Hawaii Machine Learning Meetup

Machine Learning Review

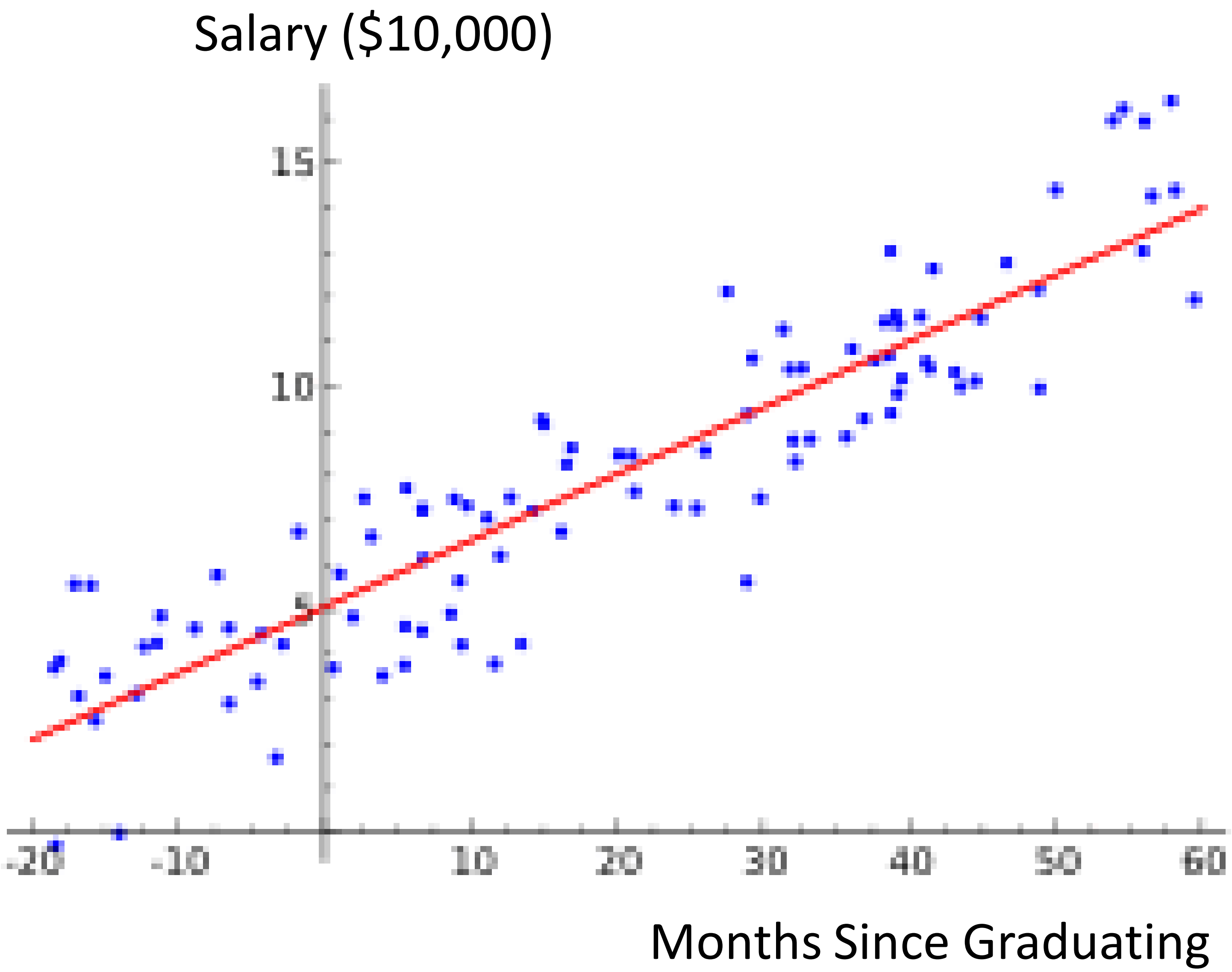
Table of Contents

Linear Regression

Linear Regression predicts a continuous variable using a linear model.

$$\hat{y} = \beta_0 + \beta_1 x_1$$

For example, predicting salary using time since graduating.



Terminology

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

- y is a continuous target variable, response, or dependent variable
- $\hat{y}$ is our prediction
- x_j are the predictors, features, or independent variables
- β_0 is the intercept or bias
- β_j are the coefficients, weights, or parameters

Finding the best coefficients

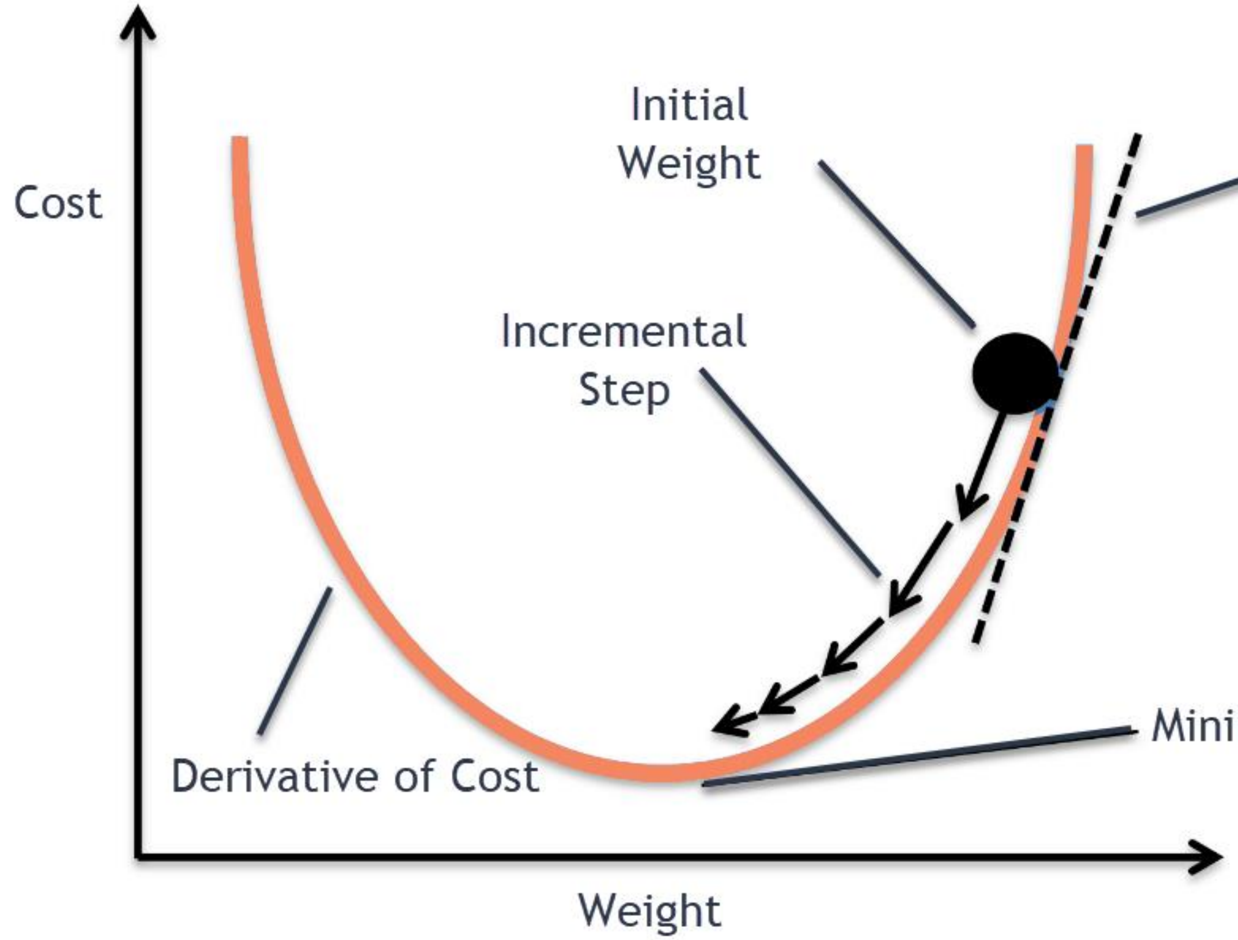
Linear Regression: $\hat{y}^{(i)} = \beta_0 + \beta_1 x_1^{(i)} + \beta_2 x_2^{(i)} + \cdots + \beta_p x_p^{(i)}$

Error/Cost/Loss Function: $RSS = \sum_{i=1}^{n} \left(\hat{y}^{(i)} - y^{(i)}\right)^2$

We find the best β_j by minimizing RSS using gradient descent.

$$\beta_j \leftarrow \beta_j - 2\alpha \cdot \sum_{i=1}^{n} \left(\hat{y}^{(i)} - y^{(i)}\right) \cdot x_j^{(i)}$$

Finding the best coefficients

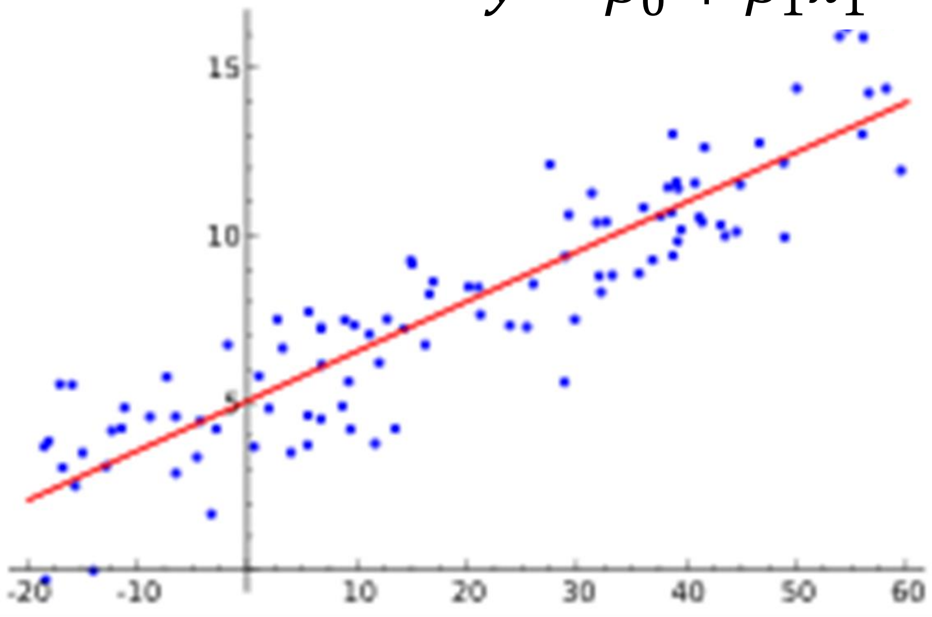


Why do we care about linear regression and what are the benefits?

Why do we care about linear regression and what are the benefits?

Salary ($10,000)

$$y = \beta_0 + \beta_1 x_1$$



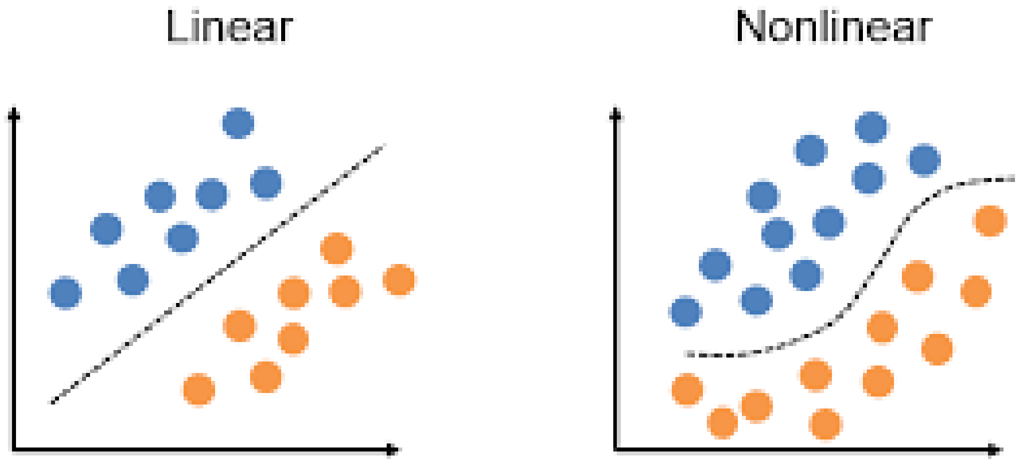
Months Since Graduating

- Relatively simple with intuitive an interpretation.

Why do we care about linear regression and what are the benefits?

- Relatively simple with intuitive an interpretation.
- Its simplicity makes it fast, scalable, and widely applicable.

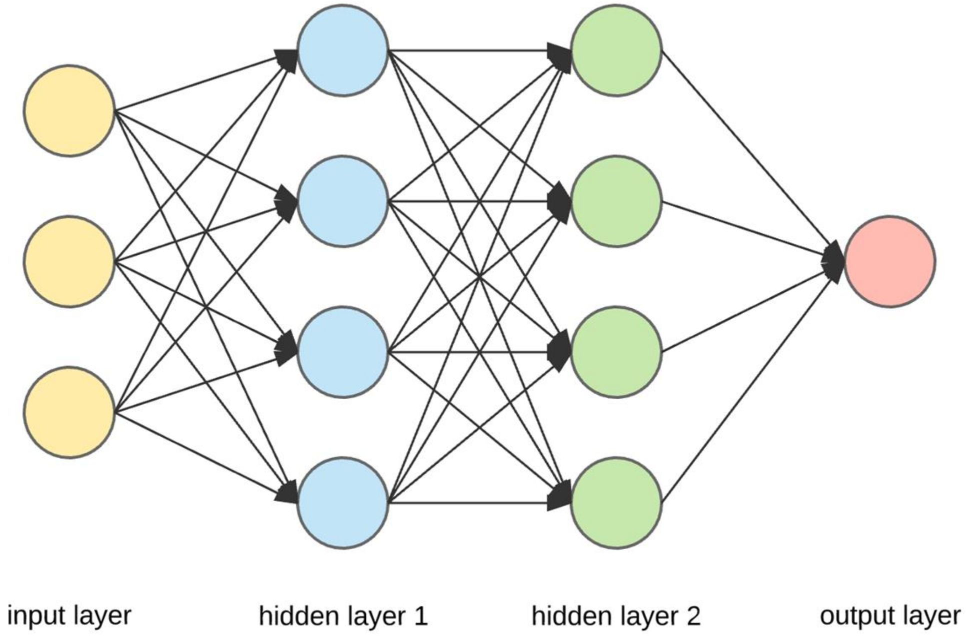
Why do we care about linear regression and what are the benefits?



- Relatively simple with intuitive an interpretation.
- Its simplicity makes it fast, scalable, and widely applicable.
- Good baseline to compare more complicated models to.

Linear Regression

Why do we care about linear regression and what are the benefits?



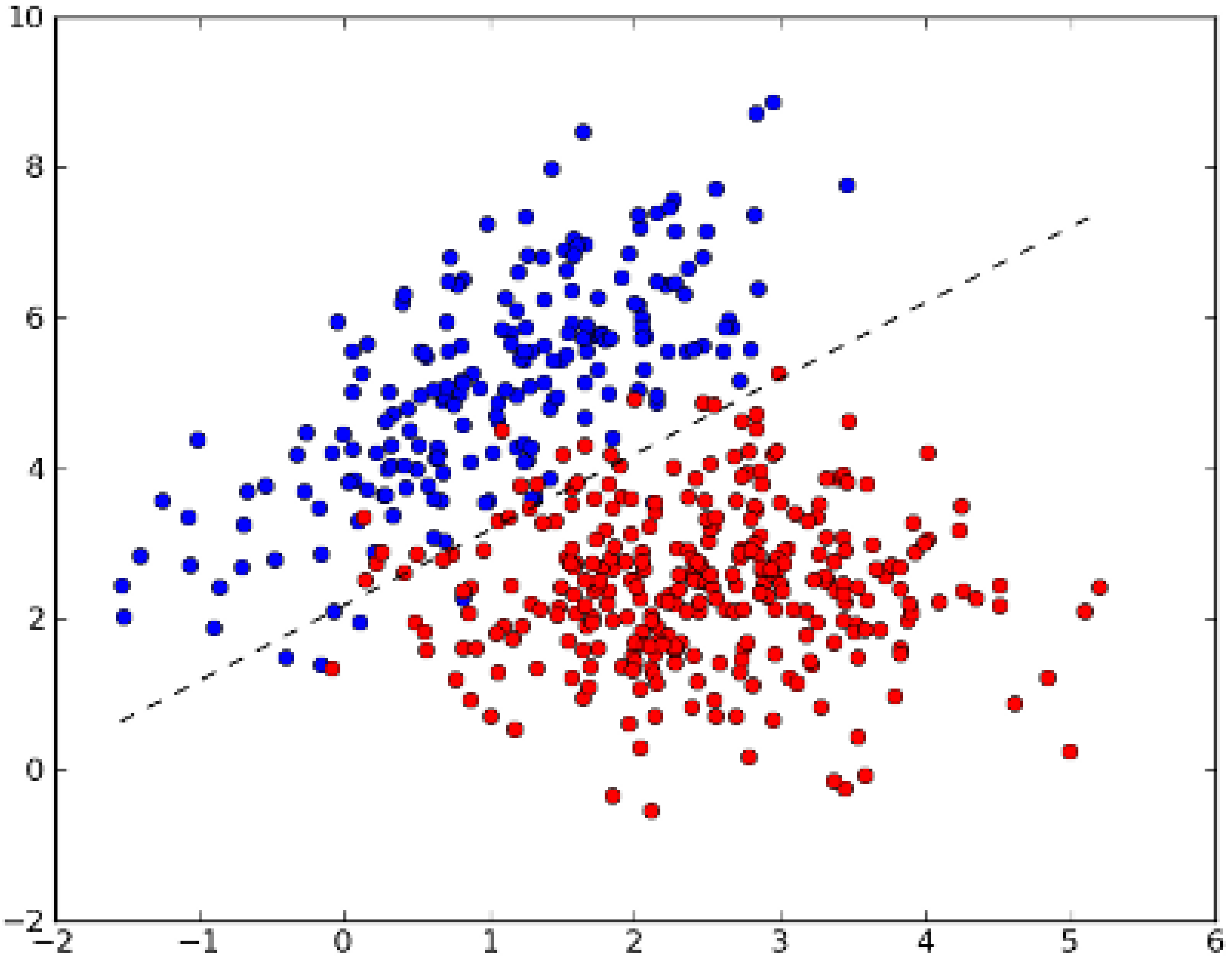
- Relatively simple with intuitive an interpretation.
- Its simplicity makes it fast, scalable, and widely applicable.
- Good baseline to compare more complicated models to.
- Building block for complex models

Classification

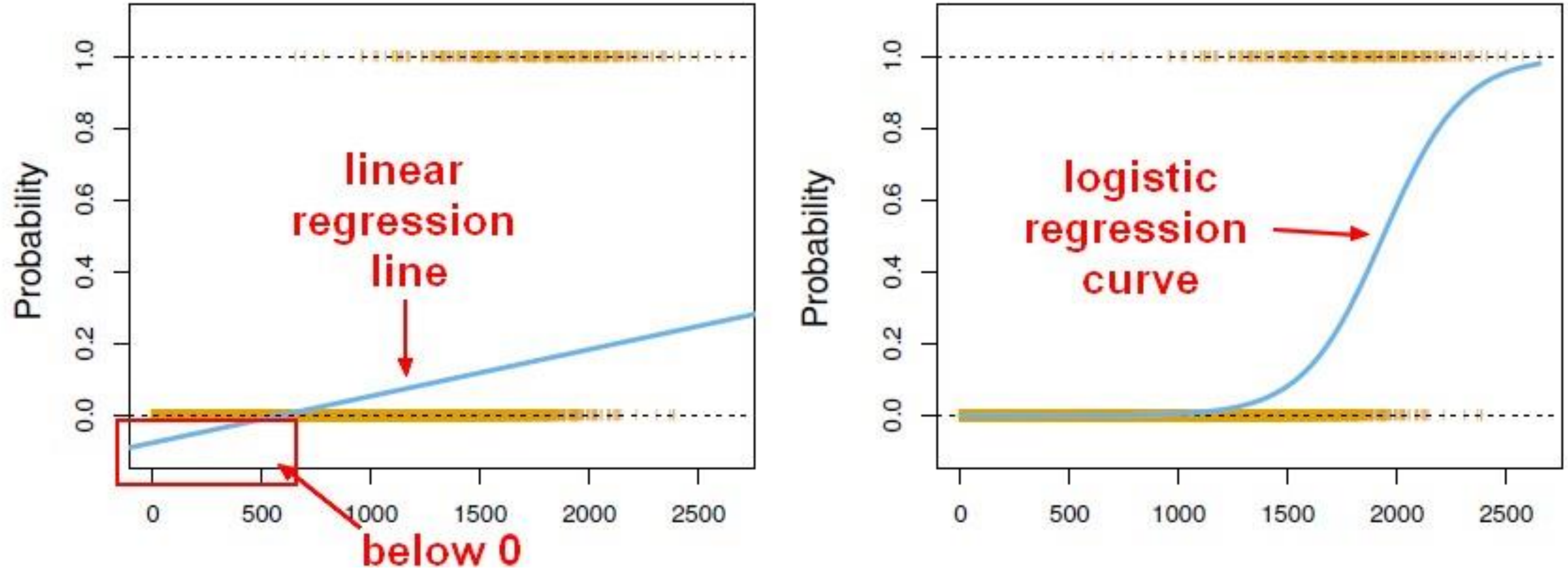
Classification predicts a categorical variable (discrete).

For example,

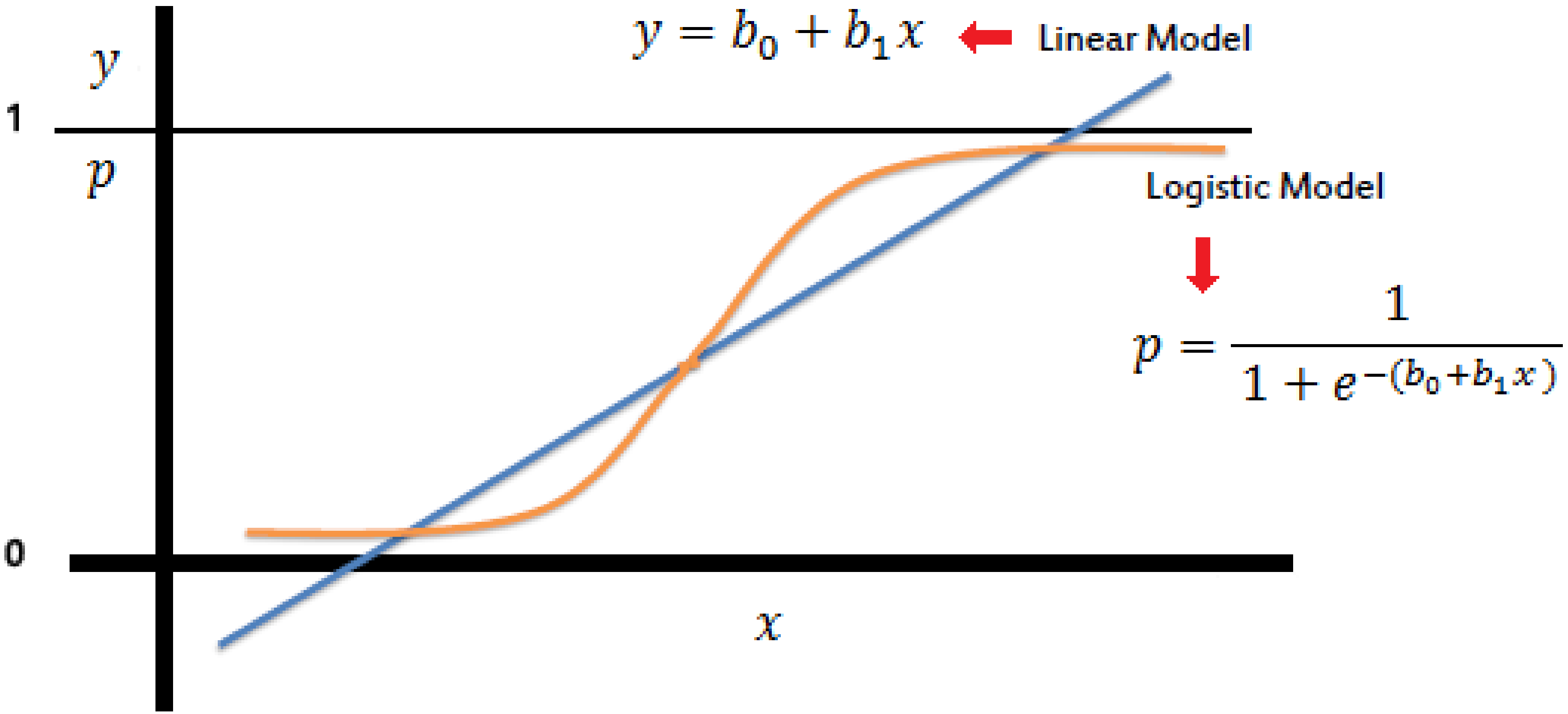
- predicting whether a person has a disease or not based on the results of lab tests
- predicting the type of objects in an image



Logistic Regression



Logistic Regression



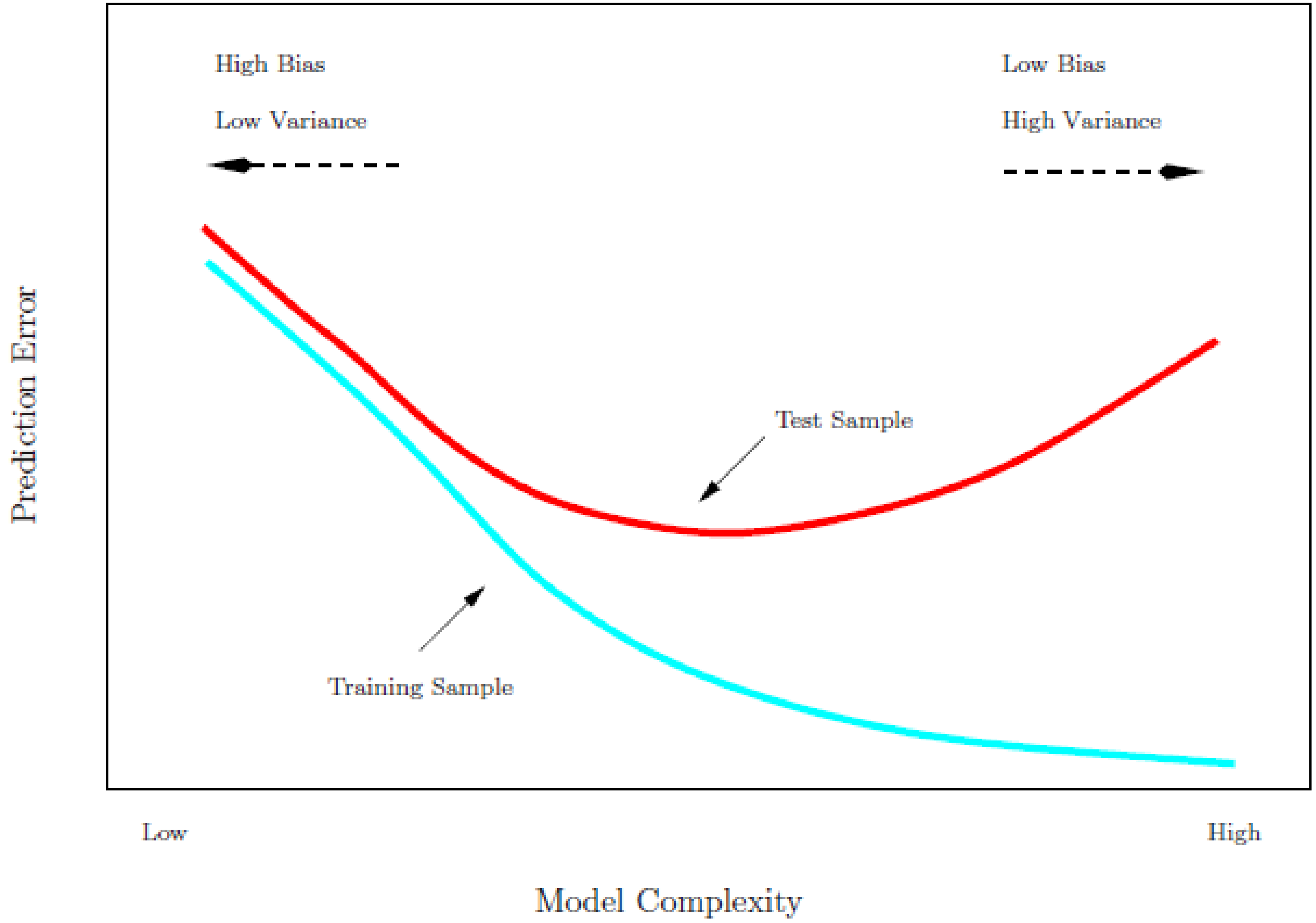
Resampling Methods

Answers: How well is the model doing?

Important Techniques:

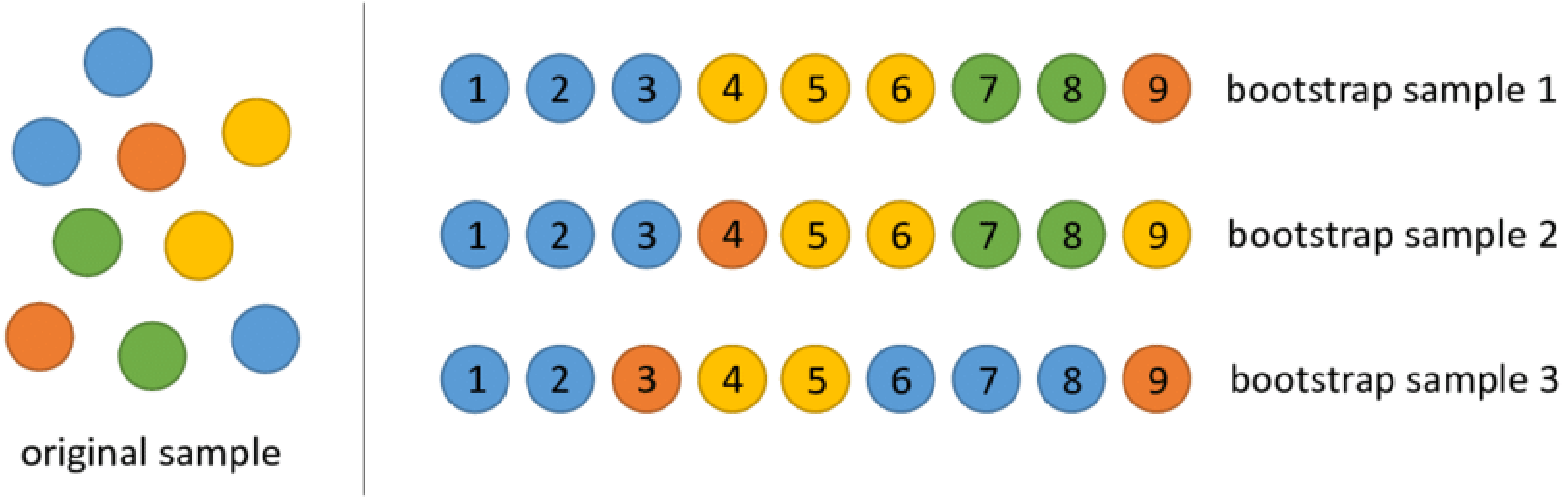
- Cross Validation
- Bootstrap Sampling

Training- versus Test-Set Performance



Cross Validation

Bootstrap Sampling

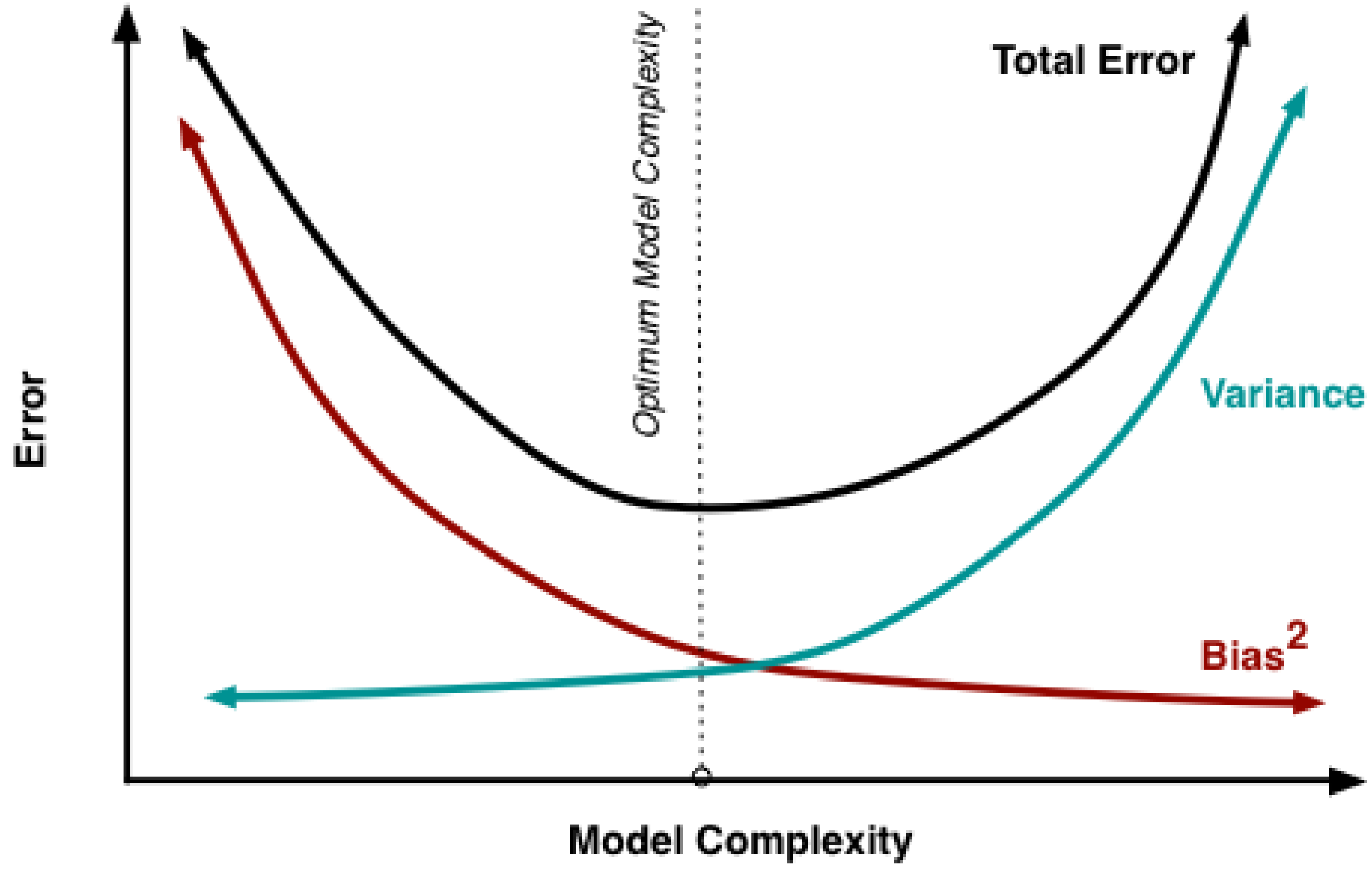


Model Selection and Regularization

Answers: How do we choose the best model?

Important Techniques:

- Subset Selection
- Shrinkage & Regularization



Subset Selection

Linear Regression:

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

Subset Selection tries to find the best combination of predictors.

- Too many predictors will increase variance unnecessarily.
- Too few predictors will make it hard to fit the data properly.

Model Selection

Best Subset Selection (brute force)

- Start with a null model
- For k = 1 through p, fit all possible models with k predictors
- Select best model using cross validation, AIC, BIC, or adjusted R^2

Forwards Stepwise Selection (null model -> greedily add)

- Start with a null model
- Add the most useful predictor one at a time

Backwards Stepwise Selection (all features -> greedily remove)

- Start with all features
- Remove the least useful predictor one at a time

Shrinkage and Regularization

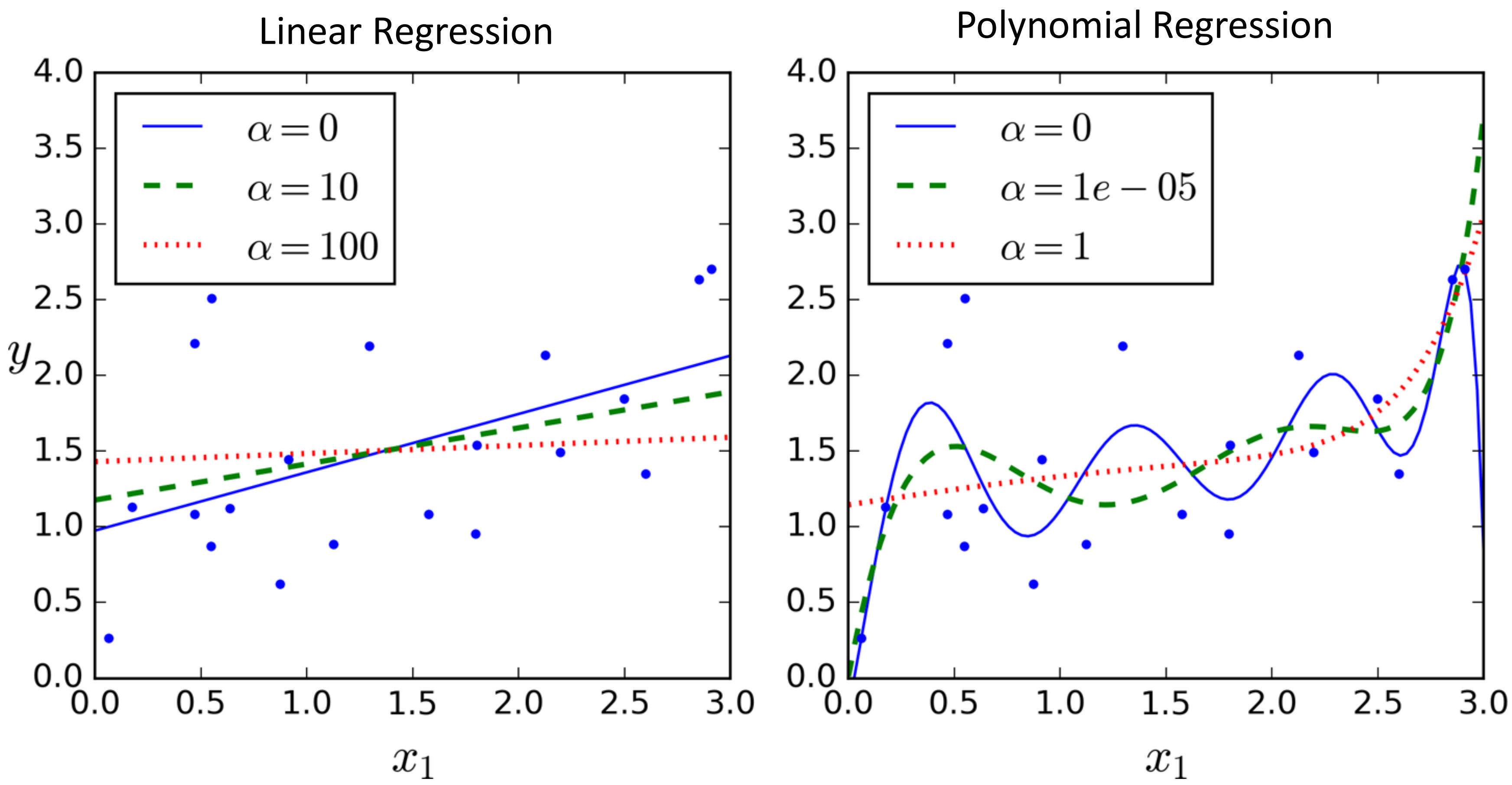
Linear Regression: $\hat{y}^{(i)} = \beta_0 + \beta_1 x_1^{(i)} + \beta_2 x_2^{(i)} + \cdots + \beta_p x_p^{(i)}$

Residual Sum of Squares: $RSS = \sum_{i=1}^{n} \left(\beta_0 + \sum_{j=1}^{p} \beta_j x_j^{(i)} - y^{(i)} \right)^2$

Ridge: $RSS + \alpha \sum_{i=1}^{n} \left| \beta_j \right|^2$ (makes coefficients small but not zero)

Lasso: $RSS + \lambda \sum_{i=1}^{n} \left| \beta_j \right|$ (makes coefficients small and equal to zero)

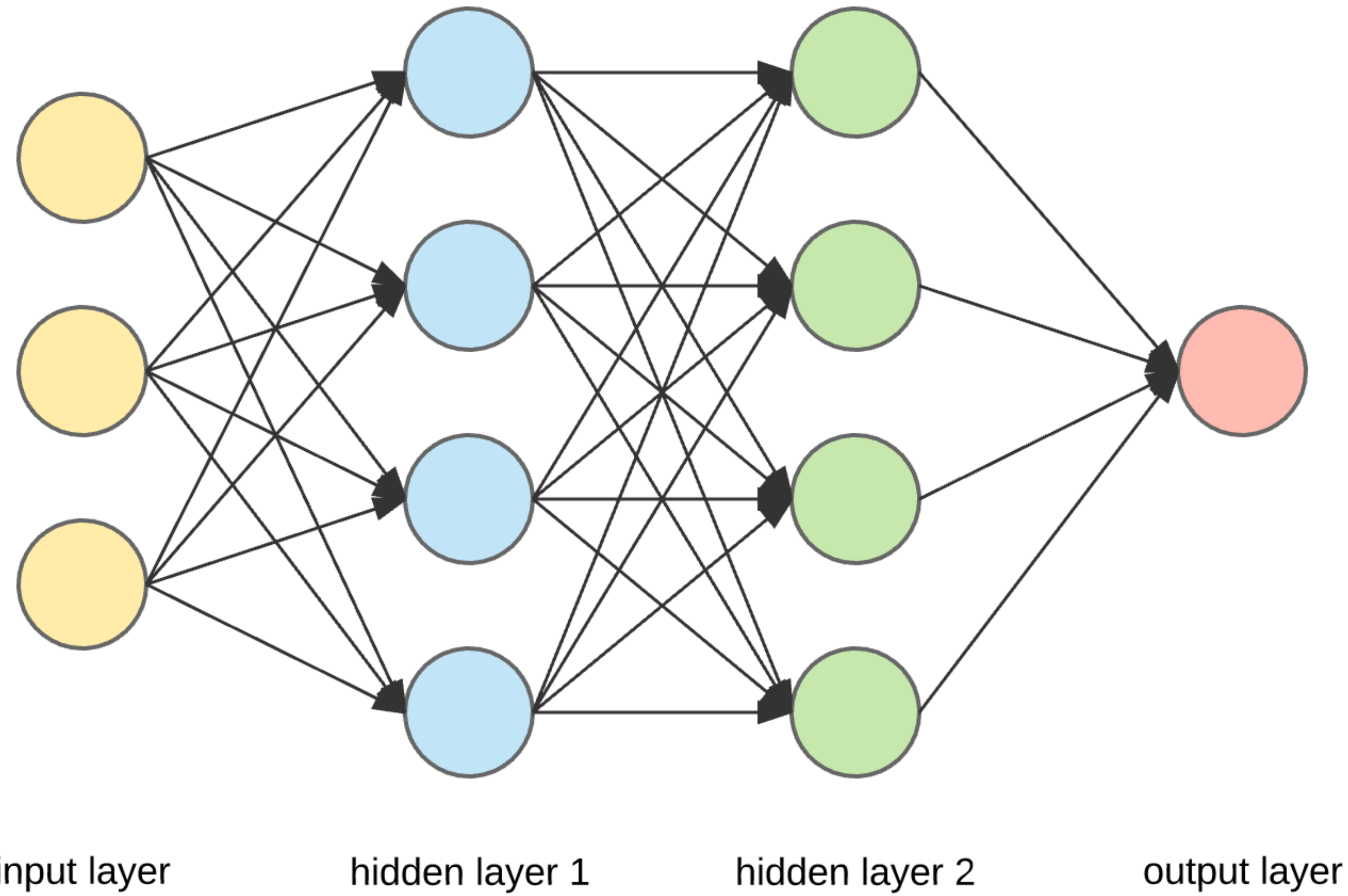
Shrinkage and Regularization



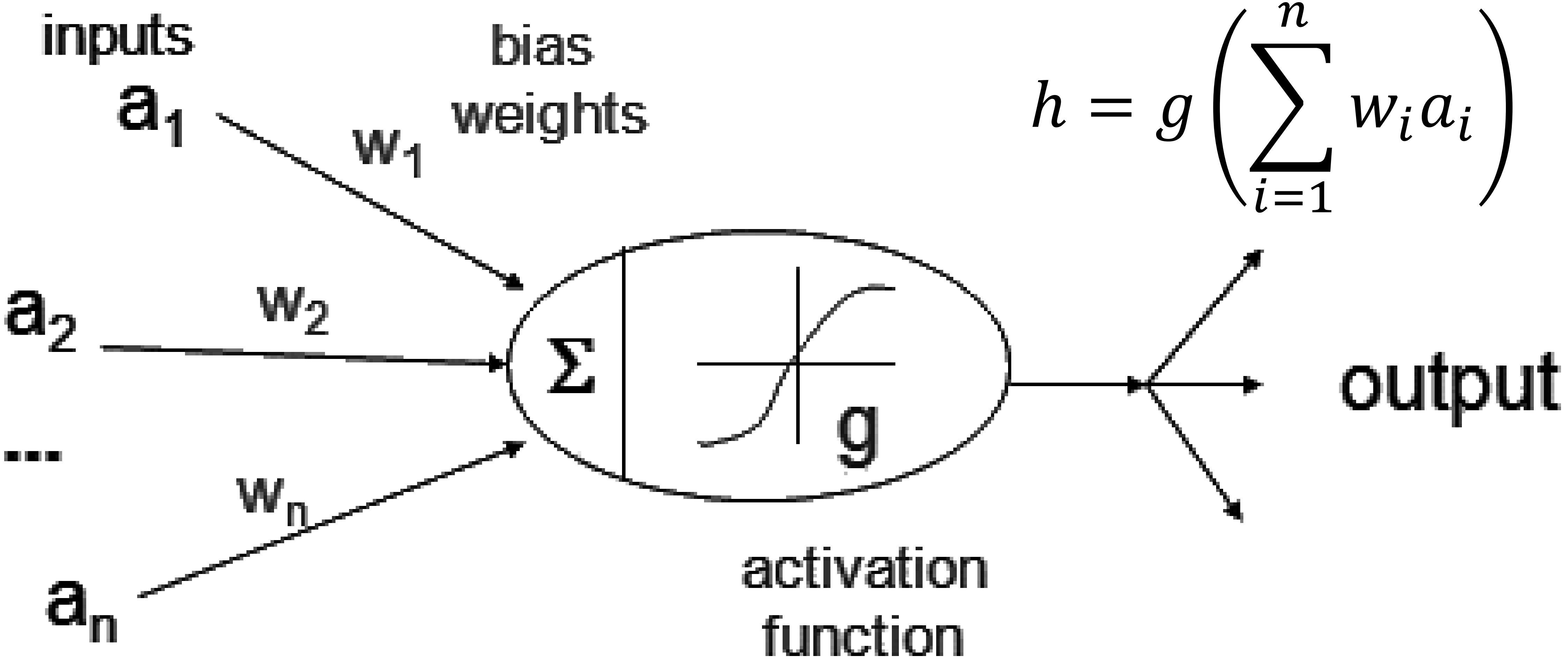
Neural Networks

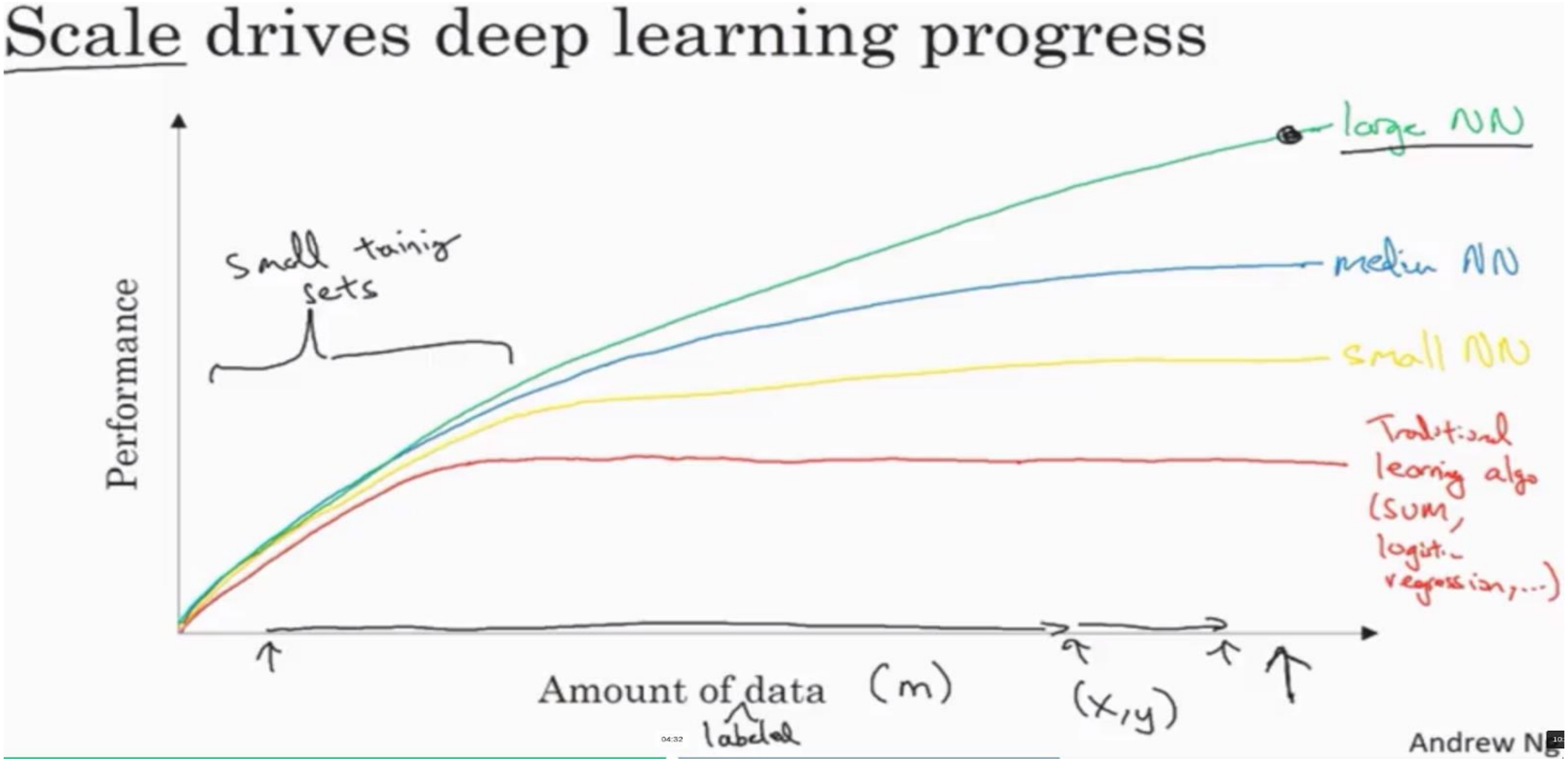
Components

- Input Layer
- Hidden Layer
- Output Layer
- Nonlinear Activations
- Neurons



$$h = g\left(\sum_{i=1}^{n} w_i a_i\right)$$




Scale drives deep learning progress
large NN
medium NN
small NN
Traditional learning algo (SVM, logistic regression, ...)
small training sets
Performance
Amount of data (m)
labeled
(x, y)
Andrew Ng

What are the benefits of fully connected neural networks?

- **Nonlinear**: Universal Approximation Theorem: Any continuous function can be modeled with a single hidden layer and a sufficient number of neurons.
- **Scalable**: Modern advances in hardware (training on GPUs or TPUs) have allowed very large neural networks to be trained.
- **Foundational**: Fully connected networks form the foundation for more complex architectures like convolutional neural networks and recurrent neural networks.

Resources

Courses

- **Introduction Statistical Learning**
  https://online.stanford.edu/courses/sohs-ystatslearning-statistical-learning-self-paced
- **Deep Learning Course 1 (Deep Learning and Neural Networks)**
  https://www.coursera.org/specializations/deep-learning

Hawaii Machine Learning Study Group

- Meet twice a month at UH’s Post building
- https://hawaiimachinelearning.github.io/studygroup/