Splines

- GLM assumption: $g(E[Y|X]) = X\beta$, is a linear function of X
- This is probably almost never the case
- Why ever assume linearity?
  - Simplicity
  - Interpretability
  - Minimal risk of overfitting
  - First-order Taylor approximation of true $E[Y|X] = f(X)$

In [1]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import statsmodels.api as sm
from patsy import dmatrix
import seaborn as sns
%matplotlib inline
```

In [2]:
```
# true regression function
def f(x):
    return(2*np.square(x)-50)

n = 500 # sample size
x = np.sort(np.random.uniform(low=0, high=10, size=n))
y = f(x)+np.random.normal(0, 10, n)
X = sm.add_constant(x)
model = sm.OLS(y, X)
results = model.fit()
results.summary()
```

Out[2]:

OLS Regression Results

Dep. Variable:	y	**R-squared:**	0.914
Model:	OLS	**Adj. R-squared:**	0.914
Method:	Least Squares	**F-statistic:**	5313.
Date:	Thu, 04 Mar 2021	**Prob (F-statistic):**	7.72e-268
Time:	16:14:57	**Log-Likelihood:**	-2158.9
No. Observations:	500	**AIC:**	4322.
Df Residuals:	498	**BIC:**	4330.