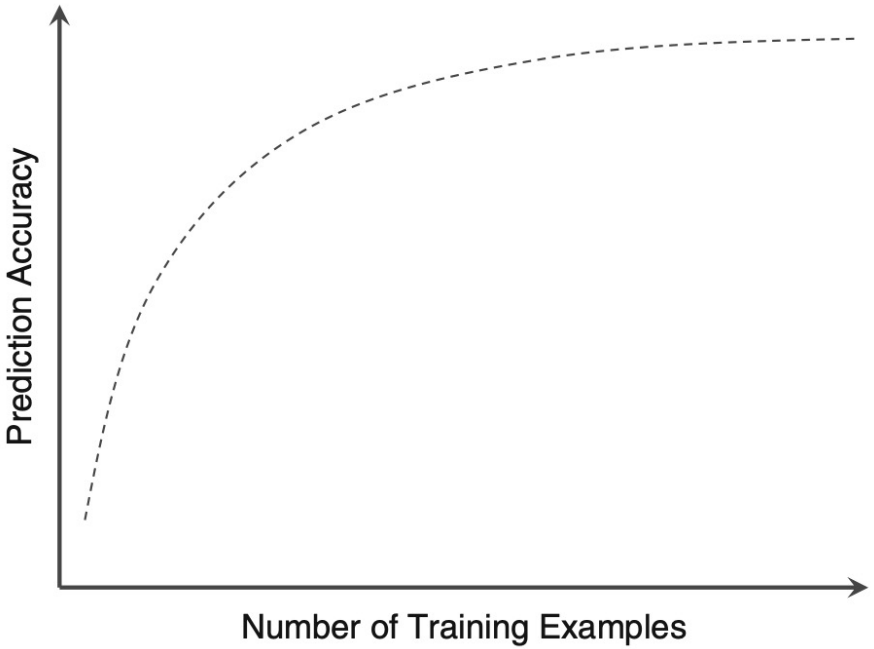
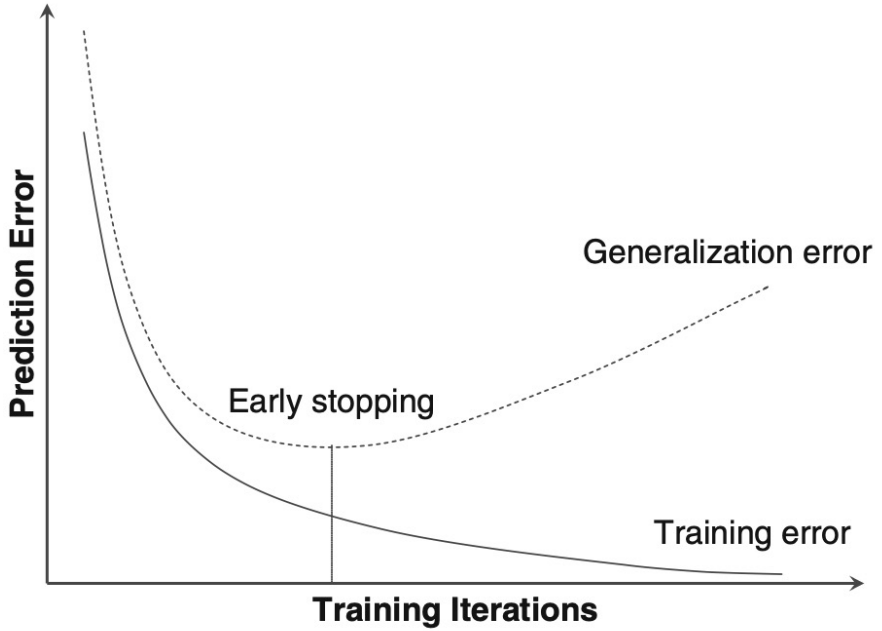
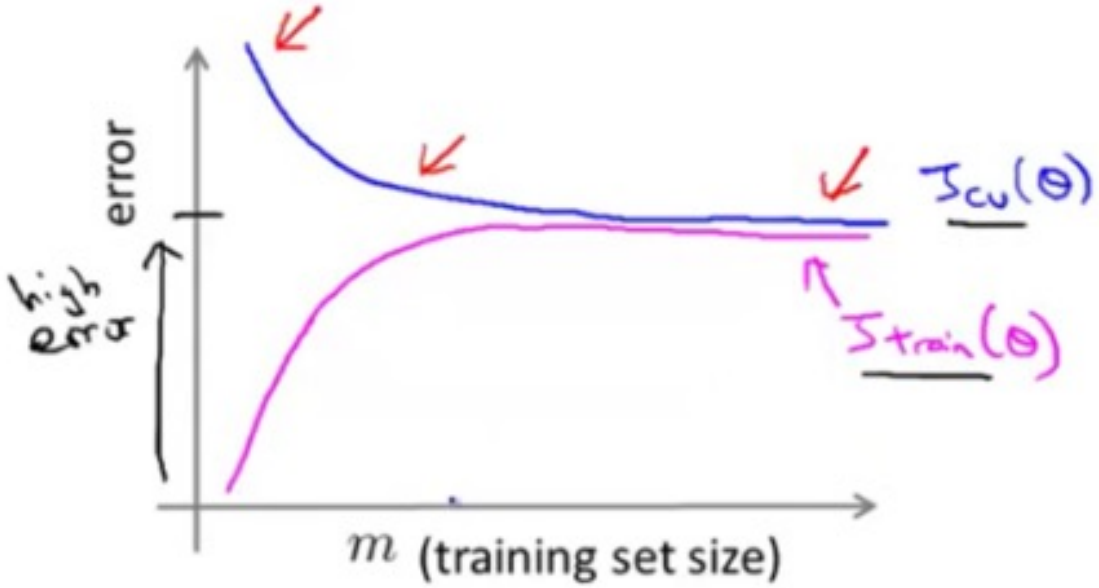
Encyclopedia of Machine Learning and Data Mining - 2nd



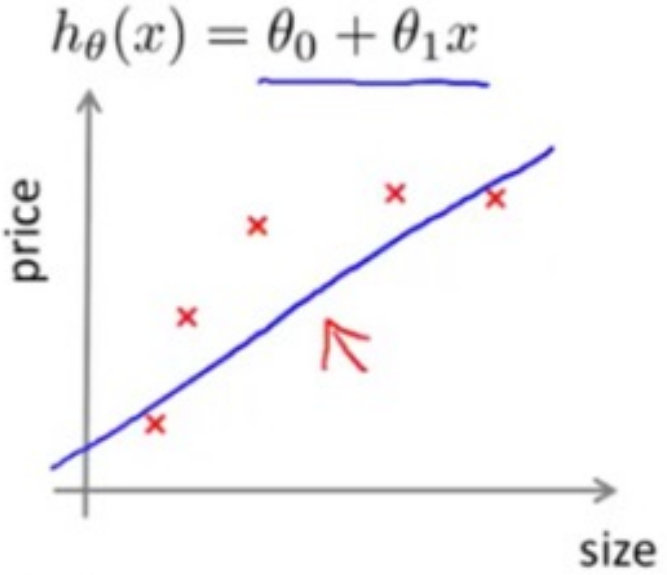



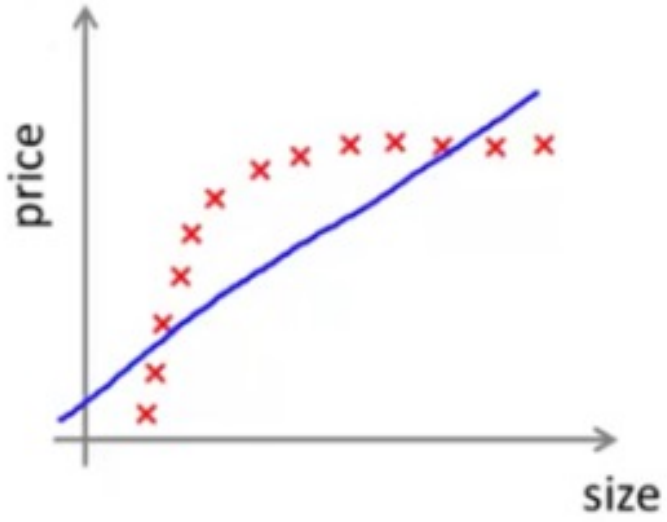
https://www.coursera.org/learn/machine-learning

## High bias

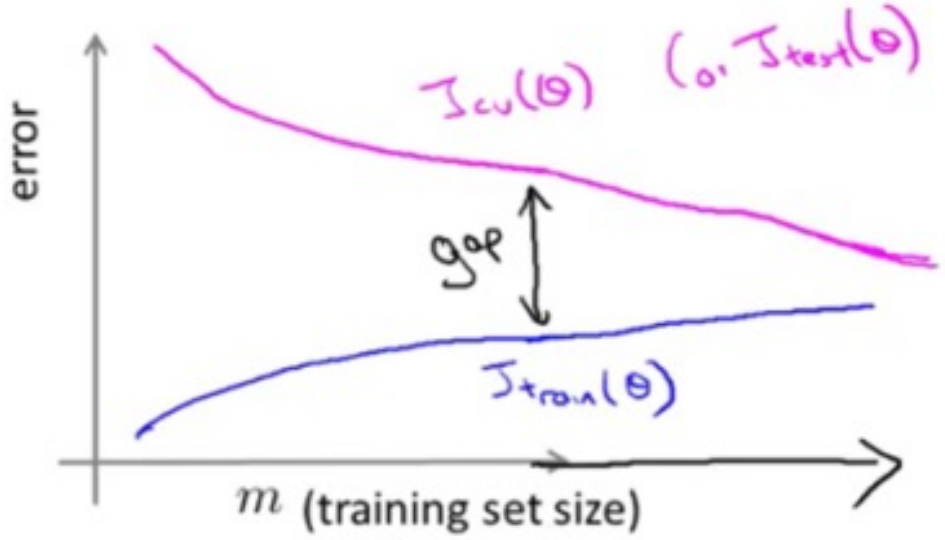



If a learning algorithm is suffering from high bias, getting more training data will not (by itself) help much.



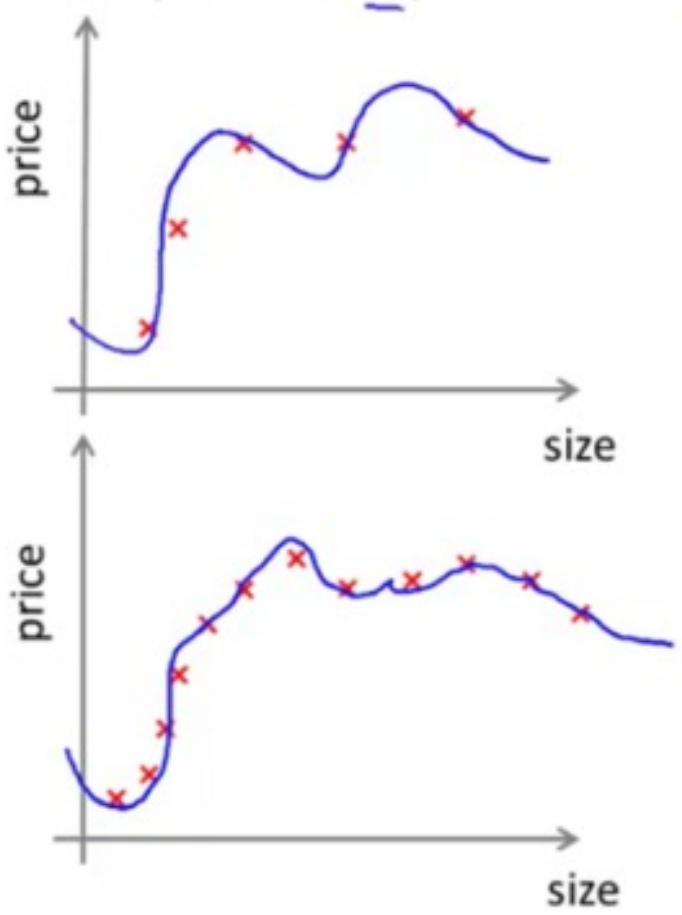
## High variance




If a learning algorithm is suffering from high variance, getting more training data is likely to help.

$$h_\theta(x) = \theta_0 + \theta_1 x + \cdots + \theta_{100} x^{100}$$

(and small $\lambda$)



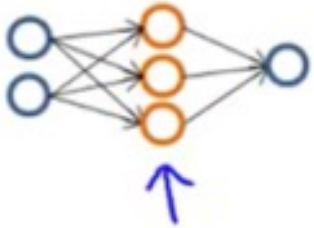
**Debugging a learning algorithm:**

Suppose you have implemented regularized linear regression to predict housing prices. However, when you test your hypothesis in a new set of houses, you find that it makes unacceptably large errors in its prediction. What should you try next?

- Get more training examples → fixes high variance
- Try smaller sets of features → fixes high variance
- Try getting additional features → fixes high bias
- Try adding polynomial features $(x_1^2, x_2^2, x_1x_2, \text{etc})$ → fixes high bias.
- Try decreasing $\lambda$ → fixes high bias
- Try increasing $\lambda$ → fixes high variance

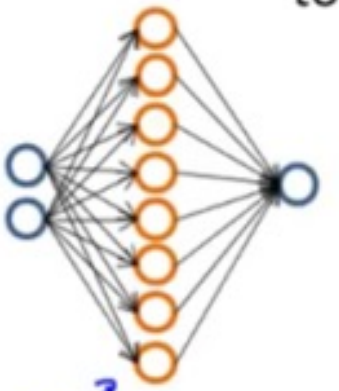
# Neural networks and overfitting

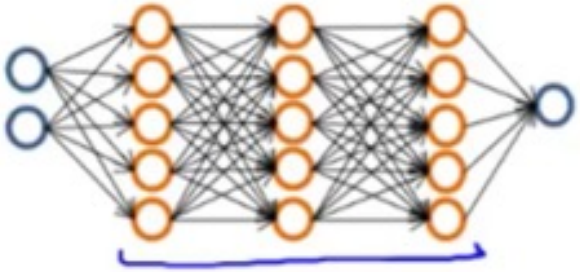
"Small" neural network
(fewer parameters; more prone to underfitting)



Computationally cheaper

"Large" neural network
(more parameters; more prone to overfitting)

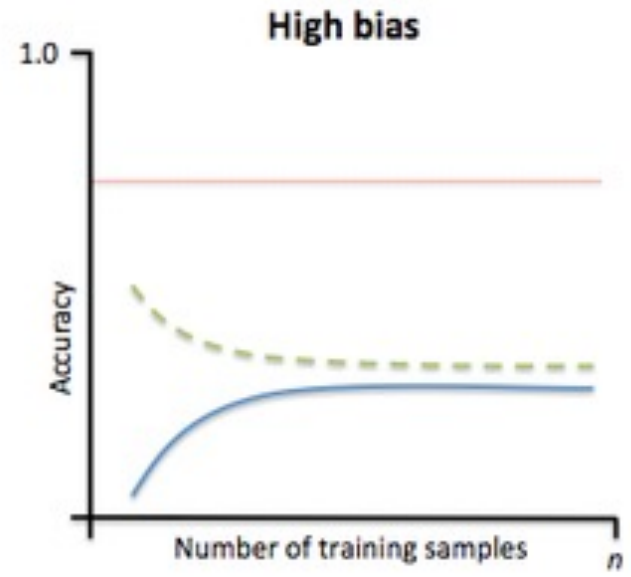


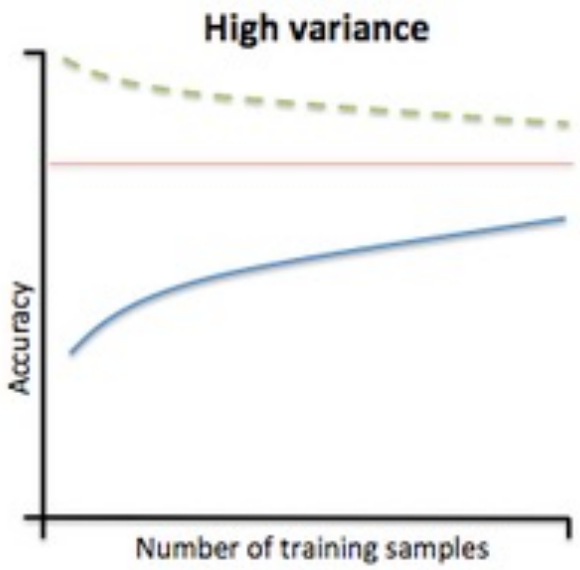


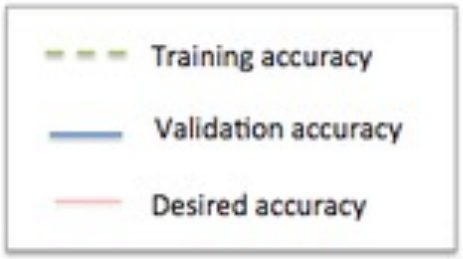
Computationally more expensive.

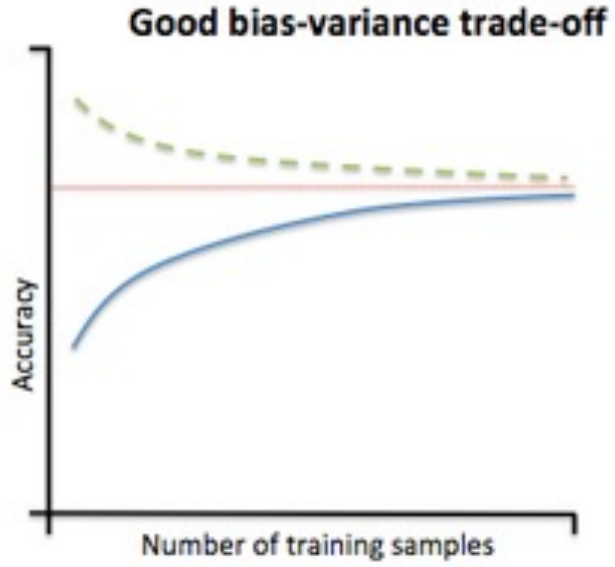
Use regularization ($\lambda$) to address overfitting.

https://sebastianraschka.com/faq/docs/ml-solvable.html



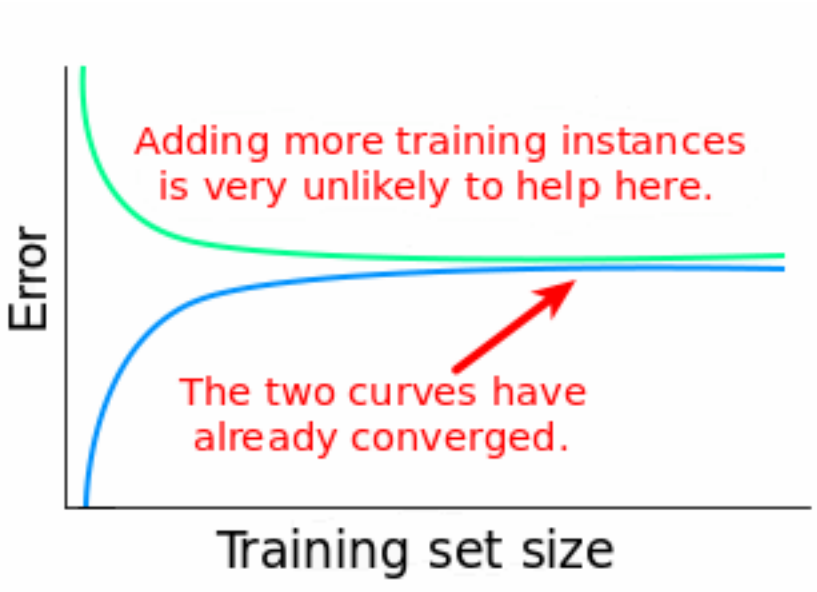



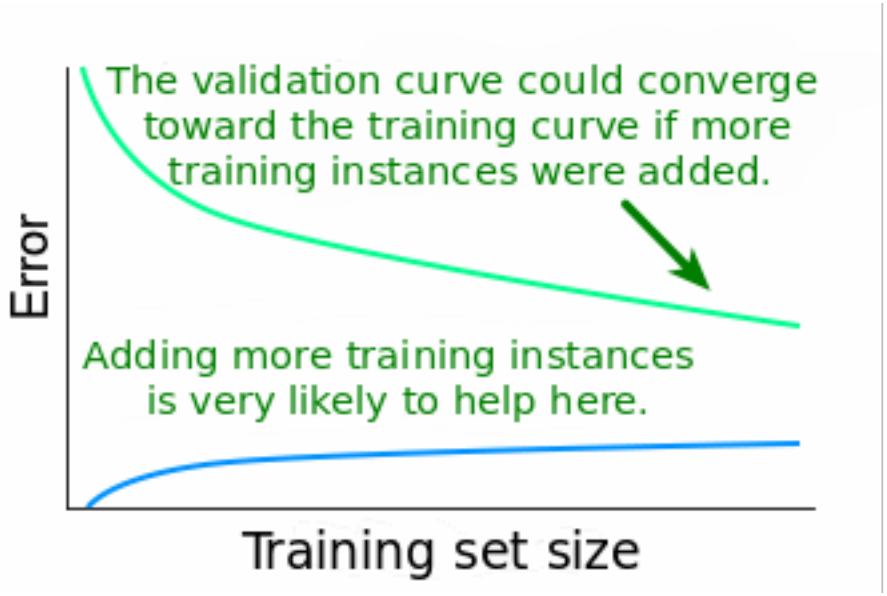



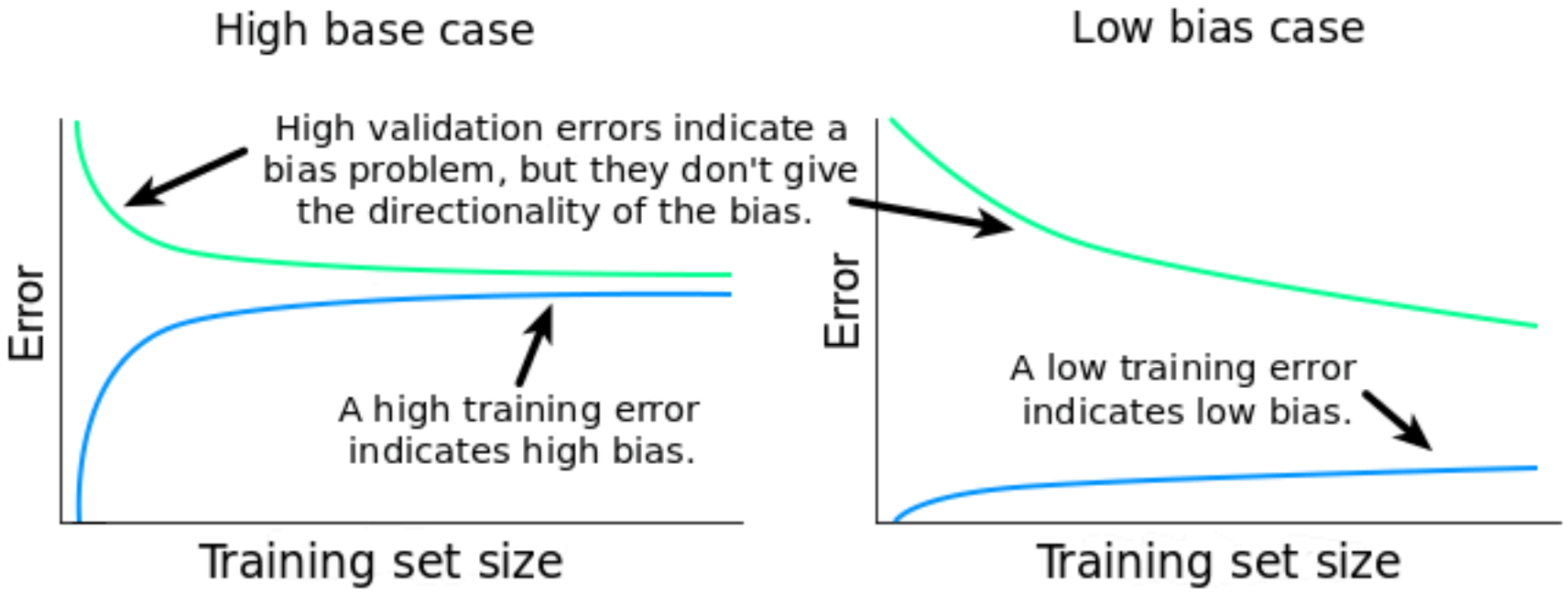



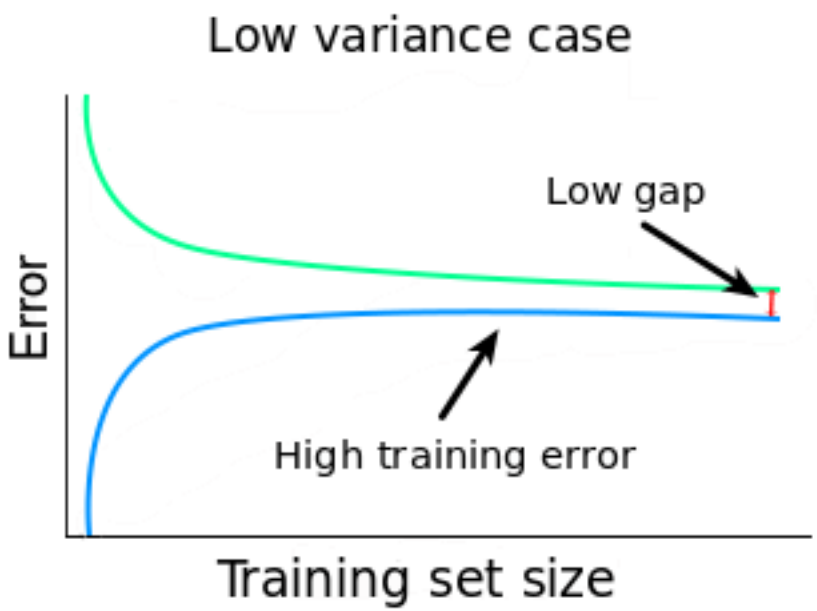
- our data is skewed
- there is a lot of noise
- there are many outliers
- our features are not informative enough
- we don’t have enough training samples

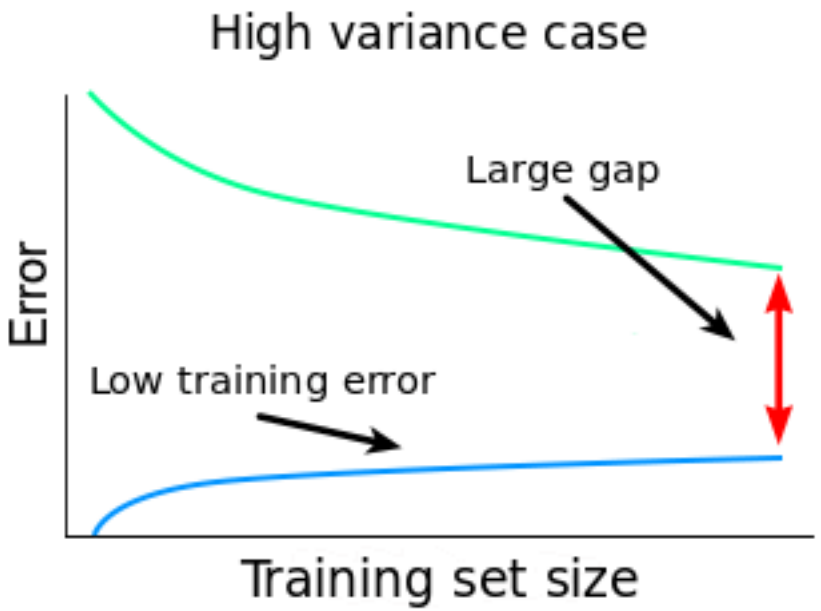
https://www.dataquest.io/blog/learning-curves-machine-learning/

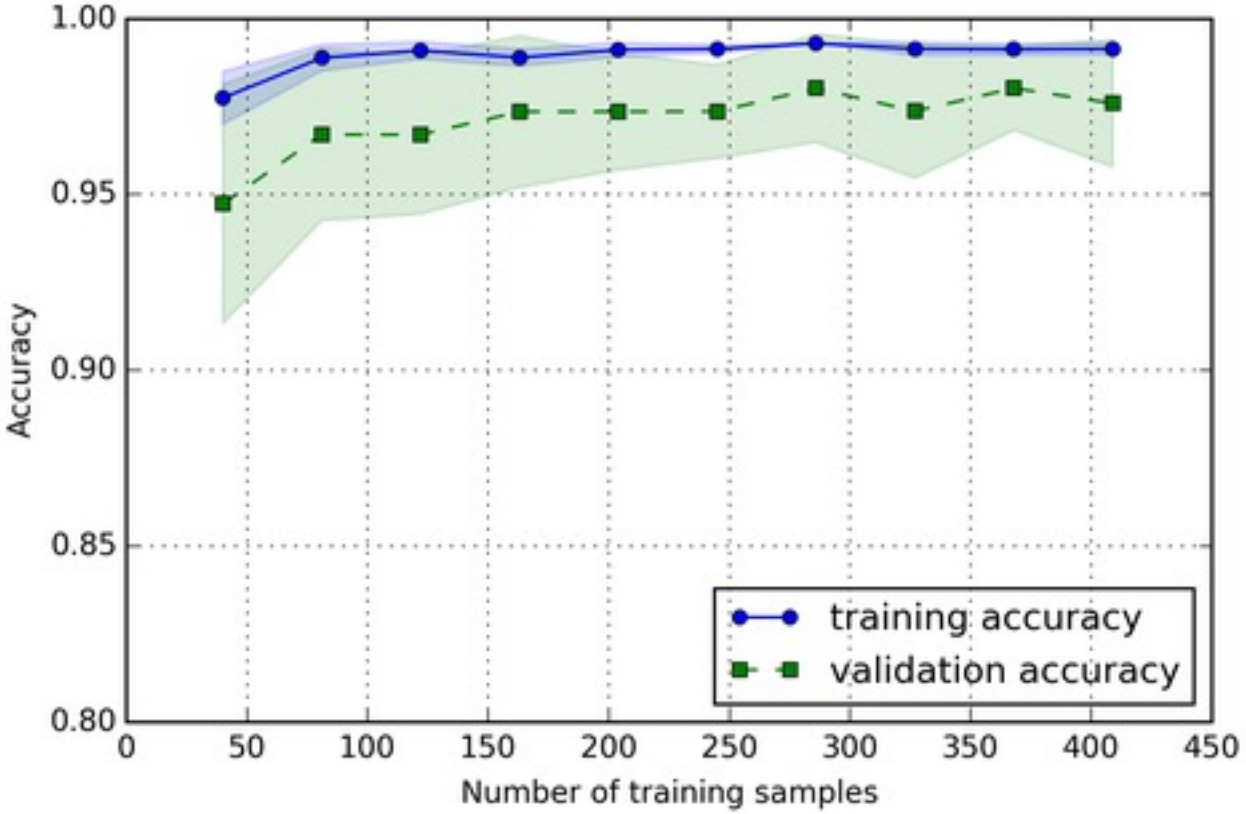


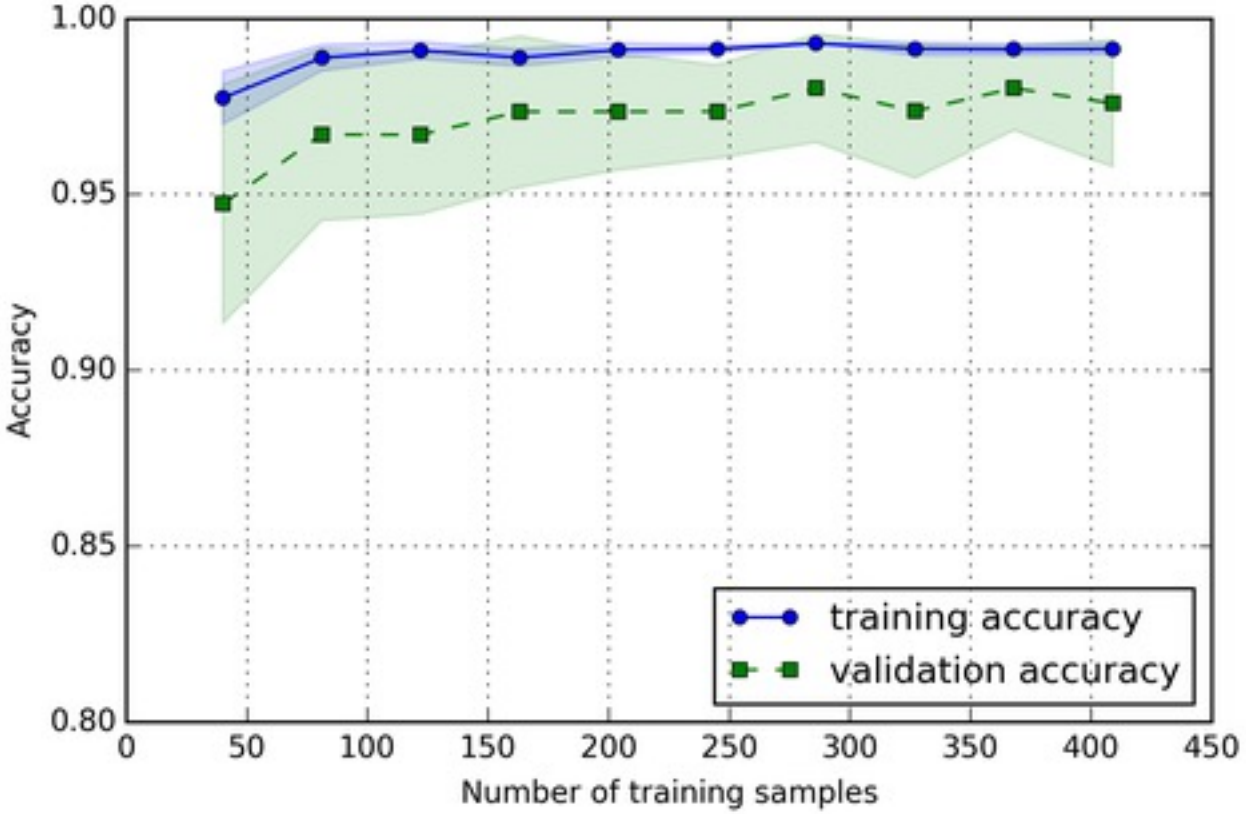




High base case
High validation errors indicate a bias problem, but they don't give the directionality of the bias.
A high training error indicates high bias.
Error
Training set size
Low bias case
A low training error indicates low bias.
Error
Training set size

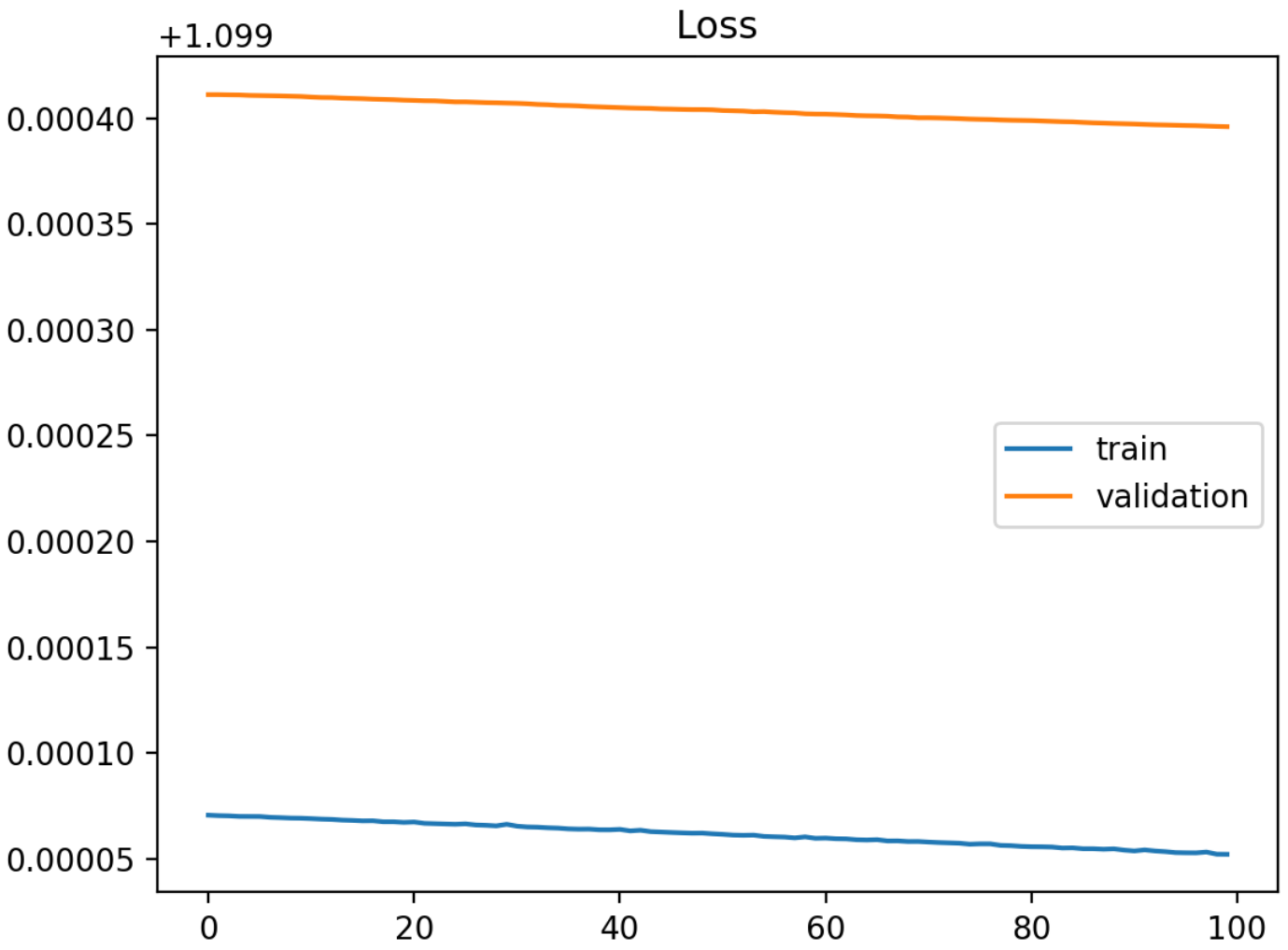

Low variance case
Low gap
Error
High training error
Training set size


High variance case
Large gap
Error
Low training error
Training set size

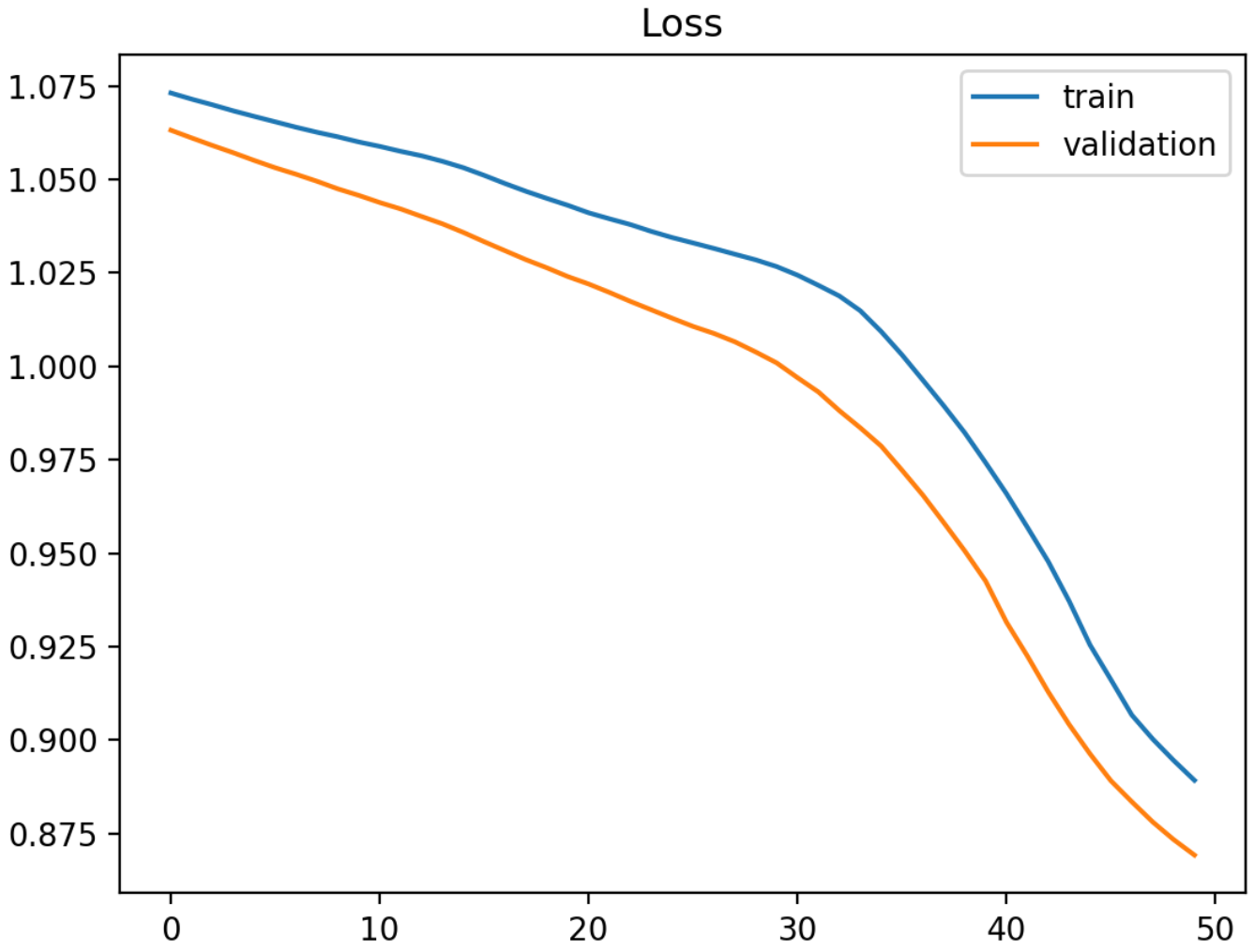

1.00
0.95
0.90
0.85
0.80
Accuracy
0
50
100
150
200
250
300
350
400
450
Number of training samples
training accuracy
validation accuracy


1.00
0.95
0.90
0.85
0.80
Accuracy
0
50
100
150
200
250
300
350
400
450
Number of training samples
training accuracy
validation accuracy

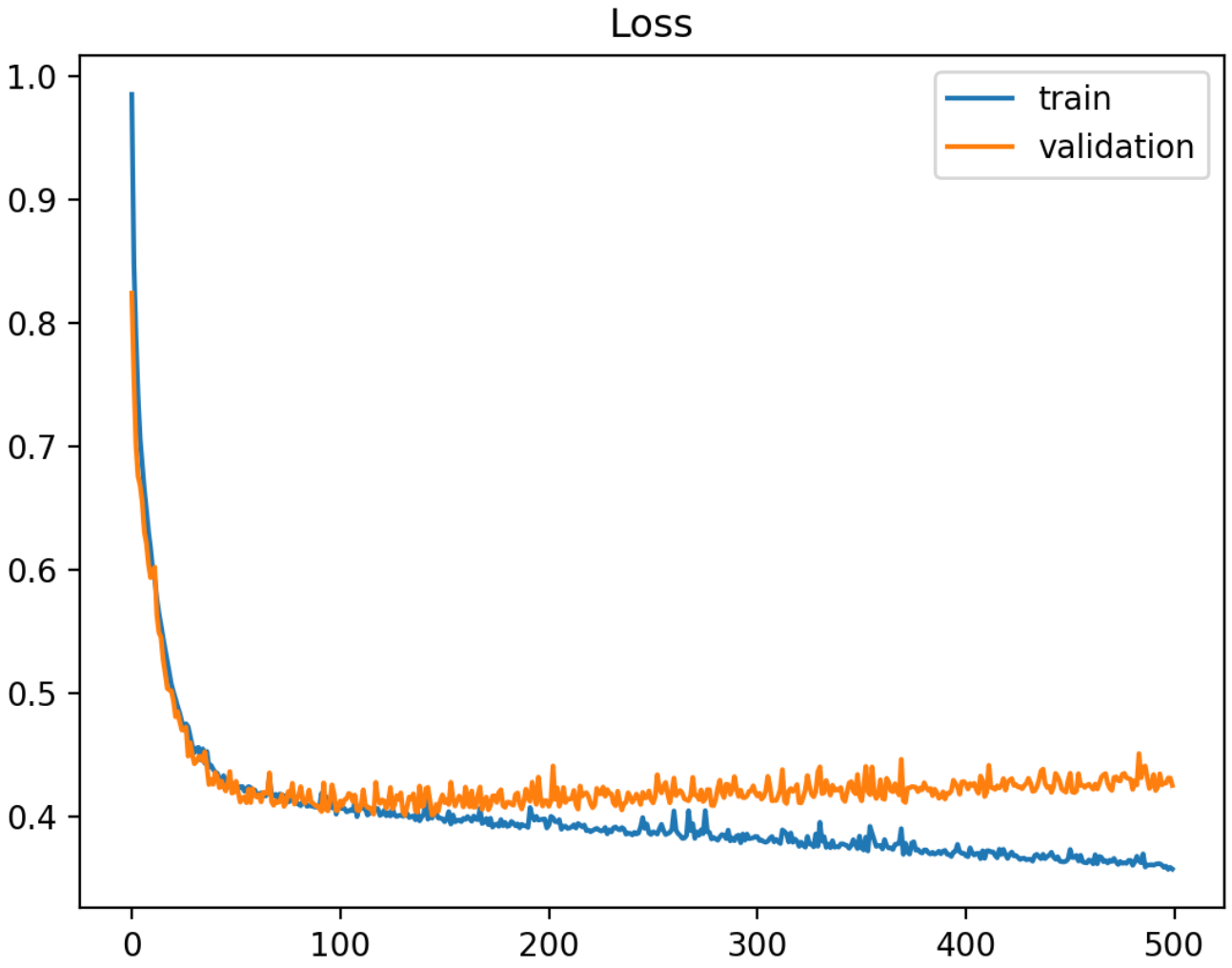
# Underfit (1)


Loss
+1.099
0.00040
0.00035
0.00030
0.00025
0.00020
0.00015
0.00010
0.00005
0
20
40
60
80
100
train
validation

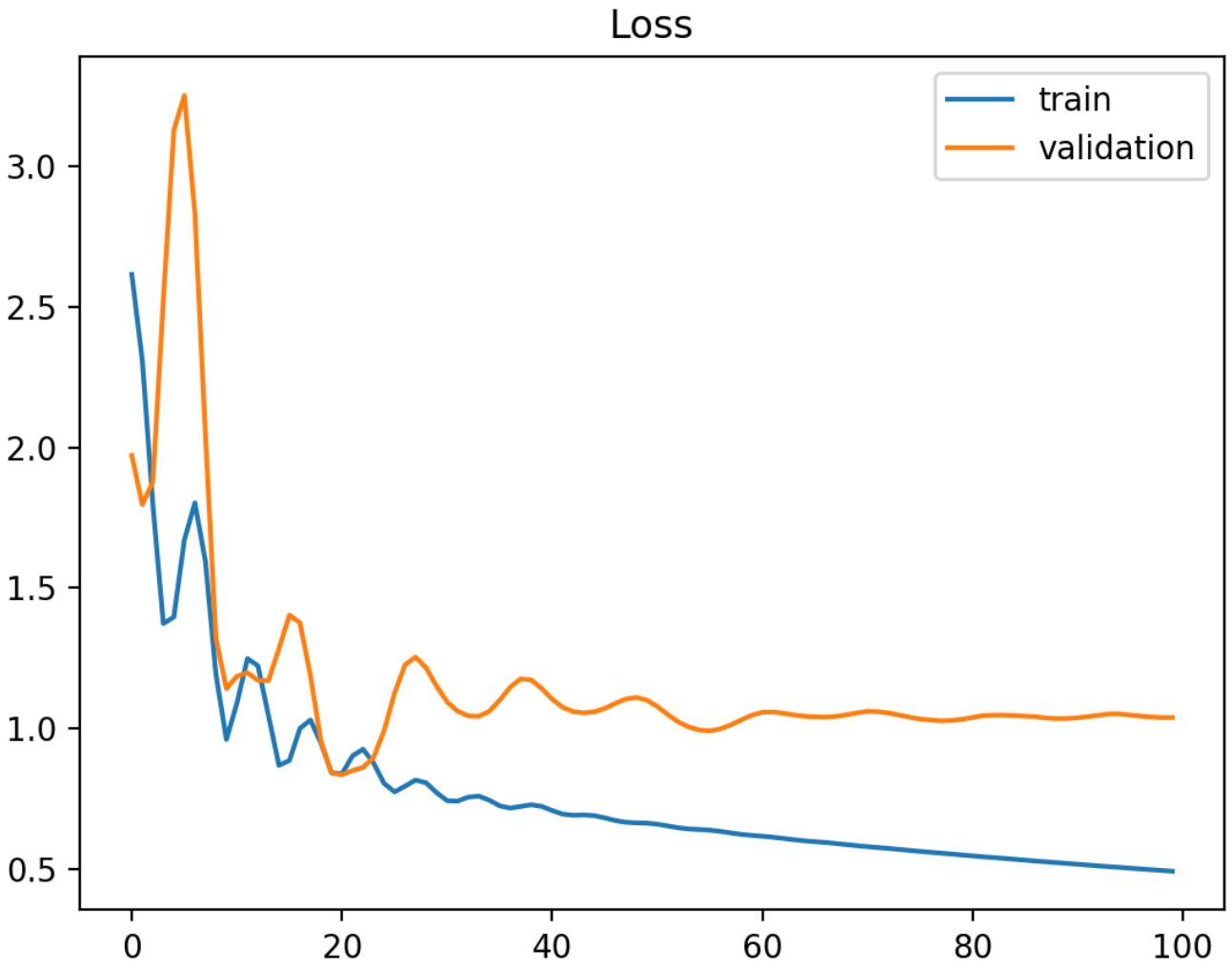
# Underfit (2)


Loss
train
validation
1.075
1.050
1.025
1.000
0.975
0.950
0.925
0.900
0.875
0
10
20
30
40
50

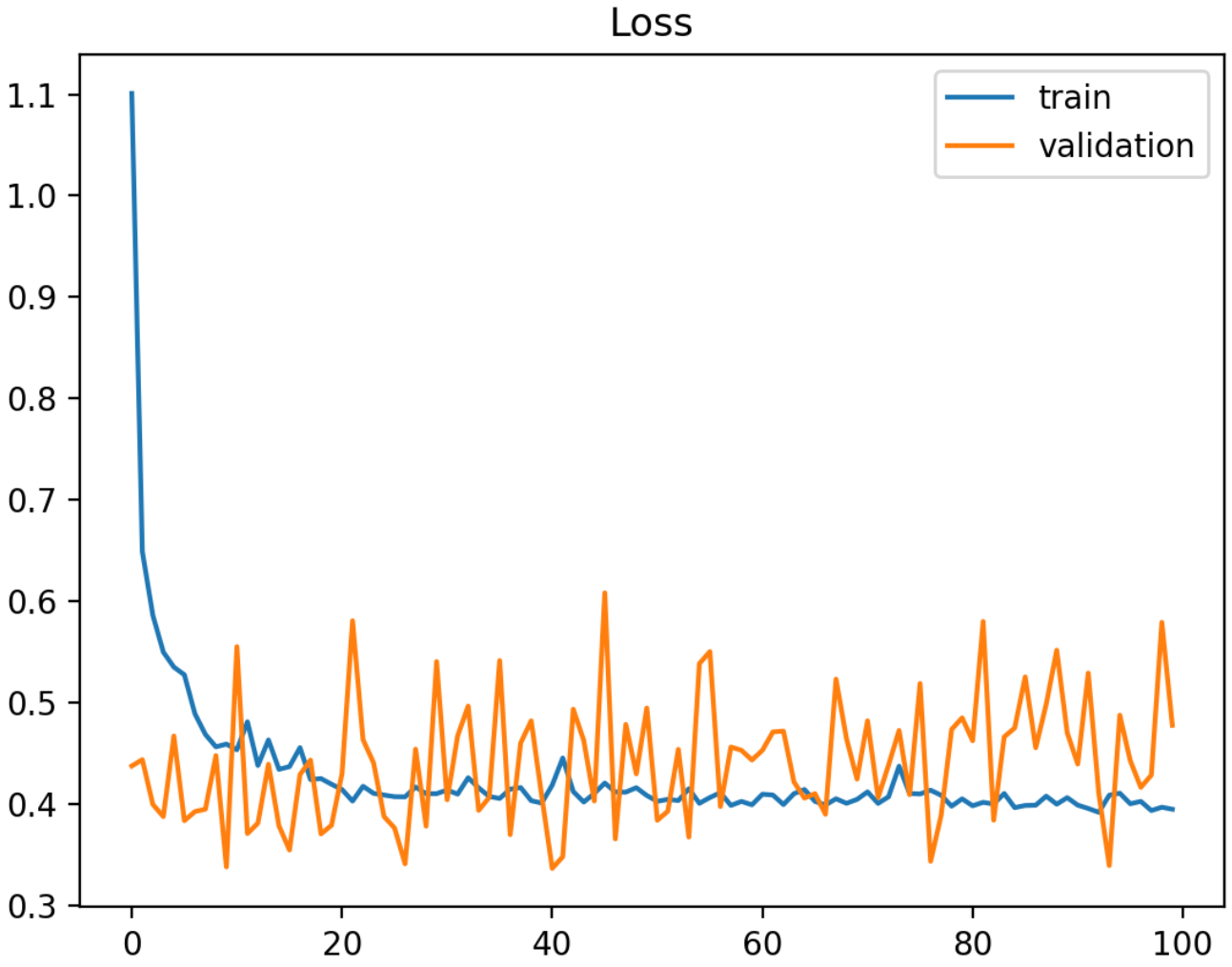
# Overfit (1)



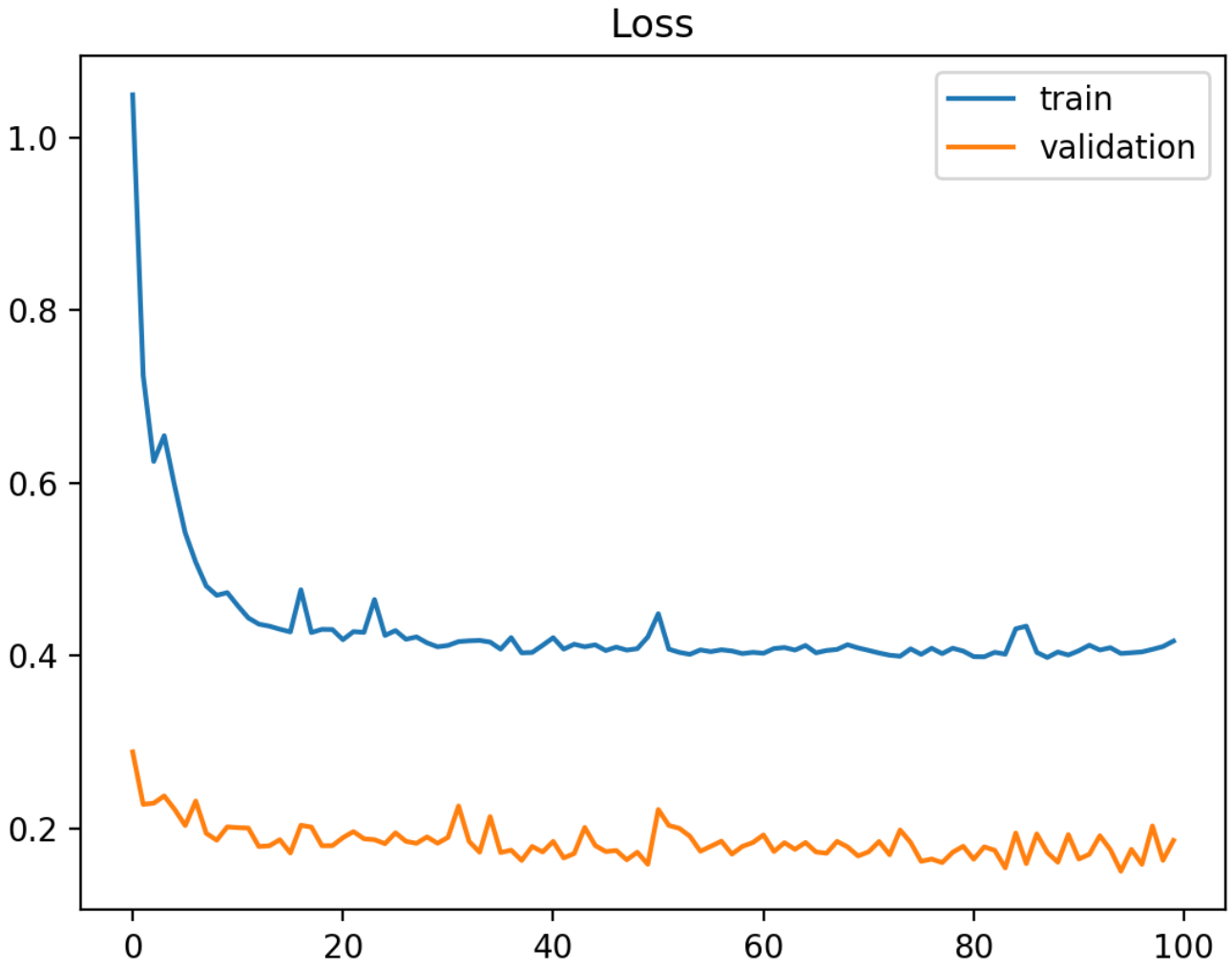
# Unrepresentative Train Dataset



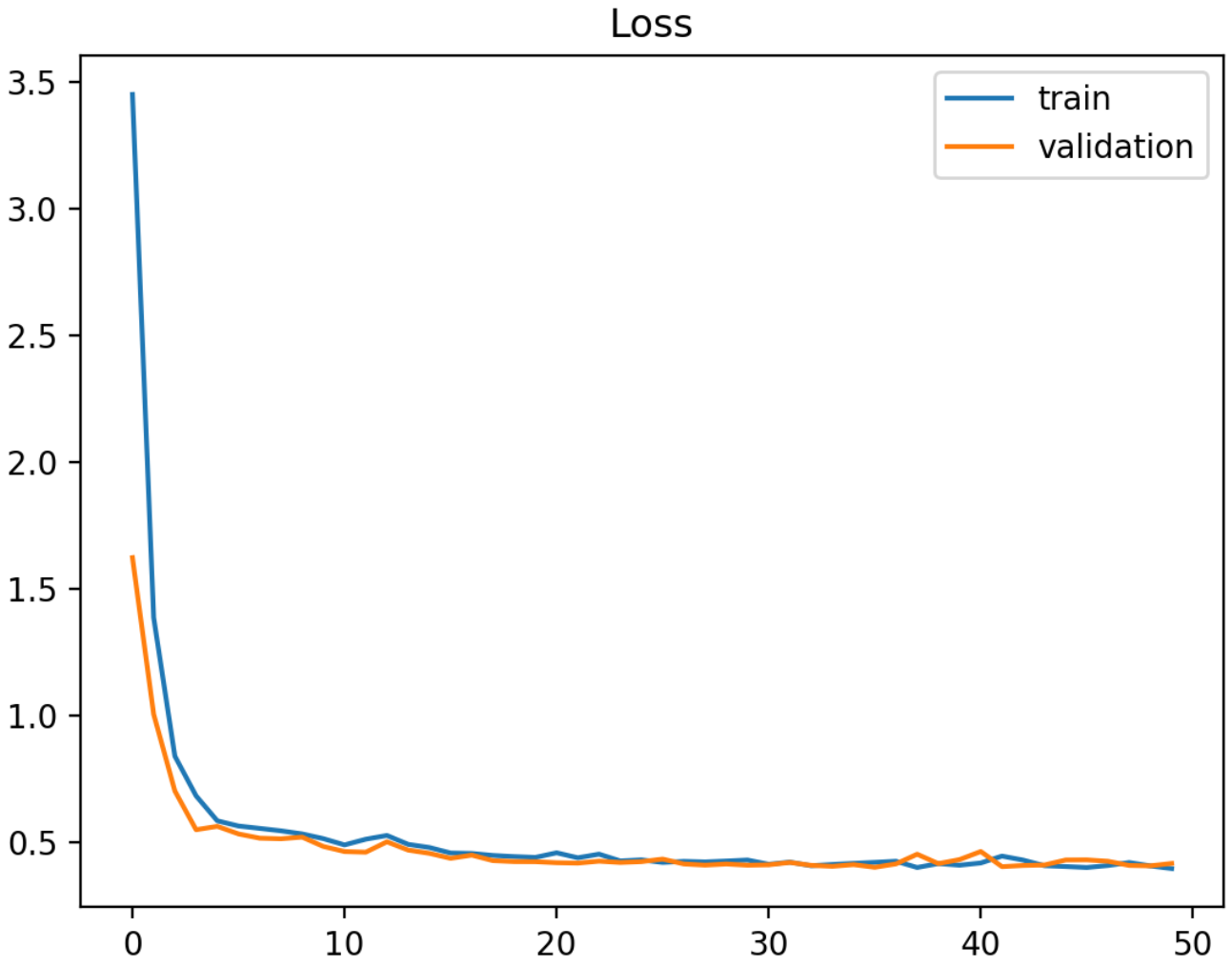
# Unrepresentative Validation Dataset (1)

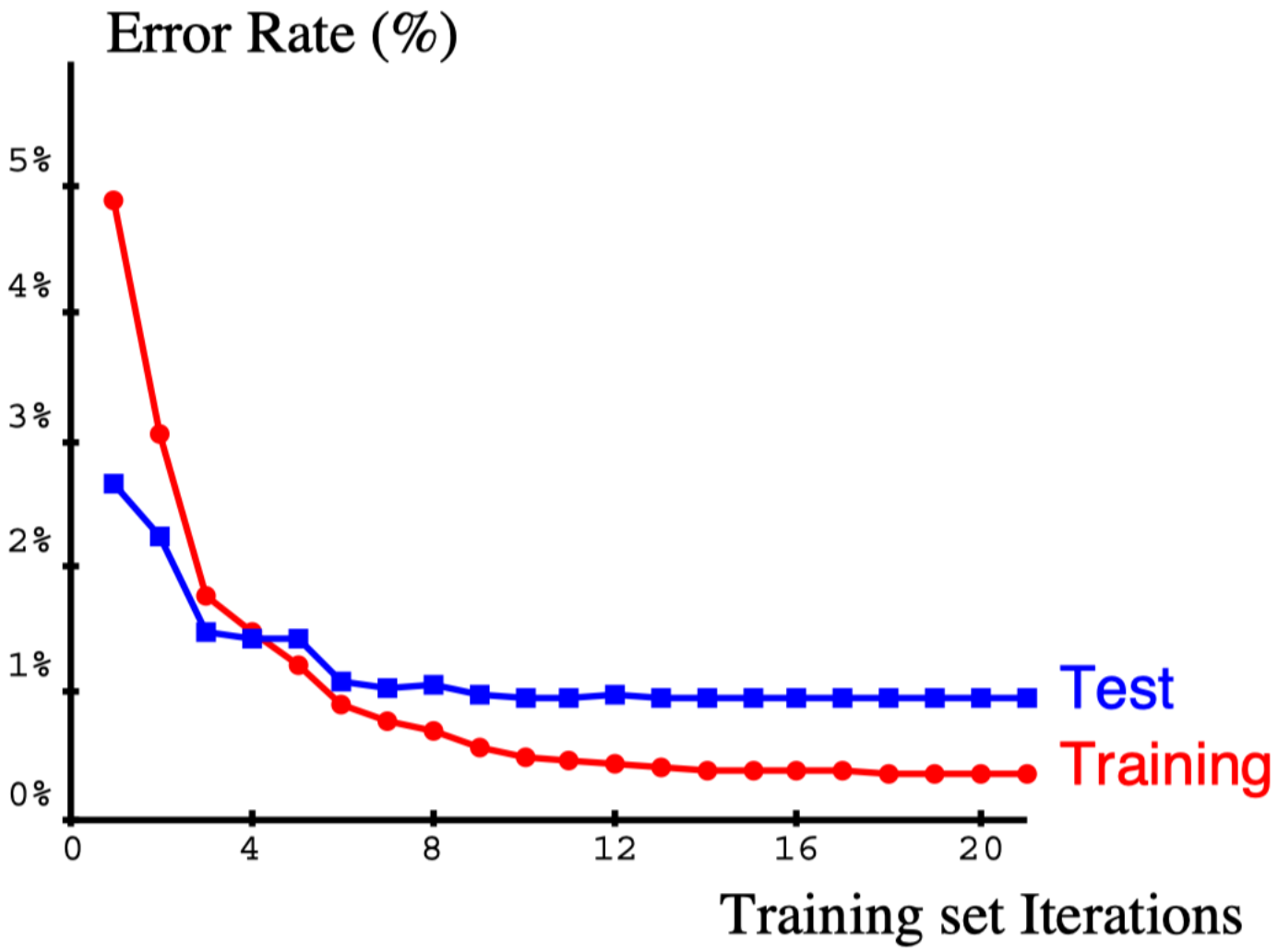


# Unrepresentative Validation Dataset (2)




the validation dataset may be easier for the model to predict than the training dataset


Loss
train
validation
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0
10
20
30
40
50


Error Rate (%)
5%
4%
3%
2%
1%
0%
0
4
8
12
16
20
Training set Iterations
Test
Training