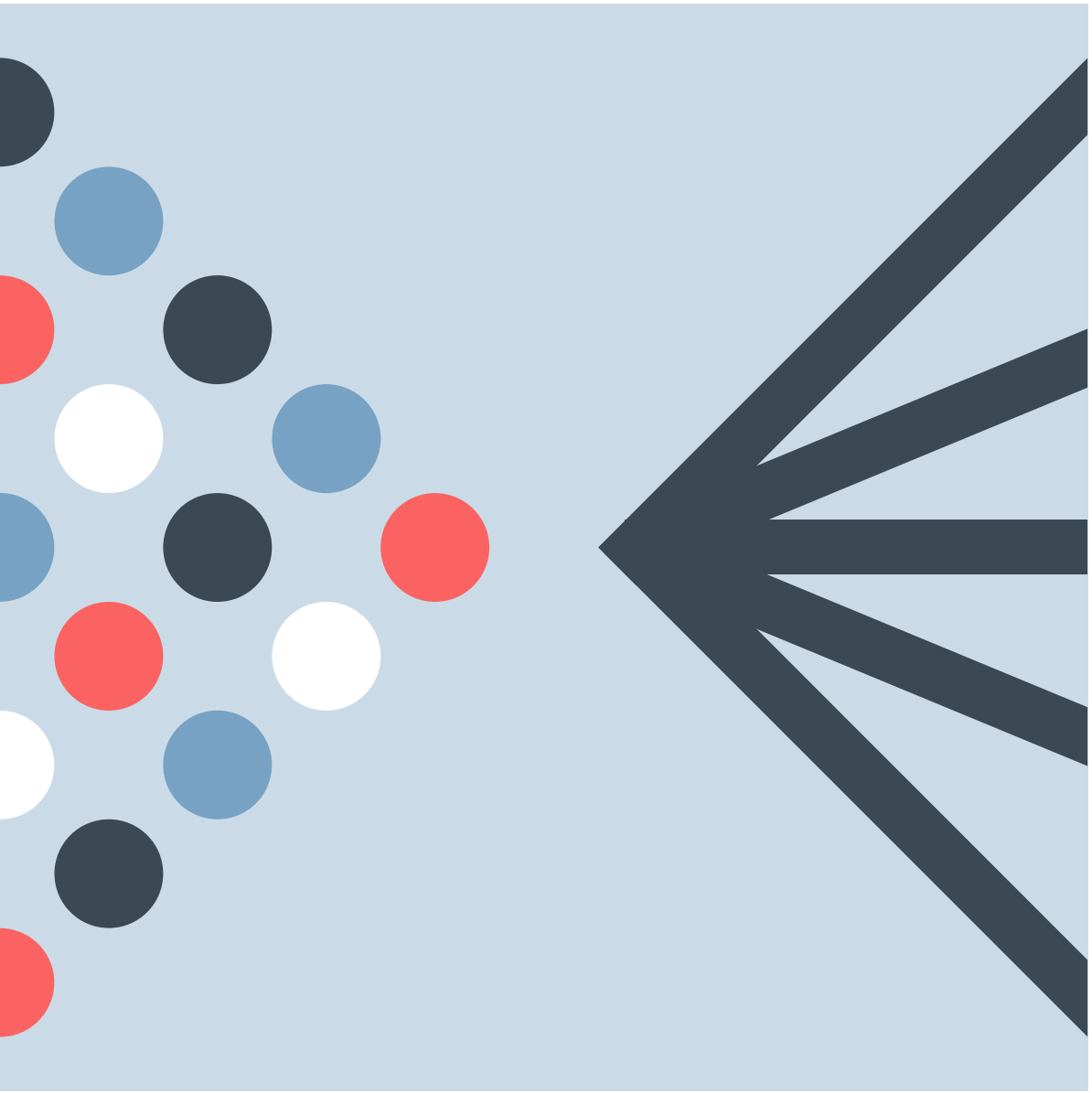


SESSION 4

SEGMENTATION & CLUSTERING
UNSUPERVISED MACHINE LEARNING

SPRING 2025

UCSD MSBA – MGTA 495

PLAN FOR TODAY

I. Segmentation

a) Why we believe it's so fundamental to marketing

b) Its role in traditional marketing frameworks

c) Examples!

II. Clustering

a) Unsupervised Machine Learning Methods

b) Latent Class Statistical Models

c) Where/how this fits into an overall segmentation project

Segmentation

WHAT IS SEGMENTATION?

> “Aggregating prospective buyers into groups with common needs and who respond similarly to a marketing action”

Business School 101: https://www.youtube.com/watch?v=lrJ1cNIfmsk

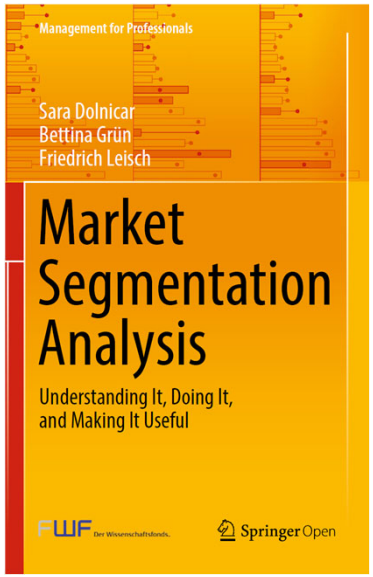
> “Market segmentation is a decision-making tool for the marketing manager in the crucial task of selecting a target market for a given product and designing an appropriate marketing mix”

Market Segmentation Analysis by Dolnicar, Grun, and Leisch. Springer, 2018.

> “The process of dividing a broad consumer or business market, normally consisting of existing and potential customers, into sub-groups of consumers based on shared characteristics”

Wikipedia: https://en.wikipedia.org/wiki/Market_segmentation

SEGMENTATION IN THE CONTEXT OF MARKETING



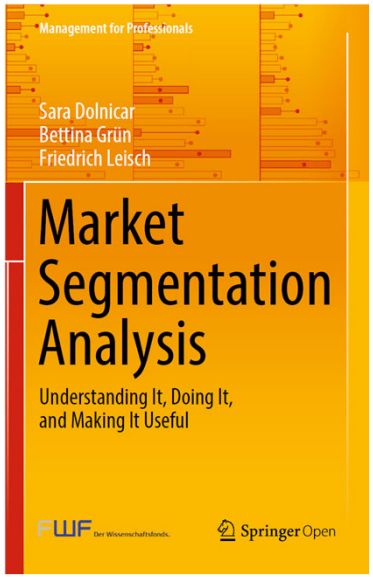
The *purpose of marketing* is to **match** the genuine needs and desires of consumers with the offers of suppliers particularly suited to satisfy those needs and desires

Marketing planning is a logical sequence and a series of activities leading to the setting of marketing objectives and the formulation of plans to achieving them; consists of two components:

1. The **strategic** plan outlines the **long-term** direction of an organization
   - Identifies consumer needs and desires, strengths and weaknesses internal to the organization, and external opportunities and threats the organization may face
   - Two key decisions: which consumers to focus on and which image of the organization to create in the market
2. The **tactical** marketing plan **translates** the long-term strategic plan into detailed instructions for **short-term** marketing action
   - the development and modification of the product in view of needs and desires of the target segment(s),
   - the determination of the price in view of cost, competition, and the willingness to pay of the target segment(s),
   - the selection of the most suitable distribution channels to reach the target segment(s),
   - the communication and promotion of the offer in a way that is most appealing to the target segment(s)

Source: *Market Segmentation Analysis* by Dolnicar, Grun, and Leisch. Springer, 2018.

SEGMENTATION IN THE CONTEXT OF MARKETING



The *purpose of marketing* is to **match** the genuine needs and desires of consumers with the offers of suppliers particularly suited to satisfy those needs and desires

Marketing planning is a logical sequence and a series of activities leading to the setting of marketing objectives and the formulation of plans to achieving them; consists of two components:

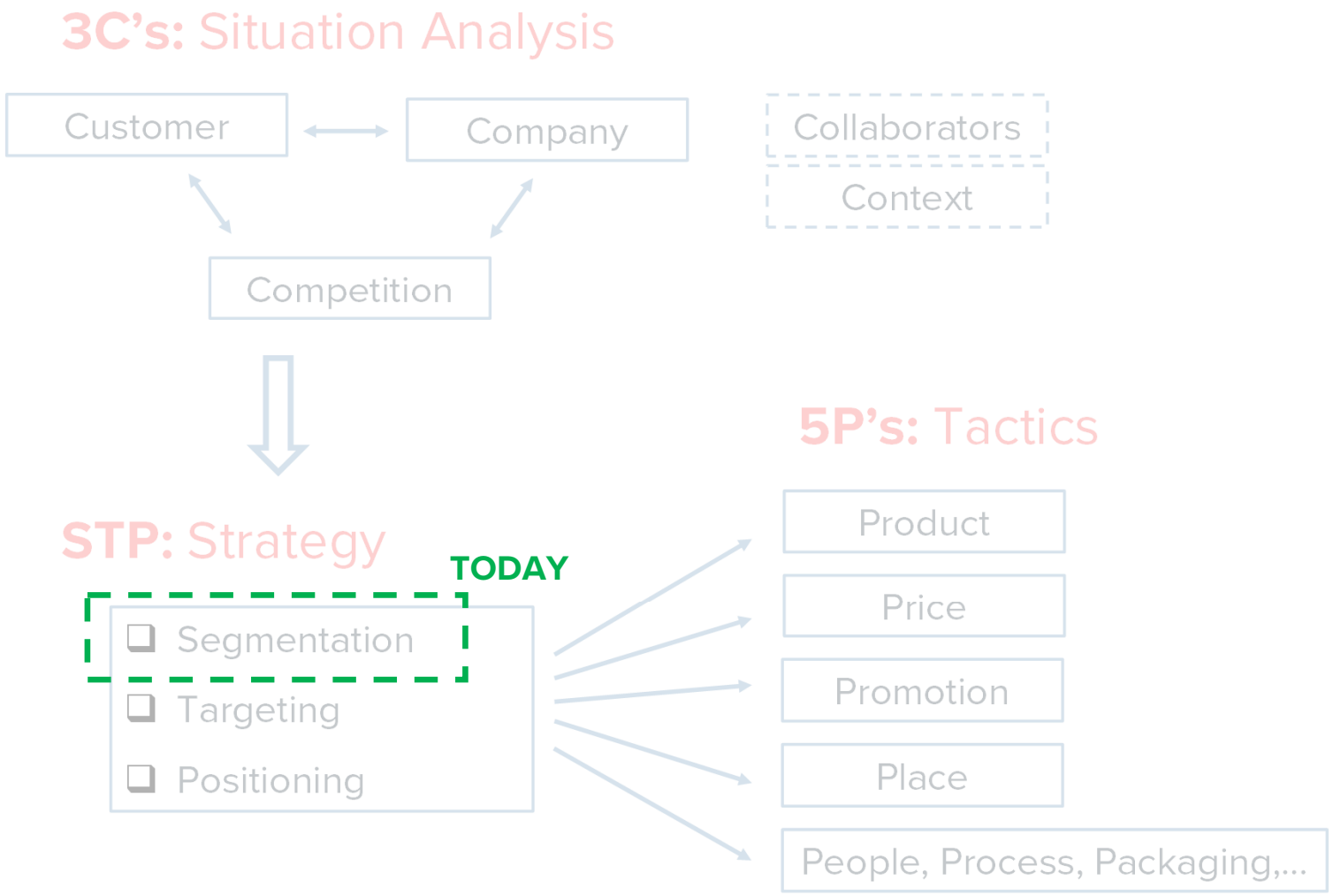
1. The **strategic** plan outlines the long-term direction of an organization — **3 C's**
   - Identifies consumer needs and desires, strengths and weaknesses internal to the organization, and external opportunities and threats the organization may face
   - Two key decisions: which consumers to focus on and which image of the organization to create in the market

   Segmentation + Targeting **Positioning**
2. The **tactical** marketing plan translates the long-term strategic plan into detailed instructions for short-term marketing action
   - the development and modification of the product in view of needs and desires of the target segment,
   - the determination of the price in view of cost, competition, and the willingness to pay of the target segment,
   - the selection of the most suitable distribution channels to reach the target segment,
   - the communication and promotion of the offer in a way that is most appealing to the target segment

   4 P's

Source: *Market Segmentation Analysis* by Dolnicar, Grun, and Leisch. Springer, 2018.

CLASSIC MARKETING STRATEGY FRAMEWORK



ADDED OPTIONAL READINGS

Wind & Bell Article

- "…no product or service appeals to **all** consumers and even those who purchase the same product may do so for **diverse reasons**."
- "this chapter is based on the premise that segmentation is the firm's response to a **fundamental** market feature – **heterogeneity**."
- "Effective marketing and business strategy therefore **requires a segmentation** of the market into homogeneous segments, an understanding of the needs and wants of these segments, the design of products and services that meet those needs, and development of marketing strategies, to effectively reach the target segments."
- "Thus focusing on segments is **at the core** of organizations' efforts to become customer-driven; it is also the **key** to effective resource allocations and deployment."

Qualtrics Blog post

- Checklists of:
  - Benefits to segmentation
  - Categories of segmentation variables
  - Types of segmentation
  - Examples
  - Characteristics of effective segments
  - Common segmentation errors

Sources:
Market Segmentation by Wind & Bell, 2007, Ch. 11.
Qualtrics: What is Segmentation available online at https://www.qualtrics.com/experience-management/brand/what-is-market-segmentation/

TL,DR VERSION

Identify groups of prospective customers who have:

- **Similar** attributes **within** the group
- **Different** attributes **between** the groups

for the purpose of effecting different marketing action toward different groups

WHY SEGMENT?

No one wants room-temperature tea!

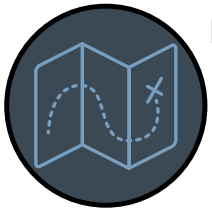
Don’t make the average product

Instead, cater to the differences among consumers

SOURCES OF CUSTOMER HETEROGENEITY

Demographics

Age, race, income, gender, language



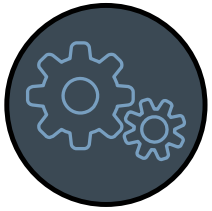
Info. & Experience

Does / doesn't understand the benefits of the product

Has purchased before

Attitudes

Positive or negative feels about social media



Psychographics

Lifestyles, values, extroversion, orientation to art / status / religion / family

Needs

Have kids = bigger car

Watch videos = bigger device screen or better resolution

Geographics

Urban vs rural

Specific zip code

Distance to retail location

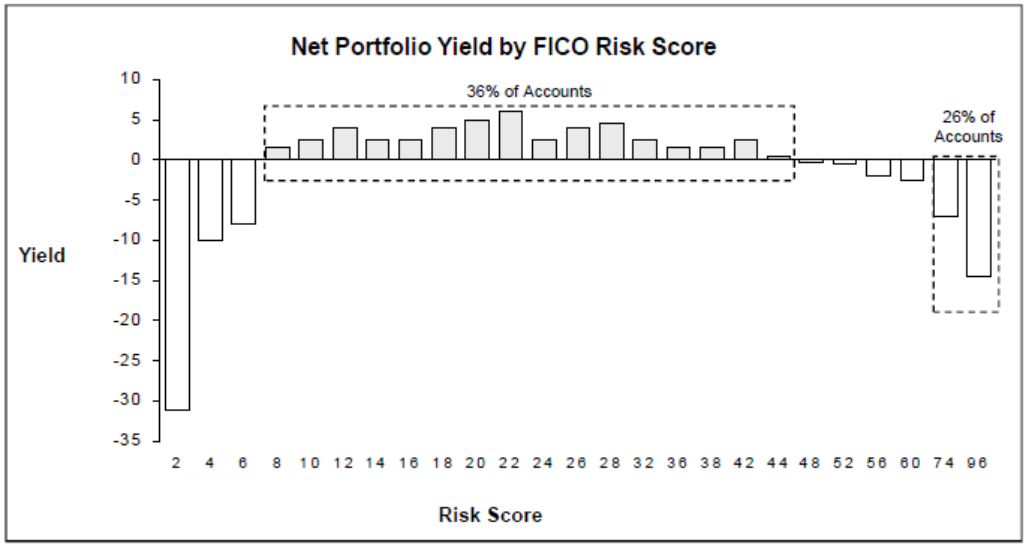
Country

WHAT MAKES FOR A GOOD SEGMENTATION SOLUTION

- ✓ **Relevant** – to firm's objectives
- ✓ **Substantial** – big enough to be profitable
- ✓ **Identifiable / Measurable** – consumers can be assigned to segments
- ✓ **Accessible** – must be able to reach them
- ✓ **Actionable** – distinct segments have differential responses to firm's marketing actions

GOOD VS BAD EXAMPLES

Good



Credit card company uses risk score to understand a meaningful dimension of its customers

Not so Good

New-style restaurant poorly blends customer groups together

A COMMENT ON DEMOGRAPHICS

Easy to collect

Easy to assign people into segments

But...

Usefulness depends on industry or product category

Prince Charles

- Male
- Born in 1948
- Raised in the UK
- Married twice
- Lives in a castle
- Wealthy & famous

British Royalty

Ozzy Osbourne

- Male
- Born in 1948
- Raised in the UK
- Married twice
- Lives in a castle
- Wealthy & famous

Former lead singer of Black Sabbath

A COMMENT ON DEMOGRAPHICS

Easy to collect

Easy to assign people into segments

But…

Usefulness depends on industry or product category

Aaron Paul

- White Male
- Born in 1979
- Raised in Idaho
- Big Break as TV Supporting Actor
- Wealthy & Famous

Chris Pratt

- White Male
- Born in 1979
- Raised in Minnesota
- Big Break as TV Supporting Actor
- Wealthy & Famous

A COMMENT ON NEEDS, ATTITUDES, PSYCHOGRAPHICS

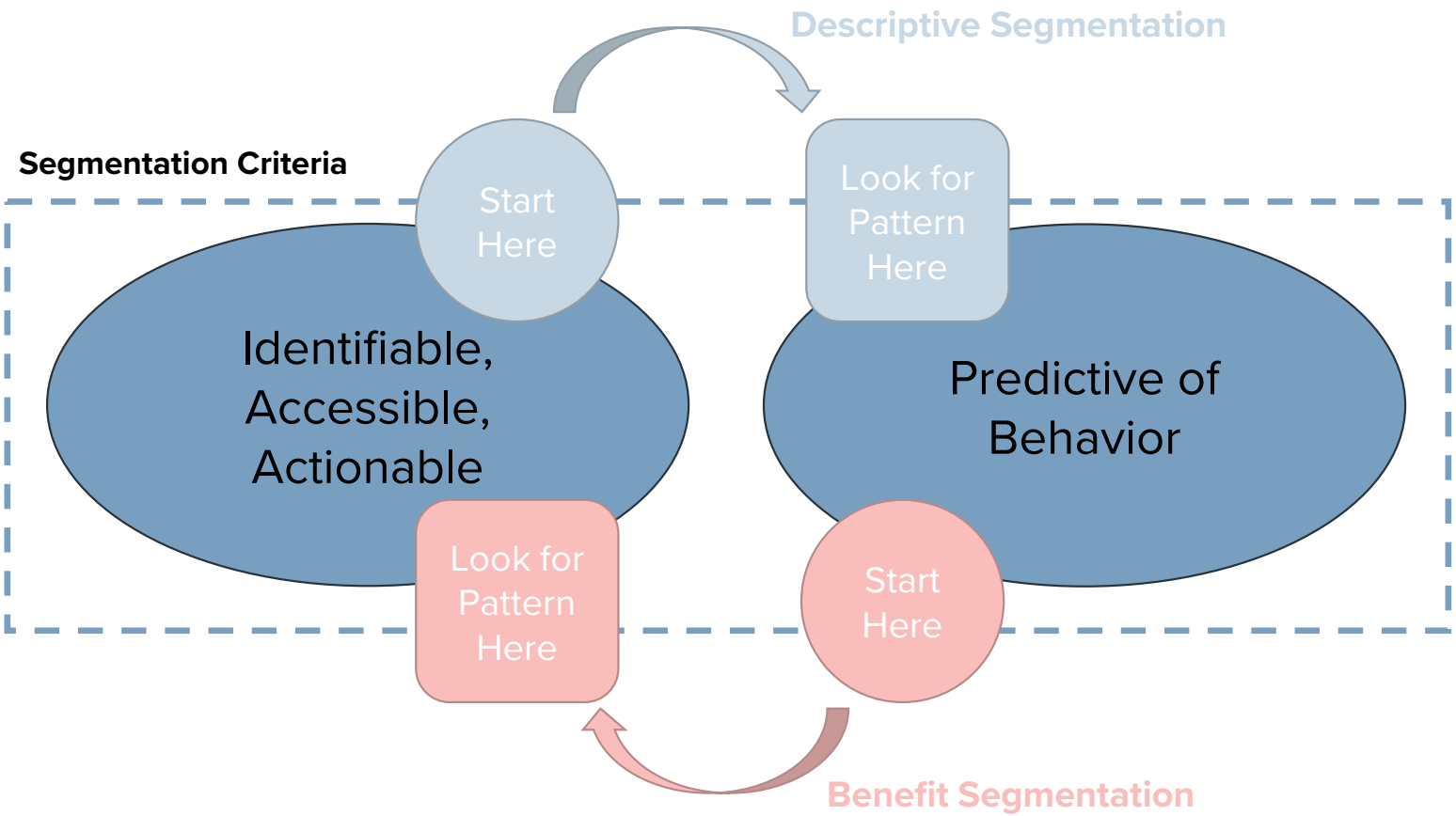
Typically leads to better segmentation solutions

But…

Can be costly to collect

Can lead to segments that are difficult to identify

HOW TO BE BOTH: A GOOD SOLUTION & IDENTIFIABLE



A COMMENT ON SUBSTANTIAL

Substantial = large enough to be worth the effort

Low-cost businesses can compete in markets with many segments

WHO SEGMENTS?

Every large business segments its markets

Sometimes there are multiple segmentations within a firm

Walmart
Save money. Live better.

Price Sensitivity

Income	Low	Med	High
Low	☑	☑	☑
Med			☑
High			☑

HOW DO FIRMS USE SEGMENTS?

Create customer response profiles to embody segments

(we'll see an example with Firefox in a bit)

Segments can drive numerous marketing decisions

- **Pricing** – price discrimination, discounts
- **Promotions** – advertising content, placement
- **Product** – design, timing of release
- **Placement** – geographic or retailer-specific
- **Brand extensions** – product portfolio

EXAMPLE: PRICING BY DISNEY

SoCal Resident 3-Day, Mon-Thurs Ticket with Admission to 1 Park Per Day
1 Ages 3+
Subtotal: **$225.00**
Edit

3-Day Ticket with Admission to 1 Park Per Day
1 Ages 10+
Subtotal: **$390.00**
Edit

February 2024

S	M	T	W	T	F	S
				1	2 $154	3 $169
4 $154	5 $119	6	7	8 $119	9 $169	10 $184
11 $169	12 $134	13 $119	14 $119	15 $134	16 $169	17 $194
18 $194	19 $194	20 $169	21 $169	22 $169	23 $169	24 $184
25 $184	26 $134	27 $119	28 $119	29 $119		

EXAMPLE: PROMOTION BY MCDONALD'S

Southwest Chicken Salad

fresh, gourmet

hearty, interesting

cilantro, lime

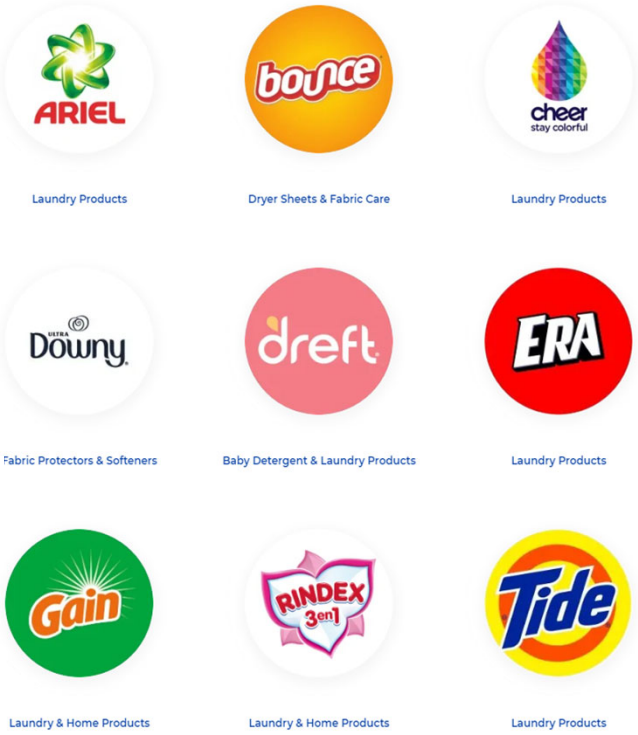
variety, new

Source: https://www.youtube.com/watch?v=tKOITmV0NfU

EXAMPLE: PRODUCTS BY P&G

Laundry detergents [edit]

- Ariel laundry detergent
- Bold laundry detergent
- Bonux laundry detergent
- Cheer laundry detergent
- Daz detergent
- Downy fabric softener[6]
- Dreft laundry detergent
- Era laundry detergent
- Fairy Non-Bio laundry detergent
- Gain laundry detergent, scent booster
- Ola laundry soap
- PMC laundry soap
- Tide laundry detergent
- Lenor fabric softener and cleaning products
- Mif (Russian: Миф) laundry detergent



Source: https://en.wikipedia.org/wiki/List_of_Procter_%26_Gamble_brands
https://us.pg.com/brands/#Fabric-Care

EXAMPLE: PRODUCT RELEASE TIMING AND PRICING BY TESLA

Roadster

Released in 2005

$100-120k

Model S

Released in 2012

$80k

Model 3

Released in 2017

$60k

EXAMPLE: PLACEMENT BY CKE RESTAURANTS

EXAMPLE: PLACEMENT BY QUIDEL

Quickvue brand home pregnancy tests

- Baby on packaging
- Near ovulation kits or in baby aisle
- 25 tests for $2.51 each

- Single woman on packaging
- Near birth control
- 50 tests for $1.67 each

EXAMPLE: BRAND EXTENSIONS BY APPLE AND COKE

Free Engraving

AirPods Max

$549.00

New

Beats Studio Pro Wireless Headphones — Black

$349.99

A CONSUMER-INSIGHTS FIRM'S APPROACH

- What do we already know?
- What can we learn from stakeholders?
- What can we learn from consumers?

"Outside-In Thinking"

- What are consumers' needs, desires, feelings and behaviors in our category?
- How might we want to profile consumers?

"Survey-Based"

- How do we select the variables to cluster on?

"Segmentation Bases"

- What method do we use for clustering?
- How do we pick the right solution?

- How big is the category and each segment?
- What are the defining characteristics of each segment?

"Segmentation and CLV Drivers"

- How do we decide which segment(s) to focus on?

"Key to Segmentation is Sacrifice"

- How do we bring the segments to life?
- How can we learn more about the segments?
- How can we implement the segmentation?

"More than Nice to Know"

A CONSUMER-INSIGHTS FIRM'S EXAMPLES

Enterprise Cohort

34%

Leading Edge Titans

24%

Stick to Tried & Tested

15%

Flying Blind

15%

Ready. Set. Grow.

12%

Growing Pains

Heavier E-Commerce

SMB Cohort

17%

Smooth Sailors

22%

Follow the Data

14%

New Brands on the Block

15%

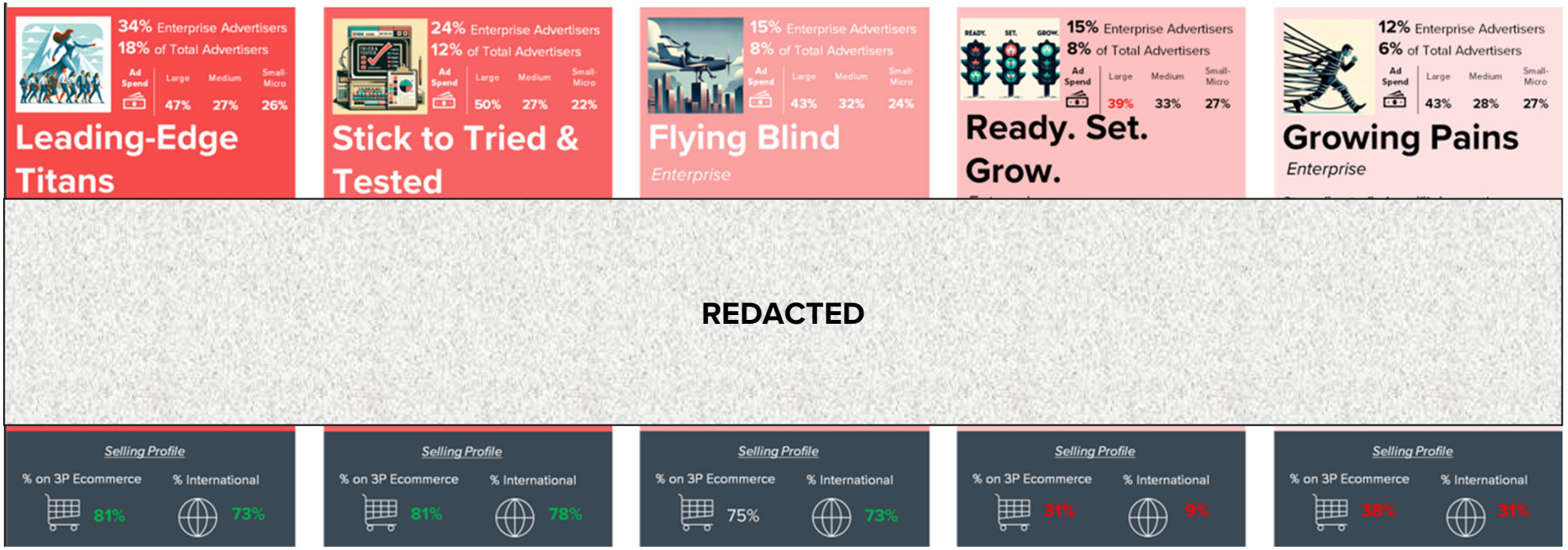
Shaky Footing

32%

Competing for the Spotlight

Heavier E-Commerce

A CONSUMER-INSIGHTS FIRM'S EXAMPLES



EXAMPLE: FIREFOX

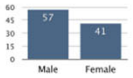
Firefox User Types
ENTHUSIASTS

I enthusiastically learn about and adopt new technology. I enjoy solving my own technology problems.

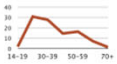
Attributes

- Enthusiastic about new technology
- Solve own technology problems
- Want to feel in control of their browsing experience
- Likely customization in the browser
- Self-confidence with new technology
- Streaming media and data synced among devices
- Vast majority are online often
- More likely to have attended college or above a college degree
- Skew younger

Population



Gender



Age

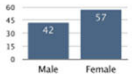
Firefox User Types
BUSY BEES

I lead a busy life and I expect technology just to work. I'm not interested in the technical details behind how the technology in my life works.

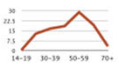
Attributes

- Utility of technology: The Internet like an appliance
- Very busy with activities in life and do not prioritize time spent on the internet for its own sake
- Not interested in the details behind technology
- Impatient with technology difficulties
- Some data integration across devices
- Skew Female
- Skew Older

Population



Gender



Age

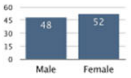
Firefox User Types
MIDDLE MANAGERS

I'm comfortable and confident with technology, esp. troubleshooting it. Technology is an important part of my day-to-day life, but I carefully evaluate new technology before I adopt it.

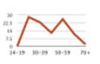
Attributes

- Comfortable and confident with technology, but not enthusiasts
- Work, school, life fully integrate technology
- Patient with solving their own technology woes
- Adopt latest tools and features, but thoughtfully
- Likely use mobile and have integrated data
- Multi-tasking and customization
- Less educational attainment than enthusiasts or wizards
- Online often

Population



Gender



Age

Source: https://blog.mozilla.org/ux/2013/08/firefox-user-types-in-north-america/

EXAMPLE: FIREFOX

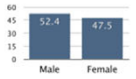
Firefox User Types
STALWARTS

I prefer to stick technology I believe works for me even if it might be outdated. I'm reluctant to upgrade most technology I use because I believe 'if it ain't broke, don't fix it.'

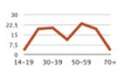
Attributes

- Change-averse
- Avoid upgrading tools or technologies unless there is a reason to do
- Prefers known to unknown
- Time spent online is a discrete activity
- Likely has older technology
- Limited mobile or smart phone usage
- Wide age distribution

Population



Gender



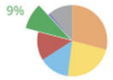
Age

Firefox User Types
EVERGREENS

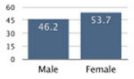
I could probably live much of my life without technology. I feel some reluctance about using some technology because I fear I may make mistakes I cannot correct.

Attributes

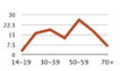
- Discomfort about new tools and customization
- Can likely live life without the Internet
- Learning technology is like vocational training
- Use new technology in tandem with older tools
- Internet is not the focal point of their life
- Rely on others for technology cues
- Less educational attainment
- Skew older



Population



Gender



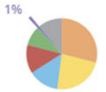
Age

Firefox User Types
WIZARDS

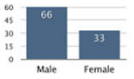
I enjoy writing software for myself and others. Technology is my life.

Attributes

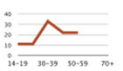
- Software developer or engineer
- Highly accurate mental model of the Internet
- Confident about installing, using, and troubleshooting technology
- Enjoys creating technology
- Generally high level of satisfaction with Firefox
- High level of expertise / tech savviness
- Higher educational attainment
- Skew male



Population



Gender



Age

EXAMPLE: URBAN OUTFITTERS

Website:

"We stock our stores with what we love, calling on our – and our customer's – interest in contemporary art, music, and fashion…"

"We offer a lifestyle-specific shopping experience for the **educated, urban-minded individual in the 18 to 30 year-old range**…"

Earnings Call:

"Our customer is from traditional homes and **advantage**, but this offers them the benefit of **rebellion**…

Our customer is exposed to new ideas and philosophies. This can be a real involvement and work, or it could be just talk.

Irreverence and **concern** can live together. Often products sell well that represents the concerns they have but also can speak to their irreverence.

Our customer leads a **pretty cloistered existence** although they deem themselves worldly… they believe that they're right and that everything that's happening to them is what's happening everywhere.

Our customer is highly involved in **mating and dating behavior**… one of the primary drives for their spending behavior… they work hard to postpone adulthood…"

EXAMPLE: ANTHROPOLOGIE

Website:

"a lifestyle brand that catered to **creative, educated, and affluent 20–45 year-old women**...

Our customer is a creative-minded woman, who wants to look like herself, not the masses. She has a sense of adventure about what she wears and, although fashion is important to her, she is too busy enjoying life to be governed by the latest trends."

Earnings Call:

"We don't think of her in terms of age or affluence or even location. We try to think of her in her **life stage** and her sensibilities.

She's **recently wed**. She's **settling down**. She's very interested **less in the mating rituals** and actually has been trying and building and creating an environment she wants to live in for herself and family.

She loves art and culture... **Clothing and her living environment** to her are her canvases in which she's able to **control her life**, whereas workplace and those things around her, she may not control.

We believe in many ways that's what's touched her and connected her to Anthropologie and why she is more loyal to us than most retailers."

EXAMPLE: BIC

Is it sexist to market different products to different genders?

★☆☆☆☆ **No good for man hands**, 16 Aug 2012

By [redacted]

This review is from: **BIC For Her Amber Medium Ballpoint Pen (Box of 12) - Black (Office Product)**

I bought this pen (in error, evidently) to write my reports of each day's tree felling activities in my job as a lumberjack. It is no good. It slips from between my calloused, gnarly fingers like a gossamer thread gently descending to earth between two giant redwood trunks.

★★★★★ **Revolutionary article - must buy!**, 20 Aug 2012

By [redacted]

This review is from: **BIC For Her Amber Medium Ballpoint Pen (Box of 12) - Black (Office Product)**

This pen is great. I bought it for all my female friends and relatives. It enabled them, finally, to write things (although they may not yet know to do so on paper; but you can only expect so much, really). I thought they were just a bit slow.

My mother, a hard-working woman who raised twelve kids single-handedly whilst doing all the ironing (as nature intended), was furtively abashed by her illiteracy. Long would she gaze upon her husband and sons' scrawlings and would dedicate five minutes a day (which she really should have spent making sandwiches) to pray that one day she would be granted the ability to create such scribbles of her own. She's still a little slow on the uptake, but this product has definitely helped start the ball rolling. We tried to give her men's pens but she used to rip the cartridges out and drink the ink. Typical woman.

Anyway, it's good that BIC are finally doing something to aid the plight of women. Hopefully a range of 'for her' paperclips is on the horizon - my wife has an awful time keeping her recipes together.

16

★★★★☆ **Only missing the paper**

Reviewed in the United States on April 12, 2013

Well at last pens for us ladies to use... now all we need is "for her" paper and I can finally learn to write!

Bic Amazon

Unsupervised Learning: Cluster Analysis

UNSUPERVISED LEARNING

Supervised learning

- Vector of outputs or responses $Y = (Y_1, \dots, Y_n)$
- Matrix of predictors, features, or inputs $\mathbf{X} = (X_1, \dots, X_J)$
- where each column $X_j = (X_{1j}, \dots X_{nj})$ or each row $X_i' = (X_{i1}, \dots, X_{iJ})$

We can evaluate the algorithm f that generates $\hat{y} = f(x)$ by some loss function $L(y, \hat{y})$ for example squared error loss $L(y, \hat{y}) = (y - \hat{y})^2$

Unsupervised learning

- No responses Y, simply a matrix of features $\boldsymbol{X}$

"It is difficult to ascertain the validity of inferences drawn from the output of most unsupervised learning algorithms. One must resort to heuristic arguments not only for motivating the algorithms, …but also for judgments as to the quality of the results."

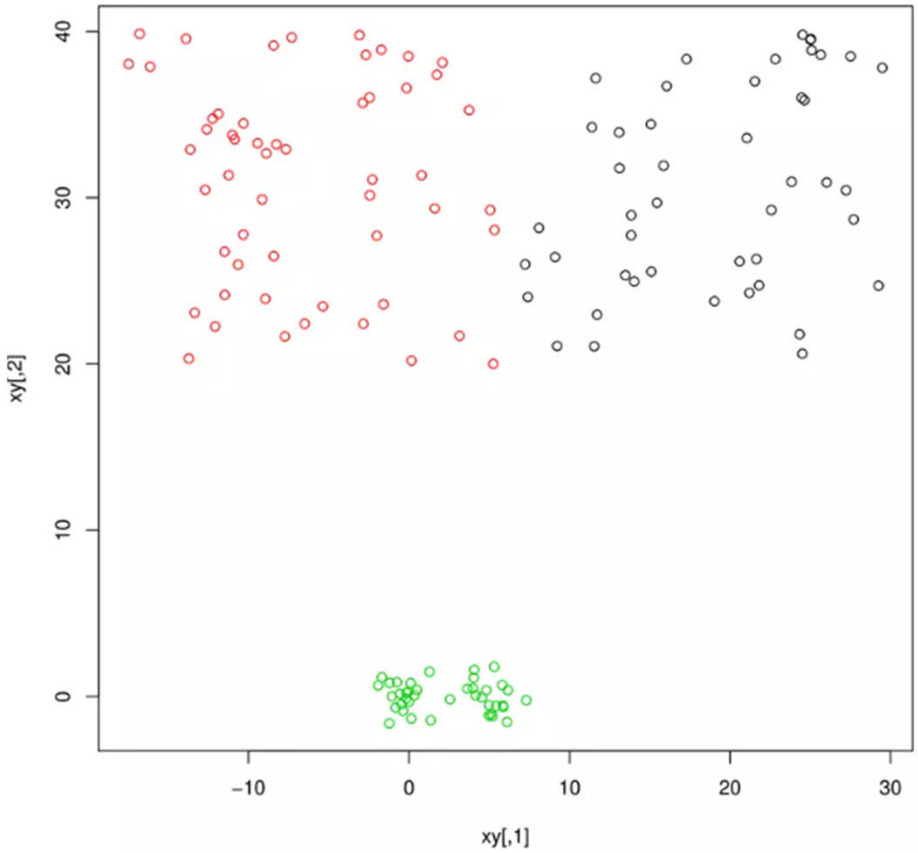
–ESL Book

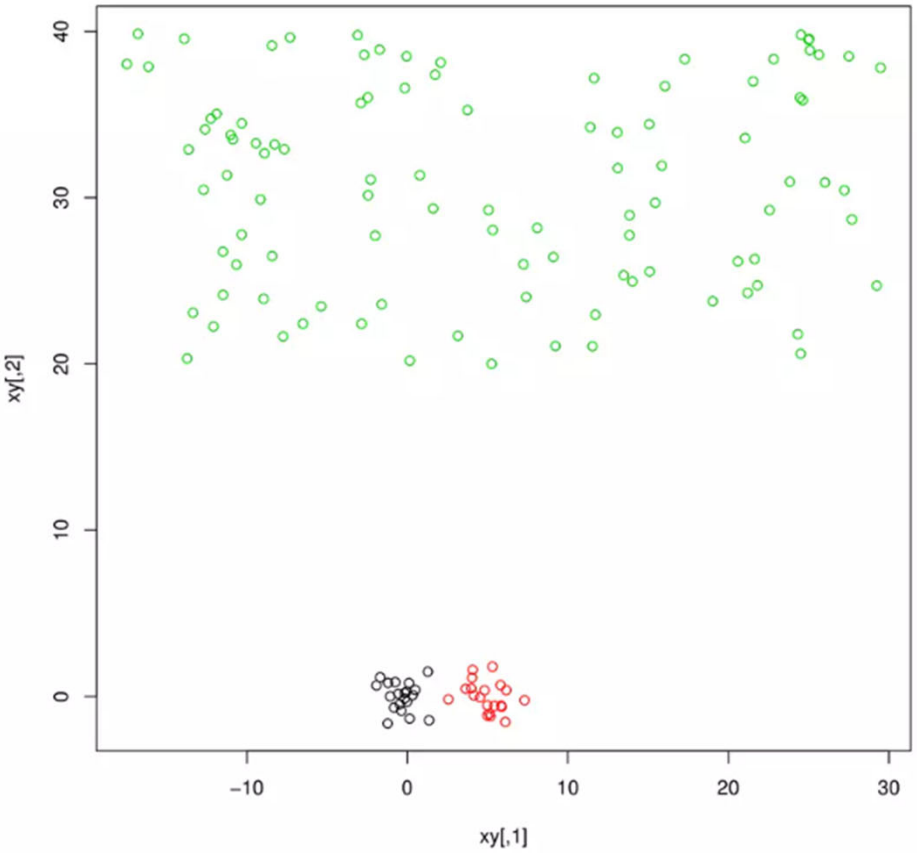
NO ONE-SIZE-FITS-ALL CRITERIA

Can evaluate cluster solutions by:

- Between-cluster separation
- Within-cluster homogeneity (low distances)
- Within-cluster homogeneous distributional shape
- Good representation of data by centroids
- Little loss of information from original distance between objects
- Clusters are regions of high density without within-cluster gaps
- Uniform cluster sizes
- Stability

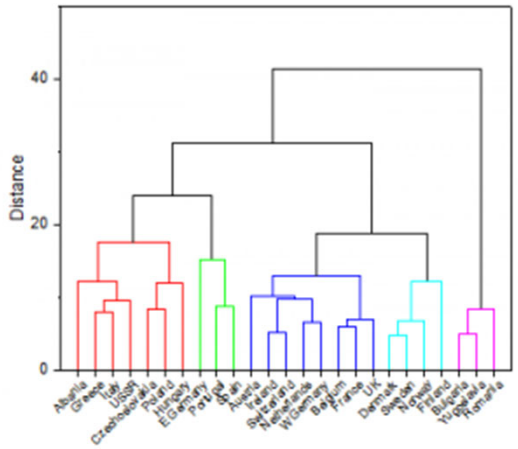
WHICH CLUSTER SOLUTION IS BETTER?



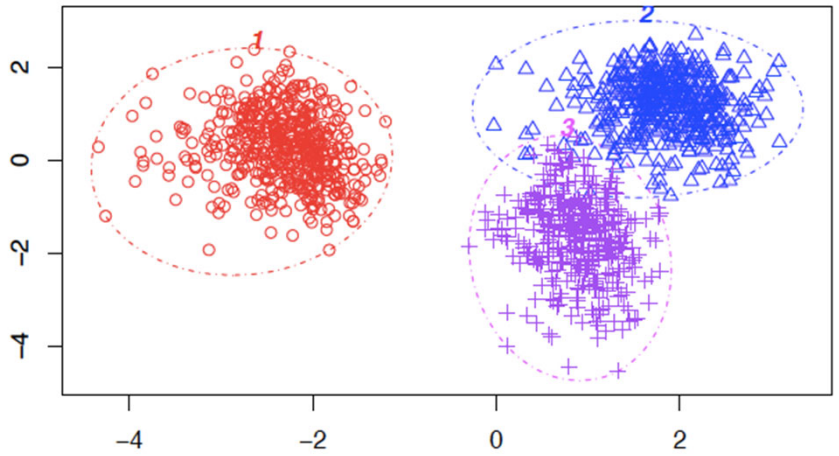


WHAT ANALYTIC METHODS ARE USED TO SEGMENT? CLUSTER ANALYSIS

Hierarchical / tree-based algorithms
(AGNES, DIANA)



Distance-based algorithms
(K-Means, PAM)



Model-based approaches
(latent-class, mixture-modeling)

$$f(\mathbf{y}_i|\mathbf{z}_i) = \sum_{x=1}^{K} P(x|\mathbf{z}_i)\, f(\mathbf{y}_i|x,\mathbf{z}_i) = \sum_{x=1}^{K} P(x|\mathbf{z}_i) \prod_{h=1}^{H} f(\mathbf{y}_{ih}|x,\mathbf{z}_i)$$

K-MEANS

K-MEANS: PARTITION A DATASET TO MINIMIZE WITHIN-CLUSTER VARIATION

K-means clustering is a simple and elegant approach for **partitioning a dataset into K distinct, non-overlapping clusters**

Let $C_1, \ldots, C_K$ denote sets containing the indices of the rows in each cluster, satisfying:

1. $C_1 \cup C_2 \cup \ldots \cup C_K = \{1, \ldots, n\}$ or that each rows belongs to at least one cluster
2. $C_k \cap C_{k'} = \emptyset$ for all $k \neq k'$ or that no row belongs to more than one cluster

The idea behind "good" clustering is to make **within-cluster variation** – a measure $W(C_k)$ of the amount by which observations differ from each other – as small as possible

$$\min_{C_1,\ldots,C_K} \left\{ \sum_{k=1}^{K} W(C_k) \right\}$$

K-MEANS: USE EUCLIDEAN DISTANCE

In order to solve the minimization problem on the prior slide, we need to define a measure of within-cluster variation. The most common is **Euclidean distance**:

- Respondent i is assigned to cluster C_k – there are K total clusters
- The cluster bases are the J variables x_j
- The middle of cluster k is the point $(\bar{x}_{k1}, \bar{x}_{k2}, \ldots, \bar{x}_{kJ})$

$$W(C_k) = \sqrt{\sum_{i \in C_k} \sum_{j=1}^{J} (x_{ij} - \bar{x}_{kj})^2}$$

Then sum up across people in cluster k

Sum up the distances across all J variables

How far is x_j from $\bar{x}_j$ in cluster k for person i

K-MEANS: GLOBAL OPTIMUM COMPLICATED, LOCAL OPTIMUM EASY

We wish to optimize this problem by minimizing within-cluster variation

$$W^* = \min_{C_1,\ldots,C_K} \left\{ \sum_{k=1}^{K} W(C_k) \right\}$$

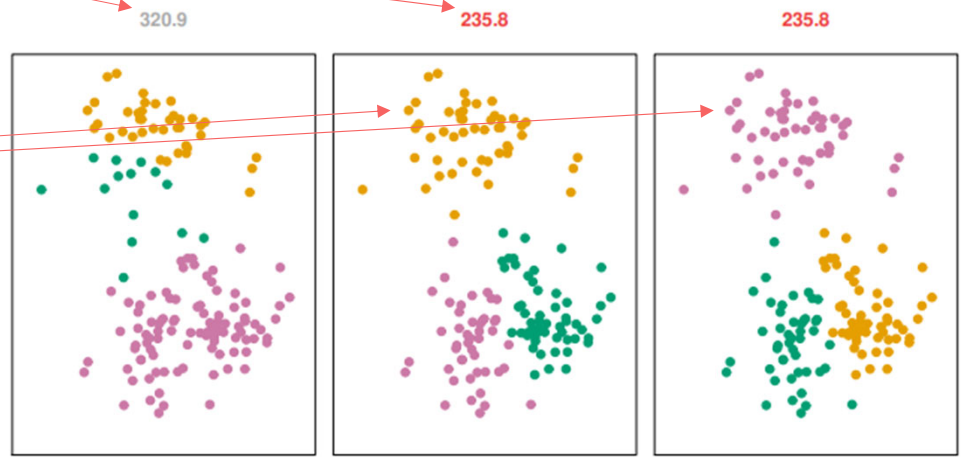
This is a very difficult problem to solve precisely, since there are almost K^n ways to partition n observations into K clusters.

But there is a **simple algorithm** that finds a local optimum:

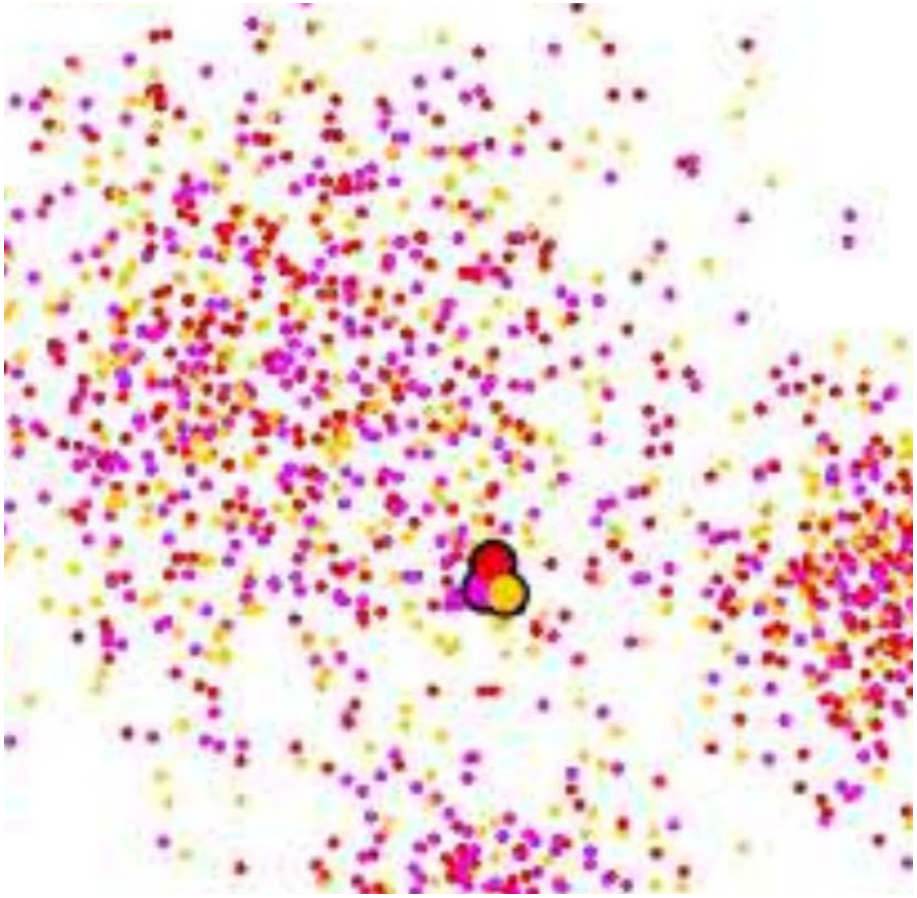
1. Randomly assign each cluster's centroid
2. Iterate the following until cluster assignments stop changing:
   - Assign observations to closest centroid
   - Compute each new centroid location as the vector of J means for the observations currently assigned to that cluster

K-MEANS: THINGS TO WATCH OUT FOR

1. We can (mostly) overcome the local-optimum issue (ie, attempt to find the global optimum) by **running the algorithm several times** from different starting positions
   - In R, use `kmeans(data, centers=k, nstarts=D)` for D different "runs"
2. The better results have the lower within-cluster variation (caveat: **increasing K always decreases** $W(C_k)$)
3. Cluster labels are **arbitrary**
4. "Distance" is unitless in R, so need to pick (or **scale** variables into) sensible units. In R, use the `scale()` function

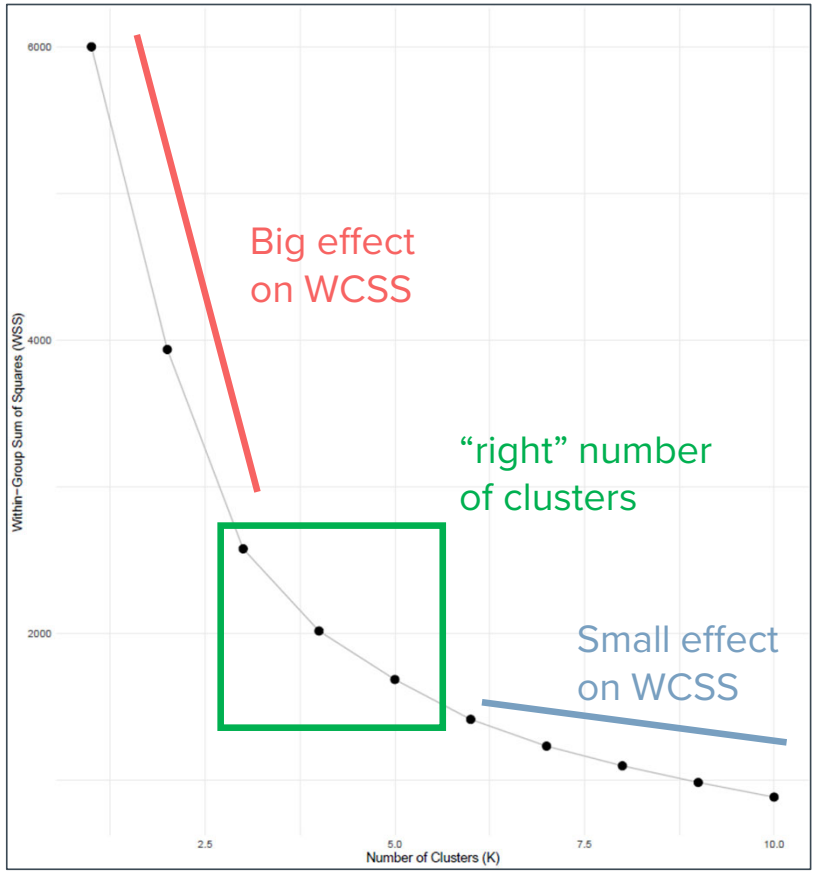


K-MEANS: ANIMATED ALGORITHM



Source: https://www.youtube.com/shorts/XCsoWZU9oN8

K-MEANS: PICKING K WITH SCREE PLOTS




```r
# let's try k=1, k=2, ..., k=10

# we'll create a vector named 'res' to store our results
res ← vector(length=10)

# we loop over k=1 through k=10
for(i in 1:10) {
    # run k means
    out ← kmeans(scl, centers=i, nstart=25)

    # grab the WSS value, store it in the i'th position of res
    res[i] ← out$tot.withinss
}
```

Intuition :

- adding a cluster when there "really is one" leads to a shard decrease in W, while
- adding an "extra" cluster simply splits an existing cluster in half, leading to a minor decrease in W.

A **Gap Statistic** can be used to determine the kink (if you don't trust your eyes to identify it on the plot).

K-MEANS: COMPARING SOLUTIONS WITH SILHOUETTE SCORES

For point i

- a_i is the average distance between point i and other points in i's cluster ("cohesion")
- b_i is the average distance to all points in the next-closest cluster ("separation")

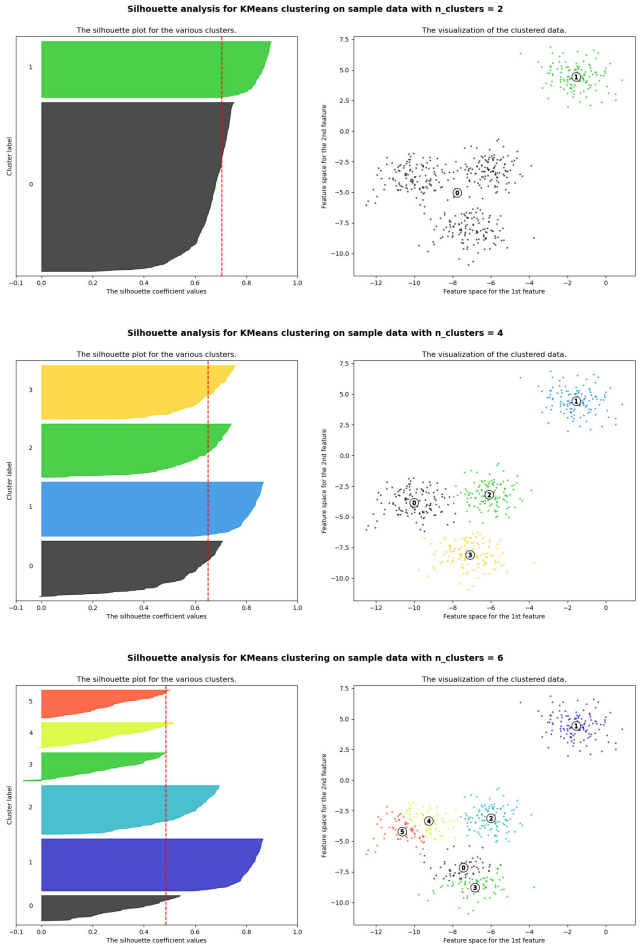
The silhouette score for point i is:

$$s_i = \frac{b_i - a_i}{\max\{a_i, b_i\}}$$

The silhouette score for the entire cluster solutions is the pointwise average

$$s = \frac{1}{n}\sum_{i=1}^{n} s_i \quad \text{where} \quad -1 \leq s \leq 1$$

Great score is > 0.5, good score > 0.2



K-MEDIODS / PAM

K-MEDIODS & GOWER'S DISTANCE

What if your variables are categorical or ordinal, rather than continuous?

- K-means represents a cluster with each cluster's center $\bar{x}_k = (\bar{x}_{k1}, \bar{x}_{k2}, \ldots \bar{x}_{kJ})$
- **K-mediods** uses the *center-most data point* of each cluster to represent that cluster

Gower's distance provides a metric for mixed variable types:

- Continuous variables are assigned Euclidean distance, as a fraction of the maximum range for that variable
- Ordinal variables are assigned a fraction of their range
- Categorical variables are assigned a distance of 0 for matches, 1 for non-matches
- Then an average is taken over the distances for the set of variables

HIERARCHICAL / TREE

TREE-BASED METHODS

AGNES: Agglomerative Nesting

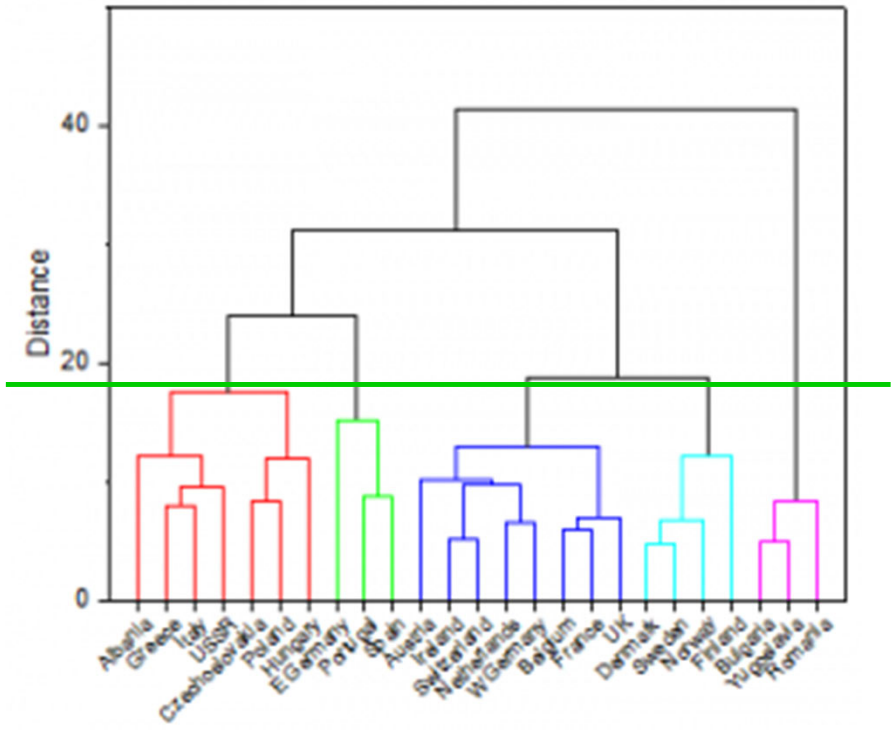
- Builds the tree from the bottom up

DIANA: Divisive Analysis

- Builds the tree from the top down

SPSS' **two-step** clustering

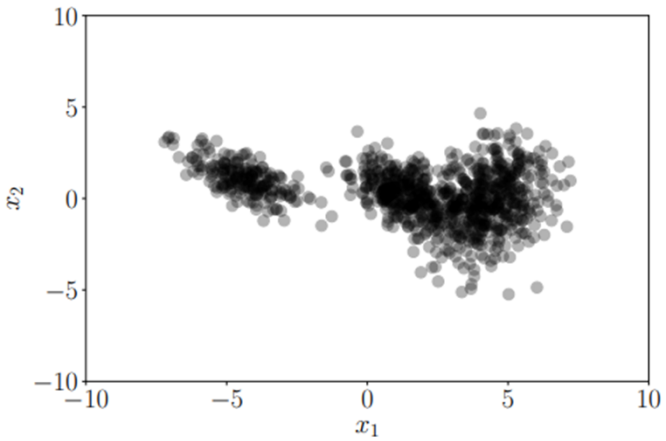
- Run K-means with big K
- Then run AGNES on the big K clusters instead of the individual data points



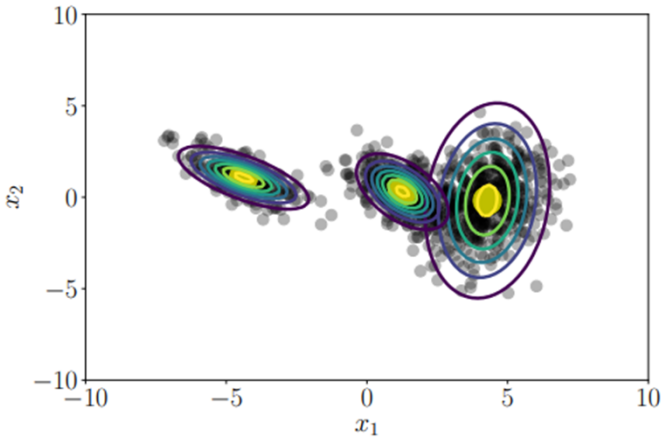
DENSITY MIXTURE MODELS

GAUSSIAN MIXTURE MODELS

Data



Fitted Model



Likelihood $p(X|\theta) = \prod_{i=1}^{n} p(x_i|\theta)$ where $p(x_i|\theta) = \sum_{k=1}^{K} \pi_k \, MVN(x_i|\mu_k, \Sigma_k)$

Source: https://mml-book.github.io/book/mml-book.pdf

LATENT-CLASS MODELS

LATENT-CLASS MNL

Recall the probability that consumer i chooses product j from the MNL model:

$$P_i(j) = \frac{e^{x_j'\beta}}{\sum_{k=1}^{J} e^{x_k'\beta}}$$

This modeling framework measure the average sensitivity to x through β

Let's now extend this so that there's two segments ($s = \{1,2\}$), each with their own sensitivities

$$P_i(j|s=1) = \frac{e^{x_j'\beta_1}}{\sum_{k=1}^{J} e^{x_k'\beta_1}} \quad \text{and} \quad P_i(j|s=2) = \frac{e^{x_j'\beta_2}}{\sum_{k=1}^{J} e^{x_k'\beta_2}}$$

And we'll denote the probability that consumer i belongs to segment s as π_s, specifically

$$\pi_1 = \frac{e^{\lambda}}{1+e^{\lambda}} \quad \text{and} \quad \pi_2 = 1 - \pi_1$$

LATENT-CLASS MNL

Now, instead of estimating just the one vector (β), we have two vectors (β_1 and β_2) and $S-1$ membership probability parameters λ_s

The unconditional choice probability is just the weighted average:

$$P_i(j) = \sum_{s=1}^{S} \pi_s \times P_i(j|s)$$

Under the standard MNL, the likelihood was:

$$L_n(\beta) = \prod_{i=1}^{n} \prod_{j=1}^{J} P_i(j)^{\delta_{ij}}$$

By comparison, the Latent-Class MNL likelihood is:

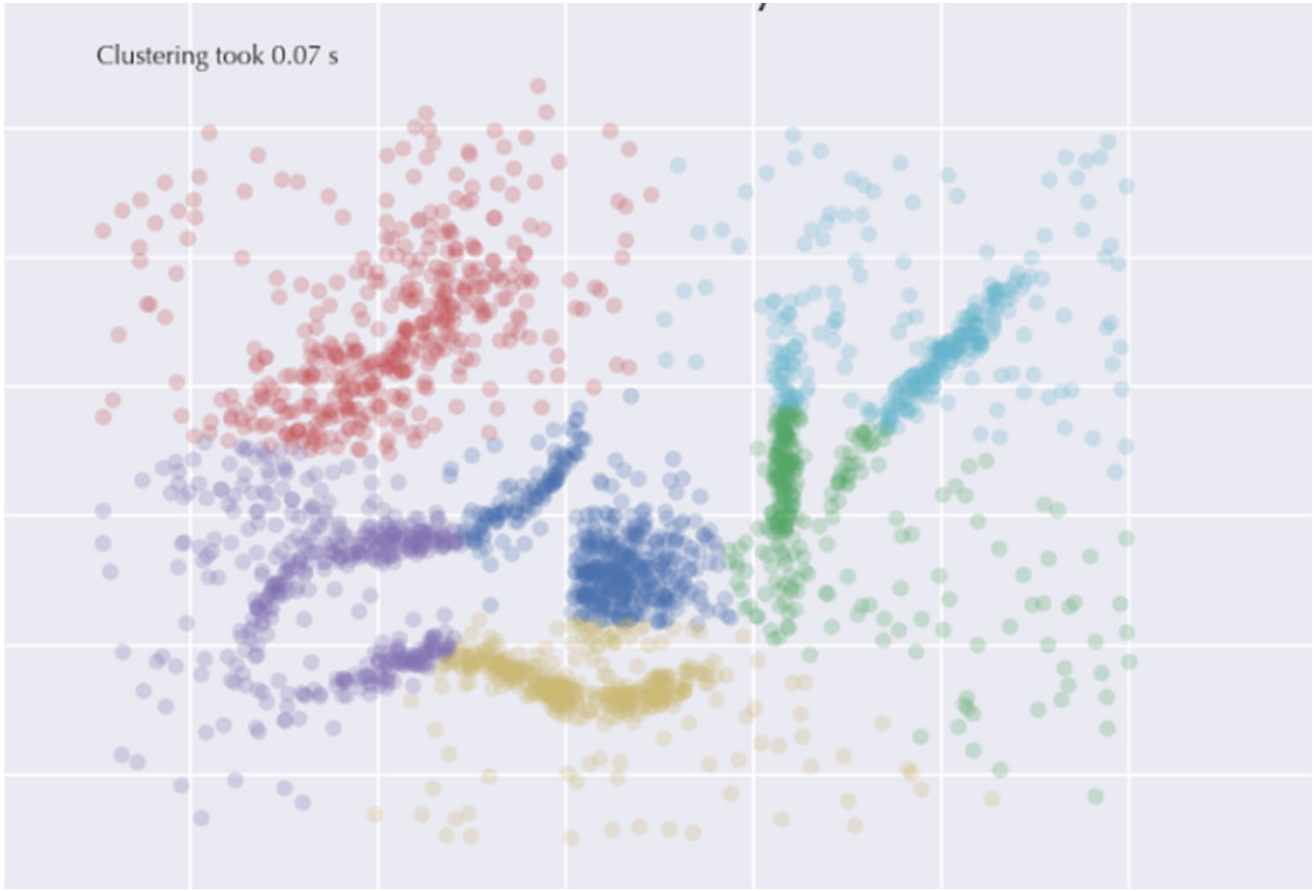
$$L_n(\beta, \lambda) = \prod_{i=1}^{n} \left(\sum\nolimits_{s=1}^{S} \pi_s \times \prod_{j=1}^{J} P_i(j)^{\delta_{ij}} \right)$$

where $\beta = (\beta_1, \beta_2, \dots, \beta_S)$ and $\lambda = (\lambda_1, \lambda_2, \dots, \lambda_{S-1})$

Sources: *Kamakura & Russel (1989) A Probabilistic Choice Model for Market Segmentation and Elasticity Structure* available online at https://journals.sagepub.com/doi/abs/10.1177/002224378902600401

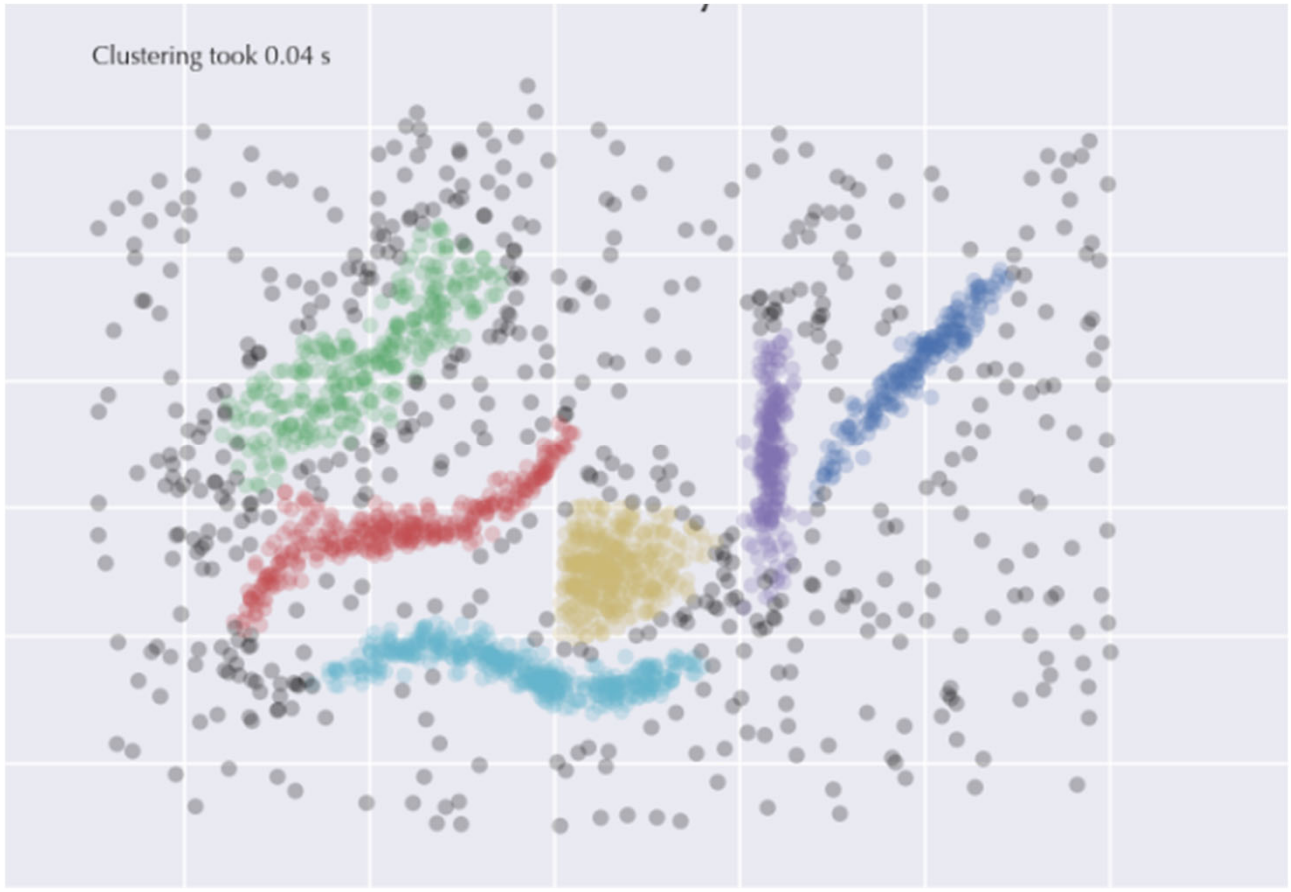
CUTTING EDGE: HDBSCAN

K-MEANS DOESN'T WORK WELL WHEN DATA ARE NON-SPHERICAL

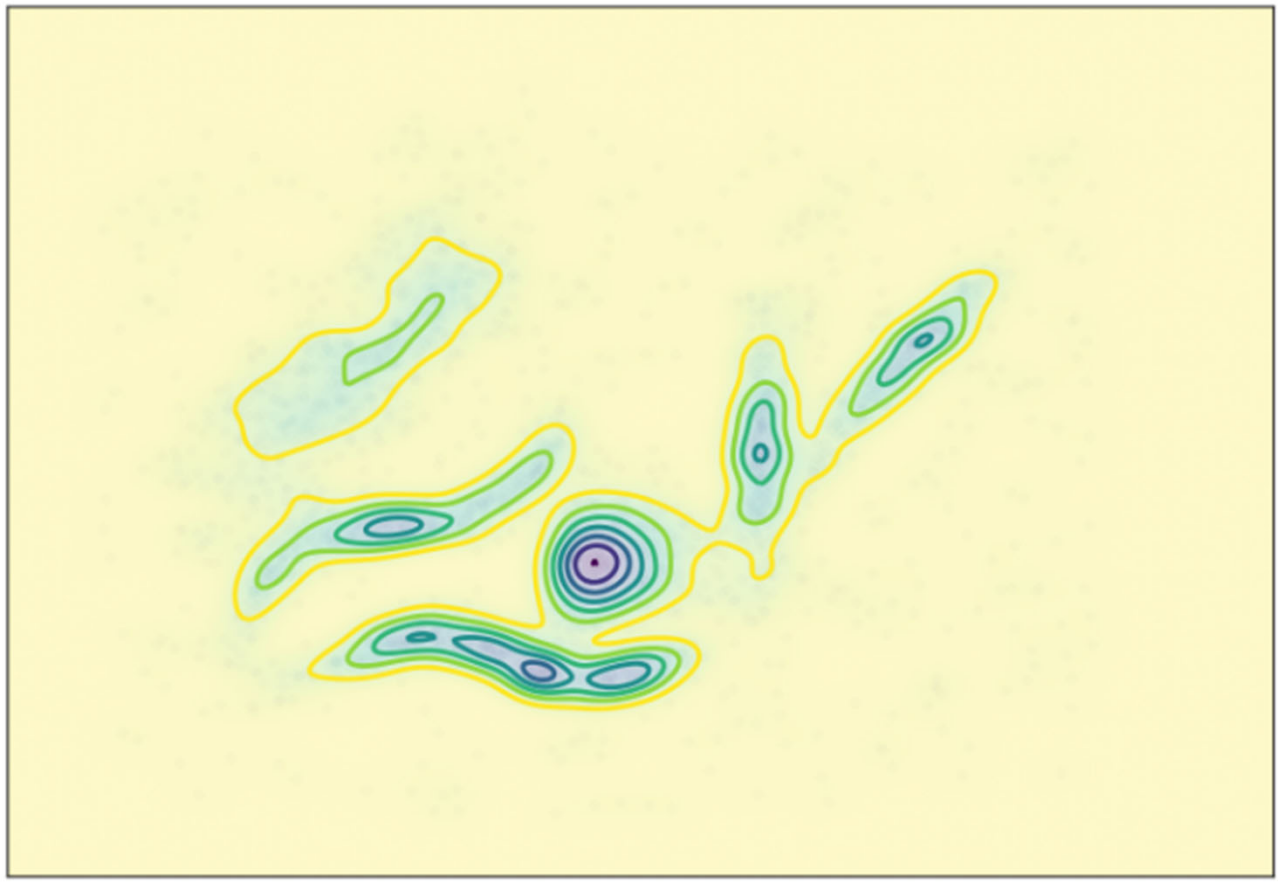


Source: John Healy and Leland McInnis, "Fast Density Based Clustering" available online at *https://www.youtube.com/watch?v=dGsxd67lFiU*

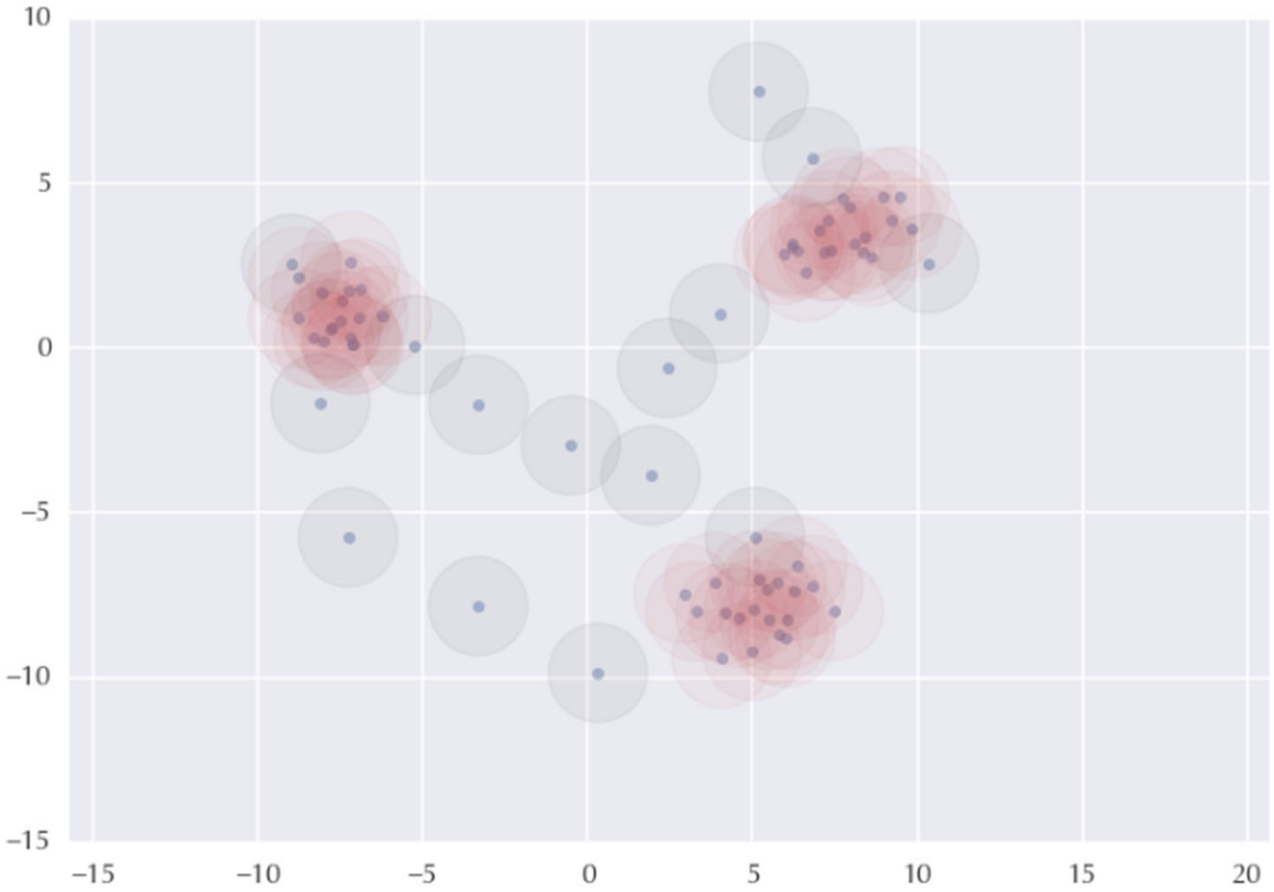
WHAT DO WE MEAN BY "CLUSTER"



WHAT DO WE MEAN BY “CLUSTER”



WHAT DO WE MEAN BY "CLUSTER"



WHAT DO WE MEAN BY "CLUSTER"

To be a cluster:

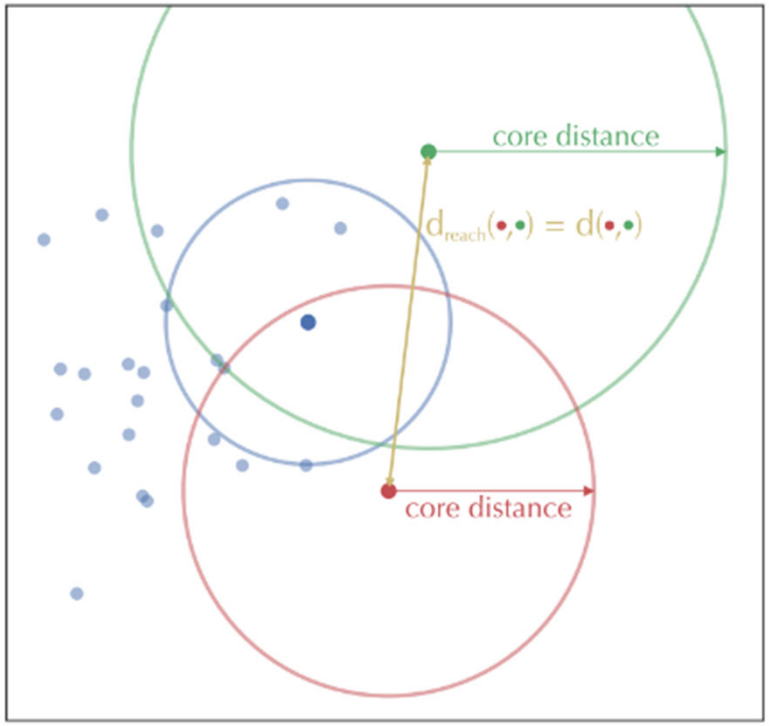
- Neet to be close in Euclidean Space

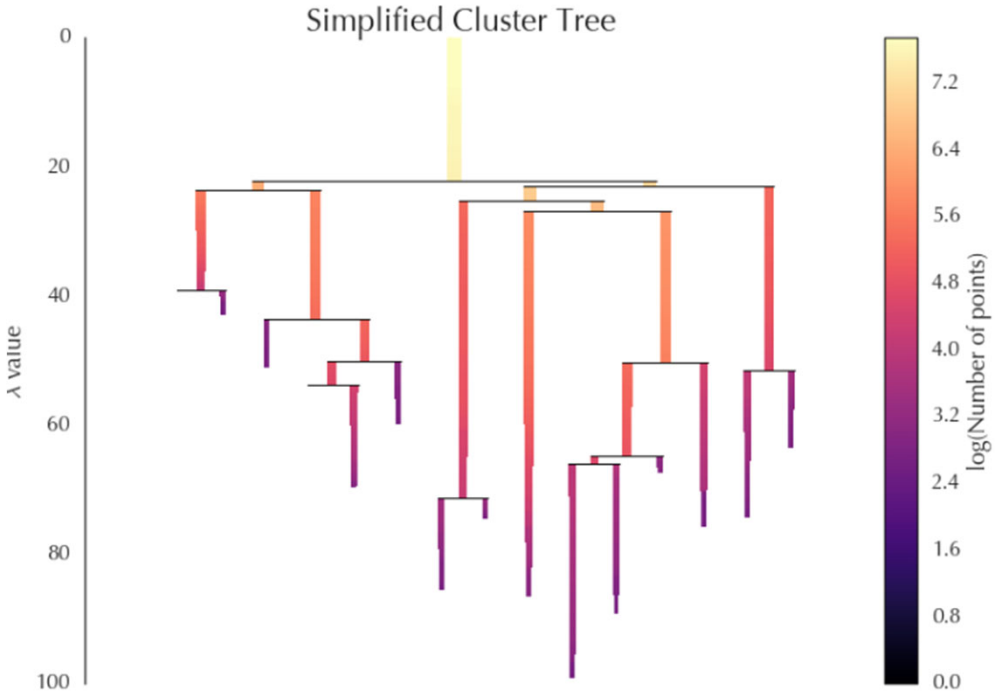
And

- Need to be Dense

WHAT DO WE MEAN BY "CLUSTER"

Mutual reachability distance





PROFILING & TYPING

PROFILING AND TYPING

Profiling: Summarizing clusters on all variables (basis variables used in the cluster analysis and additional profiling variables)

Typing: Assigning a new data point to one of the existing clusters

						Percentages by Cluster							
						Total	1	2	3	4	5	6	7
						7,018	684	1,991	1,165	628	1,327	814	409
Bas	Orde	Variable	Question	Response	Valu	100	1	2	1	9	1	1	6
	1	hs1	Hidden: to select gender for quota.	Female	1	46%	50%	47%	40%	48%	41%	51%	57%
	1	hs1	Hidden: to select gender for quota.	Male	2	53%	50%	53%	58%	52%	58%	49%	42%
	1	hs1	Hidden: to select gender for quota.	Other ((Transgender) or (Do not identify a	3	0%	0%	0%	1%	0%	0%	0%	0%
	2	h_age_qta	QUOTA dummy variable to hold key age groups for quot	18-24	1	13%	0%	6%	63%	7%	0%	2%	1%
	2	h_age_qta	QUOTA dummy variable to hold key age groups for quot	25-34	2	27%	0%	42%	37%	28%	27%	9%	2%
	2	h_age_qta	QUOTA dummy variable to hold key age groups for quot	35-44	3	24%	0%	40%	0%	23%	35%	34%	8%
	2	h_age_qta	QUOTA dummy variable to hold key age groups for quot	45-54	4	18%	0%	12%	0%	42%	21%	43%	23%
	2	h_age_qta	QUOTA dummy variable to hold key age groups for quot	55-64	5	13%	70%	0%	0%	0%	14%	10%	46%
	2	h_age_qta	QUOTA dummy variable to hold key age groups for quot	65-70	6	5%	30%	0%	0%	0%	3%	1%	20%
x	3	c9	What is your current marital status?	Now Married	1	50%	87%	91%	0%	88%	0%	64%	0%
x	3	c9	What is your current marital status?	Divorced	2	8%	0%	4%	0%	0%	0%	9%	100%
x	3	c9	What is your current marital status?	Separated	3	3%	4%	4%	1%	7%	0%	5%	0%
x	3	c9	What is your current marital status?	Widowed	4	2%	9%	1%	0%	5%	0%	2%	0%
x	3	c9	What is your current marital status?	Never married	5	38%	0%	0%	99%	0%	100%	19%	0%
x	4	children_age_grou	Children in Household	Parents - Kids <11	1	50%	90%	0%	71%	100%	78%	0%	92%
x	4	children_age_grou	Children in Household	Parents - Teens 12-17	2	37%	10%	100%	21%	0%	22%	0%	3%
x	4	children_age_grou	Children in Household	Not Parents	3	13%	0%	0%	8%	0%	0%	100%	4%
	5	income_global	What is the total annual combined income, before taxes	Low	1	17%	15%	6%	21%	15%	27%	11%	40%
	5	income_global	What is the total annual combined income, before taxes	Lower-Mid	2	28%	26%	20%	37%	27%	34%	24%	32%
	5	income_global	What is the total annual combined income, before taxes	Upper-Mid	3	19%	17%	23%	17%	16%	17%	19%	12%
	5	income_global	What is the total annual combined income, before taxes	Lower-High	4	20%	21%	27%	14%	19%	13%	23%	10%
	5	income_global	What is the total annual combined income, before taxes	Upper-High	5	17%	21%	23%	10%	23%	9%	22%	7%

ONE-SLIDE RECAP

- Good segmentations:
  - yield similarities within segments, differences between segments
  - are relevant, substantial, identifiable, accessible, and actionable
- Usually a trade-off for basis variables:
  - ease of collection vs meaningful segments
- Cluster or Latent-Class analyses often used to determine segments
  - There's no "right answer"
  - It can be difficult to "type" new customers or survey respondents into existing segments
  - Most of the work is on the front end (data prep) and back end (profiling segments)

COMING UP

Homework 3

- Due Wednesday, May 28

Next Time:

- Supervised Machine Learning and Variable Importance Measures
- Key Drivers Analysis

Pre-class work:

- Watch the Sawtooth webinar on Key Drivers Analysis
- Read up on LMG (ie, Shapley values for the linear regression model)
- Read up on tree-based methods (random forest and XGBoost)