BACS3013 Data Science

Chapter 3: Visualization and Descriptive Analytics

Content

- Visualization in data analytics
- Descriptive Analytics
  - Statistical inference
  - Association rules
  - Sequence rules
  - Segmentation

NEXT

DATA VISUALIZATION

Data Visualization

- Two basic types:

Exploration

- What the data is telling you?

Explanation

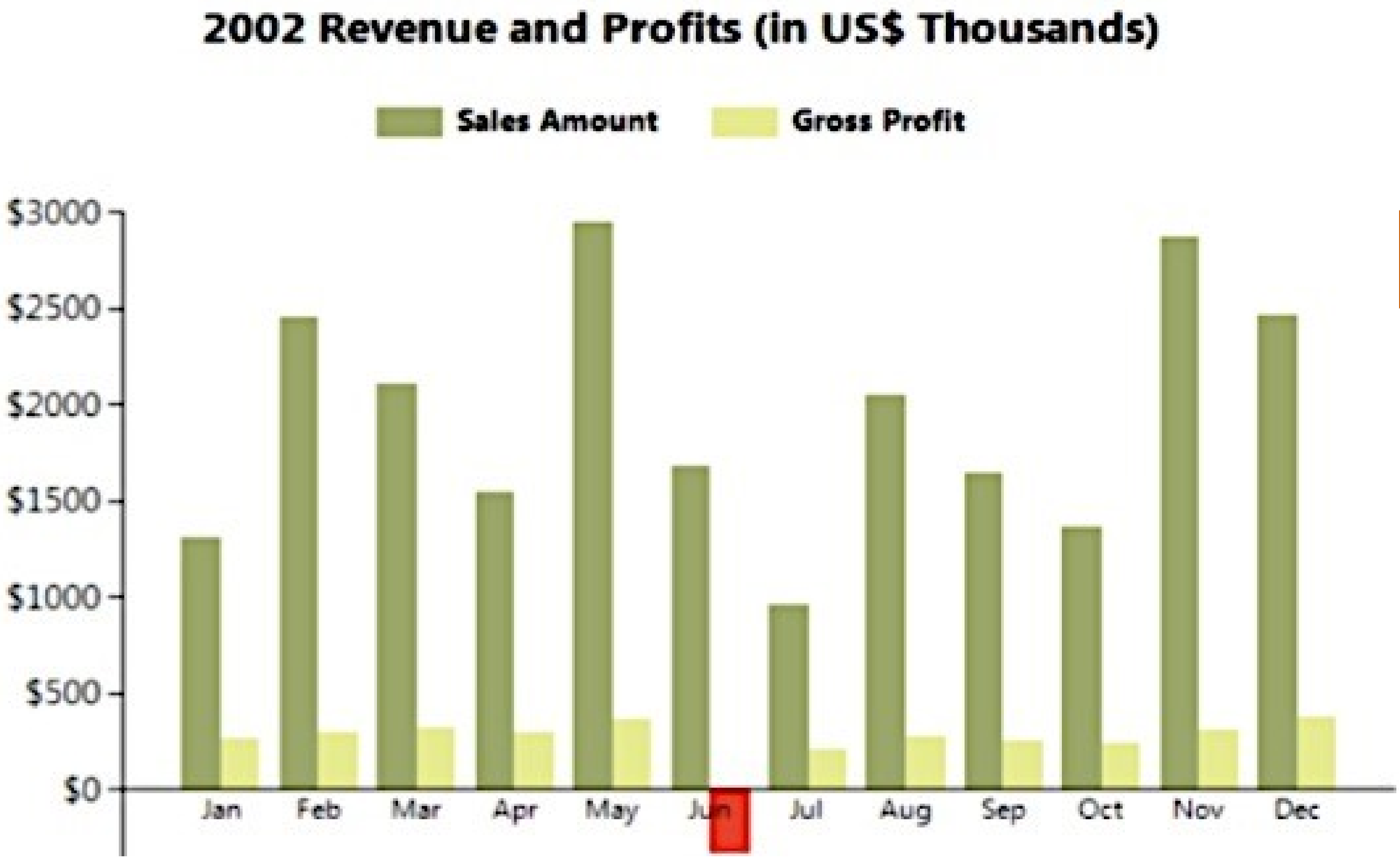
- What do you want to tell to an audience?

Example

Month of Year	Sales Amount	Total Product C...	Gross Profit Ma...	Gross Profit
January	1309863.2511	1046855.0401	0.20079058694...	263008.211
February	2451605.6244	2161789.71439...	0.11821473532...	289815.910000...
March	2099415.6158	1781531.84109...	0.15141536164...	317883.774700...
April	1546592.2292	1250946.0643	0.19115973772...	295646.164900...
May	2942672.90960...	2583467.20809...	0.12206783170...	359205.701500...
June	1678567.4193	2010739.61289...	-0.19789029012...	-332172.193599...
July	962716.741700...	754715.7636	0.21605625942...	208000.978100...
August	2044600.0034	1771778.75389...	0.13343502349...	272821.249500...
September	1639840.109	1393936.67389...	0.14995573882...	245903.43510001
October	1358050.4703	1124337.2647	0.17209463912...	233713.205600...
November	2868129.20330...	2561131.77409...	0.10703751729...	306997.42920002

https://www.slideshare.net/idigdata/data-visualization-best-practices-2013/6-Audience_Considerations_What_information_does

Example



2002 Revenue and Profits (in US$ Thousands)
Sales Amount
Gross Profit
$3000
$2500
$2000
$1500
$1000
$500
$0
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec

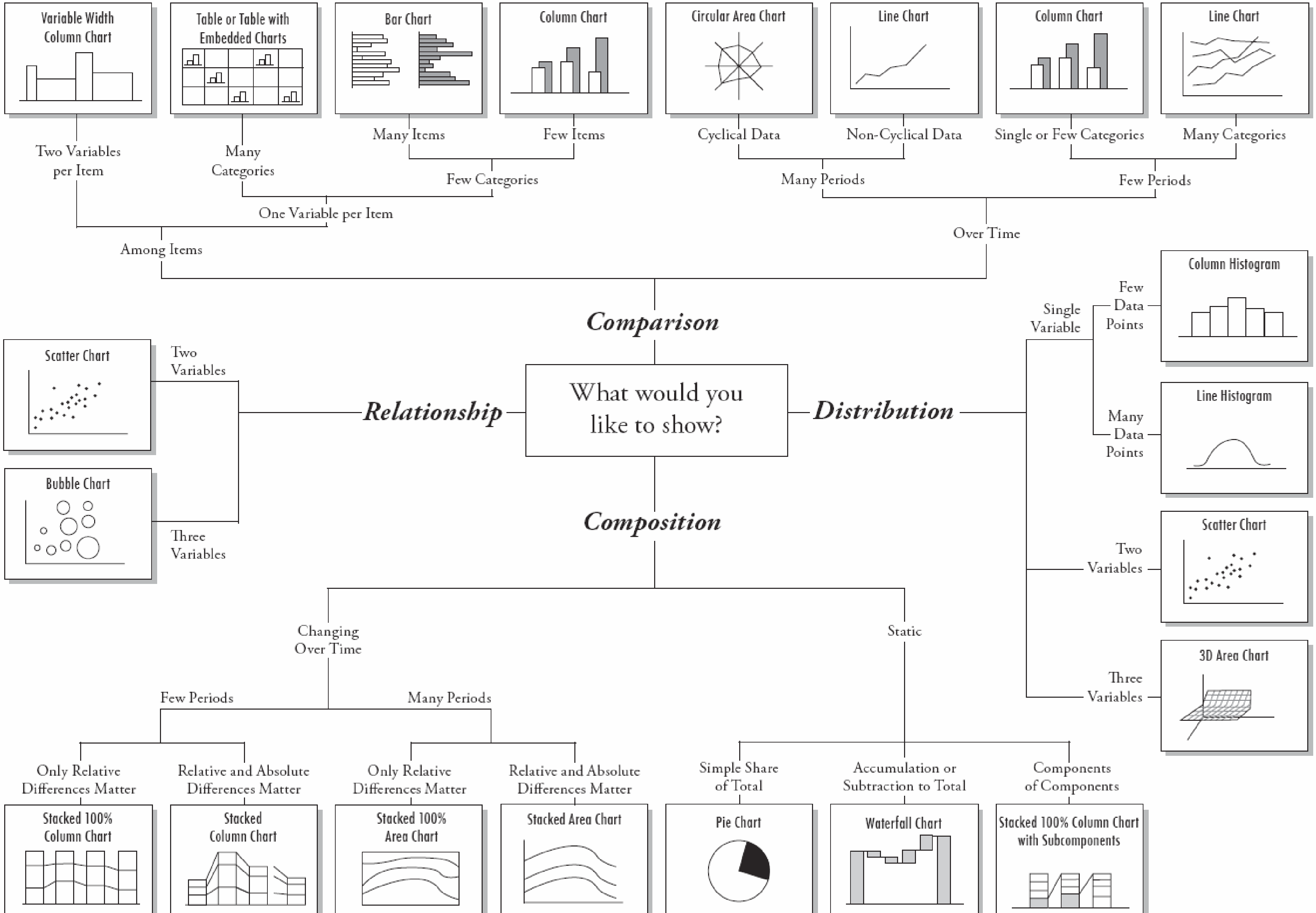
What is your audience's expectation?

- What information does the reader need?
- How much detail does the reader need?
- What action can be taken?
- What values need action?
- Any cultural assumptions that may affect the design choice?
- Other reason that may affect the design choice? E.g. color blindness

Common Data Visualization Issues

- Inappropriate display choices
- Variety for the sake of variety
- Too much information
- Poorly designed display choices
- Encoding quantitative data inaccurately
- Inconsistent ordering and placement
- Inconsistent or reversed scales
- Proportional axis scaling
- Using counts vs. percentages when comparing periods with different totals

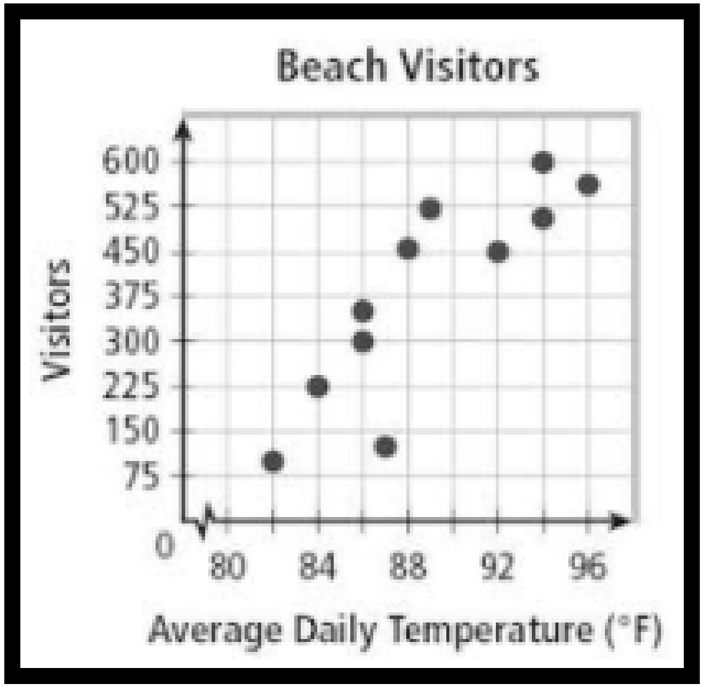
Choose the Right Chart Type for your Data



https://www.labnol.org/software/find-right-chart-type-for-your-data/6523/

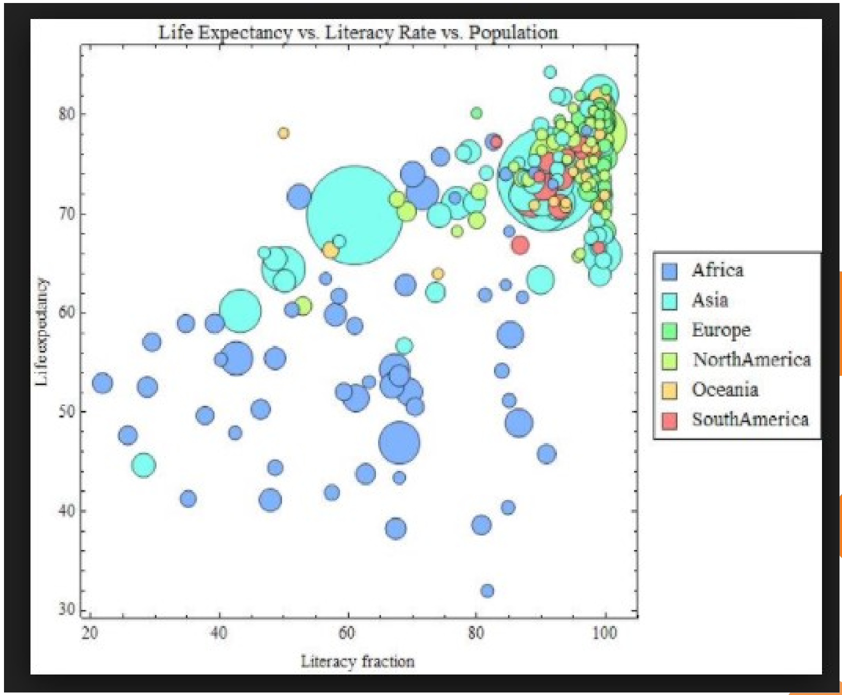
Relationship

Two variables



Scatter Chart

Three variables

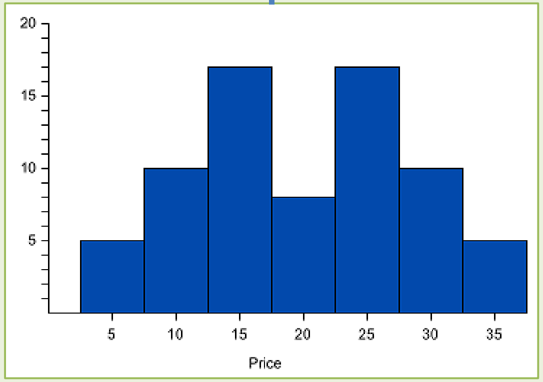


Bubble Chart

Distribution

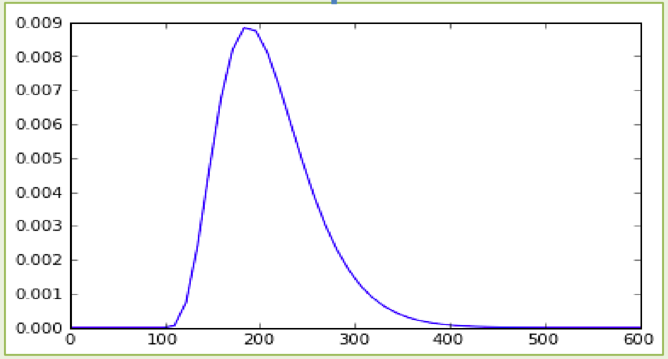
Single Variable

(few data points)



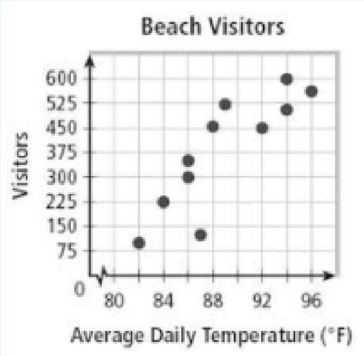
Column Histogram

(many data points)



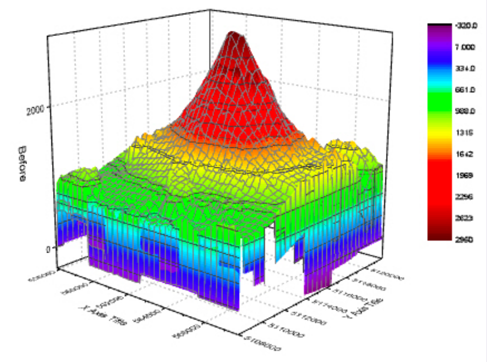
Line Histogram

Two Variables



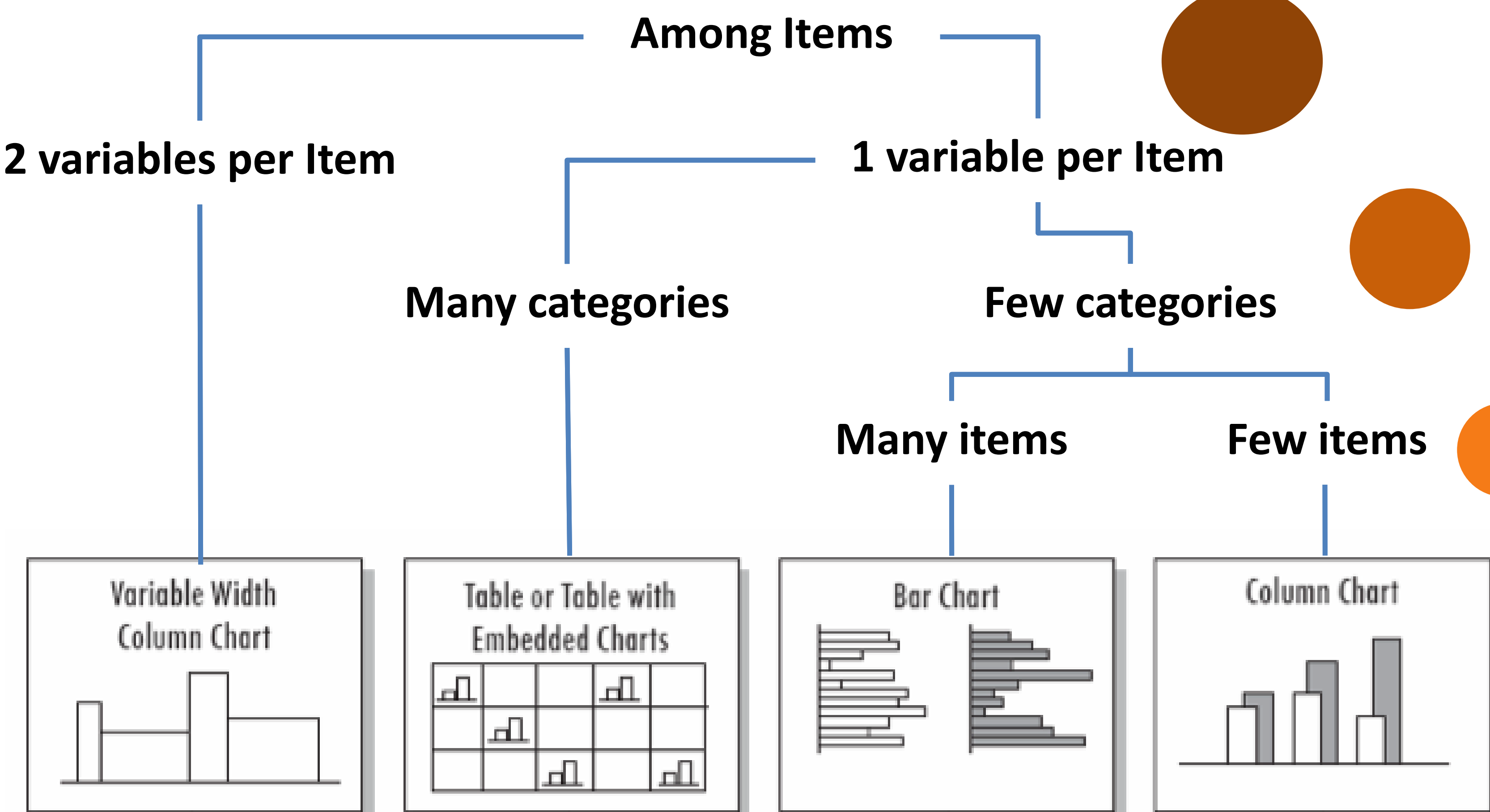
Scatter Chart

Three Variables



3D Area Chart

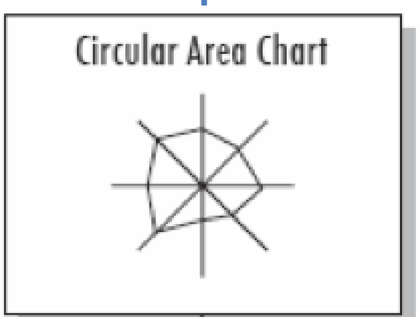
Comparisons

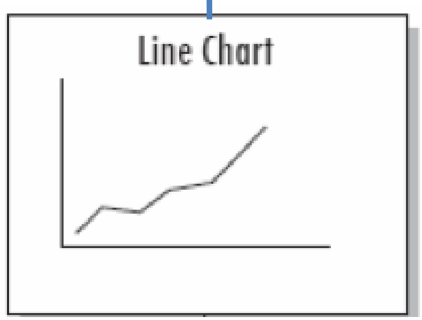


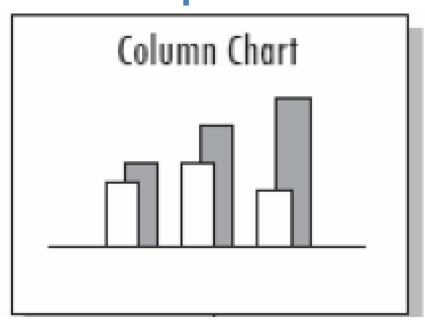
Comparisons

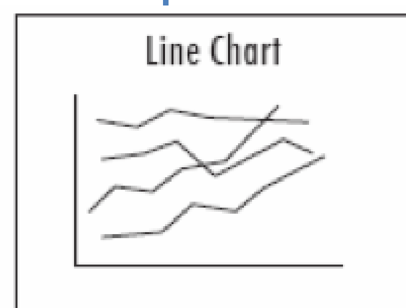
Over Time

- **Many periods**
  - **Cyclical data**
  - **Non-cyclical data**
- **Few periods**
  - **Single or few categories**
  - **Many categories**









Compositions

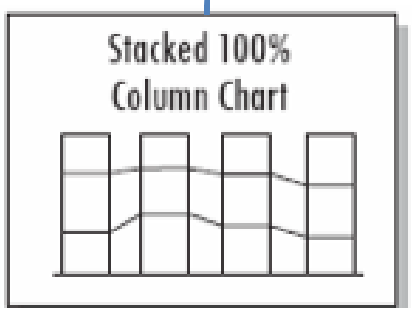
Changing over time

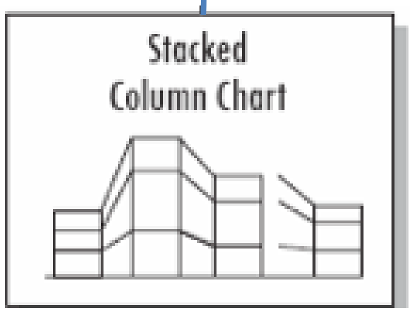
Few periods

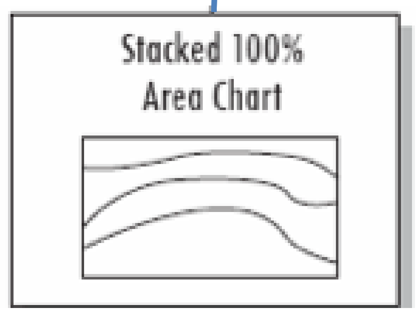
- **Only relative difference matters**
- **Relative and absolute differences matter**

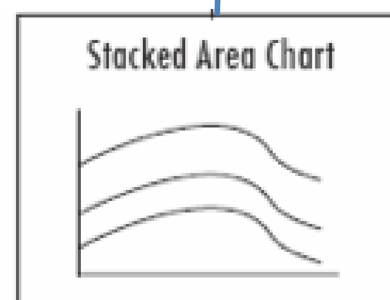
Many periods

- **Only relative difference matters**
- **Relative and absolute differences matter**









NEXT

DESCRIPTIVE ANALYTICS

Descriptive Analytics

- Statistical inference
- Association rules
- Sequence rules
- Segmentation

Descriptive and Inferential Analysis

Descriptive Analysis

- Descriptive statistical analysis limits generalization to the particular group of individuals observed. That is:
- No conclusion are extended beyond this group
- Any similarity to those outside the group cannot be assumed.
- The data describe one group and that group only.

Inferential Analysis

- Inferential analysis selects a small group (sample) out of a larger group (population) and the finding are applied to the larger group. It is used to estimate a parameter, the corresponding value in the population from which the sample is selected.
- It is necessary to carefully select the sample or the inferences may not apply to the population.

Statistical measures for descriptive data

- Measures of central tendency/average
  - Mean
  - Median
  - Mode
- Measure of spread/dispersion
  - Range
  - Variance
  - Standard deviation
- Measure of relative position
  - Standard scores
  - Percentile rank
  - Percentile score
- Measures of relationship
  - Coefficient of correlation

Association Rules

Detect frequently occurring patterns between items

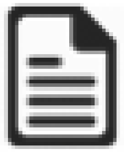
Detecting what products are frequently purchased together in a supermarket context.

Detecting what words frequently co-occur in a text document.

Detecting what elective courses are frequently chosen together in a university setting.

Sequence Rules

Detect sequences of events



Detecting sequences of purchase behavior in a supermarket context.

Detecting sequences of web page visits in a web mining context.

Detecting sequences of words in a text document.

Segmentation/Clustering

Detect homogeneous segments of observations

Differentiate between brands in a marketing portfolio.

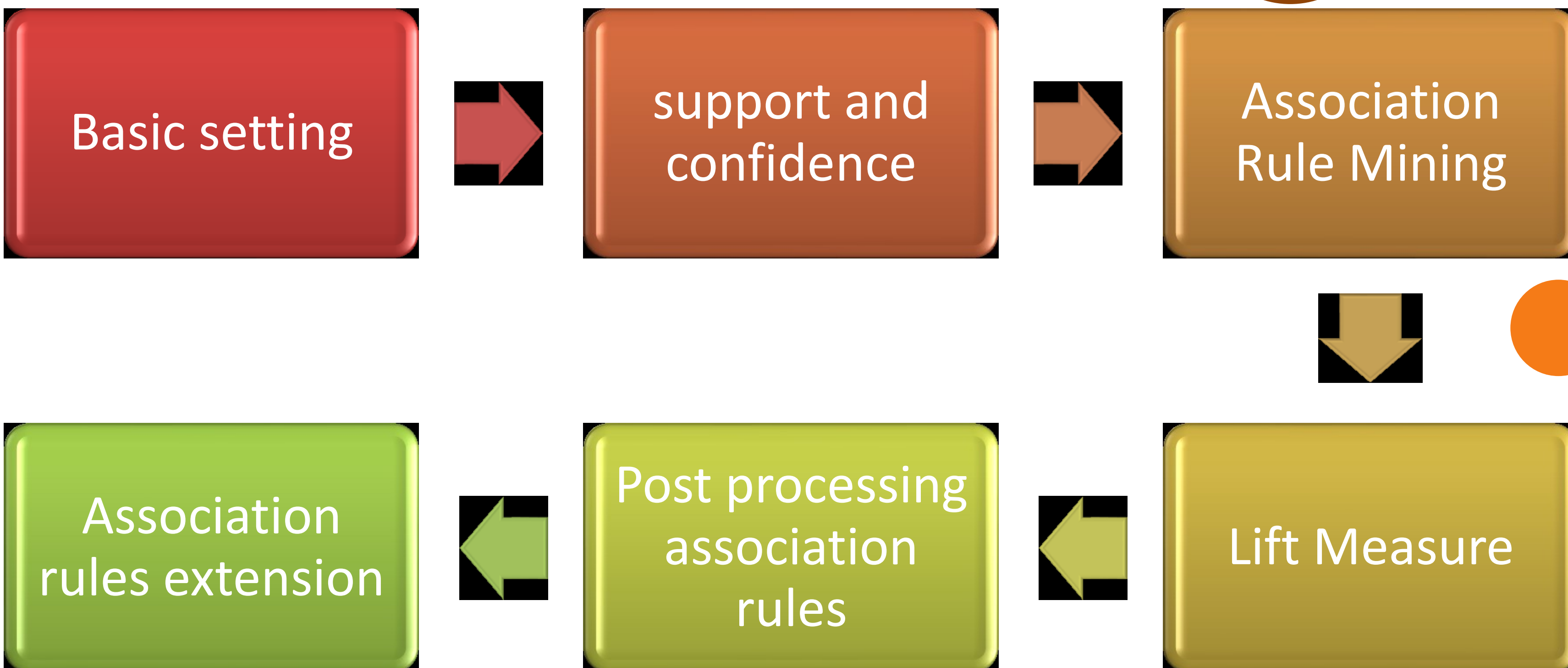
Segment customer population for targeted marketing.

NEXT

MINING ASSOCIATE RULES FROM DATA

Mining Associate Rules from Data

What will be discussed in this section:



Basic Setting

Transaction Identifier	Items
1	Beer, milk, diapers, baby food
2	Coke, beer, diapers
3 *D*	Cigarettes, diapers, baby food *I*
4	Chocolates, diapers, milk, apples
5	Tomatoes, water, apples, beer

- Association rules typically start from a database of transactions, D.
- Each transaction consists of a transaction identifier and a set of items, e.g. products) $\{i_1, i_2, \ldots, i_n\}$ selected from all possible items (I).

Basic Setting

- An association rule is the form $X \Rightarrow Y$,

  where $X \subset I$, $Y \subset I$ and $X \cap Y = \emptyset$.

Example

- If a customer buys spaghetti, then the customer buys red wine in 70% of the cases.

Write complete association rules between the X and Y given:

1. Diaper ⇒ Beer (60%)
2. Facebook ⇒ YouTube (85%)

Rules measure correlational associations between X and Y, but not as a causal effect.

Support and Confidence

- Two key features to quantify the strength of an association rule.
- The support of an item set is defined as the percentage of total transactions in the database that contains the item set, i.e.

$$support(X \cup Y) = \frac{number\ of\ transactions\ supporting\ (X \cup Y)}{total\ number\ of\ transactions}$$

- Example

Transaction Identifier	Items
1	Beer, milk, diapers, baby food
2	e, beer, diapers
3	rettes, diapers, baby food
4	Chocolates, diapers, milk, apples
5	Tomatoes, water, apples, beer

Diapers ⇒ baby food
Has support 2/5 = 40%

Support and Confidence

- A frequent item set is one for which the support is higher than a threshold (minsup) that is typically specified upfront by the business user or data analyst. A lower (higher) support will obviously generate more (less) frequent item sets.

Support and Confidence

Transaction Identifier	Items
1	Beer, milk, diapers, baby food
2	Coke, beer, diapers
3	Cigarettes, diapers, baby food
4	Chocolates, diapers, milk, apples
5	Tomatoes, water, apples, beer

- It can be formally defined as follows:

$$confidence(X \rightarrow Y) = P(Y \mid X) = \frac{support(X \cup Y)}{support(X)}$$

Diapers ⇒ Beer
Has confidence 2/4 = 50%

Again, the data analyst has to specify a minimum confidence (minconf) in order for an association rule to be considered interesting.

Association Rule Mining

Mining association rules from data is essentially a two-step process as follows:

Identification

- Identify all item sets having support above minsup ("frequent" item sets)

Discovery

- Discover all derived association rules having confidence above minconf

Step 1: Identification with Apriori algorithm

- Typically performed using the **Apriori** algorithm.
- The basic notion of a priori states that every subset of a frequent item set is frequent as well or, conversely, every superset of an infrequent item set is infrequent.
- This implies that candidate item sets with k items can be found by pairwise joining frequent item sets with $k - 1$ items and deleting those sets that have infrequent subsets.

Step 1: Identification with Apriori algorithm

Database

TID	Items
100	1, 3, 4
200	2, 3, 5
300	1, 2, 3, 5
400	2, 5

L_1

Itemsets	Support
{1}	2/4
{2}	3/4
{3}	3/4
{5}	3/4

Minsup = 50%

C_2

Itemsets	Support
{1, 2}	1/4
{1, 3}	2/4
{1, 5}	1/4
{2, 3}	2/4
{2, 5}	3/4
{3, 5}	2/4

L_2

Itemsets	Support
{1, 3}	2/4
{2, 3}	2/4
{2, 5}	3/4
{3, 5}	2/4

C_3

Itemsets	Support
{2, 3, 5}	2/4

L_3

Itemsets	Support
{2, 3, 5}	2/4

{1,3} and {2,3} give {1,2,3}, but because {1,2} is not frequent, you do not have to consider it!

Result = { {1},{2},{3},{5},{1,3},{2,3},{2,5},{3,5},{2,3,5} }

Step 2: Discovery

- Once the frequent item sets have been found, the association rules can be generated in a straightforward way, as follows:
  - –For each frequent item set k, generate all nonempty subsets of k
  - –For every nonempty subset s of k, output the rule $s \Rightarrow k - s$ if the confidence > minconf

Step 2: Discovery

- Assume that the following are given:

1. diapers, beer ⇒ baby food [conf = 75%]
2. baby food, beer ⇒ diapers [conf = 75%]
3. baby food, diapers ⇒ beer [conf = 60%]
4. beer ⇒ baby food and diapers [conf = 50%]
5. baby food ⇒ diapers and beer [conf = 43%]
6. diapers ⇒ baby food and beer [conf = 43%]

If the minconf is set to 70%, identify the association rules to be kept for further analysis.

The Lift Measure

- Consider a supermarket transactions database below.

	Tea	Not Tea	Total
Coffee	150	750	900
Not coffee	50	50	100
Total	200	800	1,000

- Assume that association rule tea ⇒ coffee.
- The support of this rule is 100/1,000, or 10%.
- The confidence of the rule is 150/200, or 75%.
- The prior probability of buying coffee equals 900/1000, or 90%.

A customer who buys tea is less likely to buy coffee than a customer about whom we have no information.

The Lift Measure

- The lift, also referred to as the *interestingness measure*, takes this into account by **incorporating the prior probability** of the rule consequent, as follows:

$$Lift(X \to Y) = \frac{support(X \cup Y)}{support(X) \bullet support(Y)}$$

A lift value < 1 indicates a **negative dependence** or **substitution** effect.

A lift value > 1 indicates a **positive dependence** or **complementary** effect.

In previous example, the lift value equals 0.89, which clearly indicates the expected substitution effect between coffee and tea.

Post Processing Association Rules

Typically, an association rule mining exercise will yield lots of association rules such that post processing will become a key activity.

Example steps that can be considered here are:

- Filter out the trivial rules that contain already known patterns (e.g., buying spaghetti and spaghetti sauce). This should be done in collaboration with a business expert.
- Perform a sensitivity analysis by varying the minsup and minconf values. Especially for rare but profitable items (e.g., Rolex watches), it could be interesting to lower the minsup value and find the interesting associations.

Post Processing Association Rules

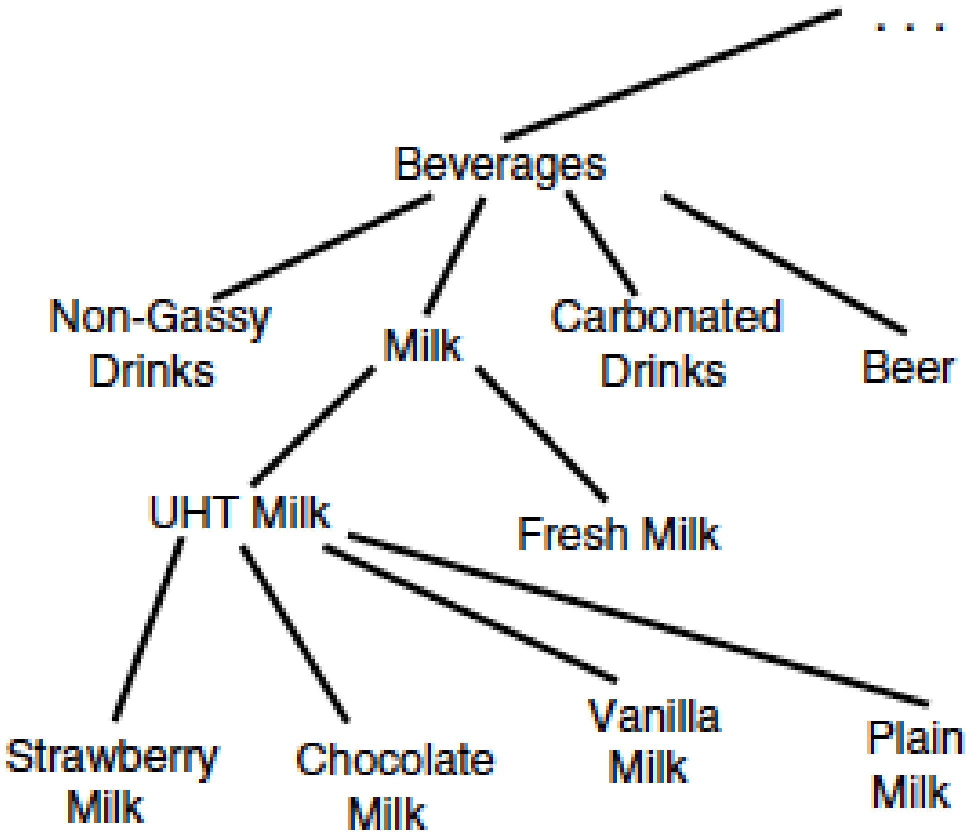
- Use appropriate visualization facilities (e.g., OLAP based) to find the unexpected rules that might represent novel and actionable behavior in the data.
- Measure the economic impact (e.g., profi t, cost) of the association rules.

Association Rule Extensions

- A first extension would be to include item quantities and/or price. This can be easily accomplished by adding discretized quantitative variables (e.g., three bottles of milk) to the transaction data set and mine the frequent item sets using the Apriori algorithm.
- Another extension is to also include the absence of items. Also, this can be achieved by adding the absence of items to the transactions data set and again mine using the Apriori algorithm.

Association Rule Extensions

- multilevel association rules mine association rules at different concept levels of a product taxonomy, as illustrated below.



Applications of Association Rules

market basket analysis

- The aim is to detect which products or services are frequently purchased together by analyzing market baskets. Finding these associations can have important implications for targeted marketing (e.g., next best offer), product bundling, store and shelf layout, and/or catalog design.

recommender systems

- These are the systems adopted by companies such as Amazon and Netflix to give a recommendation based on past purchases and/or browsing behavior.

NEXT

MINING SEQUENCE RULES FROM DATA

Mining Sequential Rules from Data

- Given a database D of customer transaction, the problem is to find the **maximal sequences** among all sequences that have certain user-specified minimum support and confidence.
- Transaction time or sequence field will be included in the analysis.
- Sequence rules are concerned about what items appear at different times (intertransaction patterns).

- Example (sequence of web page visits)

Home page ⇒ Electronics ⇒ Cameras and Camcorders ⇒ Digital Cameras ⇒ Shopping cart ⇒ Order confirmation ⇒ Return to shopping

Mining Sequential Rules from Data

- To mine the sequence rules, one can again make use of the apriori property because if a sequential pattern of length k is infrequent, its supersets of length $k + 1$ cannot be frequent.

Mining Sequential Rules from Data

- Example

Session ID	Page	Sequence
1	A	1
1	B	2
1	C	3
2	B	1
2	C	2
3	A	1
3	C	2
3	D	3
4	A	1
4	B	2
4	D	3
5	D	1
5	C	1
5	A	1

A sequential version can then be obtained as follows:

Session 1: A, B, C
Session 2: B, C
Session 3: A, C, D
Session 4: A, B, D
Session 5: D, C, A

Mining Sequential Rules from Data

Based on the sequential version obtained:

- **Session 1: A, B, C**
- Session 2: B, C
- **Session 3: A, C, D**
- Session 4: **A**, B, D
- Session 5: D, C, **A**

Consider the sequence rule A ⇒ C

Approach 1: C can appear in any subsequent stage

Support is 2/5 = 40%

Confidence is 2/4 = 50%

Approach 2: C must appear right after A

Support is 1/5 = 20%

Confidence is ¼ = 25%

Mining Sequential Rules from Data

- Recall the previous equation of confidence(c) as follows:

$$confidence(X \rightarrow Y) = P(Y \mid X) = \frac{support(X \cup Y)}{support(X)}$$

- Then the confidence of a rule of $A_1 \Rightarrow A_2$ is defined as

$$\boxed{P(A_2 \mid A_1) = support(A_1 \cup A_2)/support(A_1)}$$

- For a rule with multiple items, $A_1 \Rightarrow A_2 \Rightarrow \dots A_{n-1} \Rightarrow A_n$, the confidence is defined as

P($A_n \mid A_1 \Rightarrow A_2 \Rightarrow \dots A_{n-1} \Rightarrow A_n$) = support($A_1 \Rightarrow A_2 \Rightarrow \dots A_{n-1} \Rightarrow A_n$)/support ($A_1 \Rightarrow A_2 \Rightarrow \dots A_{n-1} \Rightarrow A_n$).

NEXT

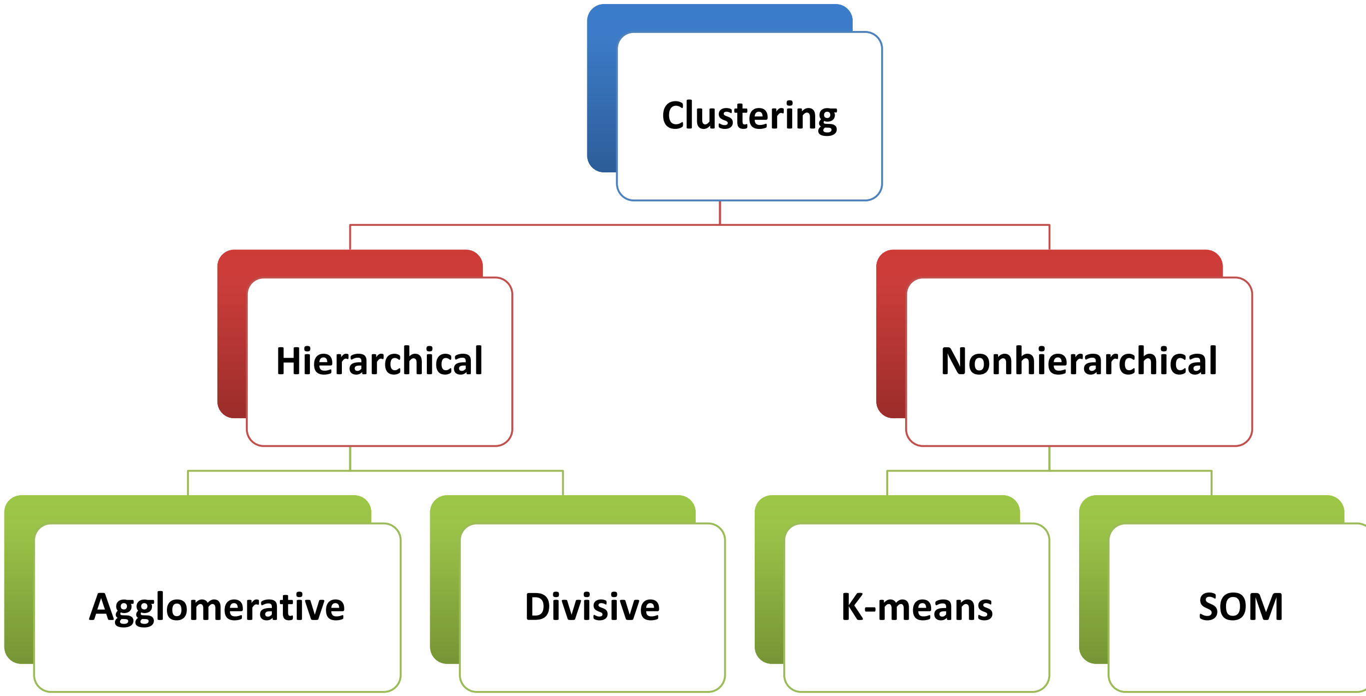
SEGMENTATION

Segmentation

- The aim of segmentation is to split up a set of customer observations into segments such that the homogeneity within a segment is maximized (cohesive) and the heterogeneity between segments is maximized (separated).
- Popular applications include:
  - –Understanding a customer population (e.g., targeted marketing or advertising [mass customization])
  - –Efficiently allocating marketing resources
  - –Differentiating between brands in a portfolio
  - –Identifying the most profitable customers
  - –Identifying shopping patterns
  - –Identifying the need for new products

Segmentation

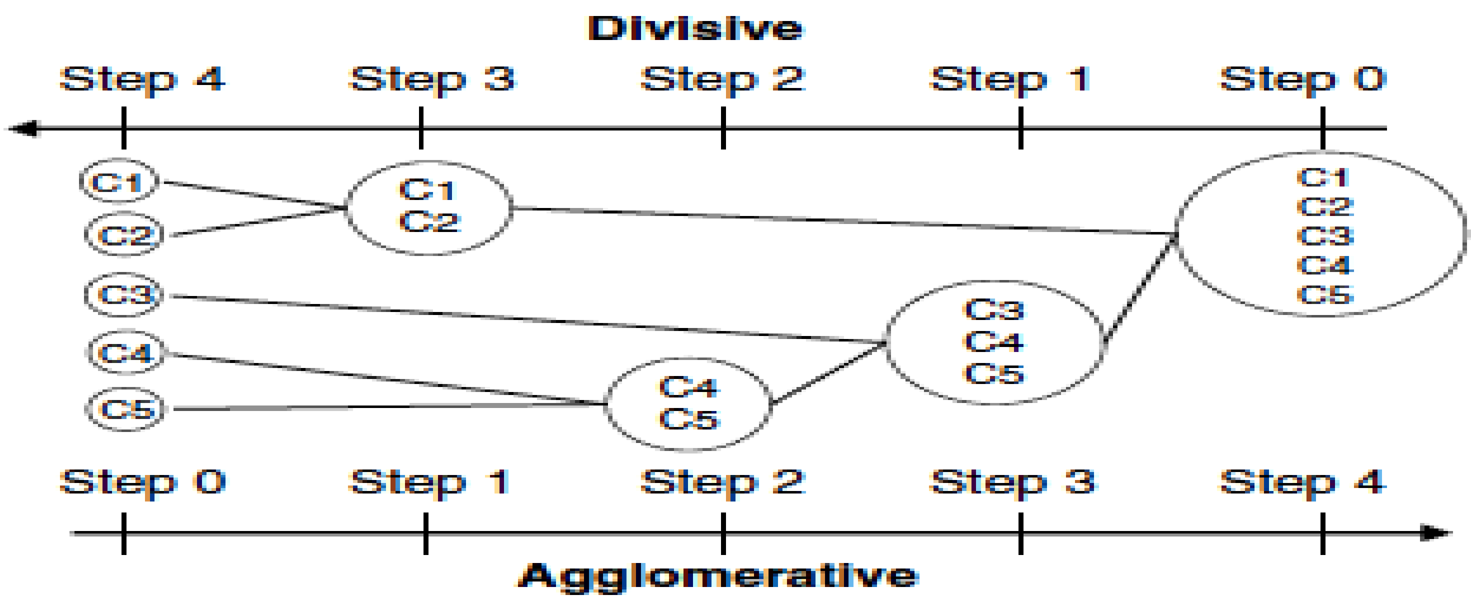
Clustering techniques can be categorized as either hierarchical or Nonhierarchical.



Various types of clustering data can be used, such as demographic, lifestyle, attitudinal, behavioral, RFM, acquisitional, social network, and so on.

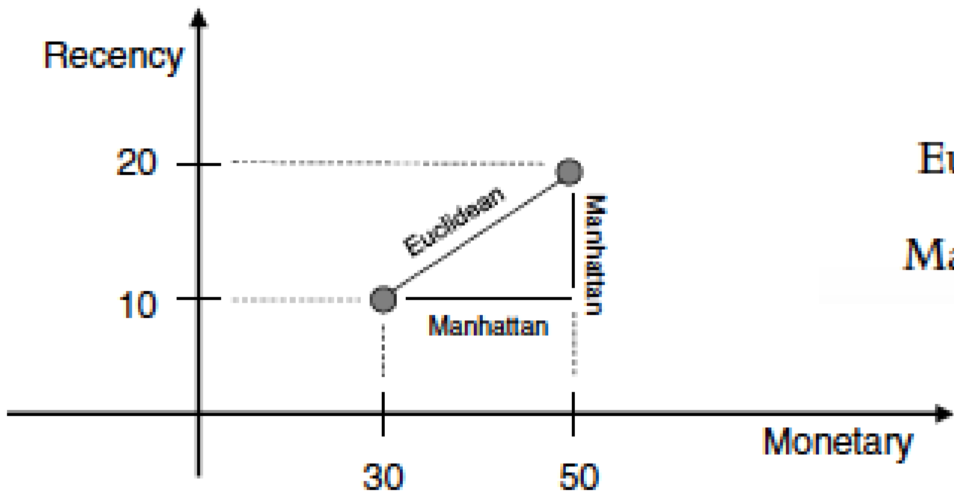
Segmentation: Hierarchical Clustering

- **Divisive hierarchical clustering** starts from the whole data set in one cluster, and then breaks this up in each time smaller clusters until one observation per cluster remains.
- **Agglomerative clustering** starts from all observations in one cluster and continuing to merge the ones that are most similar until all observations make up one big cluster.



Segmentation: Nonhierarchical Clustering

- In order to decide on the merger or splitting, a similarity rule is needed. Examples of popular similarity rules are the Euclidean distance and Manhattan (city block) distance

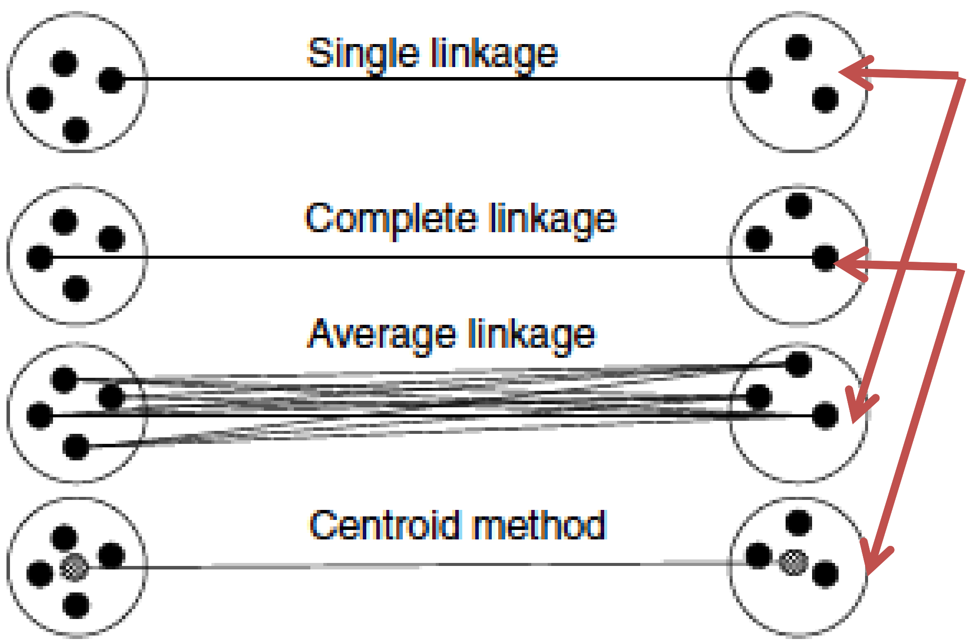


$$\text{Euclidean}: \sqrt{(50-30)^2+(20-10)^2} = 22$$

$$\text{Manhattan}: |50-30|+|20-10| = 30$$

Segmentation: Clustering

- Various schemes can be adopted to calculate the distance between 2 clusters. The single linkage method



Average linkage method calculates the average of all

Centroid method calculates the distance between the centroids of both clusters.

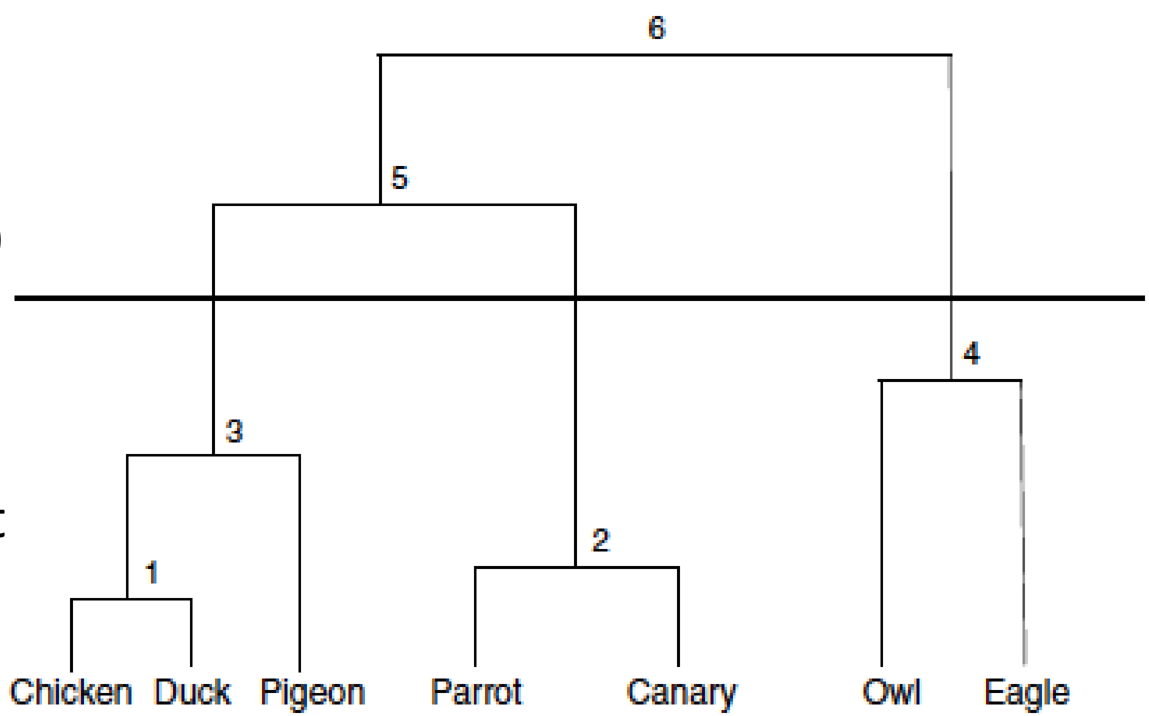
distance between the two most dissimilar

Finally, Ward's method merges the pair of clusters that leads to the minimum increase in total within-cluster variance after merging.

Segmentation: Clustering with Dendogram

- To decide on the optimal number of clusters, one could use a dendrogram or scree plot.

A dendrogram is a tree-like diagram that records the sequences of merges. The vertical (or horizontal scale) then gives the distance between two clusters amalgamated. One can then cut the dendrogram at the desired level to find the optimal clustering.



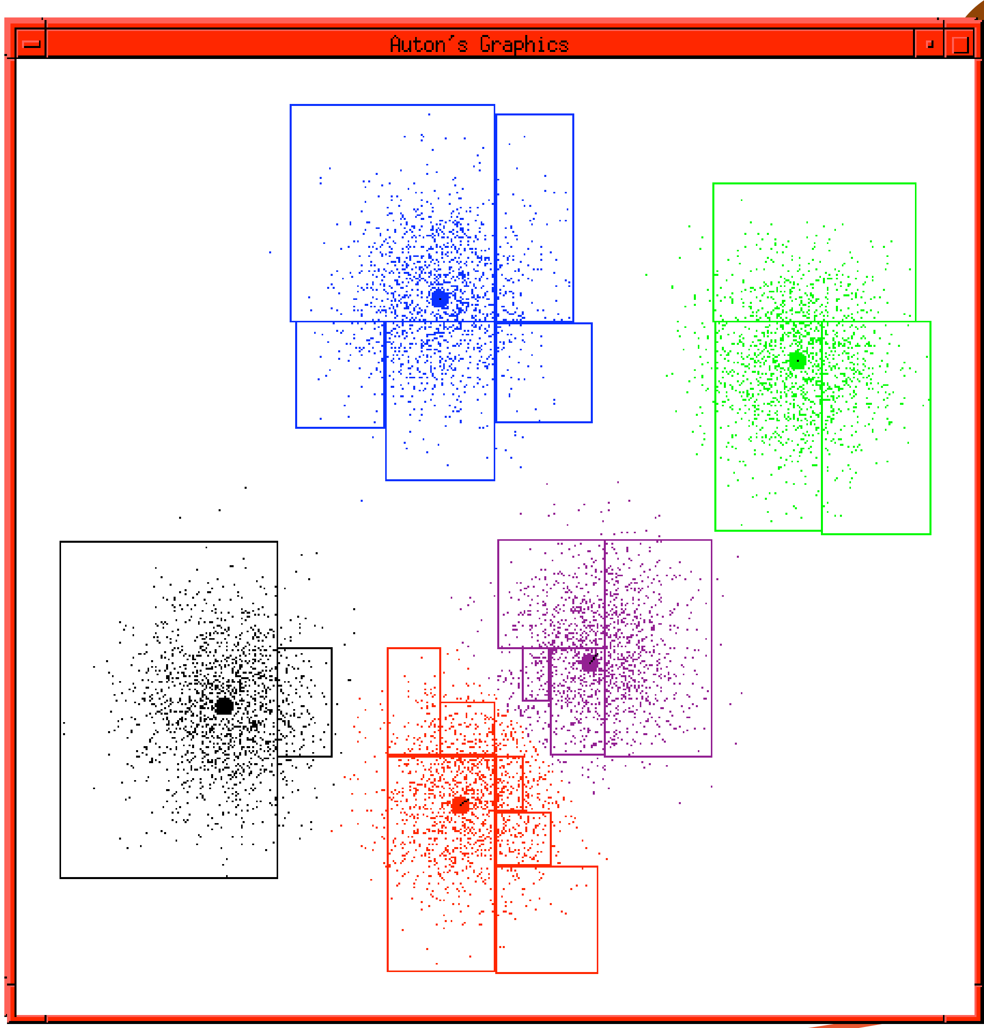
Segmentation: K-Means Clustering

K -means clustering is a nonhierarchical procedure that works along the following steps:

1. Select *k* observations as initial cluster centroids (seeds).
2. Assign each observation to the cluster that has the closest centroid (for example, in Euclidean sense).
3. When all observations have been assigned, recalculate the positions of the *k* centroids.
4. Repeat until the cluster centroids no longer change.

A key requirement here is that the number of clusters, k, needs to be specified before the start of the analysis.

Segmentation: K-Means Clustering

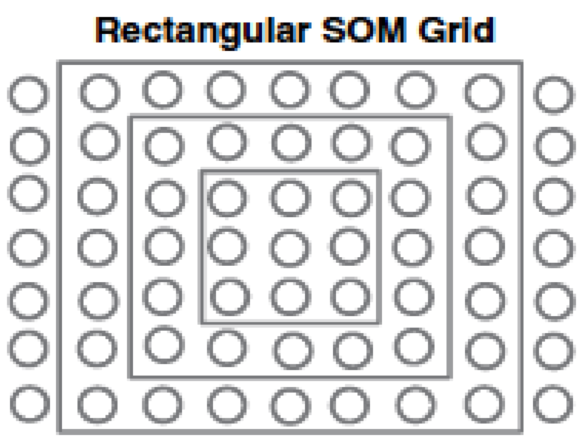


Example generated by Andrew Moore using Dan Pelleg's super-duper fast K-means system:

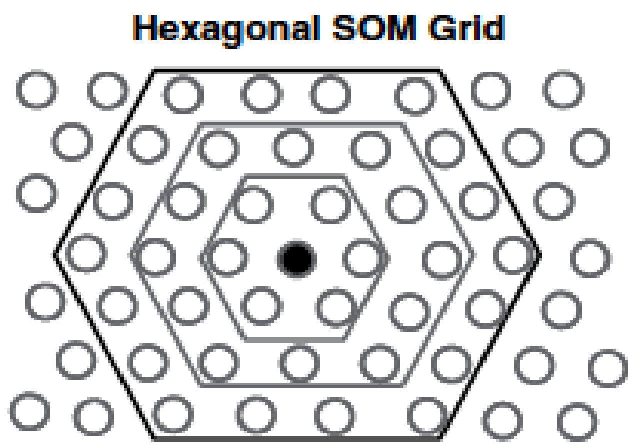
Dan Pelleg and Andrew Moore. Accelerating Exact k-means Algorithms with Geometric Reasoning. Proc. Conference on Knowledge Discovery in Databases 1999.

Segmentation: Self-Organizing Maps (SOM)

- A self-organizing map (SOM) is an unsupervised learning algorithm that allows you to visualize and cluster high-dimensional data on a low-dimensional grid of neurons.
- An SOM is a feedforward neural network with two layers. The neurons from the output layer are usually ordered in a two-dimensional rectangular or hexagonal grid.



every neuron has at most eight neighbors



every neuron has at most six neighbors

Segmentation: Self-Organizing Maps (SOM)

- Each input is connected to all neurons in the output layer with weights w = [w1 , ..., w N], with N the number of variables.
- All weights are randomly initialized. When a training vector x is presented, the weight vector w c of each neuron c is compared with x, using, for example, the Euclidean distance metric.
- The neuron that is most similar to *x* in Euclidean sense is called the best matching unit (BMU).
- The weight vector of the BMU and its neighbors in the grid are then adapted using the following learning rule:

$$w_i(t+1) = w_i(t+1) + h_{ci}(t)\left[x(t) - w_i(t)\right]$$

where t represents the time index during training, and $h_{ci}(t)$ defines the neighborhood of the BMU c, specifying the region of influence.

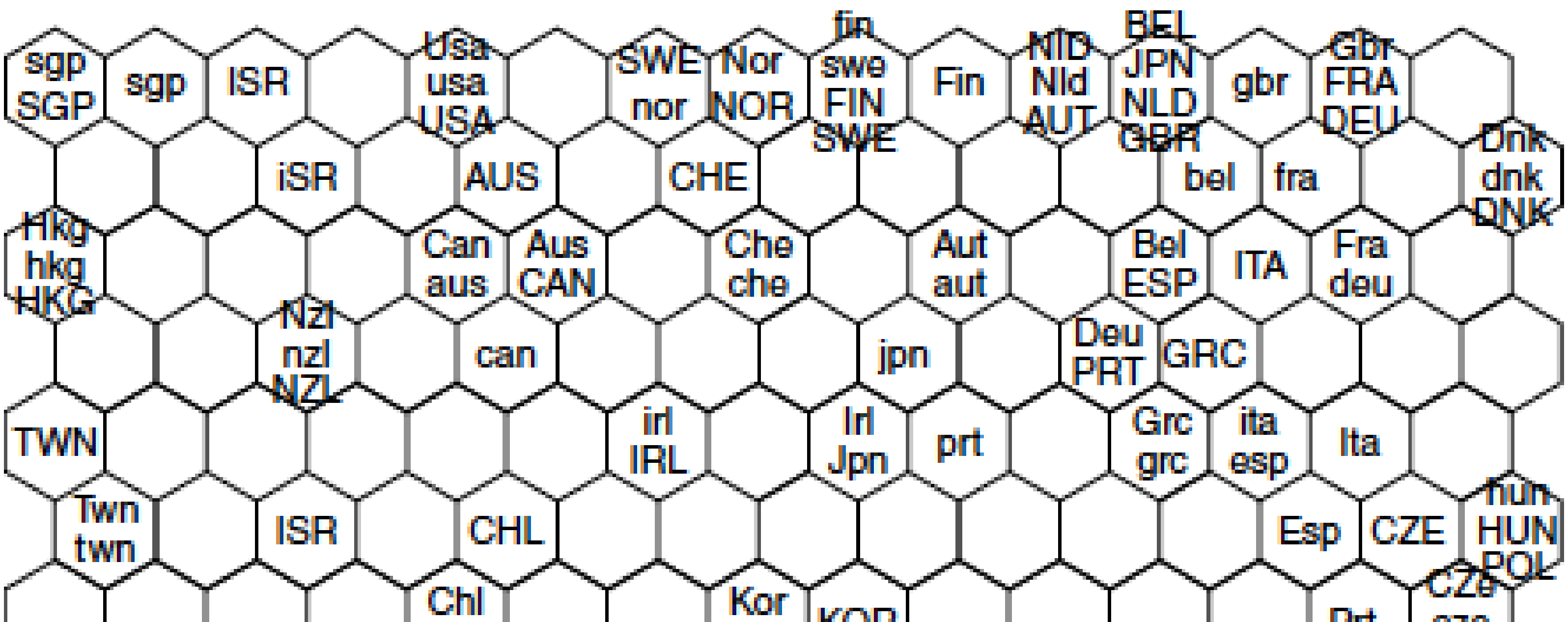
Segmentation: Self-Organizing Maps (SOM)

- The decreasing learning rate and radius will give a stable map after a certain amount of training. Training is stopped when the BMUs remain stable, or after a fixed number of iterations (e.g., 500 times the number of SOM neurons). The neurons will then move more and more toward the input observations and interesting segments will emerge.

Segmentation: Self-Organizing Maps (SOM)

- Sample visualization of SOM

Clustering Countries based on a corruption perception index (CPI), and score between 0 (highly corrupt) and 10 (highly clean) assigned to each country in the world. The CPI is combined with demographic and macroeconomic information for the years 1996, 2000 and 2004.



Segmentation: Self-Organizing Maps (SOM)

- **Limitations of SOM**
- SOMs are a very handy tool for clustering high-dimensional data sets because of the visualization facilities.
- However, since there is no real objective function to minimize, it is harder to compare various SOM solutions against each other.
- Also, experimental evaluation and expert interpretation are needed to decide on the optimal size of the SOM. Unlike k -means clustering, an SOM does not force the number of clusters to be equal to the number of output neurons.

Using and Interpreting Clustering Solutions

- In order to use a clustering scheme, one can assign new observations to the cluster for which the centroid is closest (e.g., in Euclidean or Manhattan sense).
- To facilitate the interpretation of a clustering solution, one could do the following:
  - Compare cluster averages with population averages for all variables using histograms, for example.
  - Build a decision tree with the cluster ID as the target and the clustering variables as the inputs (can also be used to assign new observations to clusters).
- It is also important to check cluster stability by running different clustering techniques on different samples with different parameter settings and check the robustness of the solution.

What you have learned

- Visualization in data analytics
- Descriptive Analytics
  - Statistical inference
  - Association rules
  - Sequence rules
  - Segmentation