Курс «Smart Data»
Лекция №1 «Введение»

План лекции:
1. Содержание модуля
2. Основные определения
3. Важность данных для процессов цифровой трансформации
4. Проблематика информации и данных
5. Виды данных
6. Пирамида данных

1. Содержание модуля

- **Лекции:**
  1. Введение, Дополнительные материалы.
  2. Графы знаний.
  3. Графы знаний (продолжение).
  4. Практическое использование SPARQL.
  5. *(на самостоятельную проработку) Big data и хранилища больших данных.*
- **Лабораторные работы.**

2. Основные определения

Определение #1
Smart Data - это такая технология обработки данных, которая позволяет интегрировать различные источники данных (включая Big Data), устанавливать взаимосвязи между данными в этих источниках, анализировать данные из различных источников совместно. Цель Smart Data – обеспечить процессы принятия решений и операционную деятельность.
Fernando Iafrate, From Big Data to Smart Data, First published:27 February 2015, Print ISBN:9781848217553 |Online ISBN:9781119116189 |DOI:10.1002/9781119116189, Copyright © 2015 John Wiley & Sons, Inc.
Многие данные являются «большими» (с точки зрения объема, скорости и т. д.), но насколько они «Smart», то есть имеют значение для бизнеса?

Умные данные следует рассматривать как набор технологий и процессов, а также связанные с ним структуры (Центры компетенции бизнес-аналитики (BICC)), которые включают все значения, взятые из данных. Такие центры могут создаваться в организациях (корпорациях) и предназначены для определения задач, ролей, обязанностей и процессов для поддержки и продвижения эффективного использования бизнес-аналитики (BI) в организации.

Strange, K. H., Hostmann, B. (22 July 2003), BI Competency Center Is Core to BI Success, Gartner Research

Определение #2

Smart Data - это данные, в которые добавлено явное семантическое содержание посредством формализация метаданных (по определению, характеристике или процессу моделирования). Термин Smart в широком смысле говорит о том, что Smart (интеллект, сообразительность) - это измеримая величина по ее степени, другими словами, существуют степени яркости, точности, аккуратности, структуры и абстракции, в которой данные могут быть формально описаны.

Определение намеренно широко в смысле, что оно применимо как к сырым, обрабатываемым приложениями, к метаданным, к моделям (данные, бизнес-процессы и другие), и даже к метамоделям. Smart Data - это продукт тщательного и открытого процесса, который описывает актуальную информацию, используемую предприятием. Полезная информация включает в себя данные, описывающие события, явления, материалы, процессы, процедуры, действия, приложения, структуры, отношения или сами данные.

James A. Rodger, Smart Data: Enterprise Performance Optimization Strategy (Wiley Series in Systems Engineering and Management), 2012

Важные замечания относительно определения «Smart Data».

*Если соотносить определение **Smart Data** с известными **V**s из определения Big Data (объем, скорость, разнообразие и достоверность), то **Smart Data** связаны с **Достоверностью (Veracity)** наряду с еще одним часто используемым V: **Ценностью (Value).***

Используя интеллектуальные данные, мы ориентируемся на ценные данные и часто меньшие наборы данных, которые можно превратить в данные для обеспечения действий или операций предприятия, и в эффективные результаты для решения проблем клиентов и бизнеса. Речь идет об анализе и интерпретации данных, чтобы мы могли сделать процесс принятия решений и бизнес-функции предприятия основанными на данных, путем помещения их в контекст целей предприятия.

***Smart Data** - это большие данные, которые превращаются в данные, которые можно использовать для действий, которые доступны в режиме реального времени для различных бизнес-результатов, будь то промышленные приложения, маркетинг на основе данных или оптимизация процессов.*

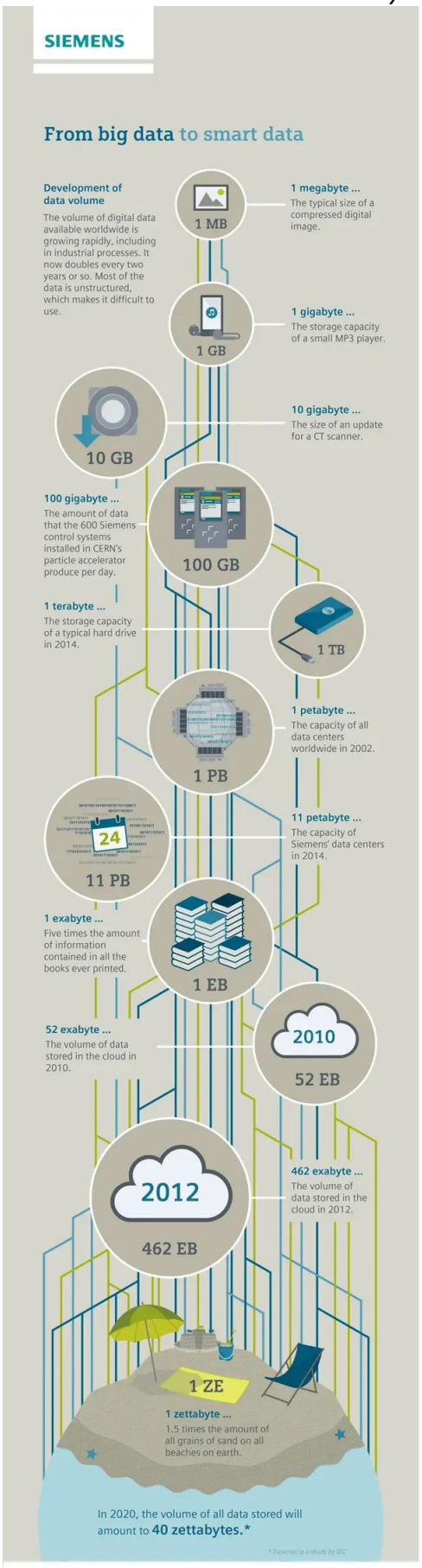
Используя интеллектуальные данные, мы действительно ищем способы избавиться от шума самого аспекта объема, так же как быстрые данные относятся к элементу скорости. Например, в контексте маркетинга и обслуживания клиентов интеллектуальные данные в основном рассматриваются с точки зрения гиперперсонализации.

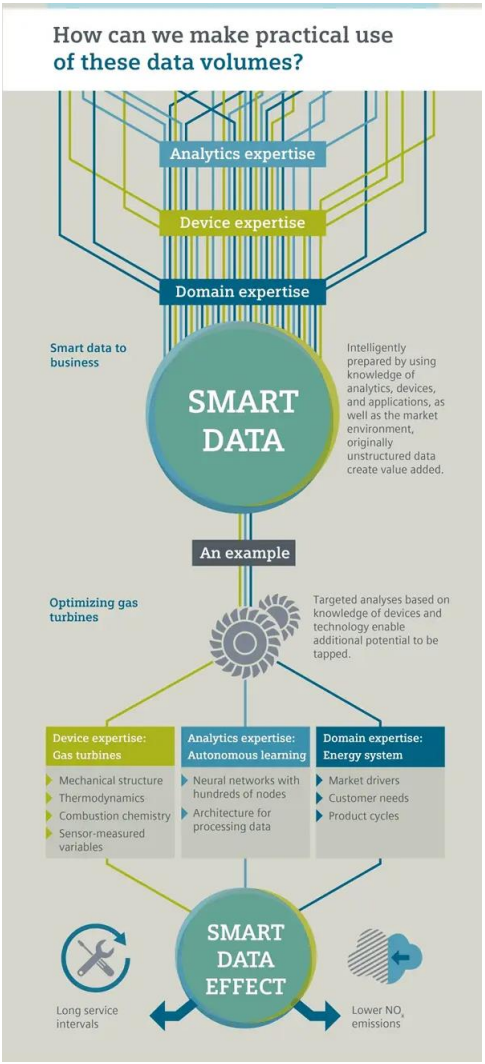
*Повышенное внимание к **Smart Data**, а не к **Big Data**, тесно связано с грядущей экономикой алгоритмов. Искусственный интеллект все чаще используется в бизнес-приложениях и для работы с **Big Data** и Интернета вещей **(IoT)**.*

Большинство имеющихся данных представляют собой неструктурированные данные, и только с помощью искусственного интеллекта и аналитики неструктурированные данные можно превратить в интеллектуальные данные и данные, пригодные для осуществления деятельности предприятия.

*Существенный вопрос в постоянно растущем объеме данных заключается в том, как использовать эти объемы на практике, и без аналитики, интерпретации и алгоритмов это невозможно. Приведенная ниже инфографика, разработанная **Siemens***

для промышленных приложений, показывает проблему объема данных и необходимость перехода от больших данных к интеллектуальным данным.





Полезность и аналитика данных

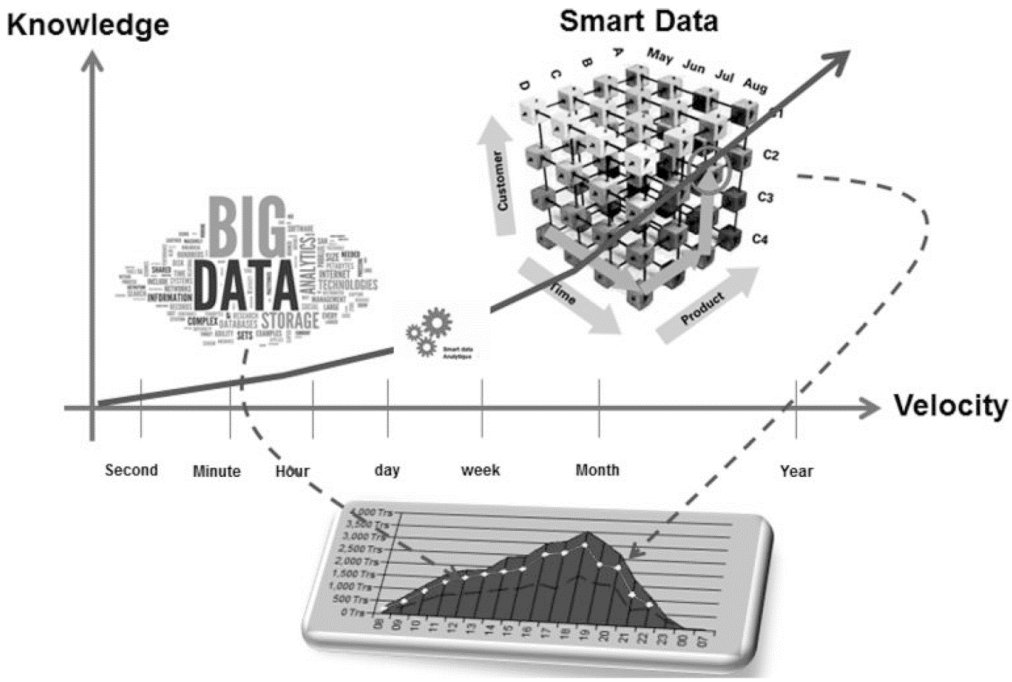
Чтобы сделать данные актуальными, важно с самого начала знать, почему мы хотим их собирать и использовать. Даже если большие данные необходимы и в нашем распоряжении есть много источников данных, умные данные - это не только объем и не только технологии. Речь идет о полезности, с многими уровнями интеллекта, построенными на том, как мы собираем, обрабатываем, анализируем, храним, интерпретируем и улучшаем данные, чтобы воздействовать на них и эффективно делать их полезными. Подумайте, например, о бумажных источниках и интеллектуальном распознавании документов. Или о том, как неструктурированные данные оптимизируются, маршрутизируются и превращаются в аналитические данные и потоки с использованием искусственного интеллекта и интеллектуального управления информацией.

В условиях быстро меняющейся динамики бизнеса скорость обработки данных стала важной и в экономике. В этом контексте существуют быстрые оперативные данные, которые соответствуют быстро меняющейся динамике рынков и новым требованиям клиентов.

Умные технологические решения

Тем не менее, очень немногие бизнес-процессы четко определены и достаточно стабильны для построения бизнес-правил и четко структурированных рабочих процессов для их

автоматизации. Большинство бизнес-процессов являются сложными и очень динамичными. Интеллектуальные технологические решения предназначены для решения этих сложных задач, которые обычно включают ряд внутренних и внешних заинтересованных сторон, необходимость подключения многоканальных входов, нескольких репозиториев данных и множества устройств.



Пример выше (см. рис.) иллюстрирует «в реальном времени» мониторинг активности процесса продаж (объем транзакций покупок в час). «Умные данные» позволят нам (от базы данных для принятия решений, в которой интегрированы исторические продажи), построить модель (жирная кривая на рисунке), которая прогнозирует торговую активность для эквивалентных дней (те же времена года, день недели и т. д.), тогда как «большие данные» будут фиксировать только текущую активность продаж (пунктирная кривая на рисунке). Очевидно, что объединение этих двух частей информации позволяет отслеживать деятельность (представленная модель сплошной кривой в качестве цели, в то время как пунктирная кривая представляет собой реальность). Сплошная кривая - это аналитическая модель, используемая для прогнозирования.

3. Стратегии использования «Smart Data»

Пользователи Smart Data.

Smart Data для бизнеса:

- Поддержка операционной деятельности (данные для операций);
- Поддержка принятия решений (данные для аналитики);
- Поддержка маркетинговых активностей (аналитика и персонализация);
- Объединение экосистем данных различных предприятий в единые экосистемы по отраслям.

Smart Data для нужд государства:

- Развитие транспортной инфраструктуры (телеметрия от транспортных средств);

- Управление малой энергетикой (данные о локальных генераторах, распределение нагрузки);
- Данные в здравоохранении (медицинские исследования, медобслуживание населения, мониторинг состояния здоровья граждан).

Smart Data для общественных объединений (communities).

Основные стратегии использования интеллектуальных данных:

- Быстрая первичная обработка данных (например, Event-driven architecture (EDA));
- Быстрая аналитика данных (подготовка первично обработанных данных для передачи далее для принятия решений или поддержки операционной деятельности);
- Быстрое принятие решений (архитектурные решения для быстрых решений)
- Поддержка операционной деятельности / исполнение принятых решений (в общем случае может быть не связано с данными).

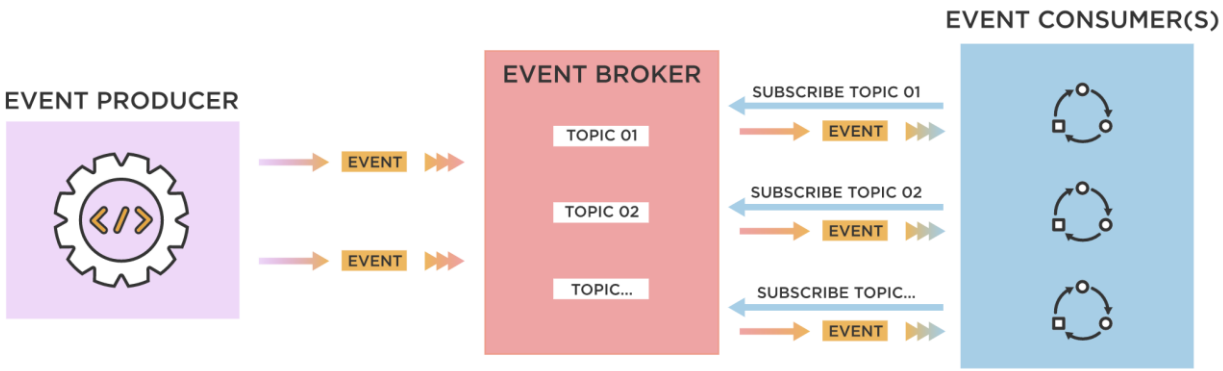
3. Первичная обработка данных

Методы первичной обработки данных.

- Маппинг данных (сопоставление входного потока данных с внутренней информационной структурой);
- Фильтрация данных (удаление данных из потока, которые не подлежат дальнейшему анализу и использованию);
- Удаление дубликатов ассетов данных;
- Простые методы валидации данных (относительно схем и правил), для некоторых случаев возможно дополнение пропущенных фрагментов и исправление ошибок (определяется применяемыми схемами данных).

Event-driven architecture (EDA)

Управляемая событиями архитектура (EDA) - это шаблон проектирования программного обеспечения, который позволяет организации обнаруживать «события» или важные бизнес-операции (такие как транзакция, посещение объекта, брошенная корзина online магазина и пр.), и действовать в соответствии с ними в режиме реального времени или почти в реальном времени. Этот шаблон заменяет традиционную архитектуру «запрос / ответ», в которой службы должны ждать ответа, прежде чем они смогут перейти к следующей задаче.



Важные особенности EDA:

- Управляемую событиями архитектуру часто называют «асинхронной» коммуникацией. Это означает, что отправителю и получателю не нужно ждать друг друга, чтобы перейти к следующей задаче. Системы не зависят от этого сообщения.
- При использовании управляемой событиями архитектуры существуют производители событий, которые генерируют и отправляют уведомления о событиях, и может быть один или несколько потребителей события, где получение события запускает логику обработки.
- Традиционно большинство систем работают в парадигме модели, ориентированной на данные, где данные являются источником истины. Переход к архитектуре, управляемой событиями, означает переход от модели, ориентированной на данные, к модели, ориентированной на события. В модели, управляемой событиями, данные по-прежнему важны, но события становятся наиболее важным компонентом.

Например, предположим, что Netflix только что загрузил новый фильм. Это может быть несколько приложений, которые прослушивают или ждут этого уведомления, которые затем запускают их собственные внутренние системы для публикации собственной информации об этом событии для своих пользователей. Это отличается от традиционного обмена сообщениями типа запрос-ответ тем, что приложения все еще работают, и даже если они могут прослушивать это событие, они не парализованы во время его ожидания. И они могут ответить, когда сообщение будет опубликовано. Таким образом, многие службы могут работать параллельно.

Как работает архитектура, управляемая событиями?

Компоненты событийно-управляемой архитектуры могут включать три части:

- Производитель событий,
- Потребитель событий,
- Брокер событий.

Брокер может быть необязательным, особенно когда есть всего один производитель и один потребитель, которые напрямую взаимодействуют друг с другом, а производитель просто отправляет события потребителю.

Примером может быть производитель, который отправляет события только в базу данных или хранилище данных, чтобы события собирались и сохранялись для анализа. Чаще всего на предприятиях присутствует несколько источников, отправляющих все типы событий, с одним или несколькими потребителями, заинтересованными в некоторых или во всех этих событиях.

4. Аналитика данных

Данные сами по себе бесполезны. Они полезны только в том случае, если можно извлечь смысл и ценность от них. Другими словами, важно то, что делается с данными, а не просто то, что они сгенерированы. Наука основана на понимании смысла и ценности данных. Масштабирование малых и больших данных создает новые проблемы. В случае масштабируемых небольших данных проблема заключается в объединение различных наборов данных для получения новых знаний и открытие данных для новых аналитических подходов, используемых в больших данных. Что касается больших данных, то задача состоит в том, чтобы справиться с их изобилием (включая значительные объемы данных с низкой полезностью и ценностью), своевременностью и динамикой, беспорядочностью и неопределенностью, высокой взаимосвязанностью, полуструктурированным или неструктурированным характером, и тем фактом, что большая часть больших данных генерируется без учета конкретной цели или является побочным продуктом другой деятельности. Действительно, до недавнего времени методы анализа данных в основном предназначались для извлечения информации из скудных, статичных, чистых и плохо структурированных наборов данных, отобранных с научной точки зрения и придерживаясь строгих допущений (таких как независимость, стационарность и нормальность), а также генерируемых и проанализированных с учетом конкретной цели.

Инструменты для объединения различных наборов данных и анализа больших данных были слабо развиты до недавнего времени, потому что они были слишком сложными в вычислительном отношении для реализации. Только в последние 40 лет регулярно анализировались очень большие наборы данных, и даже тогда они обрабатывались в рамках проектов, которые были обеспечены необходимыми ресурсами. Без мощных вычислений, становящихся более распространёнными и доступными, осмысление потока данных было бы очень дорогостоящим или трудным и отнимало бы много времени. Конечно, здесь имеет место своего рода замкнутый цикл: без развития повсеместных вычислений, такие объемы данных, не могли бы быть сгенерированы.

Тем не менее, как отмечают Хасти и др. (2009: xi), учитывая "появление компьютеров и век информации, статистические проблемы резко выросли как по размеру, так и по сложности".

Для решения проблем, связанных с обработкой и анализом масштабированных малых и больших данных, был разработан новый набор методов управления информацией и хранения, кроме этого разрабатывались методы анализа больших данных. Эти новые методы аналитики основаны на установленных статистических тестах, моделях и методах визуализации, но также происходит создание новых подходов, основанных на исследованиях в области искусственного интеллекта и экспертных систем, например машинного обучения, которое может вычислительно и автоматически выделять и обнаруживать шаблоны и строить прогнозные модели. Такая аналитика идеально подходит для обработки и извлечения информация из больших взаимосвязанных наборов данных и больших данных, и они стали важной областью инвестиций в исследования в целях быстрого расширения и создания новых методов обработки статистических данных, алгоритмов моделирования и методов визуализации. Аналитика данных, применяемая в бизнесе и науке, стремится ответить на четыре основных набора вопросов (Минелли и др. 2013):

- *Описание*: что и когда что-то произошло? Как часто это происходит?
- *Объяснение*: почему это произошло? Каково его влияние?
- *Прогноз*: что, скорее всего, произойдет дальше? Что, если бы мы сделали то или это?
- *Решение*: каков оптимальный ответ или результат? Как это достигается?

Ответы на эти вопросы получены из четырех основных классов аналитики: извлечение данных и распознавание паттернов; визуализация данных и визуальная аналитика; статистический анализ; и прогнозирование, моделирование и оптимизация. Каждый из них обсуждается вкратце далее, но сначала будет описан этап предварительной аналитики, поскольку он необходим для всех четырех видов аналитики.

Предварительная аналитика

Вся аналитика данных требует, чтобы данные, подлежащие анализу, были предварительно подготовлены и проверены.

Х. Дж. Миллер (2010) и Хан и др. (2011) изложили четыре таких процесса в отношении масштабированных и больших данные, которые обычно выполняются последовательно, хотя порядок обработки не принципиален:

> ***Выбор данных:*** *определение подмножества переменных, обладающих наибольшей полезностью, и критериев выборки для этих переменных.*
>
> ***Предварительная обработка данных:*** *очистка выбранных данных для устранения помех, ошибок или искажений или обработка отсутствующих полей или несоответствия, а также структурирование данных для дальнейшего анализа.*
>
> ***Сокращение и прогнозирование данных:*** *уменьшение размерности данных за счет преобразования (например, сглаживание, построение атрибутов, агрегирование, нормализация, использование иерархических концепций и статистические методы, такие как регрессионный анализ и анализ основных компонент) для создания эквивалентных, но более эффективных представлений.*
>
> ***Обогащение данных:*** *объединение выбранных данных с другими данными (например, данными переписи населения, рыночными данными и пр.), чтобы использовать более глубокие знания.*

Каждый из этих шагов направлен, с одной стороны, на повышение качества данных, используемых в анализе, и, с другой стороны, за исключением обогащения данных, для уменьшения объема данных. Выполнение этих задач может быть затруднено при проведении анализа в реальном времени, особенно с точки зрения попыток очистить данные. Следовательно, анализ больших объемов данных, выходящий за рамки визуализации показателей, не должен проводиться в режиме реального времени, а точнее проводят анализ на очень большой выборке очищенных, уменьшенных и обогащенных данных в формате временных рядов. Там, где требуется аналитика в реальном времени, предварительная аналитика обычно применяется к выборке заранее, чтобы установить характер данных и то, как они могут быть обработаны для выбора, уменьшения и очистки на лету.

Обогащение данных является важной задачей, поскольку оно приводит к эффекту усиления (Крэмптон и др. 2012), что позволяет получать аналитические данные, которые не могут быть получены из одного набора данных. Наряду с повторным использованием данных, обогащение является ключевым этапом для создания инфраструктур данных. В задача состоит в том, чтобы создать способы объединения данных, которые были сгенерированы для различных целей, и могут иметь различные метаданные, стандарты данных, единицы измерения, категории и масштабы, синхронность и форматов файлов, и для этого минимизировать создание потенциальных ошибок. Это не простая задача, но вычислительные методы облегчают ее за счет использования алгоритмов, которые могут искать, сопоставлять, объединять, переупаковывать (с помощью различных видов преобразований) и переформатировать данные. В результате появляются сначала новые,

затем комбинированные наборы данных, которые могут быть извлечены и проанализированы с использованием четырех рассматриваемых ниже широких классов аналитики.
Преданалитическая работа может быть чрезвычайно утомительной и трудоемкой, но, тем не менее, она жизненно важна и ее нельзя игнорировать. Учитывая взрывной рост объема данных в различных видах новых инфраструктур, предварительная аналитика станет плодородной областью исследований, поскольку исследователи данных стремятся к наиболее продуктивному, эффективному и результативному способы выполнения и особенно автоматизации такой работы.

Машинное обучение

Анализ очень большого количества записей данных может быть проведен своевременно только при помощи компьютерных алгоритмов. В то время как большая часть анализа больших данных может выполняться так же, как и анализ небольших данных, когда аналитики принимают решения о том, как анализ должен выполняться и с помощью каких алгоритмов, целью же многих других исследований является разработка автоматизированных процессов, которые могут оценивать и извлекать уроки из данных и их анализа. Такие автоматизированные процессы называются машинным обучением и они представляют собой искусственный интеллект. Машинное обучение направлено на итеративное развитие понимания набора данных; автоматическую учебу распознавать сложные закономерности и строить модели, которые объясняют и предсказывают закономерности и оптимизируют результаты (Han et al., 2011).
Машинное обучение обычно состоит из двух основных типов: контролируемое (с использованием обучающих данных) и неконтролируемое (с использованием самоорганизации). При обучении под наблюдением, модель обучается для соответствия входных данных некоторым известным результатам. Например, модель может быть обучена сопоставлению рукописных почтовых кодов с типизированными эквивалентами, или для прогнозирования определенного результата. Это "контролируется" в том смысле, что данные обучения присутствуют, чтобы направлять процесс обучения (Hastie et al., 2009). В отличие от этого, при обучении без контроля модель стремится научиться распознавать закономерности и находить структуру в данных без использования обучающих данных.
В целом это достигается за счет выявления кластеров и связей между данными, в которых характеристики сходства или ассоциаций не были известны заранее. Например, модель возможно, научится разделять клиентов на самоподобные группы и прогнозировать покупки среди этих группы (Хан и др., 2011). В обоих случаях модель создается с помощью процесса обучения, сформированного правила обучения и взвешивания, которые определяют, как строится модель по отношению к данным (Hastie et al.2009). Процесс построения модели начинается с простой конструкции, а затем ее многократно настраивают, используя правила обучения, как будто применяя "генетические мутации", пока это не превратится в надежную модель (Сигел 2013: 122). Существуют две разновидности контролируемого и неконтролируемого обучения-это полуконтролируемое обучение, которое включает использование как обучающих, так и немаркированных данных, а также активное обучение это позволяет пользователям играть активную роль в управлении моделью обучения (Han et al. 2011).
Машинное обучение используется для проведения всех четырех видов анализа больших данных, хотя оно и не является единственным методом, с помощью которого практикуется такая аналитика. Для машинного обучения аналитик остается важным для оценки и руководства процессом и оценки промежуточных результатов. Как отмечает Х. Дж. Миллер

(2010), машинное обучение - это не просто автоматизированная наука о кнопках, но, скорее, требует знаний в предметной области и вдумчивой рефлексии; навыков, с которыми человеческий разум все еще справляется лучше, чем компьютеры. Несмотря на значительный прогресс, достигнутый в разработке методов машинного обучения, это все еще развивающаяся наука и множество исследований необходимо предпринять для повышения эффективности и надежности созданных моделей.

Извлечение данных и распознавание паттернов

Извлечение данных - это процесс извлечения данных и шаблонов из больших наборов данных (Manyika et al. 2011). Это определение основывается на представлении о том, что все массивные наборы данных содержат значимую неслучайную информацию (Han et al. 2011). Зачастую используется контролируемые и неконтролируемое машинное обучение для обнаружения, классификации и сегментации значимых отношений, ассоциаций и тенденции между переменными. Это делается с использованием ряда различных техник, включая естественную обработку языка, нейронные сети, деревья решений и статистические (непараметрические и параметрические) методы. Выбор метода зависит от типа данных (структурированных, неструктурированных или полуструктурированных) и цели анализа (см. таблицу).

Table 6.1 Data mining tasks and techniques

Data mining task	Description	Techniques
Segmentation or clustering	Determine a set of implicit groups that describes the data	• Cluster analysis
Classification	Predict the class label that a set of data belongs to based on some training datasets	• Bayesian classification • Decision tree induction • Artificial neural networks • Support vector machine
Association	Find relationships among data objects; predict the value of some attribute based on the value of other attributes	• Association rules • Bayesian networks
Deviations	Find data items that exhibit unusual deviations from expectations	• Cluster analysis • Outlier detection • Evolution analysis
Trends	Lines and curves summarizing the database, often over time	• Regression • Sequence pattern extraction
Generalisations	Compact descriptions of the data	• Summary rules • Attribute-orientated induction

Source: Miller and Han (2009: 7).

Source: Miller and Han (2009: 7).

Большинство методов, перечисленных в таблице 6.1, относятся к структурированным данным, содержащимся в реляционных базах данных. Например, модели сегментации могут быть применены к базе данных розничных клиентов и их покупок, чтобы разделить их на различные профили на основе их характеристик и моделей поведения для того, чтобы предлагать каждой группе различные услуги/предложения. В анализе социальных сетей, связи между отдельными людьми анализируются, чтобы понять социальную динамику между членами и как может передаваться информация между ними. При обнаружении ассоциаций различные регрессионные модели могут использоваться для вычисления корреляций между переменными и, таким образом, выявлять скрытые закономерности, которые затем можно использовать в коммерческих целях (например, определить, на какие товары покупаются друг друга и реорганизации магазина для стимулирования покупок).

Неструктурированные данные в форме языка, изображений и звуков создают особые проблемы для интеллектуального анализа данных. Методы обработки естественного языка направлены на анализ человеческого языка, выраженного через письменное и устное слово. Они используют семантику и таксономию для распознавания шаблонов и извлечения информация из документов. Примеры включают извлечение сущностей, которые автоматически извлекаются метаданные из текста путем поиска определенных типов текста и фраз, таких как имена людей, местоположения, даты, специализированные термины и терминология продукта, а также извлечение сущностных отношений, которые автоматически определяет отношения между семантическими сущностями, связывая их вместе (например, имя человека с указанием даты или места рождения, или мнение о предмете) (McCreary 2009). Типичный применением таких методов является анализ настроений, который направлен на определение общего характера и сила мнений по какому-либо вопросу, например, что люди говорят о продукте в социальных сетях и СМИ. Используя метаданные по меткам, также можно отслеживать, где выражается такое мнение (Грэм и др., 2013) и для предотвращения распространения информации в социальных сетях, например, насколько широко веб-адреса являются избранными и разделяемыми между несколькими пользователями (Ohlhorst 2013). Такая информация полезна для компаний, таких как рекламные агентства, маркетинговые и финансовые службы, стремящиеся использовать появляющиеся тенденции и делать это своевременно (например, размещайте рекламу на подходящих историях; покупка/продажа акций в преддверии более широкой реакции рынка).

Изображения структурированы для хранения и отображения, но не для контента и поиска (Ohlhorst 2013). Обнаружение, классификация и извлечение из них паттернов, таких как распознавание лиц или мест, не просто, но решается с помощью фотограмметрии, дистанционного зондирования, обработки изображений и методов машинного зрения, включая распознавание объектов и сопоставление шаблонов с использованием обучающих наборов, методов кластеризации и нейронные сети. Проблема интеллектуального анализа изображений усугубляется, когда пытаясь извлечь, сравнить и проиндексировать шаблоны из огромного количества изображений (Zhang et al. 2001). Сегодня анализ изображений является все еще развивающейся областью, и майнинг изображений в последние годы стал более продвинутым. Для например, ImageVision (http://imagevision.com/) утверждает, что может классифицировать 50 400 минут видео в час, на сервере, с использованием алгоритмов машинного обучения для обнаружения определенных функций, таких как нагота или логотипы компаний.

Визуализация данных и визуальная аналитика

Обычно говорят, что картинка говорит тысячу слов. Как таковой, визуальный регистр уже давно используется для обобщения и описания наборов данных с помощью статистических диаграмм и графиков, диаграмм пространственной привязки, карт и анимации. Эти визуальные методы эффективно выявляют и передают структуру и тенденции переменных и их взаимосвязи. Учитывая огромные объемы и скорость больших данных, неудивительно, что визуализация оказалась популярным способом создания чувства информированности и передачи этого чувства.

Визуализации, созданные в цифровой области, могут использоваться для навигации и запроса данных, позволяя пользователям получить обзор всего набора данных, увеличить масштаб интересующих элементов, отфильтровывать неинтересные данные, выбирать элемент или группу данных и получать подробные сведения, просматривать связи между элементами и извлекать вложенные коллекции деталей, когда это необходимо (Шнейдерман 1996). При этом они позволяют видеть характеристики и структуру наборов данных, которые необходимо изучить. Более того, ее можно использовать, чтобы сделать видимыми и понятные сложные наборы данных и модели, которые трудно концептуализировать абстрактно (например, атомные, космические и трехмерные явления), и построить десятки тысяч точек данных чтобы выявить структуру, кластеры, пробелы и выбросы, которые в противном случае могли бы остаться скрытыми (Шнейдерман 1996). Например, понять смысл миллионов твитов в Twitter нелегко. Можно получить приблизительное представление об историях или темах, которые, по-видимому, находятся в тренде, но очень трудно получить краткий обзор того, как меняется тенденция отношений между людьми и местами. Решением является привязка с пространственной привязкой твитов, отфильтрованных по настроению и подходу к выбранной проблематике. Twitter создал десятки карт по конкретным темам и тенденциям в твиттере, включая карту на рисунке 6.1, которая показывает географическое распределение гомофобных твитов в Соединенных Штатах с июня 2012 года по апрель 2013. Они также сопоставили контент Википедии и метки Google. Визуализации также обычно используются в качестве средства для мониторинга динамики в реальном времени явлений, позволяющее отслеживать несколько переменных во времени и пространстве и по сравнению с одной еще и для выявления изменений. Панели мониторинга визуализированных динамических данных часто отображаются на компьютерные мониторы в современных диспетчерских, графически отображающих систему в движении для человека - оператора с графиками и диаграммами временных рядов или картами разворачивающихся событий (см. Озеро 2013, для сравнение 24 панелей мониторинга).

Например, данные по всей транспортной системе могут обеспечить карты транспортных потоков и отчеты об авариях в реальном времени; или местоположение рейсов, когда они проходят над территорией (см. рис. 6.2), или данные метеорологического радара могут предоставить карту осадков в реальном времени и анимацию последние несколько часов. Такие визуальные данные помогают не только диспетчерам управления движением или синоптикам, но и гражданам, которые могут получить доступ и отслеживать разворачивающиеся ситуации с помощью компьютера или смартфона, изменяя их поведение, чтобы избегать определенных маршрутов или одеваться соответствующим образом. Пример прототипа информационной панели public city, которая объединяет ряд данных о погоде в режиме реального времени,

загрязнение воздуха, задержки общественного транспорта, доступность общественного велосипеда, уровень реки, спрос на электроэнергию,

фондовый рынок, тенденции в Twitter и ленты дорожных камер показаны на рис. 6.3.

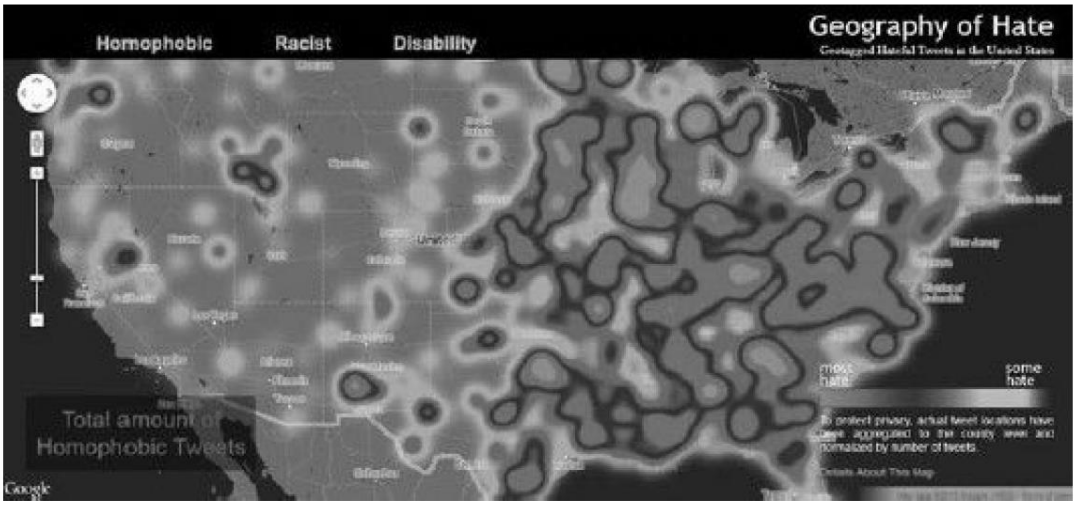


Figure 6.1 The geography of homophobic tweets in the United States

Source: http://users.humboldt.edu/mstephens/hate/hate_map.html#

Figure 6.2 Real-time flight locations

Source: http://www.flightradar24.com/

Figure 6.3 CASA's London City Dashboard

Source: http://citydashboard.org/london/

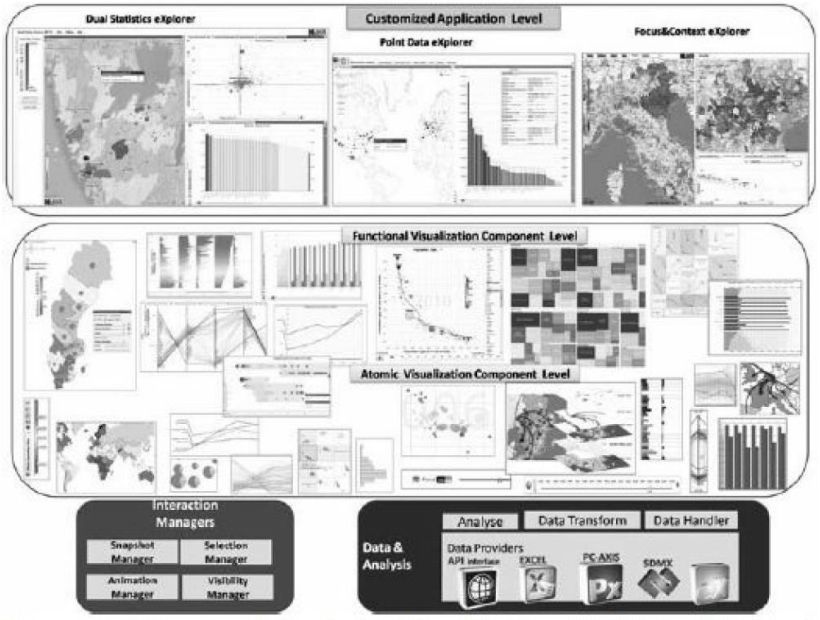


Figure 6.4 Geovisual Analytics Visualization (GAV) toolkit developed by the National Center for Visual Analytics, Linköping University

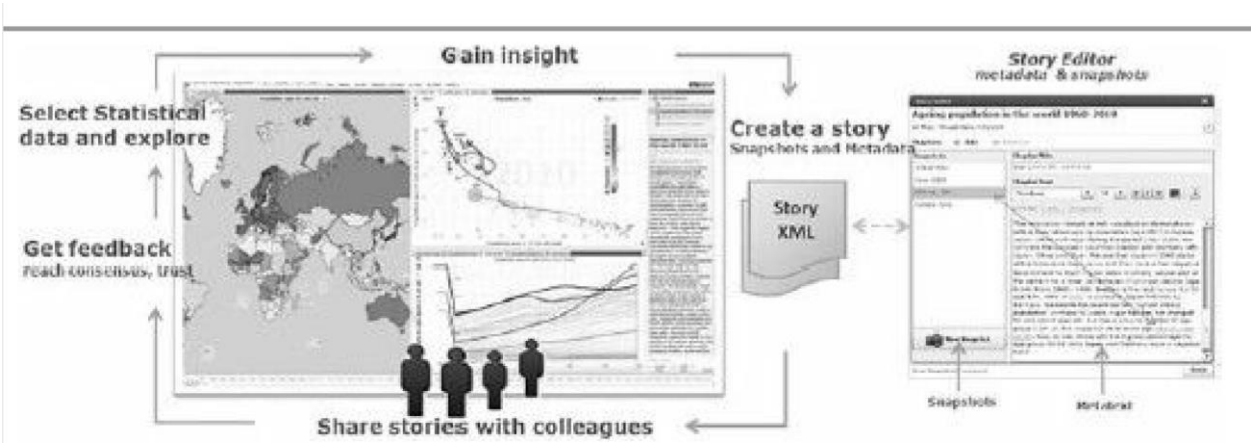


Figure 6.5 Using GAV for collaborative storytelling

Source: http://ncva.itn.liu.se/tools?l=en

Визуализацию также можно использовать в качестве формы аналитического рассуждения. В этом случае визуализация-это не просто описание или отображение данных, а также используется в качестве аналитического инструмента. Так называемый визуальный анализ, это подход часто руководствуется комбинацией людей и алгоритмов, которые согласованно работают для извлечения информации, строят визуальные модели и объяснения, а также направляют дальнейший статистический анализ (Кейм и др.2010). Визуальная аналитика стала значительной новой областью исследований, расширяющей область визуализации информации путем интеграции элементов интеллектуального анализа данных, статистики и когнитивной науки (Томас и Кук, 2006). Обычно используется несколько различных типов визуальной графики, и это в целом интерактивно и позволяет пользователю играть с изображениями и манипулировать ими для изучения и выявления закономерностей и связей. Кроме того, панели, отображаемые пользователю, как правило, взаимосвязаны таким образом, чтобы взаимодействия на одной панели воспроизводились на других, позволяя различным аспектам данных быть рассматренным с более чем одной точки зрения одновременно. Например, на рисунке 6.4 показано виды визуальной аналитики, доступные в Geovisual инструментарии визуализации Geovisual Analytics (GAV), разработанном Национальным центром визуальной аналитики при Университете Линчепинга в Швеции. Предоставляя набор аналитических инструментов, GAV обеспечивает совместное изучение и анализ данных, создание приложений, смешивание с картографическими приложениями, такими как Google Maps, и возможностью создать и поделиться примечанием. Последнее создает социальный элемент в инструментарии, позволяя коллегам и другим пользователям поделиться толкованиями смысла, касающимися визуализации (см. рис. 6.5).

Статистический анализ

Существует долгая история применения статистических методов к количественным данным для того, чтобы наделить их смыслом. *Обзор статистических методов, используемых для анализа больших данных:*

- ***Описательная статистика*** *(подробно описывает характеристики и распределение данных, а также уровни ошибок и неопределенности). Включает в себя:*

        - *Анализ временных рядов (параметры, изменяющиеся во времени);*
        - *Теория графов (математическое описание организационных и сетевых структур);*
        - *Пространственная статистика (описывает геометрию и паттерны пространственной кластеризации, дисперсии и диффузии).*
    - ***Статистика вывода (Inferential statistics)*** *(стремится объяснить, а не просто описать закономерности и взаимосвязи, которые могут существовать в рамках набор данных, а также для проверки силы и значимости ассоциаций между переменными). Включает в себя:*
        - *Параметрическая статистика:*
        - *Непараметрическая статистика;*
        - *Вероятностная статистика.*

Арсенал описательной и логической статистики, который традиционно используется для анализа небольших данных, также применяются к большим данным, хотя это не всегда просто, потому что многие из этих методов были разработаны для получения информации из относительно скудных, а неполных данных. Тем не менее, эти методы подходят для осмысления огромных объемов данных. Статистические данные регулярно используются в помощь в выделении данных, прогнозировании и оптимизации данных.

Прогнозирование, моделирование и оптимизация

Ключевым способом извлечения ценности из данных является их использование для прогнозирования того, что произойдет в других условиях. Например, компания может захотеть предсказать, как отреагируют клиенты к конкретному продукту или кампании, или местному правительству необходимо спрогнозировать, как транспортная инфраструктура может функционировать, если ее критический элемент будет закрыт или ученый пытается предсказать, когда может произойти оползень и при каких условиях. Такая информация очень полезна для организаций, планирующих действия в различных непредвиденных ситуациях, и предприятий с точки зрения обеспечения роста эффективности и увеличения прибыли. Во всех случаях модели строятся с использованием существующих знаний о том, как работает система, какие данные обрабатываются для оценки потенциального результата при различных сценариях.

Как и в случае с выделением данных, существует множество различных методов, которые можно использовать для получения прогностических моделей. У каждого из них есть свои сильные и слабые стороны, они дают меньше ошибок и более точные прогнозы в зависимости от типа проблемы и данных (Seni и Elder 2010). Тем не менее, часто трудно предугадать, какой тип модели и ее различные версии будут работать наилучшим образом для конкретного набора данных. Для решения этой проблемы был предложен ансамблевый подход, который использует преимущества огромного объема вычислительных мощностей, доступных аналитикам в настоящее время (Siegel 2013). Вместо того, чтобы выбирать единый подход и построение нескольких моделей, комплексный подход создает несколько моделей используя различные методы для прогнозирования одних и тех же явлений. Затем, вместо выбора результатов (оценки) по наиболее эффективной модели оценки, все модели объединяются для получения одного общего ответа. Агрегирование результатов приводит к более надежному результату по мере того, как процесс компенсирует недостатки в каждой модели. Например,

ансамблевый подход к прогнозированию поведения клиента может построить ряд моделей: регрессионных, нейронных сетей, моделей ближайшего соседа и модели дерева решений. Каждая модель может лучше предсказывать определенные типы потребителей, чем другие, но при объединении результатов моделей эти отклонения сглаживаются, что дает более надежный прогноз (Фрэнкс 2012; Сигел 2013). Используя ансамблевый подход, сотни различных алгоритмов могут быть применены к набору данных, что гарантирует создание наилучшей прогнозной модели.

Моделирование - это построение моделей, которые стремятся моделировать реальные процессы и системы. Цель состоит в том, чтобы определить как функционирует система и как она может вести себя в различных сценариях, а также статистическая оценка ее эффективности с целью повышения результативности работы системы (Робинсон 2003). Популярным примером является компьютерная игра SimCity, которая имитирует, как будет расти и развиваться город в условиях выбора игроков, основываясь на базовой модели известного городского процесса. Аналогичным образом, прогнозы погоды основаны на моделировании того, как будет развиваться погода, учитывая преобладающие условия и научные знания. Существует много различных видов моделирования, многие из которых используют машинное обучение для автоматического уточнения модели и решения с возникающими свойствами, такими как непредвиденные события. SimCity-это модель, основанная на агентах (Batty 2007).
Модель состоит из среды, в которой отдельные объекты, такие как здания и дороги, которым присвоены определенные характеристики. Затем эта среда заполняется агентами, которым приписываются особые качества. При запуске модели агенты стремятся решить задачу, реагируя на окружающую среду и других агентов в зависимости от их характеристик. В свою очередь, как агенты, выполняя свои задачи, они меняют окружающую среду, в данном случае город, в котором они живут, производя сложную, развивающуюся систему. Таким образом, модель работает снизу вверх, с более широкими пространственными и временными закономерностями, возникающими в результате взаимодействия отдельных агентов с окружающей средой. Такие городские имитационные модели используются вне игр для моделирования землепользования и транспортного планирования и разработки планов действий на случай чрезвычайных ситуаций (Batty 2007).

Оптимизация связана с определением оптимального курса действий для повышения производительности (как правило, снижение затрат или увеличение объема производства/оборота). Ее можно рассчитать, используя и оценивая прогностические и имитационные модели или могут быть разработаны модели с помощью других видов алгоритмов или статистического тестирования. Например, генетические алгоритмы, особый вид машинного обучения, использующий идеи из естественного отбора, такие как наследование, мутация, отбор и скрещивание, для развития и эволюции возможных решений проблемы (Митчелл, 1996). Еще один биологически вдохновленный подход-нейронный сеть, которая стремится имитировать работу человеческого мозга, используя тесно взаимосвязанные элементы обработки данных, позволяющие рассчитать, оценить и решить проблему (Picton 2000). Тестирование может использоваться на скользящей основе для оценки и настройки системы, сравнения контрольной группы с различными тестовыми группами, чтобы определить, какие методы обработки (например, текст, макеты, изображения или цвета, используемые на веб-сайте) приближают к заданной цели.