ILLINOIS
Gies College of Business

"Drowning in Data, Starving for Knowledge"

Data Exploration and Pattern Detection via Partitioning

Sridhar Seshadri

Contents

What is clustering? Why do we cluster?

How does one cluster?

Partitional Clustering: K-means

Hierarchical (agglomerative) Clustering

Clustering: What Is It and Why?

Example Questions From Data?

Types of **pages** are there on the **Web**?

Types of **customers** are there in my **market**?

Types of **people** are there on a Social **network**?

Types of **e-mails** in my **Inbox**?

Example Questions From Data?

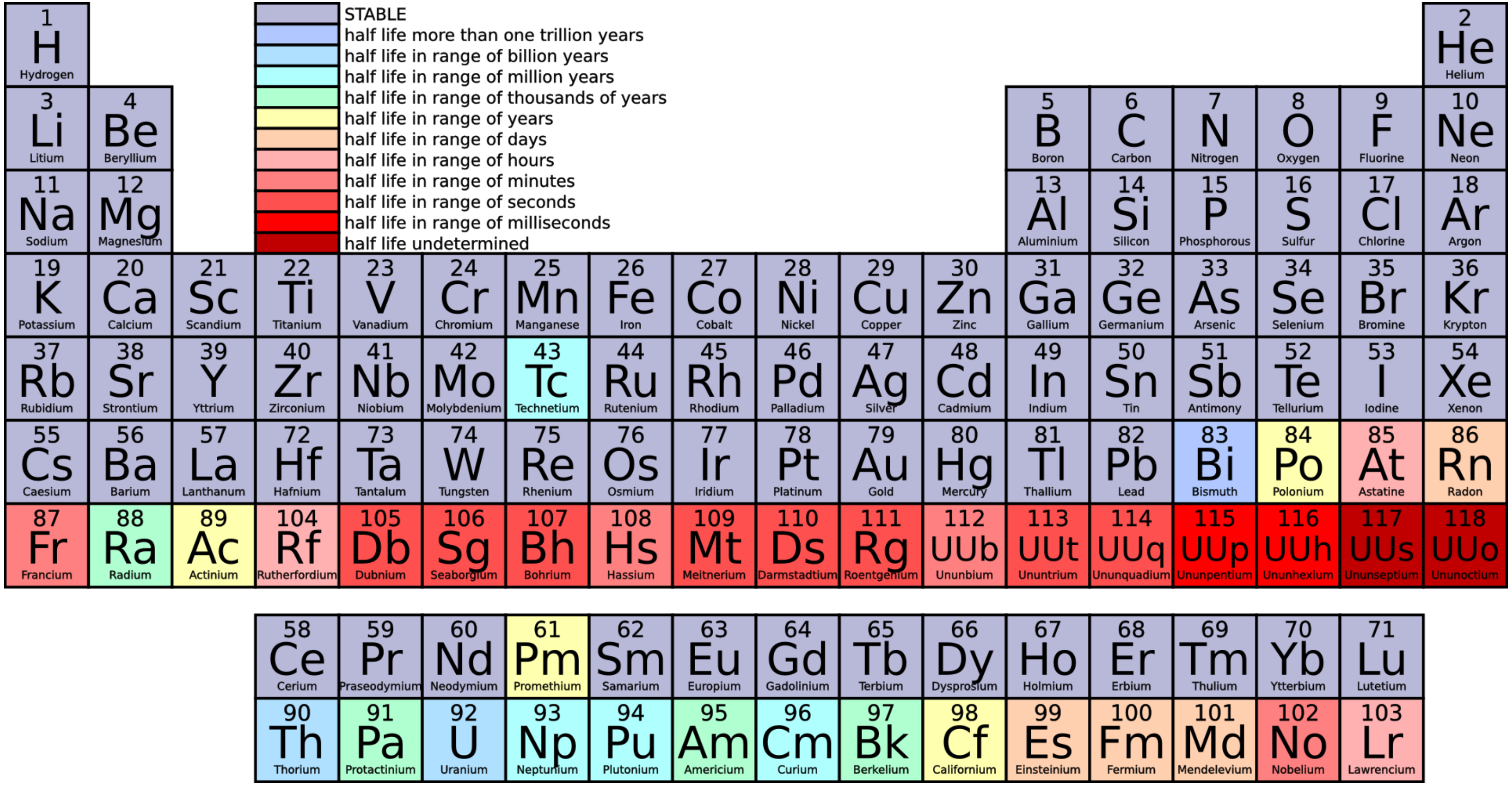
Types of **genes** the **human genome** has?

Types of **stars/planets/galaxies** are in the **universe**?

Types of **songs/movies** in my **audio library**?

Types of **animals/plants** are there on **Earth**?

Nature Is “Naturally” Organized



Hierarchy of Organization

http://vlib.org/

The WWW Virtual Library

Quick se

Agriculture
Irrigation, Livestock, Poultry Science, ...

The Arts
Art History, Classical Music, Theatre and Drama, ...

Business and Economics
Finance, Marketing, Transportation, ...

Communications and Media
Broadcasters, Publishers, Telecommunications, ...

Computing and Computer Science
Artificial Intelligence, Cryptography, Logic Programming, ...

Education
Primary, Secondary, Tertiary, ...

Engineering
Architecture, Electrical, Mechanical, ...

Humanities and Humanistic Studies
History, Languages and Linguistics, Museums, ...

Information and Libraries
Information Quality, Knowledge Management, Libraries, ...

International Affairs
International Relations and Security, Sustainable Development, ...

Law
Arbitration, Forensic Toxicology, Legal History, ...

Natural Sciences and Mathematics
Biosciences, Earth Science, Medicine and Health, Physics, ...

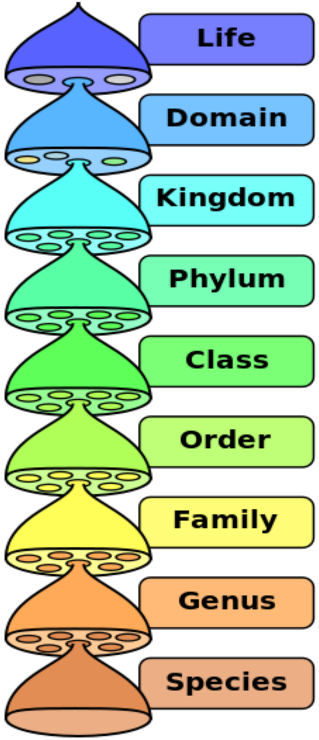
Recreation
Gardening, Recreation and Games, Sport, ...

Regional Studies
African, Asian, Latin American, European, ...

Social and Behavioural Sciences
Anthropology, Archaeology, Population and Development Studies, ...

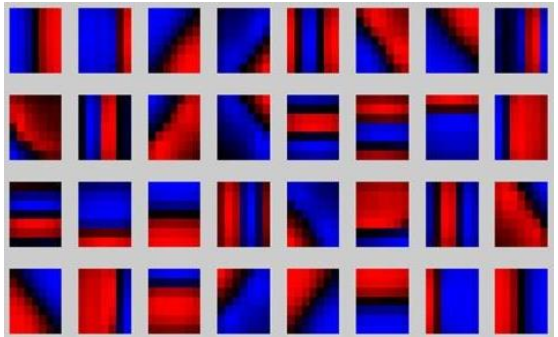
Society
Peoples, Religion, Gender Studies, ...

Biological Classification



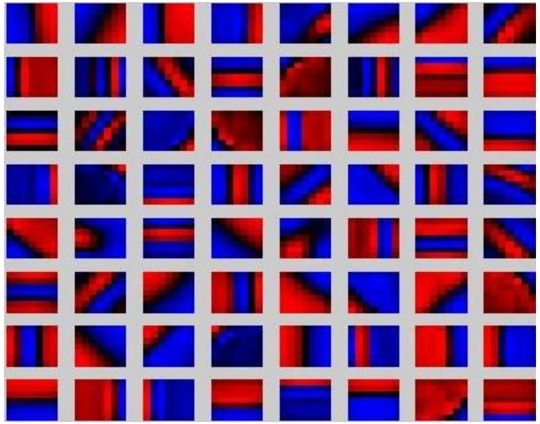
http://en.wikipedia.org/wiki/Biological_classification

Cluster to Create Features

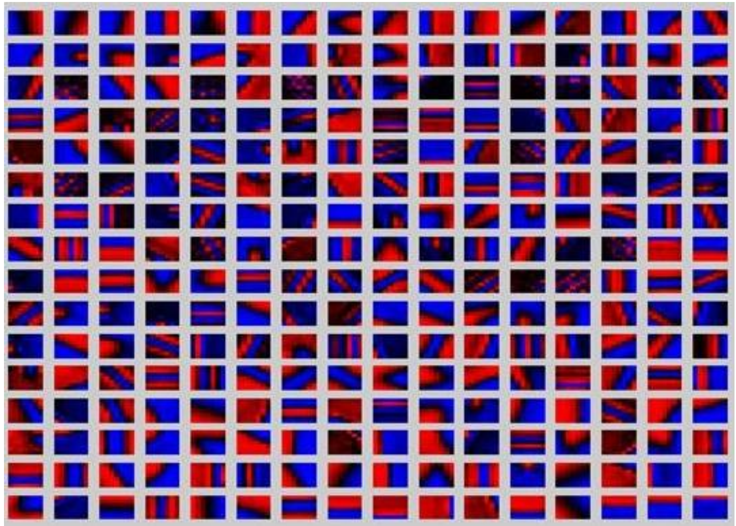


32 clusters

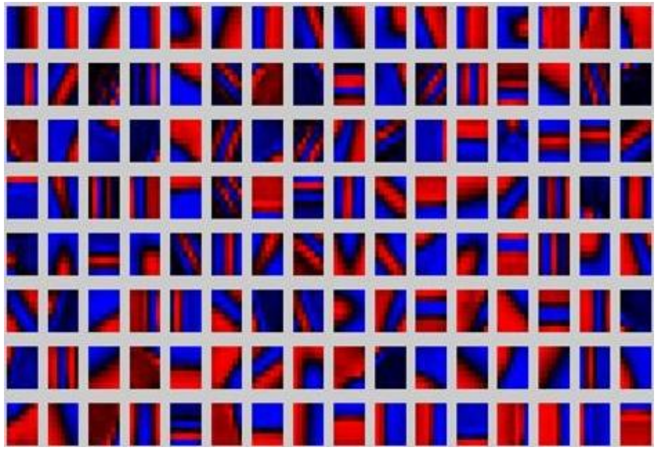
64 clusters



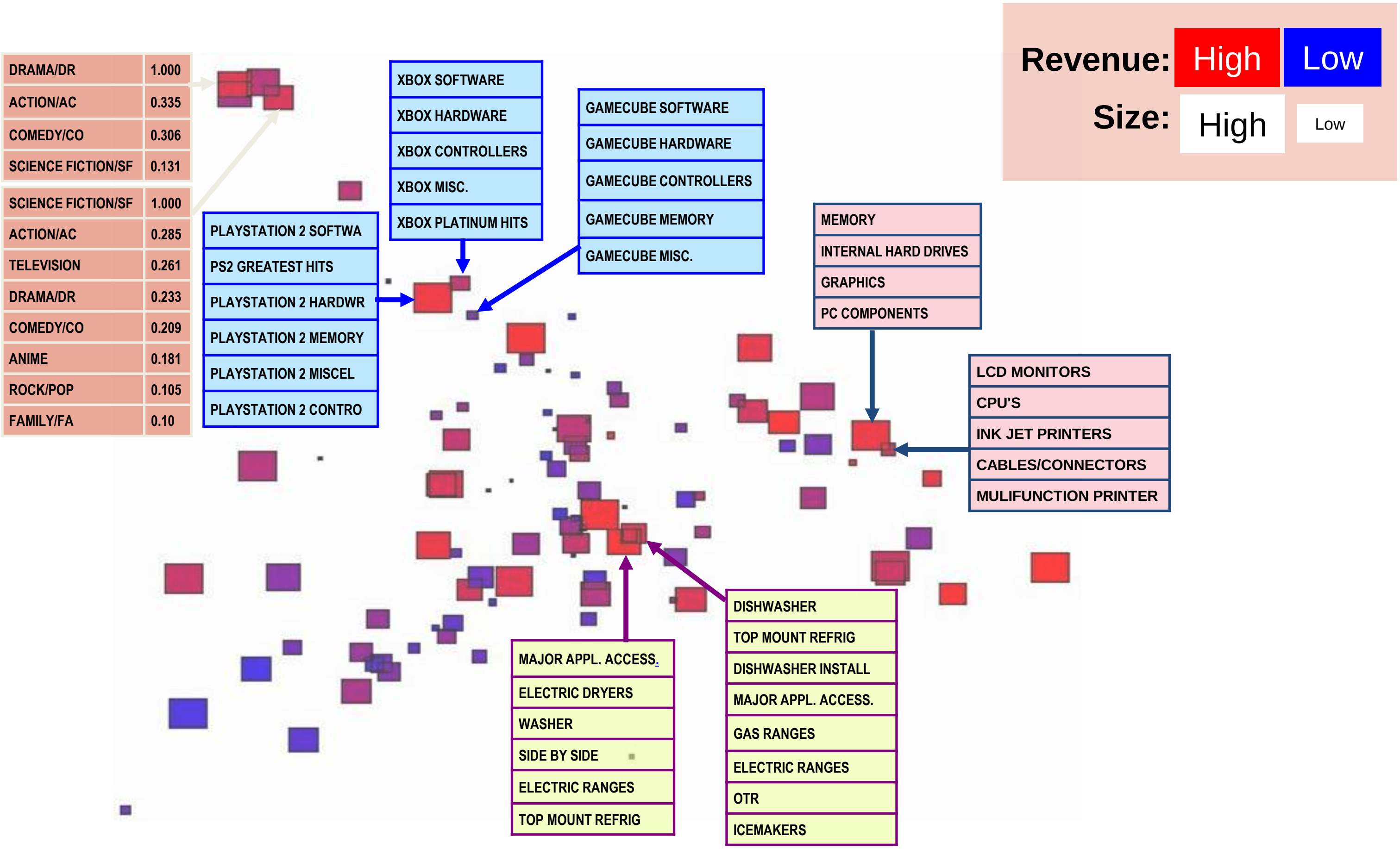
256 clusters



128 clusters



Customer Clustering by Purchase



Applications

Forecasting

Pricing

Strange patterns of residuals might imply a cluster

Missing data might mean a cluster!

So Why Clustering?

Clustering = Grouping "**Similar**" things together!

Understand/discover structure in data

Summarize data points by their "cluster center"

Compress data variability into "representative vectors"

So Why Clustering?

Clustering = Grouping "**Similar**" things together!

Extract features from data for Supervised Learning

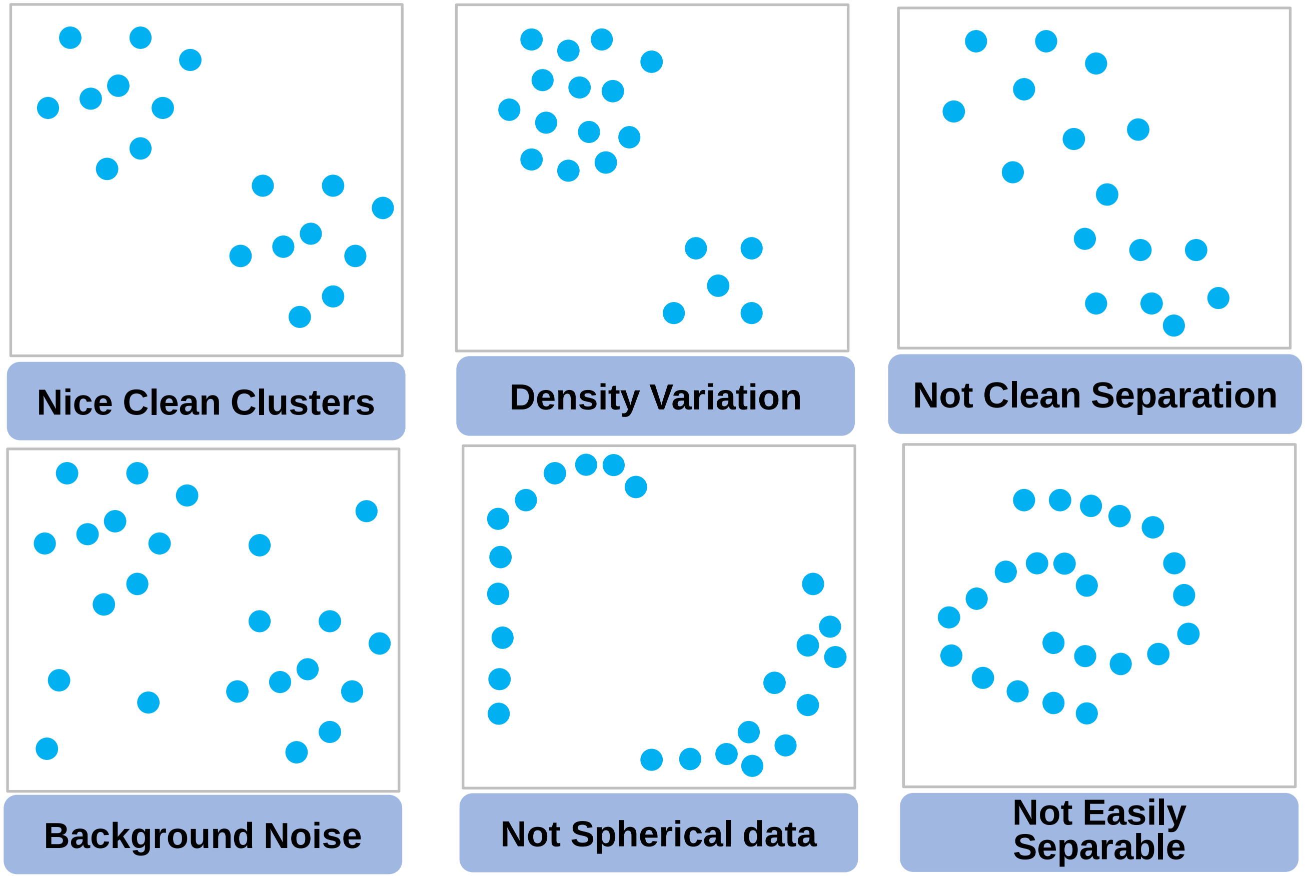
Generate class labels when not known

Business rules can be uncovered

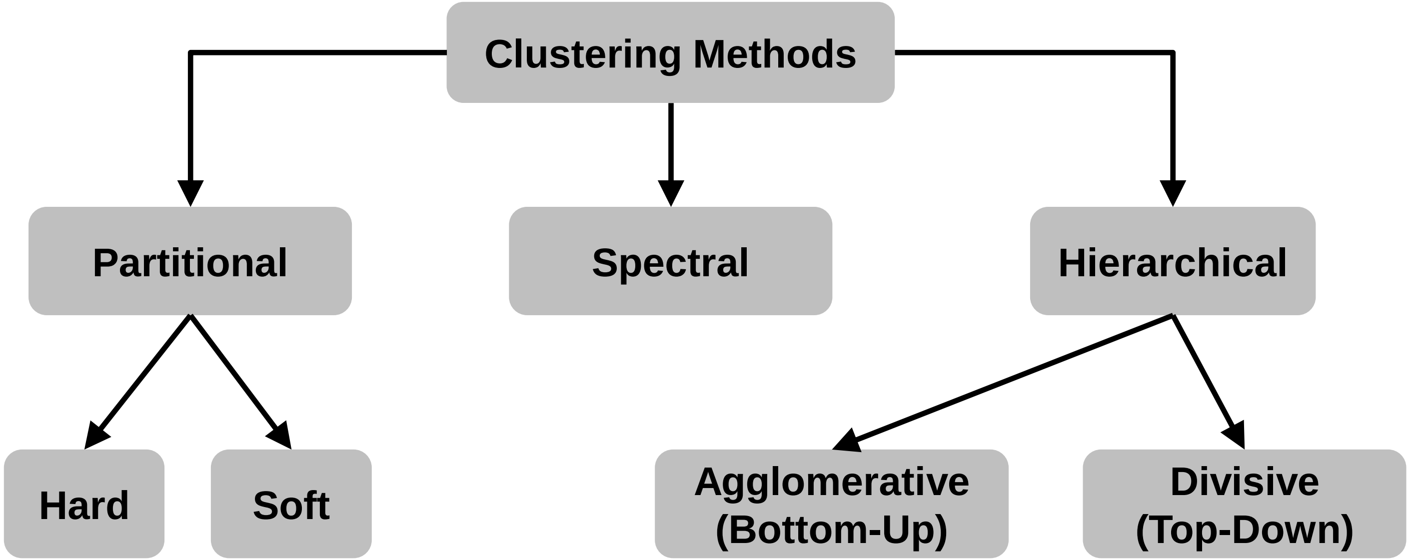
How Does One Cluster?

It depends!

Why is Clustering Non-trivial?



Clustering Methods

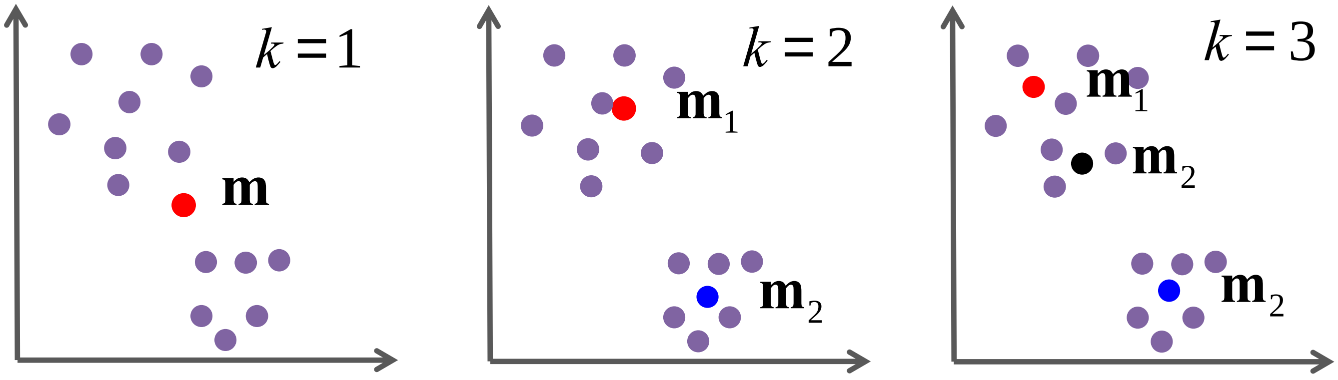


Key to Clustering is how we define “**Distance**” or “**Similarity**” between two data points

Partitional Clustering

K-Means Clustering

K-Means – Objective Function



What are the **Parameters**?

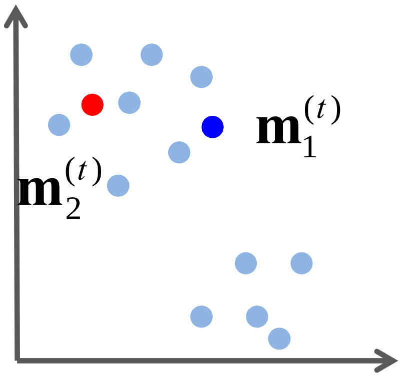
What is the **Objective Function**?

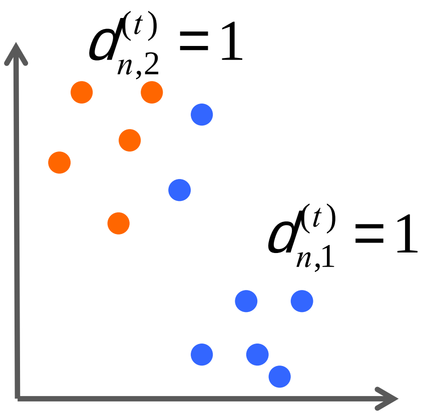
What is the **Model Complexity**?

$$J\left(\mathbf{M}=\{\mathbf{m}_k\}_{k=1}^{K}\right)=\sum_{n=1}^{N}\sum_{k=1}^{K} d_{n,k}\left\|\mathbf{x}^{(n)}-\mathbf{m}_k\right\|_2^2$$

K-Means: 1. Expectation

cluster *centers* → cluster *assignments*

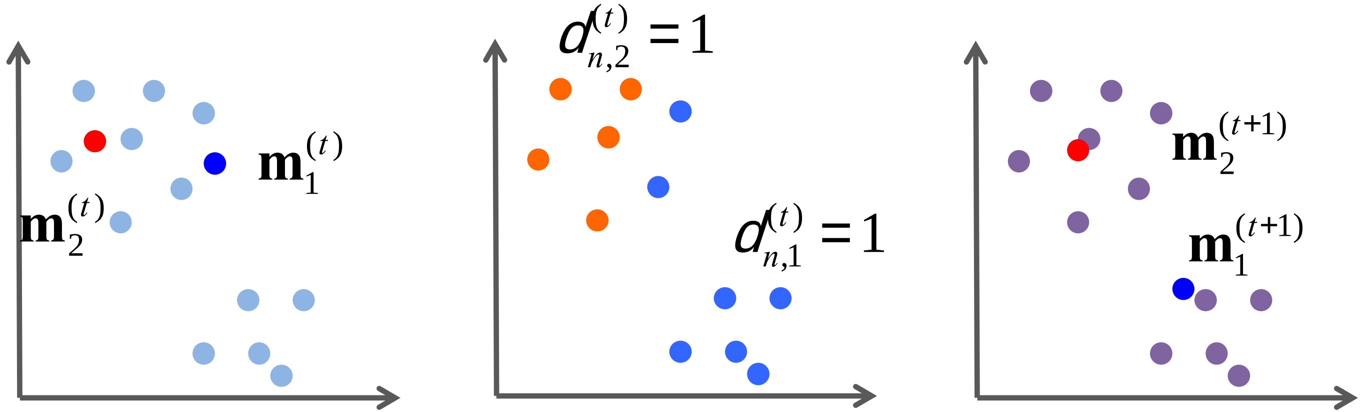




$$d_{n,k}^{(t+1)} = \left(k == \arg\min_{j=1\ldots K} \left\{ \mathrm{D}\left(\mathbf{x}^{(n)}, \mathbf{m}_j^{(t)} \right) \right\} \right)$$

K-Means: 2. Maximization

cluster *assignments* → cluster *centers*



$$\mathbf{m}_k^{(t+1)} \leftarrow \frac{\sum_{n=1}^{N} \delta_{n,k}^{(t)} \mathbf{x}^n}{\sum_{n=1}^{N} \delta_{n,k}^{(t)}}$$

Technical considerations

Distance measures

Initialization

Number of clusters

Rattle

Build a model

Improve a model

Data- BestFit's Body_Measure

Objective – Find a few groups(clusters) on the basis of the given attributes

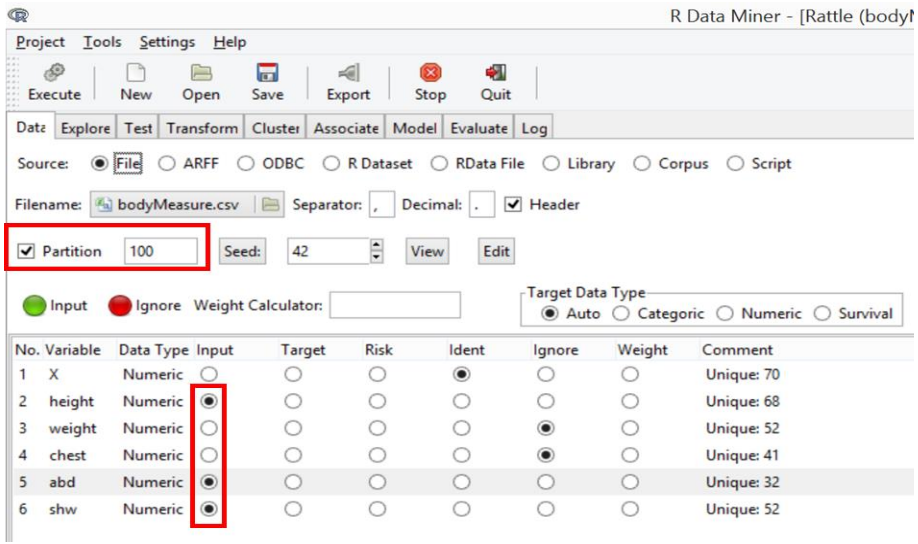
Variable	Description
Height	The height measured in (cm)
Weight	The weight measured in (kg)
Chest	The circumference of chest measured in (cm)
abd	The circumference of abdomen measured in (cm)
shw	The shoulder width measured in (cm)

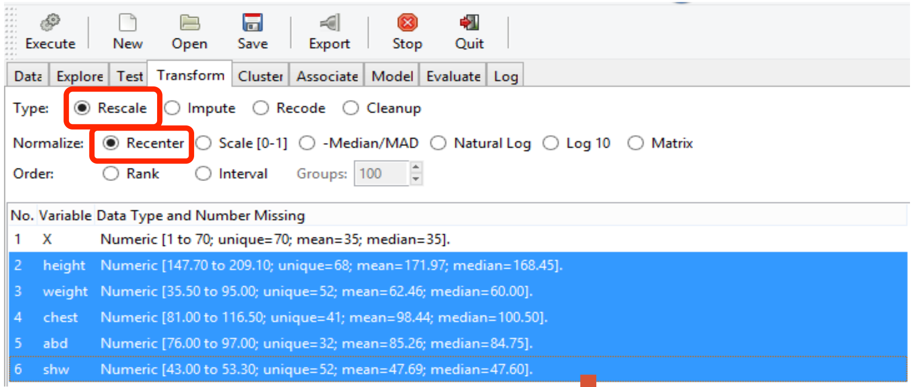
First five rows of data

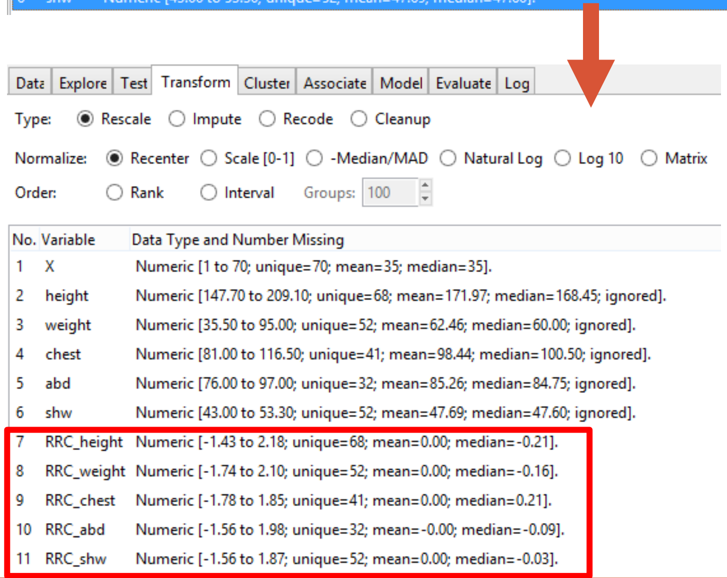
height	weight	chest	abd	shw
168	61.5	98.5	85	47.5
149.3	53	90	78	44.9
148.4	44.5	89.5	77	43.9
195.5	91.5	111	94.5	52.7
159.1	52.5	84	79.5	44.7

bodyMeasure.csv

Rescale the Data for Distance Based Algorithms (KNN)





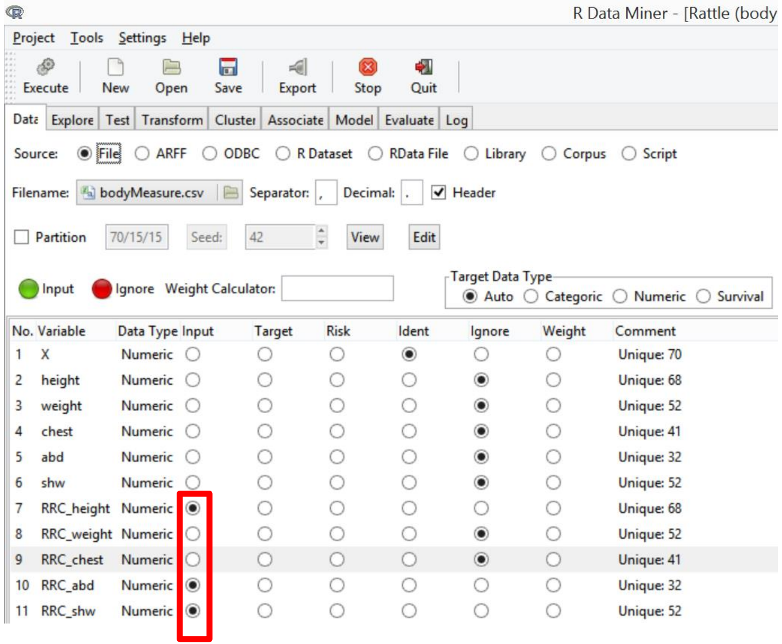


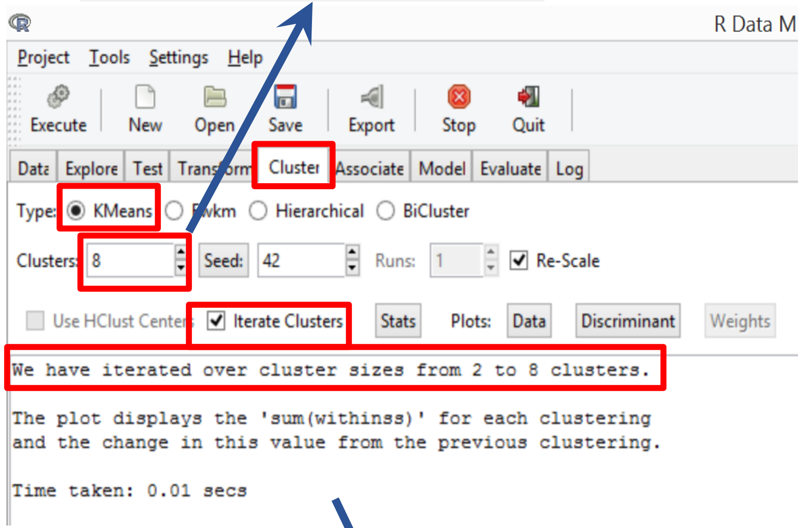
Large scale should not affect distances. We transform the data.

Source: Rattle GUI / Togaware

Select the Input Variables and Select the Best K

Max. K – user given

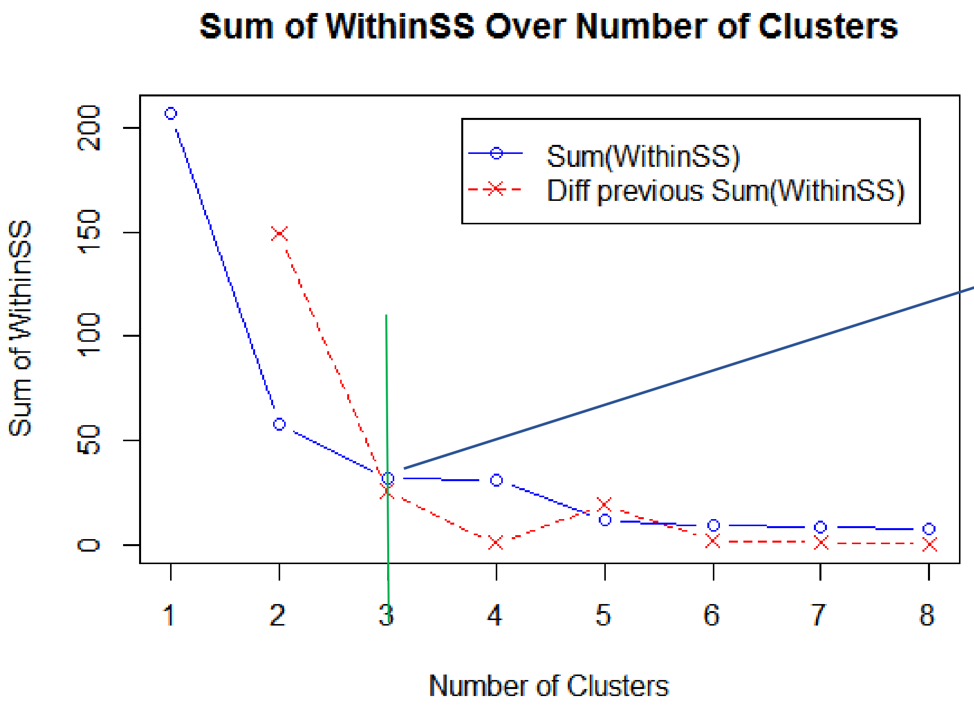




The plot for this command is reported on the next slide. Plot is in R!

Source: Rattle GUI / Togaware

Selecting the Best K

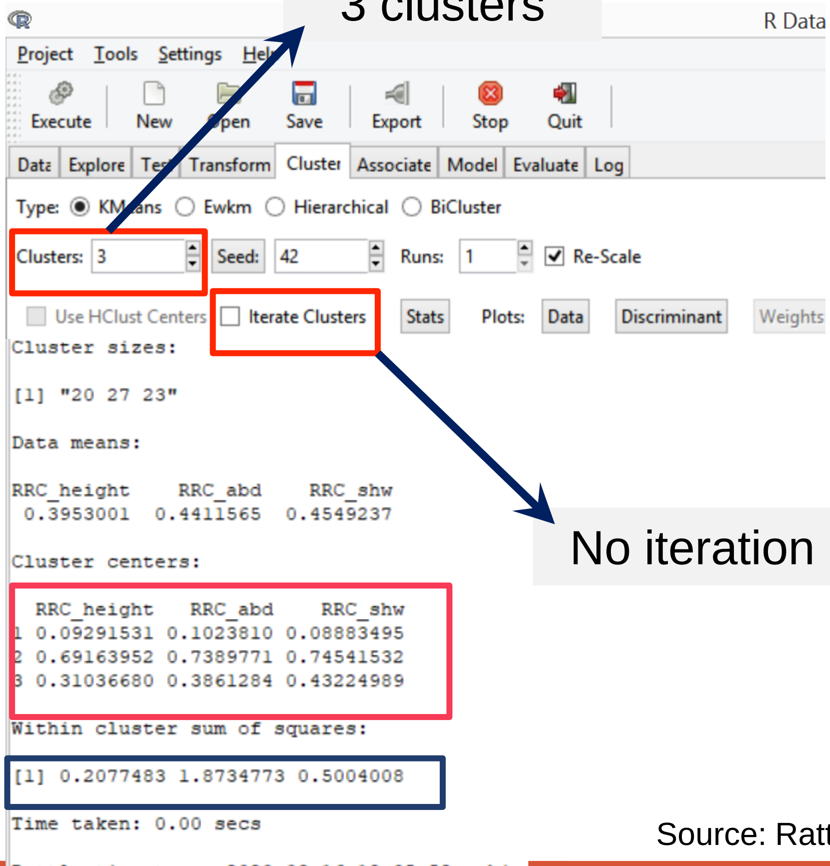


At K=3, the red line cuts the blue line from above, thus we select K=3.
This is a heuristic rule!

Output slightly different in R version 3.6.3

Running K Means for K=3

3 clusters



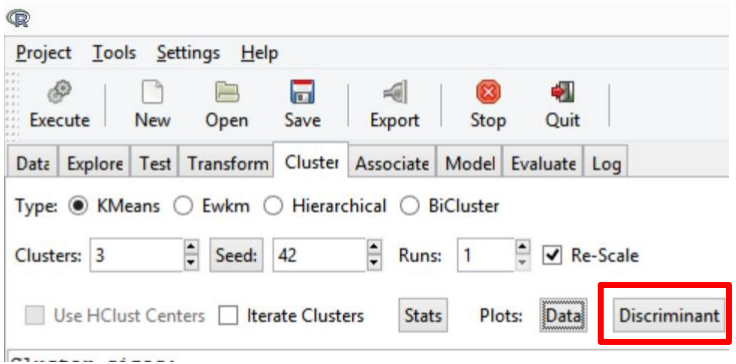
No iteration

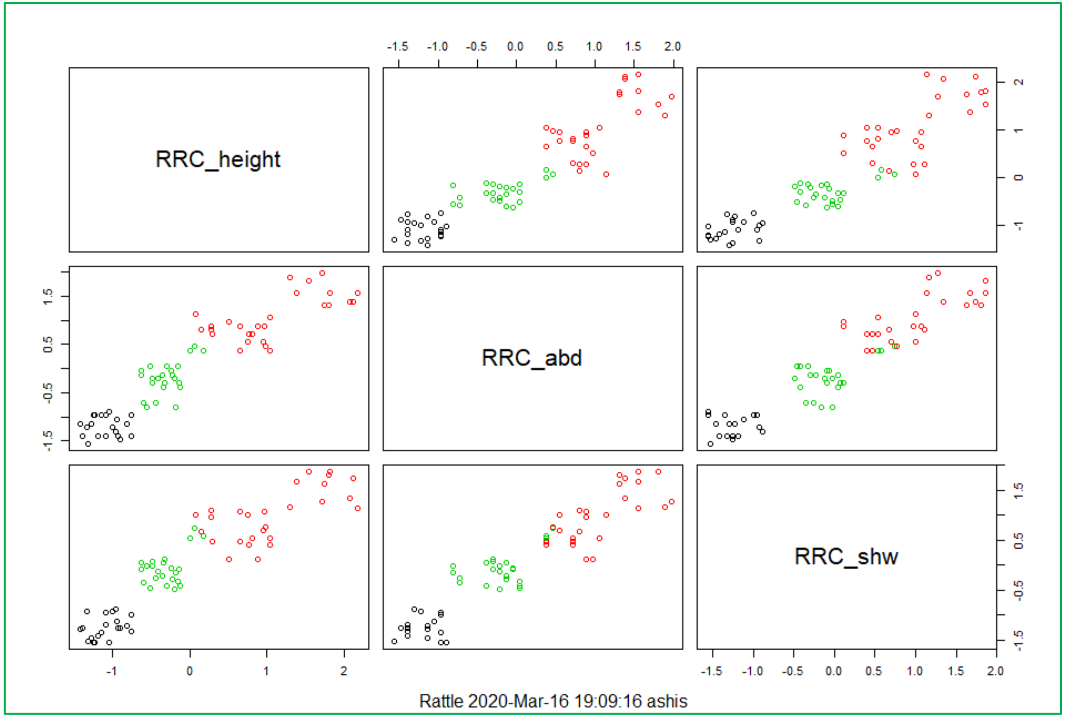
Once a model has been built, the Stats, Data Plot, and Discriminant Plot buttons become available

Source: Rattle GUI / Togaware

Kmeans – Plot

Output slightly different in R version 3.6.3

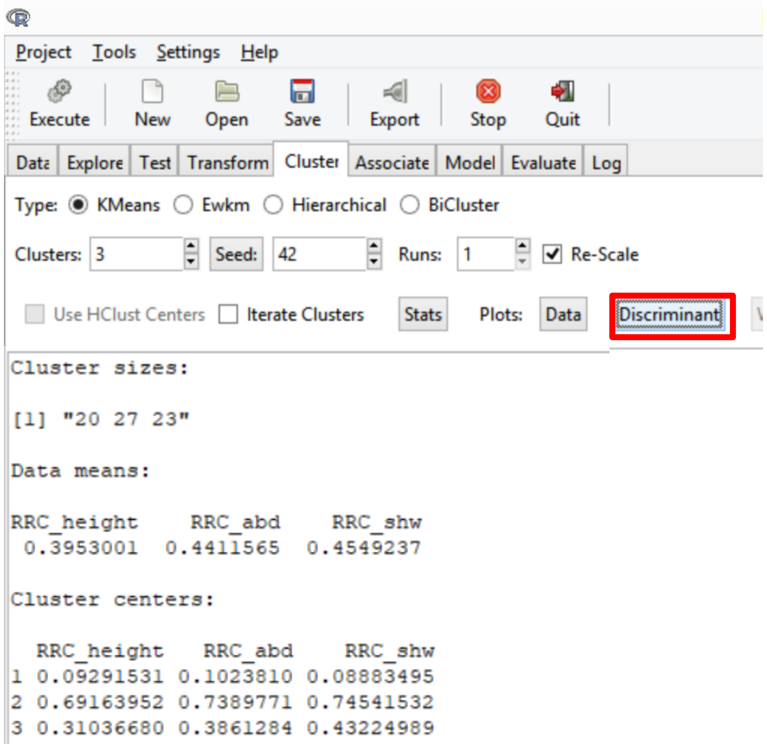


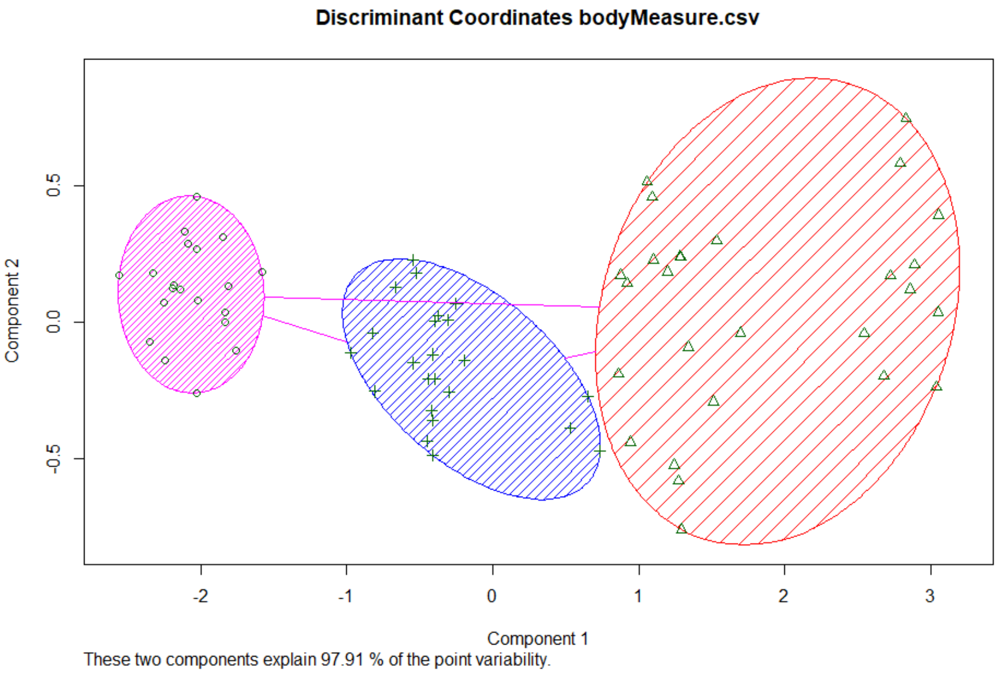


Source: Rattle GUI / Togaware

Output slightly different in R version 3.6.3

Kmeans - Discriminant Plot





Source: Rattle GUI / Togaware

Limitations

High dimensional data

Knowledge of how many clusters

Hierarchical (Agglomerative) Clustering

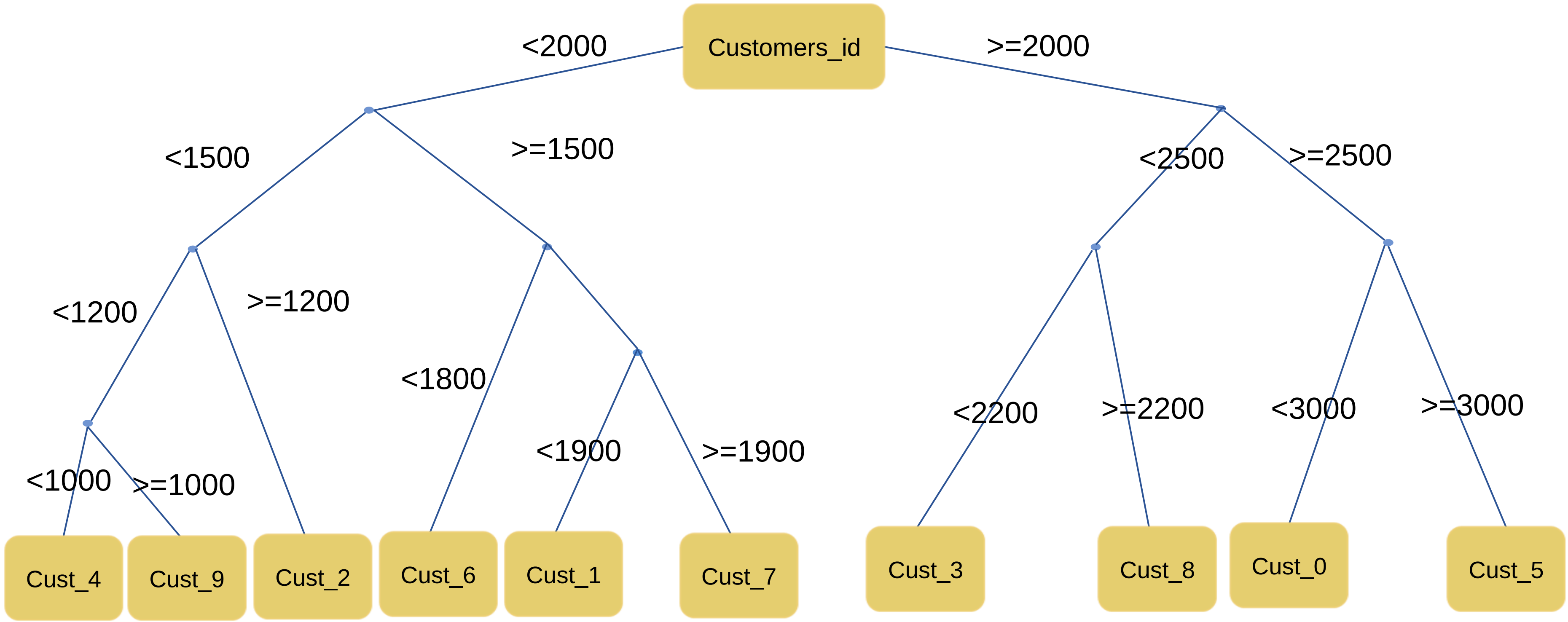
Clustering Example

A B2B firm wants to work on pricing policy for their set of customer based on their purchasing pattern. The analyst in the company selected top 10 representative accounts and analyzed their monthly purchasing (in $ 000's) and wants to group (cluster) them in some meaningful way.

Clustering Example

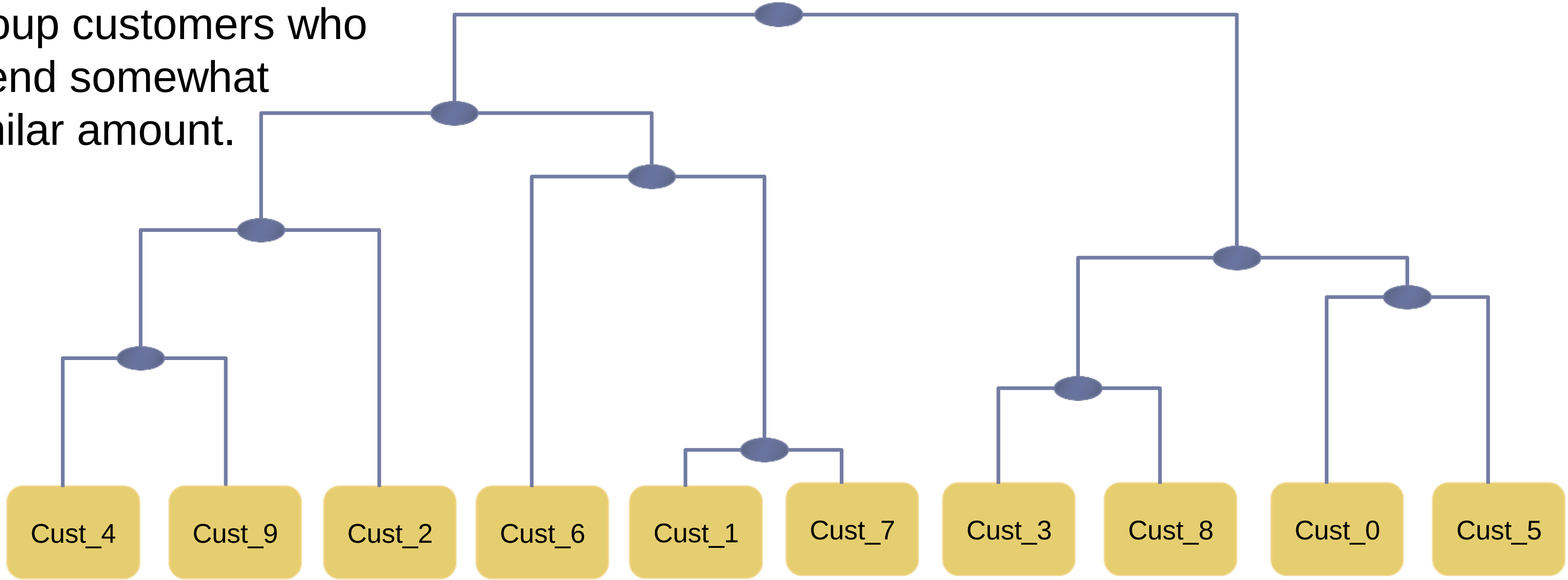
The task also involves deciding in how many groups should they be divided ?

Classification of Key Accounts on Purchase Amount ($ 000s)



Bottom-Up (Agglomerative)

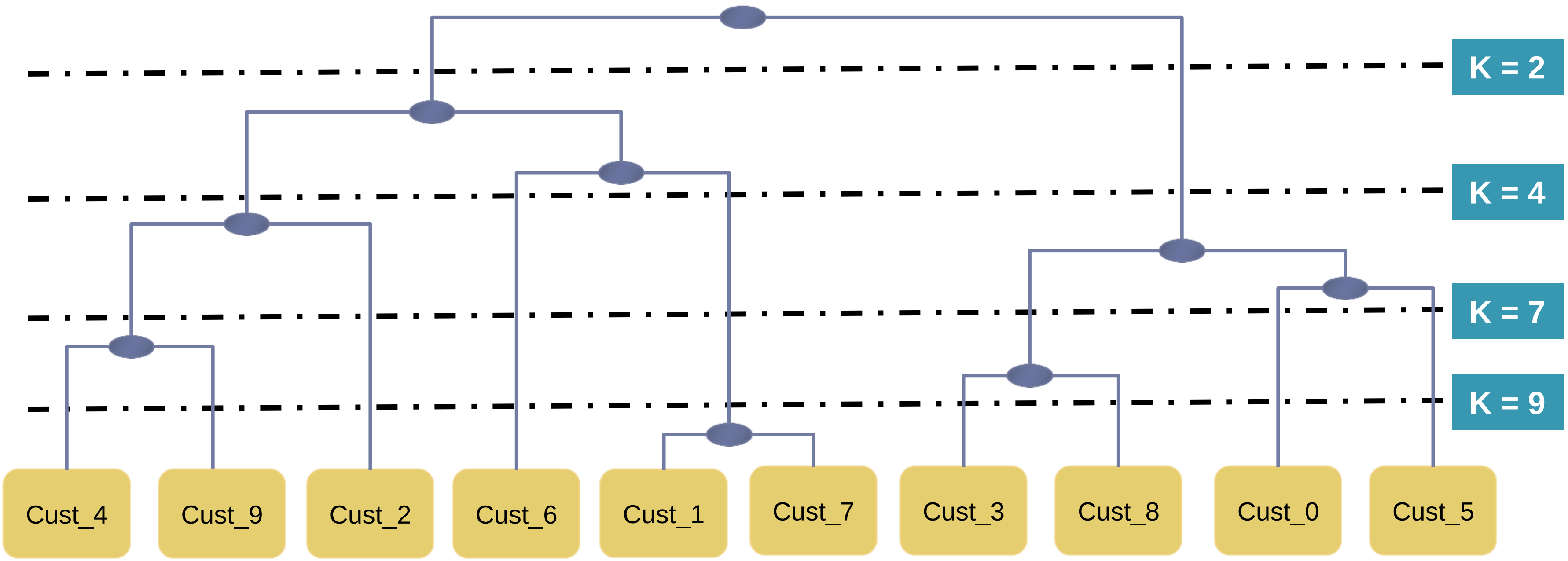
Group customers who spend somewhat similar amount.



SIMILARITY between two **DATA POINTS**? **Sim**(0, 5)

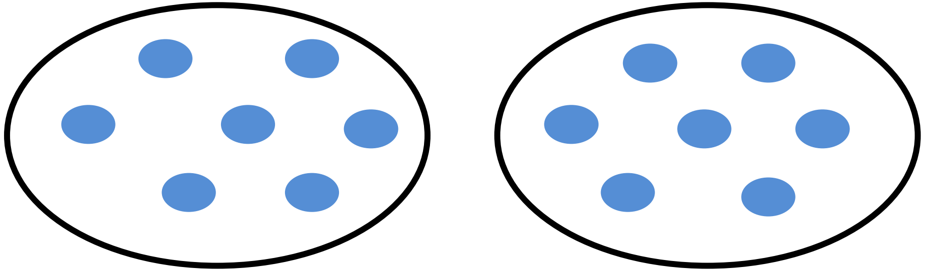
SIMILARITY between **CLUSTERS**? **Sim**({3, 8}, {0, 5})

How Many Clusters? Depends!

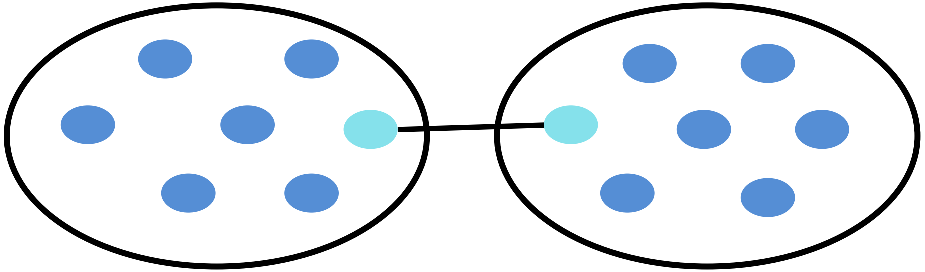


Number of clusters depends on where we CUT the dendrogram.

Similarity Between Clusters

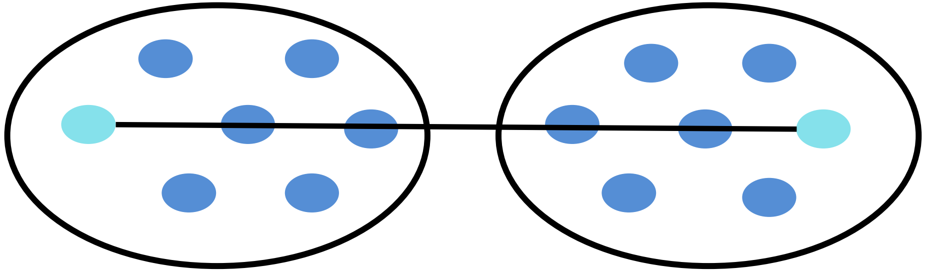
Sim()

Single Linkage
Nearest Neighbor



$$D\left(C_a, C_b\right) = \min_{\mathbf{x}\in C_a, \mathbf{y}\in C_b} \left\{ D\left(\mathbf{x}, \mathbf{y}\right) \right\}$$

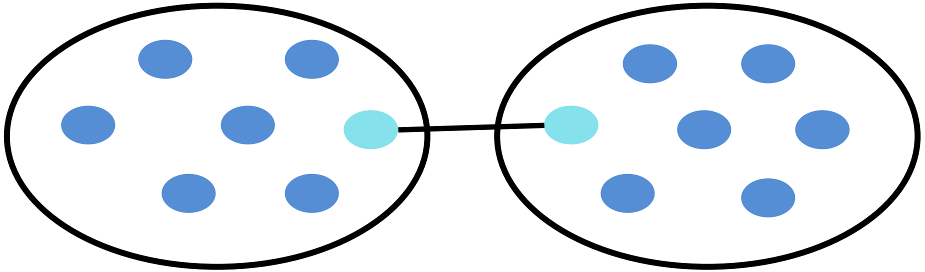
Complete Linkage
Farthest Neighbor



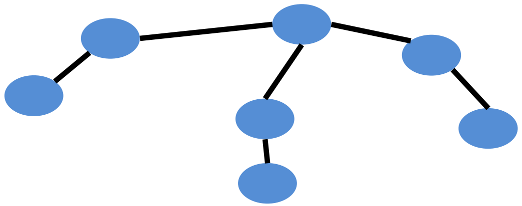
$$D\left(C_a, C_b\right) = \max_{\mathbf{x}\in C_a, \mathbf{y}\in C_b} \left\{ D\left(\mathbf{x}, \mathbf{y}\right) \right\}$$

Single Link Examples

Single Linkage
Nearest Neighbor

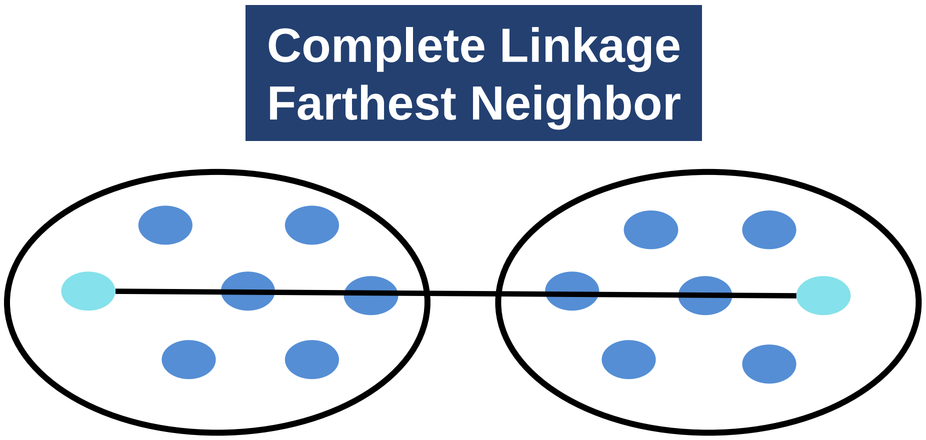


$$D(\mathbf{C}_a, \mathbf{C}_b) = \min_{\mathbf{x} \in \mathbf{C}_a, \mathbf{y} \in \mathbf{C}_b} \{ D(\mathbf{x}, \mathbf{y}) \}$$

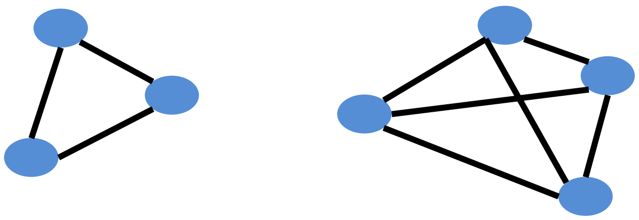


- **Elongated** Clusters
- Minimum Spanning Trees

Complete Linkage Example

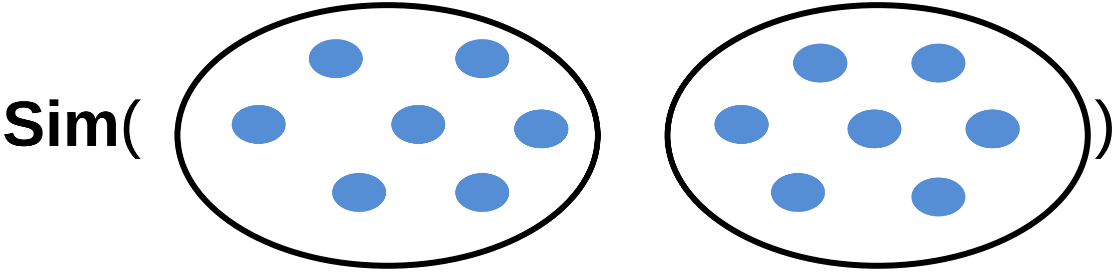


$$D\left(\mathbf{C}_a, \mathbf{C}_b\right) = \max_{\mathbf{x} \in \mathbf{C}_a, \mathbf{y} \in \mathbf{C}_b} \left\{ D\left(\mathbf{x}, \mathbf{y}\right) \right\}$$

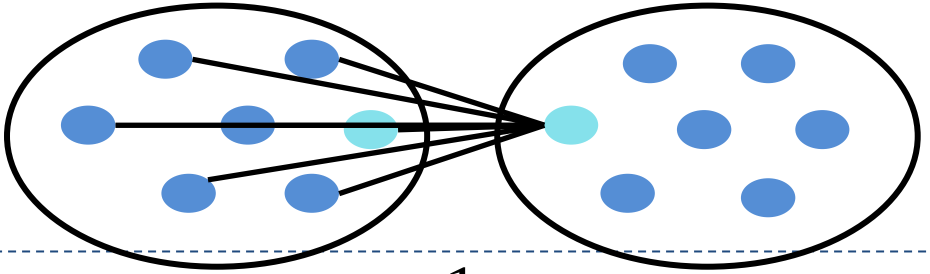


- **Compact** Clusters

Similarity between Clusters

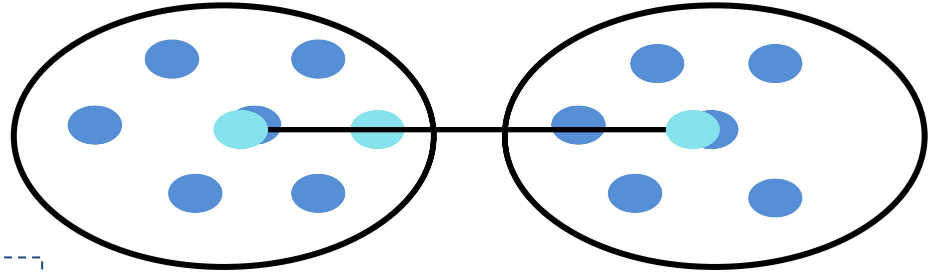


Average Linkage
All Neighbors



$$D\left(\mathbf{C}_a,\mathbf{C}_b\right)=\frac{1}{\left|\mathbf{C}_a\right|\left|\mathbf{C}_b\right|}\sum_{\mathbf{x}\in\mathbf{C}_a}\sum_{\mathbf{y}\in\mathbf{C}_b}D\left(\mathbf{x},\mathbf{y}\right)$$

Mean Linkage
Mean Neighbor



$$D\left(\mathbf{C}_a,\mathbf{C}_b\right)=\left\|\mathbf{m}_a-\mathbf{m}_b\right\|$$

Now It All Boils Down To…

$$D(\mathbf{x}, \mathbf{y})$$

Distance or **Similarity** between two points

Foundation of all Machine Learning algorithms

> "**Similar** things have **similar** properties (e.g. class/cluster labels)"

Now it all boils down to…

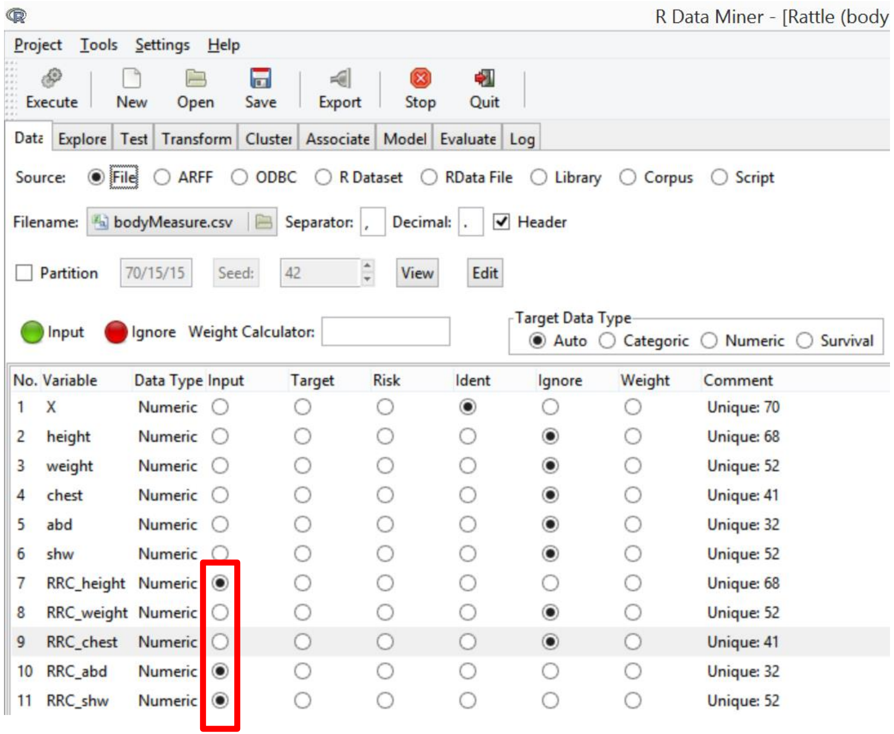
$$D(\mathbf{x}, \mathbf{y})$$

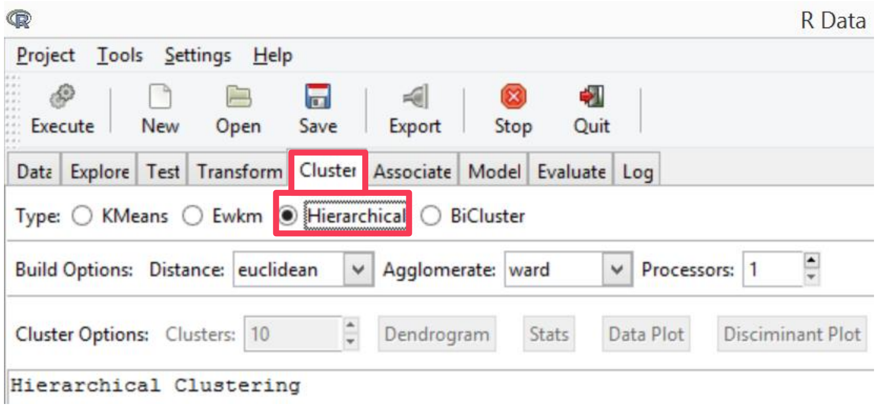
Depends on the domain

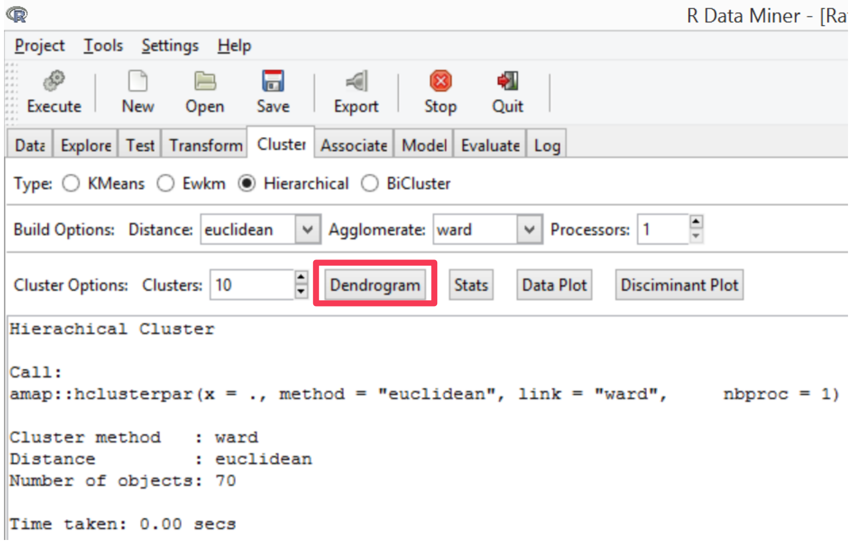
Multi-variate numeric, categorical, Text, Bag, Basket, Image,…

Can be static (constant) or dynamic (learnt)

Data: bodyMeasure

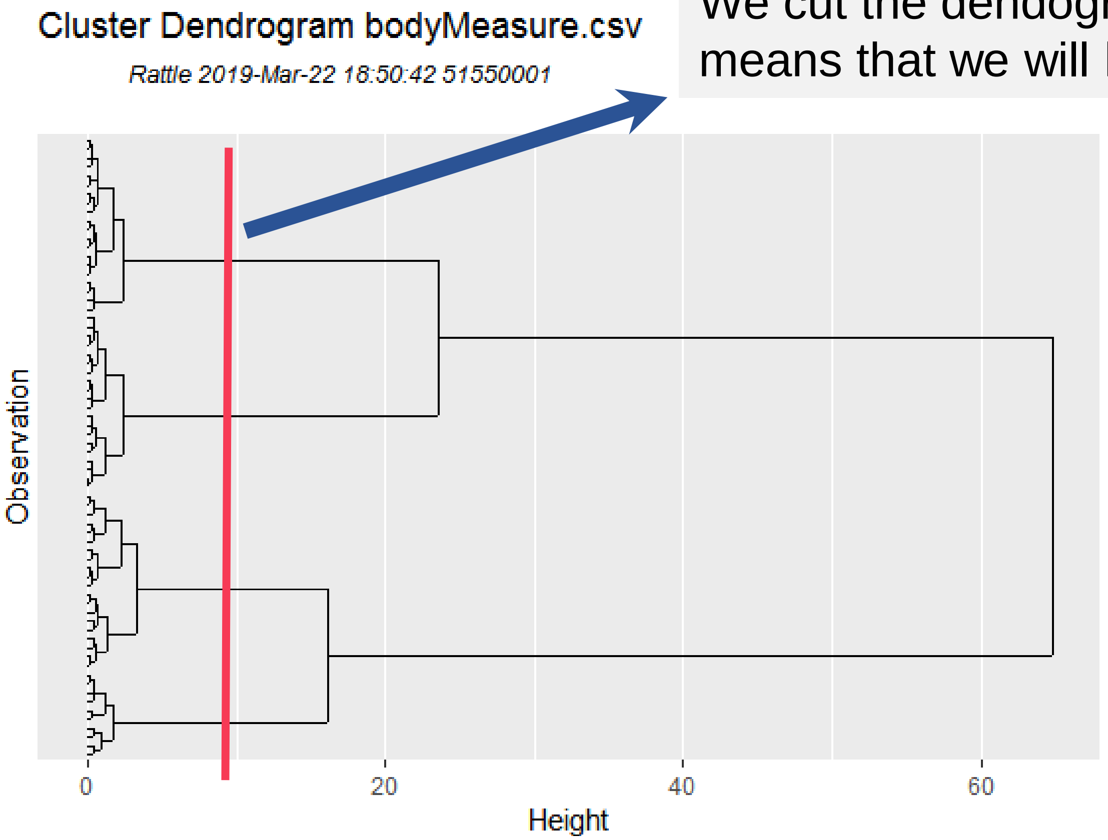






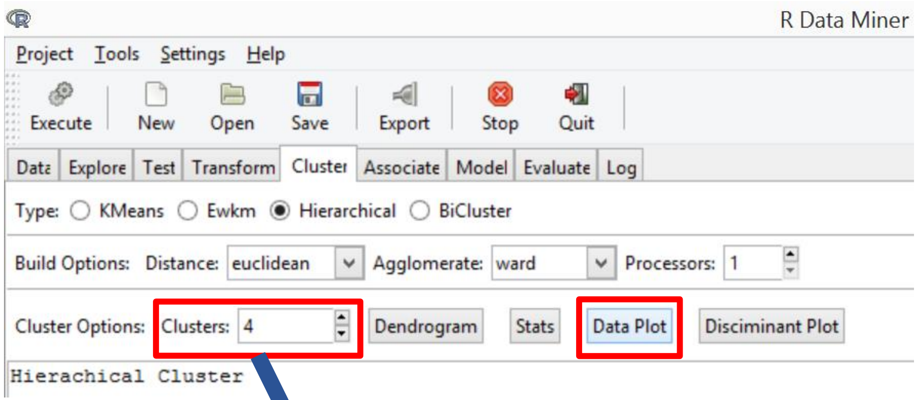
Source: Rattle GUI / Togaware

Dendogram

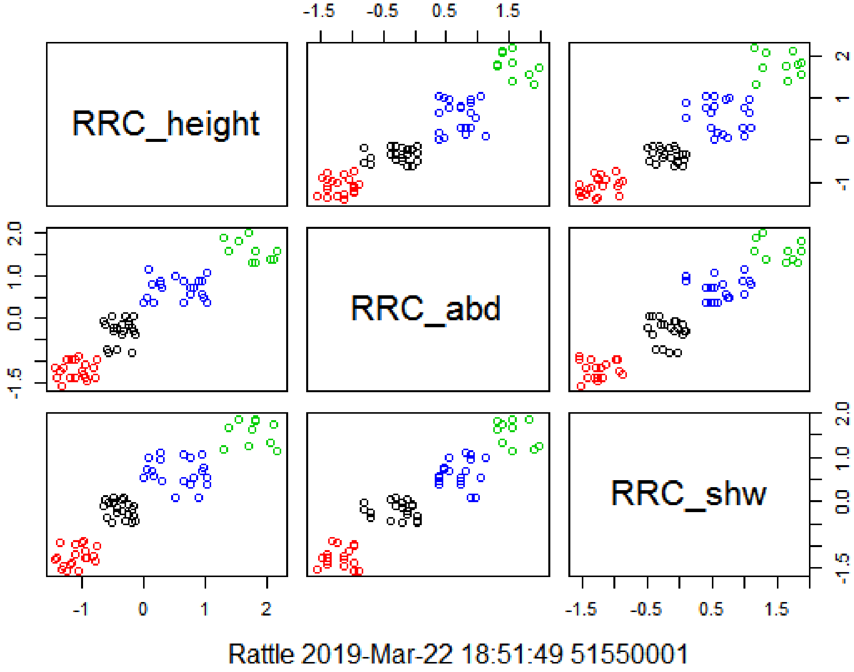


We cut the dendogram here. That means that we will have four clusters.

Data Plots

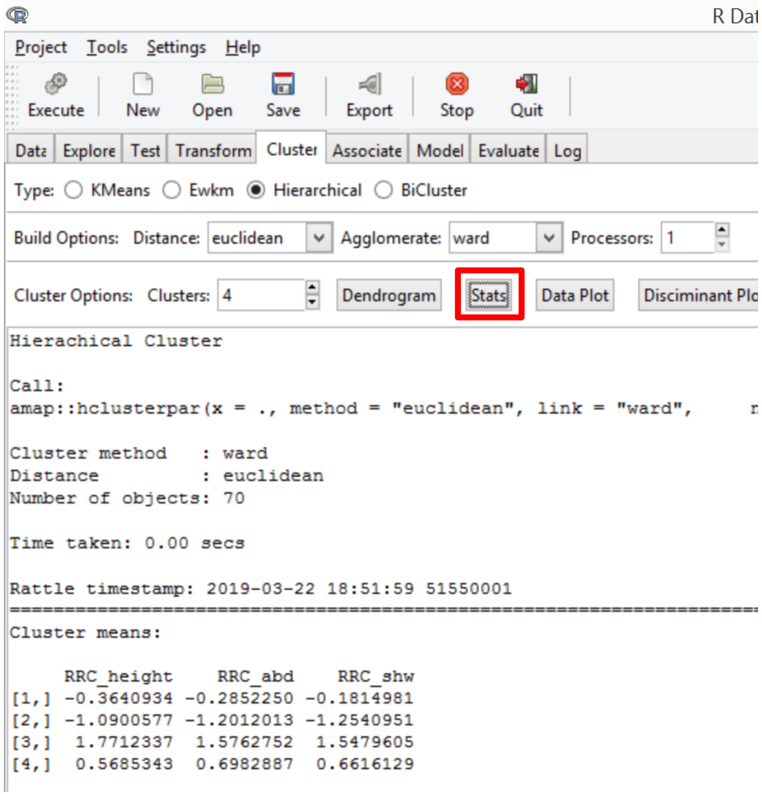


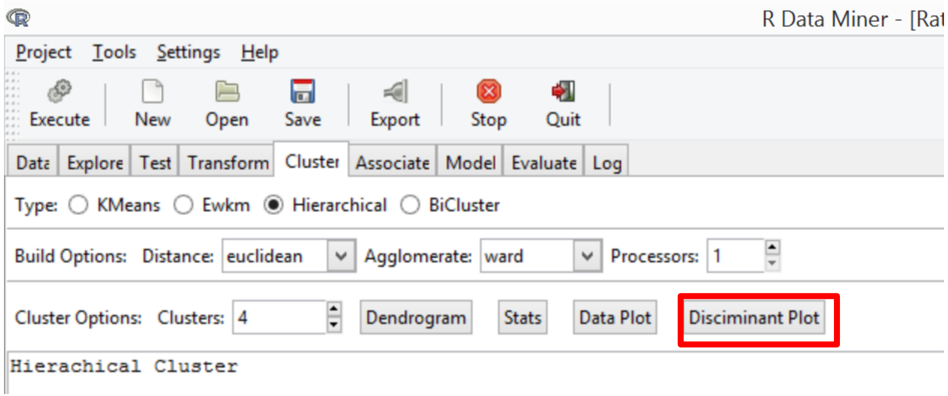
Based on Dendogram, we select K = 4

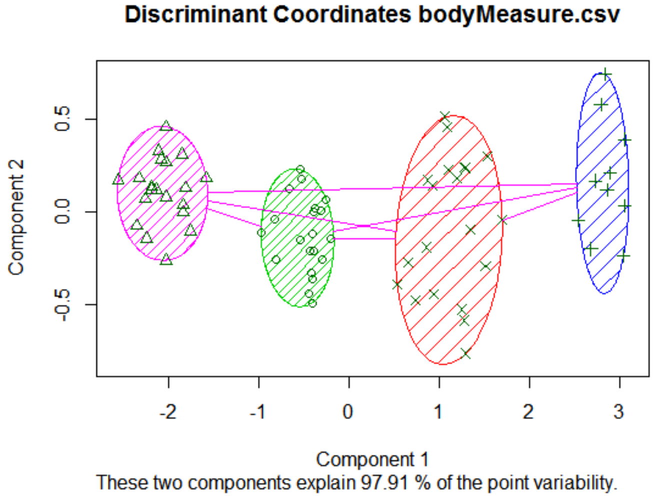


Source: Rattle GUI / Togaware

Discriminant Plot



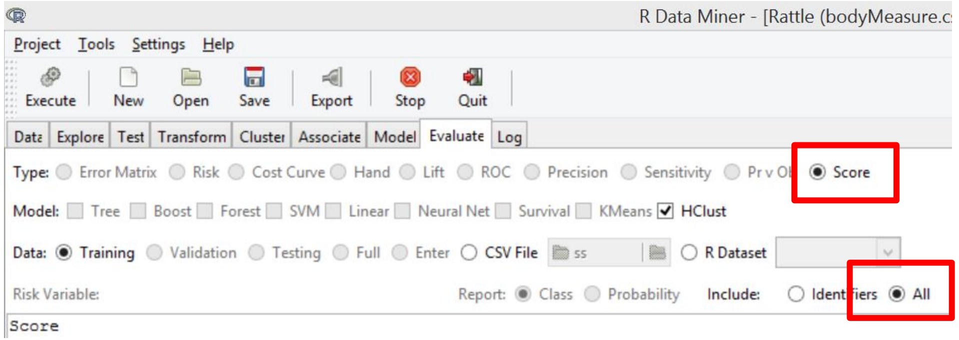


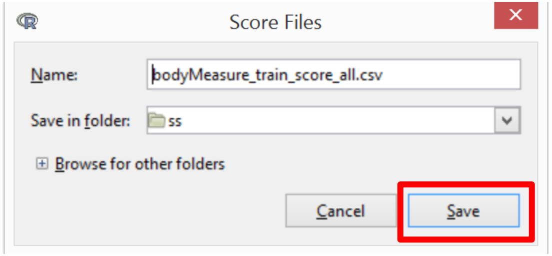


Source: Rattle GUI / Togaware

Cluster Membership

Output slightly different in R version 3.6.3





X	height	weight	chest	abd	shw	RRC_heigh	RRC_weig	RRC_chest	RRC_abd	RRC_shw	kmeans	hclust
1	168	61.5	98.5	85	47.5	-0.23317	-0.06162	0.006581	-0.04462	-0.06177	1	1
2	149.3	53	90	78	44.9	-1.33107	-0.60885	-0.86355	-1.22653	-0.9265	1	2
3	148.4	44.5	89.5	77	43.9	-1.38391	-1.15608	-0.91473	-1.39537	-1.25908	1	2
4	195.5	91.5	111	94.5	52.7	1.381391	1.869782	1.286181	1.559391	1.667692	2	3
5	159.1	52.5	84	79.5	44.7	-0.7557	-0.64104	-1.47776	-0.97326	-0.99301	1	2
6	172	74	106	87.5	49.3	0.001677	0.74313	0.774341	0.377486	0.536892	3	4

Source: Rattle GUI / Togaware

Output slightly different in R version 3.6.3

Exercise - Clustering

Perform Kmeans clustering on Universities data (use all rows and these variables: GradRate, SAT, TOP10, Accept, SFRatio and Expenses. Find the optimum number of clusters through iterative clustering. (Download from

http://users.stat.umn.edu/~kb/classes/8401/files/data/JWData5.txt)

For the chosen "K" above, run Kmeans clustering

Report the first two within cluster sum of squares (check your answer: 1.275 and 0.442 for K = 3)

You will see small difference due to random start of algorithm.

Exercise - Clustering

Perform Hierarchical clustering on Universities data (use all rows and these variables: GradRate, SAT, TOP10, Accept, SFRatio and Expenses). Find the optimum number of clusters based on the Dendrogram. Do you see a pattern in data identified by this method?

Universities data from http://users.stat.umn.edu/~kb/classes/8401/files/data/JWData5.txt

Data from Johnson & Wichtern (2002)

Summary

From data to clusters

K means and hierarchical clustering

Both problematic with high dimensional data

Summary

Hierarchical clustering cannot handle big data well but K Means clustering can due to run time being linear versus quadratic in number of points.

K Means clustering requires prior knowledge of number of clusters --in hierarchical clustering can stop by interpreting the dendrogram

Further Readings

PRACTICAL GUIDE TO CLUSTER ANALYSIS IN R

https://www.datanovia.com/en/product/practical-guide-to-cluster-analysis-in-r/

Clustering algorithms: A comparative approach (using R)

https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0210236

Nuclear Norm Clustering: a promising alternative method for clustering tasks

https://www.nature.com/articles/s41598-018-29246-4

Supervised clustering with Support Vector Machines

https://www.cs.cornell.edu/people/tj/publications/finley_joachims_05a.pdf

K-Means Complexity

E-Step: cluster **centers** → cluster **assignments**

$$d_{n,k}^{(t+1)} = \left(k == \underset{j=1\ldots K}{\arg\min} \left\{ \mathrm{D}\left(\mathbf{x}^{(n)}, \mathbf{m}_j^{(t)} \right) \right\} \right) \qquad O\left(NKD\right)$$

M-Step: cluster **assignments** → cluster **centers**

$$\mathbf{m}_k^{(t+1)} \neg \frac{\overset{N}{\underset{n=1}{\mathring{a}}} d_{n,k}^{(t)} \mathbf{x}^n}{\overset{N}{\underset{n=1}{\mathring{a}}} d_{n,k}^{(t)}} \qquad O\left(DN\right) = O\left(D \sum_{k=1}^{K} N_k \right)$$

References

Biological Classification. (n.d.). In *Wikipedia*. Retrieved May 22, 2019, from https://commons.wikimedia.org/wiki/File:Biological_classification_L_Pengo_vflip.svg

Johnson, R.A., & Wichern, D.W. (2002). *Applied Mulivariate Statistical Analysis* (5th ed.) Prentice Hall.

Rattle GUI / Togaware (https://rattle.togaware.com/)