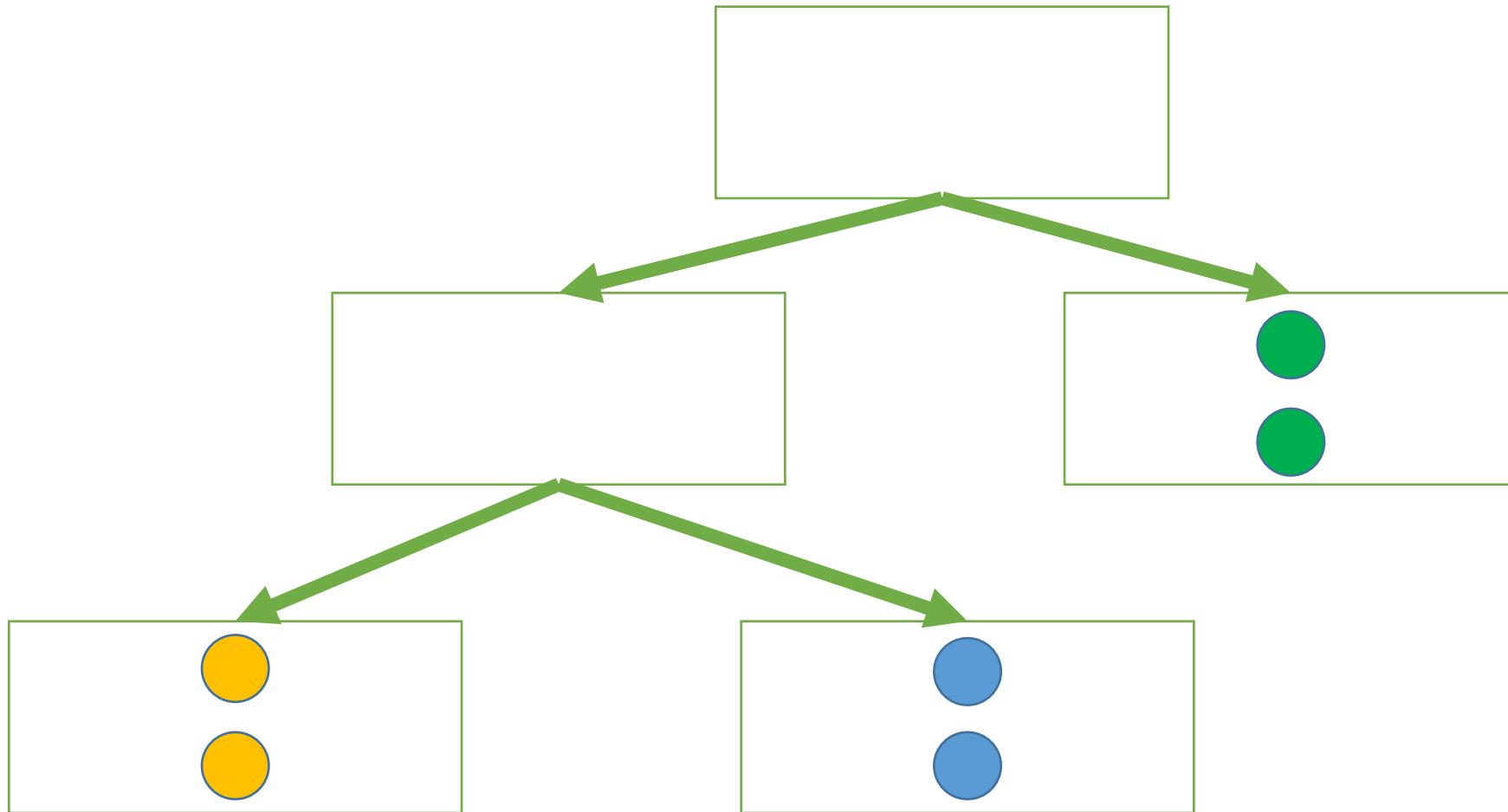


# Введение в анализ данных

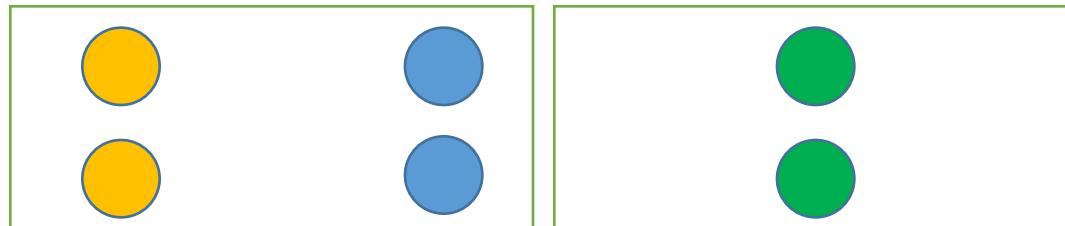
Лекция 4

Ансамбли и визуализация данных

# Жадное построение



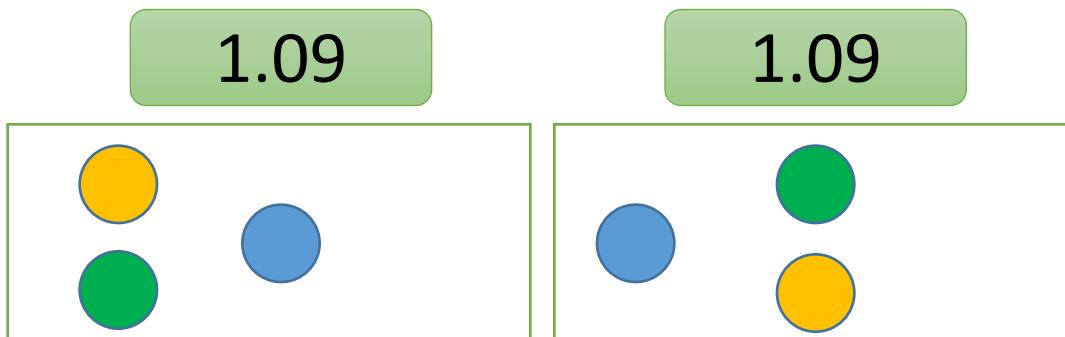
# Как сравнить разбиения?



0.693

0

- $(0.5, 0.5, 0)$  и  $(0, 0, 1)$
- $H = 0.693 + 0 = 0.693$

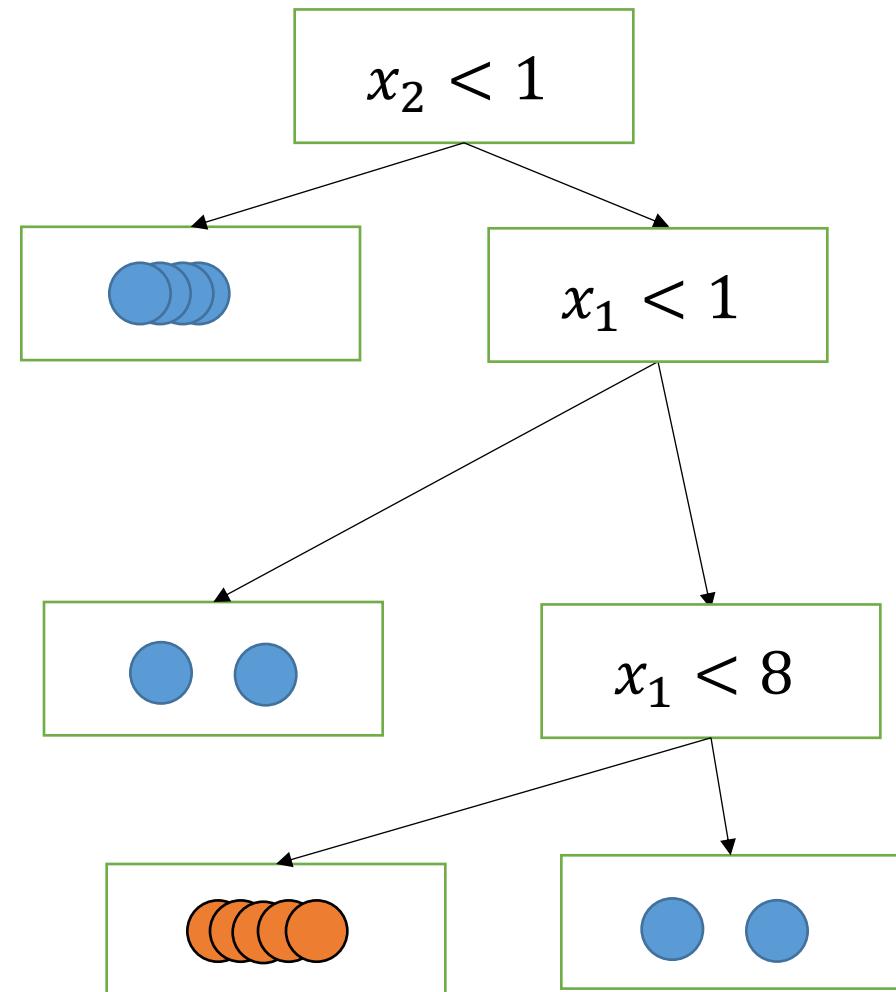
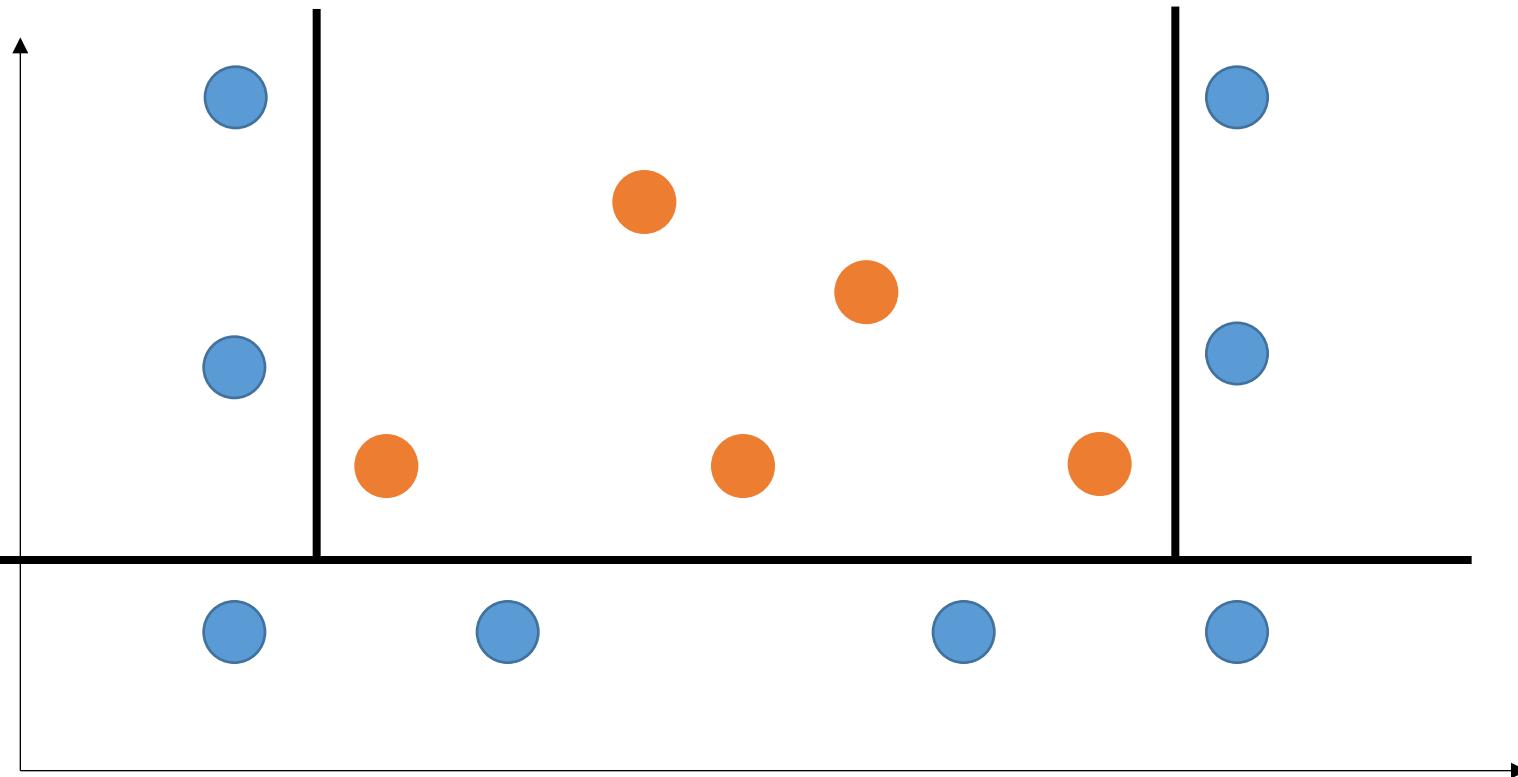


1.09

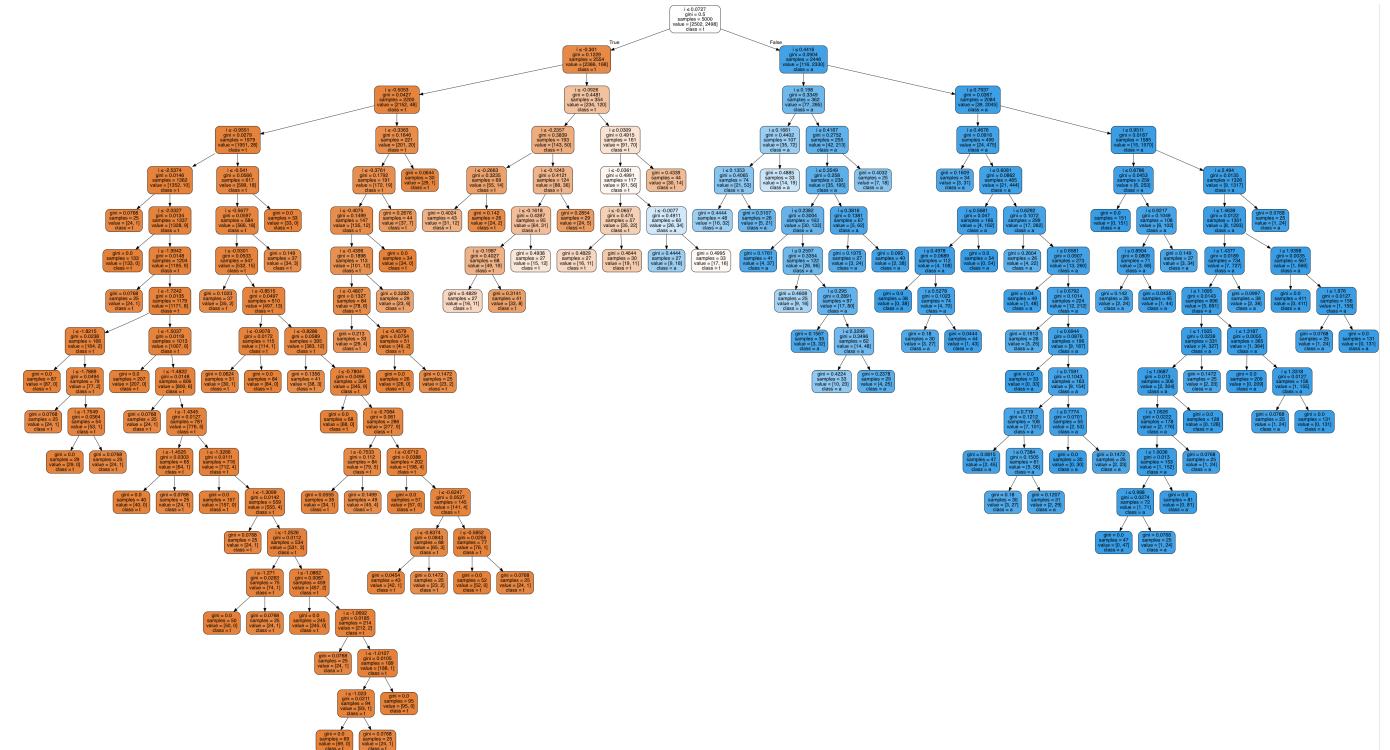
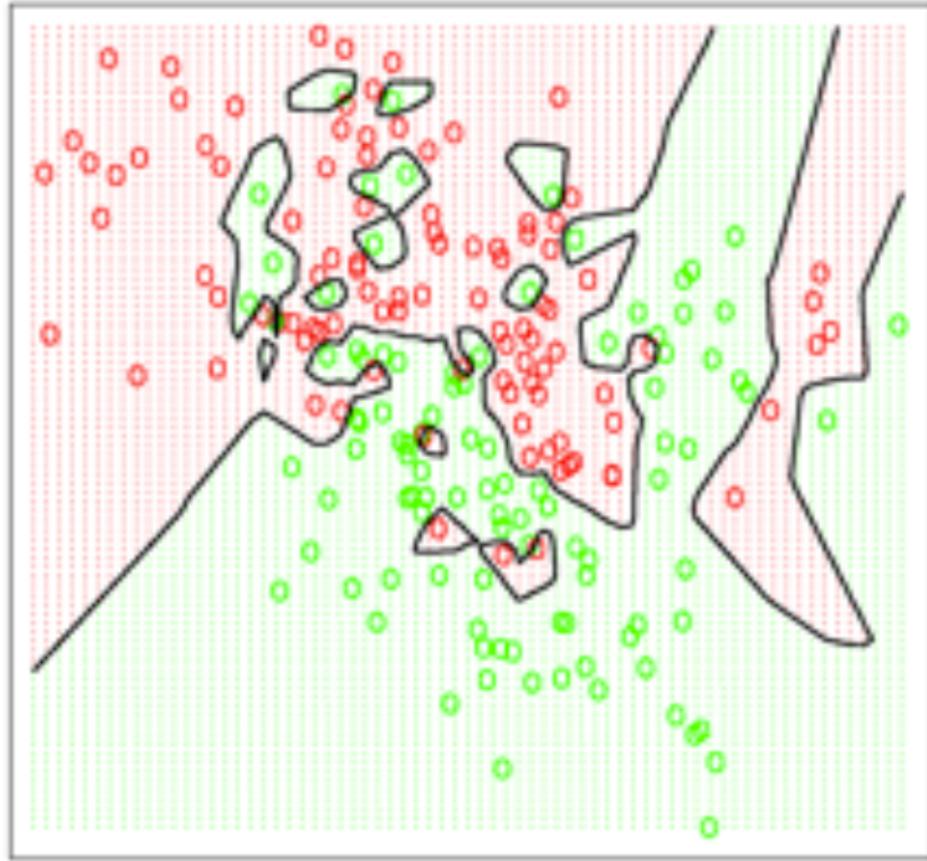
1.09

- $(0.33, 0.33, 0.33)$  и  $(0.33, 0.33, 0.33)$
- $H = 1.09 + 1.09 = 2.18$

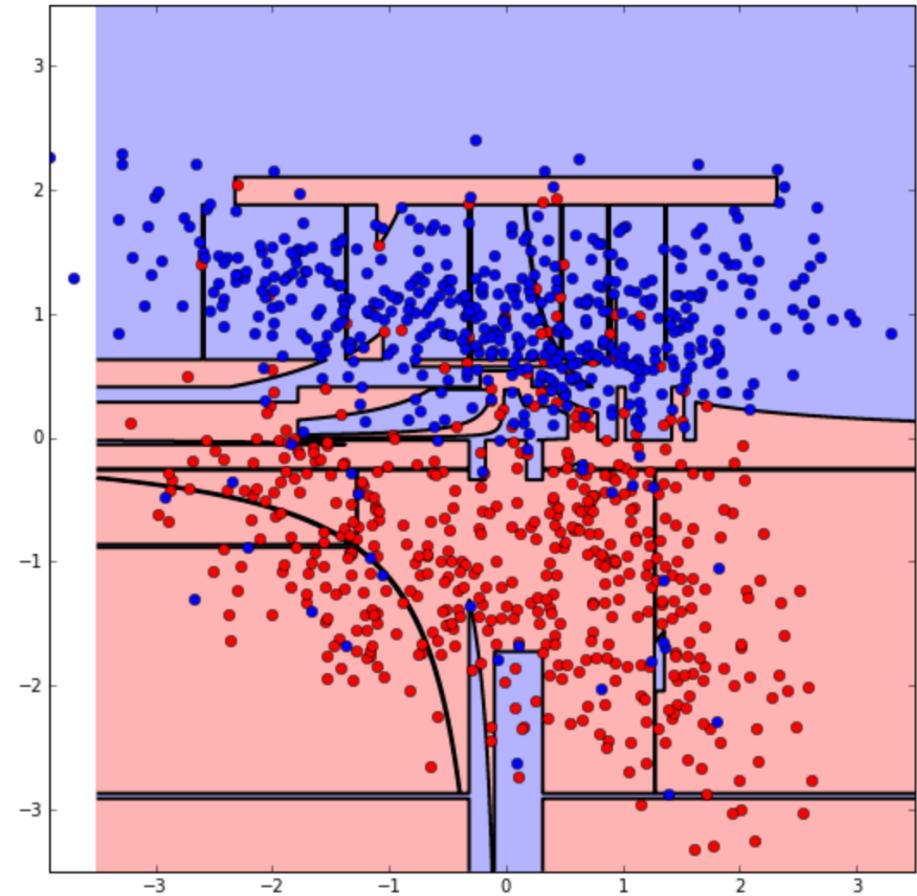
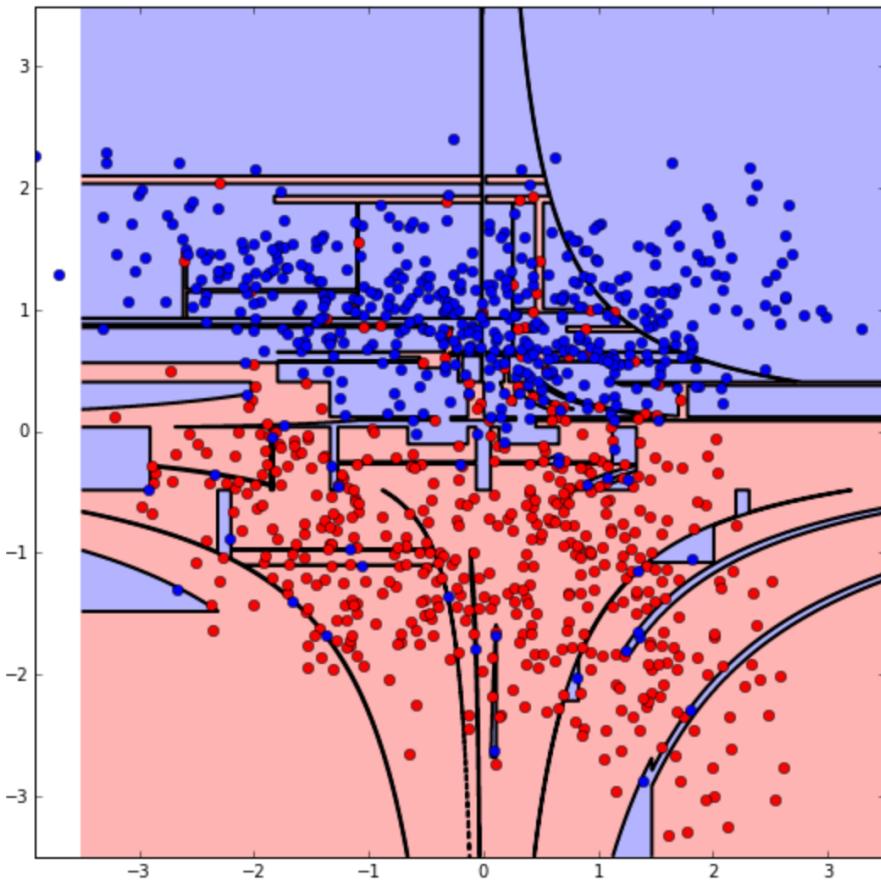
# Обучение деревьев



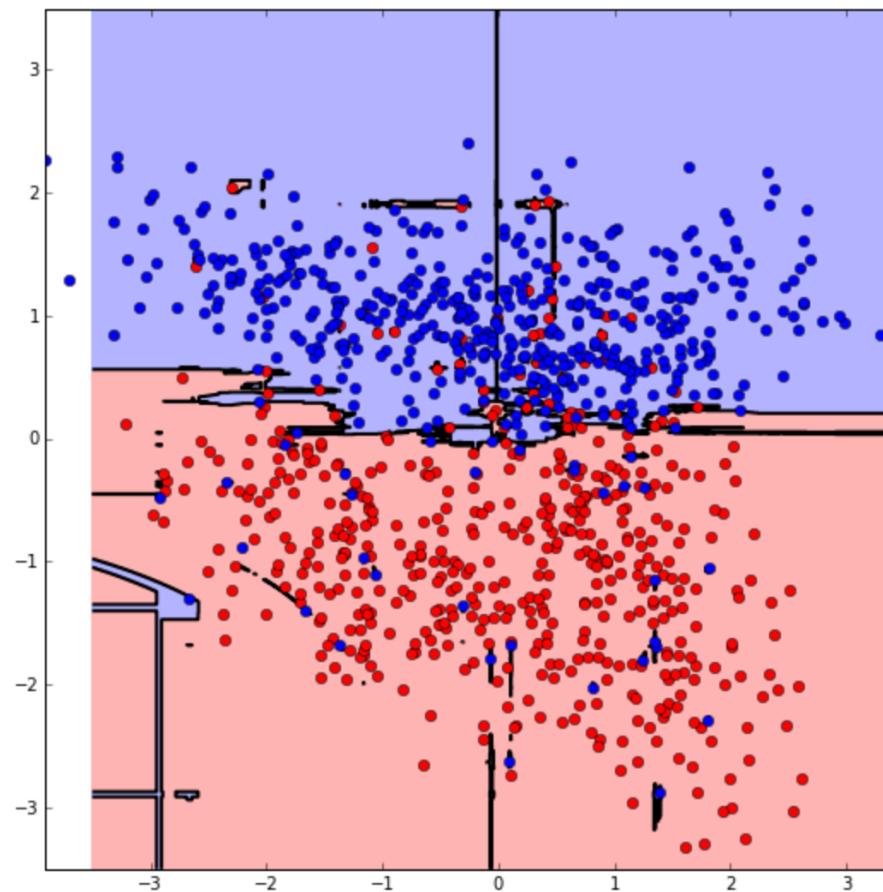
# Переобучение деревьев



# Неустойчивость деревьев



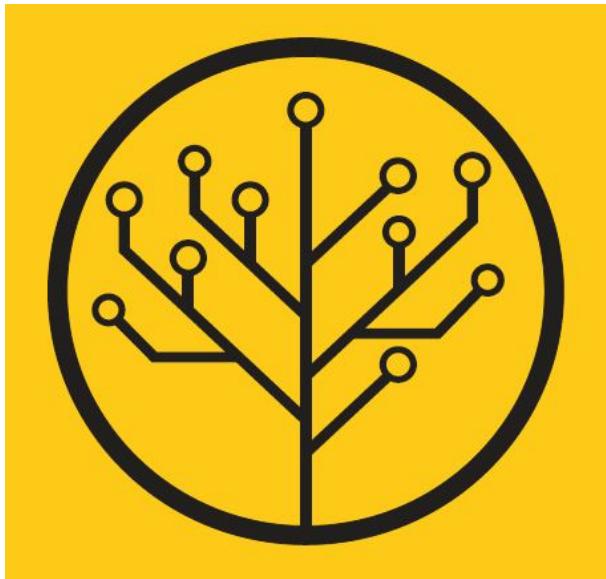
# Усреднение деревьев



# Композиции алгоритмов

# Majority vote

- Как выглядит логотип факультета компьютерных наук?



# Majority vote

- Как выглядит логотип факультета компьютерных наук?
- Какой из двух логотипов более старый?



# Majority vote

- Как выглядит логотип факультета компьютерных наук?
- Какой из двух логотипов более старый?
- Как выглядит корпус Вышки в Перми?



# Majority vote

- Дано:  $N$  базовых алгоритмов  $b_1(x), \dots, b_N(x)$
- Каждый хотя бы немного лучше случайного угадывания
- Композиция: класс, за который проголосовало больше всего базовых алгоритмов

$$a(x) = \arg \max_{y \in \mathbb{Y}} \sum_{n=1}^N [b_n(x) = y]$$

# Усреднение наблюдений

- Наблюдение: усреднение результатов повышает их точность
- Измерение артериального давления
- Измерение скорости света
- Усреднение соседних пикселей изображения

# Усреднение наблюдений

- Сколько лет факультету компьютерных наук?

# Усреднение наблюдений

- Сколько лет факультету компьютерных наук?
- Сколько метров в 1 сажени?

# Усреднение наблюдений

- Сколько лет факультету компьютерных наук?
- Сколько метров в 1 сажени?
- Сколько лет лектору?

# Усреднение наблюдений

- Сколько лет факультету компьютерных наук?
- Сколько метров в 1 сажени?
- Сколько лет лектору?
- Сколько всего стран в мире?

# Усреднение наблюдений

- Дано:  $N$  базовых алгоритмов  $b_1(x), \dots, b_N(x)$
- Каждый хотя бы немного лучше случайного угадывания
- Композиция:

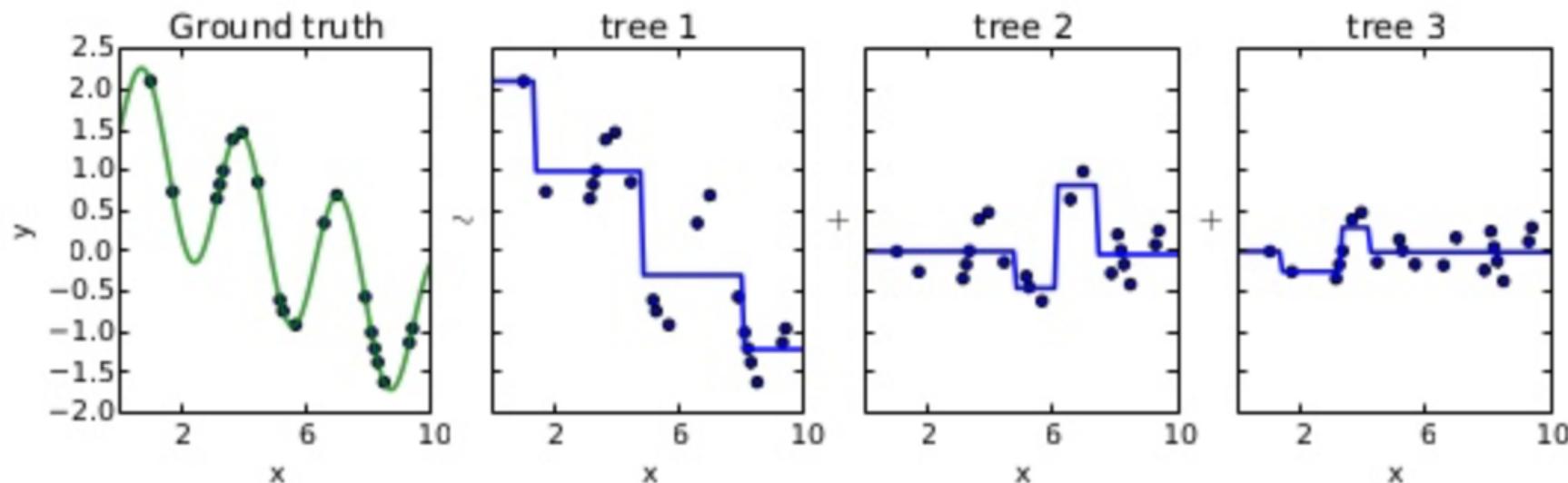
$$a(x) = \frac{1}{N} \sum_{n=1}^N b_n(x)$$

# Композиции алгоритмов

- Базовые алгоритмы:  $b_1(x), \dots, b_N(x)$
- Композиция:  $a(x)$
  
- Как по одной и той же выборке обучить  $N$  различных моделей?

# БУСТИНГ

- Каждый следующий алгоритм исправляет ошибки предыдущих
- Яркий пример: градиентный бустинг над решающими деревьями



# БЭГГИНГ

- Bagging (Bootstrap Aggregation)
- Базовые алгоритмы обучаются независимо
- Каждый обучается на подмножестве данных
- Усреднение ответов или выбор по большинству
- Яркий пример: случайный лес (random forest)

# БЭГГИНГ

Идея:

- Обучим много деревьев  $b_1(x), \dots, b_N(x)$
- Выберем ответ по большинству:

$$a(x) = \arg \max_{y \in \mathbb{Y}} \sum_{n=1}^N [b_n(x) = y]$$

# Пример

- Прогнозы деревьев:  $-1, -1, 1, -1, 1, -1$

$$a(x) = ?$$

# Пример

- Прогнозы деревьев:  $-1, -1, 1, -1, 1, -1$

$$a(x) = -1$$

# Рандомизация

- Как сделать деревья разными?
- Обучать по подвыборкам!

# Рандомизация

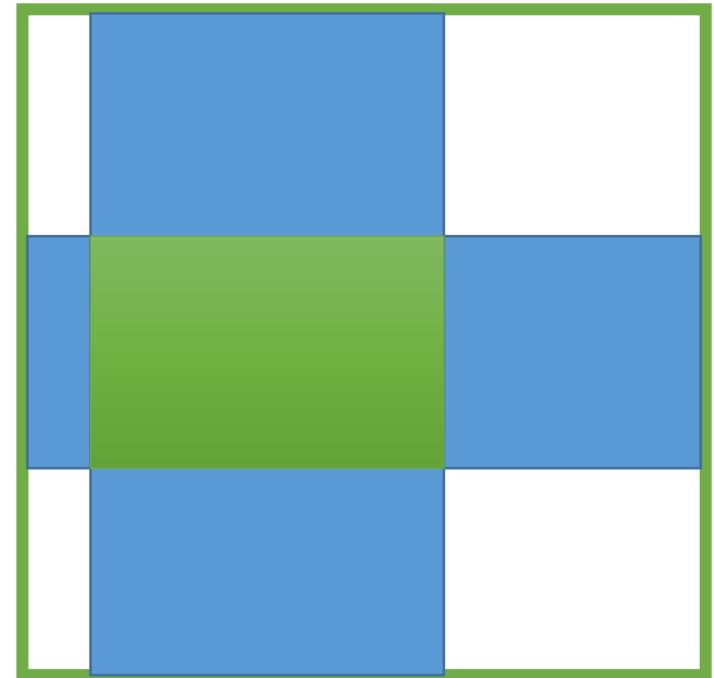
- Популярный подход: бутстрэп
- Выбираем из обучающей выборки  $\ell$  объектов с возвращением
- Пример:  $\{x_1, x_2, x_3, x_4\} \rightarrow \{x_1, x_2, x_2, x_4\}$
- Примерно  $0.632 * \ell$  различных объектов

# Рандомизация

- Другой подход: выбор случайного подмножества объектов
- Гиперпараметр: размер подмножества

# Виды рандомизации

- Бэггинг: обучаем на случайной подвыборке
- Метод случайных подпространств: обучаем на случайном подмножестве признаков
- Размер подвыборки/подмножества — гиперпараметр



# Рандомизация

- Этого недостаточно
- Как можно рандомизировать сам процесс построения дерева?

# Поиск разбиения

- Пусть в вершине  $m$  оказалась выборка  $X_m$
- $Q(X_m, j, t)$  — критерий ошибки условия  $[x^j \leq t]$
- Ищем лучшие параметры  $j$  и  $t$  перебором:

$$Q(X_m, j, t) \rightarrow \min_{j,t}$$

# Поиск разбиения

- Пусть в вершине  $m$  оказалась выборка  $X_m$
- $Q(X_m, j, t)$  — критерий ошибки условия  $[x^j \leq t]$
- Ищем лучшие параметры  $j$  и  $t$  перебором:

$$Q(X_m, j, t) \rightarrow \min_{j,t}$$

- Случайный лес: выбираем  $j$  из случайного подмножества признаков размера  $q$



# Случайный лес (Random forest)

1. Для  $n = 1, \dots, N$ :
2. Сгенерировать выборку  $\tilde{X}$  с помощью бутстрата
3. Построить решающее дерево  $b_n(x)$  по выборке  $\tilde{X}$
4. Дерево строится, пока в каждом листе не окажется не более  $n_{min}$  объектов
5. Оптимальное разбиение ищется среди  $q$  случайных признаков

# Случайный лес (Random forest)

1. Для  $n = 1, \dots, N$ :
2. Сгенерировать выборку  $\tilde{X}$  с помощью бутстрата
3. Построить решающее дерево  $b_n(x)$  по выборке  $\tilde{X}$
4. Дерево строится, пока в каждом листе не окажется не более  $n_{min}$  объектов
5. Оптимальное разбиение ищется среди  $q$  случайных признаков

Выбираются заново при каждом разбиении!

# Случайный лес

- Регрессия:

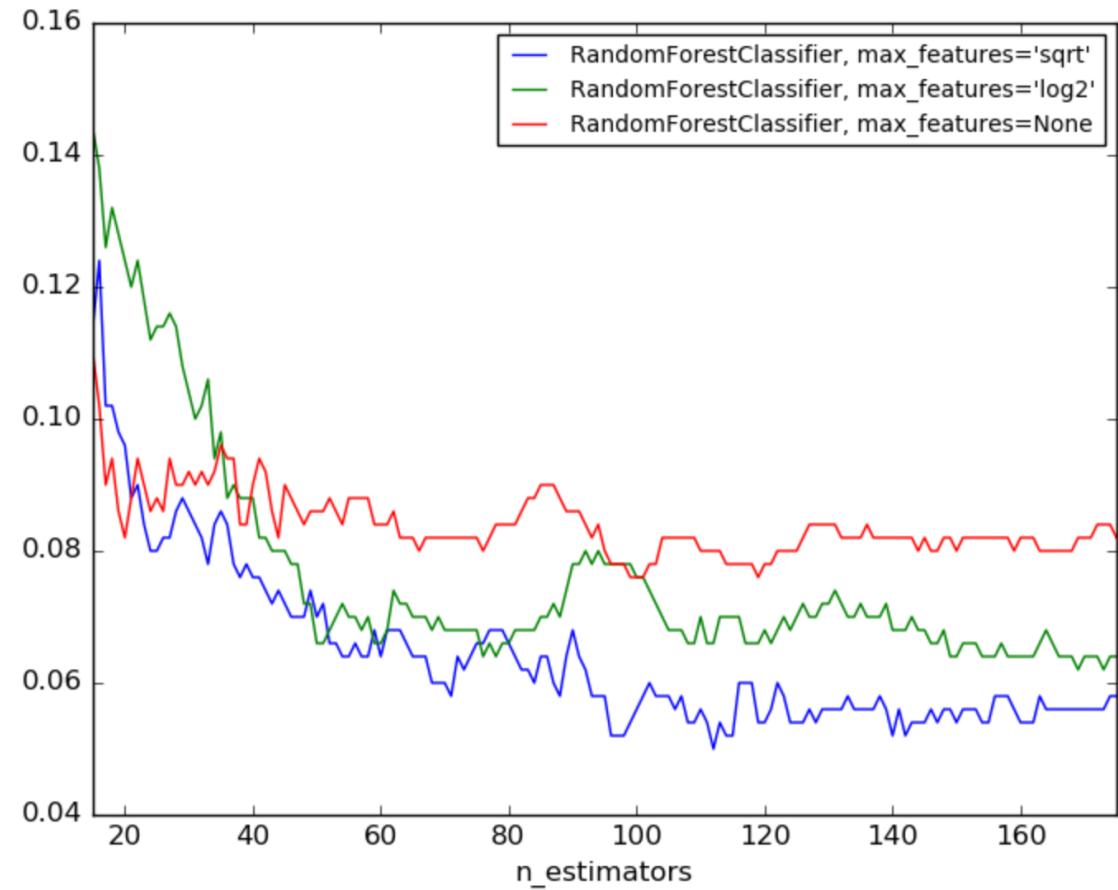
$$a(x) = \frac{1}{N} \sum_{n=1}^N b_n(x)$$

- Классификация:

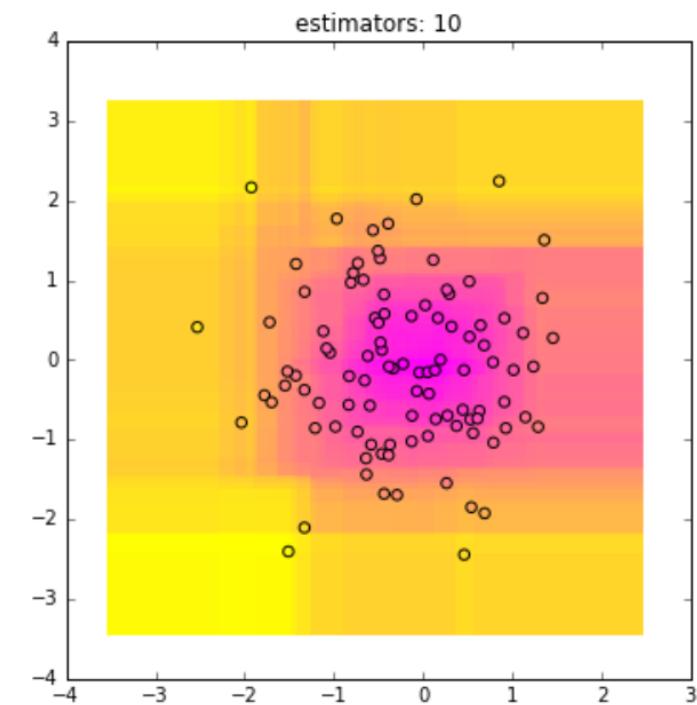
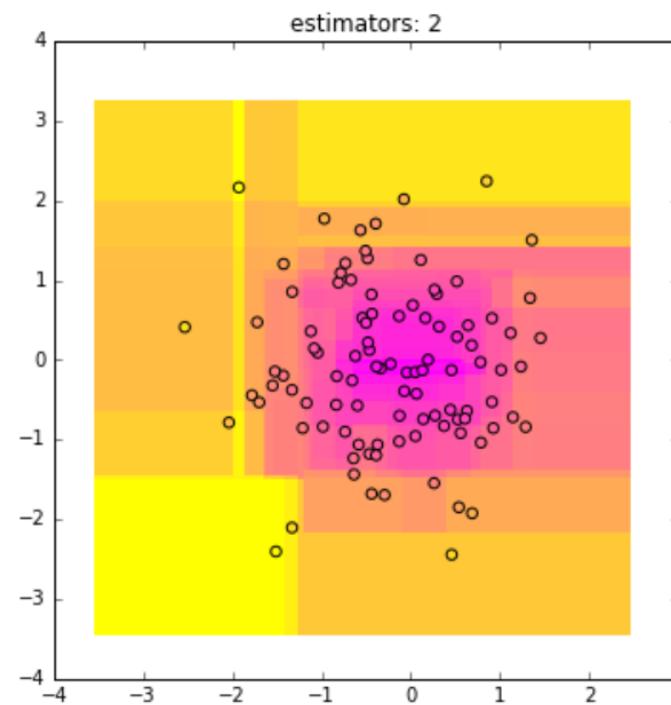
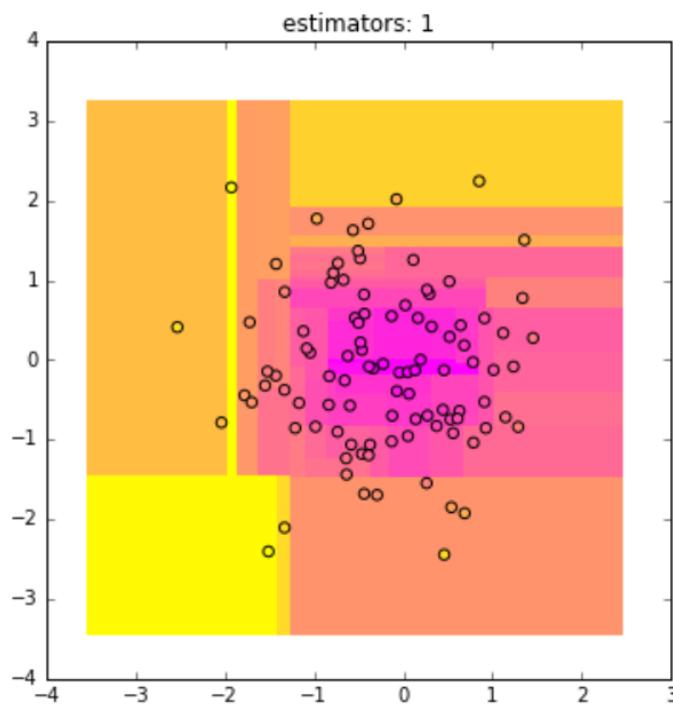
$$a(x) = \arg \max_{y \in \mathbb{Y}} \sum_{n=1}^N [b_n(x) = y]$$

# Ошибка на teste

- Ошибка сначала убывает, а затем остаётся примерно на одном уровне
- Случайный лес не переобучается при росте  $N$



# Случайный лес



# Важность признаков

Перестановочный метод:

- Проверяем важность  $j$ -го признака
- Перемешиваем соответствующий столбец в матрице «объекты-признаки» для тестовой выборки
- Измеряем качество модели
- Если оно слабо изменилось, то признак не очень важный

# Категориальные признаки

- Пример: предсказать, понравится ли пользователю фильм
- Объект: пара «пользователь-фильм»
- Признаки: ID пользователя, ID фильма, ID жанра, ID режиссёра, ID главных актёров, ID композитора, ...
- Как много фильмов снято за всю историю?

# Категориальные признаки

- Пример: предсказать, понравится ли пользователю фильм
- Объект: пара «пользователь-фильм»
- Признаки: ID пользователя, ID фильма, ID жанра, ID режиссёра, ID главных актёров, ID композитора, ...
- IMDB: >330 тысяч
- После бинарного кодирования получим миллионы признаков

# Анализ текстов

- Пример: предсказание популярности фильма по тексту его сценария
- Признаки: количество вхождений каждого слова из словаря
- Сколько слов в словаре?

# Анализ текстов

- Пример: предсказание популярности фильма по тексту его сценария
- Признаки: количество вхождений каждого слова из словаря
- Сотни тысяч признаков
- Если учитывать n-граммы, то десятки миллионов признаков

# Анализ данных ЭЭГ

- Энцефалограф: 64 датчика, частота сигнала 256 Гц
- Объект: результаты измерений для одного пациента
- За 5 секунд измерений:  $64 * 256 * 5 = 81\ 920$  признаков



# UCI Machine Learning Repository

UCI Machine Learning Repository  
Center for Machine Learning and Intelligent Systems

About Citation Policy Donate a Data Set Contact  
Search  
Repository Web Google\*

View ALL Data Sets

Browse Through: 349 Data Sets

	Name	Data Types	Default Task	Attribute Types	# Instances	# Attributes	Year
 URL Reputation	Multivariate, Time-Series	Classification	Integer, Real	2396130	3231961	2009	
 Gas sensor arrays in open sampling settings	Multivariate, Time-Series	Classification	Real	18000	1950000	2013	
 YouTube Multiview Video Games Dataset	Multivariate, Text	Classification, Clustering	Integer, Real	120000	1000000	2013	
 Gas sensor array exposed to turbulent gas mixtures	Multivariate, Time-Series	Classification, Regression	Real	180	150000	2014	
 ElectricityLoadDiagrams20112014	Time-Series	Regression, Clustering	Real	370	140256	2015	
 PEMS-SF	Multivariate, Time-Series	Classification	Real	440	138672	2011	
 Gas sensor array under flow modulation	Multivariate, Time-Series	Classification, Regression	Real	58	120432	2014	
 Bag of Words	Text	Clustering	Integer	8000000	100000	2008	

Table View List View

# Задача понижения размерности

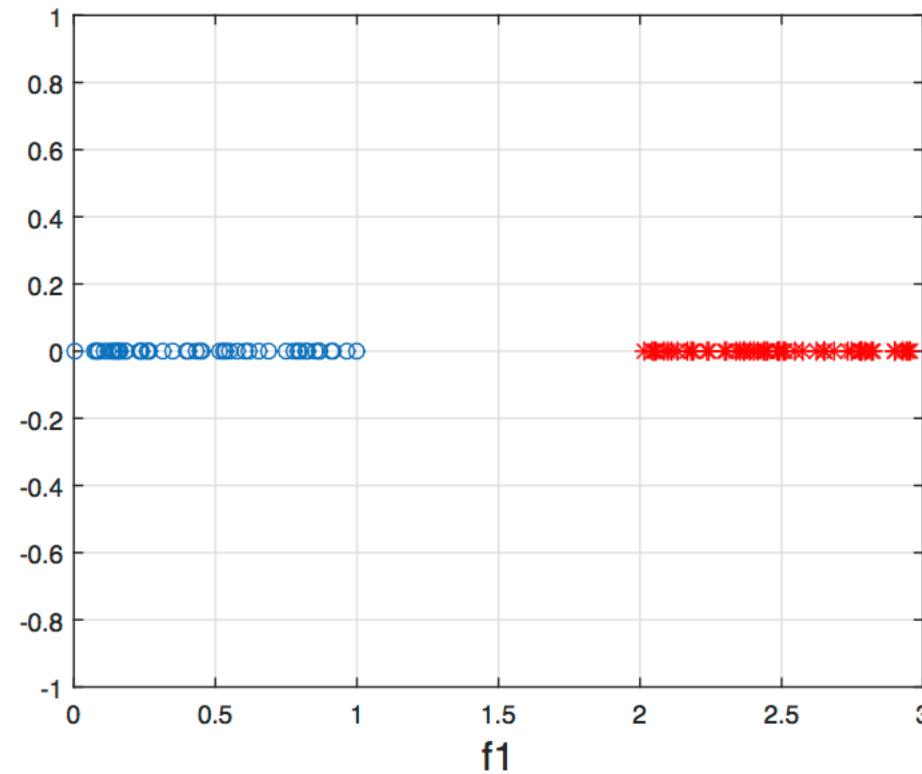
- Дано: матрица «объекты-признаки»  $X$  размера  $\ell \times D$
- Найти: новую матрицу «объекты-признаки»  $Z$  размера  $\ell \times d$
- $d < D$

# Но зачем?

- Шумовые признаки
- Переобучение
- Интерпретируемость модели
- Скорость работы модели
- Визуализация данных

# Плохие признаки

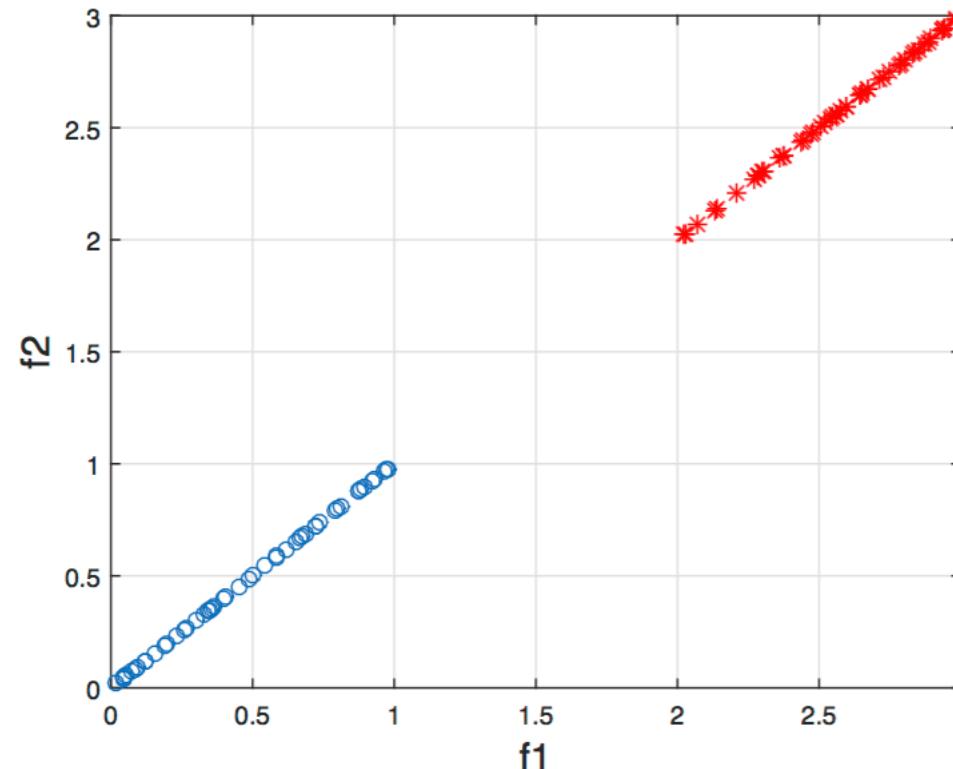
Информативный  
признак



# Плохие признаки

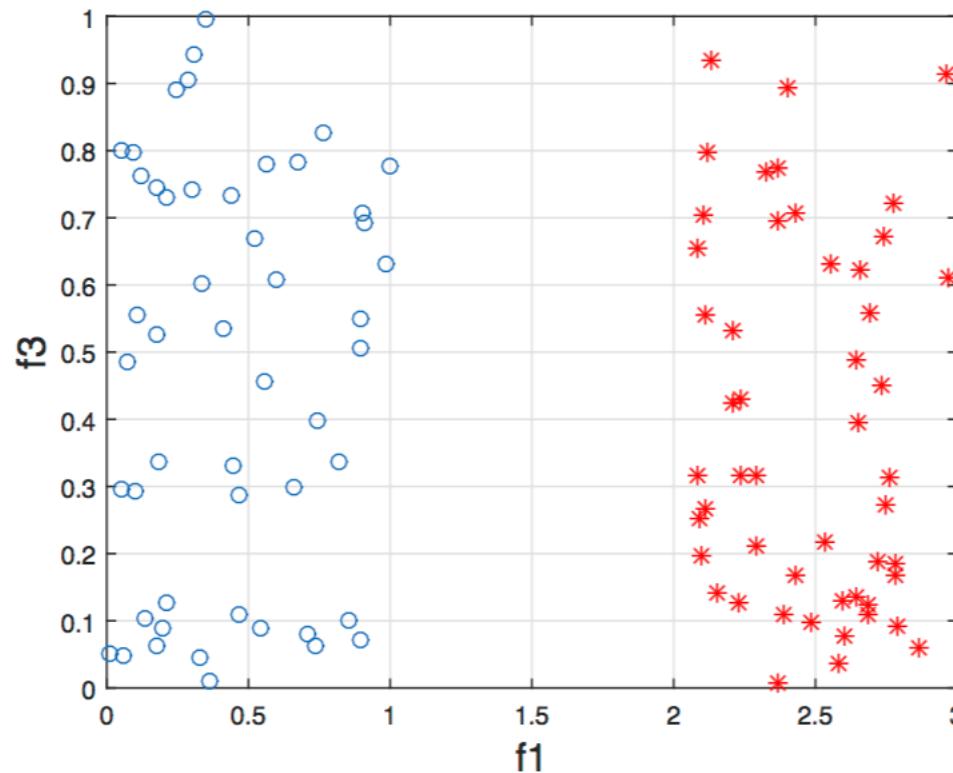
Коррелирующие  
признаки

$f_2$  — избыточный  
признак



# Плохие признаки

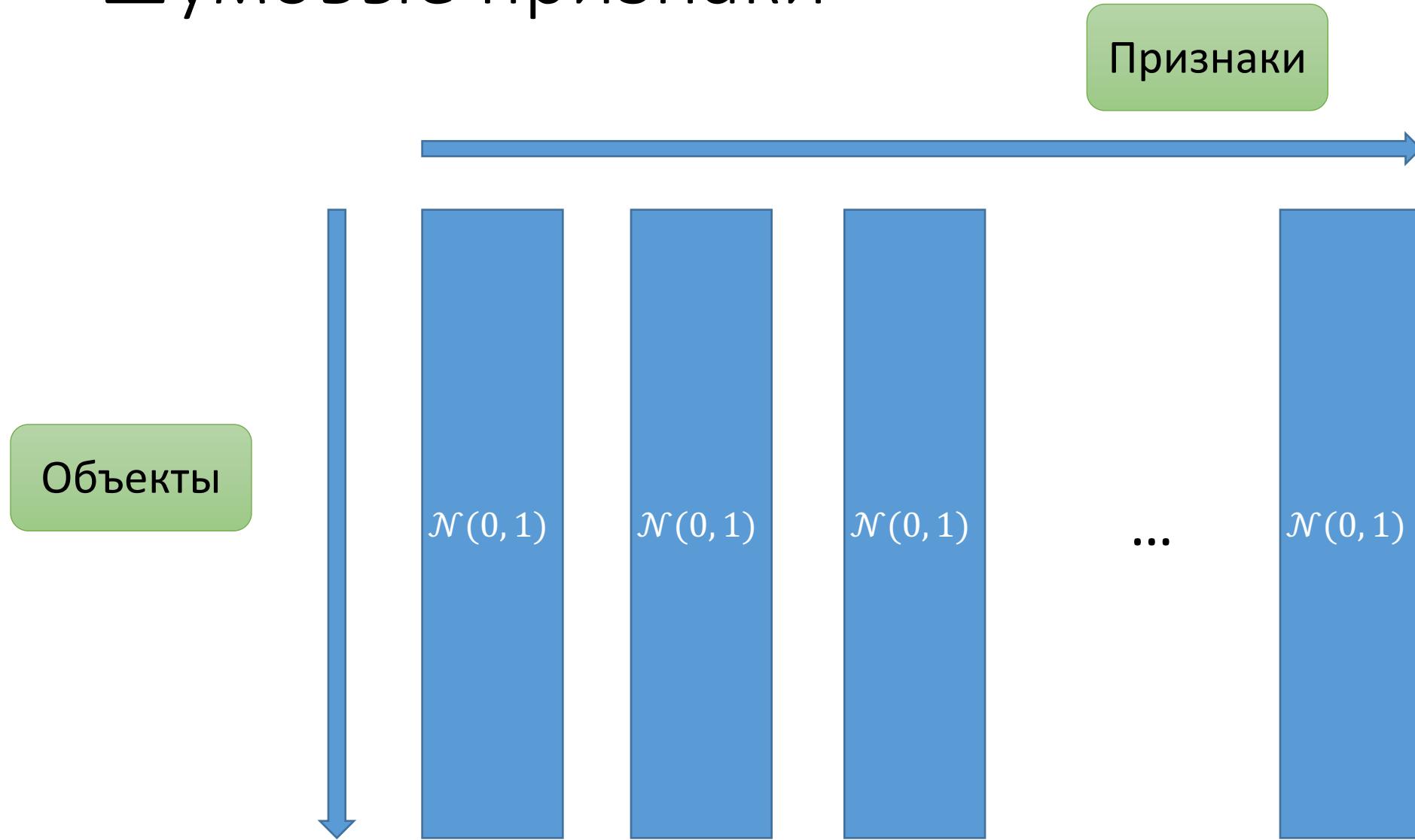
f3 — шумовой  
признак



# Шумовые признаки

- Признаки, которые никак не связаны с целевой переменной
- Но по обучающей выборке это не всегда можно понять

# Шумовые признаки



# Шумовые признаки

- Генерируем случайные признаки
- Если их много, то некоторые будут хорошо коррелировать с ответами

$y$	$x_1$	$x_2$	$x_3$	$x_4$
-1	1.11	-0.5	0.42	0.33
-1	1.22	-0.46	-1.98	-0.55
1	-1.56	0.04	0.39	-1.67
1	-0.48	1.32	0.88	-0.27

# Ускорение моделей

- Чем больше признаков, тем дольше обучаются модели
- Чем дольше обучаются модели, тем меньше экспериментов удаётся провести

# Ускорение моделей

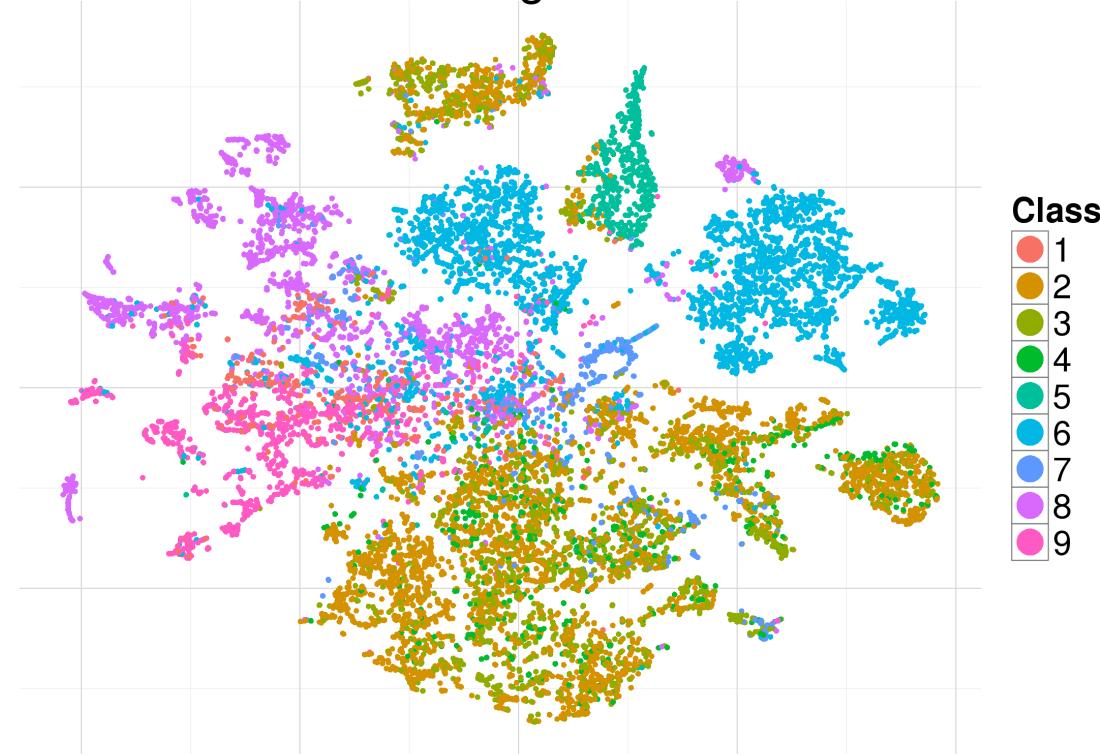
- Чем больше признаков, тем сложнее модели
- Чем сложнее модели, тем дольше они вычисляют прогнозы
- Могут быть жёсткие ограничения на скорость
- Пример: рекомендательные системы

# Визуализация

-99.99	-99.99	315.7	317.45	317.5	317.26	315.86	314.93	313.2	312.44	313.33	314.67	-99.99
315.62	316.38	316.71	317.72	318.29	318.16	316.54	314.8	313.84	313.26	314.8	315.58	315.98
316.43	316.97	317.58	319.02	320.03	319.59	318.18	315.91	314.16	313.84	315	316.19	316.91
316.93	317.7	318.54	319.48	320.58	319.77	318.57	316.79	314.8	315.38	316.1	317.01	317.64
317.94	318.56	319.68	320.63	321.01	320.55	319.58	317.4	316.25	315.42	316.69	317.7	318.45
318.74	319.08	319.86	321.39	322.24	321.47	319.74	317.77	316.21	315.99	317.12	318.31	318.99
319.57	-99.99	-99.99	-99.99	322.24	321.89	320.44	318.7	316.7	316.79	317.79	318.71	-99.99
319.44	320.44	320.89	322.13	322.16	321.87	321.39	318.8	317.81	317.3	318.87	319.42	320.04
320.62	321.59	322.39	323.87	324.01	323.75	322.39	320.37	318.64	318.1	319.79	321.08	321.38
322.06	322.5	323.04	324.42	325	324.09	322.55	320.92	319.31	319.31	320.72	321.96	322.16
322.57	323.15	323.89	325.02	325.57	325.36	324.14	322.03	320.41	320.25	321.31	322.84	323.05
324	324.42	325.64	326.66	327.34	326.76	325.88	323.67	322.38	321.78	322.85	324.12	324.63
325.03	325.99	326.87	328.14	328.07	327.66	326.35	324.69	323.1	323.16	323.98	325.13	325.68
326.17	326.68	327.18	327.78	328.92	328.57	327.34	325.46	323.36	323.57	324.8	326.01	326.32
326.77	327.63	327.75	329.72	330.07	329.09	328.05	326.32	324.93	325.06	326.5	327.55	327.45
328.55	329.56	330.3	331.5	332.48	332.07	330.87	329.31	327.51	327.18	328.16	328.64	329.68
329.35	330.71	331.48	332.65	333.09	332.25	331.18	329.4	327.43	327.37	328.46	329.57	330.25
330.4	331.41	332.04	333.31	333.96	333.6	331.91	330.06	328.56	328.34	329.49	330.76	331.15
331.75	332.56	333.5	334.58	334.87	334.34	333.05	330.94	329.3	328.94	330.31	331.68	332.15
332.93	333.42	334.7	336.07	336.74	336.27	334.93	332.75	331.59	331.16	332.4	333.85	333.9
334.97	335.39	336.64	337.76	338.01	337.89	336.54	334.68	332.76	332.55	333.92	334.95	335.51
336.23	336.76	337.96	338.89	339.47	339.29	337.73	336.09	333.91	333.86	335.29	336.73	336.85
338.01	338.36	340.08	340.77	341.46	341.17	339.56	337.6	335.88	336.02	337.1	338.21	338.69
339.23	340.47	341.38	342.51	342.91	342.25	340.49	338.43	336.69	336.86	338.36	339.61	339.93
340.75	341.61	342.7	343.57	344.13	343.35	342.06	339.81	337.98	337.86	339.26	340.49	341.13
341.37	342.52	343.1	344.94	345.75	345.32	343.99	342.39	339.86	339.99	341.15	342.99	342.78
343.7	344.5	345.28	347.08	347.43	346.79	345.4	343.28	341.07	341.35	342.98	344.22	344.42
344.97	346	347.43	348.35	348.93	348.25	346.56	344.68	343.09	342.8	344.24	345.55	345.9
346.3	346.96	347.86	349.55	350.21	349.54	347.94	345.9	344.85	344.17	345.66	346.9	347.15
348.02	348.47	349.42	350.99	351.84	351.25	349.52	348.1	346.45	346.36	347.81	348.96	348.93
350.43	351.73	352.22	353.59	354.22	353.79	352.38	350.43	348.72	348.88	350.07	351.34	351.48
352.76	353.07	353.68	355.42	355.67	355.13	353.9	351.67	349.8	349.99	351.29	352.52	352.91
353.66	354.7	355.39	356.2	357.16	356.23	354.82	352.91	350.96	351.18	352.83	354.21	354.19
354.72	355.75	357.16	358.6	359.33	358.24	356.17	354.02	352.15	352.21	353.75	354.99	355.59
355.98	356.72	357.81	359.15	359.66	359.25	357.02	355	353.01	353.31	354.16	355.4	356.37
356.7	357.16	358.38	359.46	360.28	359.6	357.57	355.52	353.69	353.99	355.34	356.8	357.04
358.37	358.91	359.97	361.26	361.68	360.95	359.55	357.48	355.84	355.99	357.58	359.04	358.89
359.97	361	361.64	363.45	363.79	363.26	361.9	359.46	358.05	357.76	359.56	360.7	360.88
362.05	363.25	364.02	364.72	365.41	364.97	363.65	361.48	359.45	359.6	360.76	362.33	362.64
363.18	364	364.56	366.35	366.79	365.62	364.47	362.51	360.19	360.77	362.43	364.28	363.76
365.33	366.15	367.31	368.61	369.3	368.87	367.64	365.77	363.9	364.23	365.46	366.97	366.63
368.15	368.87	369.59	371.14	371	370.35	369.27	366.93	364.63	365.13	366.67	368.01	368.31
369.14	369.46	370.52	371.66	371.82	371.7	370.12	368.12	366.62	366.73	368.29	369.53	369.48
370.28	371.5	372.12	372.87	374.02	373.3	371.62	369.55	367.96	368.09	369.68	371.24	371.02
372.43	373.09	373.52	374.86	375.55	375.41	374.02	371.49	370.7	370.25	372.08	373.78	373.1
374.68	375.63	376.11	377.65	378.35	378.13	376.62	374.5	372.99	373.01	374.35	375.7	375.64
376.79	377.37	378.41	380.52	380.63	379.57	377.79	375.86	374.07	374.24	375.86	377.47	377.38
378.37	379.69	380.41	382.1	382.28	382.13	380.66	378.71	376.42	376.88	378.32	380.04	379.67
381.38	382.03	382.64	384.62	384.95	384.06	382.29	380.47	378.67	379.06	380.14	381.74	381.84
382.45	383.68	384.23	386.26	386.39	385.87	384.39	381.78	380.73	380.81	382.33	383.69	383.55
385.07	385.72	385.85	386.71	388.45	387.64	386.1	383.95	382.91	382.73	383.96	385.02	385.34

# Визуализация

t-SNE 2D Embedding of Products Data



# Методы понижения размерности

- Отбор признаков (feature selection)
  - Выбрать  $d$  самых важных признаков
- Извлечение признаков (feature extraction)
  - Найти  $d$  новых признаков, выражающихся через исходные

# Методы понижения размерности

- Фильтрация (filter methods)
  - Понижение размерности без учёта модели
- Методы-обёртки (wrapper methods)
  - Выбор признаков, дающих лучшее качество для модели
- Понижение с помощью моделей (embedded methods)
  - Использование свойств моделей для оценивания важности признаков

# Одномерные методы

# Одномерные методы

- Оценивают важность каждого признака по отдельности
- Относятся к **методам фильтрации**
- Относятся к **методам отбора признаков**

# Обозначения

- $x_{ij}$  — значение  $j$ -го признака на  $i$ -м объекте
- $\bar{x}_j$  — среднее значение  $j$ -го признака
- $y_i$  — значение целевой переменной на  $i$ -м объекте
- $\bar{y}$  — среднее значение целевой переменной

# Дисперсия признаков

$$R_j = \frac{1}{\ell} \sum_{i=1}^{\ell} (x_{ij} - \bar{x}_j)^2$$

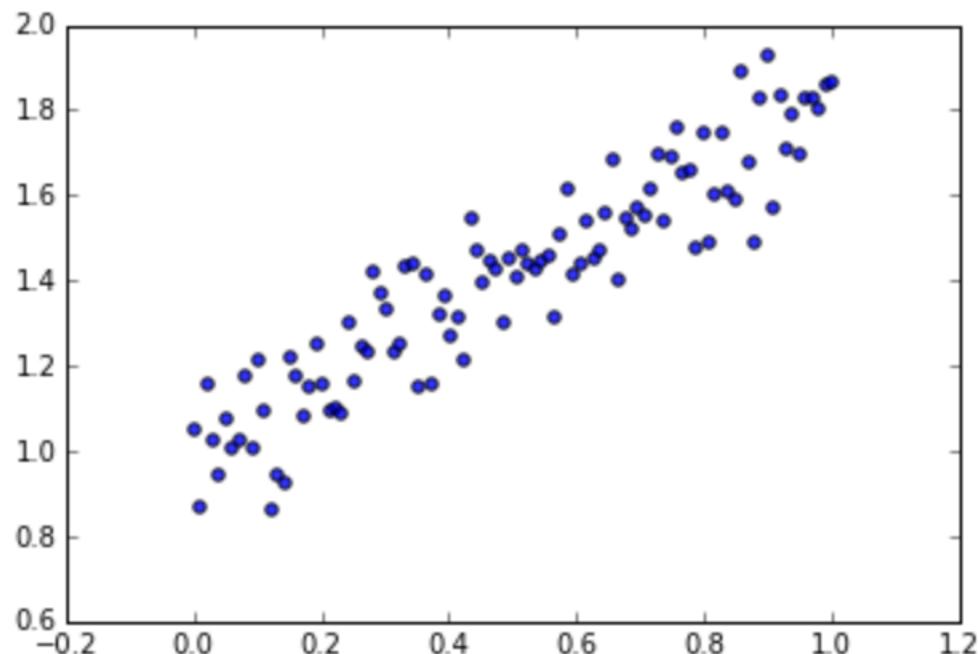
- Чем больше  $R_j$ , тем информативнее признак
- Никак не учитываются ответы
- Подходит для фильтрации константных и близких к ним признаков

# Корреляция

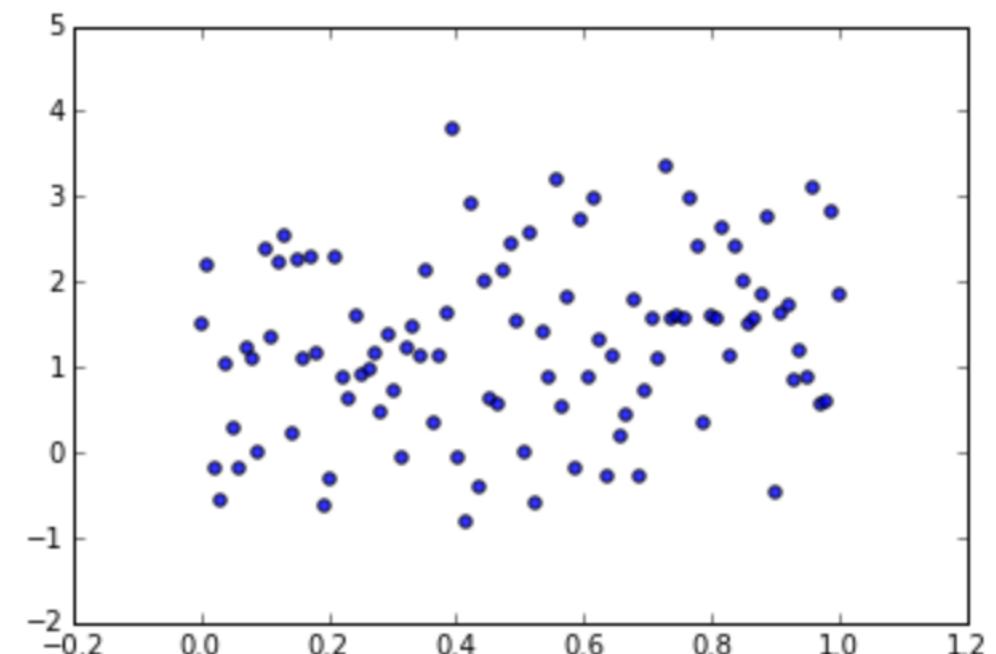
$$R_j = \frac{\sum_{i=1}^{\ell} (x_{ij} - \bar{x}_j)(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{\ell} (x_{ij} - \bar{x}_j)^2 \sum_{i=1}^{\ell} (y_i - \bar{y})^2}}$$

- Чем больше  $|R_j|$ , тем информативнее признак
- Учитывает только линейную связь

# Корреляция для регрессии

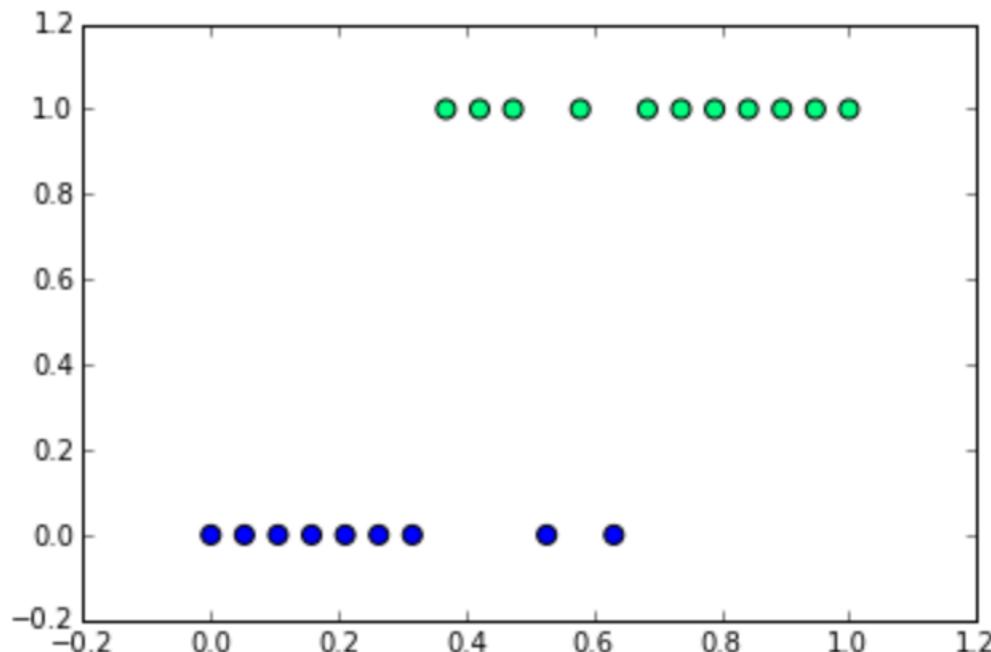


Корреляция  
0.927

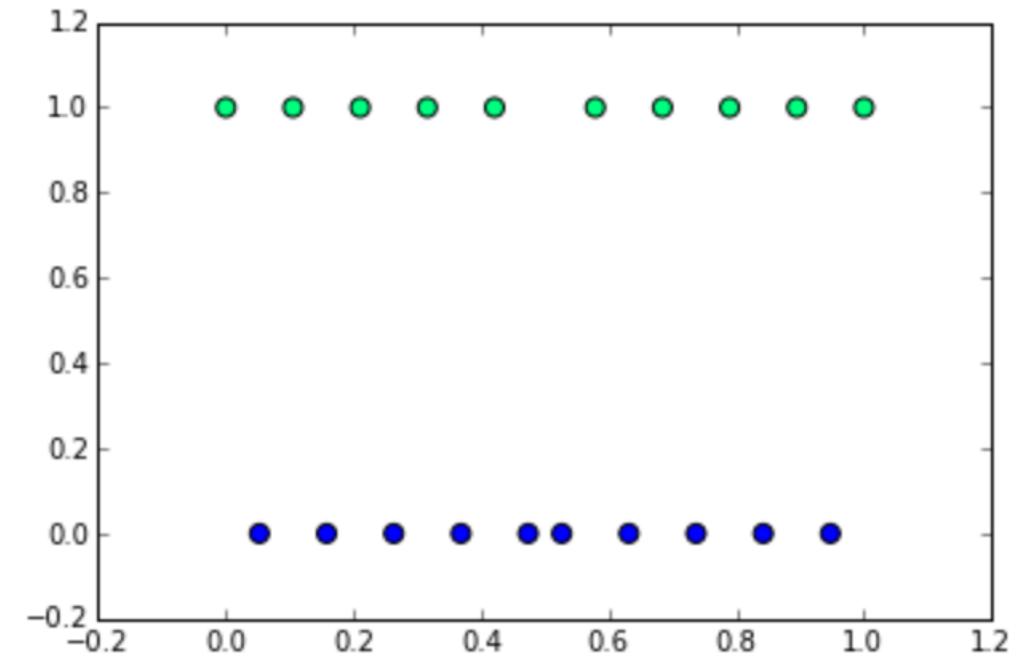


Корреляция  
0.207

# Корреляция для классификации

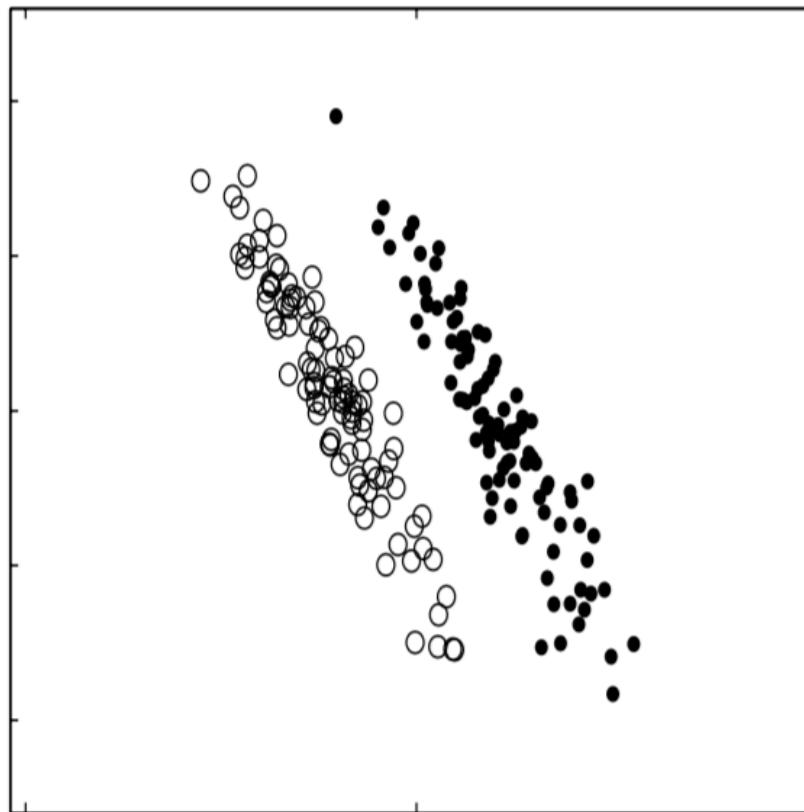


Корреляция  
0.741

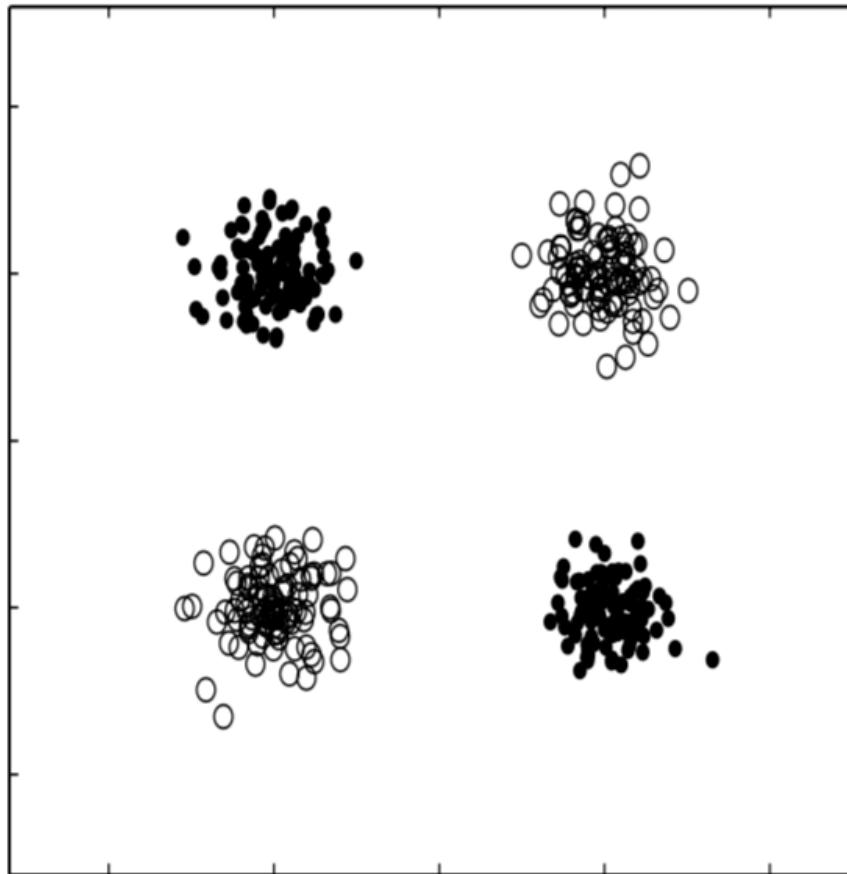


Корреляция  
0.0

# Проблемы

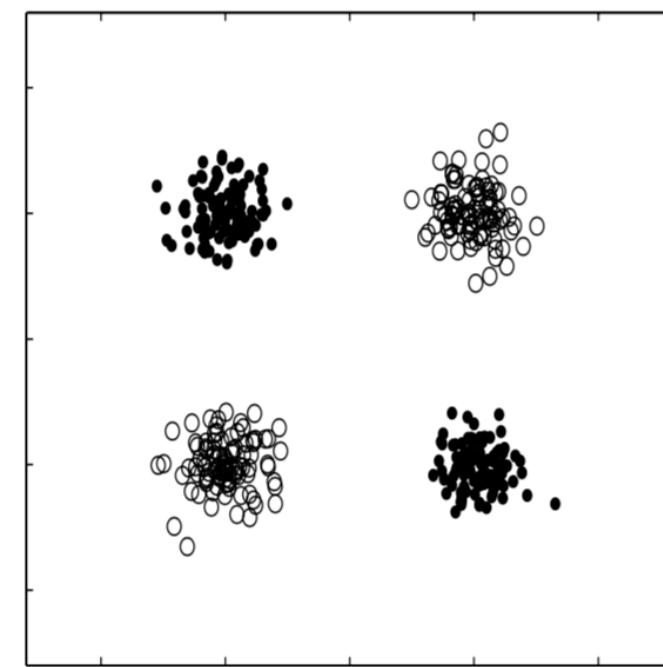
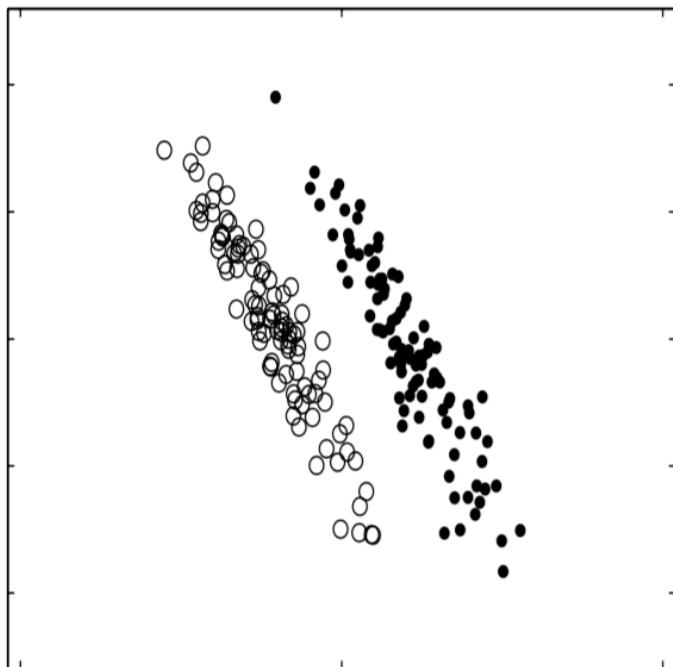


# Проблемы



# Проблемы

- Одномерные критерии не работают, если целевая переменная зависит от совокупностей признаков



Отбор с помощью моделей

# Отбор с помощью моделей

- Оценивают важность признаков, используя модели машинного обучения
- Относятся к **методам отбора признаков**

# Линейные модели

$$a(x) = \sum_{j=1}^d w_j x^j$$

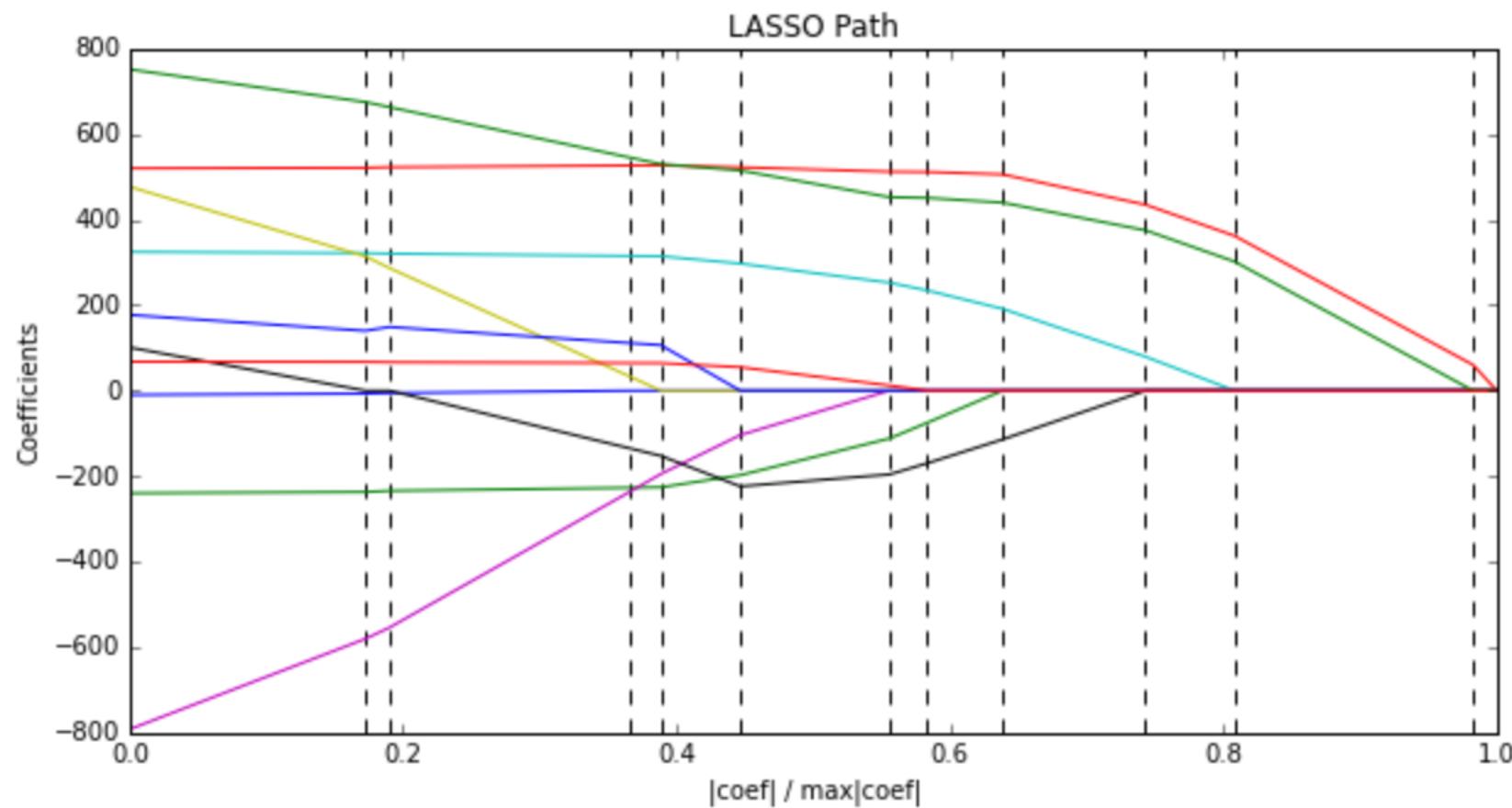
- Если признаки масштабированы, то веса можно использовать как показатели информативности
- Для повышения числа нулевых весов —  $L_1$ -регуляризация

# $L_1$ -регуляризация

$$Q(a, X) + \lambda \sum_{j=1}^d |w_j| \rightarrow \min_w$$

- Чем выше  $\lambda$ , тем больше весов зануляется
- Позволяет построить модель, использующую только самые важные признаки

# $L_1$ -регуляризация



# Решающие деревья

- Поиск лучшего разбиения:

$$Q(X_m, j, t) = H(X_m) - \frac{|X_l|}{|X_m|} H(X_l) - \frac{|X_r|}{|X_m|} H(X_r) \rightarrow \max_{j,t}$$

- $H(X)$  — критерий информативности (MSE, энтропийный)

# Решающие деревья

- Чем сильнее уменьшили  $H(X)$ , тем лучше признак
- Уменьшение критерия:

$$H(X_m) - \frac{|X_l|}{|X_m|} H(X_l) - \frac{|X_r|}{|X_m|} H(X_r)$$

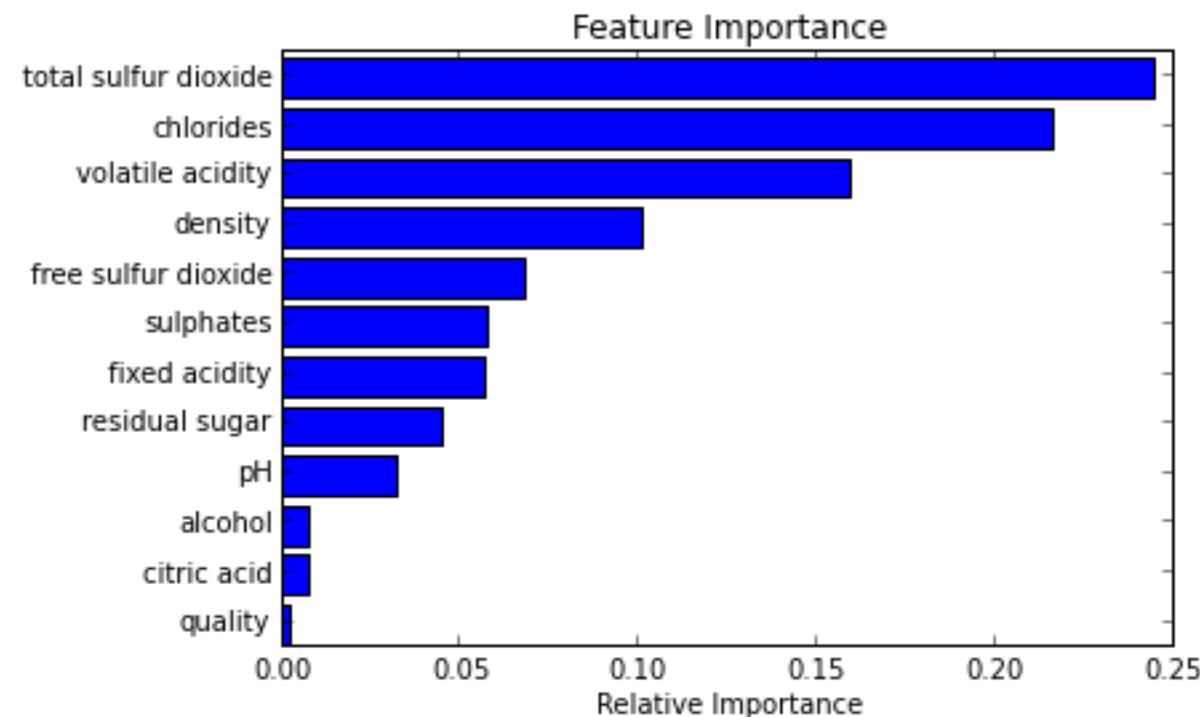
- Важность признака  $R_j$ : просуммируем уменьшения по всем вершинам, где разбиение делалось по признаку  $j$

# Случайный лес

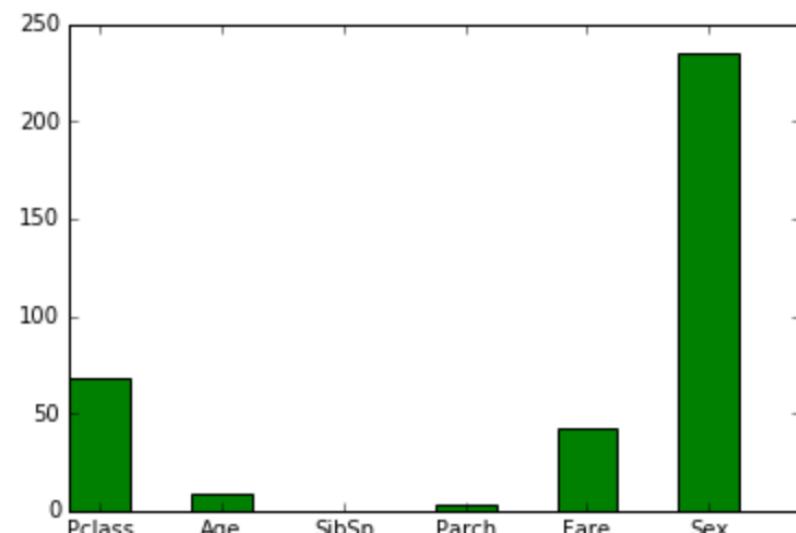
- Сумма важностей  $R_j$  по всем деревьям
- Чем больше, тем важнее признак
- Учитывается важность признаков в совокупности

# Случайные леса и отбор признаков

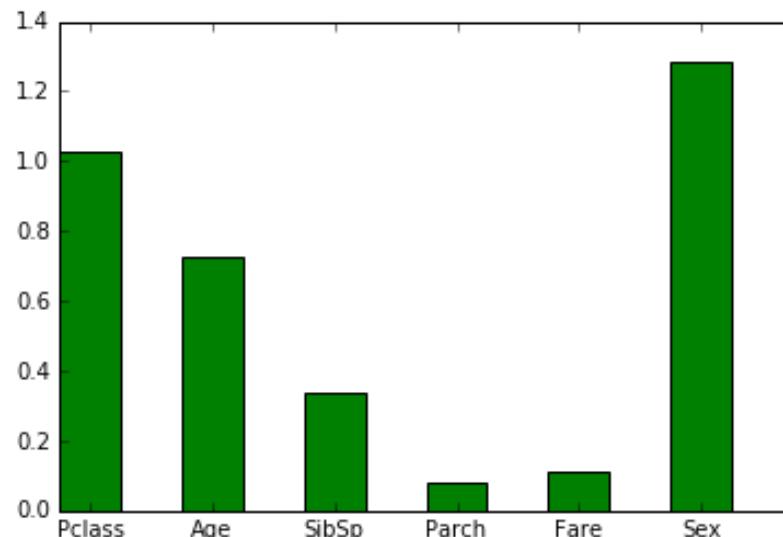
- Классификация вин на белые и красные



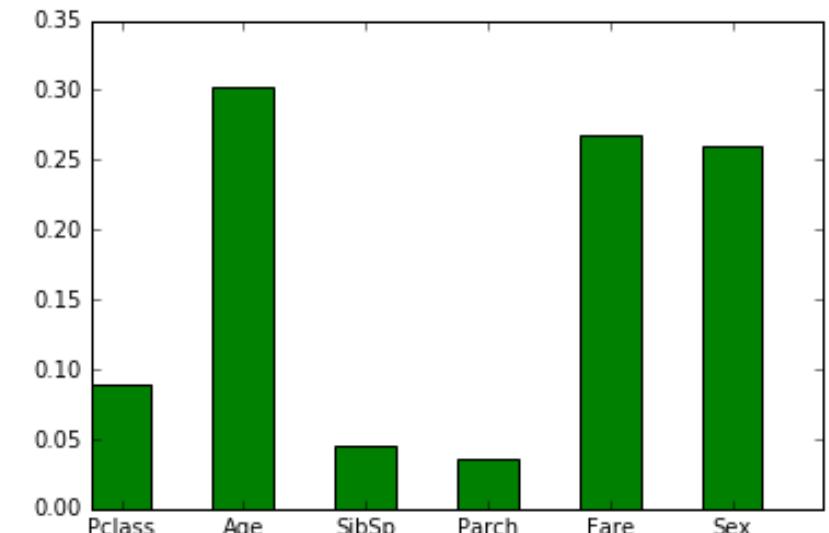
# Пример: Titanic



Одномерный отбор



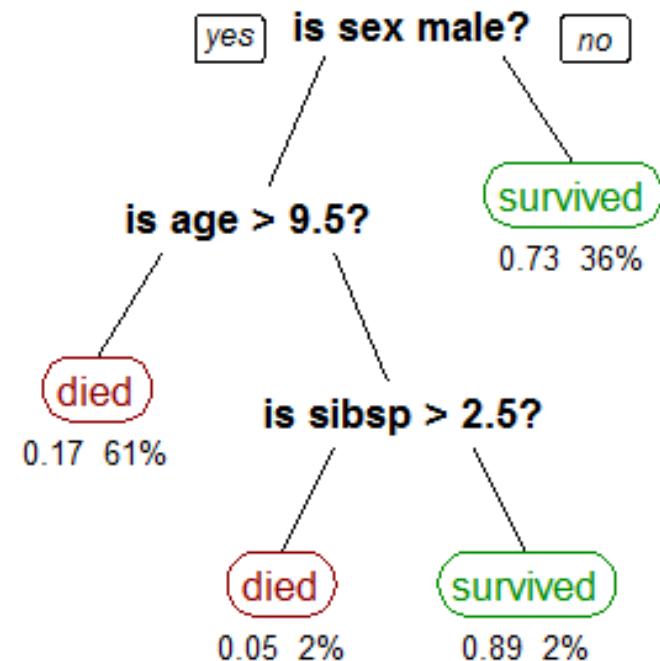
Логистическая регрессия



Случайный лес

# Пример: Titanic

- Модели выделяют признак Age как важный
- Ответы зависят от возраста только в совокупности с полом

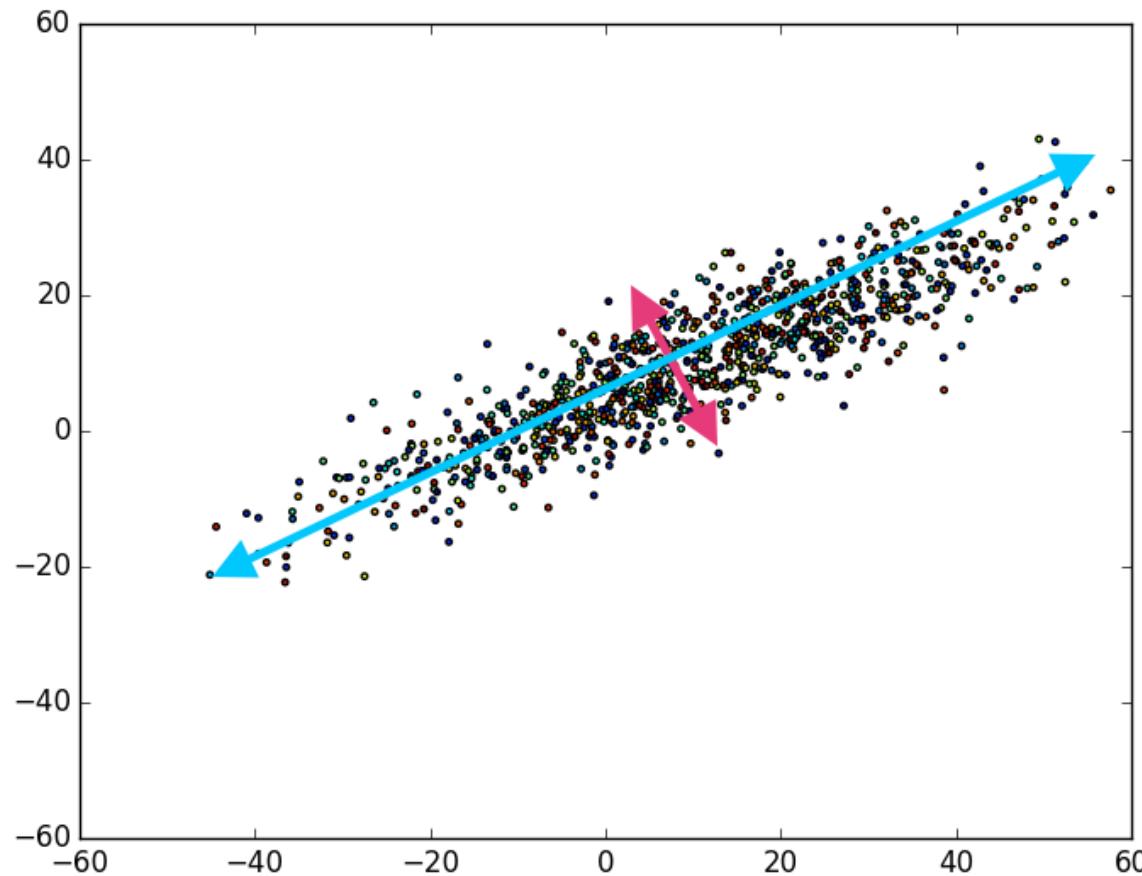


# Метод главных компонент

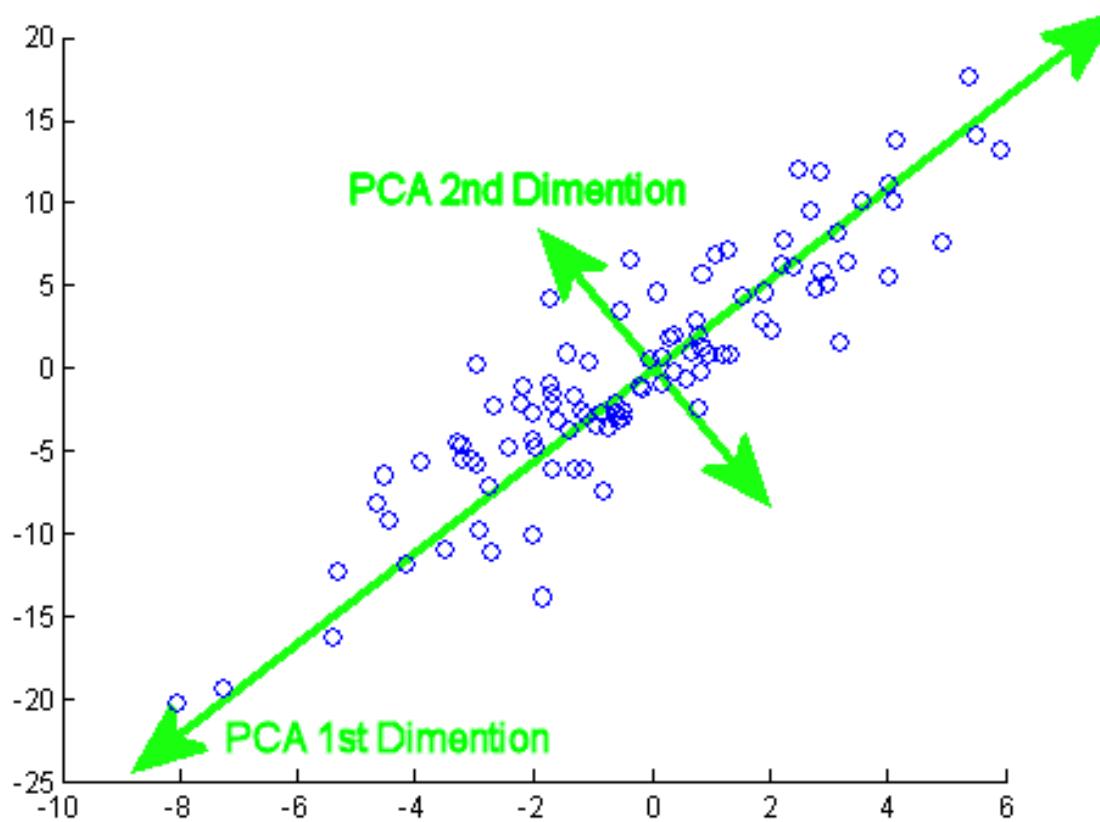
# Метод главных компонент

- Principal component analysis (PCA)
- Проецирует данные в пространство меньшей размерности
- Относится к **методам фильтрации**
- Относится к **методам извлечения признаков**

# Извлечение признаков



# Извлечение признаков



# Извлечение признаков

- Порождение новых признаков
- Их должно быть меньше
- Они должны содержать как можно больше информации из исходных признаков

# Извлечение признаков

- Линейные методы
- Каждый новый признак — линейная комбинация исходных

# Извлечение признаков

- Исходные признаки:  $x_{ik}$ ,  $D$  штук
- Новые признаки:  $z_{ij}$ ,  $d$  штук
- Линейный подход:

$$z_{ij} = \sum_{k=1}^D w_{jk} x_{ik}$$

Новые признаки

Вклад исходного  $k$ -го  
признака в новый  $j$ -й

Исходные признаки

# Метод главных компонент

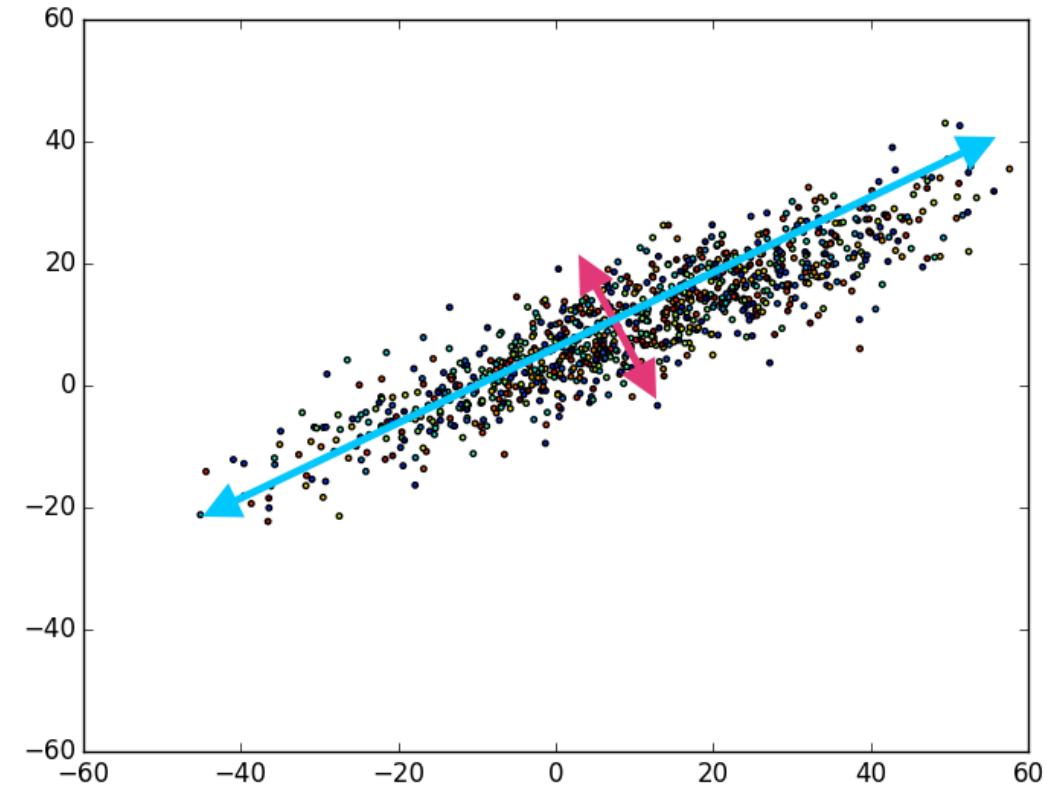
- Матричная запись:

$$Z = XW^T$$

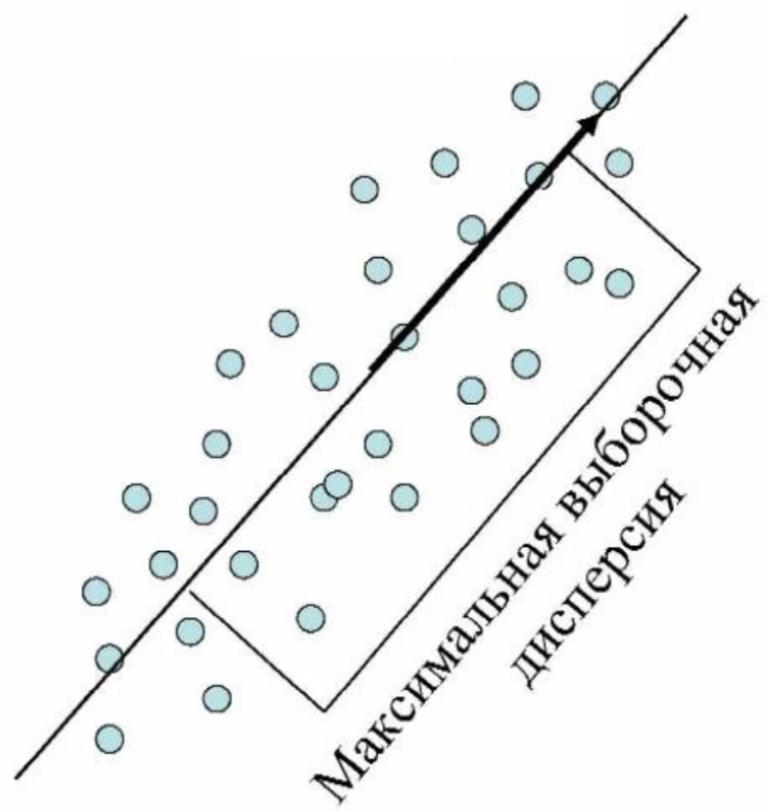
- $j$ -й столбец  $W$  — коэффициенты при исходных признаках для вычисления нового  $j$ -го признака

# Метод главных компонент

- Чем выше дисперсия выборки после проецирования, тем лучше
- Дисперсия — мера количества информации



# Метод главных компонент



# Максимизация дисперсии

$$\begin{cases} \sum_{j=1}^d w_j^T X^T X w_j \rightarrow \max_w \\ W^T W = I \end{cases}$$

# Максимизация дисперсии

$$\left\{ \sum_{j=1}^d w_j^T X^T X w_j \rightarrow \max_w \right.$$

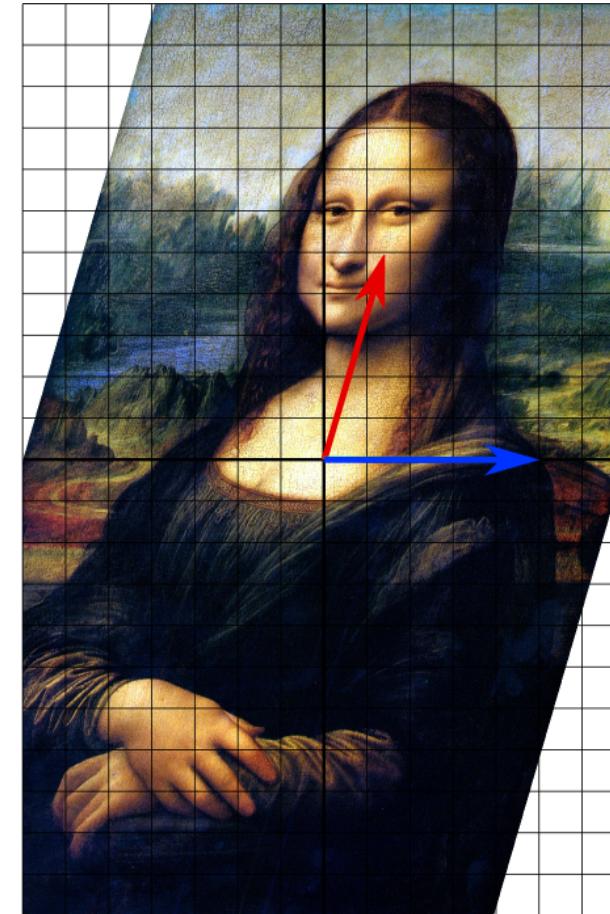
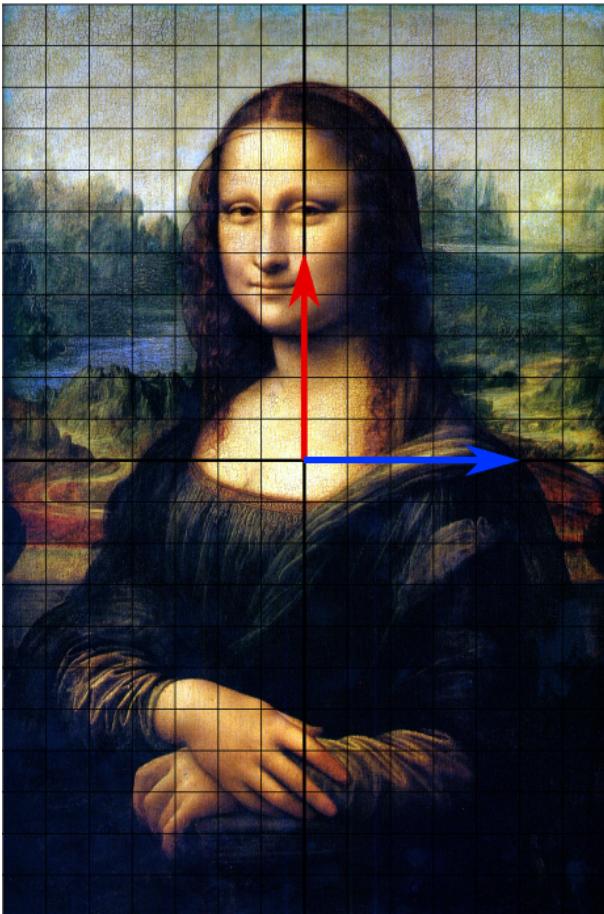
$$W^T W = I$$

Дисперсия выборки

# Собственные векторы

- $A$  — матрица размера  $n \times n$
- Пусть  $Ax = \lambda x$
- Тогда  $x$  — собственный вектор,  $\lambda$  — собственное значение
- $x$  — вектор, который не меняет направление под действием матрицы

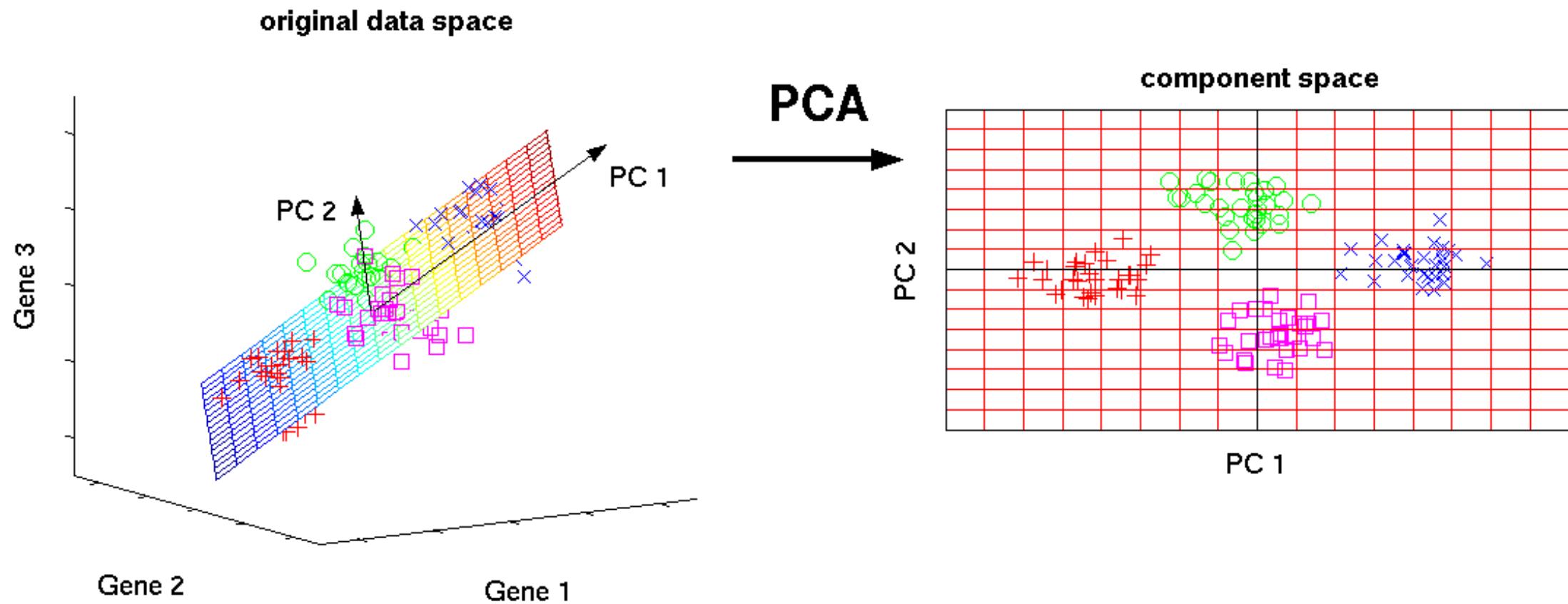
# Собственные векторы



# Решение

- Столбцы  $W$  — собственные векторы матрицы  $X^T X$ , соответствующие наибольшим собственным значениям  $\lambda_1, \lambda_2, \dots, \lambda_d$
- $\frac{\sum_{i=1}^d \lambda_i}{\sum_{i=1}^D \lambda_i}$  — доля дисперсии, сохранённой при понижении размерности

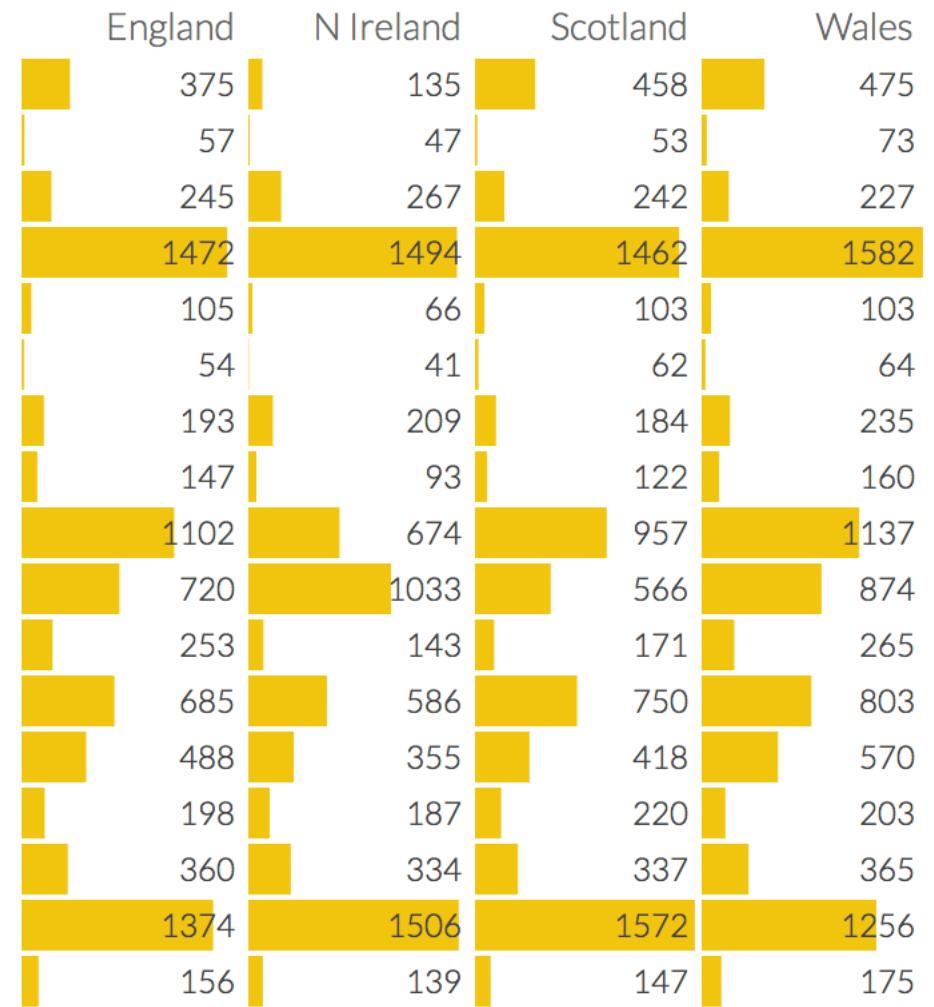
# Метод главных компонент



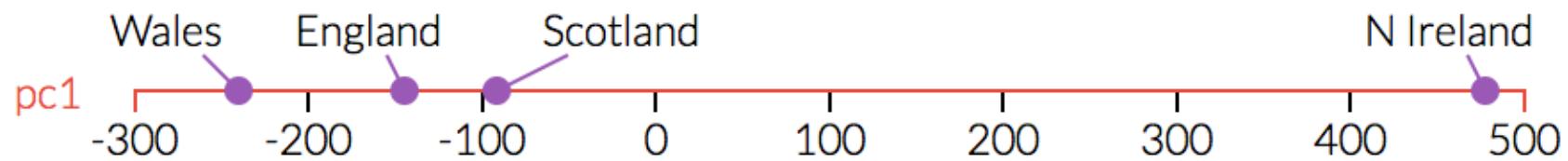
# Рацион в Великобритании

- Данные — среднее потребление продуктов в неделю в каждой провинции
- Не очень удобно смотреть на них

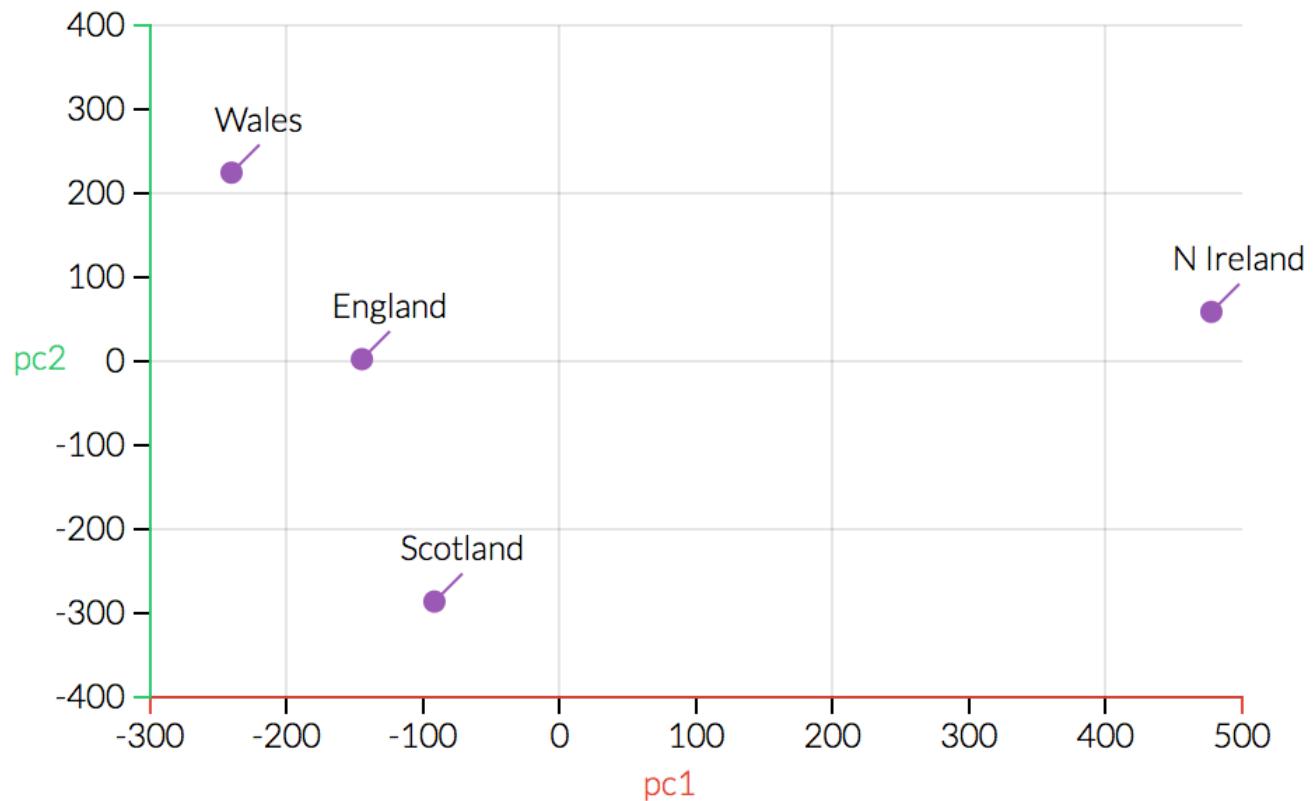
Alcoholic drinks	
Beverages	
Carcase meat	
Cereals	
Cheese	
Confectionery	
Fats and oils	
Fish	
Fresh fruit	
Fresh potatoes	
Fresh Veg	
Other meat	
Other Veg	
Processed potatoes	
Processed Veg	
Soft drinks	
Sugars	



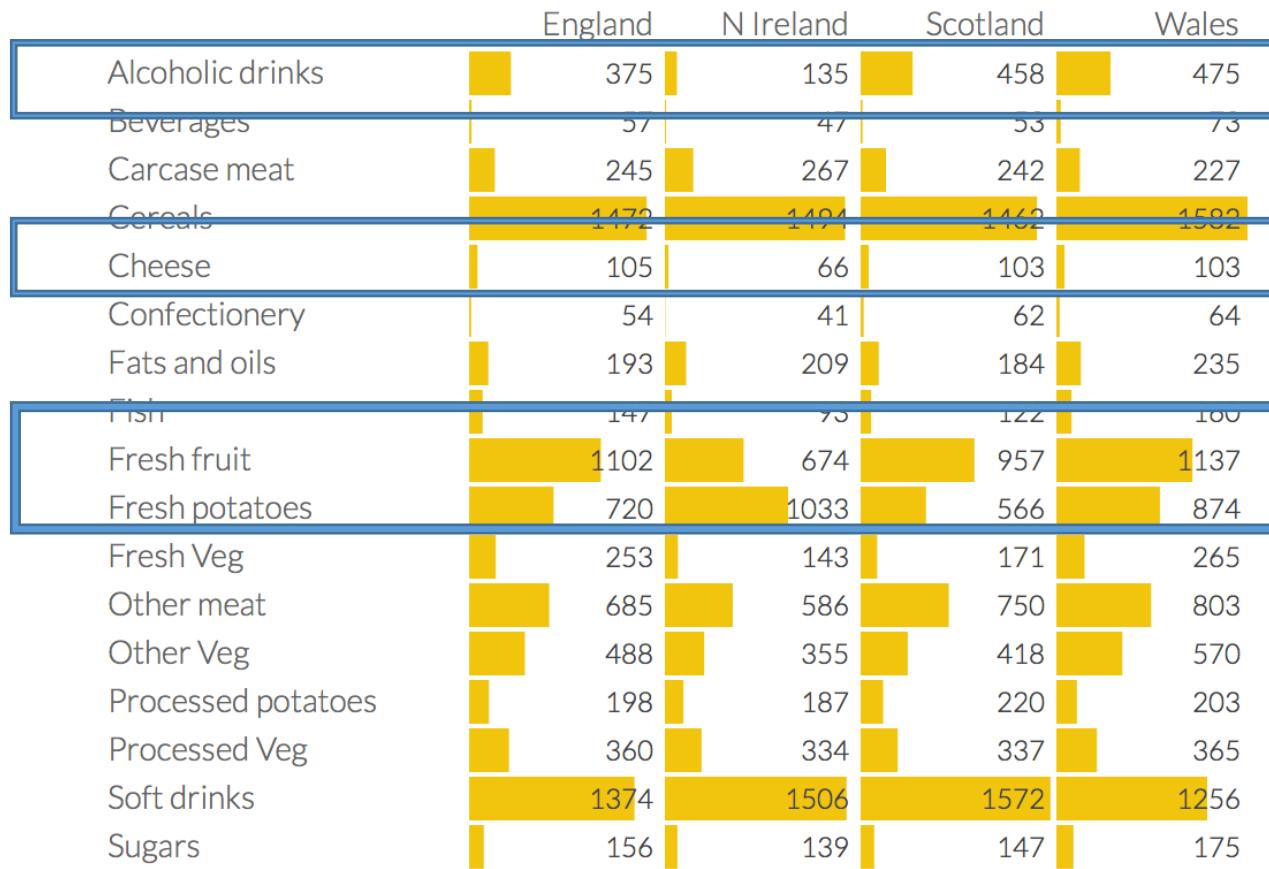
# Рационы в Великобритании



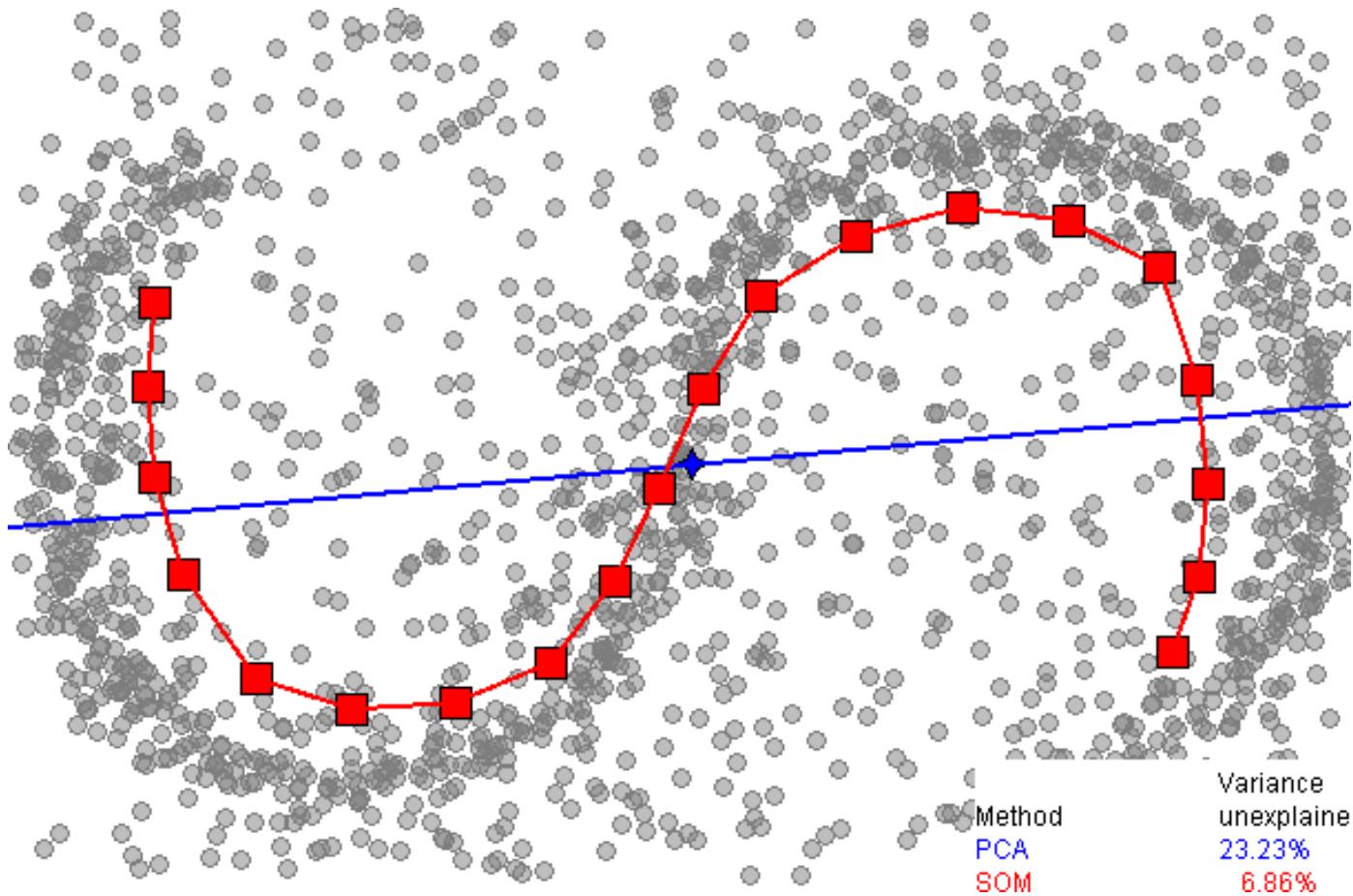
# Рационы в Великобритании



# Рацион в Великобритании



# Ограничения

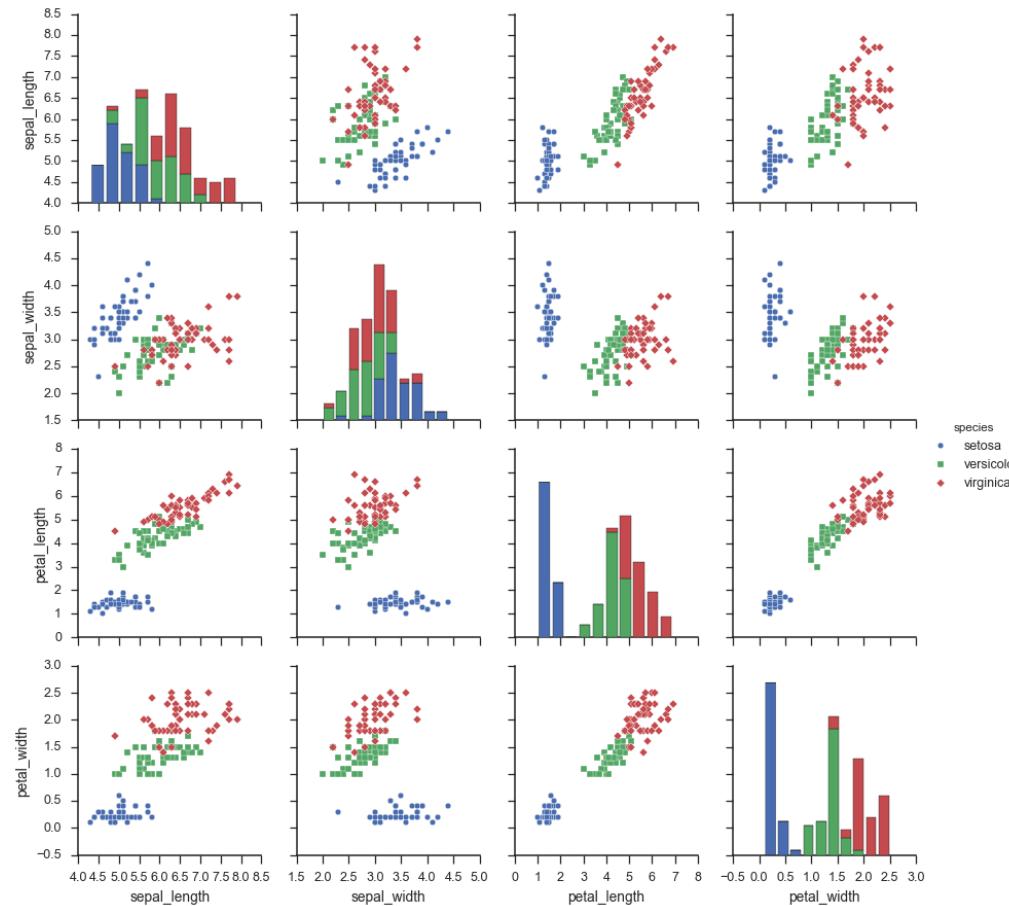


# Визуализация данных

# Визуализация данных

-99.99	-99.99	315.7	317.45	317.5	317.26	315.86	314.93	313.2	312.44	313.33	314.67	-99.99
315.62	316.38	316.71	317.72	318.29	318.16	316.54	314.8	313.84	313.26	314.8	315.58	315.98
316.43	316.97	317.58	319.02	320.03	319.59	318.18	315.91	314.16	313.84	315	316.19	316.91
316.93	317.7	318.54	319.48	320.58	319.77	318.57	316.79	314.8	315.38	316.1	317.01	317.64
317.94	318.56	319.68	320.63	321.01	320.55	319.58	317.4	316.25	315.42	316.69	317.7	318.45
318.74	319.08	319.86	321.39	322.24	321.47	319.74	317.77	316.21	315.99	317.12	318.31	318.99
319.57	-99.99	-99.99	-99.99	322.24	321.89	320.44	318.7	316.7	316.79	317.79	318.71	-99.99
319.44	320.44	320.89	322.13	322.16	321.87	321.39	318.8	317.81	317.3	318.87	319.42	320.04
320.62	321.59	322.39	323.87	324.01	323.75	322.39	320.37	318.64	318.1	319.79	321.08	321.38
322.06	322.5	323.04	324.42	325	324.09	322.55	320.92	319.31	319.31	320.72	321.96	322.16
322.57	323.15	323.89	325.02	325.57	325.36	324.14	322.03	320.41	320.25	321.31	322.84	323.05
324	324.42	325.64	326.66	327.34	326.76	325.88	323.67	322.38	321.78	322.85	324.12	324.63
325.03	325.99	326.87	328.14	328.07	327.66	326.35	324.69	323.1	323.16	323.98	325.13	325.68
326.17	326.68	327.18	327.78	328.92	328.57	327.34	325.46	323.36	323.57	324.8	326.01	326.32
326.77	327.63	327.75	329.72	330.07	329.09	328.05	326.32	324.93	325.06	326.5	327.55	327.45
328.55	329.56	330.3	331.5	332.48	332.07	330.87	329.31	327.51	327.18	328.16	328.64	329.68
329.35	330.71	331.48	332.65	333.09	332.25	331.18	329.4	327.43	327.37	328.46	329.57	330.25
330.4	331.41	332.04	333.31	333.96	333.6	331.91	330.06	328.56	328.34	329.49	330.76	331.15
331.75	332.56	333.5	334.58	334.87	334.34	333.05	330.94	329.3	328.94	330.31	331.68	332.15
332.93	333.42	334.7	336.07	336.74	336.27	334.93	332.75	331.59	331.16	332.4	333.85	333.9
334.97	335.39	336.64	337.76	338.01	337.89	336.54	334.68	332.76	332.55	333.92	334.95	335.51
336.23	336.76	337.96	338.89	339.47	339.29	337.73	336.09	333.91	333.86	335.29	336.73	336.85
338.01	338.36	340.08	340.77	341.46	341.17	339.56	337.6	335.88	336.02	337.1	338.21	338.69
339.23	340.47	341.38	342.51	342.91	342.25	340.49	338.43	336.69	336.86	338.36	339.61	339.93
340.75	341.61	342.7	343.57	344.13	343.35	342.06	339.81	337.98	337.86	339.26	340.49	341.13
341.37	342.52	343.1	344.94	345.75	345.32	343.99	342.39	339.86	339.99	341.15	342.99	342.78
343.7	344.5	345.28	347.08	347.43	346.79	345.4	343.28	341.07	341.35	342.98	344.22	344.42
344.97	346	347.43	348.35	348.93	348.25	346.56	344.68	343.09	342.8	344.24	345.55	345.9
346.3	346.96	347.86	349.55	350.21	349.54	347.94	345.9	344.85	344.17	345.66	346.9	347.15
348.02	348.47	349.42	350.99	351.84	351.25	349.52	348.1	346.45	346.36	347.81	348.96	348.93
350.43	351.73	352.22	353.59	354.22	353.79	352.38	350.43	348.72	348.88	350.07	351.34	351.48
352.76	353.07	353.68	355.42	355.67	355.13	353.9	351.67	349.8	349.99	351.29	352.52	352.91
353.66	354.7	355.39	356.2	356.17	356.23	354.82	352.91	350.96	351.18	352.83	354.21	354.19
354.72	355.75	357.16	358.6	359.33	358.24	356.17	354.02	352.15	352.21	353.75	354.99	355.59
355.98	356.72	357.81	359.15	359.66	359.25	357.02	355	353.01	353.31	354.16	355.4	356.37
356.7	357.16	358.38	359.46	360.28	359.6	357.57	355.52	353.69	353.99	355.34	356.8	357.04
358.37	358.91	359.97	361.26	361.68	360.95	359.55	357.48	355.84	355.99	357.58	359.04	358.89
359.97	361	361.64	363.45	363.79	363.26	361.9	359.46	358.05	357.76	359.56	360.7	360.88
362.05	363.25	364.02	364.72	365.41	364.97	363.65	361.48	359.45	359.6	360.76	362.33	362.64
363.18	364	364.56	366.35	366.79	365.62	364.47	362.51	360.19	360.77	362.43	364.28	363.76
365.33	366.15	367.31	368.61	369.3	368.87	367.64	365.77	363.9	364.23	365.46	366.97	366.63
368.15	368.87	369.59	371.14	371	370.35	369.27	366.93	364.63	365.13	366.67	368.01	368.31
369.14	369.46	370.52	371.66	371.82	371.7	370.12	368.12	366.62	366.73	368.29	369.53	369.48
370.28	371.5	372.12	372.87	374.02	373.3	371.62	369.55	367.96	368.09	369.68	371.24	371.02
372.43	373.09	373.52	374.86	375.55	375.41	374.02	371.49	370.7	370.25	372.08	373.78	373.1
374.68	375.63	376.11	377.65	378.35	378.13	376.62	374.5	372.99	373.01	374.35	375.7	375.64
376.79	377.37	378.41	380.52	380.63	379.57	377.9	375.86	374.07	374.24	375.86	377.47	377.38
378.37	379.69	380.41	382.1	382.28	382.13	380.66	378.71	376.42	376.88	378.32	380.04	379.67
381.38	382.03	382.64	384.62	384.95	384.06	382.29	380.47	378.67	379.06	380.14	381.74	381.84
382.45	383.68	384.23	386.26	386.39	385.87	384.39	381.78	380.73	380.81	382.33	383.69	383.55
385.07	385.72	385.85	386.71	388.45	387.64	386.1	383.95	382.91	382.73	383.96	385.02	385.34

# Визуализация данных



# Визуализация данных

- Частный случай нелинейного понижения размерности
- $d = 2$  или  $d = 3$
- Нужно сохранить структуру данных и зависимости

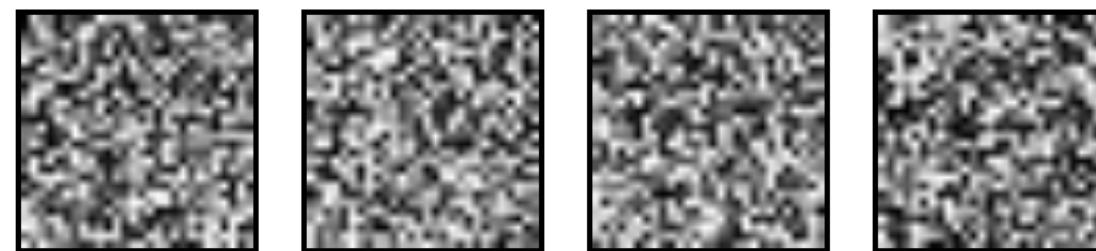
# MNIST

A selection from the 64-dimensional digits dataset

0	1	2	3	4	5	0	1	2	3	4	5	0	5
5	5	0	4	1	3	5	1	0	0	2	2	2	0
4	4	1	5	0	5	2	2	0	0	1	3	2	3
3	4	4	0	5	3	1	5	4	4	2	2	5	5
2	3	4	5	0	1	2	3	4	5	0	5	5	5
0	4	1	3	5	1	0	0	2	2	1	0	1	2
1	5	0	5	2	2	0	0	1	3	2	1	3	4
0	5	3	4	5	4	4	1	2	2	5	5	4	0
5	0	1	2	3	4	5	0	4	2	3	4	5	0
3	5	4	0	0	2	2	0	1	2	3	3	3	4
5	2	2	0	0	1	3	2	4	3	1	3	1	4
3	1	5	4	4	2	2	2	5	5	4	0	3	0
0	1	2	3	4	5	0	1	2	3	4	5	0	4
5	1	0	0	1	2	2	0	1	2	3	3	3	4
1	2	0	0	1	3	2	1	4	3	1	3	4	0
1	5	4	4	2	2	2	5	5	4	4	0	0	1
2	3	4	5	0	1	2	3	4	5	0	5	5	0
0	0	2	2	2	0	1	2	3	3	3	4	4	1
0	0	1	3	2	1	4	3	1	3	1	4	3	0
4	4	2	2	2	5	5	4	4	0	0	1	2	3

# MNIST

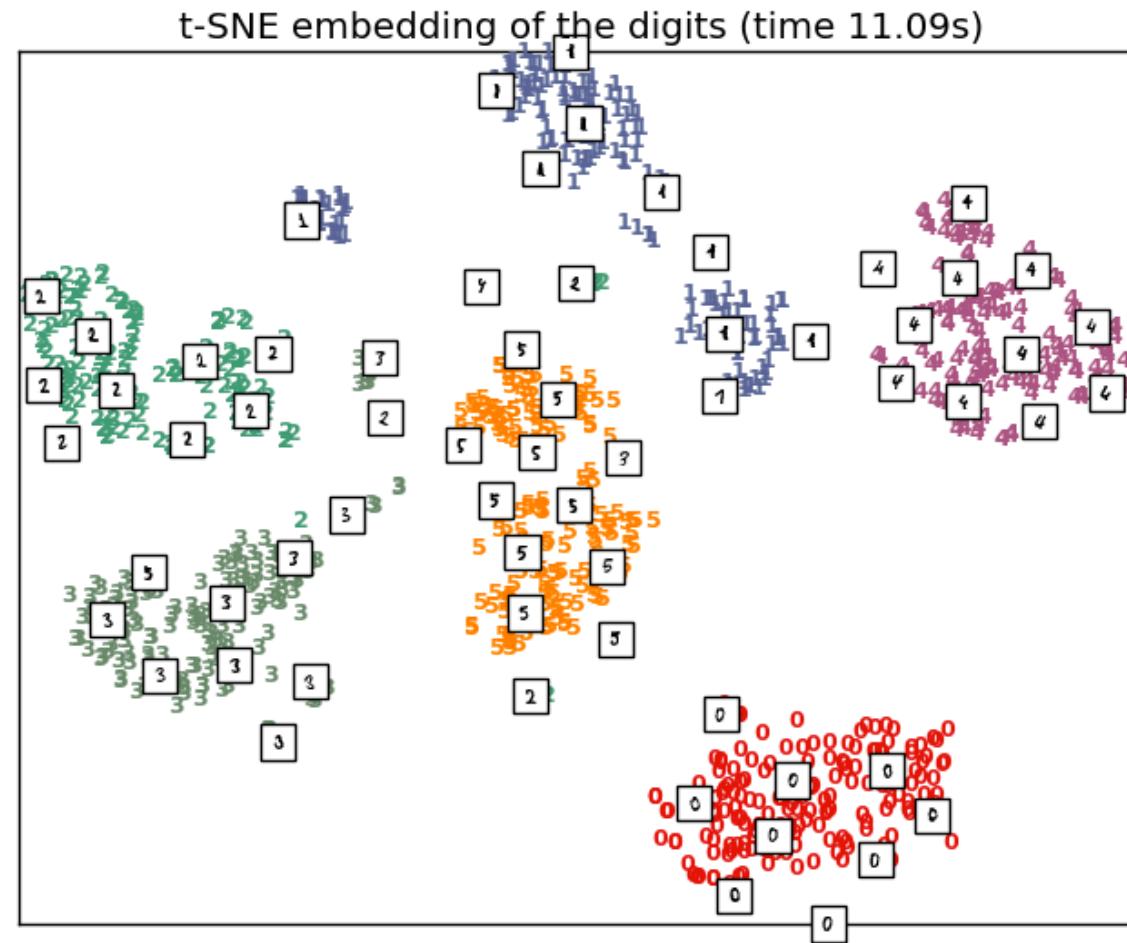
- Каждое изображение — 784 признака
- Внутренняя размерность данных гораздо ниже
- Случайное изображение такого же размера не является изображением цифры



# t-SNE

- t-Stochastic Neighbor Embedding
- Метод визуализации
- Ищет такие точки на плоскости, которые лучше всего сохраняют расстояния из исходного пространства

# MNIST

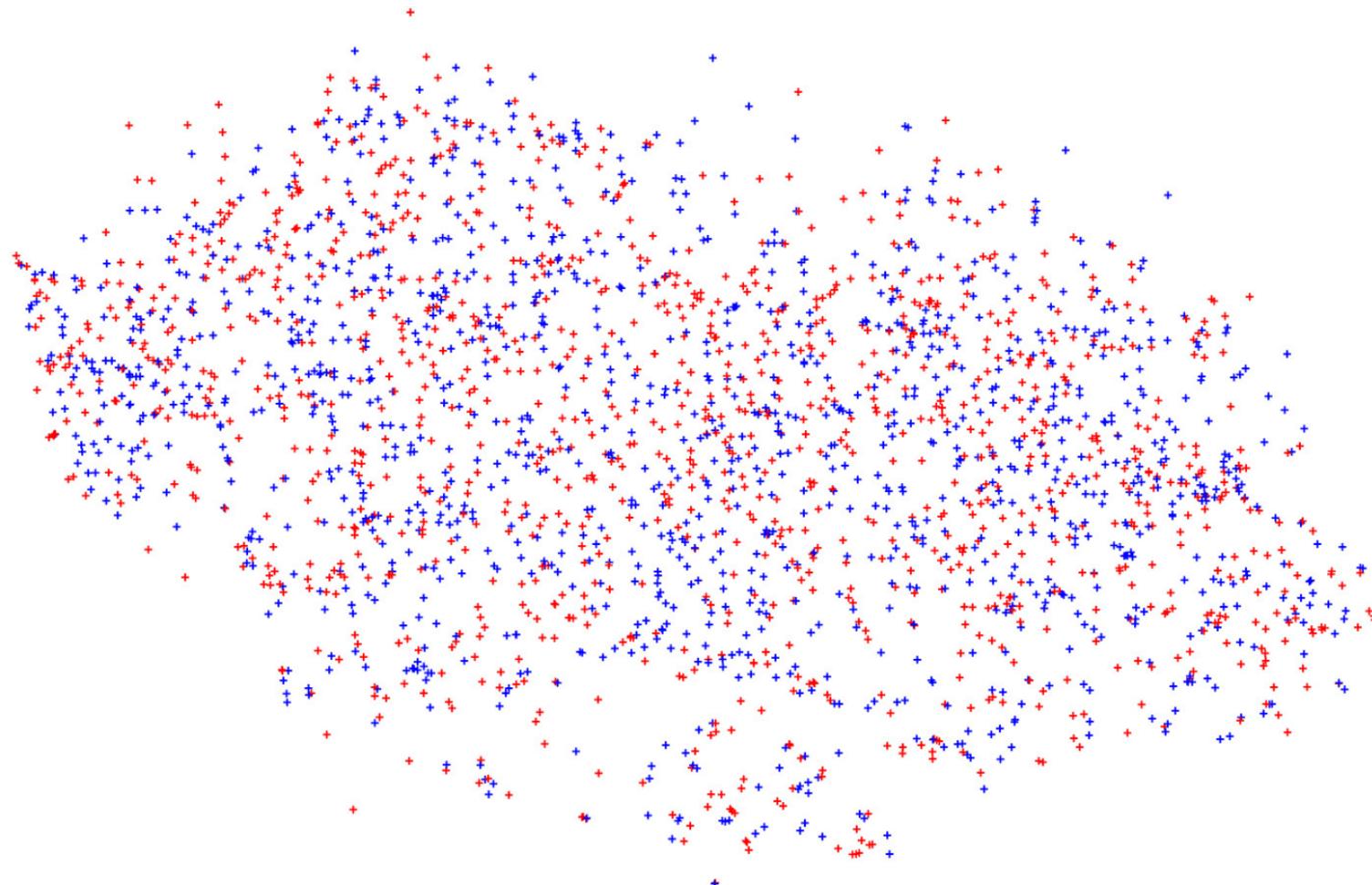


# Dogs vs. Cats



Deep Blue beat Kasparov at chess in 1997.  
Watson beat the brightest trivia minds at Jeopardy in 2011.  
Can you tell Fido from Mittens in 2013?

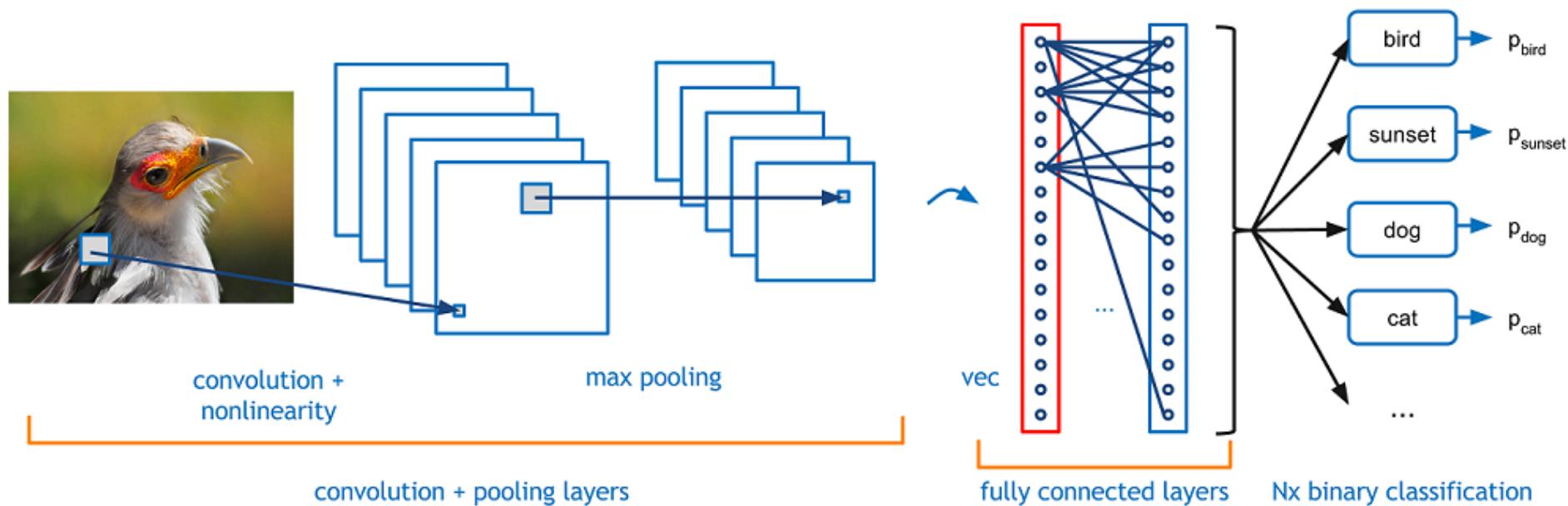
# Dogs vs. Cats



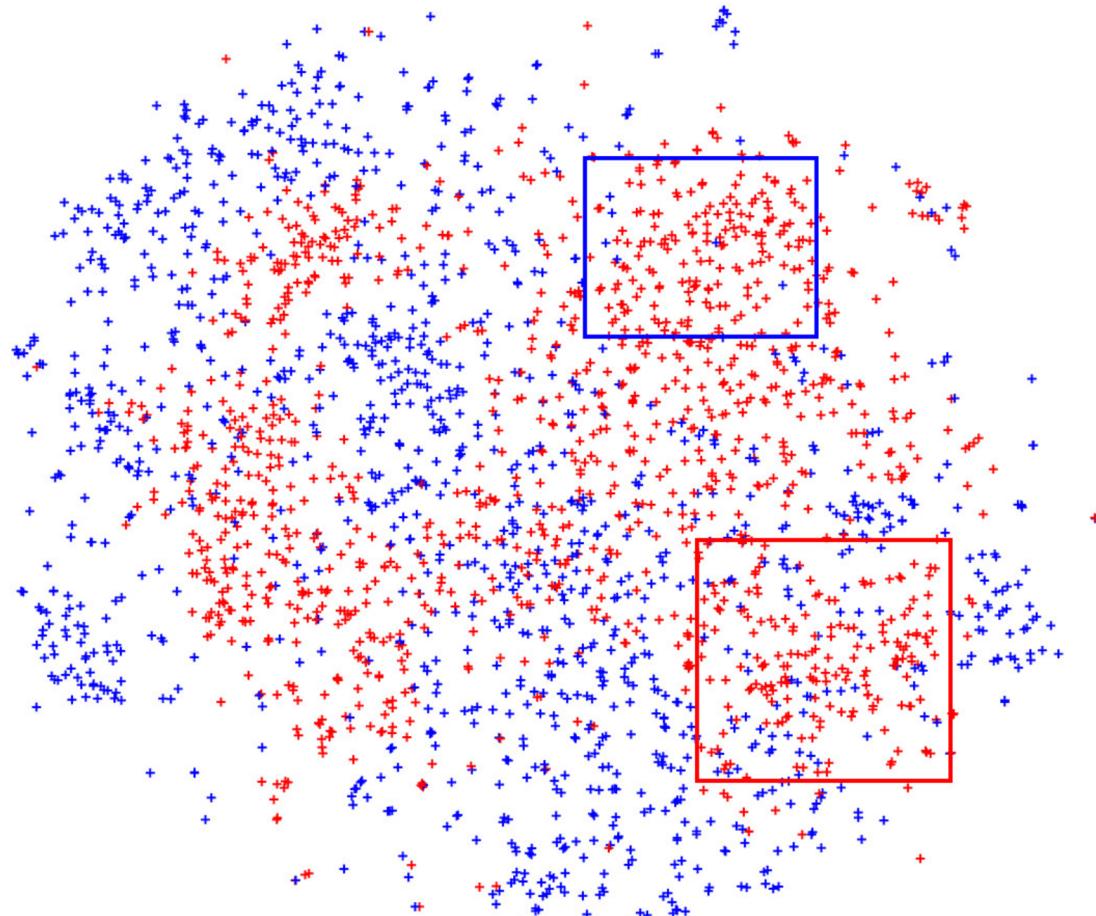
<https://indico.io/blog/visualizing-with-t-sne/>

# Dogs vs. Cats

- Визуализация не очень осмысленная
- Мы использовали интенсивности пикселей как признаки
- Современный подход — прогнать изображения через свёрточную нейронную сеть, взять выходы одного из последних слоёв



# Dogs vs. Cats



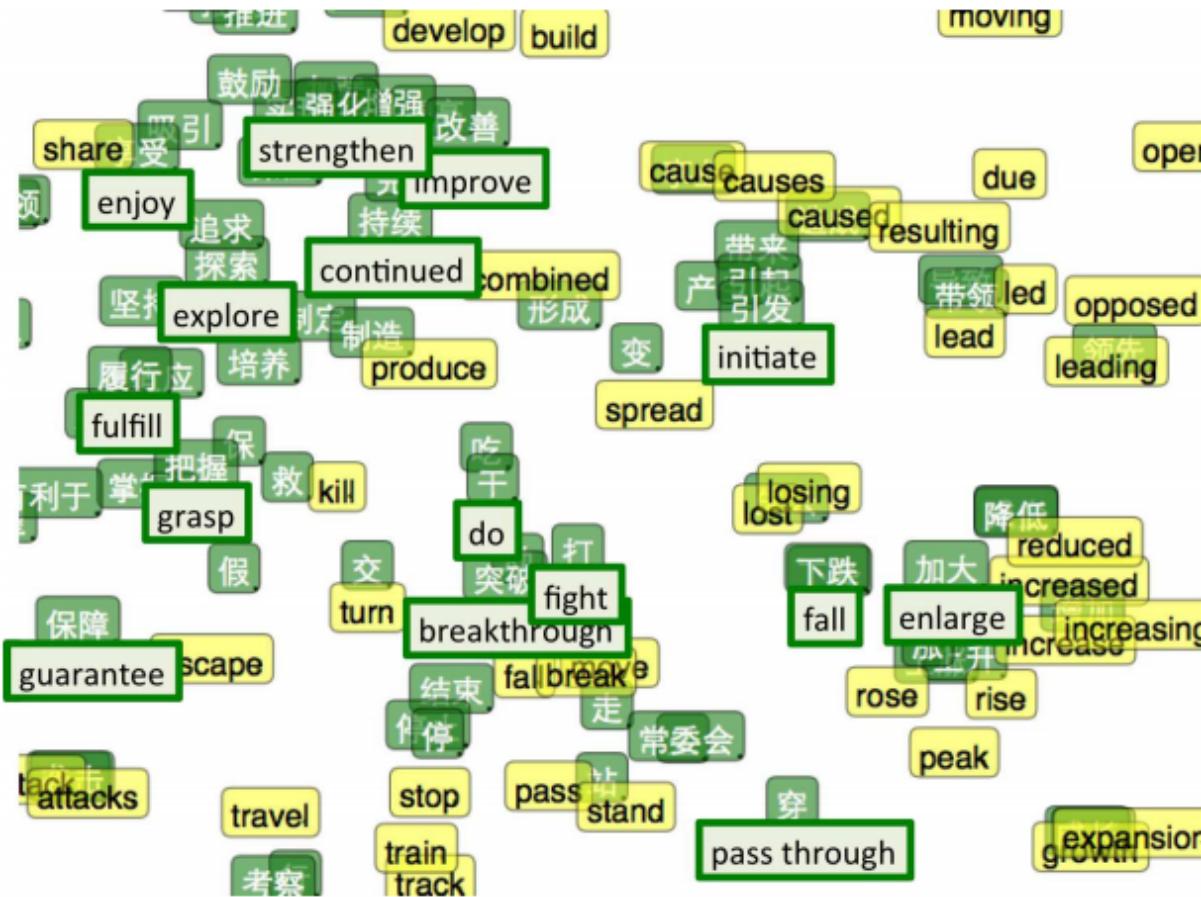
<https://indico.io/blog/visualizing-with-t-sne/>

# Dogs vs. Cats



<https://indico.io/blog/visualizing-with-t-sne/>

# Визуализация слов с помощью t-SNE



# Резюме

- Композиции алгоритмов улучшают качество
- Методы понижения размерности позволяют убрать неинформативные признаки и ускорить работу над моделями
- Классы методов: отбор признаков и извлечение признаков
- Отбор признаков: фильтрация и использование моделей
- Извлечение признаков: PCA
- Визуализация данных: t-SNE