Exercise #4

Introduction to Knowledge Management, WS2016

Camilo Thorne

(based on slides by T. Szyler and A. Melo)

Mannheim, 04.04.2016

Exercise 04 - Question 1

Supervised vs Unsupervised Learning

1. If we want to use supervised learning, what property is mandatory for the input data?

Exercise 04 - Question 1

Supervised vs Unsupervised Learning

1. If we want to use supervised learning, what property is mandatory for the input data?
   - It must be labeled!

Exercise 04 - Question 1

Supervised vs Unsupervised Learning

1. If we want to use supervised learning, what property is mandatory for the input data?
   - It must be labeled!
2. What do you think is generally cheaper: Creating a data set for supervised learning or for unsupervised learning?

Exercise 04 - Question 1

Supervised vs Unsupervised Learning

1. If we want to use supervised learning, what property is mandatory for the input data?
   - It must be labeled!
2. What do you think is generally cheaper: Creating a data set for supervised learning or for unsupervised learning?
   - Labeling data by experts is usually more expensive than collecting unlabeled data

Exercise 04 - Question 1

Supervised vs Unsupervised Learning

1. If we want to use supervised learning, what property is mandatory for the input data?
   - It must be labeled!
2. What do you think is generally cheaper: Creating a data set for supervised learning or for unsupervised learning?
   - Labeling data by experts is usually more expensive than collecting unlabeled data
3. Do you expect unsupervised or supervised learning to produce better results? Why?

Exercise 04 - Question 1

Supervised vs Unsupervised Learning

1. If we want to use supervised learning, what property is mandatory for the input data?
   - It must be labeled!
2. What do you think is generally cheaper: Creating a data set for supervised learning or for unsupervised learning?
   - Labeling data by experts is usually more expensive than collecting unlabeled data
3. Do you expect unsupervised or supervised learning to produce better results? Why?
   - Depends on the learning task. Usually supervised learning provides more accurate results while unsupervised learning often is more flexible to deal with real world data

Exercise 04 - Question 2

Agglomerative Hierarchical Clustering

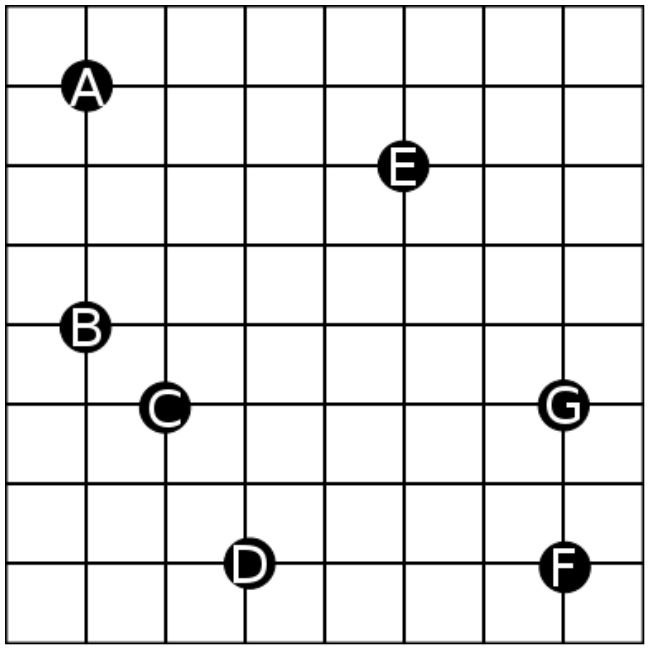


Figure : Data to be clustered

$$\Delta(c_i, c_j) =$$

Complete Link

$$\max_{\vec{x} \in c_i, \vec{y} \in c_j} distance(\vec{x}, \vec{y})$$

Single Link

$$\min_{\vec{x} \in c_i, \vec{y} \in c_j} distance(\vec{x}, \vec{y})$$

$distance(\vec{x}, \vec{y})$:
Euclidian distance of $\vec{x}$ to $\vec{y}$

Exercise 04 - Question 2

Agglomerative Hierarchical Clustering

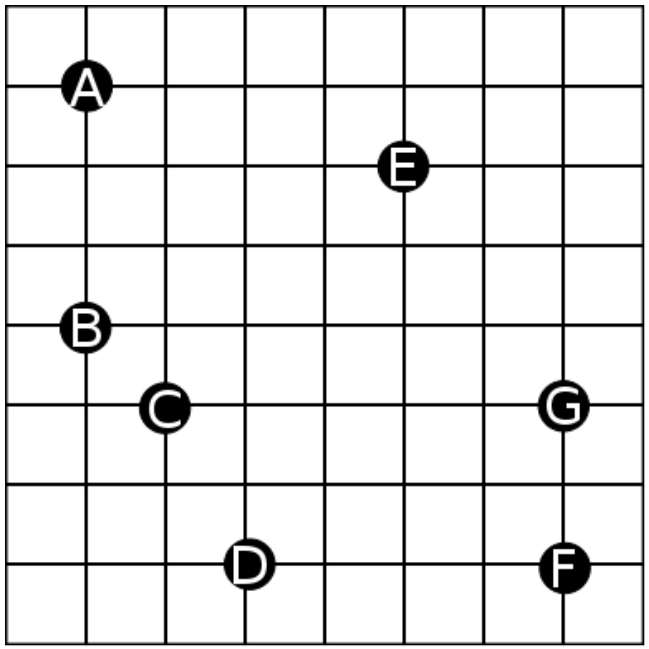


Figure : Data to be clustered

C_{BC}, C_{FG}, C_{BCD},
C_{ABCD}, C_{EFG}, $C_{ABCDEFG}$

$$\Delta(c_i, c_j) =$$

Complete Link

$$\max_{x \in c_i, y \in c_j} distance(x, y)$$

	A	B	C	D	E	F	G
A	×	3	4.12	6.32	4.12	8.49	7.21
B		×	1.41	3.61	4.47	6.71	6.08
C			×	2.24	4.24	5.39	5
D				×	5.39	4	4.47
E					×	5.39	3.61
F						×	2
G							×

Exercise 04 - Question 2

Agglomerative Hierarchical Clustering

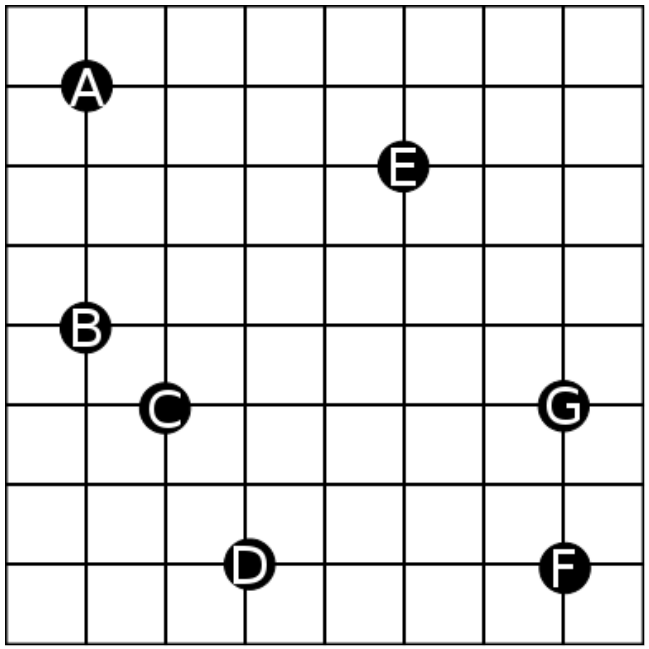


Figure : Data to be clustered

$$\Delta(c_i, c_j) =$$

Single Link

$$\min_{x \in c_i, y \in c_j} distance(x, y)$$

	A	B	C	D	E	F	G
A	×	3	4.12	6.32	4.12	8.49	7.21
B		×	1.41	3.61	4.47	6.71	6.08
C			×	2.24	4.24	5.39	5
D				×	5.39	4	4.47
E					×	5.39	3.61
F						×	2
G							×

Exercise 04 - Question 2

Agglomerative Hierarchical Clustering

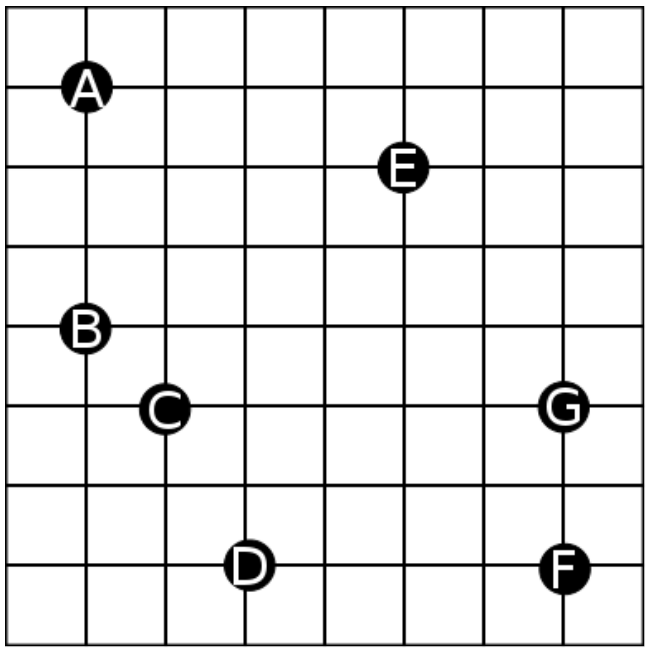


Figure : Data to be clustered

Step 1: Look for the smallest distance

	A	B	C	D	E	F	G
A	×	3	4.12	6.32	4.12	8.49	7.21
B		×	1.41	3.61	4.47	6.71	6.08
C			×	2.24	4.24	5.39	5
D				×	5.39	4	4.47
E					×	5.39	3.61
F						×	2
G							×

$\Rightarrow$ **reason over distance matrix!**

Exercise 04 - Question 2

Agglomerative Hierarchical Clustering

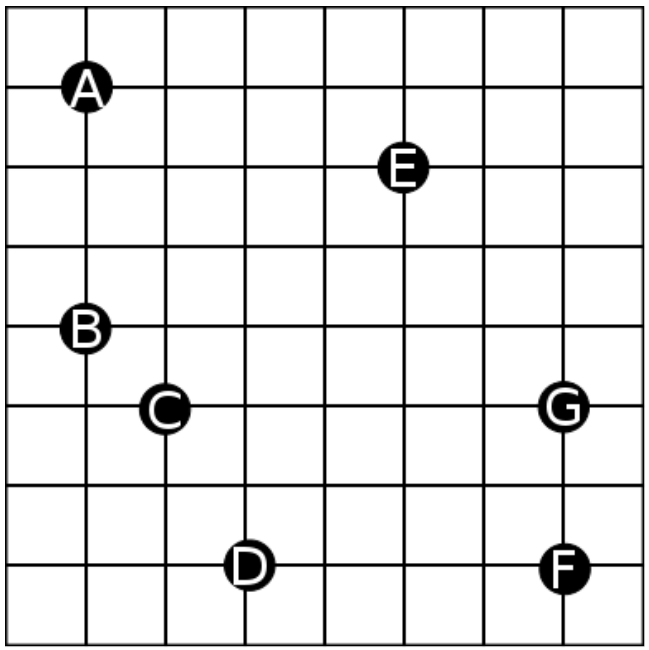


Figure : Data to be clustered

Result: C_{BC}

Step 1: Look for the smallest distance

	A	B	C	D	E	F	G
A	×	3	4.12	6.32	4.12	8.49	7.21
B		×	1.41	3.61	4.47	6.71	6.08
C			×	2.24	4.24	5.39	5
D				×	5.39	4	4.47
E					×	5.39	3.61
F						×	2
G							×

$\Rightarrow$ **reason over distance matrix!**

Exercise 04 - Question 2

Agglomerative Hierarchical Clustering

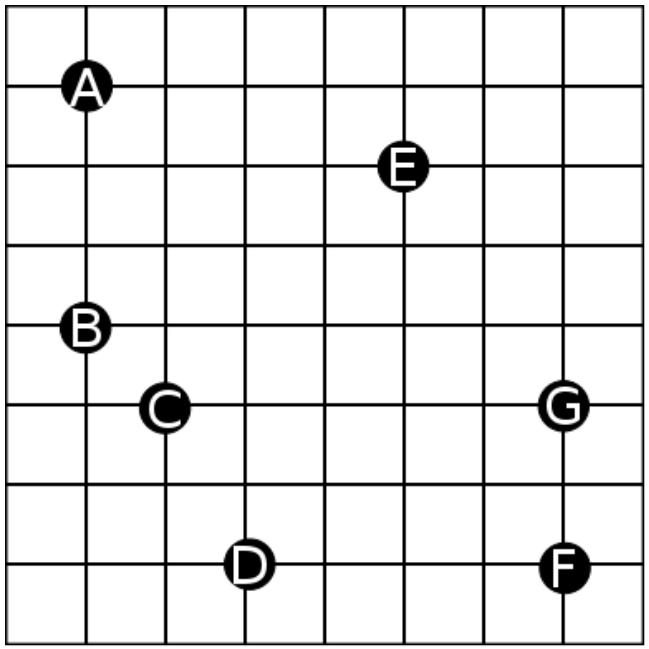


Figure : Data to be clustered

Result: C_{BC}

Step 2: Where is the smallest distance now?

	A	B C	D	E	F	G
A	×	4.12	6.32	4.12	8.49	7.21
B C		×	3.61	4.47	6.71	6.08
D			×	5.39	4	4.47
E				×	5.39	3.61
F					×	2
G						×

$\Rightarrow$ **reason over distance matrix!**

Exercise 04 - Question 2

Agglomerative Hierarchical Clustering

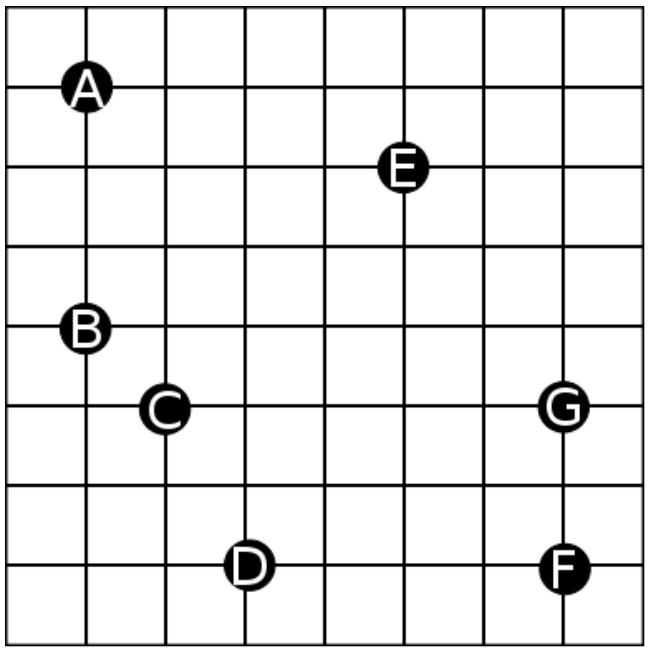


Figure : Data to be clustered

Result: C_{BC}, C_{FG}

Step 2: Where is the smallest distance now?

	A	B C	D	E	F	G
A	×	4.12	6.32	4.12	8.49	7.21
B C		×	3.61	4.47	6.71	6.08
D			×	5.39	4	4.47
E				×	5.39	3.61
F					×	2
G						×

⇒ **reason over distance matrix!**

Exercise 04 - Question 2

Agglomerative Hierarchical Clustering

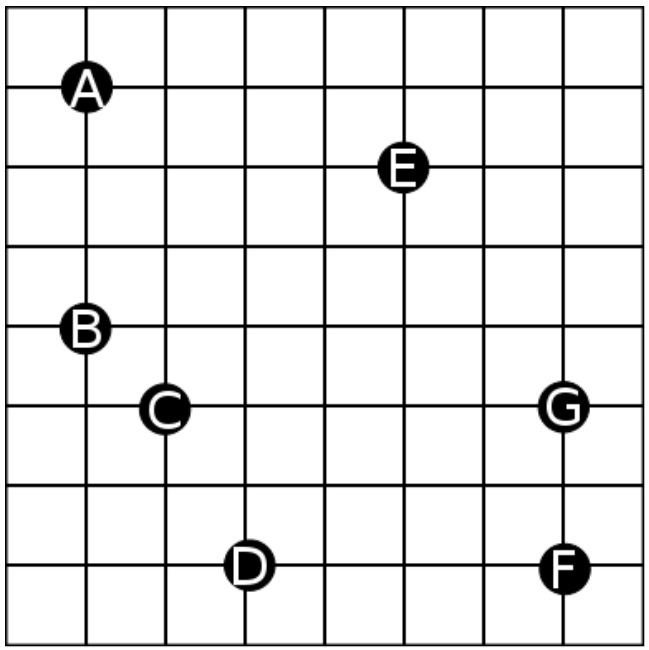


Figure : Data to be clustered

Result: C_{BC}, C_{FG}

Step 3: This time there is a cluster involved

	A	B C	D	E	F G
A	×	4.12	6.32	4.12	8.49
B C		×	3.61	4.47	6.71
D			×	5.39	4.47
E				×	5.39
F G					×

$\Rightarrow$ **reason over distance matrix!**

Exercise 04 - Question 2

Agglomerative Hierarchical Clustering

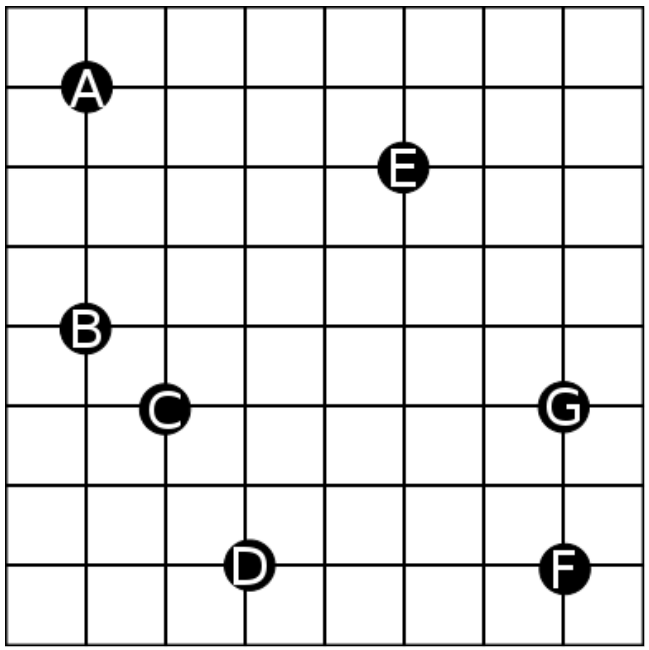


Figure : Data to be clustered

Result: C_{BC}, C_{FG}, C_{BCD}

Step 3: This time there is a cluster involved

	A	B C	D	E	F G
A	×	4.12	6.32	4.12	8.49
B C		×	3.61	4.47	6.71
D			×	5.39	4.47
E				×	5.39
F G					×

⇒ **reason over distance matrix!**

Exercise 04 - Question 2

Agglomerative Hierarchical Clustering

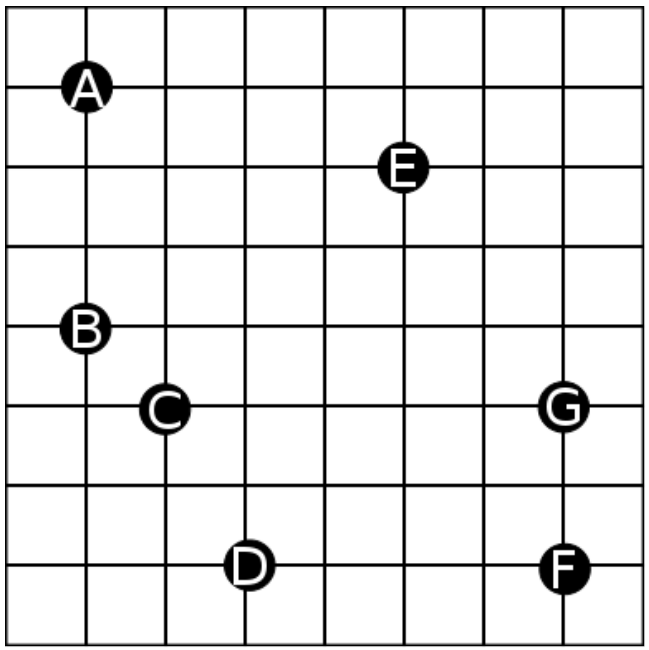


Figure : Data to be clustered

Result: C_{BC}, C_{FG}, C_{BCD}

Step 4: Always use the greatest distance if you check a cluster!

	A	B C D	E	F G
A	×	6.32	4.12	8.49
B C D		×	5.39	6.71
E			×	5.39
F G				×

⇒ **reason over distance matrix!**

Exercise 04 - Question 2

Agglomerative Hierarchical Clustering

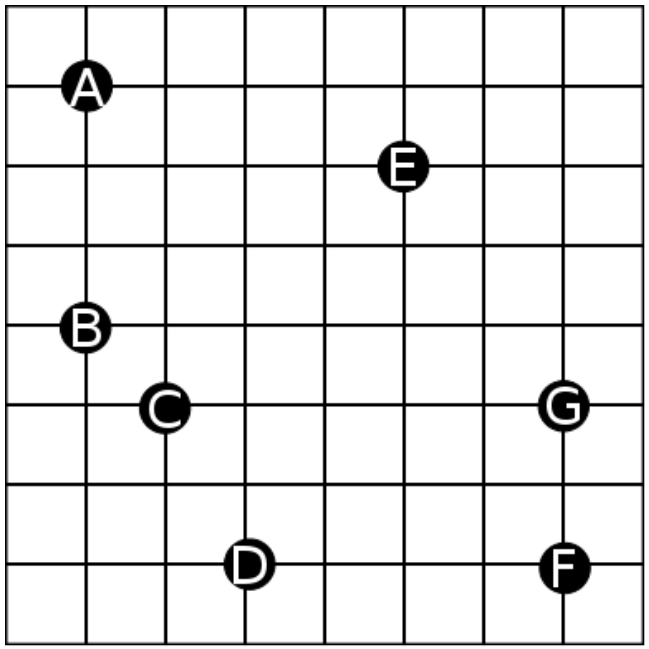


Figure : Data to be clustered

Result: C_{BC}, C_{FG}, C_{BCD}, C_{AE}

Step 4: Always use the greatest distance if you check a cluster!

	A	B C D	E	F G
A	×	6.32	4.12	8.49
B C D		×	5.39	6.71
E			×	5.39
F G				×

⇒ **reason over distance matrix!**

Exercise 04 - Question 2

Agglomerative Hierarchical Clustering

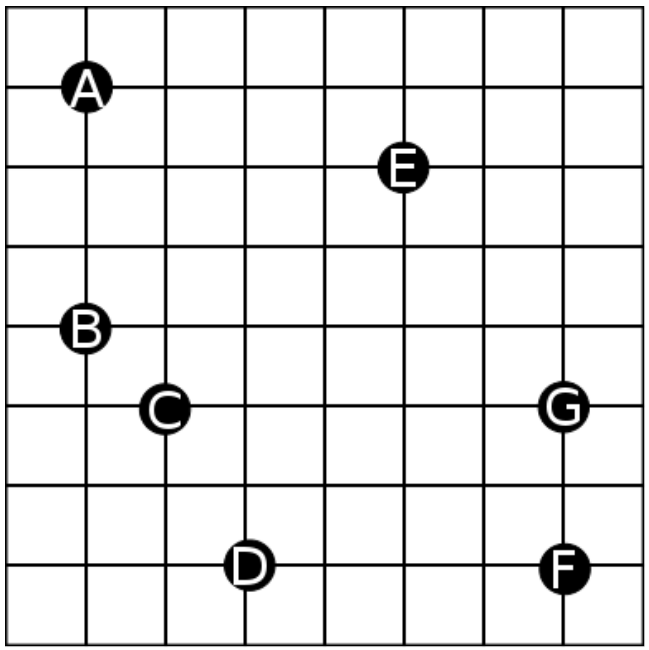


Figure : Data to be clustered

Result: C_{BC}, C_{FG}, C_{BCD}, C_{AE}

Step 4: Always use the greatest distance if you check a cluster!

	A E	B C D	F G
A E	×	6.32	8.49
B C D		×	6.71
F G			×

⇒ **reason over distance matrix!**

Exercise 04 - Question 2

Agglomerative Hierarchical Clustering

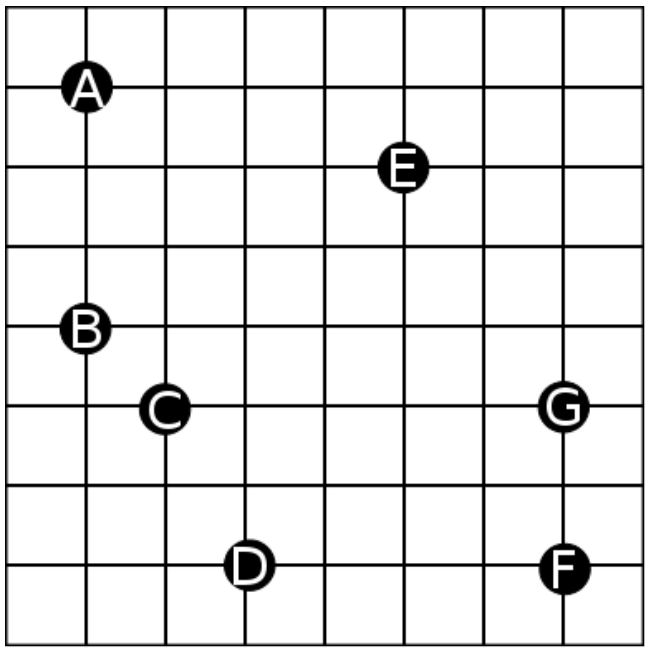


Figure : Data to be clustered

Result: C_{BC}, C_{FG}, C_{BCD}, C_{AE}, C_{ABCDE}

Step 5: ...

	A E	B C D	F G
A E	×	6.32	8.49
B C D		×	6.71
F G			×

⇒ **reason over distance matrix!**

Exercise 04 - Question 2

Agglomerative Hierarchical Clustering

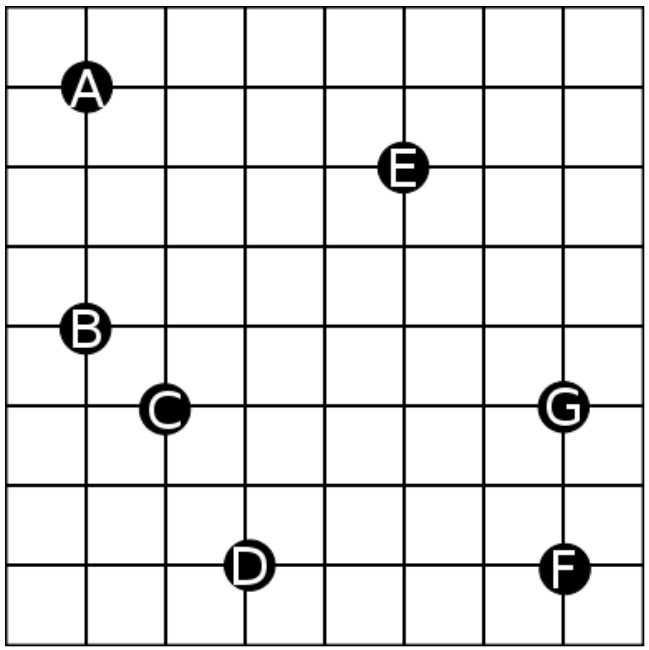


Figure : Data to be clustered

Result: C_{BC}, C_{FG}, C_{BCD}, C_{AE}, C_{ABCDE}, $C_{ABCDEFG}$

Step 6: ...

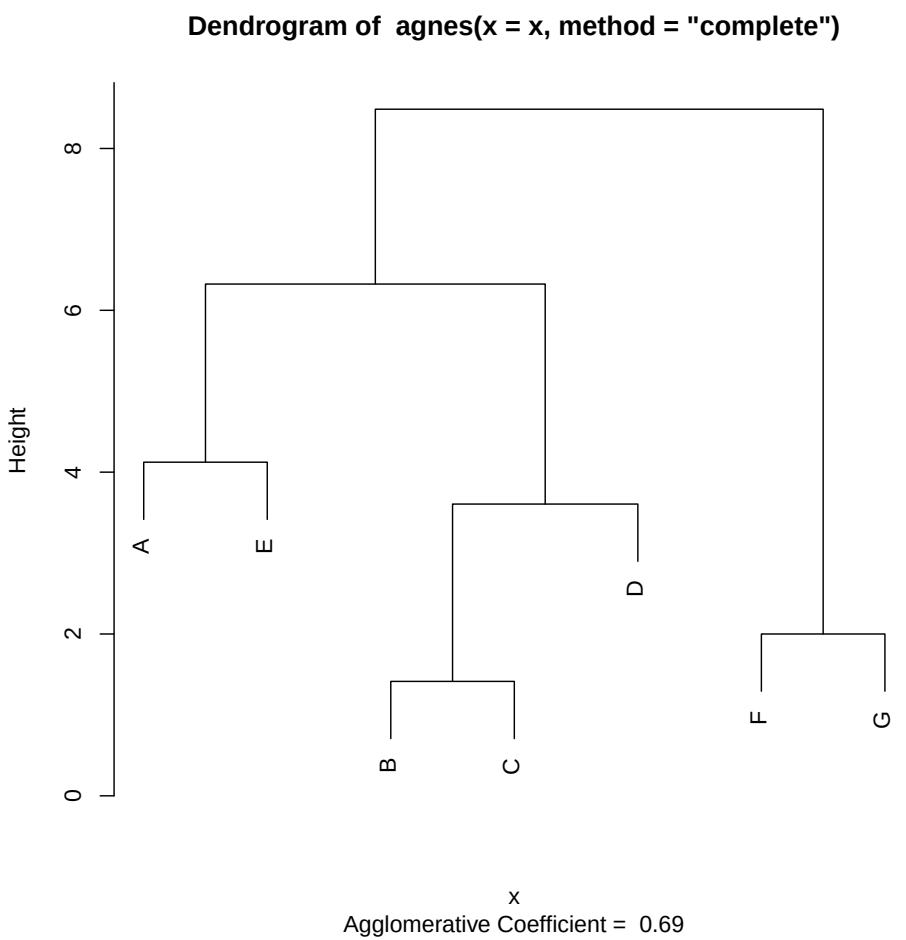
	A B C D E	F G
A B C D E	×	8.49
F G		×

$\Rightarrow$ **reason over distance matrix!**

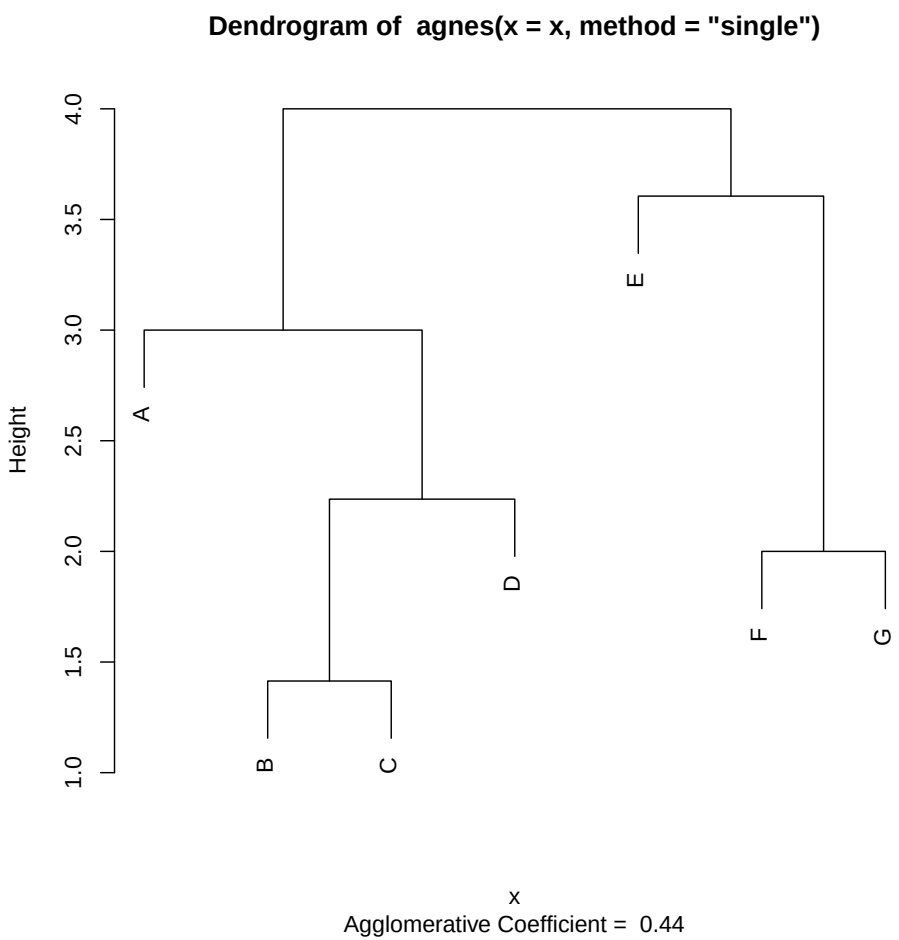
Exercise 04 - Question 2.1/2.2

Agglomerative Hierarchical Clustering

Single Link

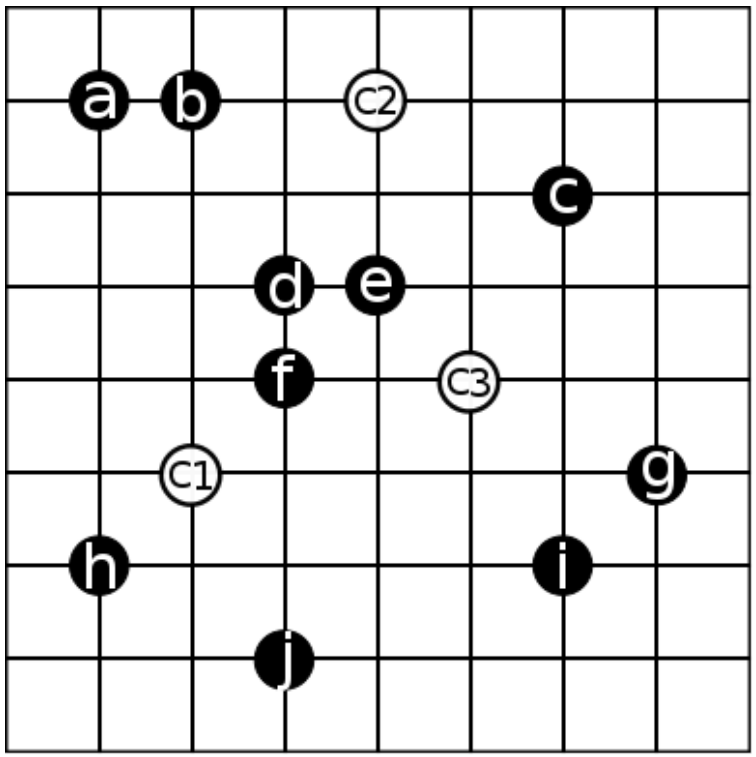


Complete Link



Exercise 04 - Question 3

k-Means Clustering



Task

Execute the k-Means algorithm with $k = 3$ based on these data points.

The initial centroids of the three clusters are:

c1(2,3); c2(4,7); c3(5,4)

Note: 2-dimensional integer space $\{0, \ldots, 8\} \times \{0, \ldots, 8\}$

Exercise 04 - Question 3

k-Means Clustering

	a	b	c	d	e	f	g	h	i	j
c1	4.12	4.0	5.0	**2.24**	2.83	**1.41**	5.0	**1.41**	4.12	**2.24**
c2	**3.0**	**2.0**	**2.24**	2.24	2.0	3.16	5.0	5.83	5.39	6.08
c3	5.0	4.24	2.24	2.24	**1.41**	2.0	**2.24**	4.47	**2.24**	3.61

Table : Distances to centroids

$$c1_{dfhj}, c2_{abc}, c3_{egi}$$

$\Rightarrow$ **reason over distance matrix!**

Exercise 04 - Question 3

k-Means Clustering - Adjust centroids

c1 (position old: $< 2, 3 >$):

$$round(\frac{\mathbf{d} + \mathbf{f} + \mathbf{h} + \mathbf{j}}{4}) = < 3, 3 >$$

c2 (position old: $< 4, 7 >$):

$$round(\frac{\mathbf{a} + \mathbf{b} + \mathbf{c}}{3}) = < 3, 7 >$$

c3 (position old: $< 5, 4 >$):

$$round(\frac{\mathbf{e} + \mathbf{g} + \mathbf{i}}{3}) = < 6, 3 >$$

Note: for $round(\cdot)$ apply the **ceiling** $\lceil \cdot \rceil$ and **floor** $\lfloor \cdot \rfloor$ functions; rounding is only necessary on integer-valued spaces ($\mathbb{N}^d$ or $\mathbb{Z}^d$), not on $\mathbb{Q}^d$ or $\mathbb{R}^d$!

Exercise 04 - Question 3

k-Means Clustering

	a	b	c	d	e	f	g	h	i	j
c1	4.47	4.12	4.24	**2.0**	**2.24**	**1.0**	4.0	**2.24**	3.16	**2.0**
c2	**1.0**	**0.0**	4.12	2.24	2.83	3.16	6.4	5.1	6.4	6.08
c3	5.83	5.0	**2.0**	3.16	2.24	3.0	**1.41**	5.39	**2.0**	4.24

Table : Distances to centroids

$$c1_{defhj}, c2_{ab}, c3_{cgi}$$

$\Rightarrow$ **reason over distance matrix!**

Exercise 04 - Question 3

k-Means Clustering - Adjust centroids

c1 (position old: $< 3, 3 >$):

$$round(\frac{\mathbf{d} + \mathbf{e} + \mathbf{f} + \mathbf{h} + \mathbf{j}}{5}) = < 3, 3 >$$

c2 (position old: $< 3, 7 >$):

$$round(\frac{\mathbf{a} + \mathbf{b}}{2}) = < 2, 7 >$$

c3 (position old: $< 6, 3 >$):

$$round(\frac{\mathbf{c} + \mathbf{g} + \mathbf{i}}{3}) = < 6, 4 >$$

Exercise 04 - Question 3

k-Means Clustering

	a	b	c	d	e	f	g	h	i	j
c1	4.47	4.12	4.24	**2.0**	**2.24**	**1.0**	4.0	**2.24**	3.16	**2.0**
c2	**1.0**	**0.0**	4.12	2.24	2.83	3.16	6.4	5.1	6.4	6.08
c3	5.83	5.0	**2.0**	3.16	2.24	3.0	**1.41**	5.39	**2.0**	4.24

Table : Distances to centroids

$$c1_{defhj}, c2_{ab}, c3_{cgi}$$

$\Rightarrow$ **reason over distance matrix!**

Exercise 04 - Question 3

k-Means Clustering - Adjust centroids

c1 (position old: $< 3, 3 >$):

$$round(\frac{\mathbf{d} + \mathbf{e} + \mathbf{f} + \mathbf{h} + \mathbf{j}}{5}) = < 3, 3 >$$

c2 (position old: $< 2, 7 >$):

$$round(\frac{\mathbf{a} + \mathbf{b}}{2}) = < 2, 7 >$$

c3 (position old: $< 6, 4 >$):

$$round(\frac{\mathbf{c} + \mathbf{g} + \mathbf{i}}{3}) = < 6, 4 >$$

Convergence Criterion

the centroids do not change anymore $\Rightarrow$ finished!

Exercise 04 - Question 4

Evaluation

c_1	c_2	c_3	c_4
a	d	h	c
b	e	j	g
	f		i

Table : Gold Standard

k_1	k_2	k_3
d	a	c
e	b	g
f		i
h		
j		

Table : k-Means result

Compute the purity score of your k-Means clustering result with respect to the given gold standard. As a reminder, the purity is calculated via this formula:

$$Purity(k_i) = \frac{1}{n_i} \times \max\{n_{ij} \mid c_j \in C\}$$

Exercise 04 - Question 4

Evaluation - Purity Measure

k_1 = {d, e, f, h ,j}
k_2 = {a, b}
k_3 = {c, g, i}

$$Purity(k_1) = \frac{1}{5} \times 3 = 0.6$$
$$Purity(k_2) = \frac{1}{2} \times 2 = 1$$
$$Purity(k_3) = \frac{1}{3} \times 3 = 1$$

Purity

$$Purity(k_i) = \frac{1}{n_i} \times \max\{n_{ij} \mid c_j \in C\}$$

where:

1. $C = \{c_1, \ldots, c_q\}$ gold standard classes
2. $K = \{k_1, \ldots, k_q\}$ clusters generated by the clustering algorithm
3. $n_i = |k_i|$
4. $n_{ij} = |k_i \cap c_j|$

Exercise 04 - Question 5

Clustering with RapidMiner (Preprocessing)

1. Load texts
2. Tokenize (mode = non letters)
3. Transform tokens to lowercase
4. Filter stopwords
5. Filter tokens by length (min = 2; max = 100)
6. Apply stemming (Snowball stemmer)
7. Store pipeline

Exercise 04 - Question 5

Clustering with RapidMiner (Bonus)

1. Apply k-Means with $k = 3$
2. Report the top 10 words of each cluster
3. Think of a headline for each cluster based on the associations you have with the top words


Thank You!

Hierarchical Clustering

Ward's Agglomerative Clustering Algorithm

1: **procedure** AGGLOMERATIVE($\{\vec{x}_1, \ldots, \vec{x}_n\}$)
2: **for** $1 \leq i \leq n$ **do**
3: $c_i \leftarrow \{\vec{x}_i\}$;
4: **end for**
5: $C \leftarrow \{c_1, \ldots, c_n\}$;
6: $C' \leftarrow C$;
7: **while** $|C| \geq 1$ **do**
8: $(c, c') \leftarrow \arg\beta\ \{\Delta(c, c') \mid (c, c') \in C \times C\}$;
9: $c'' \leftarrow c \cup c'$;
10: $C \leftarrow (C \setminus \{c, c'\}) \cup \{c''\}$;
11: $C' \leftarrow C' \cup \{c''\}$;
12: **end while**
13: **return** C';
14: **end procedure**

- Ward's Algorithm (hierarchical clustering), for $\beta \in \{\min, \max\}$
- Runs in $O(n^3)$ time, where n is the number of data points

Hierarchical Clustering

Dual Formulation in Terms of Similarity

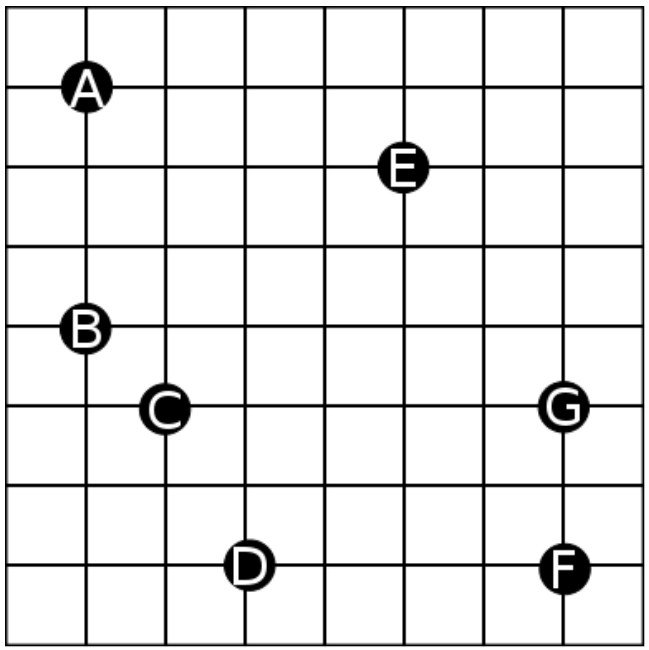


Figure : Data to be clustered

$$Sim(c_i, c_j) =$$

Single Link

$$\max_{\vec{x} \in c_i, \vec{y} \in c_j} sim(\vec{x}, \vec{y})$$

Complete Link

$$\min_{\vec{x} \in c_i, \vec{y} \in c_j} sim(\vec{x}, \vec{y})$$

$sim(\vec{x}, \vec{y})$:

- **similarity** of $\vec{x}$ to $\vec{y}$ (e.g., cosine)
- reason over **similarity matrix**
- invert $\max$ for $\min$, $\min$ for $\max$
- replace distance by similarity in Ward

k-Means Clustering

Standard k-Means Algorithm

1: **procedure** K-MEANS($\{\vec{x}_1, \ldots, \vec{x}_n\}, k$)
2: **for** $1 \leq i \leq k$ **do**
3: $\mu_i \leftarrow$ RANDOM(); $C_i \leftarrow \emptyset$;
4: **end for**
5: **repeat**
6: **for** $1 \leq i \leq k$ **do**
7: $C_i \leftarrow C_i \cup \{\arg\min_{1 \leq j \leq n} distance(\vec{x}_j, \mu_i))\}$;
8: **end for**
9: **for** $1 \leq i \leq k$ **do**
10: $\mu_i \leftarrow \frac{1}{|C_i|} \times \sum\{\vec{x} \mid \vec{x} \in C_i\}$;
11: **end for**
12: **until** $\vec{\mu}$ doesn't change
13: **return** $C_1, \ldots, C_k$
14: **end procedure**

- k-Means clustering algorithm (again we use Euclidian distance)
- Runs in $O(n^{d \times (k+1)} \times \log n)$ time, where d is the dimension of the Euclidian vector space $\mathbb{R}^d$ (**NP**-hard for $k \geq 3$)