Lecture 9. K-Means

Suzer-Gurtekin
Lecture notes adapted from Steve Miller, Richard Valliant, James Wagner and Fred Feinberg

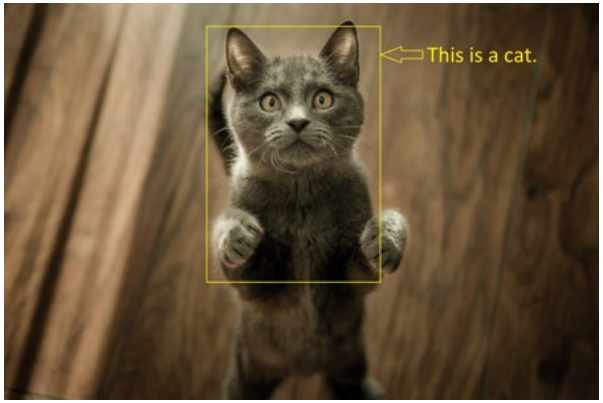
March 2025

Overview

Clustering

- Clustering is a form of **unsupervised learning**.
- Unsupervised learning involves learning from unlabelled data.
- **Supervised learning** example: Learning how to identify “cats” from labelled pictures, then identifying unlabelled pictures that also contain cats.
- Unsupervised learning looks for patterns in the data.

Cat



Clustering

The goal in clustering is to **detect patterns** in the data.

For example, patterns in customer purchasing data, tweets, etc.

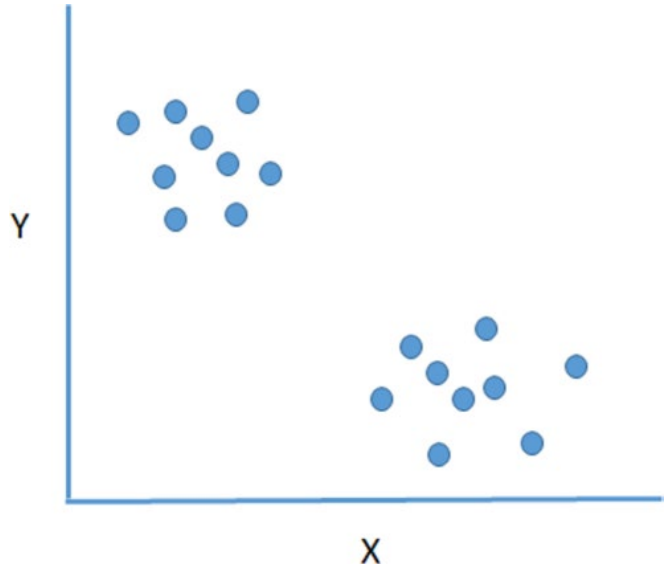
This is useful when you do not start with a specific target for prediction.

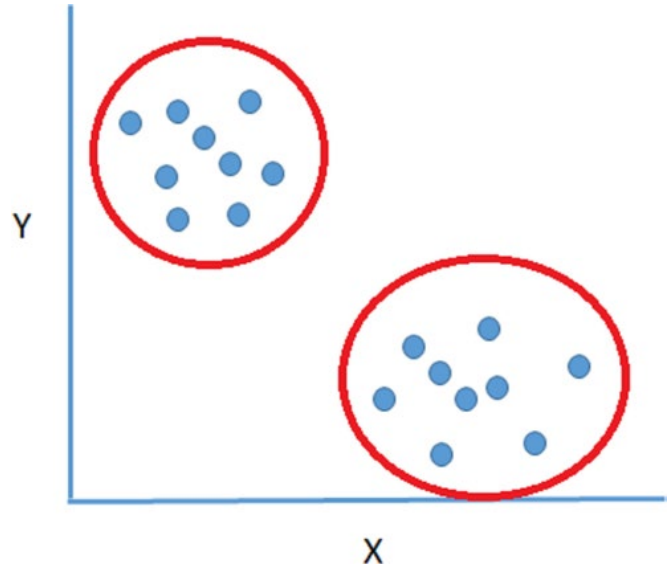
But, the risk is that results will be meaningless/useless.

Two-Dimensional Example

Let's start from a two-dimensional example.

Here, the idea is to group things together that are similar based on the two continuous dimensions.





Distance Measures

But creating clusters depends upon how you define “distance” or “similarity”.

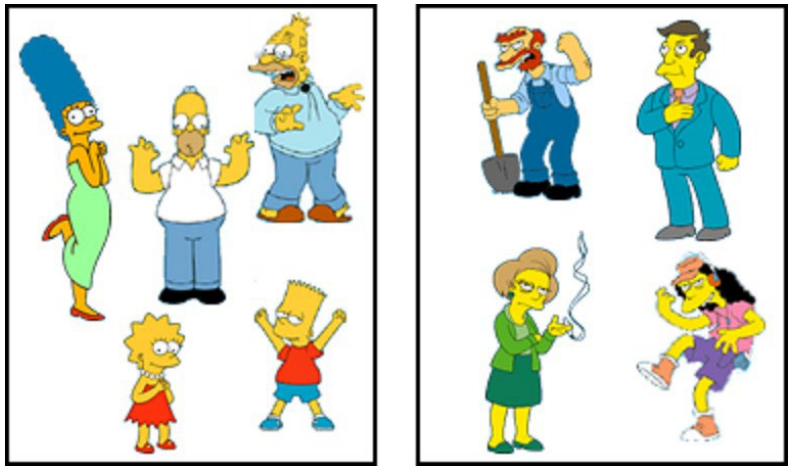
An example distance measure is the **Euclidean Distance** between two points:

$$d = \sqrt{[(x_2 - x_1)^2 + (y_2 - y)^2}$$

The results from a clustering algorithm can differ (often substantially) based upon the choice of distance measure.

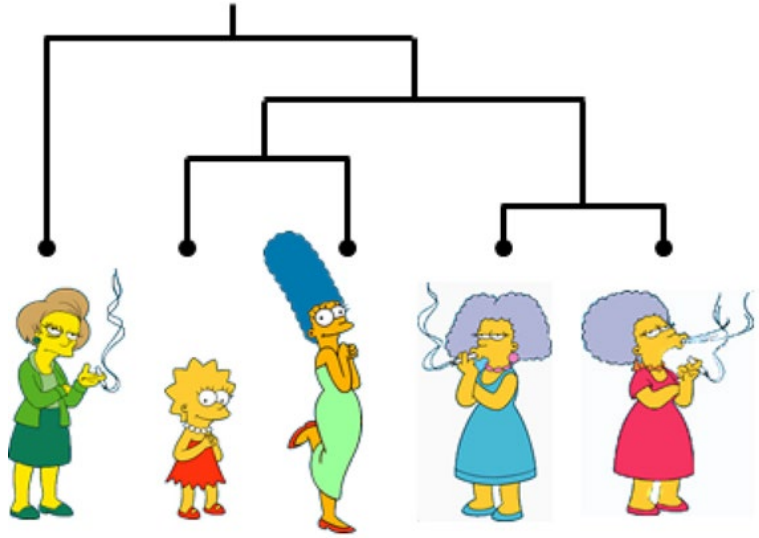
Types of Clustering Algorithms

Partition



Fixed number of clusters

Hierarchical



Hierarchy of clusters

Clustering Example 1

Example of using clustering with survey data to identify strata in the population (Ehlert, et al., 2017)

They want to **identify strata** based on the survey data in order to do qualitative work

They want all major groups or strata included in the qualitative work

Ehlert, et al., 2017

Phase 1- Collect survey data
Phase 2- Use survey data to create clusters
Phase 3- Select participants based on the clusters
Phase 4- In-depth interviews of selected participants

Clustering Example 1

They use two principal components (we will look at these later in the course) as the inputs.

Then use k-means to create two groups or strata.

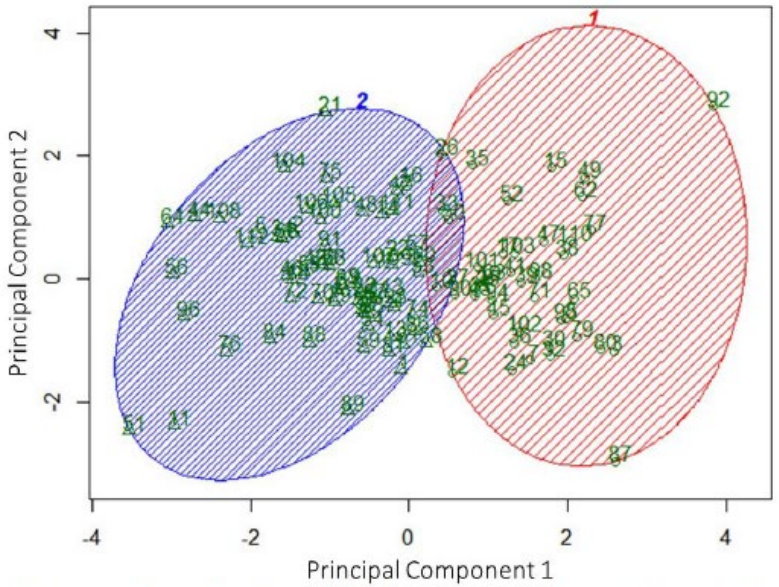


Figure 6: Two-dimensional visualization of the k-means two cluster solution for the survey data set. The red and blue ellipses are the minimum area that incorporates all the participants within the cluster.

Clustering Example 1

Table 3: Summary of the average scores for the clusters on each factor used during cluster analysis. The total number of participants in each cluster is also provided at the bottom of the table.

	Ward's		Complete Link		k-means	
Factor	Cluster 1	Cluster 2	Cluster 1	Cluster 2	Cluster 1	Cluster 2
Closed-Mindedness	2.75 ± 0.75	3.68 ± 0.73	2.69 ± 0.71	3.72 ± 0.71	2.53 ± 0.74	3.43 ± 0.75
Discomfort with Ambiguity	4.84 ± 0.92	4.74 ± 0.76	4.85 ± 0.94	4.74 ± 0.72	4.71 ± 0.93	4.87 ± 0.82
Certainty of Knowledge	2.04 ± 0.59	3.27 ± 0.72	2.06 ± 0.63	3.14 ± 0.80	1.80 ± 0.51	2.92 ± 0.76
Sources of Knowledge	3.81 ± 0.80	4.48 ± 0.74	3.76 ± 0.80	4.50 ± 0.70	3.42 ± 0.74	4.46 ± 0.61
Justification of Knowledge	3.24 ± 0.75	4.33 ± 0.60	3.21 ± 0.75	4.30 ± 0.59	2.93 ± 0.66	4.08 ± 0.67
Number of Participants (n)	71	37	68	40	44	64

Clustering Example 1

	Ward's			Complete Link			k-means			
Factor	Cluster 1	Cluster 2	Dif(Cluster2-Cluster1)	Cluster 1	Cluster 2	Dif(Cluster2-Cluster1)	Cluster 1	Cluster 2	Dif(Cluster2-Cluster1)	Average difference
Closed-Mindedness	2.75	3.68	0.93	2.69	3.72	1.03	2.53	3.43	0.9	0.95 Consistent
Discomfort with Ambiguity	4.84	4.74	-0.1	4.85	4.74	-0.11	4.71	4.87	0.16	-0.02 Consistent
Certainty of Knowledge	2.04	3.27	1.23	2.06	3.14	1.08	1.8	2.92	1.12	1.14 Consistent
Sources of Knowledge	3.81	4.48	0.67	3.76	4.5	0.74	3.42	4.46	1.04	0.82 Somewhat
Justification of Knowledge	3.24	4.33	1.09	3.21	4.3	1.09	2.93	4.08	1.15	1.11 Consistent

How would you describe cluster 1 vs. cluster 2 on these dimensions?

Clustering Example 1

	Ward's			Complete Link			k-means			
Factor	Cluster 1	Cluster 2	Dif(Cluster2-Cluster1)	Cluster 1	Cluster 2	Dif(Cluster2-Cluster1)	Cluster 1	Cluster 2	Dif(Cluster2-Cluster1)	Average difference	
Closed-Mindedness	2.75	3.68	0.93	2.69	3.72	1.03	2.53	3.43	0.9	0.95	Consistent
Discomfort with Ambiguity	4.84	4.74	-0.1	4.85	4.74	-0.11	4.71	4.87	0.16	-0.02	Consistent
Certainty of Knowledge	2.04	3.27	1.23	2.06	3.14	1.08	1.8	2.92	1.12	1.14	Consistent
Sources of Knowledge	3.81	4.48	0.67	3.76	4.5	0.74	3.42	4.46	1.04	0.82	Somewhat
Justification of Knowledge	3.24	4.33	1.09	3.21	4.3	1.09	2.93	4.08	1.15	1.11	Consistent

Cluster 2 higher on certainty of knowledge, justification of knowledge, closed-mindedness and sources of knowledge.

Clustering Example 1

	Ward's			Complete Link			k-means			
Factor	Cluster 1	Cluster 2	Dif(Cluster2-Cluster1)	Cluster 1	Cluster 2	Dif(Cluster2-Cluster1)	Cluster 1	Cluster 2	Dif(Cluster2-Cluster1)	Average difference
Closed-Mindedness	2.75	3.68	0.93	2.69	3.72	1.03	2.53	3.43	0.9	0.95 Consistent
Discomfort with Ambiguity	4.84	4.74	-0.1	4.85	4.74	-0.11	4.71	4.87	0.16	-0.02 Consistent
Certainty of Knowledge	2.04	3.27	1.23	2.06	3.14	1.08	1.8	2.92	1.12	1.14 Consistent
Sources of Knowledge	3.81	4.48	0.67	3.76	4.5	0.74	3.42	4.46	1.04	0.82 Somewhat
Justification of Knowledge	3.24	4.33	1.09	3.21	4.3	1.09	2.93	4.08	1.15	1.11 Consistent

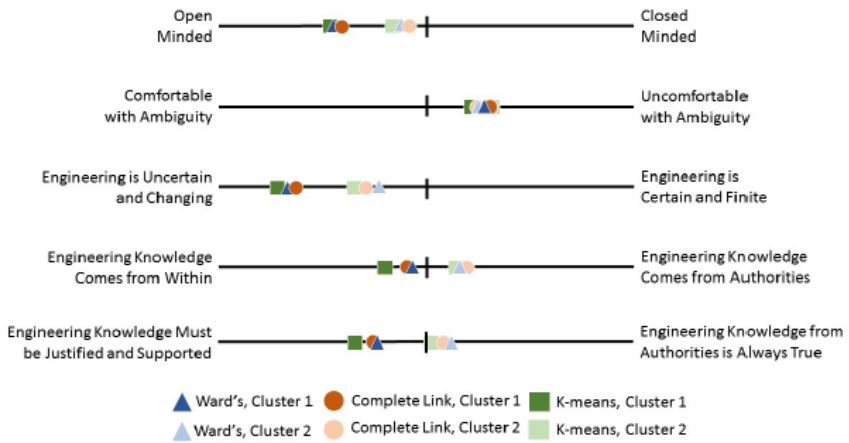


Figure 7: Visual representation of the cluster means on each factor. Cluster means are plotted on each factor scale. Ward's clusters are indicated by the blue triangles, Complete Link by the orange circles, and k-means by green squares. The dark shapes represent the mean for Cluster 1 and the light shapes represent the mean of Cluster 2. Neutral point was indicated with a black dash and descriptors of the ends are provided.

Clustering Example 2

Marketing Research: Methodological Foundations, by Gilbert A. Churchill and Dawn Iacobucci, an exceptionally thorough introduction to the subject.

Chapter 18

Clustering Example 2 (Dr. Fred Feinberg's class notes)

- This handout was prepared as an annotated guide to interpreting the results of a (hierarchical) Cluster Analysis run by a fairly complex statistical program, in this case SPSS.
- While it offers only a partial listing of the statements necessary to load the data and ask for specific parts of analyses, it does give a relatively complete picture of the actual output you would receive from such an analysis.
- The data used for the purposes of this illustrative example is attribute (caloric content, sodium content, alcohol content and wholesale price) data for 20 brands of beer. The data are given in the next slide:

Clustering Example 2 (Dr. Fred Feinberg's class notes)

Brand	Calories (12 oz.)	Sodium (mg./12 oz.)	Alcohol (%)	Price (Wholesale)
BUDWEISER	144	15	4.7	.43
SCHLITZ	151	19	4.9	.43
LOWENBRAU	157	15	4.9	.48
KRONENBOURG	170	7	5.2	.73
HEINEKEN	152	11	5.0	.77
OLD MILWAUKEE	145	23	4.6	.28
AUGSBERGER	175	24	5.5	.40
STROHS BOHEMIAN STYLE	149	27	4.7	.42
MILLER LITE	99	10	4.3	.43
BUDWEISER LIGHT	113	8	3.7	.44
COORS	140	18	4.6	.44
COORS LIGHT	102	15	4.1	.46
MICHELOB LIGHT	135	11	4.2	.50
BECKS	150	19	4.7	.76
KIRIN	149	6	5.0	.79
PABST EXTRA LIGHT	68	15	2.3	.38
HAMMS	136	19	4.4	.43
HEILEMANS OLD STYLE	144	24	4.9	.43
OLYMPIA GOLD LIGHT	72	6	2.9	.46
SCHLITZ LIGHT	97	7	4.2	.47

Clustering Example 2 (Dr. Fred Feinberg's class notes)

The first thing to notice is that these variables are on *radically* different measurement scales from one another:

	Calories	Sodium (mg.)	Alcohol (%)	Price ($)
Mean	**132**	**15**	**4.4**	**0.50**
Minimum	**68**	**6**	**2.3**	**0.28**
Maximum	**175**	**27**	**5.5**	**0.79**

This must be fixed! Cluster Analysis is a statistical technique, not a mind-reading procedure; it doesn't know what the variables *mean*.

So, to get the variables on the same scale, we do what we always do in statistics: ***standardize*** (subtract the mean and divide by the standard deviation).

[Note: if we ***do*** want some variables to 'count' more than others, we can re-weight them – scale them up or down – *after* standardizing.]

Clustering Example 2 (Dr. Fred Feinberg's class notes)

Brand	Calories (12 oz.)	Sodium (mg./12 oz.)	Alcohol (%)	Price (Wholesale)
BUDWEISER	0.383	0.008	0.342	-0.463
SCHLITZ	0.615	0.615	0.605	-0.463
LOWENBRAU	0.813	0.008	0.605	-0.115
KRONENBOURG	1.243	-1.208	1.000	1.624
HEINEKEN	0.648	-0.600	0.737	1.903
OLD MILWAUKEE	0.416	1.223	0.211	-1.506
AUGSBERGER	1.408	1.375	1.395	-0.671
STROHS BOHEMIAN STYLE	0.549	1.831	0.342	-0.532
MILLER LITE	-1.104	-0.752	-0.184	-0.463
BUDWEISER LIGHT	-0.641	-1.056	-0.974	-0.393
COORS	0.251	0.463	0.211	-0.393
COORS LIGHT	-1.005	0.008	-0.447	-0.254
MICHELOB LIGHT	0.086	-0.600	-0.316	0.024
BECKS	0.582	0.615	0.342	1.833
KIRIN	0.549	-1.360	0.737	2.042
PABST EXTRA LIGHT	-2.128	0.008	-2.817	-0.810
HAMMS	0.119	0.615	-0.053	-0.463
HEILEMANS OLD STYLE	0.383	1.375	0.605	-0.463
OLYMPIA GOLD LIGHT	-1.996	-1.360	-2.027	-0.254
SCHLITZ LIGHT	-1.170	-1.208	-0.316	-0.184

Clustering Example 2 (Dr. Fred Feinberg's class notes)

Are these variables *redundant*? If we calculate the correlations among these four variables, we find:

r	Calories	Sodium	Alcohol	Price
Calories	**1.00**			
Sodium	**0.41**	**1.00**		
Alcohol	**0.92**	**0.32**	**1.00**	
Price	**0.32**	**-0.45**	**0.33**	**1.00**

These correlations are modest, with one glaring exception: Alcohol and Calories. This is to be expected, since alcohol is the main source of the calories in alcoholic beverages. A correlation of 0.92 is *exceptionally* high: the two variables have almost identical "information content."

We must bear in mind that we are essentially *putting the same information into the Cluster Analysis twice*, and should not be surprised if our final cluster solution reflects this redundancy.

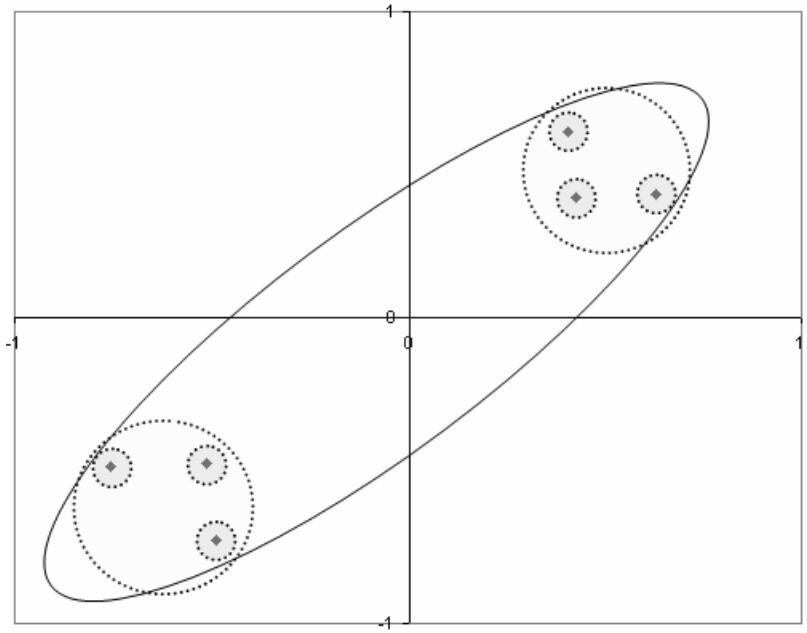
Clustering Example 2 (Dr. Fred Feinberg's class notes)

We wish to identify which of these beers are most like one another, that is, to identify ***clusters*** of beers with *similar profiles* (relative to the four variables above).

Depending on one's vantage point, all the beers, being beers and not, say, colas or motor oil, might appear similar and be deemed to be in a single cluster; conversely, a beer connoisseur might see each beer as totally unique, and place each into a separate cluster. *These are both reasonable points of view.*

Clustering Example 2 (Dr. Fred Feinberg's class notes)

For example, here are three different "cluster solutions", with one, two and six clusters; *each is perfectly valid*, depending on what the researcher is looking for.



Clustering Example 2 (Dr. Fred Feinberg's class notes)

The goal of cluster analysis is to provide some intermediate level of aggregation, so that some beers are pronounced similar and others not so similar. The appropriate number of clusters is seldom known in advance, and is typically discerned from the analysis itself.

Unfortunately, unlike in regression or Factor Analysis, where there are objective measures (e.g., r^2) of fit, one must take a more "exploratory" approach in cluster analysis, and base one's judgment on more-or-less pictorial evidence (along with a few numerical benchmarks and guides).

In the following annotated cluster analysis output, dialogue with the computer appears in **`Courier`**.

Clustering Example 2 (Dr. Fred Feinberg's class notes)

The first thing the computer does is take all the data – in this case, the four variables – and convert them to *distances* (in this case, "squared Euclidean Distances") , often called "*dissimilarities*", that the Cluster Analysis will work with (for the first 10 beers; small distances highlighted in **red**; large in **magenta**) :

Squared Euclidean Distance

Brand	1	2	3	4	5	6	7	8	9	10
1:Budweiser	0									
2:Schlitz	0.49	0								
3:Lowenbrau	0.38	0.53	0							
4:Kronenbourg	7.00	8.23	4.84	0						
5:Heineken	6.19	7.09	4.48	0.87	0					
6:Old Milwaukee	2.59	1.65	3.73	17.02	15.27	0				
7:Augsberger	4.07	1.87	3.16	12.13	11.54	3.11	0			
8:Stroh's Bohemian	3.36	1.56	3.64	14.80	12.00	1.35	2.07	0		
9:Miller Lite	3.07	5.45	5.00	11.47	9.53	7.46	13.37	9.69	0	
10:Budweiser Light	3.92	6.87	5.82	11.54	10.07	8.96	15.80	11.50	0.94	0

Clustering Example 2 (Dr. Fred Feinberg's class notes)

If we *really* want to see all the (squared) distances… here they are, with those **over 20** and **under 1** highlighted.

Squared Euclidean Distance (values over 20 and under 1 highlighted)

Case	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1:Budweiser	0	0.49	0.38	7.00	6.19	2.59	4.07	3.36	3.07	3.92	0.25	2.59	1.13	5.68	8.32	16.41	0.60	1.94	13.19	4.40
2:Schlitz	0.49	0	0.53	8.23	7.09	1.65	1.87	1.56	5.45	6.87	0.32	4.14	2.84	5.34	10.20	19.73	0.68	0.63	17.69	7.44
3:Lowenbrau	0.38	0.53	0	4.84	4.48	3.73	3.16	3.64	5.00	5.82	0.76	4.43	1.77	4.29	6.61	20.85	1.41	2.18	16.71	6.26
4:Kronenbourg	7.00	8.23	4.84	0	0.87	17.02	12.13	14.80	11.47	11.54	8.47	12.15	6.00	4.24	0.75	33.34	10.05	11.92	23.21	10.82
5:Heineken	6.19	7.09	4.48	0.87	0	15.27	11.54	12.00	9.53	10.07	6.84	9.15	4.95	1.64	0.61	28.07	7.98	9.58	19.86	9.14
6:Old Milwaukee	2.59	1.65	3.73	17.02	15.27	0	3.11	1.35	7.46	8.96	1.84	5.50	6.05	11.56	19.55	17.60	1.62	1.27	19.07	10.45
7:Augsberger	4.07	1.87	3.16	12.13	11.54	3.11	0	2.07	13.37	15.80	3.65	11.26	9.06	8.64	16.01	32.13	4.38	1.72	30.95	16.48
8:Stroh's Bohemian	3.36	1.56	3.64	14.80	12.00	1.35	2.07	0	9.69	11.50	2.00	6.44	6.87	7.07	16.96	20.55	1.82	0.31	22.35	12.74
9:Miller Lite	3.07	5.45	5.00	11.47	9.53	7.46	13.37	9.69	0	0.94	3.47	0.70	1.69	10.26	10.22	8.68	3.38	7.36	4.61	0.31
10:Budweiser Light	3.92	6.87	5.82	11.54	10.07	8.96	15.80	11.50	0.94	0	4.51	1.56	1.34	10.98	10.36	6.91	4.23	9.46	3.06	0.78
11:Coors	0.25	0.32	0.76	8.47	6.84	1.84	3.65	2.00	3.47	4.51	0	2.24	1.61	5.11	9.62	15.21	0.12	1.01	13.40	5.13
12:Coors Light	2.59	4.14	4.43	12.15	9.15	5.50	11.26	6.44	0.70	1.56	2.24	0	1.65	7.87	10.96	7.19	1.83	4.95	5.35	1.53
13:Michelob Light	1.13	2.84	1.77	6.00	4.95	6.05	9.06	6.87	1.69	1.34	1.61	1.65	0	5.43	5.97	12.22	1.79	5.08	7.92	1.99
14:Beck's 15:Kirin	5.68	5.34	4.29	4.24	1.64	11.56	8.64	7.07	10.26	10.98	5.11	7.87	5.43	0	4.10	24.68	5.64	5.96	20.52	10.90
16:Pabst Extra Light	8.32	10.20	6.61	0.75	0.61	19.55	16.01	16.96	10.22	10.36	9.62	10.96	5.97	4.10	0	29.80	10.98	13.80	19.39	9.04
17:Hamm's	16.41	19.73	20.85	33.34	28.07	17.60	32.13	20.55	8.68	6.91	15.21	7.19	12.22	24.68	29.80	0	13.18	20.01	2.82	9.04
18:Heileman's Old	0.60	0.68	1.41	10.05	7.98	1.62	4.38	1.82	3.38	4.23	0.12	1.83	1.79	5.64	10.98	13.18	0	1.08	12.32	5.13
19:Olympia Gold	1.94	0.63	2.18	11.92	9.58	1.27	1.72	0.31	7.36	9.46	1.01	4.95	5.08	5.96	13.80	20.01	1.08	0	20.12	10.01
20:Schlitz Light	13.19	17.69	16.71	23.21	19.86	19.07	30.95	22.35	4.61	3.06	13.40	5.35	7.92	20.52	19.39	2.82	12.32	20.12	0	3.64
	4.40	7.44	6.26	10.82	9.14	10.45	16.48	12.74	0.31	0.78	5.13	1.53	1.99	10.90	9.04	9.04	5.13	10.01	3.64	0

Let's see what happens when we run the *Hierarchical Cluster Analysis* on these data…

Clustering Example 2

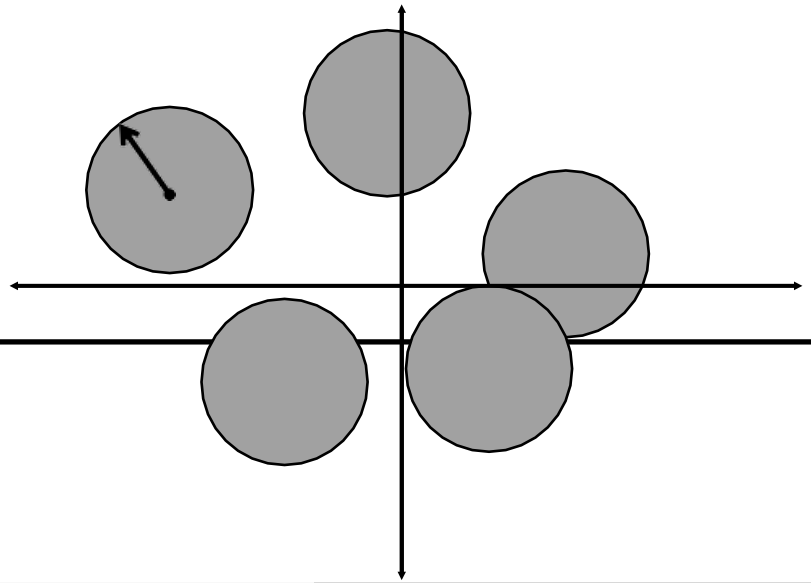
Centroid Hierarchical Cluster Analysis

Clusters Remaining	Centroid Distance	Clusters or Items Combined	
19	0.115	Coors (11)	Hamm's (17)
18	0.307	Miller Lite (9)	Schlitz Light (20)
17	0.309	Stroh's Bohemian (8)	Heileman's Old Style (18)
16	0.375	Budweiser (1)	Lowenbrau (3)
15	0.417	CL16	Schlitz (2)
14	0.483	CL15	CL19
13	0.606	Heineken (5)	Kirin (15)
12	0.658	CL13	Kronenbourg (4)
11	0.780	CL18	Budweiser Light (10)
10	1.038	CL11	Coors Light (12)
9	1.233	CL17	Old Milwaukee (6)
8	1.307	CL10	Michelob Light (13)
7	1.496	CL14	CL09
6	2.393	CL07	Augsberger
5	2.821	Pabst Extra Light (16)	Olympia Gold Light (19)
4	3.081	CL12	Beck's (14)
3	5.098	CL06	CL08
2	6.915	CL03	CL04
1	13.405	CL02	CL05

Clustering Example 2

Comment: *The output above represents the clusters derived from the attribute and price data in a rather compact form. A way to go about interpreting it is as follows. We will first be concerned with the last column, the cryptic* ***"normalized centroid distance" (NCD)****; just think of this as "radius of a circle/sphere that I put around each point once all four attribute dimensions are made equally important". If the NCD is small, then no two of the little circle/spheres around the beers* ***"overlap"*** *one another, and we are forced to conclude that there are* ***20 separate clusters, each containing one beer****.*

NCD
0.115
0.307
0.309
0.375
0.417
0.483



Clustering Example 2

However, as the NCD gets larger, eventually some pair of beers has got to overlap. The first row above indicates that, when the NCD gets up to around .115 or so, Coors (11) and Hamm's (17) are forced into the same cluster.

```
19    0.115    Coors (11)    Hamm's (17)
```

*The distinguishing feature of **HIERARCHICAL** cluster analysis (as opposed to non-hierarchical) is that once a group of objects is clustered together, they are together for life; even if at a later stage it becomes evident that two large clusters should split up a smaller one between them, **this is not allowed**, and one of the two large clusters will have to subsume the smaller one **completely**. So it is that Coors and Hamm's will remain together throughout the clustering procedure.*

Clustering Example 2

Next, Miller Lite (9) and Schlitz Light (20) form a cluster at NCD = 0.307, Stroh's Bohemian (8) and Heileman's Old Style (18) at NCD = 0.309, and Budweiser (1) and Lowenbrau (3) at NCD = 0.375.

At this stage, there would be 16 clusters, 12 with a single brand, and 4 with two brands each.

Clusters Remaining	Centroid Distance	Clusters or Items Combined	
19	0.115	Coors (11)	Hamm's (17)
18	0.307	Miller Lite (9)	Schlitz Light (20)
17	0.309	Stroh's Bohemian (8)	Heileman's Old Style (18)
16	0.375	Budweiser (1)	Lowenbrau (3)

Clustering Example 2

It is important to understand what happens next, near NCD = 0.417. Schlitz (2), rather than join up with a single other brand to form a cluster of size two (as all other clusters have been formed thus far), decides that it will join up with Cluster 16 (CL16)*, that is, with Budweiser (1) and Lowenbrau (3),* ***to form a cluster of size three****.*

Clusters Remaining	Centroid Distance	Clusters or Items Combined	
19	0.115	Coors (11)	Hamm's (17)
18	0.307	Miller Lite (9)	Schlitz Light (20)
17	0.309	Stroh's Bohemian (8)	Heileman's Old Style (18)
16	0.375	Budweiser (1)	Lowenbrau (3)
15	0.417	CL16	Schlitz (2)

Clustering Example 2

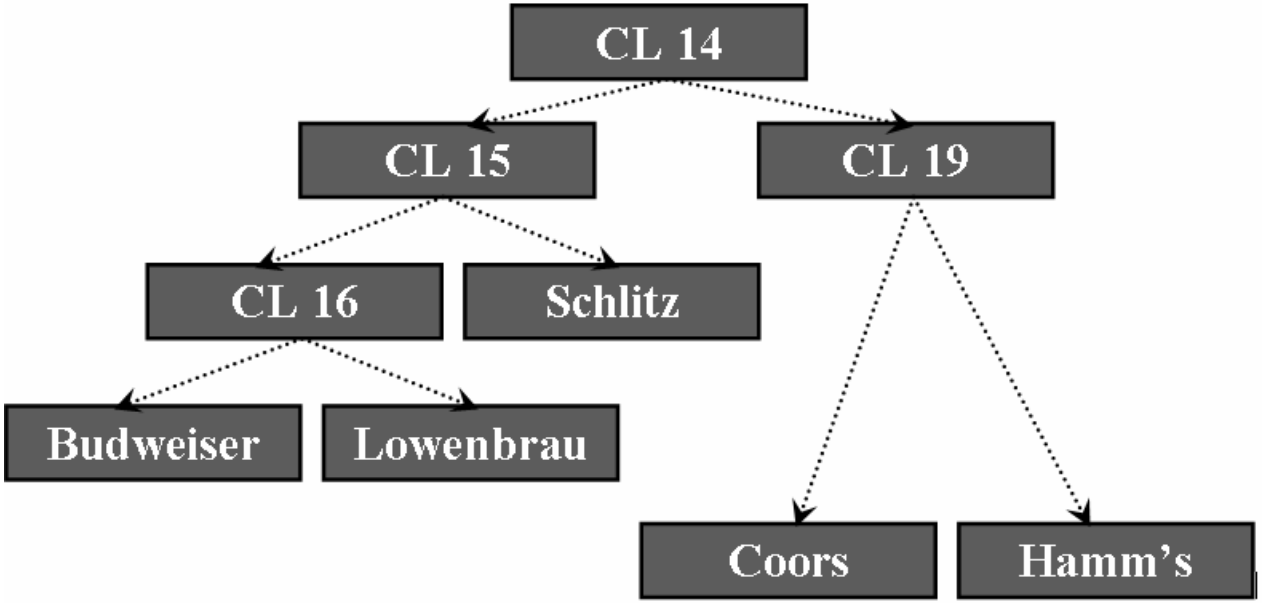
It should be apparent at this point that each stage in the clustering list above tells when a cluster, whether it have one, two or seventeen members, joins with another cluster; whenever this happens, the number of clusters decreases by one.

Thus we see that, as NCD gets larger and larger, more clusters keep joining with single-brand clusters until, at NCD = 0.483, two larger clusters, CL15 and CL19, join one another, making a single cluster of five brands: Budweiser (1), Schlitz (2), Lowenbrau (3), Coors (11) and Hamm's (17).

Clusters Remaining	Centroid Distance	Clusters or Items Combined	
14	0.483	CL15	CL19

Clustering Example 2

*A **Hierarchical** Cluster Breakdown*



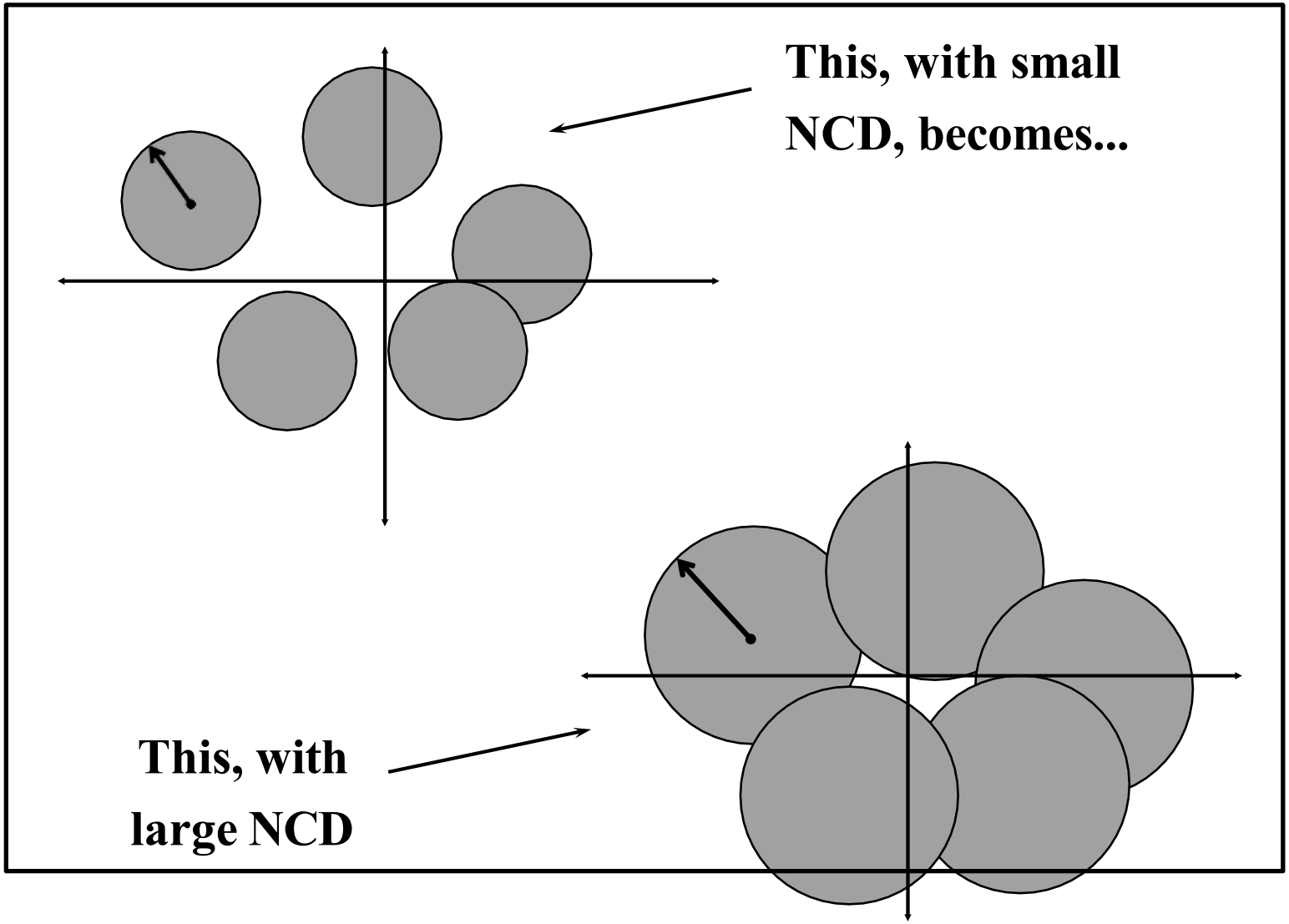
Clustering Example 2

This process continues until there are only two clusters left, which are themselves eventually joined at NCD = 13.405.

This must always happen since, if one draws exceptionally large circle/spheres around each of the brands, they will all eventually overlap.

Clusters Remaining	Centroid Distance	Clusters or Items Combined	
1	13.405	CL02	CL05

Clustering Example 2



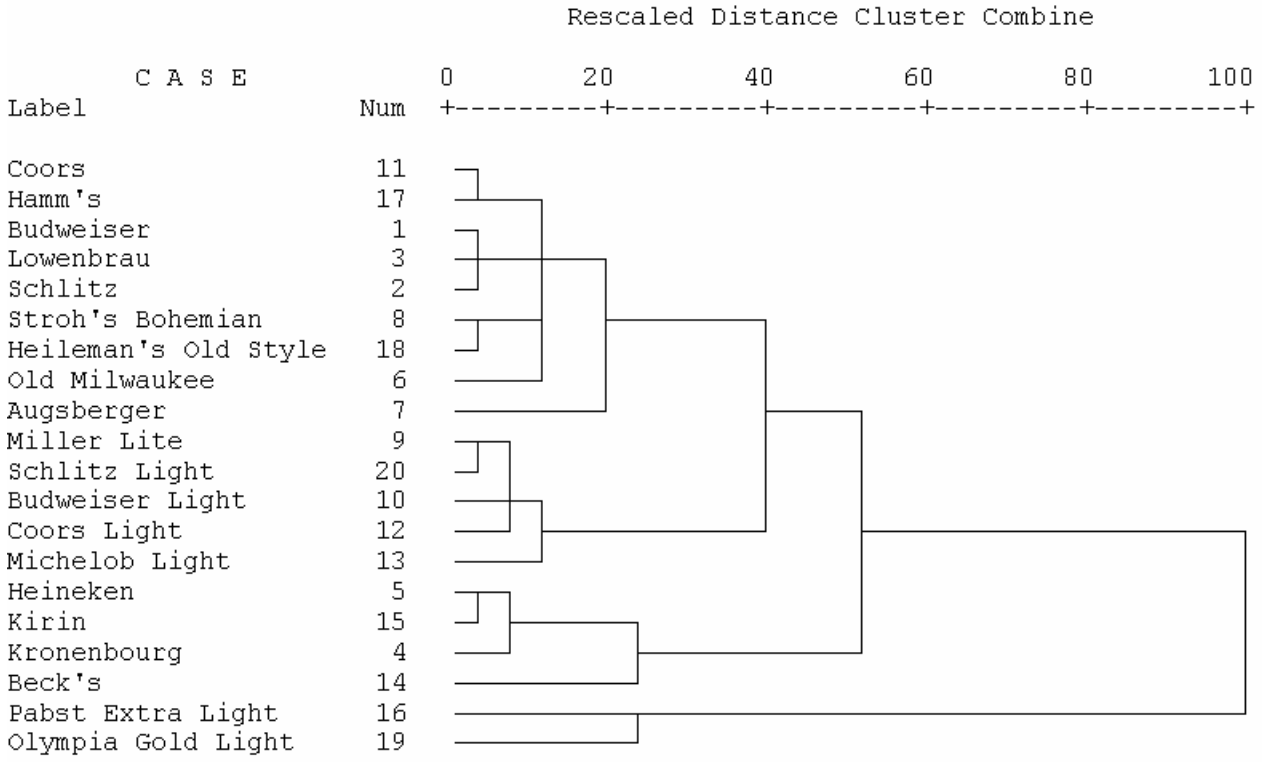
Clustering Example 2

The task of the researcher is to determine, based on knowledge about the product class or situation at hand (plus several graphical aids discussed below), how many clusters "make sense" in the end.

*Thus cluster analysis is **not** a tool for passing statistical judgment on the relative merits of different numbers of clusters, merely one for determining the clusters themselves.*

Clustering Example 2

Dendrogram using Centroid Method and Squared Euclidean Distances



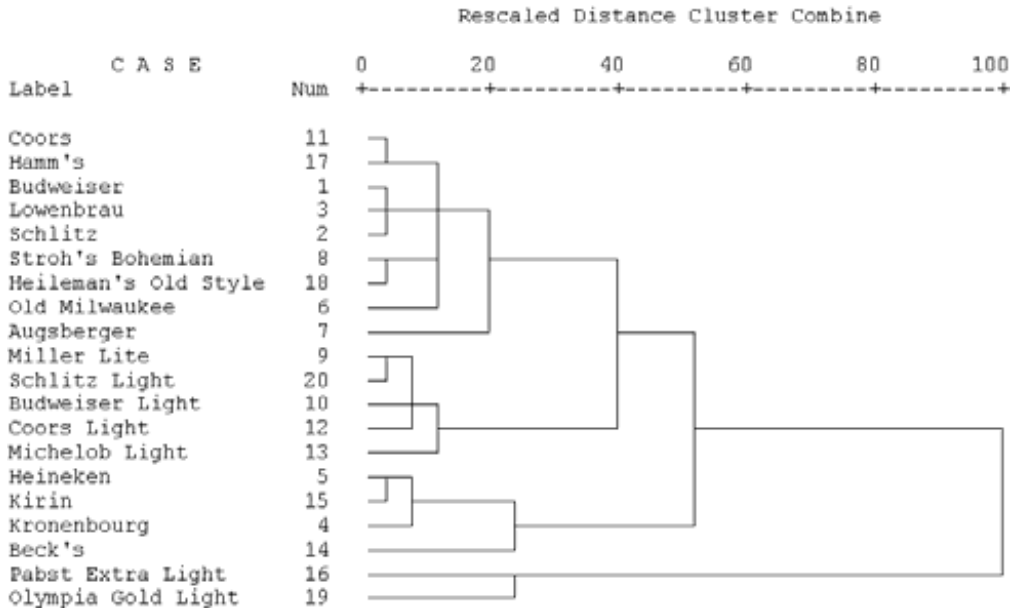
Clustering Example 2

Comment: *The graph in the previous slide – called a dendrogram– depicts, albeit in crude fashion, the hierarchical or "treelike" clustering solution. All the NCD values have been re-scaled so that they extend from 1 to 100 (easily accomplished by dividing all NCD values by the largest one, 13.405, and multiplying by 100), so that the biggest break in the set of items is at the rightmost side of the graph. This splits off Pabst Extra Light (16) and Olympia Gold Light (19) from the rest of the beers. Evidently, they are very different from all the others; let us examine the data for these two alone:*

Brand	Calories	Sodium	Alcohol	Price
Pabst Extra Light	68	15	2.3	0.38
Olympia Gold Light	72	6	2.9	0.46
Mean	132.4	15.0	4.4	0.50

Clustering Example 2

***Comment*:** As one proceeds across (from the right), one sees the "major" divisions in the graph, where larger clusters join up:



Thus, a "two cluster" solution would be {Pabst, Olympia} and {All Others}; a "three cluster" solution would be {Pabst, Olympia}, {Heineken, Kirin, Kronenbourg, Beck's} and {All Others}; etc.

Clustering Example 2

Whether there is any "obvious" way to discriminate these apparently different clusters is a task left to the researcher, perhaps with the aid of tools such as *Discriminant Analysis*.

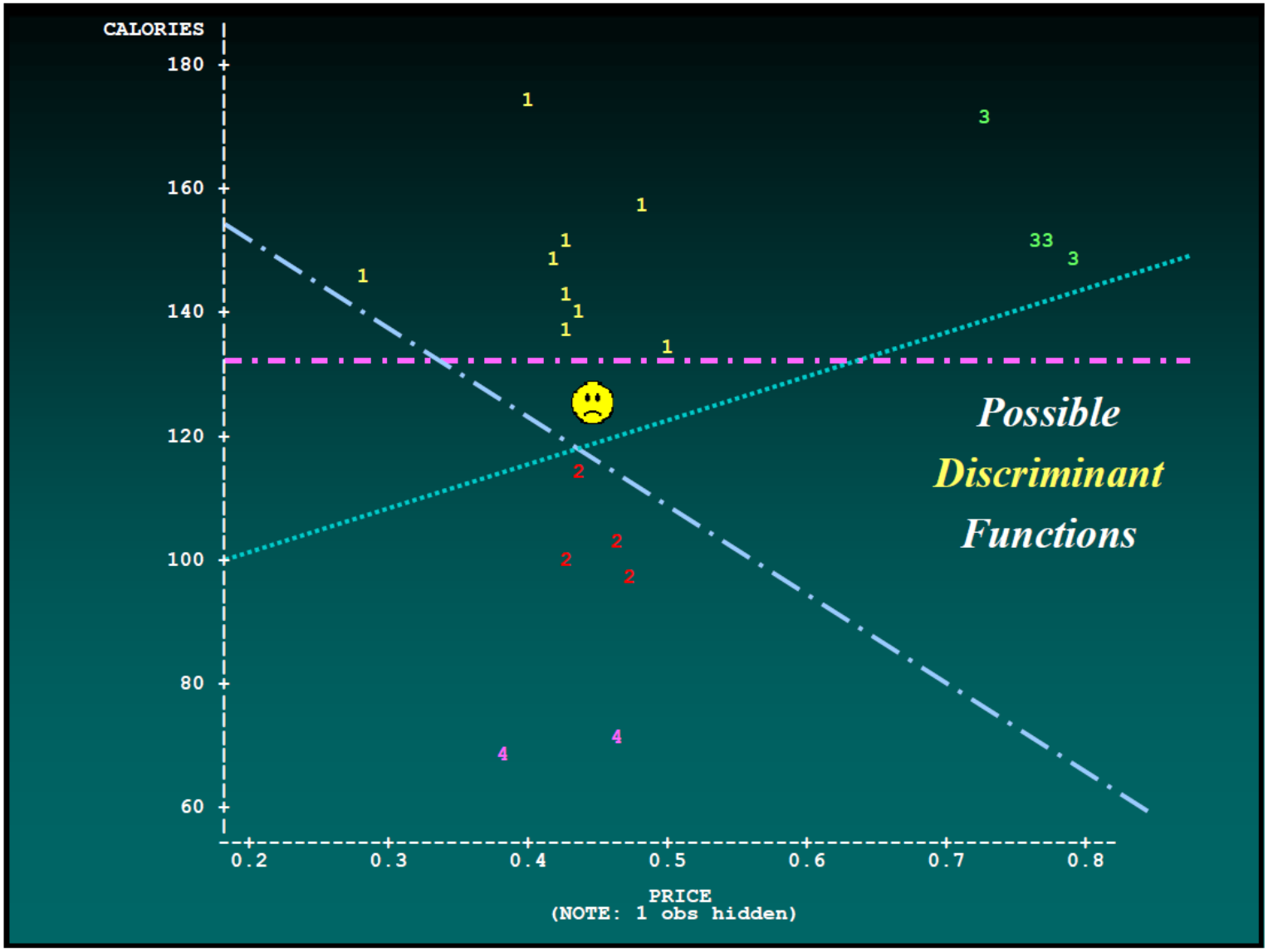
Is there any way to *explain* the resulting clustering? Recalling that the "three cluster" solution was {Pabst, Olympia}, {Heineken, Kirin, Kronenbourg, Beck's} and {All Others}, *examining* the beers themselves indicates that these are really the "super light", "mega heavy" and "all others" clusters.

What about the "four cluster solution"? Let's look…

Clustering Example 2

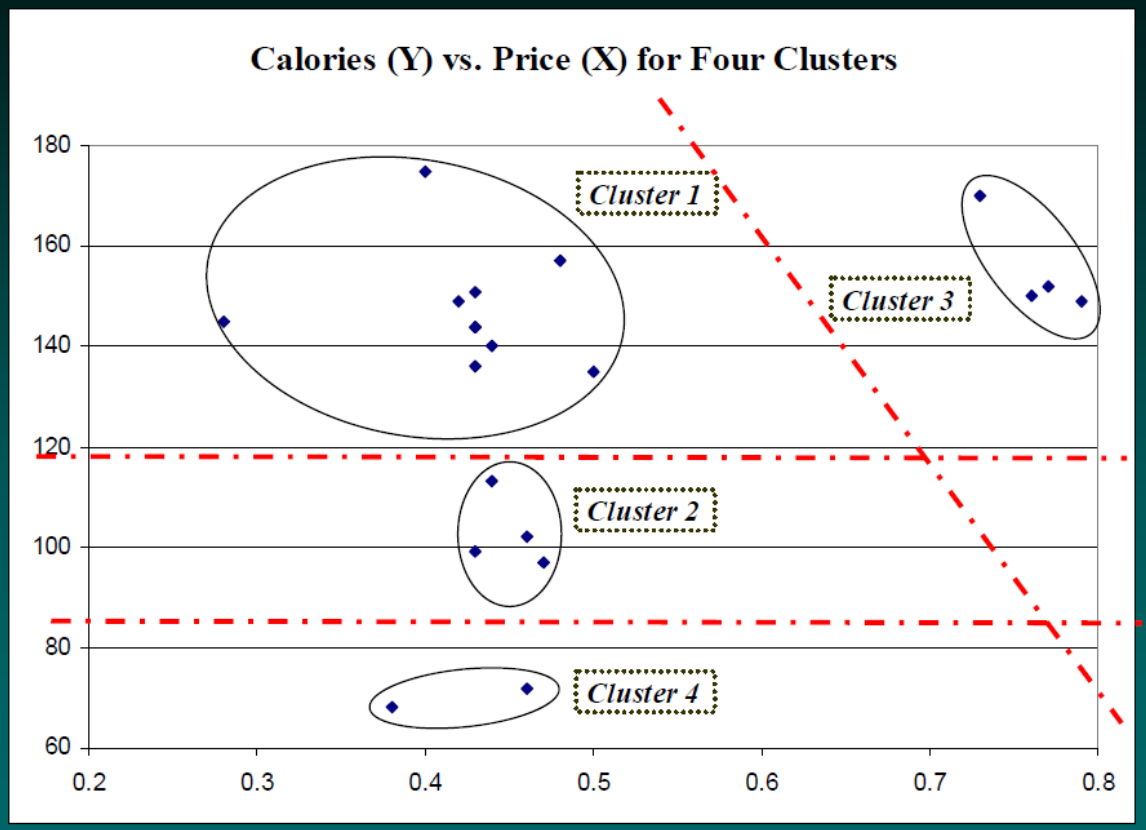
Brand	Cluster	Calories	Sodium	Alcohol	Price	Description
Coors	1	140	18	4.6	0.44	"Cheap High": High on Calories, Sodium, and Alcohol, Low on Price
Hamm's		136	19	4.4	0.43	
Budweiser		144	15	4.7	0.43	
Lowenbrau		157	15	4.9	0.48	
Schlitz		151	19	4.9	0.43	
Stroh's Bohemian Style		149	27	4.7	0.42	
Heileman's Old Style		144	24	4.9	0.43	
Old Milwaukee		145	23	4.6	0.28	
Augsberger		175	24	5.5	0.40	
Miller Lite	2	99	10	4.3	0.43	"Middle of the Road": Fairly Low on Calories, Sodium, Alcohol and Price
Schlitz Light		97	7	4.2	0.47	
Budweiser Light		113	8	3.7	0.44	
Coors Light		102	15	4.1	0.46	
Michelob Light		135	11	4.2	0.50	
Heineken	3	152	11	5.0	0.77	"Super Premium": High-Moderate Calories and Alcohol, Low Sodium, Very High Price
Kirin		149	6	5.0	0.79	
Kronenbourg		170	7	5.2	0.73	
Beck's		150	19	4.7	0.76	
Pabst Extra Light	4	68	15	2.3	0.38	"Ultra Lite": Very Low Calories, Sodium, Alcohol, Price
Olympia Gold Light		72	6	2.9	0.46	

Clustering Example 2



Clustering Example 2

Discriminating the Clusters (using only Calories and Price)



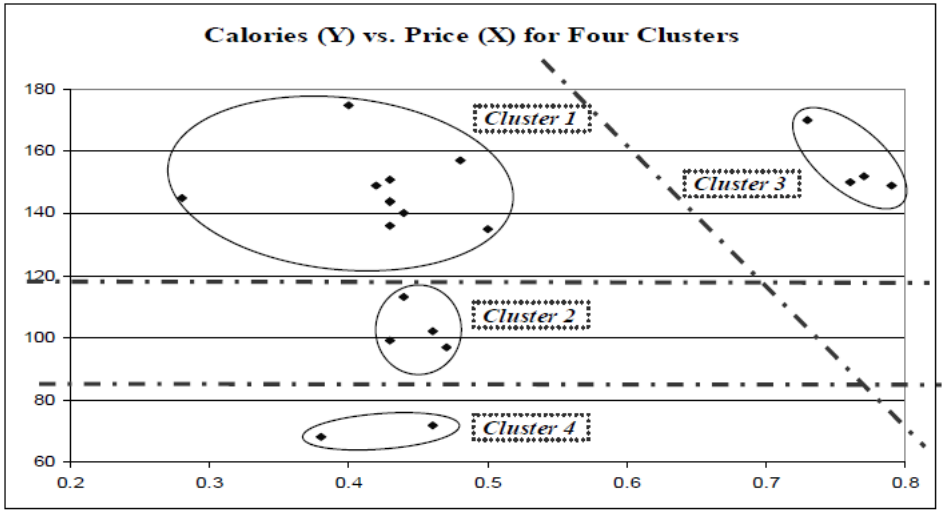
Clustering Example 2

Comment: *The graph above is a useful diagnostic aid in understanding the derived four-cluster solution; it is necessary to specify to the computer that the four-cluster solution be pictured, although* ***any*** *number of clusters, and their corresponding numbers, can be graphed.*

In this case, the researcher may have thought that the real "action", in terms of describing how the brands fall into groups, can be largely explained through just two of the four variables, Calories and Price; perhaps it was thought that sodium and alcohol content were not as able as these other two to discriminate brands in the beer market. It is therefore prudent to look at the derived four-cluster graph relative to these two variables.

Clustering Example 2

Comment: *It seems that clusters 1, 2 and 4 can be perfectly discriminated by Calories alone, without recourse to Price; recalling the strong split between light and non-light beers, this type of breakdown on the Calories variable makes intuitive sense. Only Cluster 3 appears to require some "trade off" between Price and Calories.*



Clustering Example 2

It would be misleading at this point to claim that Calories was the most important or, worse yet, the only important variable of the four, since we have considered the brands' ratings on only these two dimensions.

*A next step would be to look at the **four-cluster** solution using one or both of the omitted variables, alcohol and sodium content, to see how they fit into discriminating the clusters.*

A final step would be to look at more clusters, on various axes, to 'get a sense' of how the clustering works.

Clustering Example 2

Is there a NON-hierarchical Clustering Method?

YES: it's called **"K Means"**

Useful if you do *not* need to compare clustering solutions to one another (i.e., the 4- and 5-cluster solutions look completely different from one another; no 'structure')

K-Means is very simple to run and use. The researcher just tells the computer how many groups are required. Let's try 4, since it's such a nice number…

Clustering Example 2

The *Cluster Distances* arise directly from the Cluster Centers themselves by simple geometry (the square root of the sum-of-the-squares). Eyeballing them tells us that Cluster 2 (CL2) is very different from the other clusters, which are about equidistant:

	Cluster Center Distances			
	CL1	CL2	CL3	CL4
CL1	---			
CL2	2.38	---		
CL3	2.26	4.18	---	
CL4	2.84	4.84	2.84	---

Clustering Example 2

The *Cluster Centers* help tell us what the Clusters *mean*. Because the input variables were standardized, the values tell us *how many standard deviations above or below the mean* for that variable a Cluster's center lies:

	Cluster Centers			
	CL1	CL2	CL3	CL4
Calories	-0.77	-2.06	0.55	0.76
Sodium	-0.72	-0.68	0.83	-0.64
Alcohol	-0.45	-2.42	0.47	0.70
Price	-0.25	-0.53	-0.56	1.85
# in Cluster	5	2	9	4

Clustering Example 2

For example, Clusters 1 and 2 are both well below average on all four variables, but Cluster 2 is very, very far below (more than 2 standard deviations, which is extreme) on Calories and Alcohol.

Cluster 3 is above the mean on everything but Price, while Cluster 4 is above average on everything except Sodium; it is very highly priced, in fact, at 1.85 standard deviations above the mean for this sample of beers.

	Cluster Centers			
	CL1	CL2	CL3	CL4
Calories	-0.77	-2.06	0.55	0.76
Sodium	-0.72	-0.68	0.83	-0.64
Alcohol	-0.45	-2.42	0.47	0.70
Price	-0.25	-0.53	-0.56	1.85
# in Cluster	5	2	9	4

Clustering Example 2

It isn't difficult to see – start by comparing the number of beers in each cluster – that our K-Means solution is identical to the one obtained hierarchically.

This isn't a coincidence, nor is it the usual outcome: the hierarchical and non-hierarchical solutions often differ. *That they are identical should help us feel confident in the assignment of beers to these four groups*.

Because the two solutions are identical, we won't examine the K-means clustering in detail. Instead, let us look at the assignments obtained by running the K-Means analysis for K = 2, 3, 4, and 5 clusters:

Clustering Example 2

Comparing 2-, 3-, 4- and 5-Cluster K-Means Solutions									
		2-Cluster Solution		3-Cluster Solution		4-Cluster Solution		5-Cluster Solution	
Case	Brand	Cluster	Distance	Cluster	Distance	Cluster	Distance	Cluster	Distance
1	Budweiser	2	0.74	2	0.86	3	0.86	3	0.86
2	Schlitz	2	0.71	2	0.28	3	0.28	3	0.28
3	Lowenbrau	2	0.49	2	0.99	3	0.99	3	0.99
4	Kronenbourg	2	2.27	3	0.84	4	0.84	4	0.54
5	Heineken	2	1.98	3	0.13	4	0.13	4	0.50
6	Old Milwaukee	2	1.93	2	1.06	3	1.06	3	1.06
7	Augsberger	2	1.83	2	1.38	3	1.38	3	1.38
8	Stroh's Bohemian	2	1.68	2	1.01	3	1.01	3	1.01
9	Miller Lite	1	0.98	1	0.84	1	0.48	1	0.48
10	Budweiser Light	1	0.79	1	0.61	1	0.65	1	0.65
11	Coors	2	0.72	2	0.57	3	0.57	3	0.57
12	Coors Light	1	1.07	1	0.93	1	0.77	1	0.77
13	Michelob Light	2	1.32	1	1.46	1	0.92	1	0.92
14	Beck's	2	1.70	3	1.32	4	1.32	5	0.00
15	Kirin	2	2.52	3	0.78	4	0.78	4	0.45
16	Pabst Extra Light	1	2.05	1	2.23	2	0.84	2	0.84
17	Hamm's	2	0.99	2	0.72	3	0.72	3	0.72
18	Heileman's Old Style	2	1.26	2	0.59	3	0.59	3	0.59
19	Olympia Gold Light	1	1.29	1	1.48	2	0.84	2	0.84
20	Schlitz Light	1	0.98	1	0.87	1	0.65	1	0.65

Clustering Example 2

So… What does this all tell us?

Clustering is THE way to *rigorously break items into groups*: **SEGMENTATION**.

If you need to compare clustering solutions to one another, or want to know how big clusters "break up" into smaller ones: use **Hierarchical**.

If you just want to know the "best" 4 (or 11, or whatever) cluster solution: use **K-Means**.

Both relatively easy to use and interpret… and often give identical or near-identical answers.

K-Means: Overview

K-Means is an iterative clustering algorithm. It creates **partitions**.

K-Means Algorithm:

Initialize. User supplies K, the number of groups. Algorithm randomly selects K points as initial cluster centers.

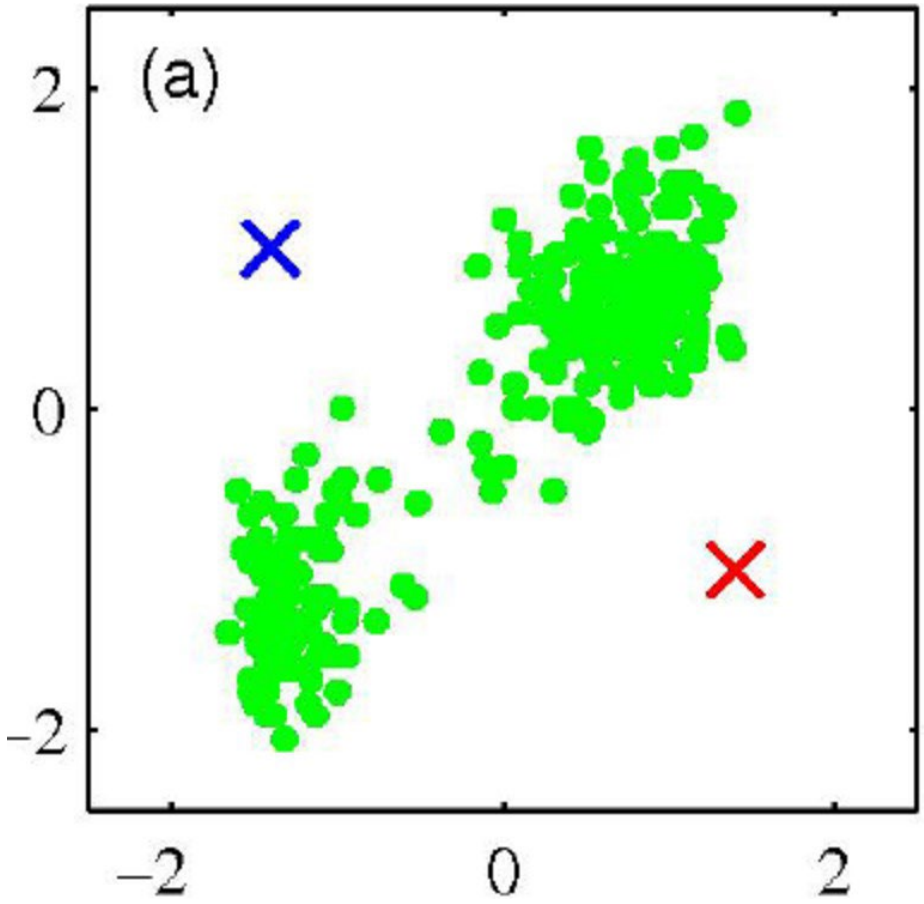
Alternate the following steps.

1. Assign each point to the *closest* cluster center.
2. Re-calculate the central point of each cluster.

- **Stop**. At any iteration when no points changes group assignment.

K-Means: Illustration

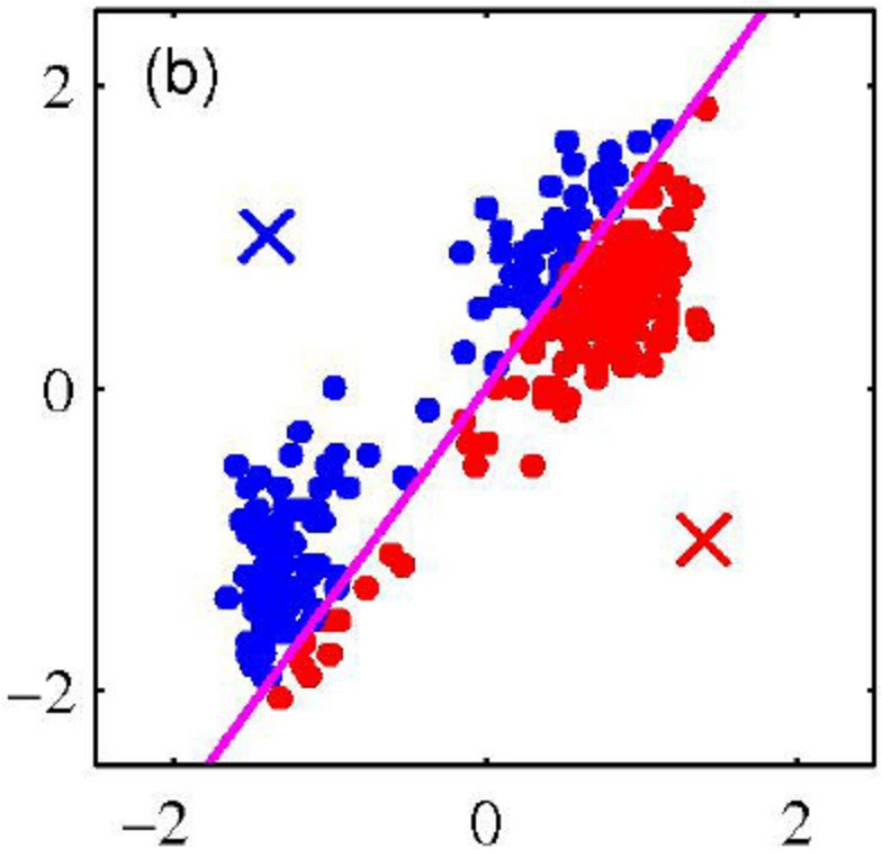
Initialize. User selects $K = 2$. Algorithm randomly picks two points.



[0]Images from David Sontag.

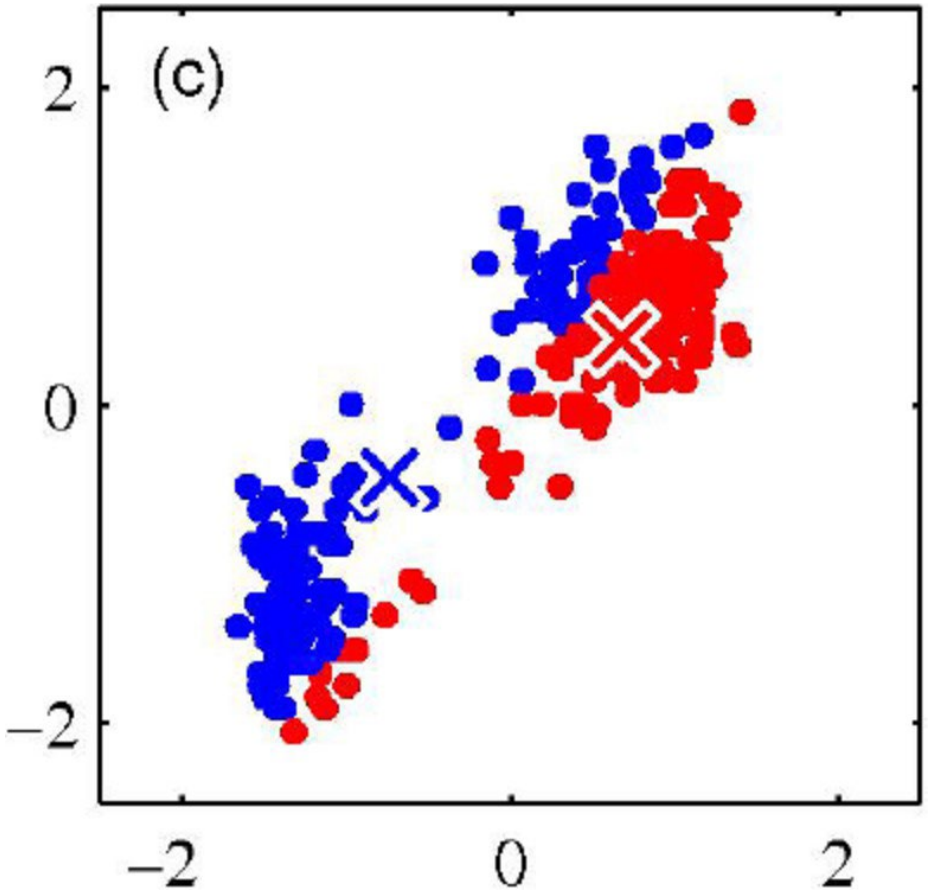
K-Means: Illustration

Alternate. Step 1. Assign Data Points to closest of the *K* points.



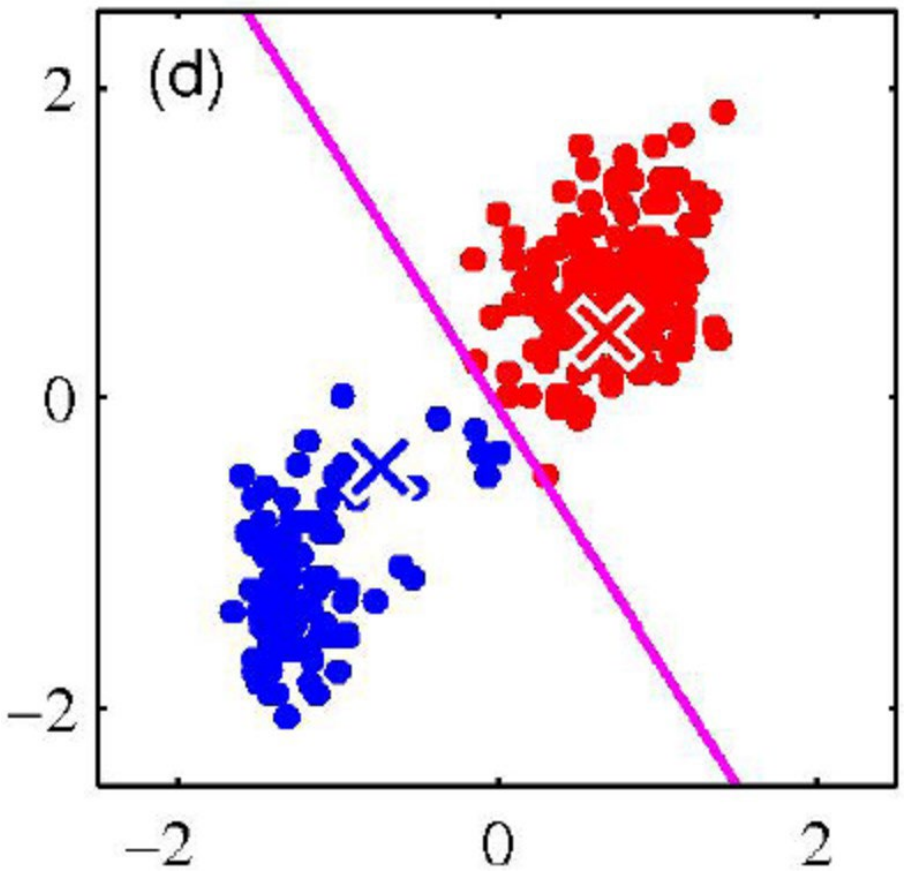
K-Means: Illustration

Alternate. Step 2. Recalculate the mean of each group (move the *x* to the location of the new mean).



K-Means: Illustration

Stop. When convergence occurs, i.e. no points change their classification.



Distance Measures

Necessary Properties of a Distance Measure:

Symmetric. For points A and B, $D(A, B) = D(B, A)$.

Positivity. $D(A, B) \geq 0$.

Self-Similarity. $D(A, B) = 0 \iff A = B$.

Triangle Inequality. $D(A, B) + D(B, C) \geq D(A, C)$.

Distance Measures

Euclidean distance meet the criteria:

$$d = \sqrt{[(x_2 - x_1)^2 + (y_2 - y)^2}$$

Other metrics are possible.

The base R function `kmeans` uses Euclidean distance. Other functions give choice of distance metric, e.g. `flexclust` package.

Algorithm

Objective function: $\min_{\mu} \min_{C} \sum_{i=1}^{k} \sum_{x \in C_i} (x - \mu_i)^2$

Now, follow the two steps described earlier:

1. Assign each point to the closest (measured via Euclidean distance) cluster center: $\min_{C} \sum_{i}^{n} (x - \mu_i)^2$
2. Re-calculate the center point of each cluster: $\min_{\mu} \sum_{i=1}^{k} \sum_{x \in C_i} (x - \mu_i)^2$

Clustering Example 2

Returning to the mall customer segmentation data described earlier.

```
>km.sol<-kmeans(Mall_Cust2_std, 4, nstart=25)
K-means clustering with 4 clusters of sizes 36, 61, 43, 60

Cluster means:
Age Annual.Income
1  0.07834805      1.4424080
2 -0.66171077      0.3756457
3 -0.81168059     -1.1750487
4  1.20743488     -0.4052331
```

Categorical Data

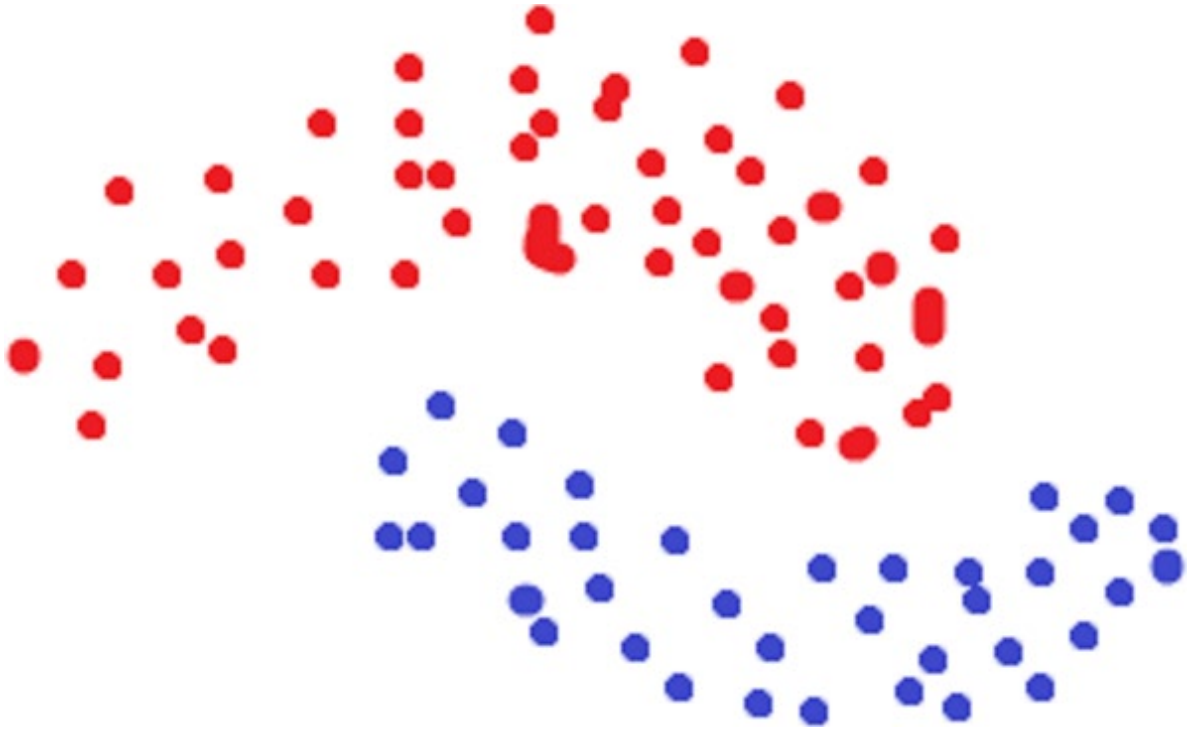
Nominal categorical variables won't work for K-means. What is the mean of such a variable indicate? For example, what does the mean of marital status indicate?

Therefore, can't really include these data.

Even binary data – already grouped. Standardizing doesn't change that.

K-Means Failures

Not all clusters are identified by K-means. For example, non-convex shaped clusters will not be found.



Summary

- Strengths of K-means
  1. Computationally efficient
  2. Requires little user input
  3. Usually identifies optimum
- Weaknesses of K-means
  1. Results sensitive to choice of K
  2. Might discover nonsense
  3. Not useful for all data (some shapes don't easily yield clusters)