

Основы искусственного интеллекта

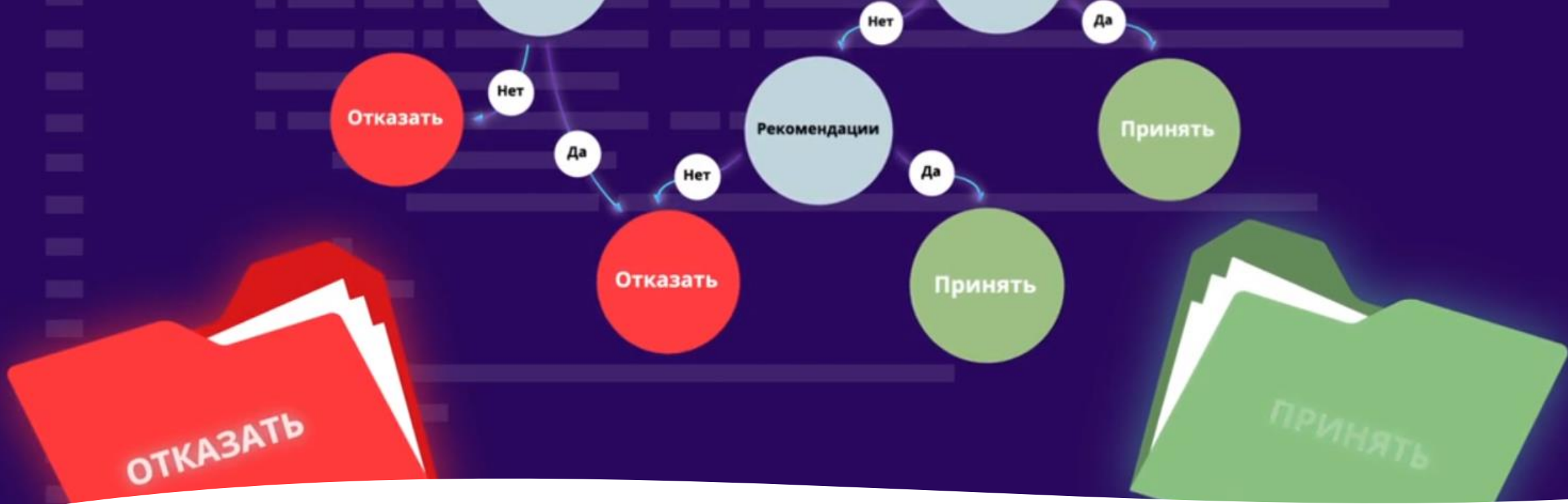
# Лекция 5. Задачи классификации и регрессии

2023/2024 учебный год

Доцент кафедры ИВЭ, Махно В.В.

©Создано при помощи <https://sberuniversity.ru/>





## Задача классификации.

Пример с вакансиями.

Необходимо построить алгоритм, который позволит системе определить, есть ли в резюме кандидата необходимые параметры. Если есть – отправить в папку «Собеседование», если нет – в папку «Отказать».

# Ответы алгоритма и реальные данные

TRUE

FALSE

NEGATIVE-

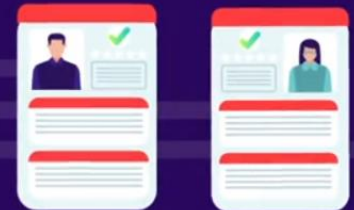
POSITIVE +

# Ответы алгоритма и реальные данные

True positive



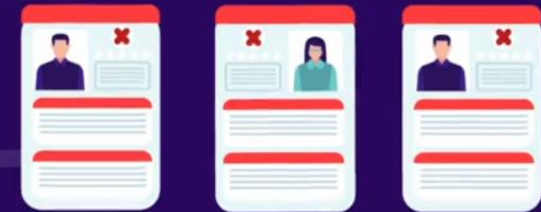
False positive



False negative



True negative



# Ответы алгоритма и реальные данные

---

## Результат классификации

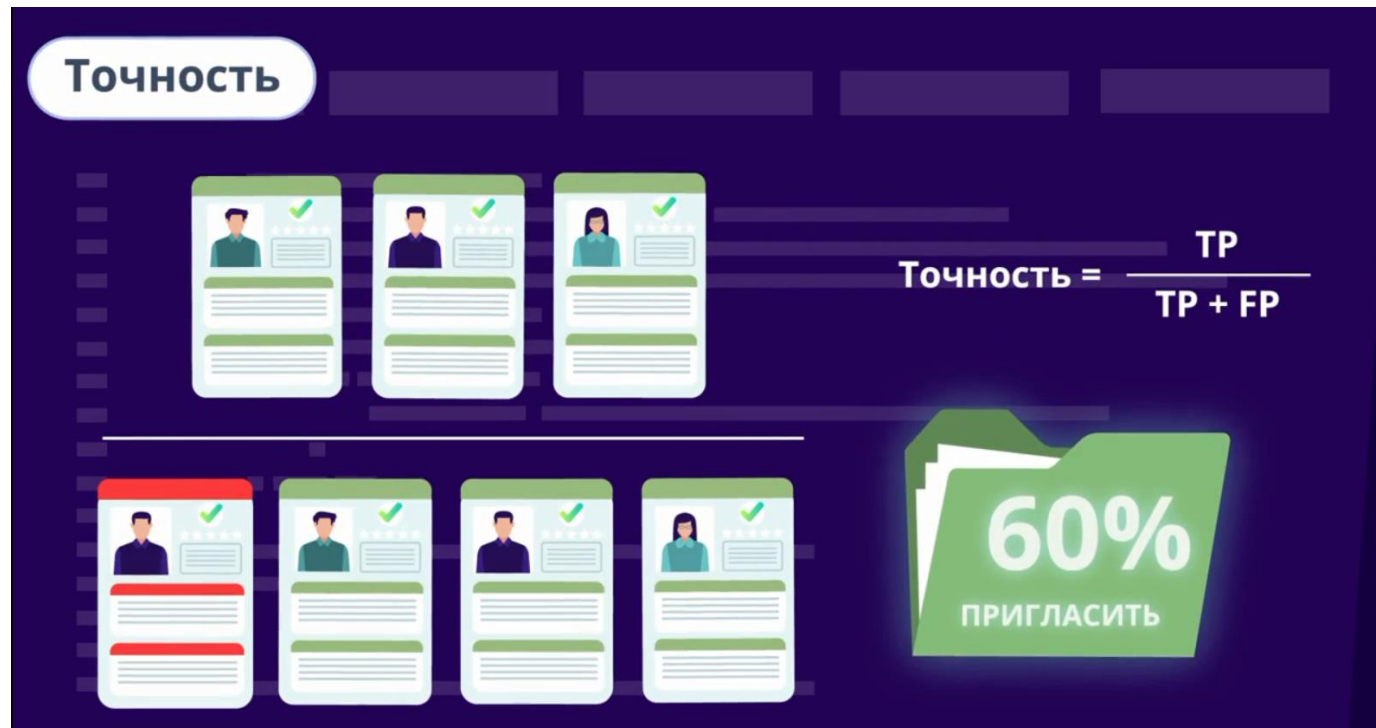
- TP — true positive, алгоритм верно пометил резюме как подходящее
- TN — true negative, алгоритм верно отнес резюме к неподходящим
- FP — false positive, алгоритм ошибочно считает подходящим резюме, в котором нет нужных качеств
- FN — false negative, алгоритм ошибочно отбраковал подходящее резюме

# Метрики

---

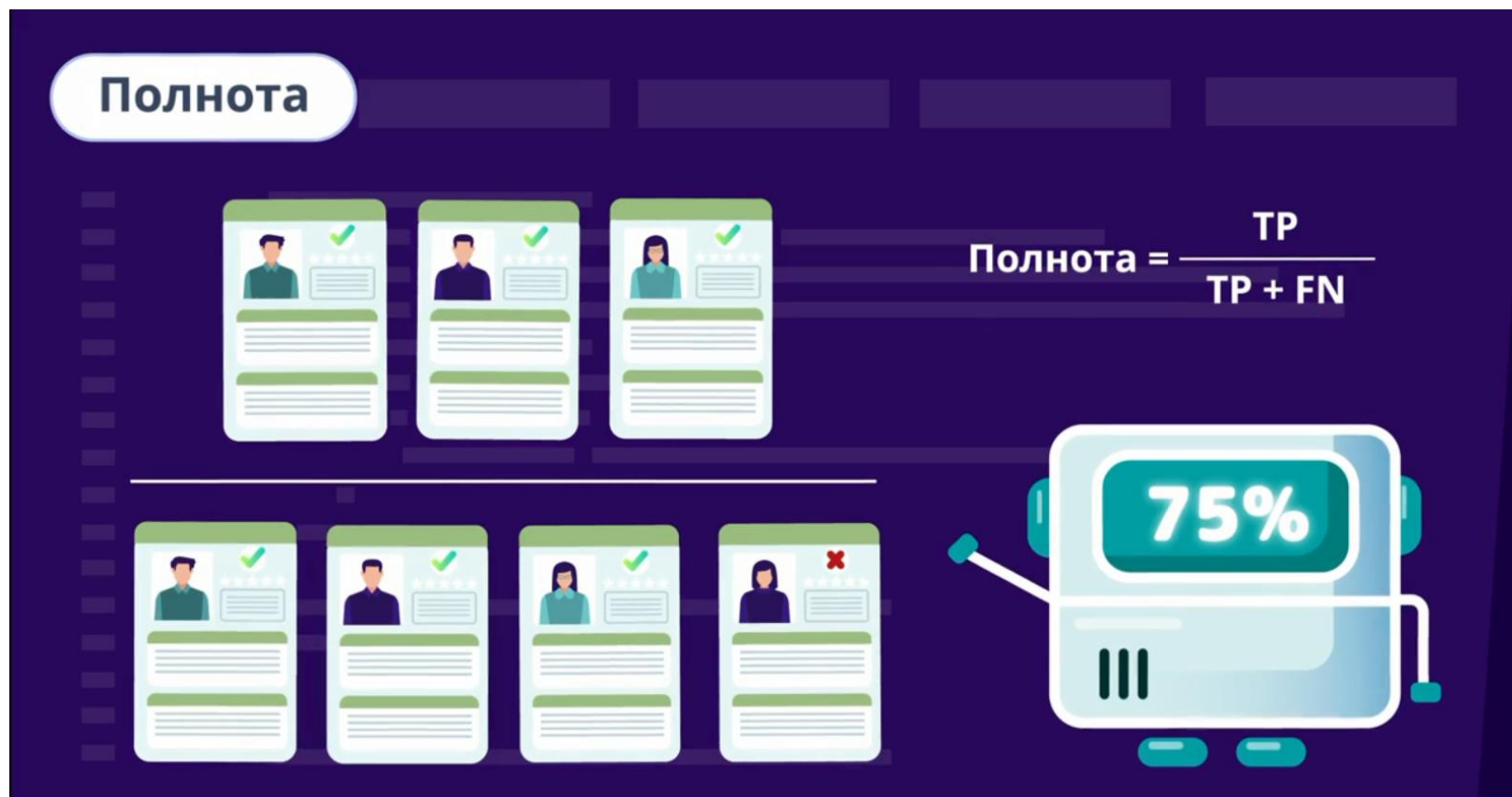
- Самая простая метрика – **доля правильных предсказаний**: сколько раз прогноз машины и разметка программиста совпали между собой.
- Другая метрика – **точность**. Она показывает отношение количества верно угаданных подходящих резюме к количеству тех, кого машина вообще отнесла к группе «собеседование».
- Кроме точности есть еще метрика **полноты**. Она показывает отношение количества верно угаданных подходящих резюме, к другому значению: количеству кандидатов, которых следовало пригласить по мнению программиста.

# Метрика точность



Точность показывает отношение количества верно угаданных подходящих резюме к количеству тех, кого машина вообще отнесла к группе «собеседование».

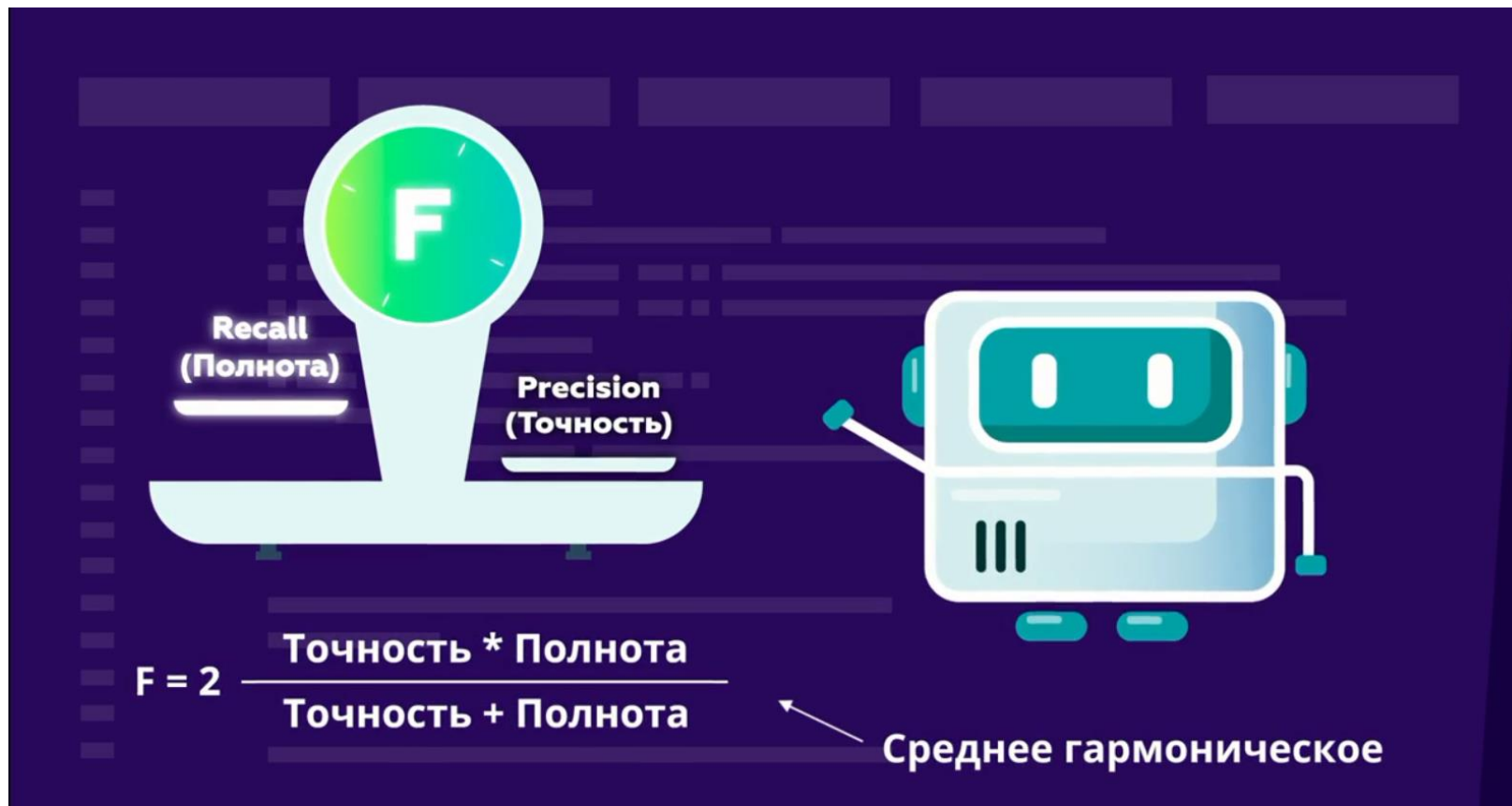
# Метрика полнота



Полнота показывает отношение количества верно угаданных подходящих резюме, к другому значению: количеству кандидатов, которых следовало пригласить по мнению программиста.



# F-метрика



# Табличные данные

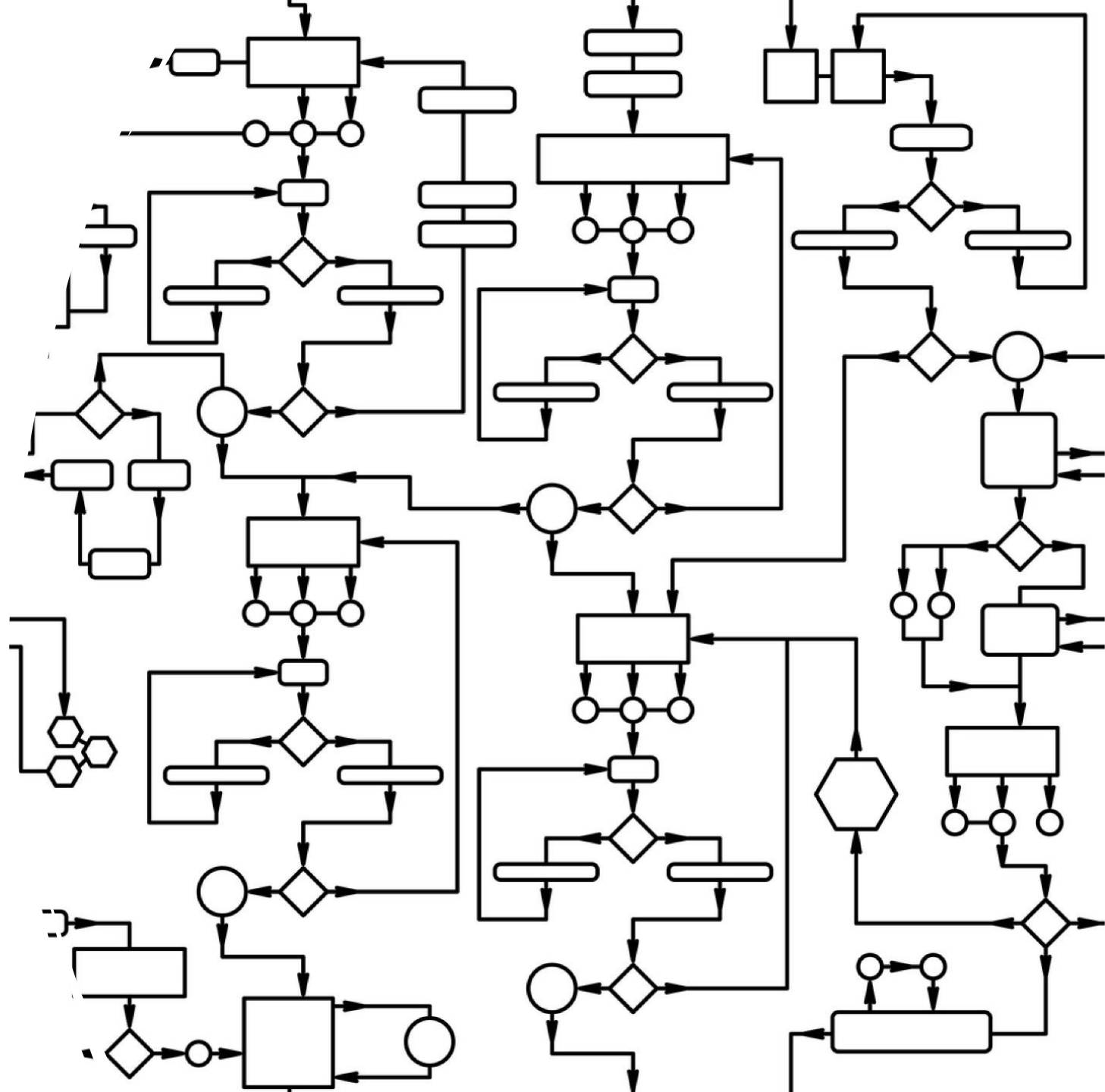
Зарботная плата	Возраст	Должность	Уровень образования	Город проживания	Стаж работы (годы)	Вернет ли клиент кредит
100000	26	Риэлтор	Высшее	Санкт- Петербург	5	Да
50000	20	Продавец- консультант	Высшее	Москва	1	Нет
35000	39	Автомеханик	Среднее специальное	Воронеж	8	Нет
25000	23	Программист	Высшее	Самара	2	Да
75000	41	Юрист	Среднее	Москва	14	Да

# Создание алгоритма классификации

---

В этой задаче каждому объекту (строке в таблице данных) соответствует класс — значение из заданного набора классов.

Задача классификации состоит в том, чтобы разработать алгоритм, который по признакам объекта будет предсказывать класс

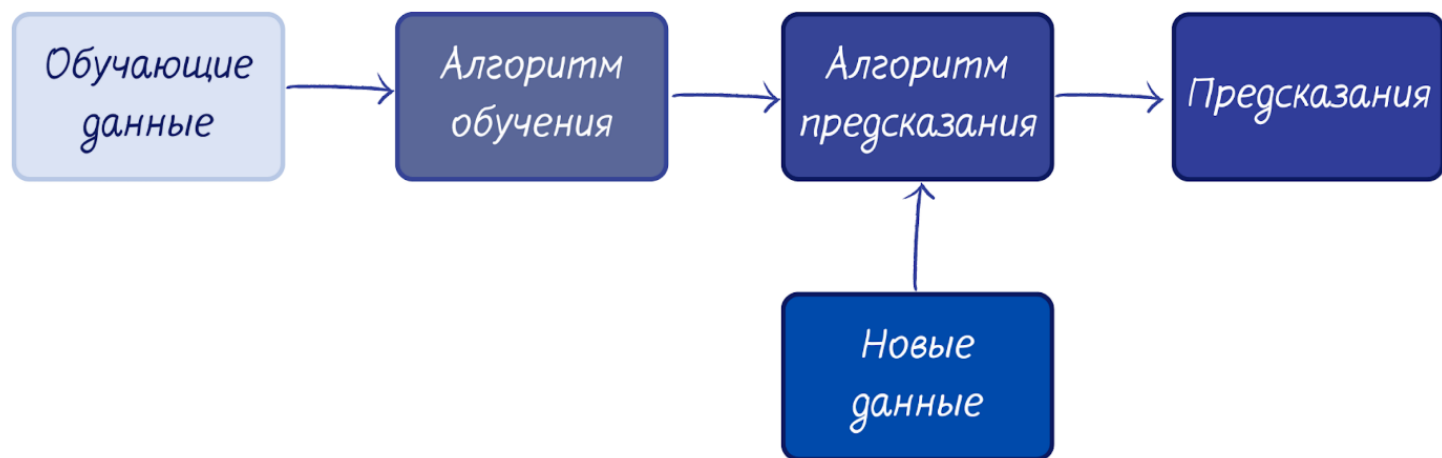


# Кейсы классификации текстов

1. Классификация текстов (определение жанра, темы, тональности)
2. Выделение именованных сущностей (наименования, даты)
3. Генерация текстов (создание текстов на заданную тематику по картинке, по 2-3 предложениям)

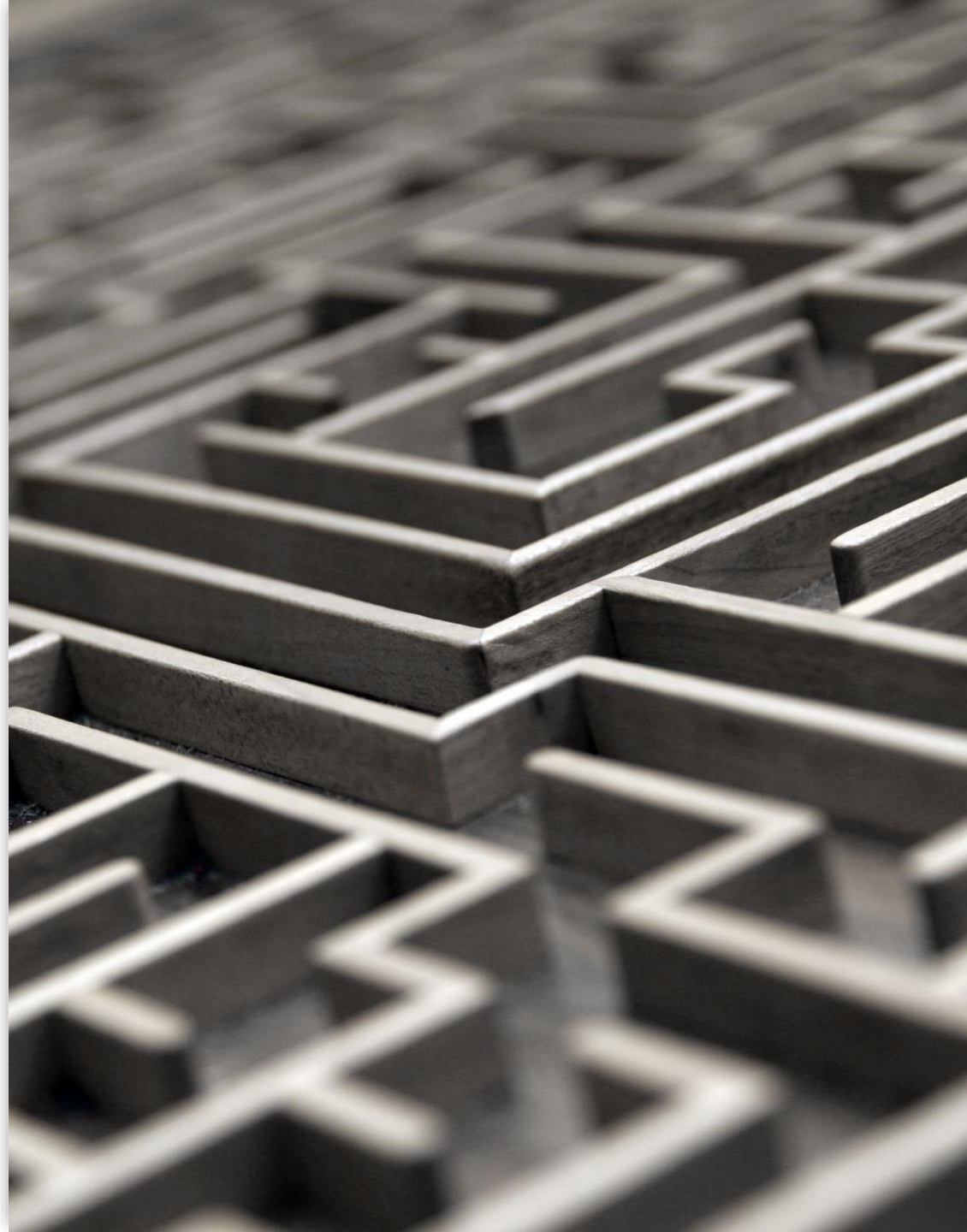


# Машинное обучение

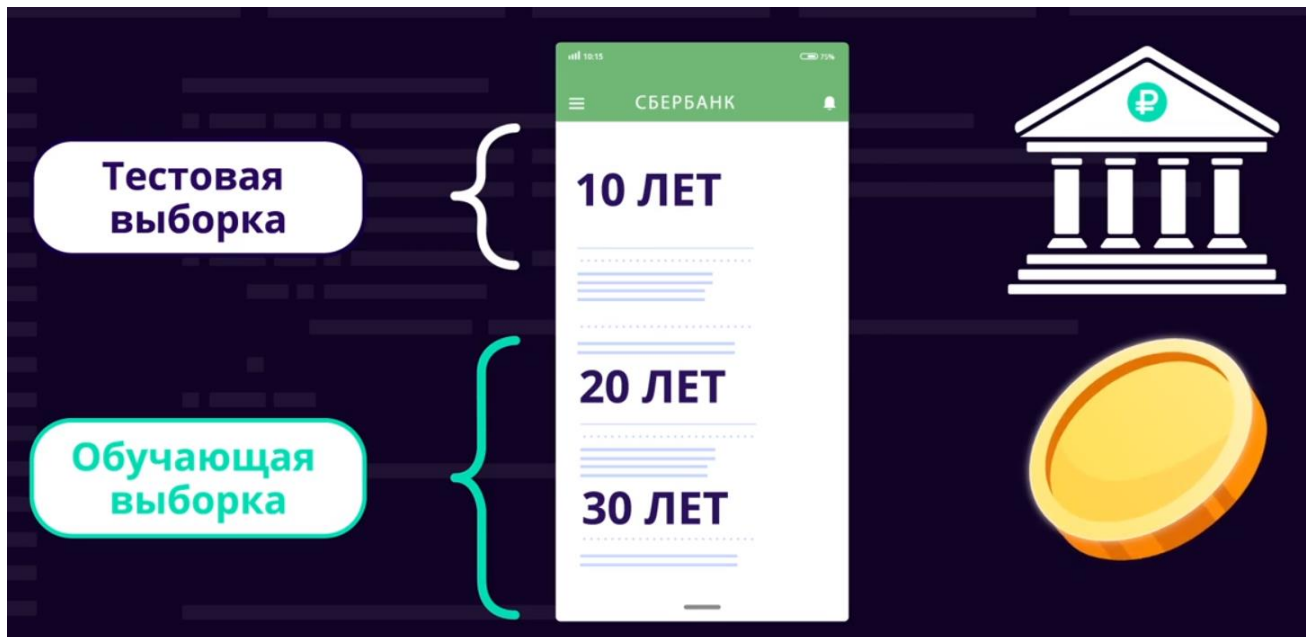


# Задача регрессии

**Задача регрессии** состоит в том, чтобы на основании различных признаков предсказать вещественный ответ, т.е. для каждого объекта нужно предсказать число.



# Пример с монетой.



# Пример с монетой.





## Задача регрессии

Пример. Задача предсказания стоимости квартиры

Площадь (м <sup>2</sup> )	Этаж	Число комнат	Число лет с последнего ремонта	Стоимость (млн)
115	3	4	2	46.0
55	5	2	5	10.3
72	6	3	12	17.7
55	20	1	0	32.0

# Поиск значений весов



# Линейные модели

Самый известный метод регрессии — это линейные модели.

Основной механизм предсказания с помощью линейной модели формулируется следующим образом: необходимо умножить значения всех признаков на веса и сложить.

Предположим, что мы хотим предсказать стоимость квартиры со следующими значениями признаков:

Площадь (м <sup>2</sup> )	Этаж	Число комнат	Число лет с последнего ремонта
70	2	3	5

Также предположим, что мы знаем веса признаков:

0,25 для площади,

1,8 для этажа,

0,5 для числа комнат и

(-0,2) для числа лет со дня ремонта.

Вес признака задает вклад признака в предсказание.

Тогда будет предсказана стоимость  $0,25 \cdot 70 + 1,8 \cdot 2 + 0,5 \cdot 3 - 0,2 \cdot 5 = 21,6$  условных единиц



## Преимущества линейных моделей

Линейные модели, как правило, решают задачу с приемлемым уровнем качества, однако уступают более мощным алгоритмам, ансамблям решающих деревьев и нейронным сетям, которые мы обсудим далее.

С другой стороны, качество линейных моделей можно значительно повысить, придумав новые признаки, вычисляемые на основе исходных признаков (например, добавив квадраты признаков), — при этом свойство интерпретируемости сохраняется. Благодаря своей интерпретируемости линейные модели очень популярны в бизнес-задачах, например в кредитном скоринге.

# Переобучение

Может случиться, что алгоритм делает хорошие предсказания только для обучающих данных. Иными словами, алгоритм запомнил, зазубрил классы/числа для обучающих объектов, но не нашел никаких зависимостей между признаками и целевой переменной (классом/числом). Такой алгоритм будет плохо работать на шаге внедрения и называется переобученным.



Онлайн-курс СберУниверситета

# Генеративное искусство

Подробнее о курсе



Бесплатный курс от Сбера  
по генеративному  
искусству

[https://courses.sberuniversity.ru/generative-art?utm\\_source=telegram&utm\\_medium=organic&utm\\_campaign=courses&utm\\_content=gen\\_i&utm\\_term=01\\_09\\_2023](https://courses.sberuniversity.ru/generative-art?utm_source=telegram&utm_medium=organic&utm_campaign=courses&utm_content=gen_i&utm_term=01_09_2023)