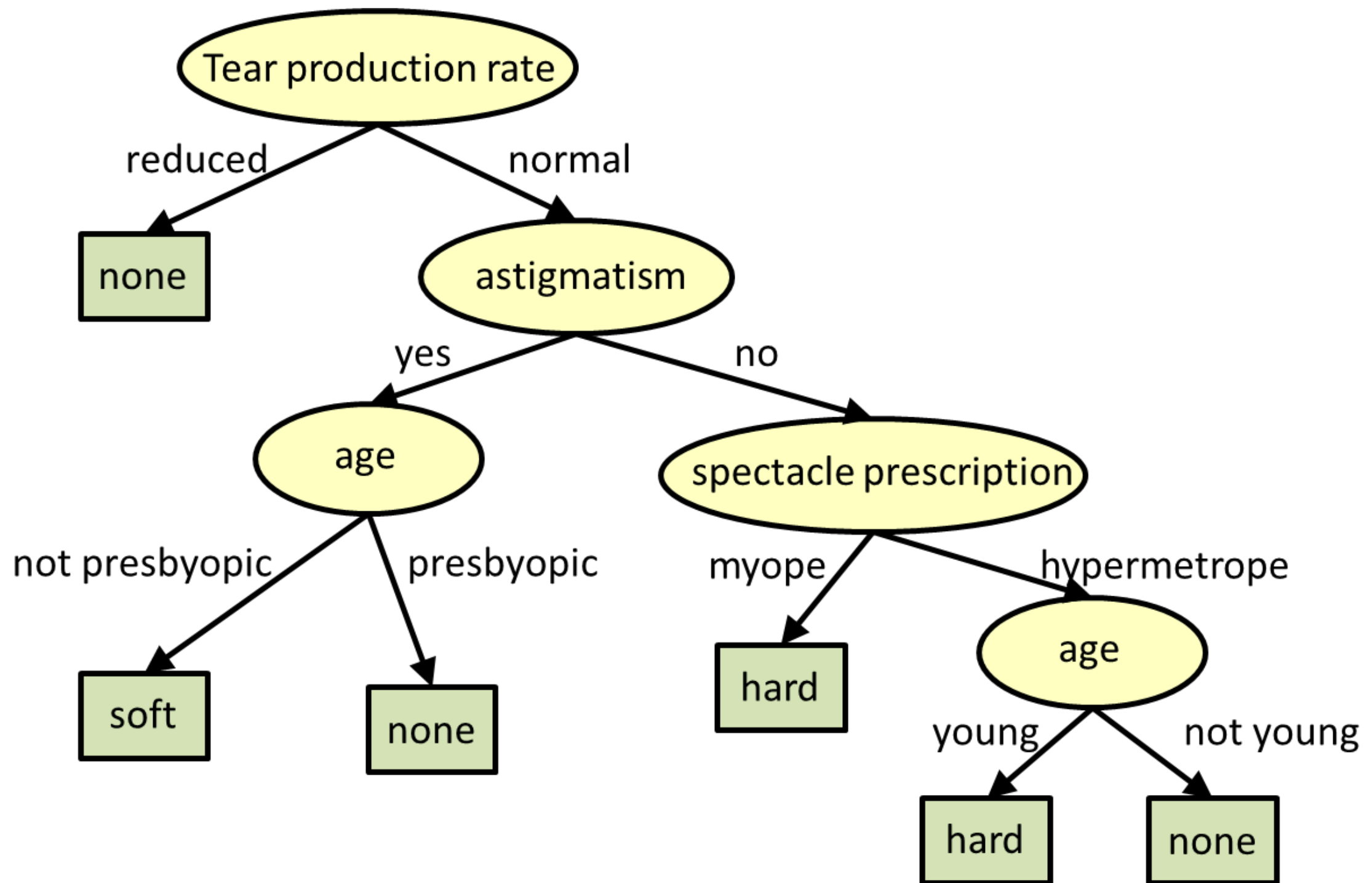


# Деревья решений



## How to win a Kaggle competition?



Anthony Goldbloom

“According to Anthony, in the history of Kaggle competitions, there are only two Machine Learning approaches that win competitions: **Handcrafted & Neural Networks.**”

## Где побеждают ансамбли деревьев решений?

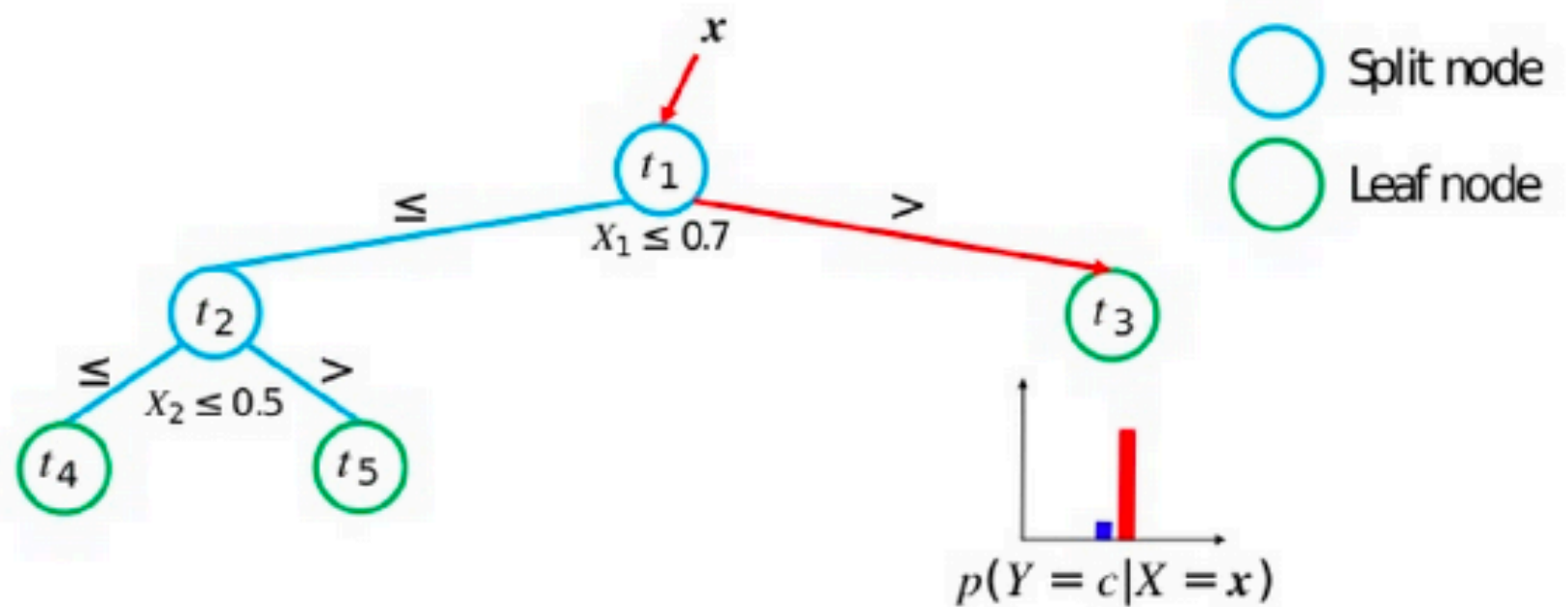
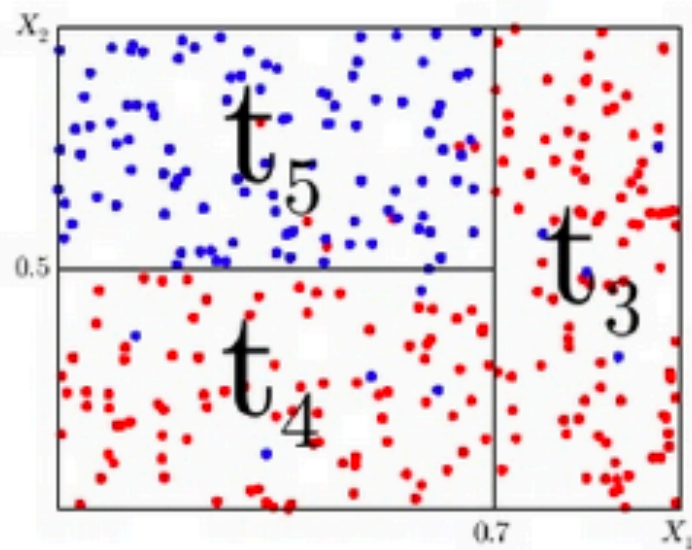
- ▶ Recommendation systems (Netflix Prize 2009)
- ▶ Learning to rank (Yahoo Learning to rank challenge 2010)
- ▶ Crowdfunder Search Results Relevance (2015)
- ▶ Avito Context Ad Clicks (2015)
- ▶ Везде :)



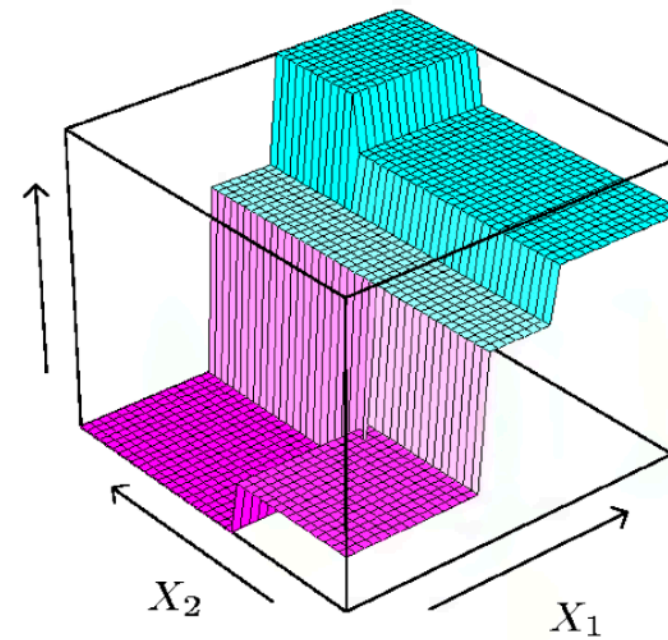
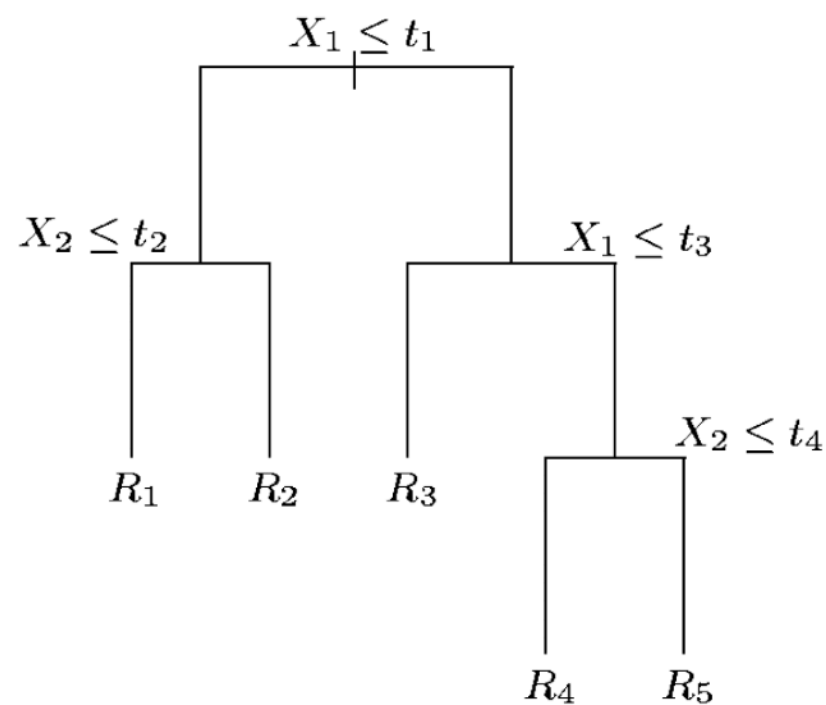
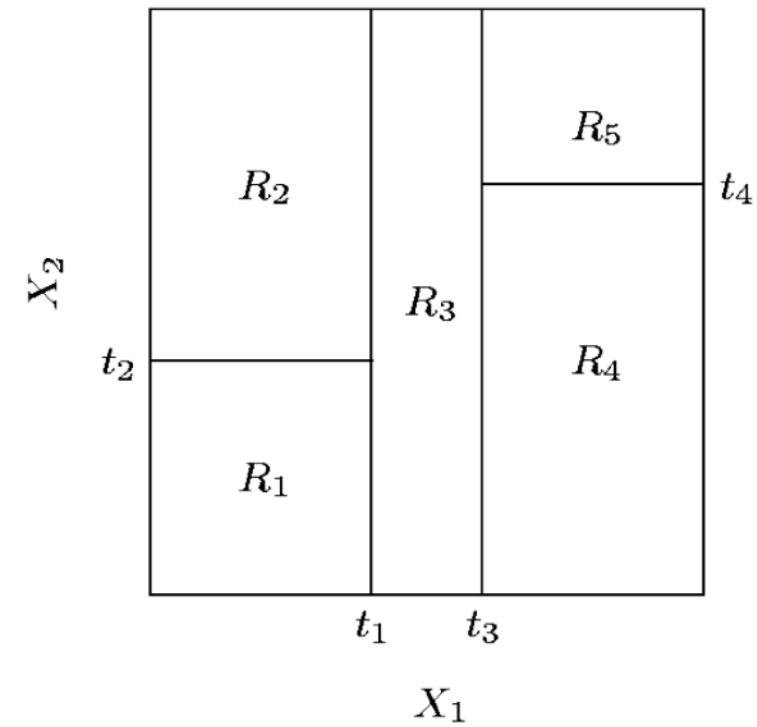
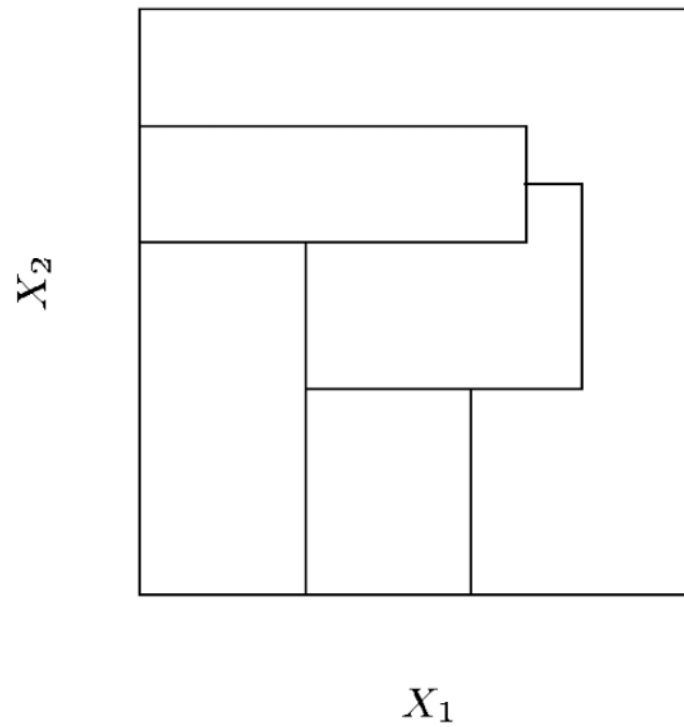
“As long as Kaggle has been around, Anthony says, it has **almost always** been **ensembles of decision trees that have won competitions.**”

# Деревья решений

## Деревья решений (принцип работы)



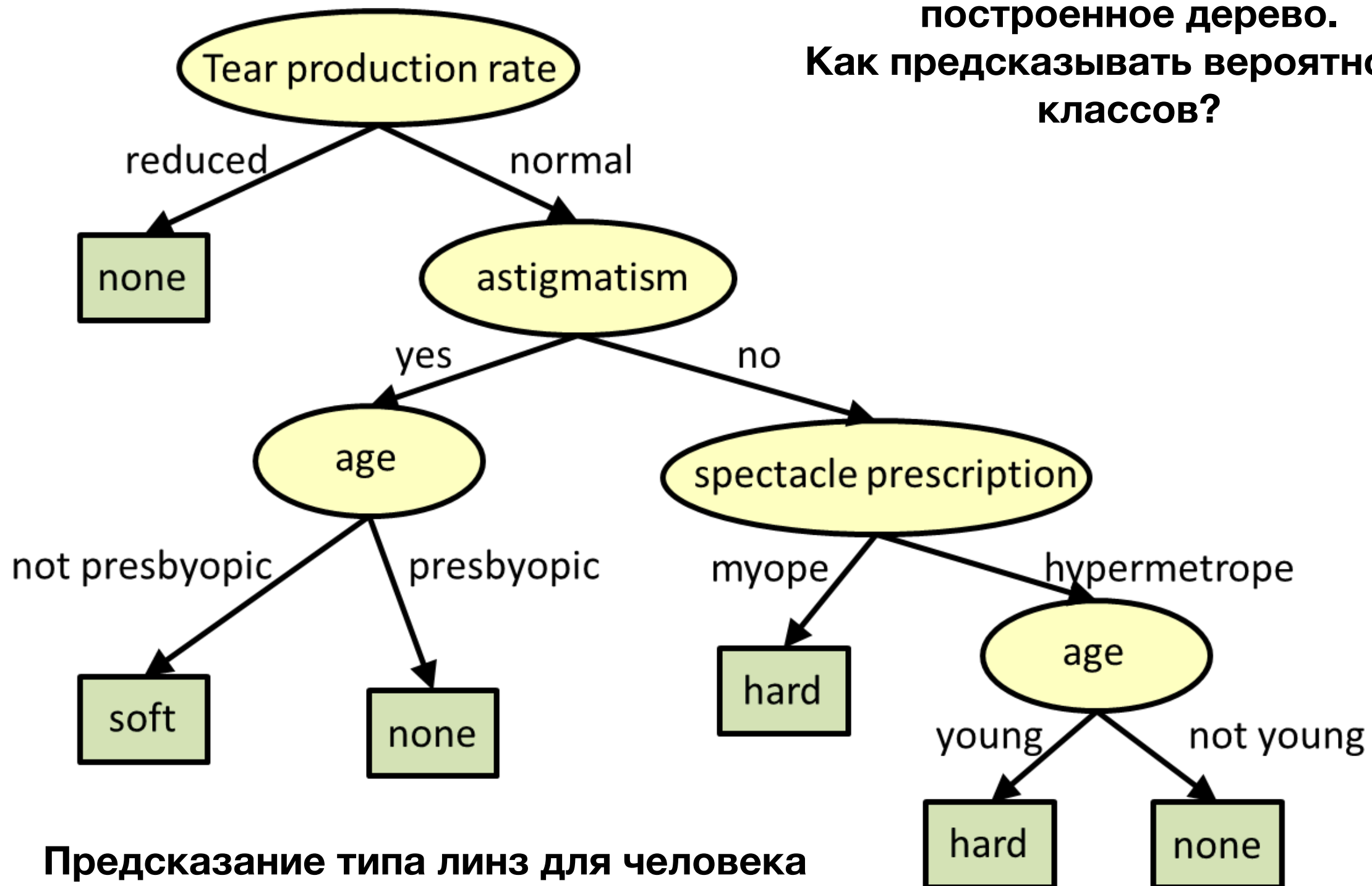
# Разбиение пространства





# Классификация

Допустим, у нас уже есть  
построенное дерево.  
Как предсказывать вероятности  
классов?



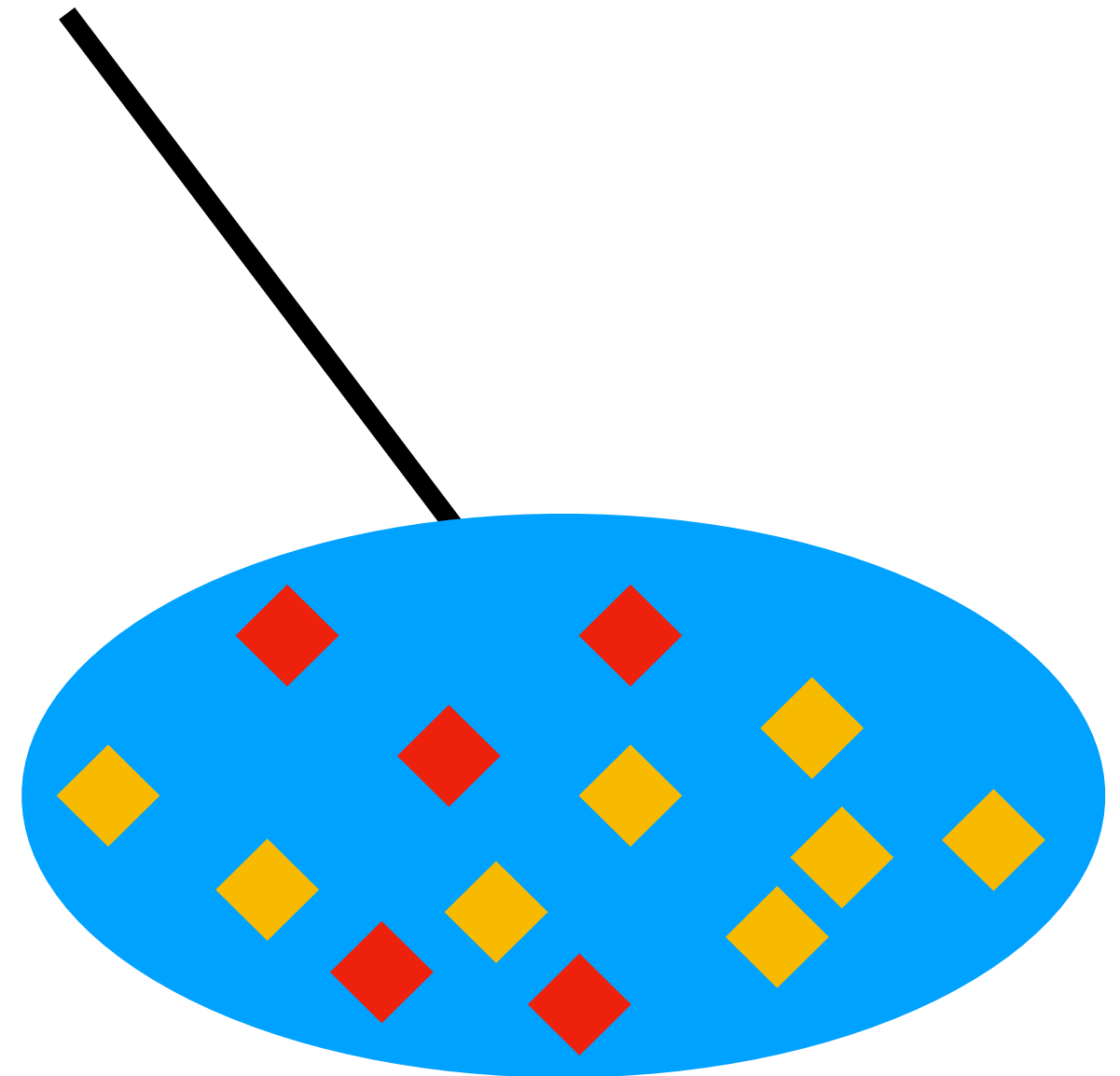
Предсказание типа линз для человека

# Классификация

◆ - объекты 1го класса из обучения

◆ - объекты 2го класса из обучения

Предсказываем вероятность - как  
ее оценить?



Лист дерева, тут нам надо сделать  
константное предсказание

# Оценка доли в популяции

$$\hat{p} = \frac{n}{N}$$

$$E(\hat{p}) = p$$

$$sd(\hat{p}) = \frac{\sqrt{p \cdot (1 - p)}}{\sqrt{N}}$$



# Классификация



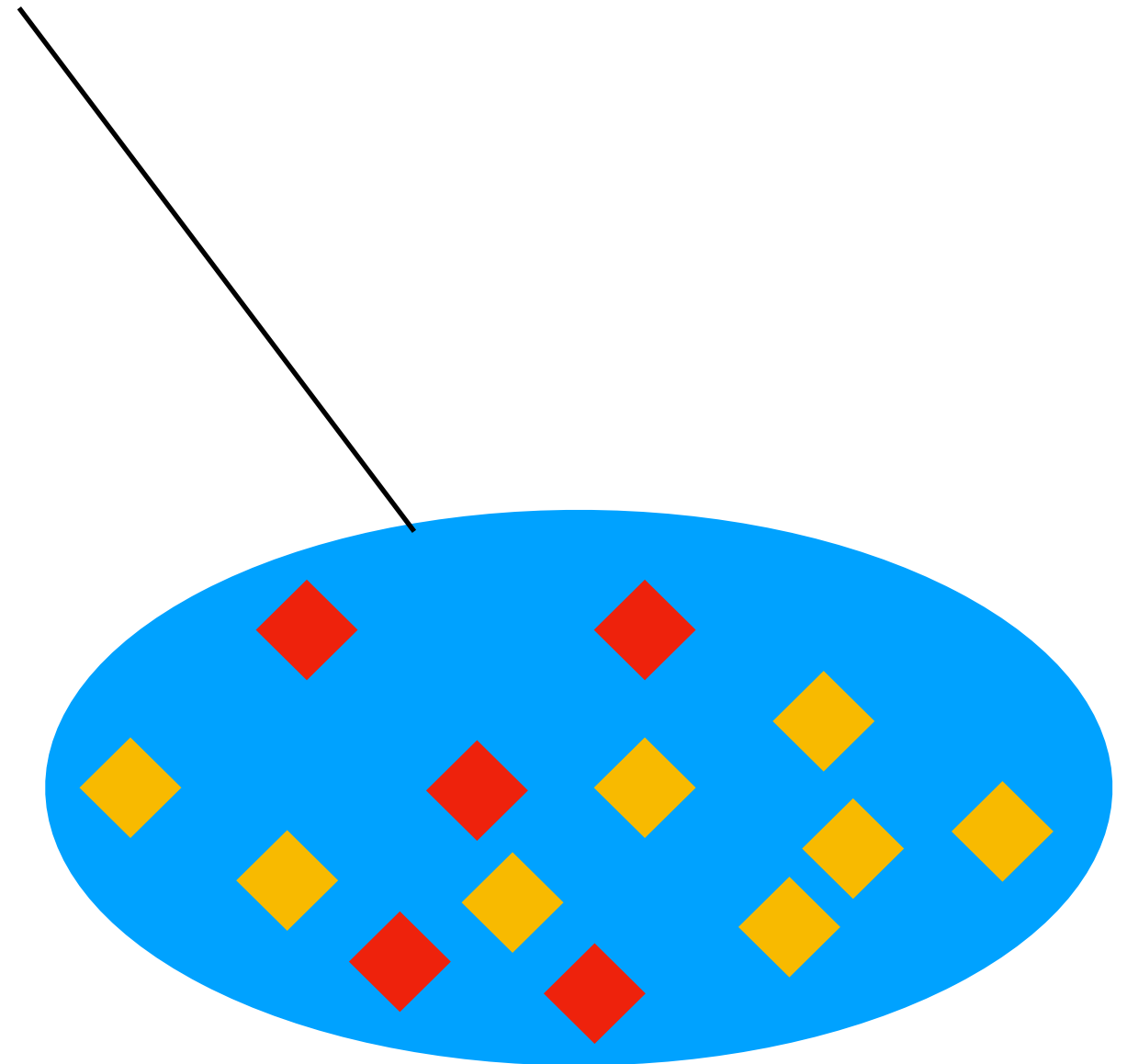
- объекты 1го класса



- объекты 2го класса

Предсказываем вероятность - как  
ее оценить?

Просто доля класса



Лист дерева, тут нам надо сделать  
константное предсказание

# Классификация

Как строить дерево?

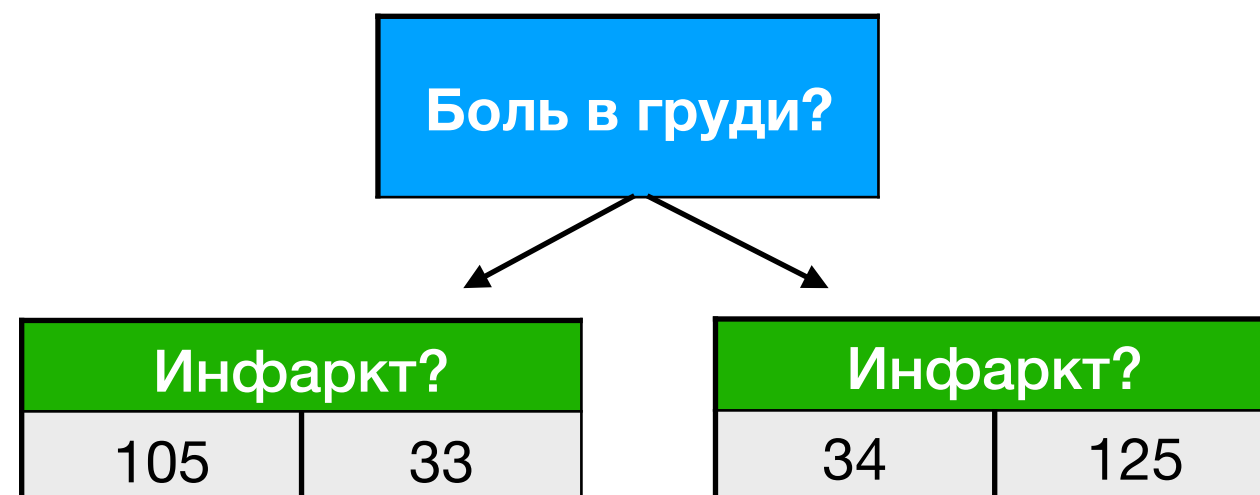
Боль в груди	Хорошо циркулирует кровь	Есть атеросклероз	Инфаркт
Нет	Нет	Нет	Нет
Да	Да	Да	Да
Да	Да	Нет	Нет
Да	Нет	Да	Да
...	...	...	...

# Классификация

Боль в груди	Хорошо циркулирует кровь	Есть атеросклероз	Инфаркт
Нет	Нет	Нет	Нет
Да	Да	Да	Да
Да	Да	Нет	Нет
Да	Нет	Да	Да
...	...	...	...

## Как строить дерево?

Попробуем построить дерево с двумя листьями,  
корень - признак - есть или нет боль в груди?



# Классификация

Как строить дерево?

Перебираем таким образом все признаки

Боль в груди	Хорошо циркулирует кровь	Есть атеросклероз	Инфаркт
Нет	Нет	Нет	Нет
Да	Да	Да	Да
Да	Да	Нет	Нет
Да	Нет	Да	Да
...	...	...	...

Боль в груди?

Инфаркт?

105

33

Инфаркт?

34

125

Хорошо циркулирует кровь

Инфаркт?

37

127

Инфаркт?

100

33

Есть атеросклероз?

Инфаркт?

92

31

Инфаркт?

45

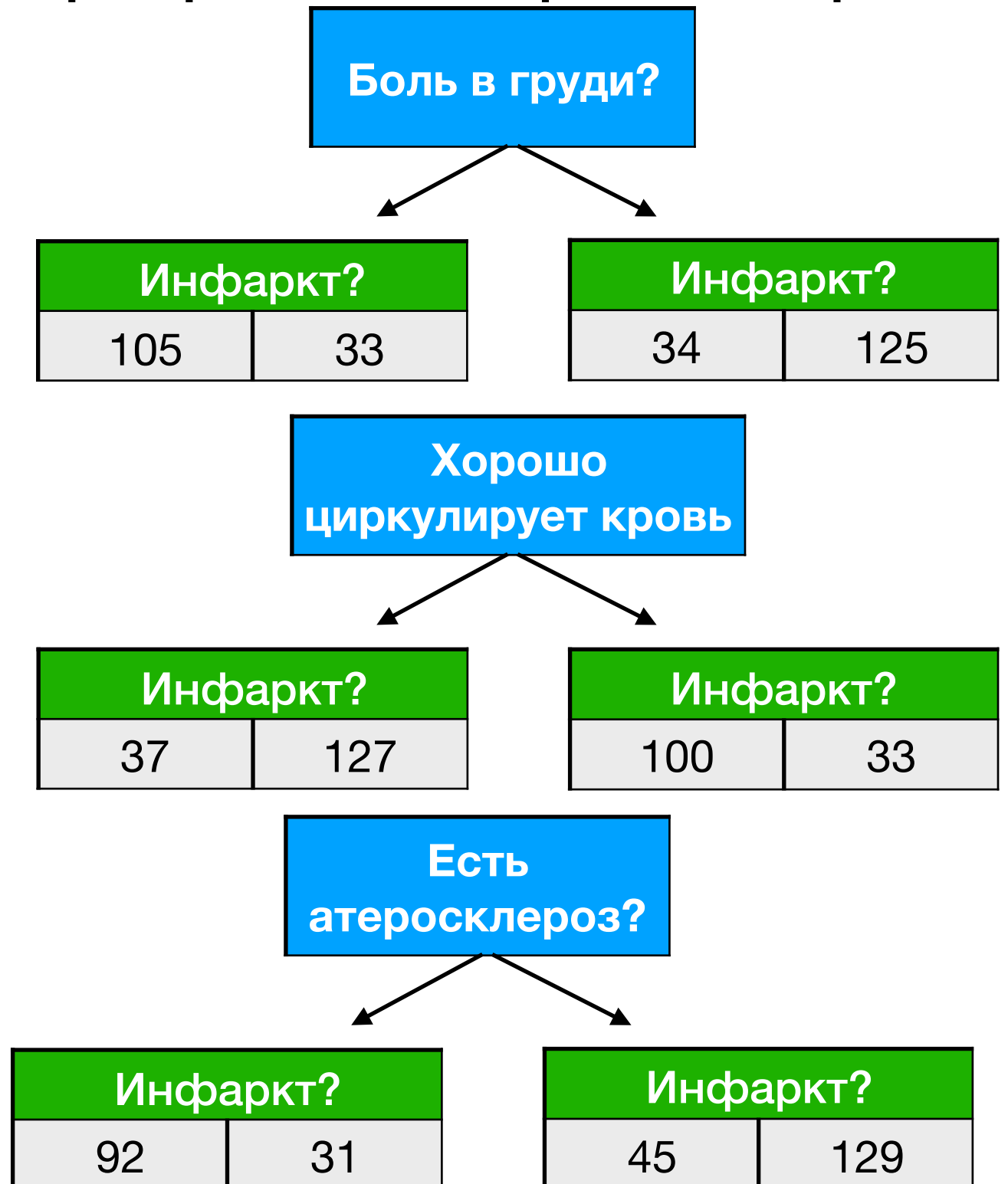
129

# Классификация

Как строить дерево?

Перебираем таким образом все признаки

Как выбрать одно из деревьев?



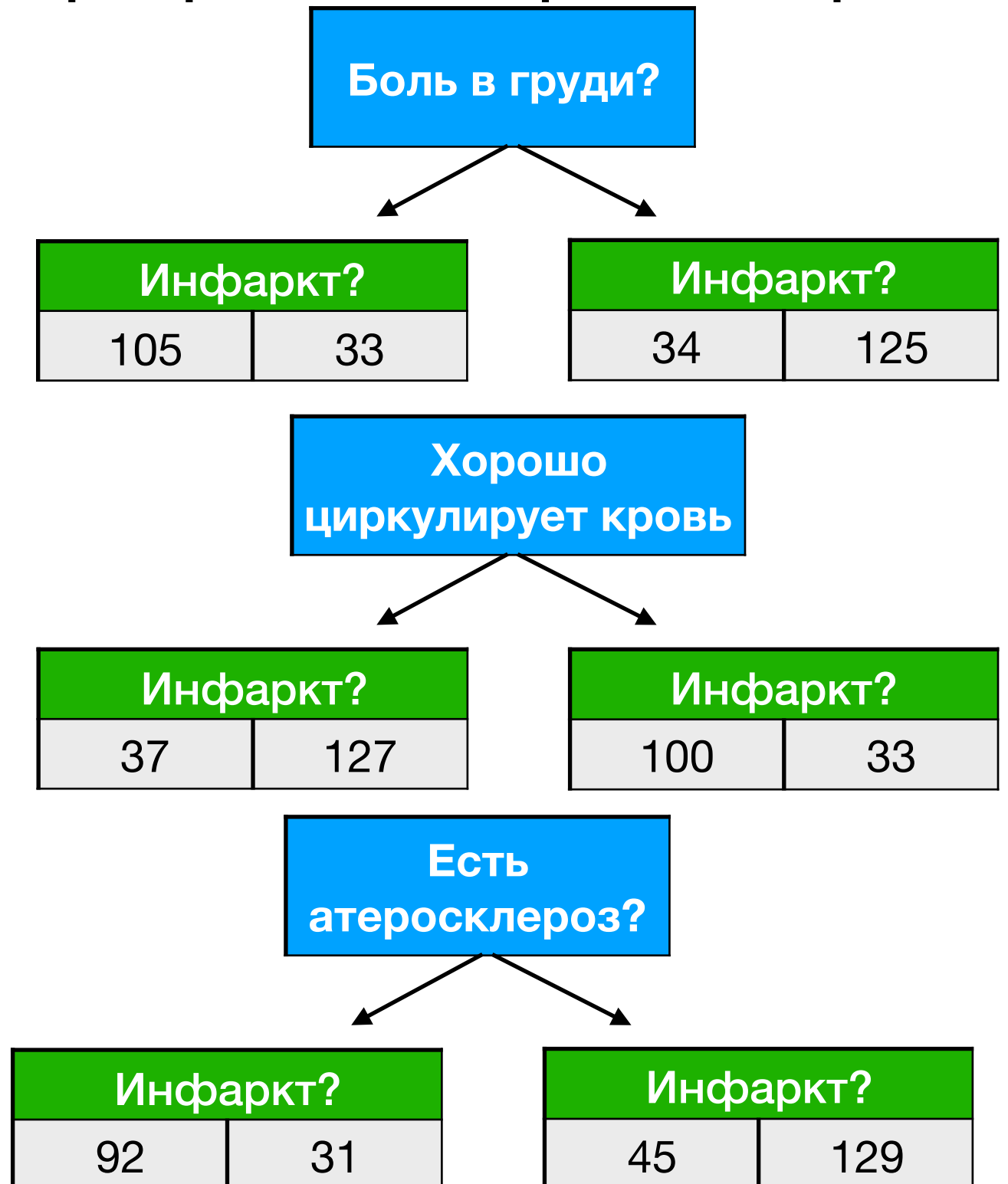
# Классификация

Как строить дерево?

Перебираем таким образом все признаки

Как выбрать одно из деревьев?

По качеству разбиения



# Классификация

Ошибка классификации - какую долю объектов неправильно классифицируем

$$\frac{1}{N} \sum_{i \in \text{train}} I(y_i \neq \hat{y}_i)$$

Gini-index - математическое ожидание ошибки классификации, если мы относим объект к классу с вероятностью, равной вероятности этого класса.

Иначе - если мы возьмем два объекта из данного листа, какова вероятность, что они будут принадлежать к **разным** классам

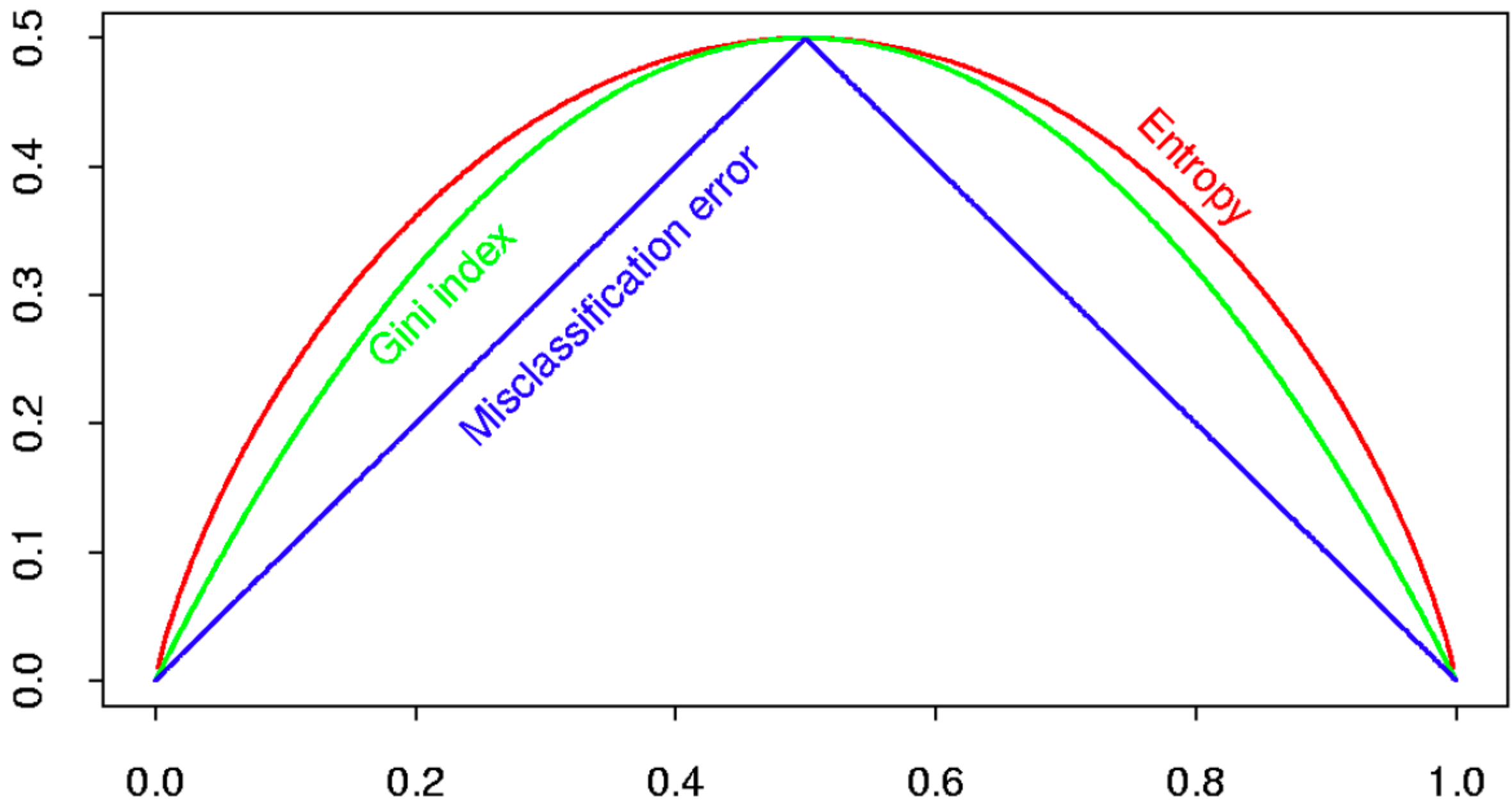
$$\sum_{k \in K} \hat{p}_k (1 - \hat{p}_k) = 1 - \sum_{k \in K} \hat{p}_k^2$$

Кросс-энтропия - сколько информации нам дает то, что лист попал в этот узел. Чем ближе кросс-энтропия к 0 - тем больше информации.

$$\sum_{k \in K} p_k \log p_k$$



# Классификация



# Классификация

Боль в груди?

Инфаркт?	
105	33

$$Gini_1 = 0.364$$

Инфаркт?

Инфаркт?	
34	125

$$Gini_2 = 0.336$$

Инфаркт?

Инфаркт?	
37	127

$$Gini_1 = 0.349$$

Инфаркт?

Инфаркт?	
100	33

$$Gini_2 = 0.373$$

Хорошо  
циркулирует кровь

Есть  
атеросклероз?

Инфаркт?

Инфаркт?	
92	31

$$Gini_1 = 0.377$$

Инфаркт?

Инфаркт?	
45	129

$$Gini_2 = 0.383$$

И что дальше?

# Качество разбиения

$$Impurity\_decrease = Gini_0 - \left( \frac{n_1}{n_1 + n_2} Gini_1 + \frac{n_2}{n_1 + n_2} Gini_2 \right)$$

**$n_1, n_2$  - число объектов в листьях**

**$Gini_0$  - чистота исходного узла**

# Классификация

$$p = \frac{105 + 33}{105 + 33 + 34 + 125} = 0.880$$

Доля объектов 0 класса до разбиения

$$Gini_0 = 0.498$$

Боль в груди?

Инфаркт?	
105	33

$$Gini_1 = 0.364$$

Инфаркт?

34	125
----	-----

$$Gini_2 = 0.336$$

Хорошо  
циркулирует кровь

Инфаркт?	
37	127

$$Gini_1 = 0.349$$

Инфаркт?

100	33
-----	----

$$Gini_2 = 0.373$$

Есть  
атеросклероз?

Инфаркт?	
92	31

$$Gini_1 = 0.377$$

Инфаркт?

45	129
----	-----

$$Gini_2 = 0.383$$

# Классификация

$$p = \frac{105 + 33}{105 + 33 + 34 + 125} = 0.880$$

Доля объектов 0 класса до разбиения

$$Gini_0 = 0.498$$

Боль в груди?

Инфаркт?	
105	33

$$Gini_1 = 0.364$$

Инфаркт?	
34	125

$$Gini_2 = 0.336$$

$$Impurity\_decrease = 0.149$$

Хорошо  
циркулирует кровь

Инфаркт?	
37	127

$$Gini_1 = 0.349$$

Инфаркт?	
100	33

$$Gini_2 = 0.373$$

$$Impurity\_decrease = 0.138$$

Есть  
атеросклероз?

Инфаркт?	
92	31

$$Gini_1 = 0.377$$

Инфаркт?	
45	129

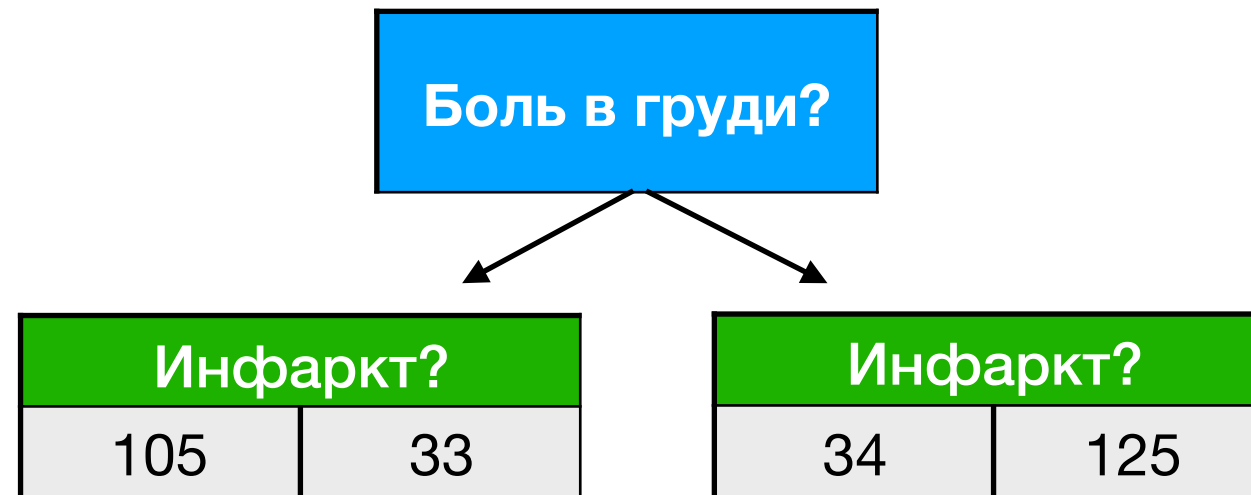
$$Gini_2 = 0.383$$

$$Impurity\_decrease = 0.117$$

# Классификация

$$p = \frac{105 + 33}{105 + 33 + 34 + 125} = 0.880 \quad \text{Доля объектов 0 класса до разбиения}$$

$$Gini_0 = 0.498$$



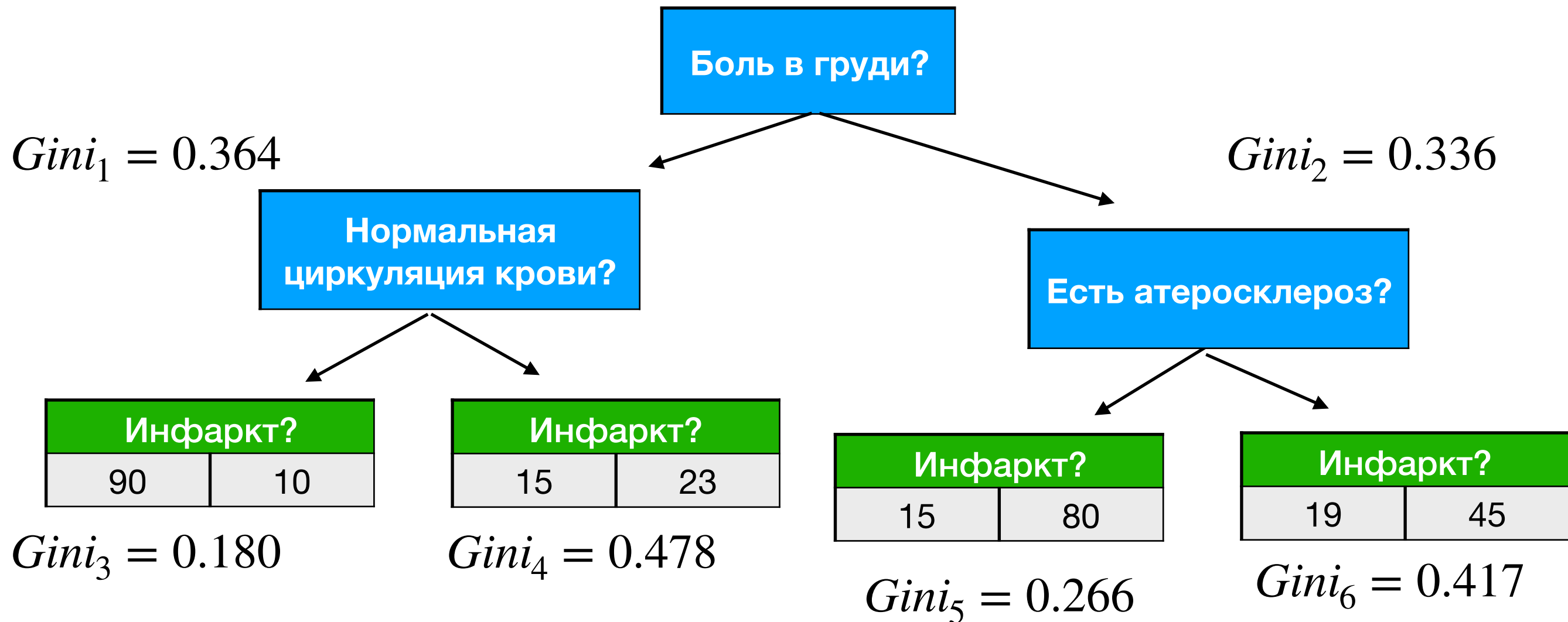
$$Gini_1 = 0.364$$

$$Gini_2 = 0.336$$

$$Impurity\_decrease = 0.149$$

Выбираем это разбиение как приводящее к наибольшему уменьшению impurity

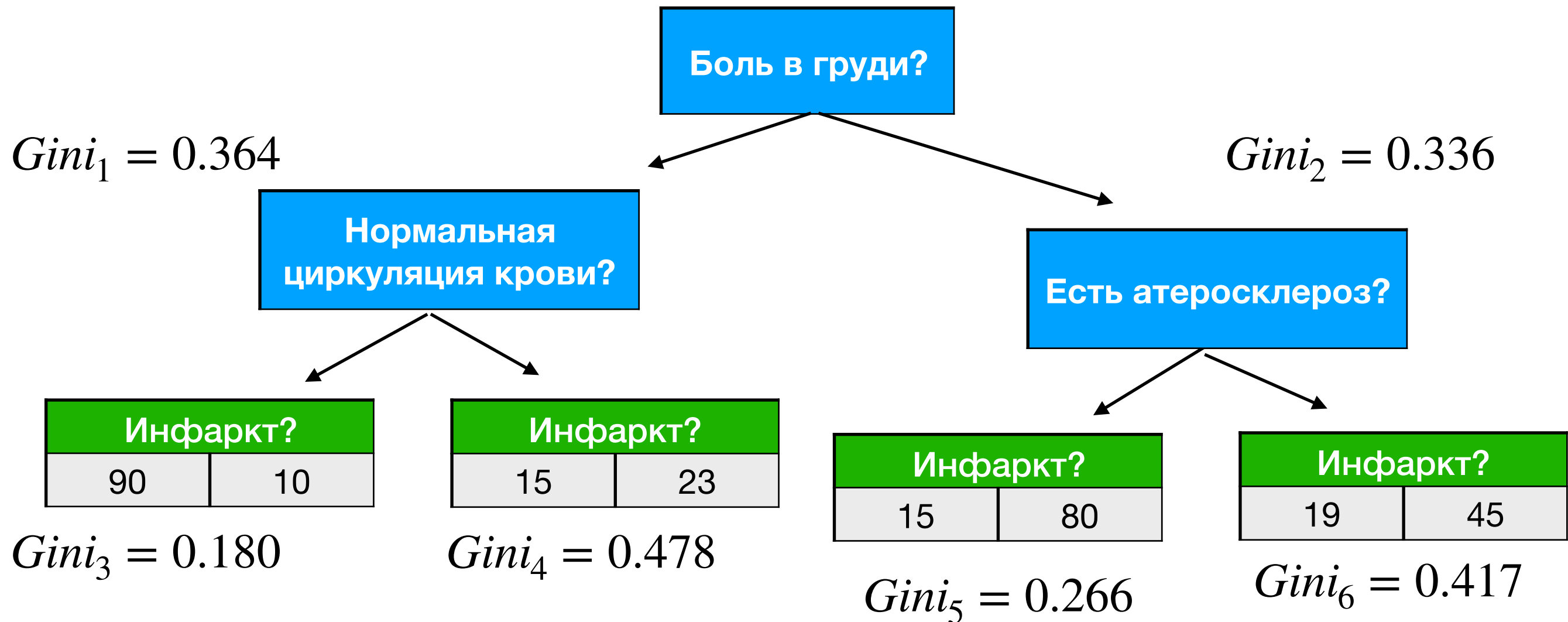
# Классификация



**Выбираем новый порог для разбиения. Для каждого узла свой, могут получиться одинаковые, могут - нет**



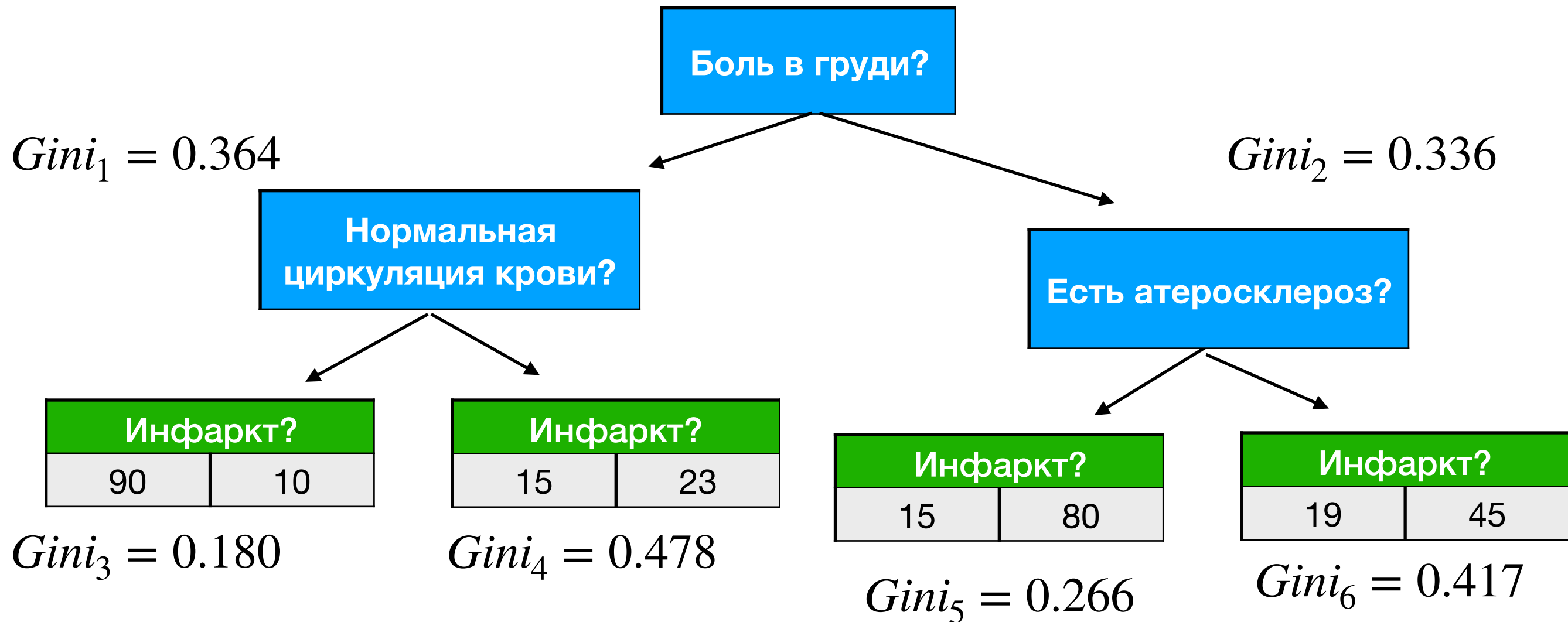
# Классификация



Выбираем новый порог для разбиения. Для каждого узла свой, могут получиться одинаковые, могут - нет

Может ли на каком-то этапе разбиение привести к ухудшению impurity?

# Классификация



Может ли на каком-то этапе разбиение  
привести к ухудшению impurity?

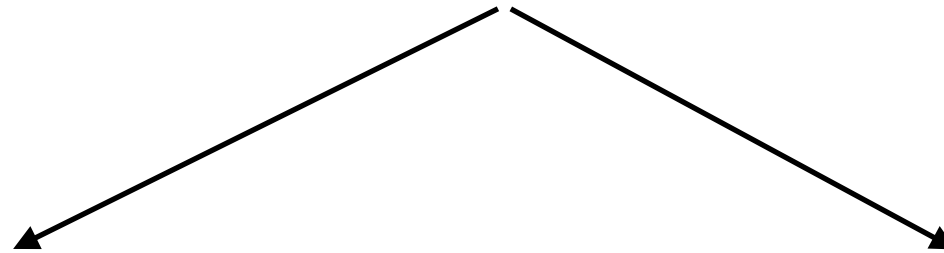
Да

# Классификация

Может ли на каком-то этапе разбиение  
привести к ухудшению *impurity*?

Да

Что делать в таком случае?



сказать, что этот узел теперь лист,  
ничего с ним не делаем (*early  
stopping*). Иногда даже ставят порог,  
что если уменьшение *impurity*  
меньше порога, то не разбивать  
узел.

Все равно бьем, в надежде, что  
следующие сплиты будут лучше

# Как работать с вещественными переменными?

...	...	Давле ние	Инфа ркт
...	...	170	Да
...	...	134	Да
...	...	50	Нет
...	...	100	Да
...	...	...	...

Сортируем по  
переменной



...	...	Давле ние	Инфа ркт
...	...	50	Нет
...	...	134	Да
...	...	...	...
...	...	170	Да
...	...	...	...

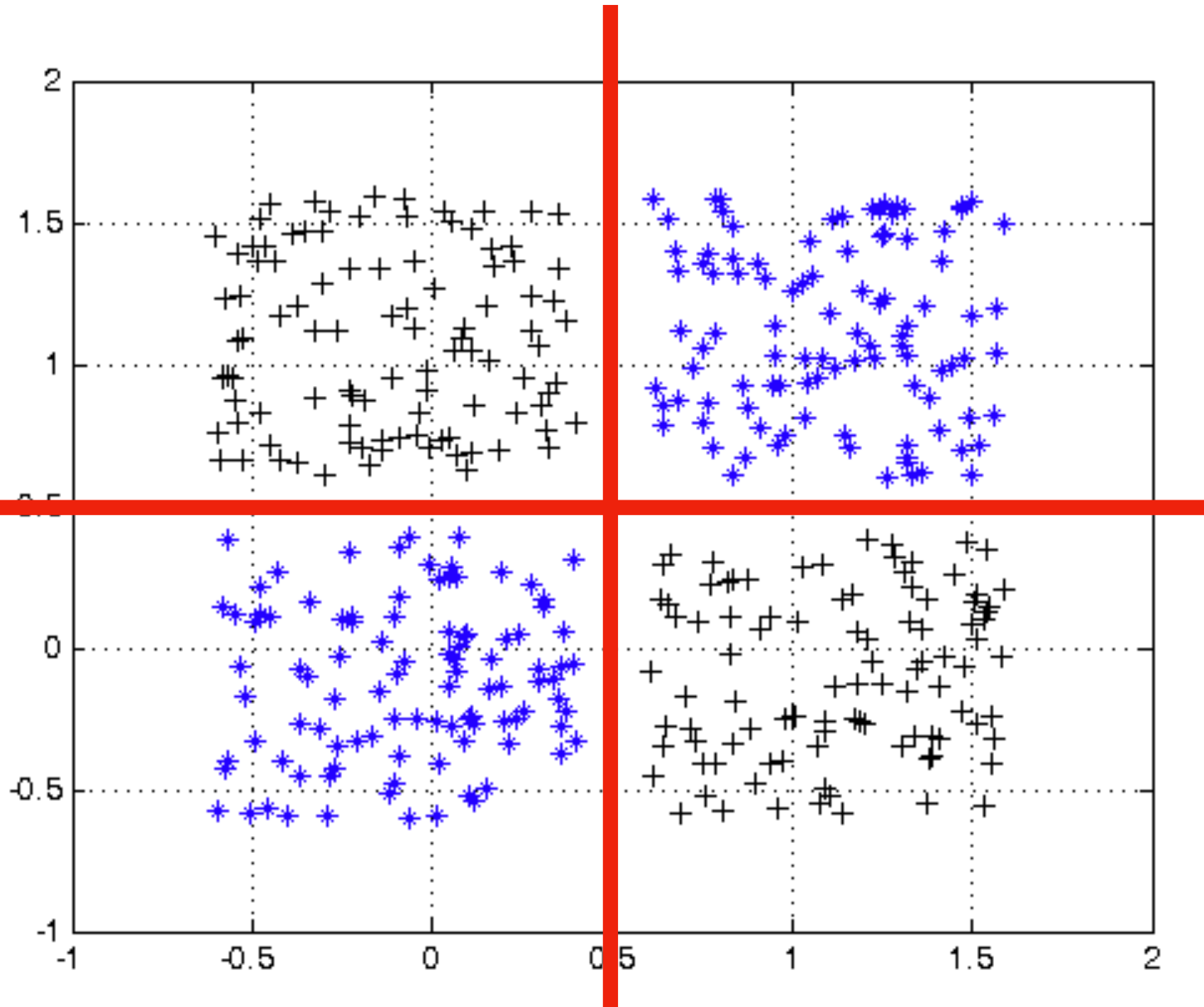
# Как работать с вещественными переменными?

...	...	Давле ние	Инфа ркт
...	...	50	Нет
...	...	134	Да
...	...	...	...
...	...	170	Да
...	...	...	...

Для каждого возможного порога  
делаем разбиение и считаем impurity

**Почему early stopping  
может быть плох?**

# Почему early stopping может быть плох?

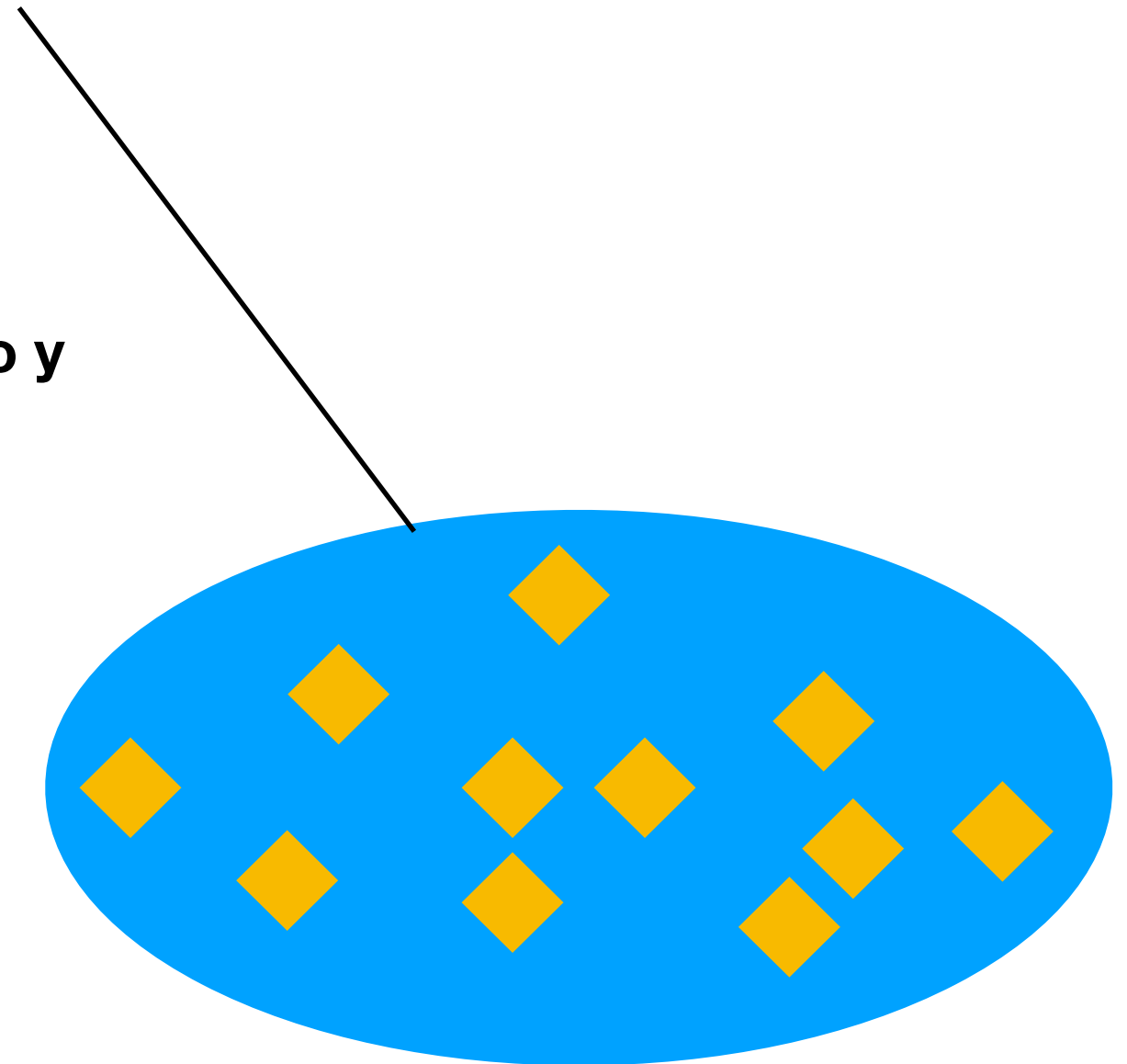


**Одной прямой  
все не  
разделить.  
Любое  
разбиение не  
улучшает  
ситуацию  
Нужно минимум  
две!**



# Регрессия

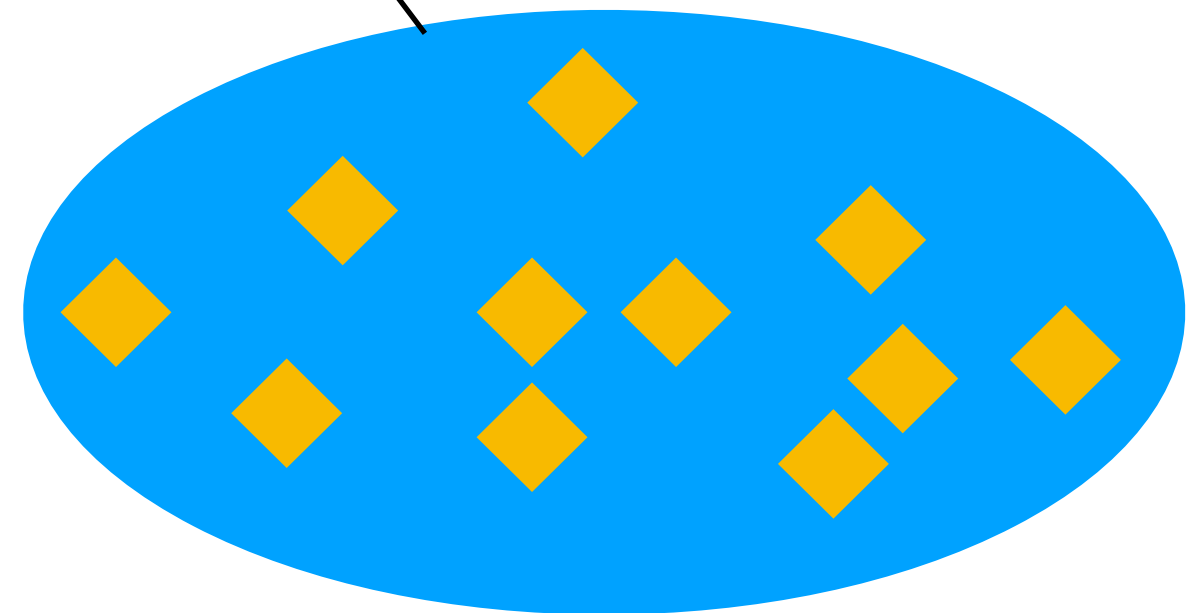
◆ - наши объекты, для каждого известно  $y$



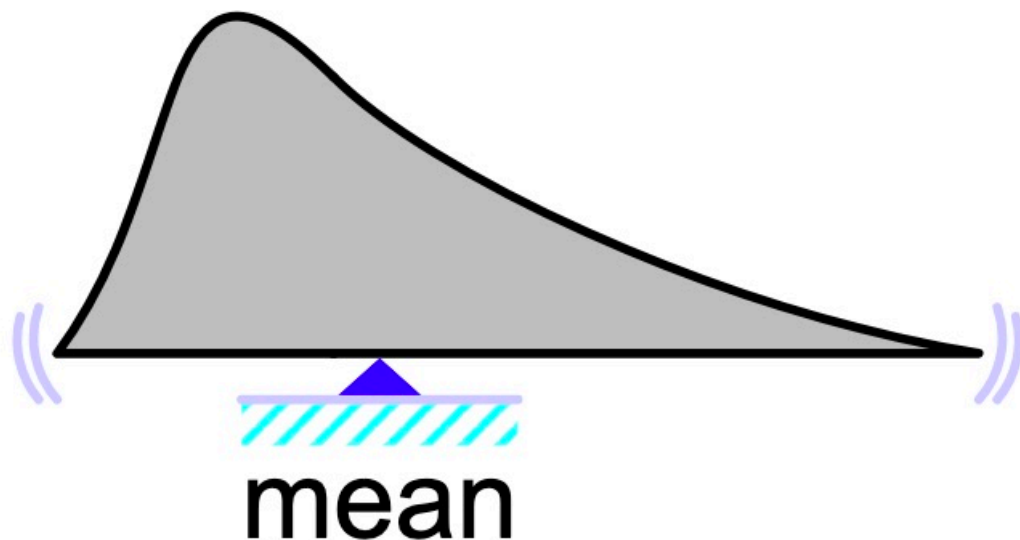
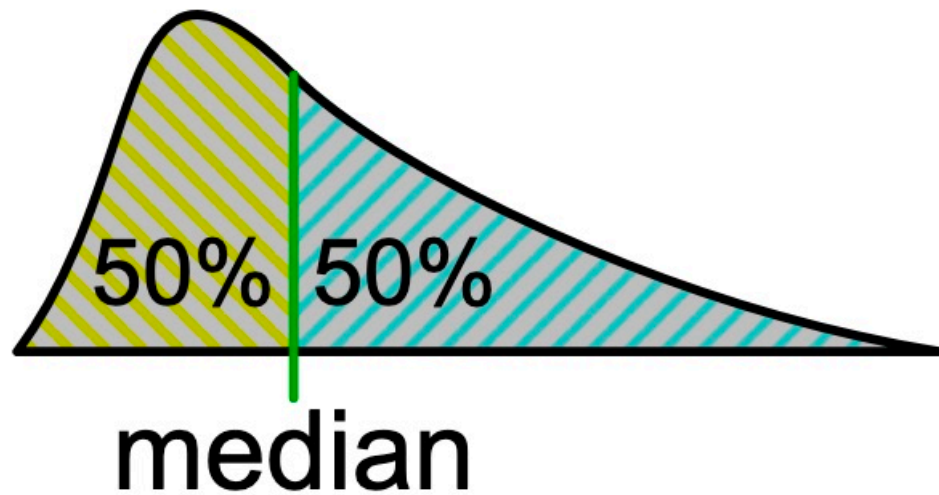
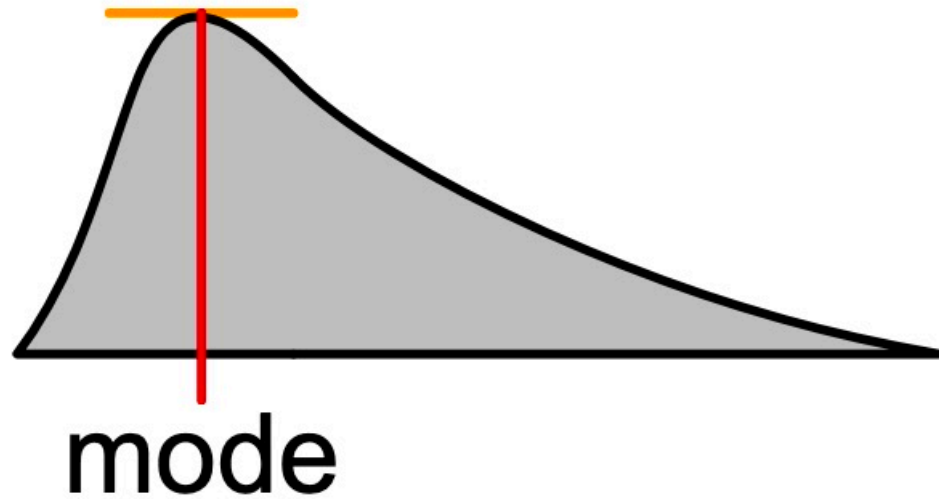
# Регрессия

◆ - наши объекты, для каждого известно  $y$

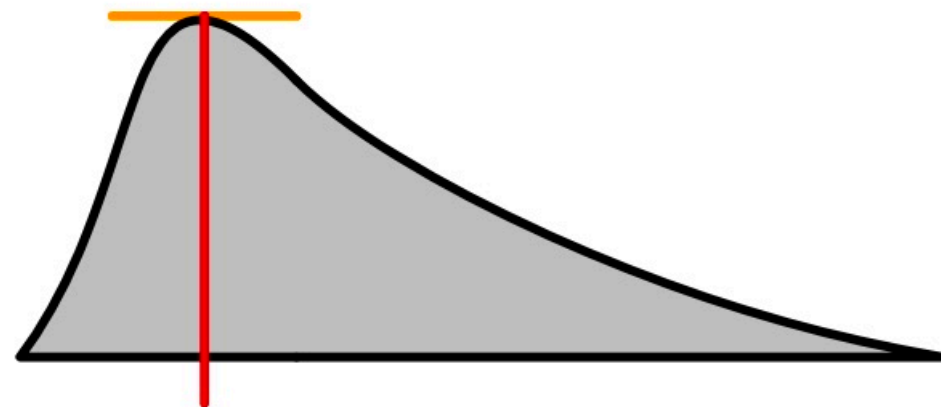
Какое число предсказать?



# Метрики регрессии

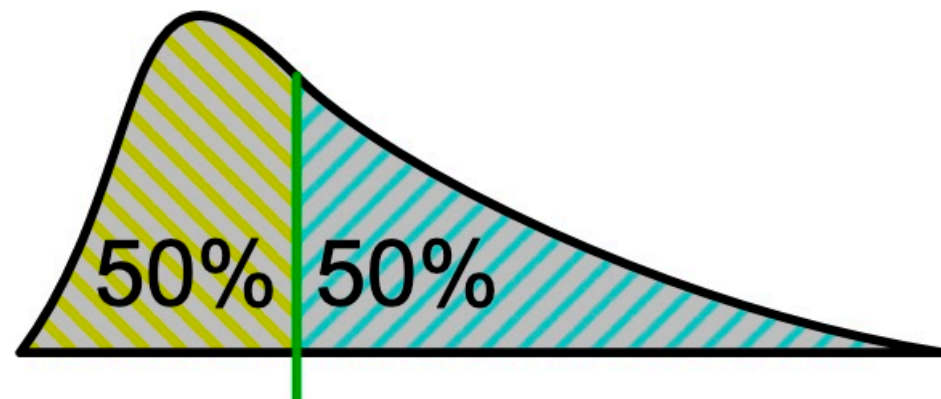


# Метрики регрессии

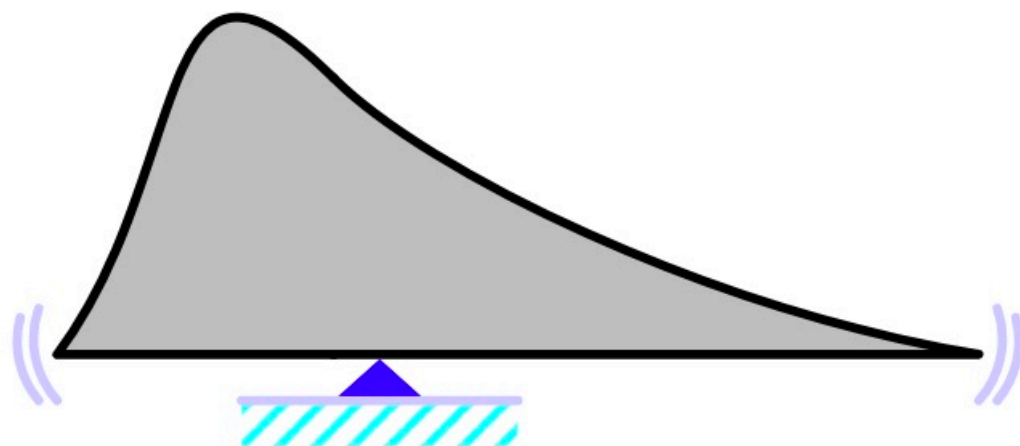


mode

Считаем число  
совпавших с реальным  
значением предсказаний

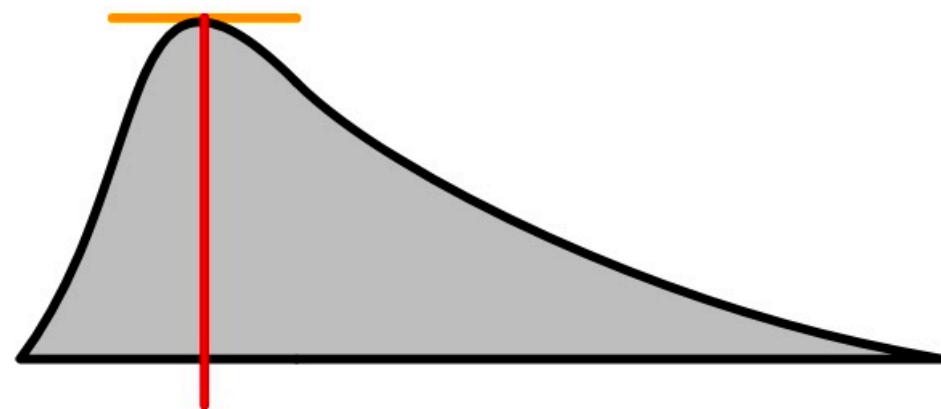


median



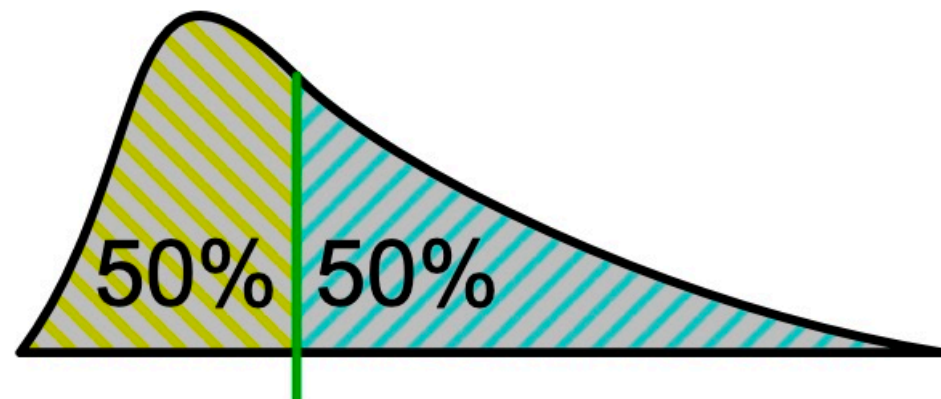
mean

# Метрики регрессии



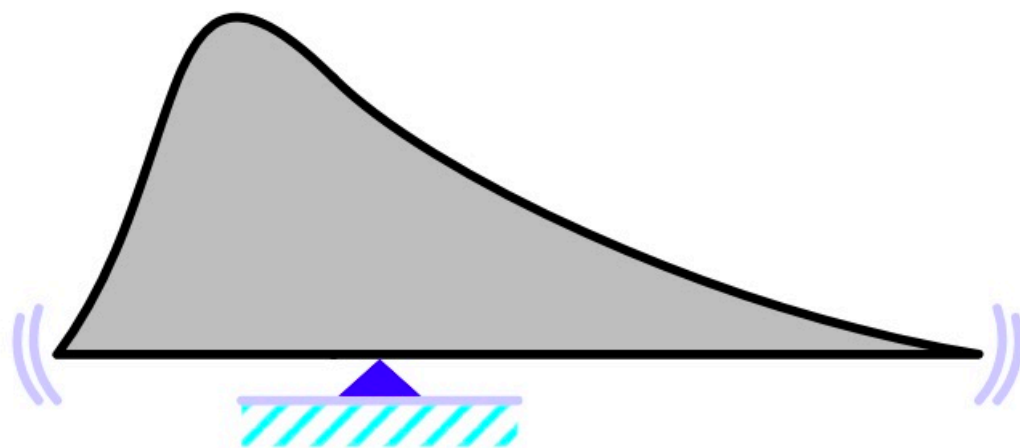
mode

Считаем число  
совпавших с реальным  
значением предсказаний



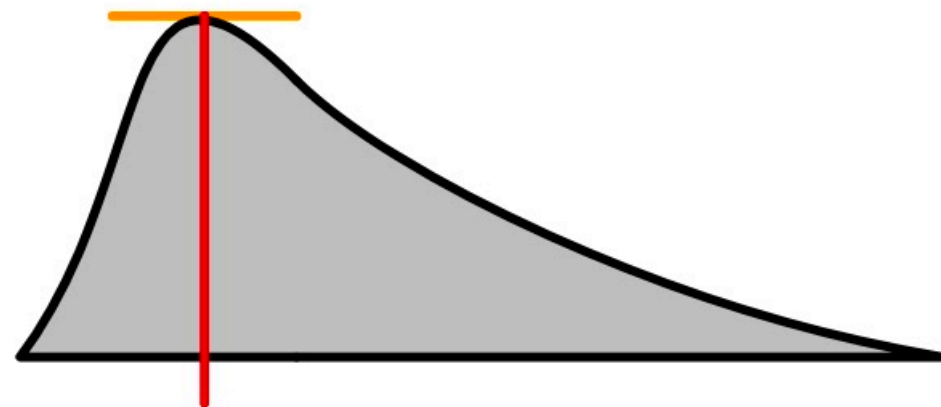
median

$$MAE = \frac{1}{N} \sum_i |y_i - \hat{y}_i|$$



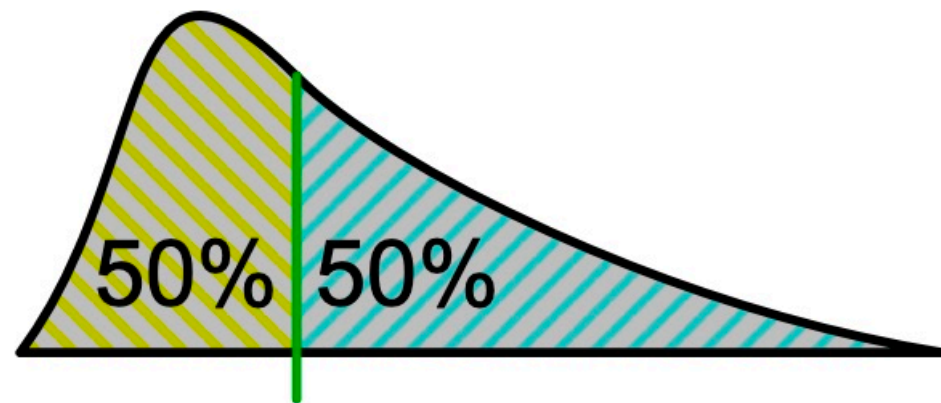
mean

# Метрики регрессии



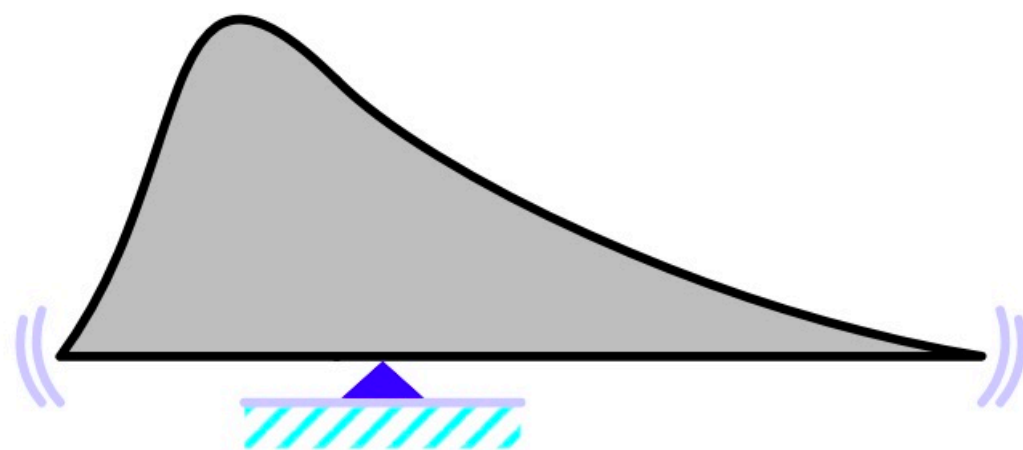
mode

Считаем число  
совпавших с реальным  
значением предсказаний



median

$$MAE = \frac{1}{N} \sum_i |y_i - \hat{y}_i|$$



mean

$$MSE = \frac{1}{N} \sum_i (y_i - \hat{y}_i)^2$$

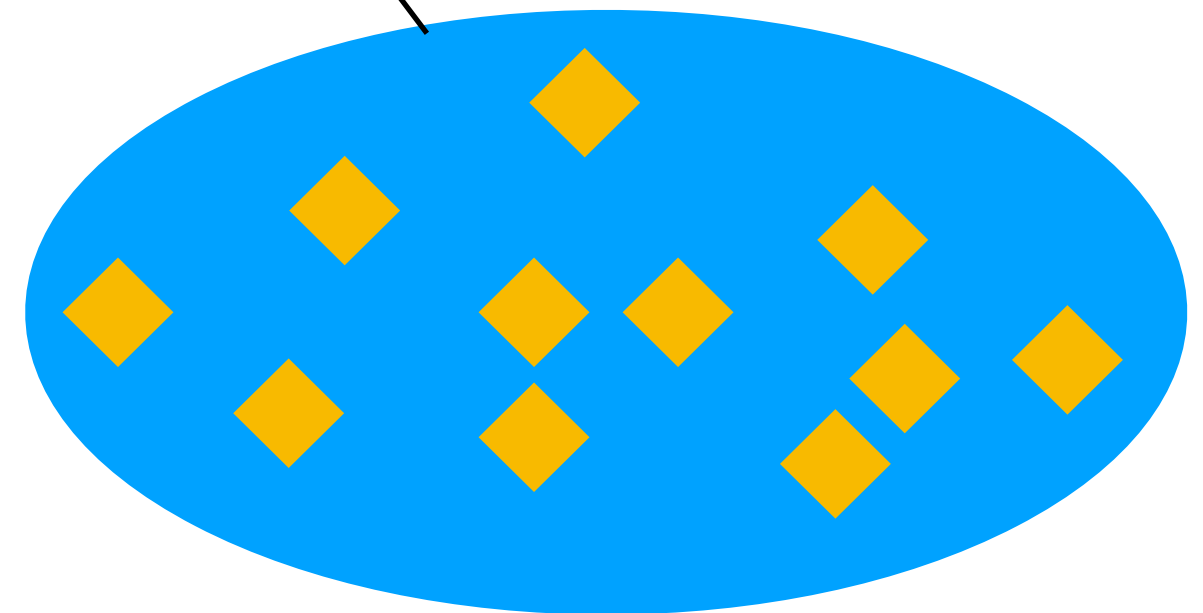
# Регрессия

◆ - наши объекты, для каждого известно  $y$

Какое число предсказать?

Обычно нам интересно MSE

Предсказываем среднее





# Оценка среднего в популяции

$$\hat{\mu} = \frac{\sum_i x_i}{N}$$

$$E(\hat{\mu}) = \mu$$

$$sd(\hat{\mu}) = \frac{\sigma}{\sqrt{N}}$$

# Разбиение регрессионного дерева

$$R_1(j, s) = \{X | X_j \leq s\} \quad \text{and} \quad R_2(j, s) = \{X | X_j > s\}.$$

Мера качества узла - дисперсия оценки  $R_0$ . Остальное - так же, как с классификацией

$$\frac{D_{R_1} \cdot N_1 + D_{R_2} \cdot N_2}{N_1 + N_2} < D_{R_0}$$

# **Категориальные признаки**

**Как разбивать категориальные признаки?**

# **Категориальные признаки**

**Как разбивать категориальные признаки?**

**Всего  $2^N$  возможных разбиений**

# Категориальные признаки

Как разбивать категориальные признаки?

Всего  $2^N$  возможных разбиений

Оказывается, можно отсортировать категории и проверить  $N$  разбиений



Что по оси?

# Категориальные признаки

Как разбивать категориальные признаки?

Всего  $2^N$  возможных разбиений

Оказывается, можно отсортировать категории и проверить  $N$  разбиений

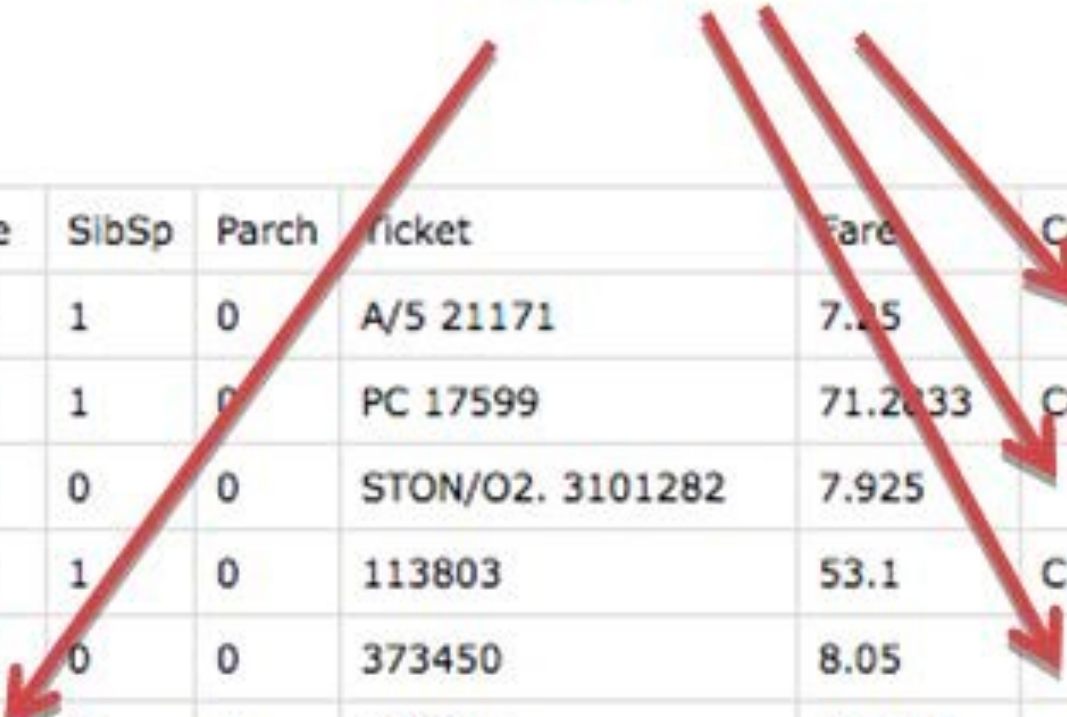


$$E_{x_i == k}(y)$$

Среднее значение  $y$  для объектов, у которых категориальная переменная  $x_i$  равно данному значению

# Пропущенные значения

Missing values



PassengerId	Survived	Pclass	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	male	22	1	0	A/5 21171	7.25		S
2	1	1	female	38	1	0	PC 17599	71.2833	C85	C
3	1	3	female	26	0	0	STON/O2. 3101282	7.925		S
4	1	1	female	35	1	0	113803	53.1	C123	S
5	0	3	male	35	0	0	373450	8.05		S
6	0	3	male		0	0	330877	8.4583		Q

Как с ними бороться?

# Пропущенные значения

Missing values

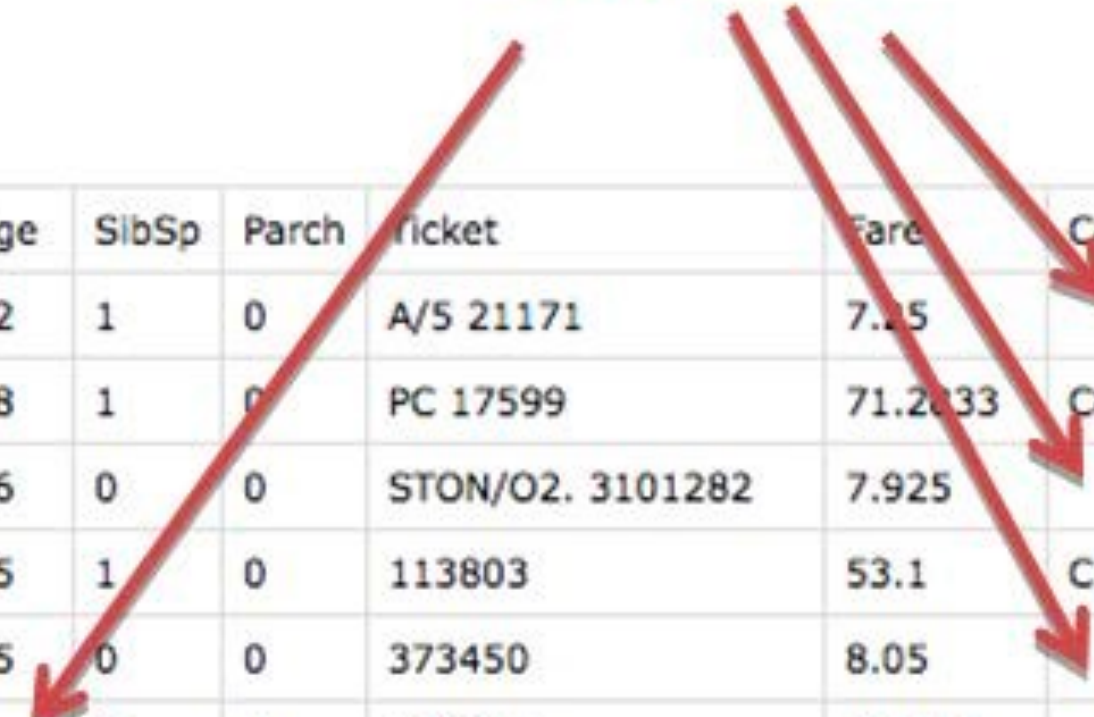
PassengerId	Survived	Pclass	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	male	22	1	0	A/5 21171	7.25		S
2	1	1	female	38	1	0	PC 17599	71.2833	C85	C
3	1	3	female	26	0	0	STON/O2. 3101282	7.925		S
4	1	1	female	35	1	0	113803	53.1	C123	S
5	0	3	male	35	0	0	373450	8.05		S
6	0	3	male		0	0	330877	8.4583		Q

Вариант1: убрать все объекты с пропущенными значениями



# Пропущенные значения

Missing values



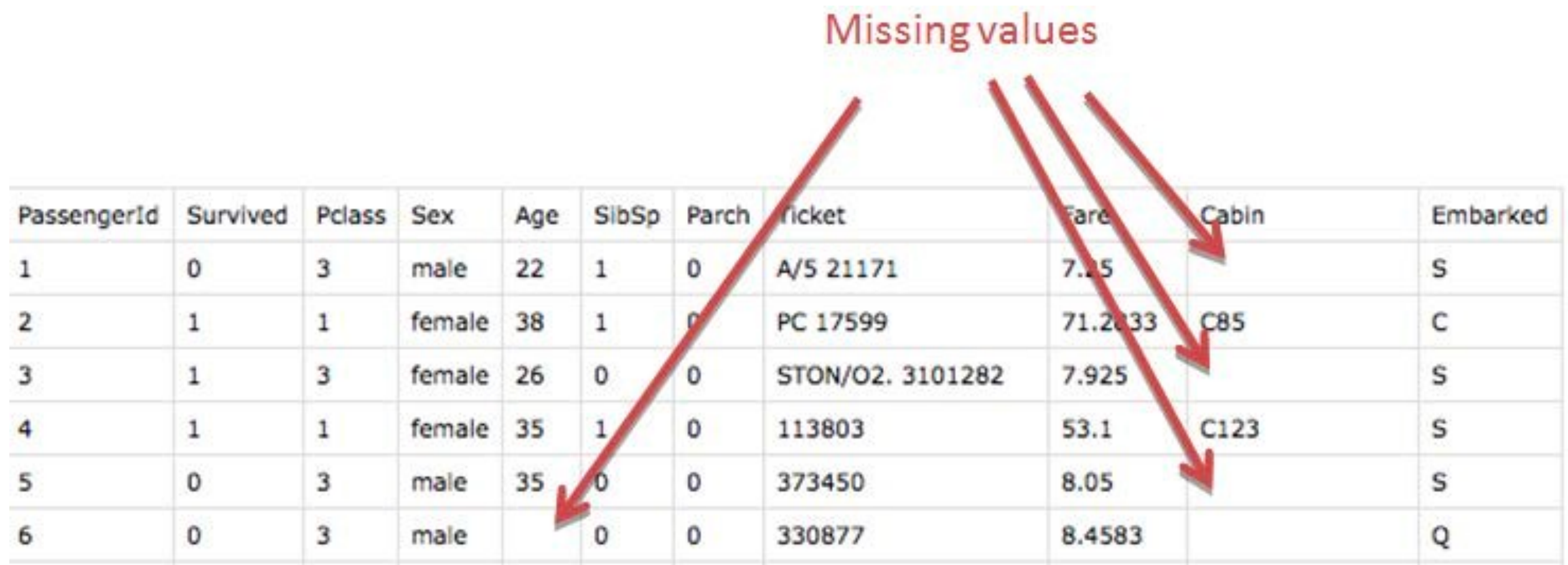
PassengerId	Survived	Pclass	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	male	22	1	0	A/5 21171	7.25		S
2	1	1	female	38	1	0	PC 17599	71.2833	C85	C
3	1	3	female	26	0	0	STON/O2. 3101282	7.925		S
4	1	1	female	35	1	0	113803	53.1	C123	S
5	0	3	male	35	0	0	373450	8.05		S
6	0	3	male		0	0	330877	8.4583		Q

Вариант 1: убрать все объекты с пропущенными значениями

А если данных мало?

# Пропущенные значения

Missing values

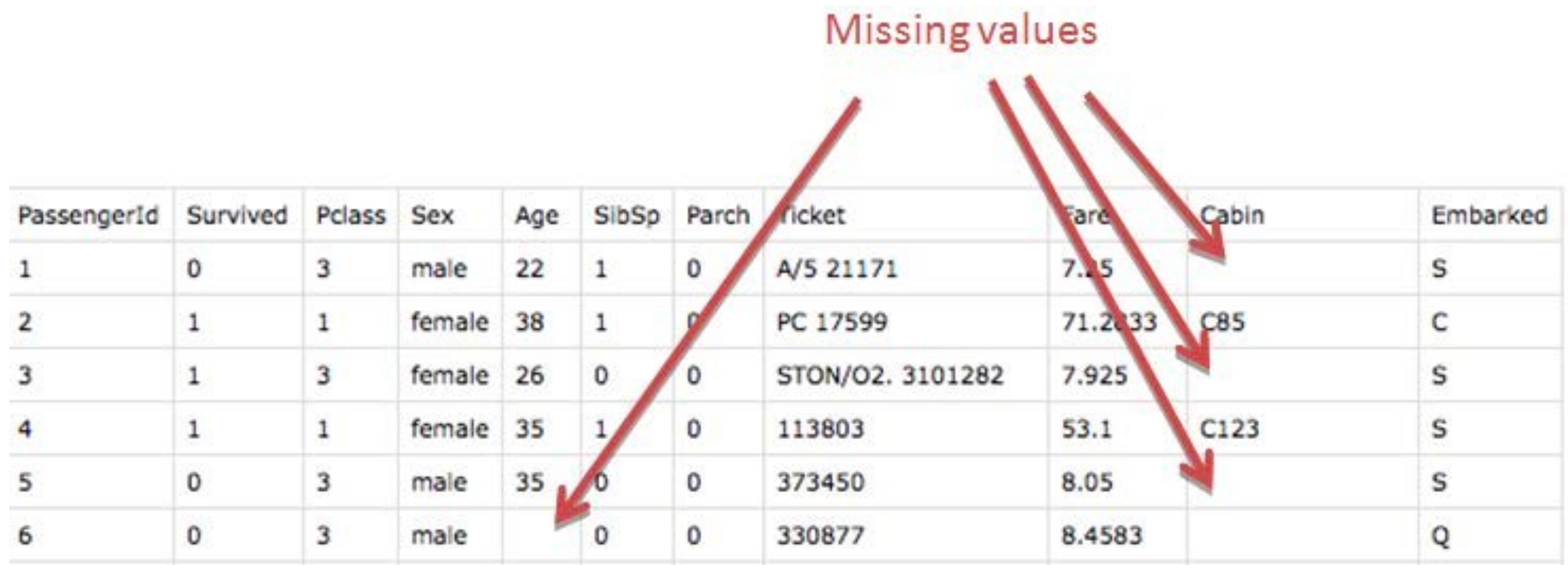


PassengerId	Survived	Pclass	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	male	22	1	0	A/5 21171	7.25		S
2	1	1	female	38	1	0	PC 17599	71.2833	C85	C
3	1	3	female	26	0	0	STON/O2. 3101282	7.925		S
4	1	1	female	35	1	0	113803	53.1	C123	S
5	0	3	male	35	0	0	373450	8.05		S
6	0	3	male		0	0	330877	8.4583		Q

Вариант 2: заполнить пропущенные значения до обучения. Как?

# Пропущенные значения

Missing values



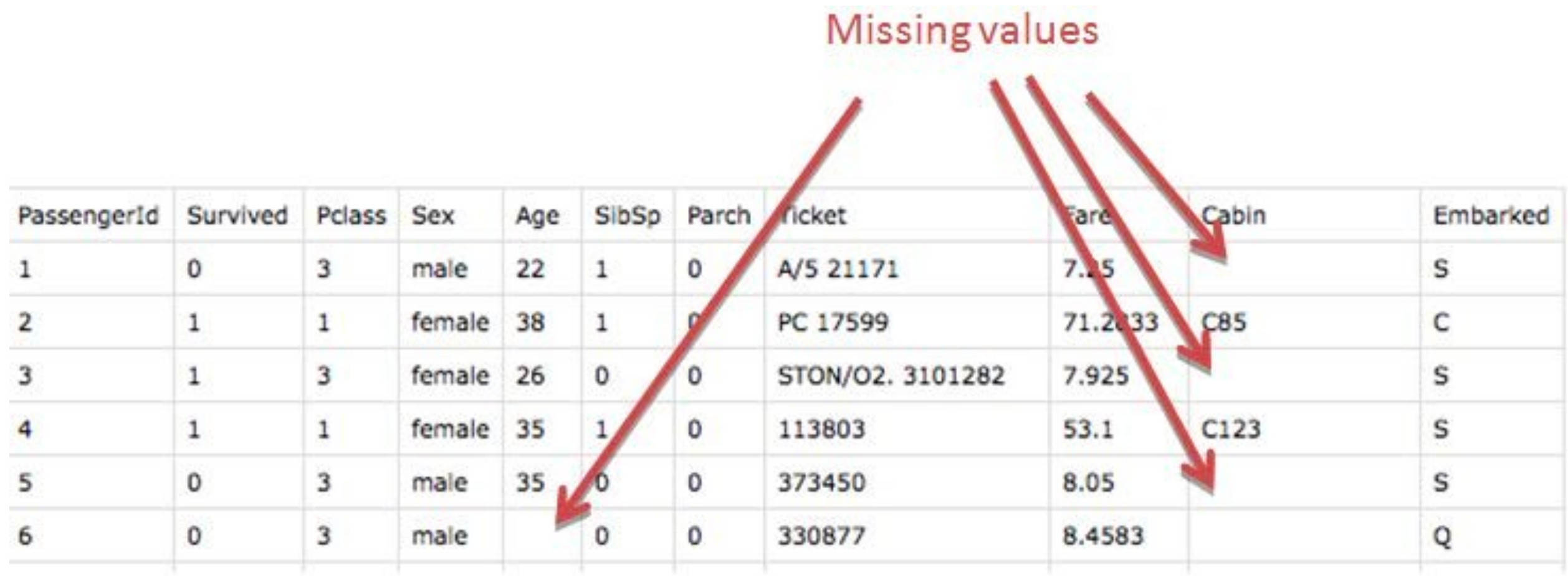
PassengerId	Survived	Pclass	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	male	22	1	0	A/5 21171	7.25		S
2	1	1	female	38	1	0	PC 17599	71.2833	C85	C
3	1	3	female	26	0	0	STON/O2. 3101282	7.925		S
4	1	1	female	35	1	0	113803	53.1	C123	S
5	0	3	male	35	0	0	373450	8.05		S
6	0	3	male		0	0	330877	8.4583		Q

Вариант 2: заполнить пропущенные значения до обучения. Как?

а) предполагаем простое распределение для наших данных, заполняем пропуски средним значением, медианой и тд

# Пропущенные значения

Missing values



PassengerId	Survived	Pclass	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	male	22	1	0	A/5 21171	7.25		S
2	1	1	female	38	1	0	PC 17599	71.2833	C85	C
3	1	3	female	26	0	0	STON/O2. 3101282	7.925		S
4	1	1	female	35	1	0	113803	53.1	C123	S
5	0	3	male	35	0	0	373450	8.05		S
6	0	3	male		0	0	330877	8.4583		Q

Вариант 2: заполнить пропущенные значения до обучения. Как?

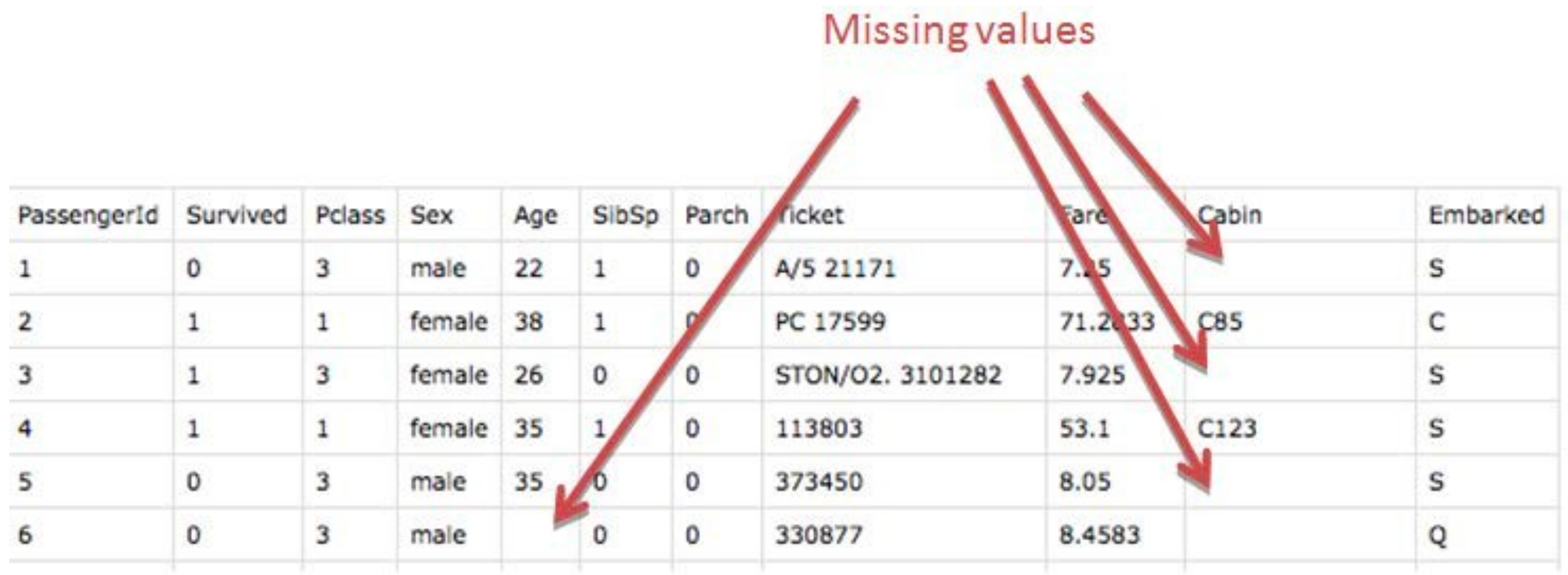
а) предполагаем простое распределение для наших данных, заполняем пропуски средним значением, медианой и тд

Часто плохо! Мы предполагаем, что пропущенные значения пропущены истинно случайно - факт пропуска не зависит от других переменных и от предсказываемой величины



# Пропущенные значения

Missing values



PassengerId	Survived	Pclass	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	male	22	1	0	A/5 21171	7.25		S
2	1	1	female	38	1	0	PC 17599	71.2833	C85	C
3	1	3	female	26	0	0	STON/O2. 3101282	7.925		S
4	1	1	female	35	1	0	113803	53.1	C123	S
5	0	3	male	35	0	0	373450	8.05		S
6	0	3	male		0	0	330877	8.4583		Q

Вариант 2: заполнить пропущенные значения до обучения. Как?

б) Учим дополнительные модели машинного обучения. Будут предсказывать нам пропущенные значения

# Пропущенные значения

Missing values

PassengerId	Survived	Pclass	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	male	22	1	0	A/5 21171	7.25		S
2	1	1	female	38	1	0	PC 17599	71.2833	C85	C
3	1	3	female	26	0	0	STON/O2. 3101282	7.925		S
4	1	1	female	35	1	0	113803	53.1	C123	S
5	0	3	male	35	0	0	373450	8.05		S
6	0	3	male		0	0	330877	8.4583		Q

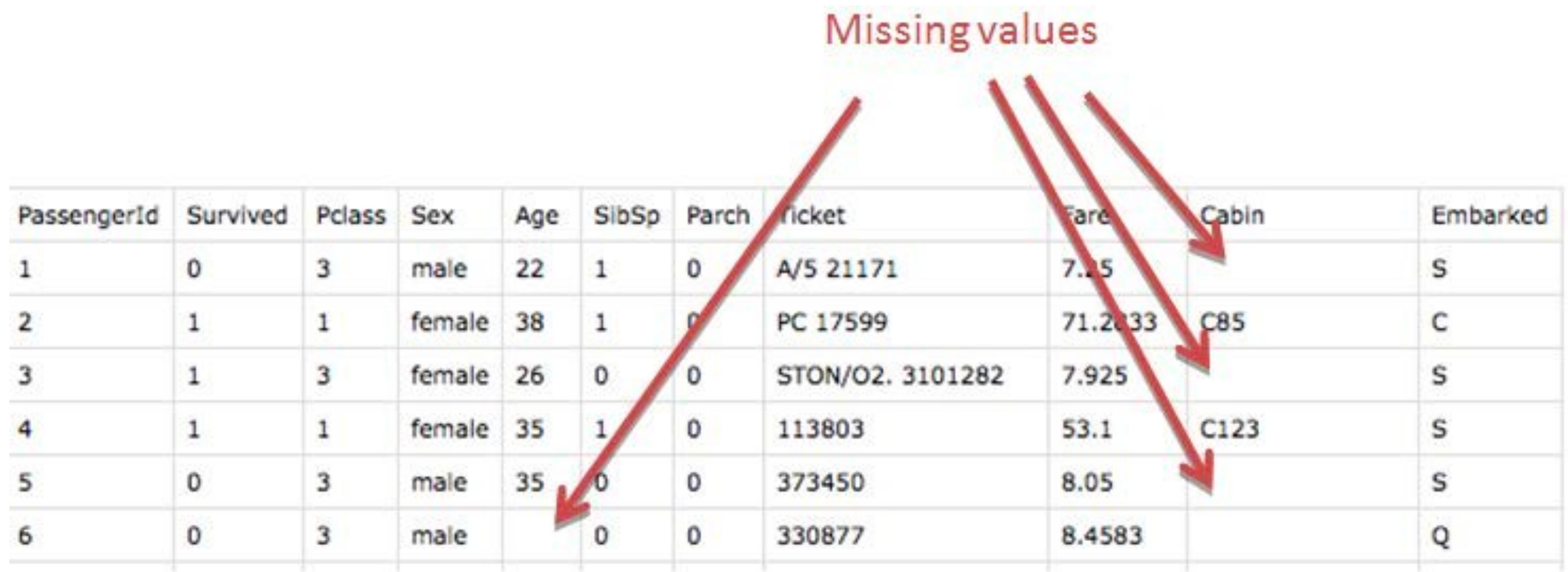
Вариант 2: заполнить пропущенные значения до обучения. Как?

б) Учим дополнительные модели машинного обучения. Будут предсказывать нам пропущенные значения

А переобучение? А если признаков очень много? А если в каждом объекте что-то да пропущено?

# Пропущенные значения

Missing values



PassengerId	Survived	Pclass	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	male	22	1	0	A/5 21171	7.25		S
2	1	1	female	38	1	0	PC 17599	71.2833	C85	C
3	1	3	female	26	0	0	STON/O2. 3101282	7.925		S
4	1	1	female	35	1	0	113803	53.1	C123	S
5	0	3	male	35	0	0	373450	8.05		S
6	0	3	male		0	0	330877	8.4583		Q

Вариант 3: использовать алгоритм, который умеет справляться с пропусками

# Пропущенные значения

Missing values

PassengerId	Survived	Pclass	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	male	22	1	0	A/5 21171	7.25		S
2	1	1	female	38	1	0	PC 17599	71.2833	C85	C
3	1	3	female	26	0	0	STON/O2. 3101282	7.925		S
4	1	1	female	35	1	0	113803	53.1	C123	S
5	0	3	male	35	0	0	373450	8.05		S
6	0	3	male		0	0	330877	8.4583		Q

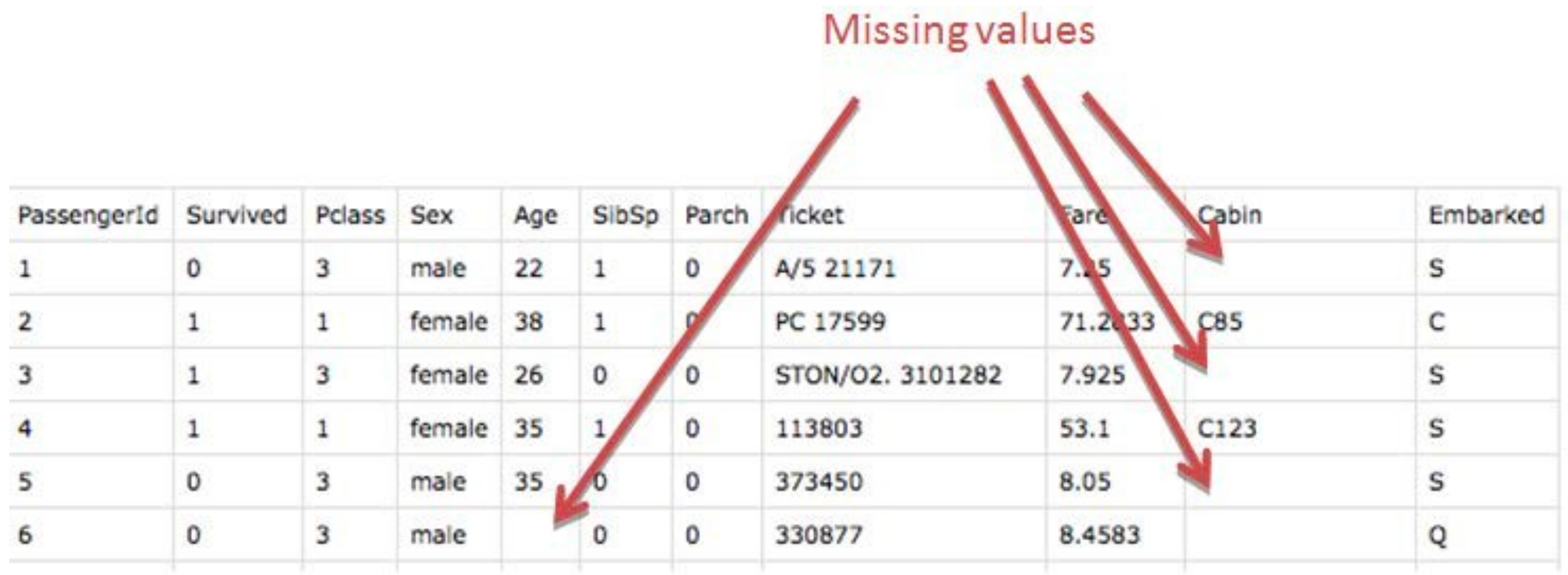
Вариант 3: использовать алгоритм, который умеет справляться с пропусками

KNN - можно брать ближайших соседей по известным признакам и на основании них восстанавливать неизвестные



# Пропущенные значения

Missing values



PassengerId	Survived	Pclass	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	male	22	1	0	A/5 21171	7.25		S
2	1	1	female	38	1	0	PC 17599	71.2833	C85	C
3	1	3	female	26	0	0	STON/O2. 3101282	7.925		S
4	1	1	female	35	1	0	113803	53.1	C123	S
5	0	3	male	35	0	0	373450	8.05		S
6	0	3	male		0	0	330877	8.4583		Q

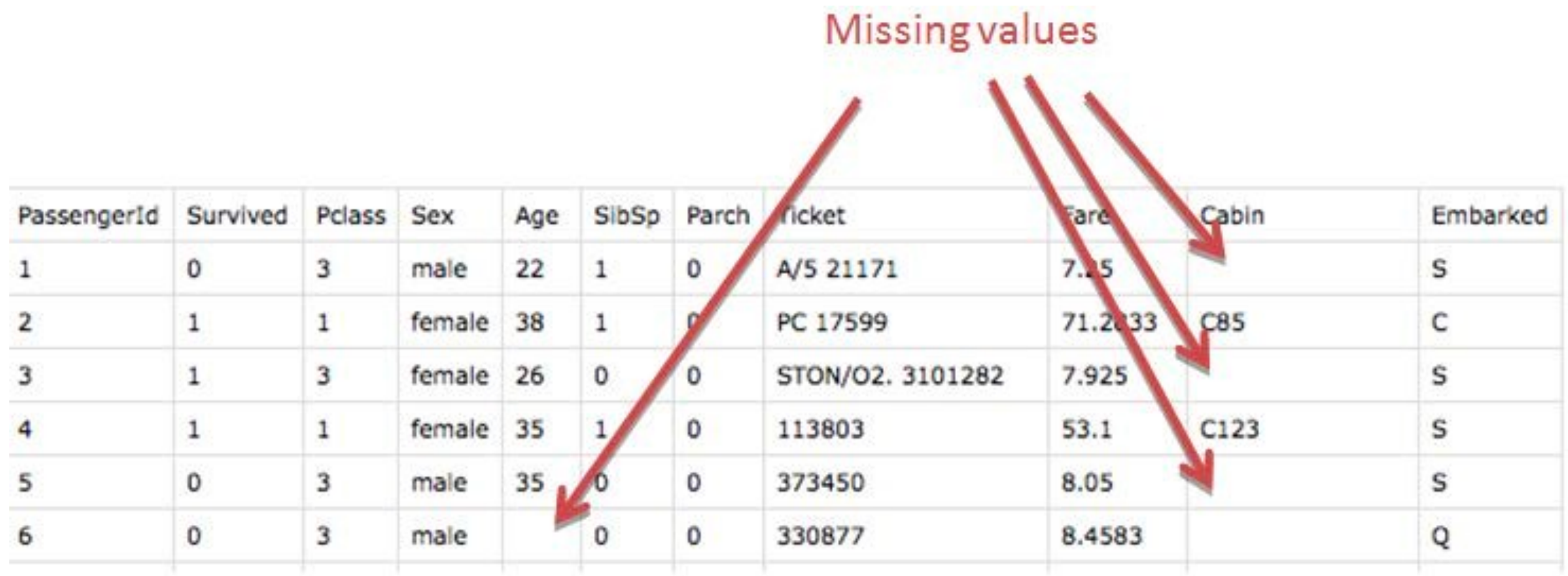
Вариант 3: использовать алгоритм, который умеет справляться с пропусками

KNN - можно брать ближайших соседей по известным признакам и на основании них восстанавливать неизвестные

Дерево решений - два способа

# Пропущенные значения

Missing values



PassengerId	Survived	Pclass	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	male	22	1	0	A/5 21171	7.25		S
2	1	1	female	38	1	0	PC 17599	71.2833	C85	C
3	1	3	female	26	0	0	STON/O2. 3101282	7.925		S
4	1	1	female	35	1	0	113803	53.1	C123	S
5	0	3	male	35	0	0	373450	8.05		S
6	0	3	male		0	0	330877	8.4583		Q

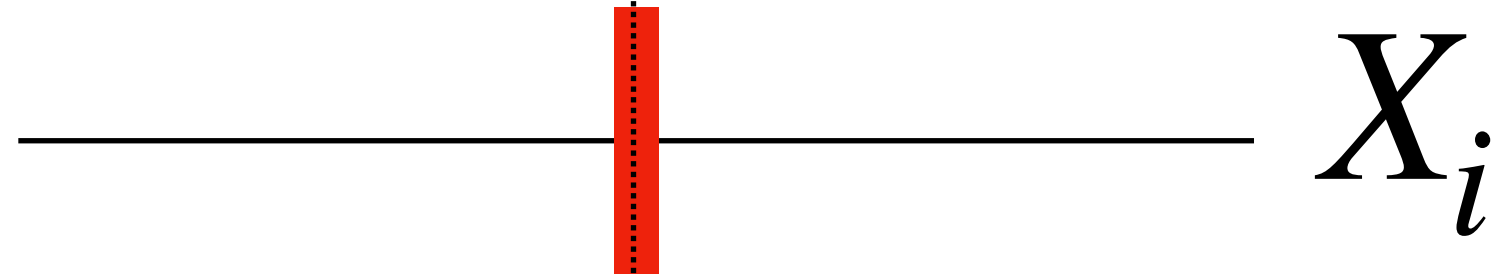
Вариант 3: использовать алгоритм, который умеет справляться с пропусками

Дерево решений - два способа

1) пропущенное значение - особая категория данных. В дереве в правилах напрямую спрашиваем - а не пропущено ли наше значение

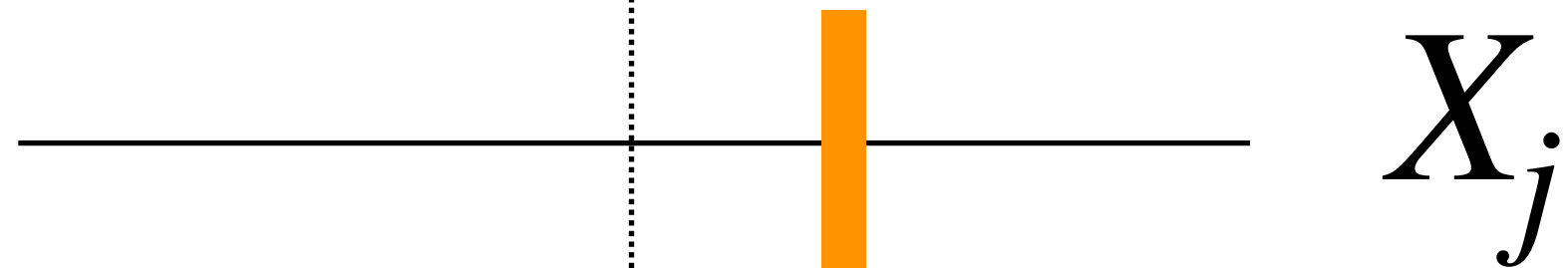
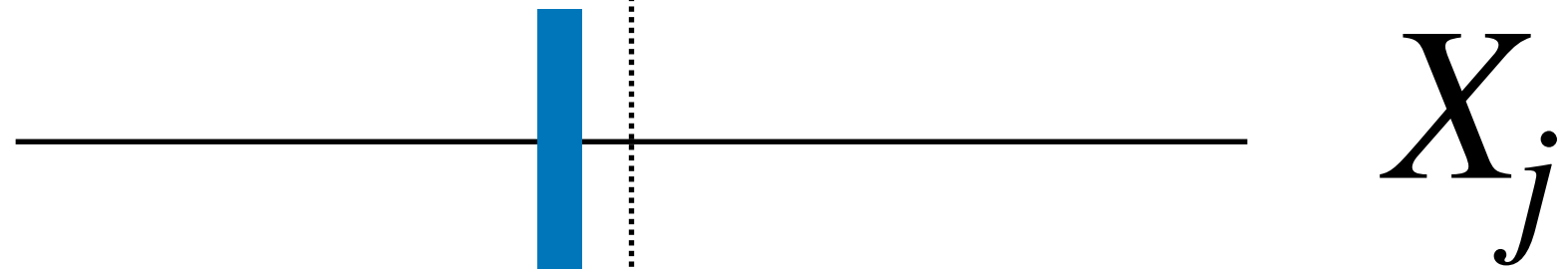
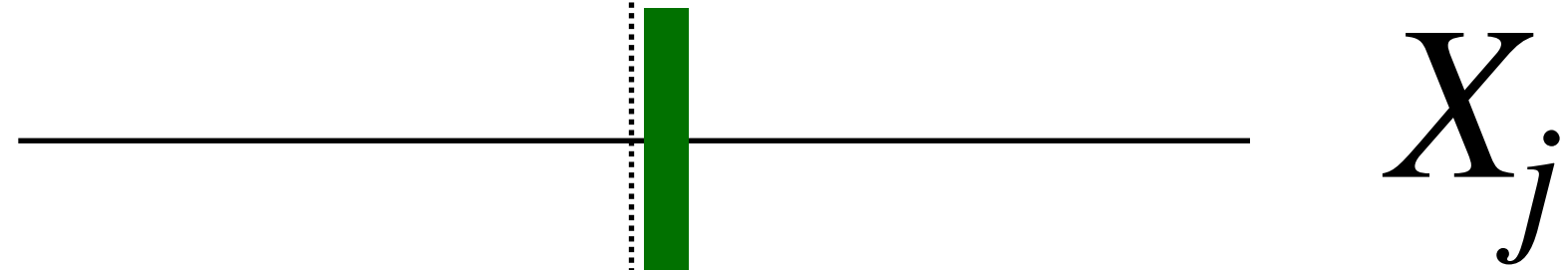
# Пропущенные значения

Дерево решений - суррогатные сплиты



Допустим в данном узле мы разбили по  $i$ -му признаку и для текущего объекта он не известен

Бьем по первому известному признаку, который дает наиболее похожее разбиение. Иногда для простоты можно просто взять признак, наиболее скорелированный с данным



# Деревья неустойчивы

- ▶ Незначительные изменения в данных приводят к значительным изменениям в топологии дерева



Взято из презентации Гулин В.,  
Техносфера

# Деревья переобучаются

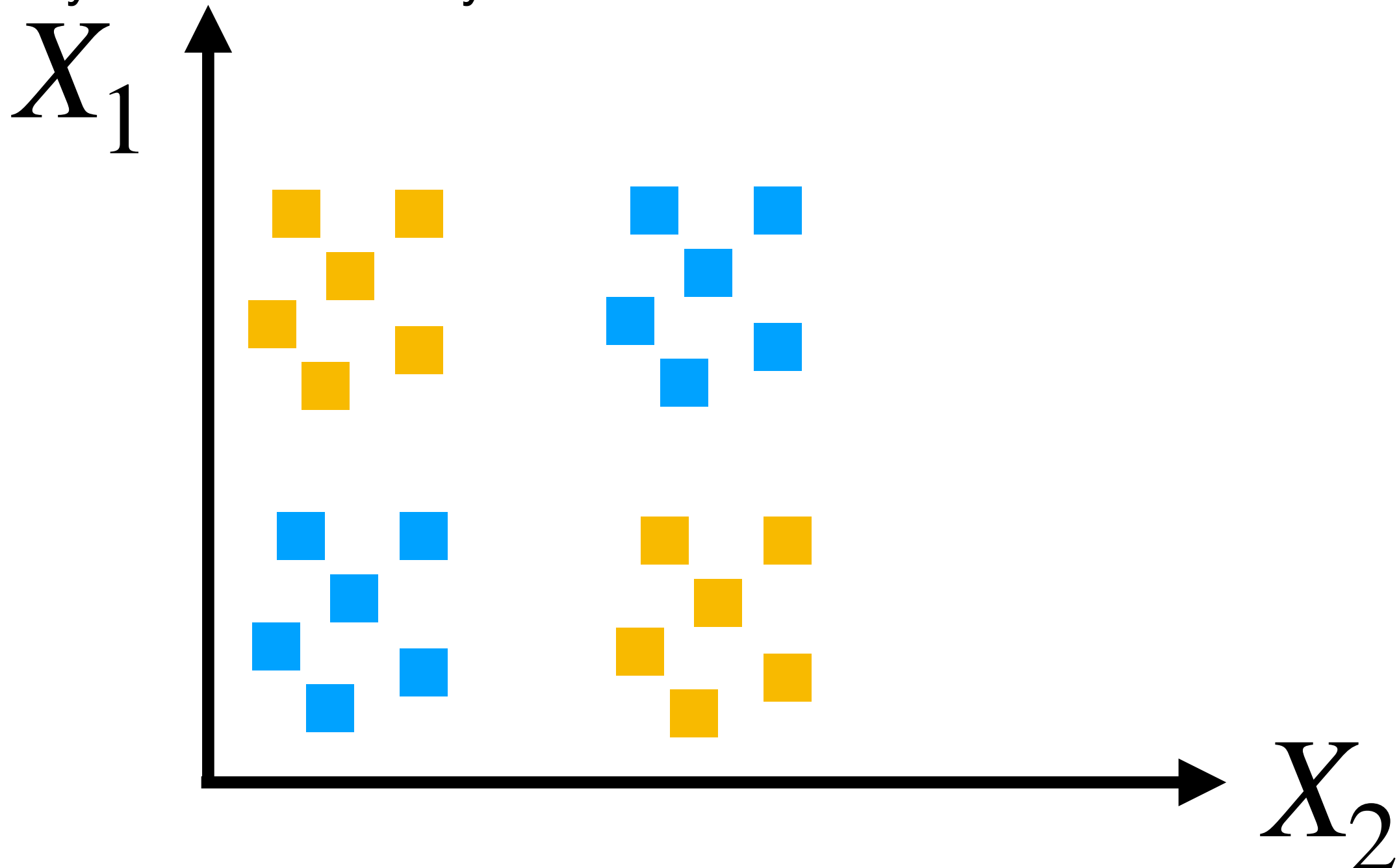
Как с этим бороться?

Можно ли просто запретить разбивать узел, если от этого качество вырастает меньше чем на какую-то константу?

# Деревья переобучаются

Как с этим бороться?

Можно ли просто запретить разбивать узел, если от этого качество вырастает меньше чем на какую-то константу?



# Деревья переобучаются

Как с этим бороться?

Можно ли просто запретить разбивать узел, если от этого качество вырастает меньше чем на какую-то константу?

Иногда дереву приходится делать плохое разбиение, чтобы далее сделать хорошее разбиение

# Деревья переобучаются

Как с этим бороться?

Можно ли ограничить число объектов в листе? Не делаем разбиение, если в результате объектов в листе будет слишком мало



# Деревья переобучаются

Как с этим бороться?

Можно ли ограничить число объектов в листе? Не делаем разбиение, если в результате объектов в листе будет слишком мало

Можно. Почему?

# Деревья переобучаются

Как с этим бороться?

Можно ли ограничить число объектов в листе? Не делаем разбиение, если в результате объектов в листе будет слишком мало

Можно. Почему?

Мы оцениваем параметр в подпространстве. Чем объектов, тем больше дисперсия оценки -> больше переобученность модели.

# Деревья переобучаются

Можно обрезать уже построенное дерево!

$$Tree\_score(T, train) = Quality(T, train) + \alpha \cdot Tree\_complexity(T)$$

# Деревья переобучаются

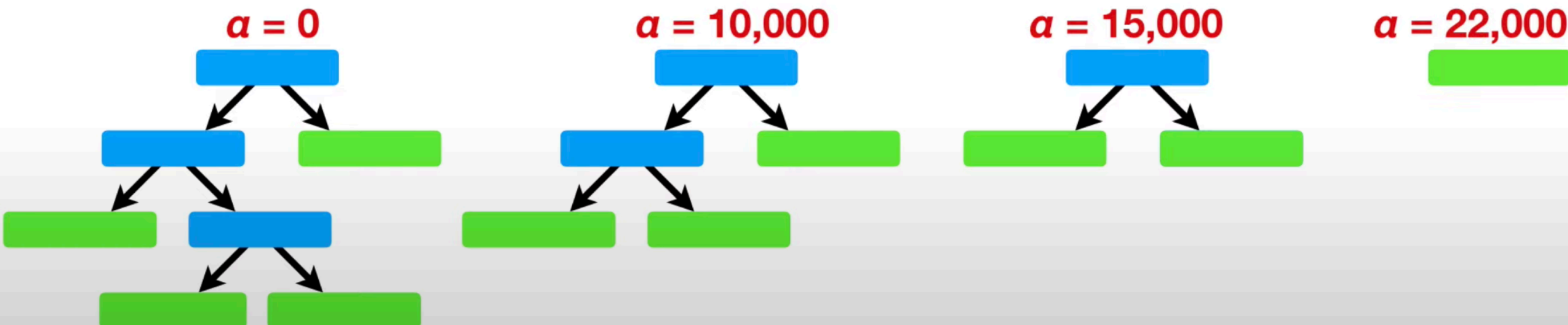
Можно обрезать уже построенное дерево!

$$Tree\_score(T, train) = Quality(T, train) + \alpha \cdot Tree\_complexity(T)$$

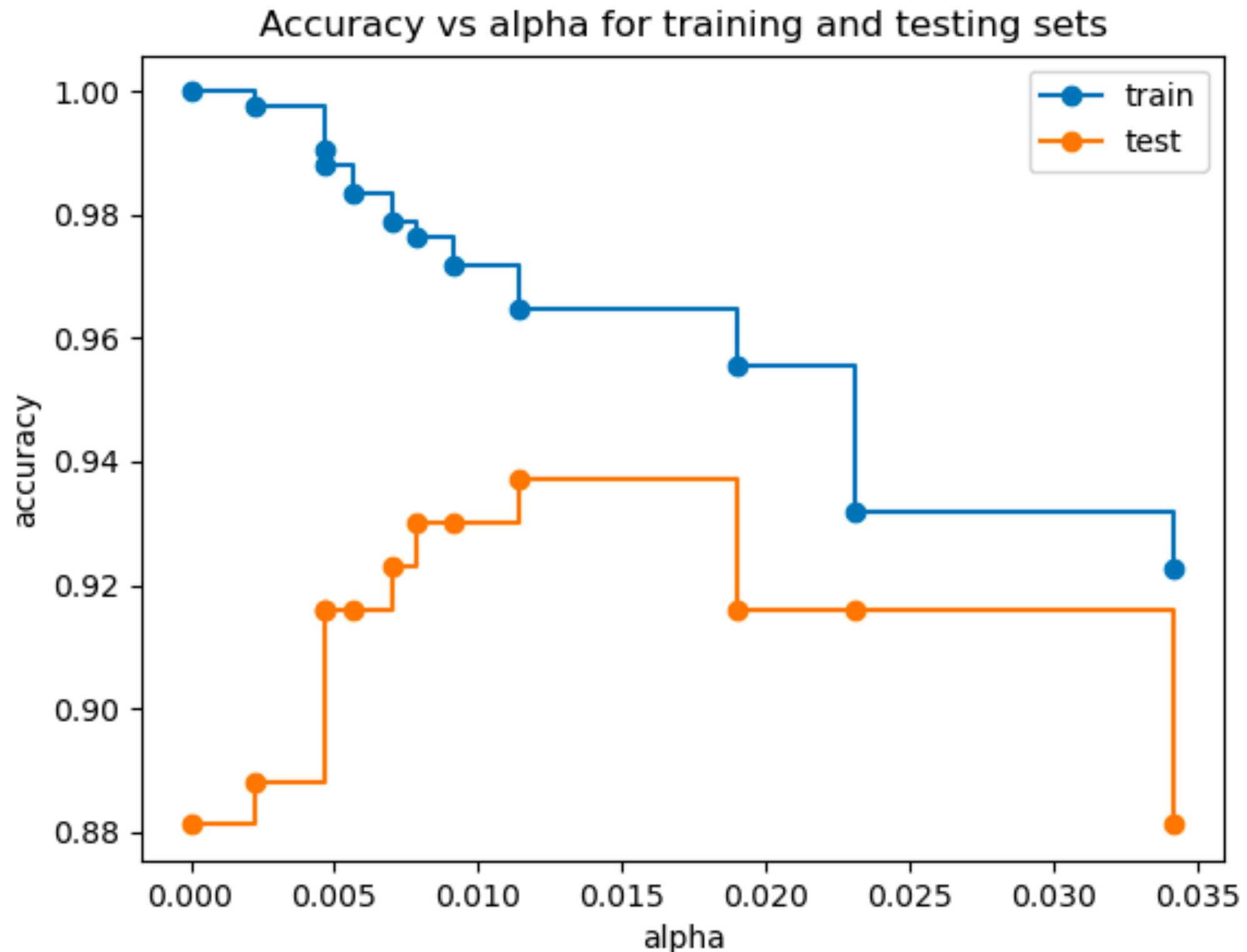
Вводим штраф на очень большие деревья

Можно эффективно перебирать все возможные деревья, получаемые прунингом из нашего, отрезая листья у узла с наименьшим весом

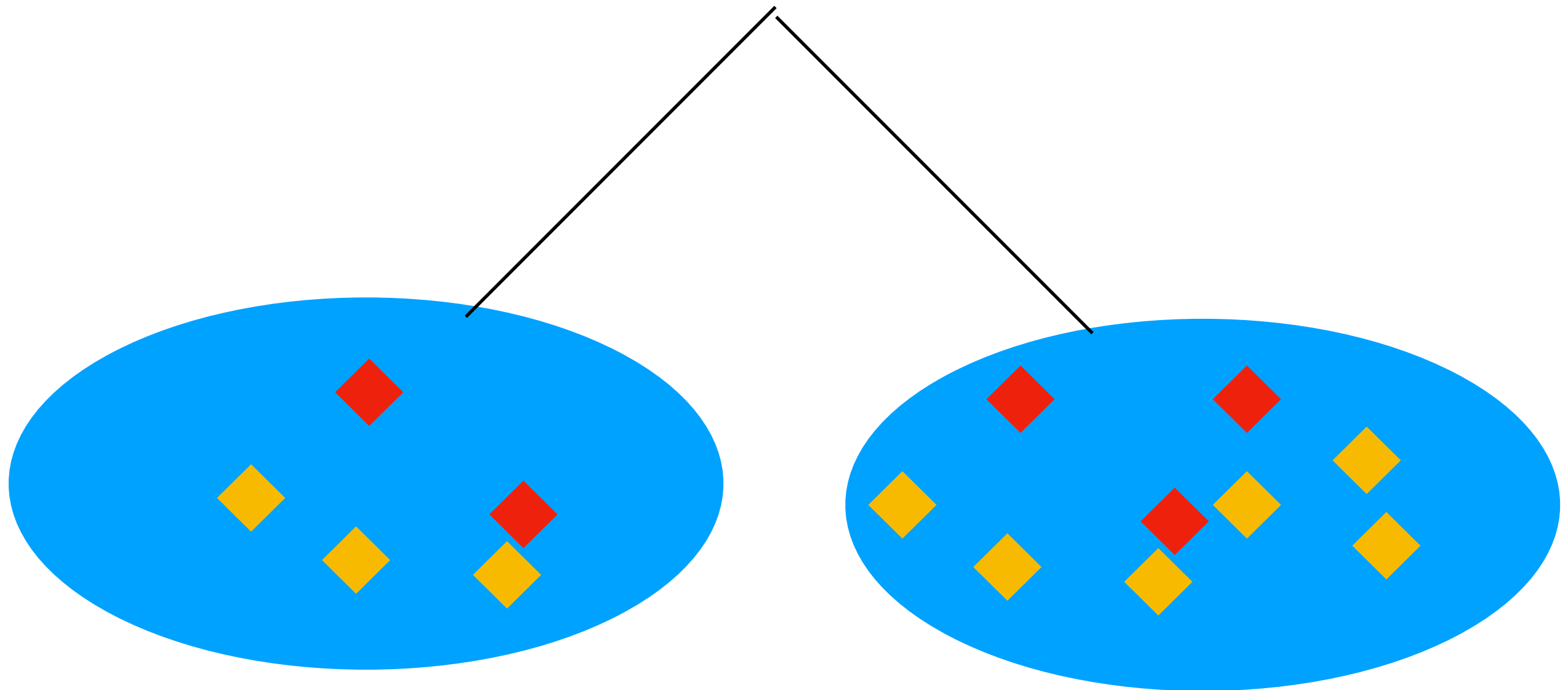
$$\alpha_{eff} = \frac{Quality(T) - Quality(T_t)}{T - 1}$$



# Где подбирать alpha?



# Почему бьем только на 2 узла?



# Решение не гладкое!

Можно ли с этим что-то сделать?

# Решение не гладкое!

Можно ли с этим что-то сделать?

Строго говоря - нет.

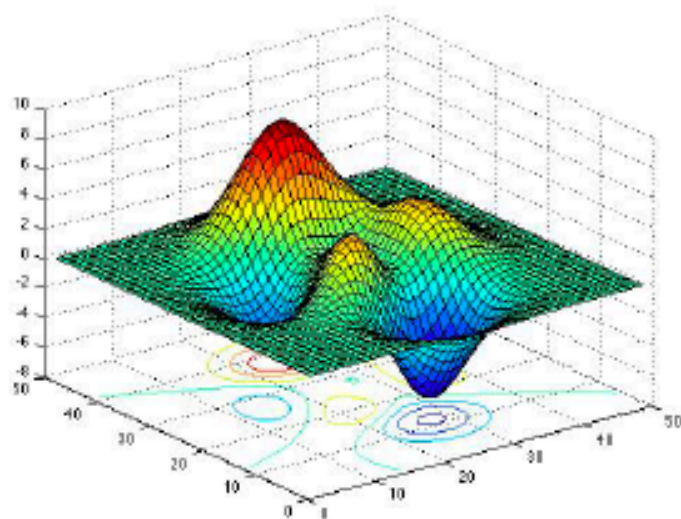
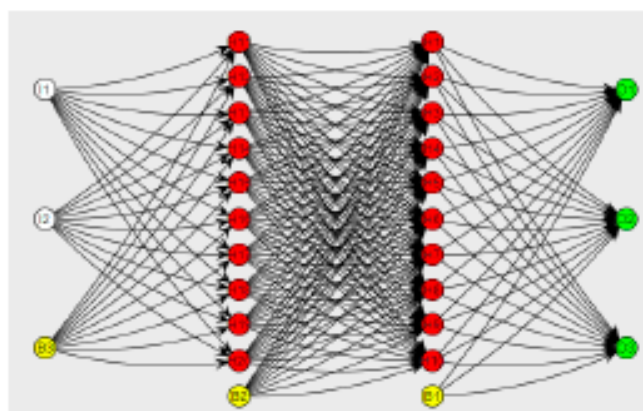
Но...



# Взгляд с точки зрения функционального анализа

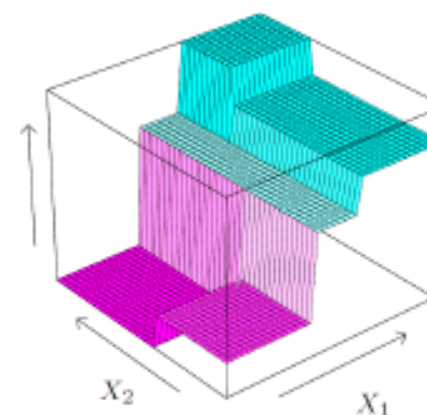
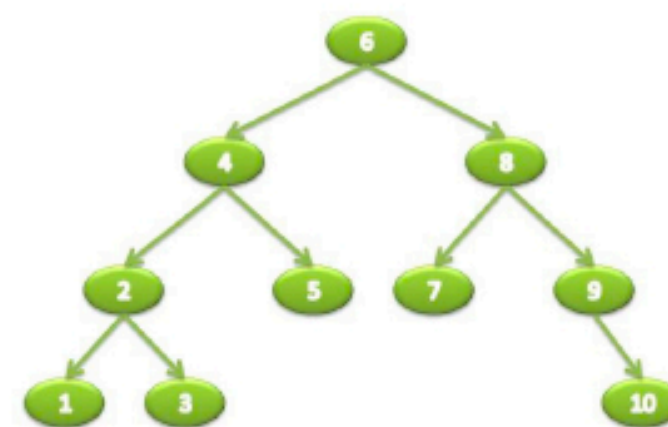
Гладкие функции

$$h(\mathbf{x}) = \sum \sigma(\dots \sum \sigma(\mathbf{w}^T \mathbf{x}))$$



Кусочно-постоянные функции

$$h(\mathbf{x}) = \sum_d c_d I\{\mathbf{x} \in R_d\}$$



Взято из презентации Гулин В.,  
Техносфера

# **Теорема об универсальном аппроксиматоре**

**С помощью нейронной сети с одним скрытым слоем можно  
аппроксимировать любую непрерывную функцию**

**С помощью дерева решений можно аппроксимировать любую  
кусочно-заданную функцию**

# Сплит по значению линейной комбинации признаков

Вместо такого:

$$X_j < thresh$$

Ищем такое:

$$\sum_j a_j X_j < thresh$$

# Сплит по значению линейной комбинации признаков

Вместо такого:

$$X_j < thresh$$

Ищем такое:

$$\sum_j a_j X_j < thresh$$

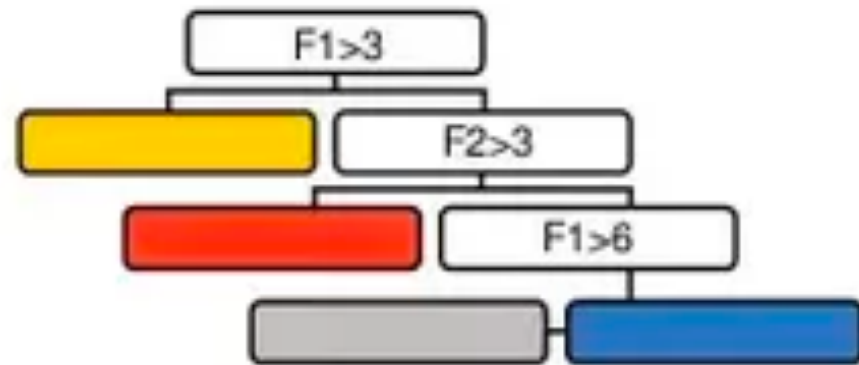
Долго считать(

# Модификации дерева решений

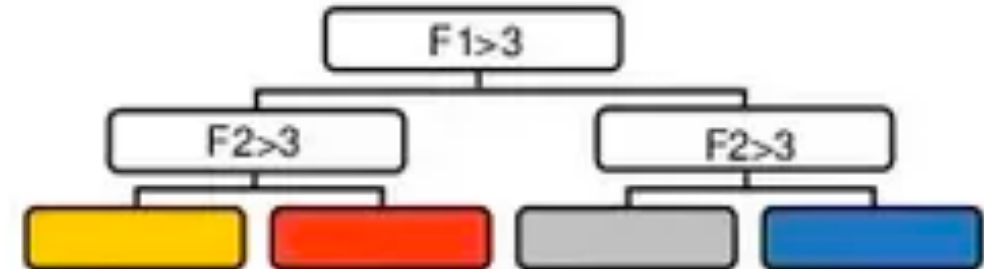
# Oblivious trees

## Regular vs oblivious trees

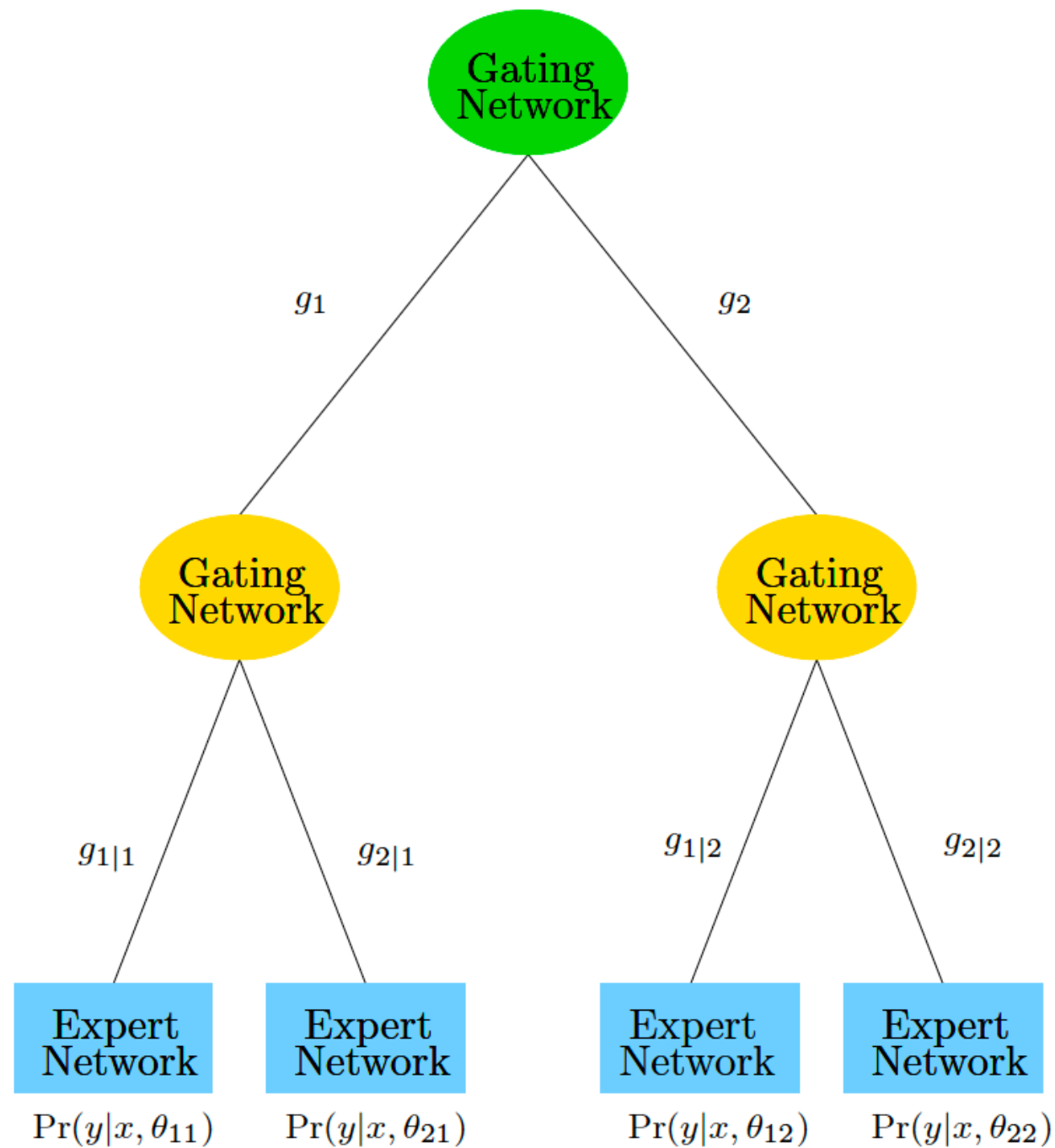
Decision Tree



Oblivious Trees

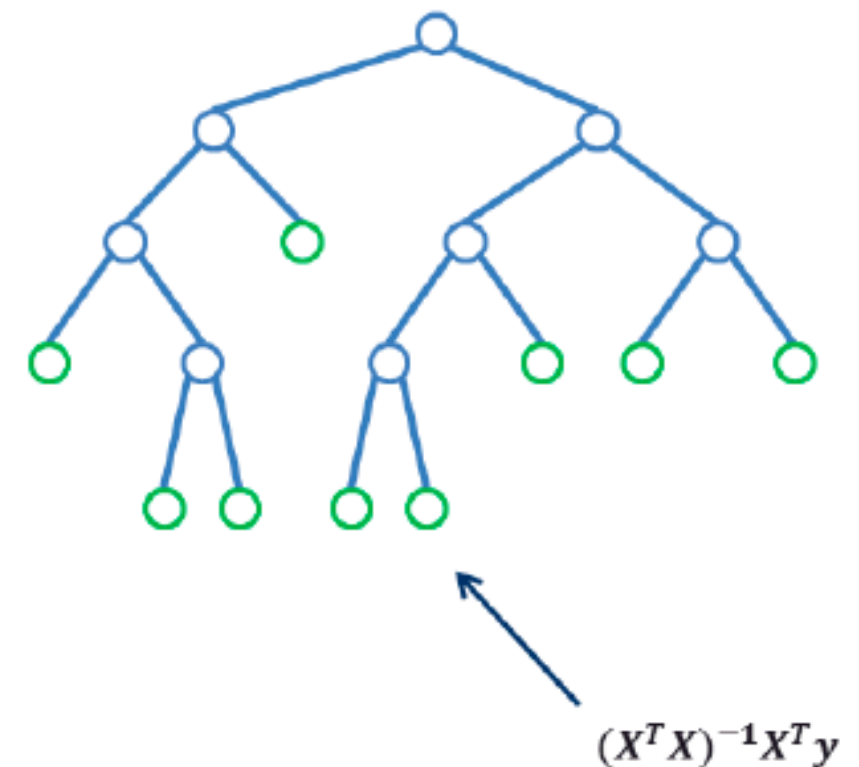
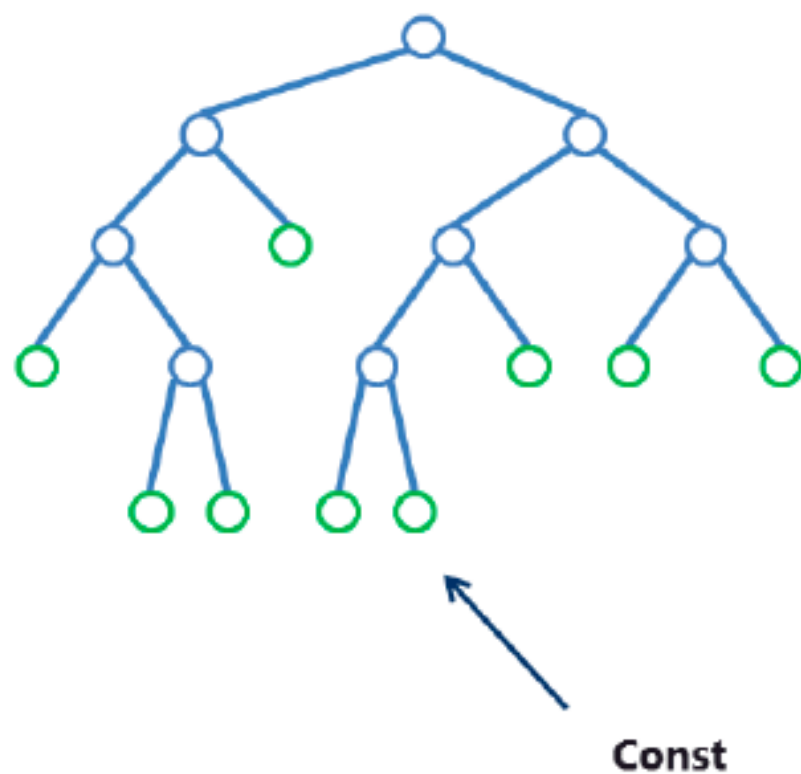


# Иерархические смесь экспертов



## Модельные деревья решений (Model decision trees)

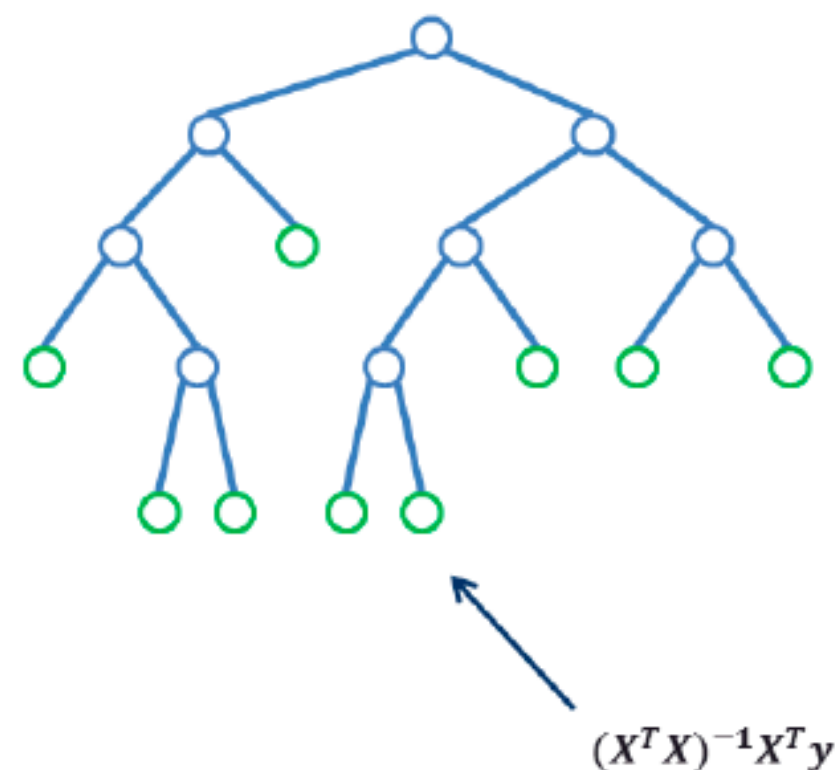
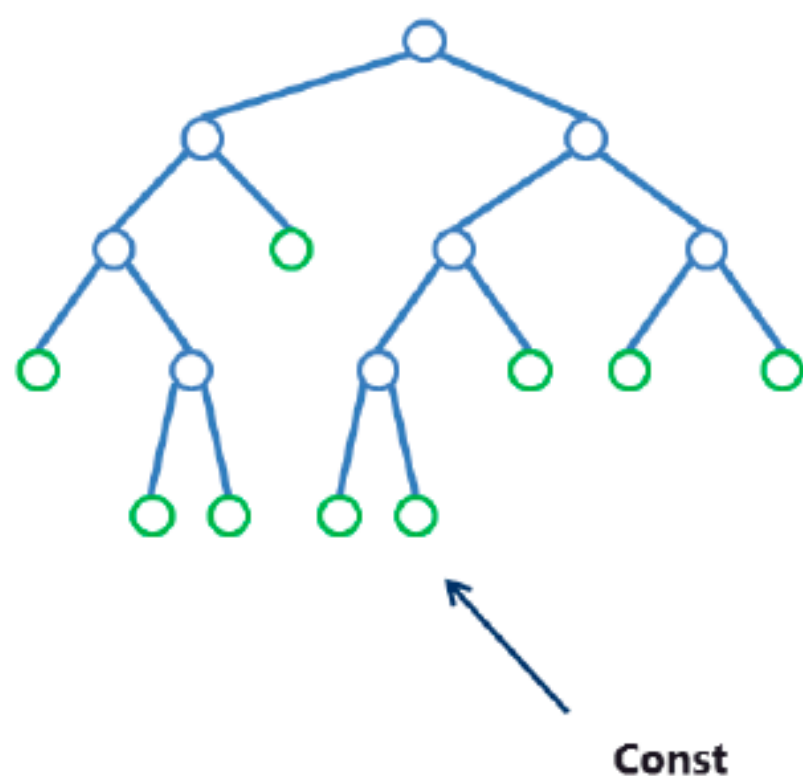
- Поместим в листья деревьев какие-нибудь алгоритмы вместо констант





## Модельные деревья решений (Model decision trees)

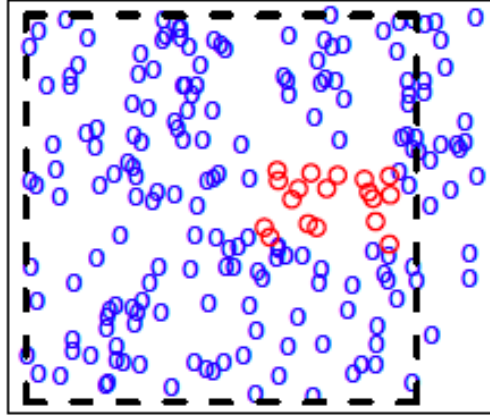
- Поместим в листья деревьев какие-нибудь алгоритмы вместо констант



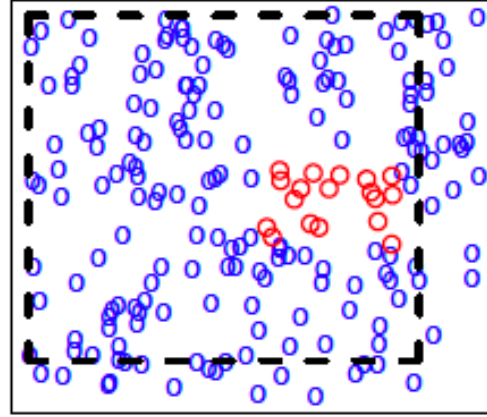
**Усложнение - пусть модели не только в листьях. Пусть каждый объект попадает в каждый лист с какой-то вероятностью. Тогда получаем иерархические экспертные модели**

# PRIM

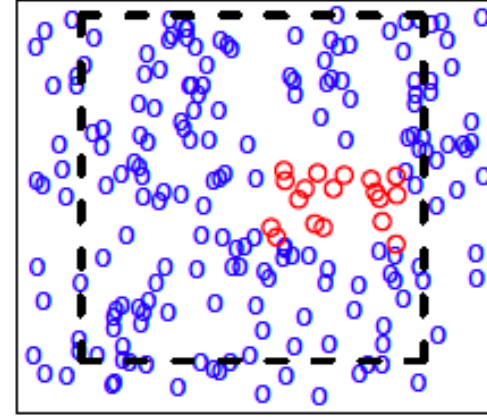
1



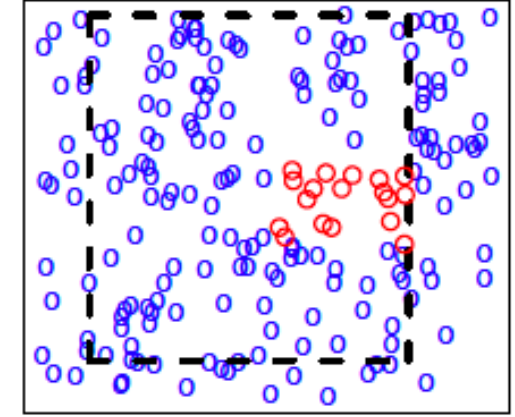
2



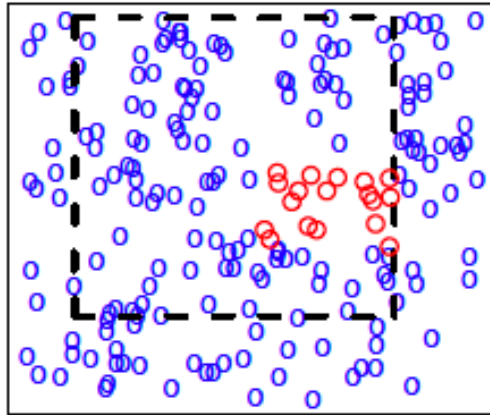
3



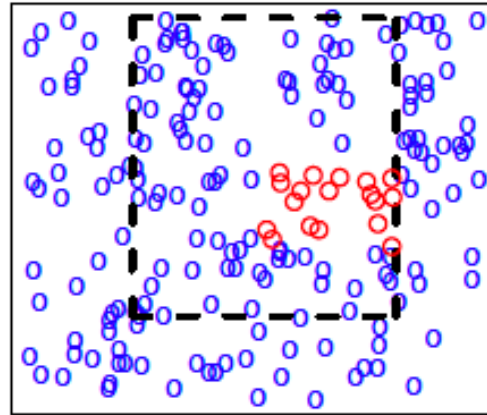
4



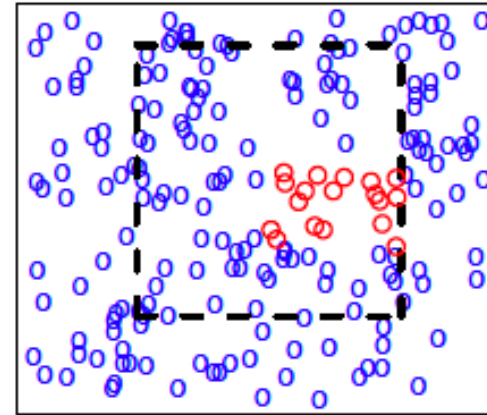
5



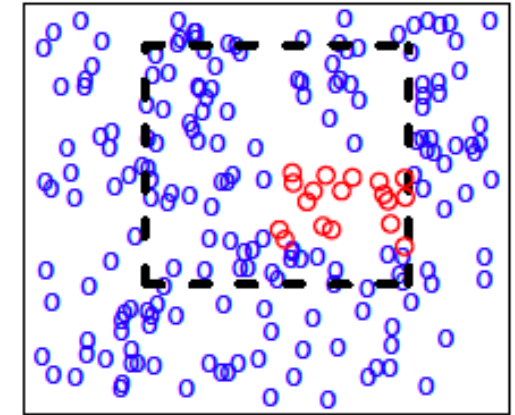
6



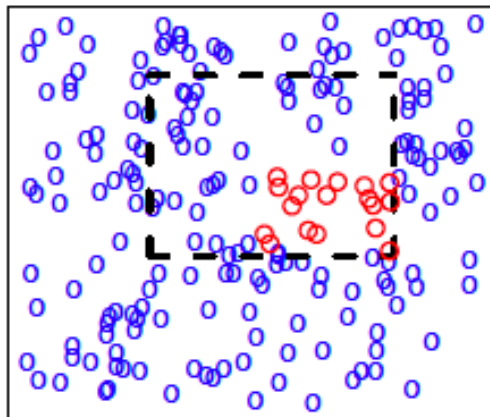
7



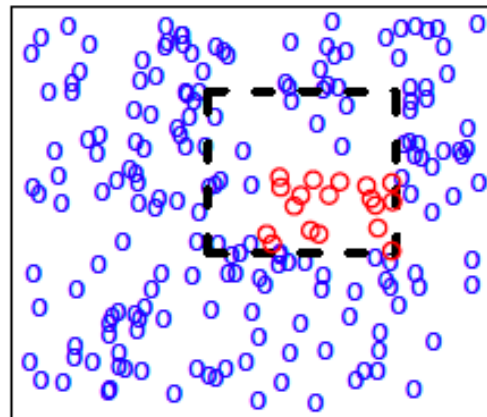
8



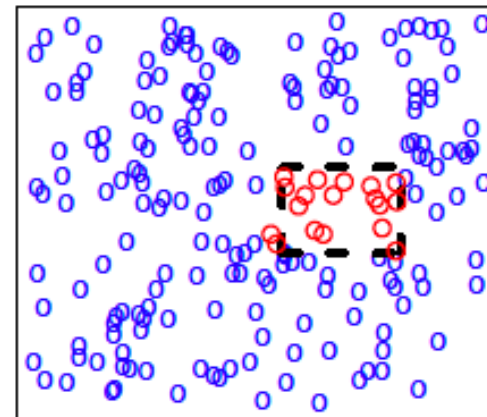
12



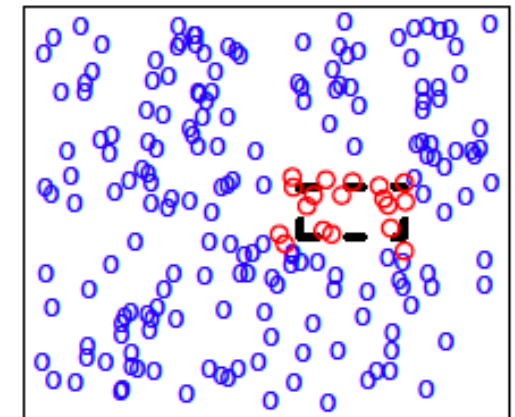
17



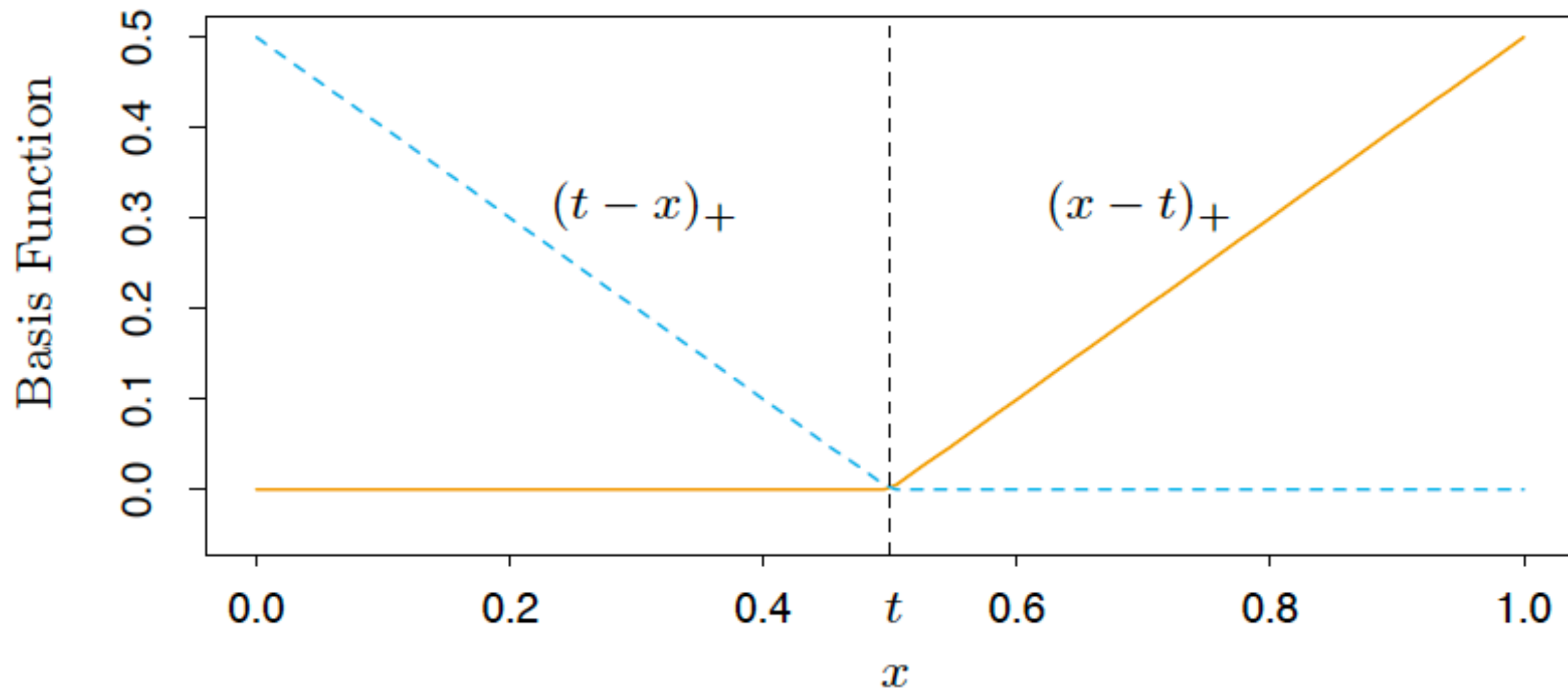
22



27



# MARS



# MARS

