IBMC
institute
of biomedical
chemistry

Bioinformatics.ru

Анализ биомедицинских данных

Михаил Пятницкий

PERSONAL BIOMEDICINE
YOUR FUTURE HEALTH

Российский национальный исследовательский медицинский университет имени Н.И. Пирогова

Данные “-омного” эксперимента

• многомерные

• имеются сложные взаимосвязи

• число переменных много больше числа образцов. Это очень затрудняет работу многих статистических методов

Типичные задачи эксперимента:

• Сравнение профилей экспрессии генов/белков между разными группами (ex: рак/норма)

• Изучение взаимосвязей между генами/`белками.

Типичные проблемы:

• Наличие пропущенных значений

• Наличие выбросов

Первым делом: разведочный анализ

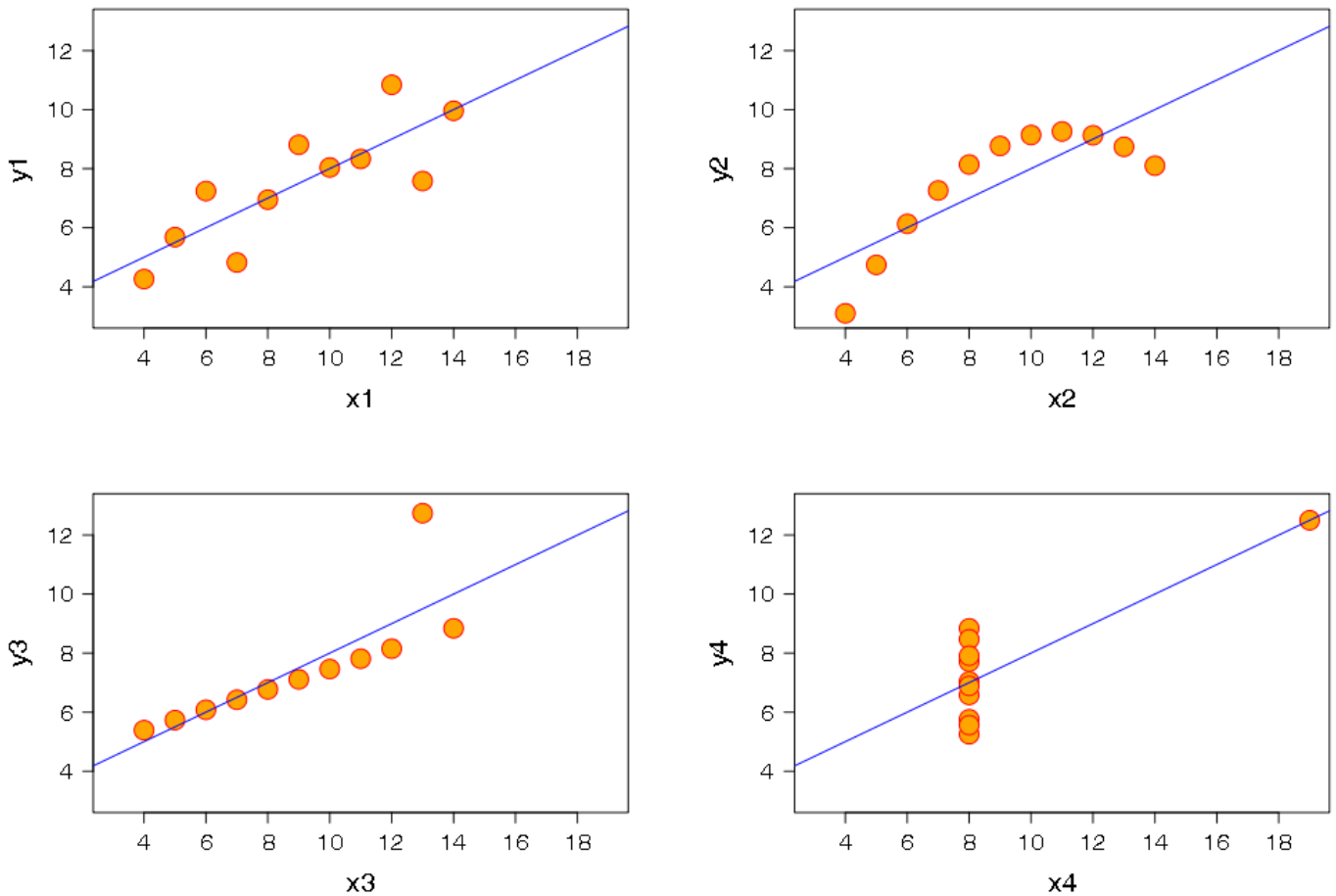
Задача: получить общее представление о данных

Не формулируем гипотез, пусть данные “говорят за себя”

- Понять сколько переменных/образцов/классов
- Посмотреть структуру и паттерны (нет ли грубых ошибок)
- Выявить аномальные наблюдения (выбросы)
- Оценить необходимость трансформаций данных

Упор на методы визуализации!

Необходимость визуализации: Anscombe's quartet



- Среднее: для x = 9, для y=7.5
- Дисперсия: x =11, для y = 4.1
- Коэффициент корреляции: 0.816
- Линейная регрессия: y = 3 + 0.5x

верно для всех(!) случаев

Важно: сперва провести разведочный анализ

- распределение каждого признака (по возможности), “ящики-с-усами”, описательная статистика

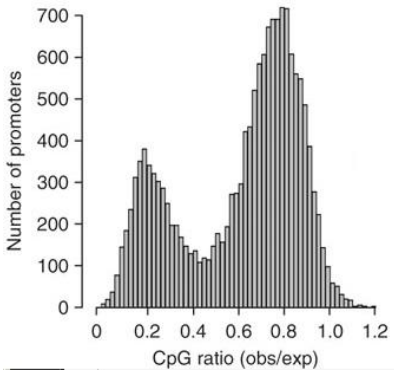
- визуализация на плоскости (PCA, MDS)

- определение необходимости трансформаций.

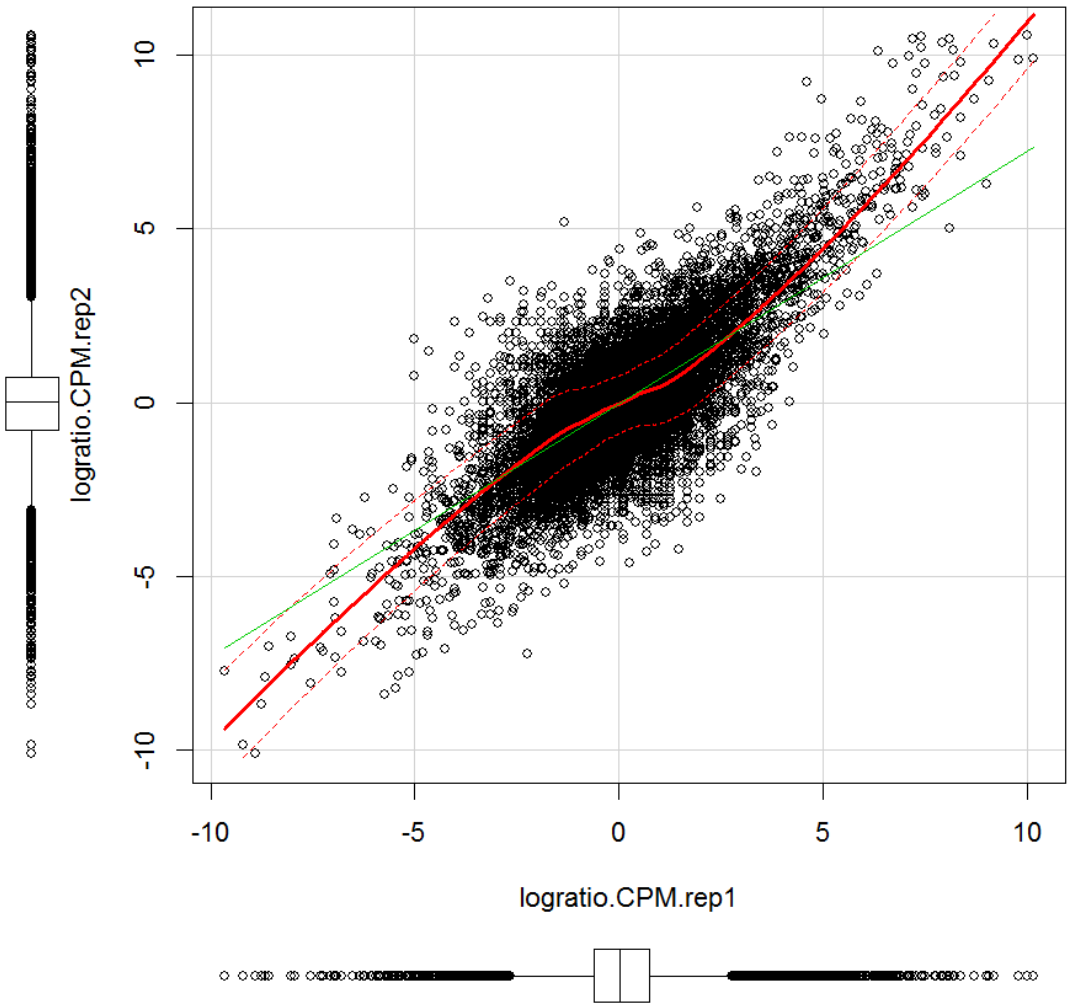
  часто: логарифмирование или z-score: $z = \frac{x - \mu}{\sigma}$

- поиск линейно-зависимых признаков

- поиск “интересных” признаков
  ex: бимодальных



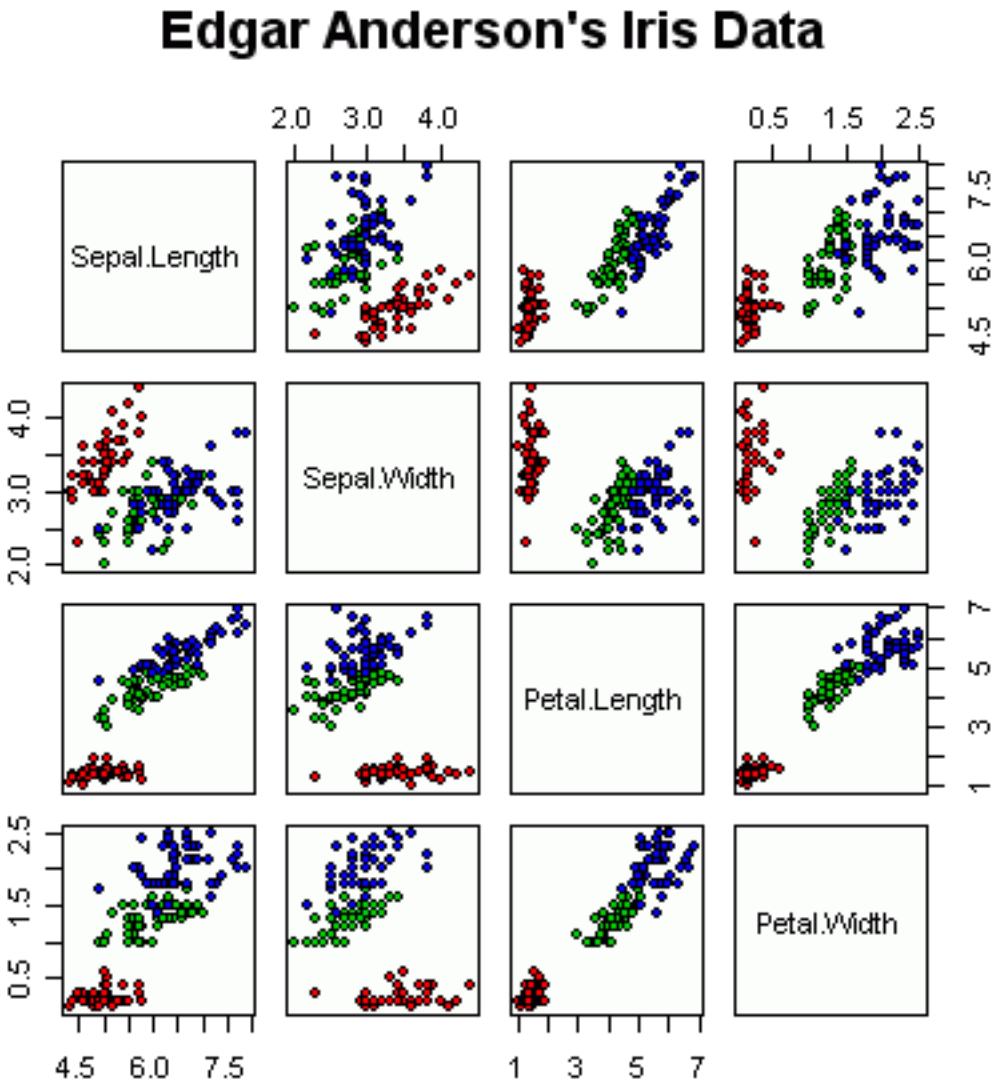
Графики рассеяния, scatterplots




```
library(car)
scatterplot()
```

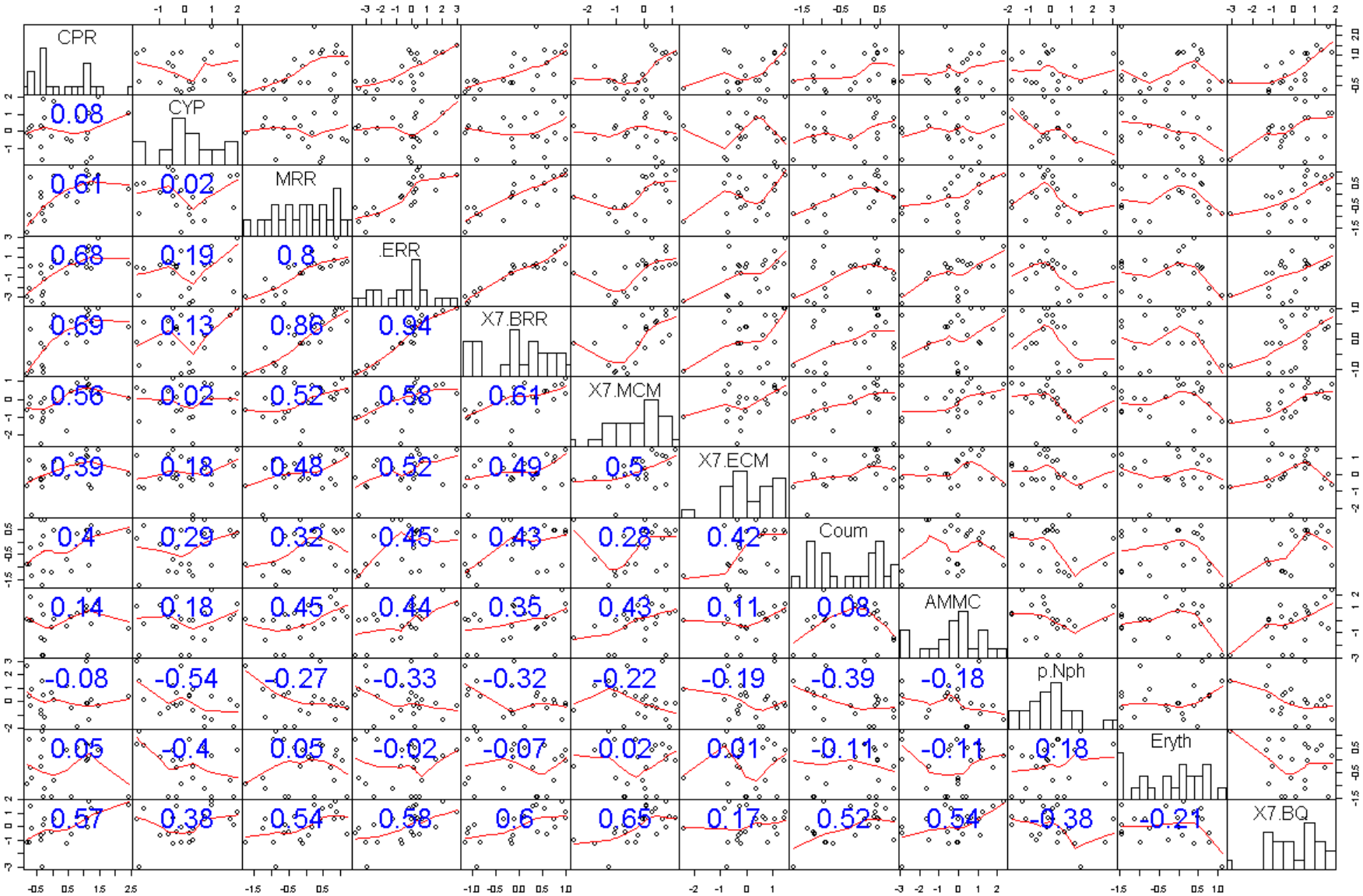
Графики рассеяния, scatterplots

Хорошо когда переменных не много (десятки)



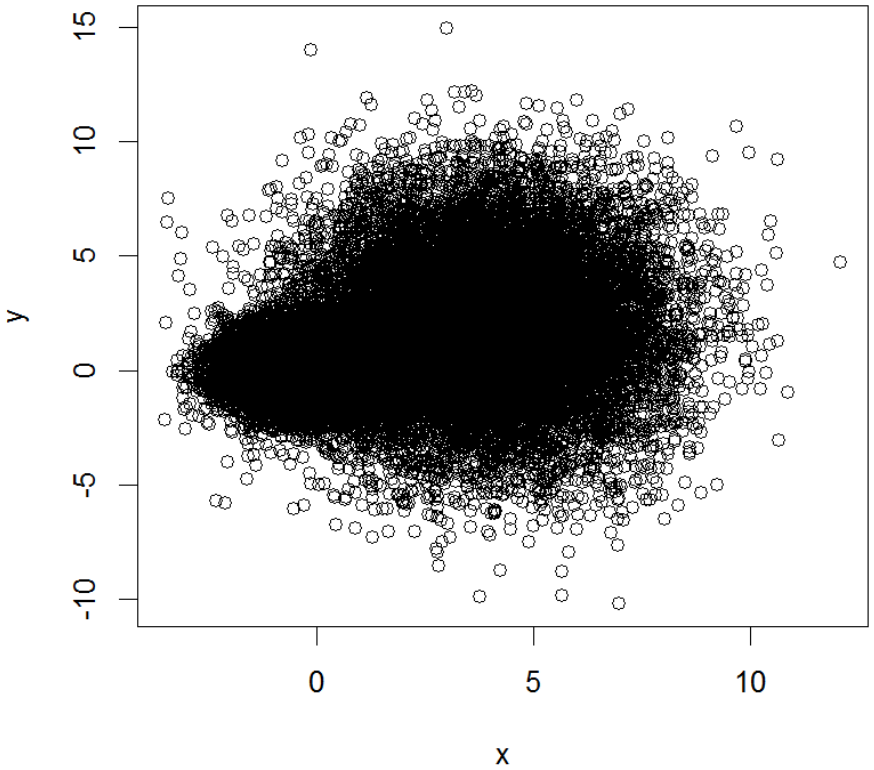

```
pairs()
```

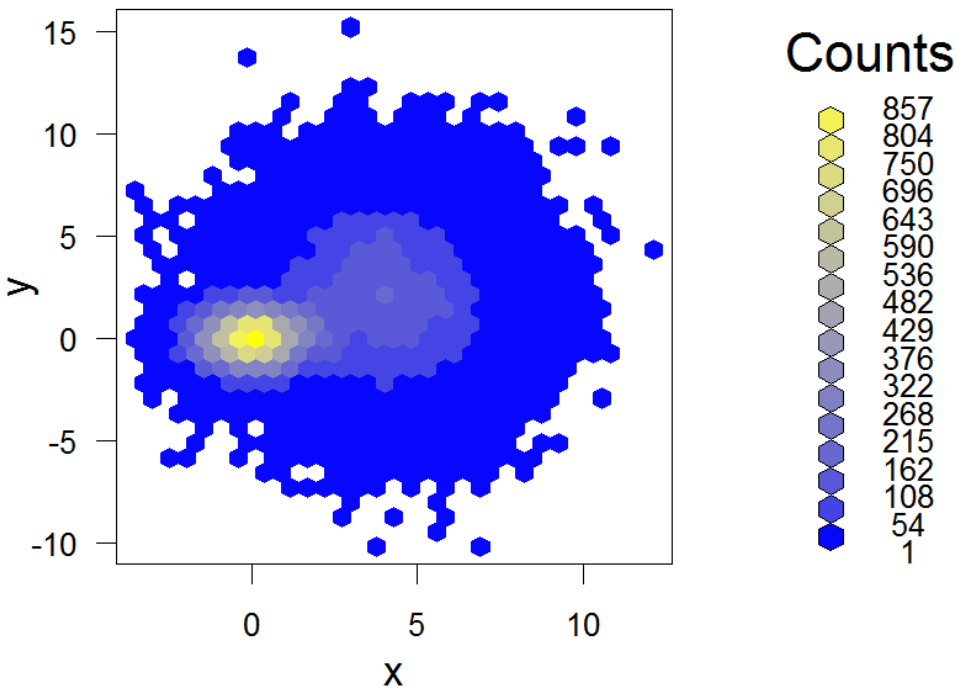
Графики рассеяния, scatterplots



matplot()

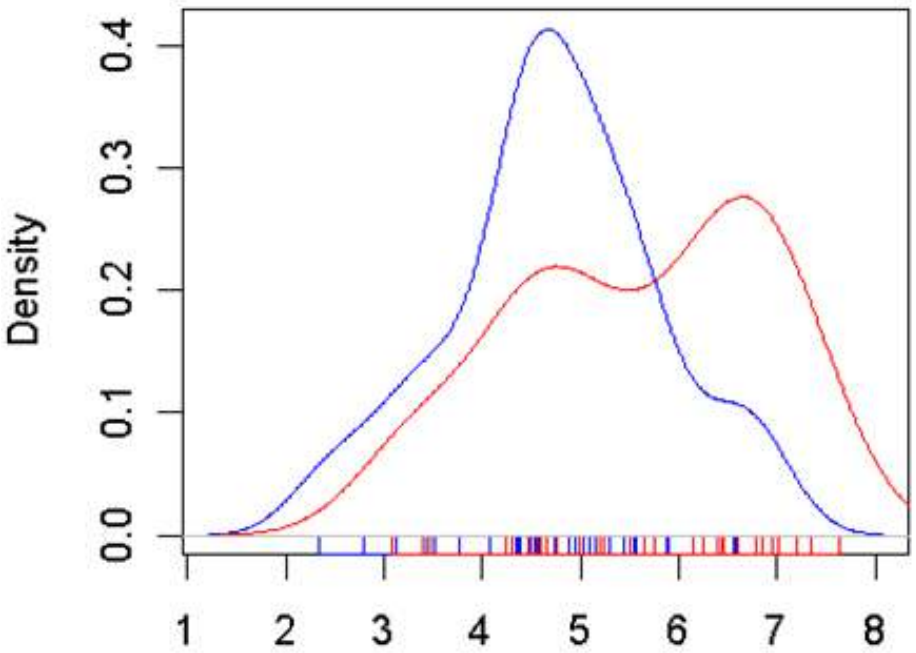
Что если точек много?

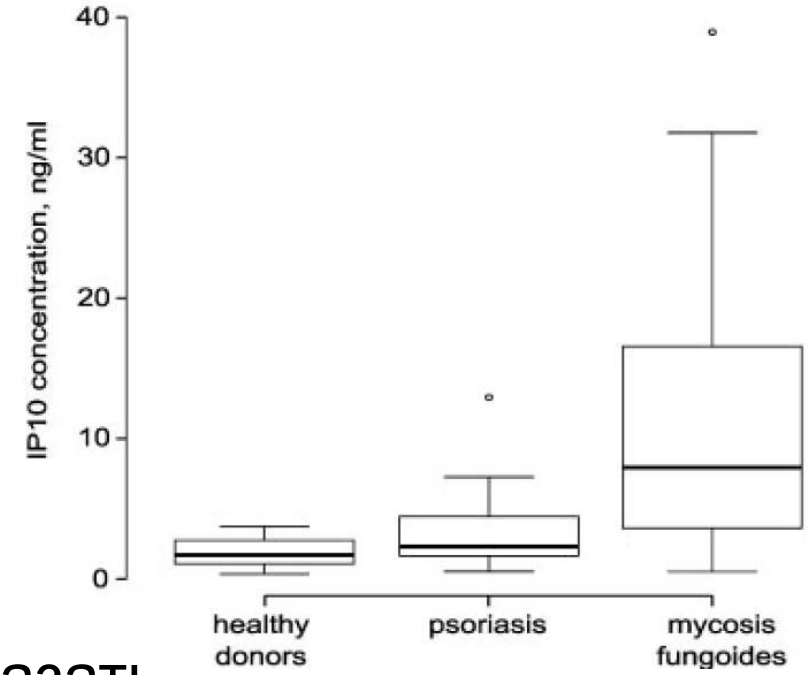





```
library(hexbin)

hexbin()
```

Анализ каждой переменной – гистограммы для каждого класса, boxplot (“ящик-с-усами”)





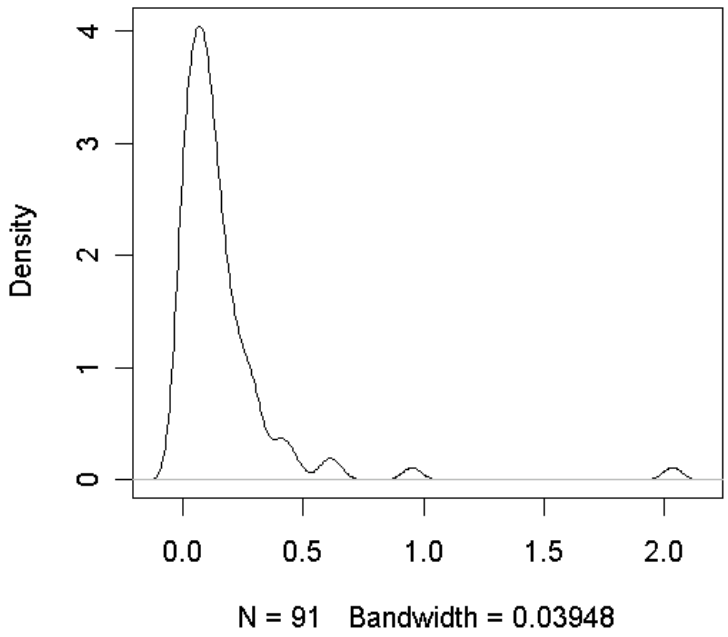
`density()`

`rug()`

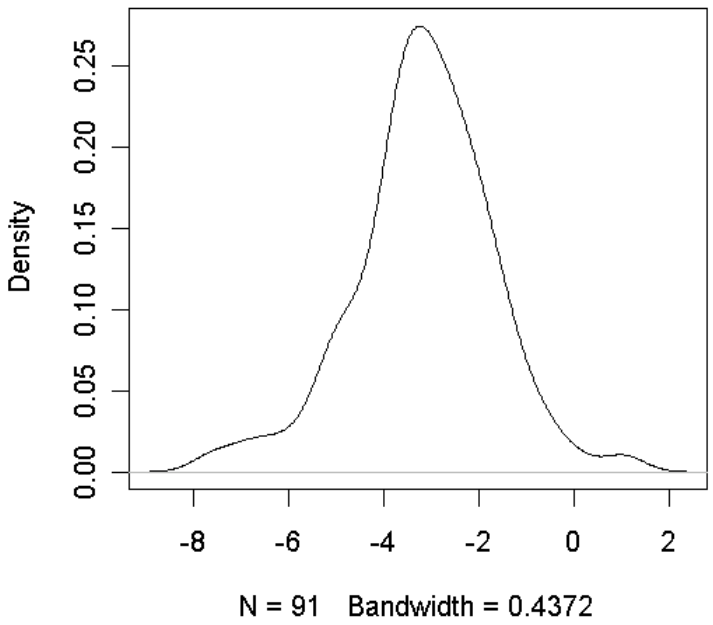
Q: что можем сказать по этому графику?

`boxplot()`

Часто требуется логарифмирование

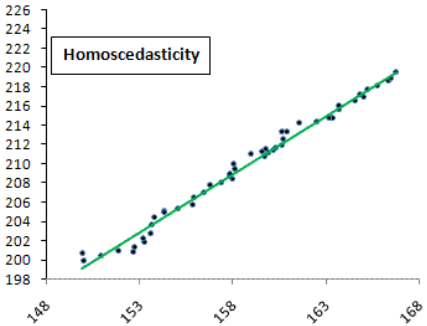


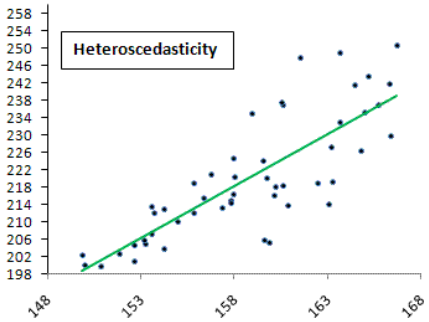
log



Многим статистическим методам нужно:

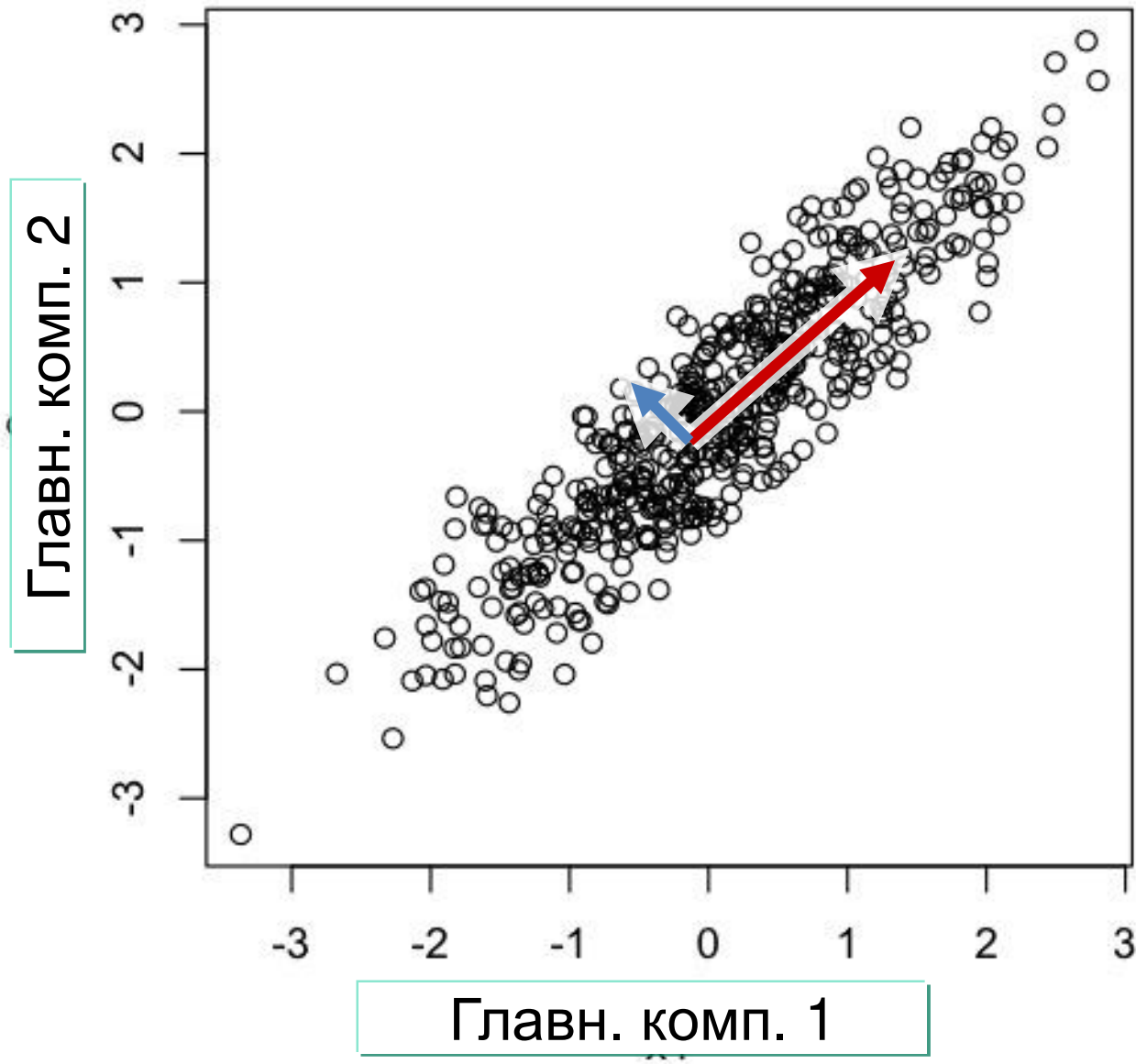
- нормальное распределение
- гомоскедастичность (разброс не зависит от среднего)





Анализ главных компонент, PCA

- Смысл – переход к новым координатам, которые содержат основную информацию
- Позволяет уменьшить кол-во признаков (обычно до 2-3)
- Быстро посмотреть на зависимости в данных
- Переменные должны быть измерены в единой шкале



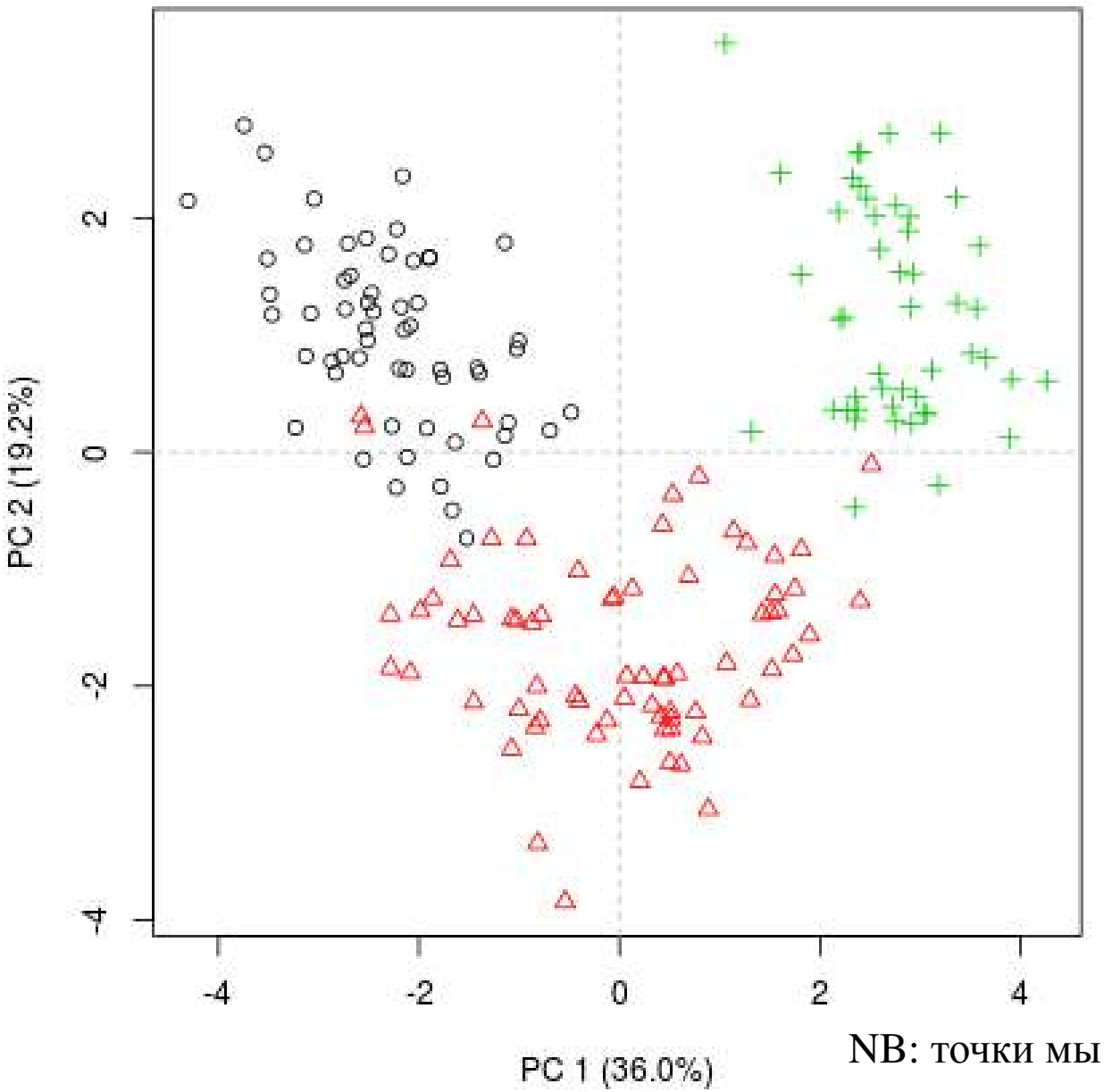
Что можно сказать про эти данные?

RGui (64-bit) - [Data Editor]

File Windows Edit Help

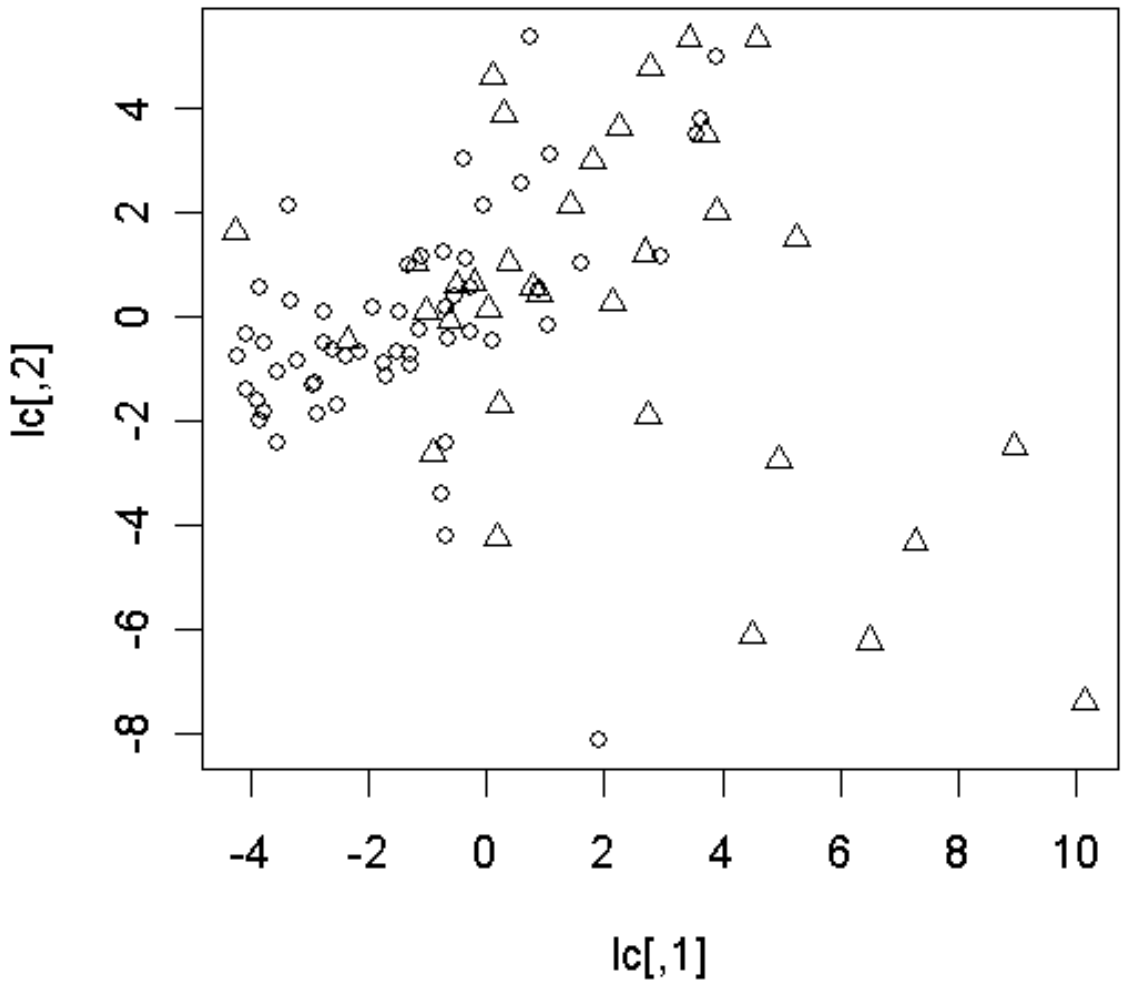
	GSM152678	GSM88980	GSM179854	GSM53099	GSM46816	GSM117584	GSM46933	GSM138028	GSM231976	GSM179837	GSM102525	GSM179791	GSM231919	GSM137936
1	4257.5	2952.9	4635.1	3860.6	2634.7	4670.4	7840.3	2562.5	2892.6	5126.9	3588.4	3227.1	4113.1	3317.1
2	447.1	382.2	983.2	568.4	464.1	294.8	709.3	835.8	528.9	1412.4	850	342.6	699.2	1113.9
3	314.2	632.5	520.7	227.4	619	192.1	628.5	571.7	539.9	1488	329.2	210.1	338.6	261.4
4	1034.5	736.6	649.2	694.3	526.6	821.4	551	1773.8	626.2	860.9	552.2	1076.7	771.6	958.3
5	62	63.1	47.3	10	36	13.7	31.6	112.4	98.7	82.5	122	93.7	36.4	98.2
6	746.5	639.5	845.2	596.9	942	1061	862.1	838.9	590.7	744.7	450.4	978.2	562.3	497.8
7	357.2	243.9	169.6	271.1	95	528.5	16.6	271	262.6	317.8	397.3	312.5	232.1	297.1
8	187.2	95	156.9	13.5	12.4	37.4	48.2	55.8	155.1	54.9	153.3	152	47.3	32
9	275.5	1887	5102.9	98.5	3235.5	2280.4	517.6	421.4	258.9	1835.6	202	1034.9	1214.1	705
10	234.8	80.9	142.3	84.2	52.6	46.4	122.5	95.6	207.6	116.7	100	114.7	167.7	59.7
11	342.5	468.6	445.1	1270.9	233.4	241.3	496.4	382.1	390.7	1482.7	751.5	302.6	530.7	740.2
12	593.2	559.6	839.6	1153.3	658.9	307.8	829.1	898.9	443.9	1141.4	458.5	659.7	482.2	714.3
13	439.9	144.3	214.2	232.5	74.4	248.1	50.5	370.7	161.7	196.7	280.3	236.4	83.9	286.5
14	626.2	924.5	2066.5	2886.9	1268.2	576.3	4528.7	3356	487.4	5177.5	943.8	1095.2	1084.3	1925.2
15	467	1474	933	2091.5	1718.1	886.2	1763.4	1380.3	905.8	2087.9	1028.4	960.3	873.7	2866.9
16	314.2	292.4	150.7	180.9	187.4	419.7	241.1	404.6	445.8	275.4	289.1	516.1	111.9	272.2
17	336.7	93.7	71.8	84	61.1	71.5	177.1	407	241.9	34.3	191.3	175.4	126.6	89.3
18	173.2	339.5	271.7	202.6	279.6	350.1	179.4	145.1	253.4	177.3	253.4	225.9	300.6	258.8
19	542.7	621.7	522.1	535.8	354.1	688.1	302	221	331.8	531.1	502.1	585.7	436.7	750.2
20	148.7	34.8	113.1	14.8	116.8	124.1	184.3	148.9	150	155.2	140.1	43	67.1	110.2
21	94.1	81.8	14.5	49.7	12.5	102.6	44.8	175.3	9.1	32	34.6	139.8	192.4	63.3
22	73.9	22.1	11.6	39.7	22.1	55.7	14.5	174.5	28.3	40.1	39.8	37.4	39.6	34.8
23	206.5	122.2	168.1	188	97.2	65.5	185.5	329.3	166.1	287.1	102.2	129.6	163.8	231.4
24	602.2	456.7	464.4	775.5	48.9	771	55.6	60.7	624.1	332	462.8	496.7	625.3	516.6
25	13.3	74.1	205.1	255.2	147.9	90.6	97.5	72.6	303.1	283.9	189.3	362.3	236.5	264.3
26	110.2	152.2	120.2	318.1	81.3	314.2	60.4	83	60.6	240.4	170	219	30.2	34.6
27	484.1	936.7	470	809.3	552.8	538.5	1017.5	797.8	369.2	1158.6	974.1	762.3	429.9	798
28	284.7	50.8	214.6	108.3	144.7	364	83.4	334.8	158	206	201.7	325.5	242.9	118.4
29	341.7	627.5	379.7	266	193.3	55.9	269.4	445.5	373.2	411.6	311.6	287.5	456.6	282.1
30	78.4	123.2	87.8	100.9	80	221.4	43.3	393.9	38.6	98.7	135.3	131.9	58.7	221
31	71.2	50.6	45.4	314.6	94.6	416.7	111.8	683.8	264.5	234.4	66.5	103.7	75.1	57.3
32	404	425.4	118.8	144.8	266.9	261.5	141	267.5	334.1	63.1	275	405.5	283	387

А теперь?



NB: точки мы сами раскрасили

Многомерное шкалирование

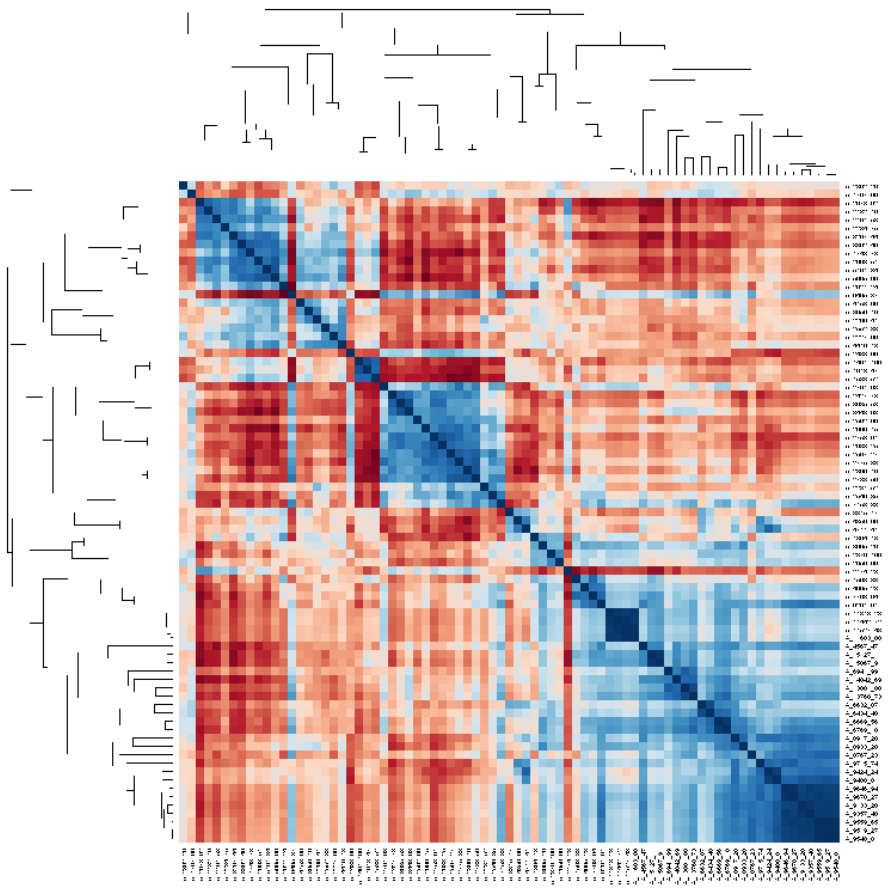


данные С.А.Мошковского

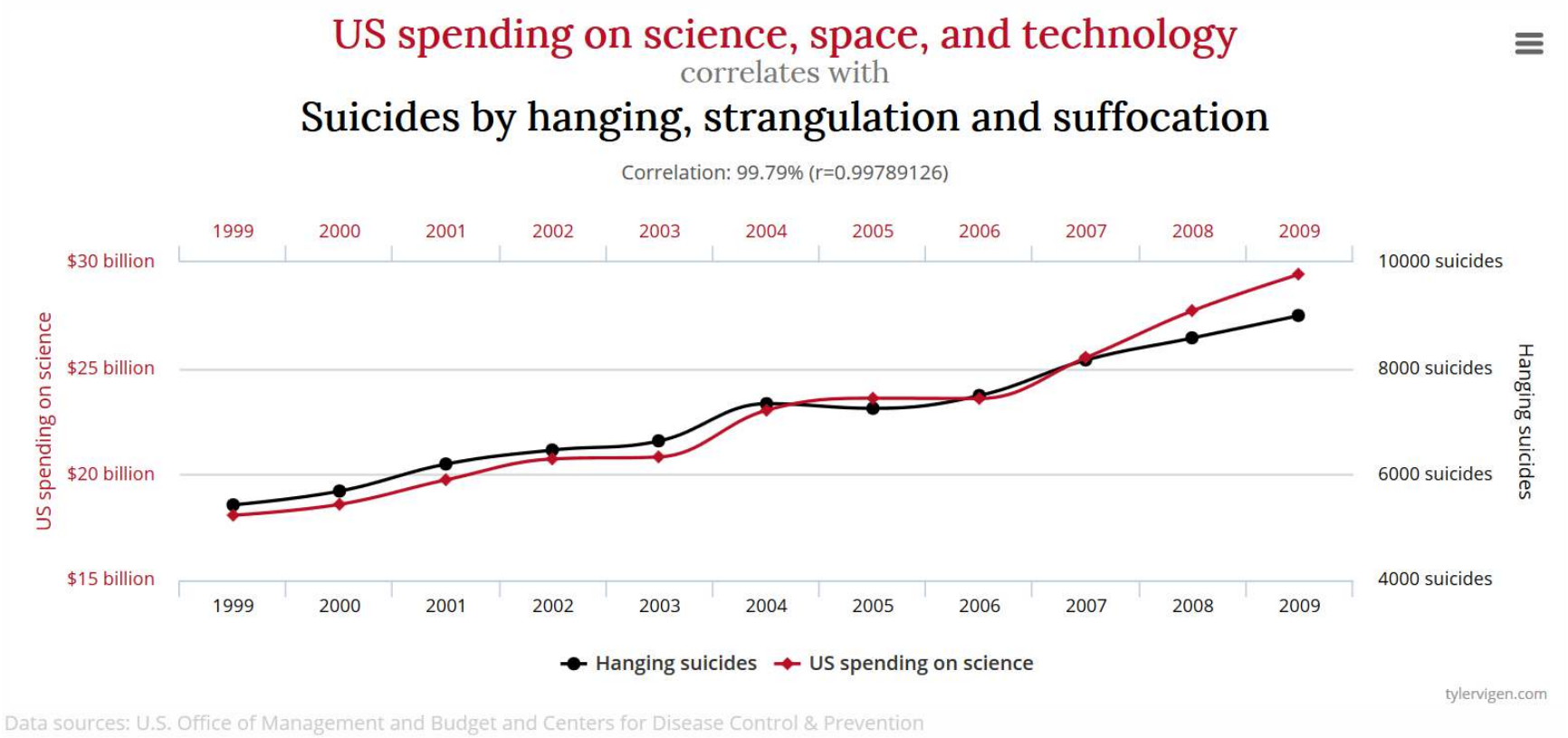
```
cmdscale()

isoMDS()
```

Поиск корреляций между переменными

Выкидываем линейно-зависимые



Не увлекайтесь поиском корреляций



Корреляция не подразумевает причинной связи

Проблема пропущенных значений

Часто встречается в протеомных экспериментах (2-DE)

Многие методы (особенно многомерные) требуют полных данных

Способы борьбы:

- убрать строки/столбцы с пропущенными значениями
- генерировать новые данные (заменять на среднее, получать из оцененного распределения)
- использовать методы нечувствительные к пропущенным данным
- поставить эксперимент еще раз!

Проблема выбросов

Выброс – образец с резко отличающимися показателями от основной группы

• может существенным образом повлиять на работу многих статистических процедур.

• трудно заметен в многомерных данных

Поиск выбросов

• вручную – анализ графиков рассеяния, PCA,

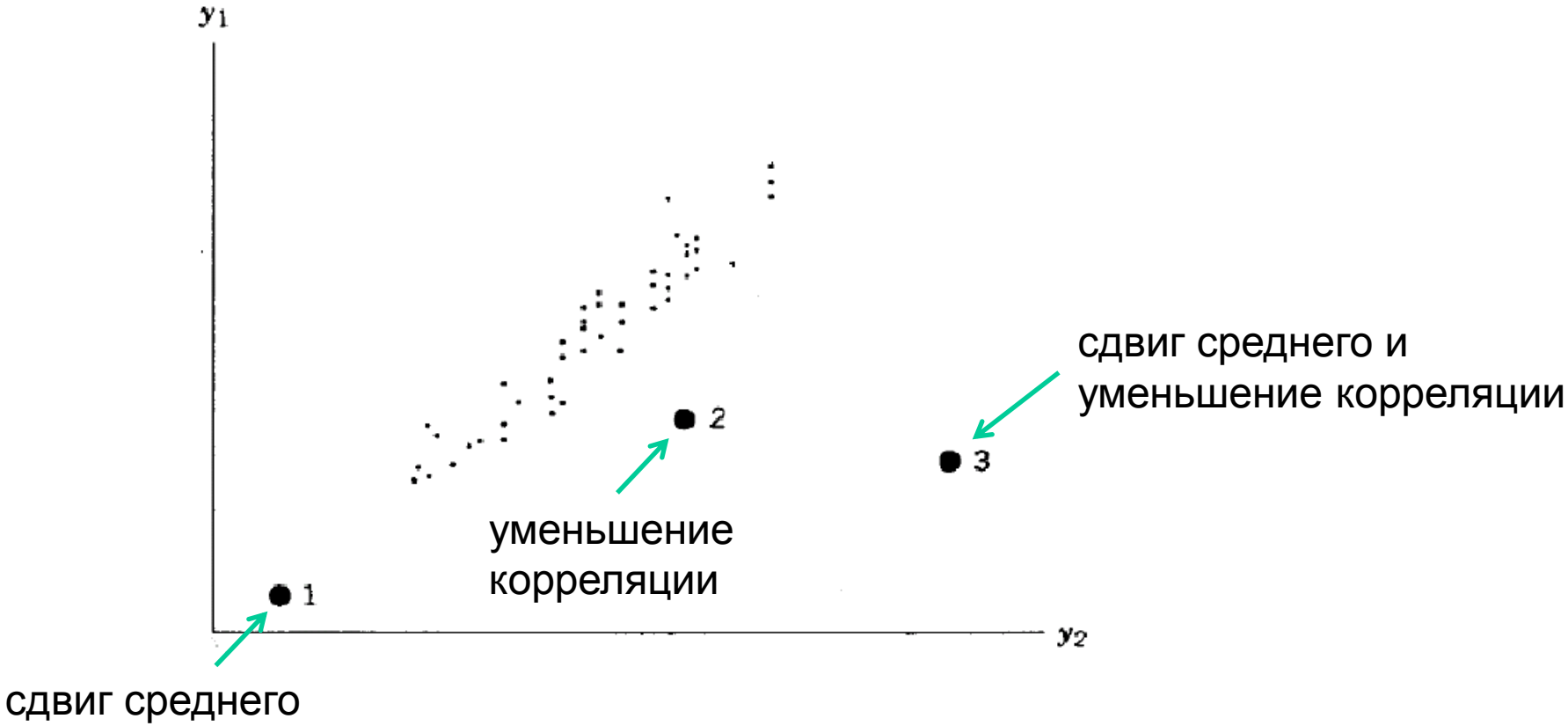
• критерии для обнаружения (Yang, Mardia, Schjwager,…)

Средства борьбы с выбросами

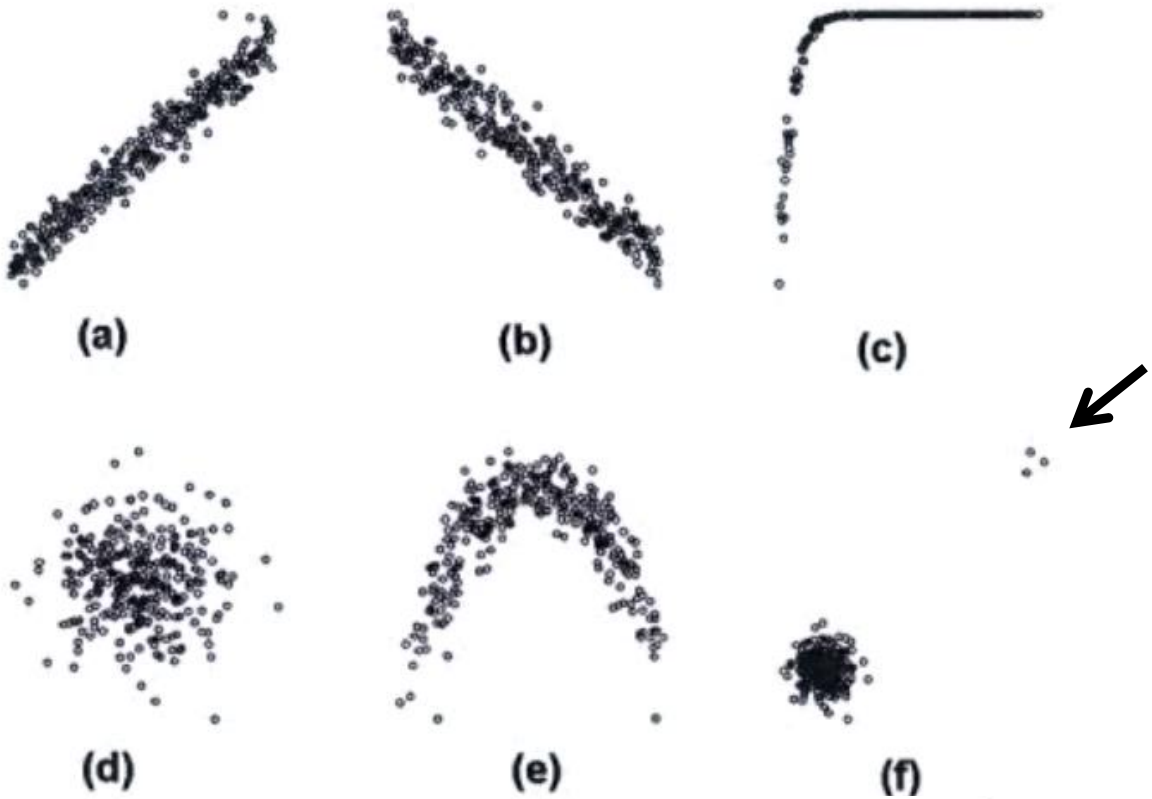
• удаление выбросов

• Гигиена: применять устойчивые к выбросам (робастные) статистически (ex: медиана вместо среднего, корреляция Спирмена вместо Пирсона)

Типы выбросов

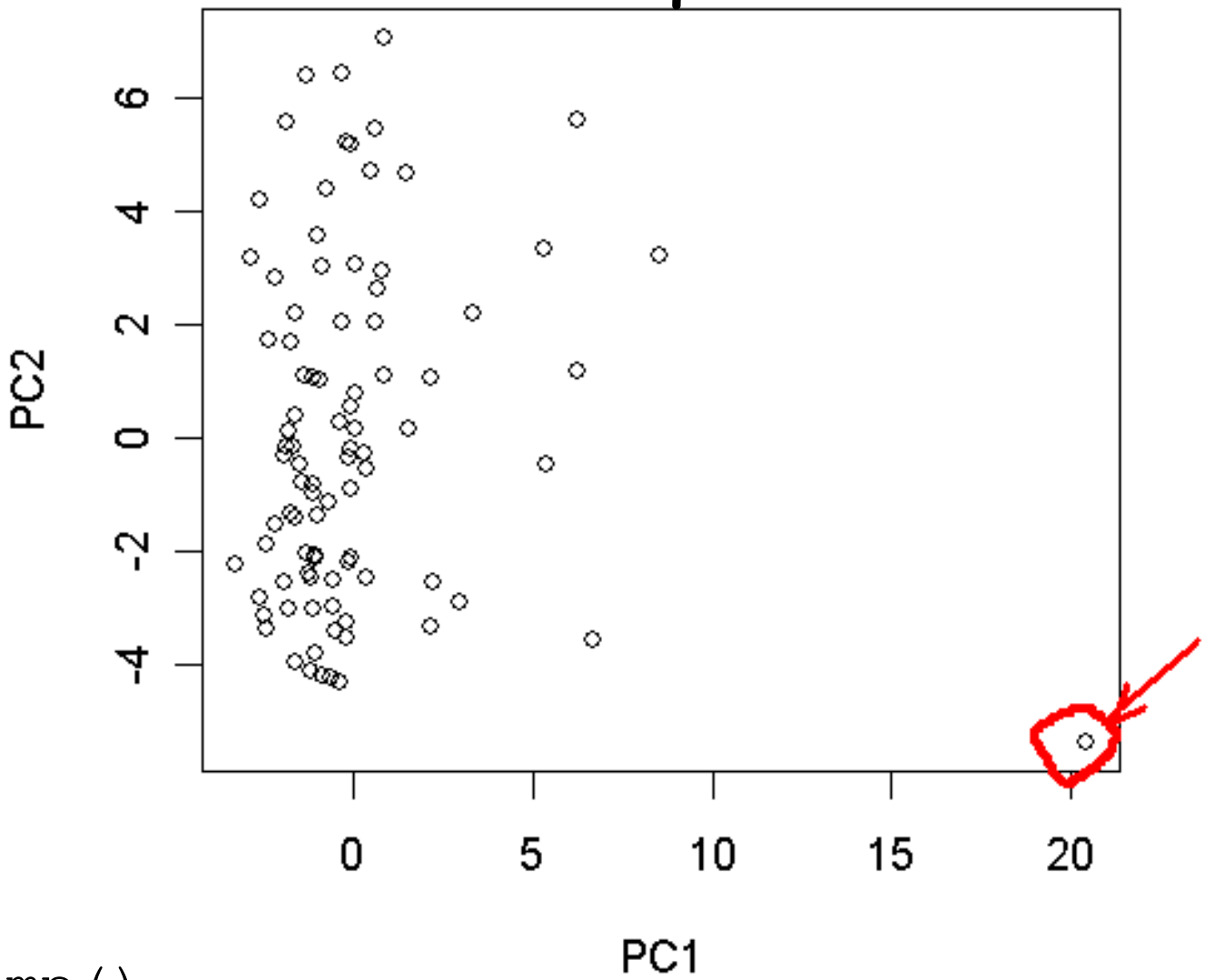


Проблема выбросов: иллюстрация на примере коэффициентов корреляций



Dataset	*Pearson's rho*	*Spearman's rho*	*Kendall's tau*
Figure 17.1(a)	0.98	0.98	0.87
Figure 17.1(b)	-0.98	-0.98	-0.87
Figure 17.1(c)	0.50	0.99	0.98
Figure 17.1(d)	-0.02	-0.03	-0.02
Figure 17.1(e)	-0.06	-0.02	-0.02
Figure 17.1(f)	0.68	0.00	0.00

Пример: обнаружение выбросов с помощью PCA

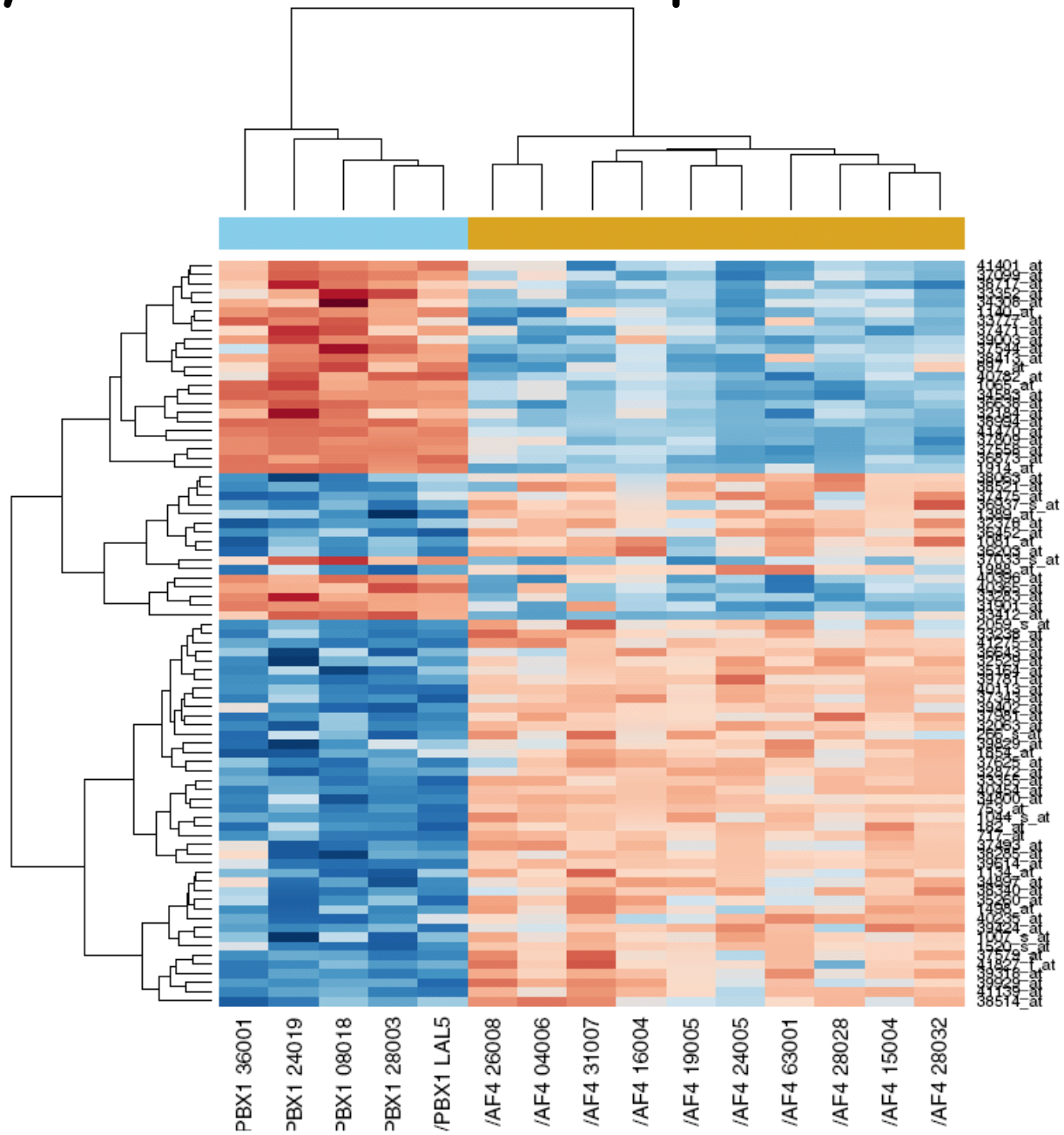


`prcomp()`

данные С.А.Мошковского

Визуализация: heatmap

Наглядно представляем взаимоотношения и переменных и образцов




```
library(gplots)

heatmap.2()
```

Машинное обучение

Машинное обучение изучает методы построения моделей и алгоритмов, способных обучаться на данных

Данные Алгоритм Модель

Calvin, I'm still confused about **cats** and **dogs**!

OK, then I will explain it once more ...

Задача: научить компьютер распознавать объекты различных классов.

Задача:
научить компьютер решать когнитивные задачи.

Дано:
- набор объектов
- признаки, характеризующие объект
- класс объекта (может быть неизвестен)

Вход: новый объект
Выход: предсказание класса для этого объекта

Ключевые слова:

- методы распознавания образов, машинное обучение, кластерный анализ, нейронные сети, дискриминантный анализ
- pattern recognition, machine learning, cluster analysis, neural networks, discriminant analysis

Примеры задач по распознаванию образов

- Распознавание рукописных букв (FineReader)
- Отпечатки пальцев, распознавание речи
- Распознавание лиц на фотографиях
- Постановка диагноза, анализ ЭКГ (медицина)
- Определение свой/чужой (армия)
- Робототехника
- Защита от спама

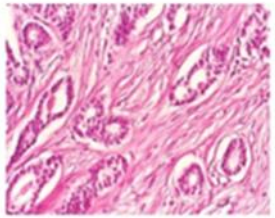
...

NB: полная автоматизация процесса распознавания!

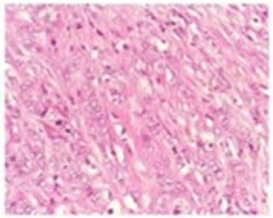
Пример: клинические решения по транскриптому

MammaPrint®

- Определение риска рецидива рака молочной железы в ближайшие 5 лет
- 70 генов, microarrray

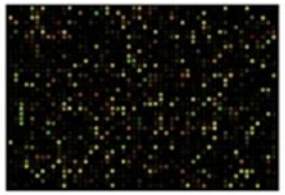


Low grade

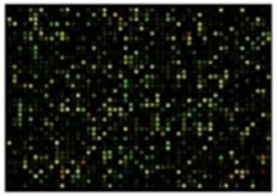


High grade

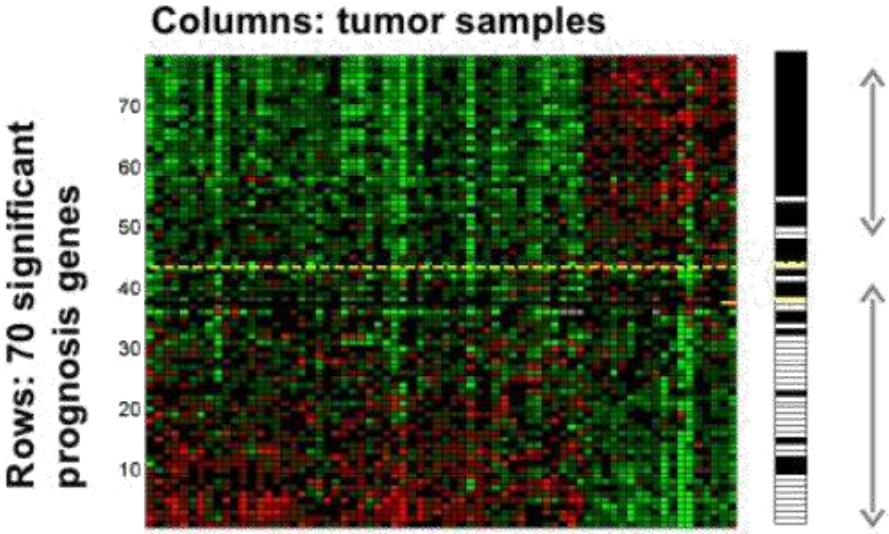
MammaPrint



Low risk



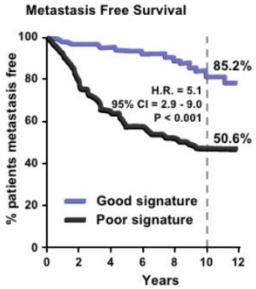
High risk

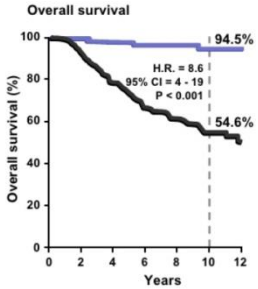


Good signature

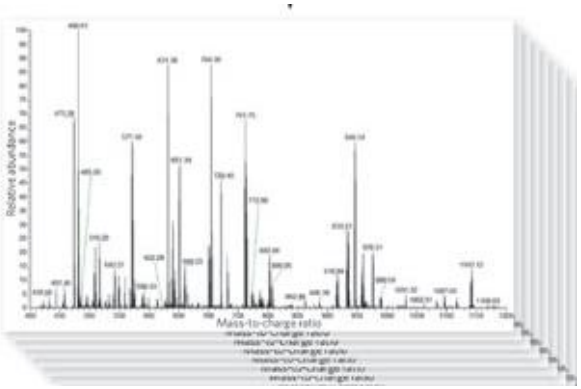
threshold

Poor signature

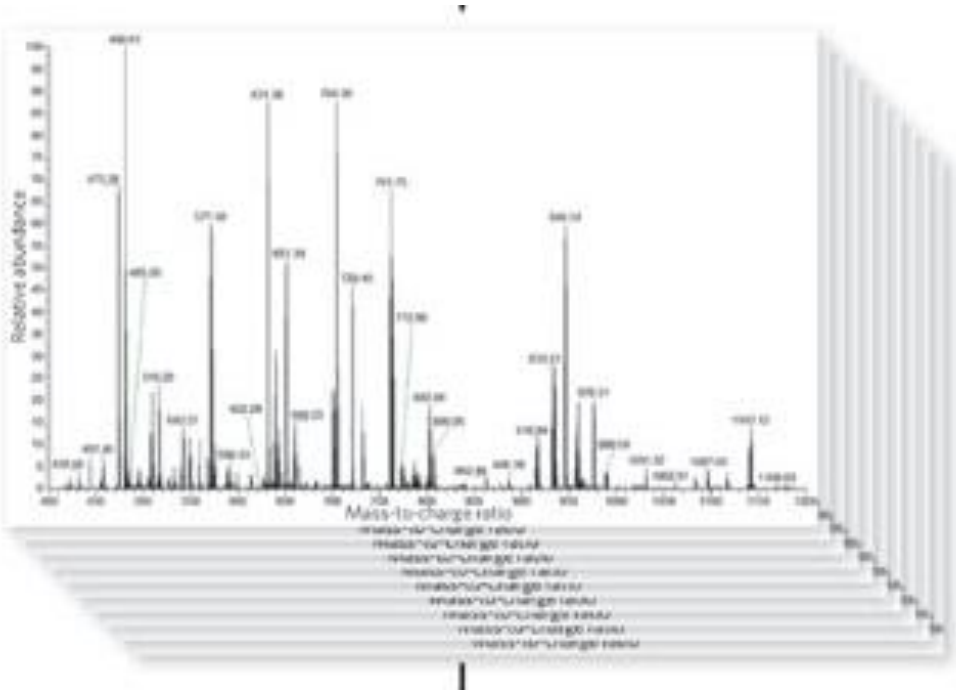




Пример: автоматическое выявление диагноза “рак яичника” по данным масс-спектрометрии



38 здоровых

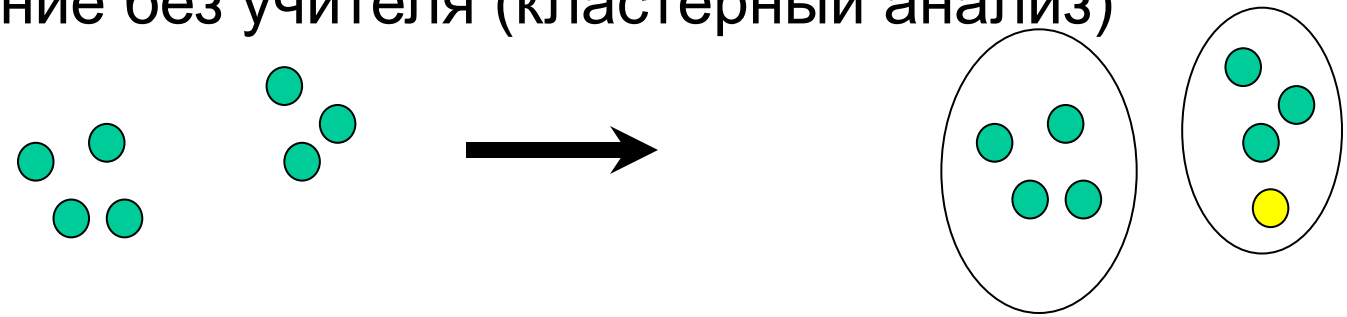


67 больных

Два типа задач (важно!)

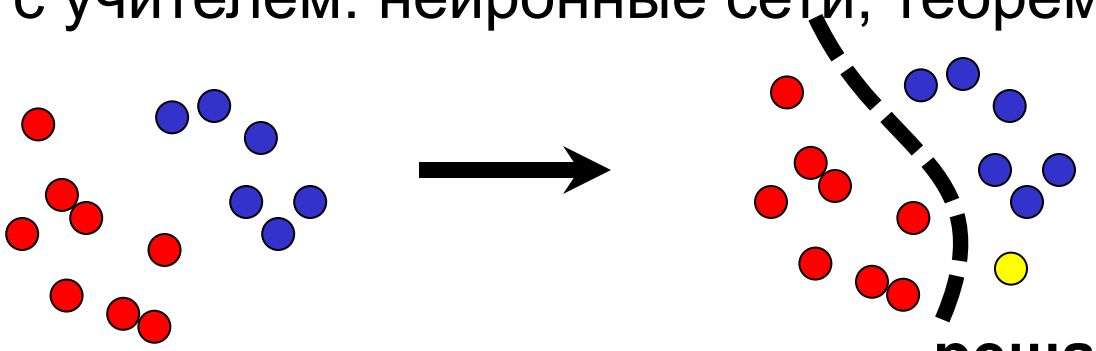
1)Есть набор объектов(образцов). Нужно сгруппировать похожие объекты между собой (если группировка осмысленна!)

Обучение без учителя (кластерный анализ)



2) Дополнительно есть информация о классах объектов. Нужно “обучиться” по данным предсказывать класс нового объекта

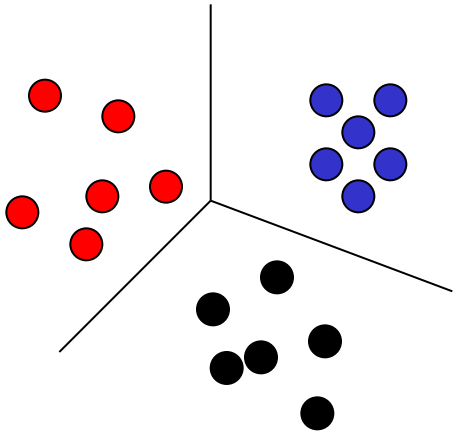
Обучение с учителем: нейронные сети, теорема Байеса, …



решающее правило

Кластеризация (обучение без учителя)

разделяющая

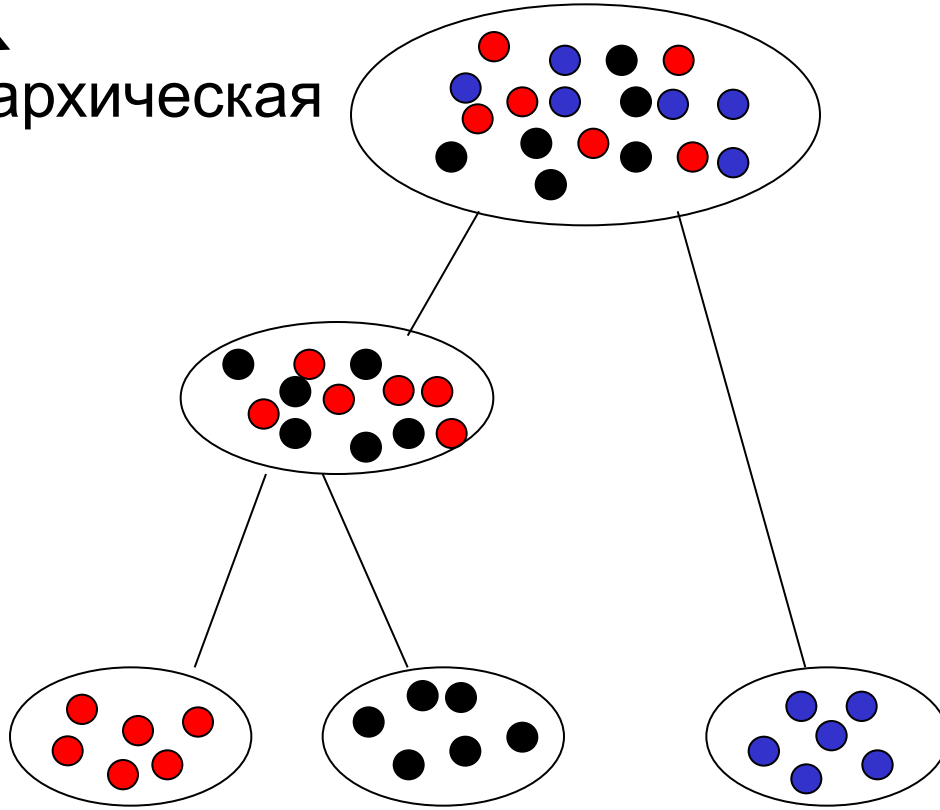


напрямую возвращает состав кластеров.

Число кластеров нужно задавать заранее.

Метод **<u>k-средних (k-means)</u>**, PAM, …

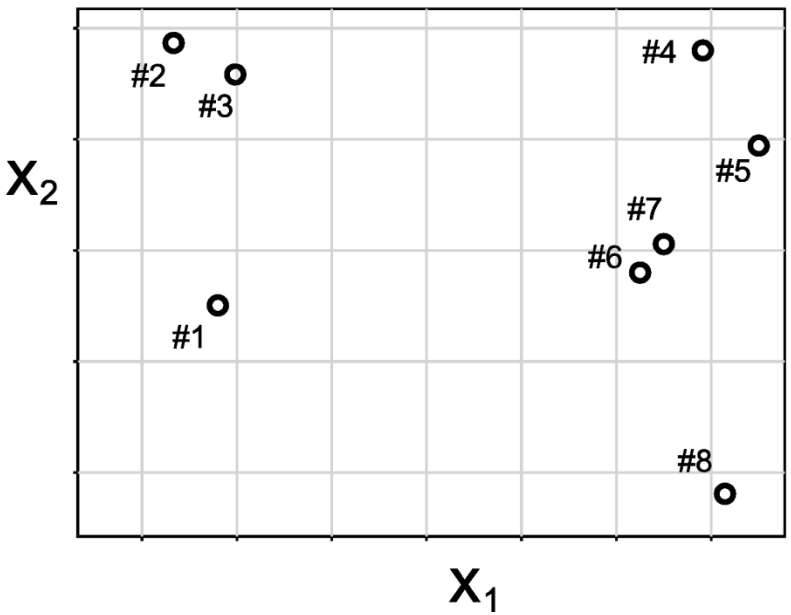
иерархическая



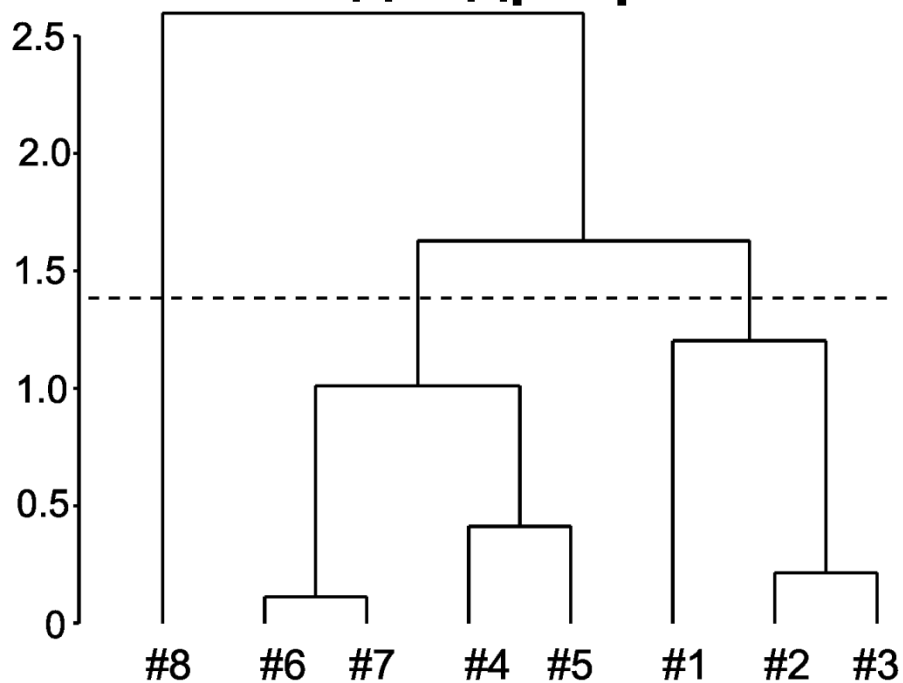
последовательное объединение (или разделение) объектов на различных уровнях сходства

<u>метод ближней связи</u>, метод полной связи, метод Уорда, …

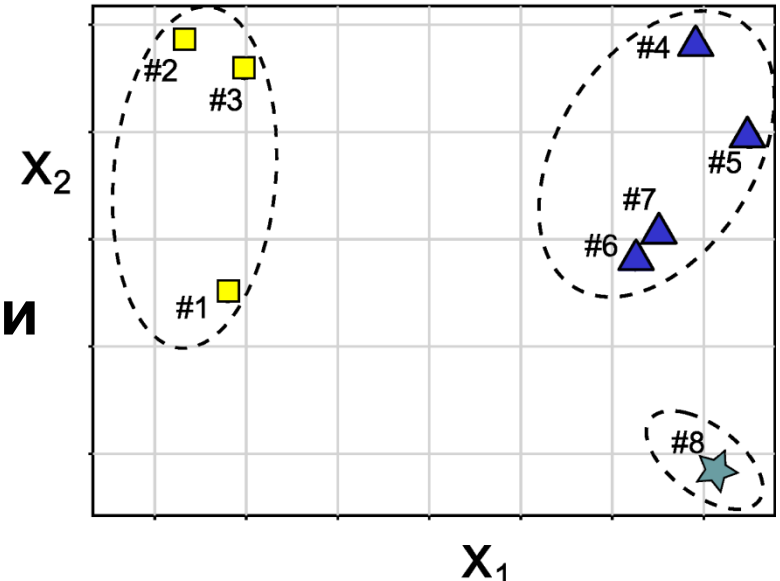
исх. данные



дендрограмма

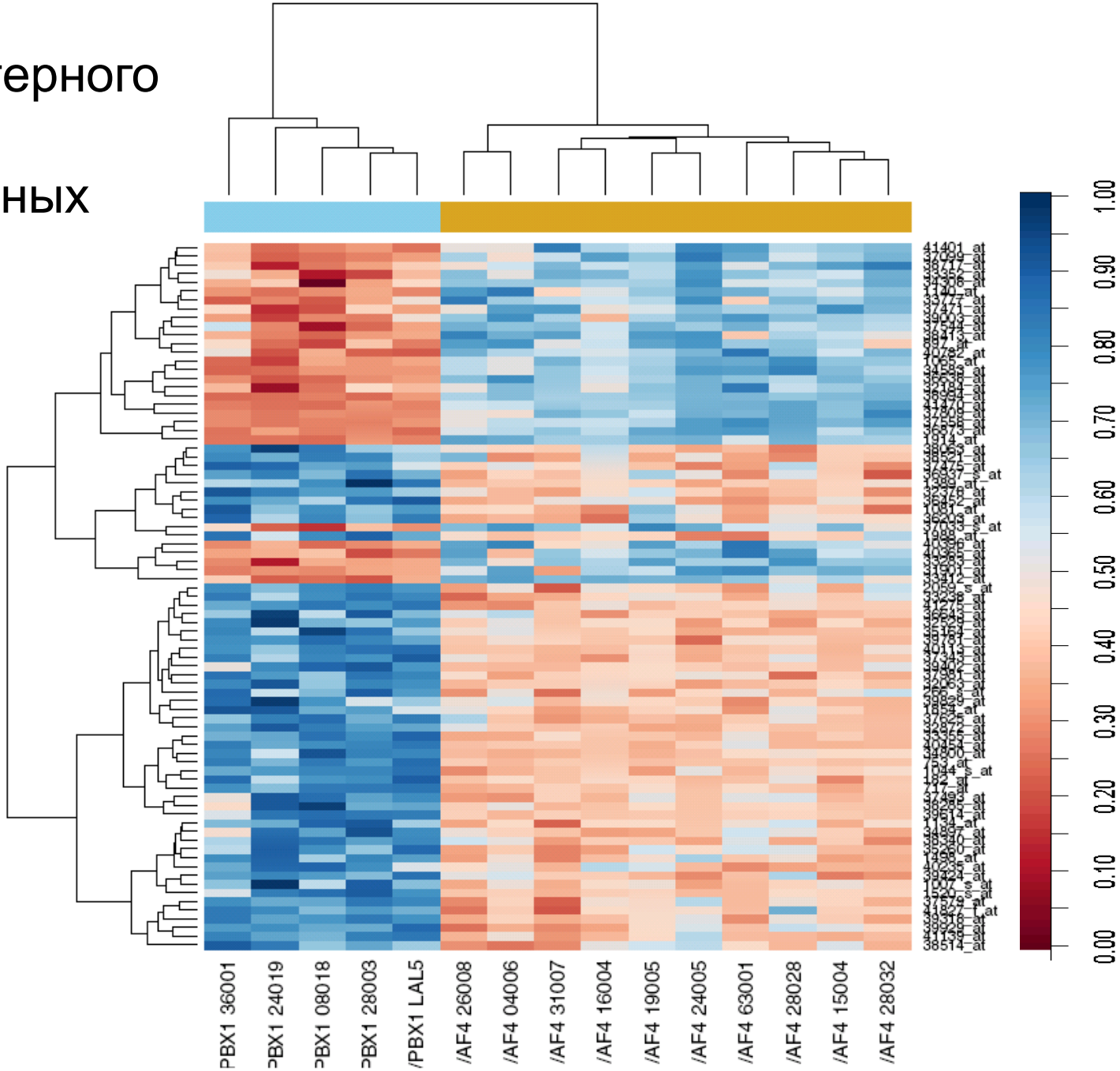


результат кластеризации



Пример кластерного анализа транскриптомных данных

(heatmap)

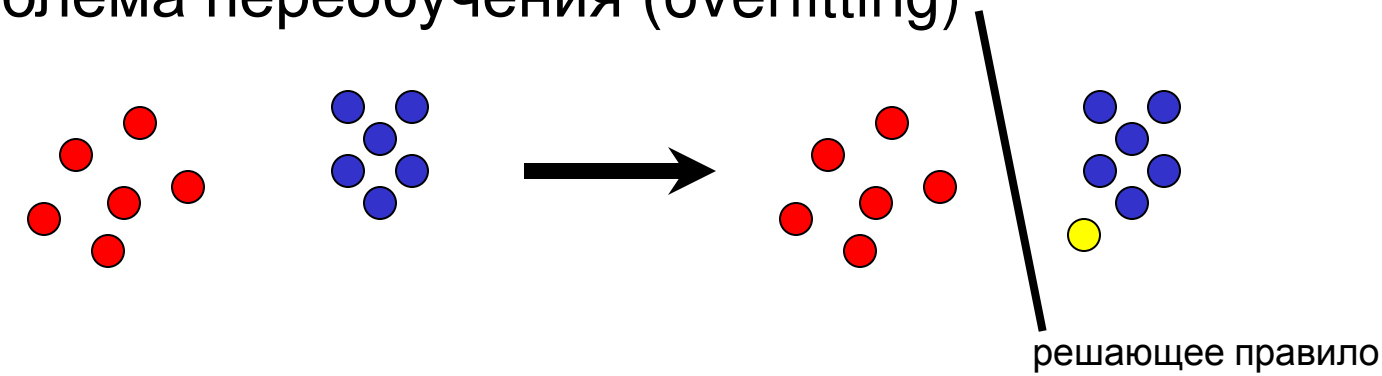


Обучение с учителем

Задача – построить решающее правило, которое позволит относить вновь обработанный образец в тот или иной класс.

ключевые слова: нейронные сети, SVM, decision trees

- **Основное отличие от кластерного анализа: требуется большая выборка с заранее отклассифицированными объектами.**
- Обычно работает лучше чем обучение без учителя.
- Кросс-валидация – при отсутствии тестовой выборки.
- Проблема переобучения (overfitting)



Обучение с учителем

$X = \{x_i, y_i\}_{i=1}^{n}$ обучающее множество

$x_i \in \Re^g$ вектор измерений (признаки объекта)

$y_i \in \{+1, -1\}$ метки (класс каждого объекта)

Задача: построить решающее правило f, которое хорошо описывает данные, т.е. найти отображение

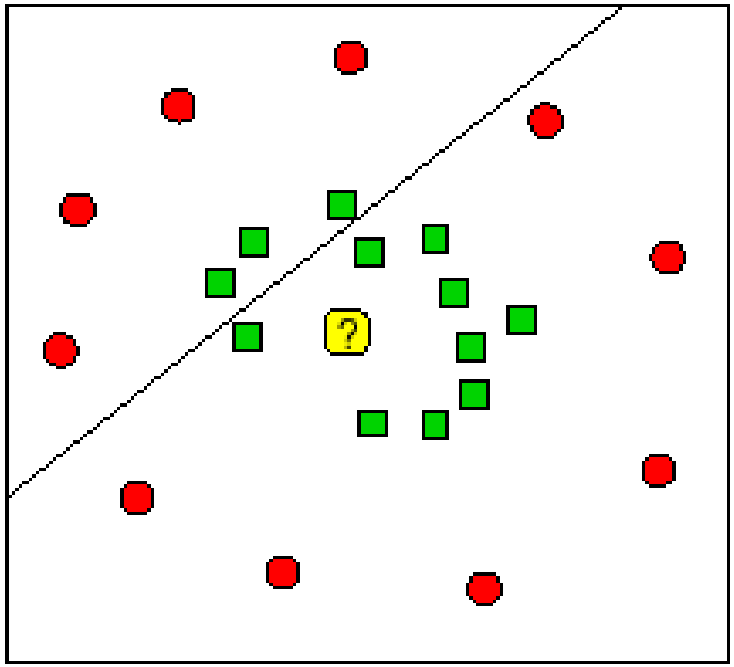
$$f_X : \Re^g \mapsto \{+1, -1\}$$

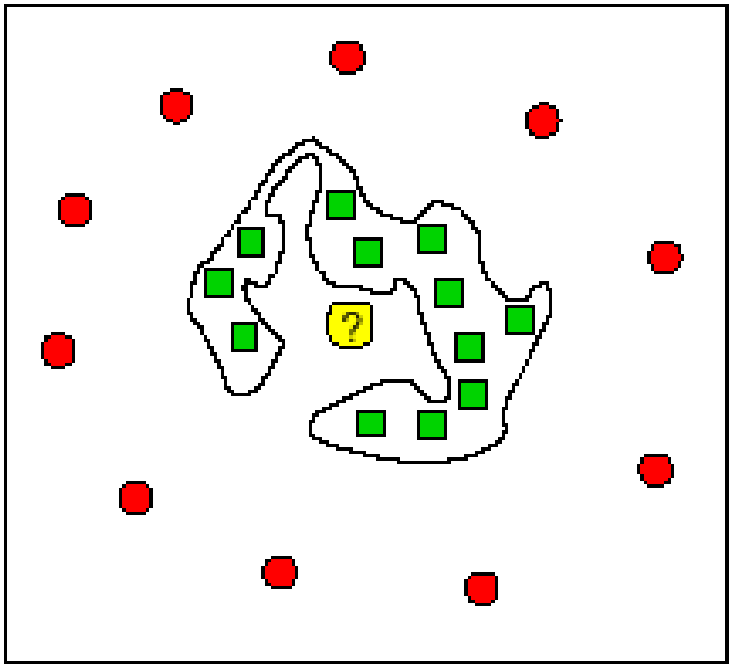
$$\text{Решение} = f_X(\textit{новый объект})$$

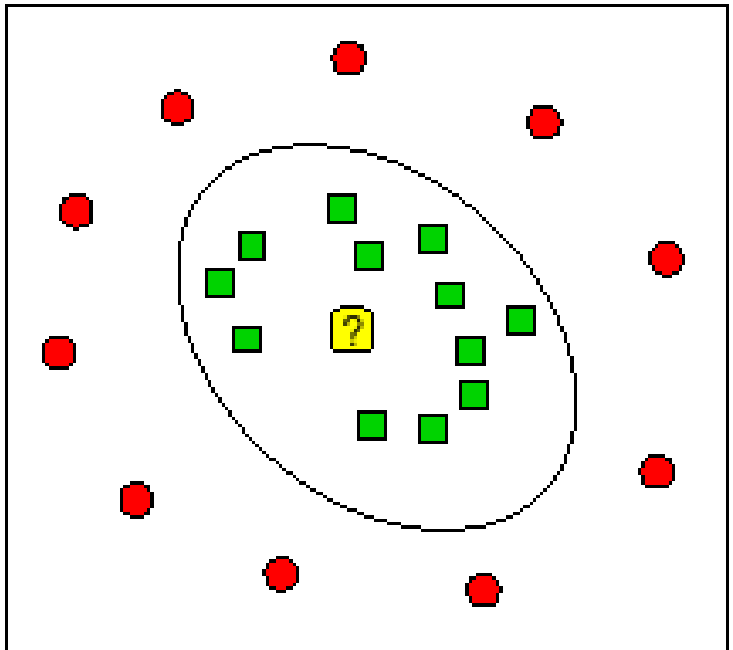
Построение решающего правила – решение сложной оптимизационной задачи (min числа ошибок)

Различные подходы к ее решению

Проблема: обобщение и переобучение





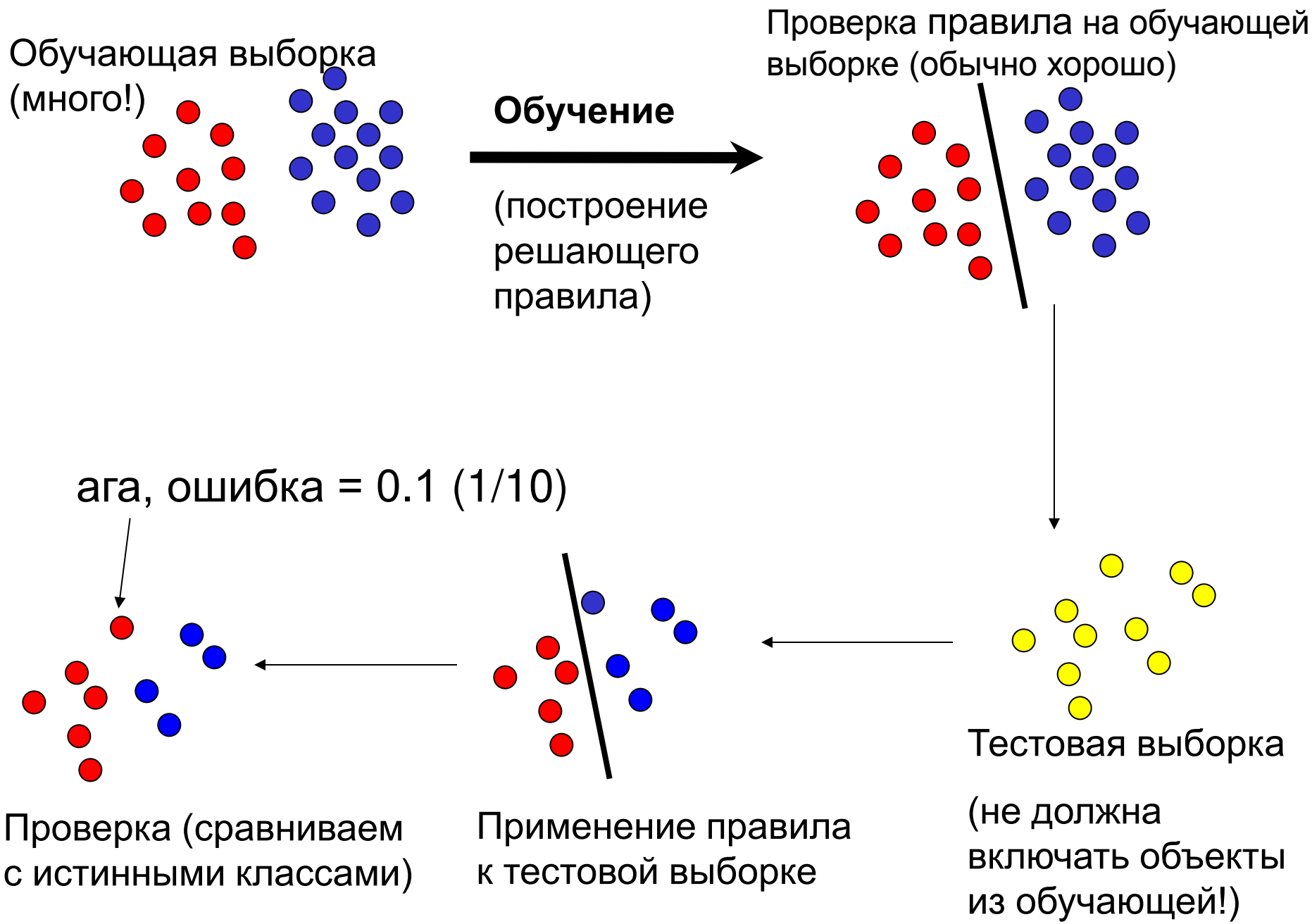


класс 1

класс2

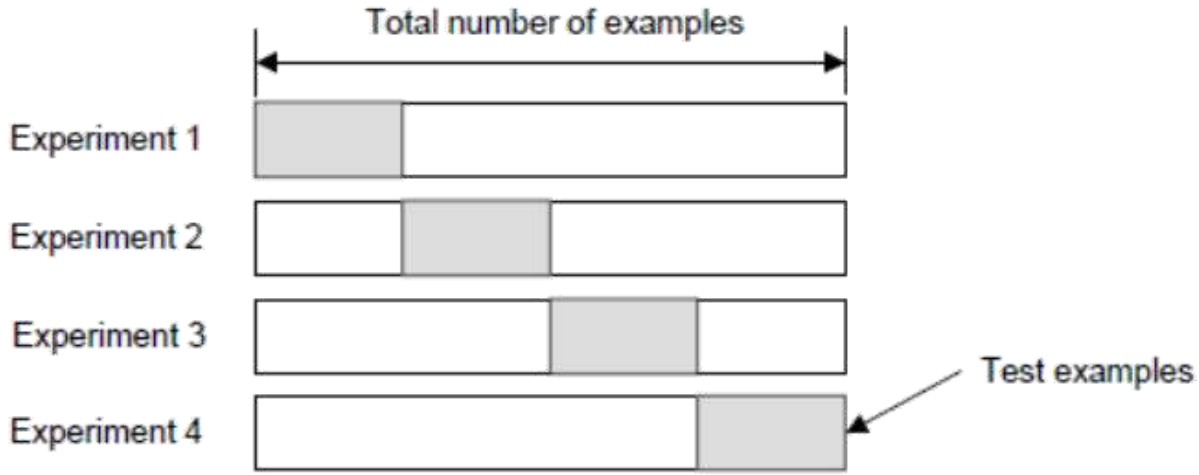
новый объект

Обучающая и тестовая выборки



Кросс-валидация: проверка классификации

- **Главное - не тестироваться на том, на чем обучались!**
  При этом оптимистическая оценка ошибки
- Применяется при малом объеме выборки (почти всегда)
- Разделить всю выборку на *k* частей, $X_i, i=1,...,K$



- Каждая k-я часть – для тестирования, остальное – на обучение

$$E = \frac{1}{K}\sum_{i=1}^{K} E_i$$ -- средняя ошибка

Оценка классификатора

		Золотой стандарт	
		−	+
Результат классифи-кации	−	Истинно-отрицательный True Negative (TN)	Ложно-отрицательный False Negative (FN) *Ошибка второго рода*
	+	Ложно-положительный False Positive (FP) *Ошибка первого рода*	Истинно-положительный True Positive (TP)

- Точность = (TP+TN) / (TP+FP+FN+TN)
- Чувствительность = TP / (TP+FN)
- Специфичность = TN / (TN+FP)

Сравнительный анализ классификаторов

- Cross-validation
- Confidence intervals for acc, spec, sens

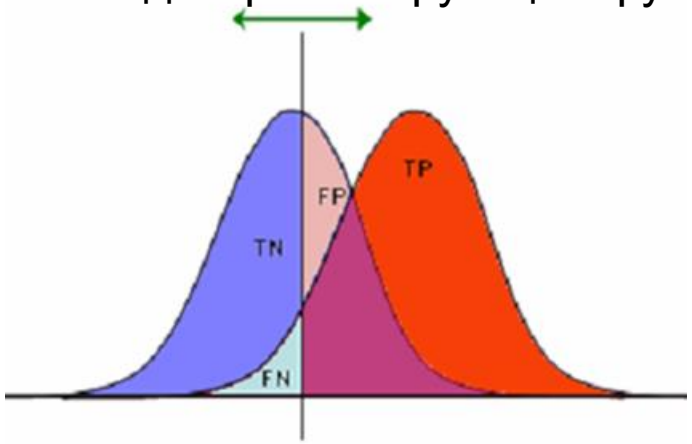
Table 3. Performance of SVM and LR classifiers in discriminating between cancer and noncancer sera estimated using 100 runs of ten-fold crossvalidation

Method	Accuracy (%)	Specificity (%)	Sensitivity (%)
SVM (CA125)[a)]	86.17 ± 0.70[b)]	98.82 ± 0.28	64.68 ± 1.67
SVM (ELISA)[c)]	86.42 ± 0.70	96.48 ± 0.49	70.27 ± 1.75
SVM RFE (MS)[d)]	89.47 ± 0.66	93.34 ± 0.70	83.57 ± 1.59
SVM (TSP(MS)+ELISA)[e)]	86.74 ± 0.70	92.84 ± 0.80	77.46 ± 1.61
SVM RFE (ELISA+MS)[f)]	95.23 ± 0.43	98.07 ± 0.35	90.84 ± 1.08
LR (CA125)	85.05 ± 0.70	95.62 ± 0.54	67.53 ± 1.72
LR (ELISA)	86.58 ± 0.69	94.30 ± 0.58	74.19 ± 1.66
LR (MS)	86.03 ± 0.74	87.52 ± 1.28	83.71 ± 1.47
LR AIC (TSP(MS)+ELISA)	90.74 ± 0.59	96.87 ± 0.47	81.24 ± 1.41
LR AIC (ELISA+MS)	91.93 ± 0.58	92.73 ± 0.71	90.69 ± 1.12

ROC – кривая

Receiver-Operating Characteristic curve

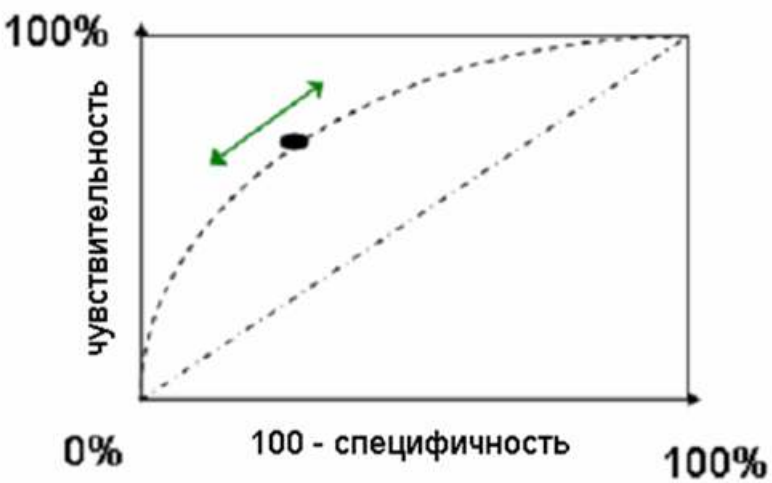
Значение дискриминирующей функции



Золотой стандарт

Модель

TP	FP
FN	TN



Чувствительность = $\frac{TP}{TP + FN}$

(ошибка на “положительных”)

Специфичность = $\frac{TN}{TN + FP}$

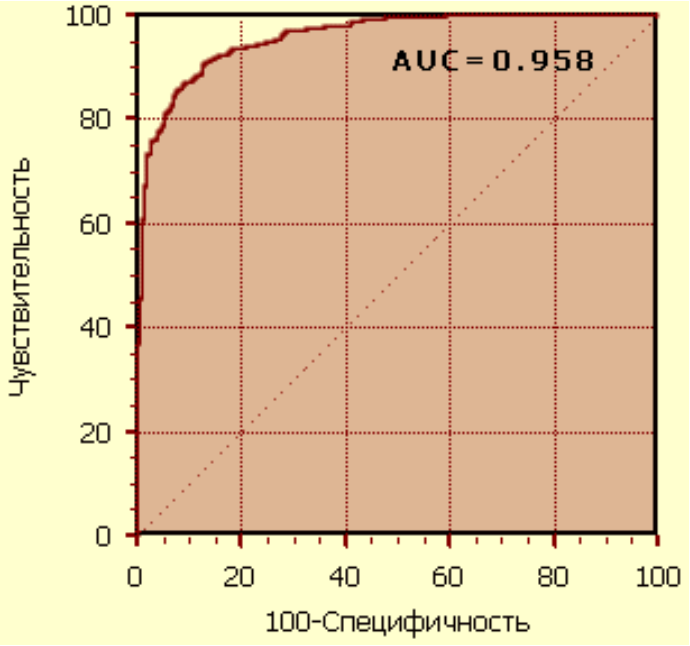
(ошибка на “отрицательных”)

ROC – кривая - summary

- Используется для представления результатов классификации на 2 класса

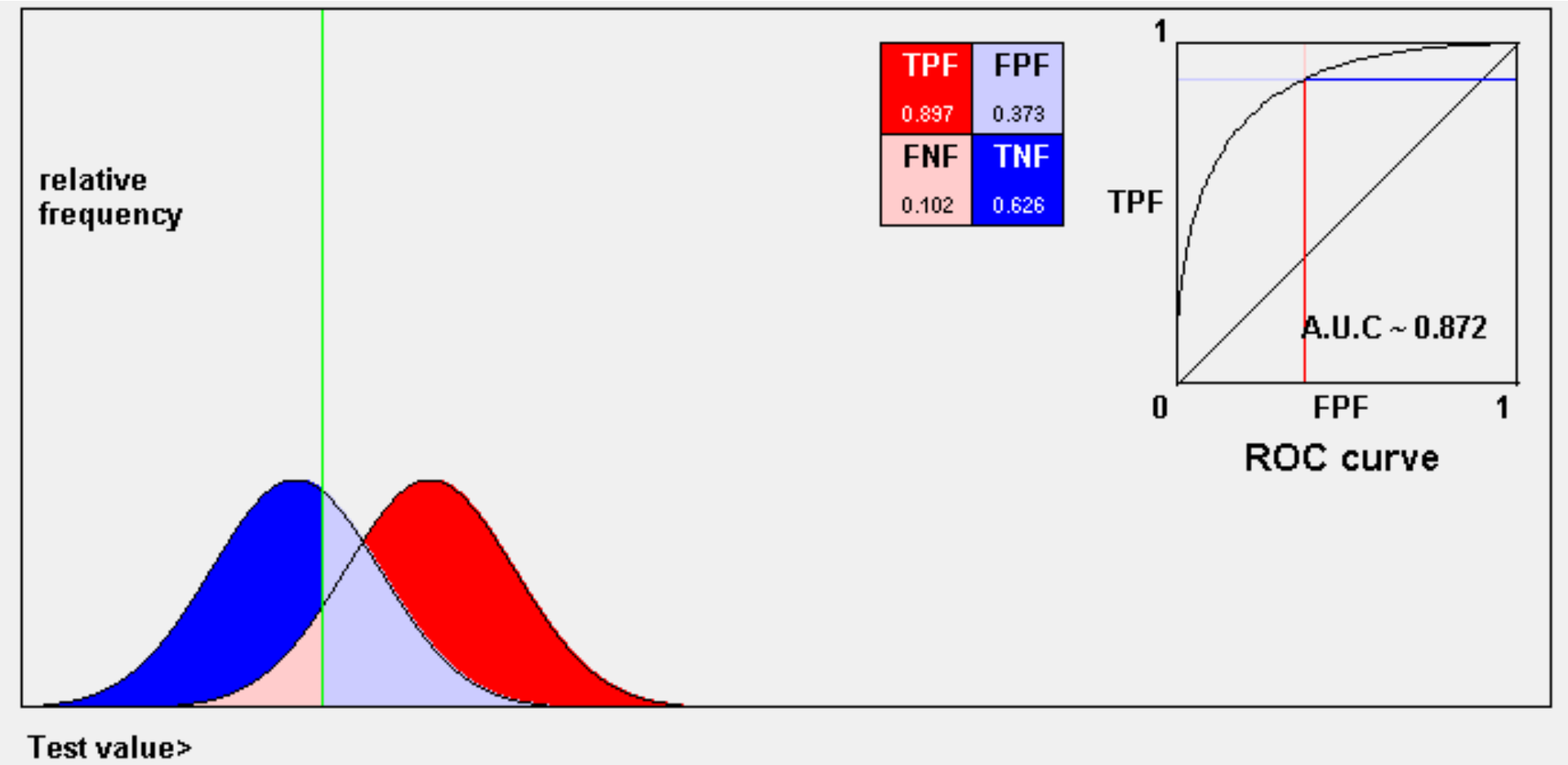
- Параметр классификатора (порог), варьируя который получается то или иное разбиение

	Фактически	
Модель	положительно	отрицательно
положительно	TP	FP
отрицательно	FN	TN

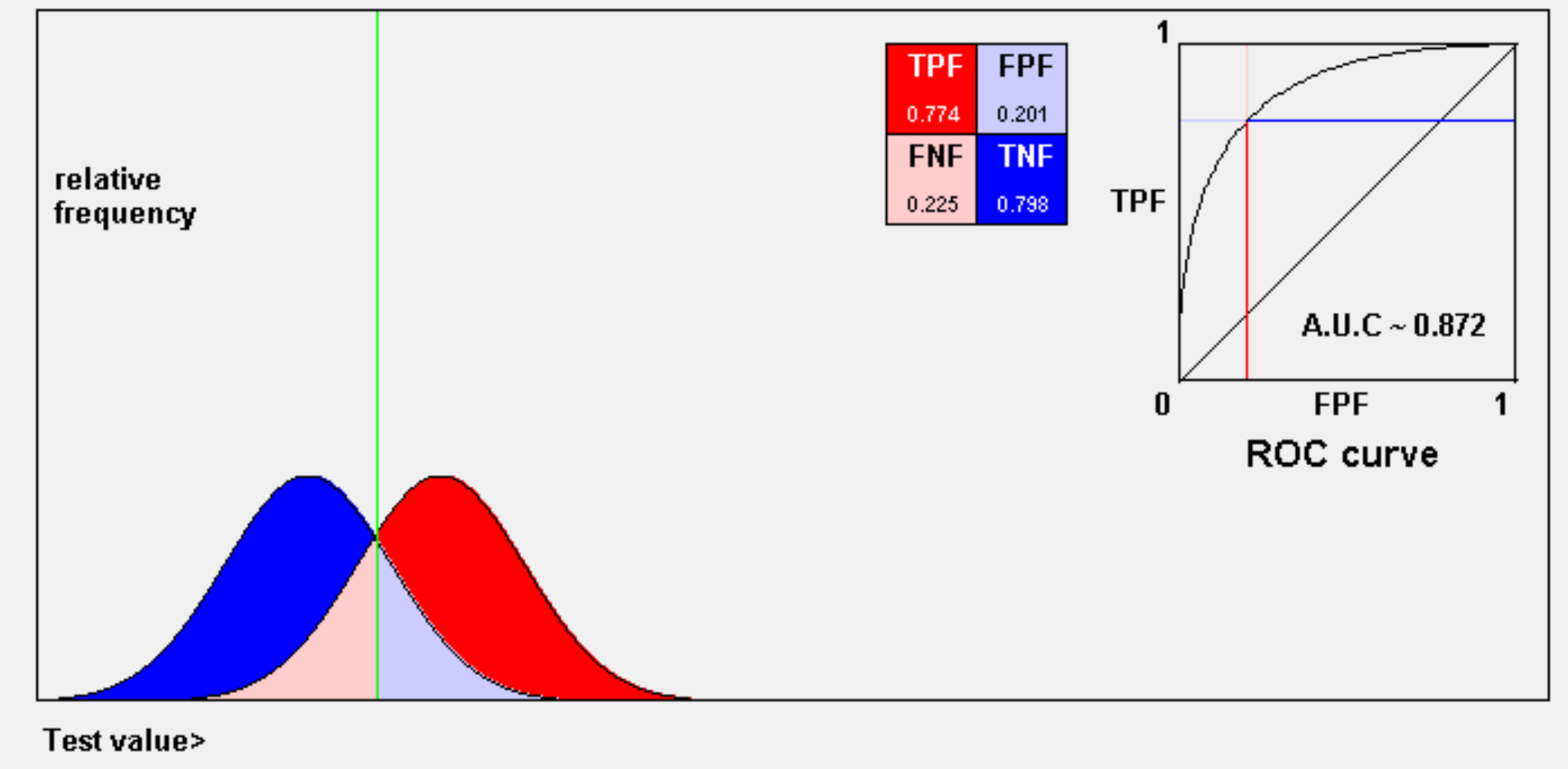


Что является “положительным” событием – зависит от условий задачи.

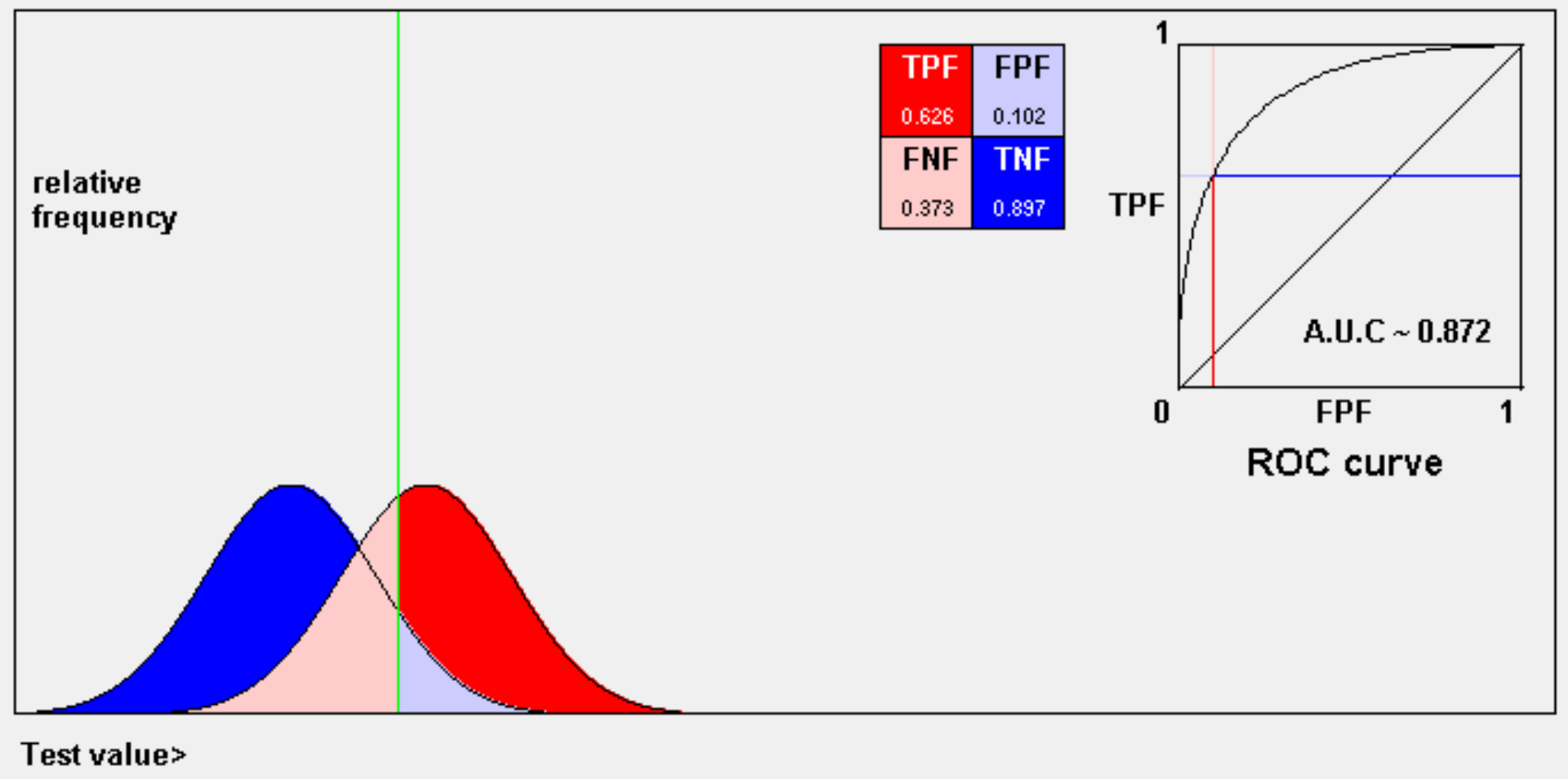
ROC-кривая: пример 1



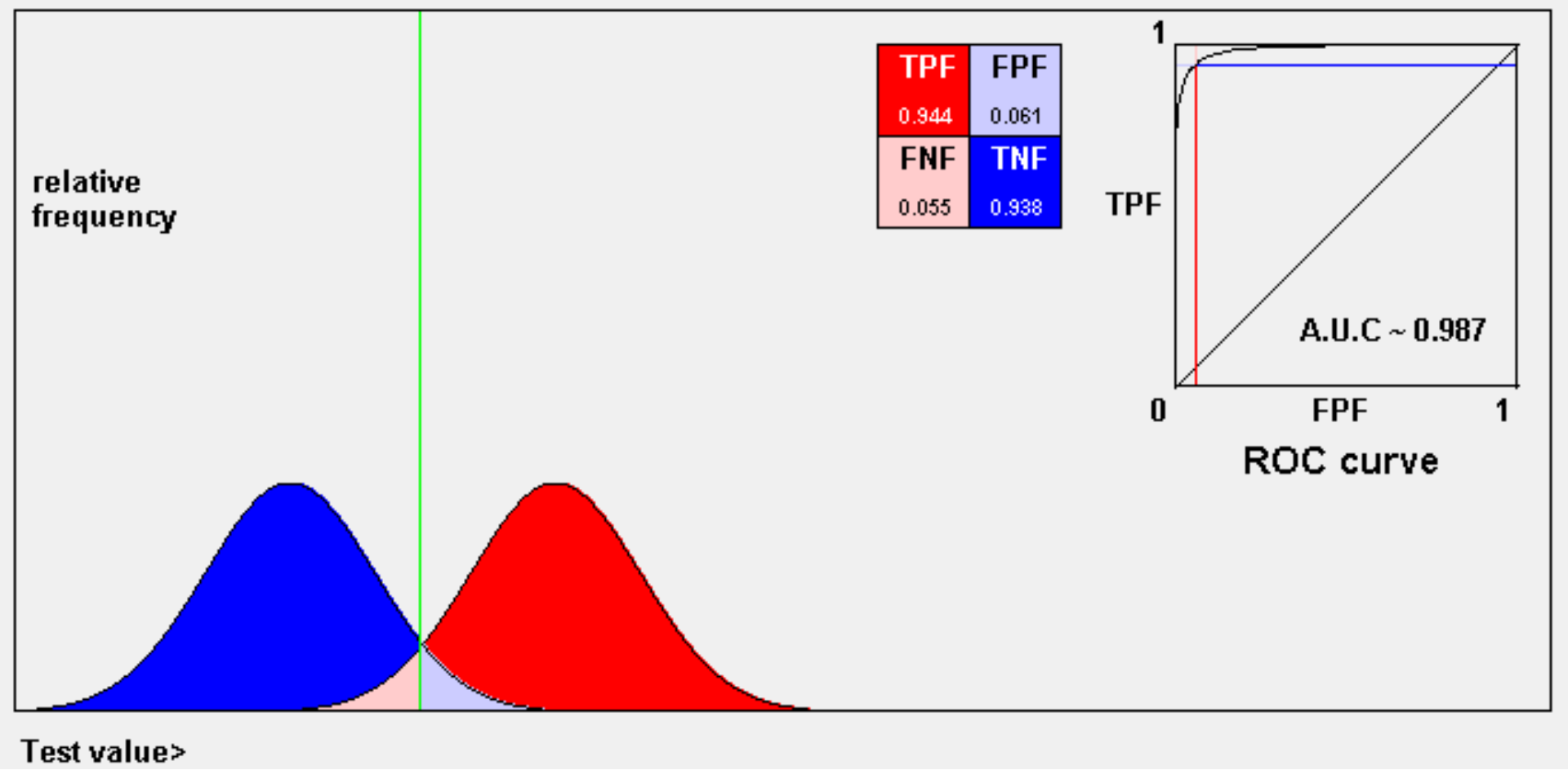
ROC-кривая: пример 1



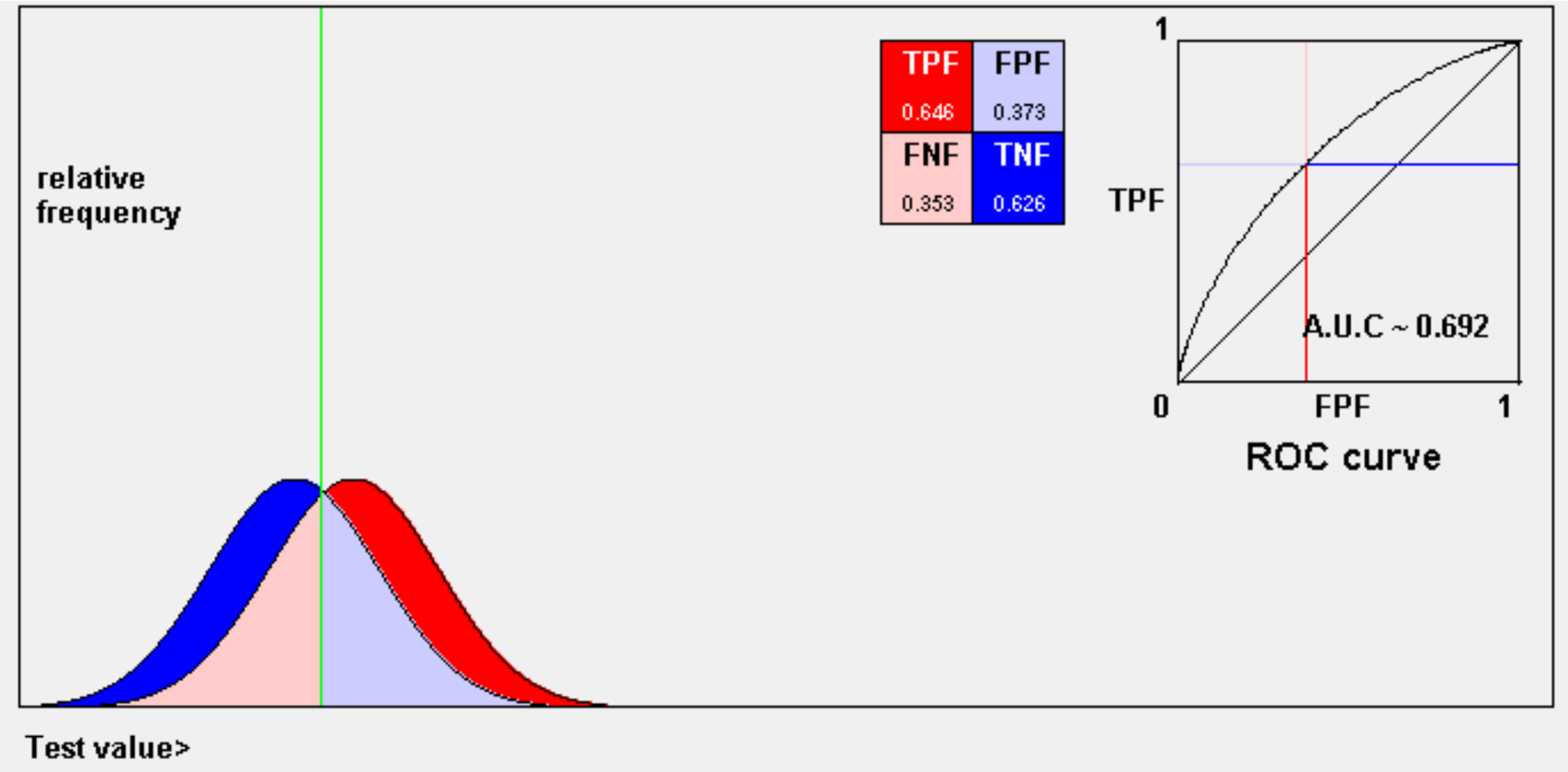
ROC-кривая: пример 1



ROC-кривая: классы хорошо разделяются



ROC-кривая: классы плохо разделяются



Важно: уменьшение числа признаков

Большое число признаков (*p*). Очень часто в современной биологии, десятки тысяч.

Число образцов (*n*) мало. Проблема $n \ll p$

- Усложняет анализ данных
- Обычно не все признаки имеют биологическую интерпретацию
- Часто целью является отбор значимых признаков

  какие гены по-разному экспрессируются в норме и патологии?
- Ухудшает работу классификации. Переобучение.
- Число признаков можно сокращать либо до классификации либо после.

NB: выбор “правильных” признаков важнее выбора “правильного” метода классификации!

Уменьшение размерности пространства (dimensionality reduction)

Filter

не используют алгоритм классификации

- корреляция
- PCA
- методы теории информации
- отбор “интересных переменных” (ex: бимодальных)

Wrapper

используют классификацию как способ оценки качества переменной

- Кросс-валидация
- Оптимально – полный перебор
- Backward, forward selection
- Генетические алгоритмы

Этапы решения практических задач

- формализация предметной области (определение признаков и набора классов)
- формирование обучающей выборки
- снижение размерности пространства признаков
- обучение («с учителем», «без учителя»)
- распознавание
- адаптация (если качество обучения признано неудовлетворительным)
- обратная задача распознавания: определение набора признаков наиболее характерных для объектов заданного класса

Распознавание образов в биологии (биоинформатика)

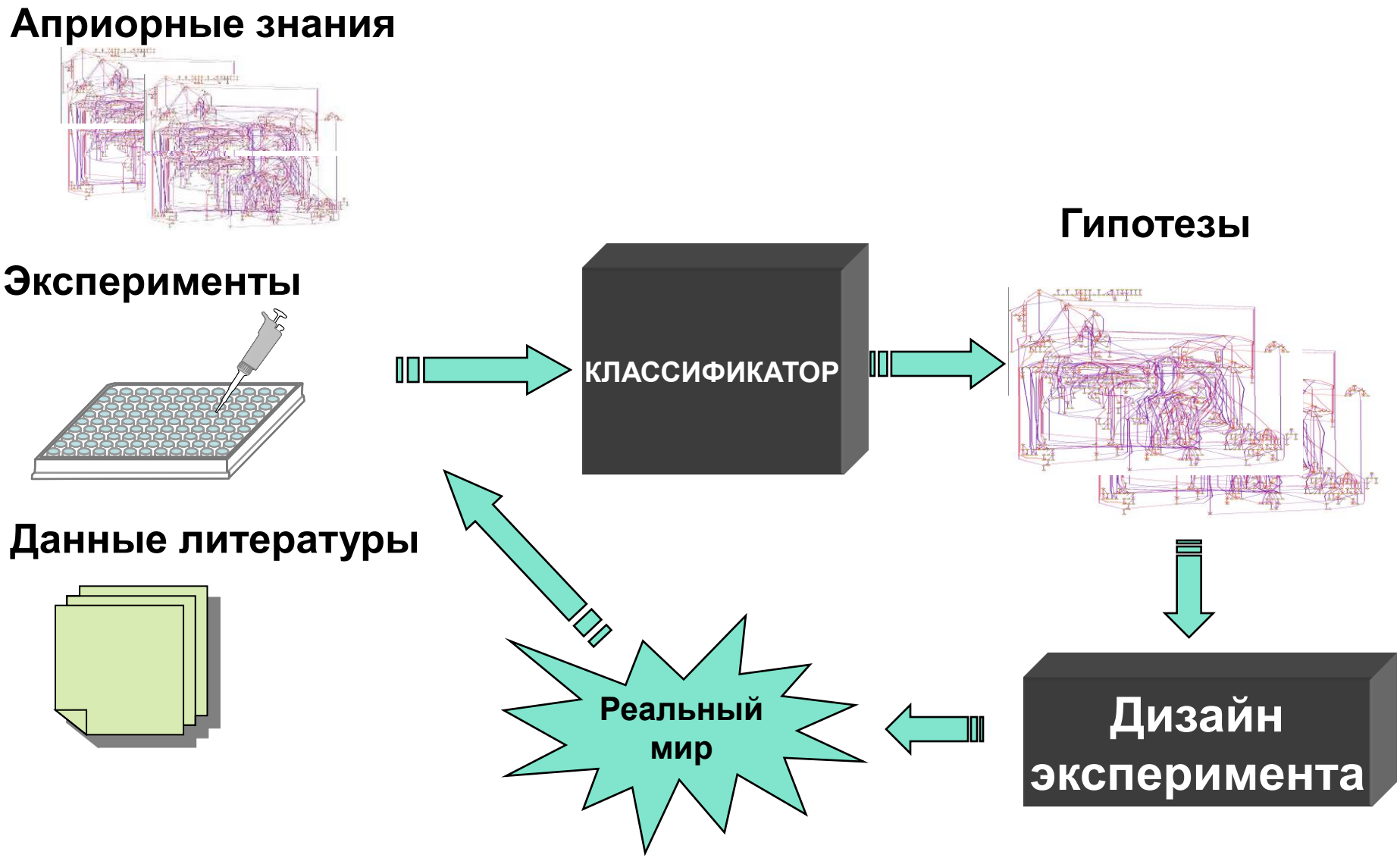
Экспериментальные технологии:

- транскриптомные ДНК микрочипы (биомаркеры!)
- протеомика: масс-спектрометрия, электрофорез
- экспресс-методы ДНК секвенирования
- системы детекции метаболитов (ЯМР, МС)
- флуоресцентная микроскопия клеток
- автоматизированный анализ тканевых срезов

Вычислительные технологии:

- предсказание 3D структуры, локализации и функций белков
- разметка экзон/интронной структуры
-методы анализа количественного соотношения структура-активность, предсказание фармакокинетики, побочных эффектов
- молекулярное моделирование и докинг

Распознавание образов в биологии



Ура!