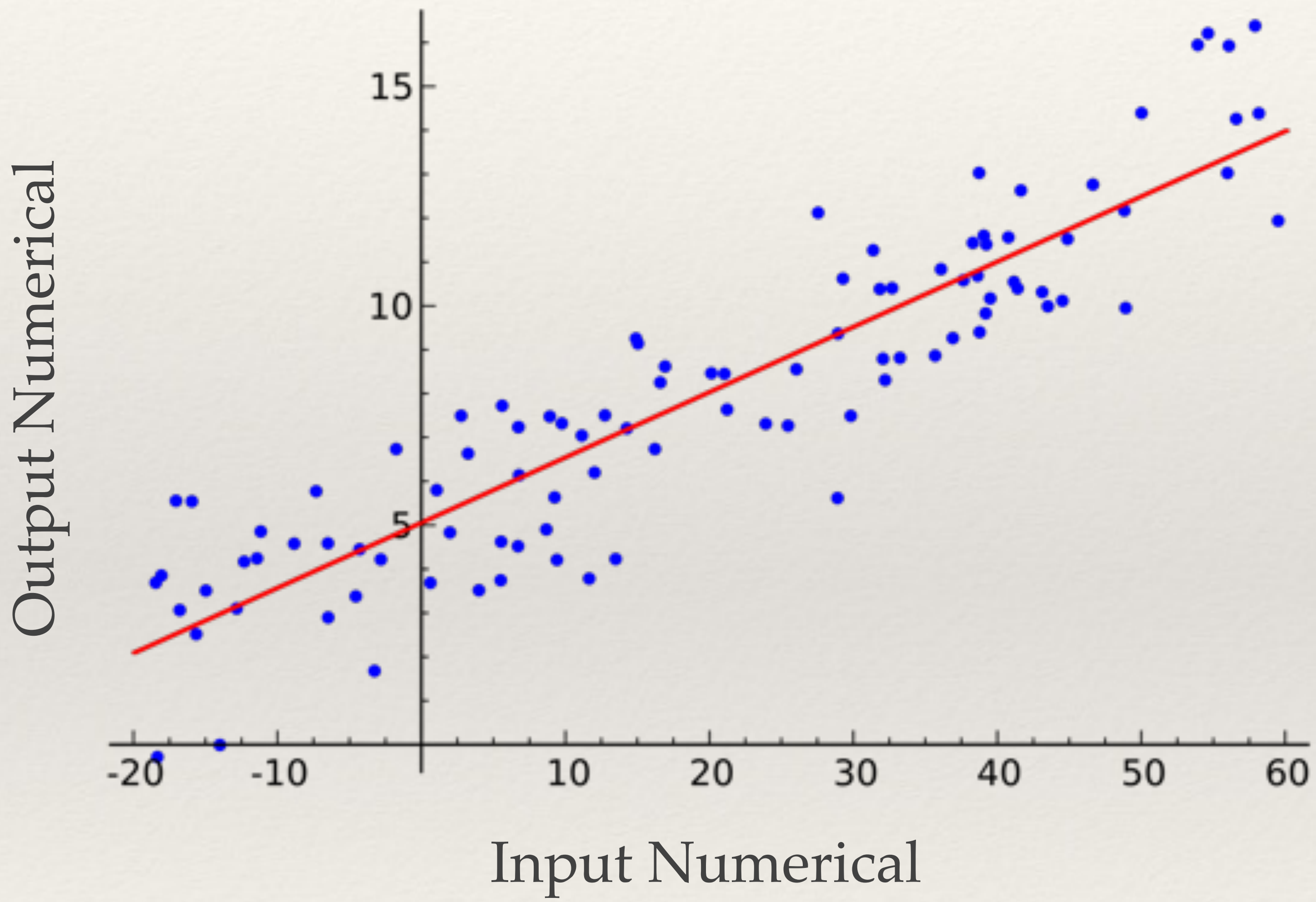
Michael Grossberg

Intro to Data Science CS59969

Regression

Regression

Regression



$$y_i = \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i = \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta} + \varepsilon_i, \qquad i = 1, \ldots, n,$$

Recall matrix formulation

$$y_i = \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i = \mathbf{x}_i^{\mathsf T}\boldsymbol\beta + \varepsilon_i, \qquad i = 1, \ldots, n,$$

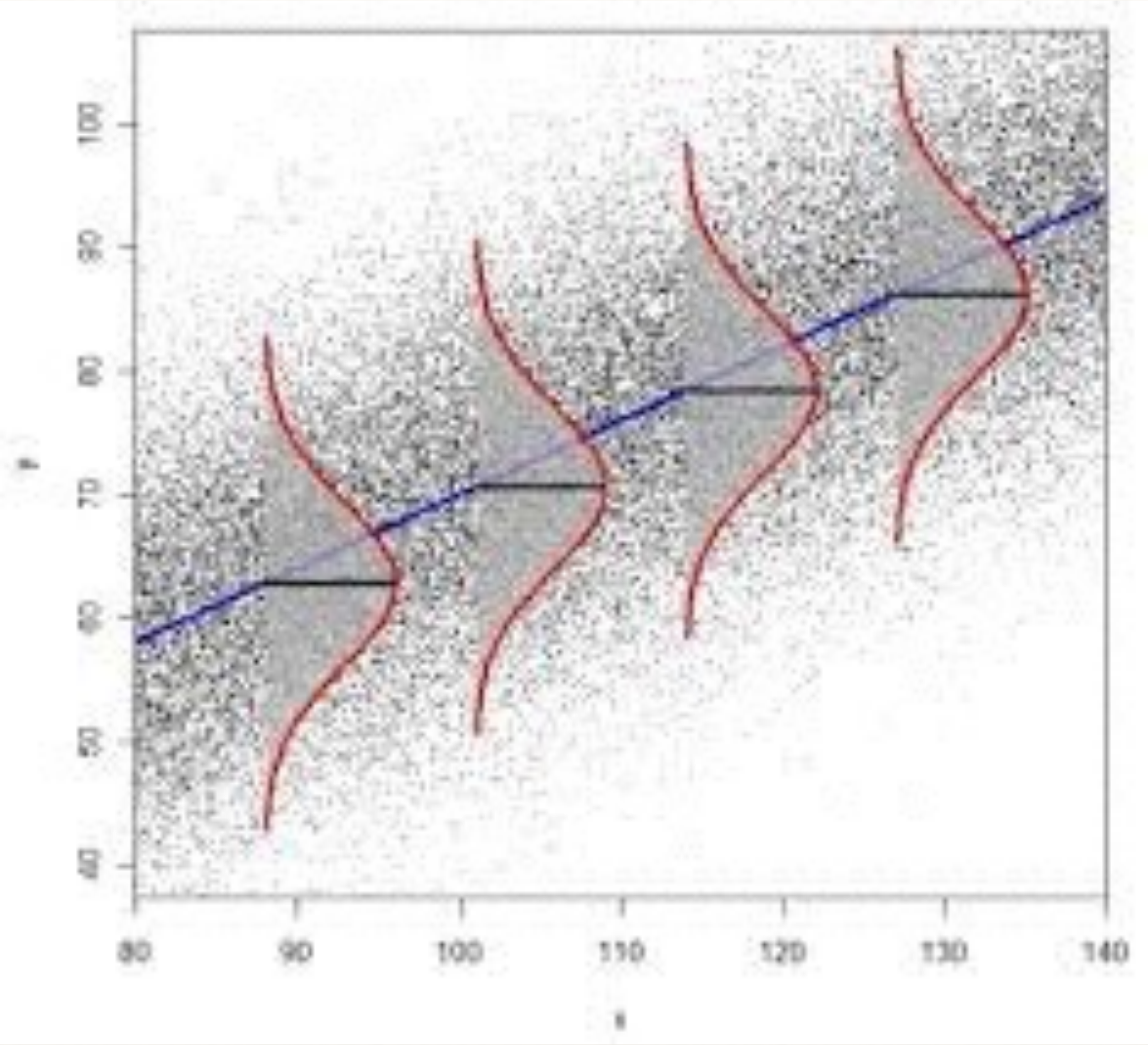
$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \quad \mathbf{X} = \begin{pmatrix} \mathbf{x}_1^{\mathsf T} \\ \mathbf{x}_2^{\mathsf T} \\ \vdots \\ \mathbf{x}_n^{\mathsf T} \end{pmatrix} = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ x_{21} & \cdots & x_{2p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix} \quad \boldsymbol\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix} \quad \boldsymbol\varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

$$\mathbf{y} = \mathbf{X}\boldsymbol\beta + \boldsymbol\varepsilon$$

Ordinary Least Squares Solution

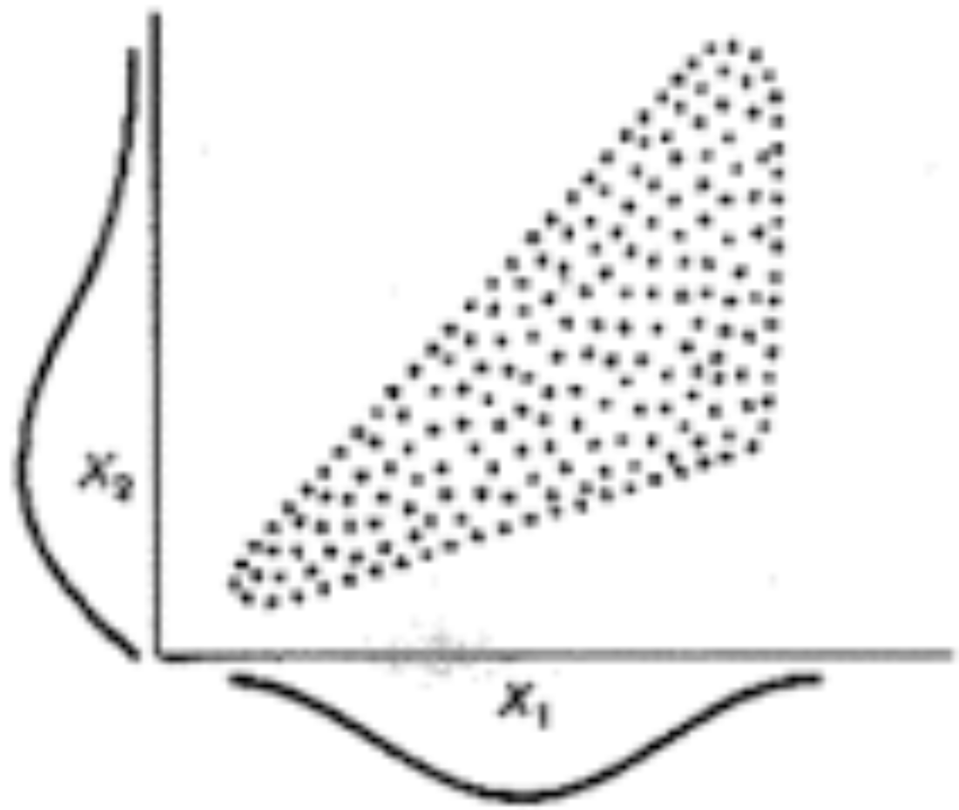
$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y} = \big(\sum \mathbf{x}_i \mathbf{x}_i^{\mathrm{T}}\big)^{-1}\big(\sum \mathbf{x}_i y_i\big)$$

OLS assumes Homoscedasticity



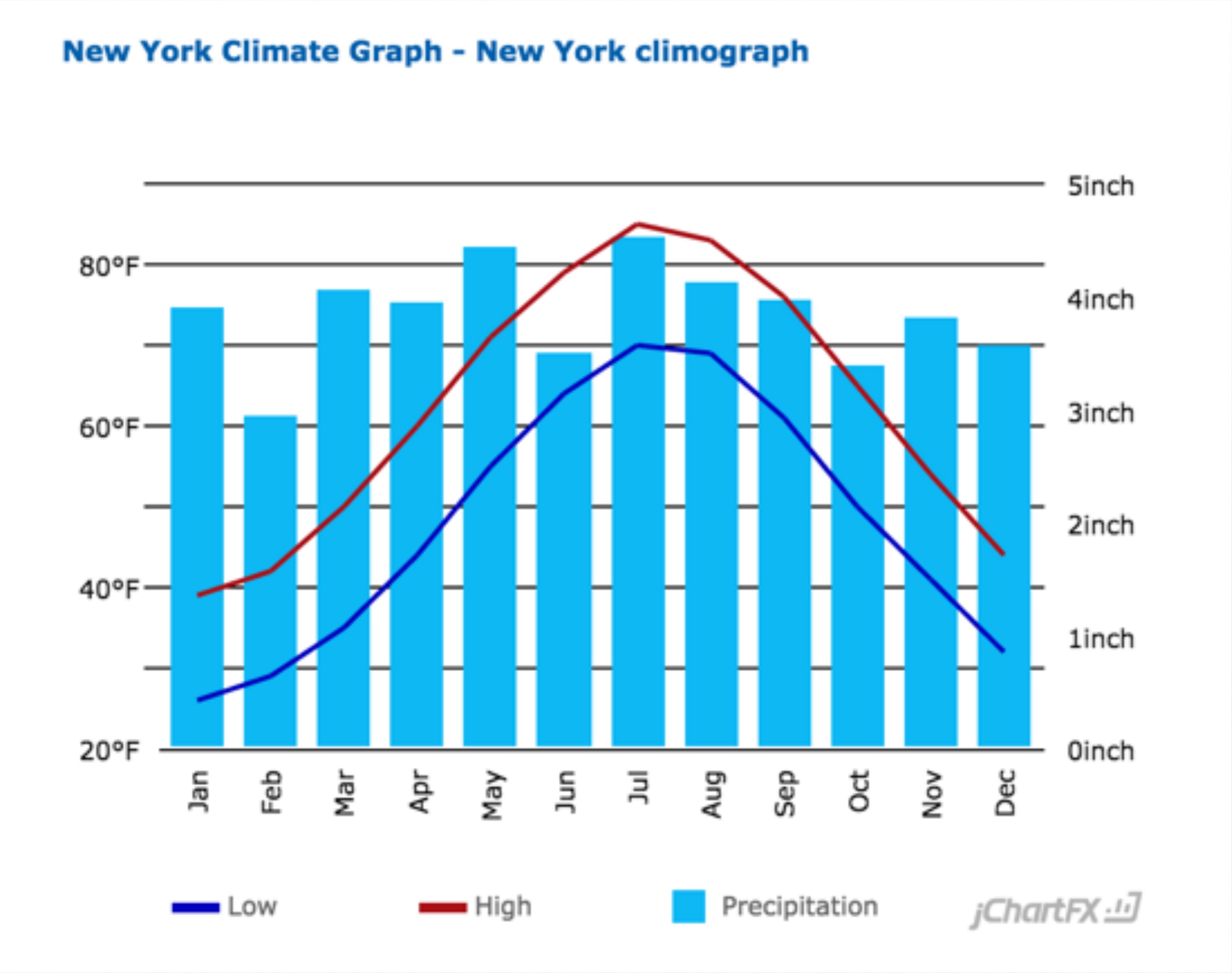
Worth checking

Some Data not Heteroscastic



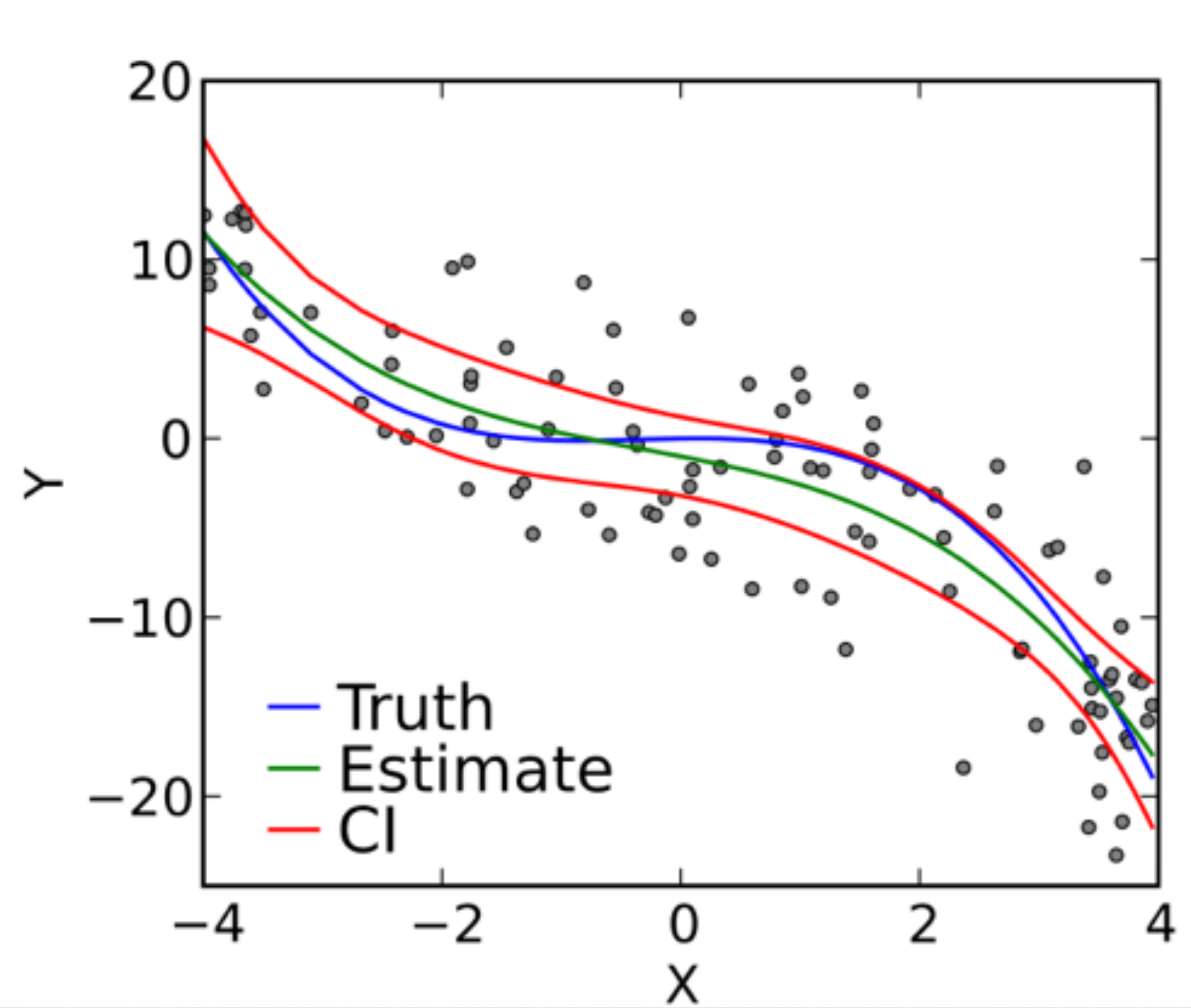
Heteroscedasticity with skewness on one variable
on one variable

What about when the data is non-linear?



http://www.usclimatedata.com/climate/new-york/united-states/3202

Polynomial Regression (still linear)



X not linear in Y

Linear Sum of non-linear functions

$$y = a_0 + a_1 x + a_2 x^2 + a_3 x^3 + \cdots + a_n x^n + \varepsilon.$$

Linear Matrix Formulation

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 & x_1^2 & \dots & x_1^m \\ 1 & x_2 & x_2^2 & \dots & x_2^m \\ 1 & x_3 & x_3^2 & \dots & x_3^m \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_n & x_n^2 & \dots & x_n^m \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ \vdots \\ a_m \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

$$\vec{y} = \mathbf{X}\vec{a} + \vec{\varepsilon}$$

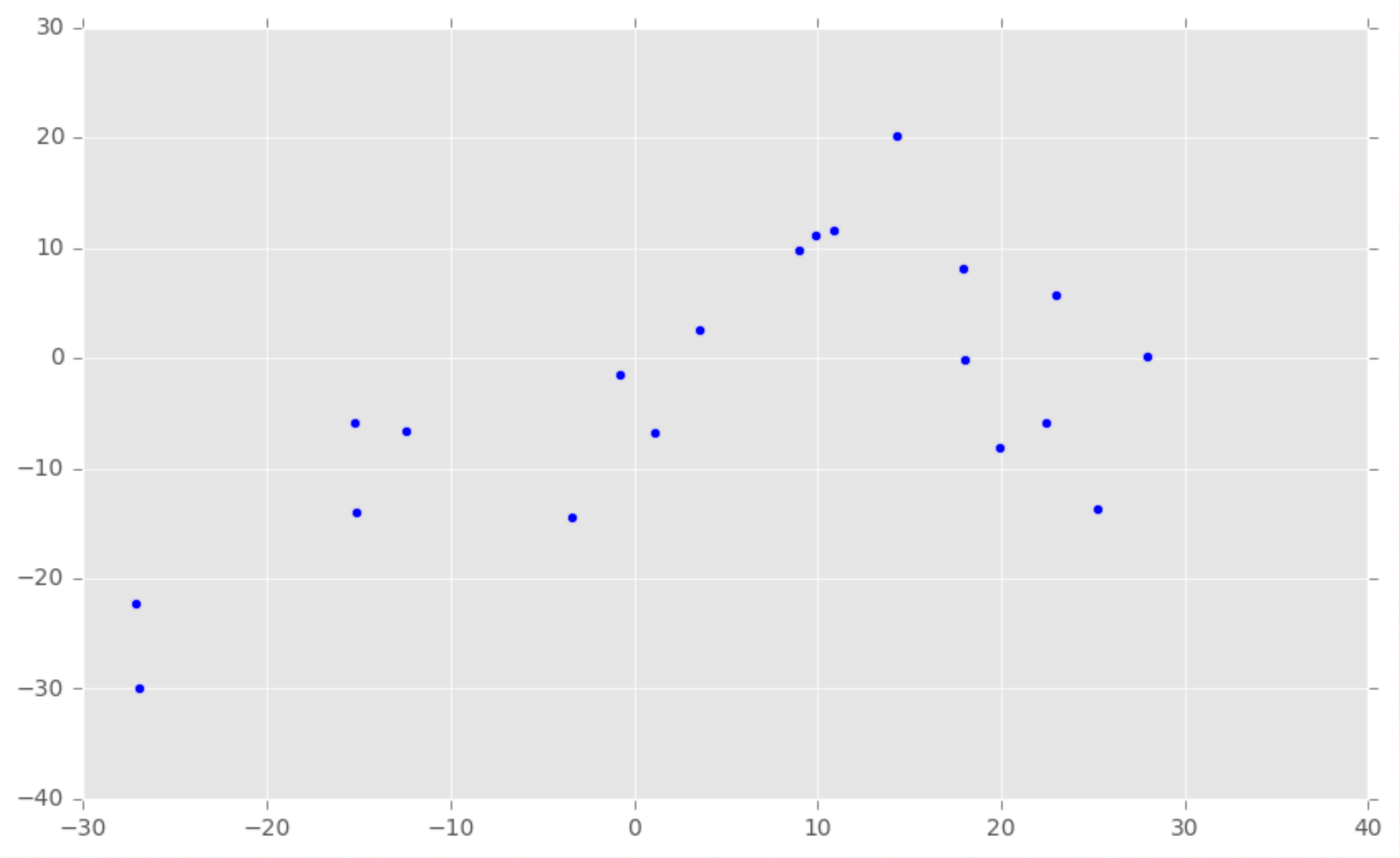
Same as before! Linear in the a vector (=beta from before)

$$\hat{\vec{a}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1} \mathbf{X}^T\vec{y}.$$

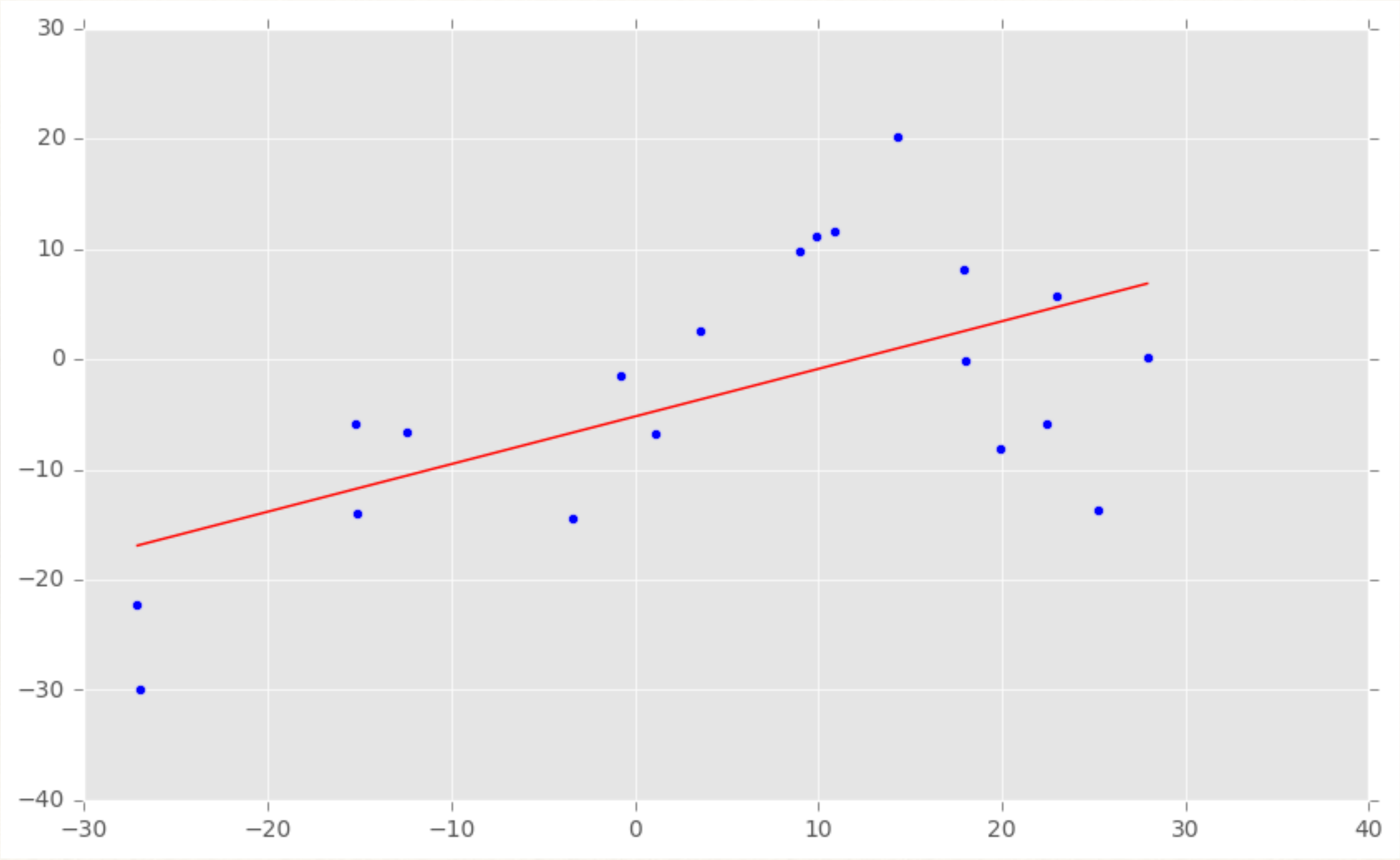
Could be sum of sin/cos or anything

$$\frac{A_0}{2} + \sum_{n=1}^{\infty} (A_n \cos nx + B_n \sin nx).$$

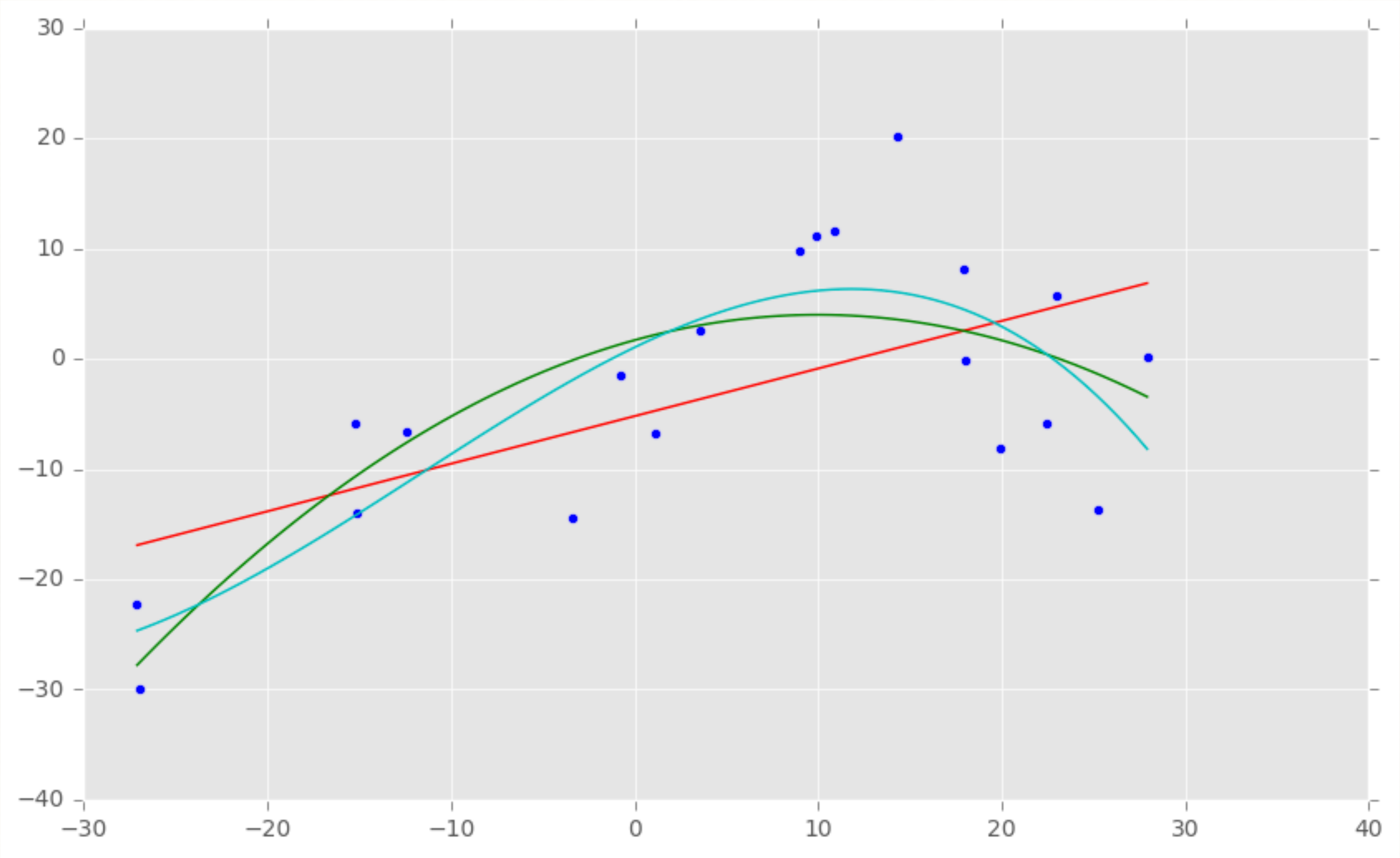
What model to fit?



Linear?

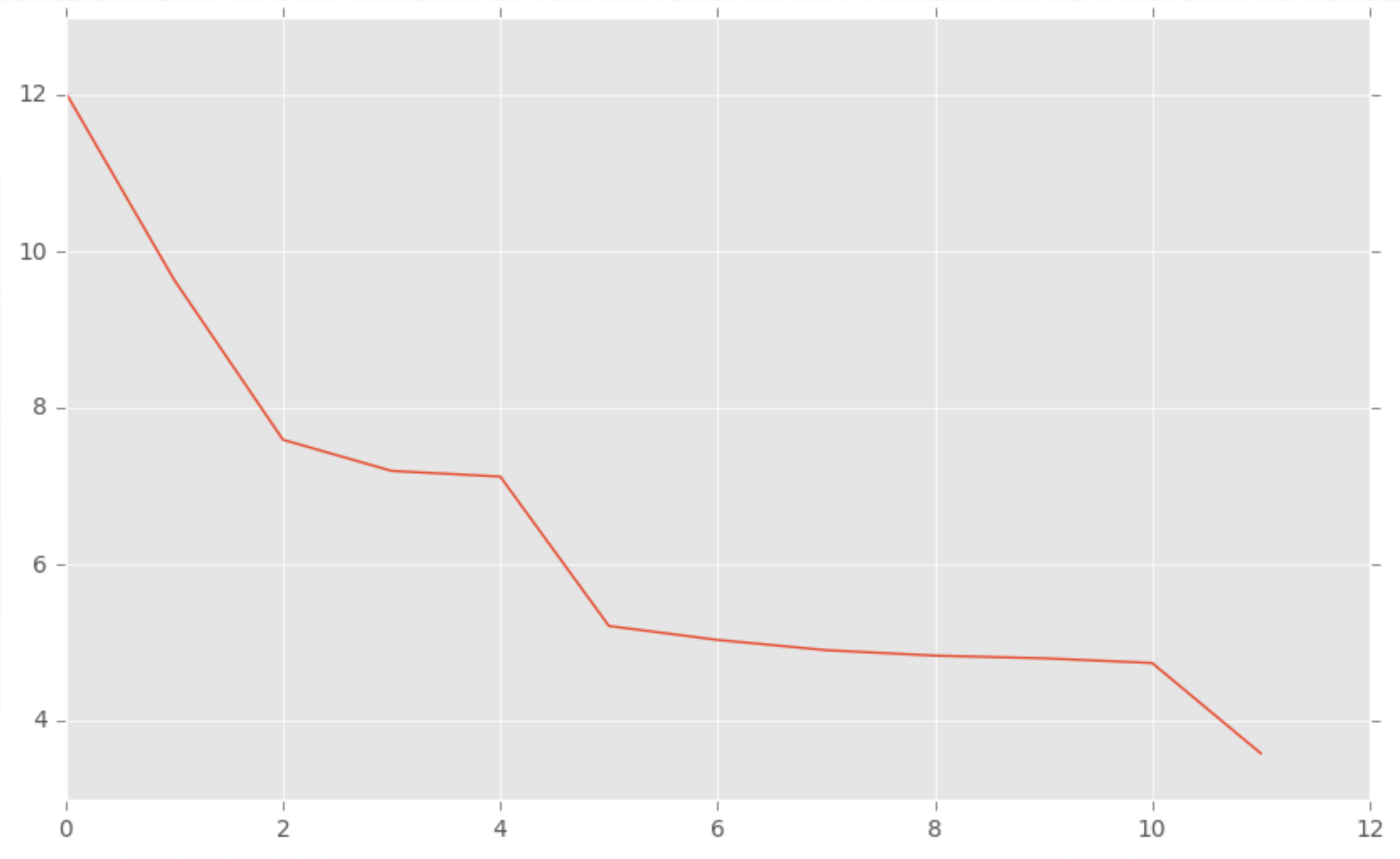


Quadratic? Cubic?

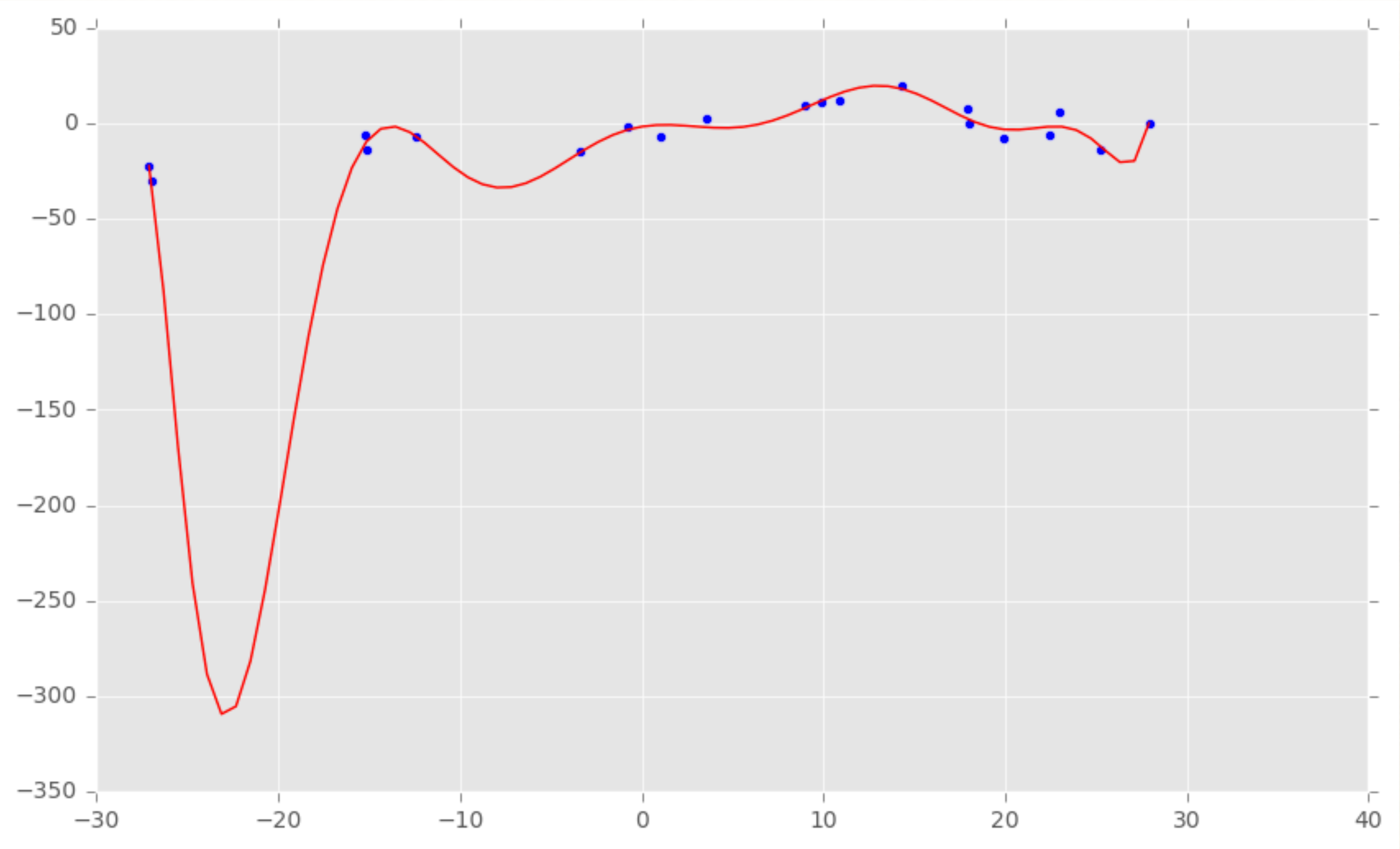


Error Keeps Dropping

Error Will Always Keep Dropping

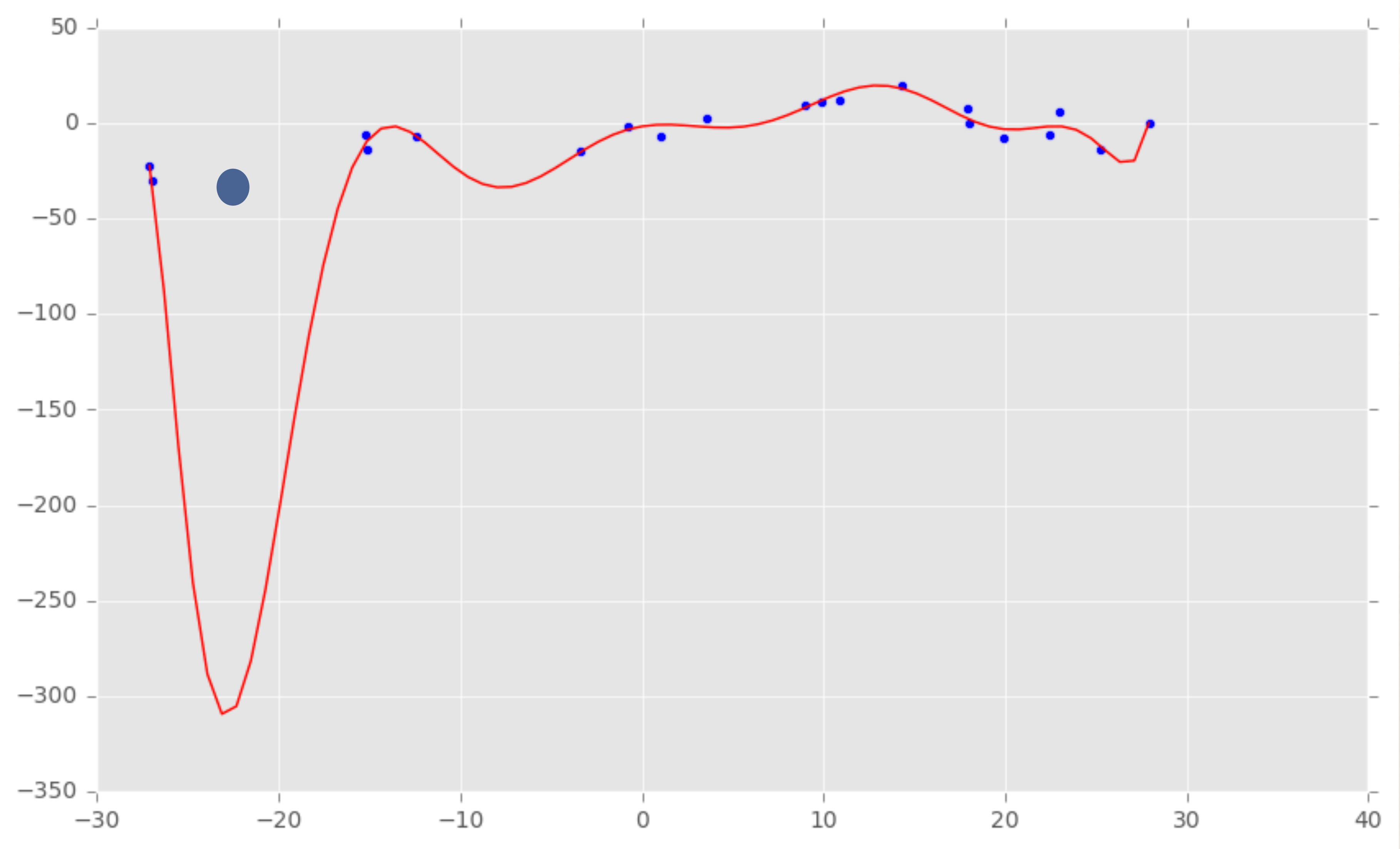


High Degree Fit Doesn't Make Sense



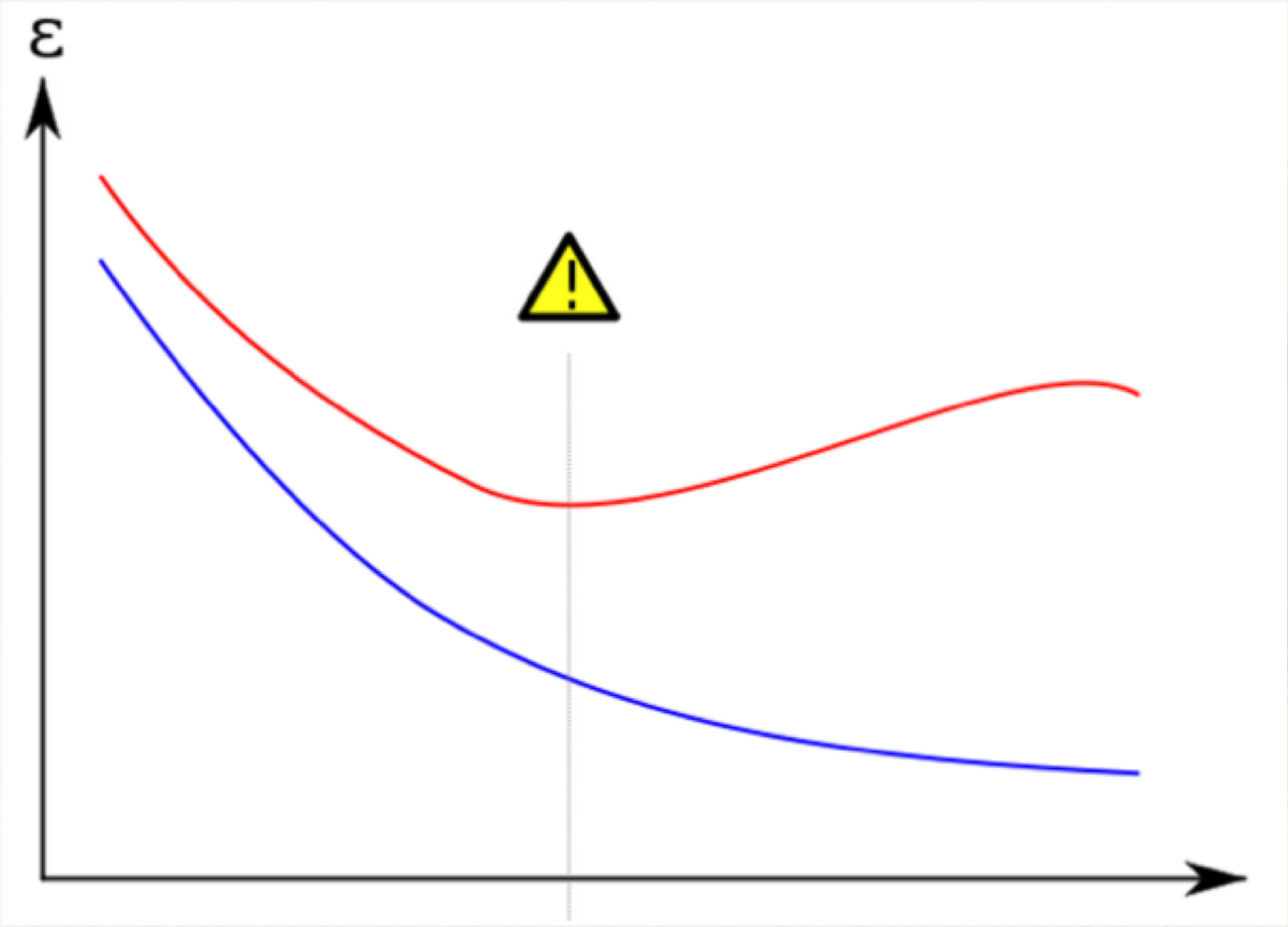
11th Degree Fit

Bad Prediction at New Data



Called Generalization Error

Overfitting: Model too Complex



Generalization Error
aka
Testing Error

Fitting Error
aka
Training Error

Analytic Solutions to Overfitting

Analytical Methods:

Akaike information criterion (AIC)

Degree

$$\mathrm{AIC} = 2k - 2\ln(L)$$

Max of Likelihood of model

Bayesian information criterion (BIC)

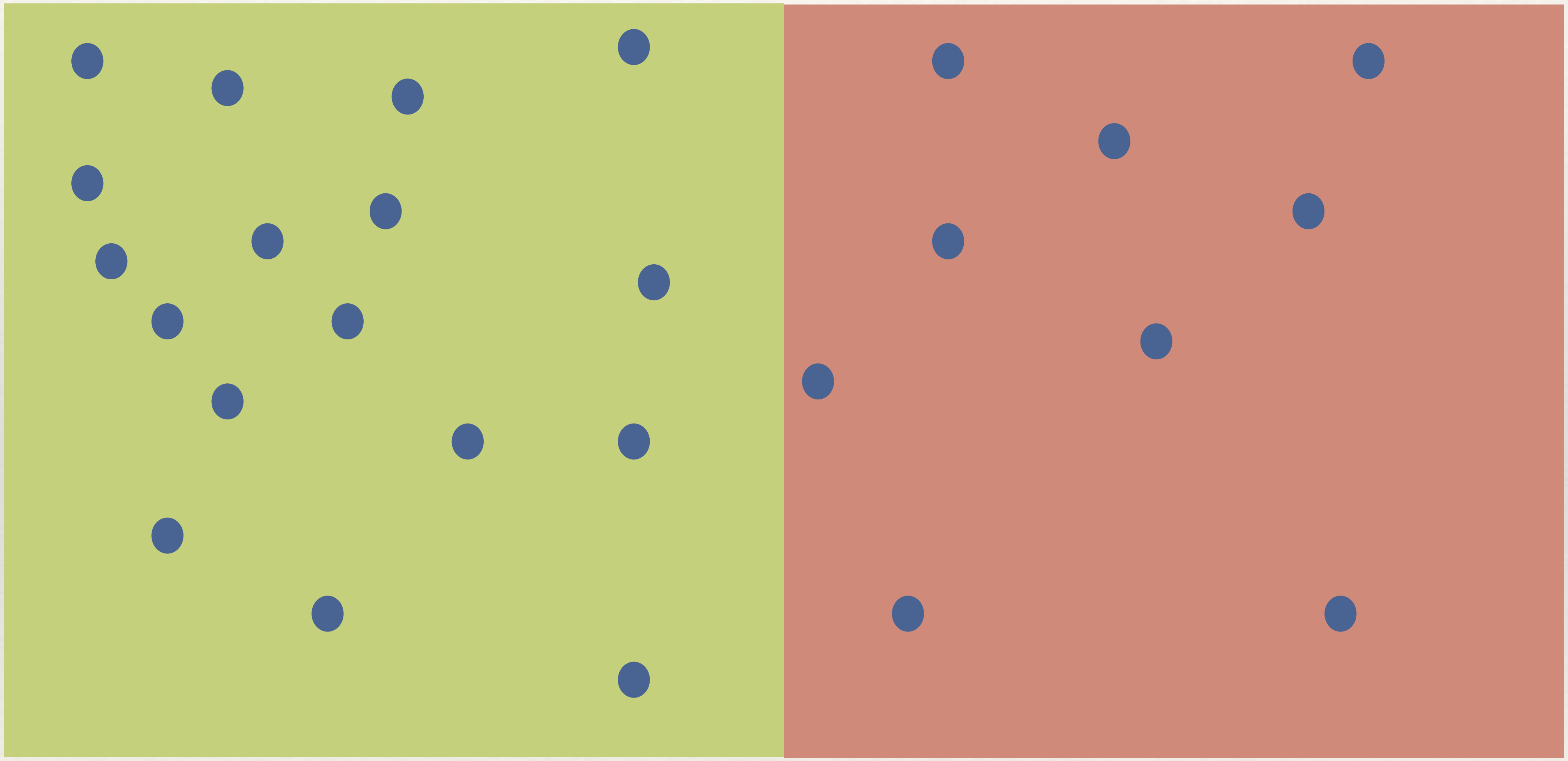
$$\mathrm{BIC} = -2 \cdot \ln \hat{L} + k \cdot \ln(n).$$

Others: Minimum Description Length (MDL)
VC Dimension

Empirical Solutions to Overfitting

Fit/Train with Some Data

Eval/Test with Separate Data



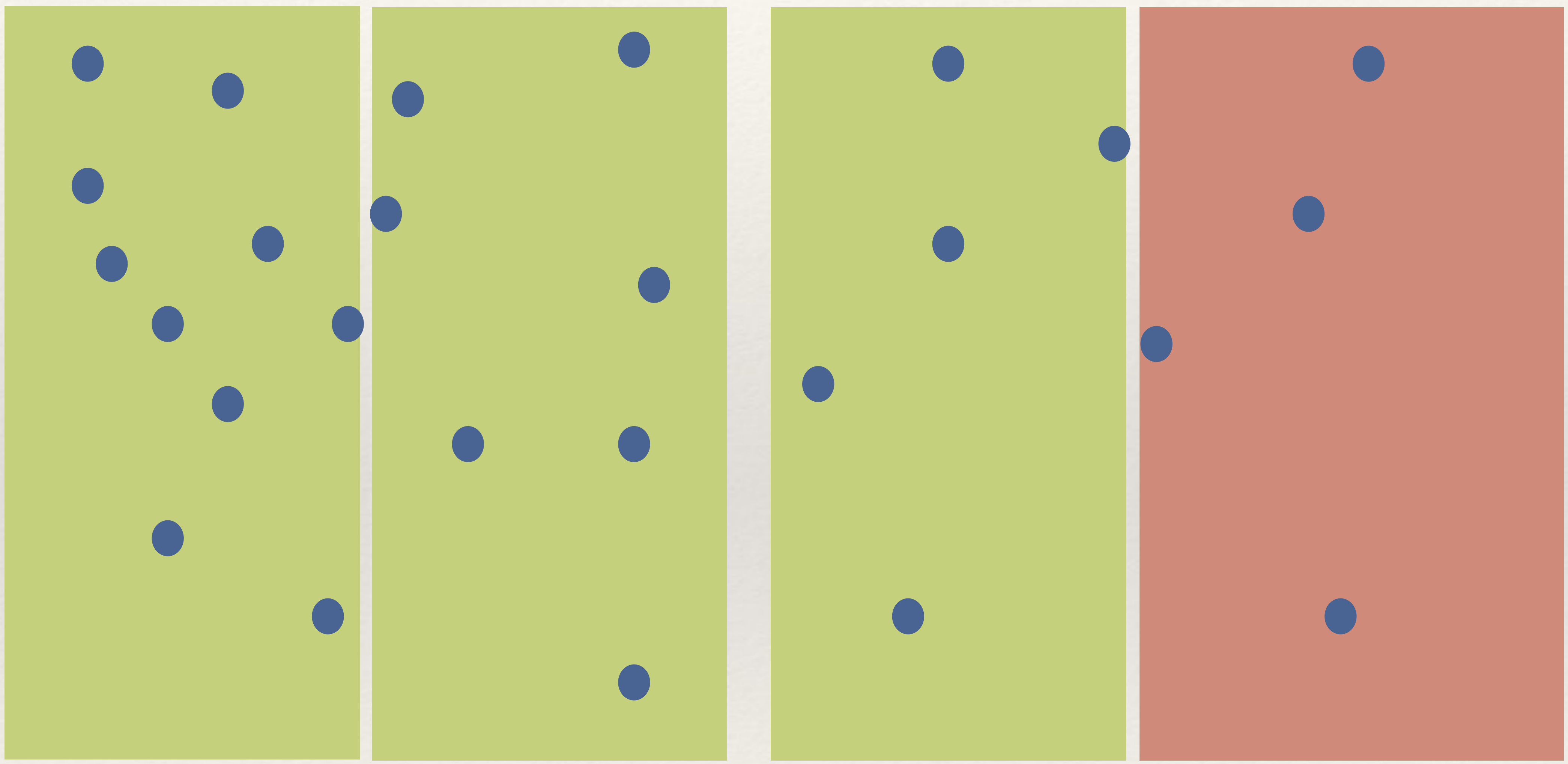
Best solution when you have lots of data!

Make sure Training/Testing representative of task:

Interpolation: Random Subsamples, Extrapolation: Past vs. "Future"

Cross Validation

Cut into k (4? 10?) Chunks



Fit on k-1

Evaluate on 1

Shuffle K and 1

Mean (over K) of Eval Error is "error" and std is "uncertainty in error estimate"