



Машинное обучение обзор



Машинное обучение - обзор

- Наконец пришло время погрузиться в детали машинного обучения!
- Этот раздел поможет Вам настроиться на нужный образ мышления, свойственный для машинного обучения
- Для начала давайте вспомним наши этапы проекта по машинному обучению...



Этапы работ по машинному обучению



Реальный
мир

Решить
задачу

Как починить или
поменять X ?

Ответить
на вопрос

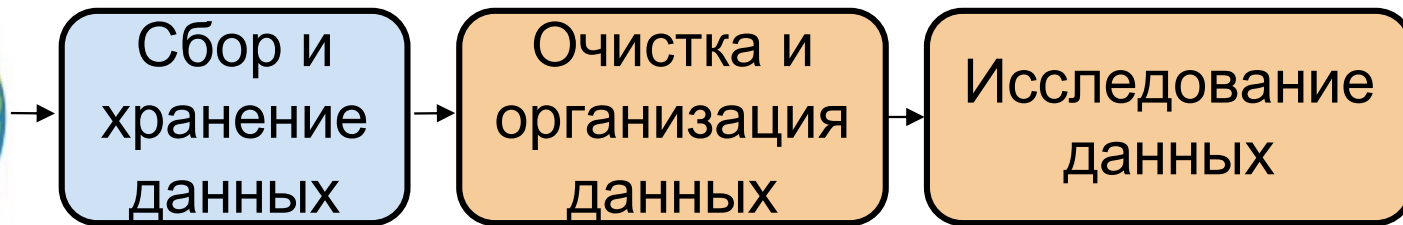
Как изменение в X
повлияет на Y ?



Этапы работ по машинному обучению



Реальный
мир



Статистический анализ,
Визуализация данных



Этапы работ по машинному обучению



Реальный
мир

Сбор и
хранение
данных

Очистка и
организация
данных

Исследование
данных

Модели
Машинного
Обучения

Supervised Learning:

Сделать предсказание

Unsupervised Learning:

Найти закономерности
в данных



Этапы работ по машинному обучению



Реальный
мир

Сбор и
хранение
данных

Очистка и
организация
данных

Исследование
данных

Модели
Машинного
Обучения

Supervised Learning:

Сделать предсказание

Unsupervised Learning:

Найти закономерности
в данных



Машинное обучение - обзор

- В этом разделе мы обсудим:
 - Какие задачи решает машинное обучение
 - Типы машинного обучения:
 - Supervised Learning
 - Unsupervised Learning
 - Процесс машинного обучения для Supervised Learning



Машинное обучение - обзор

- По мере продвижения вперёд мы будем постепенно изучать многие темы:
 - Баланс между смещением и дисперсией
 - Кросс-валидация
 - Построение признаков (Feature Engineering)
 - Scikit-learn
 - Метрики оценки модели и многое другое!



Машинное обучение - обзор

- Разделы по машинному обучению
 - Для каждого вида алгоритма
 - Интуиция и математическая теория
 - Пример написания кода для применения алгоритма
 - Расширения алгоритма
 - Упражнения с решениями



Машинное обучение - обзор

- Разделы по машинному обучению
 - Для линейной регрессии немного по-другому
 - Интуиция и математическая теория
 - Простая линейная регрессия
 - Scikit-learn и линейная регрессия
 - Регуляризация



Машинное обучение - обзор

- Разделы по машинному обучению
 - Изучение дополнительных тем по машинному обучению
 - Метрики оценки модели
 - Построение признаков (Feature Engineering)
 - Кросс-валидация
 - Применение этих тем для упражнения по линейной регрессии



Машинное обучение - обзор

- Далее мы посмотрим, зачем нужно машинное обучение, и в каких случаях оно применяется!



Зачем нужно Машинное обучение?



Зачем нужно машинное обучение

- Машинное обучение – это изучение статистических компьютерных алгоритмов, которые автоматически улучшаются на данных
- В отличие от обычных алгоритмов, где человек должен указывать компьютеру, какое направление выбрать, здесь алгоритмы сами выбирают лучший подход на основе входных данных



Зачем нужно машинное обучение

- Машинное обучение (machine learning) является частью более общего направления искусственный интеллект (artificial intelligence)
- Алгоритмы машинного обучения не программируются явно на то или иное решение
- Вместо этого, алгоритмы определяют оптимальные решения на основе данных



Зачем нужно машинное обучение

- Примеры задач, решаемых с помощью машинного обучения:
 - Кредитный скоринг
 - Риски страхования
 - Предсказание цен
 - Фильтрация спама
 - Сегментация клиентов
 - И многое другое!



Зачем нужно машинное обучение

- Подход к постановке задачи по машинному обучению:
 - На основе признаков (features) в наборе данных, найти требуемую целевую переменную (label/target)
 - Алгоритмы машинного обучения часто ещё называют словом “estimator”, поскольку они оценивают (estimate) целевую переменную



Зачем нужно машинное обучение

- Каким образом алгоритмы машинного обучения могут решать столько разных задач?
- Эти алгоритмы с помощью статистических методов обрабатывают данные, в результате чего выясняют, какие признаки являются важными в данных



Зачем нужно машинное обучение

- Простой пример:
 - Предсказать цену продажи дома на основе его характеристик (район, количество комнат и т.д.)



Зачем нужно машинное обучение

- Предсказания цены дома:
 - Обычный алгоритм – человек пишет алгоритм, в котором указывает значение важности для каждого признака
 - Алгоритм машинного обучения – алгоритм сам определяет важность признаков, на основе самих данных



Зачем нужно машинное обучение

- Зачем нужно машинное обучение?
 - Многие сложные задачи проще решить с помощью методов машинного обучения
 - Такие задачи, как выявление спама, распознавание написанного вручную текста, эффективно решаются с помощью машинного обучения



Зачем нужно машинное обучение

- Почему бы не использовать машинное обучение повсеместно?
 - Главное требование алгоритмов ML – наличие хороших данных
 - Бóльшая часть времени на построение модели тратится на очистку и подготовку данных, а вовсе не на реализацию алгоритма машинного обучения



Зачем нужно машинное обучение

- Нужно ли разрабатывать свои алгоритмы машинного обучения?
 - Это нужно очень редко, поскольку существующие алгоритмы хорошо реализованы и задокументированы



Зачем нужно машинное обучение

- Далее мы обсудим, какие бывают типы алгоритмов машинного обучения!



Типы алгоритмов машинного обучения



Типы алгоритмов машинного обучения

- В ближайших разделах мы изучим два типа алгоритмов машинного обучения:
 - Supervised Learning – обучение с учителем
 - Unsupervised Learning – обучение без учителя



Типы алгоритмов машинного обучения

- Supervised Learning – обучение с учителем
 - Используются размеченные данные или исторические данные, на основе которых делается предсказание целевой переменной
- Unsupervised Learning – обучение без учителя
 - Применяется к неразмеченным данным. Модель машинного обучения ищет возможные закономерности в данных



Типы алгоритмов машинного обучения

- Supervised Learning – обучение с учителем
 - Требуются исторические размеченные данные
 - Исторические – известные результаты из моментов в прошлом
 - Размеченные – значения целевой переменной были известны



Типы алгоритмов машинного обучения

- Supervised Learning – обучение с учителем
 - Целевые переменные бывают двух типов
 - Нужно предсказать категориальную переменную – это задача классификации
 - Нужно предсказать непрерывную переменную – это задача регрессии



Типы алгоритмов машинного обучения

- Supervised Learning – обучение с учителем
 - Задачи классификации
 - Предсказать категорию
 - Доброкачественная или раковая опухоль
 - Погашение кредита или неплатежи
 - Категория изображения – например, конкретная буква в задаче распознавания текста, написанного от руки



Типы алгоритмов машинного обучения

- Supervised Learning – обучение с учителем
 - Задачи регрессии
 - Предсказать значение непрерывной переменной
 - Будущие цены
 - Нагрузка на электросети
 - Результаты тестов



Типы алгоритмов машинного обучения

- Unsupervised Learning – обучение без учителя
 - Сгруппировать неразмеченные данные
 - Пример:
 - Сгруппировать клиентов на группы, используя их поведенческие характеристики



Типы алгоритмов машинного обучения

- Unsupervised Learning – обучение без учителя
 - Основная сложность в том, что у нас нет исторических “правильных” данных, и поэтому намного сложнее оценить правильность работы модели машинного обучения



Типы алгоритмов машинного обучения

- Разделы по машинному обучению
 - Мы начнём с алгоритмов обучения с учителем (supervised learning), чтобы изучить возможности различных алгоритмов машинного обучения
 - Далее мы перейдём к алгоритмам обучения без учителя (unsupervised learning), для кластеризации данных и уменьшения размерности данных



Типы алгоритмов машинного обучения

- Прежде чем заняться написанием кода и линейной регрессией в следующем разделе, давайте детально посмотрим на весь процесс машинного обучения с учителем (supervised learning), чтобы подготовиться к следующим разделам!



Процесс машинного обучения с учителем



Этапы работ по машинному обучению



Реальный
мир



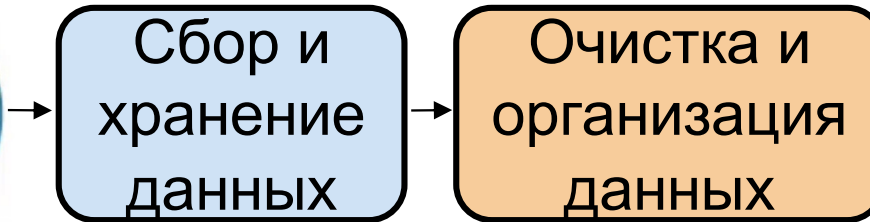
Сбор и
хранение
данных



Этапы работ по машинному обучению



Реальный
мир

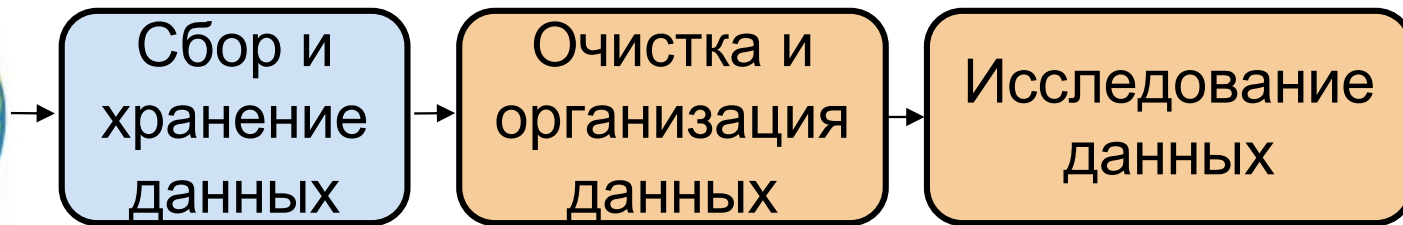




Этапы работ по машинному обучению



Реальный
мир





Этапы работ по машинному обучению



Реальный
мир

Сбор и
хранение
данных

Очистка и
организация
данных

Исследование
данных

Модели
Машинного
Обучения

Supervised Learning:

Сделать предсказание

Unsupervised Learning:

Найти закономерности
в данных



Этапы работ по машинному обучению



Реальный
мир

Сбор и
хранение
данных

Очистка и
организация
данных

Исследование
данных

Модели
Машинного
Обучения

Jupyter Notebook, NumPy, Pandas, Matplotlib, Seaborn

Scikit-learn



Этапы работ по машинному обучению



Реальный
мир

Сбор и
хранение
данных

Очистка и
организация
данных

Исследование
данных

Модели
Машинного
Обучения

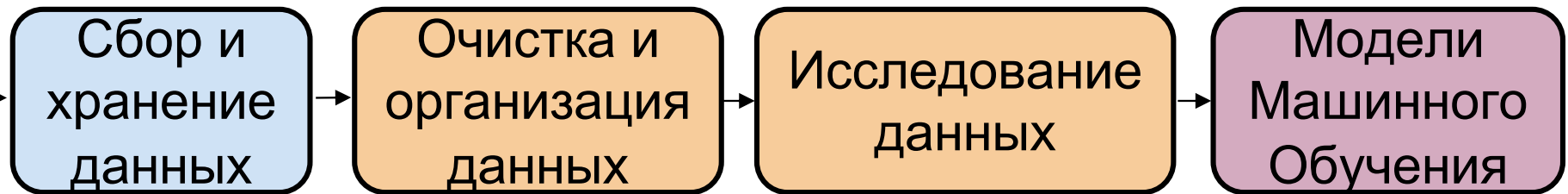
Supervised Learning:
Сделать предсказание



Этапы работ по машинному обучению



Реальный
мир



Supervised Learning:
Сделать предсказание

Например: предсказать цену продажи дома



Процесс машинного обучения с учителем

- Начинаем со сбора и структуризации исторических данных
- Предыдущие продажи домов - это размеченные данные

Area m ²	Bedrooms	Bathrooms	Price
200	3	2	\$500,000
190	2	1	\$450,000
230	3	3	\$650,000
180	1	1	\$400,000
210	2	2	\$550,000



Процесс машинного обучения с учителем

- Задача – предсказать цену продажи для нового дома, с известными значениями площади, кол-ва спален и кол-ва санузлов

Area m ²	Bedrooms	Bathrooms	Price
200	3	2	\$500,000
190	2	1	\$450,000
230	3	3	\$650,000
180	1	1	\$400,000
210	2	2	\$550,000



Процесс машинного обучения с учителем

- Продукт на основе данных: на входе параметры дома, на выходе предсказание цены (на основе исторических данных)

Area m ²	Bedrooms	Bathrooms	Price
200	3	2	\$500,000
190	2	1	\$450,000
230	3	3	\$650,000
180	1	1	\$400,000
210	2	2	\$550,000



Этапы работ по машинному обучению

- Задача предсказания цены дома



Реальный
мир

Сбор и
хранение
данных

Очистка и
организация
данных

Исследование
данных

Модели
Машинного
Обучения

Supervised Learning:
Сделать предсказание



Этапы работ по машинному обучению

- Задача предсказания цены дома

Модели
Машинного
Обучения

Supervised Learning:
Сделать предсказание



Этапы работ по машинному обучению

- Задача предсказания цены дома

Модели Машинного Обучения

Supervised Learning:

Сделать предсказание

Данные

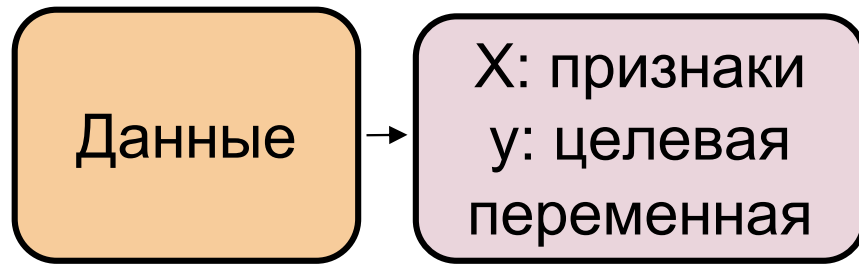


Процесс машинного обучения с учителем (supervised learning)

Данные



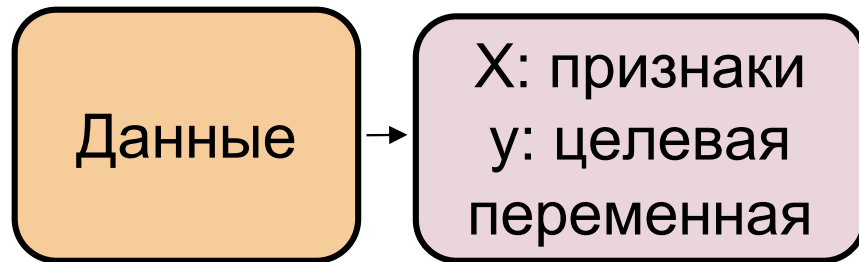
Процесс машинного обучения с учителем (supervised learning)





Процесс машинного обучения с учителем (supervised learning)

- Целевая переменная – значение, которое нам нужно предсказать

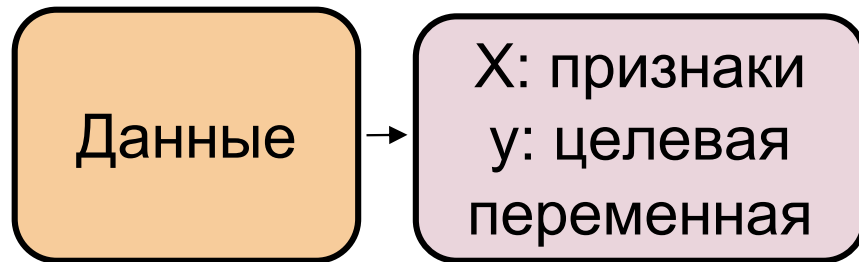


Area m ²	Bedrooms	Bathrooms	Price
200	3	2	\$500,000
190	2	1	\$450,000
230	3	3	\$650,000
180	1	1	\$400,000
210	2	2	\$550,000



Процесс машинного обучения с учителем (supervised learning)

- Целевая переменная – значение, которое нам нужно предсказать

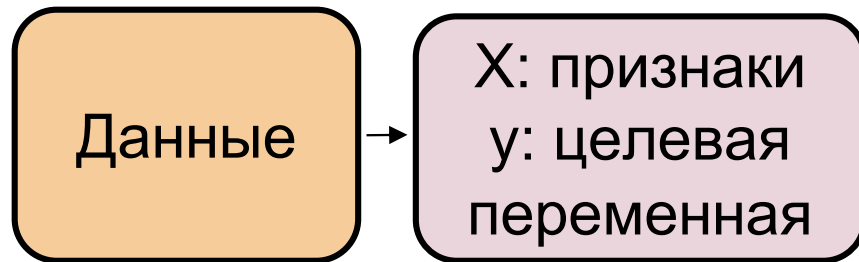


Area m ²	Bedrooms	Bathrooms	Price
200	3	2	\$500,000
190	2	1	\$450,000
230	3	3	\$650,000
180	1	1	\$400,000
210	2	2	\$550,000



Процесс машинного обучения с учителем (supervised learning)

- Признаки (features) – известные характеристики объекта

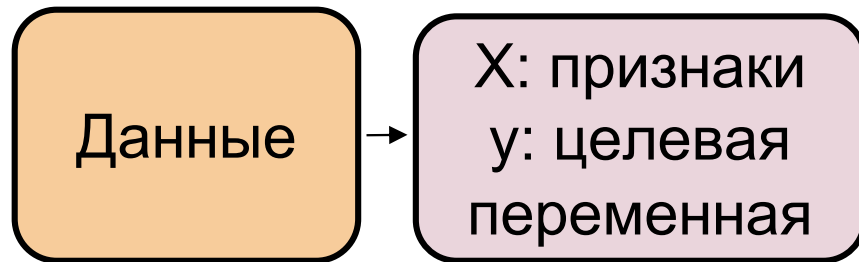


Area m ²	Bedrooms	Bathrooms	Price
200	3	2	\$500,000
190	2	1	\$450,000
230	3	3	\$650,000
180	1	1	\$400,000
210	2	2	\$550,000



Процесс машинного обучения с учителем (supervised learning)

- Признаки и целевая переменная определяются в соответствии с решаемой задачей

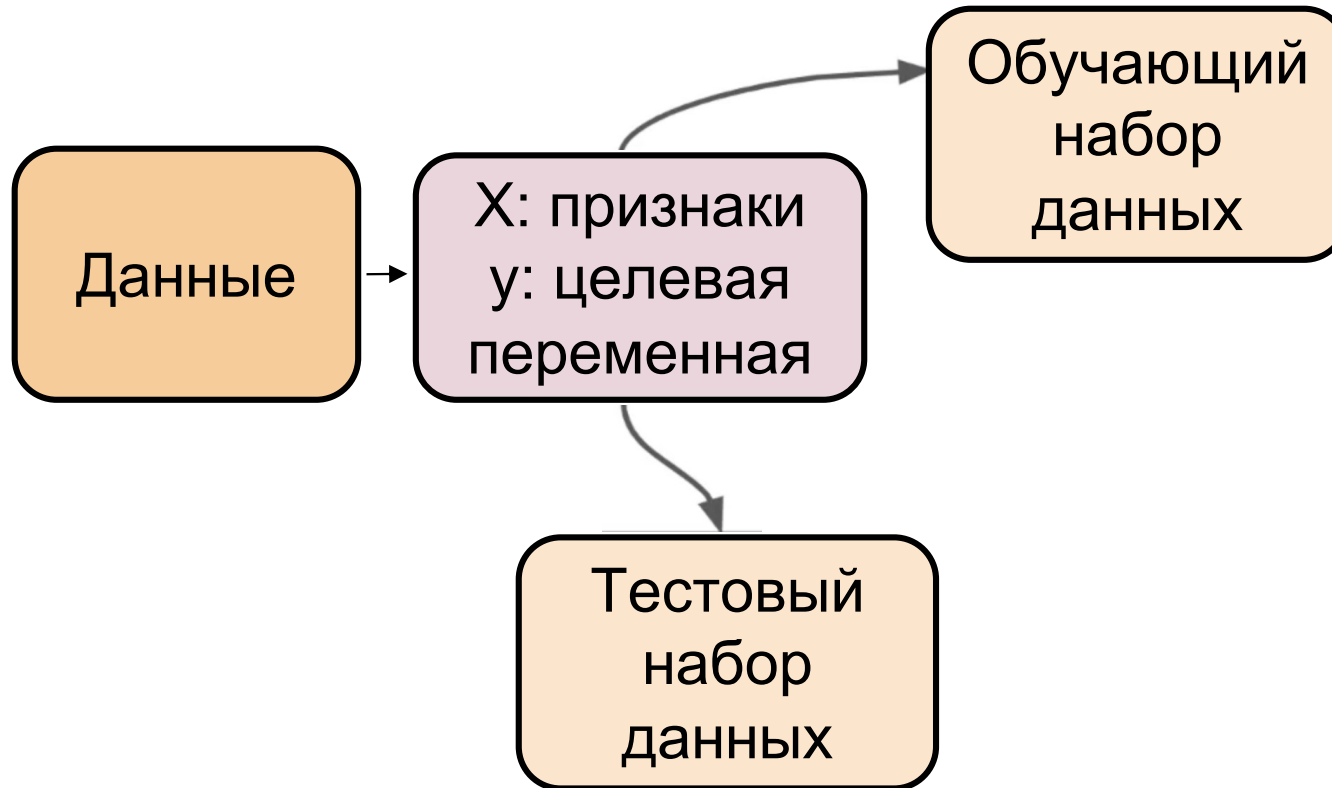


Area m ²	Bedrooms	Bathrooms	Price
200	3	2	\$500,000
190	2	1	\$450,000
230	3	3	\$650,000
180	1	1	\$400,000
210	2	2	\$550,000



Процесс машинного обучения с учителем (supervised learning)

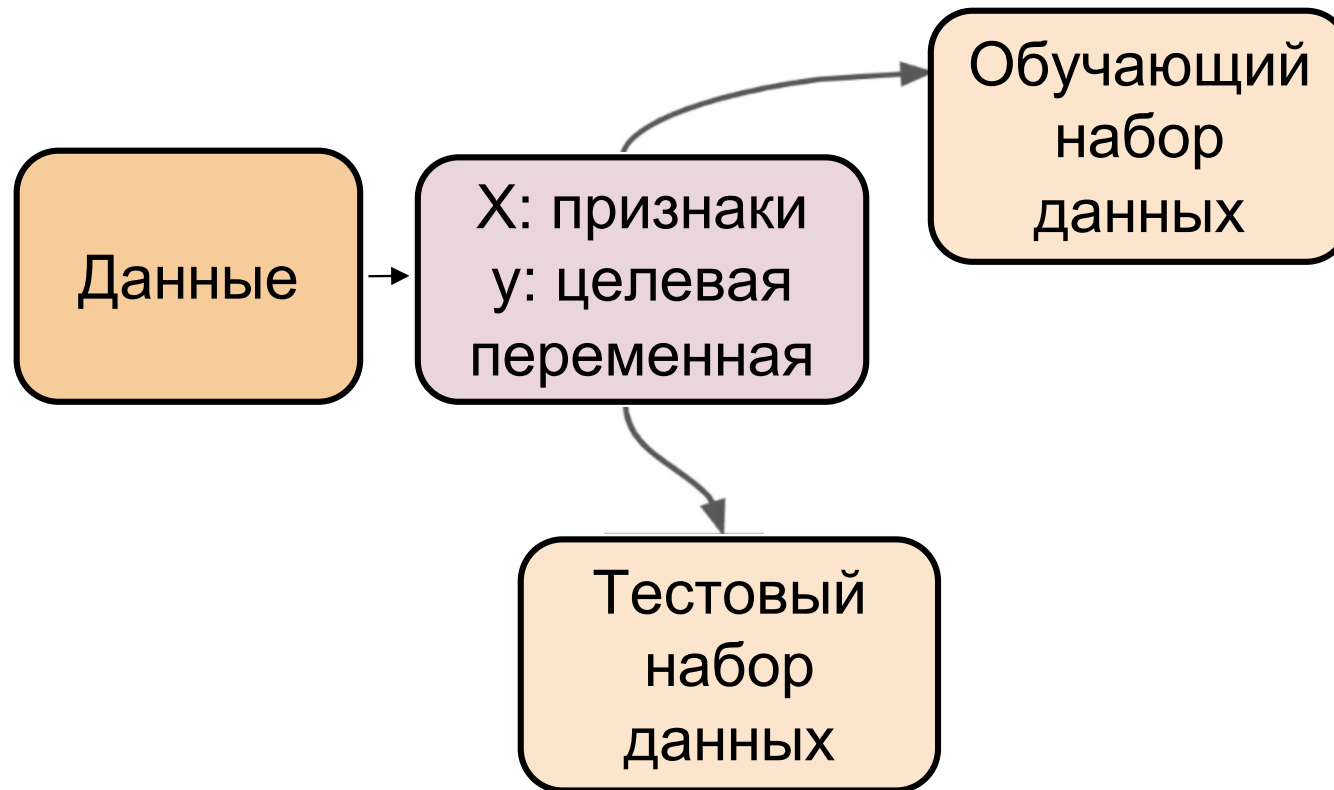
- Разбиваем данные на обучающий и тестовый наборы данных





Процесс машинного обучения с учителем (supervised learning)

- Разбиваем данные на обучающий и тестовый наборы данных

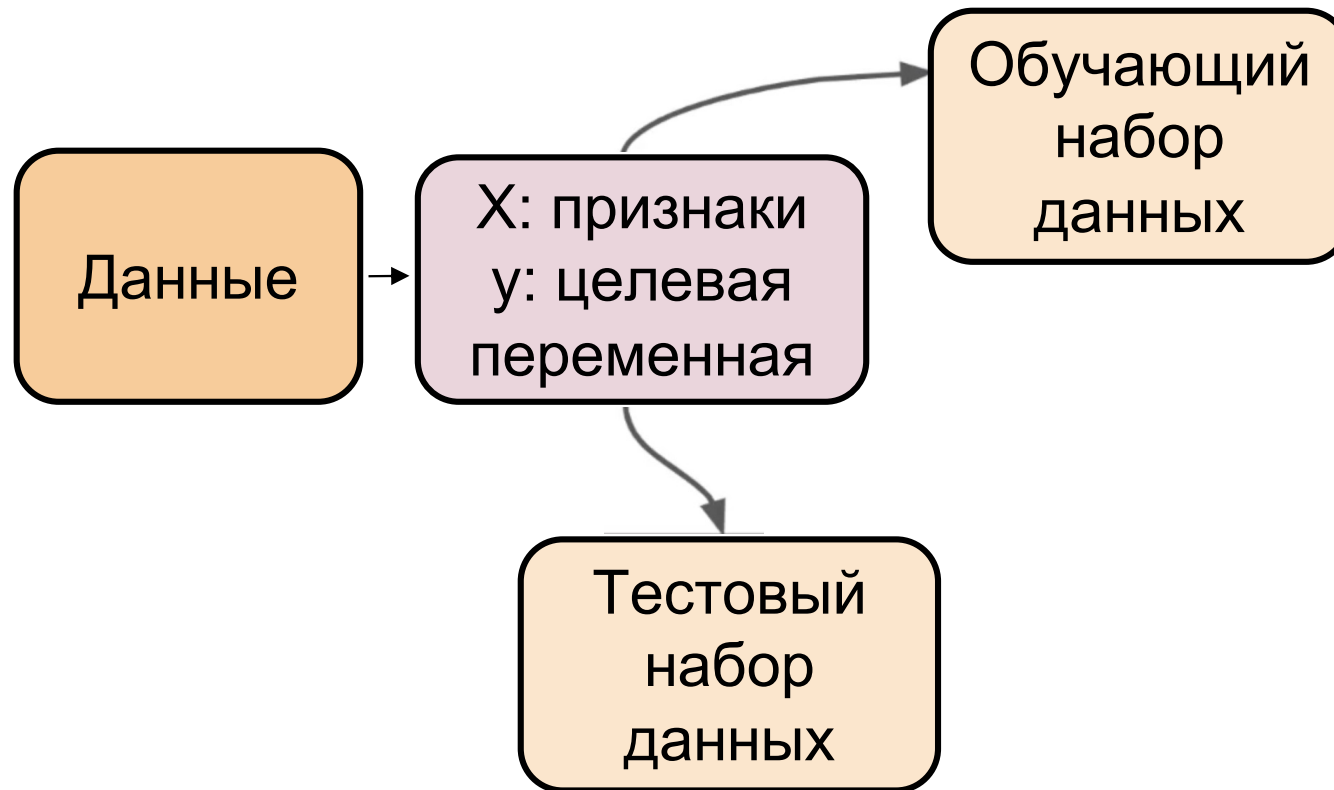


- Позже мы обсудим кросс-валидацию



Процесс машинного обучения с учителем (supervised learning)

- Зачем разбивать данные? Как это делать?





Процесс машинного обучения с учителем (supervised learning)

- Зачем разбивать данные? Как это делать?

Area m ²	Bedrooms	Bathrooms	Price
200	3	2	\$500,000
190	2	1	\$450,000
230	3	3	\$650,000
180	1	1	\$400,000
210	2	2	\$550,000



Процесс машинного обучения с учителем (supervised learning)

- Как бы Вы оценили работу обычного риэлтора?



Area m ²	Bedrooms	Bathrooms	Price
200	3	2	\$500,000
190	2	1	\$450,000
230	3	3	\$650,000
180	1	1	\$400,000
210	2	2	\$550,000



Процесс машинного обучения с учителем (supervised learning)

- Как бы Вы оценили работу обычного риелтора?
- Вы могли бы попросить риелтора посмотреть на исторические данные...



Area m ²	Bedrooms	Bathrooms	Price
200	3	2	\$500,000
190	2	1	\$450,000
230	3	3	\$650,000
180	1	1	\$400,000
210	2	2	\$550,000



Процесс машинного обучения с учителем (supervised learning)

- Затем дать риелтору характеристики нового дома, и попросить сделать оценку цены



Area m ²	Bedrooms	Bathrooms	Price
200	3	2	\$500,000
190	2	1	\$450,000
230	3	3	\$650,000
180	1	1	\$400,000
210	2	2	\$550,000



Процесс машинного обучения с учителем (supervised learning)

- Но как понять, насколько оценка цены будет корректна?



Area m ²	Bedrooms	Bathrooms	Price
200	3	2	\$500,000
190	2	1	\$450,000
230	3	3	\$650,000
180	1	1	\$400,000
210	2	2	\$550,000



Процесс машинного обучения с учителем (supervised learning)

- Но как понять, насколько оценка цены будет корректна?
Какой дом выбрать для тестирования оценки?



Area m ²	Bedrooms	Bathrooms	Price
200	3	2	\$500,000
190	2	1	\$450,000
230	3	3	\$650,000
180	1	1	\$400,000
210	2	2	\$550,000



Процесс машинного обучения с учителем (supervised learning)

- Проверить оценку для нового дома нельзя, потому что его истинная цена продажи ещё неизвестна



Area m ²	Bedrooms	Bathrooms	Price
200	3	2	\$500,000
190	2	1	\$450,000
230	3	3	\$650,000
180	1	1	\$400,000
210	2	2	\$550,000



Процесс машинного обучения с учителем (supervised learning)

- Проверить оценку для имеющихся домов тоже нельзя, потому что риелтор мог просто запомнить эти данные



Area m ²	Bedrooms	Bathrooms	Price
200	3	2	\$500,000
190	2	1	\$450,000
230	3	3	\$650,000
180	1	1	\$400,000
210	2	2	\$550,000



Процесс машинного обучения с учителем (supervised learning)

- Поэтому следует разбить данные на обучающий и тестовый наборы – давайте посмотрим, зачем...



Area m ²	Bedrooms	Bathrooms	Price
200	3	2	\$500,000
190	2	1	\$450,000
230	3	3	\$650,000
180	1	1	\$400,000
210	2	2	\$550,000



Процесс машинного обучения с учителем (supervised learning)

- Теперь мы разбиваем эти данные на обучающий набор данных (TRAIN) и тестовый набор данных (TEST)

TRAIN

Area m ²	Bedrooms	Bathrooms	Price
200	3	2	\$500,000
190	2	1	\$450,000
230	3	3	\$650,000
180	1	1	\$400,000
210	2	2	\$550,000



Процесс машинного обучения с учителем (supervised learning)

- Теперь мы разбиваем эти данные на обучающий набор данных (TRAIN) и тестовый набор данных (TEST)

	Area m ²	Bedrooms	Bathrooms	Price
TRAIN	200	3	2	\$500,000
	190	2	1	\$450,000
	230	3	3	\$650,000
TEST	180	1	1	\$400,000
	210	2	2	\$550,000



Процесс машинного обучения с учителем (supervised learning)

- У нас получилось 4 компонента:

	Area m ²	Bedrooms	Bathrooms	Price	
X TRAIN	200	3	2	\$500,000	Y TRAIN
	190	2	1	\$450,000	
	230	3	3	\$650,000	
X TEST	180	1	1	\$400,000	Y TEST
	210	2	2	\$550,000	



Процесс машинного обучения с учителем (supervised learning)

- Итак, вернёмся к вопросу о том, как мы можем проверить работу риелтора...



Area m ²	Bedrooms	Bathrooms	Price
200	3	2	\$500,000
190	2	1	\$450,000
230	3	3	\$650,000
180	1	1	\$400,000
210	2	2	\$550,000



Процесс машинного обучения с учителем (supervised learning)

- Итак, вернёмся к вопросу о том, как мы можем проверить работу риелтора...



TRAIN

TEST

Area m ²	Bedrooms	Bathrooms	Price
200	3	2	\$500,000
190	2	1	\$450,000
230	3	3	\$650,000
180	1	1	\$400,000
210	2	2	\$550,000



Процесс машинного обучения с учителем (supervised learning)

- Мы можем дать риелтору только данные из обучающего набора данных (TRAIN) – как X , так и Y



TRAIN

Area m ²	Bedrooms	Bathrooms	Price
200	3	2	\$500,000
190	2	1	\$450,000
230	3	3	\$650,000



Процесс машинного обучения с учителем (supervised learning)

- После “обучения” на обучающем наборе данных, просим риелтора сделать оценки цены дома для тестовых данных



TEST

Area m ²	Bedrooms	Bathrooms
180	1	1
210	2	2



Процесс машинного обучения с учителем (supervised learning)

- После “обучения” на обучающем наборе данных, просим риелтора сделать оценки цены дома для тестовых данных
- Предоставляем только признаки X , просим оценить y



TEST

Area m ²	Bedrooms	Bathrooms
180	1	1
210	2	2



Процесс машинного обучения с учителем (supervised learning)

- После “обучения” на обучающем наборе данных, просим риелтора сделать оценки цены дома для тестовых данных
- Предоставляем только признаки X , просим оценить y
- Риелтор ещё не видел этих данных! И цену этих домов!



TEST

Area m ²	Bedrooms	Bathrooms
180	1	1
210	2	2



Процесс машинного обучения с учителем (supervised learning)

- Риелтор даёт свои оценки – предсказания цены (predictions)



Predictions	Area m ²	Bedrooms	Bathrooms
\$410,000	180	1	1
\$540,000	210	2	2



Процесс машинного обучения с учителем (supervised learning)

- Риелтор даёт свои оценки – предсказания цены (predictions)
- Далее мы добавляем истинные цены



Predictions	Area m ²	Bedrooms	Bathrooms	Price
\$410,000	180	1	1	\$400,000
\$540,000	210	2	2	\$550,000



Процесс машинного обучения с учителем (supervised learning)

- Риелтор даёт свои оценки – предсказания цены (predictions)
- Далее мы добавляем истинные цены
- И сравниваем предсказания цены и истинные цены



Predictions	Price
\$410,000	\$400,000
\$540,000	\$550,000



Процесс машинного обучения с учителем (supervised learning)

- Часто обозначения такие – оценка $\hat{y}$ и истинное значение y



$\hat{y}$	y
Predictions	Price
\$410,000	\$400,000
\$540,000	\$550,000



Процесс машинного обучения с учителем (supervised learning)

- Часто обозначения такие – оценка $\hat{y}$ и истинное значение y
- Позже мы обсудим различные методы для понимания того, насколько эти оценки хороши!

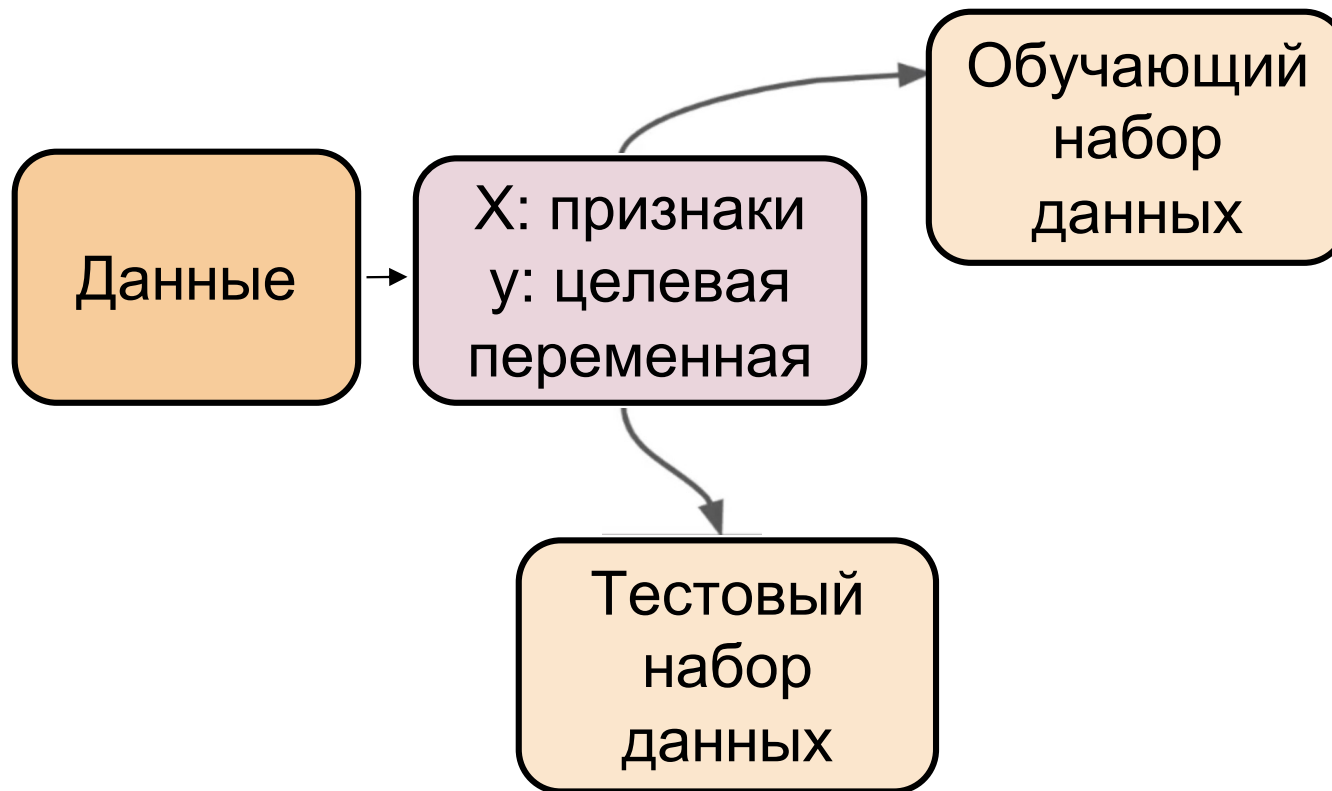


$\hat{y}$	y
Predictions	Price
\$410,000	\$400,000
\$540,000	\$550,000



Процесс машинного обучения с учителем (supervised learning)

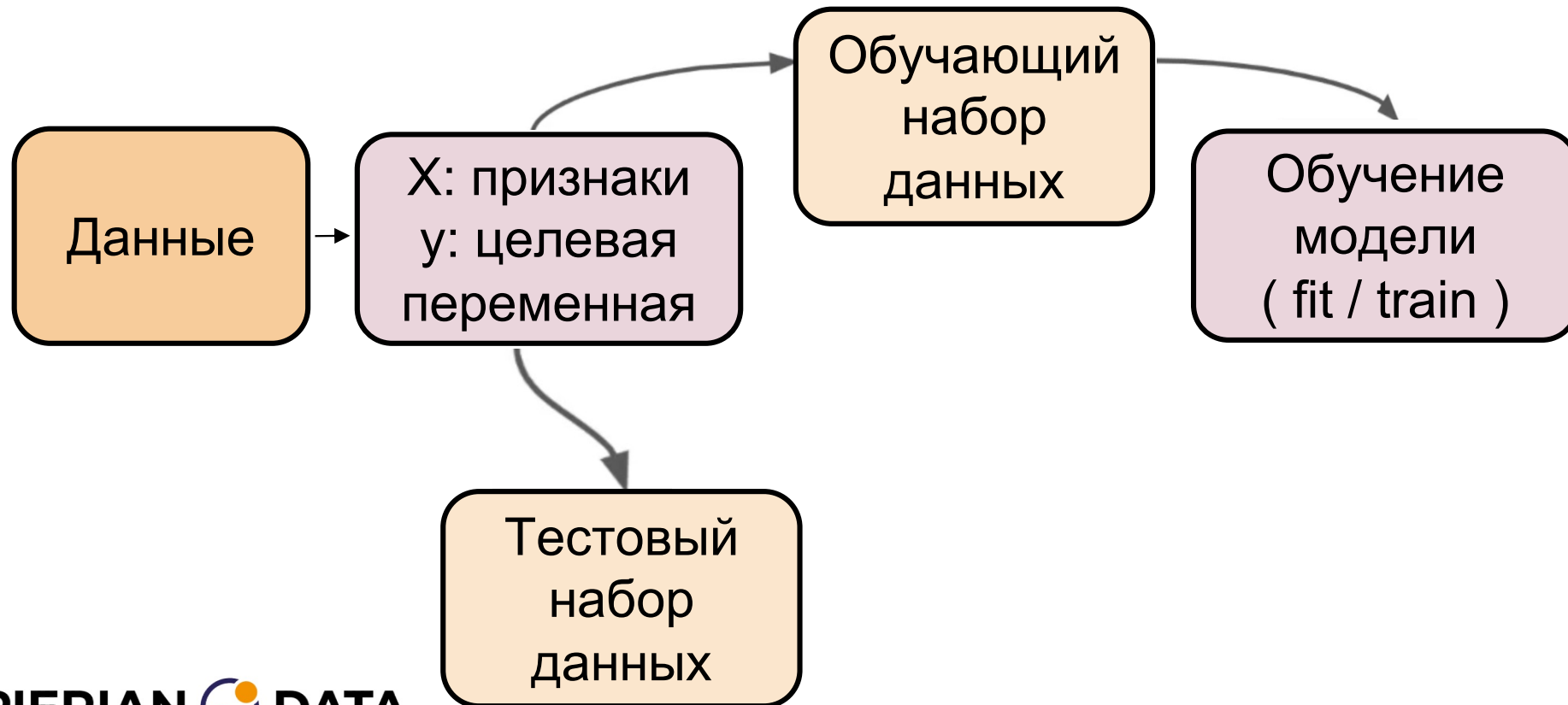
- Разбиение данных





Процесс машинного обучения с учителем (supervised learning)

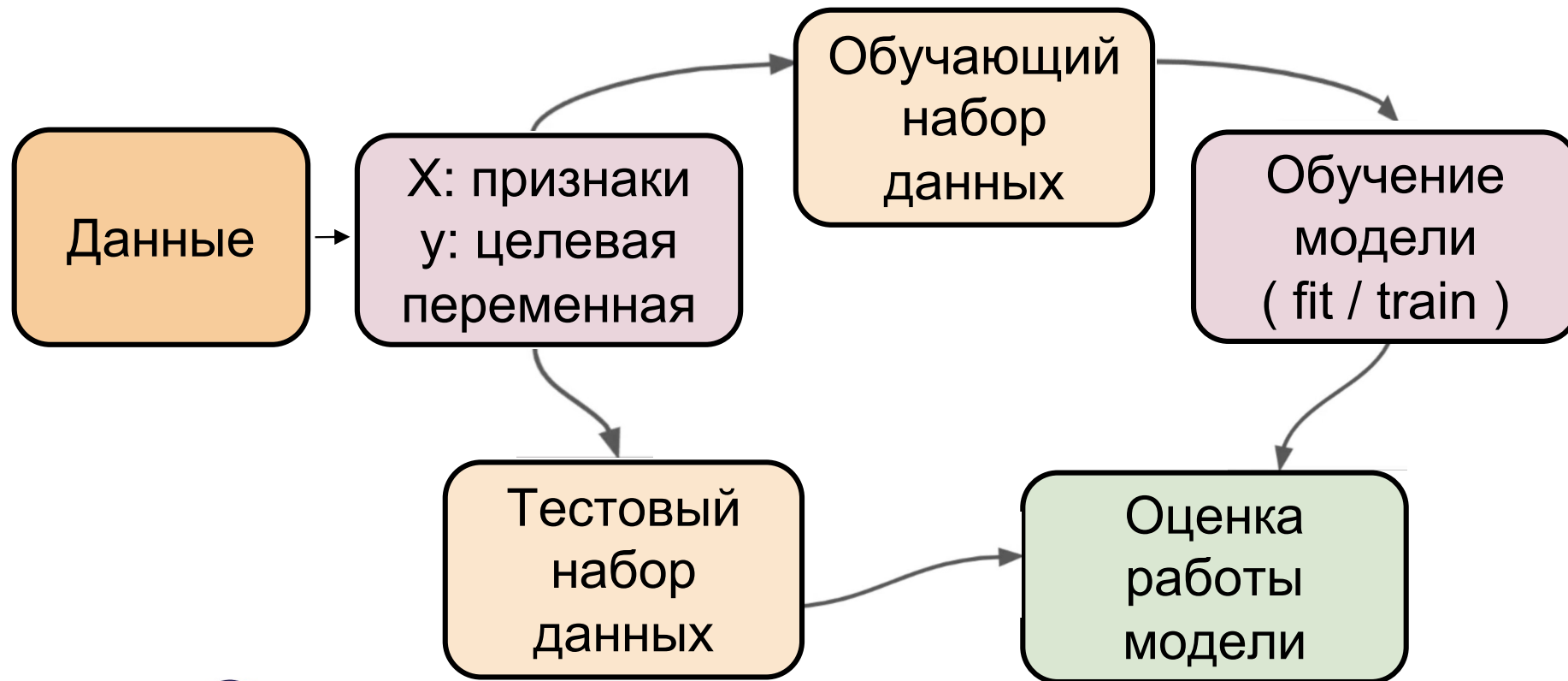
- Разбиение данных, обучение модели





Процесс машинного обучения с учителем (supervised learning)

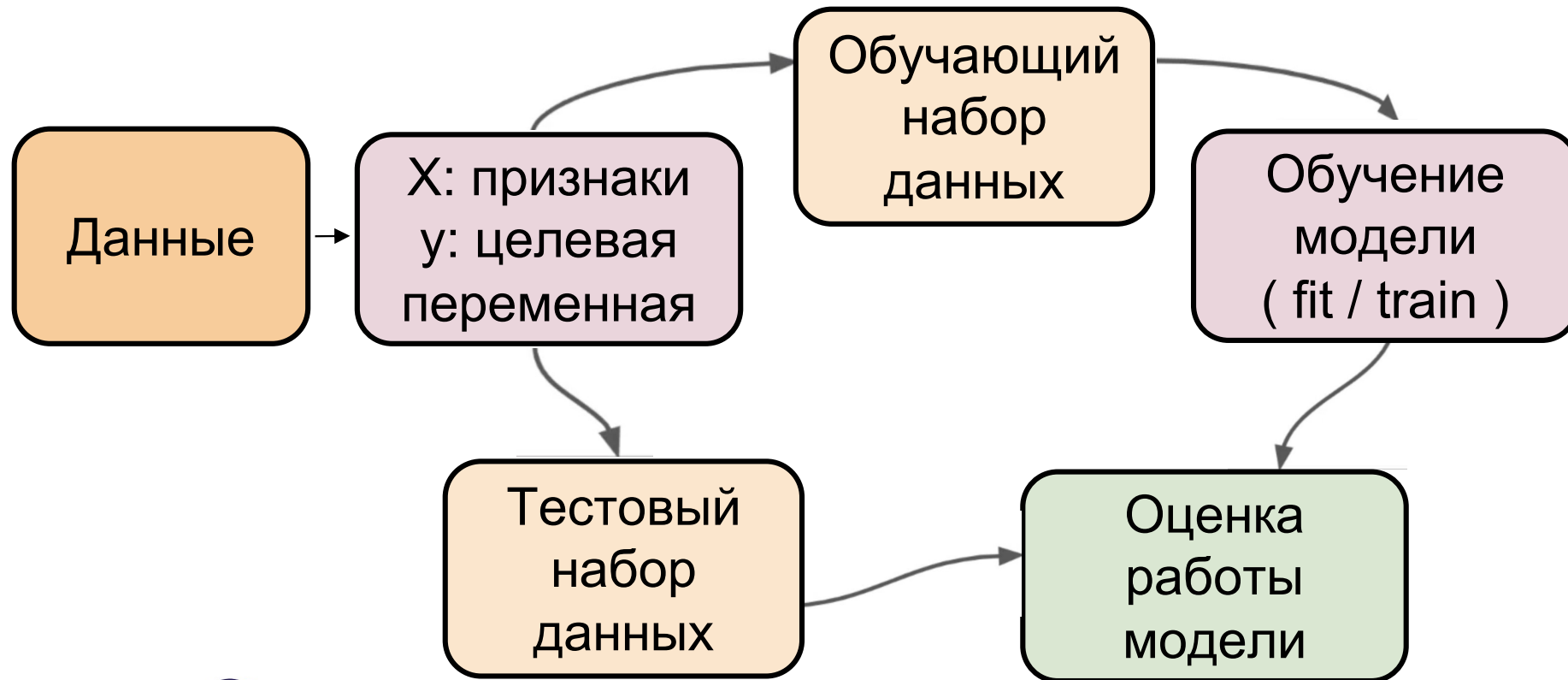
- Разбиение данных, обучение модели, оценка работы модели





Процесс машинного обучения с учителем (supervised learning)

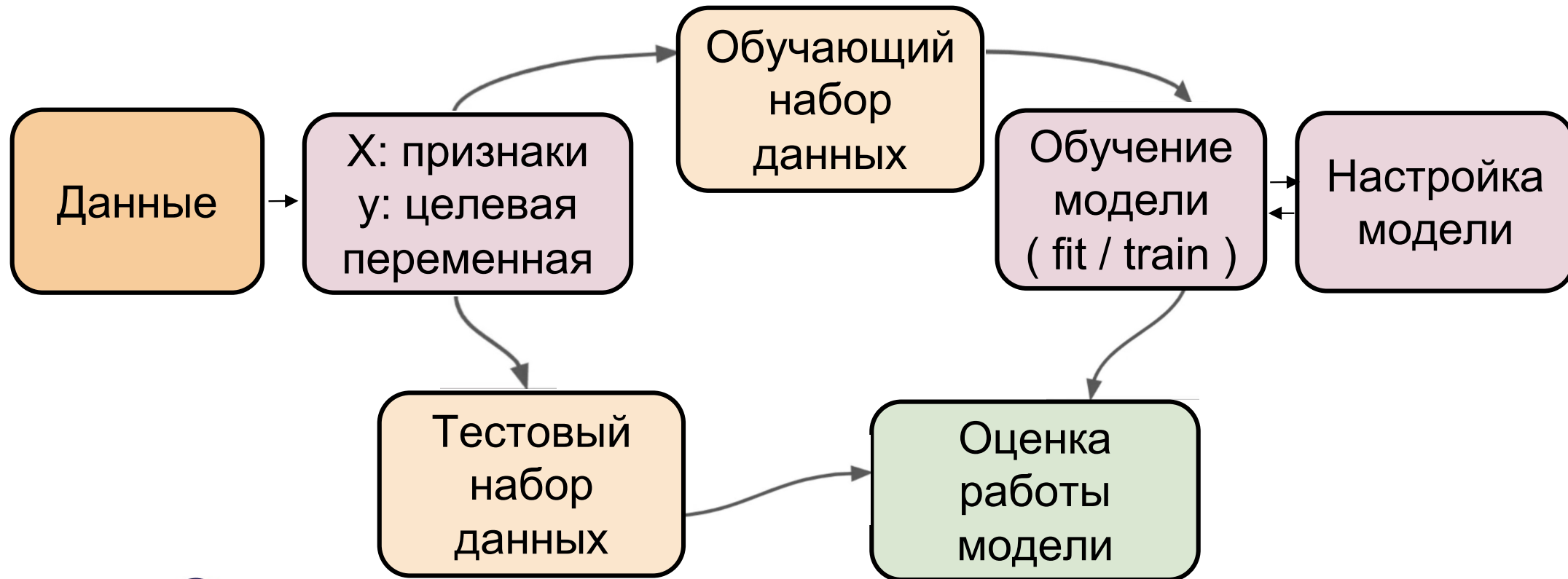
- Что делать, если оценка работы модели не устраивает?





Процесс машинного обучения с учителем (supervised learning)

- Мы можем настроить гиперпараметры модели





Процесс машинного обучения с учителем (supervised learning)

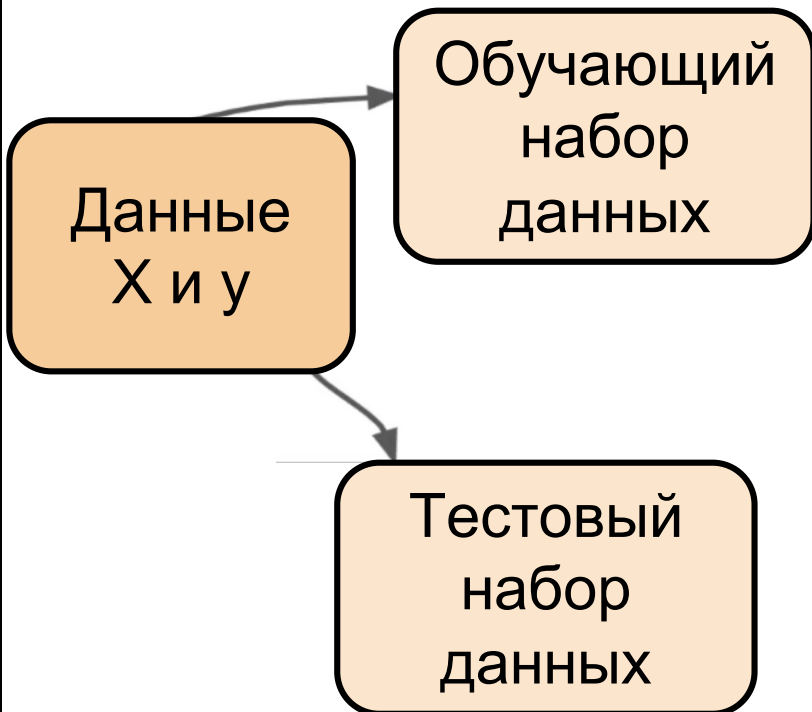
- Давайте быстро повторим весь процесс
- Получаем данные - признаки X и целевую переменную y

Данные
 X и y



Процесс машинного обучения с учителем (supervised learning)

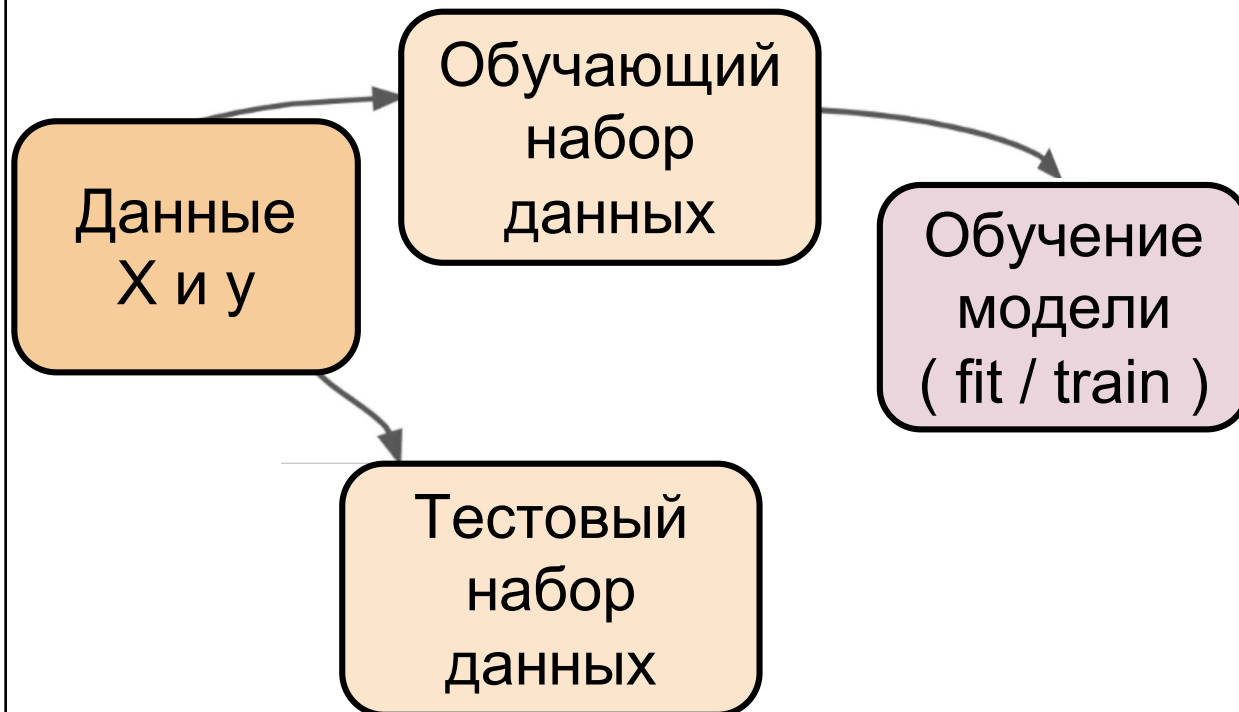
- Разбиваем данные на обучающий и тестовый наборы данных





Процесс машинного обучения с учителем (supervised learning)

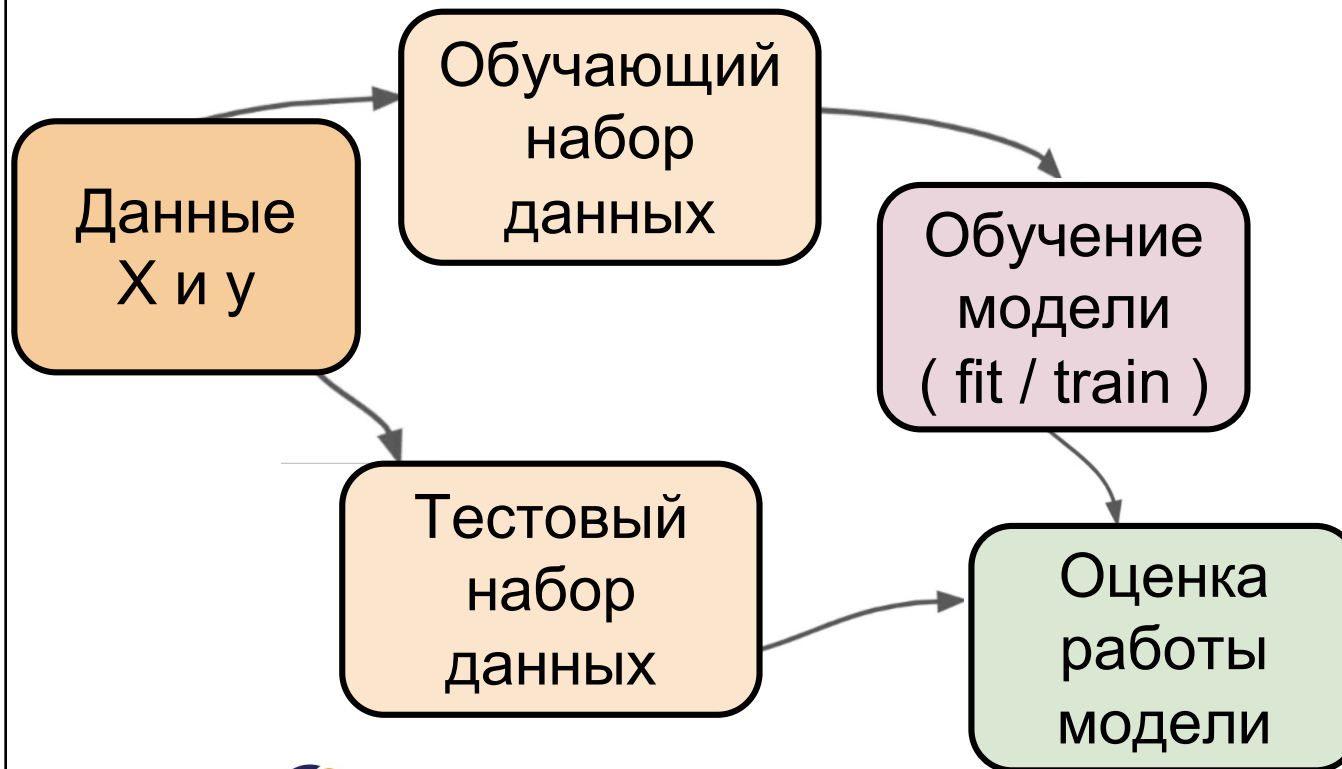
- Обучаем модель на обучающем наборе данных





Процесс машинного обучения с учителем (supervised learning)

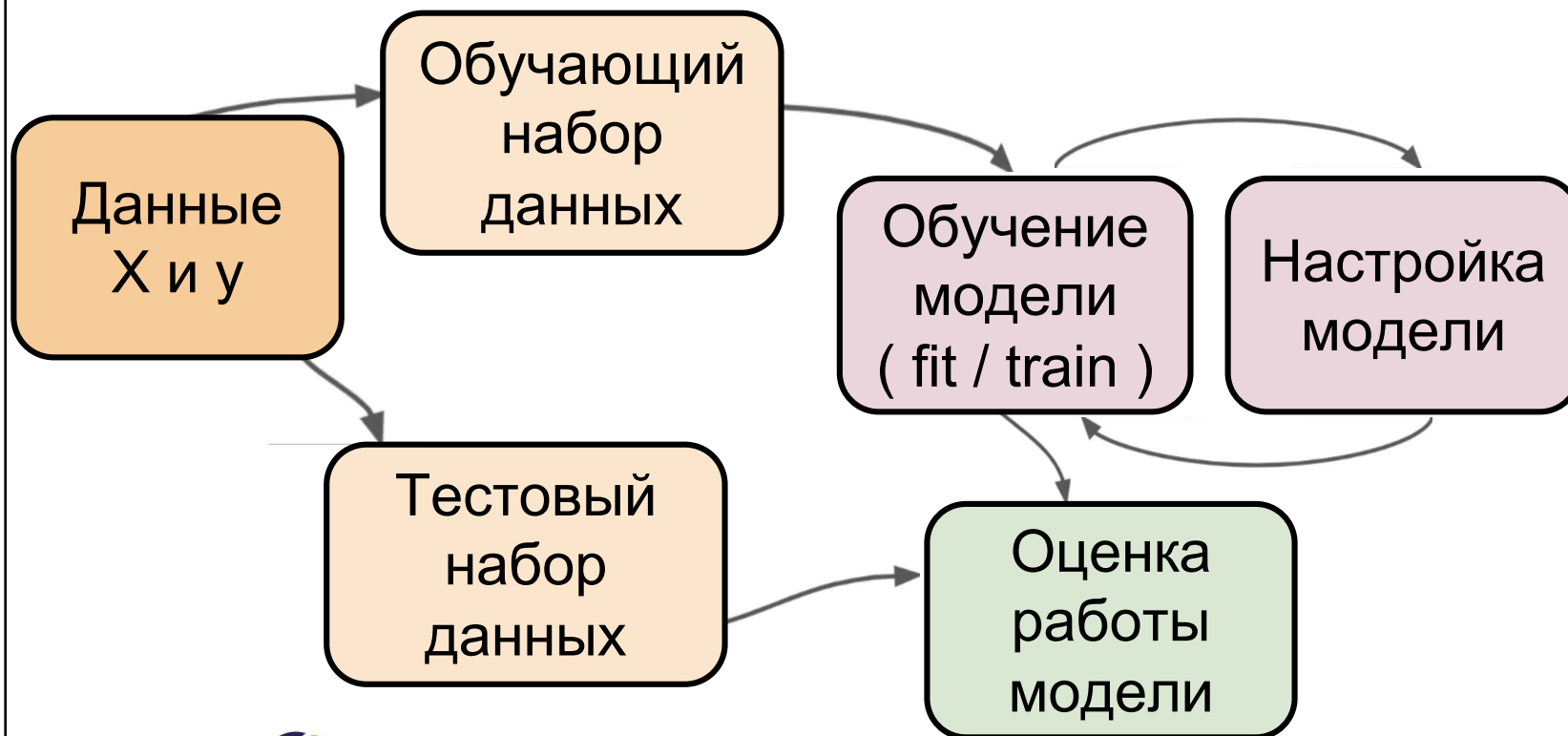
- Оцениваем работу модели (evaluate model performance)





Процесс машинного обучения с учителем (supervised learning)

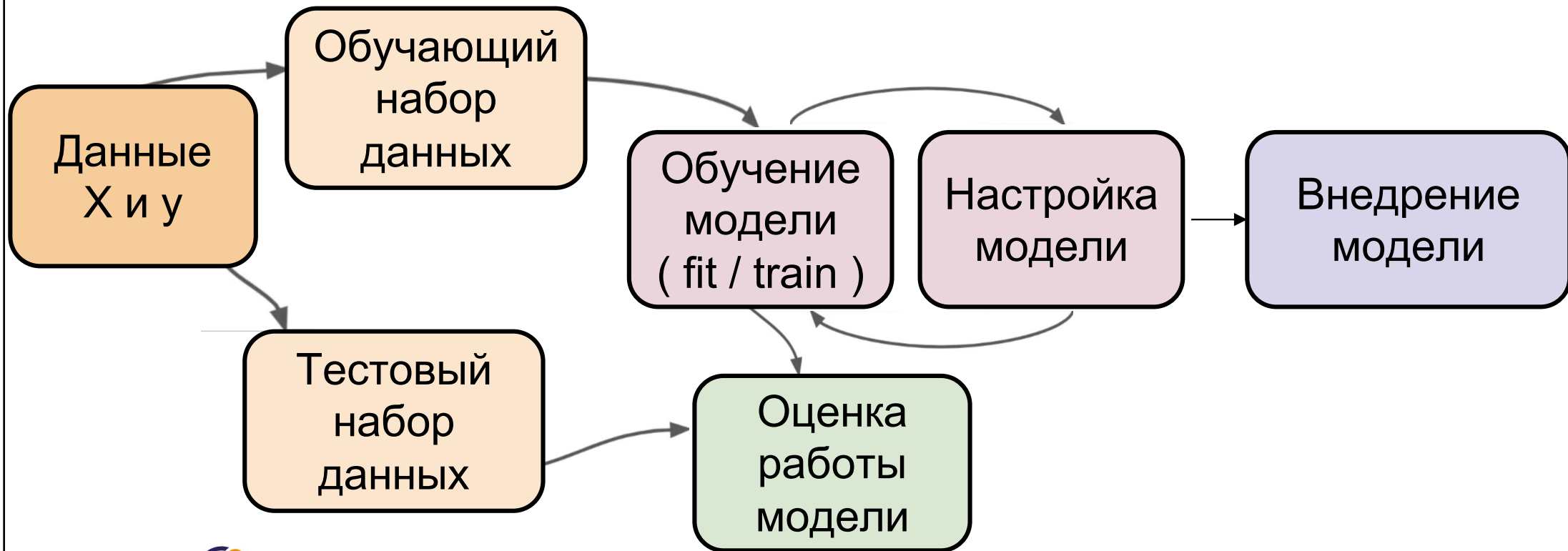
- Настраиваем гиперпараметры модели (несколько раз)





Процесс машинного обучения с учителем (supervised learning)

- Внедрение модели (deploy)





Этапы работ по машинному обучению

- Процесс машинного обучения с учителем (supervised learning)



Реальный
мир

Сбор и
хранение
данных

Очистка и
организация
данных

Исследование
данных

Модели
Машинного
Обучения

Supervised Learning:
Сделать предсказание



Этапы работ по машинному обучению

