



EMORY
UNIVERSITY

Emory
Continuing
Education

Descriptive Analytics with Python



Sridhar Pale, Ph.D.



© 2015 Consort Institute, LLC. All right reserved. This material may not be reproduced, displayed, modified or distributed in any forms by any means without the express prior written permission of Consort Institute, LLC



EMORY
UNIVERSITY

Emory Continuing
Education

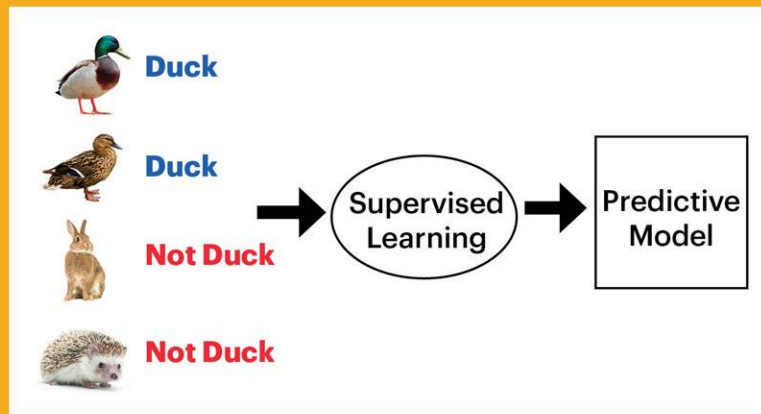
ece.emory.edu | 404.727.6000 | ece@emory.edu



Supervised vs Unsupervised

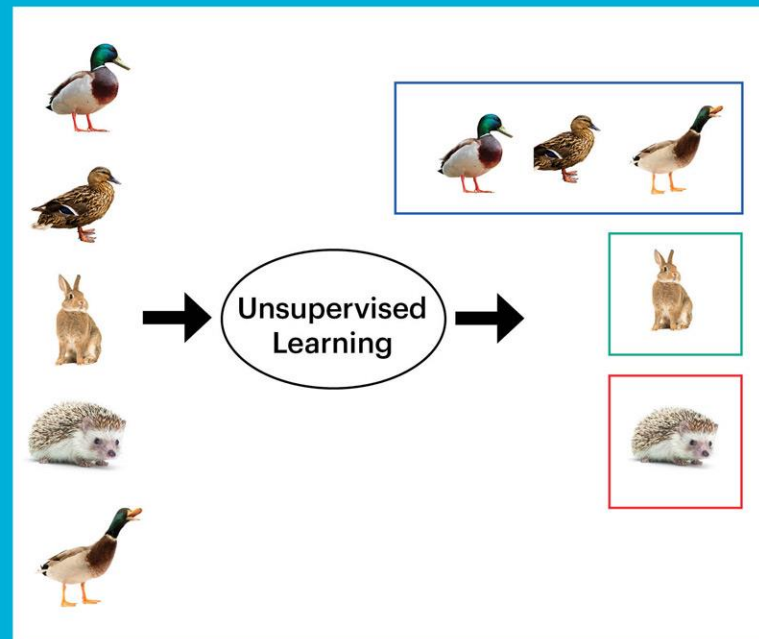
Labelled Data

Supervised Learning (Classification Algorithm)



No Labelled Data

Unsupervised Learning (Clustering Algorithm)



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



Headlines

[More Headlines](#)

That 'Green Book' Oscar win is so our country right now

CNN • 2 hours ago

- Spike Lee 'furious' when 'Green Book' won best picture Oscar, appears to try and storm out of theater: report

Fox News • today

[View full coverage](#)



Man who looks like Kim Jong Un deported from Vietnam ahead of Trump meeting with North Korean leader

USA TODAY • today

- What to expect when Trump meets North Korea's Kim Jong Un

CNN • 3 hours ago

[View full coverage](#)



Labour will back a public vote to 'prevent a damaging Tory Brexit'

CNN • 37 minutes ago

- Labour prepared to back new Brexit referendum

BBC News • one hour ago



EMORY
UNIVERSITY

Emory Continuing
Education

ece.emory.edu | 404.727.6000 | ece@emory.edu

Consort
Institute

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






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

{  \rightarrow  }

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

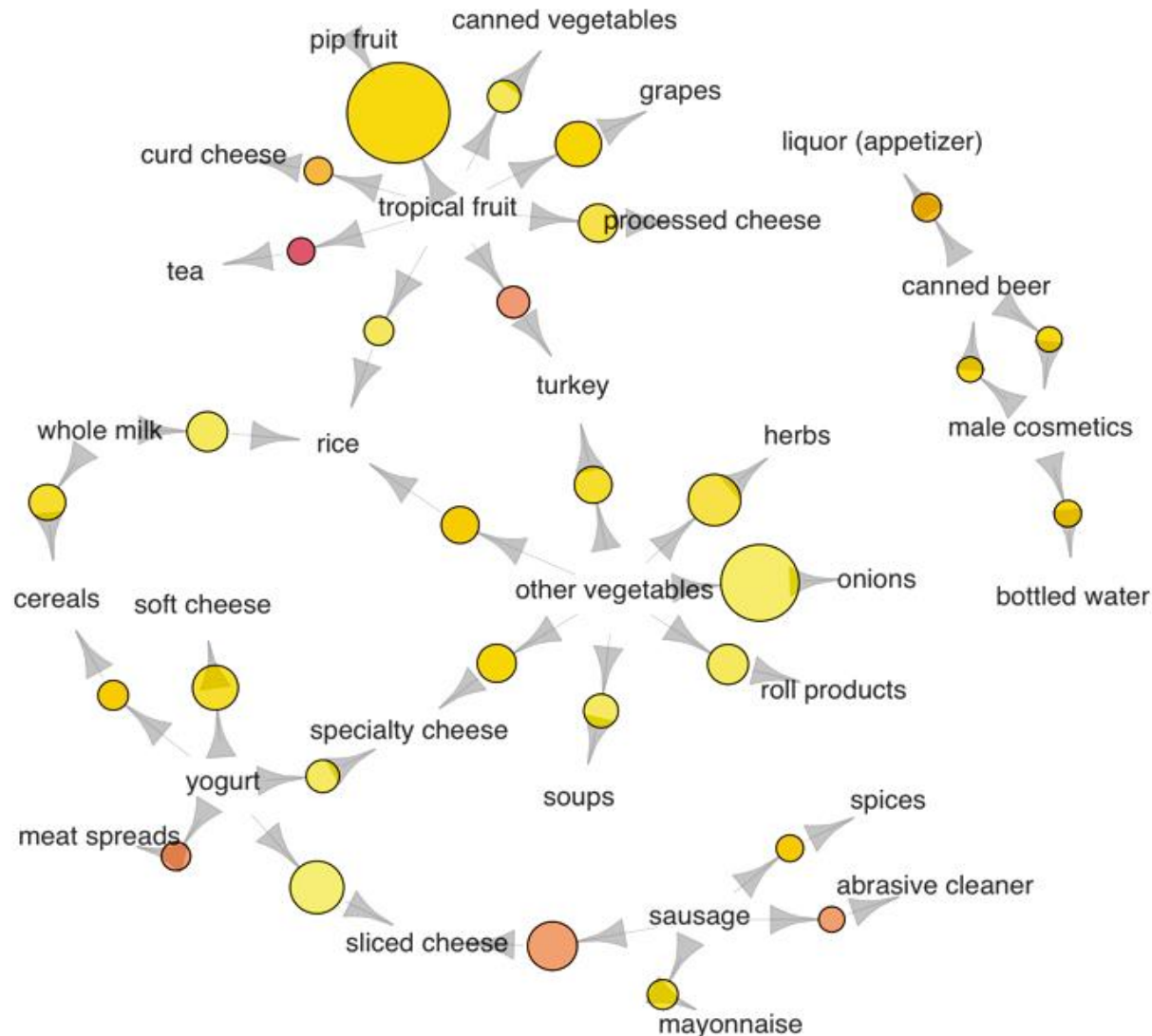
- Rule Significance

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

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

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

ARM Network Graph



Source: kdnuggets



EMORY
UNIVERSITY

Emory Continuing
Education

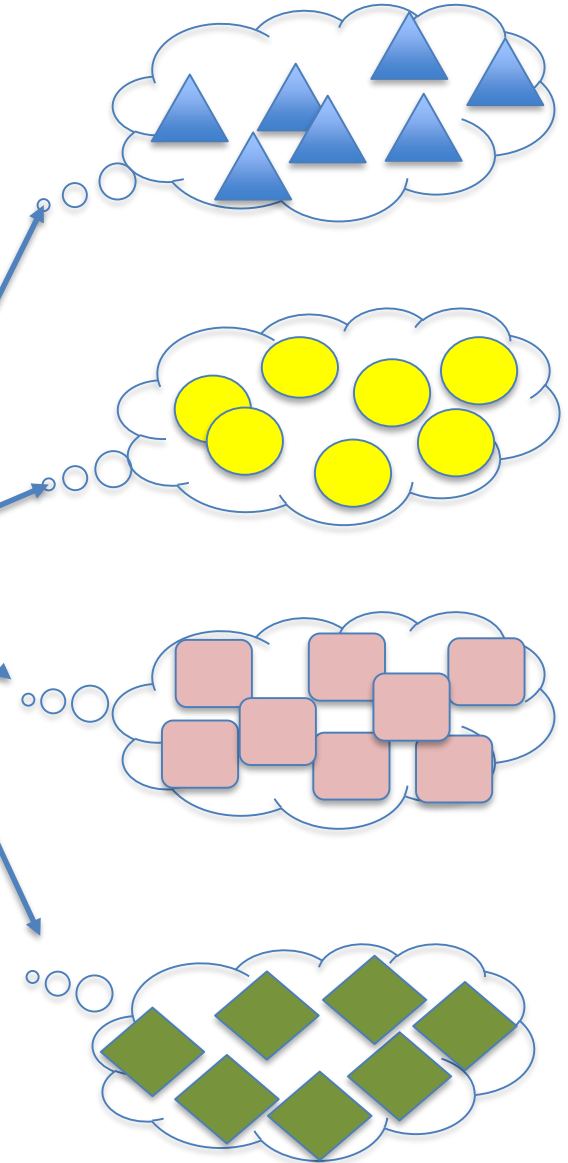
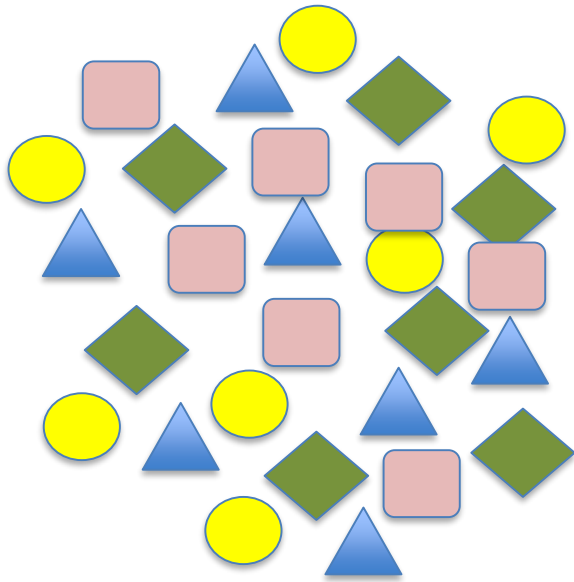
ece.emory.edu | 404.727.6000 | ece@emory.edu

Consort
Institute

Clustering

Automatically sorting data in groups of clusters

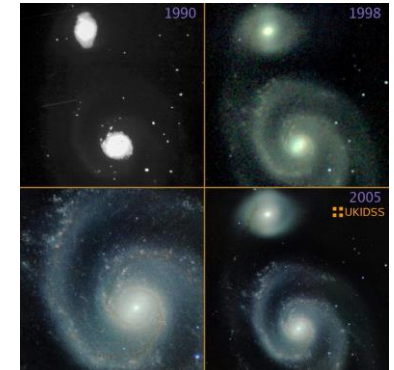
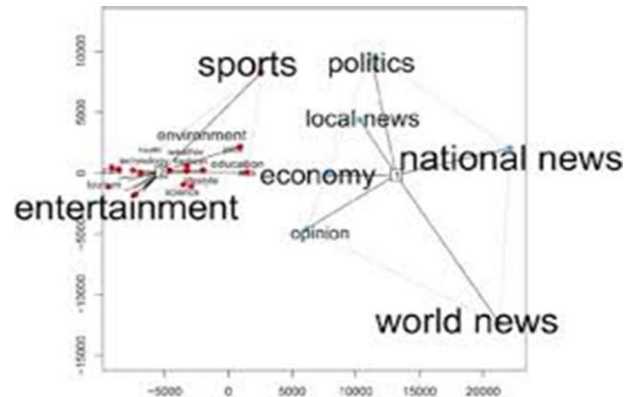
Raw Data



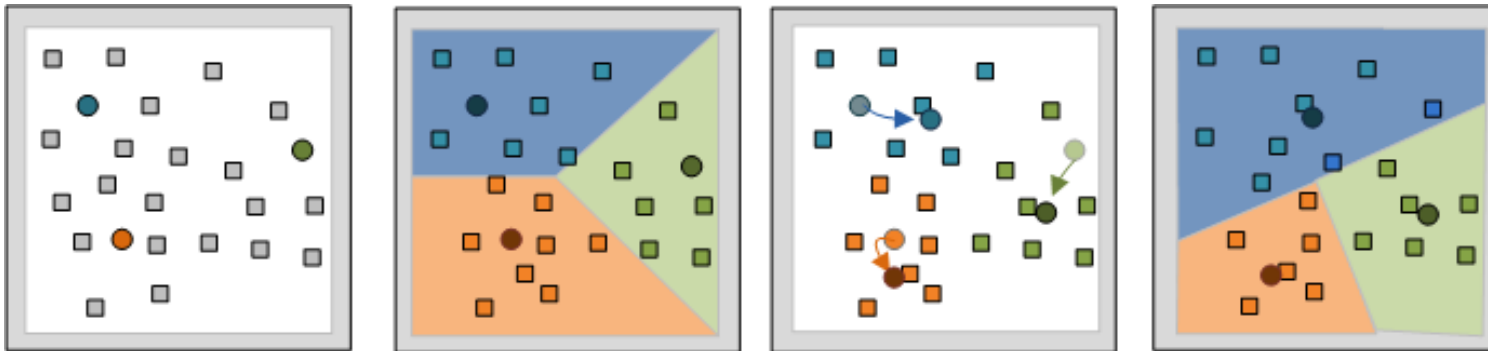
Clustering

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
 - Hierarchical clustering
 - DBSCAN



K-Means Algorithm



Descriptive Analytics with Python

- Please go to the link below
 - [Tinyurl.com/ece-bdata-python/](https://tinyurl.com/ece-bdata-python/)
- Download the folder
 - DA-Python
- Open Jupyter notebooks related to ARM and Clustering using Anaconda



Thank You 😊
Any questions?

