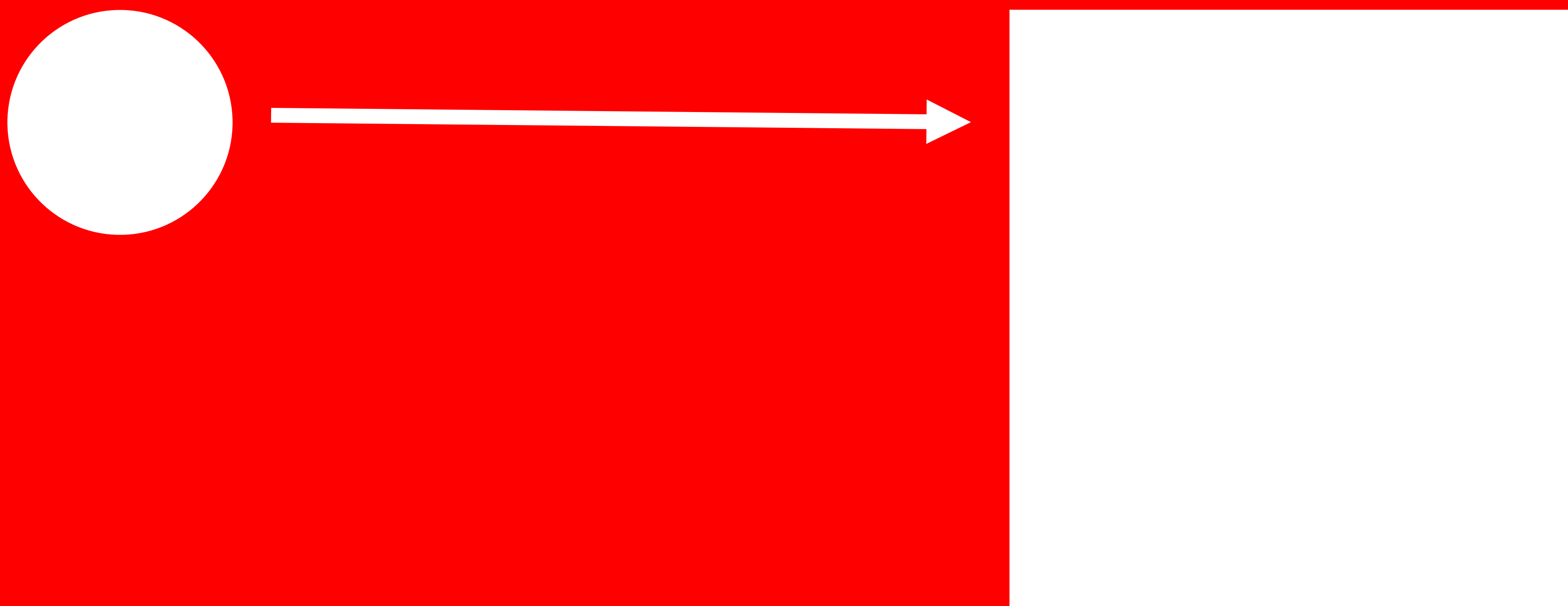
Marketing Analytics

Customer Segmentation

Master-Studiengang Business Analytics an der HS Düsseldorf im Sommersemester 2018

Prof. Dr. Christian Schwarz

Pareto Principle: *„80% of the outcome effects can be „explained“ by 20% of causes“*



„New“ Pareto Prinzip: *Super-Paretos!*

- *„Extreme distributions transcend and dominate industry. Fewer than 10% of drinkers, for example, account for over half the hard liquor sold. Even more extreme, less than 0.25% of mobile gamers are responsible for half of all in-game revenue.“*

- *„A one multibillion-euro industrial equipment company with over 2,000 SKUs determined that less than 4% of its offers were responsible for one-third of sales and roughly half of profitability.“*

„New“ Pareto Prinzip: *Supra-Paretos!*

- *Idea: Combine KPIs within the firm: What 10% of KPI clusters might explain 90% of new customer, growth, or margins? The challenge of supra-Pareto KPIs demand data-driven cross-functional collaboration.*

Pareto Prinzip

MENU

Harvard Business Review

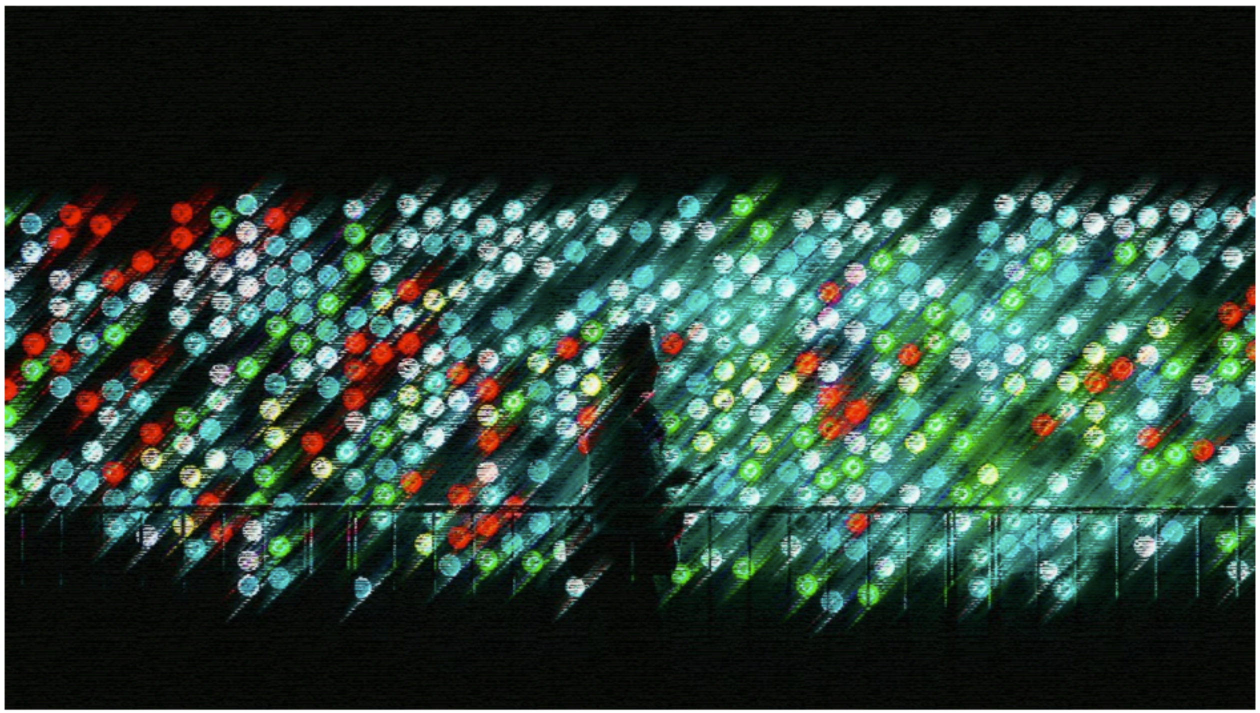
INNOVATION

AI Is Going to Change the 80/20 Rule

by Michael Schrage

FEBRUARY 28, 2017

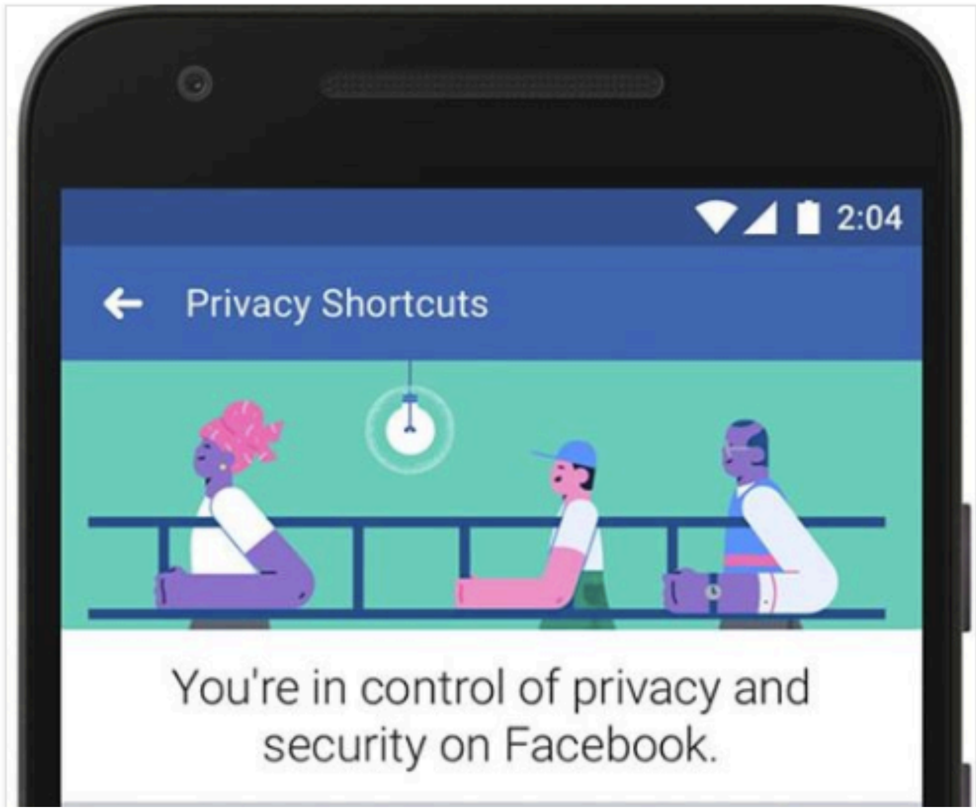
SAVE SHARE 13 COMMENT TEXT SIZE PRINT $8,95 BUY COPIES



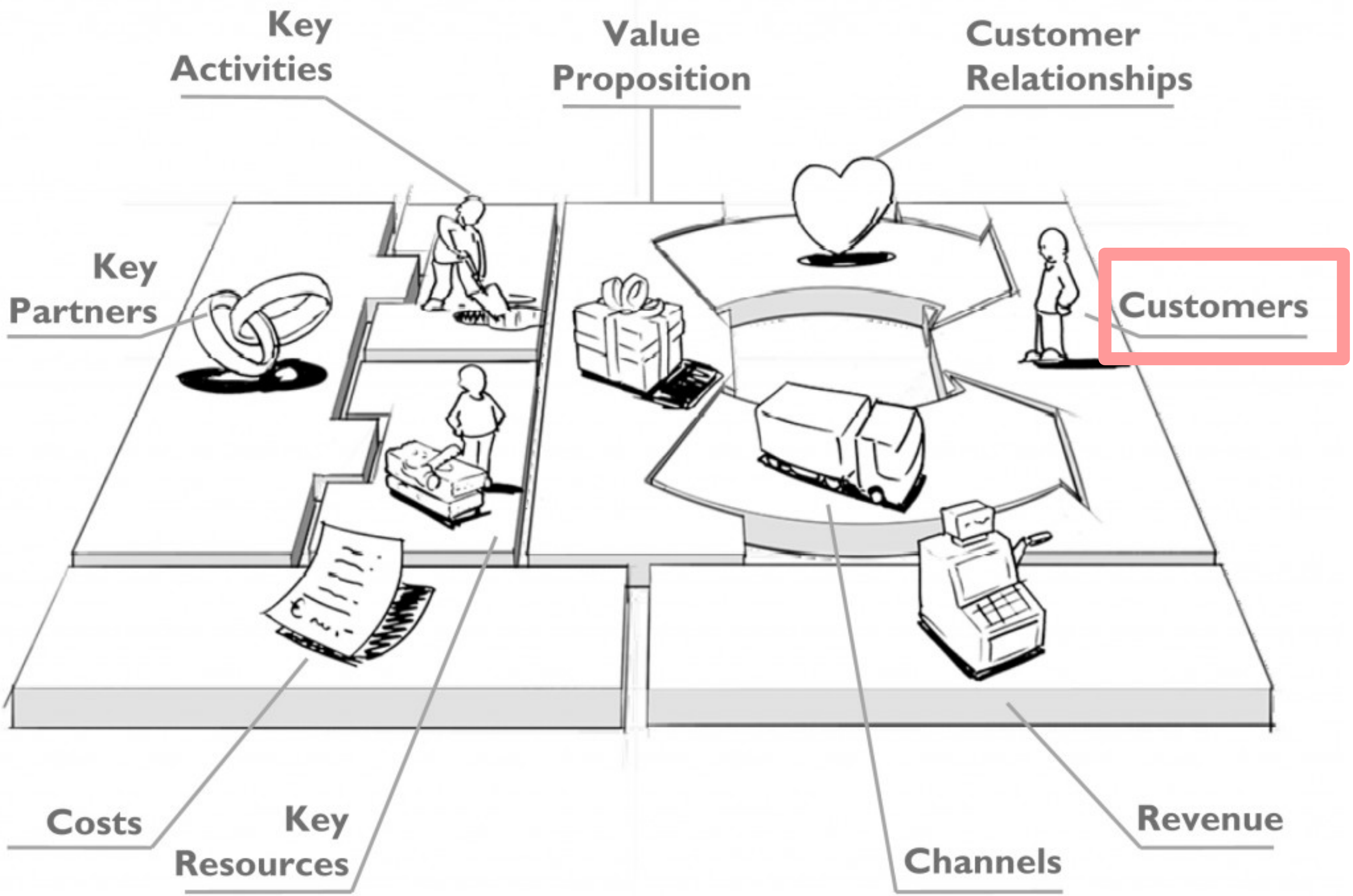
Privacy

Mark Zuckerberg
March 28 at 7:41am ·

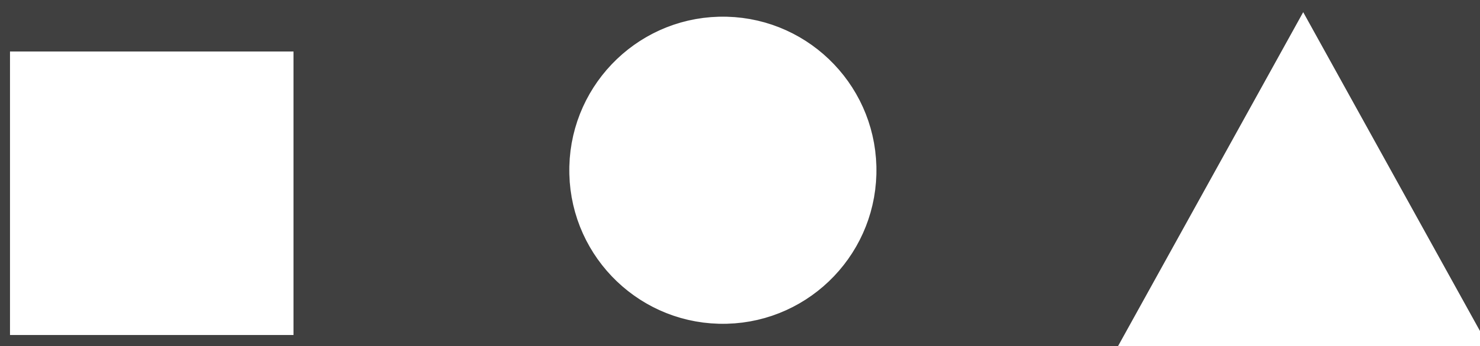
A lot of you are asking how to control what information you share on Facebook, who has access to it, and how to remove it. We recently put all your privacy and security settings in one place called Privacy Shortcuts to make it easier to use. We're going to put this in front of everyone over the next few weeks. We're also going to put a tool with all the platform apps you've signed into in at the top of your News Feed so you can easily remove any apps you no longer use.



Business Model Generation Canvas



Remember*: Customer segmentation is like...*



Cluster Analysis: „Categorize Objects into „similar“ Groups“

- *Example*: „When Procter & Gamble“ test markets a new cosmetic, it wants to group U.S. cities into groups that are similar on demographic attributes such as % of Asians, % of Blacks, % of Hipanics, median age, unemployment rate, and median income level.“

Customer Segmentation: Cluster Analysis

Data Set: 49 of America‘s largest cities

City #	City	% Black	% Hispanic	% Asian	Median Age	Unemployment rate	Per capita income(000's
1	Albuquerque	3	35	2	32	5	18
2	Atlanta	67	2	1	31	5	22
3	Austin	12	23	3	29	3	19
4	Baltimore	59	1	1	33	11	22
5	Boston	26	11	5	30	5	24
6	Charlotte	32	1	2	32	3	20
7	Chicago	39	20	4	31	9	24
8	Cincinnati	38	1	1	31	8	21
9	Cleveland	47	5	1	32	13	22
10	Columbus	23	1	2	29	3	13
11	Dallas	30	21	2	30	9	22
12	Denver	13	23	2	34	7	23
13	Detroit	76	3	1	31	9	21
14	El Paso	3	69	1	29	11	13
15	Fort Worth	22	20	2	30	9	20
16	Fresno	9	30	13	28	13	16
17	Honolulu	1	5	71	37	5	24

Customer Segmentation: Cluster Analysis

Basic Idea of the Cluster Analysis:

Identify clusters based on „a anchor city“ for each cluster. Then assign each city to the „nearest“ cluster (i.e. minimize the sum of squared distances from each city to the closest anchor)

Procedure:

- **Step 1**: Choose *n* trial anchors.
- **Step 2**: Standardize the attributes
- **Step 3**: Calculate squared distances
- **Step 4**: Find anchors which minimize squared distances

Customer Segmentation: Cluster Analysis

Step 1: Choose *n=4* trial anchors.

E.g. the first 4 cities:

- Albuquerque
- Atlanta
- Austin
- Baltimore

Customer Segmentation: Cluster Analysis

Step 2: Standardize the attributes

City #	City	% Black	% Hispanic	% Asian	Median Age	Unemploy ment rate	Per capita income(000 's)
1	Albuquerque	3	35	2	32	5	18
2	Atlanta	67	2	1	31	5	22
3	Austin	12	23	3	29	3	19
4	Baltimore	59	1	1	33	11	22
44	San Jose	5	27	20	30	8	26
45	Seattle	10	4	12	35	5	28
46	Toledo	20	4	1	32	6	19
47	Tucson	4	29	2	31	3	19
48	Tulsa	14	3	1	33	4	20
49	Virginia Beach	14	3	4	29	6	18
	Mean	24,35	14,59	6,04	31,88	7,02	20,92
	Std dev	18,11	16,47	11,14	2,00	2,69	3,33

Reading example: „*The average city has 24% blacks with a standard deviation of 18%.*“

Customer Segmentation: Cluster Analysis

Step 2: Standardize the attributes („z scores“)

City #	City	z Black	z Hispanic	z Asian	z Age	z Unemp	z income
1	Albuquerque	-1,18	1,24	-0,36	0,06	-0,75	-0,88
2	Atlanta	2,36	-0,76	-0,45	-0,44	-0,75	0,32
3	Austin	-0,68	0,51	-0,27	-1,44	-1,50	-0,58
4	Baltimore	1,91	-0,83	-0,45	0,56	1,48	0,32
5	Boston	0,09	-0,22	-0,09	-0,94	-0,75	0,92
6	Charlotte	0,42	-0,83	-0,36	0,06	-1,50	-0,28
7	Chicago	0,81	0,33	-0,18	-0,44	0,74	0,92
48	Tulsa	-0,57	-0,70	-0,45	0,56	-1,12	-0,28
49	Virginia Beach	-0,57	-0,70	-0,18	-1,44	-0,38	-0,88
	Mean	**0**	**0**	**0**	**0**	**0**	**0**
	Std dev	**1**	**1**	**1**	**1**	**1**	**1**

Reading example: *“Altanta has 2.36 standard deviations more Blacks (on a % basis) than a typical city.“*

Customer Segmentation: Cluster Analysis

Step 3: Calculate squared distance from each anchor to each data point

Distance^2 to 1	Distance^2 to 2	Distance^2 to 3	Distance^2 to 4
0	18	4	21
18	0	13	6
4	13	0	22
21	6	22	0
11	20	11	19
16	18	19	18
5	8	5	9
1	16	1	23
5	10	6	13
7	11	3	15

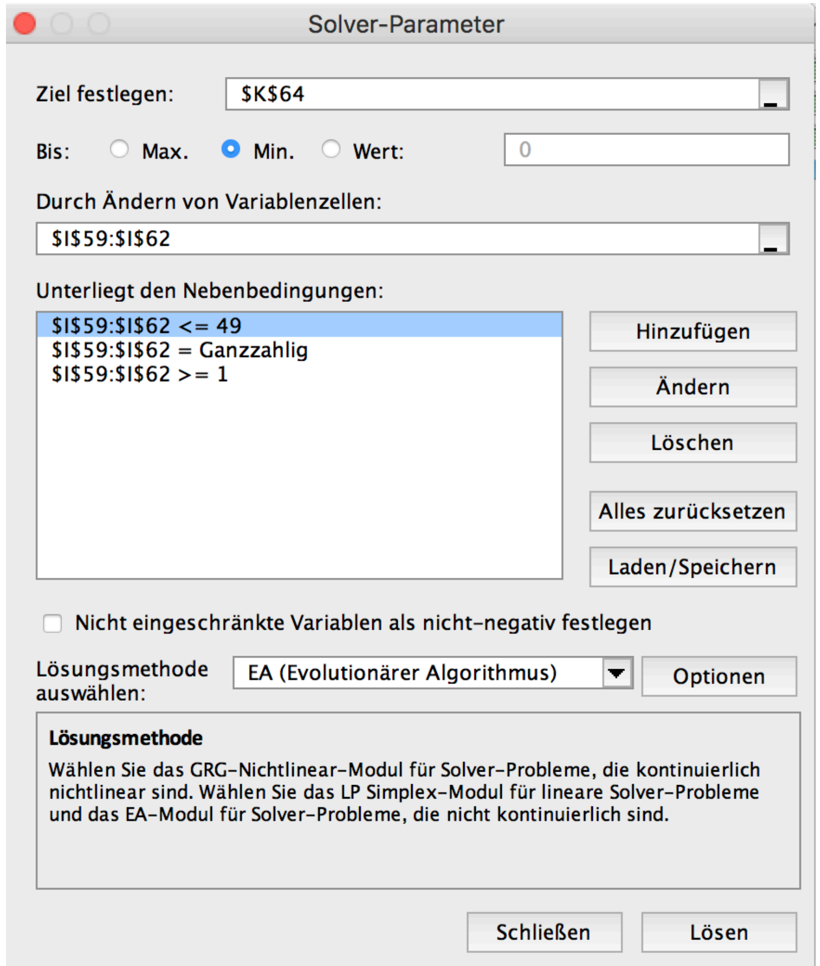
- Anchors: pick arbitrarily 1,2,3,4

City	Cluster	z Black	z Hispanic	z Asian	z Age	z Unemp	z income
Albuquerque	1	-1,18	1,24	-0,36	0,06	-0,75	-0,88
Atlanta	2	2,36	-0,76	-0,45	-0,44	-0,75	0,32
Austin	3	-0,68	0,51	-0,27	-1,44	-1,50	-0,58
Baltimore	4	1,91	-0,83	-0,45	0,56	1,48	0,32

Zielwertfunktion	Sum Dis^2	310,7033

Customer Segmentation: Cluster Analysis

Step 4: Solve via evolutionary solver (->Excel Add-ins) to minimize squared distances



Customer Segmentation: Cluster Analysis

Step 4: Solve via evolutionary solver (->Excel Add-ins) to minimize squared distances

City	Cluster	z Black	z Hispanic	z Asian	z Age	z Unemp	z income
Omaha	34	-0,63	-0,70	-0,45	0,06	-0,75	-0,28
Memphis	25	1,69	-0,83	-0,45	0,06	0,74	-0,28
San Francisco	43	-0,74	-0,04	2,06	2,07	-0,38	3,02
Los Angeles	24	-0,57	1,54	0,36	-0,44	1,48	0,02

Zielwertfunktion	Sum Dis^2	165,3482

Customer Segmentation: Cluster Analysis

Characterization of the clusters

City	Cluster	z Black	z Hispanic	z Asian	z Age	z Unemp	z income
Omaha	34	-0,63	-0,70	-0,45	0,06	-0,75	-0,28
Memphis	25	1,69	-0,83	-0,45	0,06	0,74	-0,28
San Francisco	43	-0,74	-0,04	2,06	2,07	-0,38	3,02
Los Angeles	24	-0,57	1,54	0,36	-0,44	1,48	0,02

Zielwertfunktion	**Sum Dis^2**	**165,3482**

- Omaha: approximately average income with few minorities
- Memphis: highly black cities with high unemployment rates
- San Francisco: Asian, old, high income
- Los Angeles: Hispanics with high unemployment rates.

Customer Segmentation: Cluster Analysis

Implementation in R

- Import, my name: Cluster
- Cluster <- scale(Cluster)
- fit <- kmeans(Cluster, 4)
- aggregate(Cluster,by=list(fit$cluster),FUN=mean)

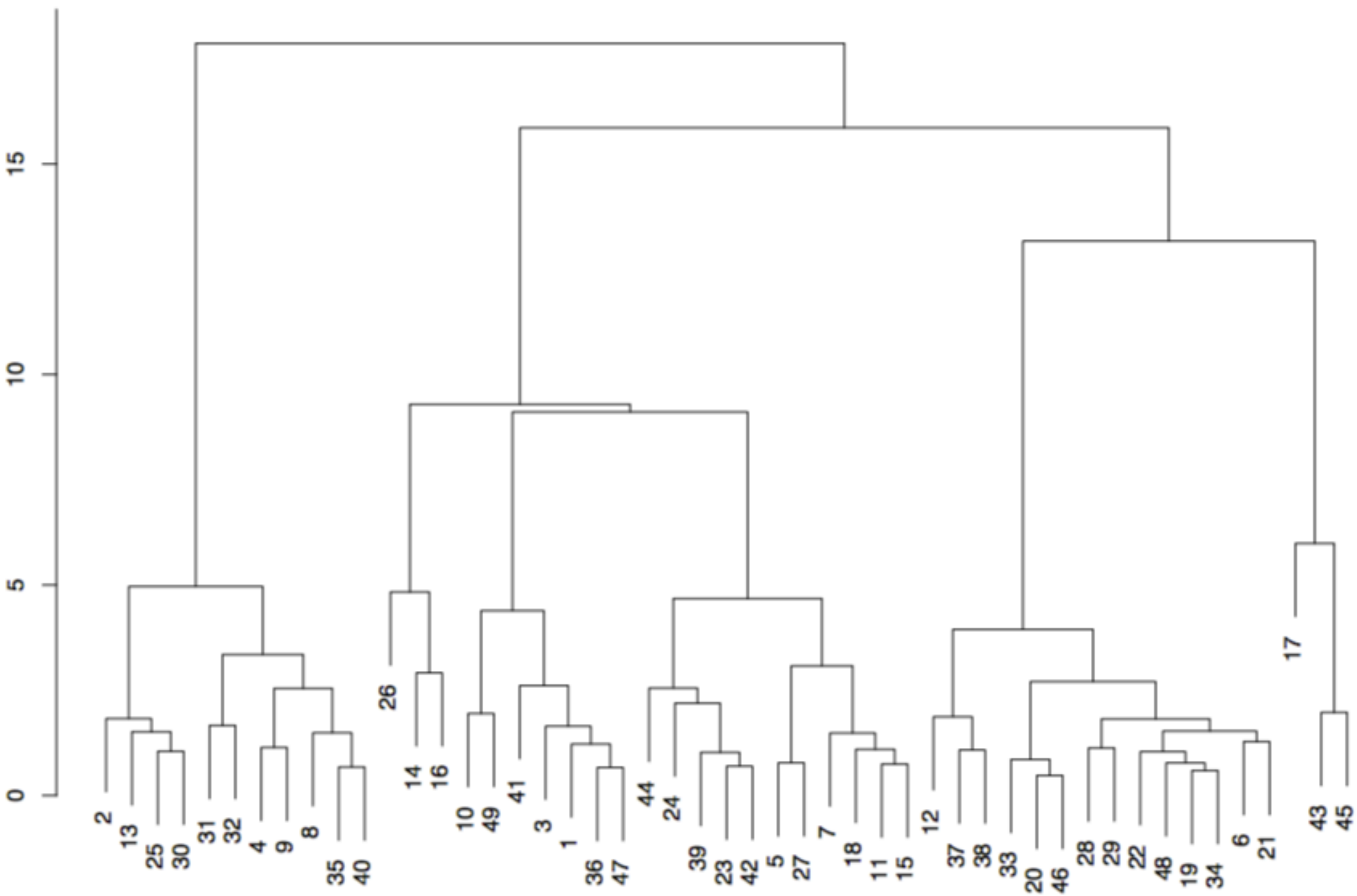
Customer Segmentation: Cluster Analysis

Illustration of Clusters with a Dendrogram

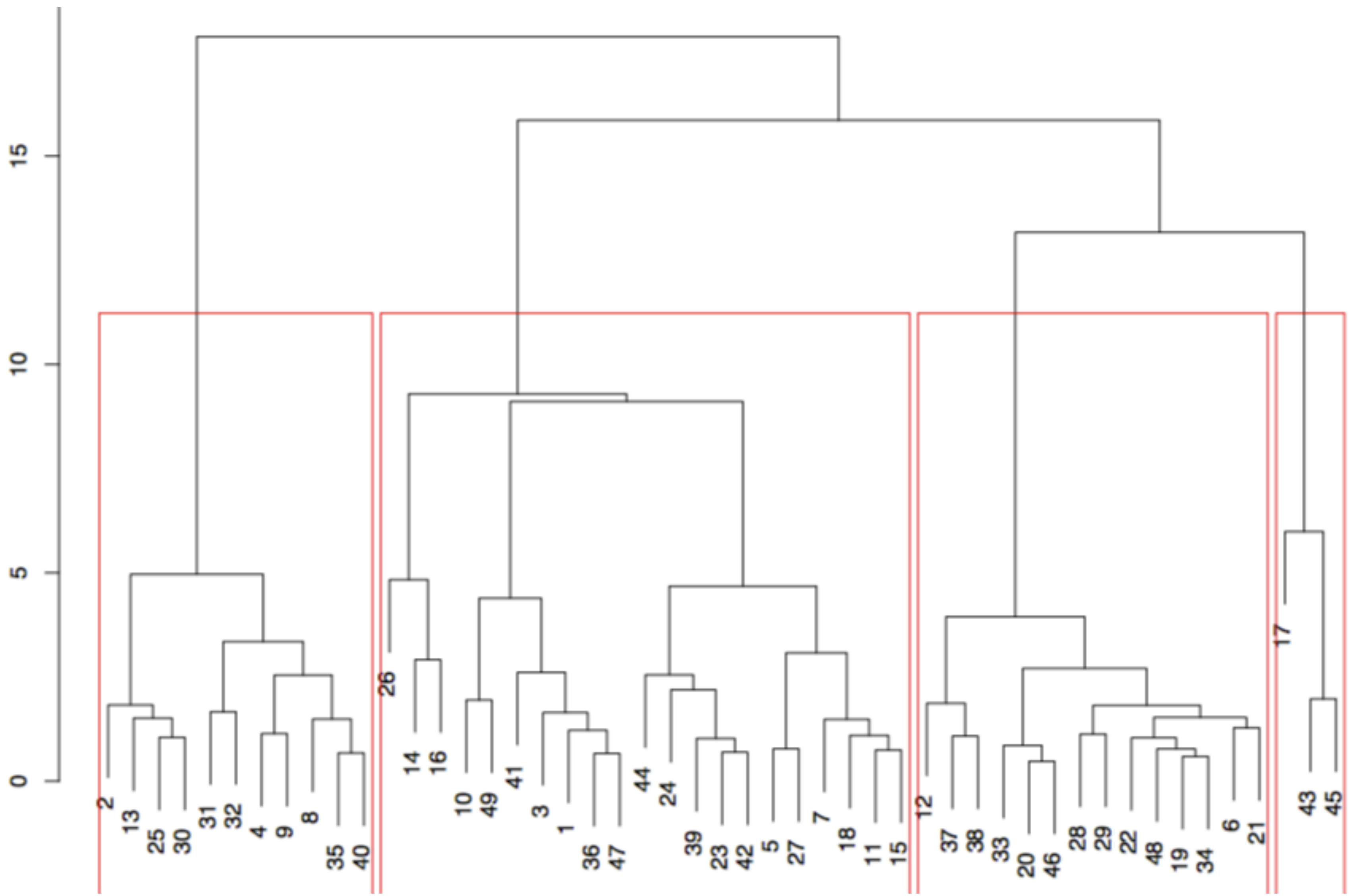
- d <- dist(Cluster, method = "euclidean") # distance matrix
- fit <- hclust(d, method="ward.D")
- plot(fit) # display dendogram

- groups <- cutree(fit, k=4) # cut tree into 4 clusters
- rect.hclust(fit, k=4, border="red")

Customer Segmentation: Dendogram



Customer Segmentation: Dendogram



Latent Class Models

Customer Segmentation: Latent Class Analysis

- Latent Class Analysis

1. Estimate a probability model that describes distribution of your data

2. Calculate probabilities that certain observation are members of certain latent classes

Customer Segmentation: Latent Class Analysis

- Latent Class Analysis

1. Estimate a probability model that describes distribution of your data

2. Calculate probabilities that certain observation are members of certain latent classes

- "Standard" Cluster Analysis

2. Algorithm "only" clusters observations according to the similarity / distance measure

1. Arbitrarily define the "similarity" / distance measure: e.g. k-means

Customer Segmentation: Latent Class Analysis

- Latent Class Analysis

1 Estimate a probability model that describes distribution of your data

2 Calculate probabilities that certain observation are members of certain latent classes

Top down, „theory“ based

- “Standard“ Cluster Analysis

2 Algorithm “only” clusters observations according to the similarity / distance measure

1 Arbitrarily define the “similarity” / distance measure: e.g. k-means

Bottum-up, „only“ mechanics

Homework

- Write a short paper explaining the difference between latent class and standard cluster analysis (800-1000 words) and illustrate with a real world example.
- Write in teams of 2; Grade is 25% of this course
- Short Paper will be posted on the HSD W Journal.
- Use the R package poLCA: **http://dlinzer.github.io/poLCA/**
- References:
  - Linzer, Drew A. and Jeffrey Lewis. 2013. "poLCA: Polytomous Variable Latent Class Analysis." R package version 1.4. http://dlinzer.github.com/poLCA.