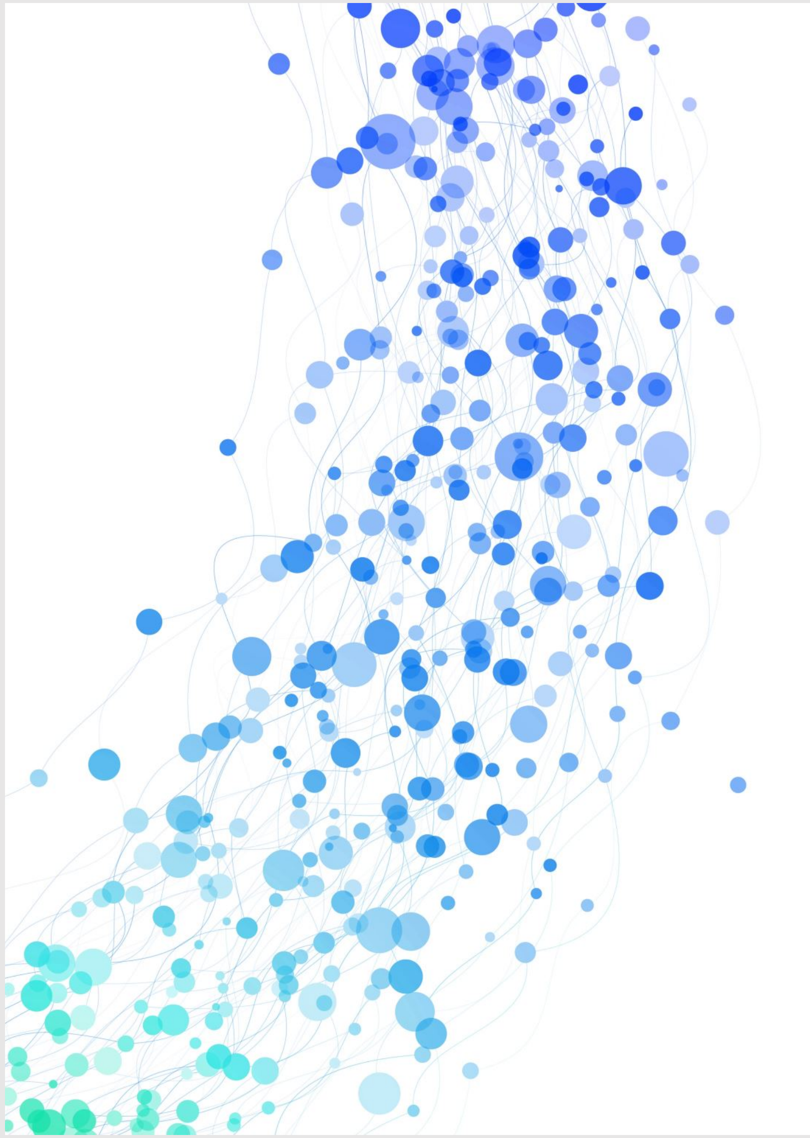


Module 2: STATISTICAL MACHINE LEARNING

Introduction, Regression, and Classification, Decision Tress, Random Forests

Reference: James, G., Witten, D., Hastie, T., Tibshirani, R. (2017). An Introduction to Statistical Learning: with Applications in R., Springer.)

What Is Statistical Learning?

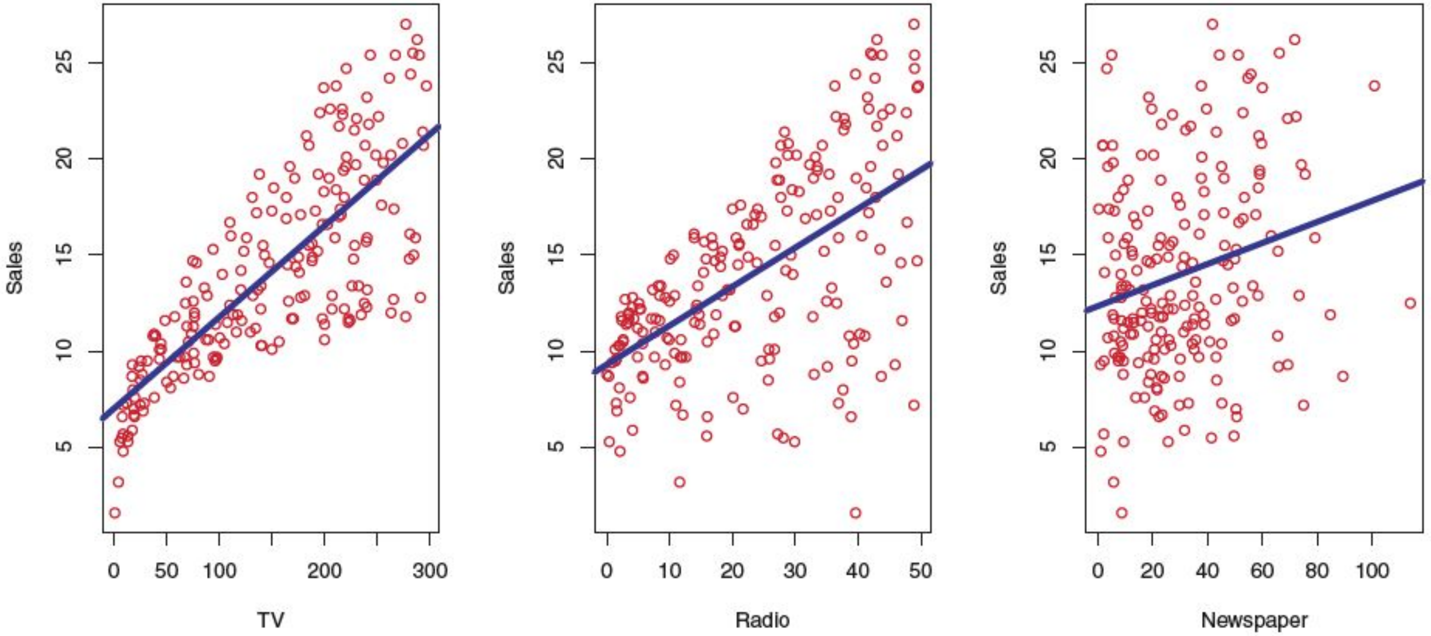


FIGURE 2.1. *The* Advertising *data set. The plot displays* sales*, in thousands of units, as a function of* TV*,* radio*, and* newspaper *budgets, in thousands of dollars, for* 200 *different markets. In each plot we show the simple least squares fit of* sales *to that variable, as described in Chapter 3. In other words, each blue line represents a simple model that can be used to predict* sales *using* TV*,* radio*, and* newspaper*, respectively.*

- In this setting, the advertising budgets are ***input*** variables while sales is an ***output*** variable
- The input variables are typically denoted using the variable symbol X, with a subscript to distinguish them.
- So X_1 might be the TV budget, X_2 the radio budget, and X_3 the newspaper budget.
- The inputs go by different names, such as ***predictors, independent variables, features***, or sometimes just ***variables***.
- The output variable—in this case, *sales*—is often called ***the response or dependent variable*** and is typically denoted using the symbol Y .

- Generally, we observe
- a quantitative response Y
- p different predictors, $X_1, X_2, \ldots, X_p$

$$Y = f(X) + \varepsilon$$

Where, $X = (X_1, X_2, \ldots, X_p)$

f is some fixed but unknown function of $X_1, X_2, \ldots, X_p$

ε is a random error term, which is independent of X and has mean zero

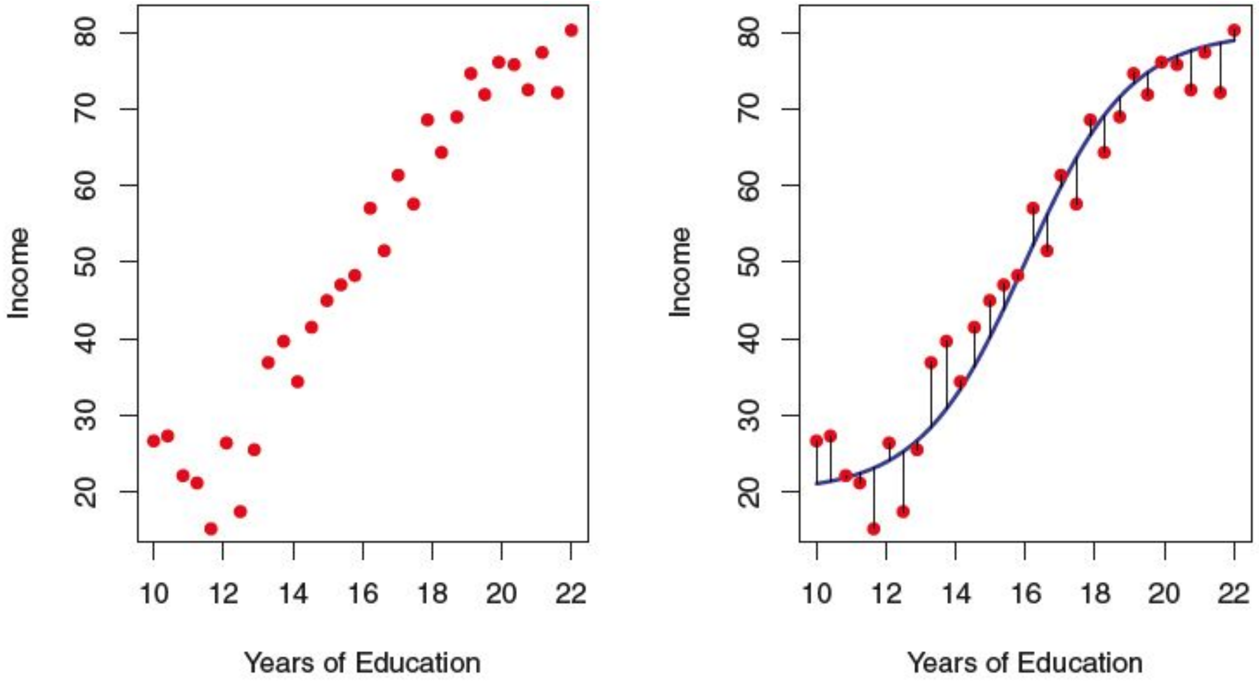


FIGURE 2.2. *The* `Income` *data set.* Left: *The red dots are the observed values of* `income` *(in tens of thousands of dollars) and* `years of education` *for* 30 *individuals.* Right: *The blue curve represents the true underlying relationship between* `income` *and* `years of education`*, which is generally unknown (but is known in this case because the data were simulated). The black lines represent the error associated with each observation. Note that some errors are positive (if an observation lies above the blue curve) and some are negative (if an observation lies below the curve). Overall, these errors have approximately mean zero.*

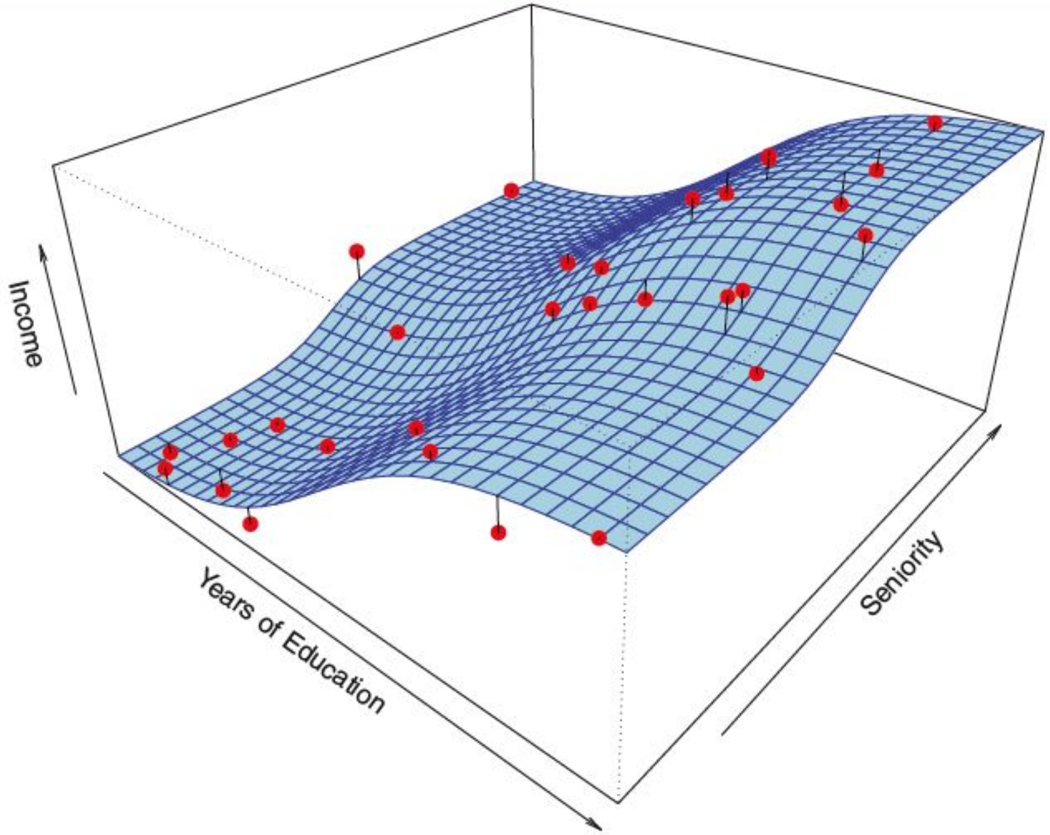


FIGURE 2.3. *The plot displays* income *as a function of* years of education *and* seniority *in the* Income *data set. The blue surface represents the true underlying relationship between* income *and* years of education *and* seniority*, which is known since the data are simulated. The red dots indicate the observed values of these quantities for* 30 *individuals.*

Why Estimate f?

- There are two main reasons that we may wish to estimate f: **prediction** and **inference**.

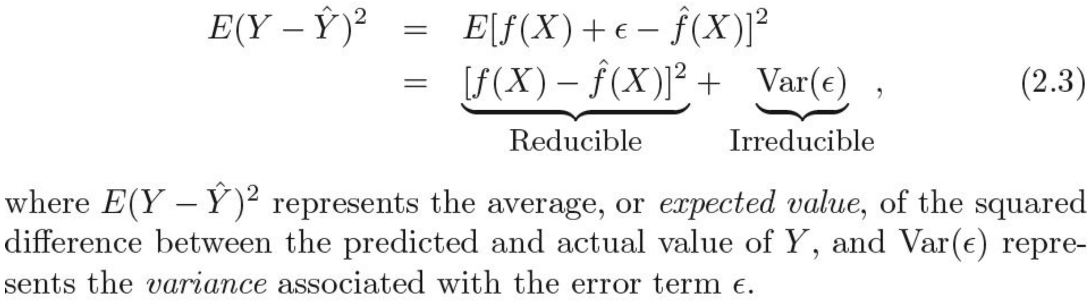
Prediction

- $$\hat{Y} = \hat{f}(X)$$

where $\hat{f}$ represents our estimate for f, and $\hat{Y}$ represents the resulting prediction for Y

The accuracy of $\hat{Y}$ as a prediction for Y depends on two quantities, which we will call the *reducible error* and the *irreducible error.*

f can be black box



$$\begin{aligned} E(Y-\hat{Y})^2 &= E[f(X)+\epsilon-\hat{f}(X)]^2 \\ &= \underbrace{[f(X)-\hat{f}(X)]^2}_{\text{Reducible}} + \underbrace{\text{Var}(\epsilon)}_{\text{Irreducible}} , \end{aligned} \qquad (2.3)$$

where $E(Y-\hat{Y})^2$ represents the average, or *expected value*, of the squared difference between the predicted and actual value of Y, and $\text{Var}(\epsilon)$ represents the *variance* associated with the error term ϵ.

Inference

- f cannot be treated as a black box, because we need to know its exact form
- Which predictors are associated with the response?
- 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?

How Do We Estimate f?

- *Training data* - train or teach our method how to estimate *f*
- Let x_{ij} represent the value of the jth predictor, or input, for observation i, where i = 1, 2, . . ., n and j = 1, 2, . . . , p
- Let y_i represent the response variable for the ith observation.
- Then our training data consist of $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ where $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^T$.

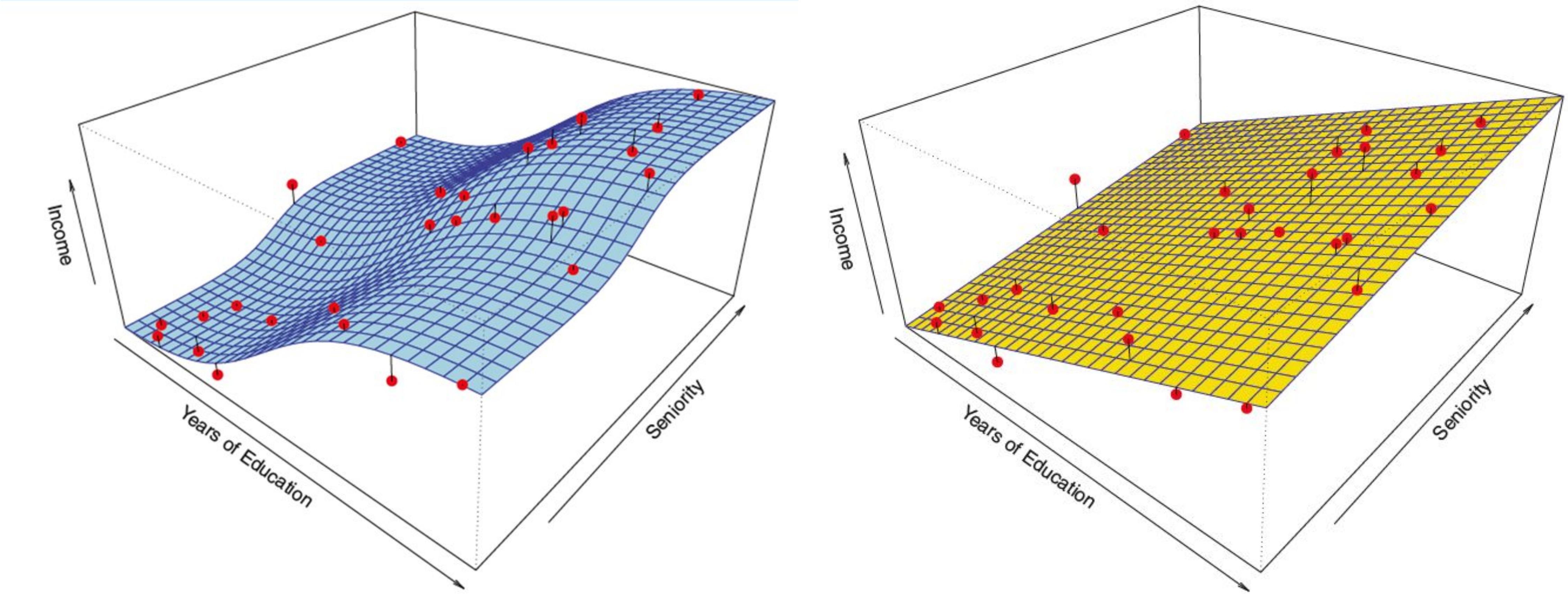
GOAL: find a function $\hat{f}$ such that $Y \approx \hat{f}(X)$ for any observation (X, Y).

- Parametric methods
- Non-parametric methods

Parametric Model

- Parametric methods involve a two-step model-based approach.
  1. make an assumption about the functional form, or shape of f
  2. After a model has been selected, we need a procedure that uses the training data to fit or train the model
     - *least squares*

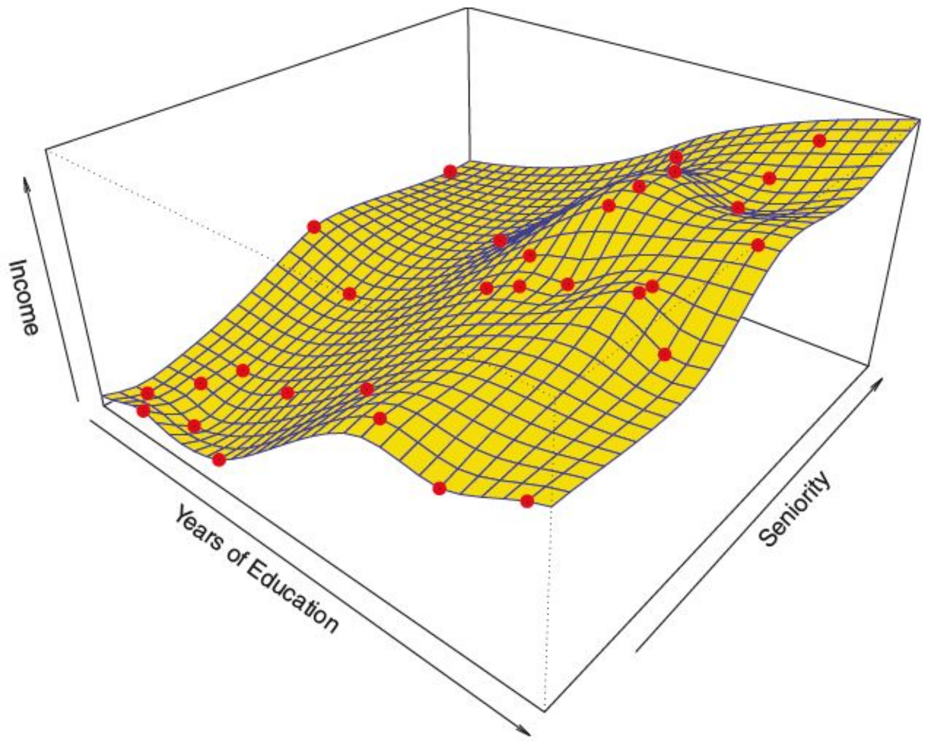
- $f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ ... + \beta_p X_p$



A linear model fit by least squares

Non-parametric Methods

- Non-parametric methods do not make explicit assumptions about the functional form of *f*
- It do not reduce the problem of estimating *f* to a small number of parameters, a very large number of observations (far more than is typically needed for a parametric approach) is required to obtain an accurate estimate for *f*.



A thin-plate spline is used to estimate *f*

Supervised Versus Unsupervised Learning

- In supervised learning, the machine is trained on a set of labeled data, which means that the input data is paired with the desired output.
- The machine then learns to predict the output for new input data.
- Supervised learning is often used for tasks such as classification, regression, and object detection.

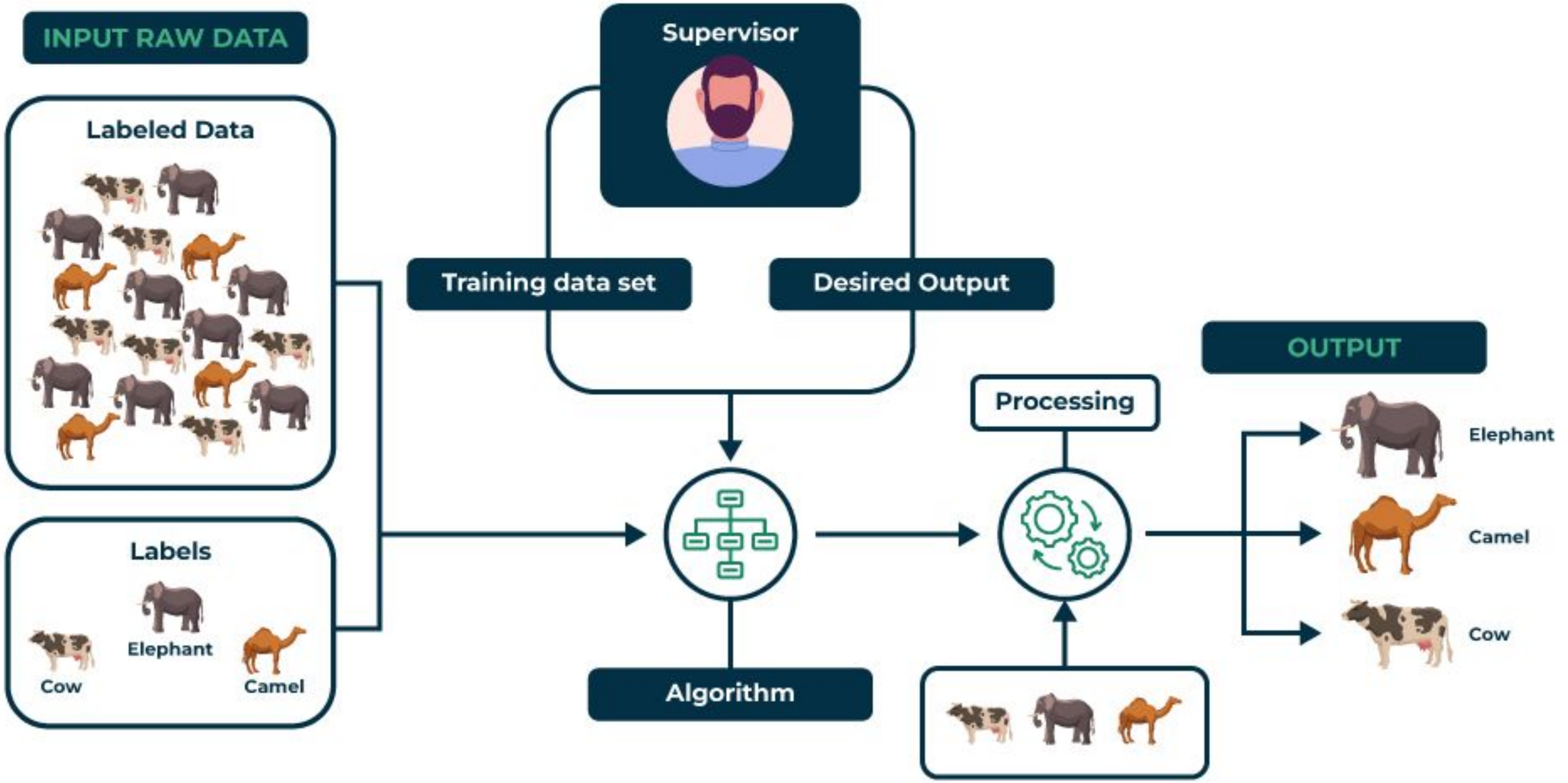
- In unsupervised learning, the machine is trained on a set of unlabeled data, which means that the input data is not paired with the desired output.
- The machine then learns to find patterns and relationships in the data.
- Unsupervised learning is often used for tasks such as clustering, dimensionality reduction, and anomaly detection

Supervised Versus Unsupervised Learning

- Supervised learning involves training a machine from labeled data.
- Labeled data consists of examples with the correct answer or classification.
- The machine learns the relationship between inputs (fruit images) and outputs (fruit labels).
- The trained machine can then make predictions on new, unlabeled data.

- Unsupervised learning allows the model to discover patterns and relationships in unlabeled data.
- Clustering algorithms group similar data points together based on their inherent characteristics.
- Feature extraction captures essential information from the data, enabling the model to make meaningful distinctions.
- Label association assigns categories to the clusters based on the extracted patterns and characteristics.

Supervised Learning



INPUT RAW DATA

Labeled Data

Labels

Cow

Elephant

Camel

Supervisor

Training data set

Desired Output

Algorithm

Processing

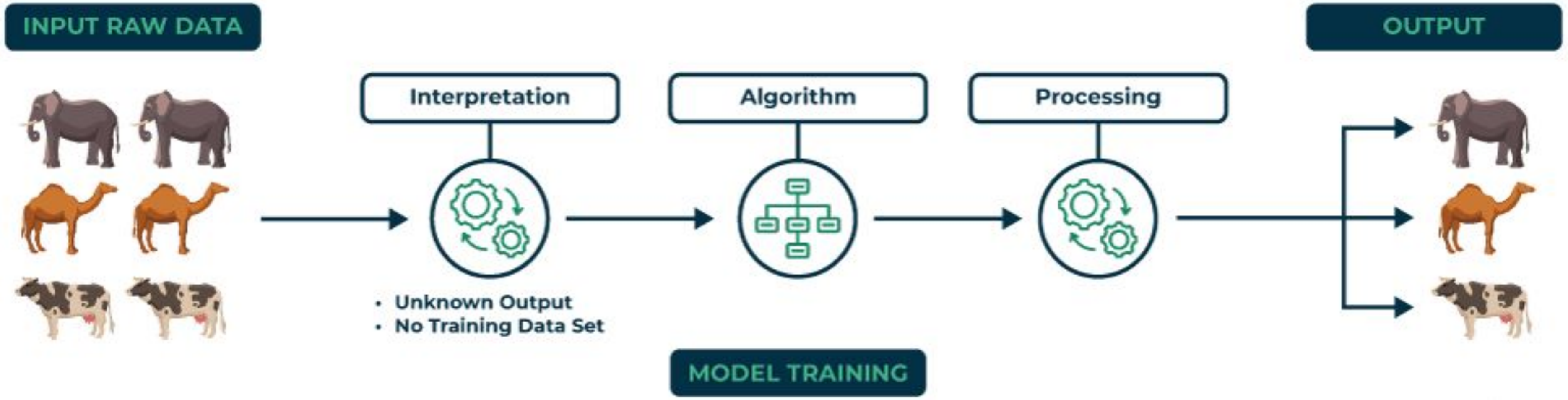
OUTPUT

Elephant

Camel

Cow

Unsupervised Learning



INPUT RAW DATA

Interpretation

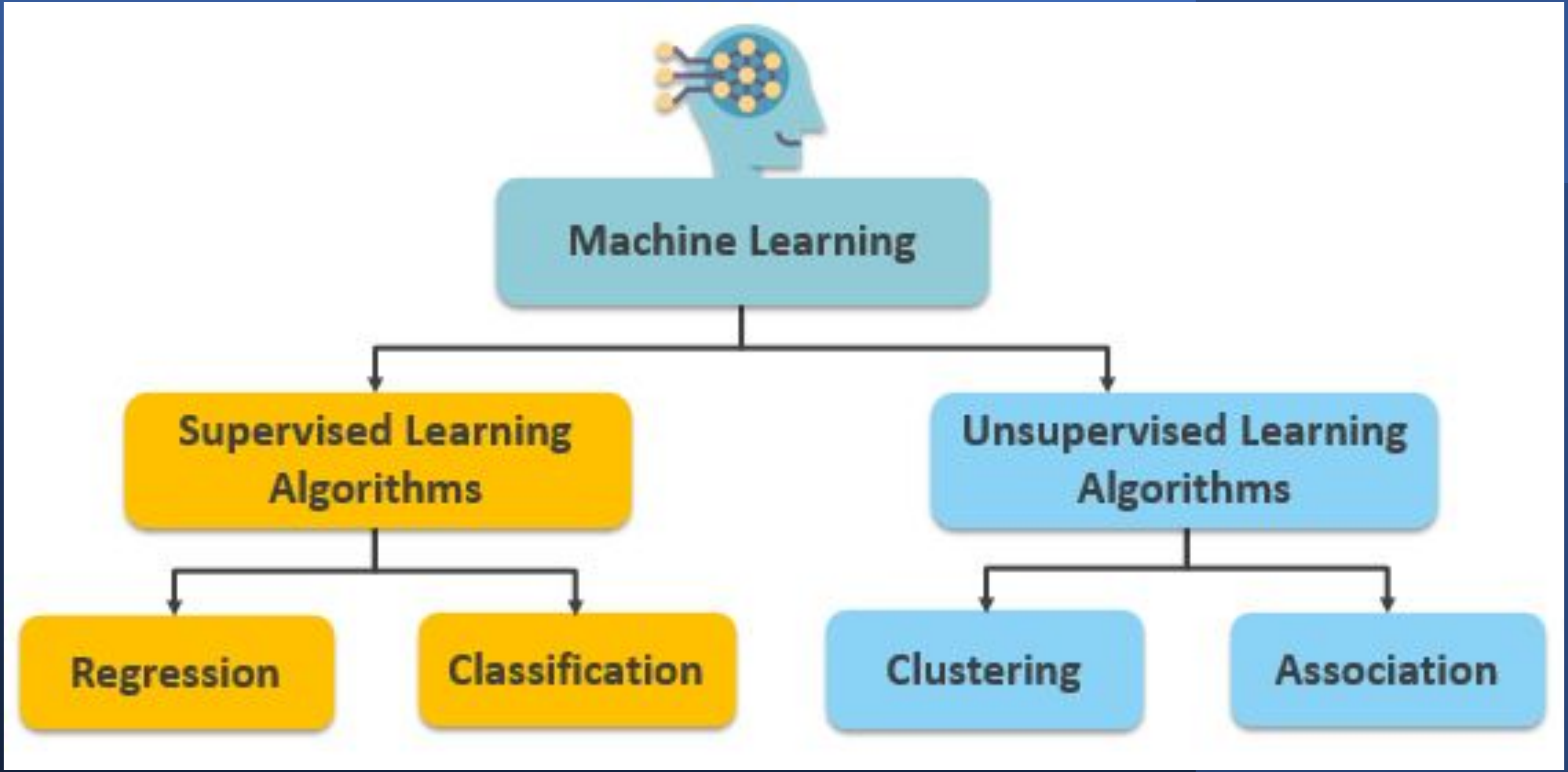
- Unknown Output
- No Training Data Set

Algorithm

Processing

MODEL TRAINING

OUTPUT


Machine Learning
Supervised Learning Algorithms
Unsupervised Learning Algorithms
Regression
Classification
Clustering
Association

Regression

- Regression is a type of supervised learning that is used to predict continuous values, such as house prices, stock prices, or customer churn.
- Regression algorithms learn a function that maps from the input features to the output value.

- Some common regression algorithms include:
  - Linear Regression
  - Polynomial Regression
  - Support Vector Machine Regression
  - Decision Tree Regression
  - Random Forest Regression

Classification

- Classification is a type of supervised learning that is used to predict categorical values, such as whether a customer will churn or not, whether an email is spam or not, or whether a medical image shows a tumor or not.
- Classification algorithms learn a function that maps from the input features to a probability distribution over the output classes.

- Some common classification algorithms include:
  - Logistic Regression
  - Support Vector Machines
  - Decision Trees
  - Random Forests
  - Naive Baye