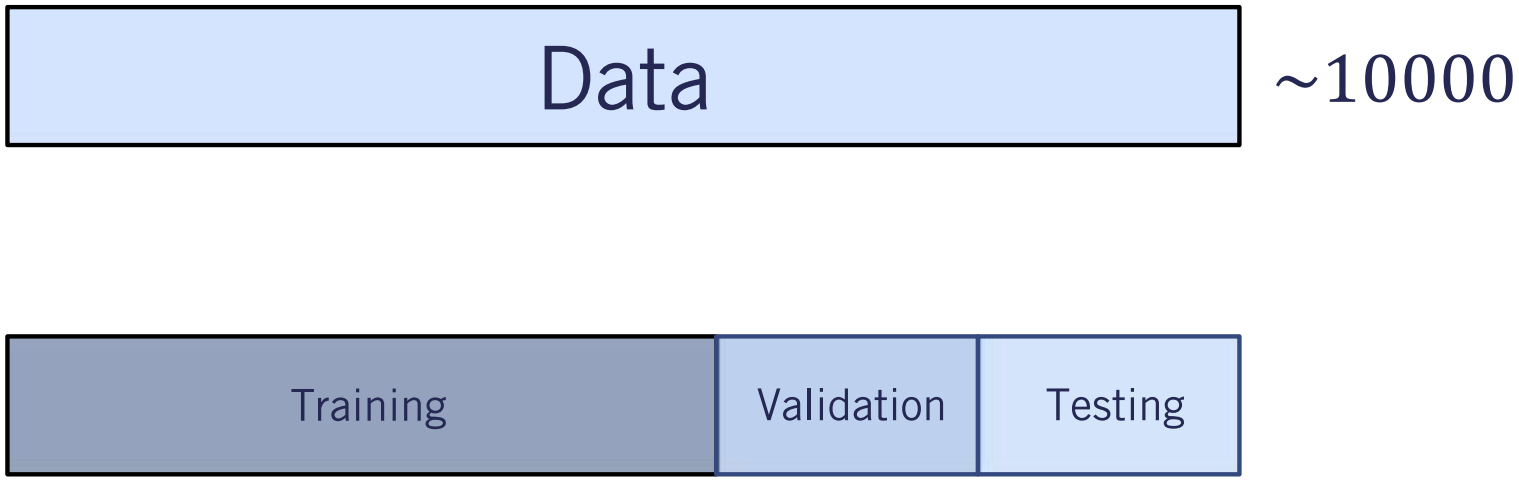
Learning Objectives

- Learn how to split a dataset for an unbiased estimate of performance

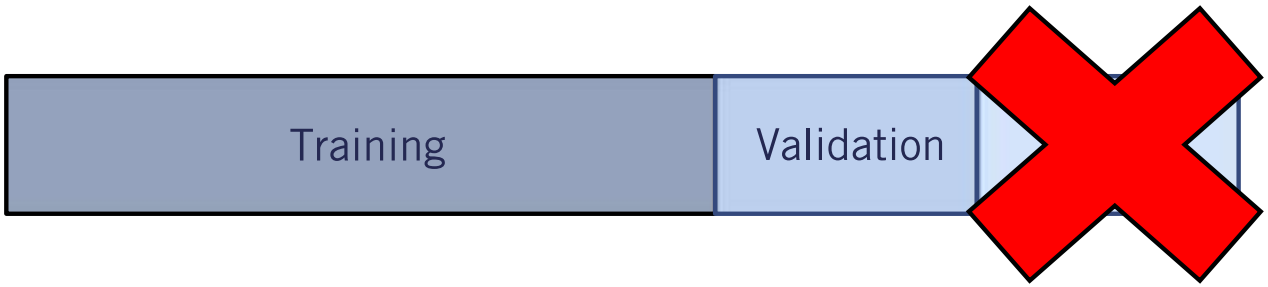
- Learn how to improve the performance of neural network by observing the difference in performance on the various data splits.

Data Splits

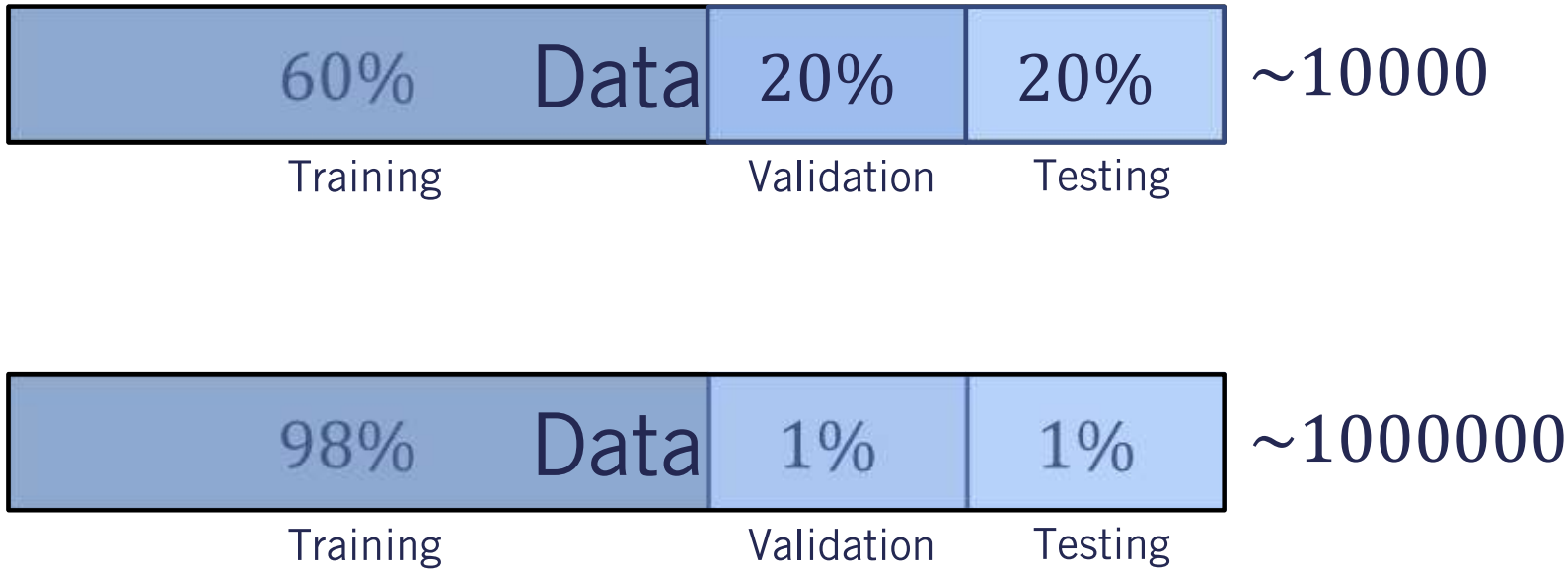


Data Splits

- **Training Split**: used to minimize the Loss Function
- **Validation Split**: used to choose best **hyperparameters**, such as the learning rate, number of layers, etc.
- **Test Split**: the neural network **never observes** this set. The developer **never** uses this set in the design process



Data Splits


60%
Data
20%
20%
~10000
Training
Validation
Testing
98%
Data
1%
1%
~1000000
Training
Validation
Testing

Behavior of Split Specific Loss Functions

	6000	2000	2000	
	Training	Validation	Testing	~10000
	$J(\theta)_{train}$	$J(\theta)_{val}$	$J(\theta)_{test}$	$J(\theta)_{Minimum}$
Good Estimator	0.21	0.25	0.30	0.18
Underfitting	1.9	1.9	2.1	
Overfitting	0.21 ⟷	2.05	2.1	

Reducing the Effect of Underfitting/Overfitting

- **Underfitting:** (Training loss is high)
  - Train longer
  - More layers or more parameters per layer
  - Change architecture
- **Overfitting:** (Generalization gap is large)
  - More training data
  - Regularization
  - Change architecture