UW
DSC.

# Clustering Workshop

Slideshow: https://bit.ly/3bDy1SW

Notebook: https://bit.ly/3bIaWP3

Presented by Jack Douglas

# Introduction

**Jack Douglas**

(he/him)

2B Software Engineering

Email: jack.douglas@uwaterloo.ca

LinkedIn: https://www.linkedin.com/in/jack-douglas-910896150/

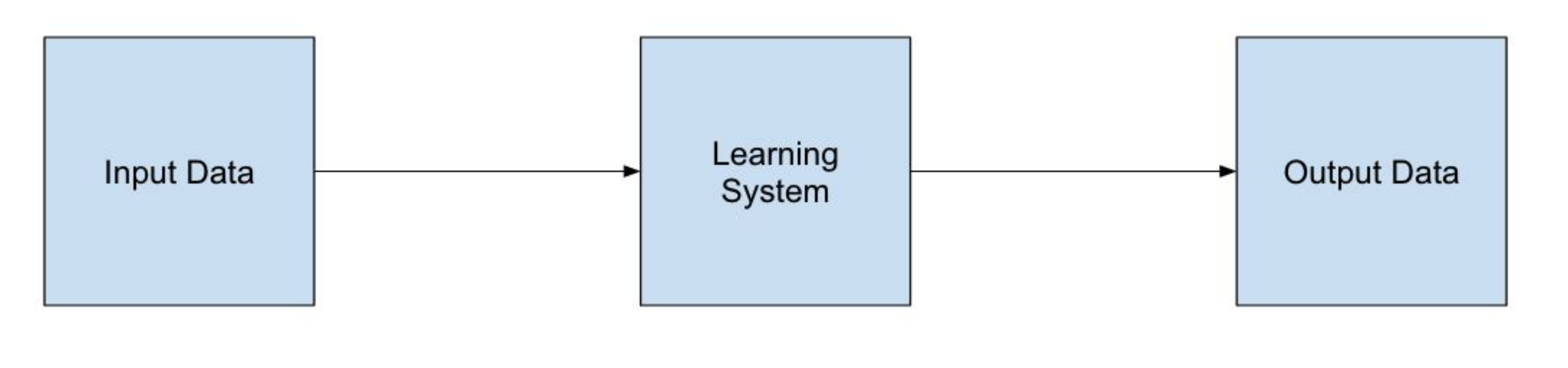
# Workshop Overview

1. Machine Learning Refresher
2. What is Clustering?
3. Types of Clustering
4. Clustering Algorithms
5. Real-World Clustering Applications
6. Notebook Demo
7. Questions

# What is Machine Learning?

UW DSC.

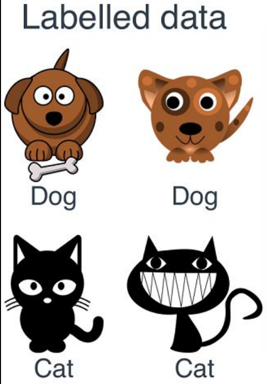
**Def'n:** "Machine learning is a branch of artificial intelligence focused on building applications that learn from data and improve their accuracy over time without being explicitly programmed to do so."



# Supervised vs. Unsupervised Learning

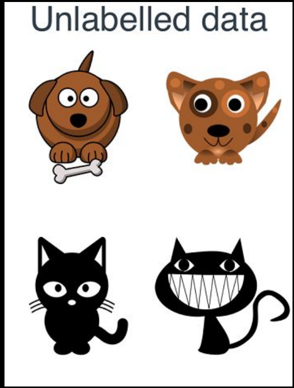
<u>Supervised</u>

**Def'n:** Refers to the technique where a model trains on *labelled* data to determine the relationship between the data (input) and its label (output).




<u>Unsupervised</u>

**Def'n:** Refers to the technique where a model trains on *unlabelled* data to determine the inherent structure of the data.



# What is Clustering?

UW DSC.

- **Def'n:** The task of grouping data in such a way that objects in the same group are more similar to each other than those in other groups
  - Each group is called a **cluster**
  - Data within a cluster is similar, each cluster has different features to each other

- **Goal:** Determine a pattern in the data using clusters such that new data can be assigned to a cluster based on this pattern

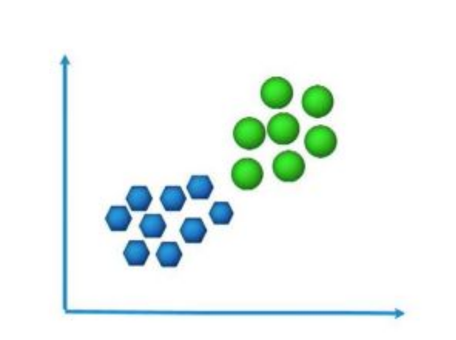
- Common unsupervised learning technique

# Types of Clustering

UW DSC.

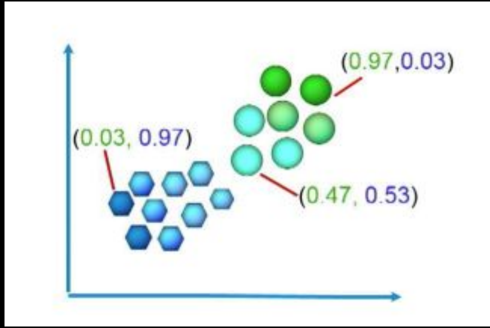
## Hard Clustering

**Def'n:** Each data point belongs to exactly one cluster.



## Soft (Fuzzy) Clustering

**Def'n:** Each data point can belong to more than one cluster to a certain degree (ie. likelihood of belong to the cluster)




Learn more about soft clustering here: https://en.wikipedia.org/wiki/Fuzzy_clustering

# Clustering Algorithms

UW DSC.

- Many different types of clustering algorithms which have different use cases depending on the data
  - Centroid-based, connectivity-based, density-based, distribution-based, etc.

- Clustering is subjective, so each clustering algorithm follows its own process for defining “similarity” between data points

# Centroid-based Clustering

UW
DSC.
_

**Basis:** The notion of similarity is derived by the *closeness* of a data point to the *centroid* of the clusters

- The **number of clusters** must be specified **beforehand**

- Centroid-based clustering models run **iteratively** to find the local optima

# Definitions and Calculations

UW DSC. –

1. *Centroid*: Central point of each cluster

$$g\left((x_1,y_1),(x_2,y_2),\ldots,(x_n,y_n)\right)=\left(\frac{\sum_n x_i}{n},\frac{\sum_n y_i}{n}\right)$$

2. Metrics of closeness
   a. *Euclidean distance:* Length of the line segment between two points

$$f\left((x_1,y_1),(x_2,y_2)\right)=\sqrt{(x_1-x_2)^2+(y_1-y_2)^2}$$

   b. Other metrics can be found here: https://www.datanovia.com/en/lessons/clustering-distance-measures/

# K-Means Clustering

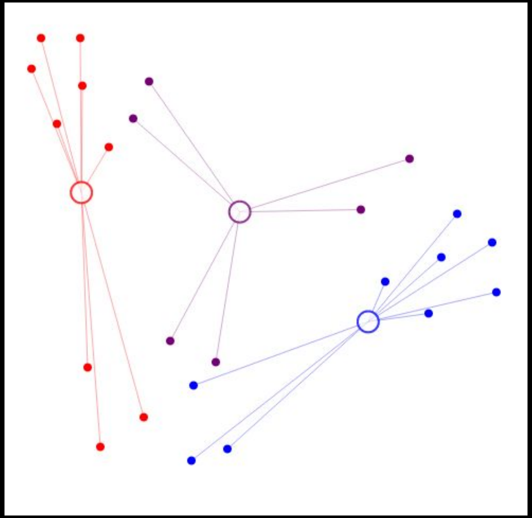
**Process:**

1. Specify the desired number of clusters, **K**
2. Randomly place **K** centroids within the data
3. Assign each data point to its nearest cluster
4. Compute cluster centroids
5. Reassign each data point to its closest cluster centroid
6. Repeat steps 4 and 5 until no improvements are possible

# K-Means Clustering Visualization

UW DSC.

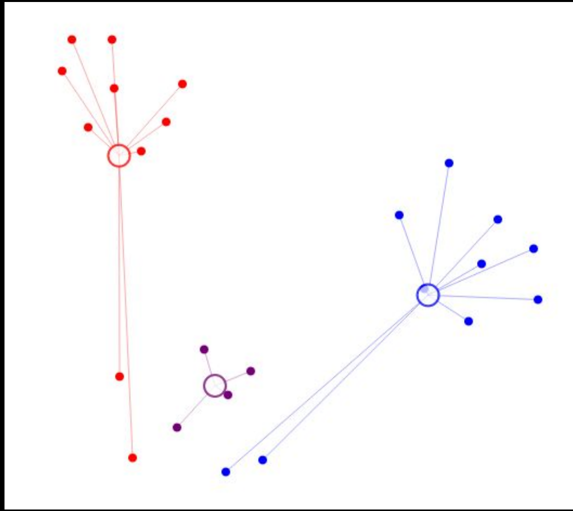
K = 3, Iteration = 0



- Centroids are place randomly
- Data is assigned to closest centroids

# K-Means Clustering Visualization

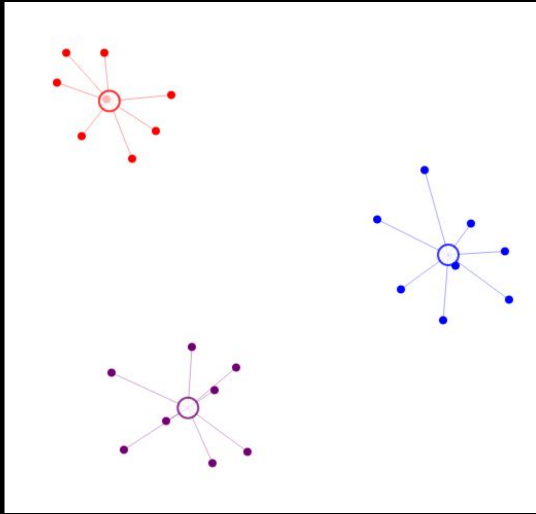
K = 3, Iteration = 1



- Centroids are recalculated within each cluster
- Data is re-assigned to closest centroids

# K-Means Clustering Visualization

K = 3, Iteration = 2



- Centroids are recalculated within each cluster
- Data is re-assigned to closest centroids

# K-Means Clustering Visualization

UW
DSC.
–

https://user.ceng.metu.edu.tr/~akifakkus/courses/ceng574/k-means/

# K-Means Characteristics

- Most common clustering algorithm
- Fast due to the small number of calculations
- Limited by having to specify number of clusters
- Results can differ between runs of the algorithm

# Connectivity-based Clustering

UW
DSC.
–

**Basis:** The notion of similarity is derived by the *closeness* of data points to each other

- Similar to centroid-based clustering, *closeness* is subjective and there are many metrics that can be used
  - Note: the average linkage function is often used for connectivity-based clustering

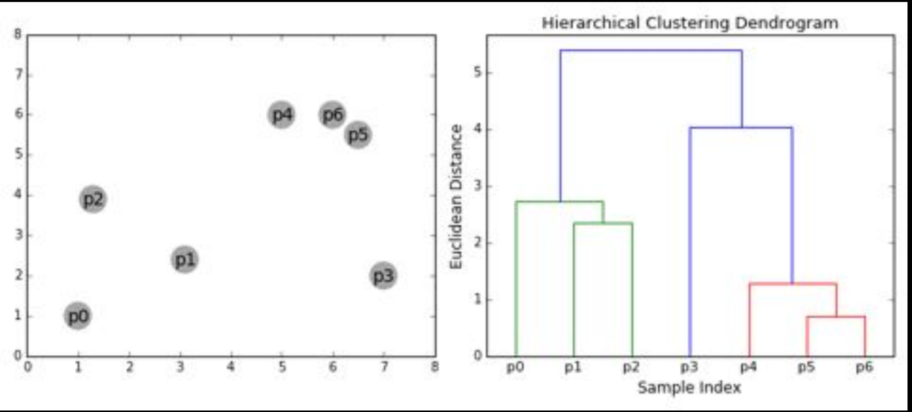
- There are two approaches: **bottom up** and **top down**
  - **Bottom up:** Every data point start as a cluster and they aggregate as distance *decreases*
  - **Top down:** All the data points start in *one* cluster  and partition as distance *increases*

# Hierarchical Agglomerative Clustering (HAC)

UW
DSC.
–

**Process:**

1. Treat all data points as their own single point clusters
2. Determine the two clusters which are the *closest* based on the distance metric have chosen
3. Combine the two clusters
4. Repeat steps 2 and 3 until we have one cluster containing all data



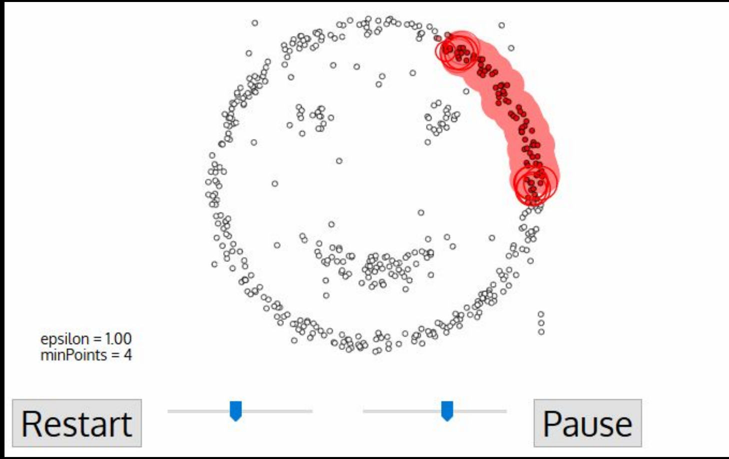
# HAC Characteristics

- Can select the best number of clusters since we are building a tree
- Not sensitive to the choice of the distance metric
- Useful when data has underlying hierarchical structure
- Comes at the cost of low efficiency and high time complexity, O(n^3)

# Density-based Clustering

**Basis:** The notion of similarity is derived by the varied density of data points

- Specify the maximum distance (epsilon) for points to be called *neighbours*
- Specify the minimum number of points in a neighbourhood to remove noise
- Every neighbourhood become a cluster, algorithm finishes once every point is visited
- More info here: https://bit.ly/3rLZk2Q




DBSCAN Visualization: https://www.naftaliharris.com/blog/visualizing-dbscan-clustering/

UW
DSC.

# Distribution-based Clustering

**Basis:** The notion of similarity is derived by how probable it is that all data points in the cluster belong to the same distribution

- This approach assumes data is composed of distributions, such as Gaussian distributions
- Must specify the number of clusters and randomly initialize the parameters
- Calculate the weighted sum of a point being in a particular cluster and recalculate parameters
- More info here: [https://bit.ly/3rLZk2Q](https://bit.ly/3rLZk2Q)

Expectation-Maximization Algorithm

# Application: Marketing Research

UW
DSC.
–

- **Goal:** Partition consumers into market segments and to better understand potential customers

- **Types of Data:**
  - Demographic
  - Geographic
  - Behavioural

- **Use Cases:**
  - Customizing advertising campaigns
  - Designing products

# Application: Image Segmentation

- **Goal:** Divide images into distinct regions

- **Types of Data:**
  - Pixels/Images

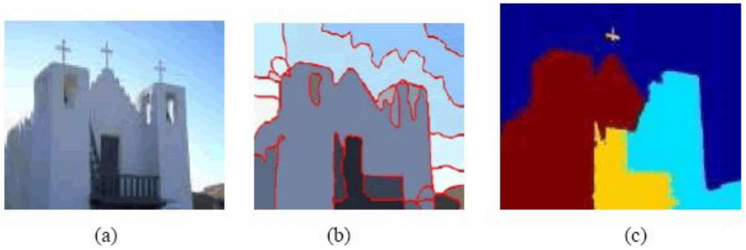



**Figure 1:** (a) is the original image; (b) and (c) are the segmentation results.

- **Use Cases:**
  - Border detection
  - Object recognition

# Application: Document Classification

UW DSC.

- **Goal:** Classify the type of a document

- **Types of Data:**
  - Document content
  - Topics
  - Tags

- **Use Cases:**
  - Categorizing documents
  - Finding similar documents

# Notebook Demo

Notebook: https://bit.ly/3bIaWP3

# Questions?