SCHOOL
OF ENERGY
& POWER ENGINEERING

Этапы обработки данных в системах, использующих машинное обучение

Сергей Владимирович Аксёнов,

Доцент отделения информационных технологий ИШИТР,

Томский политехнический университет

Томск-2023

Обработка данных в оффлане режиме

Yandex Search API

Kaggle.com API

GitHub REST API

Подготовка к исследованию

Анализ данных,
признаки,
Схема данных,

Нормализация
Соединение
Агрегация
Обработка естественного языка (NLP)

Создание репозитория

JSON
CSV
Pickle
MySQL
MongoDB
Cassandra

Конструирование признаков (Feature Engineering)

Сокращение размерности
Отбор признаков
Обработка естественного языка (NLP)
Токенизация
Лемматизация
Стемминг

Построение модели (Data Modelling)

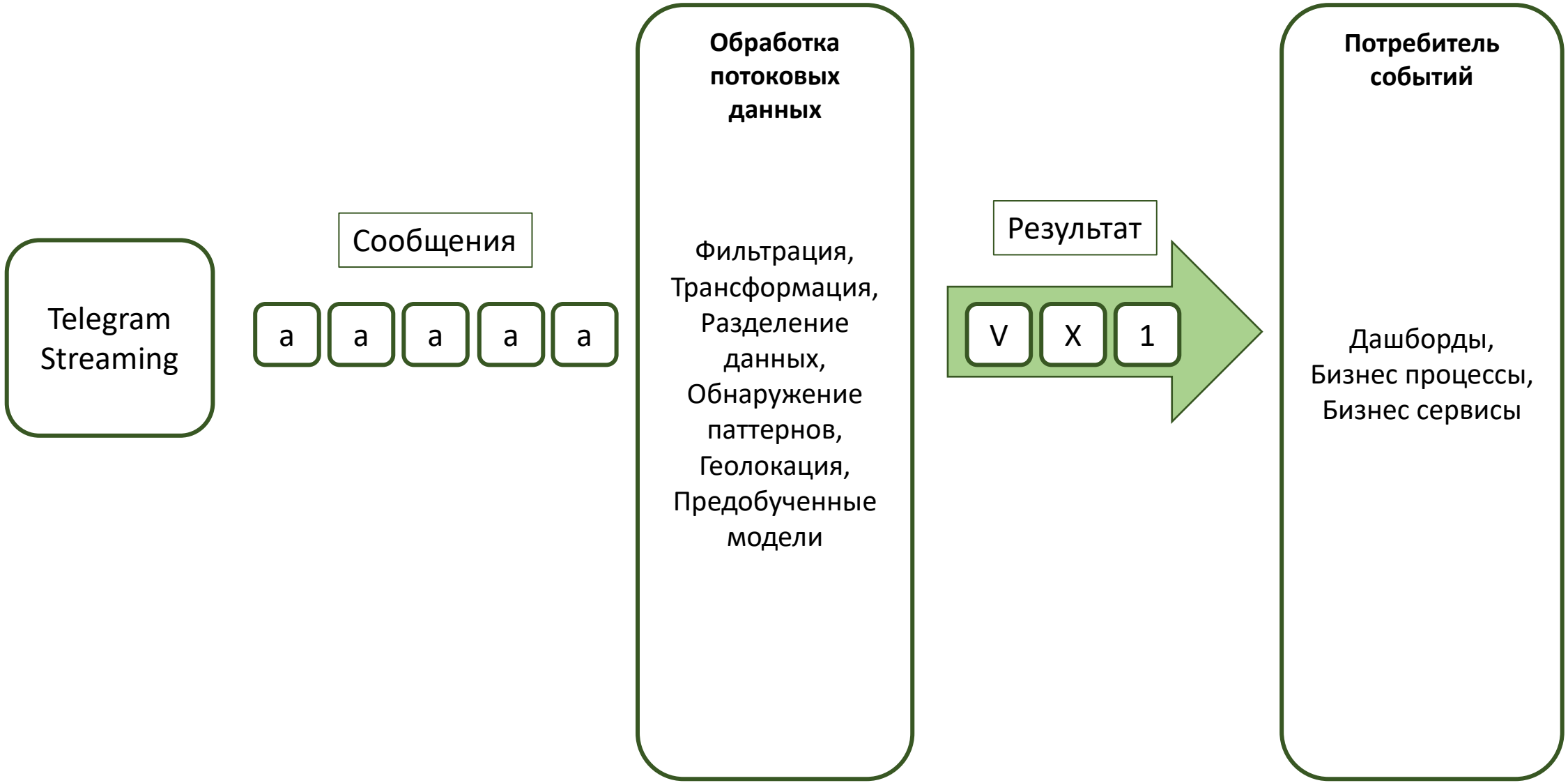
Обучение с учителем:
Классификация
Регрессия,

Обучение без учителя:
Кластеризация,
Поиск аномалий
Фильтрация

Визуализация

Отчеты,
Графики,
Гистограммы
Временные ряды,
Геолокация

Обработка данных в онлайн режиме

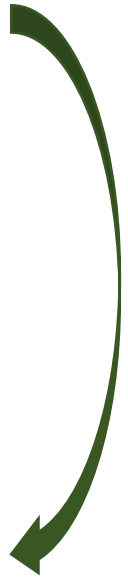


Задачи очистки данных

- ✓ Работа с отсутствующими данными
- ✓ Выявление нарушения взаимосвязи признаков
- ✓ Обработка некорректных данных
- ✓ Устранение дубликатов
- ✓ Преобразования данных
- ✓ Удаление выбросов

Удаление дубликатов

Имя пациента	Дата исследования	С-реактивный белок
Эльвира Сафина	21/08/05	
Андрей Семёнов	21/08/05	-14.3
Юлия Келлер	21/08/05	0.5
Эльвира Сафина	21/08/05	-6.1
Андрей Семенов	21/08/05	20.6
Элвира Сафина	/08/05	8.77



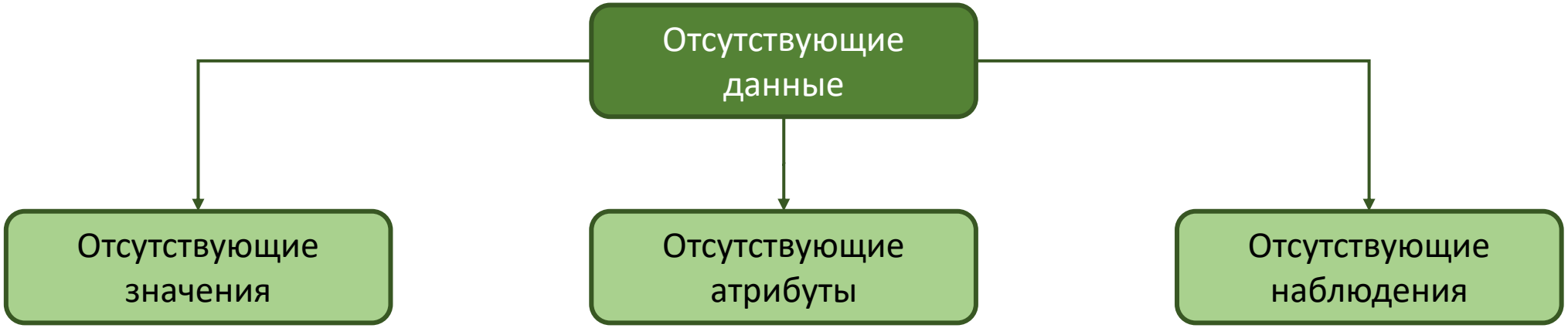
Имя пациента	Дата исследования	С-реактивный белок
Эльвира Сафина	21/08/05	8.77
Андрей Семёнов	21/08/05	20.6
Юлия Келлер	21/08/05	0.5

Баланс классов

В машинном обучении балансировка классов означает изменение числа примеров в наборе, где наблюдается разные доли примеров для каждого класса.
Перед использованием алгоритма машинного обучения важно избежать дисбаланса классов, потому что наша конечная цель – обучить модель машинного обучения, которая хорошо обобщается для всех возможных классов.

Перед использованием алгоритма машинного обучения очень важно посмотреть на распределение классов, чтобы исправить проблему балансировки классов.

Отсутствующие значения

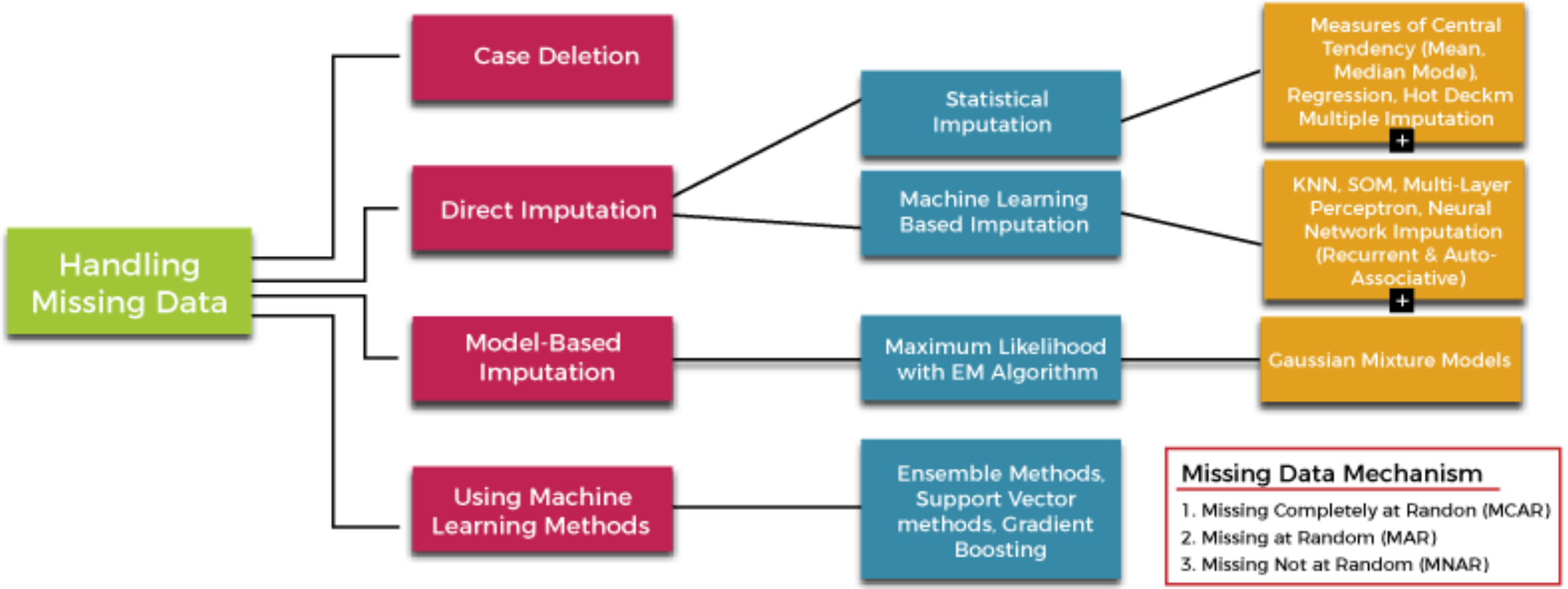


Значения данных, которые должны присутствовать, но не записываются

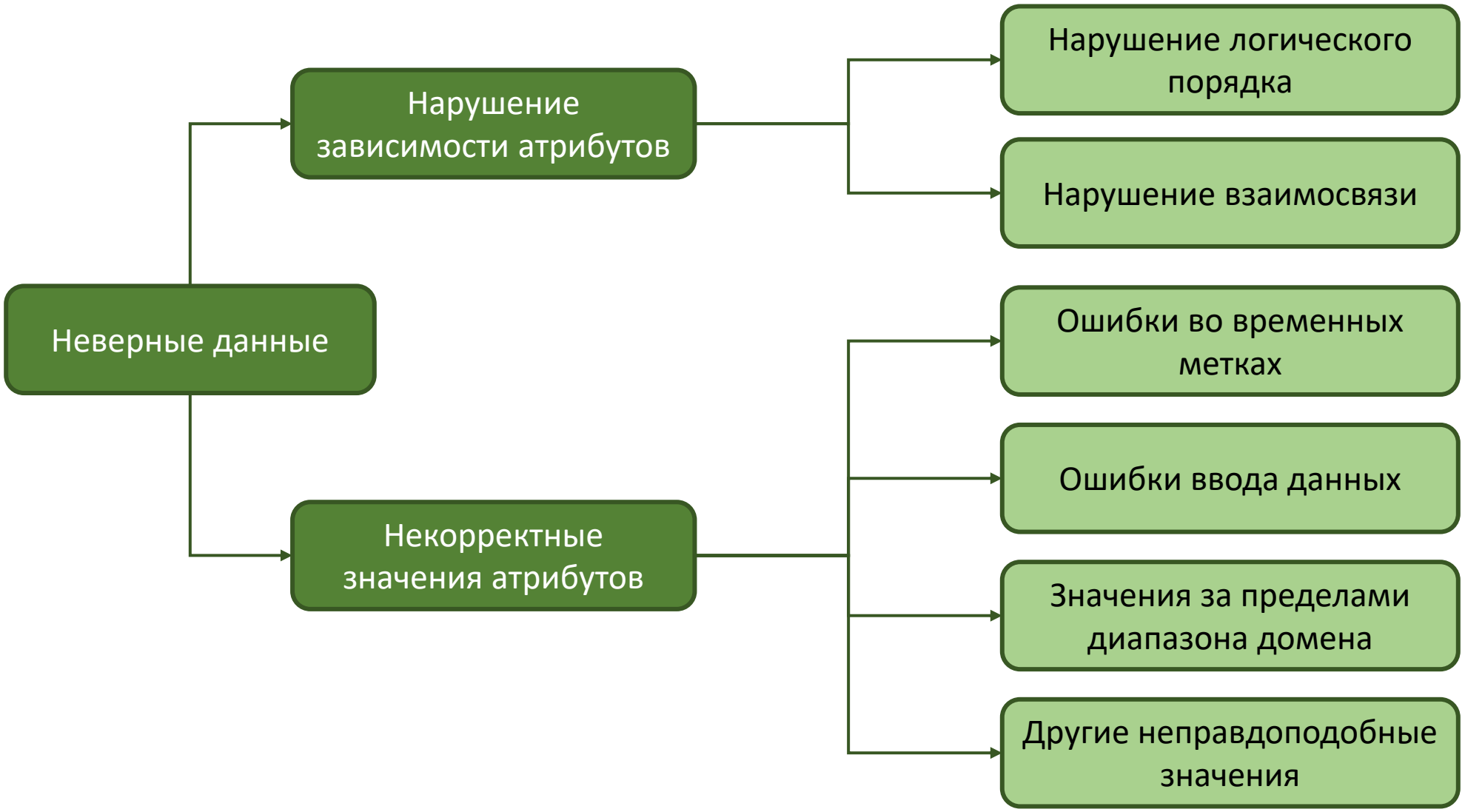
Атрибуты, которые необходимы для анализа, но отсутствуют в журнале событий

Наблюдения, которые были обработаны в действительности, не отображаются в журнале событий

Методы обработки пропущенных значений



Неверные данные



Нарушение логического порядка

Запись из истории болезни:

Заболела остро 17.09.2016г, утром, потрясающий озноб, повышение температуры до 39,5С, чувство разбитости, головная боль, тошнота, рвота, выраженная слабость. К утру 16.09.2016г появилась эритема в области правой стопы, обширная гематома на своде стопы.

Нарушение взаимосвязи

	Имя пловца	Время, сек. Вольный стиль, 50 м.	Скорость движения, м/с
✓	**Рауза Садыкова**	23	2.17
✗	**Мария Григорьева**	24	2.8
✓	**Исмаил Ахметов**	22	2.27

Ошибки во временных метках

Проведения исследования: **15.01.22 10:14**

Поступление образца в лабораторию: **15.01.22 14:47**

Исследование образца: **15.01.22 17:25**

Получения результата исследования: **16.01.22 9:30**

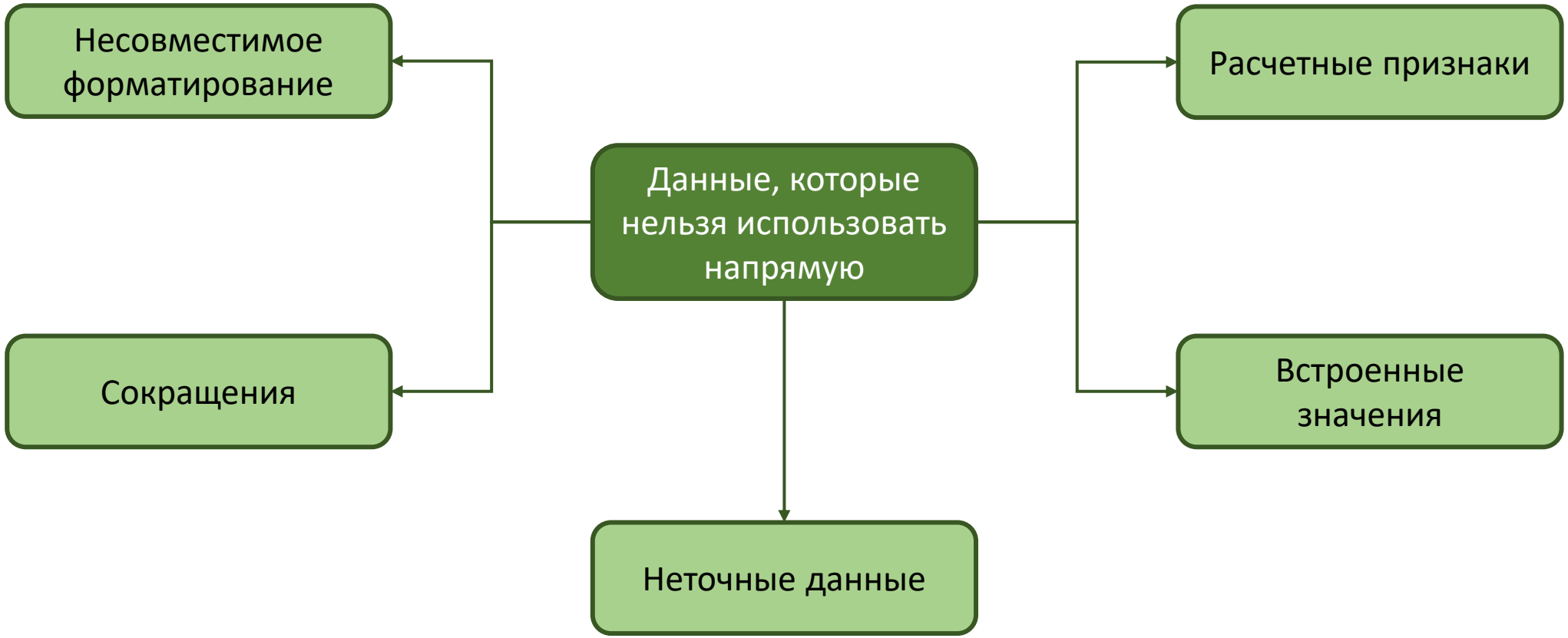
Ошибки набора данных

	Бренд	Фактура	Размер	Тип рукава
✗	**Маргарита**	вльвет	35	Без руковов
✗	**Только ты**	гладкий	399	Длинные
✓	**Savanna**	кружевной	42	Короткие
✗	**BEAUTY**	жаккард	38	Длиннные
✓	**Вирджиния**	атлас	44	Короткие

Значения за пределами диапазона домена

	Имя пациента	Холестерин	Общий белок
✗	**Роксана Ковалевская**	3.93	**-1**
✓	**Марат Денисов**	5.72	80.3
✗	**Ольга Вальтер**	6.24	**38.5**
✓	**Алия Сарымова**	3.36	86.9
✗	**Серж Оганесян**	**-0.11**	77.4

Данные, которые нельзя использовать напрямую



Несовместимое форматирование

Записи из осмотров разных пациентов:

Жалобы: На повышение температуры до 39.2С, распирающую боль в правом плече и предплечье, чувство жара, тошнота.

Жалобы: На высокую температуру до сорока градусов, отёк и гиперемия с гематомой в области свода стопы.

Жалобы: Слабость, повышение температуры до 38,9, жар, озноб, боль, гиперемия, отёк правой голени.

Сокращения

Анамнез жизни: В детстве перенесла корь, ветряную оспу, ОРВИ, грипп, 1 раз пневмония, туберкулёз отрицает. С 2007г АГII, риск 2, ГЛЖ, ИБС, Стенокардия напряжения ФКII. Хр. холецистит, на описторхоз не обследована, хр. панкреатит.

План обследовния: ОАК, ОАМ, БАК, коагулограмма, РМП на сифилис, ЭКГ

Расчетные значения

Имя клиента	Рост	Вес	Индекс массы тела
София	1.75	65	?
Дарина	1.68	73	?
Полина	1.72	69	23.3
Анна	1.81	72	24.3
Камилла	1.62	58	?

Неточные данные

Отзыв покупателя:

Сумка приехала ко мне в розовой версии, и хотя розоватый оттенок здесь действительно есть, что отмечали все мои подруги, кому я её показывала, по сути это, скорее, бежево-кремовый цвет. Впечатление от её специфическое: вроде и не белая, а какая — непонятно.

Встроенные значения

Адрес дома	Общее потребление электроэнергии	Потребление электроэнергии на освещение
Ул. Пушкина, д.7	540	?
Ул. Пушкина, д.56	730	80
Ул. Светлая, д.102	849	?
Ул. Уральская, д.6	624	145
Ул. Ахматовой, д.23	428	?

Преобразование категориальных признаков

- Порядковые (можно сравнить их и упорядочить)
- Номинальные (несравнимые)

Датасет: Футболки

	Цвет	Размер	Цена	Метка
0	зеленый	M	10.1	класс1
1	красный	L	13.5	класс2
2	синий	XL	15.3	класс1

Номинальный признак

Порядковый признак

$$XL > L > M$$

Пример из: Себастьян Рашка Python и машинное обучение, ДМК, Москва-2017, ISBN 978-5-97060-409-0 (категорически рекомендую для студентов-Beginners)

Порядковые признаки

- Для порядковых – задание соответствия вручную исходя из понимания признака

$$XL = L + 1 = M + 2$$

```
size_mapping = {   # словарь соответствий
                 'XL': 3,
                 'L' : 2,
                 'M' : 1}
df['размер'] = df['размер'].map(size_mapping)
```

	Цвет	Размер	Цена	Метка
0	зеленый	1	10.1	класс1
1	красный	2	13.5	класс2
2	синий	3	15.3	класс1

Результат

Номинальные признаки

- Для номинальных – прямое кодирование

```
pd.get_dummies(df[['цена', 'цвет', 'размер']])
```

	Цена	Размер	Цвет_зеленый	Цвет_красный	Цвет_синий
0	10.1	1	1	0	0
1	13.5	2	0	1	0
2	15.3	3	0	0	1

Алгоритмы деревья решений и случайный лес не требуют преобразований категориальных признаков

Приведение данных к одному масштабу

Кластеризация и **многие** алгоритмы обучения с учителем крайне нуждаются в приведении признаков к одной шкале, т.к. их вычисления завязаны на учет расстояния между объектами в выборке.

Нормализация:

$$x_{norm}^{(i)} = \frac{x^{(i)} - x_{min}}{x_{max} - x_{min}}$$

x_{min} - наименьшее значение

x_{max} - наибольшее значение

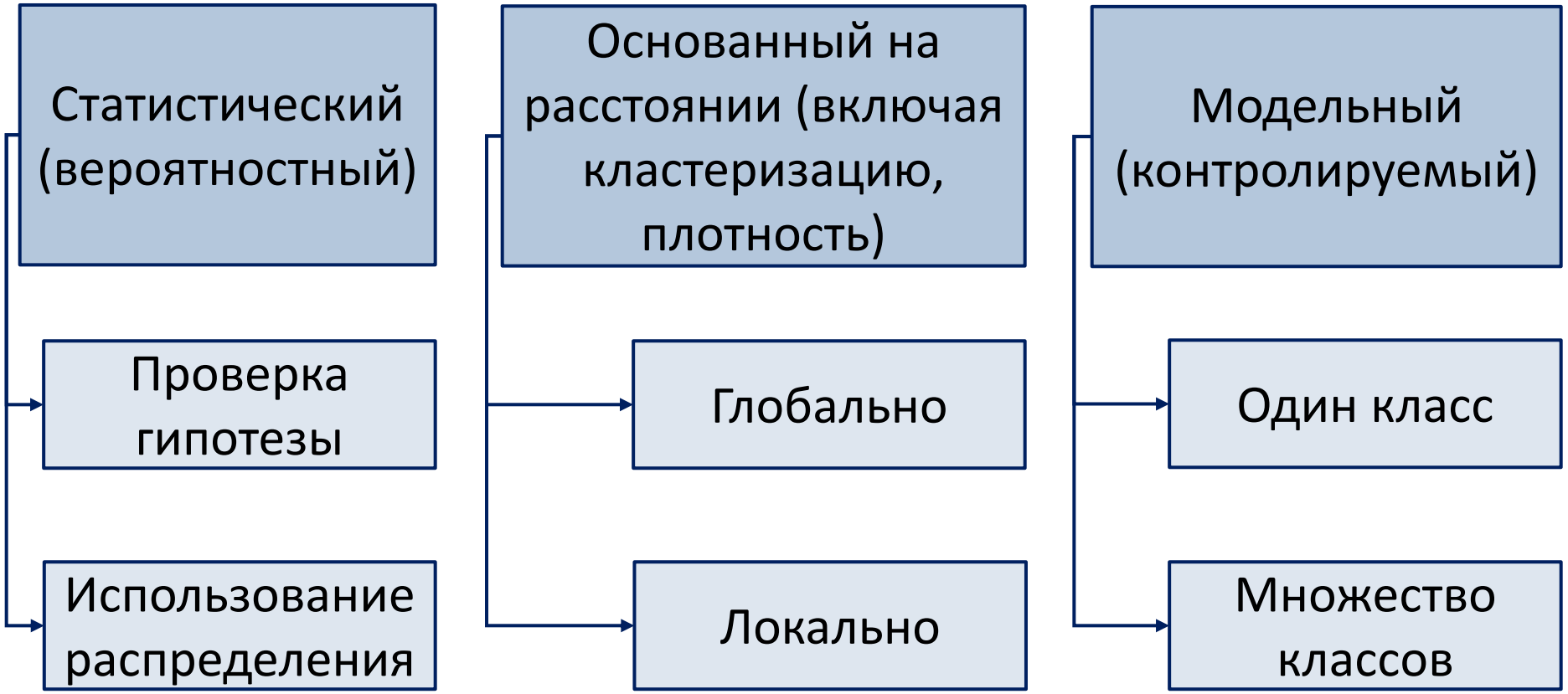
Стандартизация:

$$x_{std}^{(i)} = \frac{x^{(i)} - \mu_x}{\sigma_x}$$

μ_x - эмпирическое среднее

σ_x - стандартное отклонение

Методы обнаружения выбросов



Статистический подход

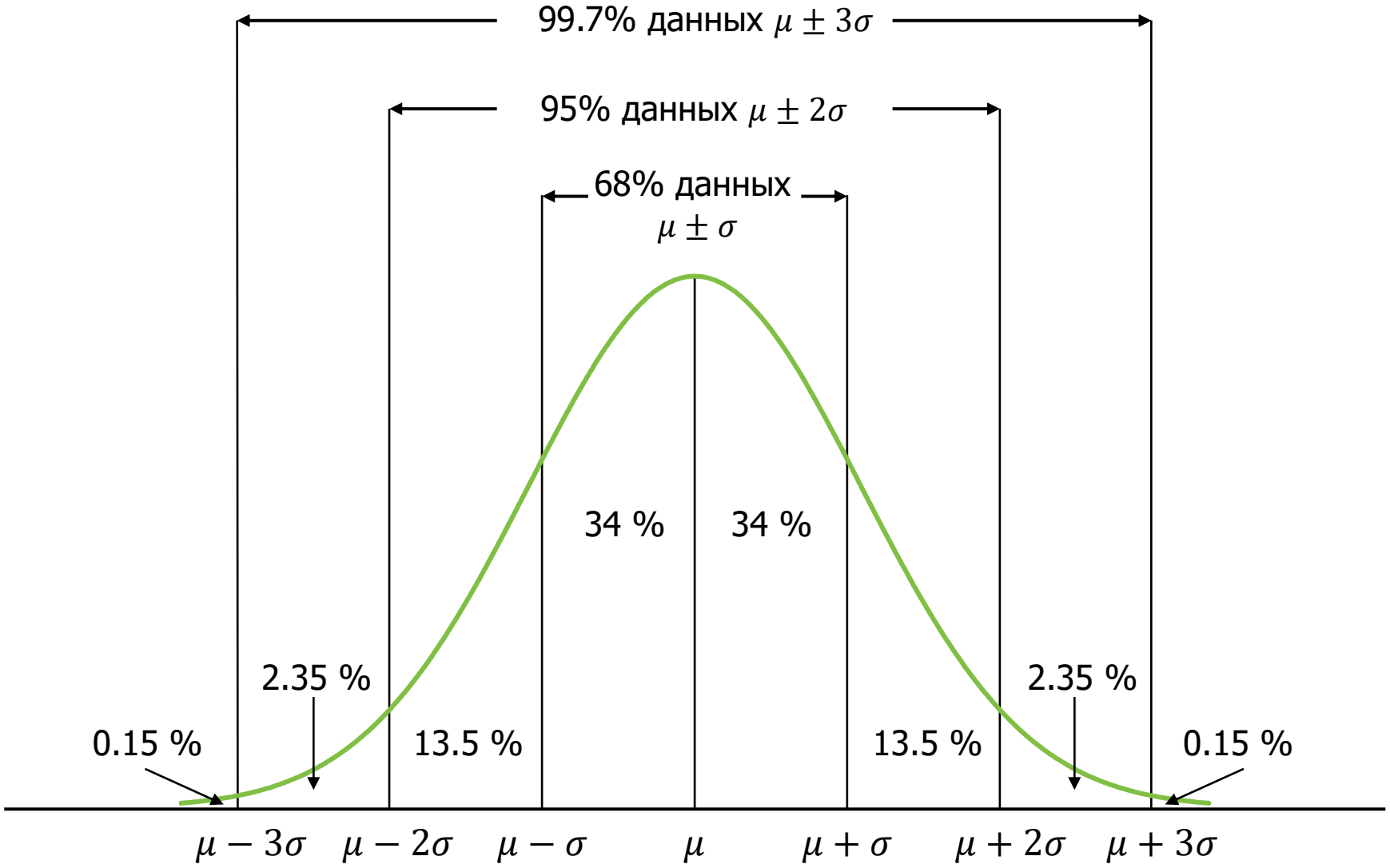
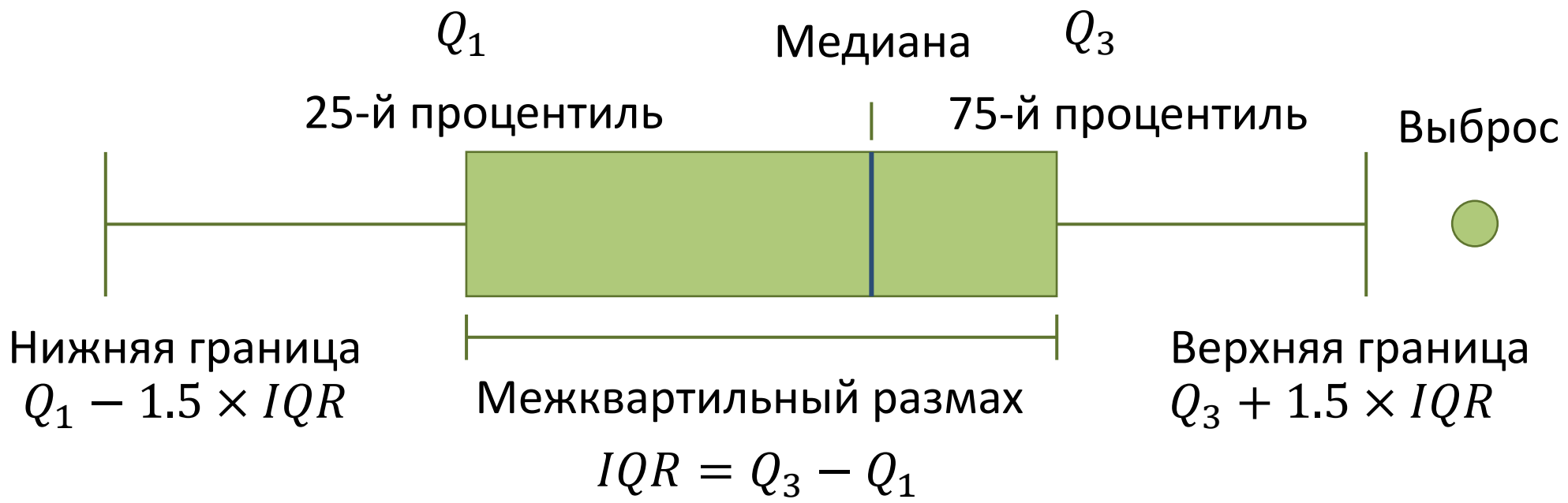
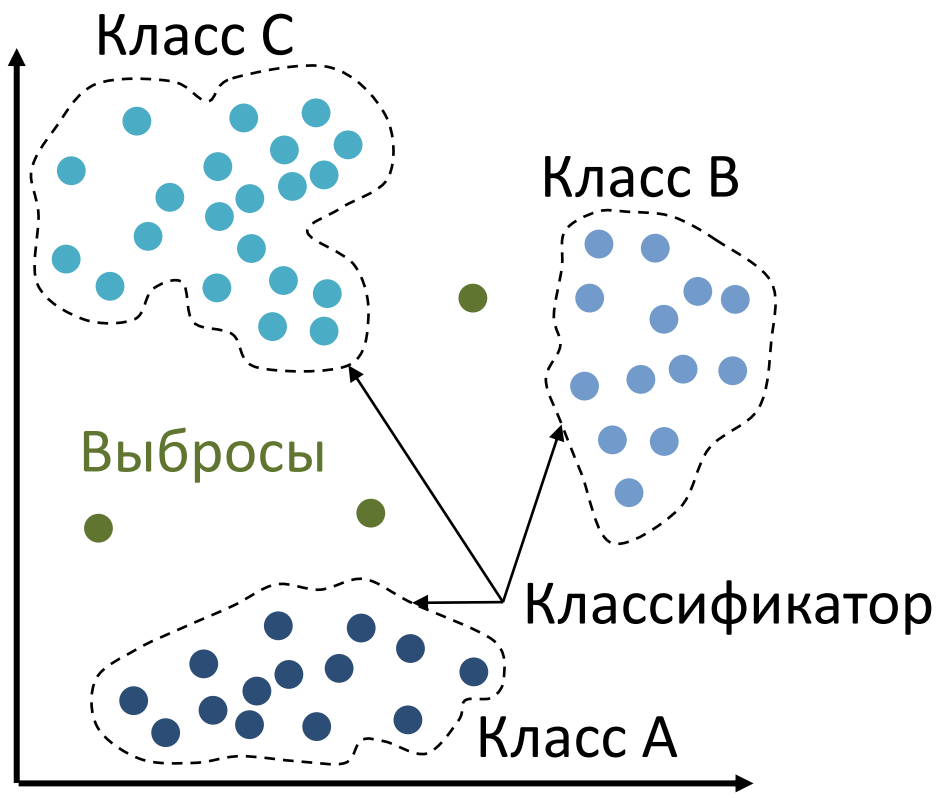


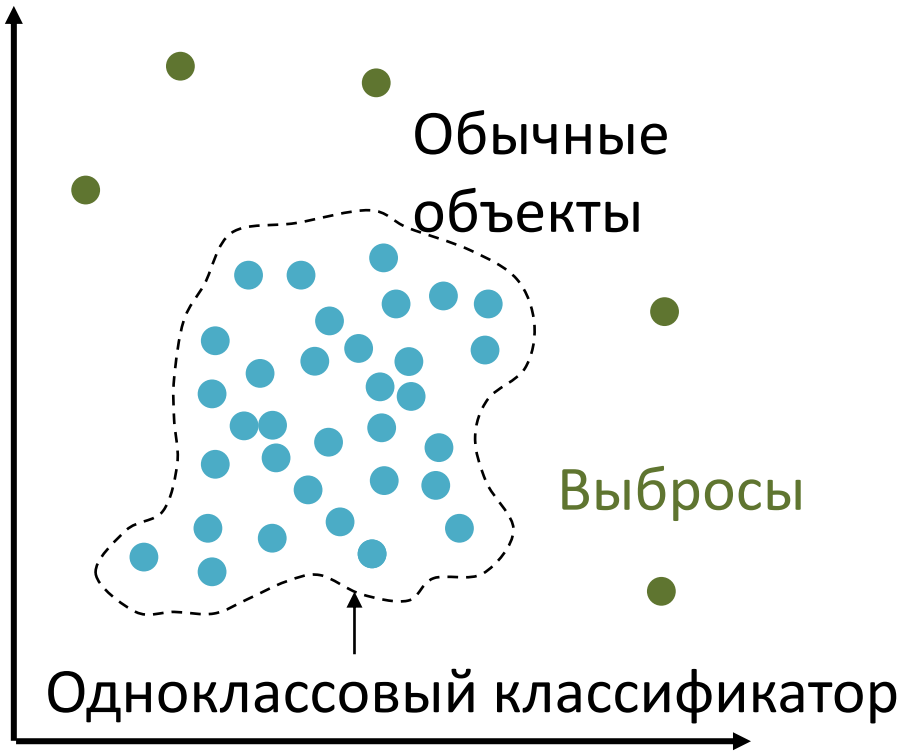
Диаграмма размаха



Методы обнаружения выбросов на основе моделей



Выделение выбросов для многих классов

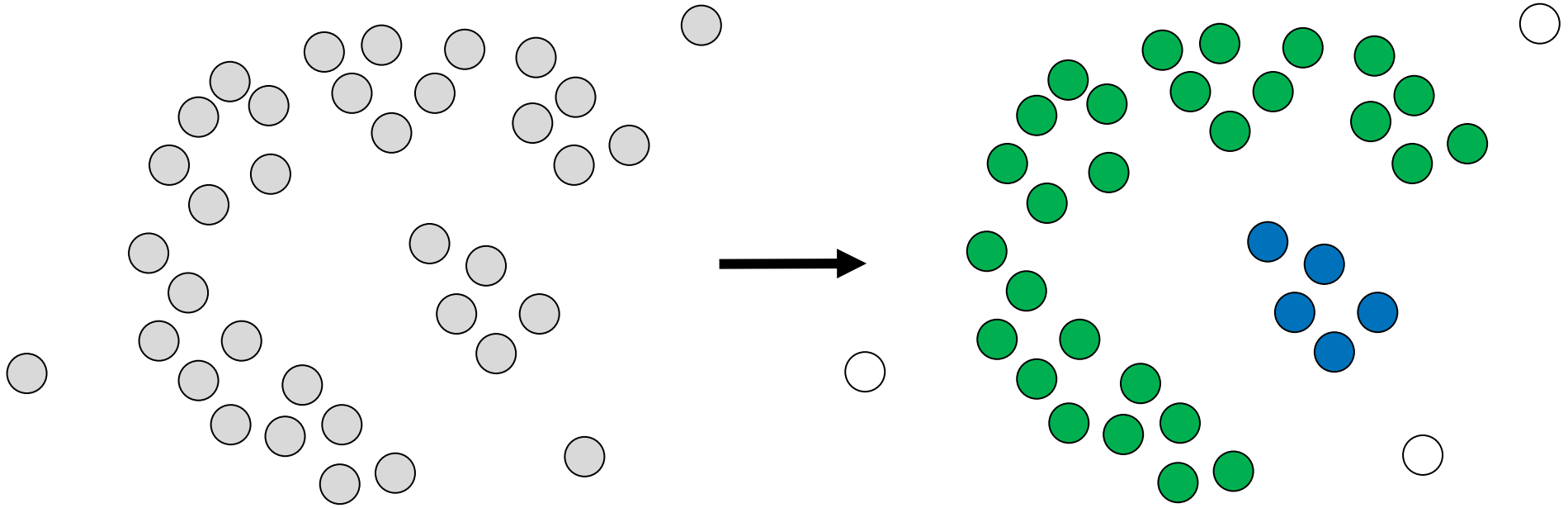


Выделение выбросов для одного класса

Использование кластеризации

Исходные данные

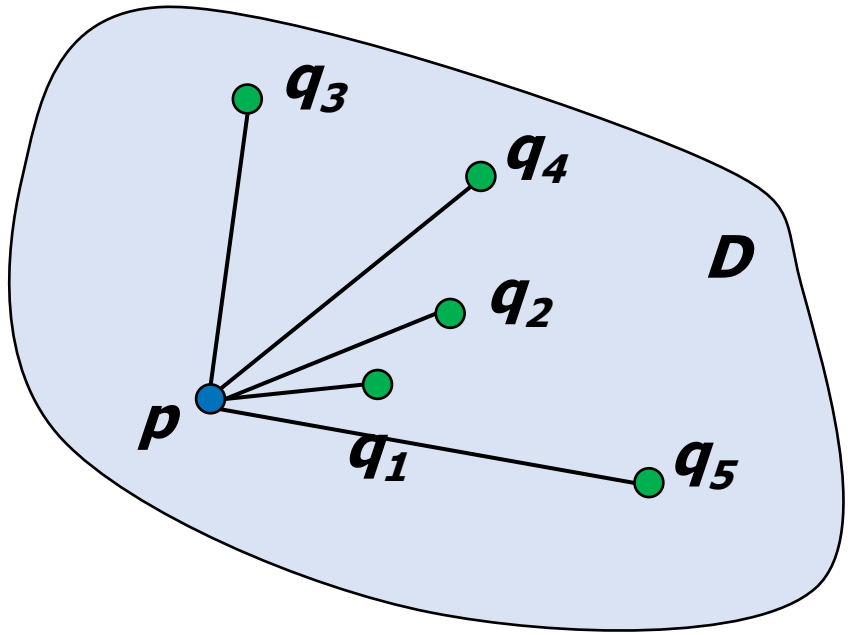
Кластеризация DBSCAN



- Кластер A
- Кластер B
- Выброс

К-расстояние от объекта p

$distance(p, q)$ или $d(p, q)$:



Расстояние между объектом p и его k-ближайшими соседями.

$d(p, q_1) \leq d(p, q_2) \leq d(p, q_3) \leq d(p, q_4) \leq d(p, q_5)$

D – множество объектов

Фактор локального выброса

Расстояние достижимости:

$$reach - dist_k(p,q) = \max(k - dist(p), dist(p,q))$$

Плотность локальной достижимости:

$$Lrd_{MinPts}(p) = 1 \Big/ \left[\frac{\sum_{q \in N_{MinPts}(p)} reach - dist_k(p,q)}{|N_{MinPts}(p)|} \right]$$

Фактор локального выброс:

$$LOF_{MinPts}(p) = \frac{\sum_{q \in N_{MinPts}(p)} \frac{Lrd_{MinPts}(q)}{Lrd_{MinPts}(p)}}{|N_{MinPts}(q)|}$$

$LOF_{MinPts}(p) \sim 1$ такая же плотность, как у соседей

$LOF_{MinPts}(p) < 1$ более высокая плотность, чем у соседей (невыброс)

$LOF_{MinPts}(p) > 1$ более низкая плотность, чем у соседей (выброс)

Эллиптическая оболочка