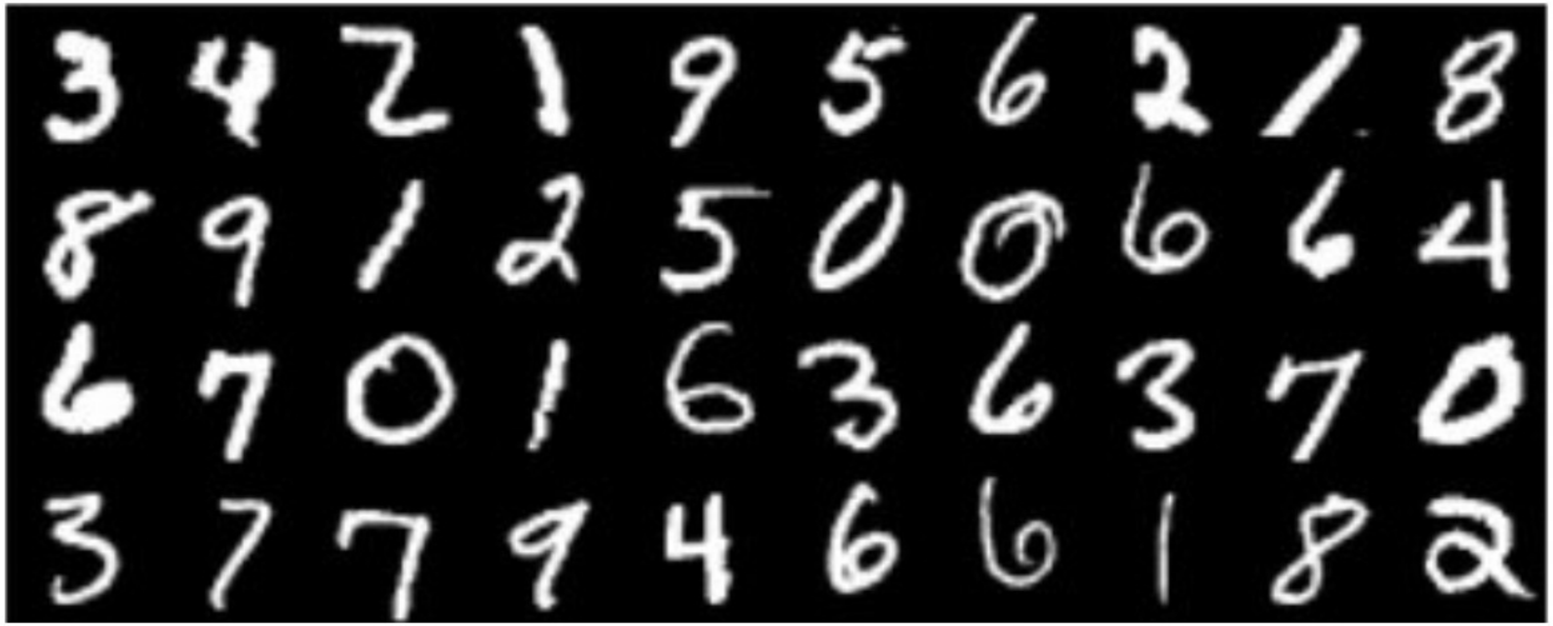
# Measuring Performance

Dr. Abdul Samad

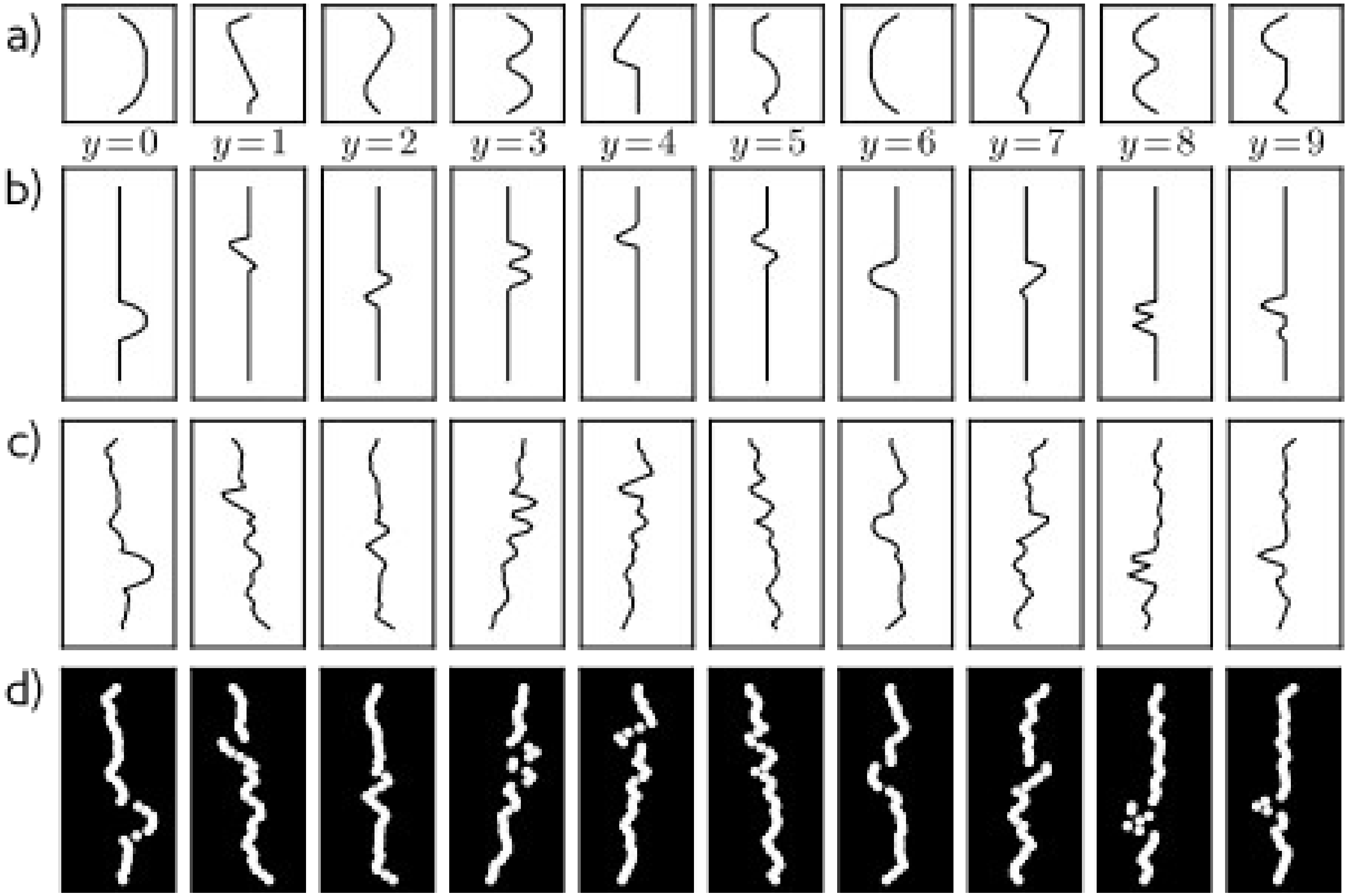
# Measuring performance

- MNIST1D dataset model and performance
- Noise, bias, and variance
- Reducing variance
- Reducing bias & bias-variance trade-off
- Double descent
- Curse of dimensionality & weird properties of high dimensional space
- Choosing hyperparameters

# MNIST Dataset



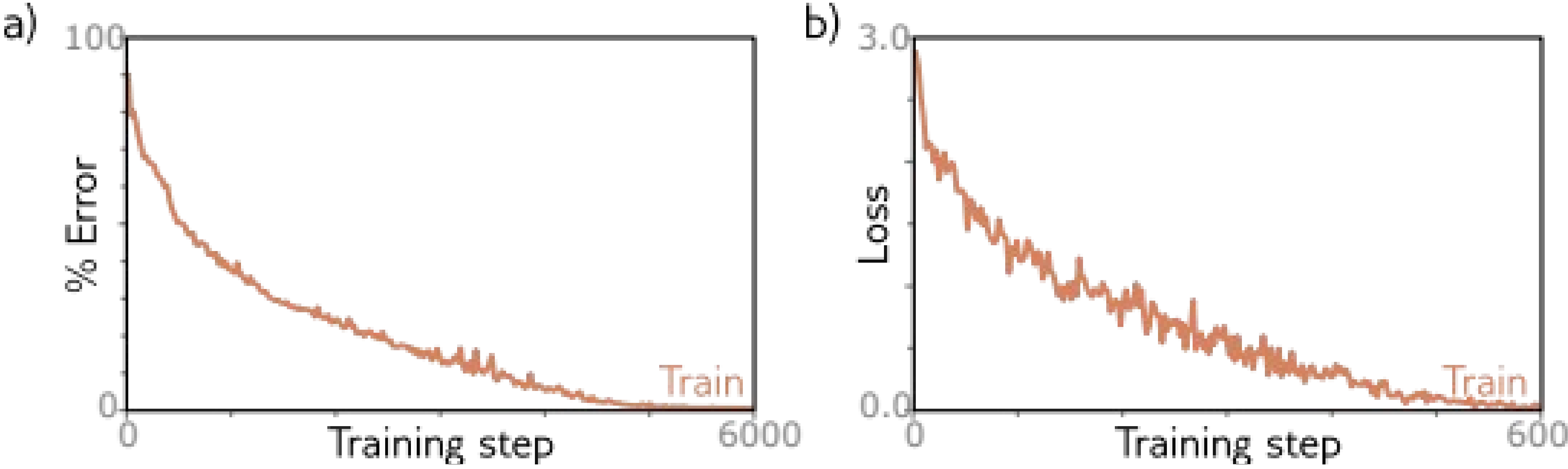
# MNIST 1D Dataset



# Network

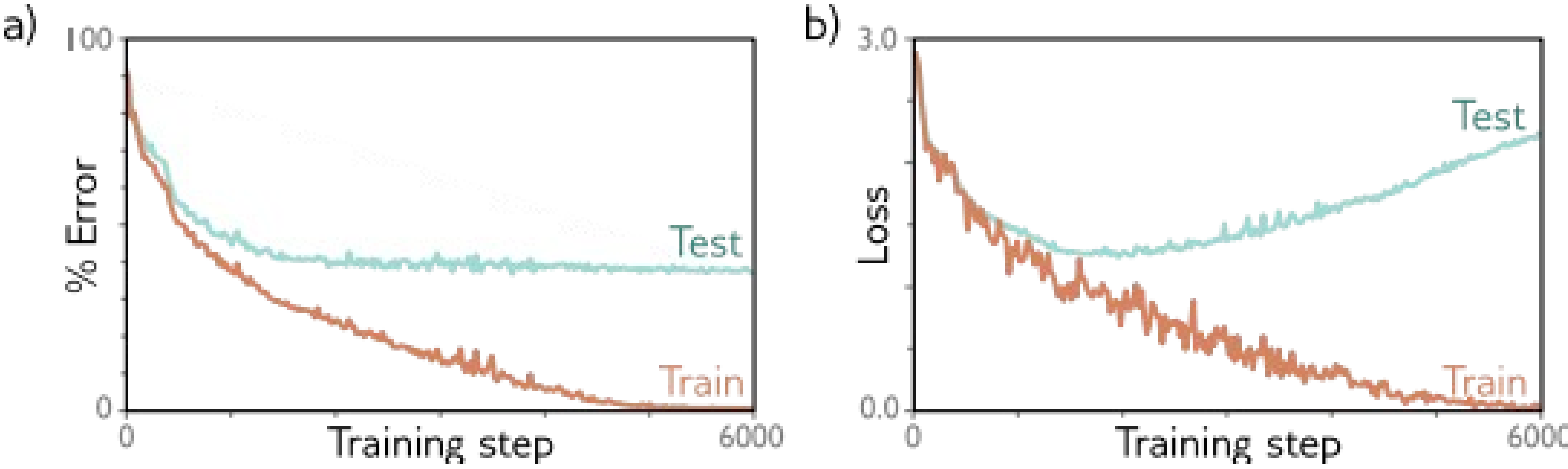
- 40 inputs
- 10 outputs
- 4000 training examples (~400 training examples per class)
- Two hidden layers
  - 100 hidden units each
- SGD with batch size 100, learning rate 0.1
- 6000 steps (?? Epochs)

# Results

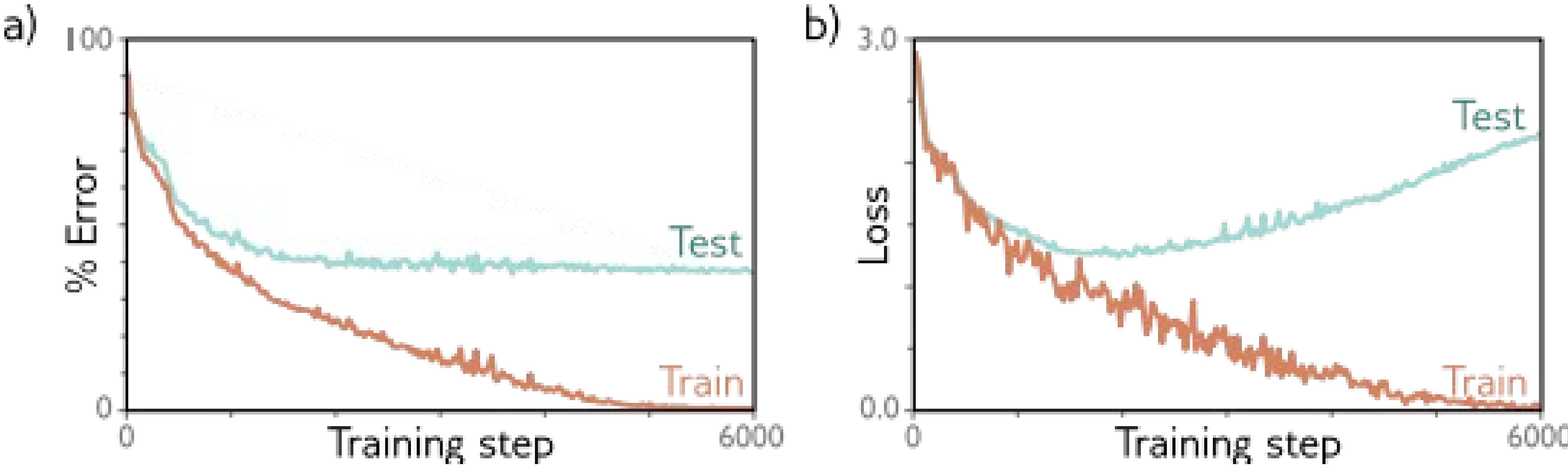


a)
100
% Error
0
0
Training step
6000
Train
b)
3.0
Loss
0.0
0
Training step
600
Train

# Need to use separate test data



# Need to use separate test data

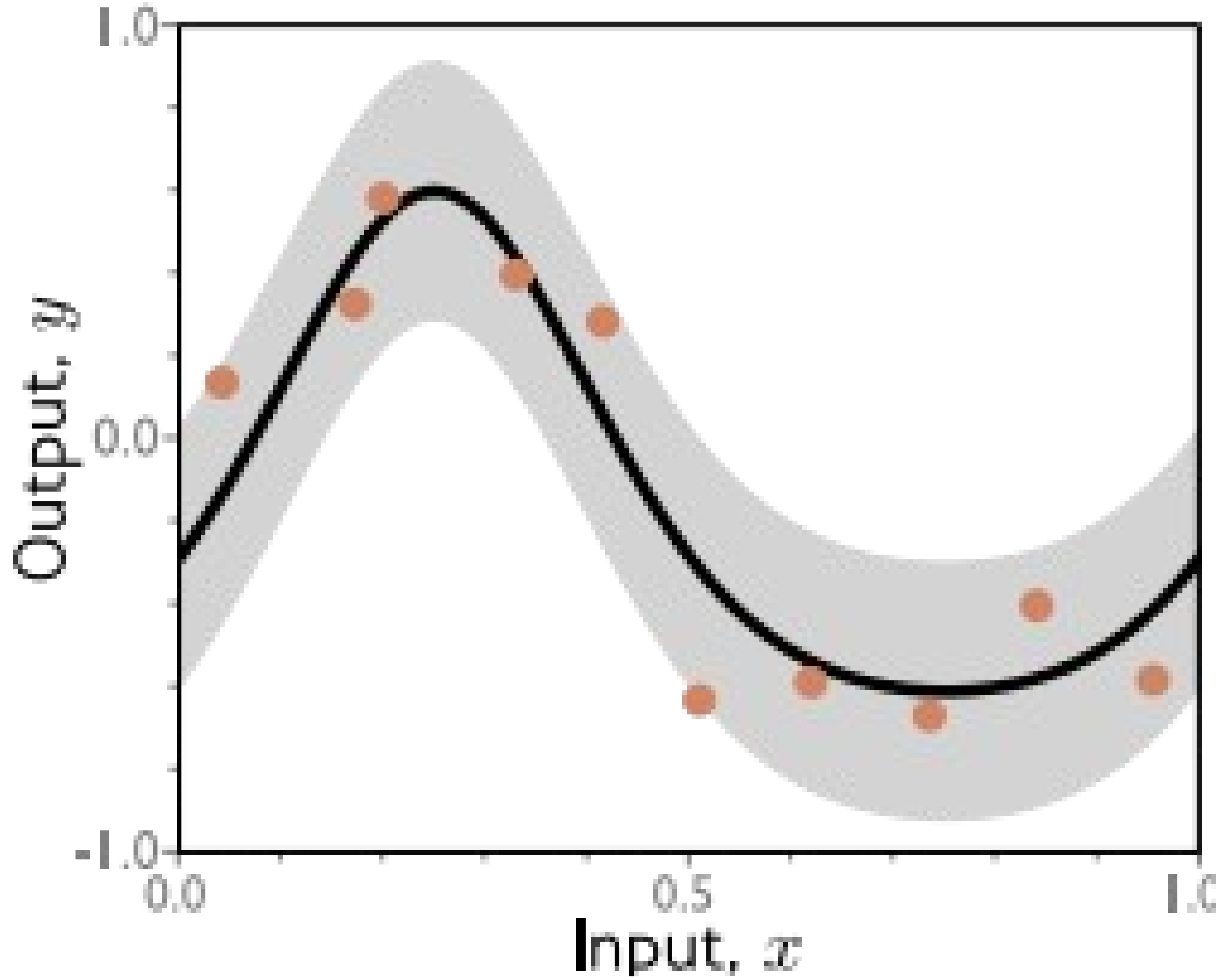



The model has not generalized well to the new data

# Measuring performance

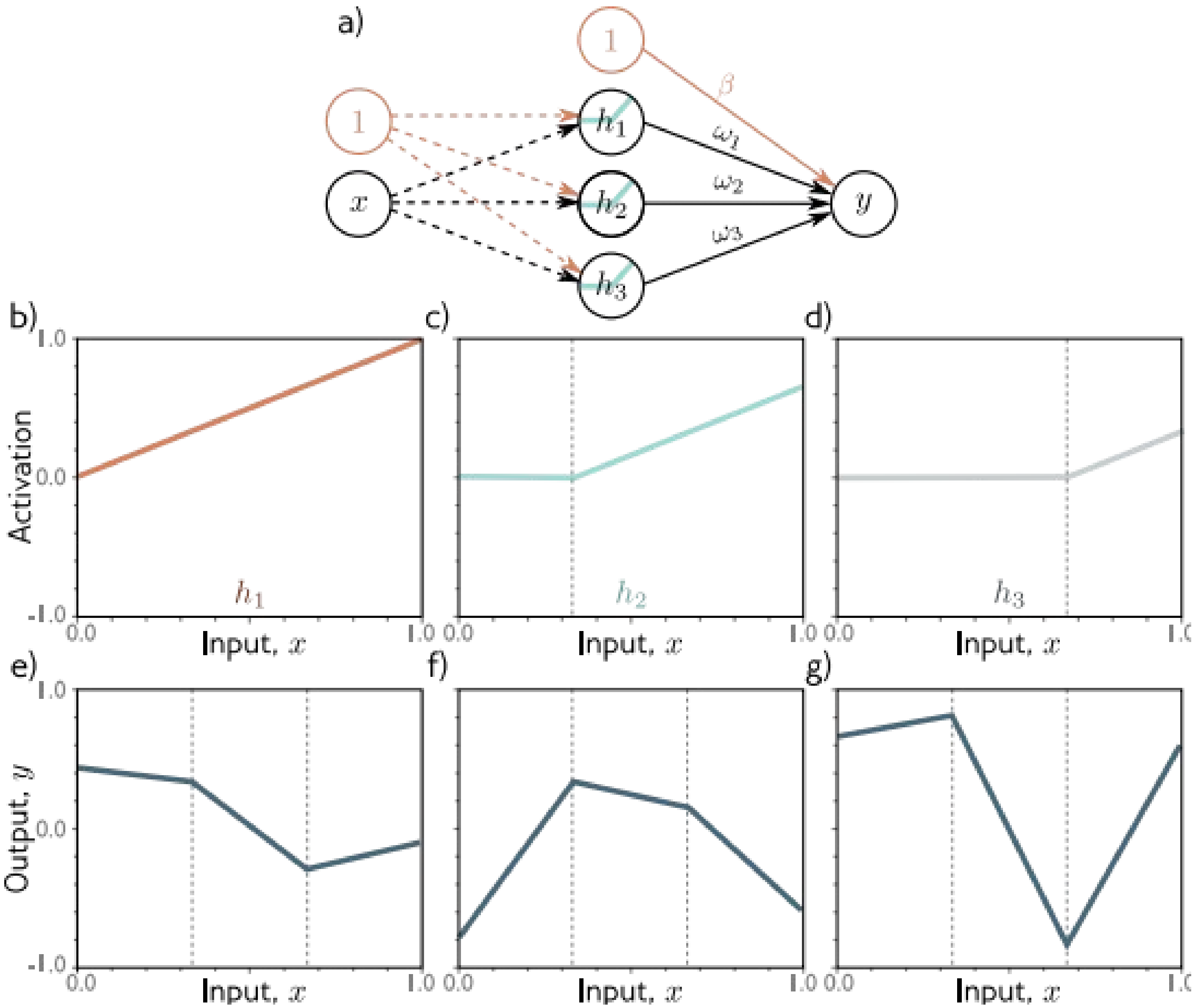
- MNIST1D dataset model and performance
- Noise, bias, and variance
- Reducing variance
- Reducing bias & bias-variance trade-off
- Double descent
- Curse of dimensionality & weird properties of high dimensional space
- Choosing hyperparameters

# Regression example

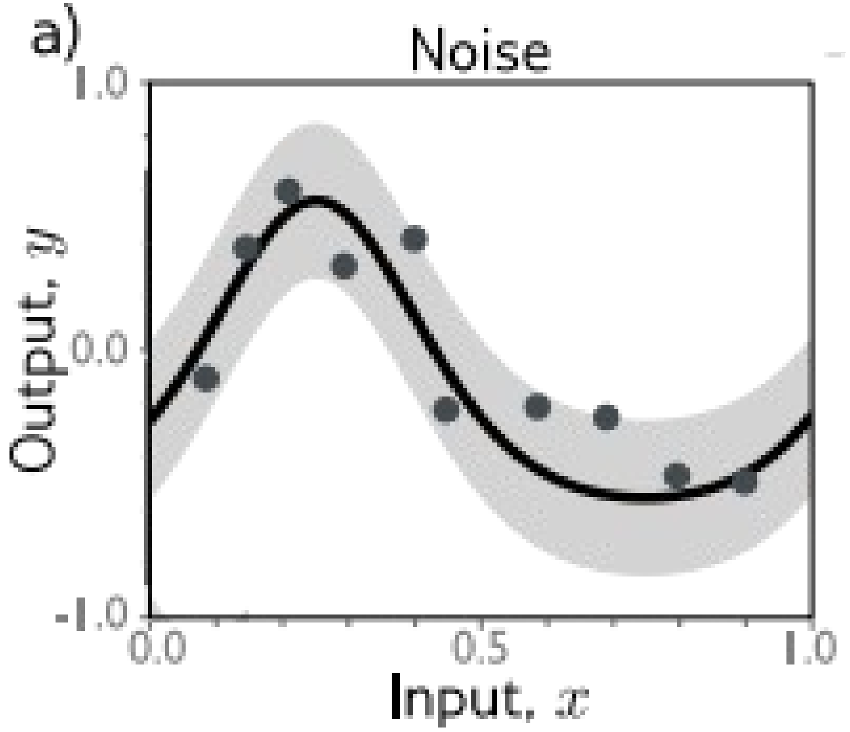


# Toy model

- K hidden units
- First layer fixed so "joints" divide interval evenly
- Second layer trained
- But… now linear in **h**
    - so convex cost function
    - can find best soln in closed-form

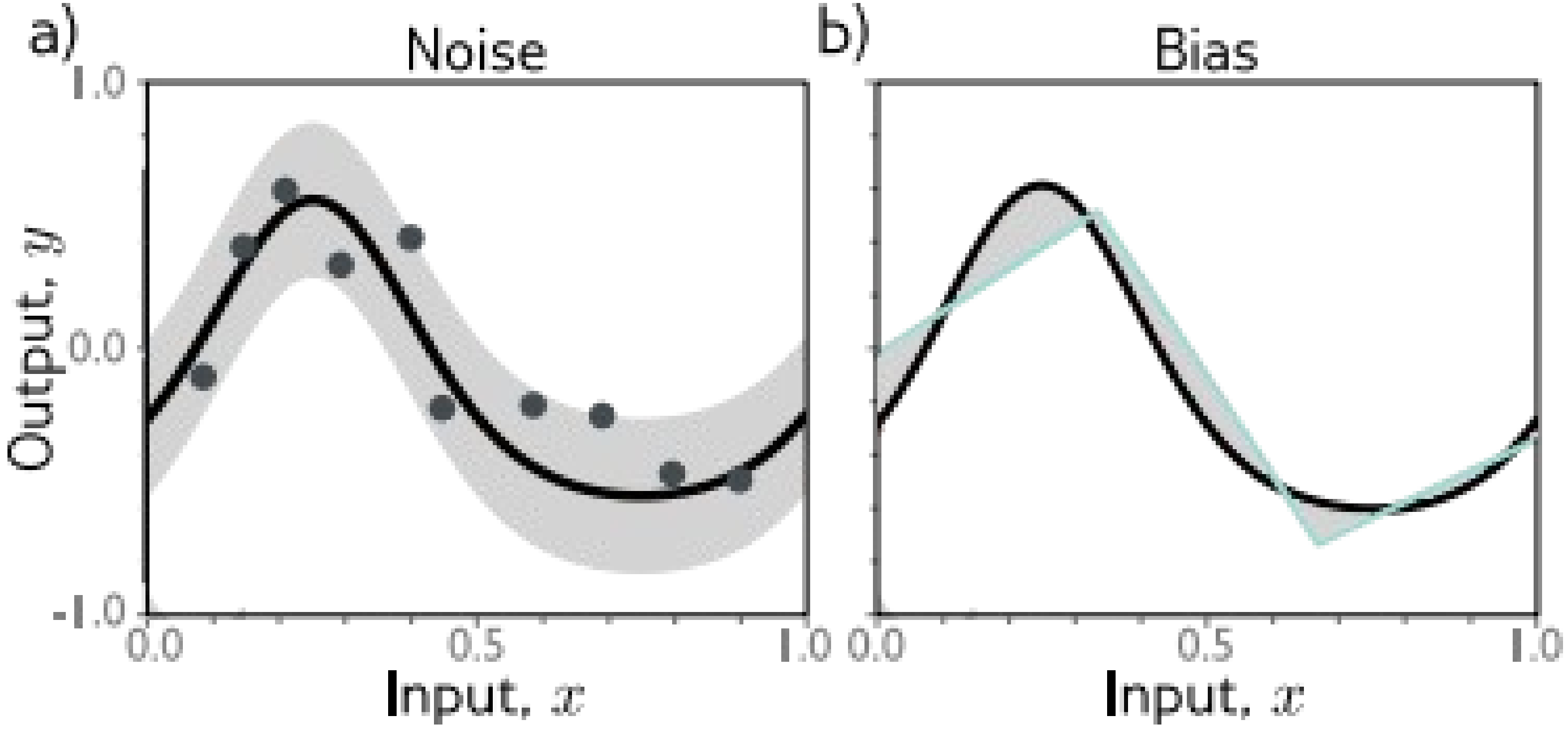


# Noise, bias, and variance

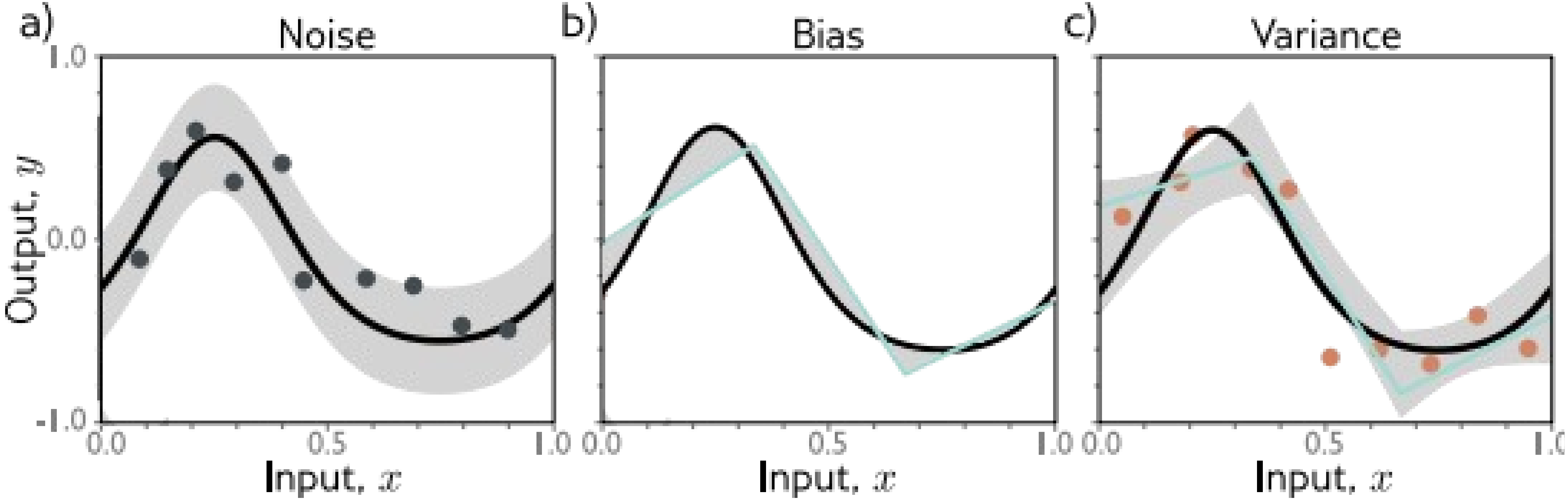



- Noise in measurements
- Some variables not observed
- Data mislabeled

# Noise, bias, and variance



# Noise, bias, and variance



# Noise, bias, and variance

- Variance is the uncertainty in fitted model due to choice of training set
- Bias is systematic deviation from the mean of the function we are modeling due to limitations in our model
- Noise is inherent uncertainty in the true mapping from input to output

# Least squares regression only

$$L[x] = \left(\mathrm{f}[x, \phi] - y[x]\right)^2$$

- We can show that:

$$\mathbb{E}_{\mathcal{D}}\Big[\mathbb{E}_y[L[x]]\Big] = \underbrace{\mathbb{E}_{\mathcal{D}}\Big[\left(\mathrm{f}[x, \phi[\mathcal{D}]] - f_{\mu}[x]\right)^2\Big]}_{\text{variance}} + \underbrace{\left(f_{\mu}[x] - \mu[x]\right)^2}_{\text{bias}} + \underbrace{\sigma^2}_{\text{noise}}$$

Expectation over noise in training data

Expectation over noise in test data

Actual model

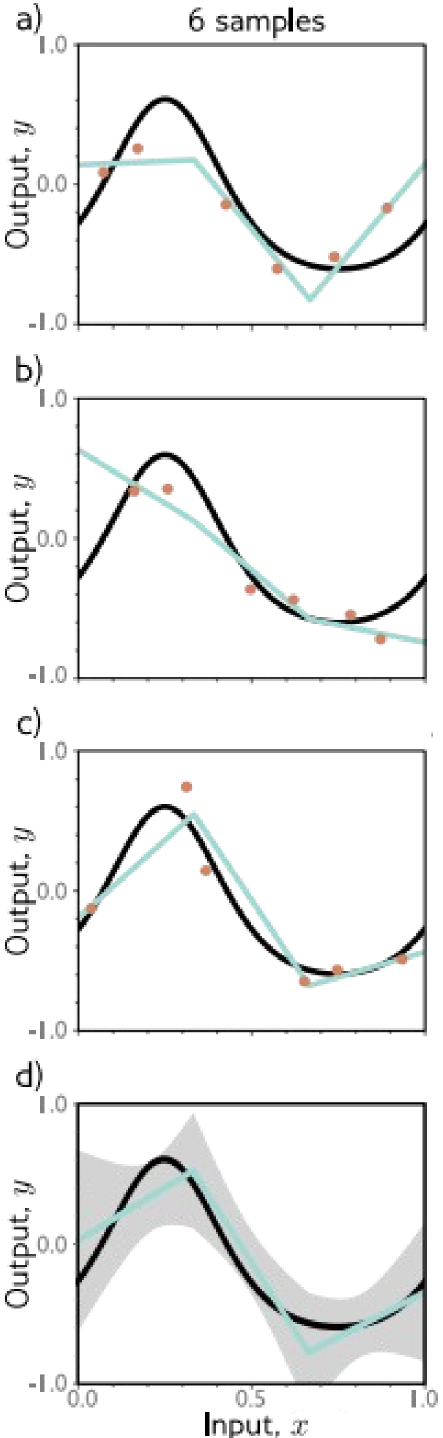
Best possible model if we had infinite data

True function

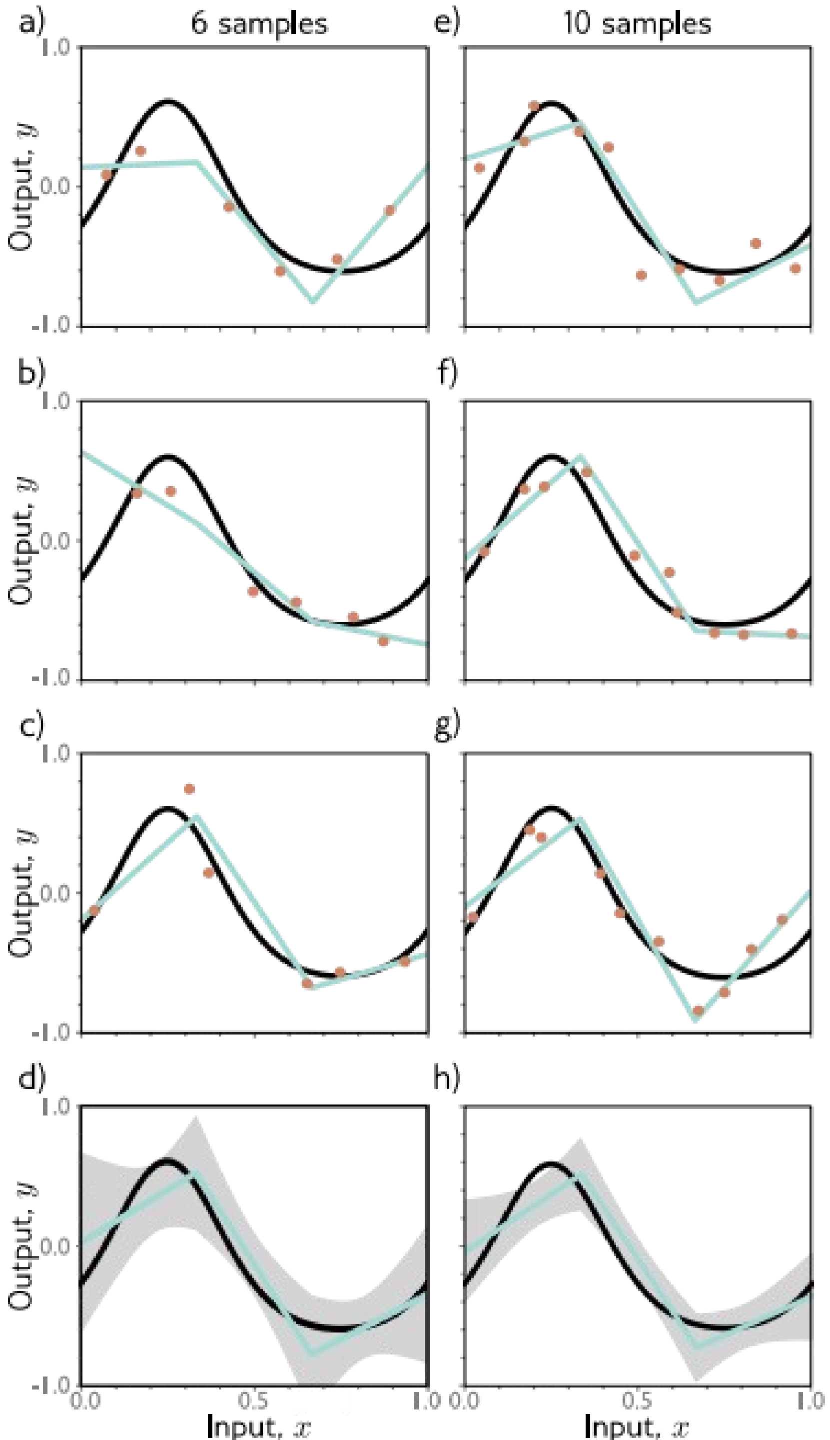
# Measuring performance

- MNIST1D dataset model and performance
- Noise, bias, and variance
- Reducing variance
- Reducing bias & bias-variance trade-off
- Double descent
- Curse of dimensionality & weird properties of high dimensional space
- Choosing hyperparameters

# Variance

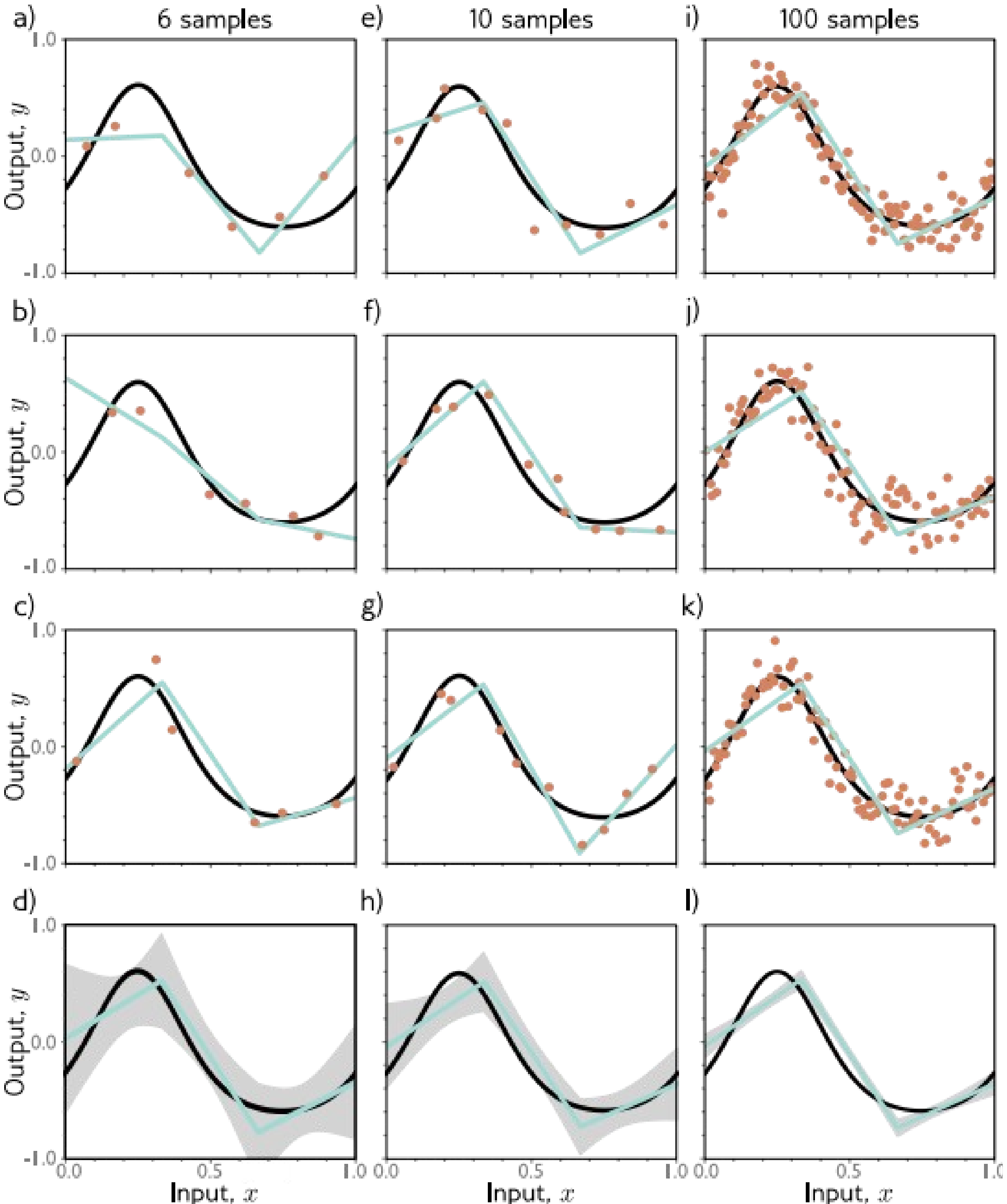


# Variance



a)
6 samples
e)
10 samples
b)
f)
c)
g)
d)
h)
Output, $y$
1.0
0.0
-1.0
Input, $x$
0.0
0.5
1.0

# Variance

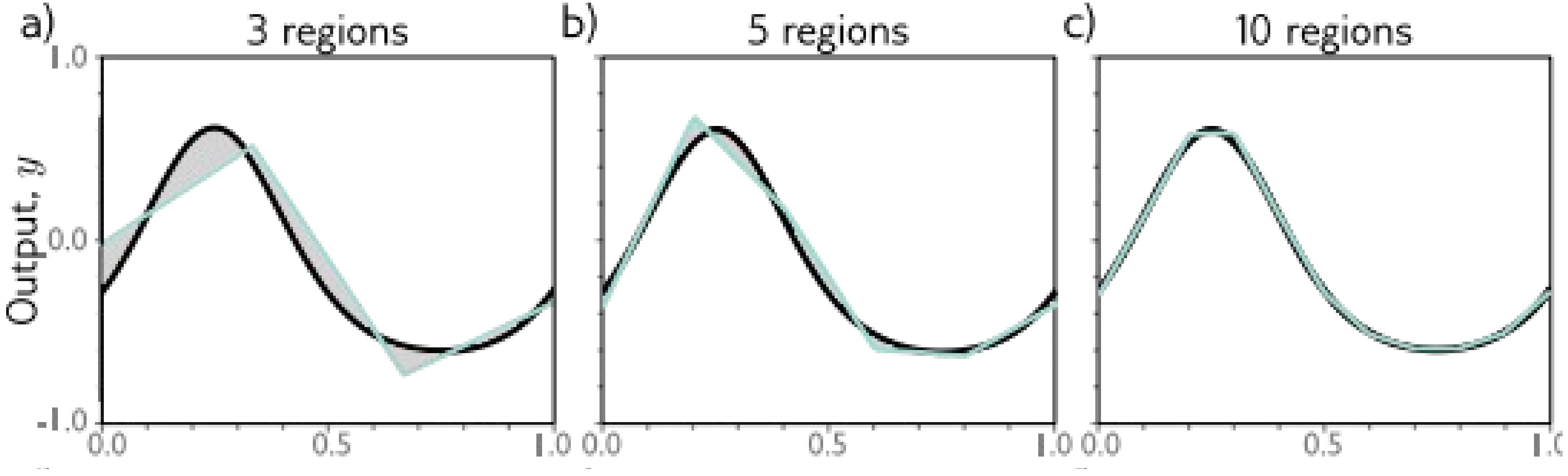



Can reduce variance by adding more samples

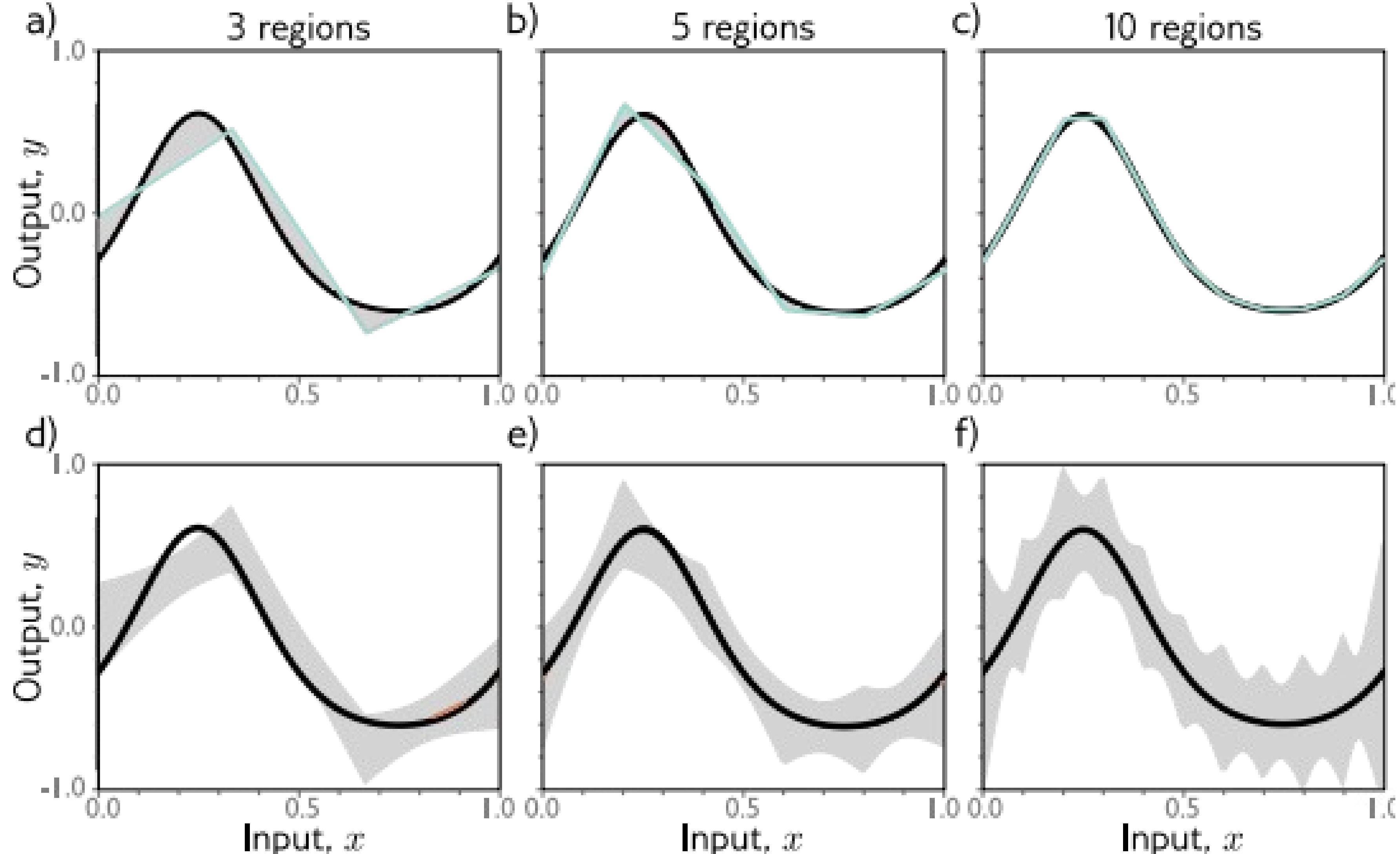
# Measuring performance

- MNIST1D dataset model and performance
- Noise, bias, and variance
- Reducing variance
- Reducing bias & bias-variance trade-off
- Double descent
- Curse of dimensionality & weird properties of high dimensional space
- Choosing hyperparameters

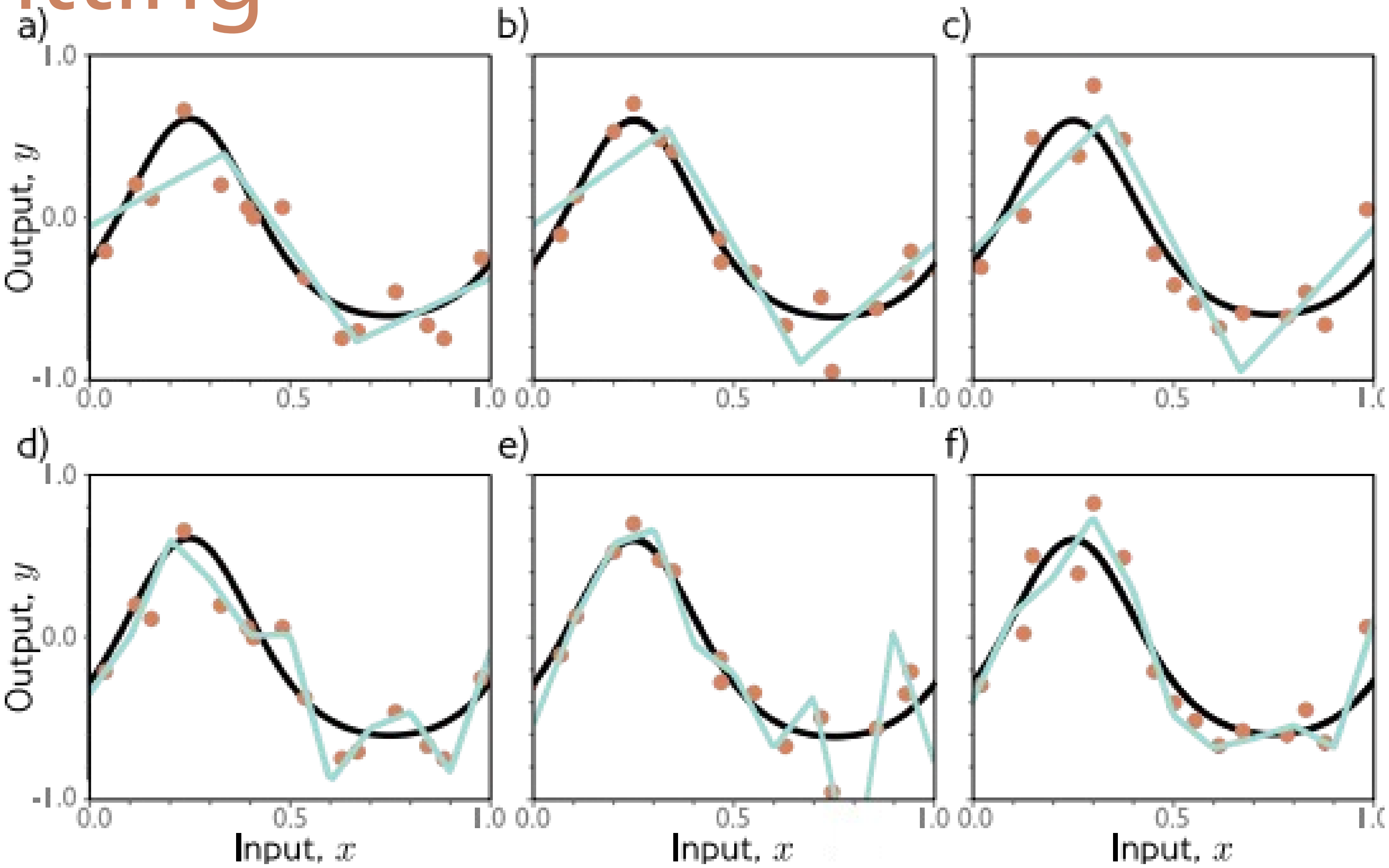
# Reducing bias



# Reducing bias

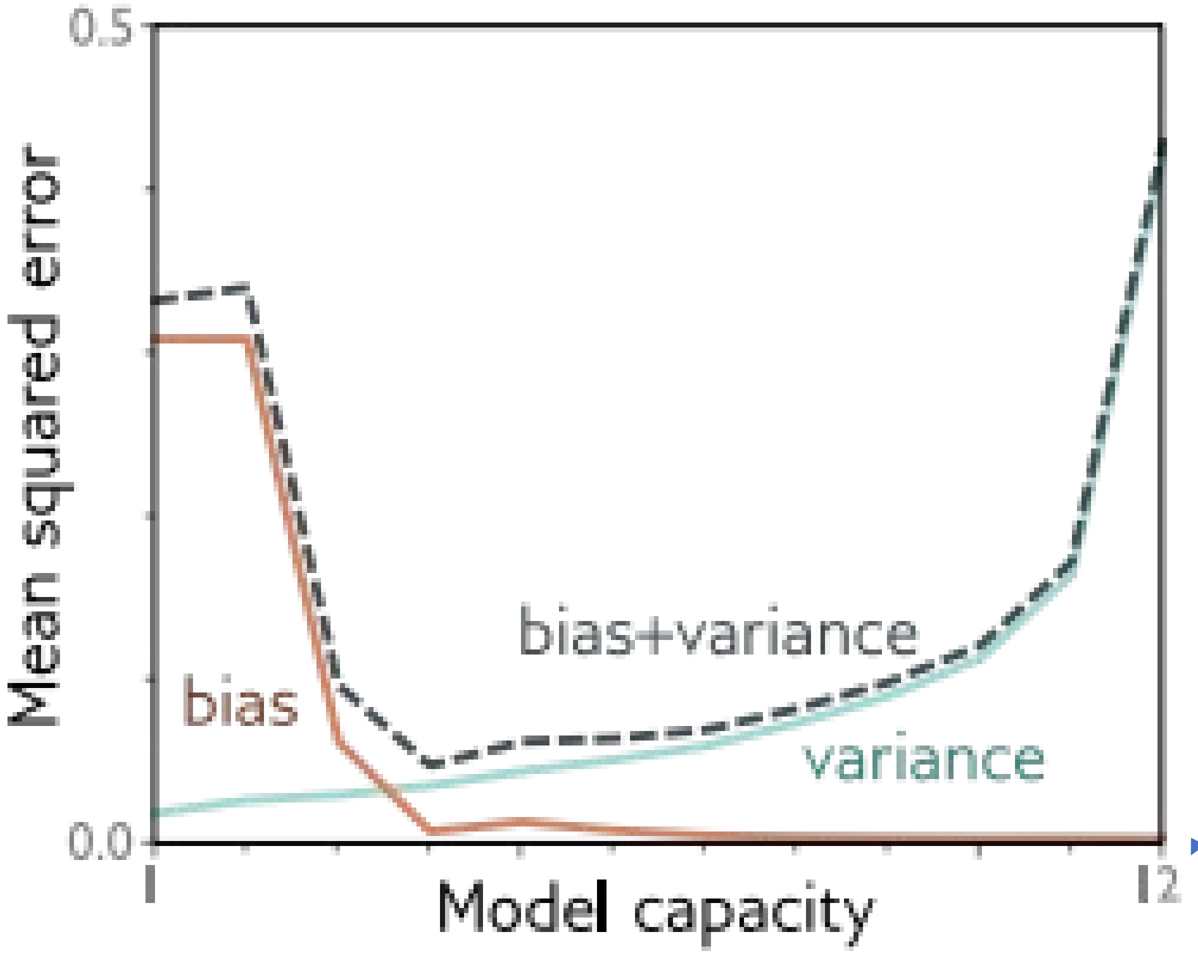


# Why does variance increase?
## Overfitting




Describes the training data better, but not the true underlying function (black curve)

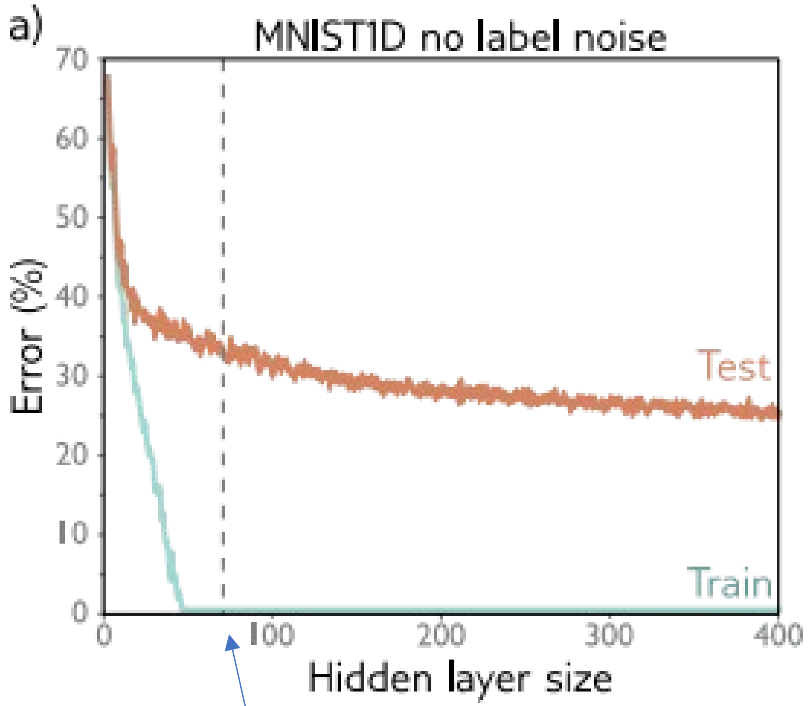
# Bias and variance trade-off



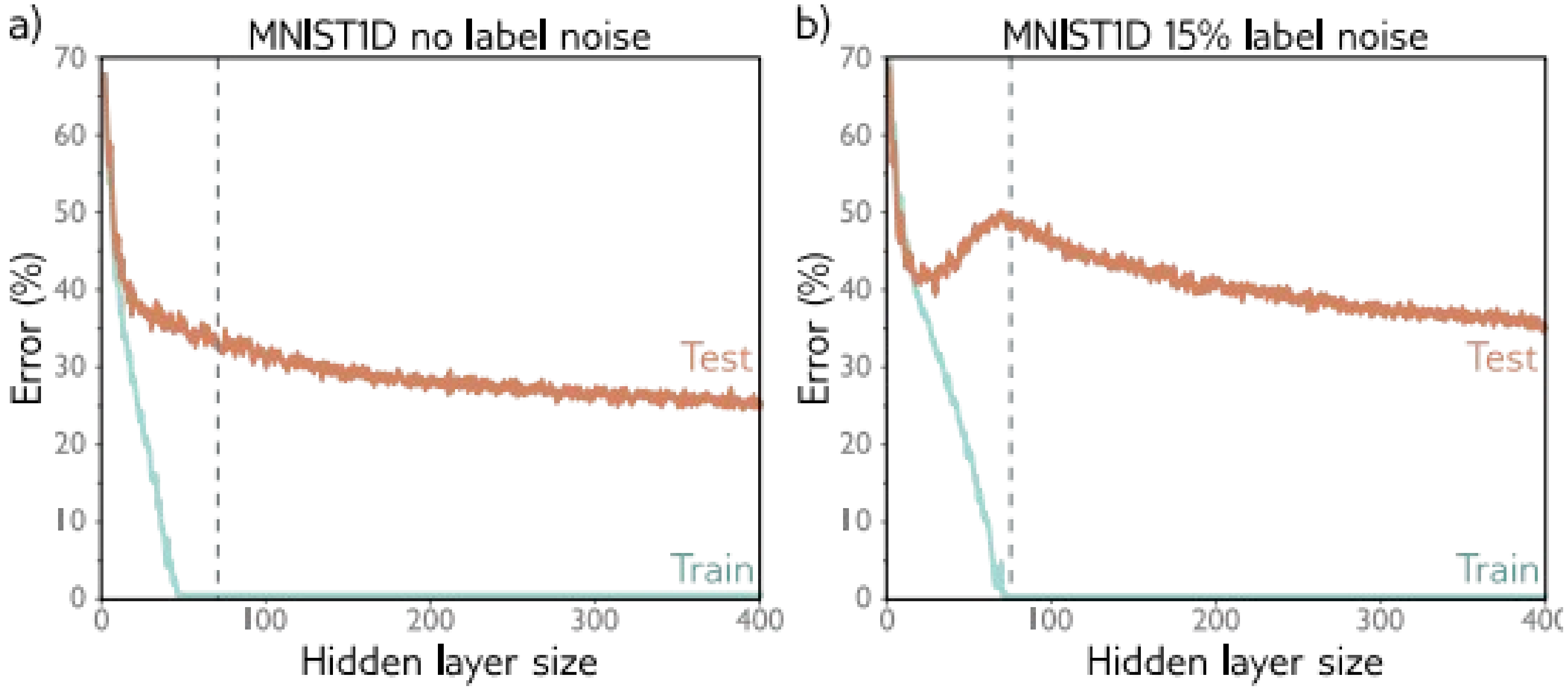

Number of datapoints

# Measuring performance

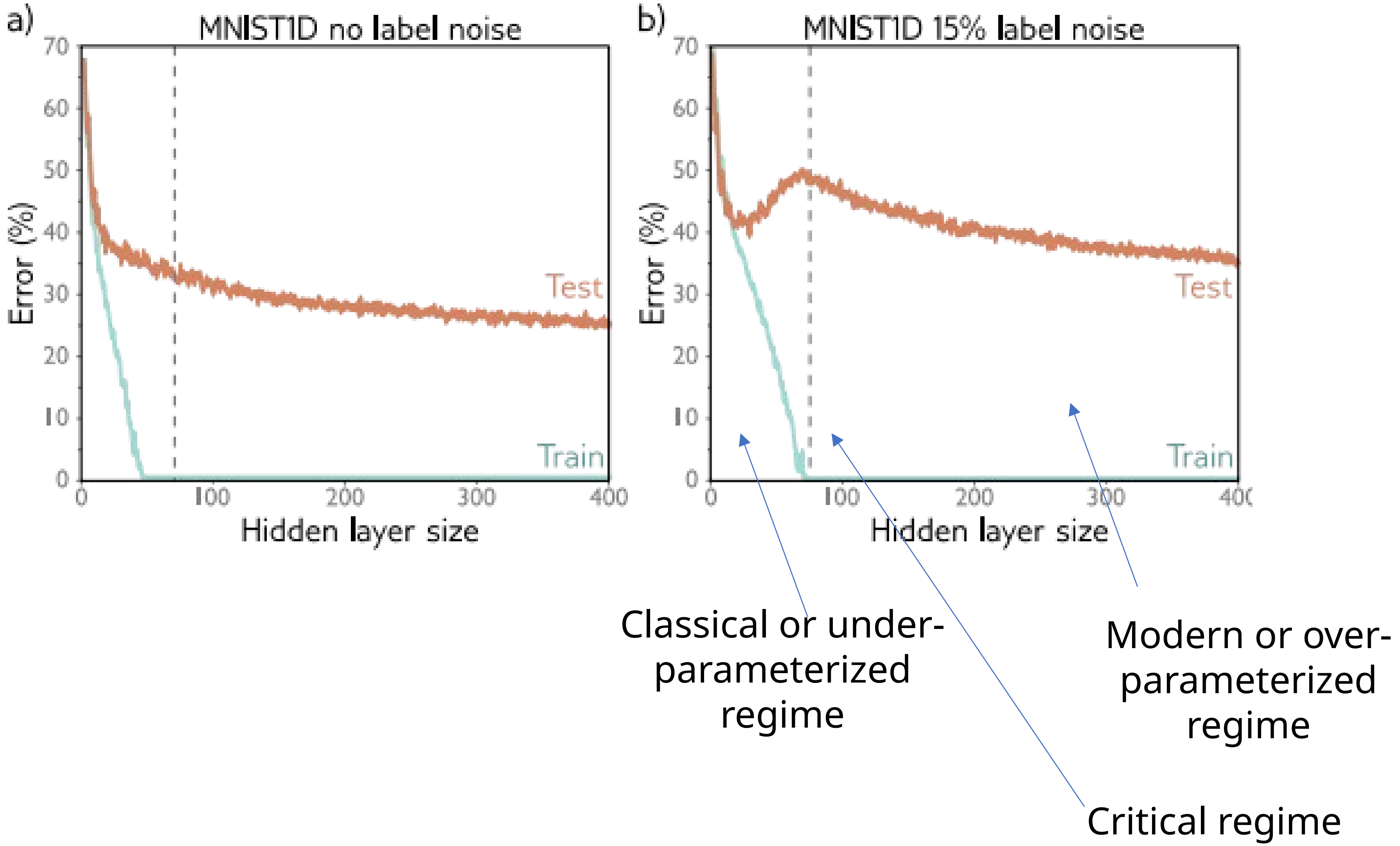
- MNIST1D dataset model and performance
- Noise, bias, and variance
- Reducing variance
- Reducing bias & bias-variance trade-off
- Double descent
- Curse of dimensionality & weird properties of high dimensional space
- Choosing hyperparameters



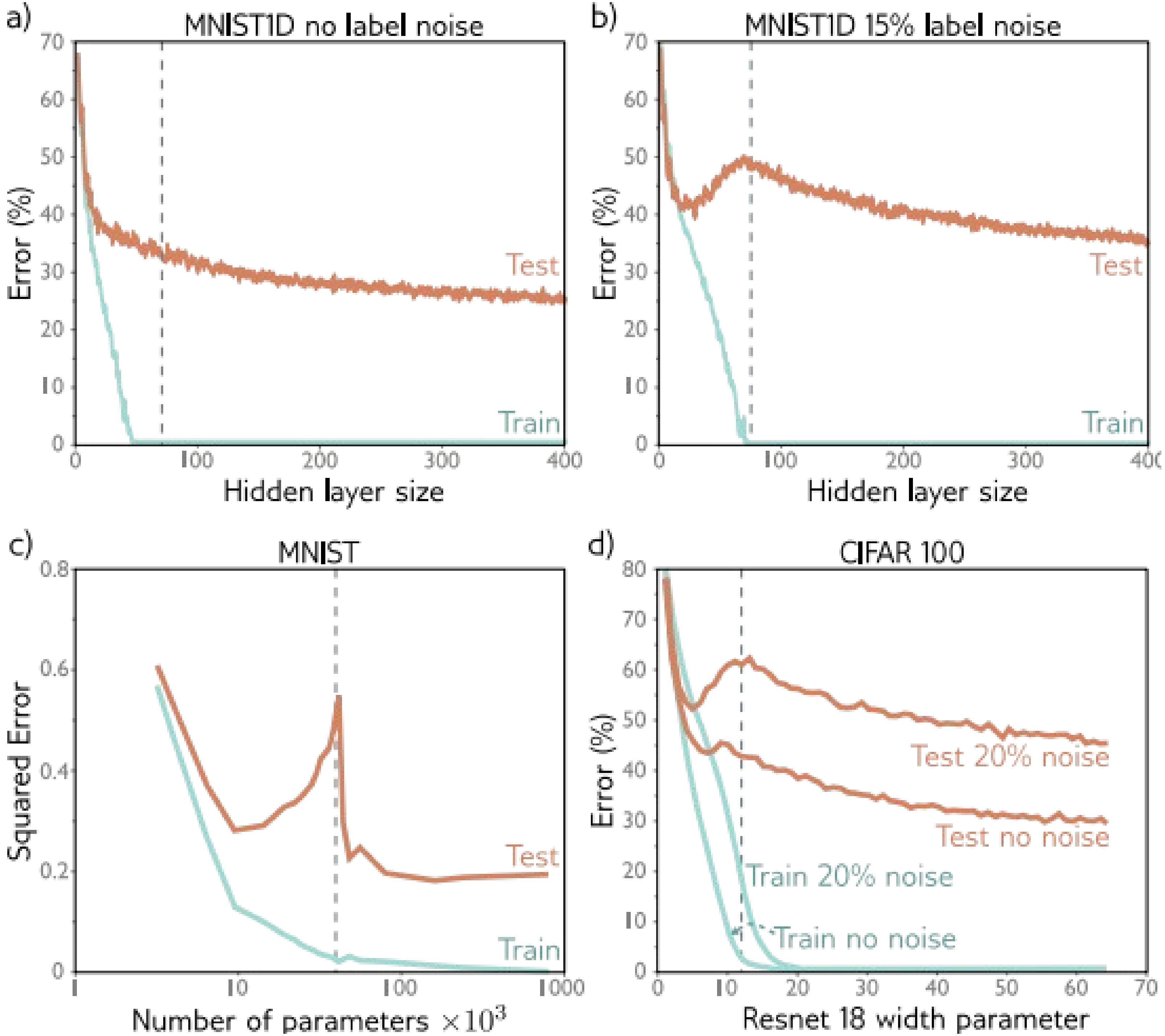

Number of
datapoints



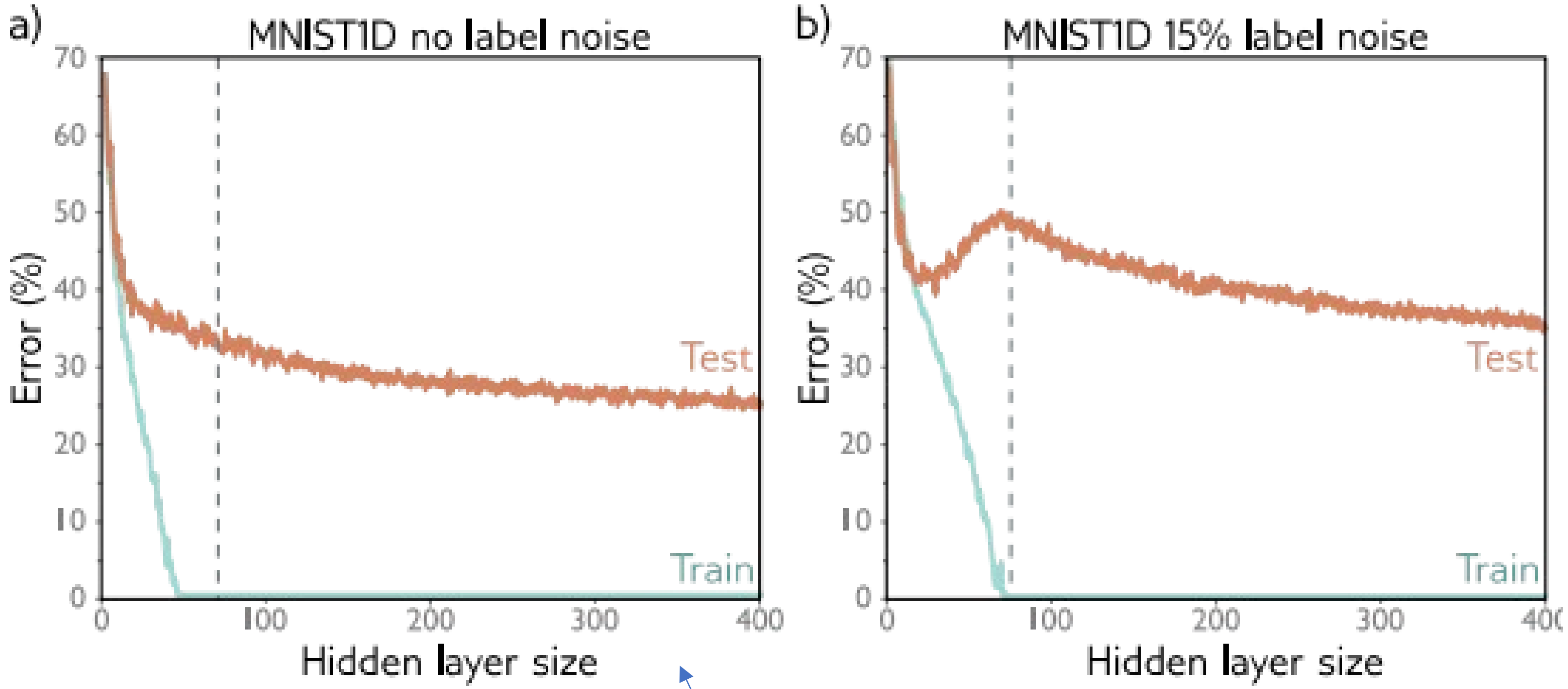

# Double descent



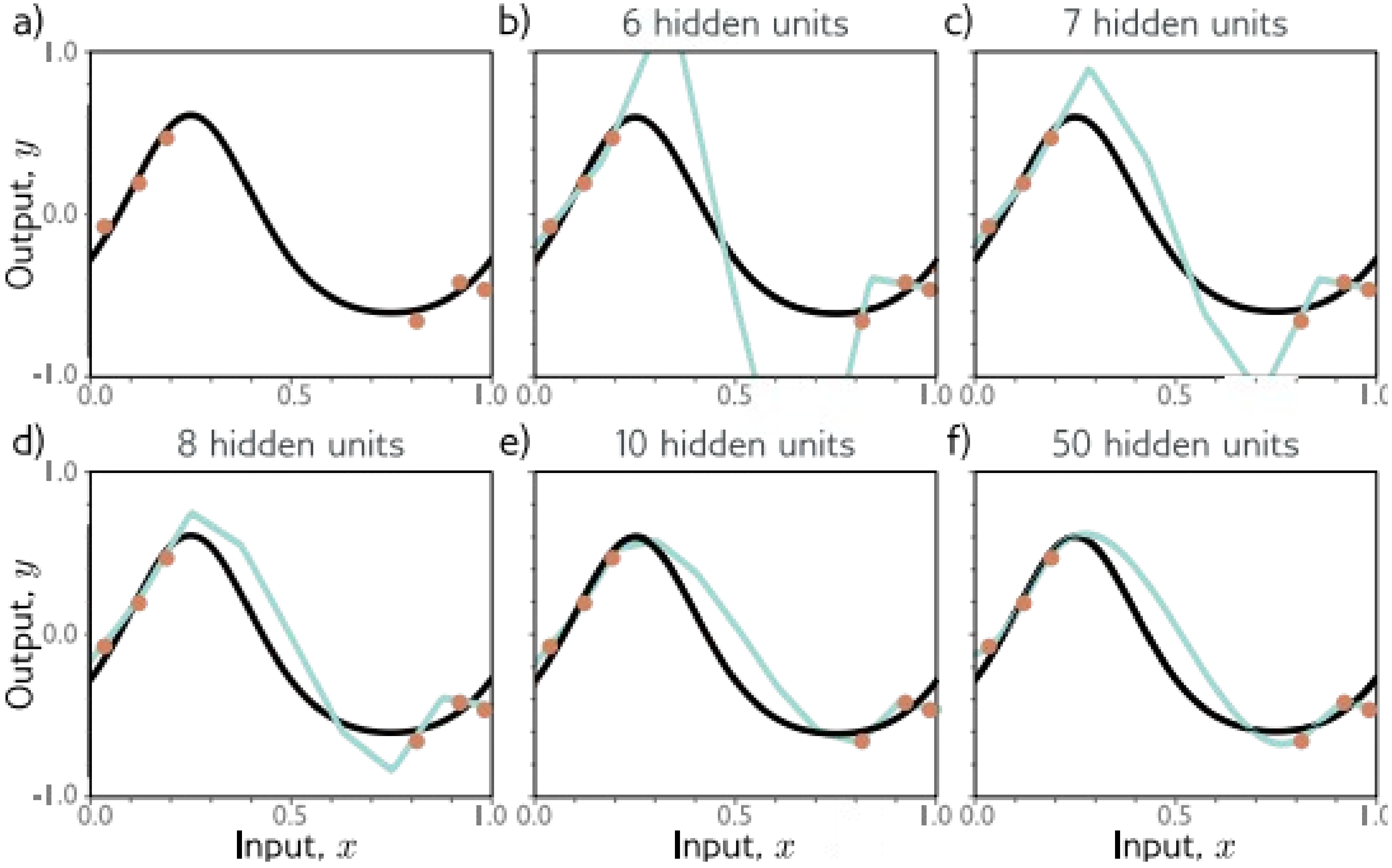
a)
MNIST1D no label noise
Error (%)
Hidden layer size
Test
Train
b)
MNIST1D 15% label noise
Error (%)
Hidden layer size
Test
Train
Classical or under-parameterized regime
Modern or over-parameterized regime
Critical regime



a)
MNIST1D no label noise
Error (%)
Hidden layer size
Test
Train
b)
MNIST1D 15% label noise
Error (%)
Hidden layer size
Test
Train
c)
MNIST
Squared Error
Number of parameters $\times 10^3$
Test
Train
d)
CIFAR 100
Error (%)
Resnet 18 width parameter
Test 20% noise
Test no noise
Train 20% noise
Train no noise



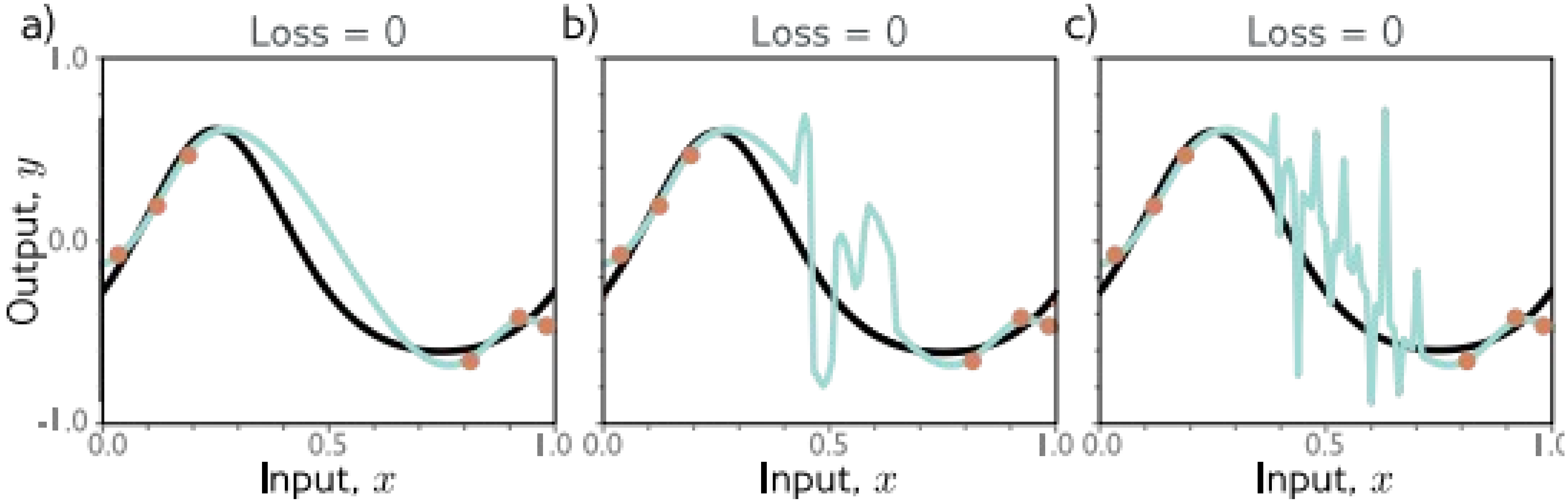

- Note that train data is very close to zero.
- Whatever is happening isn't happening at training data points
- Must be happening between the data points??




Potential explanation:

- can make smoother functions with more hidden units
- being smooth between the datapoints is a reasonable thing to do

But why?




- All of these solutions are equivalent in terms of loss.
- Why should the model choose the smooth solution?
- Tendency of model to choose one solution over another is inductive bias

# Measuring performance

- MNIST1D dataset model and performance
- Noise, bias, and variance
- Reducing variance
- Reducing bias & bias-variance trade-off
- Double descent
- Curse of dimensionality & weird properties of high dimensional space
- Choosing hyperparameters

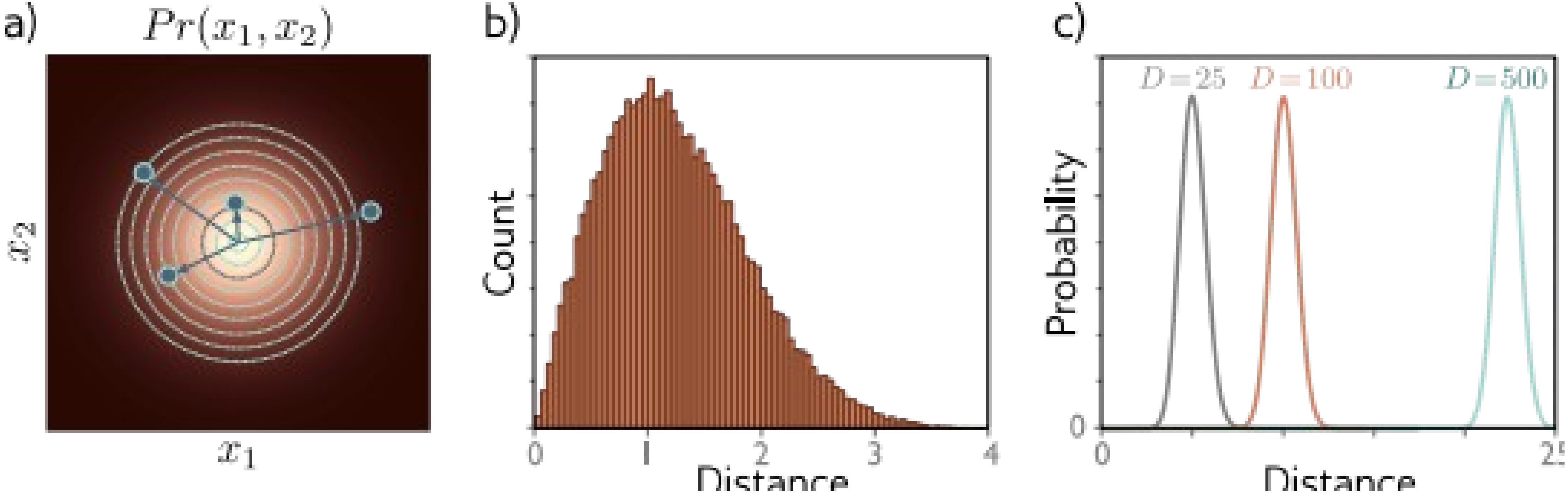
# Curse of dimensionality

- 40-dimensional data
- 10,000 data points
- Consider quantizing each dimension into 10 bins
-  bins
- 1 data point per  bins
- The tendency of high-dimensional space to overwhelm the number of data points is called the curse of dimensionality

# Weird properties of high-dimensional space

- Two randomly sampled data points from normal are at right angles to each other with high likelihood
- Distance from the origin of random samples is roughly constant
- Most of the volume of a high dimensional orange is in the peel not in the pulp
- Volume of a diameter one hypersphere becomes zero
- Generate random points uniformly in hypercube, ratio of nearest to farthest becomes close to one.

# Distance from the origin of random samples is roughly constant



# Weird properties of high-dimensional space

- Two randomly sampled data points from normal are at right angles to each other with high likelihood
- Distance from the origin of random samples is roughly constant

# Measuring performance

- MNIST1D dataset model and performance
- Noise, bias, and variance
- Reducing variance
- Reducing bias & bias-variance trade-off
- Double descent
- Curse of dimensionality & weird properties of high dimensional space
- Choosing hyperparameters

# Choosing hyperparameters

- Don’t know bias or variance
- Don’t know how much capacity to add
- How do we choose capacity in practice?
  - Or model structure
  - Or training algorithm
  - Or learning rate
- Third data set – validation set
  - Train models with different hyperparameters on training set
  - Choose best hyperparameters with validation set
  - Test once with test set