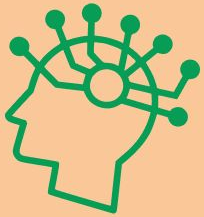


# welcome to
# ML Study Jam
## session #2

Google Developer Student Clubs
Simon Fraser University

# # Timeline

| 20 June | 27 June | 4 July | 11 July |
| --- | --- | --- | --- |
| **Session 1** | **Session 2** | **Session 3** | **Session 4** |
| Intro to Machine Learning | Feature Engineering | ML Application | Practice Project |
| Data Processing | Intro To Deep Learning | Computer Vision | |

# Quick Recap

# ML++

# Feature Engineering

# Deep Learning

# Quick Recap

# ML++

# Feature Engineering

# Deep Learning

# # Quick Recap 1

- Core Concepts and Terminologies:
  - Features: Input variables or attributes used to make predictions.
  - Labels: The output or target variable to predict.
  - Training Data: Labeled data used to train the ML model.
  - Testing Data: Unlabeled data used to evaluate the ML model's performance.
  - Prediction: Making an output or decision based on input data.
- Types of Machine Learning:
  - Supervised Learning: Learning from labeled data with input-output pairs.
  - Unsupervised Learning: Learning from unlabeled data to discover patterns or structures.
  - Reinforcement Learning: Learning through trial and error interactions with an environment.

# # Quick Recap 2

- Machine Learning Algorithms:
  - Regression: Predicting continuous numeric values.
  - Classification: Assigning data to predefined categories or classes.
  - Clustering: Grouping similar data points based on their characteristics.
- Typical Data Processing tasks:
  - Handling missing values
  - Dealing with outliers
  - Data normalization and scaling

# Quick Recap

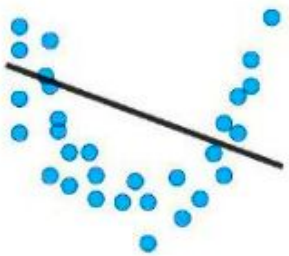
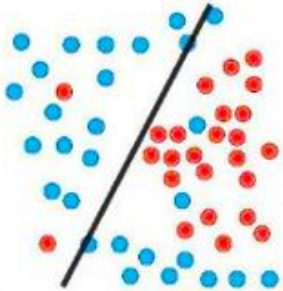
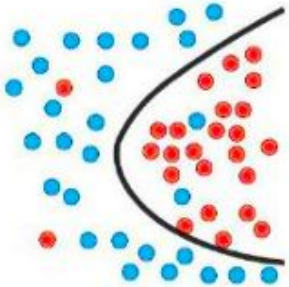
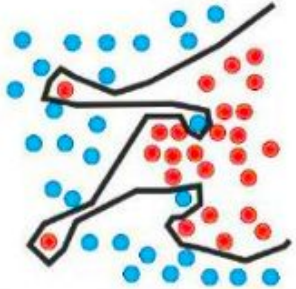
# ML++

# Feature Engineering

# Deep Learning

# # Intermediate Machine Learning

- Model Validation:
  - Assessing the performance and generalization capability of a trained model.
  - Crucial to ensure the model performs well on unseen data.
- Overfitting:
  - When a model becomes too complex and learns to fit the training data too closely.
  - Results in poor generalization to new/unseen data.
- Underfitting:
  - When a model is too simple to capture the underlying patterns in the data.
  - Results in poor performance on both training and test data.

| | Under-fitting | Optimal-fitting | Over-fitting |
|---|---|---|---|
| Regression |  |  |  |
| Classification |  |  |  |
| Deep learning |  |  |  |

# # Handling Categorical Data

- Drop Categorical Variables: simply remove categorical variables from the dataset
- Ordinal Encoding: assign each unique value to a different integer
- One-Hot Encoding: creates new columns indicating the presence (or absence) of each possible value in the original data

| fruit | fruit_label_encoded | fruit_apple | fruit_banana |
|---|---|---|---|
| apple | 0 | 1.0 | 0.0 |
| banana | 1 | 0.0 | 1.0 |
| banana | 1 | 0.0 | 1.0 |
| apple | 0 | 1.0 | 0.0 |
| banana | 1 | 0.0 | 1.0 |
| apple | 0 | 1.0 | 0.0 |

# # Data Leakage

Information from the test/validation set leaks into the training process, leading to overly optimistic results.

Types of Data Leakage:

- target leakage : when a variable that is not a feature is being used to predict the target
- train-test contamination: when we pass information from our train set to our test or vise versa

# Quick Recap

# ML++

# Feature Engineering

# Deep Learning

# # Feature Engineering

The process of creating new features or transforming existing ones to improve model performance.

Common Techniques:

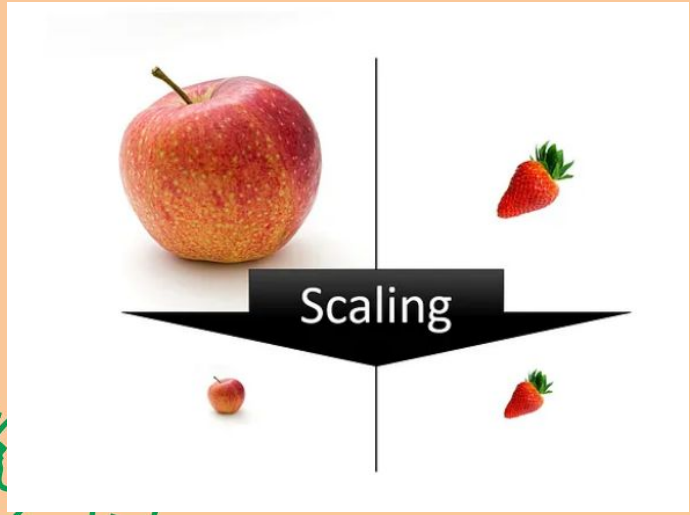
- Feature Extraction
- Feature Transformation
- Feature Selection

# Feature Extraction Techniques

- Bag-of-Words (BoW) and TF-IDF for text data.
- Principal Component Analysis (PCA) for dimensionality reduction.
- Feature encoding techniques like one-hot encoding and label encoding.

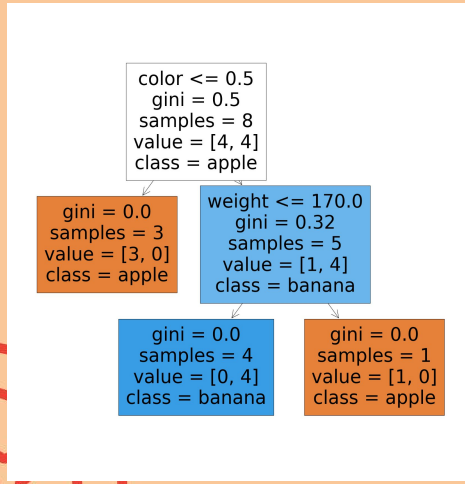
# # Feature Transformation Techniques

- Scaling and normalization for numeric features.
- Logarithmic, exponential, and power transformations.
- Binning and discretization for categorical features.



# # Feature Selection Techniques

- Univariate feature selection using statistical tests.
- Recursive Feature Elimination (RFE) based on model performance.
- L1 regularization (Lasso) for sparse feature selection.



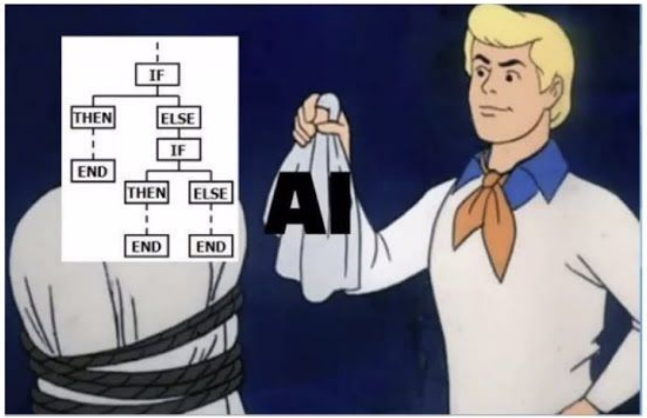
# Quick Recap

# ML++

# Feature Engineering

# Deep Learning

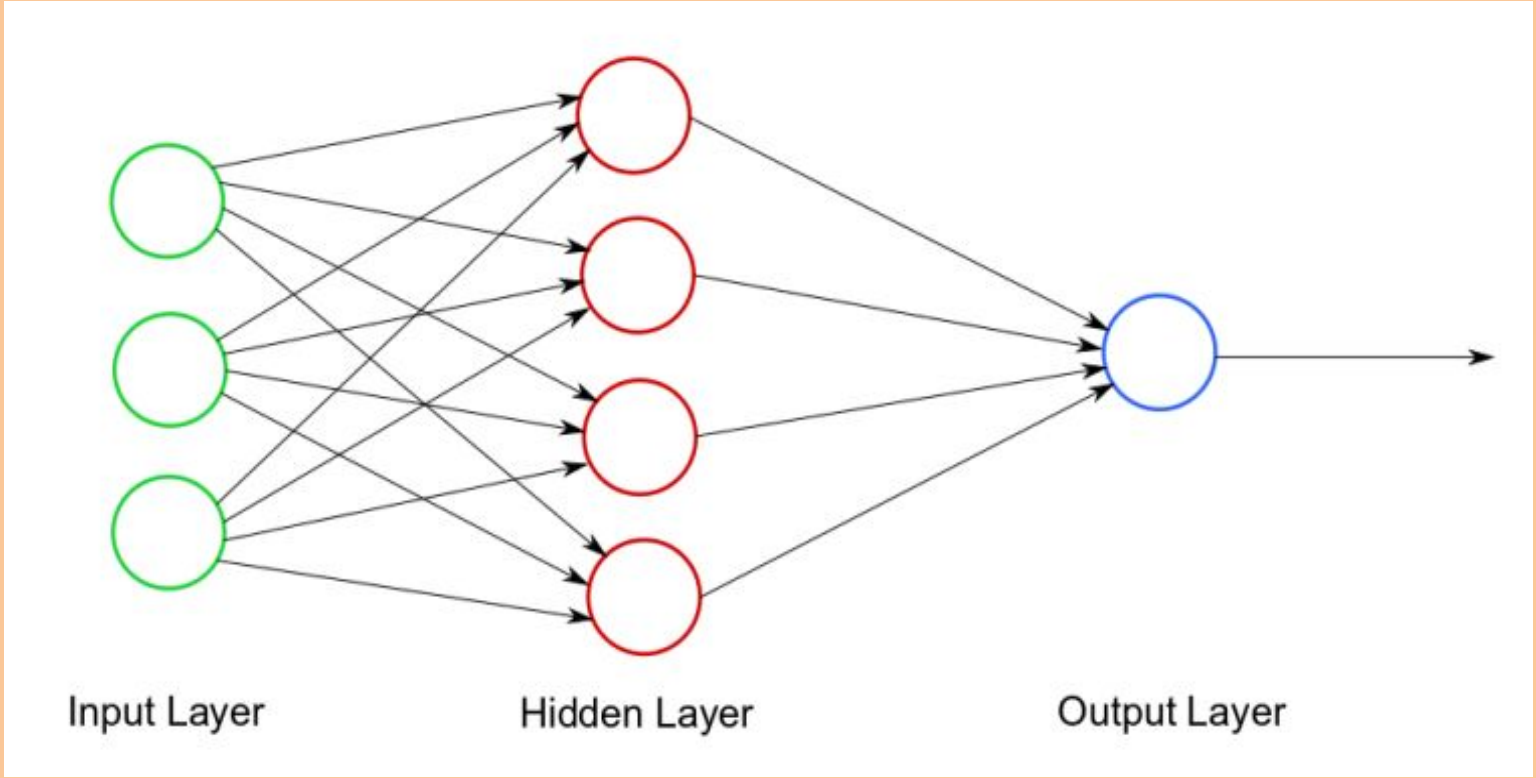
# # Intro to Deep Learning

# # Deep Learning Architecture

Structure and organization of neural networks.

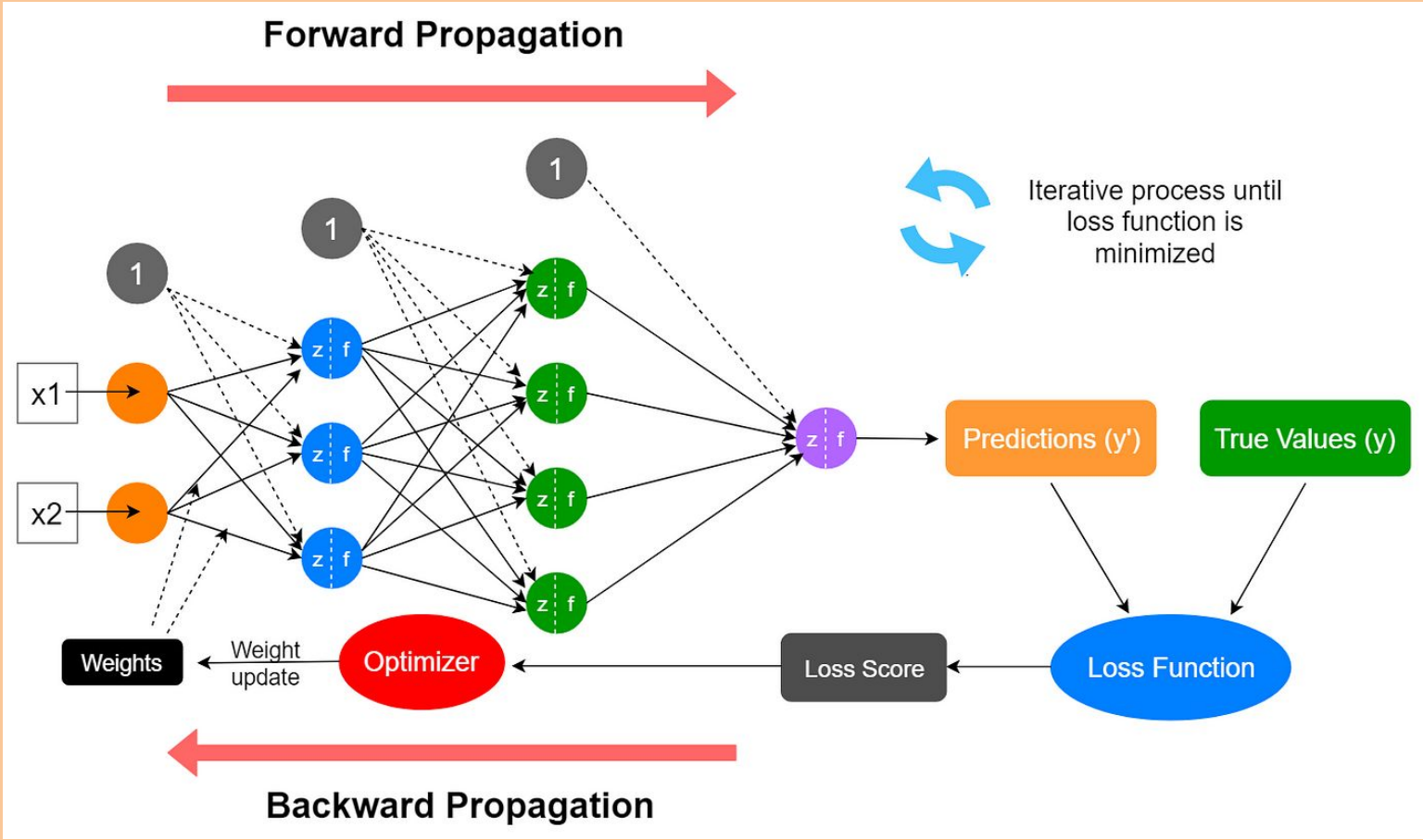
- Feedforward Neural Network:
  - Information flows in one direction, from input to output layer.
  - Common architecture for tasks like image classification and regression.
- Convolutional Neural Network (CNN):
  - Specialized for processing grid-like data, such as images.
  - Employs convolutional layers and pooling layers for feature extraction.


Input Layer
Hidden Layer
Output Layer

# # Deep Learning Process

Steps involved in building and training deep learning models.

1. Data Preparation: Collection, Preprocessing and Splitting of data
2. Model Architecture Design: Defining Neutral Network and it's parameters
3. Model Compilation: Choosing Loss, optimizer along with hyperparameters
4. Model Training: Feed training data into the model and then optimize it's weights and biases.
5. Model Evaluation: Assess model performance on validation or test set.
6. Model Fine-Tuning: Adjust hyperparameters and architecture based on performance.


Forward Propagation
1
1
1
x1
x2
z f
z f
z f
z f
z f
z f
z f
Iterative process until loss function is minimized
Predictions (y')
True Values (y)
Loss Function
Loss Score
Optimizer
Weight update
Weights
Backward Propagation

# # Exercise Time

https://colab.research.google.com/drive/15Qf76XvbDqc_IgcP9EQ8AzA7cYUTZwqn?usp=sharing

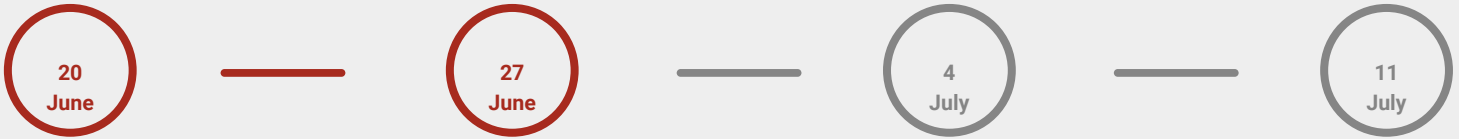
# # Kaggle Resources

- https://www.kaggle.com/learn/intermediate-machine-learning
- https://www.kaggle.com/learn/feature-engineering

# Thank you!

Google Developer Student Clubs
Simon Fraser University

# Timeline




**Session 1**

Intro to Machine Learning

Data Processing

**Session 2**

Feature Engineering

Intro To Deep Learning

**Session 3**

ML Application

Computer Vision

**Session 4**

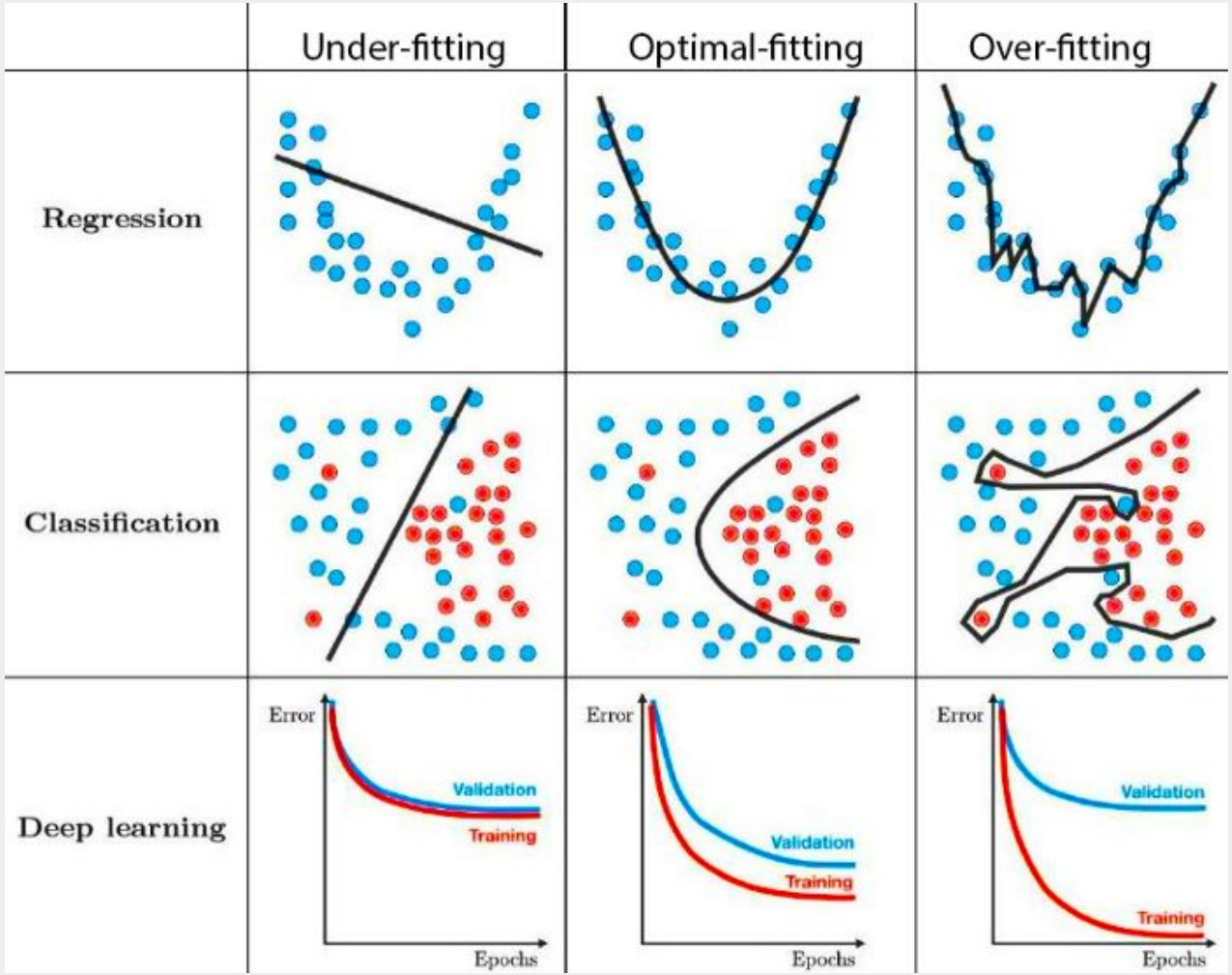
Practice Project

# Quick Recap

- Core Concepts and Terminologies:
  - Features: Input variables or attributes used to make predictions.
  - Labels: The output or target variable to predict.
  - Training Data: Labeled data used to train the ML model.
  - Testing Data: Unlabeled data used to evaluate the ML model's performance.
  - Prediction: Making an output or decision based on input data.
- Types of Machine Learning:
  - Supervised Learning: Learning from labeled data with input-output pairs.
  - Unsupervised Learning: Learning from unlabeled data to discover patterns or structures.
  - Reinforcement Learning: Learning through trial and error interactions with an environment.
- Machine Learning Algorithms:
  - Regression: Predicting continuous numeric values.
  - Classification: Assigning data to predefined categories or classes.
  - Clustering: Grouping similar data points based on their characteristics.
- Typical data Processing tasks:
  - Handling missing values
  - Dealing with outliers
  - Data normalization and scaling

# Intermediate Machine Learning

- Model Validation:
  - Assessing the performance and generalization capability of a trained model.
  - Crucial to ensure the model performs well on unseen data.
- Overfitting:
  - When a model becomes too complex and learns to fit the training data too closely.
  - Results in poor generalization to new/unseen data.
- Underfitting:
  - When a model is too simple to capture the underlying patterns in the data.
  - Results in poor performance on both training and test data.


Under-fitting
Optimal-fitting
Over-fitting
Regression
Classification
Deep learning
Error
Epochs
Validation
Training
Error
Epochs
Validation
Training
Error
Epochs
Validation
Training

# Handling Categorical Data

- Drop Categorical Variables: simply remove categorical variables from the dataset
- Ordinal Encoding: assign each unique value to a different integer
- One-Hot Encoding: creates new columns indicating the presence (or absence) of each possible value in the original data

| fruit | fruit_label_encoded | fruit_apple | fruit_banana |
|---|---|---|---|
| apple | 0 | 1.0 | 0.0 |
| banana | 1 | 0.0 | 1.0 |
| banana | 1 | 0.0 | 1.0 |
| apple | 0 | 1.0 | 0.0 |
| banana | 1 | 0.0 | 1.0 |
| apple | 0 | 1.0 | 0.0 |

# Data Leakage

Information from the test/validation set leaks into the training process, leading to overly optimistic results.

Types of Data Leakage:

- **target leakage** : when a variable that is not a feature is being used to predict the target
- **train-test contamination: when we** pass information from our train set to our test or vise versa

# Feature Engineering

The process of creating new features or transforming existing ones to improve model performance.

Common Techniques:

- Feature Extraction
- Feature Transformation
- Feature Selection

Feature Extraction Techniques:

- Bag-of-Words (BoW) and TF-IDF for text data.
- Principal Component Analysis (PCA) for dimensionality reduction.
- Feature encoding techniques like one-hot encoding and label encoding.

Feature Transformation Techniques:

- Scaling and normalization for numeric features.
- Logarithmic, exponential, and power transformations.
- Binning and discretization for categorical features.

Feature Selection Techniques:

- Univariate feature selection using statistical tests.
- Recursive Feature Elimination (RFE) based on model performance.
- L1 regularization (Lasso) for sparse feature selection.

# Intro to Deep Learning

Machine learning that focuses on artificial neural networks

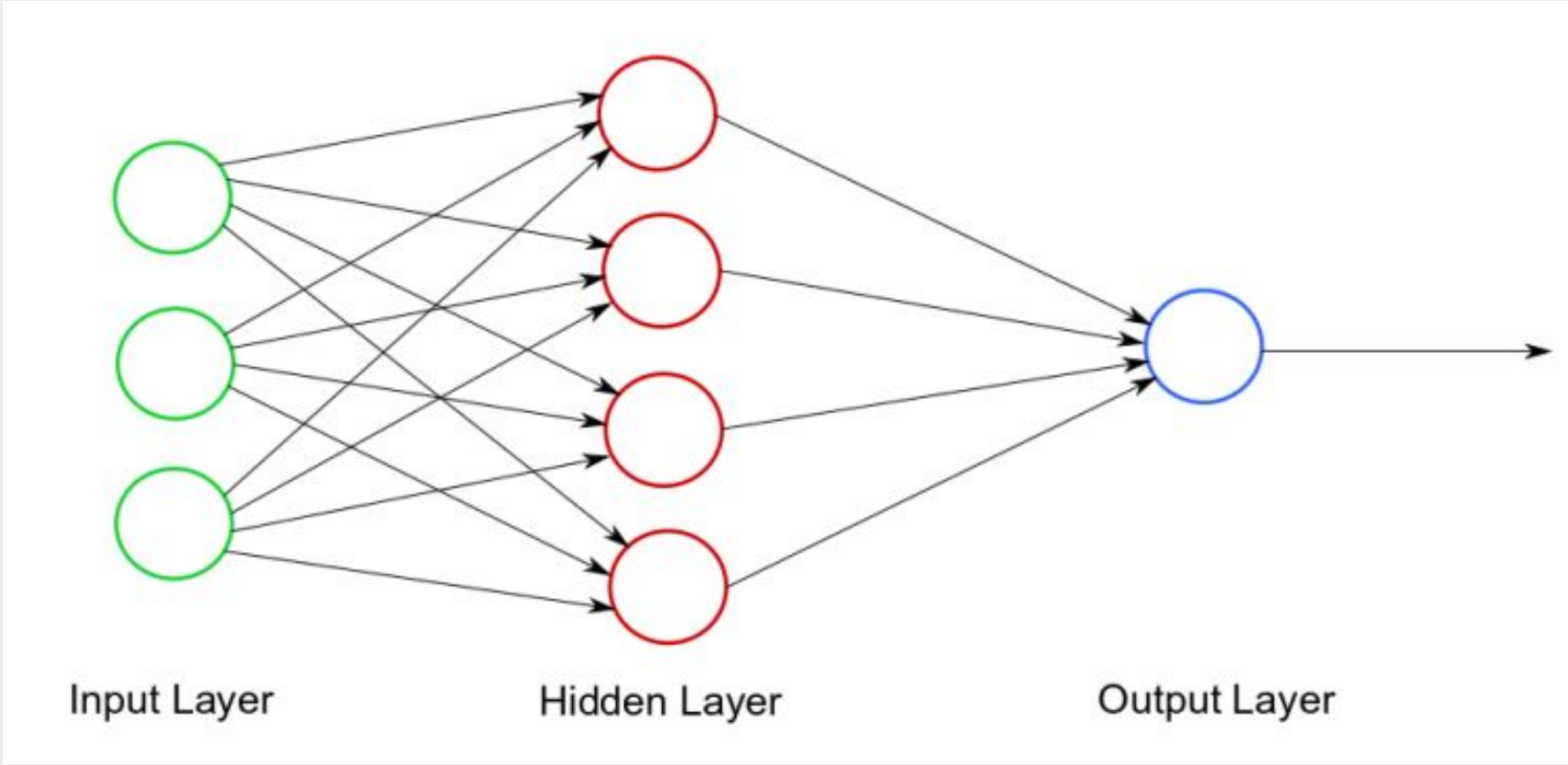
Key Components:

- Neural Networks: Interconnected layers of artificial neurons.
- Deep Neural Networks: Networks with multiple hidden layers.
- Activation Functions: Introduce non-linearity to neural network computations.
- Backpropagation: Algorithm for updating weights and biases during training.

# Deep Learning Architecture

Structure and organization of neural networks.

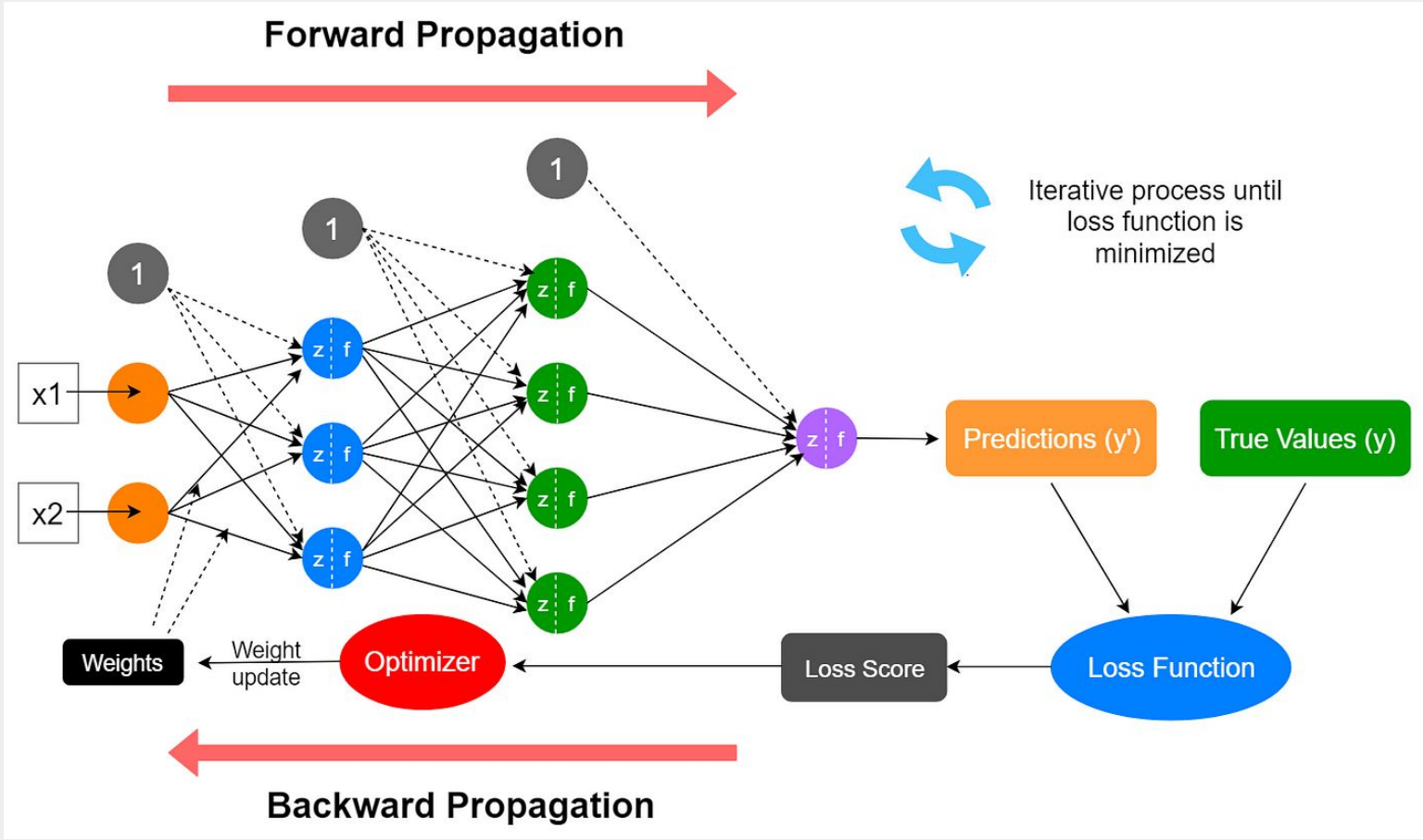
- Feedforward Neural Network:
  - Information flows in one direction, from input to output layer.
  - Common architecture for tasks like image classification and regression.
- Convolutional Neural Network (CNN):
  - Specialized for processing grid-like data, such as images.
  - Employs convolutional layers and pooling layers for feature extraction.

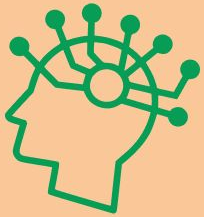

Input Layer
Hidden Layer
Output Layer

# Deep Learning Process

Steps involved in building and training deep learning models.

1. Data Preparation: Collection, Preprocessing and Splitting of data
2. Model Architecture Design: Defining Neutral Network and it’s parameters
3. Model Compilation: Choosing Loss, optimizer along with hyperparameters
4. Model Training: Feed training data into the model and then optimize it’s weights and biases.
5. Model Evaluation: Assess model performance on validation or test set.
6. Model Fine-Tuning: Adjust hyperparameters and architecture based on performance.


Forward Propagation
1
1
1
x1
x2
z f
z f
z f
z f
z f
z f
z f
Iterative process until loss function is minimized
Predictions (y')
True Values (y)
Loss Function
Loss Score
Optimizer
Weight update
Weights
Backward Propagation



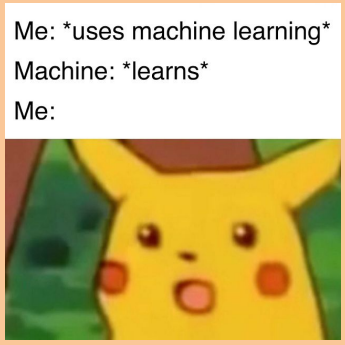
# welcome to ML Study Jam session #1

Google Developer Student Clubs
Simon Fraser University

topic subtitle sample text

# What is Machine Learning?

- Definition: Machine Learning is a subset of Artificial Intelligence that enables computers to learn and make predictions or decisions without being explicitly programmed
- ML enables systems to automatically analyze and extract patterns from data

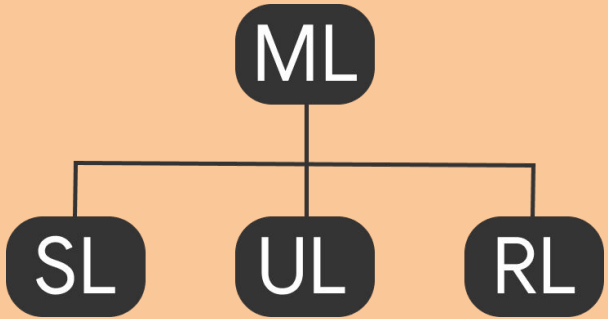


# # Core Concepts and Terminologies

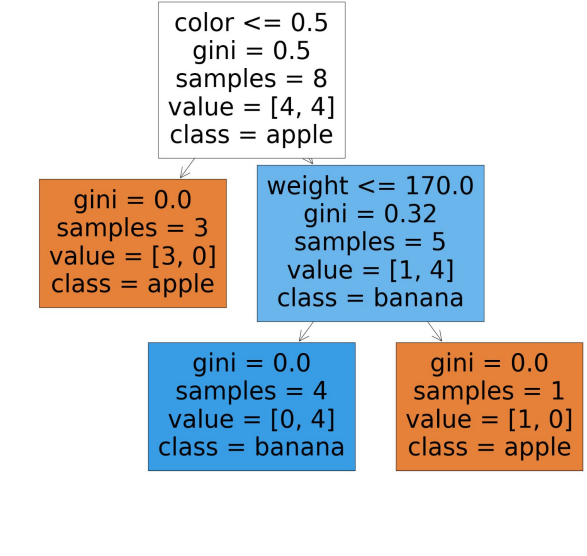
- Features: Input variables or attributes used to make predictions.
- Labels: The output or target variable to predict.
- Training Data: Labeled data used to train the ML model.
- Testing Data: Unlabeled data used to evaluate the ML model's performance.
- Prediction: Making an output or decision based on input data.

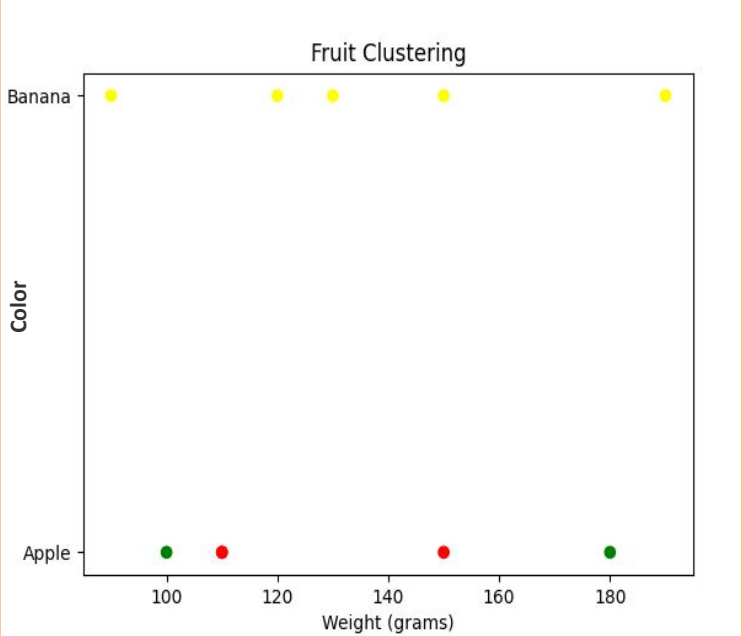
# # Types of Machine Learning

- Supervised Learning: Learning from labeled data with input-output pairs.
- Unsupervised Learning: Learning from unlabeled data to discover patterns or structures.
- Reinforcement Learning: Learning through trial and error interactions with an environment.



# # Supervised vs Unsupervised Learning



topic subtitle sample text

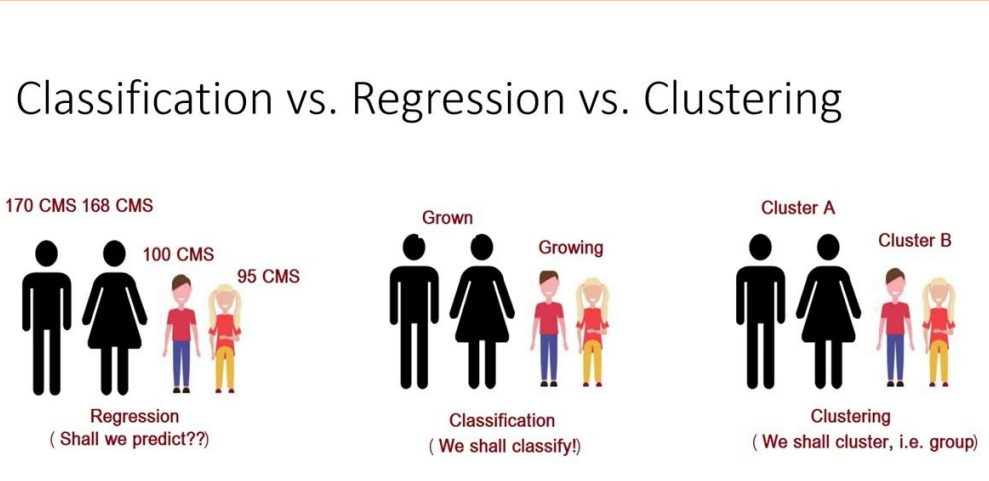
# Intro to Machine Learning

# ML Algorithms

# Data Processing

# # Machine Learning Algorithms

- Regression: Predicting continuous numeric values.
- Classification: Assigning data to predefined categories or classes.
- Clustering: Grouping similar data points based on their characteristics.



# # Popular Machine Learning Algorithms

- Linear Regression: Predicting a continuous value based on linear relationships.
- Logistic Regression: Classifying data into discrete categories using a logistic function.
- Decision Trees: Creating a tree-like model for classification or regression.
- K-means Clustering: Grouping data points into clusters based on similarity.

topic subtitle sample text

# Intro to Machine Learning

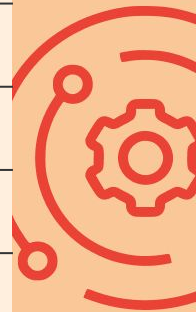
# ML Algorithms

# Data Processing

# # Data Processing

- Manipulation and transformation of raw data to make it suitable for analysis and modeling.
- Typical data Processing tasks:
  - Handling missing values
  - Dealing with outliers
  - Data normalization and scaling
- Crucial for improving the quality and performance of ML models.

| Color (Feature) | Weight (Feature ) | Fruit Name (Label) |
|---|---|---|
| Red | 150 | Apple |
| Yellow | 120 | Banana |
| Red | | Apple |
| Green | | Apple |
| 20/June/2023 | | Banana |
| | 150 | Apple |
| Yellow | 0.11 kg | Banana |
| 1 | 90 | Orange |
| Yellow | 100 | Banana |
| Red | 0.42 lb | Apple |



# # Data Processing Workflow

- Data loading and inspection.
- Data cleaning and preprocessing.
- Feature engineering and selection.
- Data transformation and normalization.
- Splitting data into training and testing sets.

# # Data Processing with Pandas

- Powerful library for data manipulation and analysis in Python.
- It provides data structures and functions for efficient data processing.
- Common data processing tasks, including data cleaning, filtering, and transformation.

# # Numerical Computing with NumPy

- Fundamental library for numerical computing in Python.
- Features and benefits:
  - N-dimensional array objects (ndarrays) for efficient storage and manipulation of large datasets.
  - Mathematical operations and functions for array-based computations.
  - Linear algebra, Fourier transforms, and random number generation capabilities.
- Provides a solid foundation for numerical operations in Machine Learning.