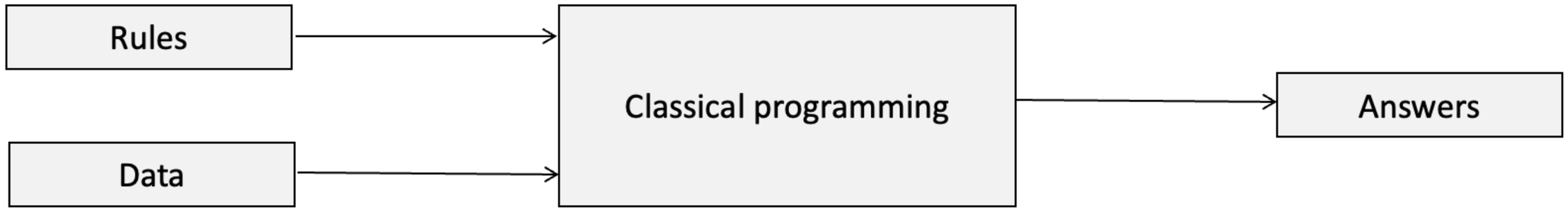
Lecture 2

- Reframing the Question of Conventional Programming vs ML
- Statistical Learning
- Introduction to Neural Networks

September 12, 2024

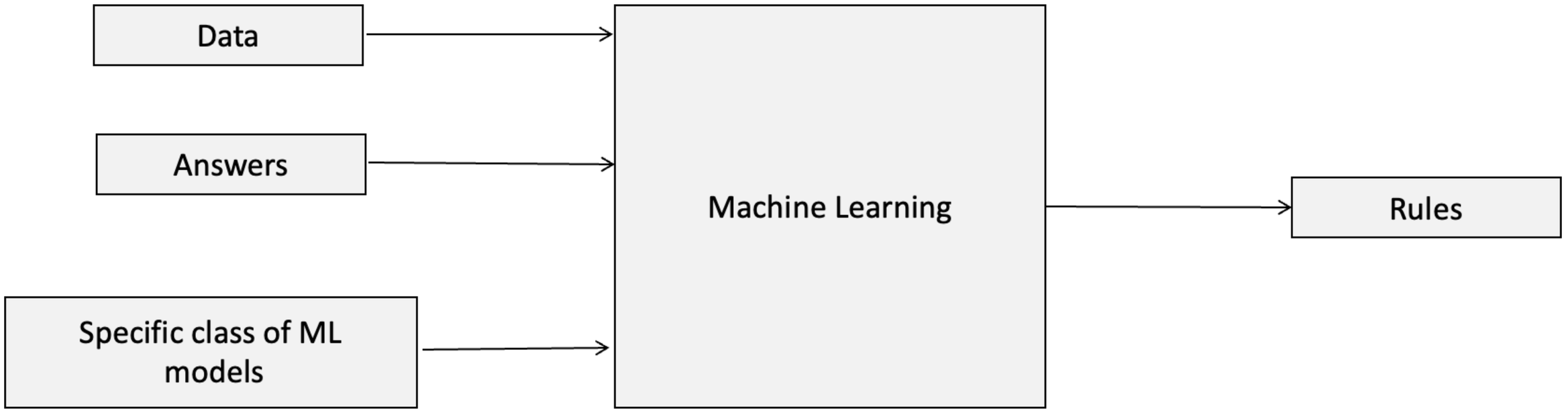
recall Lecture 1

Conventional Programming Schema

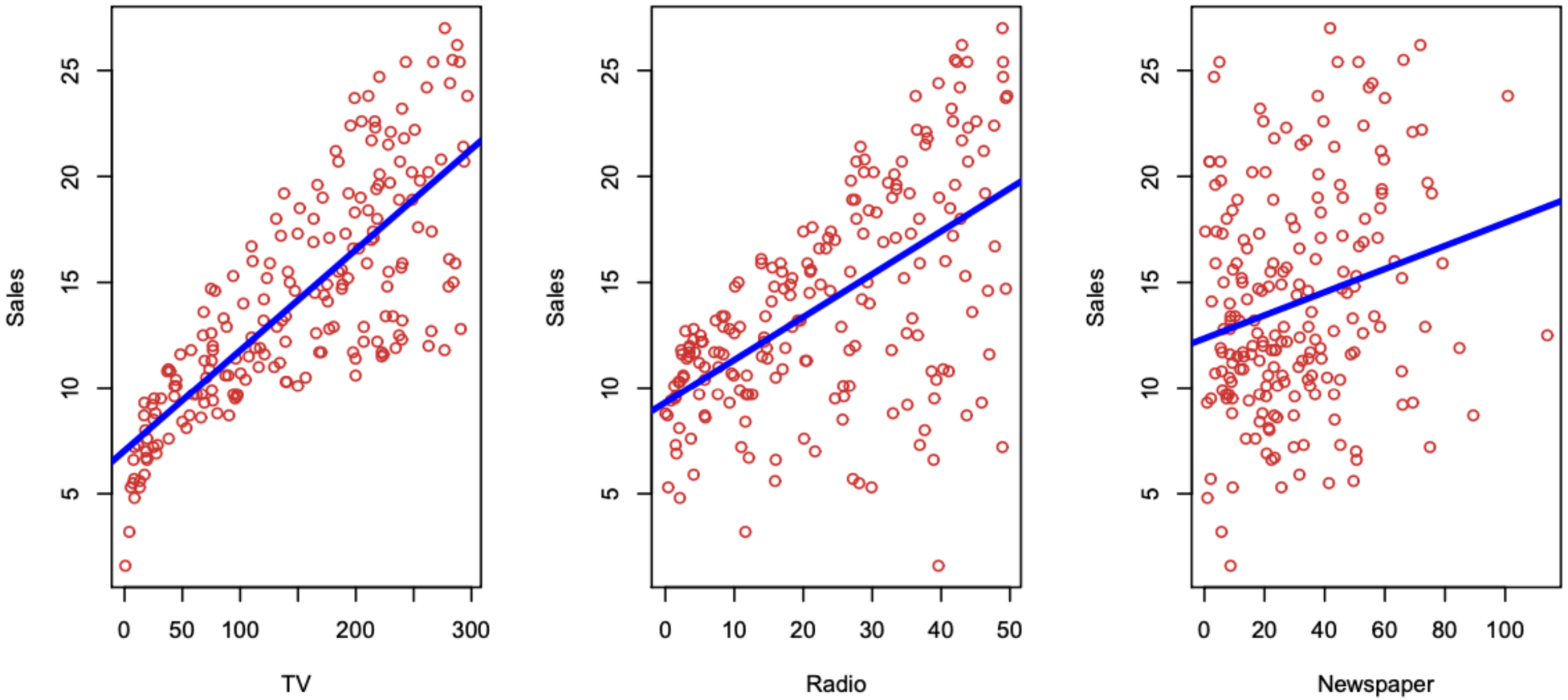


recall Lecture 1

Machine Learning Schema - expanded



Statistical Learning



Statistical Learning

- Suppose we observe a quantitative response Y and p predictors $X_1, \ldots, X_p$
- $Y = f(X) + \epsilon$ where $X = (X_1, \ldots, X_p)$
- f is an unknown, but fixed function of $X_1, \ldots, X_p$
- ϵ is independent of $X_1, \ldots, X_p$ and has a mean of 0.
- f represents the systematic information $X_1, \ldots, X_p$ provides about Y.
- What does ϵ represent?

Why estimate f? Inference

- Is there a relationship between advertising budget and sales?
- How strong is this relationship?
- Which media are associated with sales?
- How large is the association between each media and sales?
- $f(X) = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$

What is special about ordinary least squares?

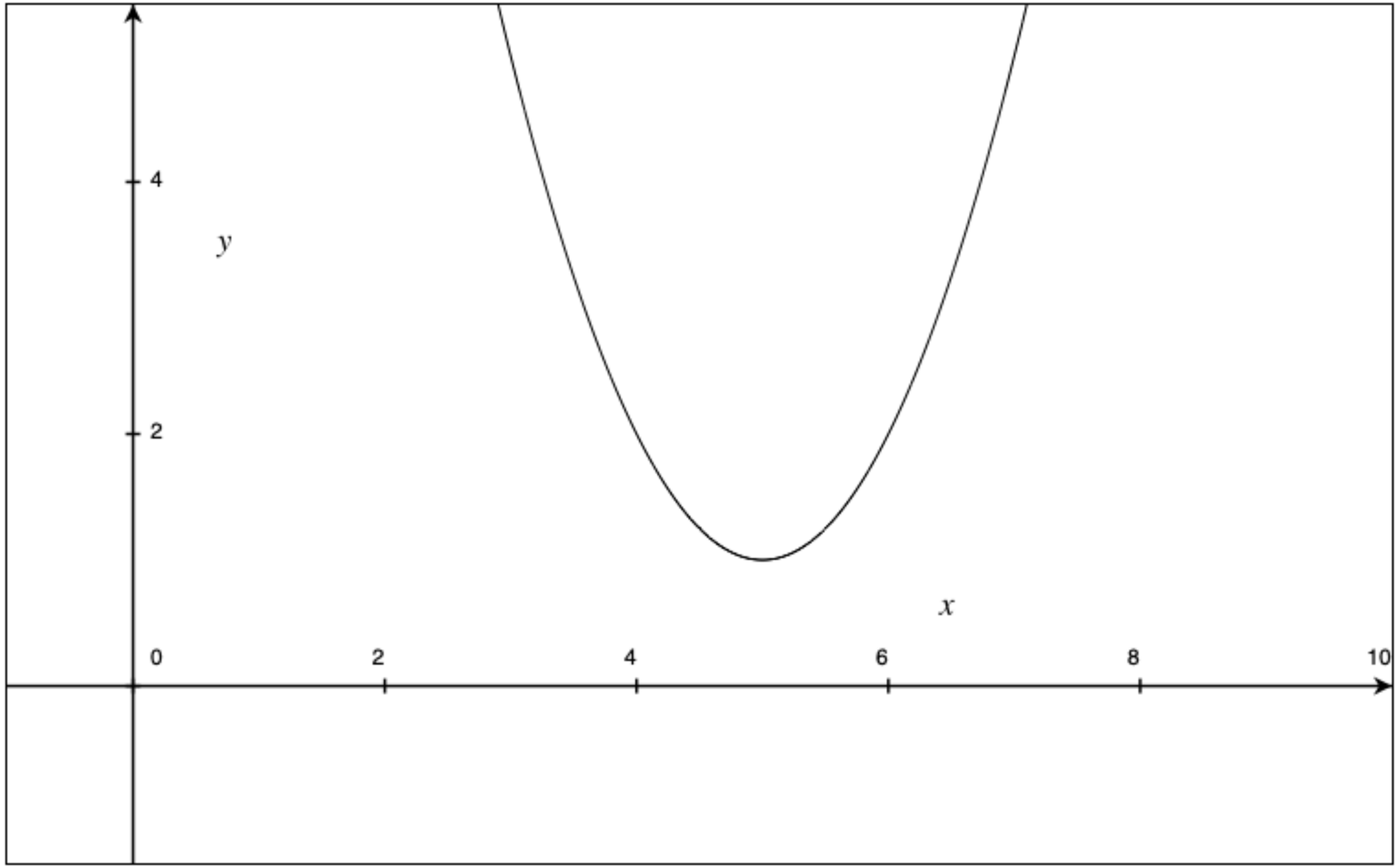
- $y_i = \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \epsilon_i$

- Quadratic minimization problem:

  Find value of β which minimizes

$$\sum_{i=1}^{n} \left| y_i - \sum_{j=1}^{p} x_{ij}\beta_j \right|^2$$

Convexity of objective function

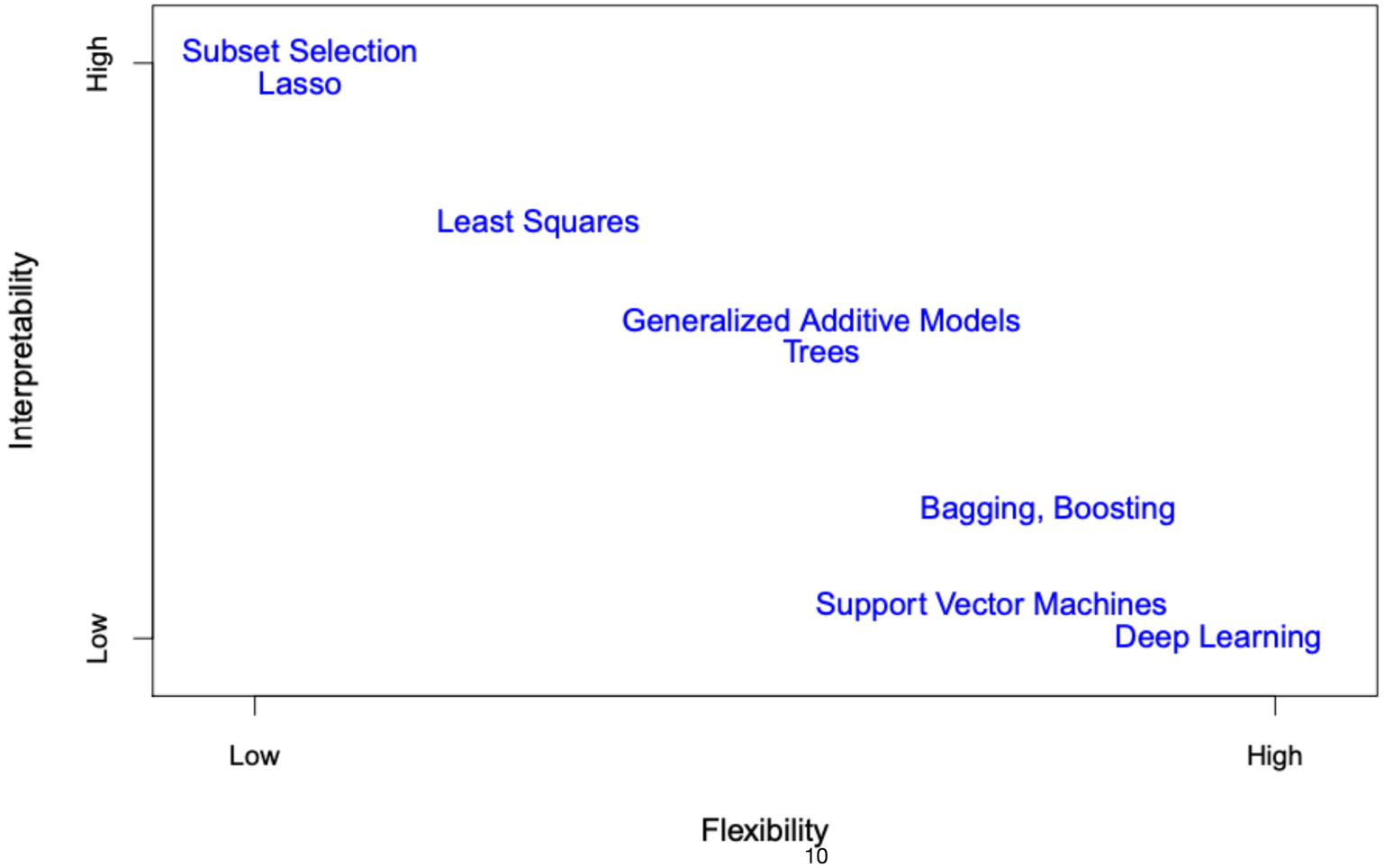


What is special about ordinary least squares (cont'd) ?

- We have formulated the problem in such a way so as to enable us to algebraically derive the unique solution
- Alternative derivations are possible
  geometric projection, maximum likelihood, general method of moments
- Best linear unbiased estimator
- It is a maximum likelihood estimator which outperforms any non-linear estimator
- Has optimal features from a variety of different approaches

- Special because its structure lends itself to
  - No statistical learning, i.e. there is no iterative process to find a solution; the solution is closed-form and can be derived
  - Gold standard for inference and interpretability

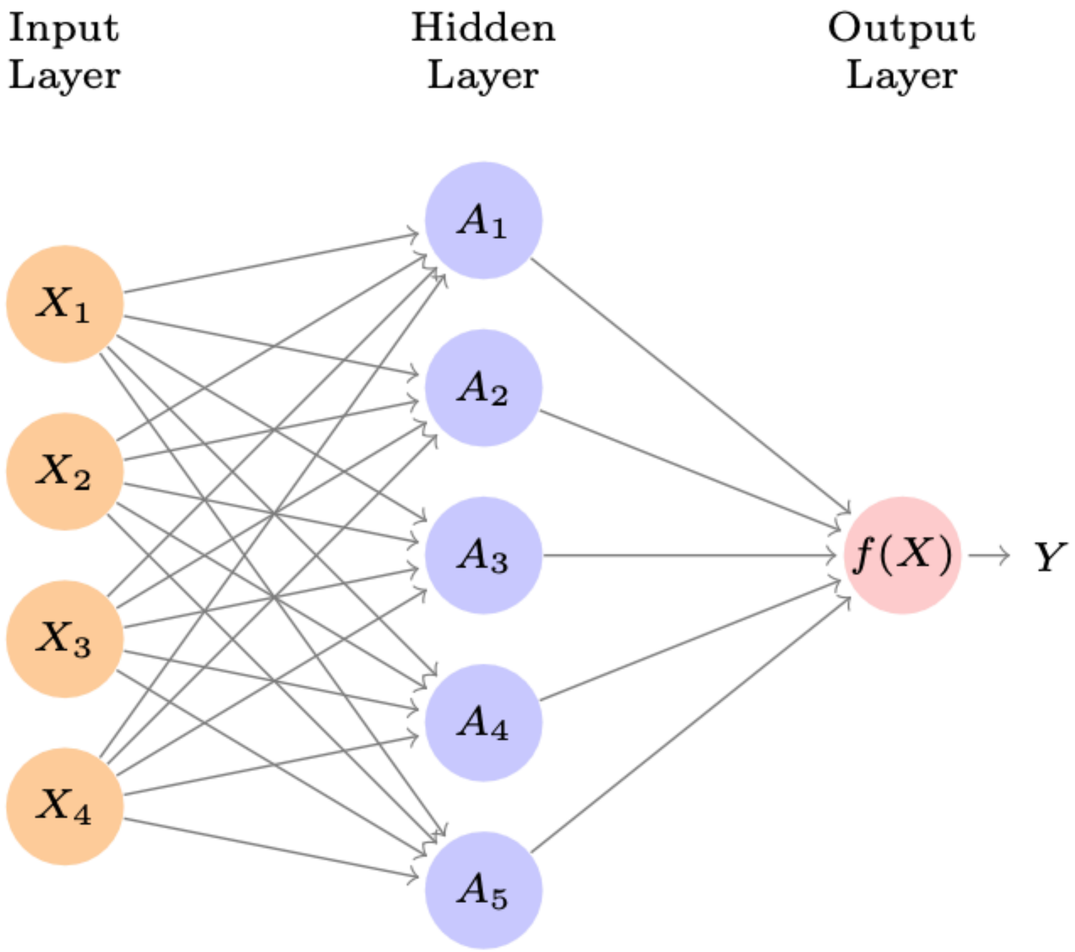
Interpretability versus Flexibility



Why estimate f? Prediction

- not concerned about structure
- observable output are not easy to produce
- $X_1, \ldots, X_p$ are characteristics of individuals' blood samples
- Y encodes individuals' severe reactions to a particular drug
- The resulting prediction comes from a black box $f(X)$

Single Layer Neural Networks



Single Layer Neural Networks

- input vector of p variables $X = (X_1, X_2, \ldots, X_p)$
  and build a non-linear function $f(X)$ to predict response Y.
- What is the structure of this model?
  - Terminology
    Say we have four features X_1, X_2, X_3, X_4 which make up input layer.
  - Each feature from the input layer feeds into each of the K hidden units (say, we choose five).
  - These K units in the hidden layer then feed into output layer,
    a linear regression in K=5 activations.

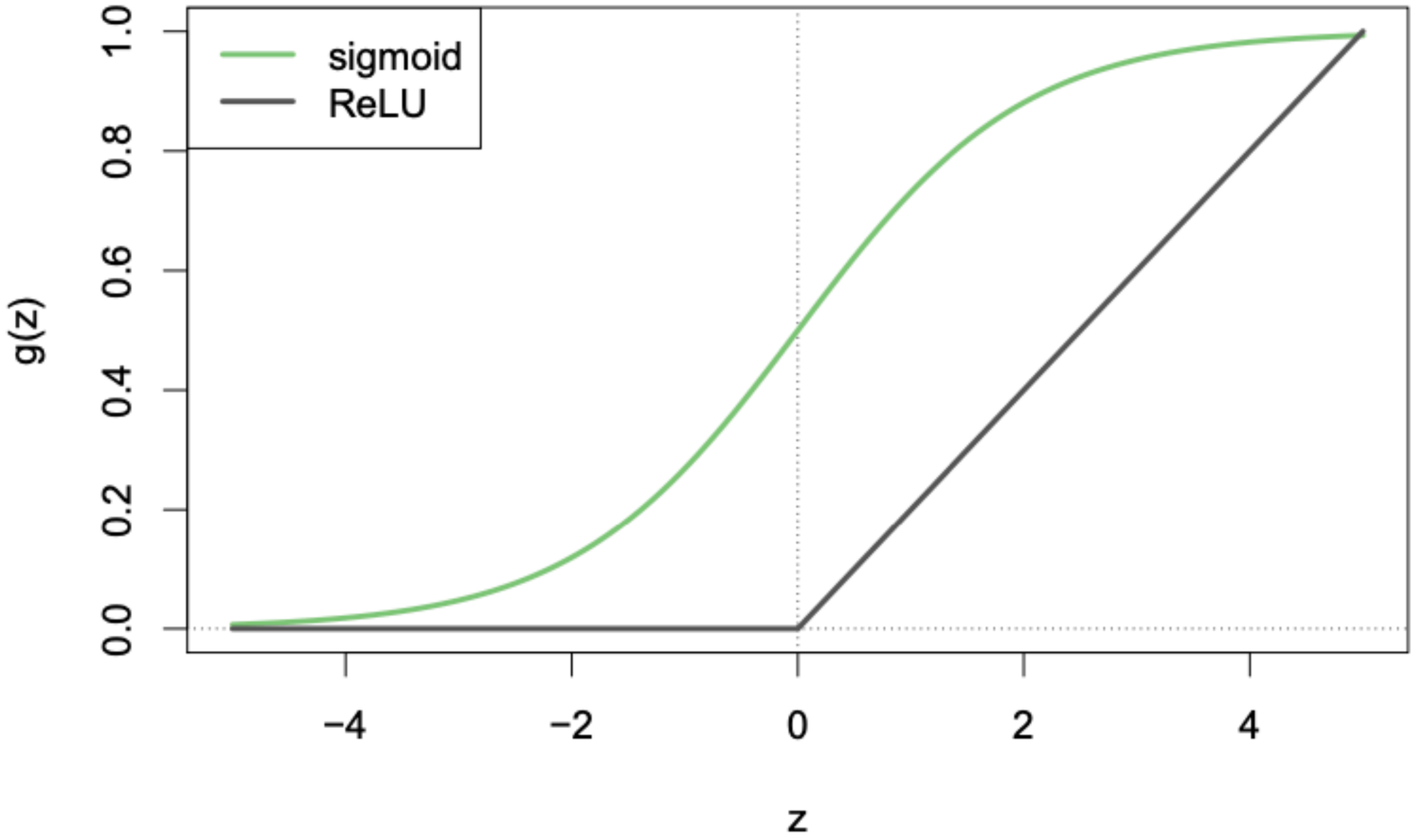
Single Layer Neural Networks

$$f(X) = \beta_0 + \sum_{k=1}^{K} \beta_k g(w_{k0} + \sum_{j=1}^{p} w_{kj} X_j)$$

K hidden layers: $g(w_{k0} + \sum_{j=1}^{p} w_{kj} X_j)$

rectified linear unit (ReLU): $g(z) = \begin{cases} 0 & \text{if } z < 0 \\ z & \text{otherwise} \end{cases}$

Single Layer Neural Network



Single Layer Neural Networks

- We derive five new features by computing by linear combinations of X
- Squash each through an ‘activation function’ $g(\cdot)$ to transform it
- Final model is linear in these derived variables