Агломеративная иерахическая кластеризация

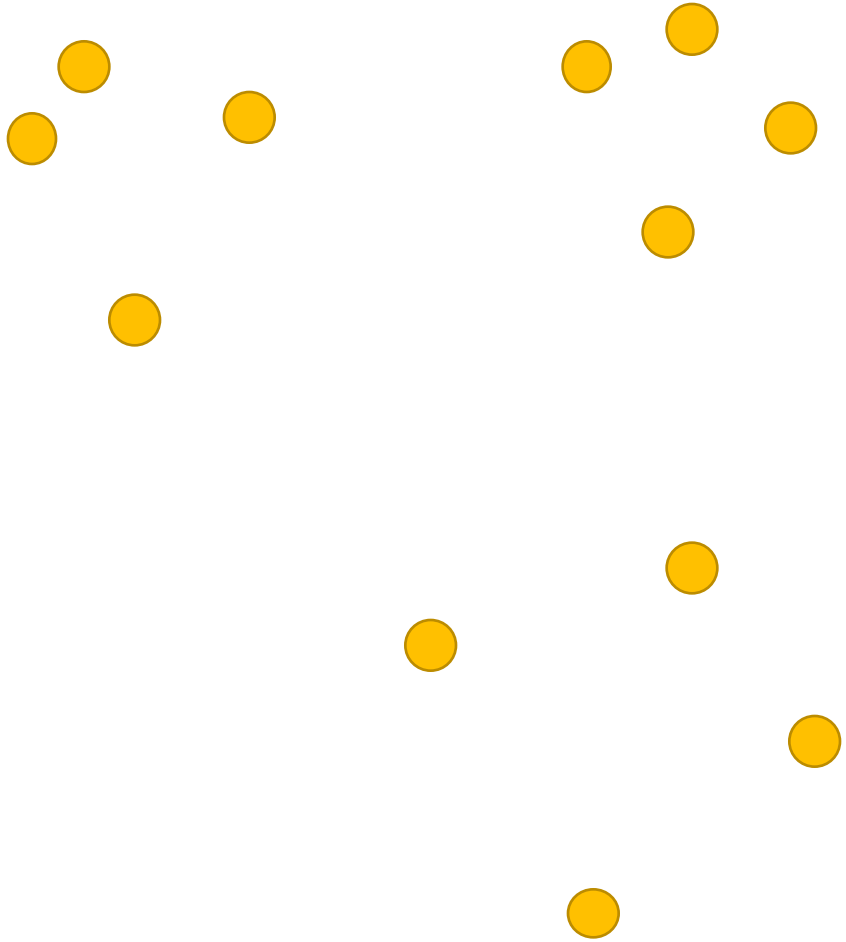
План

1. Иерархическая кластеризация
2. Как устроена агломеративная кластеризация
3. Расстояние между кластерами
4. Формула Ланса-Уильямса
5. Дендрограммы
6. Примеры работы

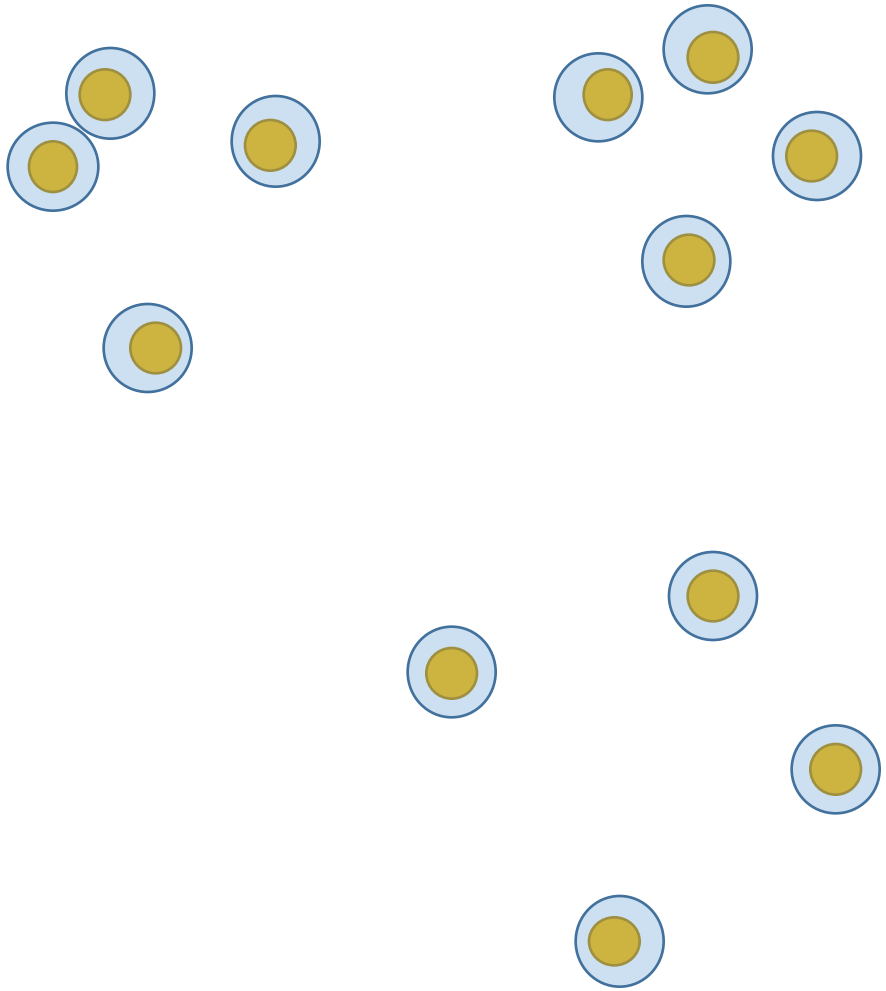
Иерархическая кластеризация

- Агломеративная
- Дивизионная или дивизимная

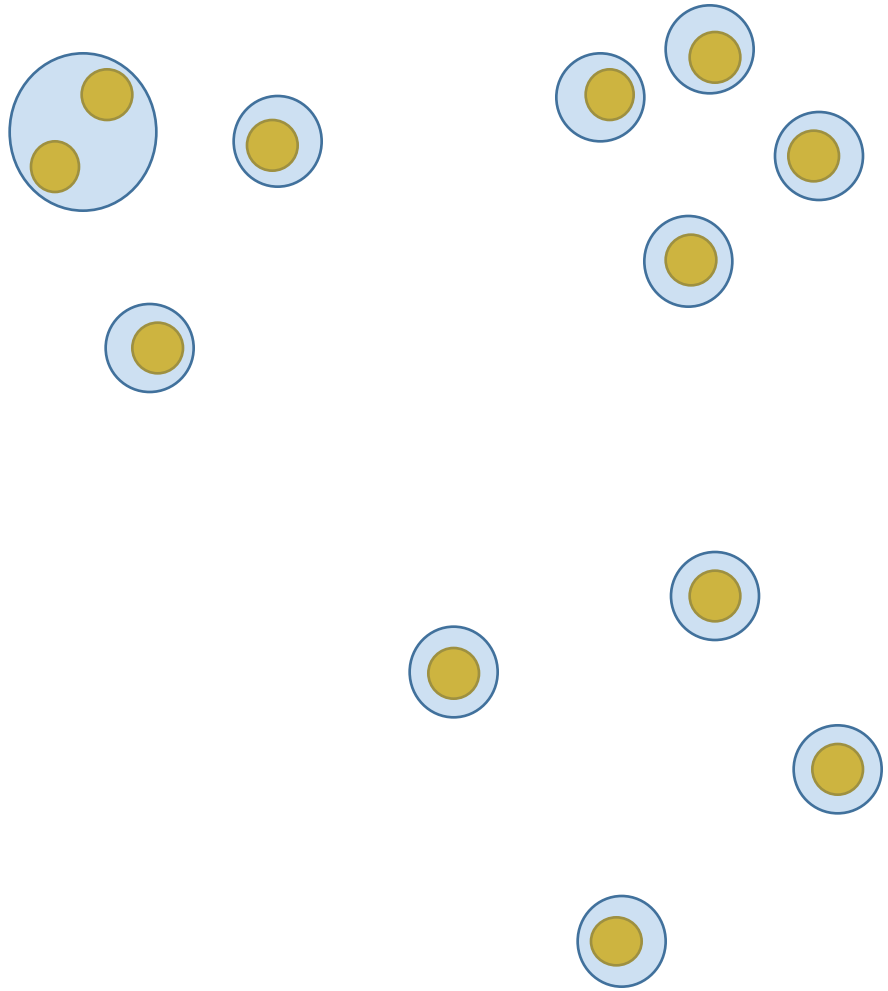
Агломеративная кластеризация



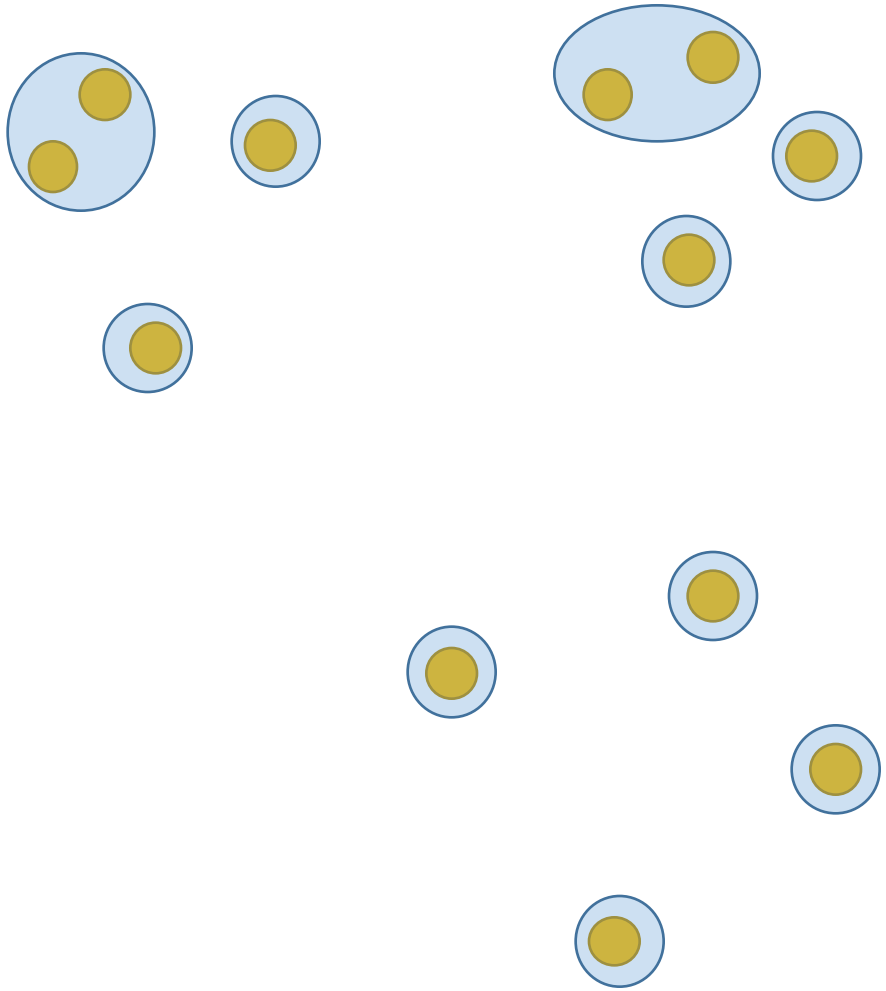
Агломеративная кластеризация



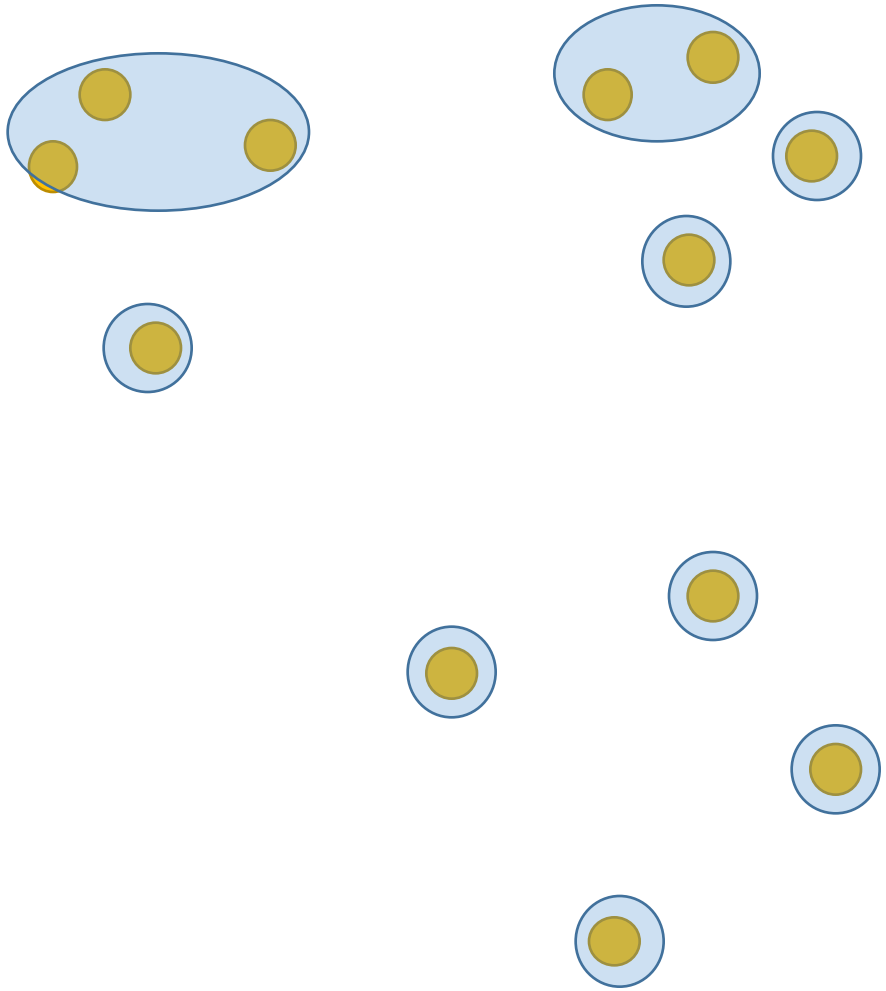
Агломеративная кластеризация



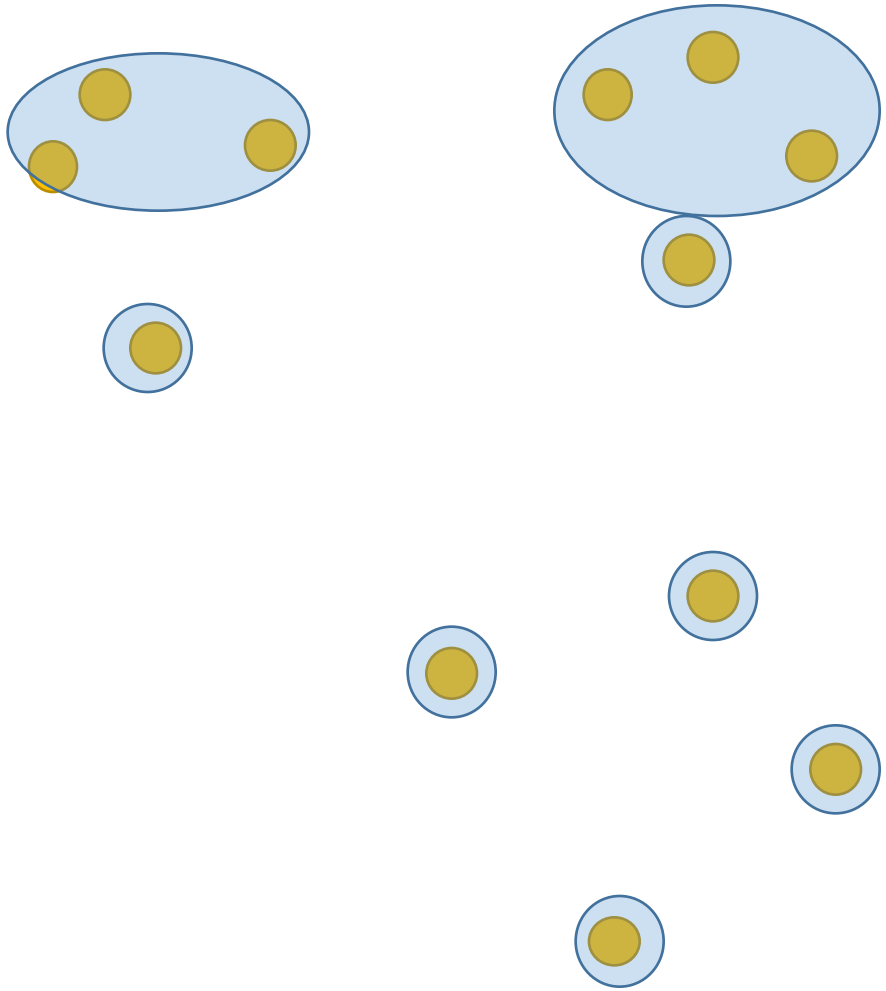
Агломеративная кластеризация



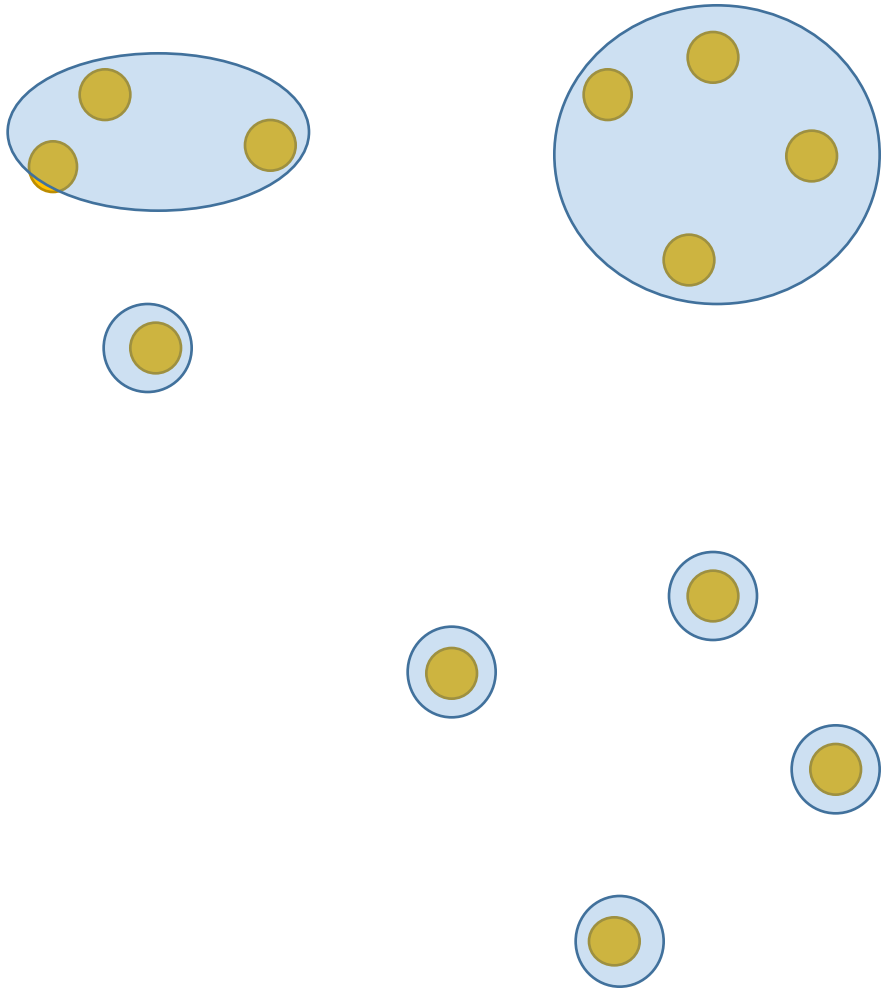
Агломеративная кластеризация



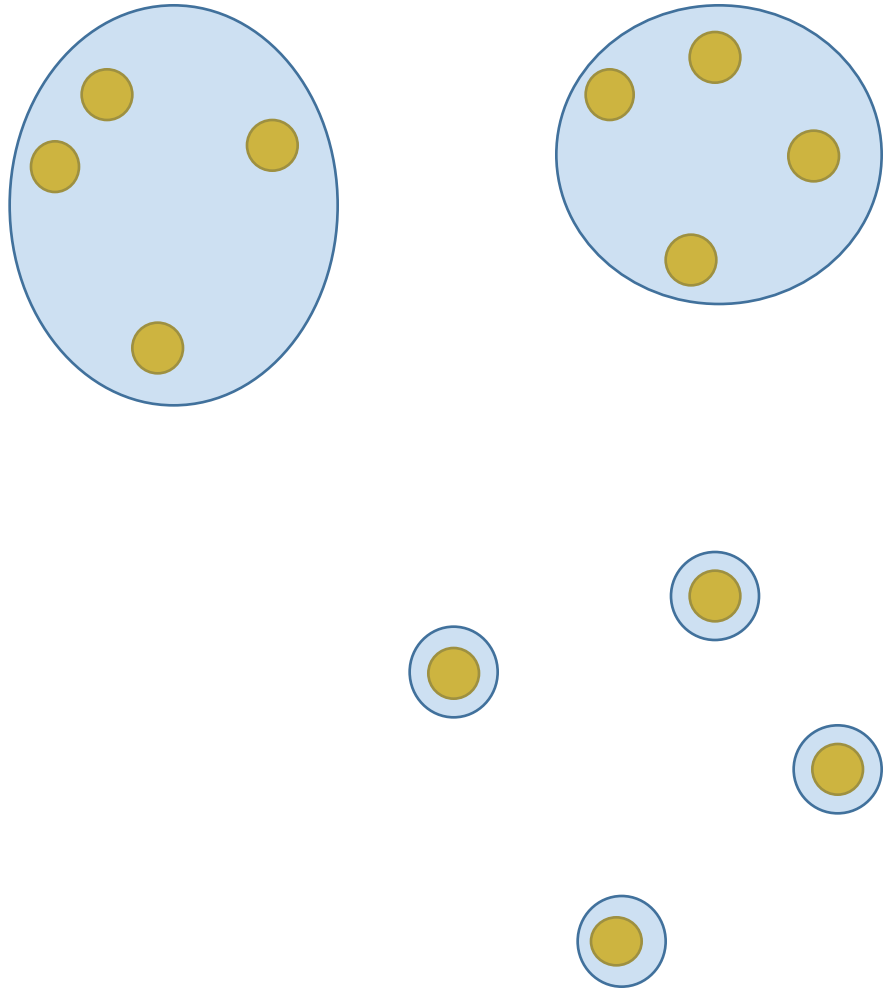
Агломеративная кластеризация



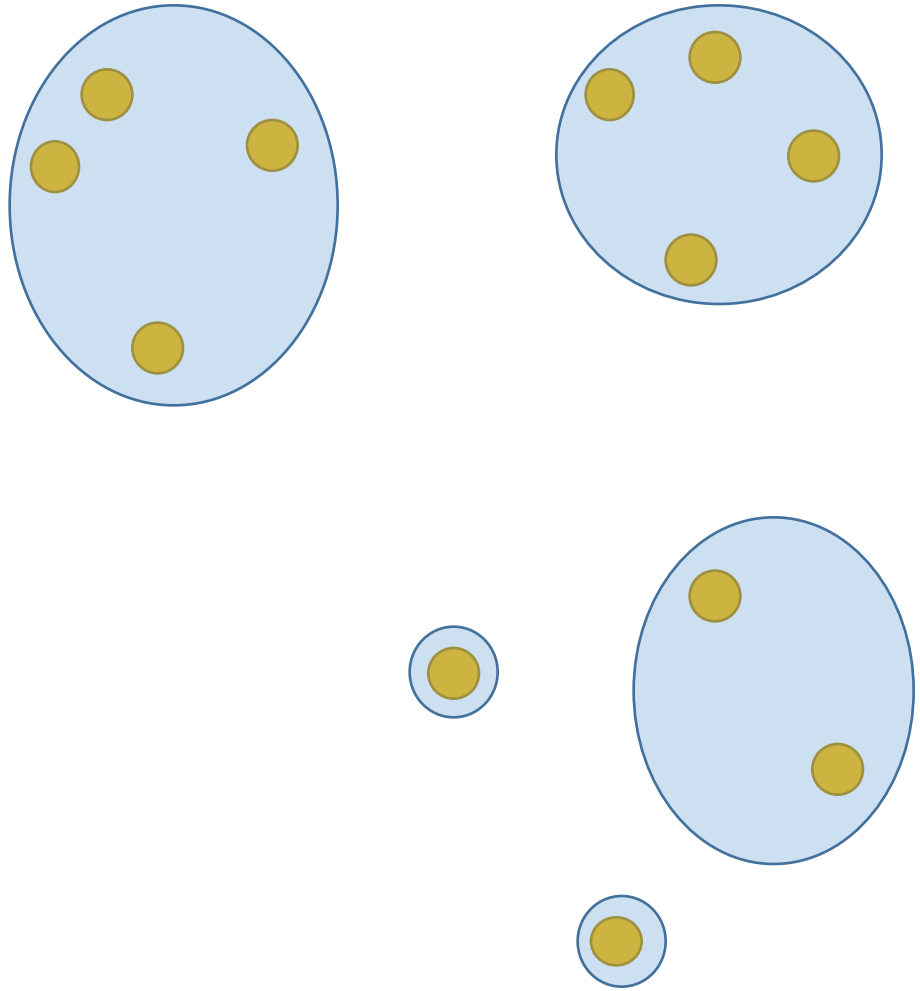
Агломеративная кластеризация



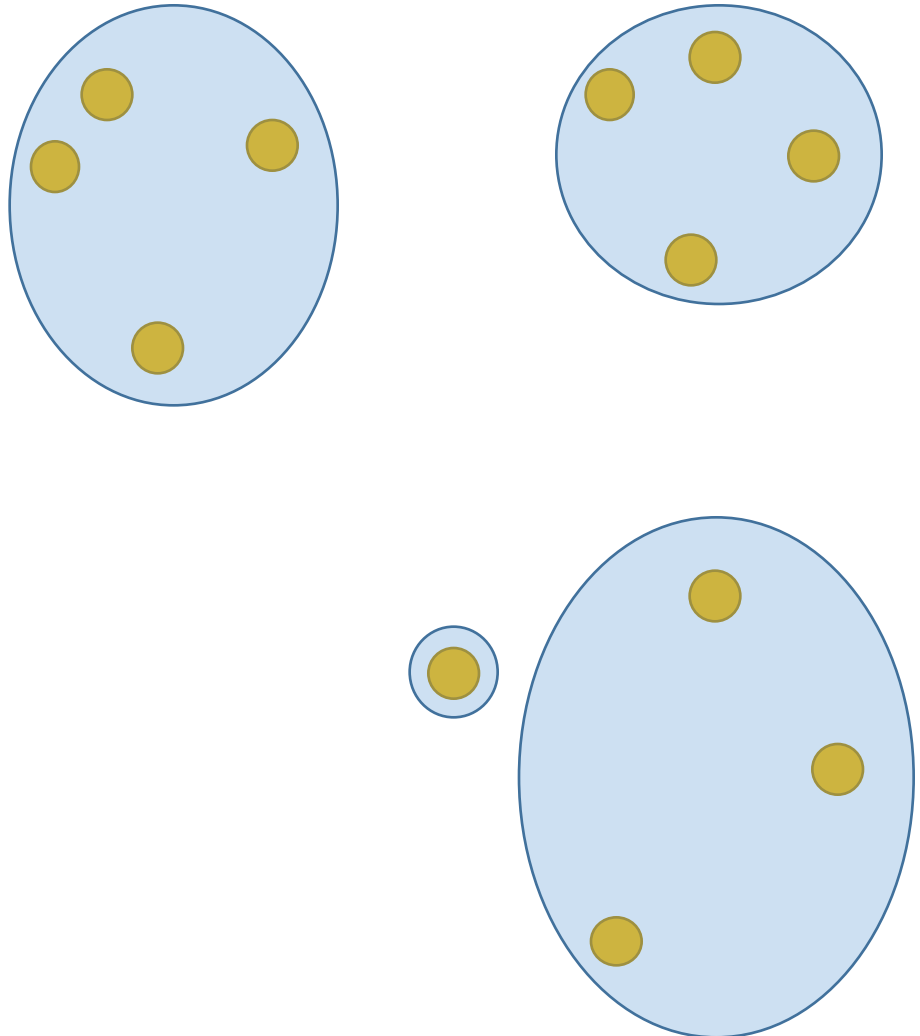
Агломеративная кластеризация



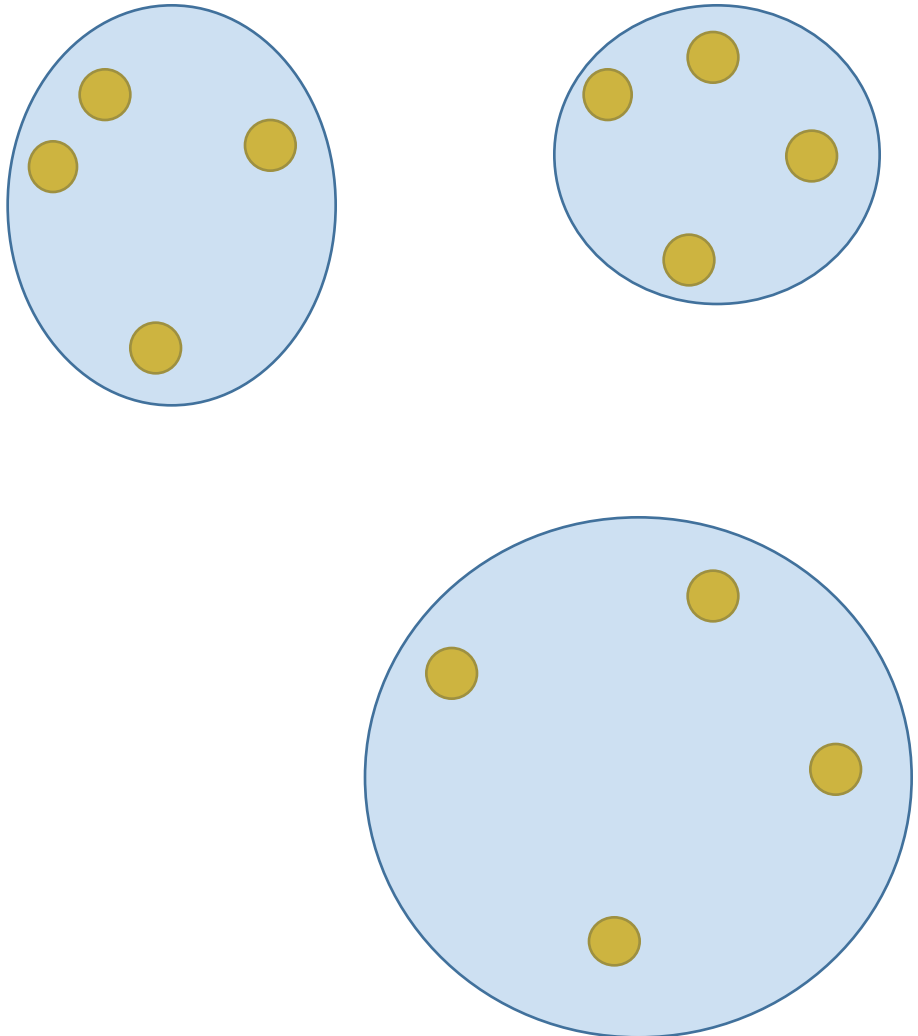
Агломеративная кластеризация



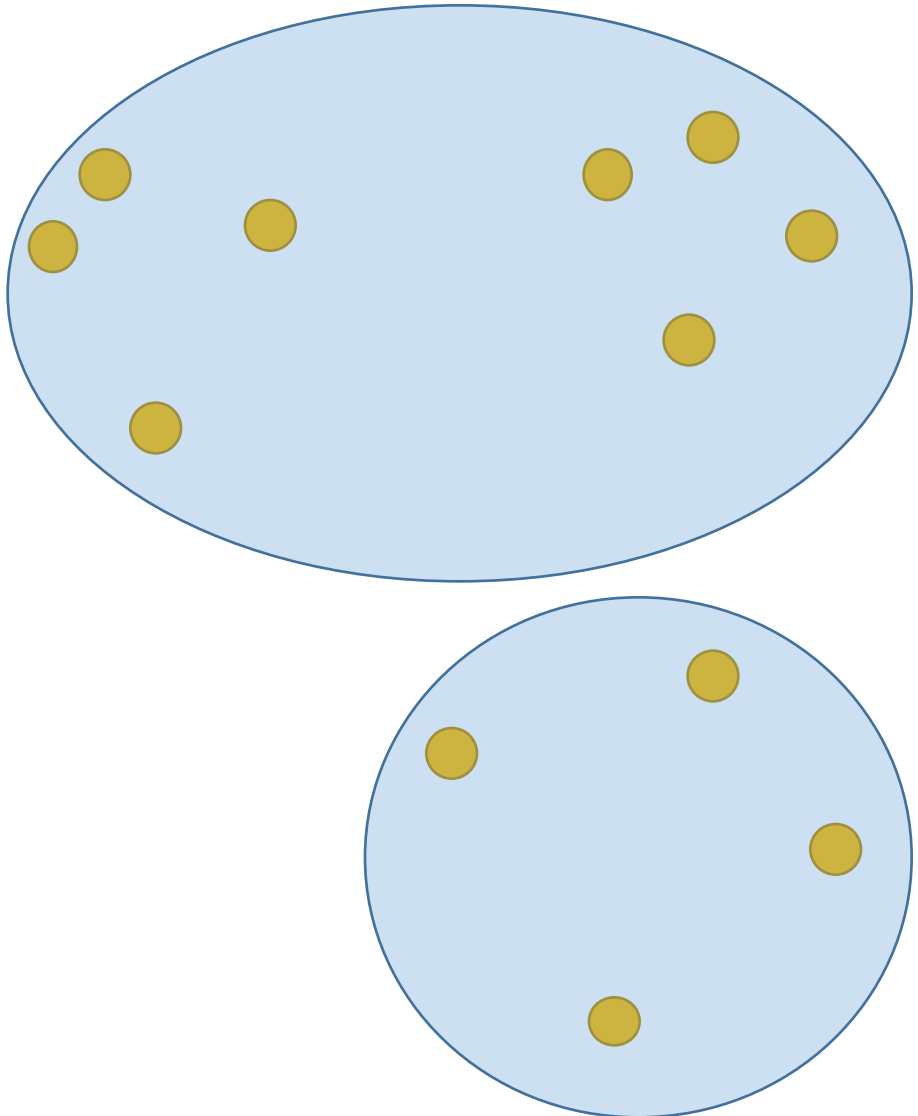
Агломеративная кластеризация



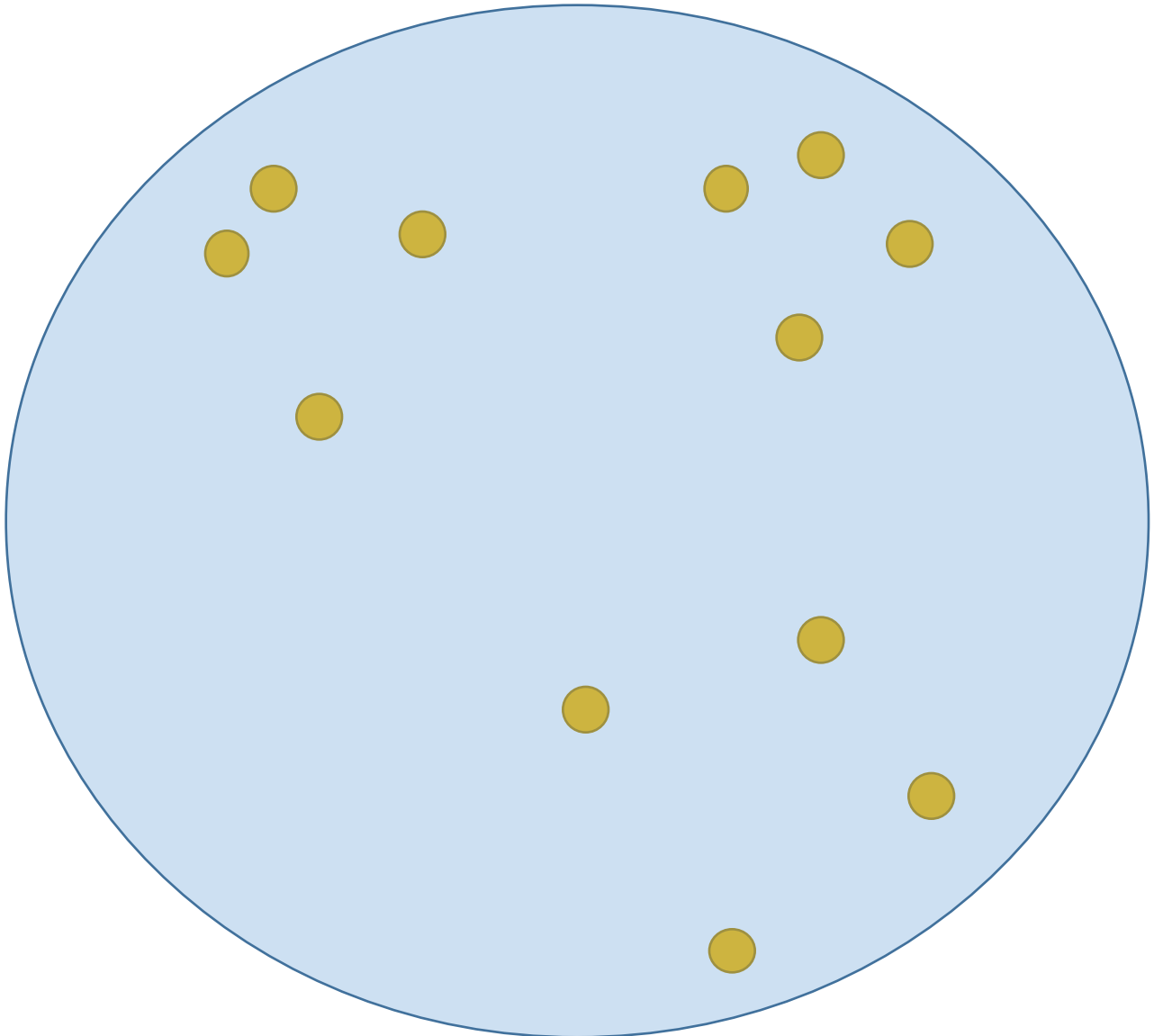
Агломеративная кластеризация



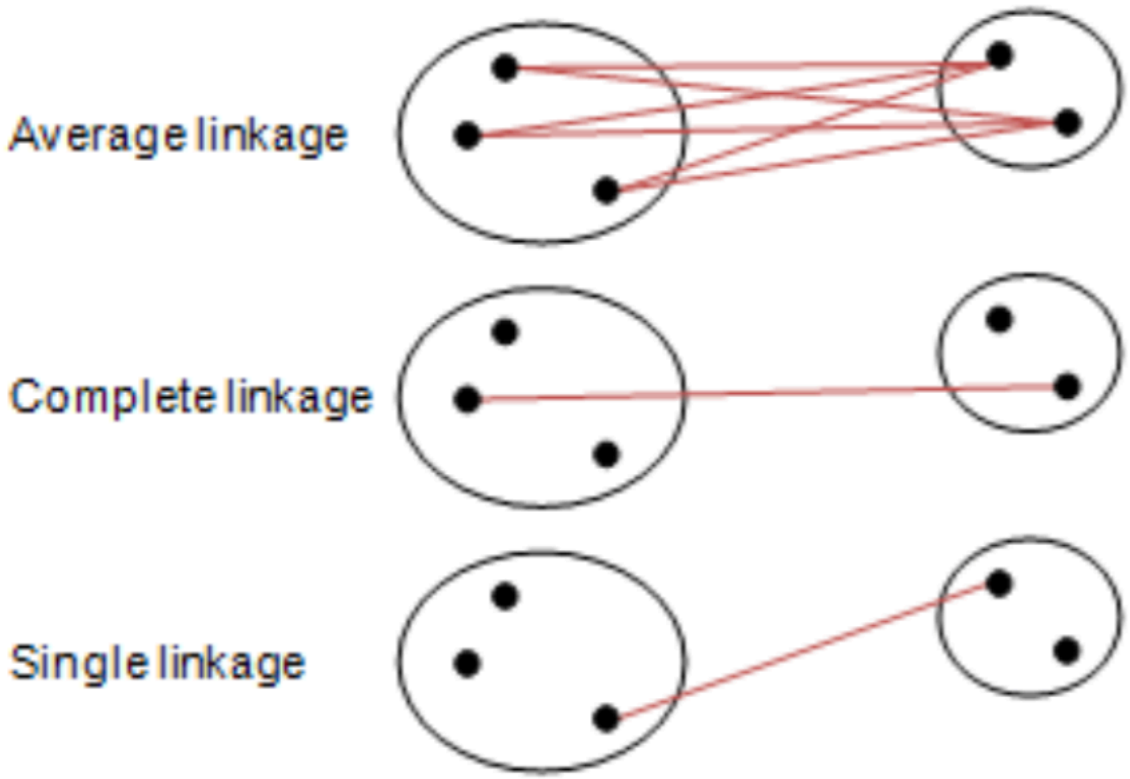
Агломеративная кластеризация



Агломеративная кластеризация



Расстояния между кластерами



Формула Ланса-Уильямса

$$R(U \cup V, S) = \alpha_U R(U, S) + \alpha_V R(V, S) + \beta R(U, V) + \gamma |R(U, S) - R(V, S)|$$

Формула Ланса-Уильямса

$$R(U \cup V, S) = \alpha_U R(U, S) + \alpha_V R(V, S) + \beta R(U, V) + \gamma |R(U, S) - R(V, S)|$$

Расстояние ближнего соседа:

$$R^{\text{б}}(W, S) = \min_{w \in W, s \in S} \rho(w, s); \qquad \alpha_U = \alpha_V = \tfrac{1}{2},\ \beta = 0,\ \gamma = -\tfrac{1}{2}.$$

Расстояние дальнего соседа:

$$R^{\text{д}}(W, S) = \max_{w \in W, s \in S} \rho(w, s); \qquad \alpha_U = \alpha_V = \tfrac{1}{2},\ \beta = 0,\ \gamma = \tfrac{1}{2}.$$

Среднее расстояние:

$$R^{\text{с}}(W, S) = \frac{1}{|W||S|} \sum_{w \in W} \sum_{s \in S} \rho(w, s); \qquad \alpha_U = \frac{|U|}{|W|},\ \alpha_V = \frac{|V|}{|W|},\ \beta = \gamma = 0.$$

Формула Ланса-Уильямса

$$R(U \cup V, S) = \alpha_U R(U, S) + \alpha_V R(V, S) + \beta R(U, V) + \gamma |R(U, S) - R(V, S)|$$

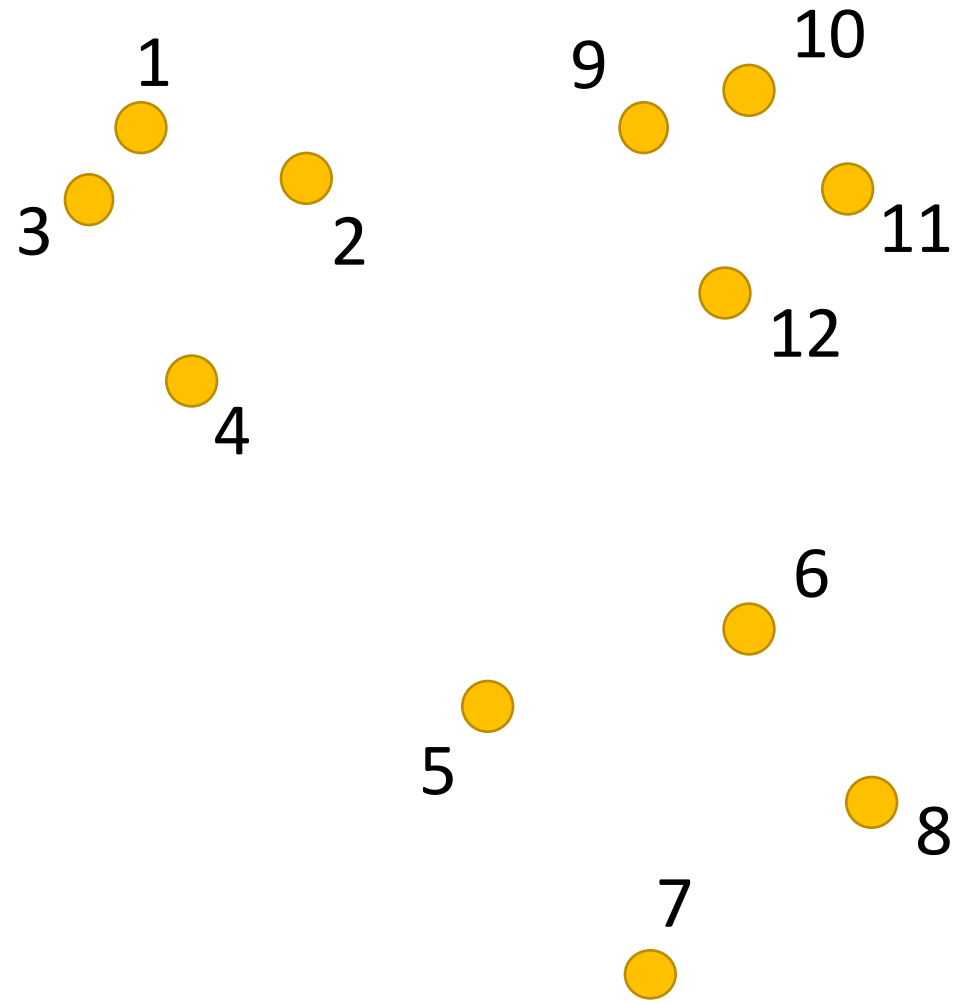
Расстояние между центрами:

$$R^{\text{ц}}(W, S) = \rho^2\Big(\sum_{w \in W} \frac{w}{|W|}, \sum_{s \in S} \frac{s}{|S|}\Big); \qquad \alpha_U = \frac{|U|}{|W|},\ \alpha_V = \frac{|V|}{|W|},\ \beta = -\alpha_U \alpha_V,\ \gamma = 0.$$

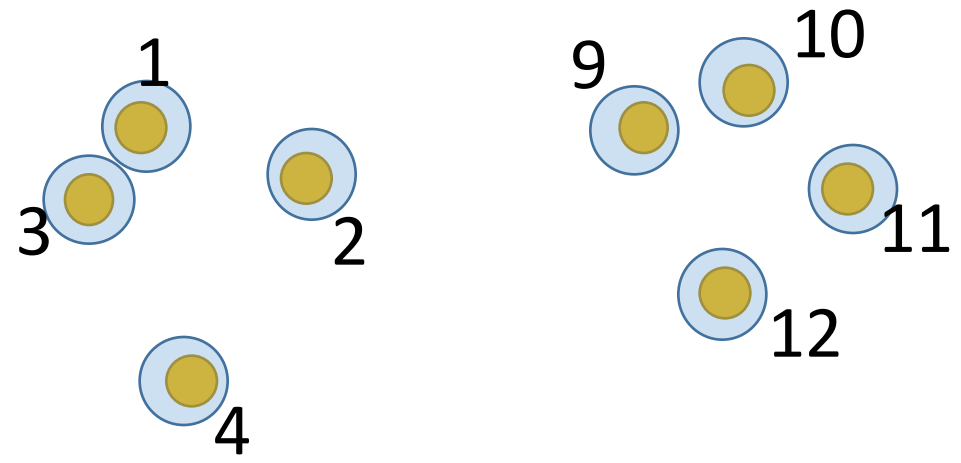
Расстояние Уорда:

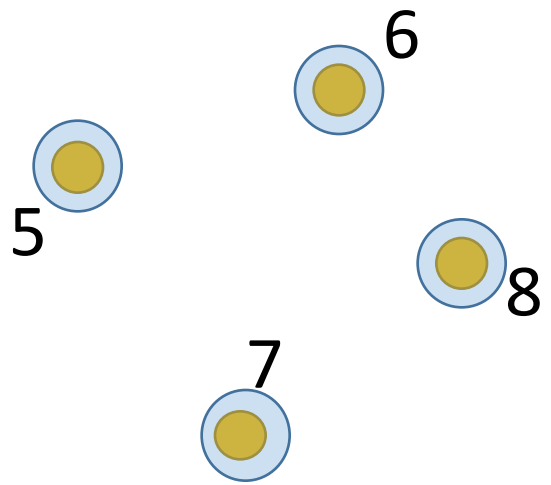
$$R^{\text{у}}(W, S) = \frac{|S||W|}{|S|+|W|}\rho^2\Big(\sum_{w \in W} \frac{w}{|W|}, \sum_{s \in S} \frac{s}{|S|}\Big); \quad \alpha_U = \frac{|S|+|U|}{|S|+|W|},\ \alpha_V = \frac{|S|+|V|}{|S|+|W|},\ \beta = \frac{-|S|}{|S|+|W|},\ \gamma = 0.$$

Дендрограмма



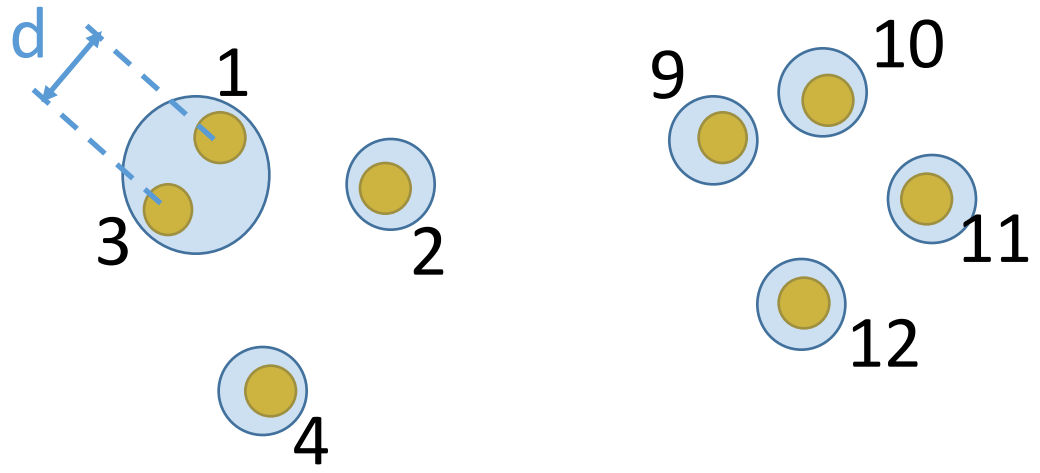
Дендрограмма

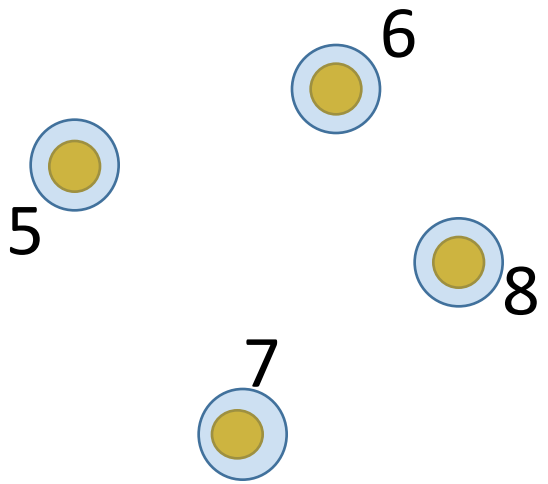




1 3 2 4 9 10 11 12 6 8 7 5

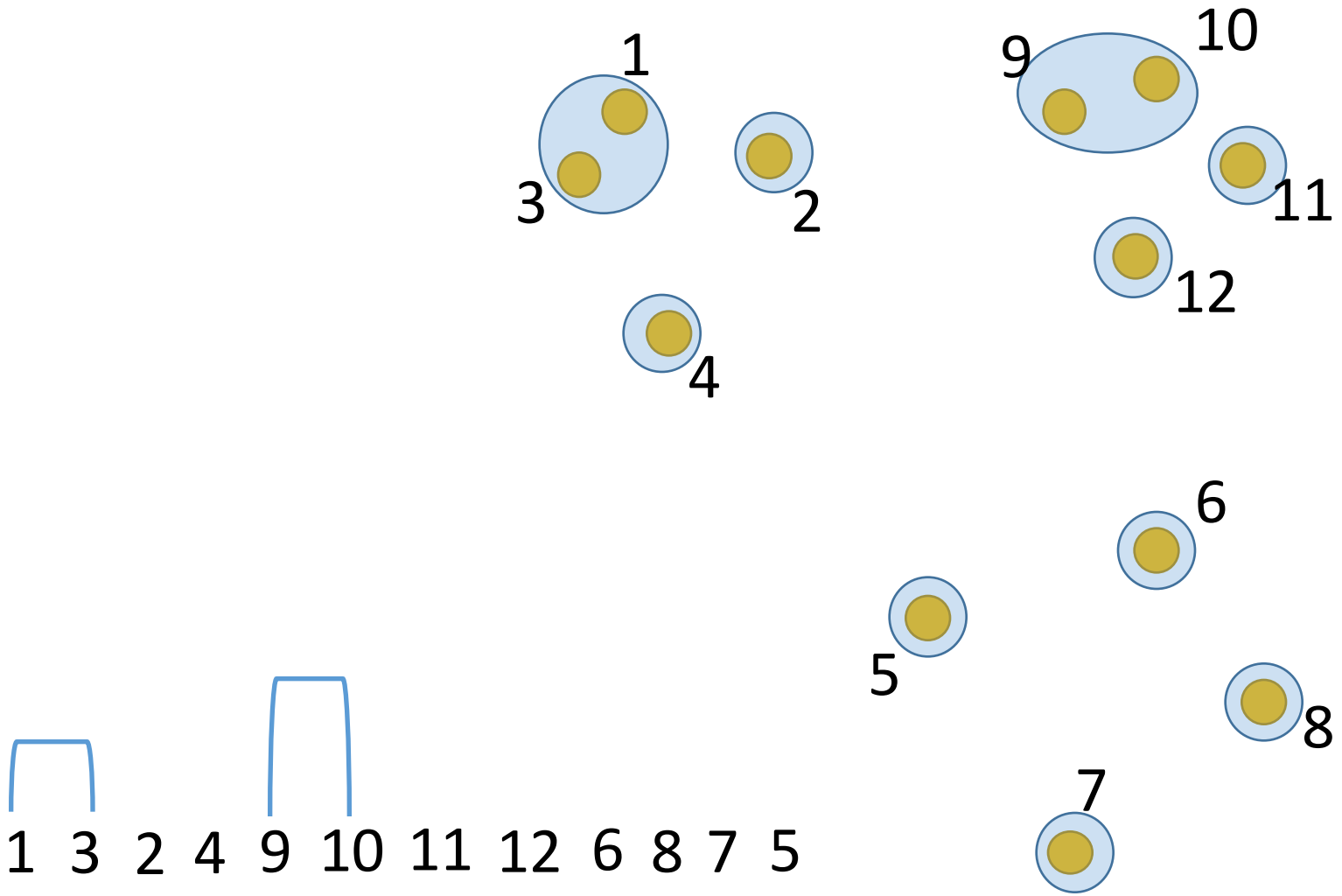
Дендрограмма



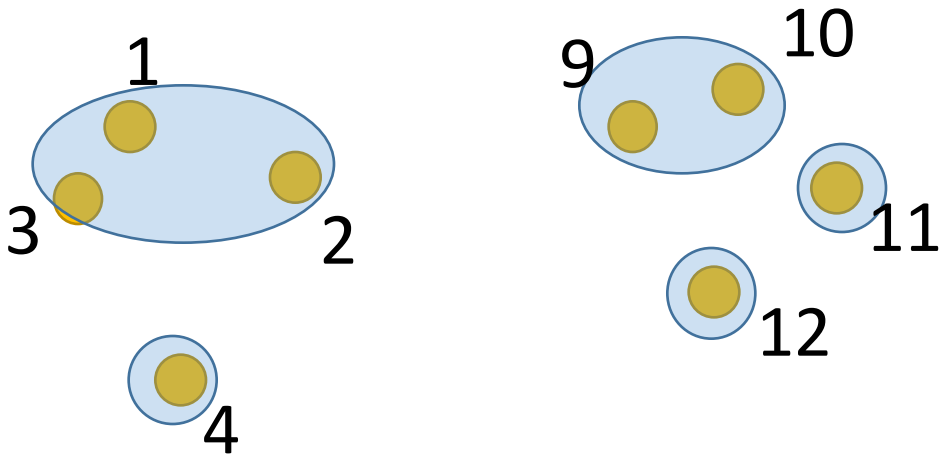


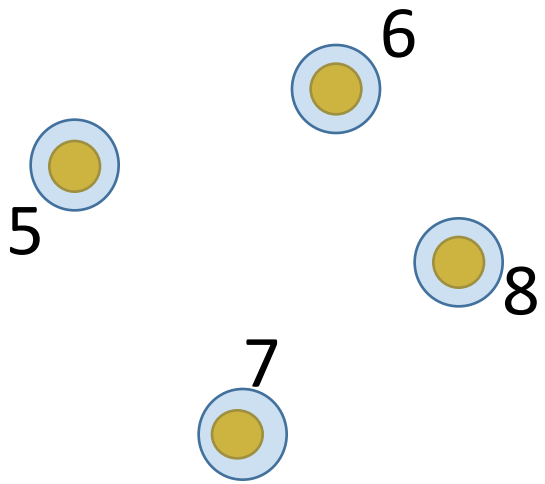
1 3 2 4 9 10 11 12 6 8 7 5

Дендрограмма

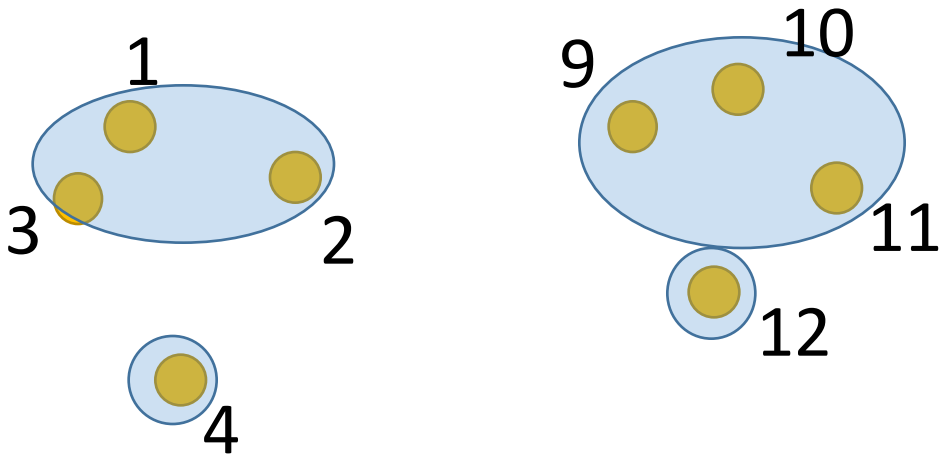


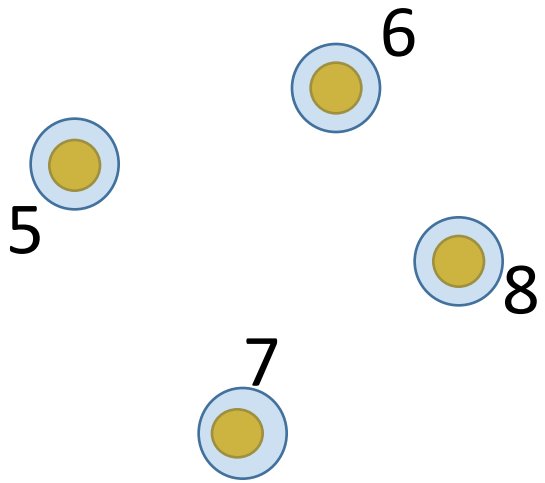
Дендрограмма

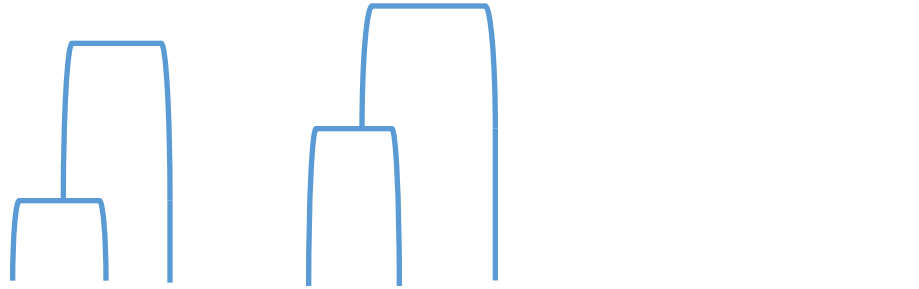




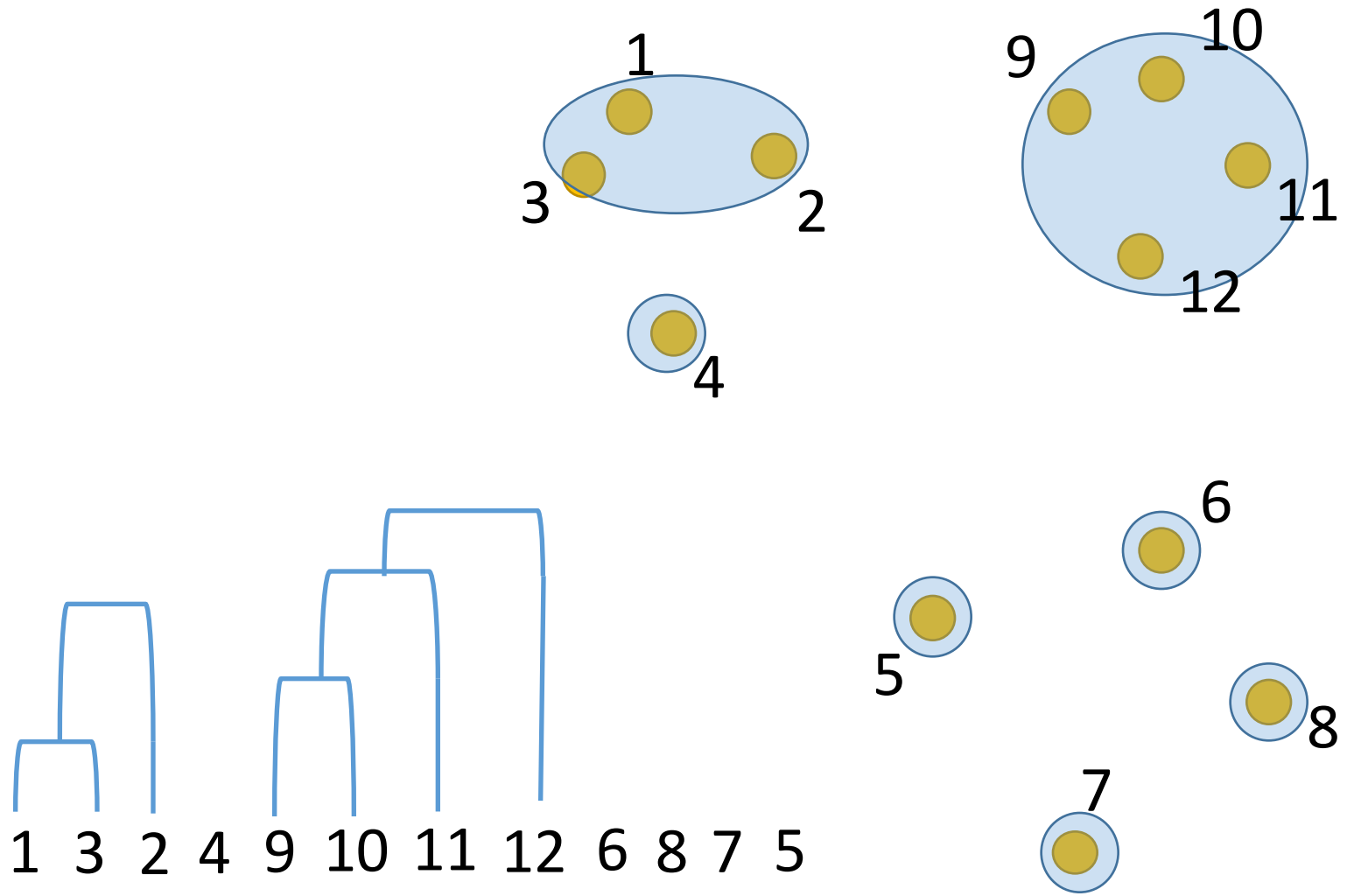
Дендрограмма



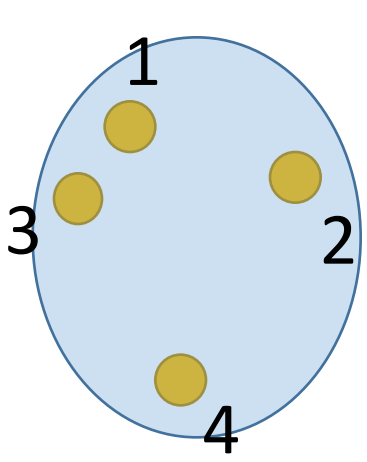


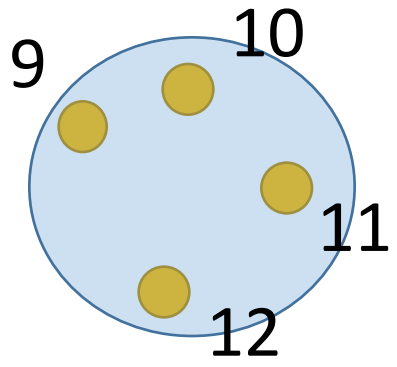


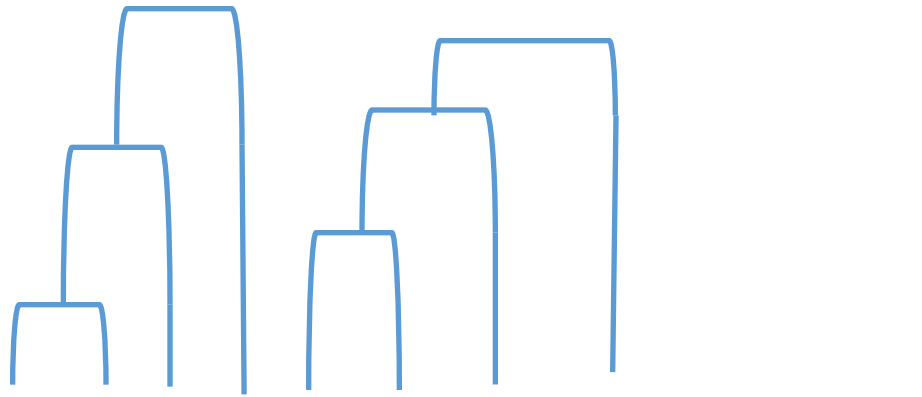
Дендрограмма

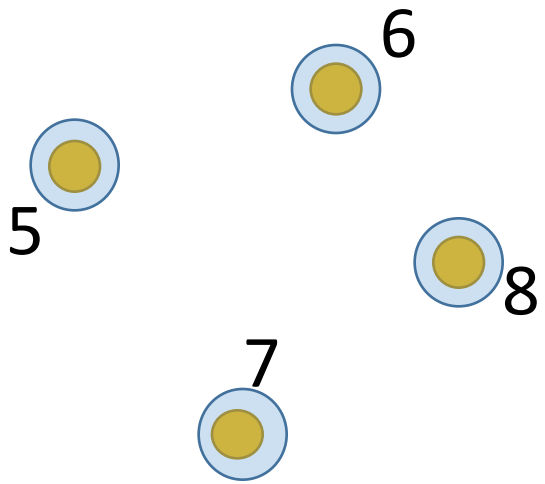


Дендрограмма

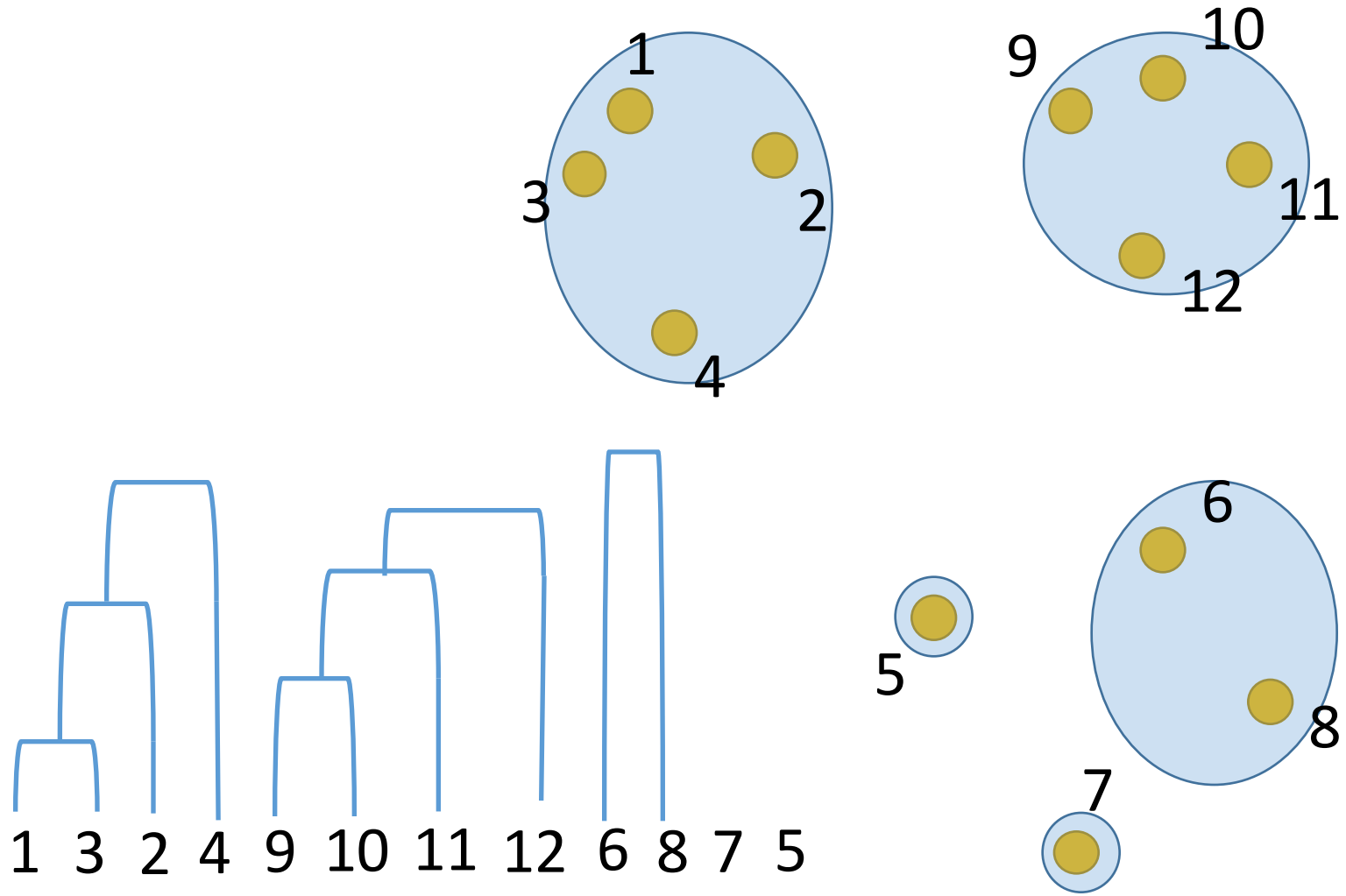




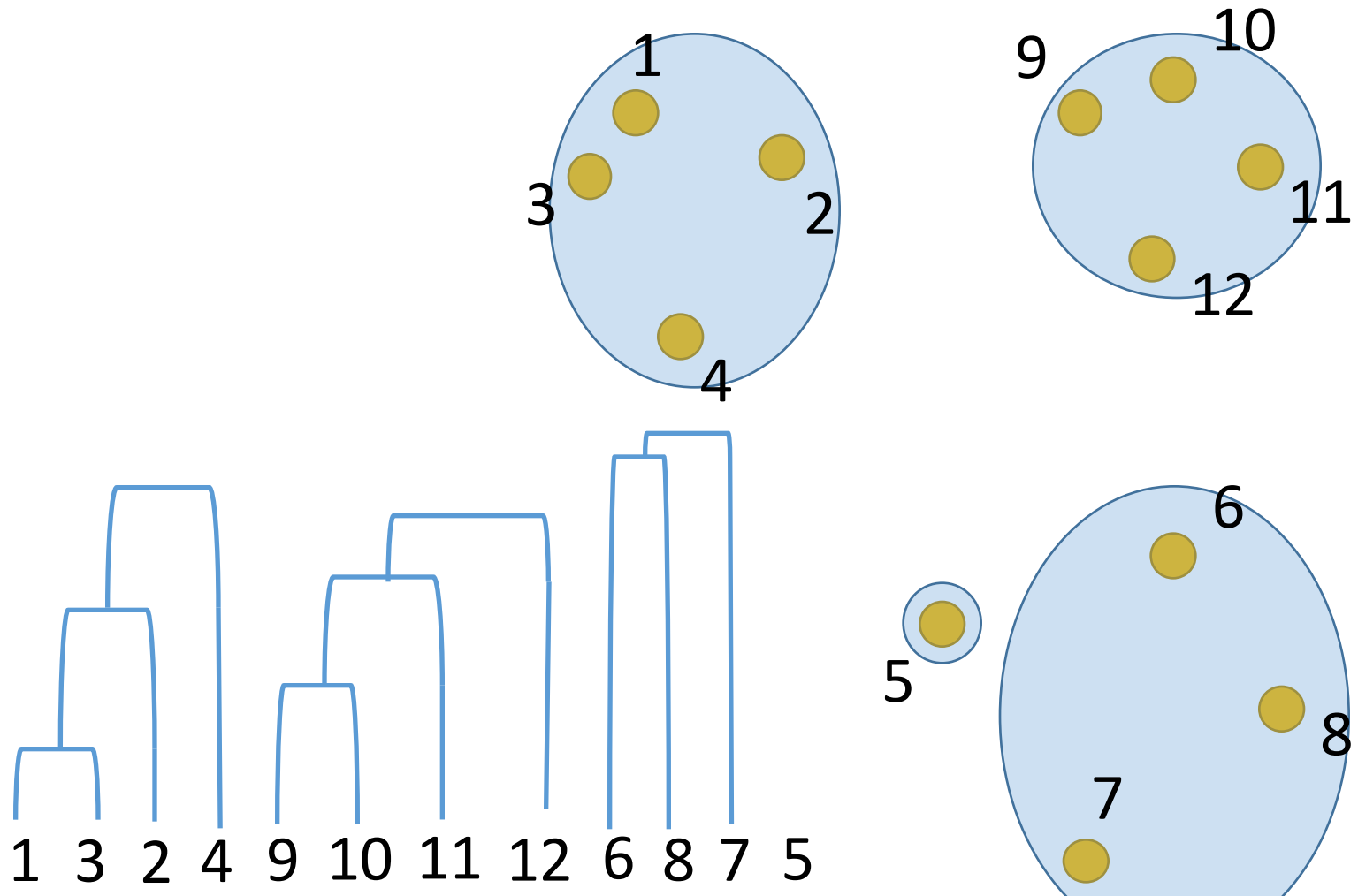




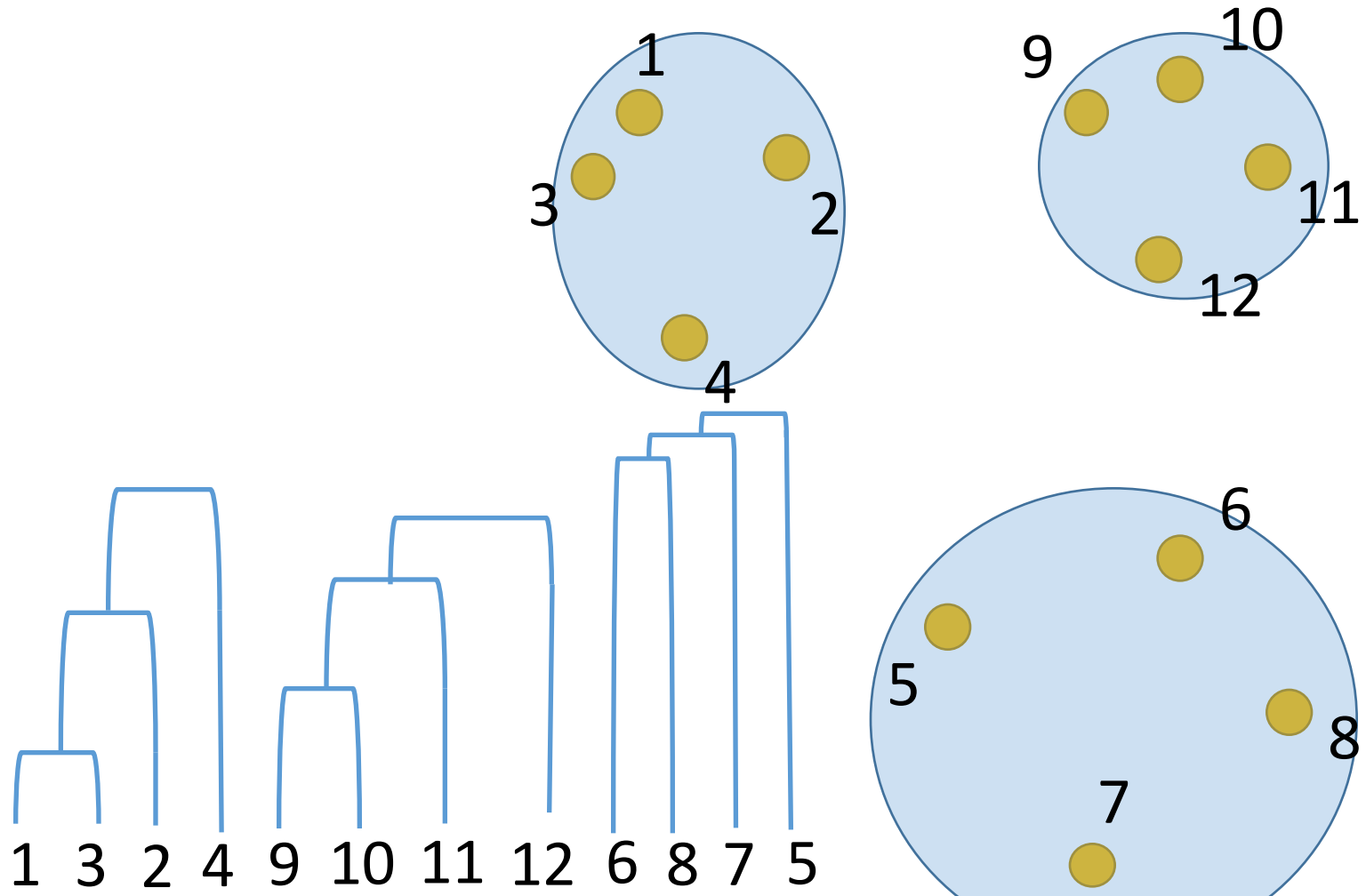
Дендрограмма



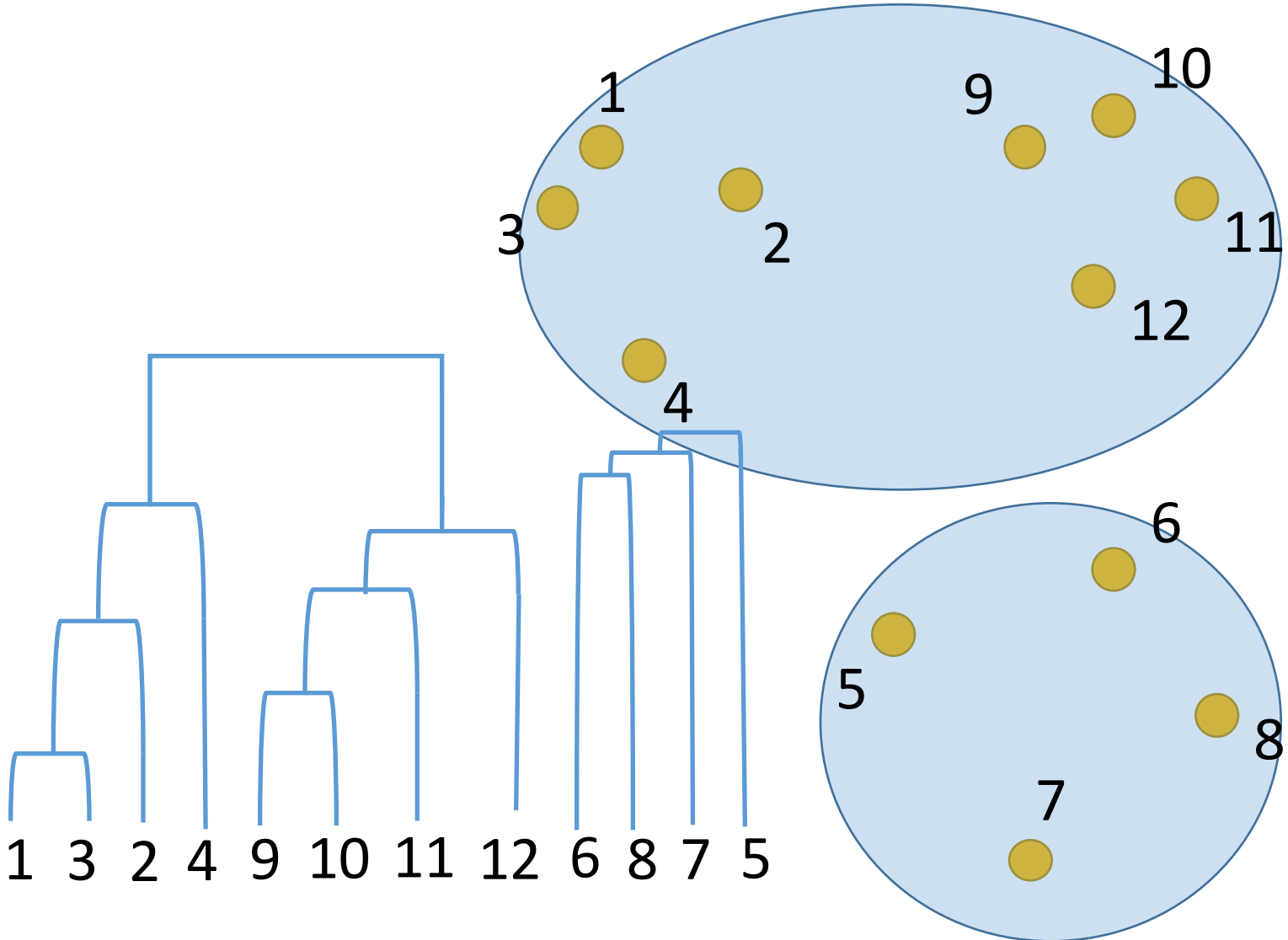
Дендрограмма



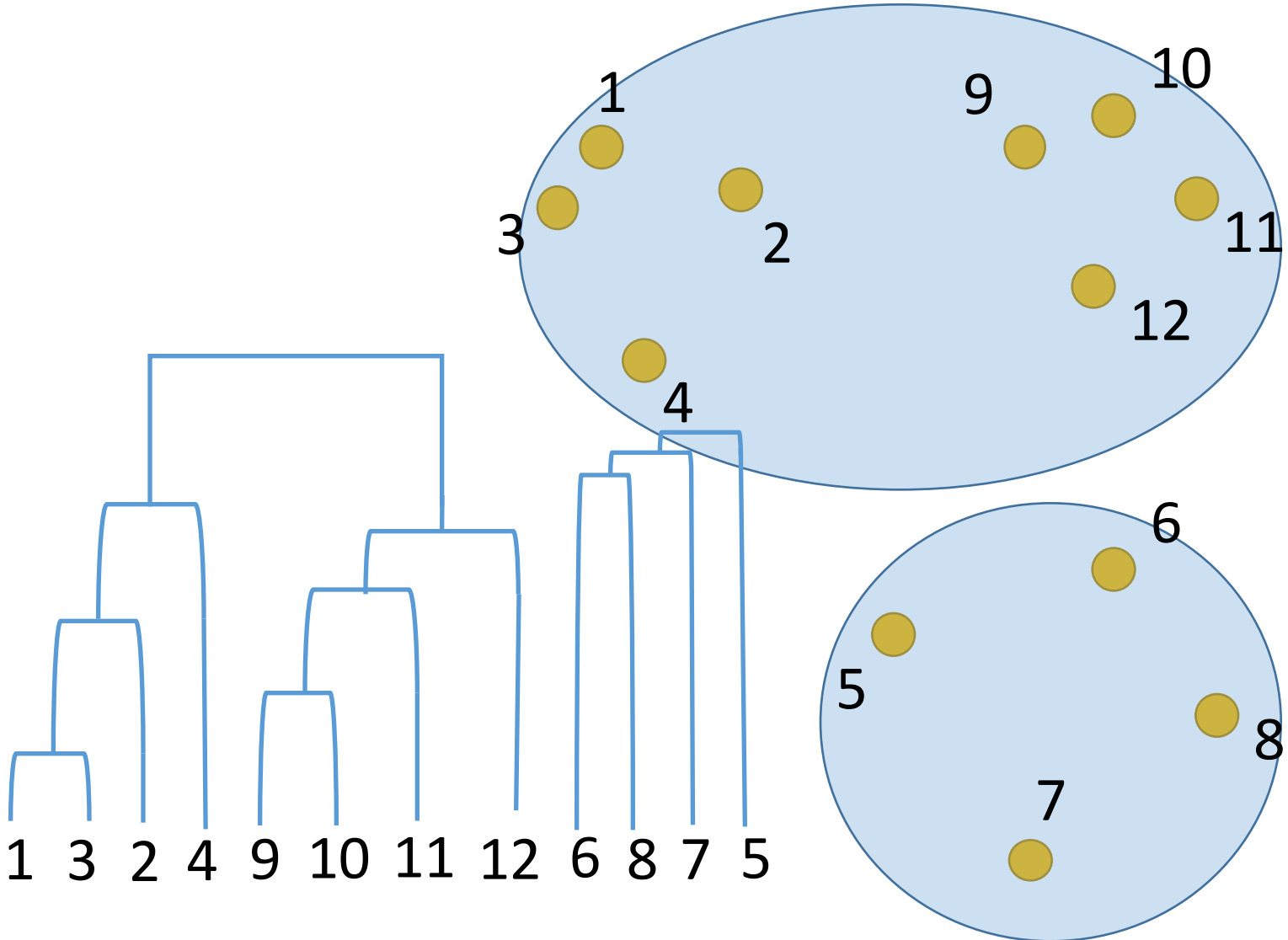
Дендрограмма



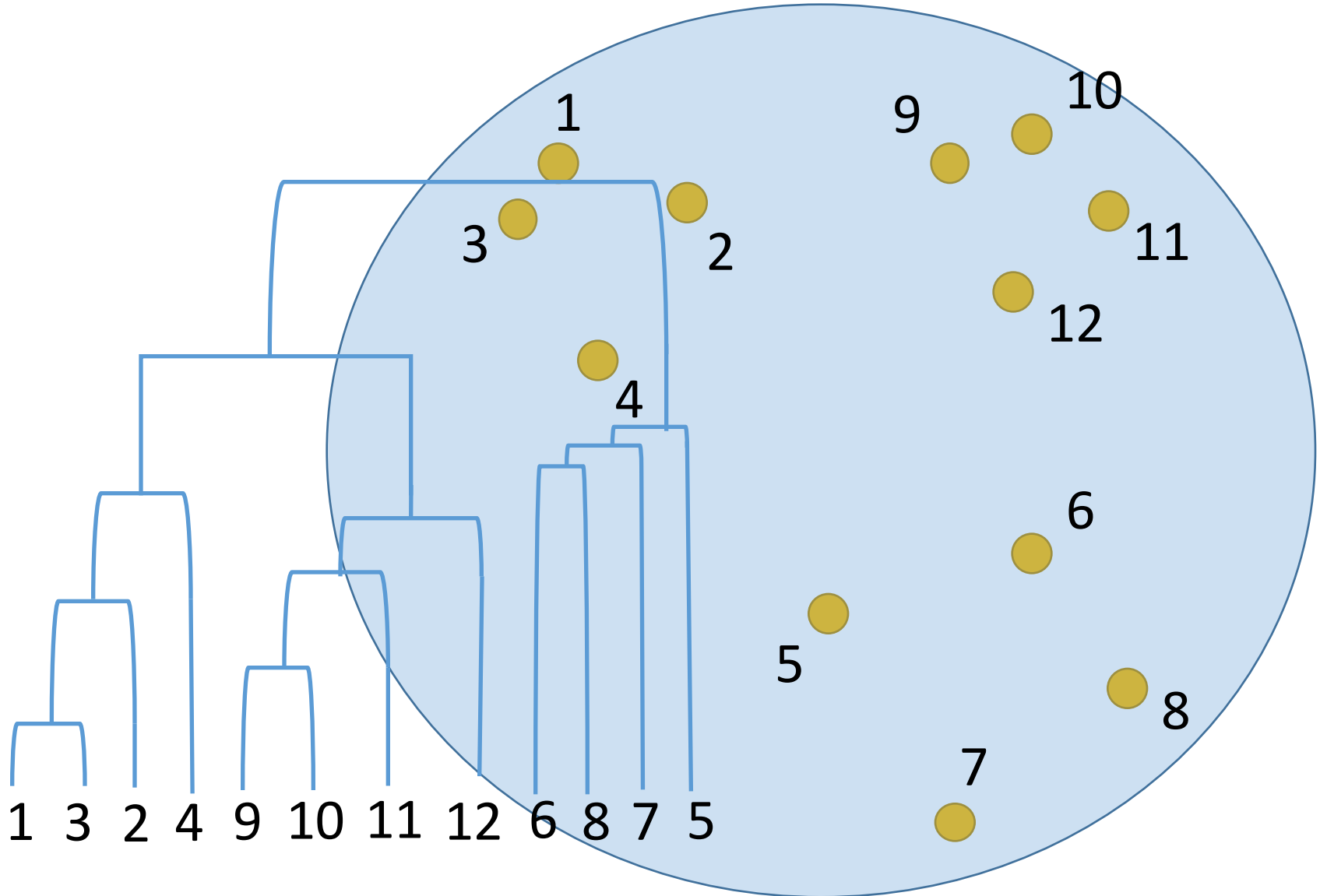
Дендрограмма



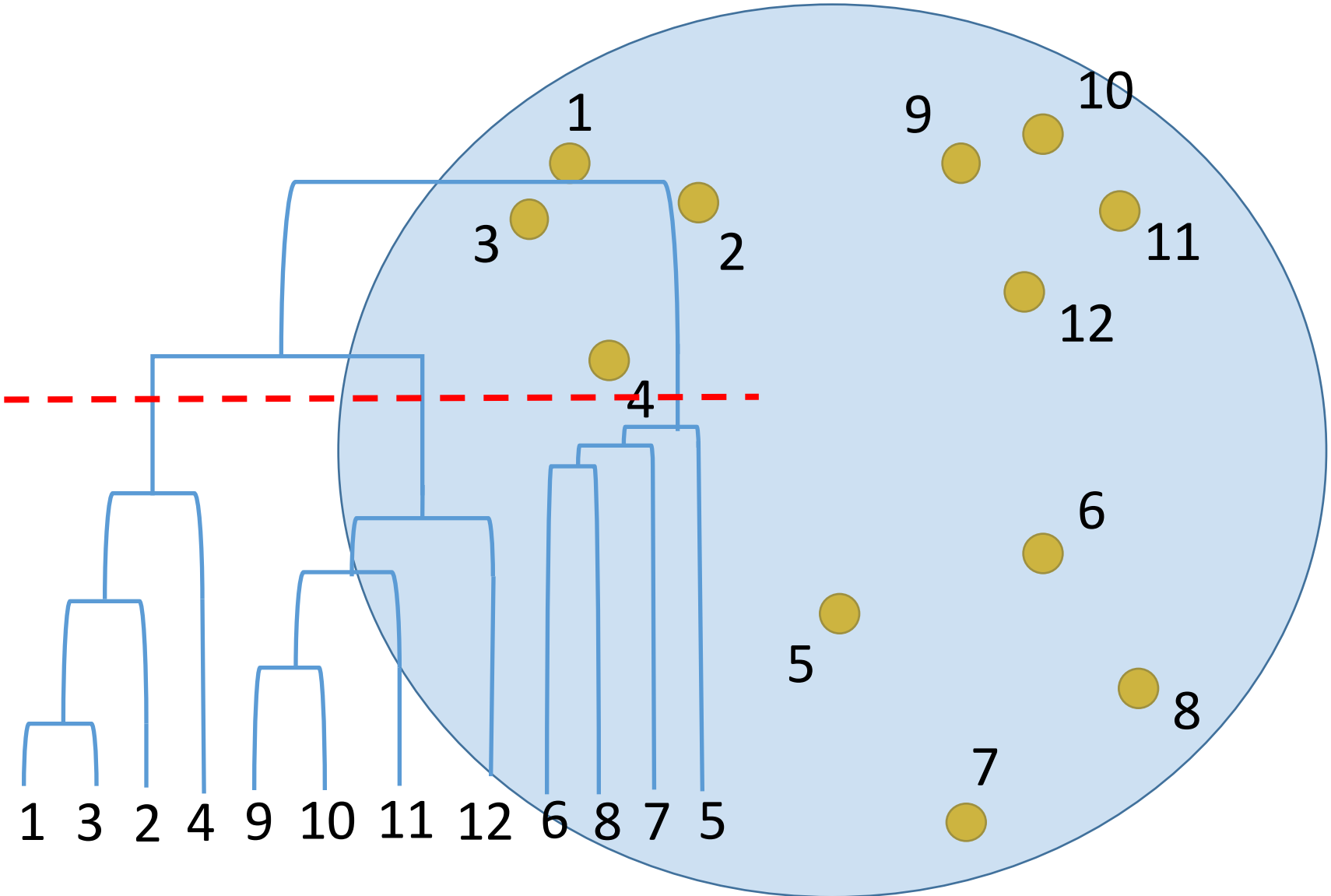
Дендрограмма



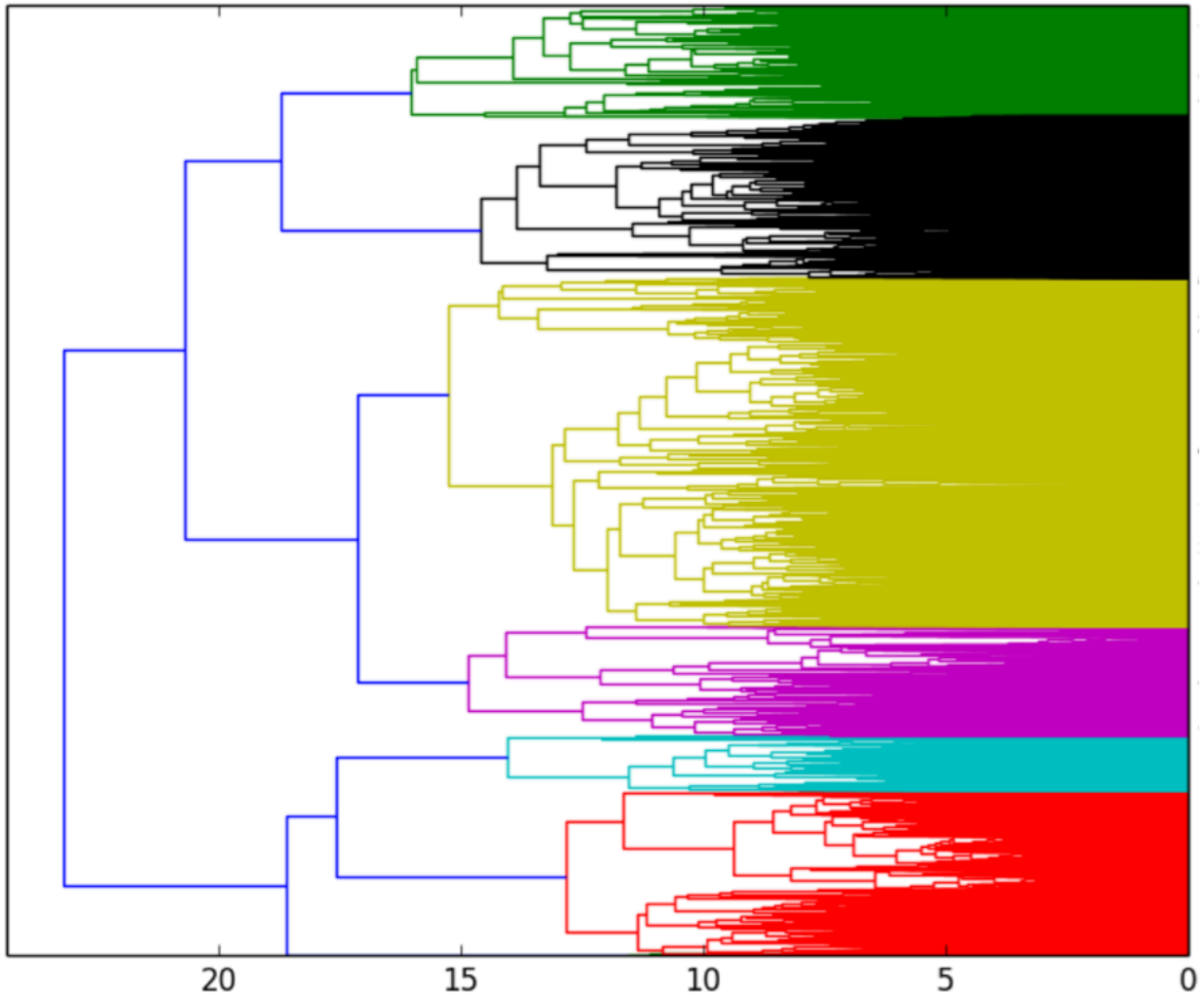
Дендрограмма



Дендрограмма

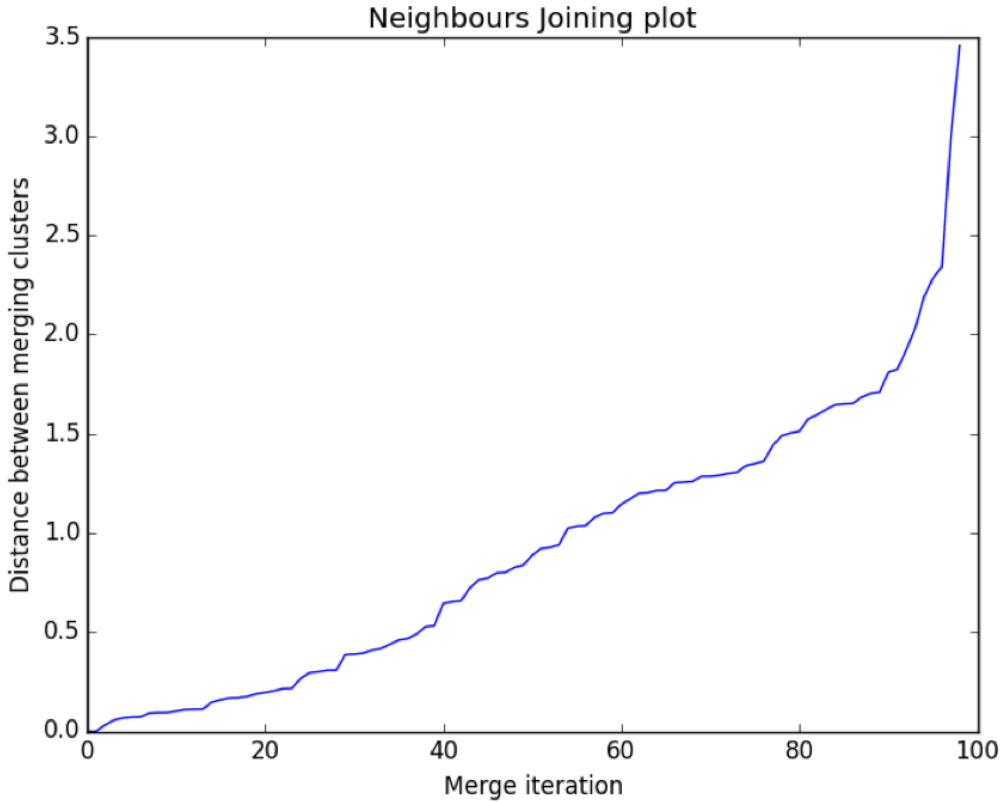


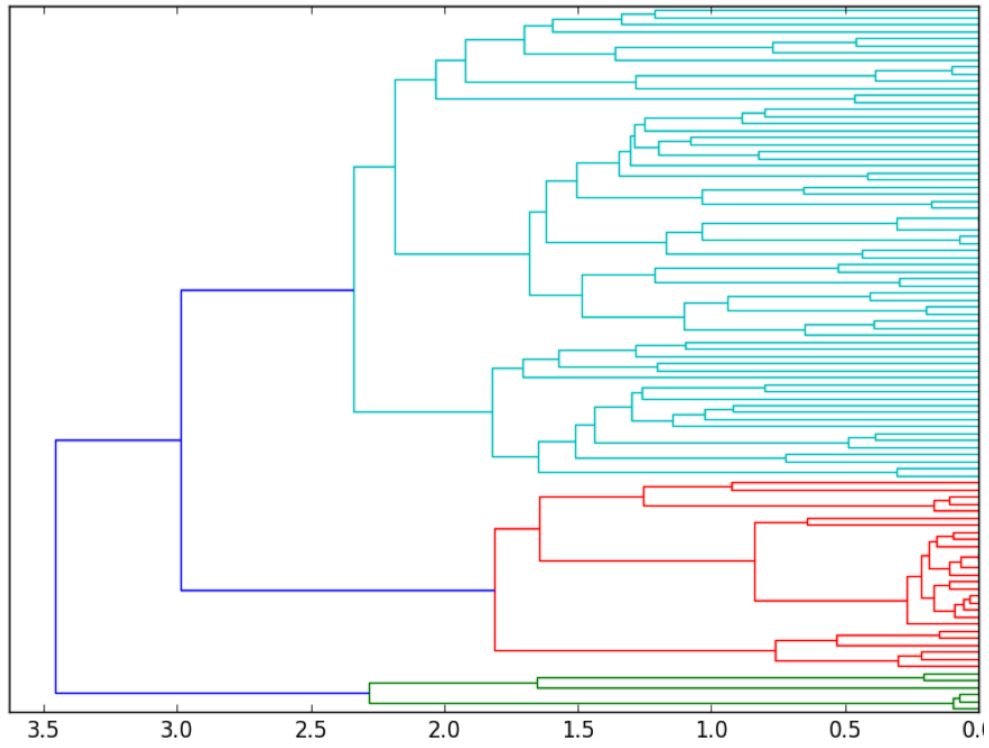
Пример: кластеризация писем



Пример: расстояние между кластерами

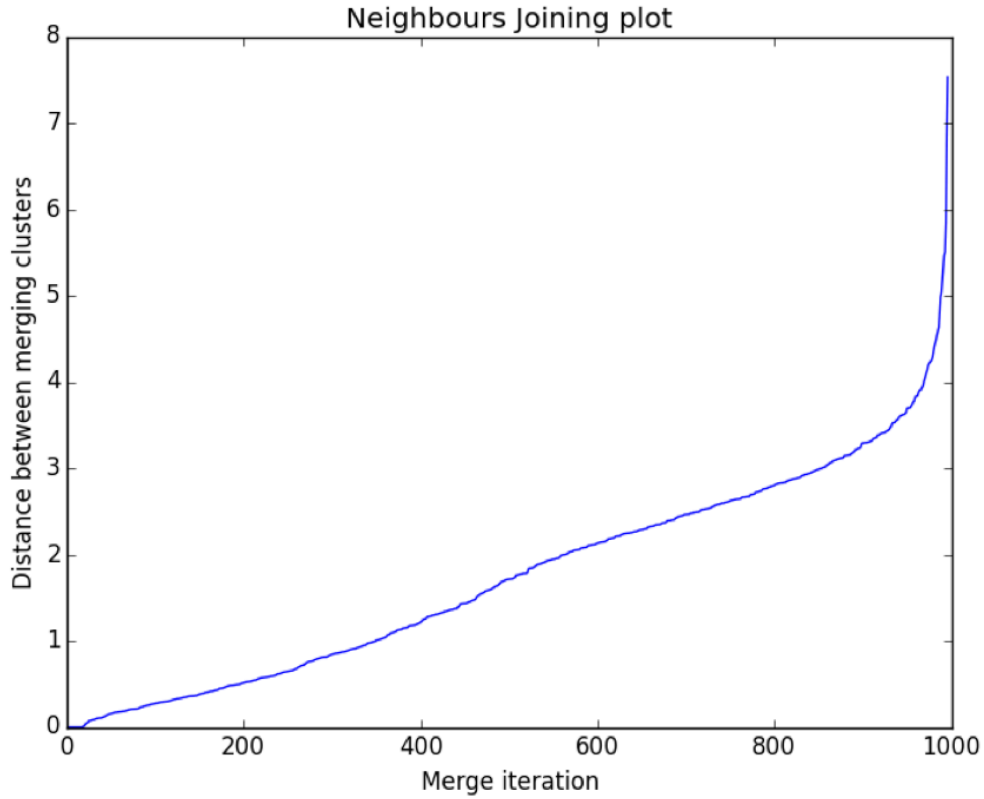
- На подвыборке из 100 писем

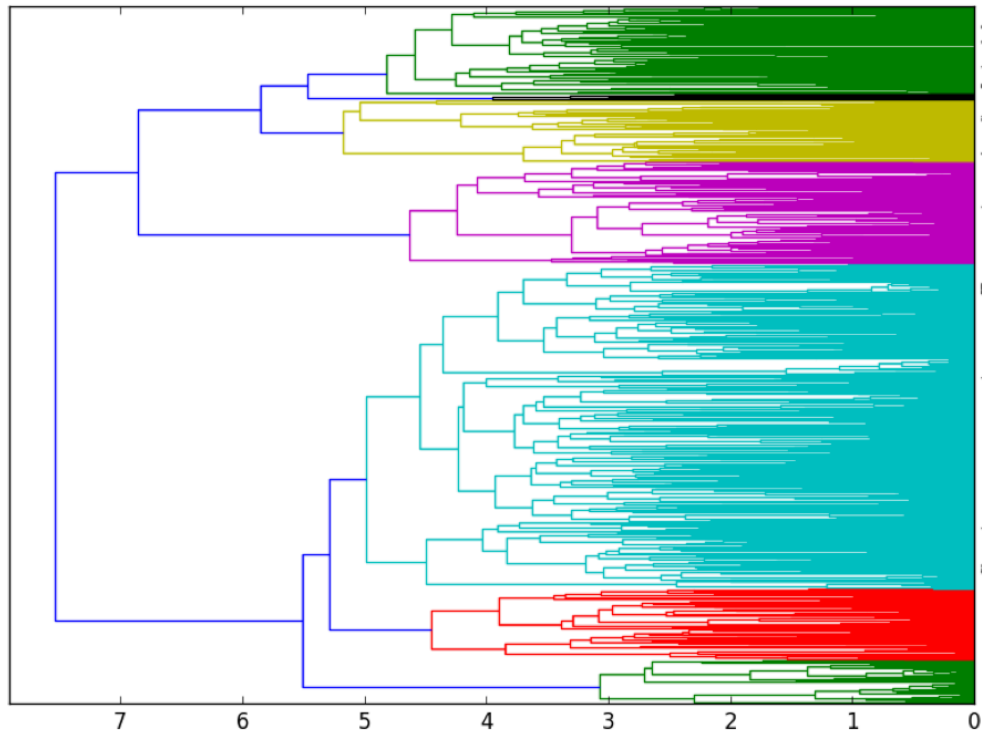




Пример: расстояние между кластерами

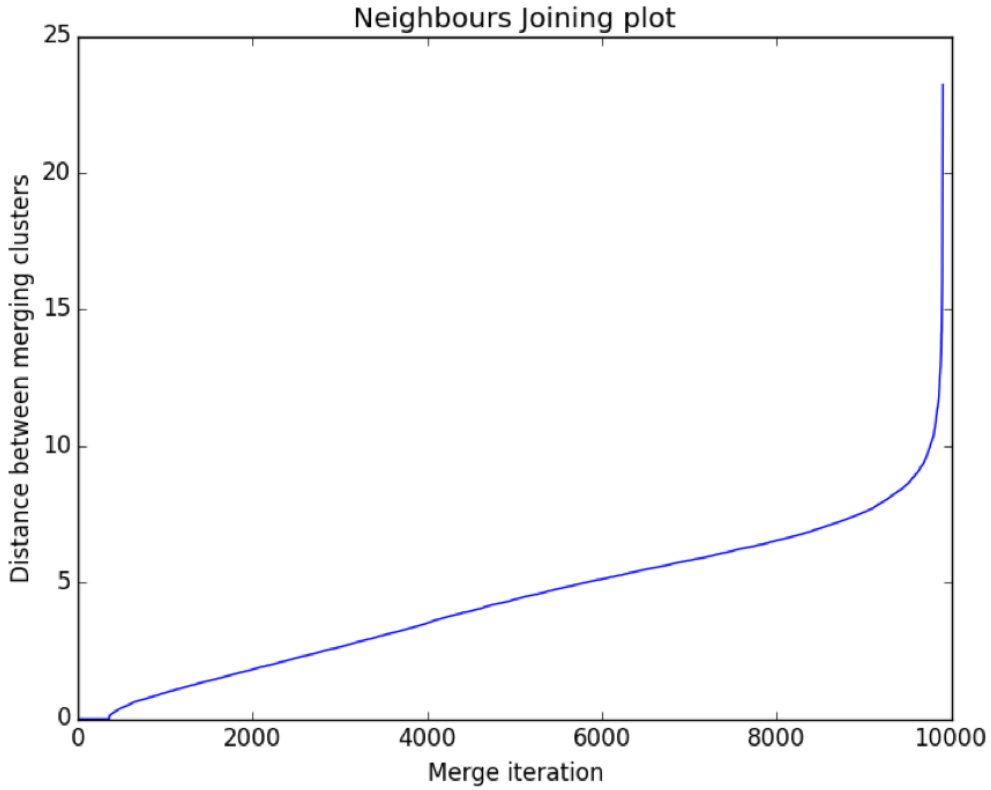
- На подвыборке из 1000 писем

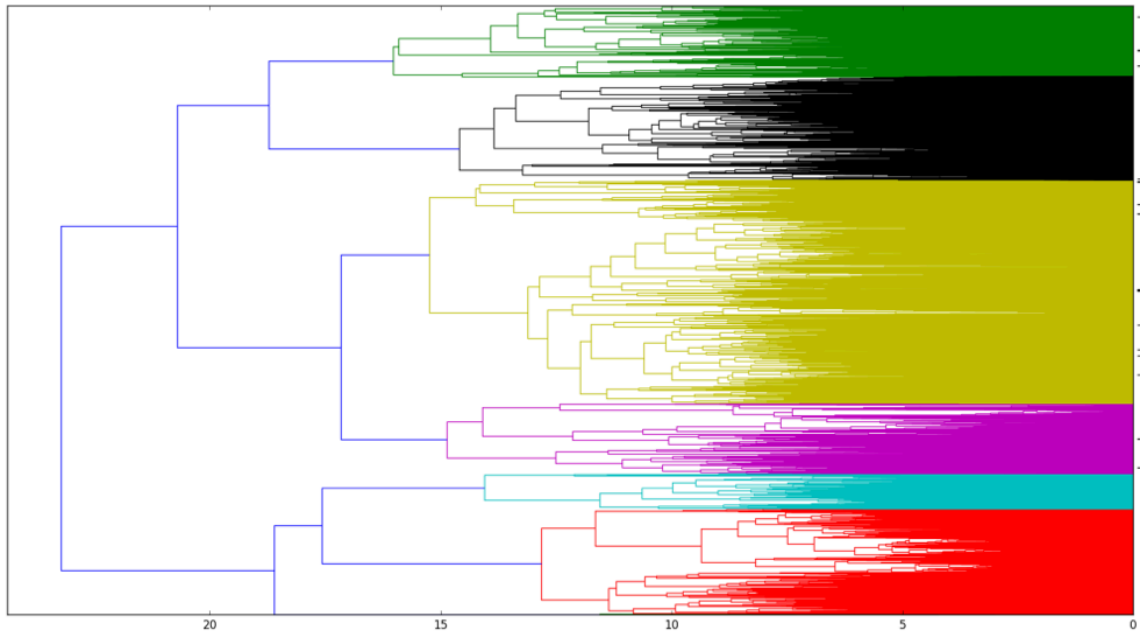




Пример: расстояние между кластерами

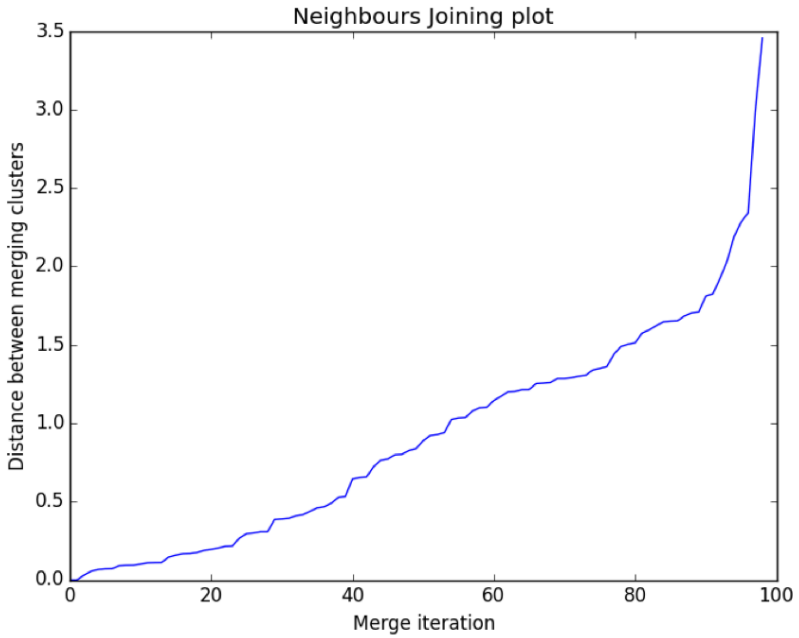
- На подвыборке из 10000 писем

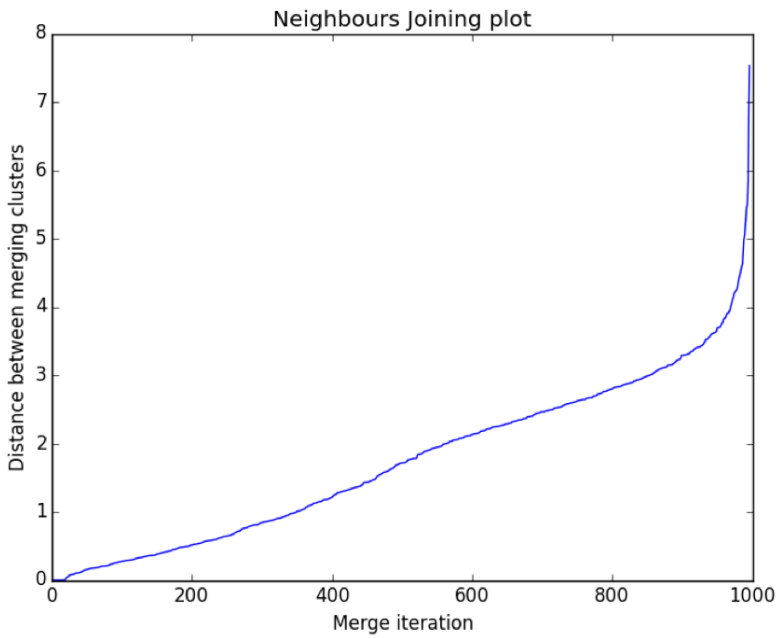


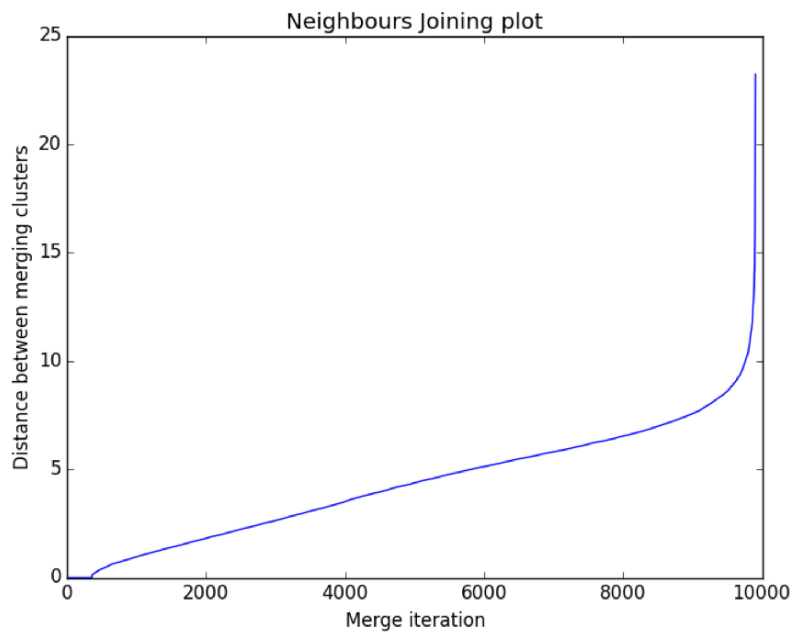


Пример: расстояние между кластерами

- Сравним графики: 100, 1000, 10000 писем

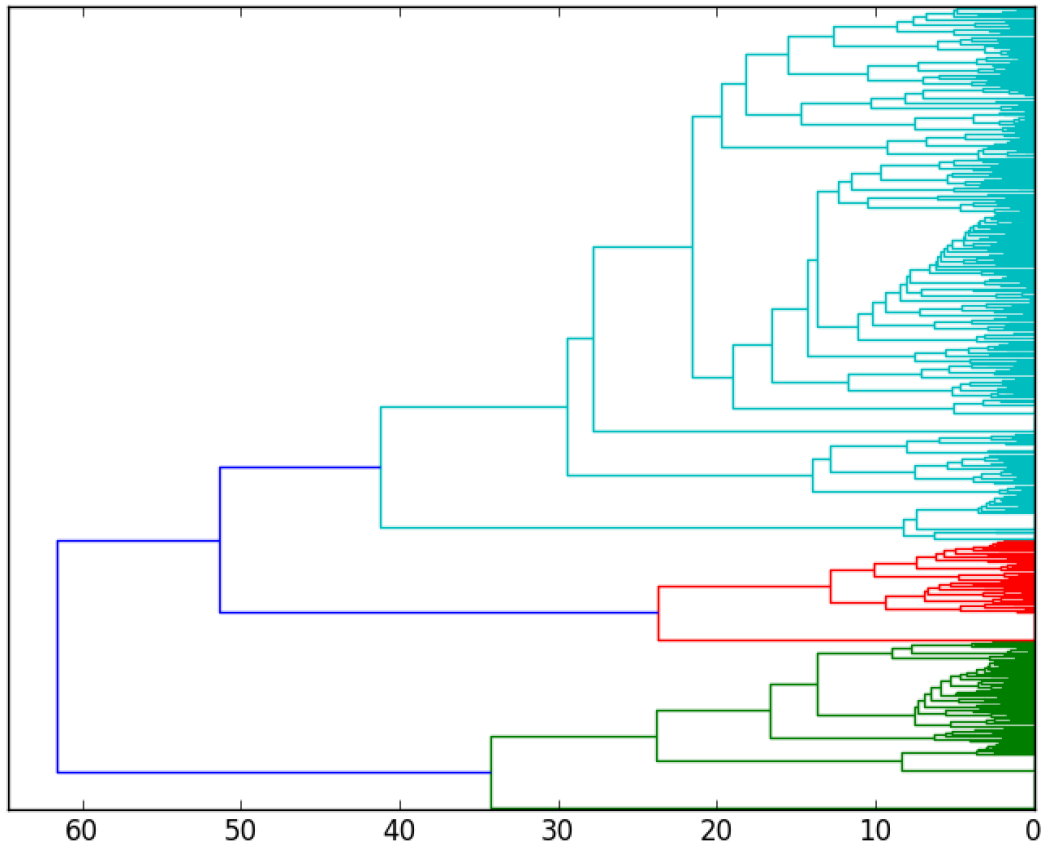




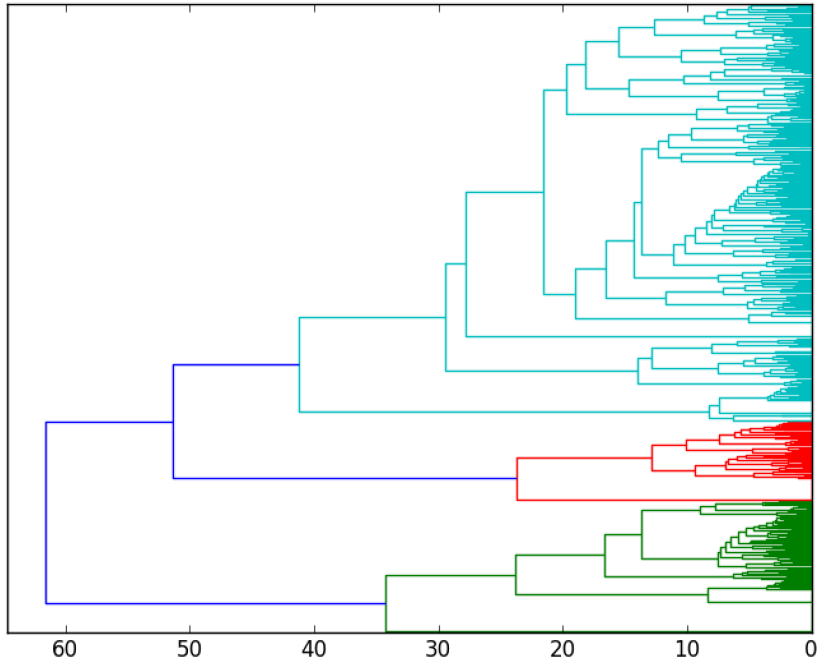


Пример: перекос в размерах кластеров

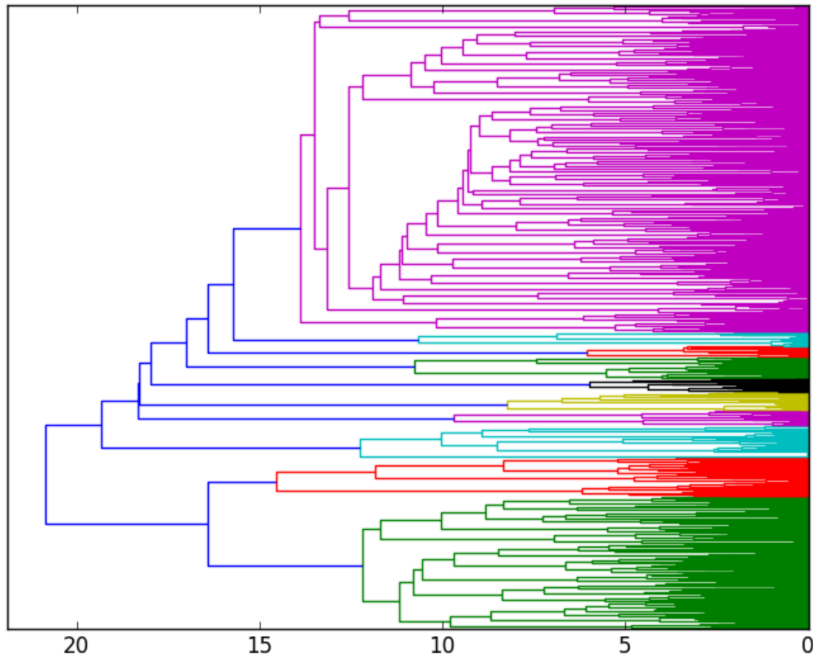
- Дендрограмма, построенная для другой выборки текстов:



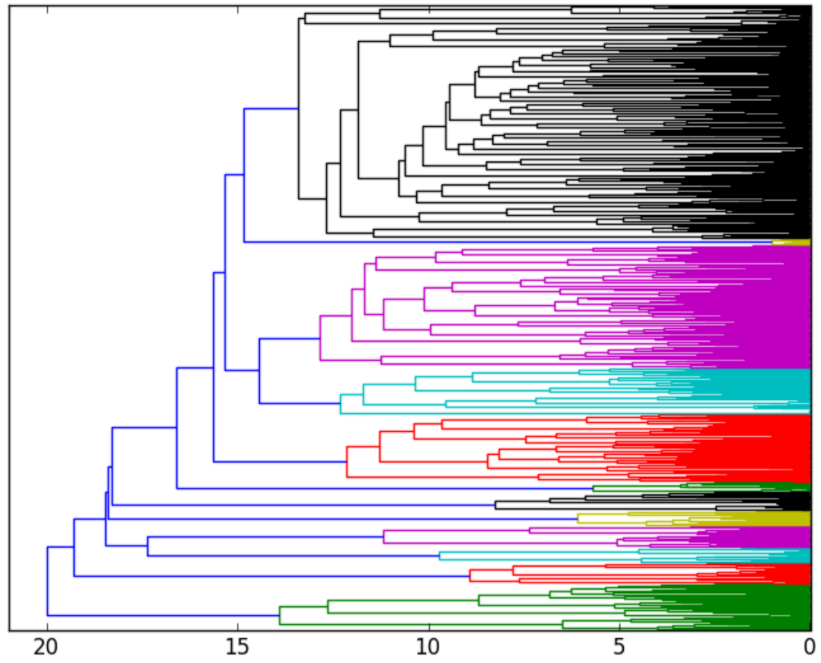
Пример: добавляем SVD



Исходные признаки

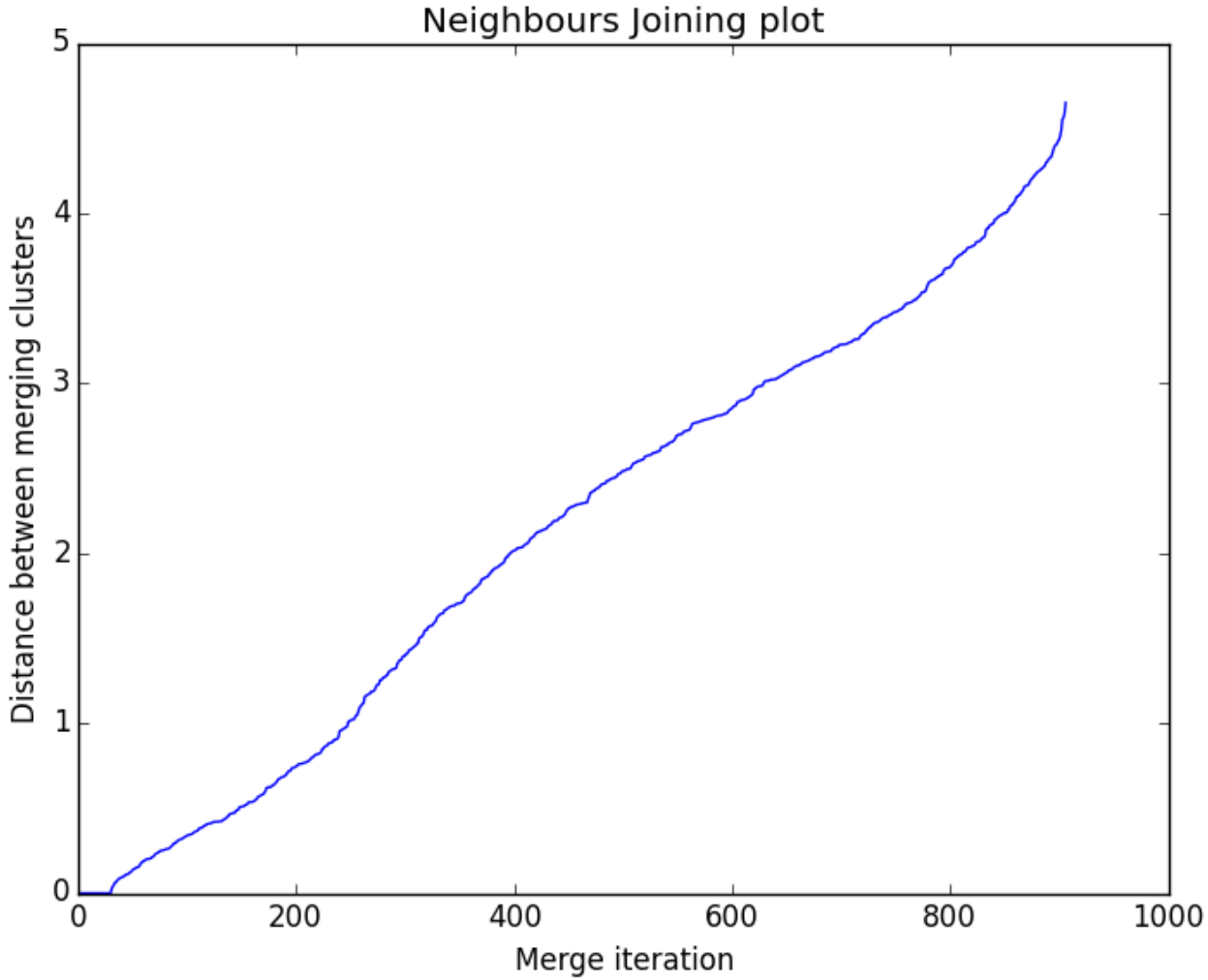


SVD



SVD (еще меньше компонент)

Пример: SVD и расстояние при слиянии



Резюме

1. Иерархическая кластеризация
2. Как устроена агломеративная кластеризация
3. Расстояние между кластерами
4. Формула Ланса-Уильямса
5. Дендрограммы
6. Примеры работы