

## Введение в анализ данных

# Содержание

<b>1</b>	<b>Основные понятия анализа данных</b>	<b>2</b>
1.1	Что такое анализ данных? . . . . .	2
1.2	Основные понятия . . . . .	2
1.3	Этапы анализа данных . . . . .	6
<b>2</b>	<b>Измерения и шкалы</b>	<b>8</b>
2.1	Номинальная шкала (nominal scale) . . . . .	9
2.2	Порядковая шкала (ordinal scale) . . . . .	9
2.3	Интервальная шкала (interval scale) . . . . .	11
2.4	Относительная шкала (ratio scale) . . . . .	12
<b>3</b>	<b>Виды данных</b>	<b>12</b>
3.1	Фактоиды (factoid) . . . . .	13
3.2	Ряды (series) . . . . .	13
3.3	Таблицы . . . . .	14
3.4	Транзакции . . . . .	16
<b>4</b>	<b>Источники данных</b>	<b>16</b>
4.1	Социологические опросы . . . . .	17
4.2	Наблюдения . . . . .	18
4.3	Документы . . . . .	19
4.4	Результаты прямых измерений . . . . .	19
4.5	Социальные сети . . . . .	20
4.6	Внешние источники данных . . . . .	20
<b>5</b>	<b>Подготовка данных</b>	<b>22</b>
5.1	Загрузка данных в хранилища . . . . .	22
5.2	Разделение данных . . . . .	23
5.3	Преобразование данных к одинаковым единицам измерения . . . . .	23
5.4	Преобразование к унифицированной лексике . . . . .	24
5.5	Объединение данных из разных источников . . . . .	25
5.6	Соединение данных из разных источников . . . . .	26
5.7	Заполнение отсутствующих численных значений . . . . .	27
5.8	Очистка данных . . . . .	29

# 1 Основные понятия анализа данных

## 1.1 Что такое анализ данных?

Мы живём в эпоху информационного общества, когда объёмы окружающих нас данных неумолимо растут. Авторы некоторых исследований утверждают, что к 2025 году объёмы данных достигнут 400 зеттабайт. Но главное – все эти накопленные данные содержат в себе важную информацию. Умение выявлять и анализировать требуемую информацию полезно, как на работе, так и в обычной жизни.

Анализ данных используется во многих отраслях науки, промышленности, сфере услуг. Это позволяет компаниям и организациям принимать лучшие решения в бизнесе и способствовать его развитию. Сегодня знаменитую фразу Натана Ротшильда *«кто владеет информацией, тот правит миром»*, можно перефразировать так *«кто понимает, как структурируется информация и владеет инструментами её обработки, тот правит миром»*. Как давно это началось?

Люди всегда старались описать мир, в котором они живут. Кто-то использовал для этого стихи и прозу или живопись, но, кроме этого, находились и те, кто описывал его с помощью чисел.

Чтобы практически применять числовые описания реального мира необходимо понимать правила и логику, которая использовалась при их создании. Числовые описания могут быть выполнены хорошо или безграмотно (например, хорошо известно ироничное описание – «средняя температура по больнице»). В менее очевидных случаях достаточно трудно отличить хорошее описание от плохого. Надеюсь, что наш курс научит вас воспринимать и критически оценивать числовые описания реального мира, а также создавать их самостоятельно и корректно использовать. Некоторые описания окажутся достаточно очевидными и простыми для понимания, другие потребуют привлечения разнообразных математических понятий статистики, вероятности и т.п. Но если вы ознакомитесь с этими возможностями для описания и анализа, вы будете вознаграждены возможностью получать нетривиальные знания об окружающем нас мире и закономерностях, которые в нем присутствуют.

Как наука, анализ данных представляет собой совокупность методов сбора данных, их организации, представления, обобщения, выявления закономерностей и интерпретации (т.е. получения выводов на основе изучаемого природного явления, связанного с этими данными).

## 1.2 Основные понятия

В рамках текущей лекции мы изучим, что именно может являться объектом исследования в рамках анализа данных, какие объекты описываются

с помощью разнообразных признаков (переменных), что принято называть генеральной совокупностью и выборкой, а также какие этапы обычно соответствуют задаче анализа данных.

### 1.2.1 Переменные (Variables)

Переменная или признак – это некоторая общая для всех изучаемых объектов характеристика или свойство, конкретные проявления которого могут меняться от объекта к объекту. Различные проявления признака называют значениями, альтернативами или градациями.

Умение «мыслить признаками», правильно определять переменные для достижения исследовательских целей является одним из важнейших качеств аналитика. На рисунке вы можете видеть некоторые переменные и их возможные значения. Так, например, переменной «профессия» соответствуют значения: «аналитик», «программист», «повар», «менеджер», «преподаватель», «врач», а переменной «размер» соответствуют значения: «5 метров», «7 метров», «100 километров»

## Переменные

переменные:	профессия	цвет	размер
значения переменных:	<ul style="list-style-type: none"><li>• аналитик</li><li>• программист</li><li>• повар</li><li>• менеджер</li><li>• преподаватель</li><li>• врач</li></ul>	<ul style="list-style-type: none"><li>• красный</li><li>• желтый</li><li>• синий</li><li>• зеленый</li><li>• кофе с молоком</li></ul>	<ul style="list-style-type: none"><li>• 5 метров</li><li>• 7 метров</li><li>• 100 километров</li></ul>



Рис. 1: Примеры переменных

### 1.2.2 Распределение значений переменных (distribution)

Предметом исследования анализа данных являются значения, которые переменные принимают на фоне описываемых объектов и явлений. Именно поэтому для каждой отдельно взятой переменной представляет интерес частота появления различных значений, которые может принимать переменная. Эту частоту принято называть распределением переменной. Она может отражаться как в виде абсолютных значений, так и в процентных соотношениях.

Распределение зависит от набора данных, на котором оно рассчитывается. Например, на рисунке вы видите распределение переменной «животные» в двух различных зоопарках. Круговая диаграмма используется для отображения, как абсолютных значений распределения переменной, так и для процентных соотношений. Хорошо видно, что распределение зверей в зоопарках сильно отличается.

## Переменные

**Распределение значений переменных** – это частота появления разных значений, которые может принимать переменная.

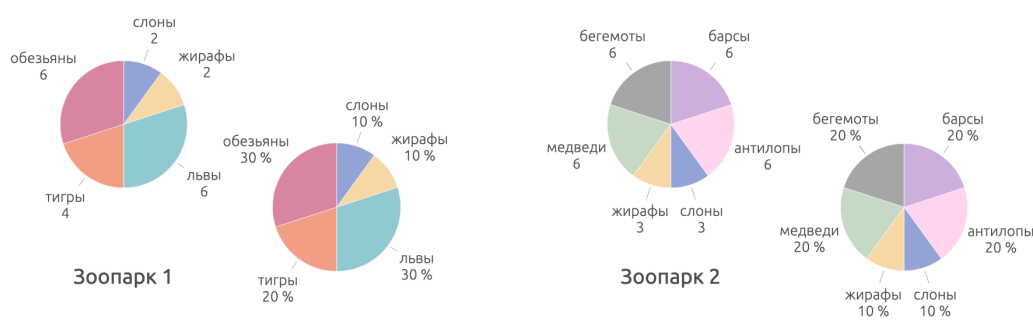


Рис. 2: Пример распределения значений переменных

### 1.2.3 Генеральная совокупность и выборка

Далеко не всегда аналитику предоставляется возможность проанализировать свойства всей совокупности изучаемых объектов. Например, требуется провести анализ общественного мнения многомиллионного города. Выяснить мнение каждого жителя – практически невозможно. Однако реально провести социологический опрос 5 тысяч жителей и по его результатам попытаться сделать выводы об общественном мнении города. Отсюда естественным образом вытекают понятия генеральной совокупности и выборки:

- **Генеральная совокупность** (population) – это вся интересующая исследователя совокупность изучаемых объектов.
- **Выборка** (sample) – это некоторая, обычно небольшая, часть генеральной совокупности, отбираемая специальным образом и исследуемая с целью получения выводов о свойствах генеральной совокупности.

Когда исследователи используют слово «выборка» они, как правило, подразумевают, так называемую, репрезентативную выборку, т.е. выборку в которой значения переменных распределены в процентном соотношении приблизительно так же, как и в генеральной совокупности. Именно поэтому при

## Генеральная совокупность и выборка

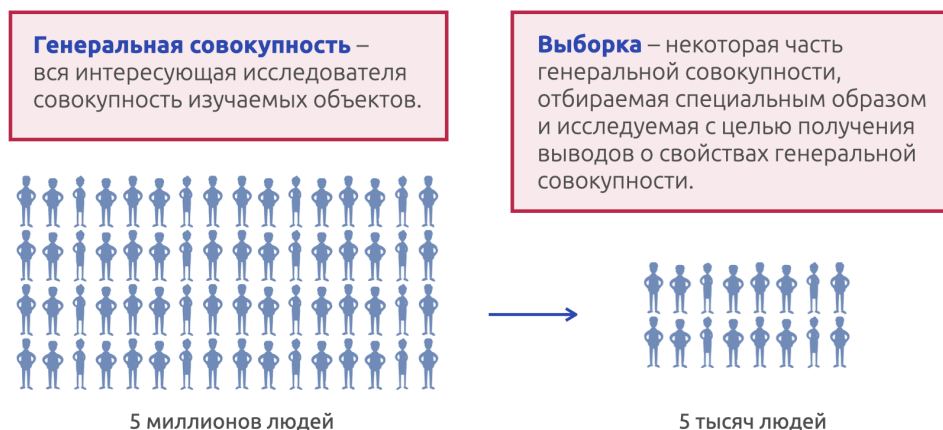


Рис. 3: Генеральная совокупность и выборка

проведении социологических опросов исследователи используют специальные технологии для формирования репрезентативной выборки.

### 1.2.4 Гипотеза (hypothesis)

Еще одно важное понятие для анализа данных – гипотеза. **Гипотеза** (hypothesis) – это предположение относительно значений переменных генеральной совокупности (которое, возможно, определяется на основе анализа выборки).

На рисунке вы можете видеть примеры гипотез. Так, например, гипотезой является предположение, что потребление электроэнергии зависит от времени года и дня недели. Еще одна гипотеза – утверждение о том, что в рабочие дни вечерний «час пик» соответствует периоду с 18:00 до 19:00.

## Гипотеза

**Гипотеза** – это предположение относительно значений переменных генеральной совокупности (которое, возможно, определяется на основе анализа выборки).

- Потребление электроэнергии зависит от времени года и от дня недели.
- В рабочие дни вечерний «час пик» соответствует периоду с 18:00 до 19:00.
- Завтра будет солнечная погода.




Рис. 4: Примеры гипотез

Предсказание ожидаемой погоды также является гипотезой, которая требует анализа существующих данных о текущей погоде и прочих метеоусловиях.

### **1.3 Этапы анализа данных**

Этапы анализа данных и порядок их выполнения могут различаться в зависимости от особенностей изучаемых объектов и явлений, но, тем не менее, есть достаточно стандартные этапы, характерные для анализа любого набора данных, например, описание изучаемых объектов, выявление переменных, сбор, подготовка данных и т.п. Рассмотрим особенности каждого из этих этапов.

#### **1.3.1 Описание изучаемых объектов**

Описание изучаемых объектов. На этом этапе полагается в письменном виде изложить теорию вопроса, то есть суть изучаемых объектов или явлений. Обычно к этому этапу привлекают экспертов предметной области. Хорошее описание будет способствовать следующим этапам анализа.

#### **1.3.2 Выявление переменных и формулировка гипотез**

Следующий этап – выявление переменных и формулировка гипотез. На этом этапе происходит операционализация понятий. То есть переход от абстрактных понятий предметной области к переменным, которые могут быть измерены количественно и качественно. На этом же этапе происходит формулировка гипотез в терминах переменных.

#### **1.3.3 Сбор и подготовка данных**

Следующий этап – сбор и подготовка данных для проверки гипотезы. Данные могут быть собраны различными путями, например, в результате социологических опросов, получены как результаты измерений, из внешних источников и т.п. На этом этапе требуется собрать всю исследуемую генеральную совокупность (что, иногда, оказывается технически возможным) или сформировать репрезентативную выборку, полученную в результате выборочного обследования.

Кроме того, этот этап подразумевает подготовку данных для загрузки в хранилище. Однако прежде, чем загрузить данные в хранилище, необходимо произвести тщательную предобработку данных, сформировать структуры для хранения, заполнить пропущенные значения, проверить правдоподобность представленных данных и т.п.

Разумеется, начиная с этого этапа в распоряжении аналитика должен быть инструмент, позволяющий загрузить данные в хранилище и выполнить все необходимые манипуляции по подготовке данных для последующего анализа.

#### 1.3.4 Разведочный анализ данных

Следующий этап – разведочный анализ данных. Разведочный анализ данных (Exploratory data analysis) – это анализ основных свойств данных, нахождение в них общих закономерностей, распределений и аномалий и, возможно, построение начальных моделей. Вообще-то у аналитиков нет твердого мнения о том, что именно следует «разведать» о данных на этом этапе.

Этот термин был введен математиком Джоном Тьюки, который сформулировал *цели разведочного анализа* следующим образом:

- максимальное «проникновение» в данные;
- выявление основных структур;
- выбор наиболее важных переменных;
- обнаружение отклонений и аномалий;
- проверка основных гипотез.

На практике этот этап, как правило, сводится к визуализации подготовленного набора данных, описанию с помощью разного рода описательных статистик и выявлению различных отклонений в данных. Именно на этом этапе может быть обнаружена необходимость в дальнейшей очистке и преобразовании данных. В некоторых случаях даже эти простейшие способы анализа дают достаточно выразительные результаты, подтверждающие выдвинутые гипотезы.

#### 1.3.5 Очистка данных от шумов и аномалий

Следующий возможный этап – очистка данных от шумов и аномалий. Реальные данные для анализа редко бывают хорошего качества. Разведочный анализ может выявить наличие шумов и разного рода аномалий в данных. Эти явления могут искажать общую картину закономерностей, присутствующих в анализируемых данных. Назначение этого этапа – избавление от выявленных шумов и аномалий.



### 1.3.6 Преобразование данных

Преобразование данных. Разведочный анализ может выявить, что значения некоторых переменных резко отличаются по масштабу. В этом случае их не только трудно анализировать, но и невозможно качественно визуализировать. Между тем, простейшие математические преобразования могут с легкостью решить эту проблему. Более того, некоторые алгоритмы анализа требуют не только соизмеримости переменных по масштабу, но и, так называемой, нормировки переменных, то есть точного попадания значений переменных в заданный диапазон возможных значений (например, от 0 до 1). Для такого рода преобразований разработаны специальные методы, с которыми мы познакомим вас в последующих лекциях.

### 1.3.7 Построение моделей

Построение моделей. Эта фаза подразумевает проверку гипотез и построение математических моделей, описывающих поведение переменных и связей между ними.

### 1.3.8 Интерпретация

И, наконец, последний этап – интерпретация. Интерпретация – процесс превращения данных в информацию, процесс придания им смысла. Этот процесс зависит от многих факторов: кто интерпретирует данные, какой информацией уже располагает интерпретатор, с каких позиций он рассматривает полученные данные и т.д.

Процесс интерпретации построенных моделей может осуществляться человеком или группой лиц, при этом он может быть неформальным, творческим или формальным (на основе метрик). Полученные модели являются, по сути, формализованными знаниями эксперта, которые можно и нужно тиражировать.

## 2 Измерения и шкалы

Как уже говорилось ранее в Анализе Данных принято представлять объекты реального мира переменными. Переменным соответствуют значения. Для того, чтобы сопоставить значение той или иной переменной, нужно провести измерение (measurement) согласно некоторому правилу.

В практической деятельности необходимо проводить измерения или классификацию различных величин, характеризующих свойства объектов, явлений и процессов. Некоторые свойства проявляются только качественно,

другие – количественно. В связи с понятием измерения переменной появляется понятие шкалы измерений.

Шкала измерений – это числовой или символьный ряд значений, отражающий допустимые вариации значений измеряемой величины.

В соответствии с логической структурой проявления свойств изучаемого объекта различают четыре основных вида шкал измерений:

- номинальная;
- порядковая;
- интервальная;
- относительная.

## 2.1 Номинальная шкала (nominal scale)

Номинальная шкала состоит из названий, имен или категорий и предназначена для классификации объектов или явлений по некоторому признаку. Примерами номинальных шкал могут служить переменные: вид животного, семейное положение, профессия, ученая степень.

Если переменная принимает только два возможных значения, например, да/нет, 0/1, true/false, знаю/не знаю, то говорят, что такие переменные являются дихотомическими (то есть это частный случай номинальной шкалы). Хорошо известный пример номинальных данных – названия отделов в супермаркете и несколько примеров дихотомических данных. Переменная наличие товара, которая принимает два возможных значения имеется или отсутствует, переменная высказывание, которое истинно либо ложно и, наконец, переменная женат, которая принимает два возможных значения: да или нет.

**Агрегирование номинальных данных.** Поскольку номинальные данные характеризуются только принадлежностью к тому или иному классу объектов, то в них отсутствует понятия нуля, единицы измерения и возможность сравнивать объекты на предмет упорядоченности. Тем не менее, количество различных номинальных значений можно подсчитать, можно определить процент от целого, однако нельзя вычислить среднее значение. Например, можно говорить о том, сколько специалистов выпустил университет: программистов, математиков, физиков, химиков, геологов, филологов, но нельзя подсчитать кого в среднем выпускает университет.

## 2.2 Порядковая шкала (ordinal scale)


Порядковая шкала (ordinal scale) означает, что числа или упорядоченный набор текстовых характеристик присваивается объектам так, чтобы обозначить относительные качественные позиции объектов. Это, с одной стороны

дает возможность классифицировать объекты, а с другой – сравнивать между собой качественные характеристики объектов.

Примером порядковой шкалы является хорошо известная классификация отелей. Вы наверняка догадываетесь, что гостиница класса «четыре звезды» по качеству лучше, чем гостиница «три звезды», однако во сколько раз она лучше сказать невозможно.

**Агрегирование порядковых данных.** На рисунке представлен еще один очень известный пример – 12-балльная шкала Бофорта для измерения силы морского ветра. Эта шкала представляет 12 категорий силы ветра. Каждой категории соответствуют баллы, характеризующие силу морского ветра. Например, ветер силой в 9 баллов (т.е. шторм) сильнее ветра силой в 0 баллов (который соответствует штилю).

### Пример: Шкала Бофорта




0 баллов штиль	4 балла умеренный ветер	8 баллов очень крепкий ветер
1 балл тихий ветер	5 баллов свежий ветер	9 баллов шторм
2 балла лёгкий ветер	6 баллов сильный ветер	10 баллов сильный шторм
3 балла слабый ветер	7 баллов крепкий ветер	11 баллов жестокий шторм
	12 баллов ураган	

Рис. 5: Пример порядковых данных

Количество различающихся порядковых значений можно подсчитать и определить процент от целого, однако нет единого мнения о том, можно ли для порядковых данных подсчитать среднее значение. С одной стороны, невозможно определить среднее значение для переменной «сила морского ветра», и даже если вы формально определите это числовое значение, оно не будет иметь фактического смысла. С другой стороны, в некоторых исследованиях разница величин между последовательными категориями приблизительно одинакова.

Например, если возможные варианты ответа в социологическом опросе 1, 2, 3, 4 и 5 и соответствуют словесным описаниям: полностью не согласен, не согласен, отношусь нейтрально, согласен и полностью согласен, то можно предположить, что разница между соседними ответами в этой шкале приблизительно одинаковая.

Это обстоятельство дает основание социологам рассчитывать средние

значения таких ответов и находить им разумную интерпретацию. Такое агрегирование сильно зависит от специфики изучаемых объектов.

В некоторых отраслях настоятельно не рекомендуется использовать порядковые данные для проведения подобных расчетов, в то время как в других – это постоянная практика.

На рисунке вы видите попытку рассчитать агрегированные значения для переменных сила ветра и результаты социологического опроса. И в первом, и во втором случае рассчитать количество удастся, а вот среднее возможно определить только для социологического опроса.

### Пример агрегирования порядковых данных

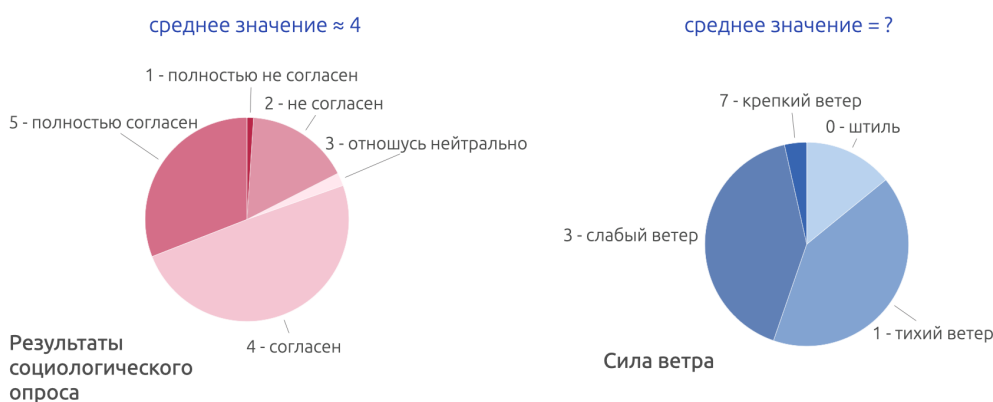


Рис. 6: Пример расчета агрегированных значений

## 2.3 Интервальная шкала (interval scale)

Шкала интервалов состоит из одинаковых интервалов, имеет единицу измерения и произвольно выбранное начало – условную нулевую точку.

Интервальная шкала (interval scale) обладает всеми свойствами номинальной и порядковой шкал, но при этом дает возможность указать количественное значение измеряемого признака.

Это позволяет находить разницу в заданных единицах измерения между двумя значениями. Недостатком служит отсутствие абсолютного нуля в качестве точки отсчета.

Пример шкалы интервалов – летоисчисление по различным календарям, в которых за начало отсчета принято считать сотворение мира (которое тоже трактуется по-разному), рождение Христова, основание Рима и другие варианты.

Еще один пример шкалы интервалов – температурные шкалы Цельсия и Фаренгейта.

Если температура воздуха 30 градусов по Цельсию, то можно говорить, что это на 10 градусов выше, чем при температуре в 20 градусов, но нельзя

сказать, что это ровно в полтора раза теплее по причине отсутствия абсолютного нуля в качестве точки отсчета. Единичные интервалы могут в свою очередь делиться на равные интервалы. Пример: шкала времени, которая делится на годы, месяцы, дни, часы, мин, секунды и т.д. Теоретически интервальная шкала делима бесконечно, что позволяет устанавливать точность в зависимости от целей исследования.

**Агрегирование интервальных данных.** Со значениями интервальной шкалы можно проводить арифметические операции: складывать, вычитать, умножать и т.п. Формально – все эти операции возможны. Однако проводить их нужно только в том случае, если у получаемого результата есть разумная интерпретация. Например, можно подсчитать среднюю температуру в июне в какой-нибудь географической точке, но нелепо считать среднюю температуру по больнице.

## 2.4 Относительная шкала (ratio scale)

Относительная шкала имеет все свойства интервальной шкалы и обладает абсолютным нулем в качестве точки отсчета.

В относительной шкале абсолютный ноль означает отсутствие того, что вы измеряете, например, скорость движения автомобиля, количество яблок в вашем холодильнике. Во всех этих случаях ноль означает, что у вас нет того, что вы измеряете, и это отличается от того условного нуля, который присутствует в интервальных данных.

**Агрегирование относительных шкал.** Для данных этой шкалы осмысленными являются все арифметические операции, включая деление. Например, цена товара измеряется в относительной шкале в рублях или другой валюте. Можно сравнивать цены на товары и говорить, что один из них в три раза дороже другого. Цена товара в 0 рублей означает, что товар бесплатный.

## 3 Виды данных

Когда вы пытаетесь превратить данные в полезную информацию, важно понимать, что именно собой представляют данные, с которыми вы собираетесь работать. Данные могут быть простыми фактоидами (результатами чьего-то предварительного анализа) или «сырыми» транзакциями, которые еще не были подвергнуты никакой обработке.

В зависимости от уровня предварительной агрегации различают следующие виды исходных данных:

- фактоиды;
- ряды;

- таблицы;
- транзакции.

Рассмотрим последовательно каждый из этих видов.

### 3.1 Фактоиды (factoid)

Фактоид – это агрегированная часть общей информации. Фактоид рассчитывается из исходных (сырых) данных и акцентирует внимание на конкретной детали. В самом слове фактоид присутствует некоторое недоверие к данным, которые кто-то и как-то проагрегировал. Тем не менее, значительная часть данных, которые мы получаем для анализа, представлены именно в виде фактоидов.

Пример фактоида – «95% туристов в Санкт-Петербурге посещают Эрмитаж». В этих данных делается акцент именно на посетителях Эрмитажа. Очевидно, что это результат чьего-то предварительного анализа, который опирался на сведения о посетителях различных музеев Санкт-Петербурга. Этот фактоид ничего не говорит о посетителях Русского музея или Петропавловской крепости. Может быть, это заключение было сделано на основе билетов, которые покупались посетителями музеев? Но эти билеты не дают сведений о статусе посетителя: житель Санкт-Петербурга или турист. Может быть – это результаты социологического опроса туристов? Нам остается только надеяться, что эти данные были получены и проагрегированы на основе достаточно репрезентативной выборки. Но некоторое недоверие к точности этих данных все же присутствует.

### 3.2 Ряды (series)

Ряд – это данные, в которых один вид информации (зависимая переменная) сопоставляется другому виду информации (независимой переменной). Информация, соответствующая зависимой переменной, может носить агрегированный характер, то есть являться фактоидом.

Например, в таблице представлен пример как выглядит процентное соотношение количества поездок по используемым видам общественного транспорта: В приведенном примере в качестве независимой переменной высту-

Вид транспорта	Количество поездок
автобус	26%
троллейбус	6%
трамвай	6%
метро	62%

пает вид общественного транспорта, а в качестве зависимой – процентное

соотношение количества поездок. Процентные соотношения в данном примере – фактоиды. Остается надеяться, что они были адекватно рассчитаны на основе сведений о реальной оплате проезда в каждом виде общественного транспорта.

Ряд называется **временным** (time series), если в качестве независимой переменной выступает время. В качестве примера можно рассмотреть статистику Санкт-Петербурга, связанную с посещением театров.

Данные, которые вы видите в таблице, были получены с официального сайта правительства Санкт-Петербурга. В приведенном примере общее количество посещений театров зависит от года. Поэтому год – это независимая

2010	2011	2012	2013	2014	2015	2016
1492600	1702200	1823300	1866300	2117200	2310680	2267129

переменная, а количество посещений театров – зависимая переменная, которая отображает количество посещений, соответствующих выбранному году (например, в 2011 году 1702200 человек посетили театры Санкт-Петербурга). Наверное, эти данные были проагрегированы на основе отчетности театров о проданных билетах и есть шанс, что они достаточно достоверны.

### 3.3 Таблицы

Табличные данные представляют особый интерес, так как сегодня большинство государственных учреждений публикуют в открытом доступе агрегированную статистику о своей деятельности именно в виде таблиц.

В табличных данных есть несколько единиц зависимой информации и одна единица независимой информации. Информация, соответствующая зависимым переменным, также может носить агрегированный характер, то есть представляться в виде фактоидов.

В таблице вы можете видеть расширенный пример с общественным транспортом. Здесь транспорт – независимая переменная, а количество поездок и средняя продолжительность поездки – зависимые. Количество поездок

Вид транспорта	Количество поездок	Средняя поездка
автобус	26%	31 мин
троллейбус	6%	22 мин
трамвай	6%	21 мин
метро	62%	28 мин

и средняя продолжительность поездок на каждом виде транспорта представлены в виде фактоидов. Однако большой вопрос – насколько достоверны эти фактоиды. Если количество поездок можно рассчитать на основе проданных

билетов, то средняя продолжительность – значительно более сложный показатель. Никакие информационные системы на общественном транспорте в настоящее время явным образом не фиксируют вход и выход пассажиров, а следовательно, это – расчетная величина, полученная на основе каких-то вероятностных алгоритмов или наблюдений. Тем не менее, такие данные явно расширяют наши знания об общественном транспорте, однако как связать эту информацию между собой для получения практически значимых результатов пока трудно представить. Агрегирование между этими зависимыми значениями неуместно, так как они даже выражены в различных единицах измерения.

Рассмотрим еще один пример табличных данных, в которых в качестве независимой переменной выступает время. Это статистика Санкт-Петербурга, в которой представлено уже знакомое вам количество посетителей театров и количество студентов в высших учебных заведениях во второй строке. Обе зависимые переменные (посетители театров и студенты высших

Год	2010	2011	2012	2013	2014	2015	2016
Театр	1492600	1702200	1823300	1866300	2117200	2310680	2267129
ВУЗ	807000	749500	707800	654500	593000	555600	553700

учебных заведений) выражены в одинаковых единицах измерения (в людях). Однако агрегирование этих данных, по-прежнему, не имеет смысла. Данные абсолютно не связаны между собой. Агрегирование между зависимыми переменными или какое-либо сравнение в данном случае лишено всякого смысла. Как можно связать между собой количество посещений театров в 2015 году (2310680 человек) и 555600 студентов в высших учебных заведениях? У нас нет никаких оснований связать эти данные между собой, несмотря на то, что они произошли в один и тот же год в одном и том же месте.

Однако картина меняется, если в качестве табличных данных будет рассмотрена статистика, в которой отображаются студенты высших и средних учебных заведений. Интуитивно становится понятно, что агрегировать эти

Год	2010	2011	2012	2013	2014	2015	2016
ВУЗ	807000	749500	707800	654500	593000	555600	553700
СУЗ	109100	109000	111900	104700	105100	101200	101300
$\Sigma$	916100	858500	819700	759200	698100	656800	655000

данные возможно. Однако хотелось бы найти более формальное обоснование. На самом деле студенты высших учебных заведений и студенты средних учебных заведений – это две категории переменной студент и именно поэтому суммировать значения, измеряемые в одинаковых единицах измерения и относящихся к категориям одной переменной вполне допустимо.



### 3.4 Транзакции

Транзакционные («сырые») записи представляют собой данные о конкретных событиях. Внешне они представлены, как правило, в виде рядов или таблиц. Однако здесь нет агрегации данных вокруг какого-либо параметра. Данные не накапливаются во времени, они одномоментные. Но именно они и представляют наибольший интерес для аналитиков.

Предварительное агрегирование убивает исходную историю этих данных, но зато значительно экономит память. Так, например, исходные транзакции о поездках пассажиров общественного транспорта Санкт-Петербурга в предыдущих примерах содержали около 60 млн. транзакций (это соответствует 2-х недельному периоду обследования). Однако, когда поездки представляются агрегировано по видам транспорта в виде количества и средней продолжительности теряется много полезной информации для последующего возможного анализа. Например, по исходным данным о поездках можно выяснить много интересного о транспортных потоках города: о скорости движения транспорта в различных районах города, о наиболее загруженных маршрутах, оценить качество транспортной инфраструктуры и т.п.

Про студентов тоже, наверняка, можно узнать много полезной информации. Какие учебные заведения, какие специальности они предпочитают и т.п. Хранение исходных транзакций требовало, да и продолжает требовать больших вычислительных ресурсов, как в смысле используемой памяти, так и скорости обработки данных. Именно поэтому, в процессе сбора информации принято использовать методы агрегации, которые позволяют значительно сократить объемы анализируемых данных и минимизировать время последующей обработки.

Но прогресс последних лет в технологиях эффективного хранения, параллельной и облачной обработке данных сделал реальностью хранение и моментальную агрегацию больших массивов данных.

Если у аналитика есть возможность выбирать между фактоидами и транзакциям, он должен предпочесть последние, так как именно они позволят ему не ограничивать возможные варианты анализа данных и использовать любые подходящие модели.

## 4 Источники данных

При получении требуемых для анализа данных возникает много вопросов. Как собирают данные для анализа? Все ли данные собирают одинаково? Кто занимается сбором данных? Есть ли открытые источники данных? Ответам на эти вопросы и посвящен этот фрагмент лекции.

Различают следующие основные источники данных:

- социологические опросы;
- наблюдения;
- документы;
- результаты прямых измерений;
- социальные сети и внешние источники.

Рассмотрим последовательно каждый из них.

## 4.1 Социологические опросы

Многие исследователи считают социологический опрос наиболее простым и доступным методом сбора первичной социологической информации. Действительно, оперативность, простота, экономичность этого метода делают его весьма популярным по сравнению с другими методами исследований. Однако эта простая доступность нередко является кажущейся. Проблема состоит в получении качественных данных. А для этого необходимы соответствующие условия и соблюдение определенных требований. Какие это условия?

Во-первых, это наличие правильно составленных анкет для опроса, во-вторых – наличие надежного инструментария для заполнения и последующего анализа анкет и, наконец, это создание благоприятной, психологически комфортной обстановки для опроса. Первый и третий пункты явно зависят от профессионализма социологов, однако вмешательство аналитиков на этом этапе тоже уместно. Необходимо убедиться в том, что данные, которые будут являться ответами на вопросы, будут непротиворечивыми и охватывать все возможные варианты. Такие категории ответов принято называть «всеобъемлющими и взаимоисключающими».

Например, ответ на вопрос: «Ваша возрастная категория?» с возможными вариантами порядковых ответов может быть задан как интервалы от: 20-30, 30-40, 40-50, 50-60. Такие варианты ответов допускают неоднозначно выбранный ответ, т.к. человек в возрасте 30 лет не будет знать, что именно из двух возможных вариантов ему выбрать. А вот у человека в возрасте 65 лет вообще не будет никакого выбора.

Правильными вариантами в таком случае можно считать такие варианты возможных интервалов, которые покрывают все возможные ответы, и у которых нет никаких пересечений. Например, меньше 20, 21-30, 31-40, 41-50 и больше 50.

Очевидно, что сегодня любой аналитик предпочтет увидеть исходные данные не на бумажном носителе, а в электронном виде. К счастью, в настоящее время в свободном доступе имеется достаточное количество ин-

струментариев, позволяющих создать любую анкету для опроса в электронном виде. Одним из лучших примеров является инструмент Google Forms (<https://docs.google.com/forms/>), позволяющий создавать такие анкеты. Вы можете создать такую анкету для опроса, распространить ссылку на анкету и ожидать получения ответов.

На рисунке приведен пример анкеты, сгенерированной в инструменте Google Forms. Кроме того, такого рода инструменты позволяют собирать и визуализировать статистику полученных ответов.

### Инструментарий для заполнения ответов

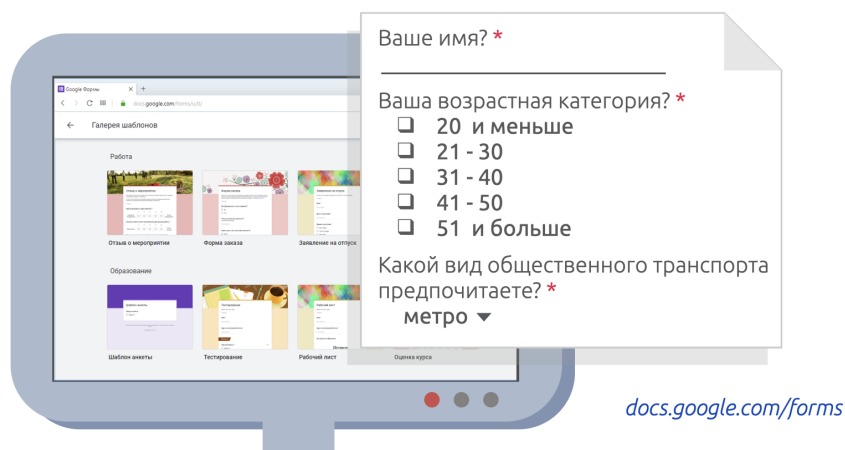


Рис. 7: Примеры Google Forms

## 4.2 Наблюдения

Наблюдения, как правило, выполняются при помощи специально нанятого персонала. Можно использовать наблюдения, чтобы понять естественное поведение. Особенно полезно использовать наблюдения, если объекты вашего интереса – не люди, и они не могут ответить на вопросы анкеты.

Один из главных недостатков этого метода состоит в том, данные, полученные в результате наблюдений, отличаются низкой достоверностью по причине наличия субъективного (т.е. человеческого) фактора.

Каждый человек видит событие со своей точки зрения, которая и обуславливает его отчет об этом событии. Конечно, эту погрешность можно уменьшить привлечением нескольких наблюдателей, чтобы собранные данные представляли различные точки зрения, но добиться высокой точности таких данных можно только с привлечением очень большого количества наблюдателей, что редко удается на практике.

Тем не менее, несмотря на указанные недостатки, в некоторых отраслях наблюдения являются единственным возможным способом сбора данных (например, наблюдения за поведением животных, растений и т.п.).

Как ни странно, этот способ получения данных до сих пор практически используется даже в тех отраслях человеческой деятельности, где уровень автоматизации процессов уже достаточно высок. Например, он широко распространен при обследовании городских транспортных потоков мегаполисов несмотря на то, что большинство из них связано с действующими информационными системами и обладают потенциалом для получения любых сведений о перемещении пассажиров.

### 4.3 Документы

Хорошо структурированные документы или хотя бы слабоструктурированные документы – прекрасный источник для получения данных, однако в большинстве случаев документы отличаются невнятной структурой, которая плохо поддается формальному разбору.

Особая беда – медицинские документы прошлых лет, которые, как правило, написаны вручную, плохим почерком, да еще и со странными обозначениями. В результате данные, накопленные врачами за десятилетия, остаются невостребованными в виду невозможности их формализации.

### 4.4 Результаты прямых измерений

Лучшие данные для аналитика – это результаты прямых измерений. Они менее других источников подвержены субъективным искажениям. Это – сырые транзакции без всякой агрегации. Такие данные, как правило, собираются при помощи всевозможных датчиков, установленных на изучаемых объектах (авиалайнеры, транспорт, бытовая техника, люди и т.п.). Эти устройства собирают невероятно большие массивы разнообразных данных. Прогресс в технологиях эффективного хранения, параллельной и облачной обработке данных сделал реальностью хранение таких массивов данных.

Пример сервиса, использующего результаты прямых измерений Flightradar24 (<https://www.flightradar24.com>). На этих данных уже основывается масса полезных сервисов, работающих в режиме реального времени.

Однако это лишь незначительная часть знаний, которые можно извлечь из указанных данных. Анализируя оперативные данные с разнообразных датчиков находящегося в полете авиалайнера можно не только отслеживать траекторию его полета, но и с заданной точностью предсказывать, какой ремонт потребуется авиалайнеру после приземления и многое другое. Специфика сбора данных, связанная с разнообразными приборами измерения, которые фиксируют транзакции в различных системах, реагируют на сторонние внешние воздействия, а иногда и выходят из строя, приводит к тому, что такие данные перед проведением анализа требуют предварительной обработки (согласования идентификации объектов в разных системах, очистки

от шумов, восстановления пропущенных значений и т.п.). Тем не менее, это обстоятельство не снижает их ценности и, главное, с каждым днем таких данных становится все больше, и нет сомнения, что они содержат немало полезных знаний, которые могут быть получены в результате содержательного анализа данных.

Надо заметить, что такие данные редко носят публичный характер и для их получения, как правило, приходится договариваться с владельцами соответствующих хранилищ данных.

## 4.5 Социальные сети

Социальные сети – еще один вид источника данных. Большинство социальных сетей предоставляет специальный интерфейс (API – application program interface) для доступа к открытым данным. Эти данные – отличный источник для анализа социальной активности и ее прогнозирования. По адресу [https://vk.com/dev/first\\_guide](https://vk.com/dev/first_guide) приведено описание такого интерфейса для сети ВКонтакте.

## 4.6 Внешние источники данных

Есть данные, которые не надо собирать, а можно поискать на одном из ресурсов публичных данных, доступных благодаря популярному в интернете движению за открытый контент и доступ. Многие правительства и организации установили политику доступности данных для обеспечения большей открытости и подотчетности обществу, а также, чтобы стимулировать развитие новых сервисов и продуктов. Источниками публичных данных могут быть: поисковые системы, и просто хранилища данных и правительственные базы данных и хранилища исследовательских учреждений. Рассмотрим конкретные примеры таких систем.

### 4.6.1 Примеры поисковых систем

Хорошо известные примеры поисковых систем – Google и Yandex. С помощью этих систем при условии правильной формулировки запроса можно найти много полезной и не очень информации, среди которой могут оказаться и нужные вам данные.

### 4.6.2 Примеры хранилищ данных

Следующий источник – открытые хранилища данных. Здесь нет никаких новостных лент, книг и журналов. Это именно разнообразные хранилища данных и ничего более. На рисунке представлены некоторые примеры таких хранилищ.

## Примеры хранилищ данных

Хранилище	Адрес
Re3data.org	<a href="http://www.re3data.org/">http://www.re3data.org/</a>
DataBib	<a href="http://databib.org/">http://databib.org/</a>
DataCite	<a href="http://www.datacite.org/">http://www.datacite.org/</a>
Dryad	<a href="http://data.dryad.org/">http://data.dryad.org/</a>
DataPortals	<a href="http://dataportals.org/">http://dataportals.org/</a>
Open Access Directory	<a href="http://oad.simmons.edu/oadwiki/Data_repositories">http://oad.simmons.edu/oadwiki/Data_repositories</a>
Gapminder	<a href="http://www.gapminder.org/data">http://www.gapminder.org/data</a>
Google Public Data Explorer	<a href="http://www.google.com/publicdata/directory">http://www.google.com/publicdata/directory</a>
IBM Many Eyes	<a href="http://www.manyeyes.com/software/analytics/manyeyes/datasets">http://www.manyeyes.com/software/analytics/manyeyes/datasets</a>
Knoema	<a href="http://www.knoema.com/atlas/">http://www.knoema.com/atlas/</a>

Рис. 8: Примеры хранилищ данных

### 4.6.3 Примеры правительственных баз данных

На рисунке примеры правительственных баз данных и среди них Портал открытых данных Российской Федерации. Здесь много полезной и хорошо структурированной информации.

## Примеры правительственных баз данных

База данных	Адрес
Всемирный банк	<a href="http://data.worldbank.org/">http://data.worldbank.org/</a>
ООН	<a href="http://data.un.org/">http://data.un.org/</a>
Open Data Index	<a href="https://index.okfn.org/">https://index.okfn.org/</a>
Open Data	<a href="http://od4d.net/">http://od4d.net/</a>
Данные правительства США	<a href="https://www.data.gov/">https://www.data.gov/</a>
Портал открытых данных Российской Федерации	<a href="https://data.gov.ru/">https://data.gov.ru/</a>

Рис. 9: Примеры правительственных баз данных

### 4.6.4 Примеры баз данных исследовательских учреждений

Следующий класс источников данных – базы данных исследовательских учреждений. В качестве примера приведем – AcademicTorrents (<http://academictorrents.com>). Здесь содержатся многочисленные наборы данных с хорошим, академическим описанием и, кроме того, здесь немало полезных учебных курсов.

Ссылаться на открытые источники данных – необходимо. Сам формат ссылки может зависеть от стиля, используемого в выбранном издании, но как

бы они не отличались друг от друга, всегда требуется указывать автора или владельца источника данных, названия и, собственно, ссылки на оригинальный источник.

## 5 Подготовка данных

Один из этапов, предваряющий анализ данных – это подготовка данных. Есть несколько распространенных операций по подготовке данных, с которыми приходится сталкиваться, особенно если исходные данные представляют собой сырые транзакции, собранные из разных источников.

К этим операциям относятся: загрузка данных в хранилища, разделение данных, приведение данных к одинаковым единицам измерения, преобразование к унифицированной лексике, объединение данных из разных источников, соединение данных из разных источников, заполнение отсутствующих числовых значений и очистка данных. Трудность состоит в том, что нет универсальных техник проведения этих операций. Каждый набор данных уникален, а некоторые техники подготовки можно использовать вообще только раз в жизни, но, тем не менее, успешное выполнение этих операций может оказать существенное влияние на результаты последующего анализа. Итак, рассмотрим основные операции подготовки данных.

### 5.1 Загрузка данных в хранилища

Это первый этап подготовки данных. Загрузка данных в единое хранилище, обладающее инструментарием, позволяющим проводить манипуляции над данными, позволит, в дальнейшем, произвести согласованное объединение данных из разных источников и провести первичную обработку данных. Как правило, в системах хранения данных существуют специальные утилиты, ориентированные на загрузку данных из внешних источников. Однако, даже на этом, этапе исследователя могут ожидать многочисленные сюрпризы: например, нечитаемые символы, типы данных не соответствующие обещанным спецификациям и т.п. В связи с этим существуют следующие рекомендации по загрузке:

- вычистить из исходных файлов все нечитаемые символы;
- загружать все исходные данные как текстовые поля (а с типами разбираться потом после загрузки в хранилище);
- саму загрузку (если данных действительно много) проводить непосредственно на сервере, где расположено хранилище.

## 5.2 Разделение данных

Простой пример задачи, с которой сталкиваются многие аналитики – это разделение имен и фамилий, а возможно и адресов. У вас могут быть исходные данные, где имена и фамилии прописаны в одной ячейке, а вам нужно их отделить друг от друга, так как специфика задачи требует знаний об именах и фамилиях по отдельности. Или у вас уже могут быть отдельные ячейки для имен и фамилий, но в некоторых случаях имена с фамилиями все равно записаны вместе и в разном порядке. Пример таких данных вы можете видеть на рисунке. Как разделить такие данные? Возможных решений много,

### Разделение данных

Diagram illustrating data separation. The left table shows data where names and surnames are combined in a single column. The right table shows the same data split into two columns, with red 'X' marks indicating incorrect splits for some entries.

Полное имя
Маша Иванова
Саша Петров
Оля Смирнова
Вася Соколов
Коля Сидоров
Аня Кузнецова

Имя	Фамилия
	Маша Иванова
Саша	Петров
Оля	Смирнова
Вася Соколов	
Коля	Сидоров
Аня	Кузнецова

Рис. 10: Пример разделения данных

но наиболее простое и достаточно эффективное – завести справочник наиболее распространенных имен и с его помощью проанализировать исходные данные. С теми именами, которые окажутся нестандартными, разбираться по отдельности. Но их окажется значительно меньше.

## 5.3 Преобразование данных к одинаковым единицам измерения

Этот этап акцентирует внимание на еще одном важном моменте при подготовке данных, а именно: необходимо проверять, чтобы все значения, относящиеся к одной переменной, были представлены в одинаковых единицах измерения. Типичный пример, медицинские данные из разных стран, где в одних странах вес измерен в фунтах, а в других – в килограммах (т.е. в разных единицах измерения). Необходимо выбрать любую из единиц измерения и конвертировать все значения к одной шкале с одинаковой единицей измерения. В противном случае их нельзя будет сравнивать и агрегировать. Пример такого преобразования вы можете видеть на рисунке.



## Преобразование данных к одинаковым единицам измерения

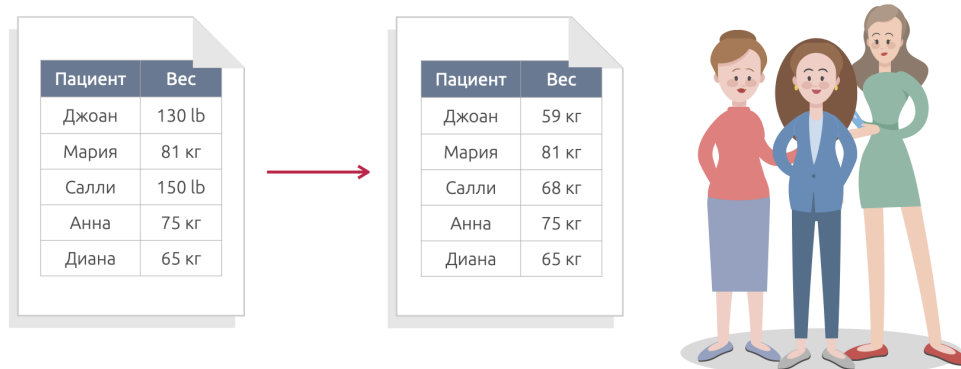


Рис. 11: Пример преобразования данных

## 5.4 Преобразование к унифицированной лексике

Этот этап обычно сопутствует вводу номинальных данных. Если номинальные данные при вводе формировались как свободный текст, а не выбирались из списка, то большие проблемы вам гарантированы. Например, если при вводе любимой дисциплины кто-то из студентов и напишет «Физкультура» с заглавной буквы, то наверняка найдутся те, кто напишет «Физ-ра» или «физкультура» не с заглавной буквы. Даже если вы знаете, что все эти

### Преобразование к унифицированной лексике

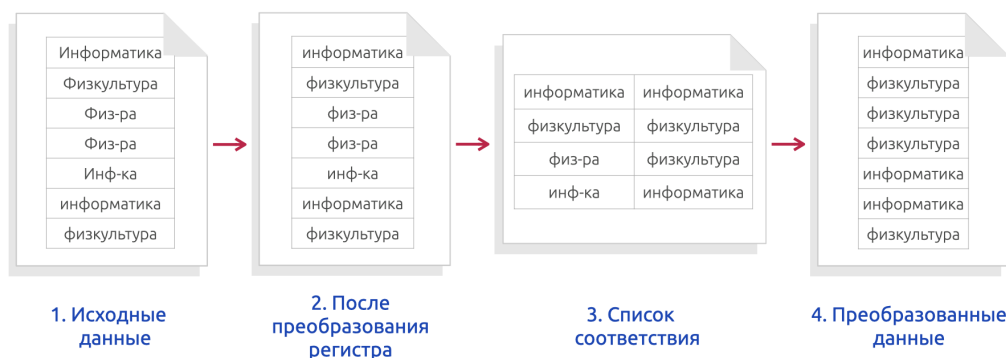


Рис. 12: Пример преобразования к унифицированной лексике

ответы обозначают одну и ту же дисциплину, они в дальнейшем ограничат ваши возможности для автоматической обработки данных и, более того, могут привести к неадекватным результатам. Необходимо преобразовывать такие данные к унифицированной лексике. Как это сделать? Возможный вариант такой: перевести все значения к единому регистру, выявить список разли-

чающихся значений, на его основе составить список соответствия, а затем провести замену значений.

## 5.5 Объединение данных из разных источников

Представьте, что вам нужно сформировать общий список жильцов дома. Но каждый подъезд дома предоставил свой список и вам необходимо на основе нескольких списков сформировать общий, т.е. выполнить операцию объединения различных сущностей. Какие проблемы могут встретиться в процессе выполнения такой операции? Напомним, что объединение – это теоретико-множественная операция, которая по определению может выполняться для элементов одинаковой структуры. Значит, для начала вам нужно убедиться, что предоставленные данные имеют требуемую структуру, а затем провести объединение. Предположим, для текущей задачи вам нужно знать имя ответственного жильца и номер квартиры. Пусть подъездов только 2 и они предоставили данные в следующем виде:

### Объединение данных из разных источников

1 подъезд

Ответств. жилец	Номер квартиры
Семен	11
Анна	12
Евгений	13

2 подъезд

Жилец	Номер квартиры	Общее количество жильцов в квартире
Сидор	14	4
Аделаида	15	2
Иннокентий	16	4



Рис. 13: Данные по подъездам

Как можно объединить такие данные? Как минимум, необходимо преобразовать данные к единой структуре. Для начала нужно провести соответствие полей, то есть понять, какие поля первой таблицы соответствуют полям второй таблицы (даже если они называются по-разному). После проведения соответствия нужно провести переименование полей к единому стилю. Следующий вопрос – что делать с полями, которые встречаются только в некоторых источниках (в нашем примере – Общее количество жильцов). Есть два варианта: удалить поля, которые встречаются не во всех источниках или наоборот, расширить определение структуры каждого источника и включить туда поля, если они присутствуют, хотя бы в одном из источников (может эти данные пригодятся для какой-то другой задачи). Затем можно проводить объединение данных. Примеры вы можете видеть на рисунках. Но

# Объединение данных из разных источников



Рис. 14: Первый вариант демонстрирует приведение данных к минимальной структуре, а затем выполнение операции объединения.

может случиться так, что объединяемые источники будут содержать данные в различных единицах измерения или не в унифицированной лексике. Разумеется, в этом случае сначала требуется произвести преобразование данных к одинаковым шкалам, единицам измерения и единой лексике, а затем проводить операцию объединения по одному из описанных ранее сценариев.

# Объединение данных из разных источников

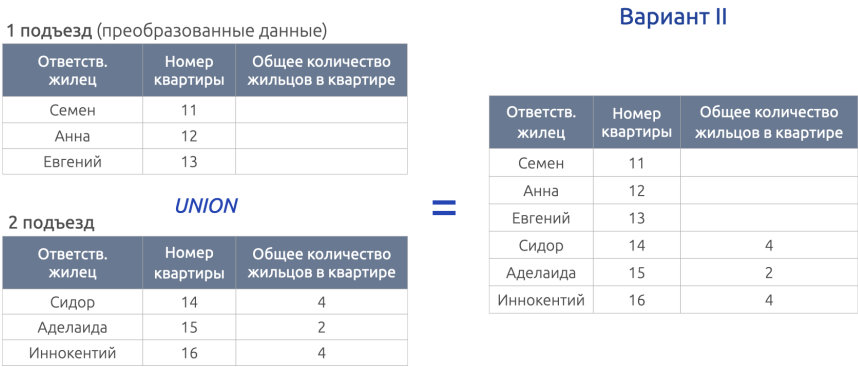


Рис. 15: Второй вариант демонстрирует преобразование данных к расширенной структуре, а затем объединение.

## 5.6 Соединение данных из разных источников

Разные источники данных об объектах реального мира дополняют друг друга и обогащают наше представление об объекте. Соединять данные из всех возможных источников необходимо, однако на этом пути есть определенные трудности.

Первая проблема – соответствие полей. Так же, как это было в задаче объединения данных из разных источников, необходимо исследовать соответствие полей и преобразовать названия к единому стилю.

Вторая проблема – преобразование данных в различных источниках к единым шкалам, единицам измерения и унифицированной лексике.

Третья проблема – идентификация данных, относящихся к одному и тому же объекту (например, выявление данных, про одного и того же покупателя в разных супермаркетах). Часто информация, поступающая из разных источников, имеет различную систему идентификацию объектов (где-то идентификатором является паспорт человека, где-то номер мобильного телефона и т.п.). Необходимо найти общие атрибуты, которые может и не являются идентификаторами в исходных системах, но, тем не менее, однозначным образом идентифицируют объект в каждой из систем (например, телефон, электронный адрес и т.п.).

И наконец, сами источники данных могут быть представлены в виде структур различных форматов (таблицы, JSON, XML и т.п.). Но это вполне решаемая задача, так как сегодняшние системы, управляющие хранилищами, зачастую позволяют выполнять запросы над данными различной структуры. В противном случае возможен дополнительный этап - конвертация данных к единому формату.

Только после всех этих шагов и можно проводить соединение данных из разных источников.

### Соединение данных из разных источников

Данные из спортивного клуба

Имя	Фамилия	Дата рождения	E-mail	Телефон	Паспорт	Вид спорта
Никита	Семенов	08.02.1998	NS@mail.ru	8(922)468-2929	4004 271492	плавание
Юра	Алексеев	03.01.1999	Y.Alex@mail.ru	8(931)852-9582	3003 262899	волейбол

Данные из супермаркета

Имя покупателя	Фамилия покупателя	E-mail	Телефон	Категория
Ник	Семенов	NS@mail.ru	8(922)468-2929	студент
Юрий	Алексеев	Y.Alex@mail.ru	8(931)852-9582	служащий

Результат соединения

Имя	Фамилия	Дата рождения	E-mail	Телефон	Паспорт	Вид спорта	Категория
Никита	Семенов	08.02.1998	NS@mail.ru	8(922)468-2929	4004 271492	плавание	студент
Юрий	Алексеев	03.01.1999	Y.Alex@mail.ru	8(931)852-9582	3003 262899	волейбол	служащий

Рис. 16: Исходные данные спортивного клуба и супермаркета.

## 5.7 Заполнение отсутствующих численных значений

Одна из типичных проблем при работе с данными – пустые или не полностью заполненные числовые поля. Если данные просто по забывчивости не были собраны, то, возможно, вы сможете вернуться к источнику и заполнить

неопределенные значения, но, возможно, что у вас больше не будет доступа к этому источнику. Например, это показания датчиков, и никаких других данных просто не будет. Есть два подхода при работе с такими данными:

- выделение таких полей специальными значениями (и исключение их из дальнейшего анализа);
- аппроксимация пропущенных значений на основе существующих данных.

### 5.7.1 Аппроксимация пропущенных значений

В большинстве случаев аппроксимация пропущенных численных значений осуществляется за счет вычисления среднего значения по всему набору данных. В то же время некоторые виды данных, например, временные ряды, требуют другого подхода и вычисляются на основе среднего значения, так называемого скользящего окна, состоящего из ближайших соседей. Ширина окна (или количество соседей), которое берется для усреднения, зависит от задачи, но, как правило, ограничиваются 4-6 соседями (слева и справа). Однако в некоторых случаях приходится пользоваться значительно ме-

#### Заполнение отсутствующих численных значений



Остановка	Время прибытия	Номер остановки
Метро Приморская	16:00	1
Наличная улица	?	2
Малый проспект	?	3
Гаванская улица	?	4
Шкиперский проток	?	5
Средний проспект	?	6
Трамвайный парк	?	7
12 линия	16:35	8

Остановка	Время прибытия	Номер остановки
Метро Приморская	16:00	1
Наличная улица	16:05	2
Малый проспект	16:10	3
Гаванская улица	16:15	4
Шкиперский проток	16:20	5
Средний проспект	16:25	6
Трамвайный парк	16:30	7
12 линия	16:35	8

Рис. 17: Пример пропущенных значений времени прибытия трамвая.

нее стандартными алгоритмами, которые сильно привязаны к анализируемой предметной области. Например, при прохождении маршрута были потеряны сведения о времени прохождения нескольких последовательных остановок. Надо восстановить это время на основе исторических данных по временам, зафиксированным до потери и после.

Если у нас, действительно, нет более никаких данных, можно вычислить временной интервал между заполненными полями и равномерно разделить это время между остановками. Но если есть сведения о том, как ранее трамвай проходил эти остановки, то следует определить соотношение по времени прохождения остановок и распределить временной интервал между остановками пропорционально этому соотношению.

## 5.8 Очистка данных

Как правило, очистка данных сводится к выполнению следующих работ:

- устранение дубликатов;
- контроль диапазонов;
- сравнение с образцами/регулярными выражениями.

### 5.8.1 Устранение дубликатов

Дубликаты могут появляться в исходных данных по причине разного рода технических сбоев. Они характерны для сырых транзакций. Наличие дубликатов может быть причиной получения неверных результатов при последующем агрегировании данных. Например, автоматический подсчет строк в справочнике, приведенном на экране, приведет к неверному результату о количестве видов транспорта.

Как правило, дубликаты легко выявляются и могут быть удалены инструментарием для манипуляций над данными, предусмотренном в хранилище.

### 5.8.2 Проверка регулярных выражений

Некоторые виды данных должны быть заданы в определенном формате. Например, электронный адрес или номер телефона. Достоверность таких данных можно проверять путем сравнения с шаблоном. Шаблон принято за-

Атрибут	Шаблон
Электронный адрес	%@%.%
Телефон	+7(DDD) DDD-DD-DD

давать в виде, так называемого, регулярного выражения (regular expressions). Способ задания шаблона варьируется в разных инструментариях, но на сегодняшний день присутствует практически в любых системах.

Разумеется, возникает вопрос – «Что делать с данными, не прошедшими проверку?». Вариантов, как минимум, три: попытаться получить адекватные данные, либо заполнить значения явно заданными символами, означающими неопределенное значение, либо вообще исключить из анализа.

### 5.8.3 Контроль диапазонов

Еще один из этапов проверки данных – контроль диапазонов. Контроль диапазонов – это на первый взгляд очень простая процедура, которую мы используем в числовых переменных, чтобы увидеть, находятся ли какие-либо

значения в этом наборе данных выше или ниже крайних допустимых значений для этой переменной. Однако возникает вопрос, как определить этот диапазон. Оптимально, если этот диапазон задается правилами предметной области. Например, это баллы, которые в принципе возможны в диапазоне от 0 до 100. Значит, если в исходных данных присутствует значение 787 – то это явная ошибка.

В этом случае уже в момент загрузки данных в хранилище мы можем тем или иным образом предусмотреть контроль загружаемых значений и не допустить значения с явно недостоверной информацией. Хуже, когда такого явного знания диапазонов у нас нет, а некоторые значения нам все равно кажутся подозрительными. Интуитивно – это значения, которые сильно отличаются от остальных. Но как формализовать понятие «сильно отличаются от остальных?» Это, так называемые, выбросы. В статистике есть формальные методы, которые позволяют обнаруживать выбросы, и мы их непременно рассмотрим, но не в рамках текущего раздела, так как для определения выбросов придется уточнить некоторые статистические понятия.

В следующей лекции мы рассмотрим некоторые инструменты для обработки данных, которые позволят в дальнейшем на практике применить полученные знания.