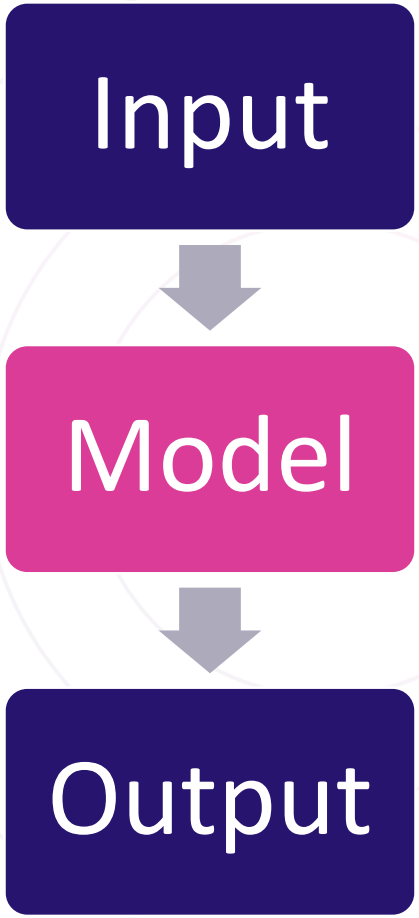
4 Basic Concepts

Model

- Software that maps input to output
  - Complicated calculator
- ML learns this mapping from data
- Mathematical formula that tries to capture real-world behavior
  - “All models are wrong, but some are useful”



Data

- Collection of information on one or multiple observation(s)

- Structured data (20%)
  - Tabular format with rows and columns
  - Examples: numbers, dates or strings
  - Stored efficiently in relational databases

- Unstructured data (80%)
  - Any digital format
  - Examples: text, image or audio
  - Requires more storage space

Structured data table

- Rows represent observations
- Columns containing information
  - Target vs features

	Feature 1	Feature 2	...	Target
Observation 1	Value 11	Value 12	...	Target value 1
Observation 2	Value 12	Value 22	...	Target value 2

	Age	Education	...	Employed
Tom	19	High School	...	no
Jon	45	Masters	...	yes

Features

- Information that you use to model/predict the target

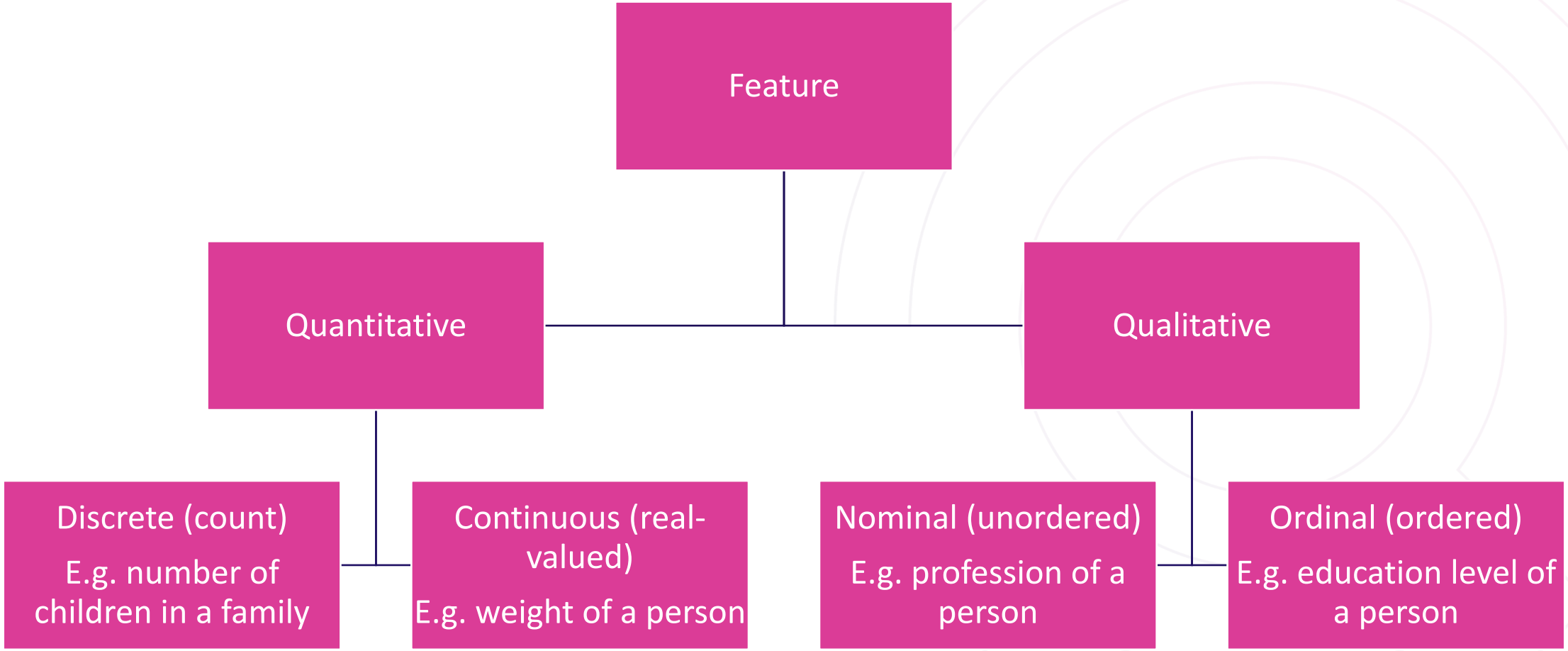
- Quantitative features
  - Can take any value in a range

- Qualitative features
  - Only a selected number of options

	Feature 1	Feature 2	...	Target
Observation 1	Value 11	Value 12	...	Target value 1
Observation 2	Value 12	Value 22	...	Target value 2

	Age	Education	...	Employed
Tom	19	High School	...	no
Jon	45	Masters	...	yes

Feature types



Target

- Information that you want to model/predict based on the available features

- Regression: quantitative target
  - House price prediction (amount)

- Classification: qualitative target
  - E-mail spam filtering (yes/no)

	Feature 1	Feature 2	...	Target
Observation 1	Value 11	Value 12	...	Target value 1
Observation 2	Value 12	Value 22	...	Target value 2

	Age	Education	...	Employed
Tom	19	High School	...	no
Jon	45	Masters	...	yes

Exercise

Problem

- Will it be cold or hot tomorrow?
- Which percentage score will the student get?
- Will my stock go up or down?
- What will the temperature be?
- Will the student pass or fail the exam?
- Which price will me stock be at?

Regression or classification?

- …
- …
- …
- …
- …
- …
- …

Classification vs regression

Classification problem

- Will it be cold or hot tomorrow?
- Will the student pass or fail the exam?
- Will my stock go up or down?

Regression problem

- What will the temperature be?
- Which percentage score will the student get?
- Which price will my stock be at?

Train vs test data

- Train data
  - Part used to learn model/function that maps features to target
- Test data
  - Part used to evaluate the model
  - Allows to check generalizations

	Age	Education	...	Employed
Tom	19	High School	...	no
Jon	45	Masters	...	yes
...	...	...	...	...
...	...	...	...	...
...	...	...	...	...
..	...	...	...	...

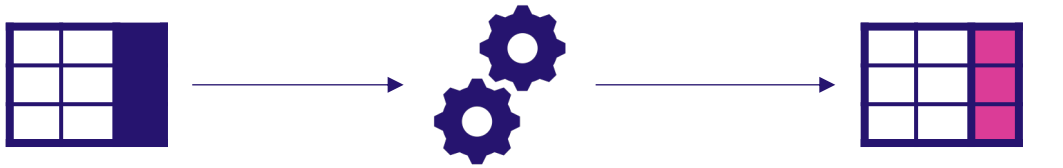
Train data

Data Algorithm Model

- Look for patterns in the data
- Model that captures relation between features and target

	Age	Education	...	Employed
Tom	19	High School	...	no
Jon	45	Masters	...	yes
...	...	...	...	...
...	...	...	...	...
...	...	...	...	...
..	...	...	...	...

Test data



Data Model Predictions

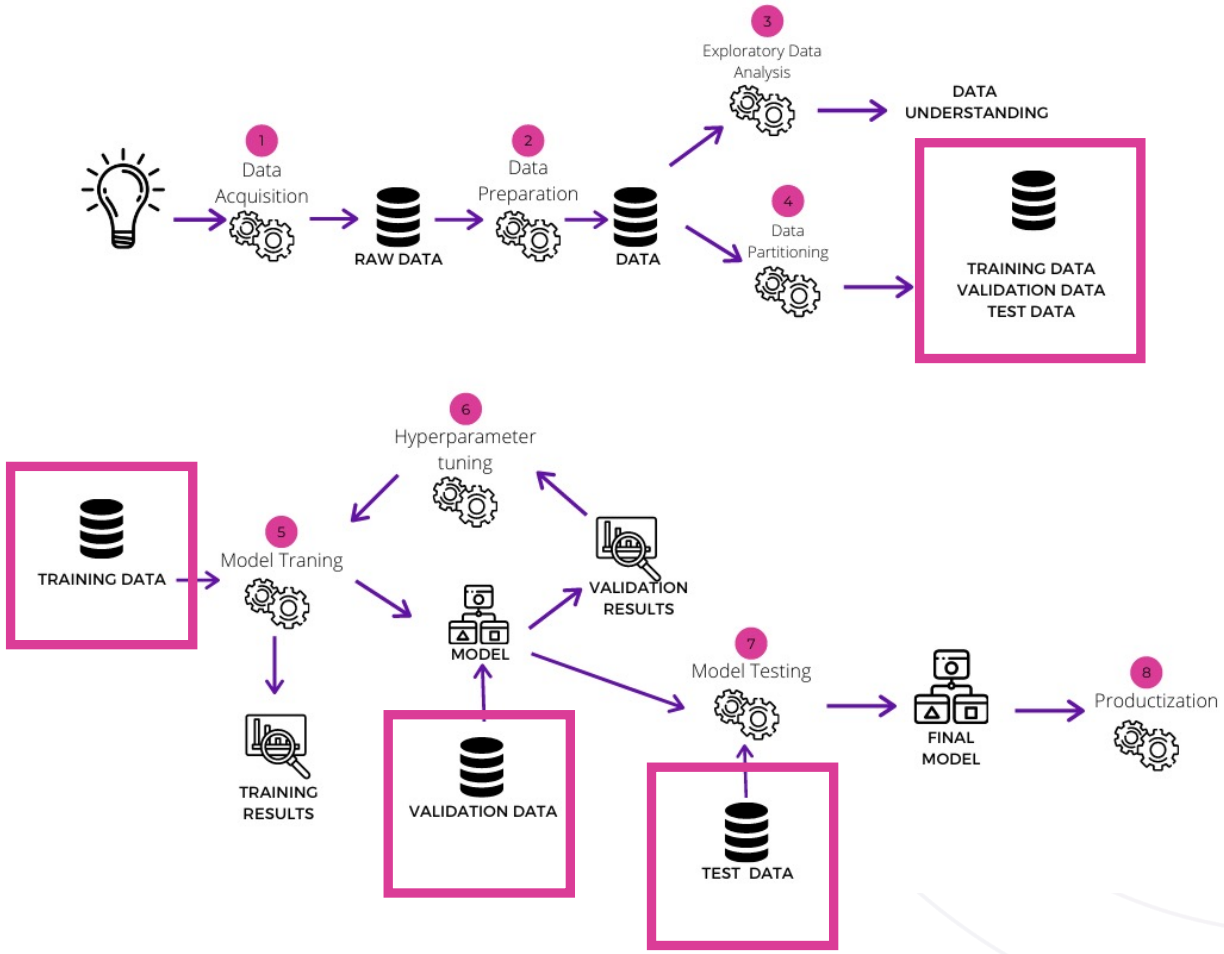
- Run learned model on new data
- Compare original targets with predictions for model evaluation

	Age	Education	...	Employed
Tom	19	High School	...	no
Jon	45	Masters	...	yes
...	...	...	...	...
...	...	...	...	...
...	...	...	...	...
..	...	...	...	...

Validation data

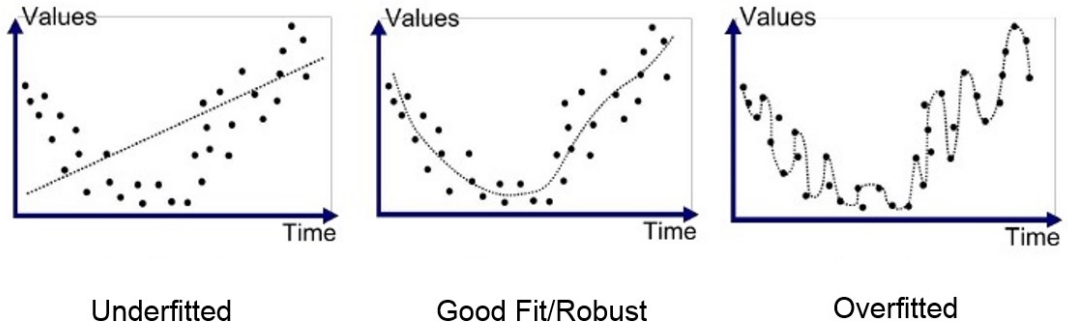
- Part of the training data used for internal model evaluation
- Difference between validation and test data?
- Validation data
  - Evaluation during model development
  - Choose the best model structure and parameter settings
- Test data
  - Evaluation after model development
  - Act as new unseen data and check for generalization performance

Different data partitions



Underfitting vs overfitting

- Data = pattern + noise
  - Want to capture the pattern without the noise
- Underfitting
  - Model too simple to capture the underlying pattern
- Overfitting
  - Model too complex such that it also captures the noise



Underfitting and overfitting explained

Evaluation criteria

Classification

Original target	Predited target	Correct?
1	1	yes
0	1	no
0	0	yes
1	1	yes
1	0	no

- Accuracy of classification
  - 3/5 = 60%

Regression

Original target	Predited target	Difference
15	12	-3
20	23	+3
50	51	+1
35	29	-6
5	9	+4

- Average of squared differences
  - (9+9+1+36+16)/5 = 14.2

Accuracy not always the best choice

- Imagine an image dataset with
  - 20% pictures of dogs
  - 80% pictures of not-dogs

- Model that always predicts not-dog has accuracy of 80%
  - Seems good right?

- However, the model is useless since it did not learn any patterns
  - Simply always predicts not-dog and does not distinguish pictures at all

Metrics for classification

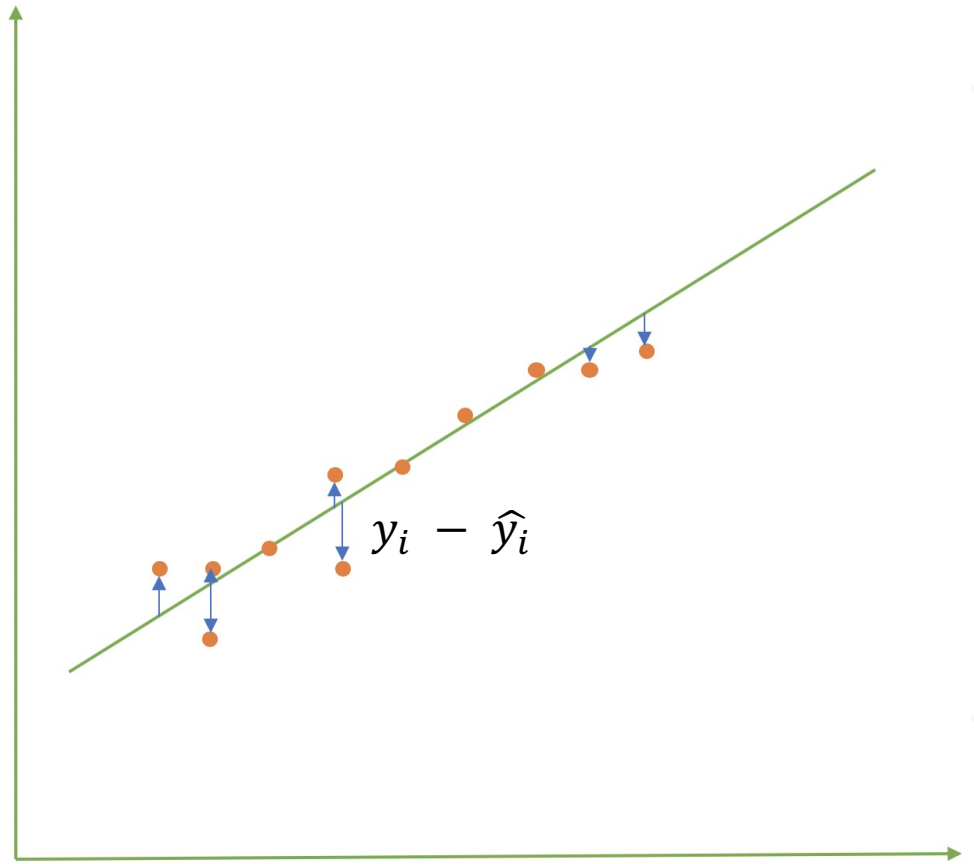
- Confusion matrix

		Prediction	
		Positive	Negative
Actual	Positive	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

- Precision = TP / (TP + FP)
  - Proportion of correctly predicted positive instances among all instances predicted as positive
- Recall = TP / (TP + FN)
  - Proportion of correctly predicted positive instances among all positive instances
- F-score = 2 x (P x R) / (P + R)
  - Combines precision and recall

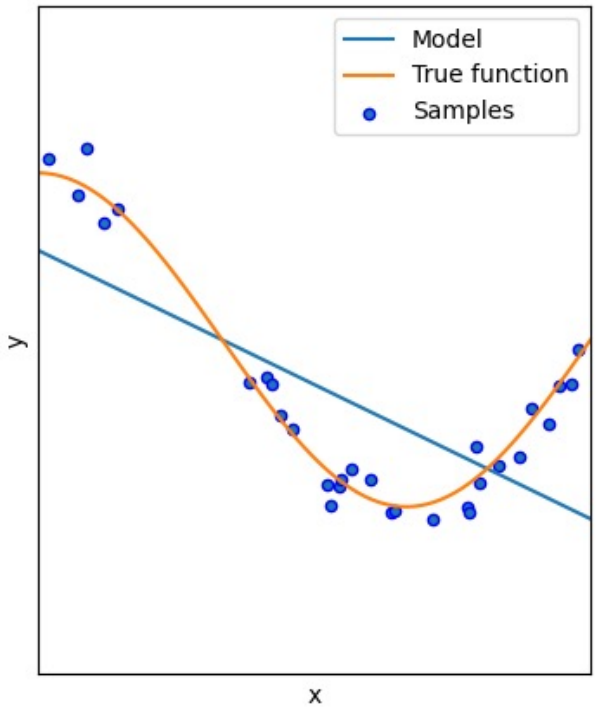
Metrics for regression

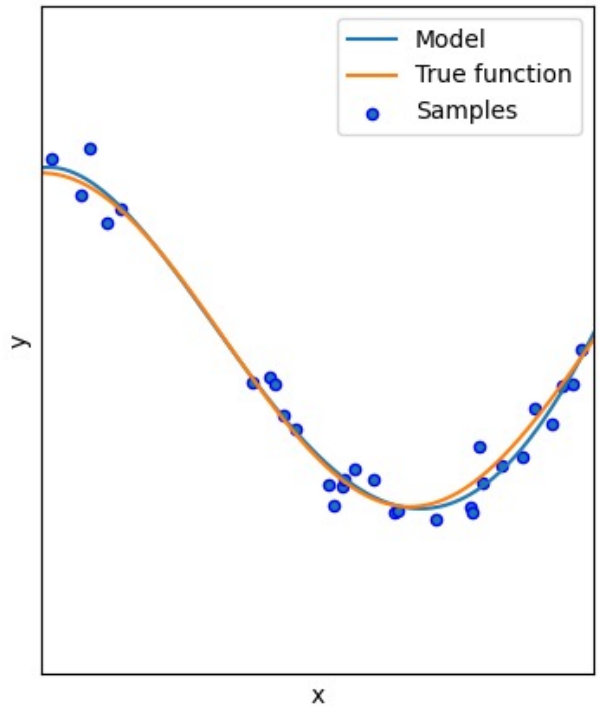
- Mean squared error
  - $MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$
- Mean absolute error
  - $MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|$
- Mean absolute percentage error
  - $MAPE = \frac{100}{N}\sum_{i=1}^{N}\left|\frac{y_i - \hat{y}_i}{y_i}\right|$

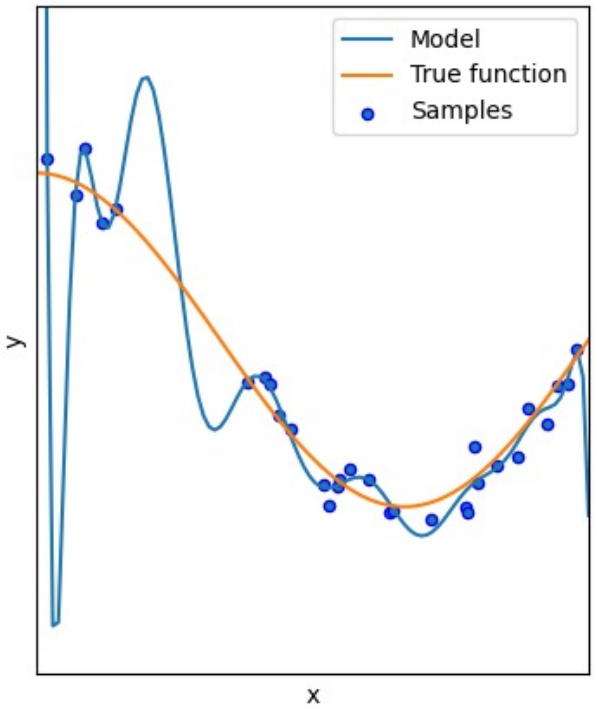


Exercise

Good fit, underfit and overfit?







Python code example