

**Тема лекции:**

**Кластерный анализ**

# ПОНЯТИЕ КЛАСТЕРИЗАЦИИ

Во многих прикладных задачах измерять степень сходства объектов существенно проще, чем формировать признаковые описания. Например, гораздо легче сравнить две фотографии и сказать, что они принадлежат одному человеку, чем понять, на основании каких признаков они схожи. Задача классификации объектов на основе их сходства друг с другом, когда принадлежность обучающих объектов каким-либо классам не задаётся, называется задачей **кластеризации**.

**Кластеризация** — это процесс автоматического разбиения некоторого множества элементов на группы на основе степени их схожести (кластеры).

**Кластерный анализ** (*cluster analysis*) — многомерная статистическая процедура, выполняющая сбор данных, содержащих информацию о выборке объектов, и затем упорядочивающая объекты в сравнительно однородные группы.

Задача кластеризации относится к статистической обработке, а также к широкому классу задач обучения без учителя.



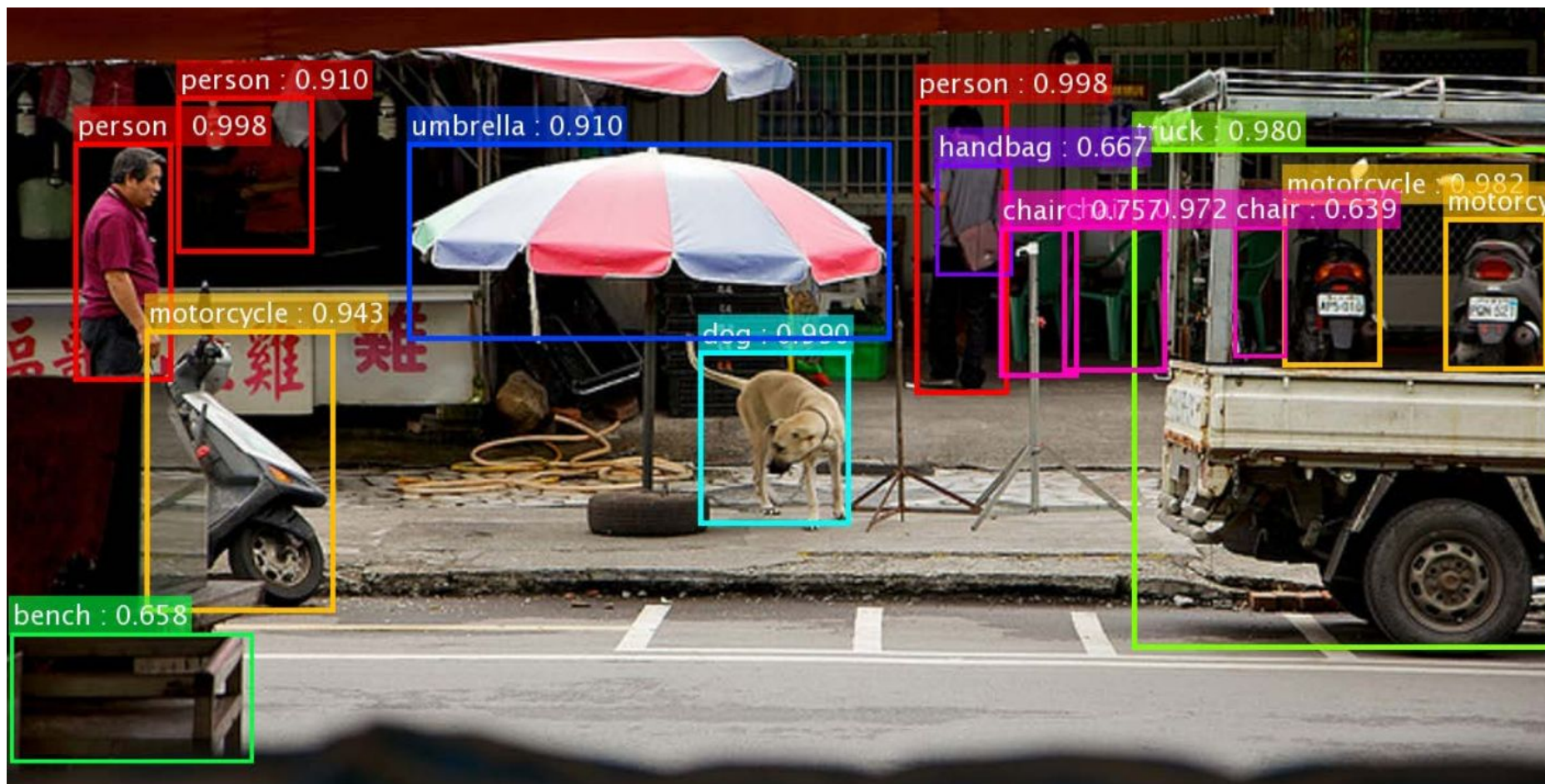
# ЗАДАЧИ И УСЛОВИЯ КЛАСТЕРИЗАЦИИ

- Понять структуру множества объектов, разбив его на группы схожих объектов. Упростить дальнейшую обработку данных и принятия решений, работая с каждым кластером по отдельности (стратегия «разделяй и властвуй»)
- Сократить объём хранимых данных в случае сверхбольшой выборки, оставив по одному наиболее типичному представителю от каждого кластера
- Выделить нетипичные объекты, которые не подходят ни к одному из кластеров. Эту задачу называют одноклассовой классификацией, обнаружением нетипичности или новизны (novelty detection)
- Вычисление значений той или иной меры сходства (или различия) между объектами

Во всех этих случаях может применяться иерархическая кластеризация, когда крупные кластеры дробятся на более мелкие, те в свою очередь дробятся ещё мельче, и т. д. Такие задачи называются задачами **таксономии** (*taxonomy*). Результатом таксономии является **не простое разбиение** множества объектов на кластеры, а **древообразная иерархическая структура**. Вместо номера кластера объект характеризуется перечислением всех кластеров, которым он принадлежит, от крупного к мелкому.

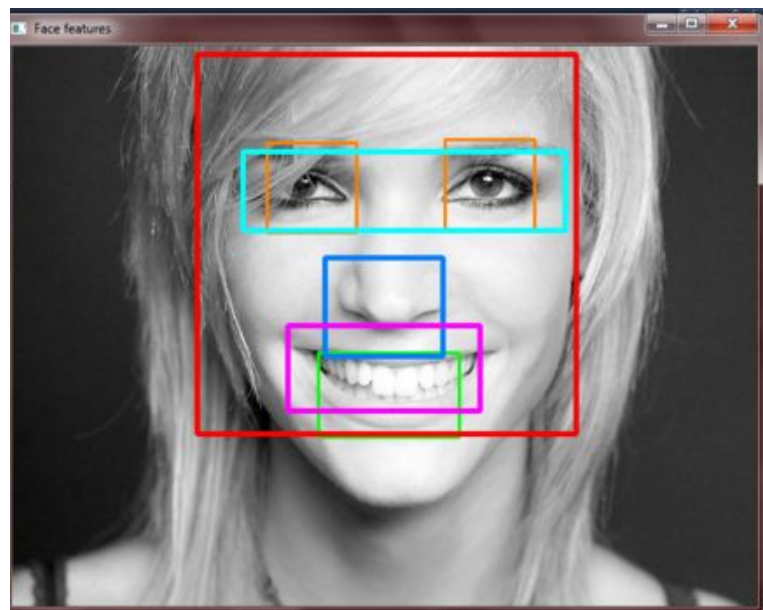
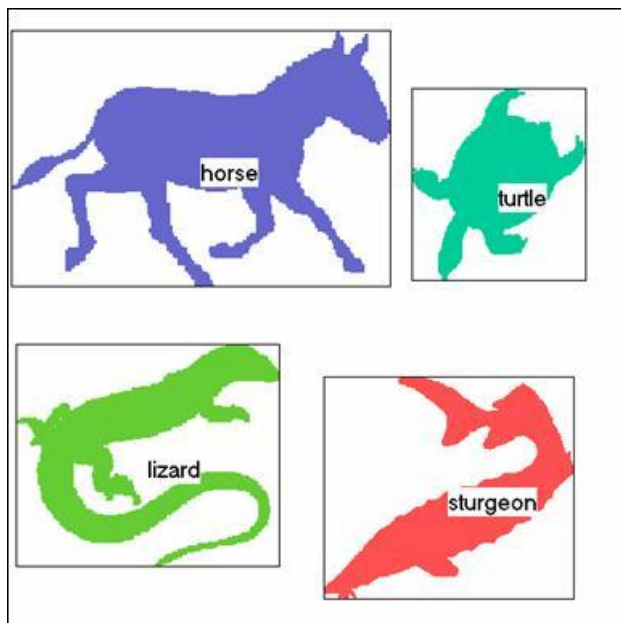
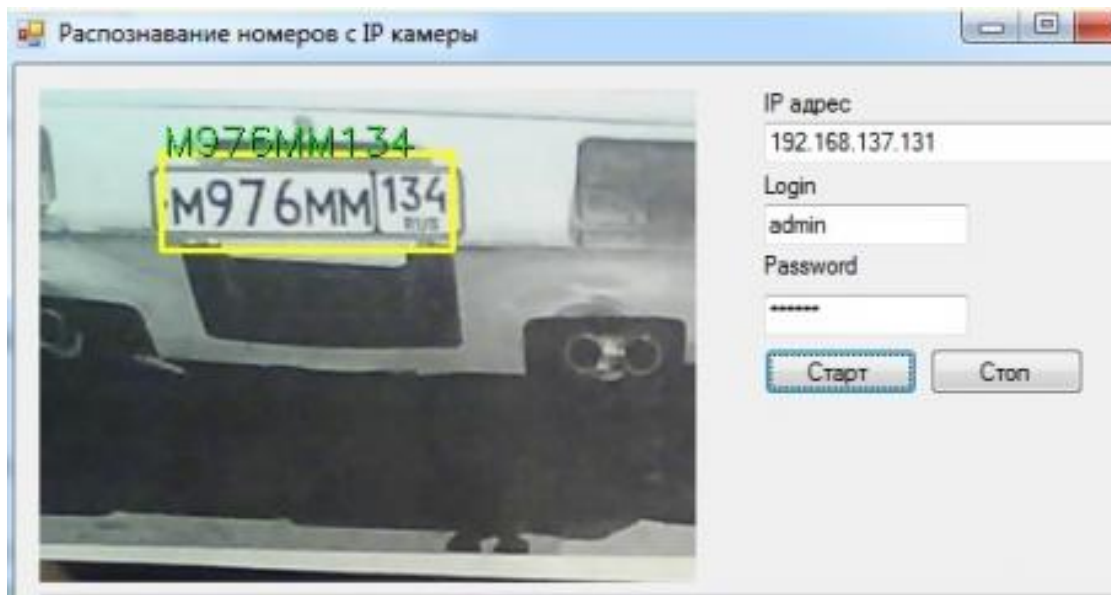
# ПРИМЕНЕНИЕ КЛАСТЕРИЗАЦИИ

## Распознавание образов



# ПРИМЕНЕНИЕ КЛАСТЕРИЗАЦИИ

## Распознавание образов



# ПРИМЕНЕНИЕ КЛАСТЕРИЗАЦИИ

## Группировка объектов

харьков театр

Рейтинг  
★★★★★

Театр им. Т.Г. Шевченко

4,9 ★★★★★ (20)

Театр · вулиця Сумська



Харьковский Академический  
Русский Драматический Театр им.  
А. С. Пушкина

4,5 ★★★★★ (26)

Театр · вулиця Чернишевська



Харьковский национальный  
академический театр оперы и  
балета имени Н. Лысенко

4,5 ★★★★★ (52)

Сцена · вулиця Сумська



Харьковский государственный  
академический театр кукол имени  
В. А. Афанасьева

4,9 ★★★★★ (17)

Театр · площа Конституції



Театр для детей и юношества (ТЮЗ)

4,6 ★★★★★ (16)

Театр · вулиця Полтавський Шлях



Театр музичної комедії

4,5 ★★★★★ (11)

Театр · вулиця Благовіщенська



Дом актера

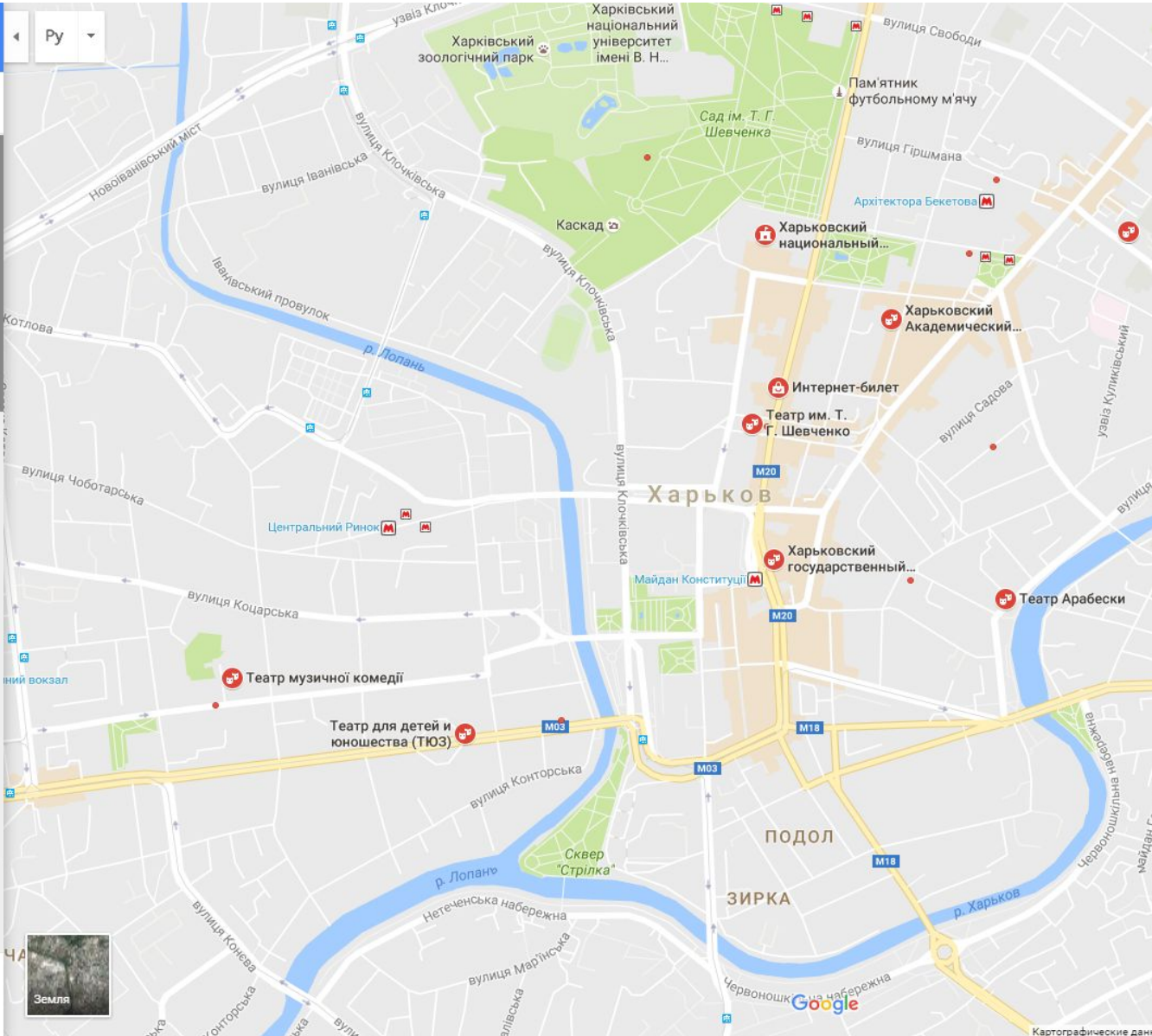
4,6 ★★★★★ (45)

Театр · вулиця Манізера



Результаты с 1 по 20

☒ Обновлять результаты при смещении карты



Map of Kharkiv, Ukraine, showing the city center and surrounding areas. The map includes the Dnieper River (р. Дніпро) and the Kharkiv River (р. Харків). Key landmarks and locations marked include:

- Харківський національний університет імені В. Н...
- Харківський зоологічний парк
- Сад ім. Т. Г. Шевченка
- Пам'ятник футбольному м'ячу
- Харківський національний...
- Харківський Академический...
- Интернет-билет
- Театр им. Т. Г. Шевченко
- Харьковский государственный...
- Театр Арабески
- Театр для детей и юношества (ТЮЗ)
- Театр музичної комедії
- Дом актера
- Харьковский государственный академический театр кукол имени В. А. Афанасьева
- Харьковский национальный академический театр оперы и балета имени Н. Лысенко
- Харьковский Академический Русский Драматический Театр им. А. С. Пушкина
- Театр им. Т.Г. Шевченко

The map also shows major roads (M20, M18, M03) and districts (Харьков, ПОДОЛ, ЗИРКА, ЧЕРВОНОШКА).

# ПРИМЕНЕНИЕ КЛАСТЕРИЗАЦИИ

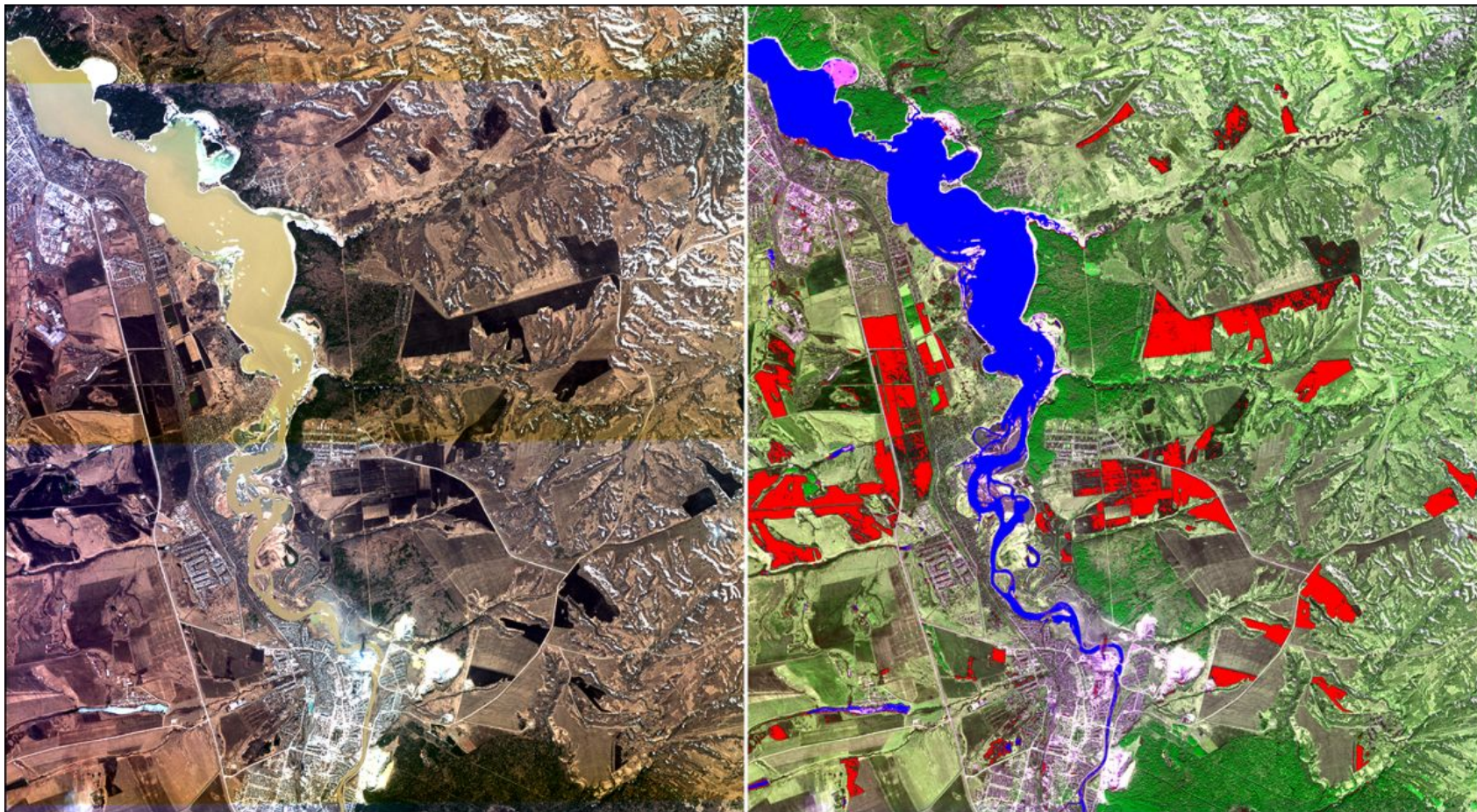
## Классификация результатов поиска





# ПРИМЕНЕНИЕ КЛАСТЕРИЗАЦИИ

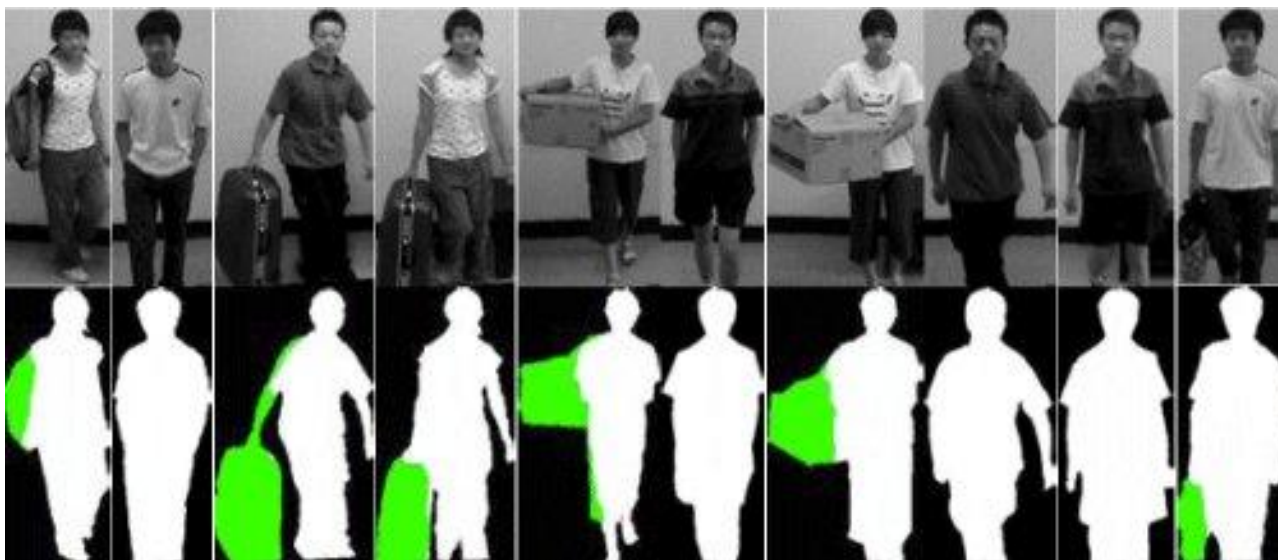
## Сегментация изображений



*RGB-композит изображения, полученного со спутника «Канопус-В»,  
и результат выделения водных объектов и переувлажненных почв*

# ПРИМЕНЕНИЕ КЛАСТЕРИЗАЦИИ

## Сегментация изображений



Image

Objects

Class



# ПРИМЕНЕНИЕ КЛАСТЕРИЗАЦИИ

**Кластеризация результатов поиска** — используется для «интеллектуальной» группировки результатов при поиске файлов, веб-сайтов, других объектов, предоставляя пользователю возможность быстрой навигации, выбора заведомо более релевантного подмножества и исключения заведомо менее релевантного — что может повысить юзабилити интерфейса по сравнению с выводом в виде простого сортированного по релевантности списка.

**Сегментация изображений** — кластеризация может быть использована для разбиения цифрового изображения на отдельные области с целью обнаружения границ (edge detection) или распознавания объектов.

**Интеллектуальный анализ данных (data mining)** — кластеризация в Data Mining приобретает ценность тогда, когда она выступает одним из этапов анализа данных, построения законченного аналитического решения. Аналитику часто легче выделить группы схожих объектов, изучить их особенности и построить для каждой группы отдельную модель, чем создавать одну общую модель для всех данных. Таким приемом постоянно пользуются в маркетинге, выделяя группы клиентов, покупателей, товаров и разрабатывая для каждой из них отдельную стратегию.

# ОБЩИЙ АЛГОРИТМ КЛАСТЕРИЗАЦИИ

- Выбор меры близости;
- Выбор алгоритма кластеризации;
- Представление полученных результатов;



# МЕРЫ БЛИЗОСТИ

**Коэффициент сходства** (также мера сходства, индекс сходства) — безразмерный показатель, применяемый для количественного определения степени сходства объектов.

- Степенное расстояние

$$\rho(x, y) = \sqrt[r]{\sum_{i=1}^n |x_i - y_i|^p}$$

Евклидово расстояние и квадрат евклидова расстояния

$$\rho(x, y) = \left( \sum_{i=1}^n (x_i - y_i)^2 \right)^{1/2}$$

$$\rho(x, y) = \sum_{i=1}^n (x_i - y_i)^2$$

- Расстояние городских кварталов (манхэттенское расстояние)

$$\rho(x, y) = \sum_{i=1}^n |x_i - y_i|$$

- Расстояние Чебышева

$$\rho(x, y) = \max_{i=1..n} |x_i - y_i|$$

# Обобщенный алгоритм Ллойда (Generalized Lloyd) или алгоритм k-средних (k-means) $O(k \cdot n \cdot ?)$

Метод k-средних – это метод кластерного анализа, цель которого является разделение  $n$  наблюдений из пространства  $R^n$  на  $k$  кластеров, при этом каждое наблюдение относится к тому кластеру, к центру (центроиду) которого оно ближе всего.

В качестве меры близости используется Евклидово расстояние.

Метод k-средних разделяет  $n$  наблюдений на  $k$  групп (или кластеров) ( $k \leq m$ ) чтобы минимизировать суммарное квадратичное отклонение точек кластеров от центроидов этих кластеров.

$$\min \left[ \sum_{i=1}^k \sum_{x^{(j)} \in S_i} \|x^{(j)} - \mu_i\|^2 \right], \text{ где } x^{(j)} \in R^n, \mu_i \in R^n$$

$\mu_i$  - центроид для кластера  $S_i$ .

1. На первом этапе центроиды кластеров выбираются случайно или по определенному правилу (например, выбрать центроиды, максимизирующие начальные расстояния между кластерами).

# Обобщенный алгоритм Ллойда (Generalized Lloyd) или алгоритм k-средних (k-means) $O(k \cdot n \cdot ?)$

**2.** Относим наблюдения к тем кластерам, чье среднее (центроид) к ним ближе всего. Каждое наблюдение принадлежит только к одному кластеру, даже если его можно отнести к двум и более кластерам.

**3.** Затем центроид каждого  $i$ -го кластера перевычисляется по следующему правилу:

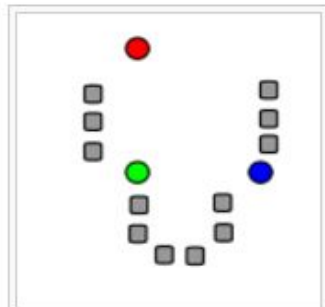
$$\mu_i = \frac{1}{s_i} \sum_{x^{(j)} \in S_i} x^{(j)}$$

Таким образом, алгоритм k-средних заключается в перевычислении на каждом шаге центроида для каждого кластера, полученного на предыдущем шаге.

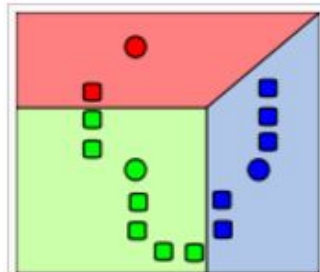
Алгоритм останавливается, когда значения  $\mu_i$  не меняются:

Неправильный выбор первоначального числа кластеров  $k$  может привести к некорректным результатам. Именно поэтому при использовании метода k-средних важно сначала провести проверку подходящего числа кластеров для данного набора данных.

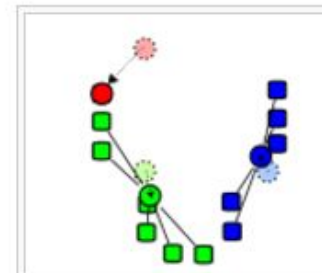
# Обобщенный алгоритм Ллойда (Generalized Lloyd) или алгоритм k-средних (k-means) $O(k \cdot n \cdot ?)$



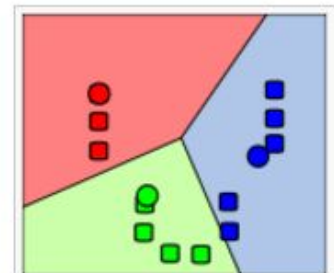
Исходные точки и  
случайно выбранные  
начальные точки.



Точки, отнесённые к  
начальным центрам.  
Разбиение на  
плоскости —  
**диаграмма Вороного**  
относительно  
начальных центров.



Вычисление новых  
центров кластеров  
(Ищется **центр масс**).



Предыдущие шаги  
повторяются, пока  
алгоритм не сойдётся.

# Кластеризация объединением ближайших соседей (pairwise nearest neighbor) $O(n^2)$

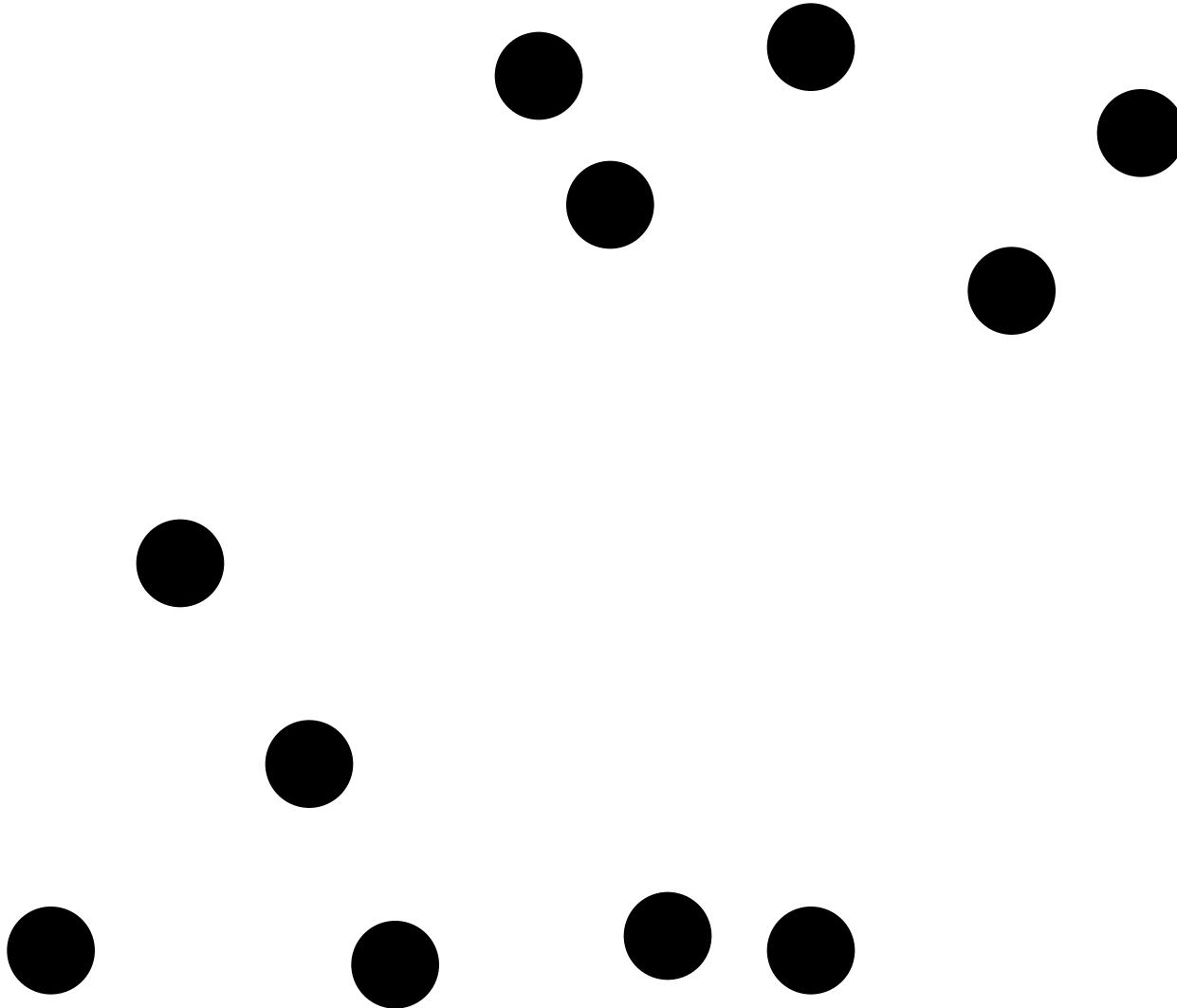
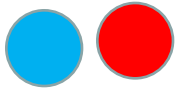
Дано:  $n$  точек  $x_i$  в многомерном пространстве, которые нужно разбить на  $k$  кластеров.

1. Считаем каждую точку отдельным кластером. Таким образом, алгоритм стартует с  $n$  кластеров. Центр каждого из этих кластеров совпадает с координатами точки, его образующей.
2. Находим два кластера, центры кластеров которых ближе всего друг к другу.
3. Объединяем эти два кластера в один и вычисляем координаты его центра, усреднением координат всех входящих в кластер точек.

Шаги 2-3 повторяются до тех пор, пока число кластеров не уменьшится до заданного числа  $k$ .

Как вариант, целевым может быть не число кластеров  $k$ , а расстояние между центрами объединяемых кластеров, которое должно быть не больше некоего заданного порога.

**Задача – разбить эти точки на два кластера**



# Кластеризация медианным сечением $O(n)$

Дано:  $n$  точек  $x_i$  в  $M$ -мерном пространстве, которые нужно разбить на  $k$  кластеров.

1. Все точки относим к одному кластеру.
2. Просматриваем все кластеры и определяем для каждого из них номер координаты  $m$  в  $M$ -мерном пространстве, для которого наблюдается наибольшая дисперсия (изменчивость) значений точек  $x$ , принадлежащих данному кластеру. В итоге определяем номер кластера  $j$  и номер координаты  $m$ , для которых эта дисперсия максимальна.
3. Сортируем по возрастанию или убыванию значения координаты  $m$  всех точки кластера  $j$ . Находим медиану  $med$  этих значений.
4. Разбиваем кластер  $j$  на два кластера так, что в один кластер помещаются точки кластера  $j$ , у которых значение координаты  $m$  меньше или равно  $med$ , а в другой – остальные.

Шаги 2-4 повторяются, пока число кластеров меньше  $k$ .