

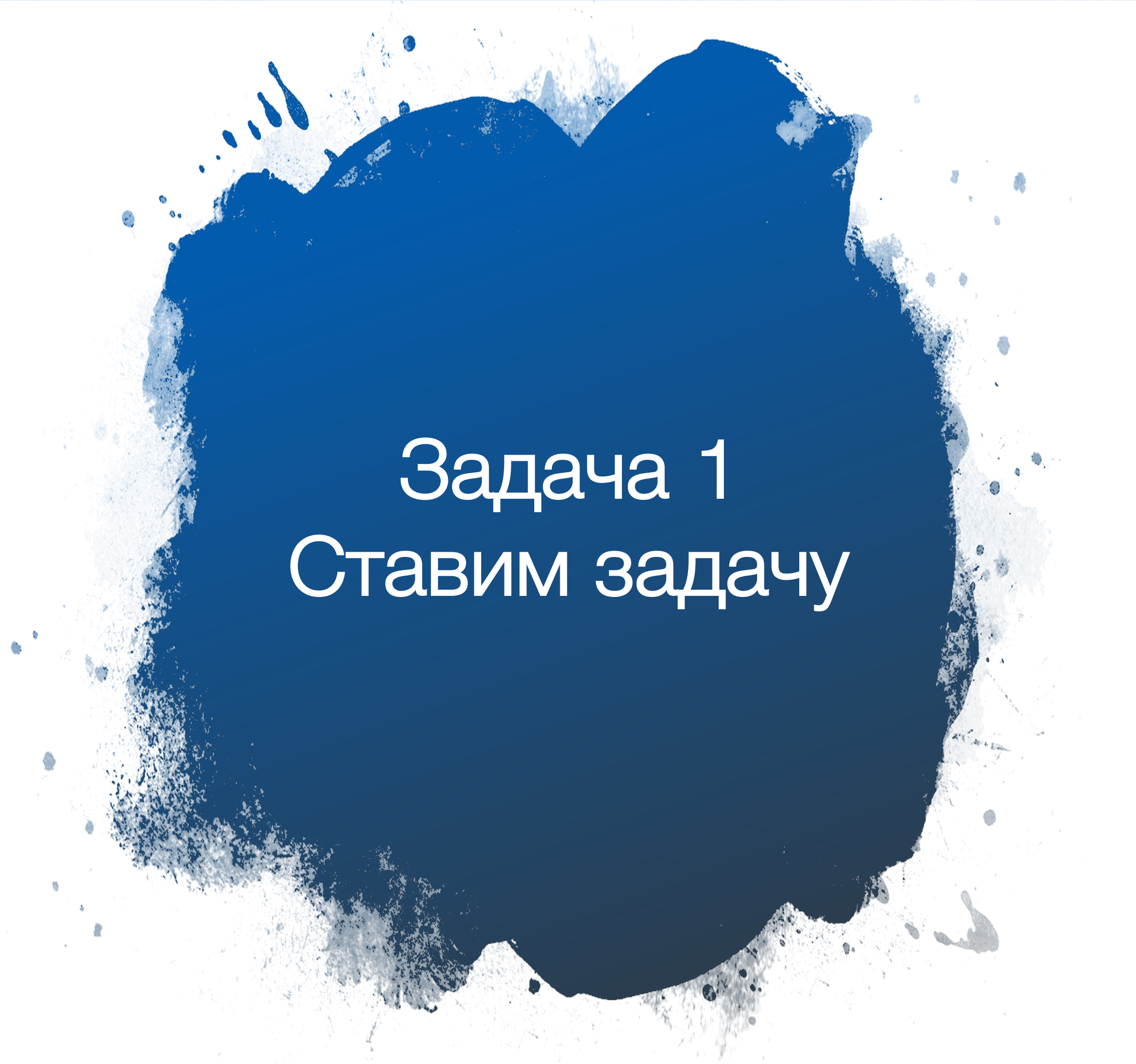


НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
УНИВЕРСИТЕТ

СЕМИНАР 5.

БММ191, 13.05.2020

Москва 2020



Задача 1

Ставим задачу

УСЛОВИЕ ЗАДАЧИ


Представьте себе, что у вас есть паблик с мемами. Вы — Хозяин мемов. Как и любой другой Хозяин мемов, вы любите лайки под мемами. Возникает желание привлечь в паблик целевую аудиторию, которая будет ставить под мемы лайки. Для этого вы хотите запустить рекламную кампанию паблика. Ясное дело, что рекламу хочется показывать не всем подряд, а только под-ходящим людям.

У вас есть данные по профилям всех тех людей, которые уже ставили в паблике лайки. По этим данным вам хочется построить модель, которая могла бы предсказать подходит ли конкретный человек для вашей рекламной компании (поставил бы ли он в паблик лайк, если бы был на него подписан).

а) Сформулируйте задачу машинного обучения. Какой должна быть целевая переменная, чтобы перед вами была задача классификации. Какой должна быть целевая переменная, чтобы это была задача регрессии?

б) Какие факторы из профилей вы бы использовали, чтобы спрогнозировать подходит ли человек для рекламной компании?

в) Приведите ещё парочку примеров задачи классификации и задачи регрессии.



Задача 2

Качество прогноза

УСЛОВИЕ ЗАДАЧИ

Добрыня, Алёша и Илья смотрят мемы и ставят на них лайки. Мы пытаемся предсказать сколько лайков они оставят под мемами на основе поведения их однокурсников.

Для этого мы оценили регрессию. Ну и она нам напредсказывала, что парни поставят 4, 20 и 110 лайков. В реальности они поставили 5, 10 и 100 лайков. Возникает вопрос: насколько сильно наша модель ошиблась в прогнозировании.

Что такое MAE, MSE, RMSE и MAPE? Посчитайте для модели все четыре метрики качества.

РЕШЕНИЕ

$$MAE = \frac{1}{\ell} \sum_{i=1}^{\ell} |y_i - \hat{y}_i|$$

$$MAPE = \frac{1}{\ell} \sum_{i=1}^{\ell} \frac{|y_i - \hat{y}_i|}{y_i}$$

$$MSE = \frac{1}{\ell} \sum_{i=1}^{\ell} (y_i - \hat{y}_i)^2$$

$$RMSE = \sqrt{\frac{1}{\ell} \sum_{i=1}^{\ell} (y_i - \hat{y}_i)^2}$$

РЕШЕНИЕ

$$MAE = 7$$

$$MSE = 67$$

$$RMSE = \sqrt{67} = 8.19$$

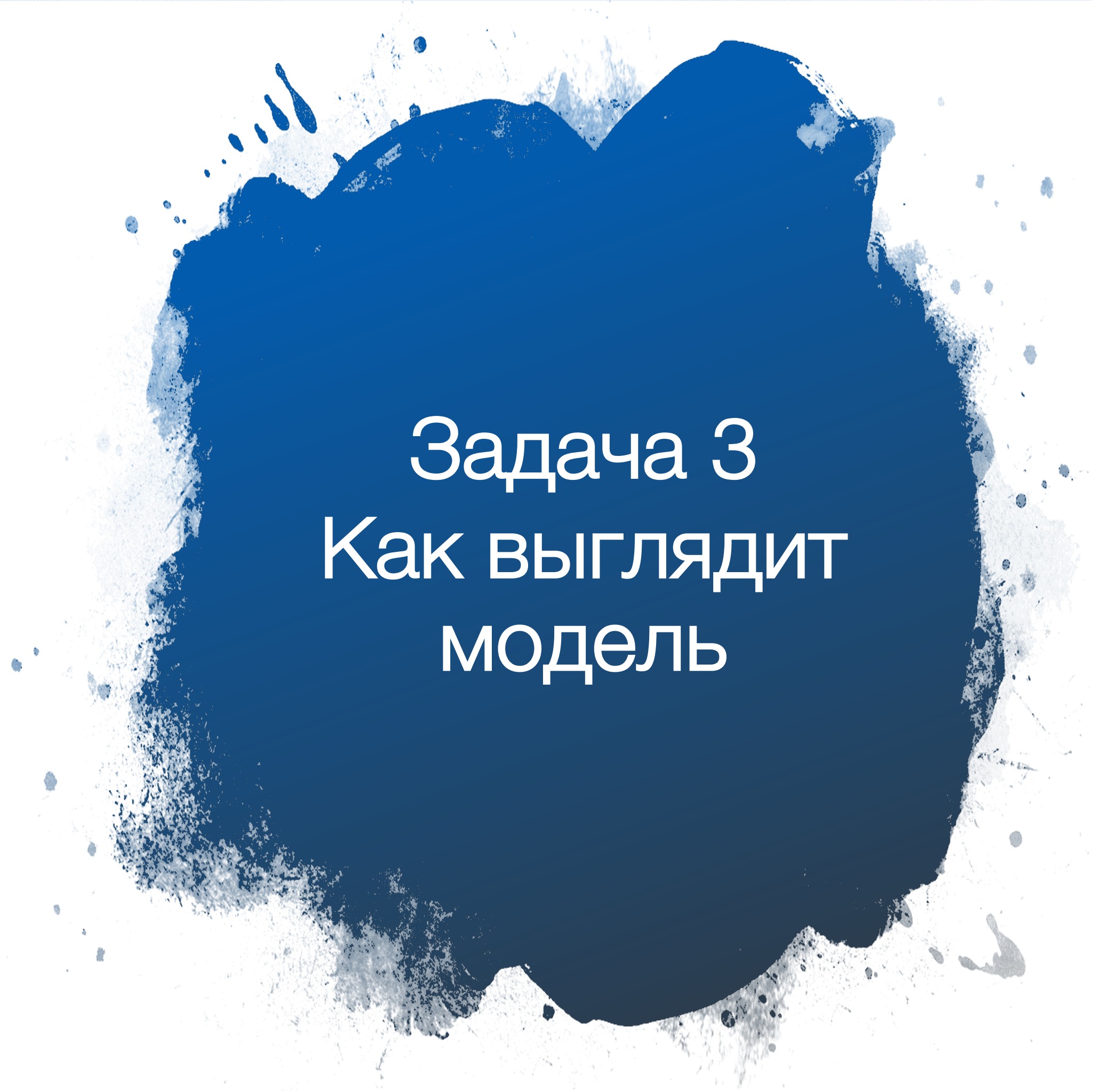
$$MAPE = 0.28 \text{ или } 28\%$$

$$R^2 = 1 - \frac{\sum_{i=1}^{\ell} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{\ell} (y_i - \bar{y})^2}$$

$$\bar{y} = 44.67$$

$$R^2 = 1 - \frac{201}{6530.7} = 0.97$$

y	\hat{y}_i	$y - \hat{y}_i$	$ y - \hat{y}_i $	$(y - \hat{y}_i)^2$	$\frac{ y_i - \hat{y}_i }{y_i}$	$(y_i - \bar{y})^2$
4	5	$= 4 - 5 = -1$	1	1	$= \frac{1}{4} = 0.25$	$(-40.67)^2$
20	10	$= 20 - 10 = 10$	10	100	$= 10/20 = 0.5$	-24.67
110	100	$= 110 - 100 = 10$	10	100	$= 10 / 110 = 0.09$	65.33



Задача 3

Как выглядит модель

УСЛОВИЕ ЗАДАЧИ

Предположим, Олег хочет купить автомобиль и считает, сколько денег ему нужно для этого накопить¹. Он пересмотрел десяток объявлений в интернете и увидел, что новые автомобили стоят около 20000, годовалые — примерно 19000, двухлетние — 18000 и так далее. В уме Олег-аналитик выводит формулу: адекватная цена автомобиля начинается от 20000 и падает на 1000 каждый год, пока не упрётся в 10000. Олег сделал то, что в машинном обучении называют регрессией — предсказал цену по известным данным. Давайте попробуем повторить подвиг Олега.

а) Как выглядит формула в случае Олега?

б) За сколько продать старый айфон? Придумайте формулу для предсказания. Проинтерпретируйте каждый коэффициент в ней.

в) Сколько одежды брать с собой в путешествие? Придумайте формулу для предсказания. Проинтерпретируйте каждый коэффициент в ней.

г) Сколько шашлыка брать на дачу? Как выглядит формула?

д) Сколько брать шашлыка, если есть друг-вегетарианец? Как можно назвать этого друга в терминах машинного обучения? Испортит ли вегетарианец формулу?

Было бы удобно иметь формулу под каждую проблему на свете. Но взять те же цены на автомобили: кроме пробега есть десятки комплектаций, разное техническое состояние, сезонность спроса и ещё столько неочевидных факторов, которые Олег, даже при всём желании, не учёл бы в голове.

УСЛОВИЕ ЗАДАЧИ

Предположим, Олег хочет купить автомобиль и считает, сколько денег ему нужно для этого накопить¹. Он пересмотрел десяток объявлений в интернете и увидел, что новые автомобили стоят около 20000, годовалые — примерно 19000, двухлетние — 18000 и так далее. В уме Олег-аналитик выводит формулу: адекватная цена автомобиля начинается от 20000 и падает на 1000 каждый год, пока не упрётся в 10000. Олег сделал то, что в машинном обучении называют регрессией — предсказал цену по известным данным. Давайте попробуем повторить подвиг Олега.

а) Как выглядит формула в случае Олега?

$$20000 - 1000n$$

$$y = 20000 - 1000 x_1$$

$$x_1 \leq 10$$

УСЛОВИЕ ЗАДАЧИ

б) За сколько продать старый айфон? Придумайте формулу для предсказания. Проинтерпретируйте каждый коэффициент в ней.


y – цена айфона (для которого мы прогнозируем)

x_1 – количество выпущенных после моделей

x_2 – возраст модели

x_3 – битый или не битый – дамми-переменные

$$y = 70\,000 - 20\,000 \cdot x_1 - 15\,000 \cdot x_2 - 24\,000 \cdot x_3$$

A large, dark blue ink splatter or blotch is centered on a white background. The splatter has irregular, feathered edges with some smaller droplets and speckles trailing off to the left and right. The text is centered within the main body of the splatter.

Задача 4

Как обучаются модели

УСЛОВИЕ ЗАДАЧИ

Давайте попробуем совсем-совсем на пальцах почувствовать, как модели обучаются. Пусть у Хозяина мемов есть две переменные: x — возраст подписчика и y — число лайков, которое он оставил. Хозяин мемов хочет оценить регрессию $y = \beta \cdot x$, то есть он хочет попытаться предсказать число лайков по возрасту подписчика. Хозяин собрал два наблюдения для оценивания модели: $x_1 = 15$, $y_1 = 10$ и $x_2 = 22$, $y_2 = 2$.

Теперь хозяину надо подобрать коэффициент β так, чтобы ошибка прогноза, измеряемая с помощью MSE оказалась поменьше.

1. Пусть $\beta = 1$. Какие значения нам спрогнозирует модель? Какая у нее будет ошибка?
2. Пусть $\beta = 0.5$. Найдите прогнозы и ошибку модели.
3. Какое значение для β нам больше подходит? Как можно найти оптимальное β ?

РЕШЕНИЕ

$$y = \beta \cdot x,$$


$$x_1 = 15, y_1 = 10 \text{ и } x_2 = 22, y_2 = 2.$$

1. Пусть $\beta = 1$. Какие значения нам спрогнозирует модель? Какая у нее будет ошибка?

y	x	$\hat{y}_i = 1 \cdot x$	
10	15	15	$(10-15)^2 = 25$
2	22	22	$(2-22)^2 = 400$
			$MSE = 425 / 2 = 212.5$

2. Пусть $\beta = 0.5$. Найдите прогнозы и ошибку модели.

y	x	$\hat{y}_i = 0.5 \cdot x$	
10	15	7.5	$(10-7.5)^2 = 6.25$
2	22	11	$(2-11)^2 = 81$
			$MSE = 87.25 / 2 = 43.625$



Задача 5

Регрессионное дерево

УСЛОВИЕ ЗАДАЧИ

Для того чтобы решать задачу регрессии и прогнозировать что-нибудь, можно пытаться искать коэффициенты в уравнениях, которые мы выписывали выше. Это один из вариантов модели. Он называется линейной регрессией. Линейной, потому что мы пытаемся провести через облако точек линию. Можно пробовать оценивать и какие-то другие, более сложные, нелинейные модели. Например, можно построить регрессионное дерево. Было бы нечестно бросать вас не обучив ручками ни одной модели. Давайте обучим!

Миша работает в маленькой кофейне. Харио Малабар Монсун — фирменный напиток этой кофейни. Мише интересно узнать, как именно ведёт себя количество заказов напитка y_i в зависимости от температуры за окном t_i . Четыре дня Миша записывал свои наблюдения:

t	y
21	1
19	2
12	8
8	8

Сегодня он решил обучить регрессионное дерево. В качестве функции потерь он использует

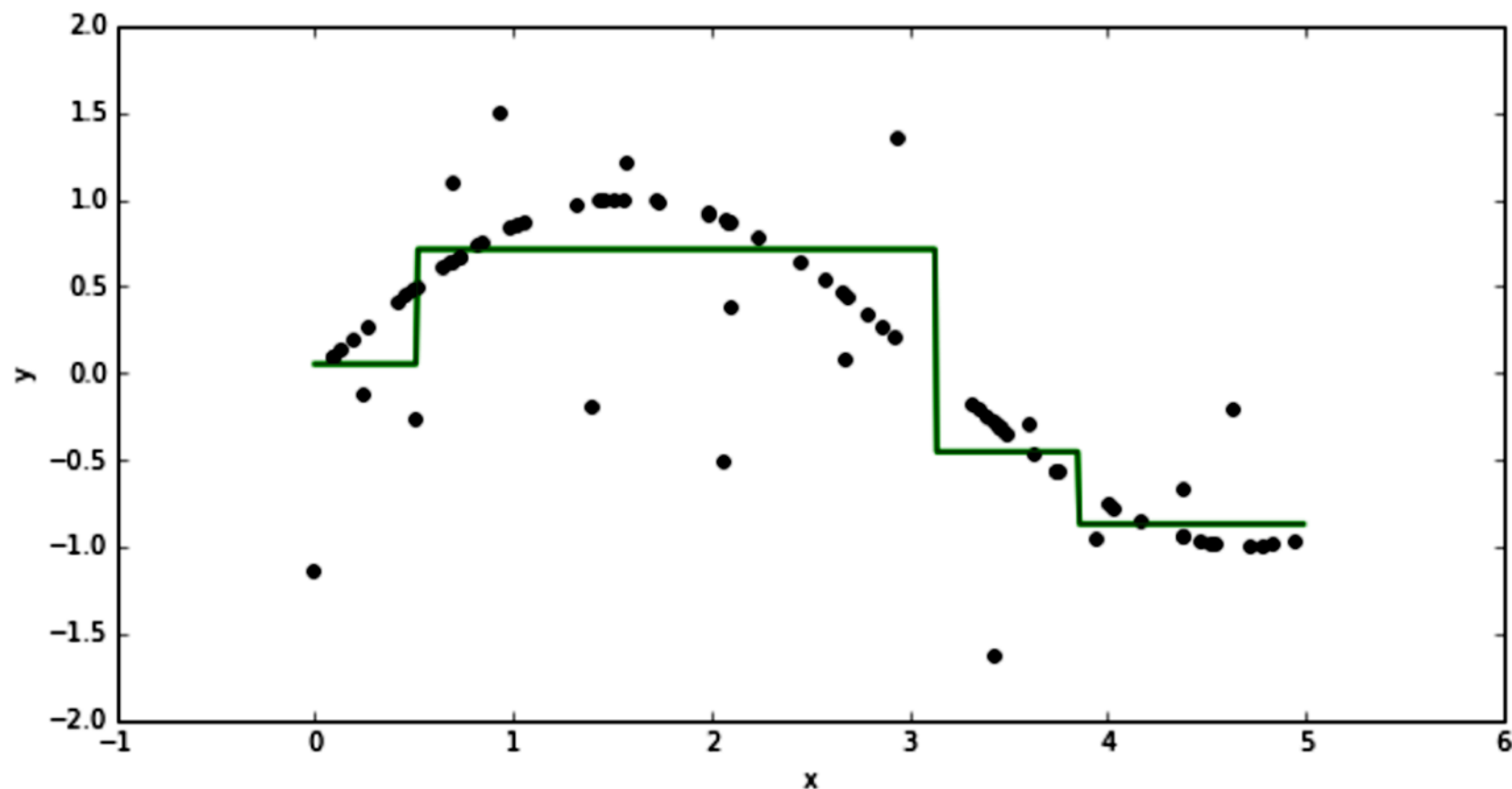
$$MSE = \frac{1}{\ell} \sum_{i=1}^{\ell} (y_i - \hat{y}_i)^2$$

- а) Обучите регрессионное дерево.
- б) Какой прогноз на сегодня сделает дерево Миши, если за окном 13 градусов?
- в) Можно ли для обучения дерева использовать MAE?

ЧТО ТАКОЕ РЕГРЕССИОННОЕ ДЕРЕВО

Модель регрессии. Такая же, как и линейная регрессия.

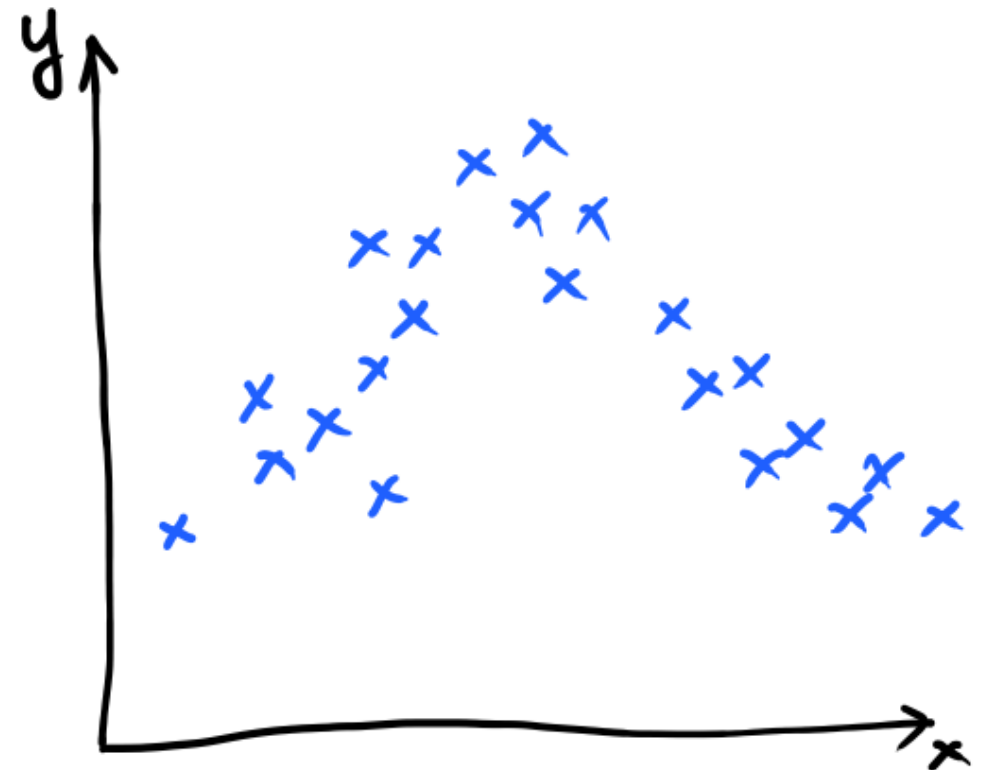
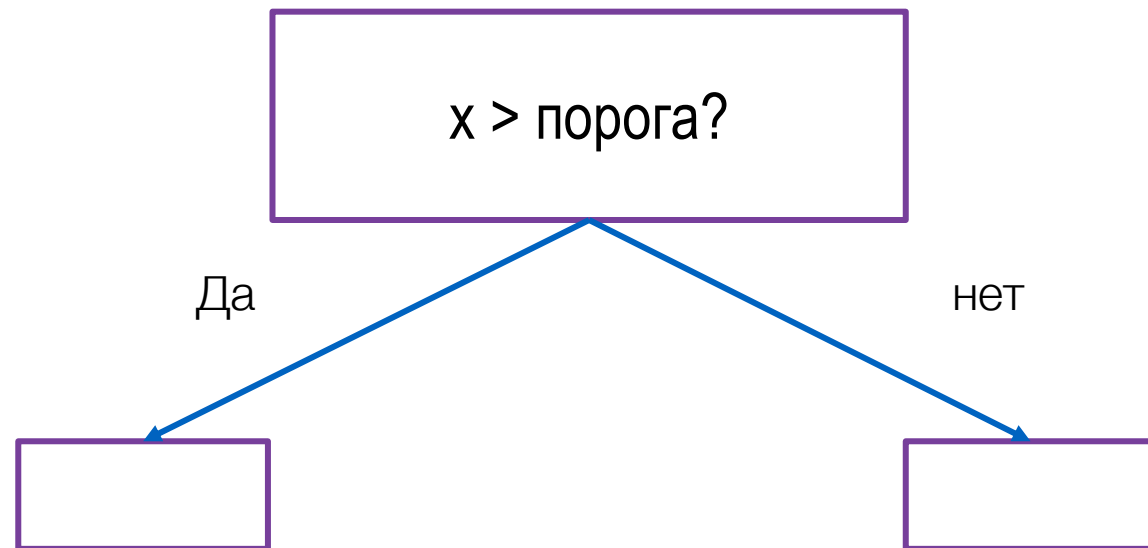
Восстанавливают не линию, а кусочно-непрерывную функцию



КАК СТРОИТСЯ ДЕРЕВО

1) Задаем вопросы по фиче:

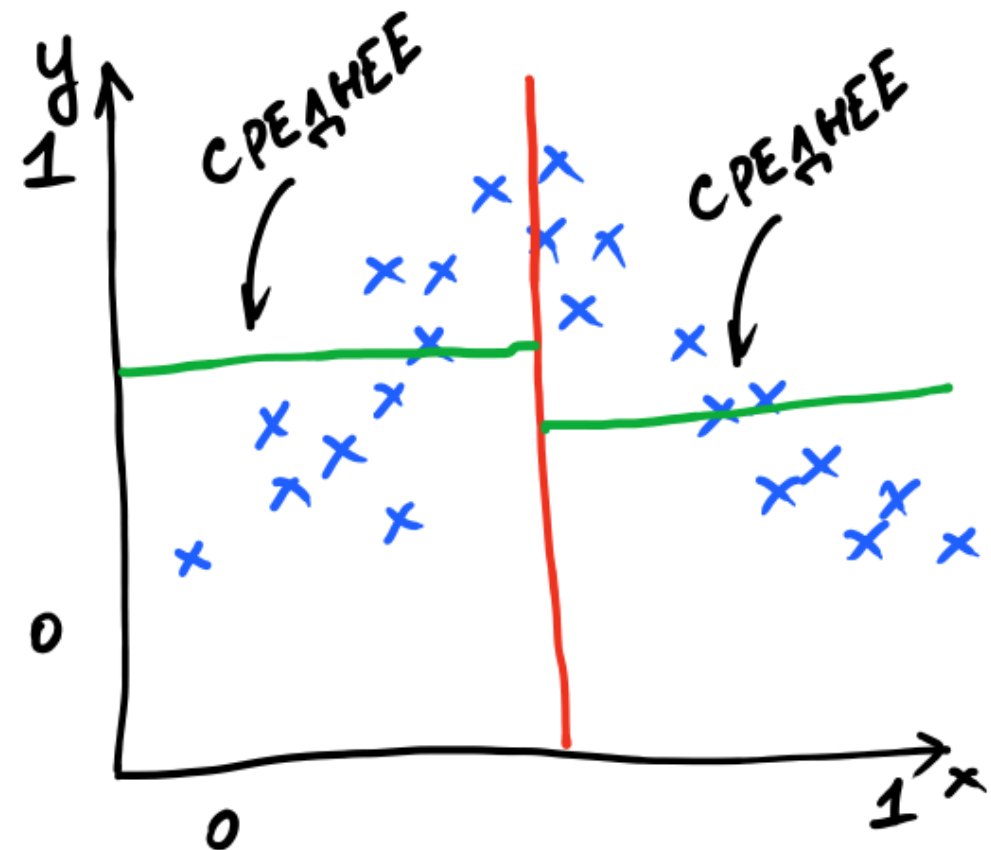
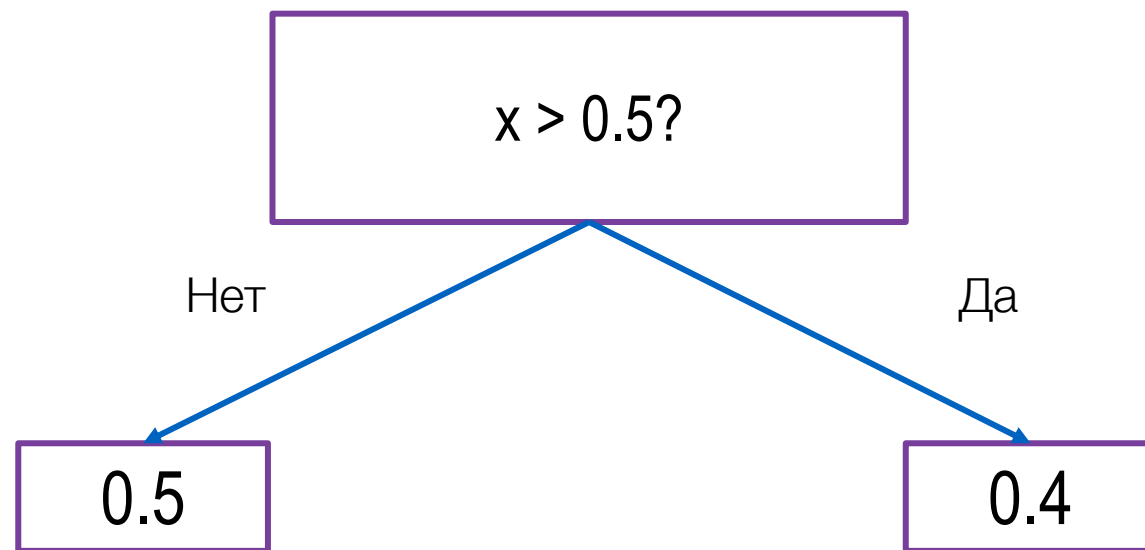
правда ли, что признак $>$ порога?



КАК СТРОИТСЯ ДЕРЕВО

1) Задаем вопросы по фиче:
правда ли, что признак $>$ порога?

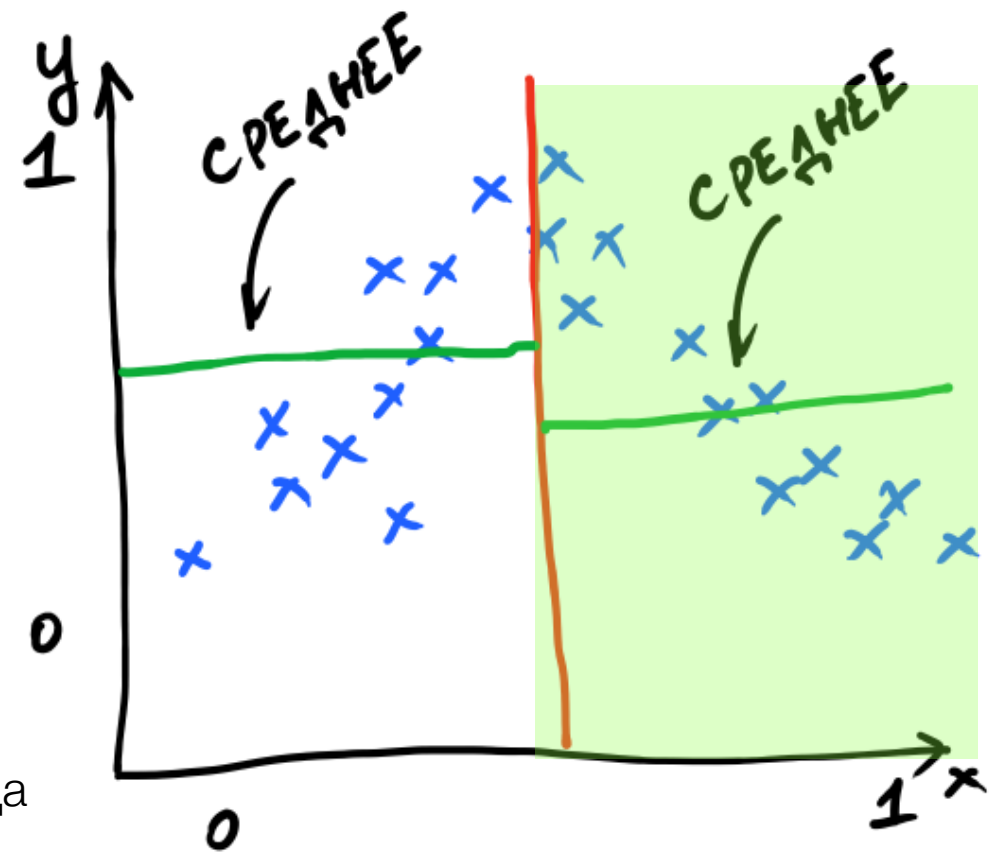
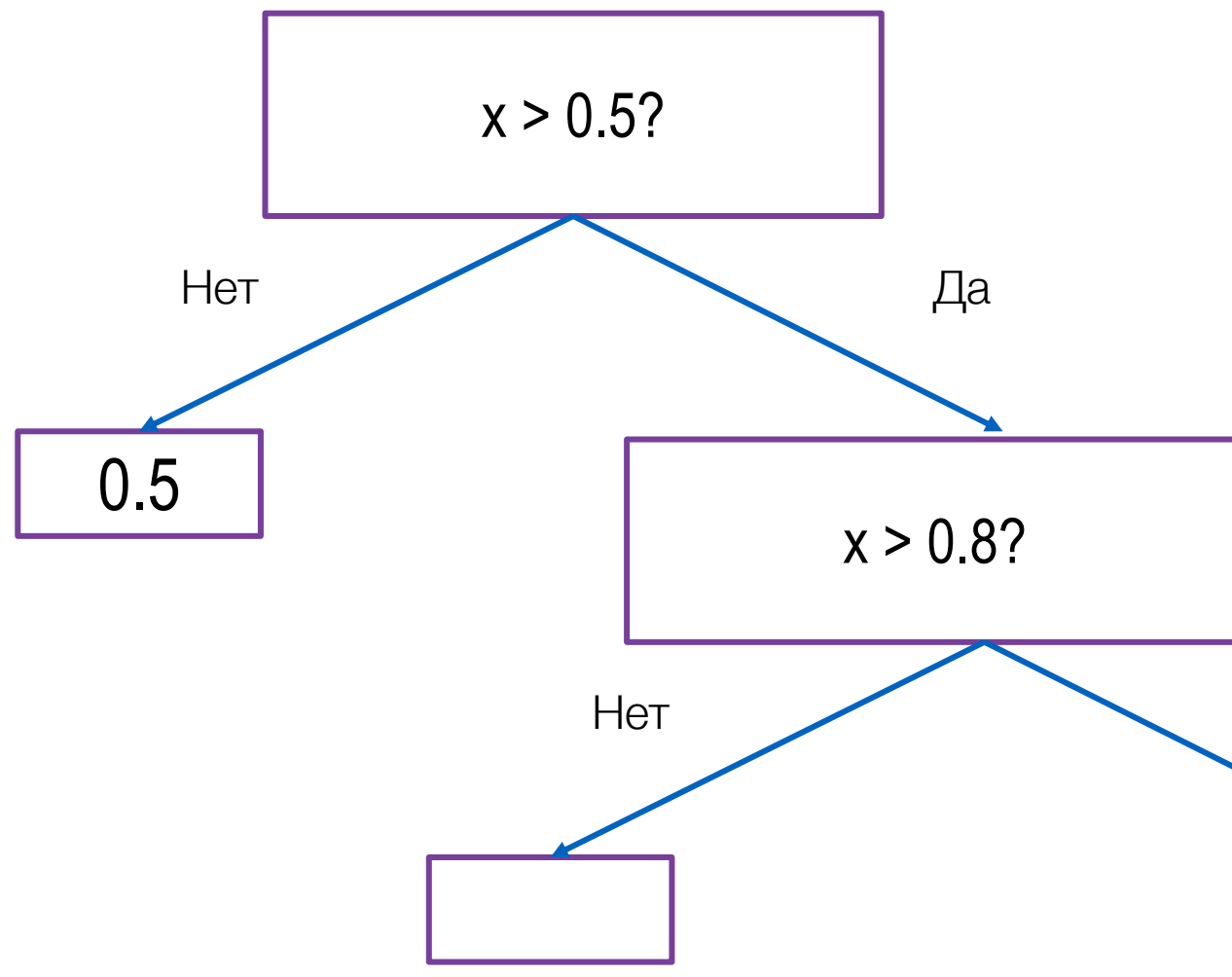
2) Делим данные на две части.



Прогноз в каждой ветке – это среднее значение всех элементов, попавших в эту ветку

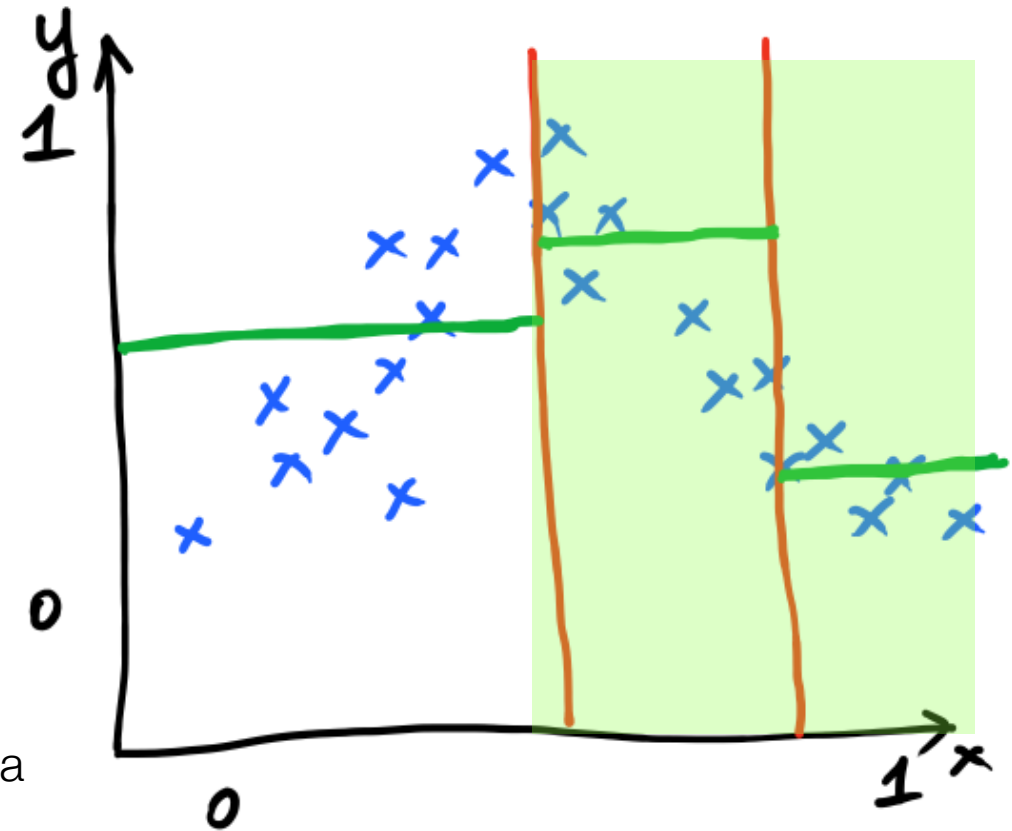
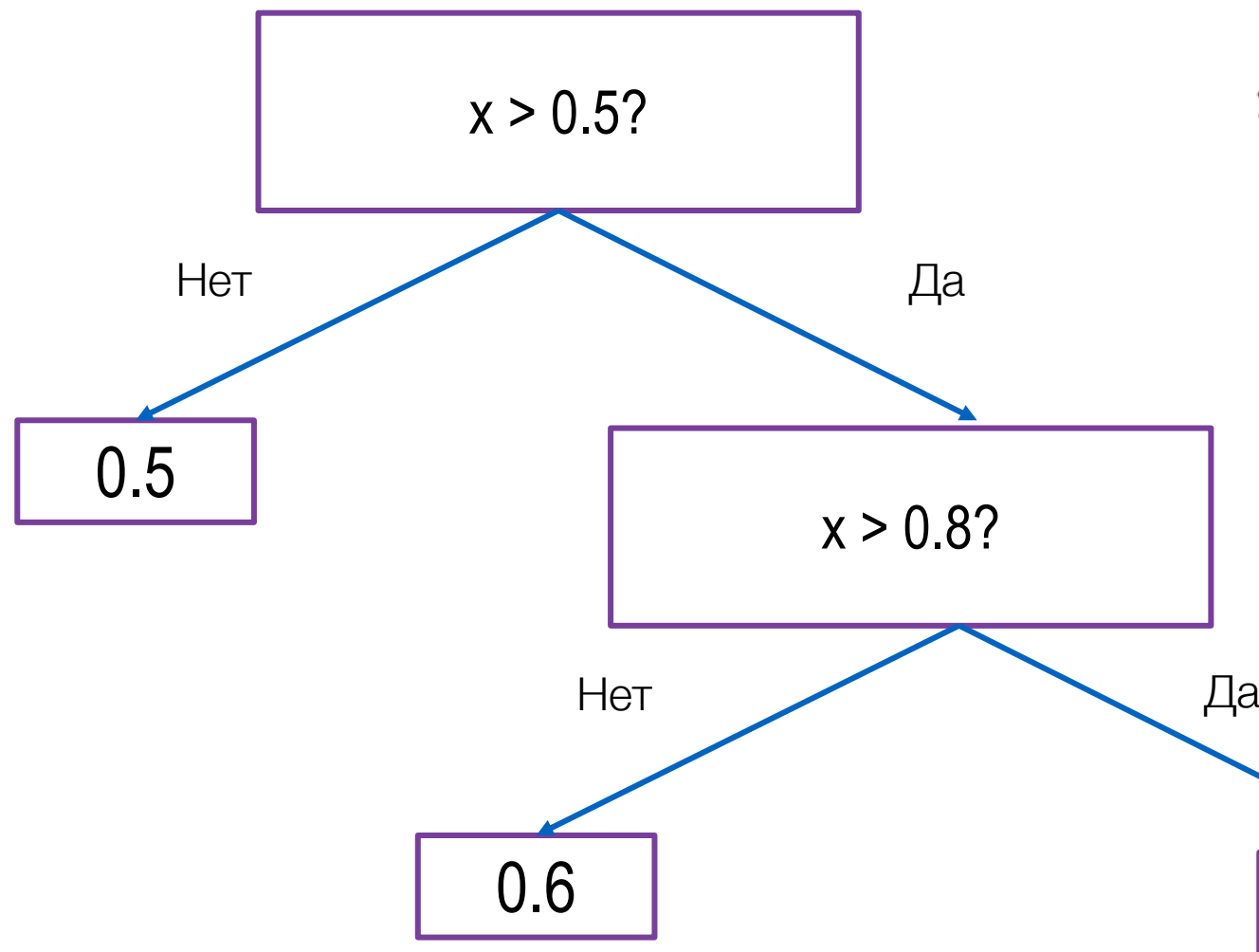
КАК СТРОИТСЯ ДЕРЕВО

3) Улучшаем прогнозы, «углубляем» дерево дальше, задавая вопросы

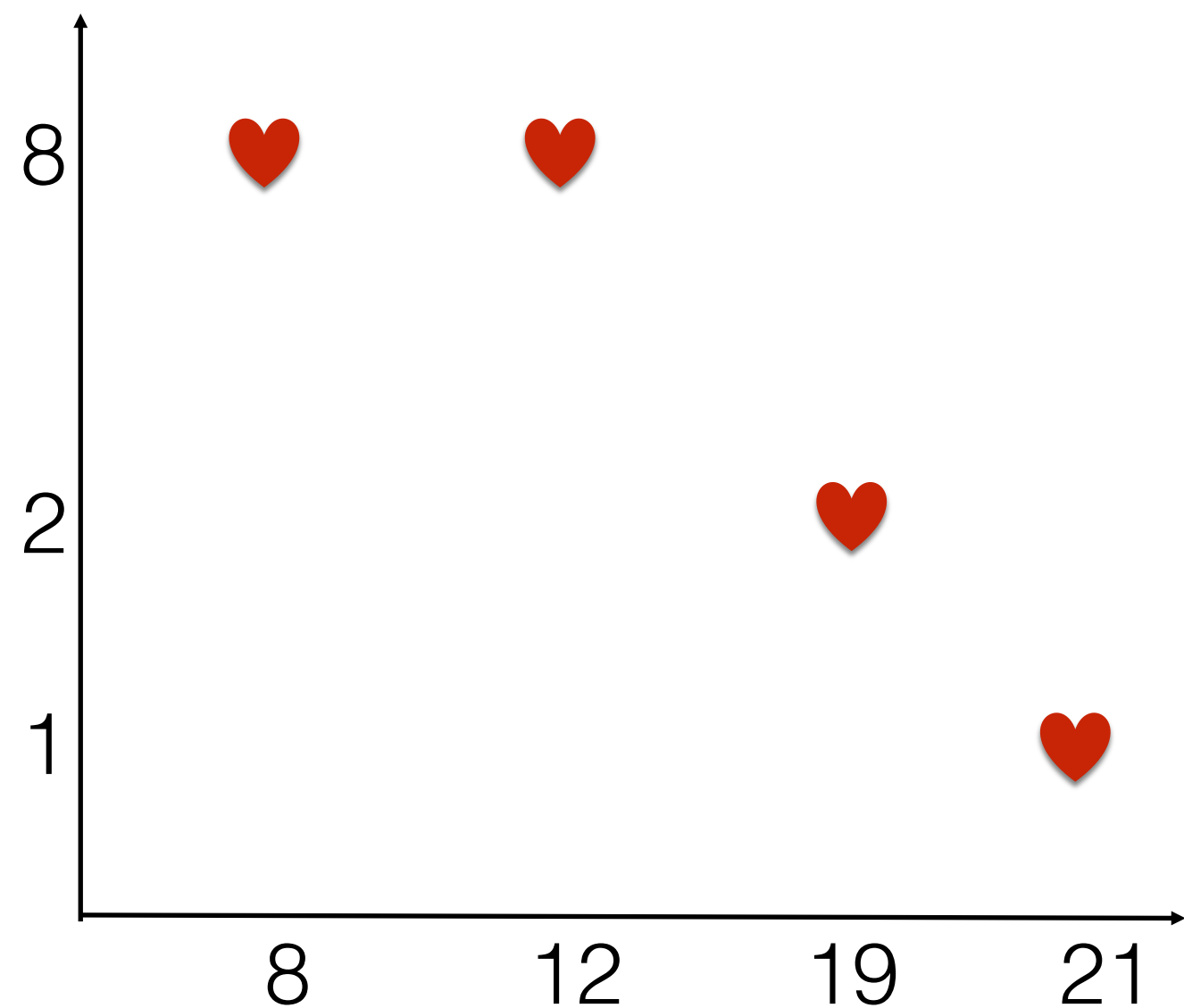


КАК СТРОИТСЯ ДЕРЕВО

- 3) »Углубляем» дерево дальше, задавая вопросы
- 4) И до победного, пока не уткнемся в какой-либо критерий останова

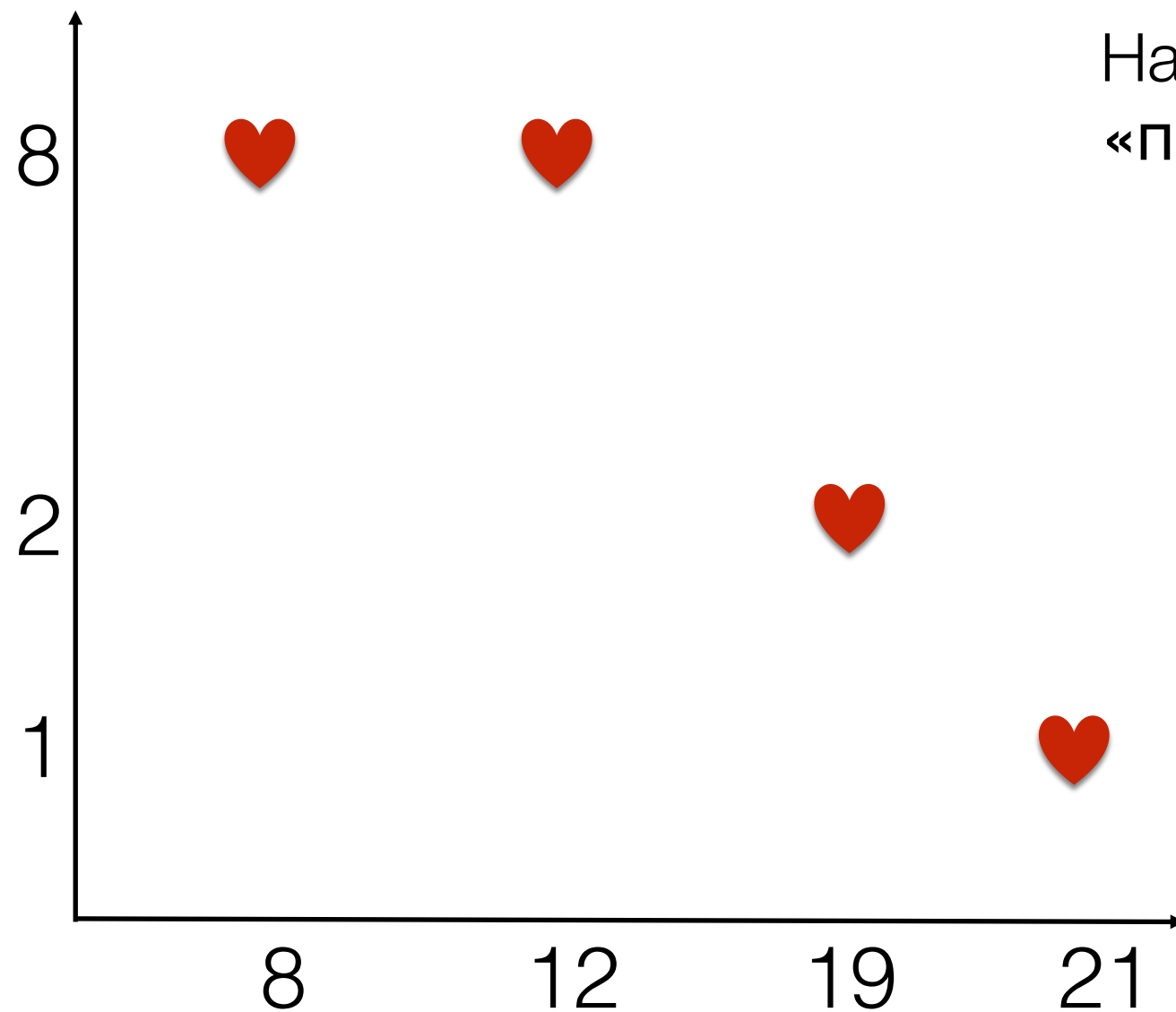


НАША ЗАДАЧКА



t	y
21	1
19	2
12	8
8	8

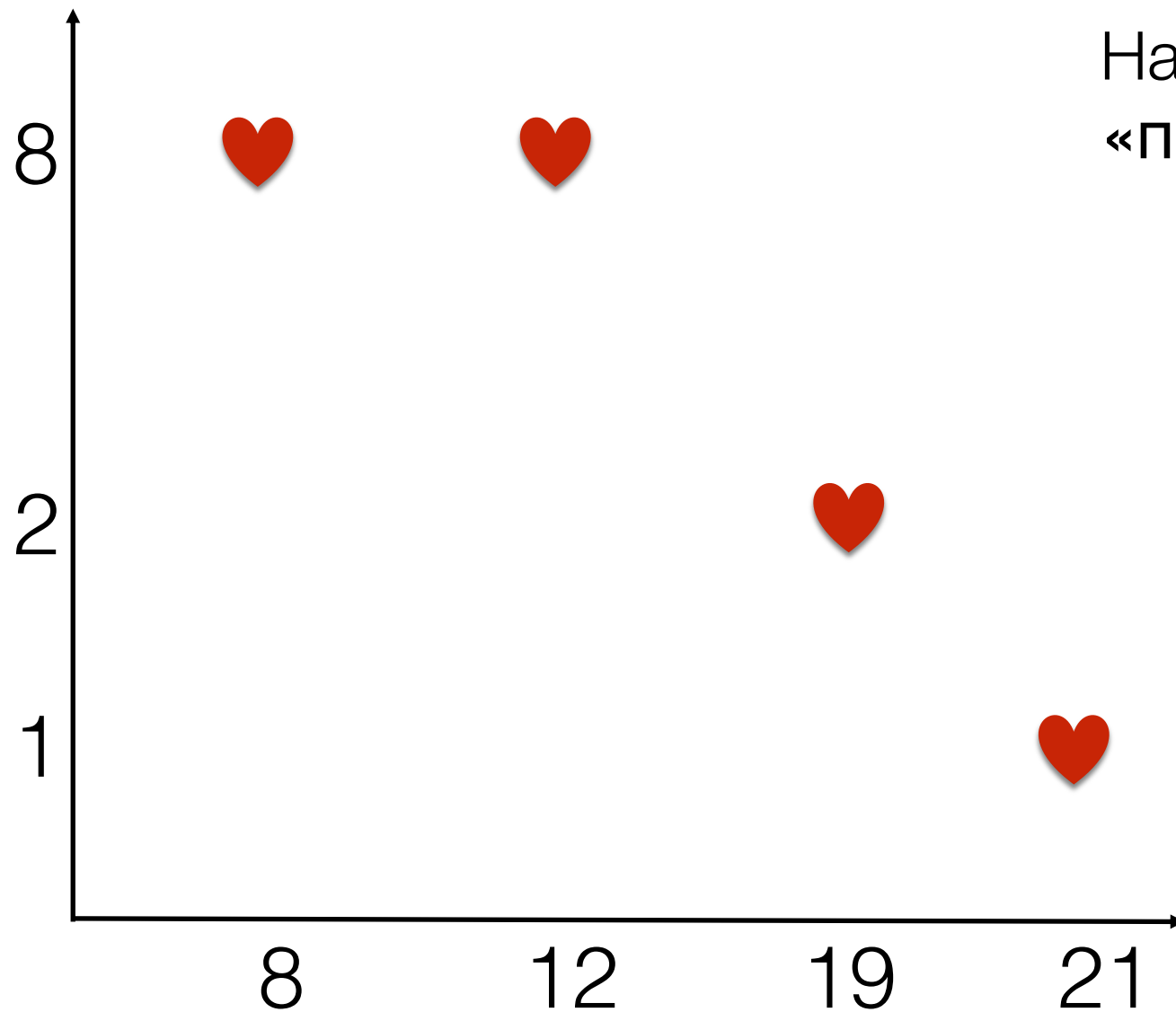
НАША ЗАДАЧКА



Надо задать вопрос вида
«правда ли, что признак $>$ порога?»

t	y
21	1
19	2
12	8
8	8

НАША ЗАДАЧКА

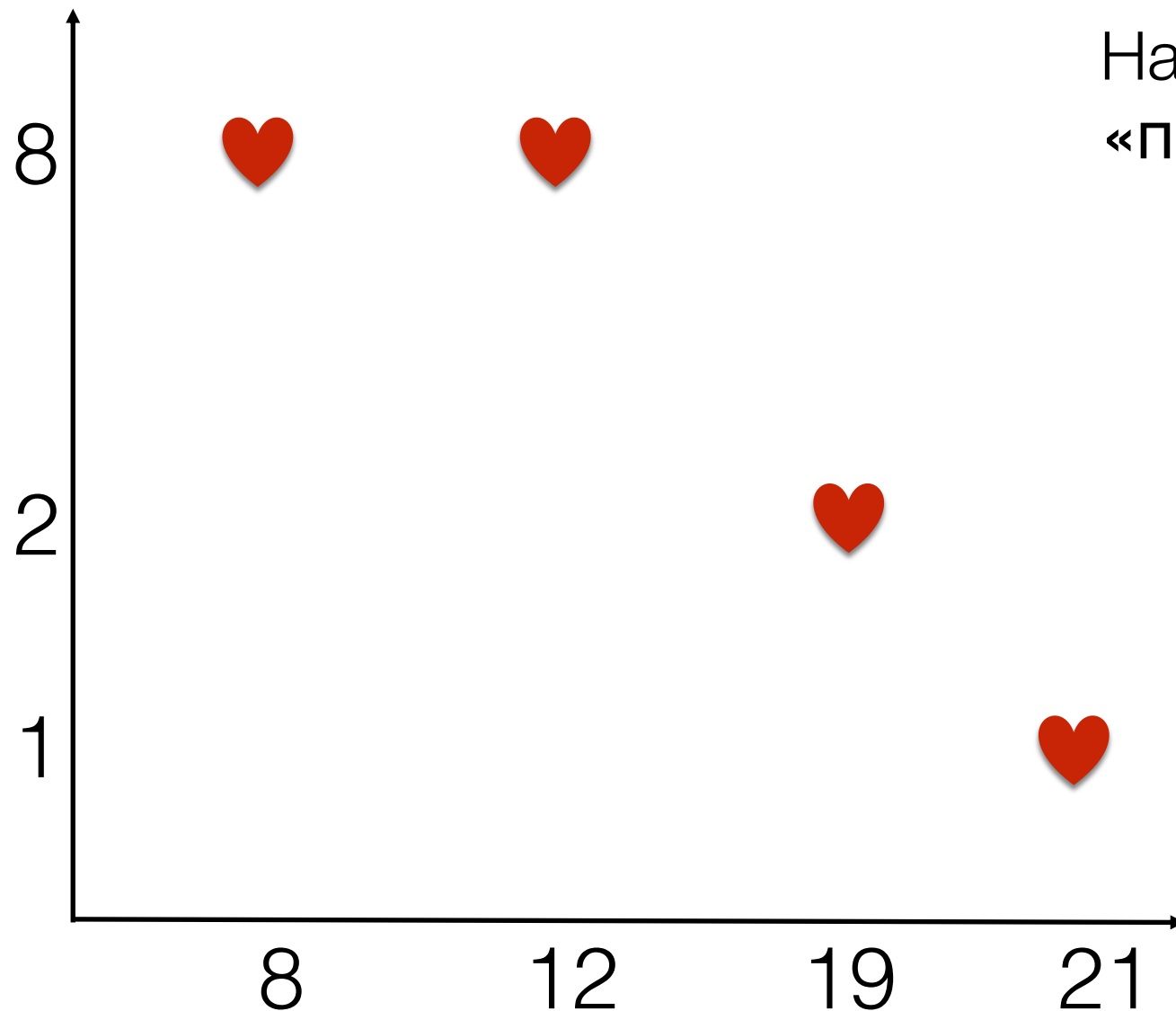


Надо задать вопрос вида
«правда ли, что признак $>$ порога?»

Где провести порог?

t	y
21	1
19	2
12	8
8	8

НАША ЗАДАЧКА



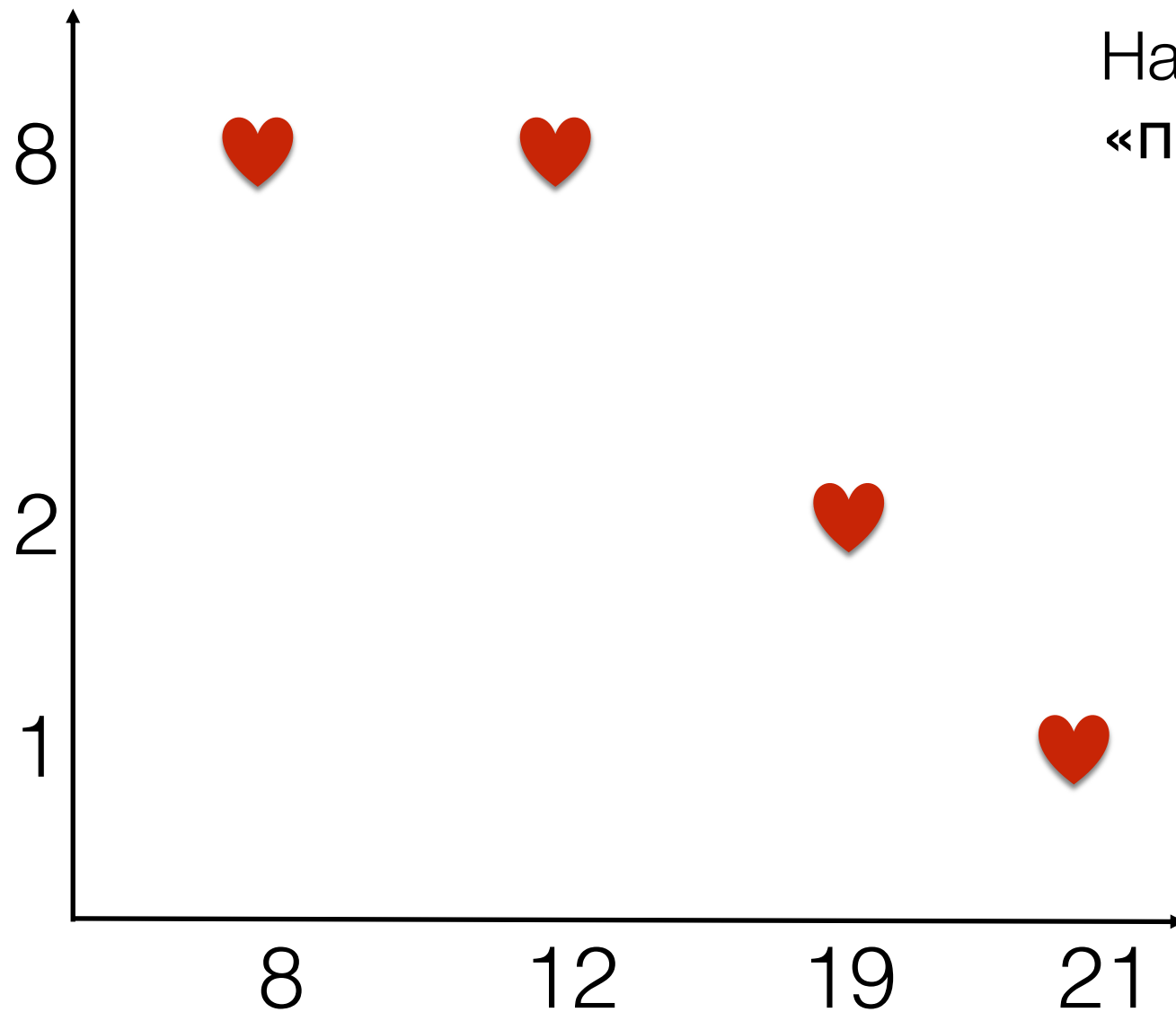
Надо задать вопрос вида
«правда ли, что признак $>$ порога?»

Где провести порог?

Придется перебрать
несколько, чтобы
найти такой, при
котором будет
минимальный MSE!

t	y
21	1
19	2
12	8
8	8

НАША ЗАДАЧКА



Надо задать вопрос вида
«правда ли, что признак $>$ порога?»

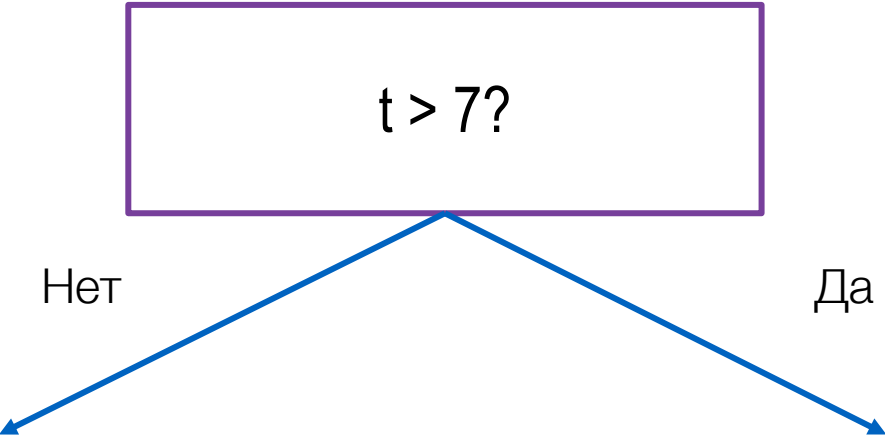
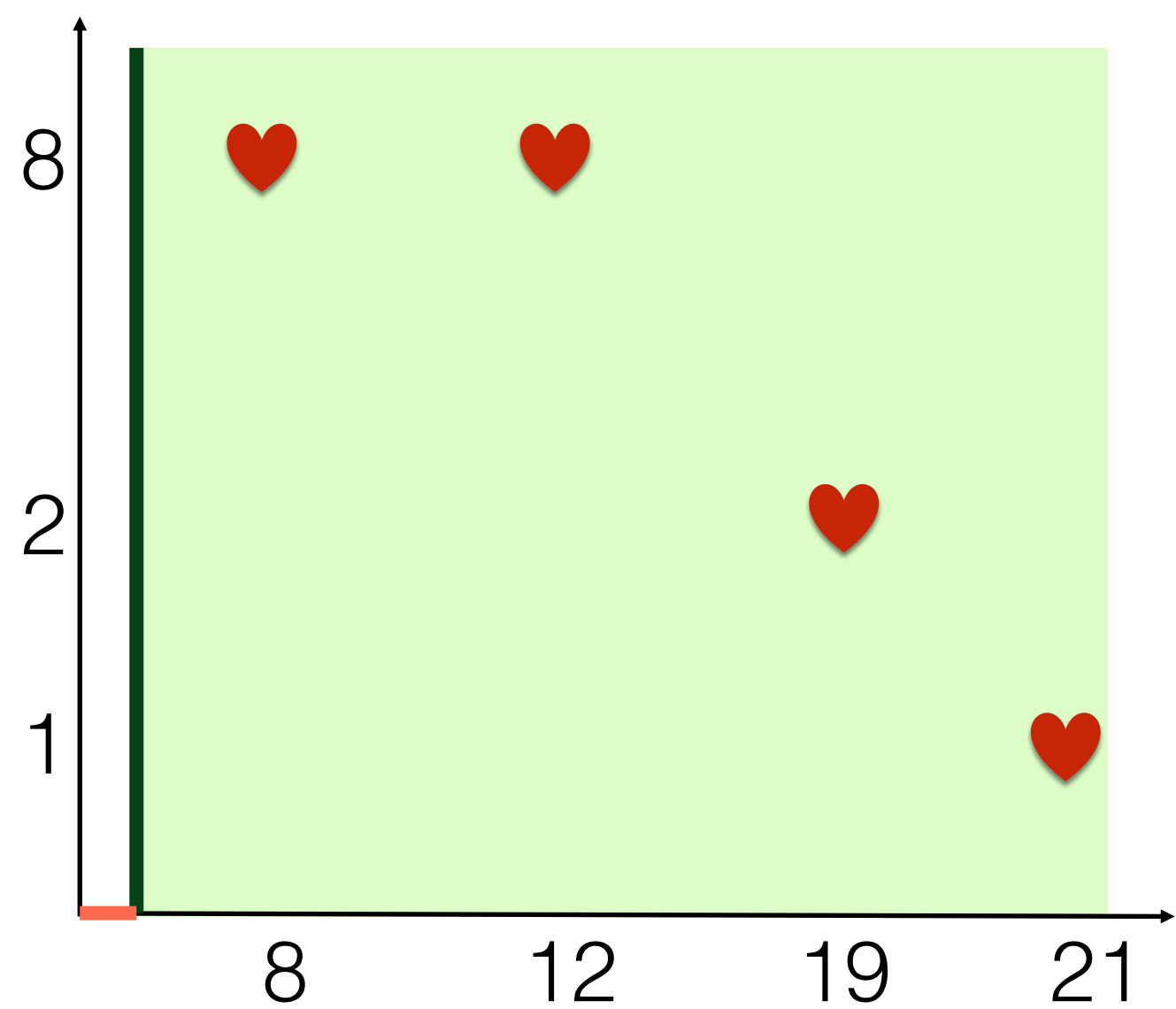
Где провести порог?

Придется перебрать
несколько, чтобы
найти такой, при
котором будет
минимальный MSE!

Пороги для пробы:
7, 9, 13, 20, 22

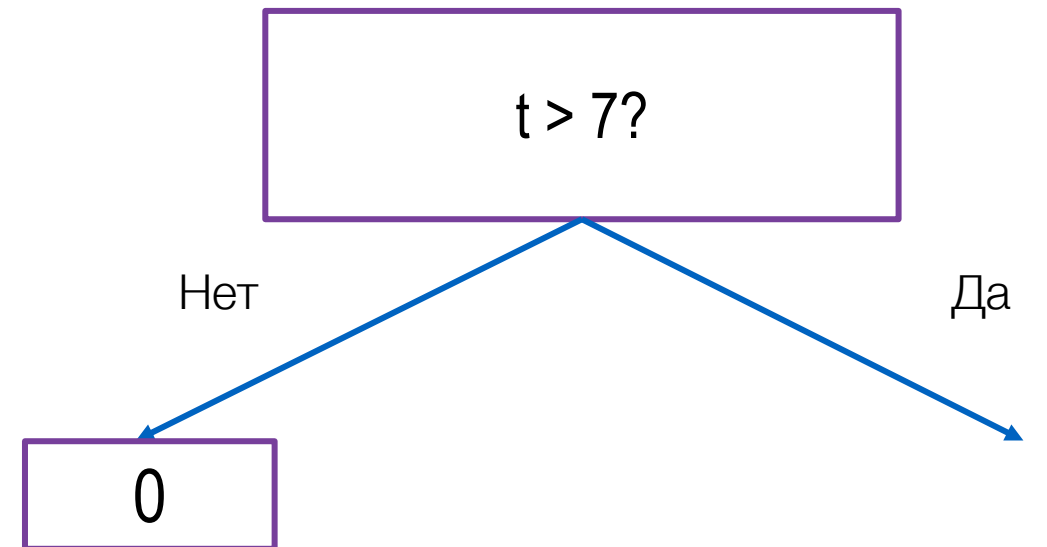
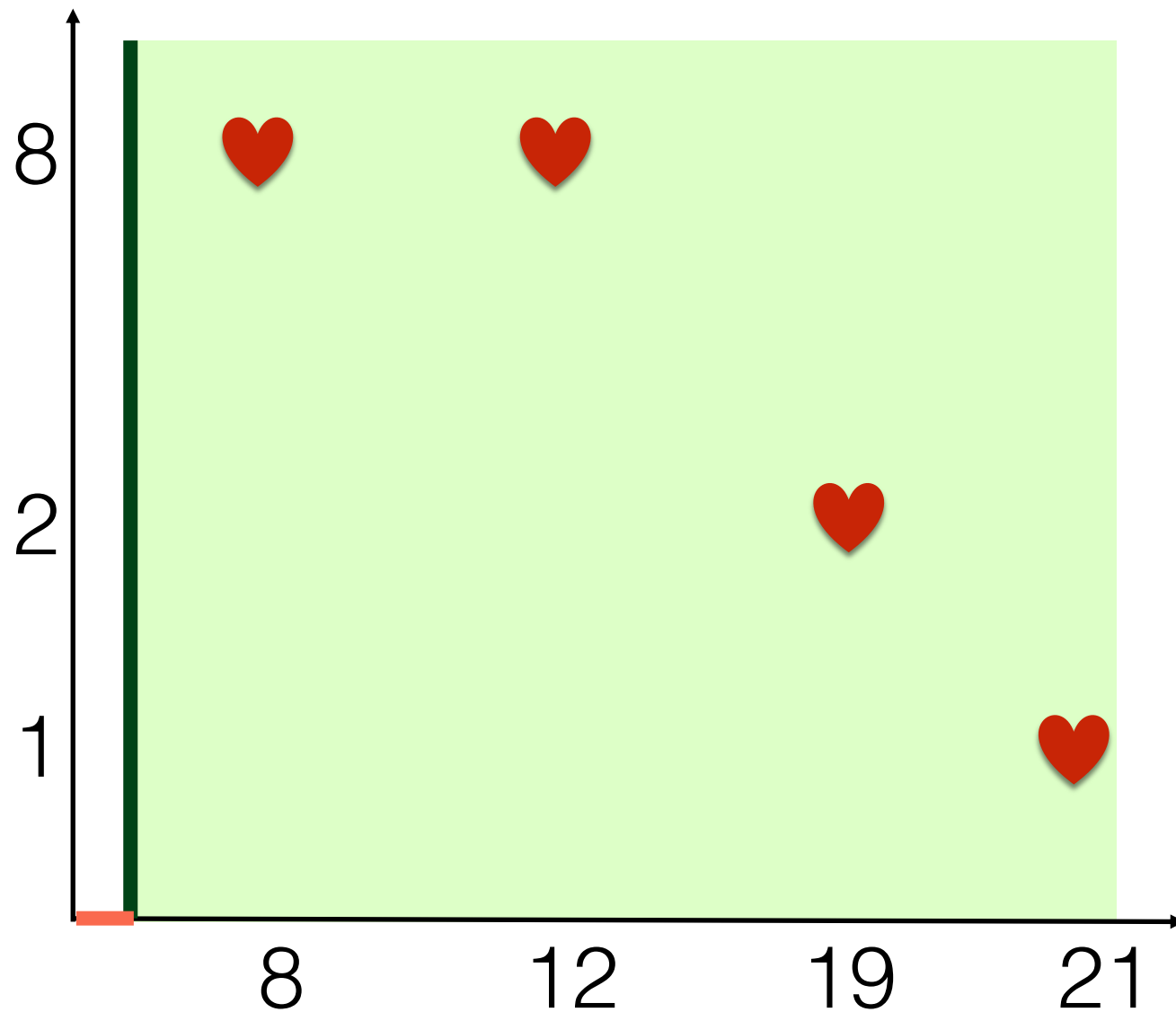
t	y
21	1
19	2
12	8
8	8

НАША ЗАДАЧКА



t	y
21	1
19	2
12	8
8	8

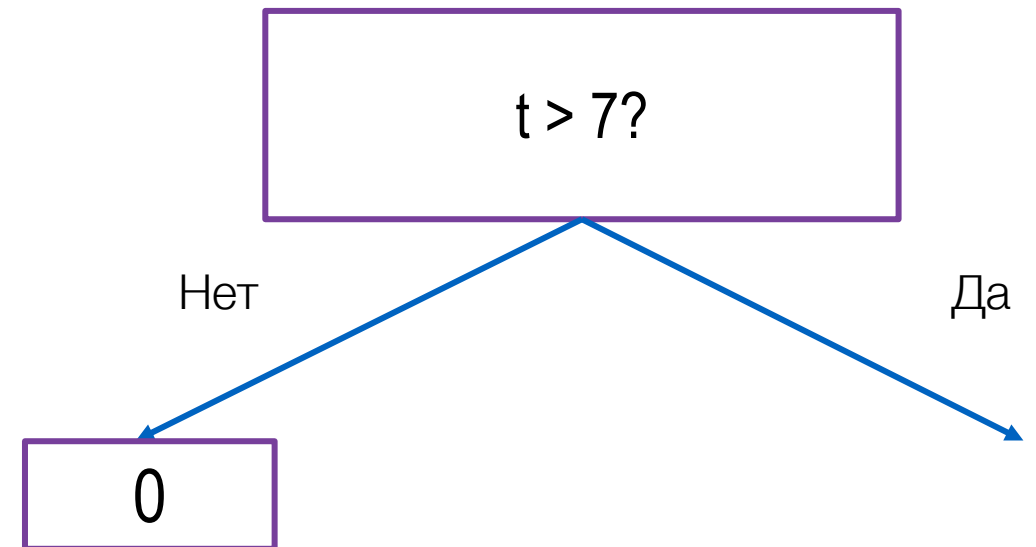
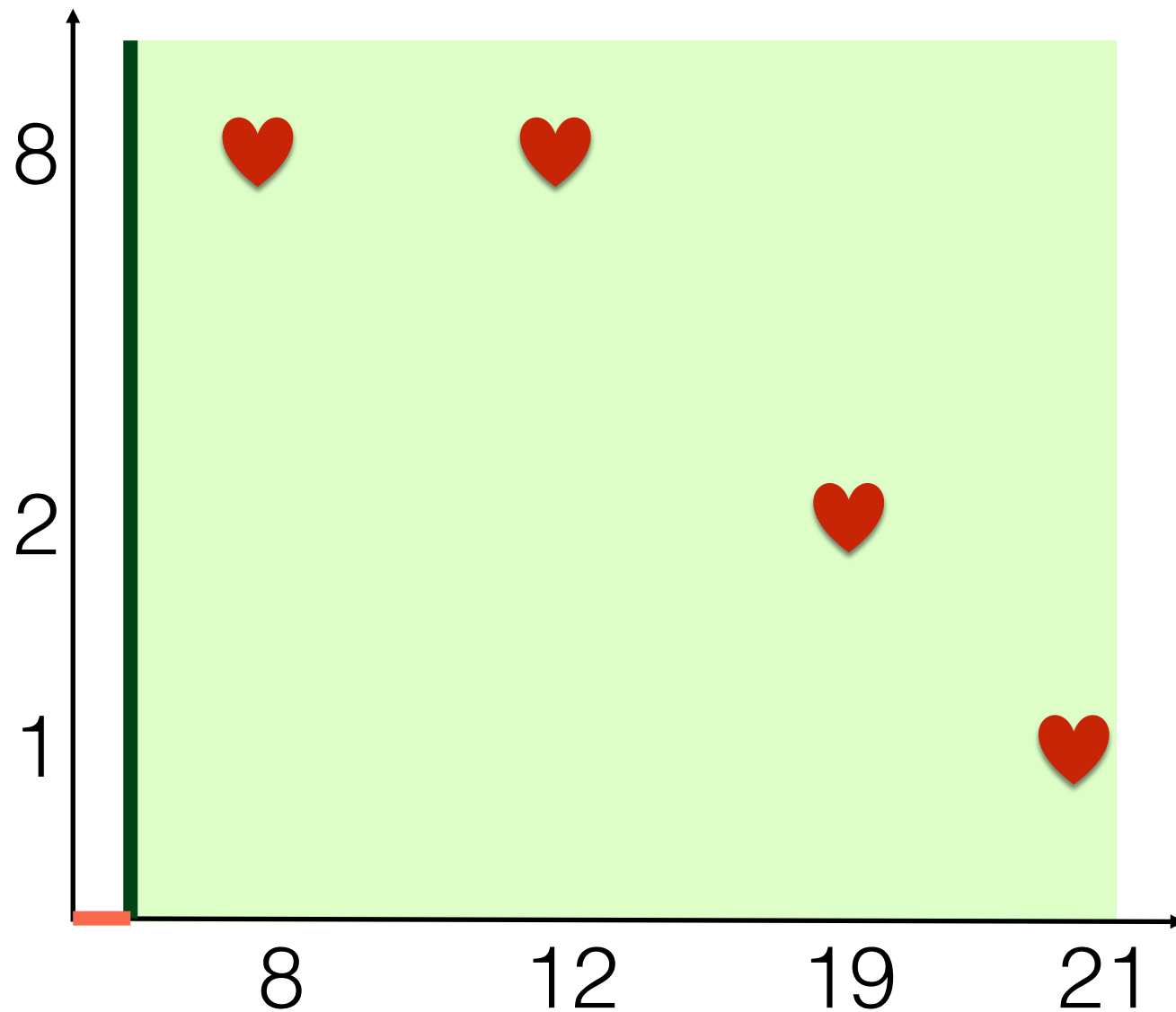
НАША ЗАДАЧКА



В ветку «нет» не попало ни одно наблюдение. Там прогноз – 0 (коралловая линия)

t	y
21	1
19	2
12	8
8	8

НАША ЗАДАЧКА

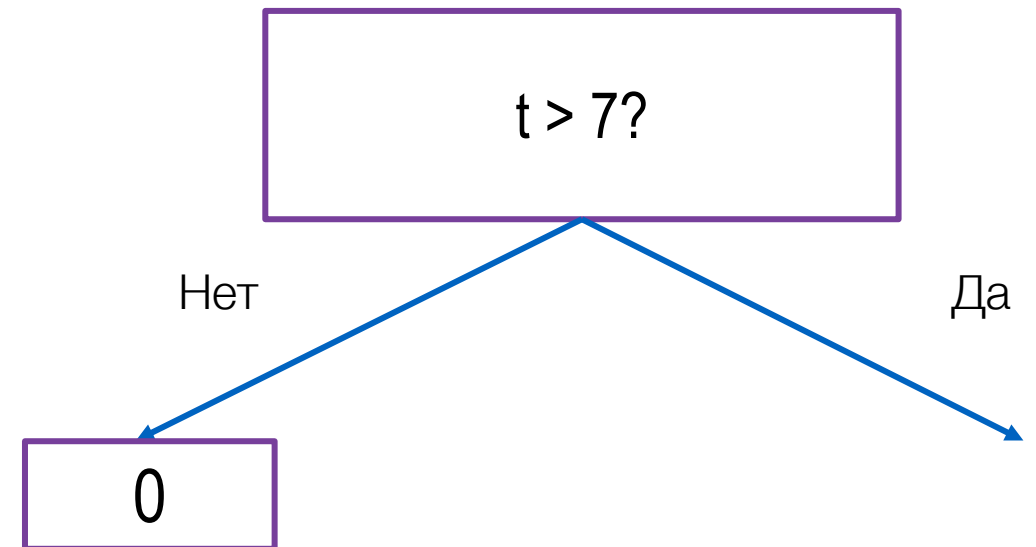
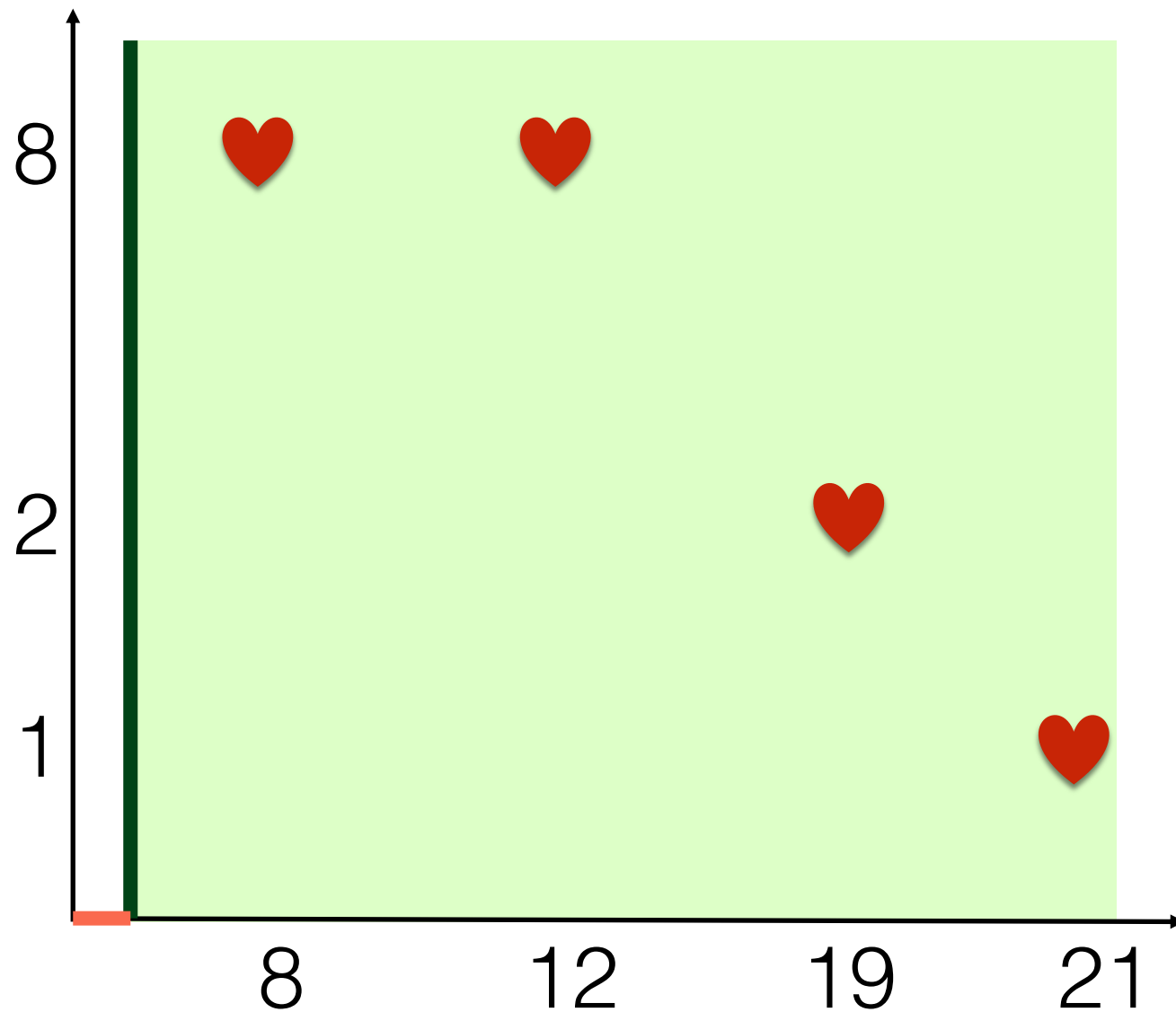


В ветку «нет» не попало ни одно наблюдение. Там прогноз – 0

В ветку да попали все. Считаем среднее – это и будет прогноз

t	y
21	1
19	2
12	8
8	8

НАША ЗАДАЧКА



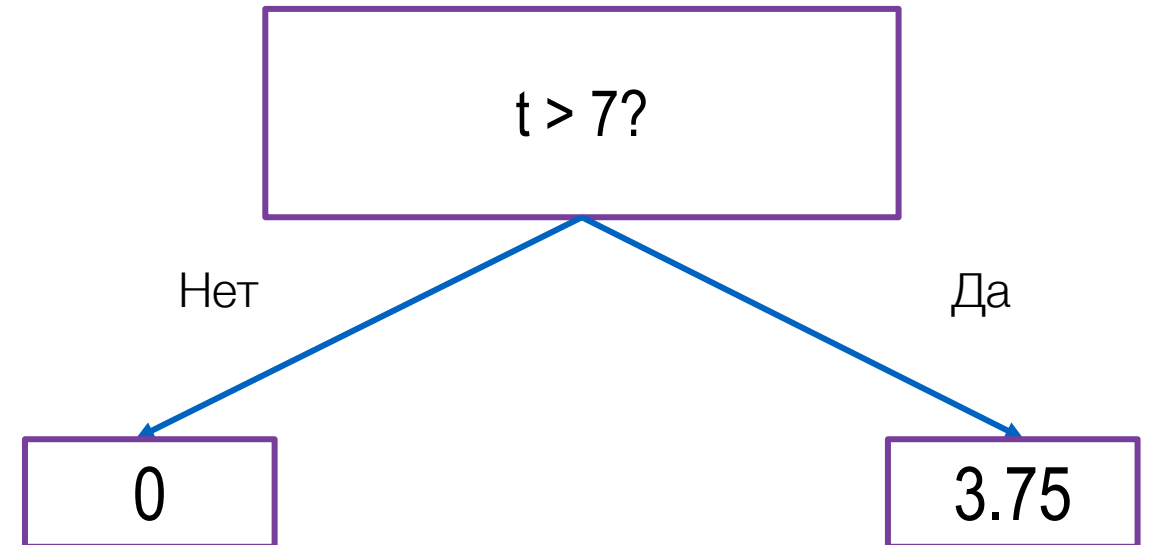
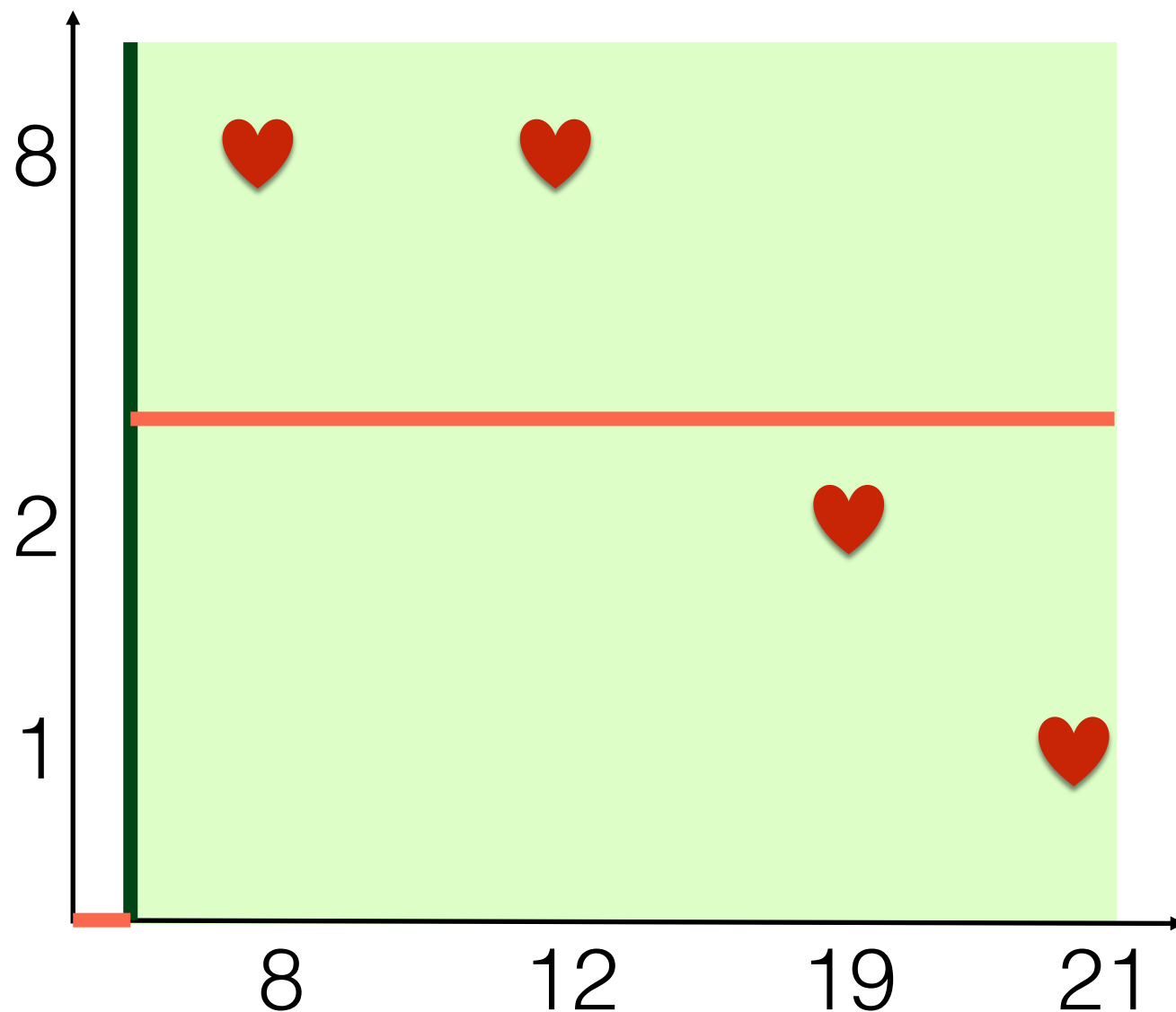
В ветку «нет» не попало ни одно наблюдение. Там прогноз – 0

В ветку да попали все. Считаем среднее – это и будет прогноз

t	y
21	1
19	2
12	8
8	8

$$\hat{y} = \frac{1}{4} (1 + 2 + 8 + 8) = 3.75$$

НАША ЗАДАЧКА



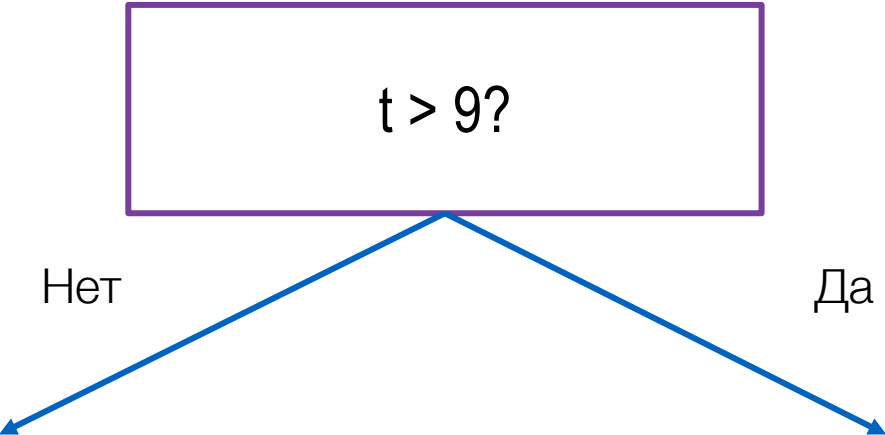
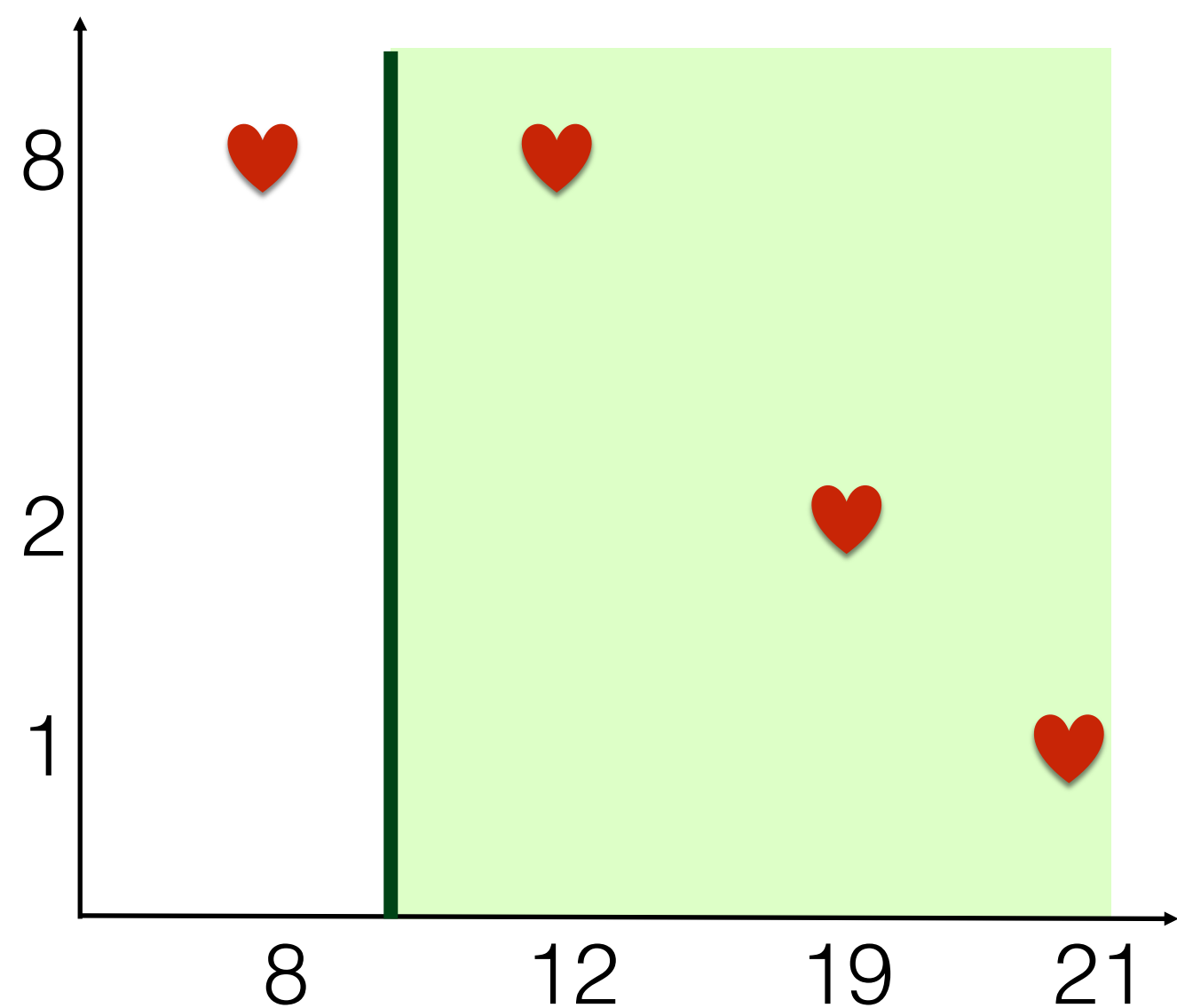
Теперь считаем ошибку прогноза.

И запоминаем эту цифру!

t	y	\hat{y}	$y - \hat{y}$	$(y - \hat{y})^2$
21	1	3.75	-2.75	7,5625
19	2	3.75	-1.75	3,0625
12	8	3.75	4.25	18,0625
8	8	3.75	4.25	18,0625

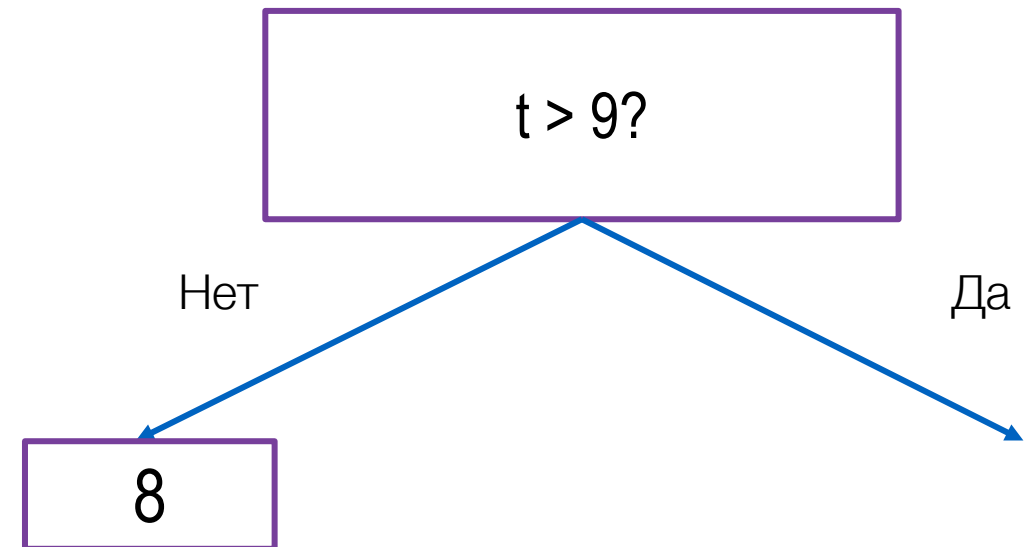
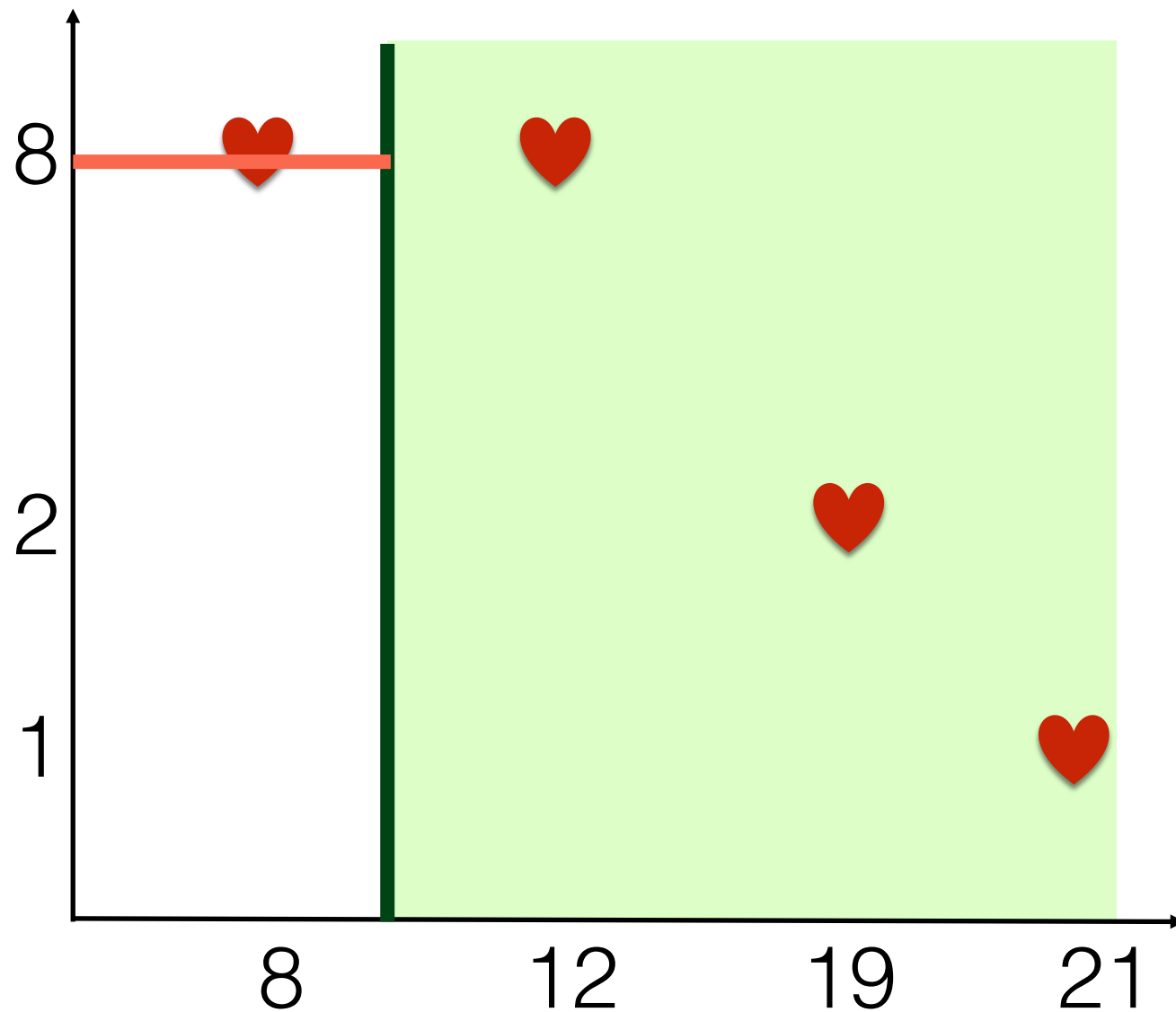
$$\text{MSE} = \frac{1}{4} (7.5625 + 3.0625 + 18.0625 + 18.0625) = 11.6875$$

НАША ЗАДАЧКА



t	y
21	1
19	2
12	8
8	8

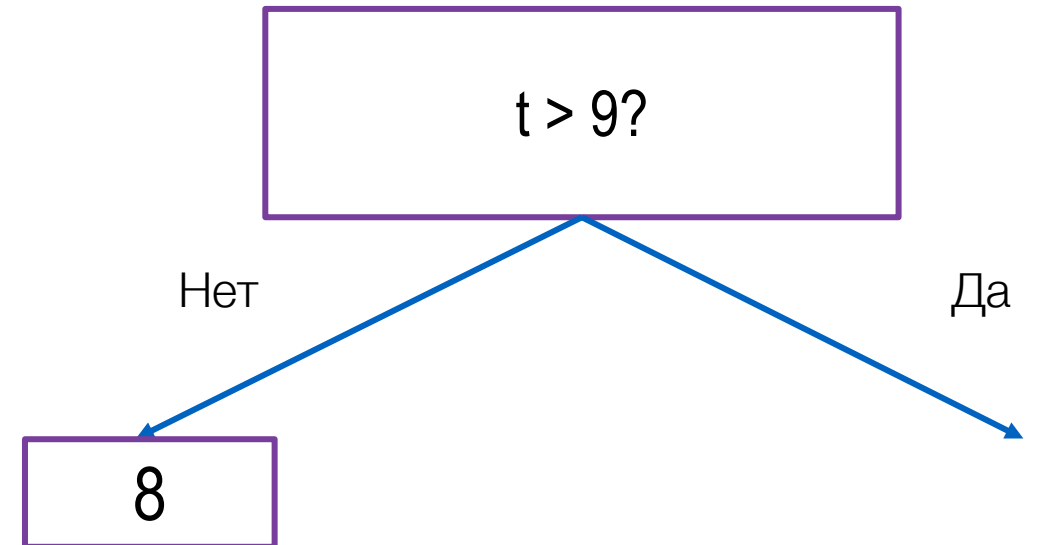
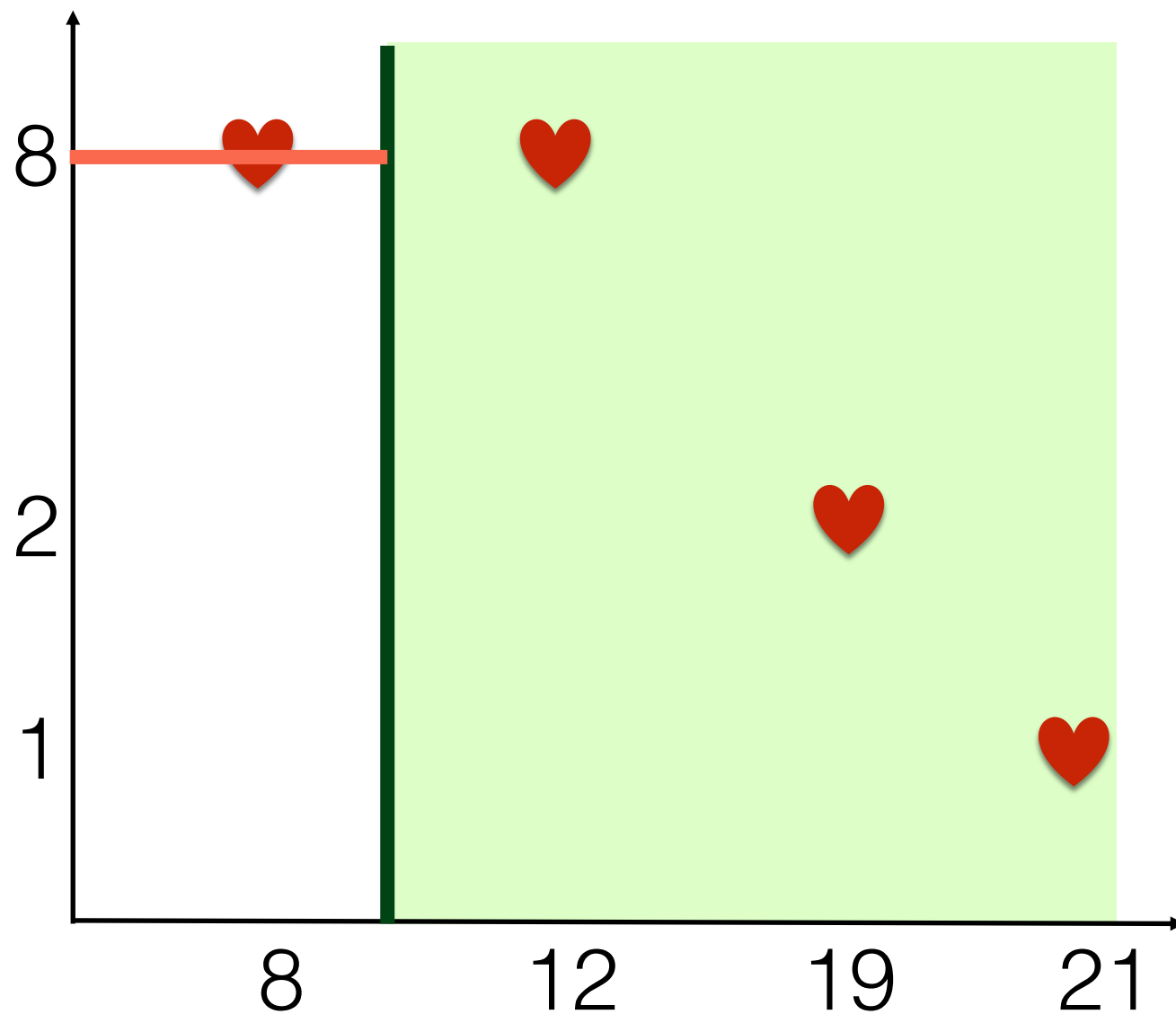
НАША ЗАДАЧКА



В ветку «нет попало одно наблюдение. Там прогноз – 8

t	y
21	1
19	2
12	8
8	8

НАША ЗАДАЧКА

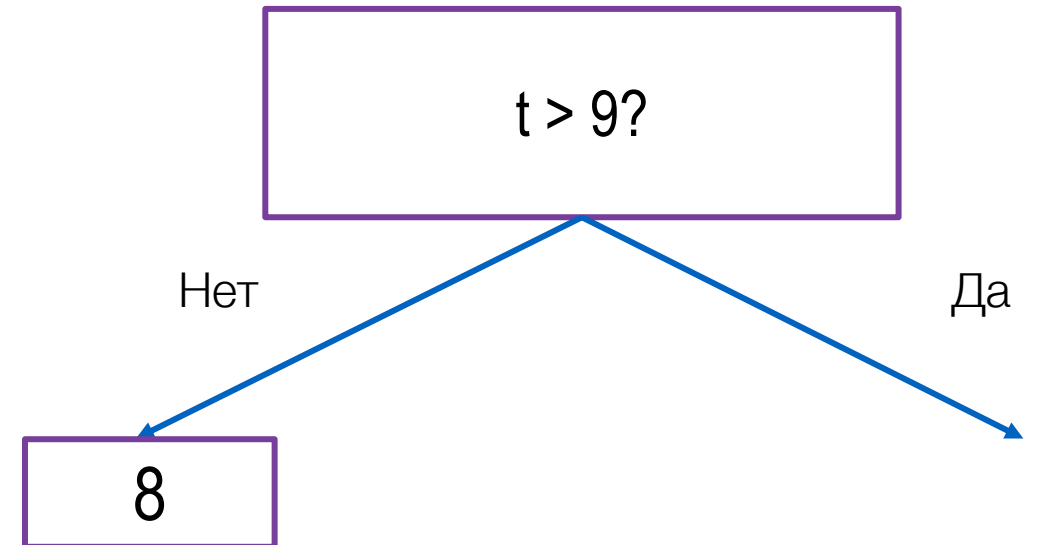
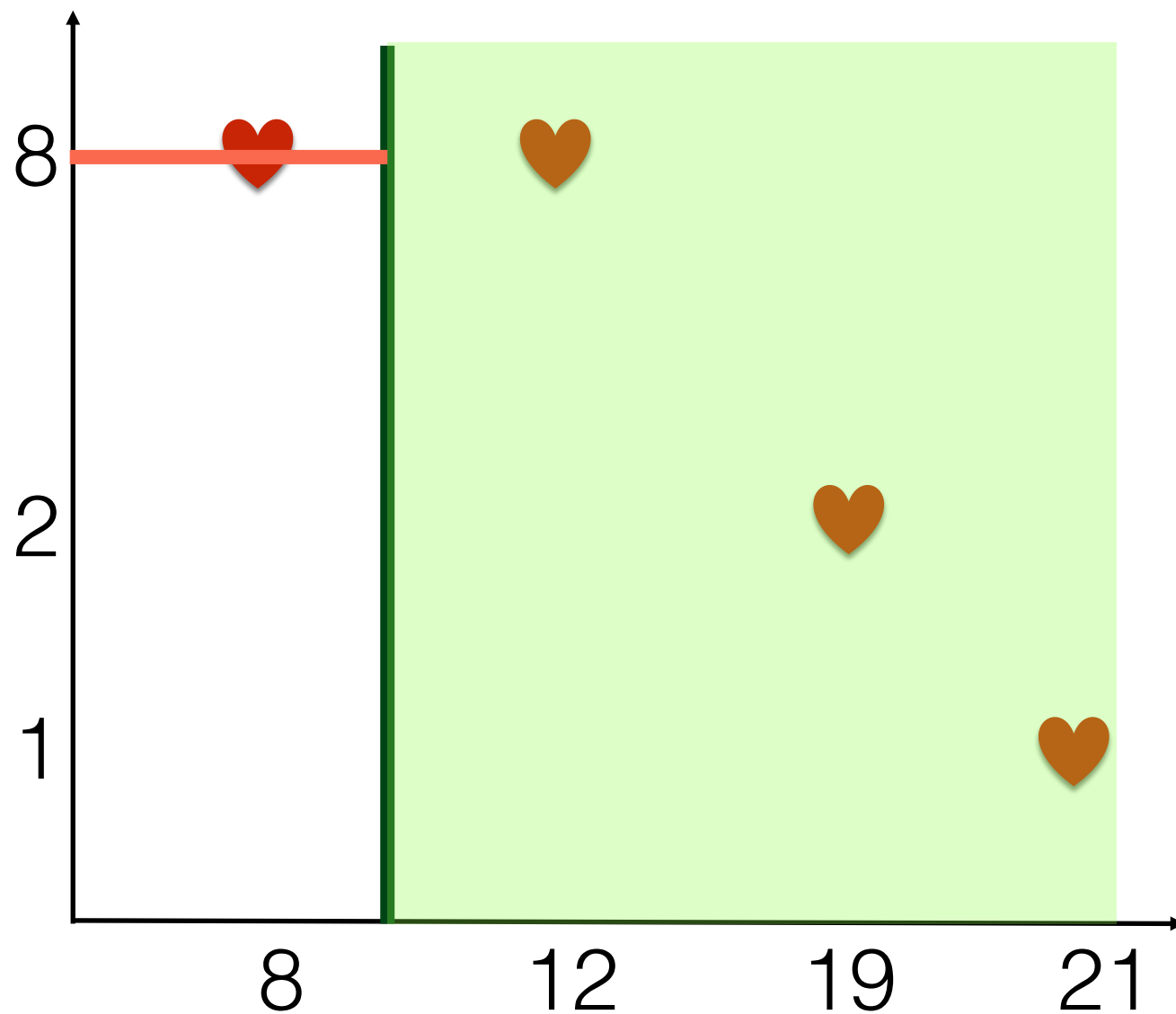


В ветку «нет» попало одно наблюдение. Там прогноз – 8

В ветку «да» попали три. Считаем среднее – это и будет прогноз

t	y
21	1
19	2
12	8
8	8

НАША ЗАДАЧКА



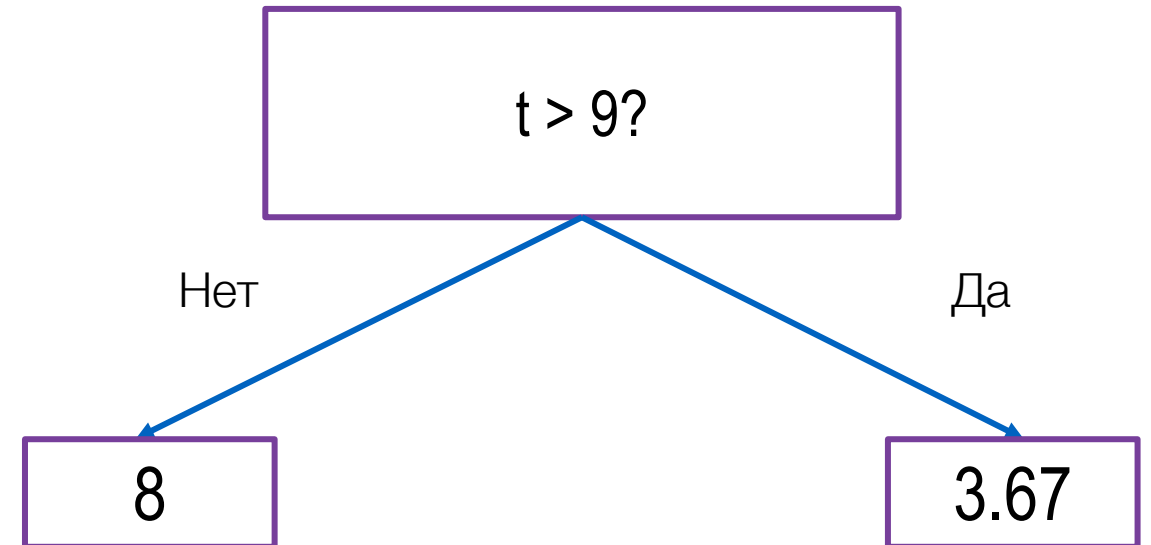
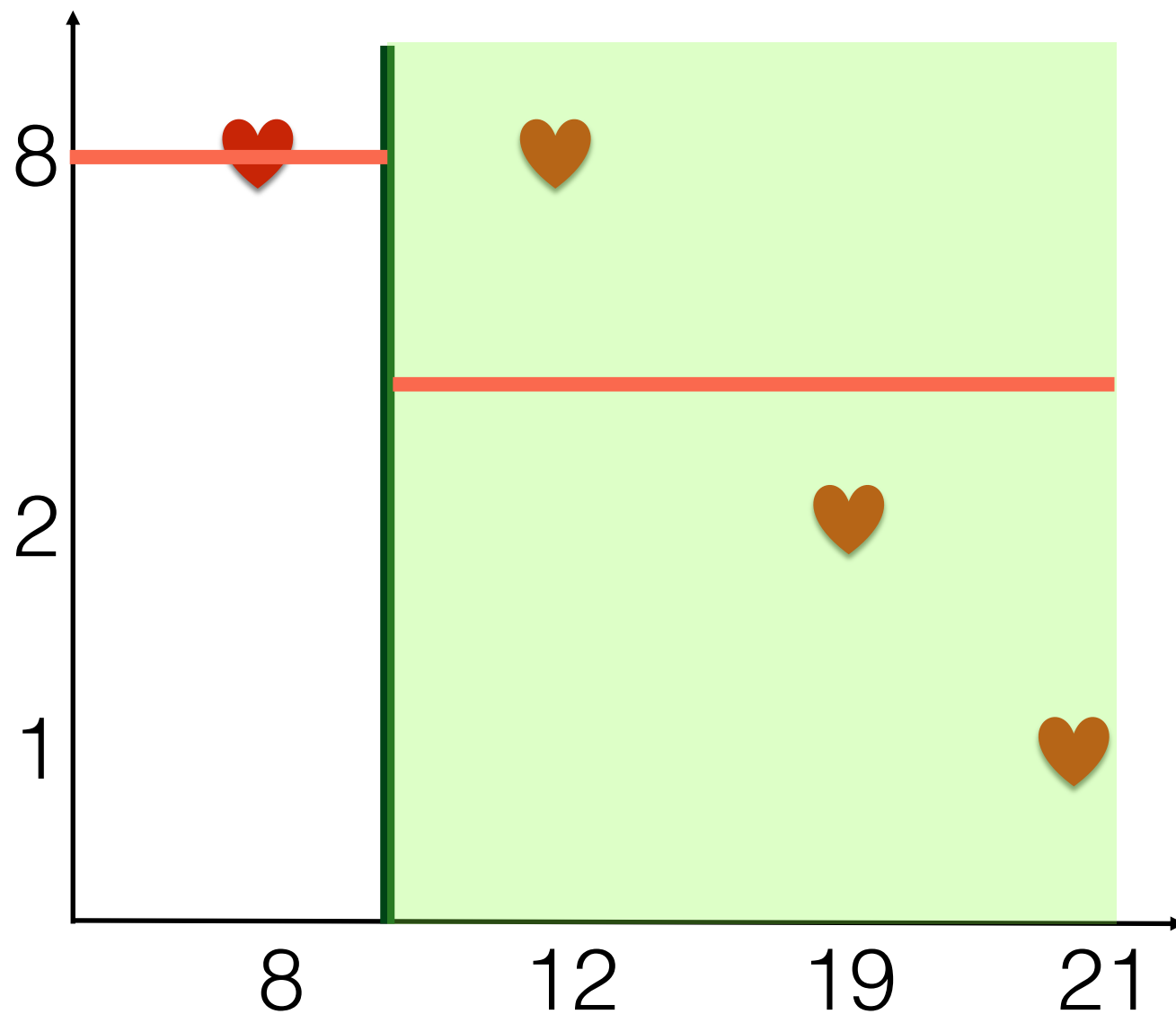
В ветку «нет» попало одно наблюдение. Там прогноз – 8

В ветку «да» попали три. Считаем среднее – это и будет прогноз

t	y
21	1
19	2
12	8
8	8

$$\hat{y} = \frac{1}{3} (1 + 2 + 8) = 3.67$$

НАША ЗАДАЧКА



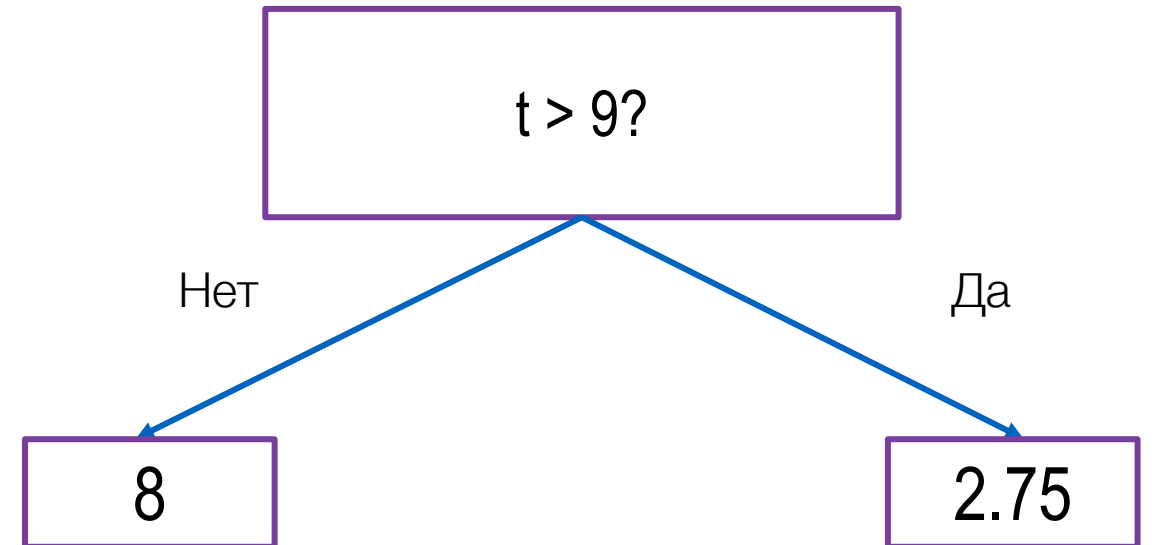
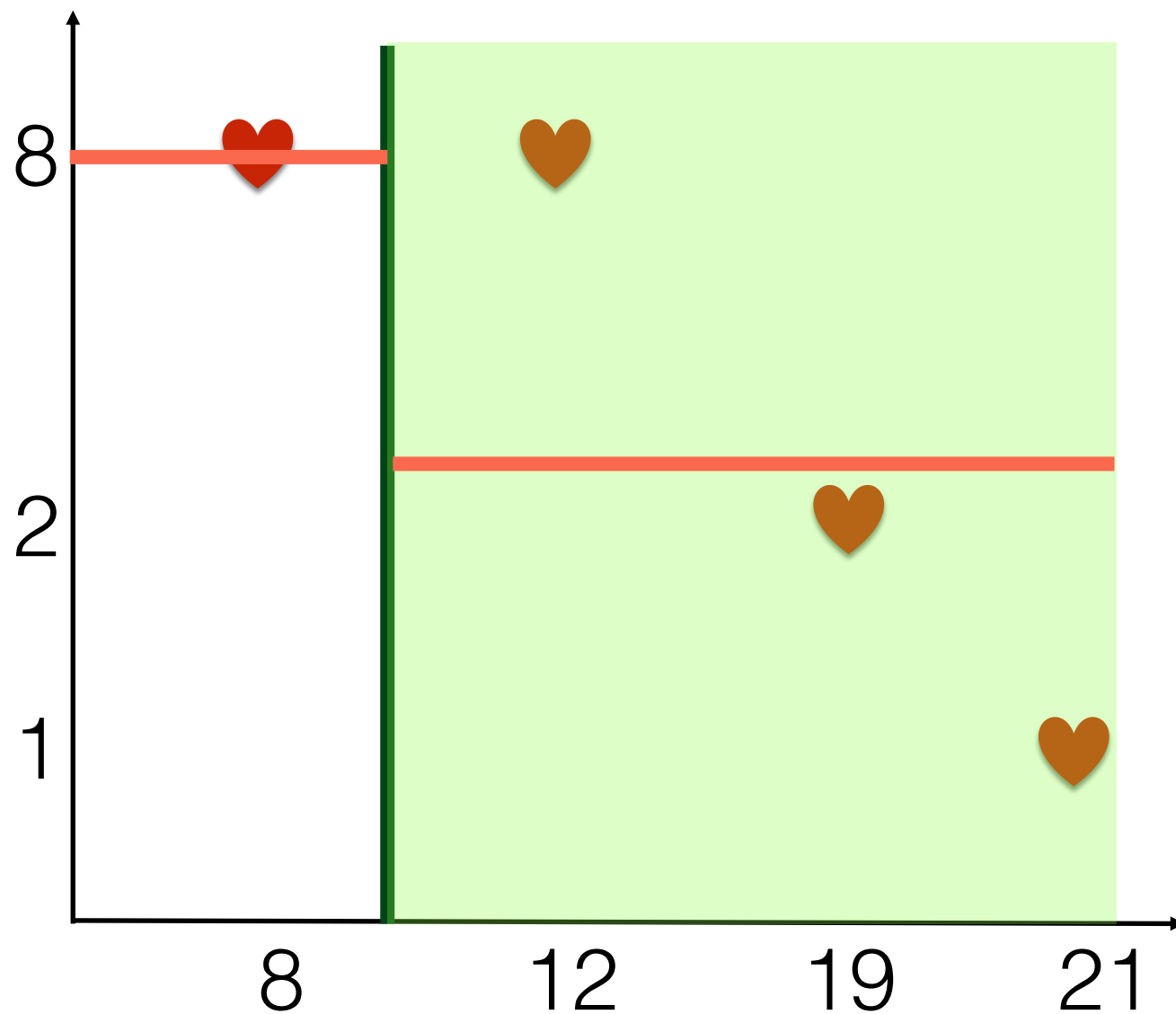
В ветку «нет» попало одно наблюдение. Там прогноз – 8

В ветку «да» попали три. Считаем среднее – это и будет прогноз

t	y
21	1
19	2
12	8
8	8

$$\hat{y} = \frac{1}{3} (1 + 2 + 8) = 3.67$$

НАША ЗАДАЧКА



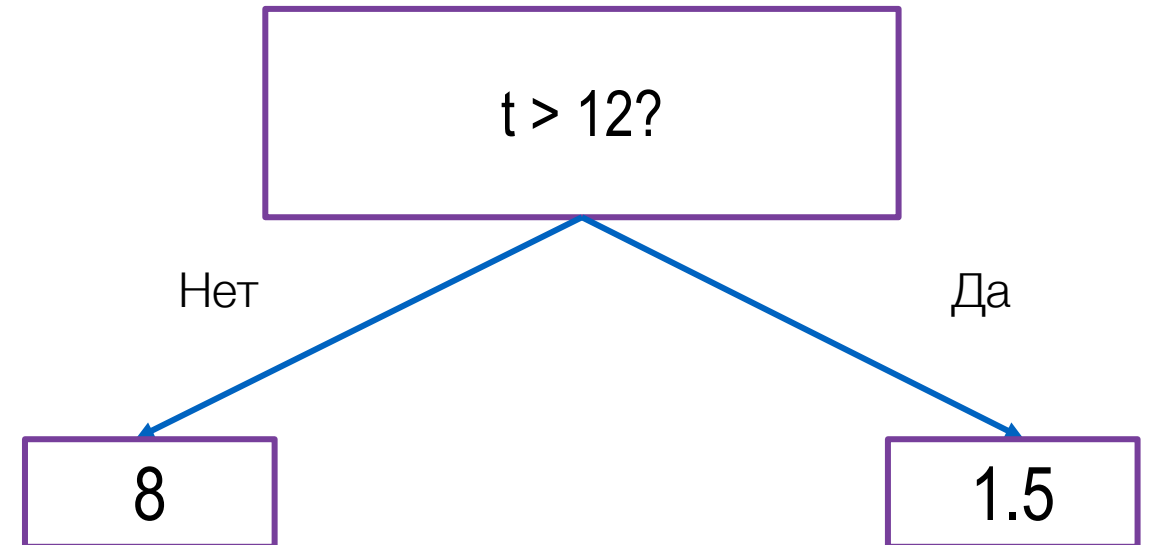
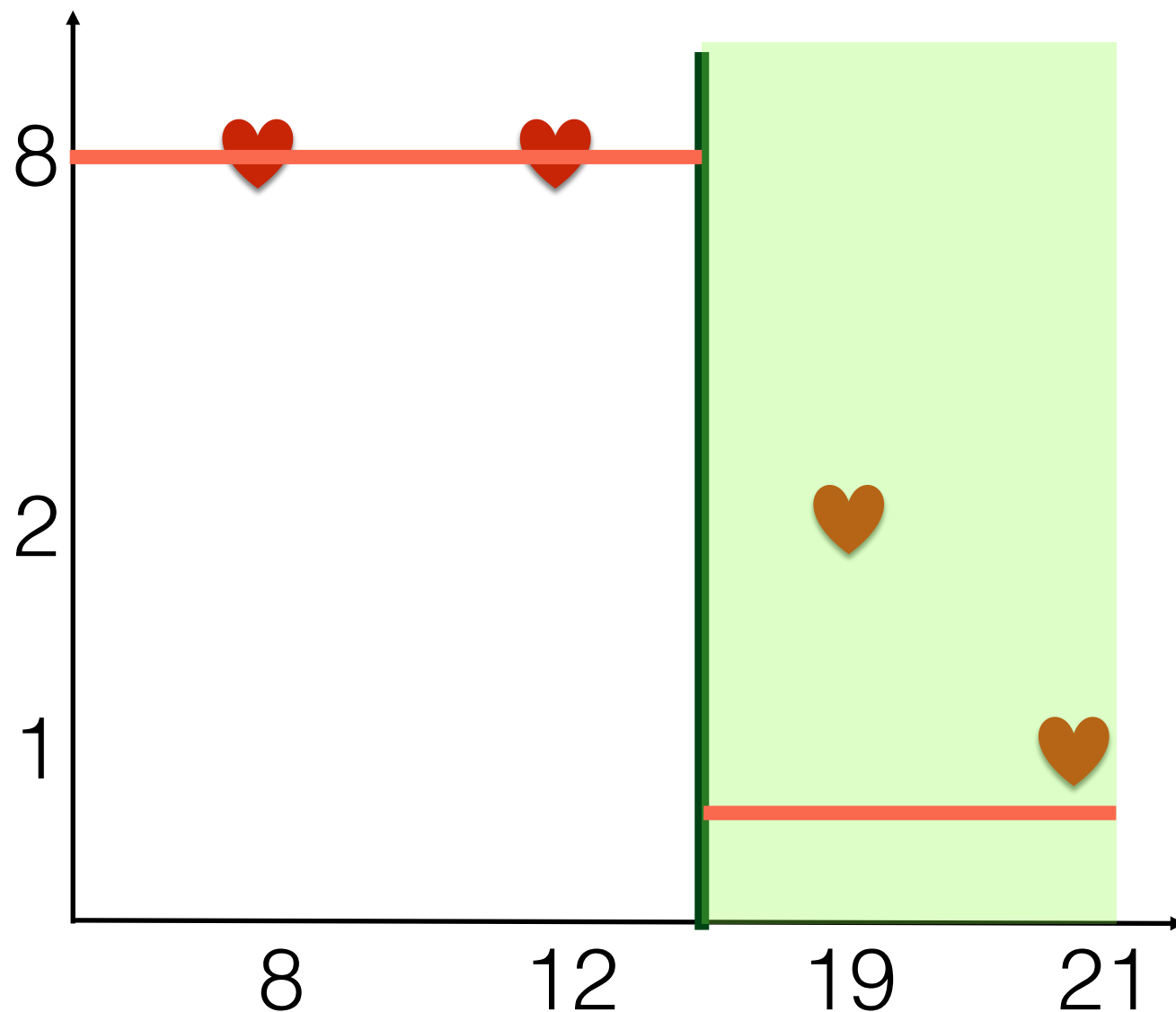
Теперь считаем ошибку прогноза.

И запоминаем эту цифру!

$$\text{MSE} = 7.17$$

t	y	\hat{y}	$y - \hat{y}$	$(y - \hat{y})^2$
21	1	3,67	-2,7	7,11
19	2	3,67	-1,7	2,78
12	8	3,67	4,33	18,8
8	8	8	0	0

НАША ЗАДАЧКА



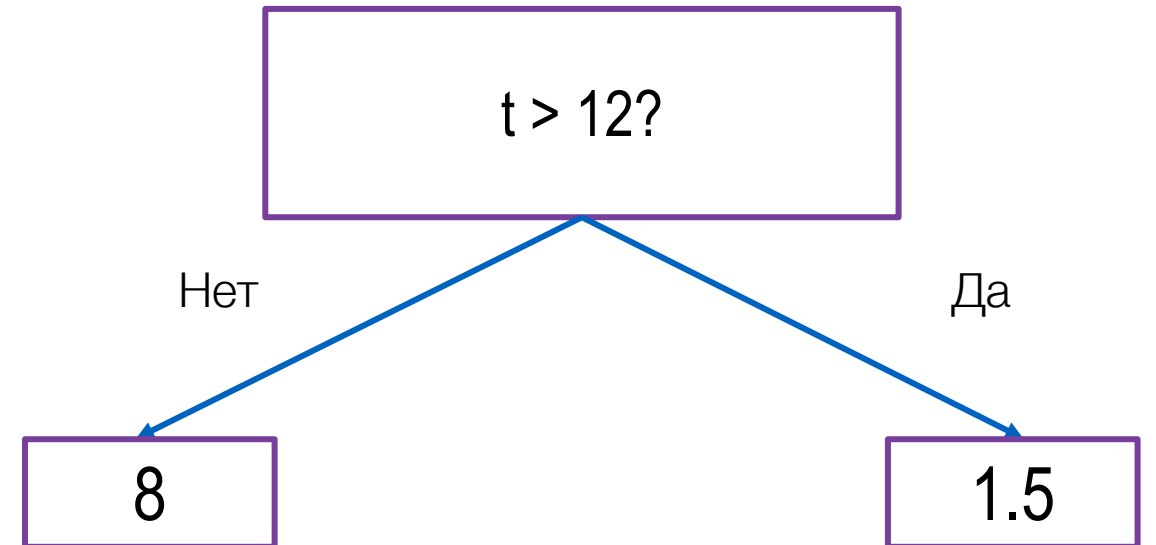
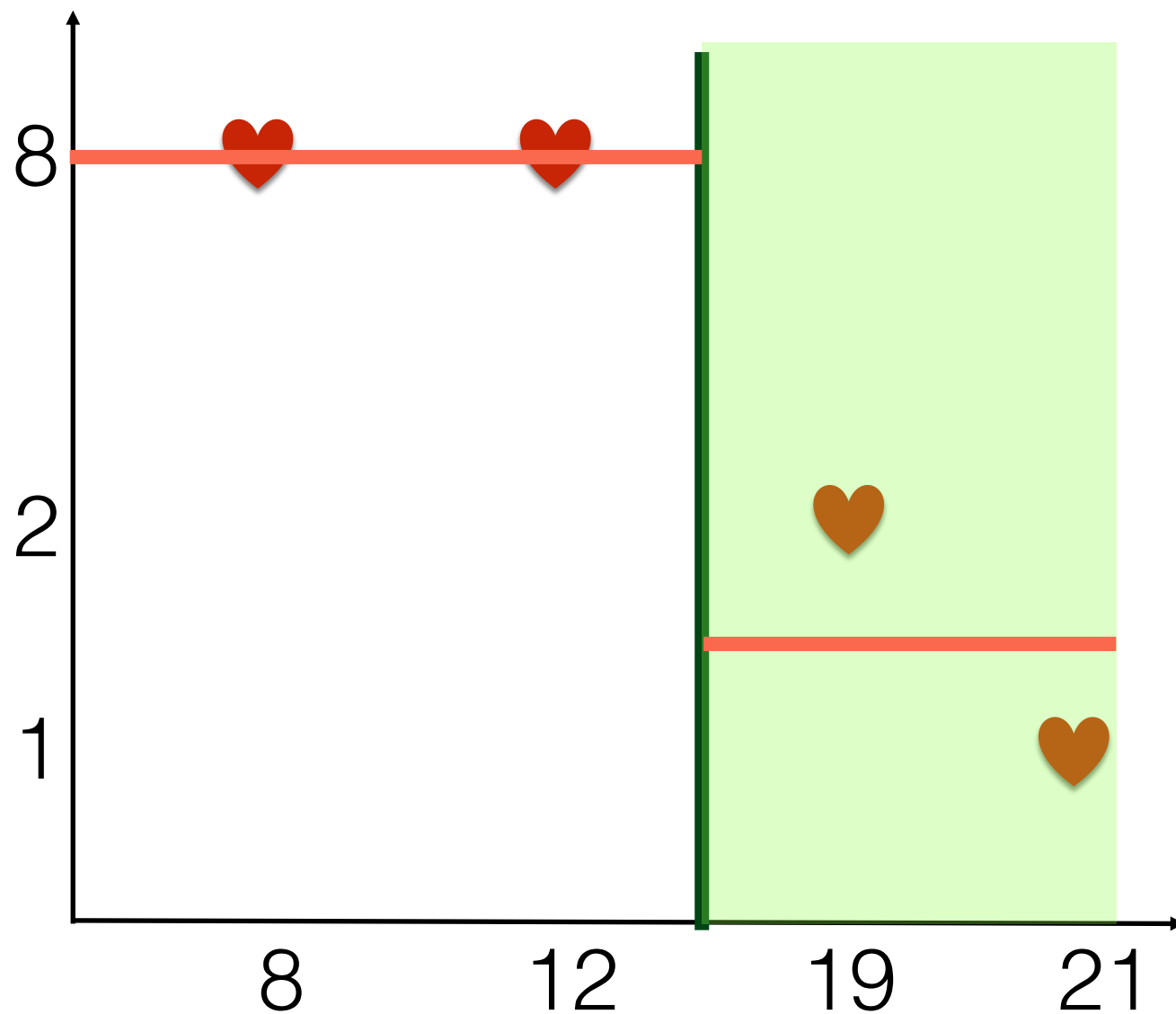
В ветку «нет» попало два наблюдения. Их среднее – 8

В ветку «да» попали два. Считаем прогноз

t	y
21	1
19	2
12	8
8	8

$$\hat{y} = \frac{1}{2} (1 + 2) = 1.5$$

НАША ЗАДАЧКА



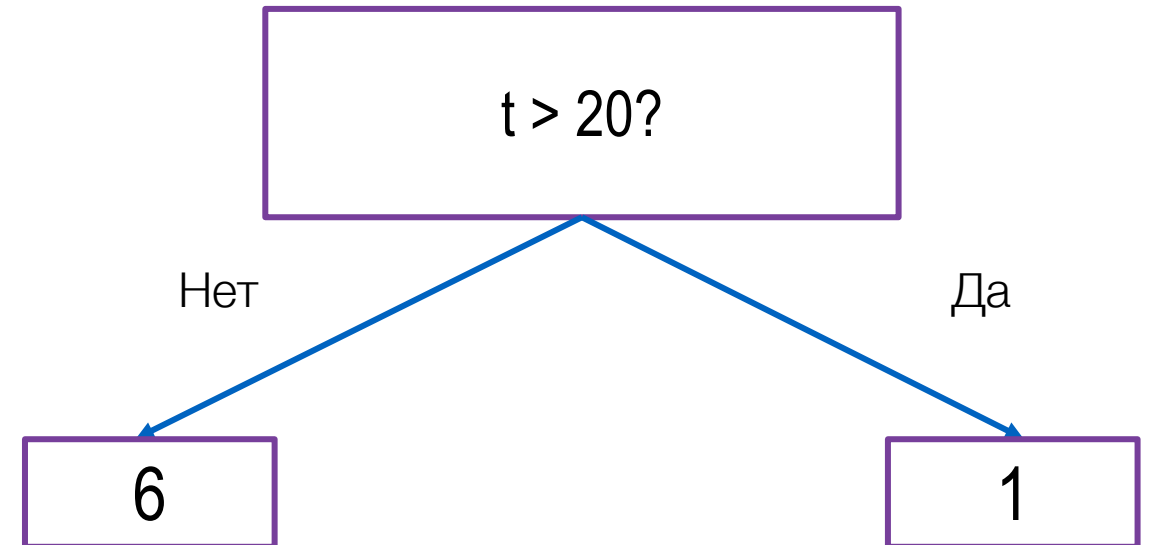
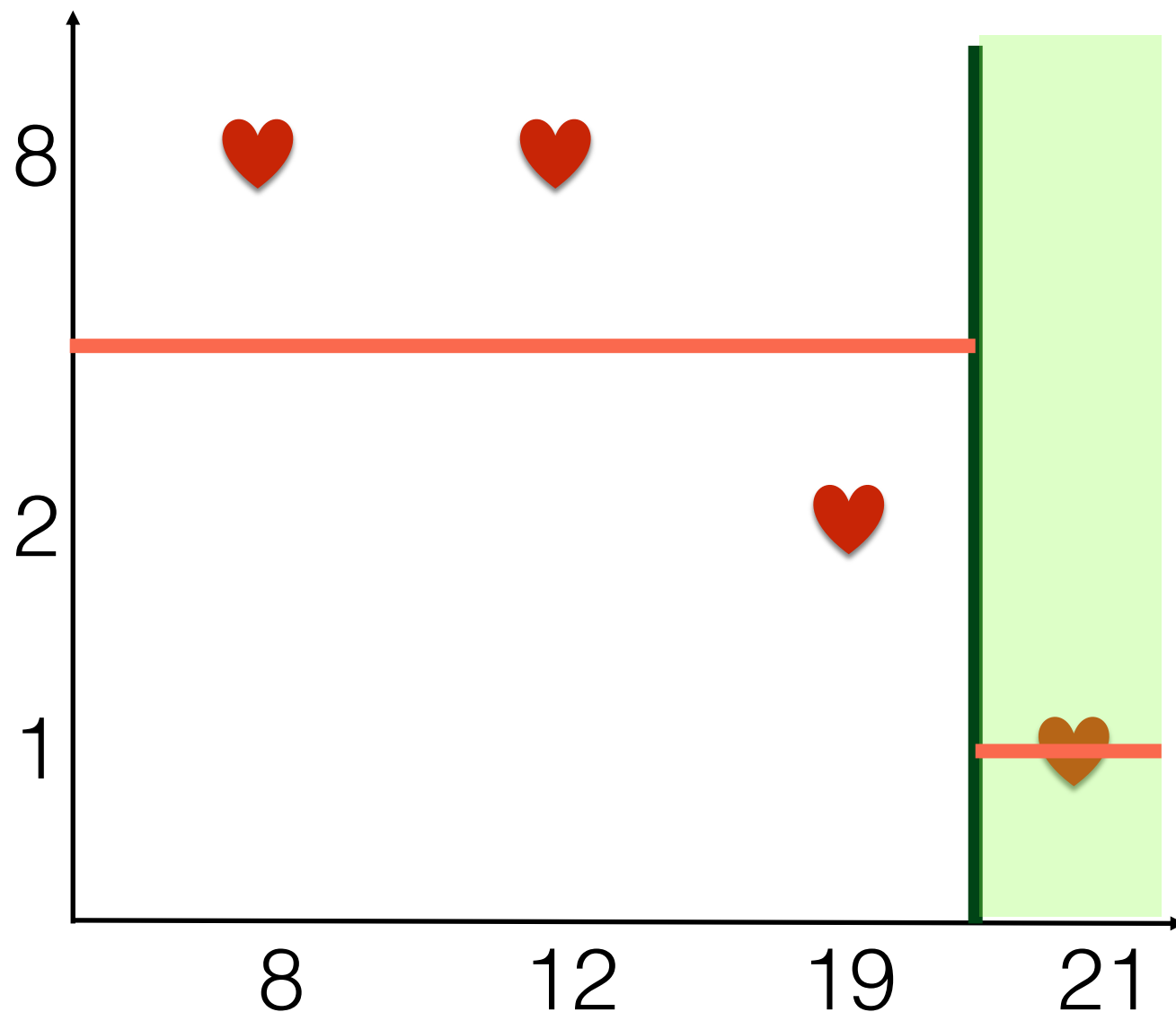
Теперь считаем ошибку прогноза.

И запоминаем эту цифру!

$$\text{MSE} = 0.125$$

t	y	\hat{y}	$y - \hat{y}$	$(y - \hat{y})^2$
21	1	0.75	-0,5	0,25
19	2	0.75	0,5	0,25
12	8	8	0	0
8	8	8	0	0

НАША ЗАДАЧКА



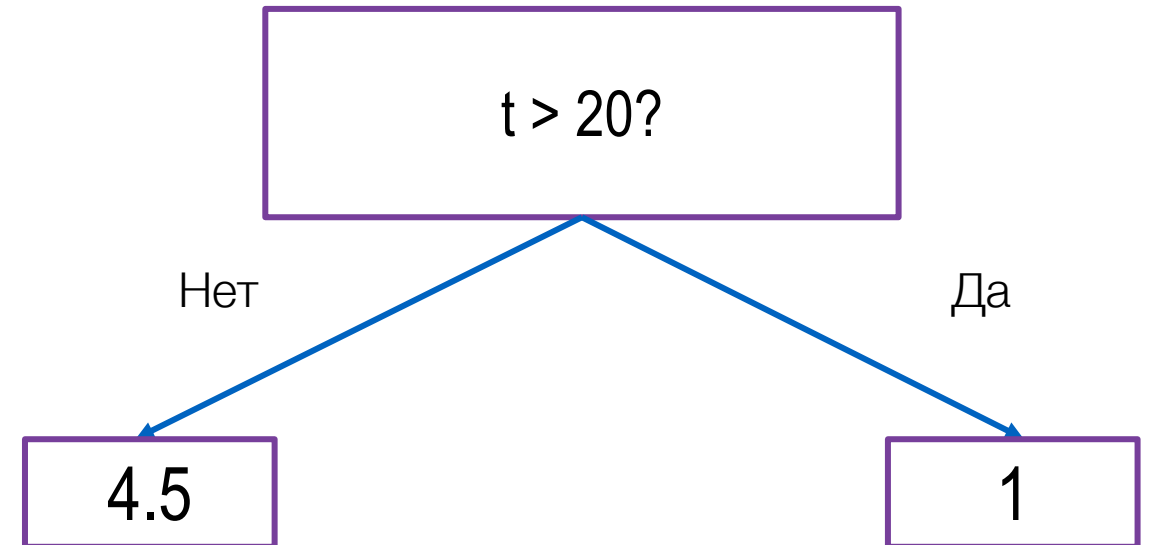
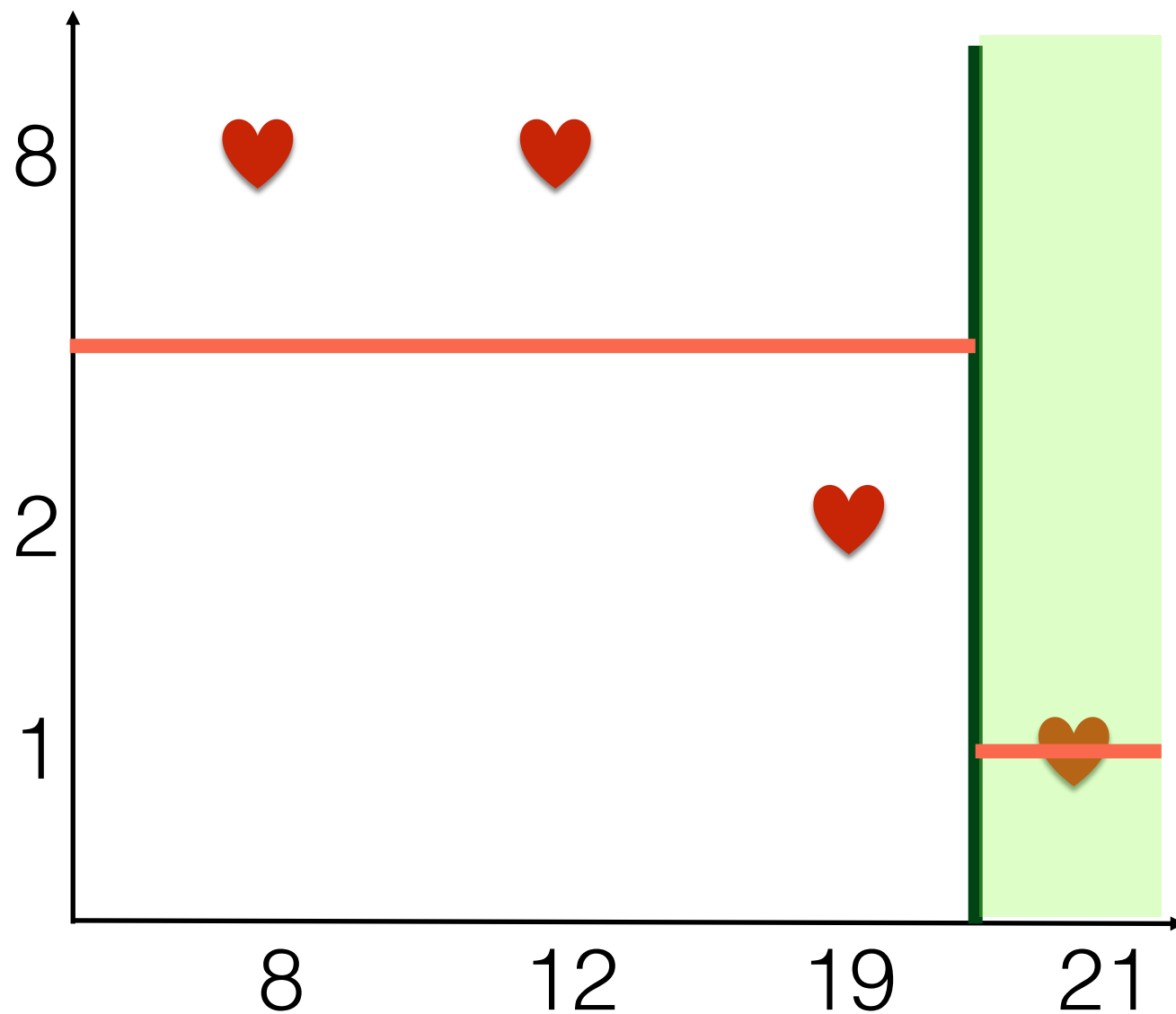
В ветку «да» попало одно наблюдение.

В ветку «нет» попали три.
Считаем прогноз

t	y
21	1
19	2
12	8
8	8

$$\hat{y} = \frac{1}{3} (2 + 8 + 8) = 6$$

НАША ЗАДАЧКА



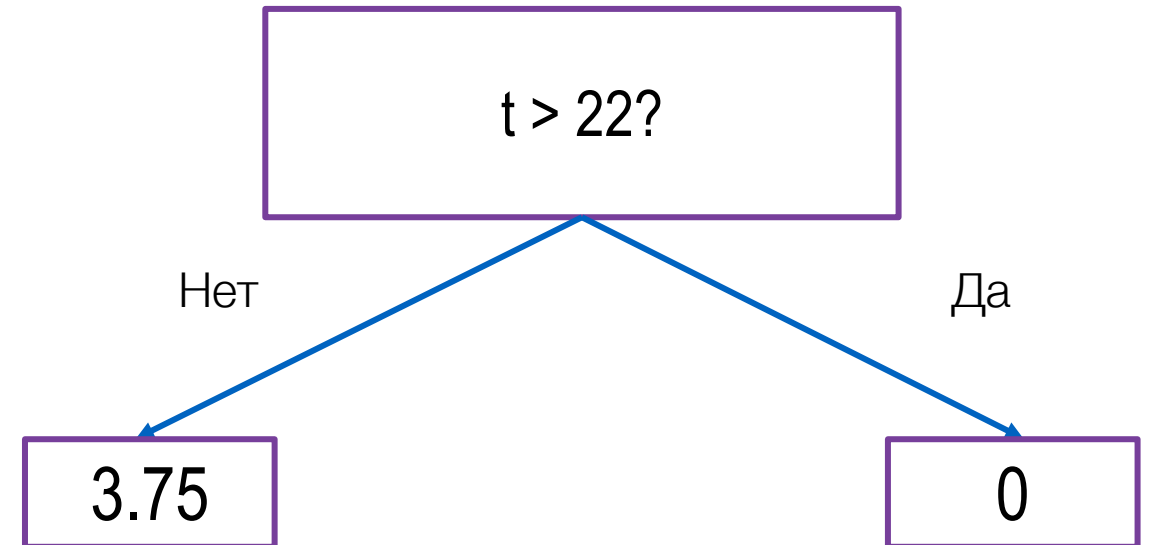
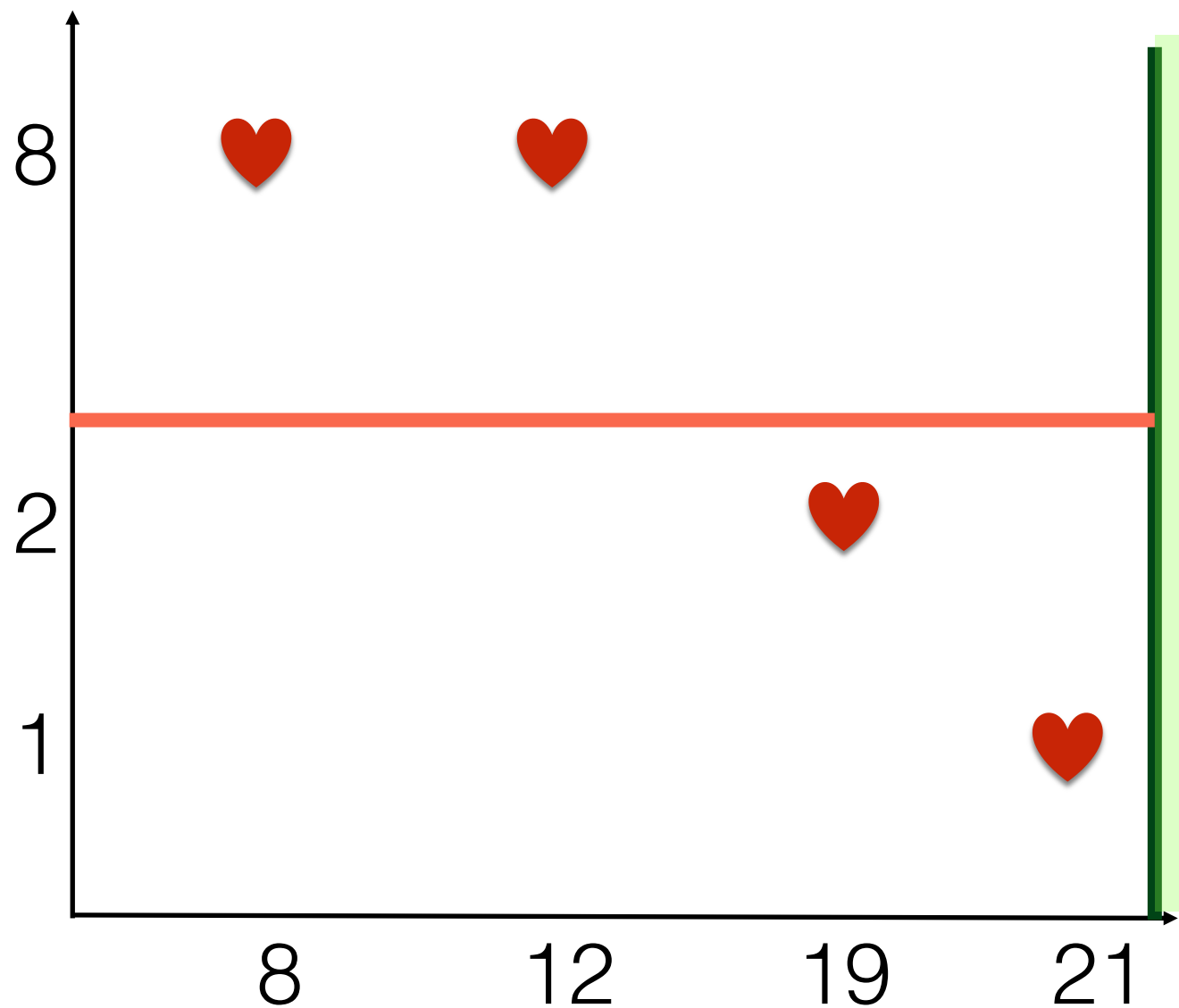
Теперь считаем ошибку прогноза.

И запоминаем эту цифру!

$$\text{MSE} = 6.25$$

t	y	\hat{y}	$y - \hat{y}$	$(y - \hat{y})^2$
21	1	1	1	1
19	2	6	-4	16
12	8	6	2	4
8	8	6	2	4

НАША ЗАДАЧКА



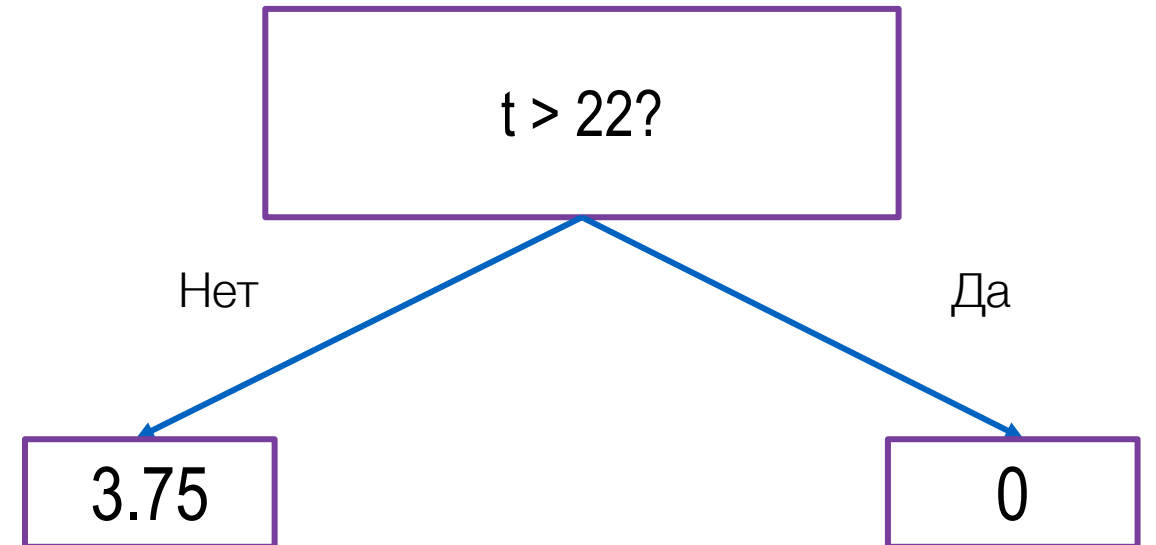
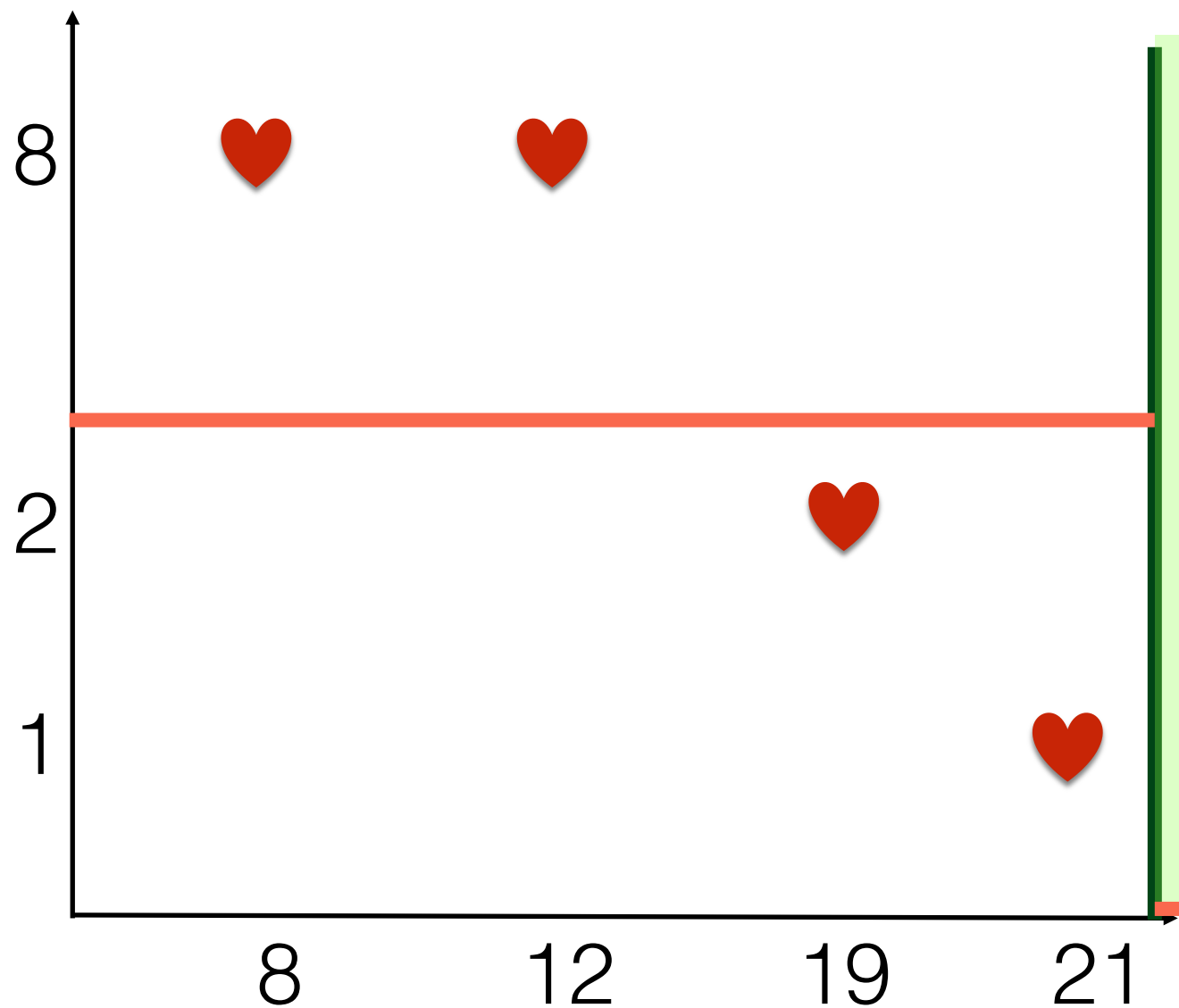
В ветку «да» попало 0 наблюдений.

В ветку «нет» попали четыре. Считаем прогноз

t	y
21	1
19	2
12	8
8	8

$$\hat{y} = \frac{1}{4} (1 + 2 + 8 + 8) = 3.75$$

НАША ЗАДАЧКА



Теперь считаем ошибку прогноза.

И запоминаем эту цифру!

t	y	\hat{y}	$y - \hat{y}$	$(y - \hat{y})^2$
21	1	3.75	-2.75	7,5625
19	2	3.75	-1.75	3,0625
12	8	3.75	4.25	18,0625
8	8	3.75	4.25	18,0625

$$\text{MSE} = \frac{1}{4} (7.5625 + 3.0625 + 18.0625 + 18.0625) = 11.6875$$

РЕЙТИНГ ПОРОГОВ

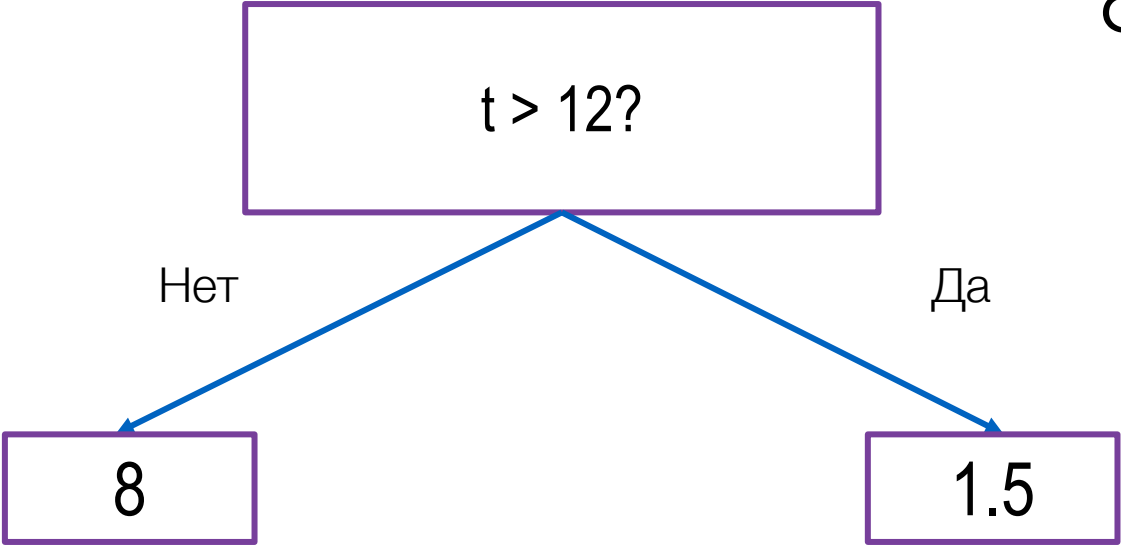
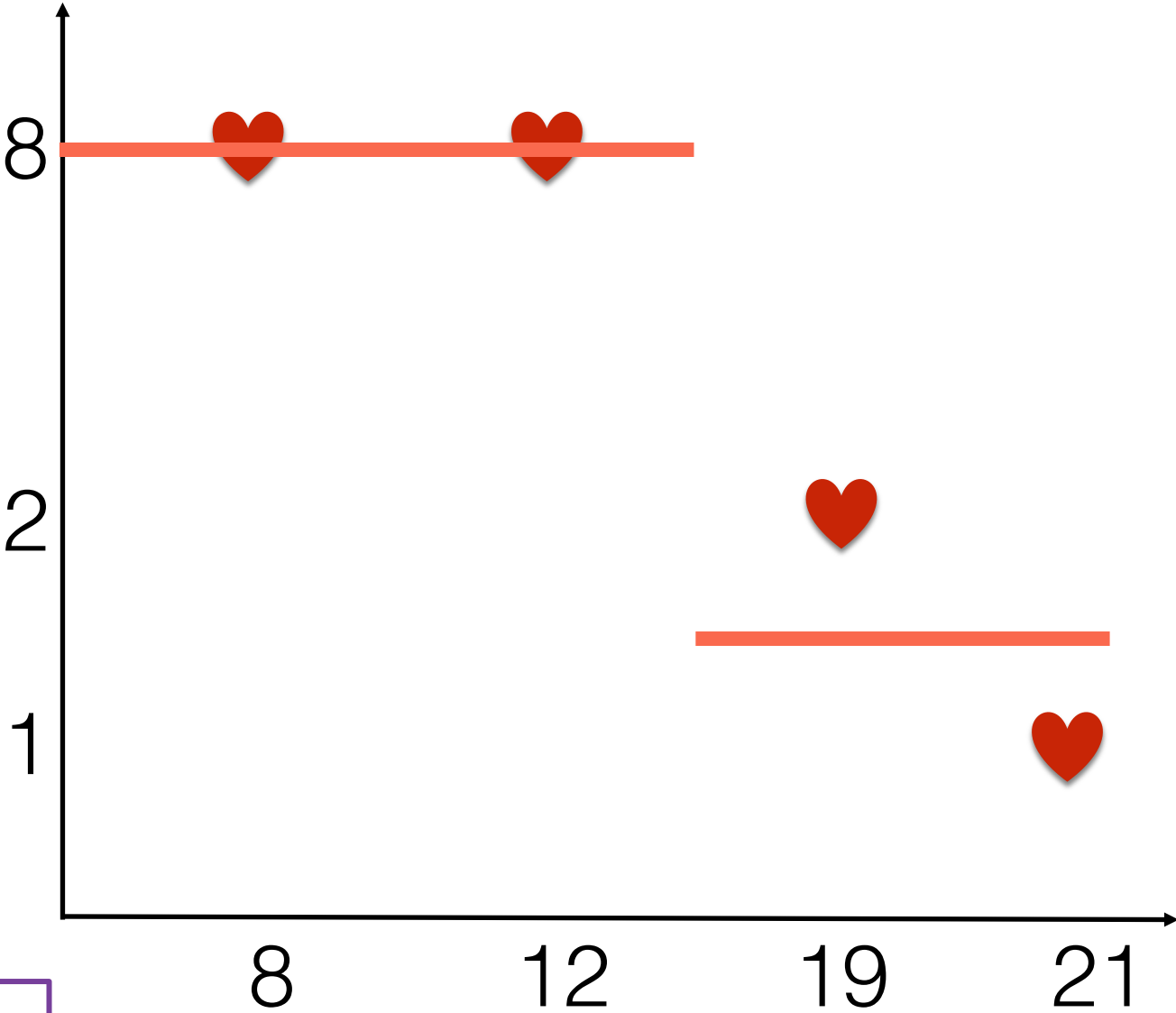
порог	MSE
7	11.6875
9	7.17
12	0.125
20	6.25
22	11.6875

РЕЙТИНГ ПОРОГОВ

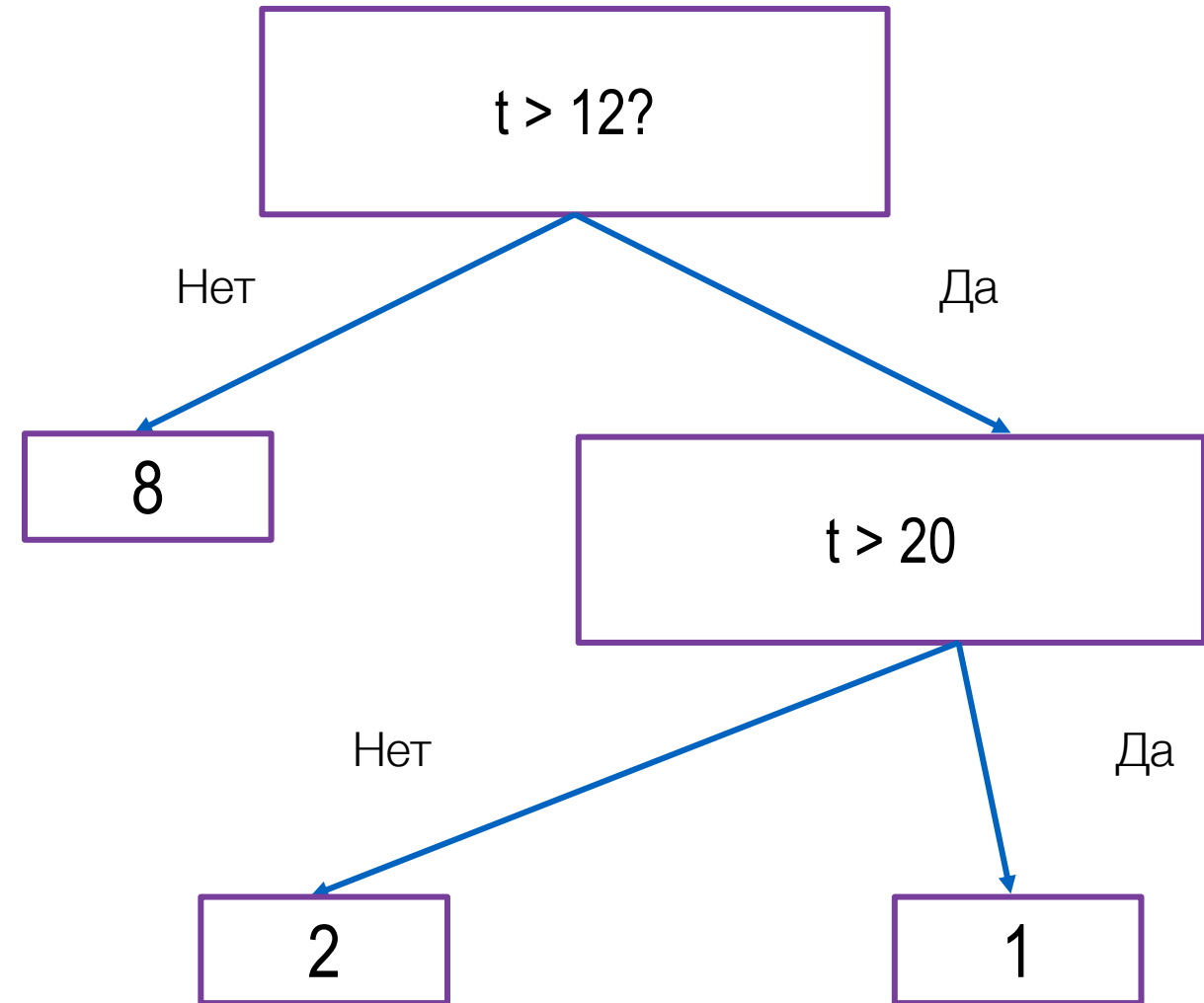
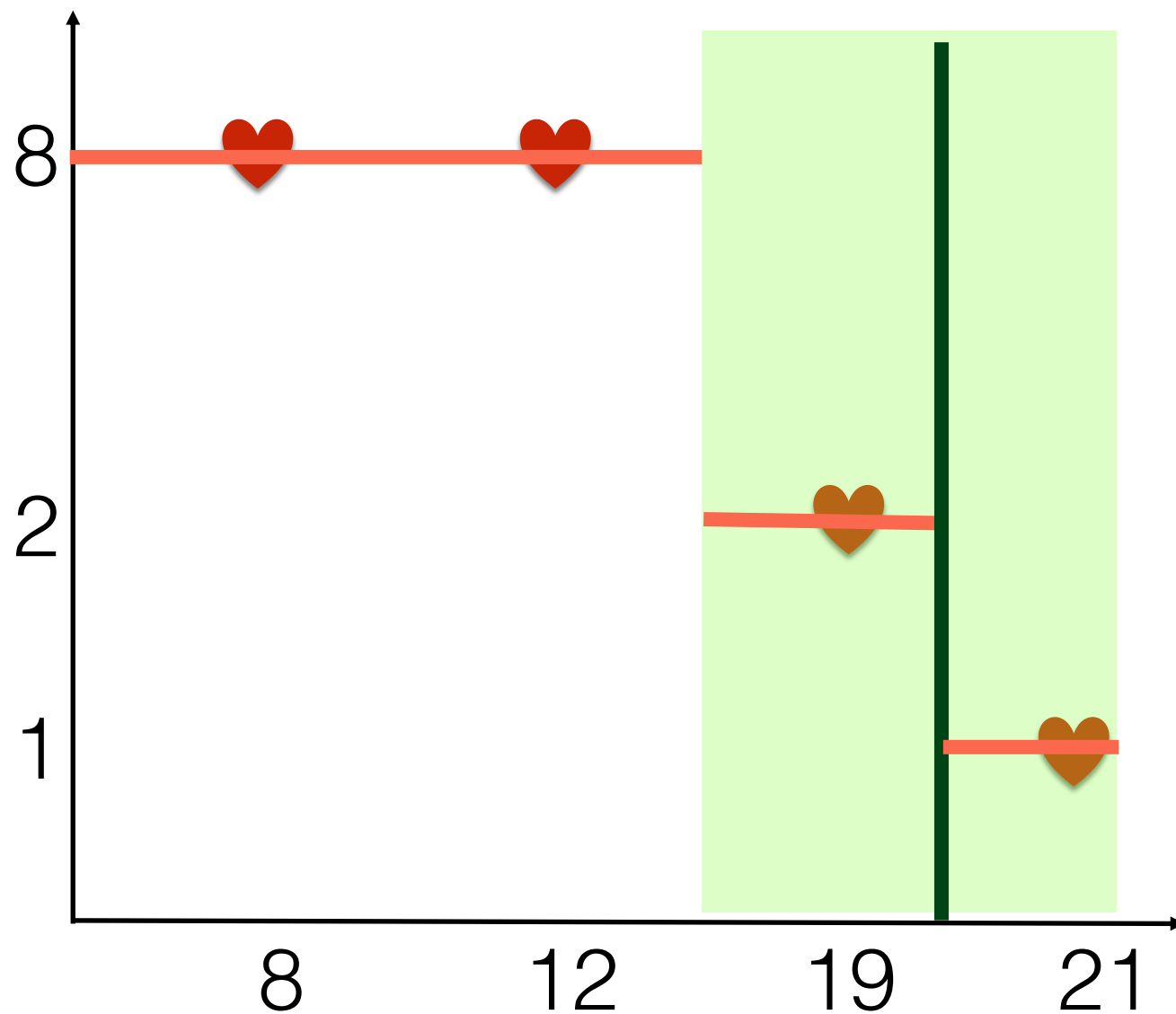
порог	MSE
7	11.6875
9	7.17
12	0.125
20	6.25
22	11.6875

РЕЙТИНГ ПОРОГОВ

порог	MSE
7	11.6875
9	7.17
12	0.125
20	6.25
22	11.6875



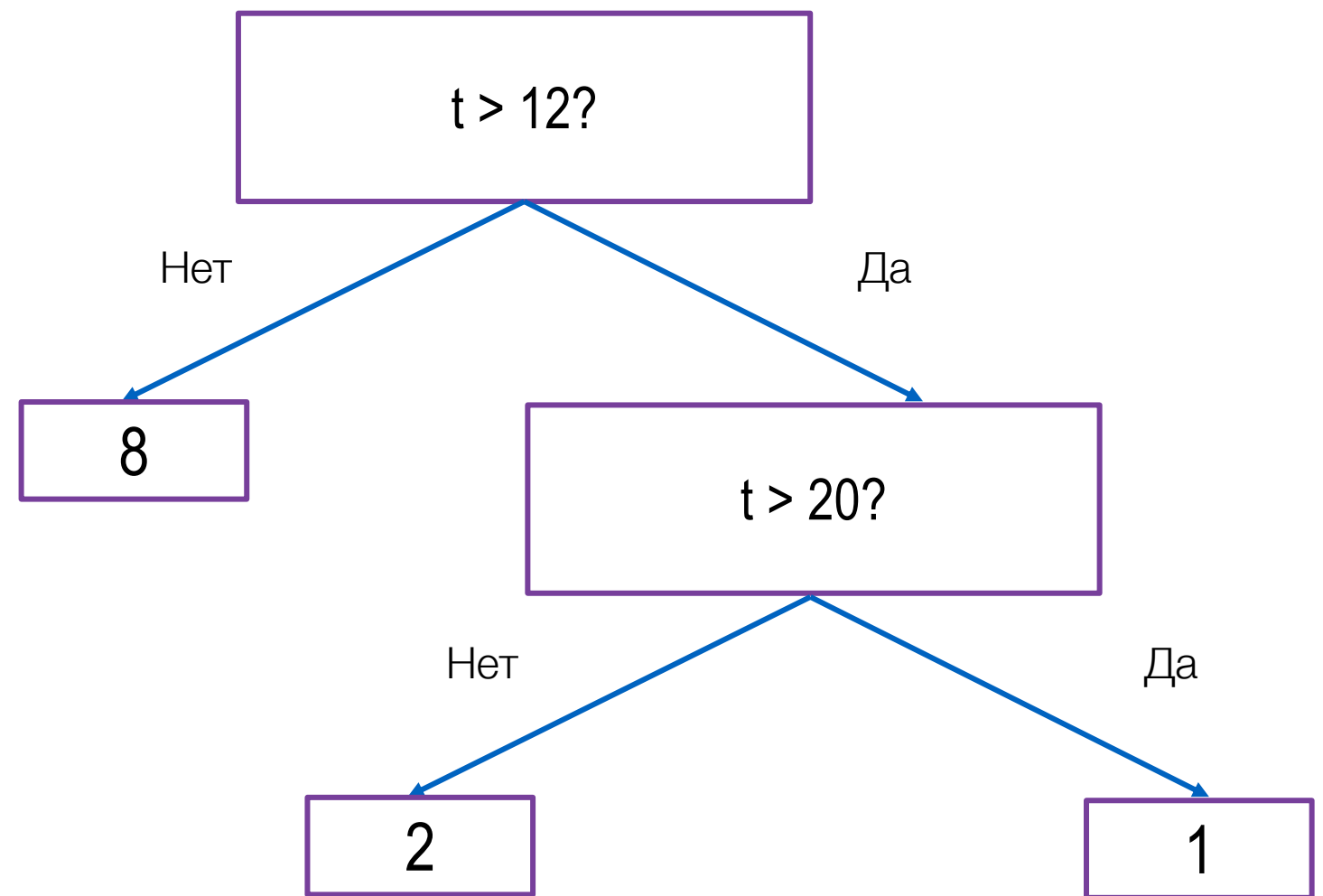
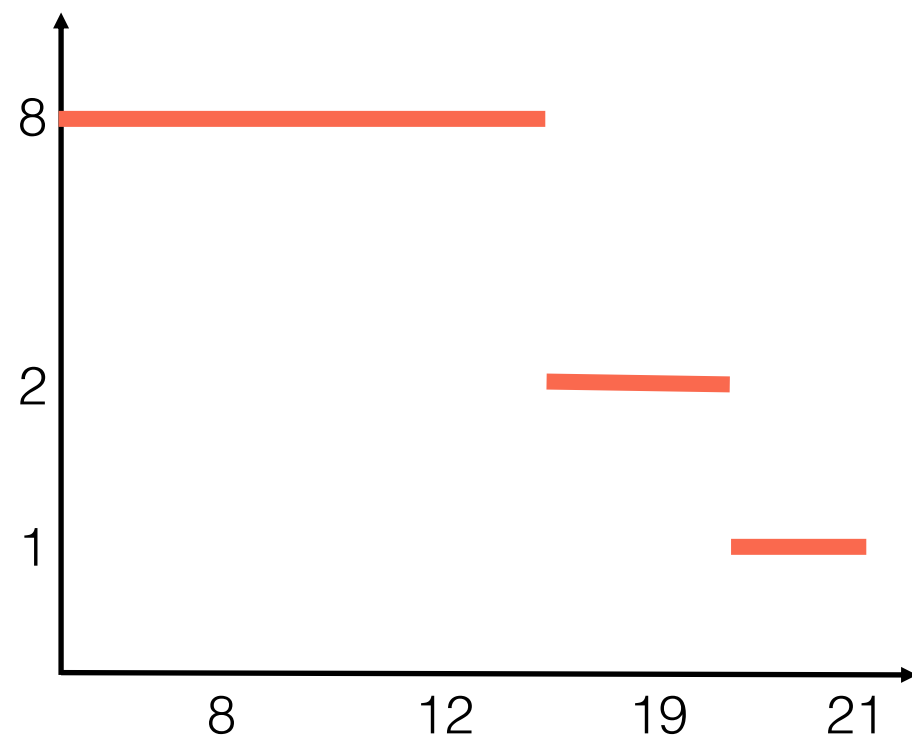
ЕЩЕ ОДИН КРУГ



$$\text{MSE} = 0$$

t	y	\hat{y}	$y - \hat{y}$	$(y - \hat{y})^2$
21	1	1	0	0
19	2	2	0	0
12	8	8	0	0
8	8	8	0	0

ИТОГОВОЕ ДЕРЕВО





НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
УНИВЕРСИТЕТ