EMORY UNIVERSITY

Emory Continuing Education

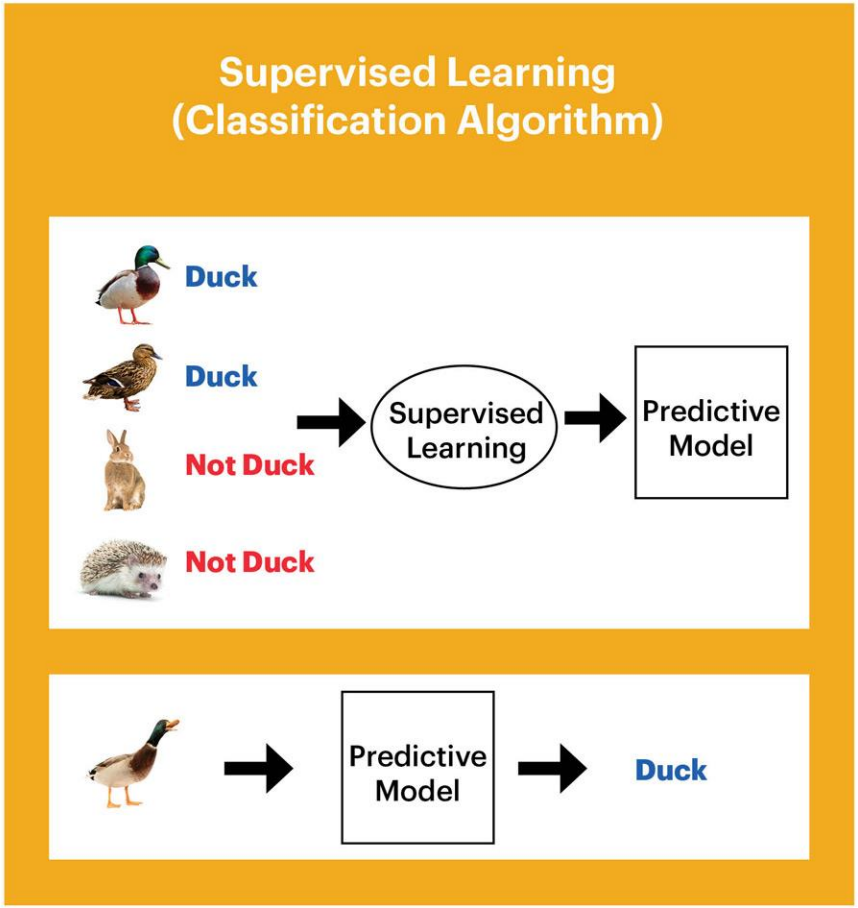
Descriptive Analytics with Python

Sridhar Palle, Ph.D.

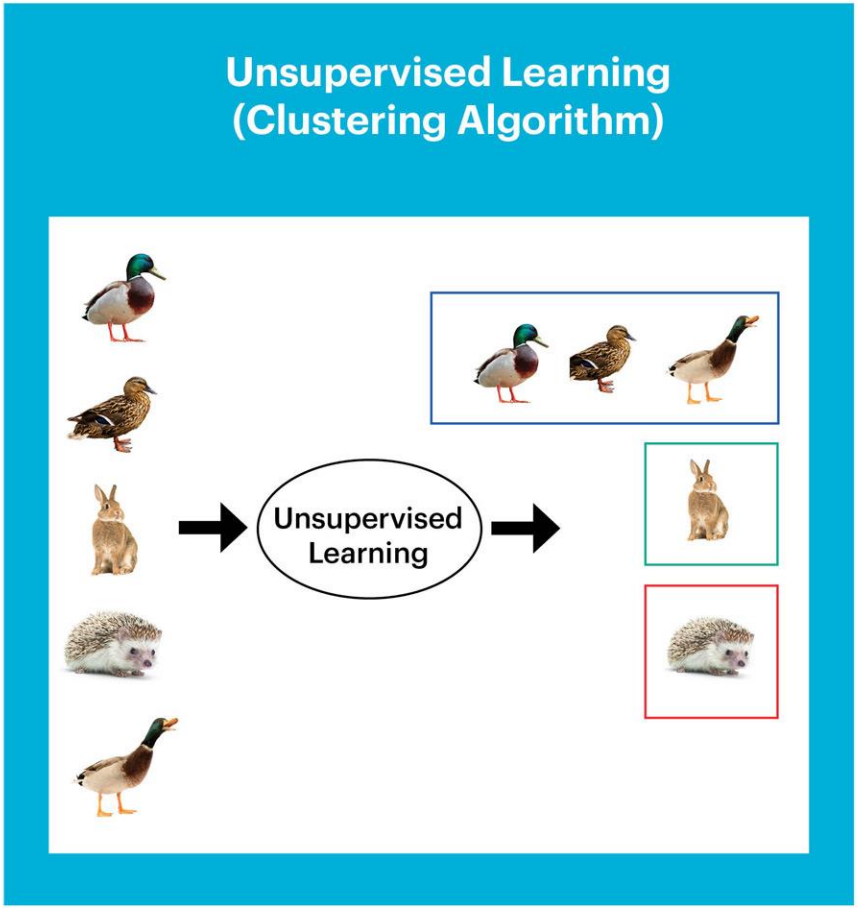
Consort Institute

Supervised vs Unsupervised

Labelled Data



No Labelled Data

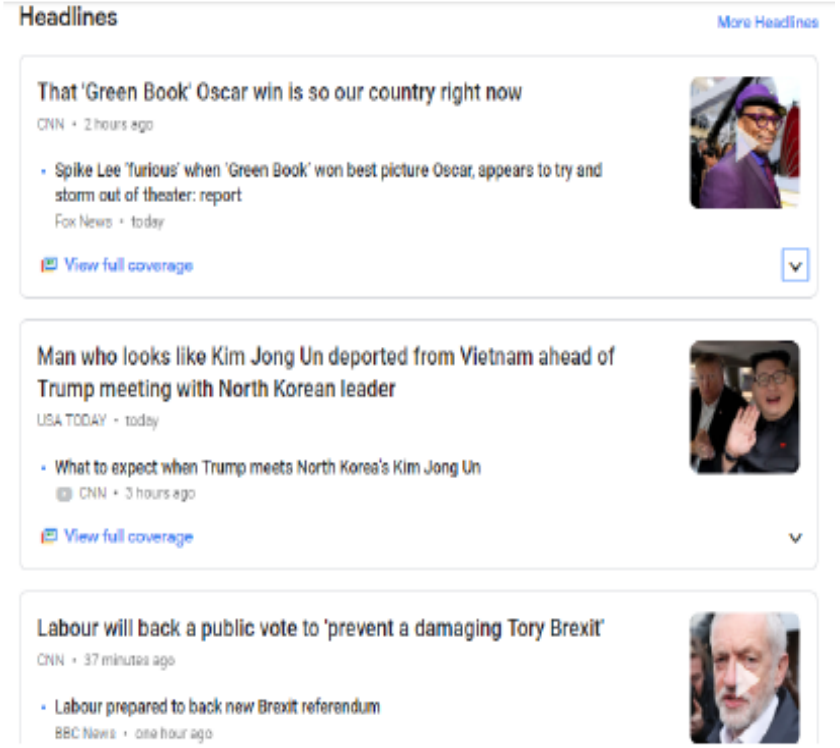


Western Digital.

Unsupervised Learning Techniques

- **Association Rules Mining (ARM)**
  - Co-occurrences of items/item sets
  - Objective is not to predict occurrence of item, but to find usable, hidden patterns

- **Clustering**
  - Automatically segregate dataset into groups
  - Ex: automatically classify news articles



ARM

- **Obvious Ones:**

- **Not so obvious**
  - Prized discoveries
  - Business can profit from

- **Advantages**
  - Bundle Pricing
  - Product placement
  - Shelf space optimization

- **Downside**
  - May also lead to spurious relationships when dealing with data with billions of transactions

ARM Procedure

Step 1: Prepare data in a particular format

Step 2: Short-list frequently occurring items (or items sets), based on some support level

Step 3: Generate relevant association rules from item sets generated in step 2 (based on some confidence parameters)

{Item A} -> {Item B}

Item A – Antecedent or premise of a rule

Item B – Consequent or conclusion of a rule.

Main challenge for association rule analysis are

- Computational time and resources
- Ex: For association analysis of 'n' items, there will be

  2^n -1 item sets, and $3^n – 2^n + 1$ association rules can be found
- Fortunately there are algorithmic approaches to efficiently find the frequent item sets and rules based on some parameters (support, confidence, lift)

Support, Confidence, Lift

- **Support (A) =** $\dfrac{Number\ of\ Occurences\ of\ Item\ A}{Total\ Number\ of\ transactions}$
- **Confidence (A → B) =** $\dfrac{Support(A,B)}{Support(A)}$
- **Lift (A → B) =** $\dfrac{Support(A,B)}{Support(A) * Support(B)}$

ARM Practical Example: Grocery Store

Step 1: Prepare data in a particular format

Step 2: Short-list frequently occurring items/item sets based on some support level

Support for Milk { } = 60%,

Support of Cookies { } = 50%,

Support of Milk and cookies { & } = 50%

Step 3: Generate Rules

{ → }

Transaction Receipts	Items
Receipt 1	
Receipt 2	
Receipt 3	
Receipt 4	
Receipt 5	
Receipt 6	
Receipt 7	
Receipt 8	
Receipt 9	
Receipt 10	

ARM: Grocery store

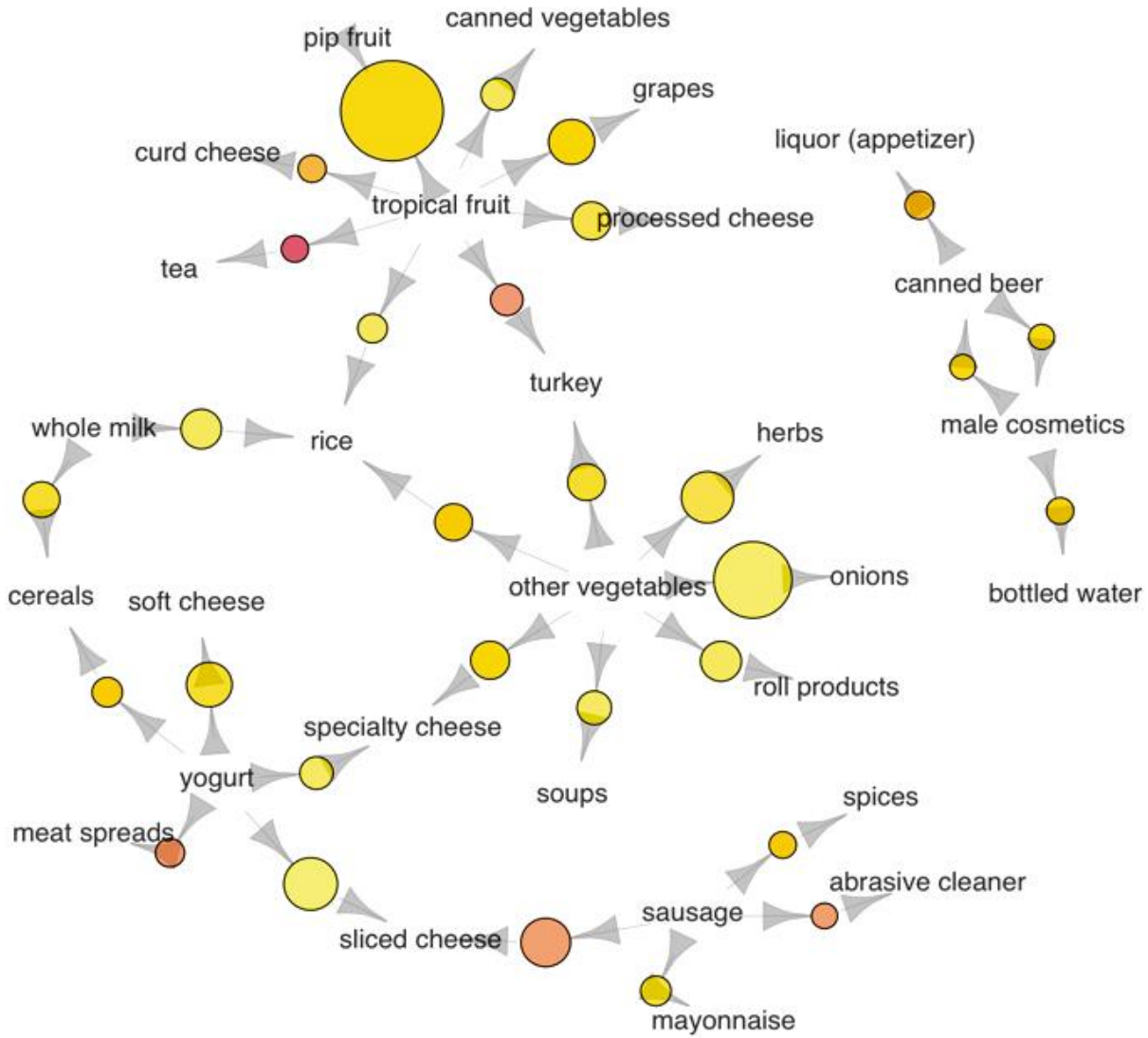
- Rule Significance

$$\text{Confidence}\{\text{milk} \rightarrow \text{cookie}\} = \frac{\text{Support of Milk and cookies}\{\text{milk \& cookie}\}}{\text{Support for Milk}\{\text{milk}\}} = 83\%$$

$$\text{Lift}\{\text{milk} \rightarrow \text{cookie}\} = \frac{\text{Support of Milk and cookies}\{\text{milk \& cookie}\}}{\text{Support of Milk}\{\text{milk}\} * \text{Support of Cookie}\{\text{cookie}\}} = 1.66$$

	Support	Confidence	Lift
Milk → Cookies	50%	83%	1.66
Milk → Eggs	40%	66%	1.11

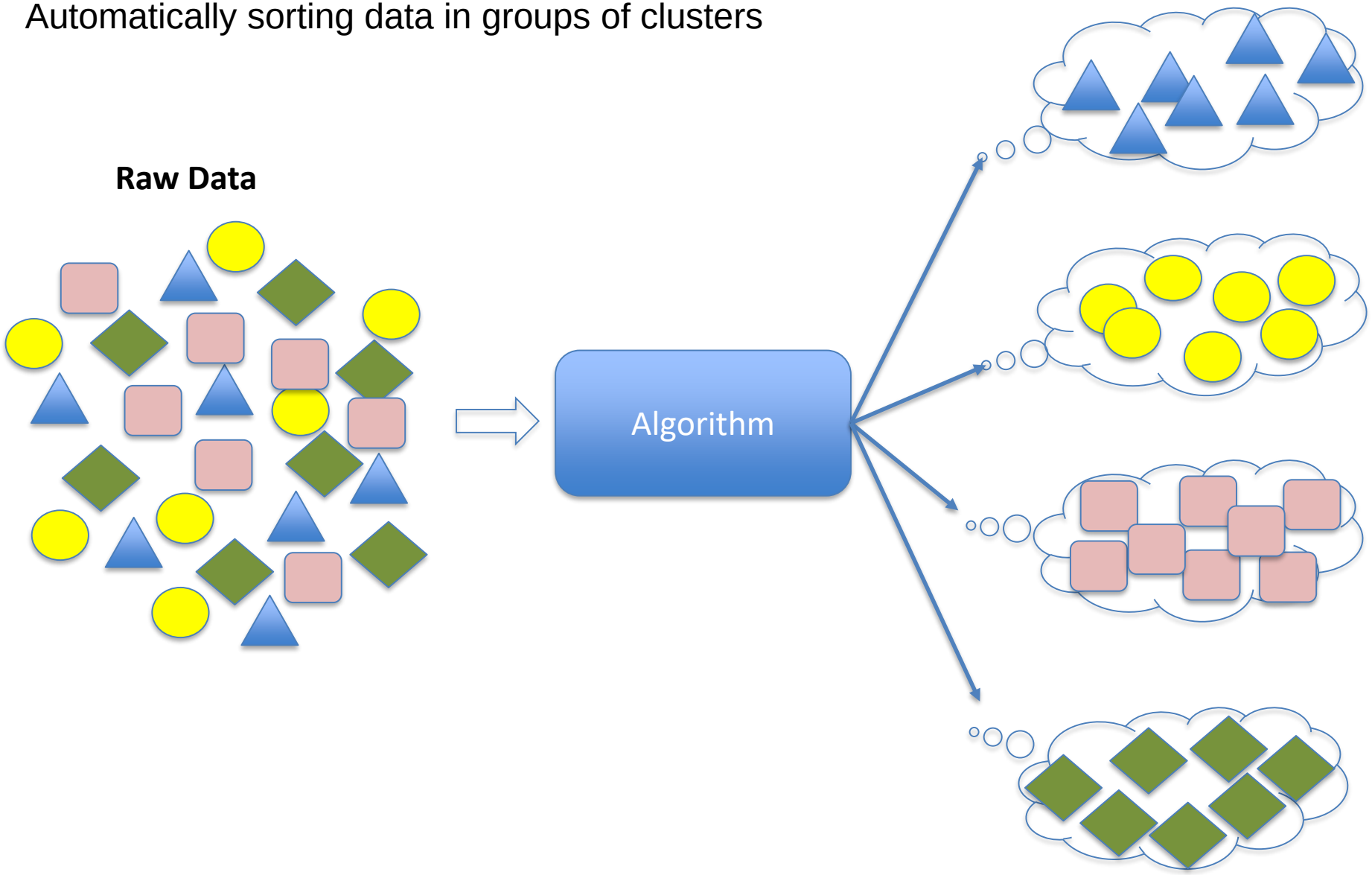
ARM Network Graph



Source: kdnuggets

Clustering

Automatically sorting data in groups of clusters

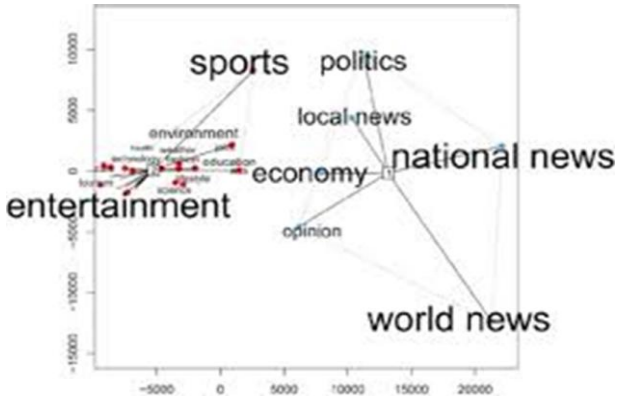


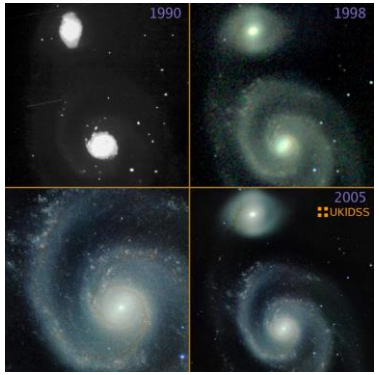
Clustering

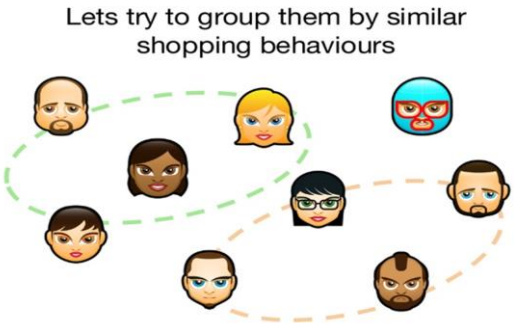
Clustering is the process of automatically segregating dataset into different groups based on commonalities (similarities).

- Applications in diverse domains

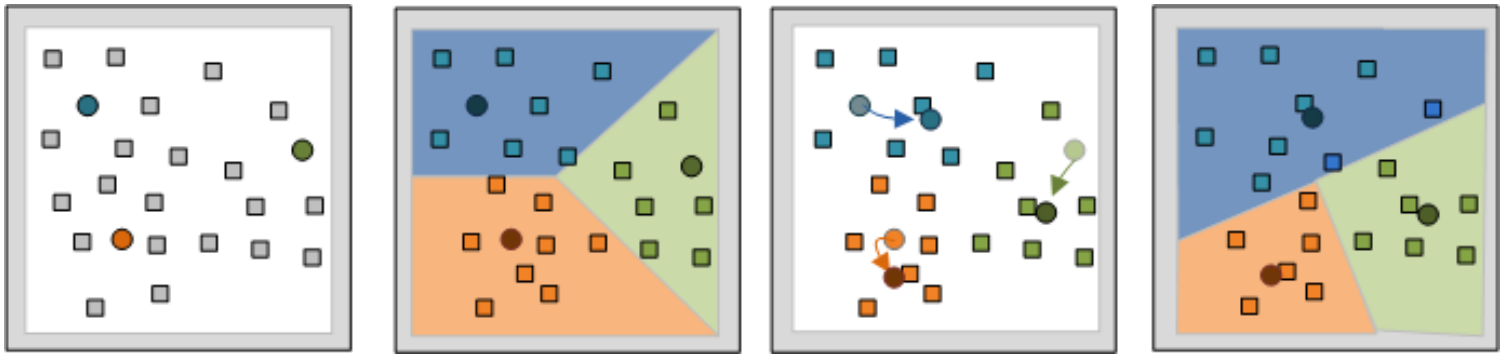
- Different clustering algorithms exist
  - Kmeans
  - Kmedoids
  - Hirerchical clustering
  - DBSCAN







K-Means Algorithm



Descriptive Analytics with Python

- Please go to the link below
  - Tinyurl.com/ece-bdata-python/
- Download the folder
  - DA-Python
- Open Jupyter notebooks related to ARM and Clustering using Anaconda

Thank You ☺

Any questions?