Joerg Niessing
Affiliate Professor of Marketing

Theos Evgeniou
Professor of Decision Sciences

# [Big]-Data Analytics for Businesses

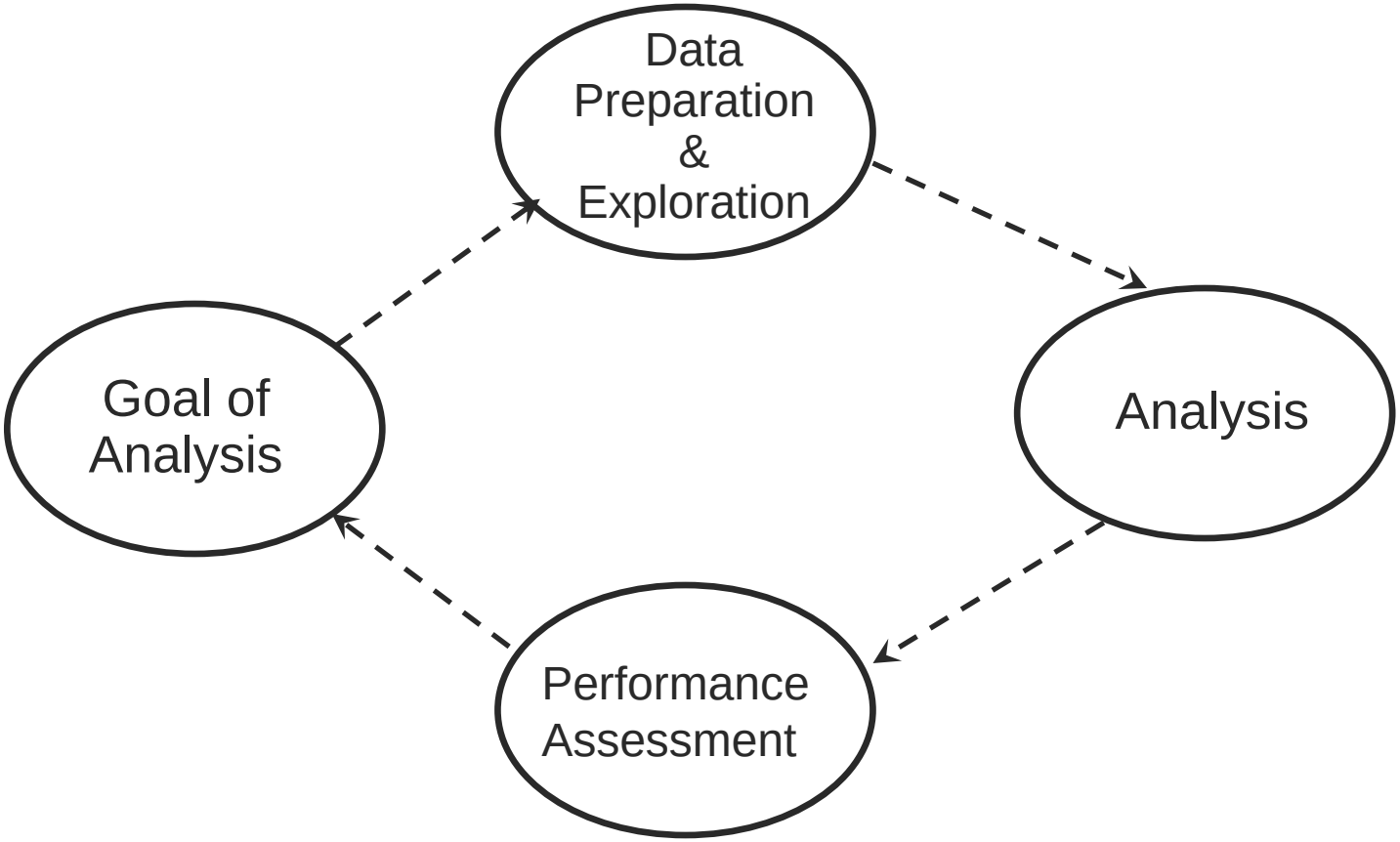
# Why this class? My three goals

1. "Big Picture": Develop your intuition about identifying data analytics opportunities and their implementation challenges

2. "Medium Picture": learn how to approach data analytics projects

3. "Dirty Hands": Learn how to perform, read, and use key data analytics methods.

# The *Iterative Process Cycle*



# Basic Types of Questions and Tools

1. Market Basket Analysis: which pairs of products are typically sold together? – "On Friday evenings, shoppers who buy diapers also buy beer".

2. Factor Analysis: Finding important dimensions ("factors") that summarize your data, and visualizing your data

3. Clustering: What are the main types of customers we have?

4. Discriminant Analysis: How can we differentiate between the "high value" and "low value" customers?

# Class Outline: three tools you will learn

1. Finding important factors that summarize your data, and visualizing your data:

   Factor Analysis (Sessions 2 and 3)

2. **Finding a few clusters of similar data:**

   **Cluster Analysis (Sessions 4 and 5)**

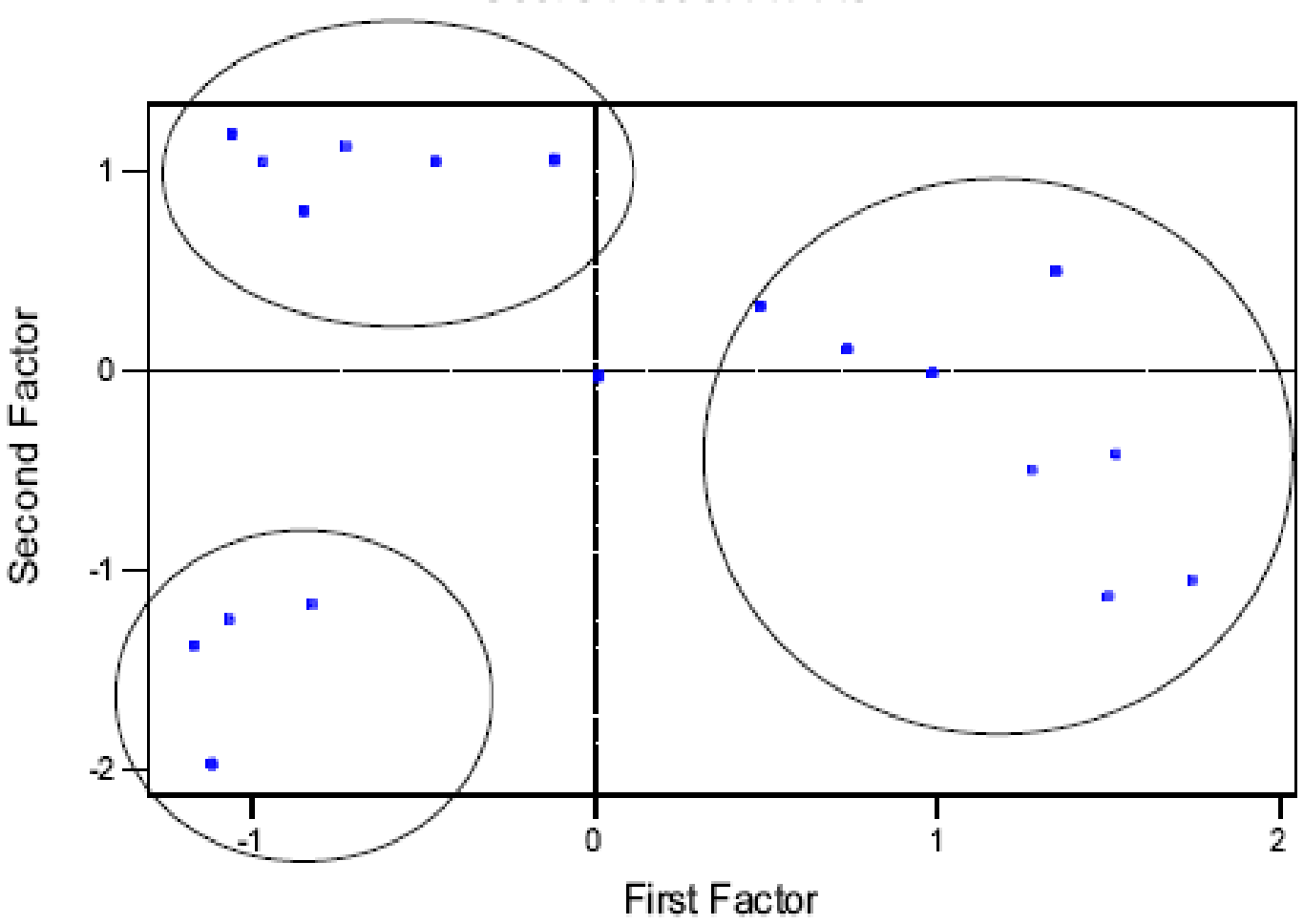
3. Discriminating among and predicting successes vs failures:

   Logistic Regression and Tree Analyses (Sessions 6 and 7)

# Today's Plan

- Cluster Analysis and Segmentation

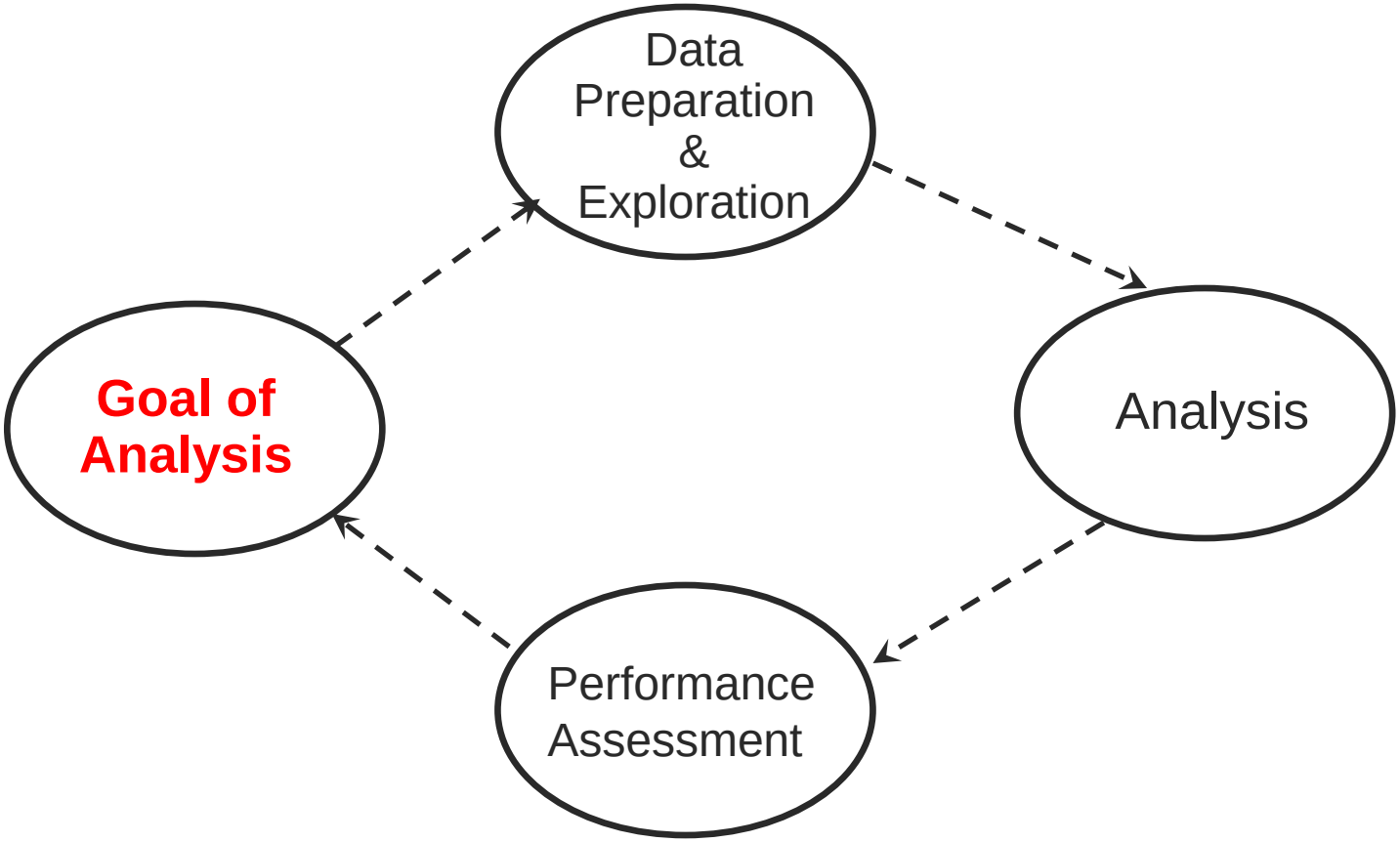
# Clustering (or “Segmentation”)



# Where is Cluster Analysis Used?

- Visualization/exploration of the data
- Major tool for market segmentation
- Recommender Systems (e.g. Amazon)
- Identifying the competitive set of products
- Identify similar financial assets
- Text Mining

# The *Iterative Process Cycle*



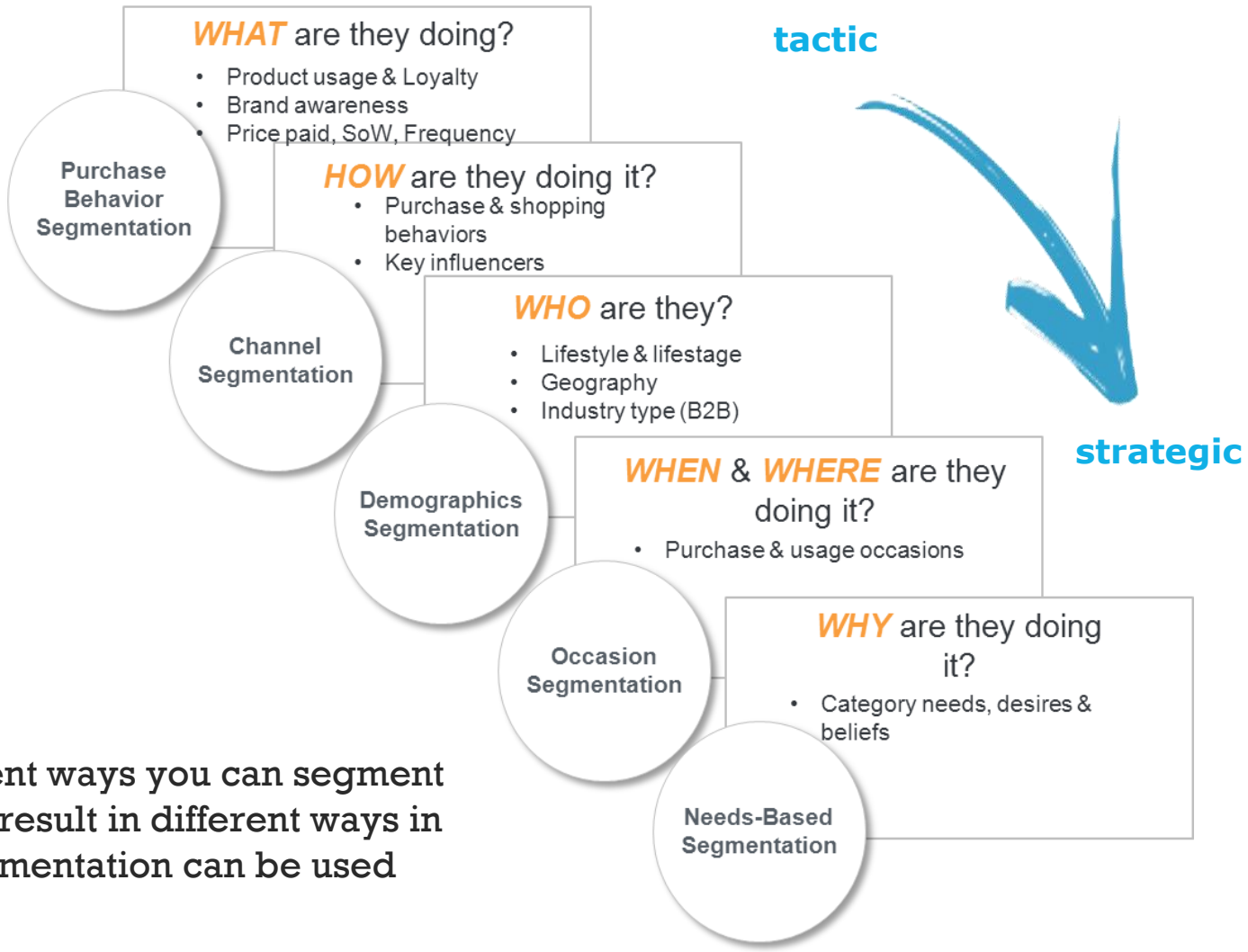
# Questions You Might Hear…








# Different Types of Data Used




There are different ways you can segment a market, which result in different ways in which the segmentation can be used

# 6 Steps in Cluster Analysis

- Data Preparation:
  - Step 1: what variables will we use for clustering?
- Modeling:
  - Step 2: how will we define the distance or similarities between observations? Do we standardize the data?
  - Step 3: selecting the clustering method
  - Step 4: how many clusters should we have?
- Evaluation:
  - Step 5: what do the clusters mean? (interpreting and profiling – a lot of subjectivity)
  - Step 6: perform sensitivity checks to assess stability of the clusters

# Cluster Analysis: Example

# Step 1: Choose Variables of Shopping Attitudes

- Based on exploratory research
- 1-7 Agree-Disagree Scale on the following
  - V1: Shopping is fun
  - V2: Shopping is bad for your budget
  - V3: I combine shopping with eating out
  - V4: I try to get the best buys while shopping
  - V5: I don't care about shopping
  - V6: You can save lots of money by comparing prices

## Step 2: How should we measure the distances between observations?

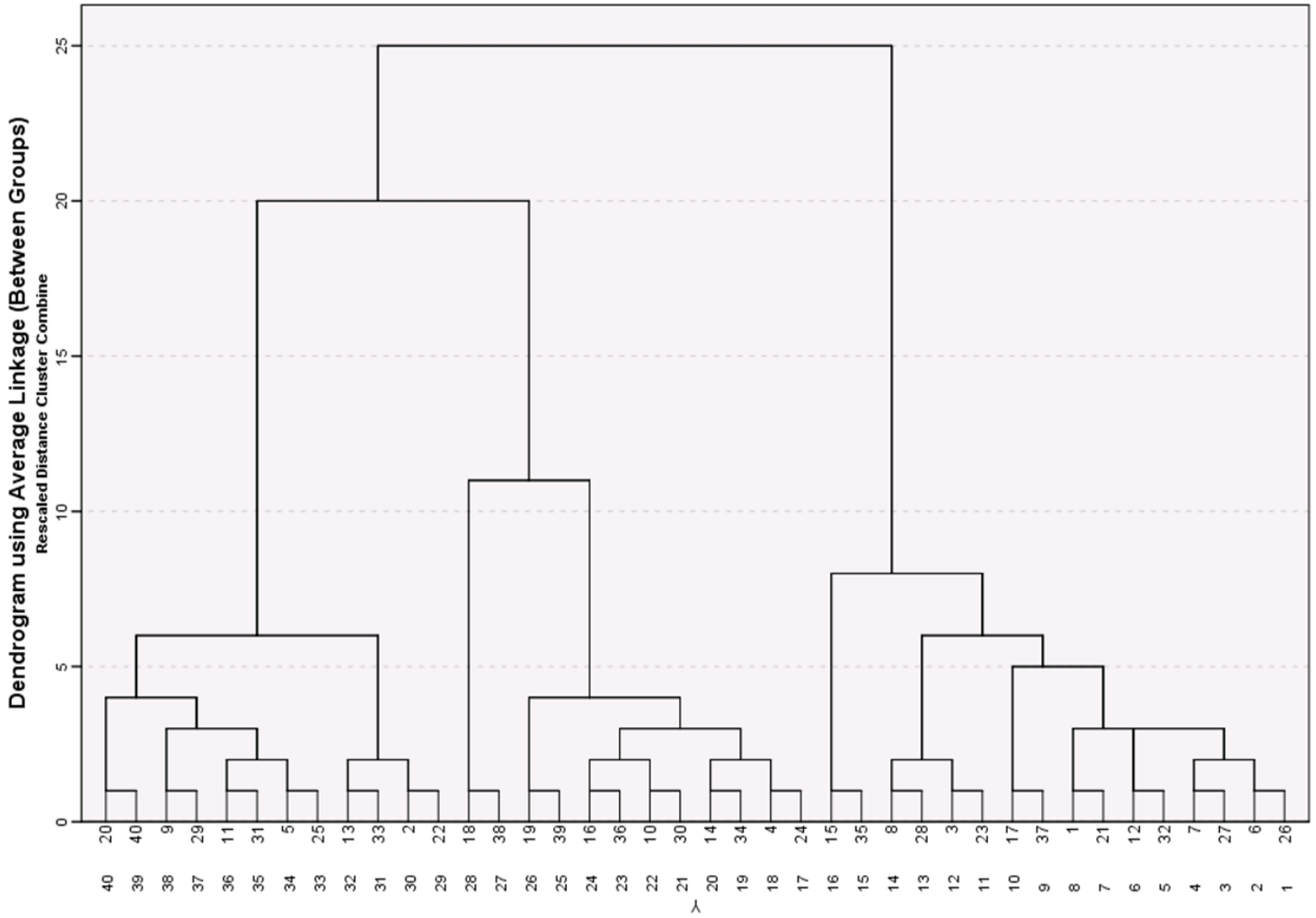
- Euclidean distance $D_{ij} = \sqrt{\sum_k \left(x_{ik} - x_{jk}\right)^2}$
- Squared Euclidean $D_{ij} = \sum_k \left(x_{ik} - x_{jk}\right)^2$
- Citi-block or Manhattan $D_{ij} = \sum_k \left|x_{ik} - x_{jk}\right|$
- Correlation $D_{ij} = \sum_k x_{ik} \cdot x_{jk}$

*You can be creative here…*

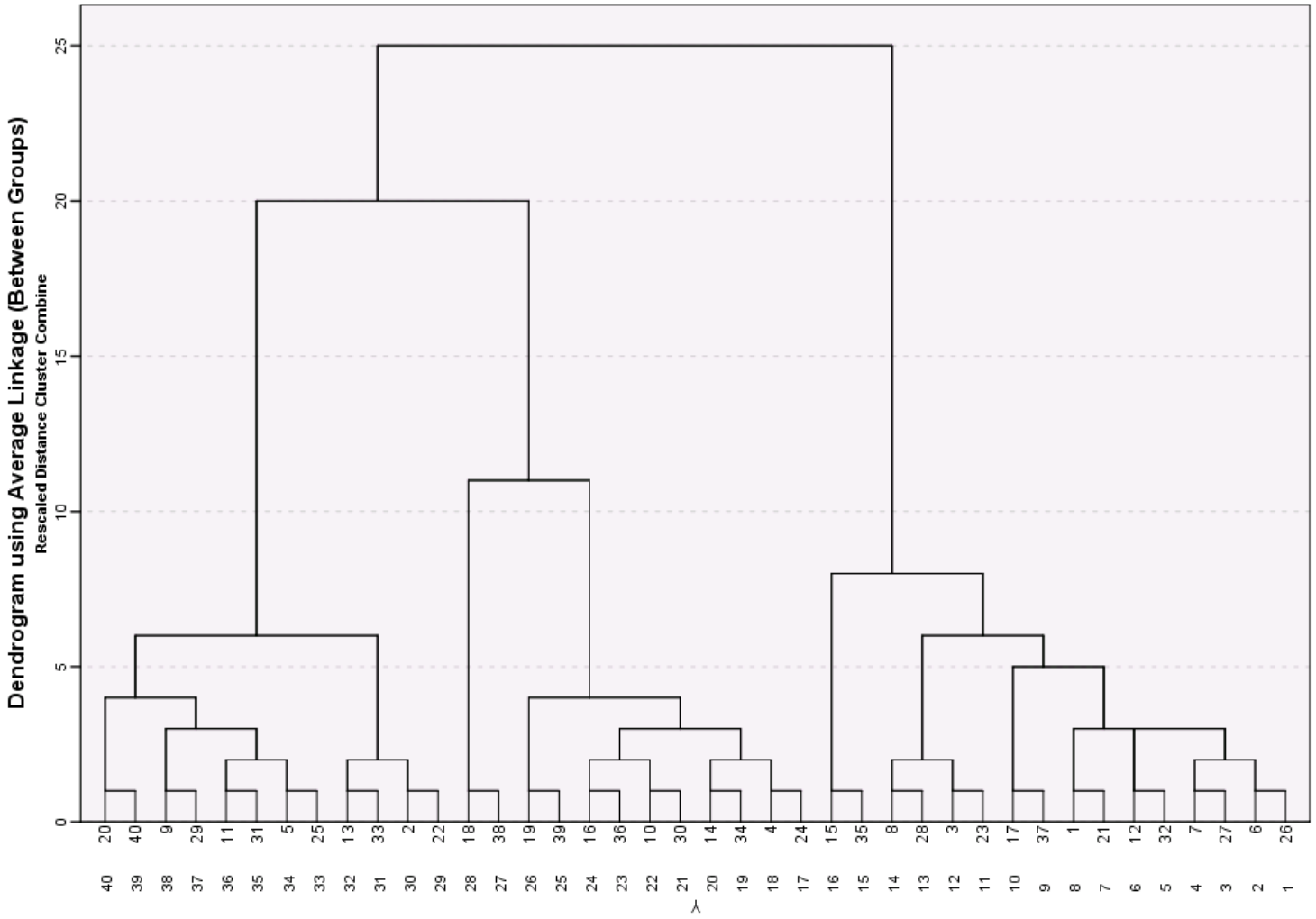
# Step 3: Choosing the Clustering Method

- Hierarchical Methods
- Non-Hierarchical Methods (e.g. k-means)

# Step 3: Method: Hierarchical Clustering



# Step 4: How Many Clusters?



# Step 5: Interpreting and Profiling the Clusters Using a plot of the centroids

V1: Shopping is fun
V2: Shopping is bad for your budget
V3: I combine shopping with eating out
V4: I try to get the best buys while shopping
V5: I don't care about shopping
V6: You can save lot of money by comparing prices

Report

| Average Linkage ... | | Income | Mall Visits | V1 | V2 | V3 | V4 | V5 | V6 |
|---|---|---|---|---|---|---|---|---|---|
| 1 | Mean | 60000.00 | 3.25 | 5.75 | 3.63 | 6.00 | 3.12 | 1.88 | 3.88 |
| | N | 16 | 16 | 16 | 16 | 16 | 16 | 16 | 16 |
| | Std. Deviation | 10954.451 | .683 | [illegible] | .885 | [illegible] | .806 | .806 | .619 |
| 2 | Mean | 42500.00 | 1.00 | 1.67 | 3.00 | 1.83 | 3.50 | 5.50 | 3.33 |
| | N | 12 | 12 | 12 | 12 | 12 | 12 | 12 | 12 |
| | Std. Deviation | 17516.226 | [illegible] | .492 | [illegible] | .718 | [illegible] | 1.000 | .778 |
| 3 | Mean | 32000.00 | 5.60 | 3.60 | 5.60 | 3.60 | 6.00 | 3.40 | 6.60 |
| | N | 10 | 10 | 10 | 10 | 10 | 10 | 10 | 10 |
| | Std. Deviation | 5374.838 | 1.075 | .516 | .516 | .516 | .667 | .843 | .516 |
| 4 | Mean | 25000.00 | [illegible] | 3.00 | [illegible] | 2.00 | 6.00 | 4.00 | 3.00 |
| | N | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| | Std. Deviation | .000 | .000 | .000 | .000 | .000 | .000 | .000 | .000 |
| Total | Mean | 46000.00 | 3.25 | 3.85 | 4.10 | 3.95 | 4.10 | 3.45 | 4.35 |
| | N | 40 | 40 | 40 | 40 | 40 | 40 | 40 | 40 |
| | Std. Deviation | 17216.569 | 1.945 | 1.875 | 1.392 | 1.986 | 1.499 | 1.739 | 1.477 |

Cluster 1: Cares and Enjoys Shopping

Cluster 2: Apathetic Shopper

Cluster 3: Economical Shopper

# Step 6: perform checks to assess validity of the clusters: Sensitivity to…

- Different Method
- Different Data Samples
- Different parameters…

**Report**

| Average Linkage … | | Income | Mall Visits | V1 | V2 | V3 | V4 | V5 | V6 |
|---|---|---|---|---|---|---|---|---|---|
| 1 | Mean | 60000.00 | 3.25 | 5.75 | 3.63 | 6.00 | 3.12 | 1.88 | 3.88 |
| | N | 16 | 16 | 16 | 16 | 16 | 16 | 16 | 16 |
| | Std. Deviation | 10954.451 | .683 | 1.000 | .885 | 1.033 | .806 | .806 | .619 |
| 2 | Mean | 42500.00 | 1.00 | 1.67 | 3.00 | 1.83 | 3.50 | 5.50 | 3.33 |
| | N | 12 | 12 | 12 | 12 | 12 | 12 | 12 | 12 |
| | Std. Deviation | 17516.226 | .853 | .492 | .853 | .718 | 1.068 | 1.000 | .778 |
| 3 | Mean | 32000.00 | 5.60 | 3.60 | 5.60 | 3.60 | 6.00 | 3.40 | 6.60 |
| | N | 10 | 10 | 10 | 10 | 10 | 10 | 10 | 10 |
| | Std. Deviation | 5374.838 | 1.075 | .516 | .516 | .516 | .667 | .843 | .516 |
| 4 | Mean | 25000.00 | 3.00 | 3.00 | 1.00 | 2.00 | 6.00 | 4.00 | 3.00 |
| | N | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| | Std. Deviation | .000 | .000 | .000 | .000 | .000 | .000 | .000 | .000 |
| Total | Mean | 46000.00 | 3.25 | 3.85 | 4.10 | 3.95 | 4.10 | 3.45 | 4.35 |
| | N | 40 | 40 | 40 | 40 | 40 | 40 | 40 | 40 |
| | Std. Deviation | 17216.569 | 1.945 | 1.875 | 1.392 | 1.986 | 1.499 | 1.739 | 1.477 |

Are the clusters stable?

# Example in SPSS: The Coffee Project

1. What variables will we use for clustering?
2. Do we standardize the data?
3. Selecting the clustering method
4. How many clusters should we have?
5. What do the clusters mean? (interpreting and profiling – a lot of subjectivity)
6. Perform sensitivity checks to assess stability of the clusters

# Boating Case: Part II

# Group Work

1. How many market segments (clusters) are there? Why?

2. How would we describe the segments?

3. How would the segments inform the strategy of CreeqBoat?

**Report**

| Centroid Method | | Q20/21. Where did you buy your (...) (...)/Where are you planning to buy your new boat? | Q29. Not including any taxes, fees or accessories, approximately how much did you pay for your (...) (...) /Not including any taxes, fees or accessories, approximately how much are you planning to spend on your new boat? | Q31. How did you pay for your (...) (...) /How are you planning to pay for your future boat purchase? | Q63/Q64. Length of Boat (in Feet) | Q78/79. I normally boat alone - Which of the following statements best describes WHO you boat with/Which of the following statements best describes WHO you would boat with? | Q78/79. I boat with my spouse or significant other - Which of the following statements best describes WHO you would boat with? | Q78/79. I boat with my family, including kids - Which of the following statements best describes WHO you boat with/Which of the following | Q78/79. I boat with my friends - Which of the following statements best describes WHO you boat with/Which of the following | How did you your rience rding ting? |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Mean | 2.76 | 4.14 | 2.03 | 23.83 | 2.42 | 3. | | | 2.26 |
| | N | 4414 | 4414 | 4414 | 4414 | 4414 | 44 | | | 4414 |
| | Std. Deviation | 1.471 | 2.340 | 1.095 | 15.154 | 1.181 | 1.1 | | | .812 |
| 2 | Mean | 3.10 | 5.27 | 1.93 | 24.17 | 2.32 | 3. | | | 1.83 |
| | N | 59 | 59 | 59 | 59 | 59 | | | | 59 |
| | Std. Deviation | 1.605 | 2.935 | 1.324 | 16.402 | 1.166 | 1.3 | | | .854 |
| 3 | Mean | 2.00 | 2.67 | 1.67 | 21.33 | 2.00 | 4.33 | 3.00 | 3.33 | 1.67 |
| | N | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| | Std. Deviation | .000 | 1.528 | 1.155 | 2.309 | 1.000 | .577 | 1.732 | .577 | .577 |
| 4 | Mean | 3.81 | 1.44 | 1.25 | 16.69 | 3.63 | 3.19 | 2.31 | 2.94 | 2.31 |
| | N | 16 | 16 | 16 | 16 | 16 | 16 | 16 | 16 | 16 |
| | Std. Deviation | 1.905 | 1.263 | .683 | 5.677 | 1.500 | 1.515 | 1.250 | 1.181 | 1.078 |
| 5 | Mean | 2.80 | 2.60 | 2.00 | 32.00 | 2.20 | 4.60 | 4.80 | 4.20 | 1.40 |
| | N | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 |
| | Std. Deviation | 1.095 | 1.817 | .707 | 10.173 | 1.789 | .894 | .447 | 1.095 | .548 |
| Total | Mean | 2.76 | 4.14 | 2.03 | 23.81 | 2.42 | 3.96 | 3.81 | 3.69 | 2.25 |
| | N | 4497 | 4497 | 4497 | 4497 | 4497 | 4497 | 4497 | 4497 | 4497 |
| | Std. Deviation | 1.475 | 2.354 | 1.097 | 15.144 | 1.184 | 1.111 | 1.125 | .910 | .815 |

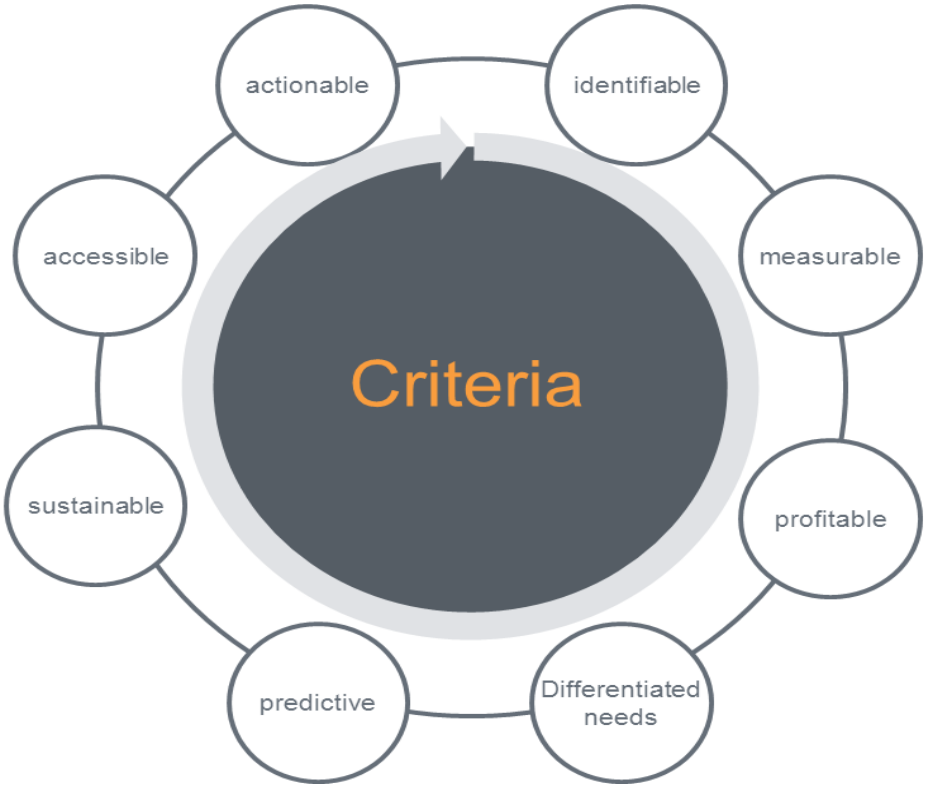
Boating results in excel

| Q81/82. Fishing - Below is a list of activities that may or may not do while boating. Using the scale provided, please indicate how often you engage in each of these activities/Below is a list of activities that you may or may not do while boating. Using | Q81/82. Swimming - Below is a list of activities that may or may not do while boating. Using the scale provided, please indicate how often you engage in each of these activities/Below is a list of activities that you may or may not do while boating. Using | Q81/82. Cruising - Below is a list of activities that may or may not do while boating. Using the scale provided, please indicate how often you engage in each of these activities/Below is a list of activities that you may or may not do while boating. Using | Q81/82. Water Sports (e.g., skiing, tubing, wakeboarding) - Below is a list of activities that may or may not do while boating. Using the scale provided, please indicate how often you engage in each of these activities/Below is a list of activities that y | Q81/82. Entertaining/socializing - Below is a list of activities that may or may not do while boating. Using the scale provided, please indicate how often you engage in each of these activities/Below is a list of activities that you may or may not do whil | Q81/82. Entertaining/rafting together - Below is a list of activities that may or may not do while boating. Using the scale provided, please indicate how often you engage in each of these activities/Below is a list of activities that you may or may not do | Q86/87. During your boating season, how many days out of the year do you typically use your boat/During your boating season, how many days out of the year do you expect to use your boat? | U1/B1/C1. Respondent's Gender | U3A/B/B3A_B_C3A_B. Past Year Household Income Before Taxes |
|---|---|---|---|---|---|---|---|---|
| 3.58 | 3.44 | 3.78 | 3.25 | 3.59 | 3.36 | 46.83 | 1.42 | 13.49 |
| 4414 | 4414 | 4414 | 4414 | 4414 | 4414 | 4414 | 4414 | 4414 |
| 1.201 | 1.050 | .942 | 1.157 | 1.029 | 1.137 | 45.373 | .494 | 4.076 |
| 2.95 | 2.80 | 3.00 | 2.71 | 3.05 | 2.88 | 40.15 | 1.42 | 11.85 |
| 59 | 59 | 59 | 59 | 59 | 59 | 59 | 59 | 59 |
| 1.419 | 1.156 | 1.339 | 1.327 | 1.209 | 1.403 | 65.977 | .498 | 5.492 |
| 1.67 | 4.33 | 4.33 | 2.33 | 3.33 | 4.33 | 56.67 | 1.33 | 13.67 |
| 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| .577 | .577 | .577 | 1.528 | 1.155 | .577 | 37.859 | .577 | 2.517 |
| 4.13 | 2.56 | 2.94 | 1.88 | 2.19 | 2.00 | 39.69 | 1.06 | 12.87 |
| 16 | 16 | 16 | 16 | 16 | 16 | 16 | 16 | 16 |
| 1.360 | 1.031 | 1.063 | 1.204 | 1.328 | 1.265 | 35.440 | .250 | 4.241 |
| 3.40 | 3.80 | 4.00 | 3.80 | 3.80 | 3.80 | 32.00 | 1.80 | 13.40 |
| 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 |
| 1.817 | 1.095 | 1.000 | 1.304 | 1.304 | 1.095 | 17.889 | .447 | 4.037 |
| 3.57 | 3.43 | 3.77 | 3.23 | 3.58 | 3.35 | 46.71 | 1.42 | 13.46 |
| 4497 | 4497 | 4497 | 4497 | 4497 | 4497 | 4497 | 4497 | 4497 |
| 1.208 | 1.055 | .954 | 1.164 | 1.037 | 1.145 | 45.641 | .494 | 4.100 |

# What Makes a “Good” Segmentation?

# What Makes a "Good" Segmentation?

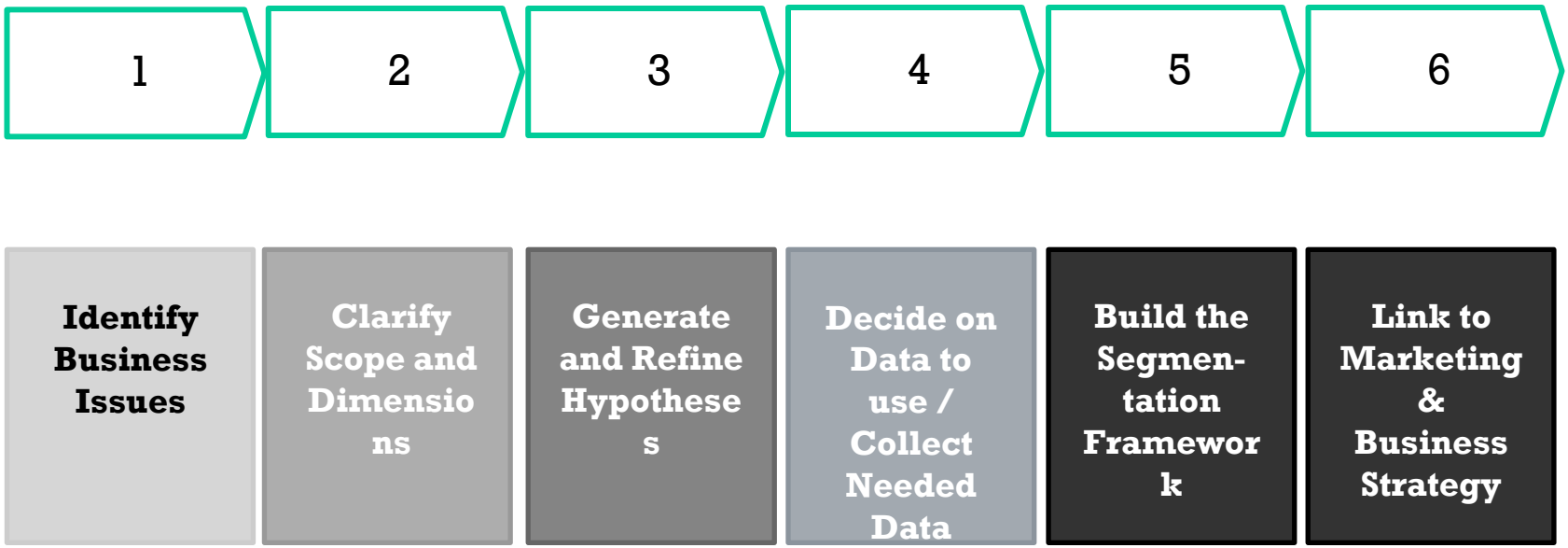
Many different evaluation systems exist.
Most deem a segmentation good if it meets these criteria



# Key Tenants on Segmentation

- No one segmentation approach will work in all situations.
- The value in segmentation does not come from the segmentation solution but from the **programs leveraging this solution.**
- Segmentation should be "**customer-in**" versus business- or product-out.
- There is both a **science and an "art"** to designing and evaluating a successful segmentation.
- Segmentation is the foundation for distinctive and **sustainable competitive advantage**.

# Segmentation Methodology – A(nother) Process



# Step 6 – Link to Business Strategy

- Developing the appropriate strategy to go after identified target segments is key for every business

# Step 6 – Link to Business Strategy

- Just building the segmentation is only half the battle

| 1 Product/Offer | 2 Pricing/ASS | 3 Marketing/Brand | 4 Dealer/Channel |
|---|---|---|---|
| Breadth & Depth of Offer | Pricing & Financing | Marketing Communication | Channel Store/Environment |
| Product Requirements | Customer Service & Support | Events | Inventory Management |
| Product Naming & Architecture | Maintenance & Repair | Licencing | Dealer Network |
| | | | Dealer Relationship Building |

There is A LOT of JUDGMENT in ANALYTICS:

Your involvmenent is CRUCIAL

# Next class: Purchase Drivers and Discrimination

- Who are most likely to click on an ad?
- Who are likely to respond to a direct mail campaign? What distinguishes those who responded to previous direct mail compared to those who do not?
- How are satisfied customers different from dissatisfied customers in terms of their demographics and attitudes towards your products' characteristics?
- Who are likely to default on a loan?
- To whom should we offer a particular promotion?
- Which transaction is most likely a fraud?
- Which applicants are most likely to fit in our organization and succeed?
- Which drug development project should we mainly invest in?

INSEAD

The Business School
for the World®