Bias-variance
and
over fitting

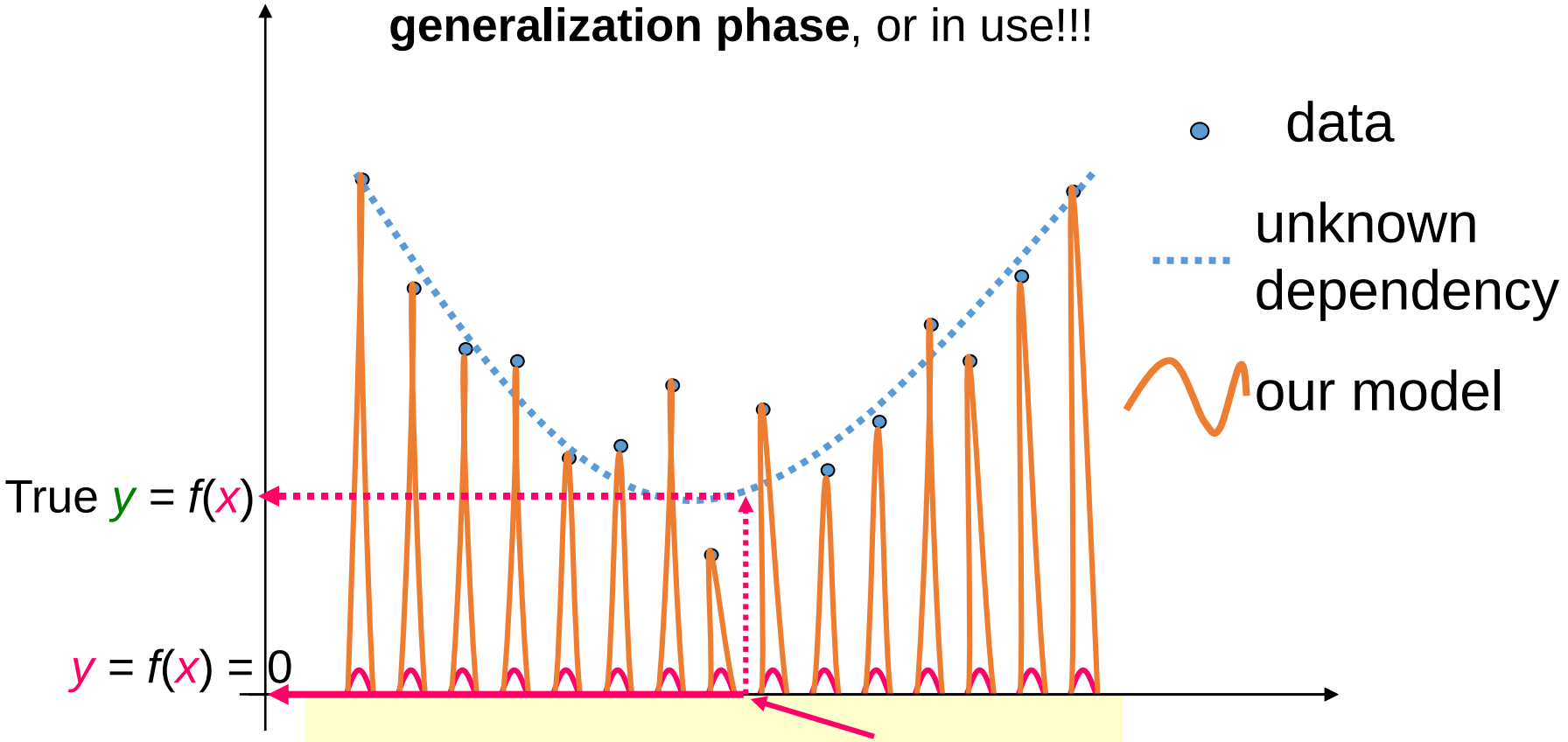
Bias – Variance - Dilemma!

It is the must piece of the knowledge in order to get an idea of the relationship between the data, models and errors!

Training and Generalization

Today, having powerful computers and good math software it is easy to be **'great** and **perfect'** on the training data set!

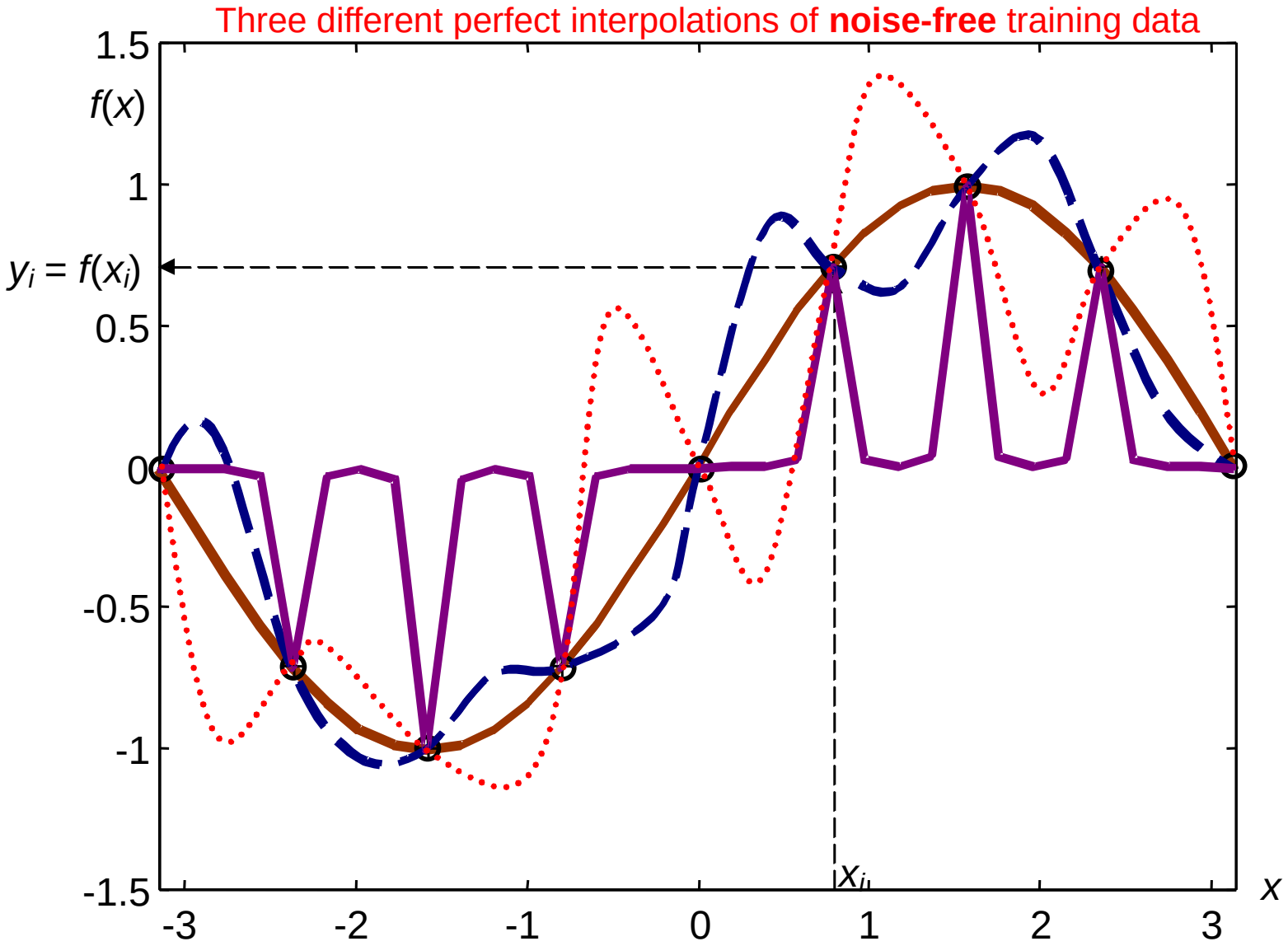
However, such a 'greatness' pays **heavy price at unseen data**, i.e., in a **generalization phase**, or in use!!!



This is (deliberately chosen) extremely bad modeling, **but real!**
The same or similar phenomena will be present in the high dimensional cases, too!

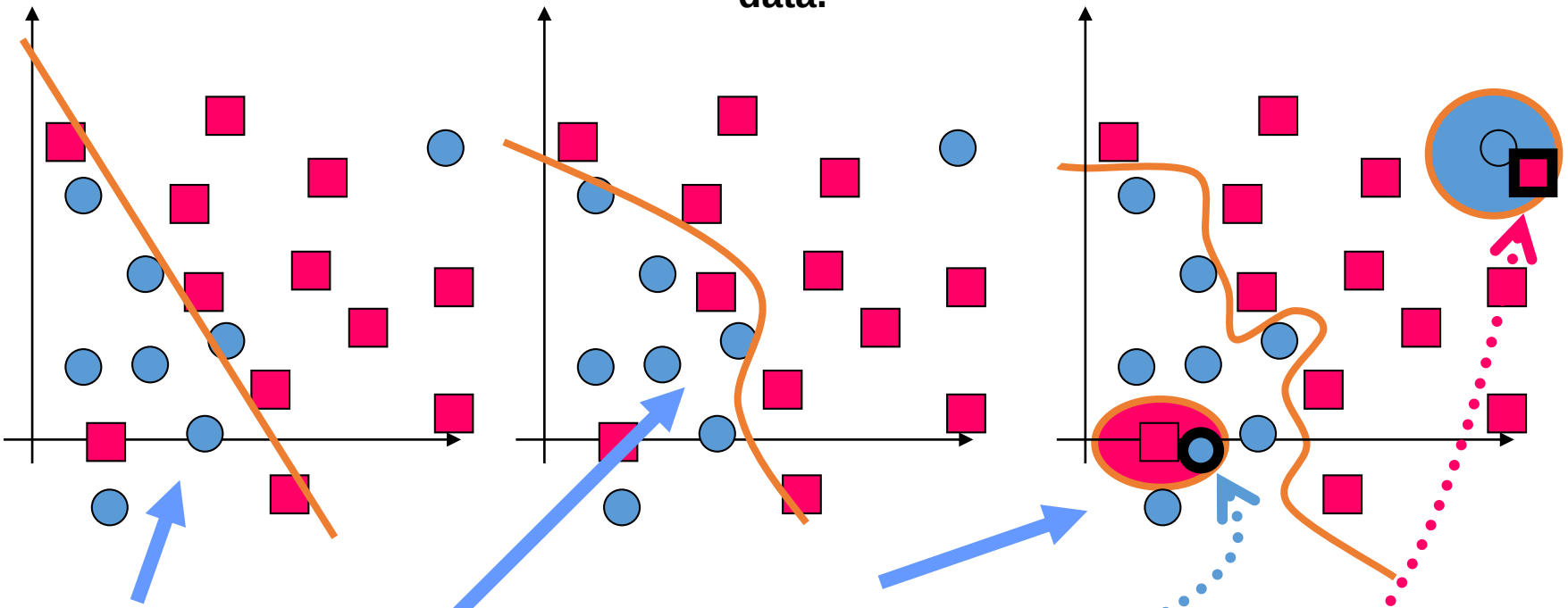
One more example showing **the perfect training results,** but **very bad generalization ones.**

Note that all three models have the training error equal zero! Bias = 0! Perfect interpolants!



One more example, but now from the PATTERN RECOGNITION (CLASSIFICATION) task, showing various models and their performances.

Note that the last model (learning machine) learns perfectly, i.e., **separates all the training data.**



On the left, the separation boundary is linear, and it misses not only the outliers, but some 'easy' points. The solution on the right does not miss anything. By having high capacity, it learns each data belongings 'by heart', but it is unlikely that it will perform well on the new data, say this one

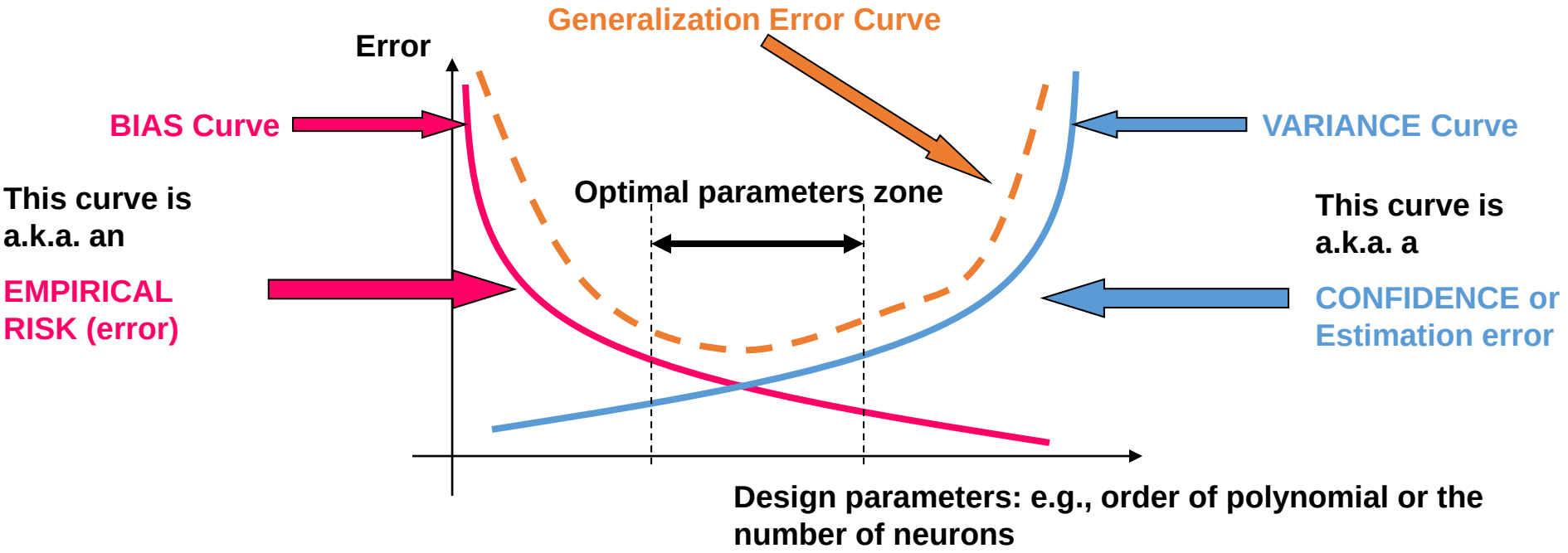
Or, this one

Central solution is of an intermediate capacity, separating most of the points, without putting too much trust into any particular training data point!!!

Obviously, we need much more than being good (or even excellent) on the training data set!

This ‘more’ means, we want that our models perform well on all future, previously unseen data, generated by the same data generator (i.e., plant, system, process, probability distribution).

The entire statistical learning fights (optimizes) the following two curves!



Although the graph looks very simple, finding the optimal modeling parameters is an

EXTREMALY DIFFICULT task!!!

and, this is due to the following facts:

- **we never (or, rarely only) know the underlying probability distribution, meaning the data generation, function,**
- **we never know the space of (target) functions, or to which class of functions our f. belongs**
- **we always have scarce (insufficient, not enough) data,**
- **our data are always high- (or/and extremely high-) dimensional,**
- **there is always the noise, or data are corrupted**

Bias & **Variance**

In modeling an **unknown dependency (regression or discrimination function)**, without knowledge of its mathematical form (**target space**), our **models (functions from hypothesis space**) produce **approximating functions**, which may be incapable of representing the **target function** behavior.

A difference between the model output and unknown target function is called **the bias.**

When there are not sufficient data, (or even if there appears to be sufficient representative data, **noise** contamination can still contribute that) **the sample of data** that is **available for training** *may not be representative of* ***average*** *data generated by the target function*.

Consequently, there may be a difference between a network output **for a particular data set**, and network function output **for the average of all data sets** produced by the target function.

The *square* of this difference is called **the variance.**

In other words, model's **bias** is a measure of **how well we can model the underlying unknown function with some function from hypothesis space H .**

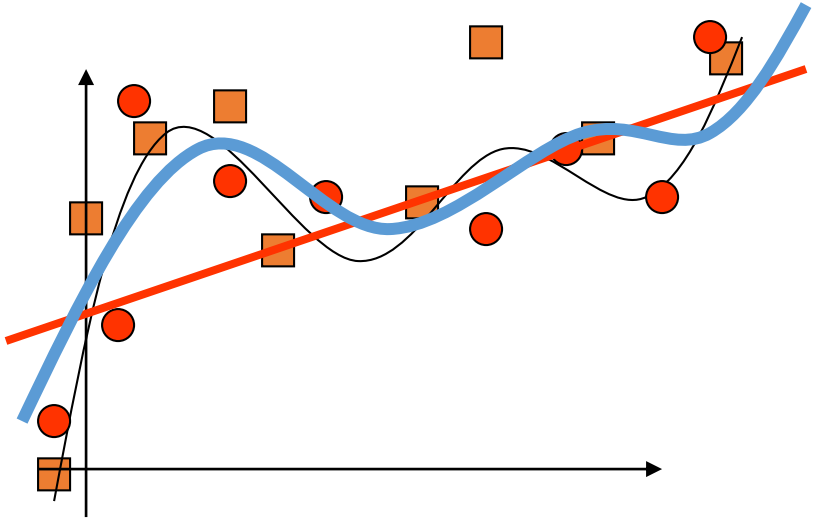
If the underlying function can be modeled perfectly with a model from our hypothesis space, i.e. if the underlying model is a member of our hypothesis space **H** , **then we say that our hypothesis space has zero model bias, or that it is unbiased.**

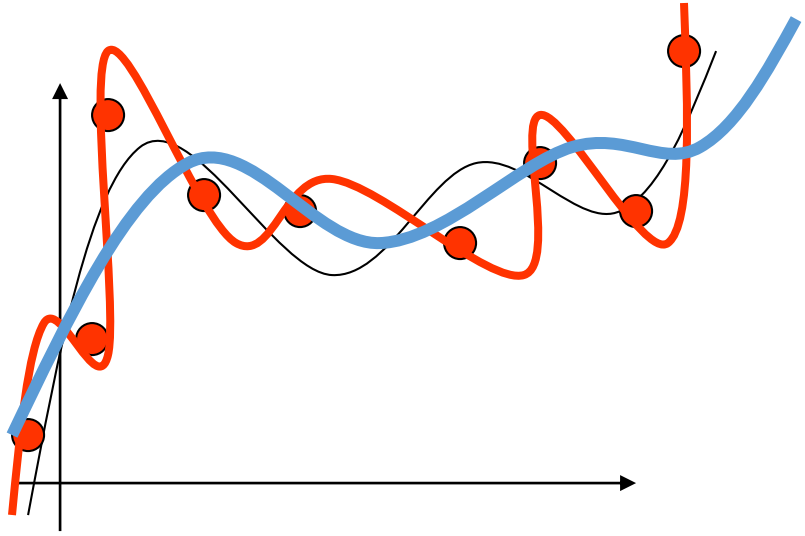
If the underlying function is not a member of the hypothesis space, then we say that **our model family is biased.**

Model's **variance** is a measure of how much our models **vary** when we train them with different training sets. If the hypothesis space **H** is very small, then there will be small differences between models trained with different training sets and we say that the model variance is small.

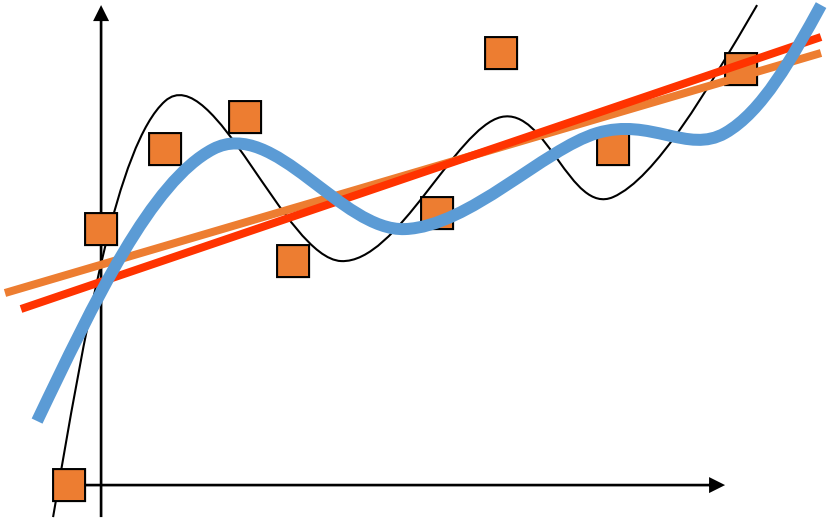
On the other hand, if the model family is large, then there can be (*and according to Murphy's law there will be* large differences between models trained with different training sets and we say that the model variance is large.

We explain the above, by presenting the geometrical (graphical) meaning of BIAS and VARIANCE! Corresponding math doesn't fit here!

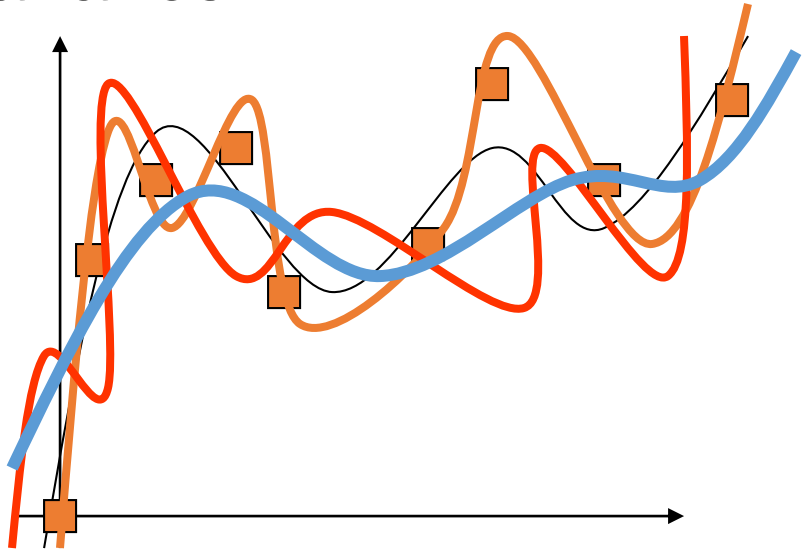




Good model designer would try to get medium both the Bias and the Variance



High Bias - Low Variance



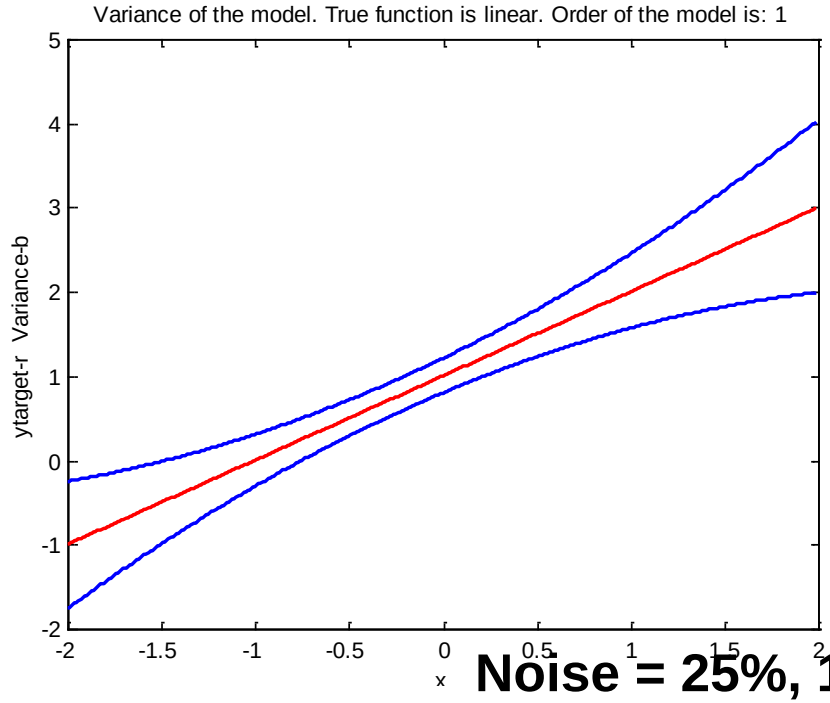
Low Bias - High Variance

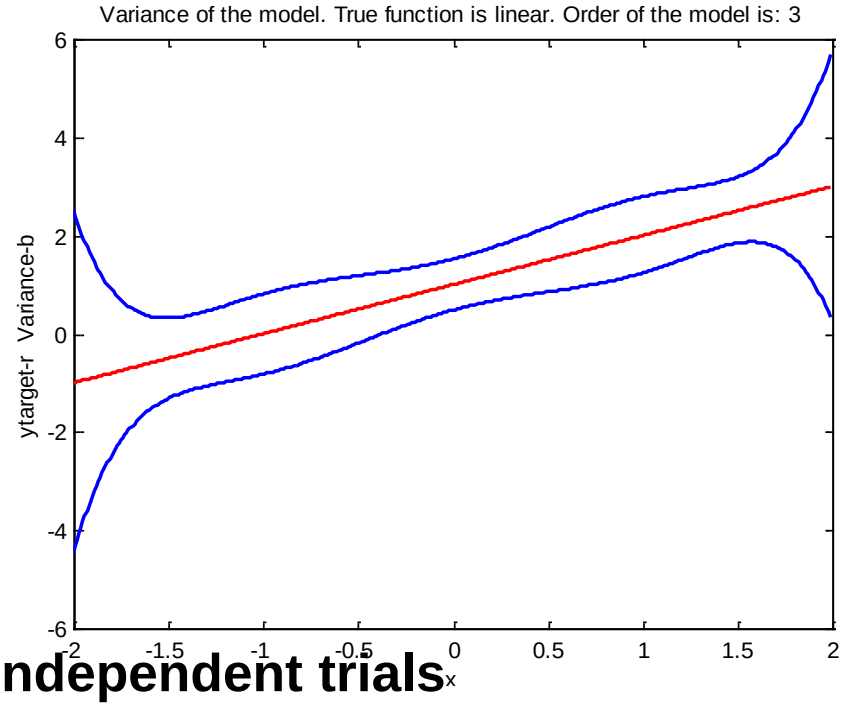
Two examples of Variance:
We explore the influence of a hypothesis space and the noise

Ex 1:
Suppose that the **unknown** underlying function ***g*** **is linear.**
Hypothesis space **H** = { all linear models } has **zero model bias** and **small model variance.**

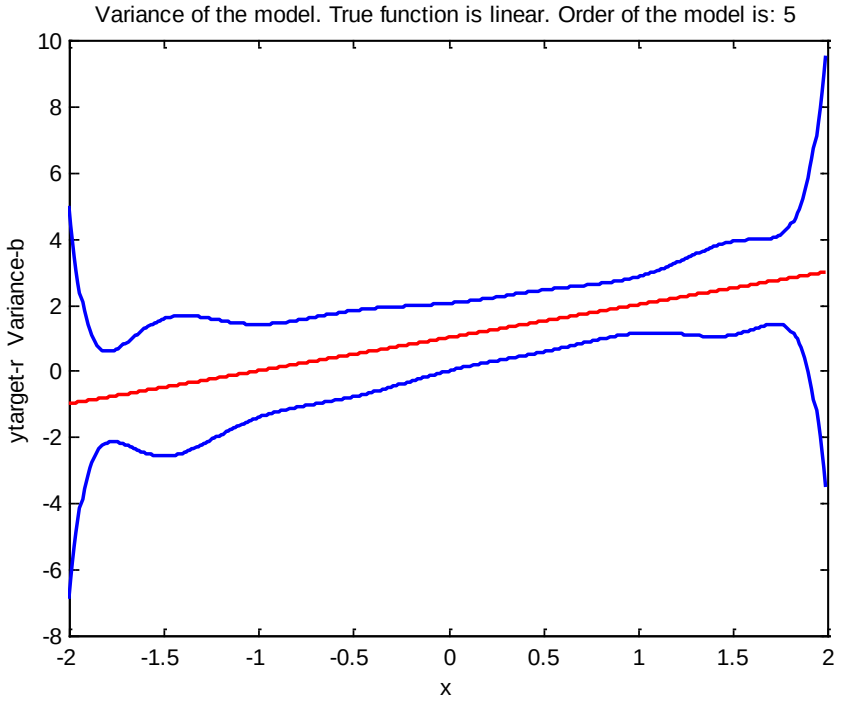
H = { all polynomial models of order >= 1 } also has **zero model bias**, but a **significantly larger model variance.**

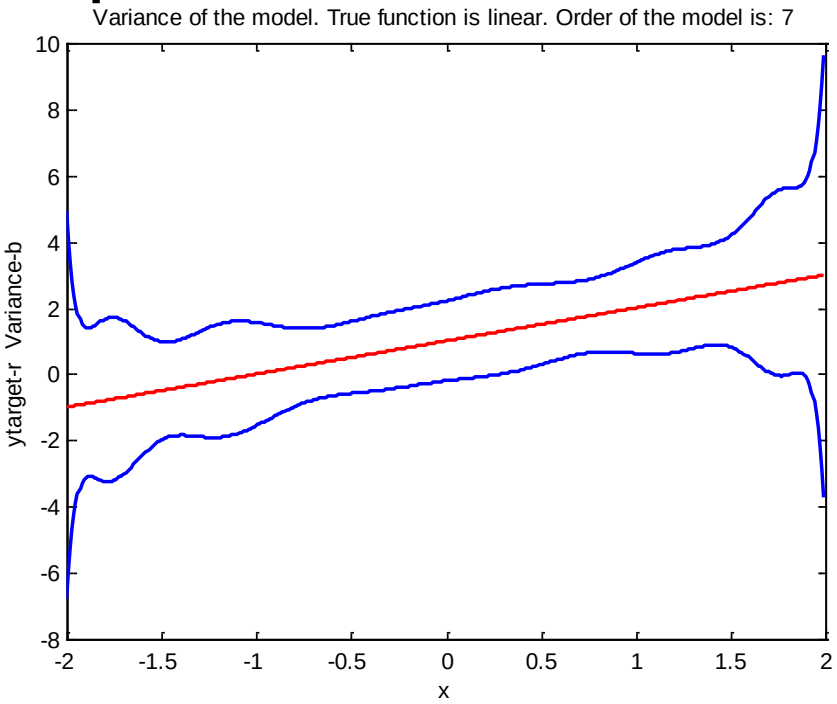
Ex 2:
Suppose that the **unknown** underlying function ***g*** **is cubic.**
H = { all linear models } has **a significant model bias** and **small model variance.**
H = { all polynomial models of order >= 1 } has **zero model bias** and **a larger model variance.**

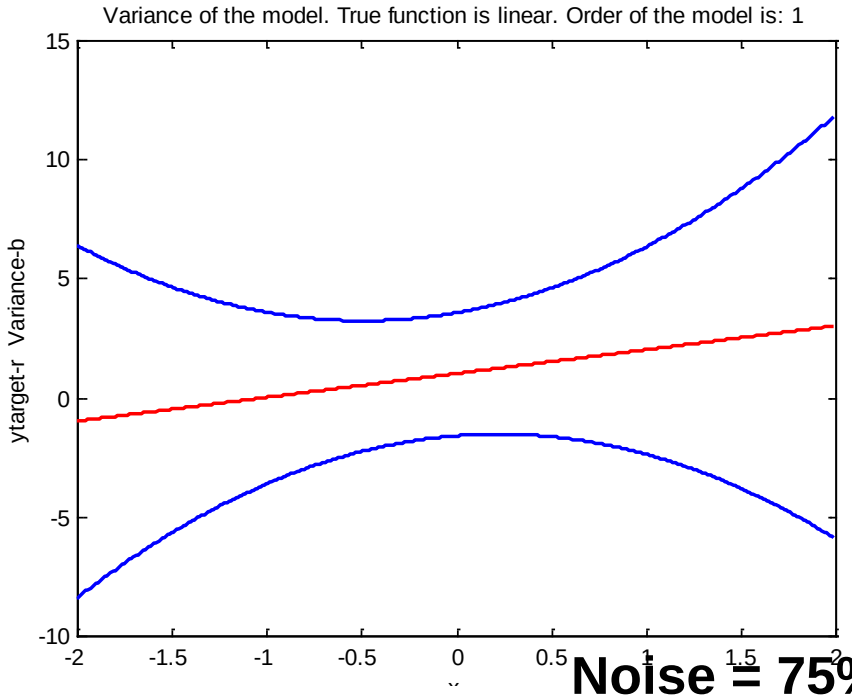


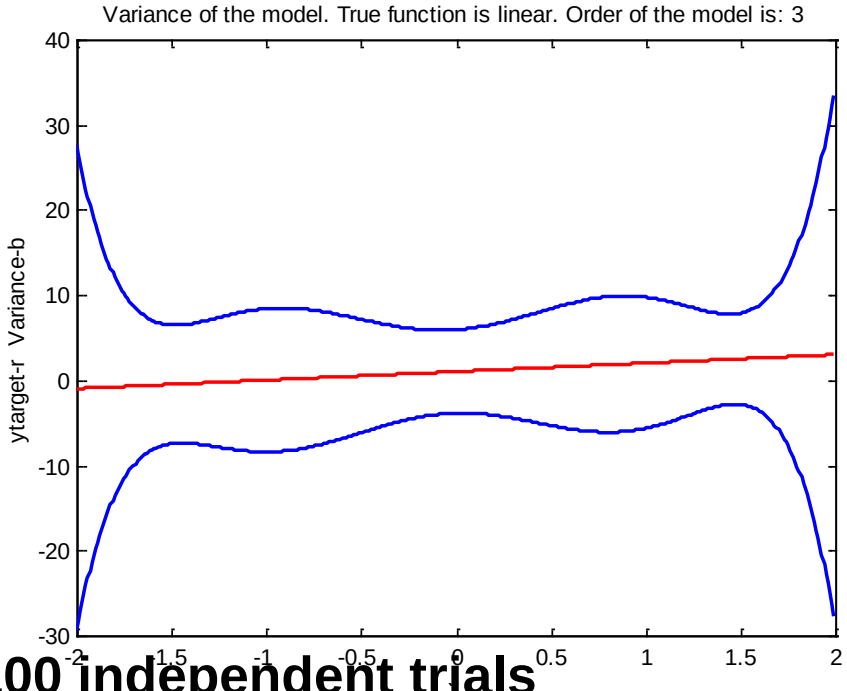


Noise = 25%, 100 independent trials

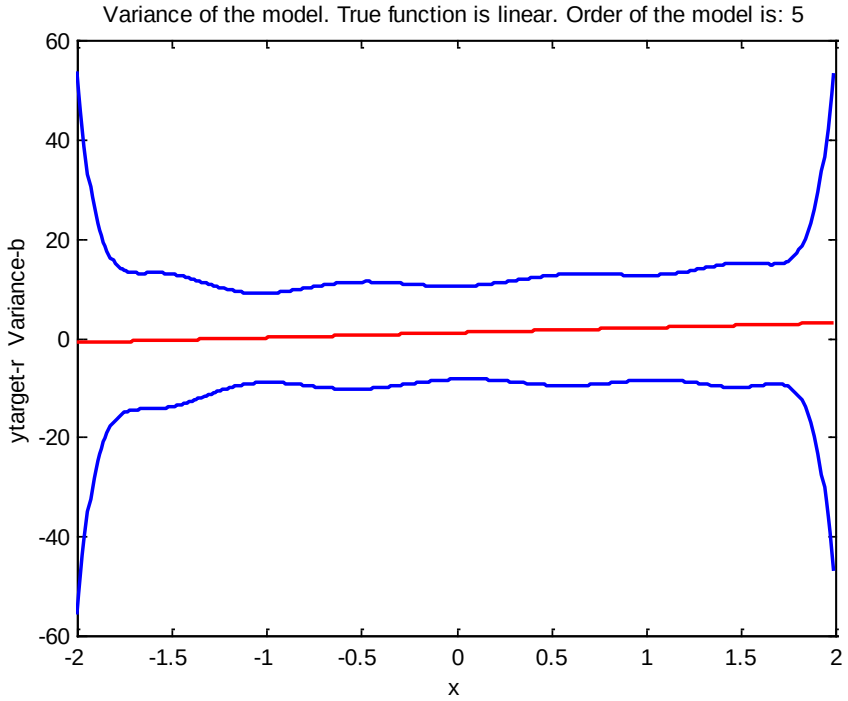


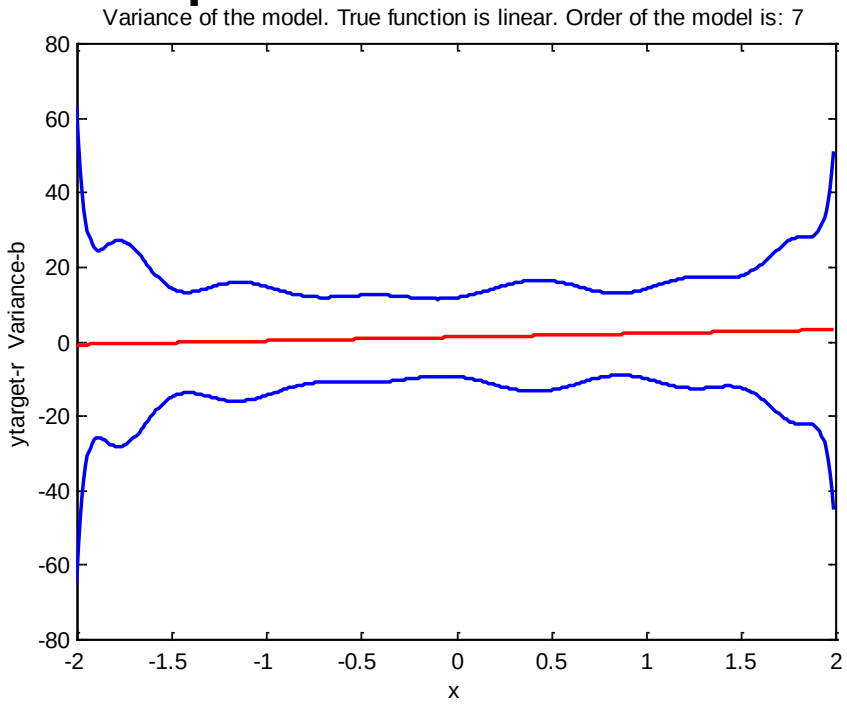


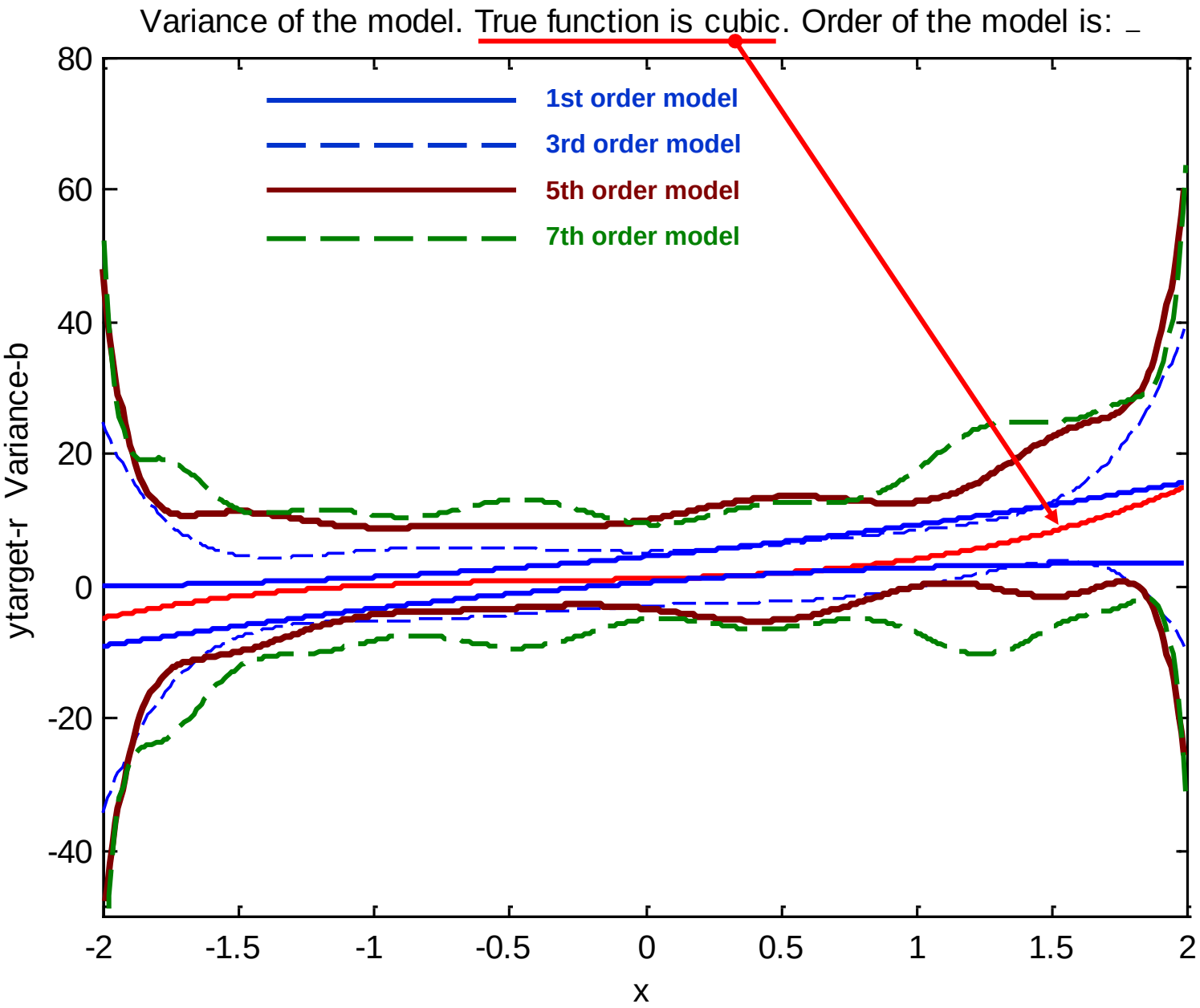




Noise = 75%, 100 independent trials







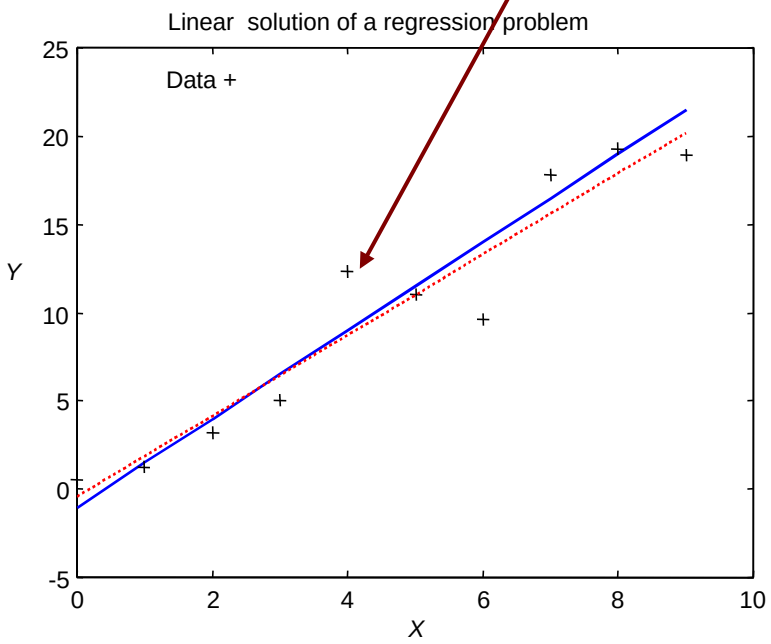
Noise = 25%, 100 independent trials

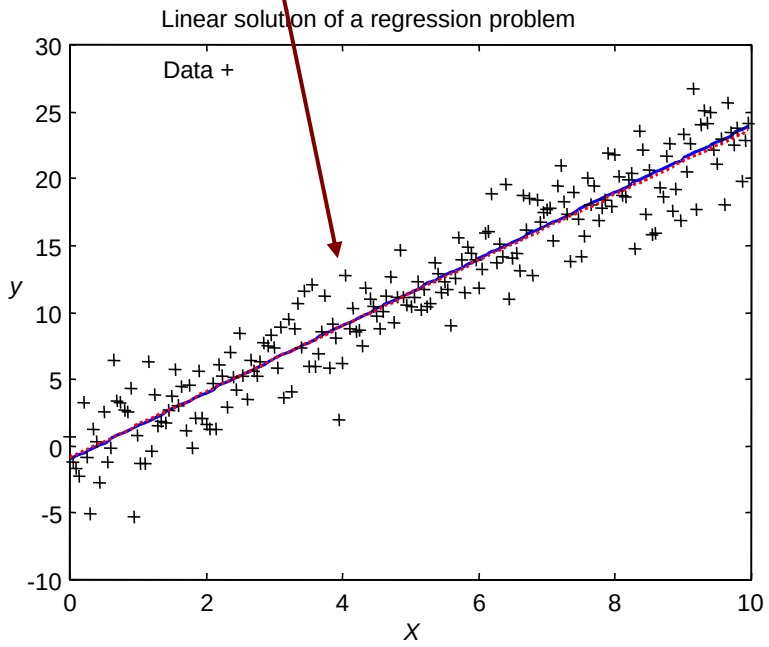
The previous three graphs were about the variance.

Now, the BIAS STORY!!!

Example: ***g* is linear,** hypothesis space **H** = {all linear models }, i.e., **unbiased estimator.** ***g*: *y* = 2.5*x* - 1 + *n***, $x \in [0,10]$, where *n* is a Gaussian random variable with a zero mean and such a variance that it corrupts the desired output *y* with 20% noise. **Left: 10 data pairs, Right: 200 data pairs**.

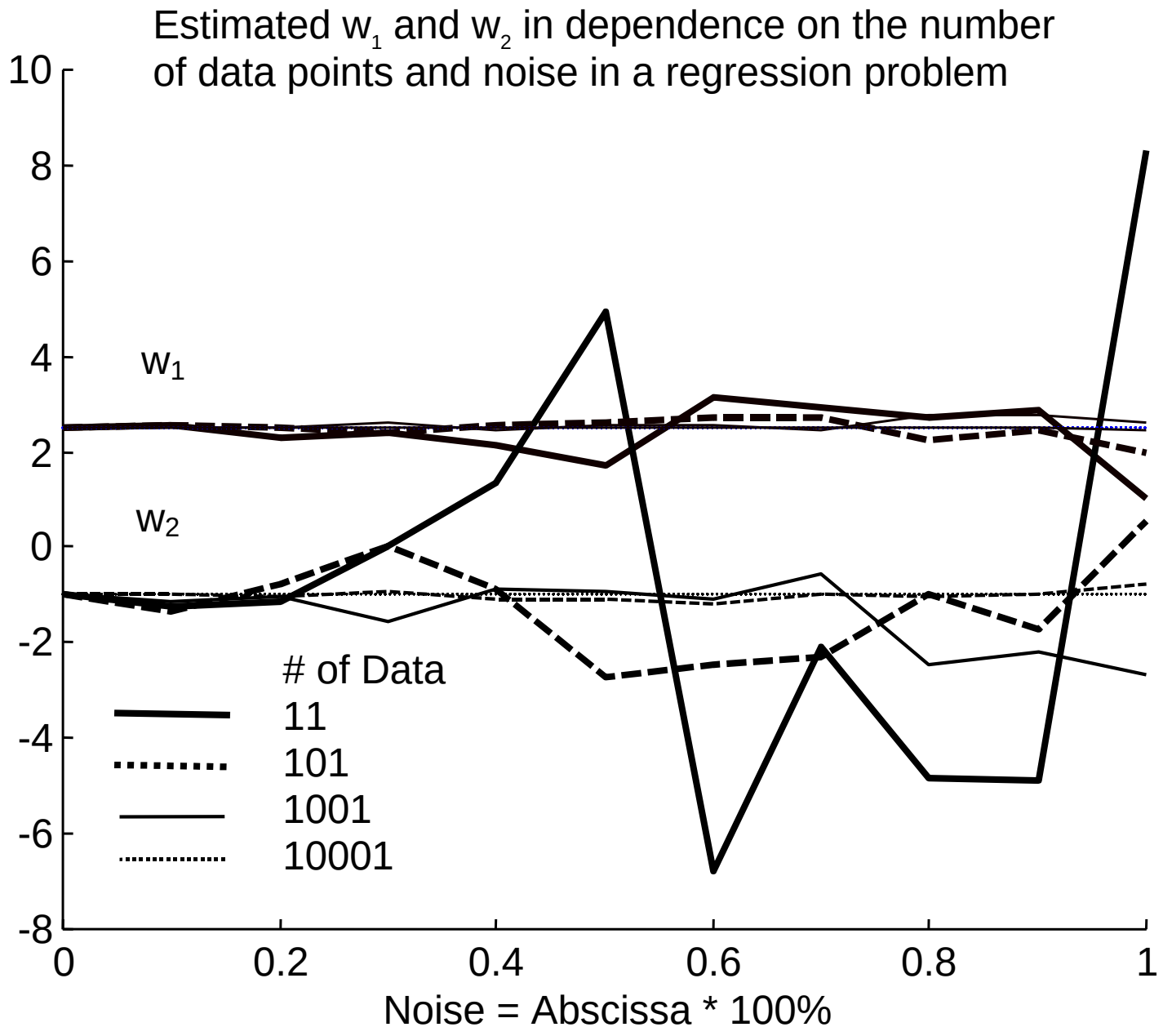
Blue line - true function, Red dashed line - regression line.





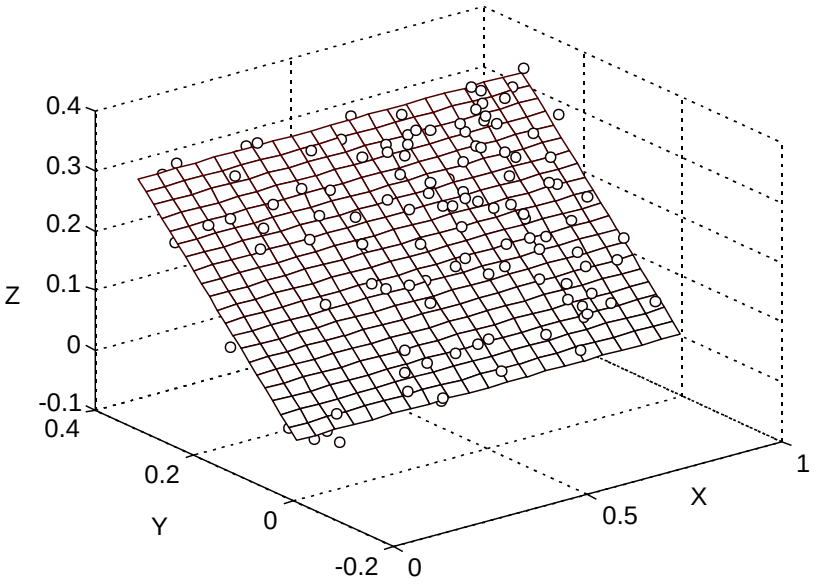
Even if **the target and hypothesis space coincide**, the errors in parameter estimation may be very high, depending heavily on the amount of training data and noise (uncertainty)

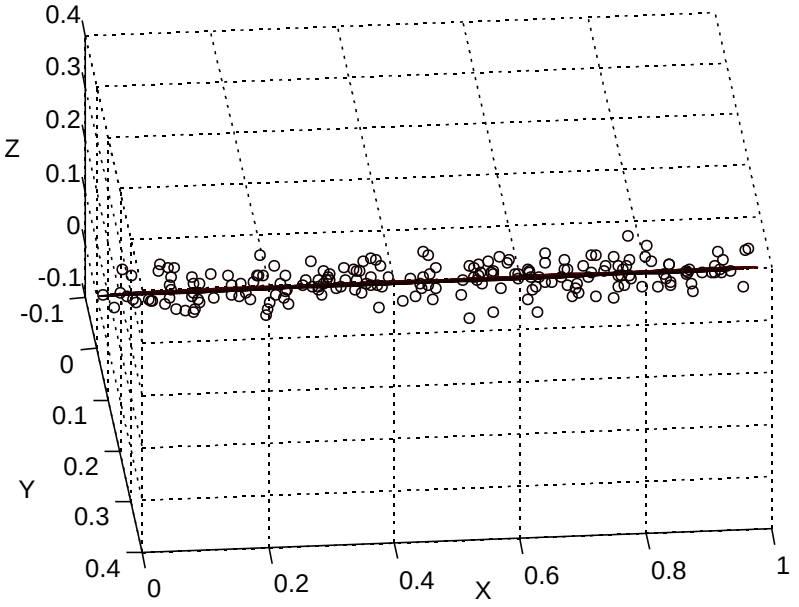
A visualization of the saying that *redundancy provides knowledge*



The basic task of the modeling problem does not change with the increase in the input space dimensionality. However, it becomes much more complicated - more data needed to 'fill the space', 'curse of dimensionality'

A rather simple example shows the result in a 'simple' case of linear both the target and the hypothesis function that is, therefore, an unbiased estimator.





Solution of the regression task for the underlying function $z = 0.004988x + 0.995y + n$ by using a training data set of 200 patterns. The picture on the right is a view along a plane. 20% Gaussian noise.