## What is Statistical Learning?

**Statistical Learning** refers to learning from data by making use of Statistics. It refers to the vast set of tools we use to understand data.

**What do we want to learn?**

* Patterns in the data.
* Relationships between variables.

**Why?**

* Inference: To understand and describe phenomena
* Prediction: To predict the future (and hence make informed decisions based on those predictions)

Most statistical learning problems can be divided into two categories: Supervised & Unsupervised Learning

**Supervised Learning:**

* Model learns from data under the supervision of a response variable.
* We need labelled data here.
* Examples: Regression tasks like predicting house prices, Classification tasks like spam email detection and Handwritten Digit Recognition, etc.
* GAMs, Random Forests, SVMs, NNs, etc.

**Unsupervised Learning:**

* Model is given a dataset, but there is no clarity on which variable is the response and which are the predictors. The model has to learn and recognise patterns & relationships between the data.
* Examples: Market Segmentation via Clustering, PCA, etc.

// Semi-Supervised Learning + Email Example

### Supervised Learning and Y = f(X) +

// Sales Data and Income Data

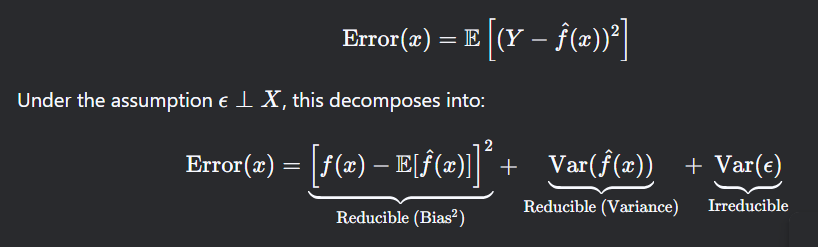
Note that we **assume** that is a zero-mean random error and that **it is independent of X.**

What does even signify?

It signifies:

* Any unmeasured variables
* Any random “chance” error or unmeasured variation
* Any measurement error in the measured variables

This assumption is what separates the error into reducible and irreducible error as follows:

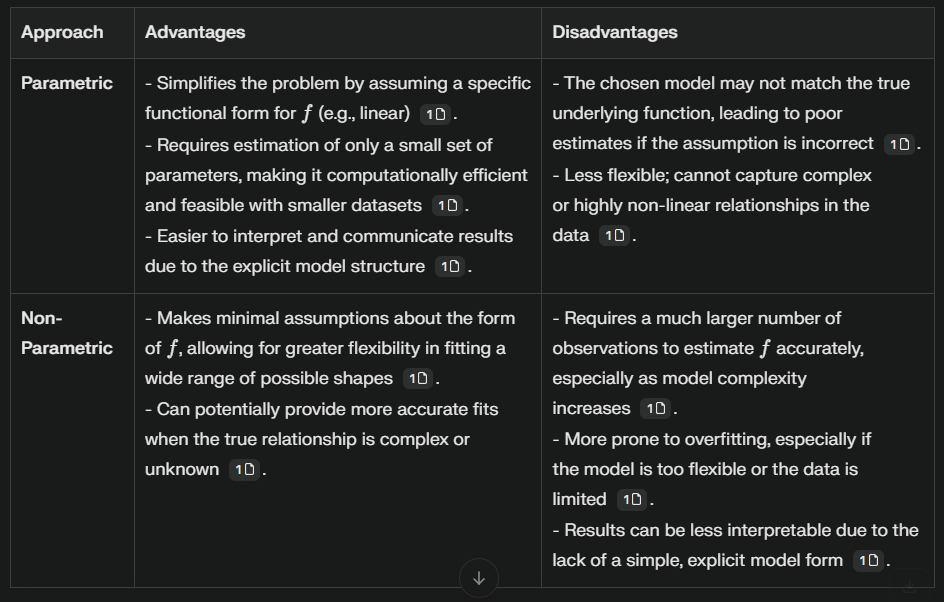


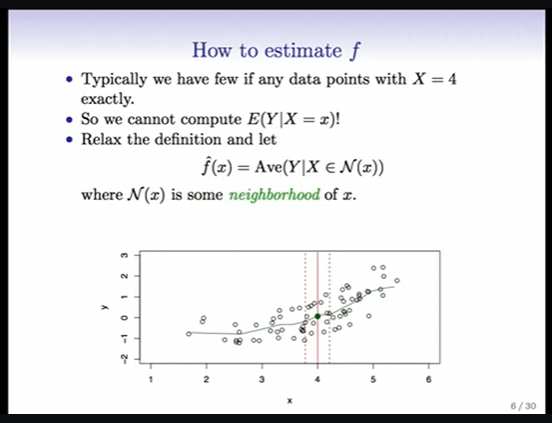
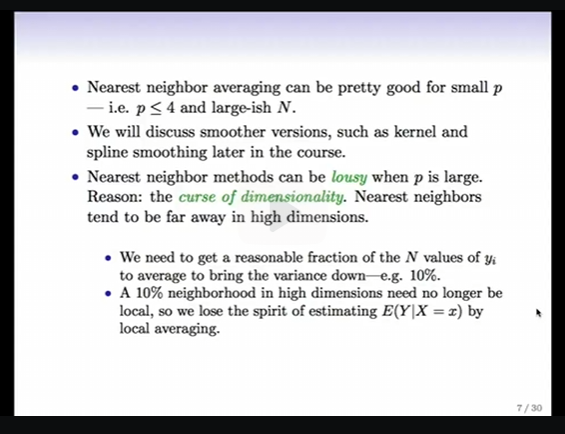
**Why Estimate f?**

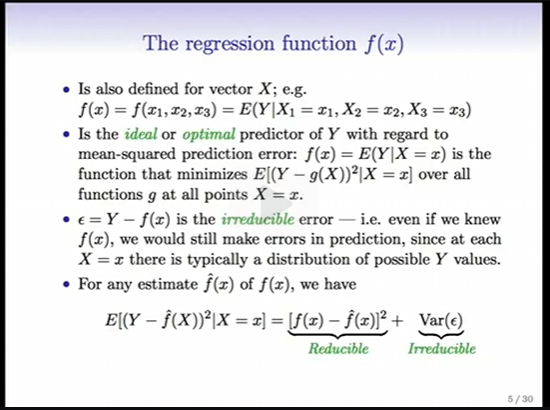
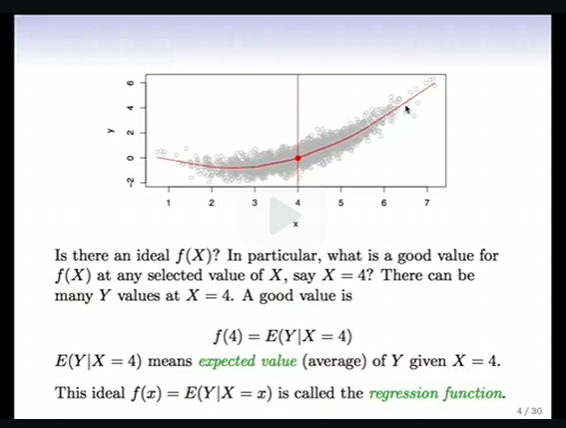
* **Prediction:** In this setting, we want to estimate f to get the best possible prediction results. Here, f is treated as a black-box and we’re not really interested in what it looks like.
* **Inference:** here, we look to infer from our model. We wish to explain the following for example:
  + Which predictors are associated with the response? // Income Example
  + What is the relationship between the response and each predictor?
  + Can the relationship between Y and each predictor be adequately summarized using a linear equation, or is the relationship more complicated?
* // Housing Example and (prediction accuracy vs interpretability)

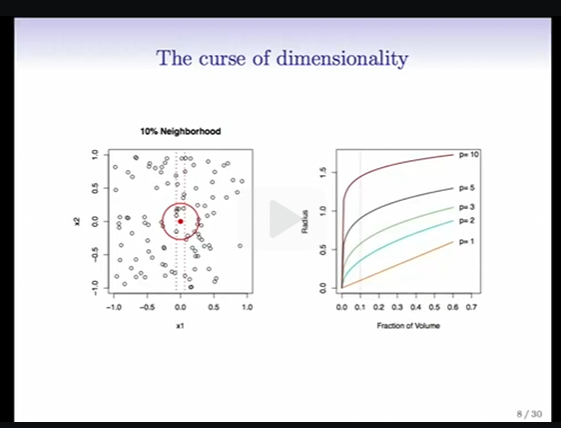
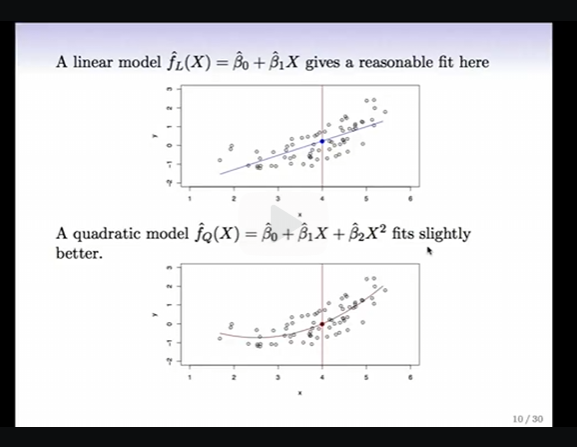
**How do we estimate f?**

* Given some training data, we’d like to estimate f, keeping in mind whether we want better prediction or inference. Broadly, we have two methods: Parametric and Non-parametric.
* **Parametric:** We reduce the problem of estimating f to estimating a set of parameters by assuming the structure of f.
  + **Step 1:** Make an assumption about the functional form of f. For example, Linear Model.
  + **Step 2:** Select a procedure to fit the model. For example, OLS, L1, L2 in linear models.
* **Non-Parametric:** No functional form of f is assumed. Example: smooth splines

**Disadvantages and Advantages in each method:**

**What is a good f?**

****

****

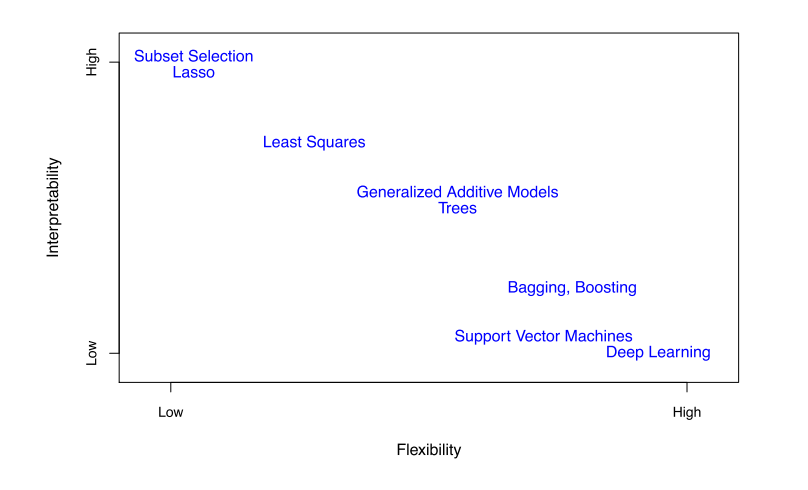
Curse of Dimensionality: <https://youtu.be/9Tf-_mJhOkU?feature=shared>

Link to the lectures: <https://learning.edx.org/course/course-v1:StanfordOnline+STATSX0001+2T2023/home>

The book Elements of Statistical Learning also has more on this.

Hence, often, we use parametric methods and give an initial structure to f to avoid encountering curse of dimensionality.

**Prediction Accuracy vs Model Interpretability:**

* There is an inverse proportionality sort of between flexibility and interpretability.
* Less flexible methods are restrictive in the sense that they shorten the range of shapes of f. For example, Linear Models assume f to be linear but that is seldom the case. On the other hand, less flexible models are quite interpretable.
* 

So, when inference is our goal, selecting lesser flexible models makes sense, since they have more interpretability.

When prediction is our goal, we might think that we should use the highest flexible model in order to have the best possible accuracy but this is not the case. This is because highly flexible models tend to overfit the training data and hence perform poorly on test data. This comes up in 2.2

We choose models based on whether we’re working with QL or QN variables. But we can always code the QL variables to be QN and use them as such.

**Variable Types**

* **Quantitative (QN):** Numerical (e.g., income, temperature).
* **Qualitative (QL):** Categorical (e.g., gender, color).
  + Can be **unordered** (red, blue) or **ordered** (low, medium, high).
  + Also called discrete variables or factors
* **Methods depend on variable type:**
  + Regression → QN response.
  + Classification → QL response.
  + Some methods (e.g., KNN, Boosting) work for both.
* We choose models based on whether we’re working with QL or QN variables. But we can always code the QL variables to be QN and use them as such.