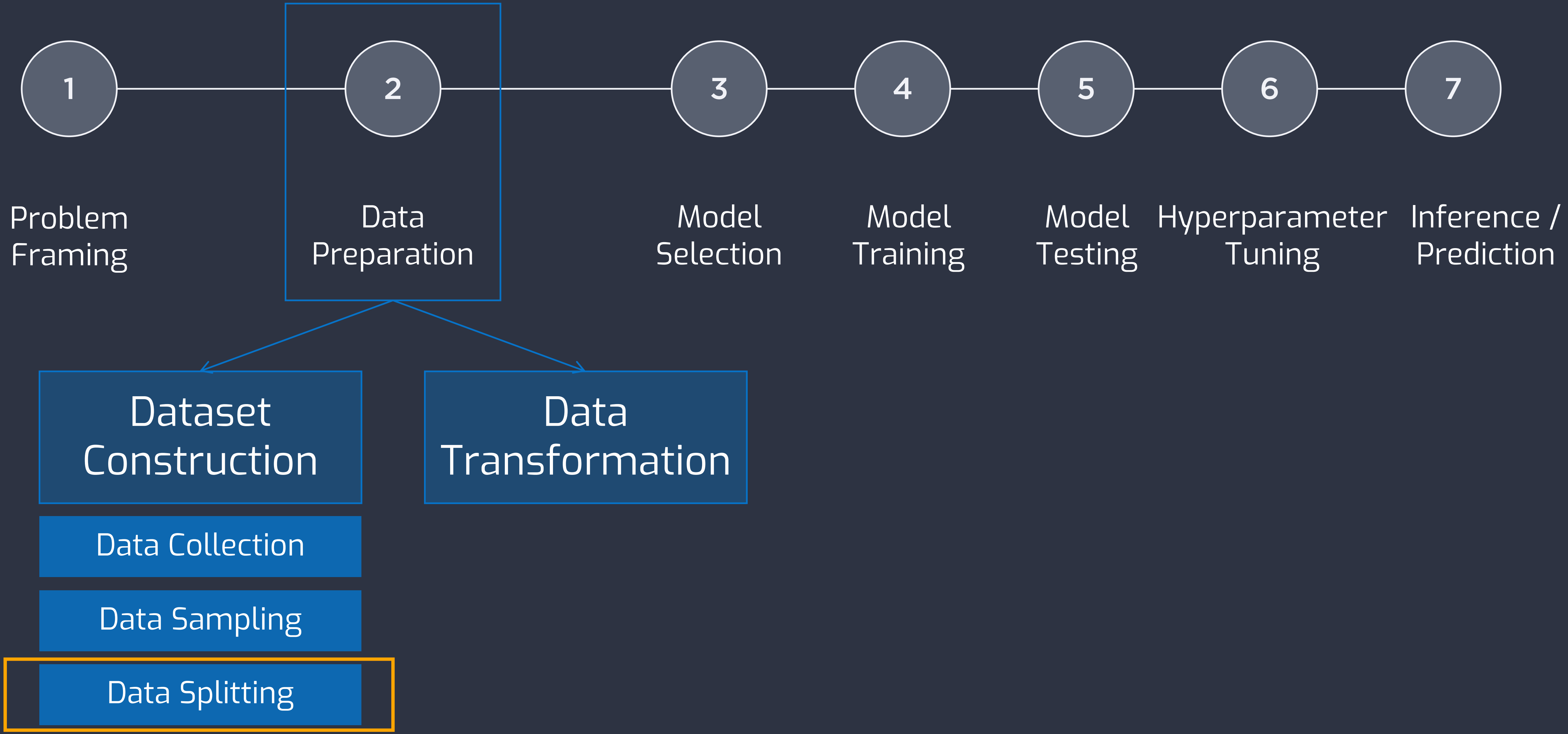
COMP2261 ARTIFICIAL INTELLIGENCE / MACHINE LEARNING

Data Splitting

Dr SHI Lei



1 Problem Framing

2 Data Preparation

3 Model Selection

4 Model Training

5 Model Testing

6 Hyperparameter Tuning

7 Inference / Prediction

Dataset Construction

- Data Collection
- Data Sampling
- Data Splitting

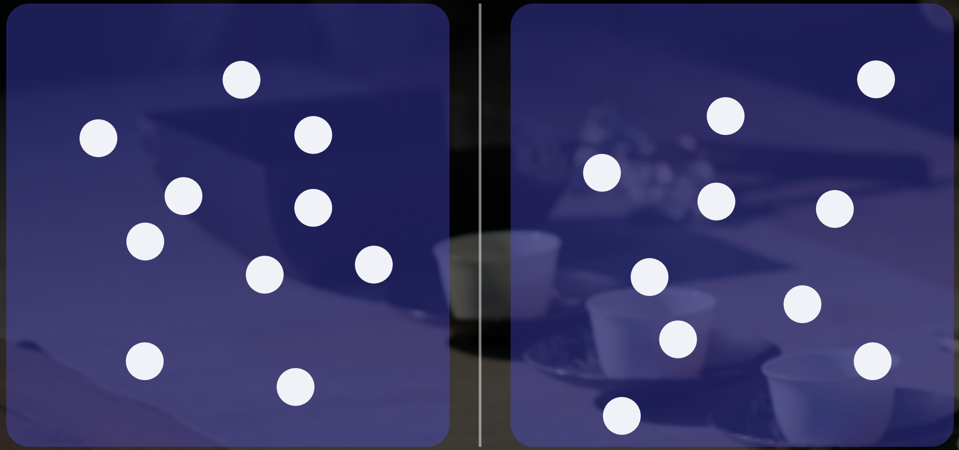
Data Transformation

Learning Objectives

- Understand what is data splitting and how to do it.
- Understand overfitting and solutions to overfitting.
- Understand what is validation set and how to use it.

EXAMPLE. to build a tea detection application

Green tea / Oolong tea



Training set

Test set

- We need to split the dataset into training set and test set.
- We need to keep them separate, as we don’t want the model to memorise the questions instead of learning from the data.

EXAMPLE. to build a tea detection application

Green tea / Oolong tea

Randomise instances

- Before splitting the dataset, we must randomise it.
- We don't want the order of the instances, which is irrelevant, to affect the model training process.

How large should we make different splits?

The larger Training Set

the better model we will be able to learn.

The larger Test Set

the better we will be able to have confidence in evaluation metrics, and tighter confidence intervals.

For now... make sure our Test Set meets the following 2 conditions:

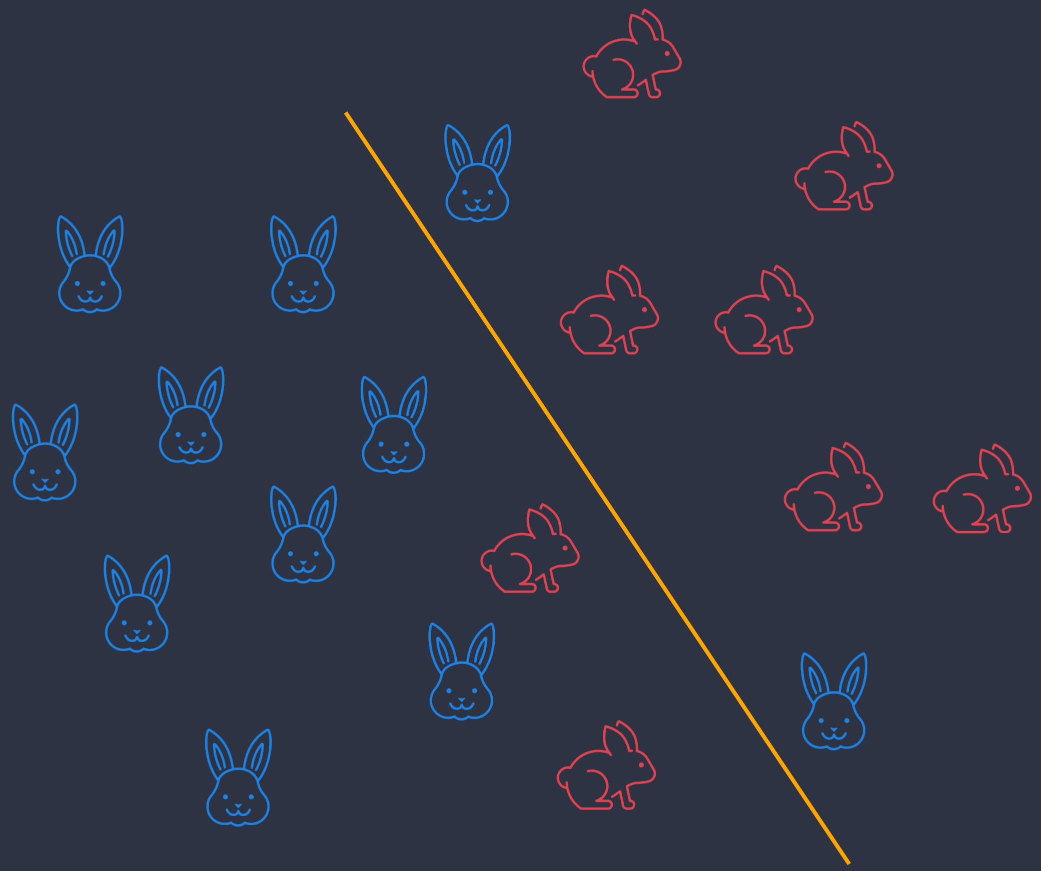
- large enough to yield statistically meaningful results.
- representative of the dataset as a whole. In other words, don't pick a Test Set with different characteristics than the Training Set.

EXAMPLE. Cute bunny detector

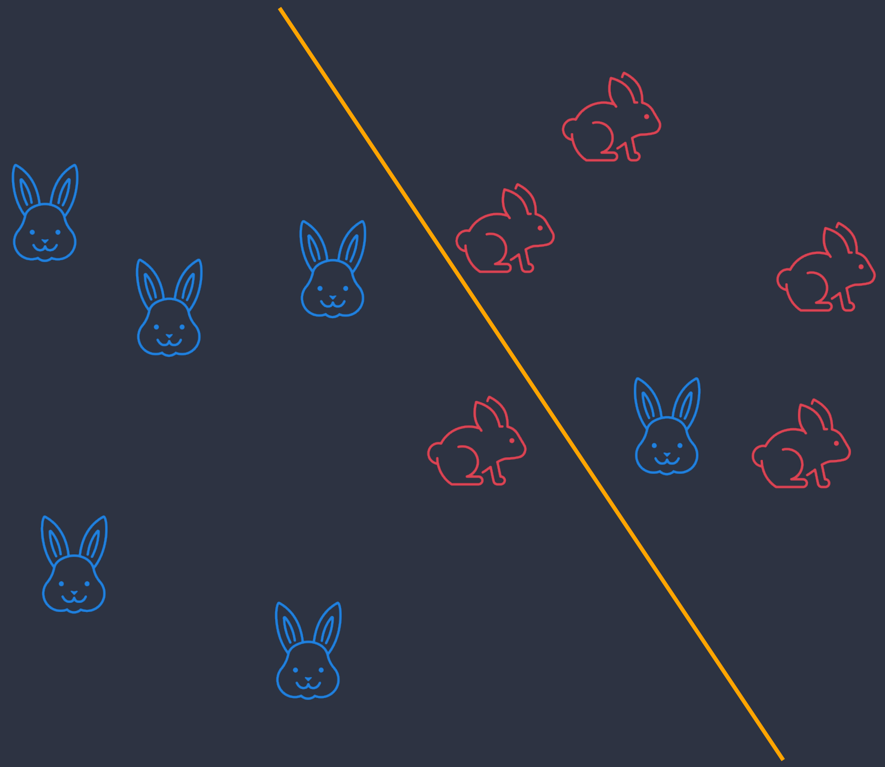
Cute bunny

Not-so-cute bunny

EXAMPLE. Cute bunny detector



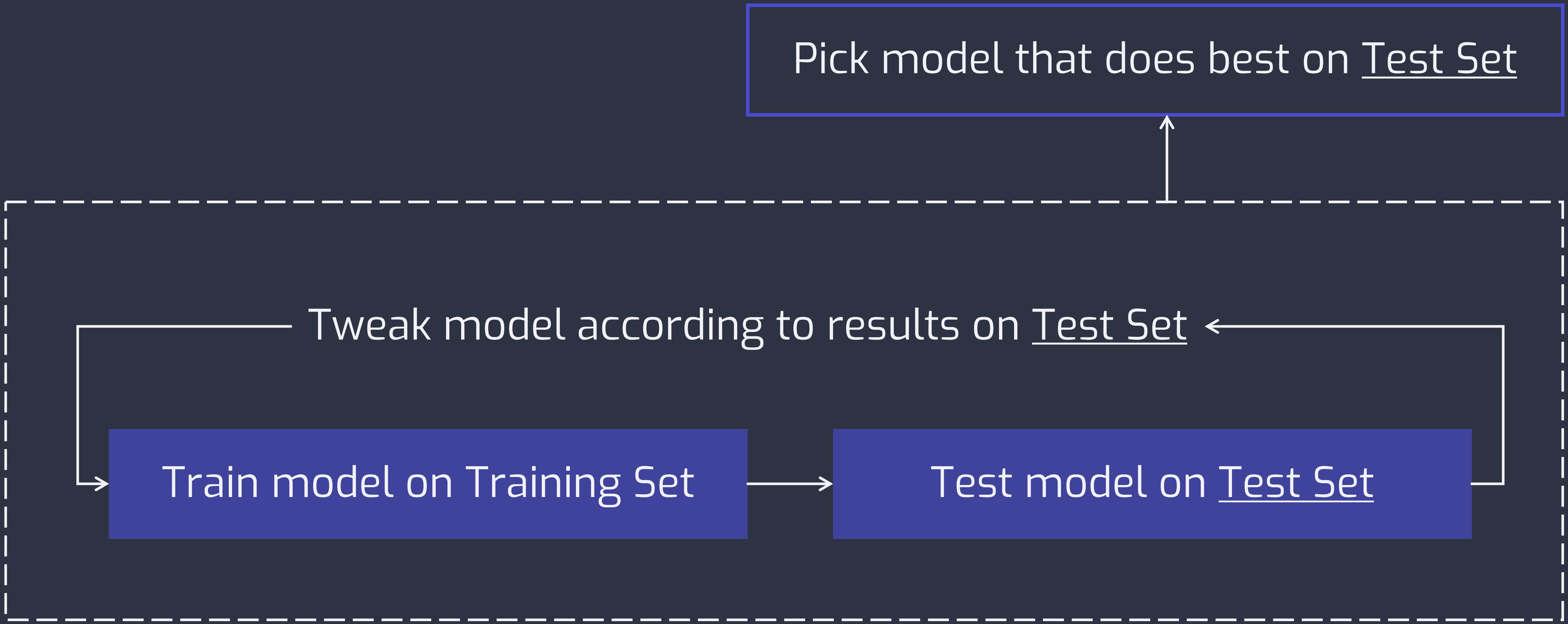
Training set



Test set

(as a proxy for new data)

Overfitting



Pick model that does best on Test Set

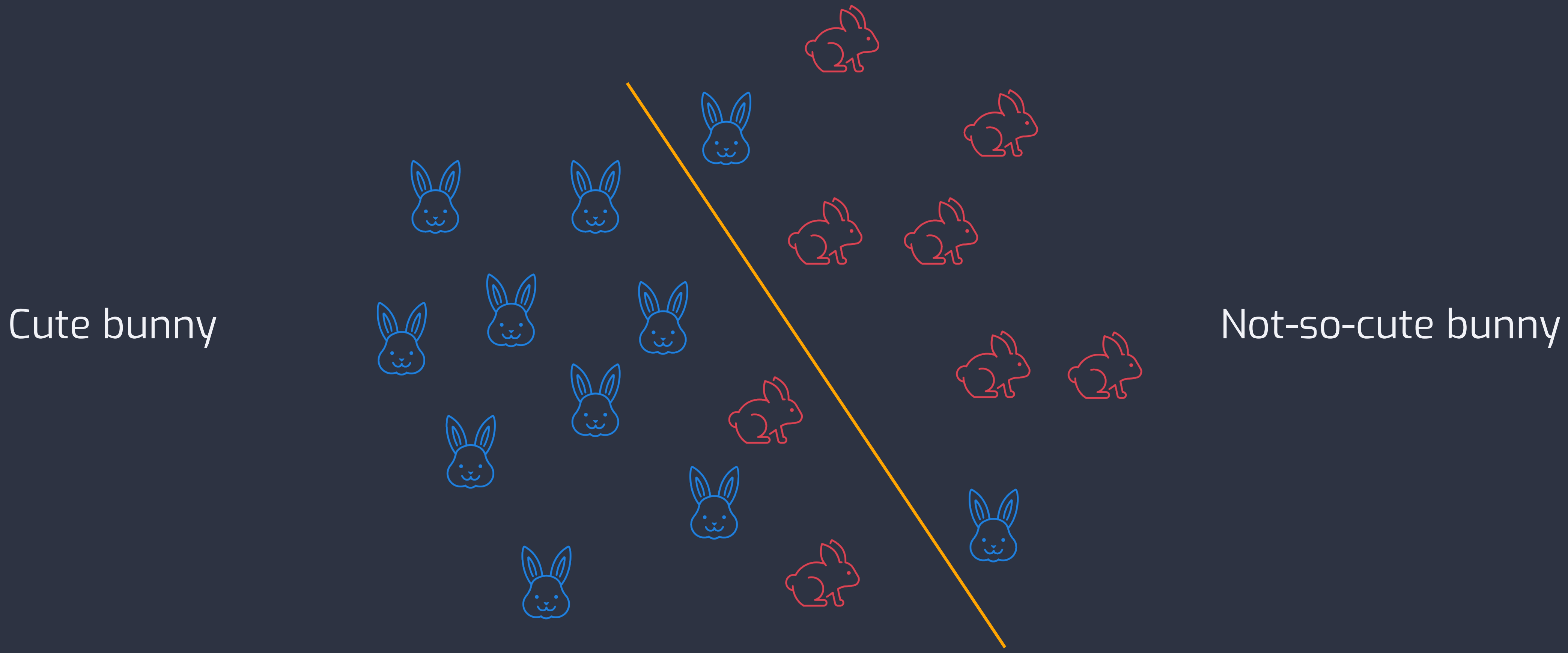
Tweak model according to results on Test Set

Train model on Training Set

Test model on Test Set

Overfitting

The result of learning corresponds too closely or exactly to a particular dataset, and may thus fail to fit previously unseen data or make reliable predictions.

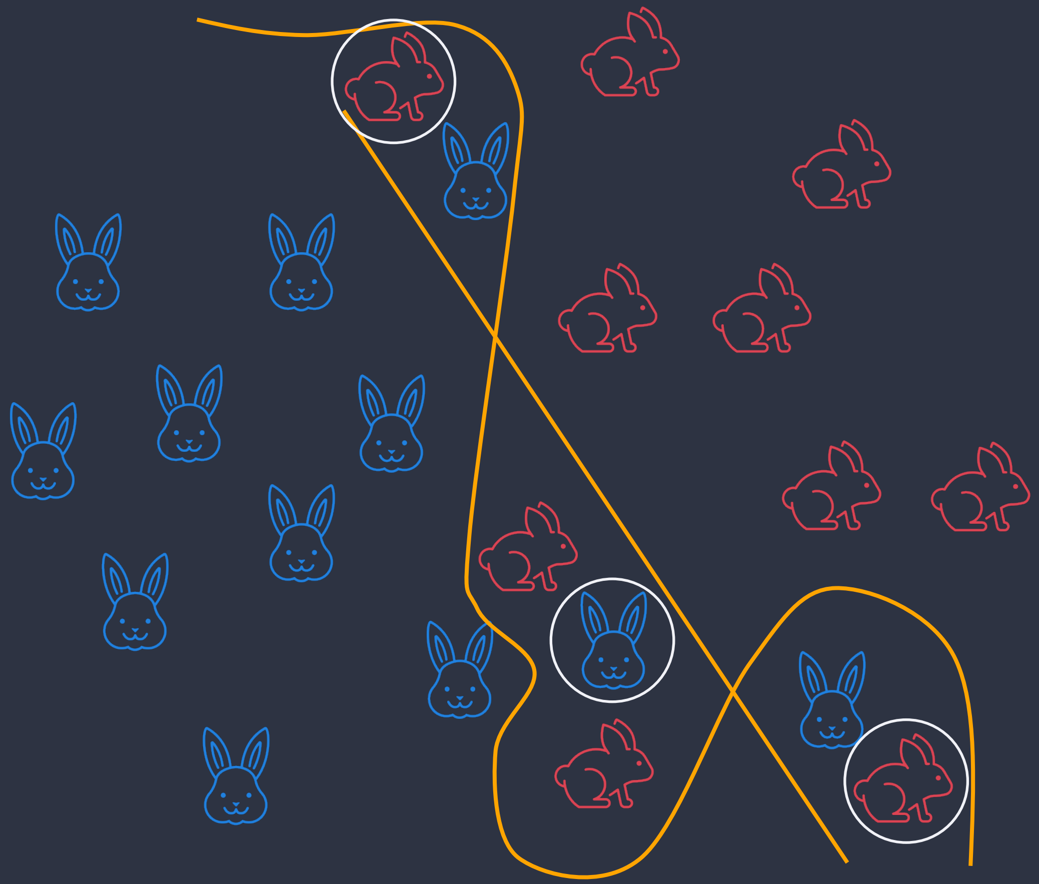


Cute bunny

Not-so-cute bunny

Overfitting

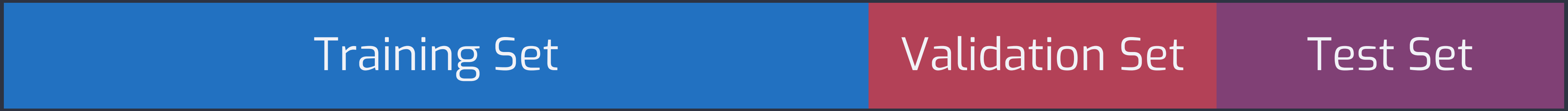
The result of learning corresponds too closely or exactly to a particular dataset, and may thus fail to fit previously unseen data or make reliable predictions.



Cute bunny

Not-so-cute bunny

A solution to Overfitting

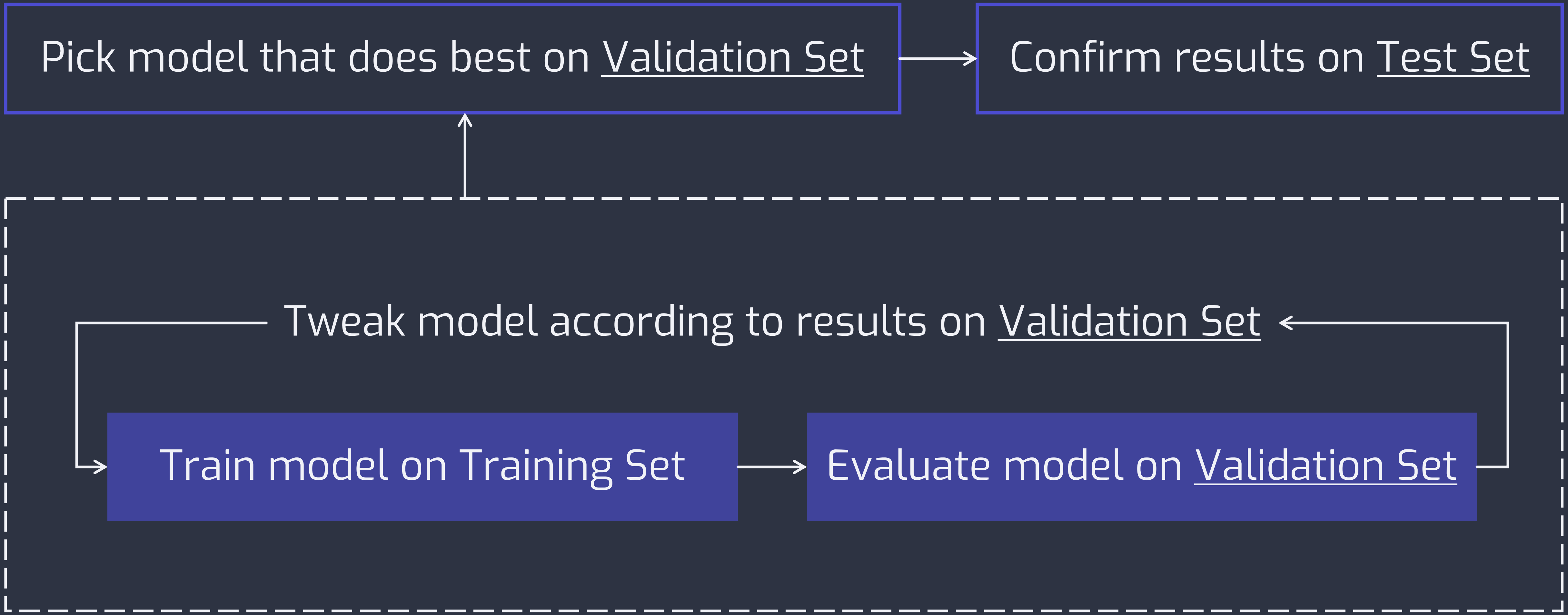


Training Set

Validation Set

Test Set

A solution to Overfitting



Pick model that does best on Validation Set

Confirm results on Test Set

Tweak model according to results on Validation Set

Train model on Training Set

Evaluate model on Validation Set

A solution to Overfitting

To summarise:

1. Put the Test Set aside and completely unused.
2. Pick the model that works best on the Validation Set.
3. Double-check that model against the Test Set.

This is a better approach because it creates <u>fewer exposures to the Test Set</u>.

Why need Validation Set AND Test Set both to evaluate model?

Why need validation set AND test set both to evaluate model?

Validation Set

Compare hyperparameter combinations

- We want to train a model whose performance depends on a set of hyperparameters e.g. learning rate.
- Validation Set is used to evaluate model performance for different combinations of hyperparameter values.

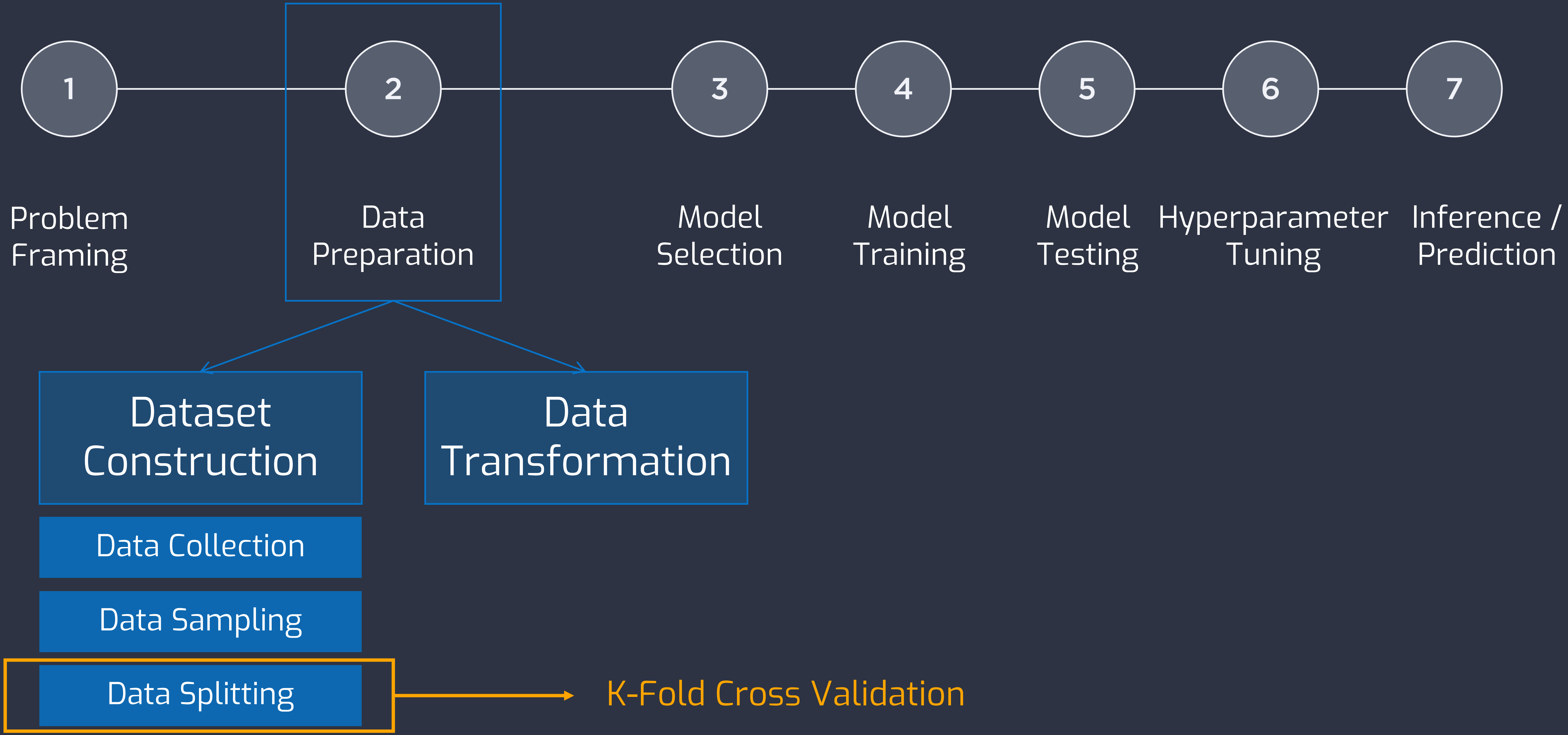
Test Set

Compare different models

- We want to compare trained models in an unbiased way, by comparing model performance using unseen data.
- Test Set is kept apart from the training process, thus being the unseen data, for comparing different trained models.

✓ Takeaway Points

- Need to split the dataset into a Training Set and a Test Set and keep the Test Set completely separate from the training process.
- Need to ensure the chosen sample does not lose statistical significance with respect to the whole population.
- Both Validation Set and Test Set are to evaluate the model, but the Validation Set is for tuning hyperparameters, and the Test Set is for comparing different trained models.



1 Problem Framing

2 Data Preparation

3 Model Selection

4 Model Training

5 Model Testing

6 Hyperparameter Tuning

7 Inference / Prediction

Dataset Construction

- Data Collection
- Data Sampling
- Data Splitting

Data Transformation

Data Splitting → K-Fold Cross Validation