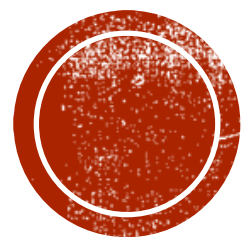


ЯЗЫКИ ПРИКЛАДНОГО ПРОГРАММИРОВАНИЯ

Лекция 3



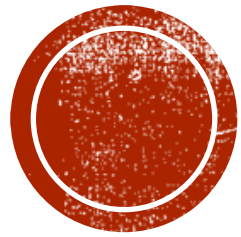
JUPITER NOTEBOOK



JUPYTER NOTEBOOK

- Интерактивный «блокнот» поддерживающий **python** и некоторые другие языки.
- Веб-среда разработки => может запускаться локально или с удаленного сервера
- В основном применяется в **ML, DS** и т.п.
- Установка:
 - `Pip install jupyter`
 - Или скачать и установить **Anaconda** (пакет **python** + набор распространенных сторонних пакетов)
- Запуск:
 - `jupyter notebook`





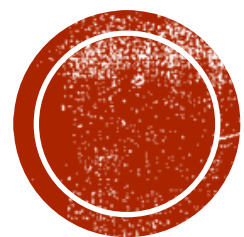
ВИЗУАЛИЗАЦИЯ ДАННЫХ



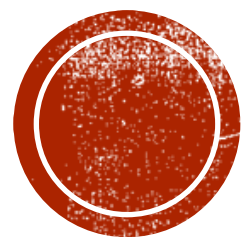
БИБЛИОТЕКИ, ИСПОЛЬЗУЕМЫЕ ДЛЯ ВИЗУАЛИЗАЦИИ ДАННЫХ

- Matplotlib
 - Seaborn
 - Pandas
 - Numpy
-
- Установка пакетов:
 - `pip install Numpy, Pandas, Matplotlib, Seaborn`





ДЕМОНСТРАЦИЯ



DATA SCIENCE



DATA SCIENCE

- **Data Science** - наука о данных. включает в себя все инструменты, методы и технологии, помогающие нам обрабатывать данные и использовать их для нашего блага. Это междисциплинарная смесь статистических выводов, анализа данных, разработки алгоритмов и технологий для решения аналитически сложных задач.



MACHINE LEARNING

- Основные задачи, которые решают алгоритмы машинного обучения — те, которые тяжело/невозможно/нерационально решать непосредственным, “явным” (explicit) программным либо аналитическим способом.
 - Кластеризация
 - Классификация
 - Регрессия
 - Определение аномалий
 - Обнаружение объектов
 - Ранжирование
 - Рекомендация
 - Прогнозирование
 - ...

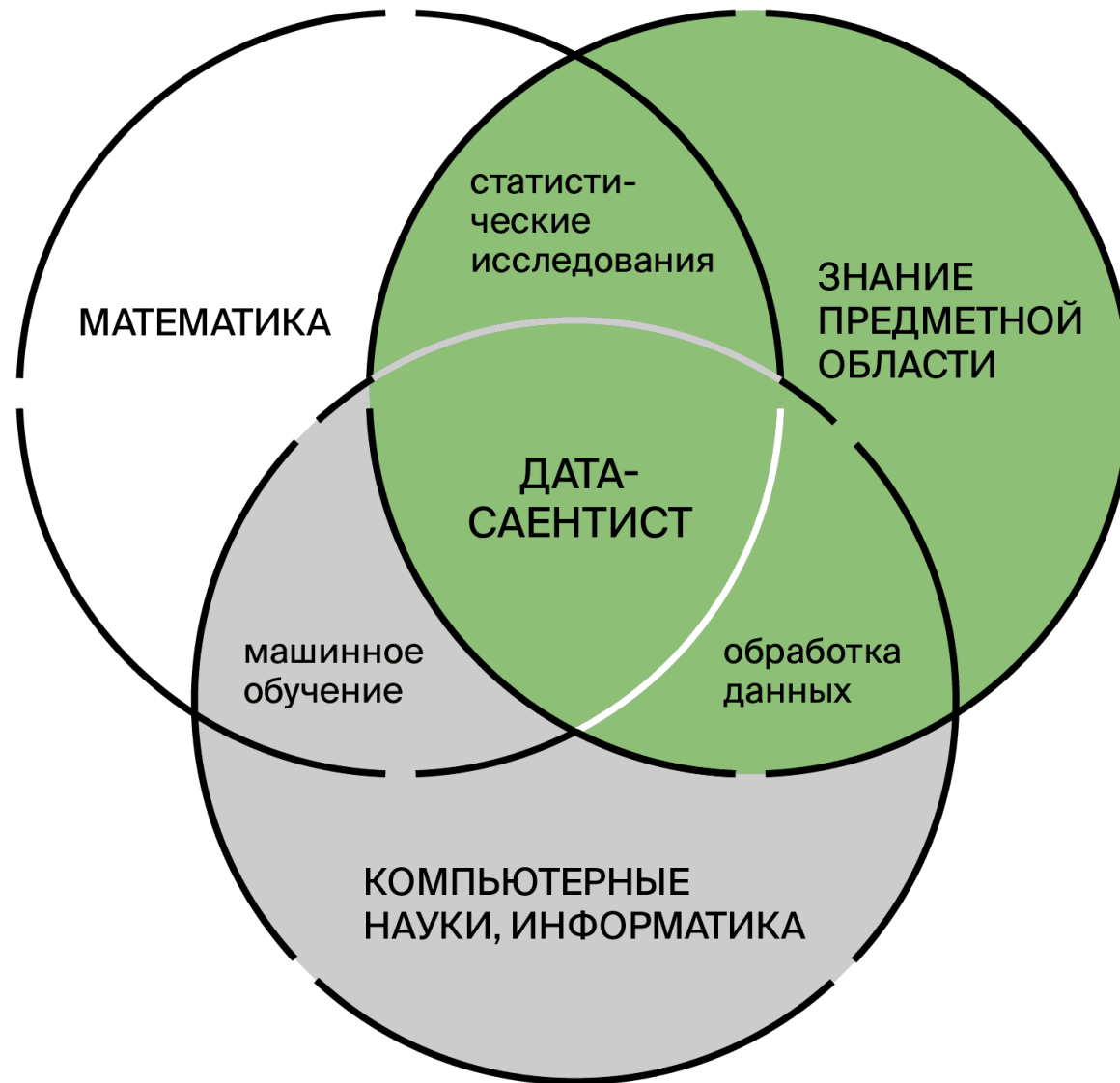


ARTIFICIAL INTELLIGENCE

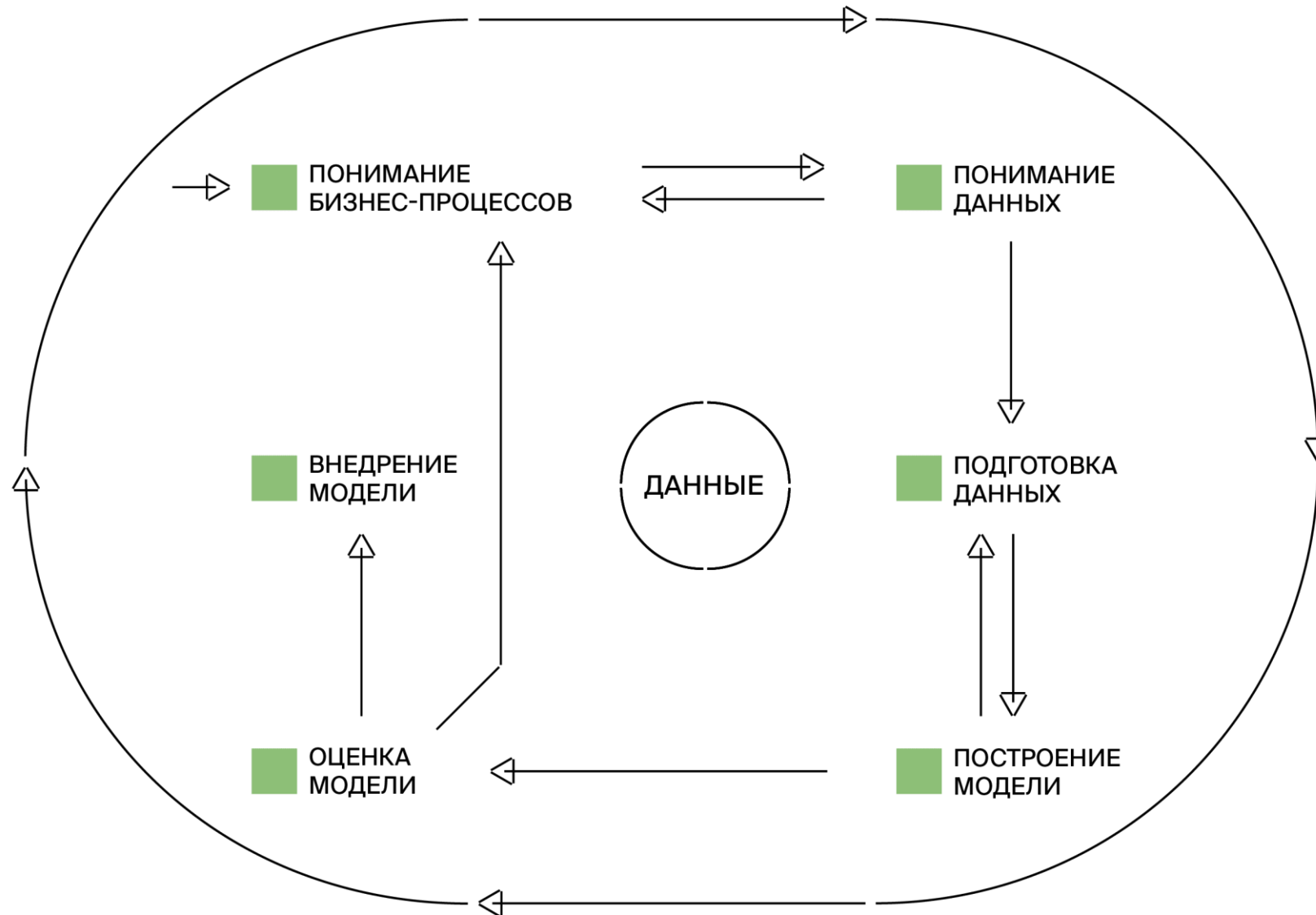
- Искусственный интеллект, ИИ (Artificial Intelligence, AI) — инженерно-математическая дисциплина, занимающаяся созданием программ и устройств, имитирующих когнитивные (интеллектуальные) функции человека, включающие, в том числе, анализ данных и принятие решений.

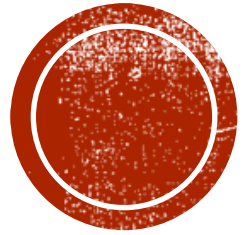


ДАТАСАЕНТИСТ: ЗНАНИЯ И НАВЫКИ



КАК РАБОТАЕТ ДАТАСАЕНТИСТ



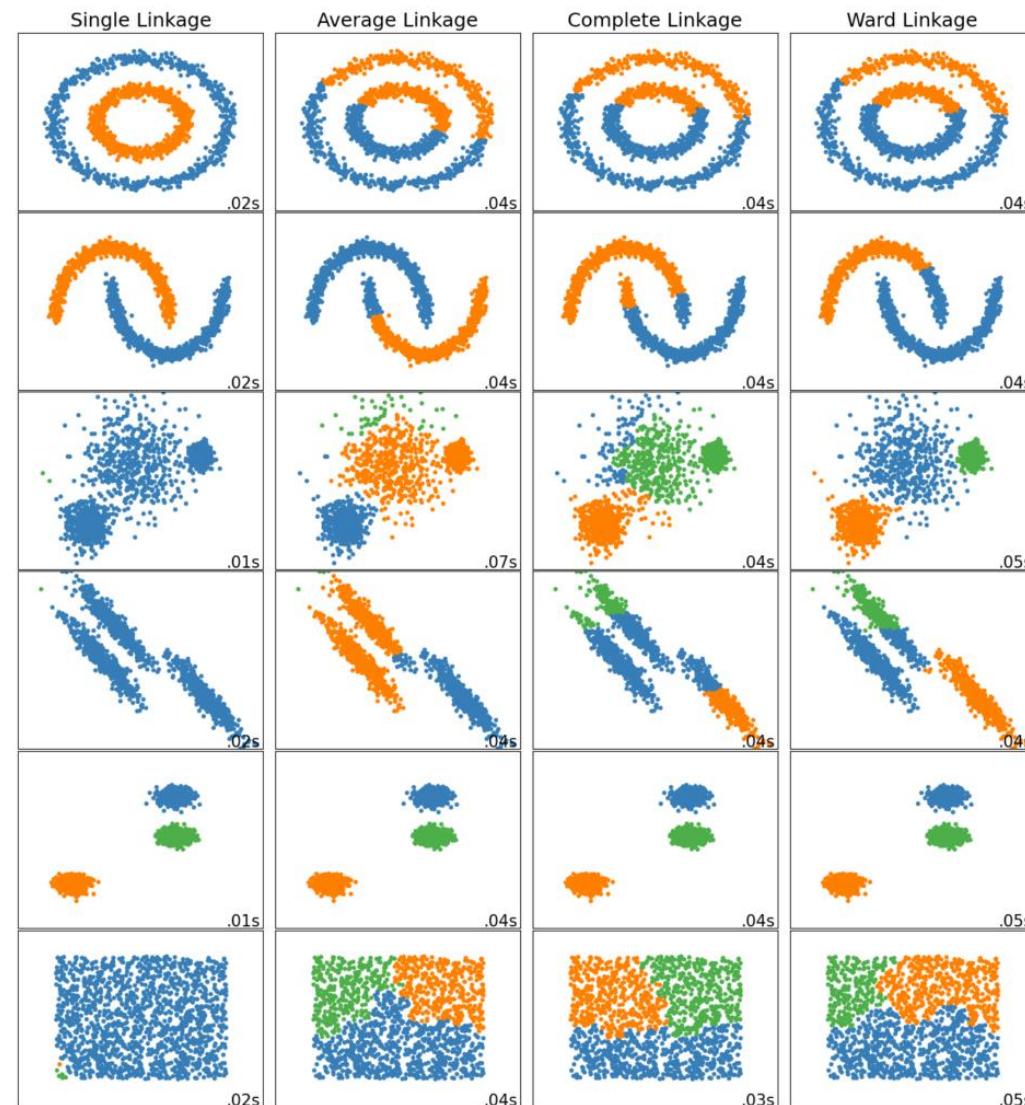


КЛАСТЕРИЗАЦИЯ ДАННЫХ



КЛАСТЕРИЗАЦИЯ

- **Кластеризация (англ. cluster analysis)** — задача группировки множества объектов на подмножества (кластеры) таким образом, чтобы объекты из одного кластера были более похожи друг на друга, чем на объекты из других кластеров по какому-либо критерию.

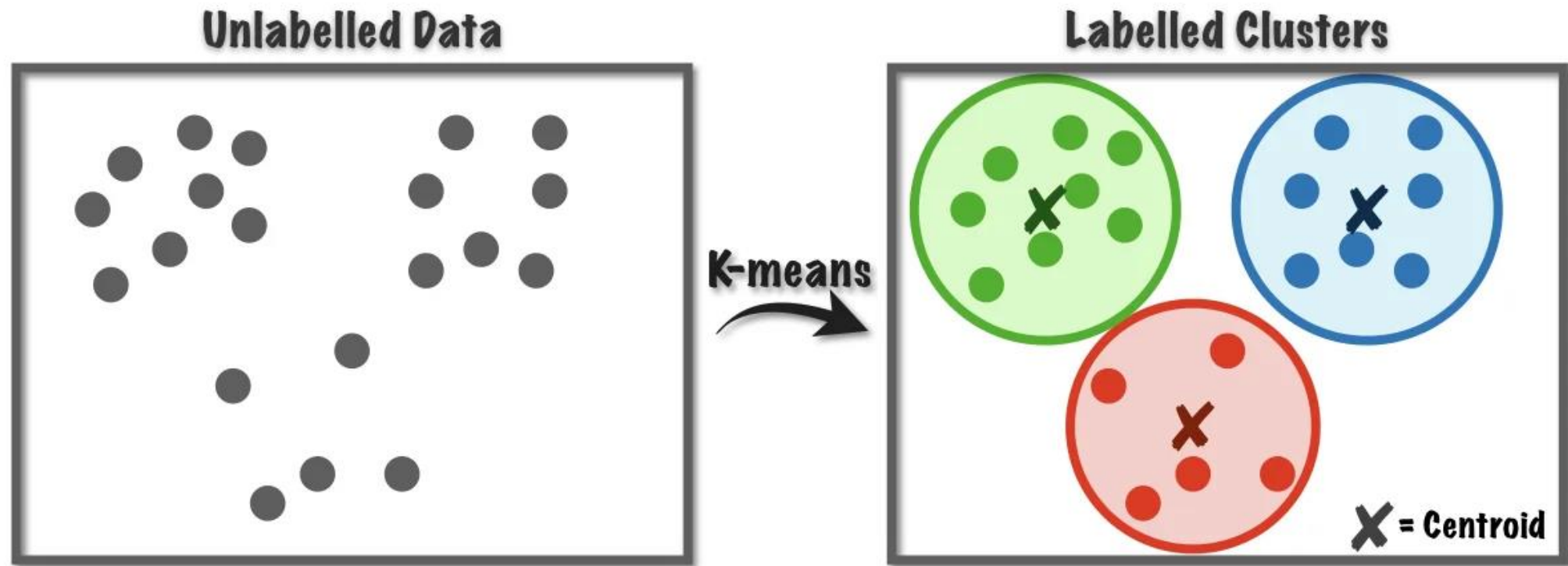


K-MEANS

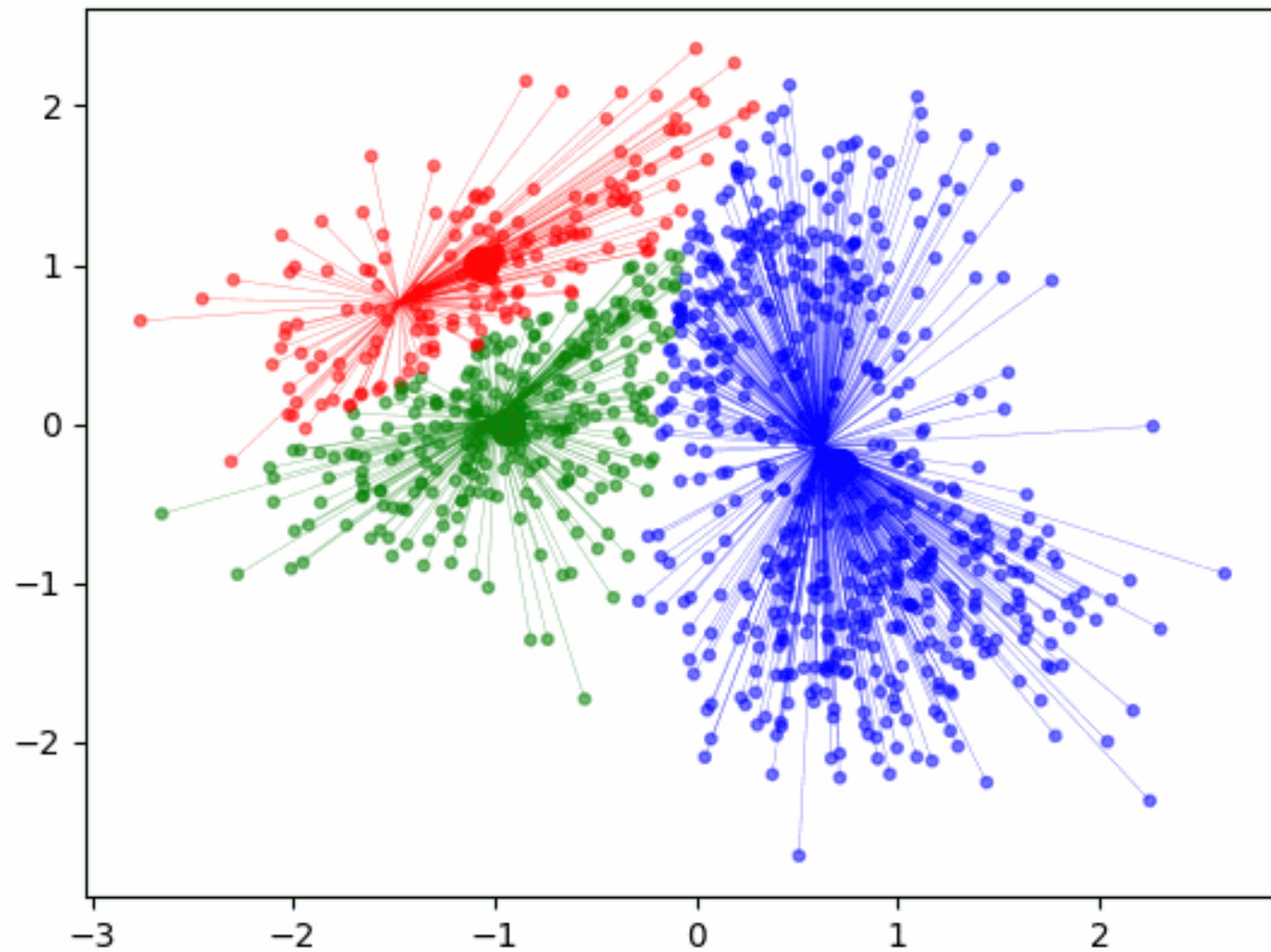
- Алгоритм k-means разбивает набор данных X на k кластеров S_1, S_2, \dots, S_k , таким образом, чтобы минимизировать сумму квадратов расстояний от каждой точки кластера до его центра.
- Основная идея: на каждой итерации перевычисляется центр масс для каждого кластера, полученного на предыдущем шаге, затем векторы разбиваются на кластеры вновь в соответствии с тем, какой из новых центров оказался ближе по выбранной метрике. Алгоритм завершается, когда на какой-то итерации не происходит изменения кластеров.
- Необходимо заранее знать количество кластеров
- Чувствителен к выбору начальных центров кластеров



K-MEANS



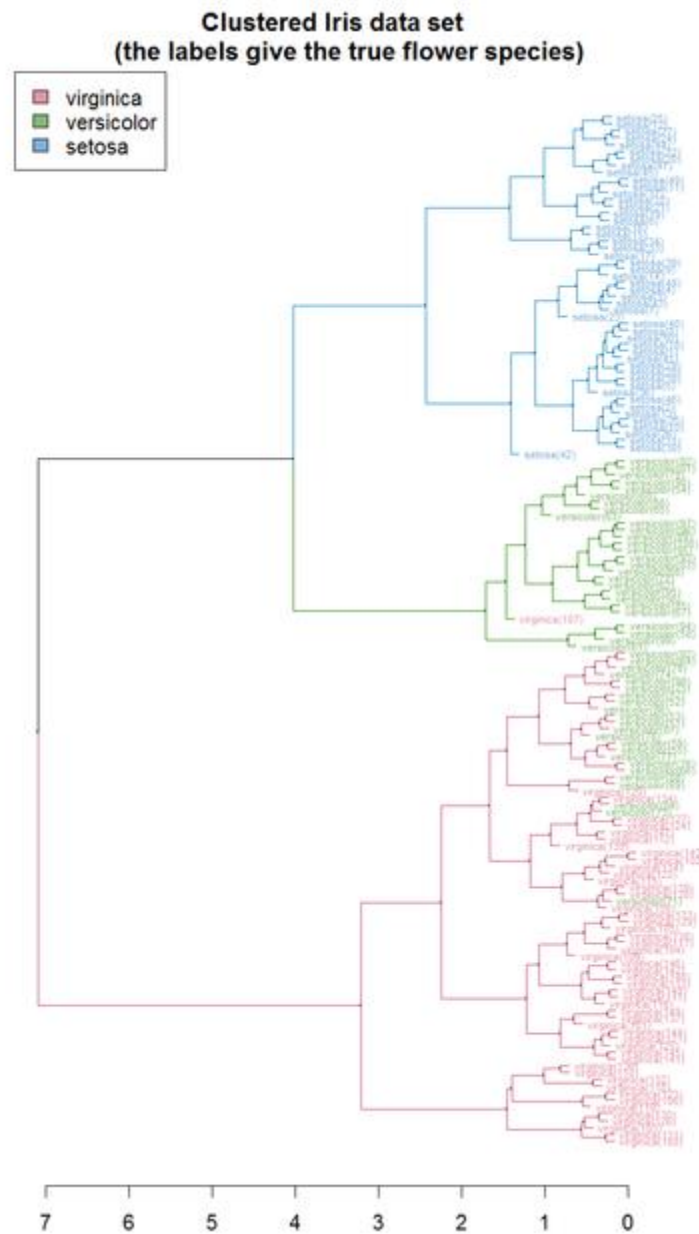
KMeans iter: 0



ИЕРАРХИЧЕСКАЯ КЛАСТЕРИЗАЦИЯ

- Иерархическая кластеризация — общее семейство алгоритмов кластеризации, которые создают вложенные кластеры путем их последовательного слияния или разделения. Эта иерархия кластеров представляется в виде дерева (дендрограммы).
- Идея: изначально – каждый элемент это отдельный кластер. Далее они объединяются при помощи определения меры близости между кластерами итеративно, пока не будет построено дерево с корнем – объединяющим все поддеревья.
- Меры близости:
 - Одиночная связность («ближний сосед»)
 - Полная связность
 - По центрам массы
 - ...





ИЕРАРХИЧЕСКАЯ КЛАСТЕРИЗАЦИЯ



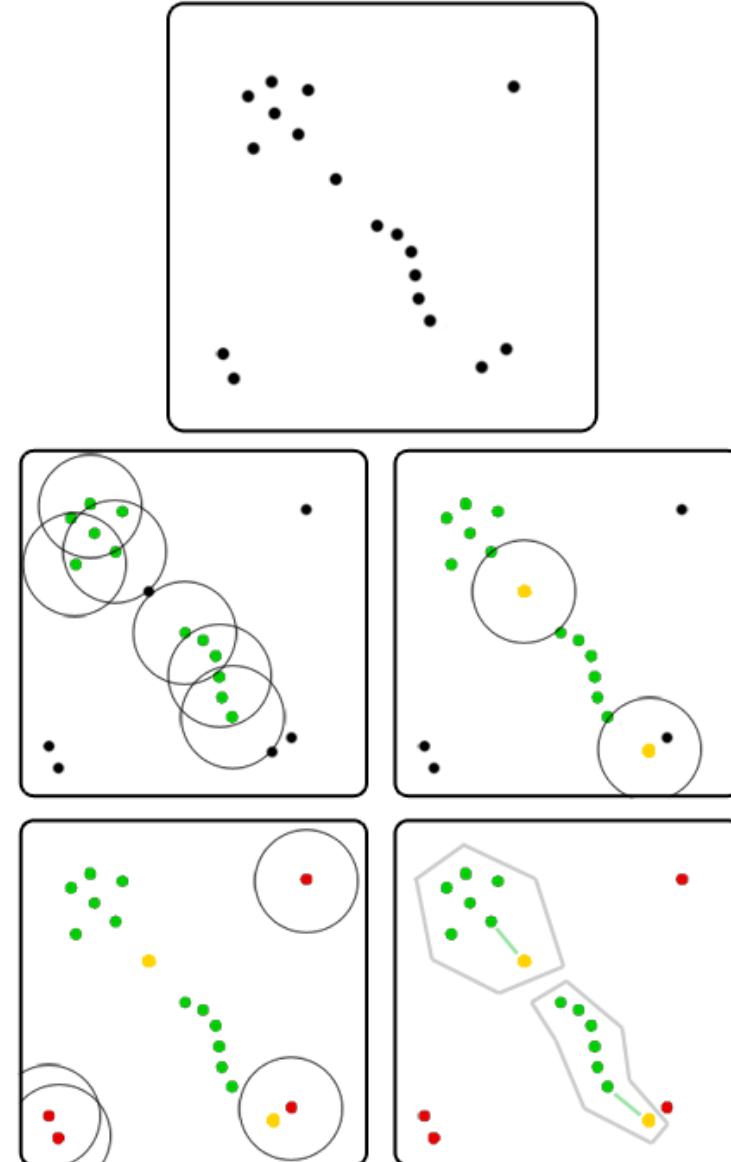
DBSCAN

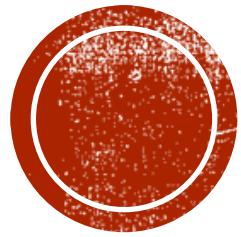
- **Density-Based Spatial Clustering of Applications with Noise** или плотностной алгоритм пространственной кластеризации с присутствием шума
- Не требует предварительных предположений о числе кластеров
- Необходимо настроить два параметра:
 - `Eps` – максимальное расстояние между точками кластера
 - `min_samples` – минимальное число элементов кластера.
- Отлично работает на плотных, хорошо отделённых друг от друга кластерах (форма не важна).
- Отлично обнаруживает кластеры малой размерности.



DBSCAN

- Зеленые – имеют 3 и более соседей – (корневые элементы).
- Желтые – имеют зеленого соседа (граница).
- Красные – не имеют зеленых соседей – (выбросы).
- Подробнее см. по [ссылке](#)





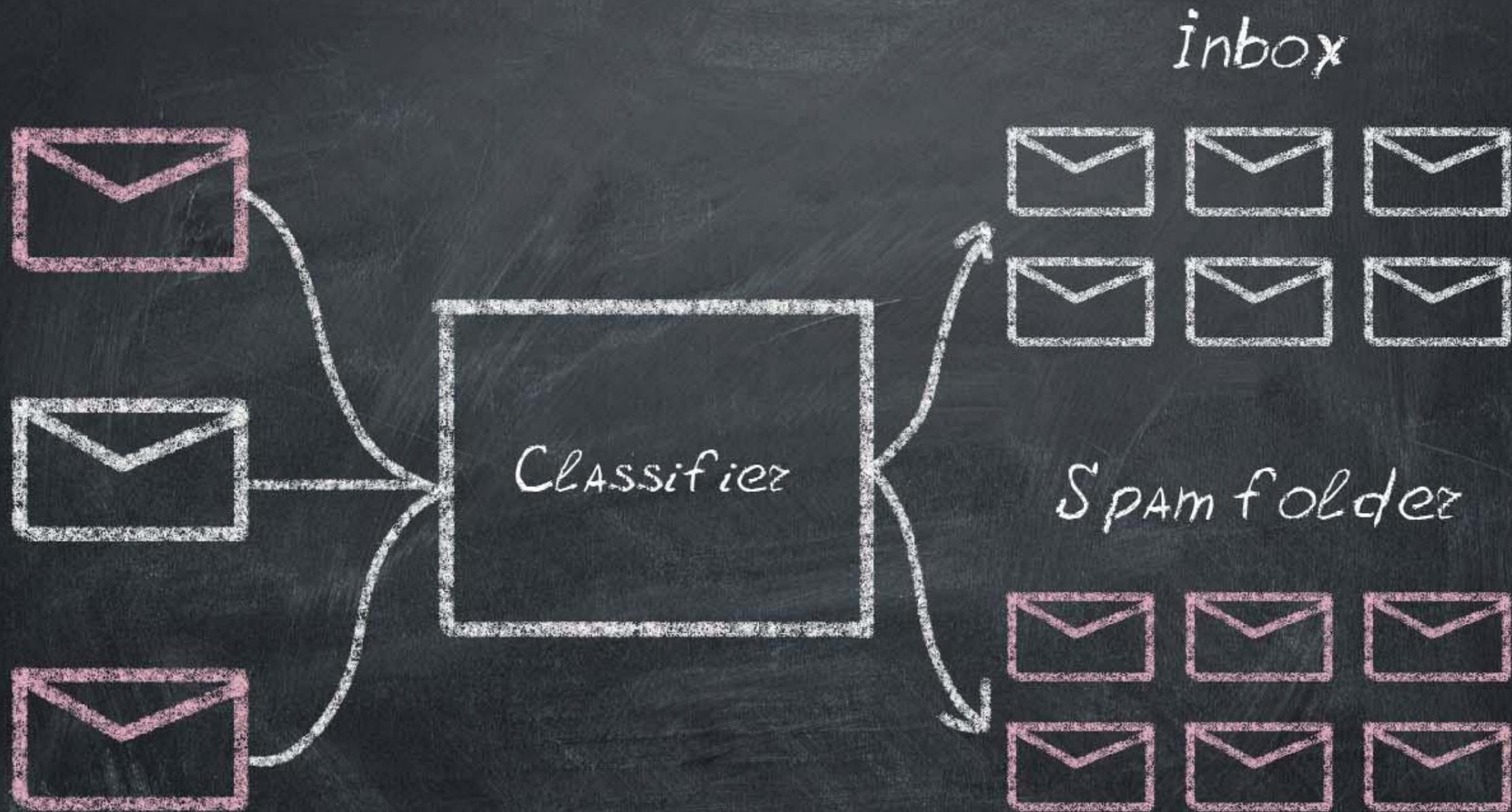
КЛАССИФИКАЦИЯ ДАННЫХ



КЛАССИФИКАЦИЯ

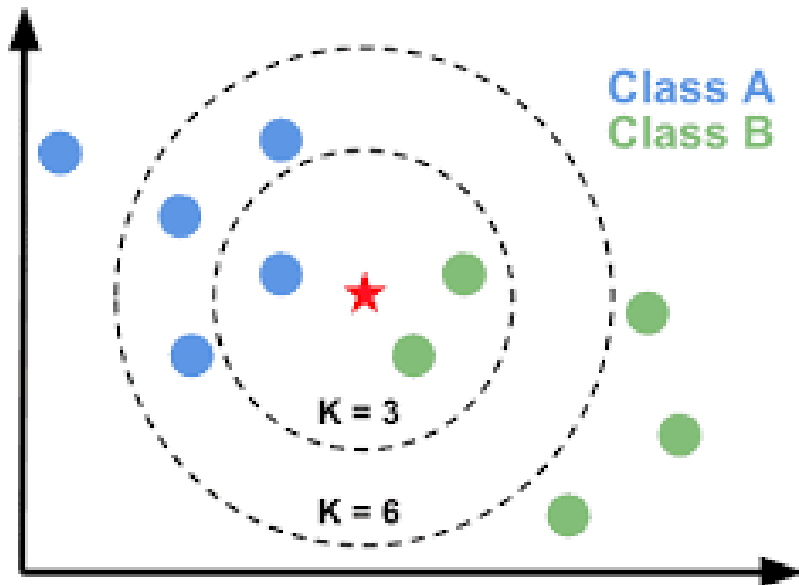
- **Классификация** – предсказание категории объекта и разделение объектов согласно определенным и заданным заранее признакам.
- Методы:
 - Линейные модели
 - Дискриминантный анализ
 - Метод опорных векторов
 - Градиентный спуск
 - Ближайшие соседи
 - Гауссовские
 - Байесовские
 - Ансамблевые
 - Нейронные сети
 - ...





МЕТОД К-БЛИЖАЙШИХ СОСЕДЕЙ

- Этот метод работает с помощью поиска кратчайшей дистанции между тестируемым объектом и ближайшими к нему классифицированными объектами из обучающего набора. Классифицируемый объект будет относиться к тому классу, к которому принадлежит ближайший объект набора.

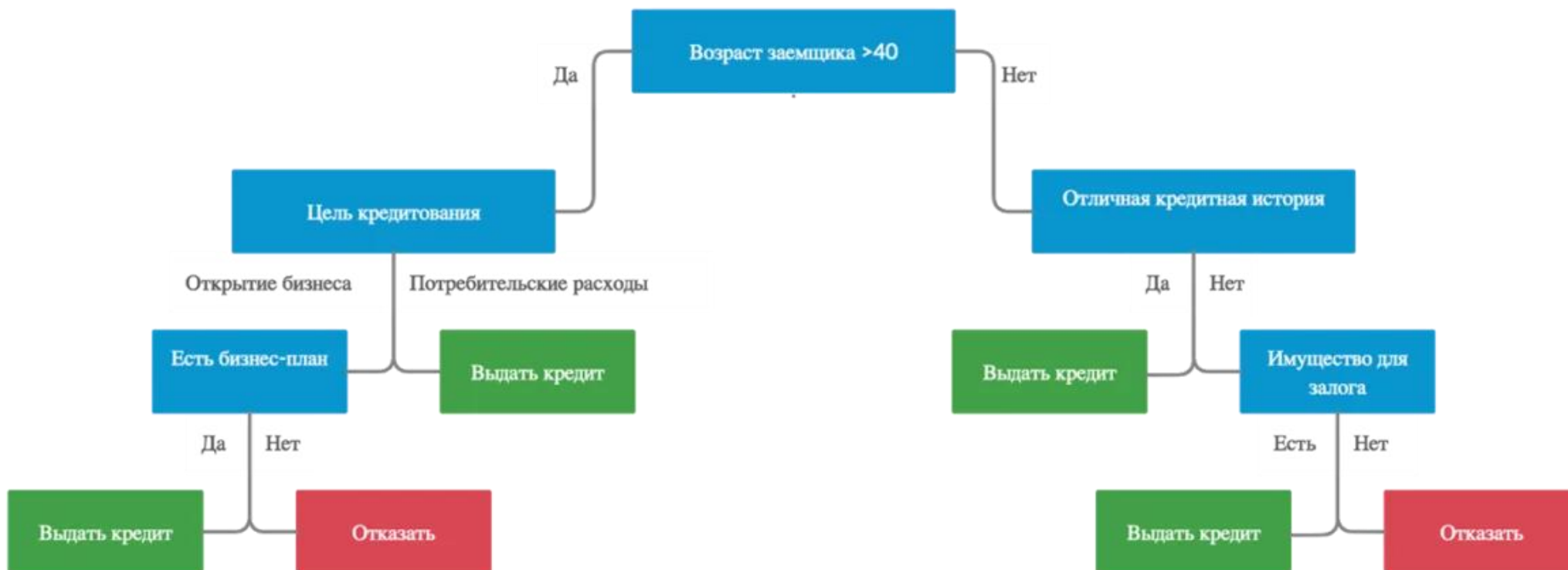


ДЕРЕВЬЯ РЕШЕНИЙ

- Этот классификатор разбивает данные на всё меньшие и меньшие подмножества на основе разных критериев, т. е. у каждого подмножества своя сортирующая категория. С каждым разделением количество объектов определённого критерия уменьшается.
- Классификация подойдёт к концу, когда сеть дойдёт до подмножества только с одним объектом.



ДЕРЕВЬЯ РЕШЕНИЙ



НАИВНЫЙ БАЙЕСОВСКИЙ КЛАССИФИКАТОР

- Такой классификатор вычисляет вероятность принадлежности объекта к какому-то классу. Эта вероятность вычисляется из шанса, что какое-то событие произойдёт, с опорой на уже произошедшие события.
- Каждый параметр классифицируемого объекта считается независимым от других параметров.
- <https://scikit-learn.ru/1-9-naive-bayes/>



ДИСКРИМИНАНТНЫЙ АНАЛИЗ

- Этот метод работает путём уменьшения размерности набора данных, проецируя все точки данных на линию. Потом он комбинирует эти точки в классы, базируясь на их расстоянии от центральной точки.
- Этот метод, как можно уже догадаться, относится к линейным алгоритмам классификации, т. е. он хорошо подходит для данных с линейной зависимостью.

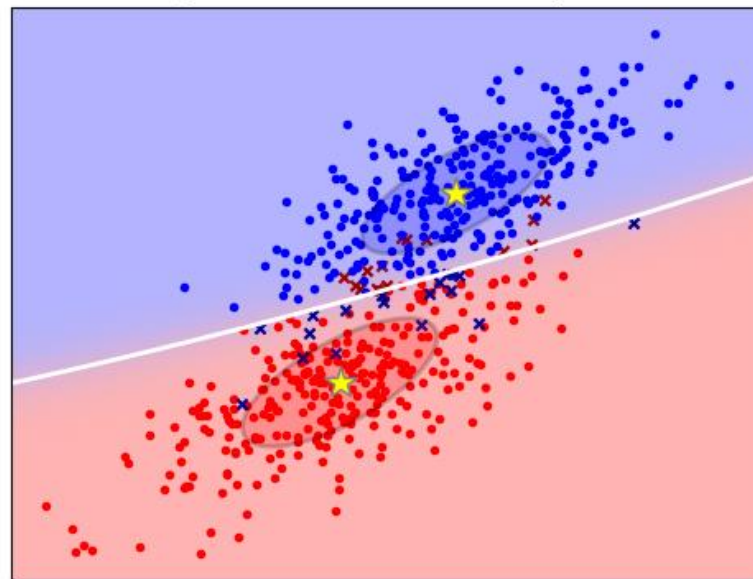
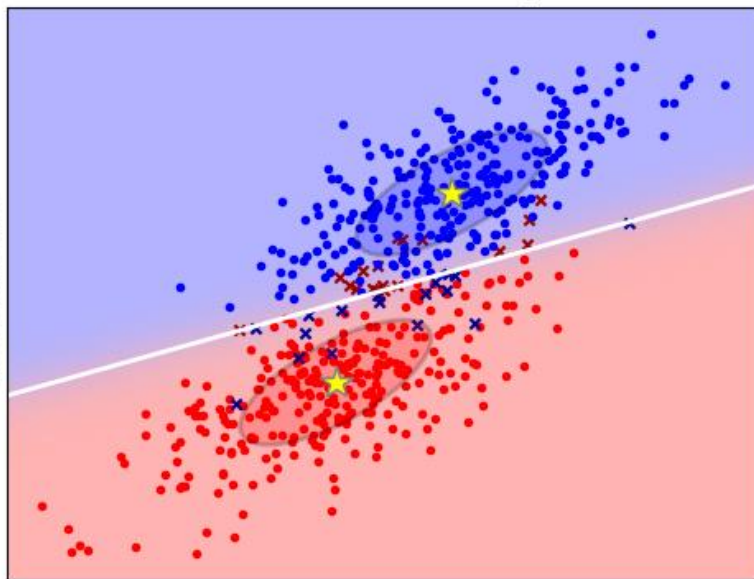


Linear Discriminant Analysis vs Quadratic Discriminant Analysis

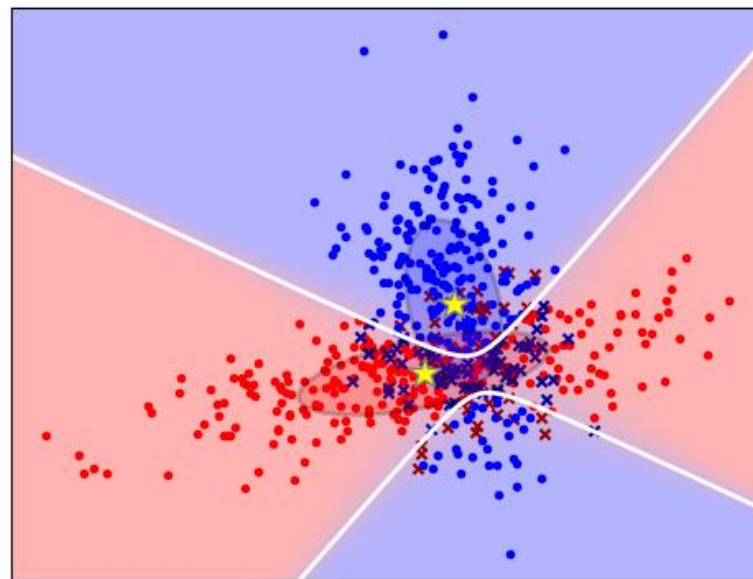
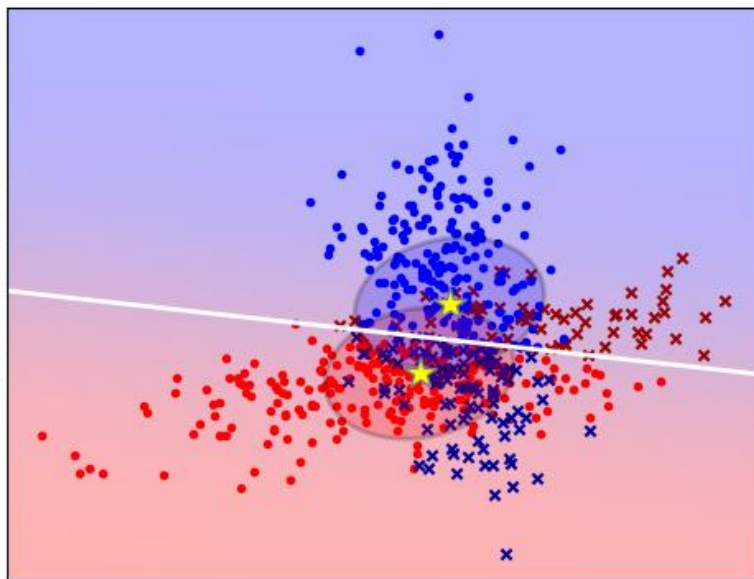
Linear Discriminant Analysis

Quadratic Discriminant Analysis

Data with
fixed covariance



Data with
varying covariances



ЧТО ИСПОЛЬЗОВАТЬ И ИЗУЧАТЬ?

- <https://scikit-learn.ru/>
- https://scikit-learn.ru/category/supervised_learning/
- <https://scikit-learn.ru/clustering/#clustering>
- <https://github.com/Sych474/BMSTU-app-programming-languages/tree/main/src/python-notebooks> - материалы демонстрации

