Train In Data

Cardinality

Cardinality definition

- The values of a categorical variable are selected from a group of categories (also called labels).

- The number of different labels is known as **cardinality**.

Cardinality examples

- The variable gender contains only 2 labels in this example
- Vehicle Make contains 9 labels in the example table
- The variables city or postcode, can contain a huge number of different labels.

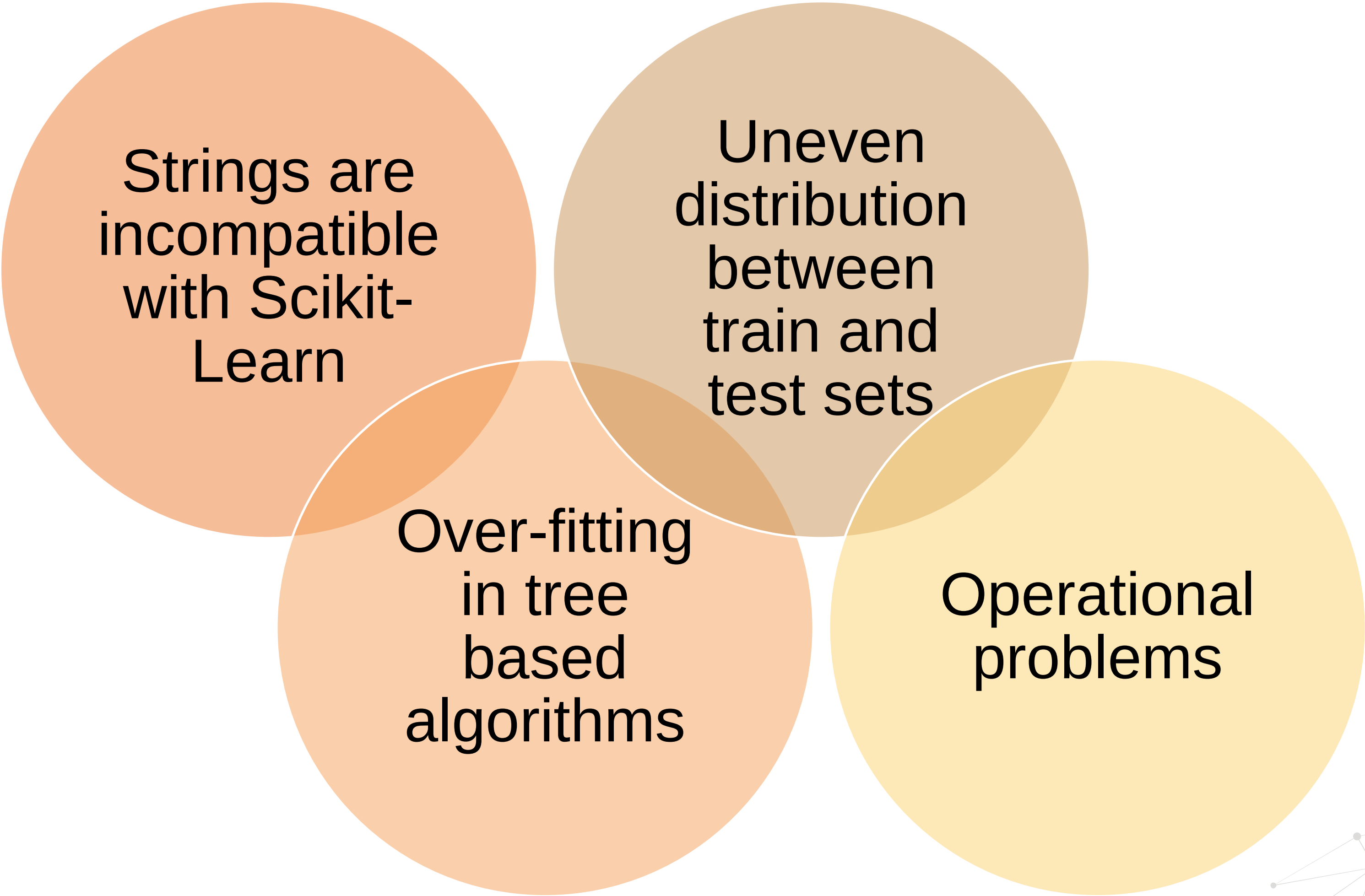
Gender	Vehicle Make
Male	Mercedes
Male	Ford
Male	Ford
Male	Renault
Male	Seat
Male	Renault
Female	Citroen
Female	Toyota
Female	Kia
Female	Kia
Female	Nissan
Female	BMW

Gender ➔ 2
Vehicle Make ➔ 9

Cardinality effects

Are multiple labels in a categorical variable a problem?

Cardinality: Impacts



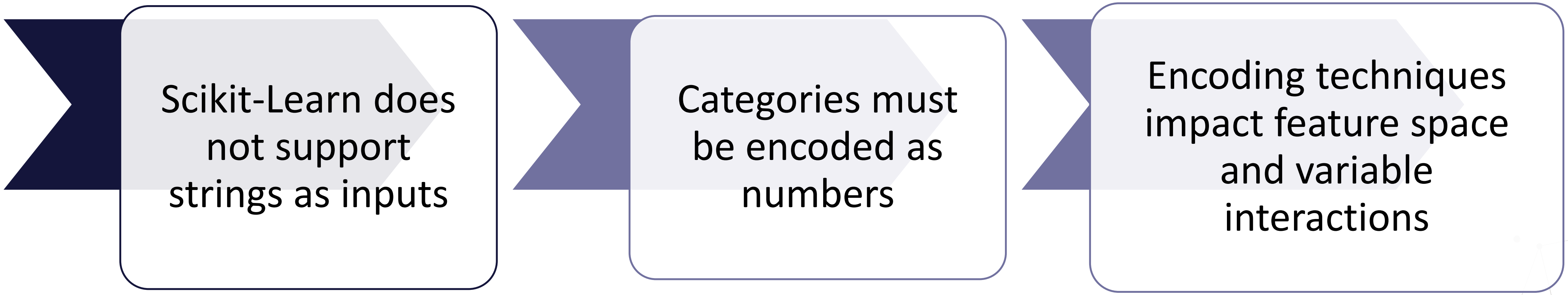
Strings are incompatible with Scikit-Learn

Uneven distribution between train and test sets

Over-fitting in tree based algorithms

Operational problems

Strings and categorical encoding



Scikit-Learn does not support strings as inputs

Categories must be encoded as numbers

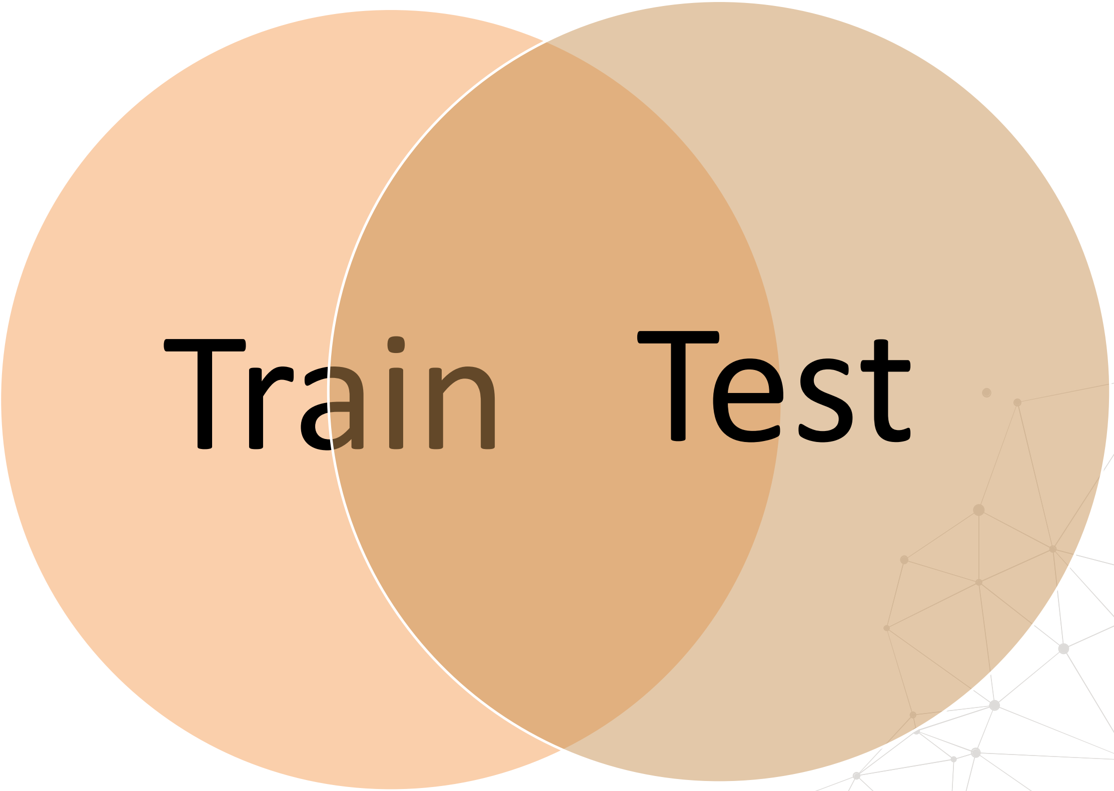
Encoding techniques impact feature space and variable interactions

More on encoding methods in a dedicated section…

Uneven distribution between train and test sets

For highly cardinal variables:

- Some labels may appear only in train set ➔ over-fitting
- Some labels may appear only in test set ➔ model will not know how to interpret the values



Uneven distribution

Obs	Vehicle Make
1	Mercedes
2	Ford
3	Ford
4	Renault
5	Seat
6	Renault
7	Citroen
8	Toyota
9	Kia
10	Kia
11	Nissan
12	BMW

Train Set

Obs	Vehicle Make
1	Mercedes
3	Ford
6	Renault
7	Citroen
9	Kia
11	Nissan

Test Set

Obs	Vehicle Make
2	Ford
5	Seat
4	Renault
8	Toyota
10	Kia
12	BMW

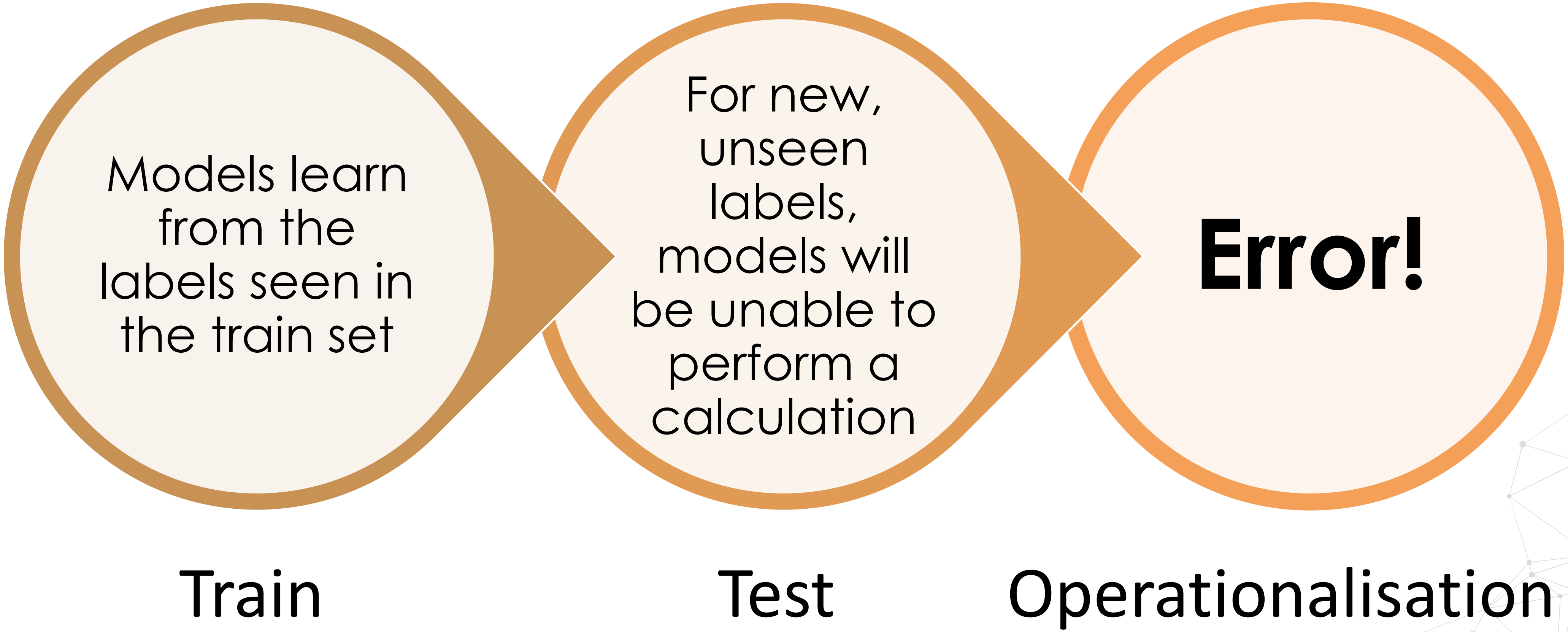
Overfitting

Cardinality and overfitting

- Variables with too many labels tend to dominate over those with fewer labels, particularly in **tree based algorithms.**
- A big number of labels within a variable may introduce noise with little, if any, information
- Reducing cardinality may help improve model performance

Operational problems

Cardinality and operationalisation



Models learn from the labels seen in the train set

For new, unseen labels, models will be unable to perform a calculation

Error!

Train

Test

Operationalisation

Summary

- Strings need to be encoded as numbers for use with Scikit-Learn

- High cardinality may cause over-fitting and operationalisation problems

- Reducing cardinality may improve model performance

Accompanying Jupyter Notebook

- Read the accompanying Jupyter Notebook

- How to quantify cardinality
- Examples of high and low cardinality variables
- Effect of cardinality when preparing train and test sets
- Effect of cardinality on Machine Learning Model performance

THANK YOU

www.trainindata.com