GENERAL ASSEMBLY

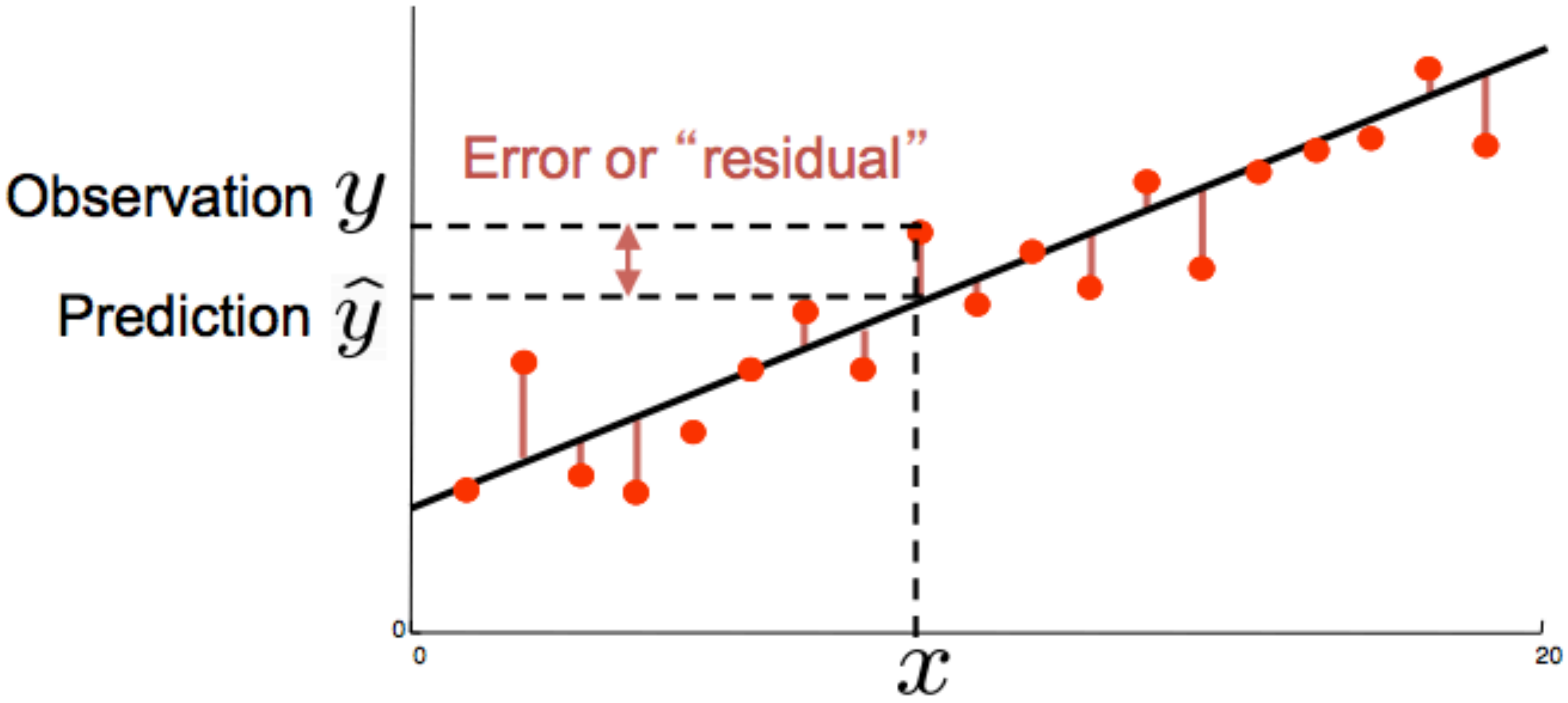
BIAS-VARIANCE TRADEOFF

Joseph Nelson, Data Science Immersive

AGENDA

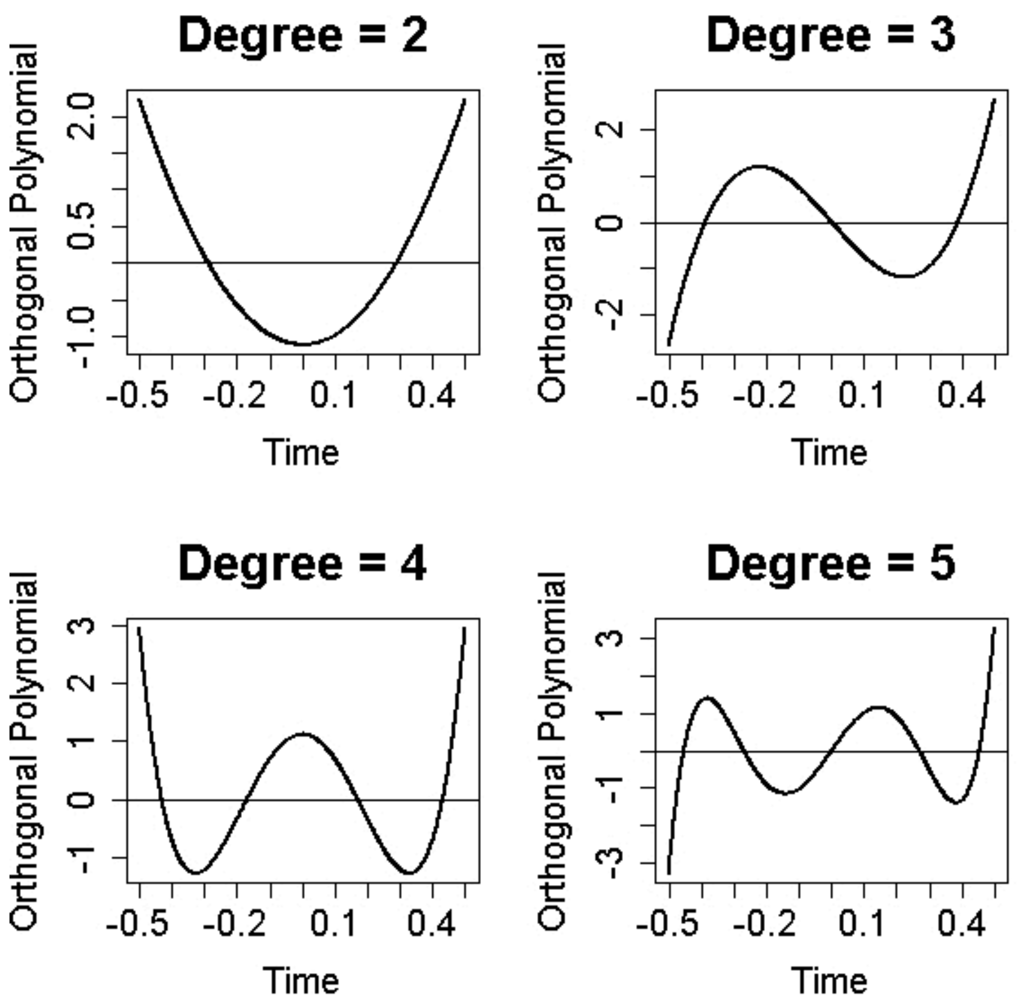
- Linear & Polynomial Regression Quick Review
- Introduction to Bias, Variance
- Regression Interactive Complexity
- Fitting Sine Exercise
- Balancing Bias and Variance
- (Bit of code)

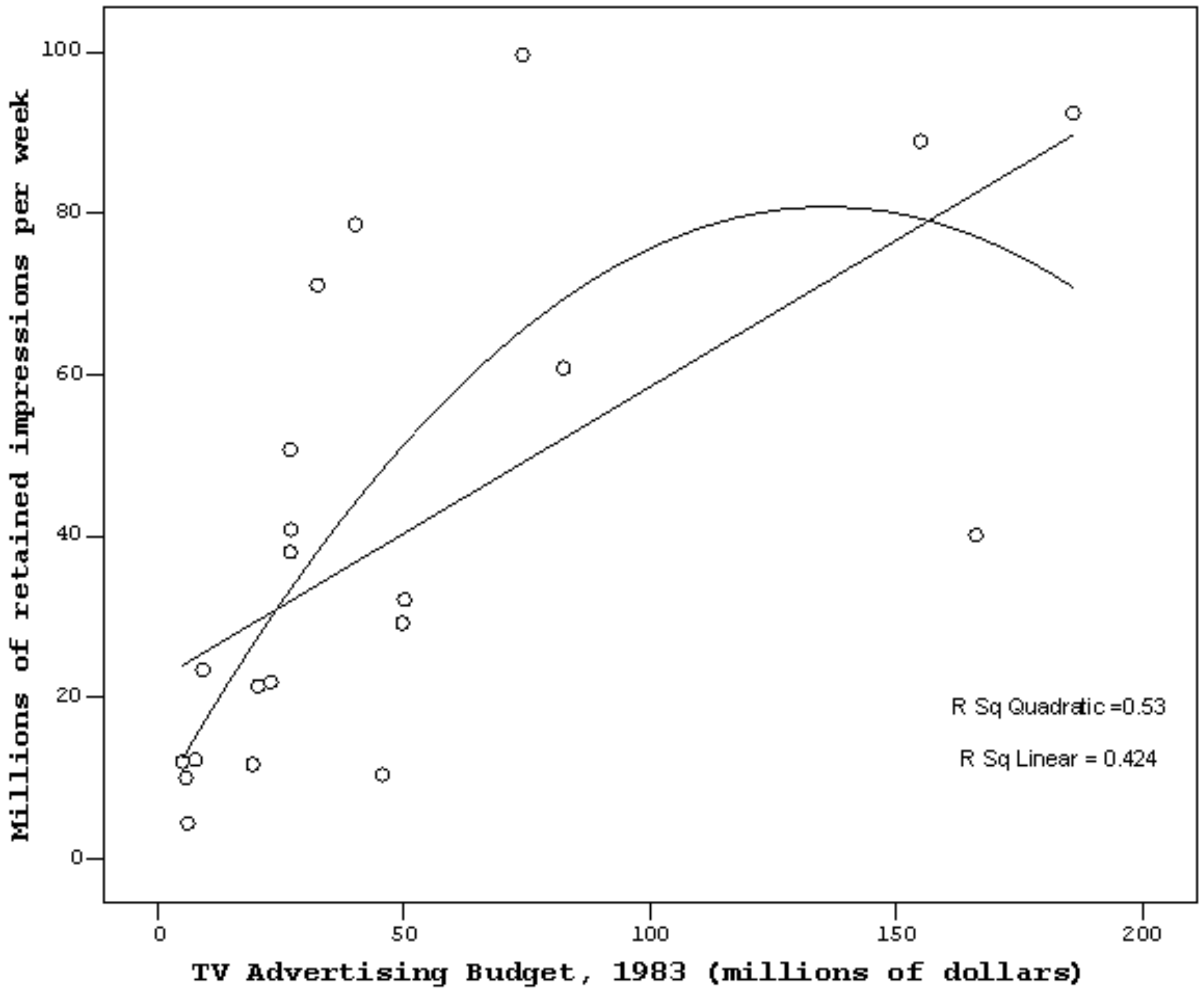
LINEAR & POLYNOMIAL REGRESSION QUICK REVIEW



LINEAR & POLYNOMIAL REGRESSION QUICK REVIEW

- Not all relationships are linear





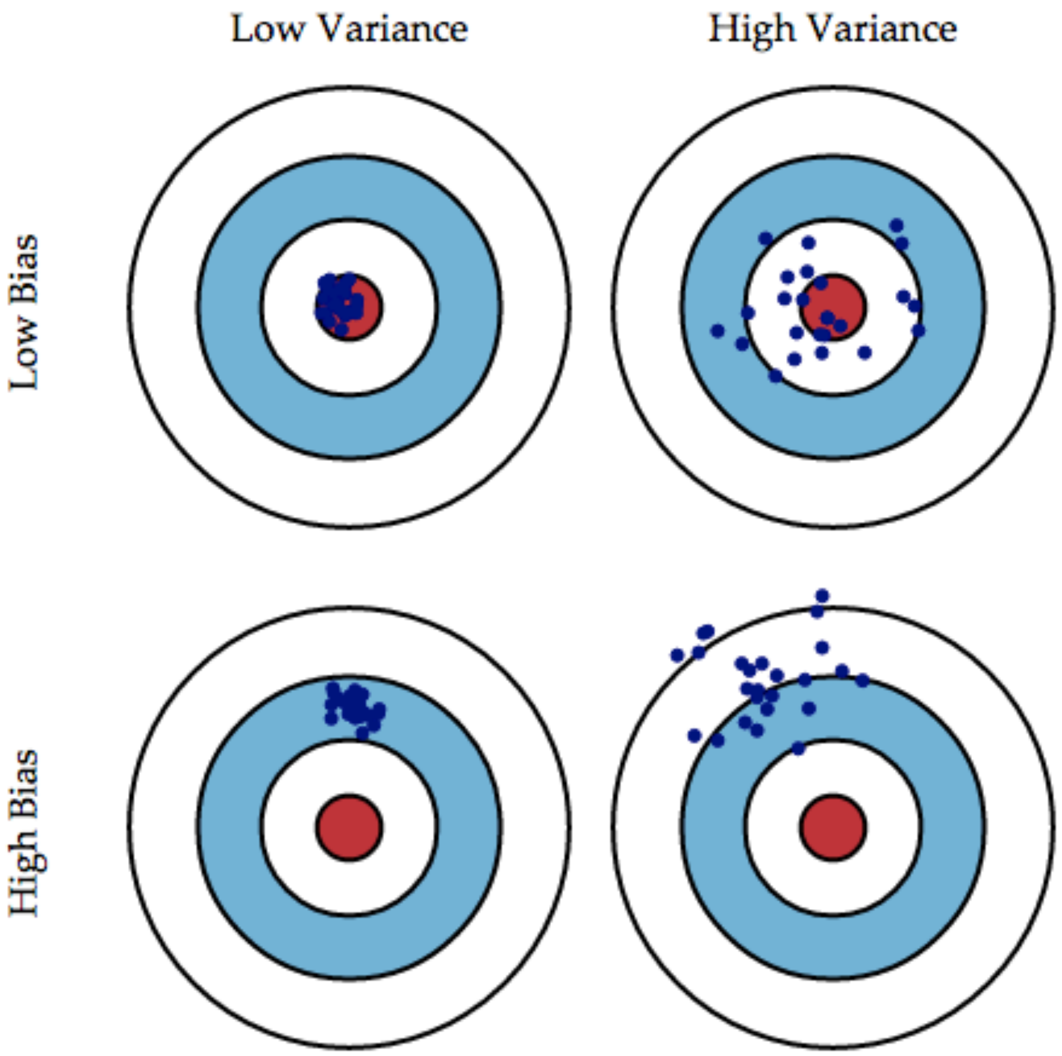
BIAS? VARIANCE?

- Conceptual Definitions
- Bias – Error that results from the correct value and the predicted value within our model
- Variance – Error due to the variability of a model prediction for a given data point

- LOOSELY:
- BIAS: Points within a model
- VARIANCE: A point between many models

BIAS? VARIANCE?

- Visually, we are building a model where the bulls-eye is the goal
- Each individual hit is one prediction based on our model
- Critically, the success of our model (low variance, low bias) depends on the training data present



BIAS? VARIANCE?

- Mathematically, we are explaining a linear relationship dependent on some function
- The error of our prediction is equal to the true outcome and our predicted outcome
- This can be decomposed into two parts: the bias and the variance:

$$Y = f(X) + \epsilon$$

$$Err(x) = E\left[(Y - \hat{f}(x))^2\right]$$

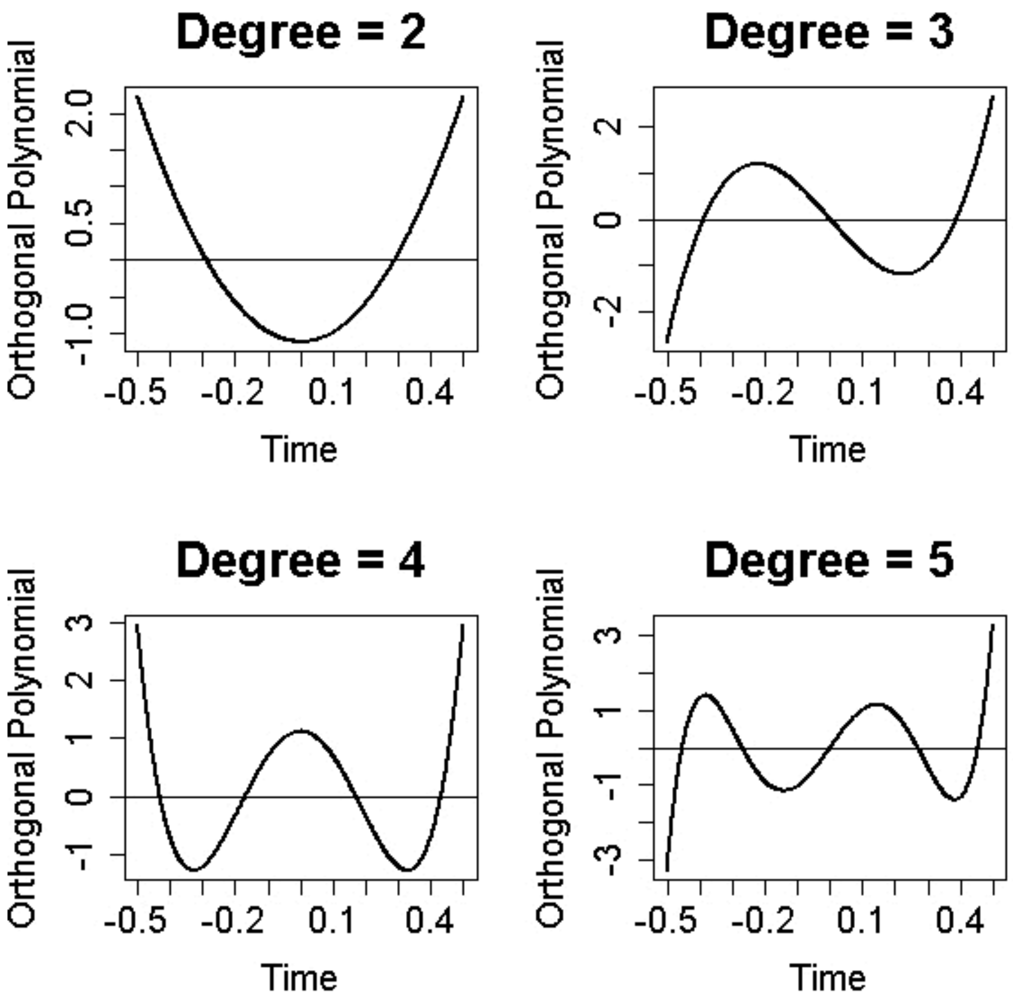
$$Err(x) = \left(E[\hat{f}(x)] - f(x)\right)^2 + E\left[\left(\hat{f}(x) - E[\hat{f}(x)]\right)^2\right] + \sigma_e^2$$

- *Err(x)* = Bias^2 + Variance + Irreducible Error

IMAGINE…

- Pretend you're predicting the outcome of an election. You random digit dial 100 individuals from a phonebook and find that of those polled, 45 are supporting Trump and 55 are supporting Clinton.
- After the election, 40% supported Clinton and 60% supported Trump.
- What happened? Is the error you suggested an example of bias in our model our variance?

REGRESSION INTERACTIVE COMPLEXITY

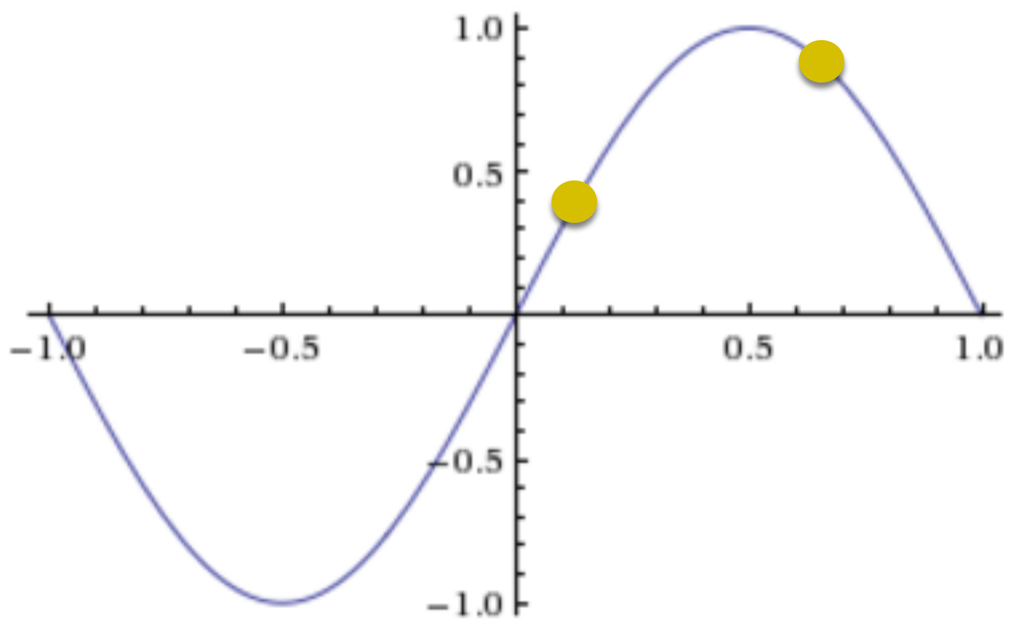


- The higher the degree of your model, the more complex ("responsive") it is.
- Play along! (Does not work in Chrome): http://mste.illinois.edu/exner/java.f/leastsquares/
- OR http://arachnoid.com/polysolve/

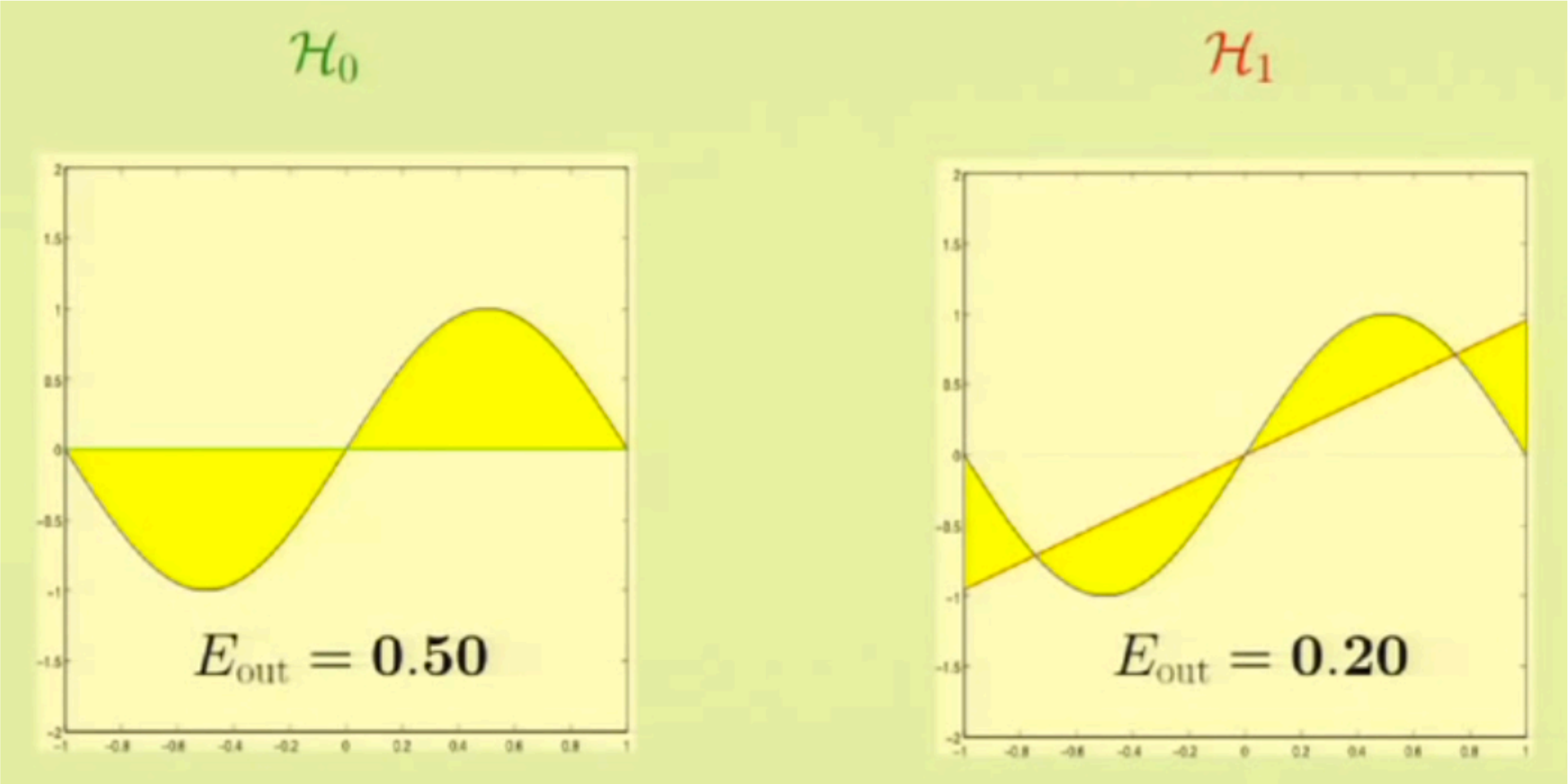
FITTING SINE EXERCISE

- Please attempt to fit one of the following two models against the sine curve.
- H(0): f(x) = b
- H(1): f(x) = a(x) + b

Plot:



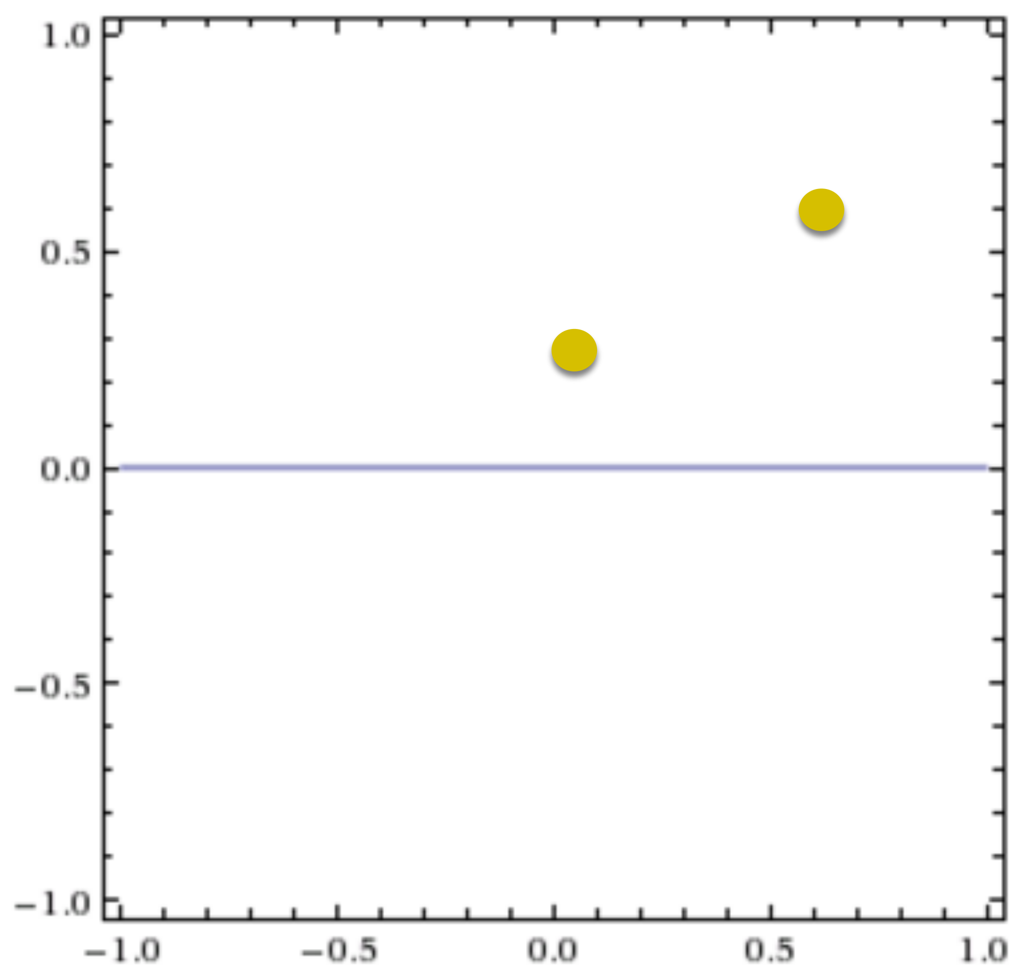
FITTING SINE EXERCISE: H0 VS H1



FITTING SINE EXERCISE

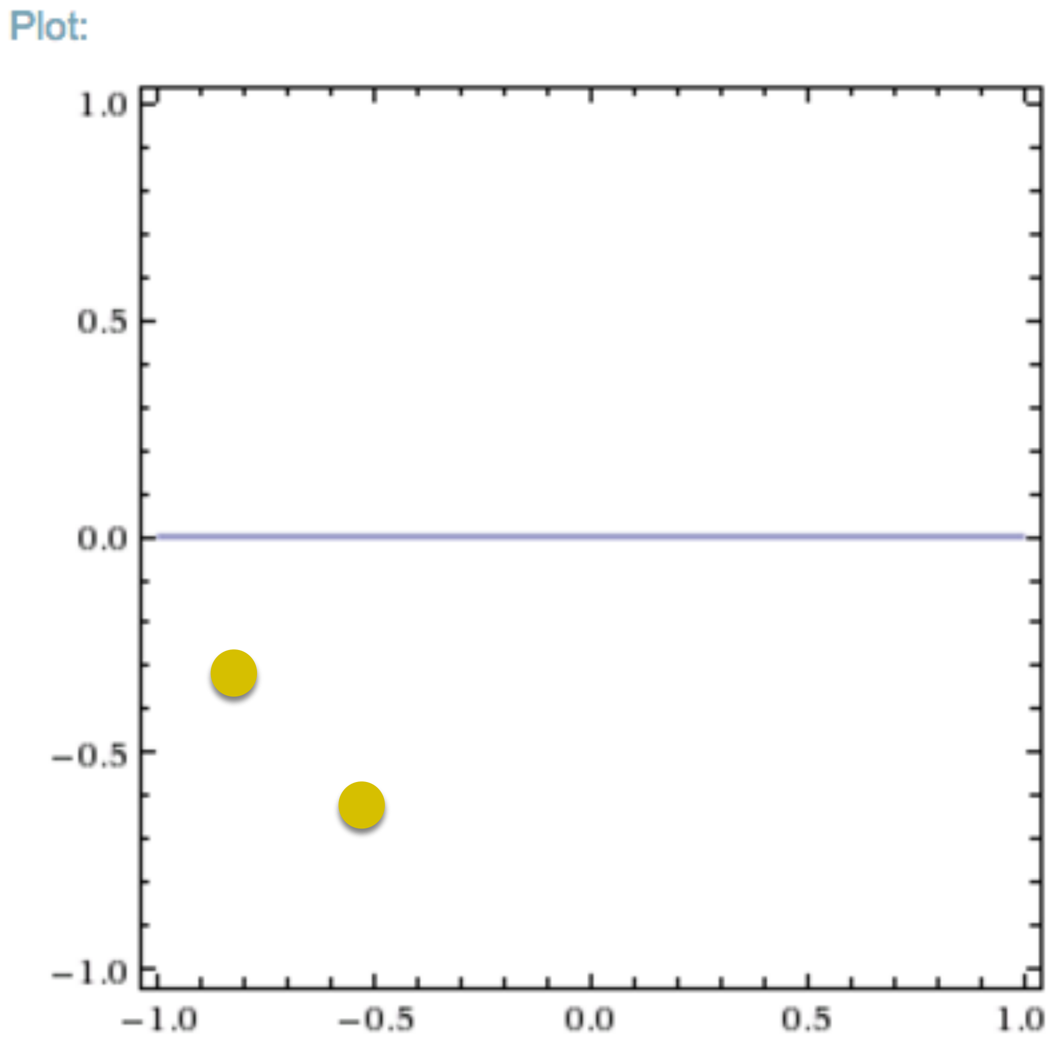
- Using the same two functions, imagine you do not have the sine curve target function. You only have the TWO POINTS given on the function.
- H(0): f(x) = b
- H(1): f(x) = a(x) + b

Plot:



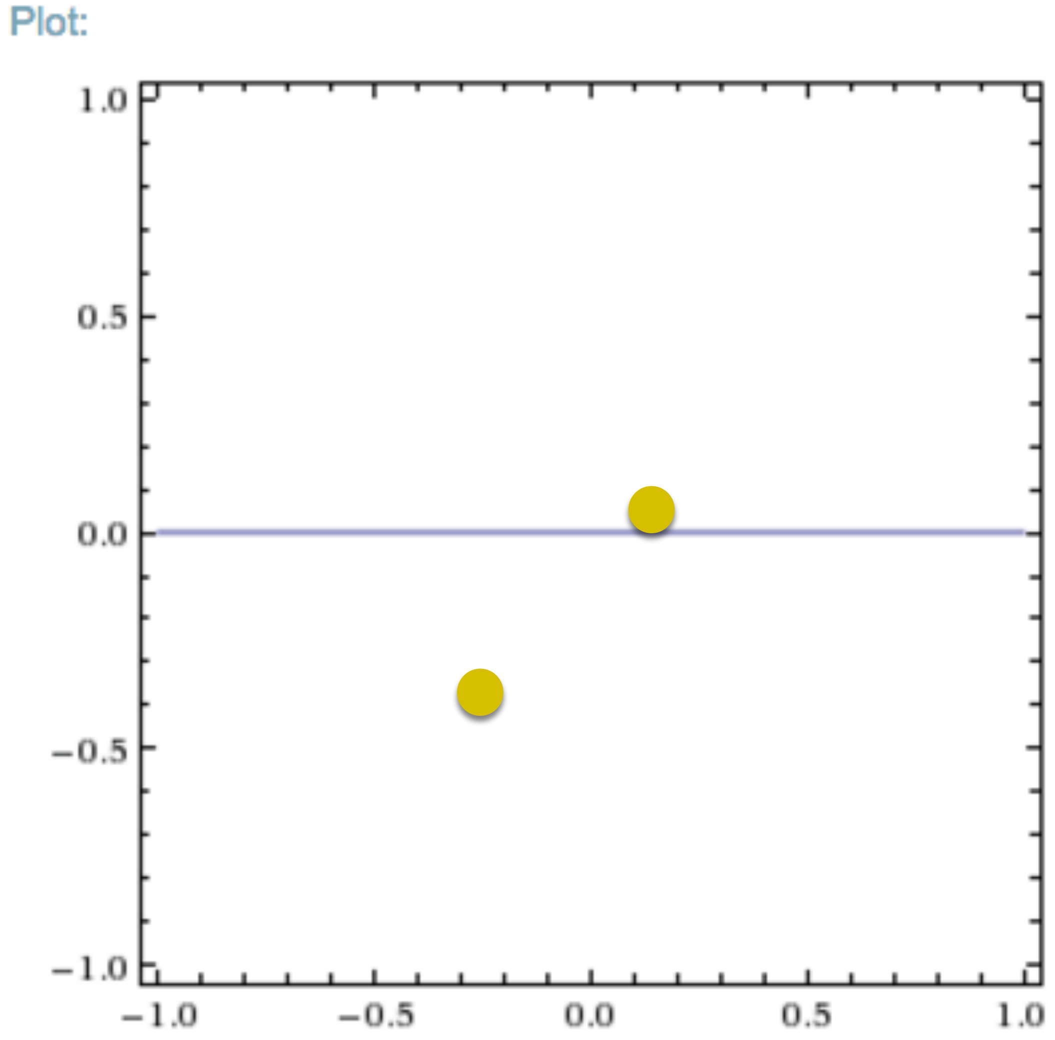
FITTING SINE EXERCISE

- Using the same two functions, imagine you do not have the sine curve target function. You only have the TWO POINTS given on the function. **Repeat your analysis with two different points.**
- H(0): f(x) = b
- H(1): f(x) = a(x) + b



FITTING SINE EXERCISE. CHALLENGE: REDUCE ERROR

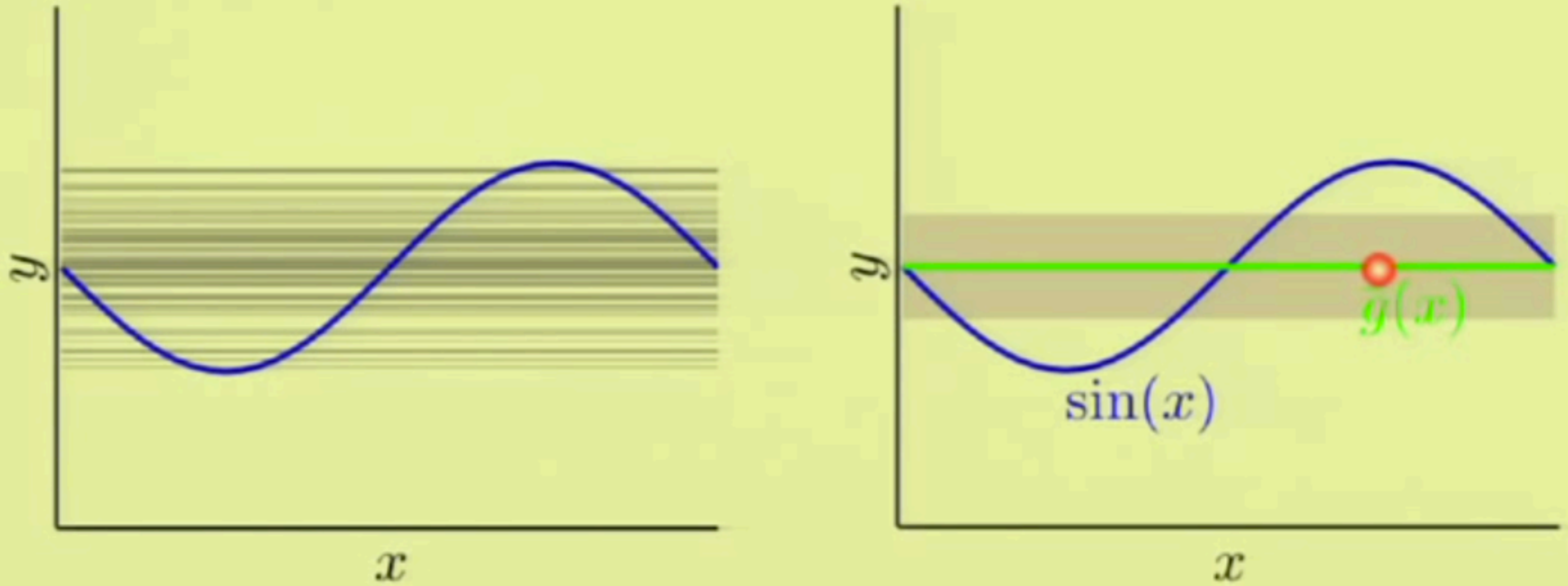
- Using the same two functions, imagine you do not have the sine curve target function. You only have the TWO POINTS given on the function. Repeat your analysis with two different points. **Attempt to plot every possible two point model.**
- H(0): f(x) = b
- H(1): f(x) = a(x) + b



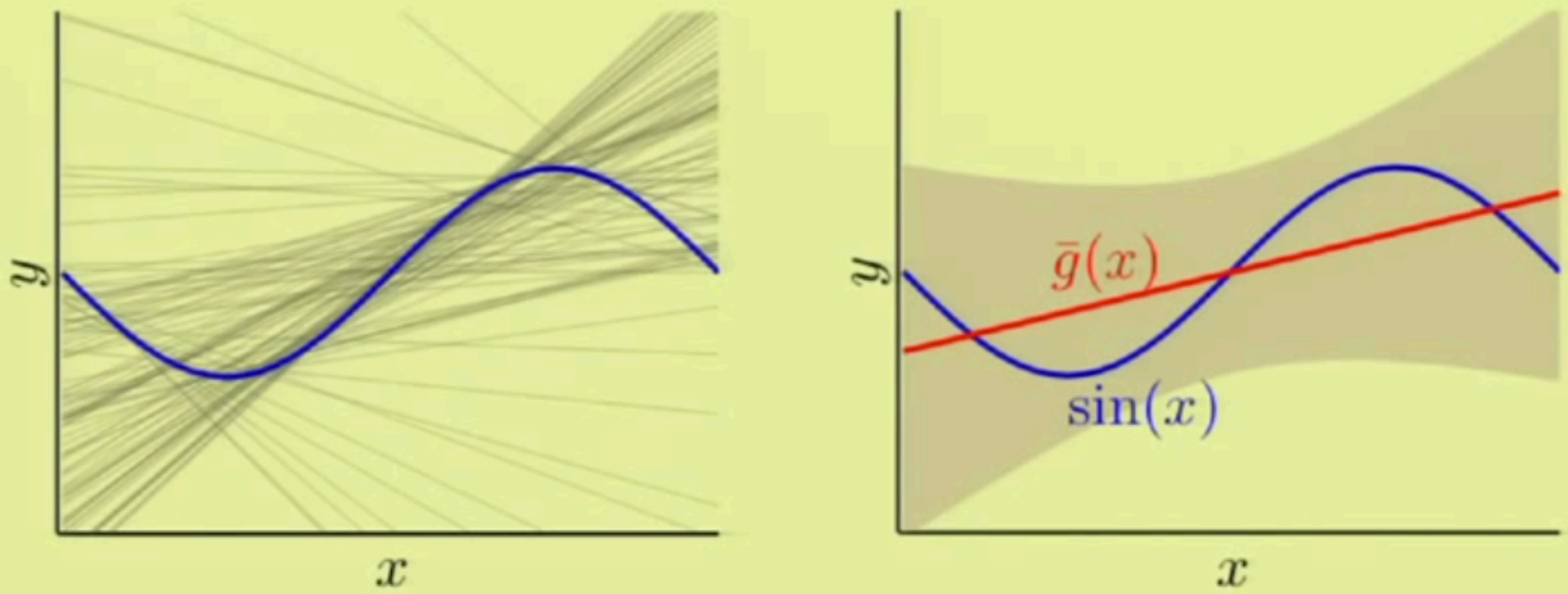
FITTING SINE EXERCISE. CHALLENGE: REDUCE ERROR

- **Which group won? Why?**

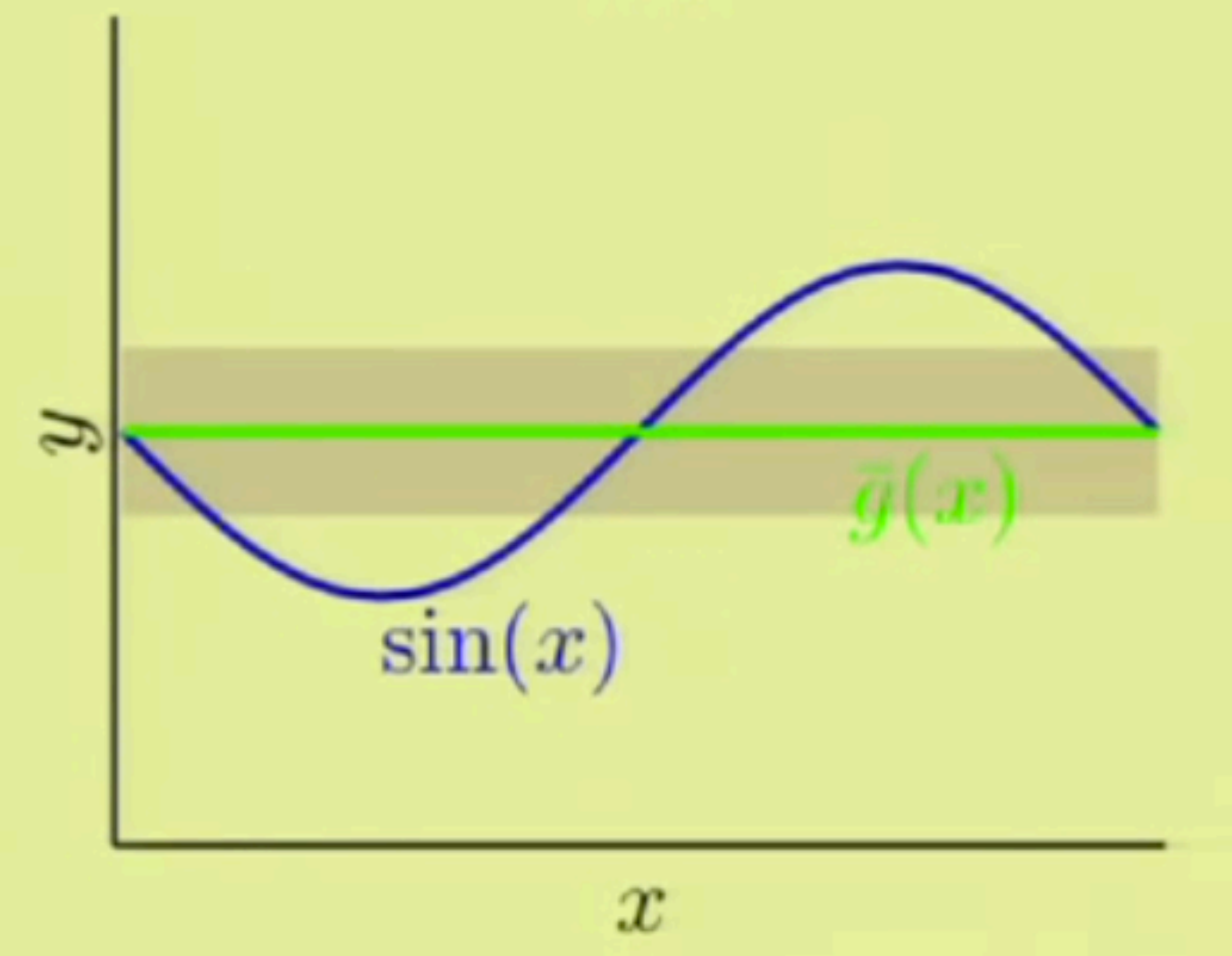
FITTING SINE EXERCISE: H0

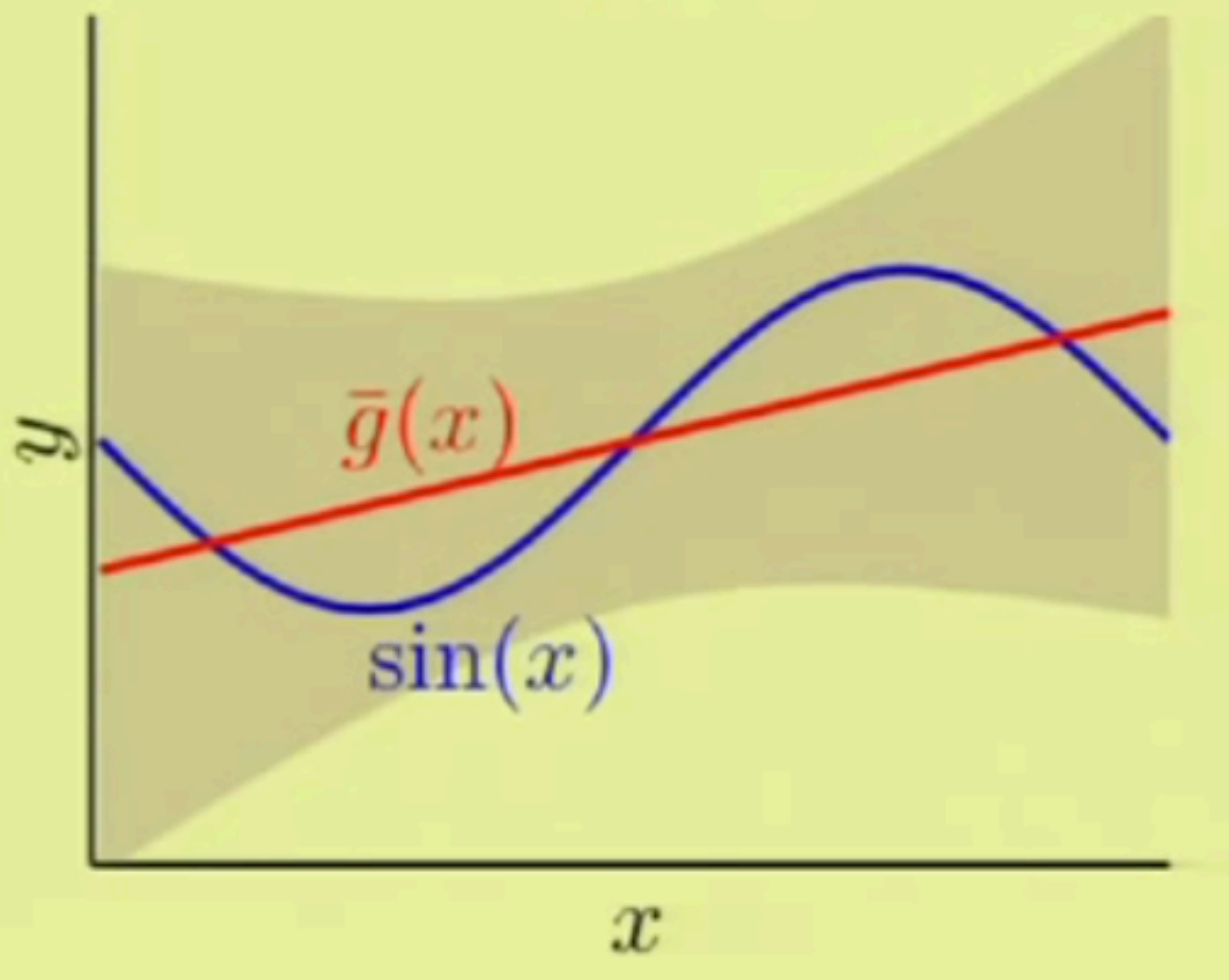


FITTING SINE EXERCISE: H1



FITTING SINE EXERCISE: H1





- Bias: 0.50
- Variance: 0.25

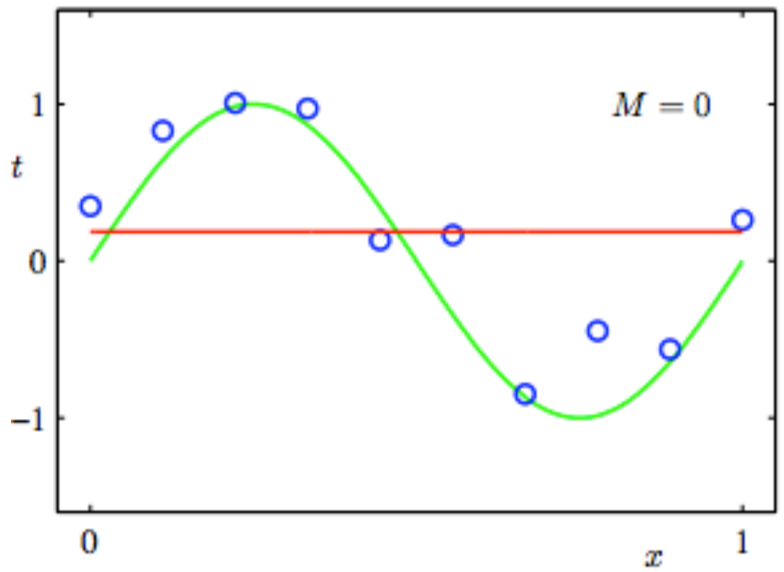
- Bias: 0.21
- Variance: 1.69

FITTING SINE EXERCISE. CHALLENGE: REDUCE ERROR

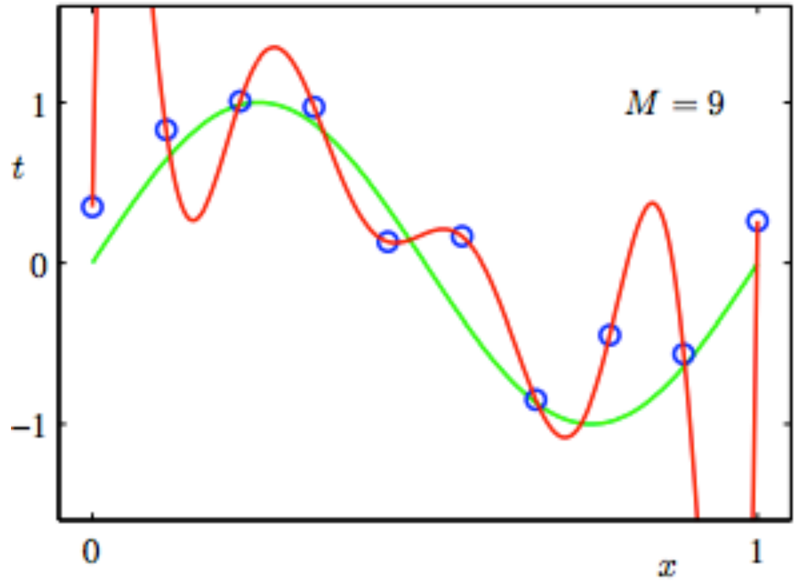
- **https://www.youtube.com/watch?v=7AZ3kYNftEs&feature=youtu.be&t=6m35s**

FITTING SINE EXERCISE. INTUITION

- Model too simple: does not fit the data well. A **BIASED** solution.



- Model too complex: small changes to the data, model changes a lot. A **HIGH-VARIANCE** solution.

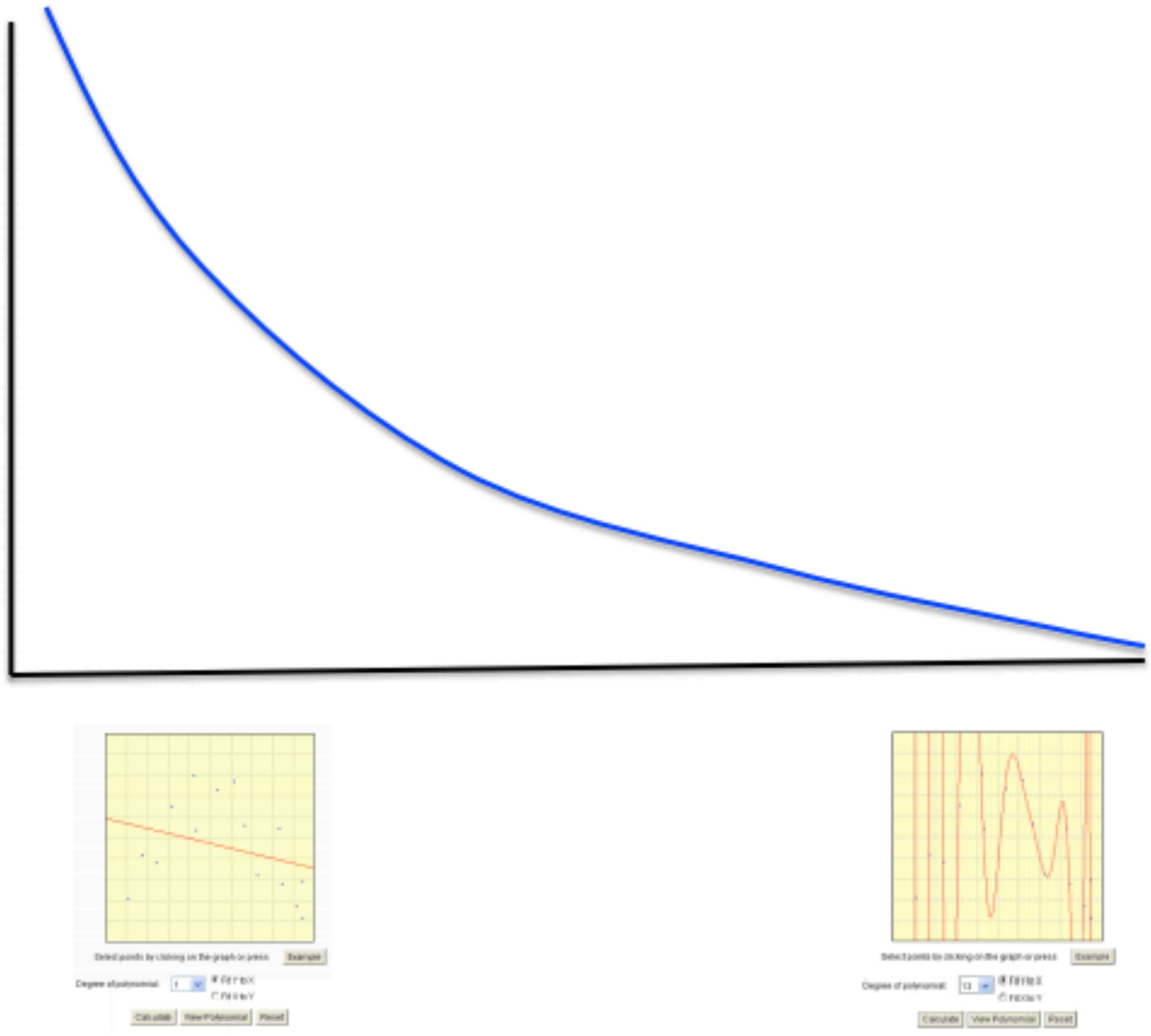


BALANCING BIAS AND VARIANCE

- We want a model that best balances bias and variance. It should match our training data well (moderate bias) yet be low-variance for out-of-sample data (moderate variance).

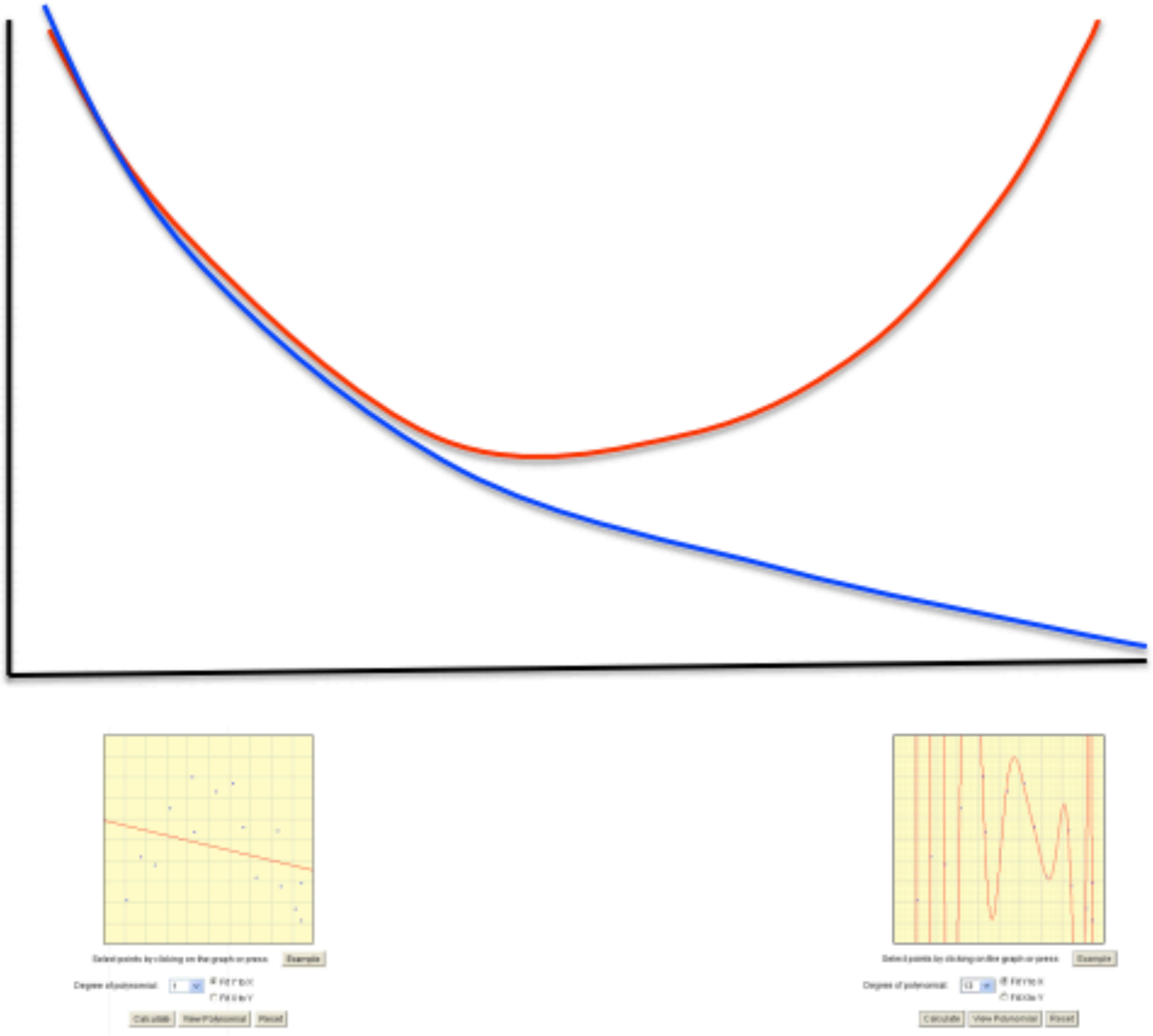
BALANCING BIAS

- Training error as a function of complexity.
- Question: why do we even care about variance if we know we can generate a more accurate model with higher complexity?



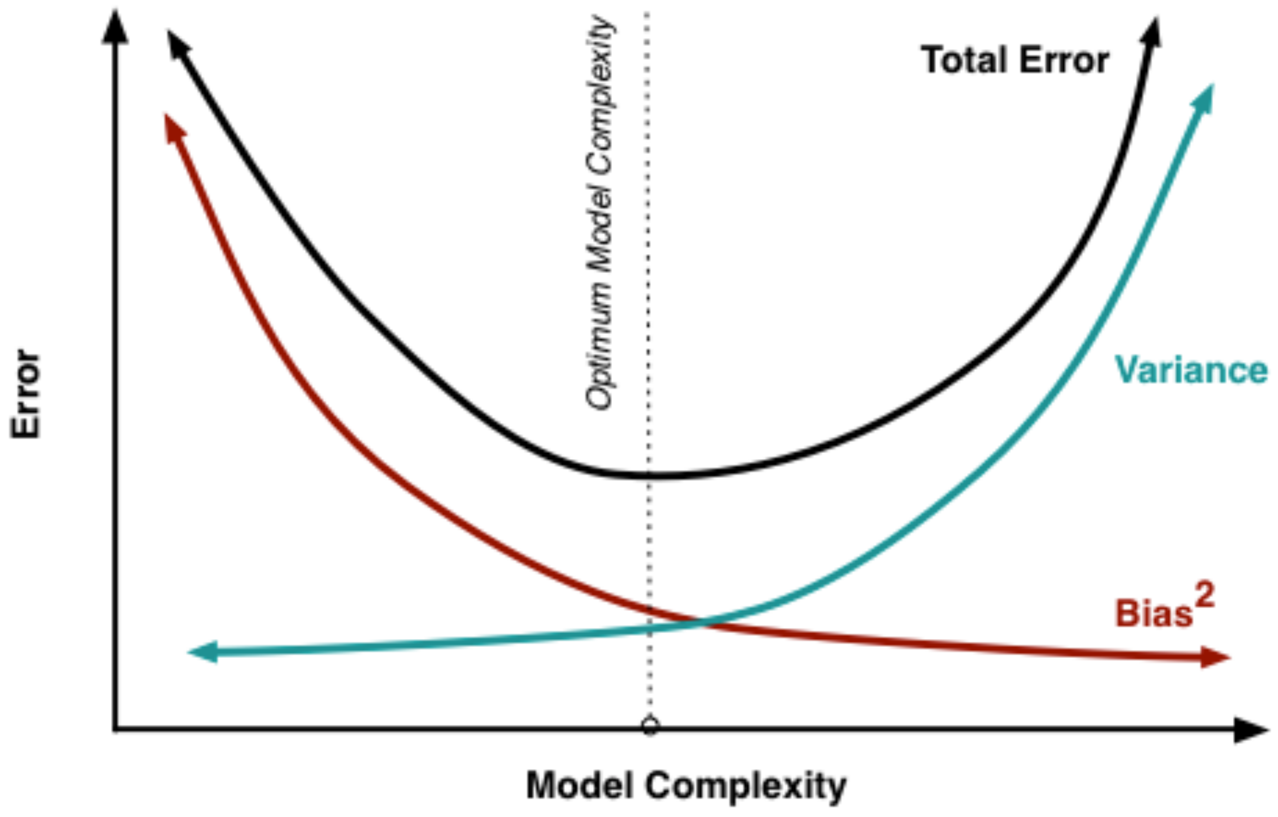
BALANCING BIAS

- Training error as a function of complexity.
- Question: why do we even care about variance if we know we can generate a more accurate model with higher complexity?
- Prediction error (red) increases. We have over-fit our model.



BALANCING BIAS AND VARIANCE

- As more parameters are added to the model, variance becomes the primary concern while bias falls.
- If our error exceeds this sweet spot, we are over-fitting. If error is below this sweet spot, we are under-fitting.
- We must use cross-validation.



RESOURCES

- http://scott.fortmann-roe.com/docs/BiasVariance.html
- https://courses.cs.washington.edu/courses/cse546/12wi/slides/cse546wi12LinearRegression.pdf