Deep dive in Hierarchical Clustering

PRESENTED BY -
Aayush Agrawal

Aayush Agrawal

Education -

- MS in Business Analytics, Carlson School of Management, University of Minnesota, 2017
- B.Tech in Electrical Engineering, Malaviya National Institute of Technology, India, 2013

Experience –

- >4 years in Data science, Currently working as Data scientist for Land O' Lakes, Inc.
- Moderator and rank 3rd at https://www.analyticsvidhya.com/
- Kaggle Expert - https://www.kaggle.com/aayushmnit

Reach me out at (@aayushmnit) –

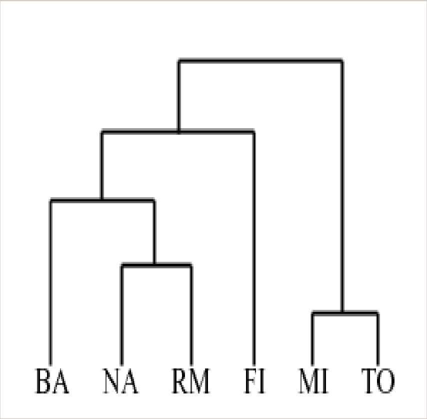
Email - aayushmnit@gmail.com

LinkedIn - https://www.linkedin.com/in/aayushmnit/

Github - https://github.com/aayushmnit

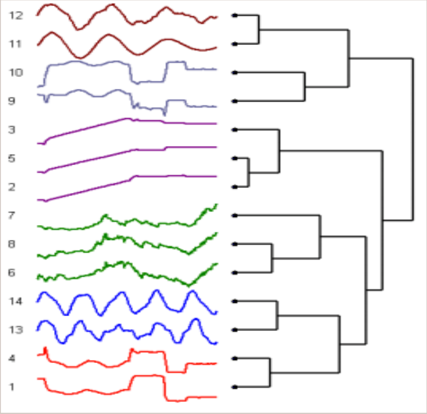
Agenda

What is Clustering



Introduction to Hierarchical clustering

Time Series Hierarchical Clustering



Examples

Clustering is grouping of data

- Unsupervised learning technique which attempts to organizes data points into homogenous groups/cluster

- Desired outcome –
  - **High Intra-similarity**

    Any two points that are assigned in same cluster are similar
  - **Low Inter-similarity**

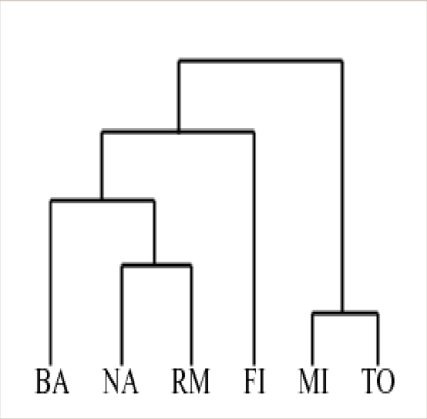
    Any two points that are assigned in different cluster are not similar to each other

- Helps to gain insight into your data – it's easier to look at few groups instead of large data

- Examples – Market segmentation, medical diagnostics, bioinformatic etc.

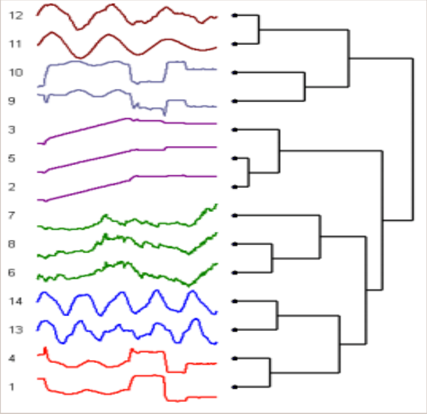
Agenda

What is Clustering



Introduction to Hierarchical clustering

Time Series Hierarchical Clustering



Examples

Types of Hierarchical clustering

Agglomerative (Bottom Up) –

- Initially each point is a cluster
- Repeatedly keeps joining two most similar clusters at a time, until only one cluster is left
- All the intermediate merges are recorded in a special kind of data structure called “Dendogram”, which is the output of the clustering
- Most commonly used

Divisive (Top down) –

- Initially every point is a single cluster
- Repeatedly keeps dividing points until only one point is left in each cluster

Understanding Dendograms

- Hierarchical clustering produces dendogram as the result which shows cluster hierarchy

- Dendrogram shows which clusters were merged at what time, indicating their similarity

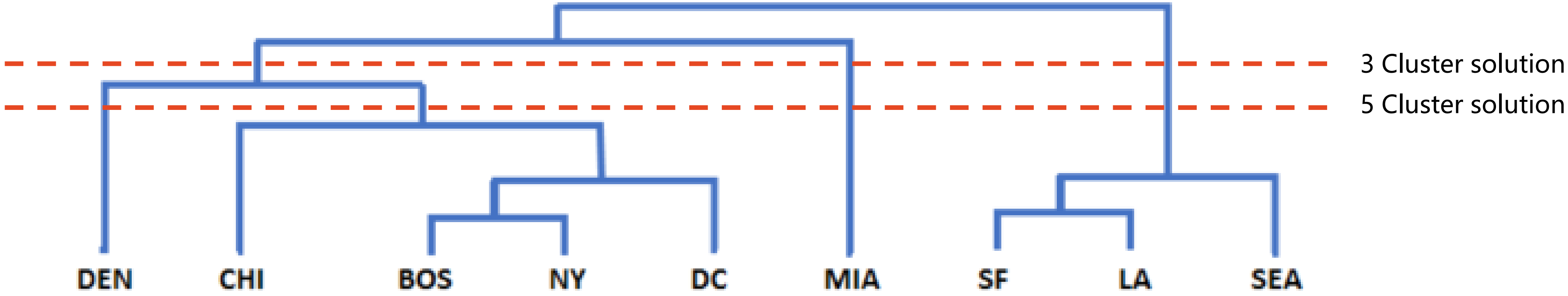


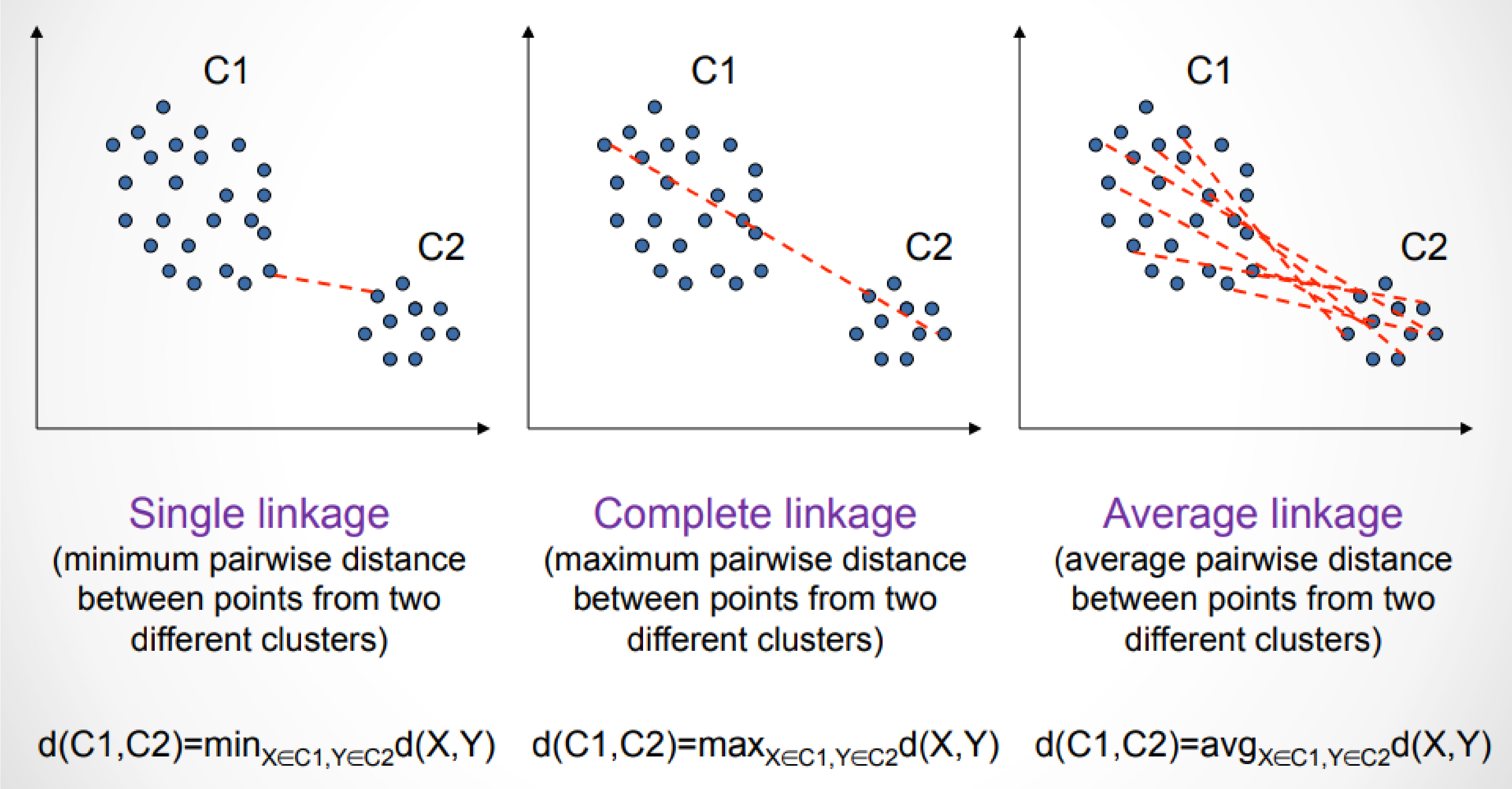
Fig: Hierarchical Clustering(Single linkage) on US major cities and their geographical distance

Hierarchical clustering by scratch

- Refer to excel file (Link)

Hierarchical
lustering by scratcl

Looking at various distance metrics between clusters



Single linkage
(minimum pairwise distance between points from two different clusters)

$d(C1,C2)=\min_{X\in C1,Y\in C2} d(X,Y)$

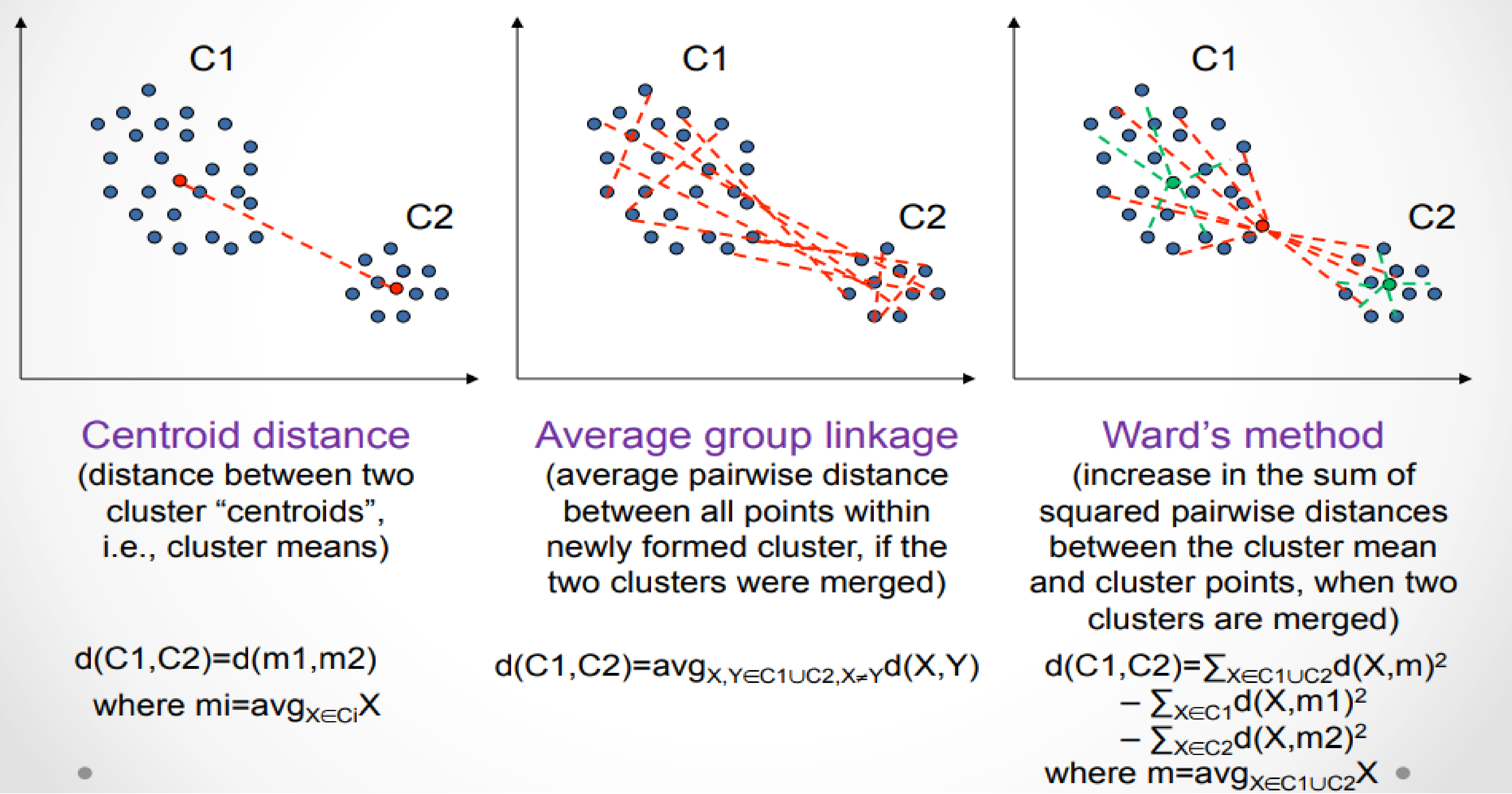
Complete linkage
(maximum pairwise distance between points from two different clusters)

$d(C1,C2)=\max_{X\in C1,Y\in C2} d(X,Y)$

Average linkage
(average pairwise distance between points from two different clusters)

$d(C1,C2)=\mathrm{avg}_{X\in C1,Y\in C2} d(X,Y)$

Looking at various distance metrics between clusters



Centroid distance

(distance between two cluster “centroids”, i.e., cluster means)

$d(C1,C2)=d(m1,m2)$

where $mi=\text{avg}_{X\in Ci}X$

Average group linkage

(average pairwise distance between all points within newly formed cluster, if the two clusters were merged)

$d(C1,C2)=\text{avg}_{X,Y\in C1\cup C2,X\neq Y}d(X,Y)$

Ward’s method

(increase in the sum of squared pairwise distances between the cluster mean and cluster points, when two clusters are merged)

$$d(C1,C2)=\sum_{X\in C1\cup C2}d(X,m)^2 - \sum_{X\in C1}d(X,m1)^2 - \sum_{X\in C2}d(X,m2)^2$$

where $m=\text{avg}_{X\in C1\cup C2}X$

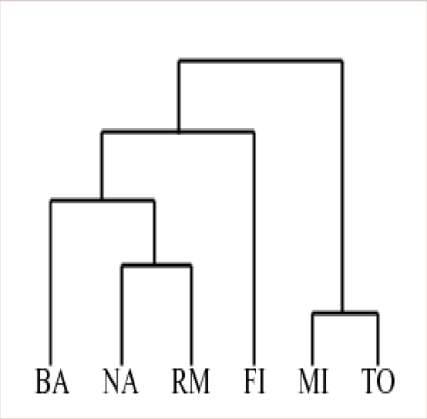
Hierarchical clustering in R

- Refer to PDF file below (Link) –

Iris Example

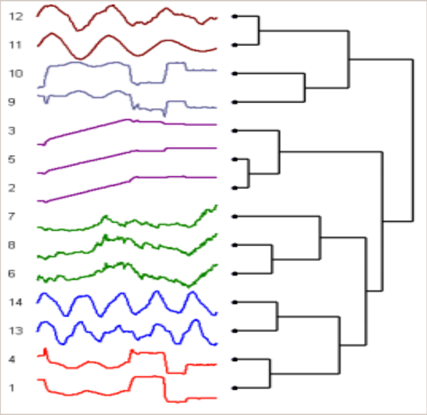
Agenda

What is Clustering



Introduction to Hierarchical clustering

Time Series Hierarchical Clustering



Examples

Scale of data makes it difficult to cluster time series data

- Time series data at different scales makes it difficult to cluster the trend and is more biased on the actual value4

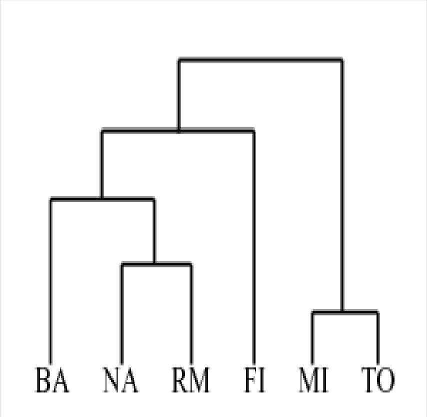
- Most of the models doesn’t account for variation in time series data and just cluster based on scale which makes the clustering exercise irrelevant

- Refer to the doc below (Link)–

Time Series
Example

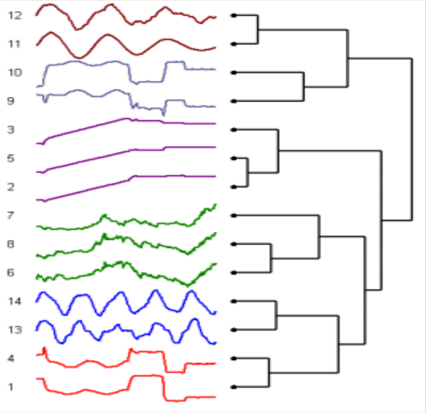
Agenda

What is Clustering



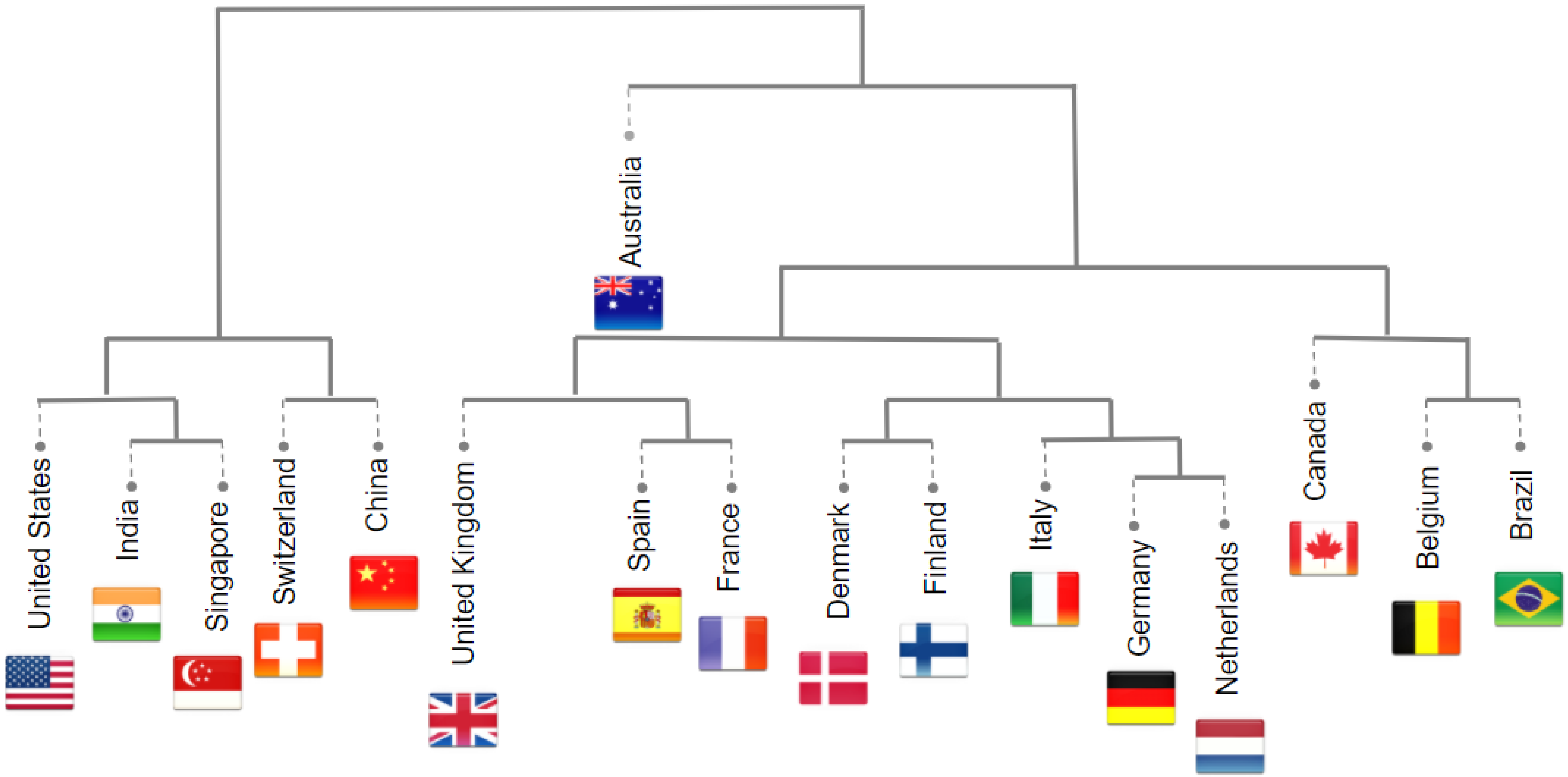
Introduction to Hierarchical clustering

Time Series Hierarchical Clustering

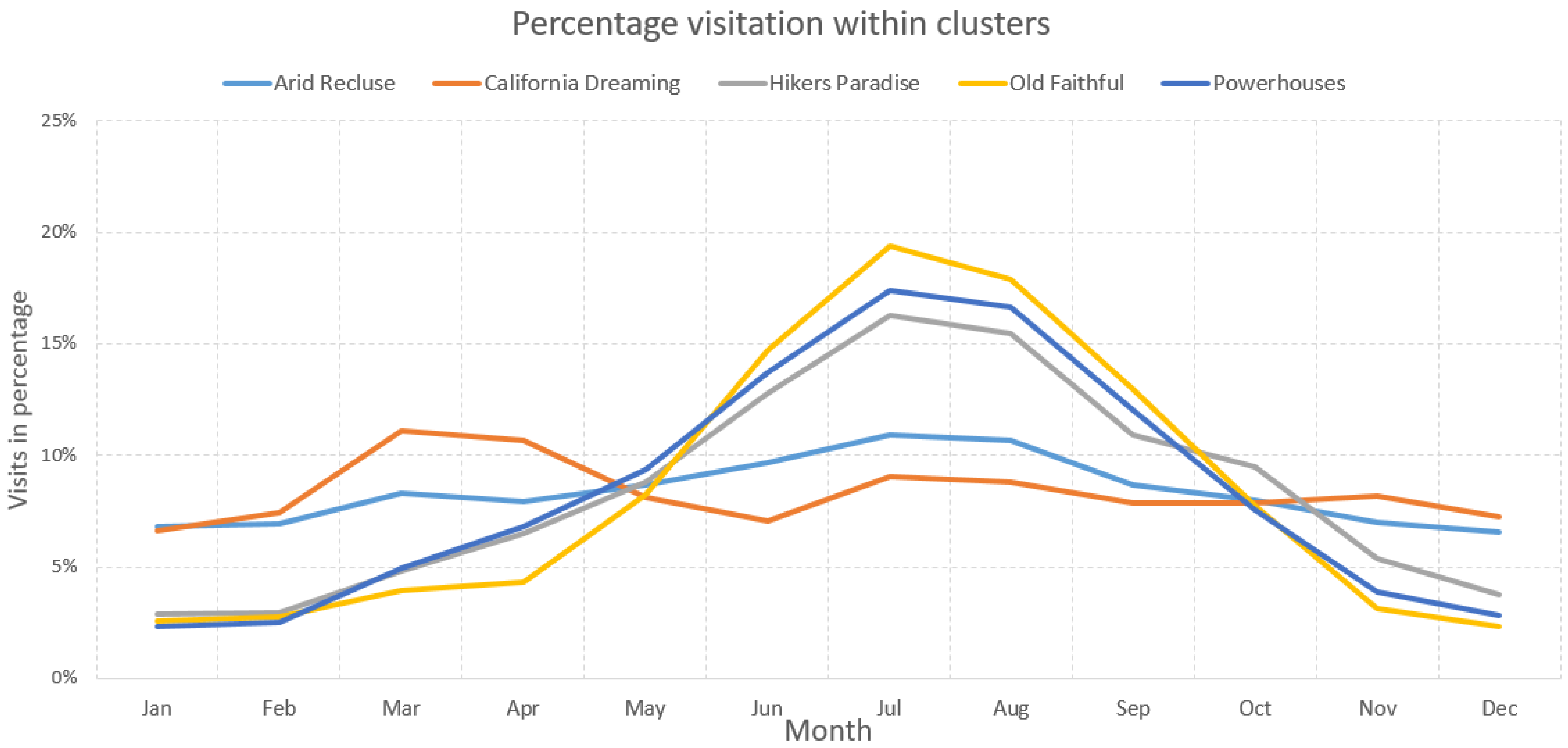


Examples

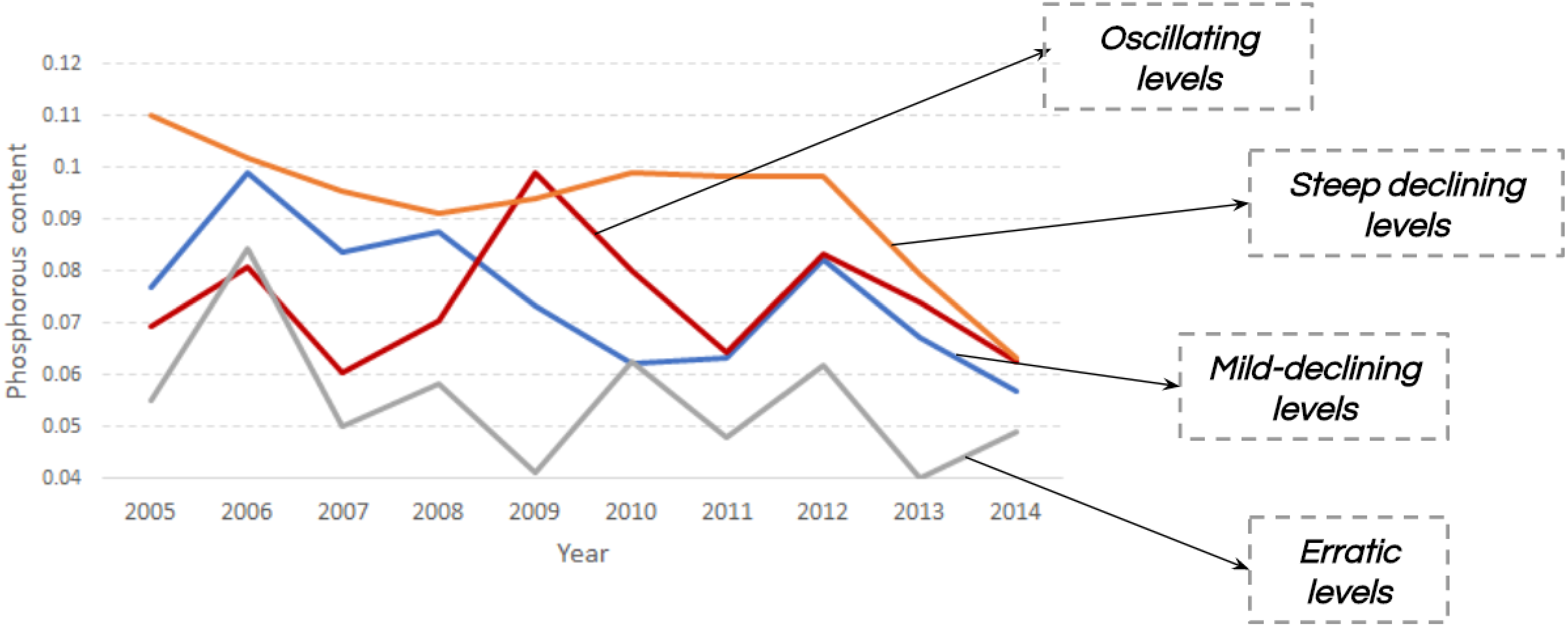
Travel management client - Country revenue clusters show economic, geographic links



National Park visitation



Lake water data challenge – Understanding quality over time



Oscillating levels

Steep declining levels

Mild-declining levels

Erratic levels

Suggested Data science track

Data science path - https://www.analyticsvidhya.com/blog/2017/01/the-most-comprehensive-data-science-learning-plan-for-2017/

My Deep learning track –

1) Machine learning by Andre NG(his first course and the most popular course in MOOC history) -> https://www.coursera.org/learn/machine-learning (Low difficulty)

2) Deep learning by Google on udacity - https://www.udacity.com/course/deep-learning--ud730 (Hard)

3) Practical deep learning for Coders by Jeremy Howard (Former Kaggle #1) - http://course.fast.ai/ (Medium/Hard)

4) A book on deep learning (Goodfellow) - http://www.deeplearningbook.org/ (If you need to understand deep math)

5) Andrew NGs deep learning track - https://www.coursera.org/specializations/deep-learning (easy/medium)

6) Just some collection of good blogs - http://colah.github.io/

Thank You!