



**Example:**

• Divide the data into 10 folds

- For  $i \equiv 1, \dots, 10$

- Using every fold except  $i$ , perform variable selection and fit the model with the selected variables

• Compute the error on fold  $i$

• Averagethe 10 test errors obtained

## **Moral of the story:**

You can and often should validate your entire data processing & learning pipeline,  
even variable selection!



**Cross Validation: Right or Wrong?**

# Cross Validation: Right or Wrong?

## Example:

- Divide the data into 10 folds
- For  $i = 1, \dots, 10$ 
  - Using every fold except  $i$ , perform variable selection and fit the model with the selected variables
  - Compute the error on fold  $i$
- Average the 10 test errors obtained

## **Moral of the story:**

You can and often should validate your entire data processing & learning pipeline,  
even variable selection!

**How many folds?**