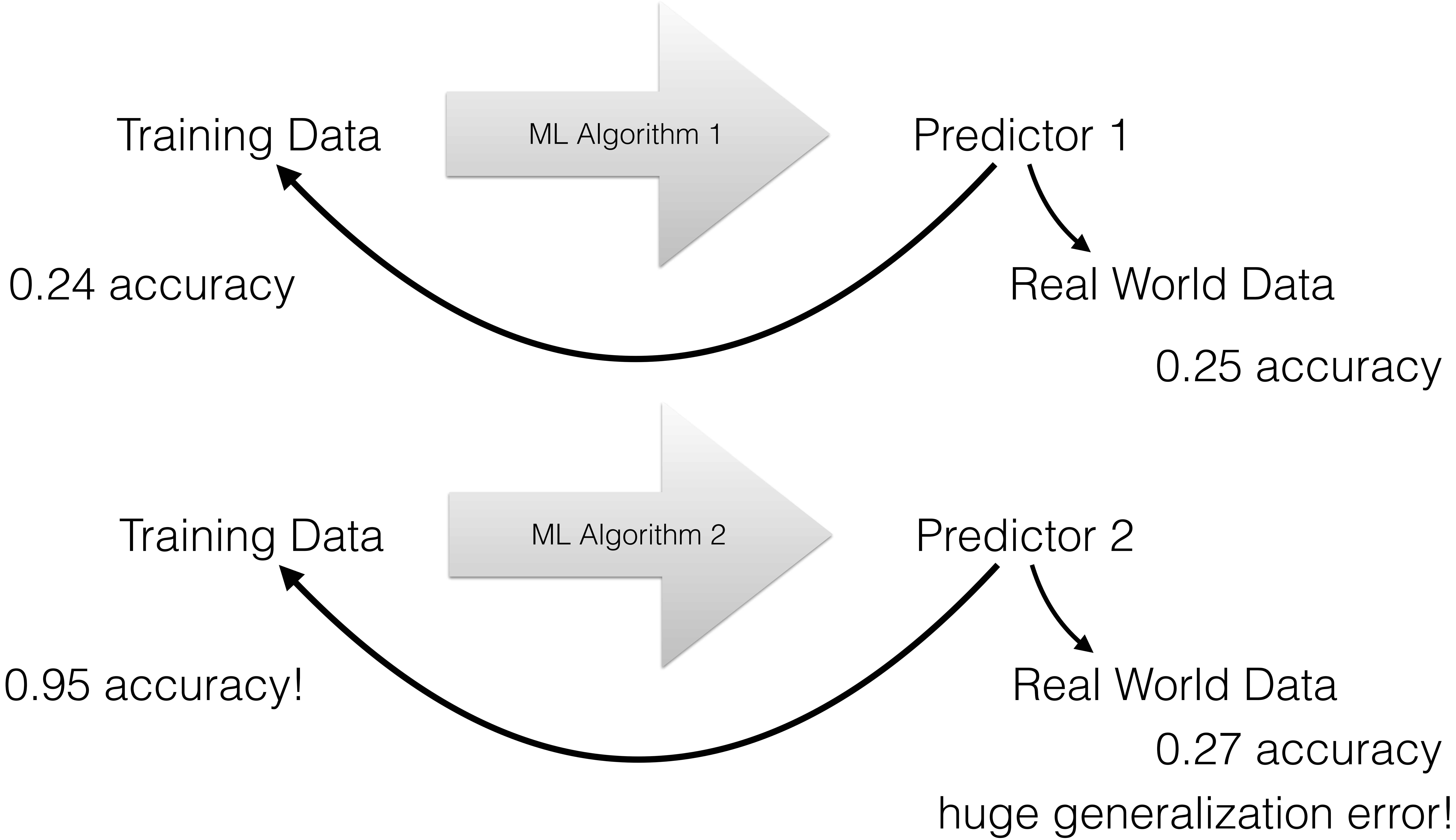
Model Selection

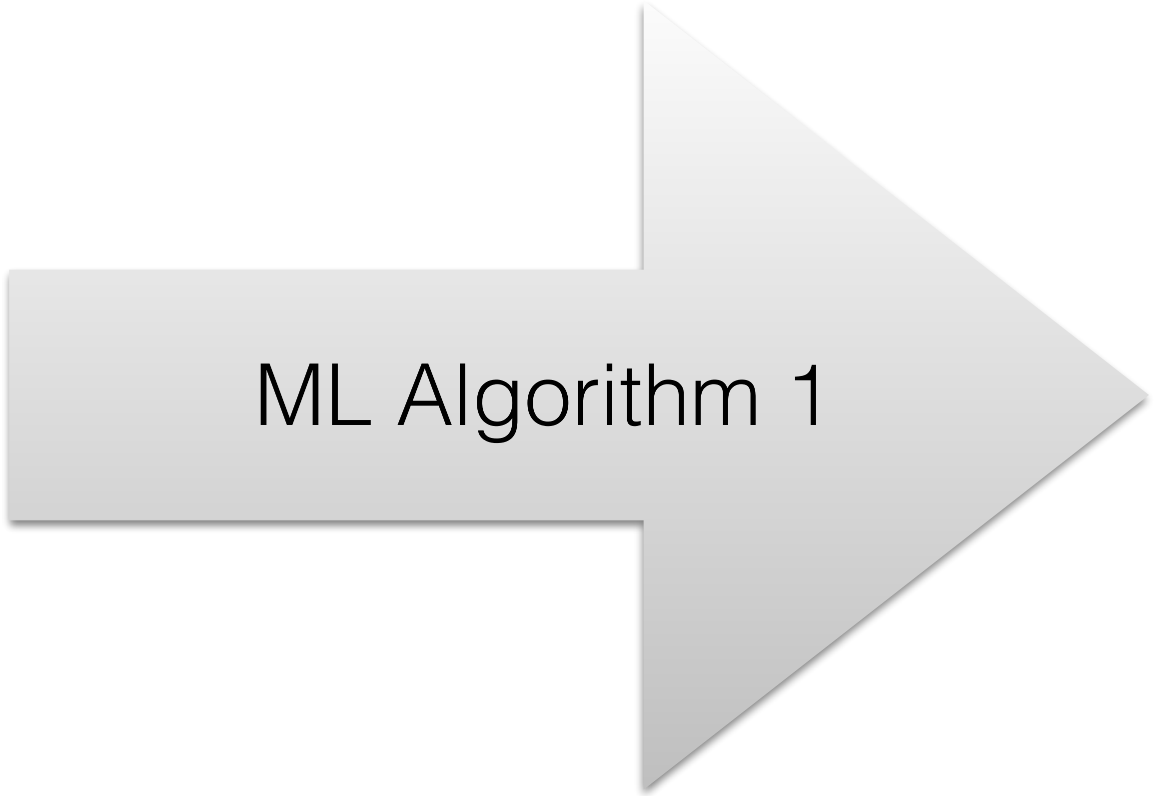
Machine Learning
CSx824/ECEx242
Bert Huang
Virginia Tech

Model Complexity

- Overfitting and underfitting
- Generalization error
- Validation for model selection

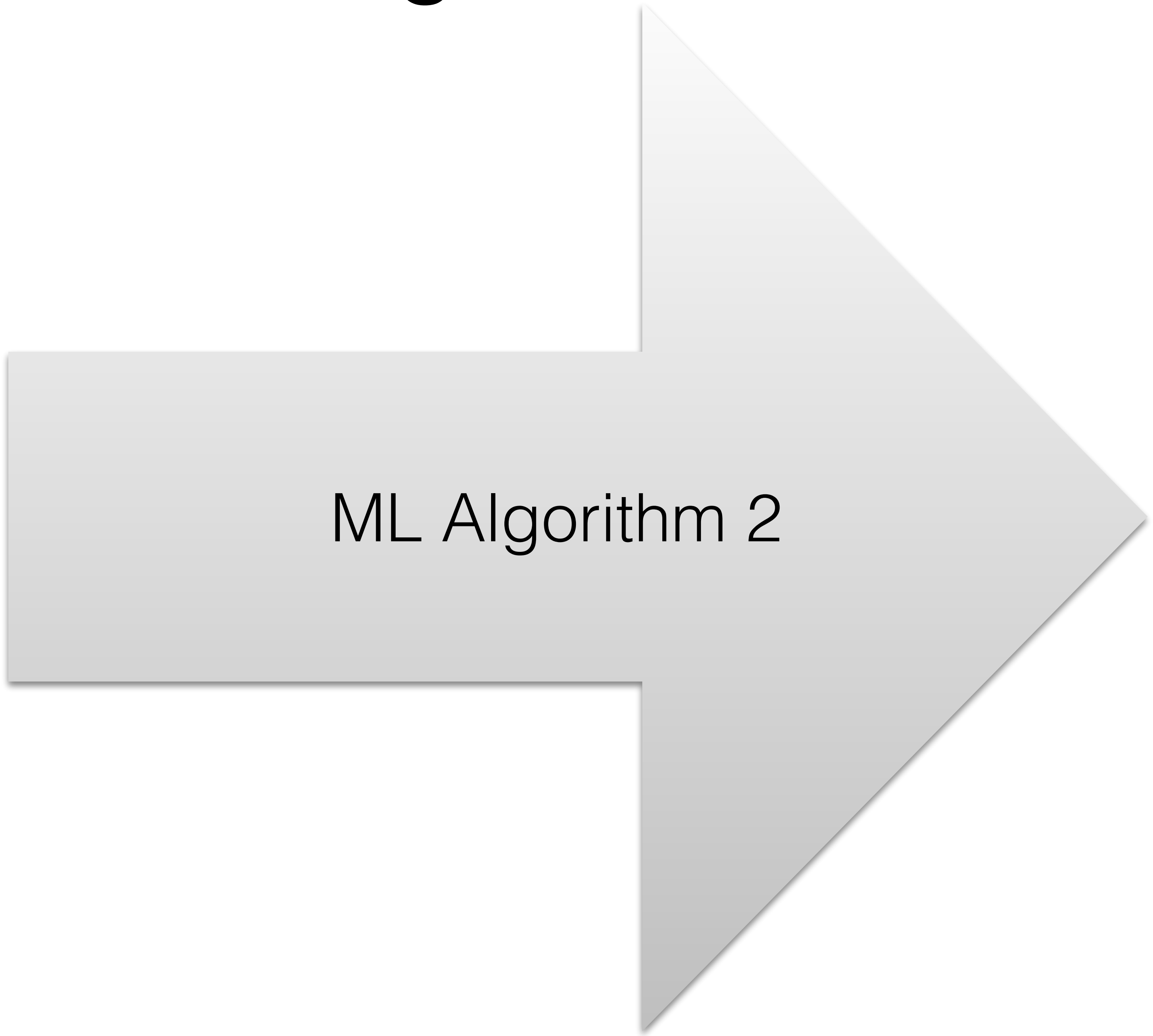

Training Data
ML Algorithm 1
Predictor 1
0.24 accuracy
Real World Data
0.25 accuracy
Training Data
ML Algorithm 2
Predictor 2
0.95 accuracy!
Real World Data
0.27 accuracy
huge generalization error!

Underfitting



- Low dimensional
- Heavily regularized
- Bad modeling assumptions

Overfitting



- High dimensional or non-parametric
- Weakly regularized
- Not enough modeling assumptions
- Not enough data

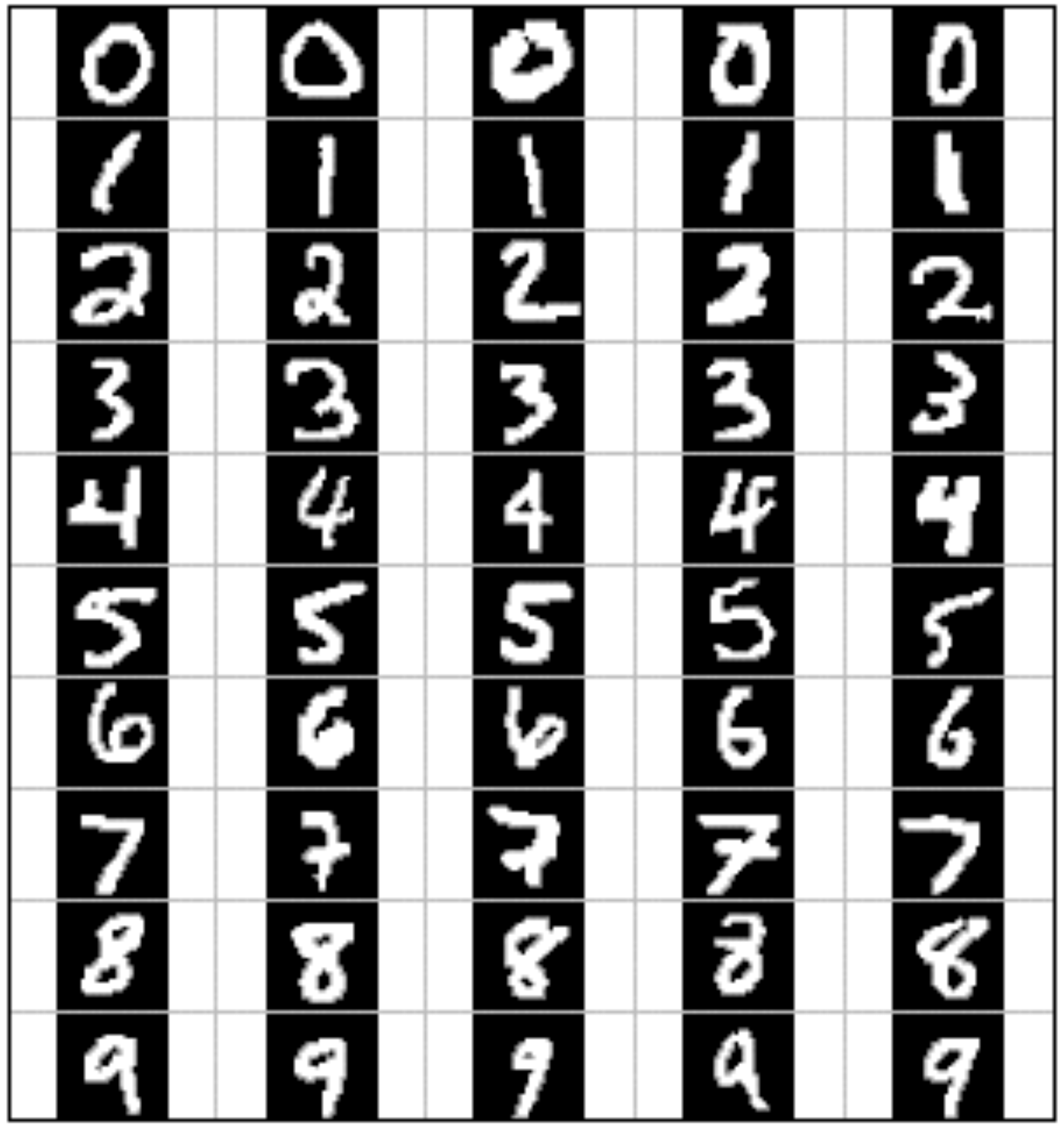

data


data
shape-preserving


data
shape-preserving
linear


data
shape-preserving
linear
quadratic

Nearest-Neighbor Classifiers



classifier = {

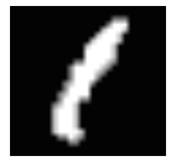
 : 0,

 : 0,

 : 0,

 : 0,

 : 0,

 : 1,

 : 1,

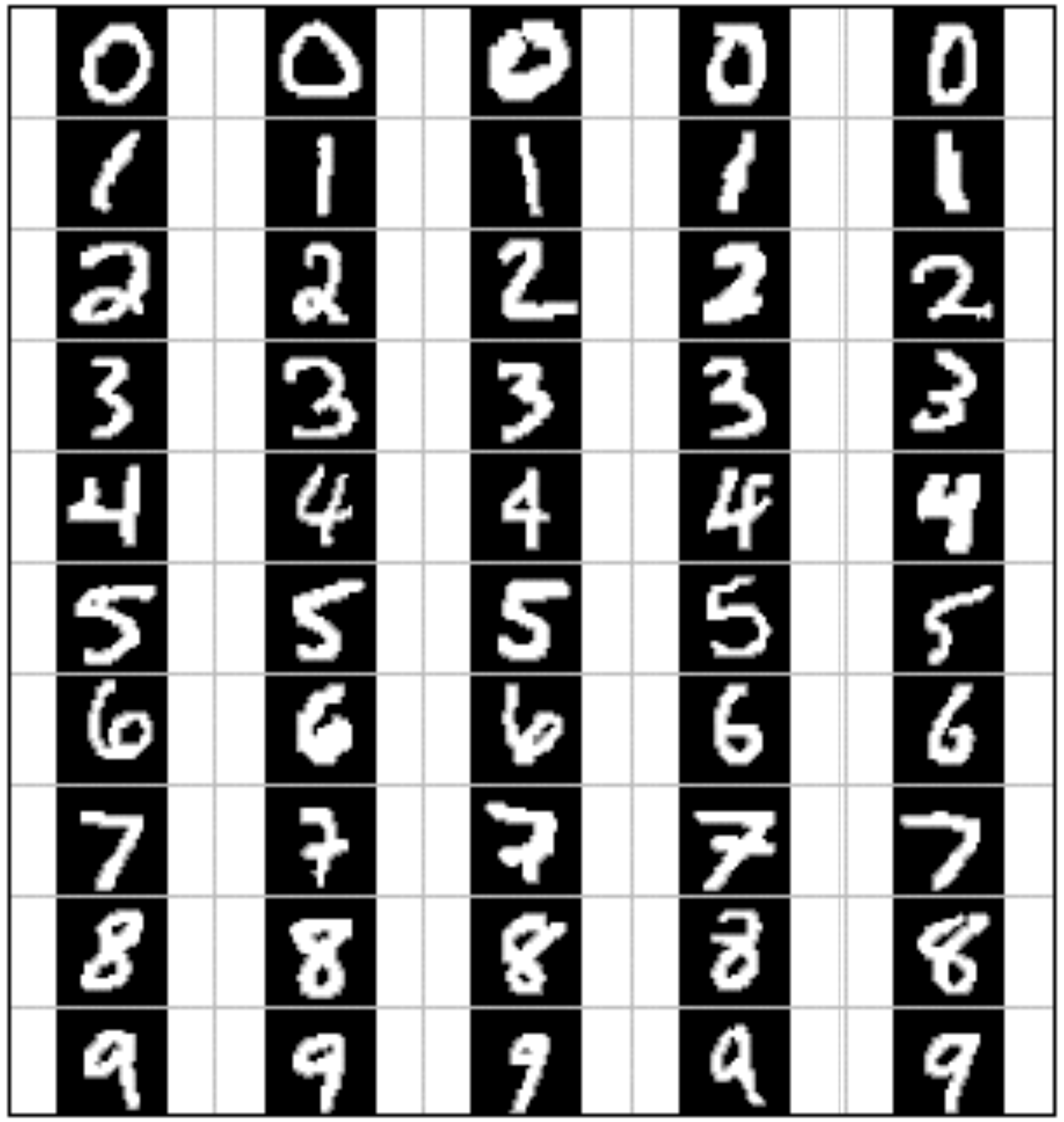
…

}

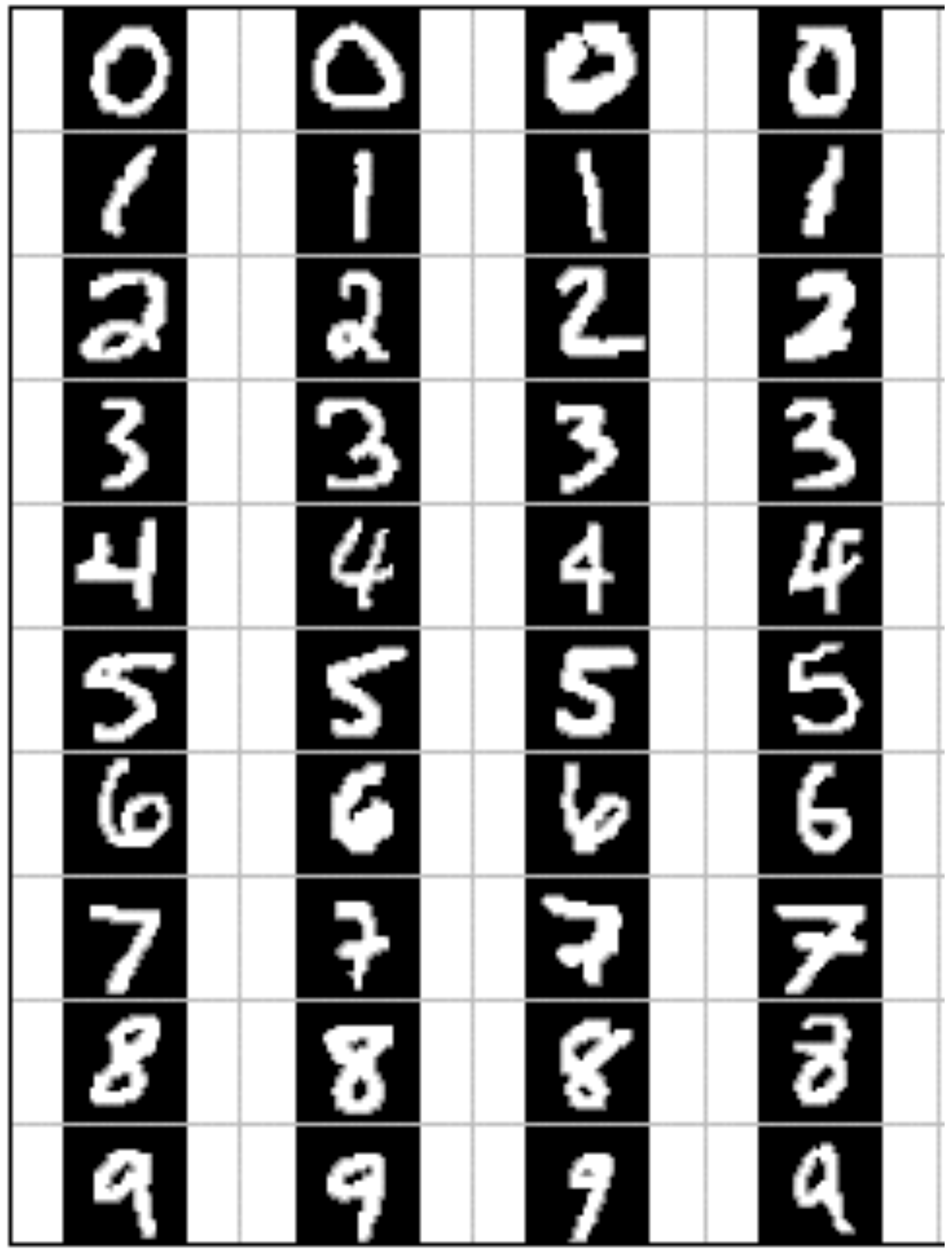
100% training accuracy!

53% testing accuracy…

Held-out Validation

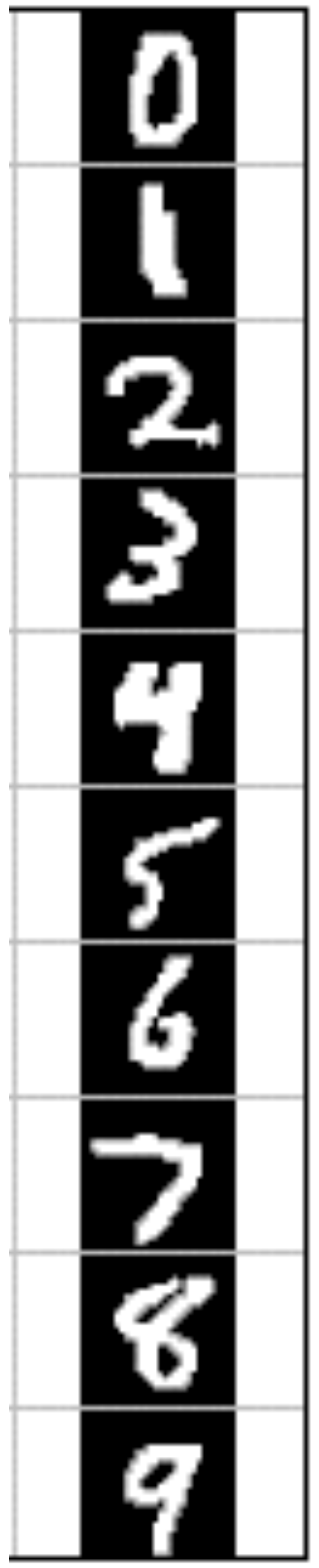


Held-out Validation



training data

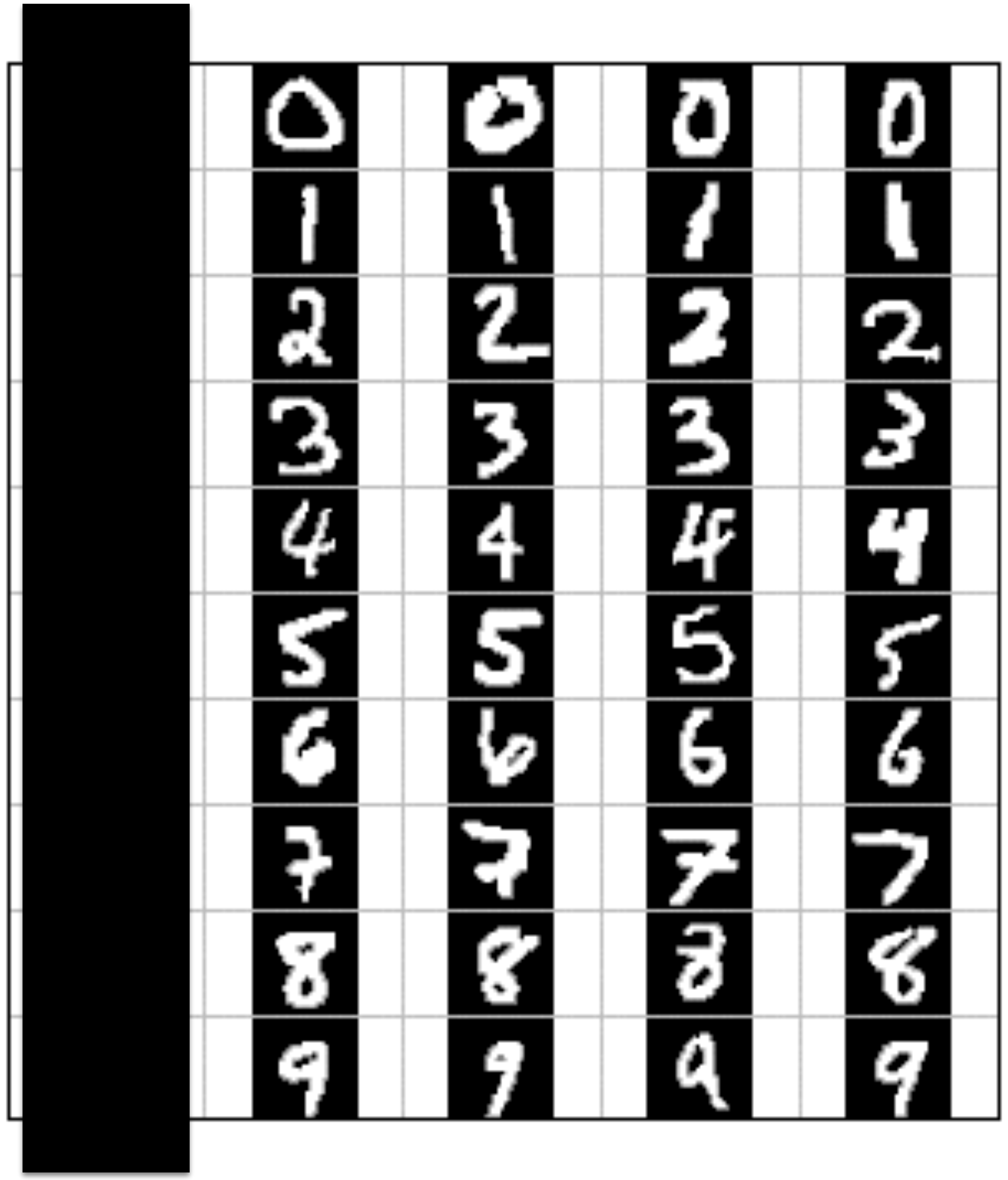
	Accuracy on training data	Accuracy on validation data
Simple	0.91	0.83
Medium	0.95	0.88
Complex	0.99	0.79
Super Complex	1.0	0.54



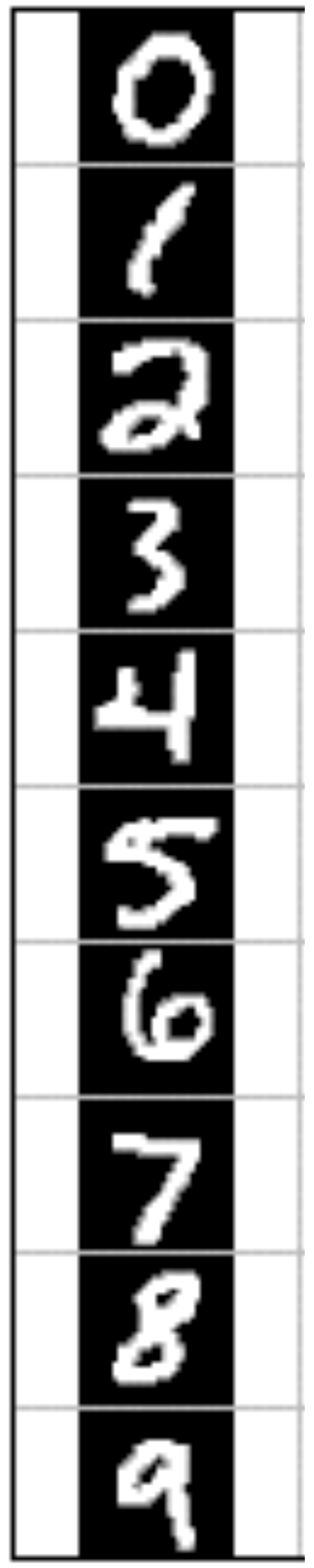
validation data

Cross Validation

Fold 1



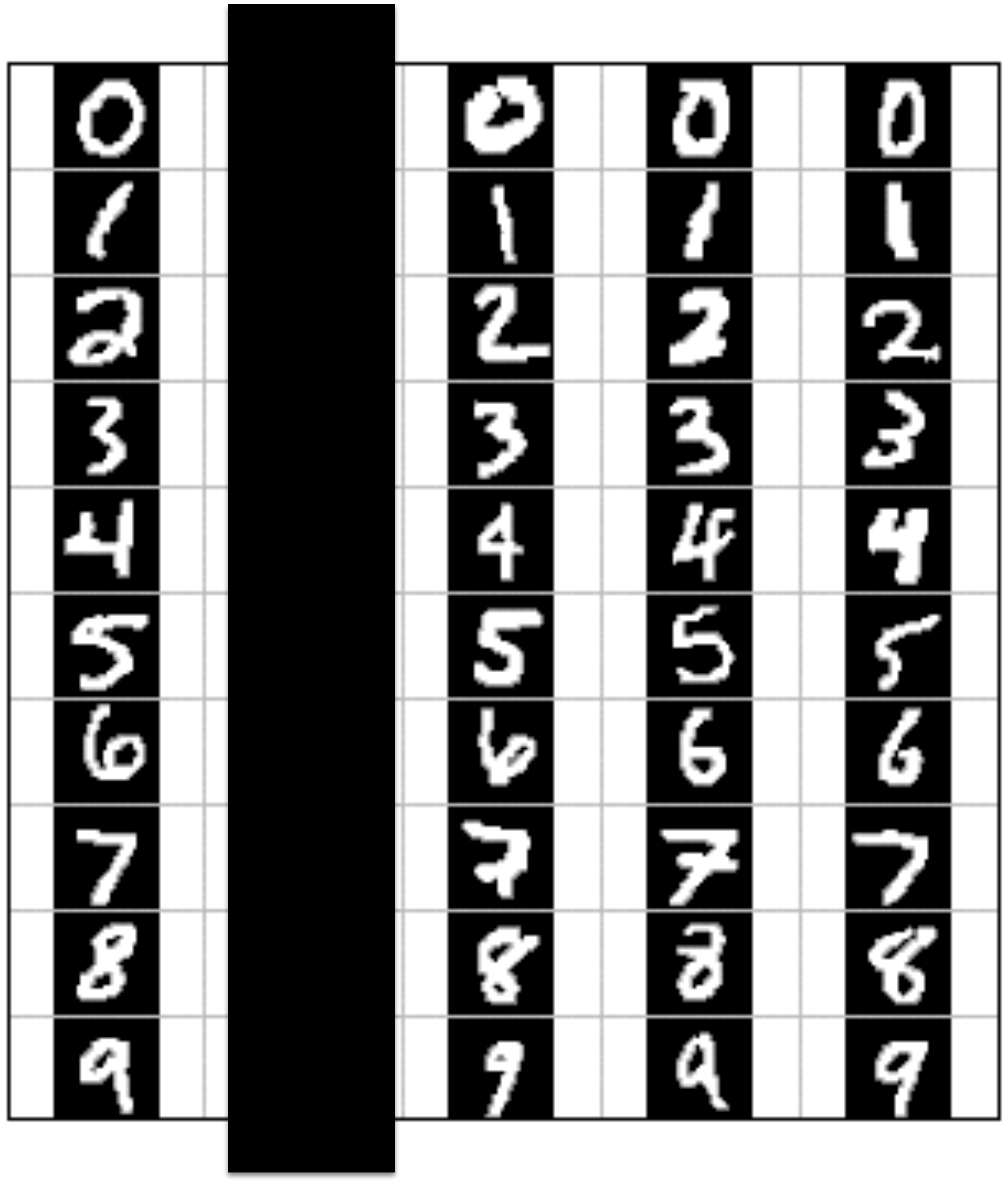
training data



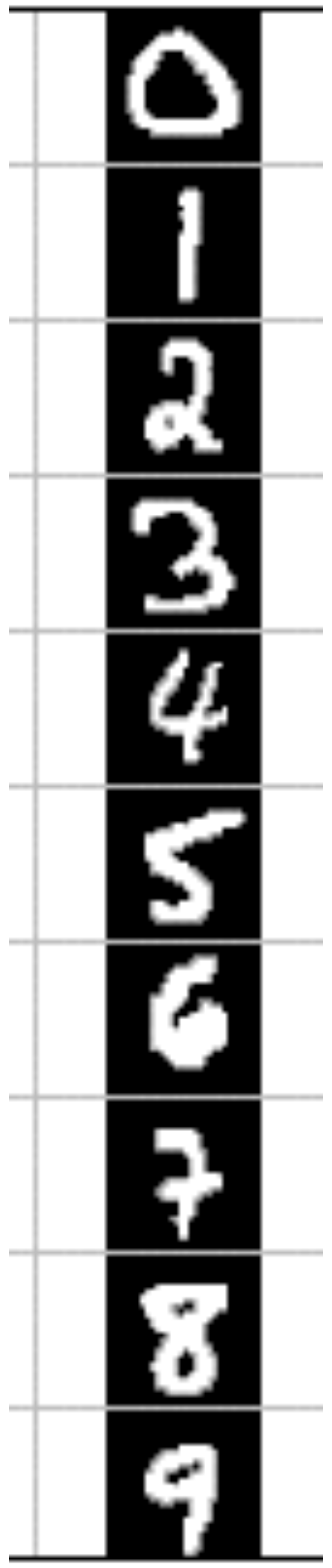
validation data

Cross Validation

Fold 2



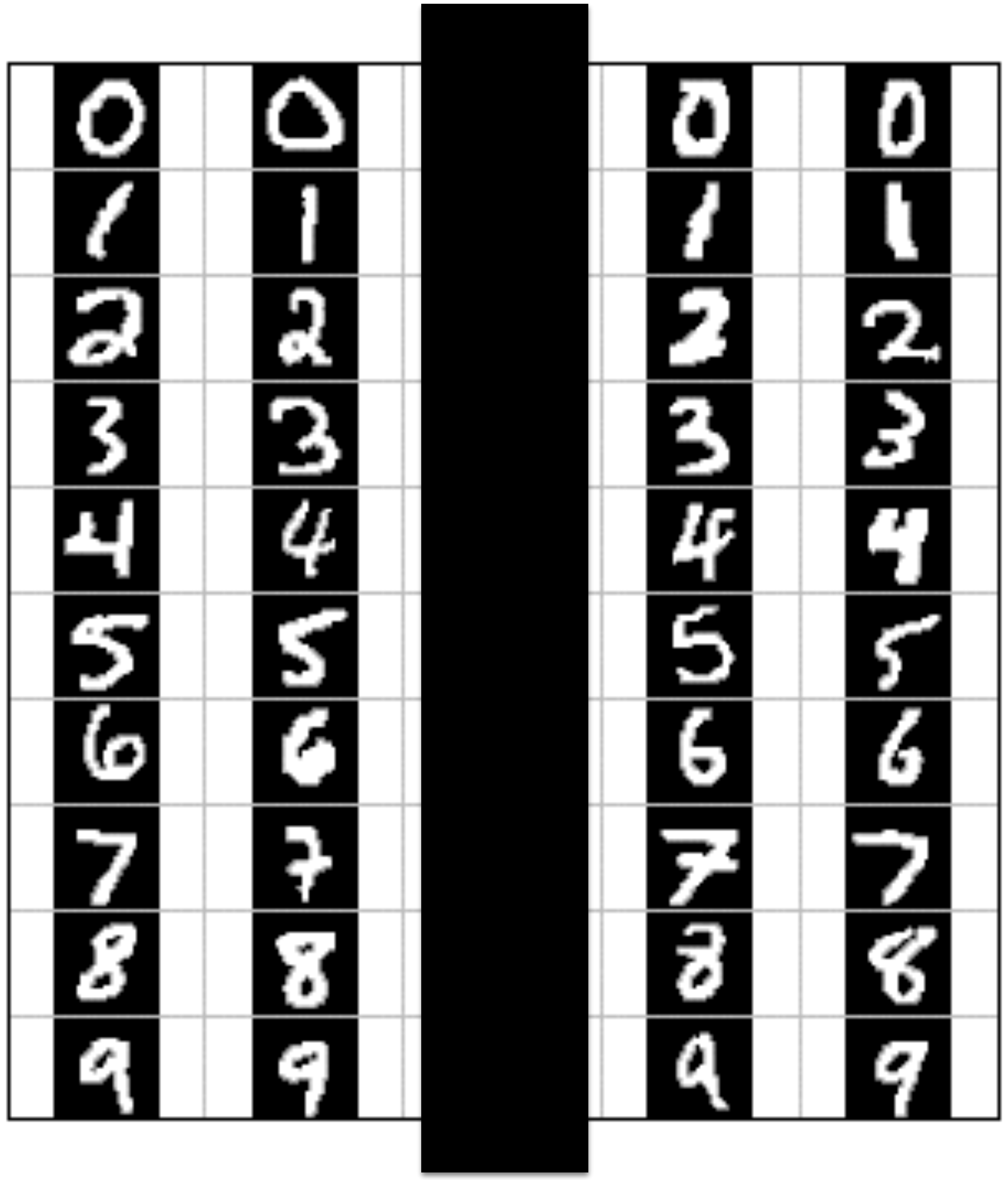
training data



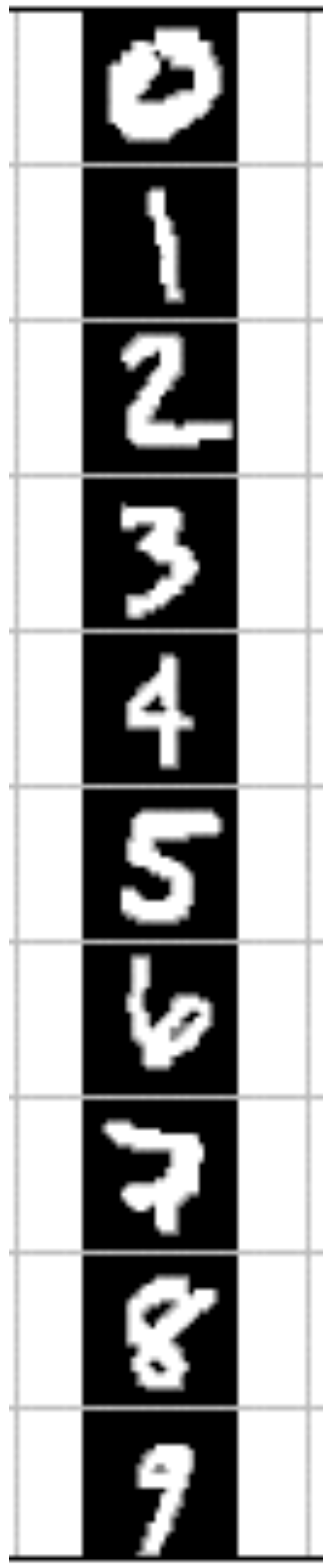
validation data

Cross Validation

Fold 3



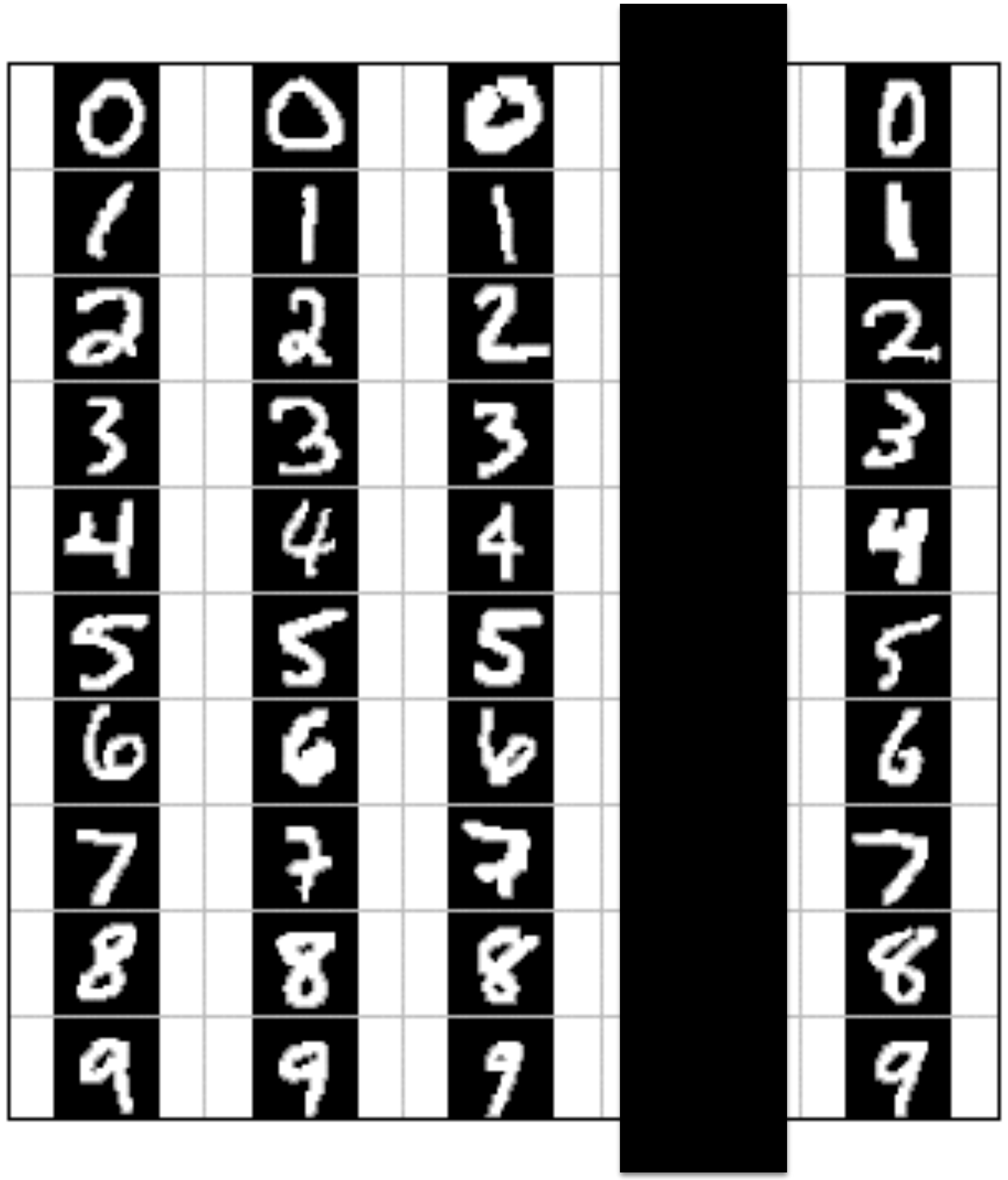
training data



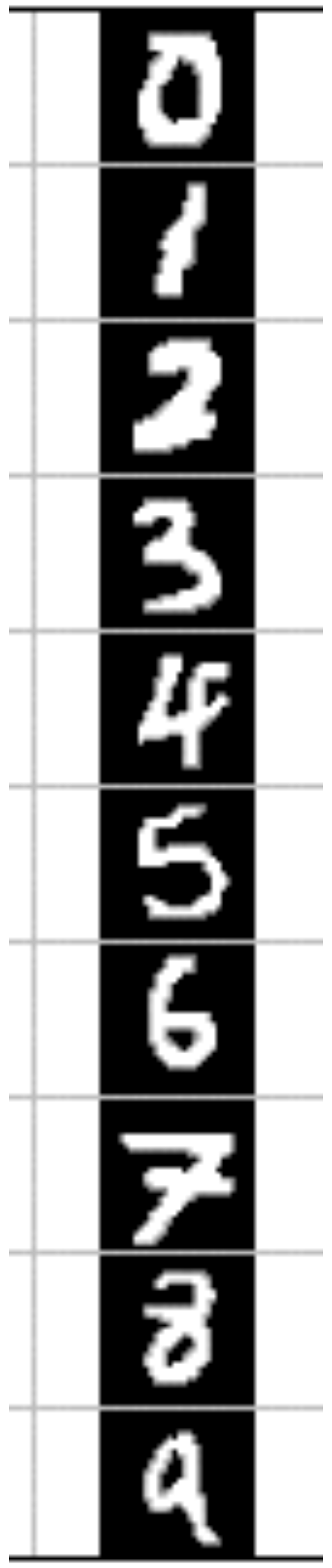
validation data

Cross Validation

Fold 4



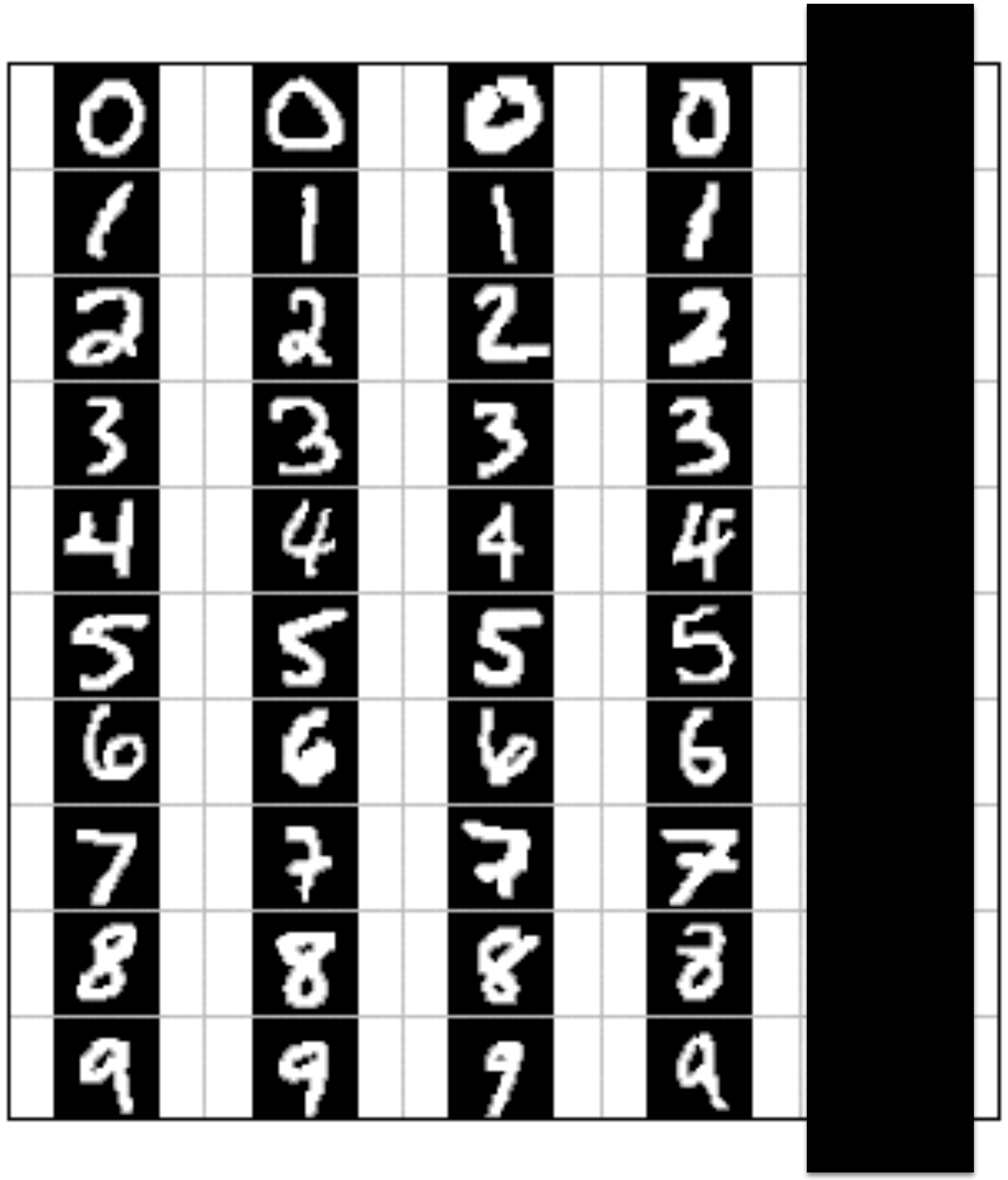
training data



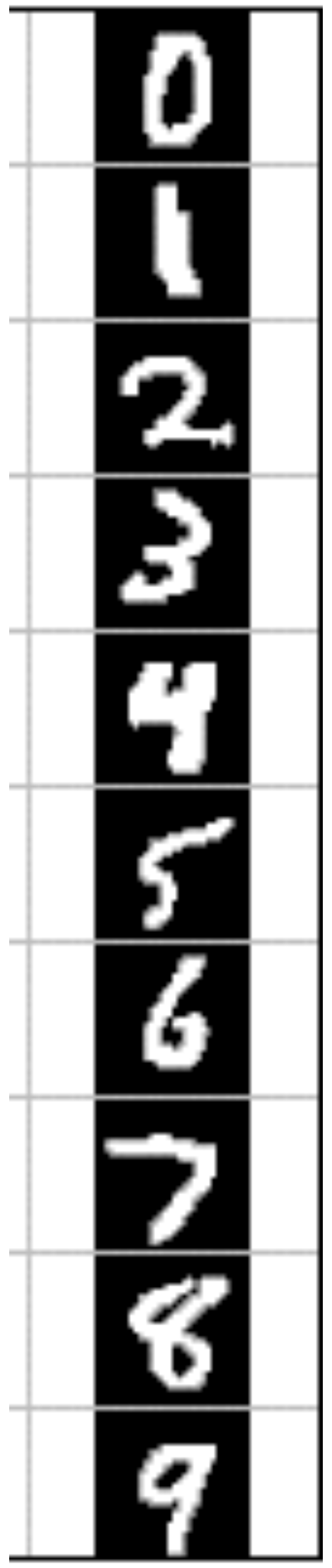
validation data

Cross Validation

Fold 5

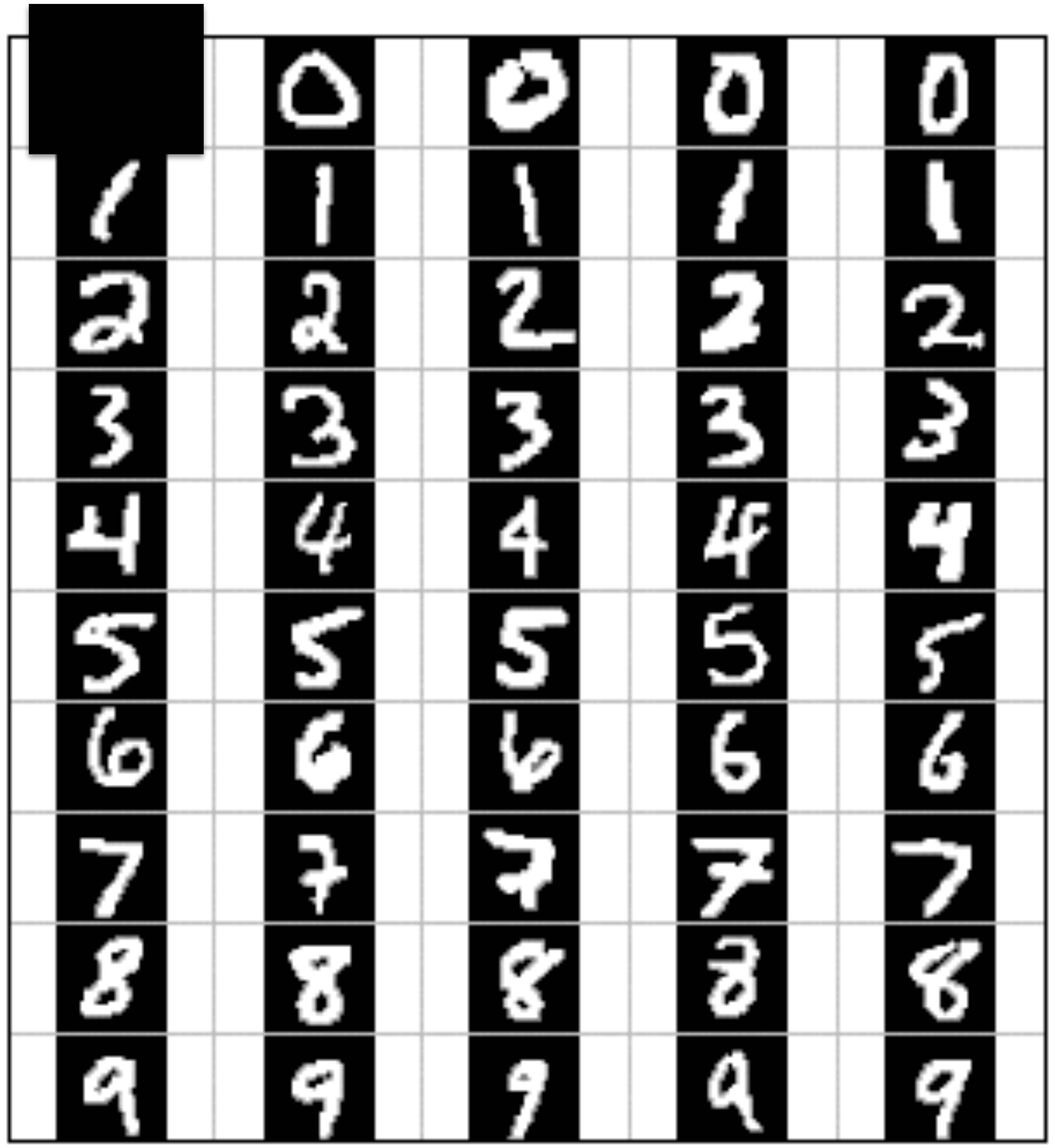


training data



validation data

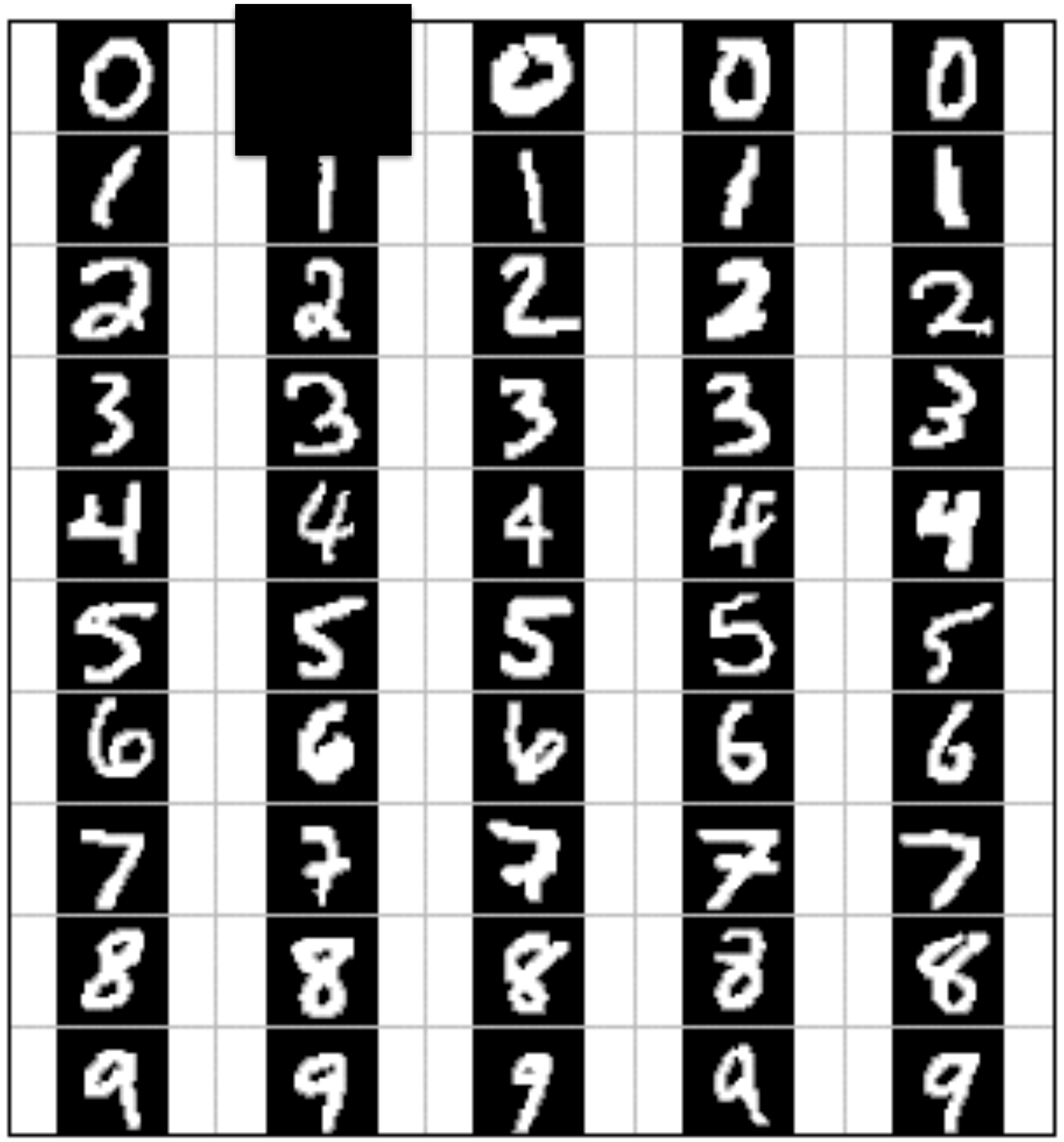
Leave-one-out Cross Validation



training data

validation data

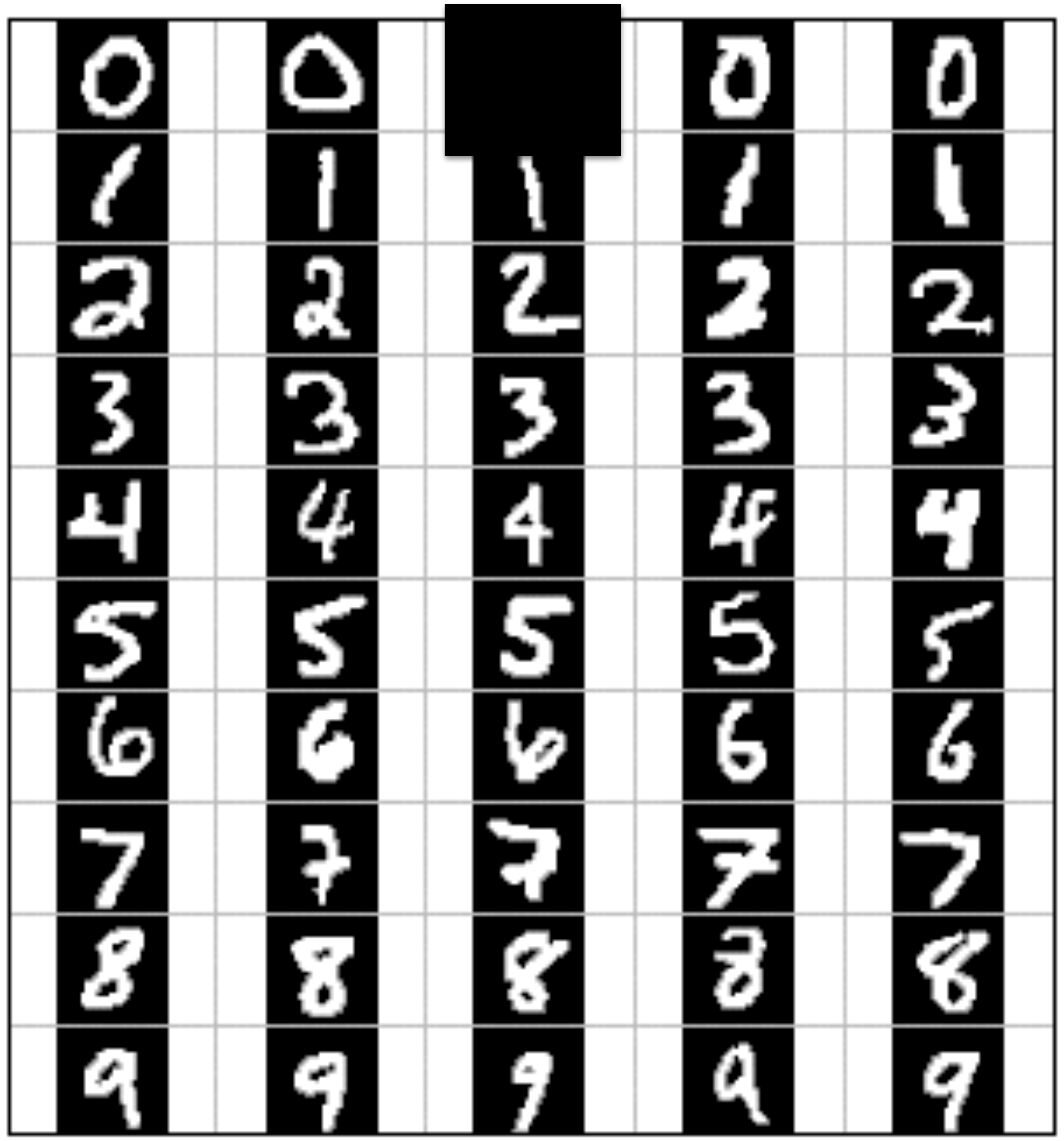
Leave-one-out Cross Validation



training data

validation data

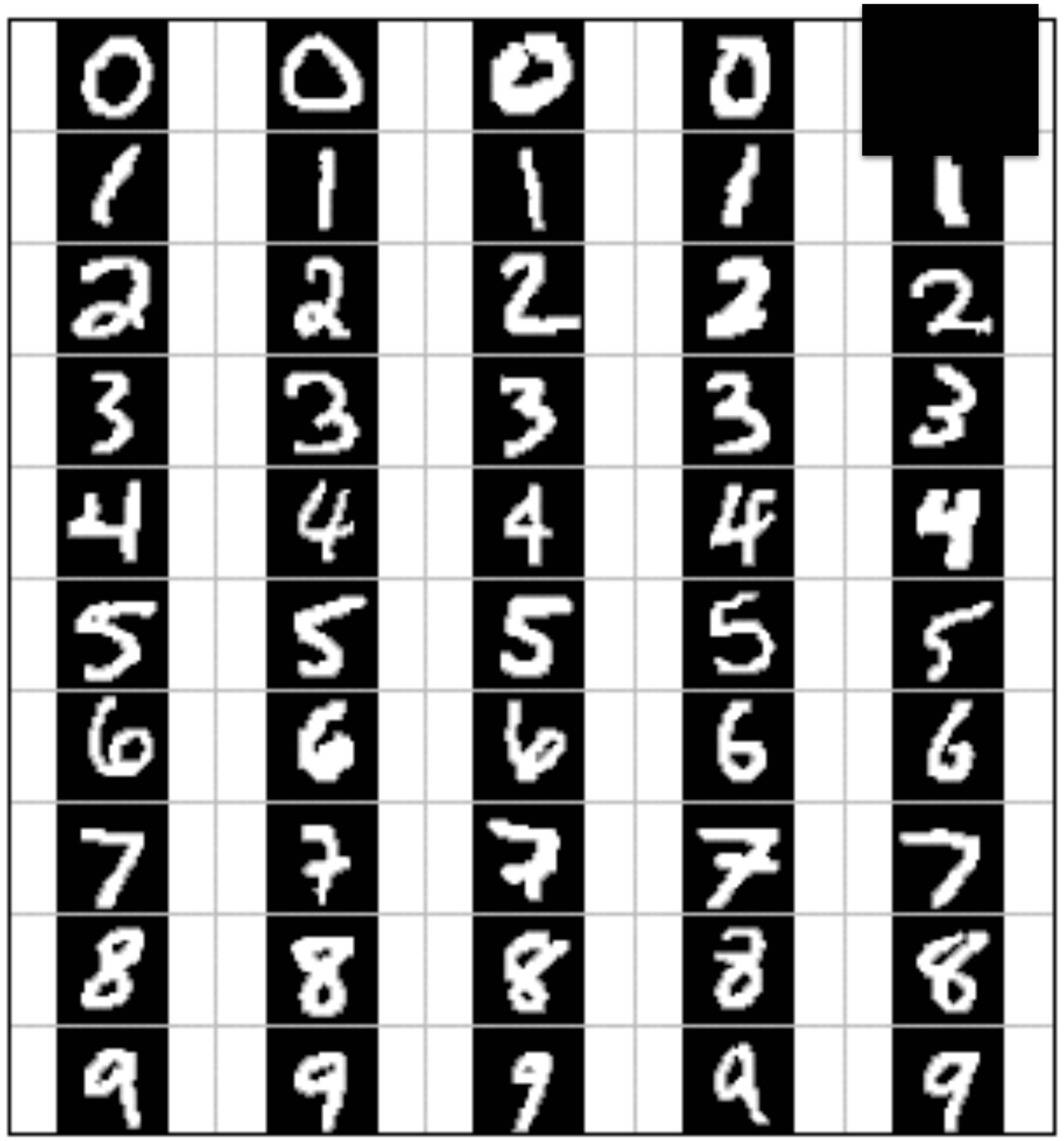
Leave-one-out Cross Validation



training data

validation data

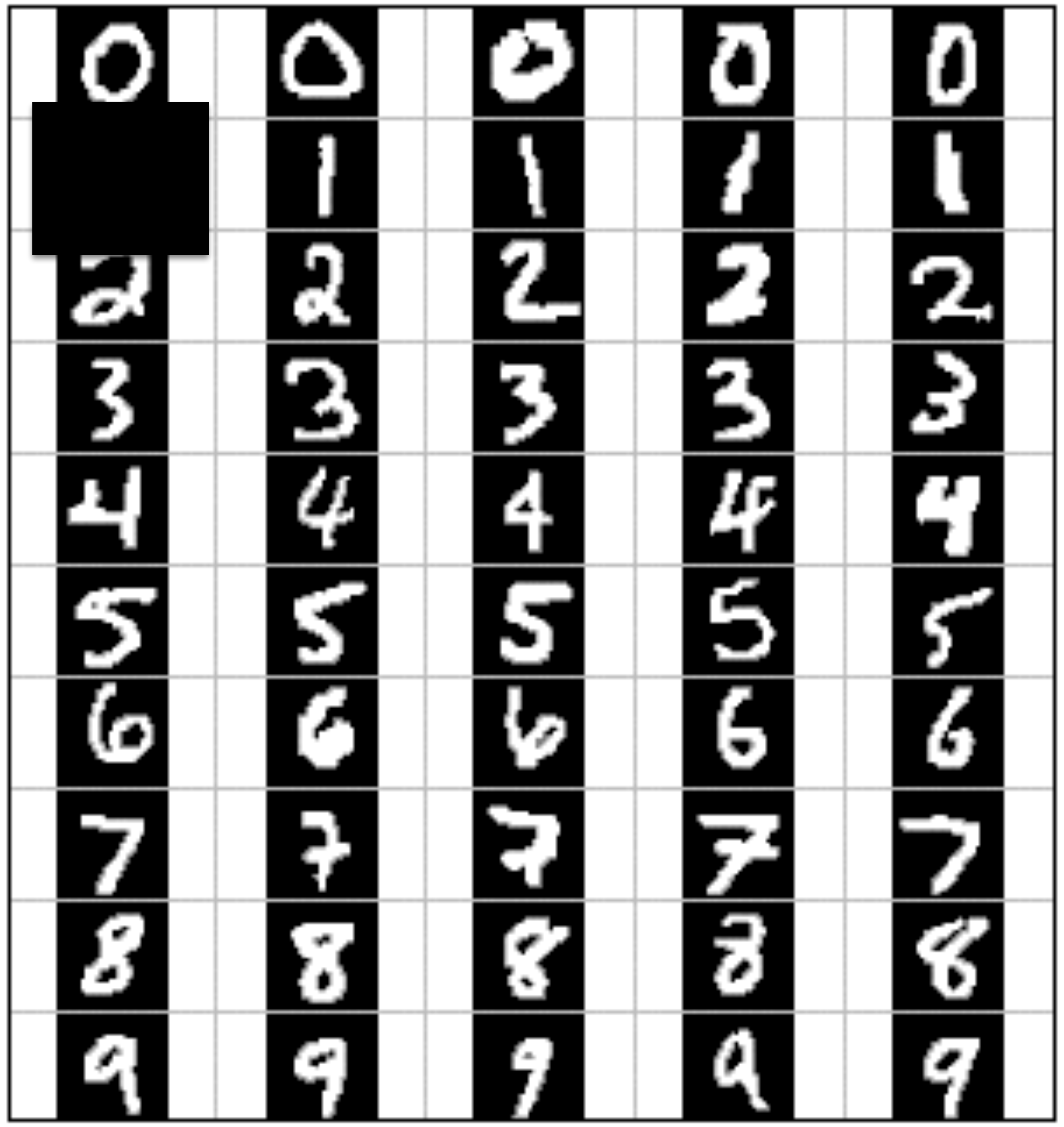
Leave-one-out Cross Validation



training data

validation data

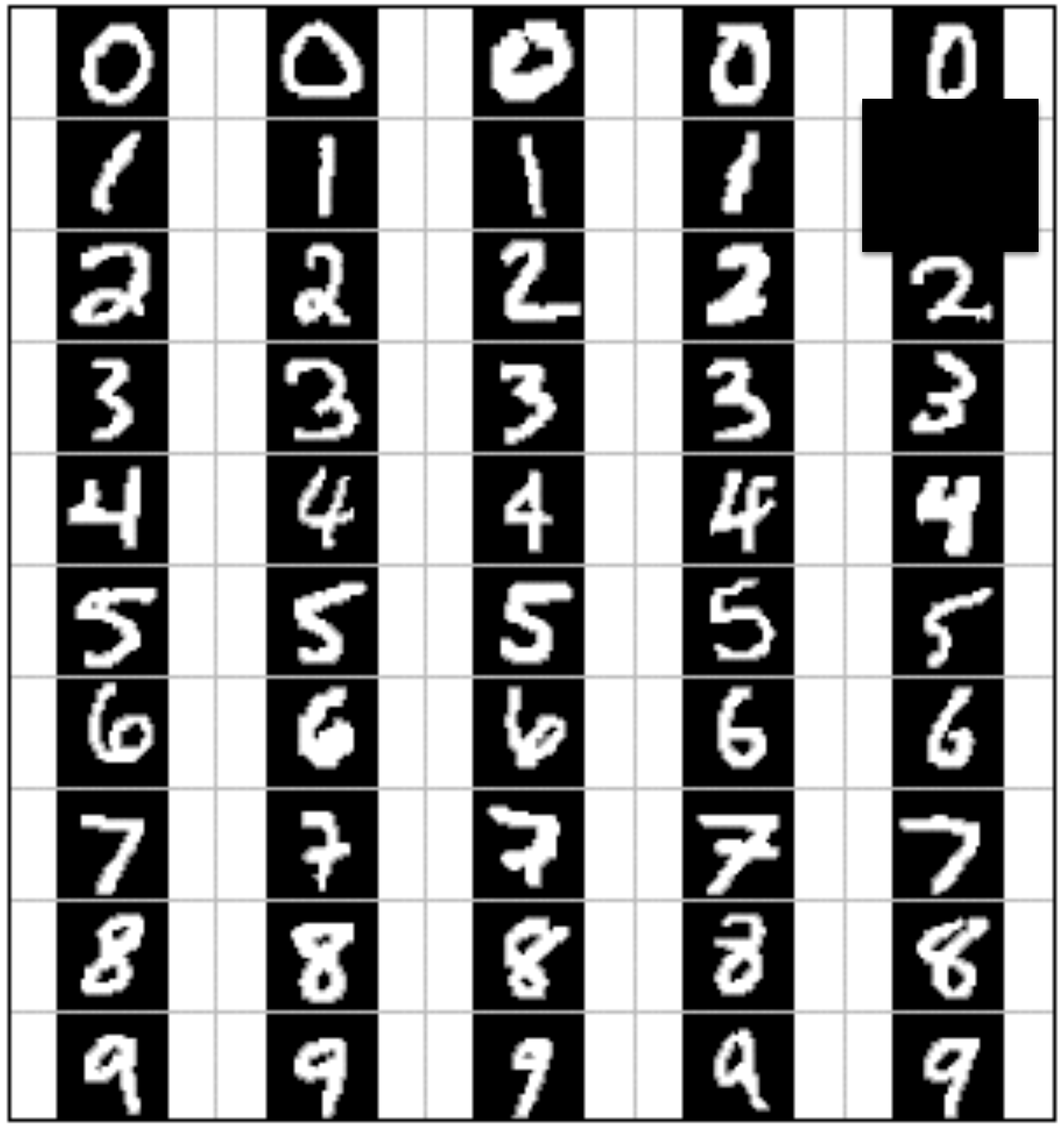
Leave-one-out Cross Validation



training data

validation data

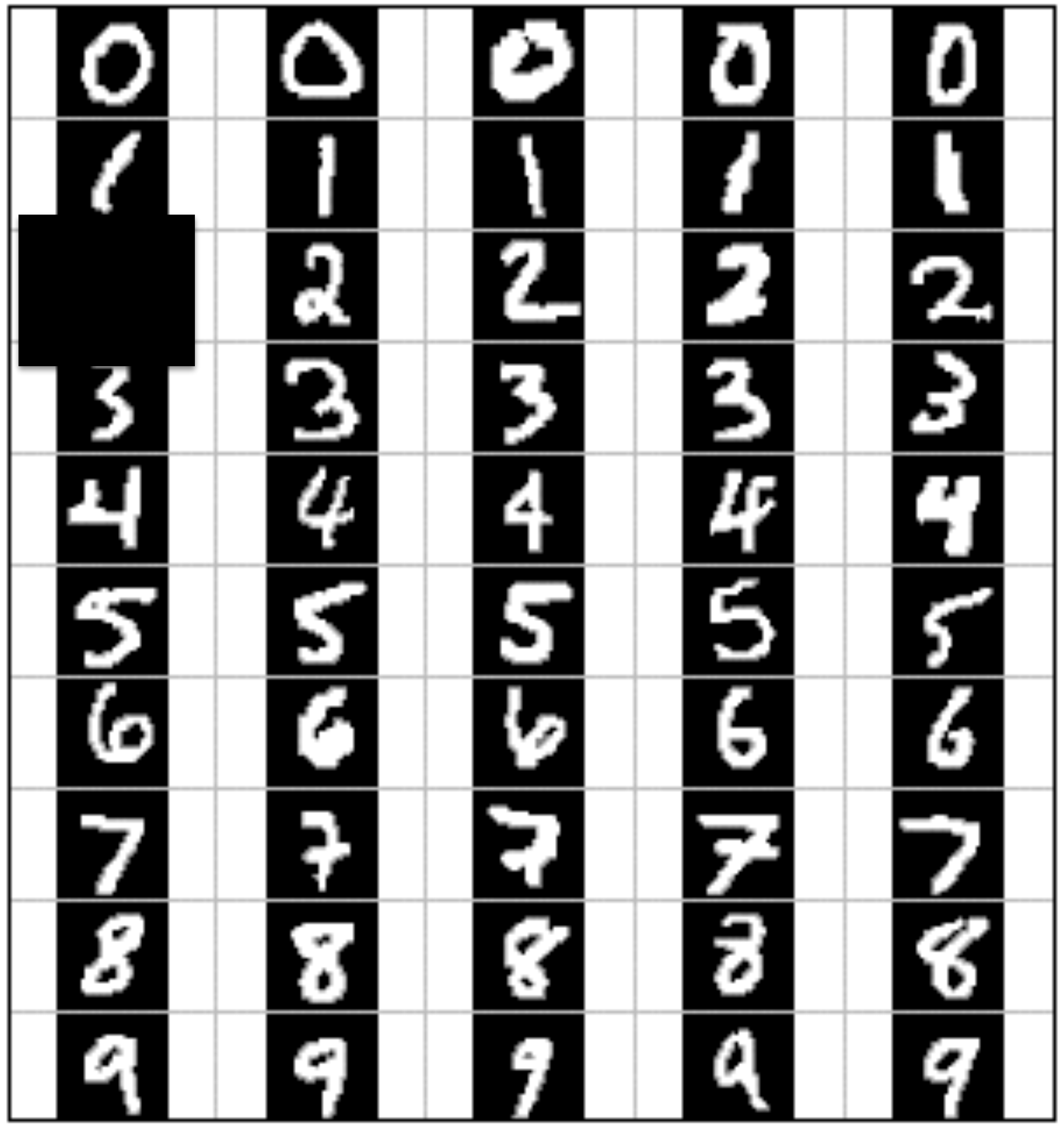
Leave-one-out Cross Validation



training data

validation data

Leave-one-out Cross Validation

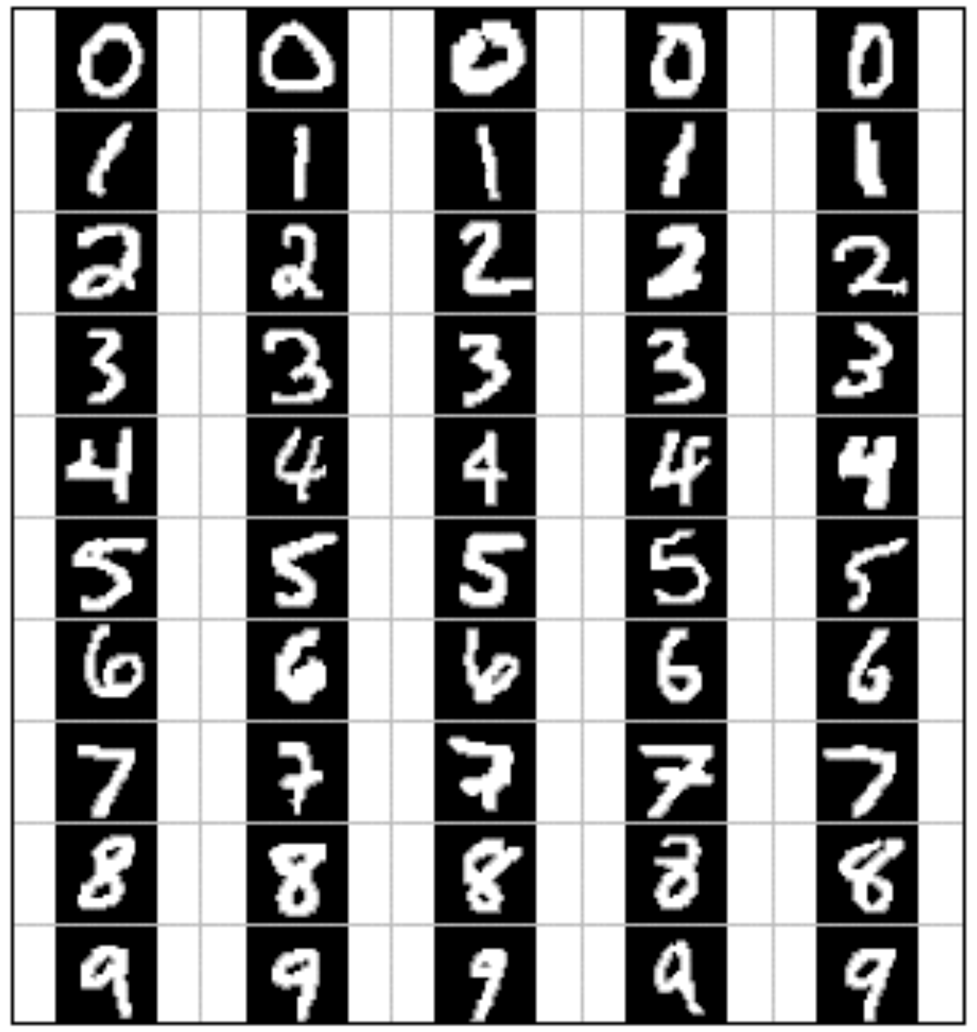


training data

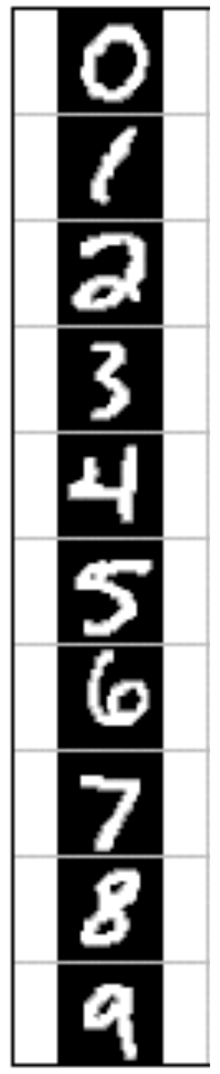
validation data

How Many Folds?

- What are the pros and cons of leave-one-out cross-validation?
- We usually train on N-1 folds and test on 1 fold. What are pros and cons of doing the inverse: train on 1 fold and test on N-1 folds?



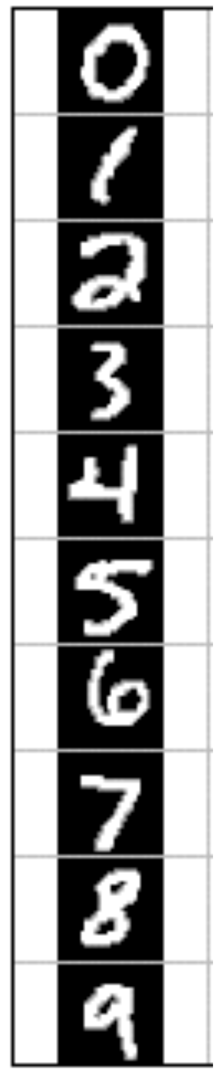
Training



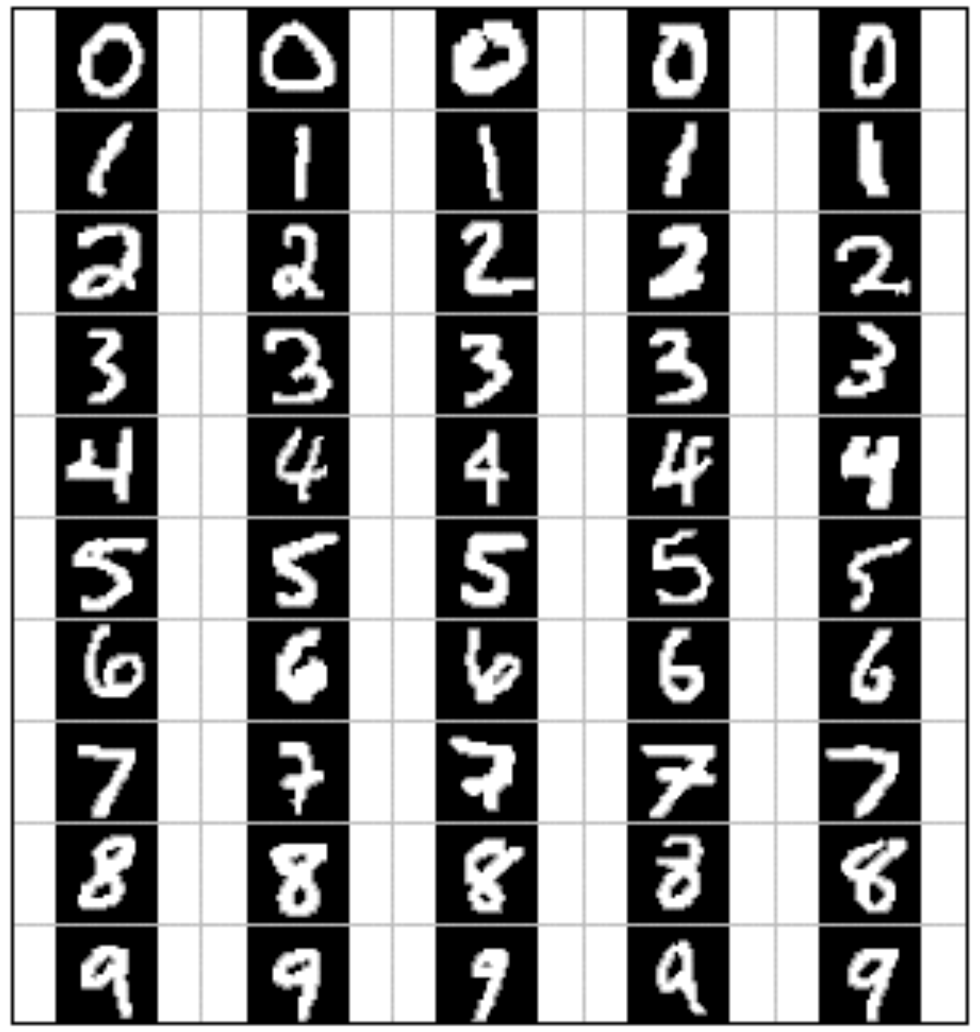
Testing

How Many Folds?

- What are the pros and cons of leave-one-out cross-validation?
- We usually train on N-1 folds and test on 1 fold. What are pros and cons of doing the inverse: train on 1 fold and test on N-1 folds?



Training



Testing

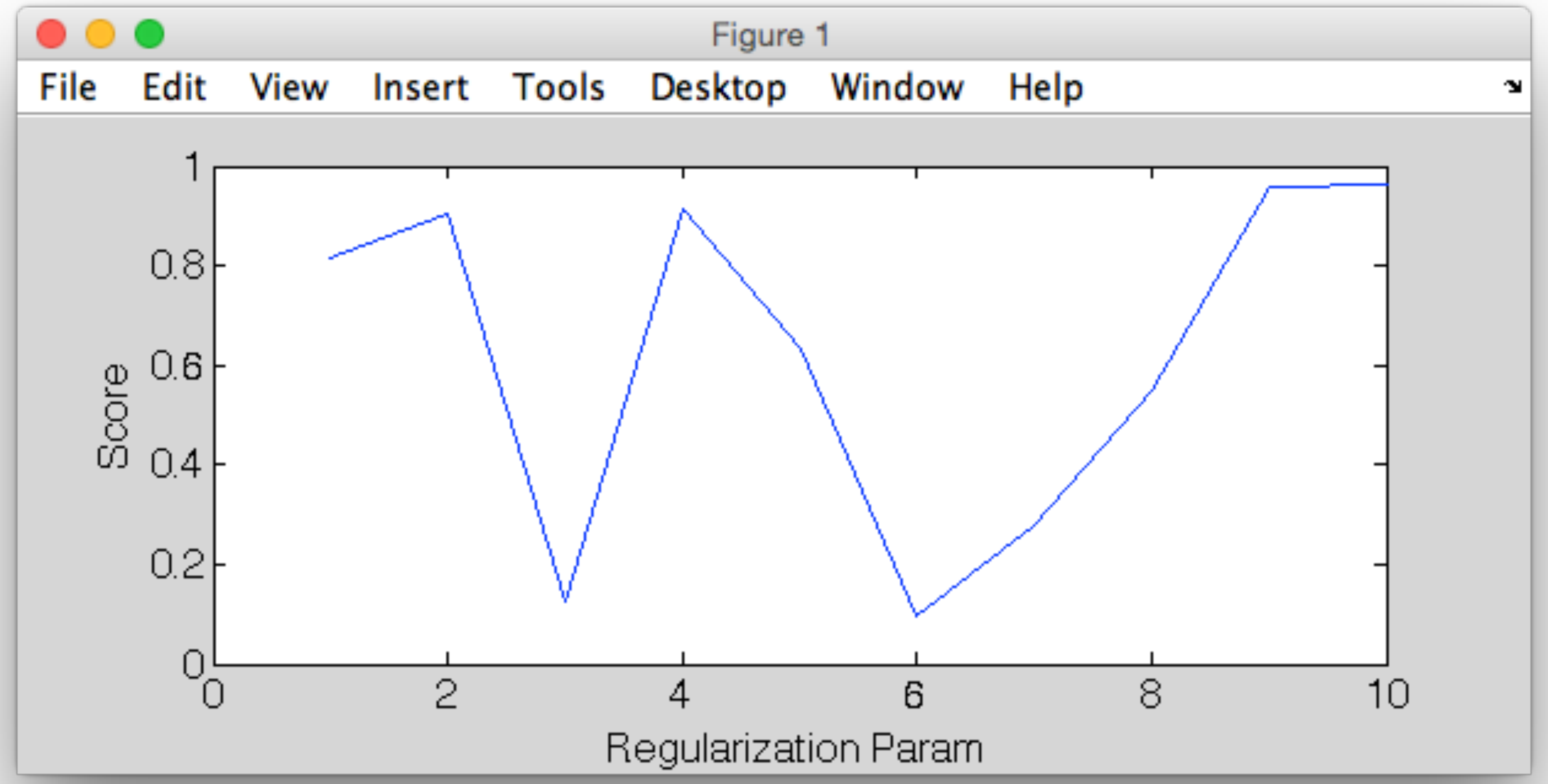
Testing versus Validation

- Best practice for experiments:
  - Hold out test set completely hidden from training
  - Use validation on training data for model (or parameter) selection
  - Evaluate on held-out test data

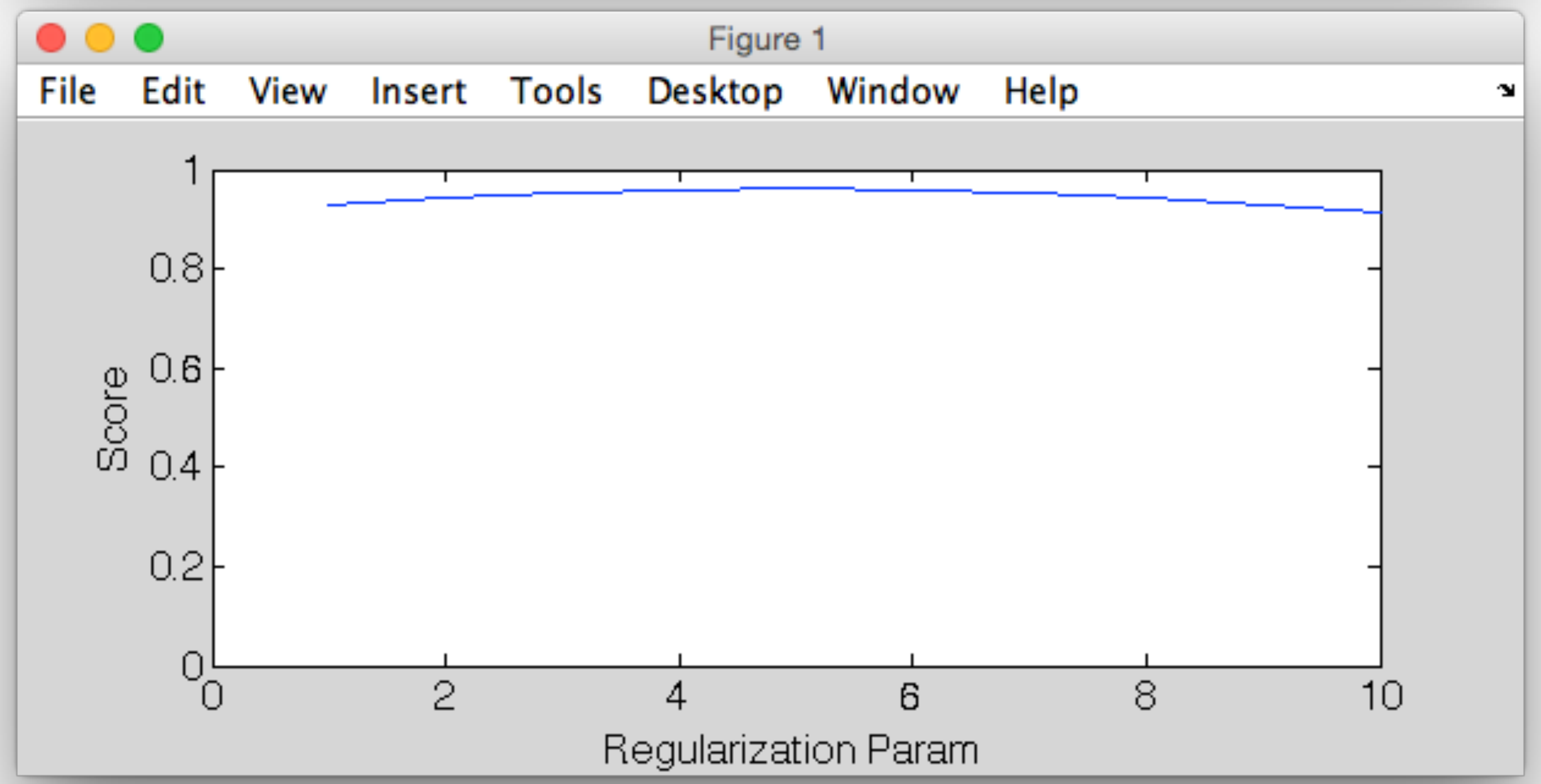
Scenarios

- Mystery ML algorithm with single complexity parameter

- Scenario 1: cross-validation scores are extremely erratic



- Scenario 2: cross-validation score is very uniform



Model Selection via Validation

- Measure performance on **held-out** training data
  - Simulate testing environment
- Rotate **folds** of held-out subsets
- Can even hold out one at a time: **leave-one-out** validation
- Use (cross) validation performance to tune extra parameters