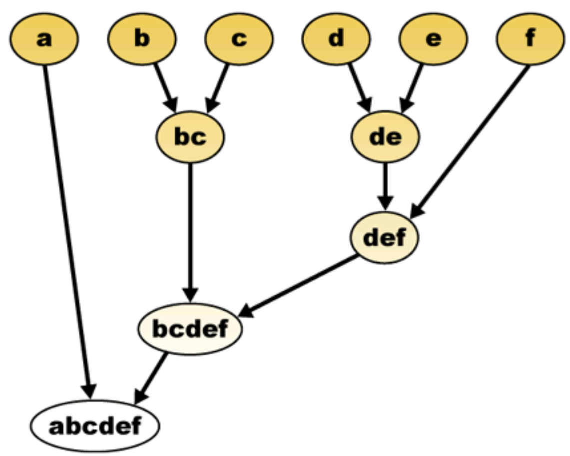
Hierarchical Clustering



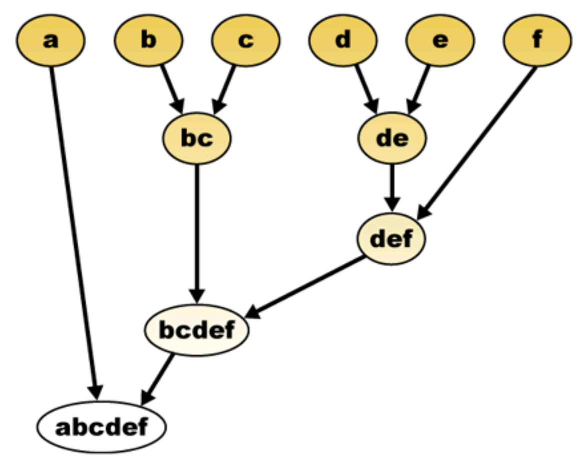
Hierarchical Clustering Algorithm

Start with *n* clusters (record = cluster)

Step 1: two closest records are merged into one cluster

At every step, pair of clusters with *smallest distance* are merged (either single record added to existing cluster, or two existing clusters are combined)

Requires a definition of **distance**



Pairwise distance between records

Single measurement case: Each record has 1 value.

Multiple measurement case : Each record has a multiple values.

$$d_{ij} = \text{distance between observations } i \text{ and } j$$

Distance Requirements:

- Non-negative ($d_{ij} > 0$)
- $d_{ii} = 0$
- Symmetry ($d_{ij} = d_{ji}$)
- Triangle inequality ($d_{ij} + d_{jk} \geq d_{ik}$)

Distance between any pair cannot exceed the sum of distances between the other two pairs

UG Business Programs

Universities Clustering.xls

Data for 25 undergraduate programs at business schools in US universities in 1995.

This dataset excludes **image variables** (student satisfaction, employer satisfaction, deans' opinions)

Student quality Program Placement

Univ	SAT	Top10	Accept	SFRatio	Expenses	GradRate
Brown	1310	89	22	13	22,704	94
CalTech	1415	100	25	6	63,575	81
CMU	1260	62	59	9	25,026	72
Columbia	1310	76	24	12	31,510	88
Cornell	1280	83	33	13	21,864	90
Dartmouth	1340	89	23	10	32,162	95
Duke	1315	90	30	12	31,585	95
Georgetown	1255	74	24	12	20,126	92
Harvard	1400	91	14	11	39,525	97
JohnsHopkins	1305	75	44	7	58,691	87
MIT	1380	94	30	10	34,870	91
Northwestern	1260	85	39	11	28,052	89
NotreDame	1255	81	42	13	15,122	94
PennState	1081	38	54	18	10,185	80
Princeton	1375	91	14	8	30,220	95
Purdue	1005	28	90	19	9,066	69
Stanford	1360	90	20	12	36,450	93
TexasA&M	1075	49	67	25	8,704	67
UCBerkeley	1240	95	40	17	15,140	78
UChicago	1290	75	50	13	38,380	87
UMichigan	1180	65	68	16	15,470	85
UPenn	1285	80	36	11	27,553	90
UVA	1225	77	44	14	13,349	92
UWisconsin	1085	40	69	15	11,857	71
Yale	1375	95	19	11	43,514	96

Notation:

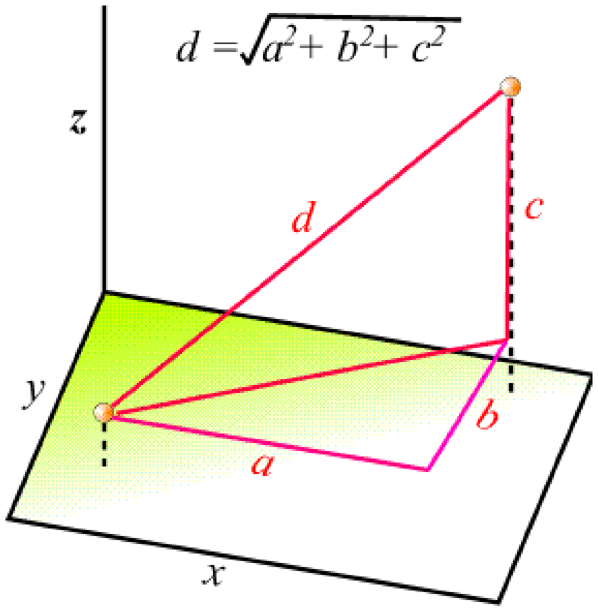
$$x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$$

$$x_j = (x_{j1}, x_{j2}, \ldots, x_{jp})$$

Caltech = (1415, 100, 25, 6, 63575, 81)

Cornell = (1280, 83, 33, 13, 21864, 90)

Euclidean Distance



$$d_{ij} = \sqrt{\left(x_{i1} - x_{j1}\right)^2 + \left(x_{i2} - x_{j2}\right)^2 + \cdots + \left(x_{ip} - x_{jp}\right)^2}$$

6-dimensional Euclidean distance between Caltech and Cornell:

$\sqrt{}$ [$(1415\text{-}1280)^2$ + $(100\text{-}83)^2$ + $(25\text{-}33)^2$ + $(6\text{-}13)^2$ +

+ $(63575\text{-}21864)^2$ + $(81\text{-}90)^2$] = 41,711.22

Standardize if multiple variables (p>1)

Euclidean distance is influenced by the **units** of the different measurements

Solution: standardize (=normalize) each variable before measuring distances

Standardizing: Example

$$Z\_SAT = \frac{SAT - mean(SAT)}{std(SAT)}$$

Univ	Z_SAT	Z_Top10	Z_Accept	Z_SFRatio	Z_Expenses	Z_GradRate
Brown	0.401994	0.644235	-0.871888	0.068840897	-0.32471667	0.80372917
CalTech	1.370988	1.210256	-0.719814	-1.65218153	2.508651168	-0.631501491
CMU	-0.059432	-0.74509	1.003685	-0.91460049	-0.16374483	-1.625122718
Columbia	0.401994	-0.024699	-0.770506	-0.17701945	0.285756214	0.141315019
Cornell	0.125139	0.335496	-0.314285	0.068840897	-0.38294938	0.362119736
Dartmouth	0.67885	0.644235	-0.821197	-0.66874014	0.330955887	0.914131529
Duke	0.448137	0.695691	-0.466359	-0.17701945	0.290955563	0.914131529
Georgetown	-0.105574	-0.127612	-0.770506	-0.17701945	-0.50343562	0.582924453

Euclidean distance between standardized Caltech and Cornell:

$$\surd[\,(1.371-1.125)^2 + (1.210-0.335)^2 + \ldots + (-.632-.362)^2]$$
$$= 3.84$$

Lots of other distance metrics

Statistical (Mahalanobis) distance

Uses correlation matrix

Manhattan distance

$$d_{ij} = \left|x_{i1} - x_{j1}\right| + \left|x_{i2} - x_{j2}\right| + \cdots + \left|x_{ip} - x_{jp}\right|$$

Distances for Binary Data

Similarity-based metrics based on 2x2 table of counts

	0	**1**
0	a	b
1	c	d

	Married?	**Smoker?**	**Manager?**
Carrie	Y	Y	Y
Sam	N	Y	N
Miranda	N	N	Y

		Miranda	
Carrie		**N**	**Y**
	N	0	0
	Y	2	1

- Binary Euclidean Distance: *(b+c)/(a+b+c+d)*
- Simple matching Coefficient: *(a+d)/(a+b+c+d)*
- Jaquard's coefficient: *d/(b+c+d)*

For >2 categories, distance =0 only if both items have same category. Otherwise =1.

Distances for Mixed (numerical + categorical) Data

Simple: standardize numerical variables to [0,1], then use Euclidian distance for all

Gower's General Dissimilarity Coefficient

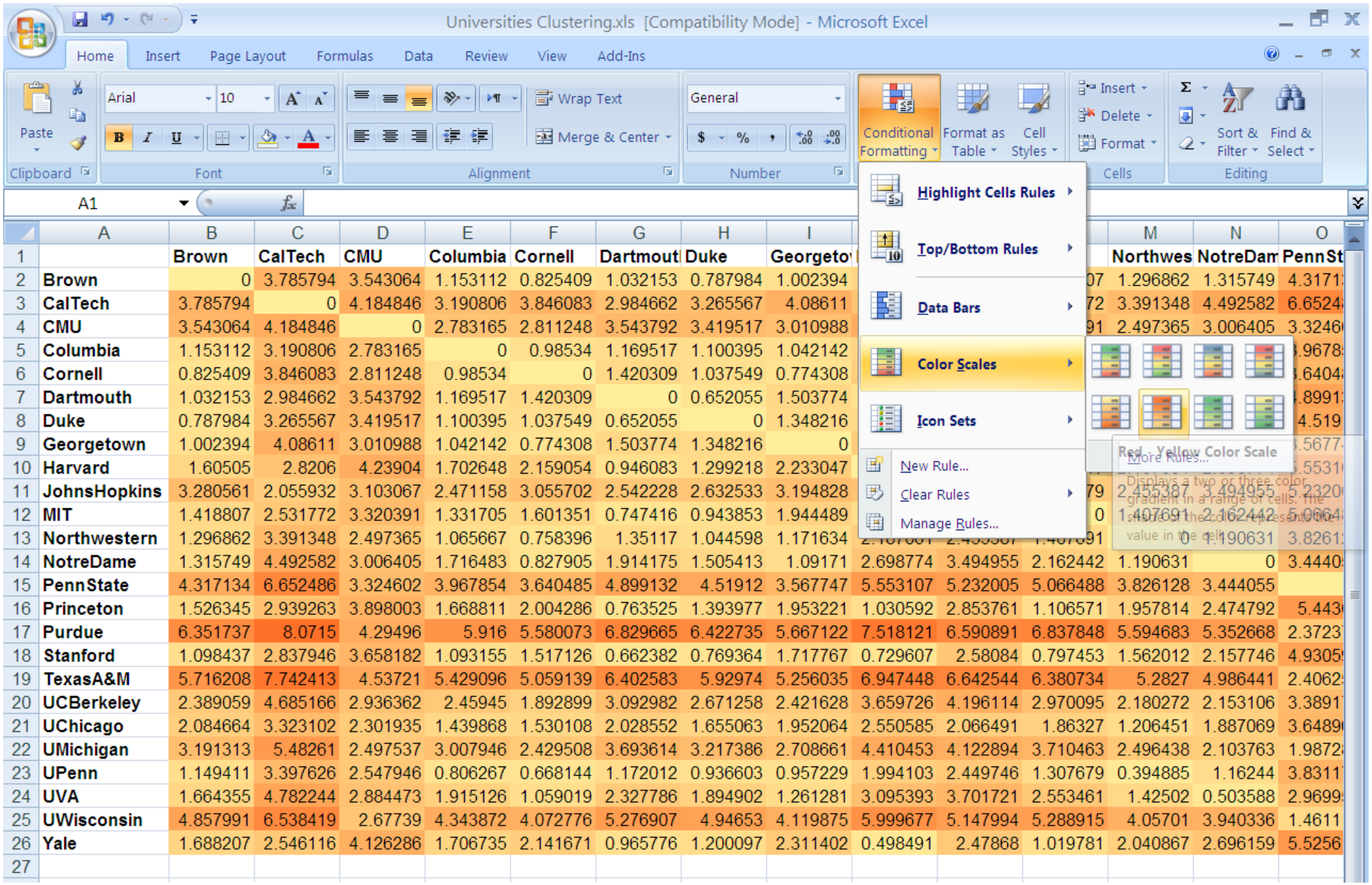
$$\mathbf{d_{ij} = \Sigma_k w_{ijk} d_{ijk} / \Sigma_k w_{ijk}}$$

$\mathbf{d_{ijk}}$ = distance provided by *k*th variable.
$\mathbf{w_{ijk}}$ = usually 1 or 0 depending whether or not the comparison is valid for the *k*th variable.

More on Gower's measure for mixed data:
www.soziologie.wiso.unierlangen.de/koeln/script/chap6.pdf

Worksheet: Euclid Distance Matrix

Distance Matrix



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1		Brown	CalTech	CMU	Columbia	Cornell	Dartmout	Duke	Georgeto	[illegible]	[illegible]	[illegible]	Northwes	NotreDam	PennSt
2	Brown	0	3.785794	3.543064	1.153112	0.825409	1.032153	0.787984	1.002394	[illegible]	[illegible]	[illegible]	1.296862	1.315749	4.3171
3	CalTech	3.785794	0	4.184846	3.190806	3.846083	2.984662	3.265567	4.08611	[illegible]	[illegible]	[illegible]	3.391348	4.492582	6.6524
4	CMU	3.543064	4.184846	0	2.783165	2.811248	3.543792	3.419517	3.010988	[illegible]	[illegible]	[illegible]	2.497365	3.006405	3.3246
5	Columbia	1.153112	3.190806	2.783165	0	0.98534	1.169517	1.100395	1.042142	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	.9678
6	Cornell	0.825409	3.846083	2.811248	0.98534	0	1.420309	1.037549	0.774308	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	.6404
7	Dartmouth	1.032153	2.984662	3.543792	1.169517	1.420309	0	0.652055	1.503774	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	.8991
8	Duke	0.787984	3.265567	3.419517	1.100395	1.037549	0.652055	0	1.348216	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	4.519
9	Georgetown	1.002394	4.08611	3.010988	1.042142	0.774308	1.503774	1.348216	0	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	.5677
10	Harvard	1.60505	2.8206	4.23904	1.702648	2.159054	0.946083	1.299218	2.233047	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	.5531
11	JohnsHopkins	3.280561	2.055932	3.103067	2.471158	3.055702	2.542228	2.632533	3.194828	[illegible]	[illegible]	79	2.455387	3.494955	5.2320
12	MIT	1.418807	2.531772	3.320391	1.331705	1.601351	0.747416	0.943853	1.944489	[illegible]	[illegible]	0	1.407691	2.162442	5.0664
13	Northwestern	1.296862	3.391348	2.497365	1.065667	0.758396	1.35117	1.044598	1.171634	[illegible]	[illegible]	[illegible]	0	1.190631	3.8261
14	NotreDame	1.315749	4.492582	3.006405	1.716483	0.827905	1.914175	1.505413	1.09171	2.698774	3.494955	2.162442	1.190631	0	3.4440
15	PennState	4.317134	6.652486	3.324602	3.967854	3.640485	4.899132	4.51912	3.567747	5.553107	5.232005	5.066488	3.826128	3.444055	
16	Princeton	1.526345	2.939263	3.898003	1.668811	2.004286	0.763525	1.393977	1.953221	1.030592	2.853761	1.106571	1.957814	2.474792	5.443
17	Purdue	6.351737	8.0715	4.29496	5.916	5.580073	6.829665	6.422735	5.667122	7.518121	6.590891	6.837848	5.594683	5.352668	2.3723
18	Stanford	1.098437	2.837946	3.658182	1.093155	1.517126	0.662382	0.769364	1.717767	0.729607	2.58084	0.797453	1.562012	2.157746	4.9305
19	TexasA&M	5.716208	7.742413	4.53721	5.429096	5.059139	6.402583	5.92974	5.256035	6.947448	6.642544	6.380734	5.2827	4.986441	2.4062
20	UCBerkeley	2.389059	4.685166	2.936362	2.45945	1.892899	3.092982	2.671258	2.421628	3.659726	4.196114	2.970095	2.180272	2.153106	3.3891
21	UChicago	2.084664	3.323102	2.301935	1.439868	1.530108	2.028552	1.655063	1.952064	2.550585	2.066491	1.86327	1.206451	1.887069	3.6489
22	UMichigan	3.191313	5.48261	2.497537	3.007946	2.429508	3.693614	3.217386	2.708661	4.410453	4.122894	3.710463	2.496438	2.103763	1.9872
23	UPenn	1.149411	3.397626	2.547946	0.806267	0.668144	1.172012	0.936603	0.957229	1.994103	2.449746	1.307679	0.394885	1.16244	3.8311
24	UVA	1.664355	4.782244	2.884473	1.915126	1.059019	2.327786	1.894902	1.261281	3.095393	3.701721	2.553461	1.42502	0.503588	2.9699
25	UWisconsin	4.857991	6.538419	2.67739	4.343872	4.072776	5.276907	4.94653	4.119875	5.999677	5.147994	5.288915	4.05701	3.940336	1.4611
26	Yale	1.688207	2.546116	4.126286	1.706735	2.141671	0.965776	1.200097	2.311402	0.498491	2.47868	1.019781	2.040867	2.696159	5.5256
27															

Feed matrix into hierarchical clustering algorithm

Once Again: The Hierarchical Clustering Algorithm

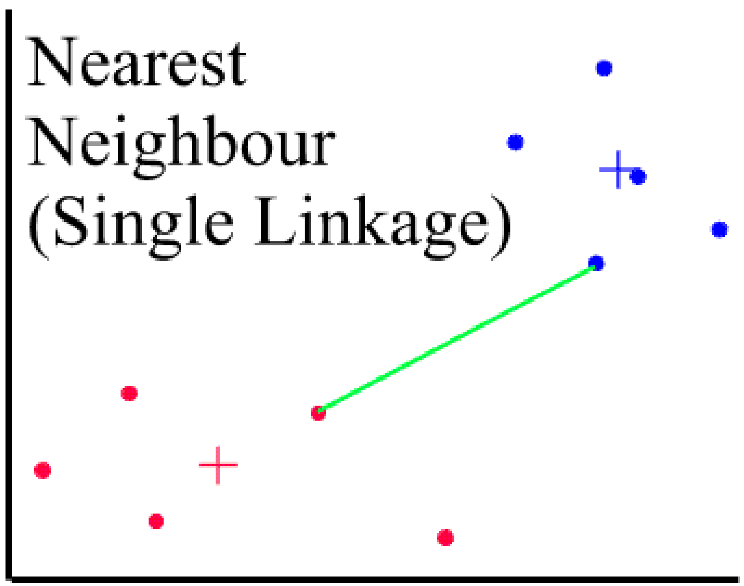
- ✓ Start with *n* clusters (record= cluster)
- ✓ Step 1: two closest records are merged into one cluster

At every step, pair of clusters with *smallest distance* are merged (either single record added to existing cluster, or two existing clusters are combined)

How to measure distances **between clusters**?

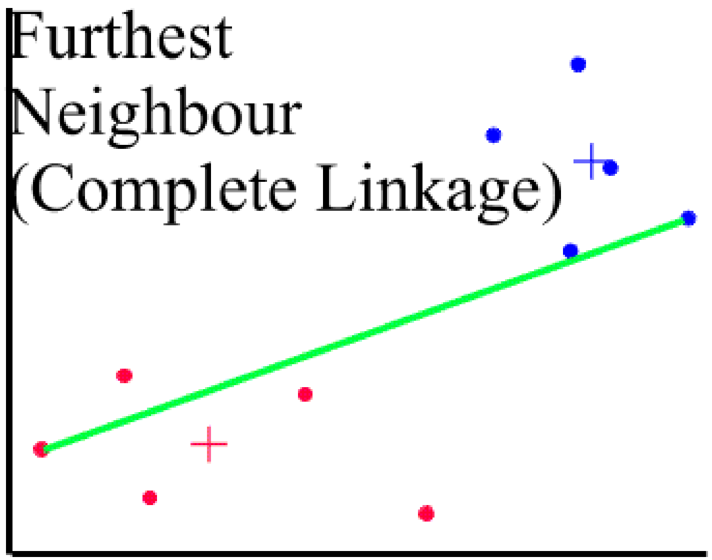
Distances Between Clusters: 'single linkage' ('nearest neighbor')

Distance between 2 clusters = **minimum distance** between members of the two clusters



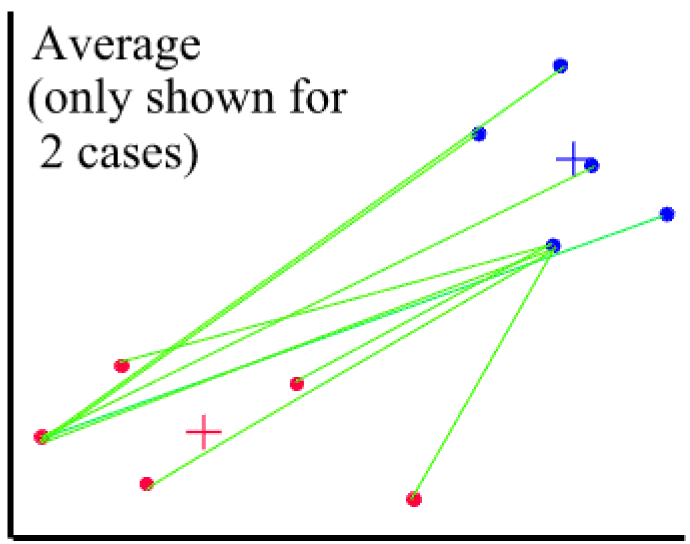
Distances Between Clusters: ‘complete linkage’ (‘farthest neighbor’)

Distance between 2 clusters = **greatest distance** between members of the two clusters



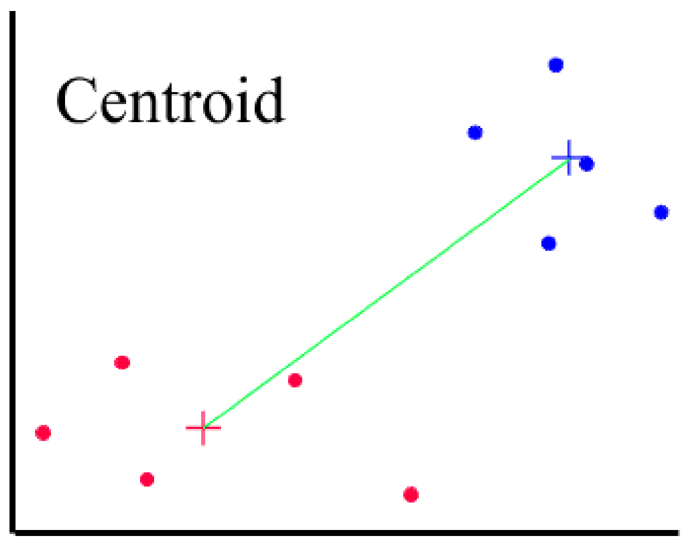
Distances Between Clusters: ‘average linkage’

Distance between 2 clusters = **average** of all distances between members of the two clusters



Distances Between Clusters: ‘centroid linkage’

Distance between 2 clusters =
distance between their **centroids** (centers)



And Again:
The Hierarchical Clustering Algorithm

- ✓ Start with *n* clusters (record = cluster)
- ✓ Step 1: two closest records are merged into one cluster

At every step, pair of clusters with *smallest distance* are merged.

At this point the distance matrix is re-computed:

- **Two rows+columns are merged into single row+column**
- **Distances to the newly merged cluster are recalculated**

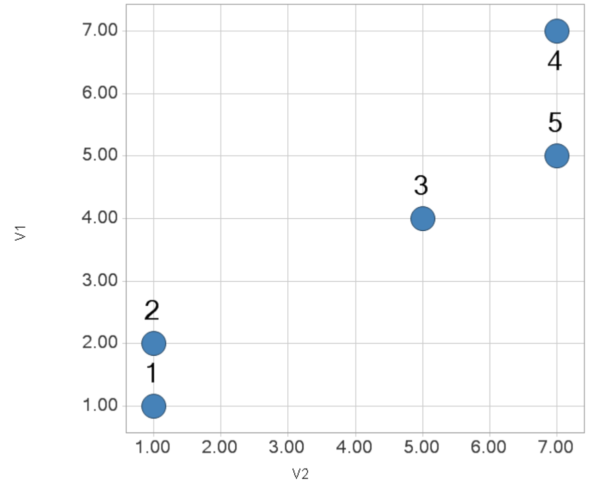
Repeat the last step until a single cluster is formed

The clustering process: example

(from http://obelia.jde.aca.mmu.ac.uk/multivar/dend.htm - no longer)

Two variables, n=5 items:

item	**v1**	**v2**
1	**1**	**1**
2	**2**	**1**
3	**4**	**5**
4	**7**	**7**
5	**5**	**7**



Euclidean distance matrix

	1	2	3	4	5
1	0.0				
2	1.0	0.0			
3	5.0	4.5	0.0		
4	8.5	7.8	3.6	0.0	
5	7.2	6.7	2.2	2.0	0.0

What happens next?

- Merge 1&2 into cluster A
- Use single linkage to compute distances from cluster A:

	1	2	3	4	5
1	0.0				
2	1.0	0.0			
3	5.0	4.5	0.0		
4	8.5	7.8	3.6	0.0	
5	7.2	6.7	2.2	2.0	0.0

→

	A	3	4	5
A	0.0			
3	4.5	0.0		
4	7.8	3.6	0.0	
5	6.7	2.2	2.0	0.0

What happens next?

Merge 4&5 (cluster B)

	A	3	B
A	0.0		
3	4.5	0.0	
B	6.7	2.2	0.0

→

Merge 3 & B

	A	B
A	0.0	
B	4.5	0.0

Finally: Summarize process in a **Dendrogram**