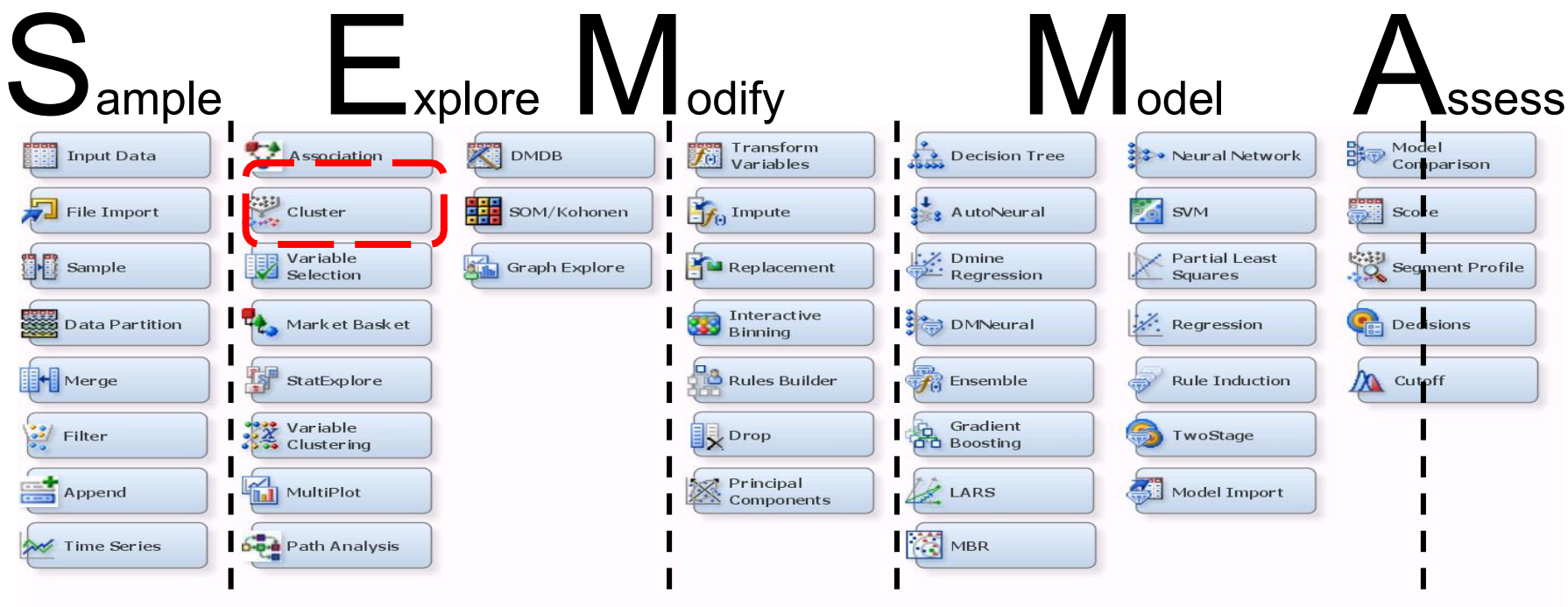
SAS ENTERPRISE MINER

КЛАСТЕРИЗАЦИЯ

КОНЦЕПЦИЯ SEMMA



ЧТО ЕСТЬ КЛАСТЕР?

- Кластер: группа «похожих» объектов
  - «похожих» между собой в группе (внутриклассовое расстояние)
  - «не похожих» на объекты других групп
  - Определение неформальное, формализация зависит от метода
- Кластерный анализ
  - Разбиение множество объектов на группы (кластеры)
- Тип моделей:
  - «описательный» (descriptive) Data mining => одна из задач - наглядное представление кластеров
  - «прогнозный» (predictive) Data mining => разбиение на кластеры, а затем «классификация» новых объектов
- Тип обучения:
  - всегда «без учителя» (unsupervised) => тренировочный набор не размечен

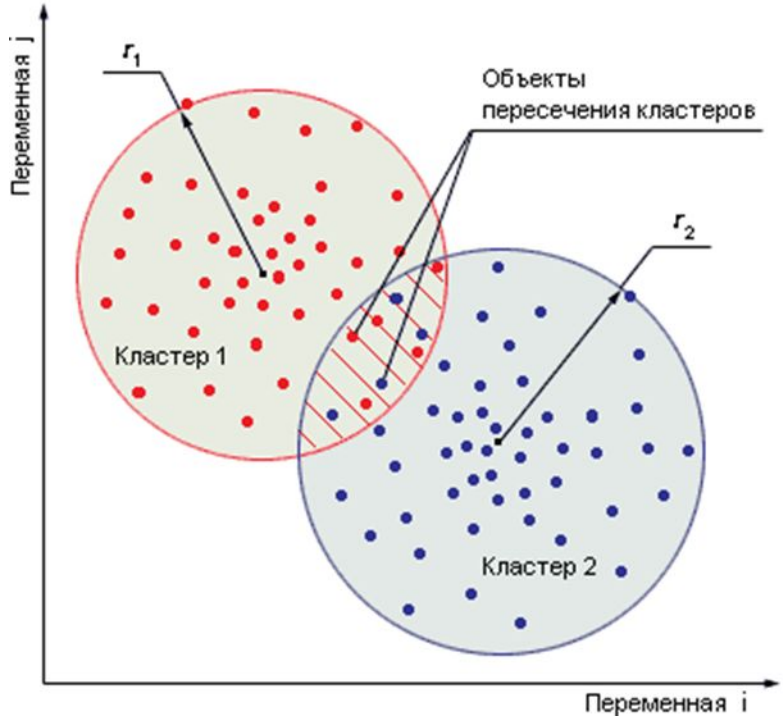
ПРИМЕНЕНИЕ МЕТОДОВ КЛАСТЕРИЗАЦИИ В ЗАДАЧАХ АНАЛИЗА ДАННЫХ

- Кластеризация ради кластеризации:
  - Выявление и описание групп (человек не способен «осознать» более 10 объектов в одной задаче, как обработать выборку с миллионами?)
  - «Сжатие» информации (особенно в обработке мультимедиа)
  - Построение различных поисковых индексов (сравниваем не со всеми, а начинаем с прототипов кластеров)
- Мощнейшее средство предобработки данных:
  - Дискретизация
  - Уменьшение размера выборки (от больших объемов к «реальным»)
  - Обработка пропущенных значений (инициализируем и итерационно «улучшаем» пропуски)
  - Поиск исключений и артефактов (что не в кластере, то под «подозрением»)

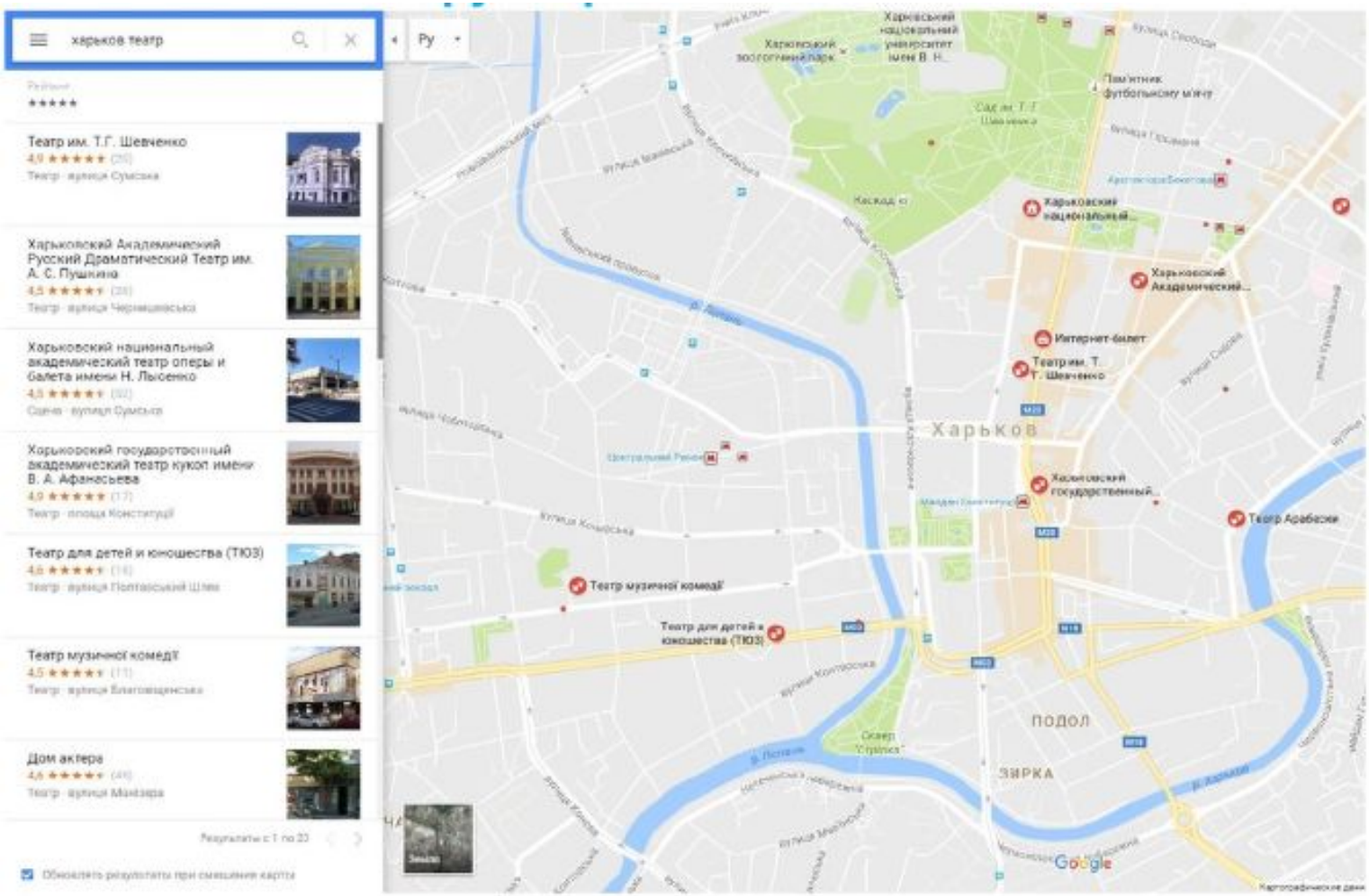
Алгоритмы кластерного анализа

Кластерный анализ включает в себя более 100 различных алгоритмов классификации для организации наблюдаемых данных в наглядные структуры.

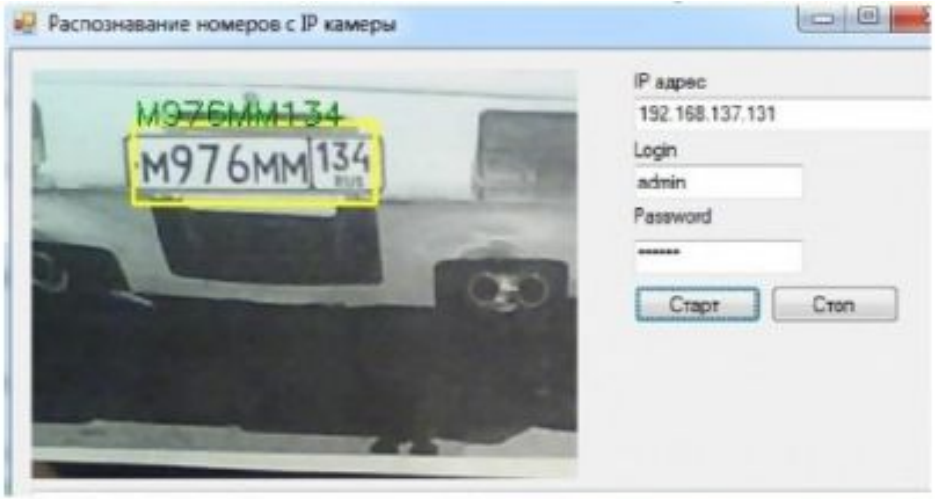
Термин «Кластерный анализ» впервые введен в 1939 году

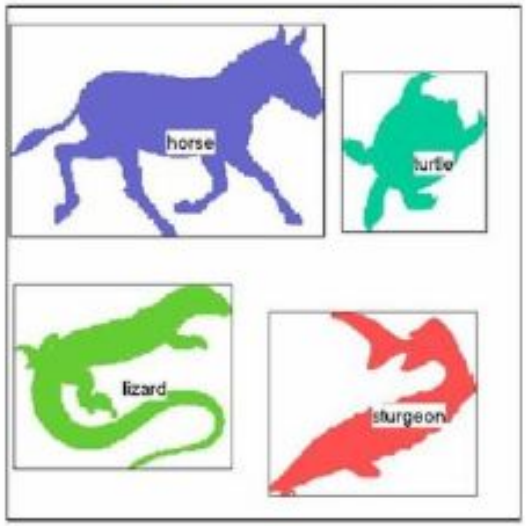


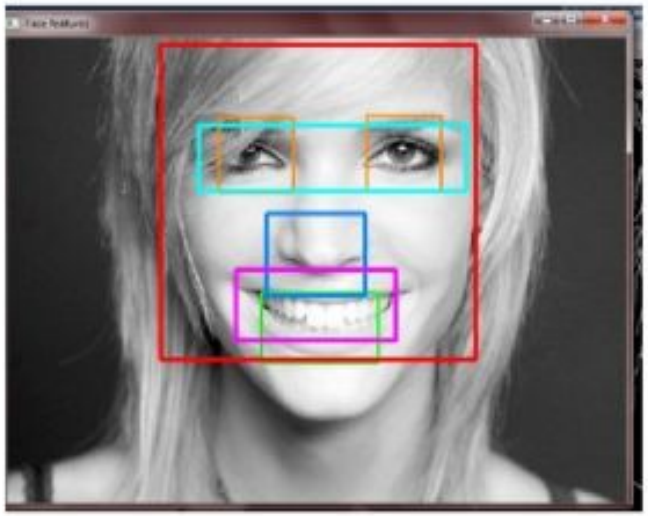
Применение кластеризации – группировка объектов



Применение кластеризации – распознавание образов







Применение кластеризации – классификация результатов поиска














Кластерный анализ – как инструмент интегральной оценки коммерческих банков

Сводные данные для проведения кластерного анализа

Банк	Коэффициент						
	K_1	K_2	K_3	K_4	K_5	K_6	K_7
«Альфабанк Башкортостан»	0,110057	0,071305	1,54347	0,185841	0,386608	0,08	0,383689
«Ашкалар»	0,022707	0,005331	4,259724	0,175054	0,044012	0,01	0,030451
«Башинвест»	0,050503	0,005259	9,602178	0,125822	0,085376	0,01	0,041801
БКС	0,047257	0,006844	6,904736	0,159047	0,0666	0,01	0,043032
«Башпром»	0,013909	0,009811	1,417634	0,036999	0,273793	0,03	0,26517
«Инвесткапитал»	0,591717	0,042821	13,81843	0,18709	0,385538	0,05	0,228879
ПТВ	0,144982	0,030409	4,767701	0,258902	0,177298	0,04	0,117455
РБР	0,038988	0,00878	4,440531	0,121867	0,155415	0,01	0,072046
СИБ	0,051182	0,005921	8,644434	0,123503	0,07047	0,01	0,047941
СКБ	0,06516	0,02003	3,253155	0,15406	0,211653	0,02	0,130014
УК	0,062037	0,01248	4,970892	0,145344	0,096735	0,02	0,085866

Кластерный анализ – как инструмент интегральной оценки коммерческих банков

Характеристика коэффициентов

Коэффициент	Обозначение	Алгоритм расчёта	Содержание показателя
Норма прибыли на капитал	K_1	$\frac{Прибыль}{Активы} \times \frac{Активы}{Капитал}$: $K_1 = K_2 \times K_3$	Характеризует прибыль, приходящуюся на 1 рубль акционерного или уставного капитала
Прибыль активов	K_2	$\frac{Процентные\ доходы}{Активы} \times \frac{Активы}{Доходы}$	Характеризует рентабельность активных операций и оценивает величину прибыли на 1 рубль активов
Достаточность капитала	K_3	$\frac{Активы}{Капитал}$	Показатель надёжности вложений
Доходность активов	K_4	$\frac{Доходы}{Активы} = \frac{Процентные\ доходы}{Активы} + \frac{Непроцентные\ доходы}{Активы}$	Характеризует эффективность размещения активов, т. е. возможность создавать доход
Доля прибыли в доходах	K_5	$\frac{Прибыль}{Доходы} = \frac{Доходы - расходы - налоги}{Доходы}$	Позволяет оценить затратность банковских операций с перераспределением затрат. не связанных с основной банковской деятельностью и существующими обязательствами банка
Рентабельность активов	K_6	$\frac{Прибыль}{Активы,\ приносящие\ доход}$	Позволяет определить основные направления работы банка по улучшению рентабельности активных операций

Кластерный анализ – как инструмент интегральной оценки коммерческих банков

В основу группировки коэффициентов положена функция Евклидова расстояния вида

$$p(x_i x_j) = \sqrt{\sum_{i=1}^{k} (x_{il} - x_{jl})^2},$$

где x_{ij} — измерения ij-объекта (банка); k — количество объектов (банков). Тогда для первого приближения минимальное отклонение показателей эффективности между банками составит:

$$p1,1 = 0;$$

$$p7,8 = \sqrt{\begin{array}{l}(0,144982 - 0,038988)^2 + \\ (0,030409 - 0,00878)^2 + \\ +(4,767701 - 4,440531)^2 + \\ +(0,258902 - 0,121867)^2 + \\ +(0,177298 - 0,155415)^2 + \\ +(0,04 - 0,01)^2 + (0,117455 - 0,072046)^2\end{array}} = 0,375449;$$

$$p_{\text{мин}} = p7,8 = 0,375\,449.$$

Кластерный анализ – как инструмент интегральной оценки коммерческих банков

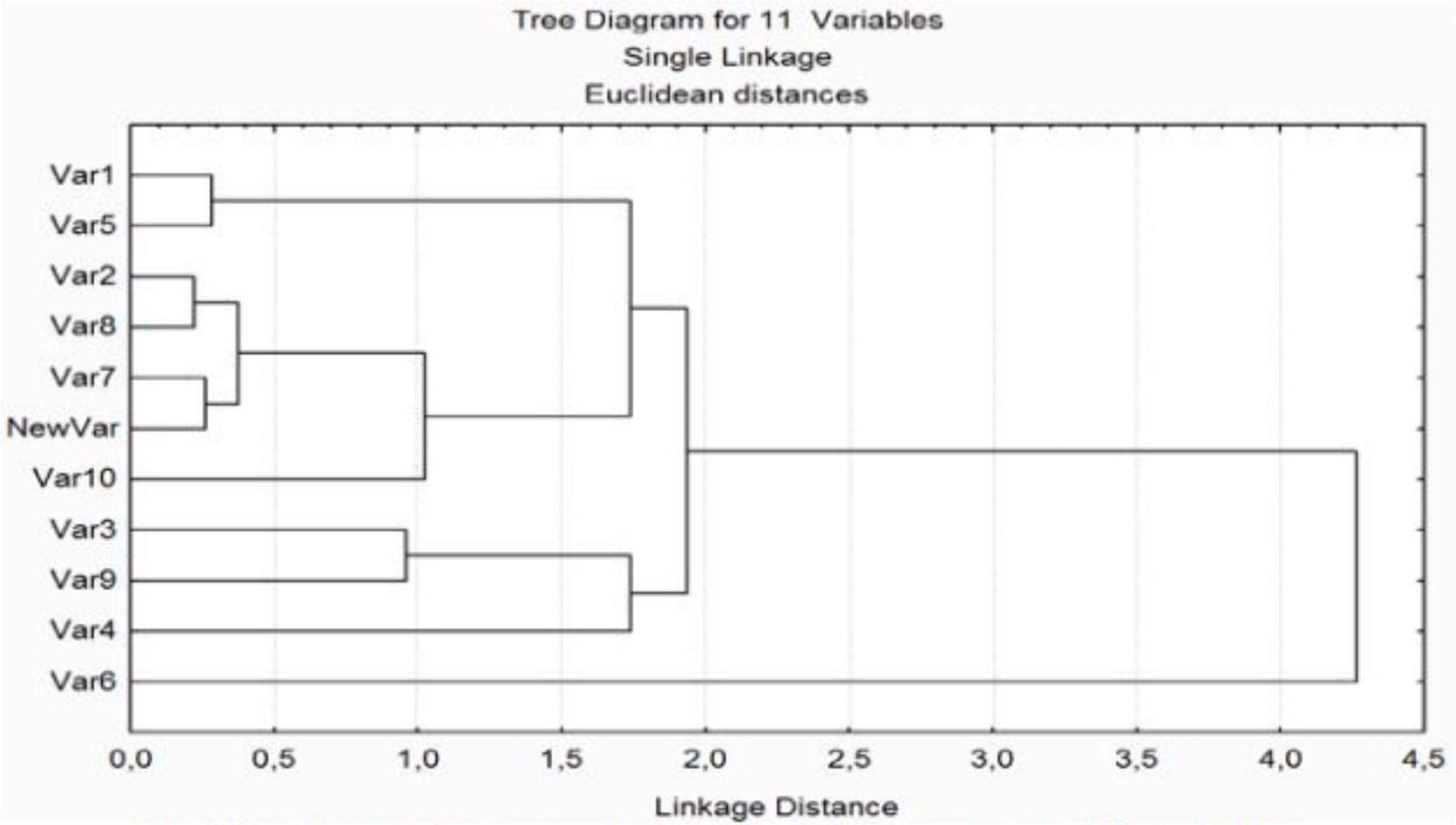


Рис. 3. Группировка банков региона на кластеры по показателям эффективности

Кластерный анализ – как инструмент интегральной оценки коммерческих банков

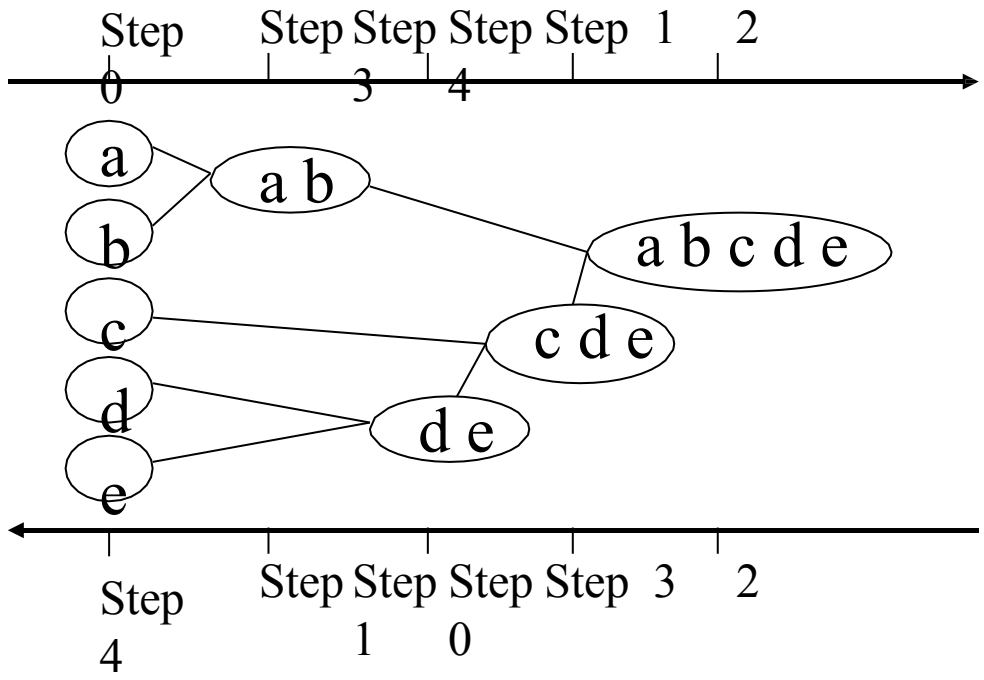
Результаты кластерного анализа финансовых результатов деятельности банков на 1 января 2007 г.

Кластер, № (количество банков)	Банки	Ликвидные активы	Прибыль	Уставный Капитал	Рентабельность ликвидных активов	Прибыль на уставный капитал	Ликвидные активы к уставному капиталу	Норма прибыли на капитал	Прибыль активов	Достаточность капитала	Доходность активов	Доля прибыли в доходах	Рентабельность активов
1 ($n = 2$)	«АФ-Банк»	557344	43152	183361	0,0774	0,235	3,0395	0,110	0,0713	1,5434	0,1858	0,3866	0,08
	«Башпромбанк»	104977	3107	150000	0,0296	0,020	0,6998	0,013	0,0098	1,4176	0,0370	0,2737	0,03
2 ($n = 2$)	«Ашкадар»	30291	735	13151	0,0242	0,055	2,3033	0,022	0,0053	4,2597	0,1750	0,0440	0,01
	РБР	310967	20669	500000	0,0664	0,041	0,6219	0,038	0,0087	4,4405	0,1218	0,1554	0,01
3 ($n = 3$)	ПТБ	141178	25358	133000	0,1796	0,190	1,0614	0,144	0,0304	4,7677	0,2589	0,1772	0,04
	СКБ	41987	11446	150000	0,2726	0,076	0,2799	0,065	0,0200	3,2531	0,1540	0,2116	0,02
	УК	103800	9791	114900	0,0943	0,085	0,9033	0,062	0,0124	4,9708	0,1453	0,0967	0,02
4 ($n = 3$)	СИБ	115134	47731	638254	0,0414	0,074	1,8038	0,051	0,0059	8,6444	0,1235	0,0704	0,01
	«Башинвест-банк»	178841	13765	131370	0,0769	0,104	1,3613	0,050	0,0052	9,6021	0,1258	0,0853	0,01
	БКС	30980	10754	81100	0,0347	0,132	3,8200	0,047	0,0068	6,9047	0,1590	0,0666	0,01
5 ($n = 1$)	«Инвесткапи-талбанк»	516173	183466	150616	0,3554	1,218	3,4270	0,591	0,0428	13,818	0,1870	0,3855	0,05

КАЧЕСТВО КЛАСТЕРИЗАЦИИ

- Хороший метод кластеризации находит кластеры
  - с высоким «внутриклассовым» сходством объектов
  - и низким «межклассовым» сходством объектов
- Оценка качества кластеризации (нет понятия «точность»)
  - необходима, так как влияет на выбор параметров метода
  - определяется либо экспертом – субъективная величина
  - либо «перекрестной» проверкой целевой функции кластеризации
- Качество кластеризации зависит:
  - от метода кластеризации
  - от меры сходства (или расстояния)

ИЕРАРХИЧЕСКАЯ КЛАСТЕРИЗАЦИЯ

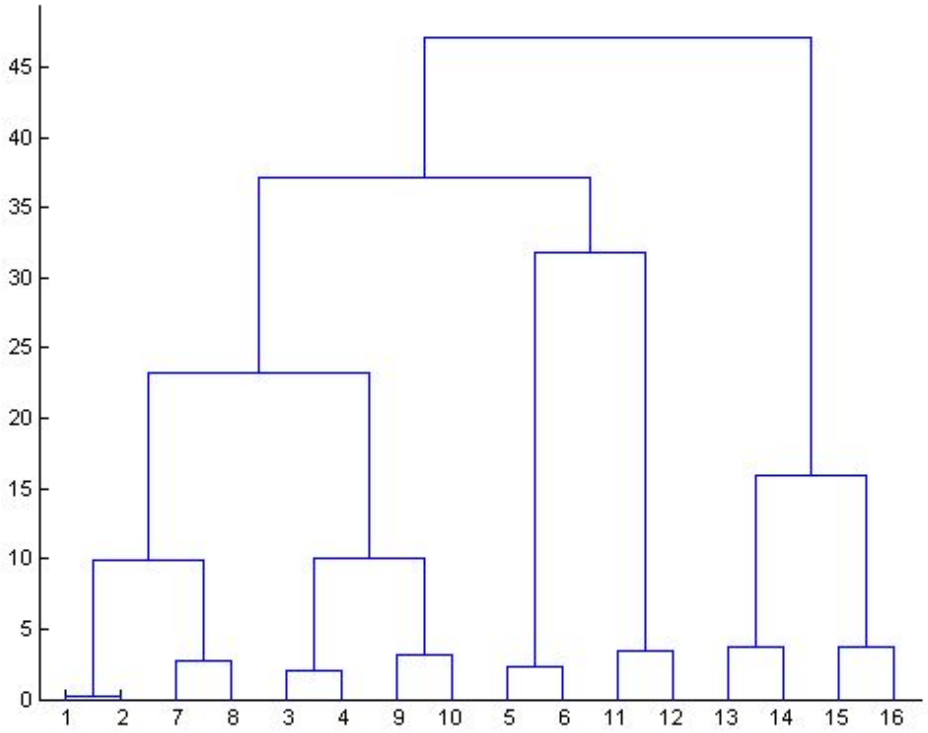


восходящая
agglomerative

нисходящая
divisive

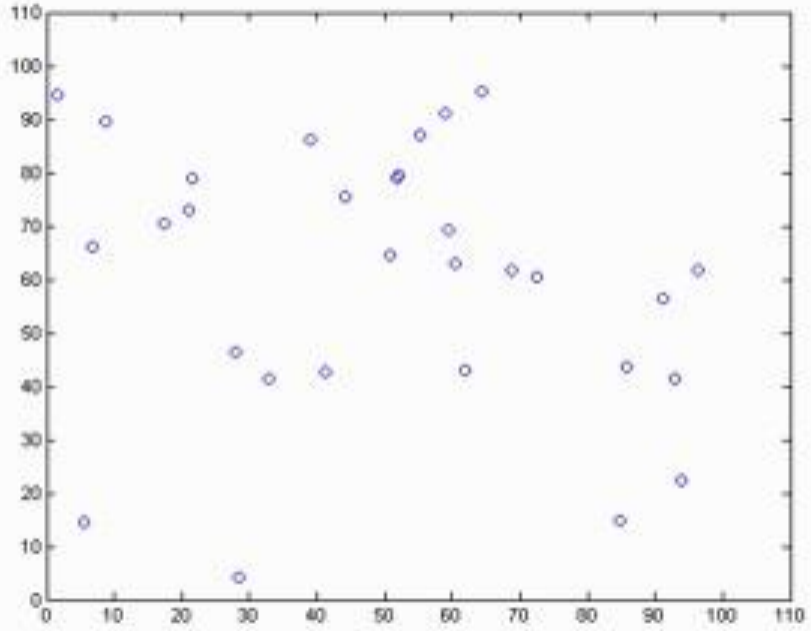
- Используется только матрица сходства (различия) и не требуется дополнительных параметров (например, числа кластеров)
- «Пошаговое» объединение ближайших кластеров (восходящая) или разбиение наиболее удаленных (нисходящая)

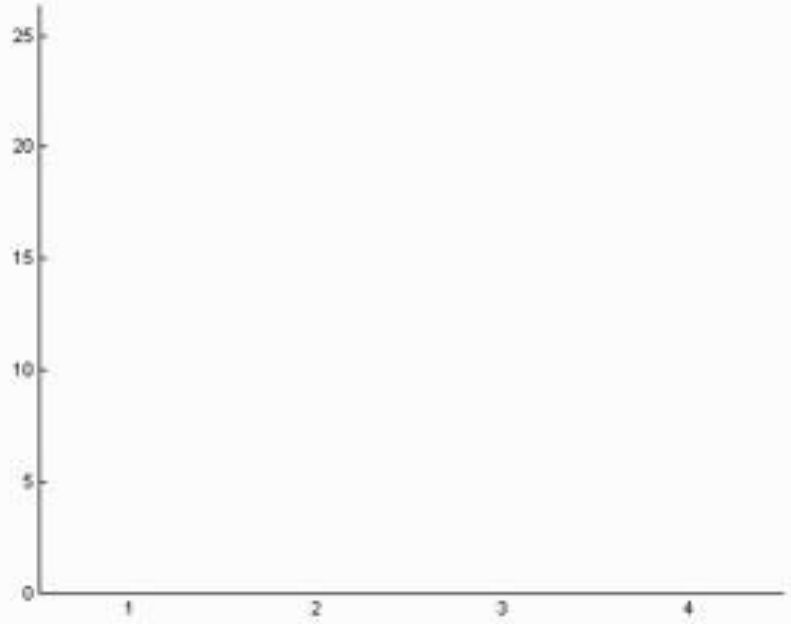
ПРЕДСТАВЛЕНИЕ ИЕРАРХИЧЕСКИХ КЛАСТЕРОВ - ДЕНДРОГРАММА



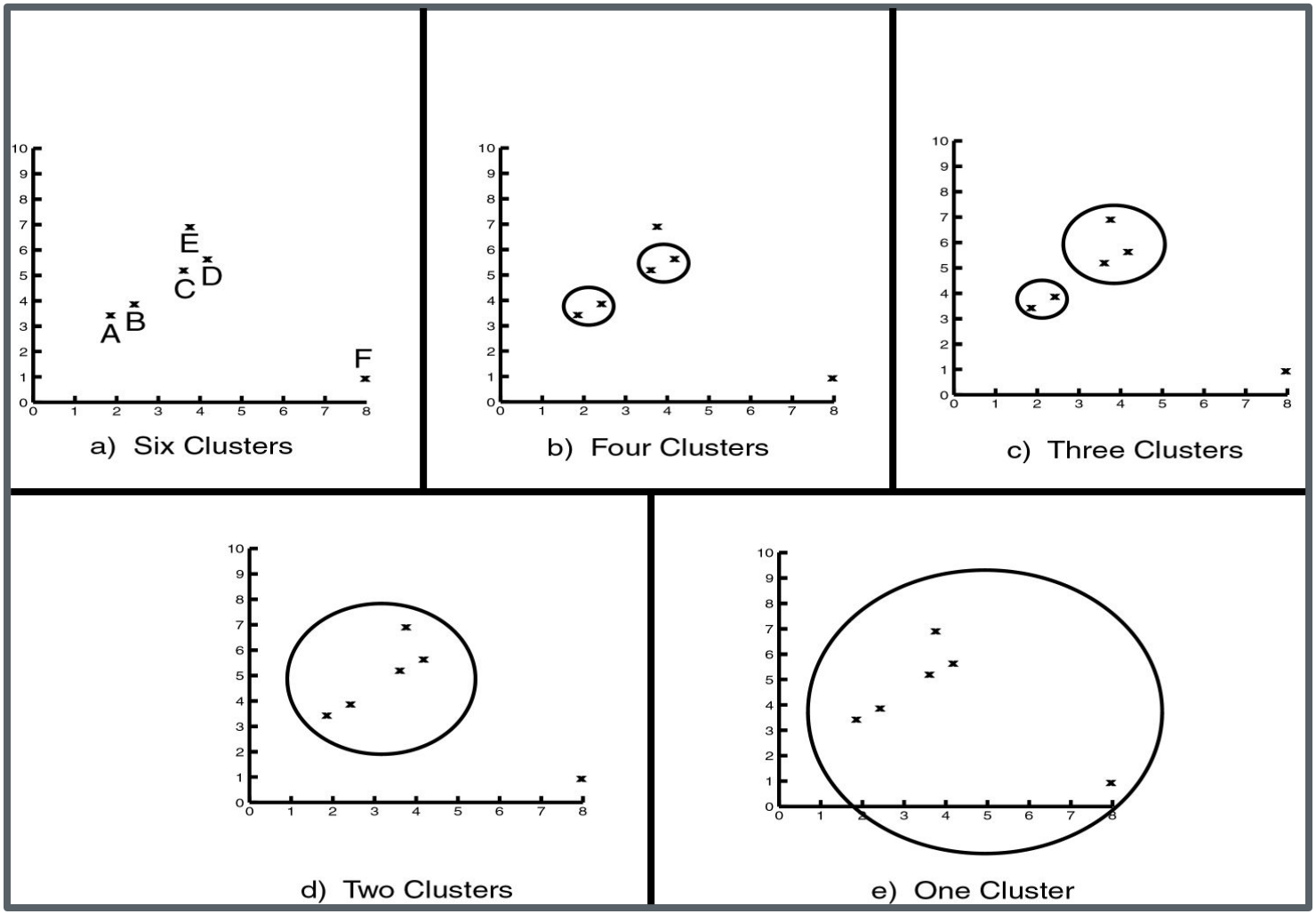
- бинарное дерево, описывающее все шаги разбиения
- Корень – общий кластер, листья - элементы
- «Высота» ветвей (до пересечения) – порог расстояния «склейки» («разделения»)
- Результат кластеризации – «срез» дендрограммы

ИЕРАРХИЧЕСКАЯ КЛАСТЕРИЗАЦИЯ - DEMO



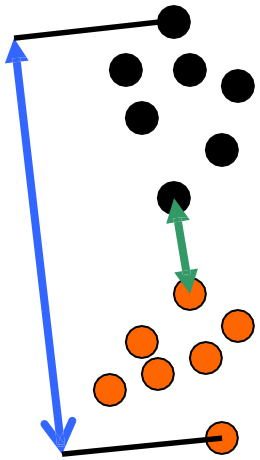


УРОВНИ КЛАСТЕРИЗАЦИИ



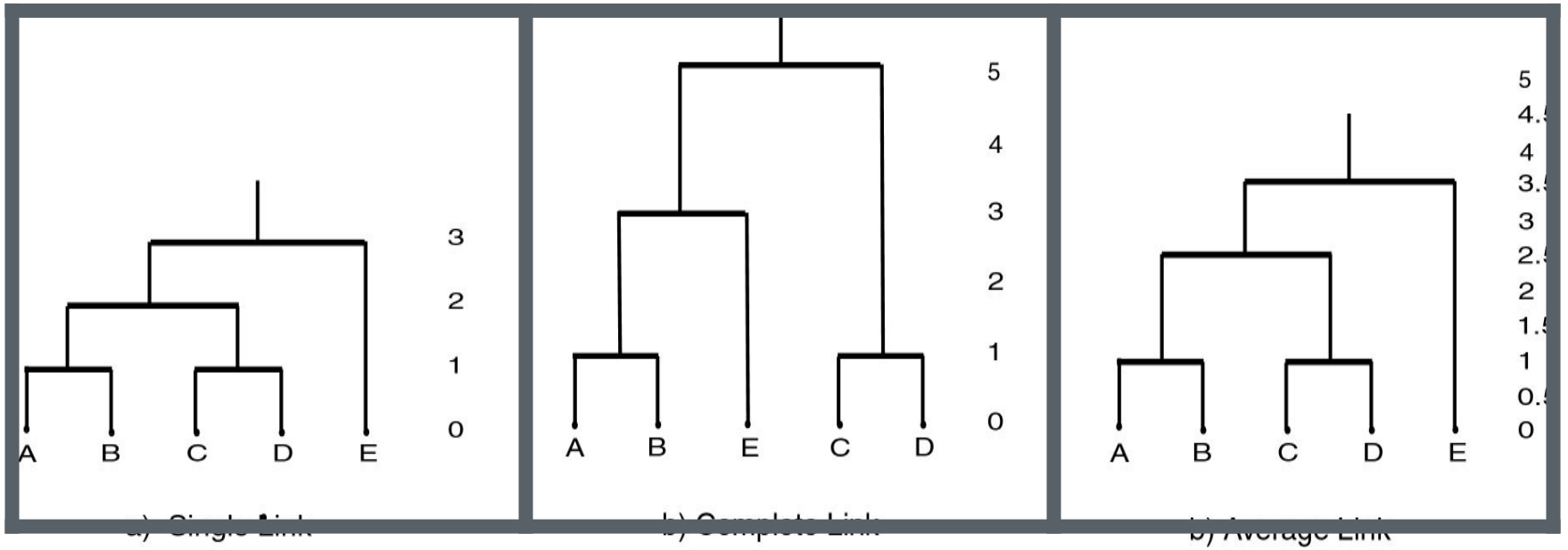
ОЦЕНКА БЛИЗОСТИ КЛАСТЕРОВ

- Расчет расстояния на основе попарных расстояний между элементами различных кластеров:
  - **Полное связывание**: наибольшее попарное расстояние. Дает компактные сферические кластеры.
  - **Среднее связывание**: усредненное попарное расстояние. Редко используется.
  - **Единственное связывание**: наименьшее попарное расстояние. Дает «растянутые» кластеры сложной формы.
  - **Центроидное связывание***:* расстояние между центрами (мат. ожидание) кластеров.
  - Другие методы (например **метод Ward'a** – минимизирует внутрикластерные дисперсии или другую целевую функцию)



Пример

	A	B	C	D	E
A	0	1	2	2	3
B	1	0	2	4	3
C	2	2	0	1	5
D	2	4	1	0	3
E	3	3	5	3	0



КЛАСТЕРИЗАЦИЯ НА ОСНОВЕ СТРОГОЙ ГРУППИРОВКИ (PARTITIONING):

- Основная задача:
  - Найти такое разбиение *C* исходного множества *X* из *N* объектов на *K* непересекающихся подмножеств C_k, покрывающих *X*, чтобы внутриклассовое расстояние было минимальным:

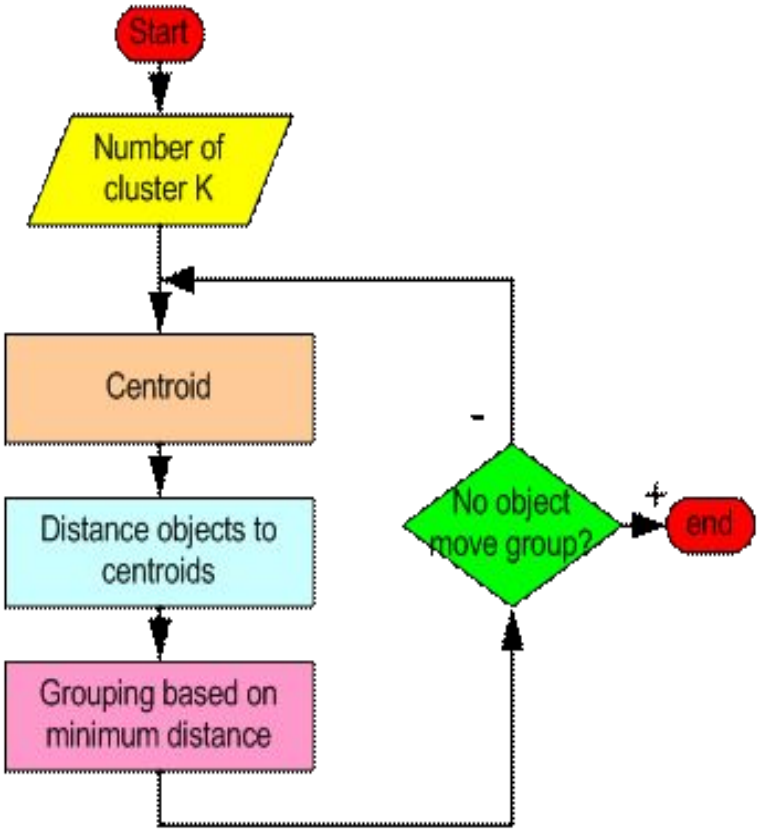
$$\min_{C_i \cap C_j = \varnothing, \bigcup C_i = X} \sum_{i=1}^{K} \sum_{x \in C_i} \sum_{x' \in C_i} d(x, x')$$

- Точное решение – перебор с отсечением
  - метод «ветвей и границ», но число комбинаций неприемлемо даже для 100 объектов:

$$S(N, K) = \frac{1}{K!} \sum_{i=1}^{K} (-1)^{K-i} \binom{K}{i} i^N$$

- Эвристические методы:
  - K-means (прототип кластера – мат. ожидание *m*), K-medoids (прототип кластера – средний элемент)
  - ищется локальный минимум $\min_{C_i \cap C_j = \varnothing, \bigcup C_i = X} \sum_{i=1}^{K} \sum_{x \in C_i} d(m_i, x)$

МЕТОД K-MEANS В ENTERPRISE MINER



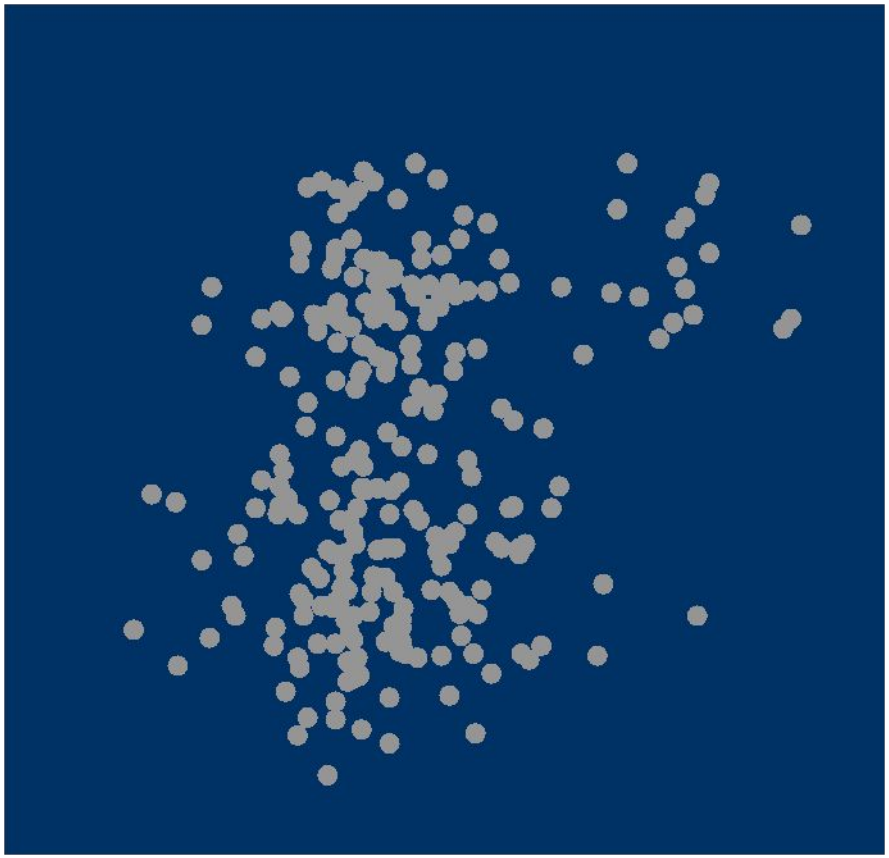
- Шаг 0. Инициализация:
  - произвольное разбиение на заданное число кластеров K (где значение K выбирается по CCC на основе иерархической кластеризации) по
- Шаг 1. Поиск центров: $m^i = \sum_{x \in C_i} x \Big/ \| C_i \|$
  - Для всех K кластеров
- Шаг 2. Расчет расстояний до центров:
  - Для всех N объектов и K кластеров $d(m^i, x) = \sum_{x \in C} x \Big/ \| C_i \|$
- Шаг 3. Выбор ближайшего кластера:

$$x \in C_i \Leftrightarrow i = \min_j d(m_j, x)$$

- Если были перестановки, то Шаг 1.

АЛГОРИТМ КЛАСТЕРИЗАЦИИ K-MEANS

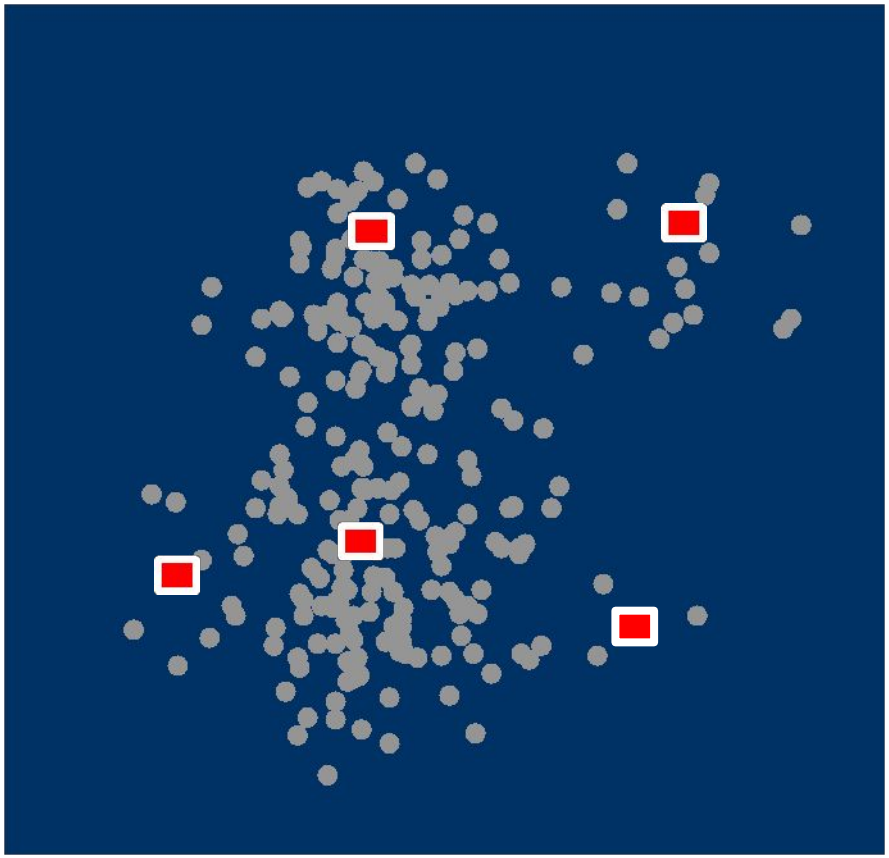
Training Data



1. **Выбор переменных.**
2. Выбор *k* центров кластеров.
3. Выбор ближайшего центра для каждого примера.
4. Пересчет центров.
5. Переход на шаг 3 пока процесс не сошелся

АЛГОРИТМ КЛАСТЕРИЗАЦИИ K-MEANS

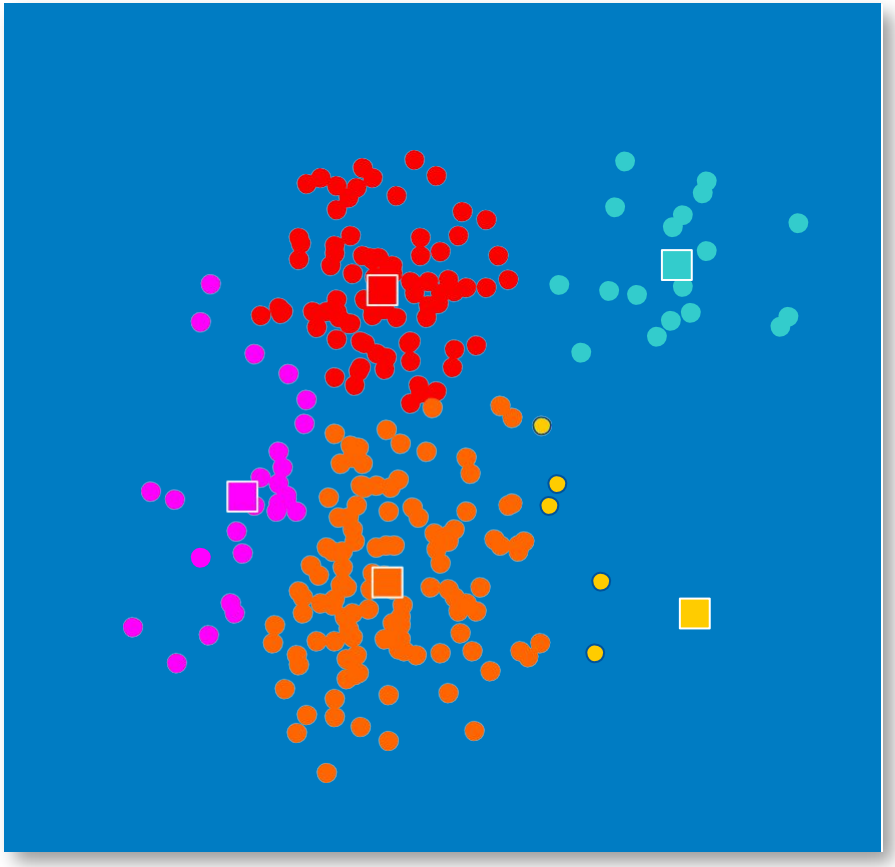
Training Data



1. Выбор переменных.
2. **Выбор *k* центров кластеров.**
3. Выбор ближайшего центра для каждого примера.
4. Пересчет центров.
5. Переход на шаг 3 пока процесс не сошелся

АЛГОРИТМ КЛАСТЕРИЗАЦИИ K-MEANS

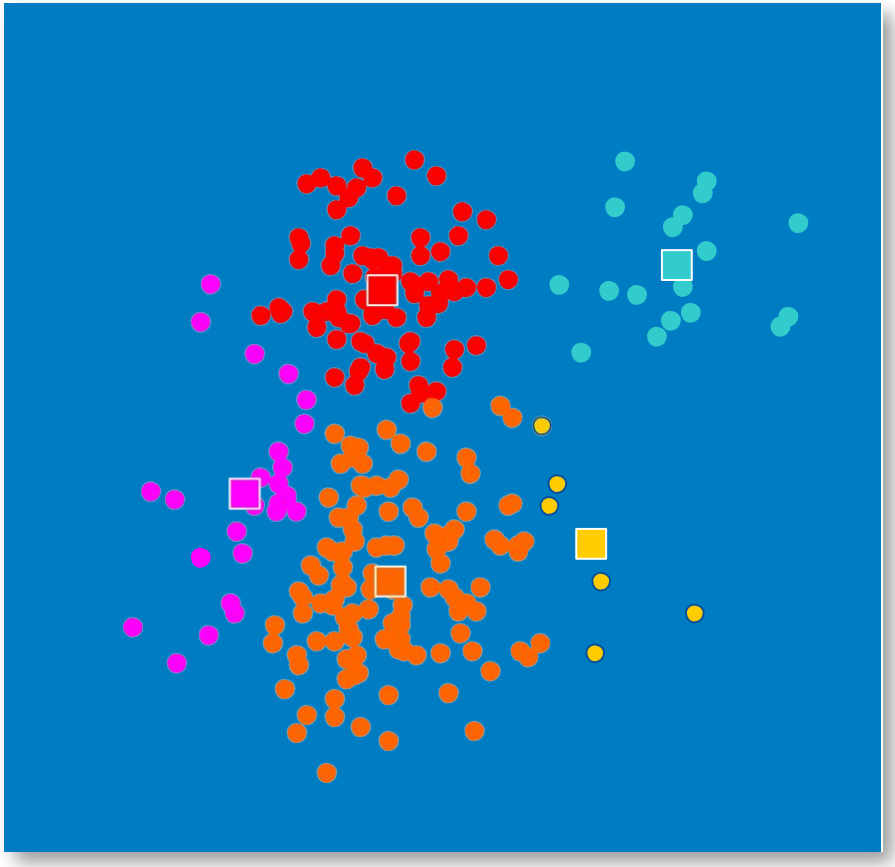
Training Data



1. Выбор переменных.
2. Выбор *k* центров кластеров.
3. **Выбор ближайшего центра для каждого примера.**
4. Пересчет центров.
5. Переход на шаг 3 пока процесс не сошелся

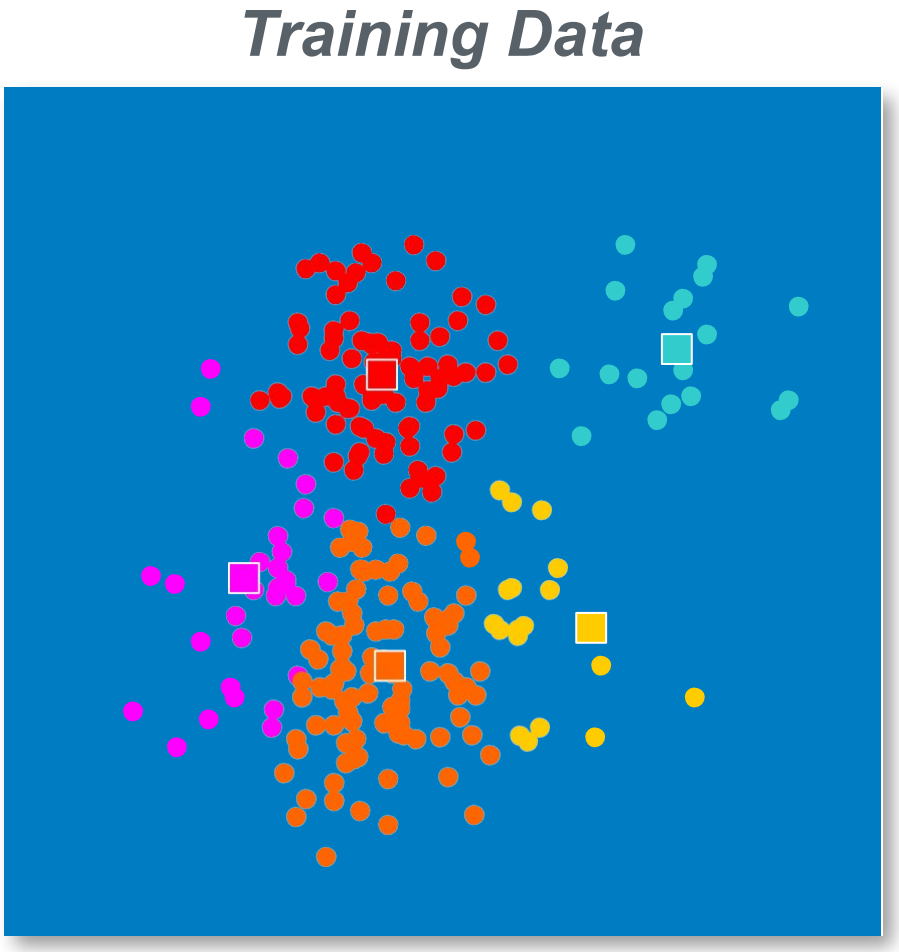
АЛГОРИТМ КЛАСТЕРИЗАЦИИ K-MEANS

Training Data



1. Выбор переменных.
2. Выбор *k* центров кластеров.
3. Выбор ближайшего центра для каждого примера.
4. **Пересчет центров.**
5. Переход на шаг 3 пока процесс не сошелся

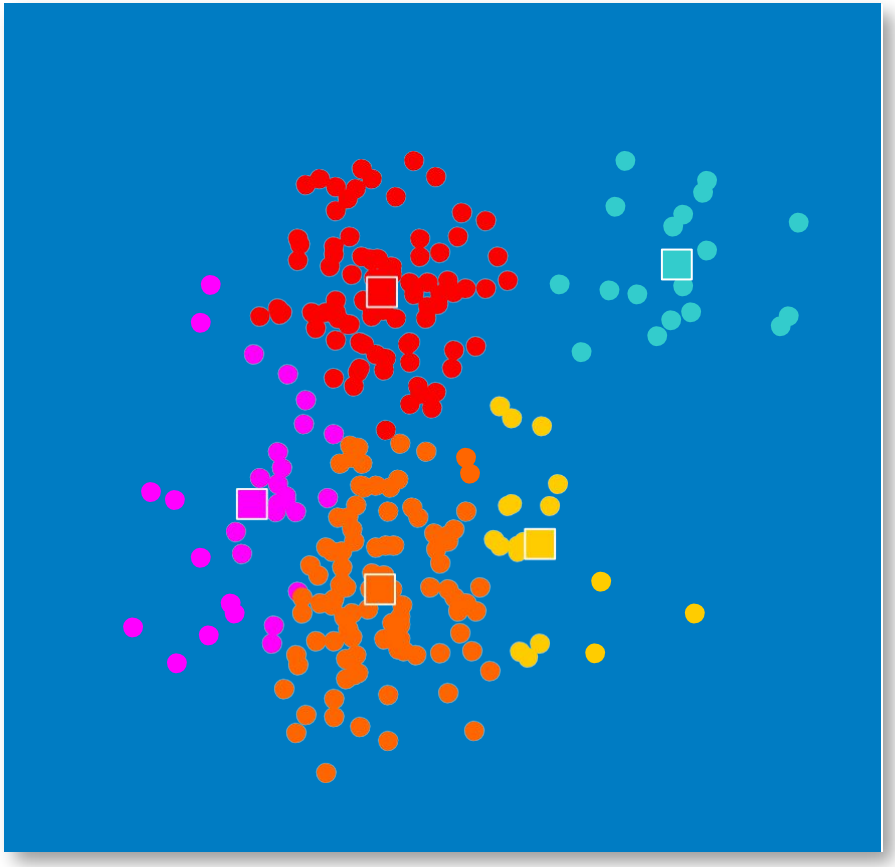
АЛГОРИТМ КЛАСТЕРИЗАЦИИ K-MEANS



1. Выбор переменных.
2. Выбор *k* центров кластеров.
3. **Выбор ближайшего центра для каждого примера.**
4. Пересчет центров.
5. Переход на шаг 3 пока процесс не сошелся

АЛГОРИТМ КЛАСТЕРИЗАЦИИ K-MEANS

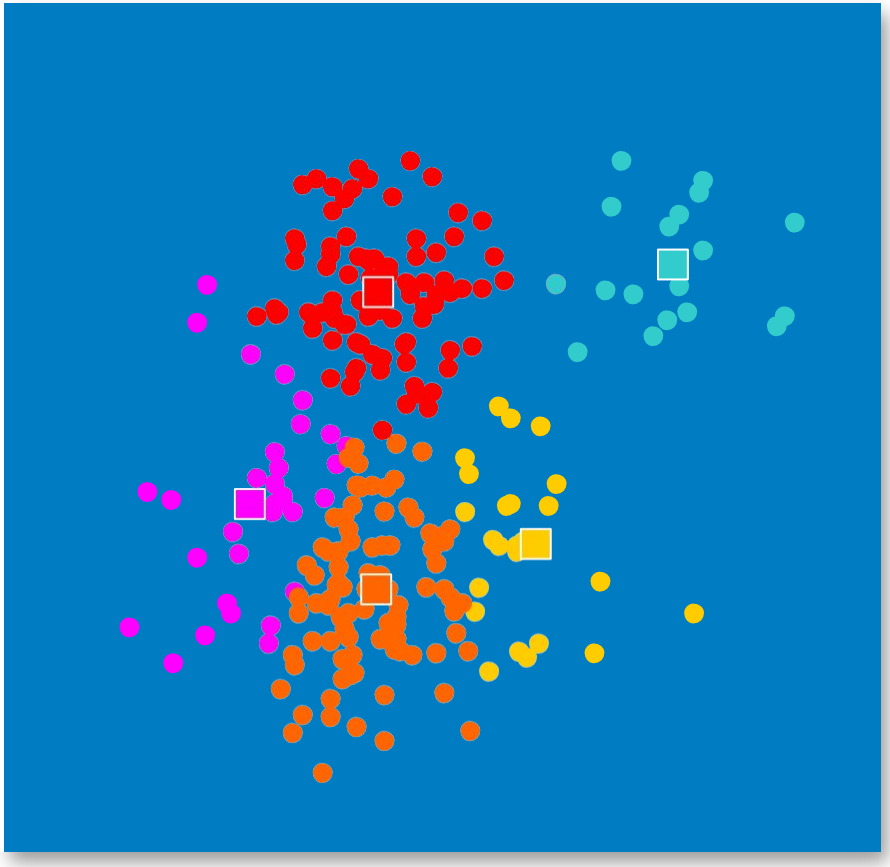
Training Data



1. Выбор переменных.
2. Выбор *k* центров кластеров.
3. Выбор ближайшего центра для каждого примера.
4. **Пересчет центров.**
5. Переход на шаг 3 пока процесс не сошелся

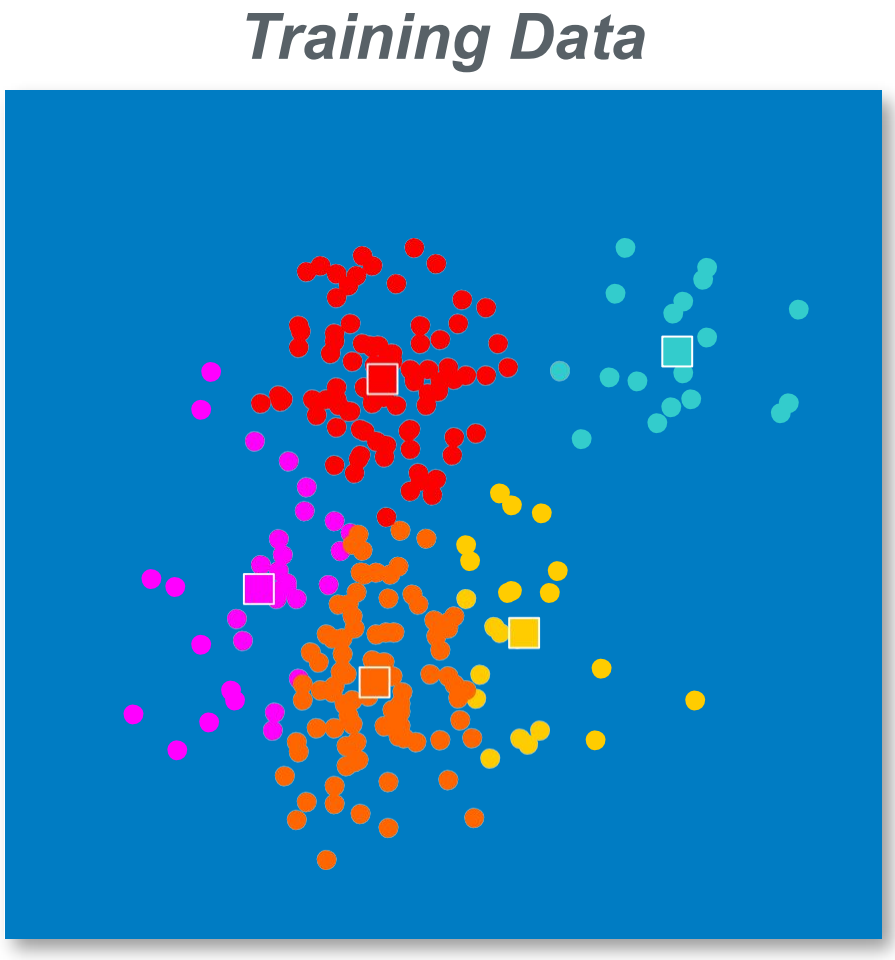
АЛГОРИТМ КЛАСТЕРИЗАЦИИ K-MEANS

Training Data



1. Выбор переменных.
2. Выбор *k* центров кластеров.
3. **Выбор ближайшего центра для каждого примера.**
4. Пересчет центров.
5. Переход на шаг 3 пока процесс не сошелся

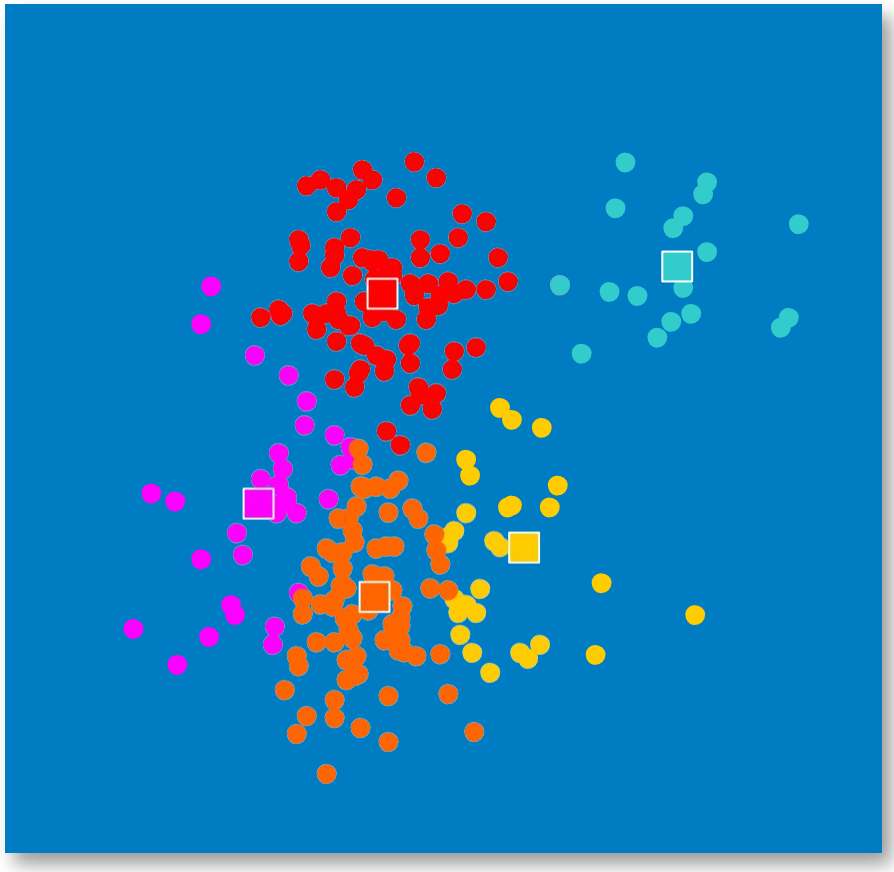
АЛГОРИТМ КЛАСТЕРИЗАЦИИ K-MEANS



1. Выбор переменных.
2. Выбор *k* центров кластеров.
3. Выбор ближайшего центра для каждого примера.
4. **Пересчет центров.**
5. Переход на шаг 3 пока процесс не сошелся

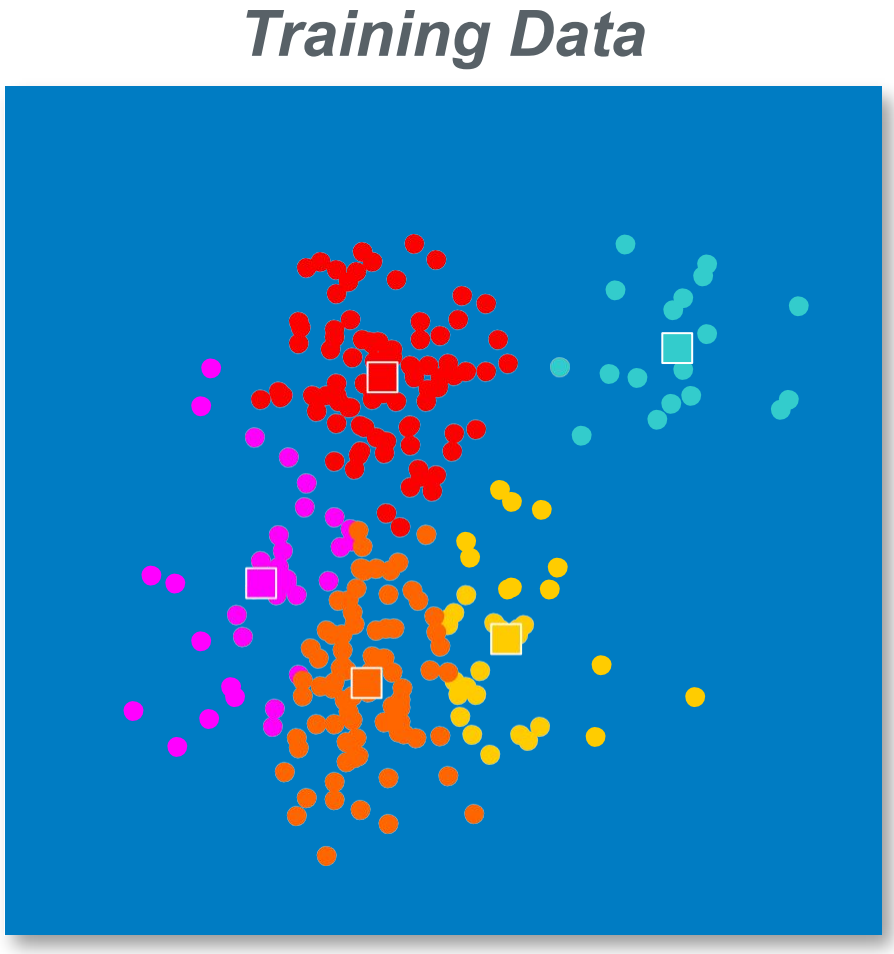
АЛГОРИТМ КЛАСТЕРИЗАЦИИ K-MEANS

Training Data



1. Выбор переменных.
2. Выбор *k* центров кластеров.
3. **Выбор ближайшего центра для каждого примера.**
4. Пересчет центров.
5. Переход на шаг 3 пока процесс не сошелся

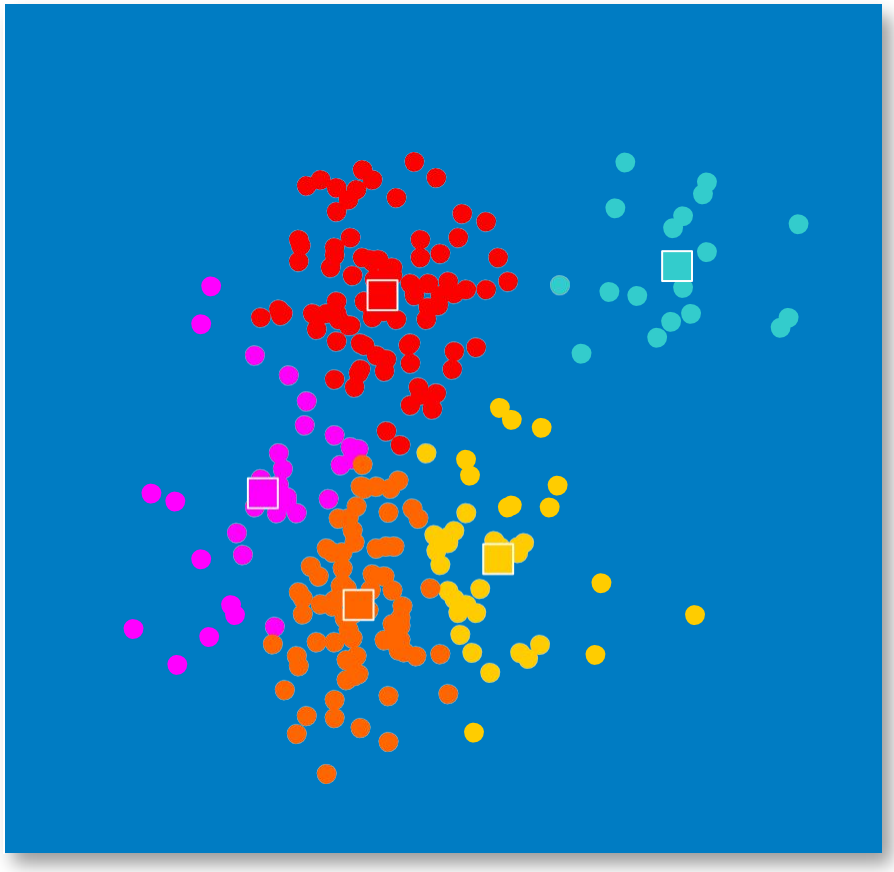
АЛГОРИТМ КЛАСТЕРИЗАЦИИ K-MEANS



1. Выбор переменных.
2. Выбор *k* центров кластеров.
3. Выбор ближайшего центра для каждого примера.
4. **Пересчет центров.**
5. Переход на шаг 3 пока процесс не сошелся

АЛГОРИТМ КЛАСТЕРИЗАЦИИ K-MEANS

Training Data



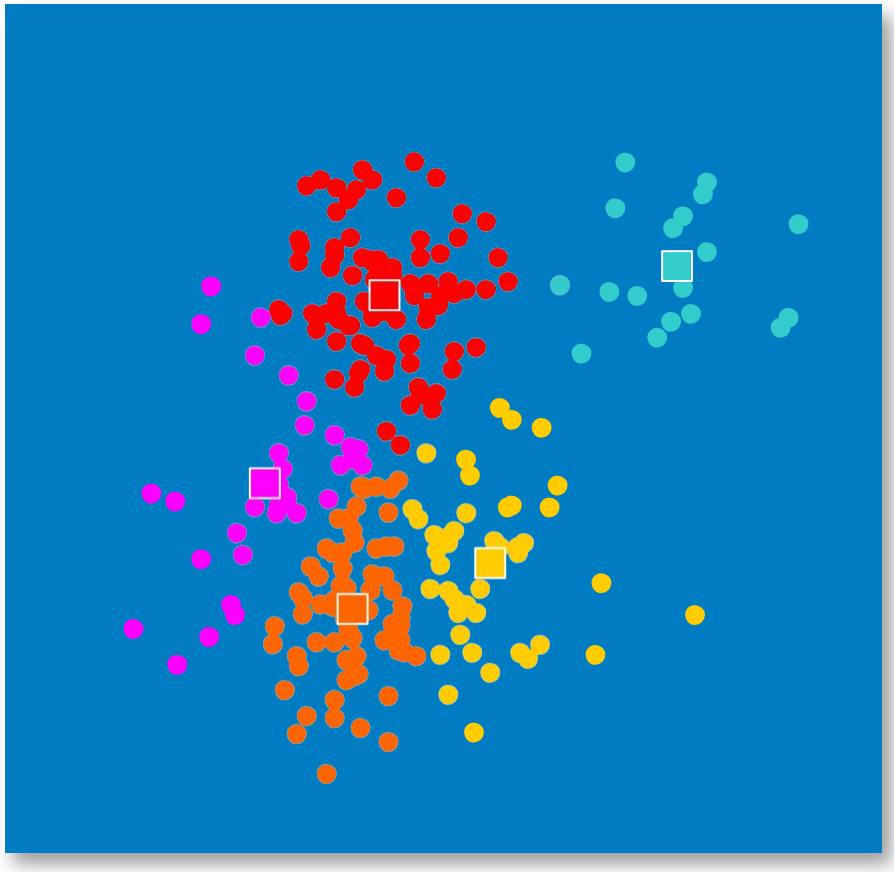
1. Выбор переменных.
2. Выбор *k* центров кластеров.
3. **Выбор ближайшего центра для каждого примера.**
4. Пересчет центров.
5. Переход на шаг 3 пока процесс не сошелся

АЛГОРИТМ КЛАСТЕРИЗАЦИИ K-MEANS

1. Выбор переменных.
2. Выбор *k* центров кластеров.
3. Выбор ближайшего центра для каждого примера.
4. **Пересчет центров.**
5. Переход на шаг 3 пока процесс не сошелся

АЛГОРИТМ КЛАСТЕРИЗАЦИИ K-MEANS

Training Data



1. Выбор переменных.
2. Выбор *k* центров кластеров.
3. Выбор ближайшего центра для каждого примера.
4. **Пересчет центров.**
5. Переход на шаг 3 пока процесс не сошелся

ОПРЕДЕЛЕНИЕ ЧИСЛА КЛАСТЕРОВ

- SAS Cubic Clustering Criterion (CCC) (Sarle, 1983)
- Основная идея: сравнение R^2 (для отображения матрицы данных с помощью индикаторной матрицы в прототипы кластеров) для заданной модели кластеризации с $E(R^2)$ для равномерно распределенного множества прототипов кластеров (как наихудший возможный вариант):

$$CCC = \ln \left[\frac{1 - E(R^2)}{1 - R^2_{clust}} \right] \times K$$

ОПРЕДЕЛЕНИЕ ЧИСЛА КЛАСТЕРОВ

1. Случайно выбираются центры для большого (50 по умолчанию) числа кластеров
2. Все наблюдения объединяются в эти случайные кластеры
3. Решается задач восходящей иерархической кластеризации, на каждом шаге считается CCC
4. По определенным правилам выбирается оптимальное число кластеров:
   - Первый локальный пик …