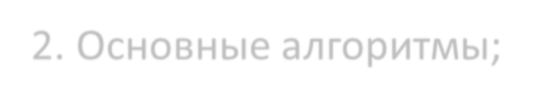
*Кластеризация*

***Вопросы занятия***

1. Задача кластеризации: постановка и примеры;



2. Основные алгоритмы;

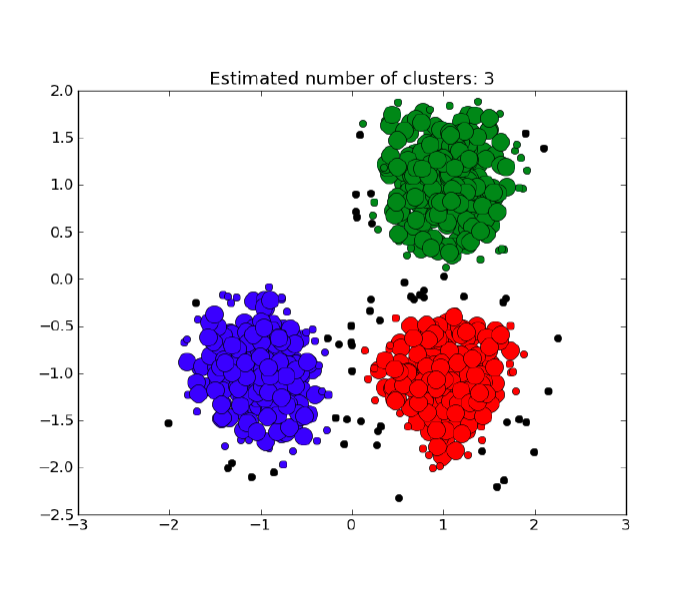
3. Метрики качества кластеризации.

***В конце занятия научимся:***

- производитькластеризациюданных;

- выбирать наиболее подходящий алгоритм для задачи.

*ЗАДАЧА КЛАСТЕРИЗАЦИИ*

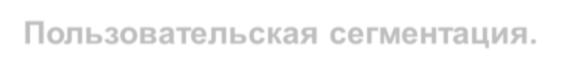
**ТИПЫ ЗАДАЧ**

- классификация - ранжирование - регрессия

**- кластеризация**

*ЗАДАЧА КЛАСТЕРИЗАЦИИ*

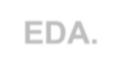
ПРИМЕРЫ ЗАДАЧ КЛАСТЕРИЗАЦИИ



**Пользовательская сегментация.** Как выглядят типичные пользователи? (находим сектора, работаем с ними отдельно)

**Логистика.** Где расположить магазины, чтобы охватить большее кол-во покупателей?

**Новости.** О чём сейчас пишут СМИ? (Новостной портал кластеризует новости и выдает их отдельными темами)

**EDA.** Есть 100млн обращений пользователей. О чём они пишут?

*ЗАДАЧА КЛАСТЕРИЗАЦИИ*

ДОПОЛНИТЕЛЬНЫЕ ПРИЛОЖЕНИЯ

**Создание дополнительных фич.** Можно дополнить имеющиеся данные метками принадлежности к одному из классов;

**Разметка данных.** Если нет проставленных классов, но нужно сделать классификатор, то в создании разметки для обучающей выборки сильно поможет кластеризация;

**Поиск структуры данных** как часть эксплоративного анализа.

*ЗАДАЧА КЛАСТЕРИЗАЦИИ* Постановка задачи кластеризации

Пусть *X* — множество объектов, *Y* — множество идентификаторов (меток) кластеров. На множестве *X* задана функция расстояния между объектами *ρ(x,x′)*.

Дана конечная обучающая выборка объектов *X\*\*m={x1,…,xm}*⊂*X*. Необходимо разбить выборку на подмножества (кластеры), то есть каждому объекту *xi*∈*X\*\*m* сопоставить метку *y*∈*Y*, таким образом чтобы объекты внутри каждого кластера были близки относительно метрики *ρ*, а объекты из разных кластеров значительно различались.

*i*

Определение

**Алгоритм кластеризации** — функция*a:X→Y,*которая любому объекту *x*∈*X* ставит в соответствие идентификатор кластера *y*∈*Y*.

*ЗАДАЧА КЛАСТЕРИЗАЦИИ*



*i*

Множество *Y* в некоторых случаях известно заранее, однако чаще ставится задача определить оптимальное число кластеров, с точки зрения того или иного критерия качества кластеризации. Кластеризация (обучение без учителя) отличается от классификации (обучения с учителем) тем, что метки объектов из обучающей выборки *y* изначально не заданы, и даже может быть неизвестно само множество *Y*.

**Решение задачи кластеризации объективно неоднозначнопо ряду причин:**

• Не существует однозначного критерия качества кластеризации. Известен ряд алгоритмов, осуществляющих разумную кластеризацию "по построению", однако все они могут давать разные результаты. Следовательно, для определения качества кластеризации и оценки выделенных кластеров необходим эксперт предметной области;

• Число кластеров, как правило, заранее не известно и выбирается по субъективным критериям. Даже если алгоритм не требует изначального знания о числе классов, конкретные реализации зачастую требуют указать этот параметр;

• Результат кластеризации существенно зависит от метрики. Однако существует ряд рекомендаций по выбору метрик для определенных классов задач.

Число кластеров фактически является гиперпараметром для алгоритмов кластеризации.

*Теорема невозможности Клейнберга*

Клейнберг постулировал три простых свойства в качестве аксиом кластеризации и доказал теорему, связывающую эти свойства.

• Алгоритм кластеризации *a* является **масштабно инвариантным**(англ. *scale-invariant*), если для любой функции расстояния *ρ* и любой константы *α>0* результаты кластеризации с использованием расстояний *ρ* и *α*⋅*ρ* совпадают.

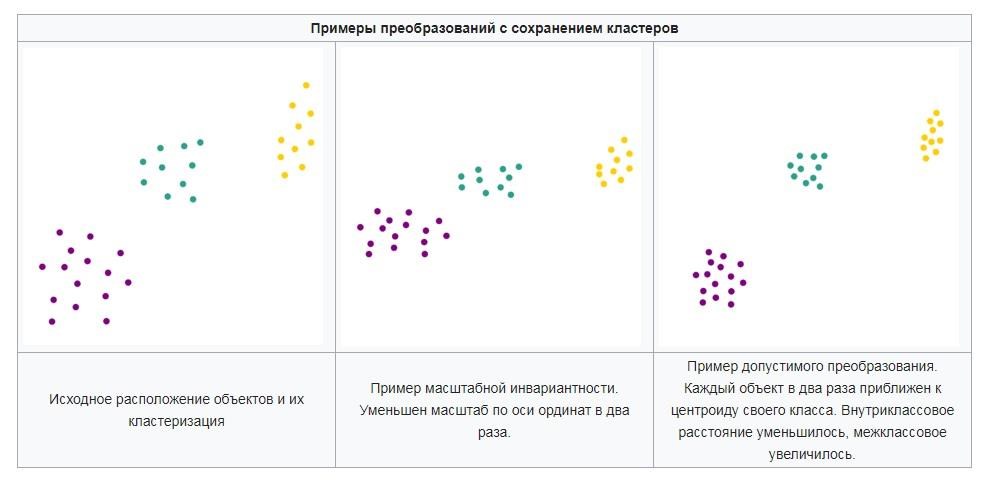
• **Полнота** (англ. *Richness*). Множество результатов кластеризации алгоритма aа в зависимости от изменения функции расстояния ρρ должно совпадать со множеством всех возможных разбиений множества объектов *X.*

• Алгоритм кластеризации является **согласованным** (англ. *consistent*), если результат кластеризации не изменяется после допустимого преобразования функции расстояния.

Исходя из этих аксиом Клейнберг сформулировал и доказал теорему:

**Для множестваобъектов, состоящего из двух и более элементов, не существует алгоритма кластеризации, который был бы одновременномасштабно-инвариантным,согласованным и полным.**

*Теорема невозможности Клейнберга*



*Типология задач кластеризации*

**Типы входных данных**

• Признаковое описание объектов. Каждый объект описывается набором своих характеристик, называемых признаками (англ. *features*). Признаки могут быть как числовыми, так и категориальными;

• Матрица расстояний между объектами. Каждый объект описывается расстоянием до всех объектов из обучающей выборки.

**Цели кластеризации**

• Классификация объектов. Попытка понять зависимости между объектами путем выявления их кластерной структуры. Разбиение выборки на группы схожих объектов упрощает дальнейшую обработку данных и принятие решений, позволяет применить к каждому кластеру свой метод анализа (стратегия «разделяй и властвуй»). В данном случае стремятся уменьшить число кластеров для выявления наиболее общих закономерностей;

• Сжатие данных. Можно сократить размер исходной выборки, взяв один или несколько наиболее типичных представителей каждого кластера. Здесь важно наиболее точно очертить границы каждого кластера, их количество не является важным критерием;

• Обнаружение новизны (обнаружение шума). Выделение объектов, которые не подходят по критериям ни в один кластер. Обнаруженные объекты в дальнейшем обрабатывают отдельно.

*Разнообразие алгоритмов кластеризации*

**Методы кластеризации**

•Графовые алгоритмы кластеризации. Наиболее примитивный класс алгоритмов. В настоящее время практически не применяется на практике;

•Вероятностные алгоритмы кластеризации. Каждый объект из обучающей выборки относится к каждому из кластеров с определенной степенью вероятности;

•Иерархические алгоритмы кластеризации. Упорядочивание данных путем создания иерархии вложенных кластеров;

•Алгоритм *k*-средних (англ. *k-means*). Итеративный алгоритм, основанный на минимизации суммарного квадратичного отклонения точек кластеров от центров этих кластеров; •Распространение похожести (англ. *affinity propagation*). Распространяет сообщения о похожести между парами объектов для выбора типичных представителей каждого кластера;

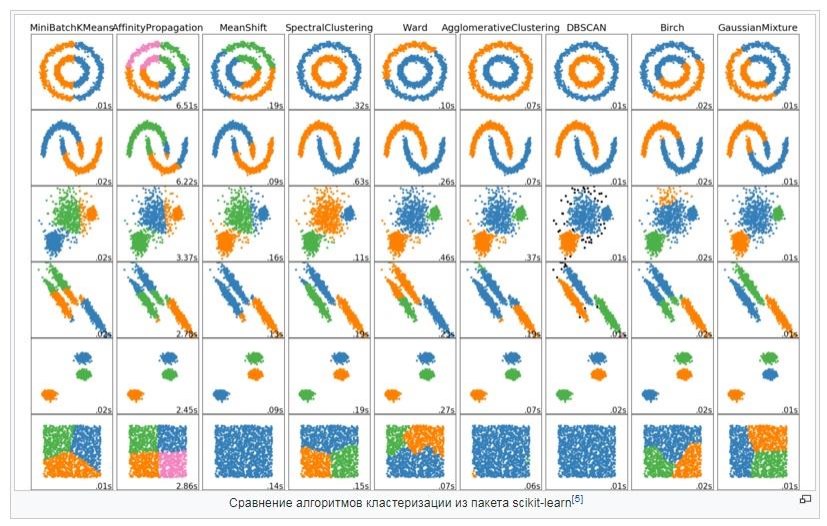
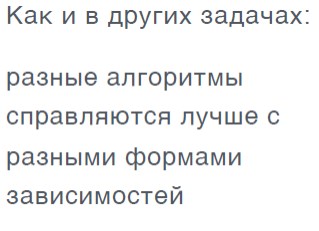
•Сдвиг среднего значения (англ. *mean shift*). Выбирает центроиды кластеров в областях с наибольшей плотностью;

•Спектральная кластеризация (англ. *spectral clustering*). Использует собственные значения матрицы расстояний для понижения размерности перед использованием другихметодов кластеризации;

•Основанная на плотности пространственная кластеризация для приложений с шумами

(англ. *Density-based spatial clustering of applications with noise*, *DBSCAN*). Алгоритм группирует в один кластер точки в области с высокой плотностью. Одиноко расположенные точки помечает как шум.

*АЛГОРИТМЫ КЛАСТЕРИЗАЦИИ*



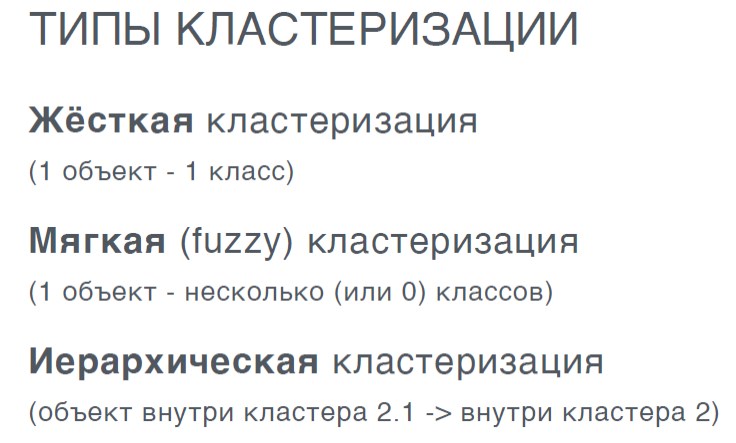
*АЛГОРИТМЫ КЛАСТЕРИЗАЦИИ*

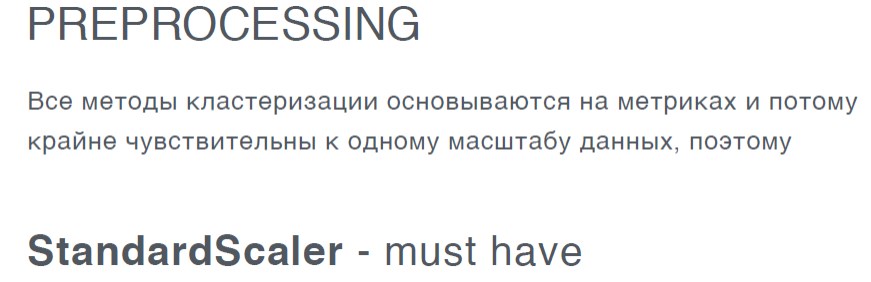


*Меры качества кластеризации*

Для оценки качества кластеризации задачуможно переформулироватьв терминах задачидискретной оптимизации. Необходима так сопоставить объектамиз множества *X* метки кластеров, чтобызначение выбранного функционала качества приняло наилучшее значение.

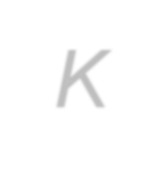
В качестве примера,стремятся достичьминимума среднеговнутрикластерного расстояния или максимума среднего межкластерногорасстояния.

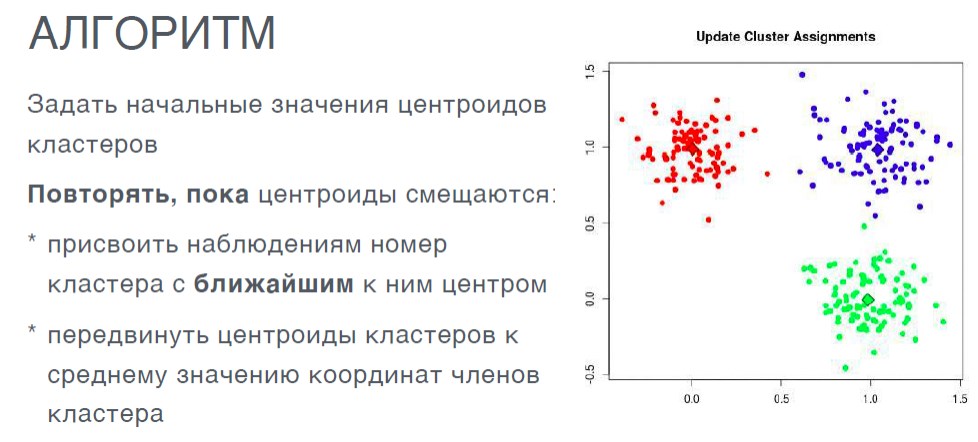
*АЛГОРИТМЫ КЛАСТЕРИЗАЦИИ*

*АЛГОРИТМЫ КЛАСТЕРИЗАЦИИ. PREPROCESSING*



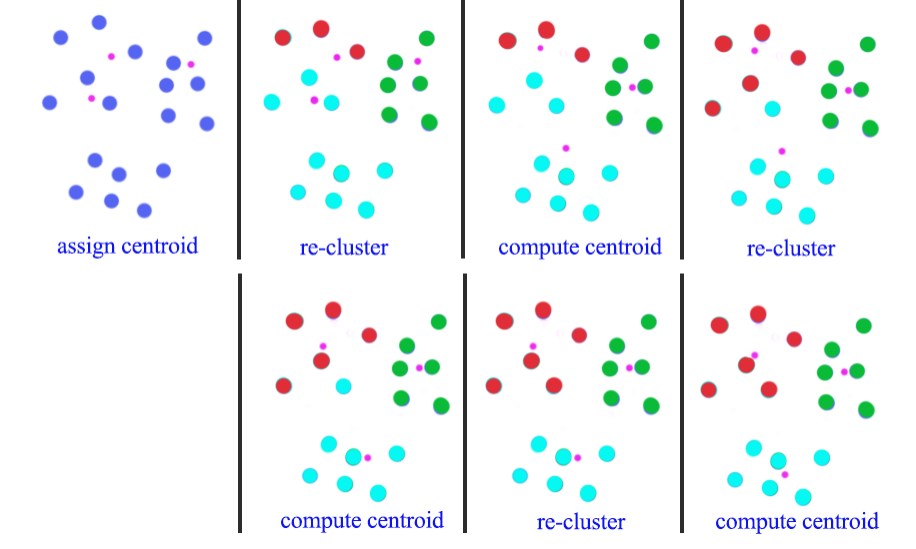
*K-MEANS*

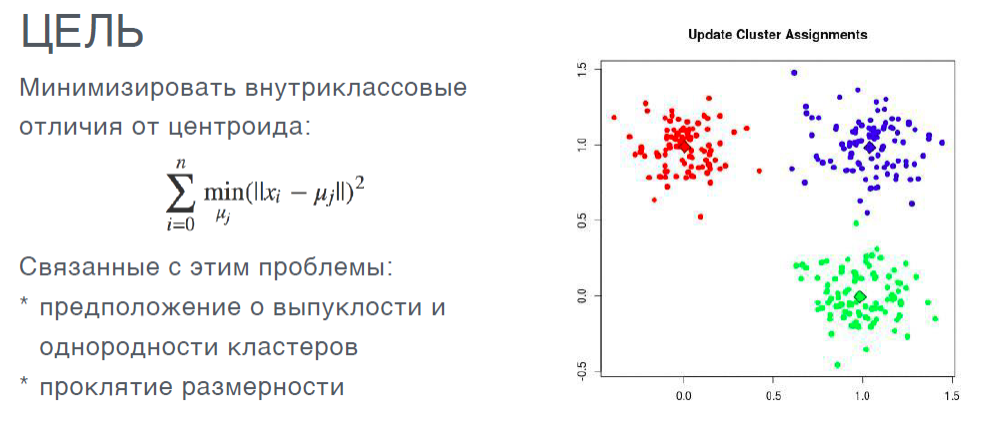


*АЛГОРИТМЫ КЛАСТЕРИЗАЦИИ. K-MEANS*



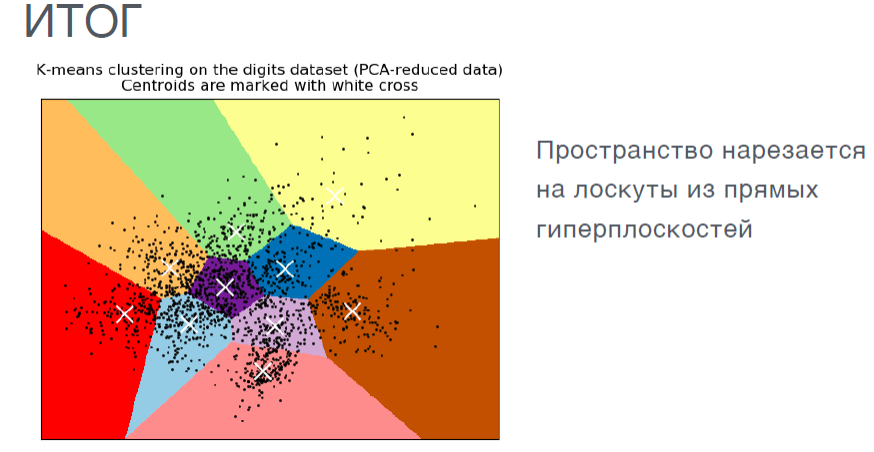
*АЛГОРИТМЫ КЛАСТЕРИЗАЦИИ. K-MEANS*

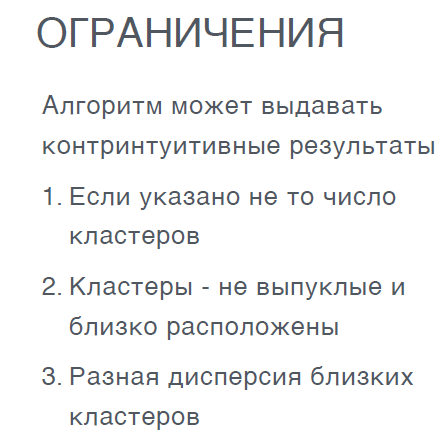


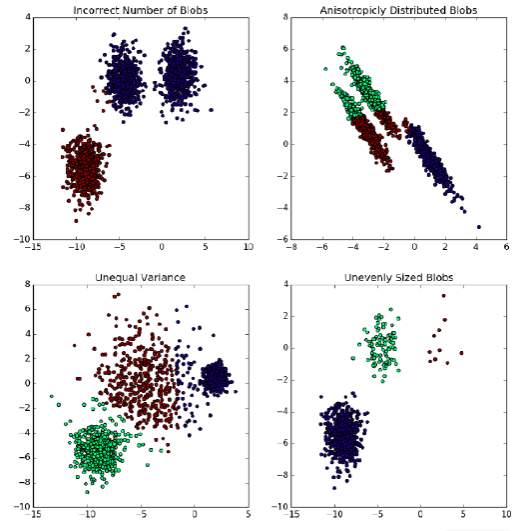
*АЛГОРИТМЫ КЛАСТЕРИЗАЦИИ. K-MEANS*

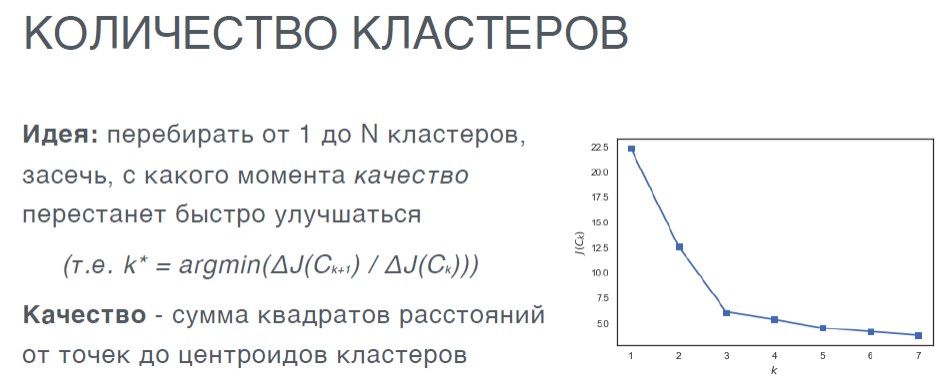


*АЛГОРИТМЫ КЛАСТЕРИЗАЦИИ. K-MEANS*



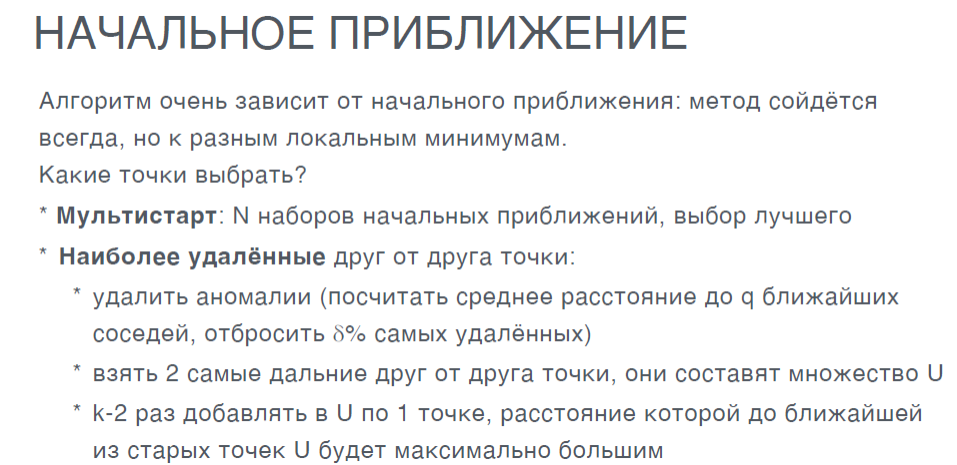
*АЛГОРИТМЫ КЛАСТЕРИЗАЦИИ. K-MEANS*



*АЛГОРИТМЫ КЛАСТЕРИЗАЦИИ. K-MEANS*



*АЛГОРИТМЫ КЛАСТЕРИЗАЦИИ. K-MEANS*

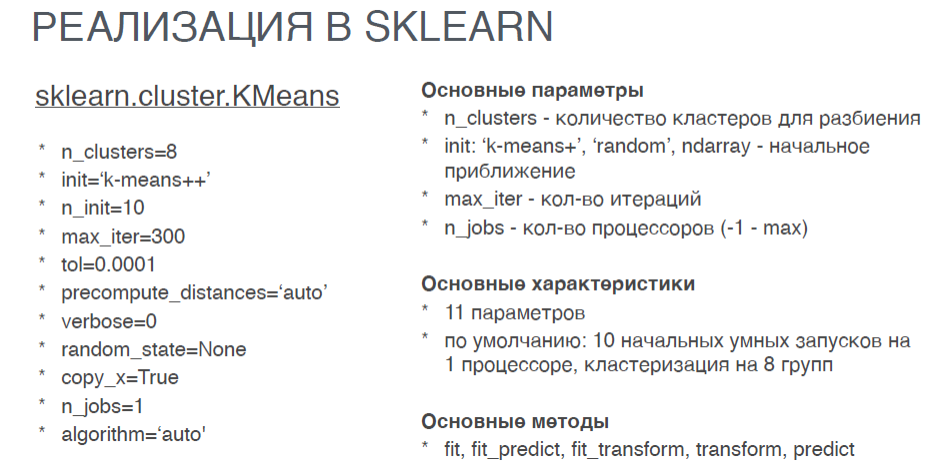


*АЛГОРИТМЫ КЛАСТЕРИЗАЦИИ. K-MEANS*



*АЛГОРИТМЫ КЛАСТЕРИЗАЦИИ. K-MEANS*

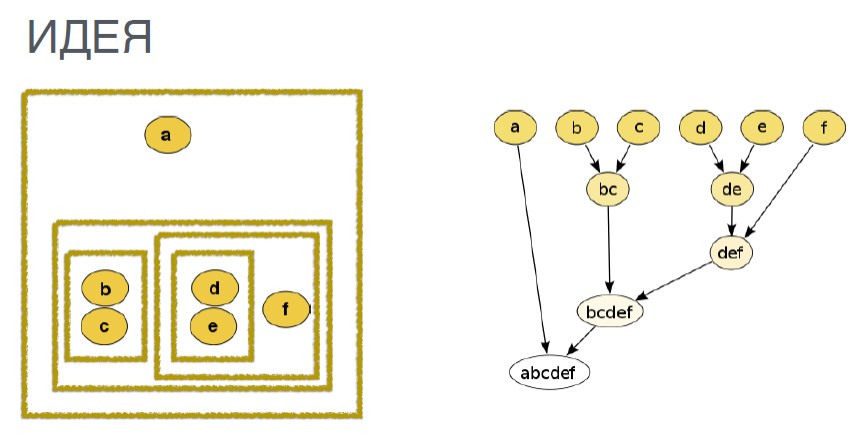


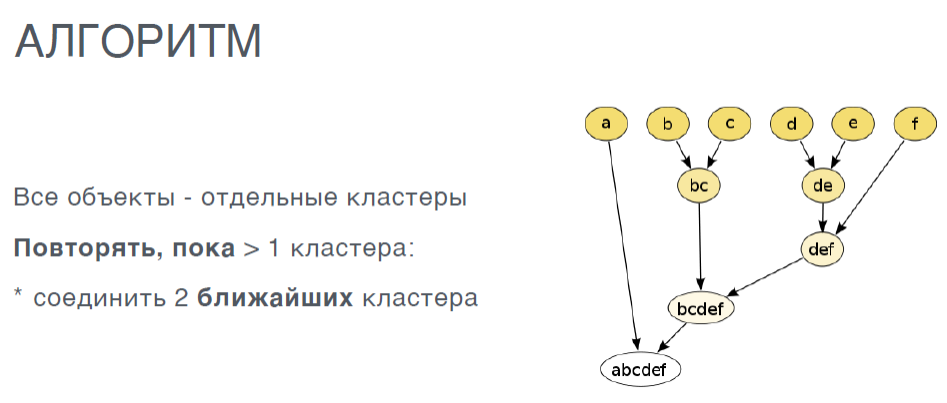
*АЛГОРИТМЫ КЛАСТЕРИЗАЦИИ. K-MEANS*



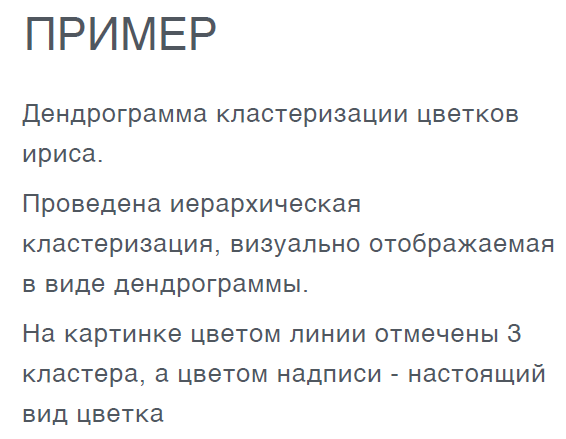
*HIERARCHICAL CLUSTERING*

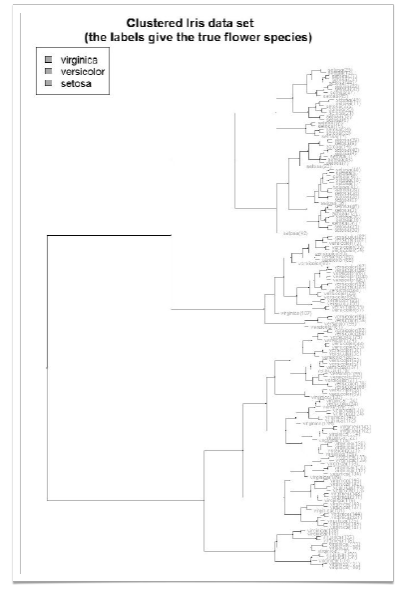
*АЛГОРИТМЫ КЛАСТЕРИЗАЦИИ. HIERARCHICAL CLUSTERING*

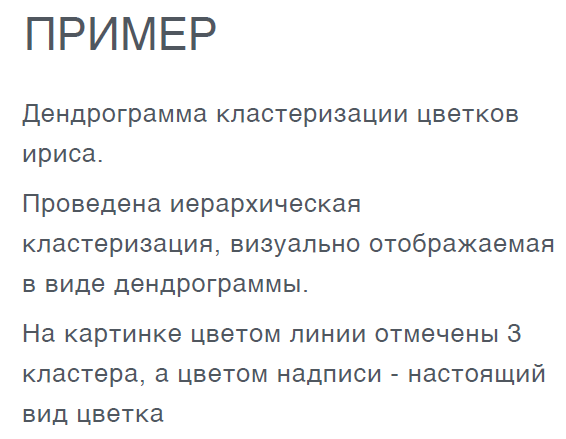


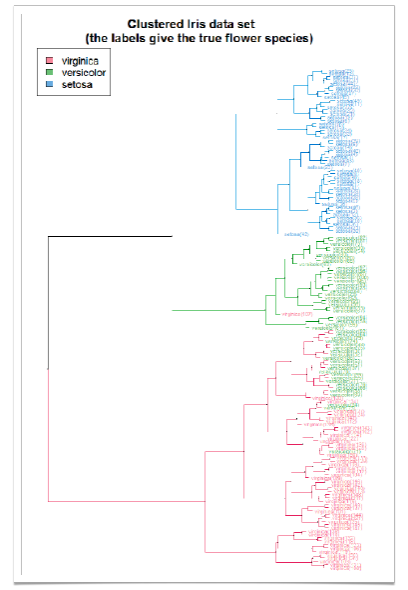
*АЛГОРИТМЫ КЛАСТЕРИЗАЦИИ. HIERARCHICAL CLUSTERING*



*АЛГОРИТМЫ КЛАСТЕРИЗАЦИИ. HIERARCHICAL CLUSTERING*



*АЛГОРИТМЫ КЛАСТЕРИЗАЦИИ. HIERARCHICAL CLUSTERING*



*АЛГОРИТМЫ КЛАСТЕРИЗАЦИИ. HIERARCHICAL CLUSTERING*

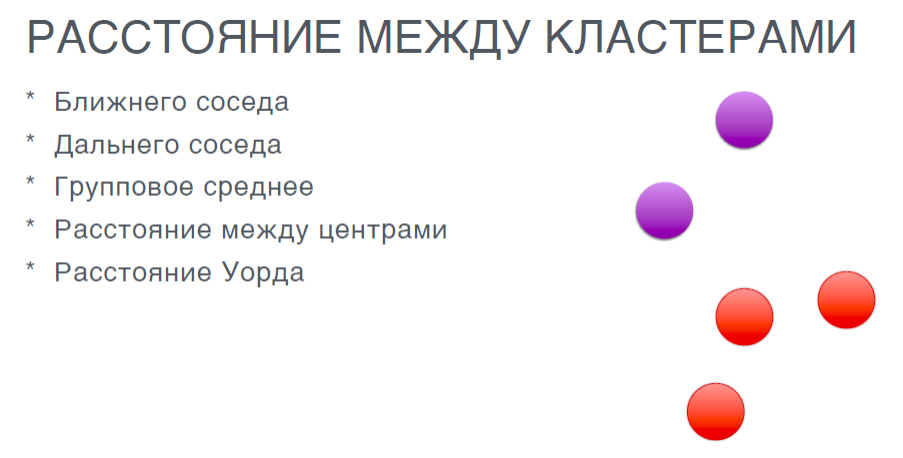


*АЛГОРИТМЫ КЛАСТЕРИЗАЦИИ. HIERARCHICAL CLUSTERING*

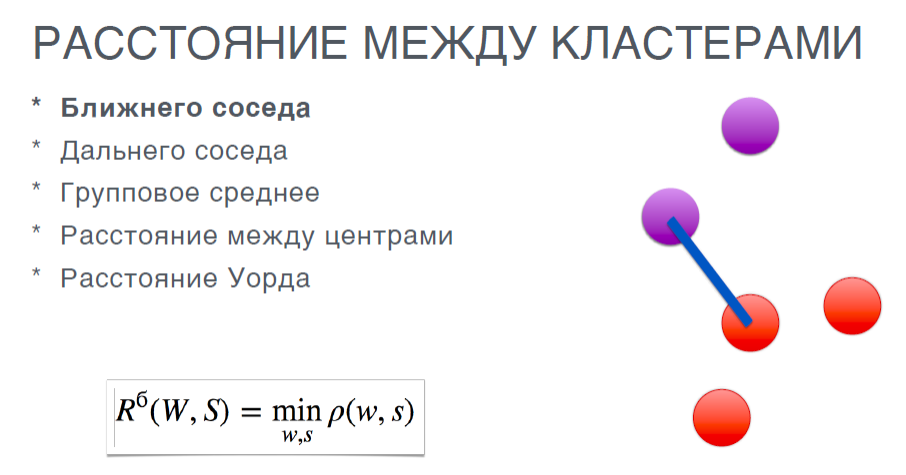


[*\* sklearn, clustering example*](https://scikit-learn.org/stable/auto_examples/cluster/plot_agglomerative_clustering_metrics.html)

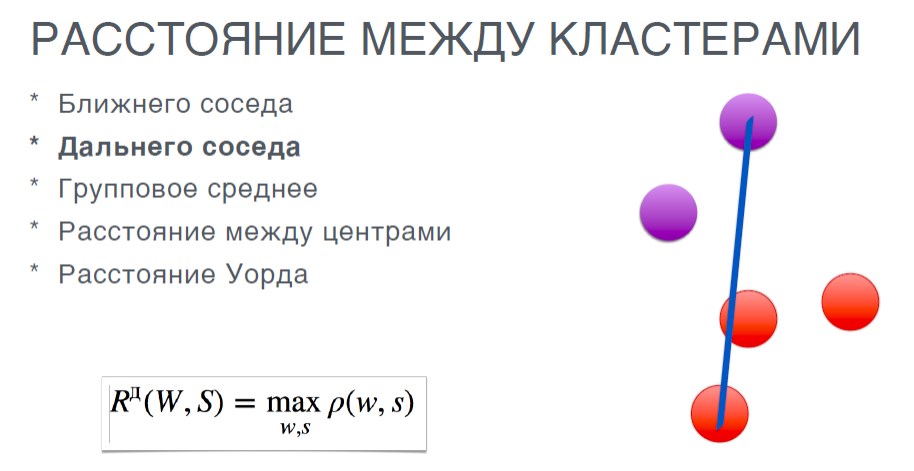


*АЛГОРИТМЫ КЛАСТЕРИЗАЦИИ. HIERARCHICAL CLUSTERING*



*АЛГОРИТМЫ КЛАСТЕРИЗАЦИИ. HIERARCHICAL CLUSTERING*

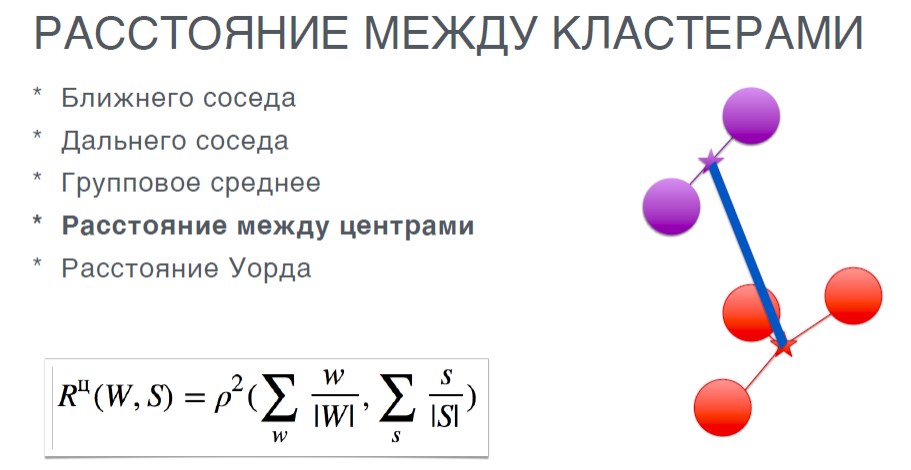


*АЛГОРИТМЫ КЛАСТЕРИЗАЦИИ. HIERARCHICAL CLUSTERING*

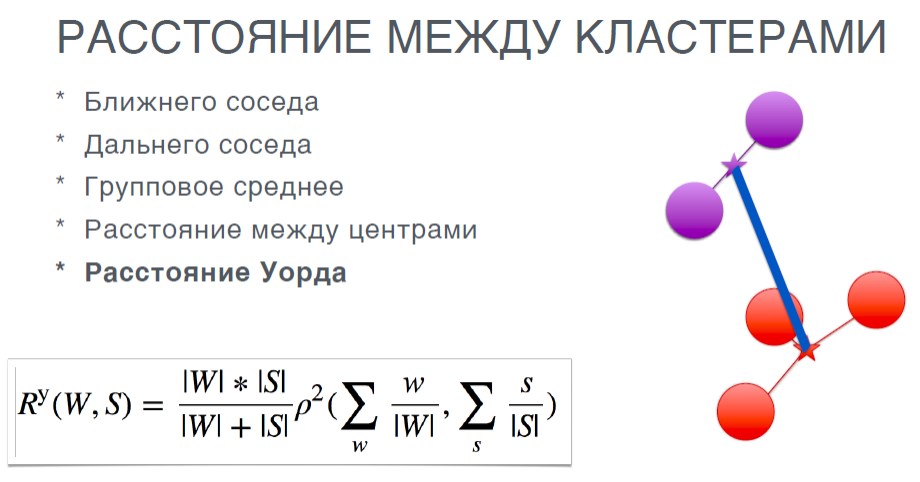


*АЛГОРИТМЫ КЛАСТЕРИЗАЦИИ. HIERARCHICAL CLUSTERING*

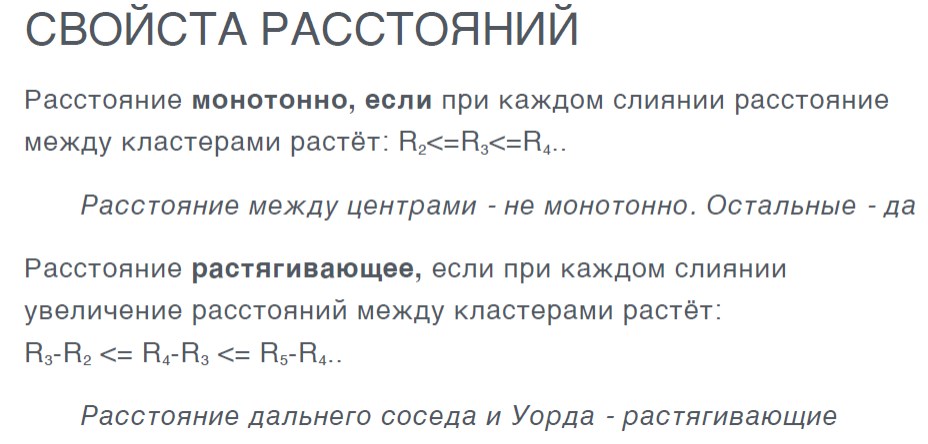


*АЛГОРИТМЫ КЛАСТЕРИЗАЦИИ. HIERARCHICAL CLUSTERING*

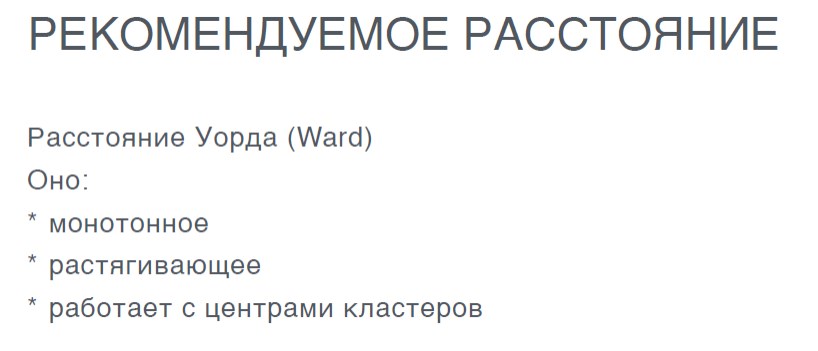


*АЛГОРИТМЫ КЛАСТЕРИЗАЦИИ. HIERARCHICAL CLUSTERING*

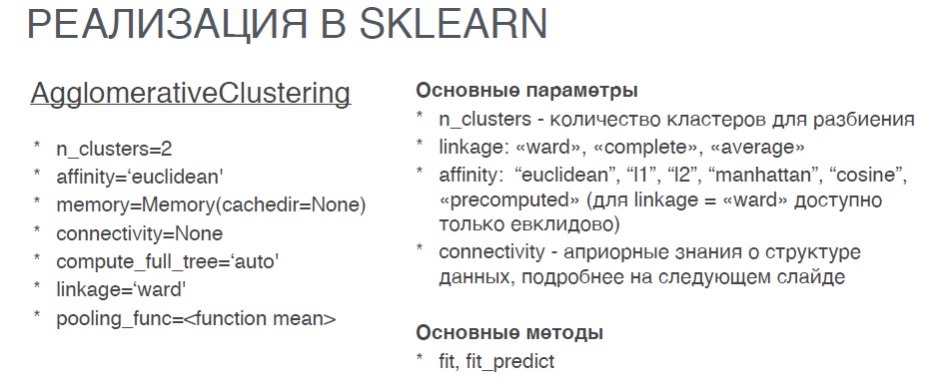


*АЛГОРИТМЫ КЛАСТЕРИЗАЦИИ. HIERARCHICAL CLUSTERING*

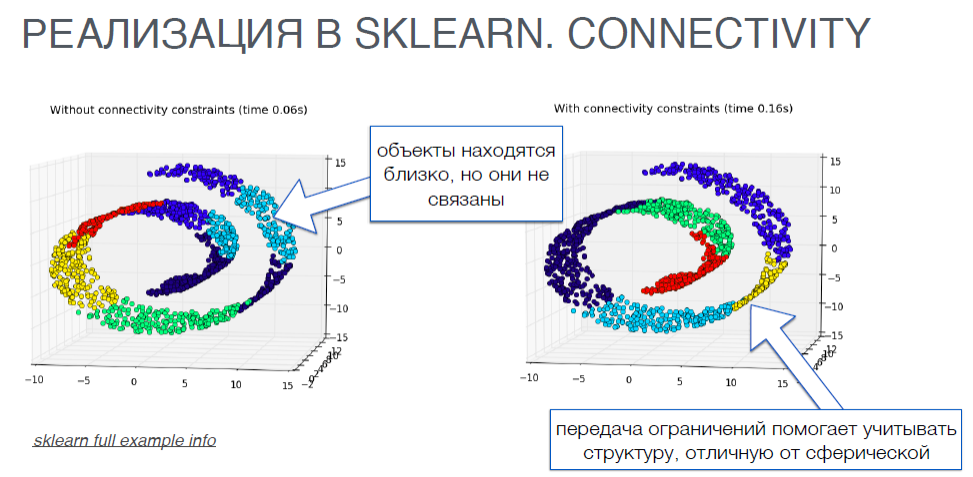


*АЛГОРИТМЫ КЛАСТЕРИЗАЦИИ. HIERARCHICAL CLUSTERING*



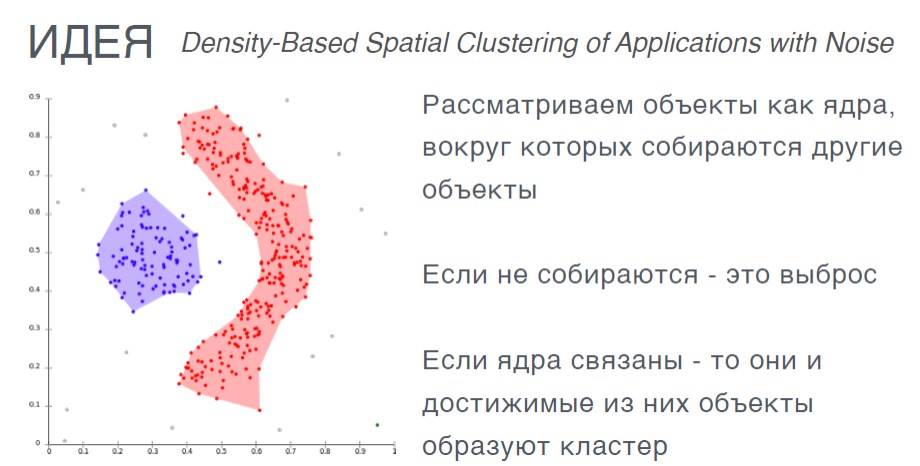
*АЛГОРИТМЫ КЛАСТЕРИЗАЦИИ. HIERARCHICAL CLUSTERING*



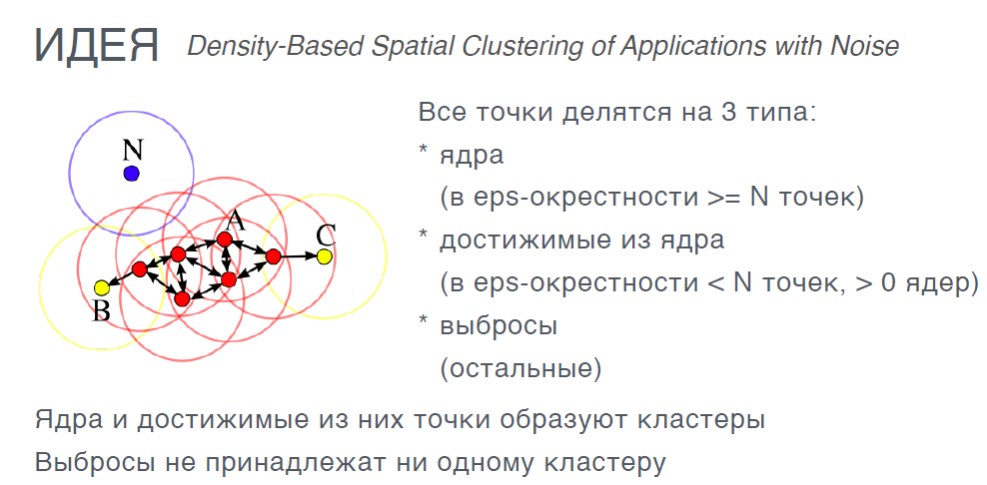
*АЛГОРИТМЫ КЛАСТЕРИЗАЦИИ. HIERARCHICAL CLUSTERING*



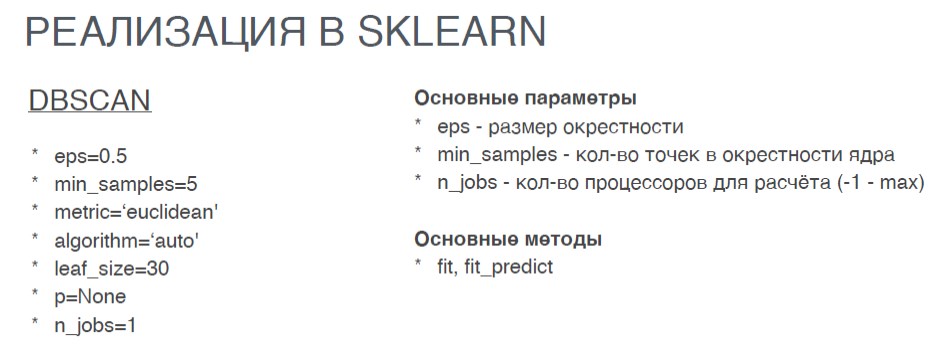
*DBSCAN*

*АЛГОРИТМЫ КЛАСТЕРИЗАЦИИ. DBSCAN*

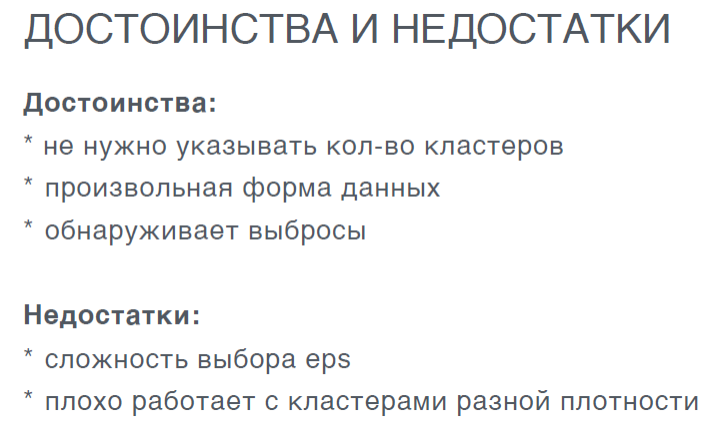


*АЛГОРИТМЫ КЛАСТЕРИЗАЦИИ. DBSCAN*



*АЛГОРИТМЫ КЛАСТЕРИЗАЦИИ. DBSCAN*



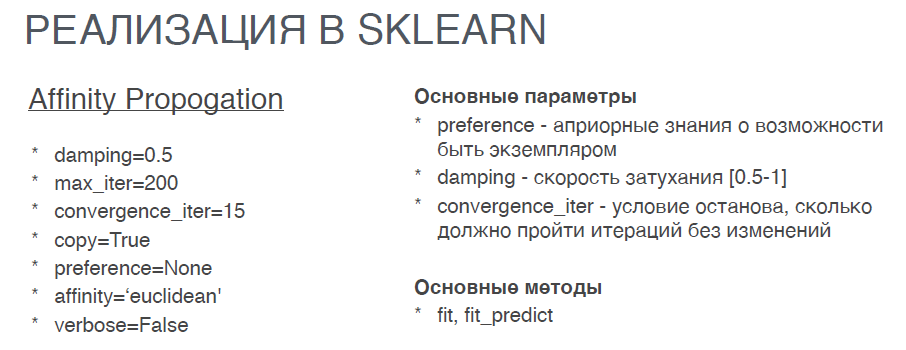
*АЛГОРИТМЫ КЛАСТЕРИЗАЦИИ. DBSCAN*



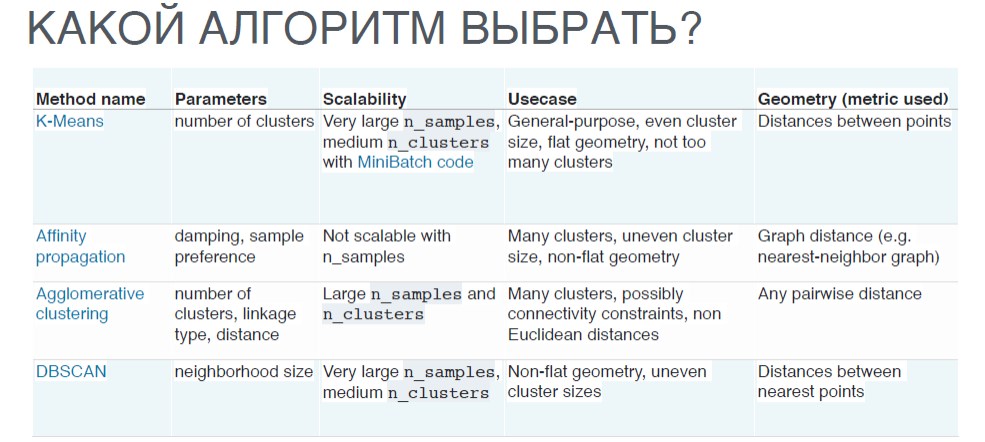
*AFFINITY PROPAGATION*

*АЛГОРИТМЫ КЛАСТЕРИЗАЦИИ. AFFINITY PROPAGATION*



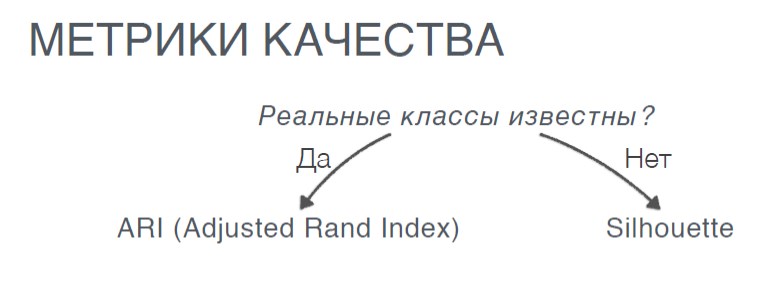
*АЛГОРИТМЫ КЛАСТЕРИЗАЦИИ. AFFINITY PROPAGATION*



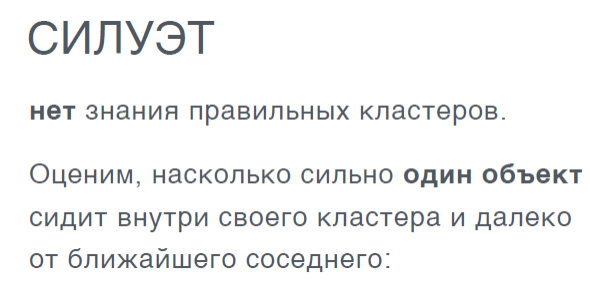
*АЛГОРИТМЫ КЛАСТЕРИЗАЦИИ. ВЫБОР АЛГОРИТМА*

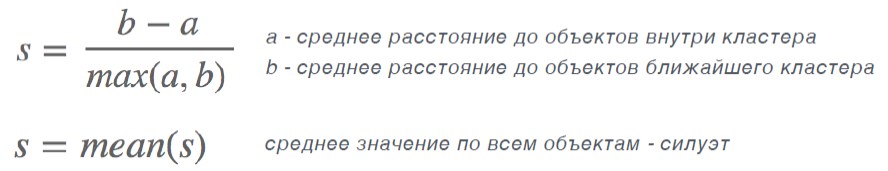


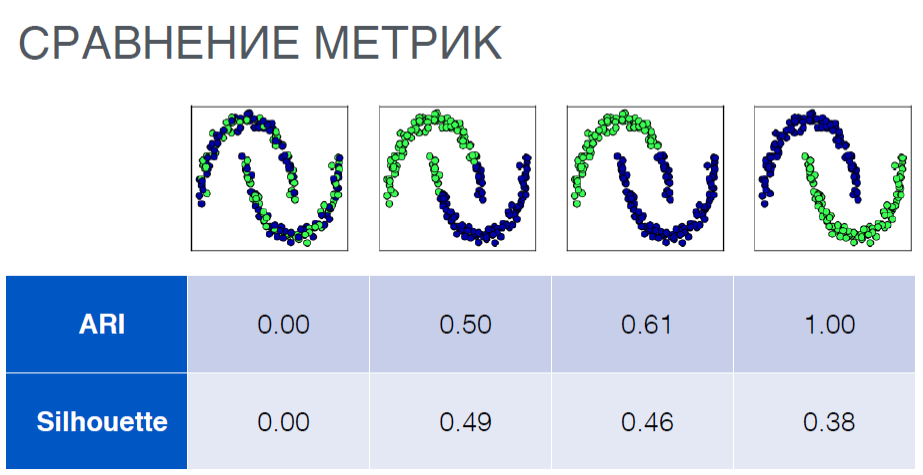
[*\* sklearn, сравнение кластеризаторов*](https://scikit-learn.org/stable/modules/clustering.html)

*МЕТРИКИ КАЧЕСТВА КЛАСТЕРИЗАЦИИ*

*МЕТРИКИ КАЧЕСТВА КЛАСТЕРИЗАЦИИ*

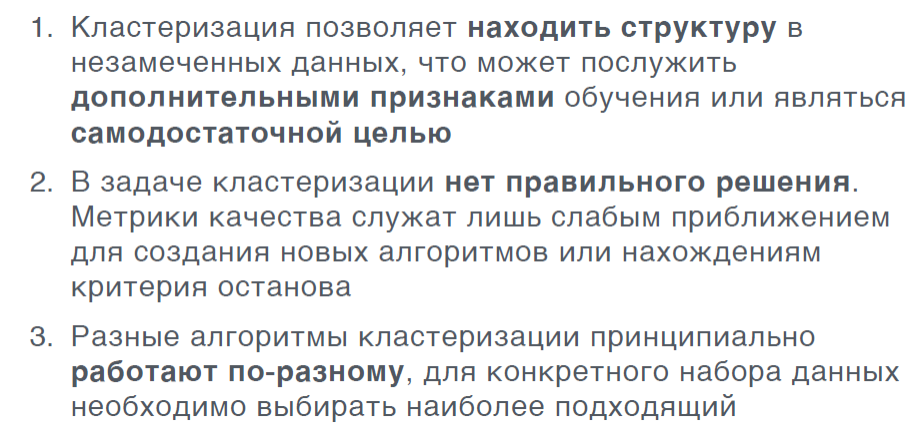
*МЕТРИКИ КАЧЕСТВА КЛАСТЕРИЗАЦИИ*

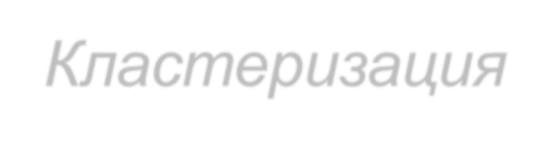


*МЕТРИКИ КАЧЕСТВА КЛАСТЕРИЗАЦИИ*

***ПРАКТИКА***

*clustering.ipynb*

*ЧТО МЫ СЕГОДНЯ УЗНАЛИ*

*Кластеризация*