Машинное обучение: классические методы и открытые вопросы

Евгений Соколов

Факультет компьютерных наук НИУ ВШЭ

О чём вообще это всё?

Правиловый машинный перевод

- Как сделать сервис для перевода, если надо прямо сейчас?

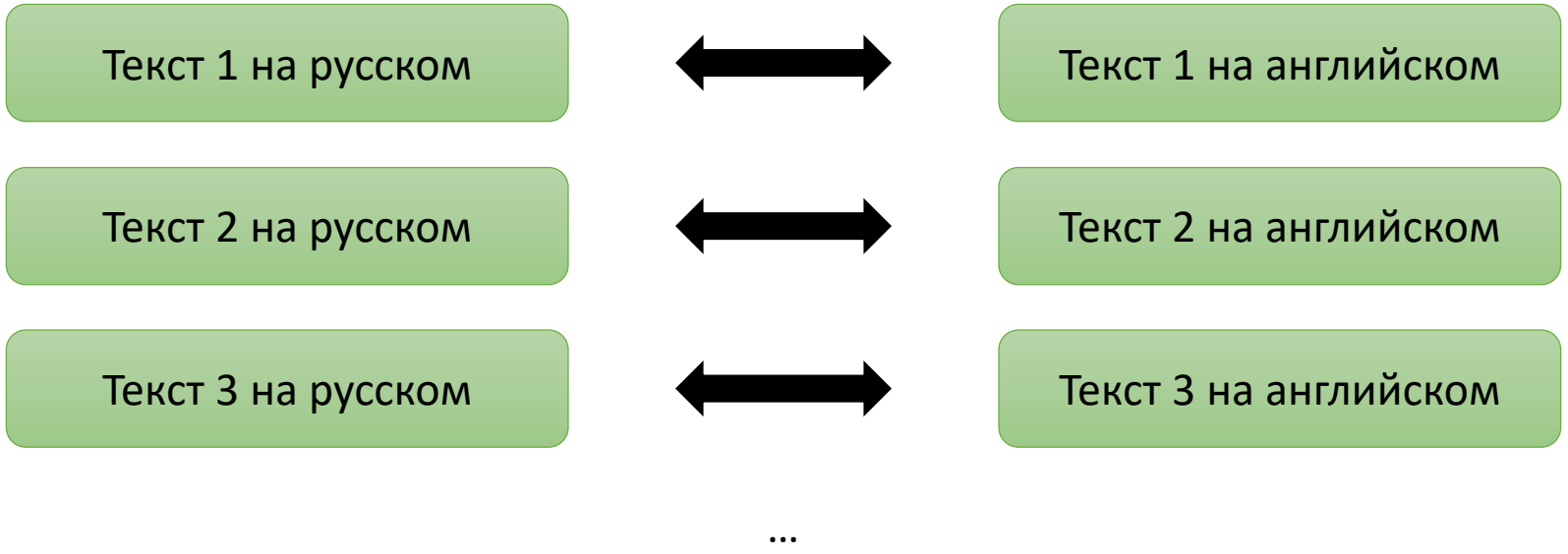
Правиловый машинный перевод

- Как сделать сервис для перевода, если надо прямо сейчас?

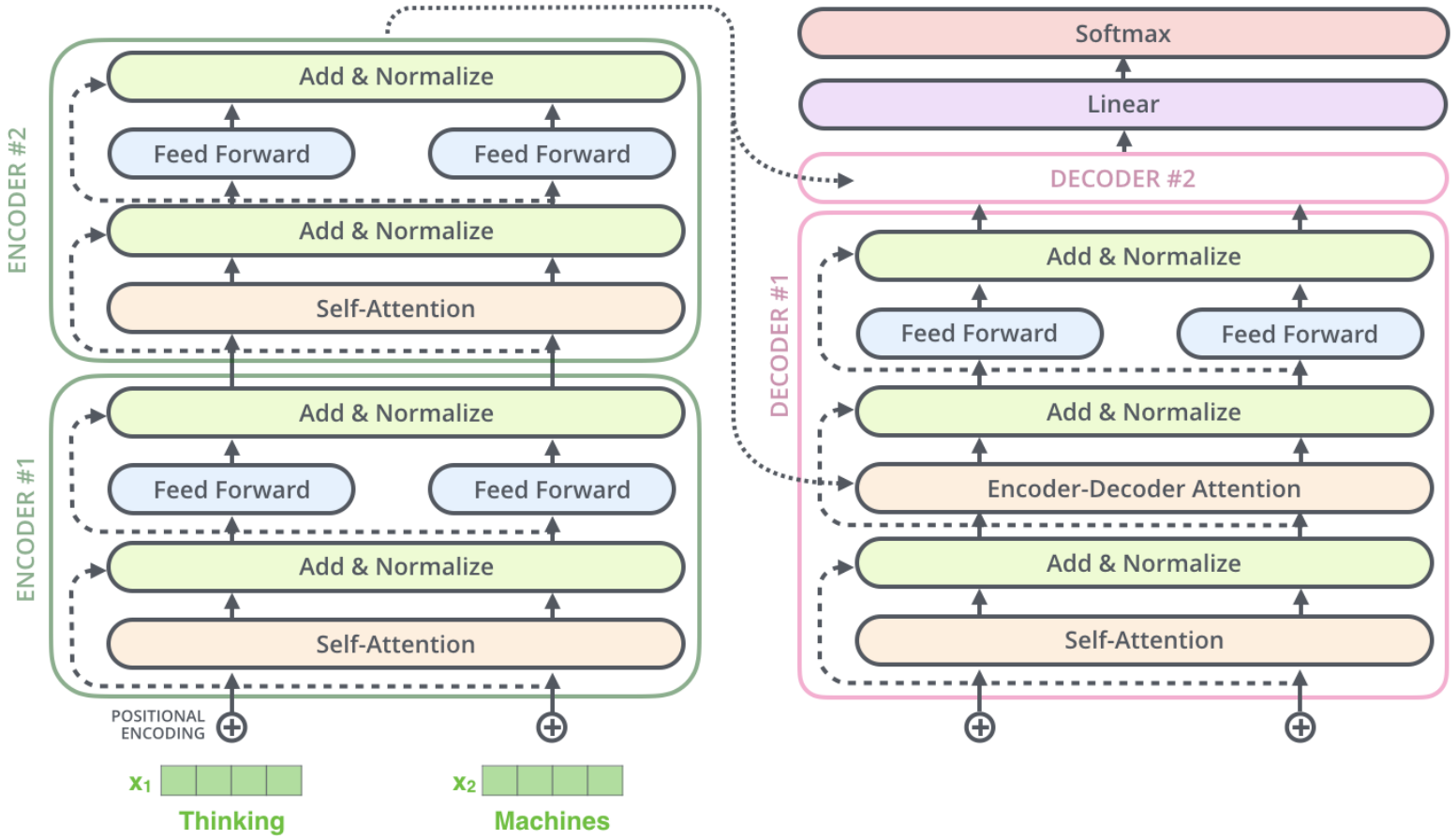
- Перевод по словам
- Грамматические правила

Правиловый машинный перевод

Идея: идти от данных, а не от правил

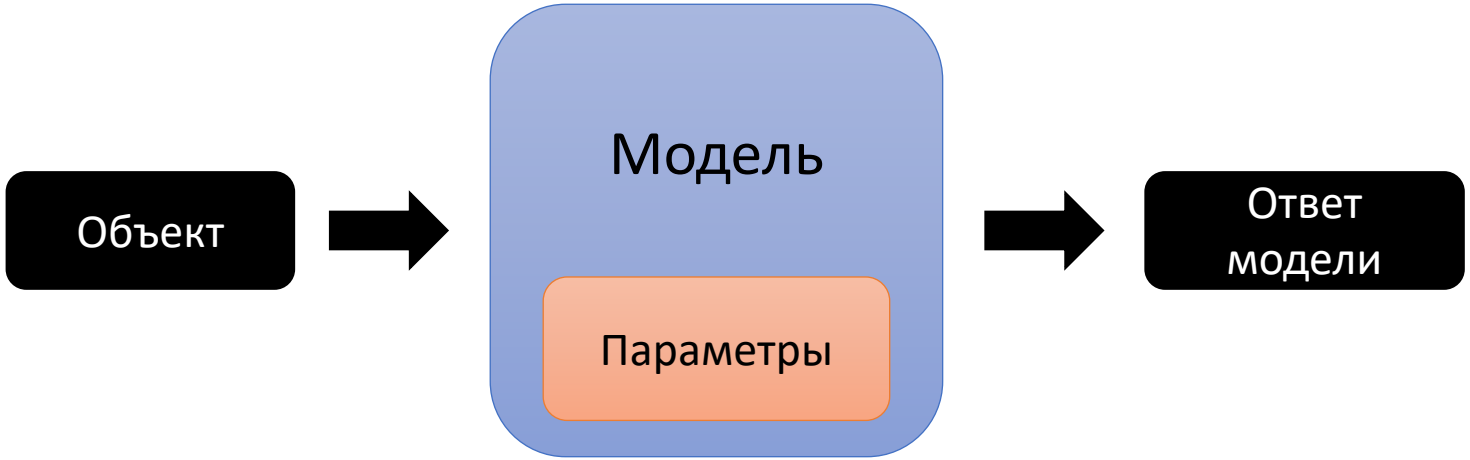


Нейросетевой машинный перевод



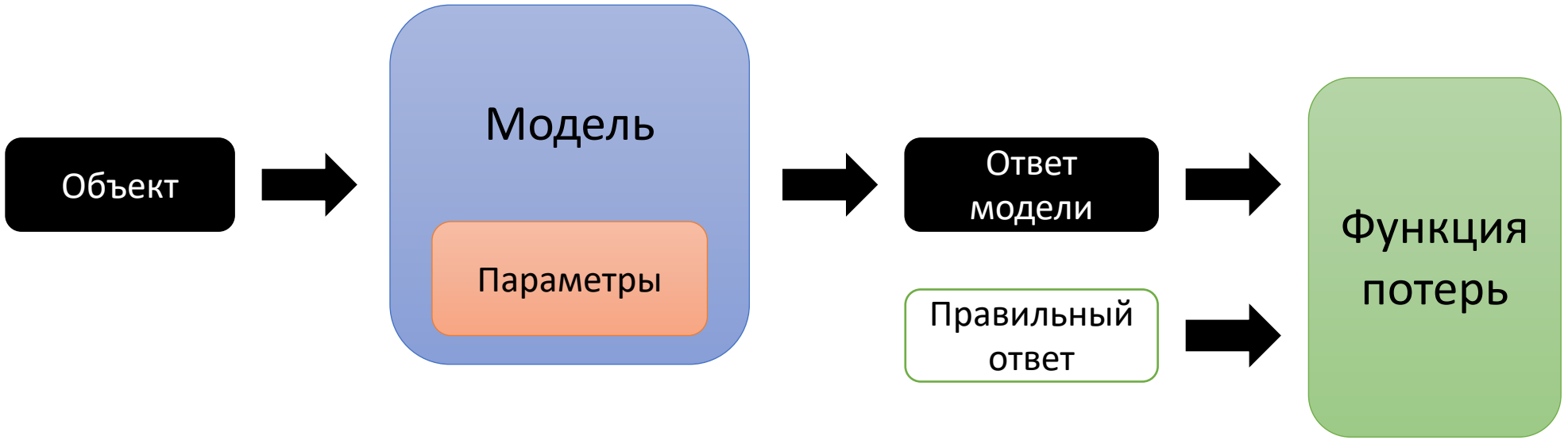
Теперь более строго

Общая схема

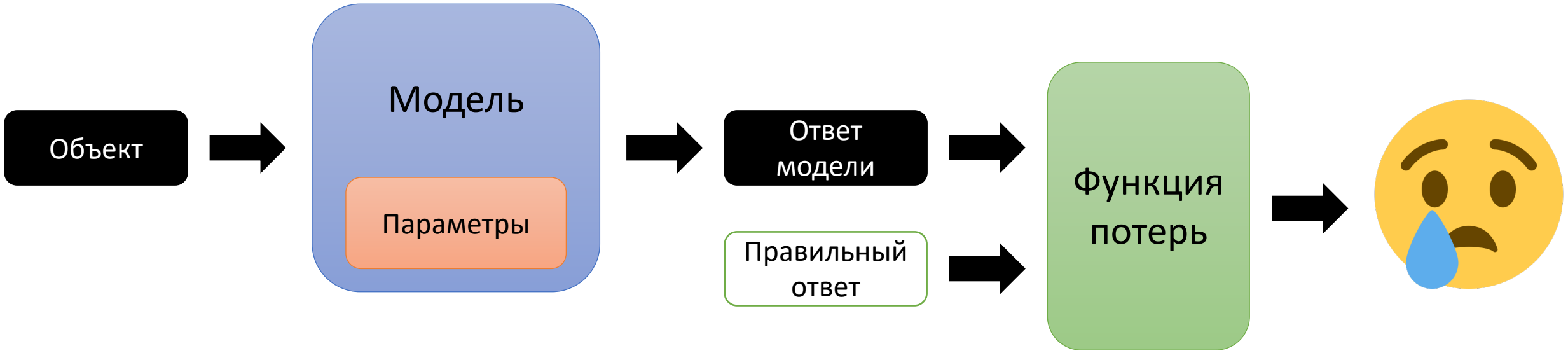


- Модель — формула или алгоритм для решения задачи
- Параметры — «ручки», влияющие на поведение модели

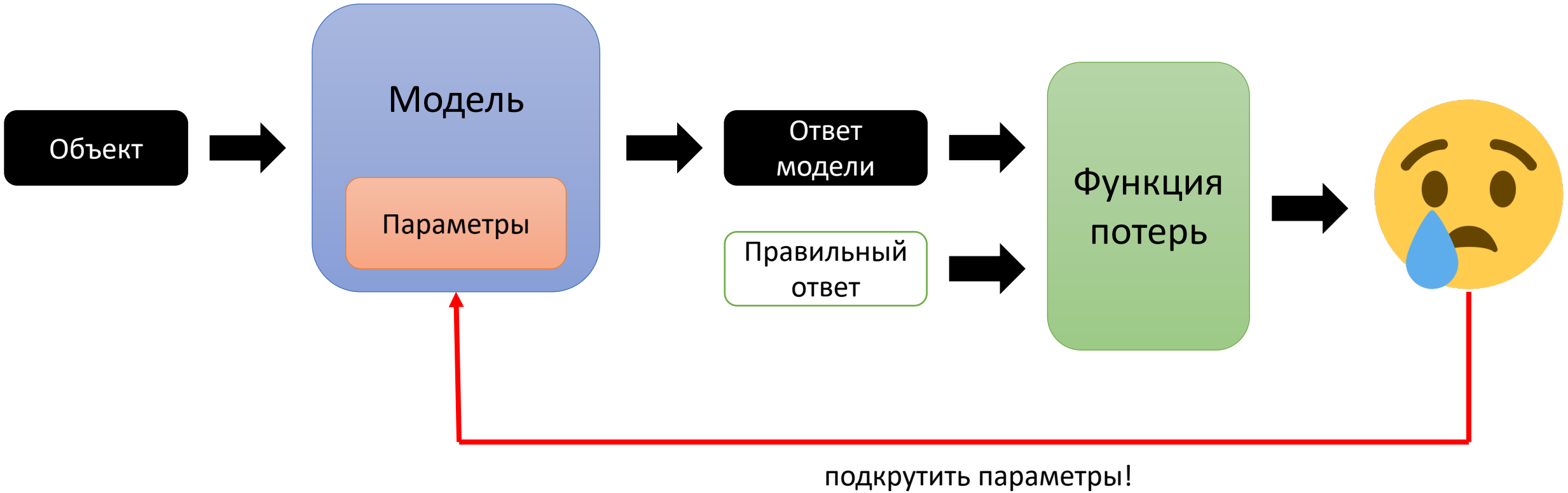
Общая схема


Объект
Модель
Параметры
Ответ модели
Правильный ответ
Функция потерь

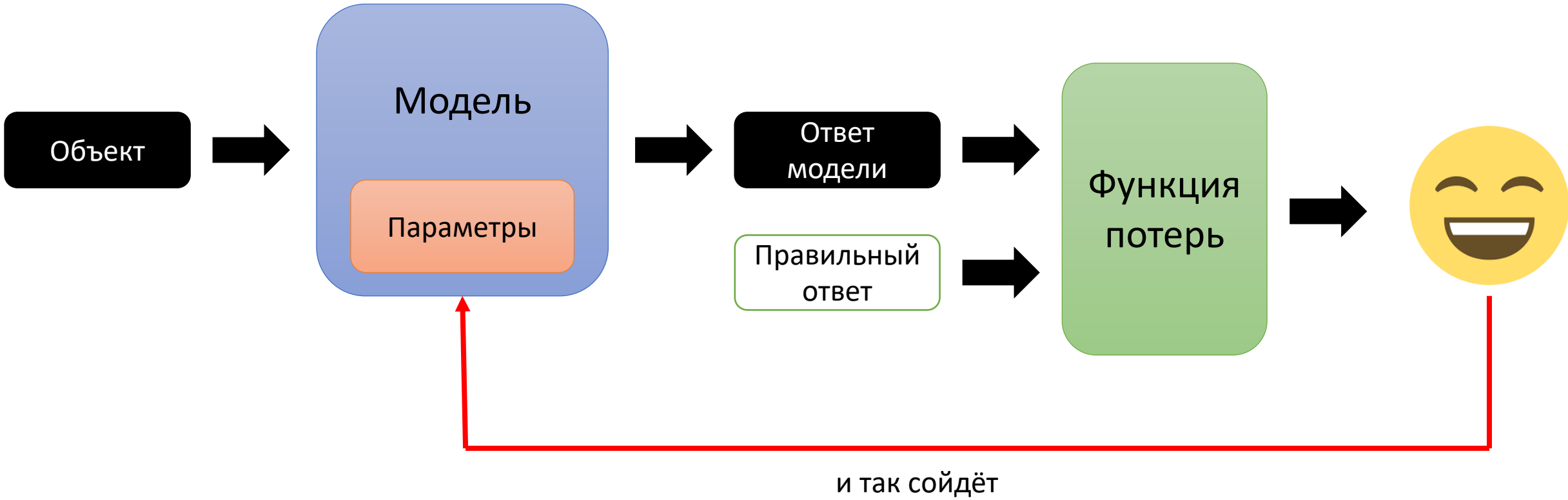
Общая схема


Объект
Модель
Параметры
Ответ модели
Правильный ответ
Функция потерь

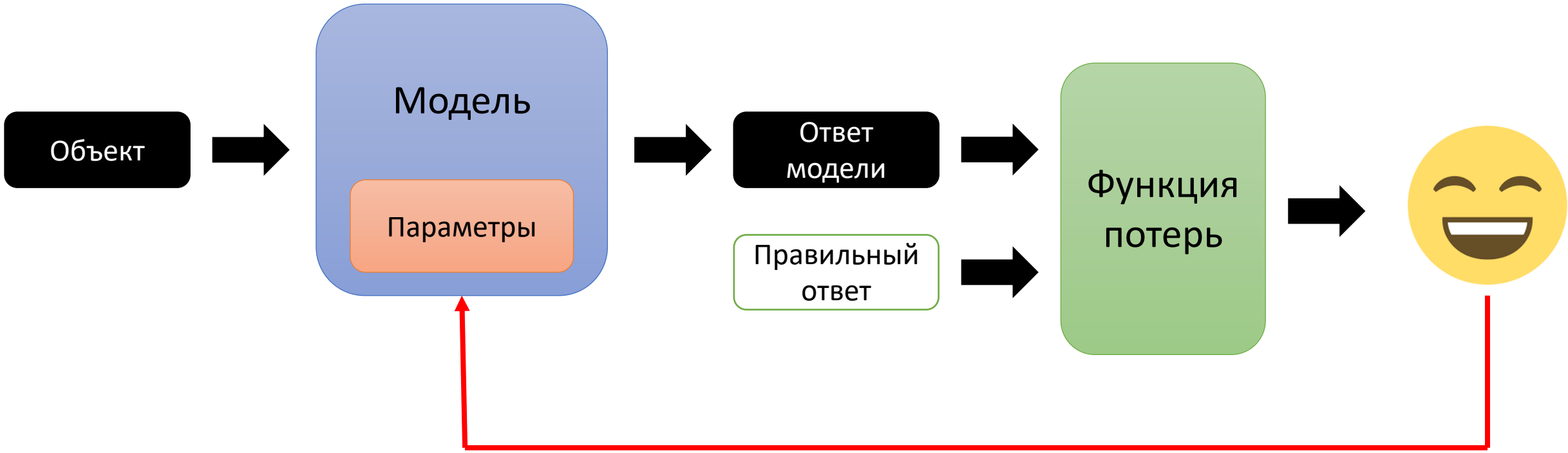
Общая схема


Объект
Модель
Параметры
Ответ модели
Правильный ответ
Функция потерь
подкрутить параметры!

Общая схема


Объект
Модель
Параметры
Ответ модели
Правильный ответ
Функция потерь
и так сойдёт

Общая схема

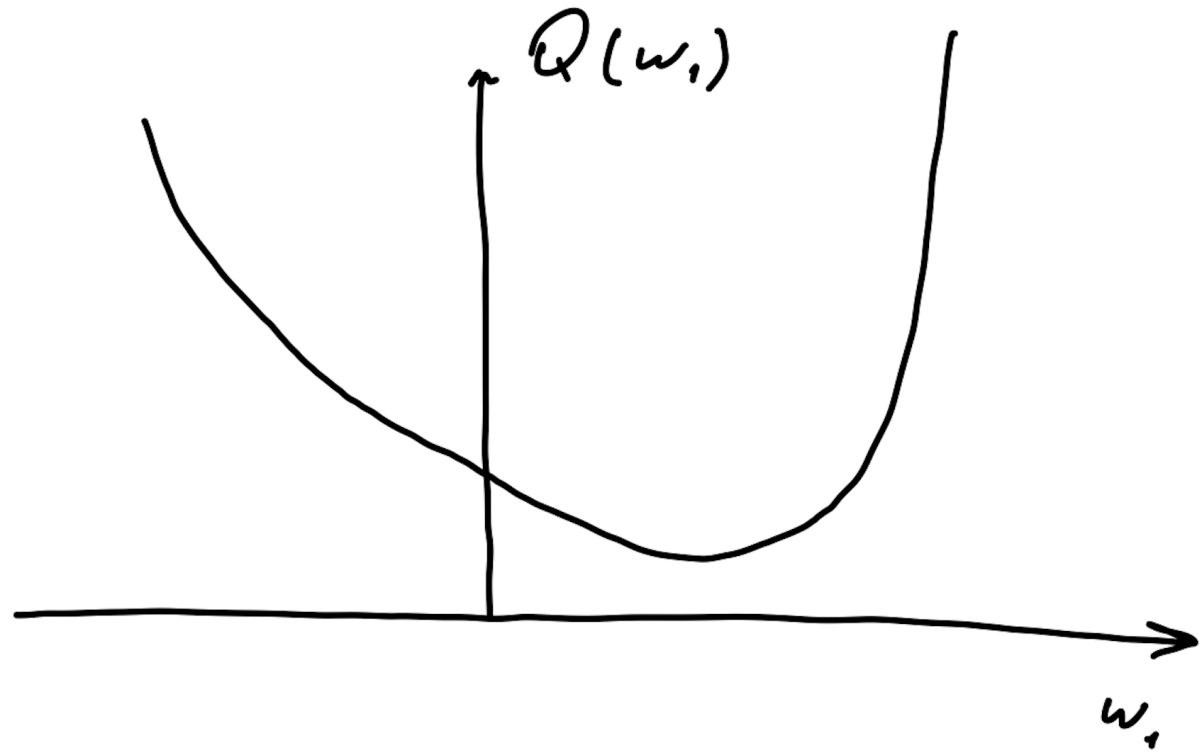


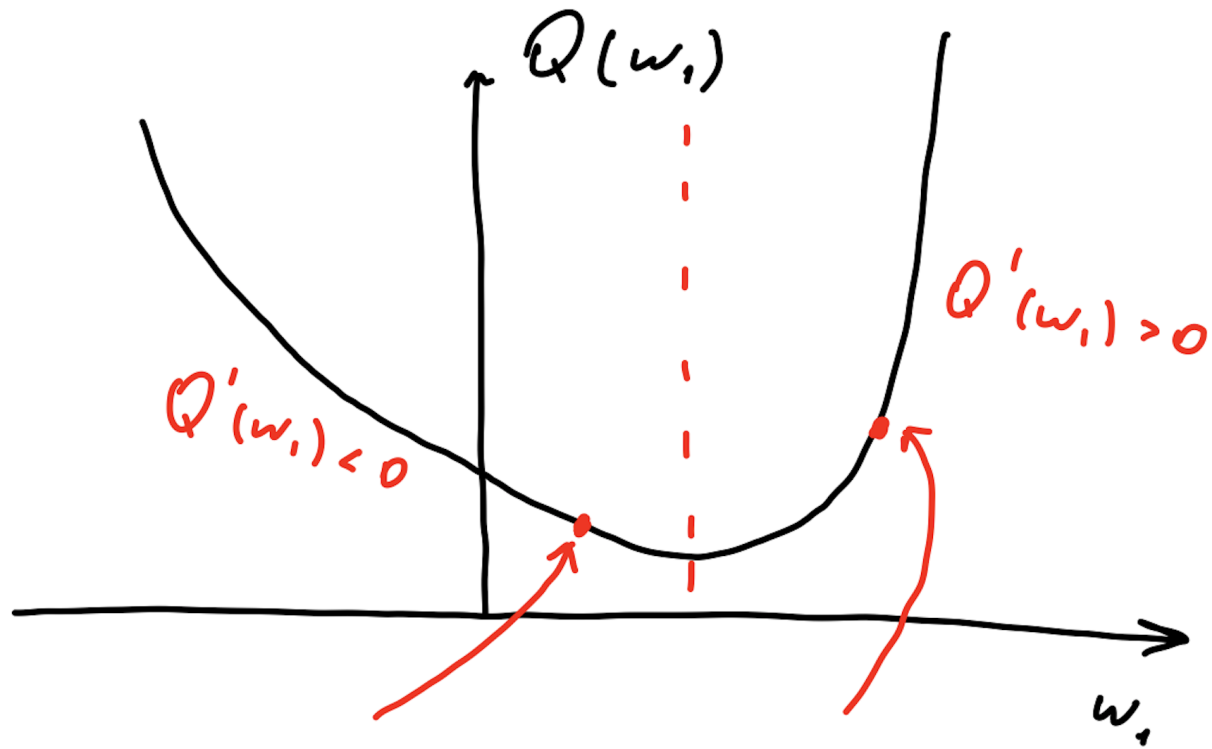
По очереди проделываем для всей выборки, пока ошибка уменьшается

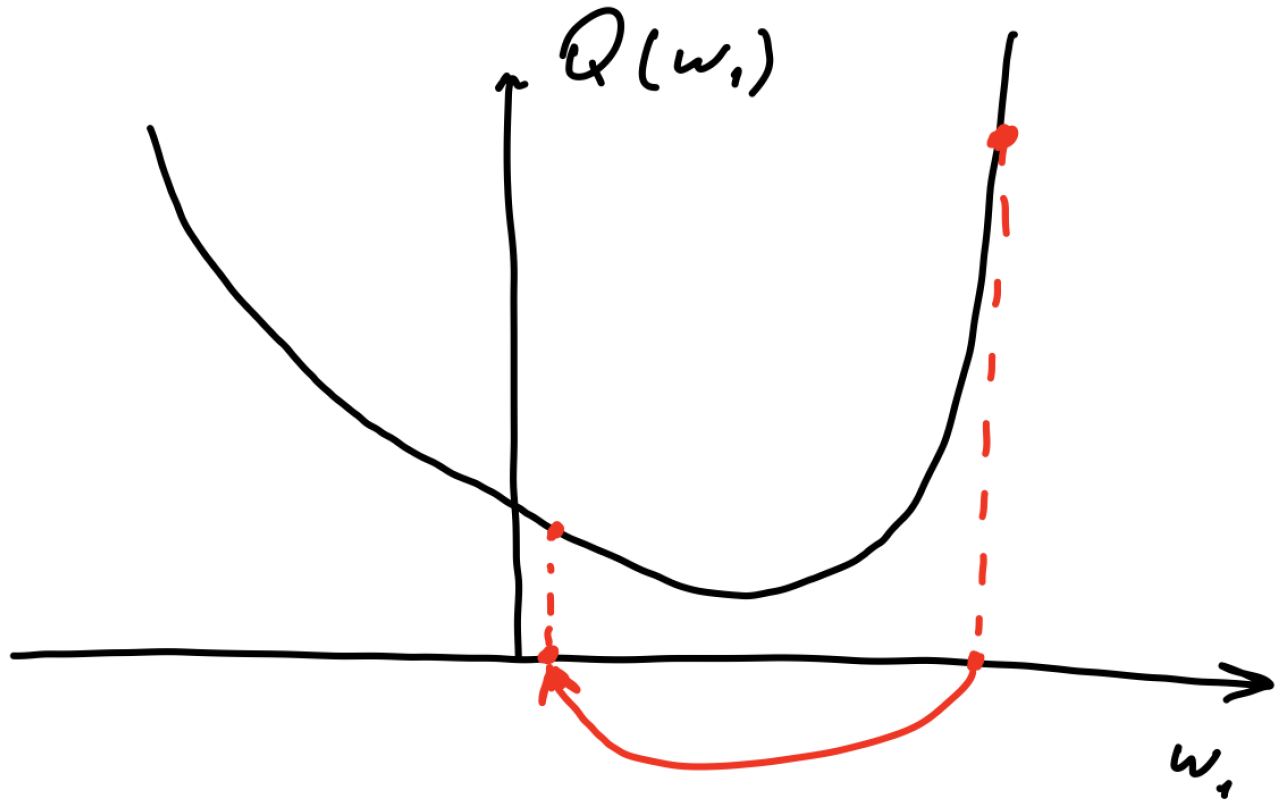
Более формально

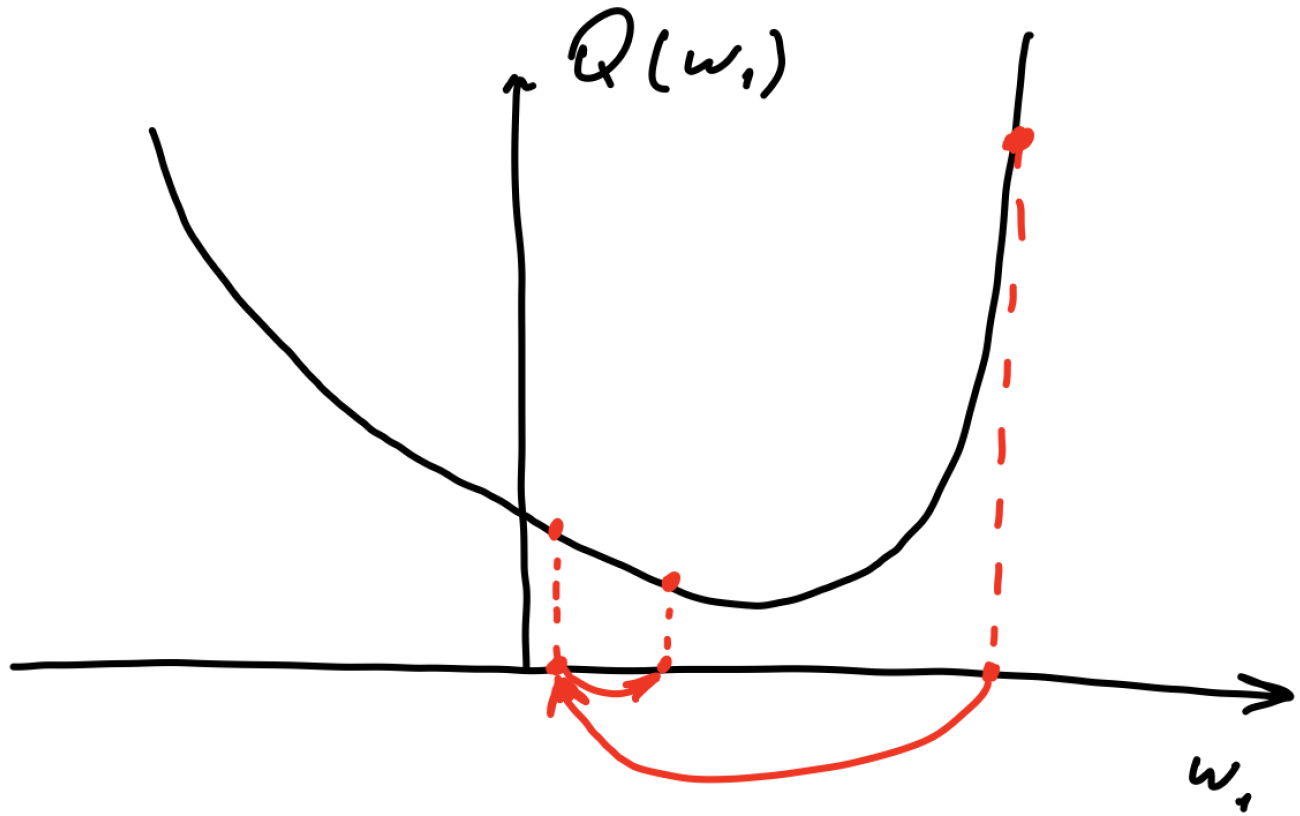
$$Q(w) = \sum_{i=1}^{\ell} L\big(y_i, a(x_i, w)\big) \to \min_w$$

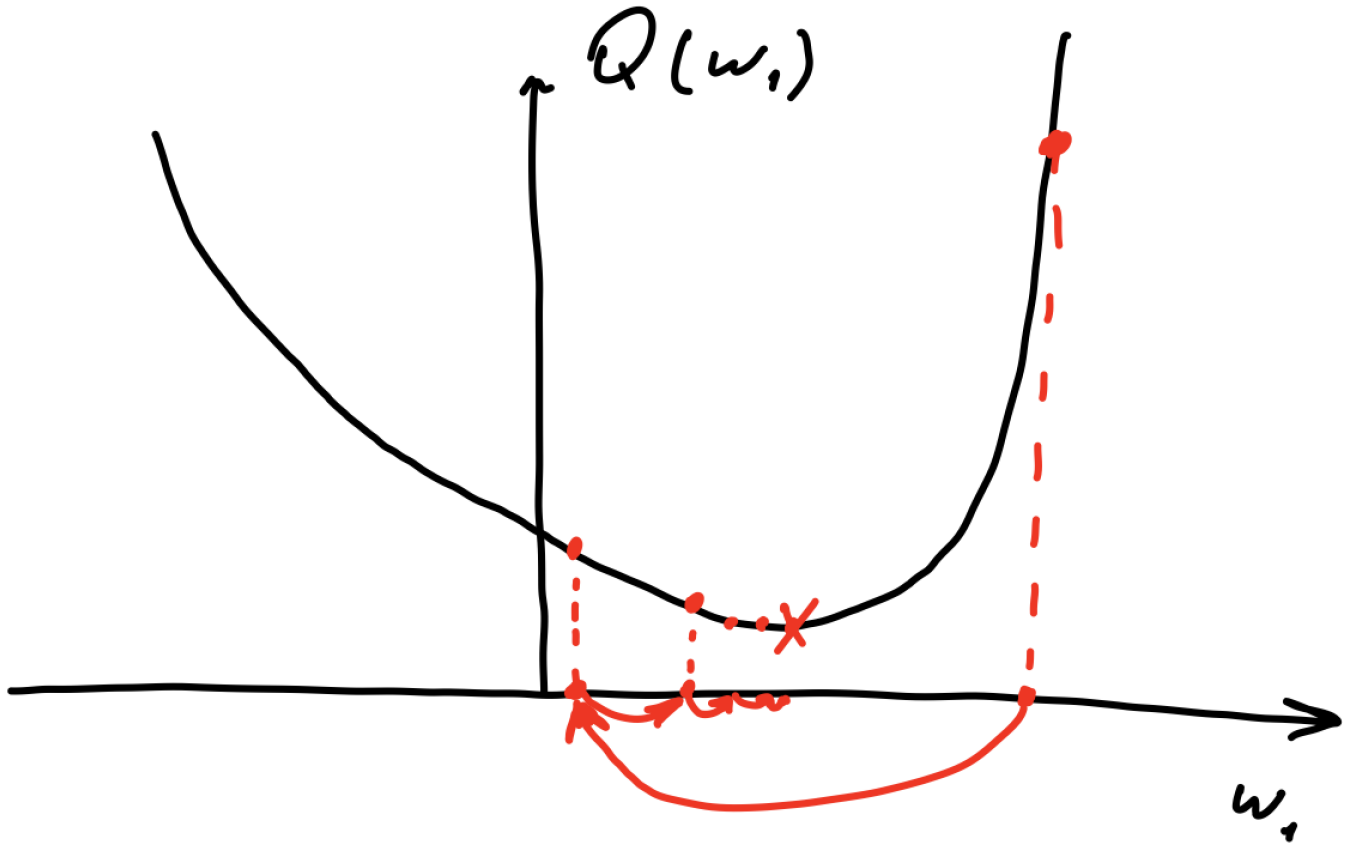
- x_i, y_i — объект и правильный ответ
- $a(x, w)$ — модель с параметрами w
- $L(y, z)$ — функция потерь

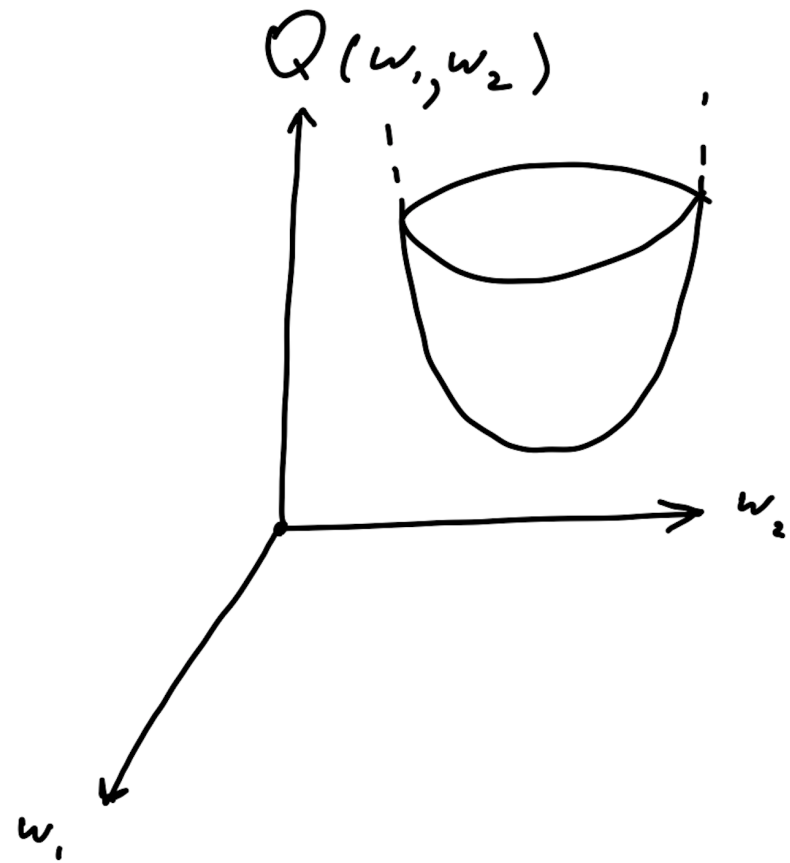

Q(w1)
w1

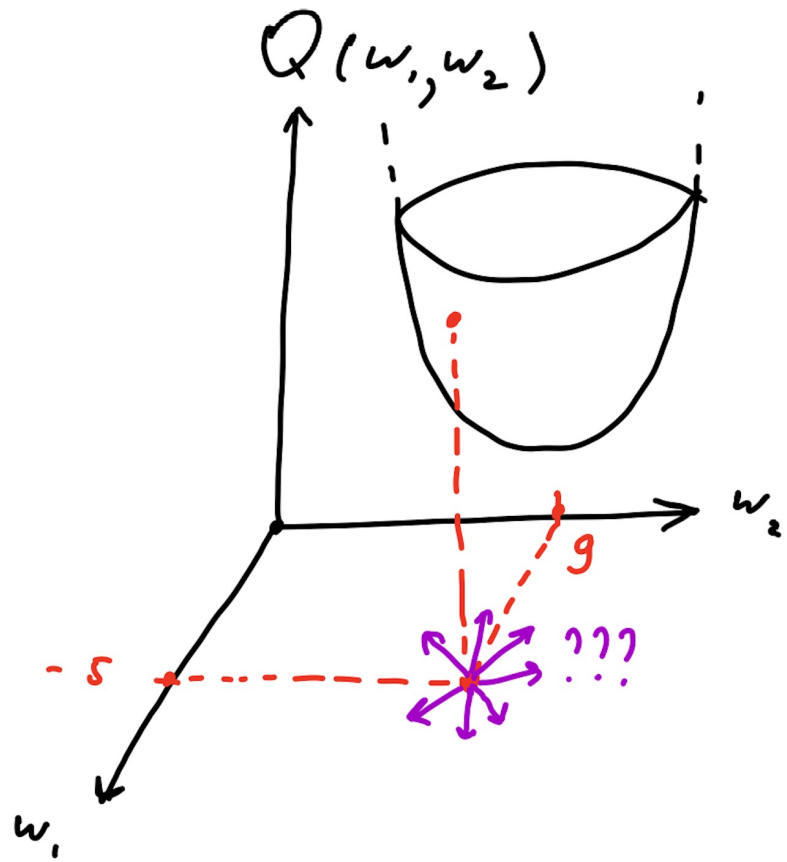

Q(w1)
Q'(w1) < 0
Q'(w1) > 0
w1

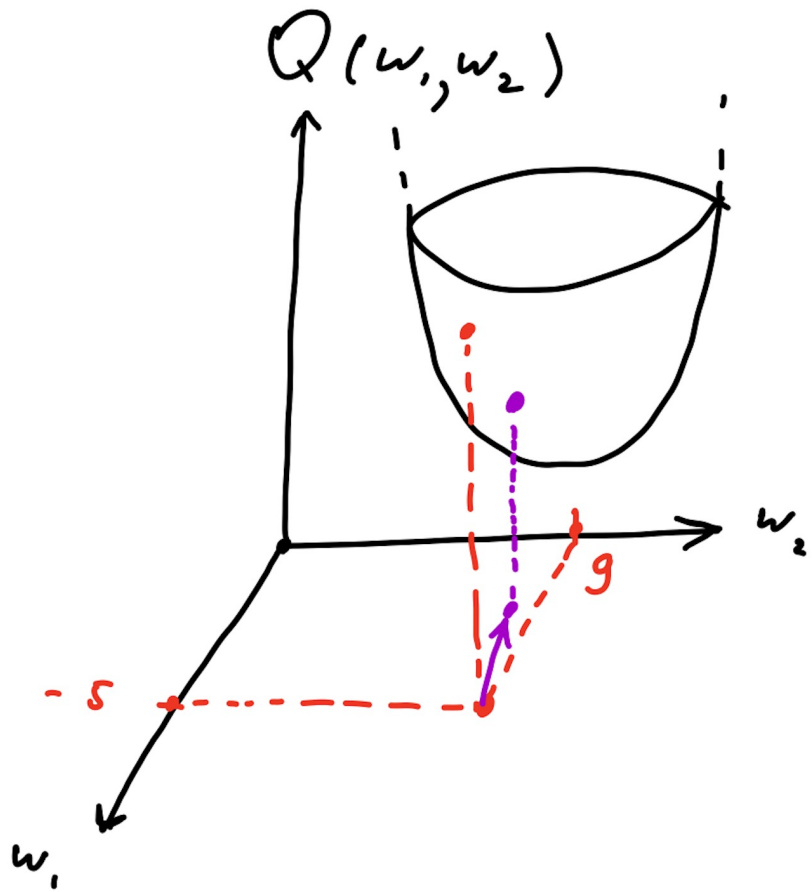

Q(w1)
w1

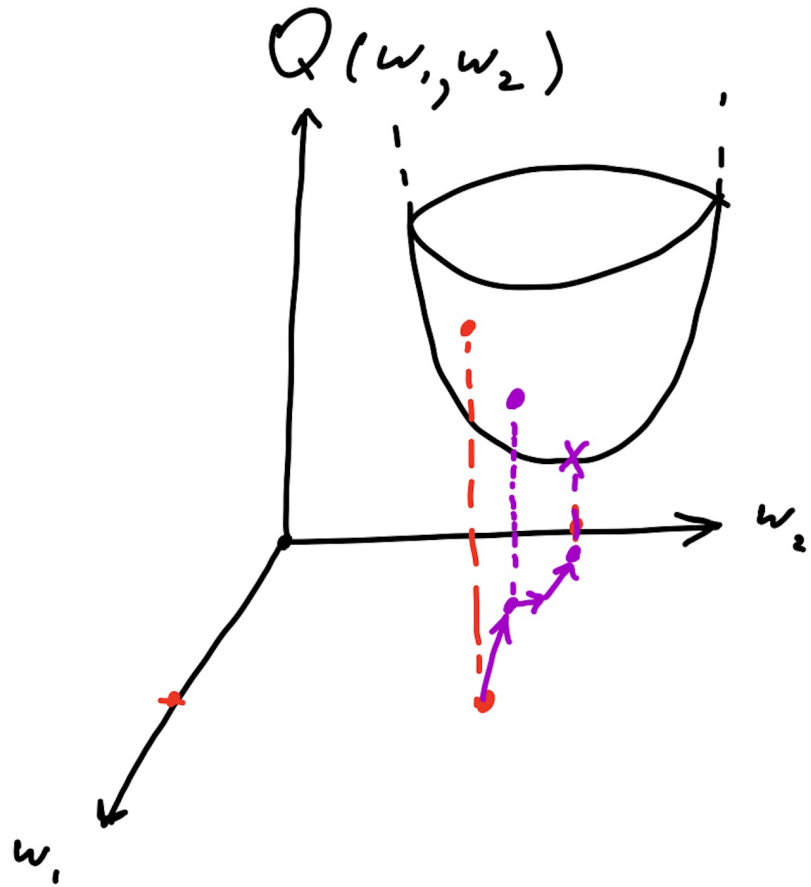

Q(w1)
w1


Q(w1)
w1


Q(w1, w2)
w2
w1


Q(w1, w2)
w2
g
- s
???
w1

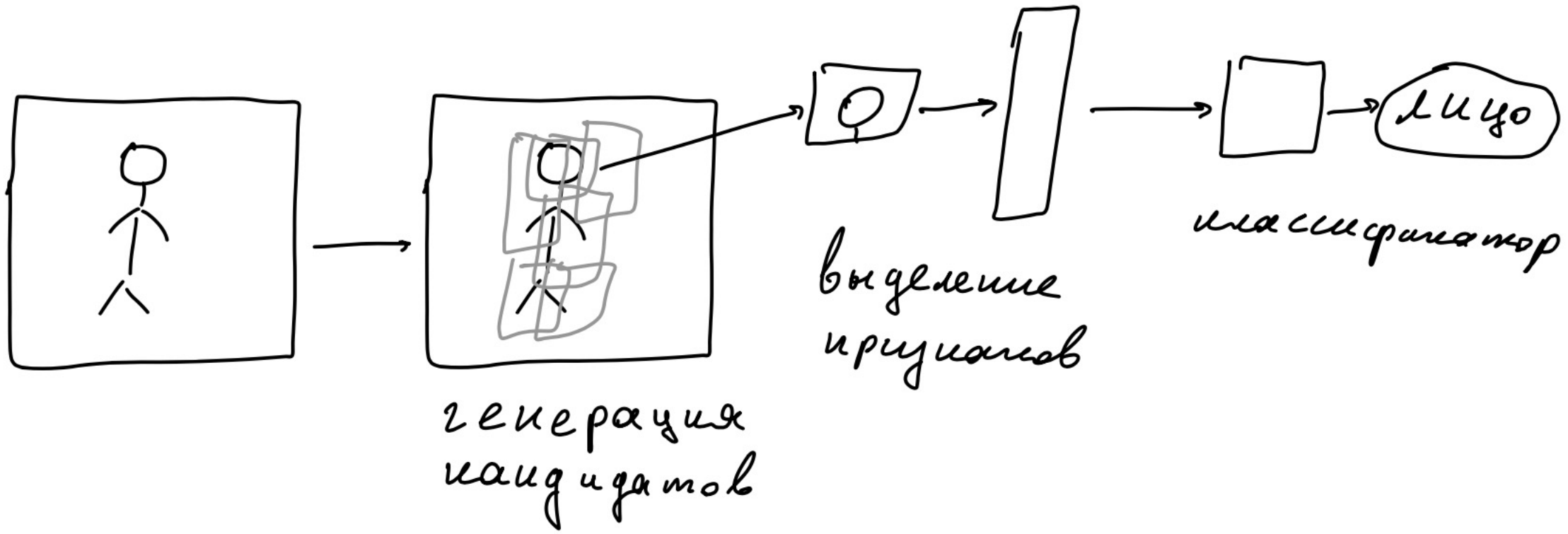

Q(w1, w2)
w2
g
- s
w1


Q(w1, w2)
w2
w1

Как работать с изображениями?

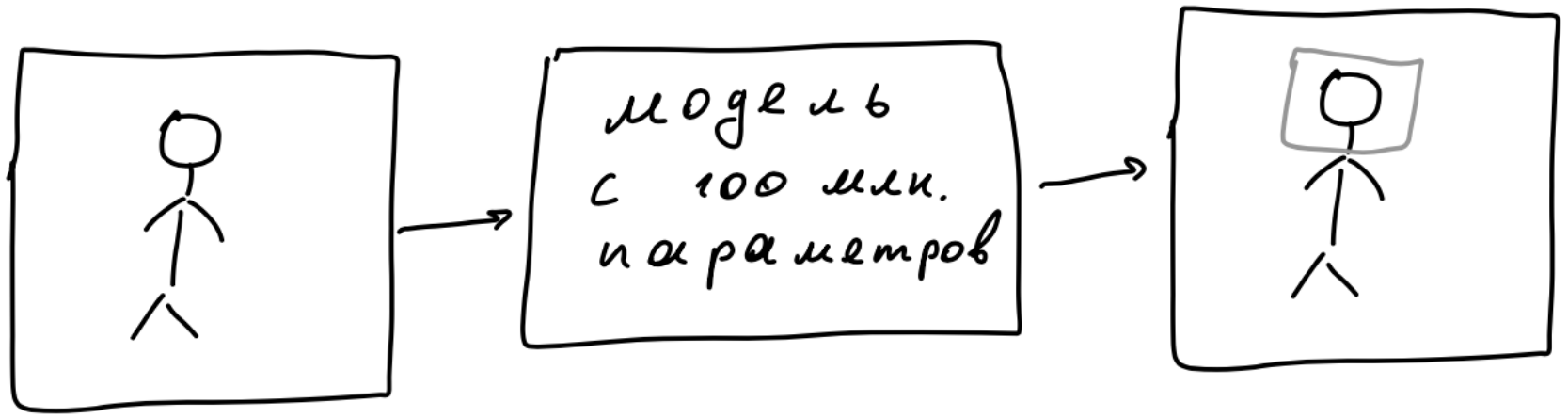
Принцип 1: end-to-end обучение

Детекция объектов раньше:



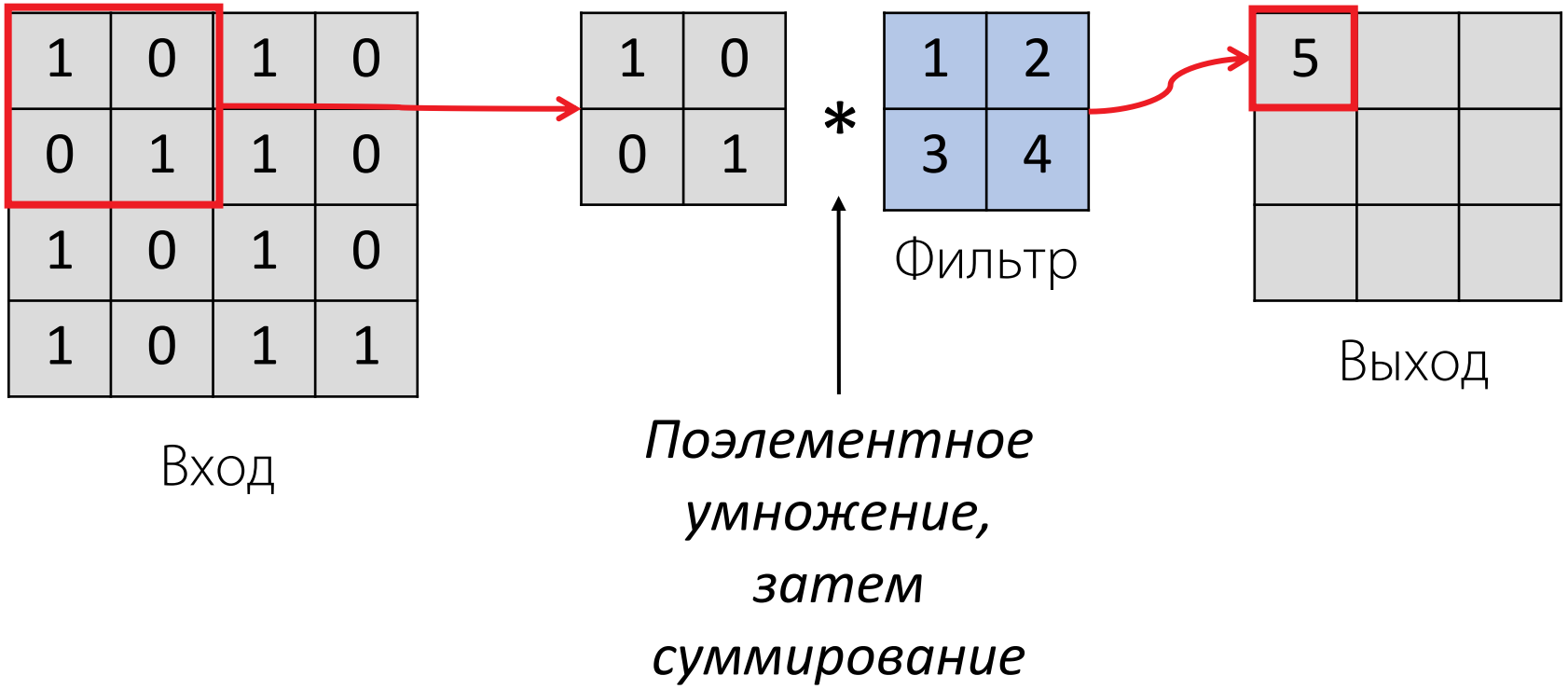
Принцип 1: end-to-end обучение

Детекция объектов сегодня:

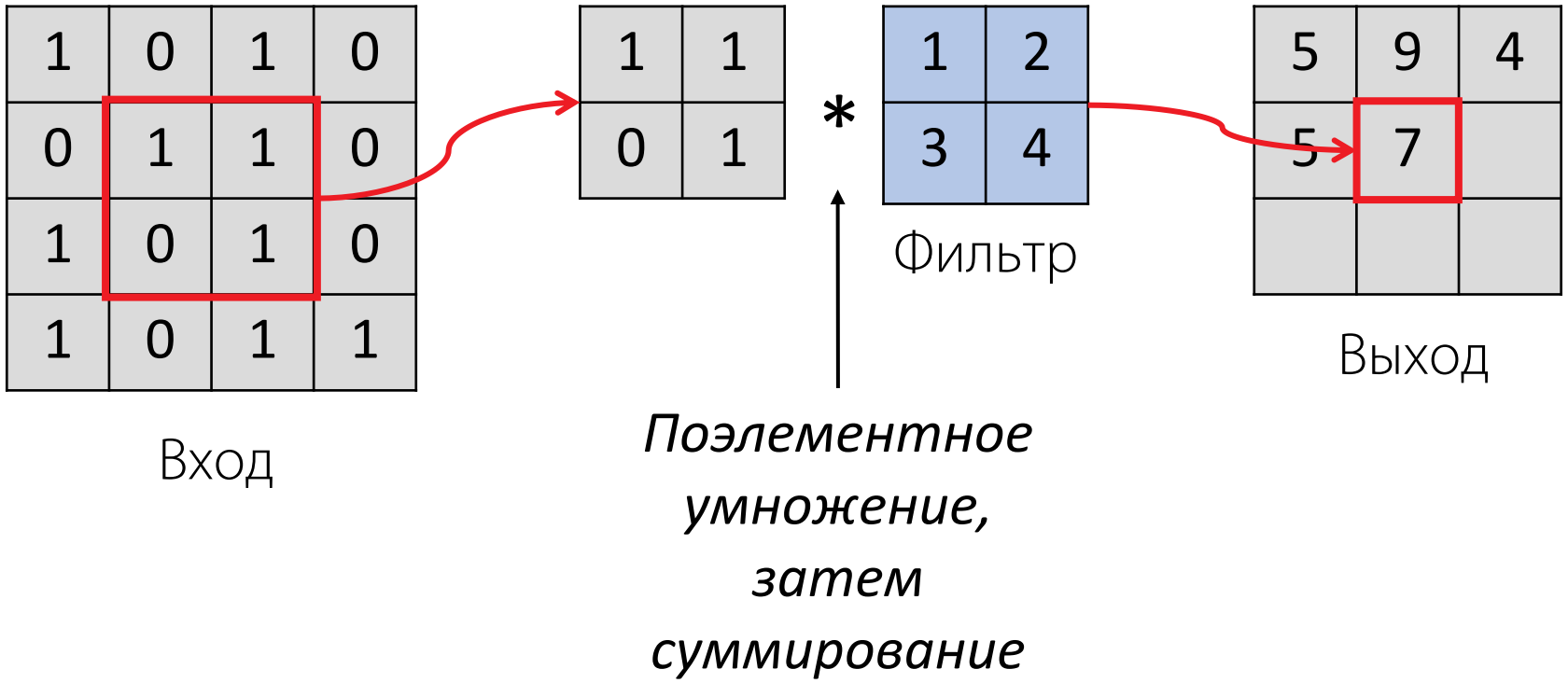


Принцип 2: многократное извлечение признаков

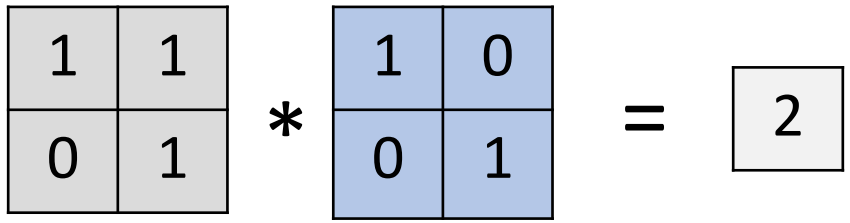
Свёртка

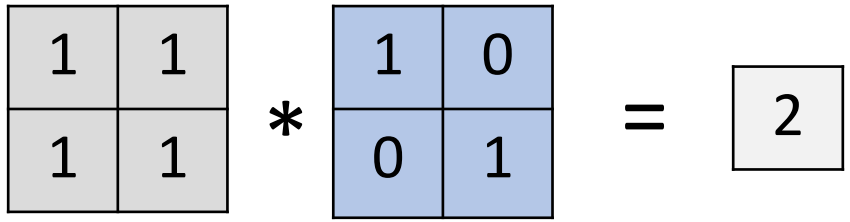

1 0 1 0
0 1 1 0
1 0 1 0
1 0 1 1
Вход
1 0
0 1
*
1 2
3 4
Фильтр
Поэлементное умножение, затем суммирование
5
Выход

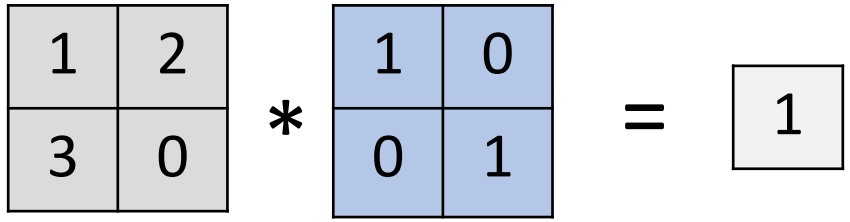
Свёртка

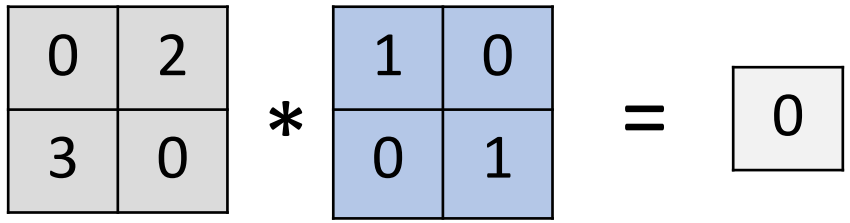

1 0 1 0
0 1 1 0
1 0 1 0
1 0 1 1
Вход
1 1
0 1
*
1 2
3 4
Фильтр
Поэлементное умножение, затем суммирование
5 9 4
5 7
Выход

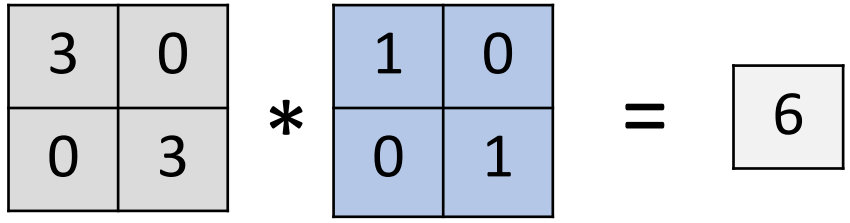
Свёртка

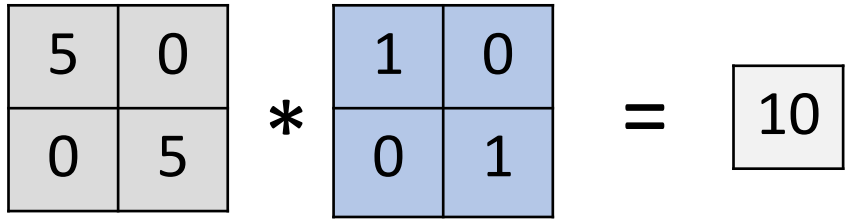

1 1
0 1
*
1 0
0 1
=
2


1 1
1 1
*
1 0
0 1
=
2

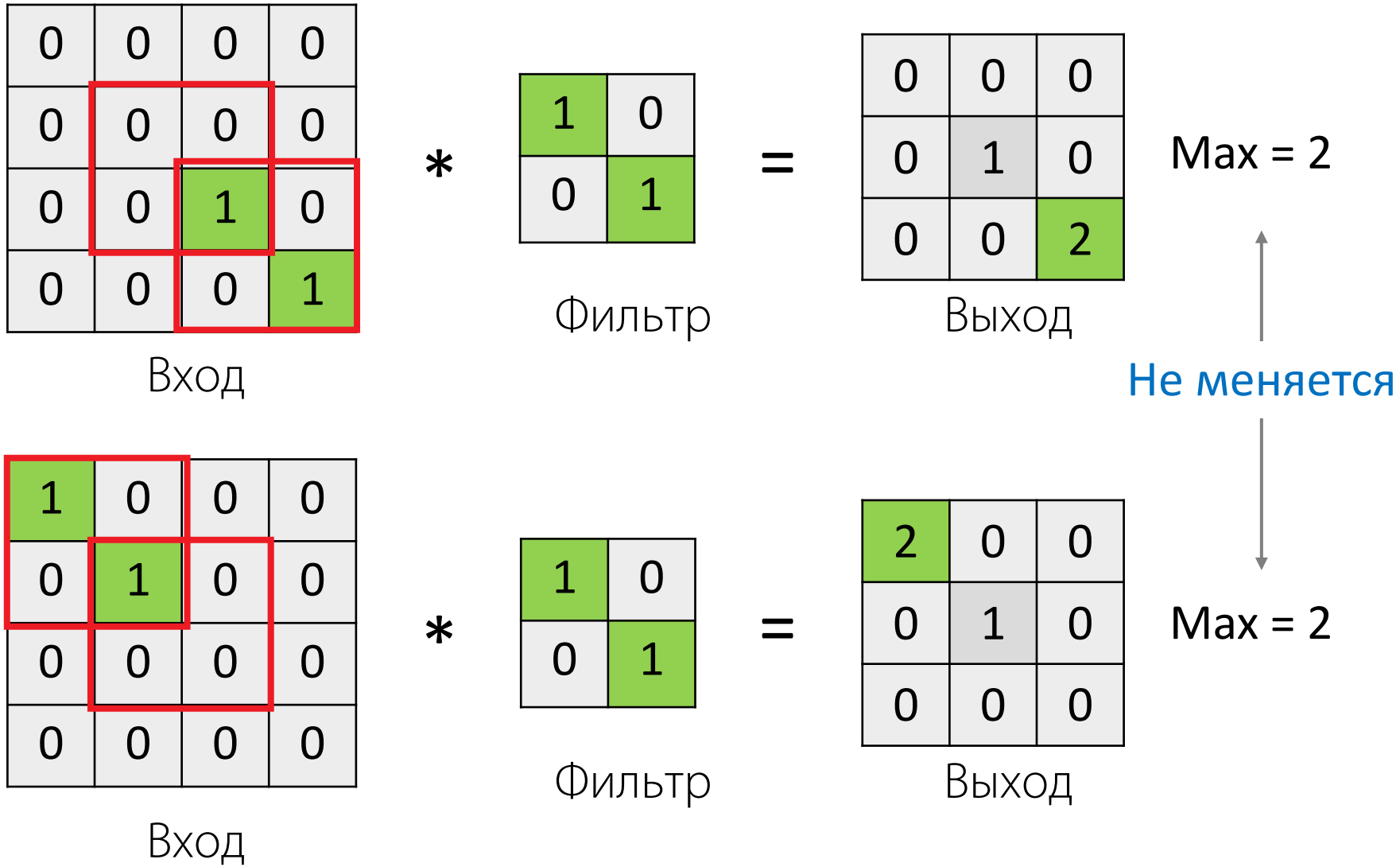

1 2
3 0
*
1 0
0 1
=
1


0 2
3 0
*
1 0
0 1
=
0

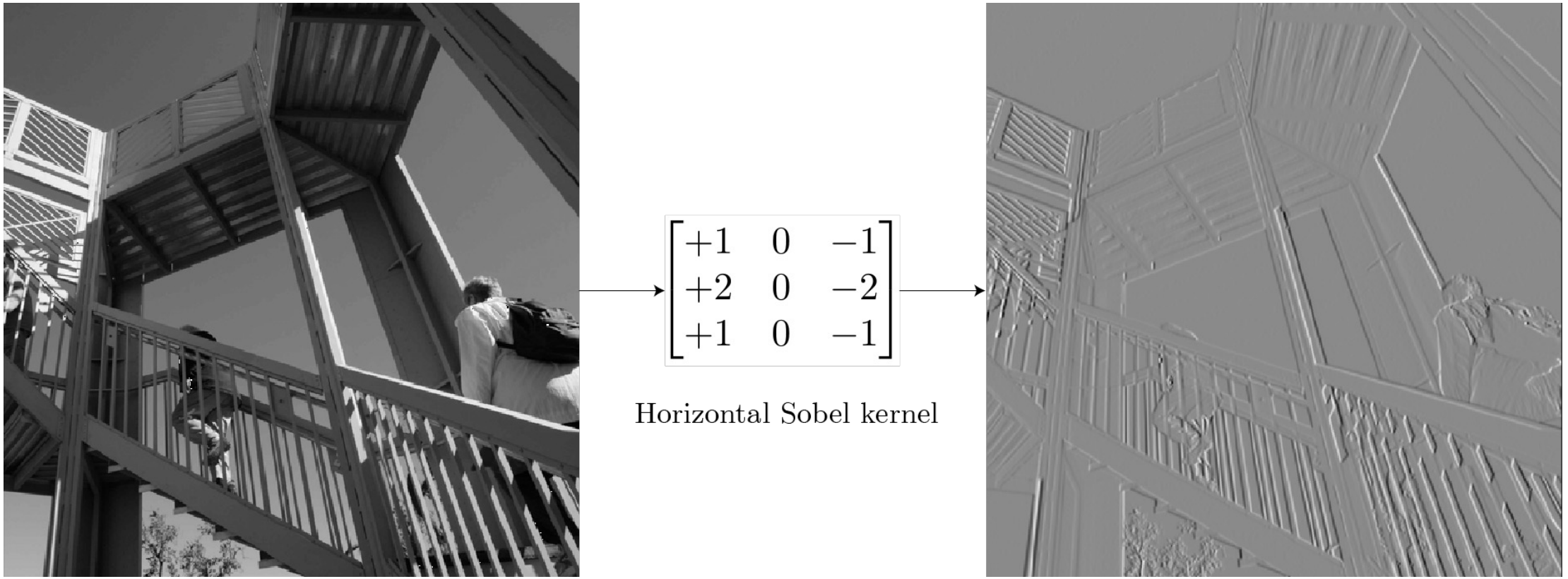

3 0
0 3
*
1 0
0 1
=
6


5 0
0 5
*
1 0
0 1
=
10

Максимум свёртки инвариантен к сдвигам



Свёртки в компьютерном зрении



$$\begin{bmatrix} +1 & 0 & -1 \\ +2 & 0 & -2 \\ +1 & 0 & -1 \end{bmatrix}$$

Horizontal Sobel kernel

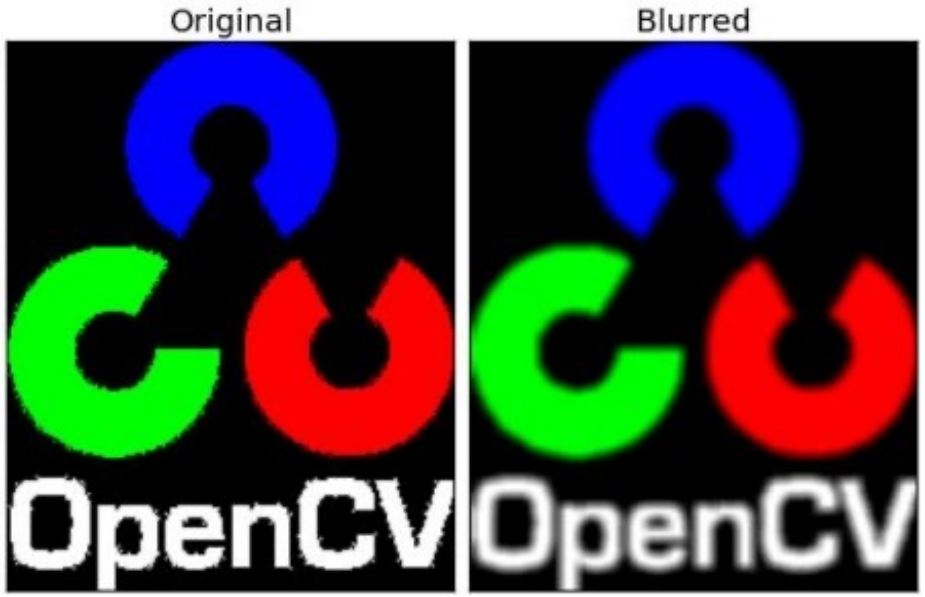
https://towardsdatascience.com/intuitively-understanding-convolutions-for-deep-learning-1f6f42faee1

Свёртки в компьютерном зрении

$$\begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} - \frac{1}{9}\begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 0 \end{bmatrix} - \frac{1}{9}\begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$$

https://ai.stanford.edu/~syyeung/cvweb/tutorial1.html

Свёртки в компьютерном зрении



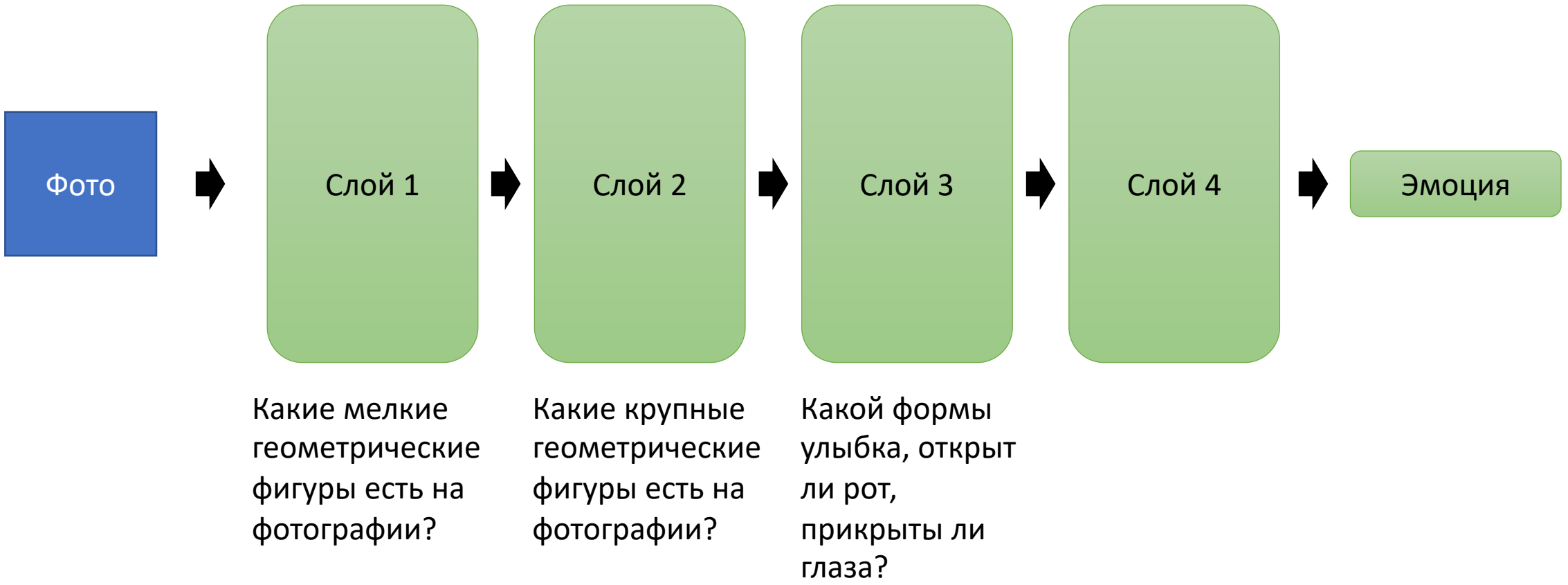
$$\frac{1}{9}\begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$$

https://docs.opencv.org/master/d4/d13/tutorial_py_filtering.html

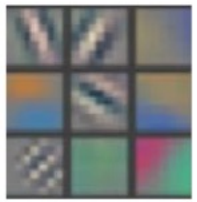
Свёрточные сети

- Из изображения выделяются всё более верхнеуровневые признаки с помощью свёрток
- Фильтры в свёртках **обучаются**

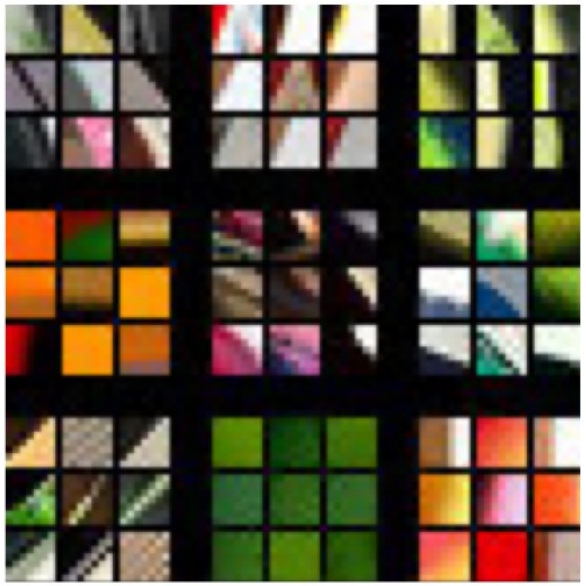
Свёрточные сети



Что видят фильтры?

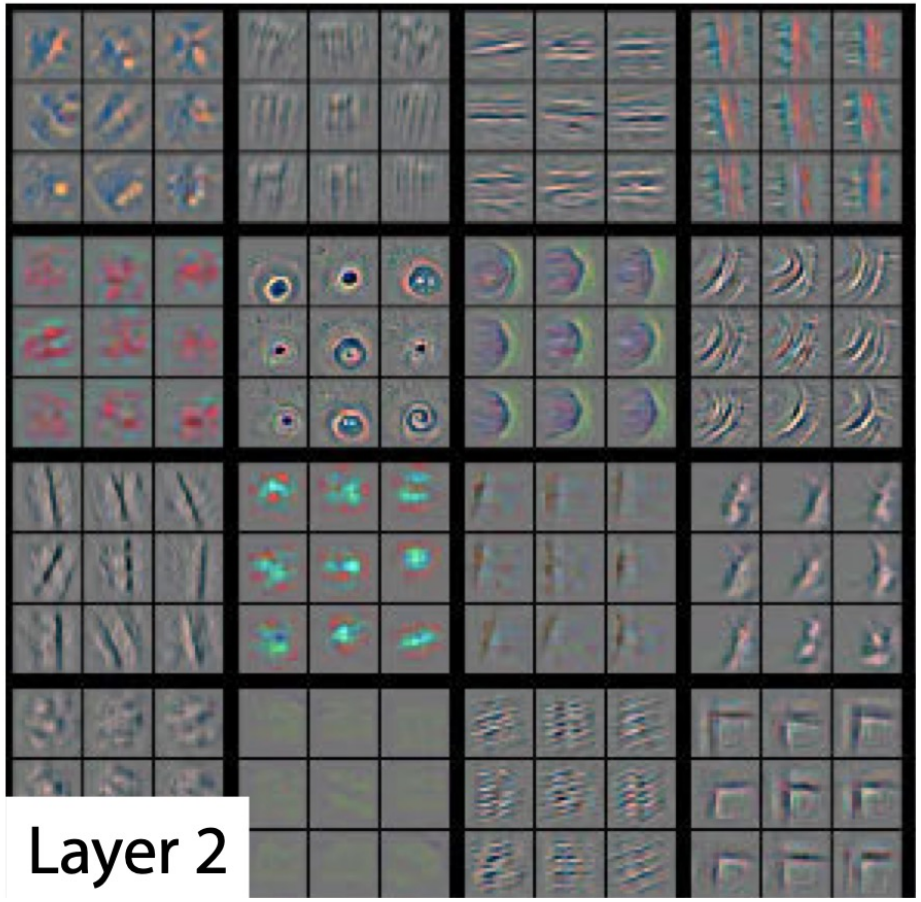


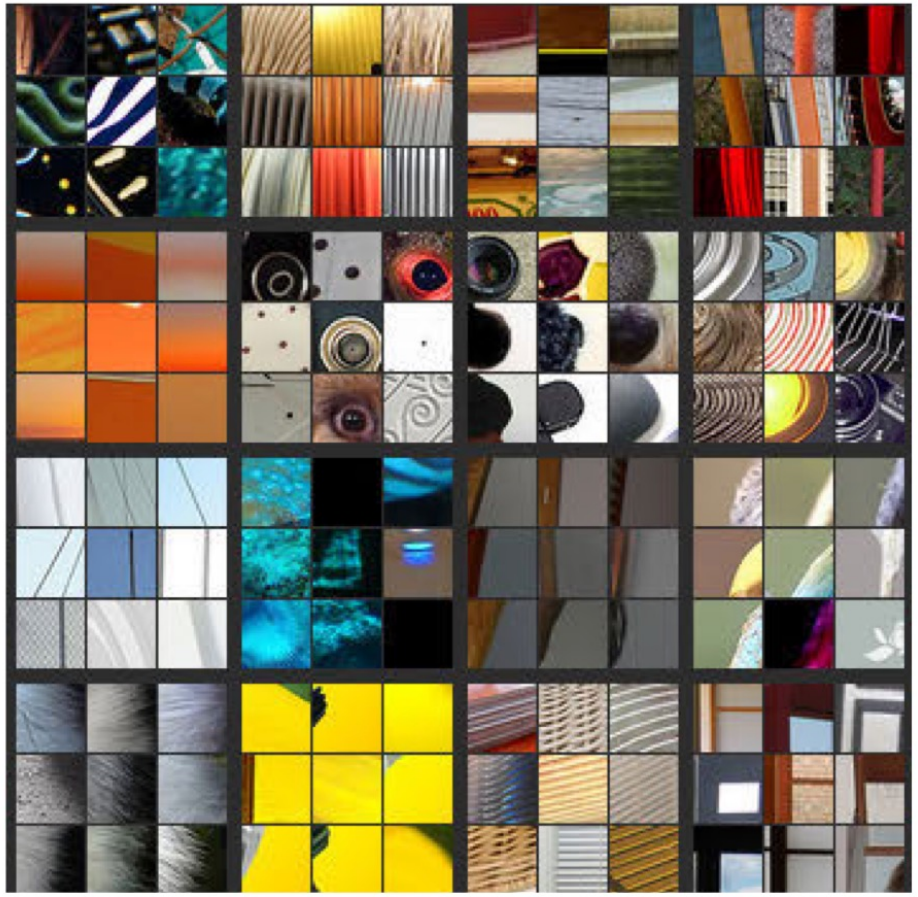
Layer 1



https://arxiv.org/abs/1311.2901

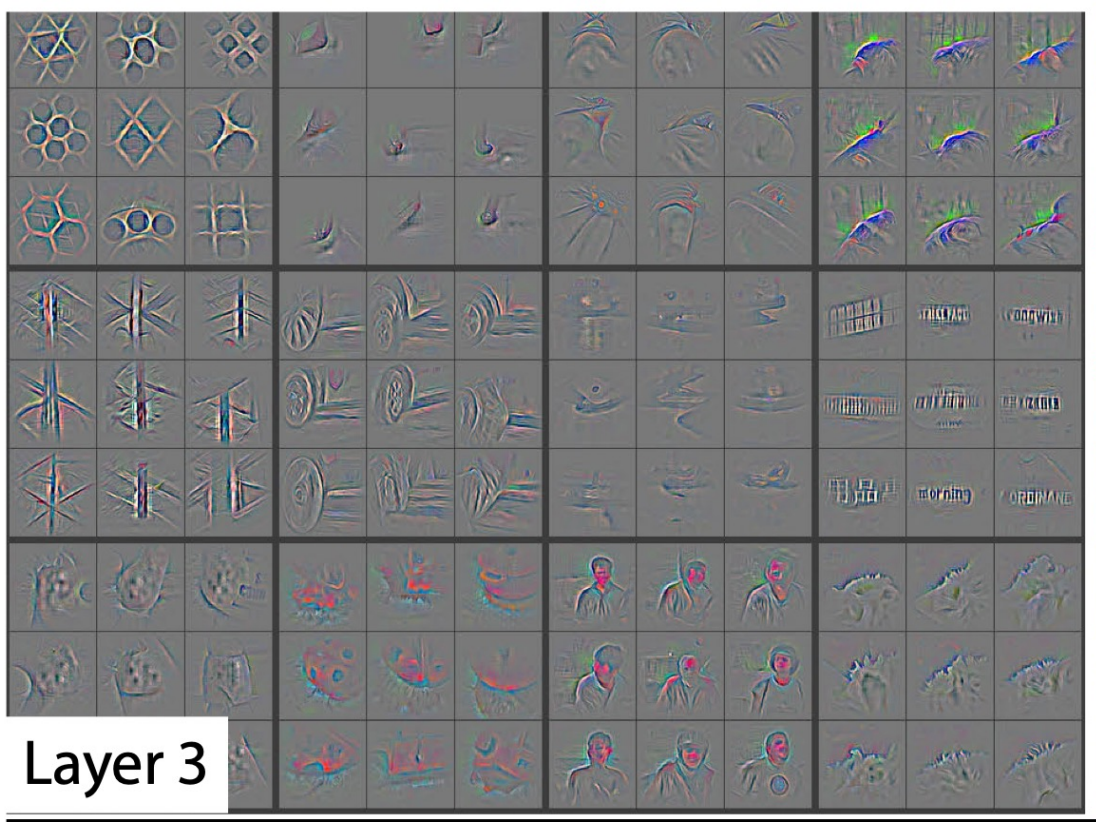
Что видят фильтры?





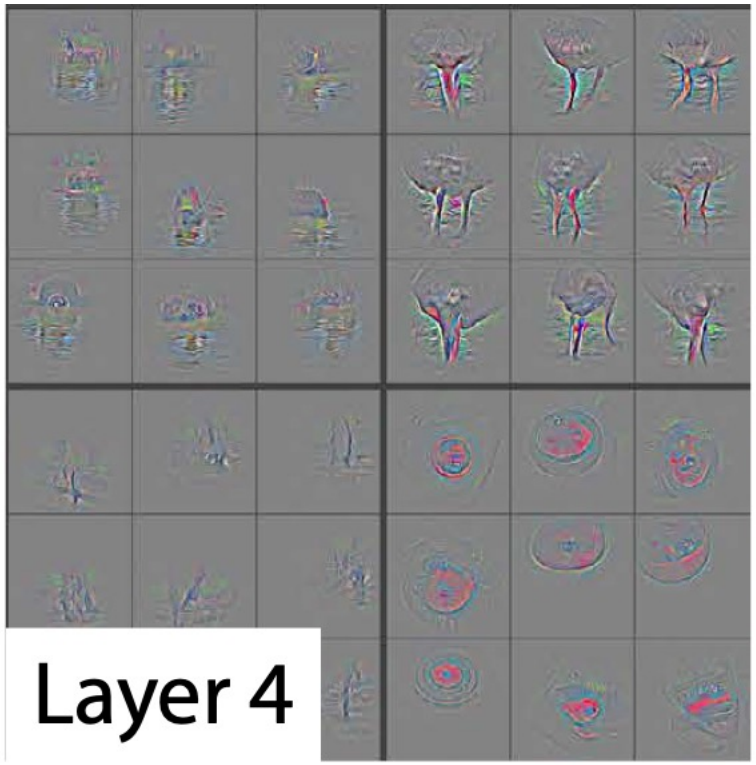
https://arxiv.org/abs/1311.2901

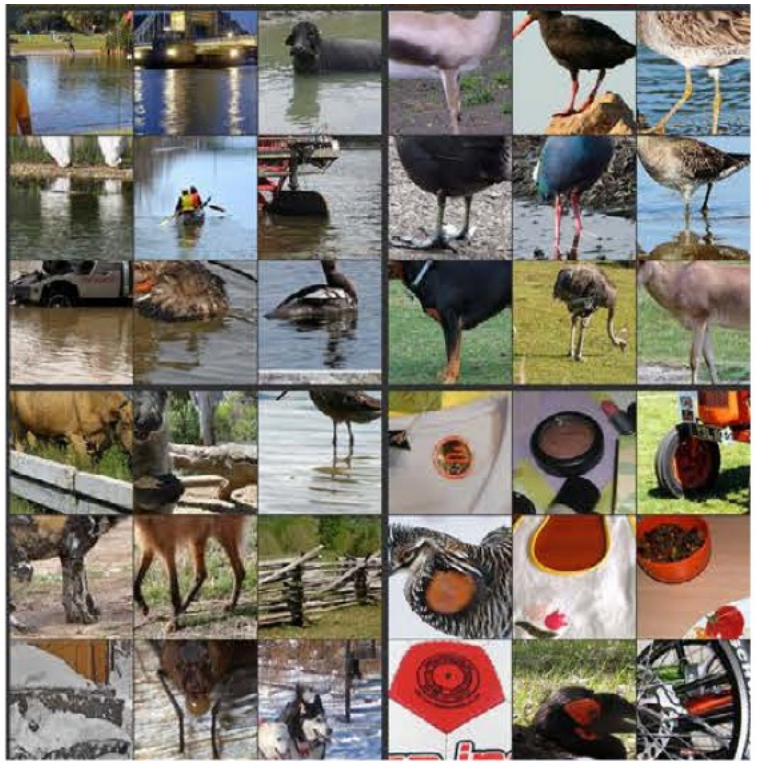
Что видят фильтры?



https://arxiv.org/abs/1311.2901

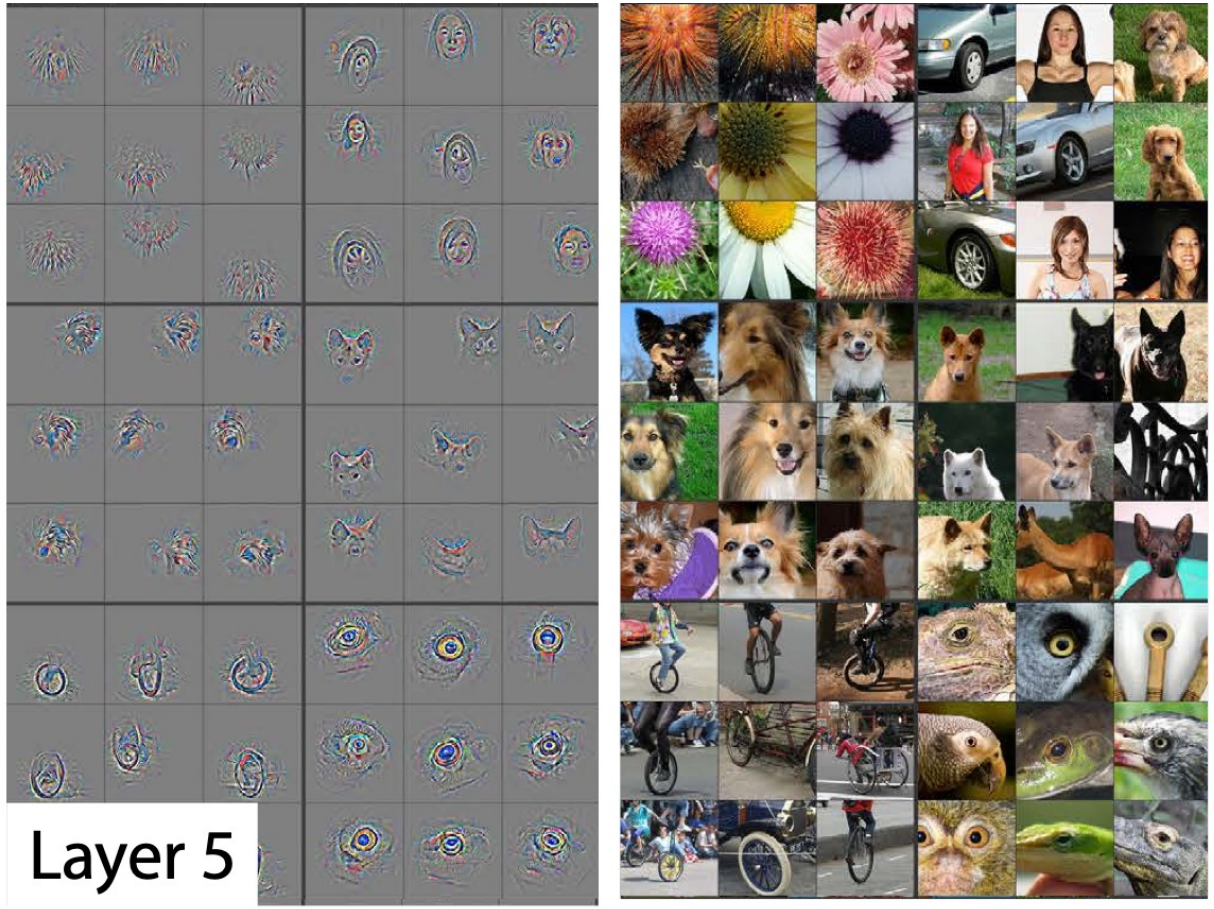
Что видят фильтры?





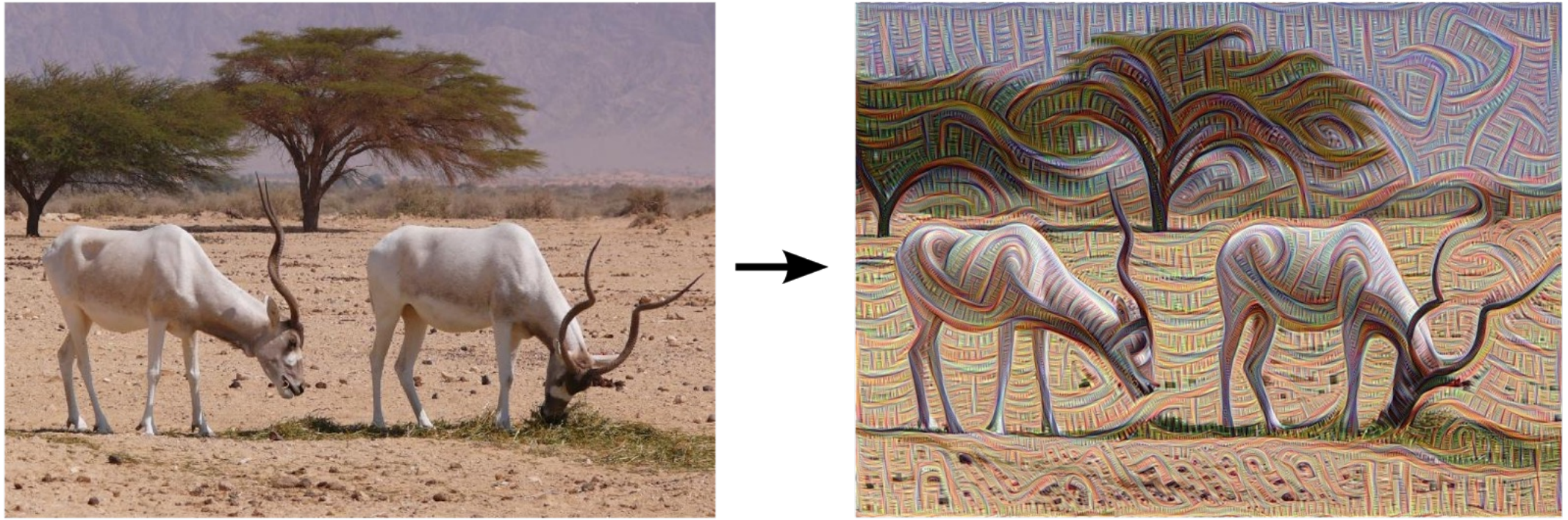
https://arxiv.org/abs/1311.2901

Что видят фильтры?



https://arxiv.org/abs/1311.2901

Максимизация вероятности класса



https://research.googleblog.com/2015/06/inceptionism-going-deeper-into-neural.html

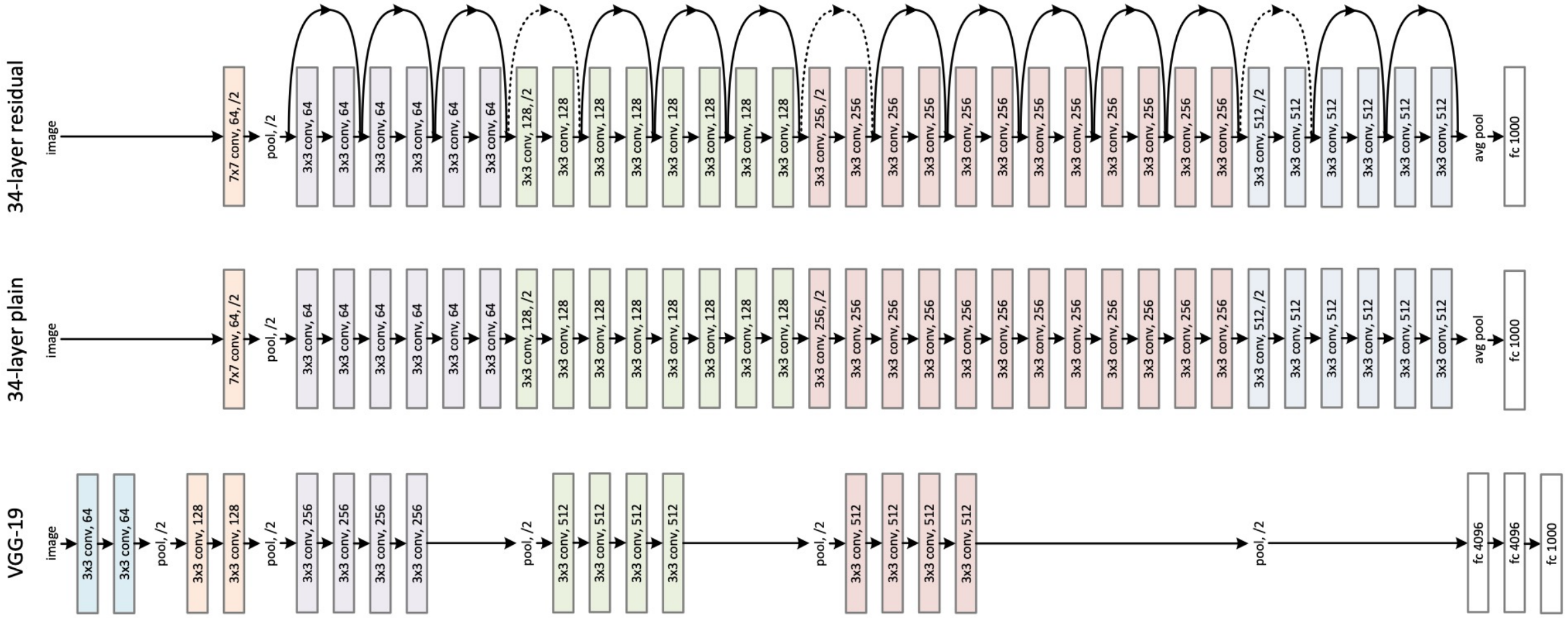
Максимизация вероятности класса

https://research.googleblog.com/2015/06/inceptionism-going-deeper-into-neural.html

Максимизация вероятности класса

https://research.googleblog.com/2015/06/inceptionism-going-deeper-into-neural.html

ResNet (2015)


VGG-19
image
3x3 conv, 64
3x3 conv, 64
pool, /2
3x3 conv, 128
3x3 conv, 128
pool, /2
3x3 conv, 256
3x3 conv, 256
3x3 conv, 256
3x3 conv, 256
pool, /2
3x3 conv, 512
3x3 conv, 512
3x3 conv, 512
3x3 conv, 512
pool, /2
3x3 conv, 512
3x3 conv, 512
3x3 conv, 512
3x3 conv, 512
pool, /2
fc 4096
fc 4096
fc 1000
34-layer plain
image
7x7 conv, 64, /2
pool, /2
3x3 conv, 64
3x3 conv, 128
3x3 conv, 128, /2
3x3 conv, 256
3x3 conv, 256, /2
3x3 conv, 512
3x3 conv, 512, /2
avg pool
fc 1000
34-layer residual
image
7x7 conv, 64, /2
pool, /2
3x3 conv, 64
3x3 conv, 128
3x3 conv, 128, /2
3x3 conv, 256
3x3 conv, 256, /2
3x3 conv, 512
3x3 conv, 512, /2
avg pool
fc 1000

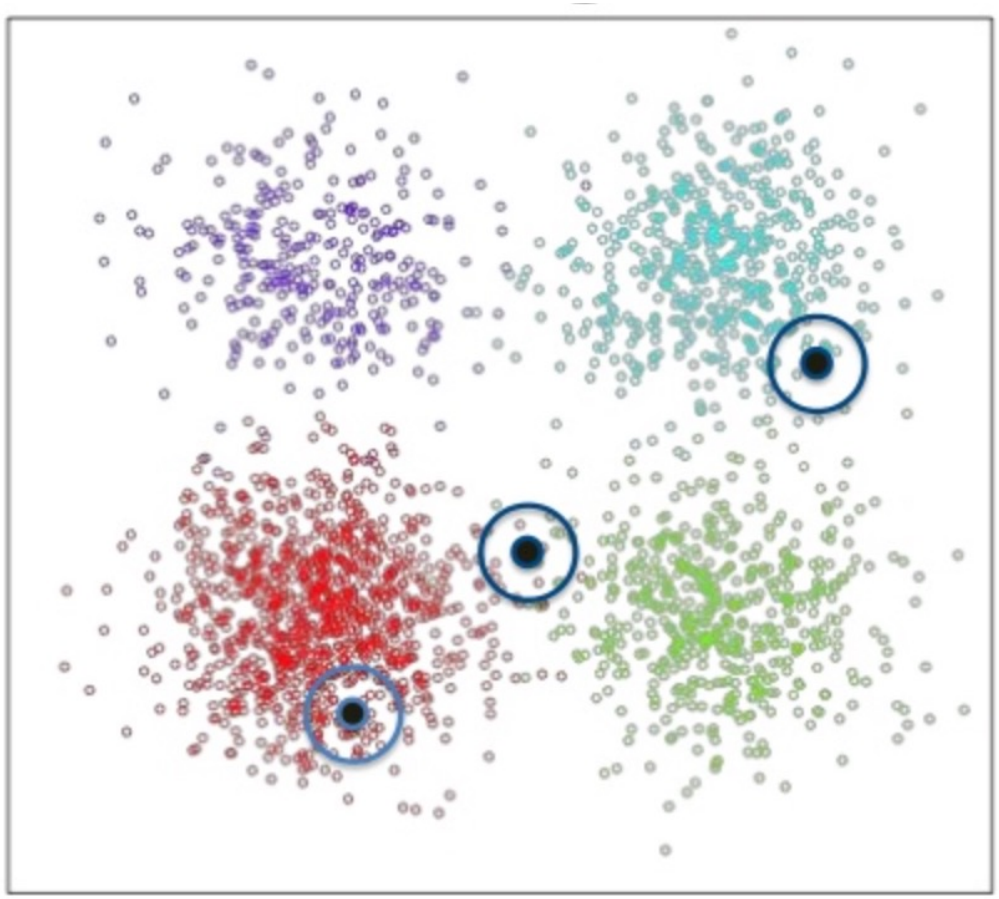
Необычные свойства функций потерь

Функция потерь

$$Q(w) = \sum_{i=1}^{\ell} L\big(y_i, a(x_i, w)\big) \to \min_w$$

- x_i, y_i — объект и правильный ответ
- $a(x, w)$ — модель с параметрами w
- $L(y, z)$ — функция потерь

Гипотеза компактности



kNN: обучение

- Дано: обучающая выборка $X = (x_i, y_i)_{i=1}^{\ell}$
- Задача классификация (ответы из множества $\mathbb{Y} = \{1, \dots, K\}$)

- Обучение модели:
  - Запоминаем обучающую выборку X

kNN: применение

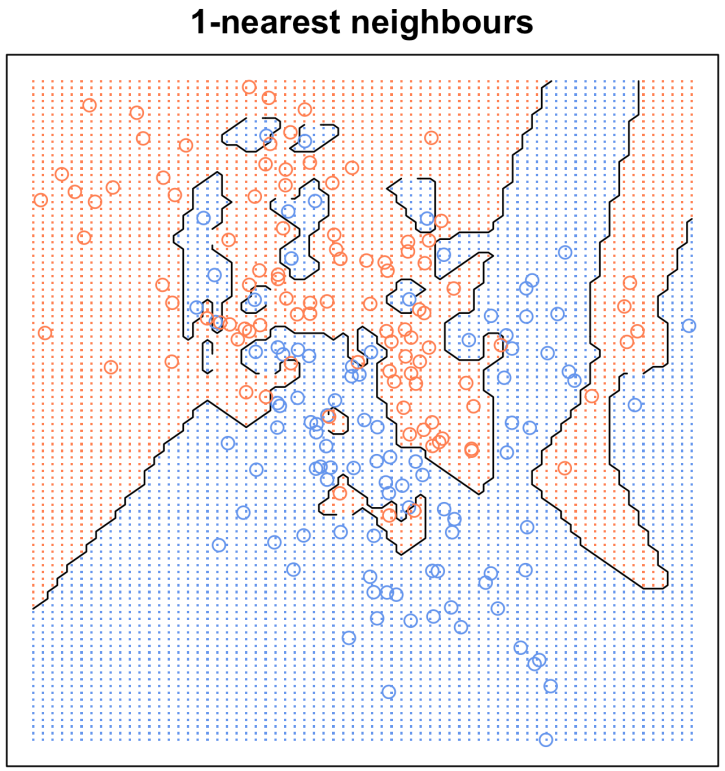
Дано: новый объект x

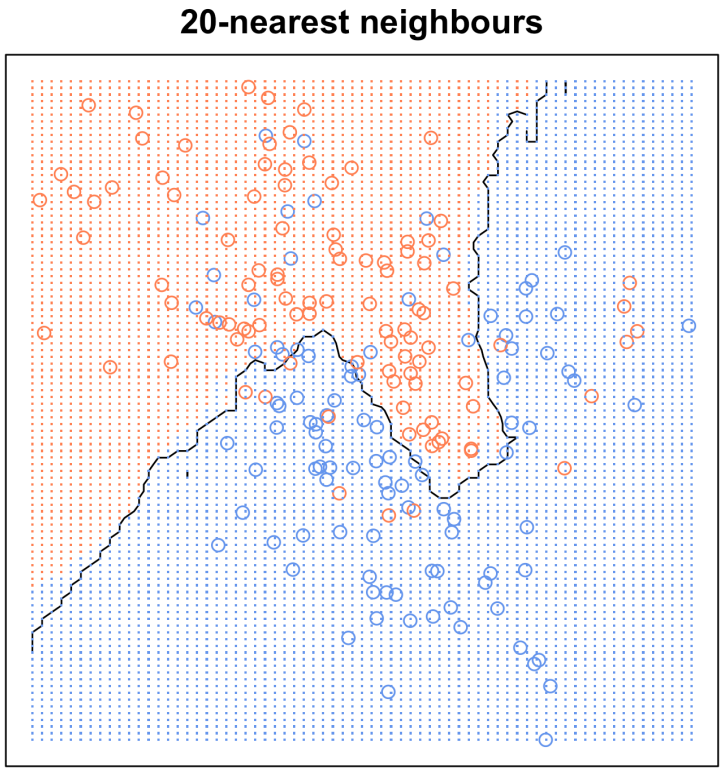
Применение модели:

- Сортируем объекты обучающей выборки по расстоянию до нового объекта: $\rho(x, x_{(1)}) \leq \rho(x, x_{(2)}) \leq \cdots \leq \rho(x, x_{(\ell)})$
- Выбираем k ближайших объектов: $x_{(1)}, \ldots, x_{(k)}$
- Выдаём наиболее популярный среди них класс:

$$a(x) = \arg\max_{y \in \mathbb{Y}} \sum_{i=1}^{k} [y_{(i)} = y]$$

Как выбрать k?





https://kevinzakka.github.io/2016/07/13/k-nearest-neighbor/

Обобщающая способность

Как готовиться к экзамену?

Заучить все примеры с занятий	Разобраться в предмете и усвоить алгоритмы решения задач

Обобщающая способность

Как готовиться к экзамену?

Заучить все примеры с занятий	Разобраться в предмете и усвоить алгоритмы решения задач
Переобучение (overfitting)	Обобщение (generalization)

Обобщающая способность

Как готовиться к экзамену?

Заучить все примеры с занятий	Разобраться в предмете и усвоить алгоритмы решения задач
Переобучение (overfitting)	Обобщение (generalization)
Хорошее качество на обучении Низкое качество на новых данных	Хорошее качество на обучении Хорошее качество на новых данных

Отложенная выборка

Обучение

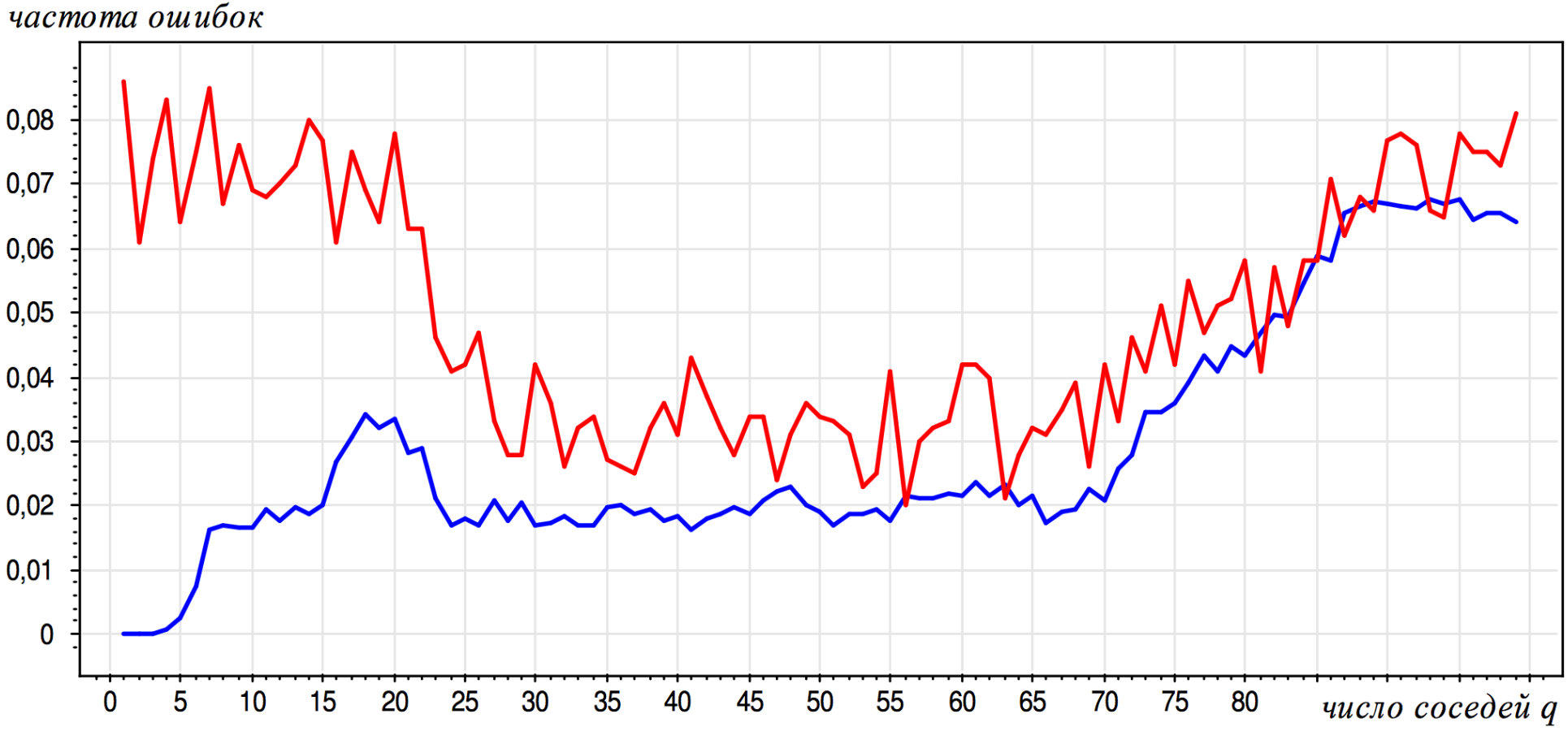
Тест

Отложенная выборка

Обучение Тест

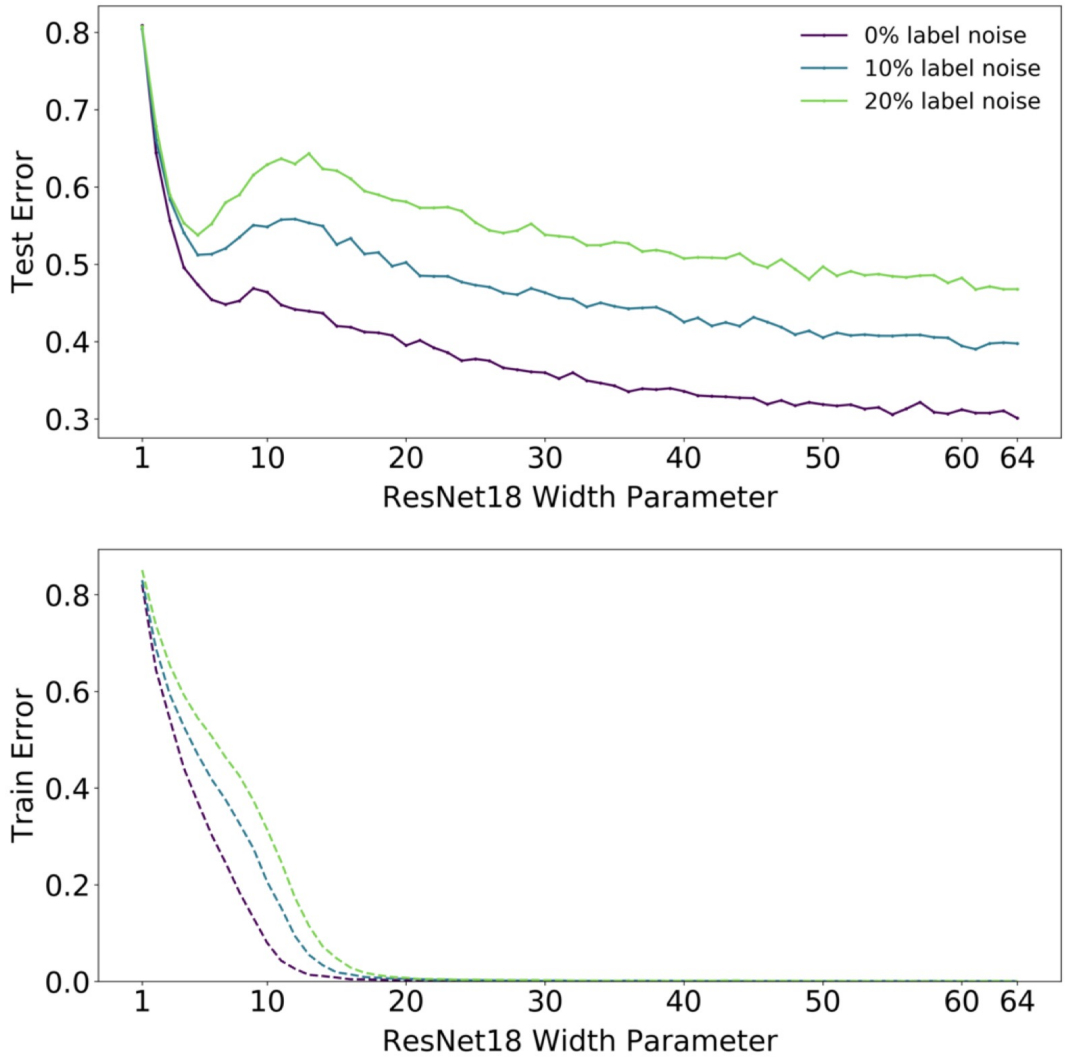
- Слишком большое обучение — тестовая выборка нерепрезентативна
- Слишком большой тест — модель не сможет обучиться
- Обычно: 70/30, 80/20

Подбор числа соседей



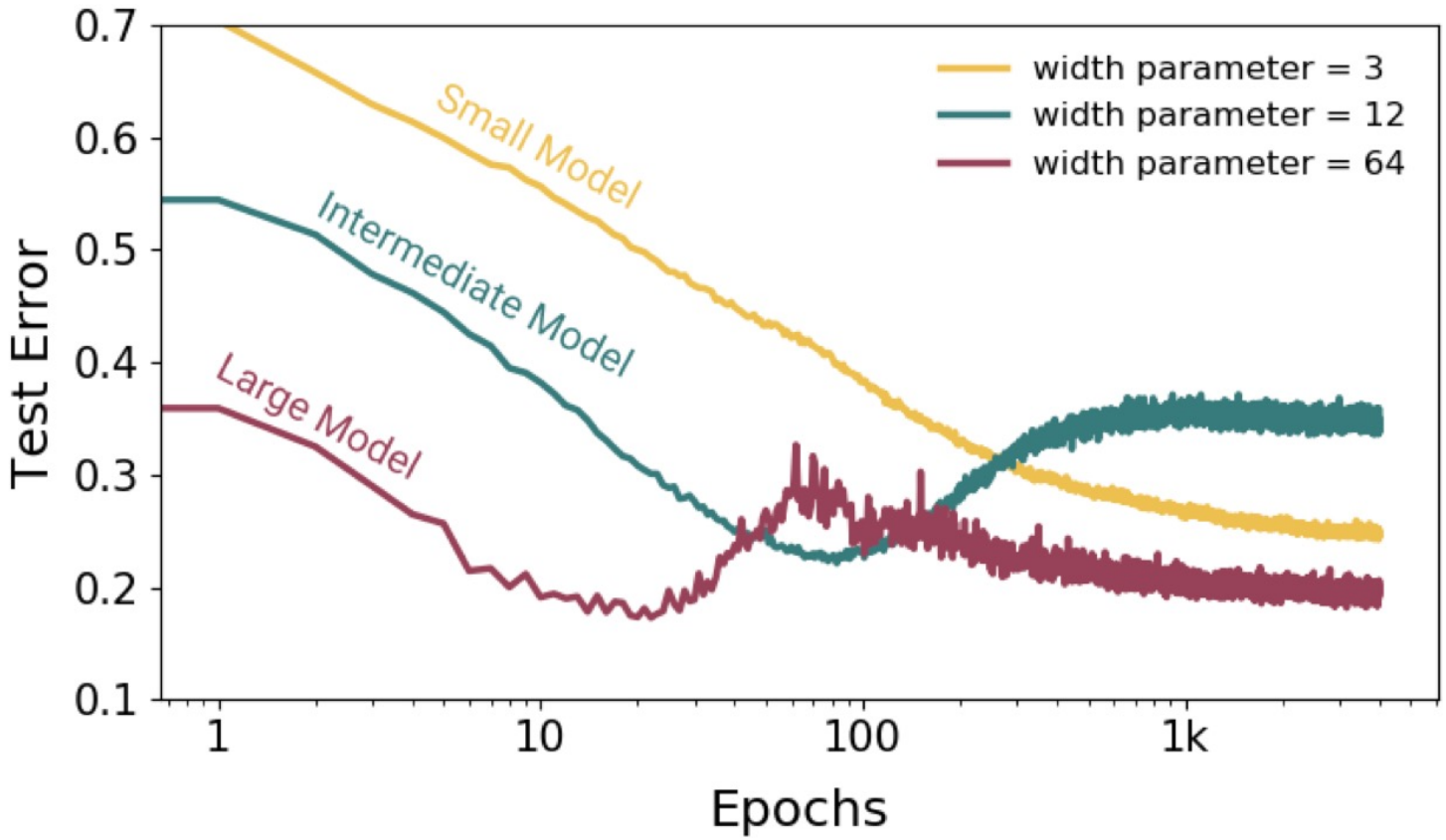
http://www.machinelearning.ru/wiki/index.php?title=MO

Что если увеличивать число параметров?



https://arxiv.org/abs/1912.02292

Что если продолжать градиентный спуск?



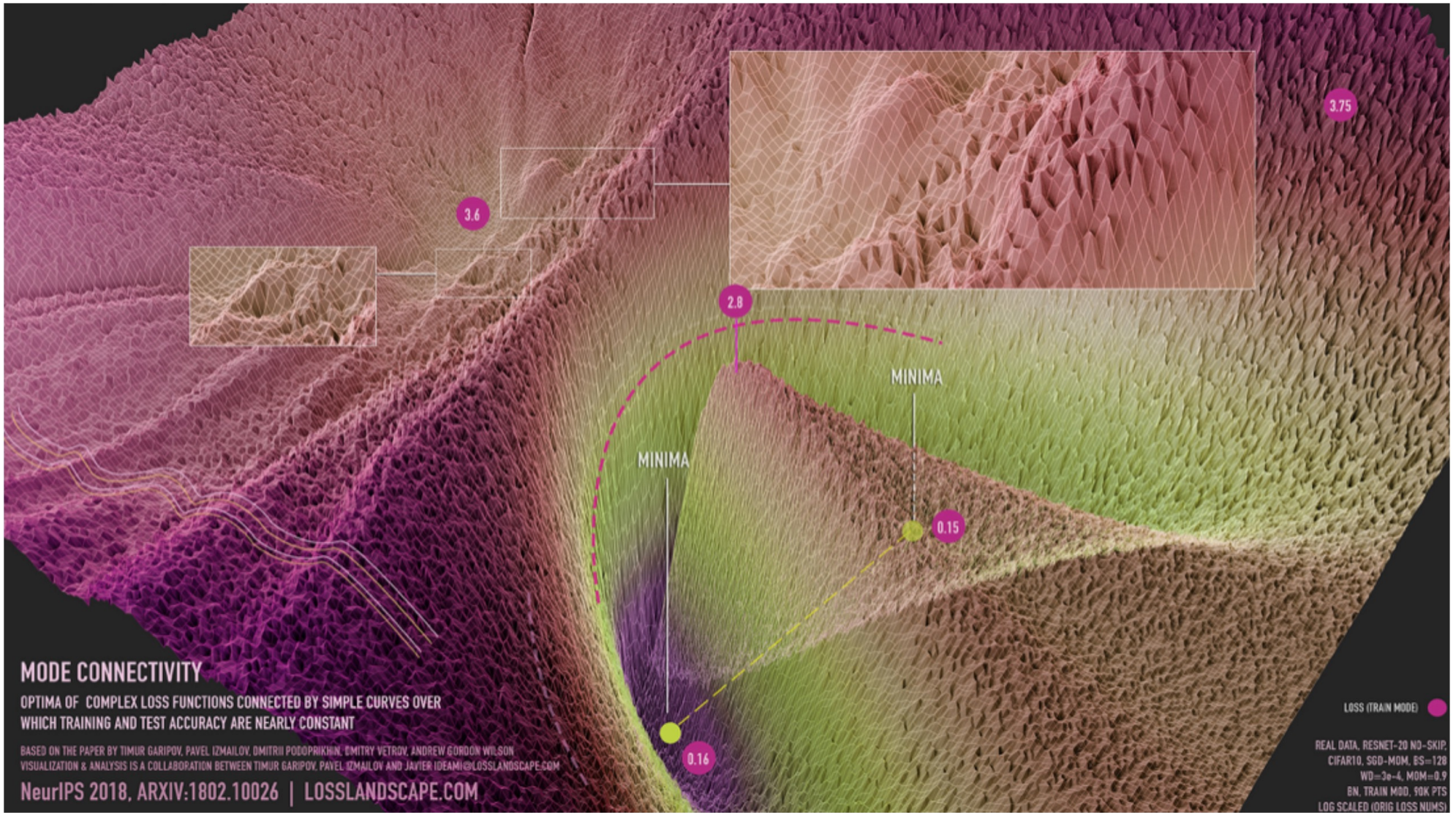
https://arxiv.org/abs/1912.02292

Эффект двойного спуска

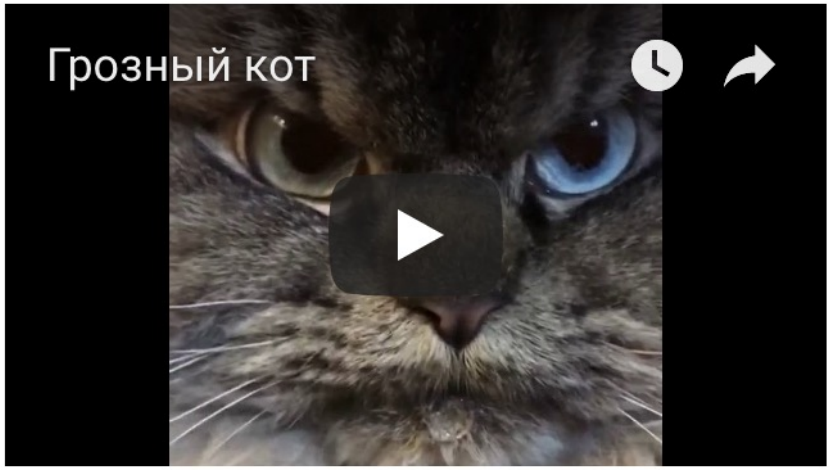
Три режима работы моделей:

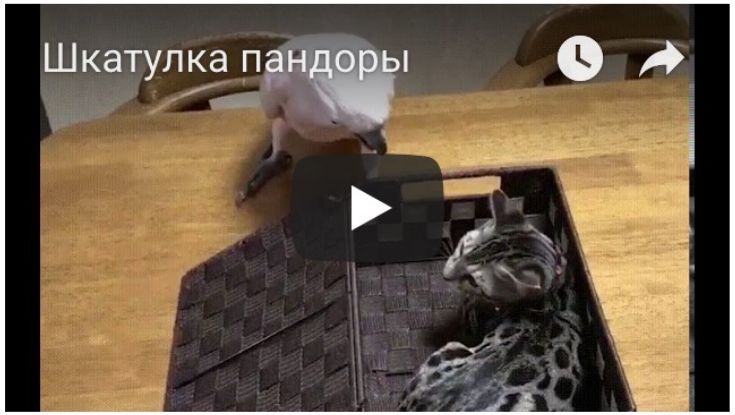
- Недопараметризованный режим
- Стандартный режим
- Перепараметризованный режим

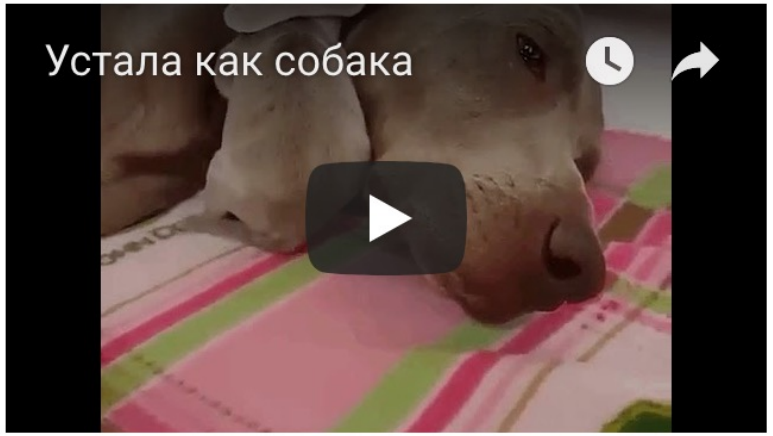
Рельеф функции потерь

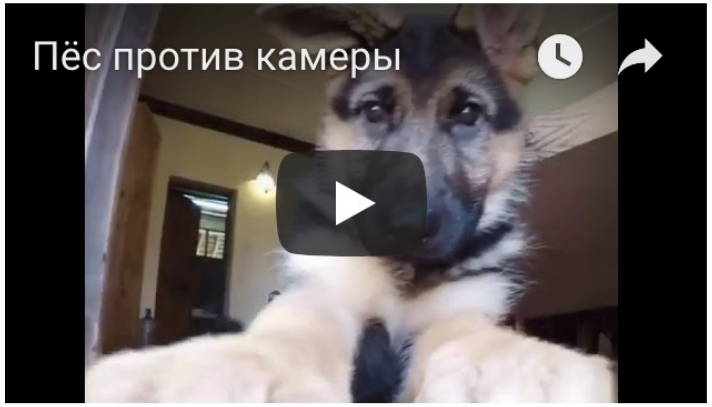


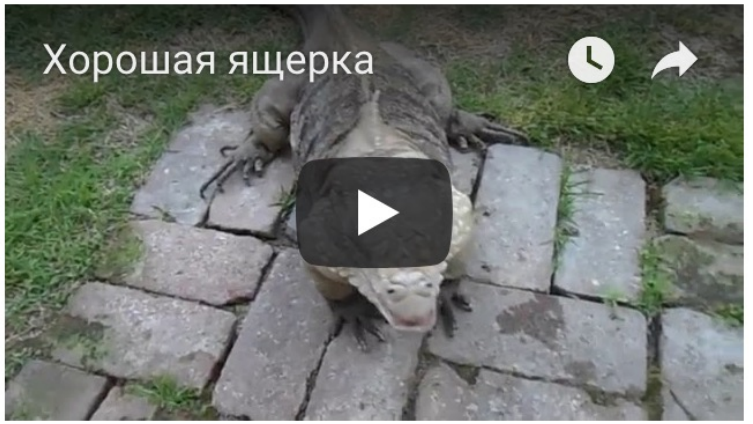
Рекомендательные системы

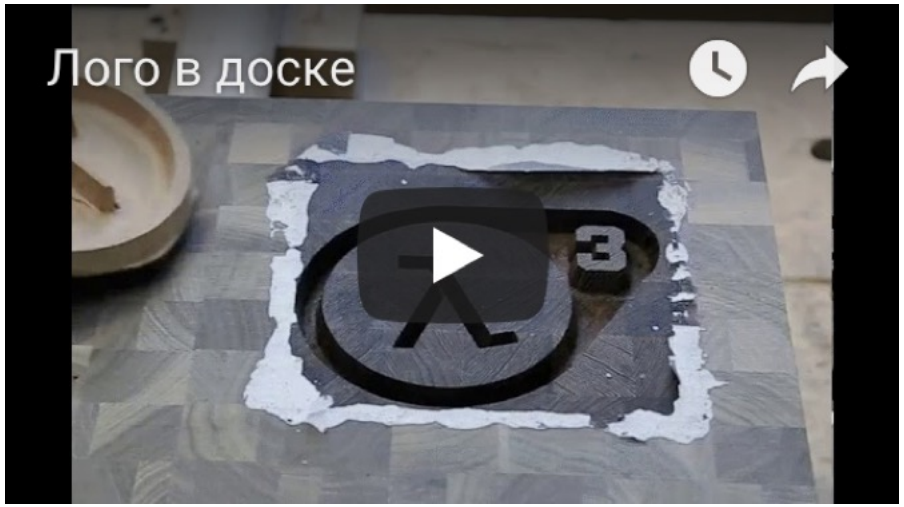

Грозный кот


Шкатулка пандоры


Устала как собака

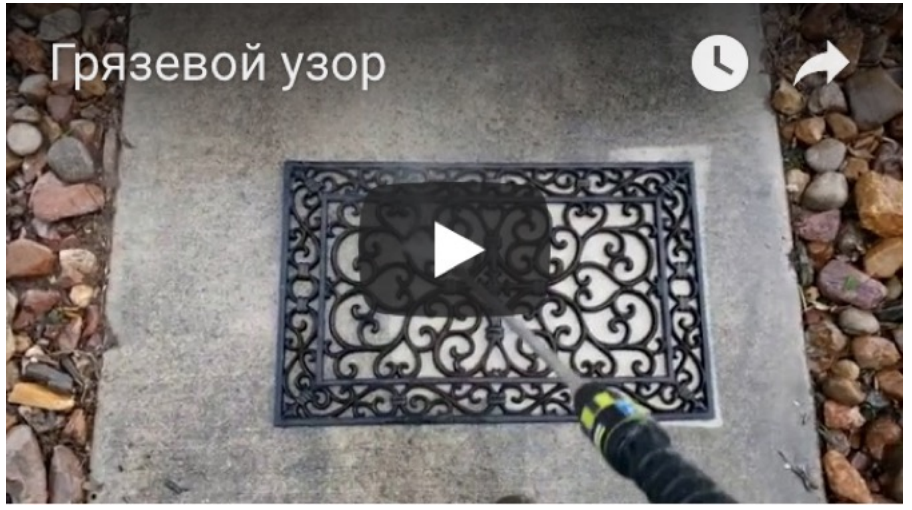

Пёс против камеры

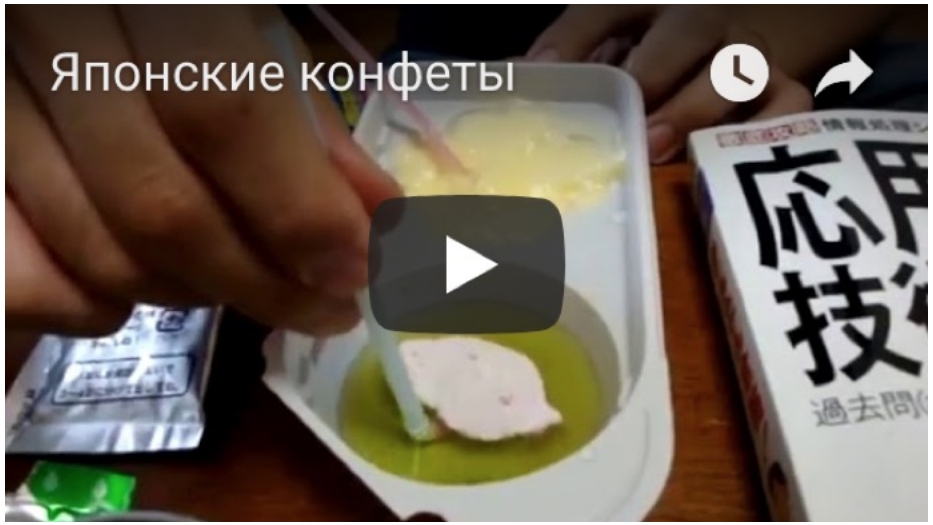

Хорошая ящерка

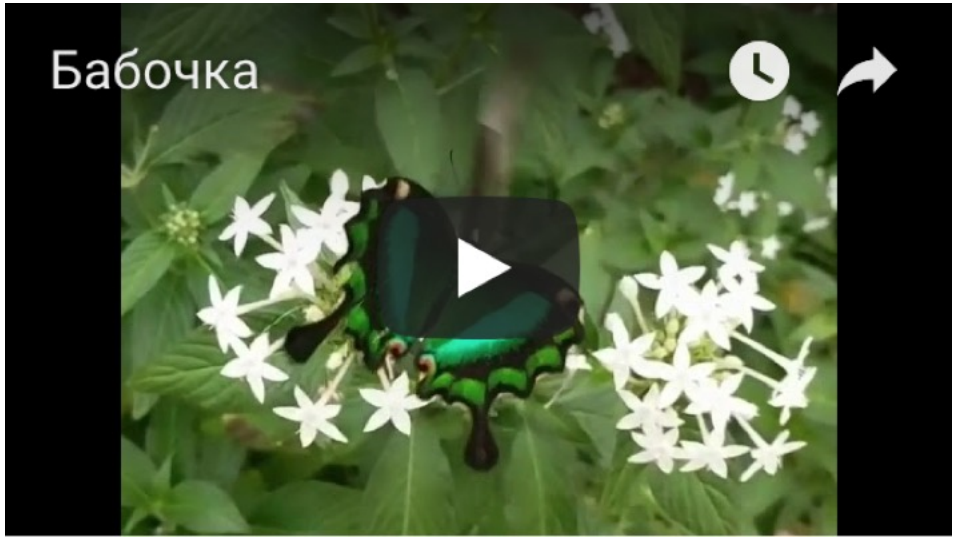

Лого в доске

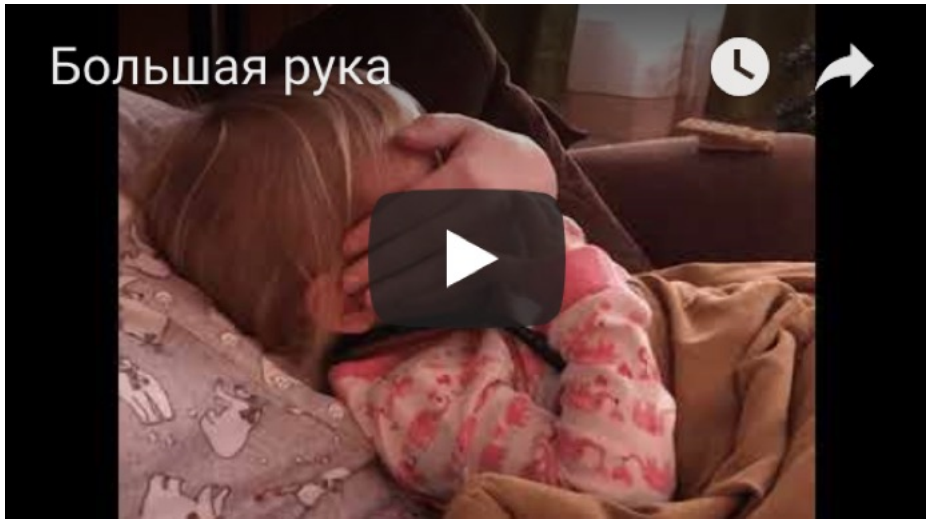

Варим металл

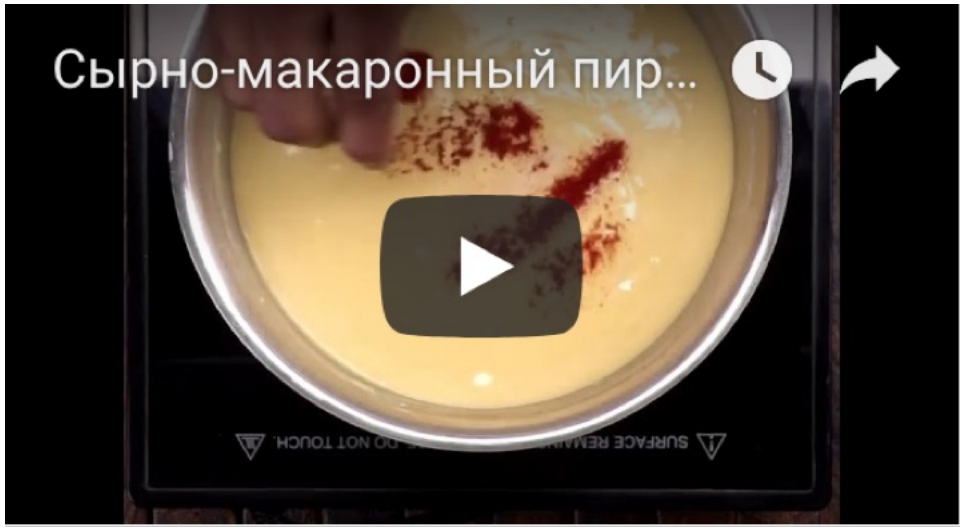

Водный сборщик мусора
SEABIN PROJECT
FOR CLEANER OCEANS


Грязевой узор


Японские конфеты


Бабочка


Большая рука


Сырно-макаронный пир...

User-based collaborative filtering

Видео

Пользователи

1	1	0		1	
0	1	1			1
			1	1	0
	1	1		0	
	1				1

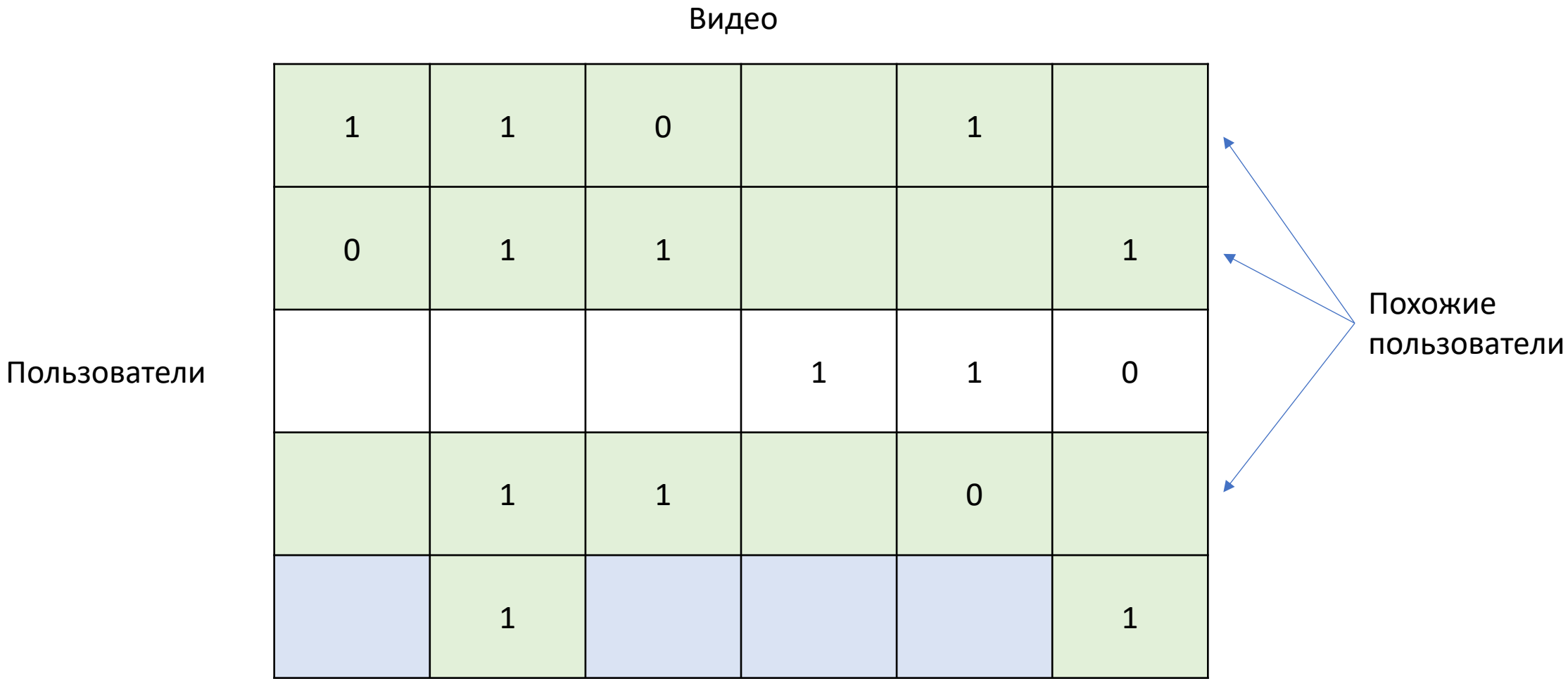
User-based collaborative filtering

Видео

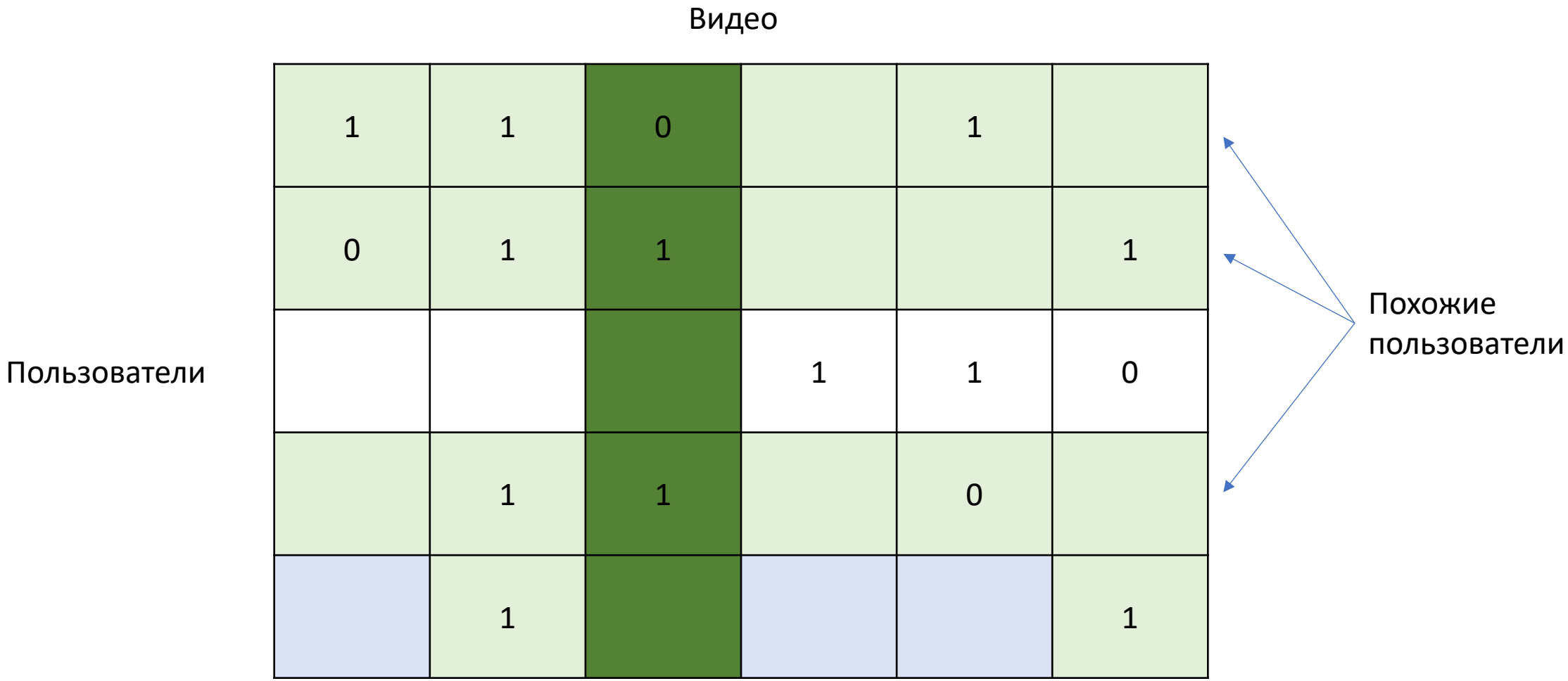
Пользователи

1	1	0		1	
0	1	1			1
			1	1	0
	1	1		0	
	1				1

User-based collaborative filtering



User-based collaborative filtering

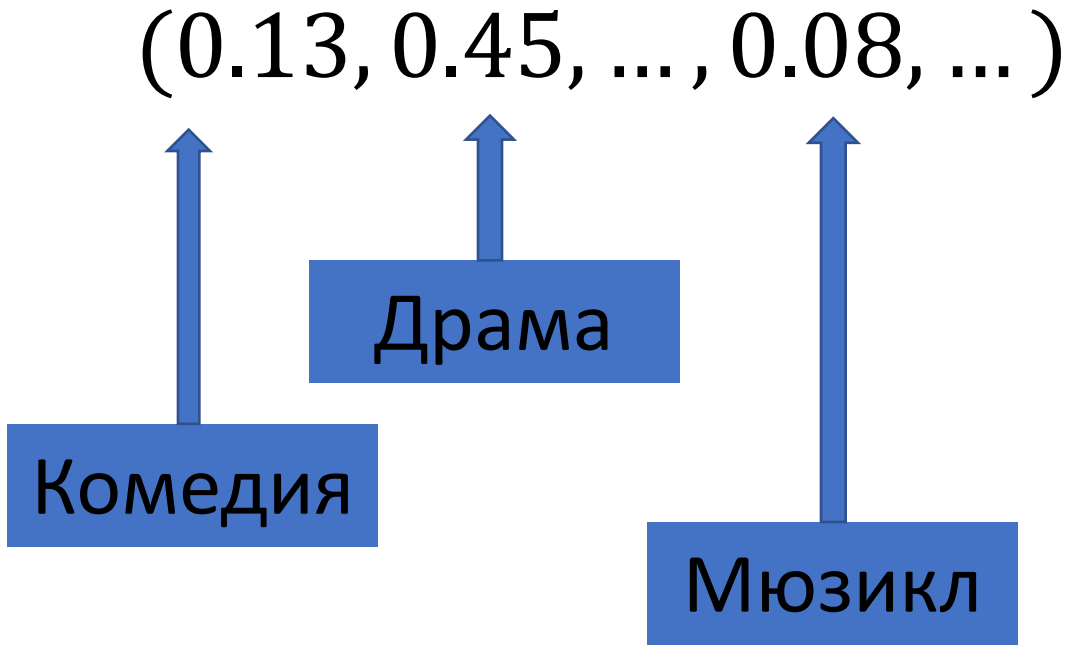


User-based collaborative filtering

- Решение чисто эвристическое
- Можно ли лучше?

Векторы интересов

- Для пользователя — насколько он интересуется каждым жанром
- Для фильма — насколько он относится к каждому жанру



Рейтинг

- Предположение: заинтересованность определяется как скалярное произведение векторов пользователя и фильма

$$(0.1, 0.5, 0.01, 0.92) \times (0, 0, 0.1, 0.95) = 0.875$$

$$(0.1, 0.5, 0.01, 0.92) \times (0.9, 0, 0, 0.1) = 0.182$$

Пользователь

Фильм

Модели со скрытыми переменными

- Обучим вектор p_u для каждого пользователя u
- Обучим вектор q_i для каждого товара i
- Оценка приближается их скалярным произведением:

$$r_{ui} \approx \langle p_u, q_i \rangle$$

- Находим векторы только по известным оценкам
- После этого можем предсказать оценку для любой пары «пользователь-товар»

Модели со скрытыми переменными

- Оптимизационная задача:

$$\sum_{(u,i)\in R} \left(r_{ui} - \bar{r}_u - \bar{r}_i - \langle p_u, q_i \rangle\right)^2 \to \min_{P,Q}$$

- Решение: градиентный спуск, Alternating Least Squares (ALS) и другие методы

Модели со скрытыми переменными

2	5	
5		4
	1	
	2	5

Модели со скрытыми переменными

	(0.9, 0.05)	(0.02, 1.1)	(1.05, 0.01)
(2.1, 5)	2	5	
(4.6, 0)	5		4
(0, 1)		1	
(4.9, 0.9)		1	5

Коллаборативная фильтрация

- Десятки миллионов пользователей и сотни тысяч видео
- Нужно уметь строить векторы очень быстро и распределённо

Резюме

- Машинное обучение — подбор алгоритма под данные
- Много интересных подходов для текстов, картинок, звука
- В основе лежит много математики и алгоритмов

tg: @esokolov

esokolov@hse.ru