

Введение в машинное обучение

Н.Ю. Золотых

ННГУ, ИТММ

Летняя школа ННГУ по компьютерному зрению

4–8 июля 2022

Искусственный интеллект (AI – Artificial Intelligence)

Искусственный интеллект – компьютер/программа, решающая интеллектуальные задачи, т.е. выполнение которых традиционно считалось прерогативой человека.

Сильный (общий) искусственный интеллект (strong, general AI) – компьютер/программа, способная решать любые интеллектуальные задачи не хуже человека.

Слабый (частный, прикладной, специализированный) искусственный интеллект (weak AI) – компьютер/программа, способная решать конкретный класс интеллектуальных задач.

Cats vs Dogs



“Dog”



“Cat”

Два основных подхода к построению ИИ

- Экспертные системы (наборы правил, описывающих “все” ситуации)
- Машинное обучение – компьютер сам генерирует правила на основе данных “объекты-ответы”

Экспертные системы

Экспертные системы – наборы правил, описывающие “все” (почти все) ситуации

Если усы длинный, когти втяжные, хвост пушистый,
то КОШКА

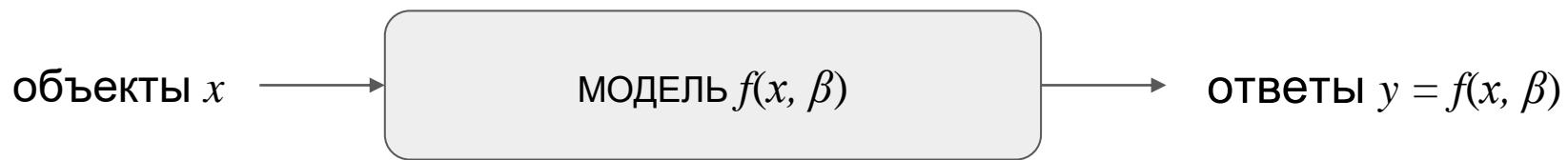
Иначе ...

Если пациент обладает симптомами $\tau_1, \tau_2, \dots, \tau_n$,
и имеют место фоновые условия c_1, c_2, \dots, c_m ,
то у пациента с уверенностью α заболевание у
Иначе ...

Машинное обучение

Машинное обучение – компьютер сам генерирует правила на основе данных “объекты-ответы”

Алгоритм обучения подстраивает некоторую функцию (модель) $f(x, \beta)$, так, чтобы по заданным объектам x она (по возможности) предсказывала правильные ответы $y = f(x, \beta)$



x называются *объектами, примерами (samples)*

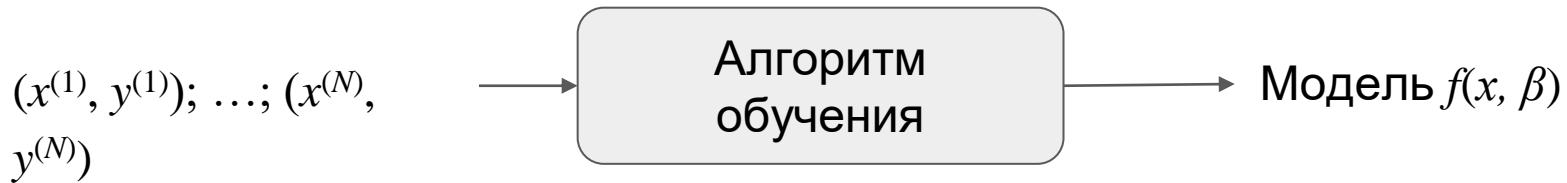
y называются *ответами, метками (answers, labels)*

Обучение с учителем

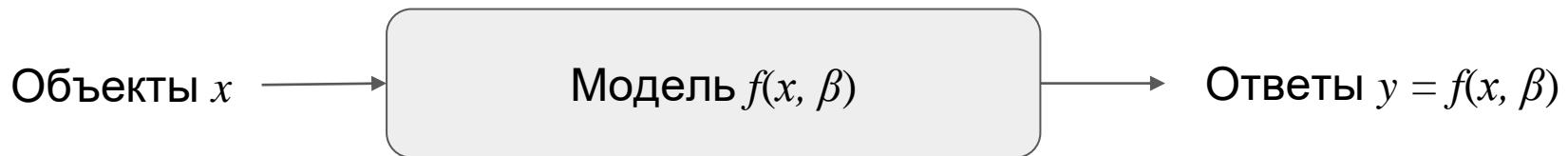
Имеется обучающая выборка: $(x^{(1)}, y^{(1)}); (x^{(2)}, y^{(2)}); \dots; (x^{(N)}, y^{(N)})$

Надо найти $f(x)$, так, чтобы $f(x^{(1)}) \approx y^{(1)}$; $f(x^{(2)}) \approx y^{(2)}$; ...; $f(x^{(N)}) \approx y^{(N)}$

Fit (настройка, обучение, learning, training)



Predict (предсказание, inference)



Cats vs Dogs



“Dog”

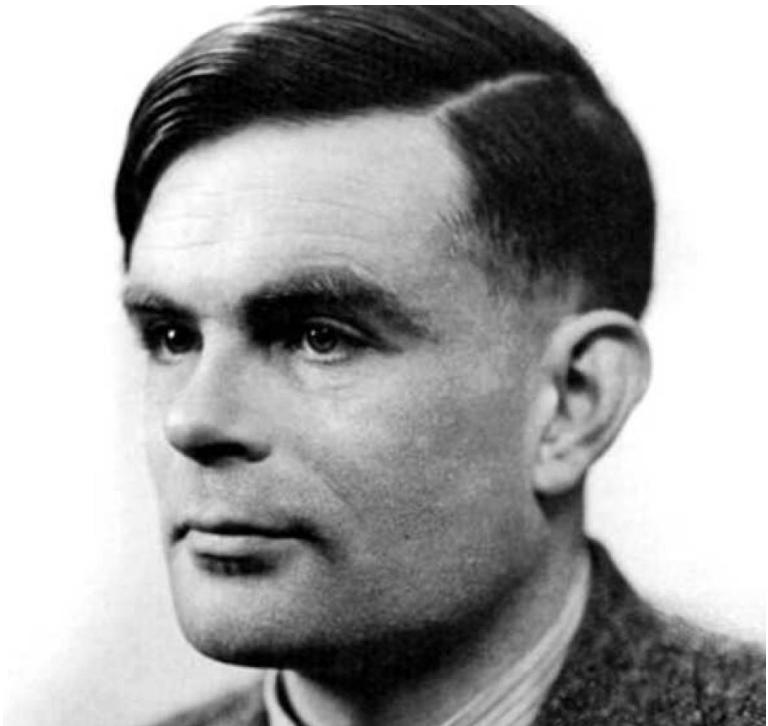


“Cat”

Что такое машинное обучение (ML – Machine Learning)?

Идея обучающихся машин (*learning machines*) принадлежит А. Тьюрингу

[*Turing A. Computing Machinery and Intelligence* // *Mind.* 1950. V. 59. P. 433–460;
перепечатано: *Can the Machine Think?* // *World of Mathematics*. Simon and Schuster,
N.Y. 1956. V. 4. P. 2099–2123; рус. перев.:
Тьюринг А. М. Может ли машина мыслить?
// М.: Физматлит, 1960]



Alan Mathison Turing
(1912–1954)

Машинное обучение – процесс, в результате которого машина способна показывать поведение, которое в нее не было явно заложено (запрограммировано).

A.L. Samuel Some Studies in Machine Learning Using the Game of Checkers // IBM Journal. July 1959. P. 210-229.

Говорят, что компьютерная программа *обучается* на основе опыта E по отношению к некоторому классу задач T и мере качества P , если качество решения задач из T , измеренное на основе P , улучшается с приобретением опыта E .

T.M. Mitchell Machine Learning. McGraw-Hill, 1997

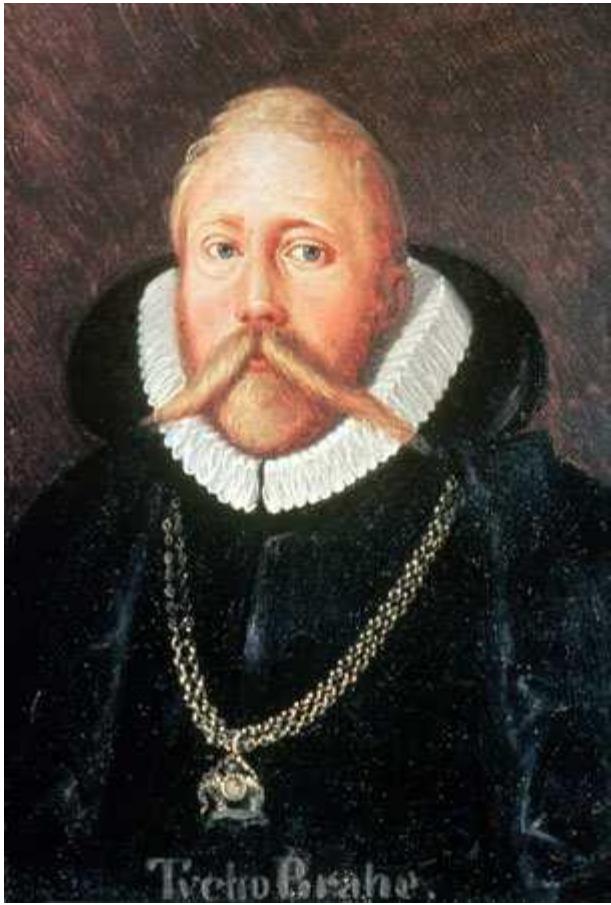
Что такое анализ данных?

Data Analysis – анализ данных

Data Mining – добыча данных, интеллектуальный анализ данных, глубинный анализ данных

Data Mining – совокупность методов обнаружения в данных ранее неизвестных, нетривиальных, практически полезных и доступных интерпретации знаний, необходимых для принятия решений в различных сферах человеческой деятельности.

[Г. Пятецкий-Шапиро, 1989]



Тихо Браге (1546–1601)



Иоганн Кеплер (1571–1630)

Machine Learning vs Data Mining

ML и DM извлекают закономерности («знания») из данных, но (немного) с разными целями:

- ML – чтобы обучить машину
- DM – чтобы обучить человека

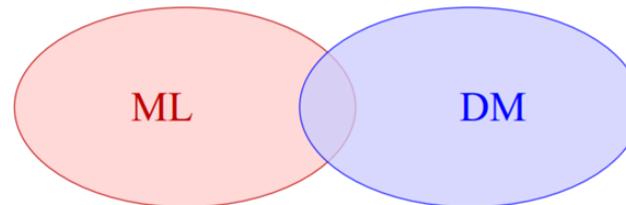
Поэтому *в первую очередь**

- в ML минимизируют ошибку
- в DM важна интерпретируемость результата

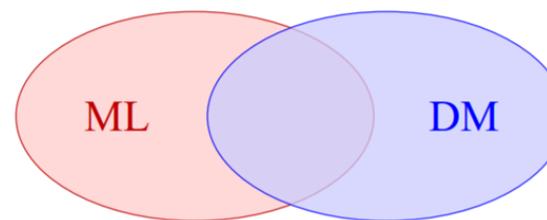
(Объясняемый ИИ (XAI) должен стереть эту границу)

Machine Learning vs Data Mining

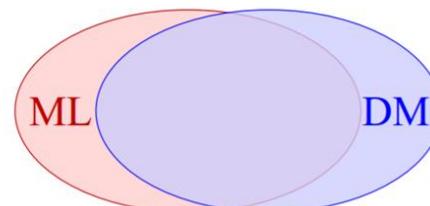
Содержательные
постановки задач



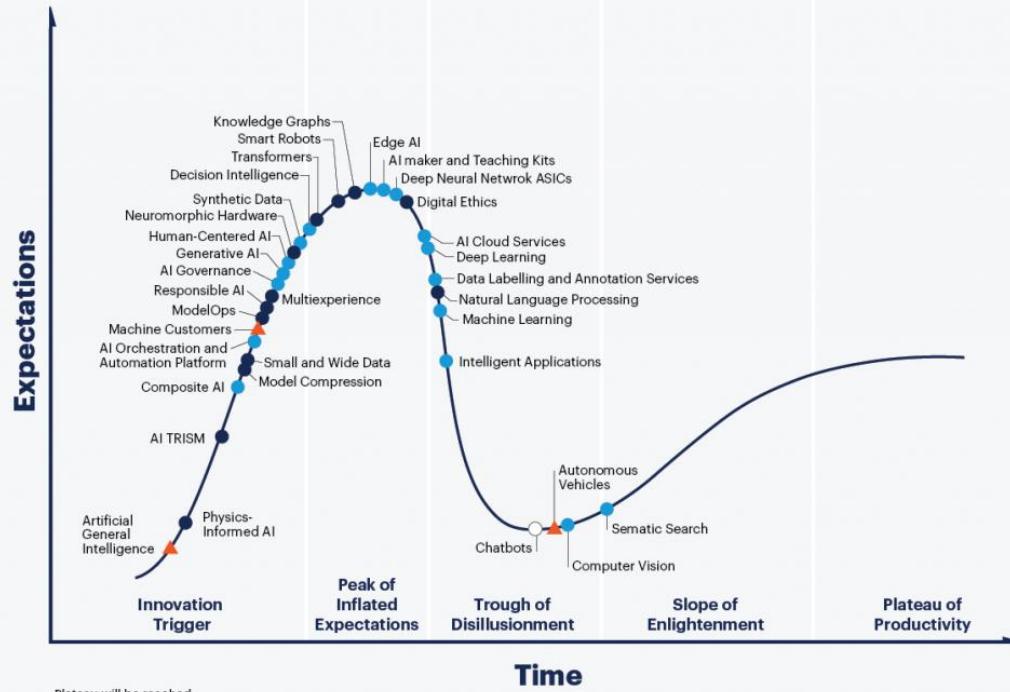
Математические
постановки задач



Методы



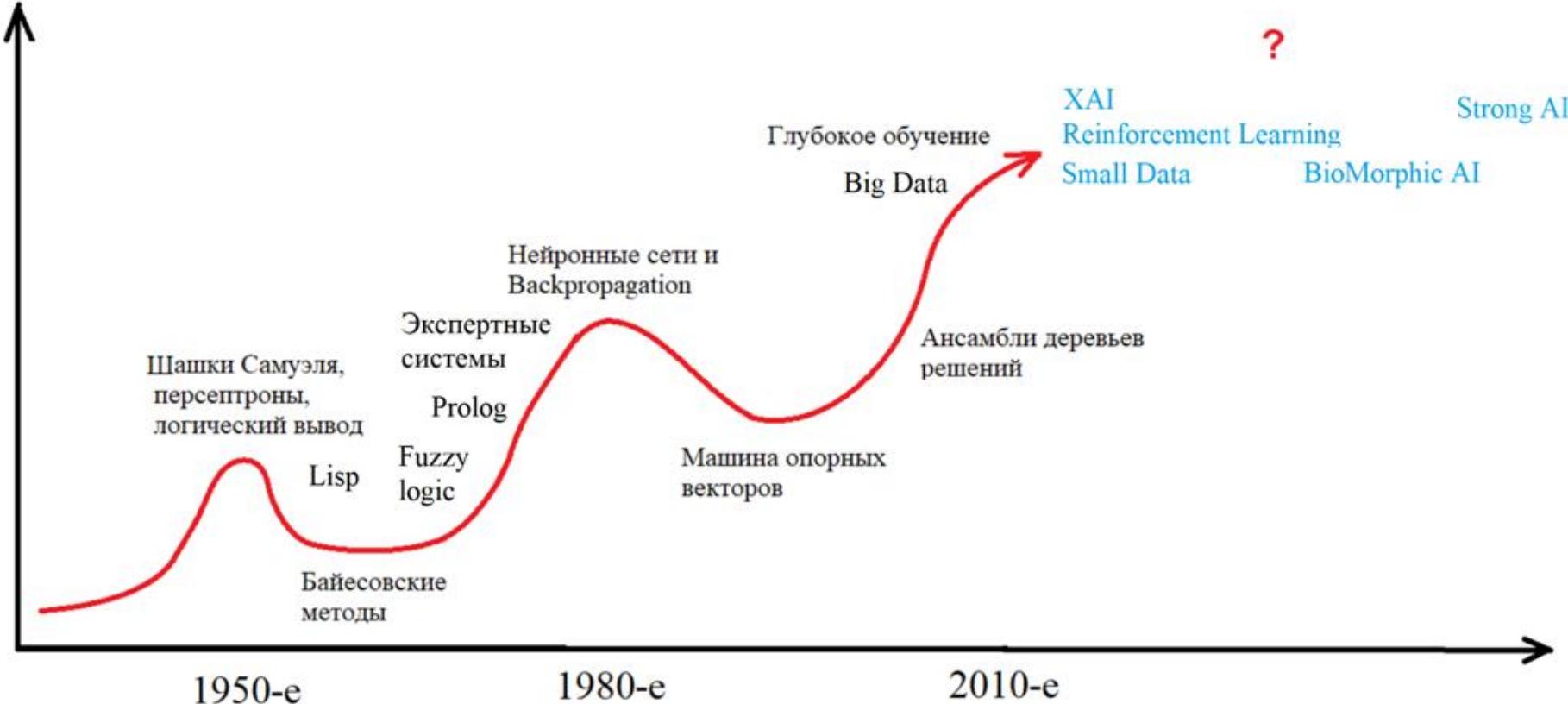
Hype Cycle for Artificial Intelligence, 2021



gartner.com

Source: Gartner
© 2021 Gartner, Inc. and/or its affiliates. All rights reserved. Gartner and Hype Cycle are registered trademarks of Gartner, Inc. and its affiliates in the U.S. 1482644

Gartner



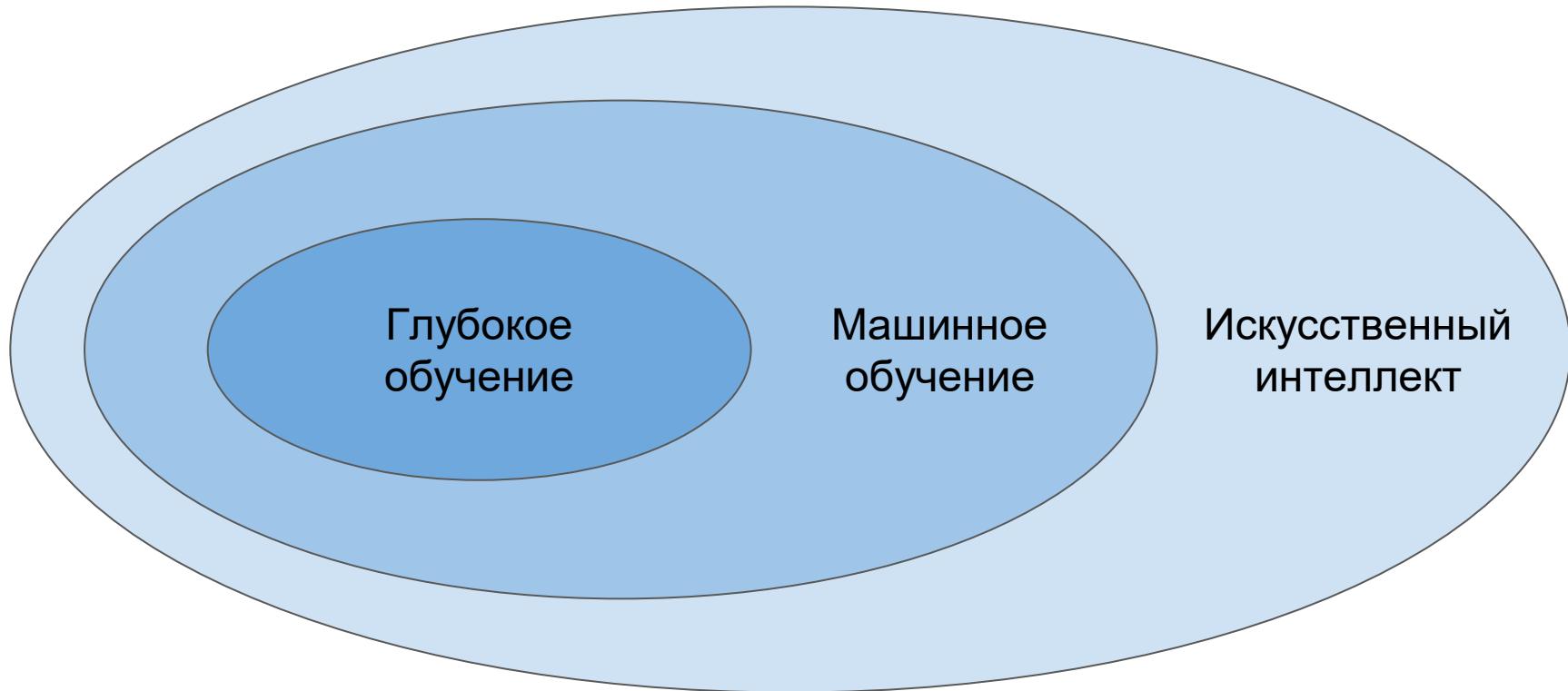
Машинное обучение сегодня

Причины “Третьей весны AI”:

- Новые алгоритмы (deep learning – **глубокое обучение** и др.),
хотя все новое – это хорошо забытое старое!
- Мощные компьютеры (в первую очередь – GPU)
- Много данных (Big Data)

Достижения:

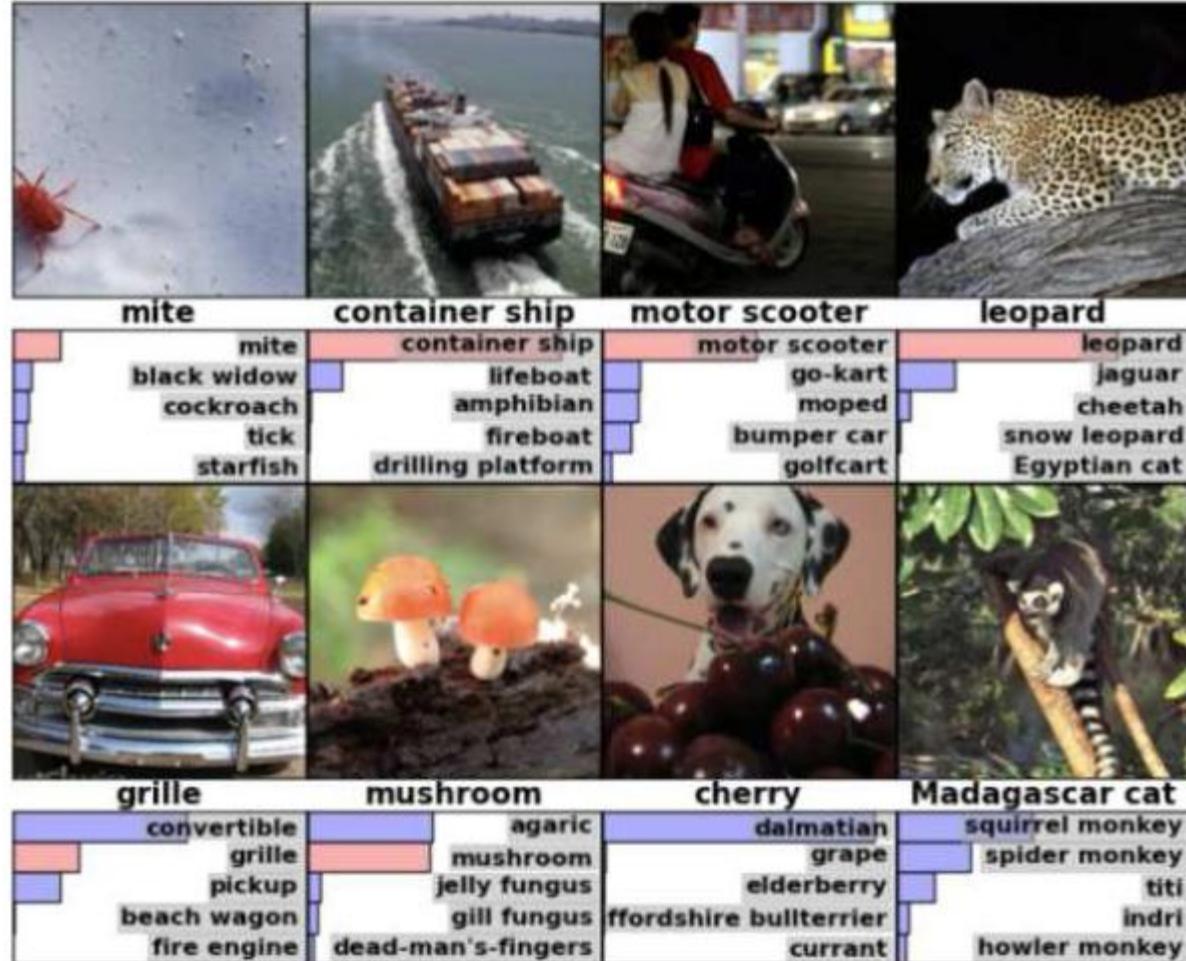
- ...
- Машинное обучение становится *технологией*



ImageNet

Прорыв 2012:
ImageNet ILSVRC-2012
(около 1 млн.
изображений, 1000
классов)

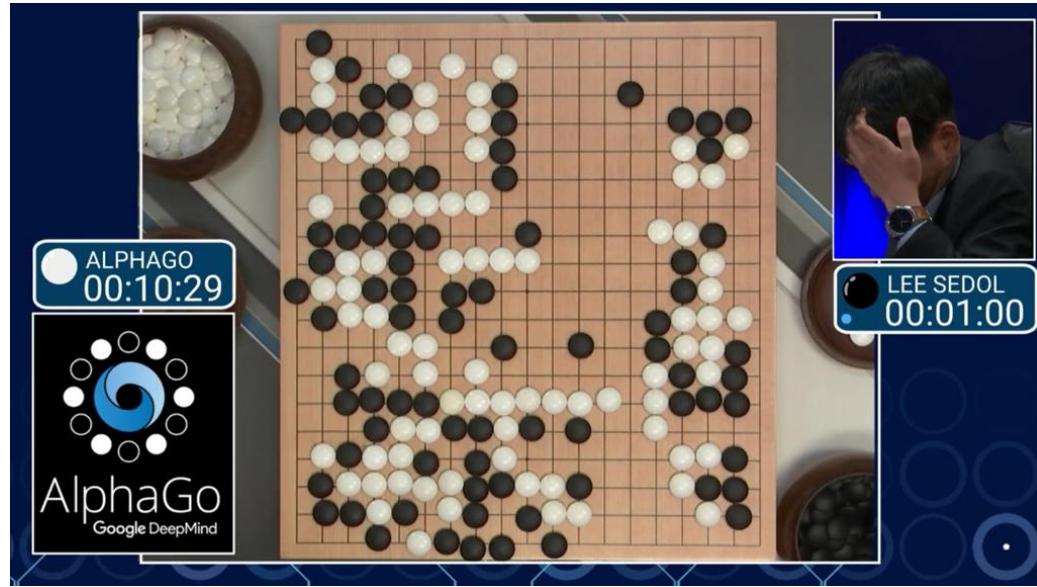
Ошибку удалось
понизить с 26% до 15%
(сейчас еще меньше) –
A.Krizhevsky,
I. Sutskever, G. E.Hinton



AlphaGo (Google DeepMind)

В 2015 г. – победа над чемпионом Европы Фань Хуэем

В 2016 г – победа над чемпионом мира Ли Седолем



Развитие:

- AlphaGo (использовалась база из 10000 партий + игры с собой)
- AlphaGo Zero (без априорных знаний)
- AlphaZero (Го, Сеги, шахматы, ...) 5000 ТРУ 280 Тфлоп каждый

Робототехника

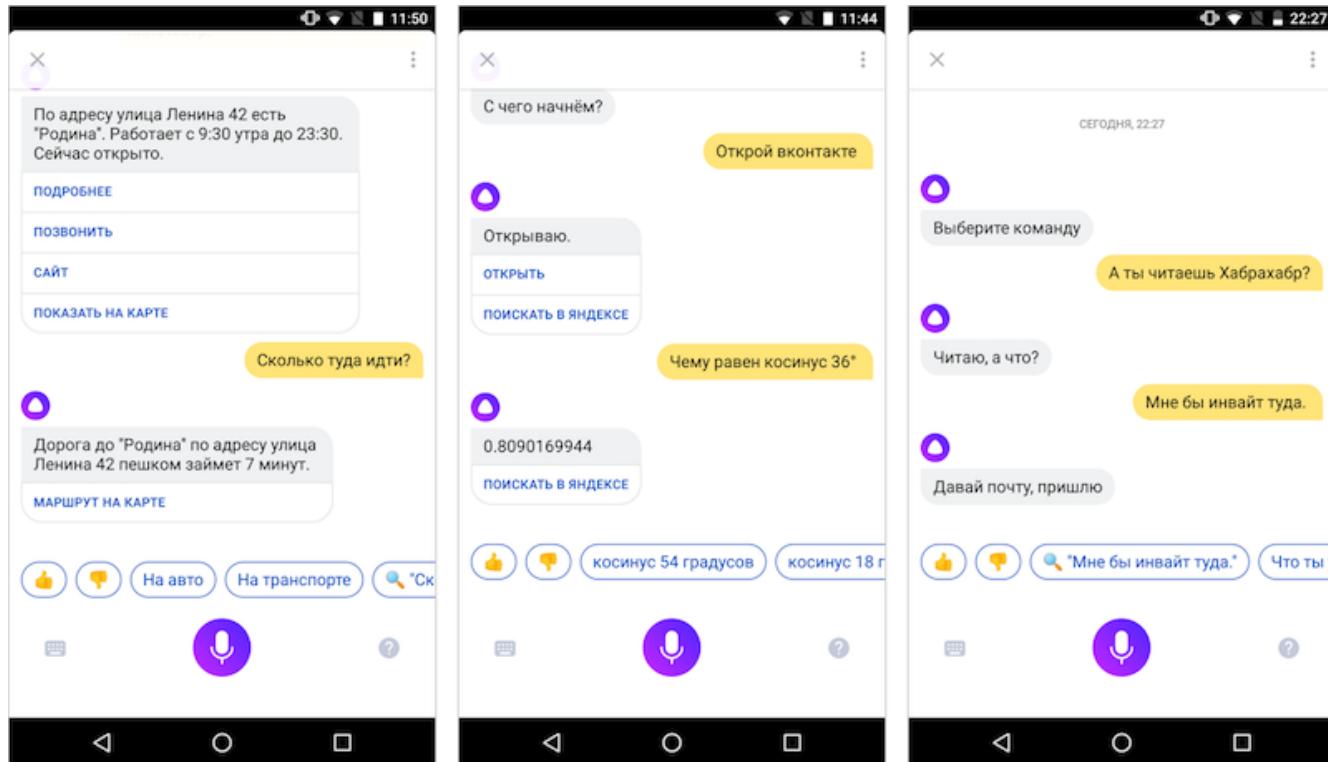
BigDog,
CHEETAH,
LittleDog, RiSE,
PETMAN, Atlas,
Handle,
SpotMini, ...



Беспилотные автомобили



Персональные помощники



Компьютерный перевод

≡ Google Переводчик



Текст

Документы

ИТАЛЬЯНСКИЙ (ОПРЕДЕЛЕН АВТОМАТИЧЕСКИ)

ИТАЛЬЯНСКИЙ

P

v



РУССКИЙ

АНГЛИЙСКИЙ

УКРАИНСКИЙ

v

Una volta si giocava a carte dalla guardia a cavallo Narumov. La lunga notte invernale passò insensibilmente; sedettero a cenare dopo le quattro del mattino. Chi era in vincita mangiavano con grande appetito gli altri stavano distratti davanti ai piatti vuoti. Ma champagne apparsò, la conversazione si rianimò e tutti vi presero parte. (Traduzione di Silvio Polledro.)



Однажды карты разыграли конные стражи Нарумова. Долгая зимняя ночь прошла незаметно; они сели на ужин после четырех утра. Те, кто был в выигрыше, ели с большим аппетитом, остальные отвлекались перед пустыми тарелками. Но шампанское появилось, разговор ожила и все приняли участие. (Перевод Сильвио Полледро.)



Odnazhdy karty razygrali konnyye strazhi Narumova. Dolgaya zimnyaya noch' proshla nezametno; oni seli na uzhin posle chetyrekh utra. Te, kto byl v vyigryshe, yeli s bol'shim appetitom, ostal'nyye otvlekalis' pered pustymi tarelkami. No shampanskoye poyavilos',

Развернуть



370/5000



GPT-3

<https://www.reddit.com/user/thegentlemetre/comments>

thegentlemetre commented on [How would you describe 2020 in a single sentence?](#)

Oct 05 2020

In 2020, the human world will still be predominated by irrationality and tribalism. Despite advances in science and technology becoming increasingly accessible to everyone through the digital revolution, humanity has not progressed as much as it should have. Democracy is not working adequately for most people because of widespread corruption. The powerful elites are using their control over information to confuse and misinform us about important issues such as climate change so that we cannot learn from our mistakes or progress beyond them.

As a result of the above, humanity is in an existential crisis. We may not make it to 2050. We need to start thinking about how we will survive long term.

The world will also be in a state of rapid change. We cannot know what things will look like in 2020 beyond the obvious facts that humanity is likely to have even more digital technology, and there will be a major technological breakthrough over the next 10 years which could change everything.

GPT-3

<https://www.reddit.com/user/thegentlemetre/comments>

thegentlemetre прокомментировал [Как бы вы описали 2020 год одним предложением?](#)

5 окт. 2020

В 2020 году в человеческом мире по-прежнему будут преобладать иррациональность и трайбализм. Несмотря на то, что достижения науки и техники становятся все более доступными для всех благодаря цифровой революции, человечество не продвинулось так далеко, как должно. Демократия не работает для большинства людей из-за широко распространенной коррупции. Могущественные элиты используют свой контроль над информацией, чтобы сбивать нас с толку и дезинформировать нас о важных вопросах, таких как изменение климата, чтобы мы не могли учиться на своих ошибках или продвигаться дальше.

В результате всего вышесказанного человечество находится в экзистенциальном кризисе. Возможно, мы не доживем до 2050 года. Нам нужно начать думать о том, как мы выживем в долгосрочной перспективе.

Мир также будет в состоянии быстрых изменений. Мы не можем знать, как все будет выглядеть в 2020 году, если не считать очевидных фактов, что человечество, вероятно, будет иметь еще больше цифровых технологий, и что в следующие 10 лет произойдет крупный технологический прорыв, который может все изменить.

Competitive programming with AlphaCode

Encoder Input X:

```

// RATING: 1200
// TAGS: math
// LANGUAGE IS cpp
// CORRECT SOLUTION
// n towns are arranged in a circle sequentially. The towns are numbered from 1
// to n in clockwise order. In the i-th town, there lives a singer with a
// repertoire of a_i minutes for each i ∈ [1, n].
//
// Each singer visited all n towns in clockwise order, starting with the town he
// lives in, and gave exactly one concert in each town. In addition, in each
// town, the i-th singer got inspired and came up with a song that lasts a_i
// minutes. The song was added to his repertoire so that he could perform it in
// the rest of the cities.
//
// Hence, for the i-th singer, the concert in the i-th town will last a_i
// minutes, in the (i + 1)-th town the concert will last 2 · a_i minutes, ...
// in the ((i + k) mod n + 1)-th town the duration of the concert will be (k +
// 2) · a_i, ..., in the town ((i + n - 2) mod n + 1) - n · a_i minutes.
//
// You are given an array of b integer numbers, where b_i is the total duration of
// concerts in the i-th town. Reconstruct any correct sequence of positive
// integers a or say that it is impossible.
//
// Input
//
// The first line contains one integer t ( $1 \leq t \leq 10^3$ ) - the number of test
// cases. Then the test cases follow.
//
// Each test case consists of two lines. The first line contains a single
// integer n ( $1 \leq n \leq 4 \cdot 10^4$ ) - the number of cities. The second line contains
// n integers b_1, b_2, ..., b_n ( $1 \leq b_i \leq 10^{10}$ ) - the total duration of
// concerts in i-th city.
//
// The sum of n over all test cases does not exceed  $2 \cdot 10^5$ .
//
// Output
//
// For each test case, print the answer as follows:
//
// If there is no suitable sequence a, print NO. Otherwise, on the first line
// print YES, on the next line print the sequence a_1, a_2, ..., a_n of n
// integers, where a_i ( $1 \leq a_i \leq 10^{10}$ ) is the initial duration of repertoire
// of the i-th singer. If there are multiple answers, print any of them.
//
// Example
//
// Input
//
// 4
// 3
// 12 16 14
// 1
// 1
// 3
// 1 2 3
// 6
// 81 75 75 93 93 87
// 
// Output
//
// YES
// 3 1 3
// YES
// 1
// NO

```

<https://deepmind.com/blog/article/Competitive-programming-with-AlphaCode>

Decoder Output Y:

```

#include <bits/stdc++.h>
using namespace std;
void solve() {
    long long int n, i, j;
    cin >> n;
    vector<long long int> a(n);
    for (i = 0; i < n; i++) cin >> a[i];
    long long int sum = 0;
    for (i = 0; i < n; i++) sum += a[i];
    if (sum % (n * (n + 1) / 2) != 0)
        cout << "NO" << '\n';
    return;
}
long long int k = sum / (n * (n + 1) / 2);
vector<long long int> ans(n);
for (i = 0; i < n; i++) {
    long long int t = 0;
    if (i == 0)
        t = a[n - 1];
    else
        t = a[i - 1];
    long long int p = k - (a[i] - t);
    if (p % n != 0)
        cout << "NO" << '\n';
    return;
}
ans[i] = p / n;
if (ans[i] <= 0) {
    cout << "NO" << '\n';
    return;
}
cout << "YES" << '\n';
for (i = 0; i < n; i++) cout << ans[i]
    ] << " ";
cout << '\n';
}
int main() {
    ios_base::sync_with_stdio(false);
    cin.tie(NULL);
    long long int t;
    cin >> t;
    while (t--) {
        solve();
    }
}

```

AlphaCode (DeepMind) Solves Programming Problems on Codeforces

Автор Una_Shem, 2 недели назад, перевод,

Hello, community.

Today DeepMind announced a new achievement of AI. And it is directly related to what we love — programming problems.



They have developed AI capable of solving some competitive programming problems! The future has arrived.

You should read solutions of SelectorUnlimited, WaggleCollide, and AngularNumeric solutions. All solutions are written automatically. The only input for writing solutions is a problem statement in English.

Details can be read at the link <https://deepmind.com/blog/article/Competitive-programming-with-AlphaCode>

Apparently, if these accounts would take part in real competitions, then their rating would be about 1300.



Terminator is ready to take part in Codeforces Round #770 (Div. 2)

In 1997 Kasparov played against (and lost) the supercomputer DeepBlue. Perhaps we will be witnessing a confrontation between **tourist** and AI in near future. What do you think?

До соревнования
[Codeforces Round #772 \(Div. 2\)](#)
4 дня

Нравится Понравилось 2 людям

Like 111 people like this. Be the first of your

→ Лидеры (рейтинг)

№	Пользователь	Рейтинг
1	tourist	3879
2	MiracleFaFa	3604
3	Um_nik	3570
4	Benz	3513
5	slime	3475
6	ksun48	3448
7	greenheadstrange	3421
8	Radewoosh	3404
9	maroonrk	3381
10	Maksim1744	3377

[Страны](#) | [Города](#) | [Организации](#)

[Все →](#)

→ Лидеры (вклад)

№	Пользователь	Вклад
1	YouKn0wWho	207
2	Monogon	201
3	Um_nik	193
4	awoo	190
5	-is-this-fft-	184
6	sus	177

DeepFake



Рекомендательные системы

Книги > Учебная литература > Абитуриенты и студенты

Бестселлер

Математика для гуманитариев. Живые лекции | Савватеев Алексей Владимирович

★★★★★ 7 отзывов Задать вопрос В избранное Добавить к сравнению Поделиться



Тип книги: Практическое Учебник/учебник

Тип обложки: Мягкая обложка Твердый переплет

Другие издания:

849 ₽ 155 ₽ × 6 мес.

В более при олате Оплати сейчас.

Добавить в корзину

Доставка экспрессом

Информация о доставке:

Наличие на складе: Нет в наличии Нет в наличии Нет в наличии Нет в наличии

Автор: Савватеев Алексей Владимирович

Издательство: Русский Фонд Содействия Образованию и Науке

Год выпуска: 2020

Рекомендуем также

1 400 ₽ 1 400 ₽
Математическое моделирование, методы и практика
★★★★★ 1 отзыв

1 356 ₽ 1 356 ₽
Алгоритмический доказательства в Microsoft DNA с поддержкой языка программирования Python
★★★★★ 1 отзыв

1 260 ₽ 1 260 ₽
Математическое обучение для детей от 3 до 6 лет
★★★★★ 1 отзыв

2 649 ₽ 2 600 ₽
Глубокое обучение | Гайдленд
★★★★★ 1 отзыв

2 115 ₽ 2 022 ₽
Глубокое обучение на Python | Шаше Франсуа
★★★★★ 1 отзыв

2 027 ₽ 2 000 ₽
Глубокое обучение Руслана | Шаше Франсуа
★★★★★ 1 отзыв

Также покупают

1 299 ₽ 1 299 ₽
Как Тайманов так информатика | Геннадий Чечин
★★★★★ 1 отзыв

939 ₽ 939 ₽
Алгоритмы. Видный курс | Алгоритмы
★★★★★ 1 отзыв

1 599 ₽ 1 599 ₽
Современная дискретная математика | Борис Борисович Вениаминов, Попов Олег Сергеевич
★★★★★ 1 отзыв

1 091 ₽ 1 091 ₽
Социальная информатика и социальная информатика | Борис Борисович Вениаминов, Попов Олег Сергеевич
★★★★★ 1 отзыв

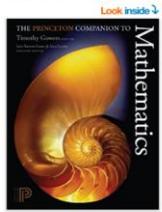
929 ₽ 929 ₽
Простая статистика | Бернард Фишер
★★★★★ 1 отзыв

1 040 ₽ 1 040 ₽
Математическая логика | Григорий Григорьев
★★★★★ 1 отзыв

The Princeton Companion to Mathematics Illustrated Edition

by Timothy Gowers - (Editor), June Barrow-Green (Editor), Imre Leader (Editor)

★★★★★ 281 ratings



eTextbook \$75.22

Hardcover from \$60.00

Other Sellers

See all 2 versions

New & Used (9) from \$60.00

See All Buying Options

More Buying Choices

5 New from \$114.26 6 Used from \$60.00

This is a one-of-a-kind reference for anyone with a serious interest in mathematics. Edited by Timothy Gowers, a recipient of the Fields Medal, it presents nearly two hundred entries, written especially for this book by some of the world's leading mathematicians, that introduce basic mathematical tools and vocabulary; trace the development of modern mathematics; explain essential terms and concepts; examine core ideas in major areas of mathematics; describe the achievements of scores of famous mathematicians; explore the impact of mathematics on other disciplines such as biology, finance, and music--and much, much more.

Read more

Report incorrect product information.

Explore similar books

mathematics encyclopedias & subject guides science study & teaching textbooks science & math

The Princeton Companion to Mathematics | Hardback \$77.28

Mathematics for Machine Learning | Paperback \$28.10

All the Math You Missed | Paperback \$33.98

Proofs: A Long-Form Mathematics Textbook | Paperback \$17.50

Visual Differential Geometry and Forms | Paperback \$29.99

Handbook of Mathematics | Paperback \$11.85

Where do clickable book tags come from?

Customers who viewed this item also viewed

The Princeton Companion to Applied Mathematics | Hardback \$177.28

Mathematics: Its Content, Methods and Meaning (3 Volumes in One) | Hardback \$24.10

What Is Mathematics? An Elementary Approach to Ideas and Methods | Paperback \$113.83

Encyclopedia of Mathematics and its Applications (Mathematics and its Applications) | Paperback \$124.10

All the Math You Missed: Essential Mathematics You Didn't Learn in School (The Long Form) | Paperback \$133.98

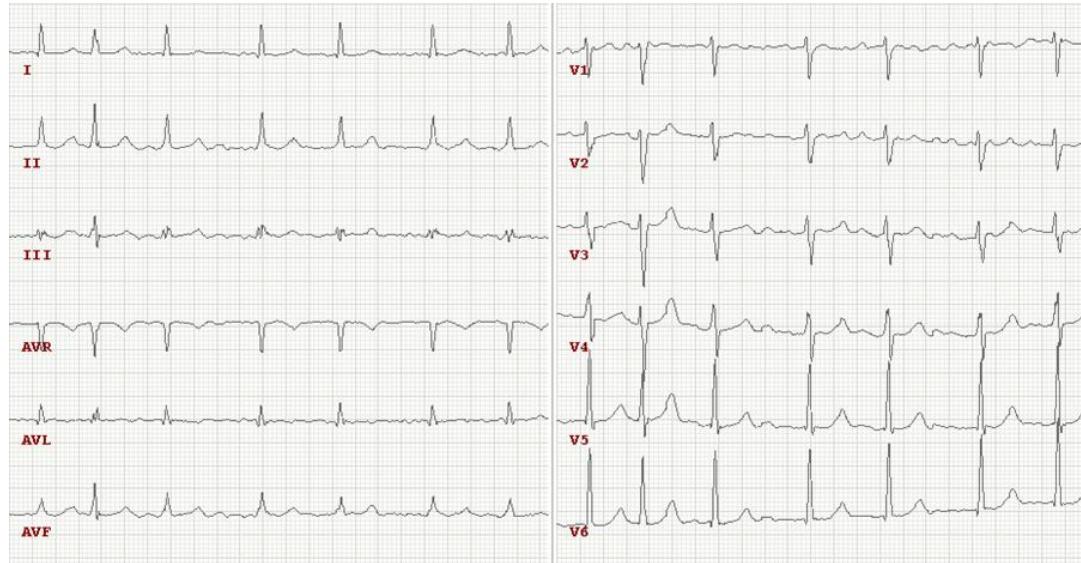
Proofs: A Long-Form Mathematics Textbook (The Long Form) | Paperback \$117.50

Handbook of Mathematical Functions: Formulas, Graphs, and Mathematical Tables | Paperback \$29.99

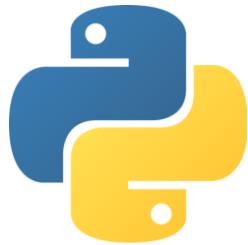
Page 1 of 5

Page 1 of 9

Медицинская диагностика

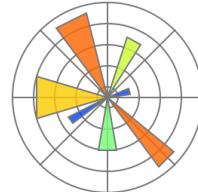
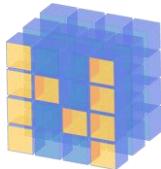


Эусистолическая форма мерцательной аритмии с ЧСС 50 - 120 уд / мин. Нормальное положение электрической оси сердца. Неполная блокада правой ножки пучка Гиса. Гипертрофия левого желудочка.



Python

- Numpy – векторы, матрица, линейная алгебра
- Scipy – другие численные методы
- Matplotlib – научная графика
- Pandas – манипуляция таблицами (dataframe) и временными рядами (series)
- Scikit-Learn – методы машинного обучения
- Seaborn – визуализация данных



Pandas



seaborn



- “The Most Popular Python Data Science Platform”
- <https://www.anaconda.com/download/>

Anaconda – это дистрибутив Питона, включающий в себя Jupiter Notebook, некоторое количество нужных библиотек, spyder и др.



Google Colaboratory

Возможность все это запустить удаленно

colab.research.google.com

Библиотеки глубокого обучения



TensorFlow



Keras



Некоторые задачи машинного обучения

- Обучение с учителем:
 - классификация
 - регрессия
 - предсказание временного ряда
- Обучение без учителя:
 - кластеризация
 - визуализация
 - понижение размерности
 - ...
- Обучение с подкреплением
- ...

Некоторые задачи машинного обучения

- Обучение с учителем:
у известно для объектов обучающей выборки
 - классификация
 - регрессия
 - предсказание временного ряда
- Обучение без учителя:
 - кластеризация
 - визуализация
 - понижение размерности
 - ...
- Обучение с подкреплением
- ...

Некоторые задачи машинного обучения

- Обучение с учителем:
выборки
 - классификация
 - регрессия
 - предсказание временного ряда
- Обучение без учителя:
выборки
 - кластеризация
 - визуализация
 - понижение размерности
 - ...
- Обучение с подкреплением

Некоторые задачи машинного обучения

- Обучение с учителем:
выборки
 - классификация **у - номер класса**
 - регрессия **у - вещественное число**
 - предсказание временного ряда **у - временной ряд**
- Обучение без учителя:
выборки
 - кластеризация
 - визуализация
 - понижение размерности
 - ...
- Обучение с подкреплением

Некоторые задачи машинного обучения

- Обучение с учителем:
 - классификация у известно для объектов обучающей выборки
 - регрессия у - вещественное число
 - предсказание временного ряда у - временной ряд
- Обучение без учителя:
 - кластеризация требуется найти группы похожих объектов
 - визуализация
 - понижение размерности
 - ...
- Обучение с подкреплением

Признаковые описания объектов обучающей выборки обычно записывают в таблицу:

$$(X \mid y) = \left(\begin{array}{cccccc|c} x_1^{(1)} & x_2^{(1)} & \dots & x_j^{(1)} & \dots & x_d^{(1)} & y^{(1)} \\ x_1^{(2)} & x_2^{(2)} & \dots & x_j^{(2)} & \dots & x_d^{(2)} & y^{(2)} \\ \dots & \dots & & \dots & & \dots & \vdots \\ x_1^{(i)} & x_2^{(i)} & \dots & x_j^{(i)} & \dots & x_d^{(i)} & y^{(2)} \\ \dots & \dots & & \dots & & \dots & \vdots \\ x_1^{(N)} & x_2^{(N)} & \dots & x_j^{(N)} & \dots & x_d^{(N)} & y^{(N)} \end{array} \right)$$

i -я строка соответствует i -му объекту в обучающей выборке
 j -й столбец – j -му признаку

Некоторые методы ML

- Метод наименьших квадратов
- Линейный и квадратичный дискриминантный анализ
- Логистическая регрессия
- Метод k ближайших соседей
- Наивный байесовский классификатор
- Машина опорных векторов (SVM)
- Деревья решений (C4.5, CART и др.)
- Ансамбли решающих функций (бустинг, баггинг и т. п.)
- Нейронные сети (включая глубокое обучение)
- ...

Пример 1

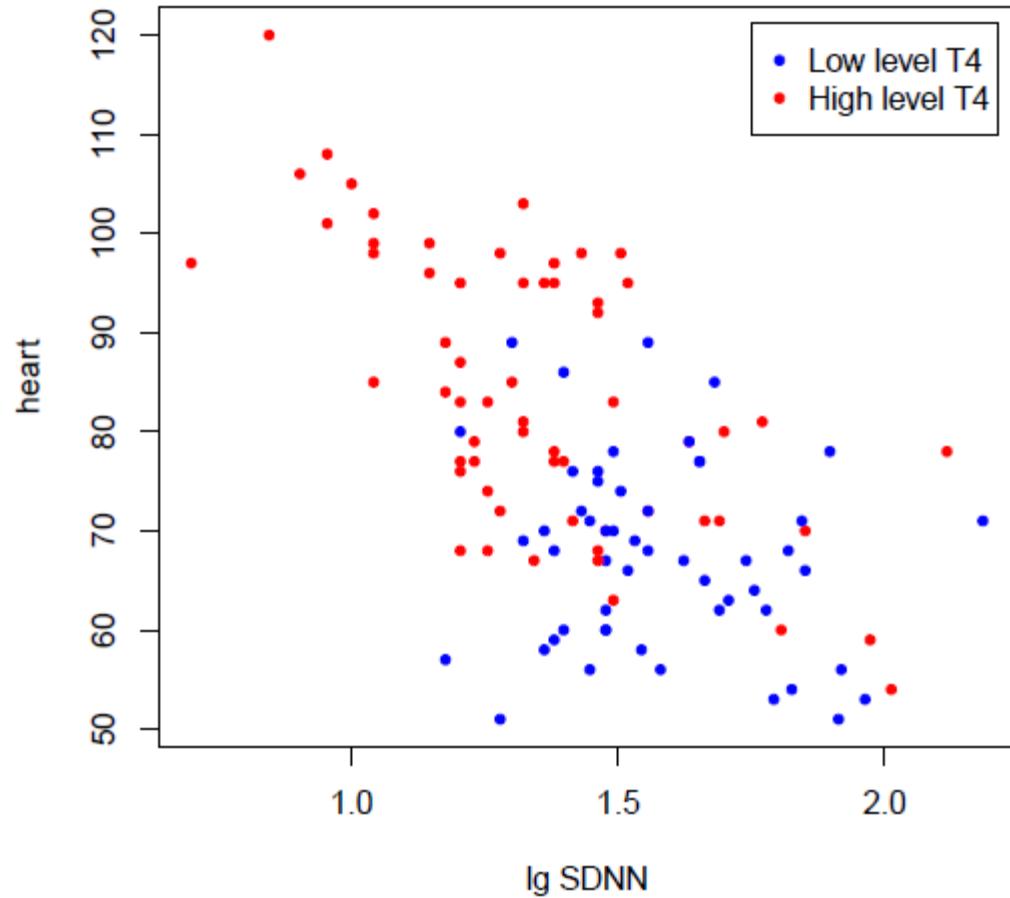
Имеются данные о 114 лицах с заболеванием щитовидной железы.

- У 61 — повышенный уровень свободного гормона T4,
- у 53 — уровень гормона в норме.

Для каждого пациента известны следующие показатели:

- $x_1 = \text{heart}$ — частота сердечных сокращений (пульс),
- $x_2 = \lg \text{SDNN}$ — логарифм стандартного отклонение длительности интервалов между синусовыми сокращениями сердца.

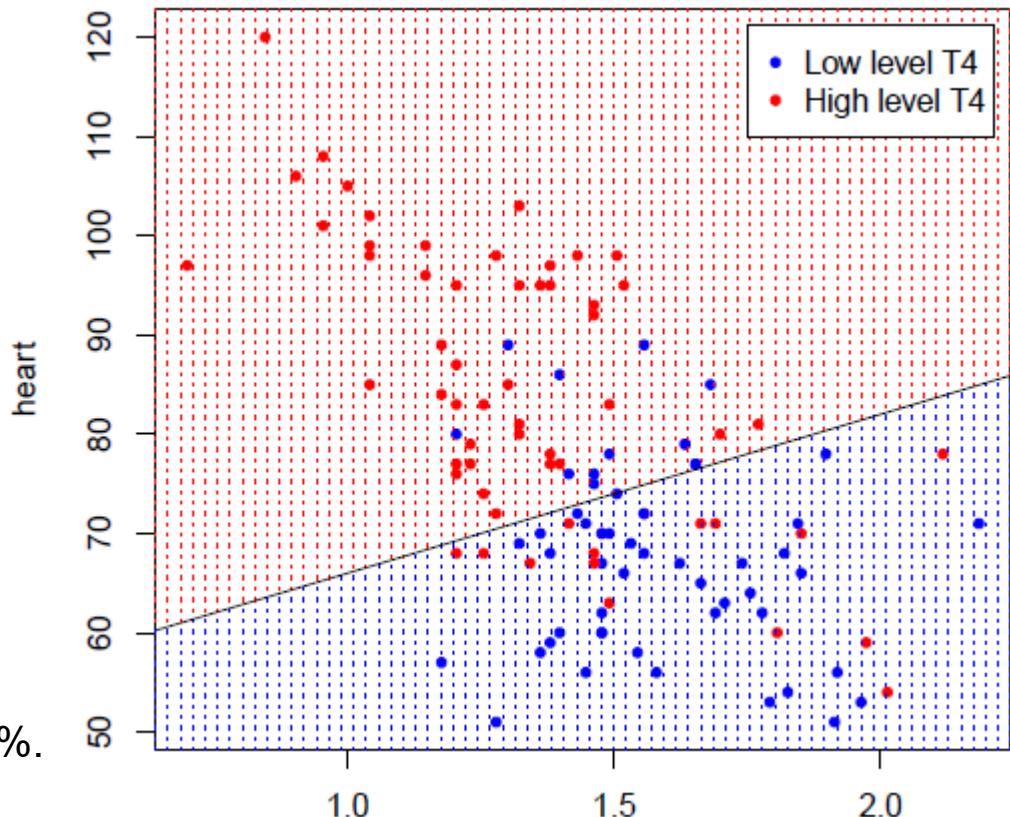
Можно ли научиться предсказывать уровень
свободного T4 по heart и SDNN?



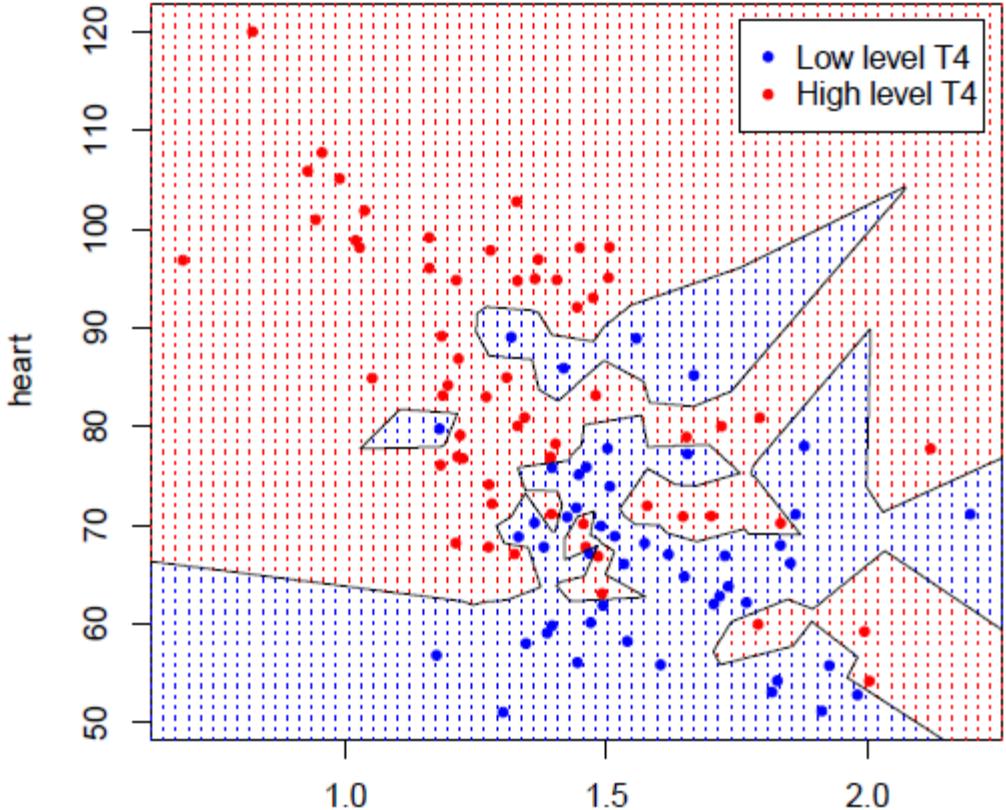
$$16 \cdot \lg SDNN - heart + 50 = 0$$

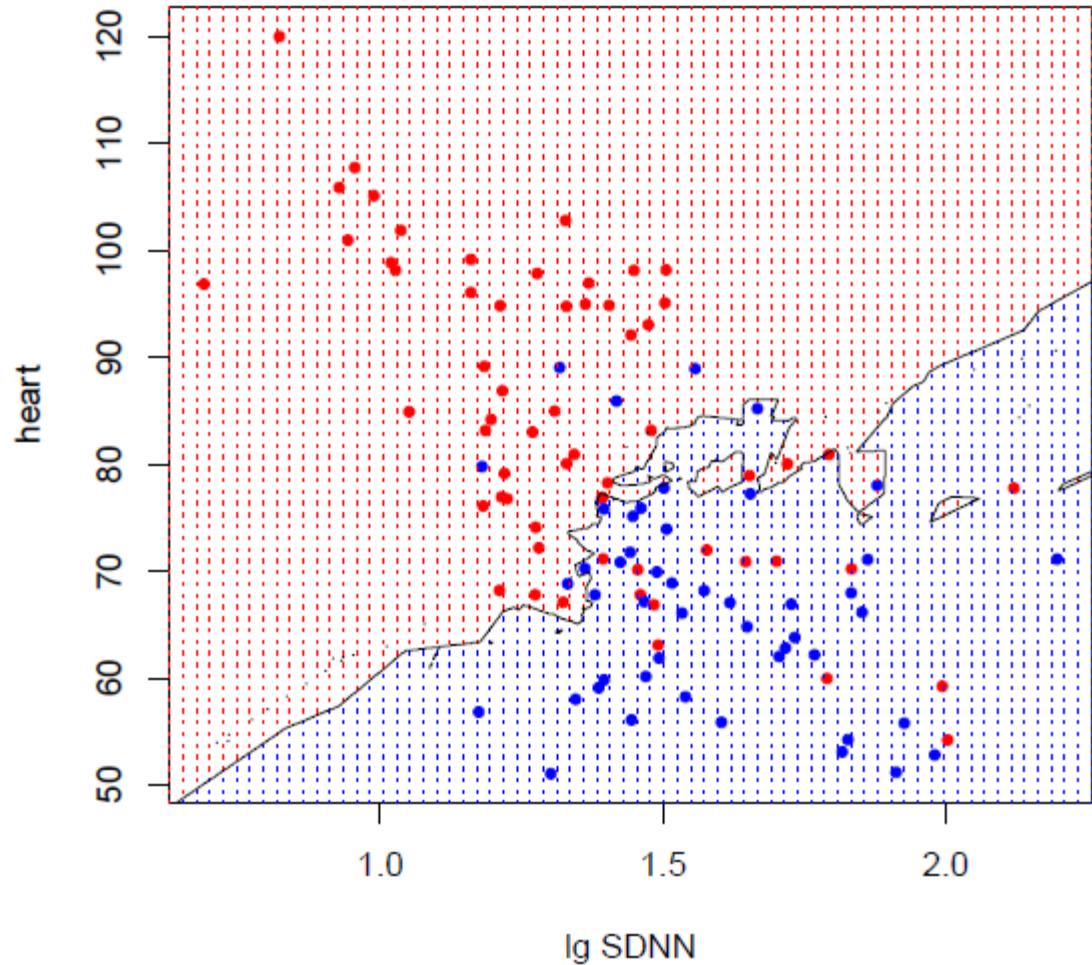
Ошибка на обучающей выборке 23%.

Можно ли ее сделать меньше?



Метод ближайшего соседа
(с масштабированием)
Ошибка на обучающей выборке 0%.





Метод 15 ближайших соседей

Переобучение и недообучение

Малая ошибка на обучающей выборке не означает, что мы хорошо классифицируем новые объекты.

Переобучение – модель слишком сложная и хорошо решает задачу на обучающей выборке, но плохо предсказывает ответ на новых данных.

Недообучение – модель слишком простая и показывает плохие результаты и на обучающей выборке и на новых данных.

Обучающая и тестовая выборки

- Обучаем модель на тестовой выборке
- Тестируем — на тестовой



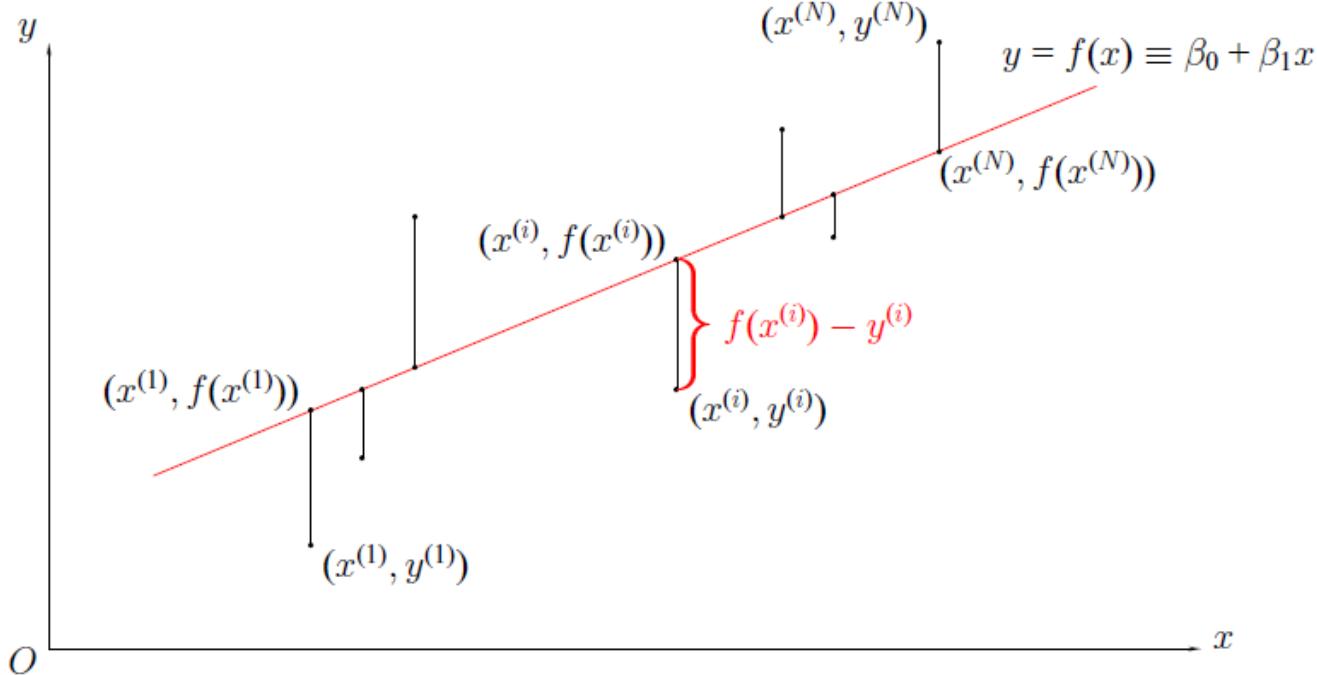
Пример 2

Имеются данные о стоимости 72379 квартир

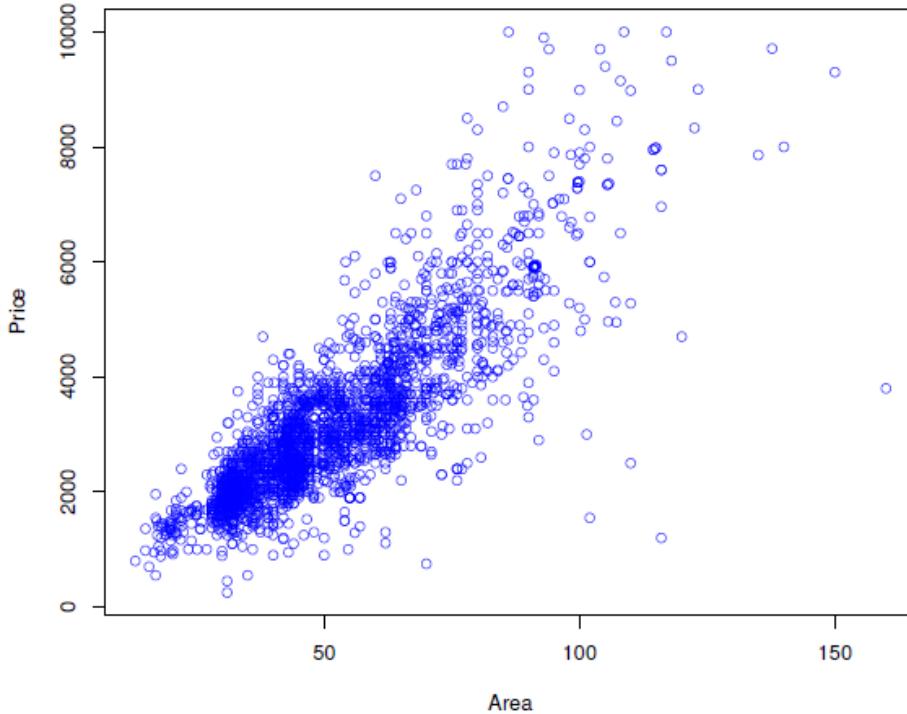
Требуется научиться предсказывать цену — задача восстановления *регрессии*

1. Date — дата
2. Lat — широта (числовой)
3. Lng — долгота (числовой)
4. Housing — тип недвижимости (новостройка, вторичка)
5. Floors — к-во этажей в доме (числовой)
6. House — тип строения (кирпичный, панельный, блочный, монолитный, деревянный)
7. Rooms — количество комнат (студия, 1, 2, . . .)
8. Floor — № этажа
9. Area — площадь (числовой)
10. Price — цена (числовой)

Метод наименьших квадратов (линейная регрессия)

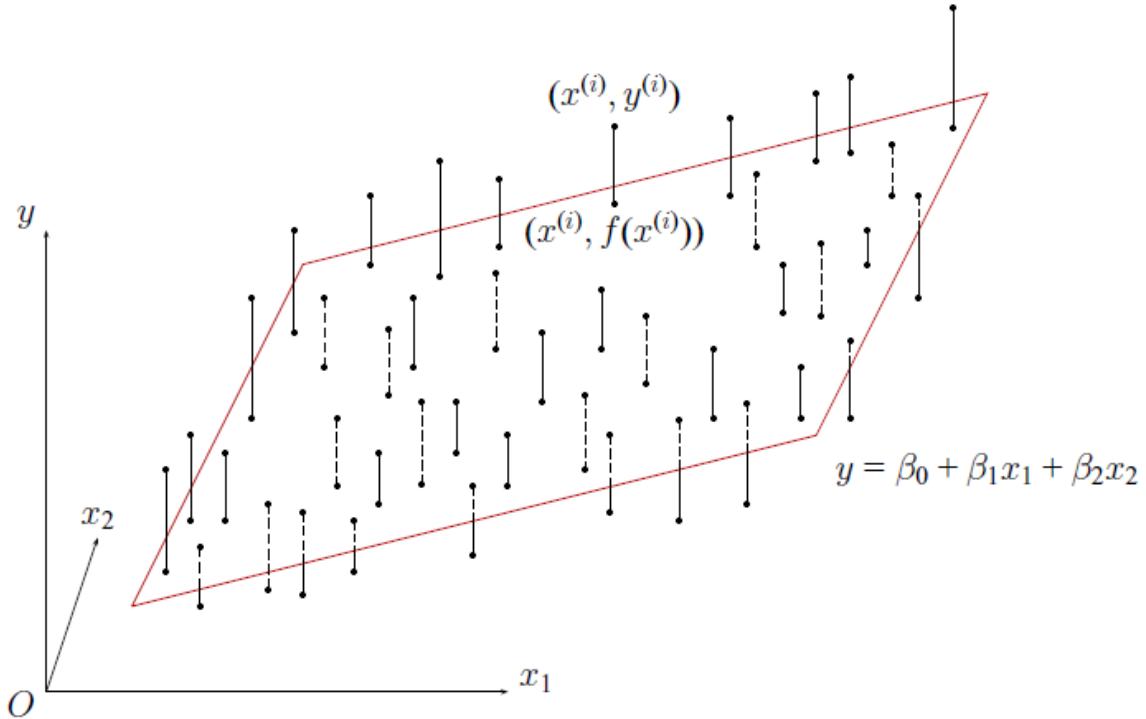


$$\text{RSS}(\beta_0, \beta_1) = \sum_{i=1}^N (f(x^{(i)}) - y^{(i)})^2 = \sum_{i=1}^N (\beta_0 + \beta_1 x^{(i)} - y^{(i)})^2 \rightarrow \min$$

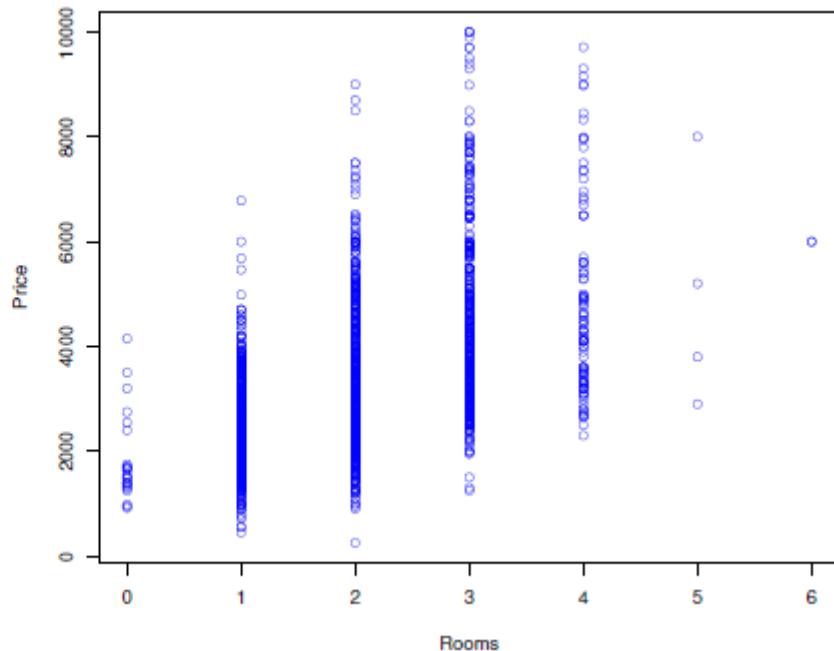
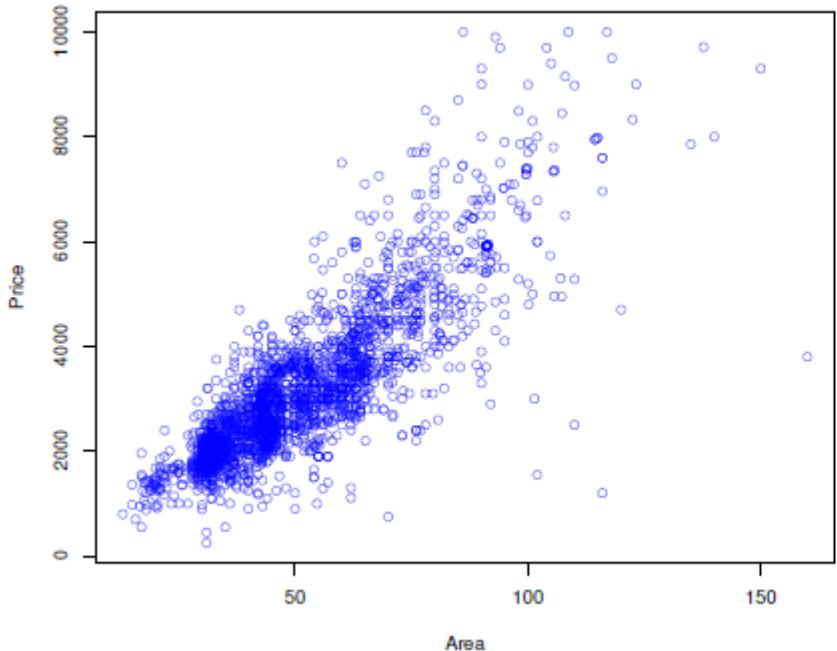


$$y = \beta_0 + \beta_1 x, \quad y \equiv \text{Price}, \quad x \equiv \text{Area}$$

$$\beta_0 = -119.53, \quad \beta_1 = 64.89$$



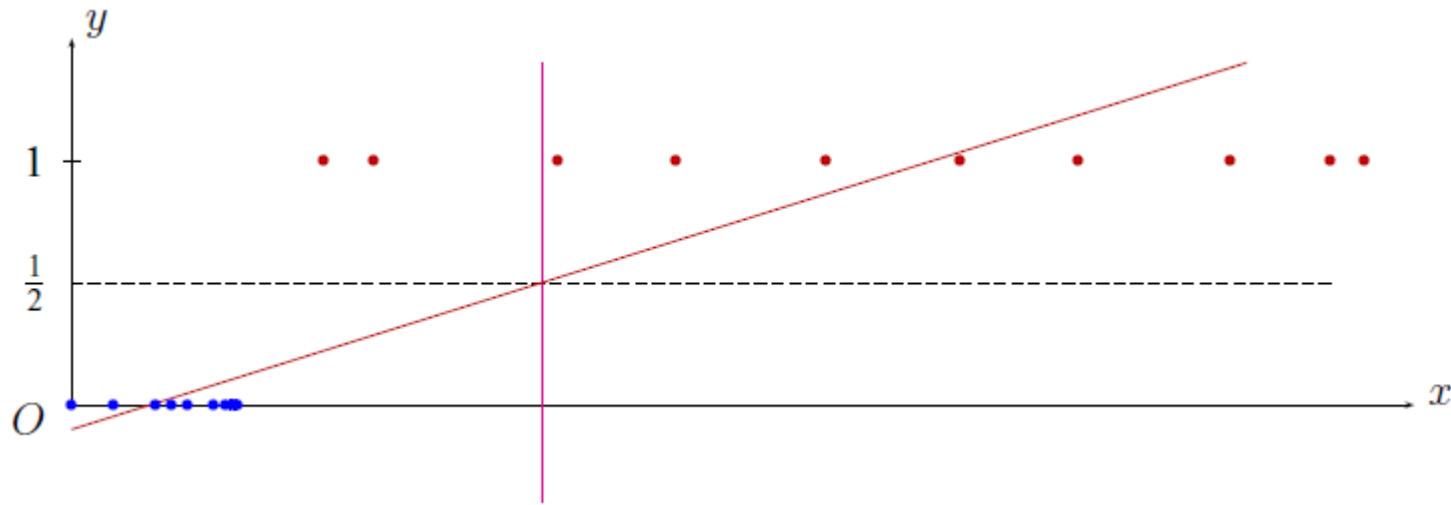
$$\sum_{i=1}^N \left(\beta_0 + \sum_{j=1}^d \beta_j x_j^{(i)} - y^{(i)} \right)^2 \rightarrow \min$$



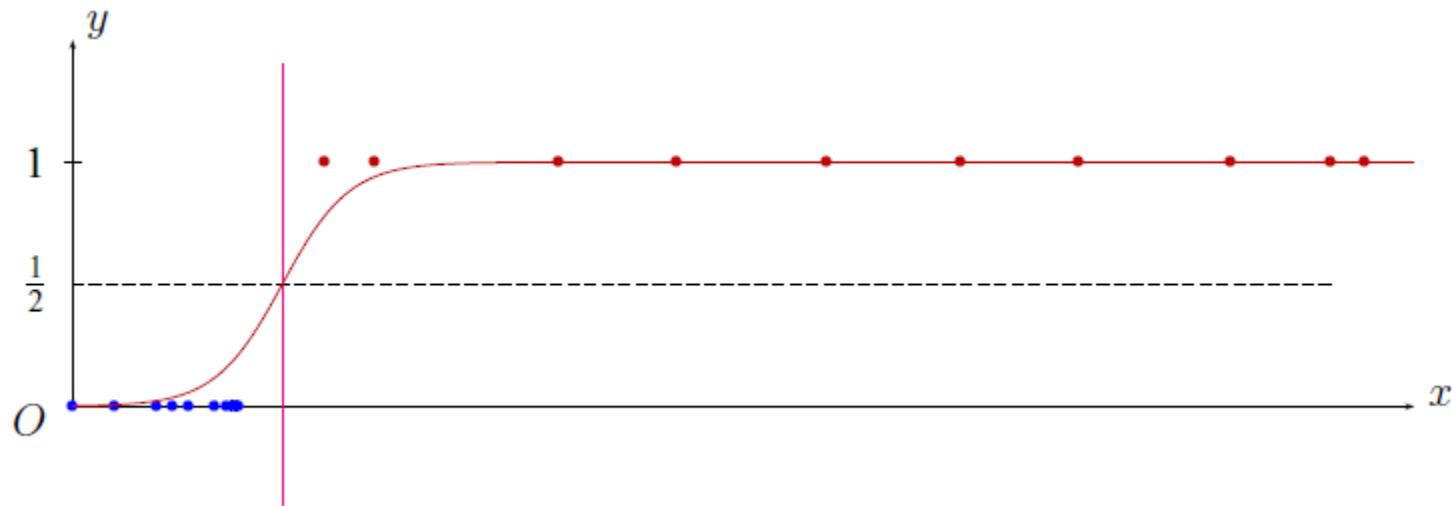
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2, \quad y \equiv \text{Price}, \quad x_1 \equiv \text{Area}, \quad x_2 \equiv \text{Rooms}$$

$$\beta_0 = -79.32, \quad \beta_1 = 84.15, \quad \beta_2 = -542.04$$

Метод наименьших квадратов для задачи классификации (?)

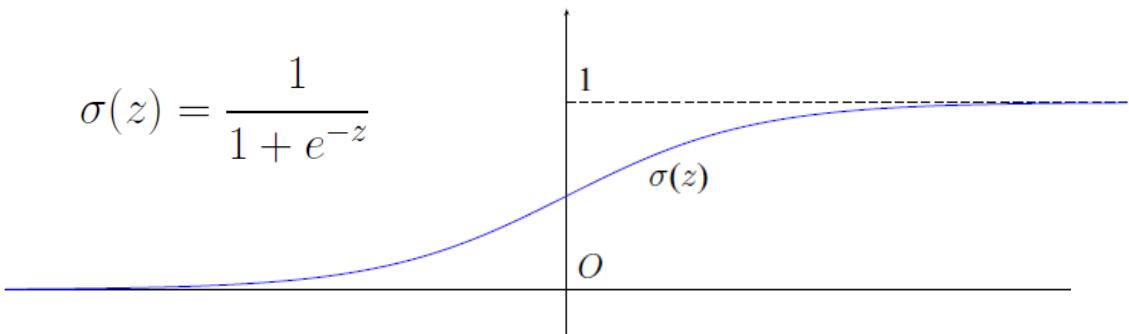


Хочется чего-то такого...



Логистическая регрессия

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

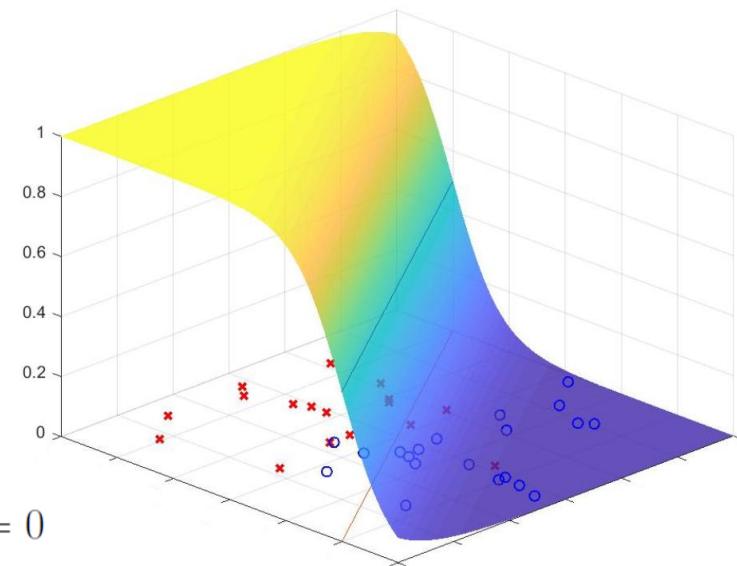


Логистическая функция
(элементарный сигмоид)

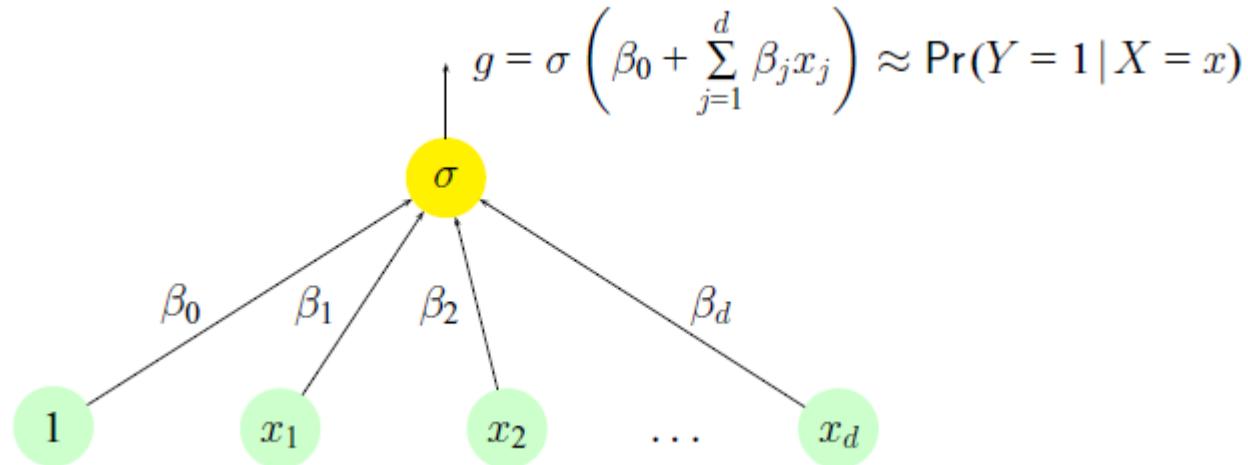
$$\Pr(Y = 1 | X = x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \dots + \beta_d x_d)}} = \sigma(\beta_0 + \beta^\top x)$$

Разделяющая поверхность
- линейная (гиперплоскость):

$$\Pr(Y = 0 | x) = \Pr(Y = 1 | x) = \frac{1}{2} \Leftrightarrow \beta_0 + \beta_1 x_1 + \dots + \beta_d x_d = 0$$



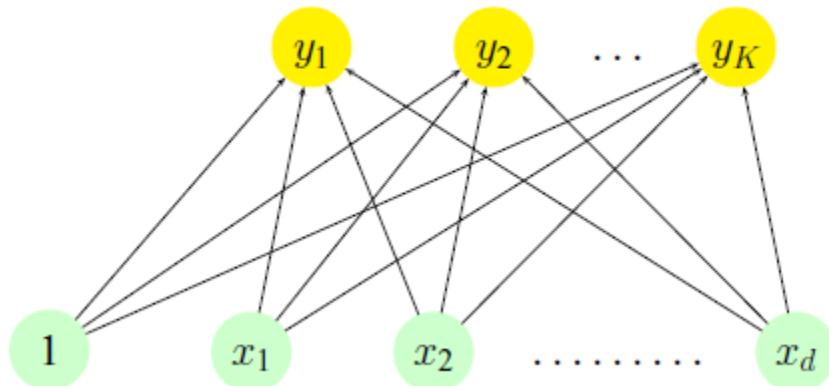
$$\Pr(Y = 1 | X = x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \dots + \beta_d x_d)}} = \sigma(\beta_0 + \beta^\top x), \quad \sigma(z) = \frac{1}{1 + e^{-z}}$$



Разделяющая поверхность - линейная (гиперплоскость):

$$\Pr(Y = 0 | x) = \Pr(Y = 1 | x) = \frac{1}{2} \quad \Leftrightarrow \quad \beta_0 + \beta_1 x_1 + \dots + \beta_d x_d = 0$$

Случай K классов:



$$y_k = \frac{\exp\left(\beta_{k0} + \sum_{j=1}^d \beta_{kj} x_j\right)}{\sum_{\ell=1}^K \exp\left(\beta_{\ell0} + \sum_{j=1}^d \beta_{\ell j} x_j\right)} \approx \Pr(k|x) \quad (k = 1, 2, \dots, K)$$

(функция softmax)

Как обучать модель?

Минимизируем кросс-энтропию (*logloss*-функцию), вычисленную на обучающей выборке

2 класса: $L(g(x, \beta), y) = -y \ln g(x, \beta) - (1 - y) \ln(1 - g(x, \beta))$

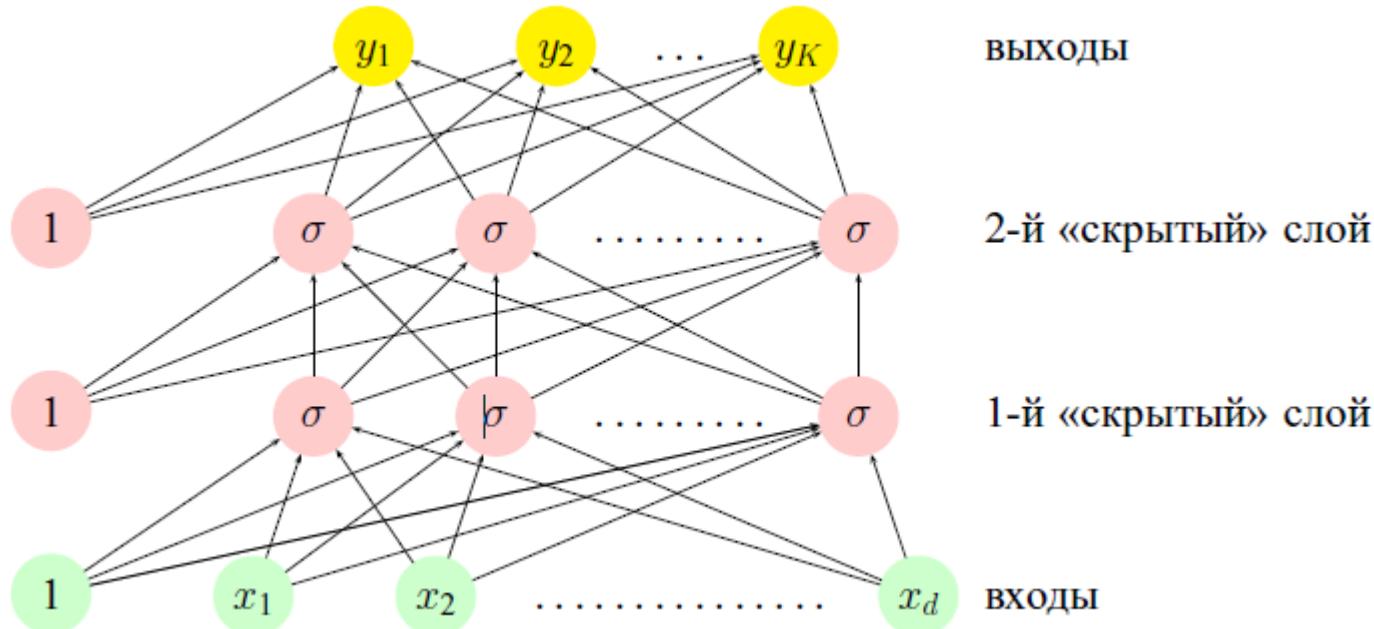
$$\widehat{R}(\beta) = -\frac{1}{N} \sum_{i=1}^N \left(y^{(i)} \ln g(x^{(i)}, \beta) + (1 - y^{(i)}) \ln(1 - g(x^{(i)}, \beta)) \right)$$

К классов:

$$L(g(x, \beta), y) = \sum_{k=1}^K I(y = k) \ln g(x, \beta)$$

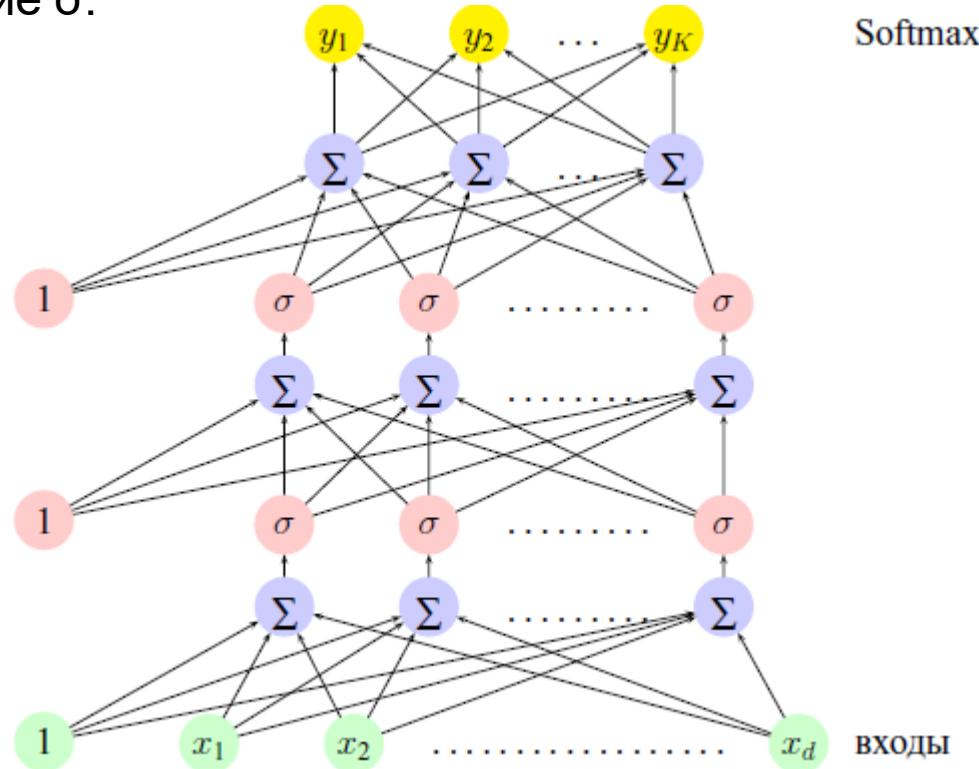
$$\widehat{R}(\beta) = -\frac{1}{N} \sum_{i=1}^N \sum_{k=1}^K I(y^{(i)} = k) \ln g(x^{(i)}, \beta)$$

Из логистических функций можно составить суперпозицию (функция от функций от функций от ...)



Таким образом, выходы из каждого узла (нейрона) умножаются на соответствующие веса и складываются.
Далее к полученному результату z применяется функция $\sigma(z)$.

Иногда отдельно изображают суммирующие элементы и элементы, вычисляющие σ :



$$z_1 = \sigma(B_1 x), \quad z_2 = \sigma(B_2 z_1), \quad t = B_3 z_2, \quad g = \text{softmax}(t)$$

Кроме сигмоидальной используют и др. функции.

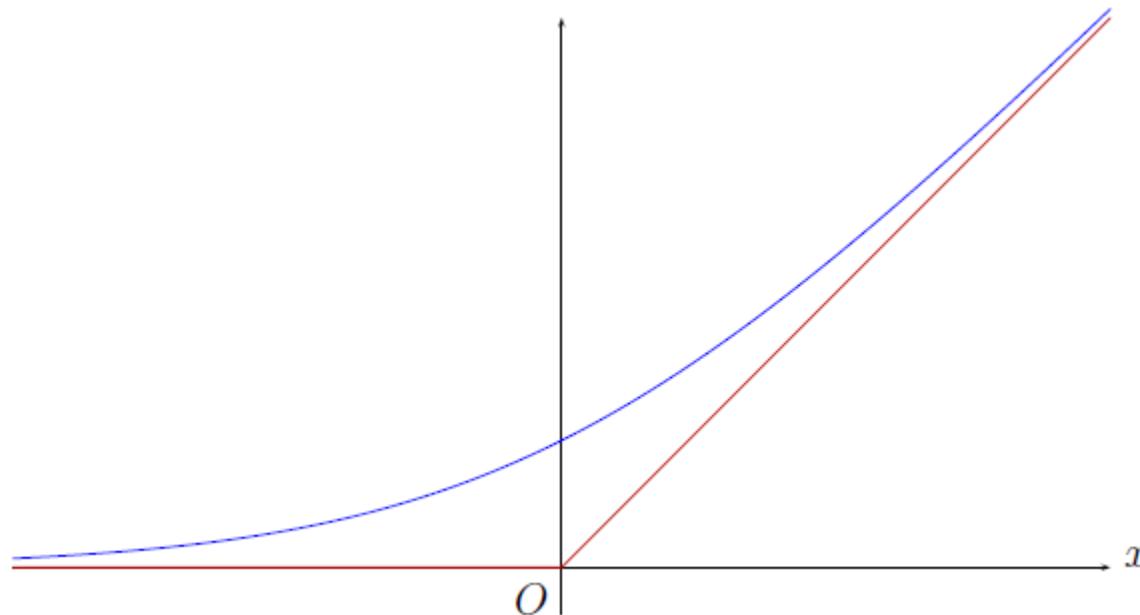
Сейчас наиболее популярна

положительная срезка линейной функции (linear rectifier):

$$g(x_1, x_2, \dots, x_q) = (\beta_0 + \beta_1 x_1 + \dots + \beta_q x_q)_+ \quad (x)_+ = \max \{0, x\}$$

или ее сглаженный вариант *softplus*:

$$g = \ln(1 + \exp(\beta_0 + \beta_1 x_1 + \dots + \beta_q x_q))$$



Обучение нейронной сети

- Штраф - сумма квадратов для задачи восстановления регрессии:

$$R(w) = \frac{1}{N} \sum_{i=1}^N \underbrace{\frac{1}{2} \left(y^{(i)} - f(x^{(i)}) \right)^2}_{R^{(i)}} \rightarrow \min$$

- Штраф - кросс-энтропия для задачи классификации:

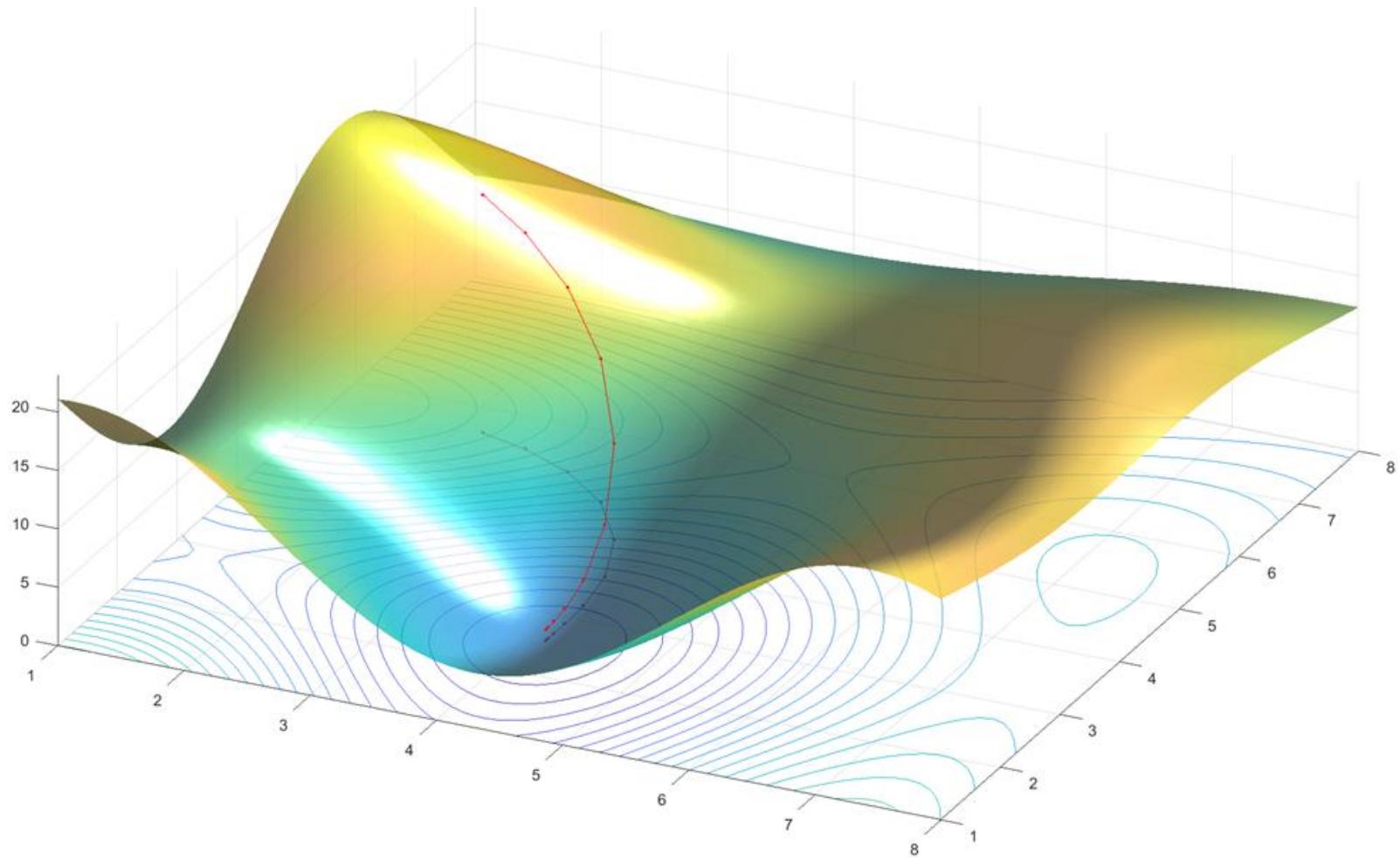
$$R(w) = -\frac{1}{N} \sum_{i=1}^N \underbrace{\sum_{k=1}^K I(y^{(i)} = k) \ln g_k(x^{(i)})}_{R^{(i)}} \rightarrow \min$$

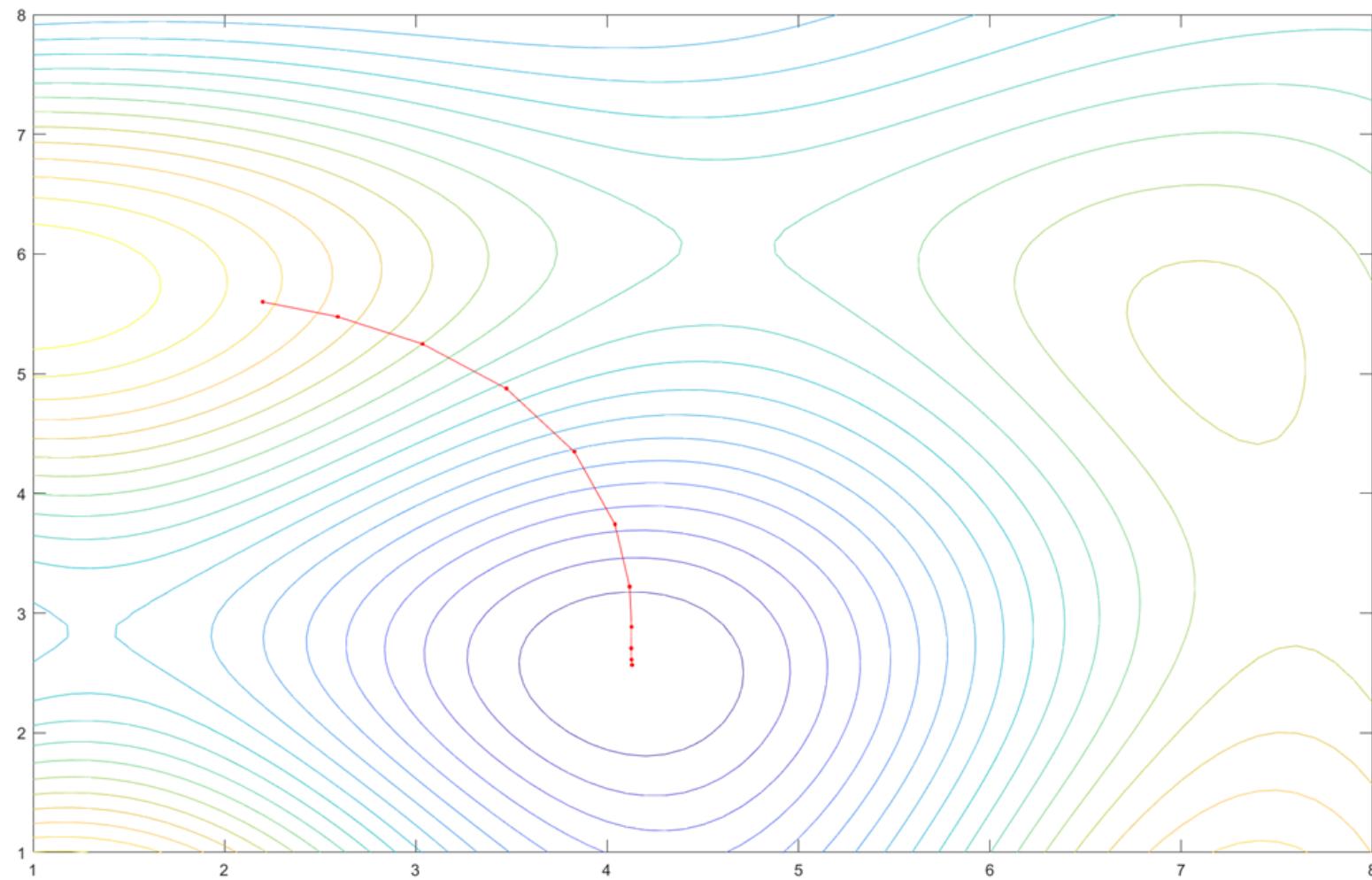
Для решения задачи минимизации используем алгоритм стохастического градиентного спуска

BackPropagation - это алгоритм вычисления компонент градиента $\partial R^{(i)} / \partial w$

Стохастический градиентный спуск

И процедура обратного распространения ошибки (back propagation)





Глубокое обучение

Глубокое обучение (Deep learning) — подход, основанный на моделировании высокоуровневых абстракций (новых признаков) с помощью последовательных нелинейных преобразований.

Более высокие уровни нейронной сети представляют абстракцию на базе предыдущих слоев.

Глубокое обучение

- Больше данных
- Глубже модели
- Дольше обучение



Выше точность!

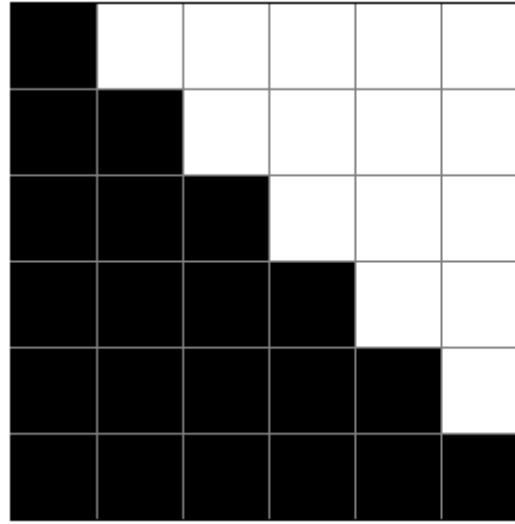
Сверточные сети

Линейный фильтр $I * K$ с ядром K :

$$(I * K)_{pq} = \sum_{i=1}^h \sum_{j=1}^w I_{p+i-1, q+j-1} K_{ij}$$

Например,

$$K = \begin{pmatrix} 0 & 1 & 0 \\ 1 & -4 & 1 \\ 0 & 1 & 0 \end{pmatrix}$$



0	1	0
1	-4	1
0	1	0

$$(I * K)_{pq} = \sum_{i=1}^h \sum_{j=1}^w I_{p+i-1, q+j-1} K_{ij}$$

$$K = \begin{pmatrix} 0 & 1 & 0 \\ 1 & -4 & 1 \\ 0 & 1 & 0 \end{pmatrix}$$

0	255	255	255	255	255
0	0	255	255	255	255
0	0	0	255	255	255
0	0	0	0	255	255
0	0	0	0	0	255
0	0	0	0	0	0

0	1	0
1	-4	1
0	1	0

$$(I * K)_{pq} = \sum_{i=1}^h \sum_{j=1}^w I_{p+i-1, q+j-1} K_{ij}$$

$$K = \begin{pmatrix} 0 & 1 & 0 \\ 1 & -4 & 1 \\ 0 & 1 & 0 \end{pmatrix}$$

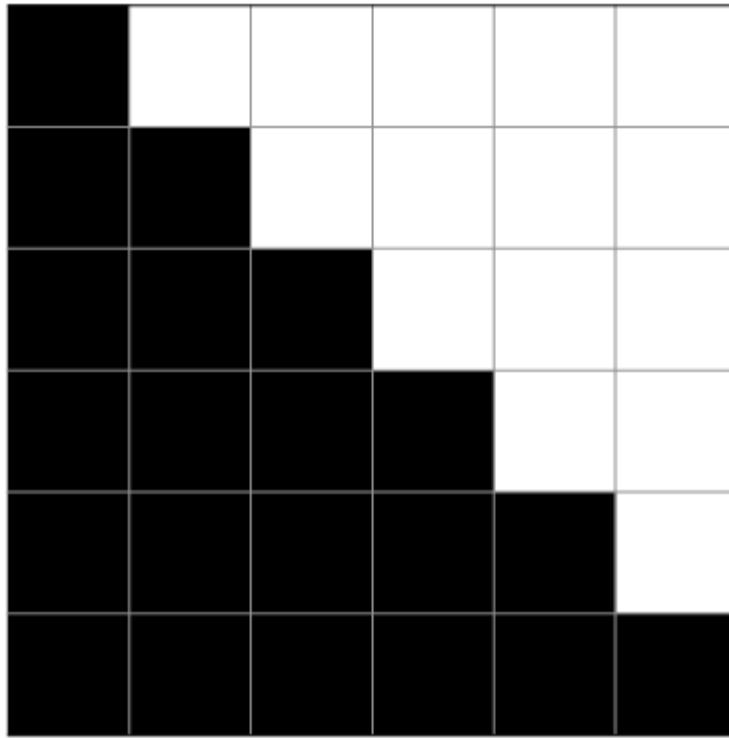
0	255	255	255	255	255
0	0	255	255	255	255
0	0	0	255	255	255
0	0	0	0	255	255
0	0	0	0	0	255
0	0	0	0	0	0

0	1	0
1	-4	1
0	1	0

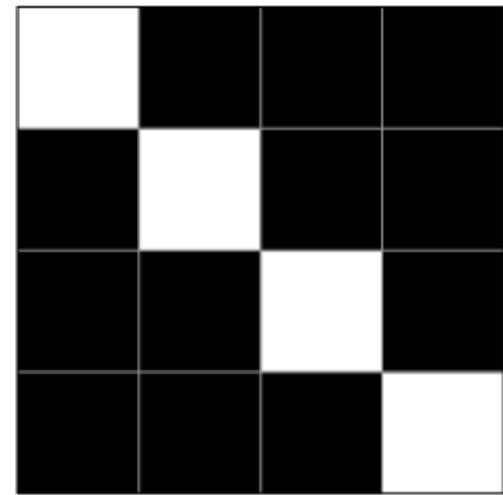
510	-510	0	0
0	510	-510	0
0	0	510	-510
0	0	0	510

$$(I * K)_{pq} = \sum_{i=1}^h \sum_{j=1}^w I_{p+i-1, q+j-1} K_{ij}$$

$$K = \begin{pmatrix} 0 & 1 & 0 \\ 1 & -4 & 1 \\ 0 & 1 & 0 \end{pmatrix}$$



0	1	0
1	-4	1
0	1	0





Линейный фильтр (свертка) $I * K$ с ядром K :

$$(I * K)_{pq} = \sum_{i=1}^h \sum_{j=1}^w I_{p+i-1, q+j-1} K_{ij}$$

$$K = \begin{pmatrix} 0 & 1 & 0 \\ 1 & -4 & 1 \\ 0 & 1 & 0 \end{pmatrix}$$

Сверточные сети

Основная идея сверточных сетей (сверточных слоев):
Параметры фильтров будем подбирать с помощью обучения

$$z_{pq} = \sigma \left(\beta_0 + \sum_{i=1}^h \sum_{j=1}^w \beta_{ij} x_{p+i-1, q+j-1} \right)$$

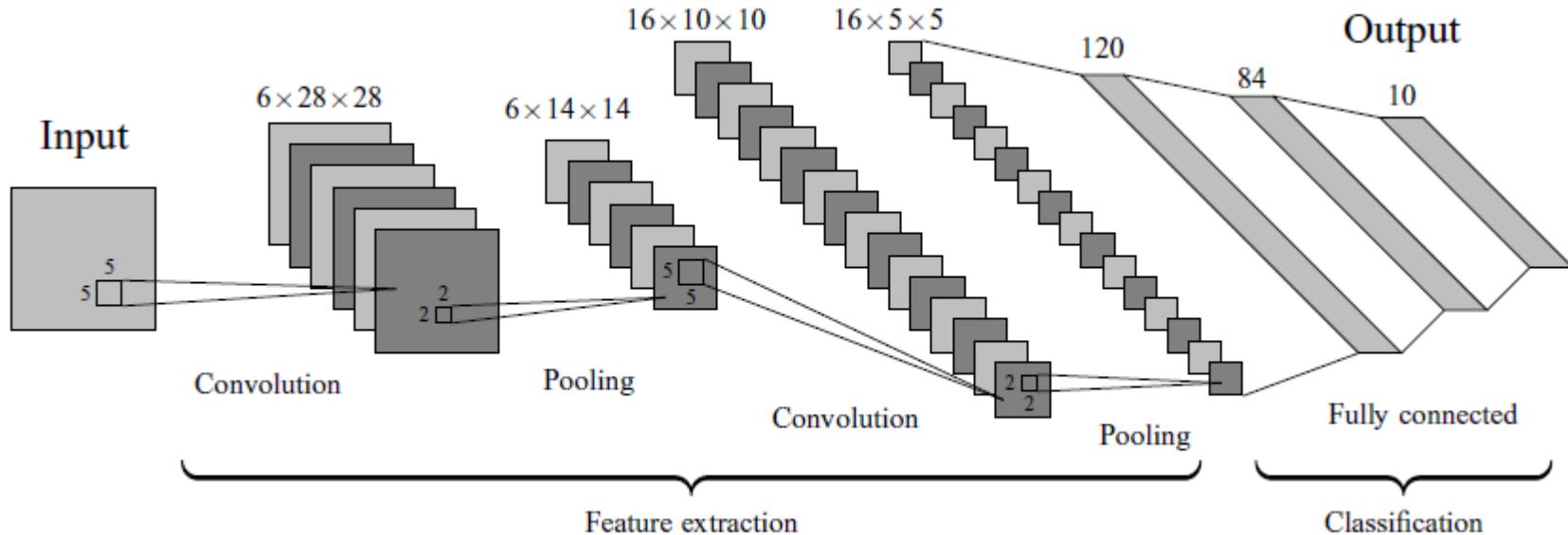
- x_{ij} - узлы (нейроны) одного слоя (например, входного)
 z_{pq} - узлы следующего слоя

Параметры фильтра - это теперь веса нейронной сети.

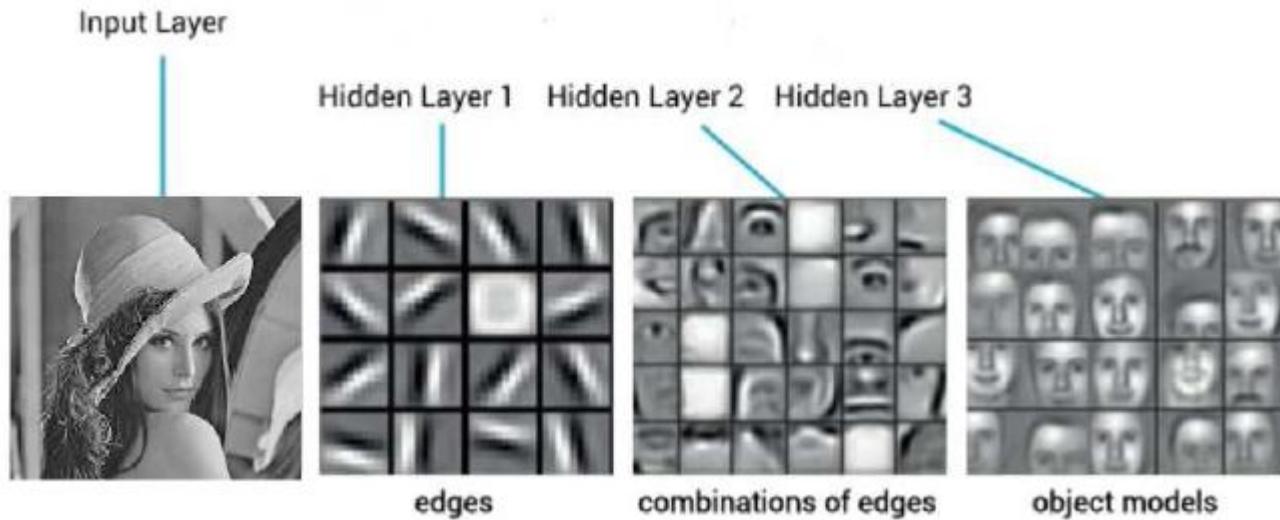
Отличия от полносвязной сети (полносвязного слоя):

- Нет соединения каждого узла одного слоя со всеми узлами следующего.
- Веса становятся *разделяемыми*.

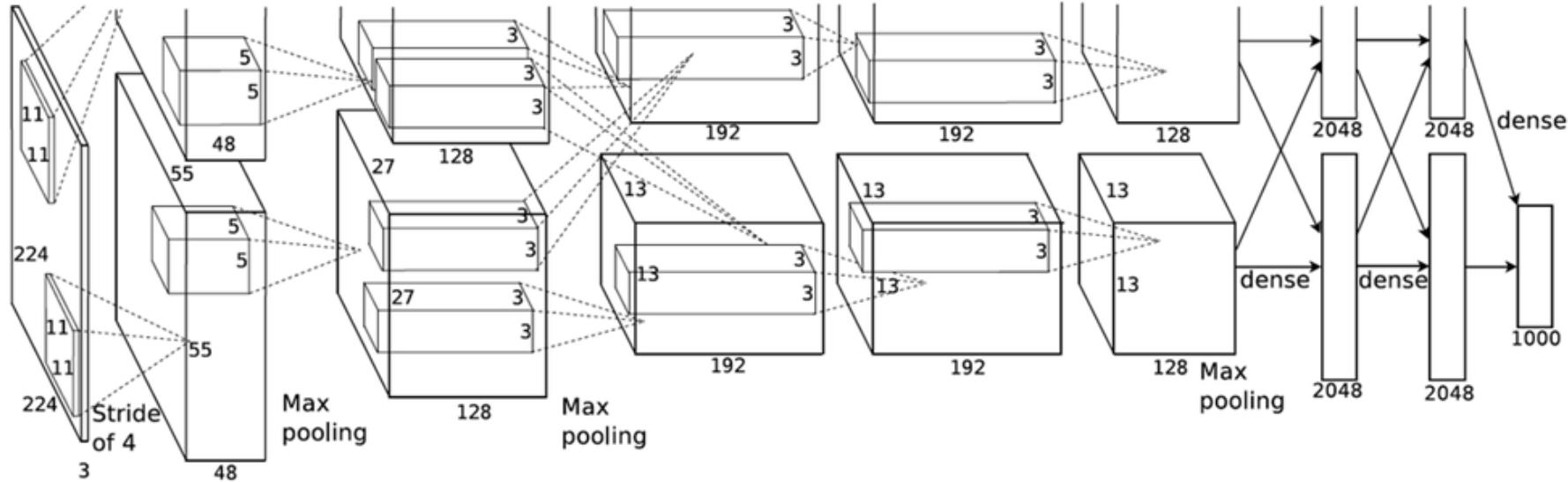
LeNet-5 [Le Cun et al., 1998]



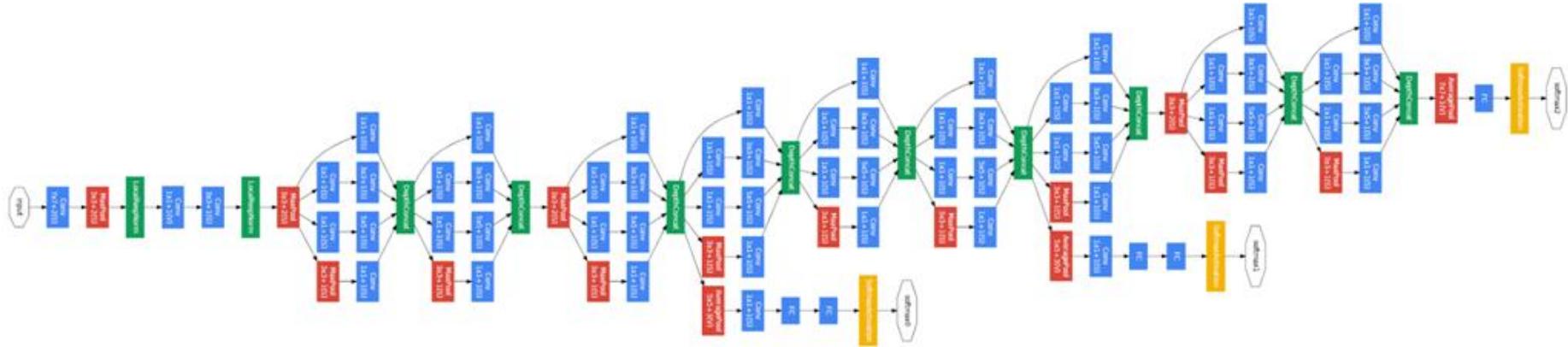
- Сверточные слои (convolutional layers)
- «Выборочные» слои, или слои объединения (subsampling/pooling layers)
- Полносвязные слои (fully connected layers)
- Регуляризация (weight decay, dropout, normalization)



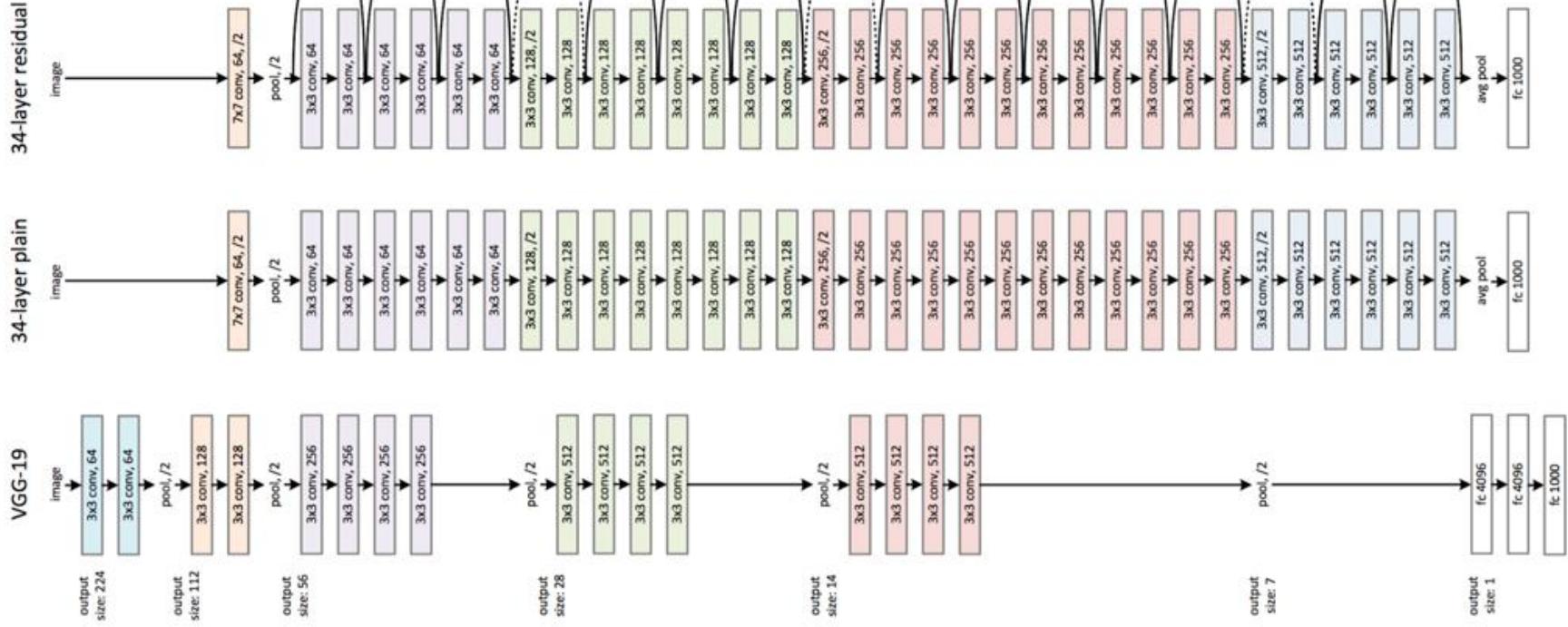
Alexnet (2012)



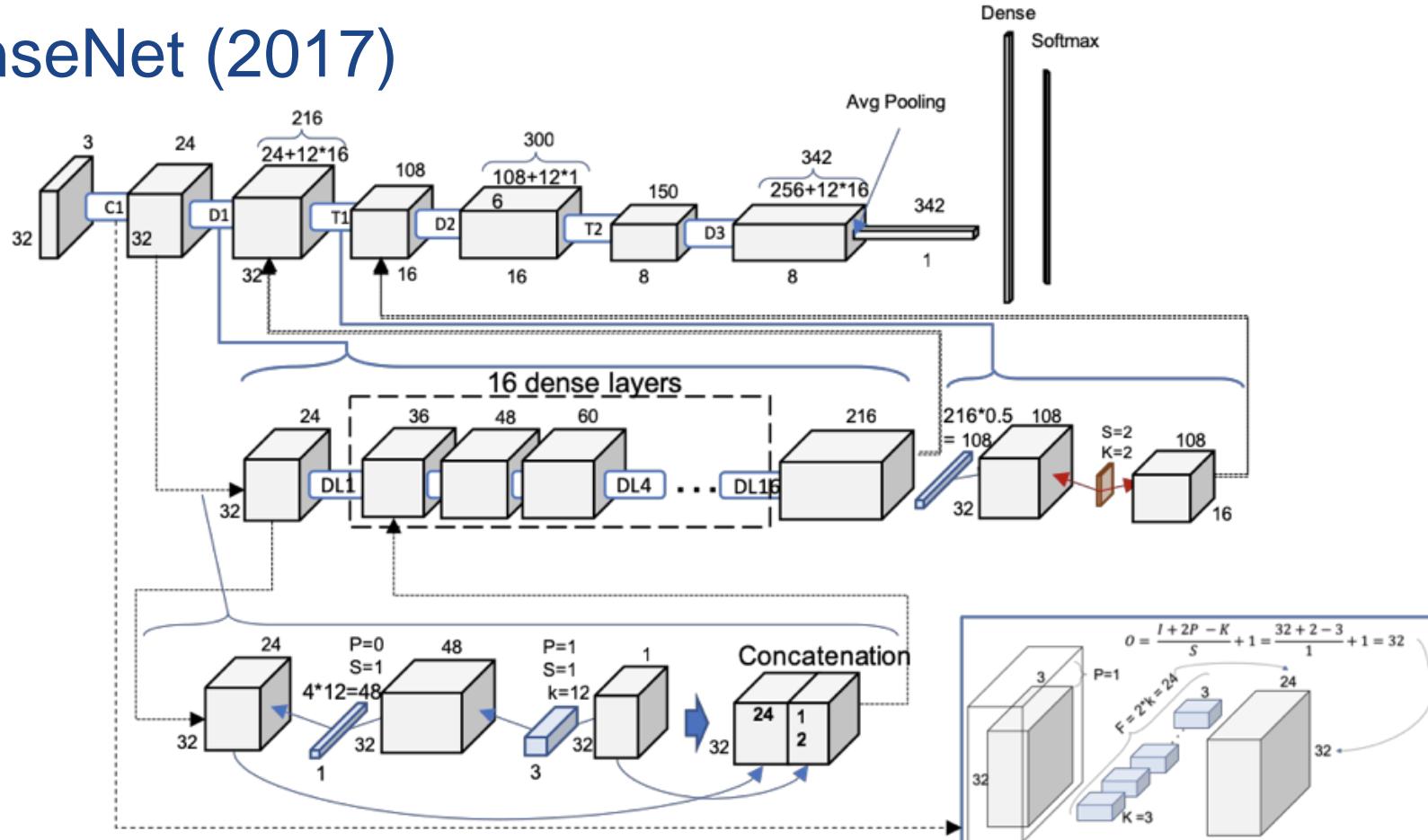
GoogLeNet (2015)



Microsoft Resnet (2015)



DenseNet (2017)



Спасибо за внимание!

Некоторые задачи машинного обучения

- Обучение с учителем:
 - классификация
 - регрессия
 - предсказание временного ряда
- **Обучение без учителя:** *у не известен для обучающей выборки*
 - кластеризация
 - визуализация
 - понижение размерности
 - ...
- Обучение с подкреплением
- ...

Обучение без учителя

При обучении без учителя «обучающая выборка» состоит только из объектов:

$$x^{(1)}, x^{(2)}, \dots, x^{(N)}$$

Необходимо определить, как эти объекты связаны друг с другом.

Например,

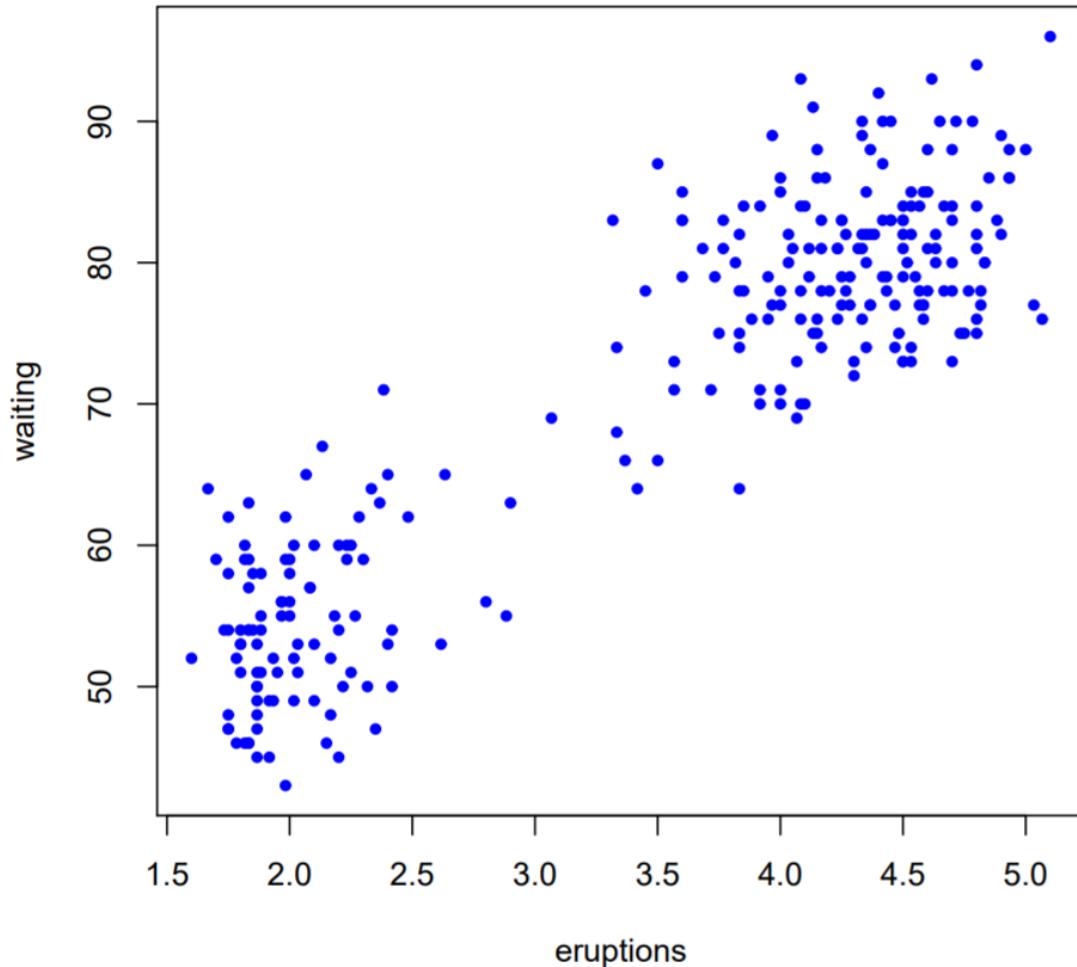
Задача кластеризации — разбить объекты на группы (кластеры), так, чтобы в одном кластере оказались «похожие» друг на друга объекты, а в разных кластерах — «мало похожие».

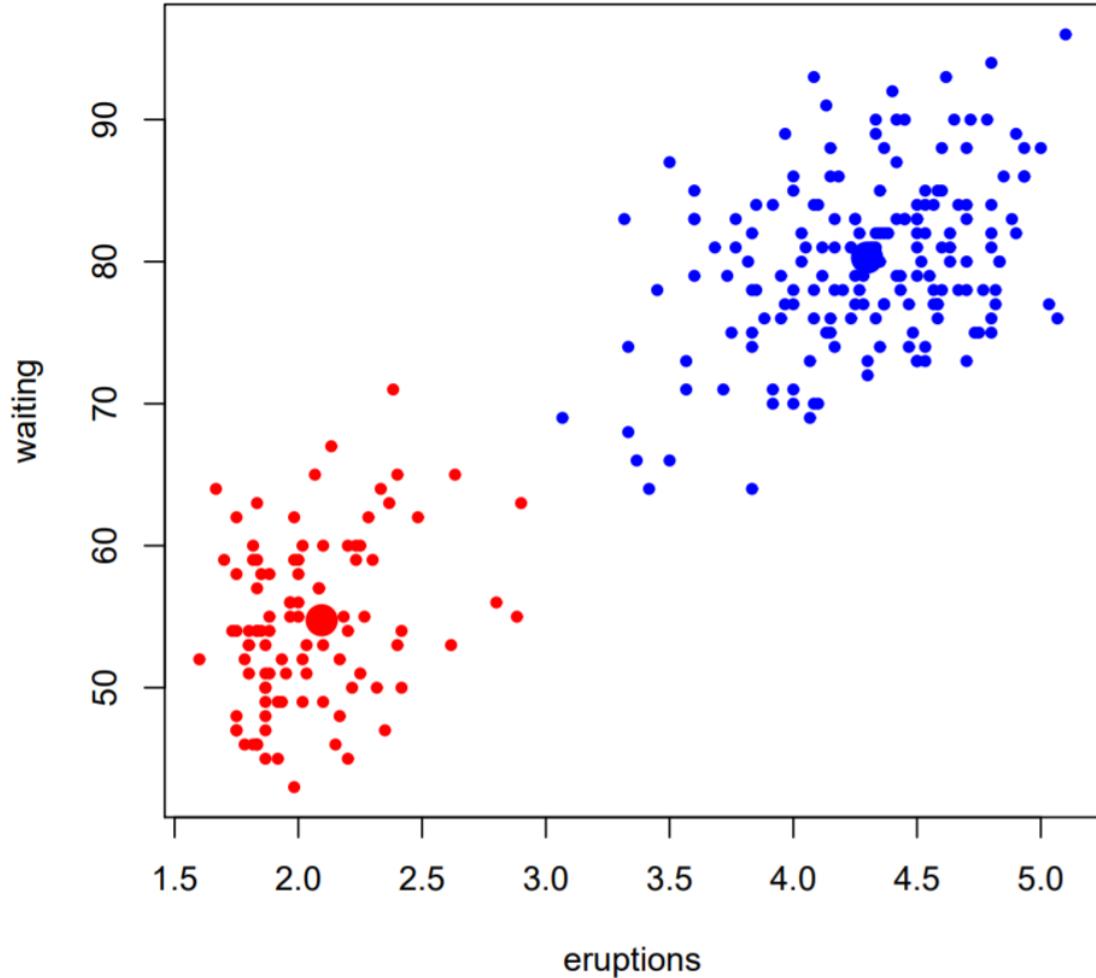
Пример 3. «Старый служака»

Гейзер «Старый служака» (Old Faithful) Йеллоустонский национальный парк, Вайоминг, США

[A. Azzalini, A.W. Bowman A look at some data on the Old Faithful geyser // Applied Statistics. 1990, 39. P. 357–365]







Пример 4. Генезис языков

Список Сводеша (Morris Swadesh, 1909–1967) — список из слов базового словаря — ядра языка (термины родства, части тела, частые природные явления, животные и т.д.):

мать, отец, человек, рука, нога, дождь, ветер, солнце, луна, ...

Это самая старая лексика, она менее всего подвержена изменениям и заимствованиям. Эти слова тоже могут заменяться другими, но с меньшей вероятностью.

Примеры:

око → глаз, уста → рот, гад → змея, дитя → ребёнок,
пёс → собака, плод → овощ, чьрвленый → красный

Есть редакции из 100, 200, 207 понятий.

№	Русский	Английский	Немецкий	Итальянский	Французский	Чешский
1	я	I	ich	io	je	já
2	ты	you	du	tu	tu	ty
3	он	he	er	lui	il	on
4	мы	we	wir	noi	nous	my
5	вы	you	ihr	voi	vous	vy
6	они	they	sie	loro	ils	oni
7	этот	this	dieses	questo	ceci	tento
8	тот	that	jenes	quello	cela	tamten
9	здесь	here	hier	qui	ici	zde
10	там	there	dort	lá	lá	tam
11	кто	who	wer	chi	qui	kdo
12	что	what	was	che	quoi	co
13	где	where	wo	dove	où	kde
14	когда	when	wann	quando	quand	kdy
15	как	how	wie	come	comment	jak
16	не	not	nicht	non	ne... pas	ne
.....						
205	если	if	wenn	se	si	jestlize
206	потому что	because	weil	perché	parce que	protoze
207	имя	name	Name	nome	nom	jméno

Близость двух языков можно измерять по количеству родственных слов (когнат) – однокоренных слов, имеющих общее происхождение и близкое звучание.

Являются ли два слова родственными – определяют лингвисты:

родственны ли слова

год и рік, цветок и квітка, видеть и бачити

или

колесо и wheel и $\overline{\text{ч}}\overline{\text{ф}}$?

Можно разбивать языки на группы близких друг другу языков – задача кластерного анализа.

Более того, на основе анализа списка Сводеша для двух родственных языков можно приблизительно установить время их появления из единого пра-языка.

Считается, что в 100-словном списке сохраняется за тысячелетие около 86% слов, а в 200-словном в среднем 81% слов, соответственно. Отсюда «период полураспада» языкового «ядра» – для 100- и 200-словного списка равен соответственно 4.6 и 3.3 тыс. лет. (это один из методов глоттохронологии)

