Prediction and overfitting

NHH
TECH3

In-sample and out-of-sample

- In-sample performance
  - Refers to the data used to fit (train) the model
  - Model performance here shows how well the model explains the data it already saw.
  - Good in-sample performance doesn't guarantee generalization.
  - Can be misleading if the model is overfitting.
- Out-of-sample performance
  - Refers to new, unseen data not used during model training.
  - Used to assess how well the model generalizes to new situations.
  - Commonly evaluated using a test set or via cross-validation.
  - Crucial for detecting **overfitting** or poor generalization.

NHH
TECH3

Fitted values vs predictions

- Predictions

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

- Fitted values

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

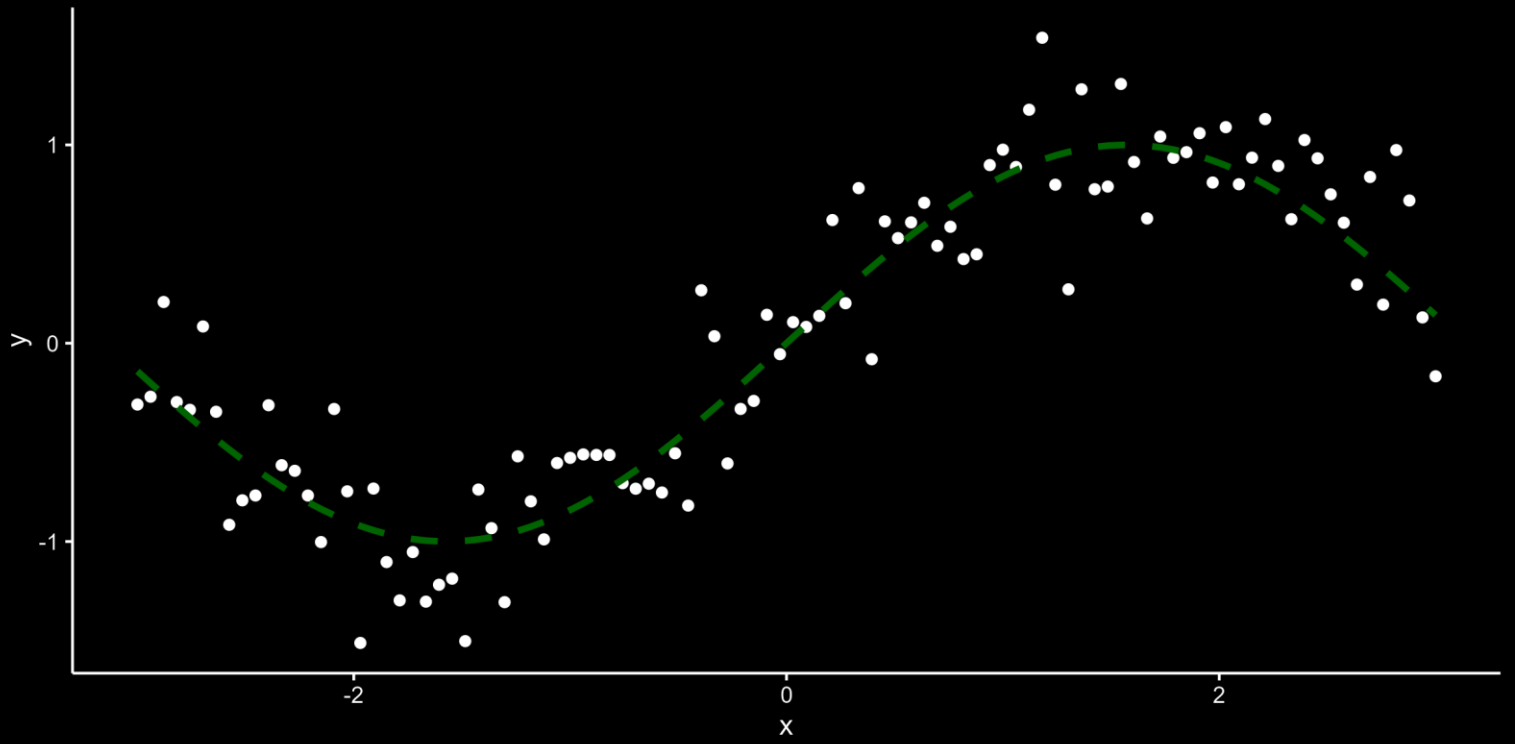
NHH
TECH3

What is prediction?

- Ability to estimate the value of some variable in advance of seeing the data
- The fit of a model to the dataset used to obtain the parameters will nearly always be better than the fit of the model to a new dataset

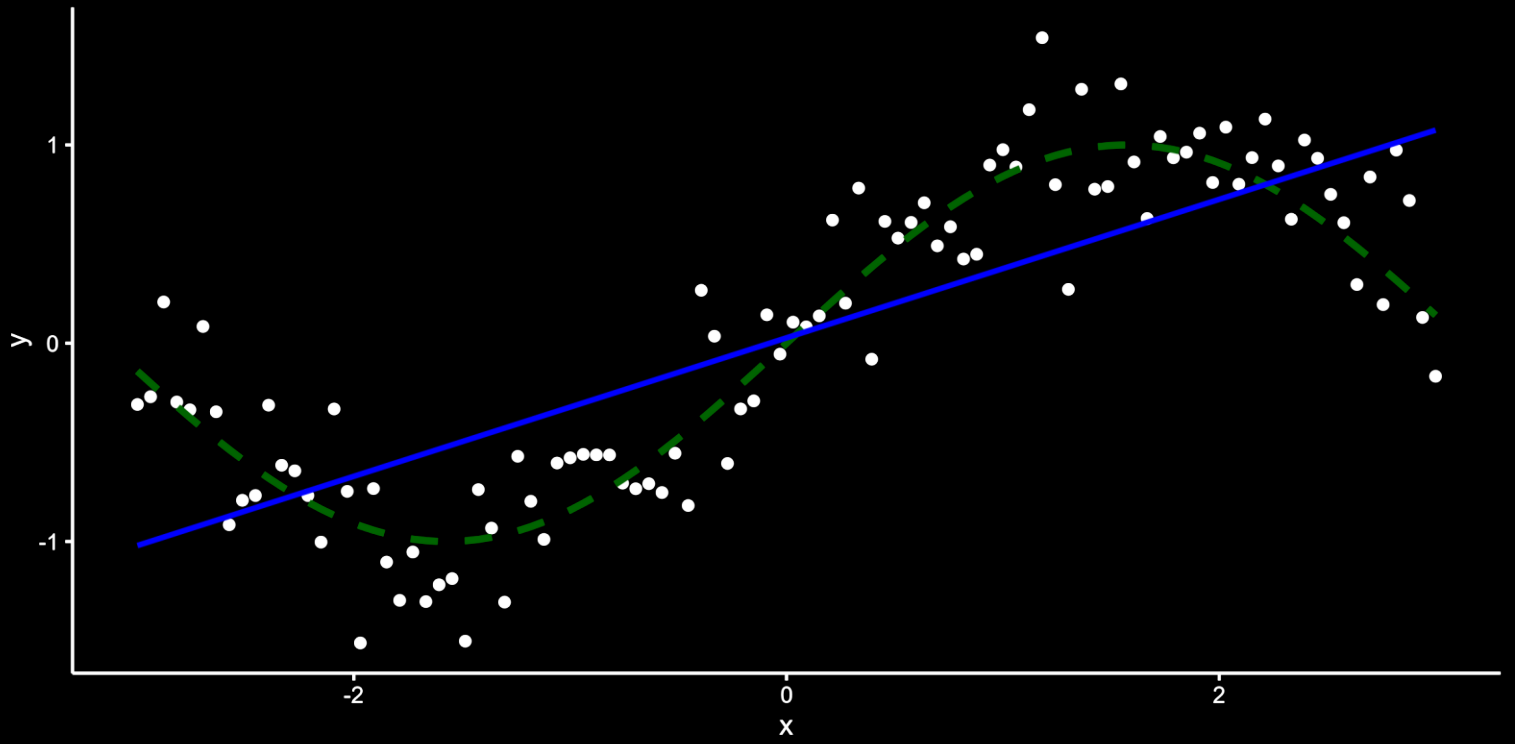
NHH
TECH3

Overfitting



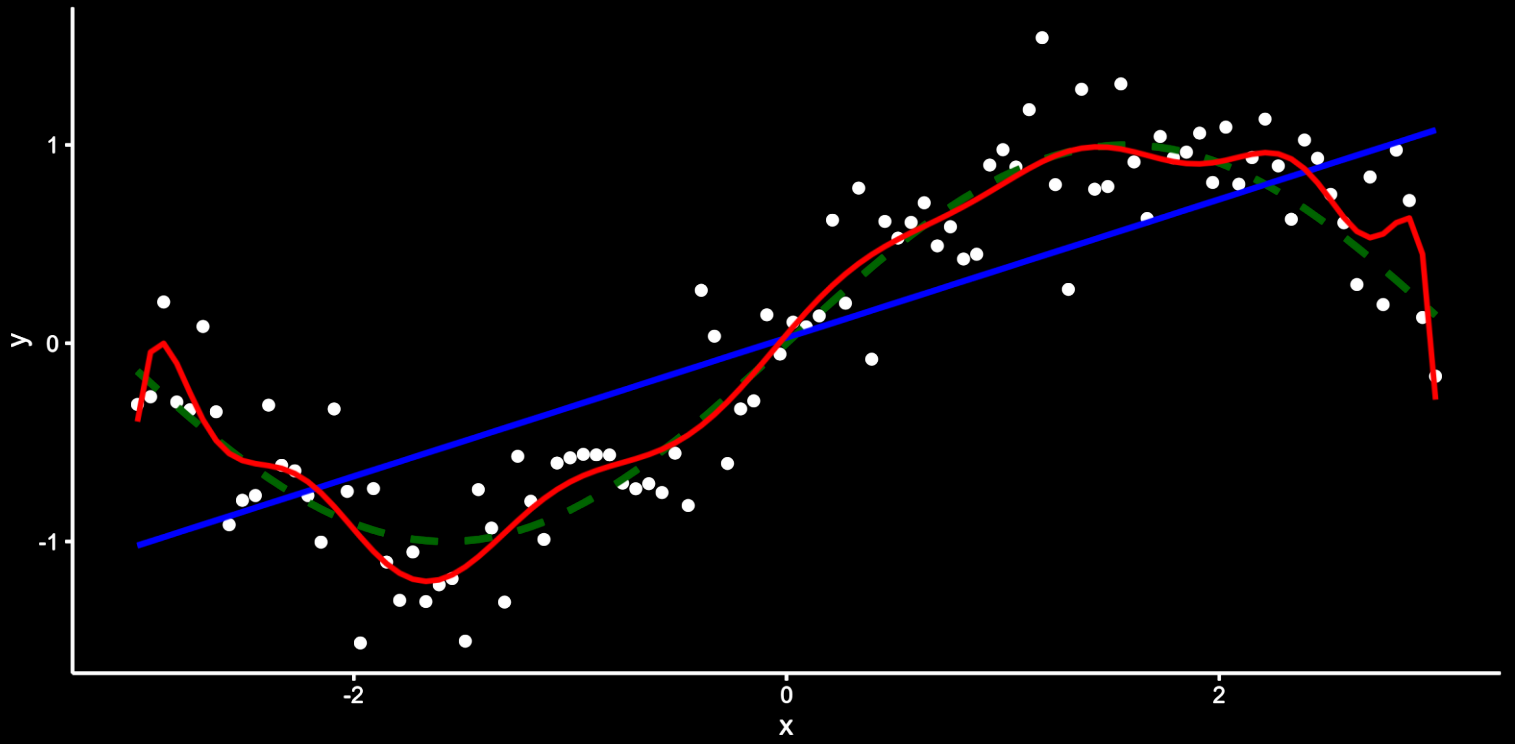
NHH
TECH3

Overfitting



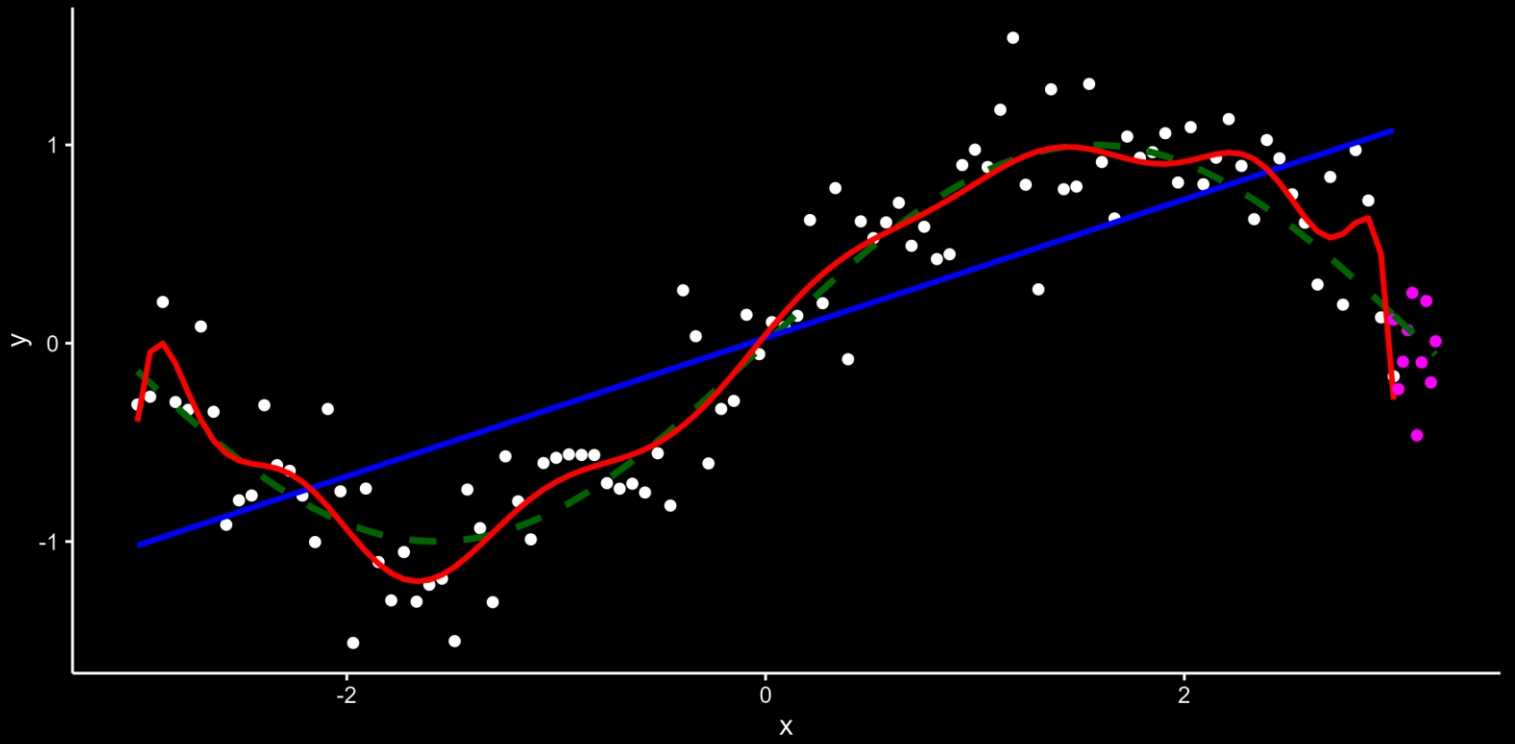
NHH
TECH3

Overfitting



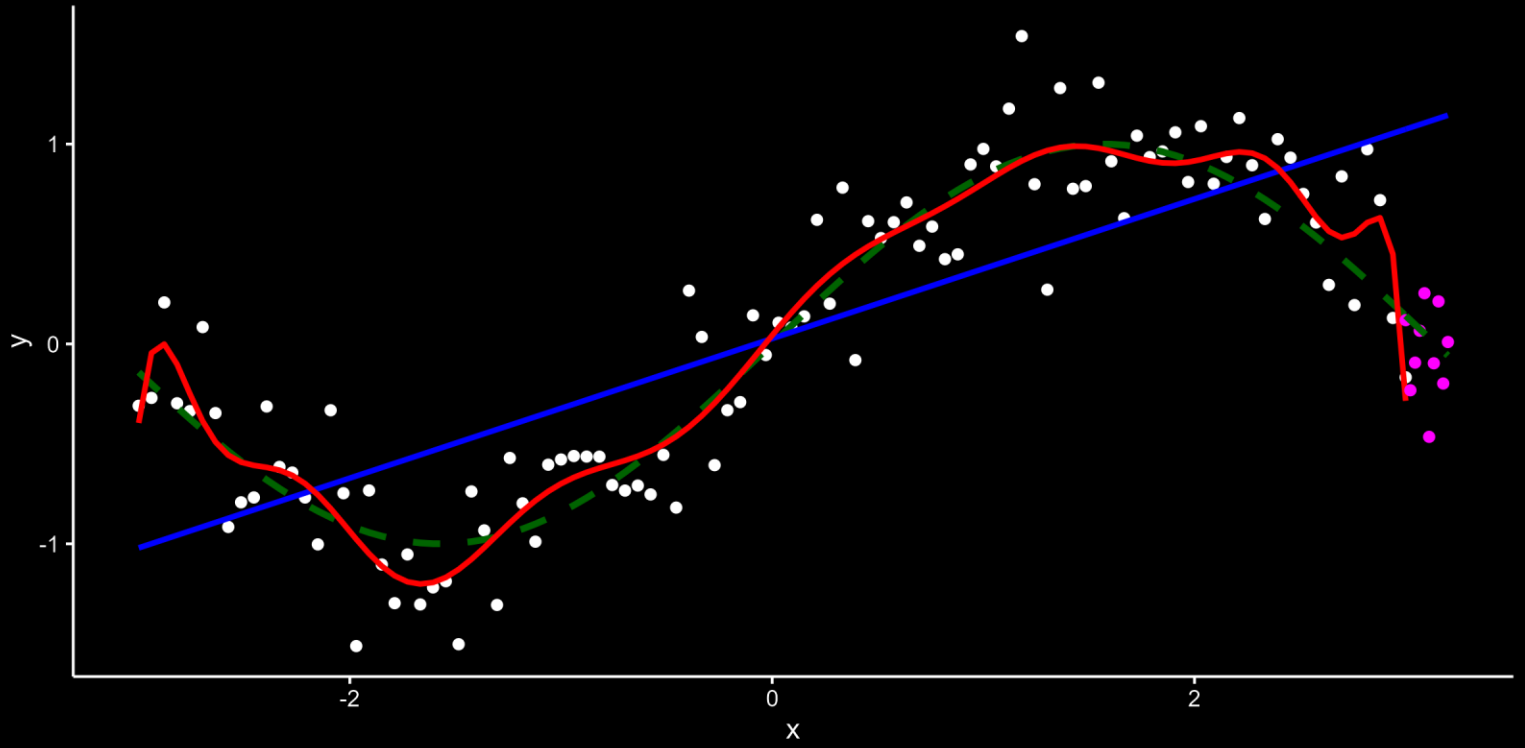
NHH
TECH3

Overfitting



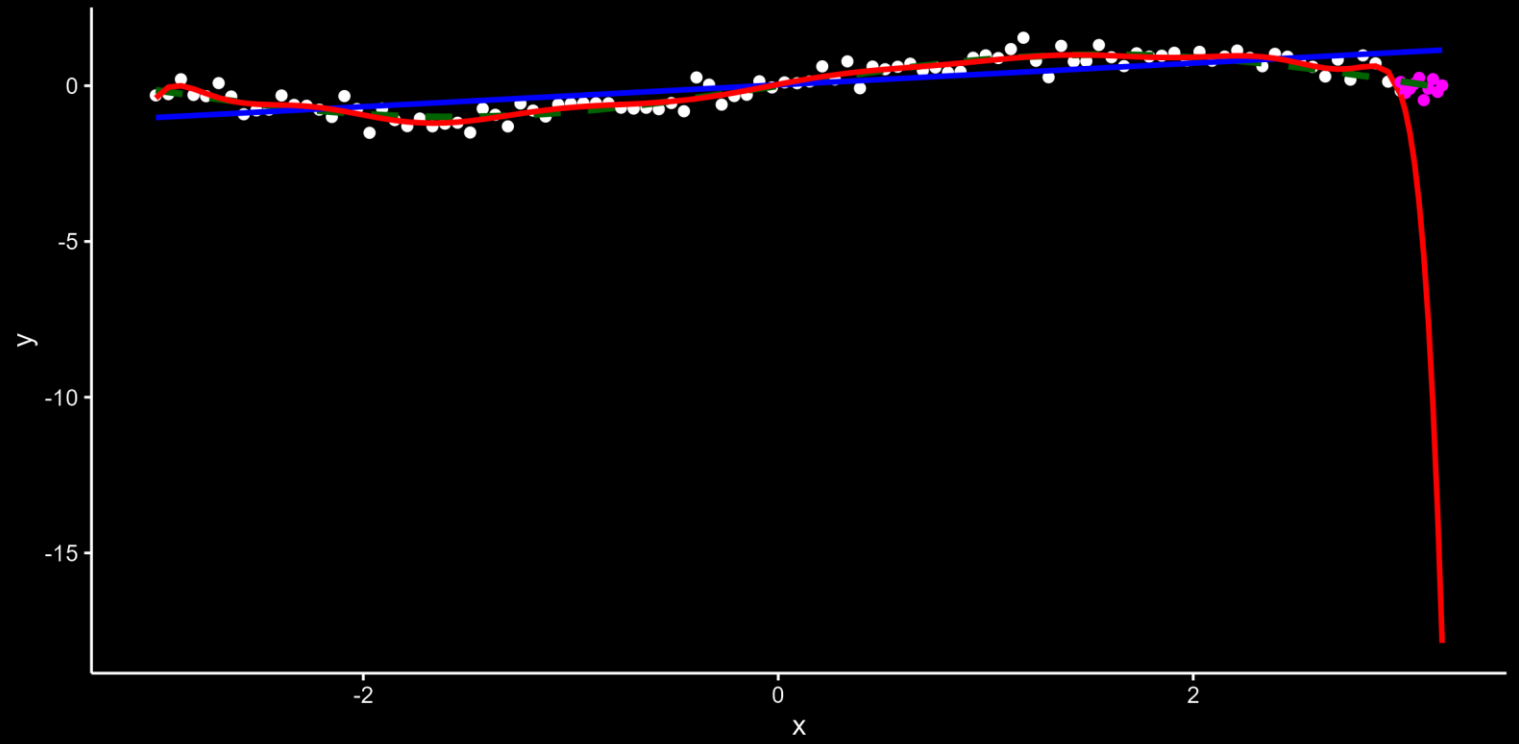
NHH
TECH3

Overfitting



NHH
TECH3

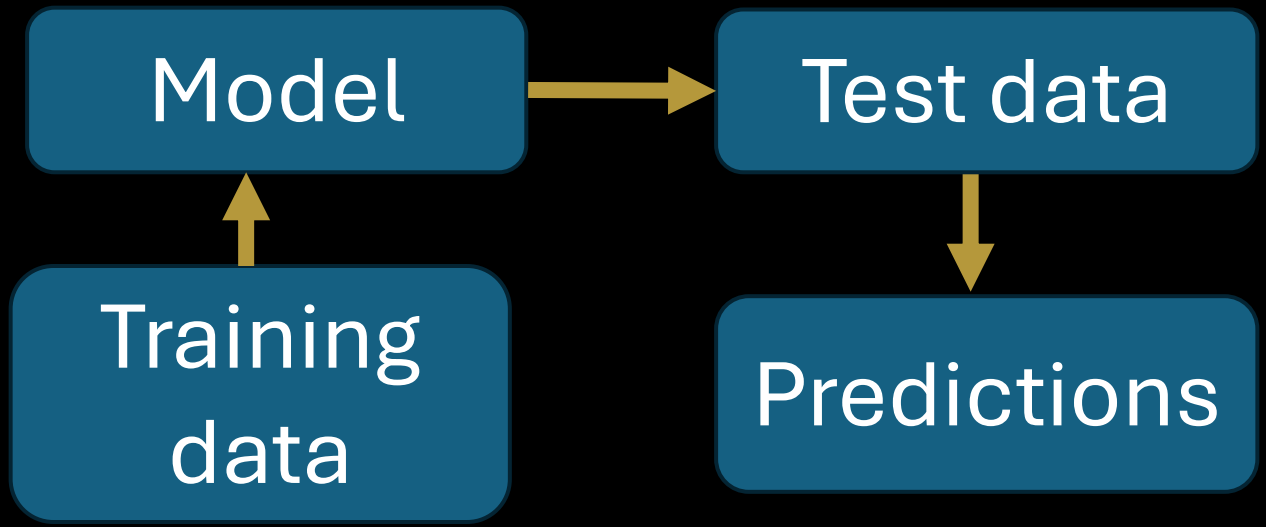
Overfitting



NHH
TECH3

Models that generalize

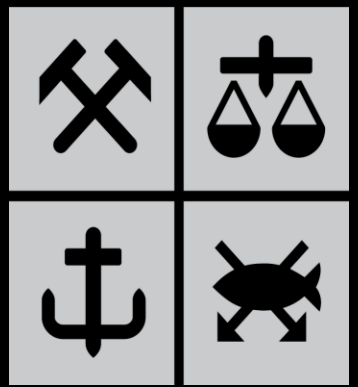
- If you are using the model to predict unseen data, it is important that the model generalize to the new dataset
- It should therefore not overfit the training data



NHH TECH3

NHH

TECH3



Sondre Hølleland
Geir Drage Berentsen