# МАШИННОЕ ОБУЧЕНИЕ

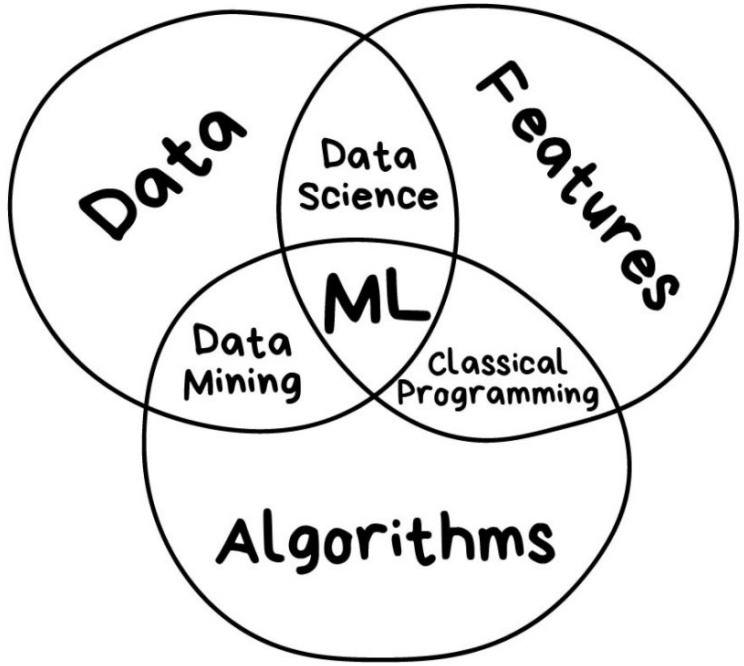
РГГУ ФиКЛ-2021

А. Ивойлова

# ТРИ СОСТАВЛЯЮЩИЕ МО

Цель машинного обучения — предсказать результат по входным данным.

- Данные
- Признаки
- Алгоритм



# СТРУКТУРА ОБЛАСТИ ЗНАНИЙ


**Машина может**

Предсказывать

Запоминать

Воспроизводить

Выбирать лучшее

**Машина не может**

Создавать новое

Резко поумнеть

Выйти за рамки задачи

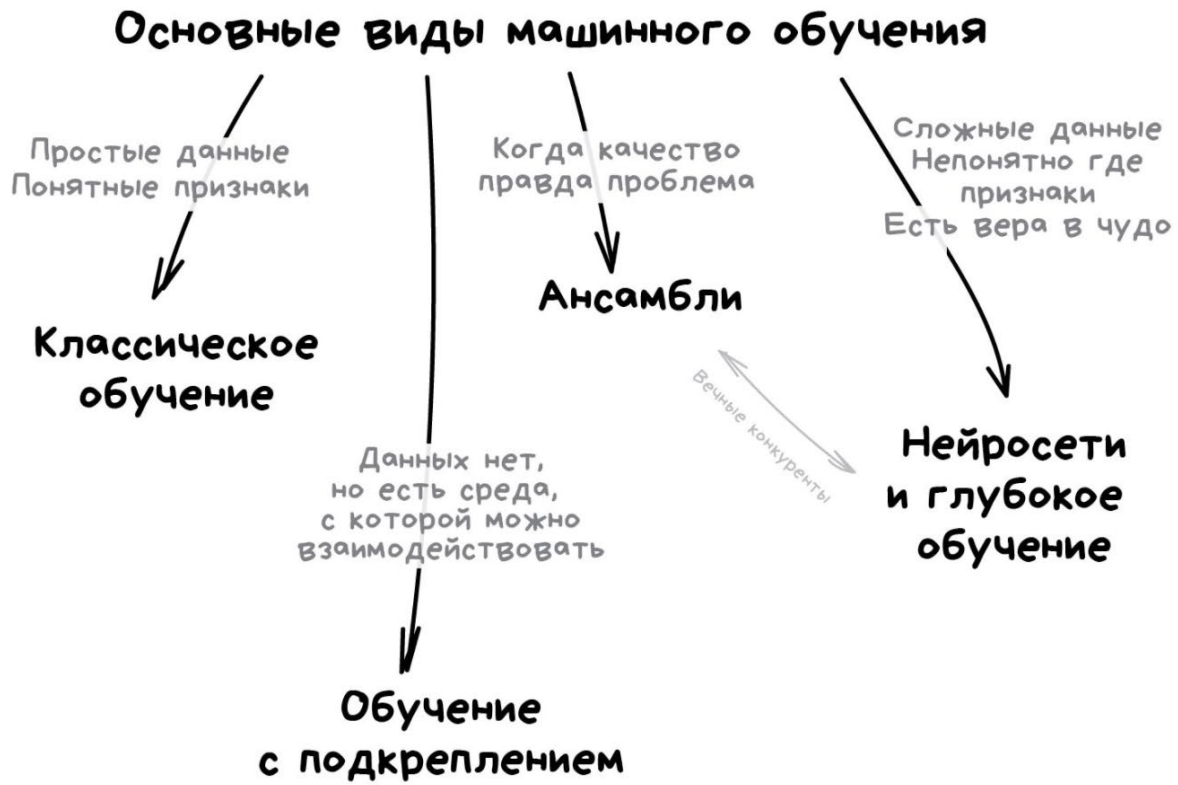
Убить всех людей

Вот о чем сегодня будем говорить ->

В следующем семестре будем изучать (если нам позволят):

- Базовые алгоритмы МО
  - Классификация
  - Регрессия
- Понятие разбиения выборки
- Метрики

# ОСНОВНЫЕ ВИДЫ МО



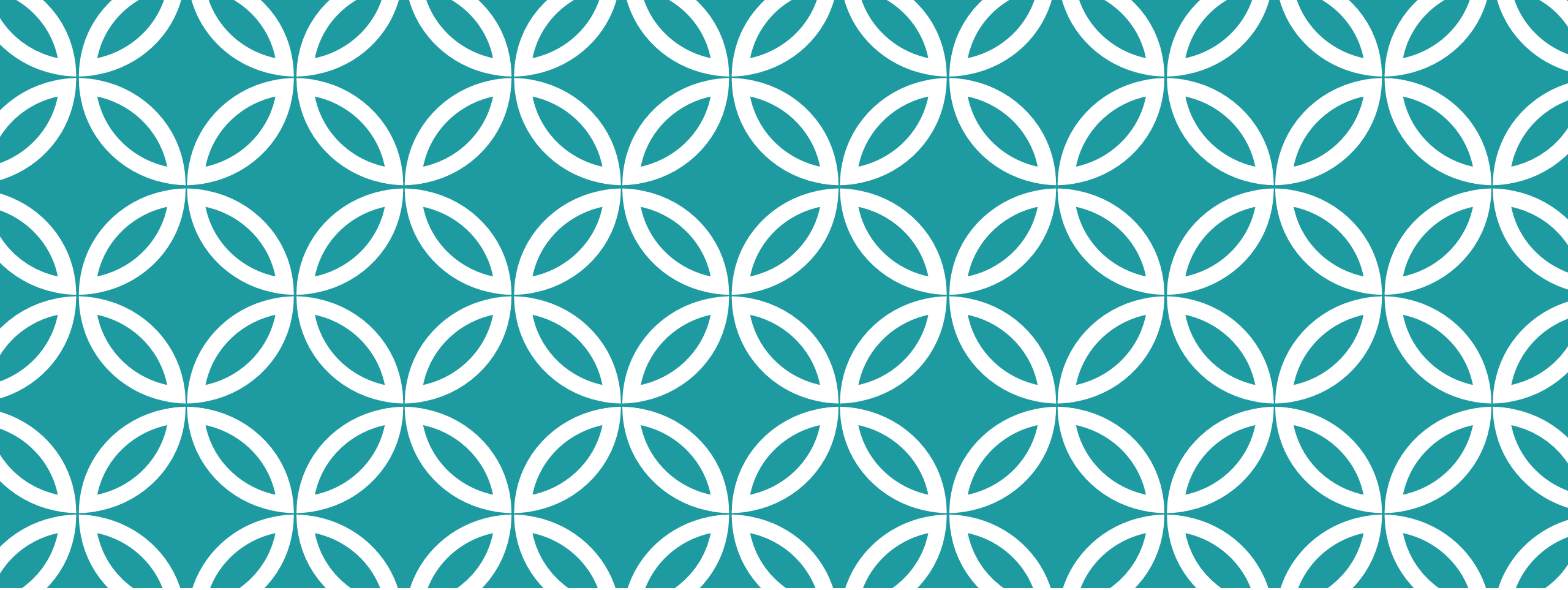
# КЛАССИЧЕСКОЕ ОБУЧЕНИЕ

Мы обсудим пока только обучение с учителем. Оно делится на два вида задач:

**Классификация** — предсказание категории объекта
**Регрессия** — предсказание места на числовой прямой

Основная библиотека:

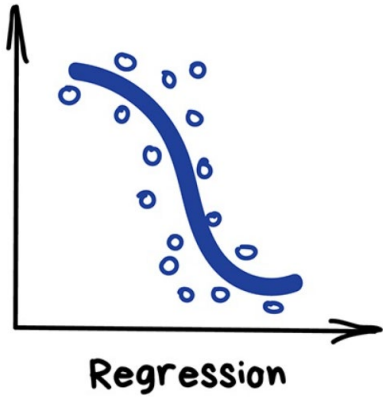
Scikit-learn



# ОБУЧЕНИЕ С УЧИТЕЛЕМ

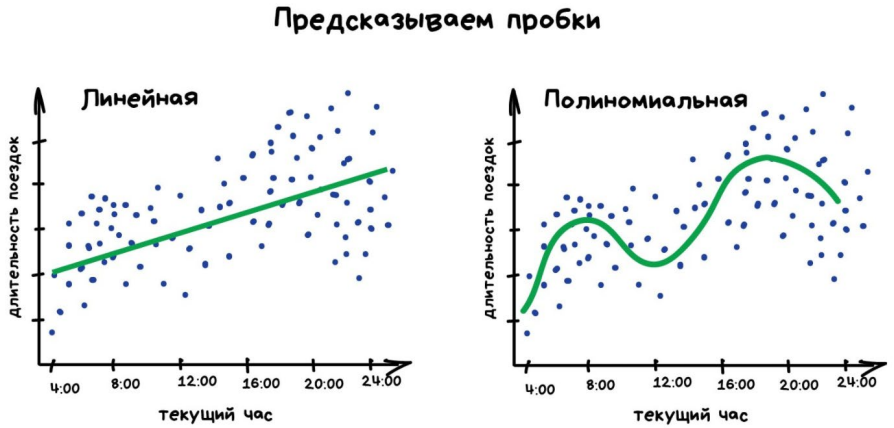
Много размеченных данных, еще больше размеченных данных! ...Рабы-практиканты

# РЕГРЕССИЯ




Сегодня используют для:

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

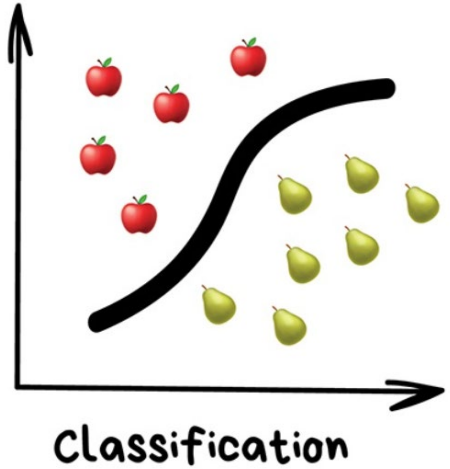



Регрессия

Используемые алгоритмы:

- Линейная регрессия
- Полиномиальная регрессия
- Ридж/Лассо (с регуляризацией весов)

# КЛАССИФИКАЦИЯ




Сегодня используют для:

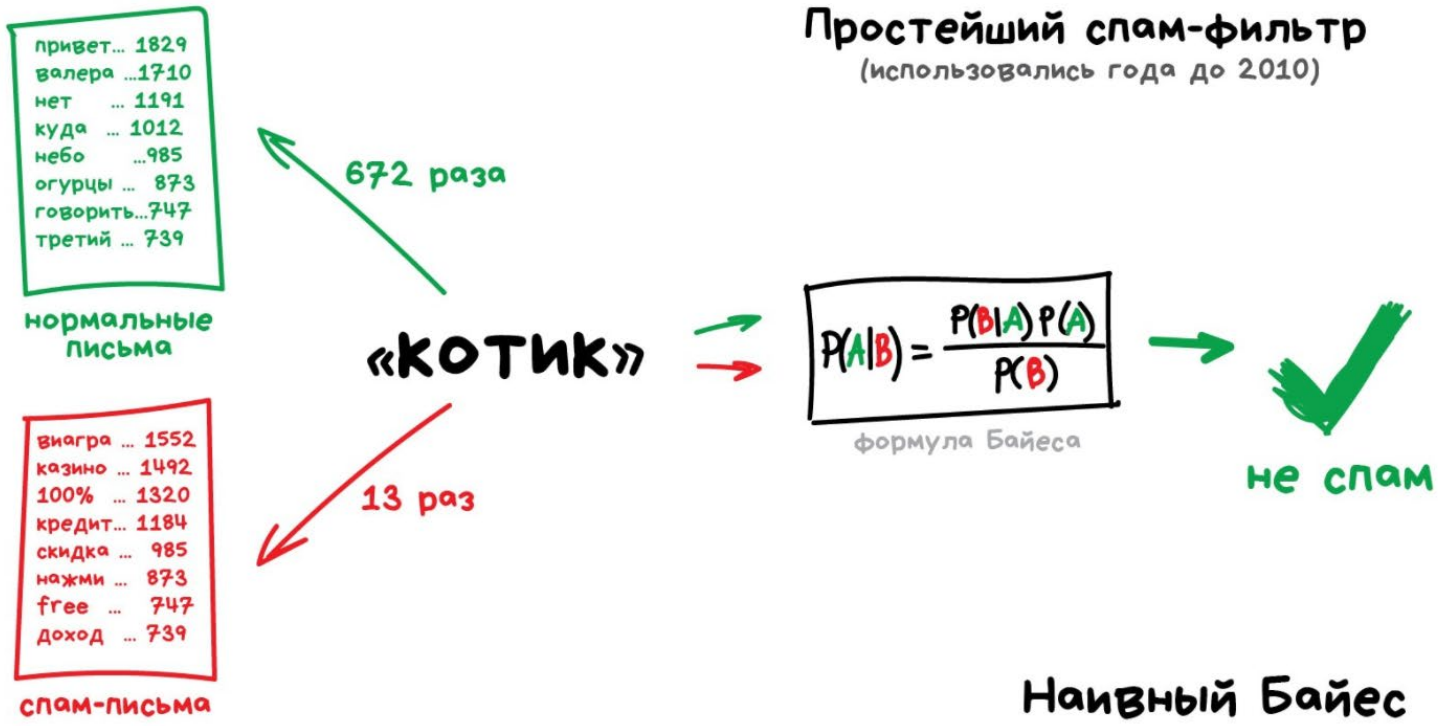
- ✓ Спам-фильтры
- ✓ Определение языка
- ✓ Поиск похожих документов
- ✓ Анализ тональности
- ✓ Распознавание рукописных букв и цифр
- ✓ Определение подозрительных транзакций

Используемые алгоритмы:

- Наивный Байес
- Деревья Решений
- Логистическая Регрессия
- К-ближайших соседей
- Метод Опорных Векторов

# НАИВНЫЙ БАЙЕС

Хорошо работает с текстами!



# ДЕРЕВЬЯ РЕШЕНИЙ

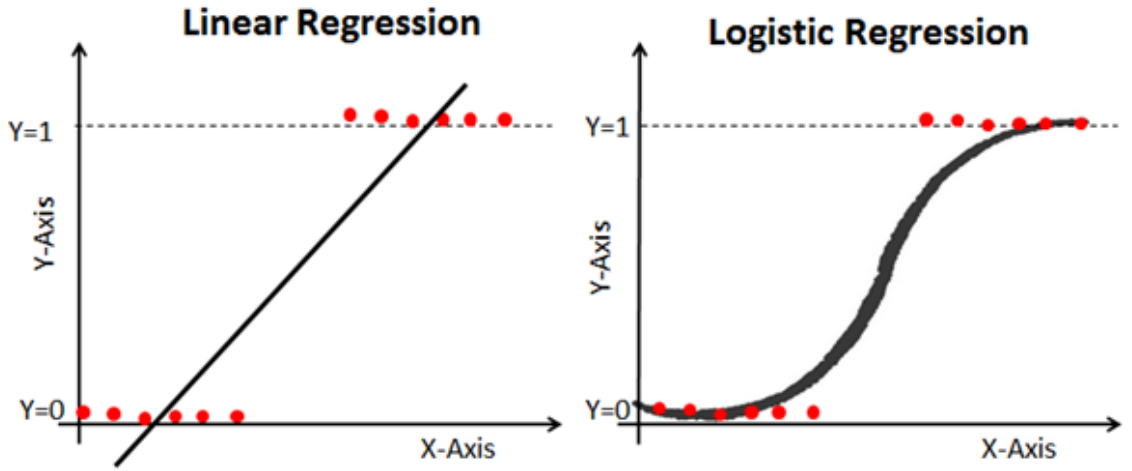
Машина автоматически разделяет все данные по вопросам, ответы на которые «да» или «нет». Вопросы могут быть не совсем адекватными с точки зрения человека, например «*зарплата заёмщика больше, чем 25934 рубля?*», но машина придумывает их так, чтобы на каждом шаге разбиение было самым точным.


Дерево Решений

# ЛОГИСТИЧЕСКАЯ РЕГРЕССИЯ

Предсказывает вероятность класса.
Является классификатором!

Результаты регрессии пропускаются через сигмоиду, которая все загоняет в интервал от 0 до 1.



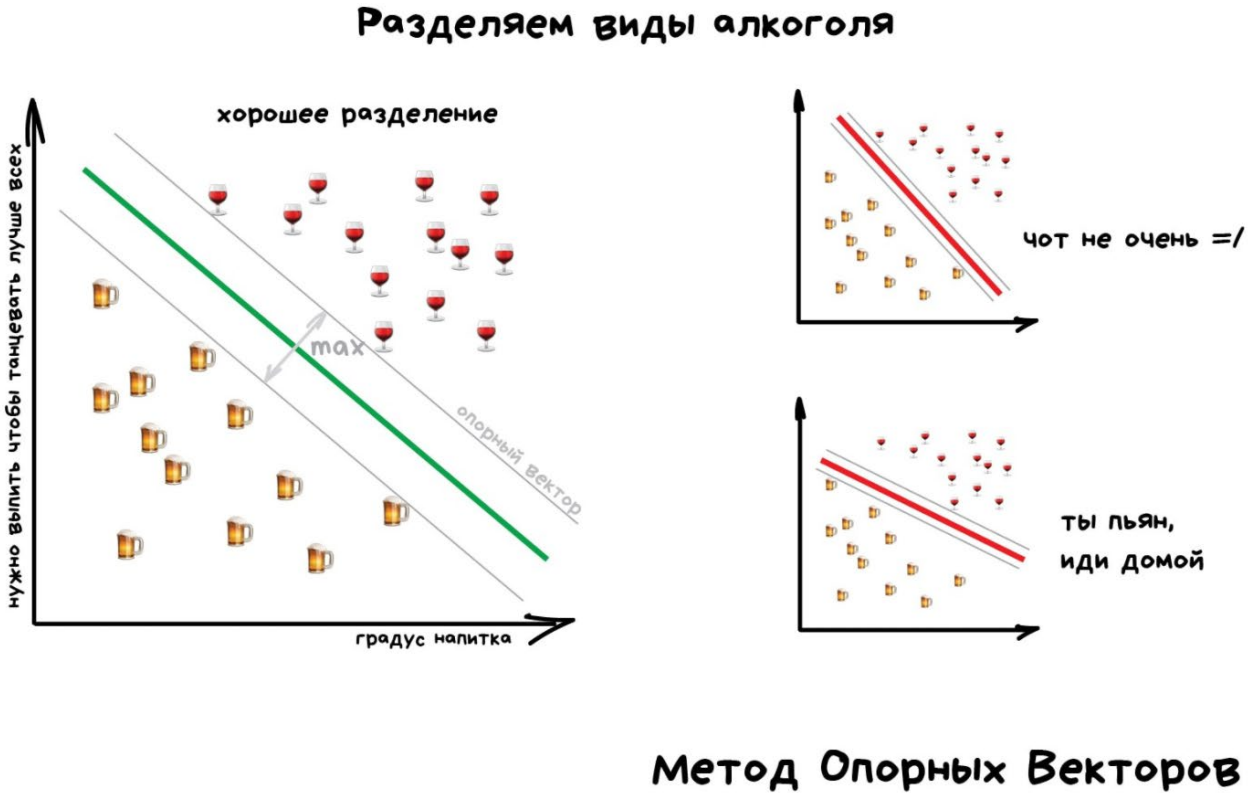
# МЕТОД ОПОРНЫХ ВЕКТОРОВ

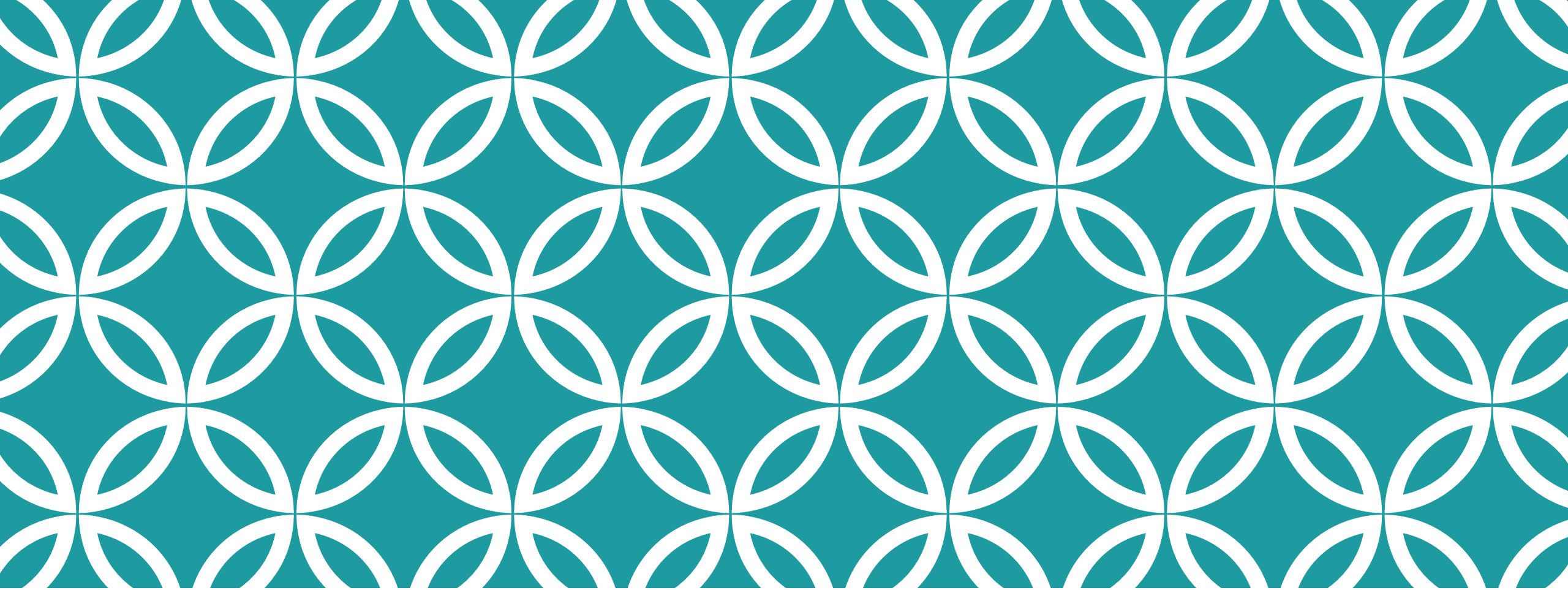
Ищет, как так провести две прямые между категориями, чтобы между ними образовался наибольший зазор.

Полезен (как и любая классификация) для поиска аномалий.

Можно сделать классификатор лиц! (но зачем?..)

На нем до сих пор работают спам-фильтры.





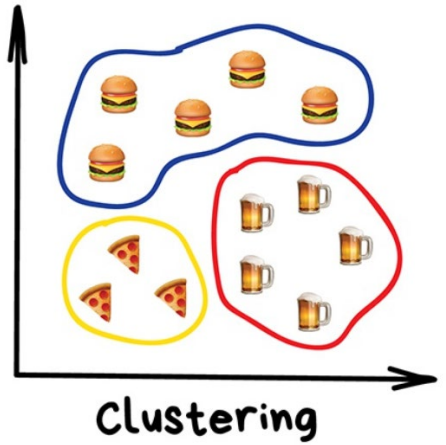
# ОБУЧЕНИЕ БЕЗ УЧИТЕЛЯ

Обычно используется для анализа данных, потому что работает так себе

# КЛАСТЕРИЗАЦИЯ

Сегодня используют для:

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

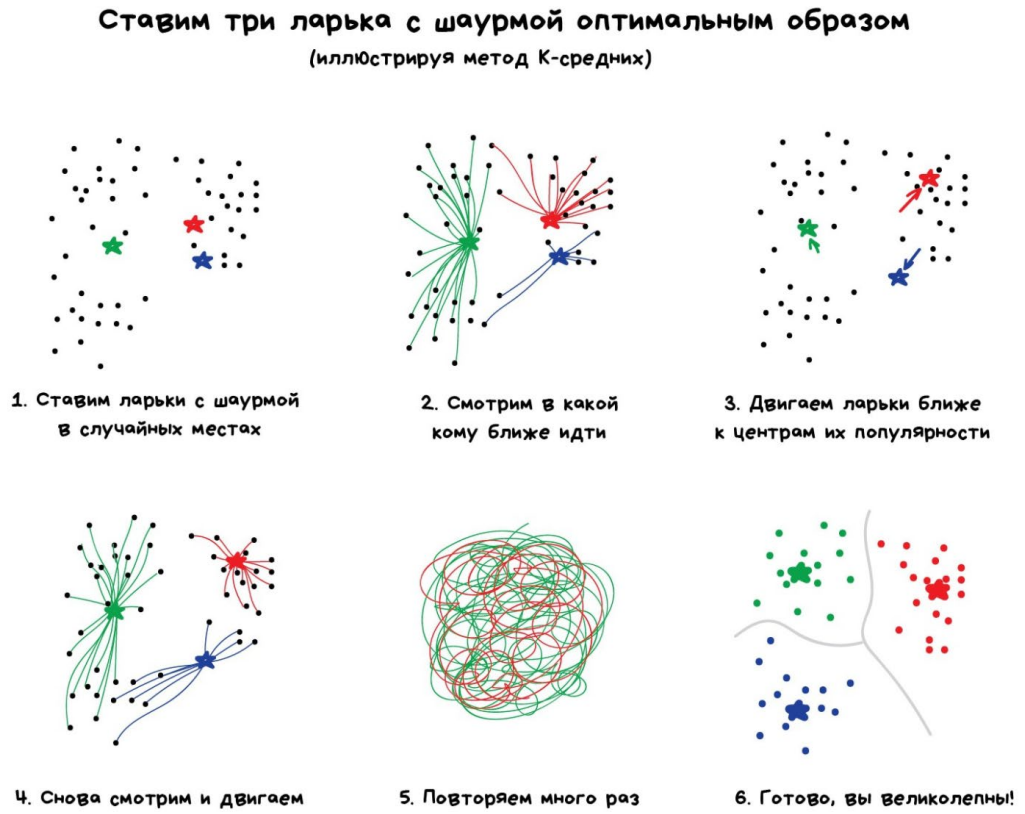



Используемые алгоритмы:

- ✓ Метод K-средних
- ✓ Mean-Shift
- ✓ DBSCAN

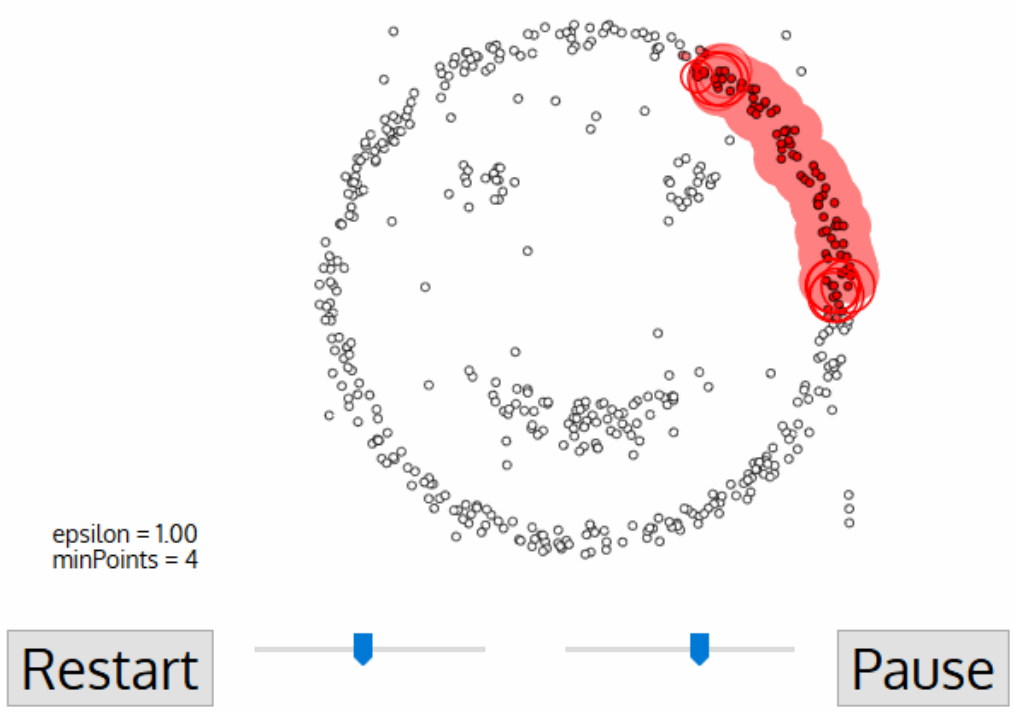
# МЕТОД К-СРЕДНИХ

Лагутин это будет вам объяснять в следующем семестре (подробно!)

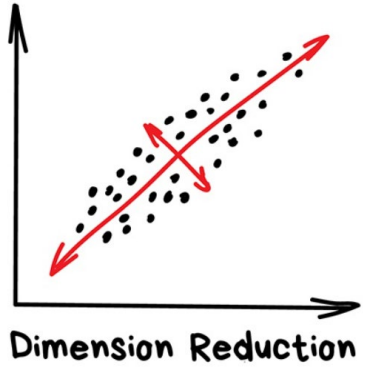


# DBSCAN

Точки — это люди на площади. Находим три любых близко стоящих человека и говорим им взяться за руки. Затем они начинают брать за руку тех, до кого могут дотянуться. Так по цепочке, пока никто больше не сможет взять кого-то за руку — это и будет первый кластер. Повторяем, пока не поделим всех. Те, кому вообще некого брать за руку — это выбросы, аномалии.



# УМЕНЬШЕНИЕ РАЗМЕРНОСТИ




Сегодня используют для:

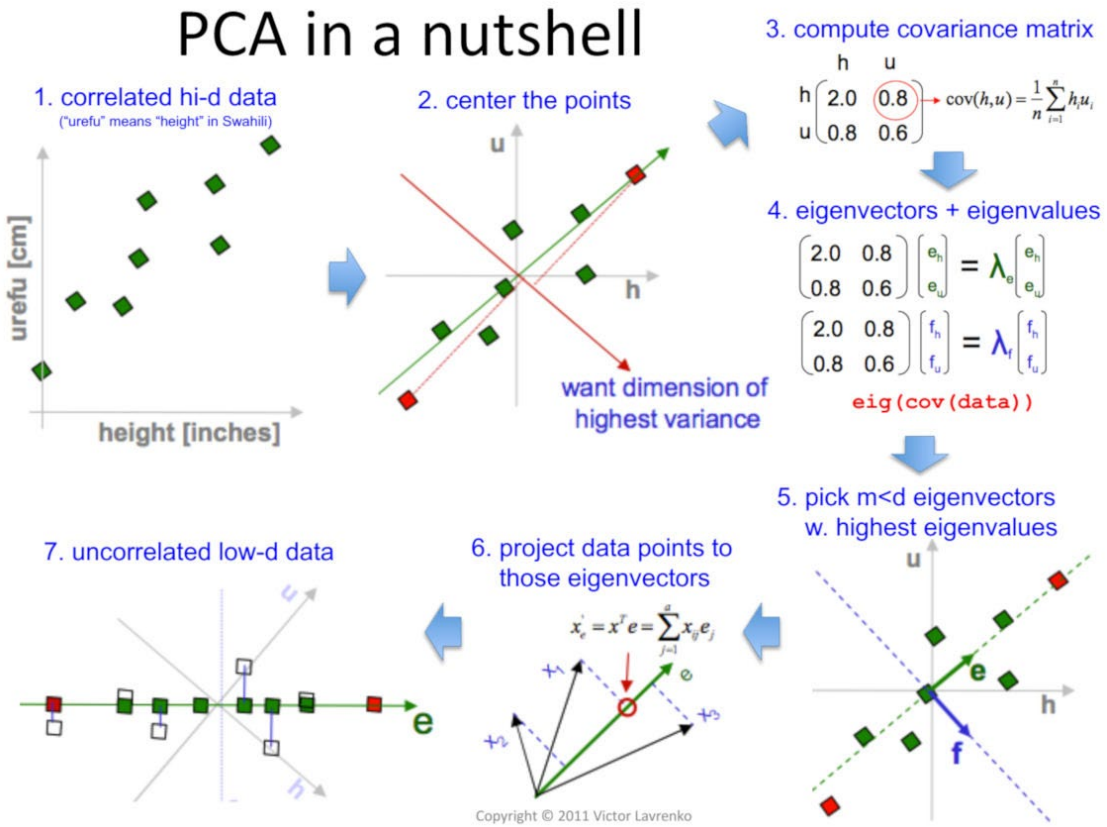
- ✓ Рекомендательные Системы
- ✓ Красивые визуализации
- ✓ Определение тематики и поиска похожих документов
- ✓ Анализ фейковых изображений
- ✓ Риск-менеджмент

Используемые алгоритмы:

- ✓ Метод главных компонент (PCA)
- ✓ Сингулярное разложение (SVD)
- ✓ Латентное размещение Дирихле (LDA)
- ✓ Латентно-семантический анализ (LSA, pLSA, GLSA)
- ✓ t-SNE (для визуализации)

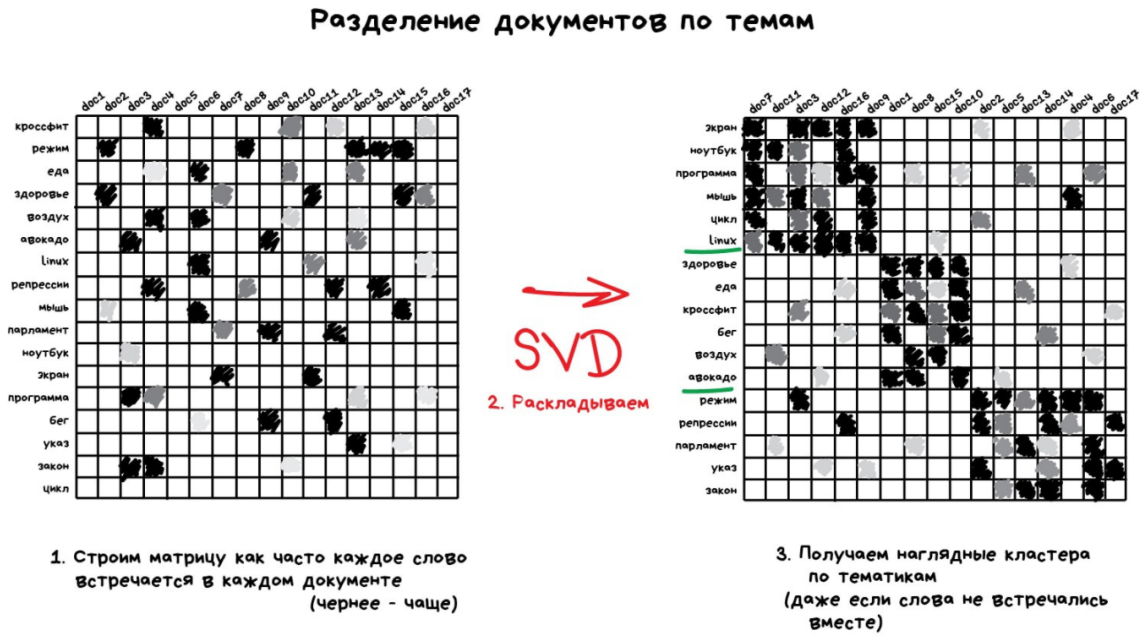
# МЕТОД ГЛАВНЫХ КОМПОНЕНТ

Хорошо объясняет Лагутин (в конце 3 семестра)



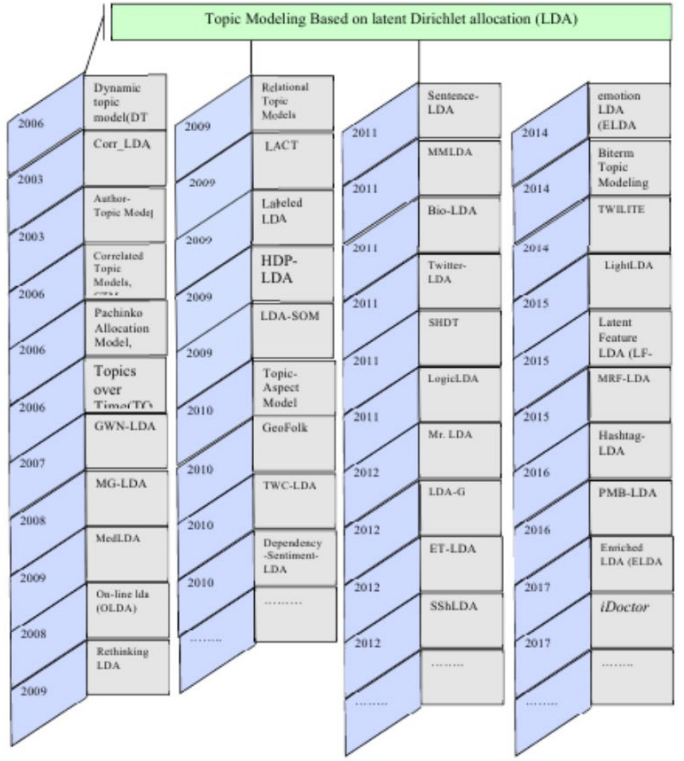
# SVD

В математические детали лучше не вдаваться (страшные), но на основе сингулярного разложения матриц когда-то делали эмбеддинги из слов (до 2013 года, когда придумали word2vec), а заодно и латентно-семантический анализ с ее помощью можно делать



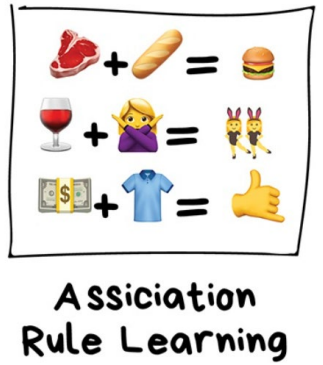
# LDA

Позволяет с удивительно неплохим качеством извлекать из корпуса текстов списки слов, наиболее характерных для определенной тематики. Требует только указывать количество желаемых тем (кластеров).




1. автор издание искусство картина литература литературный отец перевод писать письмо поэт произведение рассказ роман стих творчество художник
2. александр владимир генерал деятель иван князь михаил народный николай орден память писатель польский пётр р. род. ум.
3. где задержанный заявление имя мвд нарушение обвинение отмечать отношение полицейский преступление район следствие случай сми территория убийство уголовный центр январь
4. актриса актёр альбом выйти герой кино концерт музыка музыкальный музыкант персонаж песня премия режиссёр роль сериал серия сцена театр фильм
5. вечер взять далёкий дверь домой ехать ждать мама минута неделя нога ночь остаться пара прийти рядом сидеть увидеть ходить

# ПОИСК ПРАВИЛ (АССОЦИАЦИЯ)

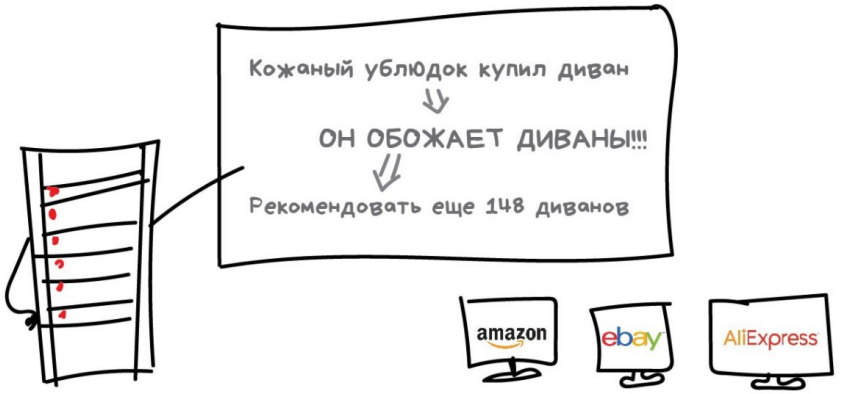



Сегодня используют для:

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

Используемые алгоритмы:

- ✓ Apriori
- ✓ Euclat
- ✓ FP-growth



# ОБУЧЕНИЕ С ПОДКРЕПЛЕНИЕМ

Сегодня используют для:

- ✓ Самоуправляемых автомобилей
- ✓ Роботов пылесосов
- ✓ Игр
- ✓ Автоматической торговли
- ✓ Управления ресурсами предприятий

Используемые алгоритмы:

- ✓ Q-Learning
- ✓ SARSA
- ✓ DQN
- ✓ A3C
- ✓ Генетический Алгоритм



Цель алгоритма – минимизировать ошибки, а не рассчитать все ходы.
Такой алгоритм обыграл человека в го, где все ходы просчитать невозможно.

# Q-LEARNING

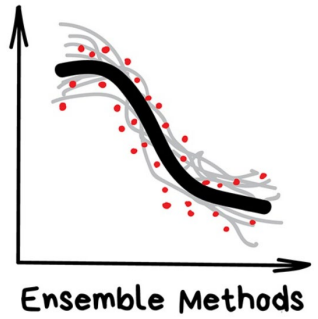
Машина выбирает лучший выход из каждой ситуации.

Эта идея лежит в основе алгоритма Q-learning и его производных (SARSA и DQN). Буква Q в названии означает слово Quality, то есть робот учится поступать наиболее качественно в любой ситуации, а все ситуации он запоминает как простой марковский процесс.

DQN – это Q-Learning на нейронках.

# АНСАМБЛИ

*Несколько алгоритмов дополняют друг друга*




Сегодня используют для:

- ✓ Всего, где подходят классические алгоритмы (но работают точнее)
- ✓ Поисковые системы
- ✓ Компьютерное зрение
- ✓ Распознавание объектов

Используемые алгоритмы:

- ✓ Random Forest
- ✓ Gradient Boosting

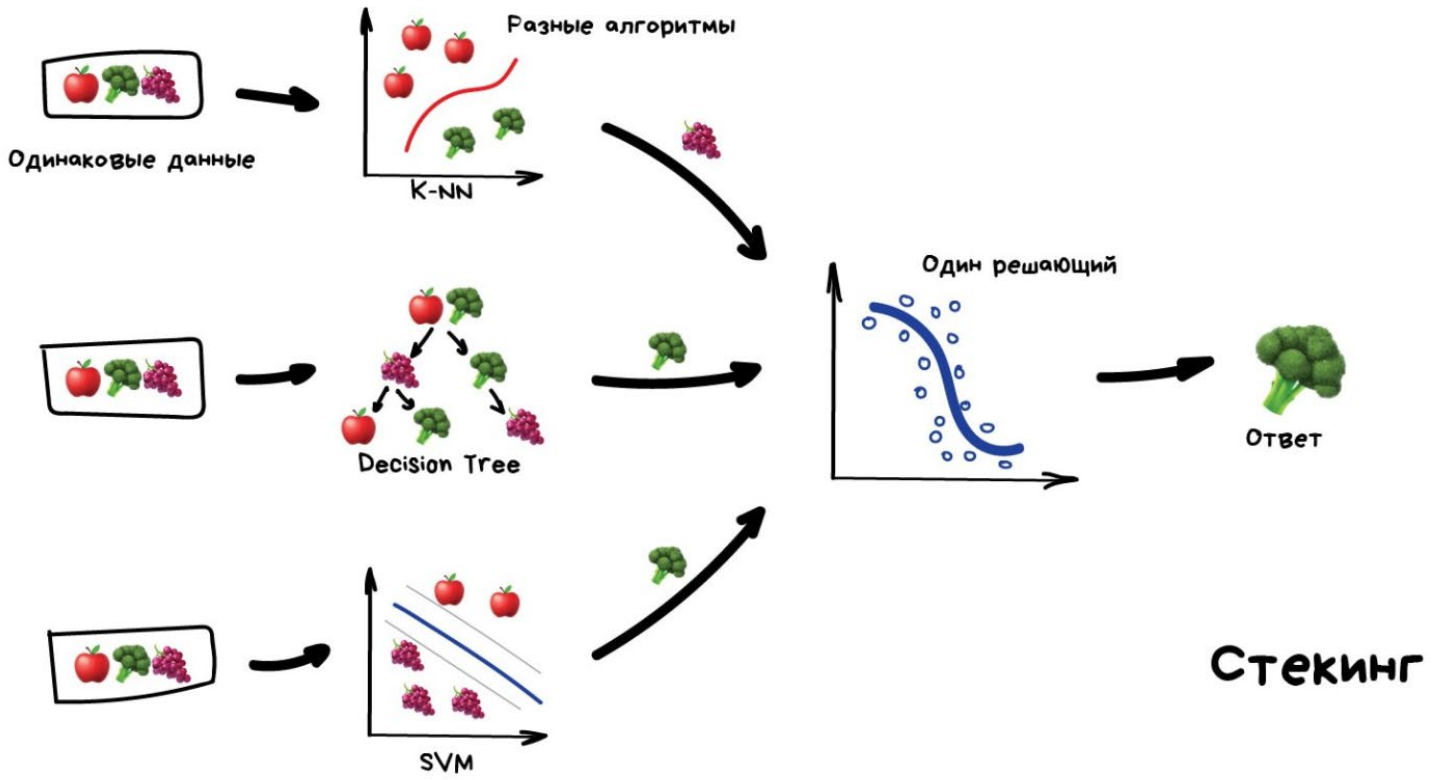
Ансамбль можно собрать как угодно, хоть случайно нарезать в тазик классификаторы и залить регрессией. За точность, правда, тогда никто не ручается. Потому есть три проверенных способа делать ансамбли:

- ✓ Стекинг
- ✓ Беггинг
- ✓ Бустинг

# СТЕКИНГ

Алгоритмы должны быть разные, а один – решающий на основе их ответов.

На практике применяется реже двух других, потому что менее точный.

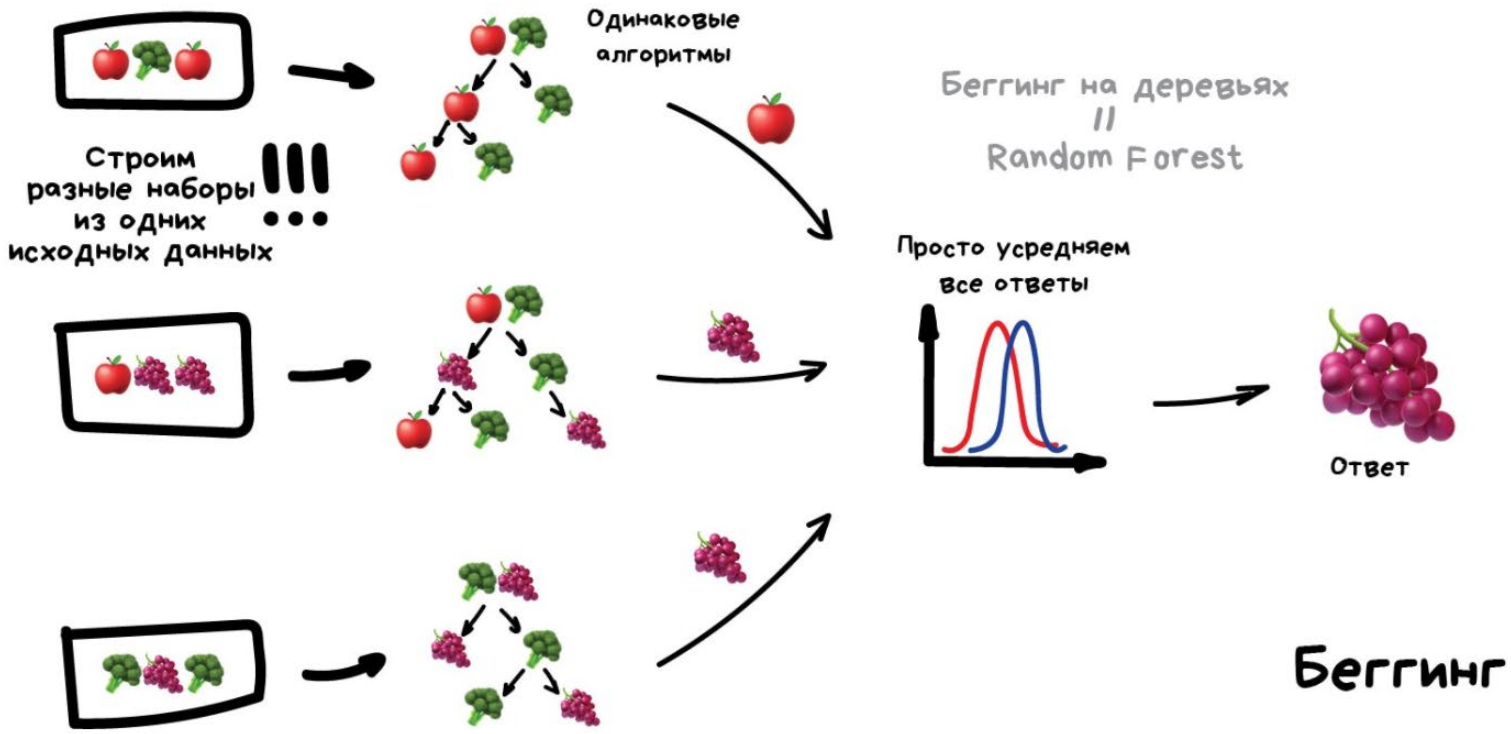


# БЕГГИНГ (BOOTSTRAP AGGREGATING)

Обучаем один алгоритм много раз на случайных выборках из исходных данных. В самом конце усредняем ответы.

Random Forest – самый популярный алгоритм беггинга (он на картинке)

Можно гонять параллельно => профит!

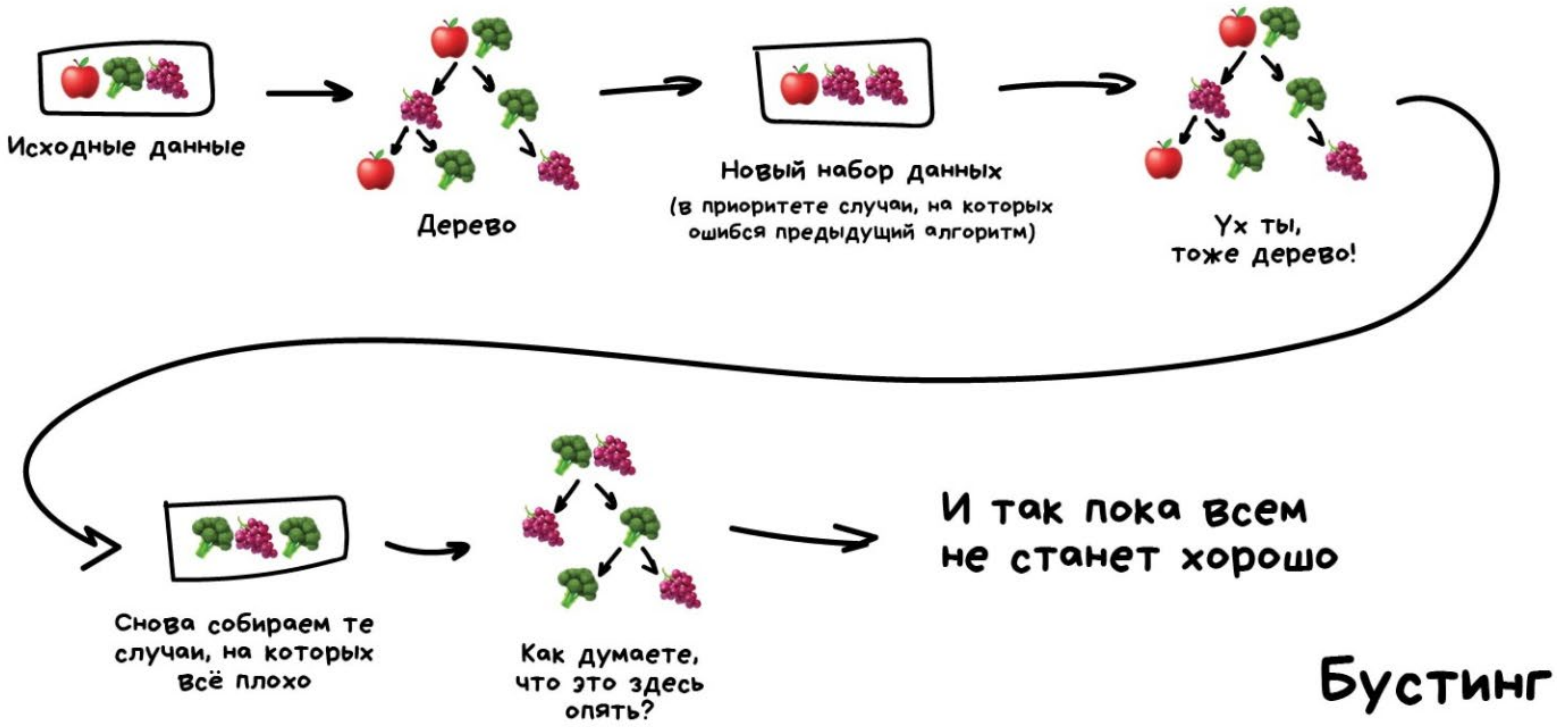


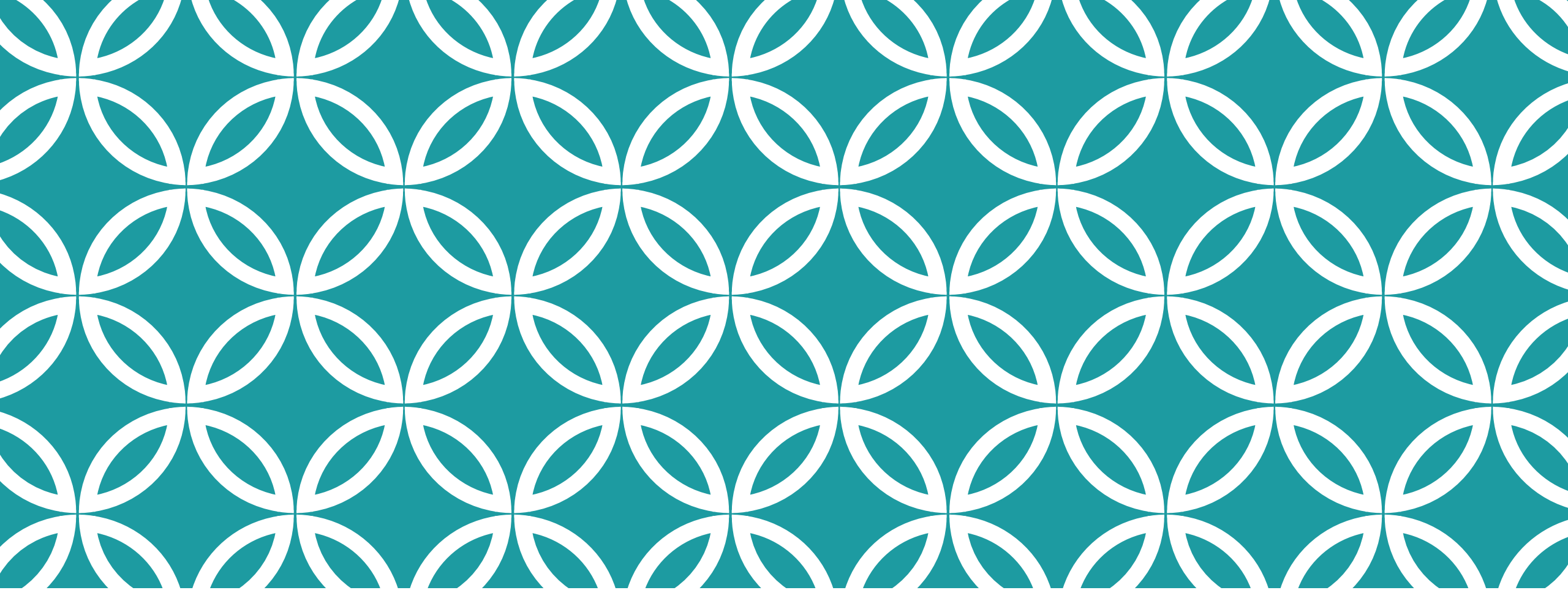
# БУСТИНГ

Обучаем алгоритмы последовательно, каждый следующий уделяет особое внимание тем случаям, на которых ошибся предыдущий.

Очень точный. На нем работает поисковик Яндекса.

Подробно можно почитать: https://towardsdatascience.com/catboost-vs-light-gbm-vs-xgboost-5f93620723db

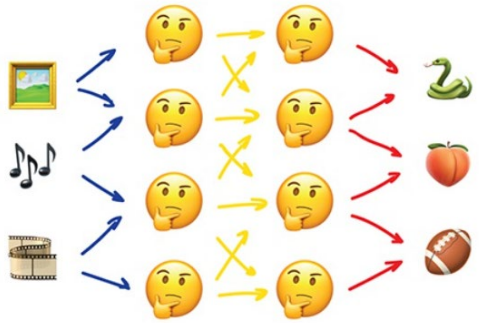




# НЕЙРОСЕТИ И ГЛУБОКОЕ ОБУЧЕНИЕ

Какое обучение самое глубокое, уже никто точно не знает, просто говорят о разных архитектурах

# НЕЙРОСЕТИ




Сегодня используют для:

- ✓ Вместо всех вышеперечисленных алгоритмов вообще
- ✓ Определение объектов на фото и видео
- ✓ Распознавание и синтез речи
- ✓ Обработка изображений, перенос стиля
- ✓ Машинный перевод

Популярные архитектуры:

- ✓ Перцептрон
- ✓ Свёрточные Сети (CNN)
- ✓ Рекуррентные Сети (RNN)
- ✓ Генеративно-состязательные (GAN)
- ✓ Автоэнкодеры
- ✓ Трансформеры

Библиотеки:

- Keras
- Tensorflow
- PyTorch
- AllenNLP

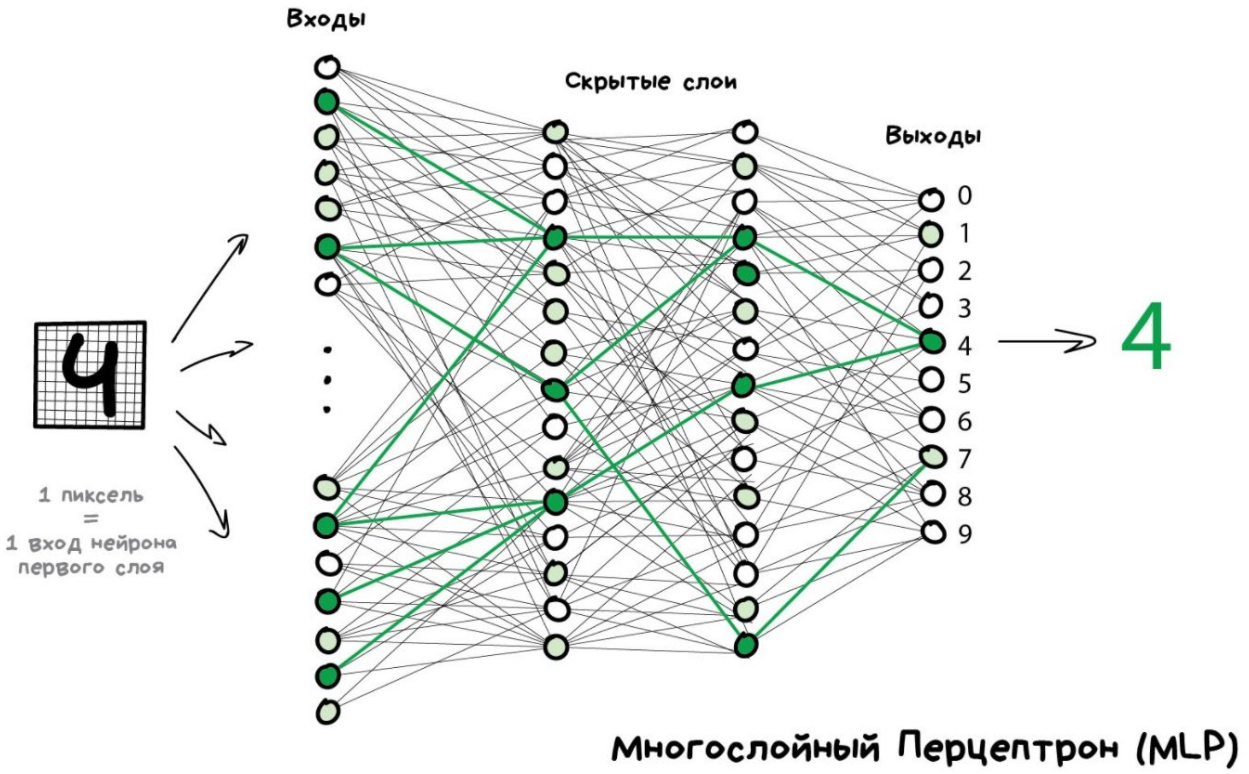
# ПЕРЦЕПТРОН

Был придуман 80 лет назад Ф. Розенблаттом, который написал книжку, а эту книжку потом критиковал Минский с коллегами: Минский доказал, что нейронные сети не могут предсказывать нелинейную зависимость (это потом обошли).

Используется для:

- ✓ Обучение студентов

Всплески радости и волны отчаяния можно посмотреть на Википедии: https://en.wikipedia.org/wiki/Timeline_of_machine_learning



# CNN


Используются для:

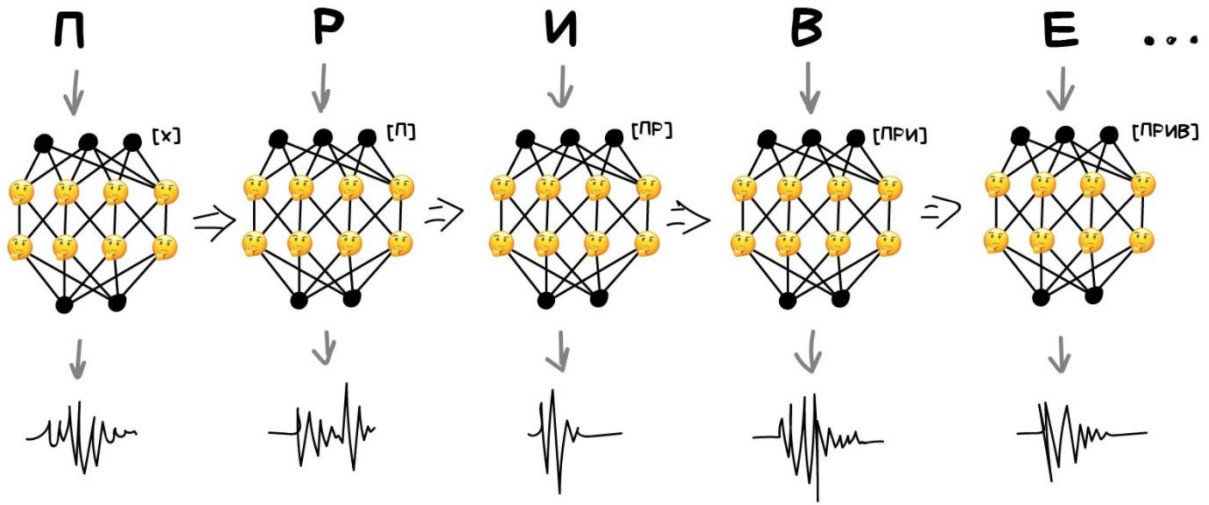
- ✓ поиск объектов на фото и видео
- ✓ распознавание лиц
- ✓ перенос стиля
- ✓ генерация и дорисовка изображений
- ✓ создание эффектов типа слоу-мо
- ✓ улучшение качества фотографий

Изображение делится на маленькие кусочки размером 8x8 пикселей, для каждого кусочка выбирается доминирующая линия: горизонтальная, вертикальная или диагональная. Потом берутся кусочки побольше, нейронка смотрит, как на них сочетаются эти палочки и т.д. Это и есть свертка.

# RNN

Хорошо работает с любыми последовательностями, а значит, используется для:

- ✓ Машинного перевода
- ✓ Синтеза речи
- ✓ И других задач, связанных со звуковыми или текстовыми последовательностями…




Рекуррентная Нейросеть (RNN)

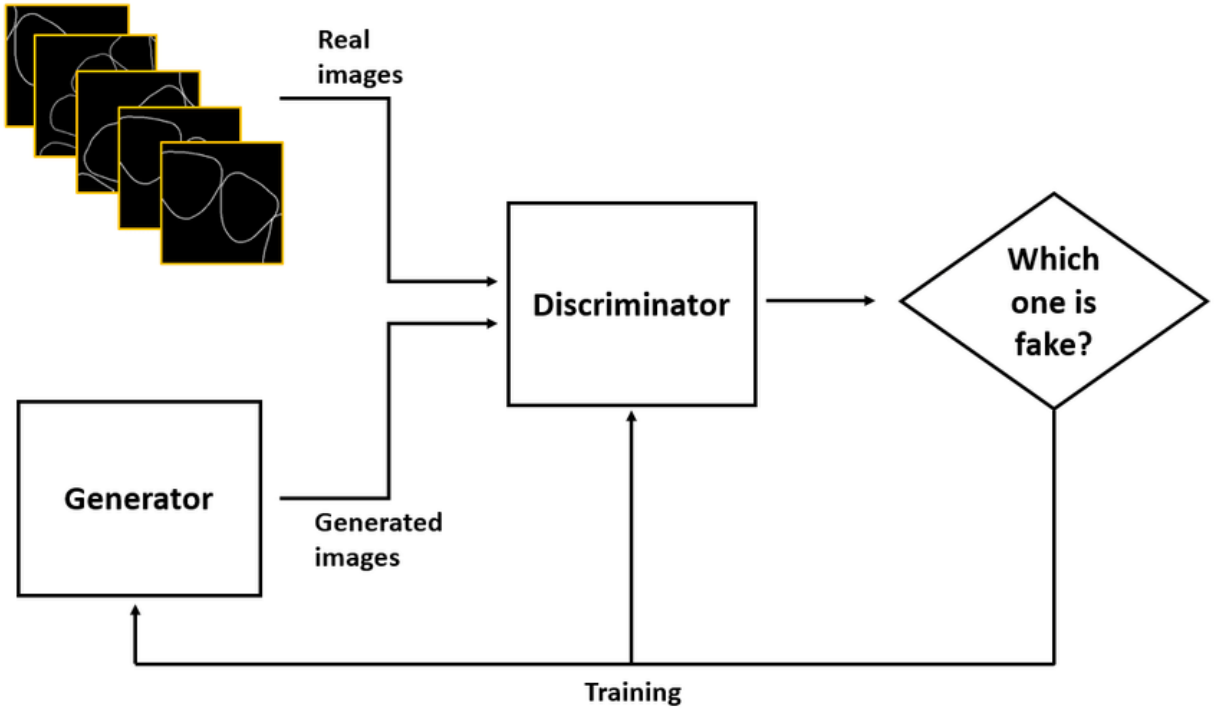
У рекуррентной сети есть память (LSTM – long-short term memory), поэтому, когда она учит последовательности, обращает внимание на предыдущие элементы.

# GAN

Состоит из двух частей: генератор и дискриминатор.

Генератор подсовывает дискриминатору фейковые сгенерированные картинки вместе с реальными, а дискриминатор должен угадать, где подделки.

Используется для генерации всякого разного.




Поиграем в дискриминаторы: какая из фоток фейковая?

# TRANSFORMERS

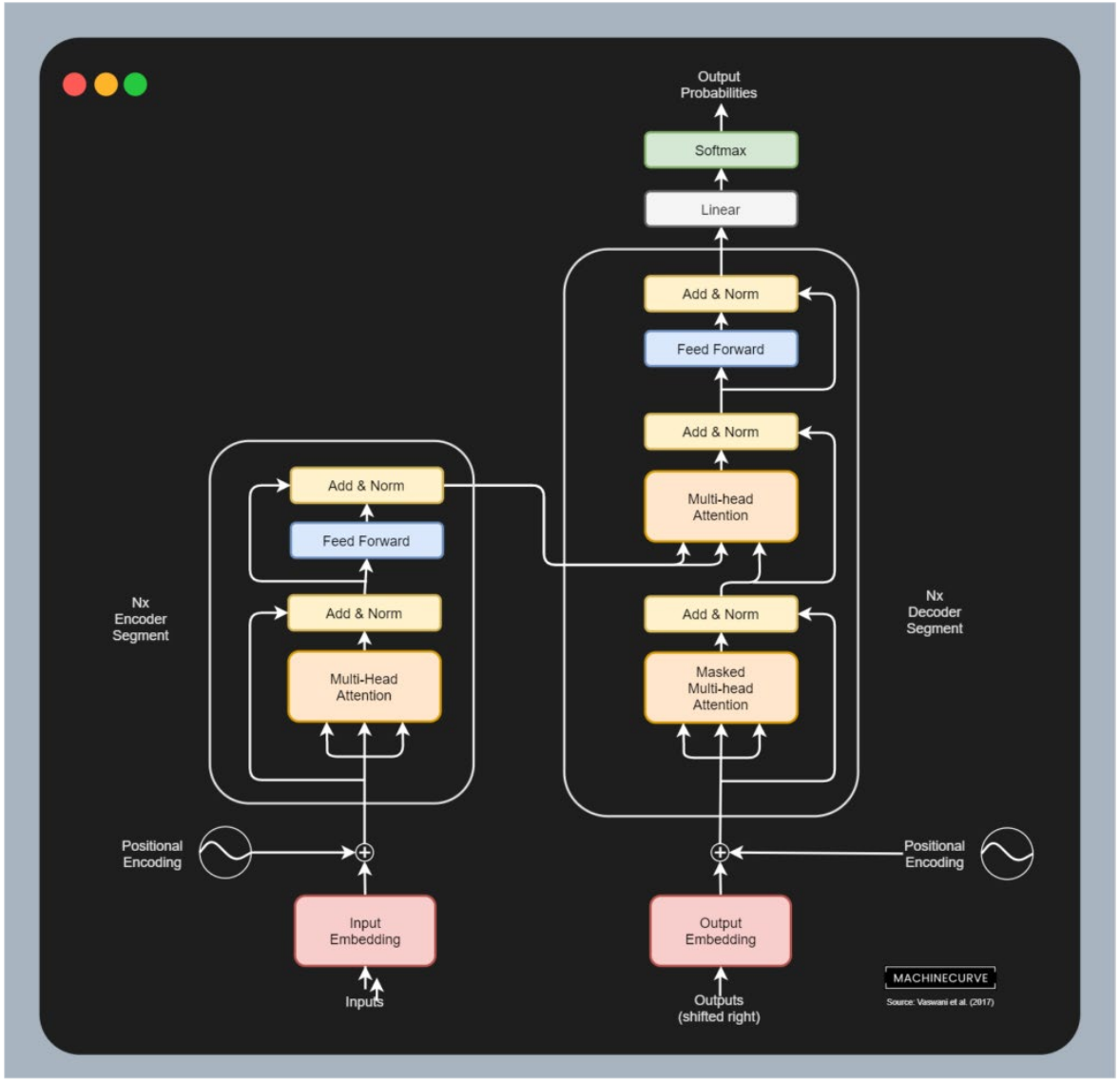
Появились в 2017 году.
Используются для:

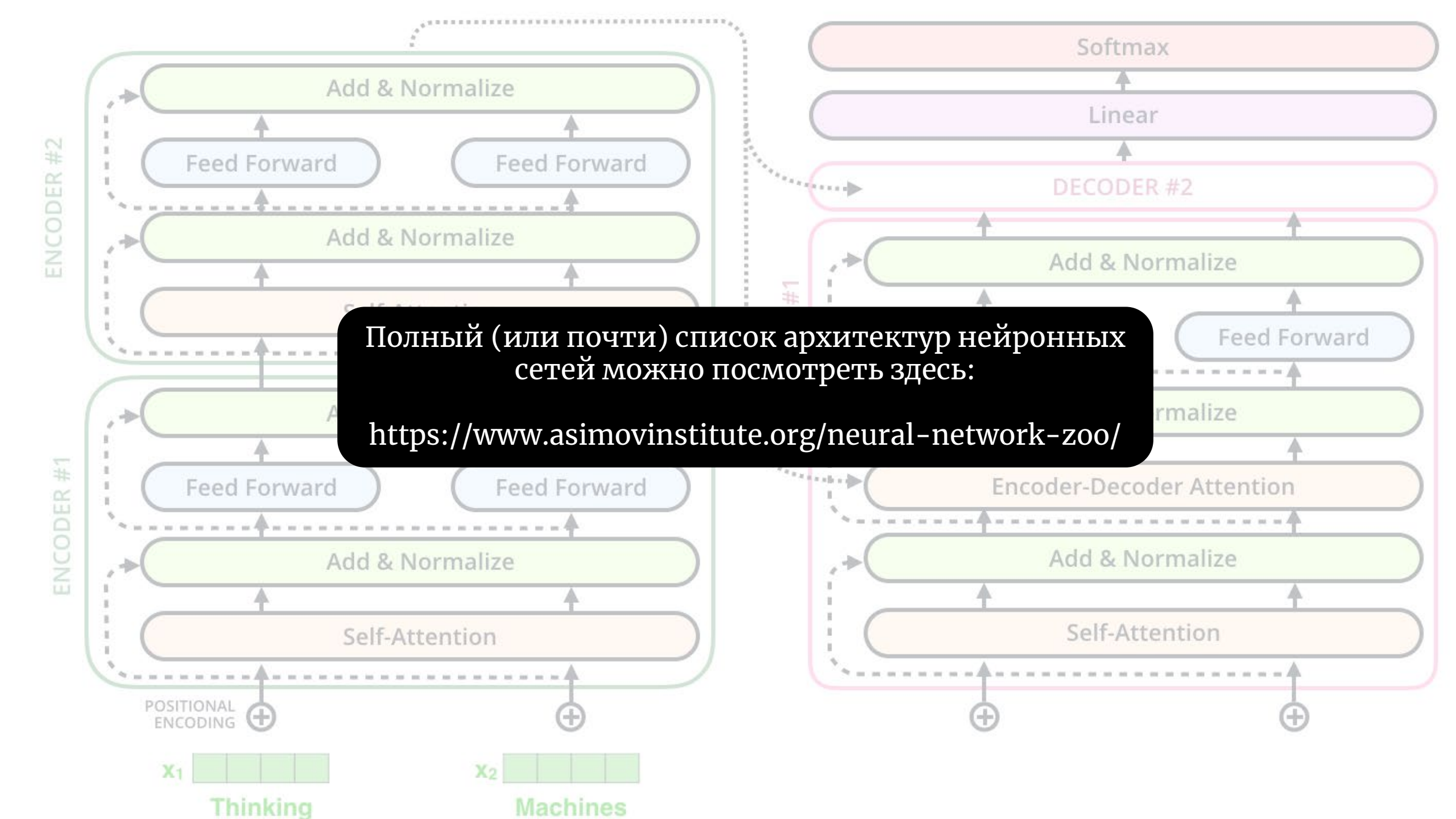
- ✓ NLP-задач
- ✓ Компьютерного зрения (CV)

Используют механизм внимания, который вычисляет веса всех элементов последовательности (это совсем сложное).

К ним относятся BERT & GPT.

Есть модуль питона transformers.






Полный (или почти) список архитектур нейронных сетей можно посмотреть здесь:

https://www.asimovinstitute.org/neural-network-zoo/