Univ.AI

Learning a Model

Last Time

1. What is x, f, y, and that damned hat?
2. Fitting simple models by minimizing loss
3. The generation story and how it gives us the loss
4. Noise and Sampling
5. Prediction and uncertainty

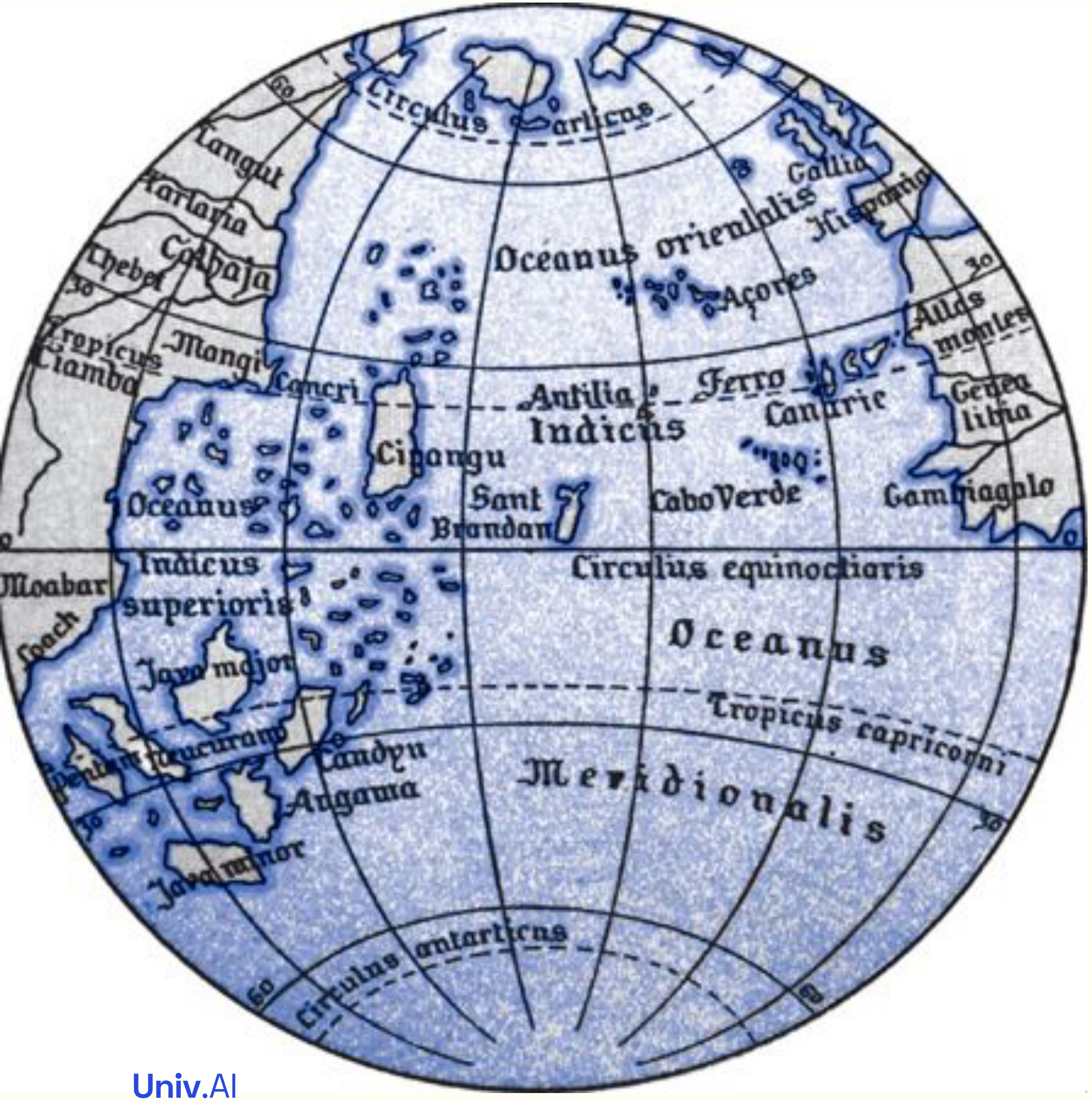
This time

1. SMALL World vs BIG World
2. Approximation
3. THE REAL WORLD HAS NOISE
4. Complexity amongst Models
5. Training, Testing, and Validation

1. SMALL World

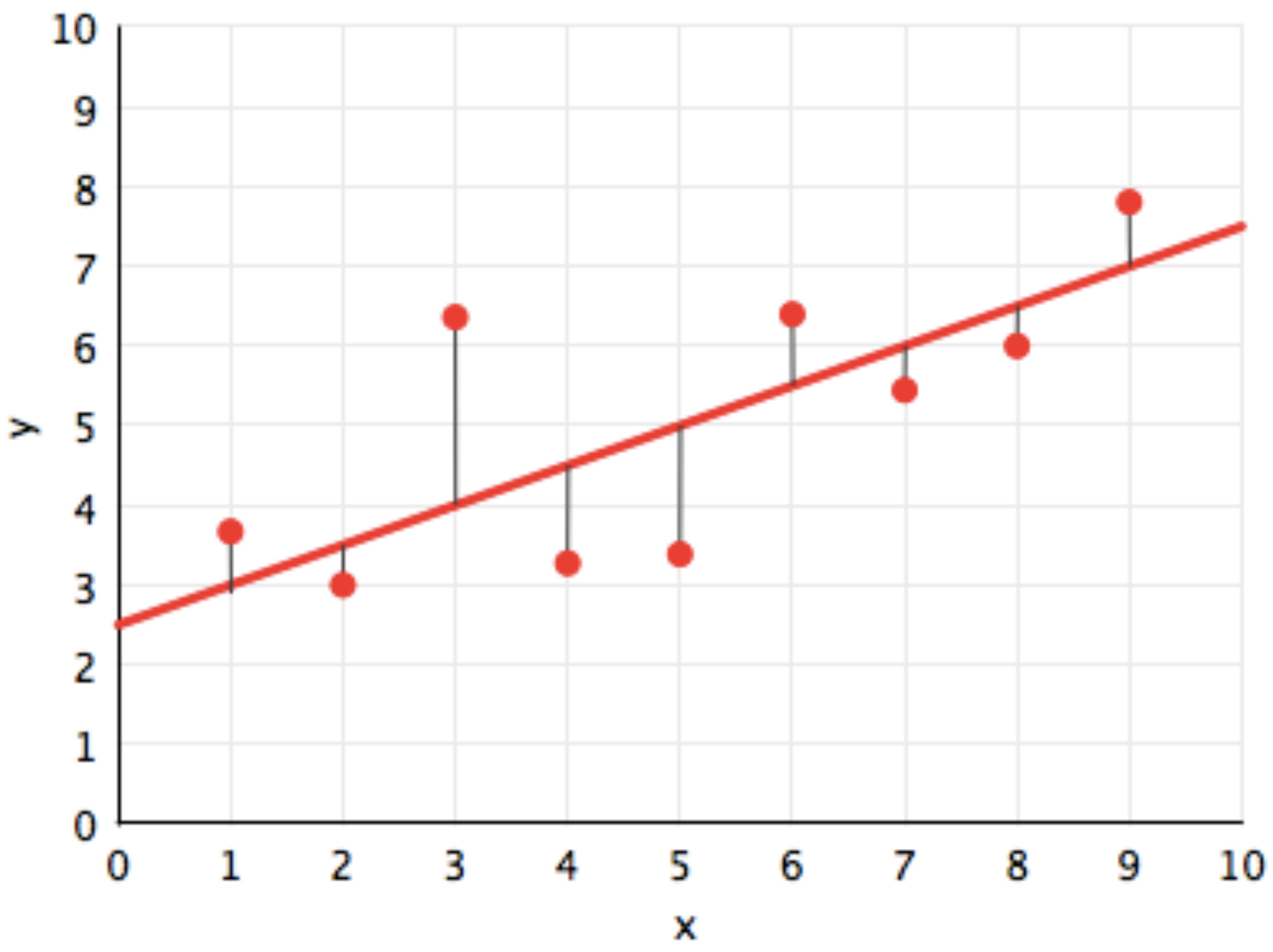
vs

BIG World



- *Small World* given a map or model of the world, how do we do things in this map?
- *BIG World* compares maps or models. Asks: whats the best map?

(Behaim Globe, 21 inches (51 cm) in diameter and was fashioned from a type of papier-mache and coated with gypsum. (wikipedia))



RISK: What does it mean to FIT?

Minimize distance from the line?

$$R_{\mathcal{D}}(h_1(x)) = \frac{1}{N} \sum_{y_i \in \mathcal{D}} (y_i - h_1(x_i))^2$$

Minimize squared distance from the line.
Empirical Risk Minimization.

$$g_1(x) = \arg \min_{h_1(x) \in \mathcal{H}_1} R_{\mathcal{D}}(h_1(x)).$$

Get intercept w_0 and slope w_1.

HYPOTHESIS SPACES

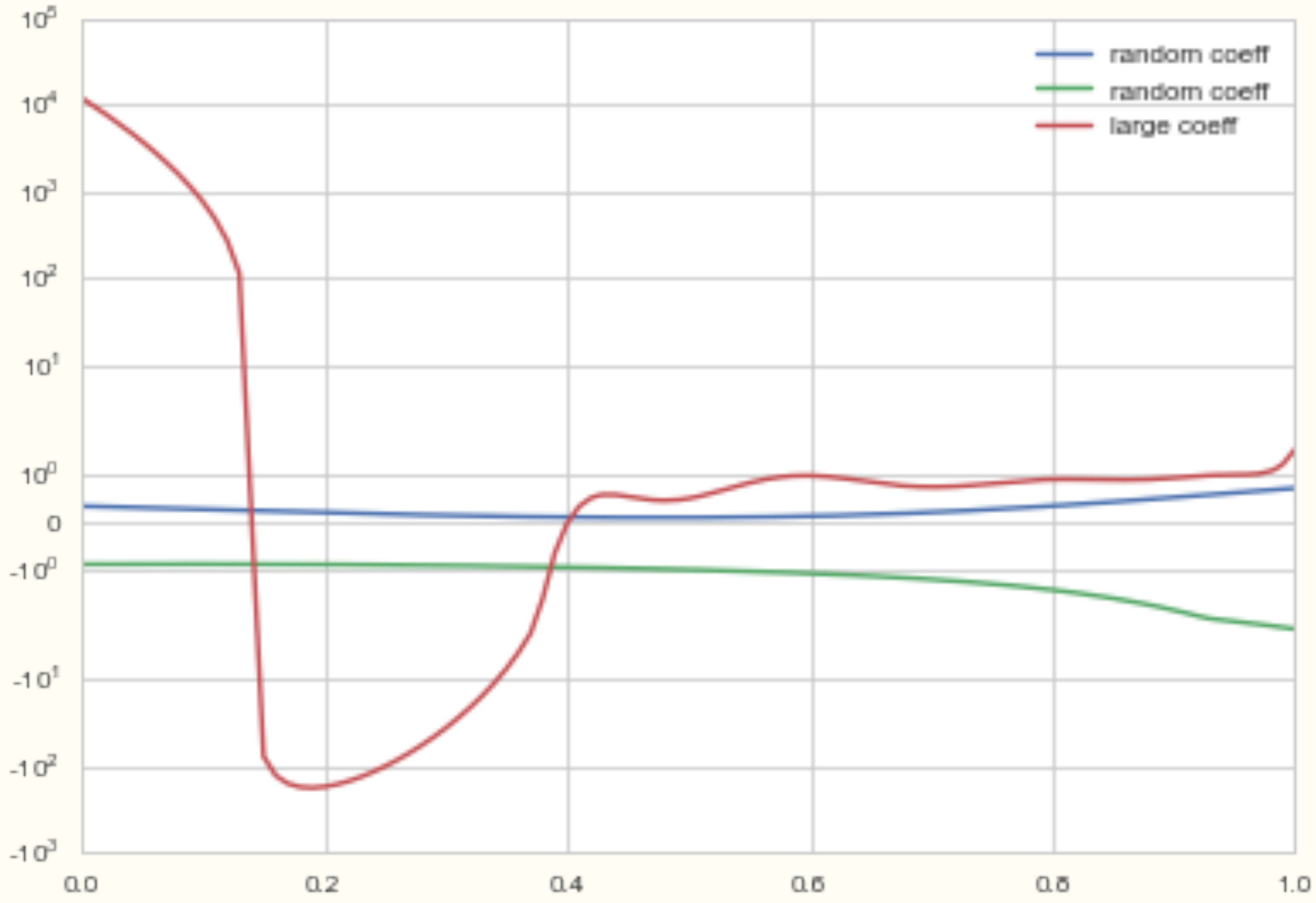
For example, a polynomial looks so:

$$h(x) = \theta_0 + \theta_1 x^1 + \theta_2 x^2 + \ldots + \theta_n x^n = \sum_{i=0}^{n} \theta_i x^i$$

All polynomials of a degree or complexity d constitute a hypothesis space.

$$\mathcal{H}_1 : h_1(x) = \theta_0 + \theta_1 x$$

$$\mathcal{H}_{20} : h_{20}(x) = \sum_{i=0}^{20} \theta_i x^i$$



Small World vs Big World, redux

Small World answers the question: given a model class (i.e. a Hypothesis space, whats the best model in it). Thus its looking for a particular $h(x)$ in a particular $\mathcal{H}$.

BIG World compares model spaces. Wants to find the true f(x), or at least the **best** $h(x)$ in the best $\mathcal{H}$ amongst the Hypothesis spaces we test.

Why not test ALL hypothesis spaces?

2. Approximation

Learning Without Noise...

Constructing a sample from a population

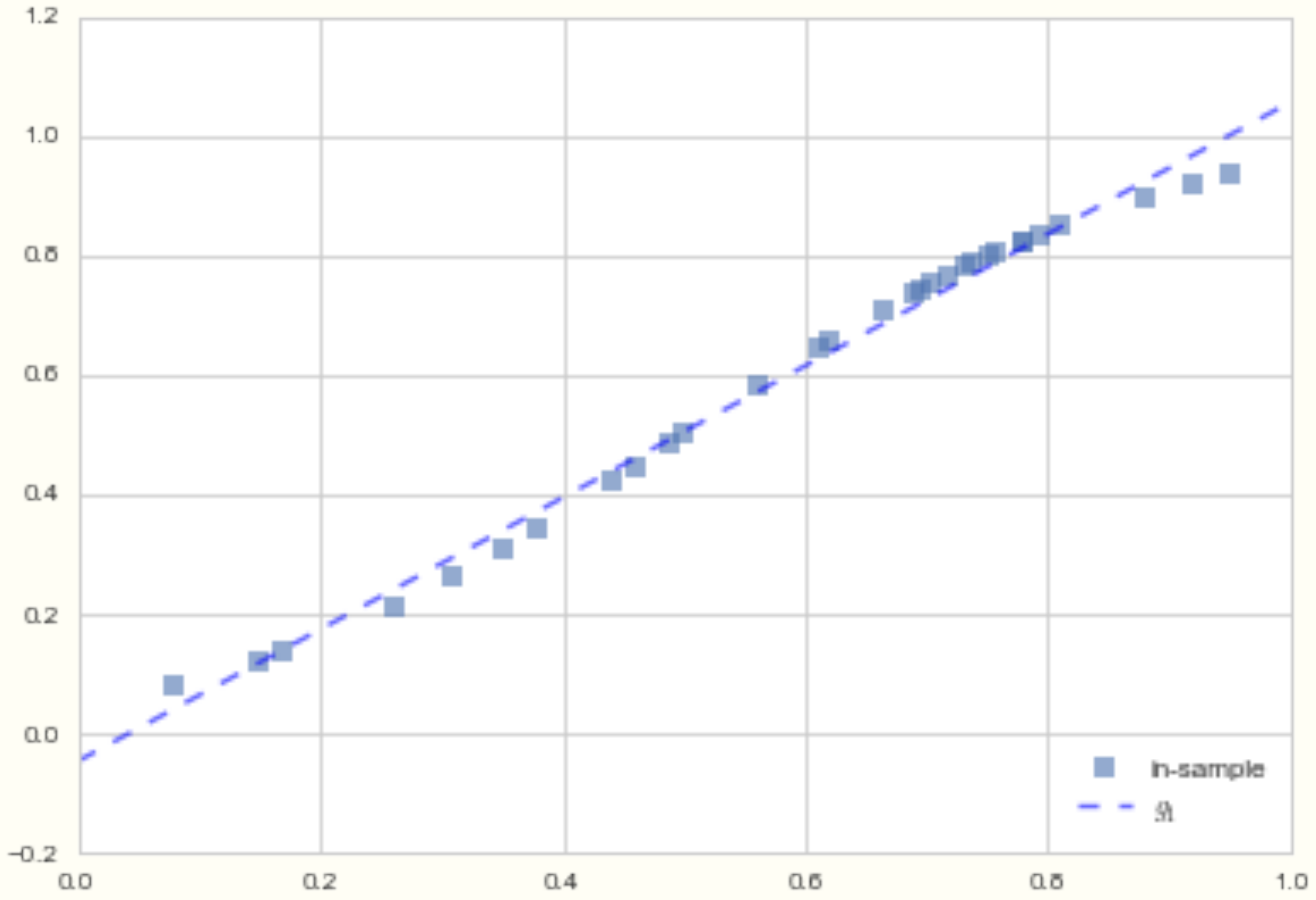
Well usually you are only given a sample. What is it?

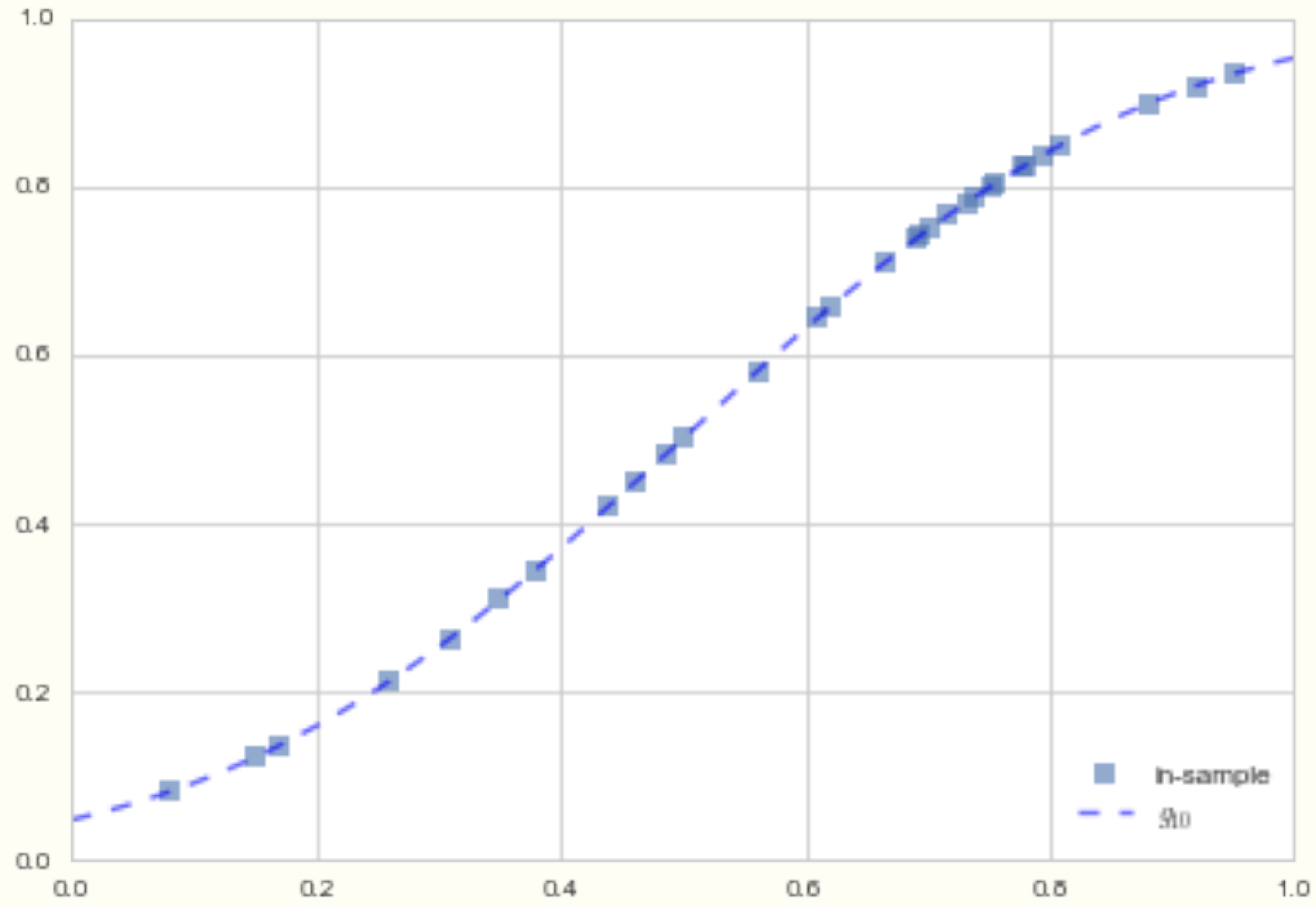
Its a set of (x, y) points chosen from the population.

If you had the population you could construct many samples of a smaller size by randomly choosing subsamples of such points.

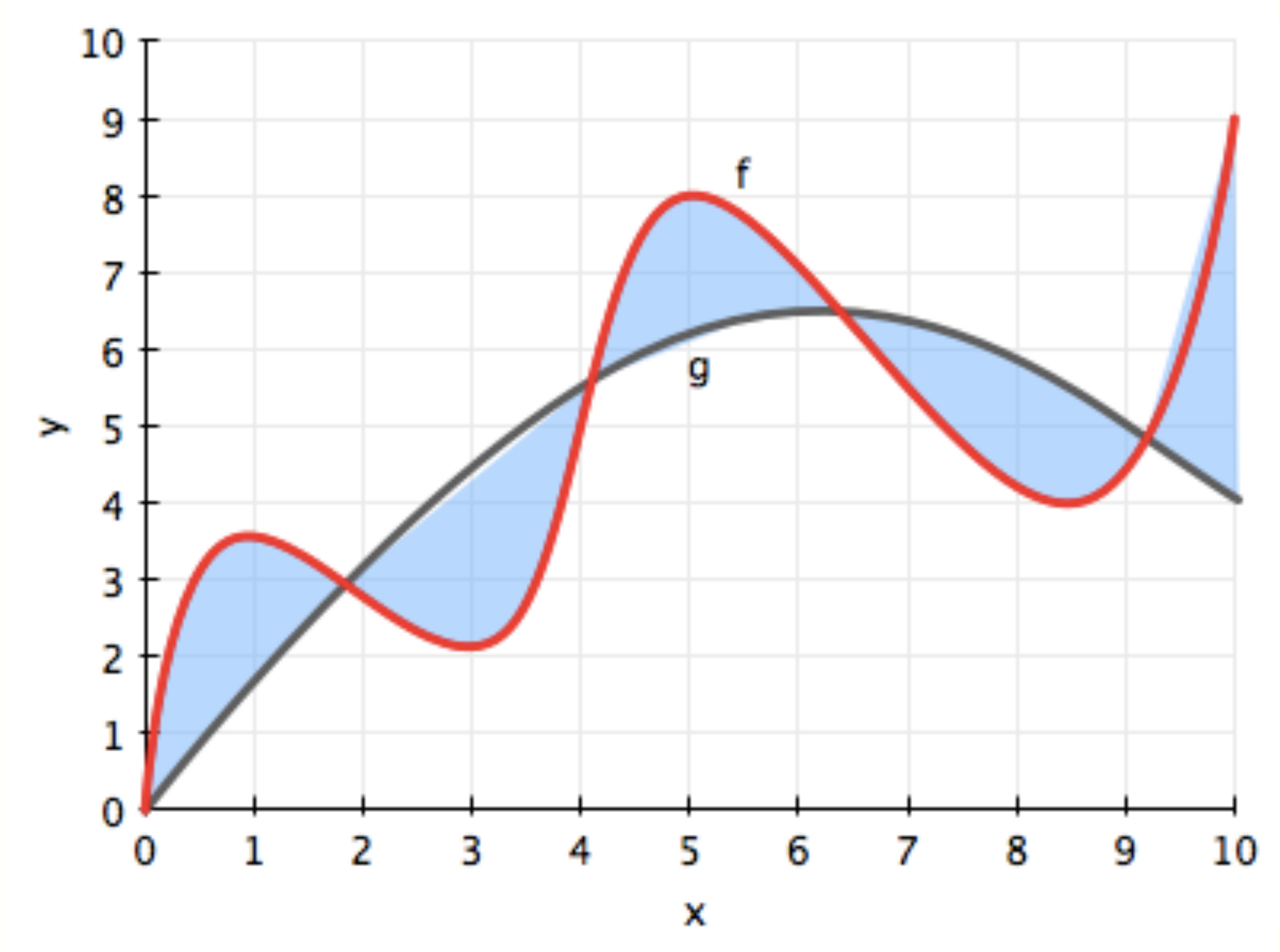
This is a game we will play...

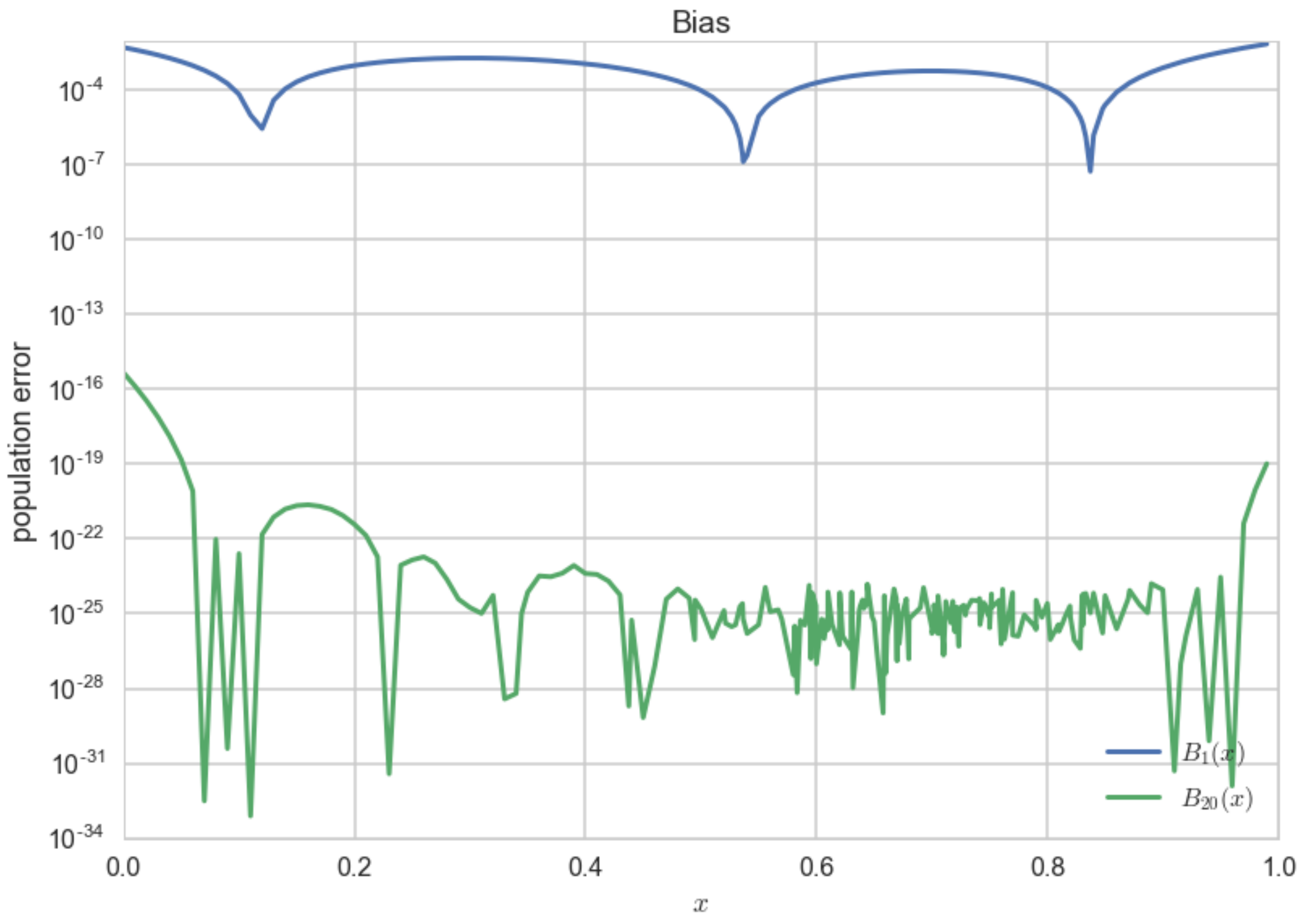
A sample of 30 points of data. Which fit is better? Line in $\mathcal{H}_1$ or curve in $\mathcal{H}_{20}$?



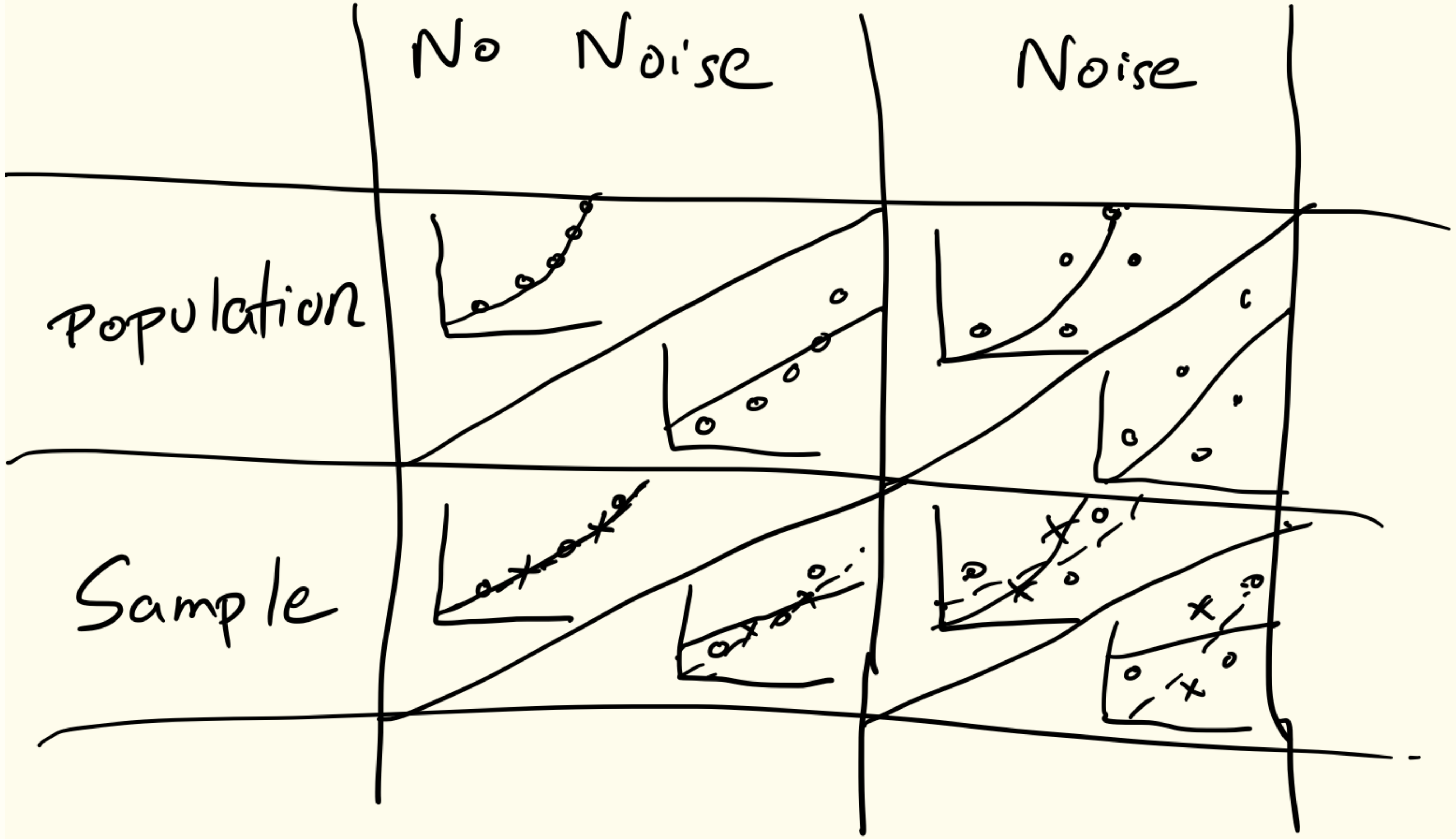


Bias or Mis-specification Error





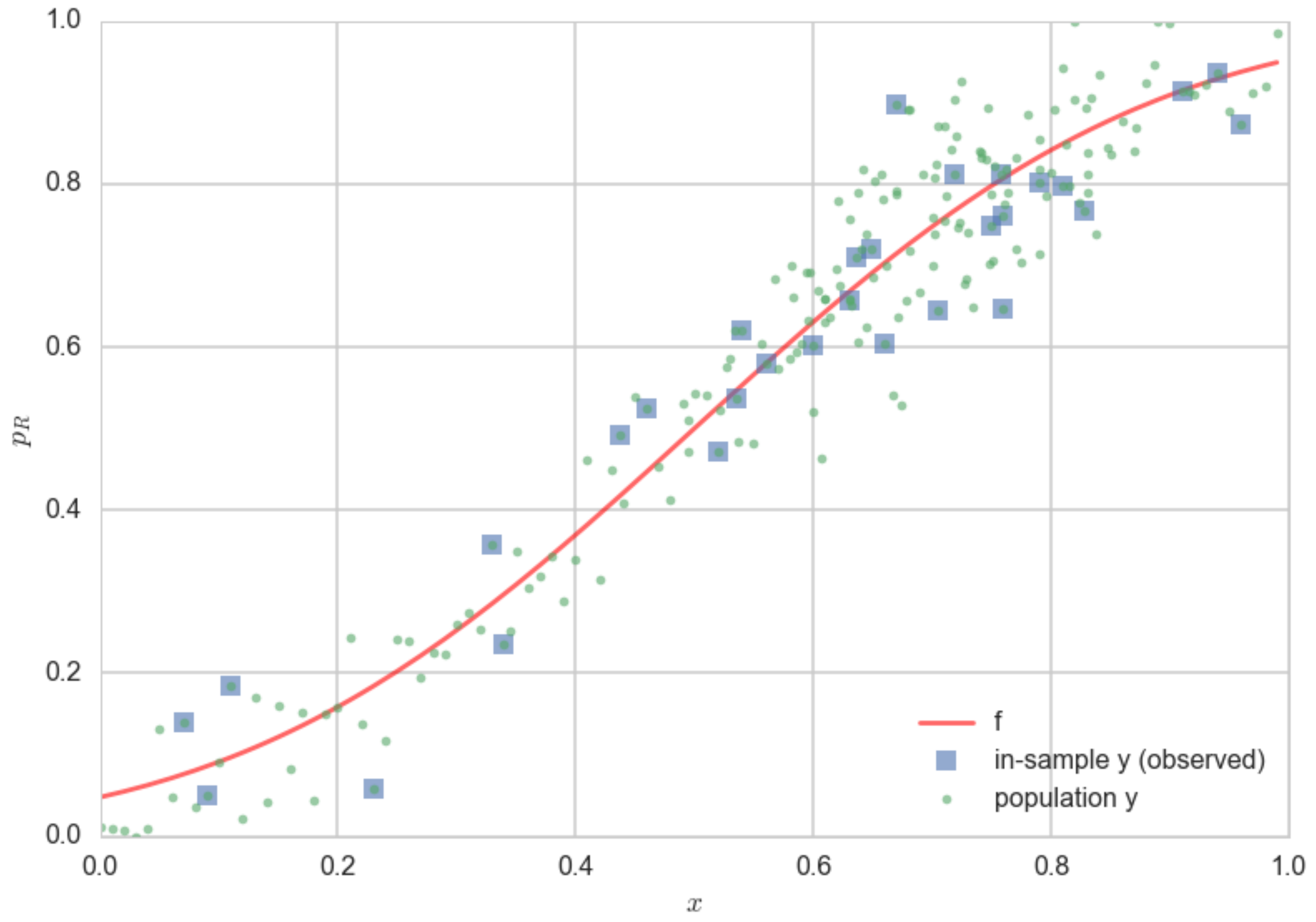
Interplay of (a) irreducible noise (b) bias (c) sampling.



3. THE REAL WORLD

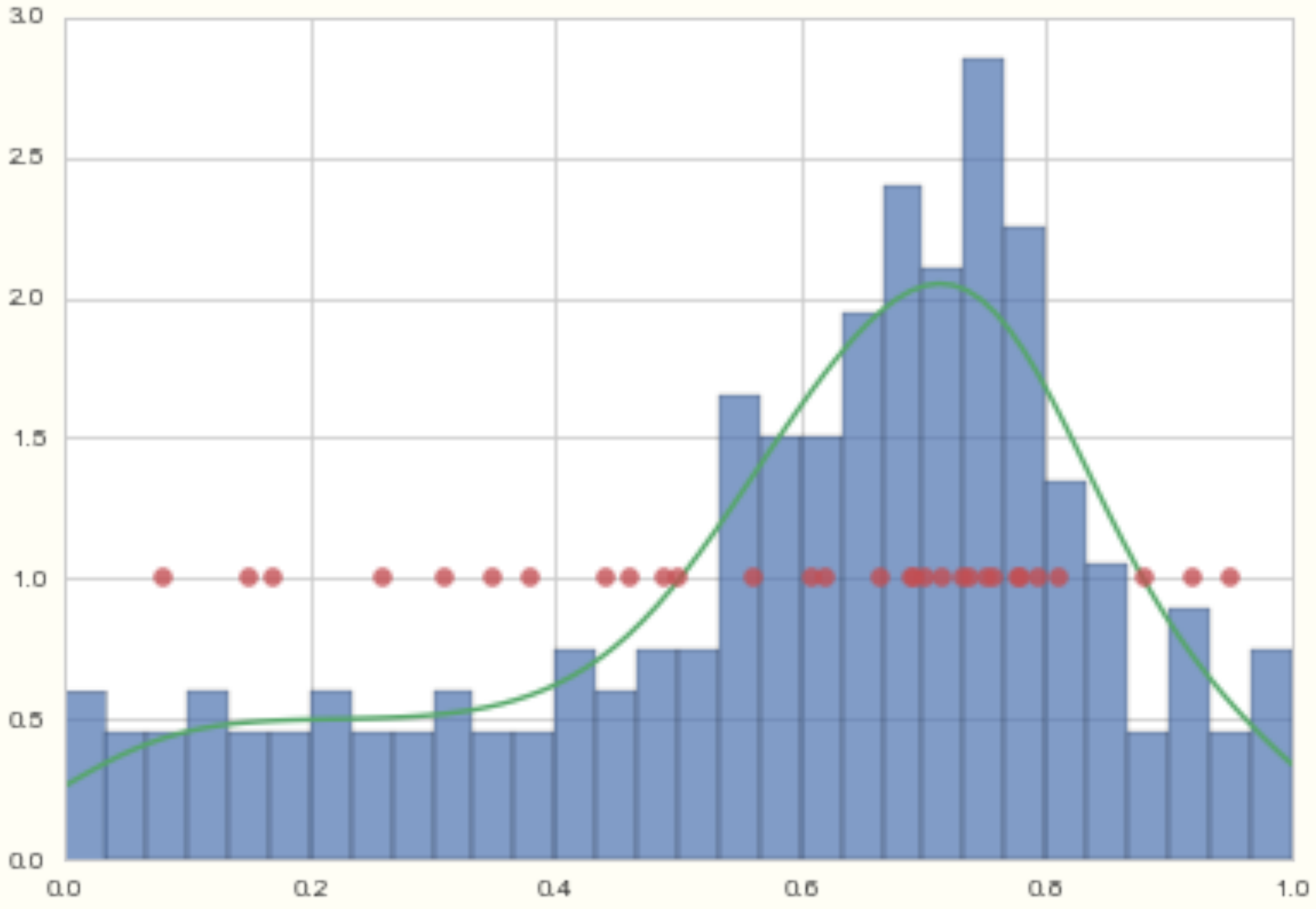
HAS NOISE

(or finite samples, usually both)



1.0
0.8
0.6
0.4
0.2
0.0
p_R
0.0
0.2
0.4
0.6
0.8
1.0
x
f
in-sample y (observed)
population y

Statement of the Learning Problem



The sample must be representative of the population!

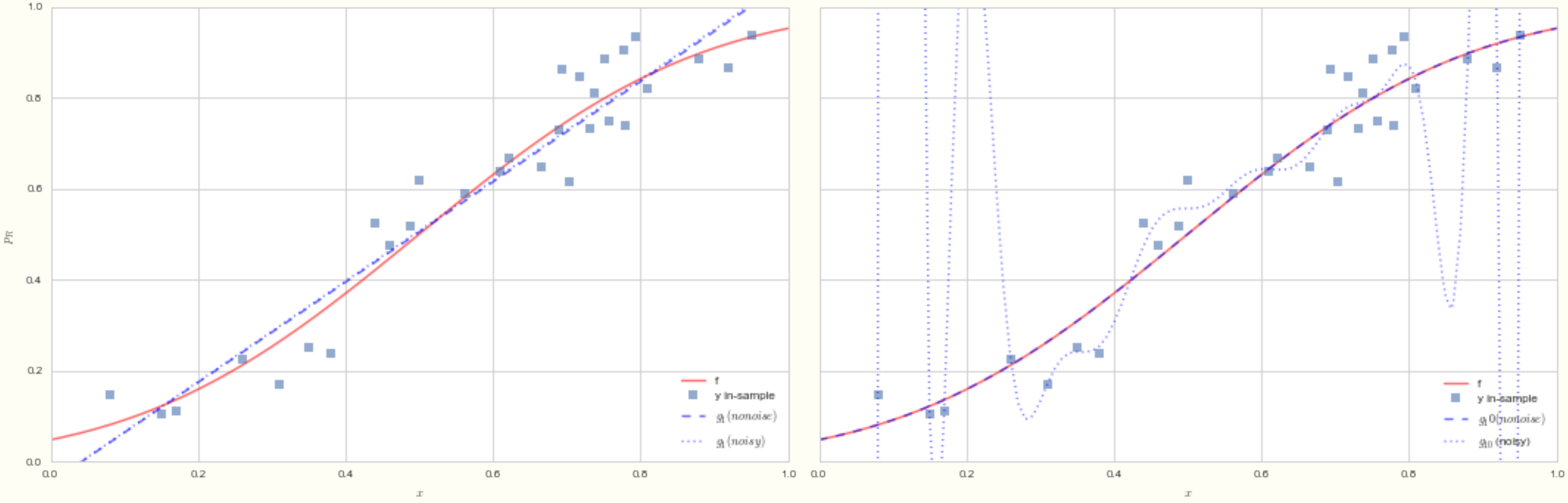
$$A : R_{\mathcal{D}}(g) \ \textit{smallest on}\ \mathcal{H}$$
$$B : R_{out}(g) \approx R_{\mathcal{D}}(g)$$

A: In-sample risk is small

B: Population, or out-of-sample risk is WELL estimated by in-sample risk. Thus the out of sample risk is also small.

Which fit is better now?

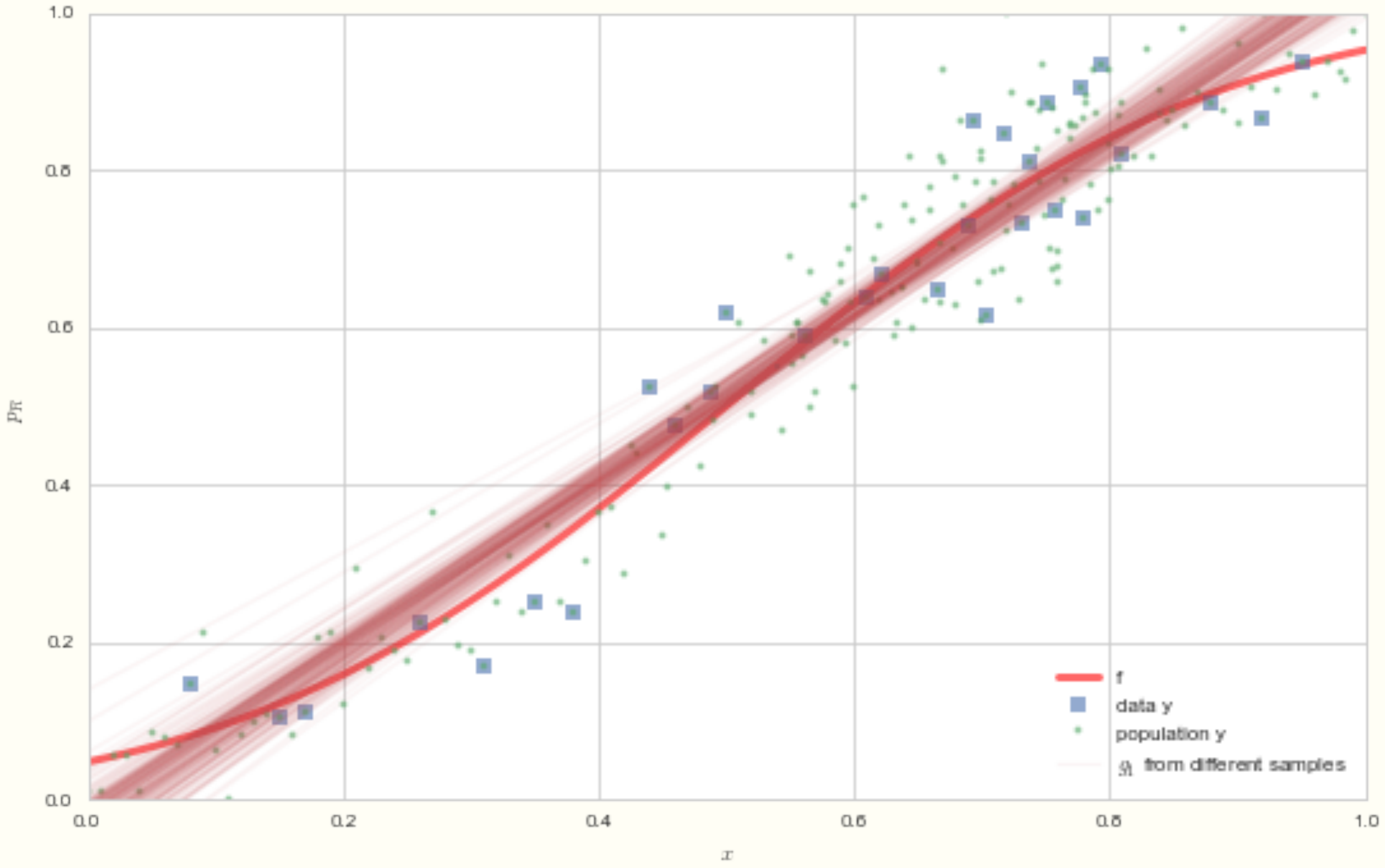
The line or the curve?

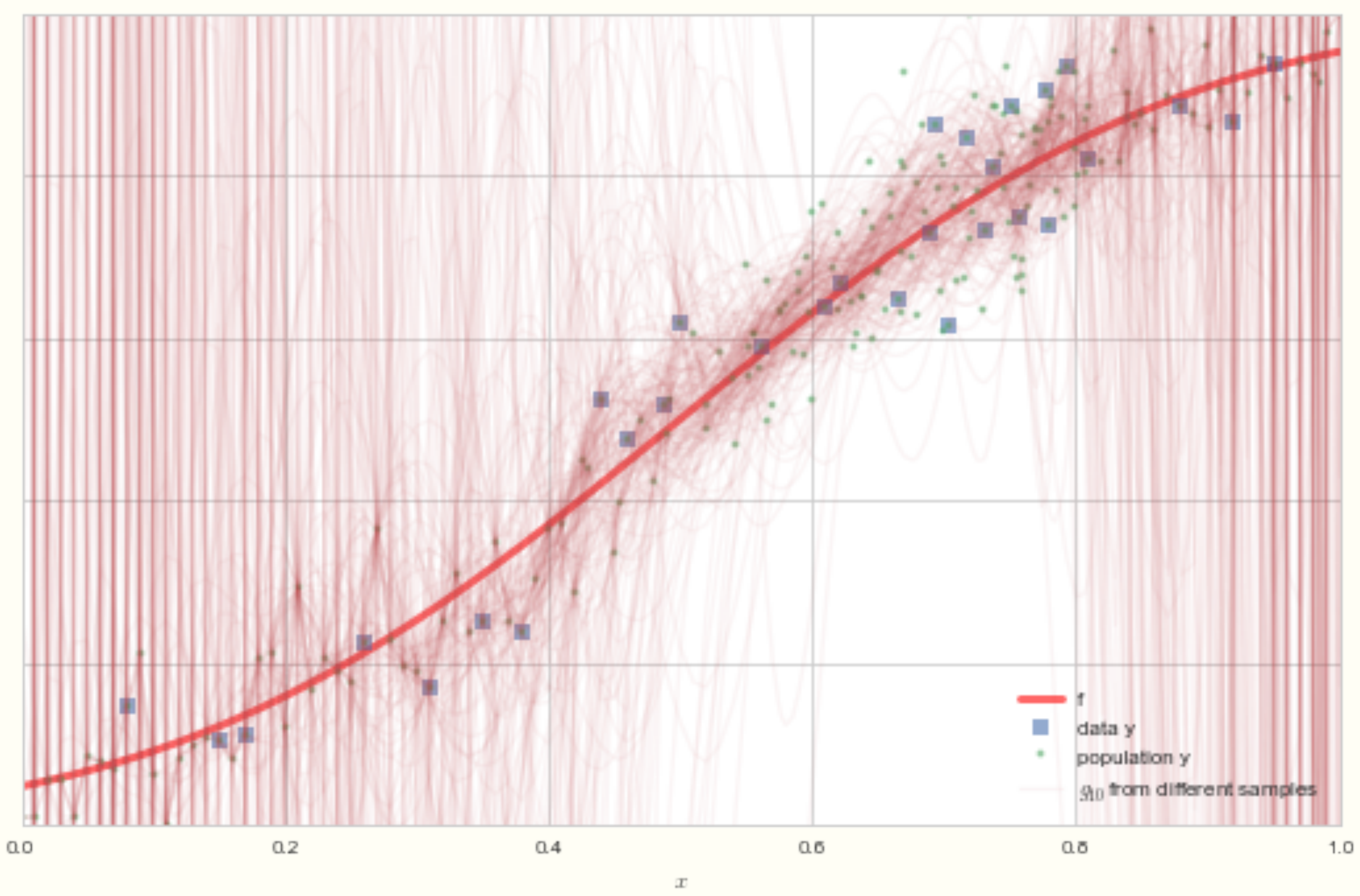


Training sets

- look at fits on different "training sets $\mathcal{D}$"
- in other words, different samples
- in real life we are not so lucky, usually we get only one sample
- but lets pretend, shall we?

UNDERFITTING (Bias) vs OVERFITTING (Variance)





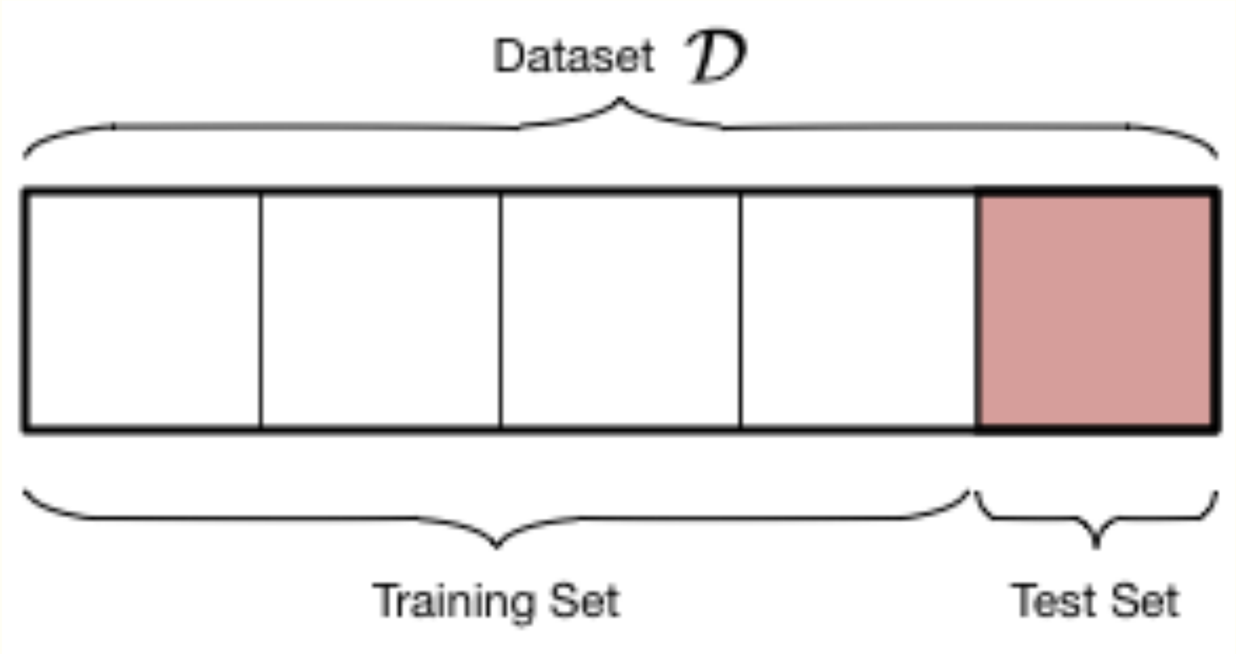
Univ.AI

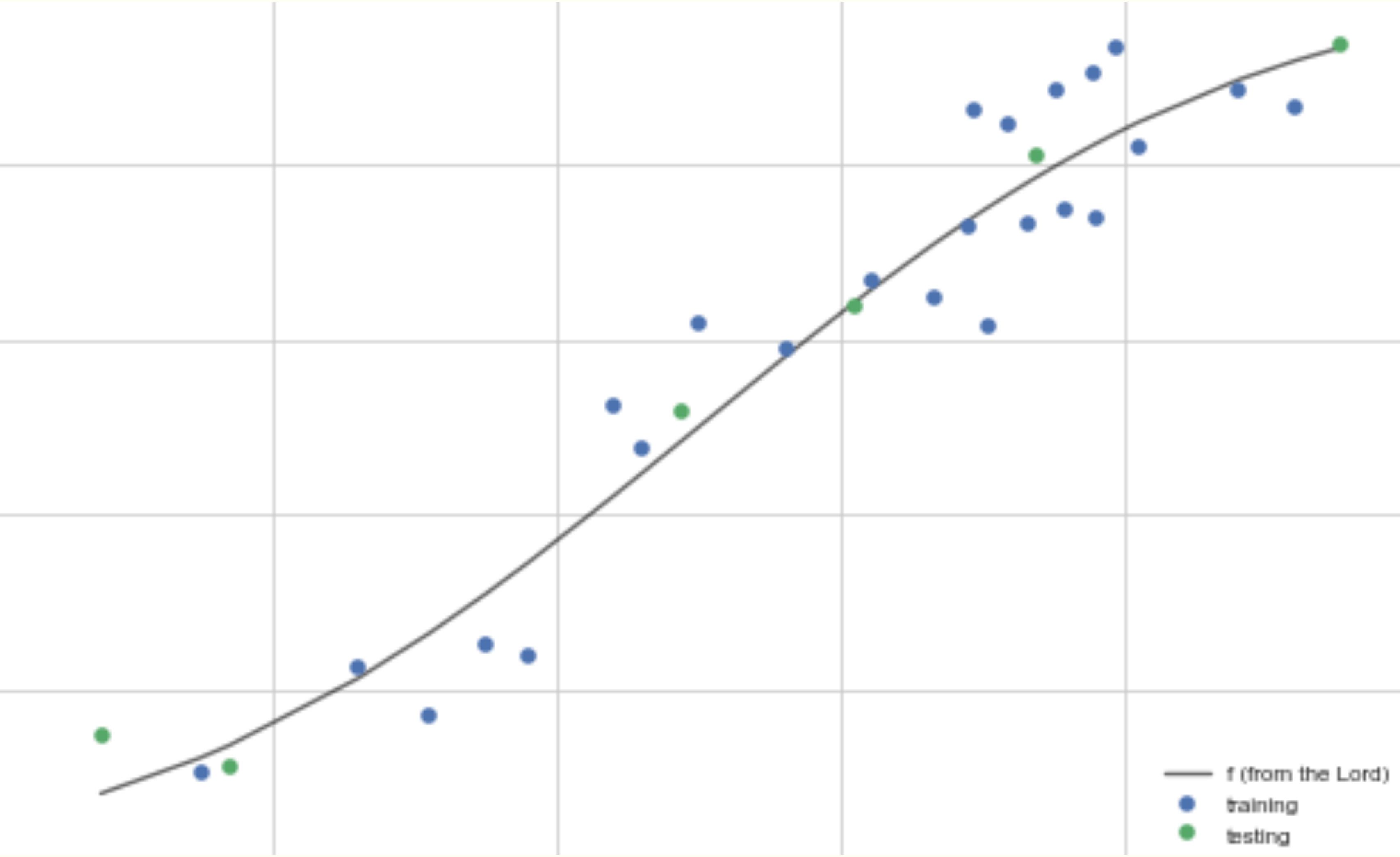
4. Complexity amongst Models

How do we estimate

out-of-sample or population error R_{out}

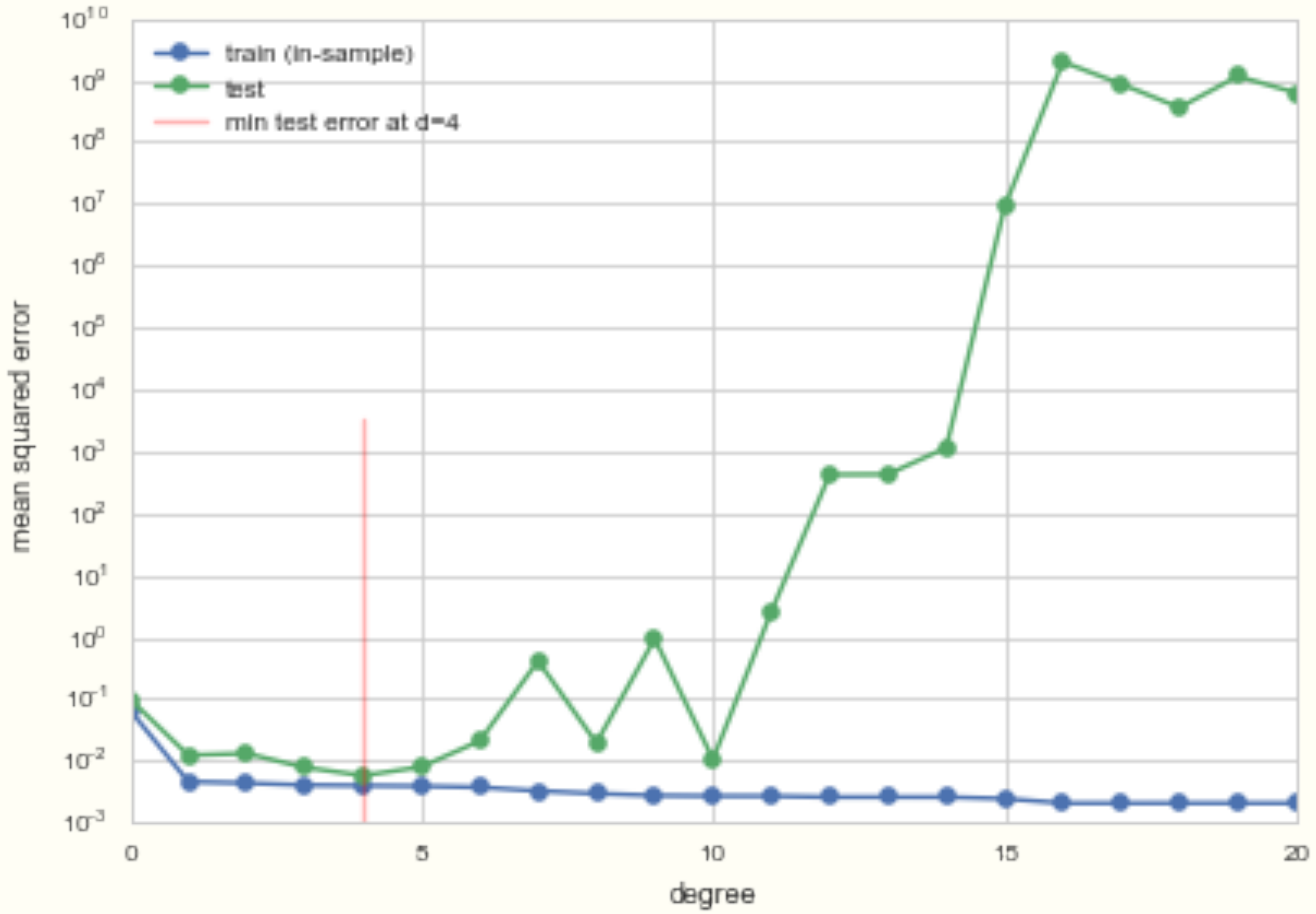
TRAIN AND TEST



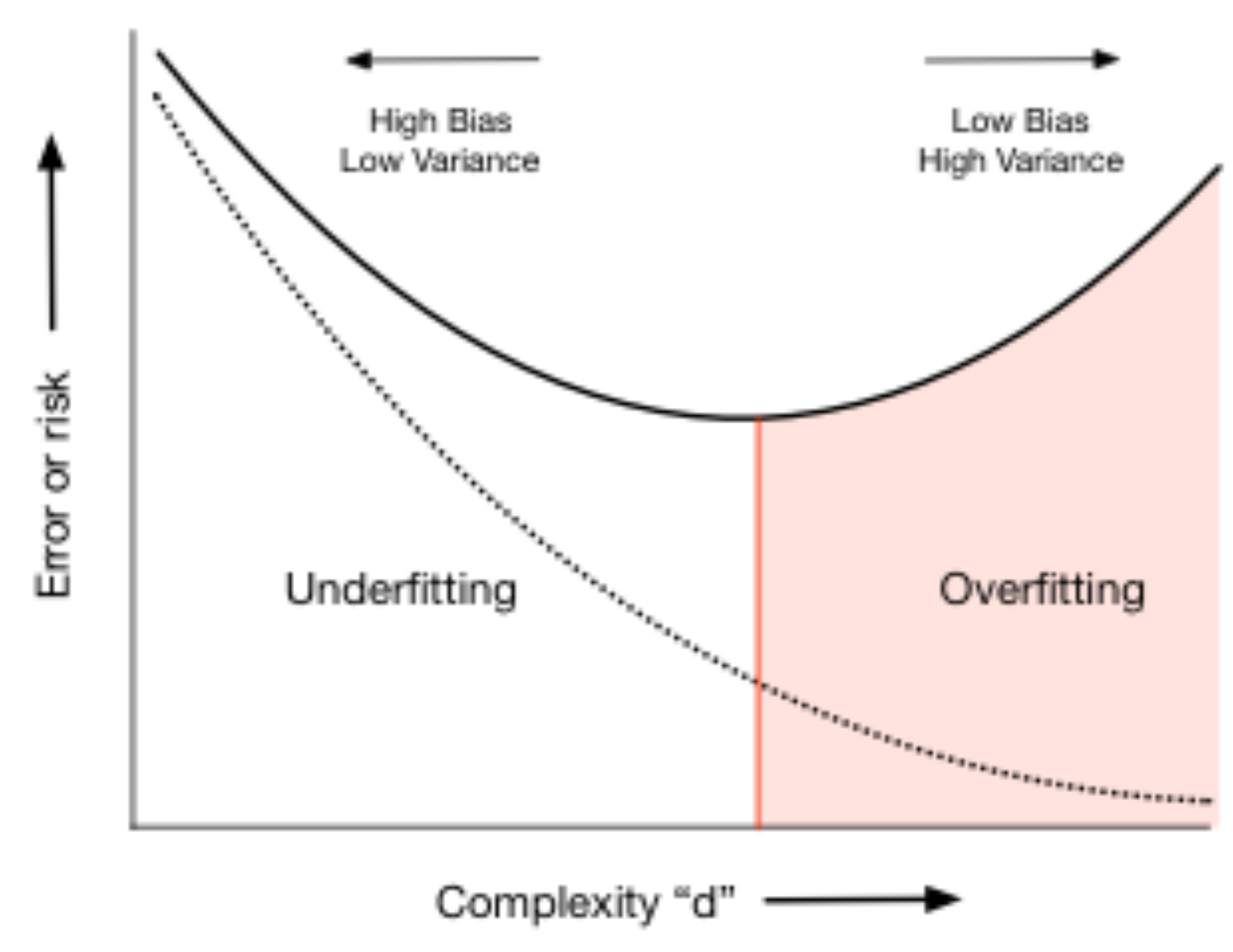


MODEL COMPARISON: A Large World approach

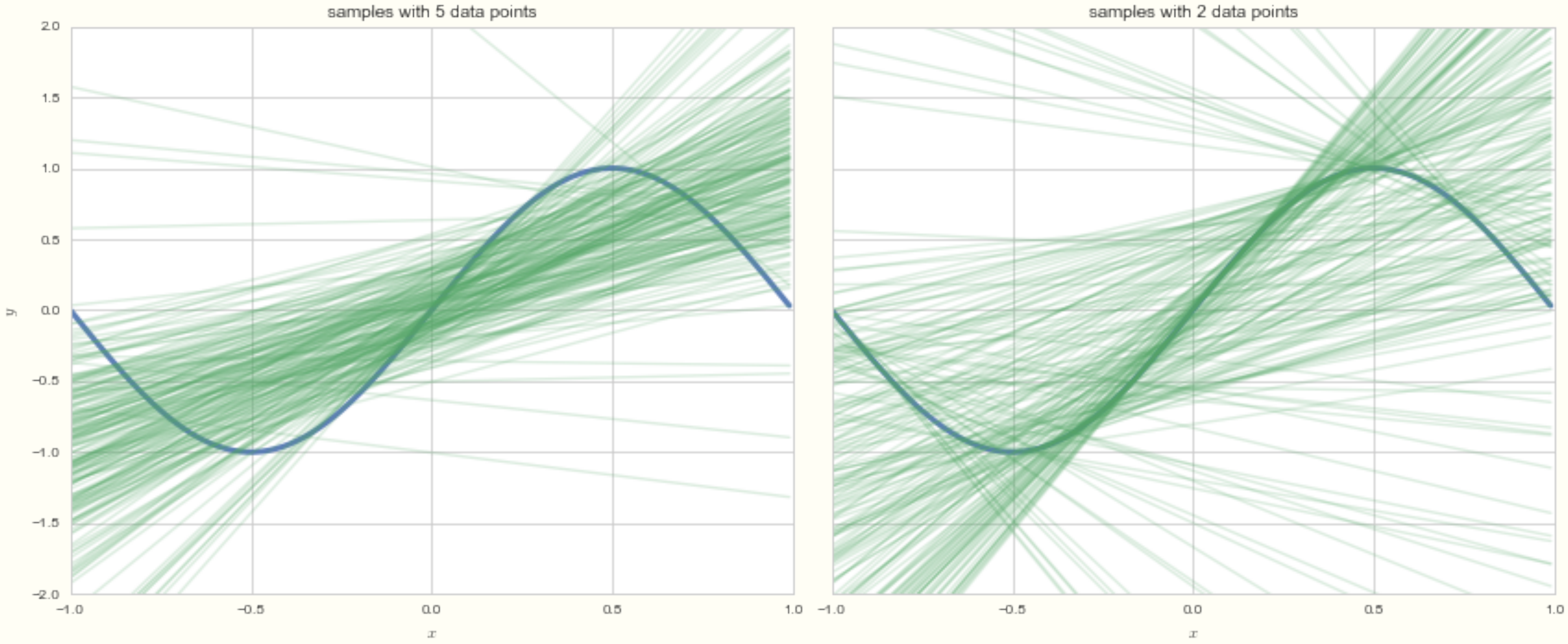
- want to choose which Hypothesis set is best
- it should be the one that minimizes risk
- but minimizing the training risk tells us nothing: interpolation
- we need to minimize the training risk but not at the cost of generalization
- thus only minimize till test set risk starts going up



Complexity Plot



DATA SIZE MATTERS: straight line fits to a sine curve



Corollary: Must fit simpler models to less data! This will motivate the analysis of learning curves later.

5. Validation

Do we still have a test set?

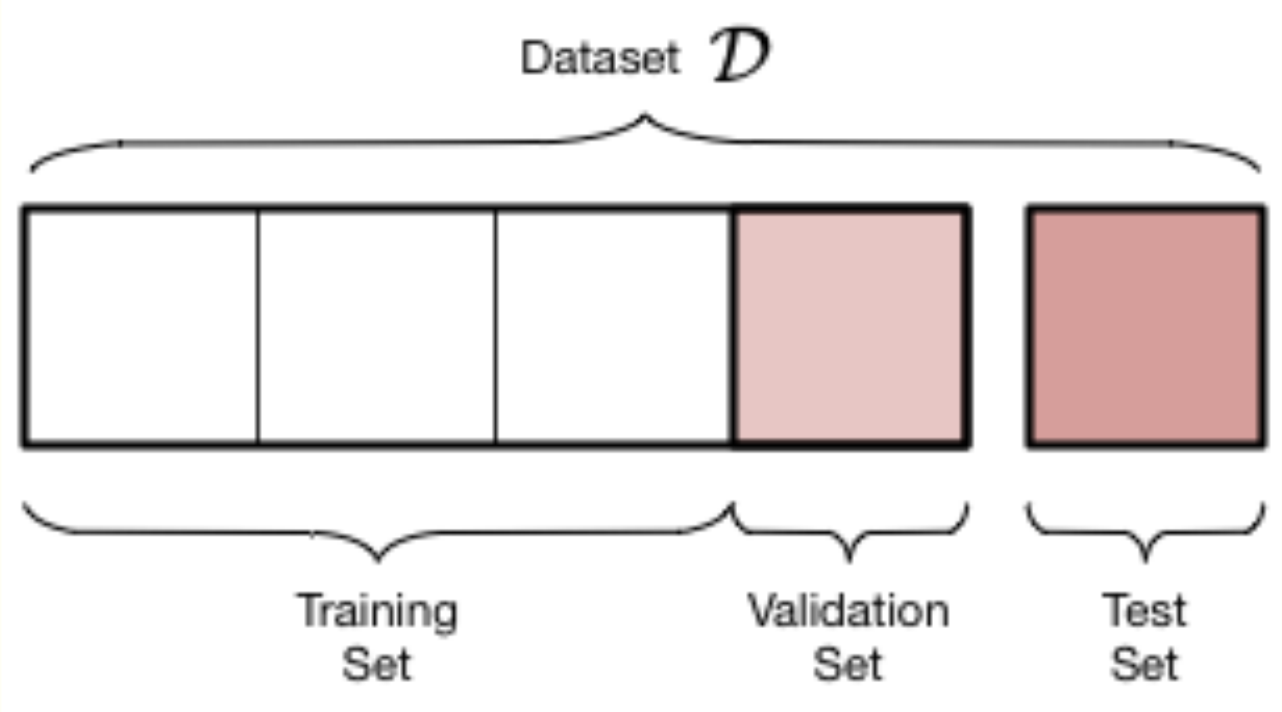
Trouble:

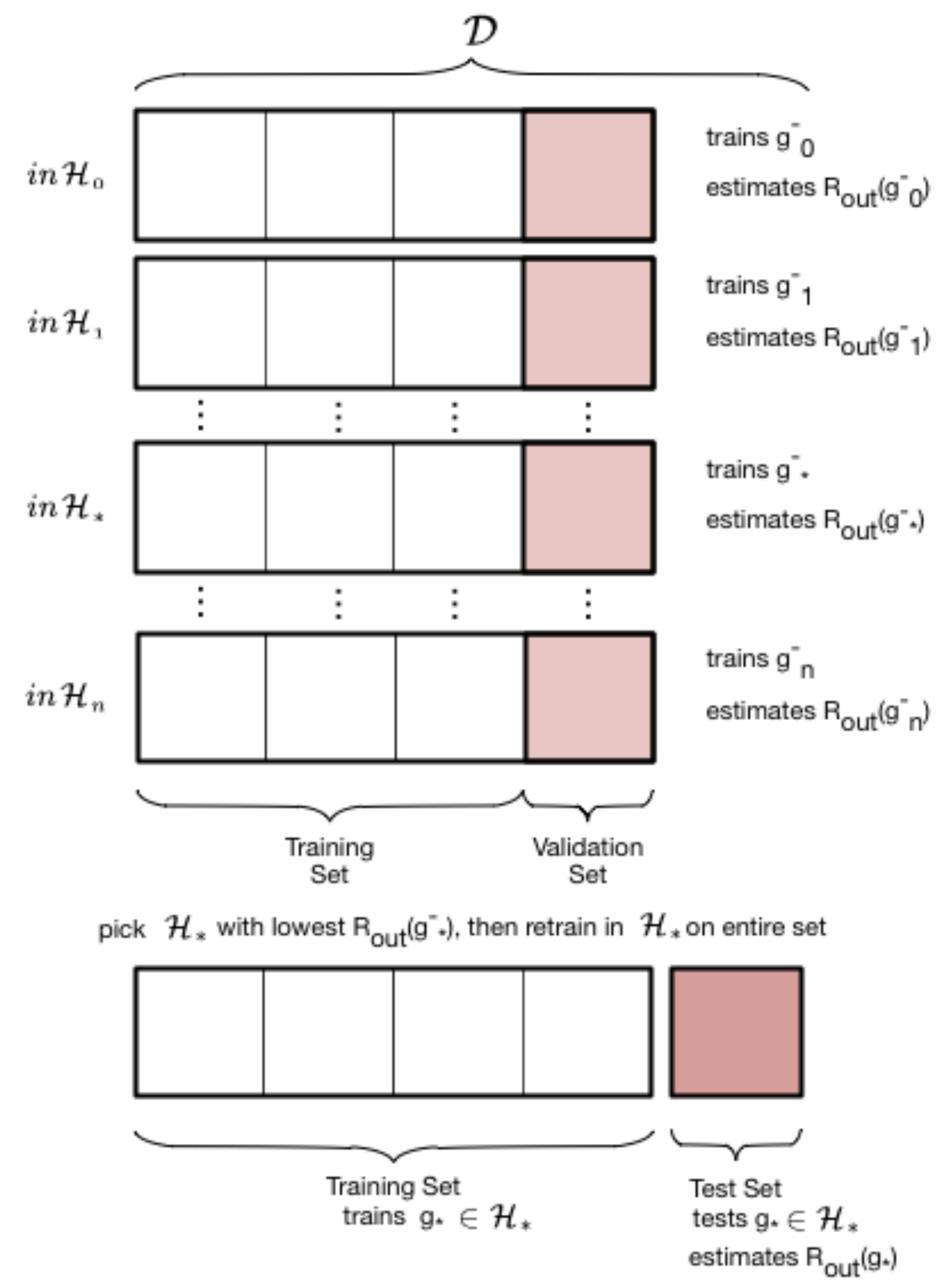
- no discussion on the error bars on our error estimates
- "visually fitting" a value of $d \implies$ contaminated test set.

The moment we **use it in the learning process, it is not a test set**.

VALIDATION

- train-test not enough as we *fit* for d on test set and contaminate it
- thus do train-validate-test





usually we want to fit a hyperparameter

- we **wrongly** already attempted to fit d on our previous test set.
- choose the d, g^{-*} combination with the lowest validation set risk.
- $R_{val}(g^{-*}, d^*)$ has an optimistic bias since d effectively fit on validation set

Then Retrain on entire set!

- finally retrain on the entire train+validation set using the appropriate d^*
- works as training for a given hypothesis space with more data typically reduces the risk even further.