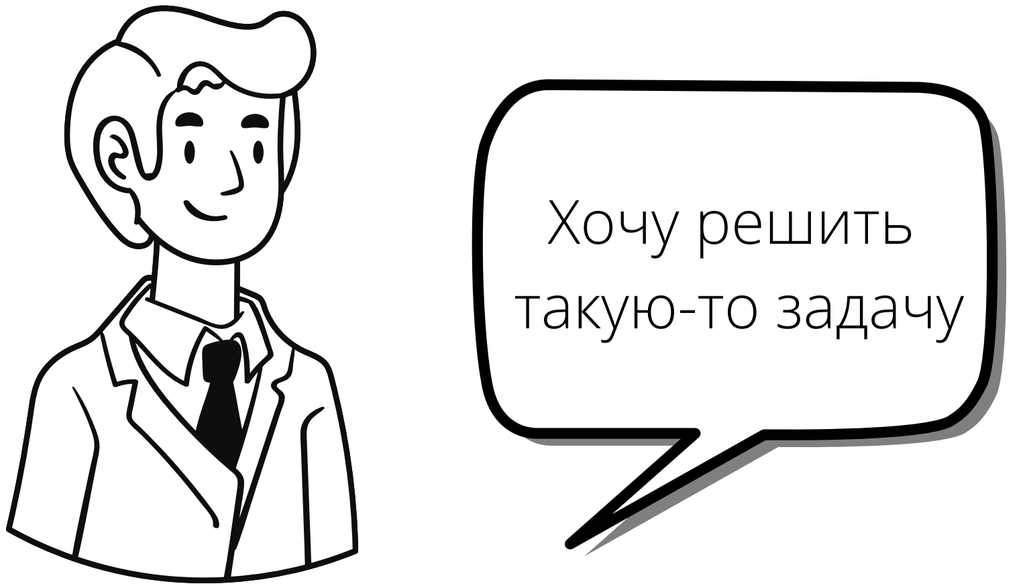
Пайплайн ML-моделей

Заказчик

Заказчик



Заказчик

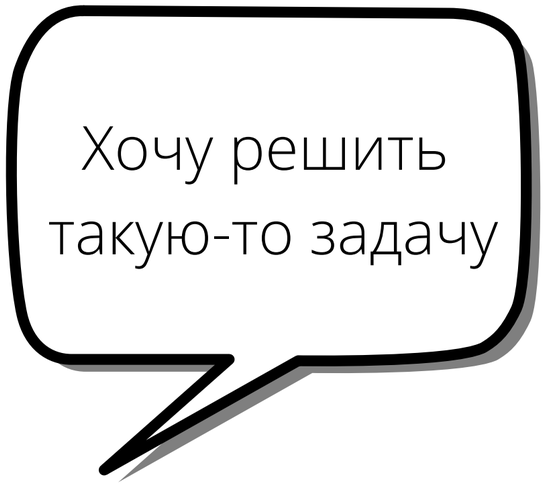
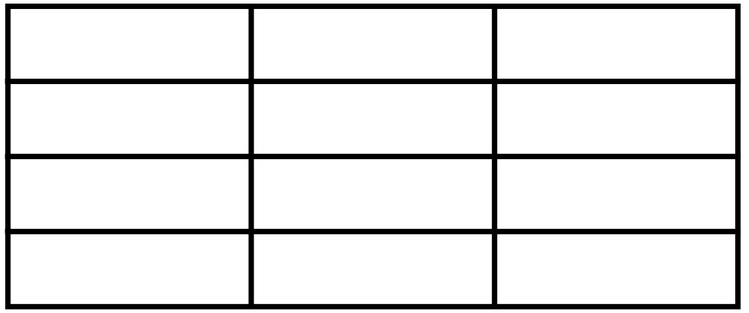


Таблица данных

Заказчик

Хочу решить
такую-то задачу

Таблица данных



Набор файлов

Заказчик

Хочу решить
такую-то задачу

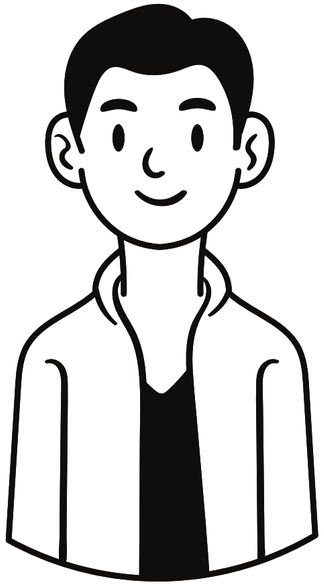
Таблица данных

Набор файлов

Нет данных

Заказчик

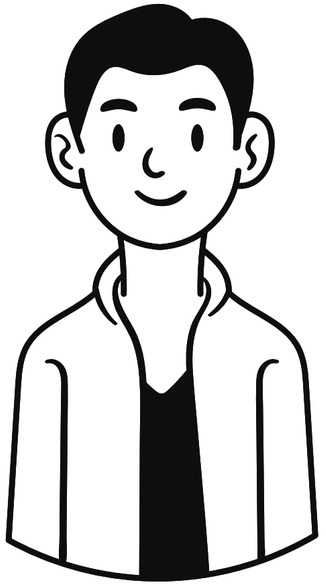
Аналитик



Заказчик

Какое качество для тебя приемлемо?

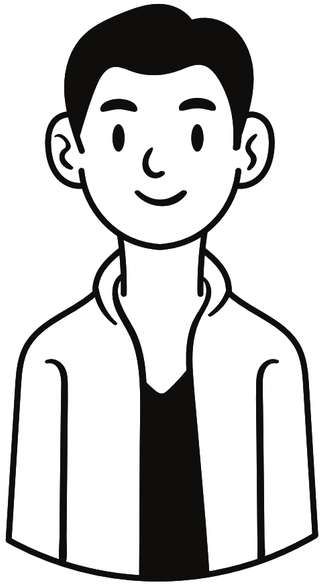
Аналитик



Заказчик

Хотим, чтобы продажи выросли на 2%.

Аналитик



Сбор данных

Задача:

построить рекомендательную систему

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.
-

Метрика

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

Метрика

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

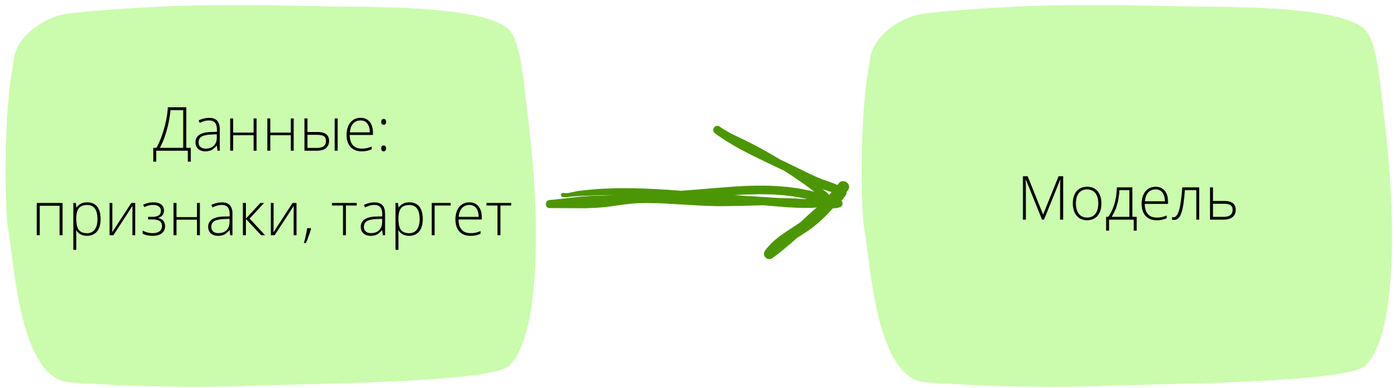
Задача для DS:

Точность ⟶ **max**

На чем считать точность?

Данные:
признаки, таргет

На чем считать точность?



На чем считать точность?

Сравниваем предсказания
с реальными значениями

На чем считать точность?

Сравниваем предсказания
с реальными значениями

Точность → **max**

Решение задачи

Решение задачи

Обучение:

- тупо запомнить все данные

Решение задачи

Обучение:

- тупо запомнить все данные

Применение:

- если *x* был в данных, выдать его метку

Решение задачи

Обучение:

- тупо запомнить все данные

Применение:

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

Решение задачи

Обучение:

- тупо запомнить все данные

Применение:

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

Точность = 100%

Решение задачи

Обучение:

- тупо запомнить все данные

Применение:

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

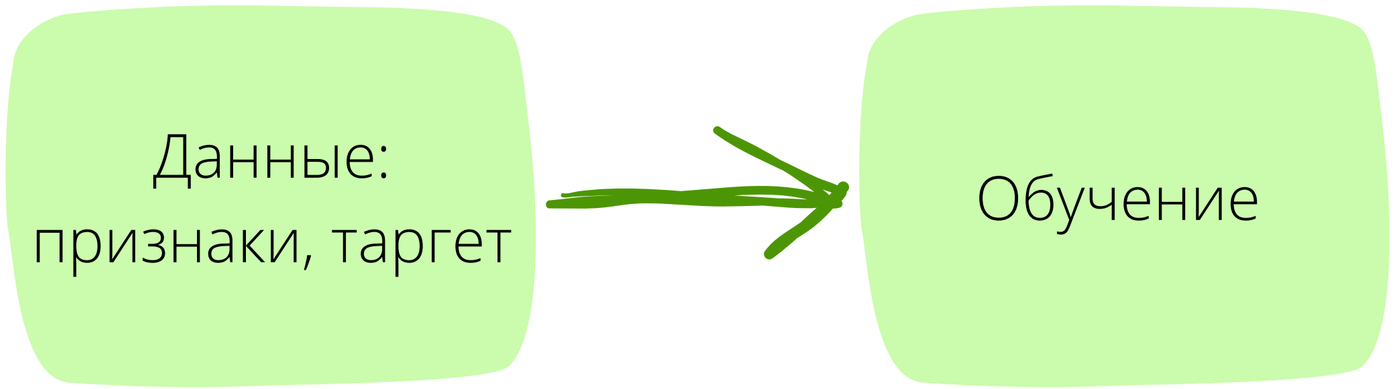
Точность = 100%

Полезность = 0

Как на самом деле устроен процесс?

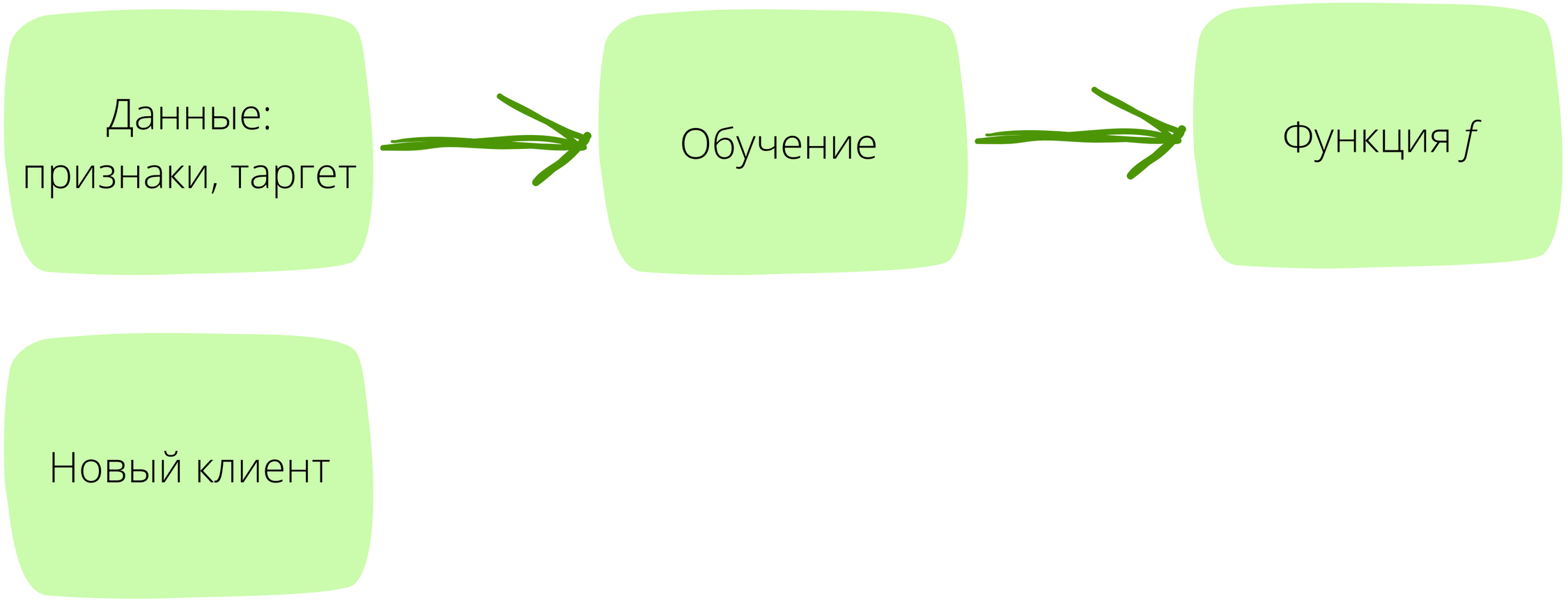
Данные:
признаки, таргет

Как на самом деле устроен процесс?

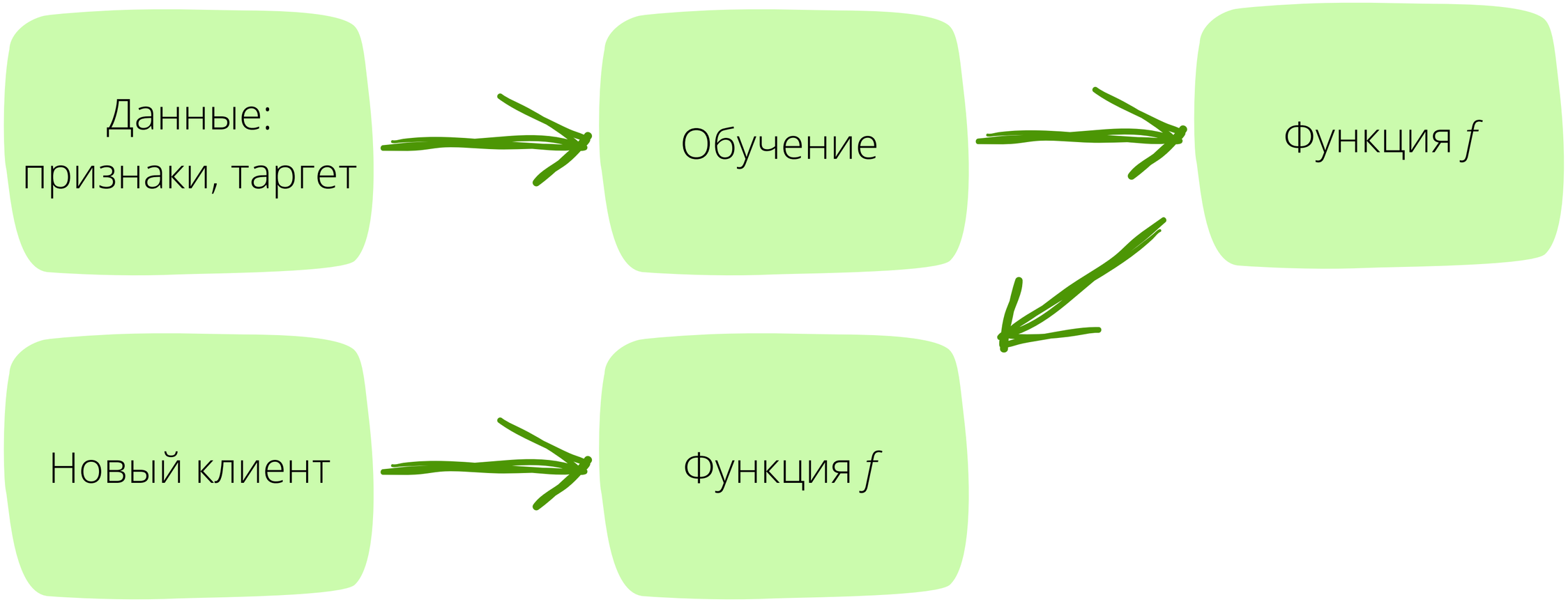


Как на самом деле устроен процесс?

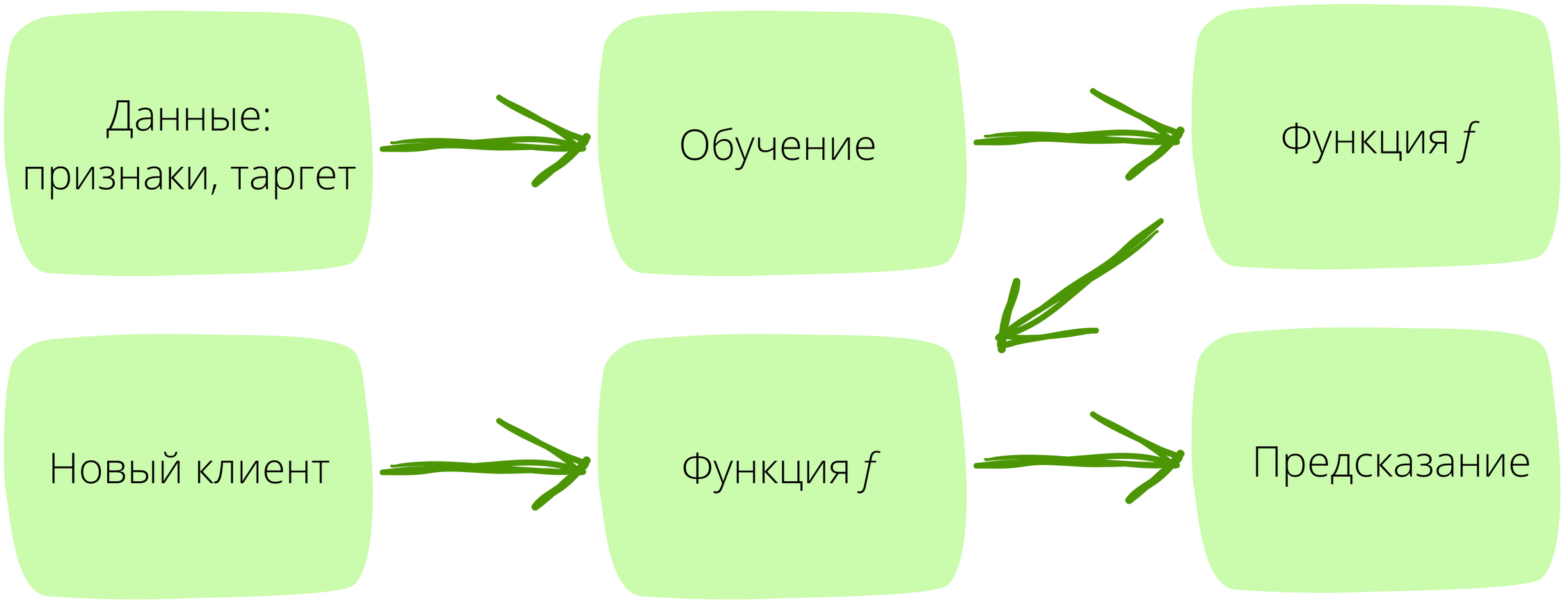
Как на самом деле устроен процесс?



Как на самом деле устроен процесс?

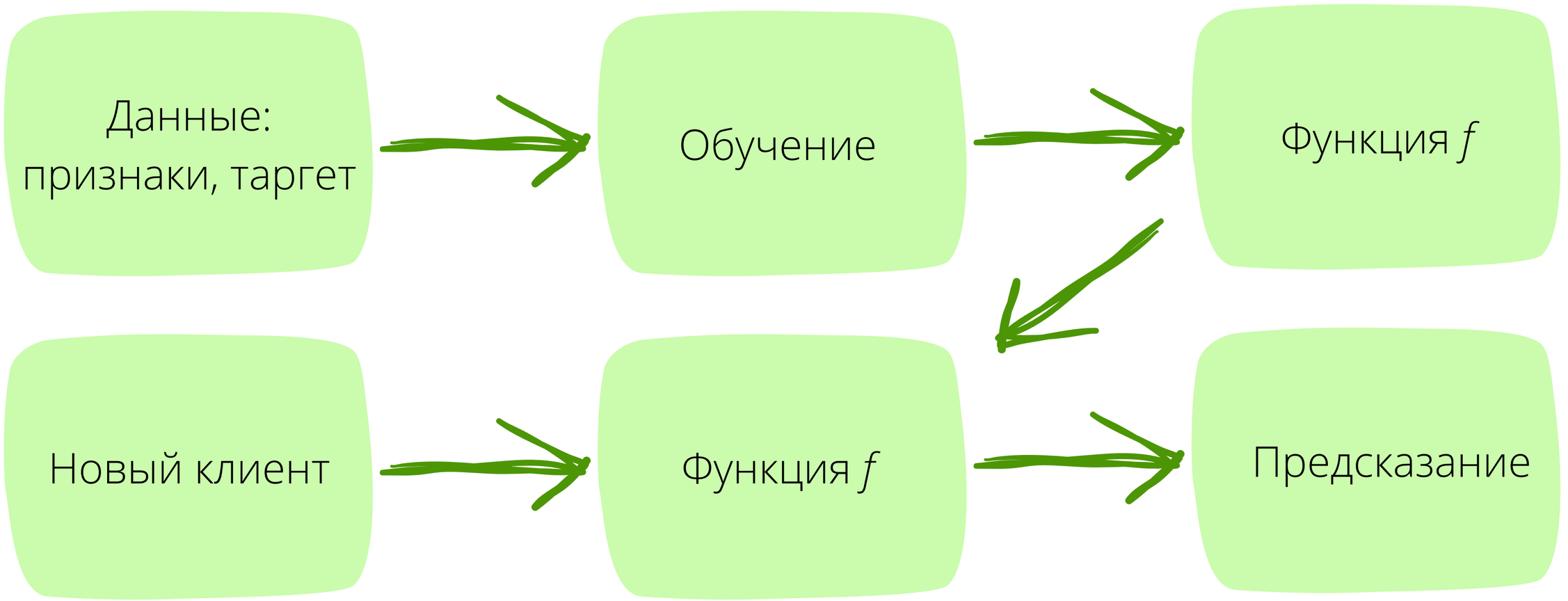


Как на самом деле устроен процесс?



Вывод:

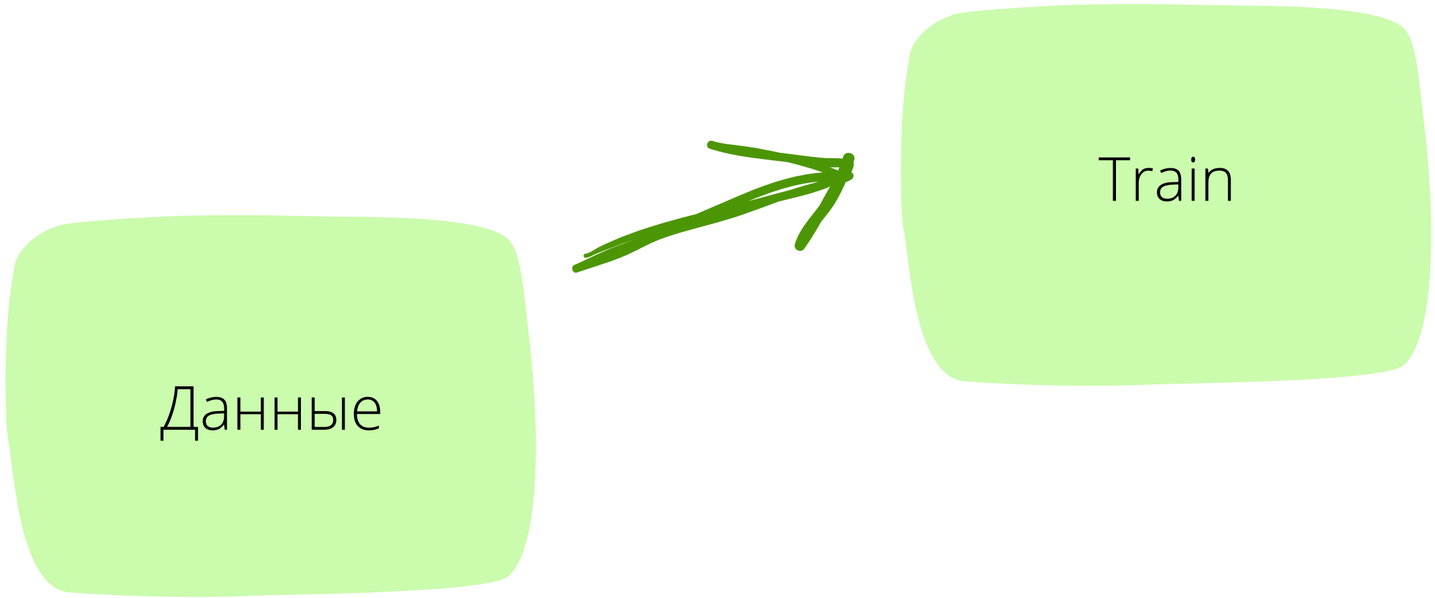
Как на самом деле устроен процесс?



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

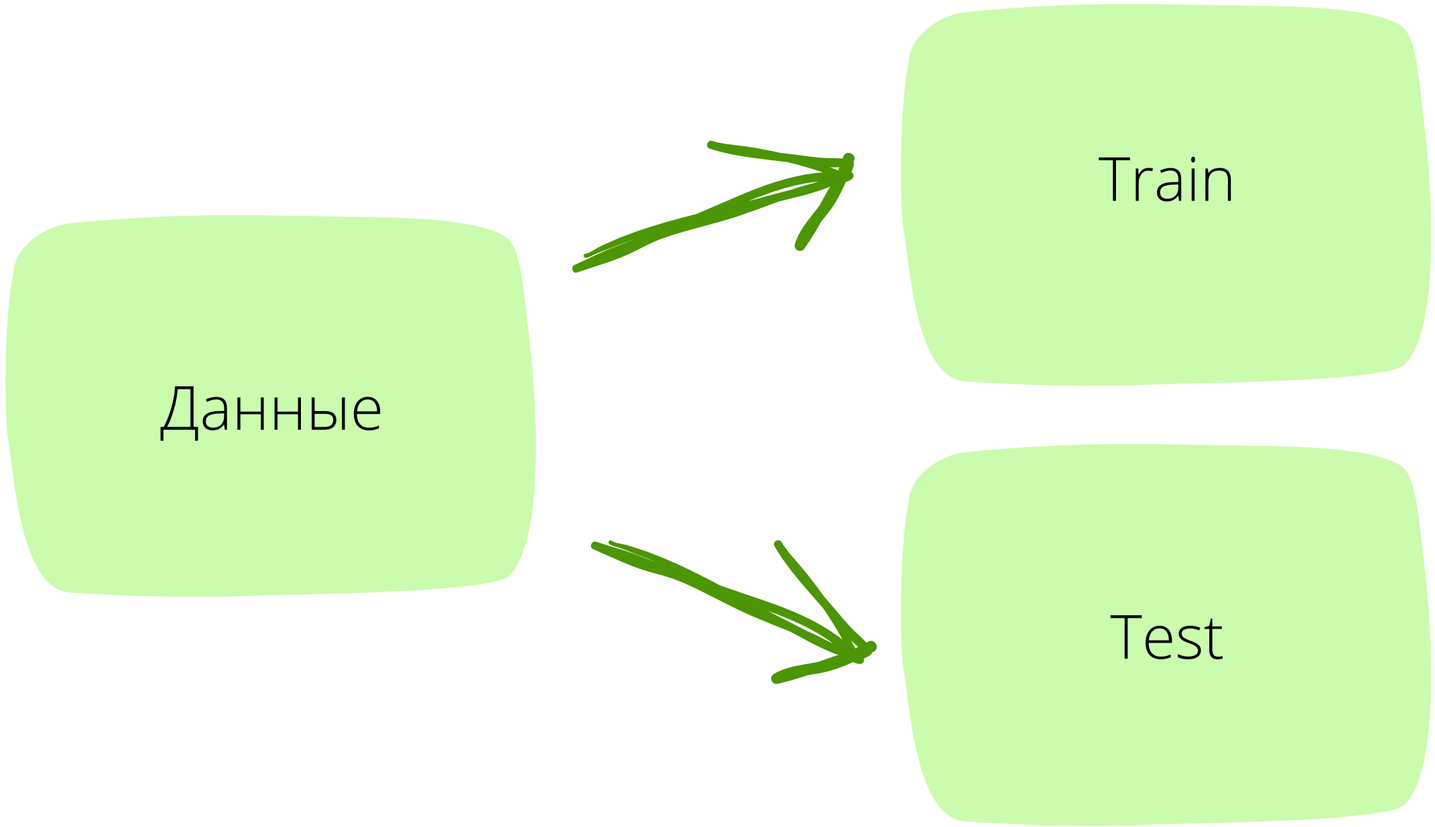
Деление данных

Деление данных



Обучаемся, оптимизируя какой-то *функционал*

Деление данных



Обучаемся, оптимизируя какой-то *функционал*

Считаем *метрику*

Точность (test) ⟶ **max**

Данные нужно приготовить

Данные нужно приготовить

Сырые данные

Данные нужно приготовить

Сырые данные

Подготовленные данные

Данные нужно приготовить

Сырые данные

Правило
определяем
только по train

Подготовленные данные

Данные нужно приготовить

Сырые данные

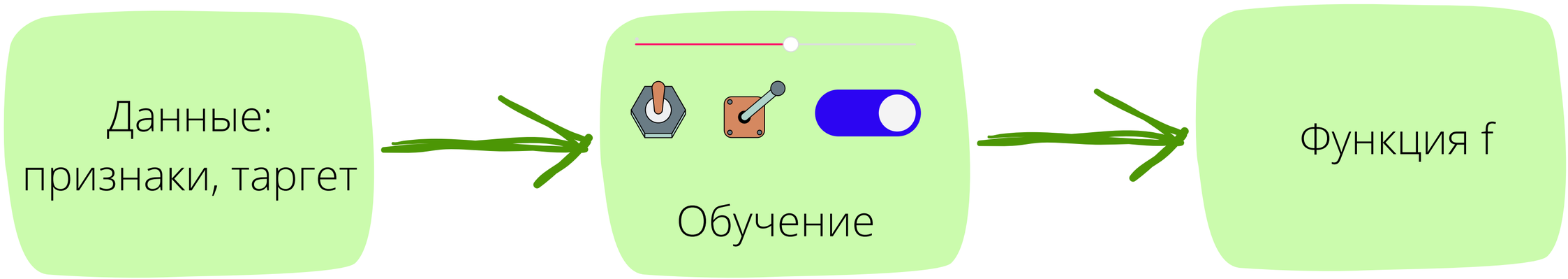
Правило
определяем
только по train

Подготовленные данные

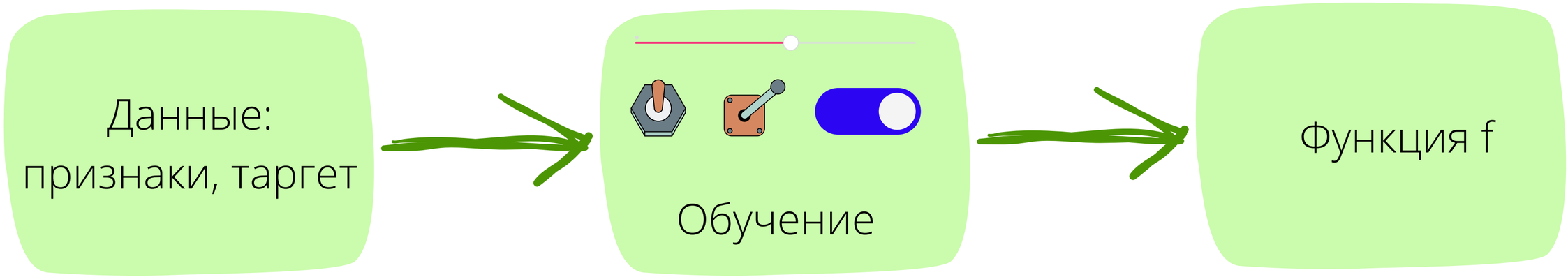
Подготовка данных занимает 80% времени аналитика

Обучение

Обучение



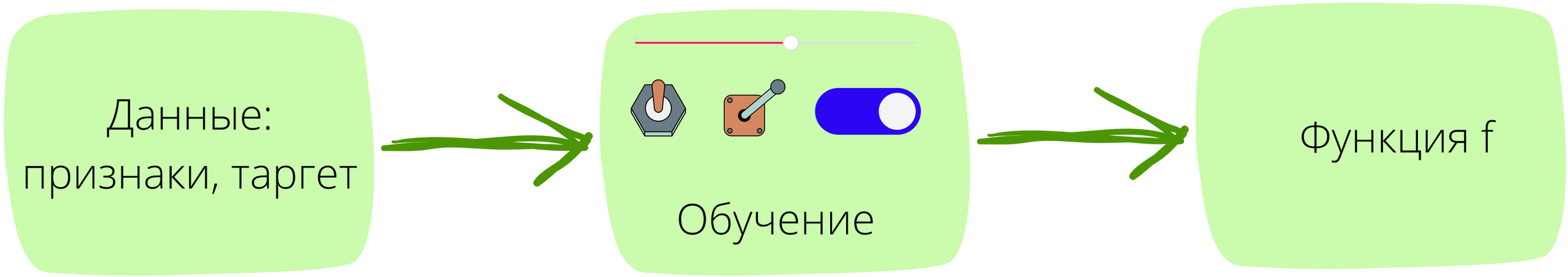
Обучение



Обычно метод обучения имеет разные параметры:

-

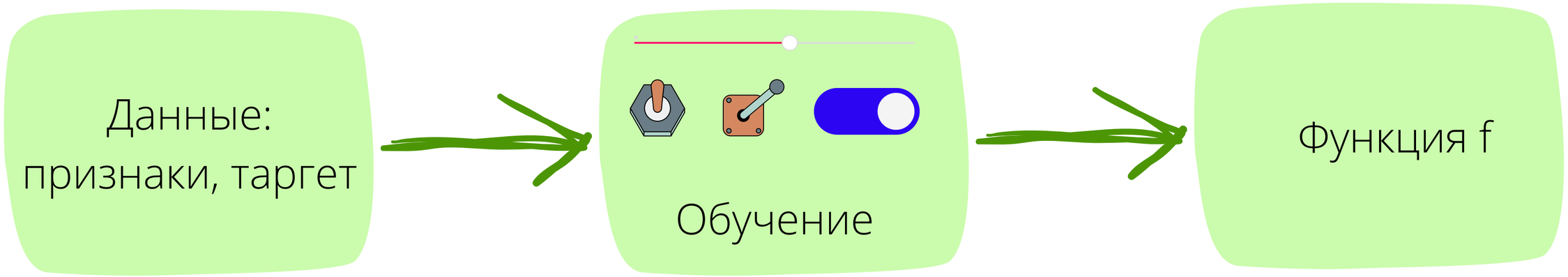
Обучение



Обычно метод обучения имеет разные параметры:

- **настраиваемые** методом параметры
  с помощью оптимизации *функционала*
-

Обучение

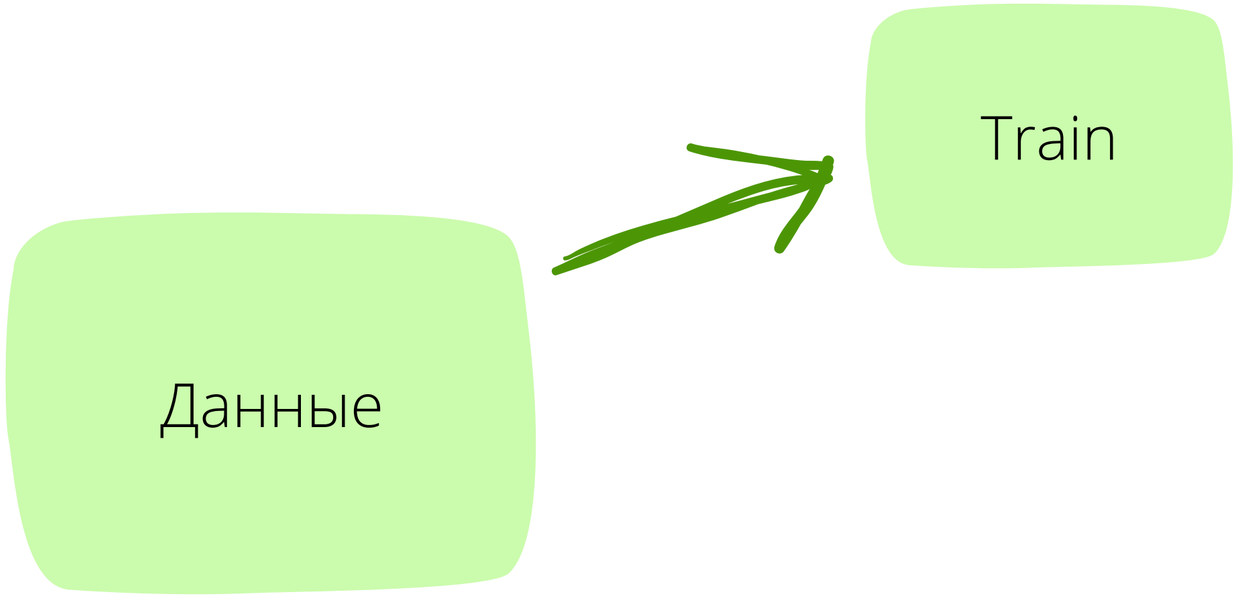


Обычно метод обучения имеет разные параметры:

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

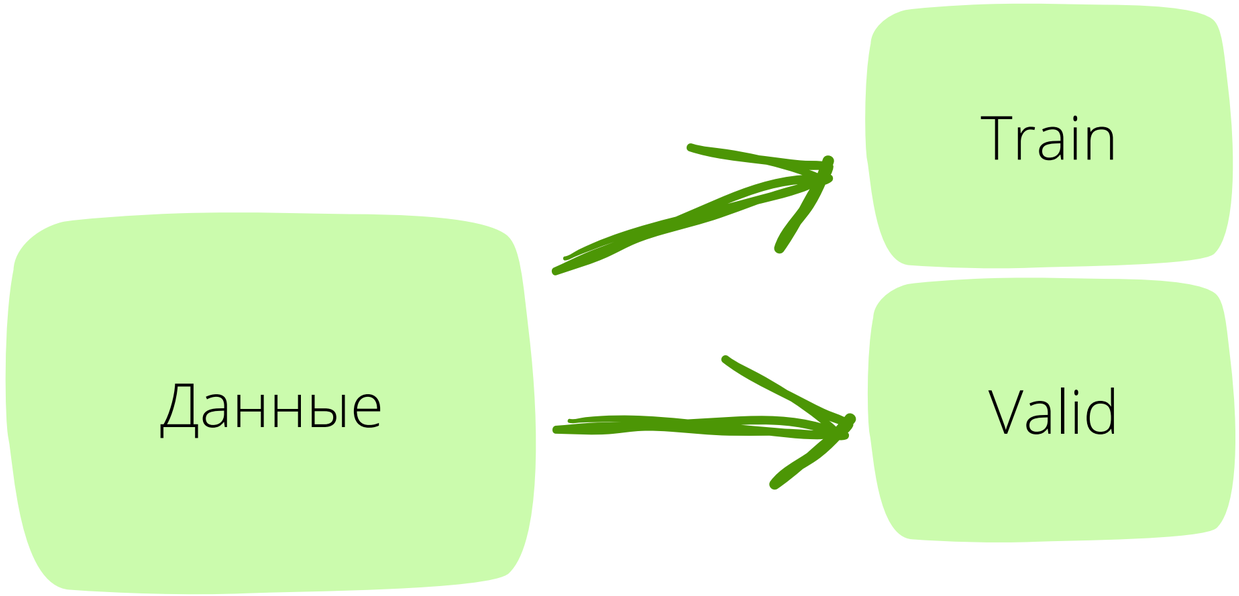
Валидация

Валидация



Обучаемся, оптимизируя
какой-то *функционал*

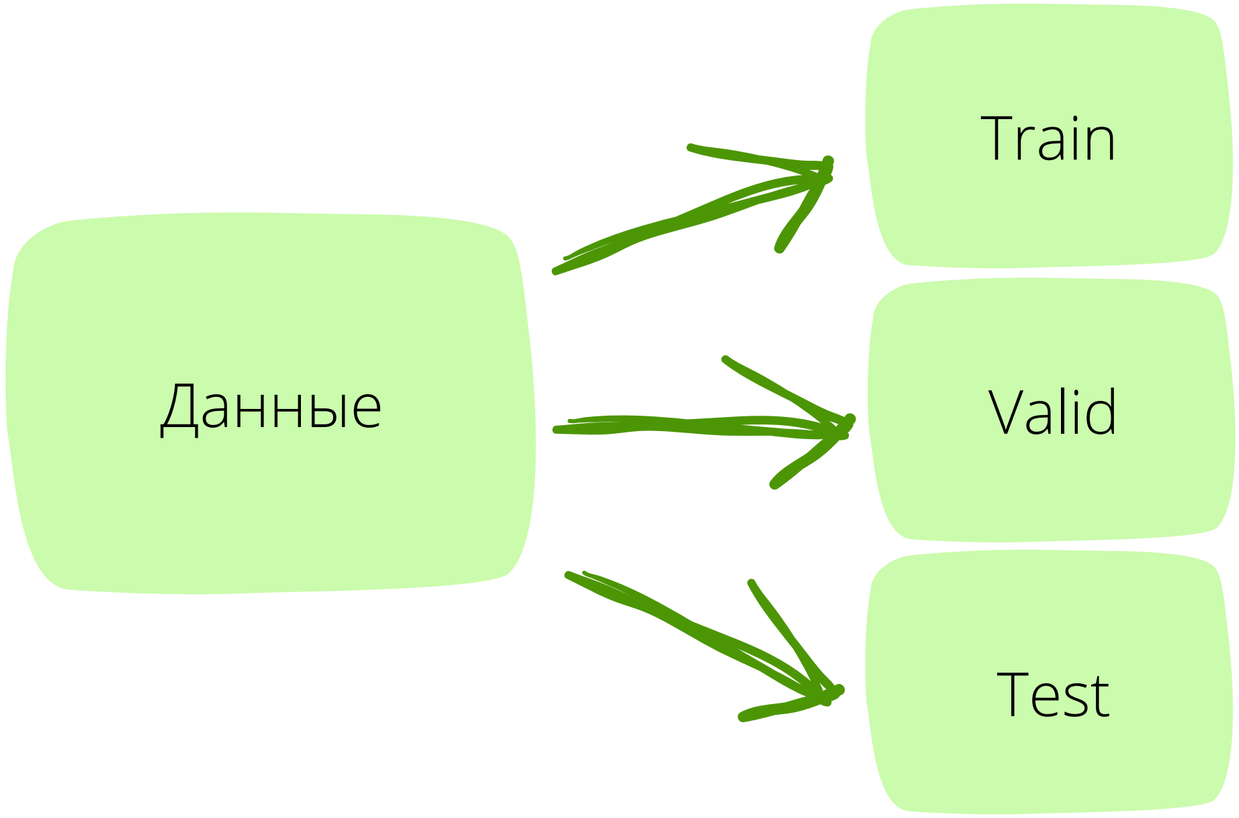
Валидация



Обучаемся, оптимизируя
какой-то *функционал*

Считаем *метрику*,
подбираем гиперпараметры

Валидация

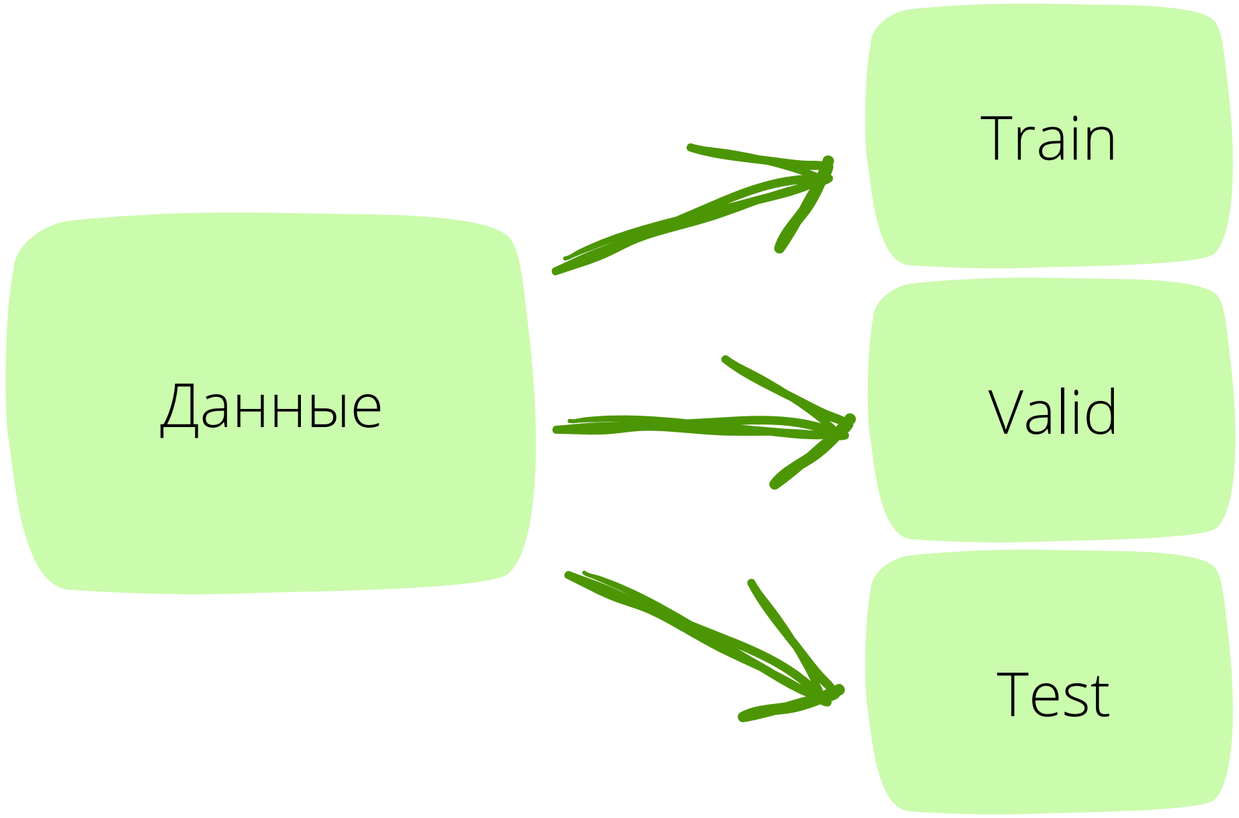


Обучаемся, оптимизируя
какой-то *функционал*

Считаем *метрику*,
подбираем гиперпараметры

Считаем *метрику*,
отдаем заказчику

Валидация



Обучаемся, оптимизируя
какой-то *функционал*

Считаем *метрику*,
подбираем гиперпараметры

Считаем *метрику*,
отдаем заказчику

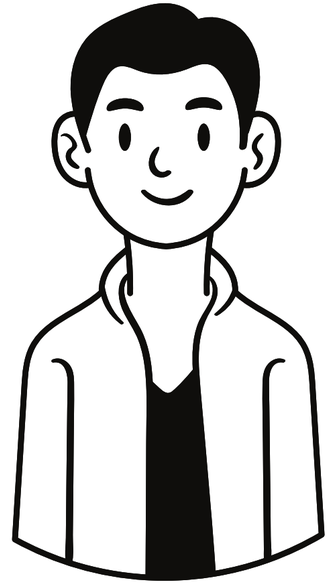
Точность (valid) → **max**

Точность (test) отдаем заказчику

Заказчик

Вот, готово.
Потратил на нее
2 месяца

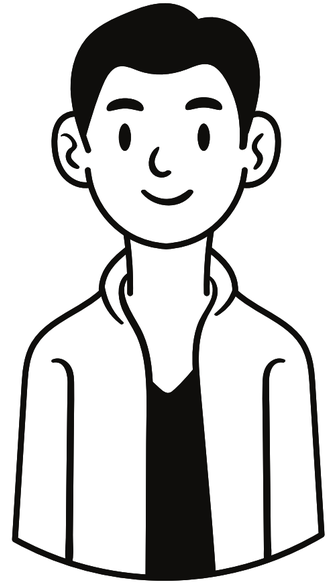
Аналитик



Заказчик

Вот, готово.
Потратил на нее
2 месяца

Аналитик

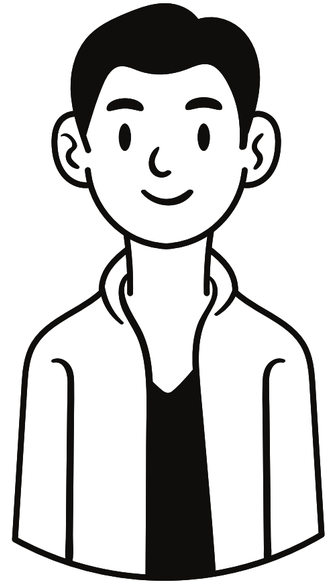


Модель — черный ящик

Заказчик

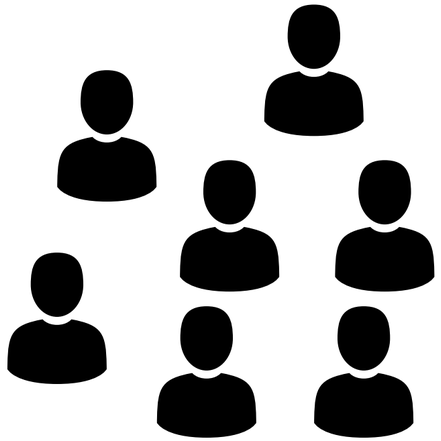
Хорошо,
а продажи
увеличись?

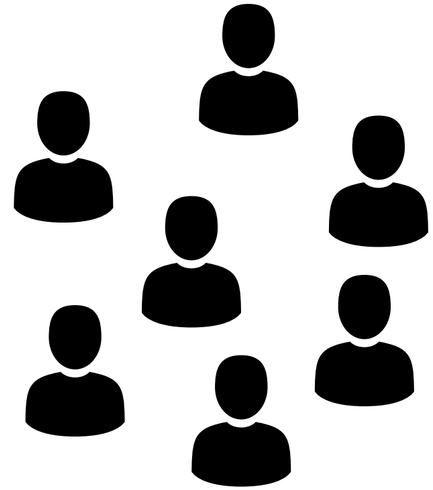
Аналитик



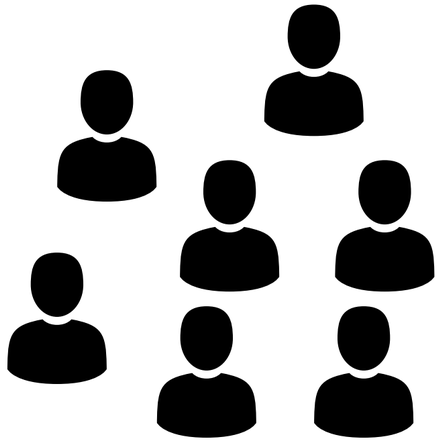
Проводим AB-тест

Проводим АВ-тест



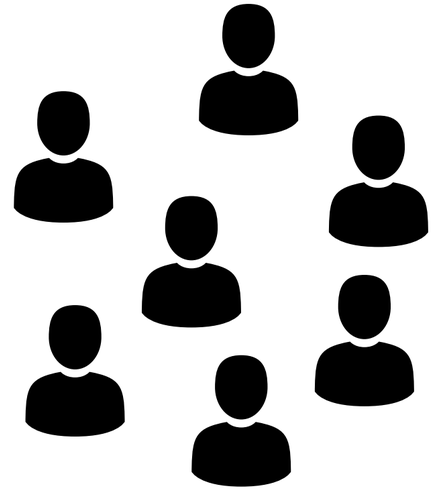


Проводим AB-тест



Группа A

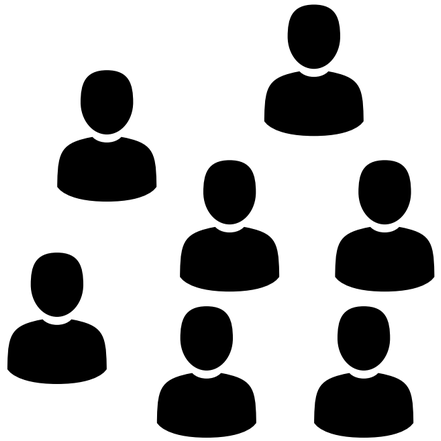
Старая модель



Группа B

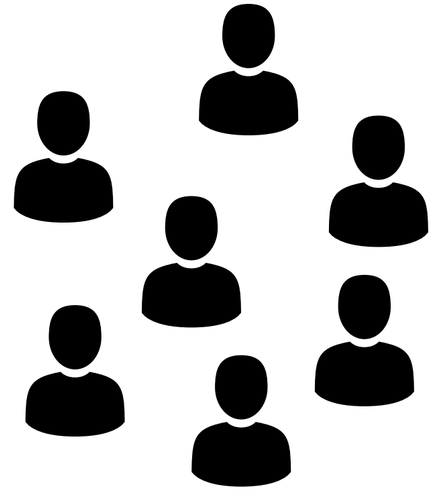
Новая модель

Проводим AB-тест



Группа A

Старая модель



Группа B

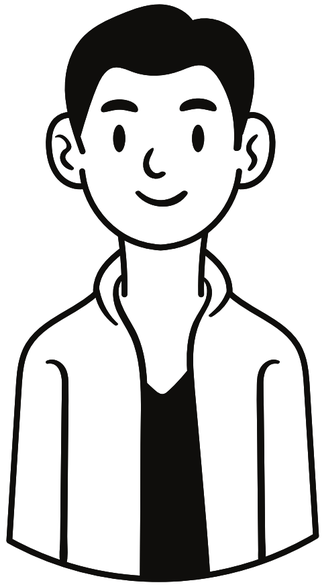
Новая модель

Отличаются ли продажи между группами?

На вопрос отвечают статистические критерии.

Заказчик

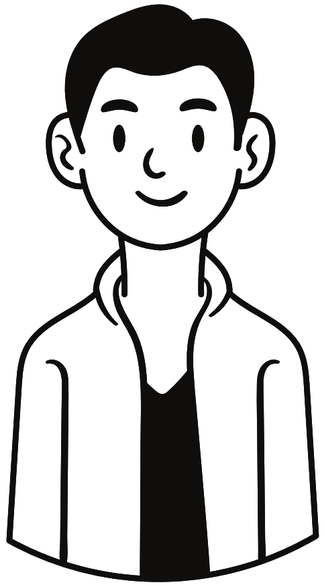
Аналитик



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

Заказчик

Аналитик



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

Круто!
Выкатываем
модель.