



Пайплайн ML-моделей



Заказчик



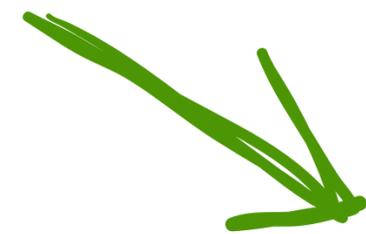
Хочу решить
такую-то задачу



Таблица данных



Набор файлов



Нет данных

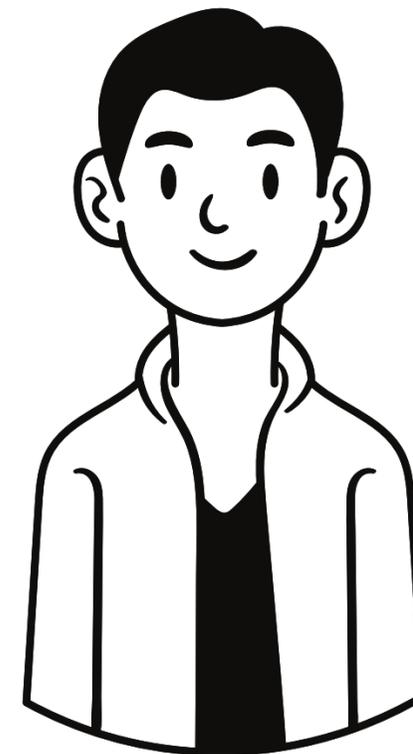


Заказчик



Какое качество
для тебя
приемлемо?

Аналитик



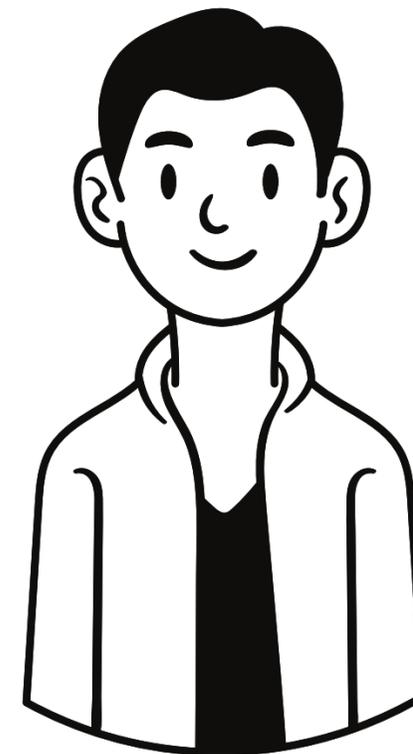


Заказчик



Хотим, чтобы
продажи выросли
на 2%.

Аналитик





Сбор данных

Задача:

построить рекомендательную систему



Признаки:

- история покупок пользователя
- характеристики товаров
- персональные данные пользователя
- признаки на основе даты

Ожидаем, что продажи
вырастут на 2%



Метрика

- Чисто по данным количество продаж не посчитать.
- Поэтому считают какую-то DS-метрику.
- Например точность - доля правильных предсказаний.

Задача для DS:

Точность \longrightarrow **max**





На чем считать точность?



Сравниваем предсказания с реальными значениями

Точность → **max**



Решение задачи

Обучение:

- тупо запомнить все данные

Применение:

- если x был в данных, выдать его метку
- иначе бросить монетку

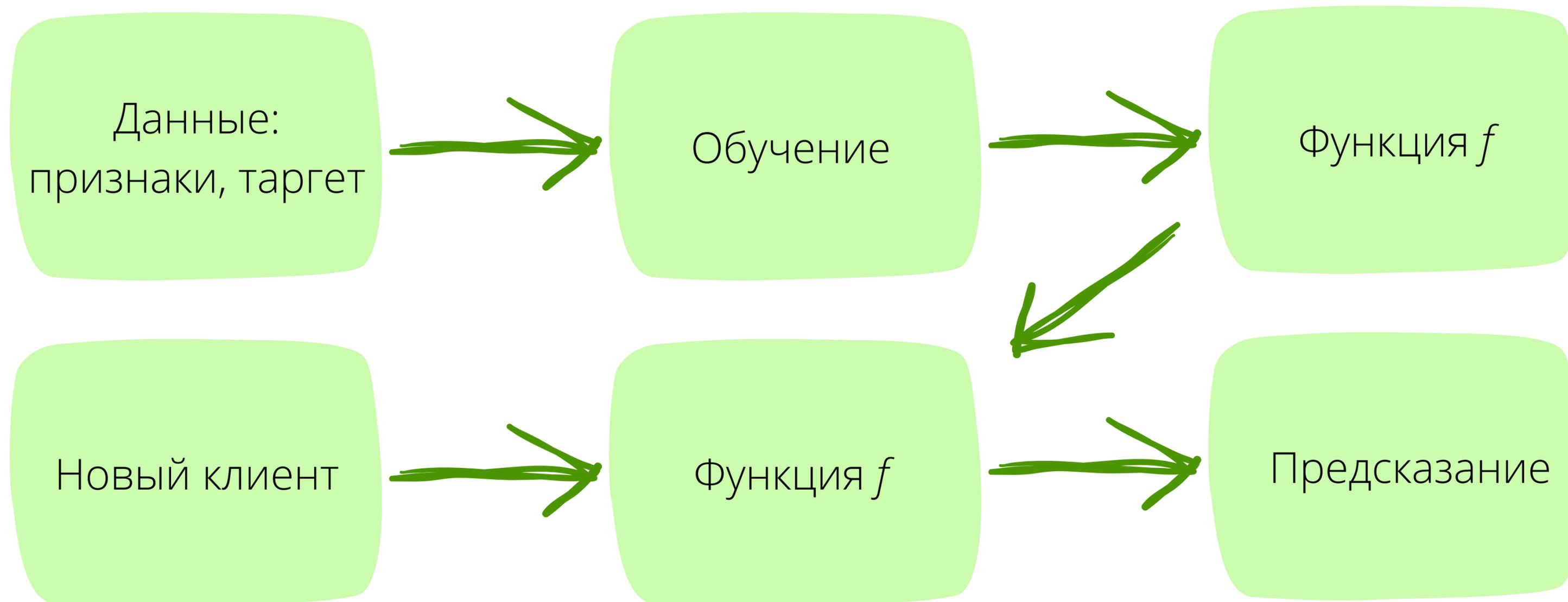
Точность = 100%

Полезность = 0





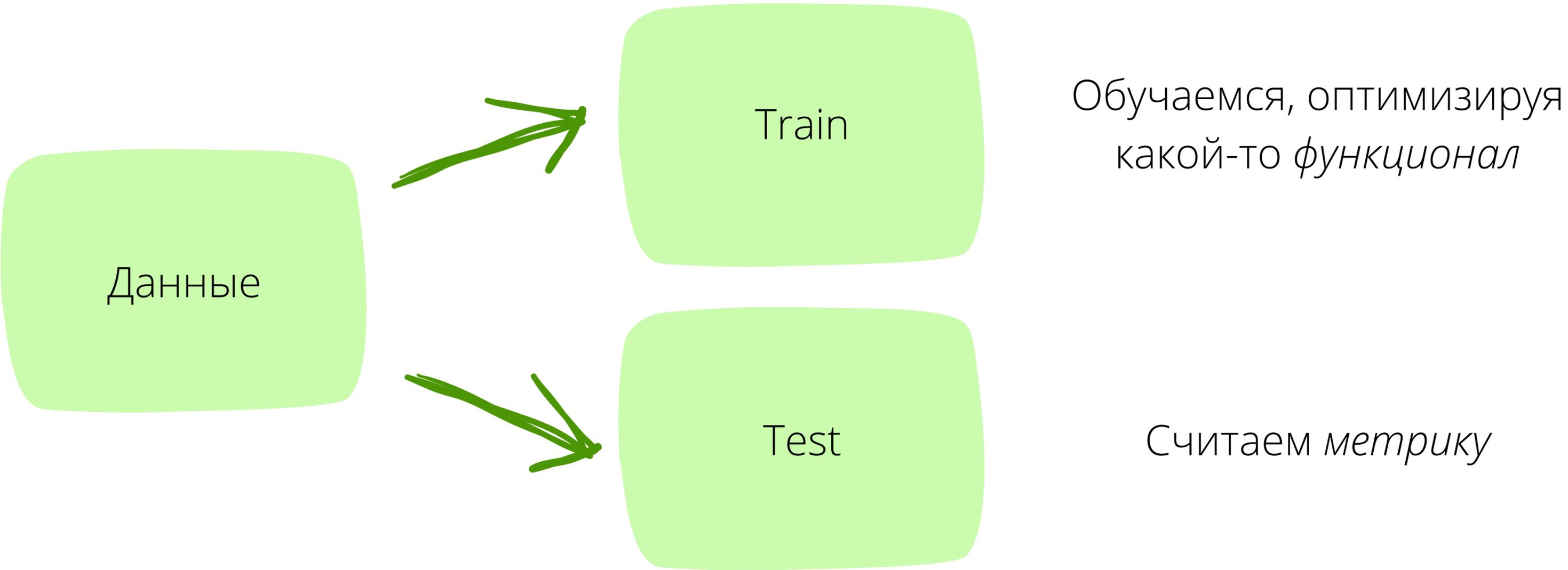
Как на самом деле устроен процесс?



Вывод: необходимо, чтобы модель хорошо работала на новых данных.



Деление данных



Точность (test) → max



Данные нужно приготовить

Сырые данные



Правило
определяем
только по train

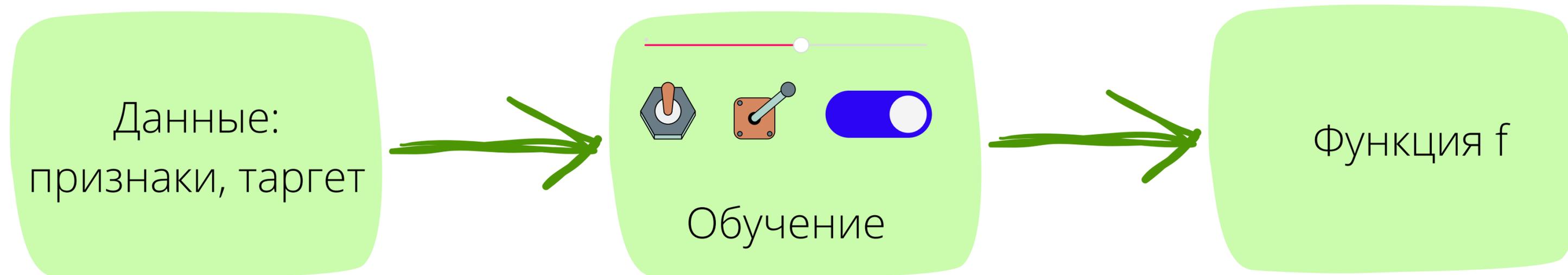
Подготовленные данные



Подготовка данных занимает 80% времени аналитика



Обучение

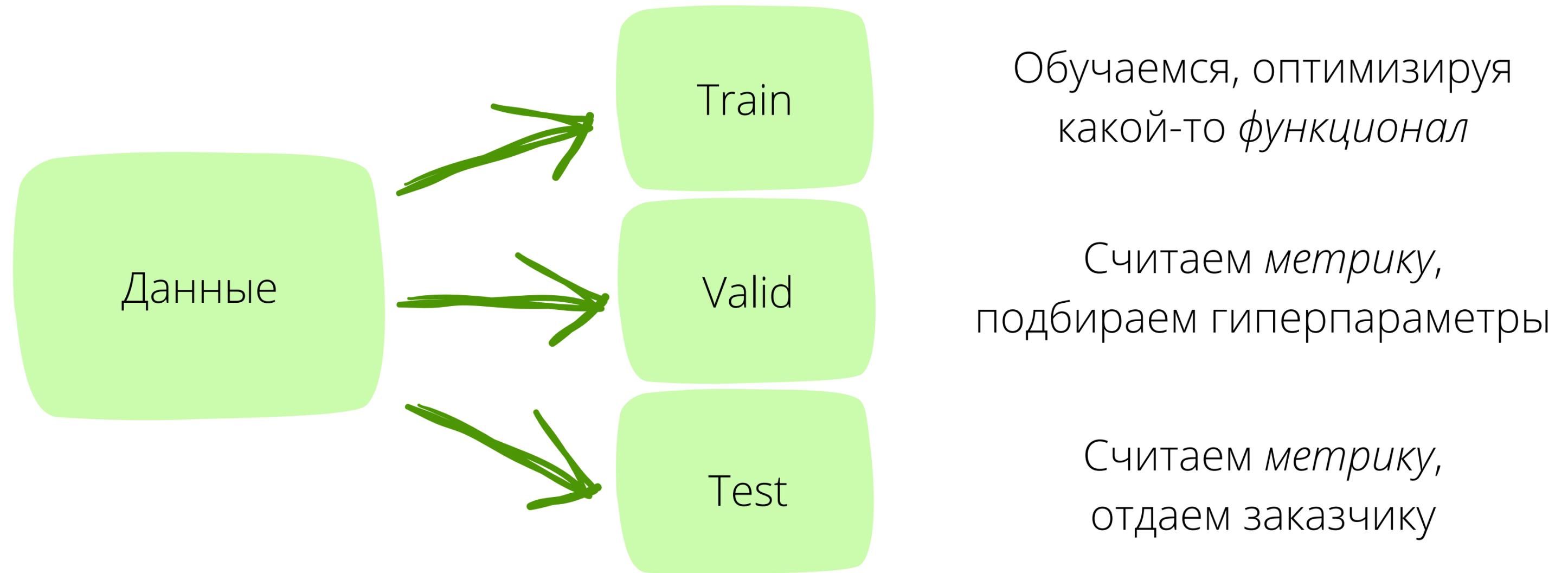


Обычно метод обучения имеет разные параметры:

- **настраиваемые** методом параметры с помощью оптимизации *функционала*
- **гиперпараметры**, которые задаются пользователем



Валидация



Точность (valid) \longrightarrow **max**

Точность (test) отдаем заказчику



Заказчик



Вот, готово.
Потратил на нее
2 месяца

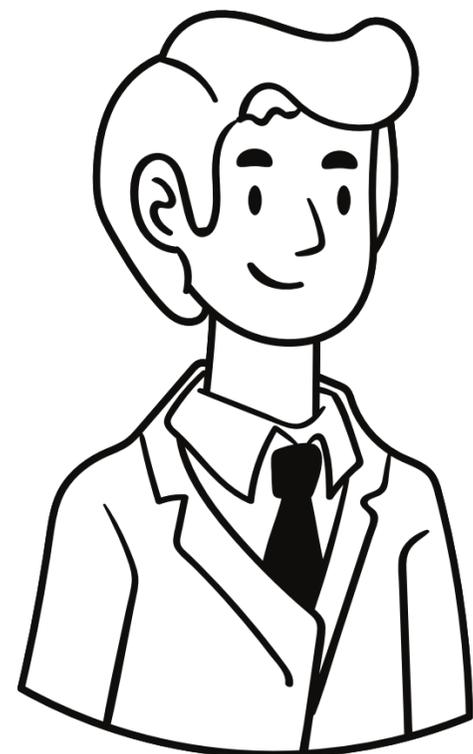
Аналитик



Модель - черный ящик



Заказчик

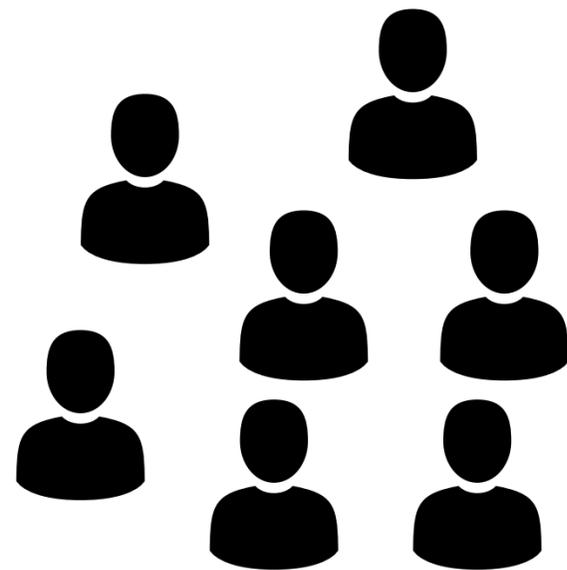


Аналитик



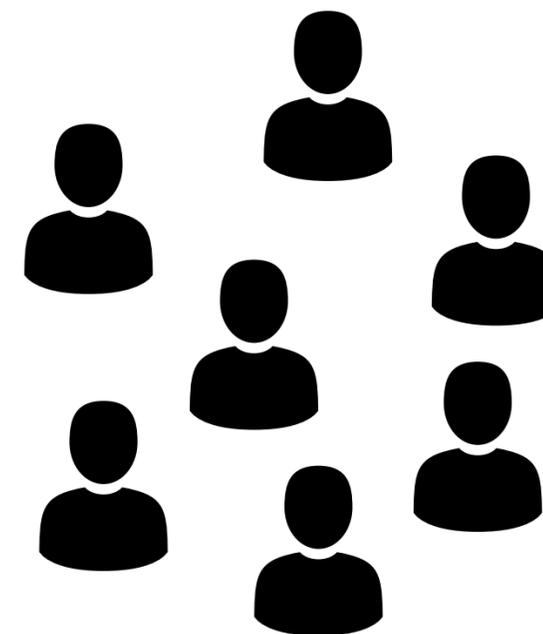


Проводим АВ-тест



Группа А

Старая модель



Группа В

Новая модель

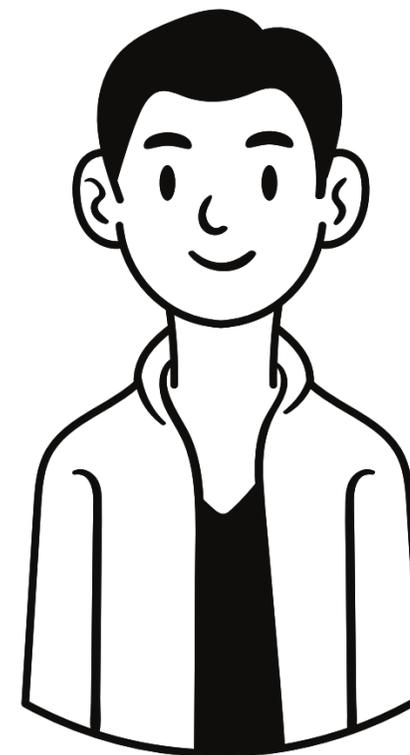
Отличаются ли продажи между группами?
На вопрос отвечают статистические критерии.



Заказчик



Аналитик



Провели АВ-тест.
На тестовой группе
увеличение продаж
на 3%, результат
статистически значим.

Круто!
Выкатываем
модель.



Слово студентам DS-потока



Елена Велегурина
3 курс, DS-поток
tg: @LenkaVelegurina