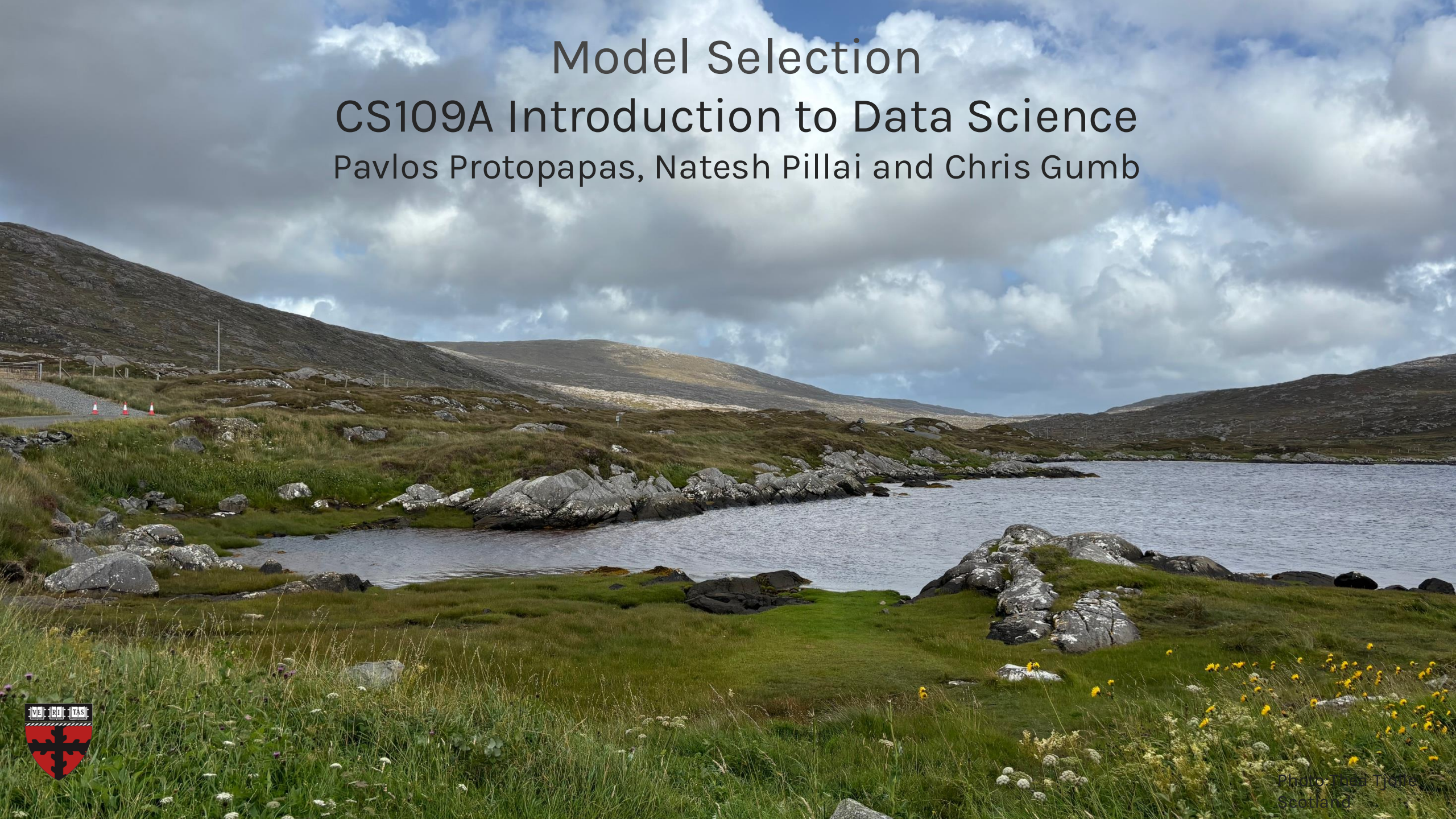
Model Selection

CS109A Introduction to Data Science

Pavlos Protopapas, Natesh Pillai and Chris Gumb



Lecture Outline

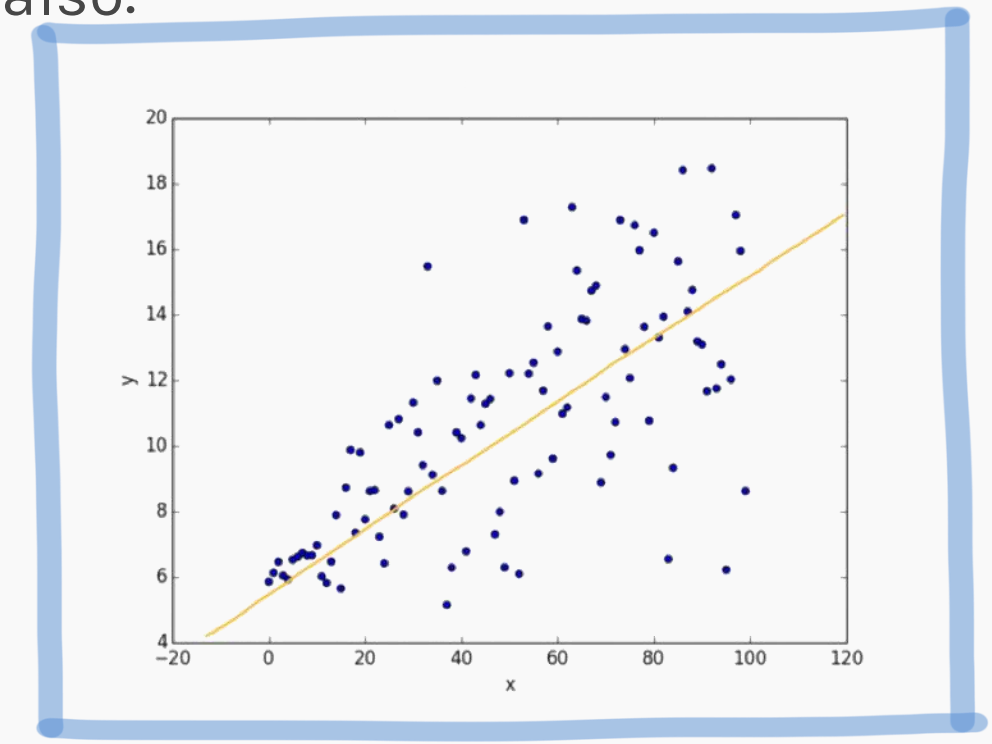
Interaction Effects in Regression Models

Polynomial Regression: Extending Linear Models

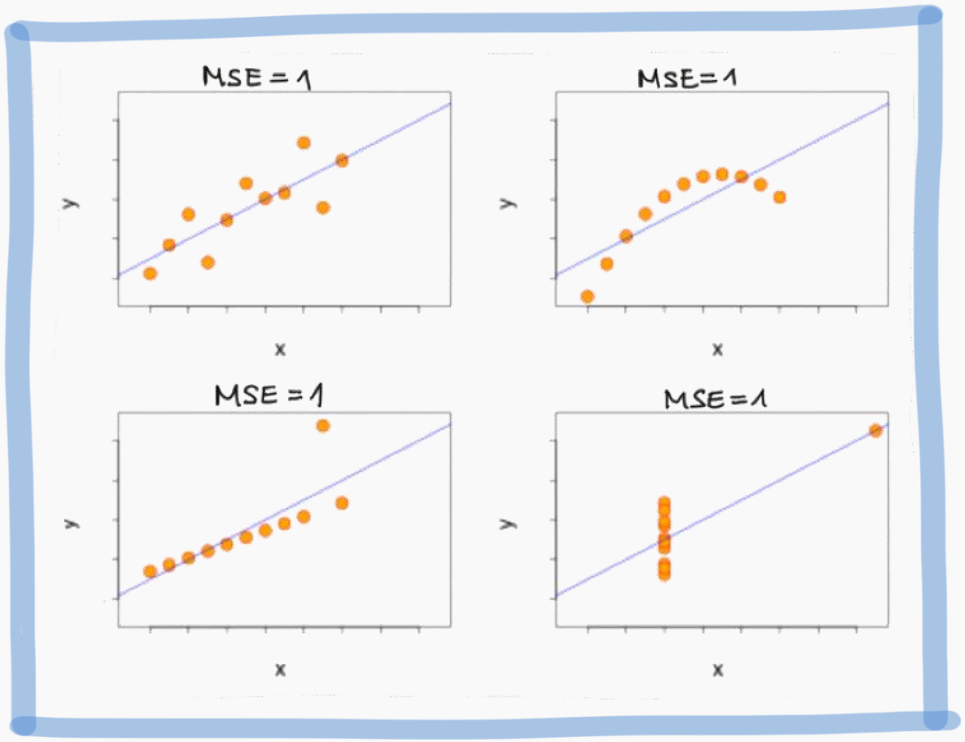
Model Selection Techniques: Focus on Cross-Validation

Evaluation: Training Error

Just because we found the model that minimizes the squared error it doesn't mean that it's a good model. We could investigate the R^2 but also:



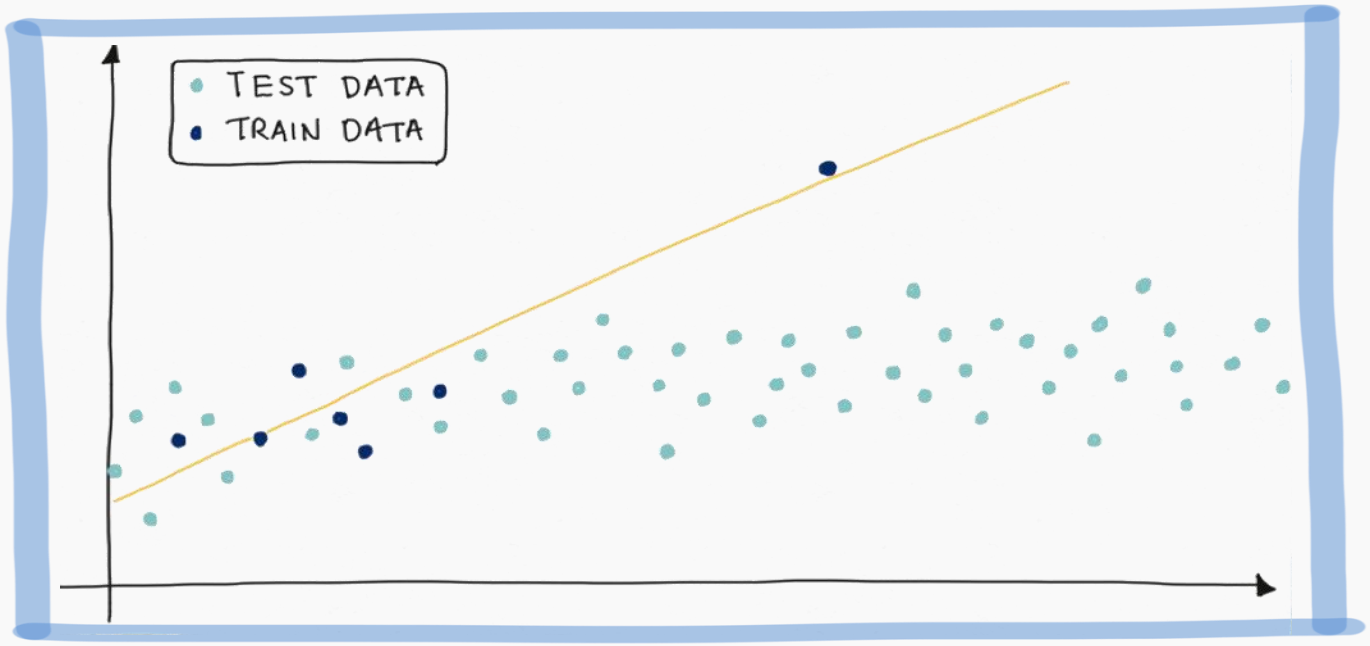
The MSE is high due to noise in the data.



The MSE is high in all four models, but the models are not equal.

Evaluation: Test Error

We need to evaluate the fitted model on new data, data that the model did not train on, the test data.



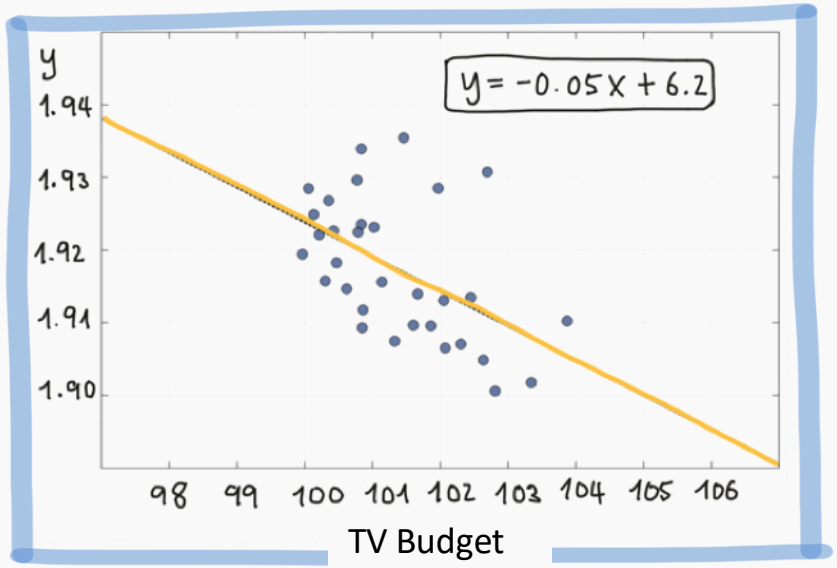
The training MSE here is 2.0 where the test MSE is 12.3.

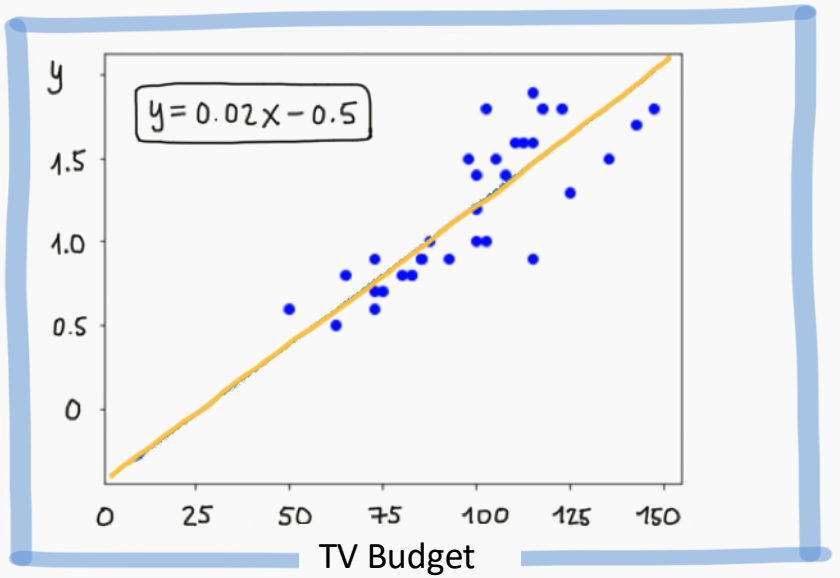
The training data contains a strange point – an outlier – which confuses the model.

Fitting to meaningless patterns in the training is called **overfitting.**

Evaluation: Model Interpretation

For linear models it's important to interpret the parameters





The MSE of this model is very small. But the slope is -0.05. That means the larger the budget the less the sales.

The MSE is very small, but the intercept is -0.5 which means that for very small budget we will have negative sales.

Generalization Error

We know to evaluate the model on both train and test data, because models that do well on training data may do poorly on new data (overfitting).

The ability of models to do well on new data is called generalization.

The goal of model selection is to choose the model that generalizes the best.

Model Selection

Model selection is the application of a principled method to determine the complexity of the model, e.g., choosing a subset of predictors, choosing the degree of the polynomial model etc.

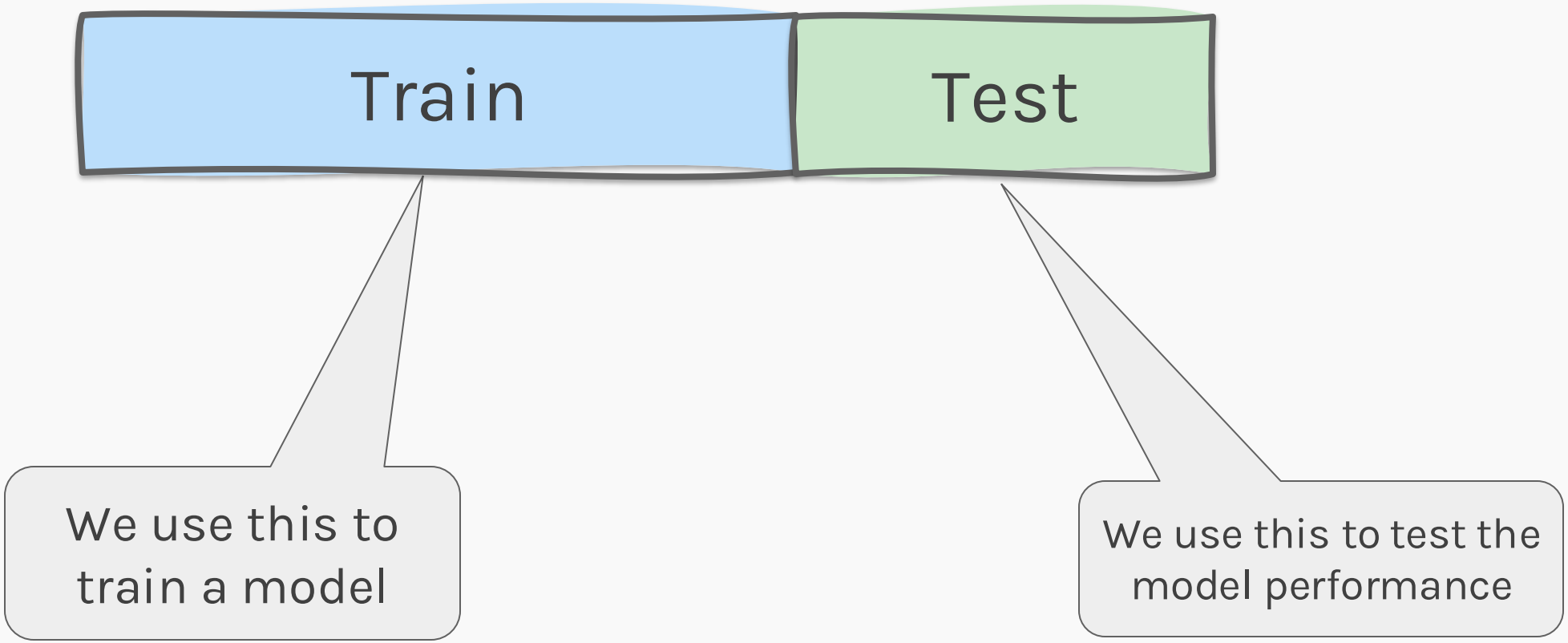
A strong motivation for performing model selection is to avoid **overfitting**, which we saw can happen when:

- there are too many predictors:
  - the feature space has high dimensionality
  - the polynomial degree is too high
  - too many cross terms are considered
- the coefficients values are too **extreme (we have not seen this yet)**

Train-Test split

How do we select a model?

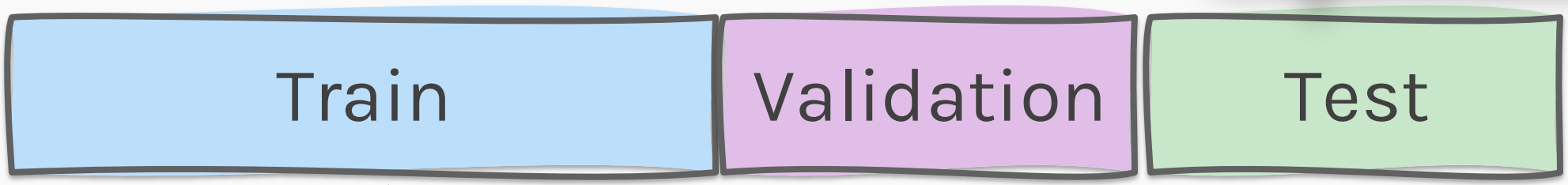
So far, we have been using train/test splits



Train-Validation-Test

We introduce a different sub-set, which we called validat
to select the model.

The test set should never be touched for model training or selection.



We use this to train a model

We use this to select model

We use this to test the model performance

Model Selection

Ways of model selection:

- Exhaustive search
- Greedy algorithms
- Fine tuning hyper-parameters
- Regularization

Model Selection

Ways of model selection:

- **Exhaustive search**
- Greedy algorithms
- Fine tuning hyper-parameters
- Regularization

Model Selection: How many models?

Can you prove this?

Question:

How many different models when considering J predictors (only linear terms) do we have?

Example: 3 predictors (X_1, X_2, X_3)

- Models with 0 predictor:
  - M0:
- Models with 1 predictor:
  - M1: X_1
  - M2: X_2
  - M3: X_3
- Models with 2 predictors:
  - M4: $\{X_1, X_2\}$
  - M5: $\{X_2, X_3\}$
  - M6: $\{X_3, X_1\}$
- Models with 3 predictors:
  - M7: $\{X_1, X_2, X_3\}$

2^J models

Model Selection

Ways of model selection:

- Exhaustive search
- **Greedy algorithms**
- Fine tuning hyper-parameters
- Regularization

Stepwise Variable Selection and Validation

Selecting optimal subsets of predictors (including choosing the degree of polynomial models) through:

- stepwise variable selection - iteratively building an optimal subset of predictors by optimizing a fixed model evaluation metric each time.

- selecting an optimal model by evaluating each model on validation set.

Stepwise Variable Selection: Forward method

In **forward selection**, we find an 'optimal' set of predictors by iterative building up our set.

1. **Start Simple**

Begin with the empty set P_0, construct the null model M_0

2. Add Predictors One by One

For $k = 1, \dots, J$:

2.1 Use the best model so far, M_{k-1}, based on $k-1$ predictors

2.2 Pick one predictor not yet P_{k-1} that improves the model the most according to validation MSE

2.3 Create a new model, M_k, using the selected predictors

3. Select the model M amongst $\{M_0, M_1, \dots, M_J\}$ that optimizes validation MSE

Stepwise Variable Selection Computational Complexity

How many models did we evaluate?

- 1st step, ***J* Models**
- 2nd step, ***J-1* Models** (add 1 predictor out of *J-1* possible)
- 3rd step, ***J-2* Models** (add 1 predictor out of *J-2* possible)
- …

$$O(J^2) \ll 2^J \text{ for large } J$$

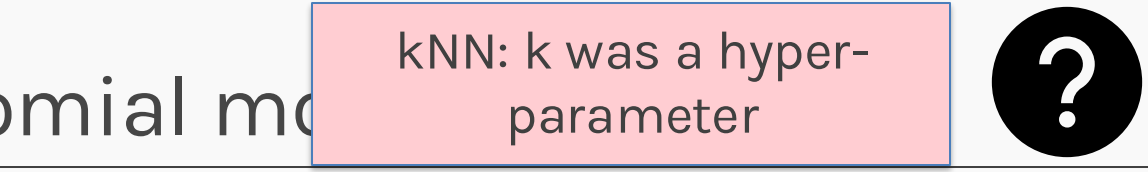
Model Selection

Ways of model selection:

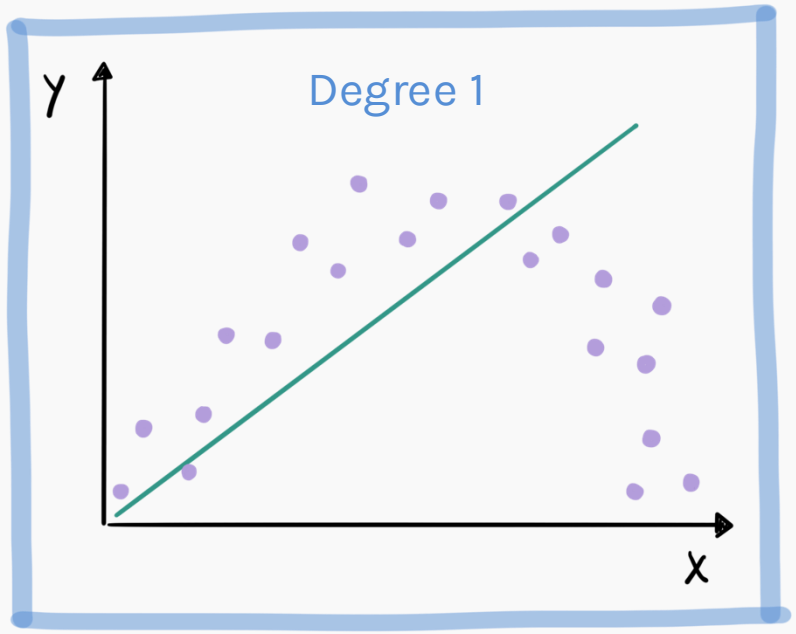
- Exhaustive search
- Greedy algorithms
- **Fine tuning hyper-parameters**
- Regularization

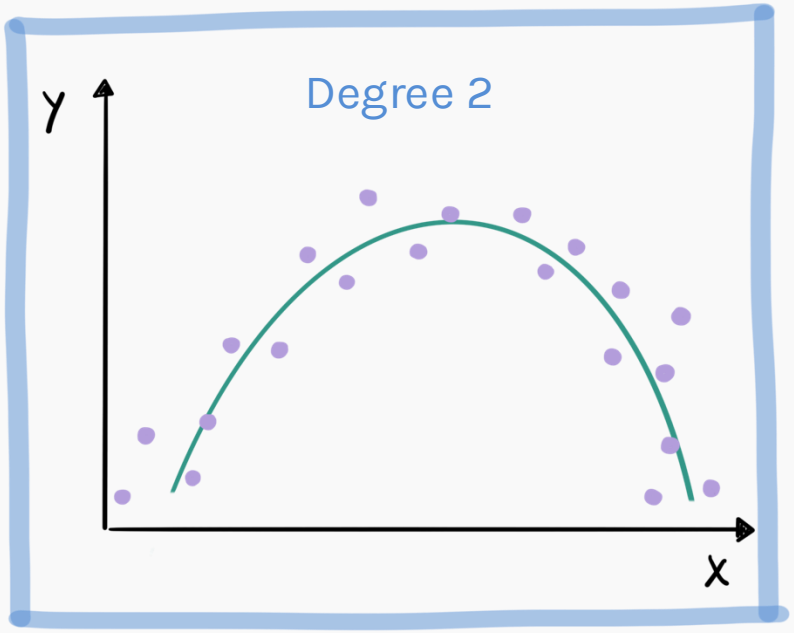
Choosing the degree of the polynomial mo

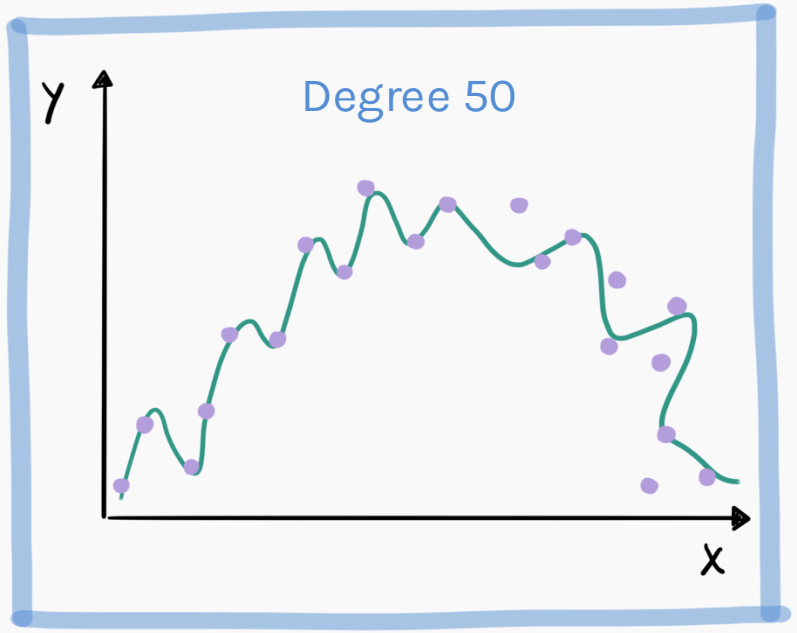
kNN: k was a hyper-parameter



We turn model selection into choosing a hyper-parameter. For example, polynomial regression requires choosing a degree – this can be thought as model selection – and we select the model by tuning the hyper-parameter.





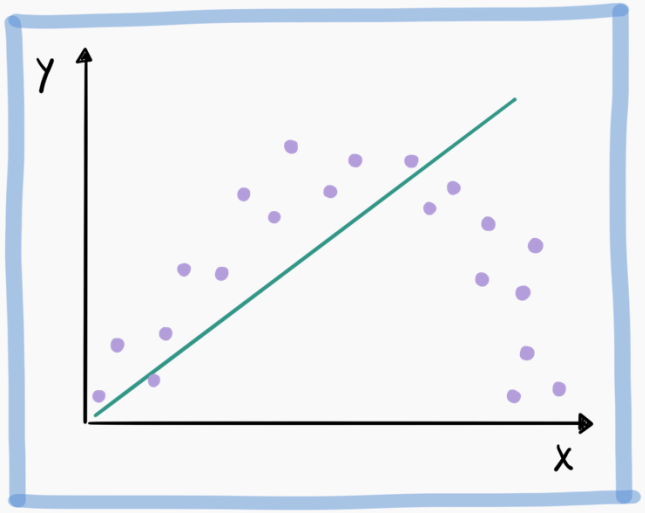


Underfitting: when the degree is too low, the model cannot fit the trend.

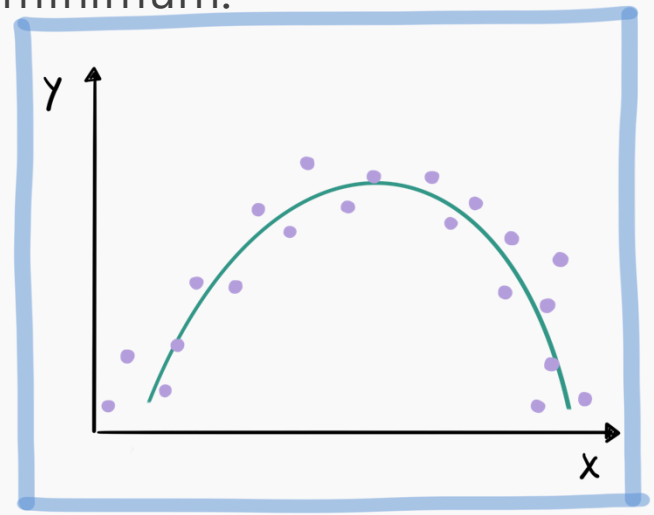
We want a model that fits the trend and ignores the noise.

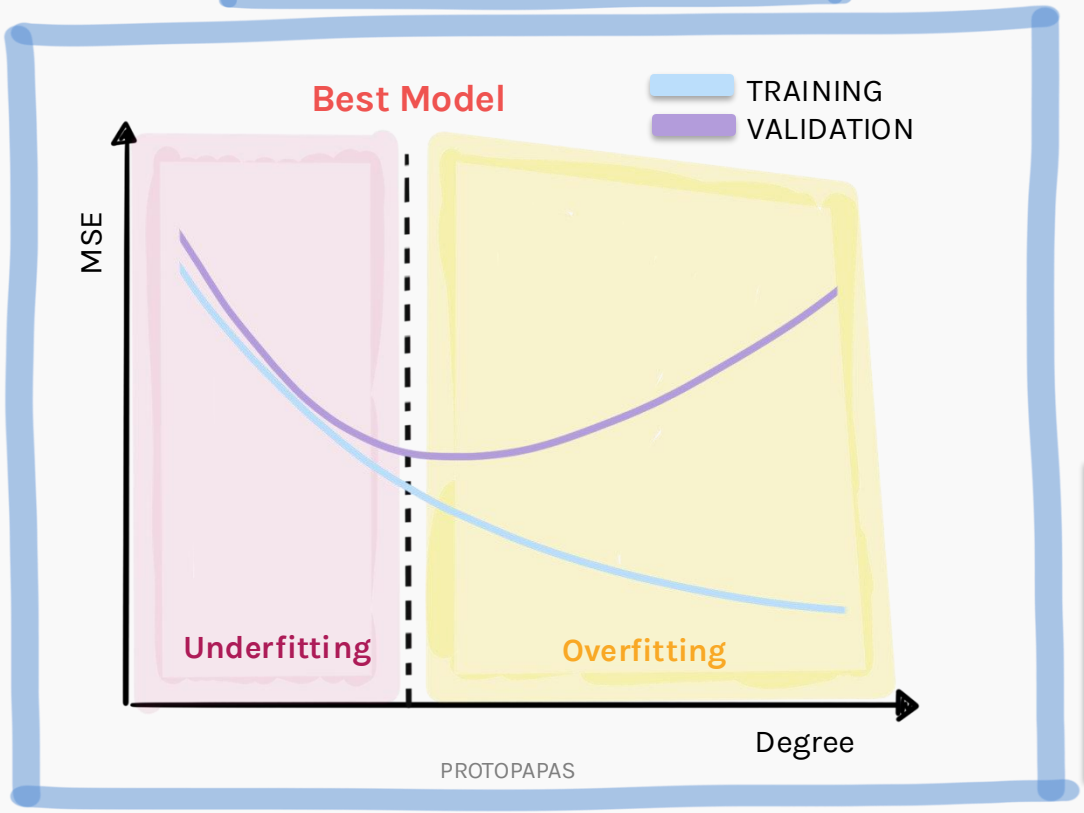
Overfitting: when the degree is too high, the model fits all the noisy data points.

Underfitting: train and validation error is high.

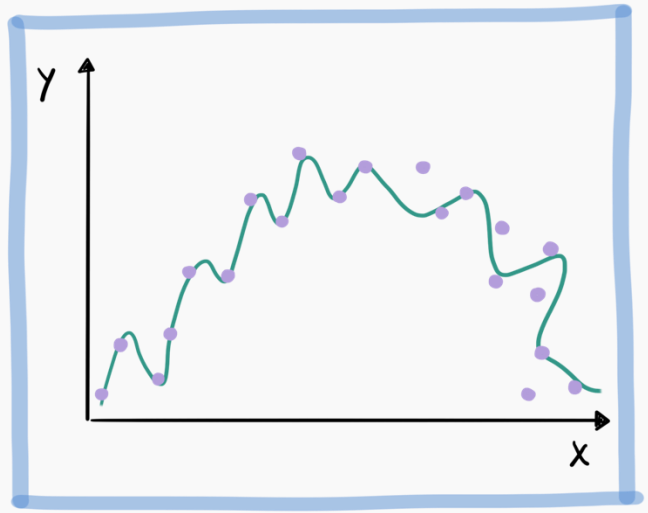


Best model: validation error is minimum.





Overfitting: train error is low, validation error is high.



What are the parameters of the models and what are the hyperparameters?