- **Вопрос. Методы прогнозного моделирования.**

Анализ данных — область математики и информатики, занимающаяся построением и исследованием наиболее общих математических методов и вычислительных алгоритмов извлечения знаний из экспериментальных (в широком смысле) данных; процесс исследования, фильтрации, преобразования и моделирования данных с целью извлечения полезной информации и принятия решений. Анализ данных имеет множество аспектов и подходов, охватывает разные методы в различных областях науки и деятельности (Wiki).

Примеры:

- **Традиционный статистический анализ**, описательный(размер выборки, мода, медиана, среднее, min,max, отклонения)
- **Разведочный анализ** - опровержение или подтверждение гипотез (изучение и визуализация входных данных, выявление закономерностей)

Это все описательная статистика

- В узком смысле

- **Data mining** (рус. *добыча данных, интеллектуальный анализ данных, глубинный анализ данных*) — собирательное название, используемое для обозначения совокупности методов обнаружения в данных ранее неизвестных, нетривиальных, практически полезных и доступных интерпретации знаний, необходимых для принятия решений в различных сферах человеческой деятельности. Термин введён Григорием Пятецким-Шапиро в 1989 году.
- Knowledge Discovery in Databases, KDD- обнаружение знаний в базах данных

О терминологии

What is the difference between Data Analytics, Data Analysis, Data Mining, Data Science, Machine Learning, Big Data?

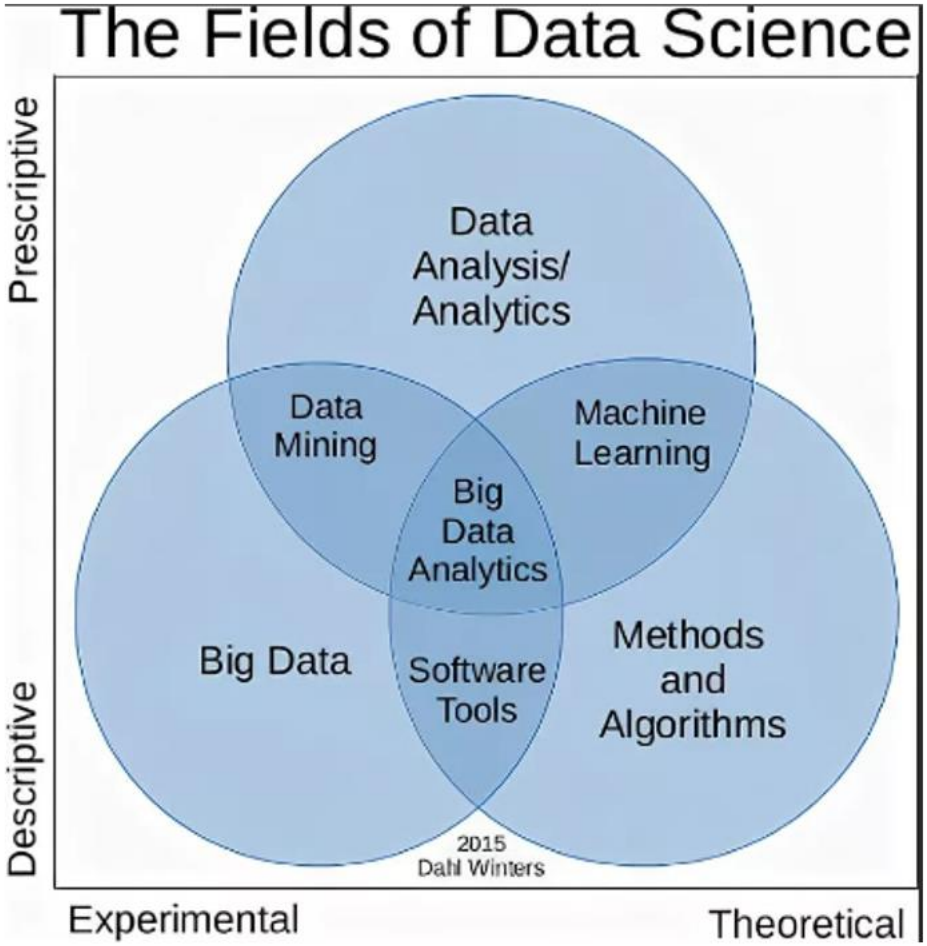
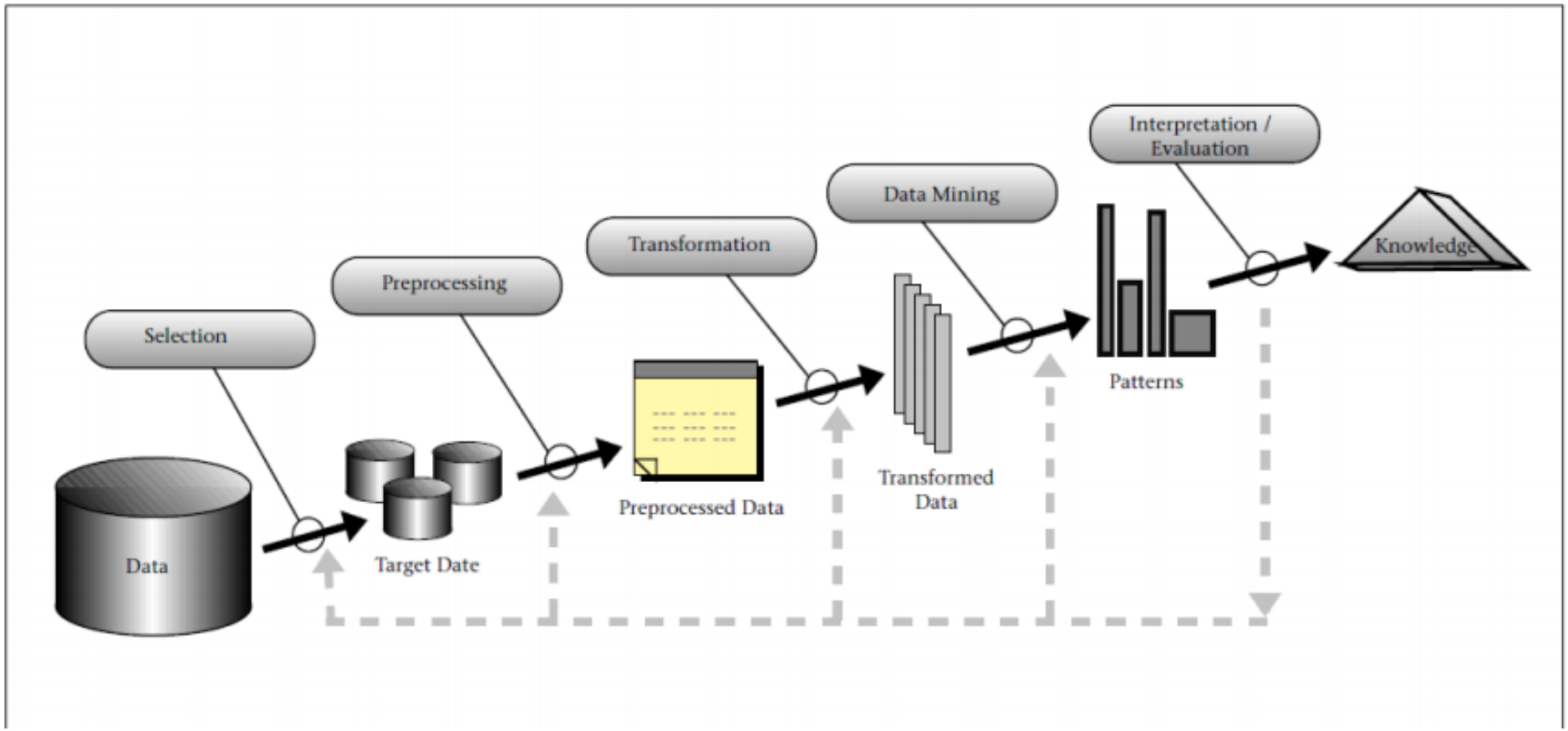


Схема процесса обнаружения знаний в данных

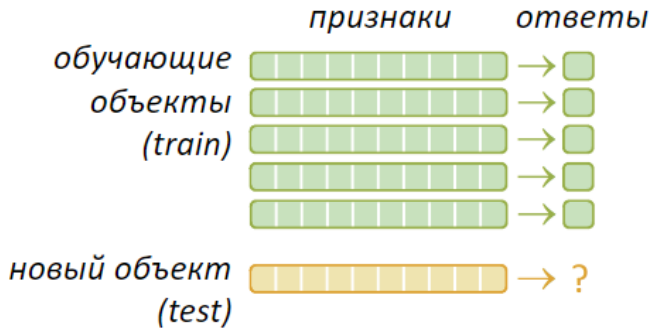


Задача машинного обучения с учителем

Этап №1 – обучение с учителем

- **На входе:**
  данные – выборка прецедентов «*объект* → *ответ*»,
  каждый объект описывается набором *признаков*
- **На выходе:**
  модель, предсказывающая ответ по объекту

Если нет данных,
то нет
и машинного
обучения



Этап №2 – применение

- **На входе:**
  данные – новый объект
- **На выходе:**
  предсказание ответа на новом объекте

Классическое Обучение

Данные заранее категоризированы или численные

Данные никак не размечены

С учителем

Предсказать категорию

Классификация

«Разложи носки по цвету»

Предсказать значение

Регрессия

«Разложи галстуки по длине»

Без учителя

Разделить по схожести

Кластеризация

«Разложи похожие вещи по кучкам»

Найти зависимости

Уменьшение Размерности (обобщение)

«Собери из вещей лучшие наряды»

Выявить последовательности

Ассоциация

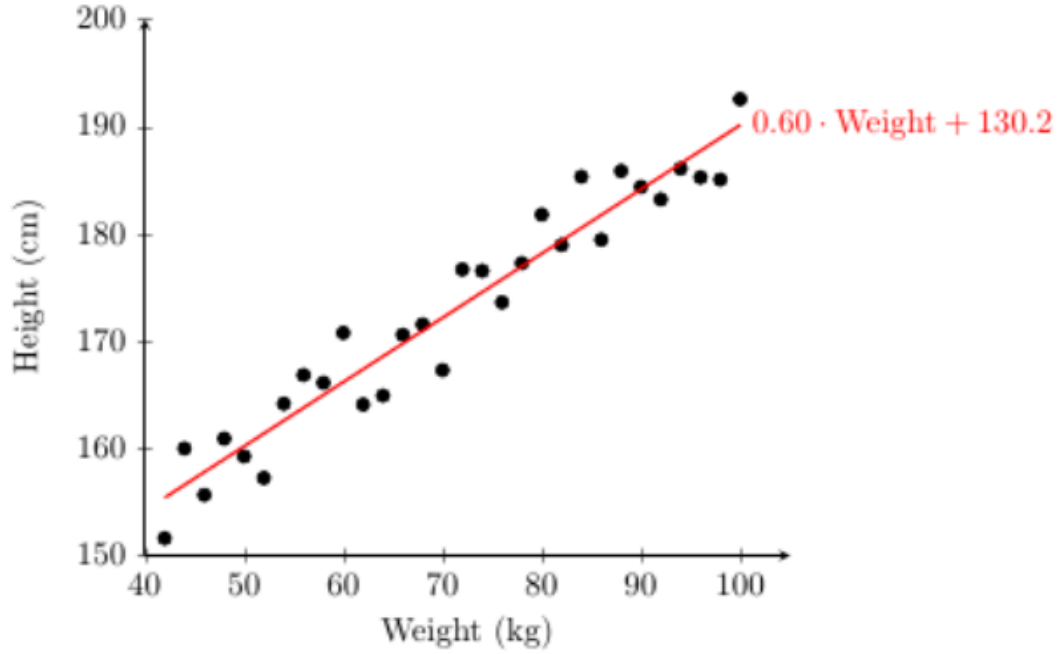
«Найди какие вещи я часто ношу вместе»

Есть две основные задачи машинного обучения с учителем:

- классификация (classification)
- регрессия (regression)

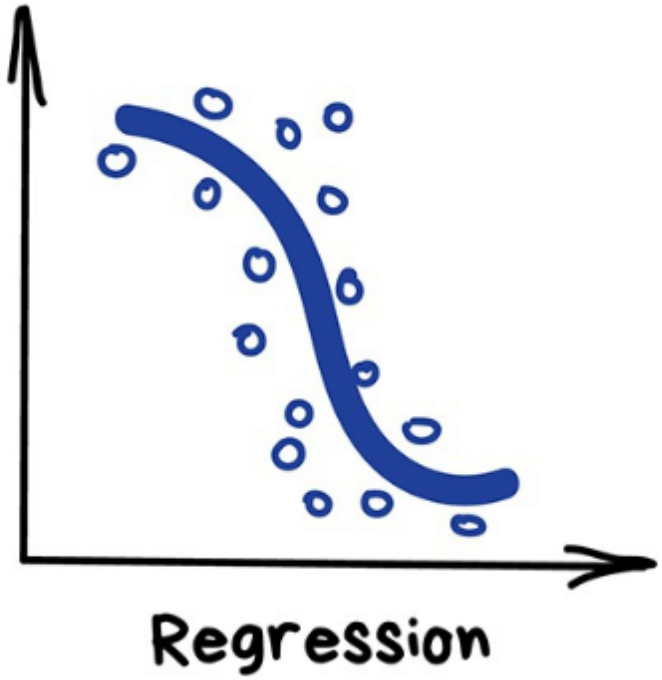
Регрессия

- Вещественные ответы: $\mathbb{Y} = \mathbb{R}$
- (вещественные числа — числа с любой дробной частью)
- Пример: предсказание роста по весу



Регрессия

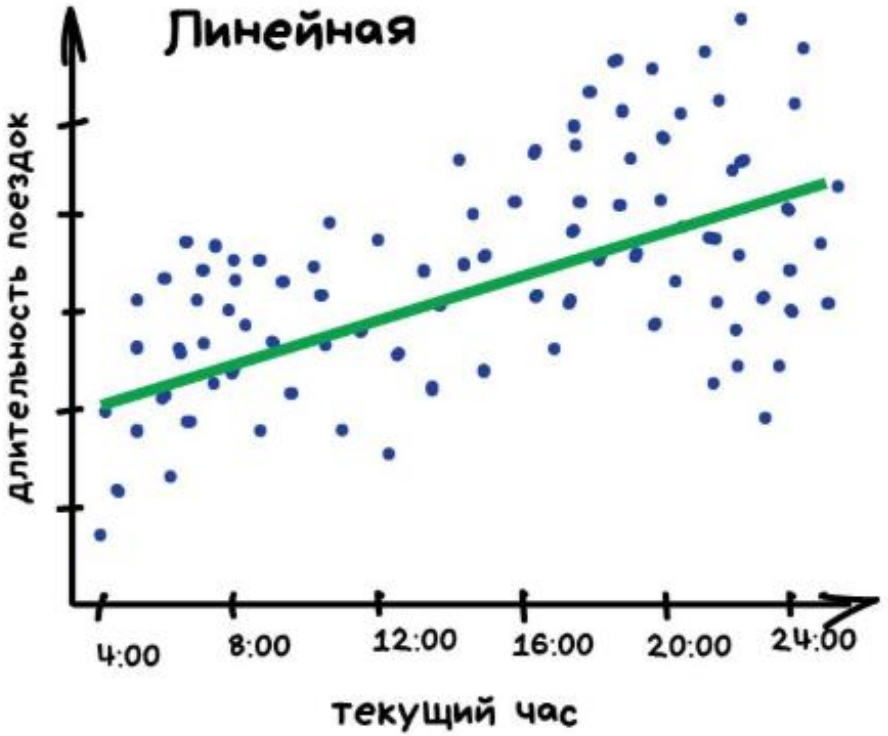
«Нарисуй линию вдоль моих точек.
Да, это машинное обучение»

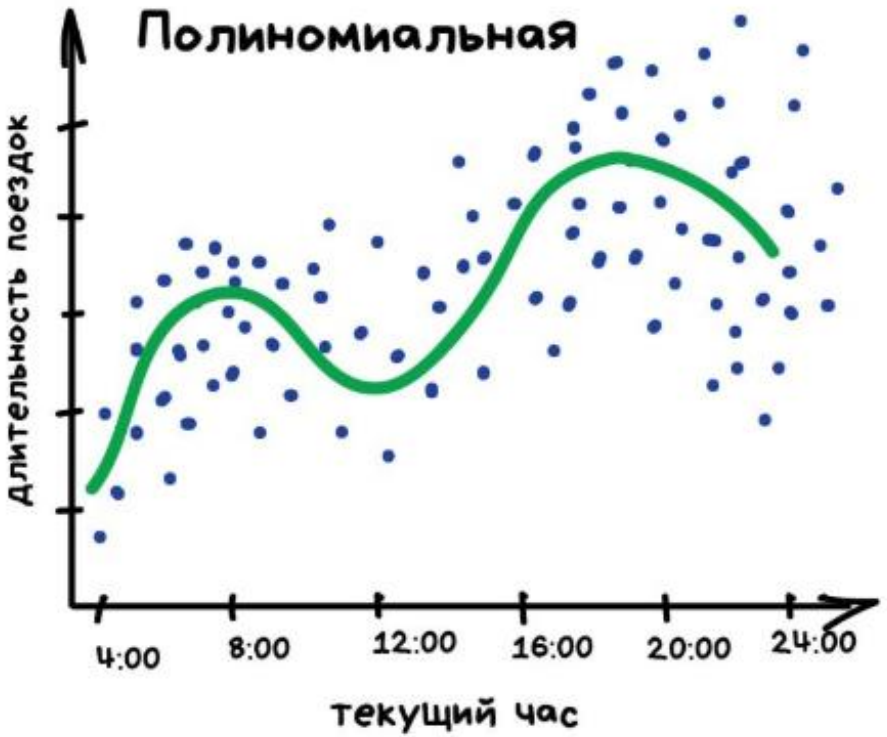


Сегодня используют для:

- Прогноз стоимости ценных бумаг
- Анализ спроса, объема продаж
- Медицинские диагнозы
- Любые зависимости числа от времени

Предсказываем пробки

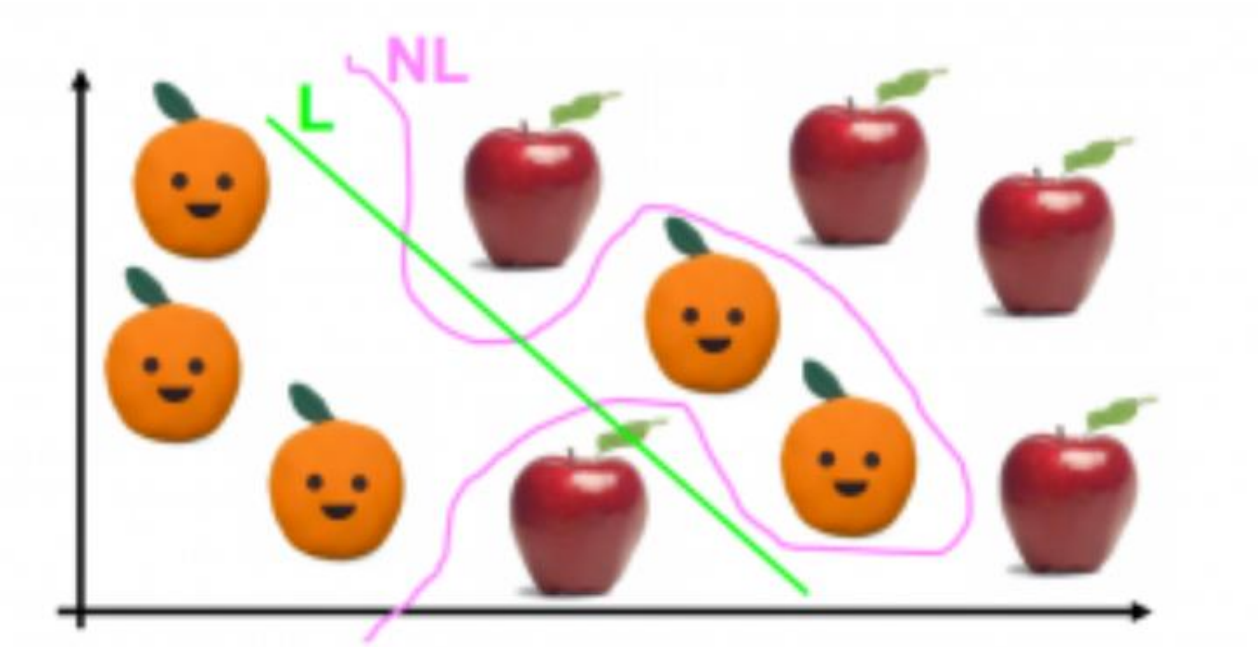




Регрессия

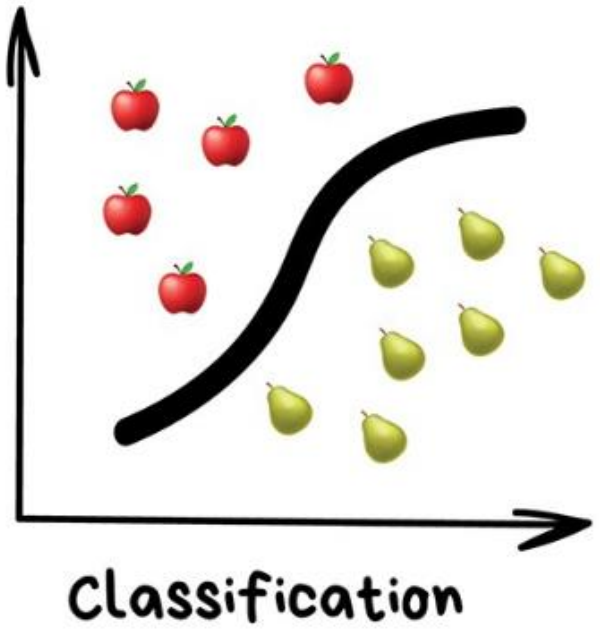
Классификация

- Конечное число ответов: $|\mathbb{Y}| < \infty$
- Бинарная классификация: $\mathbb{Y} = \{-1, +1\}$



Классификация

«Разделяет объекты по заранее известному признаку. Носки по цветам, документы по языкам, музыку по жанрам»



Сегодня используют для:

- Спам-фильтры
- Определение языка
- Поиск похожих документов
- Анализ тональности
- Распознавание рукописных букв и цифр
- Определение подозрительных транзакций

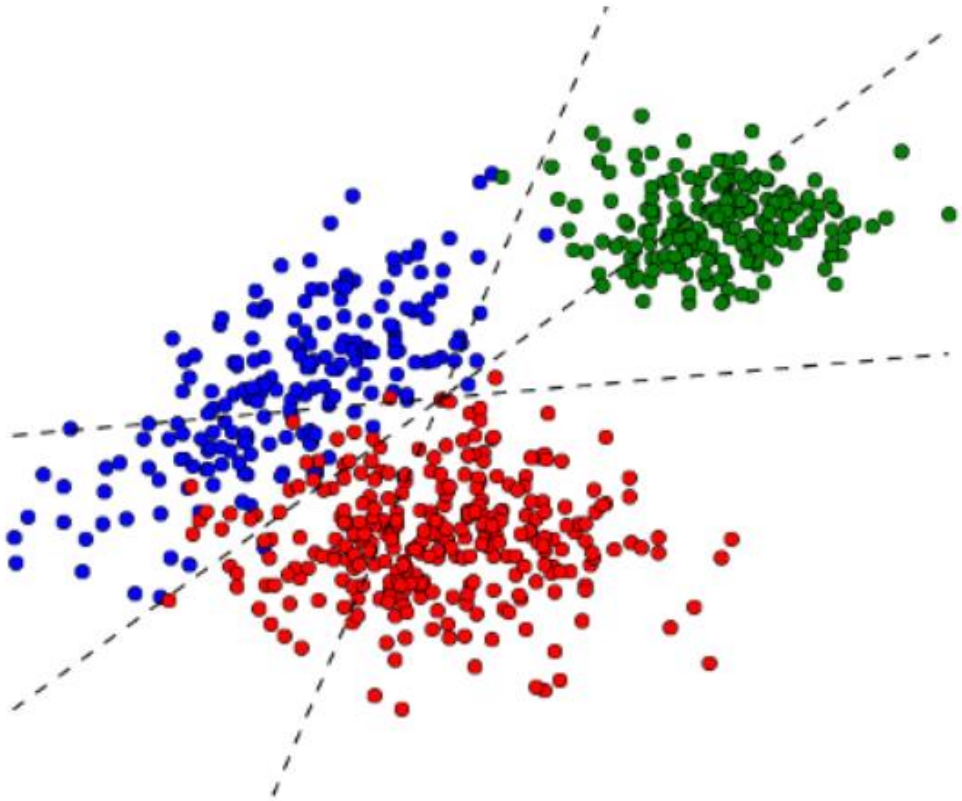
1. $\mathbb{Y} = \{0, 1\}$ — бинарная классификация. Например, мы можем предсказывать, кликнет ли пользователь по рекламному объявлению, вернет ли клиент кредит в установленный срок, сдаст ли студент сессию, случится ли определенное заболевание с пациентом (на основе, скажем, его генома).

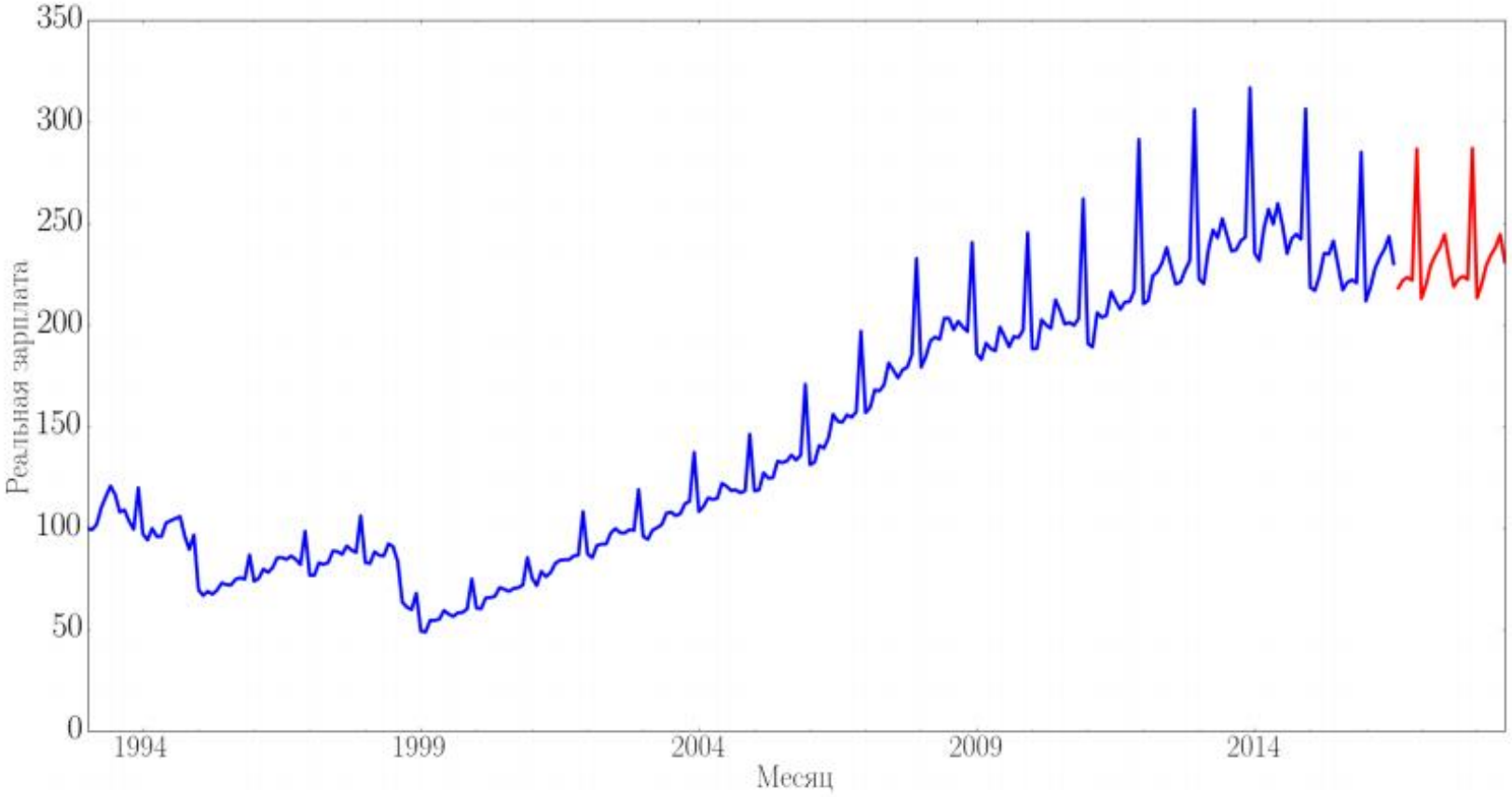
2. $\mathbb{Y} = \{1, \ldots, K\}$ — многоклассовая (multi-class) классификация. Примером может служить определение предметной области для научной статьи (математика, биология, психология и т.д.).

3. $\mathbb{Y} = \{0, 1\}^K$ — многоклассовая классификация с пересекающимися классами (multi-label classification). Примером может служить задача автоматического проставления тегов для ресторанов (логично, что ресторан может одновременно иметь несколько тегов).

Классификация

- Многоклассовая классификация: $\mathbb{Y} = \{1, 2, \dots, K\}$





Прогноз реальной заработной платы в России на два года вперёд (красным).

Задача кредитного скоринга

Объект — заявка на выдачу кредита.

Классы — bad или good.

Примеры признаков:

- **бинарные:** пол, наличие телефона, и т. д.
- **номинальные:** место проживания, профессия, работодатель, и т. д.
- **порядковые:** образование, должность, и т. д.
- **количественные:** возраст, зарплата, стаж работы, доход семьи, сумма кредита, и т. д.

Особенности задачи:

- нужно оценивать вероятность дефолта $P(\text{bad})$.

Объект — абонент в определённый момент времени.

Классы — уйдёт или не уйдёт в следующем месяце.

Примеры признаков:

- **бинарные:** корпоративный клиент, включение услуг, и т. д.
- **номинальные:** тарифный план, регион проживания, и т. д.
- **количественные:** длительность разговоров (входящих, исходящих, СМС, межгород, и т. д.), частота оплаты, и т. д.

Особенности задачи:

- нужно строить признаки по потоку действий абонентов;
- нужно оценивать вероятность ухода;
- сверхбольшие выборки.

Задача регрессии: прогноз стоимости недвижимости

Объект — квартира в Москве.

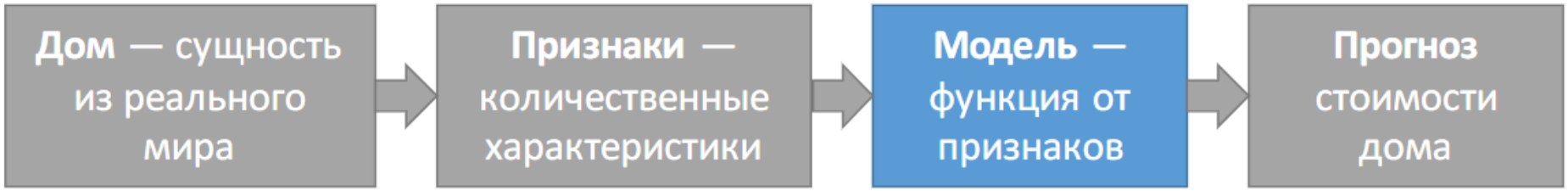
Примеры признаков:

- **бинарные:** наличие балкона, лифта, мусоропровода, охраны, гаража, чердака, и т. д.
- **номинальные:** район города, тип дома (кирпичный/панельный/блочный/монолит), и т. д.
- **количественные:** число комнат, жилая площадь, расстояние до центра, до метро, возраст дома, и т. д.

Особенности задачи:

- выборка неоднородна, стоимость меняется со временем;
- разнотипные признаки;
- для линейной модели нужны преобразования признаков.

Предсказание стоимости дома



Предсказание стоимости дома

Обучающая выборка:

Площадь	Цена
50	250
60	340
10	20
90	800

Возможные признаки:

- площадь
- площадь^2
- площадь^3
- $\sin(\text{площадь})$
- $\sqrt{\text{площадь}}$
- и так далее

Возможные модели:

- $w_1 * \text{площадь}$
- $w_1 * \text{площадь}^2$
- $w_1 * \text{площадь} + w_2 * \text{площадь}^2$
- и так далее

Вид модели — работа эксперта либо полный перебор.

Выбор весов w_1, w_2 — автоматический процесс (на основе данных)

Предсказание стоимости дома

Модель $a(x) = 5 * \text{площадь}$

Площадь	Прогноз	Цена	$(a - y)^2$
50	250	250	0
60	300	340	1600
10	50	20	900
90	450	800	122500

MSE: 31 250

RMSE: 176,78

Модель $a(x) = 0.1 * \text{площадь}^2$

Площадь	Прогноз	Цена	$(a - y)^2$
50	250	250	0
60	360	340	400
10	10	20	100
90	810	800	100

MSE: 150

RMSE: 12,25

Предсказание стоимости дома

Признаков может быть больше:

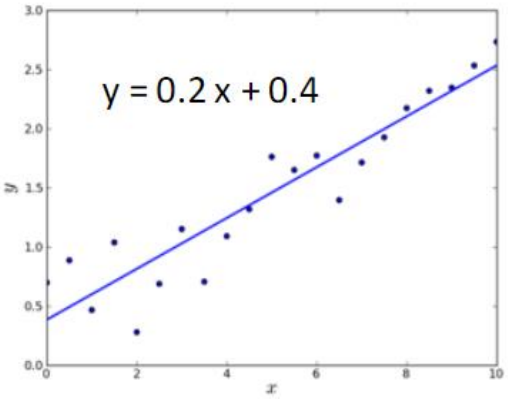
- Площадь
- Год постройки
- Наличие бассейна
- Число комнат
- Удалённость от центра
- Рейтинг полицейского участка
- И так далее

Возможные модели:

- Линейная: $w_1 * \text{площадь} + w_2 * \text{год} + w_3 * \text{бассейн} + w_4 * \text{комнаты} + w_5 * \text{удалённость} + w_6 * \text{полиция}$
- Решающие деревья
- Нейронные сети
- Метод k ближайших соседей
- И так далее

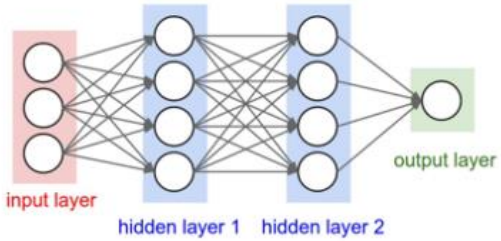
Основные классы моделей

Линейные модели



Решающие деревья

Нейронные сети



Вопрос. Методы поиска закономерностей (ассоциации и кластеризация).

Цели обучения без учителя

- в *Data Mining:*
  выявлять структуру в данных для лучшего их понимания;
- в *Machine Leanring:*
  как предварительный этап при решении задачи обучения с учителем (например, сокращение размерности (PCA и др.) или решаем задачу кластеризации, а потом в каждом кластере — свою задачу классификации и т. п.).

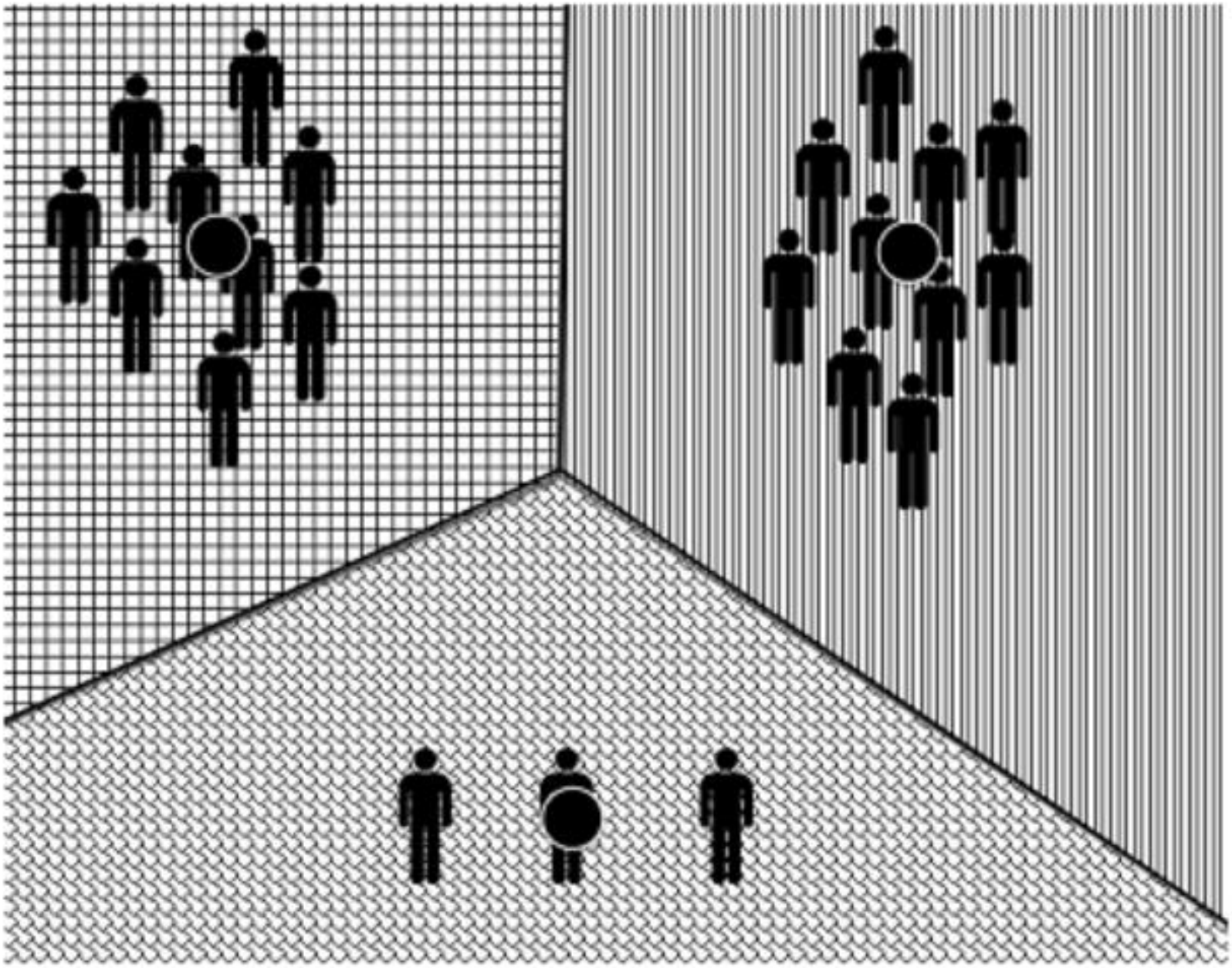
- 1. **Кластеризация** — задача разделения объектов на группы, обладающие некоторыми свойствами. Примером может служить кластеризация документов из электронной библиотеки или кластеризация абонентов мобильного оператора.
- 2. **Восстановление плотности** — задача приближения распределения объектов. Примером может служить задача обнаружения аномалий, в которой на этапе обучения известны лишь примеры «правильного» поведения игроков на бирже, а в дальнейшем требуется обнаруживать случаи незаконного поведения игроков. В таких задачах сначала оценивается распределение «правильных» объектов, а затем аномальными объявляются все объекты, которых в рамках этого распределения получают слишком низкую вероятность.
-

Кластеризация

«Разделяет объекты по неизвестному признаку. Машина сама решает как лучше»

Сегодня используют для:

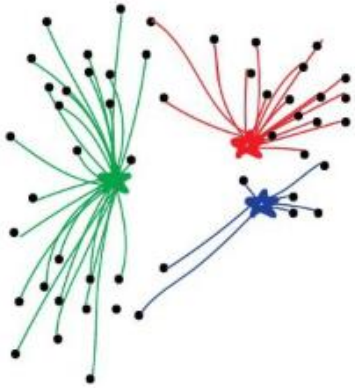
- Сегментация рынка (типов покупателей, лояльности)
- Объединение близких точек на карте
- Сжатие изображений
- Анализ и разметки новых данных
- Детекторы аномального поведения



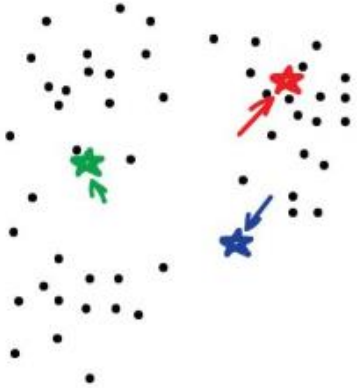
Ставим три ларька с шаурмой оптимальным образом

(иллюстрируя метод K-средних)

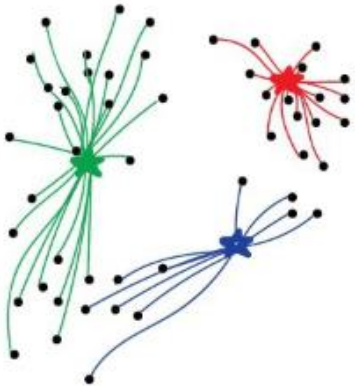
1. Ставим ларьки с шаурмой в случайных местах



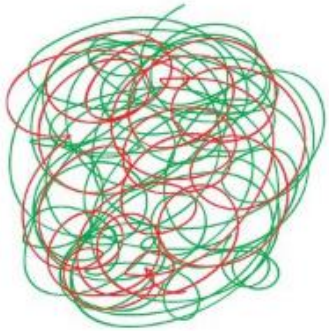
2. Смотрим в какой кому ближе идти



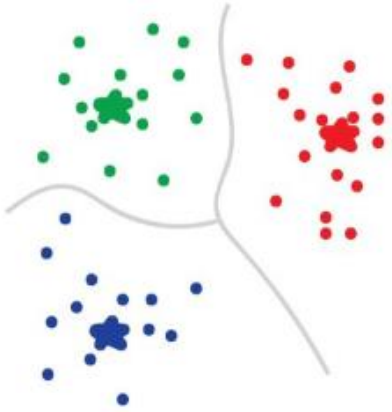
3. Двигаем ларьки ближе к центрам их популярности



4. Снова смотрим и двигаем



5. Повторяем много раз



6. Готово, вы великолепны!

- **Примеры применения**
- **Сегментация** и построение профилей клиентов. С помощью кластеризации можно выделить сегменты с группами "похожих" объектов. Данный алгоритм дает возможность выделить характерные признаки и персональные предпочтения клиентов, оценить наиболее и наименее доходные или активные сегменты. Это позволяет решить задачи разработки маркетинговых акций, направленных на определенные сегменты клиентов, повышает эффективность работы с ними.

- **Выявление целевой аудитории** – наиболее ценной, перспективной, влиятельной группы потребителей, на которую, в первую очередь, будет направлена маркетинговая стратегия. Позволяет решить задачи разработки рекламного сообщения и подбора медиаканалов для его размещения, позиционирования, выбора товарного ассортимента и каналов дистрибуции… Концентрация усилий на целевой аудитории обеспечит максимизацию прибыли в сегменте.

- **Каннибализация товаров:** продукты, находящиеся в одной рыночной нише, "поедают" друг друга, то есть конкурируют за потребителя между собой. Алгоритм дает возможность выделять товары, находящиеся в «зоне риска», прогнозировать эффект каннибализации и управлять им.

- **Анализ миграции клиентов** – перемещение клиентов между поставщиками товаров и услуг, причиной которой является изменение их запросов со временем. Рассматриваемые алгоритмы позволяют прогнозировать миграцию клиентов, визуализировать ее, оценить изменение их ценности для компании, определить причину миграции. В результате происходит укрепление отношений с ценными клиентами и противодействие оттоку.

-

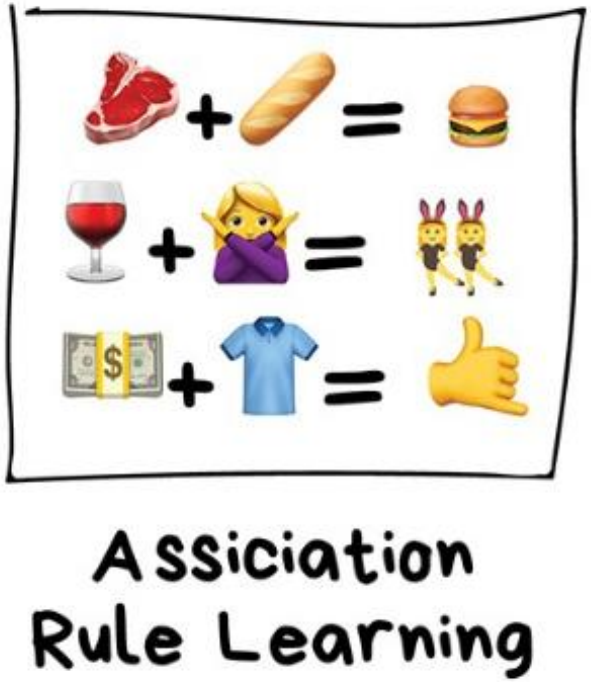
- 3. **Поиск ассоциаций-** это метод обучения машин на базе правил обнаружения интересующих нас связей между переменными в большой базе данных..

Поиск правил (ассоциация)

«Ищет закономерности в потоке заказов»

Сегодня используют для:

- Прогноз акций и распродаж
- Анализ товаров, покупаемых вместе
- Расстановка товаров на полках
- Анализ паттернов поведения на веб-сайтах



Поиск ассоциаций.

Множество объектов I — это молоко, хлеб, масло, пиво, памперсы, и в таблице выше показана маленькая база данных, содержащая объекты, в которой значение 1 означает наличие объекта в соответствующей транзакции, а значение 0 означает отсутствие объекта в транзакции.

Примером правила для супермаркета может служить {масло, хлеб} => {молоко}, что означает, что, если куплены масло и хлеб, покупатель также купит и молоко.

Пример базы данных с 5 транзакциями и 5 элементами

ID транзакции	молоко	хлеб	масло	пиво	памперсы
1	1	1	0	0	0
2	0	0	1	0	0
3	0	0	0	1	1
4	1	1	1	0	0
5	0	1	0	0	0

Замечание: этот пример крайне мал. В практических приложениях, правило должно удовлетворяться в нескольких сотнях тысяч транзакций, прежде чем его будут считать статистически значимым, а базы данных часто содержат тысячи или миллионы транзакций.

Сюда входят все методы анализа продуктовых корзин, стратегий маркетинга и других последовательностей.

Предположим, покупатель берёт в дальнем углу магазина пиво и идёт на кассу. Стоит ли ставить на его пути орешки? Часто ли люди берут их вместе? Орешки с пивом, наверное да, но какие ещё товары покупают вместе? Когда вы владелец сети гипермаркетов, ответ для вас не всегда очевиден, но одно тактическое улучшение в расстановке товаров может принести хорошую прибыль.

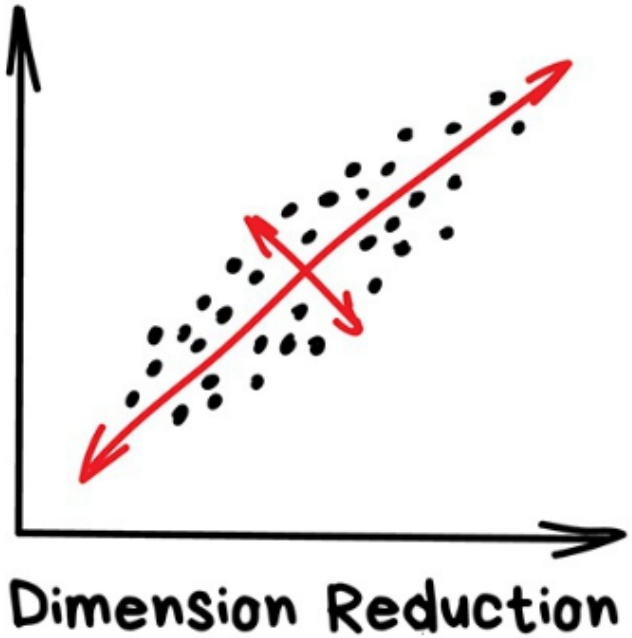
То же касается интернет-магазинов, где задача еще интереснее — за каким товаром покупатель вернётся в следующий раз?

- 3. Визуализация — задача изображения многомерных объектов в двумерном или трехмерном пространстве таким образом, что сохранялось как можно больше зависимостей и отношений между ними.

- 4. Понижение размерности — задача генерации таких новых признаков, что их меньше, чем исходных, но при этом с их помощью задача решается не хуже (или с небольшими потерями качества, или лучше — зависит от постановки). К этой же категории относится задача построения латентных моделей, где требуется описать процесс генерации данных с помощью некоторого (как правило, небольшого) набора скрытых переменных.

Уменьшение Размерности (Обобщение)

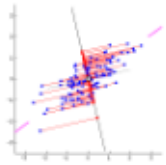
«Собирает конкретные признаки в абстракции более высокого уровня»



Сегодня используют для:

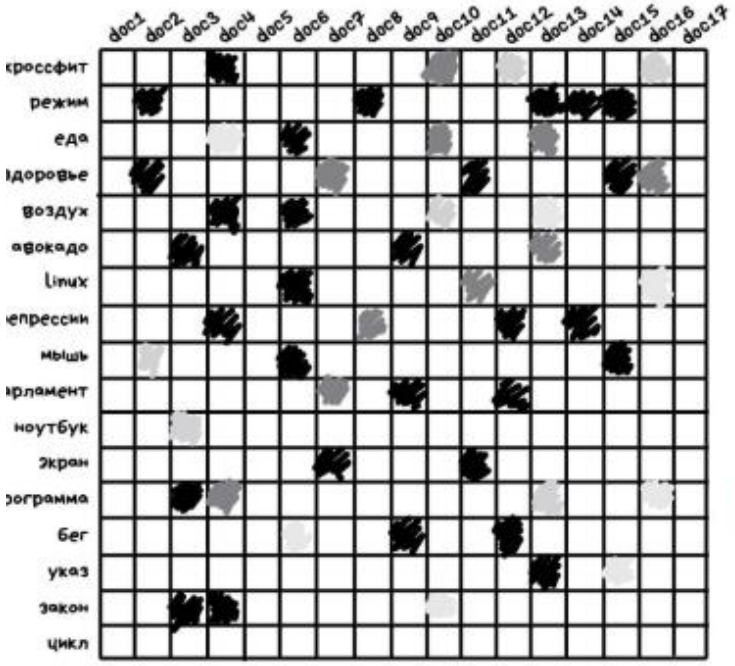
- Рекомендательные Системы (★)
- Красивые визуализации
- Определение тематики и поиска похожих документов
- Анализ фейковых изображений
- Риск-менеджмент

Для нас практическая польза их методов в том, что мы можем объединить несколько признаков в один и получить абстракцию. Например, собаки с треугольными ушами, длинными носами и большими хвостами соединяются в полезную абстракцию «овчарки». Да, мы теряем информацию о конкретных овчарках, но новая абстракция всяко полезнее этих лишних деталей. Плюс, обучение на меньшем количестве размерностей идёт сильно быстрее.



Проецируем 2D-данные на прямую (PCA)

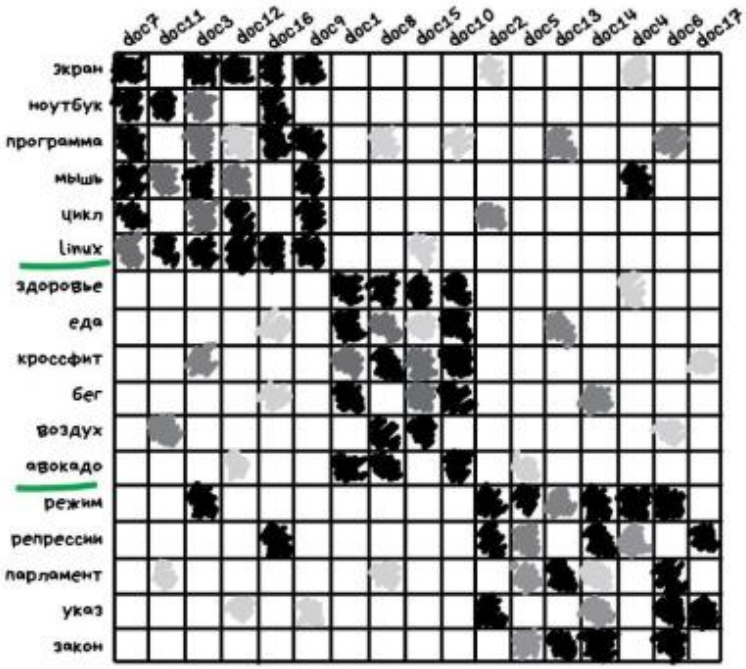
Разделение документов по темам



1. Строим матрицу как часто каждое слово встречается в каждом документе (чернее - чаще)

SVD

2. Раскладываем



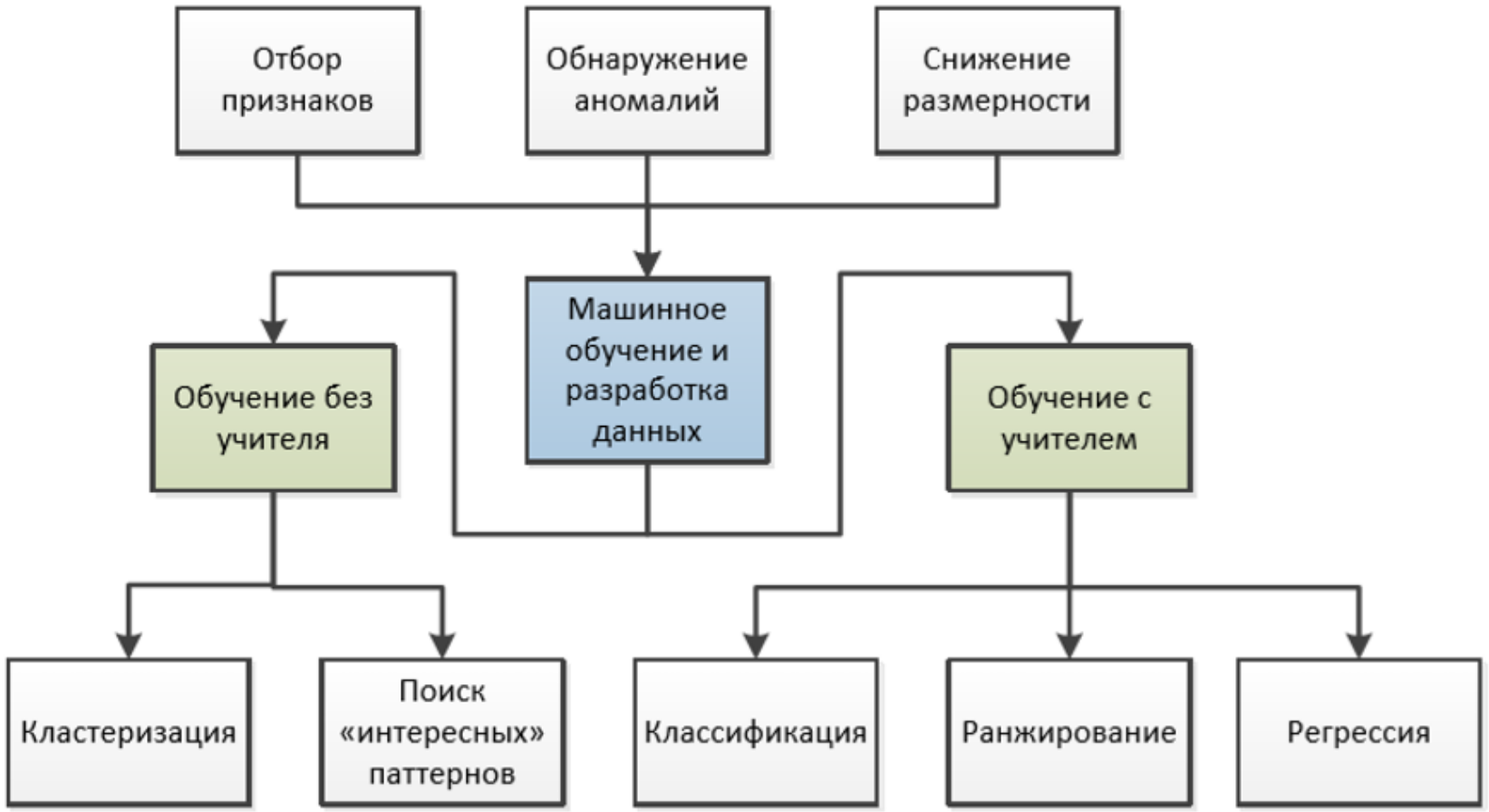
3. Получаем наглядные кластера по тематикам (даже если слова не встречались вместе)

Что предсказываем?

Два типа обучения:

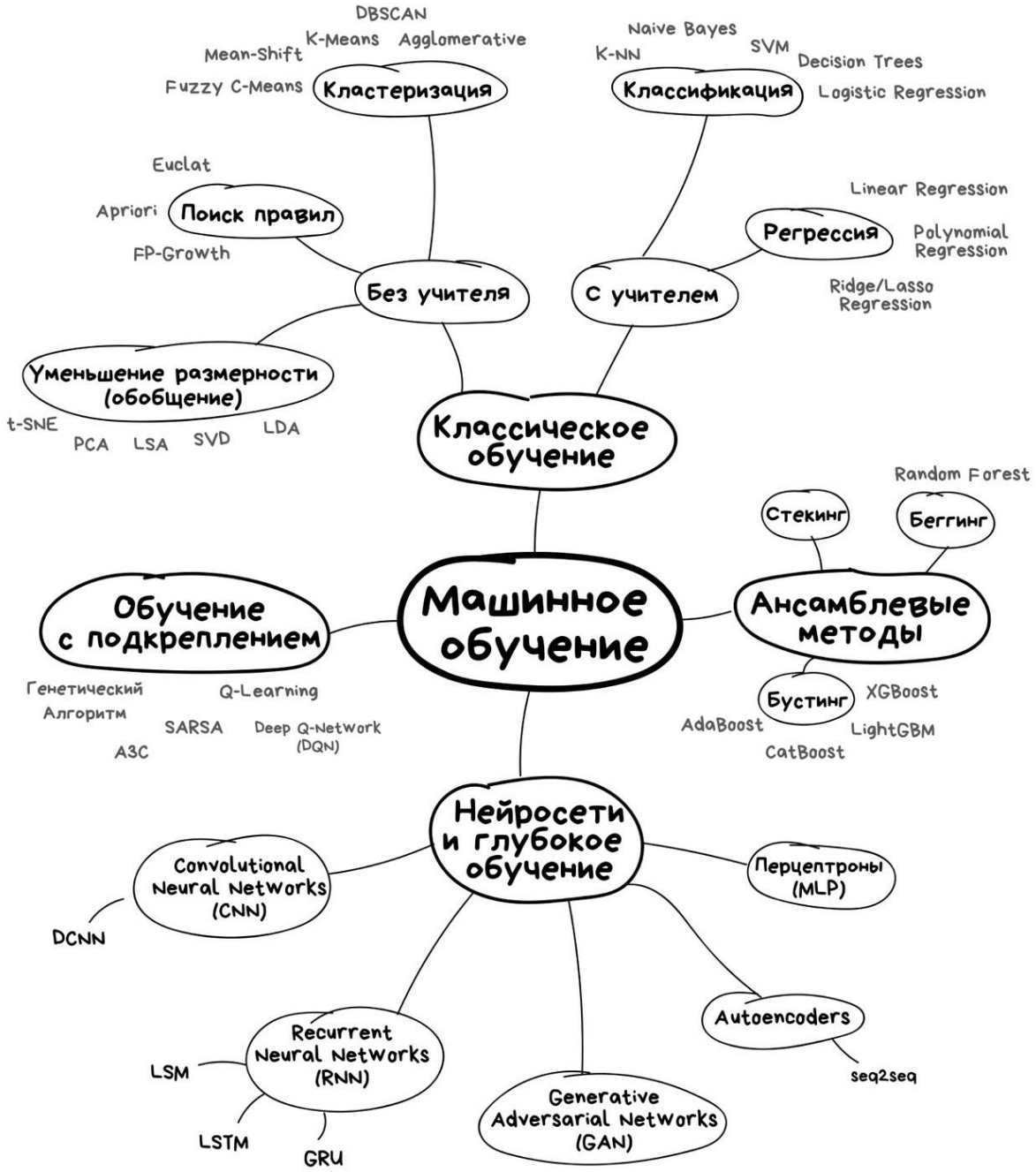
- Обучение с учителем (пытаемся понять, как зависят ответы, известные на объектах обучающей выборки, от входных данных):
  - Классификация (бинарная, multiclass, multilabel)
  - Регрессия
  - Прогнозирование временных рядов
  - Рекомендации
  - ...
- Обучение без учителя (как можем формализуем, что хотим найти в данных, и ищем).
  - Кластеризация
  - Понижение размерности
  - Визуализация
  - ...

Таксономия методов DM & ML



Давать ли кредит?

Дерево Решений


Машинное обучение
Классическое обучение
Без учителя
Кластеризация
DBSCAN
K-Means
Agglomerative
Mean-Shift
Fuzzy C-Means
Поиск правил
Euclat
Apriori
FP-Growth
Уменьшение размерности (обобщение)
t-SNE
PCA
LSA
SVD
LDA
С учителем
Классификация
Naive Bayes
K-NN
SVM
Decision Trees
Logistic Regression
Регрессия
Linear Regression
Polynomial Regression
Ridge/Lasso Regression
Обучение с подкреплением
Генетический Алгоритм
Q-Learning
SARSA
Deep Q-Network (DQN)
A3C
Ансамблевые методы
Стекинг
Беггинг
Random Forest
Бустинг
XGBoost
AdaBoost
LightGBM
CatBoost
Нейросети и глубокое обучение
Convolutional Neural Networks (CNN)
DCNN
Перцептроны (MLP)
Recurrent Neural Networks (RNN)
LSM
LSTM
GRU
Generative Adversarial Networks (GAN)
Autoencoders
seq2seq

Задачи, возможности и инструменты интеллектуального анализа данных в пакете SAS .