



IT-ОБРАЗОВАНИЕ
В ПЕТЕРБУРГЕ
И УДАЛЕННО

Пайплайн ML

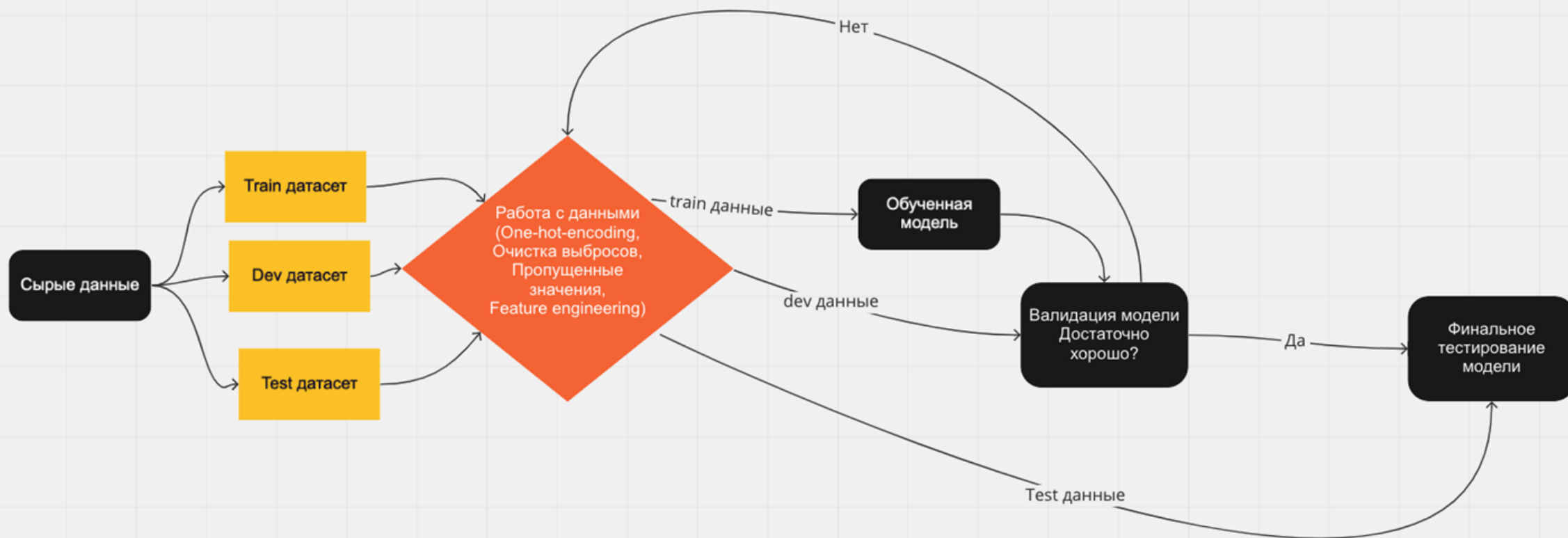
Преподаватель: Зубоченко Антон

Пайплайн машинного обучения

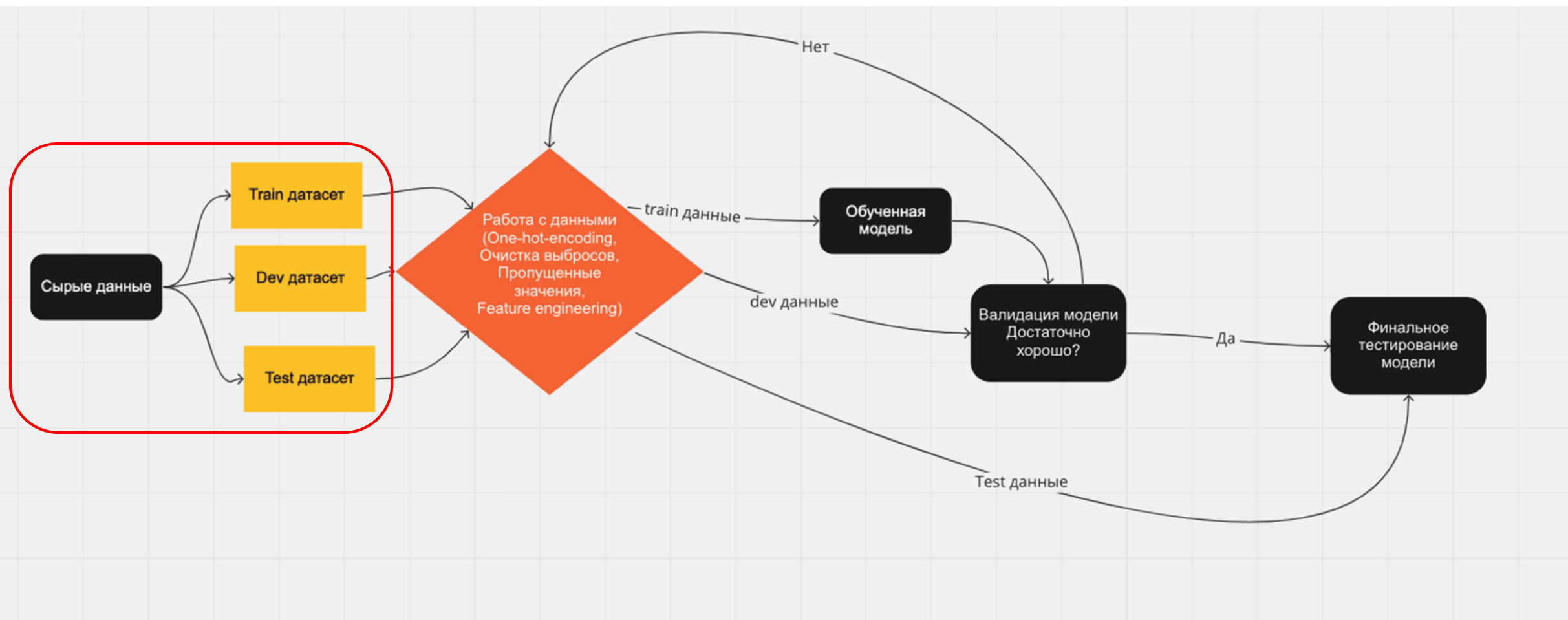
План занятия

- Схема создания алгоритма машинного обучения
 - Обработка данных
 - Кросс-валидация

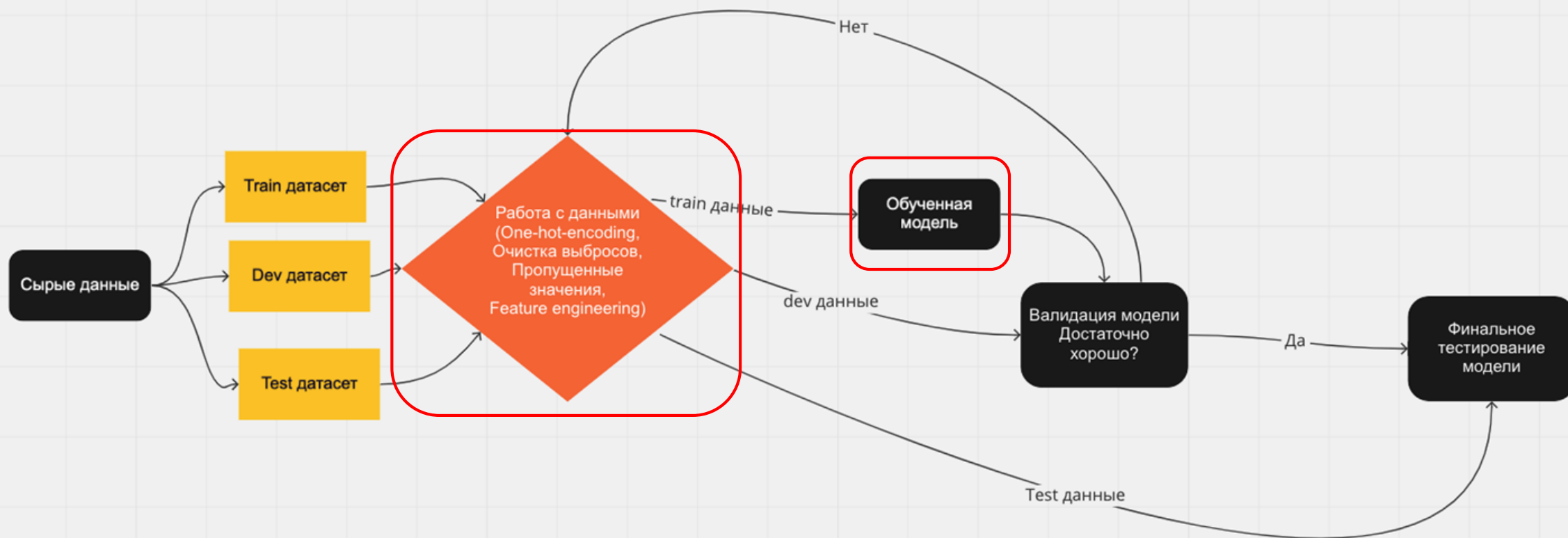
Текущий пайплайн МЛ



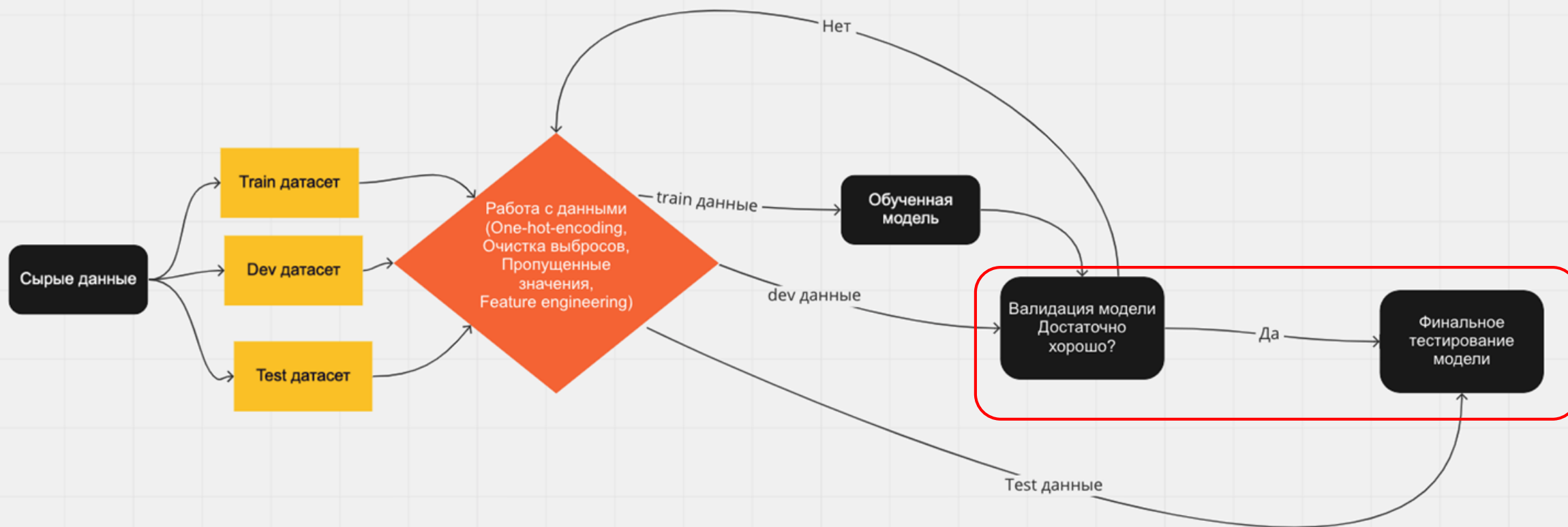
Текущий пайплайн МЛ



Текущий пайплайн МЛ



Текущий пайплайн МЛ



Базовая обработка обучающих данных



Утечки данных (Data leakage)

- Задача классификации на больных/не больных пневмонией

got_pneumonia	age	weight	male	Принимает антибиотики	
False	65	100	False	False	...
False	72	130	True	False	...
True	58	100	False	True	...

Утечки данных (Data leakage)

- Задача классификации на больных/не больных пневмонией
- Колонка “принимает антибиотики” почти полностью определяет ответ.
- Модель будет давать 100% качество на валидационном и тестовом датасете, но не применима в реальности.

got_pneumonia	age	weight	male	Принимает антибиотики	
False	65	100	False	False	...
False	72	130	True	False	...
True	58	100	False	True	...

Утечки данных (Data leakage)

- Задача классификации на больных/не больных пневмонией
- Колонка “принимает антибиотики” почти полностью определяет ответ.
- Модель будет давать 100% качество на валидационном и тестовом датасете, но не применима в реальности.

- | got_pneumonia | age | weight | male | Принимает антибиотики | ... |
|---------------|-----|--------|-------|-----------------------|-----|
| False | 65 | 100 | False | False | ... |
| False | 72 | 130 | True | False | ... |
| True | 58 | 100 | False | True | ... |

Пропущенные значения

- Задача классификации пассажиров Титаника
- Пропущенные значения могут быть разного вида
- Большинство алгоритмов не умеют работать с пропущенными

Missing values

PassengerId	Survived	Pclass	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	male	22	1	0	A/5 21171	7.25	Not known	S
2	1	1	female	38	1	0	PC 17599	71.2833	C85	C
3	1	3	female	26	0	0	STON/O2. 3101282	7.925	Not known	S
4	1	1	female	35	1	0	113803	53.1	C123	S
5	0	3	male	35	0	0	373450	8.05	Not known	S
6	0	3	male		0	0	330877	8.4583	Not known	Q

Численные пропущенные значения

Для численных признаков:

- Заполнение средним/медианой по признаку

Missing values



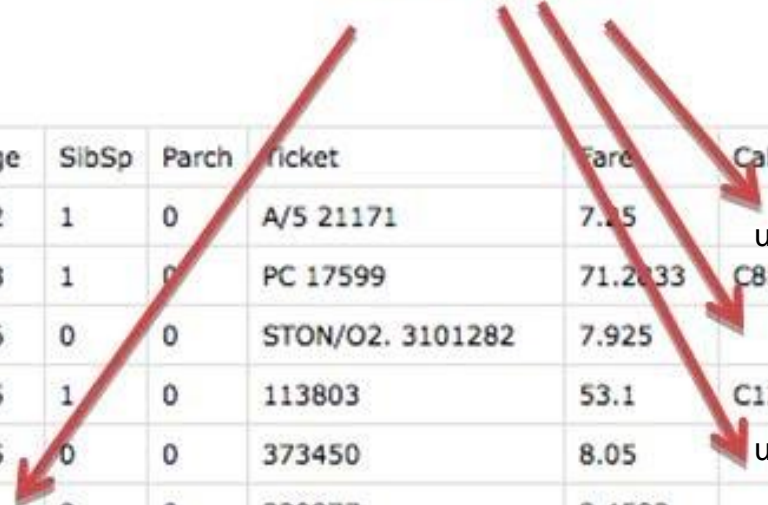
PassengerId	Survived	Pclass	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	male	22	1	0	A/5 21171	7.25		S
2	1	1	female	38	1	0	PC 17599	71.2833	C85	C
3	1	3	female	26	0	0	STON/O2. 3101282	7.925		S
4	1	1	female	35	1	0	113803	53.1	C123	S
5	0	3	male	35	0	0	373450	8.05		S
6	0	3	male	31.2	0	0	330877	8.4583		Q

Категориальные пропущенные значения

Для категориальных признаков:

- Заполнение самым частым значением признака
- Заполнение новым значением

Missing values



PassengerId	Survived	Pclass	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	male	22	1	0	A/5 21171	7.25	unknown	S
2	1	1	female	38	1	0	PC 17599	71.2833	C85	C
3	1	3	female	26	0	0	STON/O2. 3101282	7.925	unknown	S
4	1	1	female	35	1	0	113803	53.1	C123	S
5	0	3	male	35	0	0	373450	8.05	unknown	S
6	0	3	male		0	0	330877	8.4583	unknown	Q

Пропущенные значения

- Много пропущенных значений → выбрасываем колонку
- Мало пропущенных значений → выбрасываем строки с пропущенными значениями

Missing values

PassengerId	Survived	Pclass	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	male	22	1	0	A/5 21171	7.25		S
2	1	1	female	38	1	0	PC 17599	71.2833	C85	C
3	1	3	female	26	0	0	STON/O2. 3101282	7.925		S
4	1	1	female	35	1	0	113803	53.1	C123	S
5	0	3	male	35	0	0	373450	8.05		S
6	0	3	male		0	0	330877	8.4583		Q

One-hot-encoding



Признаки до обработки

Дубай	[1, 0, 0]
Москва	[0, 1, 0]
Амстердам	[0, 0, 1]

Признаки после обработки

One-hot-encoding

	City	Weight	Name
0	0	88	0
1	1	110	1
2	2	-10	2
3	1	56	3
4	1	43	4

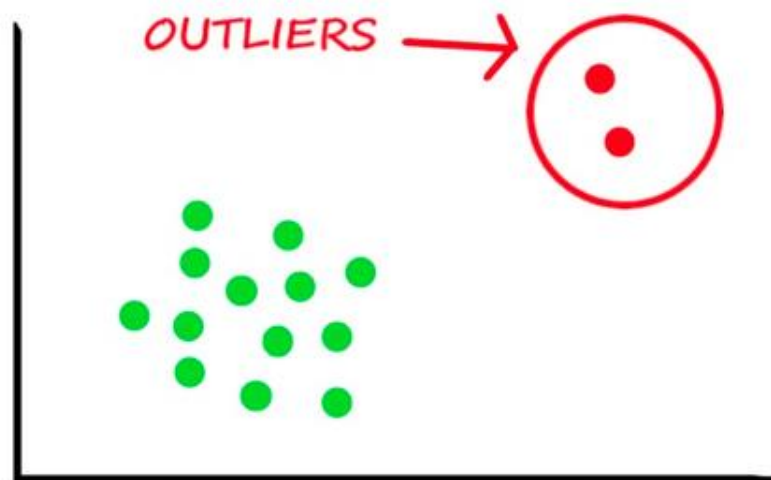
Признаки до обработки

	Weight	City_0	City_1	City_2	Name_0	Name_1	Name_2	Name_3	Name_4
0	88	1	0	0	1	0	0	0	0
1	110	0	1	0	0	1	0	0	0
2	-10	0	0	1	0	0	1	0	0
3	56	0	1	0	0	0	0	1	0
4	43	0	1	0	0	0	0	0	1

Признаки после обработки

Выбросы

- Выбросы — это ошибки в данных или экстраординарные события
- Выбросы могут повлиять на работу алгоритма



Очистка от выбросов

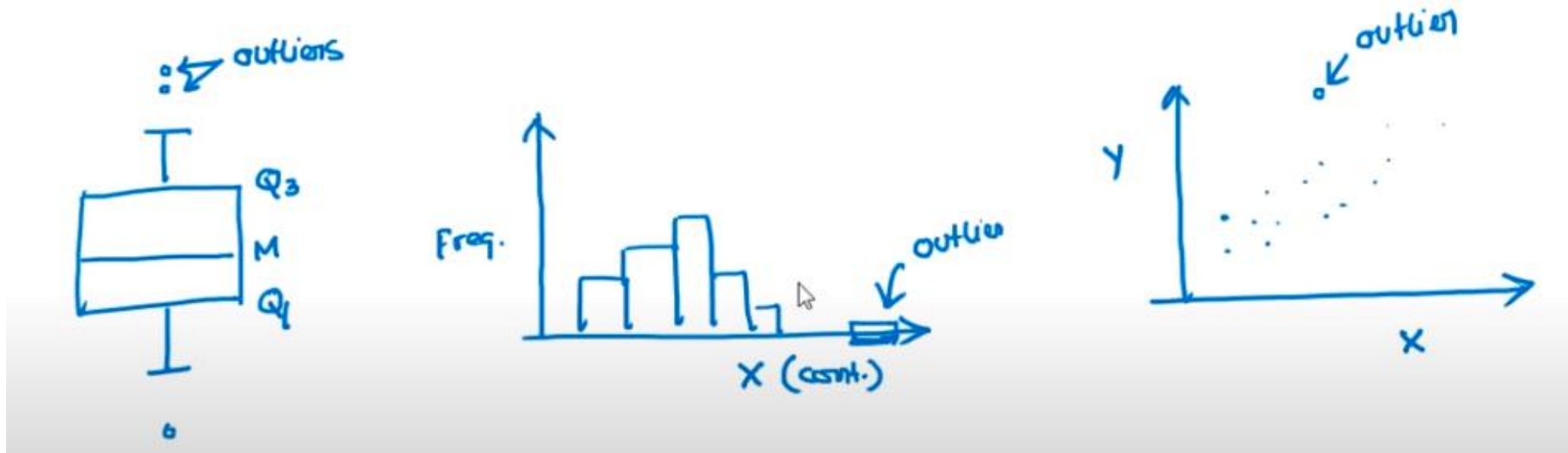
Математические методы

- Позволяют находить выбросы в больших количествах в многомерных данных
- Требуют точной калибровки

Очистка от выбросов

Визуализации

- Не требует калибровки
- Нужно построить хорошую визуализацию
- Не работает, когда выбросов или признаков много



Кросс валидация

Проблема. Жаль выделять данные на валидационную выборку

Алгоритм кросс-валидации

- Разбиваем выборку на n частей
- Для i от 0 до $n-1$
 - Обучаем модель на всех частях кроме i -ой
 - Считаем качество предсказаний на i -ой части
- Усредняем все полученные оценки качества



Кросс валидация

- Получаем более точную оценку качества
- Понимаем, насколько большой разброс у оценки качества
- Перед финальным тестированием можем переобучить модель на всех данных

Итоги занятия

- Можем использовать пайплайн машинного обучения для решения задач
- Научились делать базовую обработку входных данных
- Обсудили кросс валидацию для оценки качества моделей.



IT-ОБРАЗОВАНИЕ
В ПЕТЕРБУРГЕ
И УДАЛЕННО

Спасибо за внимание!