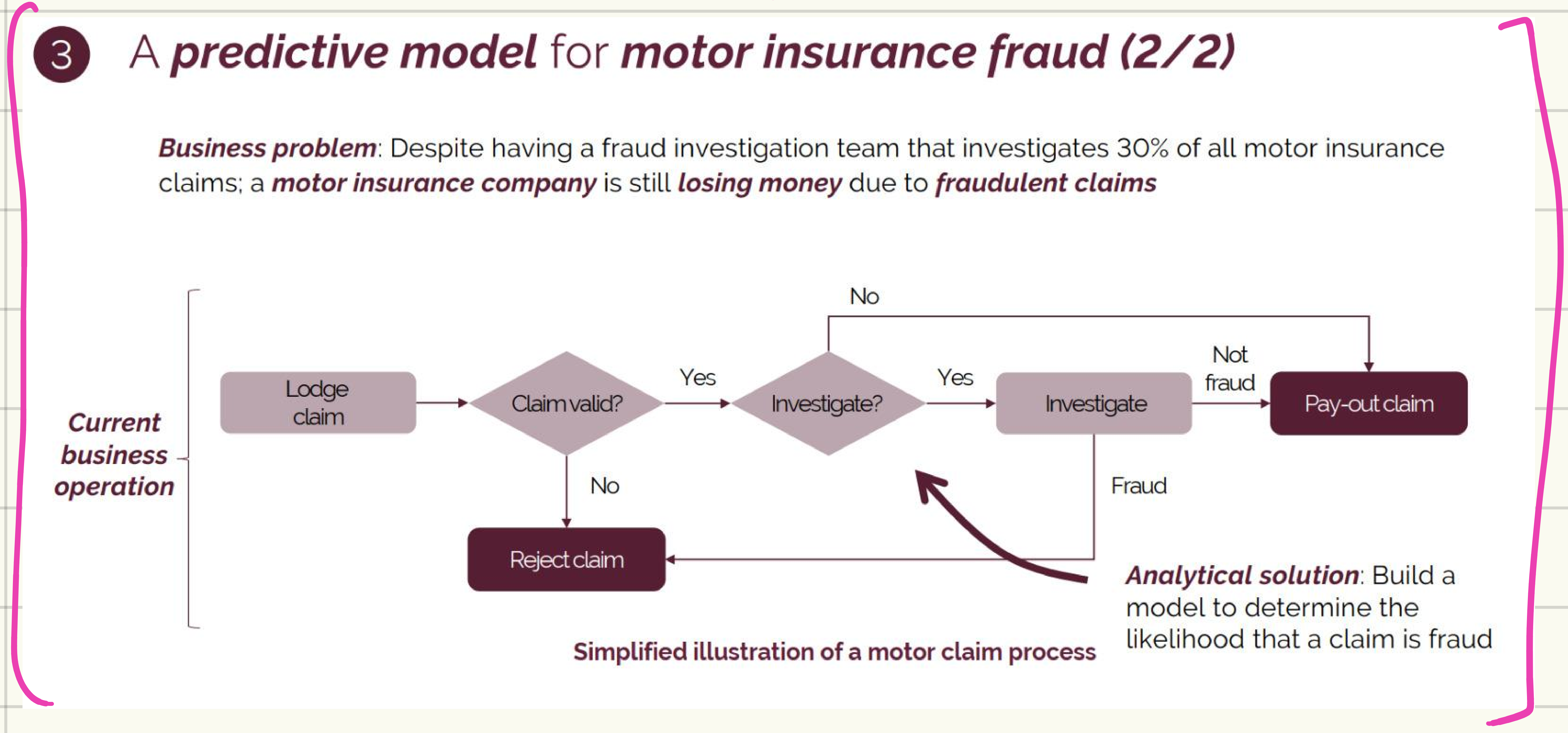
		Actual	
		Maintenance	Not
Predict	Maintenance	2	5
	Not	1	22

1+2 = 3

∴ 3×$5000

Answer: 15000 ✓

Slide example: analytical process



3 A ***predictive model*** for ***motor insurance fraud (2/2)***

Business problem: Despite having a fraud investigation team that investigates 30% of all motor insurance claims; a ***motor insurance company*** is still ***losing money*** due to ***fraudulent claims***

Analytical solution: Build a model to determine the likelihood that a claim is fraud

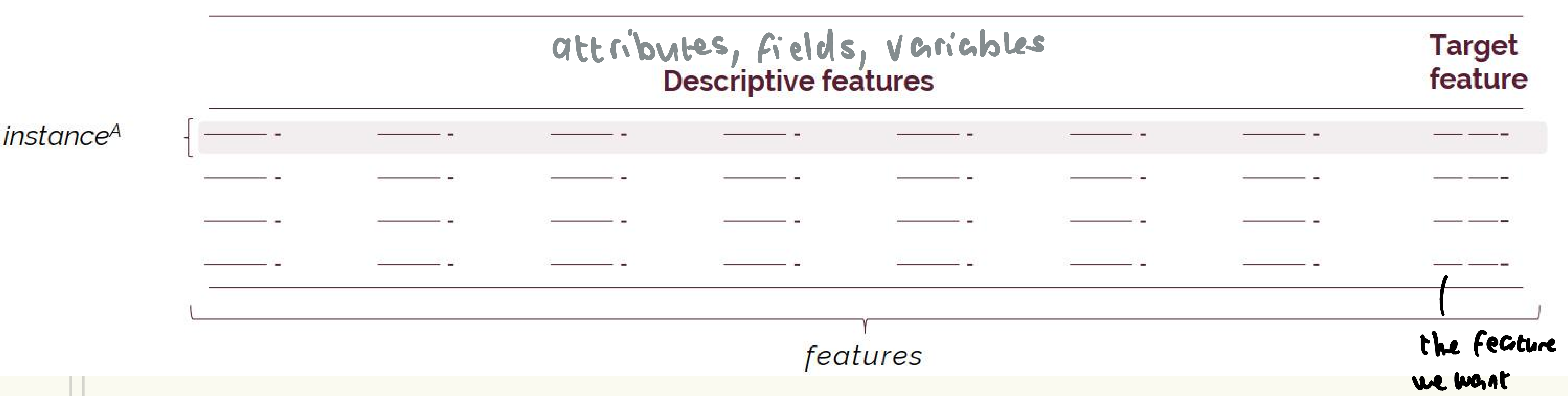
Simplified illustration of a motor claim process

constructed from multiple data sources, undergoes organization.

THE ANALYTICAL BASE TABLE (ABT)

ROW = case (row-per-subject)

column = feature → measurable property of object we want to analyse.



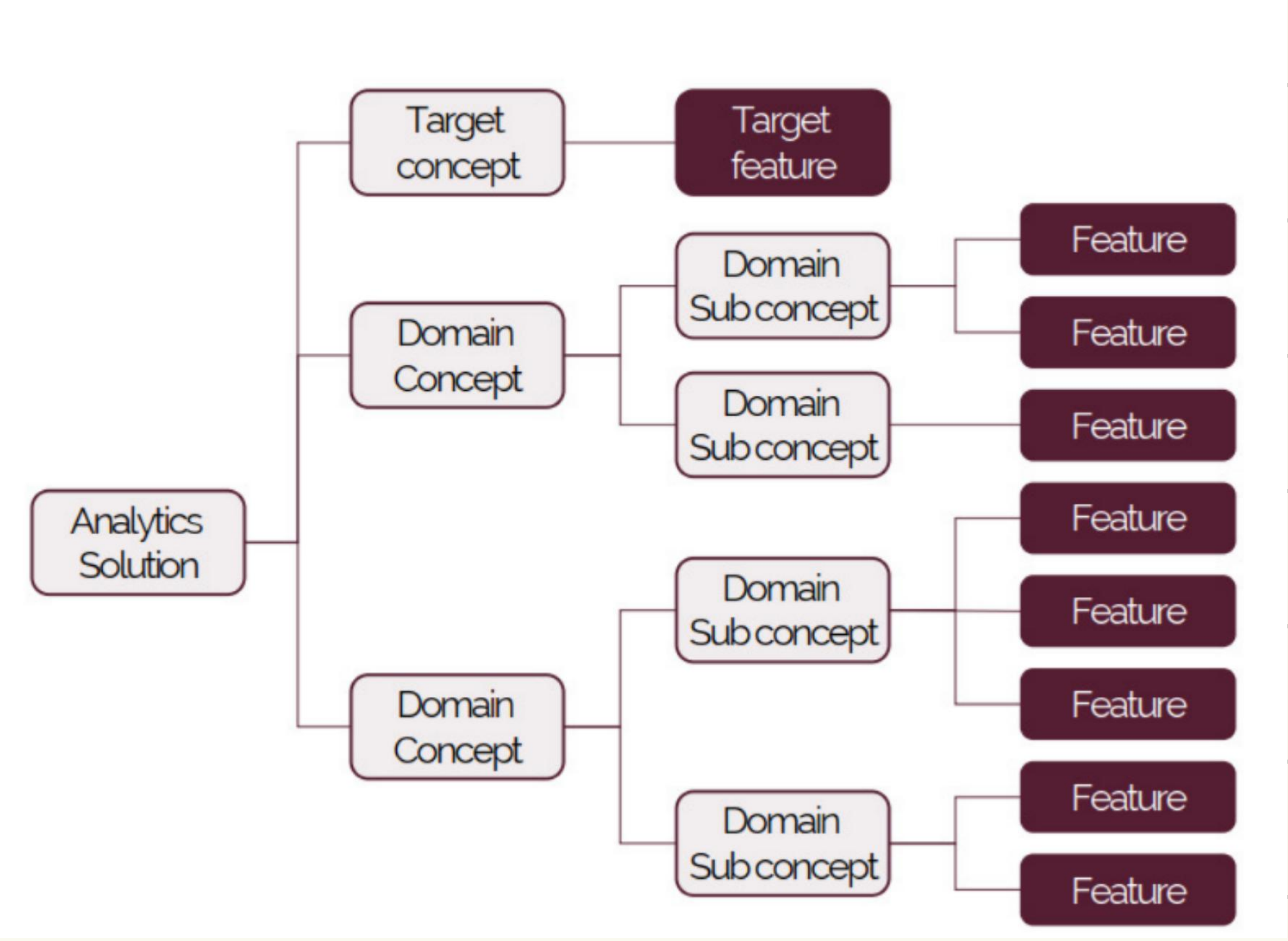
the feature we want our model to determine

FEATURE IDENTIFICATION

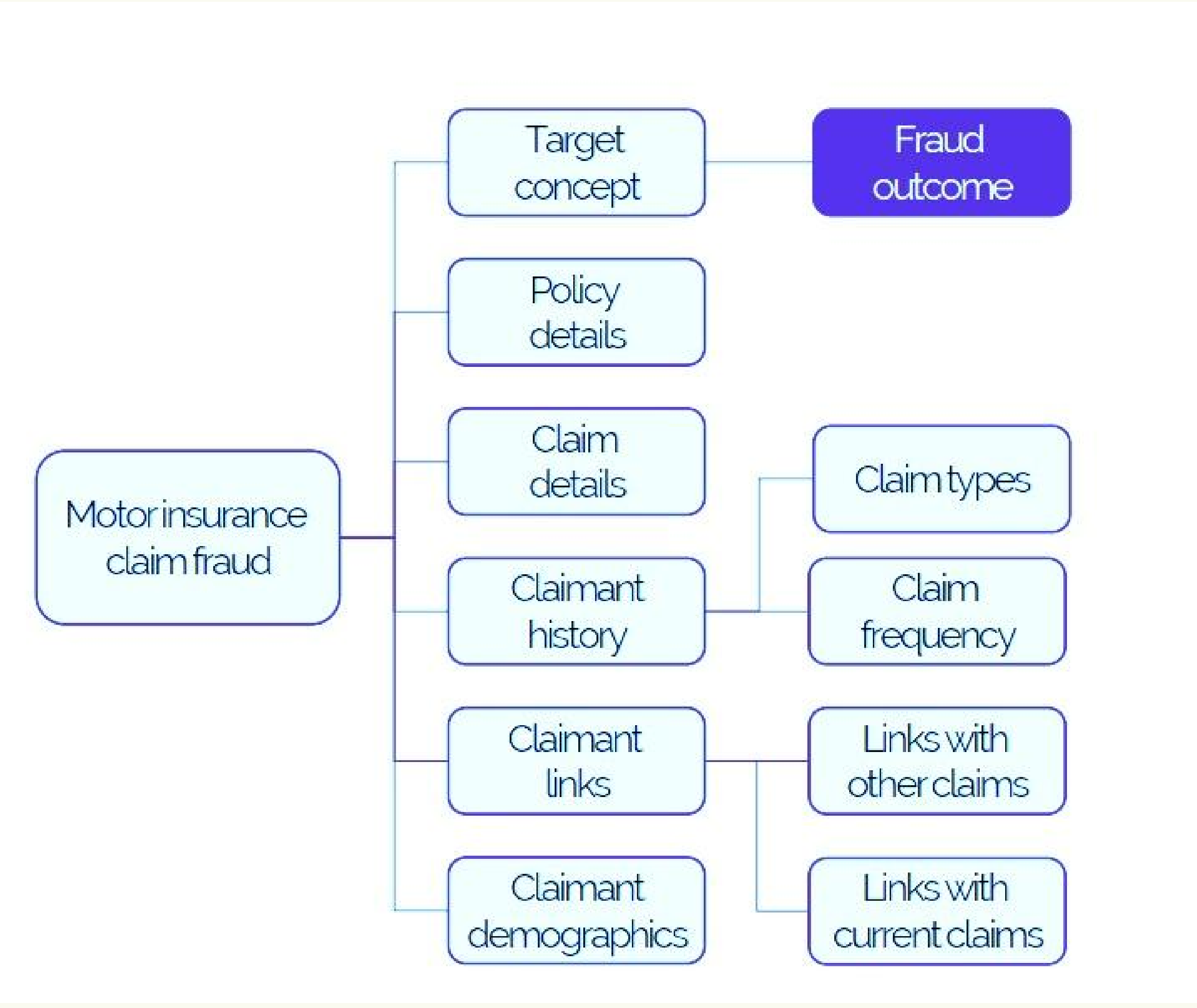
feature selection → based on: • domain knowledge

• analysis of relationship between features

→ to identify features, identify set of domain concepts → high-level abstraction that describes some characteristics of the prediction subject from which we derive a set of concrete features.



e.g. ⇒



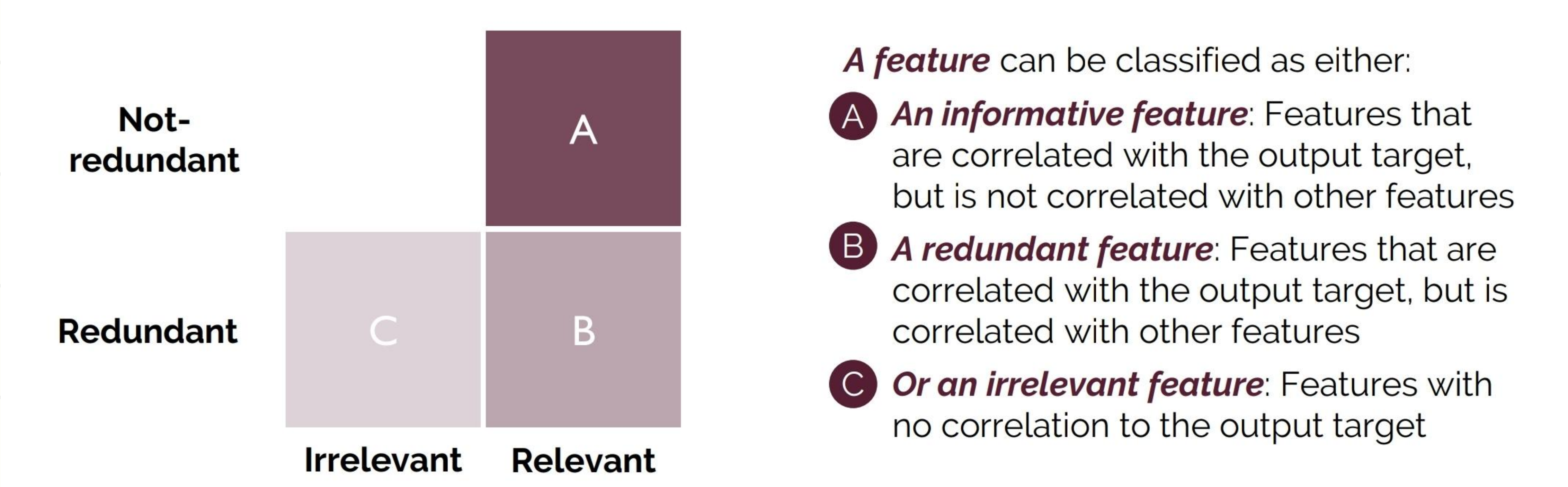
FEATURES
- raw features : directly copied from source data
- derived features: constructed from 1+ raw data sources

Feature design considerations

Considerations	Description	Examples
Availability	data must be available	The domain expert, advises that the voltage applied to the equipment terminal could be a potential descriptive feature. Sensors that record these readings are however not installed at the plant.
Timing	data must be available B4 target feat. is known	You are interested in creating a feature, time since the last maintenance was performed
Longevity	data can become stale	Historic data was collected for the period 2015 to 2017. At the start of 2016, the company started to perform preventative maintenance every week, rather than monthly

FEATURE SELECTION

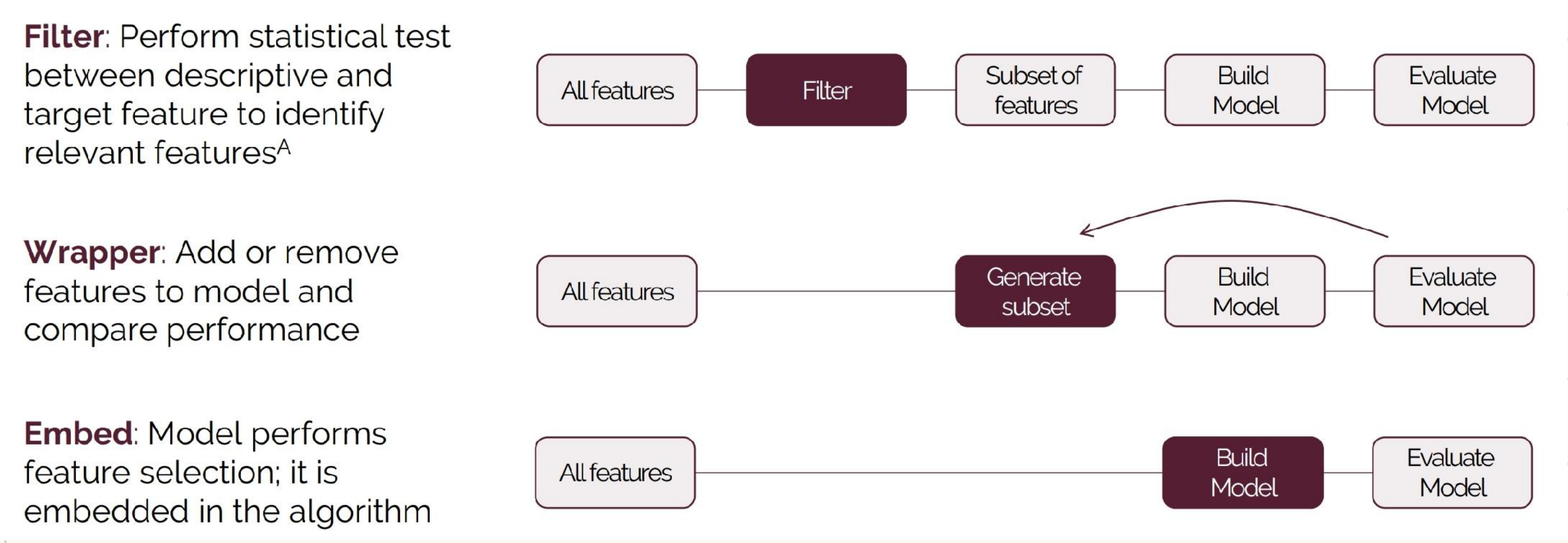
→ process of selecting a subset of the most informative features for use in model construction.



A feature can be classified as either:

- (A) ***An informative feature***: Features that are correlated with the output target, but is not correlated with other features
- (B) ***A redundant feature***: Features that are correlated with the output target, but is correlated with other features
- (C) ***Or an irrelevant feature***: Features with no correlation to the output target

fewer :
- simpler models
- shorter training time
- reduce potential to overfit



Filter: Perform statistical test between descriptive and target feature to identify relevant features^A

Wrapper: Add or remove features to model and compare performance

Embed: Model performs feature selection; it is embedded in the algorithm