

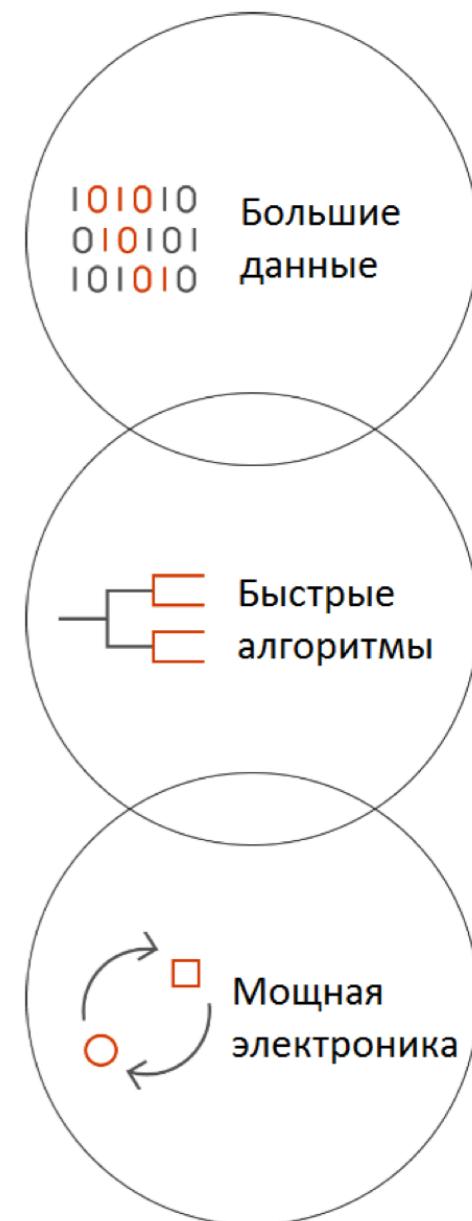
Deep Learning & Machine Learning

Баранов Максим Александрович

Три предпосылки бума ИИ

– три перехода количества в качество:

- Повсеместное применение компьютерных технологий
→ *накопление больших выборок данных*
в частности, ImageNet
- Развитие математических методов и алгоритмов
→ *накопление критической массы опыта*
в частности, Deep Neural Networks
- Достижения микроэлектроники
→ *рост вычислительных мощностей по закону Мура*
в частности, GPU



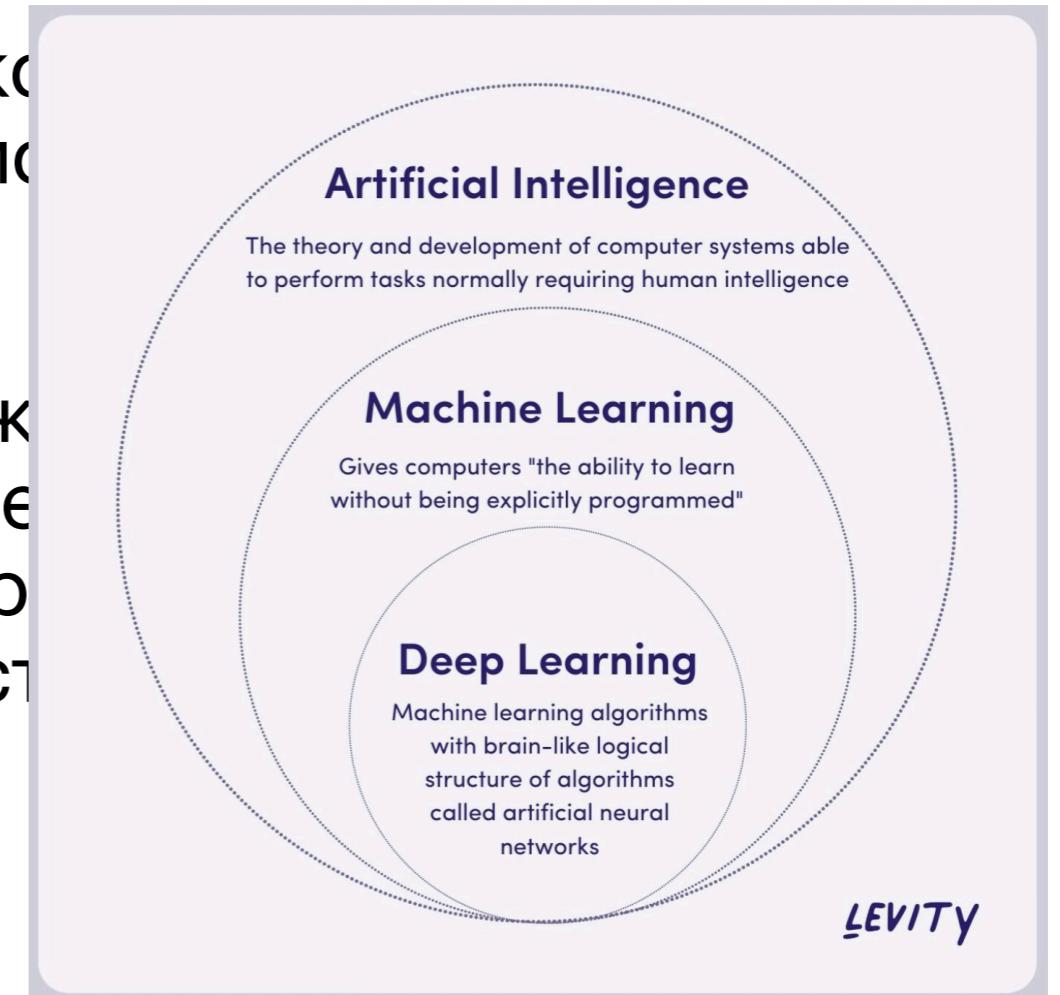
Основные выгоды ИИ

- Сокращение издержек и повышение производительности труда
- Автоматизация банковских и финансовых услуг (FinTech)
- Автоматизация юридических услуг (LegalTech)
- Автоматизация посреднической деятельности, распределённая экономика
- Роботизация производств, автономный транспорт
- Оптимизация логистики и цепей поставок
- Оптимизация энергетических и транспортных сетей
- Сенсорные сети, мониторинг сельского хозяйства
- Персональная медицина, улучшение клинических практик
- Персональные образовательные траектории, социальная инженерия
- Автономные системы вооружени

В чем разница?

Машинное обучение означает, что компьютеры обрабатывают данные с использованием алгоритмов, не требующих явного программирования.

Глубокое обучение использует сложные алгоритмы, смоделированные на человеческом мозге, что позволяет обрабатывать неструктурированные данные, такие как документы, изображения и тексты.



Машинное обучение

Машинное обучение — это общий термин, обозначающий, когда компьютеры учатся на данных. Он описывает пересечение информатики и статистики, где алгоритмы используются для выполнения конкретной задачи без явного программирования; вместо этого они распознают закономерности в данных и делают прогнозы после поступления новых данных.

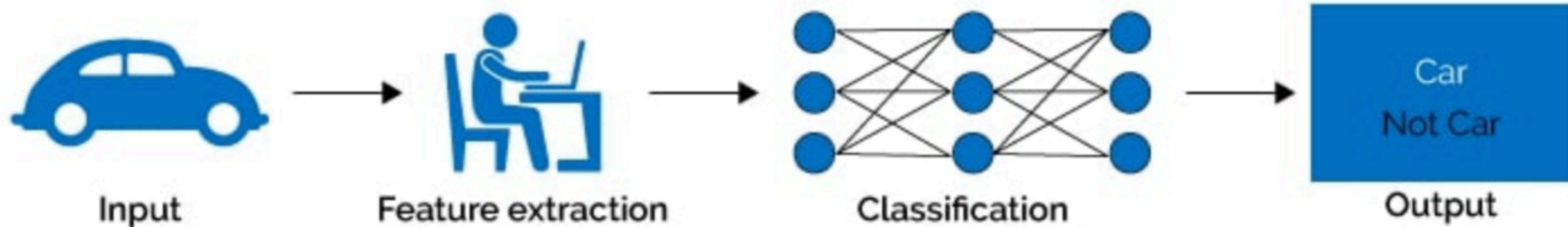
В общем, процесс обучения этих алгоритмов может быть контролируемым или неконтролируемым, в зависимости от данных, используемых для подачи алгоритмов.

Глубокое обучение

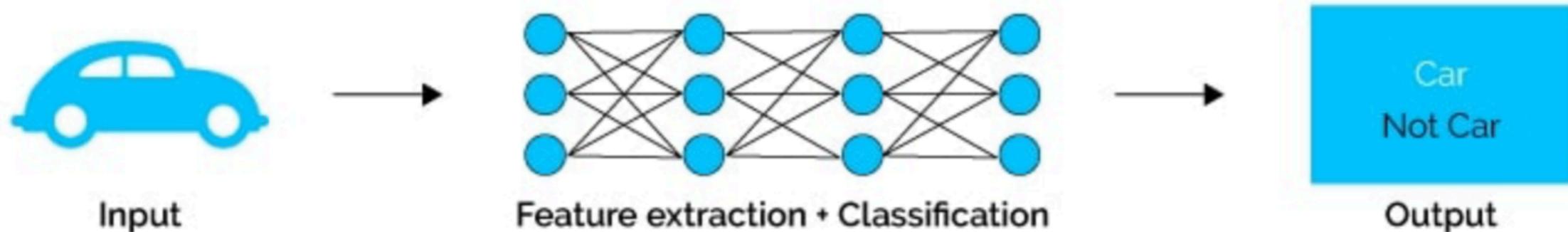
Алгоритмы глубокого обучения можно рассматривать как сложную и математически сложную эволюцию алгоритмов машинного обучения. В последнее время этой области уделяется много внимания, и не зря: недавние разработки привели к результатам, которые раньше считались невозможными.

Глубокое обучение описывает алгоритмы, которые анализируют данные с логической структурой, аналогичной тому, как человек делает выводы. Обратите внимание, что это может произойти как при контролируемом, так и при неконтролируемом обучении. Для достижения этой цели приложения глубокого обучения используют многоуровневую структуру алгоритмов, называемую искусственной нейронной сетью (ИНС).

Machine Learning

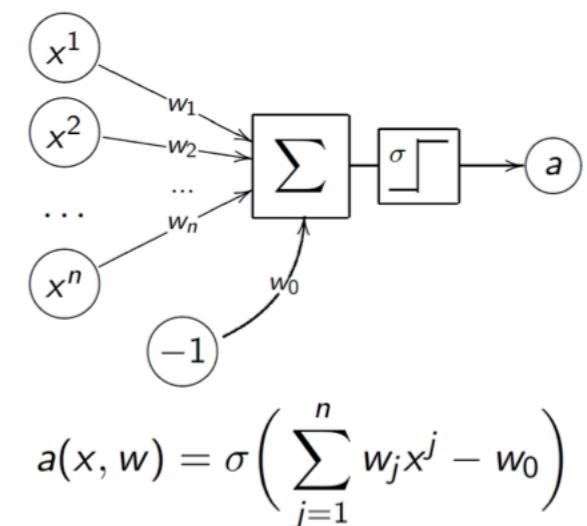
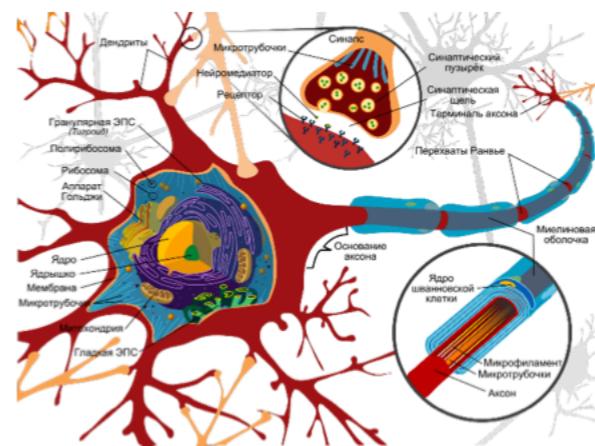


Deep Learning

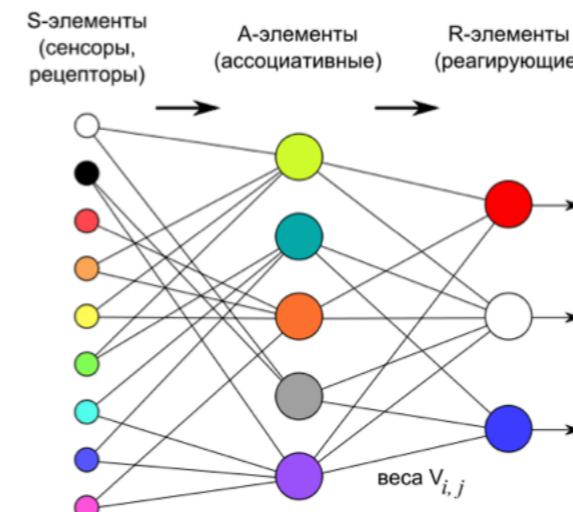


Что такое «искусственные нейронные сети»

Математическая модель нейрона
(МакКаллок и Питтс, 1943)



Первый нейрокомпьютер Mark-1
(Фрэнк Розенблatt, 1960)



Глубокие нейронные сети

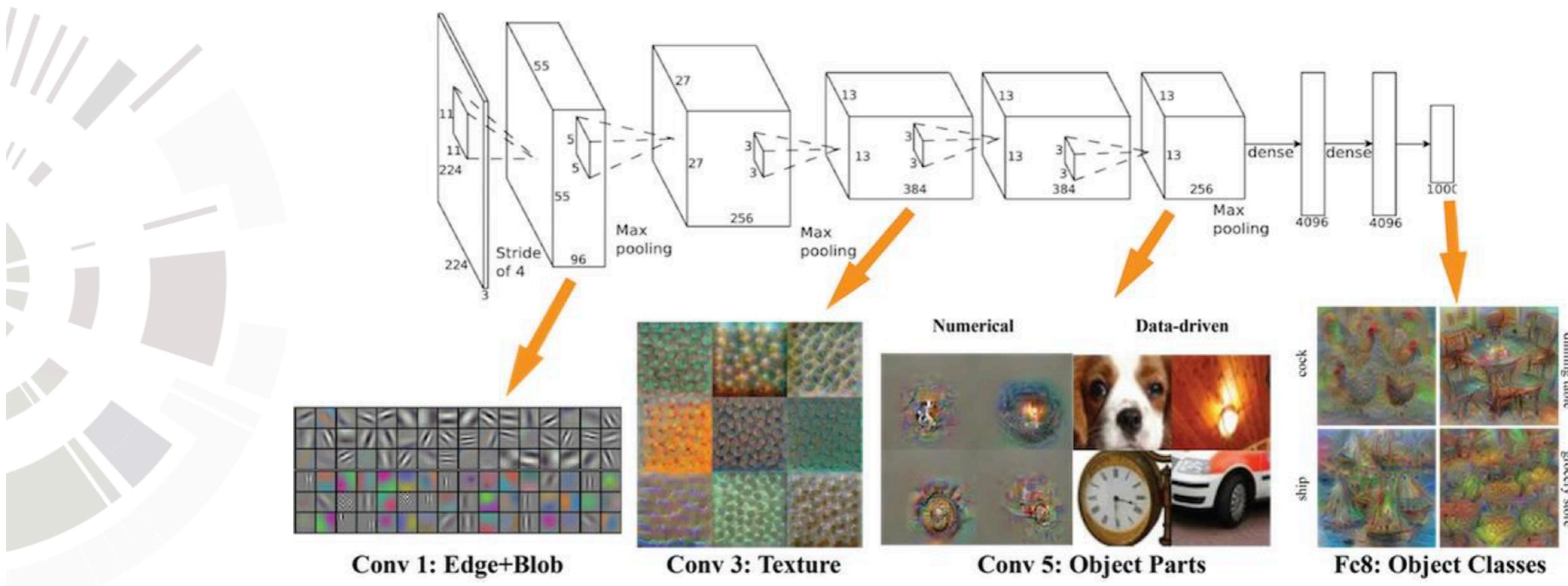
ImageNet: открытая выборка
15M размеченных изображений



Google: Распознавание кадров
с котами на видео из Youtube



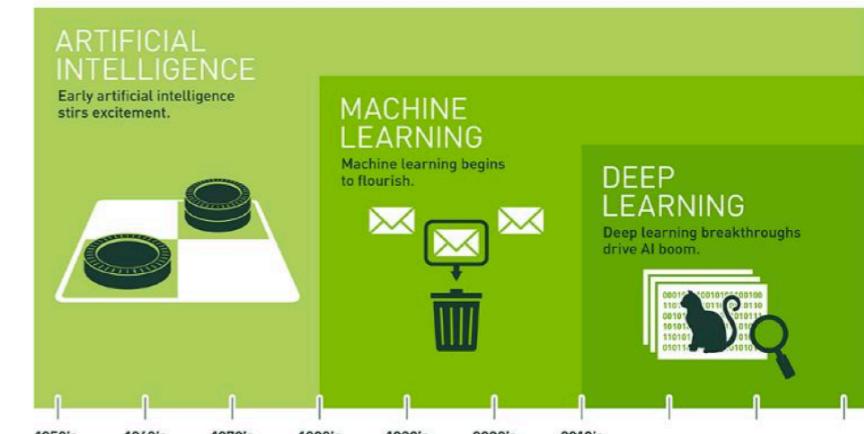
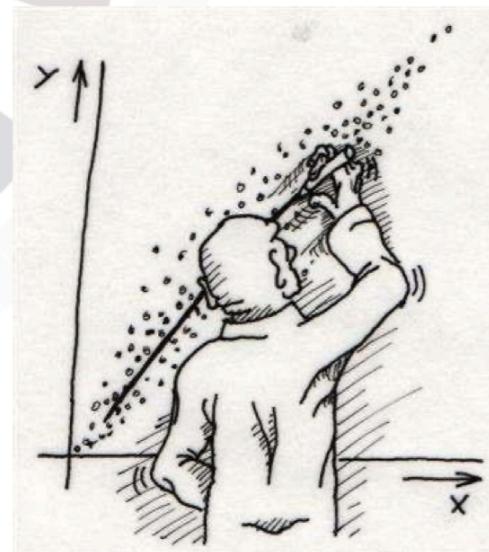
Что такое «глубокие нейронные сети»



A.Krizhevsky, I.Sutskever, G.Hinton. ImageNet classification with deep convolutional neural networks.
Communications of the ACM. 60 (6): 84–90.

Машинное обучение (Machine Learning, ML)

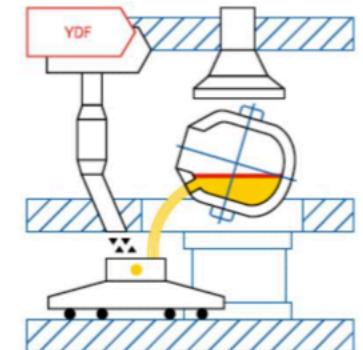
- одна из ключевых информационных технологий будущего
- наиболее успешное направление ИИ, вытеснившее экспертные системы и инженерию знаний



- **проведение функции через заданные точки в сложно устроенных пространствах**
- математическое моделирование в условиях, когда знаний мало, данных много
- тысячи различных методов и алгоритмов
- около 100 000 научных публикаций в год

Примеры задач машинного обучения

- **Медицинская диагностика:**
объект – данные о пациенте на текущий момент
ответ – диагноз / лечение / риск исхода
- **Поиск месторождений полезных ископаемых:**
объект – данные о геологии района
ответ – есть/нет месторождение
- **Управление технологическими процессами:**
объект – данные о сырье и управляющих параметрах
ответ – количество/качество полезного продукта



Примеры задач машинного обучения

- **Кредитный scoring:**

объект – данные о заемщике

ответ – решение по кредиту & вероятность дефолта



- **Предсказание оттока клиентов:**

объект – данные о клиенте на момент времени t

ответ – уйдет ли клиент к моменту времени $t + \Delta$



- **Прогнозирование объемов продаж:**

объект – данные о продажах на момент времени t

ответ – объем спроса в интервале от t до $t + \Delta$



Примеры задач машинного обучения

- **Информационный поиск в Интернете:**
объект – данные о паре «запрос и документ»
ответ – оценка релевантности документа запросу
- **Продажа рекламы в Интернете:**
объект – данные о тройке «пользователь, страница, баннер»
ответ – оценка вероятности клика
- **Рекомендательные системы в Интернете / TV:**
объект – данные о паре «пользователь, товар / фильм»
ответ – оценка вероятности покупки / просмотра



Примеры задач с данными сложной структуры

- **Статистический машинный перевод:**
объект – предложение на естественном языке
ответ – его перевод на другой язык
- **Перевод речи в текст:**
объект – аудиозапись речи человека
ответ – текстовая запись речи
- **Компьютерное зрение:**
объект – динамика сцены в видеопоследовательности
ответ – решение (объехать, остановиться, игнорировать)

Прогресс в этих областях связан с «**большими данными**» (англ. «*Big Data*»)

...очень важное уточнение:
с аккуратными
большими данными

Задача машинного обучения с учителем

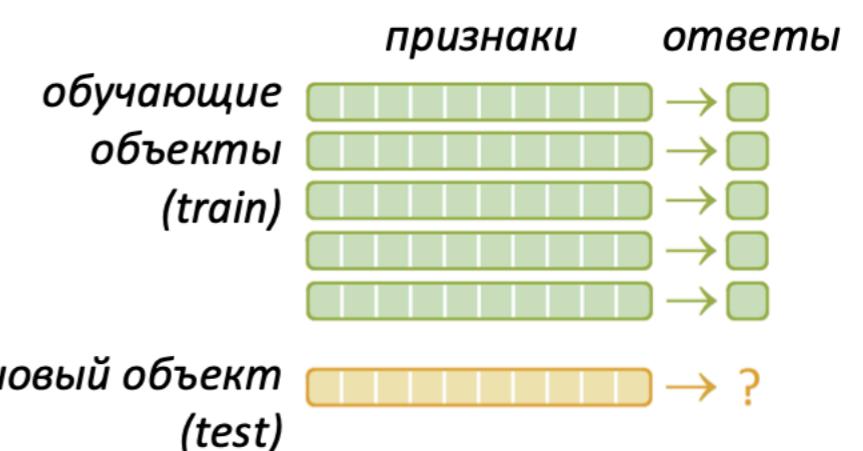
Этап №1 – обучение с учителем

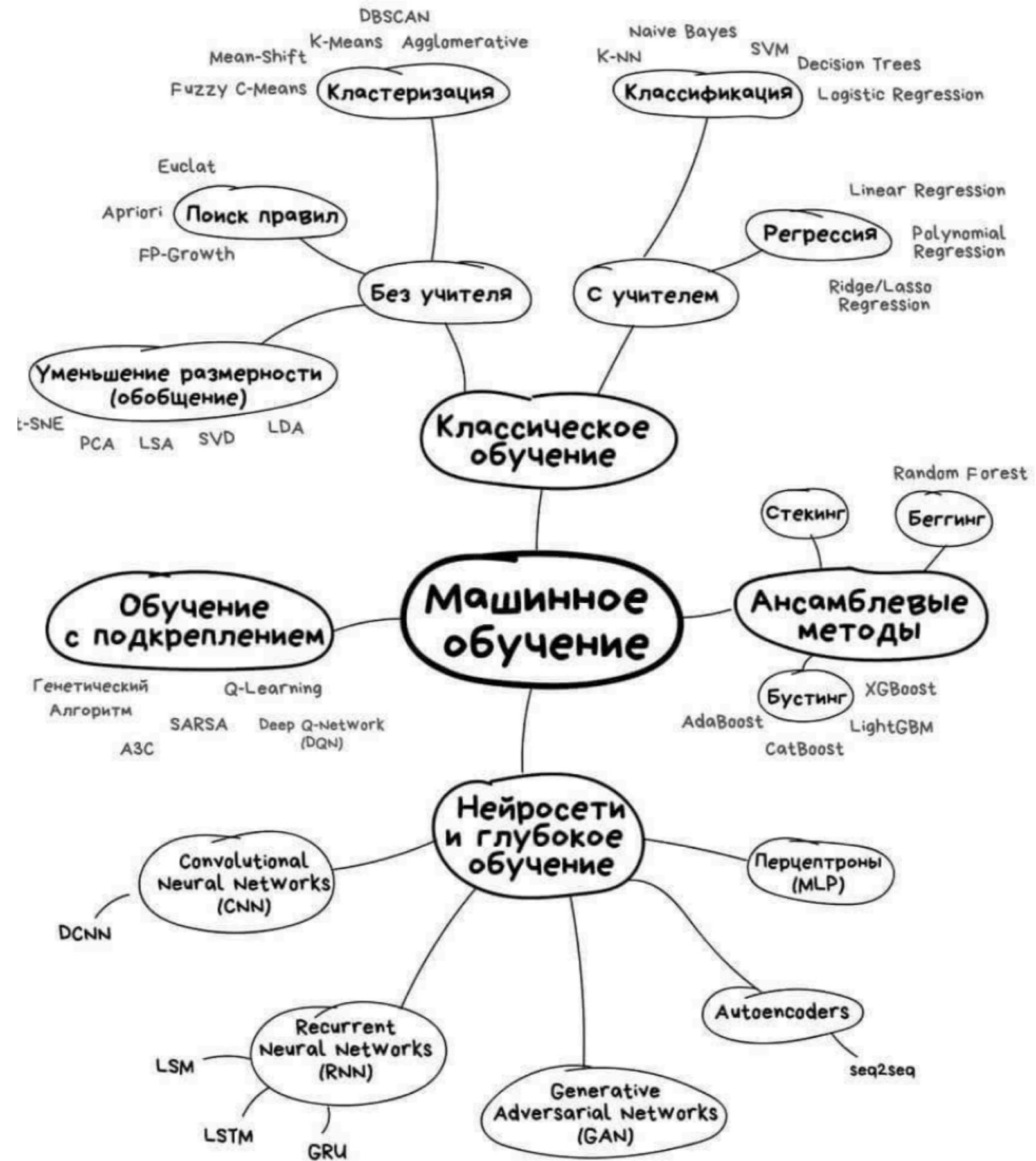
- **На входе:**
данные – выборка прецедентов «объект \rightarrow ответ»,
каждый объект описывается набором признаков
- **На выходе:**
модель, предсказывающая ответ по объекту

Если нет данных,
то нет
и машинного
обучения

Этап №2 – применение

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

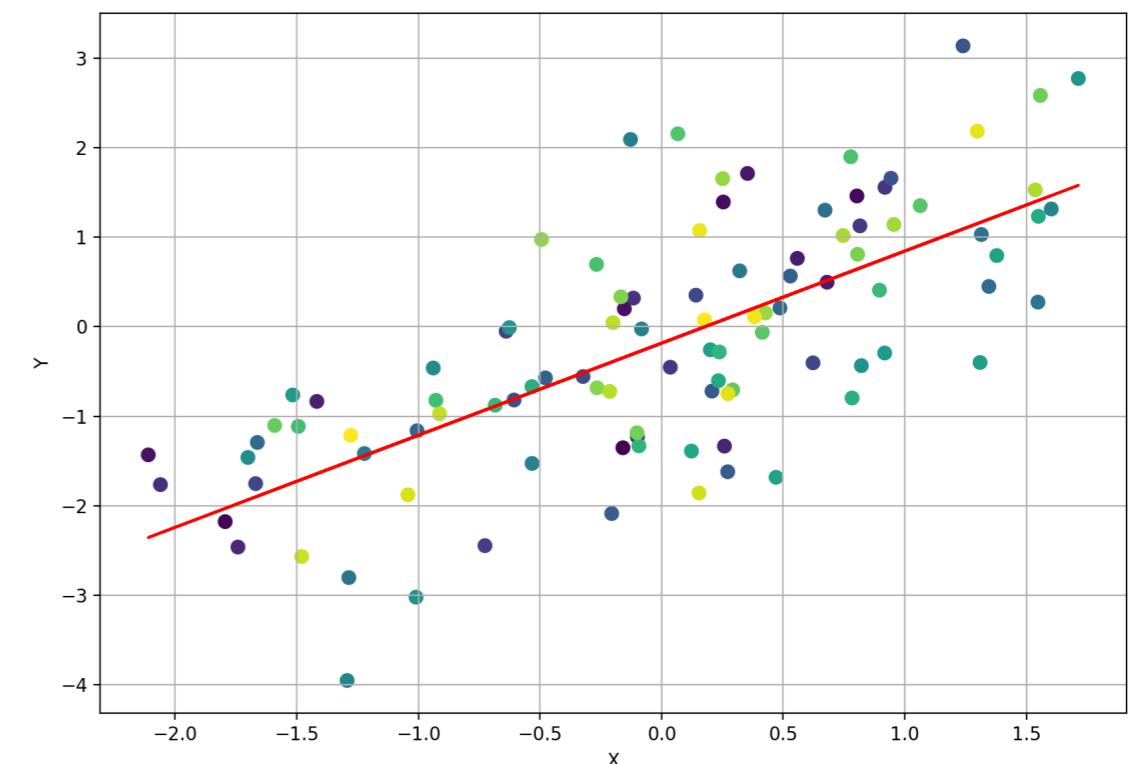




Линейная регрессия

$$y = f(x, b) + \varepsilon, E(\varepsilon)$$

Это контролируемый метод машинного обучения, который находит линейное уравнение, лучше всего описывающее корреляцию зависимых переменных с независимыми. Это достигается путем вписывания линии в данные с помощью метода наименьших квадратов. минимизировать сумму квадратов



Классификация

Один из разделов машинного обучения, посвященный решению следующей задачи. Имеется множество объектов (ситуаций), разделённых некоторым образом на классы. Задано конечное множество объектов, для которых известно, к каким классам они относятся. Это множество называется обучающей выборкой. Классовая принадлежность остальных объектов не известна. Требуется построить алгоритм, способный классифицировать произвольный объект из исходного множества.

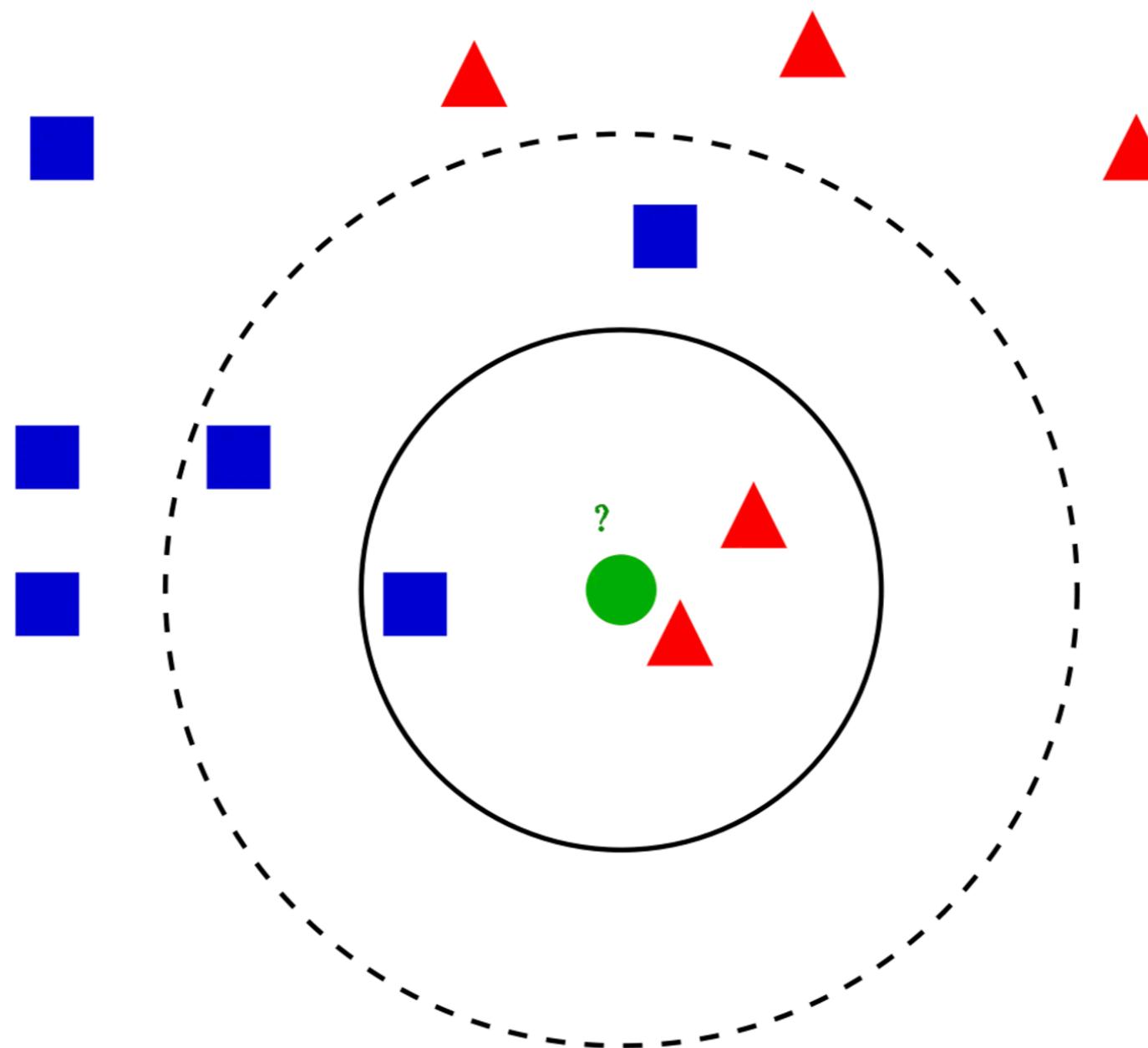
Классификация

Типы классификаторов

- Метод k-ближайших соседей (K-Nearest Neighbors);
- Метод опорных векторов (Support Vector Machines);
- Классификатор дерева решений (Decision Tree Classifier) / Случайный лес (Random Forests);
- Наивный байесовский метод (Naive Bayes);
- Линейный дискриминантный анализ (Linear Discriminant Analysis);
- Логистическая регрессия (Logistic Regression);

Метод k-ближайших соседей (K-Nearest Neighbors)

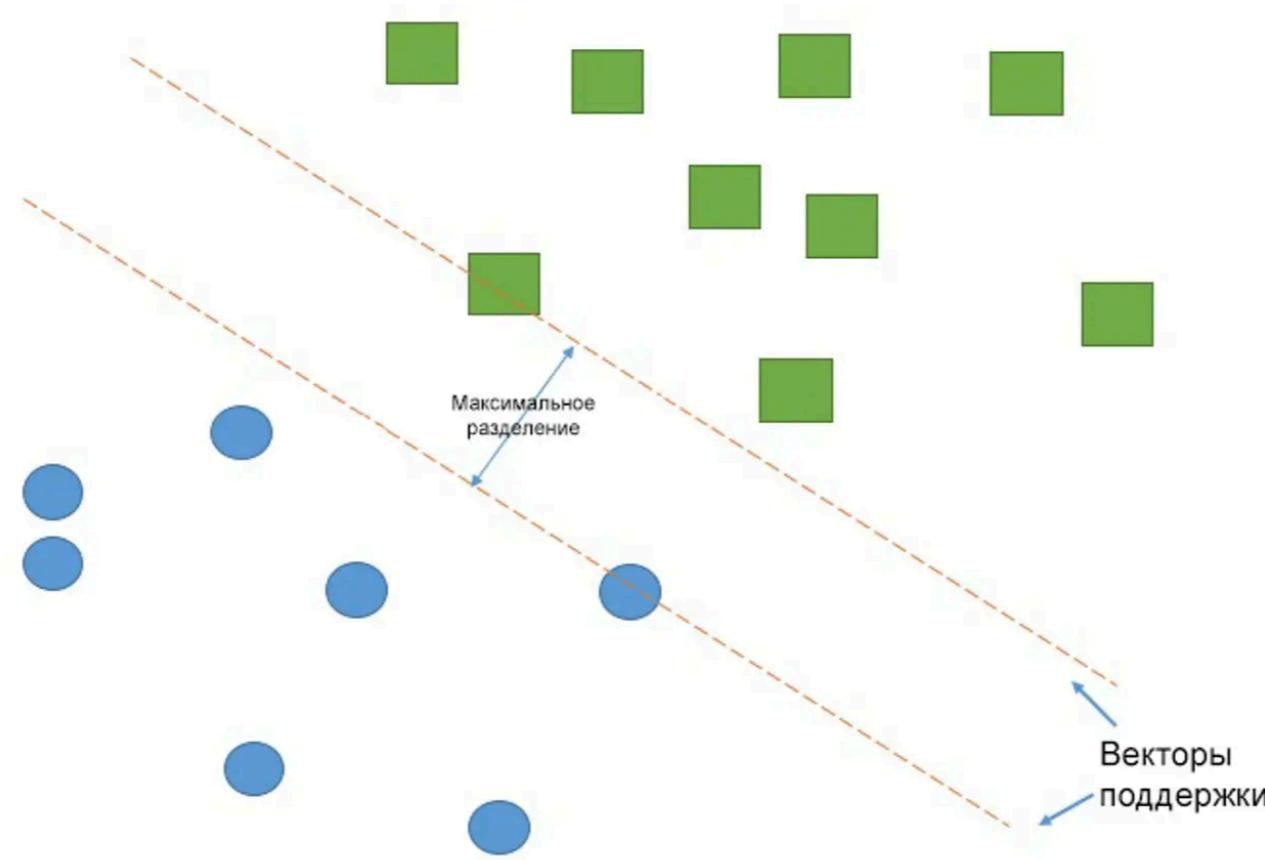
Этот метод работает с помощью поиска кратчайшей дистанции между тестируемым объектом и ближайшими к нему классифицированным объектами из обучающего набора. Классифицируемый объект будет относится к тому классу,



Метод опорных векторов (Support Vector Machines)

Работа метода опорных векторов заключается в рисовании линии между разными кластерами точек, которые нужно сгруппировать принадлежащ классу.

Классифициатс рисуемыми ли увеличить свс точки построее класс, которо



будут точки, — к другому

яние между ах, чтобы са. Когда все т — это

Рынок труда в области анализа данных

Инженер по данным (Data Engineer)

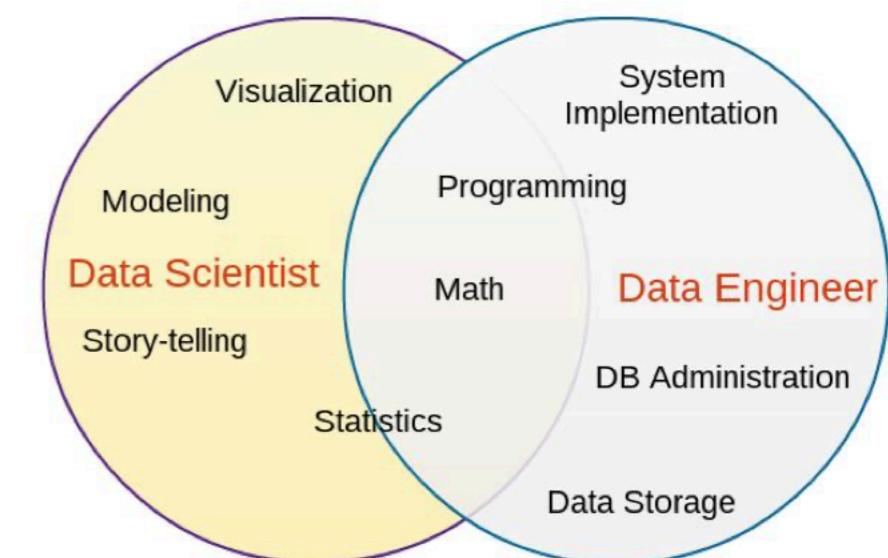
- Понимает бизнес-процессы, порождающие данные
- Работает с данными в различных форматах
- Визуализирует, понимает, очищает, готовит данные

Исследователь данных (Data Scientist)

- Моделирует, строит признаки (feature engineering)
- Выбирает модели и методы, оценивает решения
- Ходит по кругу CRISP-DM

Менеджер проектов по анализу данных

- Организует бизнес-процессы сбора и очистки данных
- Видит бизнес задачи и формализует их в терминах «Дано-Найти-Критерий»
- Организует открытые конкурсы и пилотные проекты
- Адекватно оценивает сложность задач и трудозатраты



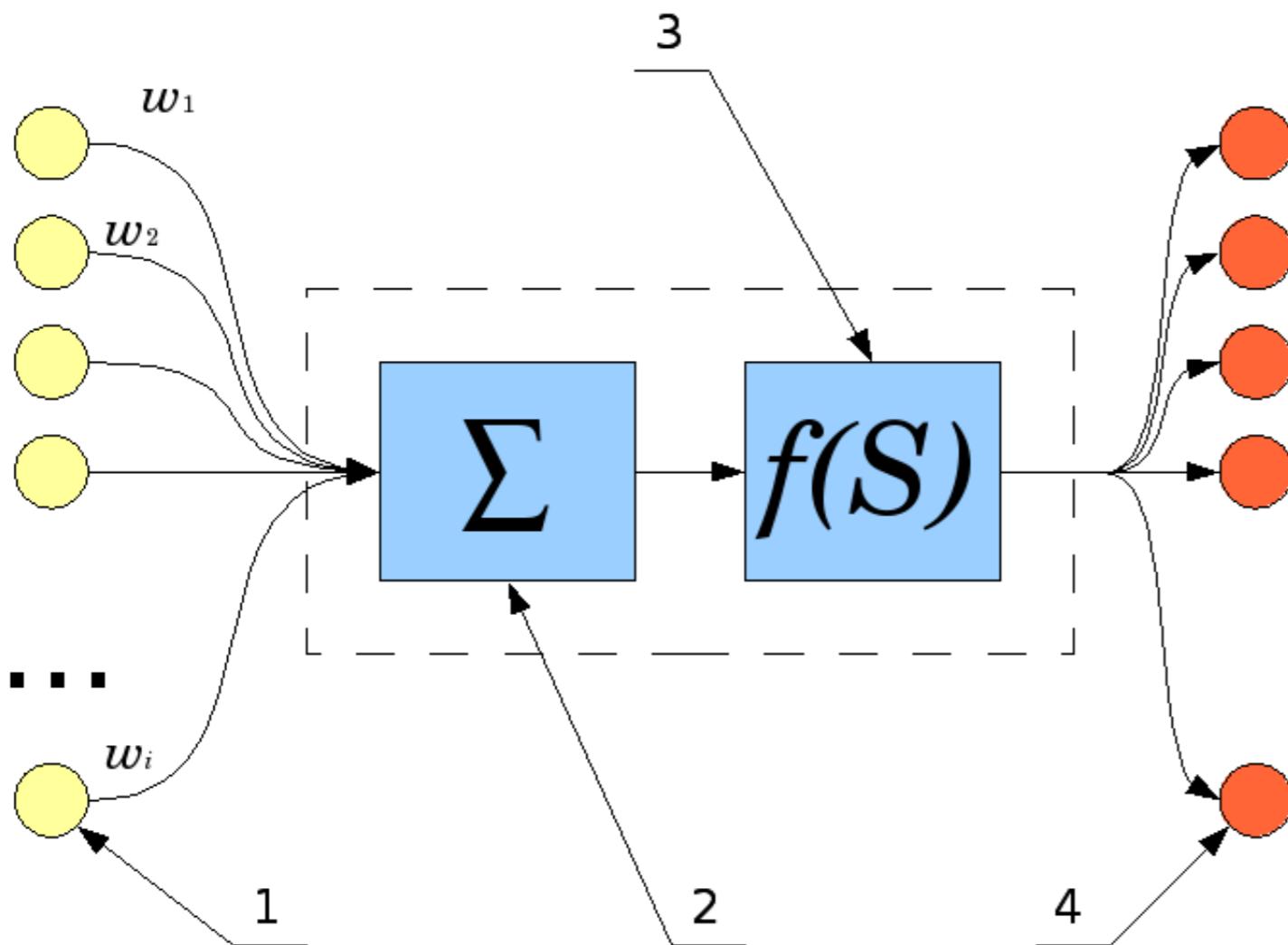
Нейронные сети

Определение

Вычислительные системы или машины, созданные для моделирования аналитических действий, совершаемых человеческим мозгом.

Нейронные сети используются для решения сложных задач, которые требуют аналитических вычислений подобных тем, что делает человеческий мозг.

Математический нейрон



Модель математического
нейрона Мак-Каллока-Питтса

- Уоррен Мак-Каллок и Уолтер Питтс в 1943 году предложили модель математического нейрона
- 1958 год. Френк Розенблatt на основе нейрона Мак-Каллока-Питтса создал компьютерную программу, а затем и физическое устройство – перцептрон

Классификация нейронных сетей

- Нейронные сети прямого распространения (Feed forward neural networks, FFNN)
- Сверточные нейронные сети (Convolutional neural network, CNN), слои:
 - А. входной;
 - Б. свертывающий;
 - С. объединяющий;
 - Д. подключенный;
 - Е. выходной.
- Рекуррентные нейронные сети (Recurrent neural network, RNN)

Задачи нейронных сетей

Классификация — распределение данных по параметрам. и т.д.

Предсказание — возможность предсказывать следующий шаг. Например, рост или падение акций, основываясь на ситуации на фондовом рынке.

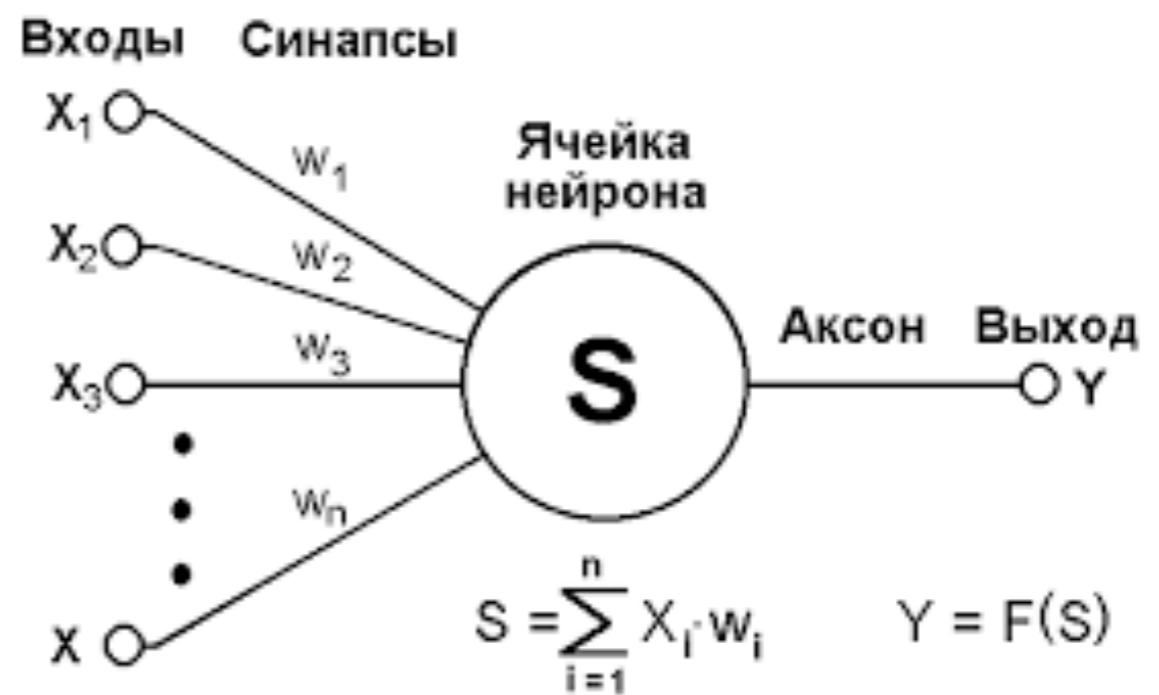
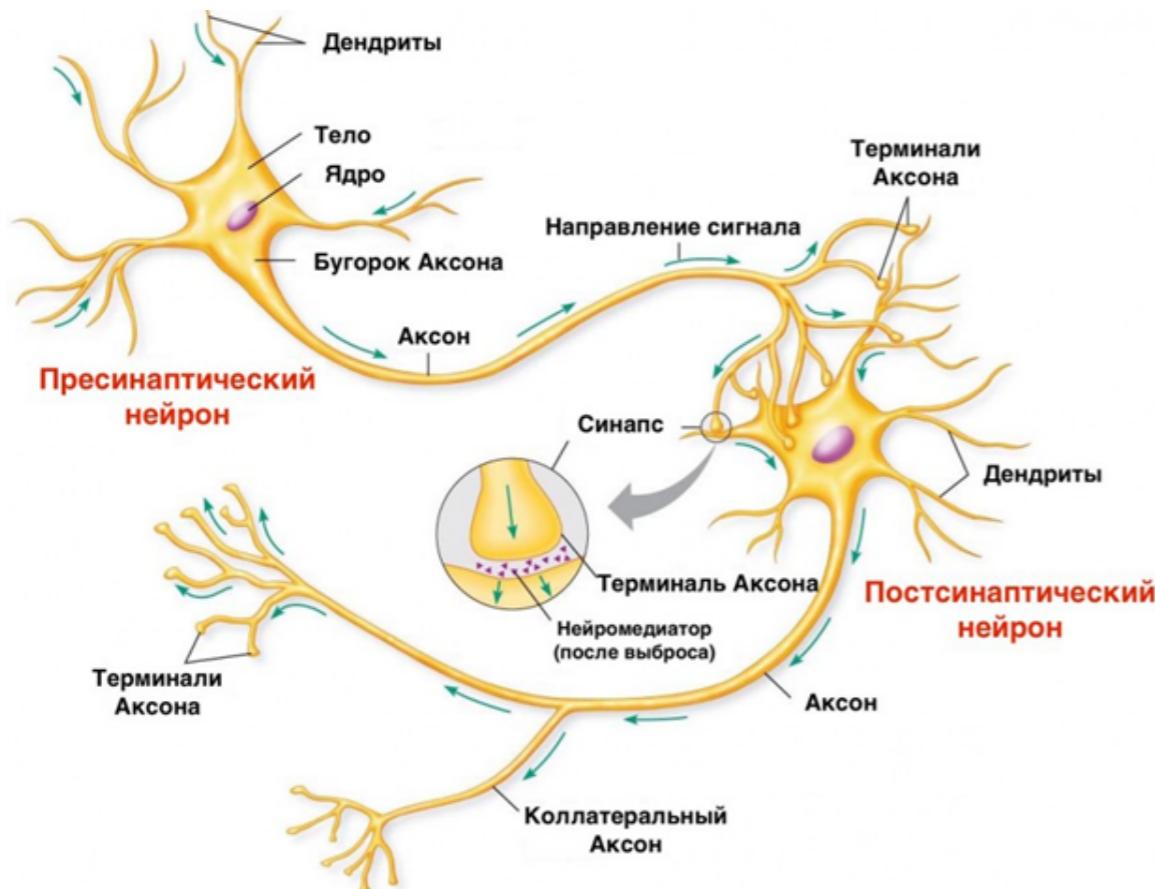
Распознавание — в настоящее время, самое широкое применение нейронных сетей. Используется в Google, когда вы ищете фото или в камерах телефонов, когда оно определяет положение вашего лица и выделяет его и многое другое.

Это как в биологии?

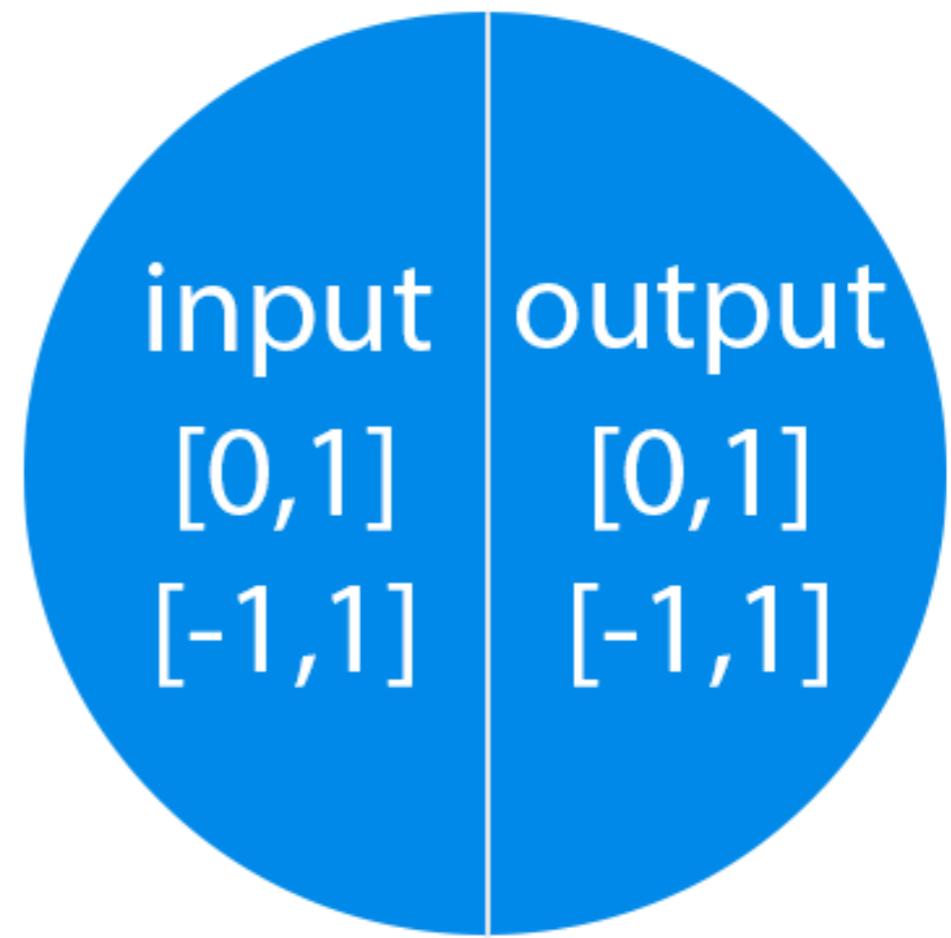
- Нейрон — это вычислительная единица, которая получает информацию, производит над ней простые вычисления и передает ее дальше.
- Синапс это связь между двумя нейронами.



Это как в биологии?

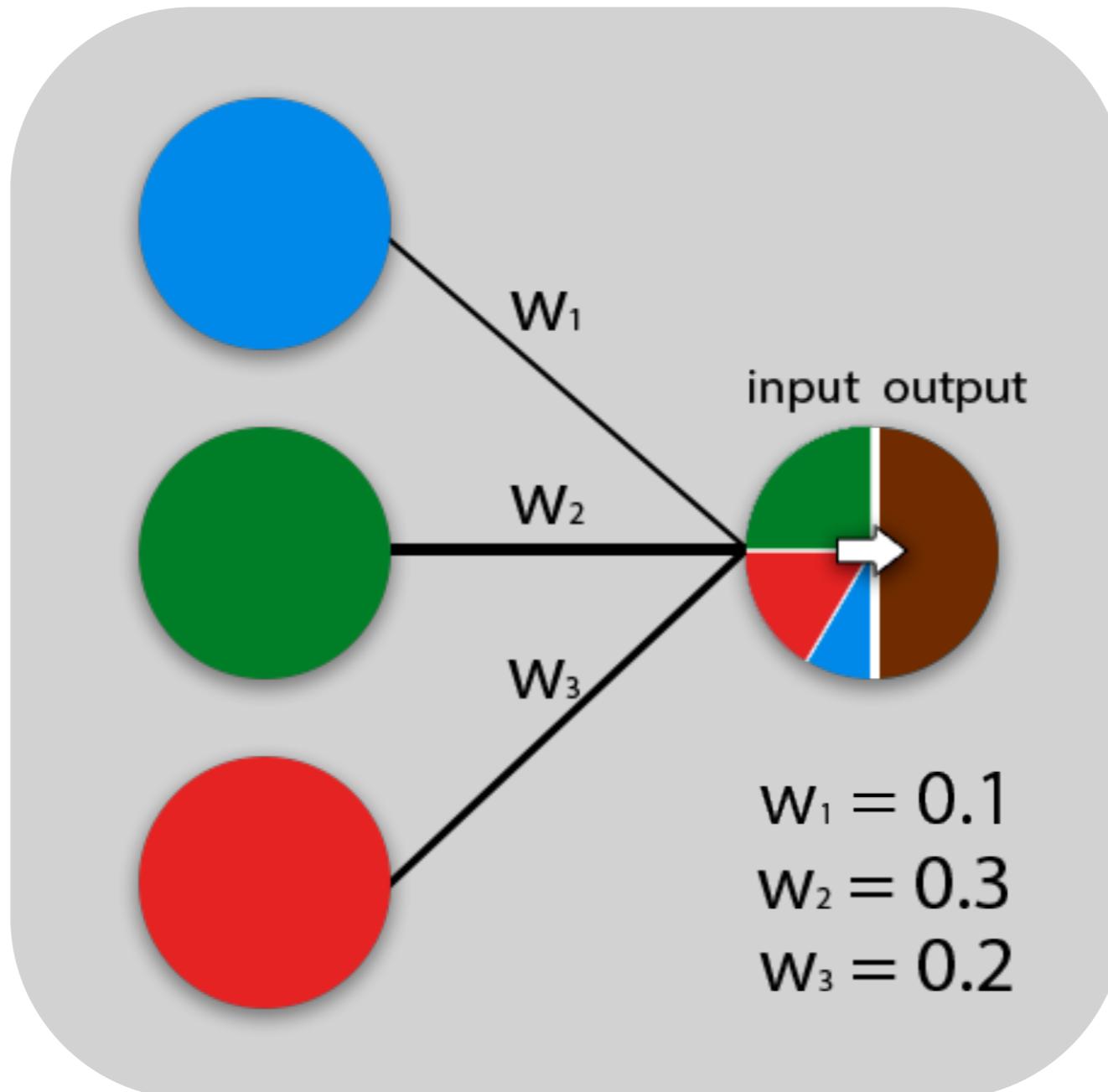


Нейроны



Важно помнить, что нейроны оперируют числами в диапазоне [0,1] или [-1,1]. А как же, вы спросите, тогда обрабатывать числа, которые выходят из данного диапазона? На данном этапе, самый простой ответ — это разделить 1 на это число.

Синапс



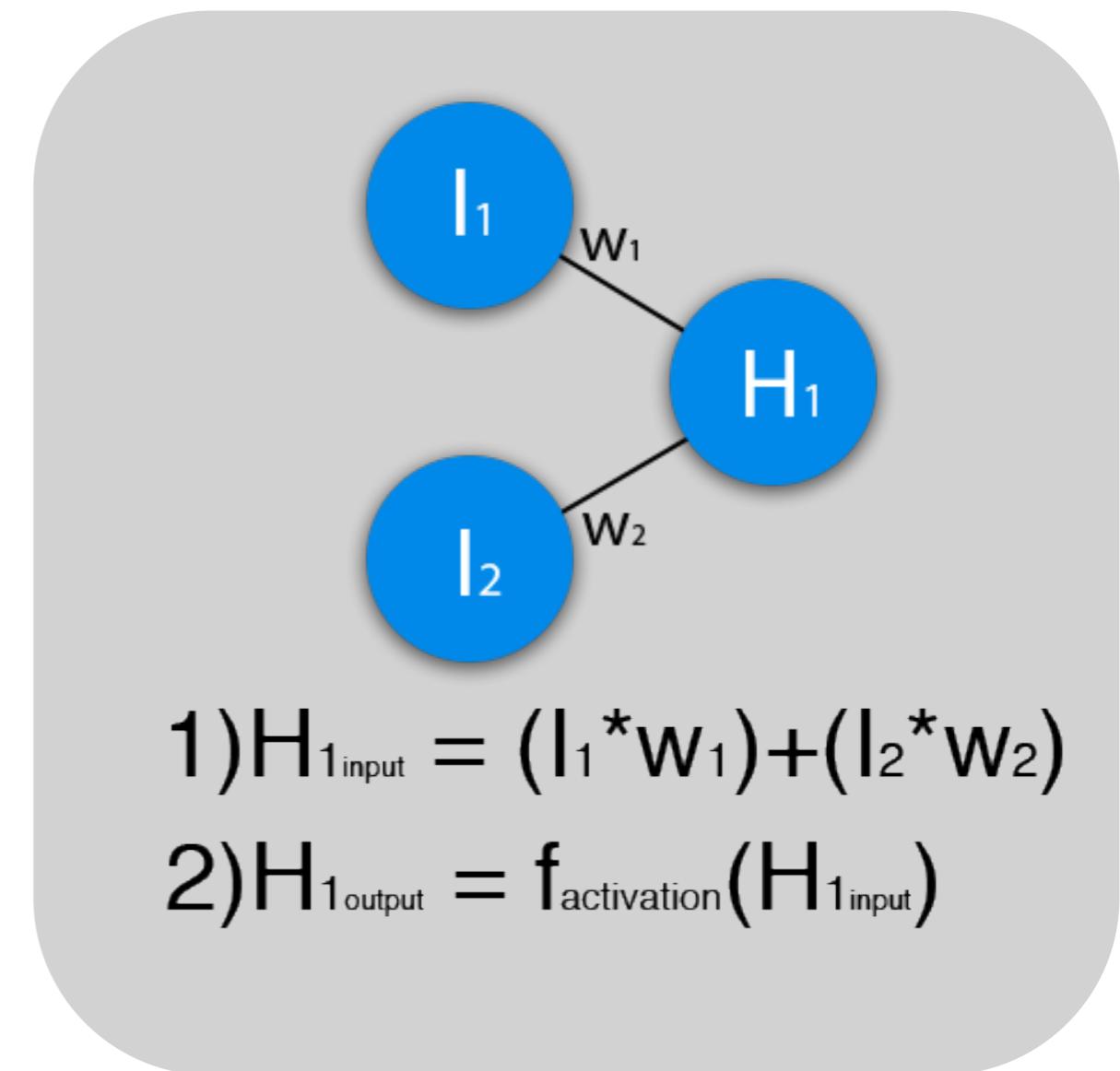
Как работает нейронная сеть

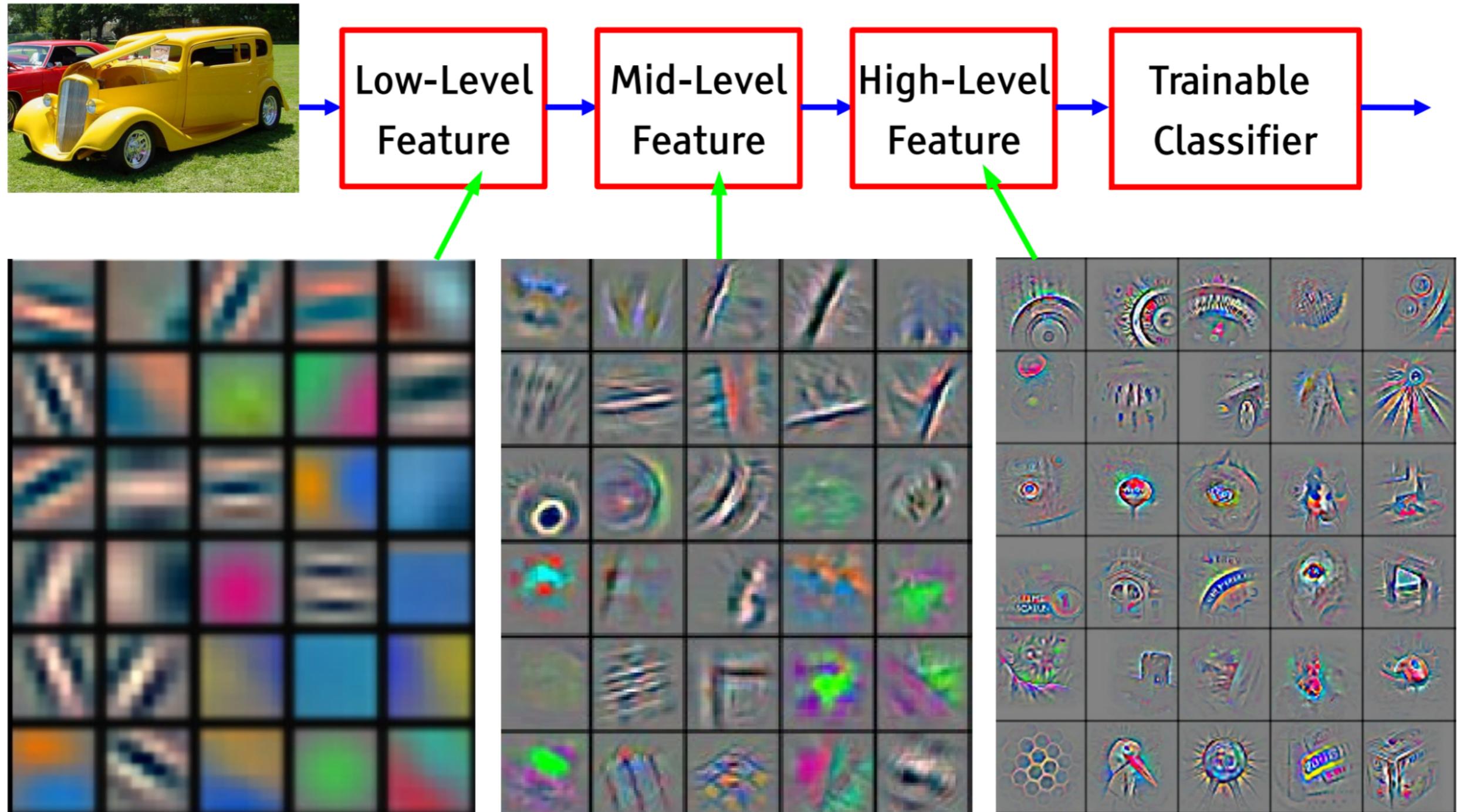
Вход: 1 и 0

Веса: $w_1=0.4$ и $w_2 = 0.7$

Входные данные нейрона H_1 :

$$1 \cdot 0.4 + 0 \cdot 0.7 = 0.4$$

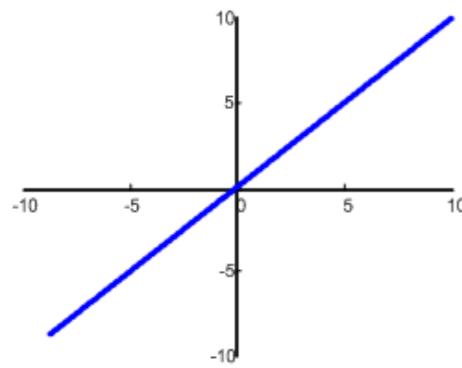




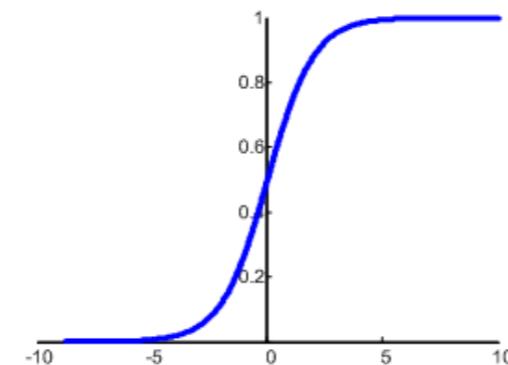
Функция активации

это способ нормализации входных данных

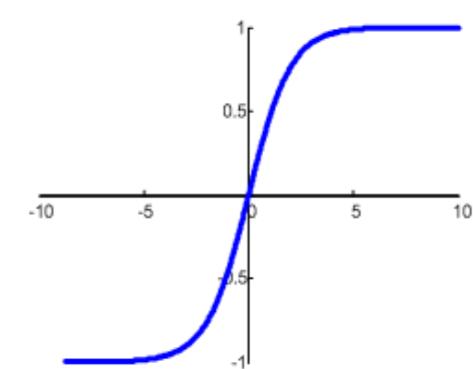
$$f(x) = x$$



$$f(x) = \frac{1}{1+e^{-x}}$$

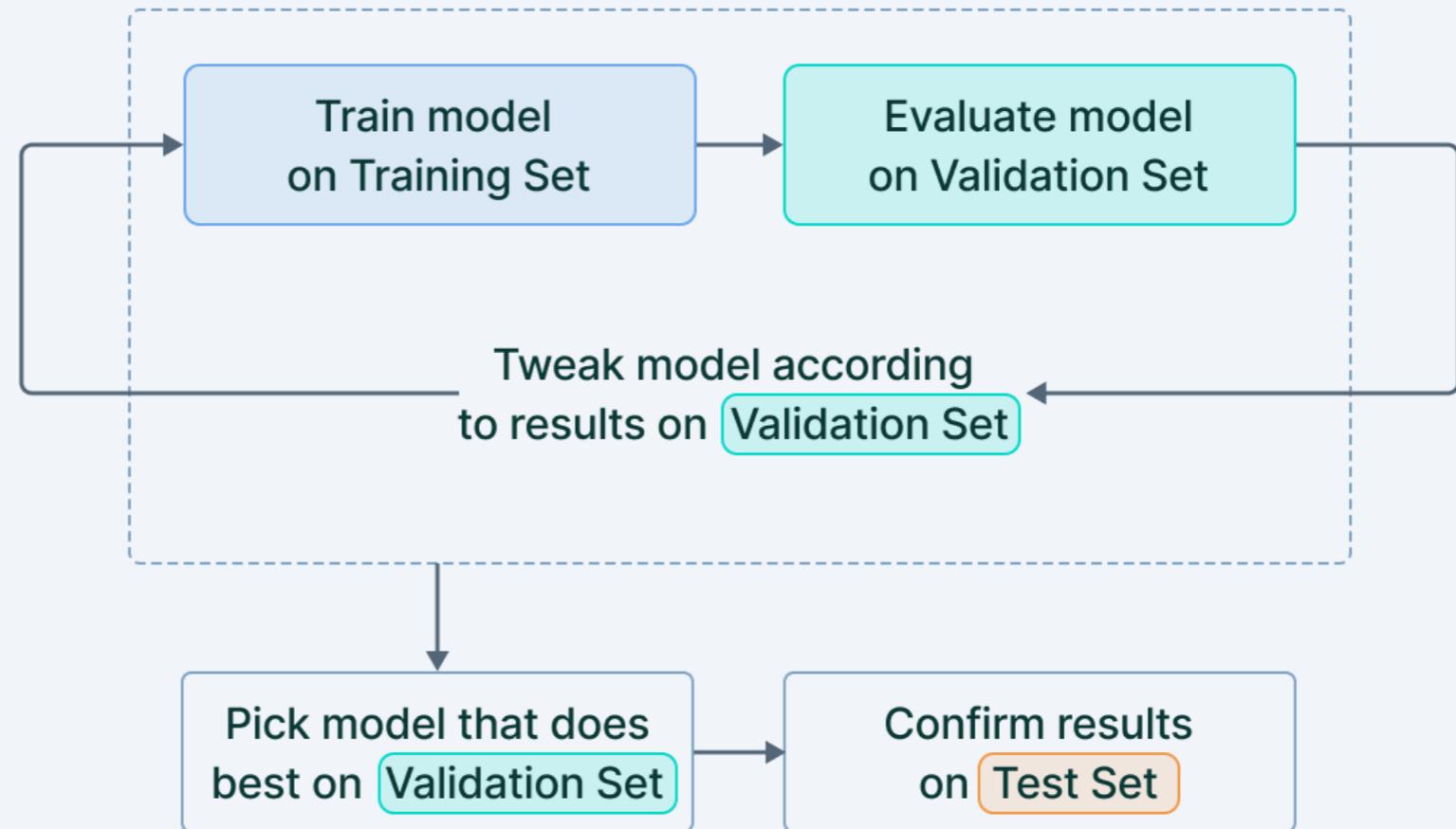


$$f(x) = \frac{e^{2x} - 1}{e^{2x} + 1}$$



Сеты

Training data/validation/test



V7 Labs

Итерации и эпохи

- Это своеобразный счетчик, который увеличивается каждый раз, когда нейронная сеть проходит один тренировочный сет.
- Эпоха увеличивается каждый раз, когда мы проходим весь набор тренировочных сетов.

✓ for (int i=0;i<maxEpoch;i++)
 for (int j=0;j<trainSet;j++)

✗ for (int j=0;j<trainSet;j++)
 for (int i=0;i<maxEpoch;i++)

Ошибки

Ошибка — это процентная величина, отражающая расхождение между ожидаемым и полученным ответами. Ошибка формируется каждую эпоху и должна идти на спад.

MSE

$$\frac{(i_1-a_1)^2 + (i_2-a_2)^2 + \dots + (i_n-a_n)^2}{n}$$

Root MSE

$$\sqrt{\frac{(i_1-a_1)^2 + (i_2-a_2)^2 + \dots + (i_n-a_n)^2}{n}}$$

Arctan

$$\frac{\arctan^2(i_1-a_1) + \dots + \arctan^2(i_n-a_n)}{n}$$

Этапы создания НН

- Сбор данных
- Препроцессинг
- Построение модели
- Анализ качества и интерпретации модели

класс: (категория)	цвет: (категория)	вкус: (категория)	вес: (число)	твёрдый: (bool)
-	красный	кислый	4.23	да
-	зеленый	горький	3.15	нет
+	зеленый	горький	4.14	да
+	синий	сладкий	4.38	нет
-	зеленый	соленый	3.62	нет

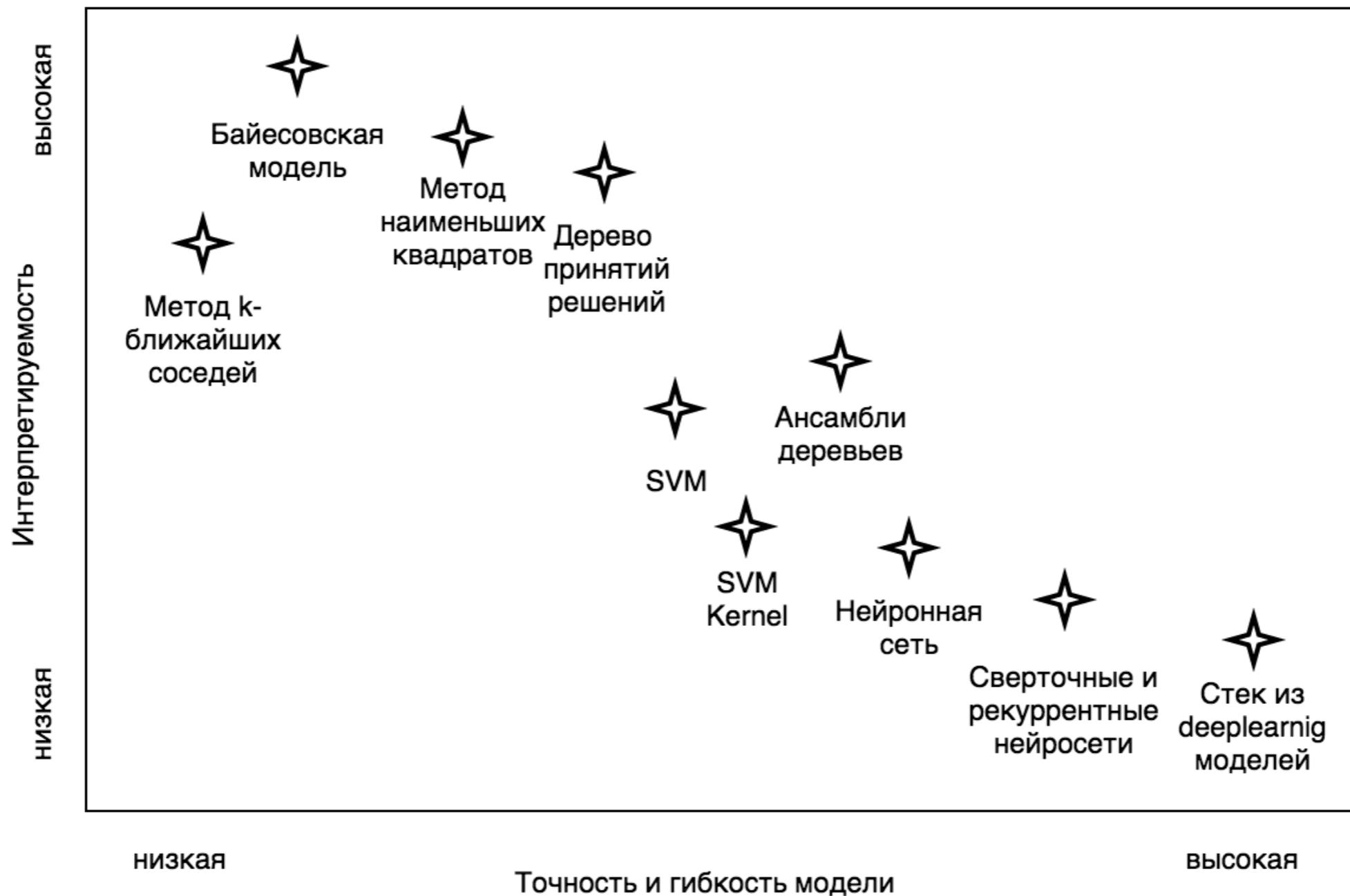
Препроцессинг

- Создание векторного пространства признаков;
- Нормализация данных;
- Изменение размерности векторного пространства.

```
def normalize(X):  
    return (X-X.mean())/  
          X.std()
```

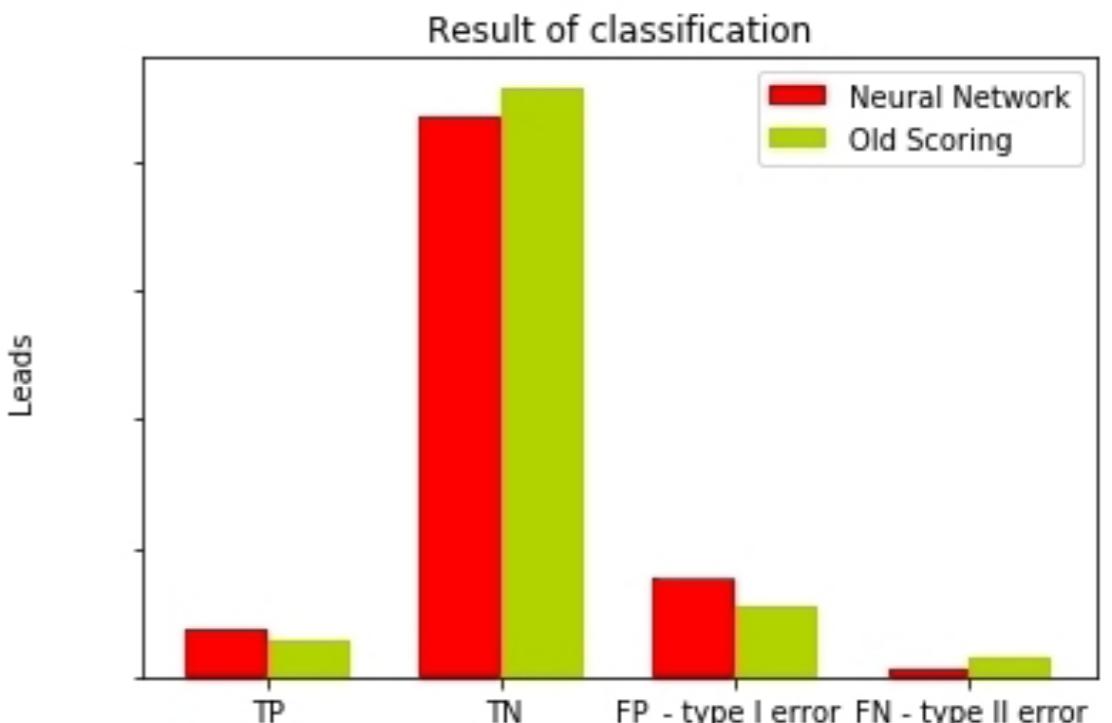
class:	red:	green:	blue:	bitter:	sweet:	salti:	sour:	weight:	solid:
0	1	0	0	0	0	0	1	0.23	1
0	0	1	0	1	0	0	0	-0.85	0
1	0	1	0	1	0	0	0	0.14	1
1	0	0	1	0	1	0	0	0.38	0
0	0	1	0	0	0	1	0	-0.48	0

Выбор модели



Оценка качества модели

- TP (True Positive) — истиноположительный. Классификатор решил, что клиент купит, и он купил.
- FP (False Positive) — ложноположительный. Классификатор решил, что клиент купит, но он не купил. FN (False Negative) — ложноотрицательный. Классификатор решил, что клиент не купит, а он мог купить (или уже купил).
- TN (True Negative) — истиноотрицательный. Классификатор решил, что клиент не купит, и он не купил.



Полезные ссылки

https://www.youtube.com/playlist?list=PLHIAngK_uV8FJlh04cUMZMPWNHryN7Ma

<https://mattmazur.com/2015/03/17/a-step-by-step-backpropagation-example/>

<http://stevenmiller888.github.io/mind-how-to-build-a-neural-network/>

https://www.youtube.com/watch?v=uSUOdu_5MPc&index=2&list=FLeakN3yOlVASrAnu779jzQ&ab_channel=TED

https://www.youtube.com/watch?v=0qVOUD76JOg&list=FLeakN3yOlVASrAnu779jzQ&index=4&ab_channel=TEDxTalks