Кластеризация и визуализация

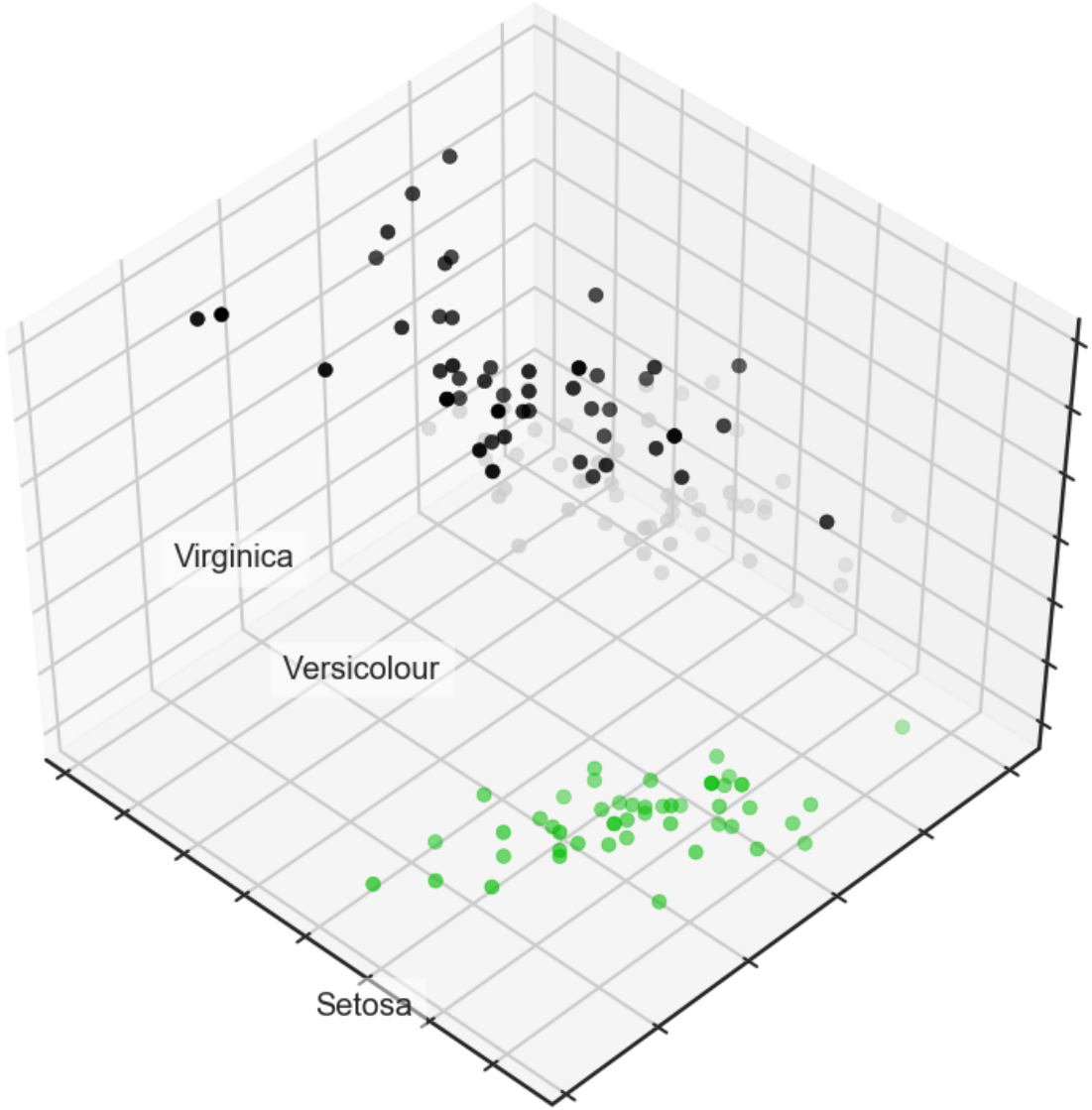
Московский физико-технический институт, МФТИ

Москва

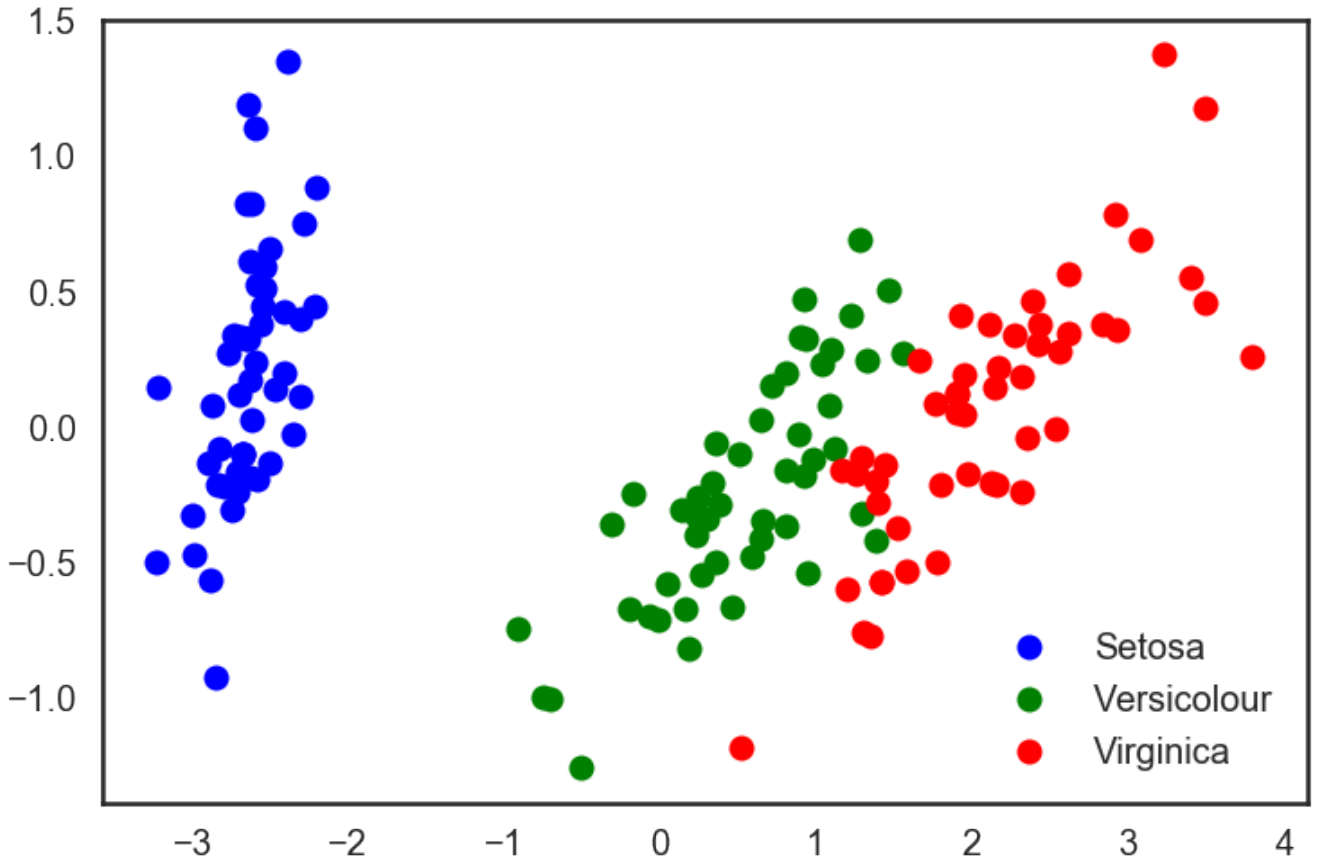
План лекции

- визуализация, метод главных компонент (PCA), t-SNE;
- алгоритм k- средних;
- алгоритм FOREL;
- статистический EM;
- агломеративная кластеризация.

Цветки ириса 3D

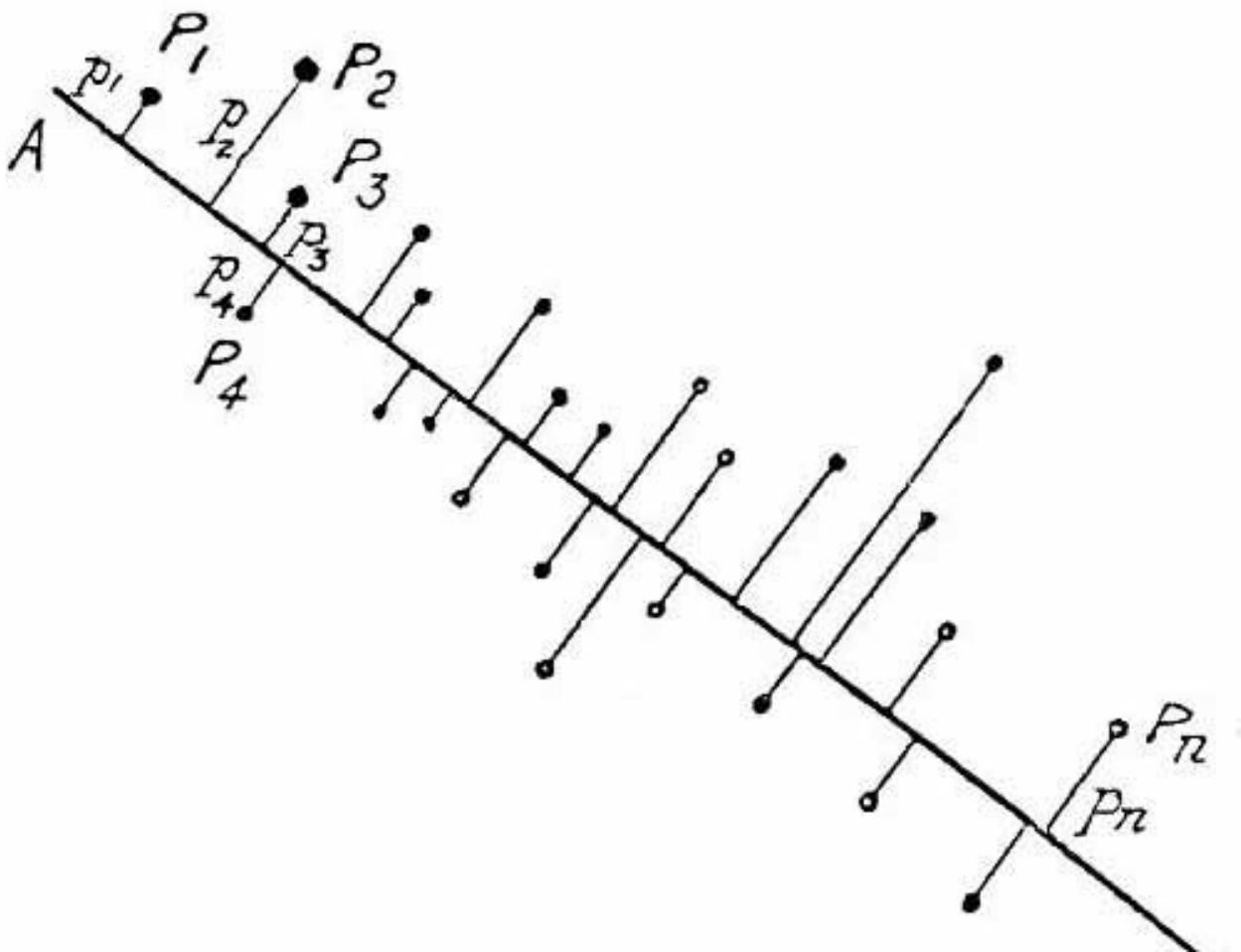


Цветки ириса 3D

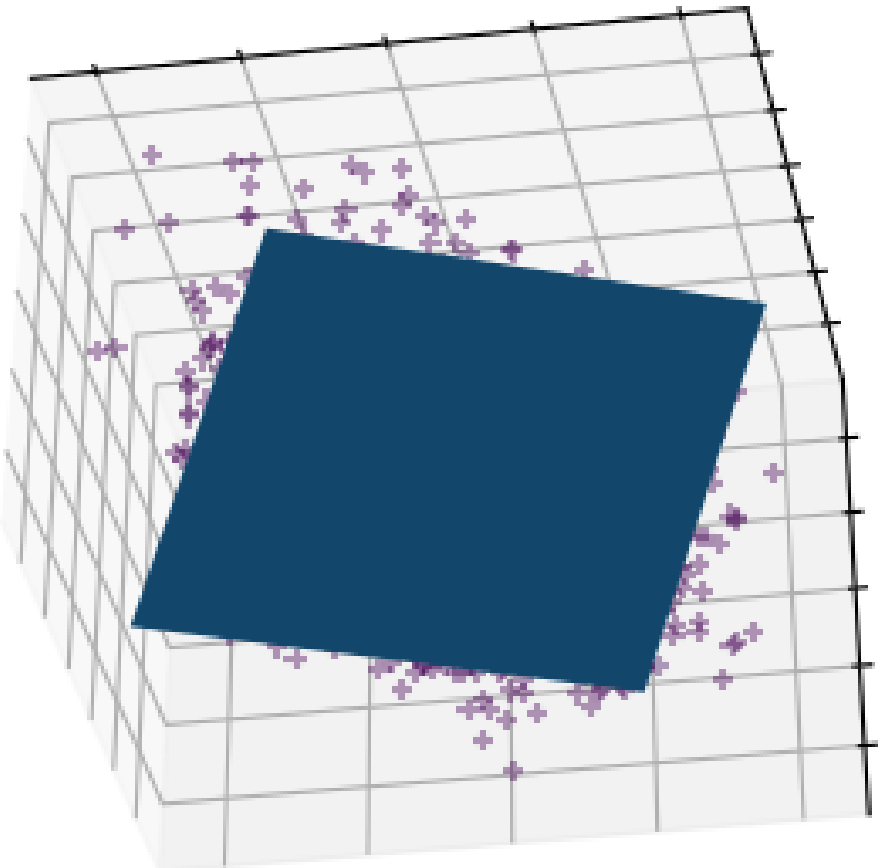


Мат основы метода PCA

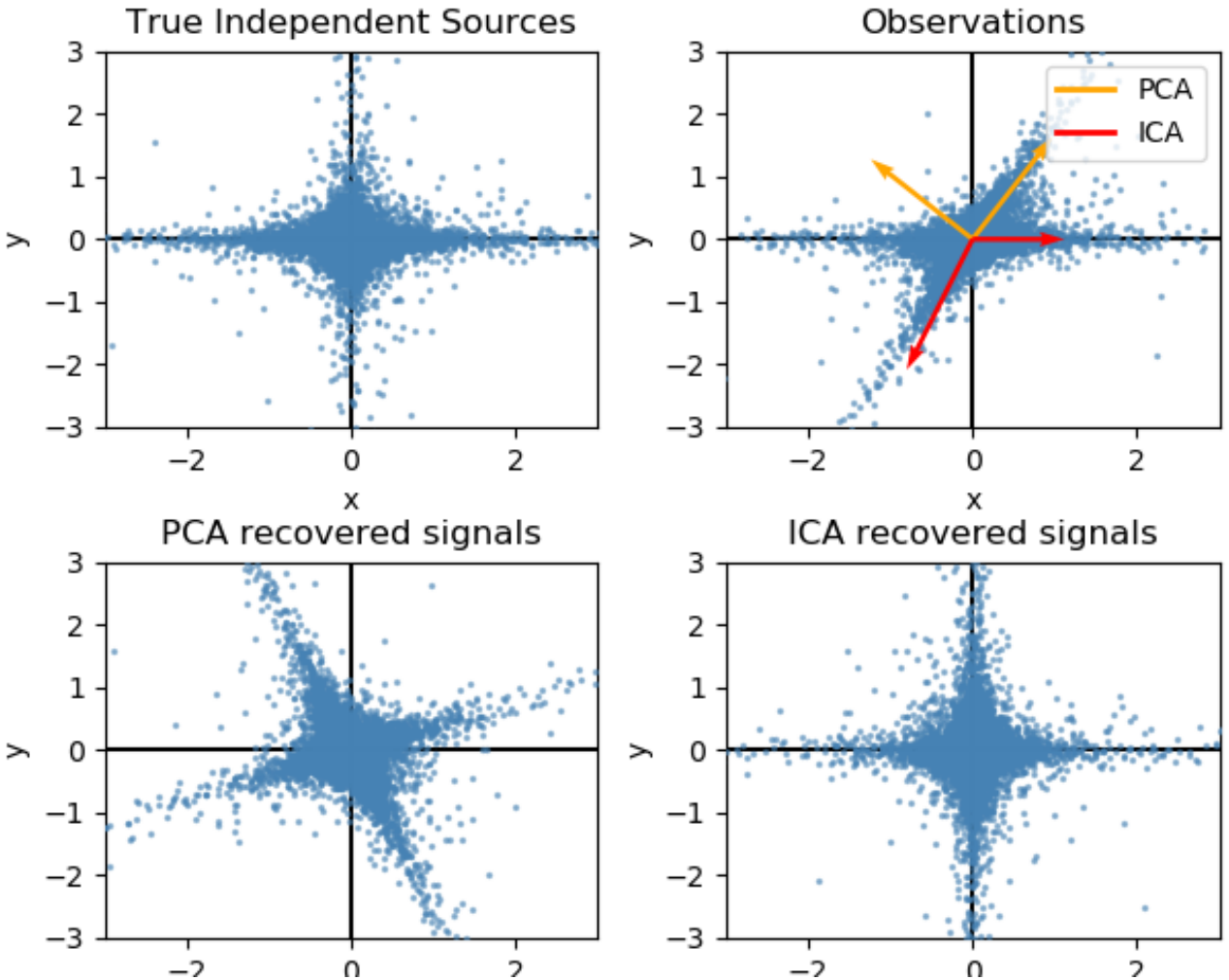
1. Поиск ортогональных проекций с наибольшим рассеянием
2. Диагонализация ковариацинной матрицы
3. сингулярное разложение матрицы данных



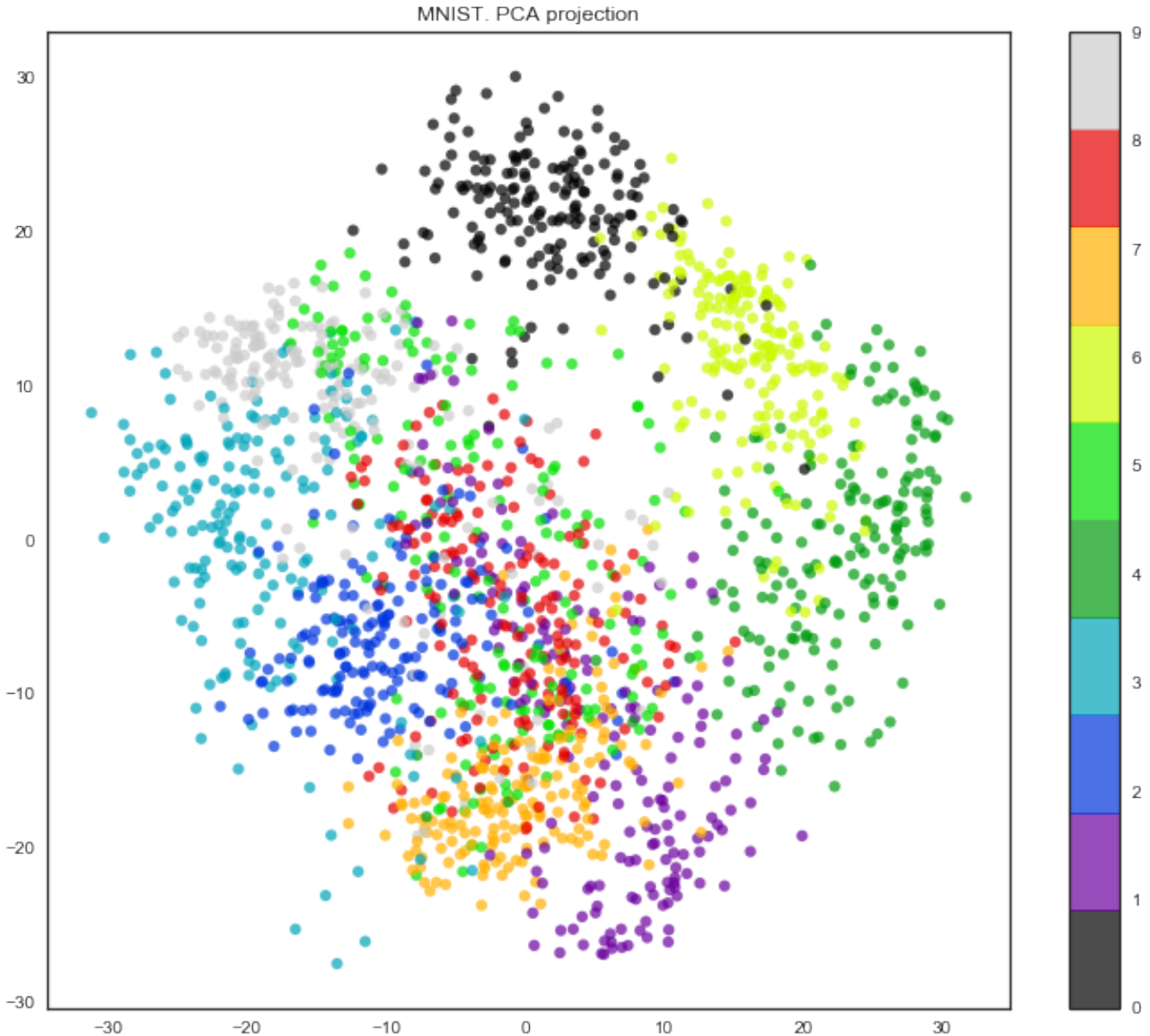
PCA 3d



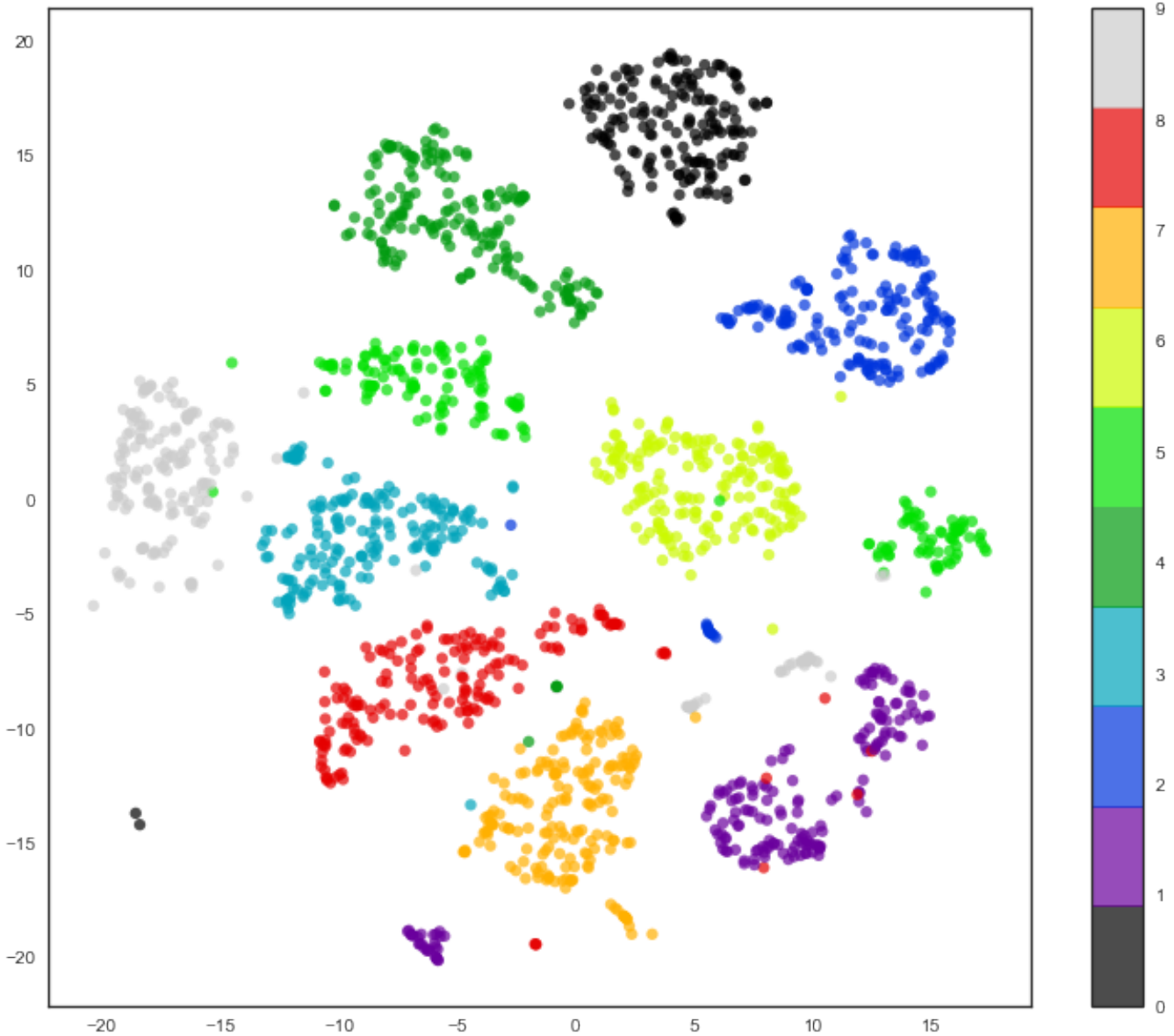
ICA vs PCA


True Independent Sources
Observations
PCA
ICA
PCA recovered signals
ICA recovered signals
x
y

Набор данных рукописных чисел PCA



Набор данных рукописных чисел t-SNE



Мат основа метода t-SNE

Постановка задачи: Есть набор данных с точками, описываемые многомерное переменной с размерностью пространства существенно больше трех. Необходимо получить новоую переменную, существующую в двумерном или трехмерном пространстве, которая бы в максимальной степени сохранила структуру и закономерность в исходных данных. Сопоставим каждой дистанции между точками x_i исходного многомерного пространства X ее веротностный аналог:

$$p_{j|i} = \frac{\exp(-||x_i - x_j||/2\sigma_i^2)}{\sum_{k \neq i} \exp(-||x_i - x_k||/2\sigma_i^2)},$$

в проекционом пространстве размерности 2 или 3 точкам x_i, x_j сопоставим y_i, y_j:

$$q_{j|i} = \frac{\exp(-||y_i - y_j||)}{\sum_{k \neq i} \exp(-||y_i - y_k||)}.$$

Мат основа метода t-SNE

Если одни вероятности $p_{j|i}$ эквивалентны другим $q_{j|i}$, то в виде меры качечтва предлагается использовать расстояние Кульбака-Лейблера:

$$QL = \sum_i \sum_j p_{j|i} \log \frac{p_{j|i}}{q_{j|i}},$$

градиент которой считается достаточно просто,

$$\frac{\partial QL}{\partial y_i} = 2 \sum_j (p_{j|i} - q_{j|i} + p_{i|j} - q_{i|j})(y_i - y_j),$$

а итеритивный процесс проходит с учетом сохранения моментов:

$$Y^t = Y^{t-1} + \eta \frac{\partial QL}{\partial Y} + \alpha(t)(Y^{t-1} - Y^{t-2}).$$

Математическое описание задачи

Исходные данные: X - пространство обьектов, $X^l = \{x_i\}_{i=1}^l$ - обучающая выборка, $\rho : X \times X \rightarrow$ - метрика расстояния между обьектами.
Требуется определдить: Y, $a : X -> Y$-алгоритм кластеризации, такой что – каждый кластер состоит из близких обьектов,
– обьекты разных кластеров существенно отличаются. Другое название - **обучение без учителя**.

Проблема неоднозначности задачи кластеризации

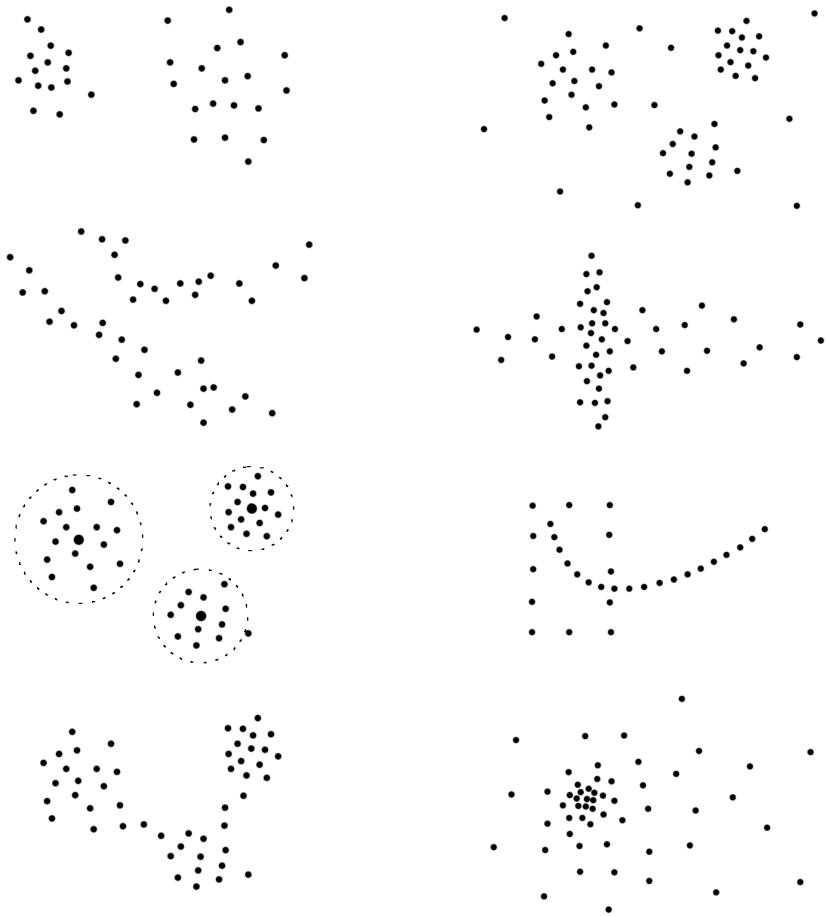
- нет точной постановки задачи кластеризации;
- существует много критериев качетва кластеризации (близко- различные)
- существует много эвристических методов кластеризации;
- число кластеров $|Y|$ как правило не извстно зараннее;
- результат кластеризации существенно зависит от метрики ρ, которую эксперт задает субьективно.

Цели кластеризации

- **Упростить дальнейшую обработку данных**, разбить множество выборки на группы схожих обьектов, чтобы работать с каждой группой в отдельности;
- **Сократить обьем хранимых данных** оставив по одному представителю от каждого кластера;
- **Выделить нетипичные обьекты**, которые не подходят ни к одному из кластеров;
- **Построение иерархии обьектов**.

Какие цели кластеризации??

Разновидности кластеров



Чем могут отличаться задачи кластеризации

- формы кластеров, которые нужно выделять;
- необходимость вложеннности кластеров;
- размер кластеров;
- конечная задача или вспомогательная;
- жесткая или мягкая кластеризация;

Алгоритм K средних (K-means)

K-means итеративно минимизирует среднее внутрикластерное расстояние.

1. Обьект присваивается к тому кластеру, центр которого ближе
2. центр кластера перемещается в центр среднего арифметического координато векторов.

K-means, выбор центров

В зависимости от начального положения цетров - разные результаты и время сходимости. Варианты выбора центов: случайно, подальше друг от друга, для двух кластеров, выбор начального приближения:

- первый центр выбираем случайно из равномерного распределения на выборке;
- каждый следующий центр выбирается из оставшихся точек так, что вероятность выбрать каждую точку была пропорциональна квадрату расстояния от нее до ближайшего центра.

Функционалы качества кластеризации

Для **метрических пространств** можно ввести следующие функционалы:
Среднее внутрикластерное расстояние

$$F_0 = \frac{\sum_{i<j}[y_i = y_j]\rho(x_i, x_j)}{\sum_{i<j}[y_i = y_j]} \to \min,$$

Среднее межкластерное расстояние:

$$F_1 = \frac{\sum_{i<j}[y_i \neq y_j]\rho(x_i, x_j)}{\sum_{i<j}[y_i \neq y_j]} \to \max,$$

Также можно рассмотреть метрику

$$F_0/F_1 \to 0.$$

Функционалы качества кластеризации

Для **линейных пространств** можно посчитать найти центры кластеров μ, $y \in Y$, тогда для внутрикластерных расстояний:

$$\Phi_0 = \sum_{y \in Y} \frac{1}{|K_y|} \sum_{i:y_i=y} \rho^2(x_i, \mu_y) \to \min,$$

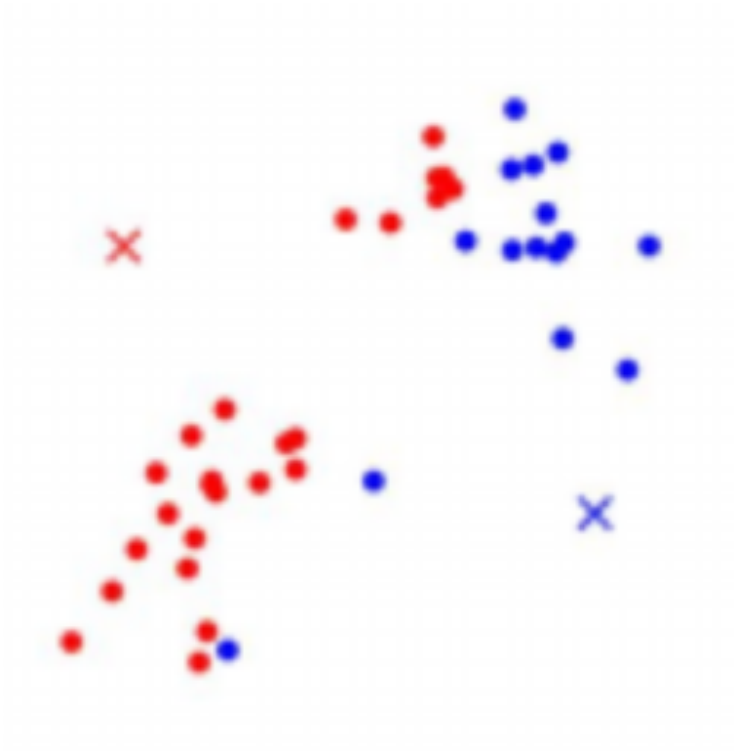
аналоично максимизируем межкалстерные расстояния:

$$\Phi_0 = \sum_{y \in Y} \rho^2(\mu_y, \mu) \to \max.$$

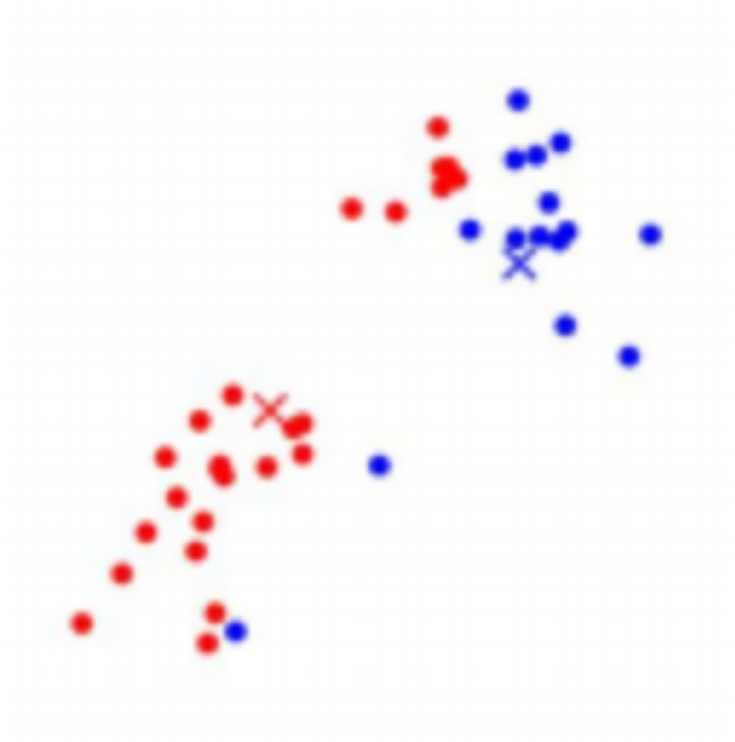
Как работает K Means

Как работает K Means

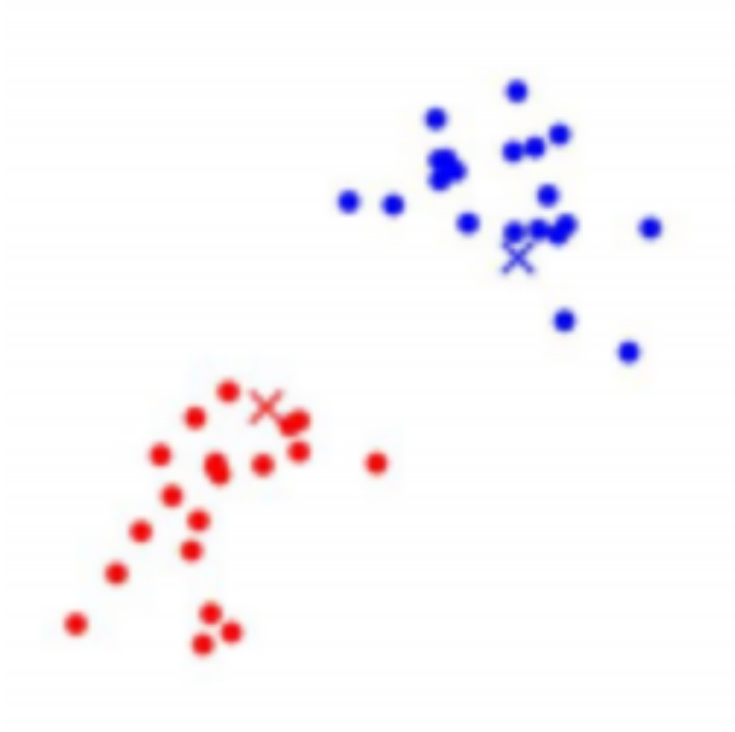
Как работает K Means



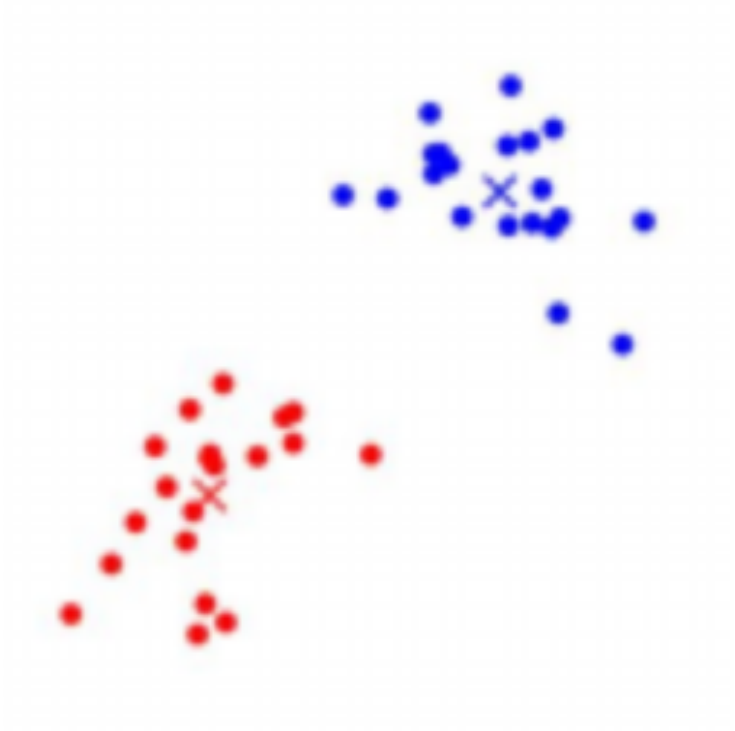
Как работает K Means



Как работает K Means



Как работает K Means



Алгортм FOREL

1. $U = X^l$ - множество некластеризованных точек;
2. **While** в выборке есть некластеризованные точки $U \neq 0$ взять случайную точку $x_0 \in U$
3. **Repeat**
4. образовать кластер с центром в x_0 и радиусом R

$$K_0 = \{x_i \in U | \rho(x_i, x_0) \leq R\}$$

5. переместить центр x_0 в центр масс кластера:

$$x_0 = \frac{1}{|K_0|} \sum_{x_i \in K_0} x_i$$

6. **While** состав класса K_0 не стабилизируется
7. $U = U \setminus K_0$;
8. применить алгоритм один из графовых алгоритмов, например кратчайшего незамкнутого пути к множеству центров кластеров;
9. каждый приписать кластеру с ближайшим центром.

EM - распределения

Входные данные: априорные вероятности кластеров $w_1, .., w_k$, плотности распределения кластеров $p_1(x), .., p_k(x)$
Тогда плотность распределения вектора признаков x можно записать в виде

$$p(x) = \sum_{j=1}^{K} w_j p_j(x).$$

Сделаем предположение о том, что плотности имеют Гауссов вид:

$$p_j = (2\pi)^{n/2}(\sigma_{j1}...\sigma_{j1})^{-1} \exp\left(-0.5\rho_y^2(x, \mu_y))\right),$$

матрица ковариаций Σ_y диагональная,
$\rho_y^2 = \sum_k \sigma_{yk}^{-2}|f_j(x) - f_j(x')|^2$.
Задача: по выборке оценить параметры модели: $w_1, .., w_k$, $p_1(x), .., p_k(x)$.

Описание алгоритма EM 1

1. Начальное приближение для всех кластеров

$$w_y = 1/|Y|,$$

μ_y - случайный обьект выборки

$$\sigma_{yj}^2 = \frac{1}{l|Y|}\sum_{i=1}^{l}(f_j(x_i) - \mu_{yj})^2,\ j = 1..n$$

2. **Loop**
3. E-step

$$g_{iy} = \frac{w_y p_y(x_i)}{\sum_{z \in Y} w_z p_z(x_i)},\ y \in Y,\ i = 1..l.$$

Описание алгоритма EM 2

5. M-step

$$w_y = \frac{1}{l}\sum g_{iy},\ y \in Y;$$

$$\mu_{yj} = \frac{1}{l w_y}\sum_{i=1}^{l} g_{iy} f_j(x_i),\ y \in Y,\ j = 1..n$$

$$\sigma^2_{yj} = \frac{1}{l w_y}\sum_{i=1}^{l} g_{iy}(f_j(x_i) - \mu_{yj})^2,\ j = 1..n$$

6. Отнести обьекты к кластерам по байесовскому решающему правилу

$$y_i = \arg\max_{y \in Y} g_{iy},\ i = 1..l.$$

7. **While** y_i меняется

Сравнение и развитие данных методов

1. вариант Болла-Холла;
2. вариант МакКина: при переходе обьектов из кластеров их центры пересчитываются;

Отличия EM и k-means

1. EM: мягкая кластеризация $g_{iy} = P\{y_i = y\}$; k-m: $g_{iy} = [y_i = y]$;
2. EM: формула кластеров эллиптическая, настраиваемя; k-m: формлуа кластеров жестко определяется метрикой ρ;

Гибридный вариант на пути упрощения EM

1. EM с жесткой кластеризацией на E-шаге;
2. EM без настройки дисперсий.

Недостатки k-means - чувствительность к выбору начального приближения;
- необходимость задавать k;
Как улучшить? - несколько случайных кластеризаций, выбор лучшей по функц качества;
- постепенно наращивание числа кластеров k.

Иерархическая кластеризация

Строится не одно разбиение выборки, а система **вложенных** разбиений. Рассмотрим аглемеративные методы, или **восходящие** алгоритмы, в которых обьекты обьединяются во все более и более крупные кластеры.

Формула Ланса-Уильямса

На каждой итерации образуется новый кластер $W = U \cup V$, расстояние от нового кластера W до любого другого кластера S вычисляется по расстояниям $R(U, V)$:

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

Способы вычисления расстояния между кластерами 1

1. Расстояние ближнего соседа

$$R^n(W,S) = \min_{w \in W, s \in S} \rho(w,s),$$

$$\alpha_U = \alpha_V,\ \beta = 0, \gamma = -0.5,$$

2. расстояние дальнего соседа

$$R^f(W,S) = \min_{w \in W, s \in S} \rho(w,s),$$

$$\alpha_U = \alpha_V,\ \beta = 0, \gamma = -0.5,$$

3. среднее расстояние

$$R^f(W,S) = \frac{1}{|W||S|} \sum_{w \in W} \sum_{s \in S} \rho(w,s),$$

$$\alpha_U = \frac{|U|}{|W|}, \alpha_V = \frac{|V|}{|W|},\ \beta = \gamma = 0,$$

Способы вычисления расстояния между кластерами 2

5. расстояние между центрами (худший вариант!)

$$R^f(W,S) = \rho(\sum_{w\in W}\frac{w}{|W|}, \sum_{s\in S}\frac{s}{|S|}),$$

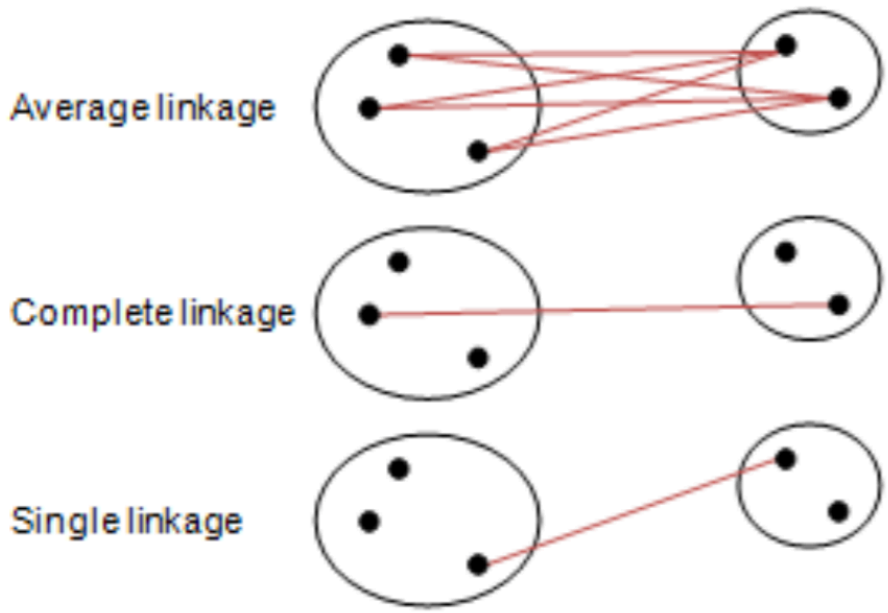
$$\alpha_U = \frac{|U|}{|W|}, \alpha_V = \frac{|V|}{|W|}, \beta = -\alpha_U\alpha_V, \gamma = 0,$$

6. расстояние Уорда

$$R^f(W,S) = \frac{|W||S|}{|W|+|S|}\rho(\sum_{w\in W}\frac{w}{|W|}, \sum_{s\in S}\frac{s}{|S|}),$$

$$\alpha_U = \frac{|S|+|U|}{|S|+|W|}, \alpha_V = \frac{|S|+|V|}{|S|+|W|}, \beta = \frac{-|S|}{|S|+|W|}\,\gamma = 0.$$

Расстояния между кластерами



Агломеративная кластеризация Ланса - Уильямса

1. Инициализация множества кластеров

$$C_t = \{x_1, .., x_l\},$$

2. **Forall**
3. найти в C_{t-1} два ближайшиз кластера

$$(U, V) = \arg \min_{U \neq V} R(U, V),$$

$$R_t = R(U, V)$$

4. изьять кластеры и добавить слитые кластеры $W = U \cup V$

$$C_t = C_{t-1} \cup \{W\} \setminus \{U, V\}$$

5. **Forall** вычислить расстояние $R(W, S)$ по формуле Ланса -Уильямса.

Свойство иерархической кластеризации

Монотонность: функция расстояния обладает монотонностью, если при каждом слиянии расстояние между обьединяемыми кластерами только увеличивается $R_2 \leq R_2 \leq .. \leq R_l$
Редуктивность - некоторое гемоетрическое свойство обьетков.

Теорема

Кластеризация монотонна, если выплонены условия

$$\alpha_U \geq 0,\ \alpha_v \geq 0,\ \alpha_U + \alpha_V + \beta \geq 1,\ \min\{\alpha_u, \alpha_V\} + \gamma \geq 0.$$

Быстрый редуктивный алгоритм кластеризации

Самая трудоемкая операция в алгоритме - поиск ближайших кластеров $O(l^2)$ операций, всего $O(l^3)$.

$$(U, V) = \arg \min_{U \neq V} R(U, V).$$

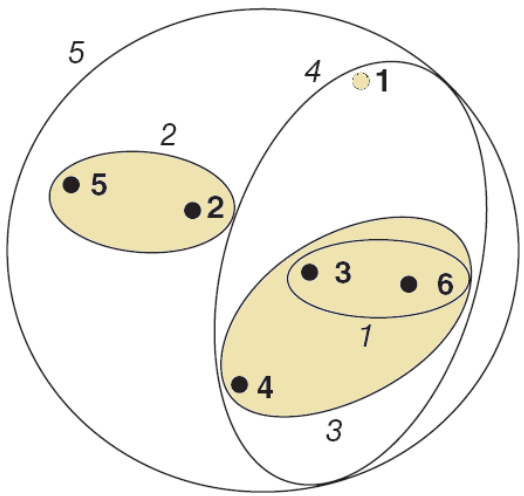
для повышения эффективности будем перебирать лишь наиболее близкие пары:

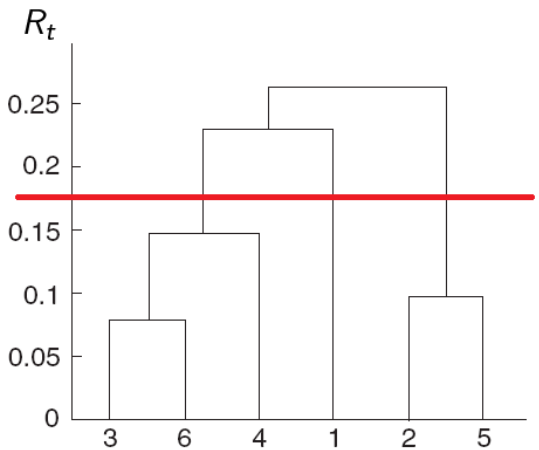
$$(u, v) = \arg \min_{r(u,v) \leq \delta} r(u, v),$$

параметр δ периодически увеличивается.
один раз строим множество $\{r(u, v) \leq \delta\}$ за $o(l^2)$ операций, потом его используем

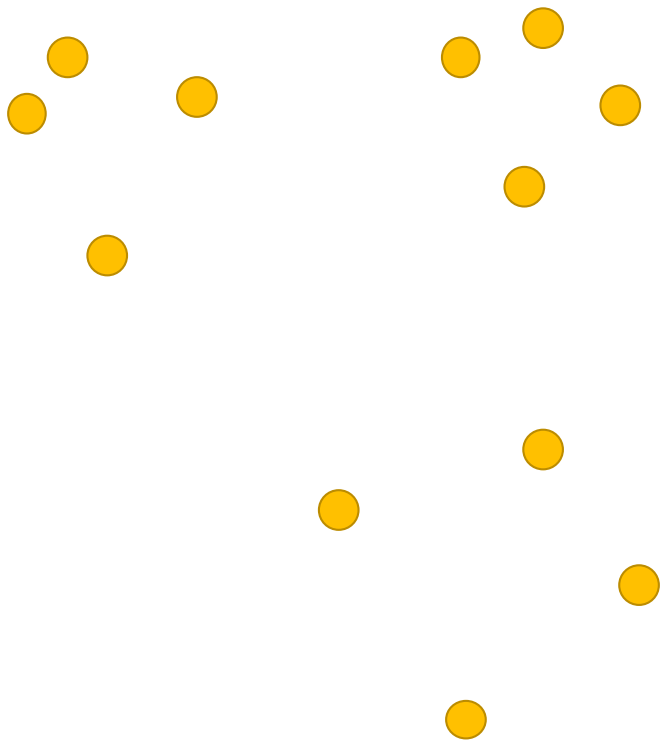
Замечания к иерархическому методу

- рекомендуется пользоваться расстоянием Уорда R^u;
- по практике строят несколько разбиений и выбирают лучшую визуально по дендрограмме;
- определение числа кластеров - по максимуму $|R_{t+1} - R_t|$, тогда за наилучшее множество кластеров можно взять C_t.

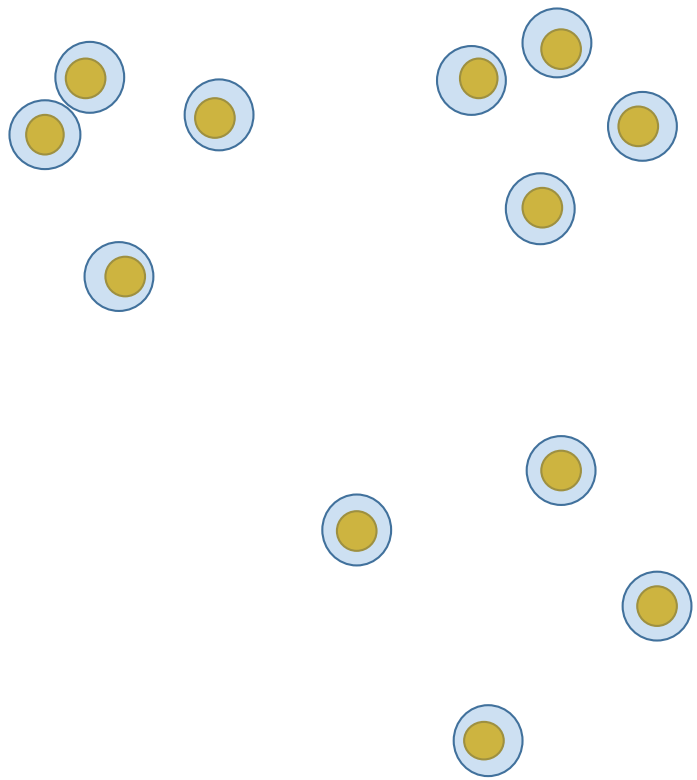




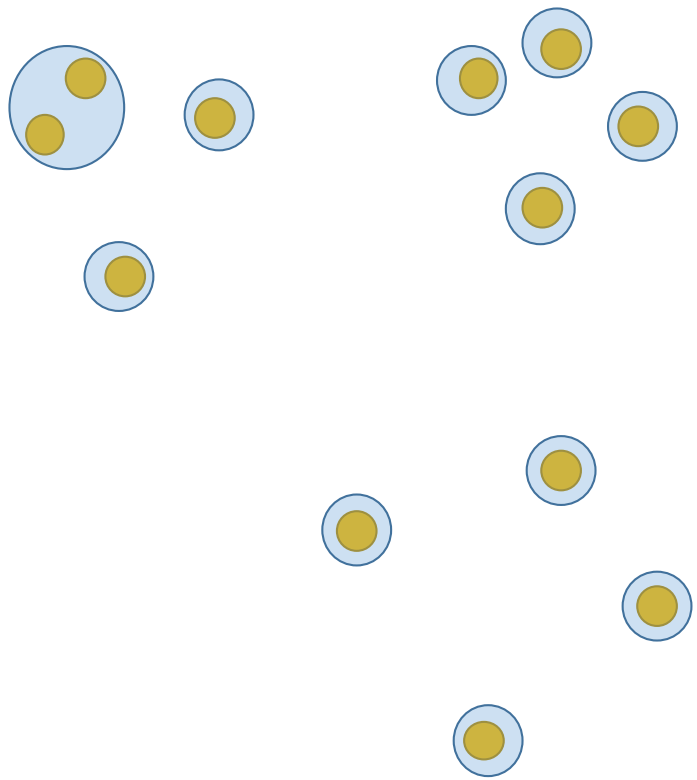
Агломеративная кластеризация



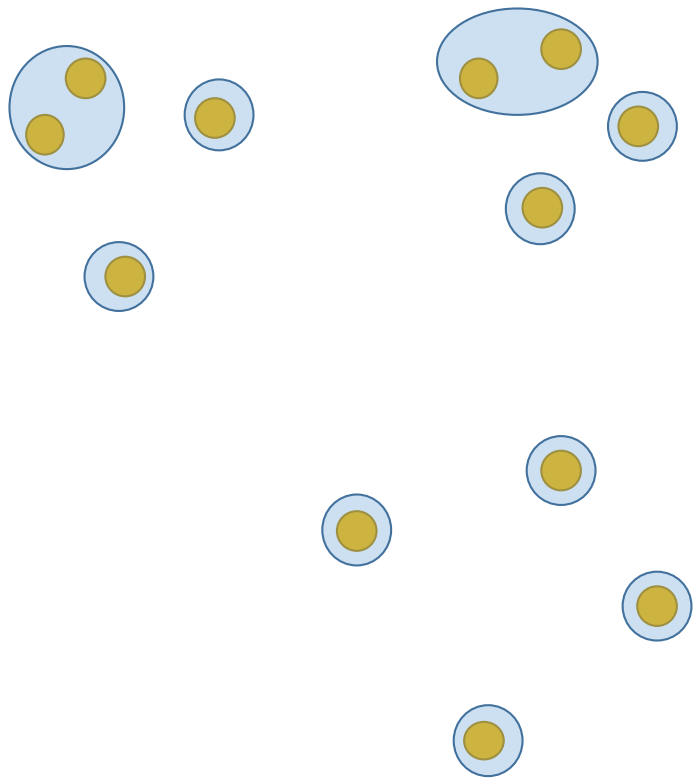
Агломеративная кластеризация



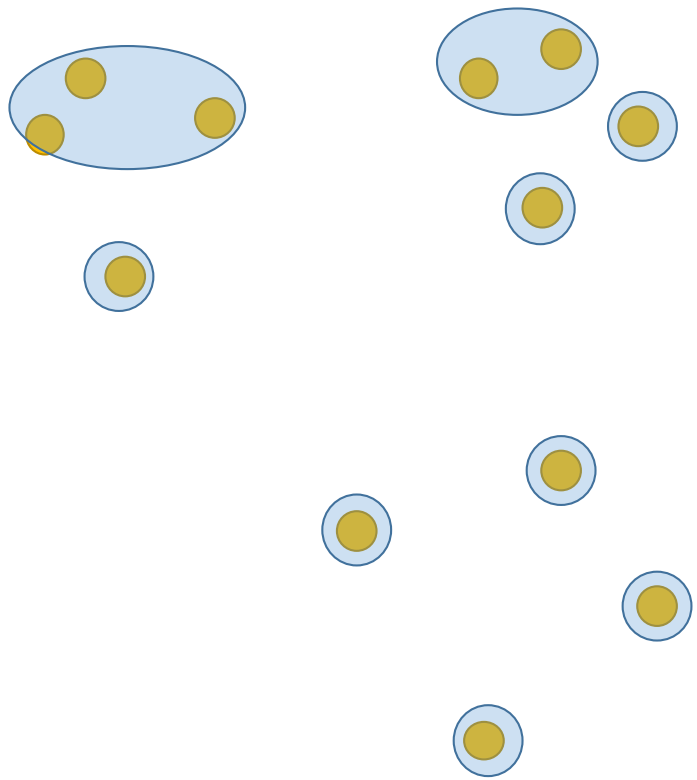
Агломеративная кластеризация



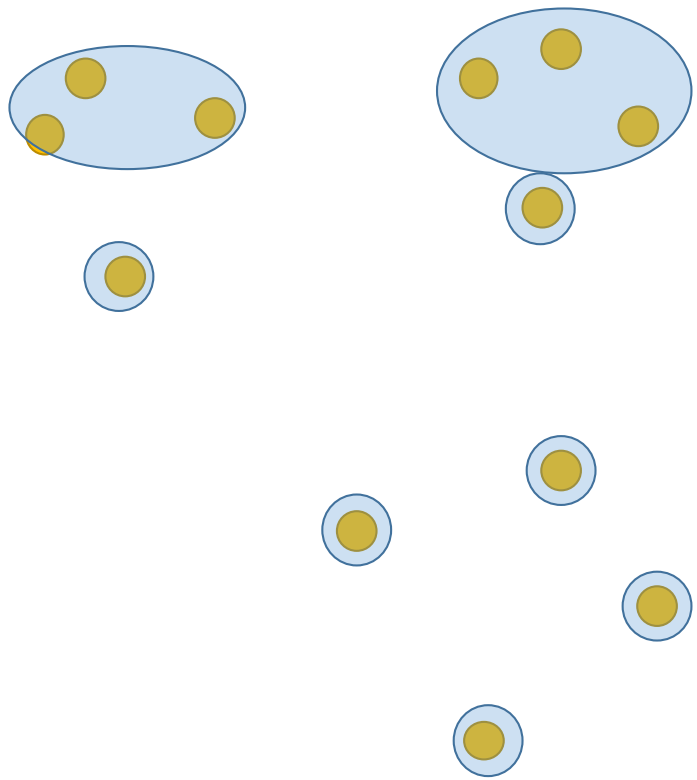
Агломеративная кластеризация



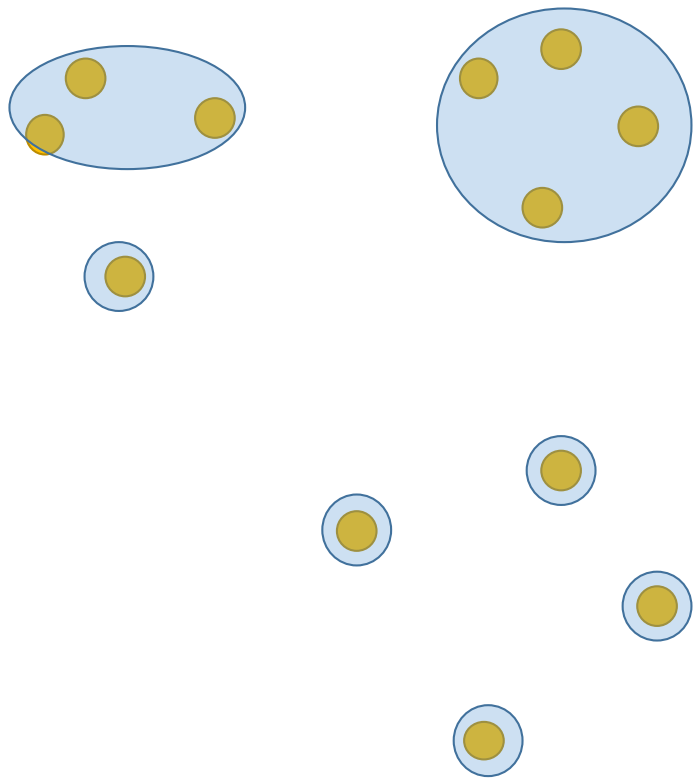
Агломеративная кластеризация



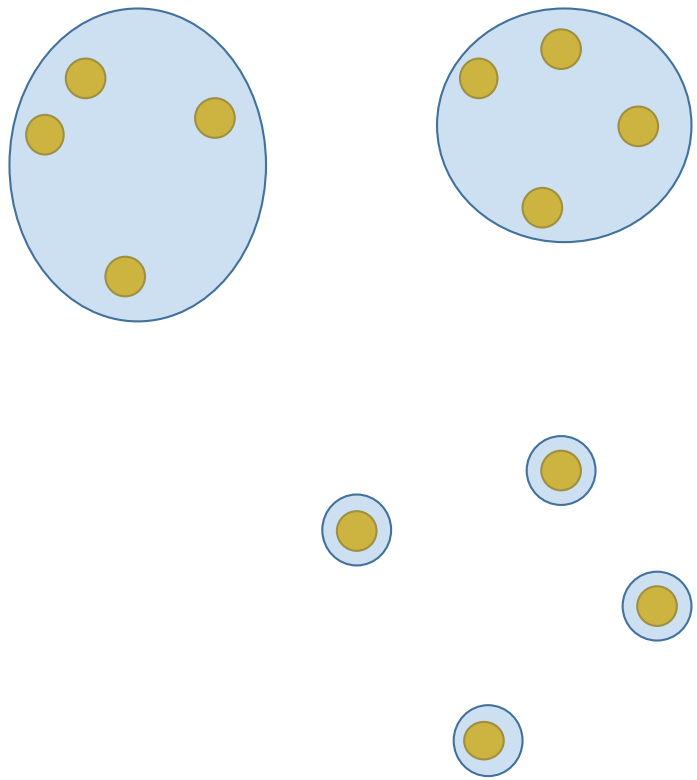
Агломеративная кластеризация



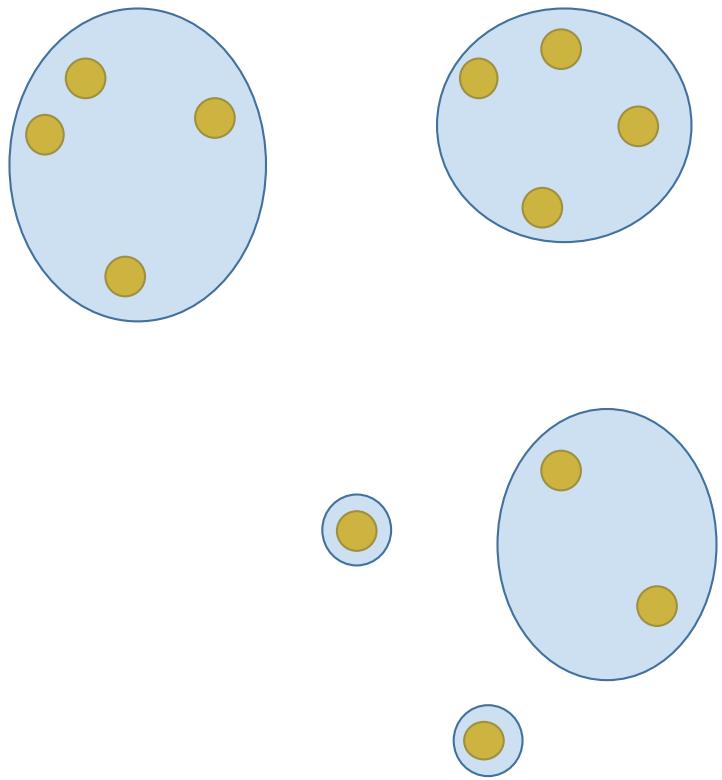
Агломеративная кластеризация



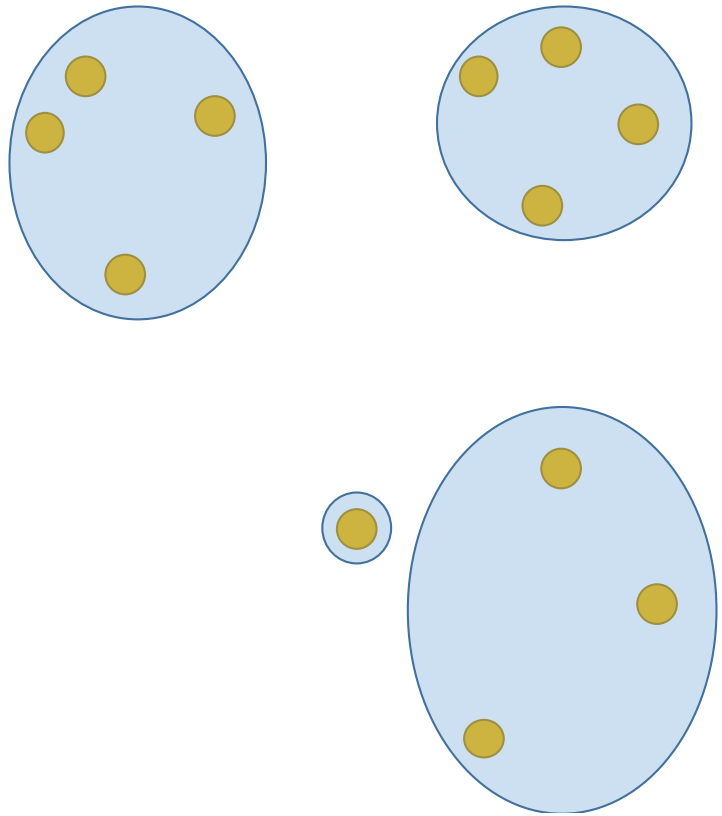
Агломеративная кластеризация



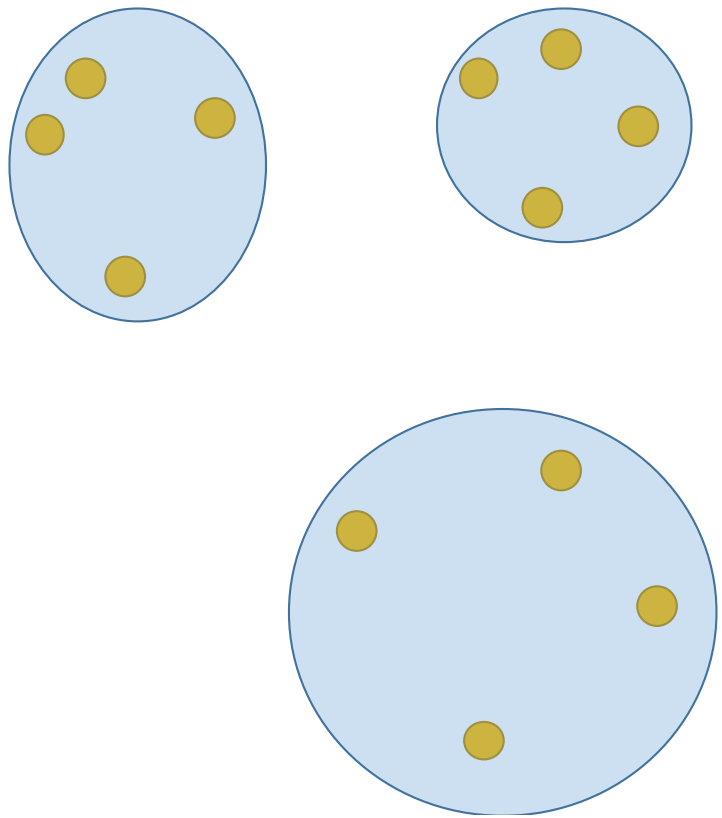
Агломеративная кластеризация



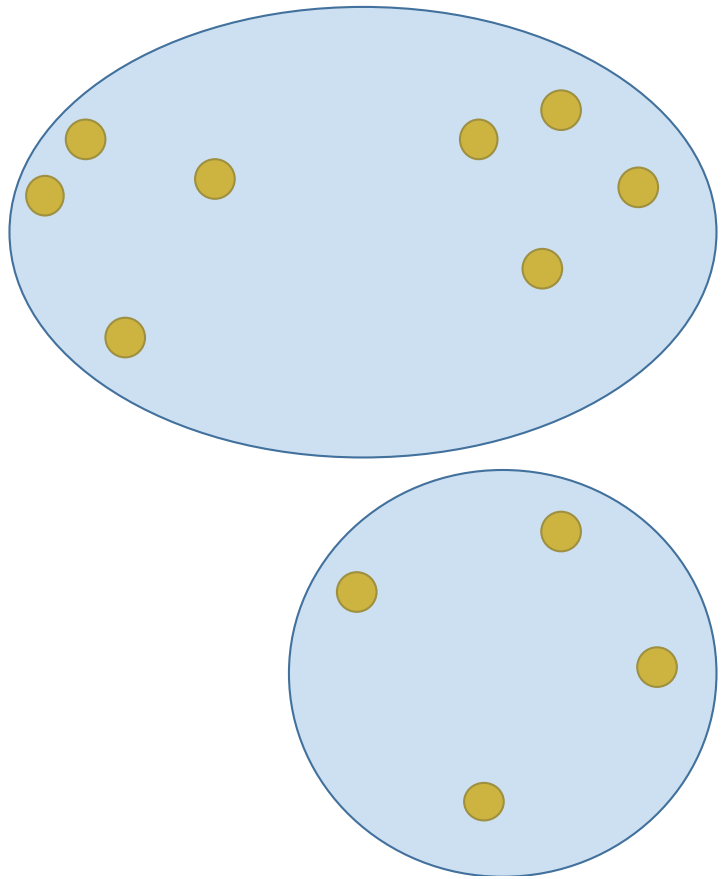
Агломеративная кластеризация



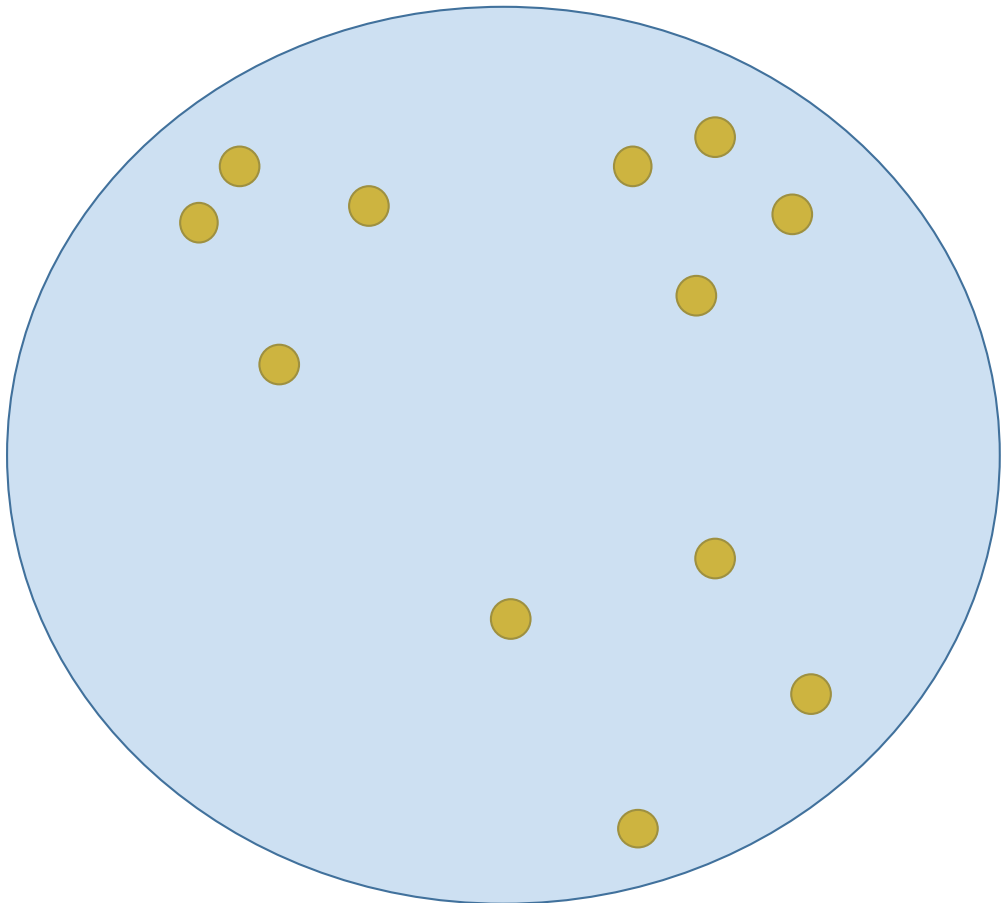
Агломеративная кластеризация



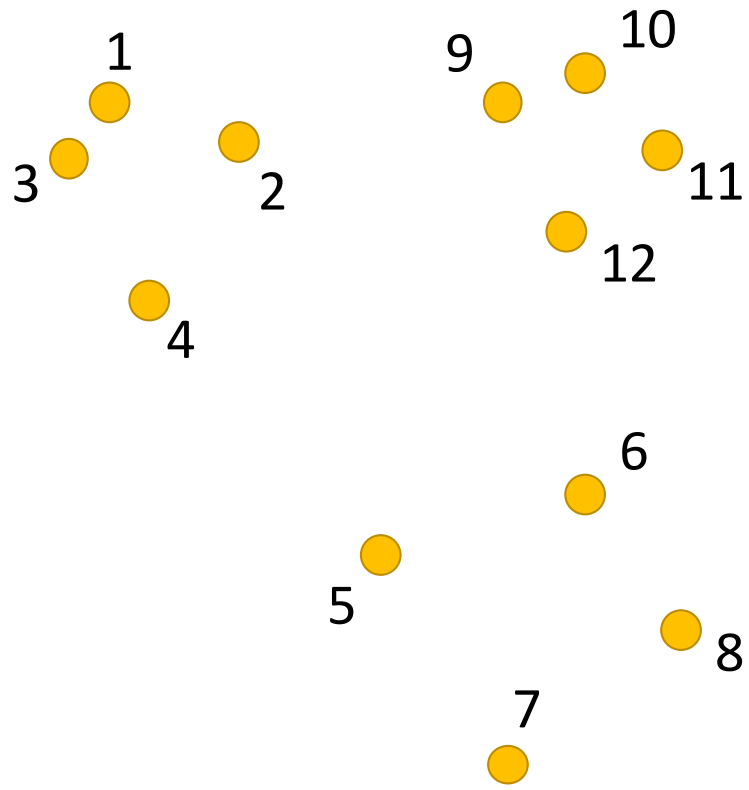
Агломеративная кластеризация



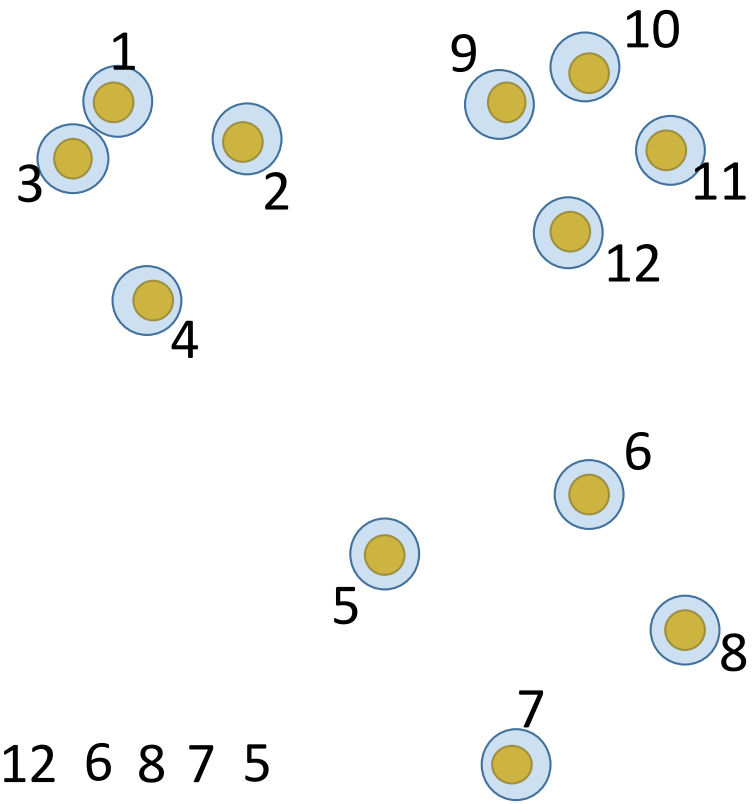
Агломеративная кластеризация



Дендрограмма

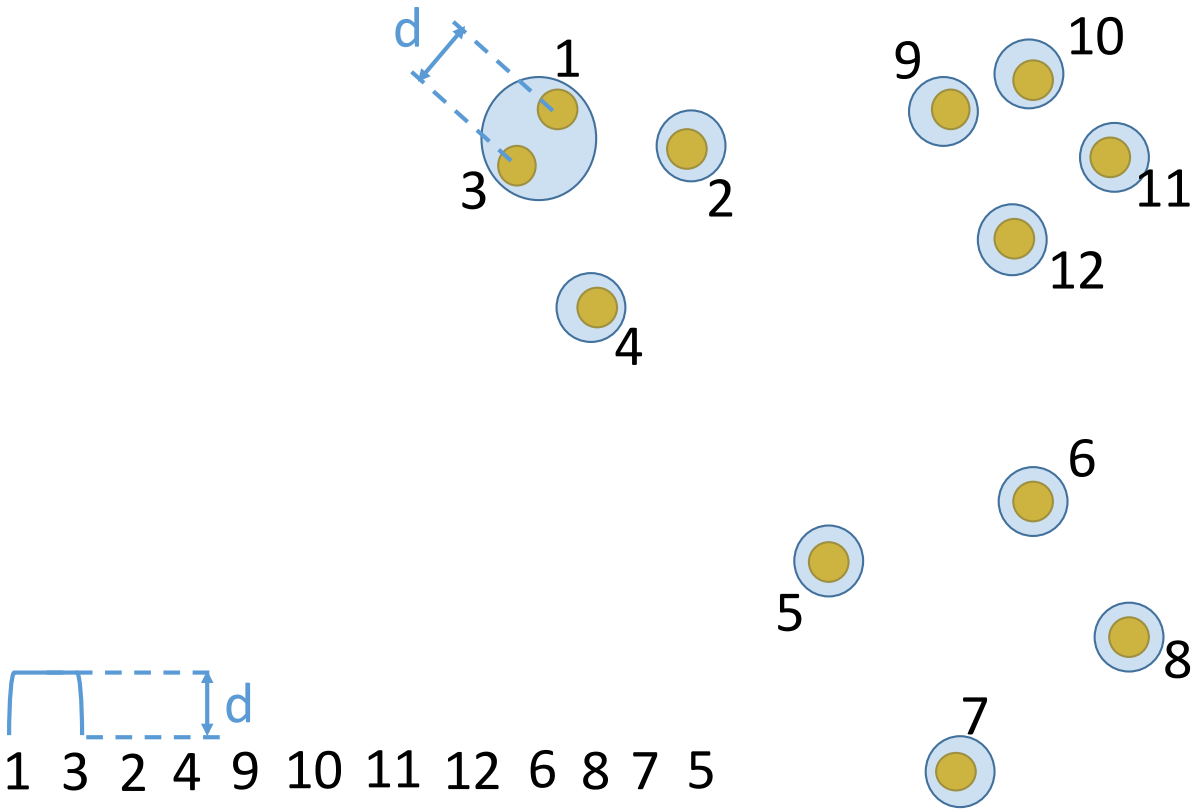


Дендрограмма

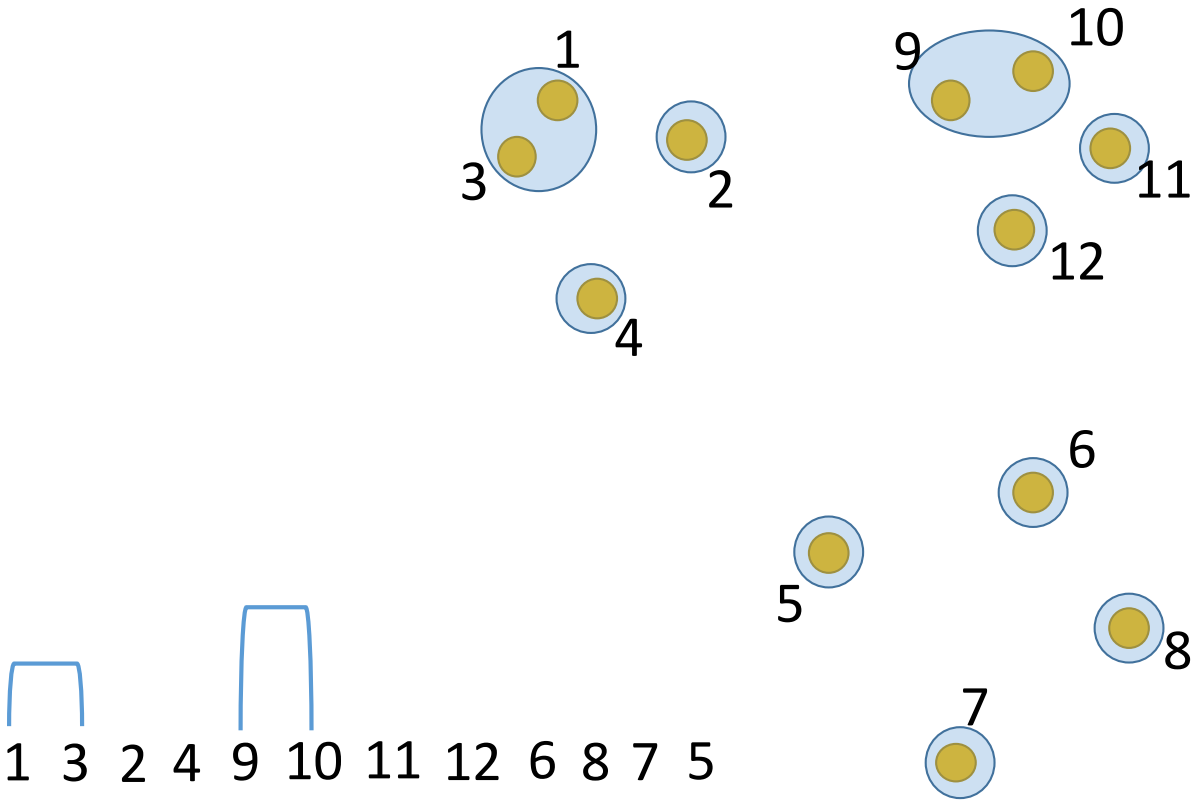


1 3 2 4 9 10 11 12 6 8 7 5

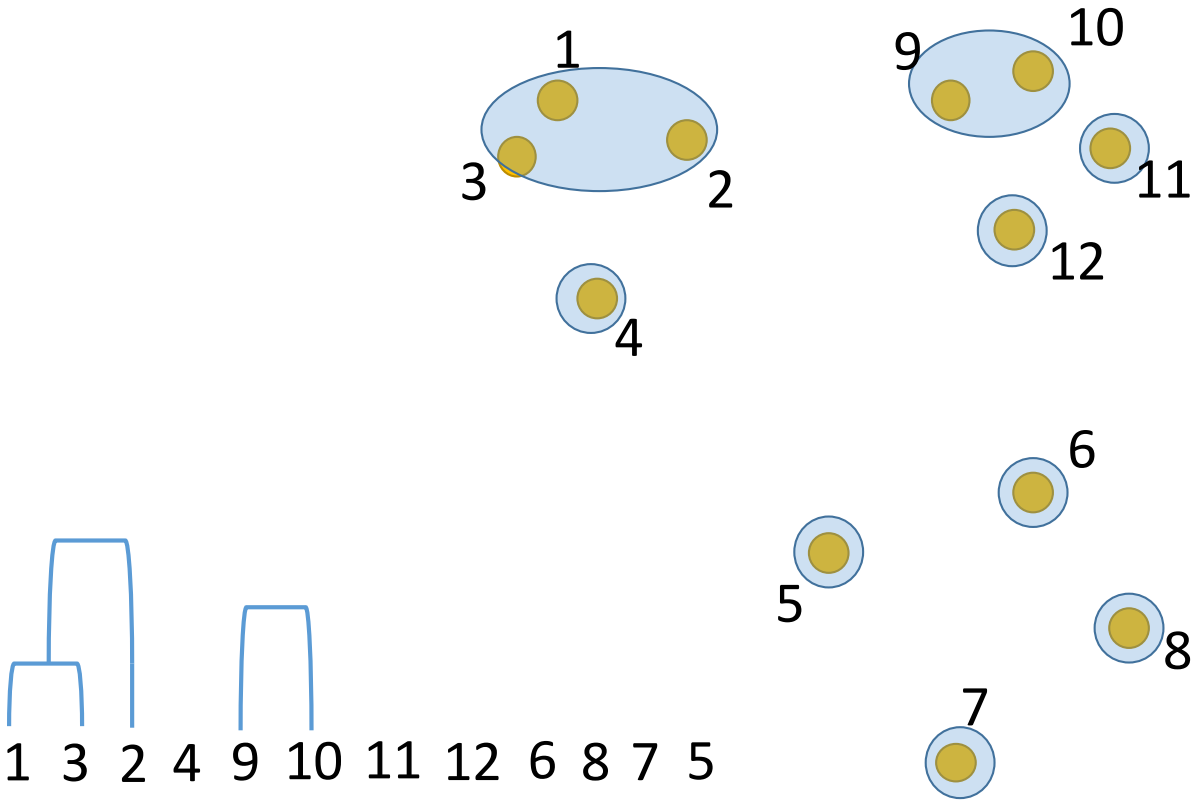
Дендрограмма



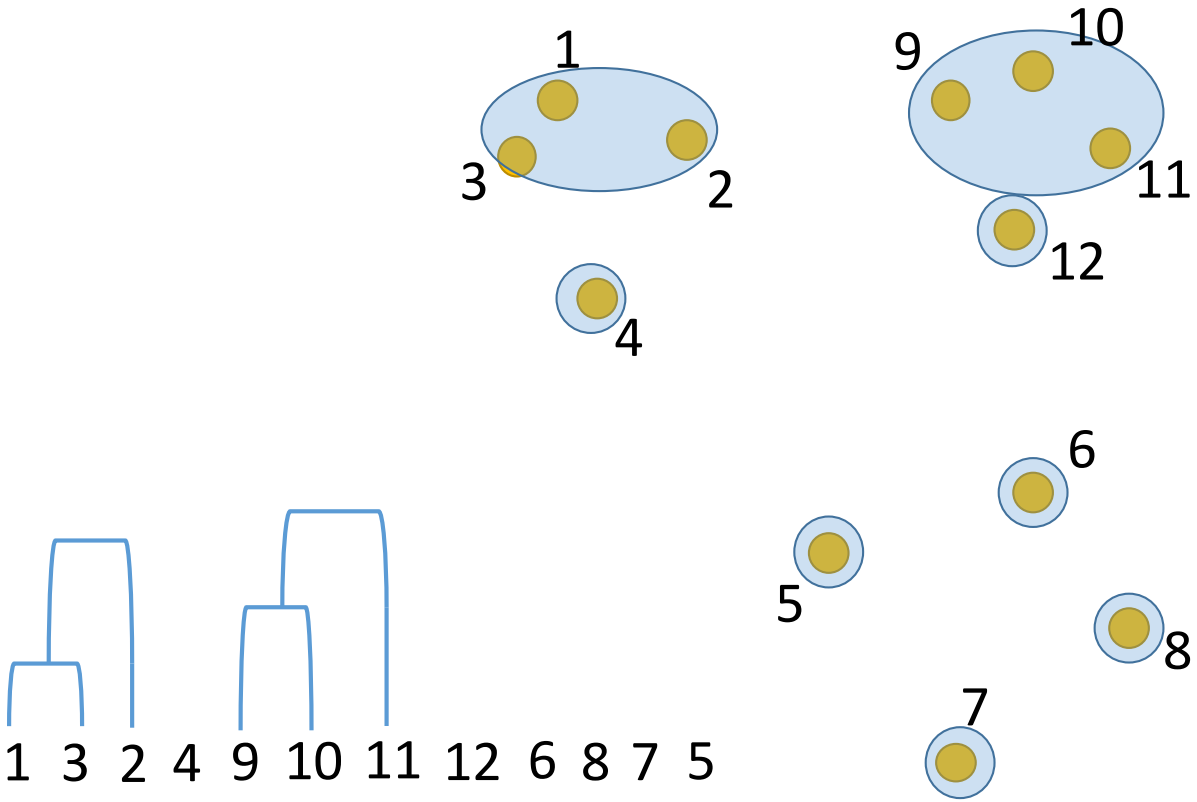
Дендрограмма



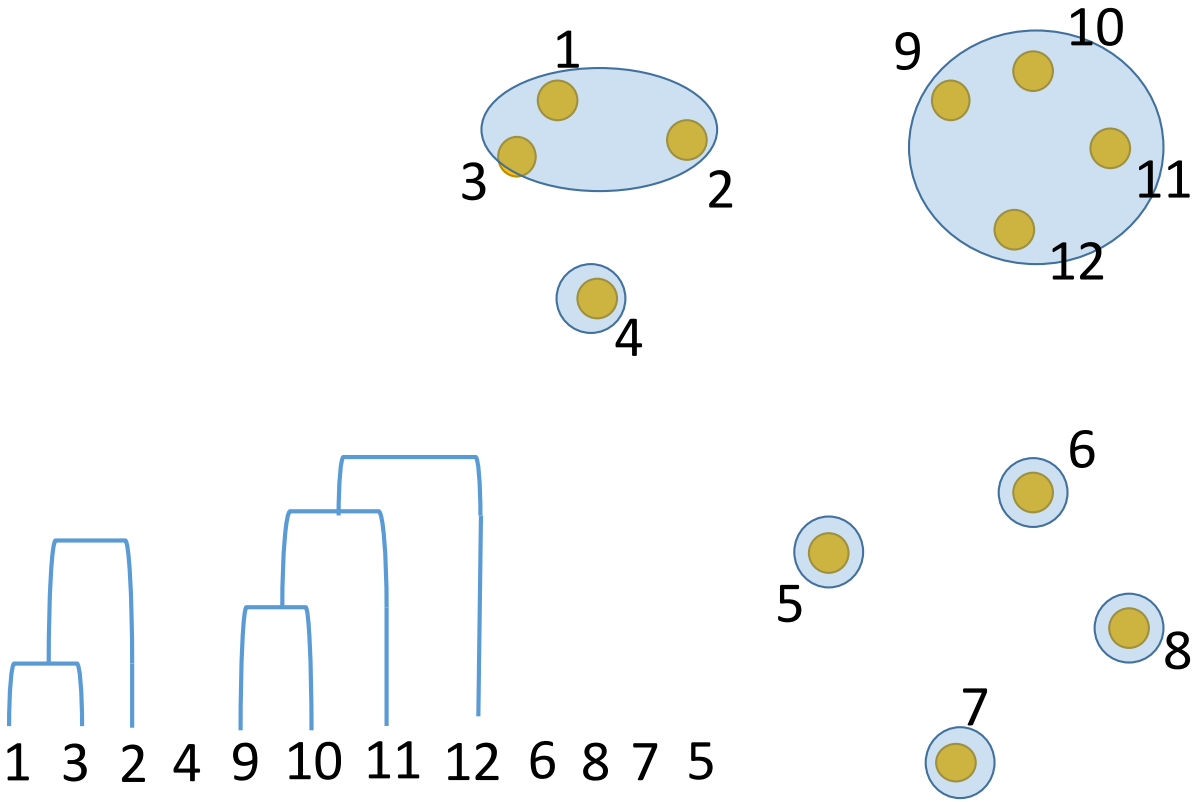
Дендрограмма



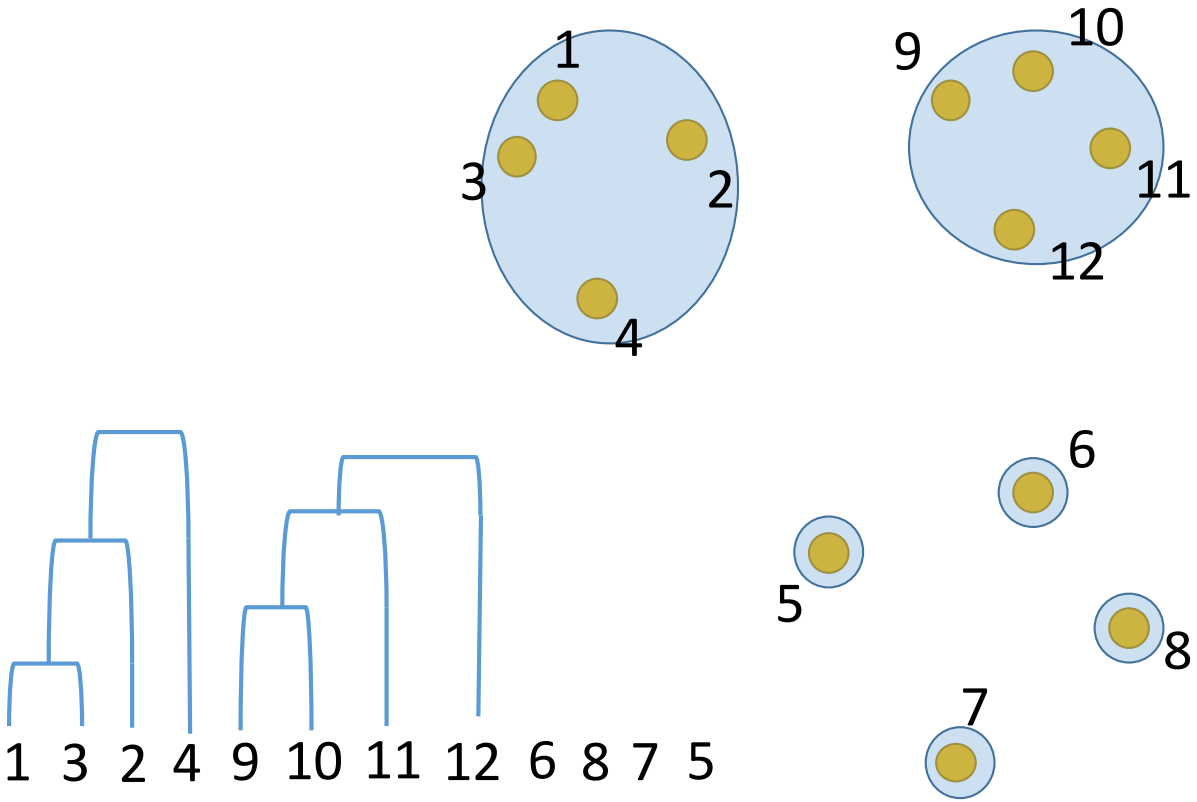
Дендрограмма



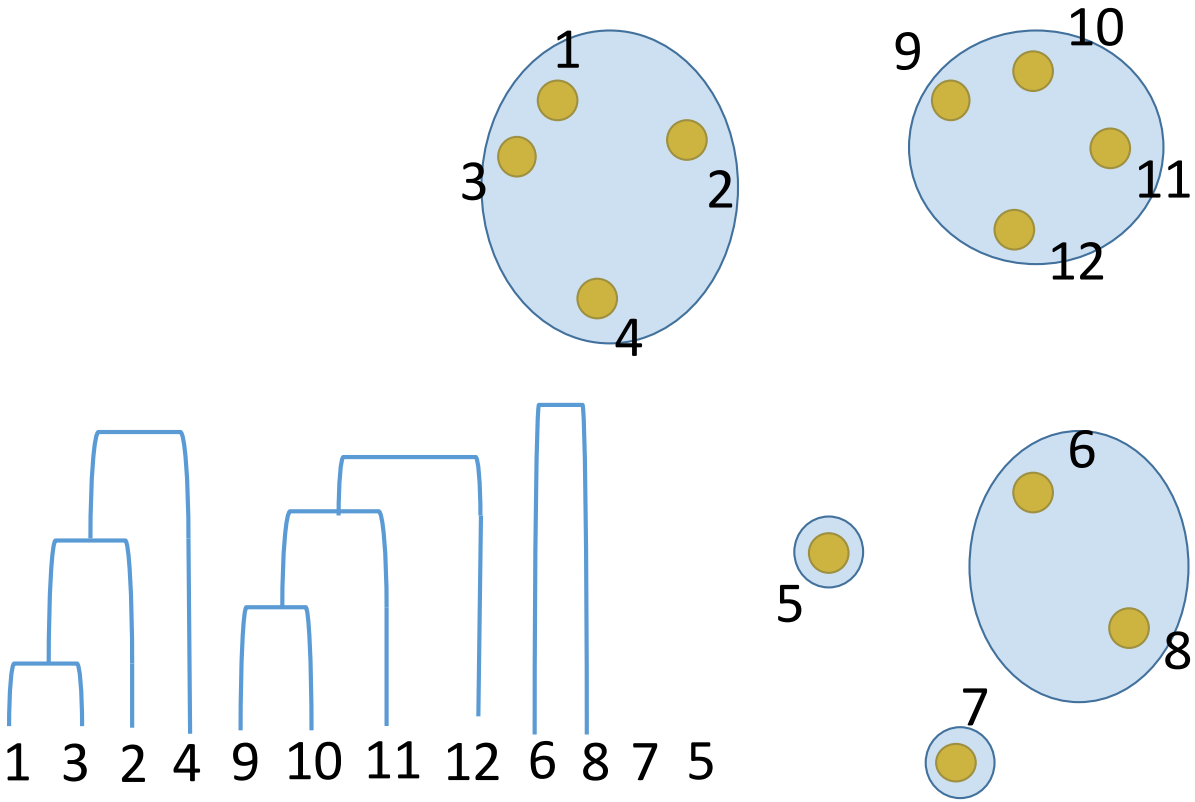
Дендрограмма



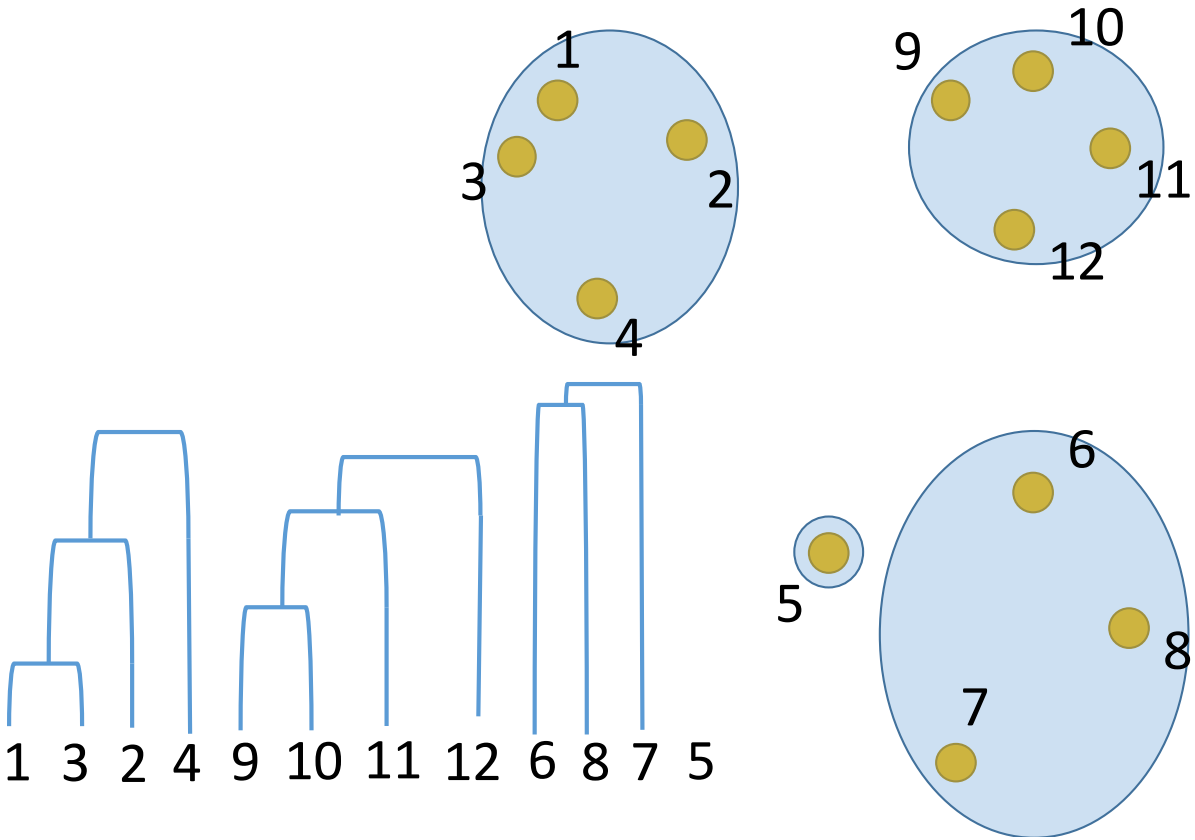
Дендрограмма



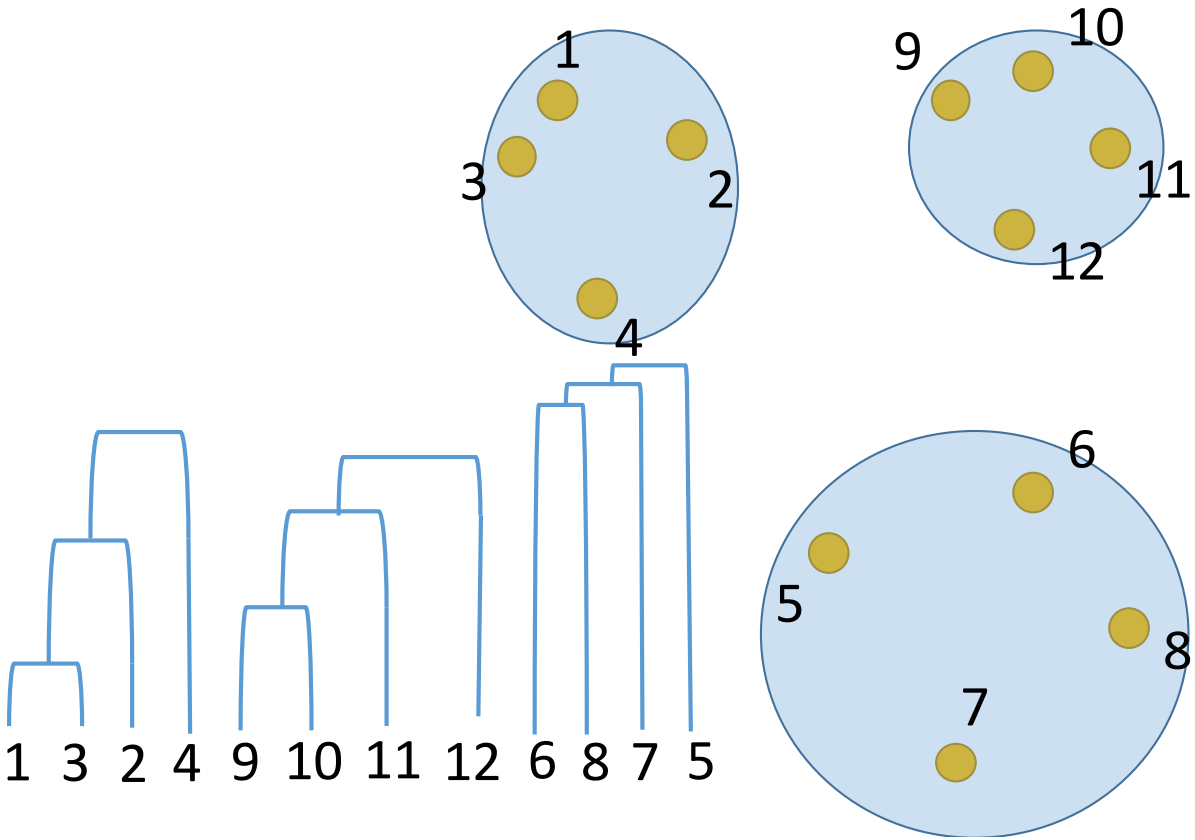
Дендрограмма



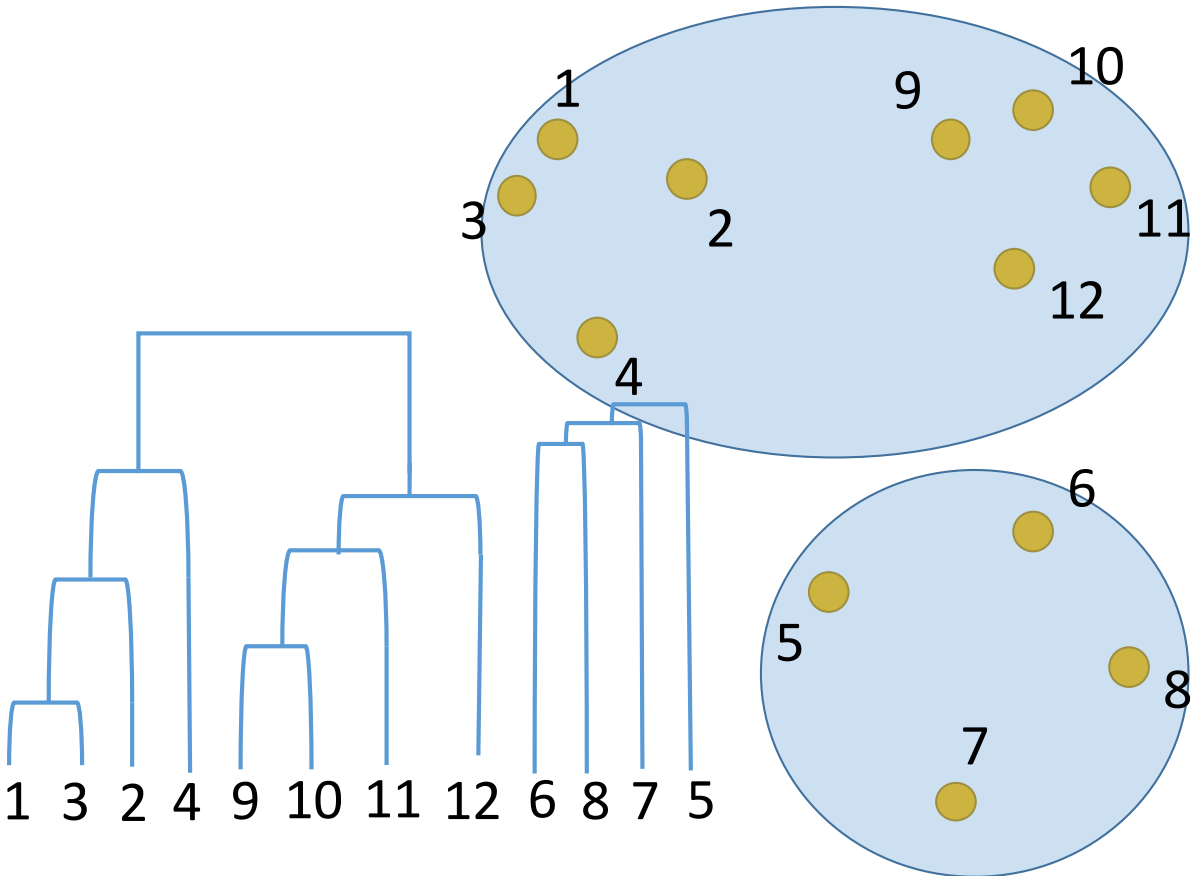
Дендрограмма



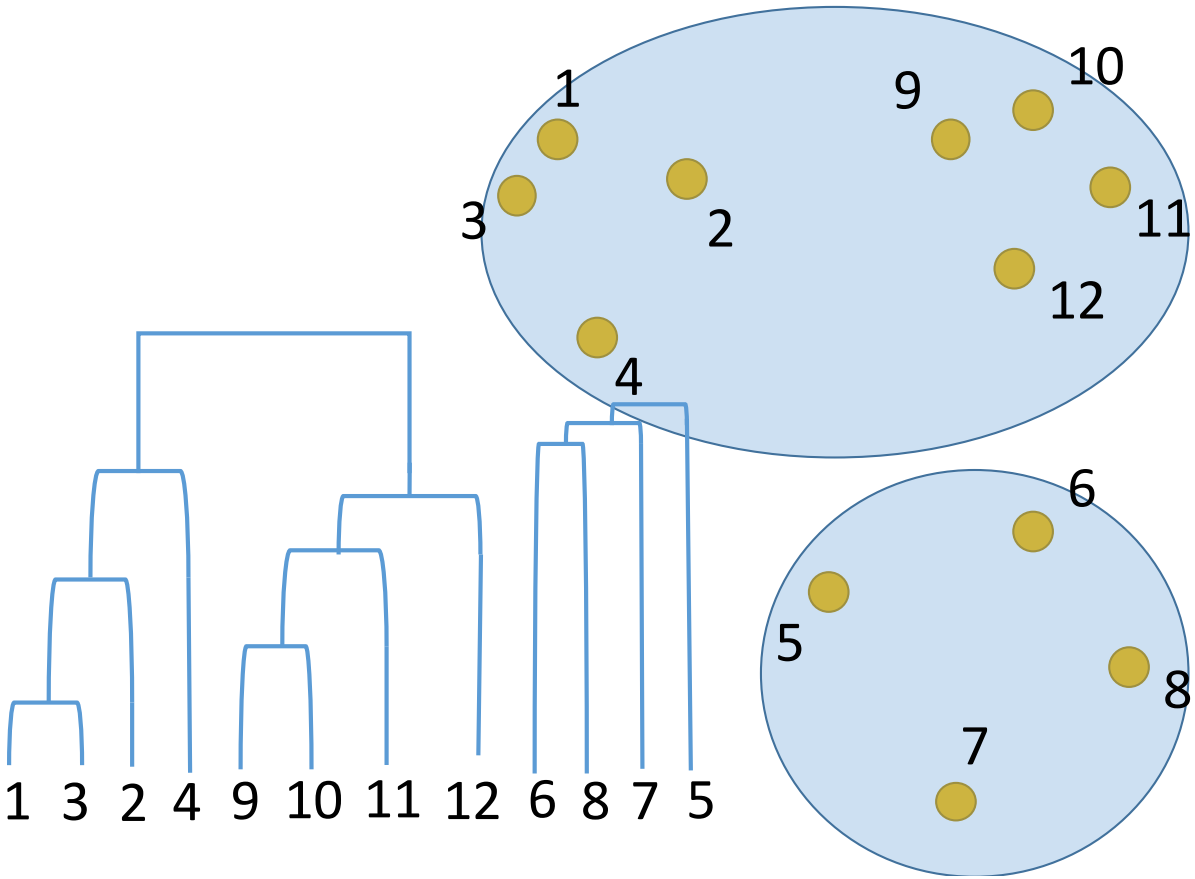
Дендрограмма



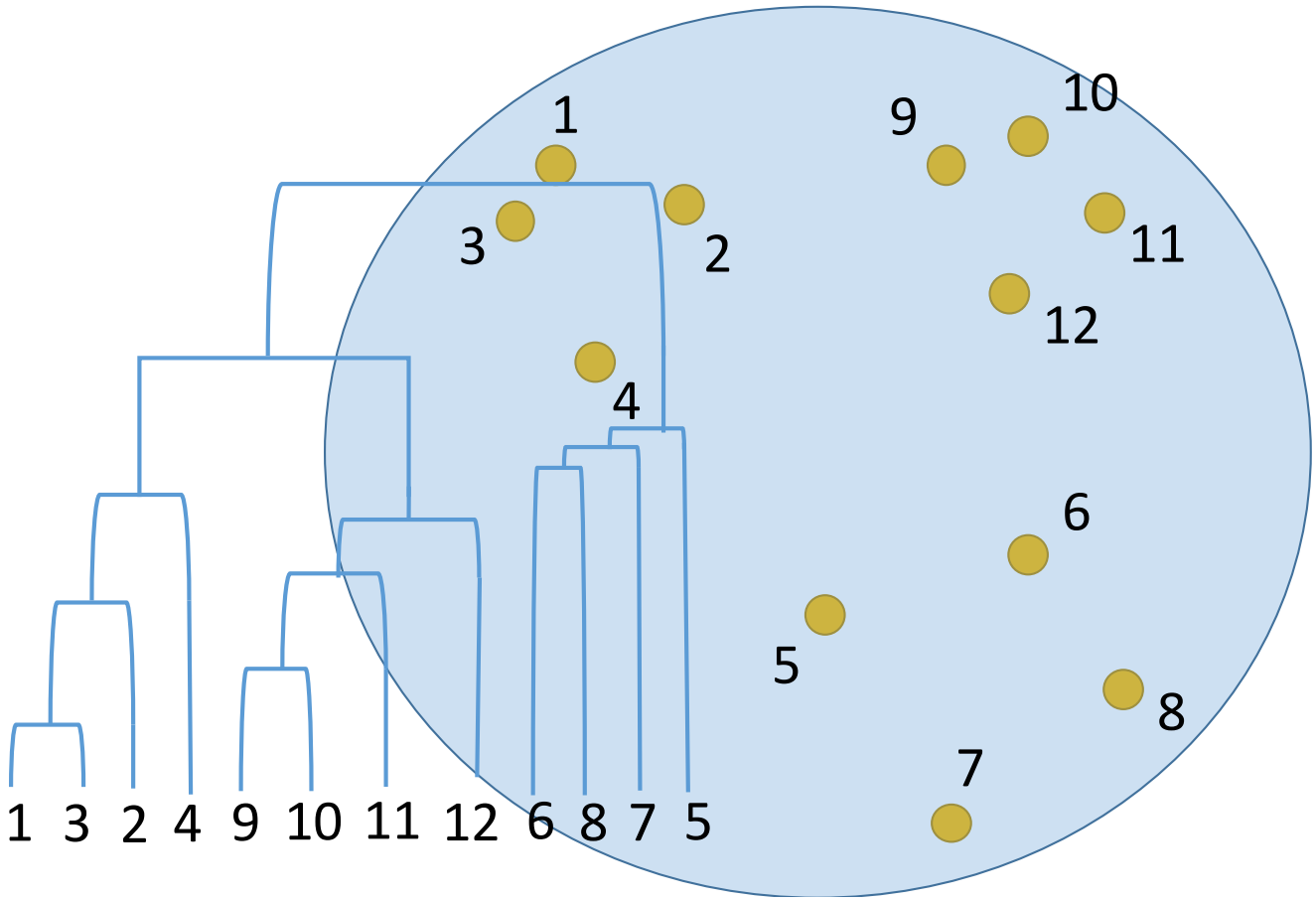
Дендрограмма



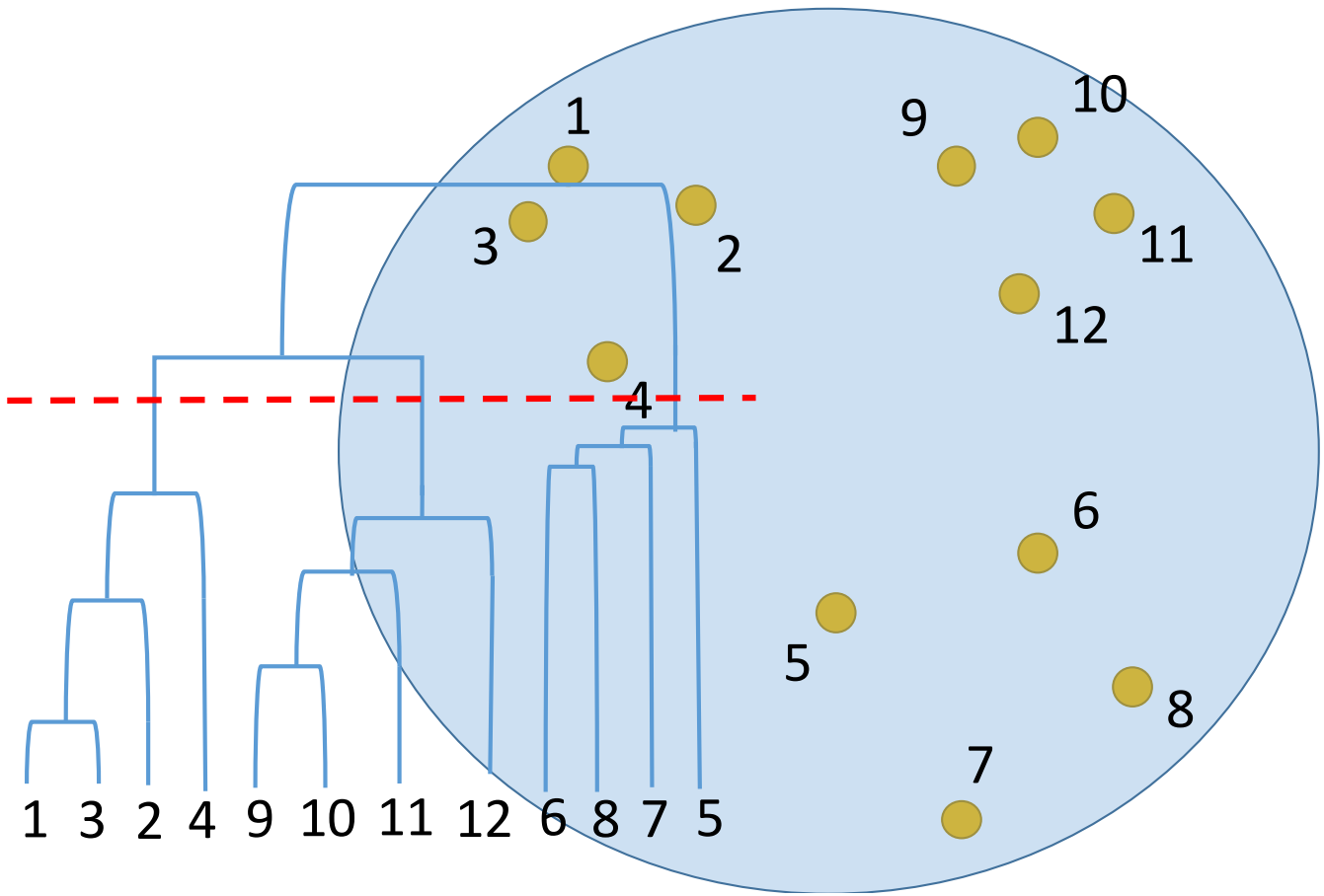
Дендрограмма



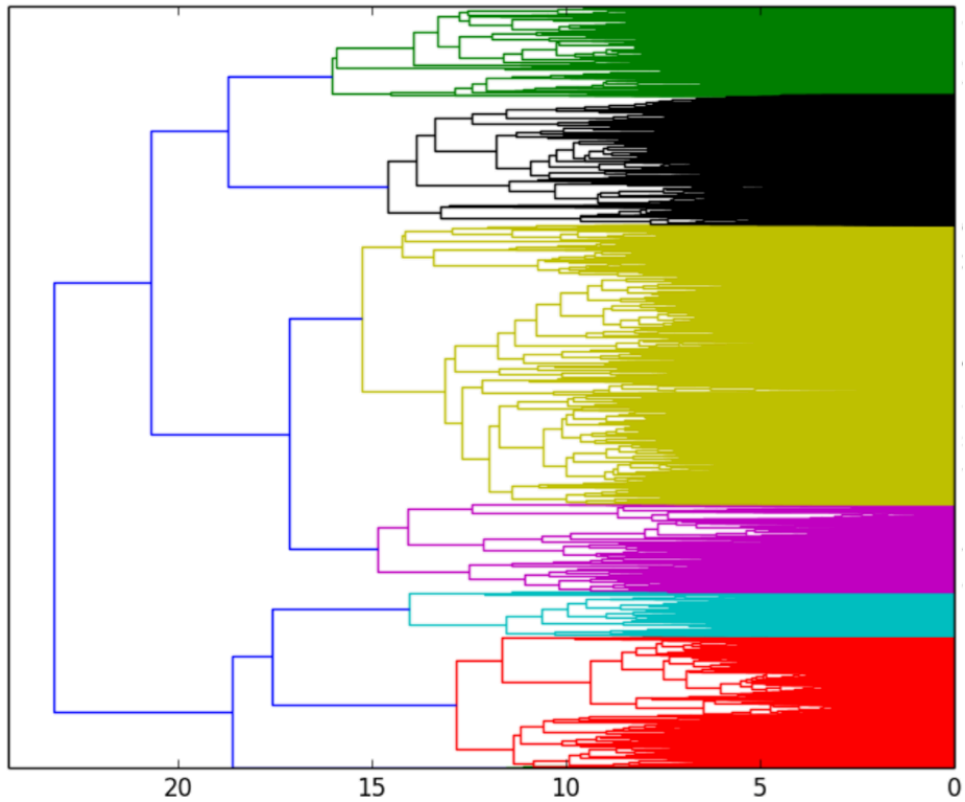
Дендрограмма



Дендрограмма

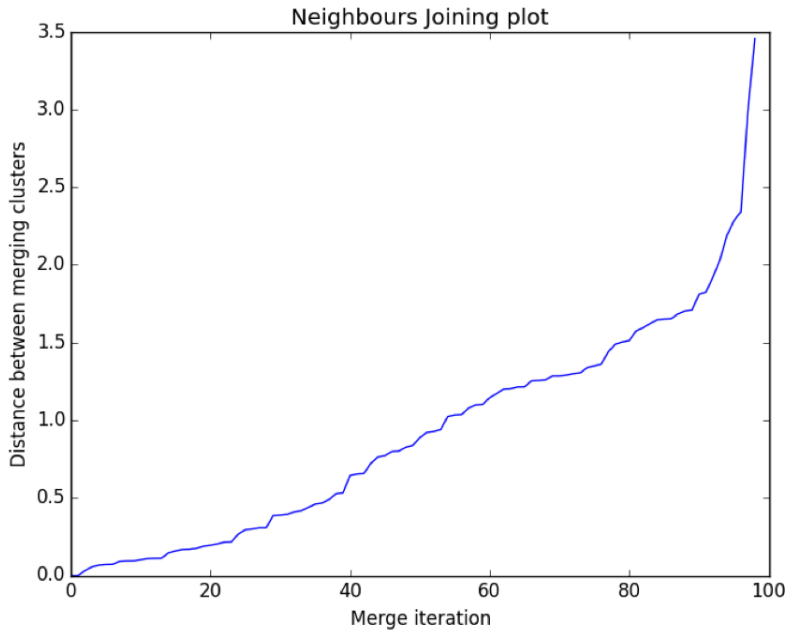


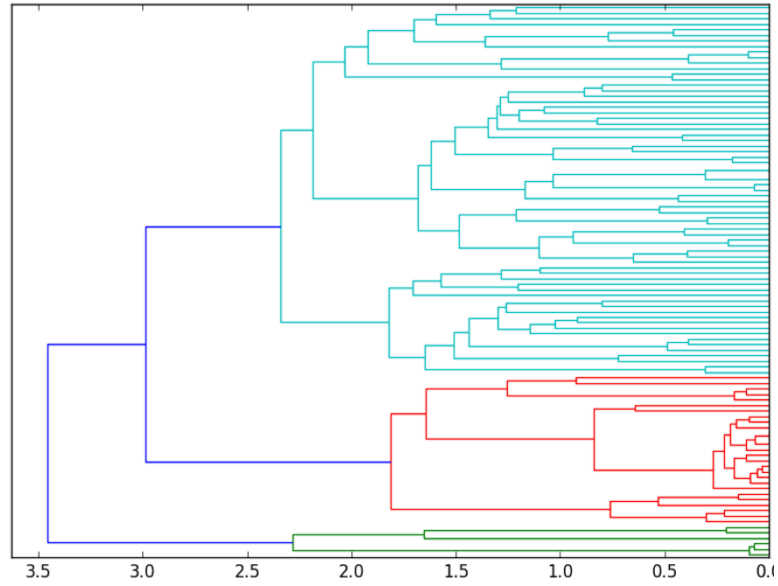
Пример: кластеризация писем



Пример: расстояние между кластерами

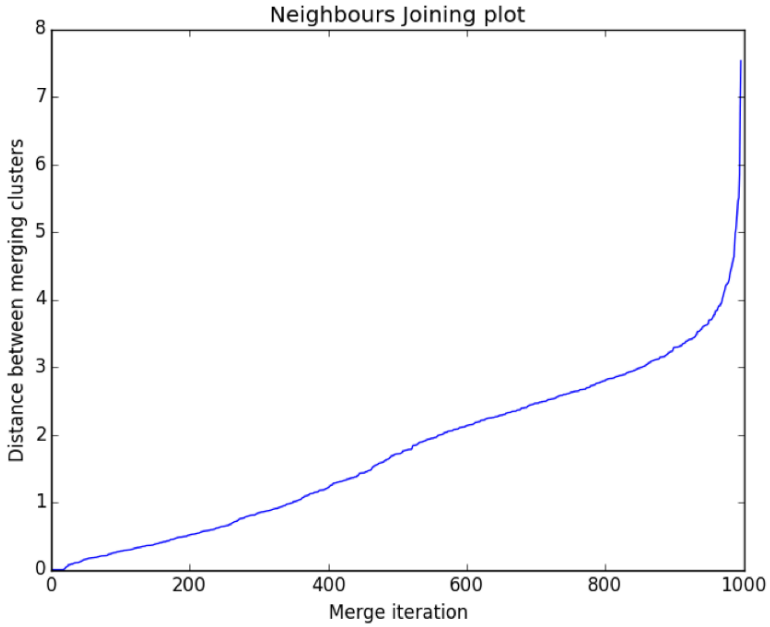
- На подвыборке из 100 писем

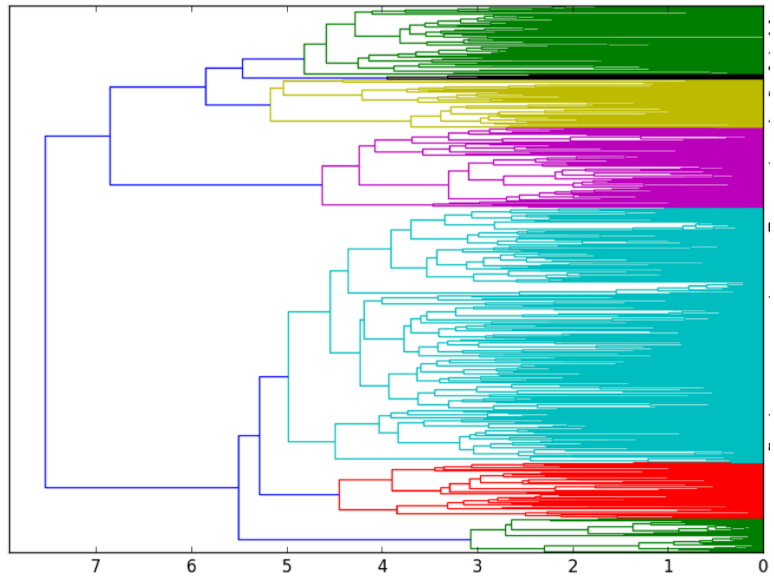




Пример: расстояние между кластерами

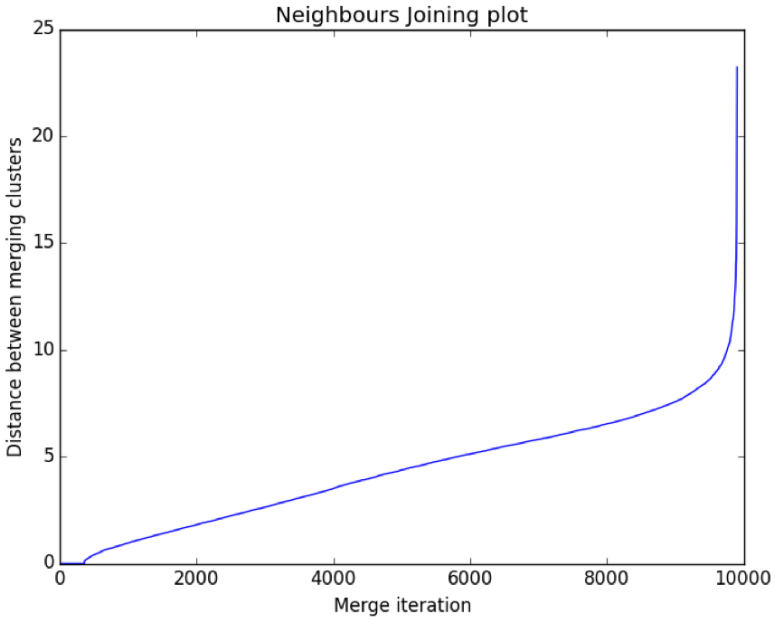
- На подвыборке из 1000 писем

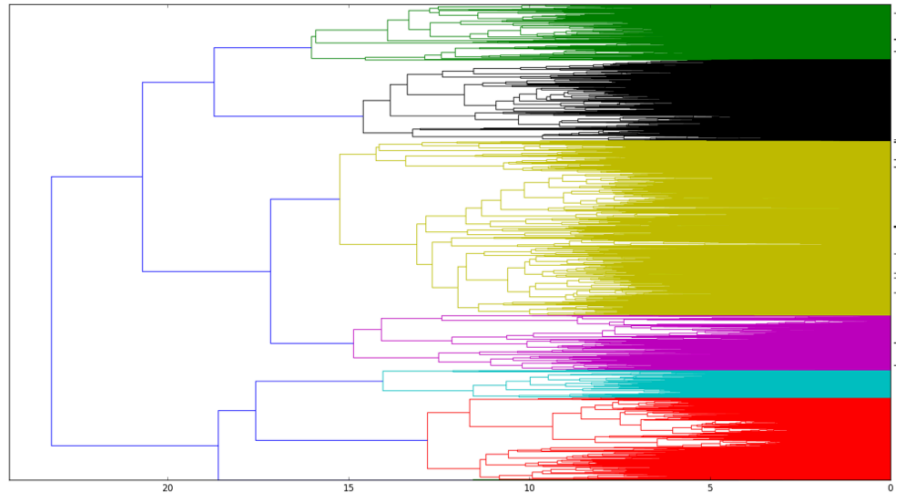




Пример: расстояние между кластерами

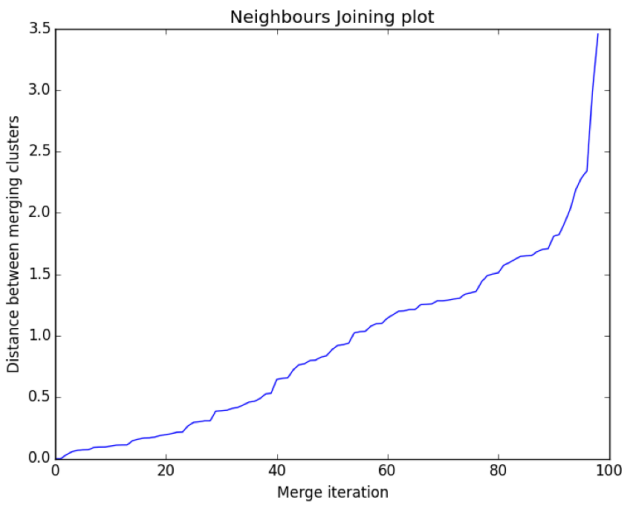
- На подвыборке из 10000 писем

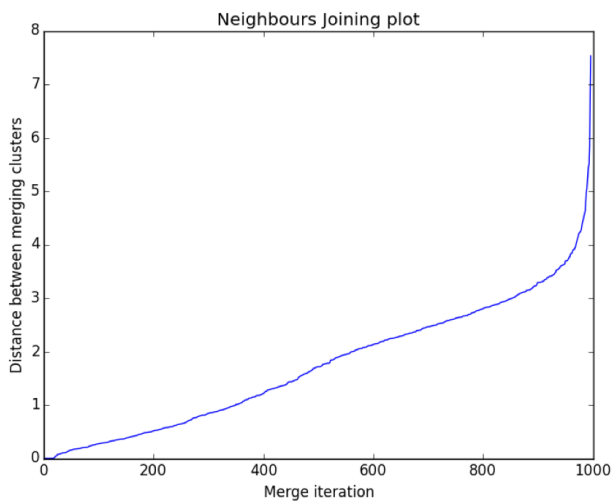


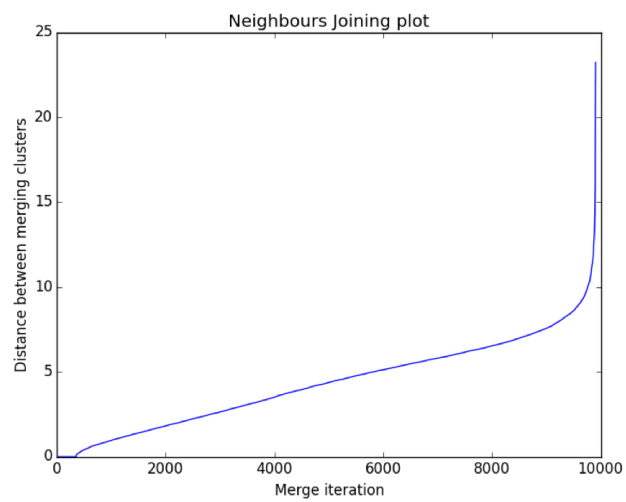


Пример: расстояние между кластерами

- Сравним графики: 100, 1000, 10000 писем

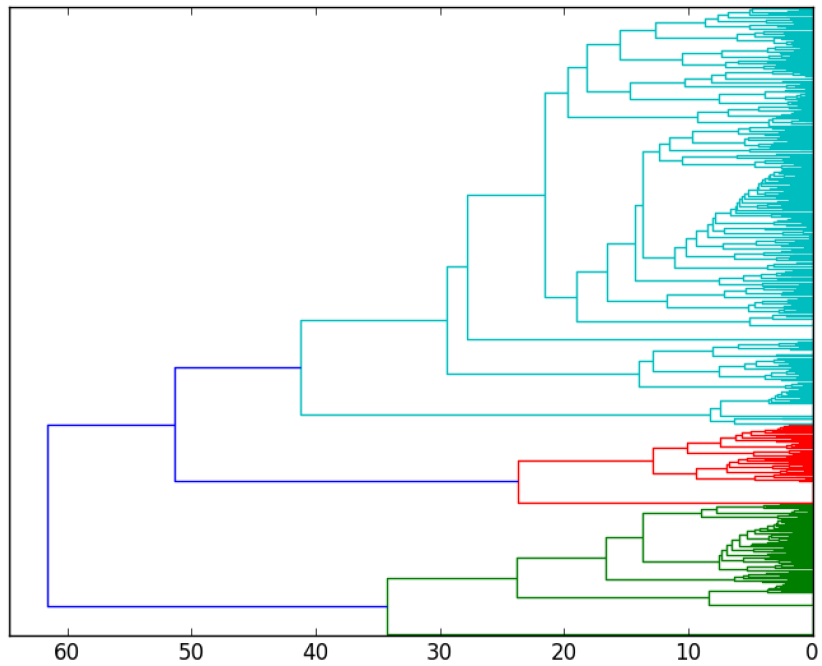




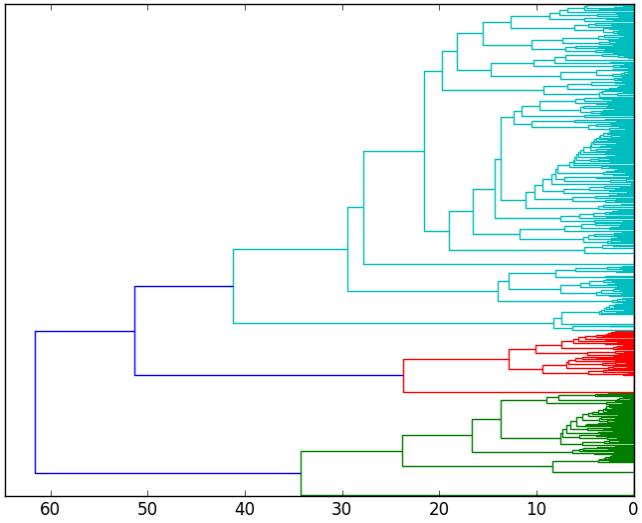


Пример: перекос в размерах кластеров

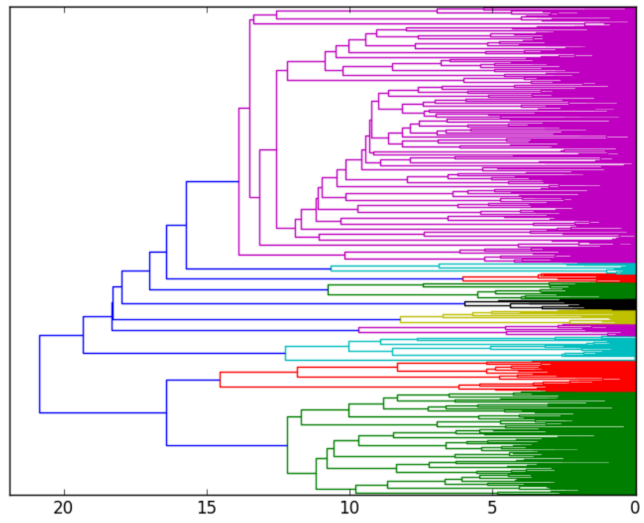
- Дендрограмма, построенная для другой выборки текстов:



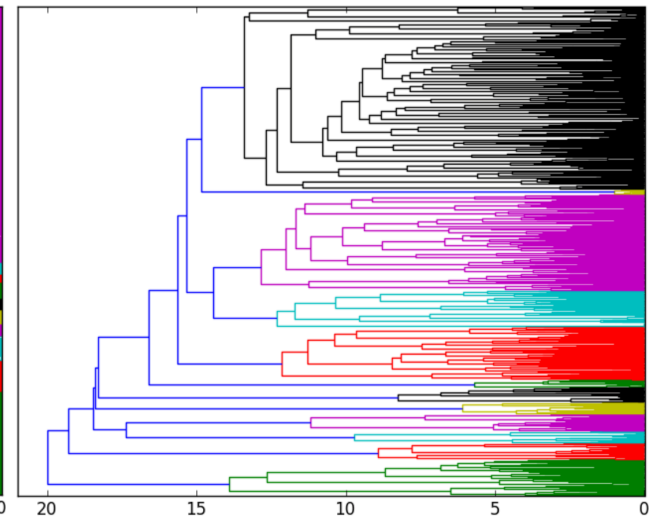
Пример: добавляем SVD



Исходные признаки



SVD



SVD (еще меньше компонент)

Пример: SVD и расстояние при слиянии