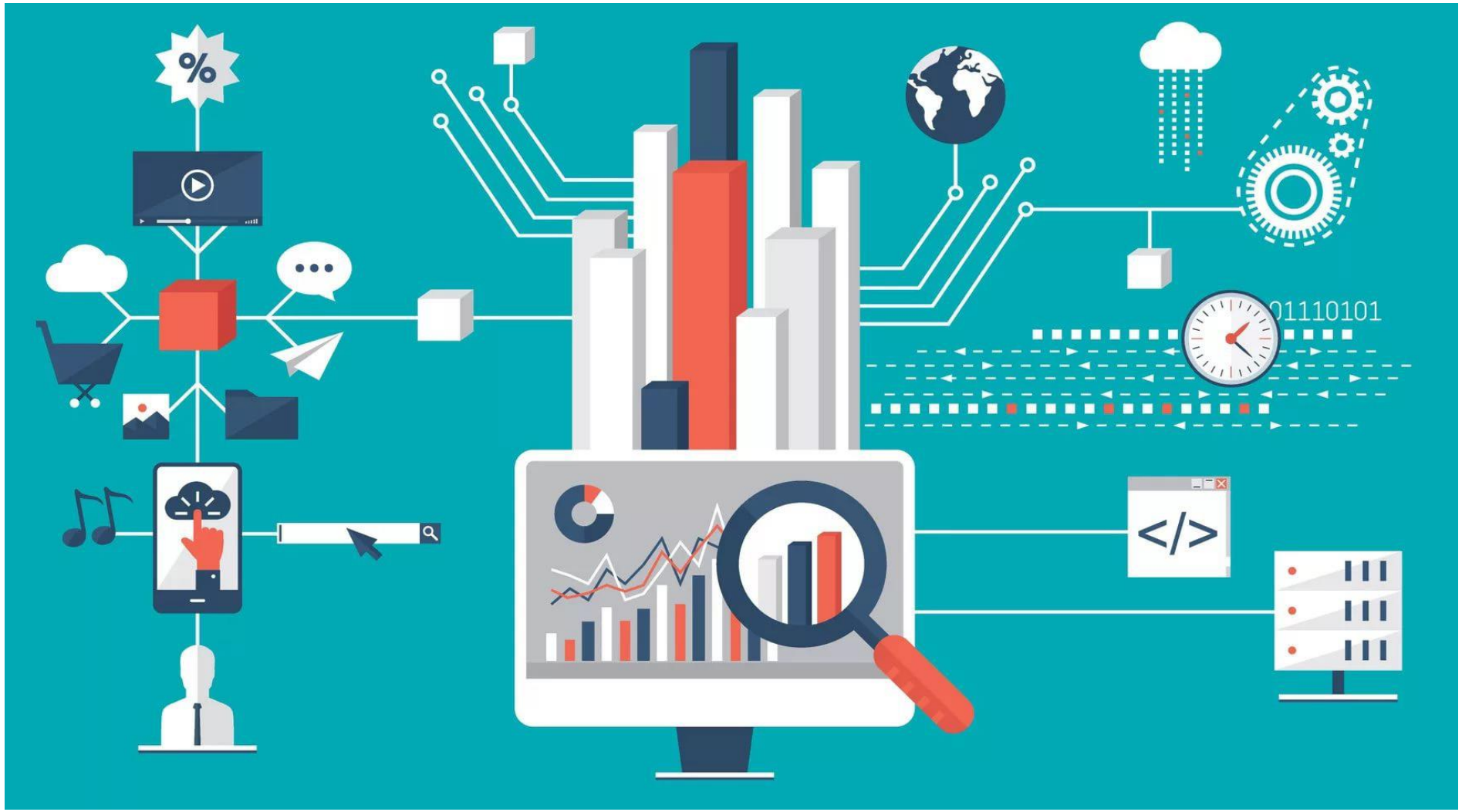
# Анализ больших наборов данных




**2020**

# Анализ больших наборов данных

## Данные в цифровой экономике

Мы живём в мире, насыщенным данными.

Веб-сайты порождают данные при  любом нажатие любого пользователя.

Смартфоны и телефоны накапливают сведения о вашем местоположении и скорости в ежедневном и ежесекундном режиме.

"Оцифрованные" селферы (личные фото и другие данные) шагомеры, которые не переставая записывают сердечные ритмы, особенности движения, схемы питания и сна.

"Умные" авто собирают сведения о манерах вождения своих владельцев.

"Умные"  дома накапливают данные об образе жизни своих обитателей.

"Умные" маркетологи накапливают данные  о наших покупательских привычках.

Сам Интернет представляет собой огромный граф знаний, который, среди всего прочего, содержит обширную гипертекстовую энциклопедию, специализированные базы данных о фильмах, музыке, спортивных результатах, игровых автоматах и слишком много статистических отчетов (причем некоторые почти соответствуют действительности!) от слишком большого числа государственных исполнительных органов, и все это для того, чтобы вы объяли необъятное.

В этих данных кроятся ответы на бесчисленные вопросы, которые никто даже не думает задавать.

# Анализ больших наборов данных

## Наука о данных

Существует шутка, что аналитик данных — это тот, кто знает статистику лучше, чем специалист в области информатики, а информатику — лучше, чем специалист в области статистики.

Некоторые имеют учёные степени доктора наук с впечатляющей историей публикаций, в то время как другие никогда не читали академических статей.

В значительной мера неважно, как определять понятие науки о данных, потому что всегда можно найти практикующих аналитиков данных, для которых это определение будет всецело и абсолютно неверным .

Определим, что аналитика данных - это тот, кто извлекает ценные наблюдения из запутанных данных. В наши дни мир переполнен людьми, которые пытаются превратить данные в ценные наблюдения.

Например, компания Facebook просит вас указывать свой родной город и нынешнее местоположение, якобы чтобы облегчить вашим друзьям находить вас и связываться с вами. Но она также анализирует эти местоположения, чтобы определить схемы глобальной миграции и места проживания фанатов различных футбольных команд.

Крупный оператор розничной торговли Target отслеживает покупки и взаимодействия онлайн и в магазине. Он использует данные, чтобы строить прогнозные модели продаж.

# Анализ больших наборов данных

## Примеры больших наборов данных

В области подбора кадров используют личностную аналитику (people analytics) и глубокий анализ текста для отбора кандидатов, отслеживания настроения работников и изучения неформальных связей среди коллег. Применение статистики позволяет нанимать эффективных сотрудников. Финансовые учреждения используют большие наборы данных для прогнозирования рынка ценных бумаг, вычисления риска предоставления ссуд и привлечения новых клиентов. В настоящее время по меньшей мере 50% торговых сделок по всему миру выполнялось автоматически на основании алгоритмов, разработанных специалистами по использованию больших данных и методов науки о данных.

# Анализ больших наборов данных

## Примеры больших наборов данных

Правительственные организации также хорошо осведомлены о ценности данных. Многие правительственные организации не только используют собственных аналитиков для поиска ценной информации, но и выкладывают свои результаты в открытый доступ. Данные могут использоваться для глубокого анализа или построения приложений, управляемых данными. Университеты используют большие наборы данных в своих исследованиях и для повышения качества учебного процесса. Бурное развитие массовых открытых дистанционных курсов породило большой объем данных, на основании которых университеты могут изучать, как этот тип обучения дополняет традиционные программы.

# Анализ больших наборов данных

## Категории наборов данных

В области больших наборов данных встречается много разных типов данных, для каждого из которых требуются свои инструменты и методы. Основные категории данных следующие:

- Структурированные.
- Неструктурированные.
- На естественном языке.
- Машинные.
- Графовые.
- Аудио, видео и графика.
- Потоковые.

# Анализ больших наборов данных

## Структурированные данные

Структурированные данные зависят от модели данных и хранятся в фиксированном поле внутри записи. Соответственно, структурированные данные часто бывает удобно хранить в таблицах, в таблицах Excel, базах данных SQL, ADABAS, они являются основным средством управления и обращения с запросами к данным, хранящимся в базах данных. Также иногда встречаются структурированные данные, которые достаточно трудно сохранить в традиционной реляционной базе данных (один из примеров — иерархические данные, например генеалогическое дерево).

В реальности, мир не состоит из структурированных данных; просто это представление удобно для человека и программ ЭВМ.

# Анализ больших наборов данных

## Неструктурированные данные

Неструктурированные данные трудно подогнать под конкретную модель данных, потому что их содержимое зависит от контекста или имеет переменный характер. Один из примеров неструктурированных данных — обычные сообщения электронной почты. Хотя сообщение содержит структурированные элементы (отправитель, заголовок, тело), одни и те же задачи могут решаться множеством разных способов, например, существует бесчисленное количество вариантов упоминания конкретного человека в сообщениях.

Проблема дополнительно усложняется существованием тысяч языков и диалектов.

Сообщение электронной почты, написанное человеком, также является идеальным примером данных на естественном языке.

# Анализ больших наборов данных

## Данные на естественном языке

Данные на естественном языке составляют особую разновидность неструктурированных данных; обработка таких данных достаточно сложна, потому что она требует знания как лингвистики, так и специальных методов науки о данных.

В обработке данных на естественном языке добилось успеха в области распознавания сущностей, распознавания тематических областей, обобщения, завершения текста и анализа эмоциональной окраски, но модели, адаптированные для одной предметной области, плохо обобщаются для других областей.

Сама концепция смысла выглядит спорно. Два человека слушают один разговор; вынесут ли они одинаковый смысл из него? Даже смысл отдельных слов может изменяться в зависимости от настроения говорящего.

# Анализ больших наборов данных

## Машинные данные

К машинным данным относится информация, автоматически генерируемая компьютером, процессом, приложением или устройством без вмешательства человека. Машинные данные становятся одним из основных источников информации, и ситуация вряд ли изменится. Считается, что рыночная стоимость промышленного Интернета в 2020 году составит приблизительно 540 миллиардов долларов. По разным оценкам, количество узлов сети в 2020 году в 26 раз превысит численность населения. Эта сеть часто называется Интернетом вещей.

Анализ машинных данных из-за их громадных объемов и скоростей сильно зависит от инструментов с высокой масштабируемостью. К примерам машинных данных относятся журналы веб-серверов, записи детализации звонков, журналы сетевых событий и телеметрии.

# Анализ больших наборов данных

## Графовые, или сетевые данные

Под «графом» в данном случае имеется в виду понятие графа из математической теории графов — математическая структура для моделирования попарных отношений между объектами. В графовых, или сетевых, данных особое внимание уделяется связям между объектами. Графовые структуры данных используют узлы, ребра и свойства для представления и хранения графических данных. Графовые данные естественным образом подходят для представления социальных сетей, а их структура позволяет вычислять такие специфические метрики, как влияние участников и кратчайший путь между двумя людьми. Графовые данные встречаются на многих веб-сайтах социальных сетей. Для хранения графовых данных используются графовые базы данных, а для построения запросов к ним — такие специализированные языки запросов, как SPARQL.

# Анализ больших наборов данных

## Аудио, видео и графика

Аудио, видео и графика — типы данных, ставящие непростые задачи перед специалистом Big Date и Bata Ыcience. Задачи, тривиальные с точки зрения человека (например, распознавание объекта на картинке), оказываются сложными для компьютера.

Уже созданы алгоритм который получает на входе содержимое экрана и учится интерпретировать эти данные в сложном процессе глубокого обучения. Алгоритм обучения получает данные, генерируемые компьютерной игрой, т. е. потоковые данные.

# Анализ больших наборов данных

## Потоковые данные

Потоковые данные могут принимать почти любую из перечисленных форм, однако у них имеется одно дополнительное свойство. Данные поступают в систему при возникновении некоторых событий, а не загружаются в хранилище данных большими массивами.

И хотя формально они не являются отдельной разновидностью данных, целесообразно выделять их в особую категорию, потому что придется приспособить свой рабочий процесс для работы с потоковой информацией.

Примерами потоковых данных могут служить раздел «Что происходит?» в Твиттере, прямые трансляции спортивных и музыкальных мероприятий и т.п.

# Анализ больших наборов данных

## Методы анализа больших наборов данных

Будем рассматривать и анализировать данные очень большого объема, не помещающихся в оперативную намять. Подобные данные относятся к вебу или к данным, полученным из веба. Интересен алгоритмический подход т.е. применение алгоритмов к данным, а не использование данных для «обучения» той или иной машины.

## Рассмотрим следующие методы анализа:

1. Распределенные файловые системы и технология распределения-редукции (map-reduce) как средство создания параллельных алгоритмов, успешно справляющихся с очень большими объемами данных.
2. Поиск по сходству, в том числе такие важнейшие алгоритмы, как MinHash и хэширование с учетом близости (locality sensitive hashing).
3. Обработка потоков данных и специализированные алгоритмы для работы с данными, которые поступают настолько быстро, что либо обрабатываются немедленно, либо теряются.
4. Принципы работы поисковых систем, в том числе алгоритм Google PageRank, распознавание ссылочного спама и метод авторитетных и хаб- документов.
5. Частые предметные наборы, в том числе поиск ассоциативных правил, ана¬лиз корзины, алгоритм Apriori и его усовершенствованные варианты.

# Анализ больших наборов данных

## Методы анализа больших наборов данных

6. Алгоритмы кластеризации очень больших многомерных наборов данных.
7. Две важные для веб-приложений задачи: управление рекламой и рекомендательные системы.
8. Алгоритмы анализа структуры очень больших графов, в особенности графов социальных сетей.
9. Методы получения важных свойств большого набора данных с помощью понижения размерности, в том числе сингулярное разложение и латентно-семантическое индексирование.
10. Алгоритмы машинного обучения, применимые к очень большим наборам данных, в том числе перцептроны, метод опорных векторов и градиентный спуск.

# Анализ больших наборов данных

## Машинное обучение

Машинное обучение заключается в извлечении знаний из данных.

Это научная область, находящаяся на пересечении статистики, искусственного интеллекта и компьютерных наук, также известная как прогнозная аналитика или статистическое обучение.

В последние годы применение методов машинного обучения в повседневной жизни стало обыденным явлением. Многие современные веб-сайты и электронные устройства используют алгоритмы машинного обучения, начиная с автоматических рекомендаций по просмотру фильмов, заказа еды или покупки продуктов и заканчивая персонализированными онлайн-радиотрансляциями и распознаванием друзей на фотографиях.

Когда вы видите сложный сайт типа Facebook или Amazon, то весьма вероятно, что каждый раздел сайта содержит несколько моделей машинного обучения.

# Анализ больших наборов данных

## Задачи машинного обучения

Наиболее успешные алгоритмы машинного обучения — это те, которые автоматизируют процессы принятия решений путем обобщения известных примеров.

В этих методах, известных как обучение с учителем или контролируемое обучение, пользователь предоставляет алгоритму пары “объект-ответ”, а алгоритм находит способ получения ответа по объекту. В частности, алгоритм способен выдать ответ для объекта, которого он никогда не видел раньше, причем без какой-либо помощи человека.

Пример классификации спама с использованием машинного обучения, изначально пользователь предъявляет алгоритму большое количество писем (**объекты**) вместе с информацией о том, является каждое из этих писем спамом или нет (**ответы**). После проведения обучения для любого нового электронного письма алгоритм самостоятельно сможет вычислить вероятность, с которой это письмо можно отнести к спаму.

Алгоритмы машинного обучения, которые учатся на парах “**объект-ответ**”, называются алгоритмами обучения **с учителем**, так как “учитель” предоставляет алгоритму правильный ответ для каждого наблюдения, по которому происходит обучение.

# Анализ больших наборов данных

## Примеры реальных задач машинного обучения с учителем

*Определение почтового индекса по рукописным цифрам на конверте*

Здесь объектом будет сканированное изображение почерка, а ответом — фактические цифры почтового индекса. Чтобы создать набор данных для построения модели машинного обучения, вам нужно собрать большое количество конвертов. Далее вам потребуется самостоятельно просмотреть почтовые индексы и сохранить цифры в виде ответов.

*Обнаружение мошеннической деятельности в сделках по кредитным картам*

Здесь объект — запись о транзакции по кредитной карте, а ответ — информация о том, является ли транзакция мошеннической. Предположим, вы — учреждение, выдающее кредитные карты, сбор данных подразумевает сохранение всех транзакций и запись сообщений клиентов о мошеннических транзакциях.

# Анализ больших наборов данных

## Принцип Бонферрони

Пусть имеются какие-то данные, и мы ищем в них события определенного вида. Можно ожидать, что такие событие встретятся, даже если данные выбраны абсолютно случайно, а количество событий будет расти вместе с объемом данных. Эти события «фиктивные» в том смысле, что у них нет никакой причины, помимо случайности данных, а в случайных данных всегда встретится какое-то количество необычных признаков, которые, хотя и выглядят значимыми, на самом деле таковыми не являются.

Теорема математической статистики, известная под названием ***поправка Бонферрони,*** дает статистически корректный способ избежать большинства таких ложноположительных ответов на поисковый запрос.

Не вдаваясь в технические детали, мы предложим ее неформальный вариант, *принцип* ***Бонферрони,*** который поможет избежать трактовки случайных фактов как реальных.

Вычислите ожидаемое число искомых событий в предположении, что данные случайны. Если это число существенно больше количества реальных событий, которые вы надеетесь обнаружить, то следует ожидать, что почти все найденные события фиктивные, т. е. являются статистическими артефактами, а не свидетельством в пользу того, что вы ищете. Это наблюдение и есть неформальный принцип Бонферрони.

# Анализ больших наборов данных

## Пример применения принципа Бонферрони

Допустим, мы полагаем, что где-то действуют «злоумышленники», и хотим их обнаружить. Допустим также, что есть основания полагать, что злоумышленники периодически встречаются в гостинице, чтобы спланировать свой злой умысел. Сделаем следующие предположения о размере задачи:

Есть миллиард людей, среди которых могут быть злоумышленники.

Любой человек останавливается в гостинице один день из 100.

Гостиница вмещает 100 человек. Следовательно, 100000 гостиниц будет достаточно, чтобы разместить 1 % от миллиарда людей, которые останавливаются в гостинице в каждый конкретный день.

Мы изучаем данные о регистрации в гостиницах за 1000 дней.

Чтобы найти в этих данных злоумышленников, мы будем искать людей, которые в два разных дня останавливались в одной и той же гостинице. Допустим, однако, что в действительности никаких злоумышленников нет. То есть все ведут себя случайным образом, с вероятностью 0,01 решая в данный день остановиться в какой-то гостинице и при этом случайно выбирая одну из $10^5$ гостиниц.

Найдем ли мы пары людей, которые выглядят как злоумышленники?

Найдём, потребуются знания математики.

# РЕКОМЕНДУЕМАЯ ЛИТЕРАТУРА


1. Юре Лесковец, Ананд Раджараман, Джеффри Д. Ульман Анализ больших наборов данных. / Пер. с англ. Слинкин А. А. - М.: ДМК Пресс, 2016. - 498 с.: ил. ISBN 978-5-97060-190-7

2. Форман Дж. Много цифр: Анализ больших данных при помощи Excel / Джон Форман ; Пер. с англ. А. Соколовой. - М.: Альпина Паблишер, 2016. — 461 с. ISBN 978-5-9614-5032-3

3. Вайгенд, Андреас. BIG DATA. Вся технология в одной книге / Андреас Вайгенд ; [пер. с англ. С. Богданова]. — Москва : Эксмо, 2018. - 384 с. (Top Business Awards). ISBN 978-5-04-094117-9

4. Сенько А. Работа с BigData в облаках. Обработка и хранение данных с примерами из Microsoft Azure. — СПб.: Питер, 2019. — 448 с.: ил. — (Серия «Для профессионалов»). ISBN 978-5-4461-0578-6

5. О'Нил, Кэти. Убийственные большие данные. Как математика превратилась в оружие массового поражения / Кэти О'Нил; [перевод с английского В. Дегтяревой]. — Москва: Издательство АСТ, 2018. — 320 с. — (Цифровая экономика и цифровое будущее). ISBN 978-5-17-982583-8

# РЕКОМЕНДУЕМАЯ ЛИТЕРАТУРА


6. Силен Дэви, Мейсман Арно, Али Мохамед Основы Data Science и Big Data. Python и наука о данных. - СПб.: Питер, 2018. — 336 с.: ил. (Серия «Библиотека программиста»). ISBN 978-5-496-02517-1

7. Грас Дж. Data Science. Наука о данных с нуля: Пер. с англ. — СПб.: БХВ-Петербург, 2017, —336 с.: ил. ISBN 978-5-9775-3758-2

8. Мюллер, Андреас, Гвидо, Сара. Введение в машинное обучение с помощью Python. Руководство для специалистов по работе с данными.: Пер. с англ. — СПб.: ООО “Альфа-книга”, 2017. — 480 с.: ил. — Парал. тит. англ. ISBN 978-5-9908910-8-1 (рус.)

9. Ын Анналин, Су Кеннет Теоретический минимум по Big Data. Всё, что нужно знать о больших данных. — СПб.: Питер, 2019. — 208 с.: ил. — (Серия «Библиотека программиста»). ISBN 978-5-4461-1040-7