

Data Mining in Action

# Работа с признаками

Виктор Кантор



# Виды признаков

Какие бывают признаки:

1. Числовые
2. Порядковые
3. Категориальные
4. Даты и время
5. Координаты

# Даты и время

1. Количество прошедших секунд

например, с 00:00:00 UTC, 1 January 1970

2. Использование периодичности

а. номер дня в году, в месяце, в неделе

б. час, минута, секунда

3. Время до/после важных событий

Например, количество дней, оставшихся до ближайшего праздника

# Координаты

1. Повороты системы координат на 45 градусов, 22.5 градусов, etc
2. Добавление расстояний до:
  - a. Других объектов из выборки
  - b. Центров кластеров
  - c. Инфраструктурных зданий - магазинов, школ, больниц

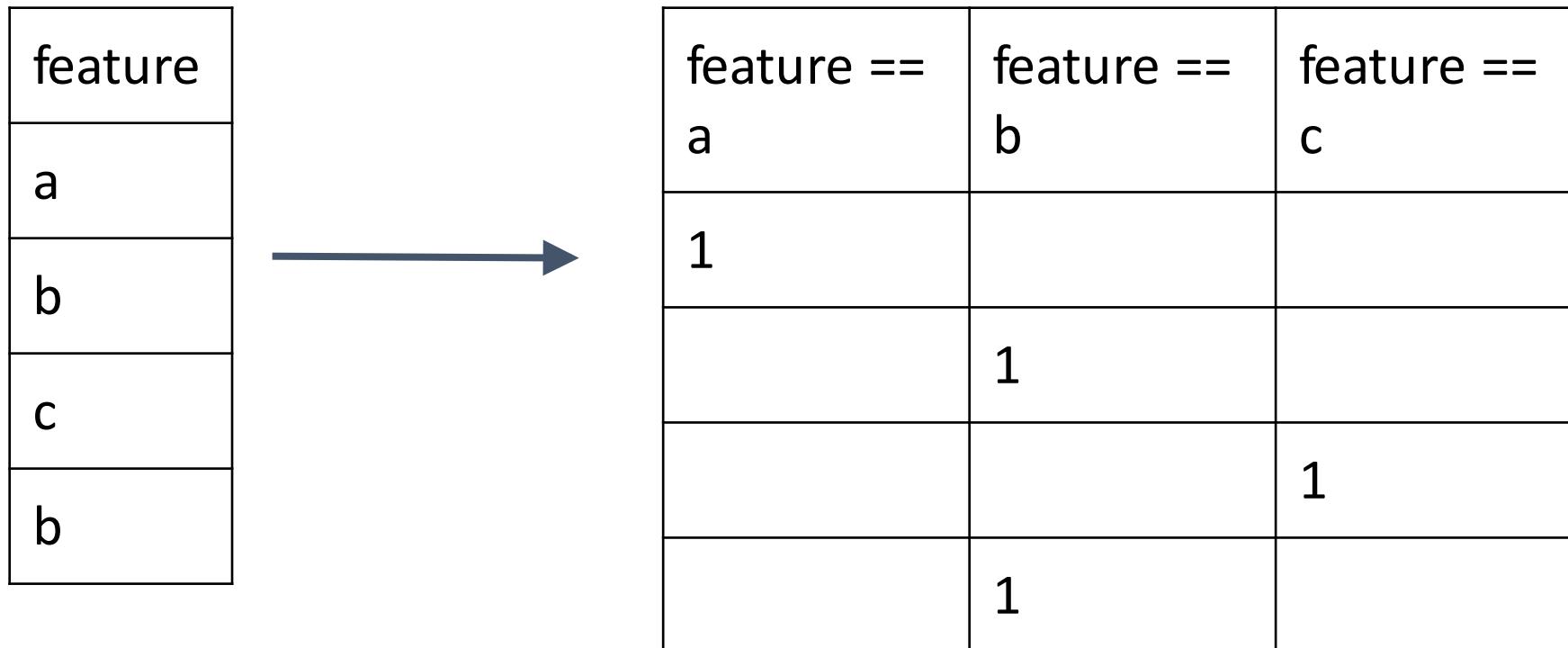
# Категориальные признаки (строки)

Из колонок “name”, “ticket”, “cabin” можно сгенерировать новые признаки

	A	B	C	D	E	F	G	H	I	J	K
1	survived	pclass	name	sex	age	sibsp	parch	ticket	fare	cabin	embarked
2	0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171	7.25		S
3	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38	1	0	PC 17599	71.2833	C85	C
4	1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2.	7.925		S
5	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35	1	0	113803	53.1	C123	S
6	0	3	Allen, Mr. William Henry	male	35	0	0	373450	8.05		S
7	0	3	Moran, Mr. James	male		0	0	330877	8.4583		Q
8	0	1	McCarthy, Mr. Timothy J	male	54	0	0	17463	51.8625	E46	S
9	0	3	Palsson, Master. Gosta Leonard	male	2	3	1	349909	21.075		S
10	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27	0	2	347742	11.1333		S
11	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14	1	0	237736	30.0708		C
12	1	3	Sandstrom, Miss. Marguerite Rut	female	4	1	1	PP 9549	16.7	G6	S
13	1	1	Bonnell, Miss. Elizabeth	female	58	0	0	113783	26.55	C103	S

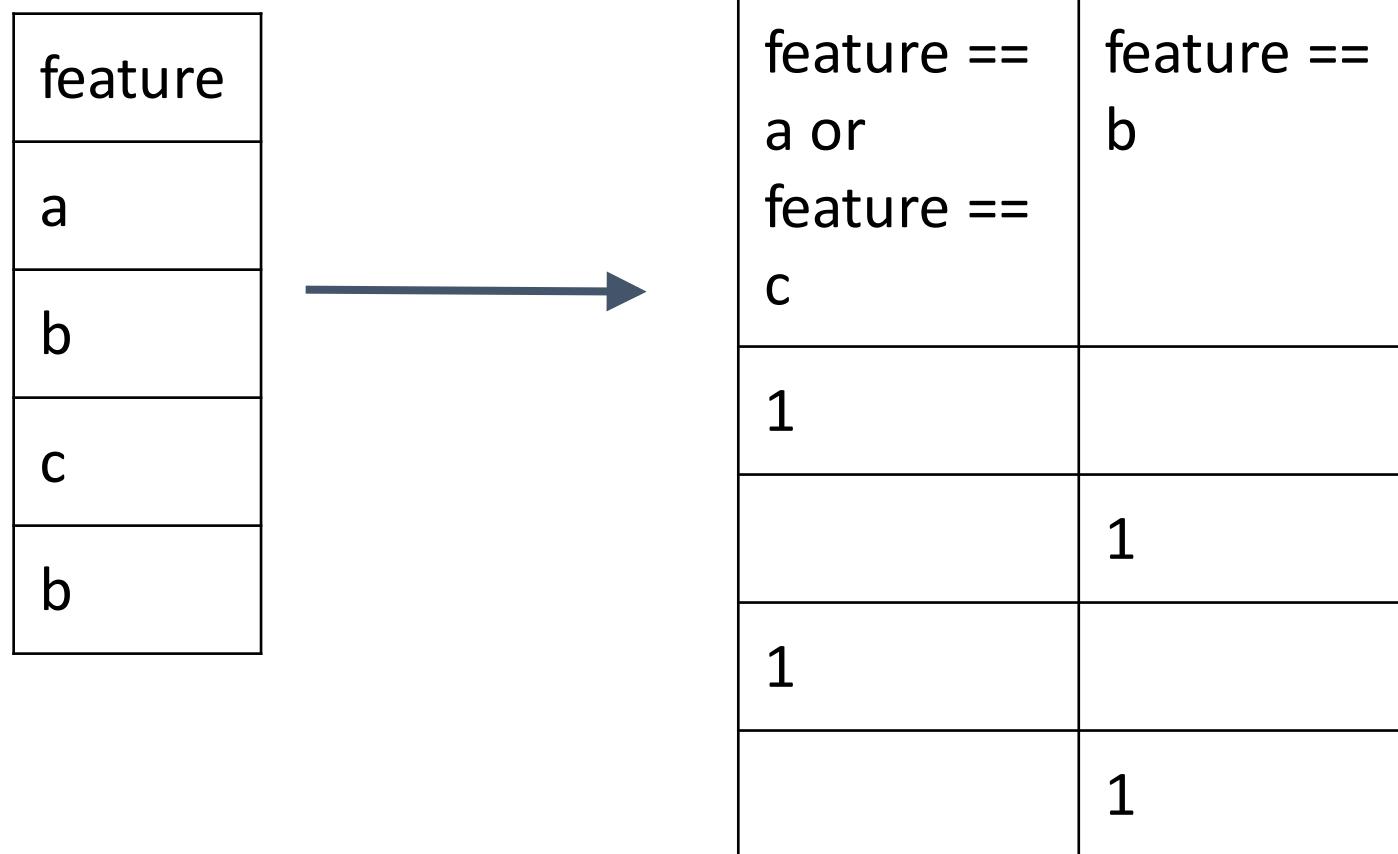
# Категориальные признаки

## Бинаризация



# Категориальные признаки

Hashing trick



# Категориальные признаки

№ склада	Город	...	Продано вина (ящиков)
2343	Москва	...	56000
185	Самара	...	10500
121	Ростов	...	11300
...	...	...	...

# Категориальные признаки

№ склада	В среднем продают в городе	...	Продано вина (ящиков)
2343	59000	...	56000
185	11200	...	10500
121	12100	...	11300
...	...	...	...

# Метапризнаки

Использование ответов других алгоритмов

	xgb_prediction	knn_prediction	svm_prediction	target
train	0.192	0.293	0.122	0
train	0.789	0.890	0.670	1
test	0.542	0.310	0.173	?

Осторожно с переобучением: используйте KFold, LOO

# Feature engineering

- Выделение признаков (feature extraction) – генерация признаков по известным данным
- Отбор признаков (feature selection) – ранжирование признаков по «полезности» и выкидывание наименее полезных (или даже наоборот «вредных»)
- Преобразование признаков (feature transform) – создание новых признаков на основе имеющихся

# Генерация признаков

Для решения задачи нужно использовать разные типы данных

**Пример:** задача рекомендации музыки

1. Музыкальные треки
2. Тексты песен
3. Плейлисты

**Проблема:** нужно преобразовать к одному формату - матрице  
“объекты-признаки”

# Пример 1: текстовые признаки

- Dataset: 20news\_groups
- Электронные письма, разбитые по 20 темам (классам)
- Попробуем придумать классификатор, который различает две темы:
- **auto** и **politics.mideast**

# Извлечение текстовых признаков

- Пример письма 1:

From: carl\_f\_hoffman@cup.portal.com

Newsgroups: rec.autos

Subject: 1993 Infiniti G20

Message-ID: <78834@cup.portal.com>

Date: Mon, 5 Apr 93 07:36:47 PDT

Organization: The Portal System (TM)

Lines: 26

I am thinking about getting an Infiniti G20.

In consumer reports it is ranked high in many  
catagories including highest in reliability index for compact cars.  
Mitsubishi Galant was second followed by Honda Accord.)

# Извлечение текстовых признаков

- Пример письма 2:

From: Bob.Waldrop@f418.n104.z1.fidonet.org (Bob Waldrop)

Subject: Celebrate Liberty! 1993

Message-ID: <1993Apr5.201336.16132@dsd.es.com>

Followup-To:talk.politics.misc

Announcing... Announcing... Announcing... Announcing...

CELEBRATE LIBERTY!  
1993 LIBERTARIAN PARTY NATIONAL CONVENTION  
AND POLITICAL EXPO

THE MARRIOTT HOTEL AND THE SALT PALACE  
SALT LAKE CITY, UTAH

# Текстовые признаки: bag-of-words



the world of **TOTAL**

**all about the company**

Our energy exploration, production, and distribution operations span the globe, with activities in more than 100 countries.

At TOTAL, we draw our greatest strength from our fast-growing oil and gas reserves. Our strategic emphasis on natural gas provides a strong position in a rapidly expanding market.

Our expanding refining and marketing operations in Asia and the Mediterranean Basin complement already solid positions in Europe, Africa, and the U.S.

Our growing specialty chemicals sector adds balance and profit to the core energy business.

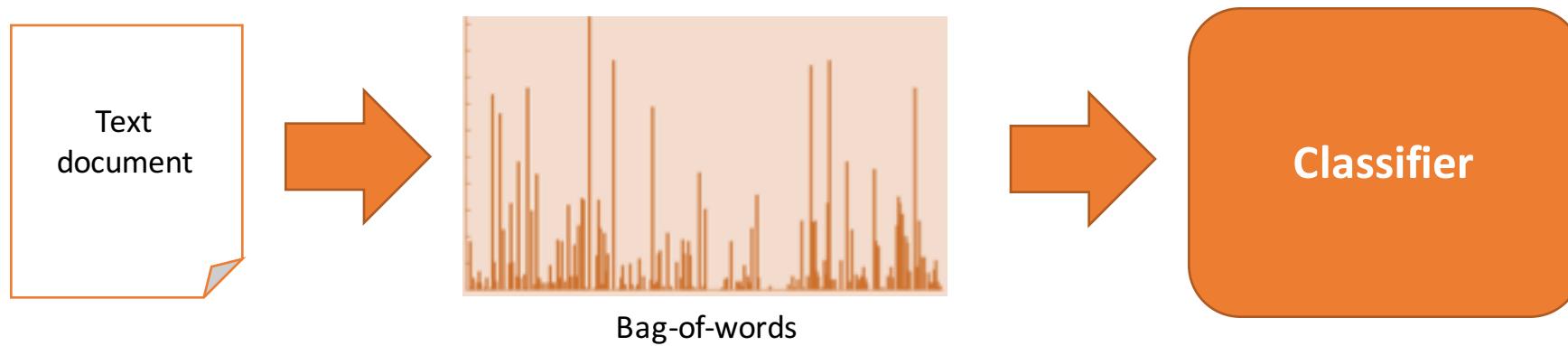
**All About The Company**

- Global Activities
- Corporate Structure
- TOTAL's Story
- Upstream Strategy
- Downstream Strategy
- Chemicals Strategy
- TOTAL Foundation
- Homepage



aardvark	0
about	2
all	2
Africa	1
apple	0
anxious	0
...	
gas	1
...	
oil	1
...	
Zaire	0

# Простой классификатор текстов



# Взвешивание частот слов в текстах

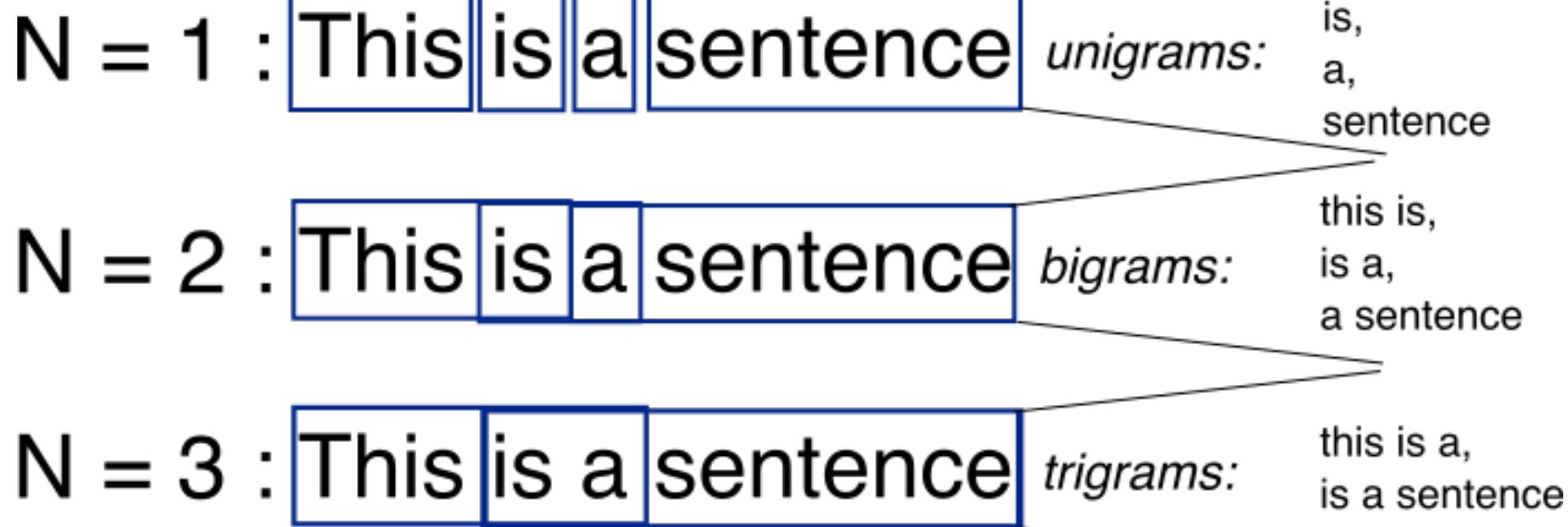
Term Frequency

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

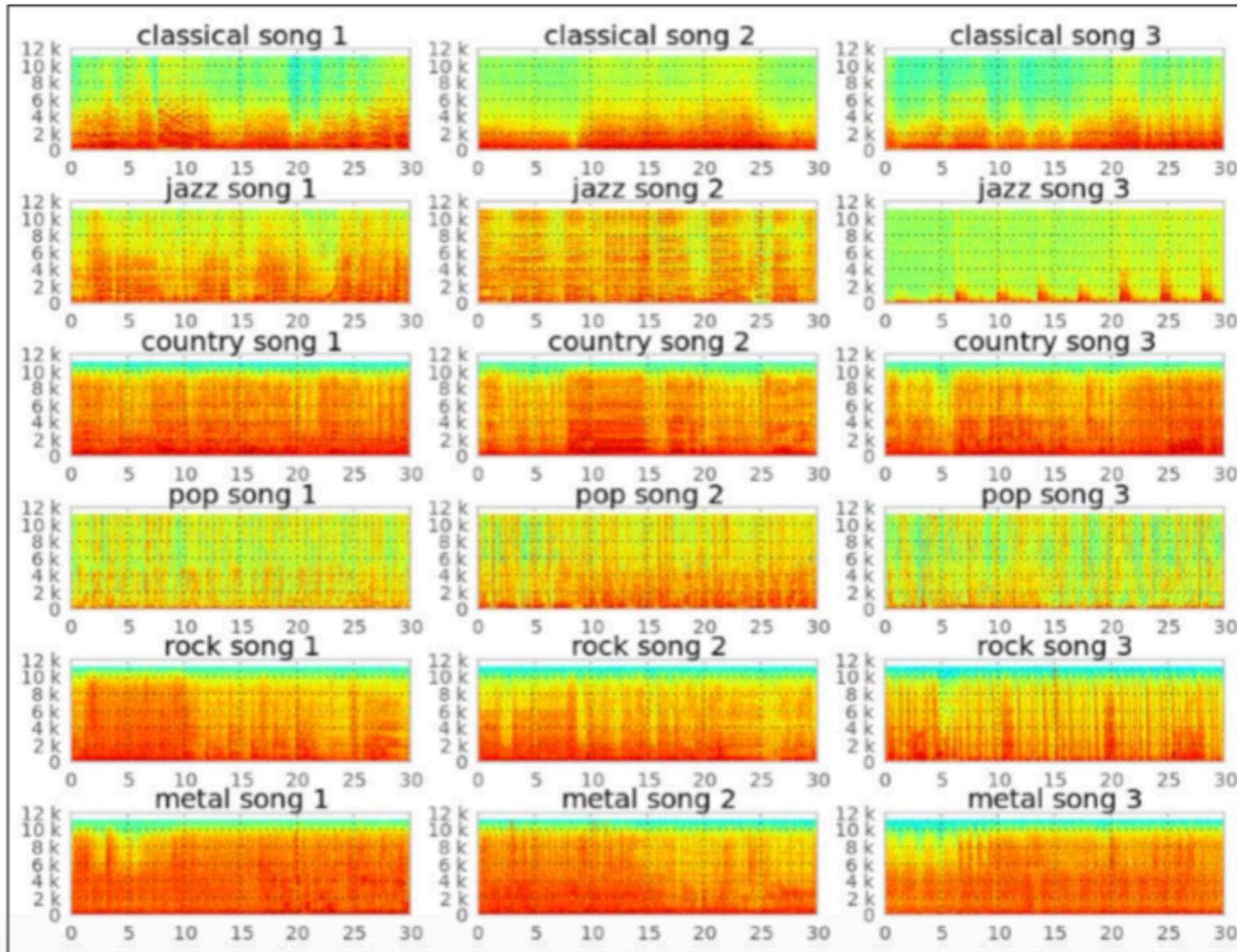
Inverse Document Frequency

$$idf_i = \log \frac{|D|}{|\{d : t_i \in d\}|}$$

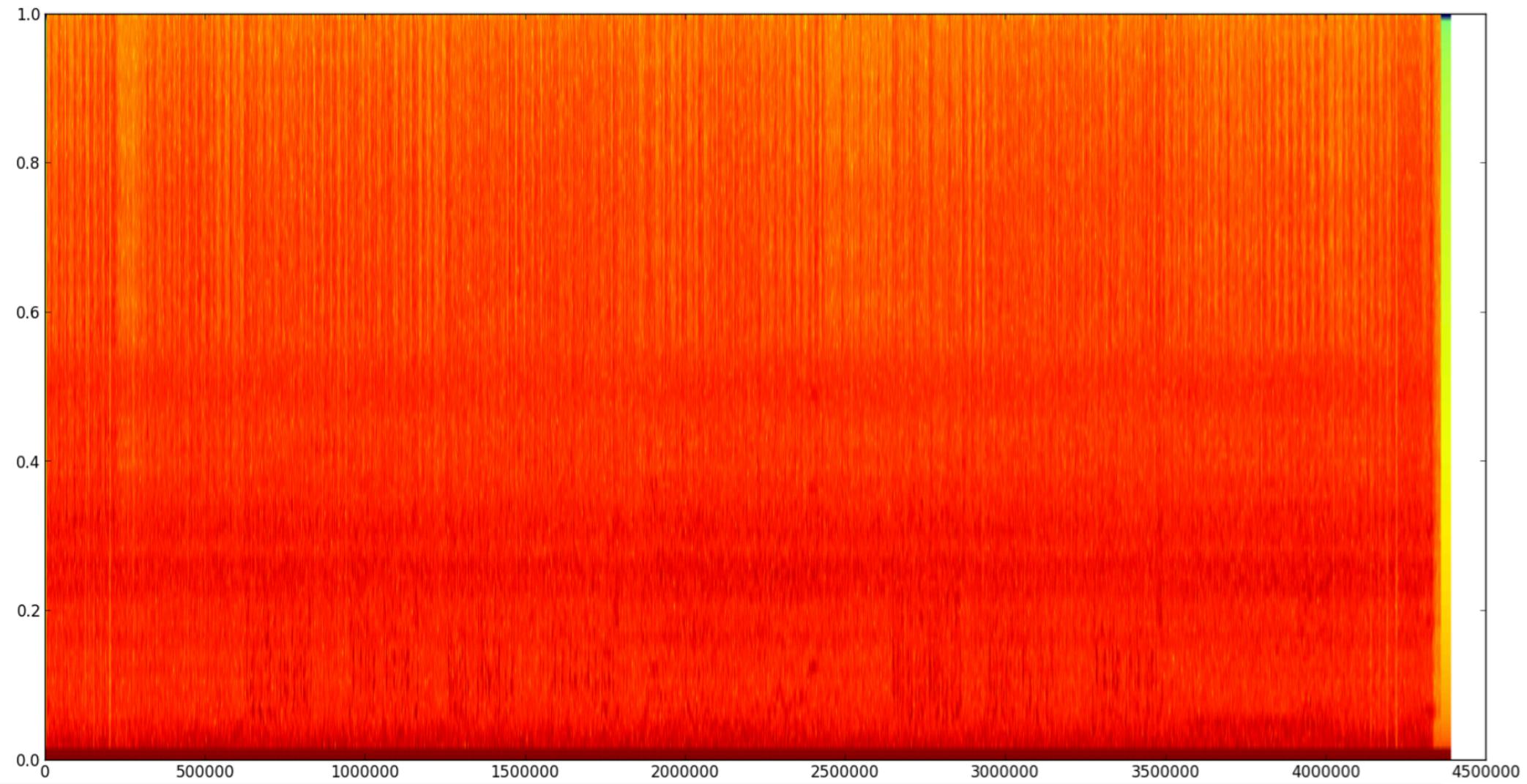
# Частоты N-грамм



## Пример 2: признаки аудиофайла

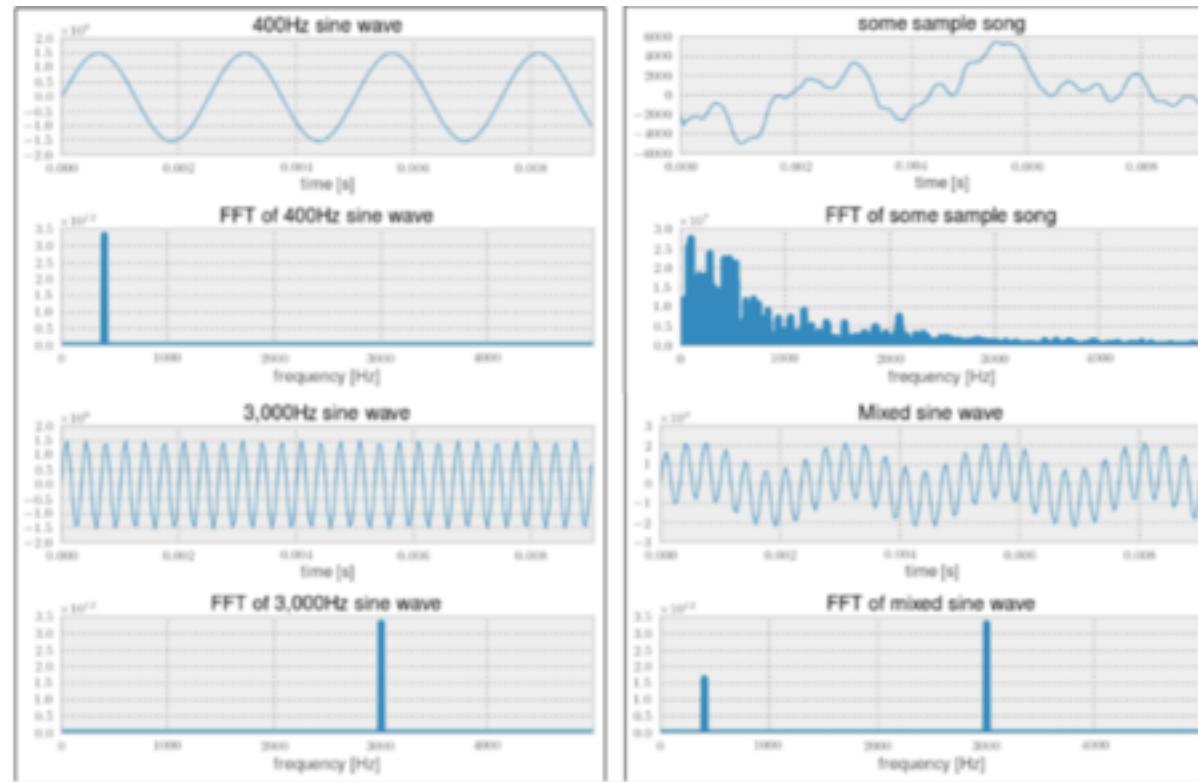


## Пример 2: признаки аудиофайла



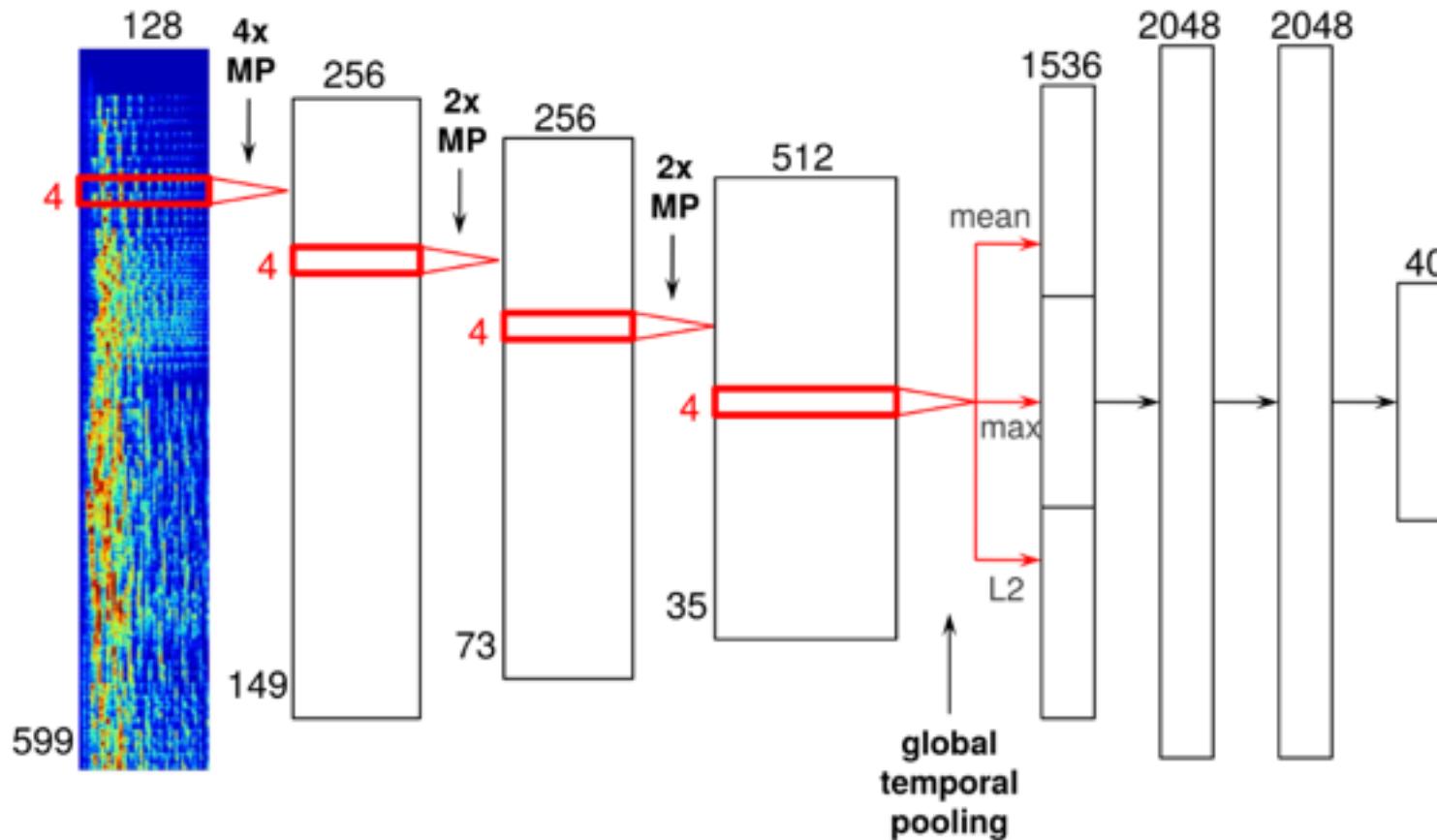
# Пример 2: признаки аудиофайла

MFCC - преобразование Фурье логарифма спектра



# Пример 2: признаки аудиофайла

Embeddings с помощью нейронных сетей:



# Пример 3: признаки изображения

Input image

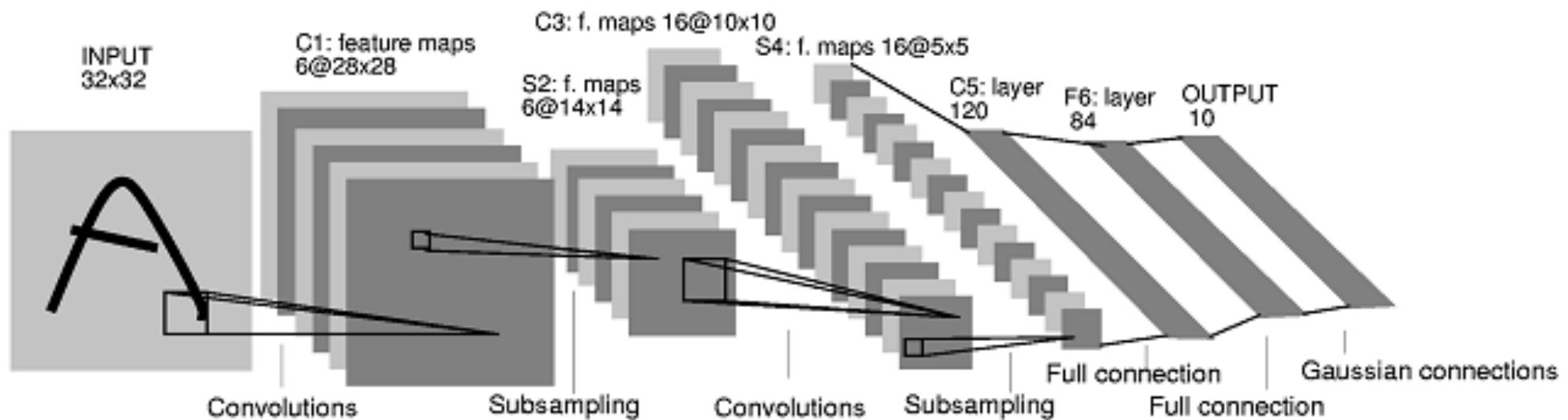


Histogram of Oriented Gradients



# Пример 3: признаки изображения

## Выходы слоев из нейросети



# Отбор признаков

1. Статистические методы
2. С помощью регуляризации L1
3. Жадный отбор
4. С помощью моделей

# Отбор признаков по статистическим критериям

**Пример:** критерий хи-квадрат позволяет отобрать лучшие бинарные признаки для каждого класса

	Значение признака 1	Значение признака 0
Объект принадлежит классу	A	B
Объект не принадлежит классу	C	D

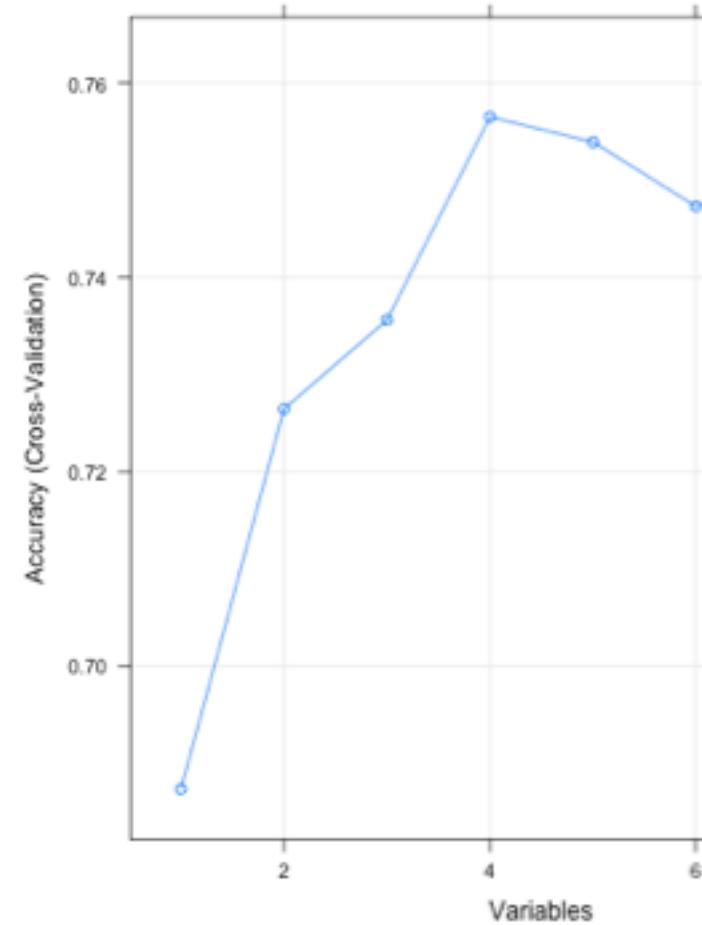
$$\chi^2 = \frac{N(AD - BC)^2}{(A + B)(A + C)(B + D)(C + D)}$$

# Отбор признаков с помощью l1-регуляризации

$$\sum_{i=1}^l L(M_i) + \gamma \sum_{k=1}^m |w_k| \rightarrow \min$$

# Жадный отбор признаков

Чередование добавления и удаления  
признаков

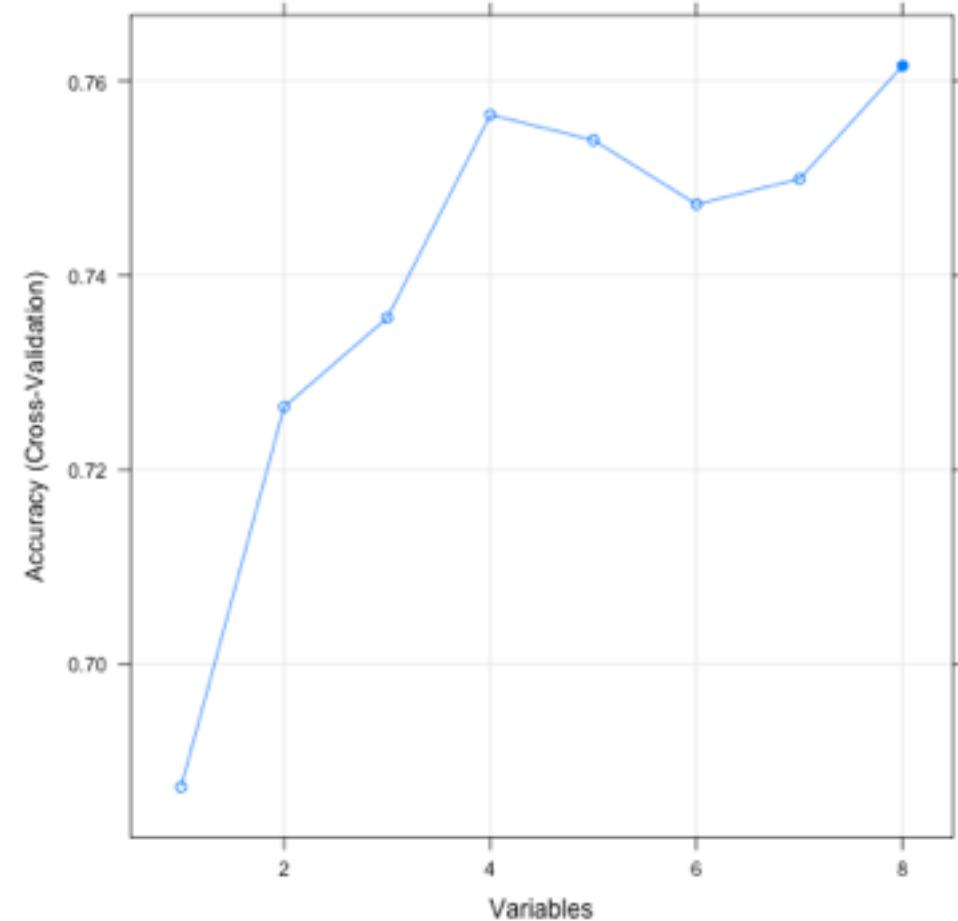


# Жадный отбор признаков

Чередование добавления и удаления  
признаков

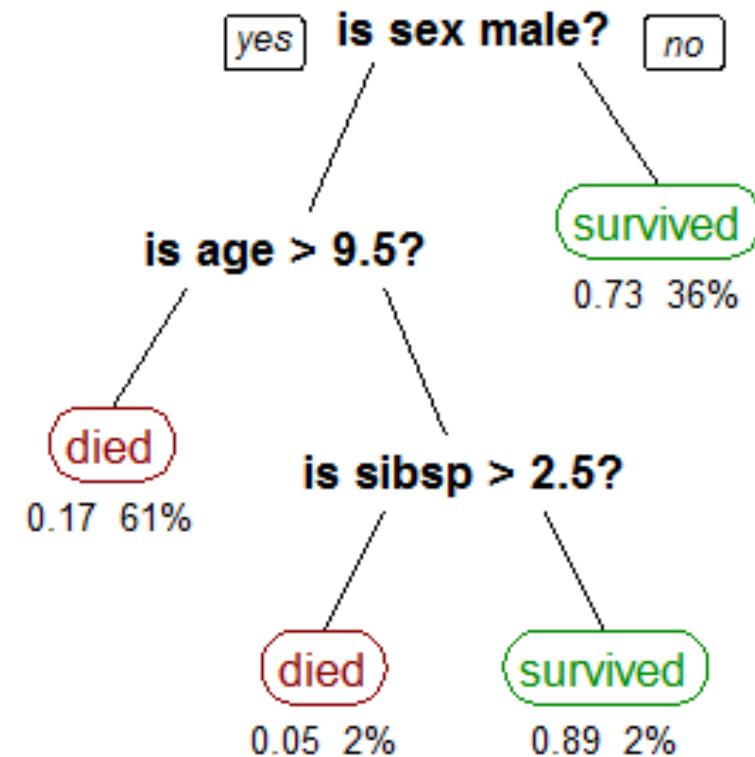
Этап добавления: добавляем лучшие  
признаки

Этап удаления: удаляем худшие  
признаки



# Отбор признаков с помощью моделей

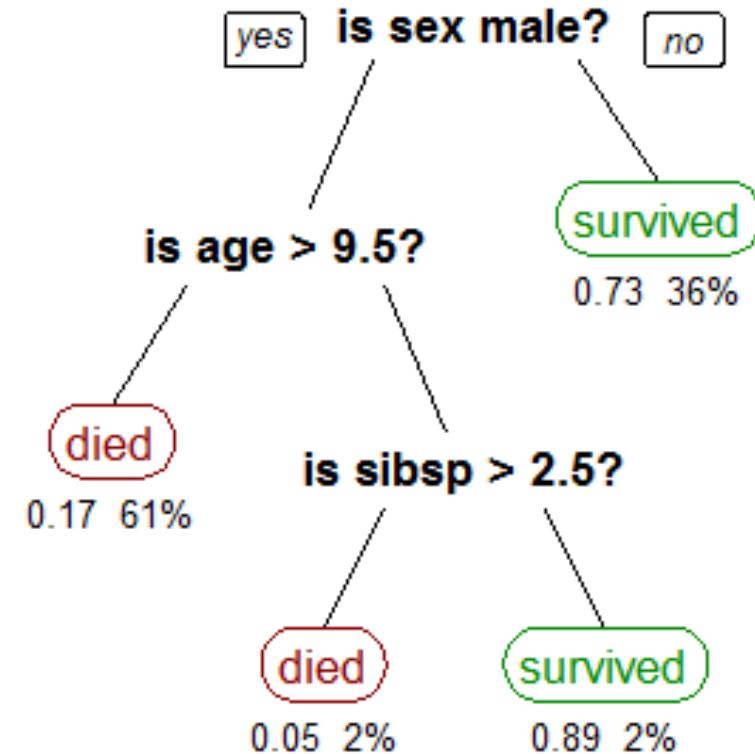
**Вопрос:** как можно оценивать важность признака в решающих деревьях?



# Отбор признаков с помощью моделей

**Вопрос:** как можно оценивать важность признака в решающих деревьях?

А в линейных моделях?



# Преобразование признаков

1. Задача понижения размерности
2. Метод главных компонент и SVD
3. Manifold learning

# Как выглядит обучающая выборка

Fisher's Iris Data

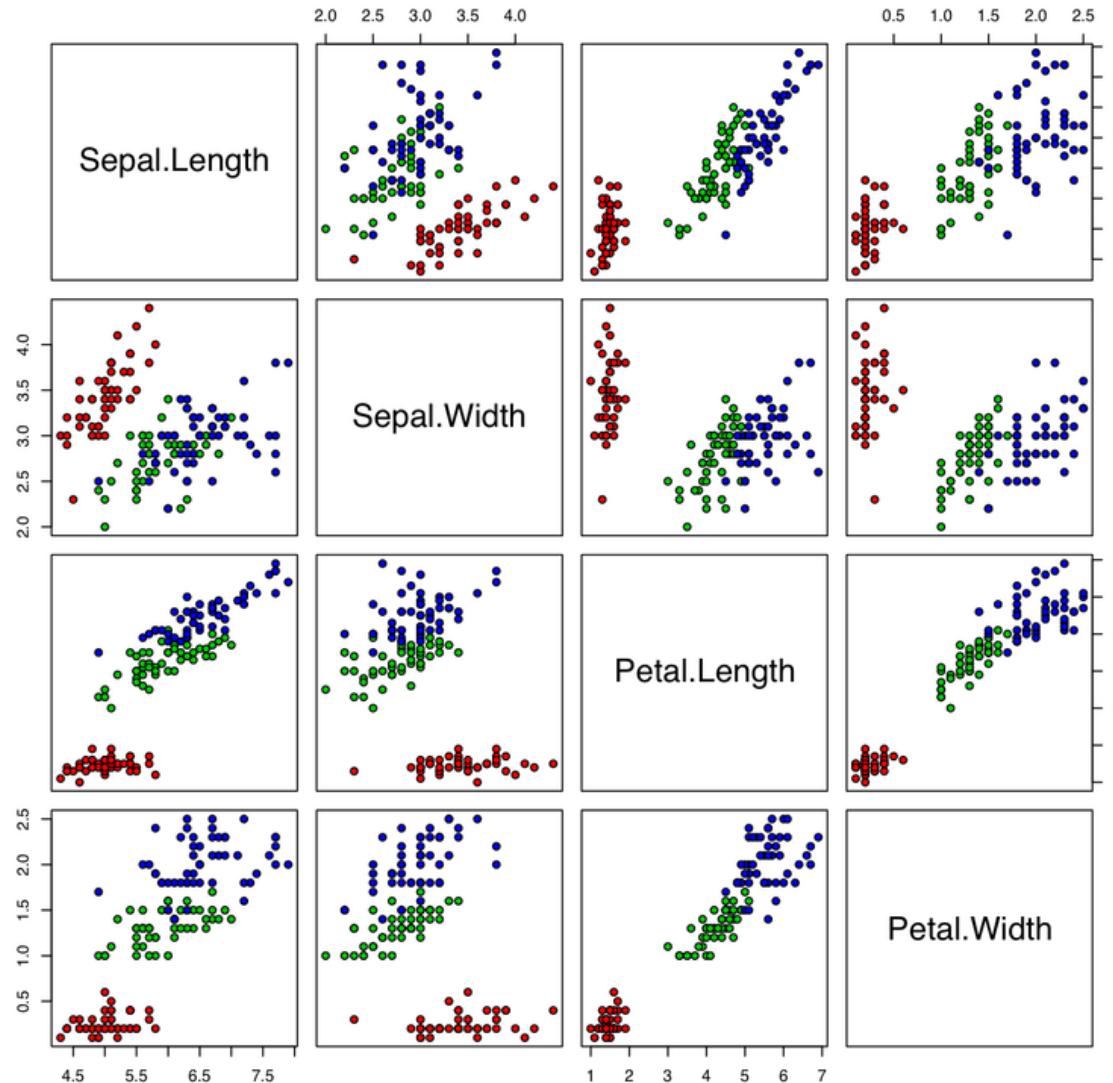
Sepal length	Sepal width	Petal length	Petal width	Species
5.0	2.0	3.5	1.0	<i>I. versicolor</i>
6.0	2.2	5.0	1.5	<i>I. virginica</i>
6.2	2.2	4.5	1.5	<i>I. versicolor</i>
6.0	2.2	4.0	1.0	<i>I. versicolor</i>
6.3	2.3	4.4	1.3	<i>I. versicolor</i>
5.5	2.3	4.0	1.3	<i>I. versicolor</i>
5.0	2.3	3.3	1.0	<i>I. versicolor</i>
4.5	2.3	1.3	0.3	<i>I. setosa</i>
5.5	2.4	3.8	1.1	<i>I. versicolor</i>
5.5	2.4	3.7	1.0	<i>I. versicolor</i>
4.9	2.4	3.3	1.0	<i>I. versicolor</i>
6.7	2.5	5.8	1.8	<i>I. virginica</i>
5.7	2.5	5.0	2.0	<i>I. virginica</i>
6.3	2.5	5.0	1.9	<i>I. virginica</i>
6.3	2.5	4.9	1.5	<i>I. versicolor</i>
4.9	2.5	4.5	1.7	<i>I. virginica</i>

# Что хотелось бы уметь

- Визуализировать обучающую выборку, когда признаков больше трёх
- Уменьшать количество признаков, переходя к новым, более информативным

# Визуализируем выборку

Iris Data (red=setosa,green=versicolor,blue=virginica)

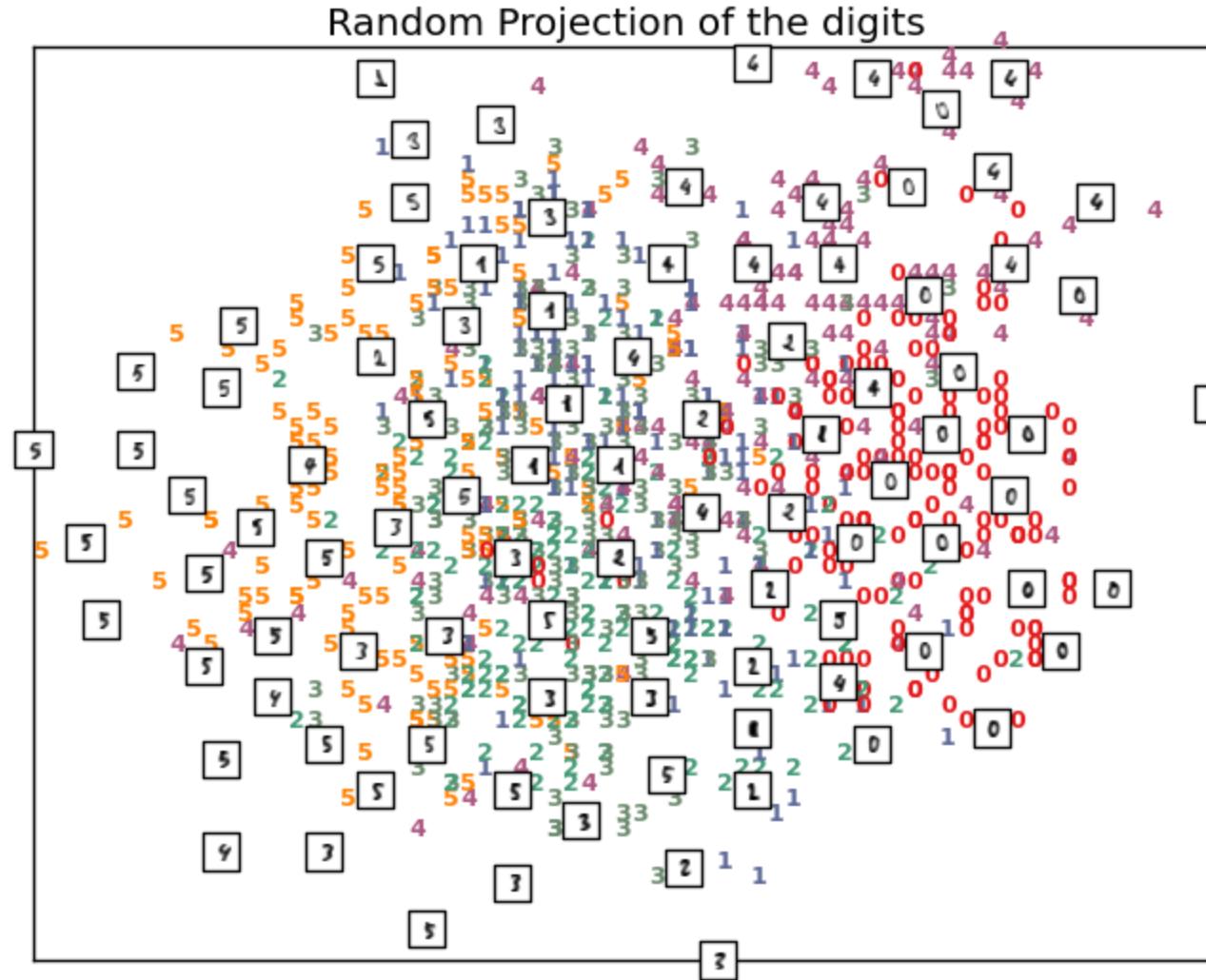


# Более сложный случай

Что делать, если признаков еще больше?

1	6	4	12	5	5	3	4	1	67	3	2	1	2	1	0	0	1	0	0	1	0	0	1	1
2	48	2	60	1	3	2	2	1	22	3	1	1	1	1	0	0	1	0	0	1	0	0	1	2
4	12	4	21	1	4	3	3	1	49	3	1	2	1	1	0	0	1	0	0	1	0	0	1	1
1	42	2	79	1	4	3	4	2	45	3	1	2	1	1	0	0	0	0	0	0	0	0	0	1
1	24	3	49	1	3	3	4	4	53	3	2	2	1	1	1	0	1	0	0	0	0	0	0	1
4	36	2	91	5	3	3	4	4	35	3	1	2	2	1	0	0	1	0	0	0	0	1	0	1
4	24	2	28	3	5	3	4	2	53	3	1	1	1	1	0	0	1	0	0	0	1	0	0	1
2	36	2	69	1	3	3	2	3	35	3	1	1	2	1	0	1	1	0	1	0	0	0	0	1
4	12	2	31	4	4	1	4	1	61	3	1	1	1	1	0	0	1	0	0	1	0	1	0	1
2	30	4	52	1	1	4	2	3	28	3	2	1	1	1	1	0	1	0	0	1	0	0	0	2
2	12	2	13	1	2	2	1	3	25	3	1	1	1	1	1	0	1	0	0	1	0	0	0	1
1	48	2	43	1	2	2	4	2	24	3	1	1	1	1	0	0	1	0	0	1	0	0	0	1
2	12	2	16	1	3	2	1	3	22	3	1	1	2	1	0	0	1	0	0	1	0	0	0	1
1	24	4	12	1	5	3	4	3	60	3	2	1	1	1	1	0	1	0	0	1	0	0	1	2
1	15	2	14	1	3	2	4	3	28	3	1	1	1	1	1	0	1	0	0	1	0	0	0	1
1	24	2	13	2	3	2	2	3	32	3	1	1	1	1	1	0	0	1	0	0	1	0	0	2
4	24	4	24	5	5	3	4	2	53	3	2	1	1	1	0	0	1	0	0	1	0	0	0	1
1	30	0	81	5	2	3	3	3	25	1	3	1	1	1	1	0	0	1	0	0	1	0	0	1
2	24	2	126	1	5	2	2	4	44	3	1	1	2	1	0	1	1	0	0	0	0	0	0	2
4	24	2	34	3	5	3	2	3	31	3	1	2	2	1	0	0	1	0	0	1	0	0	0	1
4	9	4	21	1	3	3	4	3	48	3	3	1	2	1	1	0	1	0	0	1	0	0	1	1
1	6	2	26	3	3	3	3	1	44	3	1	2	1	1	0	0	1	0	0	1	0	0	1	1
1	10	4	22	1	2	3	3	1	48	3	2	2	1	2	1	0	1	0	0	1	0	0	1	1
2	12	4	18	2	2	3	4	2	44	3	1	1	1	1	0	1	1	0	0	1	0	0	0	1
4	10	4	21	5	3	4	1	3	26	3	2	1	1	2	0	0	1	0	0	1	0	0	0	1
1	6	2	14	1	3	3	2	1	36	1	1	1	2	1	0	0	1	0	0	1	0	0	1	1
4	6	0	4	1	5	4	4	3	39	3	1	1	1	1	0	0	1	0	0	1	0	1	0	1
3	12	1	4	4	3	2	3	1	42	3	2	1	1	1	1	0	0	1	0	0	1	0	0	1
2	7	2	24	1	3	3	2	1	34	3	1	1	1	1	1	0	0	0	0	0	1	0	0	1
1	60	3	68	1	5	3	4	4	63	3	2	1	2	1	0	0	1	0	0	0	1	0	0	2
2	18	2	19	4	2	4	3	1	36	1	1	1	2	1	0	0	1	0	0	0	1	0	0	1
1	24	2	40	1	3	3	2	3	27	2	1	1	1	1	1	0	0	1	0	0	0	1	0	1

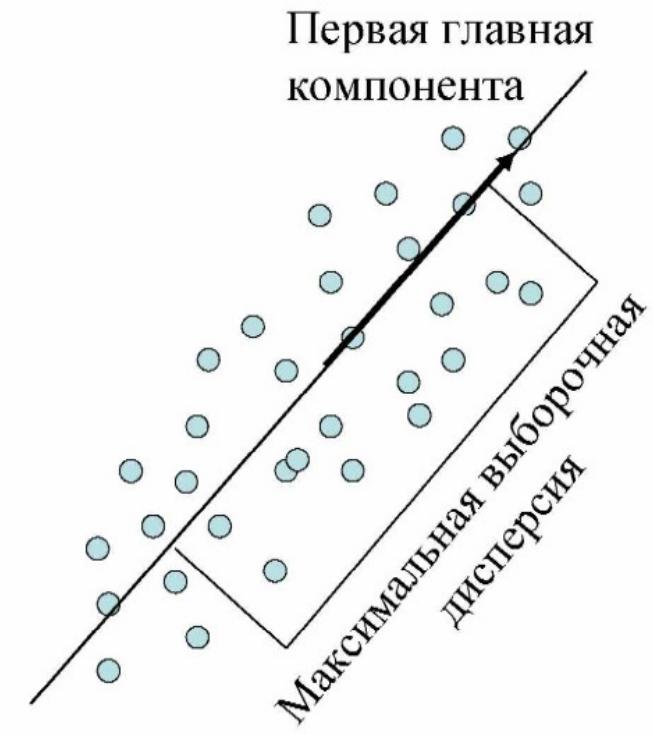
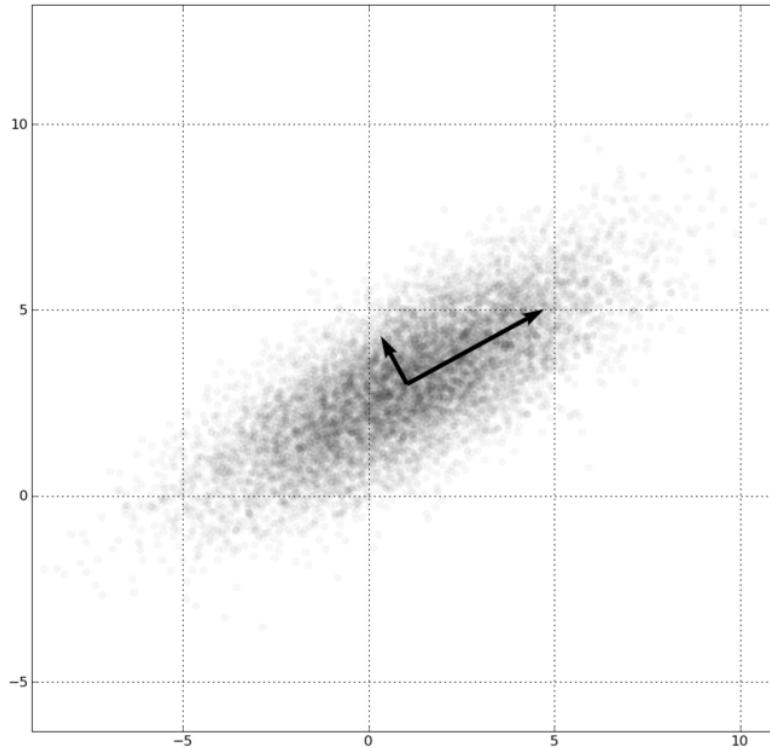
# Пример случайной проекции для рукописных цифр



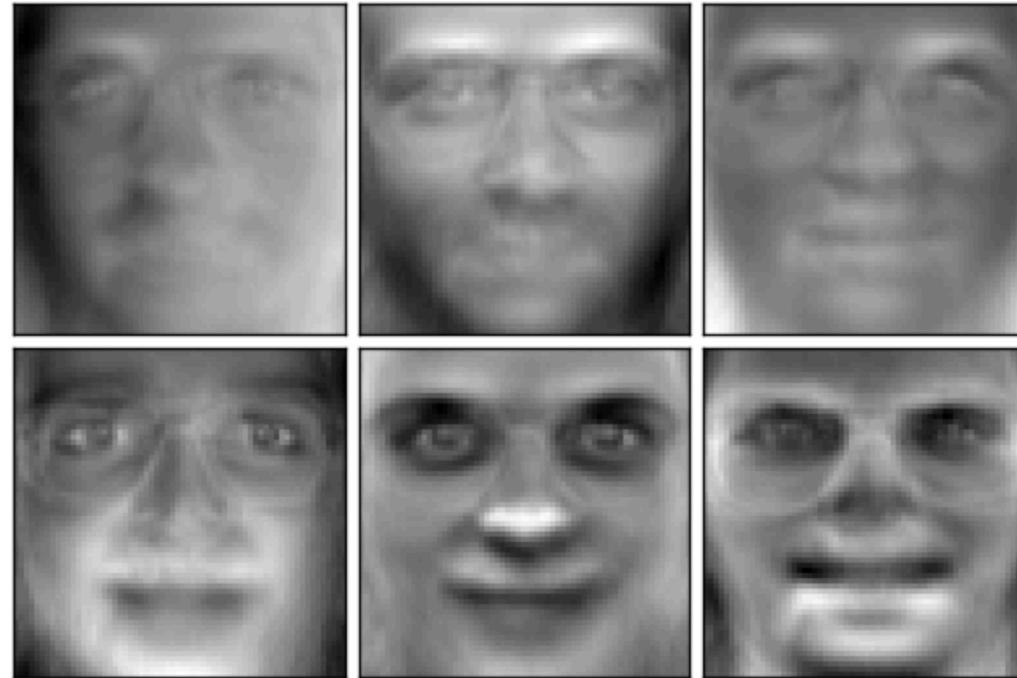
# Principal Component Analysis

- Идея: давайте выделять в пространстве признаков направления, вдоль которых разброс точек наибольший (они кажутся наиболее информативными)

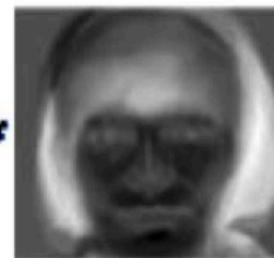
# PCA



# Пример: eigenfaces

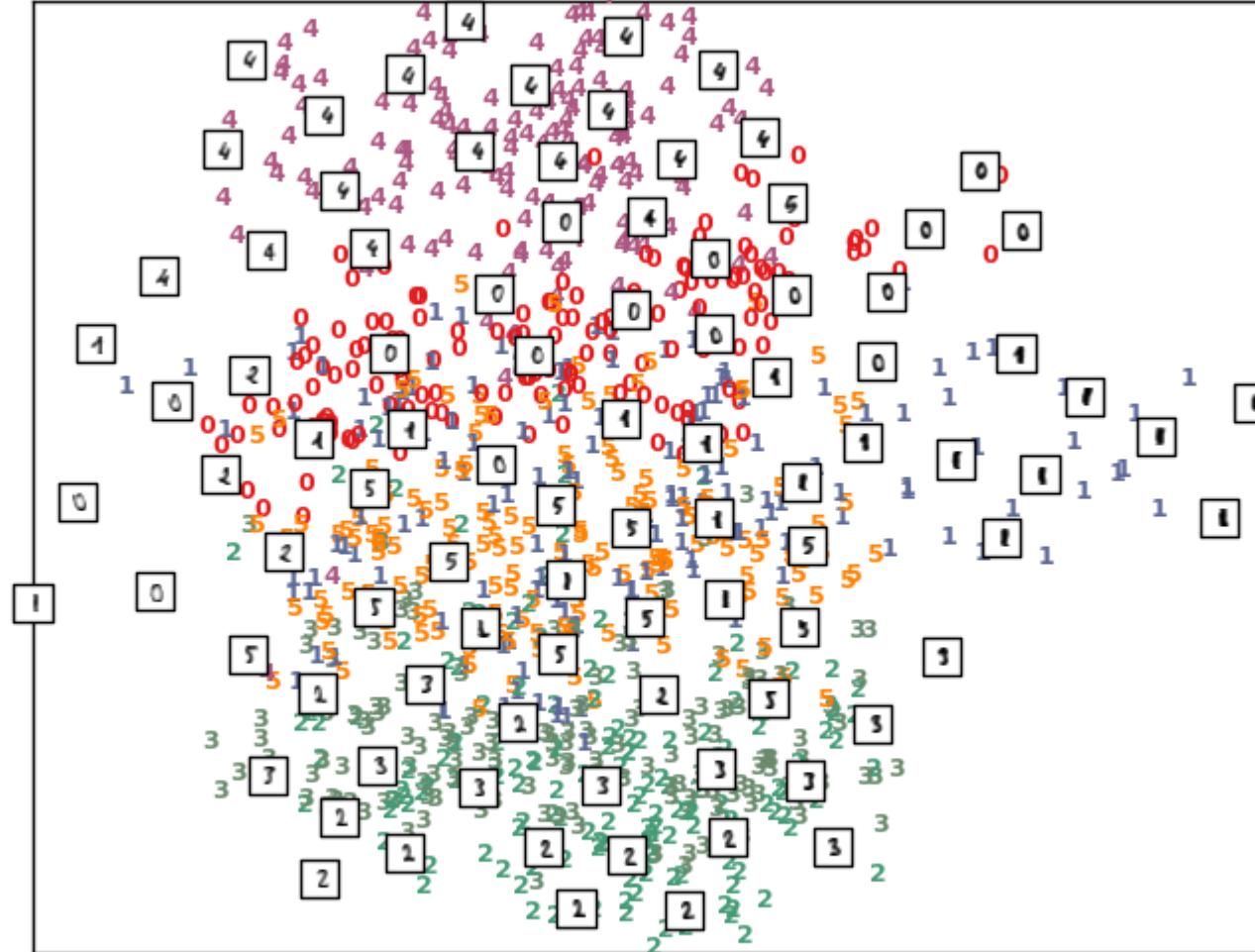


$$= \text{mean} + 0.9 * \text{eigenface}_1 - 0.2 * \text{eigenface}_2 + 0.4 * \text{eigenface}_3 + \dots$$



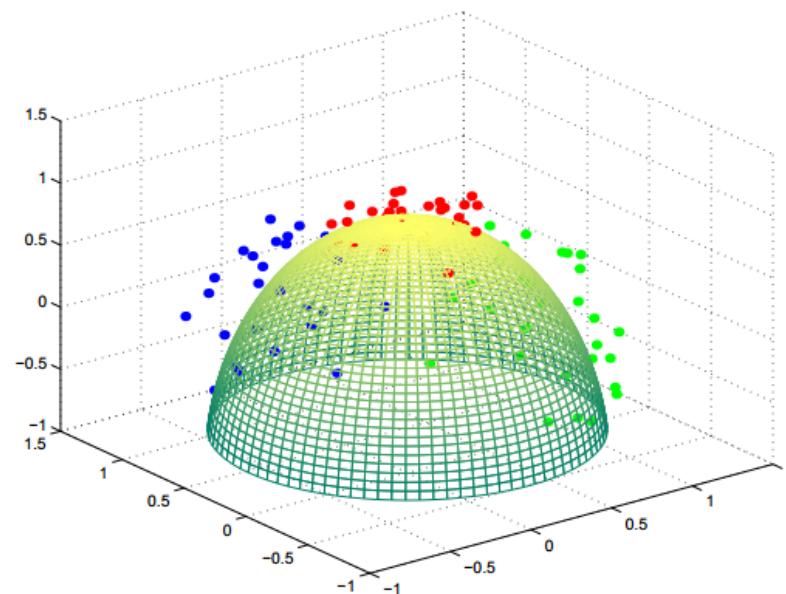
# Рукописные цифры: проекция на главные компоненты

Principal Components projection of the digits (time 0.02s)



# А что, если линейных преобразований признаков мало?

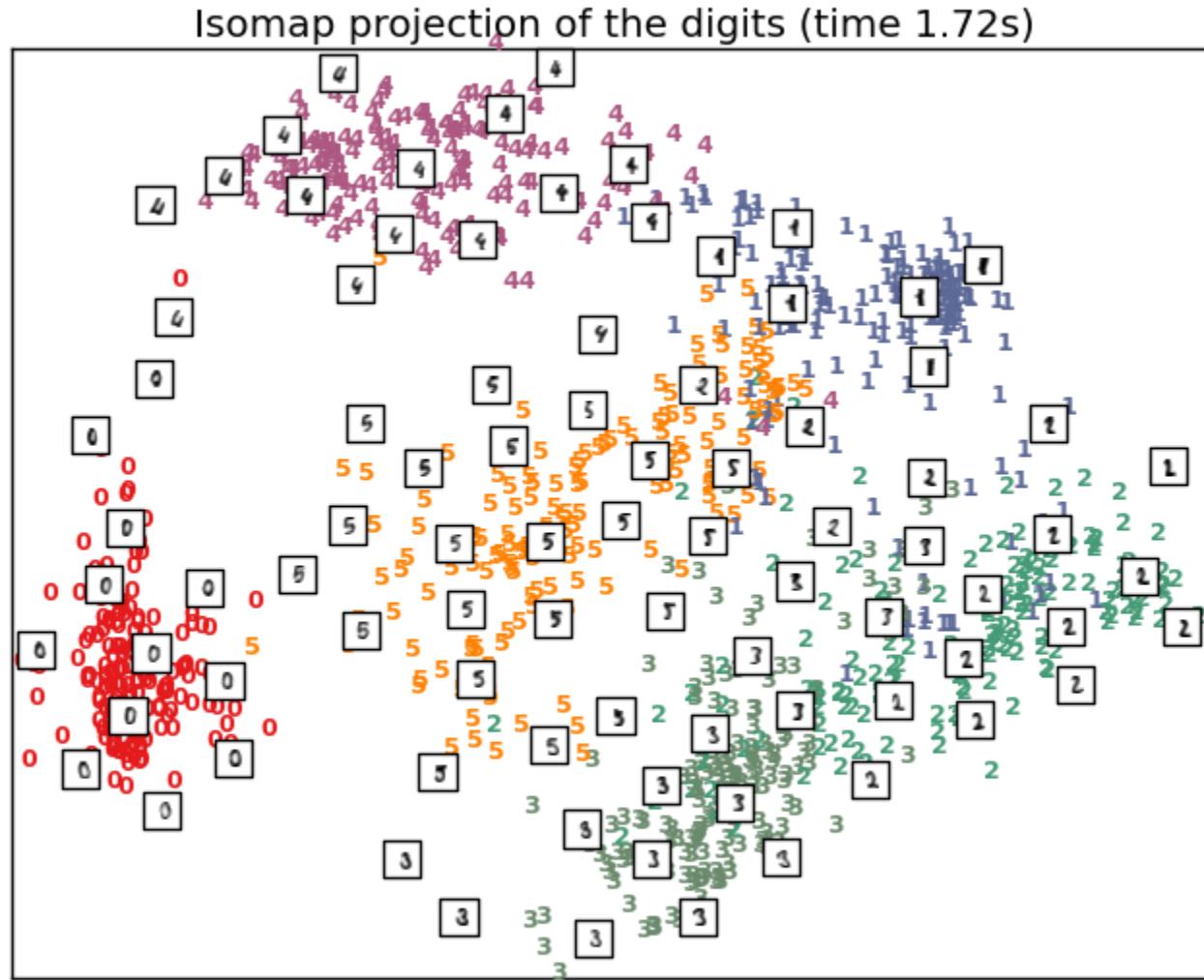
- Идея 1: объекты могут лежать в пространстве признаков на поверхности малой размерности.
- Идея 2: эта поверхность может быть нелинейной.



# Нелинейное преобразование признаков

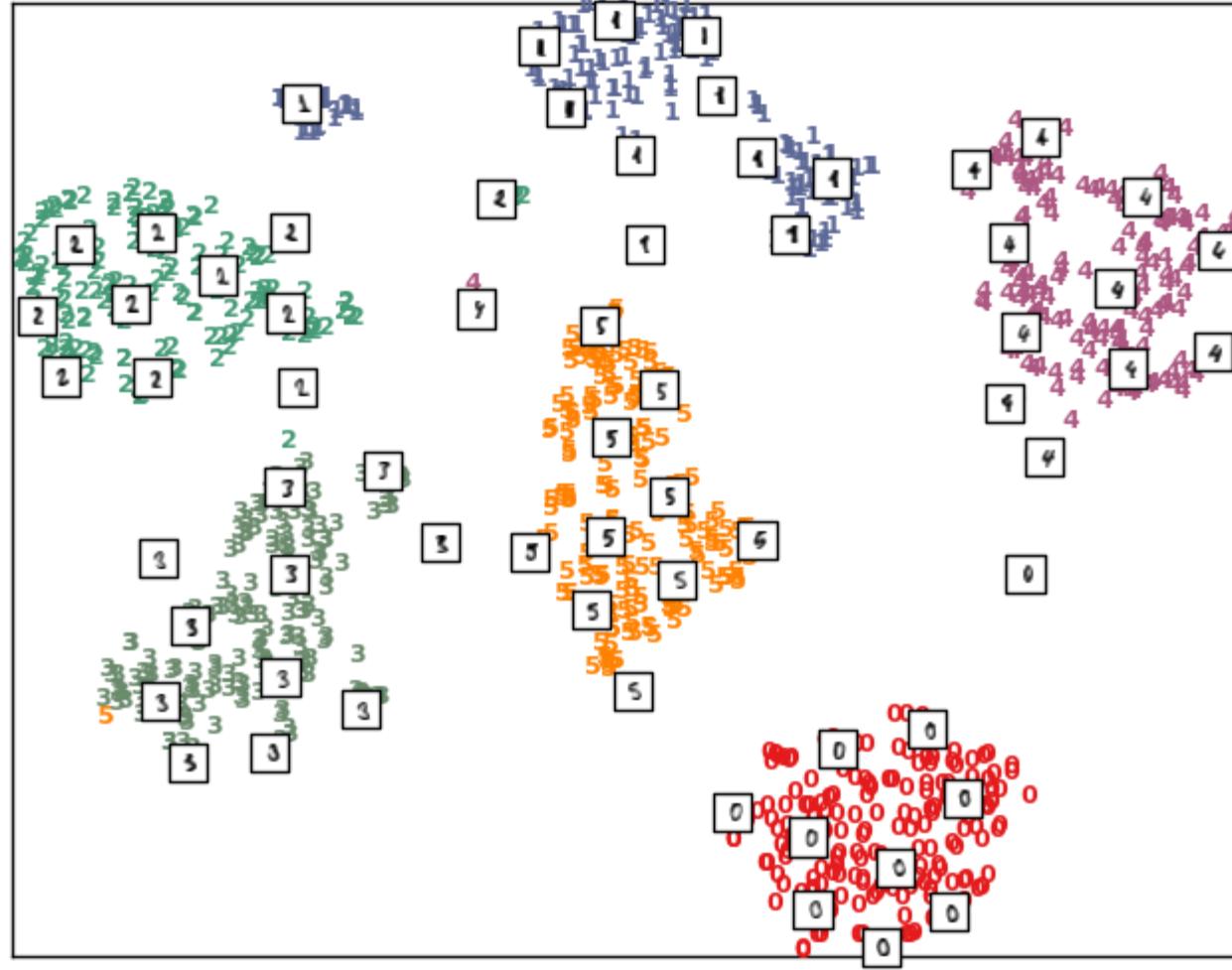
- SOM (Self-Organizing Maps) – самоорганизующиеся карты Кохонена. Не самый новый алгоритм, но идейно очень прост.
- Есть целое направление Manifold Learning

# Manifold learning: Isomap



# Manifold learning: t-SNE

t-SNE embedding of the digits (time 23.50s)



# Резюме

- Категориальные признаки
- Извлечение признаков из текстов, изображений и звука
- Отбор признаков
- Преобразование признаков