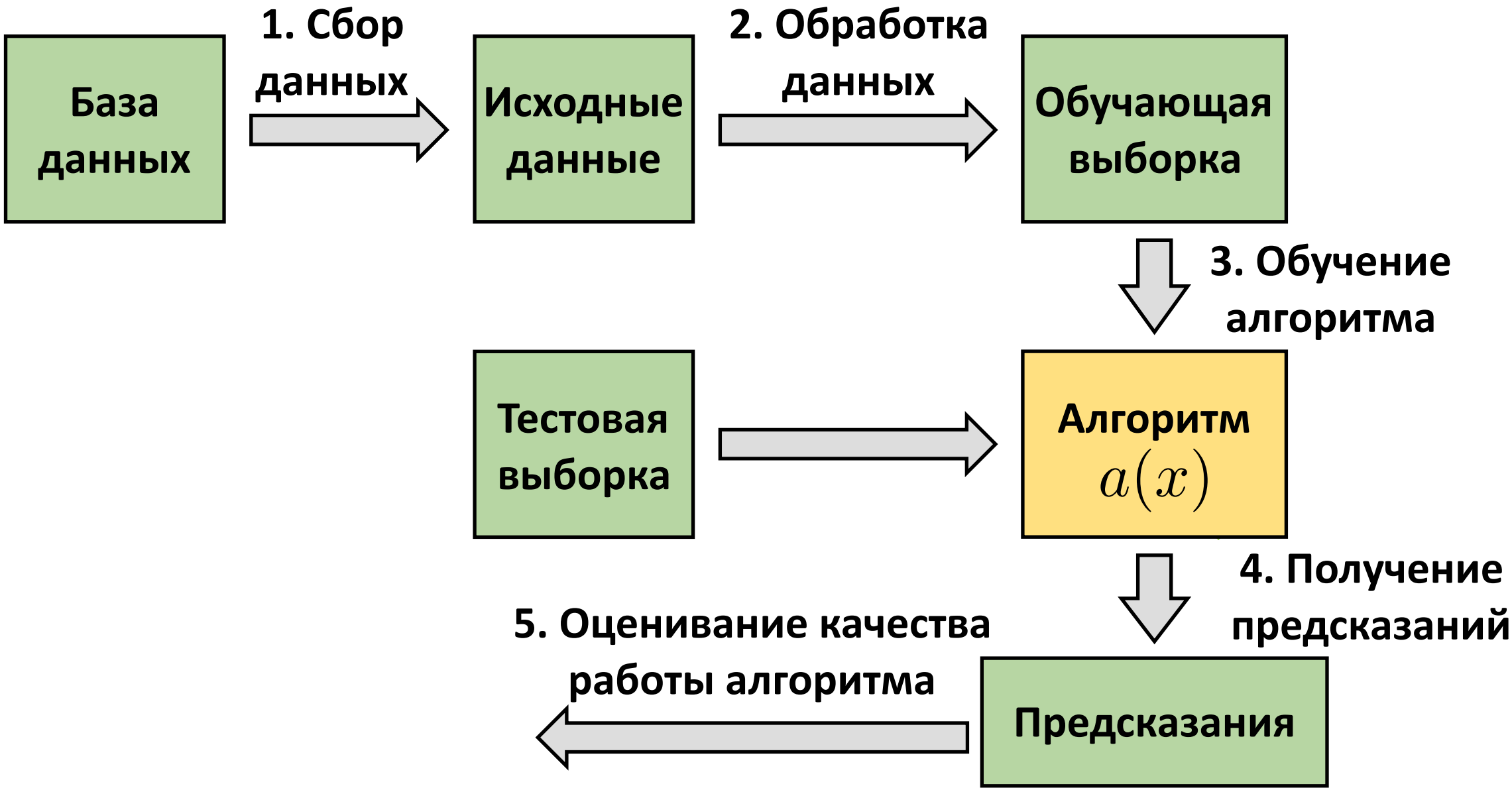
Машинное обучение: напоминание

Схема работы машинного обучения



Напоминание: обучающая выборка

Обучающая выборка

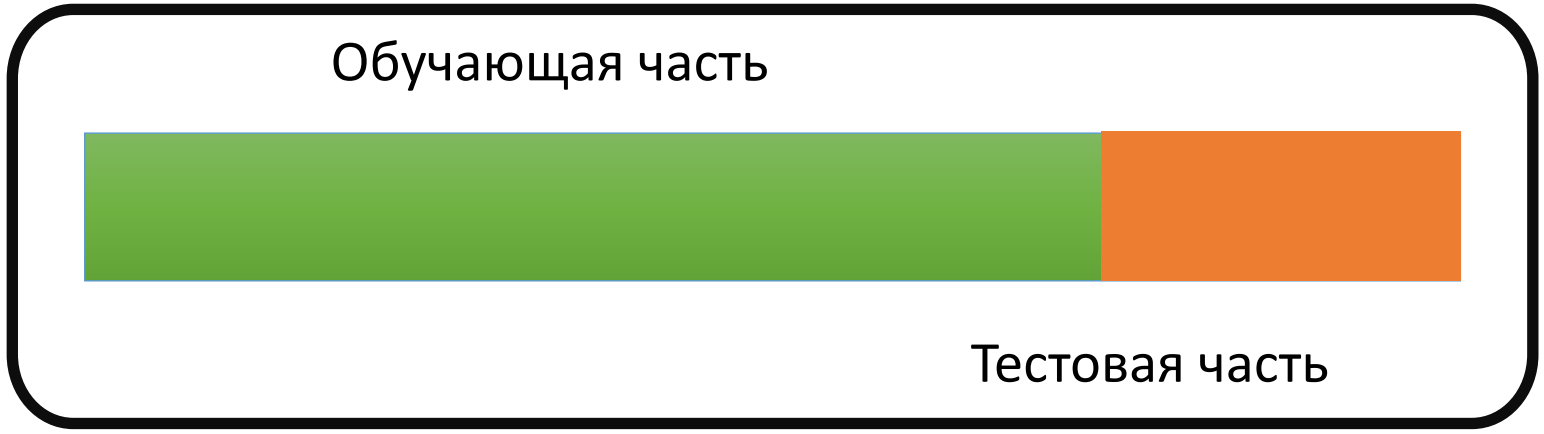
объект x_i

Площадь	Год постройки	Число комнат	Цена
45	1995	1	7000000
60	2005	2	9900000
35	2010	1	5500000

ответ y_i

Обобщающая способность алгоритма

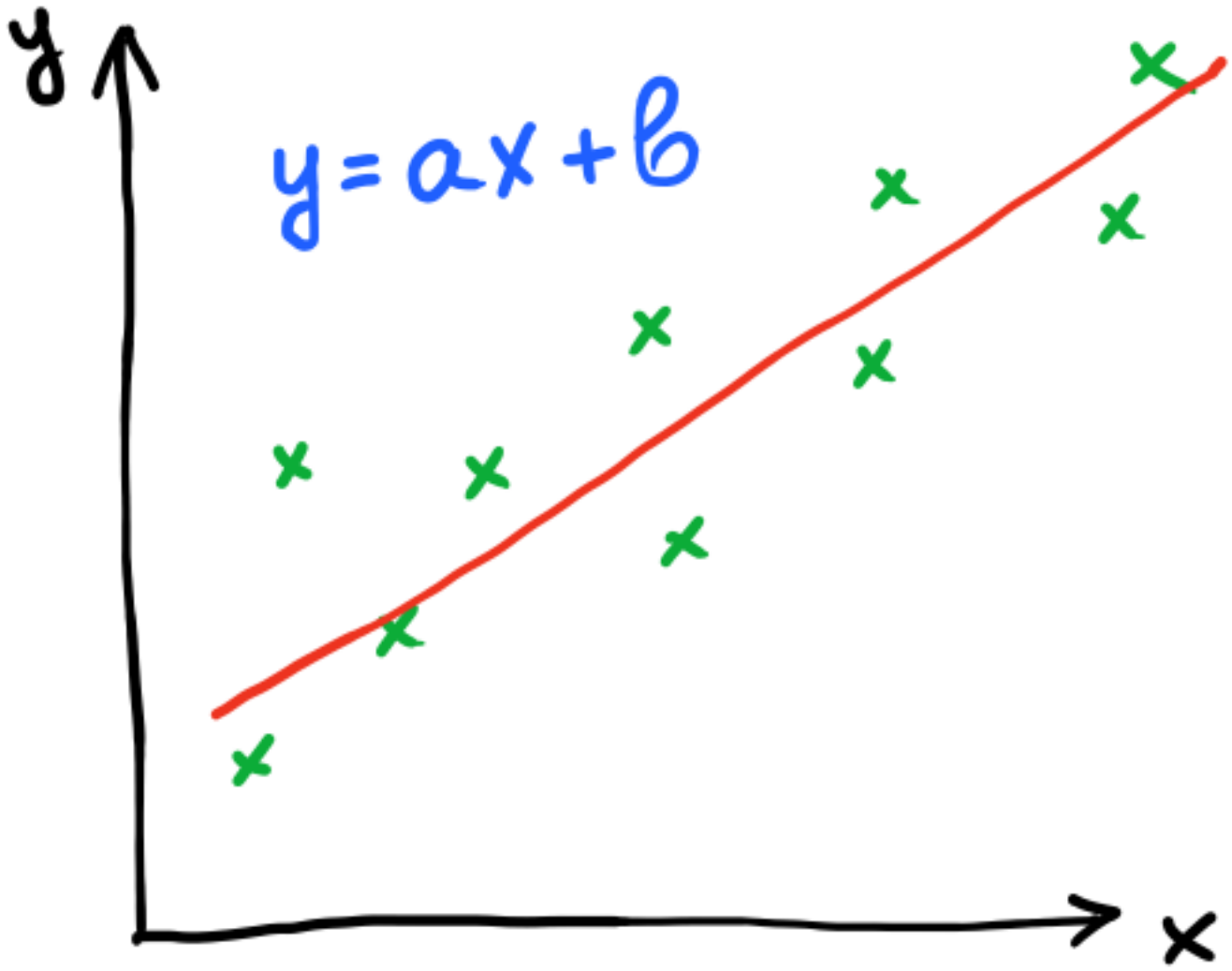
- Обучаем алгоритм на **обучающих данных**, измеряем качество на **тесте**:



Все прецеденты

Линейная модель для регрессии

x	y
1	2
3	5
-1	-2
5	?



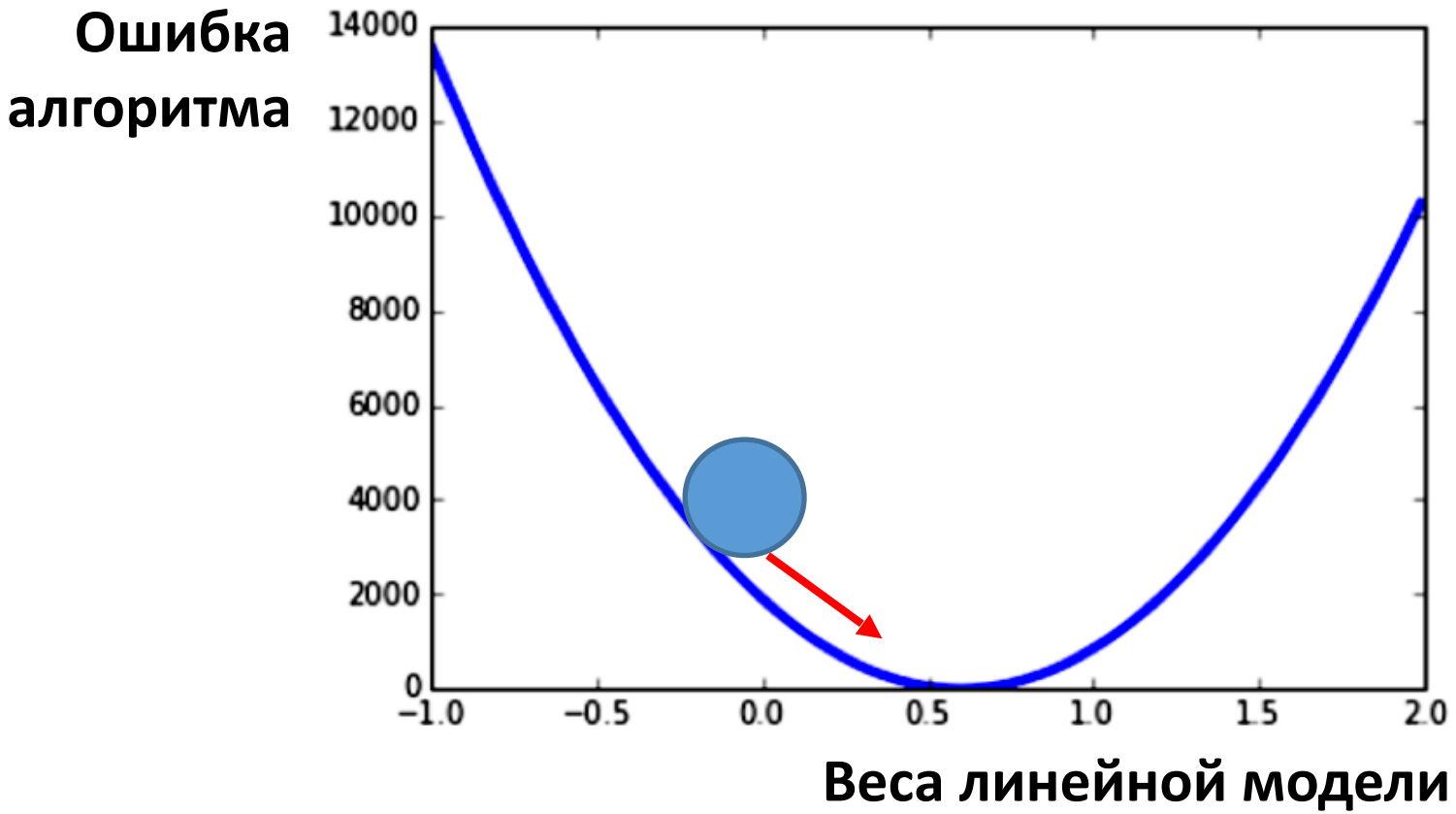
Линейные модели для задачи регрессии

Линейная модель суммирует значения всех признаков с некоторыми весами
Веса при признаках — параметры, которые необходимо настраивать в процессе обучения

$$a(x) = w_0 + w_1 x_1 + \ldots w_d x_d$$

d — число признаков

Обучение линейной модели



Метод k ближайших соседей

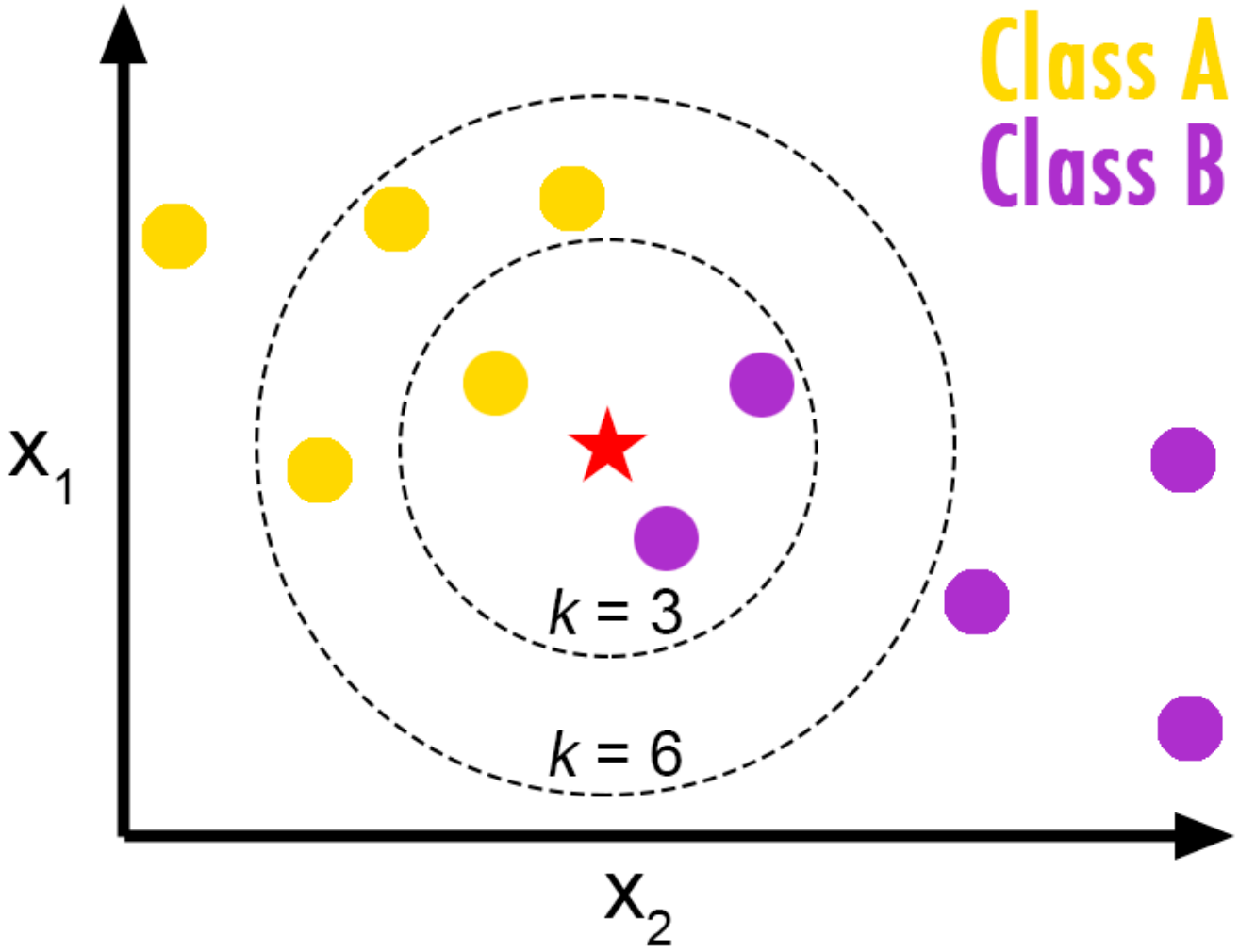
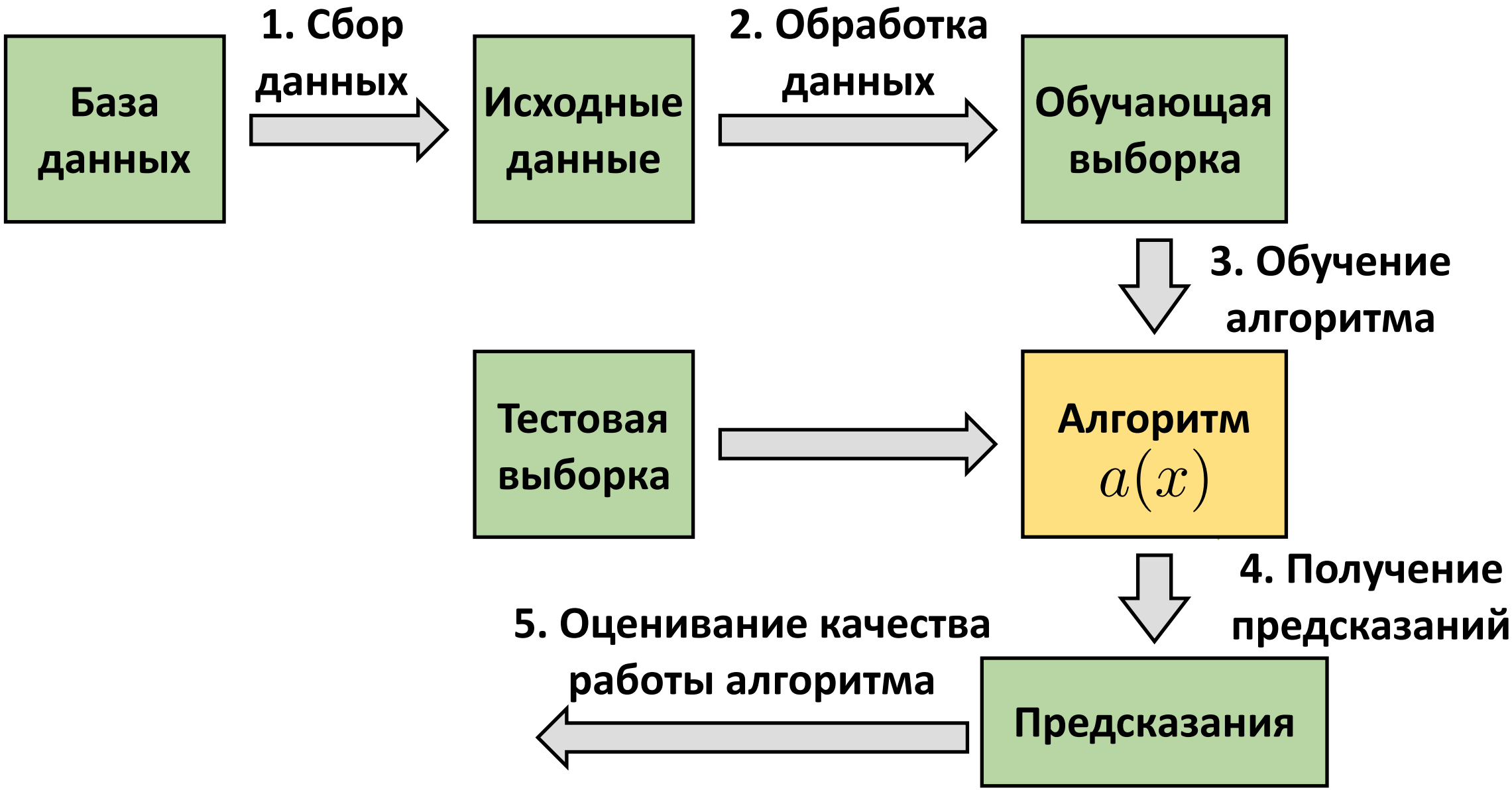


Схема работы машинного обучения



Измерение качества в регрессии

a(x)	y	отклонение?
11	10	?
9	10	?
20	10	?
1	10	?

Измерение качества в регрессии

a(x)	y	отклонение
11	10	1
9	10	-1
20	10	10
1	10	-9

Измерение качества в регрессии

a(x)	y	**|a(x) - y|**
11	10	**1**
9	10	**1**
20	10	**10**
1	10	**9**

Измерение качества в регрессии

Среднее абсолютное отклонение, или MAE (Mean Absolute Error)

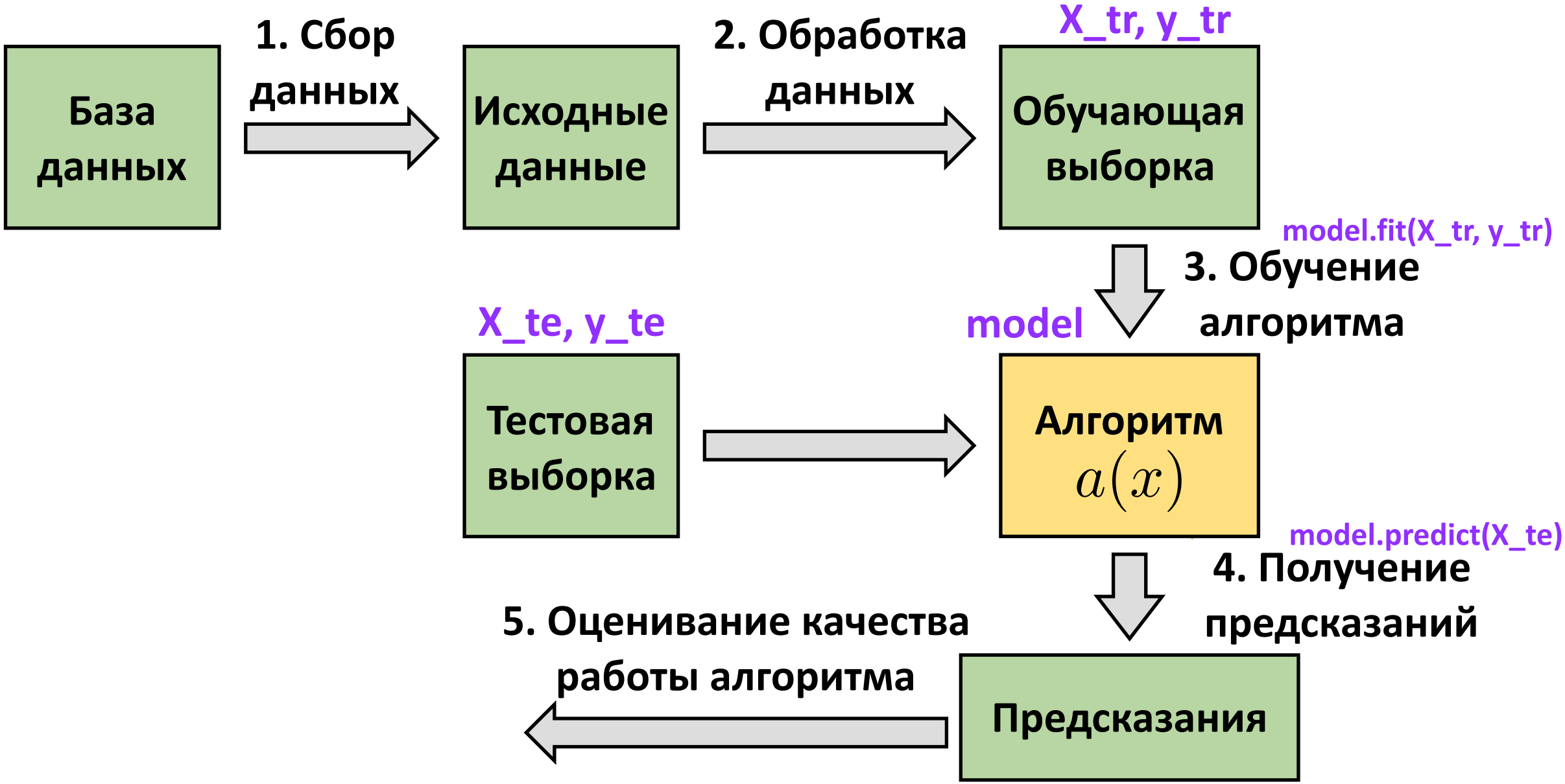
$$MSE = \frac{1}{\ell}\sum_{i=1}^{\ell}|a(x_i) - y_i|$$

ℓ — число прецедентов в выборке

В примере:

$$MAE = \frac{1}{4}(1 + 1 + 10 + 9) = 5.25$$

Схема работы машинного обучения



Типы столбцов

- **Числовые (numerical)**: в них хранятся целые числа (integer) или числа с плавающей точкой (real).

ФИО	Город	Возраст	Пол
Иванова Иванка Ивановна	Москва	32	Ж
Петров Пётр Петрович	Москва	18	М
Сидоров Сидр Сидорович	Пермь	54	М

Типы столбцов

- **Категориальные (nominal)**: здесь хранится какое-то значение из справочника.

ФИО	Город	Возраст	Пол
Иванова Иванка Ивановна	Москва	32	Ж
Петров Пётр Петрович	Москва	18	М
Сидоров Сидр Сидорович	Пермь	54	М
	(poly-)		(bi-)

Анализируют по-разному

- **Числовые** столбцы можно умножать на коэффициенты, а **категориальные** нельзя
  - Москва x 5 = ?
  - (номер Москвы в справочнике) x 5 = ?

Анализируют по-разному

- **Числовые** столбцы можно умножать на коэффициенты, а **категориальные** нельзя
  - Москва x 5 = ?
  - (номер Москвы в справочнике) x 5 = ?
- Что делать с **категориальными** столбцами?
  - можно считать средние значения по категориям (средний возраст в городе)
  - можно бинаризовать (“город = Москва?”, “город = Казань?” …)

Измерение качества в классификации

a(x)	y	отклонение?
яблоко	яблоко	?
апельсин	яблоко	?
яблоко	яблоко	?
апельсин	апельсин	?

Измерение качества в классификации

a(x)	y	**совпадение**
яблоко	яблоко	**да**
апельсин	яблоко	**нет**
яблоко	яблоко	**да**
апельсин	апельсин	**да**

Измерение качества в классификации

Доля правильных ответов (accuracy):

$$Accuracy = \frac{1}{\ell}\sum_{i=1}^{\ell}[a(x_i) = y_i]$$

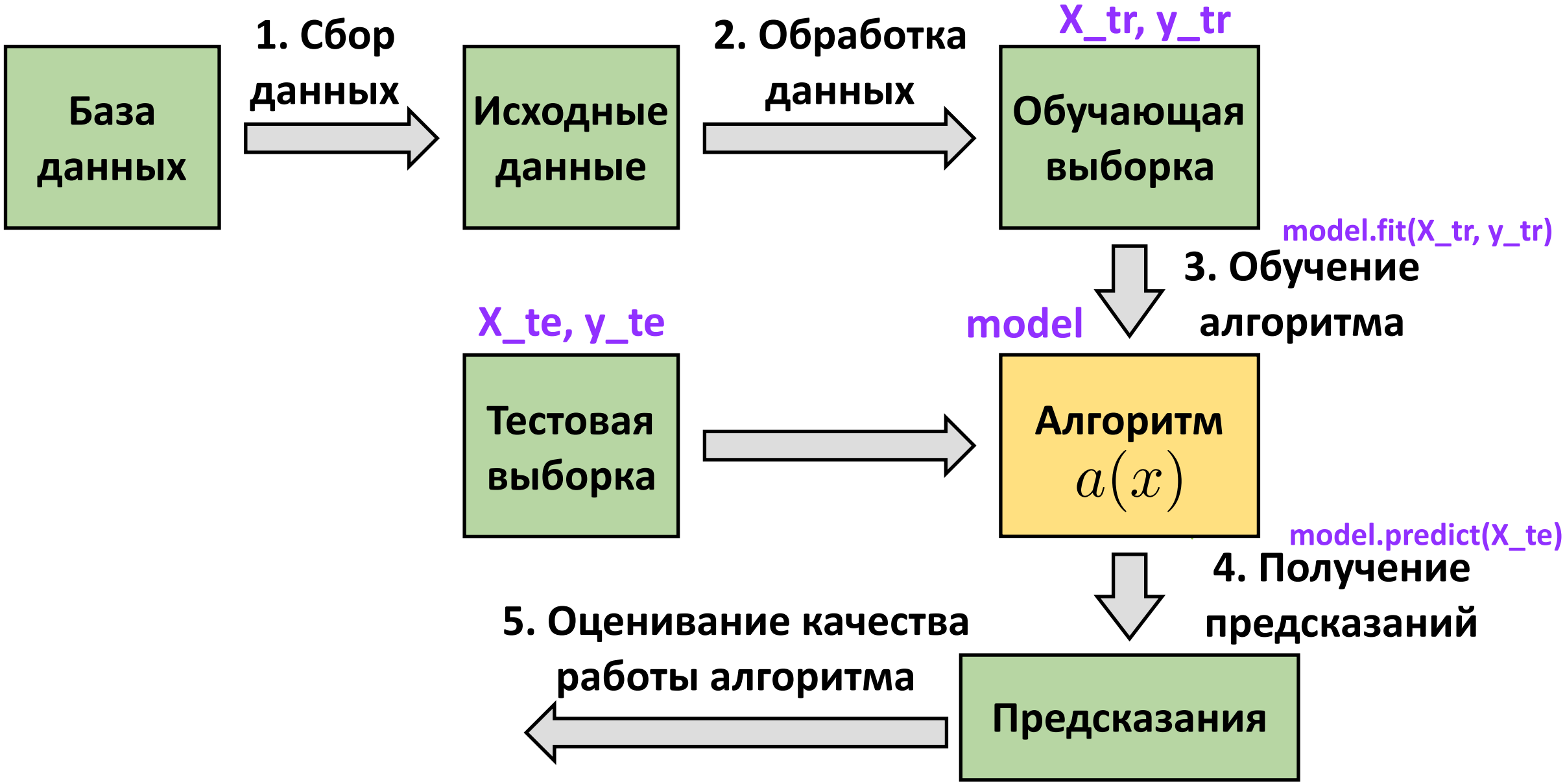
ℓ — число прецедентов в выборке

* выражение [x] равно единице, если x является верным утверждением, и нулю иначе

В примере:

$$Accuracy = \frac{1}{4}(1 + 0 + 1 + 1) = 0.75$$

Схема работы машинного обучения



Типы столбцов

- **Текстовые (text)**: текст может быть из любой последовательности символов, нельзя представить в виде значения из справочника.

ФИО	Город	Возраст	Пол
Иванова Иванка Ивановна	Москва	32	Ж
Петров Пётр Петрович	Москва	18	М
Сидоров Сидр Сидорович	Пермь	54	М

Тексты: мешок слов

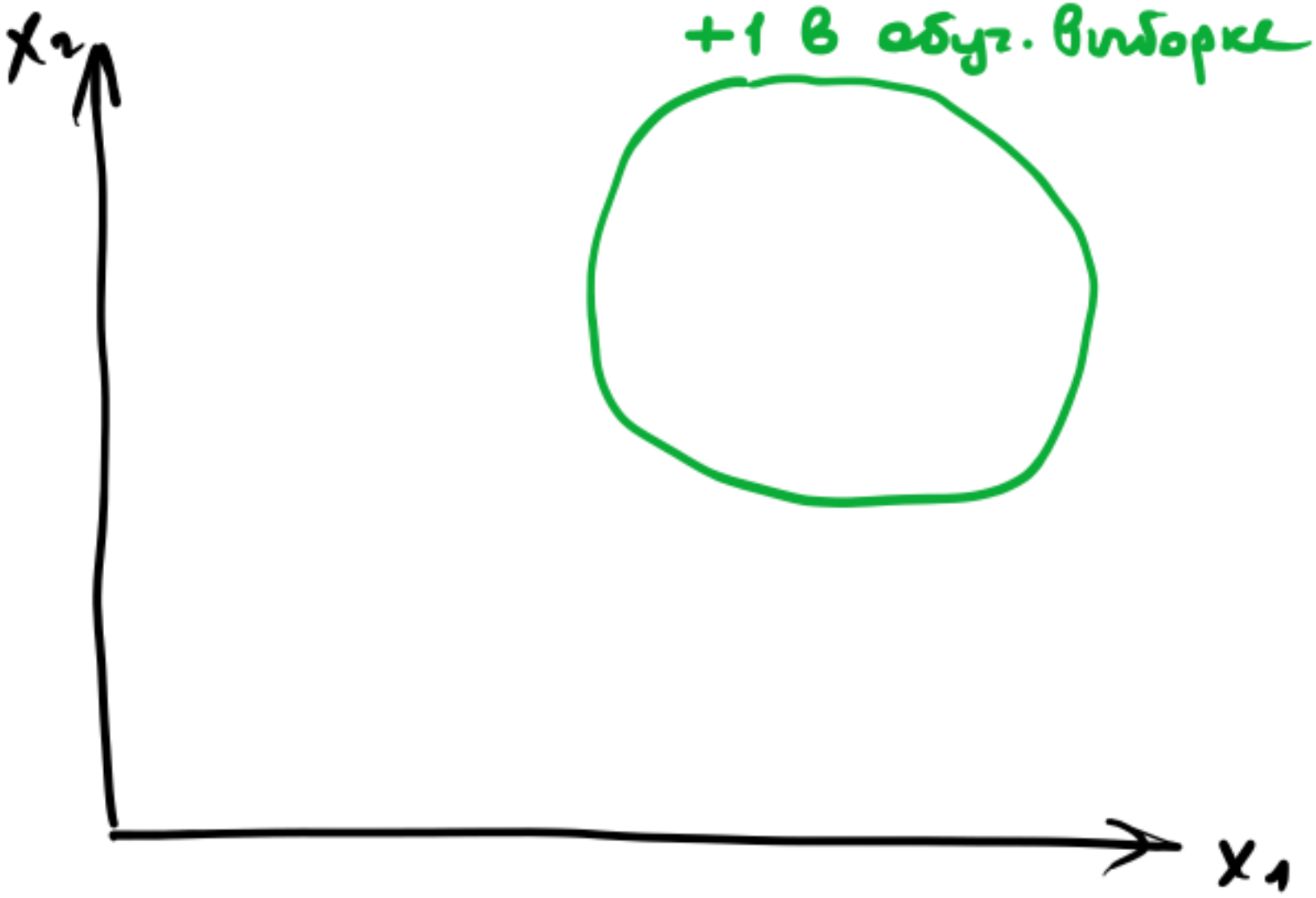
хватит денег
денег не хватит
много денег

хватит	денег	не	много
1	1	0	0
1	1	1	0
0	1	0	1

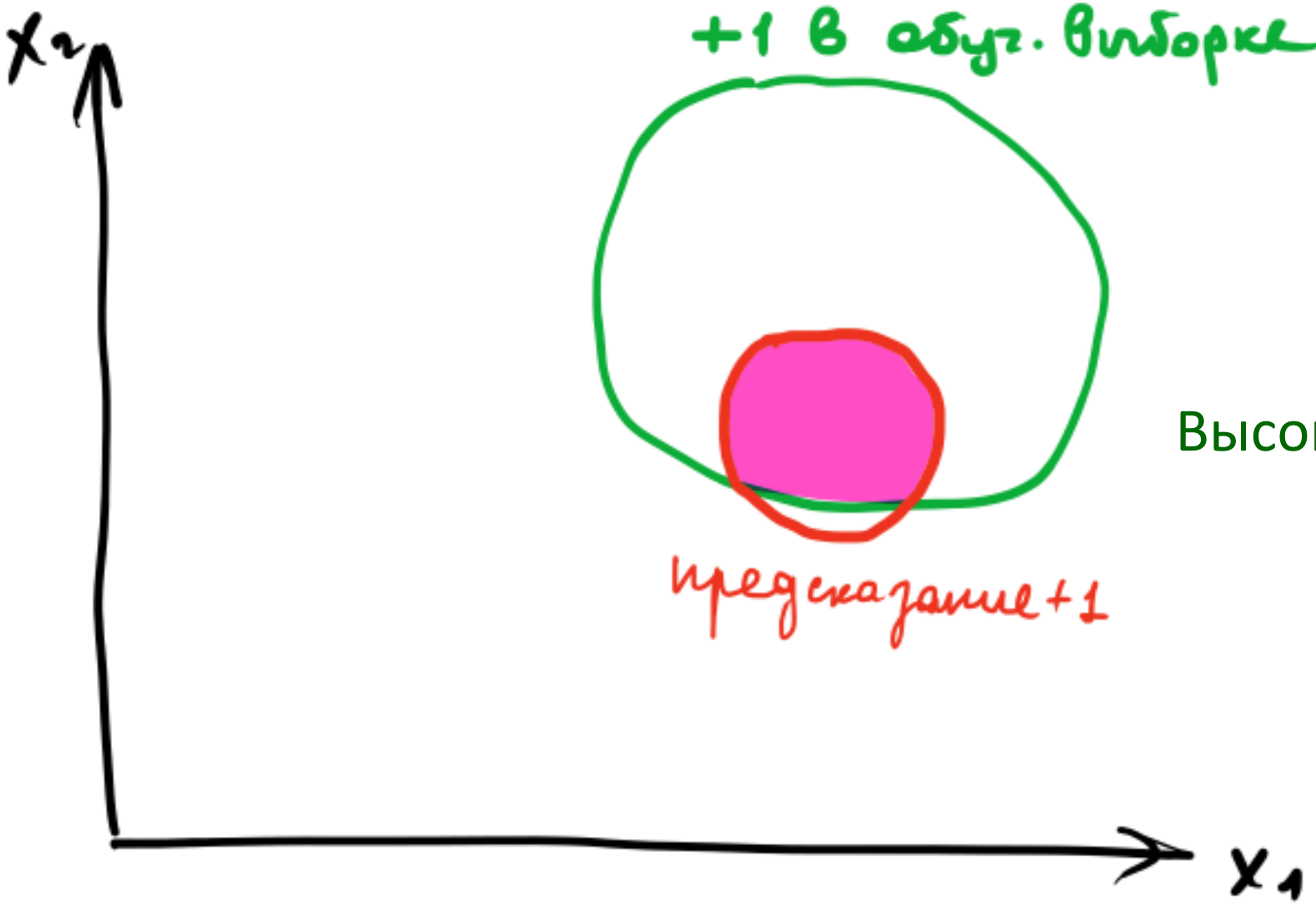
Тексты

Числа (модель *мешка слов*)

Точность и полнота (precision, recall)

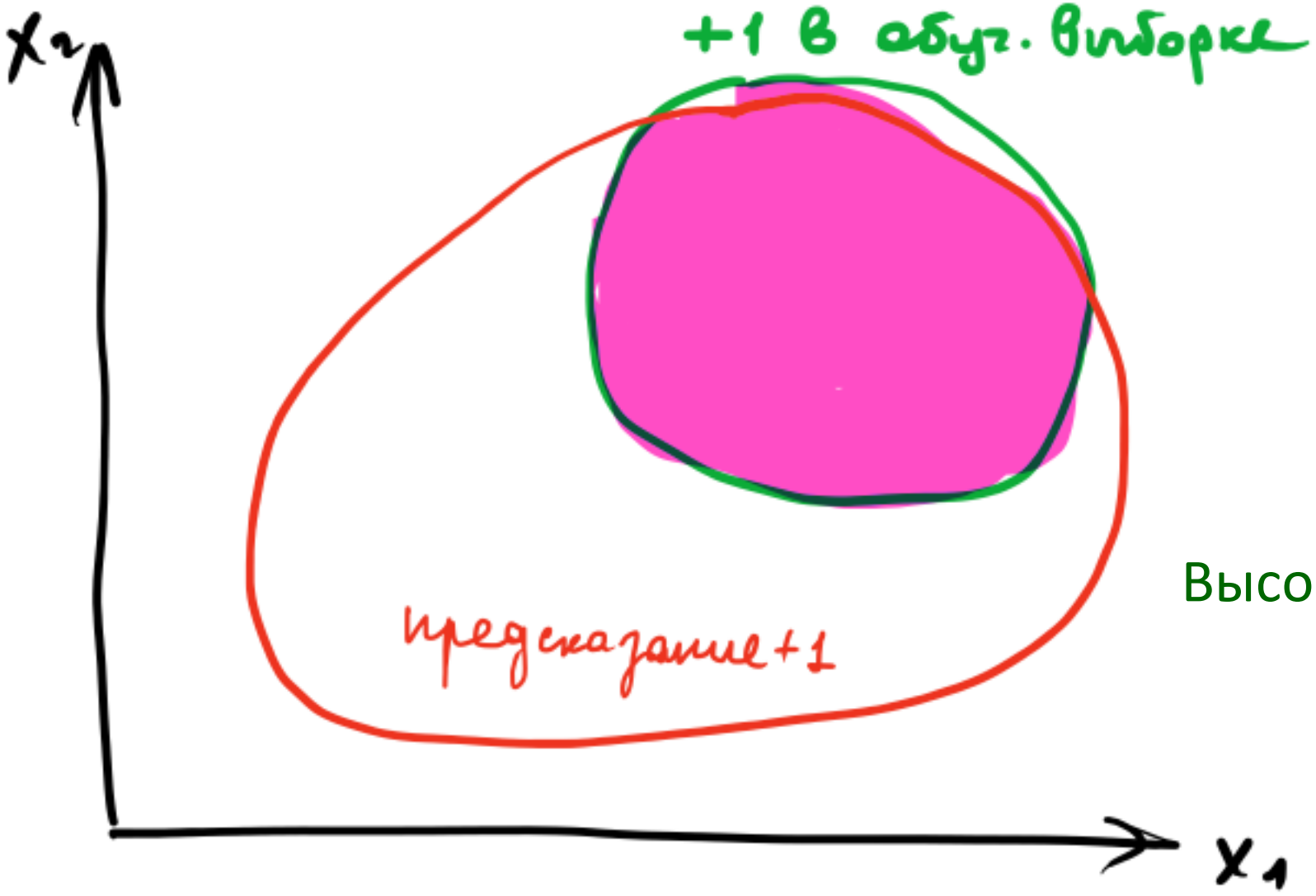


Точность и полнота (precision, recall)



Высокая точность!

Точность и полнота (precision, recall)



Высокая полнота!