Special Lecture

The ML Pipeline/Concept Map

DATA	FEATURES	FEATURE XFORMATIONS	ML PROBLEM	ALGORITHMS	PERFORM. METRICS
• **Structured** (NUM, CAT, ATTR) • **Digital Logs** (Tweets, SMS) • **RawData/ Sensors** (IMG/Speech) • **Others** User Behavior, etc.	• Intuitive User defined • Raw data itself • Statistics (Histograms, PCA) • Signal Processing (Fourier Xform)	•Feature Selection •Feature Extraction •Dimensionality Reduction Eg. PCA	1. Classification a. Binary b. Multiclass 2. Regression 3. Clustering 4. Prediction (time series)	1. KNN 2. Naïve Bayes 3. Perceptron 4. Linear 5. Decision Tree	• Accuracy • Confusion Matrix • Precision • Recall • AP • True Positive, etc.

Different types of data representation

Raw data

3072 X 1 Vector

Feature Vector
32 X 32 X 3 = 3072 Dimension
Per Image (d=3072)

CONCERNS:
- Too big?
- May be redundancy?

Hand Crafting Features

MIN RED
MAX RED
MEAN RED
MIN GREEN
MAX GREEN
MEAN GREEN
MIN BLUE
MAX BLUE
MEAN BLUE

9 X 1
FEATURE VECTOR
PER IMAGE

Concerns:

- **Too naïve to capture the visual content?**
- **Too small to represent information?**

Deep Learning Features

Deep Learning = End to End Learning (Raw data to labels)
Deep Learning = Feature Learning!!

Raw Image

Initial Stages of the Deep Neural Networks
Many linear and nonlinear operations

Final Stages

Classifier

1000 Labels For a 1000 class classification

An intermediate representation from a popular "DeepNet", which was designed and trained for solving a "general" 1000 class classification.

Training and Testing

Creating and Evaluating Models

The ML Pipeline/Concept Map

DATA

- **Structured** (NUM, CAT, ATTR)
- **Digital Logs** (Tweets, SMS)
- **RawData/ Sensors** (IMG/Speech)
- **Others** User Behavior, etc.

FEATURES

- Intuitive User defined
- Raw data itself
- Statistics (Histograms, PCA)
- Signal Processing (Fourier Xform)

FEATURE XFORMATIONS

- Feature Selection
- Feature Extraction
- Dimensionality Reduction Eg. PCA

ML PROBLEM

1. Classification
   a. Binary
   b. Multiclass
2. Regression
3. Clustering
4. Prediction (time series)

ALGORITHMS

1. KNN
2. Naïve Bayes
3. Perceptron
4. Linear
5. Decision Tree

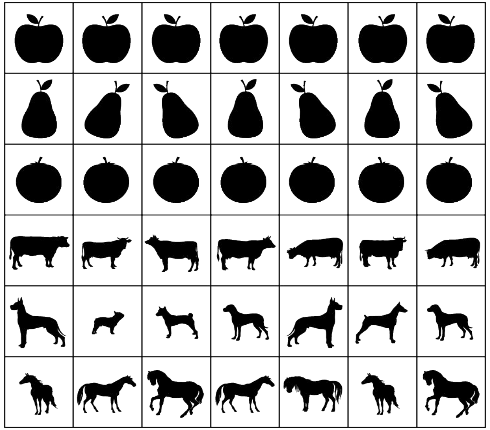
PERFORM. METRICS

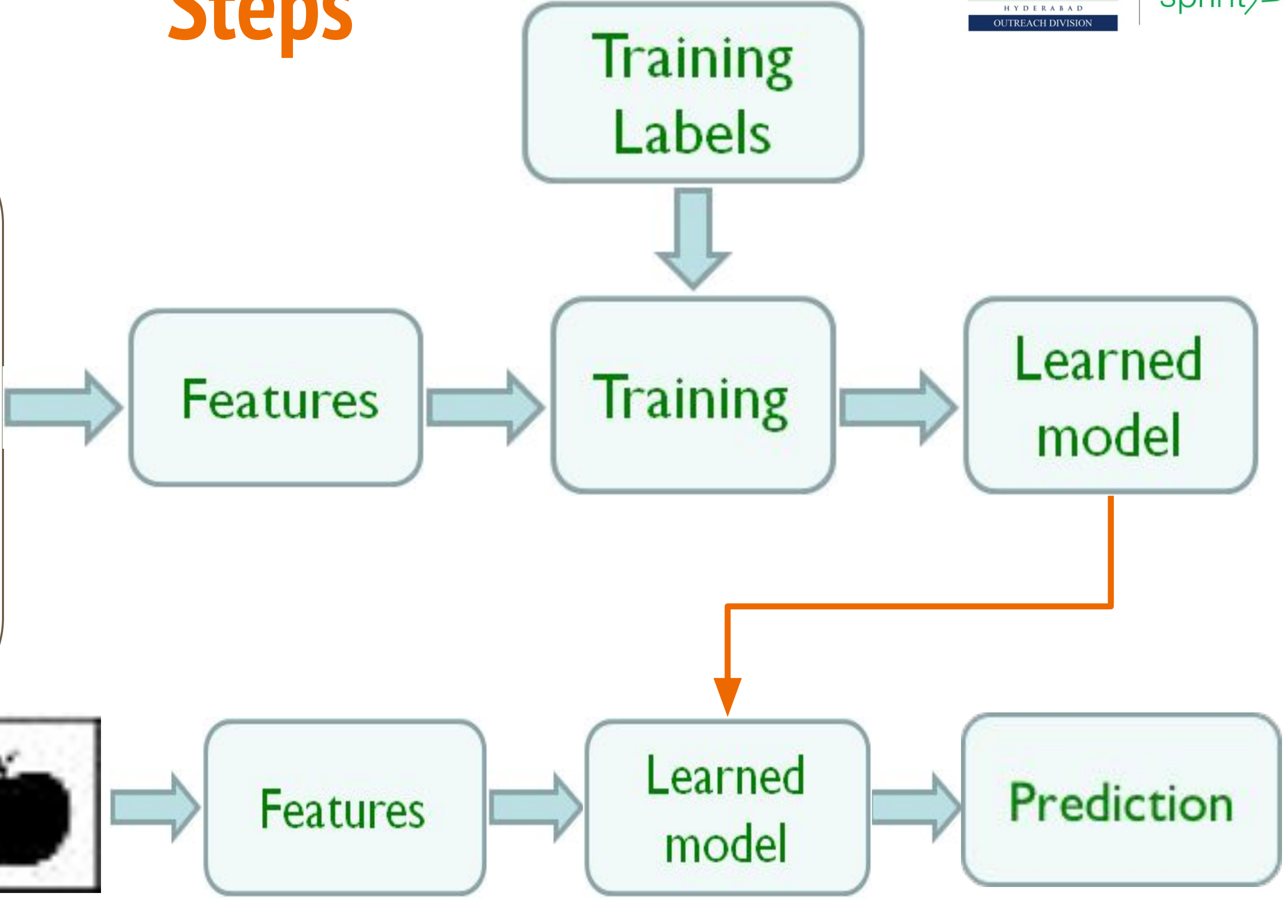
- Accuracy
- Confusion Matrix
- Precision
- Recall
- AP
- True Positive, etc.

Steps

- Training

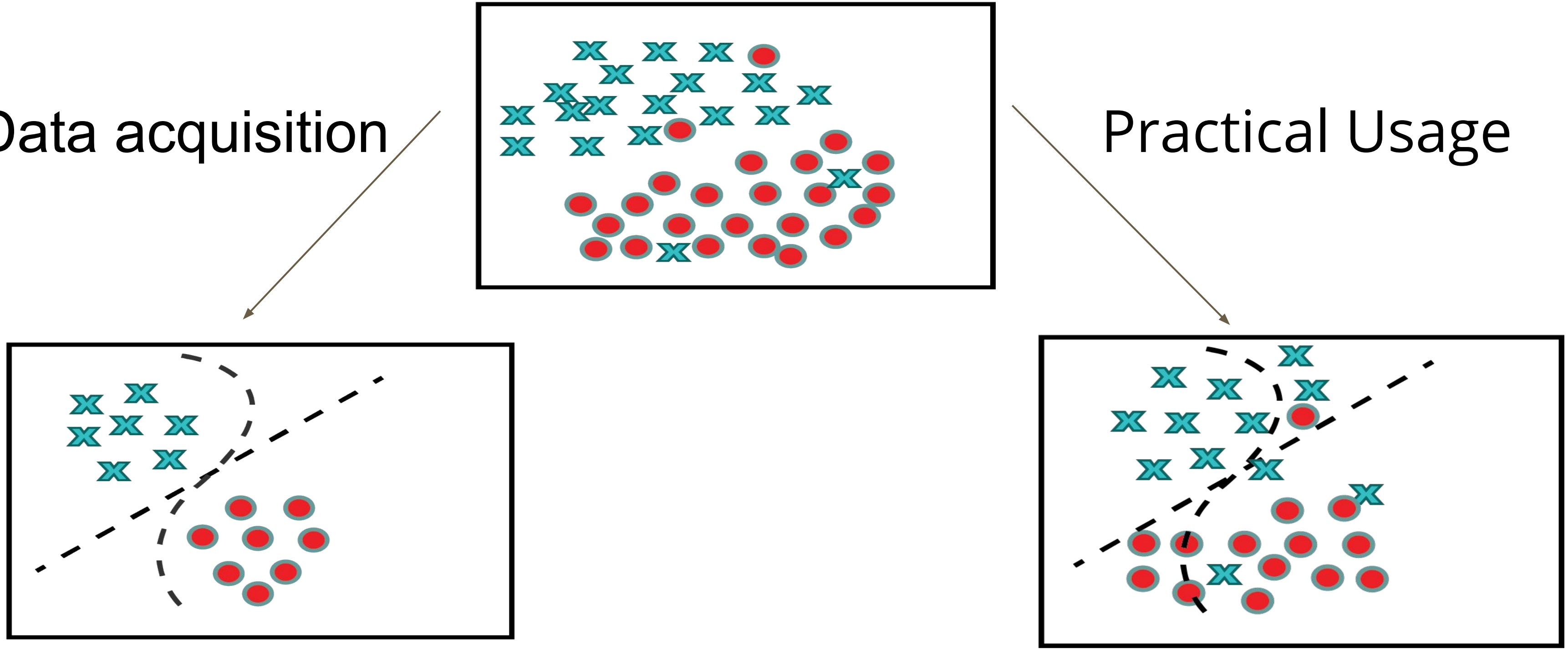
Training Data





- Testing

Training and testing

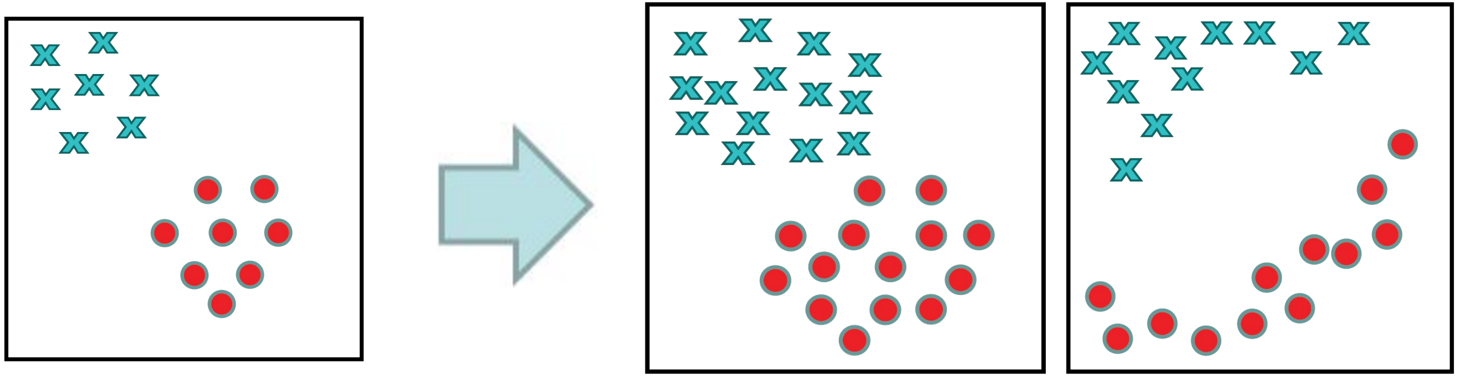


Training Set
(Observed)

Testing Set
(Unobserved)

Training and testing

- Training is the process of making the system able to learn
- Assumptions:
  - Training set and testing set come from the same distribution
  - Need to make some assumptions or bias



Performance Evaluation Metrics

Accuracy, Precision, Recall

The ML Pipeline/Concept Map

DATA
- **Structured** (NUM, CAT, ATTR)
- **Digital Logs** (Tweets, SMS)
- **RawData/ Sensors** (IMG/Speech)
- **Others** User Behavior, etc.

FEATURES
- Intuitive User defined
- Raw data itself
- Statistics (Histograms, PCA)
- Signal Processing (Fourier Xform)

FEATURE XFORMATIONS
- Feature Selection
- Feature Extraction
- Dimensionality Reduction Eg. PCA

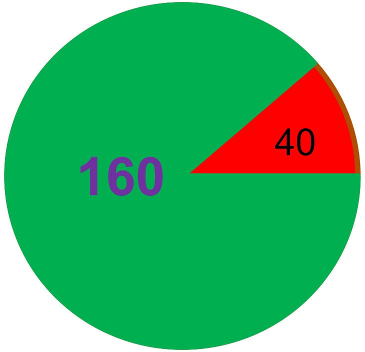
ML PROBLEM
1. Classification
   a. Binary
   b. Multiclass
2. Regression
3. Clustering
4. Prediction (time series)

ALGORITHMS
1. KNN
2. Naïve Bayes
3. Perceptron
4. Linear
5. Decision Tree

PERFORM. METRICS
- Accuracy
- Confusion Matrix
- Precision
- Recall
- AP
- True Positive, etc.

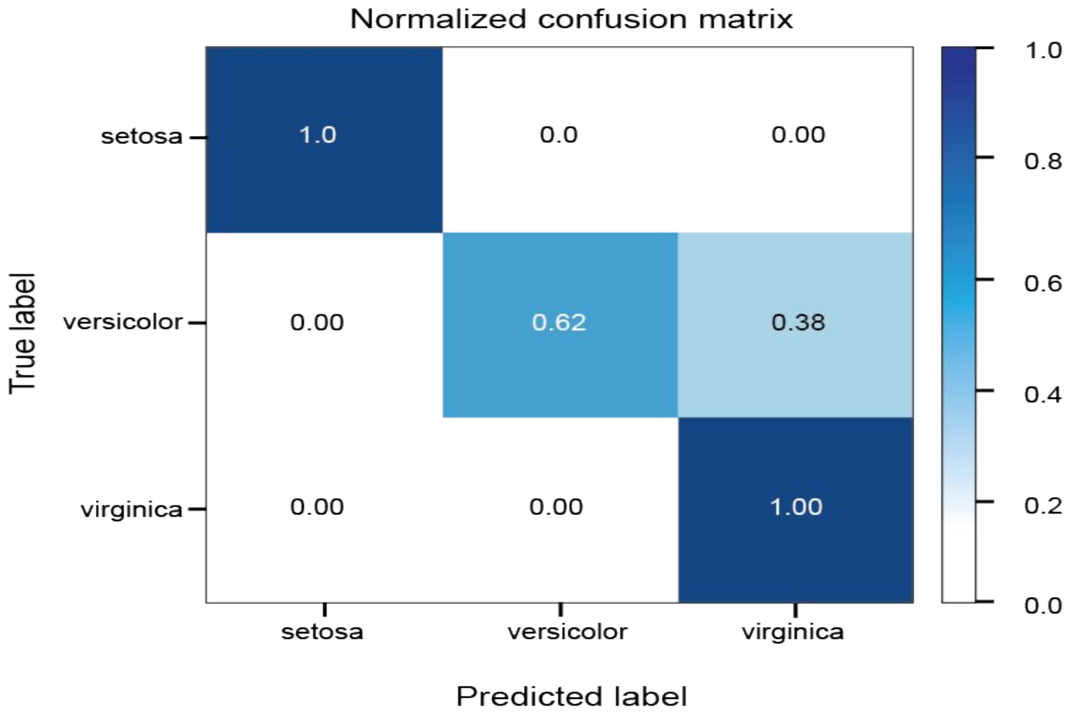
Accuracy

- Simple, Intuitive performance measure
  - The ratio of correctly predicted observations to the total number of observations
  - $\frac{Number\ of\ Correctly\ Classified\ Test\ Samples}{Total\ Number\ of\ Test\ Samples}$
  - $Accuracy = \frac{160}{200} = 0.8\ \ (80\%)$
  - Does not consider class labels



Test Samples

Confusion Matrix



A Specific Case (Binary)

$$Accuracy = \frac{(100 + 50)}{165} = 0.91$$

$$Misclassification = \frac{(10 + 5)}{165} = 0.09$$

$$TruePositiveRate(TP) = \frac{(100)}{105} = 0.95$$

$$FalsePositiveRate(FP) = \frac{(10)}{60} = 0.17$$

n=165	Predicted: NO	Predicted: YES	
Actual: NO	TN = 50	FP = 10	60
Actual: YES	FN = 5	TP = 100	105
	55	110	

A Specific Case (Binary)...

$$TrueNegativeRate(TN) = \frac{(50)}{60} = 0.833$$

$$FalseNegativeRate(FN) = \frac{(5)}{105} = 0.048$$

n=165	Predicted: NO	Predicted: YES	
Actual: NO	TN = 50	FP = 10	60
Actual: YES	FN = 5	TP = 100	105
	55	110	

A Specific Case (Binary)

- When you do cancer screening what do you care?
  - High TP
- When you classify between “apple” and “orange”
  - High Accuracy or High TP or High TN
- Automatic Firing on detecting a violation.
  - Very low FP

Clustering

Unsupervised Machine Learning

The ML Pipeline/Concept Map

DATA

- **Structured** (NUM, CAT, ATTR)
- **Digital Logs** (Tweets, SMS)
- **RawData/ Sensors** (IMG/Speech)
- **Others** User Behavior, etc.

FEATURES

- Intuitive User defined
- Raw data itself
- Statistics (Histograms, PCA)
- Signal Processing (Fourier Xform)

FEATURE XFORMATIONS

- Feature Selection
- Feature Extraction
- Dimensionality Reduction Eg. PCA

ML PROBLEM

1. Classification
   a. Binary
   b. Multiclass
2. Regression
3. Clustering
4. Prediction (time series)

ALGORITHMS

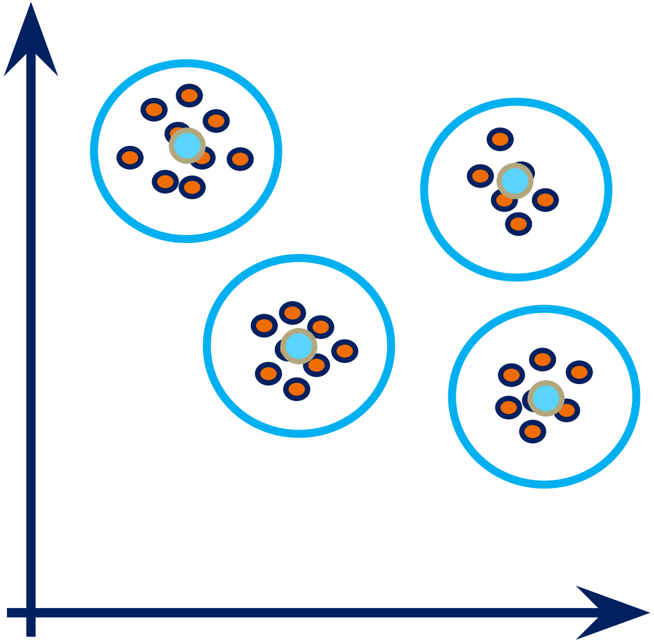
1. KNN
2. Naïve Bayes
3. Perceptron
4. Linear
5. Decision Tree

PERFORM. METRICS

- Accuracy
- Confusion Matrix
- Precision
- Recall
- AP
- True Positive, etc.

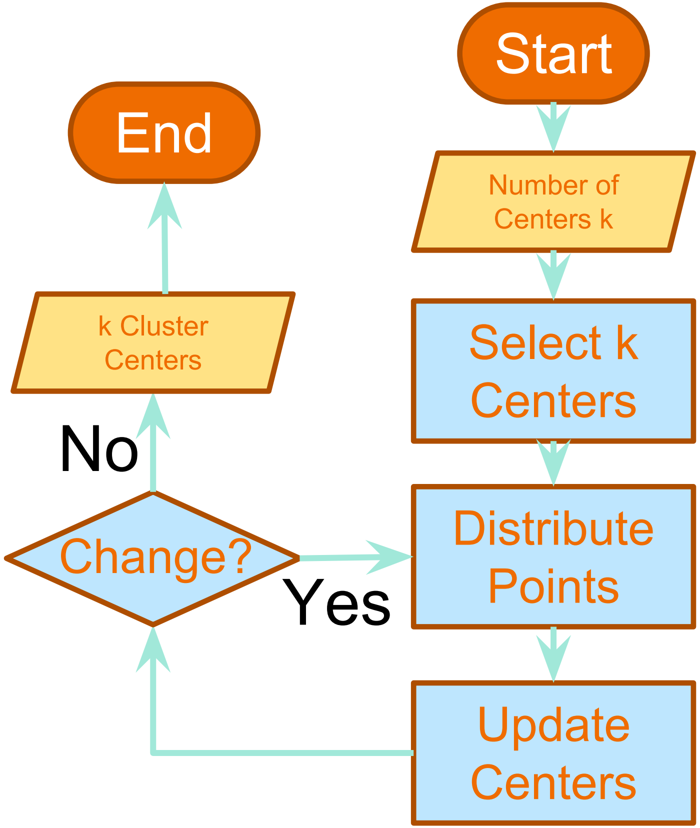
K-Means

- You are given N points
- How do we find k clusters?
    - What if we know the cluster centers?
- How do we find the cluster centers?
    - What if we know the k clusters?

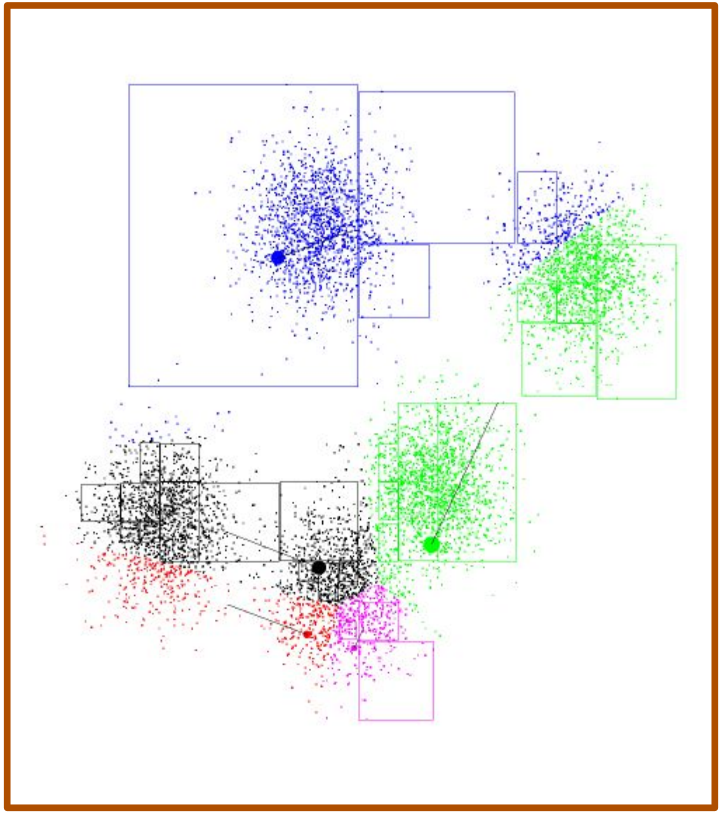


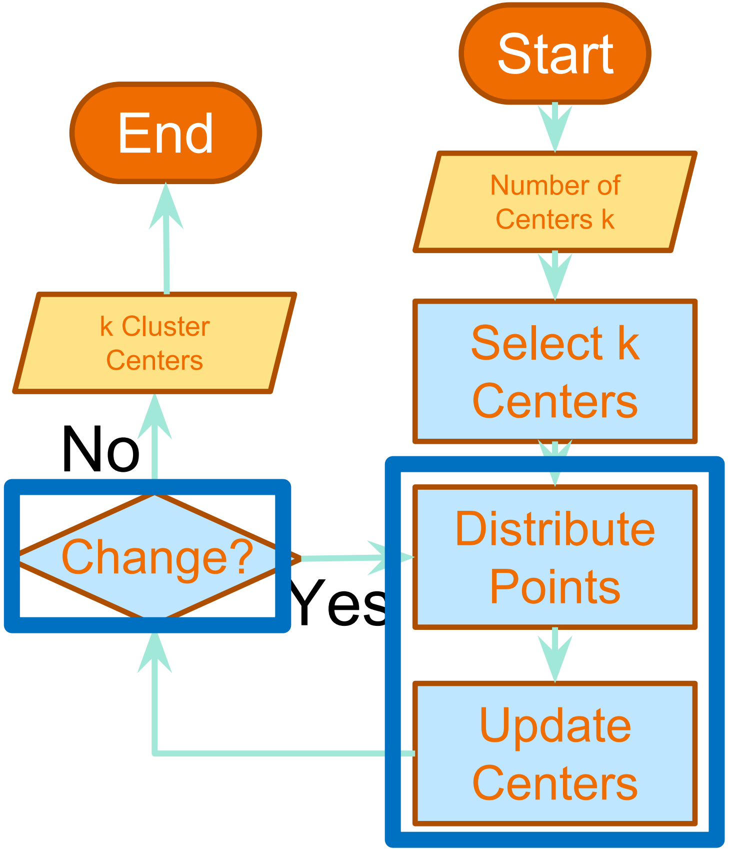
K-Means

1. Input: k (number of clusters)
2. Randomly select k centers
    3. Distribute Points
    4. Update Centers
    5. Repeat 3,4 till convergence
6. Output: Cluster centers

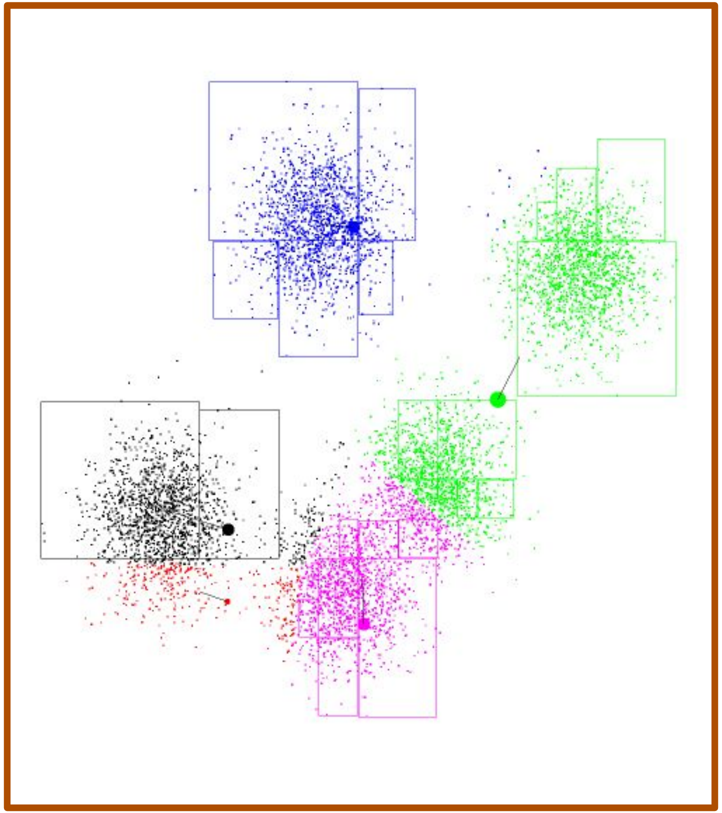


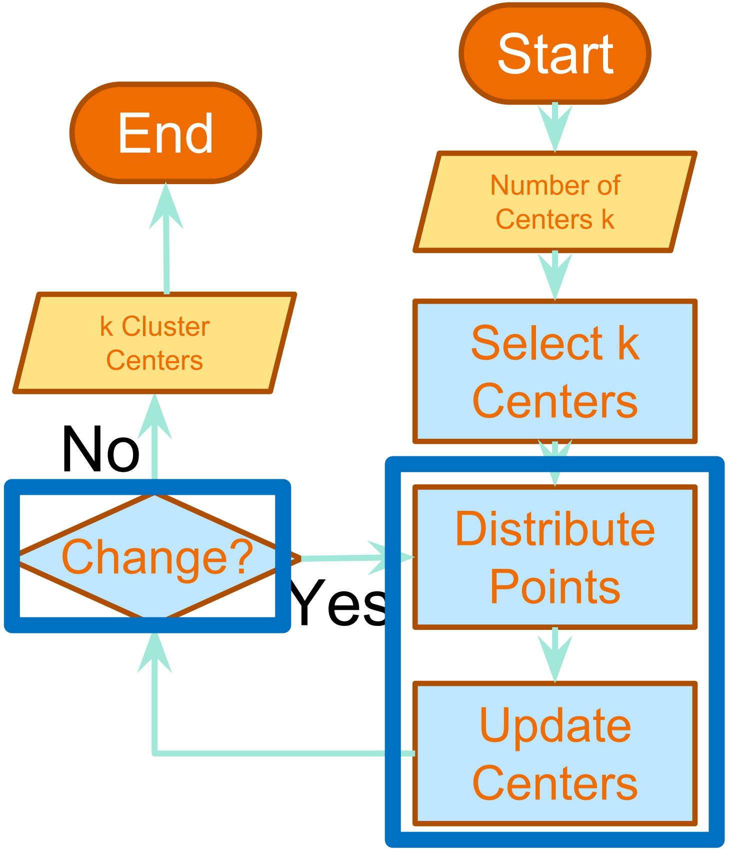
K-Means



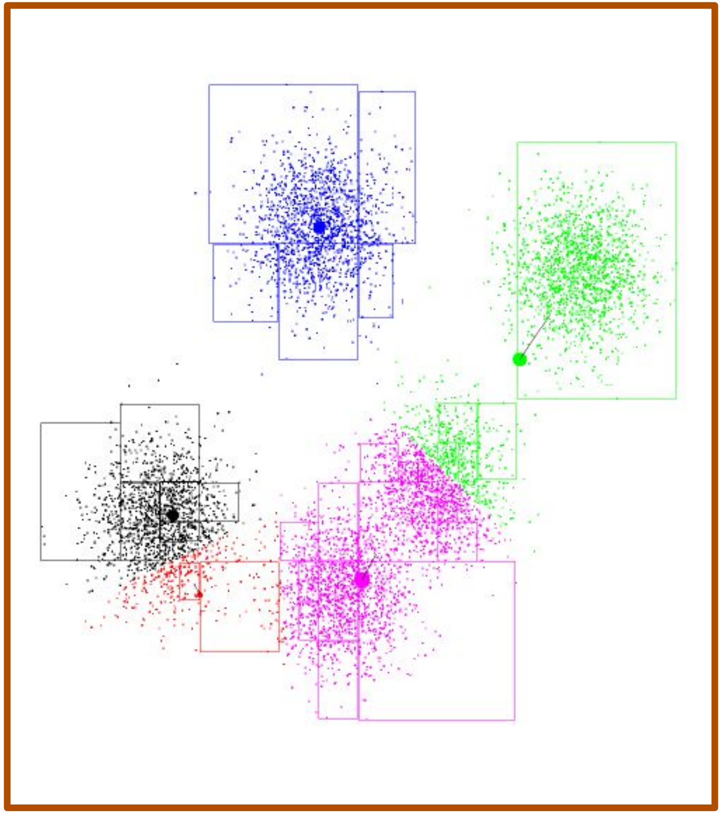

Start
Number of Centers k
Select k Centers
Distribute Points
Update Centers
Change?
Yes
No
k Cluster Centers
End

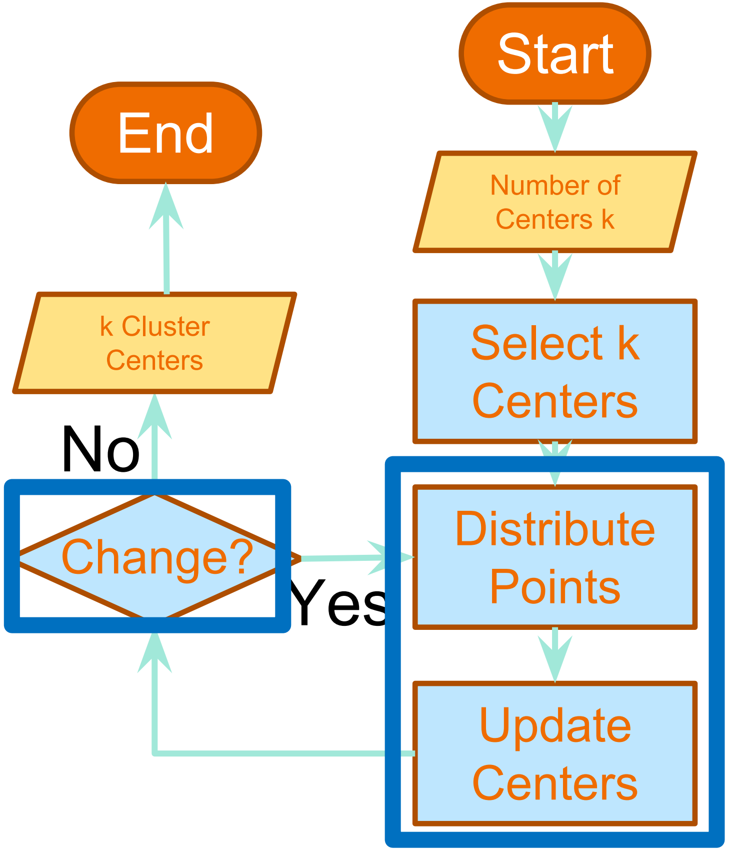
K-Means: Update 1



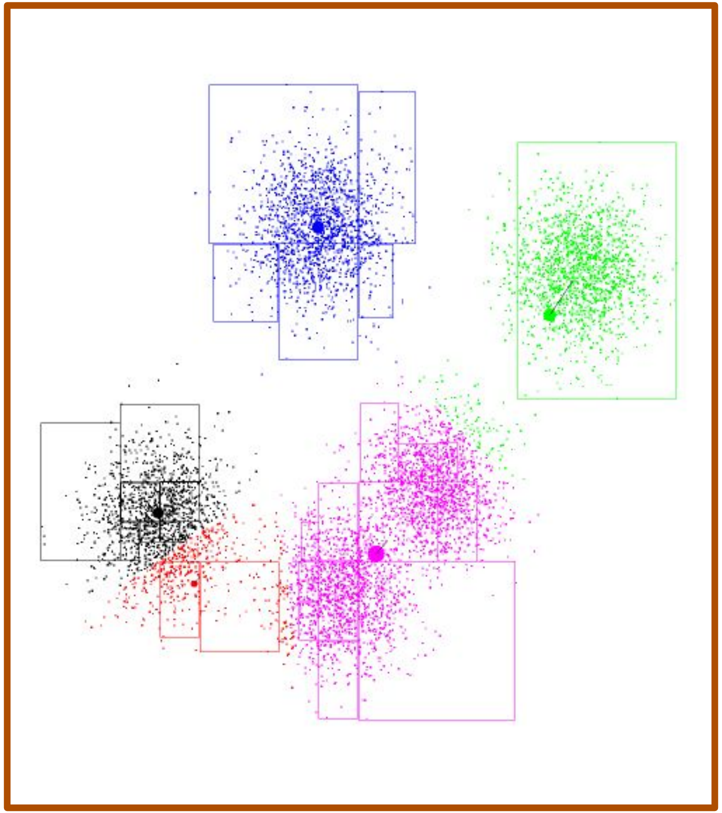


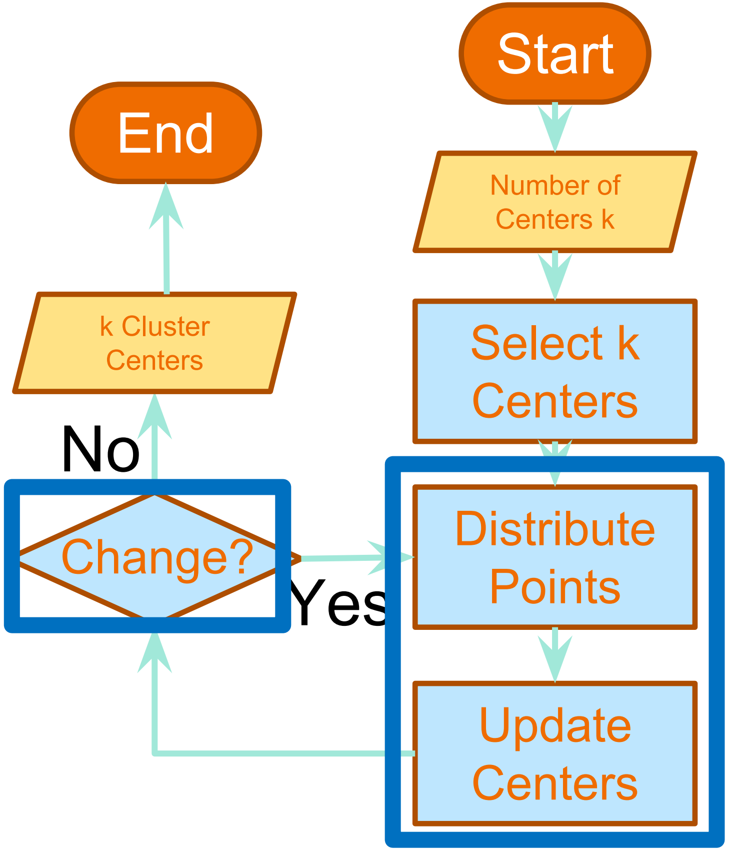
K-Means: Update 2



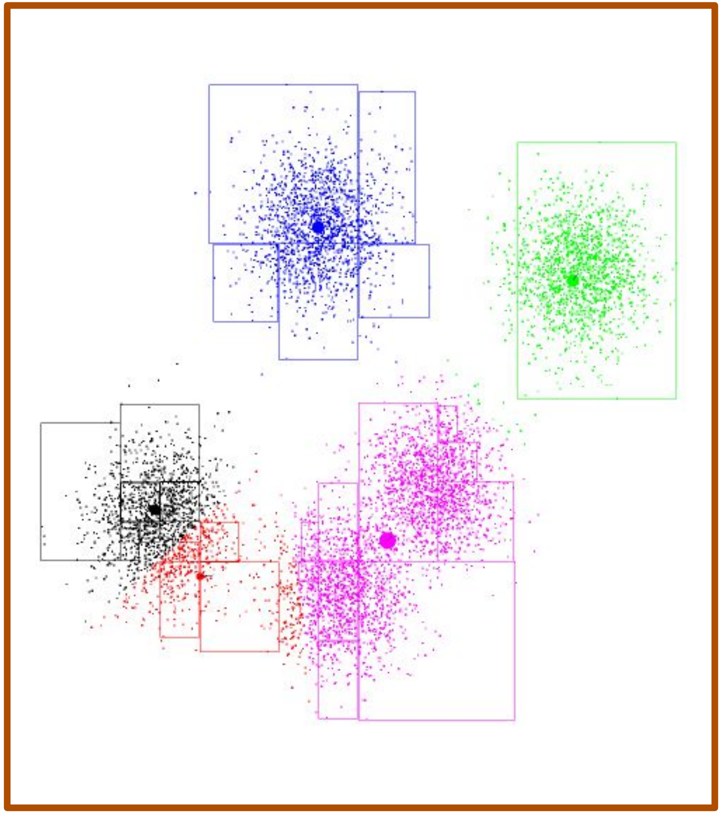


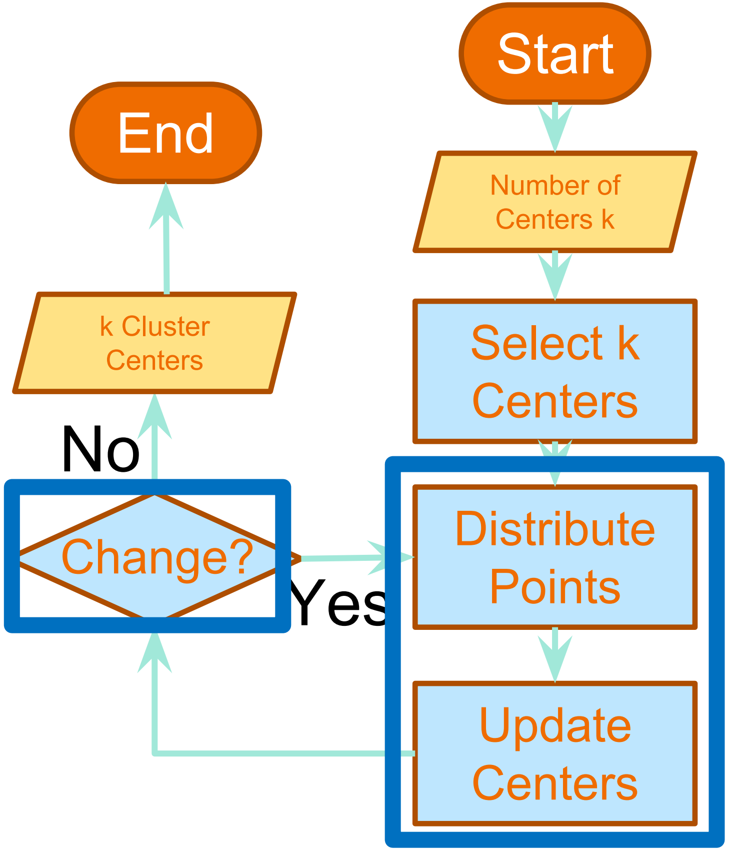
K-Means: Update 3



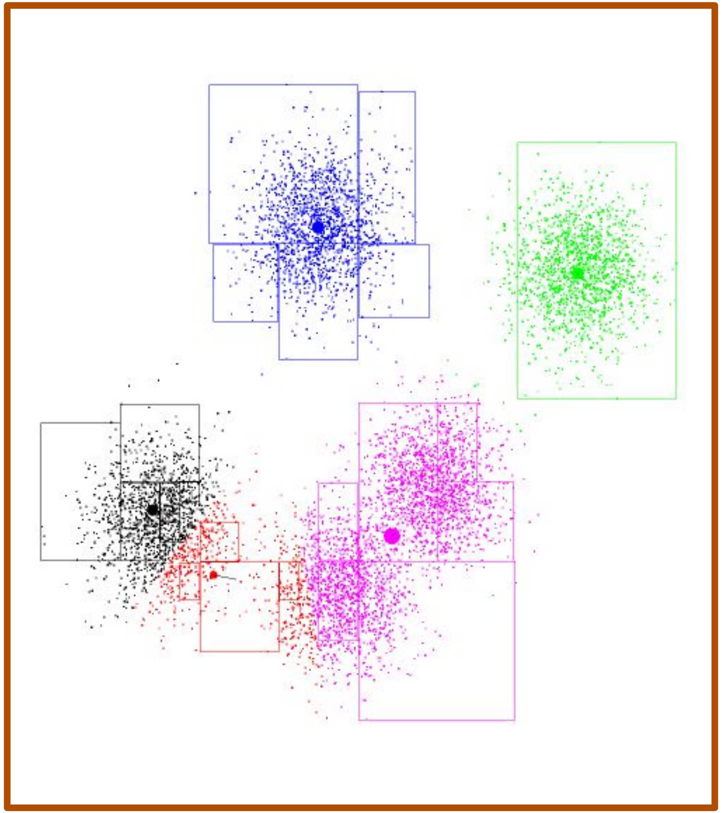


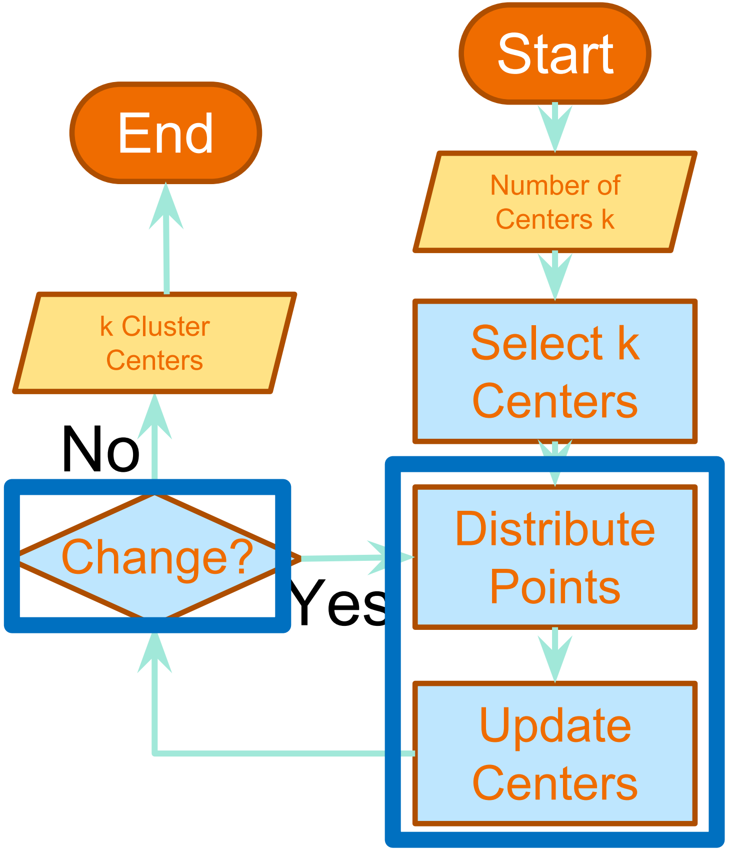
K-Means: Update 4



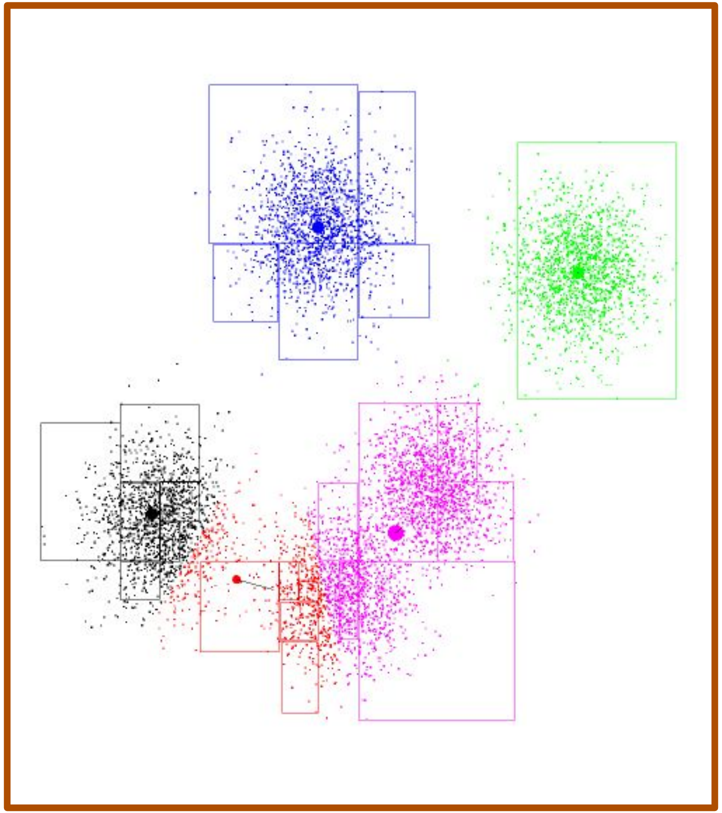


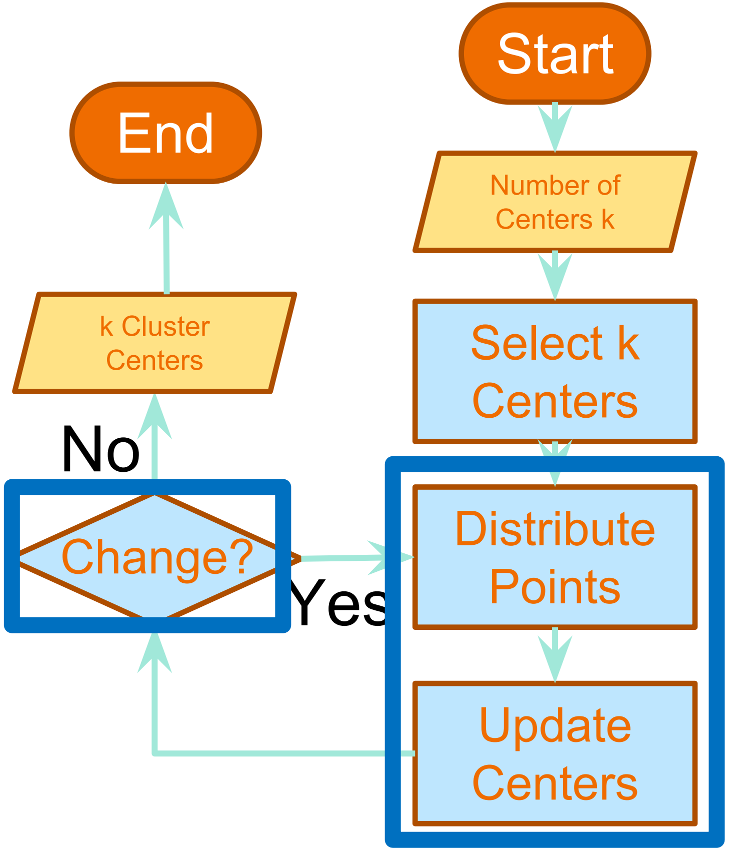
K-Means: Update 5



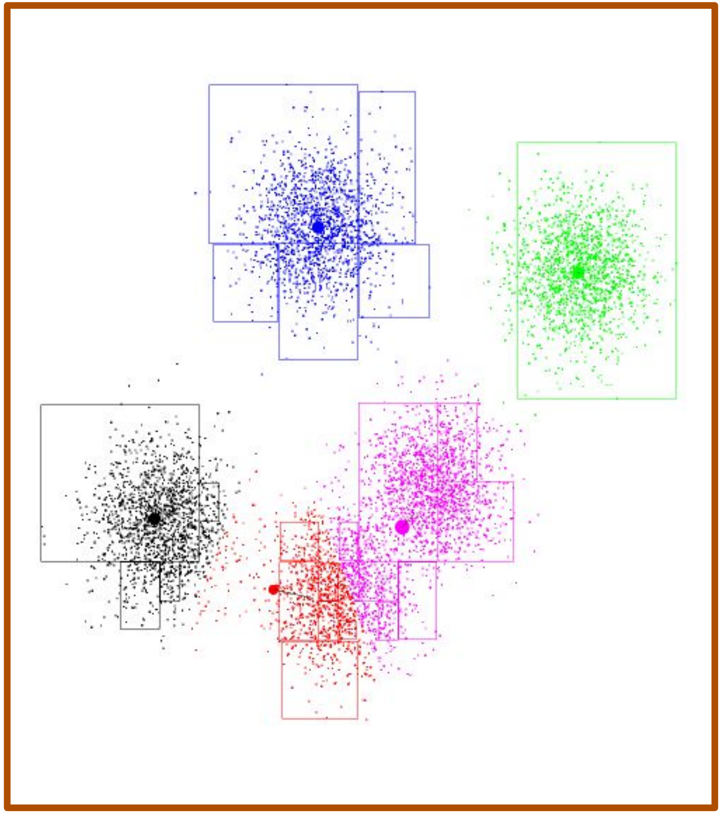


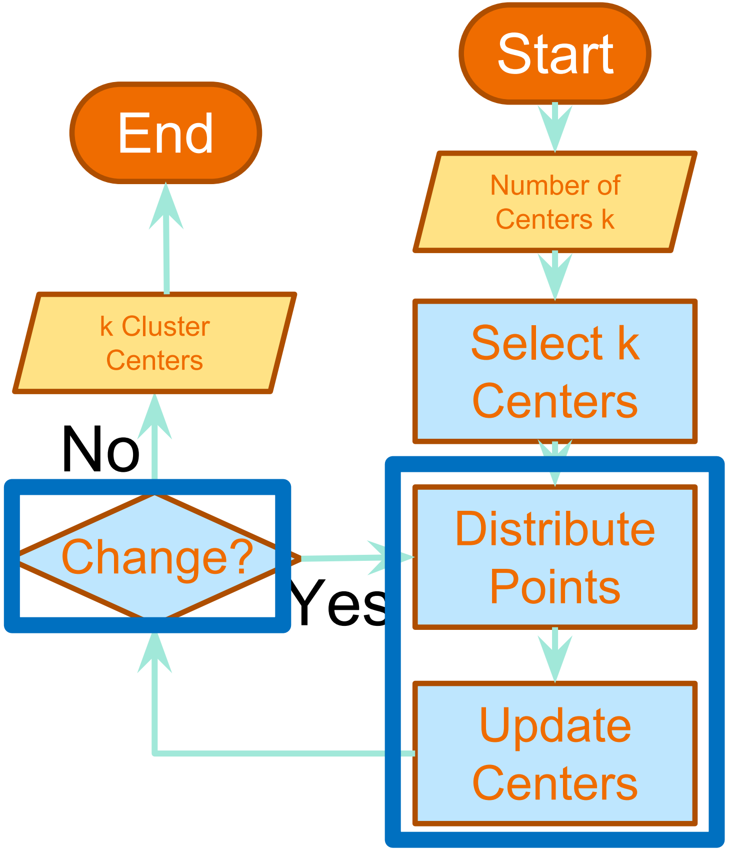
K-Means: Update 6



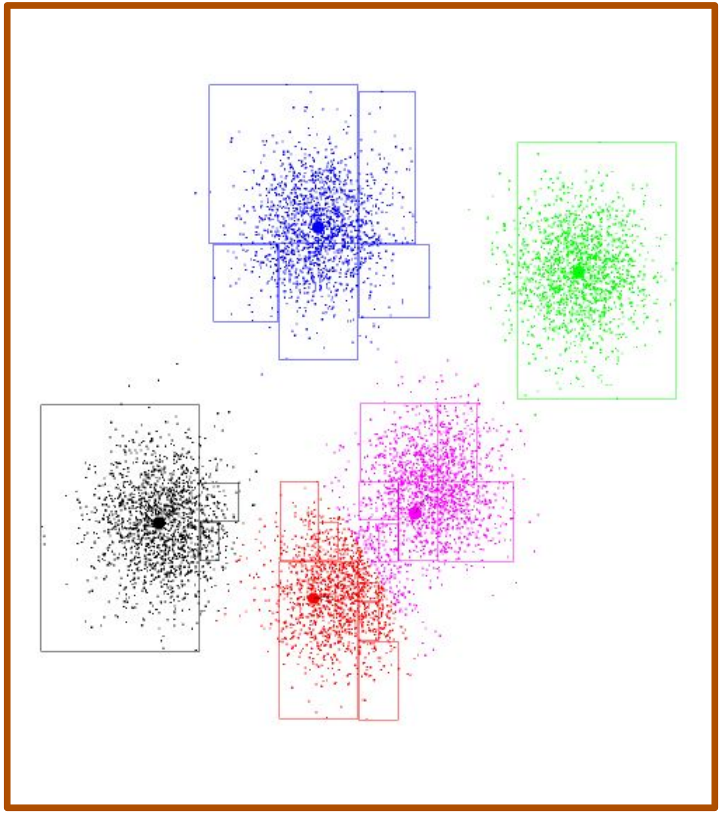


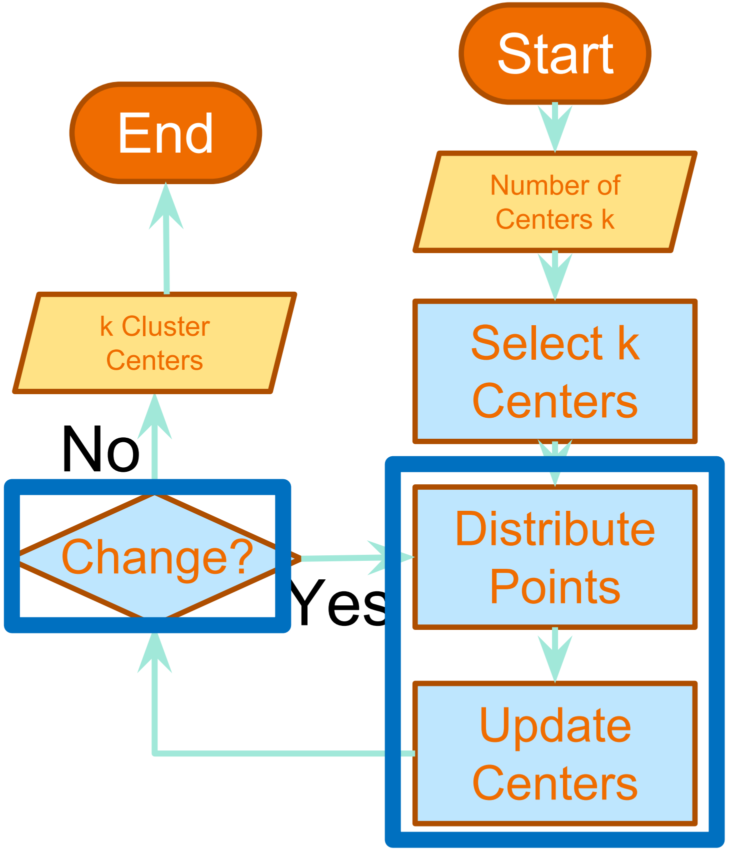
K-Means: Update 7



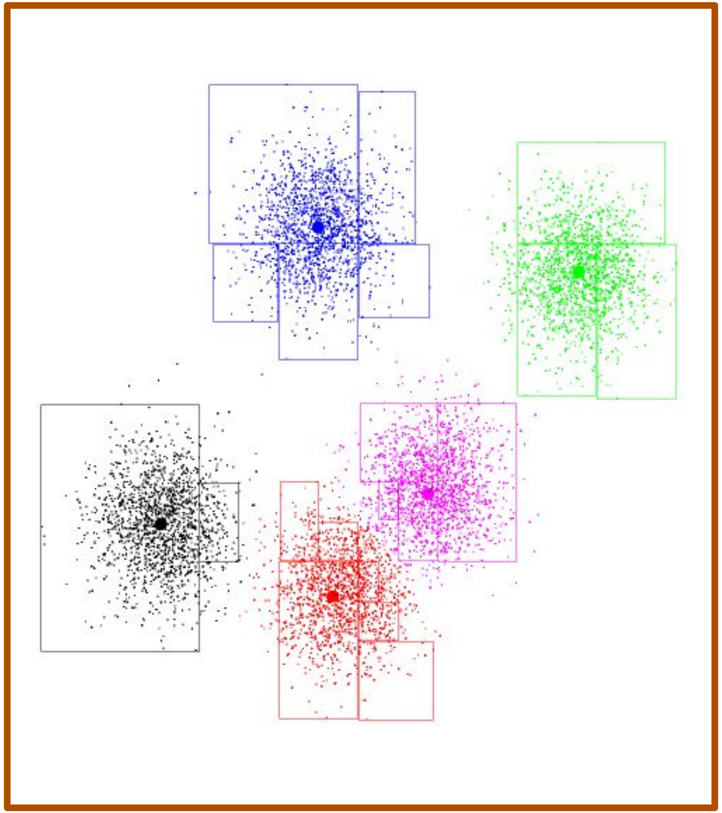


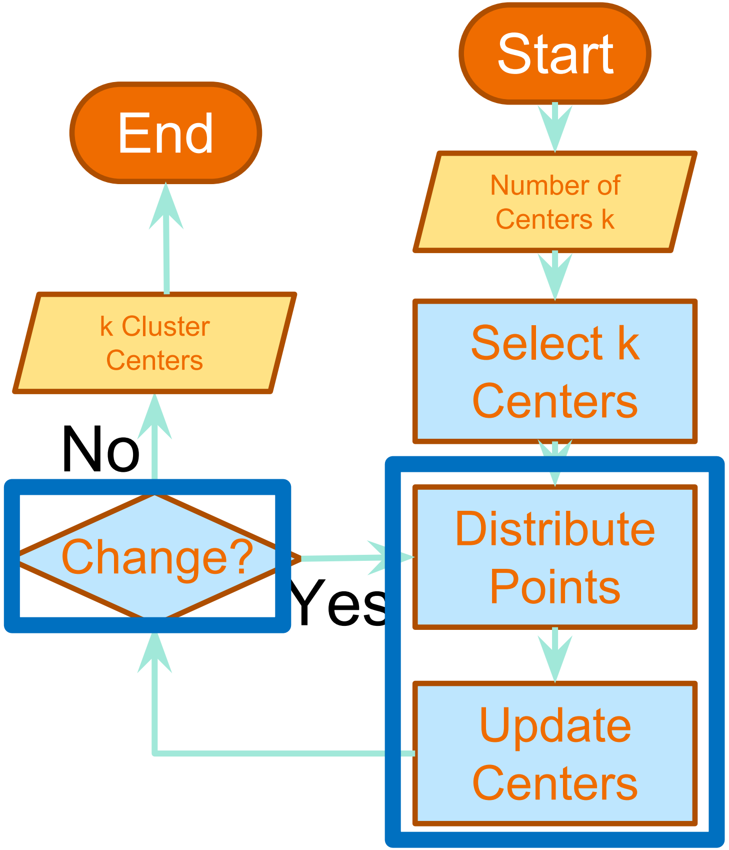
K-Means: Update 8





K-Means: Update 9





Hierarchical Clustering

- Data in the world is not flat
  - Animals, Trees, Birds, Fish, Rocks
  - Types of Animals, Species, Subspecies, Types of rocks,...
- Can we recover the hierarchical structure from clustering
  - Agglomerative vs. Divisive
  - Bottom-up vs. Top-down

User’s Dilemma!

- Which similarity measure and which features to use?
- How many clusters?
- Which is the “best” clustering method?
- Are the individual clusters and the partition valid?
- How to choose algorithmic parameters?

Data Clustering: Jain and Dubes.

Sequence Prediction/ Time Series

The ML Pipeline/Concept Map

DATA	FEATURES	FEATURE XFORMATIONS	ML PROBLEM	ALGORITHMS	PERFORM. METRICS
• **Structured** (NUM, CAT, ATTR) • **Digital Logs** (Tweets, SMS) • **RawData/ Sensors** (IMG/Speech) • **Others** User Behavior, etc.	• Intuitive User defined • Raw data itself • Statistics (Histograms, PCA) • Signal Processing (Fourier Xform)	•Feature Selection •Feature Extraction •Dimensionality Reduction Eg. PCA	1. Classification a. Binary b. Multiclass 2. Regression 3. Clustering 4. Prediction (time series)	1. KNN 2. Naïve Bayes 3. Perceptron 4. Linear 5. Decision Tree	• Accuracy • Confusion Matrix • Precision • Recall • AP • True Positive, etc.

Definition of Time series

Any feature that contains data ordered by time, is said to be Time series feature.

Converting time series to supervised form

The key idea is to be able to convert the time series into supervised data.

Day of week	Accidents at time T
1	10
2	11
3	12
4	13
5	14
6	15
7	16

Converting time series to supervised form

Notice how the entire data in yellow is switched one step upwards to create the T+1(red) column

Day of week	Accidents at time T	Accidents at time T+1
1	10	11
2	11	12
3	12	13
4	13	14
5	14	15
6	15	16
7	16	NaN

Number of time steps to choose

The number of time steps can be increased further based on the ‘memory’ that we intend to preserve.

Day of week	Accidents at time T	Accidents at time T+1	Accidents at time T+2
1	10	11	12
2	11	12	13
3	12	13	14
4	13	14	15
5	14	15	16
6	15	16	NaN
7	16	NaN	NaN

Univariate and Multivariate

- The examples above show how a single variable varies with time. It is possible however to have a multiple variables that vary together to be modeled as supervised learning problem.

Day of week	Rainfall at time T	Accidents at time T	Rainfall at time T+1	Accidents at time T+1
1	200	10	201	11
2	201	11	202	12
3	202	12	203	13
4	203	13	204	14
5	204	14	205	15
6	205	15	206	16
7	206	16		

Using the series-to-supervised form

- We can further use this new supervised representation of series data, for building models such as Linear regression, Recurrent neural networks (RNN) etc.

Linear Regression

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY HYDERABAD OUTREACH DIVISION | talent sprint

The ML Pipeline/Concept Map

DATA

- **Structured** (NUM, CAT, ATTR)
- **Digital Logs** (Tweets, SMS)
- **RawData/ Sensors** (IMG/Speech)
- **Others** User Behavior, etc.

FEATURES

- Intuitive User defined
- Raw data itself
- Statistics (Histograms, PCA)
- Signal Processing (Fourier Xform)

FEATURE XFORMATIONS

- Feature Selection
- Feature Extraction
- Dimensionality Reduction Eg. PCA

ML PROBLEM

1. Classification
   a. Binary
   b. Multiclass
2. Regression
3. Clustering
4. Prediction (time series)

ALGORITHMS

1. KNN
2. Naïve Bayes
3. Perceptron
4. Linear
5. Decision Tree

PERFORM. METRICS

- Accuracy
- Confusion Matrix
- Precision
- Recall
- AP
- True Positive, etc.

The Machine Learning Framework

$$Y = f(x)$$

Output — **prediction function** — **feature representation**

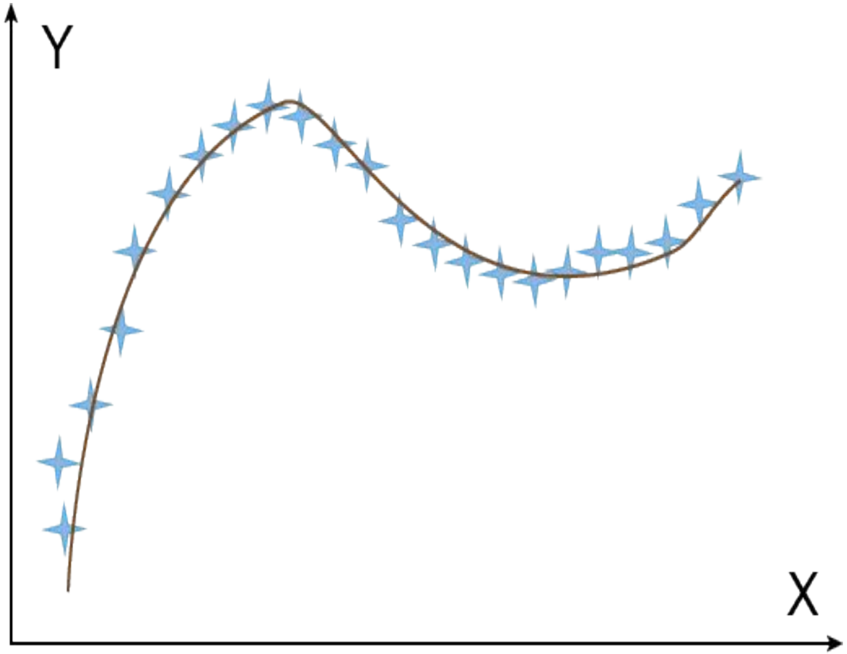
Training: given a training set, estimate the prediction function **f** by minimizing the prediction error

Testing: apply **f** to unknown test sample ***x*** and predicted value(output) is **y**

Find a function to fit the data

- Discover hidden structure in the data, given the samples
  Why?
- A functional form is usable for interpolation and extrapolation

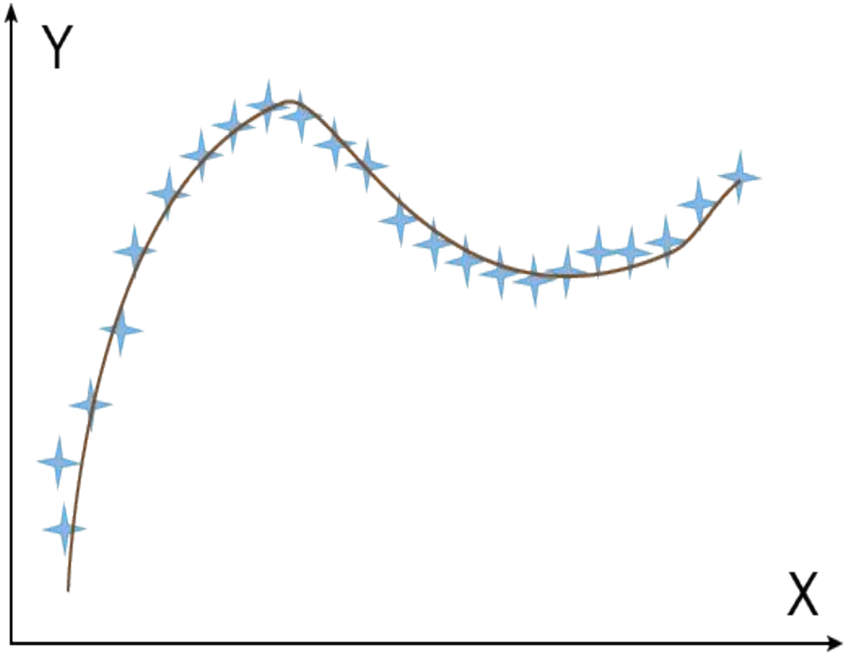
$$y = f(x)$$



Scalar or vector valued x, y
Multivariate, when x is a vector x

Find a function to fit the data

- Functional forms: Linear, Polynomial, Gaussian, etc
- Such problems are called regression in general

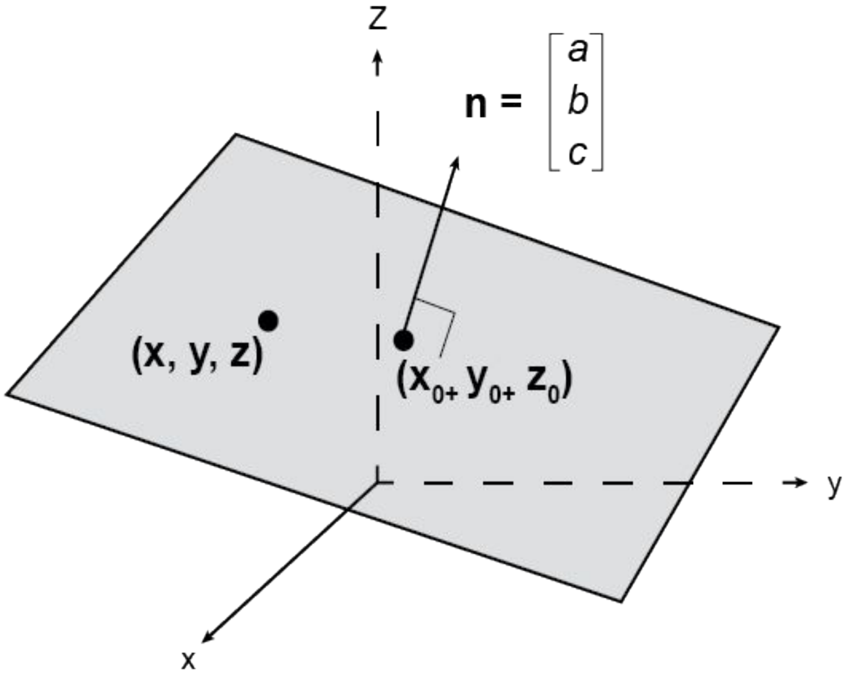


$$y = f(x)$$

Scalar or vector valued x, y
Multivariate, when x is a vector x

Linear Model

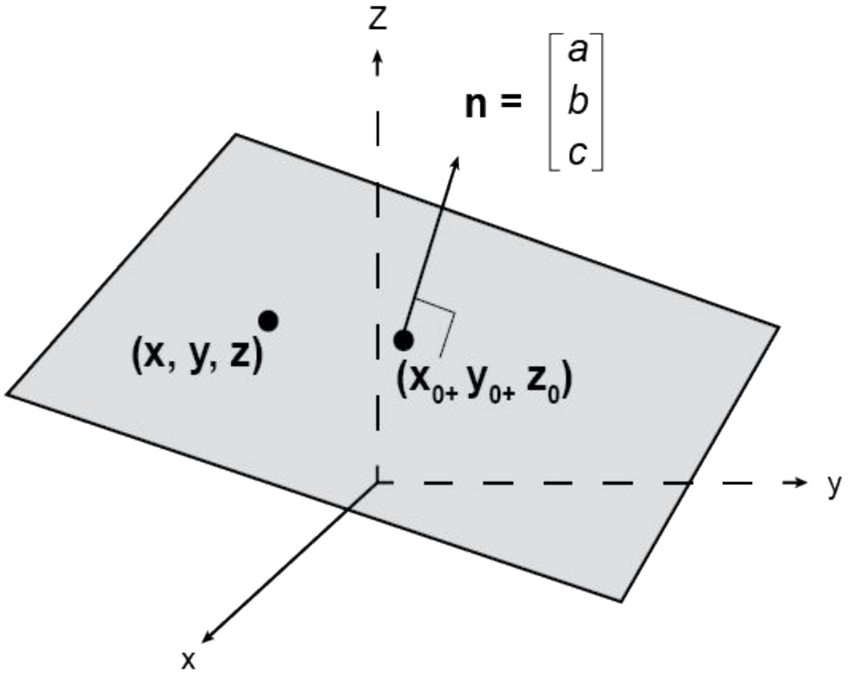
- Linear when f is a line
- In general,a hyperplane
  - $y = ax + b$
    - a, b : parameters of the model
  - $y = w^T x = w.x$ when multivariate
    - $x = [1 \; x_1 \; x_2 \; ... \; x_d]^T$
    - w :parameters of model



Linear Model

- Line: Only 2 parameters in 2D. d parameters in a d-dimensional space

$$y = f_w(x) = w^T x$$



Thanks!!

Questions?