Big Data Analytics

Введение

Зрелов П.В.

Лаборатория информационных технологий ОИЯИ
Лаборатория облачных технологий и аналитики больших данных
РЭУ им. Плеханова

GRID and Advanced Information Systems. 2-6 ноября 2015.
Дубна

Big Data Analytics

Введение

Содержание лекции

- История появления термина Big Data
- Что же такое Big Data?
- Источники Big Data
- Объемы Big Data
- Примеры использования Big Data
- Понятие Big Data
- Датификация
- Литература


BIG
DATA
ALSO
TOOLS
DISK
TARGET
APPLIED
SENSOR
MAY
EXAMPLES
EVERY
SHARED
DEFINITION
MOVING
MANAGE
WITHIN
ARCHIVES
PRACTITIONERS
CAPTURE
TECHNOLOGIES
INTERNET
BUSINESS
SETS
GENOMICS
NEEDED
TIME
ONE
LARGE
MANAGEMENT
TERABYTES
INFORMATION
PETABYTES
INCLUDE
SOFTWARE
ABILITY
PERFORMANCE
SYSTEMS
TOLERABLE
STORE
SIZE
RESEARCH
STORAGE
LOGS
PARALLEL
SAN
MASSIVELY
GROW
SEARCH
DIFFICULTY
GARTNER
WORKING
RECORDS
NETWORKS
DATABASES
HUNDREDS
CAPACITY
BIOLOGICAL
PROCESSING
ANALYTICS
TENS
COMPLEX
CITATION
LARGER
CONTINUES
SOCIAL
RELATIONAL
SET
USE
ELAPSED
COST
PROCESS
USING
DISTRIBUTED
CASE

История появления термина Big Data

Считается, что первые упоминания термина относятся к **2005** году в изданиях компании **O'Reilly media** в связи с необходимостью хоть как-то определить те данные, с которыми традиционные технологии управления и обработки данных не справлялись в силу их сложности и большого объема.

В **2008** году термин Big Data использовался в специальном номере журнала **Nature**, посвященном теме «Как могут повлиять на будущее науки технологии, открывающие возможности работы с большими объемами данных?». В номере были собраны материалы о феномене взрывного роста объемов и многообразия обрабатываемых данных. Там же обсуждались технологические перспективы в парадигме вероятного скачка «от количества к качеству»

В **2009** году термин широко распространился в деловой прессе, а к **2010** году относят появление первых продуктов и решений. К 2011 году большинство крупнейших поставщиков информационных технологий используют понятие Больших данных, в том числе IBM, Oracle, Microsoft, Hewlett-Packard, EMC.

В **2011** году компания **Gartner** дала прогноз, что внедрение технологий Больших данных окажет влияние на подходы в области информационных технологий в производстве, здравоохранении, торговле и государственном управлении.

Что же такое Big Data?

Big Data – это группа технологий и методов производительной обработки очень больших объемов данных, в том числе неструктурированных, в распределенных информационных системах, обеспечивающих организацию качественно новой полезной информации.

Технологии **Big Data** предоставляют услуги, позволяющие раскрыть потенциал мегамассивов данных за счет выявления скрытых закономерностей и фактов.

Под «очень большими» наборами данных подразумеваются данные объемом от терабайт до сотен петабайт. Например, фото и видео хранилище на Facebook оценивается как минимум в 100 петабайт.

Полезно напомнить, что
1 PB = 10^15 bytes (пета-), 1 EB = 10^18 bytes (экса-), 1 ZB = 10^21 bytes (зета-)

Что же такое Big Data?

Big Data – это наборы данных такого объема, когда традиционные инструменты не способны осуществлять их захват, управление и обработку за приемлемое для практики время.

Big Data – это группа технологий и методов производительной обработки очень больших объемов данных, в том числе неструктурированных, в распределенных информационных системах, обеспечивающих организацию качественно новой полезной информацией.

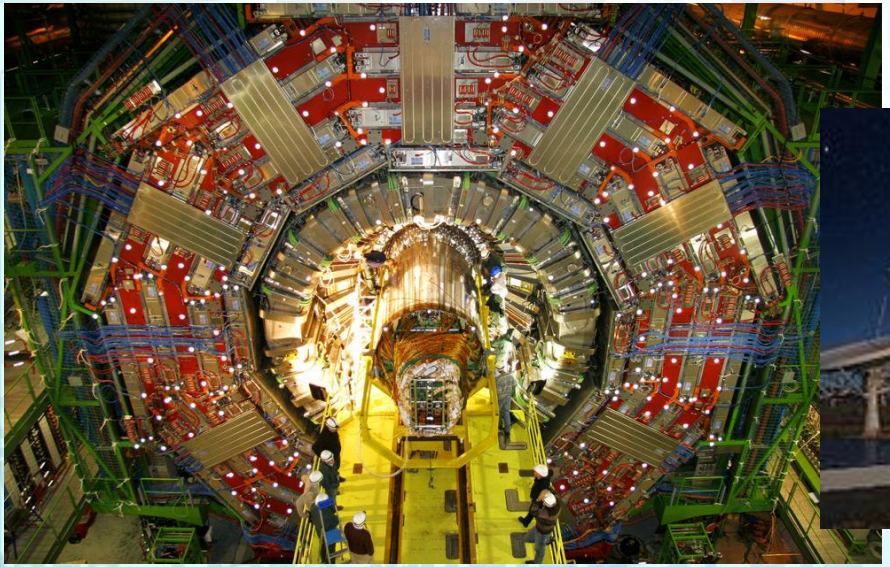
Технологии **Big Data** предоставляют услуги, позволяющие раскрыть потенциал мега массивов данных за счет выявления скрытых закономерностей и фактов.

Под «очень большими» наборами данных подразумеваются данные объемом от терабайт до сотен петабайт. Например, хранилище фото и видео на Facebook оценивается как минимум в 100 петабайт.

Полезно напомнить, что
1 PB = 10^15 bytes (пета-), 1 EB = 10^18 bytes (экса-), 1 ZB = 10^21 bytes (зета-)

Источники Больших данных

- Торговля
- Промышленность
- Экономика
- Наука



Объемы Больших данных

Walmart

Каждый час собирает данные о сделках с клиентами > 2,5 PB

Google™

processes 20 PB a day (2008)
crawls 20B web pages a day (2012)

JPMorganChase

150 PB on 50k+ servers
running 15k apps (6/2011)

ebay

>10 PB data, 75B DB
calls per day (6/2012)

INTERNET ARCHIVE

Wayback Machine: 240B
web pages archived, 5 PB
(1/2013)

>100 PB of user data +
500 TB/day (8/2012)

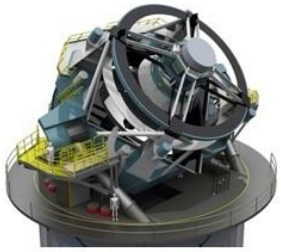
facebook

LHC: ~15 PB a year

CERN

amazon web services™

S3: 449B objects, peak 290k
request/second (7/2011)
1T objects (6/2012)



LSST: 6-10 PB a year
(~2015)

Large Synoptic Survey Telescope

Square Kilometre Array

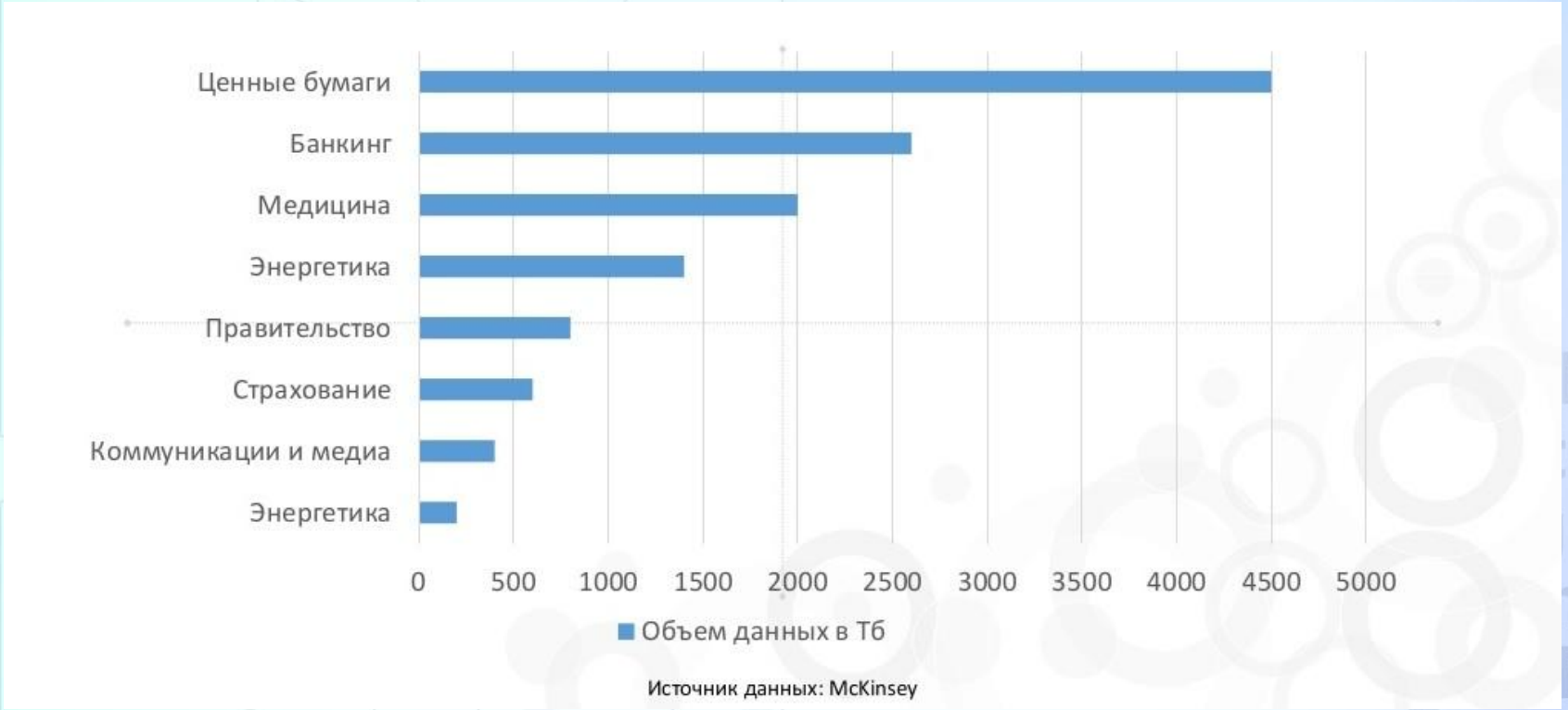
radio telescope

SKA: 0.3 – 1.5 EB
per year (~2020)

at&t

AT&T передает 30Pb в день

Объем данных корпораций по отраслям



Представленные данные относятся к 2012 году и конечно быстро меняются. Диаграмма интересна соотношениями между отраслями

Понятие Big Data

Определений больших данных очень много. Одно из самых распространенных:
Большие данные – это данные, которые описываются с помощью **четырех Vs**:

Volume (объем),
Velocity (скорость),
Variety (разнообразие)
Veracity (достоверность)

Объем.

Реально большие объемы данных в физическом смысле. Тот объем данных, который раньше накапливался годами, теперь генерируется каждую минуту.

Новые инструменты больших данных используют распределенные системы, так что данные можно хранить и анализировать в нескольких географически распределенных базах данных.

Скорость.

Сообщения в социальных сетях расходится по всему интернету в считанные секунды. Современные технологии позволяют анализировать данные на лету, даже не размещая их в базах данных.

Понятие Big Data

Разнообразие.

В недавнем прошлом рассматривались только **структурированные данные**, аккуратно встроенные в таблицы реляционных баз данных, например, финансовые данные. Но, фактически, **80%** мирового объема данных являются **неструктурированными** (текст, изображения, видео, голос и др.)

С технологиями больших данных теперь есть возможность проанализировать и свести воедино **данные разных типов**, такие как сообщения, разговоры в социальных сетях, фотографии, данные с датчиков, видео или голосовые записи.

Достоверность.

Для значительного множества данных их **качество и точность** являются **слабо** контролируемыми (сообщения в Твиттере, сокращения и ошибки в разговорной речи, ненадежность и неточности контента). **Новая технология позволяет** теперь работать и с этим типом данных.

Понятие Big Data

Новые технологии, такие как облачные вычисления и распределенные системы, вместе с последними разработками программного обеспечения и современными методами анализа данных позволяют использовать все виды данных одновременно, чтобы получать дополнительные знания.

Современные **технологии** делают возможным обработку и анализ огромного количества данных, в некоторых случаях – **всех данных**, касающиеся того или иного явления (не полагаясь на случайные выборки) в их первозданном виде – **структурированные, неструктурированные, потоковые**.

Big Data Analytics

Применения (по отраслям)

Отрасли экономики	Применение (анализ)
• Финансы	кредитные карты
• Страхование	запросы, выявление мошенничества
• Телекоммуникации	записи звонков
• Транспорт	управление логистикой
• Потребительские товары	продвижение товаров
• Научные исследования	изображения, видео, речь
• Коммунальные услуги	энергопотребление

Big Data

Применения больших данных? Пример 1.

Лучше понять и нацелить клиентов:

Чтобы лучше понять и нацелить клиентов, компании дополняют свои БД данными из социальных сетей, браузеров, данными датчиков и т.д., чтобы получить более полную картину о своих клиентах. Главной целью является создание прогнозных моделей. С помощью больших данных телекоммуникационные компании теперь могут лучше прогнозировать отток клиентов; розничные торговцы могут предсказывать, какие продукты будут продавать, а автомобильные страховые компании понять, насколько хорошо их клиенты на самом деле управляют автомобилем.

Big Data

Применения больших данных? Пример 2.

Понимать и оптимизировать бизнес-процессы:

Большие данные все шире используются для оптимизации бизнес-процессов. Ритейлеры имеют возможность оптимизировать свои запасы на основании моделей прогноза, сгенерированных из данных социальных сетей, тенденций интернет запросов и прогнозов погоды. Другим примером является оптимизация дорожного движения с использованием данных GPS адиочастотных датчиков

Big Data

Применения больших данных? Пример 3.

Здравоохранение

Вычислительные мощности, созданные для анализа больших данных, позволяют находить новые подходы и методы лечения, лучше понимать и предсказать болезни. Теперь стало возможным на основании данных от смарт-часов, других носимых устройств лучше понять связь между образом жизни и различными заболеваниями.

Аналитика больших данных позволяют следить и прогнозировать эпидемии и вспышки заболеваний, просто послушав, что люди говорят, например, “плохо себя чувствую – в постели с простудой” или ищут в Интернете, например, “лекарства от гриппа”.

Big Data

Применения больших данных? Пример 4.

Повышение безопасности и укрепление законопорядка:

Службы безопасности используют анализ больших данных для срыва террористических заговоров и выявления кибератак.

Спецслужбы используют инструменты больших данных, чтобы поймать преступников и даже предугадывать преступные намерения.

Банки используют аналитику больших данных для выявления мошенничества с помощью анализа операций по картам.

Big Data

Применения больших данных? Пример 5.

Улучшение спортивных результатов:

Наиболее элитарные виды спорта в настоящее время уже используют анализ больших данных. Многие используют видео-аналитику для отслеживания эффективности игроков в футболе или бейсболе, сенсорная технология встроена в спортивный инвентарь: баскетбольные мячи или клюшки для гольфа, и многие элитные спортивные команды контролируют своих участников вне спортивной среды – с использованием смарт-технологий для отслеживания питания и сна, также как и разговоры в социальных сетях для мониторинга эмоционального состояния.

Big Data

Применения больших данных? Пример 6.

«Совершенствование и оптимизация» городов и стран:

"Большие данные" используется для улучшения многих аспектов жизни наших городов и стран. Например, это позволяет городу оптимизировать транспортные потоки на основе информации о дорожном движении, получаемой в реальном времени, данных из социальных сетей и данных о погодных условиях . В настоящее время целый ряд городов используют анализ больших данных, чтобы превратить себя в «умные города», где транспортная инфраструктура и коммунальные процессы объединены. Где автобус будет ждать поезда в случае его опоздания, и где светофоры предвидят транспортные потоки и работают в режиме, который минимизирует пробки.

Big Data
Источники

Оперативные данные

Даже такие простые занятия, как прослушивание музыки или чтение книги сейчас генерируют данные. Цифровые музыкальные плееры и электронные книги собирают информацию о нашей деятельности. Ваш смартфон собирает данные о том, как вы его используете и ваш веб-браузер собирает информацию о том, что вы ищете. Компания, выпустившая вашу кредитную карту, собирает данные о том, где находится ваш магазин, а ваш магазин собирает данные о том, что вы покупаете. Трудно представить деятельность, которая не генерирует данные.

Big Data

Источники

Данные разговоров

Наши разговоры теперь записываются в цифровом формате. Все началось с электронных писем, но в наше время большинство наших разговоров оставляют цифровой след. Даже многие наши телефонные разговоры теперь записаны в цифровом формате.

Фото и видео данные

Только подумайте обо всех фотоснимках, которые мы сделали на наши смартфоны или цифровые камеры. Мы загружаем сотни тысяч фотографий в секунду на сайты социальных сетей. Увеличивающееся количество камер видеонаблюдения снимает видео изображения, и мы каждую минуту загружаем на YouTube и другие сайты сотни часов видео-данных.

Big Data

Источники

Данные датчиков

Мы все чаще попадаем в окружение датчиков, которые собирают и передают данные. Ваш смартфон содержит датчик глобального позиционирования, чтобы отслеживать, где именно вы находитесь в данную секунду, у него есть акселерометр, чтобы отслеживать скорость и направление ваших путешествий. Теперь сенсоры есть во многих устройствах и товарах.

Big Data

Источники

Интернет вещей

Сейчас у нас есть смарт-телевизоры, которые способны собирать и обрабатывать данные, у нас есть умные часы, умные холодильники, умные будильники. Интернет вещей, или всеобъемлющий Интернет, соединяет эти устройства так, чтобы, например, датчики загруженности дорог могут отправить данные на ваш будильник. Будильник разбудит вас раньше, чем вы планировали, по причине перекрытия дороги для того, чтобы вы могли уйти раньше и успеть на ваше утреннее заседание в 9.

Датификация

Большие данные способны **обращать в «цифру»** то, что никогда раньше не оценивалось количественно: для это введен в оборот термин **датификация**. **Датификация** обеспечивает беспрецедентный поток данных в плане объема, скорости, разнообразия и достоверности.

Примеры:

1) Местоположение объекта на поверхности Земли стало возможным датифицировать с изобретением спутниковых систем глобальной навигации (GPS, ГЛОНАСС).

2) Слова превращаются **в цифры**, когда «компьютеры раскапывают в старинных книгах наслоения эпох».

3) Дружеские отношения и **симпатии** датифицируются в социальных сетях через «**лайки**».

Особенности подхода Big Data в науке

- Подход Big Data обязан своим рождением экономике и бизнесу. Там он, прежде всего, и используется. Причина популярности - потенциал для развития бизнеса.

- Применение в науке имеет много общего с применением к бизнес-данным. Однако есть отличие, заключающееся в том, что существует большое количество накопленных знаний (в отличие от данных) и научных теорий. Таким образом, существует гораздо меньше шансов найти новые знания прямо из данных.

  Однако, эмпирические результаты могут быть ценны в новых областях знаний, в прикладных областях, граничащих с техникой, или при моделировании сложных явлений.

Особенности подхода Big Data в науке

- Еще одно отличие заключается в том, что в торговле, бизнесе в целом, правила «мягкие», социологические, культурные, отражающие определенные традиции поведения (в частности, покупателя).

- Например, правдоподобный миф о том, что «30% людей, покупающих подгузники для младенцев, одновременно покупают пиво», вряд ли отражает какие-нибудь фундаментальные положения и имеет характер закона природы, но его можно с пользой применить в практике продаж.

- С другой стороны, научные правила или законы, в принципе, проверяемы объективно.

Любые результаты применения методов Big Data должны находиться в пределах существующих знаний конкретной предметной области.

Привлечение эксперта предметной области имеет решающее значение для процесса интеллектуального анализа данных.

Big Data в научных областях

1 Astrophysics - астрофизика
2 Biology - биология
3 Nanoscience - нанотехнологии
4 Power and Communication Networks – электрические и коммуникационные сети
5 Climate Systems Modeling – моделирование климата
6 Fusion Physics – термоядерный синтез
7 Accelerator Physics – физика на ускорителях
8 Cybersecurity - кибербезопасность
9 Combustion – процессы горения

Mathematics for Analysis of Petascale Data. Report on a Department of Energy Workshop. June 3–5, 2008

Big Data, Big Data Analytics and Data Mining

В настоящий момент нет различия в употреблении терминов Big Data и Big Data Analytics. Эти термины описывают как сами данные, так и технологии управления и методы анализа.

Big Data Analytics является развитием концепции Data Mining. Одни и те же задачи, сферы применения, источники данных, методы и технологии.

За годы, прошедшие с момента появления концепции Data Mining до наступления эры Больших данных, революционным образом изменились объемы анализируемых данных, появились системы высокопроизводительных вычислений, новые технологии, в том числе MapReduce и ее многочисленные программные реализации. С появлением социальных сетей появились и новые задачи.

Data Mining

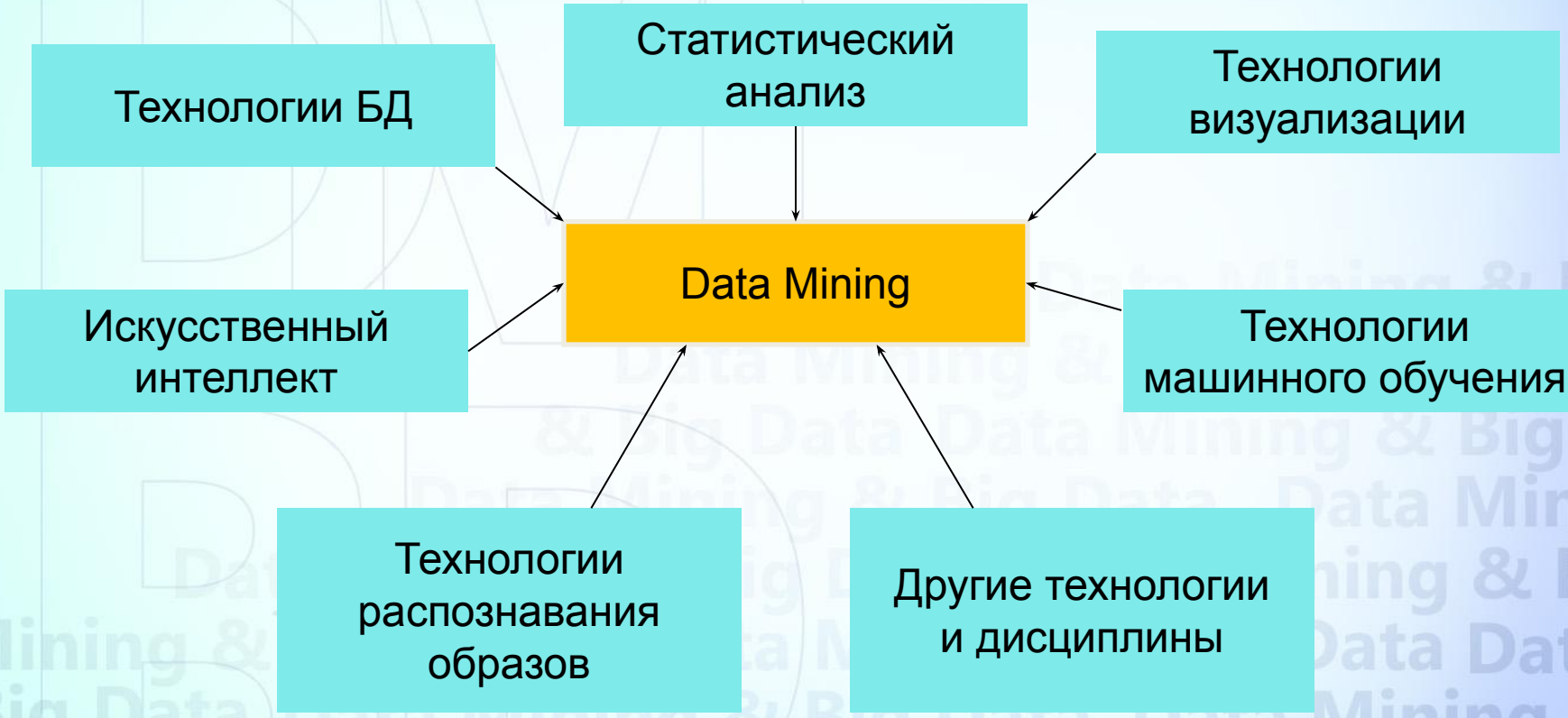
Data Mining - это процесс **поддержки принятия решений**, основанный на поиске в сырых данных скрытых закономерностей, ранее неизвестных, нетривиальных, практически полезных и доступных интерпретации знаний, необходимых для принятия решений в различных сферах человеческой деятельности.

Data Mining – это особый подход к анализу данных. Акцент делается не только на извлечении фактов, но и на **генерацию гипотез**. Созданные в процессе гипотезы следует проверять с помощью обычного анализа в рамках привычных схем и/или с привлечением экспертов предметной области.

В данном подходе используются традиционные инструменты анализа, такие как математическая статистика (регрессионный, корреляционный, кластерный, факторный анализ, анализ временных рядов, деревья решений и др.), а также те, что связаны с искусственным интеллектом (машинное обучение, нейронные сети, генетические алгоритмы, нечеткие логики и др.).

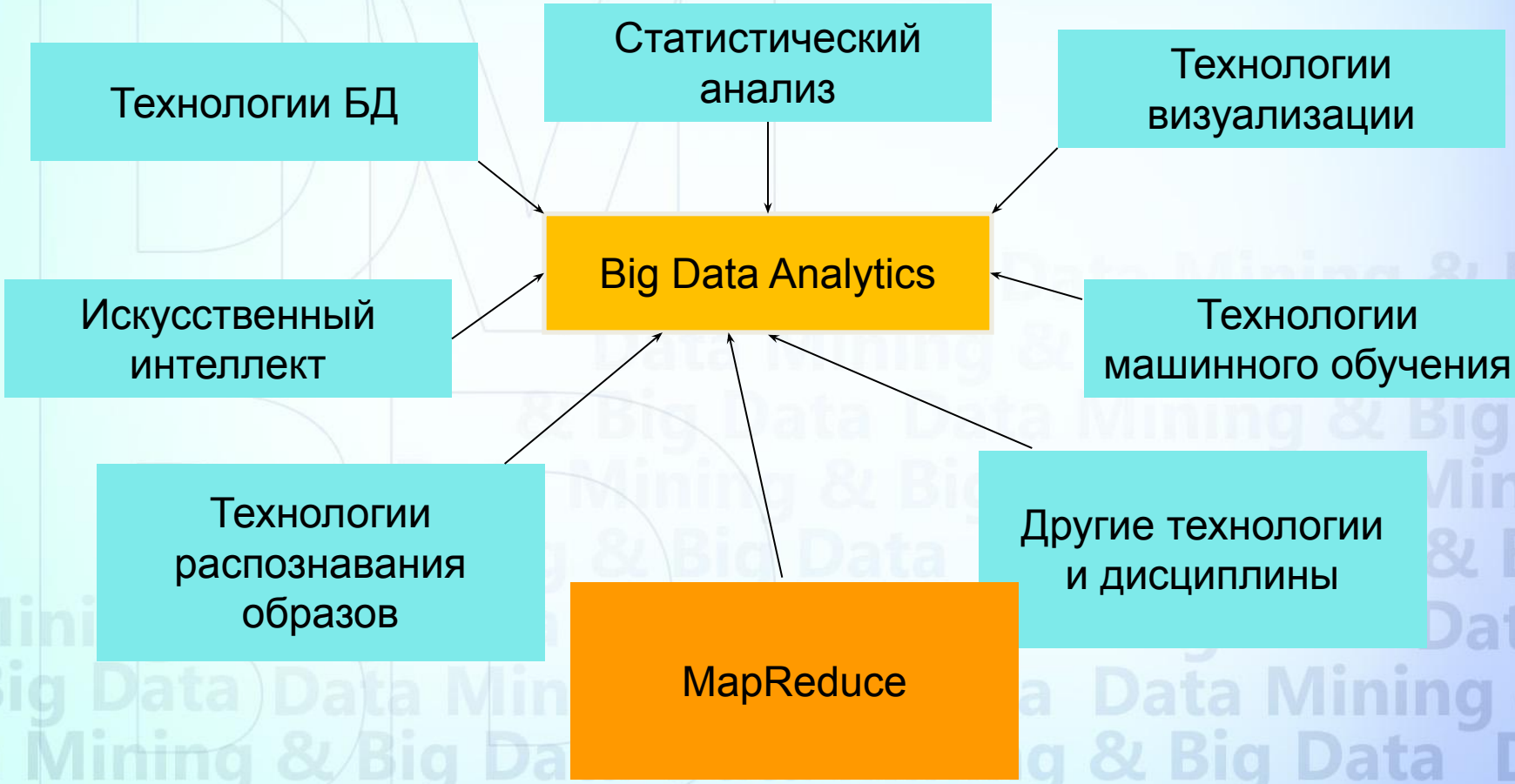
Data Mining

Data Mining – это «сплав» нескольких дисциплин и технологий



Big Data Analytics

Если схему дополнить технологией MapReduce и требованием 4V, она отразит функциональные связи Big Data Analytics



MapReduce
Simplied Data Processing on Large Clusters

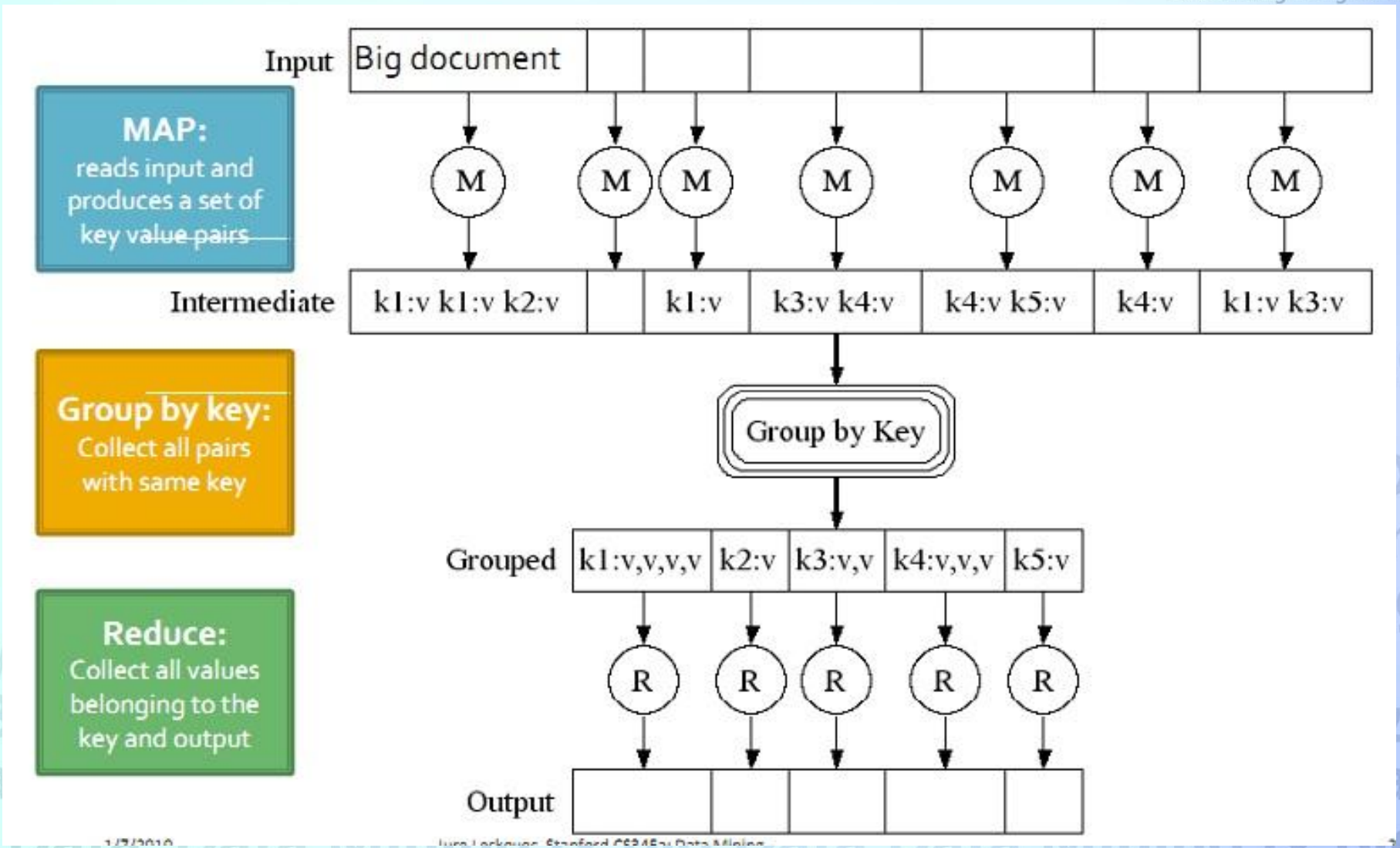
MapReduce - это модель программирования для обработки и генерации больших наборов данных. В настоящий момент типовой подход параллельной обработки больших объемов сырых данных. Разработана **Google**.

Многие практические задачи могут быть реализованы в данной модели программирования.

Работа **MapReduce** состоит из двух шагов: Map и Reduce.
На **Map-шаге** происходит предварительная обработка входных данных. Для этого один из компьютеров (называемый главным узлом — **master node**) получает входные данные задачи, разделяет их на части и передает другим компьютерам (рабочим узлам — **worker node**) для предварительной обработки.
На **Reduce-шаге** происходит свертка предварительно обработанных данных. Главный узел получает ответы от рабочих узлов и на их основе формирует результат — решение задачи, которая формулировалась изначально.

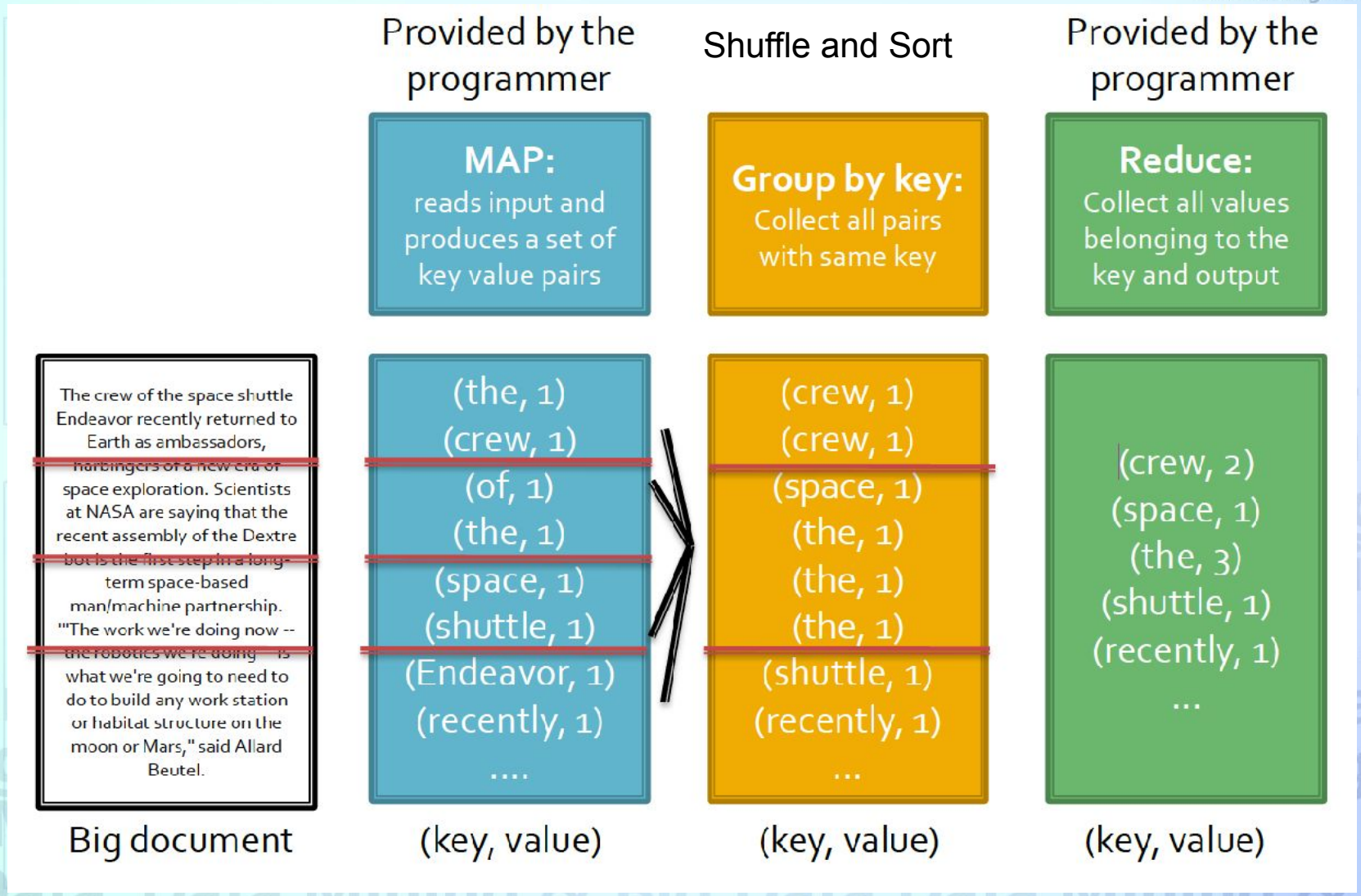
Пользователи задают функцию **Map**, которая обрабатывает пары ключ/значение для генерации набора промежуточных пар ключ/значение, и функцию **Reduce**, которая объединяет все промежуточные значения, связанные с одним и тем же промежуточным ключом.

MapReduce
Диаграмма



MapReduce

подсчет статистики по словам



Примеры заданий для MapReduce

Распределенный Grep: Map функция выдает строку, если она совпадает с заданным шаблоном. Reduce функция в этом случае просто копирует промежуточные данные в выходной файл.

Подсчет частоты доступа к URL: Функция Map обрабатывает логи запросов к веб-странице и выдает <URL; 1>. Функция Reduce суммирует все значения для одних и тех же URL и выдает пары <URL; общее количество>.

Инвертированный индекс: Функция Map анализирует каждый документ и формирует последовательность пар <слово; идентификатор документа>. Функция Reduce принимает все пары для данного слова, сортирует соответствующие идентификаторы документов и формирует пары <слово; список(идентификатор документа)>. Множество всех таких пар образует простой инвертированный индекс.

Пример обучения с учителем на MapReduce

Обучение модели Neural Network на данных эмпирической выборки

$$L(w) = \frac{1}{N}\sum_i L(y_i, f_w(x_i)) \to \min_w$$

Например решать методом градиентного спуска

$$\nabla L(w) = \frac{1}{N}\sum_i \nabla L(y_i, f_w(x_i))$$

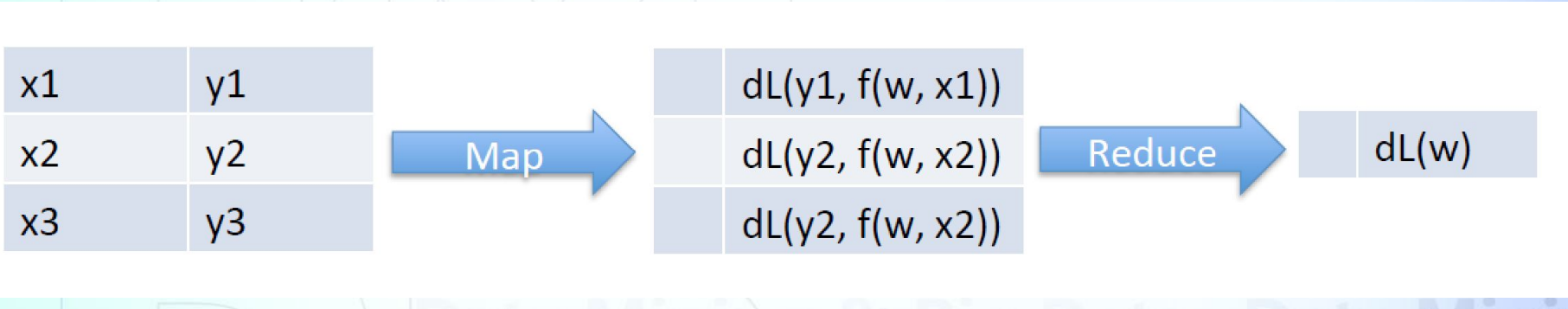
Веса сети корректируются в соответствии с

$$w \leftarrow w - \alpha \nabla L(w)$$

N может быть очень большим. Тогда каждый шаг спуска будет требовать вычисления и суммирования большого числа членов.

Пример обучения с учителем на MapReduce

Каждый шаг градиентного спуска можно выполнить с помощью map и reduce:



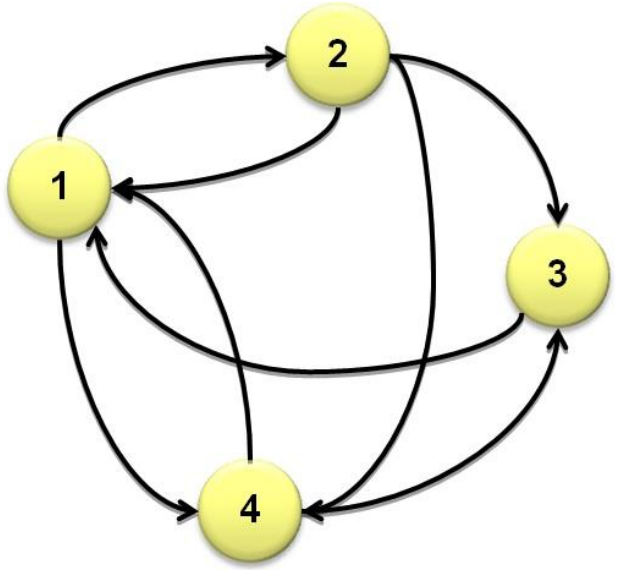
Графы и MapReduce

Граф G = (V, E) можно представить посредством:

1) Матрицы смежности (Adjacency matrix)
2) Списка смежности (Adjacency list)

Матрица смежности. Представляет граф как *n* x *n* квадратную матрицу *M.*
n = |V|, M*ij* = 1 означает наличие ребра от узла *i* к узлу *j.*

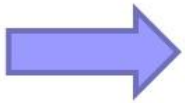
	1	2	3	4
1	0	1	0	1
2	1	0	1	1
3	1	0	0	0
4	1	0	1	0



Графы и MapReduce

Список смежности. Из матрицы смежности… «вытряхиваются» все нули….

	1	2	3	4
1	0	1	0	1
2	1	0	1	1
3	1	0	0	0
4	1	0	1	0



1: 2, 4
2: 1, 3, 4
3: 1
4: 1, 3

Программные реализации MapReduce

Google реализовал MapReduce на C++ реализовал MapReduce на C++ с интерфейсами на языках Python реализовал MapReduce на C++ с интерфейсами на языках Python и Java.
Greenplum — коммерческая реализация с поддержкой языков Python — коммерческая реализация с поддержкой языков Python, Perl — коммерческая реализация с поддержкой языков Python, Perl, SQL и других.
GridGain — бесплатная реализация с открытым исходным кодом на языке Java.
Apache Hadoop — бесплатная реализация MapReduce с открытым исходным кодом на языке Java.
Phoenix — реализация MapReduce на языке Си с использованием разделяемой памяти.
Qt Concurrent — упрощённая версия фреймворка, реализованная средствами Qt — упрощённая версия фреймворка, реализованная средствами Qt на C++, которая используется для распределения задачи между несколькими ядрами одного компьютера.
CouchDB использует MapReduce для определения представлений поверх распределённых документов
MongoDB позволяет использовать MapReduce для параллельной обработки запросов на нескольких серверах
Skynet — реализация с открытым исходным кодом на языке Ruby
Disco — реализация, созданная компанией Nokia — реализация, созданная компанией Nokia, её ядро написано на языке Erlang, а приложения для неё можно писать на языке Python.
Apache Hive — надстройка с открытым исходным кодом от Facebook — надстройка с открытым исходным кодом от Facebook, позволяющая комбинировать Hadoop и доступ к данным на SQL-подобном языке.

Поиск похожих объектов

Многие задачи могут быть озвучены, как «найти похожие объекты»

Примеры:

- Веб - страницы с похожими словами (классификация, распределение дубликатов)
- Покупатели с «похожими интересами»
- Изображения с «похожими признаками»
- Пользователи, которые посещают один и тот же веб-сайт
- и т.д.

Оценка «похожести объектов» после их датификации возможна на основе сравнения их как математических объектов с помощью так называемых метрик расстояния

Поиск похожих объектов
Метрики расстояний

L_2 norm: d(p,q)

$$\mathrm{d}(\mathbf{p},\mathbf{q}) = \mathrm{d}(\mathbf{q},\mathbf{p}) = \sqrt{(q_1-p_1)^2+(q_2-p_2)^2+\cdots+(q_n-p_n)^2} = \sqrt{\sum_{i=1}^{n}(q_i-p_i)^2}.$$

L_1 norm:

Сумма абсолютных разниц по каждому измерению

$$d_1(\mathbf{p},\mathbf{q}) = \|\mathbf{p}-\mathbf{q}\|_1 = \sum_{i=1}^{n}|p_i-q_i|,$$

Манхетенновское расстояние (в честь решетчатой структуры некоторых районов Нью-Йорка) (можно двигаться только вдоль осей)

L_∞ norm: d(x,y)

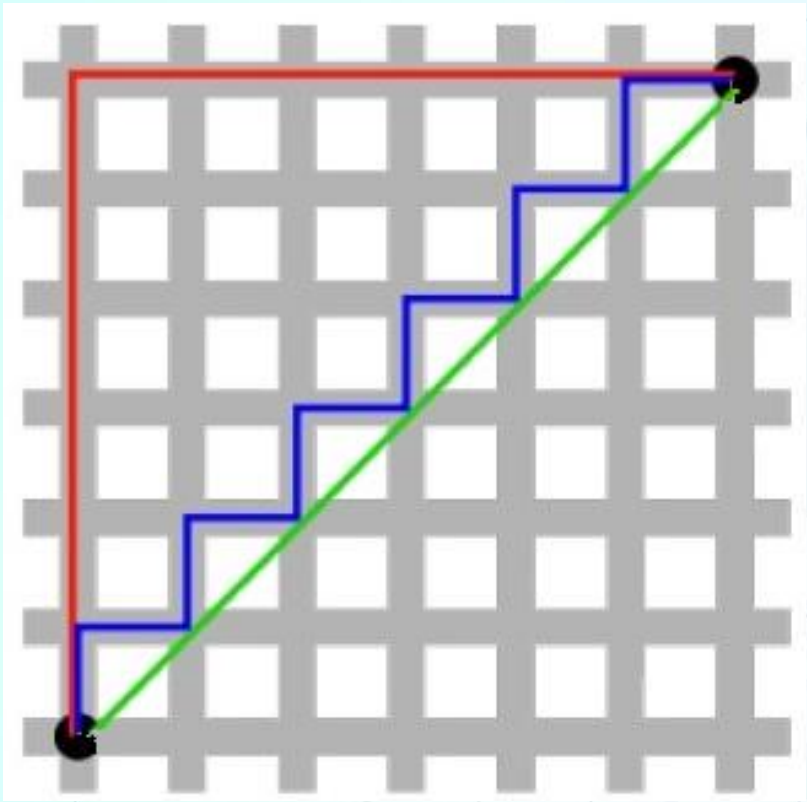
$$l_\infty(\vec{x},\vec{y}) = \max_{i=1,\ldots,n}|x_i-y_i|$$

Чебышевское расстояние

L_p norm:

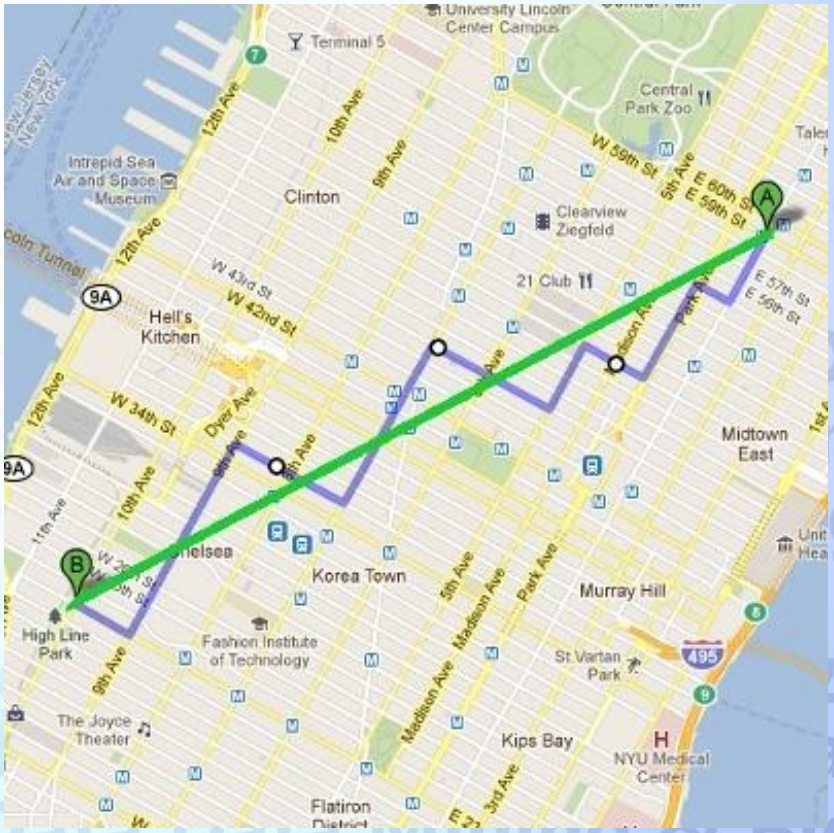
$$d([x_1,x_2,\ldots,x_n],\ [y_1,y_2,\ldots,y_n]) = (\sum_{i=1}^{n}|x_i-y_i|^r)^{1/r}$$

Метрики расстояний
Пример



1) Длина зеленого отрезка (L2) ≈ 8, 435

2) Синей ломаной (L1) = 12



3) Красной = 12

(Чебышевское расстояние = 6)

4) L4 ≈ 7, 135

Другие метрики расстояний

1. **Косинусное расстояние (Cosine Distance)** = это угол между векторами.

$$\cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}}$$

например, A = 00111; B = 10011

- $A \cdot B = 2$; $\|A\| = \|B\| = \sqrt{3}$
- $\cos(\theta) = 2/3$;

2. **Edit distance** = число вставок, удалений, которое нужно чтобы преобразовать одну строку в другую.

$$d(x,y) = |x| + |y| - 2|LCS(x,y)|$$

LCS (longest common subsequence) = наибольшая общая подпоследовательность (последовательность символов, следующих слева направо, но необязательно в порядке «друг за другом»)

Пример, x = **abcde**; y = **bcduve**

LCS(x,y) = **bcde**, d(x,y) = 5 + 6 – 2 * 4 = 3

Edit distance
Пример из биоинформатики

Анализ первичных последовательностей

Азотистые основания, входящие в ДНК:

A – аденин, **C** – цитозин, **G** – гуанин, **T** – тимин

S1 = AAACCGTGAGTTATTCGTTCTAGAA (25 символов)

S2 = CACCCCTAAGGTACCTTTGGTTC (23 символа)

Выделяем последовательность LSG (красным)

S1 = AAACCGTGAGTTATTCGTTCTAGAA
S2 = CACCCCTAAGGTACCTTTGGTTC

LSG(S1,S2) = ACCTAGTACTTTG (13 символов)

Edit Distance D(S1,S2) = 25 + 23 – 2*13 = 48 – 26 **= 22**

Другие метрики расстояний

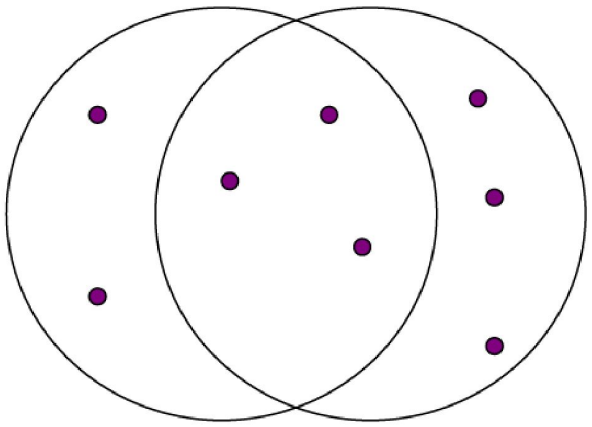
3. **Расстояние Хемминга** (Hamming Distance) = число позиций, в которых соответствующие символы двух слов одинаковой длины различны.

Пример, x = 10101, y = 10011

Hamming Distance = d(x,y) = 2

4. **Jaccard Distance** между двумя наборами – это 1 минус «размер их пересечения»/ «размер их объединения»

- $d(C_1, C_2) = 1 - |C_1 \cap C_2| / |C_1 \cup C_2|$



Пример

размер пересечения = 3
размер объединения = 8

Jaccard Distance = 1 – 3/8 = 5/8

Big Data Analytics

Big data is generally understood to refer to techniques developed to analyse data sets which are either too big, too complex or too lacking in structure to be analysed using standard approaches. A common misconception around big data is the expectation that acquiring powerful computer infrastructure will immediately provide a business advantage. Instead information technology, computer science and mathematical science must go hand in hand. Infrastructure is necessary, but achieving value from big data also requires more sophisticated data analysis methods.

New approaches to analysing data have to be found and, where appropriate, existing methods have to be scaled. This is where the mathematical sciences can make a considerable contribution: building on the foundations of current statistical methods and identifying new techniques to augment or replace old ones that are less appropriate, making the analytics efficient, and most importantly making sure the correct inferences are drawn from the data available.

«Data Science: Exploring the Mathematical Foundations». Smith Institute. UK. 2014

Big Data Analytics

С точки зрения бизнеса

IMPLEMENTING BIG DATA: 7 TECHNIQUES TO CONSIDER

Whether your business wants to discover interesting correlations, categorize people into groups, optimally schedule resources, or set billing rates, a basic understanding of the seven techniques mentioned above can help Big Data work for you.

1) Association rule learning
2) Classification tree analysis
3) Genetic algorithms
4) Machine learning
5) Regression analysis
6) Sentiment analysis
7) **Social network analysis**

«7 Big Data Techniques That Create Business Value» By Debbie Stephenson - Jan 18, 2013
http://www.firmex.com/thedealroom/

Big Data Analytics

1. ASSOCIATION RULE LEARNING

Are people who purchase tea more or less likely to purchase carbonated drinks?

Association rule learning is a method for discovering interesting correlations between variables in large databases. It was first used by major supermarket chains to discover interesting relations between products, using data from supermarket point-of-sale (POS) systems.

Люди, которые покупают чай больше или меньше шансов приобрести газированные напитки?

Обучение по ассоциативным правилам-это метод для обнаружения интересных корреляций между переменными в больших базах данных. Впервые он был использован крупными сетями супермаркетов, чтобы обнаружить интересные взаимоотношения между изделиями, используя данные из супермаркета в точках продаж (POS) системы.

Big Data Analytics

2. CLASSIFICATION TREE ANALYSIS

Which categories does this document belong to?

Statistical classification is a method of identifying categories that a new observation belongs to. It requires a training set of correctly identified observations – historical data in other words.

Statistical classification is being used to:
- automatically assign documents to categories
- categorize organisms into groupings
- develop profiles of students who take online courses

Big Data Analytics

3. GENETIC ALGORITHMS

Which TV programs should we broadcast, and in what time slot, to maximize our ratings?

Genetic algorithms are inspired by the way evolution works – that is, through mechanisms such as inheritance, mutation and natural selection. These mechanisms are used to "evolve" useful solutions to problems that require optimization.

Genetic algorithms are being used to:

- schedule doctors for hospital emergency rooms
- return combinations of the optimal materials and engineering practices required to develop fuel-efficient cars
- generate "artificially creative" content such as puns and jokes

Big Data Analytics

4. MACHINE LEARNING

Which movies from our catalogue would this customer most likely want to watch next, based on their viewing history?

Machine learning includes software that can learn from data. It gives computers the ability to learn without being explicitly programmed, and is focused on making predictions based on known properties learned from sets of “training data.”

Machine learning is being used to help:

- distinguish between spam and non-spam email messages
- learn user preferences and make recommendations based on this information
- determine the best content for engaging prospective customers
- determine the probability of winning a case, and setting legal billing rates

Big Data Analytics

5. REGRESSION ANALYSIS

How does your age affect the kind of car you buy?

At a basic level, regression analysis involves manipulating some independent variable (i.e. background music) to see how it influences a dependent variable (i.e. time spent in store). It describes how the value of a dependent variable changes when the independent variable is varied. It works best with continuous quantitative data like weight, speed or age.

Regression analysis is being used to determine how:

- levels of customer satisfaction affect customer loyalty
- the number of supports calls received may be influenced by the weather forecast given the previous day
- neighbourhood and size affect the listing price of houses
- to find the love of your life via online dating sites

Big Data Analytics

6. SENTIMENT ANALYSIS

How well is our new return policy being received?

Sentiment analysis helps researchers determine the sentiments of speakers or writers with respect to a topic.

Sentiment analysis is being used to help:

- improve service at a hotel chain by analyzing guest comments
- customize incentives and services to address what customers are really asking for
- determine what consumers really think based on opinions from social media

Big Data Analytics

7. SOCIAL NETWORK ANALYSIS

How many degrees of separation are you from Kevin Bacon?

Social network analysis is a technique that was first used in the telecommunications industry, and then quickly adopted by sociologists to study interpersonal relationships. It is now being applied to analyze the relationships between people in many fields and commercial activities. Nodes represent individuals within a network, while ties represent the relationships between the individuals.

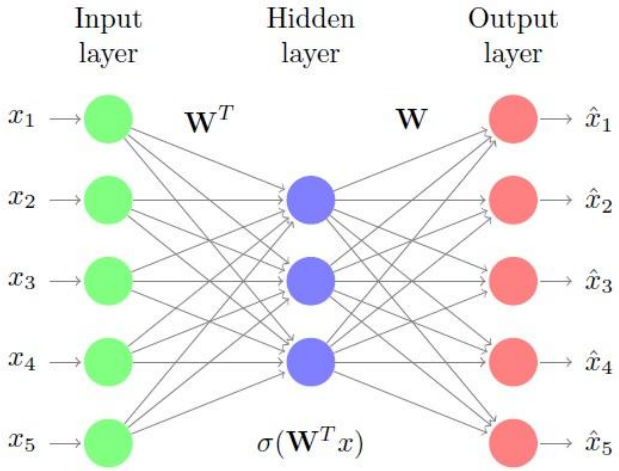
Social network analysis is being used to:

- see how people from different populations form ties with outsiders
- find the importance or influence of a particular individual within a group
- find the minimum number of direct ties required to connect two individuals
- understand the social structure of a customer base

Big Data Analytics

Autoencoders

- In deep learning, multiple In the neural network literature, an autoencoder generalizes the idea of principal components. Figure below provides a simple illustration of the idea, which is based on a reconstruction idea.



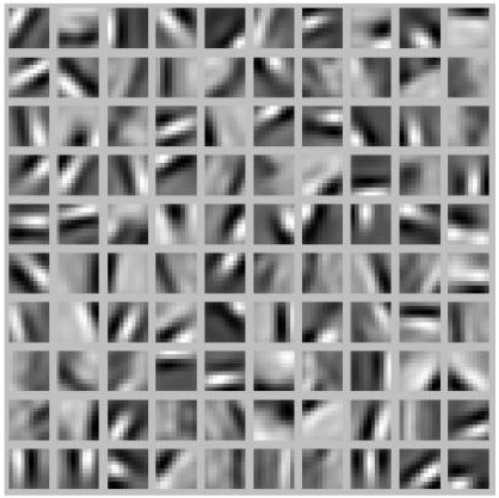


Figure : Left: Network representation of an autoencoder used for unsupervised learning of nonlinear principal components. The middle layer of hidden units creates a bottleneck, and learns nonlinear representations of the inputs. The output layer is the transpose of the input layer, and so the network tries to reproduce the input data using this restrictive representation. Right: Images representing the estimated columns of $\mathbf{W}$ in an image modeling task.

Mathematics for Analysis of Petascale Data

needs of various scientific domains

Application Domains

1 Astrophysics
2 Biology
3 Nanoscience
4 **Power and Communication Networks**
5 Earth and Climate Systems Modeling
6 Fusion Physics
7 Accelerator Physics
8 Cybersecurity
9 Combustion
10 Visualization

Mathematics Research Findings

1 Scalability
2 Distributed Data
3 Architectures
4 Data and Dimension Reduction
5 Models
6 Uncertainty

Big Data Analytics

- Statistics
- Optimization
- Uncertainty quantification
- Machine learning
- Network and graph analysis
- Analysis of streaming data
- Data reduction (which includes dimension reduction, feature extraction, and topological methods).

Big Data Analytics

The analysis of extensive quantities of data and the need to grasp value out of individual behaviors require processing methods that go beyond the traditional statistical techniques.

Both Manyika et al. (2011) and Chen (2012) propose a list of Big Data Analytical Methods, that include (in alphabetical order):
A/B testing, Association rule learning, Classification, Cluster analysis, Data fusion and data integration, Ensemble learning, Genetic algorithms, Machine learning, Natural Language Processing, Neural networks, Network analysis, Pattern recognition, Predictive modelling, Regression, Sentiment Analysis, Signal Processing, Spatial analysis, Statistics, Supervised and Unsupervised learning, Simulation, Time series analysis and Visualization.

List of Big Data Analytical Methods

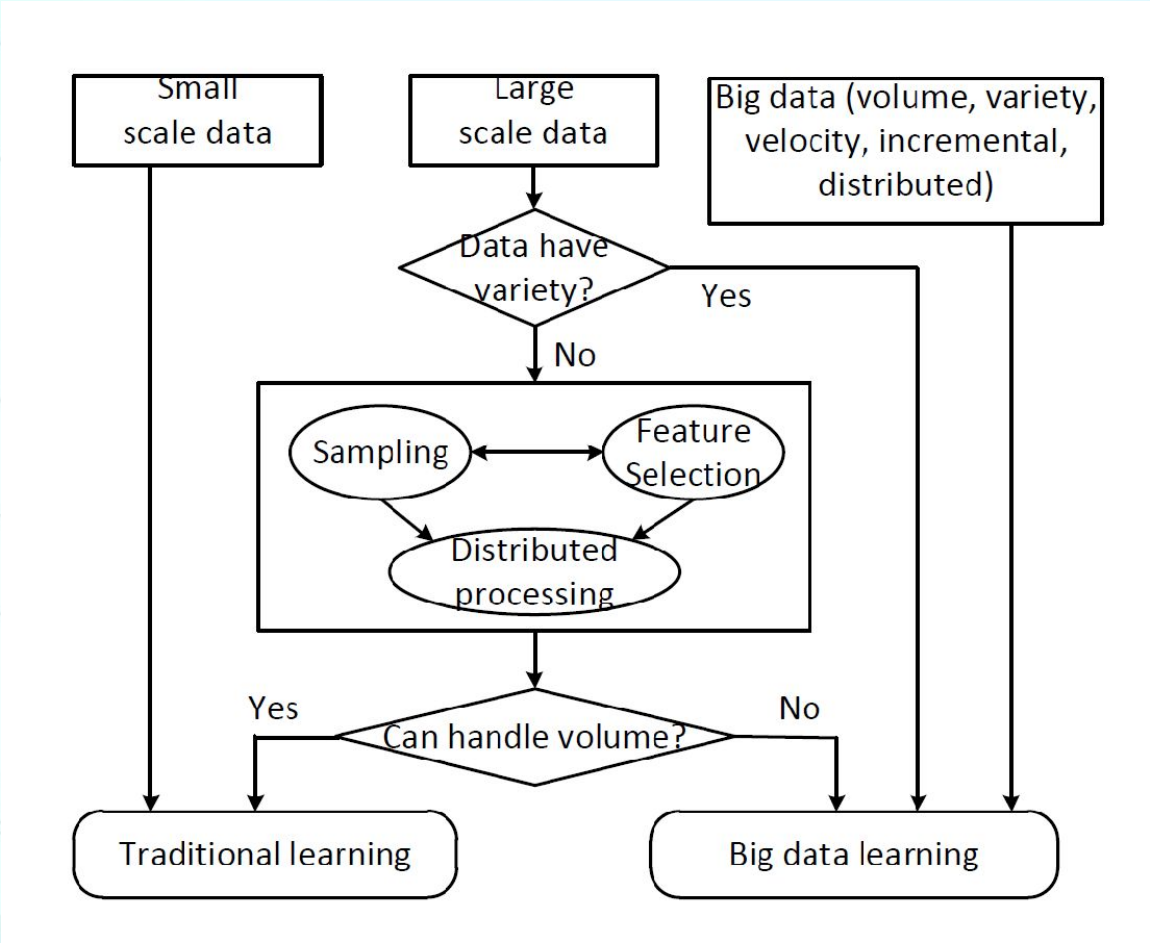
1) A/B testing
2) Association rule learning
3) Classification
4) Cluster analysis
5) Data fusion and data integration
6) Ensemble learning
7) Genetic algorithms
8) Machine learning
9) Natural Language Processing
10) Neural networks
11) Network analysis
12) Pattern recognition
13) Predictive modelling
14) Regression
15) Sentiment Analysis
16) Signal Processing
17) Spatial analysis
18) Statistics
19) Supervised and Unsupervised learning
20) Simulation
21) Time series analysis
22) Visualization

Big Data Analytics

Being aware of the limitations of Big Data Methods and potential methodological issues is a fundamental resource for organizations who want to drive data-based decision making: for example, predictions should always be accompanied by valid confidence intervals in order to avoid the false sense of precision that the apparent sophistication of some Big Data applications can suggest. Analysts should also be capable of avoiding models' overfitting that would facilitate apophenia, i.e. the tendency of humans to "see patterns where none actually exist simply because enormous quantities of data can offer connections that radiate in all directions", (Boyd & Crawford 2012).

Andrea De Mauro et al.«What is Big Data? A Consensual Definition and a Review of Key Research Topics». AIP Proceedings", 2014.

Big Data Analytics



Traditional data mining and mining of Big Data

Big Data Analytics

Задачи

- Классификация — отнесение входного вектора (объекта, события, наблюдения) к одному из заранее известных классов.
- Кластеризация — разделение множества входных векторов на группы (кластеры) по степени «похожести» друг на друга.
- Сокращение описания — для визуализации данных, лаконизма моделей, упрощения счета и интерпретации, сжатия объемов собираемой и хранимой информации.
- Ассоциация — поиск повторяющихся образцов. Например, поиск «устойчивых связей в корзине покупателя» (market basket analysis) — вместе с пивом часто покупают орешки.
- Прогнозирование
- Анализ отклонений — например, выявление нетипичной сетевой активности позволяет обнаружить вредоносные программы.
- Визуализация

Big Data Analytics

Методы и примеры

- Классификация и предсказание (classification and prediction)
  Пример – целенаправленный найм (focused hiring)
- Кластерный анализ (cluster analysis)
  Пример – сегментирование рынка
- Анализ выбросов (outlier analysis)
  Пример – обнаружение мошенничества
- Анализ скрытых закономерностей (association analysis)
  Пример – анализ рыночной корзины
- Эволюционные алгоритмы (evolution analysis, genetic algorithms)
  Пример – прогнозирование индекса фондового рынка с помощью анализа временных рядов

Data Mining
Применения

Data Mining для анализа финансовых данных

Проектирование и строительство хранилищ данных для многомерного анализа данных и Data Minig

Прогнозирование платежей по кредиту и анализ кредитной политики

Классификация и кластеризация клиентов для целевого маркетинга

Выявление случаев отмывания денег и других финансовых преступлений

Интеллектуальный анализ данных для розничной торговли

Data Mining
Применения

Data Mining в розничной торговле

Проектирование и построение хранилищ данных на основе использования преимуществ технологий интеллектуального анализа данных

Многомерный анализ продаж, клиентов, продуктов, времени и региона

Анализ эффективности сбытовых кампаний

Удержание клиентов – анализ лояльности клиентов

Рекомендация продуктов

Data Mining
Применения

Data Mining в телекоммуникациях

Многомерный анализ телекоммуникационных данных

Выявление необычных паттернов и определение мошенничества

Сервисы мобильной связи

Использование средств визуализации в анализе телекоммуникационных данных

Big Data Analytics

Примеры

Методы классификации и прогнозирования. Деревья решений

Метод деревьев решений (decision trees) является одним из наиболее популярных методов решения задач классификации и прогнозирования.

Деревья решений – довольно старый метод, он предложен в конце 50-х годов прошлого века.

В наиболее простом виде дерево решений – это способ представления правил в иерархической, последовательной структуре. Основа такой структуры – ответы «да» или «нет» на ряд вопросов.

Алгоритмы конструирования деревьев решений состоят из этапов «создание» дерева (tree building) и «сокращение» дерева (tree pruning). В ходе создания дерева решаются вопросы выбора критерия расщепления и остановки обучения (если это предусмотрено алгоритмом). В ходе этапа сокращения дерева решается вопрос отсечения некоторых его ветвей.

Метод деревьев решений часто называют «наивным» подходом.

Data Mining

Примеры

Деревья решений. Играть ли в гольф?

Пусть решается задача, в которой надо ответить на вопрос: «Играть ли в гольф?». Чтобы решить задачу текущую ситуацию к одному из известных классов (в данном случае – «играть» или «не играть»).

Для этого требуется ответить на ряд вопросов, которые находятся в узлах этого дерева, начиная с его корня.

В результате прохождения от корня дерева до его вершины решается задача классификации.

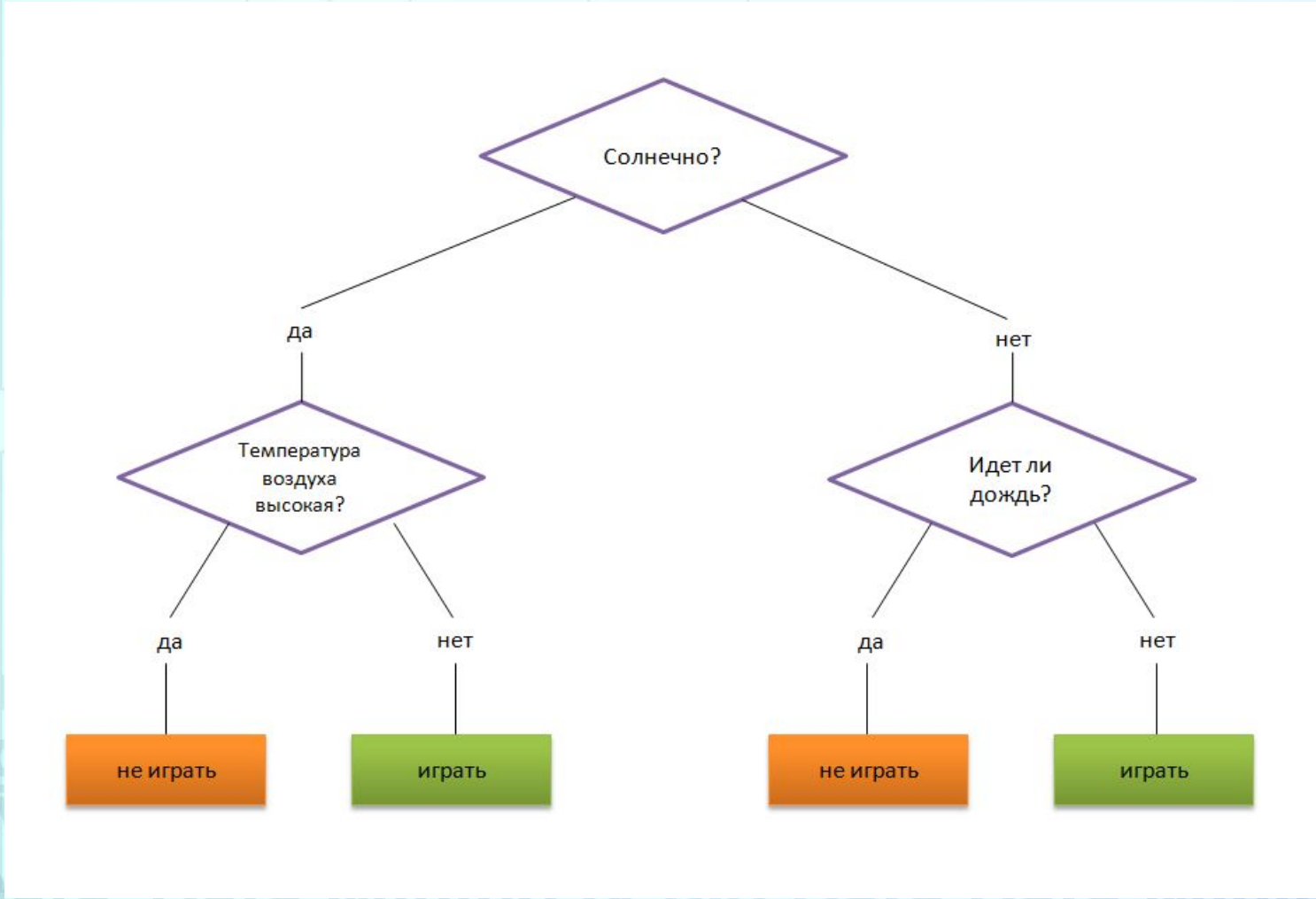
Рассмотренный пример относится к задачам бинарной классификации, т.е. создается дихотомическая классификационная модель.

В узлах бинарных деревьев ветвление может вестись только в двух направлениях, т.е. существует возможность только двух ответов на поставленный вопрос («да» или «нет).

Data Mining

Примеры

Играть ли в гольф?



Big Data Analytics

Примеры

Деревья решений. Задача об оценке кредитного риска.

База данных содержит ретроспективные данные о клиентах банка, являющиеся её атрибутами: годовой доход, долги, займы, кредитная история и т.д.

Такая задача классификации решается в два этапа: построение классификационной модели и её использование.

Атрибуты базы данных являются внутренними узлами дерева.

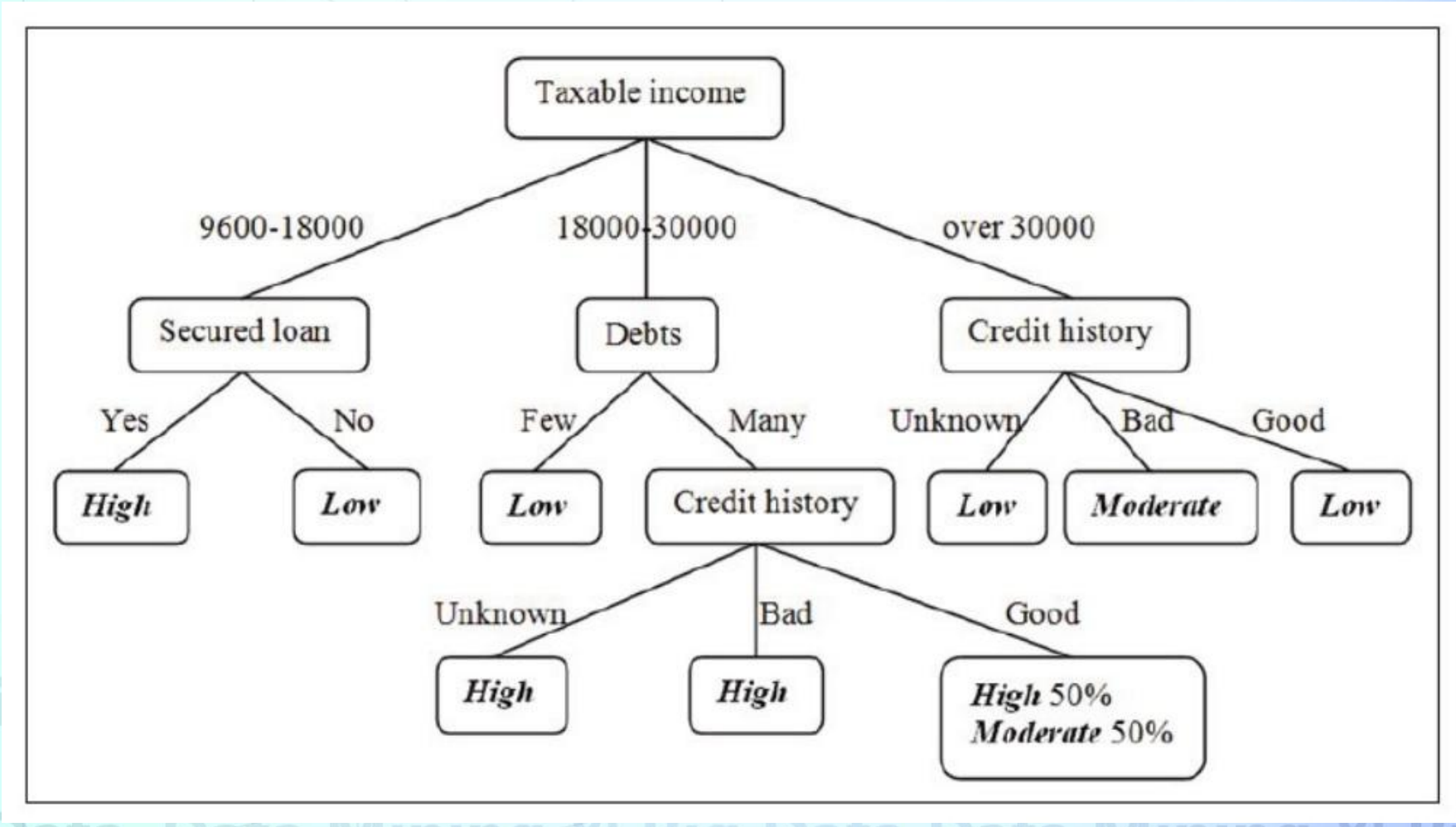
Эти атрибуты называют прогнозирующими, или атрибутами расщепления (splitting attribute).

Конечные узлы дерева, или листы, именуются метками класса, являющимися значениями зависимой категориальной переменной «кредитный риск»: Low, Moderate, High.

Big Data Analytics

Примеры

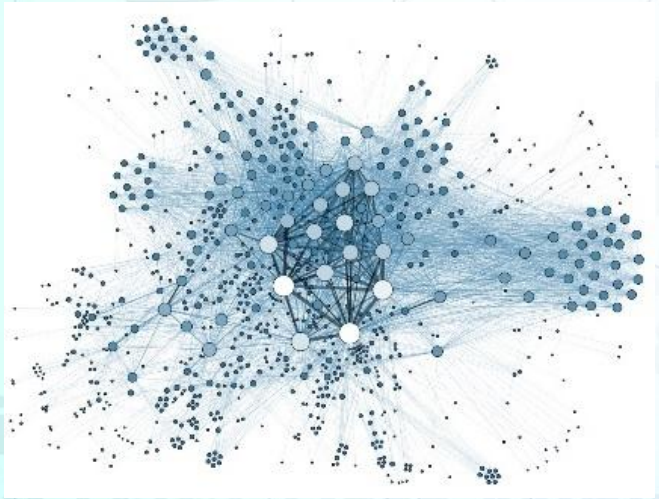
Методы классификации и прогнозирования. Деревья решений

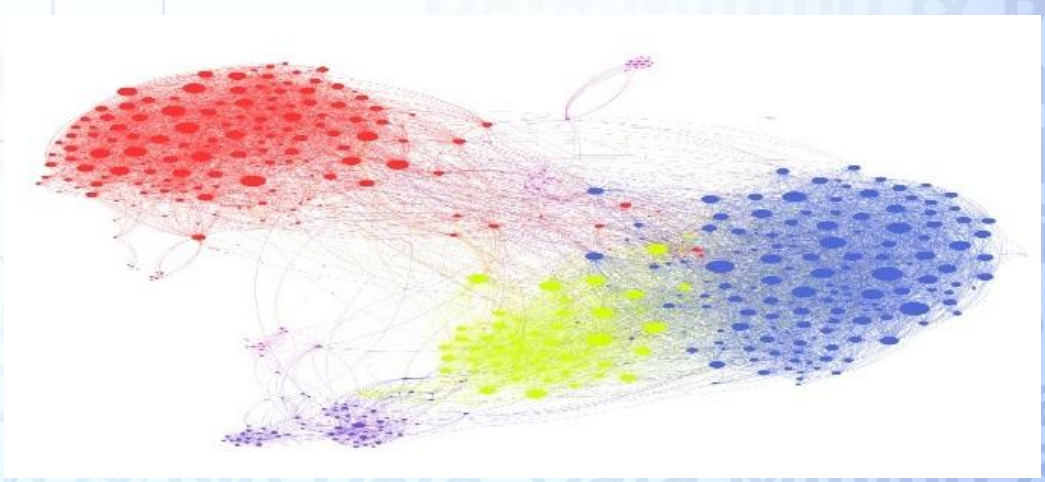


Big Data

Большие данные в разных отраслях экономики

1) Выделение архетипа пользователя
2) Выделение связей между группами пользователей
3) Выделение сообществ
4) Анализ круга общения
5) Выделение нетипичных пользователей
6) Прогнозирование новых связей

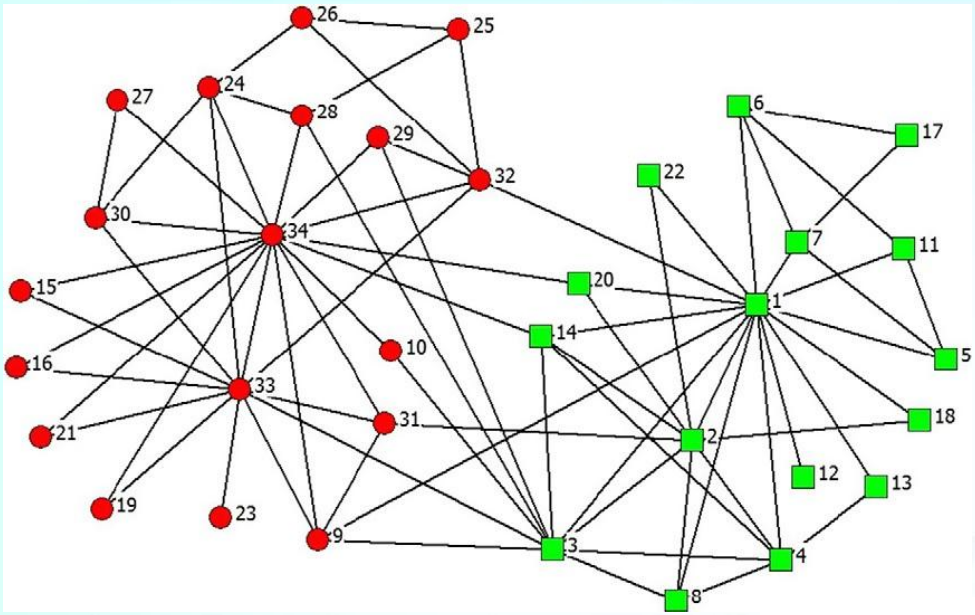




Примеры

Задачи кластеризации на графах

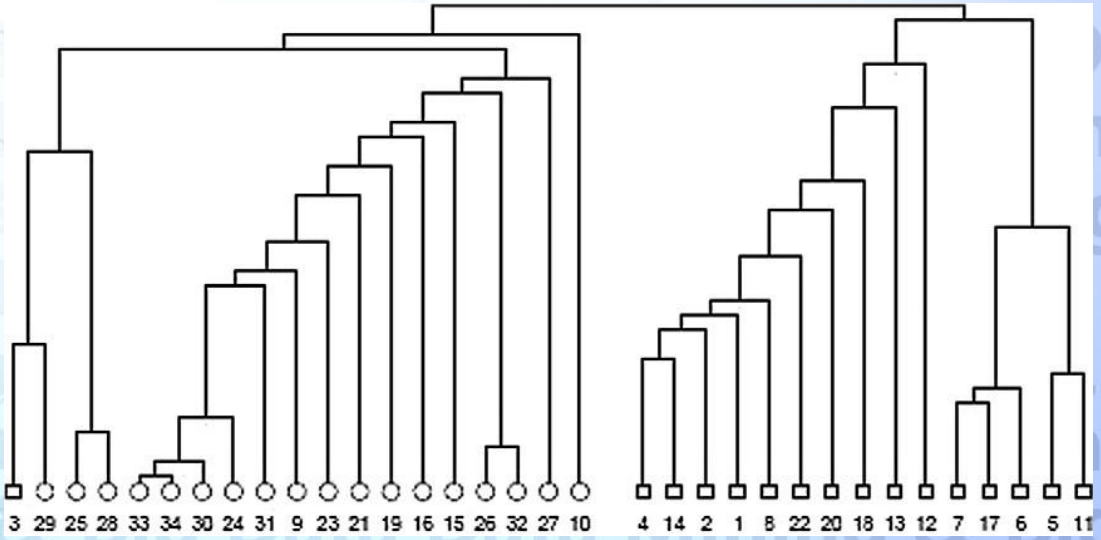
применение алгоритма Girvan and Newman



The network of friendships in the karate club study

Социальная сеть, известная как “клуб карате”, построенная Zachary. В течении 2 лет он наблюдал за 34 членами клуба. В течение этого срока члены клуба разделились на две группы вследствие споров между администратором клуба и тренером. Члены одной из групп основали свой собственный клуб.

Result of Girvan and Newman algorithm

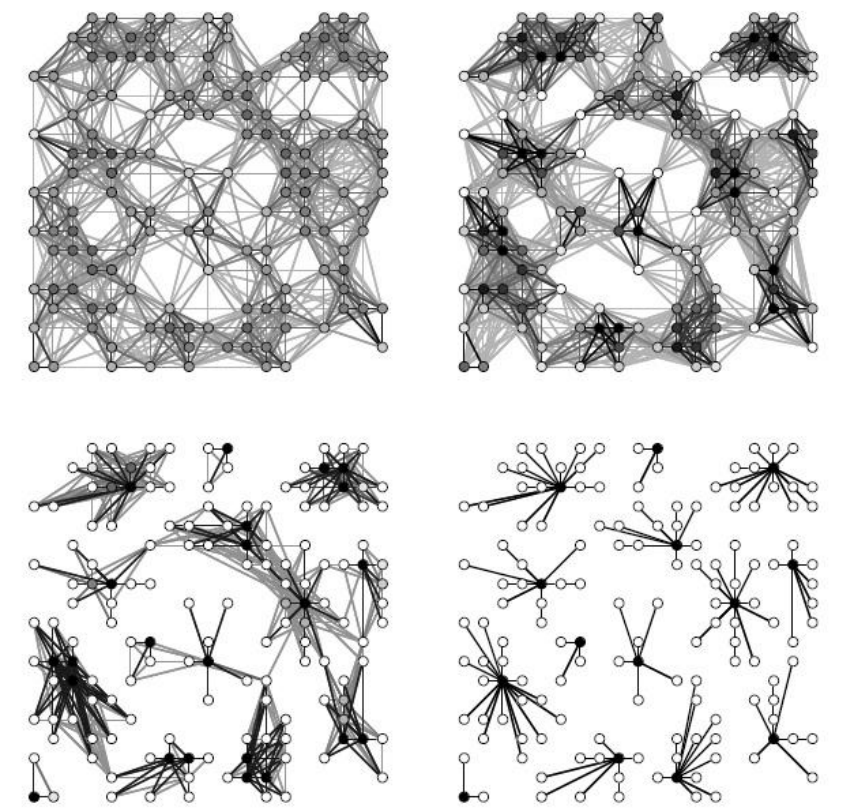


Zachary построил простой не взвешенный граф, чтобы отразить отношения дружбы между каждой парой членов клуба. Каждый член клуба представляется на графе узлом, а ребро появляется между узлами, если эти члены клуба являются друзьями вне пределов клуба.

Примеры

Задачи кластеризации на графах

Результаты применения метода MLP (Markov Cluster Algorithm)



Метод Dynamic Quantum Clustering

Авторы метода ставят задачу весьма парадоксальным образом: «Как искать иголку в многомерном стоге сена, не зная, как она выглядит, и, не зная, есть ли она в этом стоге». И отвечают, что подобная постановка требует смены парадигмы поиска в сторону «пусть данные говорят о себе сами».

Разработанная для анализа Больших многомерных данных методология «Dynamic Quantum Clustering» (DQC) реализует указанную парадигму.

Метод DQC (как и многие другие методы аналитики Больших данных) «работает» без предварительного знания о тех «структурах», их типе и топологии, которые могут быть «скрыты» в данных и выявлены в результате его применения. Метод хорошо работает с многомерными данными, и, что очень важно, время анализа линейно зависит от размерности

Метод Dynamic Quantum Clustering

В ***n***-мерном признаковом пространстве строится функция ***φ***, являющаяся суммой гауссовых функций с центрами в каждой точке данных (Парзеновская функция).

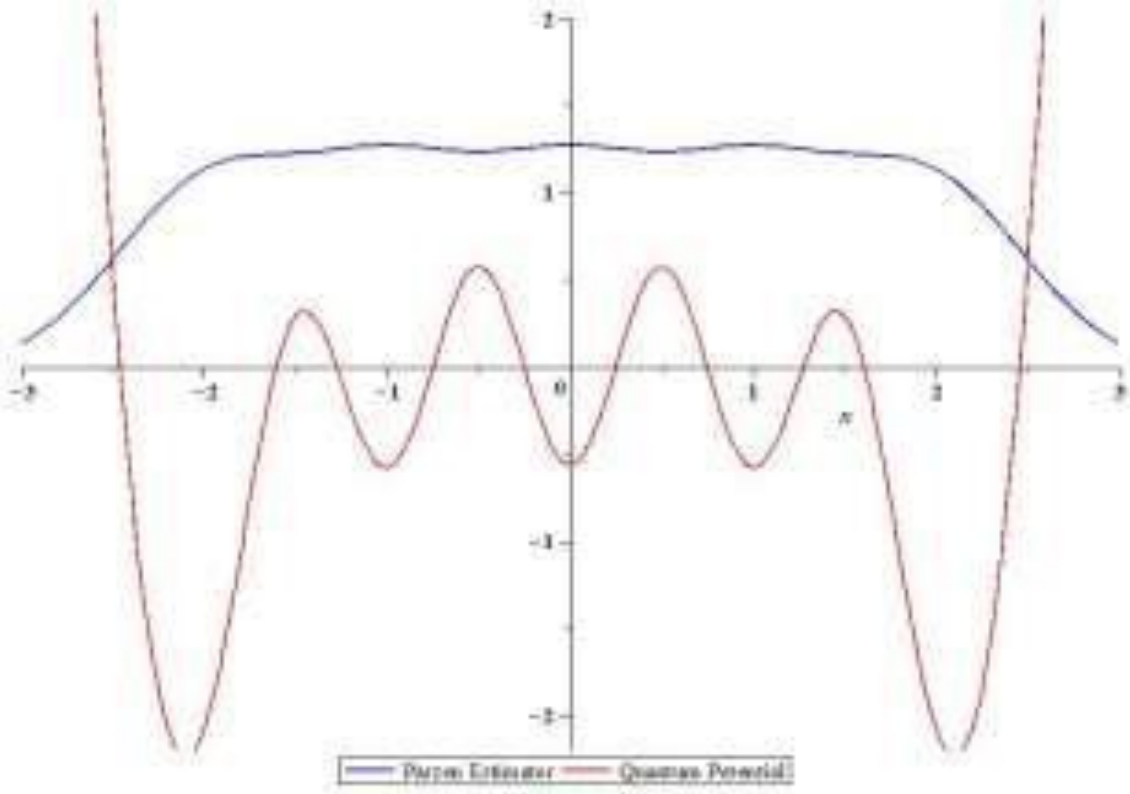
$$\varphi(\vec{x}) = \sum_{l=1}^{m} e^{-\frac{1}{2\sigma^2}(\vec{x}-\vec{x_l})\cdot(\vec{x}-\vec{x_l})}$$

Вычисляется функция квантового потенциала ***V***, удовлетворяющая уравнению Шредингера для ***φ***.

$$-\frac{1}{2\sigma^2}\nabla^2\varphi + V(\vec{x})\varphi = E\varphi = 0$$

Локальные минимумы функции ***V*** соответствуют локальным максимумам ***φ***, кроме того функция ***V*** может иметь минимумы там, где у ***φ*** нет максимума. Функция ***V*** лучше выявляет структуру данных, чем Парзеновская функция. Затем для каждого гауссиана, связанного с определенной точкой данных, задается его эволюция путем умножения его на квантовый время-эволюционный оператор. Вычисляются новые центры гауссианов, и процедура повторяется. Доказано, что новые центры стремятся к ближайшим минимумам потенциальной функции ***V***.

Метод Dynamic Quantum Clustering

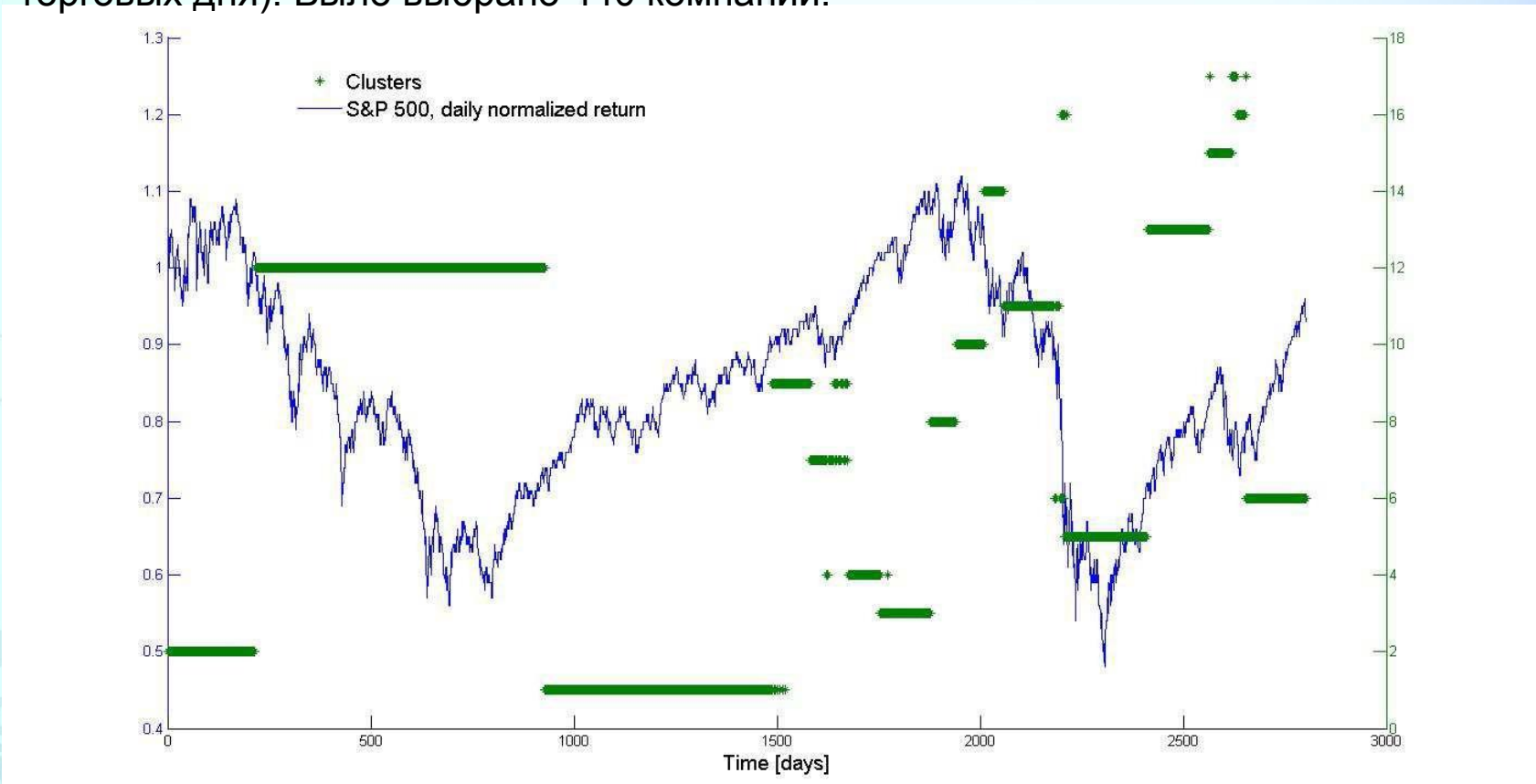


Парзеновская функция (синяя кривая) и соответствующий ей квантовый потенциал (красная кривая). Парзеновская функция является суммой 5 гауссианов с центрами (-2,-1,0,1,2)

Примеры

Результаты применения метода Quantum Clustering на примере данных фондового рынка

Анализ цен акций компаний, входящих в лист индекса Standard and Poor's S&P500 за период 1 января 2000 года по 24 февраля 2011 года (всего 2803 торговых дня). Было выбрано 440 компаний.



01.01.2000 – 24.03.2011

Примеры

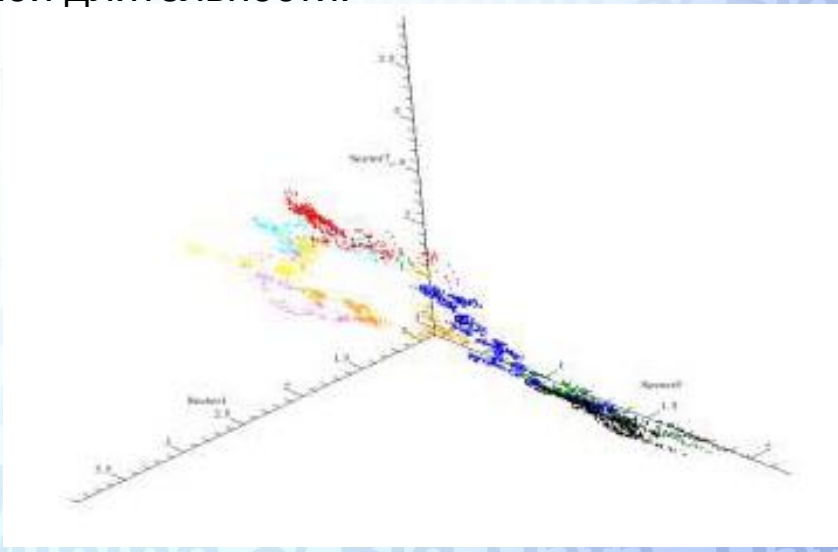
Результаты применения метода Quantum Clustering на примере данных фондового рынка

С математической точки зрения анализу подвергалась матрица размером 2803 x 440. Каждая строка матрицы содержит информацию о ценах всех 440 акций за один день.

Прежде всего, анализ подтвердил очевидный результат, что цены акций коррелированны в соответствии с принадлежностью к одному из 9 рыночных секторов («энергетический», «финансовый», «промышленный» и т.д.).

Однако главным результатом стало, что в результате анализа были выявлены «временные» кластеры, которые авторы назвали «рыночными эпохами». Всего за указанный период было выявлено 17 эпох различной длительности.

Результаты дальнейшего анализа показали, что каждый из «временных» кластеров имеет свои собственные характеристики. Это хорошо видно на рисунке (цветом выделены события, принадлежащие различным «эпохам»), где каждая точка представляет собой вектор из средних дневных цен акций 3 рыночных секторов, представленных осями координат.



01.01.2000 – 24.03.2011

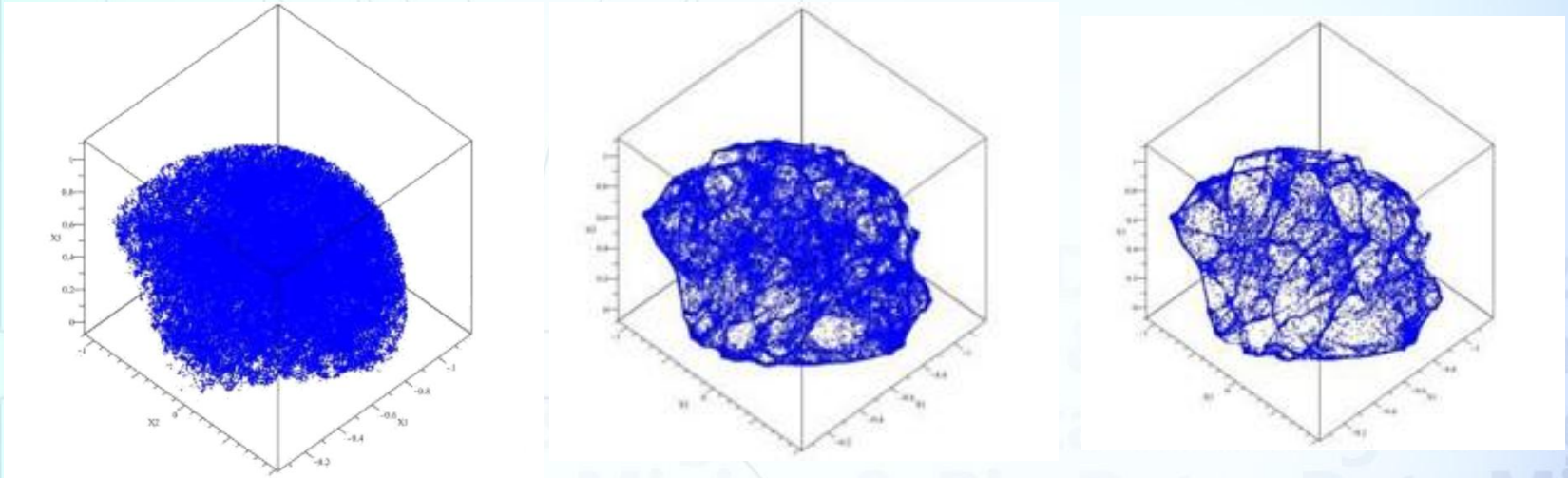
Примеры

Результаты применения метода Dynamic Quantum Clustering (DQC) на примере астрономических данных

Анализируется пространственное распределение 139798 галактик (каталогизированные данные из Sloan Digital Sky Survey (SDSS). Для каждой галактики известно три координаты – два угла φ и θ и величина так называемого «красного смещения» z, играющего роль расстояния до галактики. То, что галактики распределены во вселенной неравномерно – известный факт, их распределение напоминает полотно из волокон и пустот. Этот факт и был проверен с помощью метода DQC.

Примеры

Результаты применения метода Dynamic Quantum Clustering (DQC) на примере астрономических данных



Хорошо видна эволюция начального распределения (слева) к структуре, которая действительно напоминает полотно из волокон и пустот (справа)

"Гусеница" - будущий классический метод анализа временных рядов

- ***- Каковы принципиальные отличия метода "Гусеница" от других методов анализа временных рядов? В чем его уникальность?***
- - Первой идеей, лежащей в основе метода, является создание повторности путем перехода от временного ряда (последовательности некоторых измерений или характеристик в равноотстоящие моменты времени) к последовательности векторов, состоящих из отрезков временного ряда выбранной длины. Таким образом, получается что-то вроде многомерной выборки, так как если исходный ряд имел какую-то структуру, то и его отрезки наследуют эту структуру. Второй идеей является анализ полученной многомерной выборки (траекторной матрицы) с помощью ее сингулярного разложения или, используя статистические аналогии, анализа главных компонент. Тем самым получается разложение исходного временного ряда (точнее, его траекторной матрицы) по базису, порождаемому им самим.
- Мне кажется, одной из отличительных черт метода является его естественность. Метод не навязывает изначально какую-либо модель исследуемого временного ряда. Но при этом он позволяет так разложить ряд на элементарные составляющие, что по ним оказывается возможным воссоздать структуру ряда, например, выделить трэнд или найти периодические составляющие. Кроме этого, метод дает замечательную возможность очищать сигнал от шумовой составляющей.
- ***- "Гусеница" используется не только для анализа временных рядов, но еще и для их прогнозирования. Если сравнивать получаемые с его помощью прогнозы, а также результаты работы других методов, то какие из моделей оказываются наиболее точными?***
- - Проблема корректного сравнения различных методов не так проста, как кажется на первый взгляд. Одним из основных препятствий является интерактивность метода "Гусеница", что не позволяет проводить сравнение автоматически, на основе большого числа промоделированных или реальных данных. Другой аспект - это некорректность сравнения методов безотносительно к классу временных рядов. Например, для рядов, удовлетворяющих заданной модели, скорее всего, лучше будет метод, настроенный на эту модель.
- Совсем другое дело, если модель априори неизвестна. Тогда будут, в среднем, лучше проявлять себя методы, настроенные на более широкий класс рядов. К таким методам мы и относим "Гусеницу", которая применима для достаточно широкого класса рядов, но проигрывает незначительно, если его сравнивать с рядом известных параметрических методов (такими, как, например, линейная регрессия или разложение Фурье).

Заключение

Важнейшим условием успешного развития мировой экономики на современном этапе становится возможность фиксировать и анализировать огромные массивы и потоки информации. Существует точка зрения, что страны, которые овладеют наиболее эффективными методами работы с Большими данными, ждет новая индустриальная революция. Направление «BigData» концентрирует усилия в организации хранения, обработки, анализа огромных массивов данных.

У России с ее колоссальным научным и образовательным потенциалом есть все шансы занять достойное место среди тех экономик, где извлечение полезных знаний из больших объемов данных различной природы поставлено на службу индустриальному прогрессу.

Заключение

Важнейшим условием успешного развития мировой экономики на современном этапе становится возможность фиксировать и анализировать огромные массивы и потоки информации. Существует точка зрения, что страны, которые овладеют наиболее эффективными методами работы с Большими данными, ждет новая индустриальная революция. Направление «BigData» концентрирует усилия в организации хранения, обработки, анализа огромных массивов данных.

Big Data. Bibliography

1) Bernard Marr. “Big Data: Using SMART Big Data, Analytics and Metrics To Make Better Decisions and Improve Performance”. John Wiley & Sons Ltd, 2015.

2) Andrea De Mauro, Marco Greco and Michele Grimaldi. “What is Big Data? A Consensual Definition and a Review of Key Research Topics”. In “AIP Proceedings”2014, “4th International Conference on Integrated Information”.

3) Sofia Berto Villas-Boas. “Big Data in Firms and Economic Research”. Applied Economics and Finance, Vol. 1, No. 1; May 2014.

4) Liran Einav, Jonathan Levin. “The Data Revolution and Economic Analysis”. NBER Working Paper No. 19035, Issued in May 2013.

5) Тезисы докладов конференции «Большие данные в национальной экономике», Москва, 21 октября 2014 г.

6) Тезисы докладов конференции «Большие данные в национальной экономике», Москва, 22 октября 2013 г.

7) А. Климентов, А. Ваняшин, В. Кореньков. «За большими данными следит ПАНДА». Суперкомпьютеры, 15-2013, стр. 56.

Big Data. Bibliography

8) Денис Серов. “Аналитика “больших данных”– новые перспективы”. “Storage News”, №1 (49), 2012.

9) Zhanpeng Huang, Pan Hui, Christoph Peylo. “When Augmented Reality Meets Big Data”. arXiv:1407.7223v1.

10) Patrick J. Wolfe. “Making sense of big data”. PNAS. November 5, 2013, vol. 110, no. 45, 18031–18032.

11) Jure Leskovec, Anand Rajaraman, Jeffrey D. Ullman. “Mining of Massive Datasets”. Cambridge University Press. 2012.

12) M. Weinstein, F. Meirer, A. Hume, Ph. Sciau, G. Shaked, R. Hofstetter, E. Persi, A.Mehta, and D. Horn. “Analyzing Big Data with Dynamic Quantum Clustering”. arXiv:1310.2700.

13) Marvin Weinstein and David Horn, “Dynamic quantum clustering: A method for visual exploration of structures in data”. PHYSICAL REVIEW E 80, 066117 (2009).

Big Data. Bibliography

14) David Horn and Assaf Gottlieb. “The Method of Quantum Clustering”. Proceedings of the Neural Information Processing Systems: NIPS’01, 2001, pp. 769–776.

15) Vijay Gadepally & Jeremy Kepner. “Big Data Dimensional Analysis”. arXiv:1408.0517v1.

16) MOHAMED-ALI BELABBAS AND PATRICK J. WOLFE. “On landmark selection and sampling in high-dimensional data analysis”. Phil. Trans. R. Soc. A (2009) 367, 4295–4312.

17) Yonathan Aflalo and Ron Kimmel. “Spectral multidimensional scaling”. PNAS, November 5, 2013, vol. 110, no. 45, 18052–18057.

18) Shahar Ronen, Bruno Gonçalves, Kevin Z. Hu, Alessandro Vespignani, Steven Pinker, and César A. Hidalgo. “Links that speak: The global language network and its association with global fame”. PNAS. 2014. Vol. 111. No.52, pp. E5616-E5622.