

“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



1 H Hydrogen	<div><div>STABLE</div><div>half life more than one trillion years</div><div>half life in range of billion years</div><div>half life in range of million years</div><div>half life in range of thousands of years</div><div>half life in range of years</div><div>half life in range of days</div><div>half life in range of hours</div><div>half life in range of minutes</div><div>half life in range of seconds</div><div>half life in range of milliseconds</div><div>half life undetermined</div></div>																2 He Helium														
3 Li Lithium	4 Be Beryllium																	5 B Boron	6 C Carbon	7 N Nitrogen	8 O Oxygen	9 F Fluorine	10 Ne Neon								
11 Na Sodium	12 Mg Magnesium																	13 Al Aluminium	14 Si Silicon	15 P Phosphorous	16 S Sulfur	17 Cl Chlorine	18 Ar Argon								
19 K Potassium	20 Ca Calcium	21 Sc Scandium	22 Ti Titanium	23 V Vanadium	24 Cr Chromium	25 Mn Manganese	26 Fe Iron	27 Co Cobalt	28 Ni Nickel	29 Cu Copper	30 Zn Zinc	31 Ga Gallium	32 Ge Germanium	33 As Arsenic	34 Se Selenium	35 Br Bromine	36 Kr Krypton														
37 Rb Rubidium	38 Sr Strontium	39 Y Yttrium	40 Zr Zirconium	41 Nb Niobium	42 Mo Molybdenum	43 Tc Technetium	44 Ru Rutenium	45 Rh Rhodium	46 Pd Palladium	47 Ag Silver	48 Cd Cadmium	49 In Indium	50 Sn Tin	51 Sb Antimony	52 Te Tellurium	53 I Iodine	54 Xe Xenon														
55 Cs Caesium	56 Ba Barium	57 La Lanthanum	72 Hf Hafnium	73 Ta Tantalum	74 W Tungsten	75 Re Rhenium	76 Os Osmium	77 Ir Iridium	78 Pt Platinum	79 Au Gold	80 Hg Mercury	81 Tl Thallium	82 Pb Lead	83 Bi Bismuth	84 Po Polonium	85 At Astatine	86 Rn Radon														
87 Fr Francium	88 Ra Radium	89 Ac Actinium	104 Rf Rutherfordium	105 Db Dubnium	106 Sg Seaborgium	107 Bh Bohrium	108 Hs Hassium	109 Mt Meitnerium	110 Ds Darmstadtium	111 Rg Roentgenium	112 Uub Ununbium	113 Uut Ununtrium	114 Uuq Ununquadium	115 Uup Ununpentium	116 Uuh Ununhexium	117 Uus Ununseptium	118 Uuo Ununoctium														
																		58 Ce Cerium	59 Pr Praseodymium	60 Nd Neodymium	61 Pm Promethium	62 Sm Samarium	63 Eu Europium	64 Gd Gadolinium	65 Tb Terbium	66 Dy Dysprosium	67 Ho Holmium	68 Er Erbium	69 Tm Thulium	70 Yb Ytterbium	71 Lu Lutetium
																		90 Th Thorium	91 Pa Protactinium	92 U Uranium	93 Np Neptunium	94 Pu Plutonium	95 Am Americium	96 Cm Curium	97 Bk Berkelium	98 Cf Californium	99 Es Einsteinium	100 Fm Fermium	101 Md Mendelevium	102 No Nobelium	103 Lr Lawrencium

Hierarchy of Organization



<http://vlib.org/>

The WWW Virtual Library

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

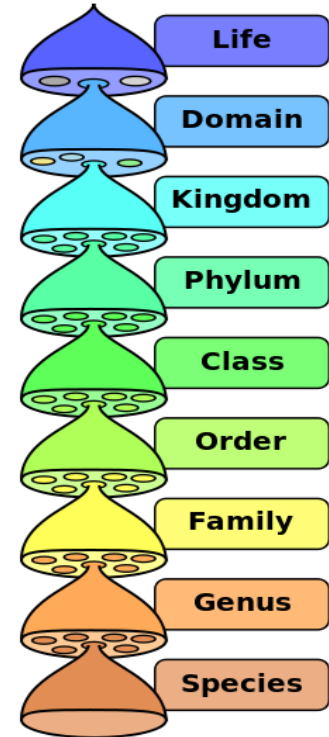
Anthropology, Archaeology, Population and Development Studies, ...

Society

Peoples, Religion, Gender Studies, ...

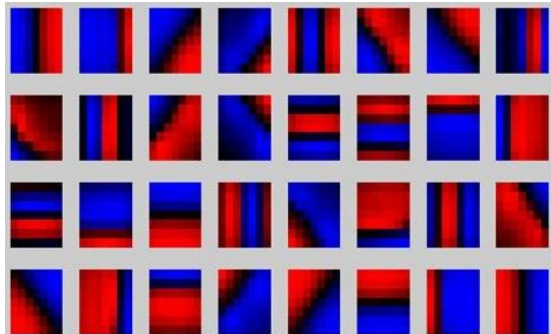
Quick se

Biological Classification



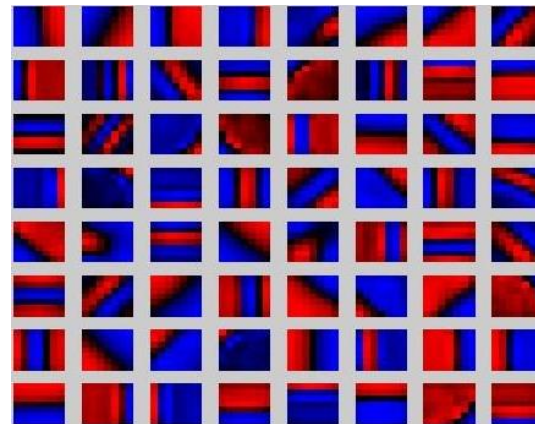
http://en.wikipedia.org/wiki/Biological_classification

Cluster to Create Features

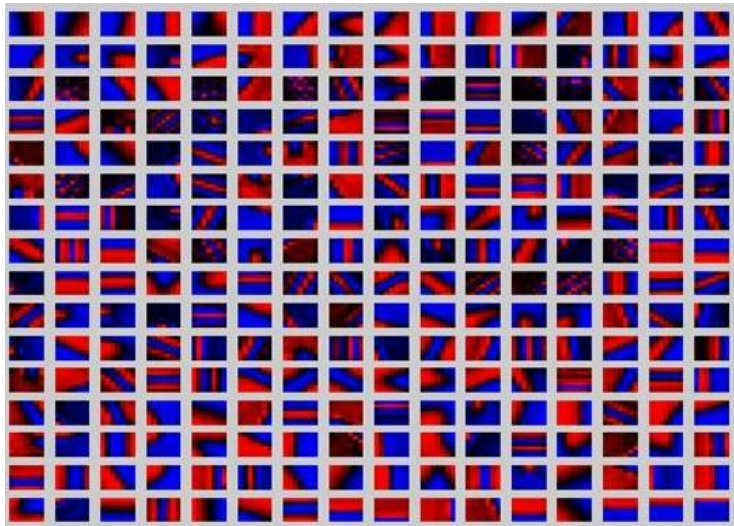


32 clusters

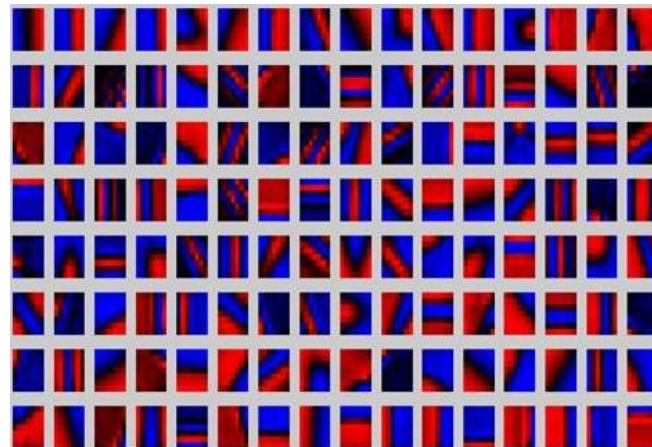
64 clusters



256 clusters



128 clusters



Customer Clustering by Purchase

I

DRAMA/DR	1.000
ACTION/AC	0.335
COMEDY/CO	0.306
SCIENCE FICTION/SF	0.131
SCIENCE FICTION/SF	1.000
ACTION/AC	0.285
TELEVISION	0.261
DRAMA/DR	0.233
COMEDY/CO	0.209
ANIME	0.181
ROCK/POP	0.105
FAMILY/FA	0.10

PLAYSTATION 2 SOFTWA
PS2 GREATEST HITS
PLAYSTATION 2 HARDWR
PLAYSTATION 2 MEMORY
PLAYSTATION 2 MISCEL
PLAYSTATION 2 CONTRO

XBOX SOFTWARE
XBOX HARDWARE
XBOX CONTROLLERS
XBOX MISC.
XBOX PLATINUM HITS

GAMECUBE SOFTWARE
GAMECUBE HARDWARE
GAMECUBE CONTROLLERS
GAMECUBE MEMORY
GAMECUBE MISC.

MEMORY
INTERNAL HARD DRIVES
GRAPHICS
PC COMPONENTS

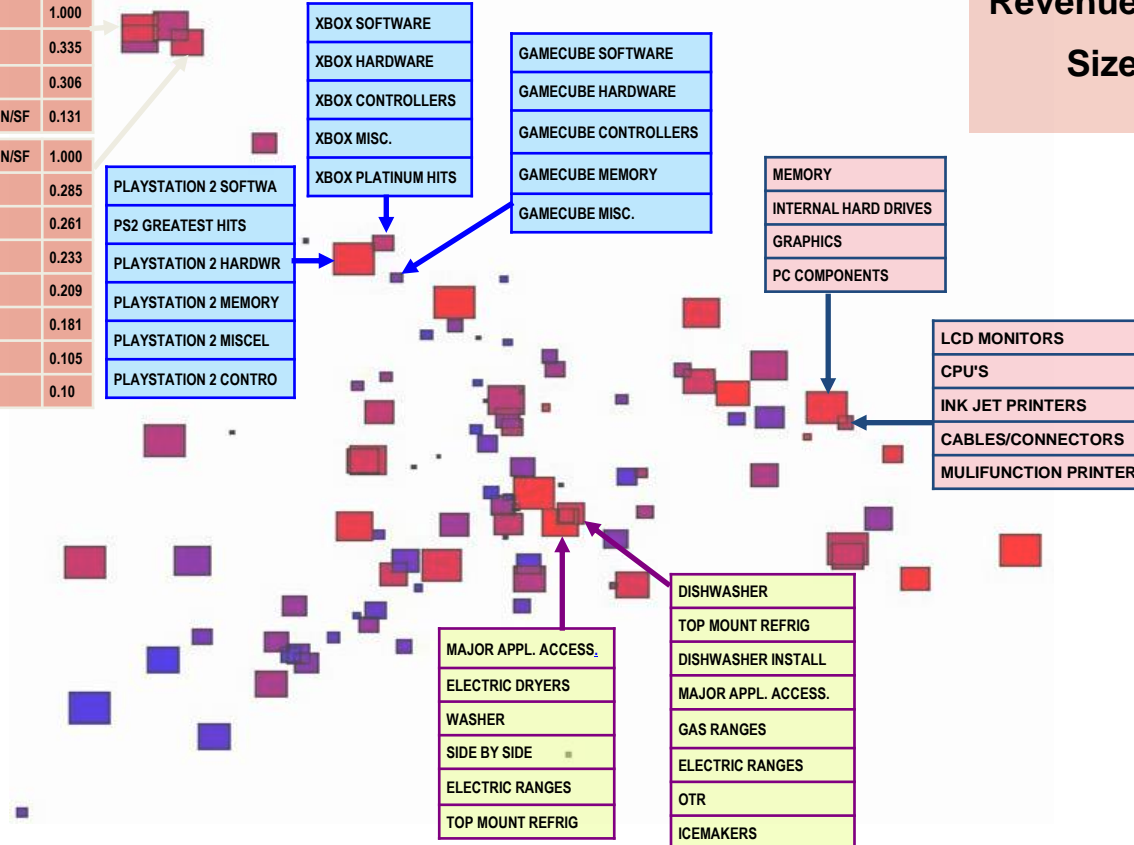
LCD MONITORS
CPU'S
INK JET PRINTERS
CABLES/CONNECTORS
MULIFUNCTION PRINTER

DISHWASHER
TOP MOUNT REFRIG
DISHWASHER INSTALL
MAJOR APPL. ACCESS.
GAS RANGES
ELECTRIC RANGES
OTR
ICEMAKERS

MAJOR APPL. ACCESS.
ELECTRIC DRYERS
WASHER
SIDE BY SIDE
ELECTRIC RANGES
TOP MOUNT REFRIG

Revenue: High Low

Size: High Low



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

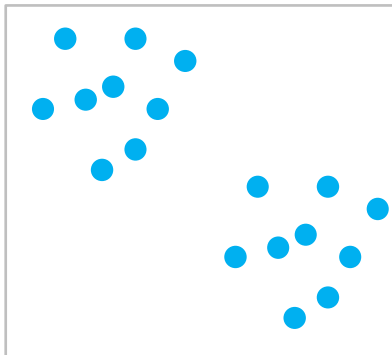
Business rules can be uncovered

How Does One Cluster?

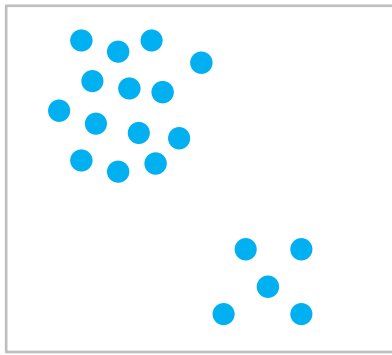


It depends!

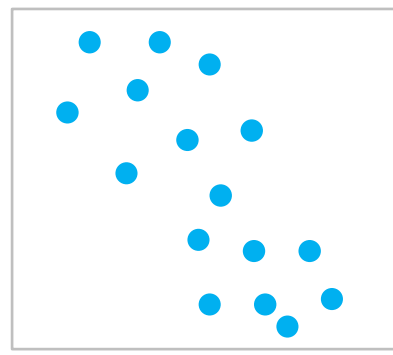
Why is Clustering Non-trivial?



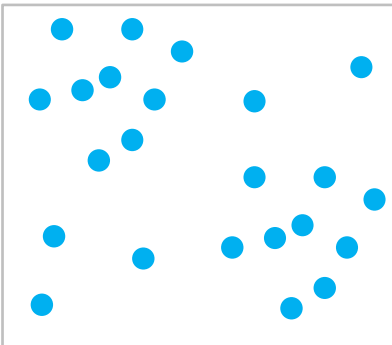
Nice Clean Clusters



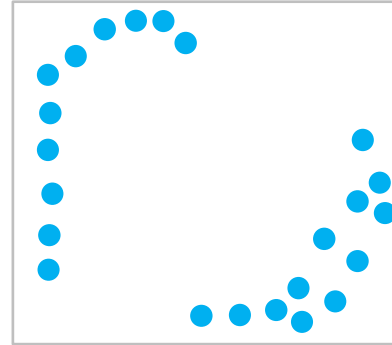
Density Variation



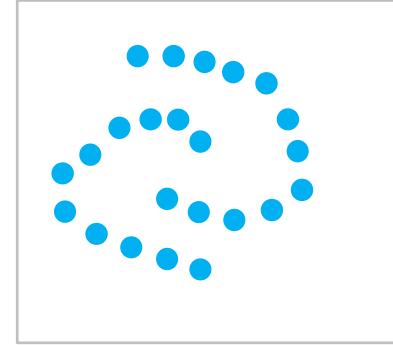
Not Clean Separation



Background Noise

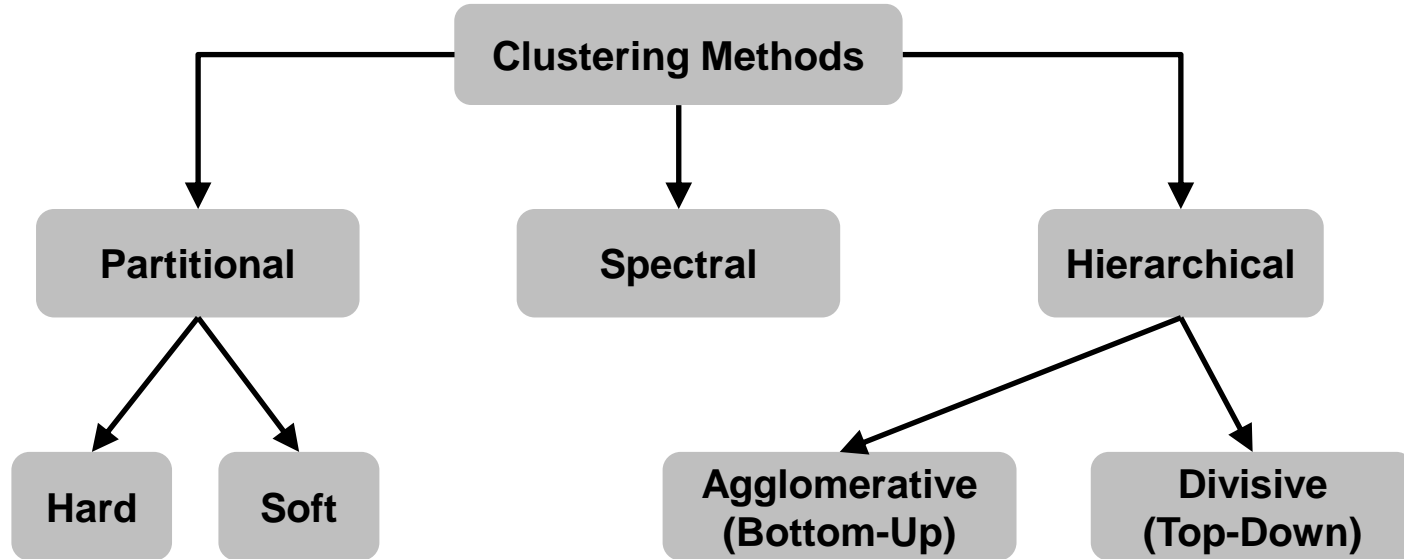


Not Spherical data



Not Easily Separable

Clustering Methods



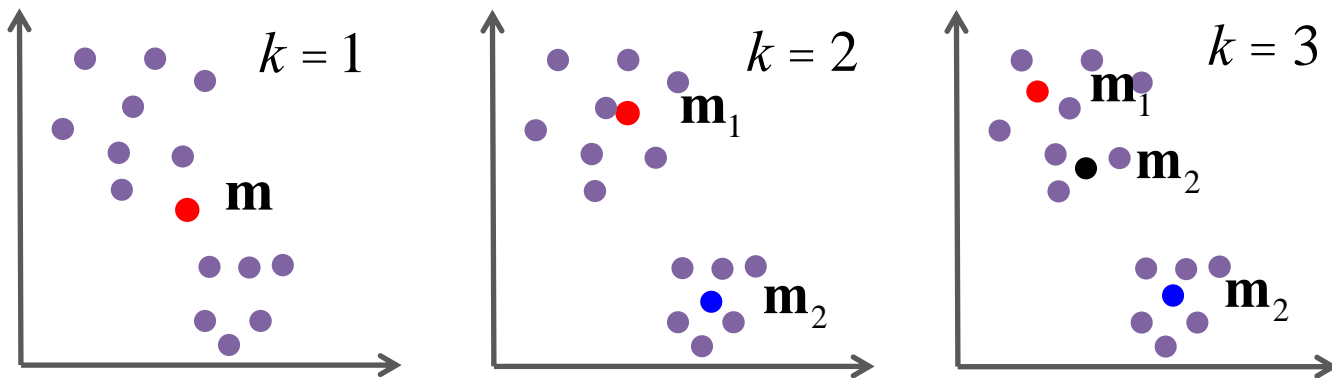
Key to Clustering is how we define “**Distance**” or “**Similarity**” between two data points

Partitional Clustering



K-Means Clustering

K-Means – Objective Function



What are the **Parameters**?

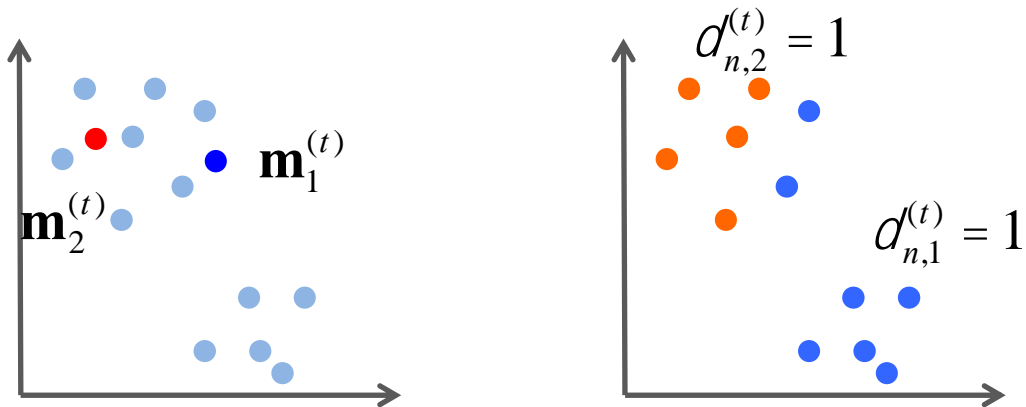
What is the **Objective Function**?

What is the **Model Complexity**?

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

K-Means: 1. Expectation

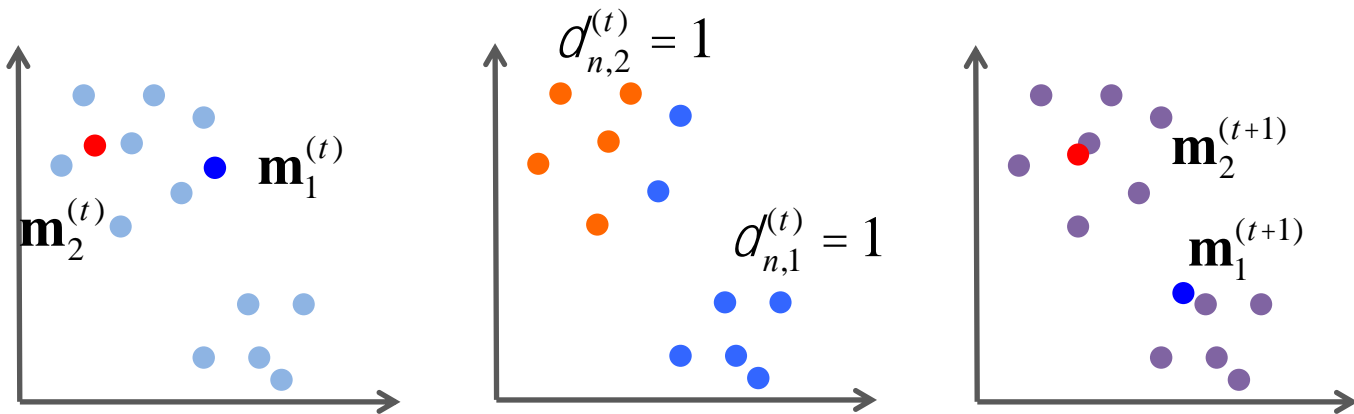
cluster *centers* → cluster *assignments*



$$\mathcal{d}_{n,k}^{(t+1)} = \left(k == \arg \min_{j=1 \dots K} \left\{ D\left(\mathbf{x}^{(n)}, \mathbf{m}_j^{(t)} \right) \right\} \right)$$

K-Means: 2. Maximization

cluster *assignments* → cluster *centers*



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

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)

bodyMeasure.csv

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

Rescale the Data for Distance Based Algorithms (KNN)

I

R Data Miner - [Rattle (bodyf)]

Project Tools Settings Help

Execute New Open Save Export Stop Quit

Data Explore Test Transform Cluster Associate Model Evaluate Log

Source: ☒ File ☐ ARFF ☐ ODBC ☐ R Dataset ☐ RData File ☐ Library ☐ Corpus ☐ Script

Filename: Separator: Decimal: ☒ Header

☒ Partition Seed: View Edit

☒ Input ☐ Ignore Weight Calculator: Target Data Type: ☒ Auto ☐ Categorical ☐ Numeric ☐ Survival

No. Variable	Data Type	Input	Target	Risk	Ident	Ignore	Weight	Comment
1 X	Numeric	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	Unique: 70
2 height	Numeric	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Unique: 68
3 weight	Numeric	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	Unique: 52
4 chest	Numeric	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	Unique: 41
5 abd	Numeric	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Unique: 32
6 shw	Numeric	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Unique: 52

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

Source: Rattle GUI / Togaware

Execute New Open Save Export Stop Quit

Data Explore Test Transform Cluster Associate Model Evaluate Log

Type: ☒ Rescale ☐ Impute ☐ Recode ☐ Cleanup

Normalize: ☒ Recenter ☐ Scale [0-1] ☐ -Median/MAD ☐ Natural Log ☐ Log 10 ☐ Matrix

Order: ☐ Rank ☐ Interval Groups:

No. Variable Data Type and Number Missing

1 X	Numeric	[1 to 70; unique=70; mean=35; median=35].
2 height	Numeric	[147.70 to 209.10; unique=68; mean=171.97; median=168.45].
3 weight	Numeric	[35.50 to 95.00; unique=52; mean=62.46; median=60.00].
4 chest	Numeric	[81.00 to 116.50; unique=41; mean=98.44; median=100.50].
5 abd	Numeric	[76.00 to 97.00; unique=32; mean=85.26; median=84.75].
6 shw	Numeric	[43.00 to 53.30; unique=52; mean=47.69; median=47.60].

Data Explore Test Transform Cluster Associate Model Evaluate Log

Type: ☒ Rescale ☐ Impute ☐ Recode ☐ Cleanup

Normalize: ☒ Recenter ☐ Scale [0-1] ☐ -Median/MAD ☐ Natural Log ☐ Log 10 ☐ Matrix

Order: ☐ Rank ☐ Interval Groups:

No. Variable Data Type and Number Missing

1 X	Numeric	[1 to 70; unique=70; mean=35; median=35].
2 height	Numeric	[147.70 to 209.10; unique=68; mean=171.97; median=168.45; ignored].
3 weight	Numeric	[35.50 to 95.00; unique=52; mean=62.46; median=60.00; ignored].
4 chest	Numeric	[81.00 to 116.50; unique=41; mean=98.44; median=100.50; ignored].
5 abd	Numeric	[76.00 to 97.00; unique=32; mean=85.26; median=84.75; ignored].
6 shw	Numeric	[43.00 to 53.30; unique=52; mean=47.69; median=47.60; ignored].
7 RRC_height	Numeric	[-1.43 to 2.18; unique=68; mean=0.00; median=-0.21].
8 RRC_weight	Numeric	[-1.74 to 2.10; unique=52; mean=0.00; median=-0.16].
9 RRC_chest	Numeric	[-1.78 to 1.85; unique=41; mean=0.00; median=0.21].
10 RRC_abd	Numeric	[-1.56 to 1.98; unique=32; mean=-0.00; median=-0.09].
11 RRC_shw	Numeric	[-1.56 to 1.87; unique=52; mean=0.00; median=-0.03].

Select the Input Variables and Select the Best K

R Data Miner - [Rattle (body)

Project Tools Settings Help

Execute New Open Save Export Stop Quit

Data Explore Test Transform Cluster Associate Model Evaluate Log

Source: ☒ File ☐ ARFF ☐ ODBC ☐ R Dataset ☐ RData File ☐ Library ☐ Corpus ☐ Script

Filename: bodyMeasure.csv Separator: , Decimal: . ☒ Header

☐ Partition 70/15/15 Seed: 42 View Edit

☒ Input ☐ Ignore Weight Calculator: Target Data Type: ☒ Auto ☐ Categorical ☐ Numeric ☐ Survival

No.	Variable	Data Type	Input	Target	Risk	Ident	Ignore	Weight	Comment
1	X	Numeric	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	Unique: 70
2	height	Numeric	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	Unique: 68
3	weight	Numeric	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	Unique: 52
4	chest	Numeric	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	Unique: 41
5	abd	Numeric	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	Unique: 32
6	shw	Numeric	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	Unique: 52
7	RRC_height	Numeric	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Unique: 68
8	RRC_weight	Numeric	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	Unique: 52
9	RRC_chest	Numeric	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	Unique: 41
10	RRC_abd	Numeric	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Unique: 32
11	RRC_shw	Numeric	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Unique: 52

Max. K – user given

R Data M

Project Tools Settings Help

Execute New Open Save Export Stop Quit

Data Explore Test Transform Cluster Associate Model Evaluate Log

Type: ☒ KMeans ☐ Pwkm ☐ Hierarchical ☐ BiCluster

Clusters: 8 Seed: 42 Runs: 1 ☒ Re-Scale

☐ Use HClust Centers ☒ Iterate Clusters Stats Plots: Data Discriminant Weights

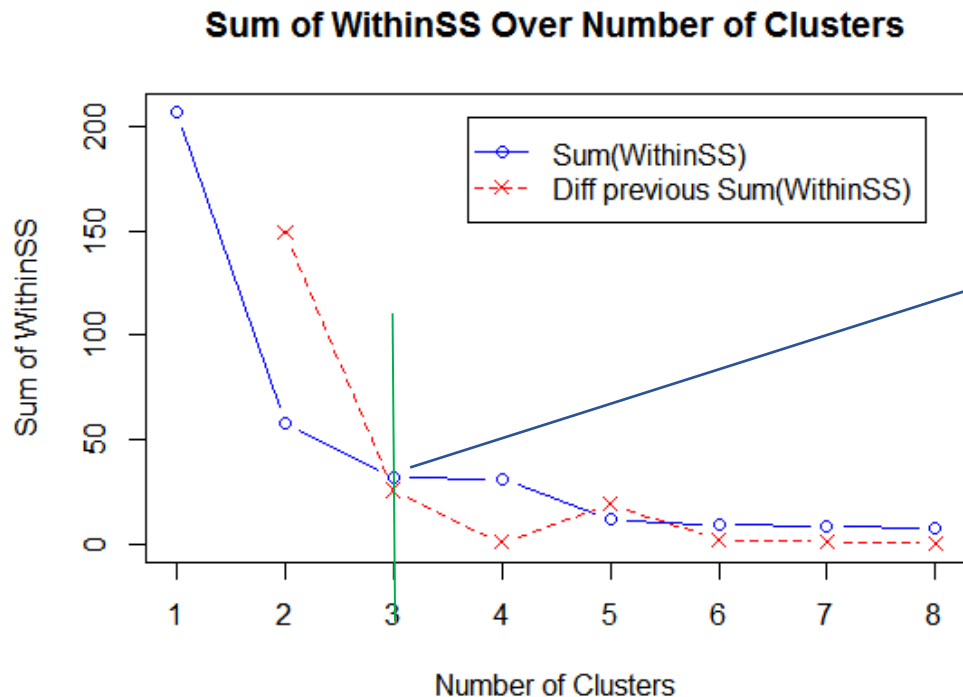
We have iterated over cluster sizes from 2 to 8 clusters.

The plot displays the 'sum(withinss)' for each clustering and the change in this value from the previous clustering.

Time taken: 0.01 secs

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

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!

Running K Means for K=3

3 clusters

R Data

Project Tools Settings Help

Execute New Open Save Export Stop Quit

Data Explore Test Transform Cluster Associate Model Evaluate Log

Type: ☒ KMeans ☐ Ewkm ☐ Hierarchical ☐ BiCluster

Clusters: 3 Seed: 42 Runs: 1 ☒ Re-Scale

☐ Use HClust Centers ☐ Iterate Clusters Stats Plots: Data Discriminant Weights

Cluster sizes:

```
[1] "20 27 23"
```

Data means:

RRC_height	RRC_abd	RRC_shw
0.3953001	0.4411565	0.4549237

Cluster centers:

	RRC_height	RRC_abd	RRC_shw
1	0.09291531	0.1023810	0.08883495
2	0.69163952	0.7389771	0.74541532
3	0.31036680	0.3861284	0.43224989

Within cluster sum of squares:

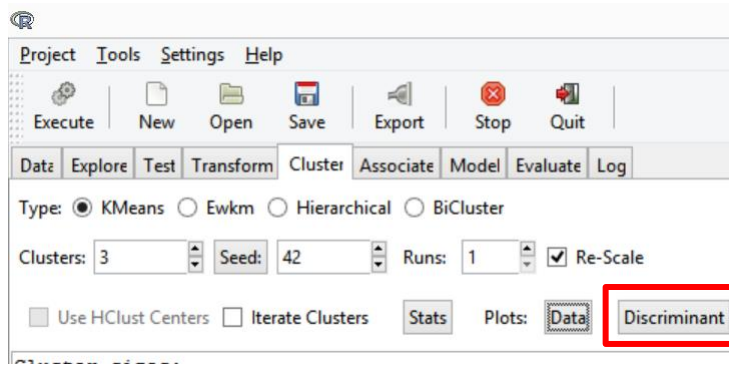
```
[1] 0.2077483 1.8734773 0.5004008
```

Time taken: 0.00 secs

No iteration

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

Output slightly different in R version 3.6.3



Kmeans - Discriminant Plot

Output slightly different in R version 3.6.3



Project Tools Settings Help

Execute New Open Save Export Stop Quit

Data Explore Test Transform Cluster Associate Model Evaluate Log

Type: ☒ KMeans ☐ Ewkm ☐ Hierarchical ☐ BiCluster

Clusters: 3 Seed: 42 Runs: 1 ☒ Re-Scale

☐ Use HClust Centers ☐ Iterate Clusters Stats Plots: Data **Discriminant**

Cluster sizes:

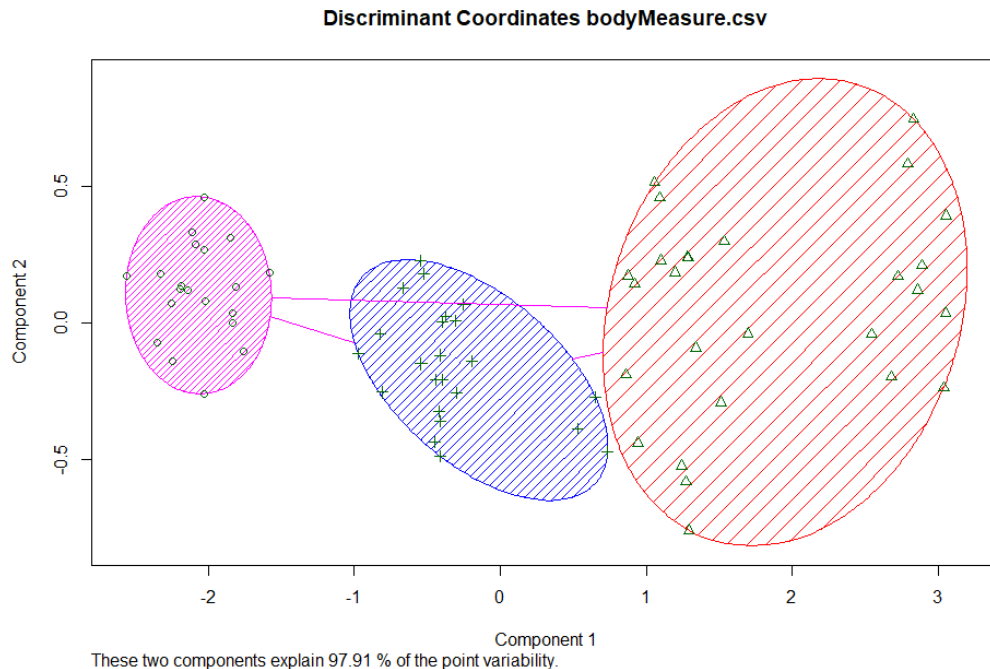
```
[1] "20 27 23"
```

Data means:

RRC_height	RRC_abd	RRC_shw
0.3953001	0.4411565	0.4549237

Cluster centers:

	RRC_height	RRC_abd	RRC_shw
1	0.09291531	0.1023810	0.08883495
2	0.69163952	0.7389771	0.74541532
3	0.31036680	0.3861284	0.43224989



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.

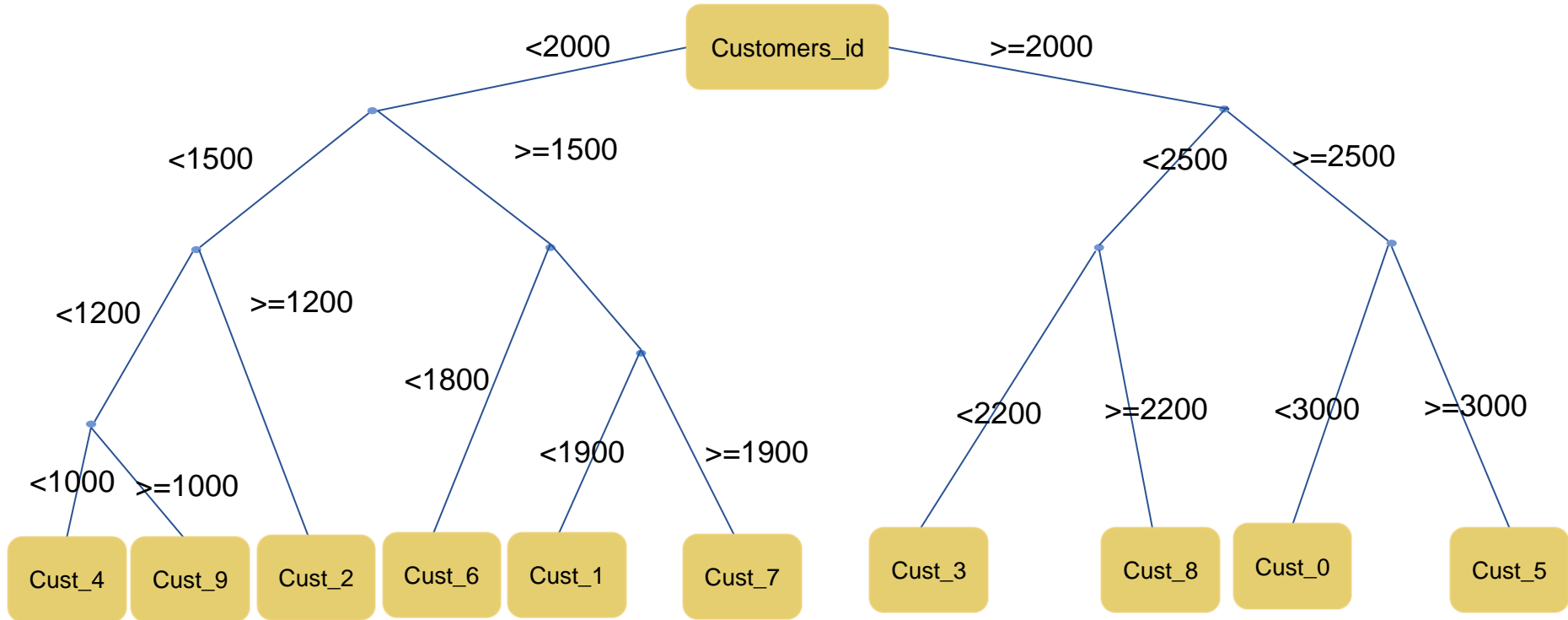
Clustering Example



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

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

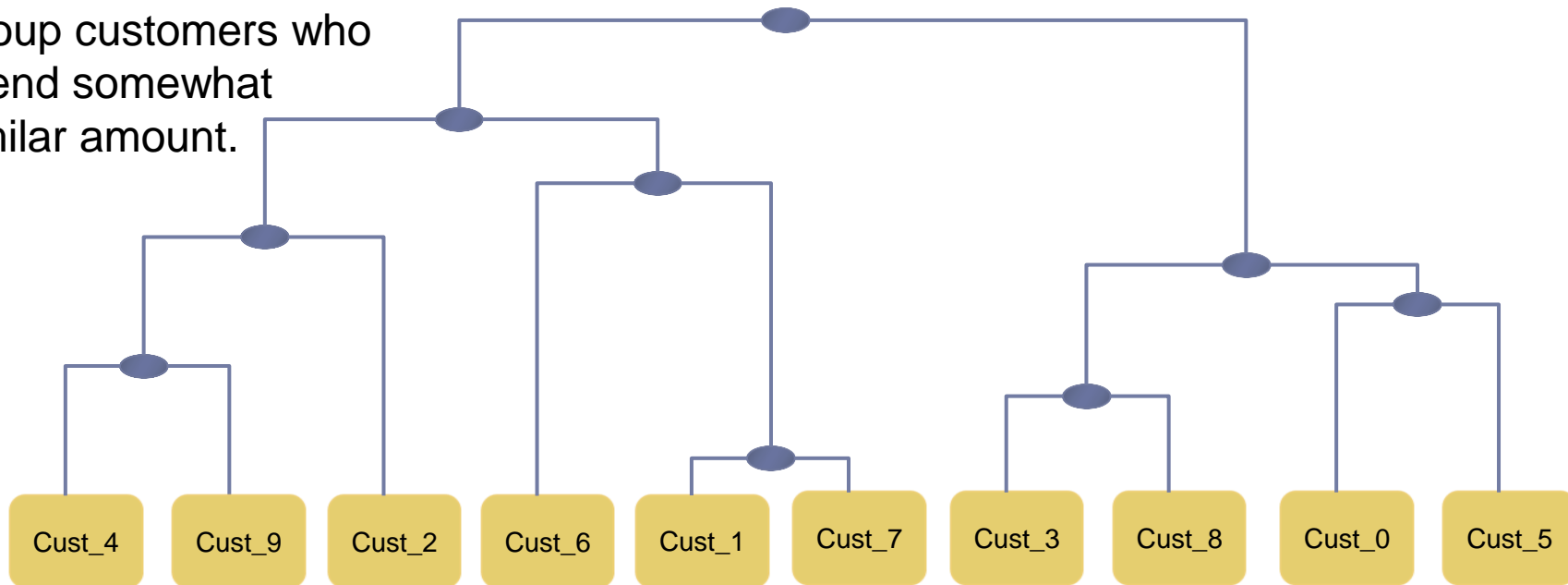
I



Bottom-Up (Agglomerative)

I

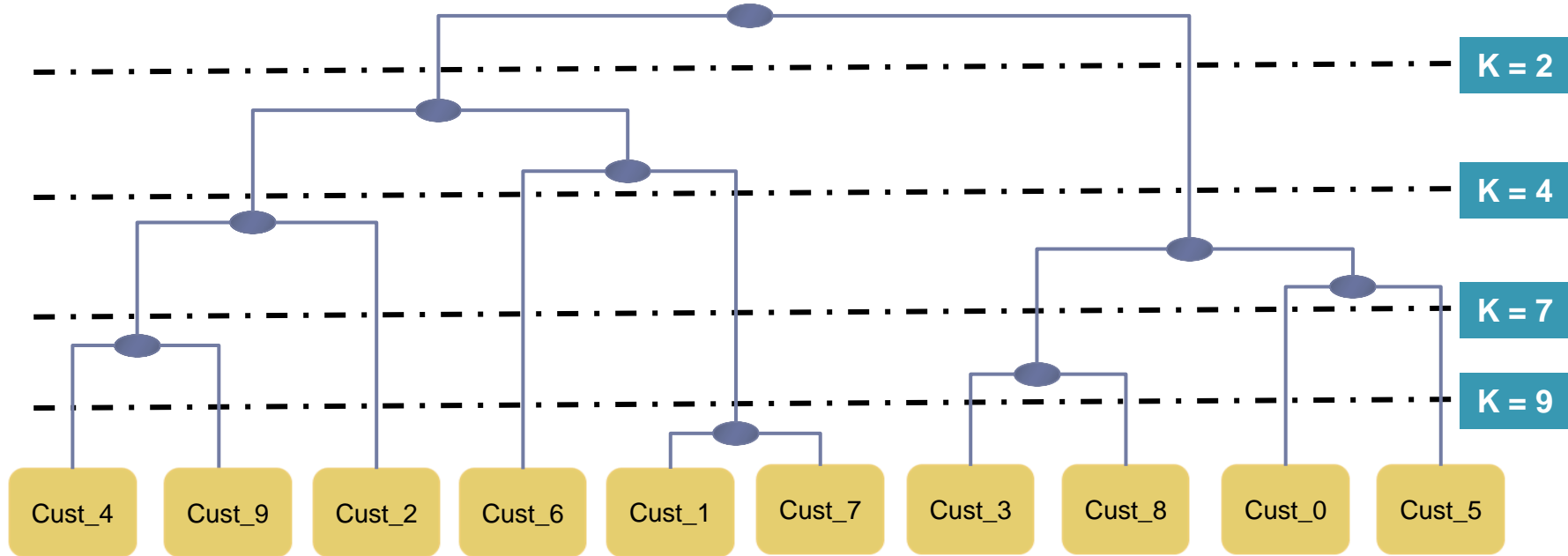
Group customers who
spend somewhat
similar amount.



SIMILARITY between two **DATA POINTS**? $\text{Sim}(0, 5)$

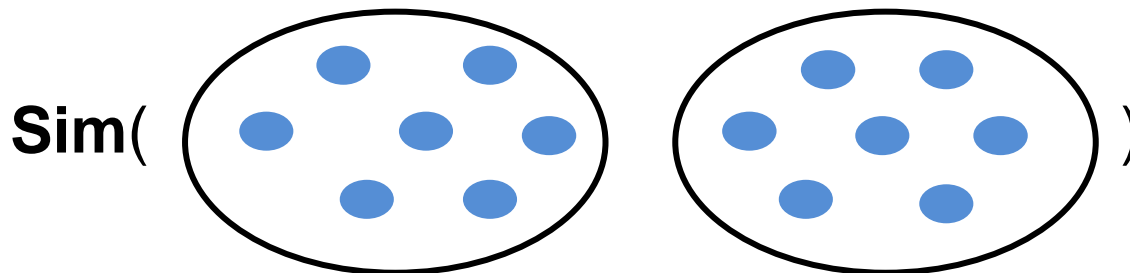
SIMILARITY between **CLUSTERS**? $\text{Sim}(\{3, 8\}, \{0, 5\})$

How Many Clusters? Depends!

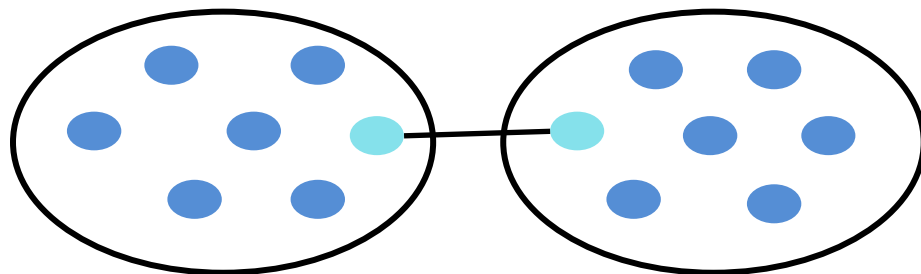


Number of clusters depends on where we CUT the dendrogram.

Similarity Between Clusters

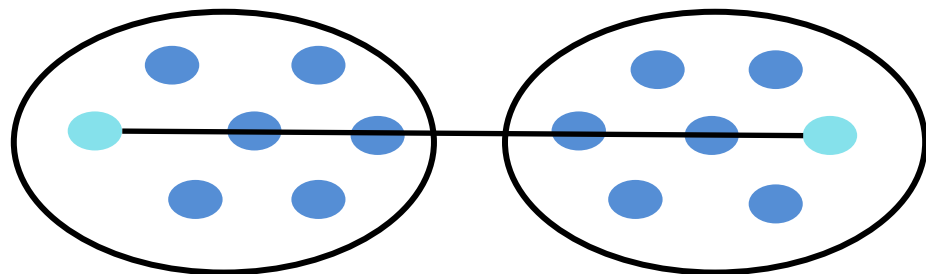


Single Linkage
Nearest Neighbor



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

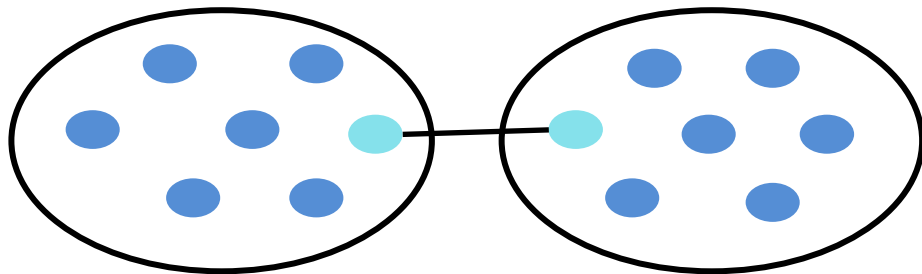
Complete Linkage
Farthest Neighbor



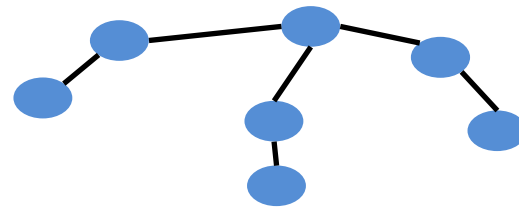
$$D(\mathbf{C}_a, \mathbf{C}_b) = \max_{\mathbf{x} \in \mathbf{C}_a, \mathbf{y} \in \mathbf{C}_b} \{D(\mathbf{x}, \mathbf{y})\}$$

Single Link Examples

Single Linkage
Nearest Neighbor



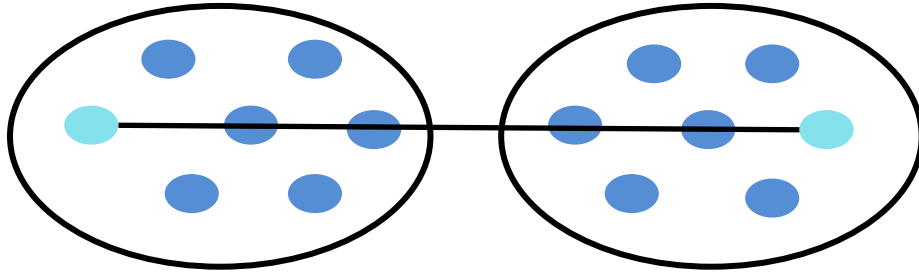
$$D(\mathbf{C}_a, \mathbf{C}_b) = \min_{\mathbf{x} \in \mathbf{C}_a, \mathbf{y} \in \mathbf{C}_b} \{D(\mathbf{x}, \mathbf{y})\}$$



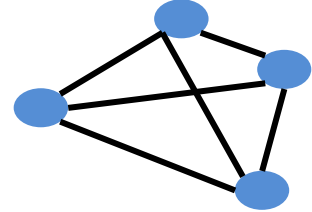
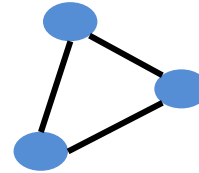
- **Elongated** Clusters
- Minimum Spanning Trees

Complete Linkage Example

Complete Linkage
Farthest Neighbor

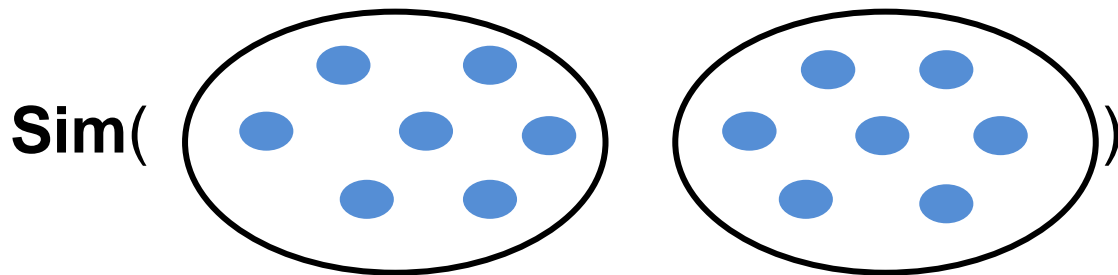


$$D(\mathbf{C}_a, \mathbf{C}_b) = \max_{\mathbf{x} \in \mathbf{C}_a, \mathbf{y} \in \mathbf{C}_b} \{D(\mathbf{x}, \mathbf{y})\}$$

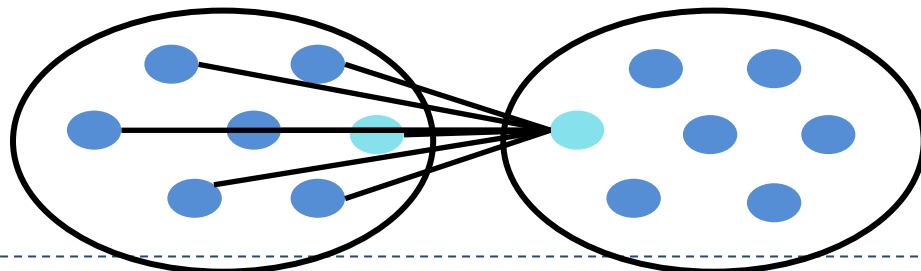


- **Compact Clusters**

Similarity between Clusters

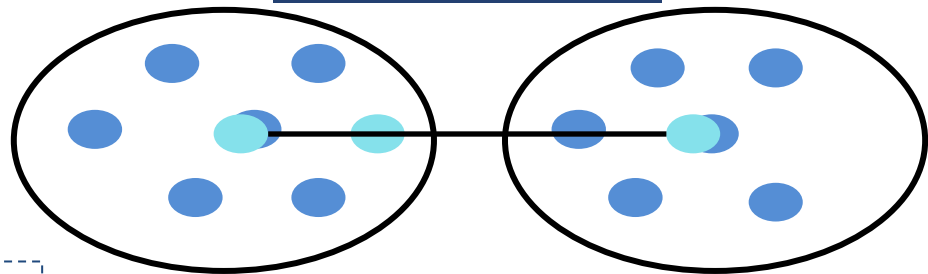


**Average Linkage
All Neighbors**



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

**Mean Linkage
Mean Neighbor**



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

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...



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

Depends on the domain

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

Can be static (constant) or dynamic (learnt)

Data: bodyMeasure

R Data Miner - [Rattle (body)

Project Tools Settings Help

Execute New Open Save Export Stop Quit

Date Explore Test Transform Cluster Associate Model Evaluate Log

Source: ☒ File ☐ ARFF ☐ ODBC ☐ R Dataset ☐ RData File ☐ Library ☐ Corpus ☐ Script

Filename: bodyMeasure.csv Separator: , Decimal: . ☒ Header

☐ Partition 70/15/15 Seed: 42 View Edit

☒ Input ☒ Ignore Weight Calculator: Target Data Type: ☒ Auto ☐ Categorical ☐ Numeric ☐ Survival

No.	Variable	Data Type	Input	Target	Risk	Ident	Ignore	Weight	Comment
1	X	Numeric	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	Unique: 70
2	height	Numeric	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	Unique: 68
3	weight	Numeric	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	Unique: 52
4	chest	Numeric	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	Unique: 41
5	abd	Numeric	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	Unique: 32
6	shw	Numeric	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	Unique: 52
7	RRC_height	Numeric	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Unique: 68
8	RRC_weight	Numeric	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	Unique: 52
9	RRC_chest	Numeric	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	Unique: 41
10	RRC_abd	Numeric	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Unique: 32
11	RRC_shw	Numeric	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Unique: 52

Source: Rattle GUI / Togaware

R Data Miner - [Rattle (body)

Project Tools Settings Help

Execute New Open Save Export Stop Quit

Date Explore Test Transform Cluster Associate Model Evaluate Log

Type: ☐ KMeans ☐ Ewkm ☒ Hierarchical ☐ BiCluster

Build Options: Distance: euclidean Agglomerate: ward Processors: 1

Cluster Options: Clusters: 10 Dendrogram Stats Data Plot Discriminant Plot

Hierarchical Clustering

R Data Miner - [Rattle (body)

Project Tools Settings Help

Execute New Open Save Export Stop Quit

Date Explore Test Transform Cluster Associate Model Evaluate Log

Type: ☐ KMeans ☐ Ewkm ☒ Hierarchical ☐ BiCluster

Build Options: Distance: euclidean Agglomerate: ward Processors: 1

Cluster Options: Clusters: 10 Dendrogram Stats Data Plot Discriminant Plot

Hierarchical Cluster

```
Call:
amap::hclusterpar(x = ., method = "euclidean", link = "ward", nbproc = 1)

Cluster method : ward
Distance       : euclidean
Number of objects: 70

Time taken: 0.00 secs
```

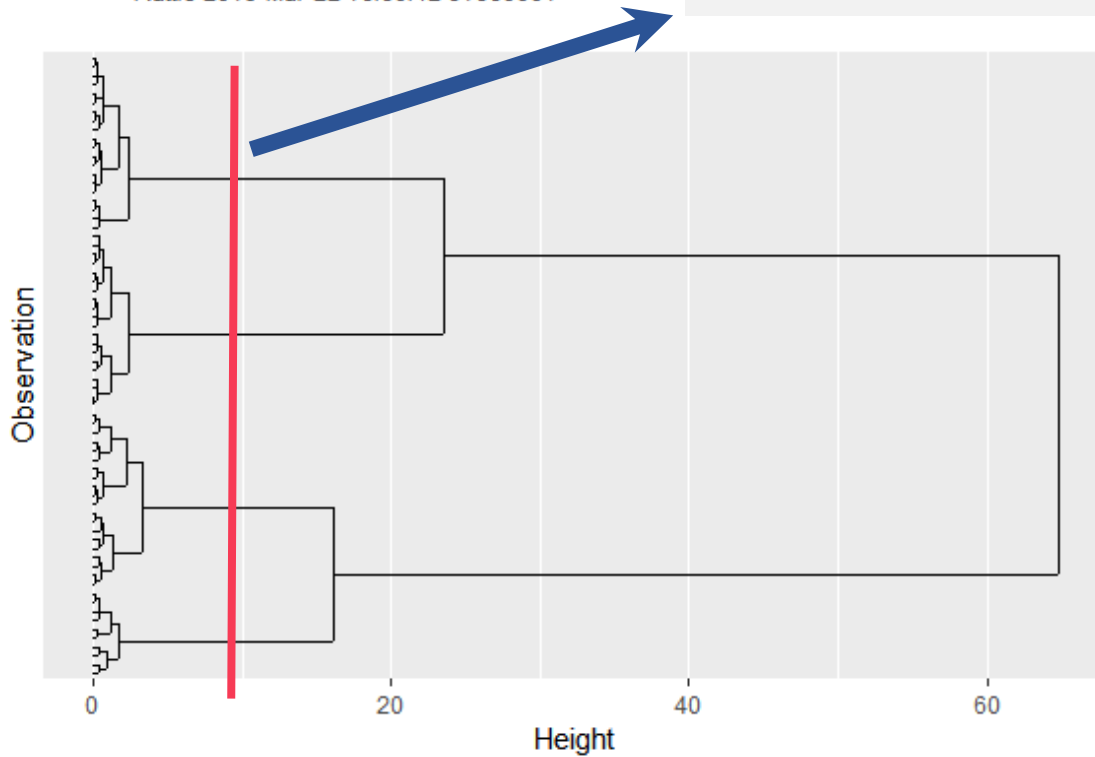
Dendrogram



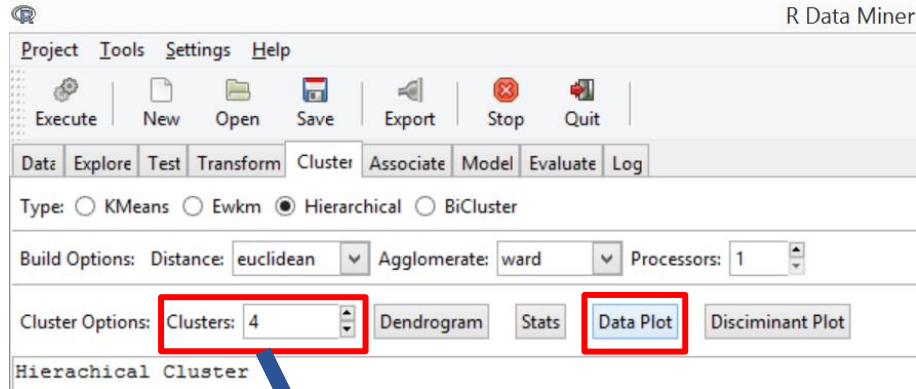
Cluster Dendrogram bodyMeasure.csv

Rattle 2019-Mar-22 18:50:42 51550001

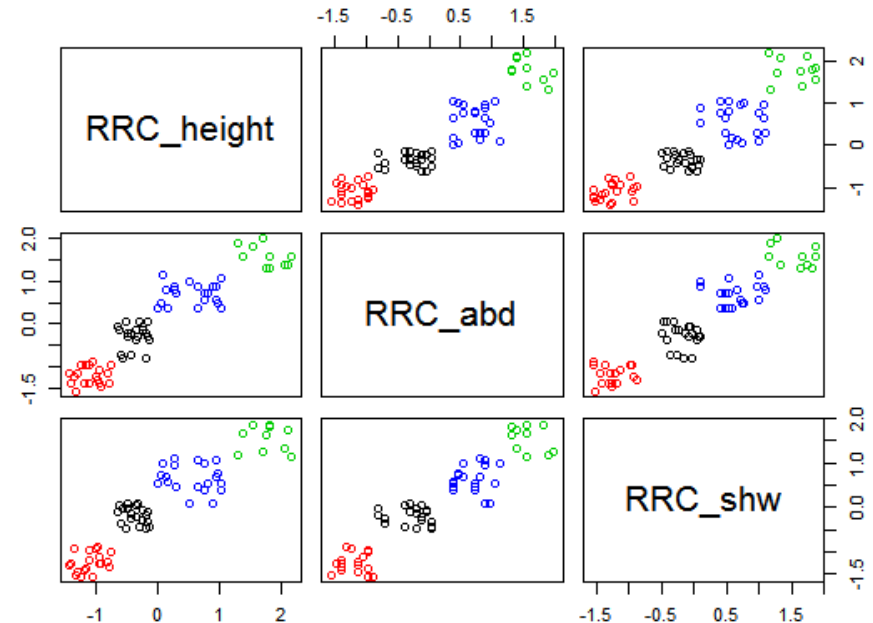
We cut the dendrogram here. That means that we will have four clusters.



Data Plots



Based on Dendrogram,
we select $K = 4$



Rattle 2019-Mar-22 18:51:49 51550001

Discriminant Plot



R Data Miner - [Rat]

Project Tools Settings Help

Execute New Open Save Export Stop Quit

Date Explore Test Transform Cluster Associate Model Evaluate Log

Type: ☐ KMeans ☐ Ewkm ☒ Hierarchical ☐ BiCluster

Build Options: Distance: euclidean Agglomerate: ward Processors: 1

Cluster Options: Clusters: 4 Dendrogram **Stats** Data Plot Discriminant Plot

Hierarchical Cluster

Call:

```
amap::hclclusterpar(x = ., method = "euclidean", link = "ward",
```

Cluster method : ward
Distance : euclidean
Number of objects: 70

Time taken: 0.00 secs

Rattle timestamp: 2019-03-22 18:51:59 51550001

Cluster means:

	RRC_height	RRC_abd	RRC_shw
[1,]	-0.3640934	-0.2852250	-0.1814981
[2,]	-1.0900577	-1.2012013	-1.2540951
[3,]	1.7712337	1.5762752	1.5479605
[4,]	0.5685343	0.6982887	0.6616129

R Data Miner - [Rat]

Project Tools Settings Help

Execute New Open Save Export Stop Quit

Date Explore Test Transform Cluster Associate Model Evaluate Log

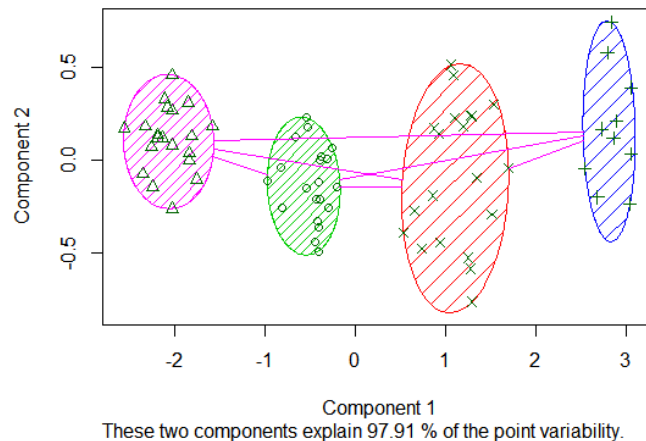
Type: ☐ KMeans ☐ Ewkm ☒ Hierarchical ☐ BiCluster

Build Options: Distance: euclidean Agglomerate: ward Processors: 1

Cluster Options: Clusters: 4 Dendrogram Stats Data Plot **Discriminant Plot**

Hierarchical Cluster

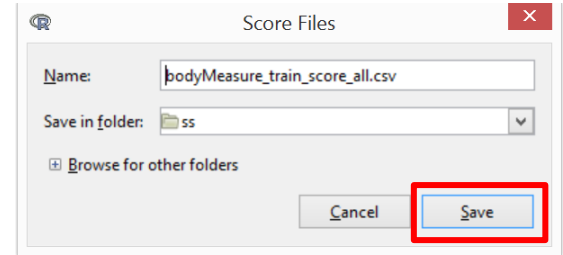
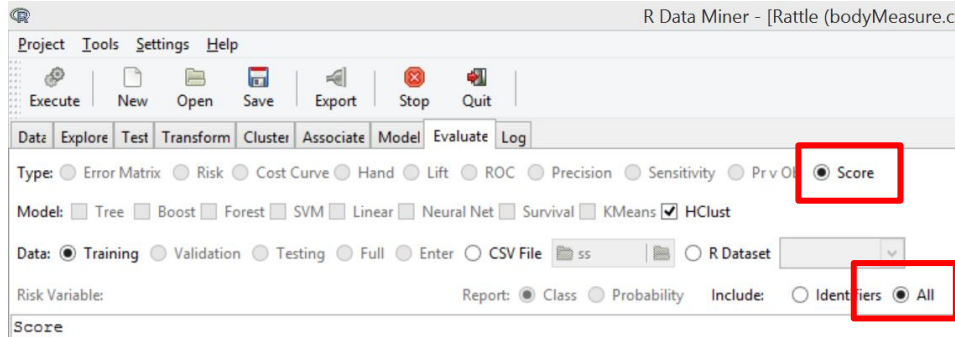
Discriminant Coordinates bodyMeasure.csv



Cluster Membership

Output slightly different in R version 3.6.3

I



X	height	weight	chest	abd	shw	RRC_height	RRC_weight	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

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/JWD_ata5.txt](http://users.stat.umn.edu/~kb/classes/8401/files/data/JWD_ata5.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>

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** \rightarrow cluster **assignments**

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

M-Step: cluster **assignments** \rightarrow cluster **centers**

$$\mathbf{m}_k^{(t+1)} \leftarrow \frac{\sum_{n=1}^N d_{n,k}^{(t)} \mathbf{x}^n}{\sum_{n=1}^N d_{n,k}^{(t)}} \quad O(DN) = O\left(D \sum_{k=1}^K N_k\right)$$

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 Multivariate Statistical Analysis* (5th ed.)
Prentice Hall.

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