IDS 702

Cross validation

Cross validation

- The train/test method of model validation is often referred to as cross validation

- Splitting the data randomly into two sets can have a big impact on the train and test MSE, particularly in small samples

- **K-fold cross validation** is a type of cross validation that aims to address this sensitivity

K-fold cross validation

- Split the data into K mutually exclusive groups (folds)

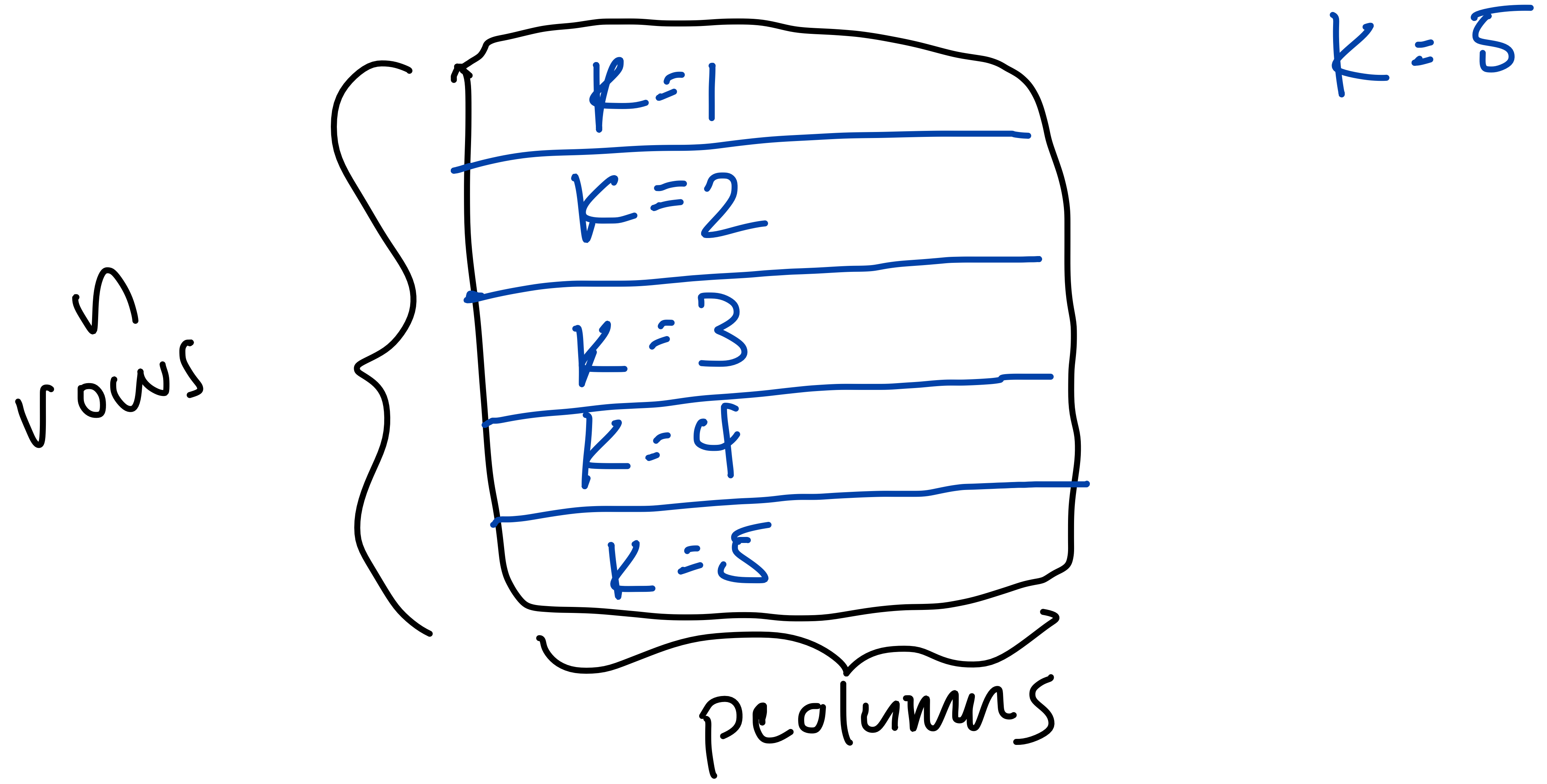
- For the kth fold, with $k = 1,\ldots,K$, fit the model on all the remaining data excluding the kth fold (that is, all the other folds combined) and use the kth fold as the test set

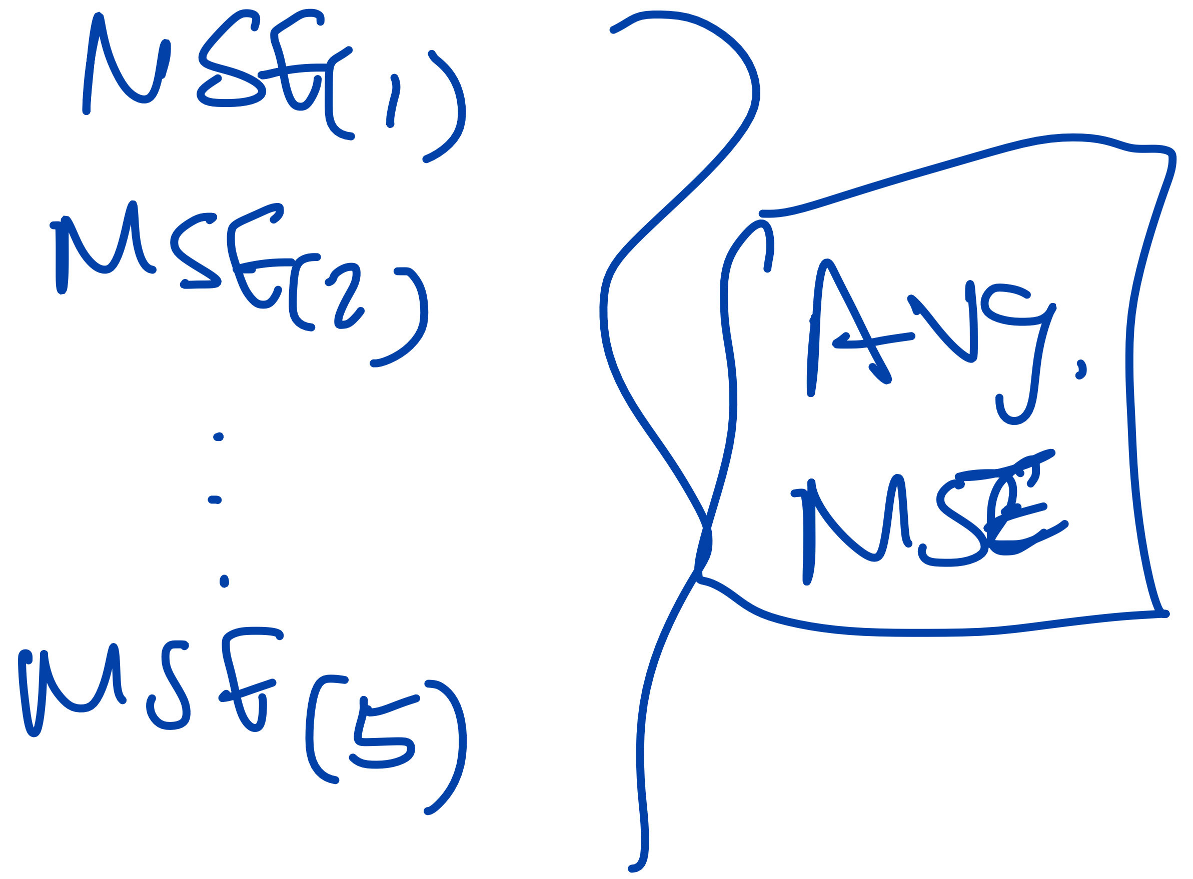
- Repeat for each k, obtain MSE for each k, and summarize using the average over K

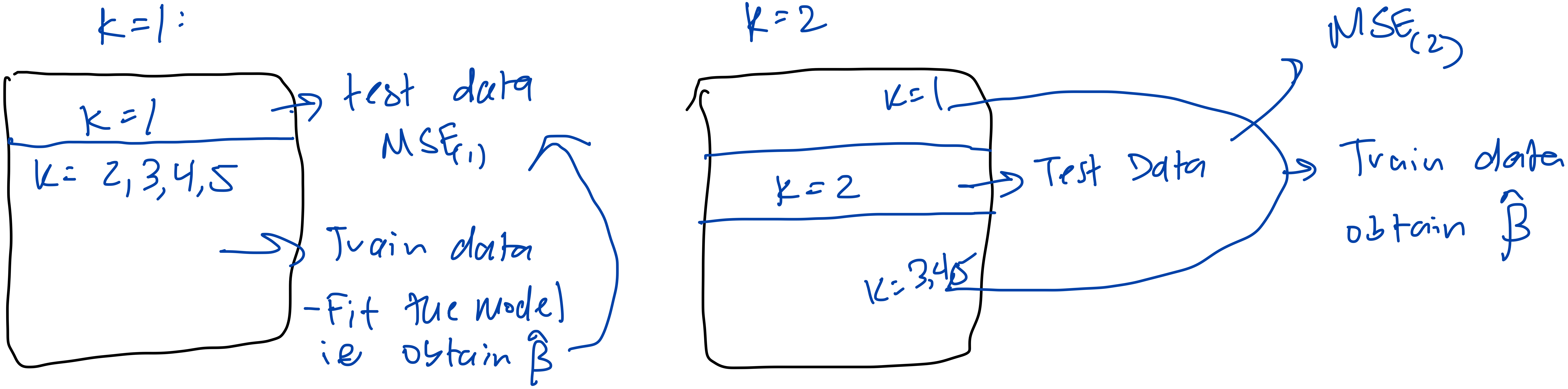
- $$AvgMSE = \frac{1}{K}\sum_{i=1}^{K} MSE_{test}^{(k)}$$

MSE is an appropriate metric for an average (p-values are not))

K-fold cross validation







What should K be?

- Leave-one-out cross validation: $K = N$ (computationally intensive)
- $k = 5, \underline{\underline{k = 10}}$ are common choices

Consider:

- Sample size
- How many models to compare (computation resources)