RUDN University

Faculty of Physics, Mathematics and Natural Sciences

Foreign Languages Department

Professional translation

A. Gandomi, M. Haider. Ted Rogers School of Management, Ryerson University, Toronto, Ontario M5B2K3, Canada (2015). Beyond the hype: Big data concepts, methods and analytics. *International Journal of Information Management, 35*(2)*, 137-144*. *https://doi.org/10.1016/j.ijinfomgt.2014.10.007*

Выполнила

Студент группы НФИмд-01-22

Логинов Сергей Андреевич

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| phrases | dictionary | 10 sentences | 3 questions | translation | Σ |
| 2 | 1 | 1 | 1 | 1 | 6 |
|  |  |  |  |  |  |

**Dictionary**

|  |  |  |
| --- | --- | --- |
| Word | Transcription | translation |
| Revolutionary | rɛvəluˈʃɛnəri | Революционный |
| Profound | prəˈfaʊnd | Глубокий |
| Implications | ɪmplɪˈkeɪʃənz | Последствия/выводы |
| Evolutionary | ˌiːvəˈluːʃənri | Эволюционный |
| Methodology | ˌmɛθəˈdɒlədʒi | Методология |
| Correlation | ˌkɒrɪˈleɪʃən | Корреляция |
| Algorithm | ˈælɡərɪðəm | Алгоритм |
| Heterogeneity | ˌhɛtərəʊdʒɪˈniːəti | Гетерогенность |
| Econometric | ɛˌkɒnəˈmɛtrɪk | Эконометрика |
| Multivariate | ˌmʌltɪˈvɛəriət | Множественный,многомерный |

**Sentences for translation**

|  |  |  |
| --- | --- | --- |
| № | Предложение | Перевод |
| 1 | While big data has the potential to drive innovation and improve decision-making across a range of industries, it also presents significant challenges due to its high volume, velocity, and variety. | Хотя большие данные имеют потенциал для стимулирования инноваций и улучшения процесса принятия решений в различных отраслях, они также представляют значительные вызовы из-за их большого объема, скорости и разнообразия. |
| 2 | The application of statistical modeling, machine learning, and data visualization to big data requires special consideration due to the scale and complexity of the data, as well as the need for computational power and storage. | Применение статистического моделирования, машинного обучения и визуализации данных к большим данным требует особого внимания из-за масштаба и сложности данных, а также необходимости в вычислительной мощности и хранилищах. |
| 3 | Data quality is a critical factor in producing meaningful insights from big data, as inaccuracies, inconsistencies, and incompleteness can lead to erroneous conclusions and decision-making. | Качество данных является критическим фактором в получении значимых выводов из больших данных, поскольку неточности, несоответствия и неполнота могут привести к ошибочным выводам и принятию решений. |
| 4 | Effective data management is essential in big data analysis, including strategies for data collection, storage, retrieval, processing, and integration from various sources and formats. | Эффективное управление данными является необходимым в анализе больших данных, включая стратегии сбора, хранения, получения, обработки и интеграции из различных источников и форматов. |
| 5 | The ethical and social implications of big data analysis, including privacy, security, and bias, need to be carefully considered to avoid harm to individuals and communities and promote responsible use of data. | Этические и социальные последствия анализа больших данных, включая конфиденциальность, безопасность и предвзятость, должны быть тщательно рассмотрены, чтобы избежать вреда для отдельных лиц и сообществ и способствовать ответственному использованию данных. |
| 6 | Collaboration and interdisciplinary approaches are important in big data research to integrate diverse perspectives, skills, and knowledge from various fields and stakeholders. | Сотрудничество и междисциплинарные подходы важны в исследовании больших данных для интеграции различных перспектив, навыков и знаний из разных областей и заинтересованных сторон. |
| 7 | Big data analysis has the potential to generate significant benefits, such as improved healthcare, environmental sustainability, and social justice, but also poses risks and challenges that require careful consideration and management. | Анализ больших данных имеет потенциал для генерации значительных преимуществ, таких как улучшение здравоохранения, экологической устойчивости и социальной справедливости, но также представляет риски и вызовы, требующие тщательного рассмотрения и управления. |
| 8 | As big data continues to shape society, it is essential to promote data literacy and education to enable individuals and communities to understand and use data effectively and responsibly. | Поскольку большие данные продолжают формировать общество, важно продвигать грамотность в области данных и образование, чтобы позволить отдельным лицам и сообществам понимать и использовать данные эффективно и ответственно. |
| 9 | The rapid evolution of big data technologies and applications requires continuous learning, adaptation, and innovation to keep pace with emerging trends and challenges. | Быстрое развитие технологий и приложений больших данных требует непрерывного обучения, адаптации и инноваций, чтобы не отставать от новейших тенденций и вызовов. |
| 10 | The development of ethical frameworks, guidelines, and regulations for big data analysis can help ensure that data is used in ways that align with societal values, principles, and goals, and promote trust and accountability in data-driven decision-making. | Разработка этических основ, руководящих принципов и регламентов для анализа больших данных может помочь гарантировать, что данные используются в соответствии с общественными ценностями, принципами и целями, а также способствовать доверию и подотчетности при принятии решений, основанных на данных. |

**Questions for answering**

1. What are the three Vs of big data, and how do they impact the analysis of large datasets?

Answer: The three Vs of big data are volume, velocity, and variety. They impact the analysis of large datasets by making it challenging to collect, process, and analyze data at scale and in real-time.

1. What are some commonly used methods and tools in big data analysis, and how do they differ from traditional data analysis approaches?

Answer: Some commonly used methods and tools in big data analysis include statistical modeling, machine learning, and data visualization. They differ from traditional data analysis approaches by requiring special considerations due to the scale and complexity of the data.

1. Why is data quality important in producing meaningful insights from big data, and what are some strategies for ensuring data quality?

Answer: Data quality is important in producing meaningful insights from big data because inaccuracies, inconsistencies, and incompleteness can lead to erroneous conclusions and decision-making. Some strategies for ensuring data quality include data validation, cleaning, normalization, and integration.

1. How does effective data management impact big data analysis, and what are some challenges associated with managing large volumes of data?

Answer: Effective data management is essential in big data analysis because it enables data to be collected, stored, retrieved, processed, and integrated from various sources and formats. Some challenges associated with managing large volumes of data include storage, security, and privacy concerns.

1. What are some ethical and social implications of big data analysis, and how can they be addressed?

Answer: Some ethical and social implications of big data analysis include privacy, security, and bias concerns. They can be addressed through the development of ethical frameworks, guidelines, and regulations, as well as promoting responsible use of data and promoting transparency in data-driven decision-making.

1. Why is collaboration and interdisciplinary approaches important in big data research, and what are some benefits of these approaches?

Answer: Collaboration and interdisciplinary approaches are important in big data research because they integrate diverse perspectives, skills, and knowledge from various fields and stakeholders. Some benefits of these approaches include improved innovation, problem-solving, and decision-making.

1. What are some potential benefits of big data analysis, and how can they be realized?

Answer: Some potential benefits of big data analysis include improved healthcare, environmental sustainability, and social justice. They can be realized through effective data management, ethical and responsible use of data, and collaboration across sectors and stakeholders.

1. How can individuals and communities become more data literate and use data effectively and responsibly?

Answer: Individuals and communities can become more data literate and use data effectively and responsibly through education, training, and access to tools and resources that promote data literacy, such as data visualization and interpretation tools.

1. How does the rapid evolution of big data technologies and applications impact the field of big data research, and what are some strategies for keeping up with emerging trends and challenges?

Answer: The rapid evolution of big data technologies and applications impacts the field of big data research by requiring continuous learning, adaptation, and innovation to keep pace with emerging trends and challenges. Some strategies for keeping up with these trends and challenges include ongoing training and education, collaboration and networking, and involvement in industry conferences and events.

1. What are some challenges associated with the development of ethical frameworks, guidelines, and regulations for big data analysis, and how can they be addressed?

Answer: Some challenges associated with the development of ethical frameworks, guidelines, and regulations for big data analysis include balancing the need for data privacy and security with the potential benefits of data sharing and use, as well as promoting transparency and accountability in data-driven decision-making. They can be addressed through ongoing dialogue

**Translation**

|  |  |
| --- | --- |
| **English** | **Russian** |
| **Beyond the hype: Big data concepts, methods and analytics** | **Погружение в концепции больших данных, их методов и аналитики** |
| **Abstract**  Size is the first, and at times, the only dimension that leaps out at the mention of big data. This paper attempts to offer a broader definition of big data that captures its other unique and defining characteristics. The rapid evolution and adoption of big data by industry has leapfrogged the discourse to popular outlets, forcing the academic press to catch up. Academic journals in numerous disciplines, which will benefit from a relevant discussion of big data, have yet to cover the topic. This paper presents a consolidated description of big data by integrating definitions from practitioners and academics. The paper's primary focus is on the analytic methods used for big data. A particular distinguishing feature of this paper is its focus on analytics related to unstructured data, which constitute 95% of big data. This paper highlights the need to develop appropriate and efficient analytical methods to leverage massive volumes of heterogeneous data in unstructured text, audio, and video formats. This paper also reinforces the need to devise new tools for predictive analytics for structured big data. The statistical methods in practice were devised to infer from sample data. The heterogeneity, noise, and the massive size of structured big data calls for developing computationally efficient algorithms that may avoid big data pitfalls, such as spurious correlation. | **Аннотация** Размер — это первое, а иногда и единственное измерение, которое бросается в глаза при упоминании больших данных. В данной статье предпринята попытка предложить более широкое определение больших данных, которое охватывает другие их уникальные и определяющие характеристики. Быстрое развитие и внедрение больших данных в промышленности привело к тому, что этот дискурс перекочевал в популярные издания, вынуждая академическую прессу догонять его. Академические журналы по многочисленным дисциплинам, которые выиграют от актуального обсуждения больших данных, еще не освещали эту тему. В данной статье представлено сводное описание больших данных путем интеграции определений, полученных от практиков и ученых. Основное внимание в статье уделено аналитическим методам, используемым для работы с большими данными. Отличительной особенностью данной статьи является ее фокус на аналитике, связанной с неструктурированными данными, которые составляют 95% больших данных. В статье подчеркивается необходимость разработки подходящих и эффективных аналитических методов для использования огромных объемов разнородных данных в форматах неструктурированного текста, аудио и видео. В работе также подчеркивается необходимость разработки новых инструментов предиктивной аналитики для структурированных больших данных. Существующие на практике статистические методы были разработаны для того, чтобы делать выводы на основе выборочных данных. Неоднородность, шум и огромный размер структурированных больших данных требуют разработки эффективных с вычислительной точки зрения алгоритмов, которые позволят избежать "подводных камней" больших данных, таких как ложная корреляция. |
| **Keywords:**  Big data analytics, Big data definition, Unstructured data analytics, Predictive analytics | **Ключевые слова:**  Аналитика больших данных, определение больших данных, неструктурированный анализ данных, предиктивная аналитика |
| **Introduction**  This paper documents the basic concepts relating to big data. It attempts to consolidate the hitherto fragmented discourse on what constitutes big data, what metrics define the size and other characteristics of big data, and what tools and technologies exist to harness the potential of big data.  From corporate leaders to municipal planners and academics, big data are the subject of attention, and to some extent, fear. The sudden rise of big data has left many unprepared. In the past, new technological developments first appeared in technical and academic publications. The knowledge and synthesis later seeped into other avenues of knowledge mobilization, including books. The fast evolution of big data technologies and the ready acceptance of the concept by public and private sectors left little time for the discourse to develop and mature in the academic domain. Authors and practitioners leapfrogged to books and other electronic media for immediate and wide circulation of their work on big data. Thus, one finds several books on big data, including Big Data for Dummies, but not enough fundamental discourse in academic publications.  The leapfrogging of the discourse on big data to more popular outlets implies that a coherent understanding of the concept and its nomenclature is yet to develop. For instance, there is little consensus around the fundamental question of how big the data has to be to qualify as ‘big data’. Thus, there exists the need to document in the academic press the evolution of big data concepts and technologies.  A key contribution of this paper is to bring forth the oft-neglected dimensions of big data. The popular discourse on big data, which is dominated and influenced by the marketing efforts of large software and hardware developers, focuses on predictive analytics and structured data. It ignores the largest component of big data, which is unstructured and is available as audio, images, video, and unstructured text. It is estimated that the analytics-ready structured data forms only a small subset of big data. The unstructured data, especially data in video format, is the largest component of big data that is only partially archived.  This paper is organized as follows. We begin the paper by defining big data. We highlight the fact that size is only one of several dimensions of big data. Other characteristics, such as the frequency with which data are generated, are equally important in defining big data. We then expand the discussion on various types of big data, namely text, audio, video, and social media. We apply the analytics lens to the discussion on big data. Hence, when we discuss data in video format, we focus on methods and tools to analyze data in video format.  Given that the discourse on big data is contextualized in predictive analytics frameworks, we discuss how analytics have captured the imaginations of business and government leaders and describe the state-of-practice of a rapidly evolving industry. We also highlight the perils of big data, such as spurious correlation, which have hitherto escaped serious inquiry. The discussion has remained focused on correlation, ignoring the more nuanced and involved discussion on causation. We conclude by highlighting the expected developments to realize in the near future in big [data analytics](https://www.sciencedirect.com/topics/social-sciences/data-analytics). | **Введение**  В этом документе изложены основные понятия, связанные с большими данными.  В нем предпринята попытка объединить до сих пор разрозненные рассуждения о том, что такое большие данные, какие метрики определяют размер и другие характеристики больших данных, и какие инструменты и технологии существуют для того, чтобы  использовать потенциал больших данных.  От руководителей корпораций до муниципальных планировщиков и ученых,  большие данные являются предметом внимания и, в некоторой степени, страха.  Внезапный рост объема больших данных заставил многих оказаться неподготовленными. В прошлом новые технологические разработки сначала появлялись в технических и  академических изданиях. Позже знания и синтез просачивались в другие способы мобилизации знаний, включая книги. Быстрая эволюция технологий больших данных и готовое принятие этой концепции государственным и частным секторами оставили мало времени для развития и становления дискурса в академической сфере.  Авторы и практики перепрыгнули на книги и другие электронные носители для немедленного и широкого распространения своих идей. СМИ для немедленного и широкого распространения своих работ по большим данным. Таким образом, можно найти несколько книг о больших данных, включая «Big Data для чайников», но недостаточно фундаментальных рассуждений в академических публикациях.  Переход дискурса о больших данных в более популярные подразумевает, что последовательное понимание концепции и номенклатуры еще не сформировалось. Например, не существует достаточного консенсус вокруг фундаментального вопроса о том, насколько большими должны быть данные, чтобы квалифицироваться как "большие данные". Таким образом, существует необходимость документировать в научной прессе эволюцию концепций и технологий больших данных.  Ключевой вклад данной работы заключается в том, чтобы привлечь внимание к часто  пренебрегаемые аспекты больших данных. Популярный дискурс о больших данных,  в котором доминируют и на который влияют маркетинговые усилия крупных  разработчиков программного и аппаратного обеспечения, сосредоточены на предиктивной аналитике и структурированных данных. При этом игнорируется самый большой компонент больших данных, который является неструктурированной и доступен в виде аудио, изображений, видео и неструктурированного текст. По оценкам, структурированные данные, пригодные для анализа, составляют лишь небольшую часть больших данных. Неструктурированные данные, особенно данные в формате видео, являются крупнейшим компонентом больших данных, которые архивируются лишь частично.  Данная статья организована следующим образом. В начале статьи мы даем определение  определения больших данных. Мы подчеркиваем тот факт, что размер является лишь одним из нескольких измерений больших данных. Другие характеристики, такие как частота, с которой генерируются данные, не менее важны для определения  больших данных. Затем мы расширяем обсуждение различных типов больших данных, а именно: текст, аудио, видео и социальные медиа.  данных, а именно: текст, аудио, видео и социальные сети. Мы применяем аналитический подход к обсуждению больших данных. Таким образом, когда мы обсуждаем  обсуждаем данные в видеоформате, мы фокусируемся на методах и инструментах для анализа данных в видеоформате.  Учитывая, что рассуждения о больших данных производятся в контексте предиктивной аналитики, мы обсуждаем, как аналитика захватила воображение лидеров бизнеса и правительства и описываем состояние практики в быстро развивающейся отрасли. Мы также подчеркиваем опасности больших данных, такие как ложная корреляция,  которые до сих пор оставались вне серьезного исследования. Дискуссия по-прежнему сосредоточена на корреляции, игнорируя более тонкие моменты и обсуждение причинно-следственных связей. В заключение мы подчеркиваем ожидаемые изменения, которые произойдут в ближайшем будущем в области больших данных и аналитики. |
| 2. Defining Big Data.  While it is ubiquitous today, however, ‘big data’ as a concept is nascent and has uncertain origins. Diebold (2012) argues that the term “big data … probably originated in lunch-table conversations at Silicon Graphics Inc. (SGI) in the mid-1990s, in which John Mashey figured prominently”. Despite the references to the mid-nineties, Fig. 1 shows that the term became widespread as recently as in 2011. The current hype can be attributed to the promotional initiatives by IBM and other leading technology companies who invested in building the niche analytics market.  Изображение выглядит как диаграмма  Автоматически созданное описание  Figure 1 Frequency distribution of documents containing the term "big data" in ProQuest Research Library  Big data definitions have evolved rapidly, which has raised some confusion. This is evident from an online survey of 154 C-suite global executives conducted by Harris Interactive on behalf of SAP in April 2012 (“Small and midsize companies look to make big gains with big data,” 2012). Fig. 2 shows how executives differed in their understanding of big data, where some definitions focused on what it is, while others tried to answer what it does.  Изображение выглядит как диаграмма  Автоматически созданное описание  Figure 2 Definitions of big data based on an online survey of 154 global executives in April 2012  Clearly, size is the first characteristic that comes to mind considering the question “what is big data?” However, other characteristics of big data have emerged recently. For instance, Laney (2001) suggested that Volume, Variety, and Velocity (or the Three V's) are the three dimensions of challenges in data management. The Three V's have emerged as a common framework to describe big data (Chen et al., 2012, [Kwon et al., 2014](https://www.sciencedirect.com/science/article/pii/S0268401214001066" \l "bib0105)). For example, Gartner, Inc. defines big data in similar terms:  “Big data is high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making.” (“Gartner IT Glossary, n.d.”)  Similarly, TechAmerica Foundation defines big data as follows:  “Big data is a term that describes large volumes of high velocity, complex and variable data that require advanced techniques and technologies to enable the capture, storage, distribution, management, and analysis of the information.” (TechAmerica Foundation's Federal Big Data Commission, 2012)  We describe the Three V's below.  Volume refers to the magnitude of data. Big data sizes are reported in multiple terabytes and petabytes. A survey conducted by IBM in mid-2012 revealed that just over half of the 1144 respondents considered datasets over one terabyte to be big data (Schroeck, Shockley, Smart, Romero-Morales, & Tufano, 2012). One terabyte stores as much data as would fit on 1500 CDs or 220 [DVDs](https://www.sciencedirect.com/topics/social-sciences/optical-discs), enough to store around 16 million Facebook photographs. Beaver, Kumar, Li, Sobel, and Vajgel (2010) report that Facebook processes up to one million photographs per second. One petabyte equals 1024 terabytes. Earlier estimates suggest that Facebook stored 260 billion photos using storage space of over 20 petabytes.  Definitions of big data volumes are relative and vary by factors, such as time and the type of data. What may be deemed big data today may not meet the threshold in the future because storage capacities will increase, allowing even bigger data sets to be captured. In addition, the type of data, discussed under variety , defines what is meant by ‘big’. Two datasets of the same size may require different data management technologies based on their type, e.g., tabular versus video data. Thus, definitions of big data also depend upon the industry. These considerations therefore make it impractical to define a specific threshold for big data volumes.  Variety refers to the structural heterogeneity in a dataset. Technological advances allow firms to use various types of structured, semi-structured, and unstructured data. Structured data, which constitutes only 5% of all existing data (Cukier, 2010), refers to the tabular data found in spreadsheets or relational databases. Text, images, audio, and video are examples of unstructured data, which sometimes lack the structural organization required by machines for analysis. Spanning a continuum between fully structured and unstructured data, the format of semi-structured data does not conform to strict standards. Extensible Markup Language (XML), a textual language for exchanging data on the Web, is a typical example of semi-structured data. XML documents contain user-defined data tags which make them machine-readable.  A high level of variety, a defining characteristic of big data, is not necessarily new. Organizations have been hoarding unstructured data from internal sources (e.g., sensor data) and external sources (e.g., social media). However, the emergence of new data management technologies and analytics, which enable organizations to leverage data in their business processes, is the innovative aspect. For instance, facial recognition technologies empower the brick-and-mortar retailers to acquire intelligence about store traffic, the age or gender composition of their customers, and their in-store movement patterns. This invaluable information is leveraged in decisions related to product promotions, placement, and staffing. Clickstream data provides a wealth of information about customer behavior and browsing patterns to online retailers. Clickstream advises on the timing and sequence of pages viewed by a customer. Using big [data analytics](https://www.sciencedirect.com/topics/social-sciences/data-analytics), even small and medium-sized enterprises (SMEs) can mine massive volumes of semi-structured data to improve website designs and implement effective cross-selling and personalized product recommendation systems.  Velocity refers to the rate at which data are generated and the speed at which it should be analyzed and acted upon. The proliferation of digital devices such as smartphones and sensors has led to an unprecedented rate of data creation and is driving a growing need for real-time analytics and evidence-based planning. Even conventional retailers are generating high-frequency data. Wal-Mart, for instance, processes more than one million transactions per hour (Cukier, 2010). The data emanating from mobile devices and flowing through mobile apps produces torrents of information that can be used to generate real-time, personalized offers for everyday customers. This data provides sound information about customers, such as geospatial location, demographics, and past buying patterns, which can be analyzed in real time to create real customer value.  Given the soaring popularity of smartphones, retailers will soon have to deal with hundreds of thousands of streaming data sources that demand real-time analytics. Traditional data management systems are not capable of handling huge data feeds instantaneously. This is where big data technologies come into play. They enable firms to create real-time intelligence from high volumes of ‘perishable’ data.  In addition to the three V's, other dimensions of big data have also been mentioned. These include:  Veracity. IBM coined Veracity as the fourth V, which represents the unreliability inherent in some sources of data. For example, customer sentiments in social media are uncertain in nature, since they entail human judgment. Yet they contain valuable information. Thus the need to deal with imprecise and uncertain data is another facet of big data, which is addressed using tools and analytics developed for management and mining of uncertain data.  Variability (and complexity). SAS introduced Variability and Complexity as two additional dimensions of big data. Variability refers to the variation in the data flow rates. Often, big data velocity is not consistent and has periodic peaks and troughs. Complexity refers to the fact that big data are generated through a myriad of sources. This imposes a critical challenge: the need to connect, match, cleanse and transform data received from different sources.  Value. Oracle introduced Value as a defining attribute of big data. Based on Oracle's definition, big data are often characterized by relatively “low value density”. That is, the data received in the original form usually has a low value relative to its volume. However, a high value can be obtained by analyzing large volumes of such data.  The relativity of big data volumes discussed earlier applies to all dimensions. Thus, universal benchmarks do not exist for volume, variety, and velocity that define big data. The defining limits depend upon the size, sector, and location of the firm and these limits evolve over time. Also important is the fact that these dimensions are not independent of each other. As one dimension changes, the likelihood increases that another dimension will also change as a result. However, a ‘three-V tipping point’ exists for every firm beyond which traditional data management and analysis technologies become inadequate for deriving timely intelligence. The Three-V tipping point is the threshold beyond which firms start dealing with big data. The firms should then trade-off the future value expected from big data technologies against their implementation costs. | 2. Определение больших данных  Несмотря на повсеместное распространение сегодня, "большие данные" как понятие  зарождается и имеет неопределенное происхождение. Диболд (2012) утверждает, что  термин "большие данные . ... вероятно, возник в разговорах за обеденным столом в Silicon Graphics Inc.(SGI) в середине 1990-х годов, в которых Джон Мэши занимал видное место". Несмотря на ссылки на середину девяностых, рис. 1 показывает, что термин получил широкое распространение совсем недавно в 2011 году. Нынешний ажиотаж можно объяснить рекламными инициативам IBM и других ведущих технологических компаний, которые инвестировали в создание нишевого рынка аналитики.  Изображение выглядит как диаграмма  Автоматически созданное описание  Рисунок Частотное распределение документов, содержащих термин "большие данные" в ProQuest Research Library  Определения больших данных быстро эволюционировали, что вызвало некоторую  путаницу. Об этом свидетельствуют результаты онлайн-опроса 154 руководителей высшего звена, проведенного компанией Harris Interris Interris, проведенного компанией Harris Interactive по заказу SAP в апреле 2012 года ("Малые и средние компании стремятся добиться больших успехов с помощью больших данных", 2012). На рис. 2 показано, как различаются руководители в своем понимании больших данных, где одни определения фокусировались на том, что это такое, а другие пытались ответить на вопрос что они делают.  Изображение выглядит как диаграмма  Автоматически созданное описание  Рисунок Различные определения больших данных, основанные на онлайн-опросе 154 руководителей компаний мира в апреле 2012 года  Очевидно, что размер — это первая характеристика, которая приходит на ум  при рассмотрении вопроса "что такое большие данные?". Однако существуют и другие характеристики больших данных появились недавно. Например, Лейни (2001) предположил, что Объем, Разнообразие и Скорость (или Три V) являются тремя измерениями проблем в управлении данными.  Три "V" стали общей основой для описания больших данных (Chen, Chiang, & Storey, 2012; Kwon, Lee, & Shin, 2014). Например, компания Gartner, Inc. определяет большие данные в похожих терминах:  "Большие данные — это большие объемы, высокая скорость и большое разнообразие информационные активы, которые требуют экономически эффективных, инновационных форм обработки информации для улучшения понимания и принятия решений". ("Gartner IT Glossary, n.d.").  Аналогично, Фонд TechAmerica определяет большие данные следующим образом:  "Большие данные — это термин, который описывает большие объемы высокоскоростных, сложных и изменчивых данных, которые требуют передовых методов и технологий для обеспечения сбора, хранения, распределения, управления и анализа информации". (TechAmerica Федеральная комиссия по большим данным Фонда TechAmerica, 2012).  Ниже мы приводим описание трех "V".  *Объем* относится к величине данных. Размеры больших данных измеряются терабайтами и петабайтами. Исследование, проведенное компанией IBM в середине 2012 года, показал, что чуть более половины из 1144 респондентов считают большими данными наборы данных объемом более одного терабайта. (Schroeck, Shockley, Smart, Romero-Morales, & Tufano, 2012). Один терабайт хранит столько данных, сколько поместилось бы на 1500 CD или 220 DVD, его достаточно для хранения около 16 миллионов фотографий Facebook. Бобер, Kumar, Li, Sobel и Vajgel (2010) сообщают, что Facebook обрабатывает  обрабатывает до миллиона фотографий в секунду. Один петабайт равен 1024 терабайтам. По более ранним оценкам, Facebook хранит 260 миллиардов фотографий, используя пространство для хранения более 20 петабайт. Определения объемов больших данных являются относительными и зависят от таких факторов, как время и тип данных. То, что может считаться большими данными сегодня, может не соответствовать порогу в будущем, поскольку объемы хранения данных будут увеличиваться, что позволит собирать еще большие массивы данных быть захваченными. Кроме того, тип данных, рассмотренный в разделе "Разнообразие", определяет, что подразумевается под большими данными. Эти данные определяют, что подразумевается под "большими". Два набора данных одинакового размера могут потребовать различных технологий управления данными в зависимости от их типа, например, табличные или видеоданные. Таким образом, определения большие данные также зависят от отрасли. Эти соображения делают нецелесообразным определение конкретного порога для больших данных.  Разнообразие относится к структурной неоднородности набора данных. Технологические достижения позволяют фирмам использовать различные типы структурированных, полу структурированные и неструктурированные данные. Структурированные данные, которые составляют лишь 5% всех существующих данных (Cukier, 2010), относятся к  табличные данные, содержащиеся в электронных таблицах или реляционных базах данных.  Текст, изображения, аудио и видео являются примерами неструктурированных данных, которые иногда не имеют структурной организации данных, необходимой  машинам для анализа. Охватывая континуум между полностью структурированными и неструктурированными данными, формат полу структурированных данных не соответствует строгим стандартам. Язык расширяемой разметки (XML), текстовый язык для обмена данными в Интернете, является типичным примером полу структурированных данных. XML документы содержат определяемые пользователем теги данных, которые делают их машиночитаемыми. Высокий уровень разнообразия, являющийся определяющей характеристикой больших данных, не обязательно является чем-то новым. Организации накапливали неструктурированные данные из внутренних источников (например, данные датчиков) и внешних источников (например, данных датчиков) и внешних источников (например, социальных сетей). Однако появление новых технологий управления данными  Однако появление новых технологий управления данными и аналитики, которые позволяют организациям использовать данные в своих бизнес-процессах, является инновационным аспектом. Например, технологии распознавания лиц позволяют  например, технологии распознавания лиц позволяют розничным торговцам получать информацию о посещаемости магазинов, возрастном и гендерном составе покупателей и  их передвижения по магазину. Эта бесценная информация используется при принятии решений, связанных с продвижением товаров, размещением,  и подбора персонала. Данные о потоке кликов предоставляют богатую информацию  о поведении покупателей и шаблонах просмотра интернет-магазинам.  компаниям. Clickstream дает рекомендации по времени и последовательности страниц.  просматриваемых покупателем. Используя аналитику больших данных, даже малые и  средние предприятия (МСП) могут добывать огромные объемы полу структурированных данных для улучшения дизайна веб-сайтов и внедрения эффективных перекрестных продаж и персонализированных рекомендаций продуктов системы.  Скорость относится к скорости, с которой генерируются данные, и к скорости, с которой они должны анализироваться. Распространение цифровых устройств, таких как смартфоны и датчики, привело к беспрецедентной скорости создания данных.  беспрецедентным темпам создания данных и вызывает растущую потребность  в анализе в режиме реального времени и планировании на основе фактических данных. Даже традиционные розничные компании генерируют данные с высокой частотой. WalMart, например, обрабатывает более одного миллиона транзакций в час (Cukier, 2010). Данные, поступающие с мобильных устройств и проходящие через мобильные приложения, создают потоки информации, которую можно использовать для создания персонализированных предложений в реальном времени для повседневных пользователей. Эти данные содержат достоверную информацию о клиентах,  например, местоположение, демографические характеристики и прошлые покупательские паттерны, которые могут быть проанализированы в режиме реального времени для создания реальной потребительской ценности.  Учитывая стремительный рост популярности смартфонов, в скором времени розничные торговцы будут иметь дело с сотнями тысяч потоковых источников данных.  которые требуют анализа в режиме реального времени. Традиционные системы управления данными не способны мгновенно обрабатывать огромные потоки данных.  Именно здесь в игру вступают технологии больших данных. Они позволяют  фирмам создавать интеллектуальные данные в реальном времени из больших объемов "скоропортящихся" данных.  В дополнение к трем «V» были упомянуты и другие аспекты больших данных. К ним относятся:  *Правдивость*. IBM назвала правдивость четвертым V, которое представляет собой  ненадежность, присущую некоторым источникам данных. Например, настроения клиентов, настроения пользователей в социальных сетях являются неопределенными по своей природе, поскольку они связаны с человеческими суждениями. Тем не менее, они содержат ценную информацию. Таким образом, необходимость работы с неточными и неопределенными данными является еще одним аспектом больших данных, который решается с помощью инструментов и аналитики, разработанные для управления и добычи неопределенных данных.  *Изменчивость* (и сложность). SAS ввел понятия изменчивости и сложности как два дополнительных измерения больших данных. Изменчивость относится  к вариациям в скорости потока данных. Часто скорость больших данных  не является постоянной и имеет периодические пики и спады. Сложность  относится к тому факту, что большие данные генерируются из огромного количества  источников. Это создает серьезную проблему: необходимость соединять, сопоставлять, очищать и преобразовывать данные, полученные из различных источников.  источников.  *Ценность*. Компания Oracle ввела понятие ценности как определяющий атрибут больших данных. Согласно определению Oracle, большие данные часто характеризуются  относительно "низкой плотностью ценности". То есть, данные, полученные в  исходном виде, обычно имеют низкую ценность по отношению к своему объему. Как-  однако, высокая ценность может быть получена при анализе больших объемов  таких данных. Относительность объемов больших данных, о которой говорилось ранее, применима ко всем измерениям. Таким образом, универсальных эталонов для объема, разнообразия и скорости, определяющих большие данные, не существует. Определяющие пределы зависят от размера, сектора и местоположения компании, и эти пределы меняются с течением времени. Также важен тот факт, что эти параметры не зависят друг от друга. При изменении одного измерения увеличивается вероятность того, что другое измерение также изменится в результате этого. Однако для каждой компании существует "переломный момент с тремя V", после которого традиционные технологии управления и анализа данных становятся неадекватными для управления и анализа данных. Технологии становятся неадекватными для получения своевременной оперативной информации. Три-V — это порог, после которого компании начинают работать с большими данными. Затем компании должны сопоставить будущую ценность, ожидаемую от технологий больших данных, с затратами на их внедрение. |
| 3. Big Data Analytics  Big data are worthless in a vacuum. Its potential value is unlocked only when leveraged to drive decision making. To enable such evidence-based decision making, organizations need efficient processes to turn high volumes of fast-moving and diverse data into meaningful insights. The overall process of extracting insights from big data can be broken down into five stages (Labrinidis & Jagadish, 2012), shown in [Fig. 3](https://www.sciencedirect.com/science/article/pii/S0268401214001066" \l "fig0015). These five stages form the two main sub-processes: data management and analytics. Data management involves processes and supporting technologies to acquire and store data and to prepare and retrieve it for analysis. Analytics, on the other hand, refers to techniques used to analyze and acquire intelligence from big data. Thus, big data analytics can be viewed as a sub-process in the overall process of ‘insight extraction’ from big data.  Изображение выглядит как текст  Автоматически созданное описание  Figure 3 Process for extracting insights from big data  In the following sections, we briefly review big data analytical techniques for structured and unstructured data. Given the breadth of the techniques, an exhaustive list of techniques is beyond the scope of a single paper. Thus, the following techniques represent a relevant subset of the tools available for big data analytics.  3.1 Text analytics  Text analytics (text mining) refers to techniques that extract information from textual data. Social network feeds, emails, blogs, online forums, survey responses, corporate documents, news, and call center logs are examples of textual data held by organizations. Text analytics involve statistical analysis, [computational linguistics](https://www.sciencedirect.com/topics/social-sciences/computational-linguistics), and machine learning. Text analytics enable businesses to convert large volumes of human generated text into meaningful summaries, which support evidence-based decision-making. For instance, text analytics can be used to predict stock market based on information extracted from financial news (Chung, 2014). We present a brief review of text analytics methods below.  Information extraction (IE) techniques extract structured data from unstructured text. For example, IE algorithms can extract structured information such as drug name, dosage, and frequency from medical prescriptions. Two sub-tasks in IE are Entity Recognition (ER) and Relation Extraction (RE) (Jiang, 2012). ER finds names in text and classifies them into predefined categories such as person, date, location, and organization. RE finds and extracts semantic relationships between entities (e.g., persons, organizations, drugs, genes, etc.) in the text. For example, given the sentence “Steve Jobs co-founded Apple Inc. in 1976”, an RE system can extract relations such as FounderOf [Steve Jobs, Apple Inc.] or FoundedIn [Apple Inc., 1976].  Text summarization techniques automatically produce a succinct summary of a single or multiple documents. The resulting summary conveys the key information in the original text(s). Applications include scientific and news articles, advertisements, emails, and blogs. Broadly speaking, summarization follows two approaches: the extractive approach and the abstractive approach. In extractive summarization, a summary is created from the original text units (usually sentences). The resulting summary is a subset of the original document. Based on the extractive approach, formulating a summary involves determining the salient units of a text and stringing them together. The importance of the text units is evaluated by analyzing their location and frequency in the text. Extractive summarization techniques do not require an ‘understanding’ of the text. In contrast, abstractive summarization techniques involve extracting semantic information from the text. The summaries contain text units that are not necessarily present in the original text. In order to parse the original text and generate the summary, abstractive summarization incorporates advanced [Natural Language Processing](https://www.sciencedirect.com/topics/social-sciences/natural-language-processing) (NLP) techniques. As a result, abstractive systems tend to generate more coherent summaries than the extractive systems do (Hahn & Mani, 2000). However, extractive systems are easier to adopt, especially for big data.  Question answering (QA) techniques provide answers to questions posed in natural language. Apple's Siri and IBM's Watson are examples of commercial QA systems. These systems have been implemented in healthcare, finance, marketing, and education. Similar to abstractive summarization, QA systems rely on complex NLP techniques. QA techniques are further classified into three categories: the information retrieval (IR)-based approach, the knowledge-based approach, and the hybrid approach. IR-based QA systems often have three sub-components. First is the question processing, used to determine details, such as the question type, question focus, and the answer type, which are used to create a query. Second is document processing which is used to retrieve relevant pre-written passages from a set of existing documents using the query formulated in question processing. Third is answer processing, used to extract candidate answers from the output of the previous component, rank them, and return the highest-ranked candidate as the output of the QA system. Knowledge-based QA systems generate a semantic description of the question, which is then used to query structured resources. The Knowledge-based QA systems are particularly useful for restricted domains, such as tourism, medicine, and transportation, where large volumes of pre-written documents do not exist. Such domains lack data redundancy, which is required for IR-based QA systems. Apple's Siri is an example of a QA system that exploits the knowledge-based approach. In hybrid QA systems, like IBM's Watson, while the question is semantically analyzed, candidate answers are generated using the IR methods.  Sentiment analysis (opinion mining) techniques analyze opinionated text, which contains people's opinions toward entities such as products, organizations, individuals, and events. Businesses are increasingly capturing more data about their customers’ sentiments that has led to the proliferation of sentiment analysis (Liu, 2012). Marketing, finance, and the political and social sciences are the major application areas of sentiment analysis. Sentiment analysis techniques are further divided into three sub-groups, namely document-level, sentence-level, and aspect-based. Document-level techniques determine whether the whole document expresses a negative or a positive sentiment. The assumption is that the document contains sentiments about a single entity. While certain techniques categorize a document into two classes, negative and positive, others incorporate more sentiment classes (like the Amazon's five-star system) (Feldman, 2013). Sentence-level techniques attempt to determine the polarity of a single sentiment about a known entity expressed in a single sentence. Sentence-level techniques must first distinguish subjective sentences from objective ones. Hence, sentence-level techniques tend to be more complex compared to document-level techniques. Aspect-based techniques recognize all sentiments within a document and identify the aspects of the entity to which each sentiment refers. For instance, customer product reviews usually contain opinions about different aspects (or features) of a product. Using aspect-based techniques, the vendor can obtain valuable information about different features of the product that would be missed if the sentiment is only classified in terms of polarity.  3.2 Audio analytics  Audio analytics analyze and extract information from unstructured audio data. When applied to human spoken language, audio analytics is also referred to as [*speech*](https://www.sciencedirect.com/topics/social-sciences/logopedics)analytics. Since these techniques have mostly been applied to spoken audio, the terms audio analytics and speech analytics are often used interchangeably. Currently, customer call centers and healthcare are the primary application areas of audio analytics.  Call centers use audio analytics for efficient analysis of thousands or even millions of hours of recorded calls. These techniques help improve customer experience, evaluate agent performance, enhance sales turnover rates, monitor compliance with different policies (e.g., privacy and security policies), gain insight into customer behavior, and identify product or service issues, among many other tasks. Audio analytics systems can be designed to analyze a live call, formulate cross/up-selling recommendations based on the customer's past and present interactions, and provide feedback to agents in real time. In addition, automated call centers use the Interactive Voice Response (IVR) platforms to identify and handle frustrated callers.  In healthcare, audio analytics support diagnosis and treatment of certain medical conditions that affect the patient's communication patterns (e.g., depression, schizophrenia, and cancer) (Hirschberg, Hjalmarsson, & Elhadad, 2010). Also, audio analytics can help analyze an infant's cries, which contain information about the infant's health and emotional status (Patil, 2010). The vast amount of data recorded through speech-driven clinical documentation systems is another driver for the adoption of audio analytics in healthcare.  Speech analytics follows two common technological approaches: the transcript-based approach (widely known as large-vocabulary continuous speech recognition, LVCSR) and the phonetic-based approach. These are explained below.  LVCSR systems follow a two-phase process: indexing and searching. In the first phase, they attempt to transcribe the speech content of the audio. This is performed using automatic speech recognition (ASR) algorithms that match sounds to words. The words are identified based on a predefined dictionary. If the system fails to find the exact word in the dictionary, it returns the most similar one. The output of the system is a searchable index file that contains information about the sequence of the words spoken in the speech. In the second phase, standard text-based methods are used to find the search term in the index file.  Phonetic-based systems work with sounds or [*phonemes*](https://www.sciencedirect.com/topics/social-sciences/phoneme). Phonemes are the perceptually distinct units of sound in a specified language that distinguish one word from another (e.g., the phonemes/k/and/b/differentiate the meanings of “cat” and “bat”). Phonetic-based systems also consist of two phases: phonetic indexing and searching. In the first phase, the system translates the input speech into a sequence of phonemes. This is in contrast to LVCSR systems where the speech is converted into a sequence of words. In the second phase, the system searches the output of the first phase for the phonetic representation of the search terms.  3.3 Video Analytics  Video analytics, also known as video [content analysis](https://www.sciencedirect.com/topics/social-sciences/program-content) (VCA), involves a variety of techniques to monitor, analyze, and extract meaningful information from video streams. Although video analytics is still in its infancy compared to other types of data mining (Panigrahi, Abraham, & Das, 2010), various techniques have already been developed for processing real-time as well as pre-recorded videos. The increasing prevalence of closed-circuit television (CCTV) cameras and the booming popularity of video-sharing websites are the two leading contributors to the growth of computerized video analysis. A key challenge, however, is the sheer size of video data. To put this into perspective, one second of a high-definition video, in terms of size, is equivalent to over 2000 pages of text (Manyika et al., 2011). Now consider that 100 hours of video are uploaded to YouTube every minute (YouTube Statistics, n.d.).  Big data technologies turn this challenge into opportunity. Obviating the need for cost-intensive and risk-prone manual processing, big data technologies can be leveraged to automatically sift through and draw intelligence from thousands of hours of video. As a result, the big data technology is the third factor that has contributed to the development of video analytics.  The primary application of video analytics in recent years has been in automated security and surveillance systems. In addition to their high cost, labor-based surveillance systems tend to be less effective than automatic systems (e.g., Hakeem et al., 2012 report that security personnel cannot remain focused on surveillance tasks for more than 20 minutes). Video analytics can efficiently and effectively perform surveillance functions such as detecting breaches of restricted zones, identifying objects removed or left unattended, detecting loitering in a specific area, recognizing suspicious activities, and detecting camera tampering, to name a few. Upon detection of a threat, the surveillance system may notify security personnel in real time or trigger an automatic action (e.g., sound alarm, lock doors, or turn on lights).  The data generated by CCTV cameras in retail outlets can be extracted for business intelligence. Marketing and operations management are the primary application areas. For instance, smart algorithms can collect demographic information about customers, such as age, gender, and ethnicity. Similarly, retailers can count the number of customers, measure the time they stay in the store, detect their movement patterns, measure their dwell time in different areas, and monitor queues in real time. Valuable insights can be obtained by correlating this information with customer demographics to drive decisions for product placement, price, assortment optimization, promotion design, cross-selling, layout optimization, and staffing.  Another potential application of video analytics in retail lies in the study of buying behavior of groups. Among family members who shop together, only one interacts with the store at the cash register, causing the traditional systems to miss data on buying patterns of other members. Video analytics can help retailers address this missed opportunity by providing information about the size of the group, the group's demographics, and the individual members’ buying behavior.  Automatic video indexing and retrieval constitutes another domain of video analytics applications. The widespread emergence of online and offline videos has highlighted the need to index multimedia content for easy search and retrieval. The indexing of a video can be performed based on different levels of information available in a video including the metadata, the soundtrack, the transcripts, and the visual content of the video. In the metadata-based approach, relational database management systems (RDBMS) are used for video search and retrieval. Audio analytics and text analytics techniques can be applied to index a video based on the associated soundtracks and transcripts, respectively. A comprehensive review of approaches and techniques for video indexing is presented in Hu, Xie, Li, Zeng, and Maybank (2011).  In terms of the system architecture, there exist two approaches to video analytics, namely server-based and edge-based:  Server-based architecture. In this configuration, the video captured through each camera is routed back to a centralized and dedicated server that performs the video analytics. Due to bandwidth limits, the video generated by the source is usually compressed by reducing the frame rates and/or the image resolution. The resulting loss of information can affect the accuracy of the analysis. However, the server-based approach provides [economies of scale](https://www.sciencedirect.com/topics/social-sciences/economies-of-scale) and facilitates easier maintenance.  Edge-based architecture. In this approach, analytics are applied at the ‘edge’ of the system. That is, the video analytics is performed locally and on the raw data captured by the camera. As a result, the entire content of the video stream is available for the analysis, enabling a more effective content analysis. Edge-based systems, however, are more costly to maintain and have a lower processing power compared to the server-based systems.  3.4 Social media analytics  Social media analytics refer to the analysis of structured and unstructured data from social media channels. Social media is a broad term encompassing a variety of online platforms that allow users to create and exchange content. Social media can be categorized into the following types: Social networks (e.g., Facebook and LinkedIn), blogs (e.g., Blogger and WordPress), microblogs (e.g., Twitter and Tumblr), social news (e.g., Digg and Reddit), social bookmarking (e.g., Delicious and StumbleUpon), media sharing (e.g., Instagram and YouTube), wikis (e.g., Wikipedia and Wikihow), question-and-answer sites (e.g., Yahoo! Answers and Ask.com) and review sites (e.g., Yelp, TripAdvisor) ([Barbier and Liu, 2011](https://www.sciencedirect.com/science/article/pii/S0268401214001066" \l "bib0010), Gundecha and Liu, 2012). Also, many mobile apps, such as Find My Friend, provide a platform for social interactions and, hence, serve as social media channels.  Although the research on social networks dates back to early 1920s, nevertheless, social media analytics is a nascent field that has emerged after the advent of Web 2.0 in the early 2000s. The key characteristic of the modern social media analytics is its data-centric nature. The research on social media analytics spans across several disciplines, including psychology, sociology, anthropology, computer science, mathematics, physics, and economics. Marketing has been the primary application of social media analytics in recent years. This can be attributed to the widespread and growing adoption of social media by consumers worldwide (He, Zha, & Li, 2013), to the extent that Forrester Research, Inc., projects social media to be the second-fastest growing marketing channel in the US between 2011 and 2016 (VanBoskirk, Overby, & Takvorian, 2011).  User-generated content (e.g., sentiments, images, videos, and bookmarks) and the relationships and interactions between the network entities (e.g., people, organizations, and products) are the two sources of information in social media. Based on this categorization, the social media analytics can be classified into two groups:  Content-based analytics. Content-based analytics focuses on the data posted by users on social media platforms, such as customer feedback, product reviews, images, and videos. Such content on social media is often voluminous, unstructured, noisy, and dynamic. Text, audio, and video analytics, as discussed earlier, can be applied to derive insight from such data. Also, big data technologies can be adopted to address the data processing challenges.  Structure-based analytics. Also referred to as social network analytics, this type of analytics are concerned with synthesizing the structural attributes of a social network and extracting intelligence from the relationships among the participating entities. The structure of a social network is modeled through a set of nodes and edges, representing participants and relationships, respectively. The model can be visualized as a graph composed of the nodes and the edges. We review two types of network graphs, namely social graphs and activity graphs (Heidemann, Klier, & Probst, 2012). In social graphs, an edge between a pair of nodes only signifies the existence of a link (e.g., friendship) between the corresponding entities. Such graphs can be mined to identify communities or determine hubs (i.e., the users with a relatively large number of direct and indirect social links). In activity networks, however, the edges represent actual interactions between any pair of nodes. The interactions involve exchanges of information (e.g., likes and comments). Activity graphs are preferable to social graphs, because an active relationship is more relevant to analysis than a mere connection.  Various techniques have recently emerged to extract information from the structure of social networks. We briefly discuss these below.  Community detection, also referred to as community discovery, extracts implicit communities within a network. For online social networks, a community refers to a sub-network of users who interact more extensively with each other than with the rest of the network. Often containing millions of nodes and edges, online social networks tend to be colossal in size. Community detection helps to summarize huge networks, which then facilitates uncovering existing behavioral patterns and predicting emergent properties of the network. In this regard, community detection is similar to clustering (Aggarwal, 2011), a data mining technique used to partition a data set into disjoint subsets based on the similarity of data points. Community detection has found several application areas, including marketing and the World Wide Web (Parthasarathy, Ruan, & Satuluri, 2011). For example, community detection enables firms to develop more effective product recommendation systems.  Social influence analysis refers to techniques that are concerned with modeling and evaluating the influence of actors and connections in a social network. Naturally, the behavior of an actor in a social network is affected by others. Thus, it is desirable to evaluate the participants’ influence, quantify the strength of connections, and uncover the patterns of influence diffusion in a network. Social influence analysis techniques can be leveraged in viral marketing to efficiently enhance brand awareness and adoption.  A salient aspect of social influence analysis is to quantify the importance of the network nodes. Various measures have been developed for this purpose, including degree centrality, betweenness centrality, closeness centrality, and eigenvector centrality (for more details refer to Tang & Liu, 2010). Other measures evaluate the strength of connections represented by edges or model the spread of influence in social networks. The Linear Threshold Model (LTM) and Independent Cascade Model (ICM) are two well-known examples of such frameworks (Sun & Tang, 2011).  Link prediction specifically addresses the problem of predicting future linkages between the existing nodes in the underlying network. Typically, the structure of social networks is not static and continuously grows through the creation of new nodes and edges. Therefore, a natural goal is to understand and predict the dynamics of the network. Link prediction techniques predict the occurrence of interaction, collaboration, or influence among entities of a network in a specific time interval. Link prediction techniques outperform pure chance by factors of 40–50, suggesting that the current structure of the network surely contains latent information about future links (Liben-Nowell & Kleinberg, 2003).  In biology, link prediction techniques are used to discover links or associations in biological networks (e.g., protein–protein interaction networks), eliminating the need for expensive experiments (Hasan & Zaki, 2011). In security, link prediction helps to uncover potential collaborations in terrorist or criminal networks. In the context of online social media, the primary application of link prediction is in the development of recommendation systems, such as Facebook's “People You May Know”, YouTube's “Recommended for You”, and Netflix's and Amazon's recommender engines.  3.5 Predictive analytics  Predictive analytics comprise a variety of techniques that predict future outcomes based on historical and current data. In practice, predictive analytics can be applied to almost all disciplines – from predicting the failure of jet engines based on the stream of data from several thousand sensors, to predicting customers’ next moves based on what they buy, when they buy, and even what they say on social media.  At its core, predictive analytics seek to uncover patterns and capture relationships in data. Predictive analytics techniques are subdivided into two groups. Some techniques, such as moving averages, attempt to discover the historical patterns in the outcome variable(s) and extrapolate them to the future. Others, such as linear regression, aim to capture the [interdependencies](https://www.sciencedirect.com/topics/social-sciences/interdependencies) between outcome variable(s) and explanatory variables, and exploit them to make predictions. Based on the underlying methodology, techniques can also be categorized into two groups: regression techniques (e.g., multinomial logit models) and machine learning techniques (e.g., neural networks). Another classification is based on the type of outcome variables: techniques such as linear regression address continuous outcome variables (e.g., sale price of houses), while others such as Random Forests are applied to discrete outcome variables (e.g., credit status).  Predictive analytics techniques are primarily based on statistical methods. Several factors call for developing new statistical methods for big data. First, conventional statistical methods are rooted in statistical significance: a small sample is obtained from the population and the result is compared with chance to examine the significance of a particular relationship. The conclusion is then generalized to the entire population. In contrast, big data samples are massive and represent the majority of, if not the entire, population. As a result, the notion of statistical significance is not that relevant to big data. Secondly, in terms of computational efficiency, many conventional methods for small samples do not scale up to big data. The third factor corresponds to the distinctive features inherent in big data: heterogeneity, noise accumulation, spurious correlations, and incidental endogeneity (Fan, Han, & Liu, 2014). We describe these below.  Heterogeneity. Big data are often obtained from different sources and represent information from different sub-populations. As a result, big data are highly heterogeneous. The sub-population data in small samples are deemed outliers because of their insufficient frequency. However, the sheer size of big data sets creates the unique opportunity to model the heterogeneity arising from sub-population data, which would require sophisticated statistical techniques.  Noise accumulation. Estimating predictive models for big data often involves the simultaneous estimation of several parameters. The accumulated estimation error (or noise) for different parameters could dominate the magnitudes of variables that have true effects within the model. In other words, some variables with significant explanatory power might be overlooked as a result of noise accumulation.  Spurious correlation. For big data, spurious correlation refers to uncorrelated variables being falsely found to be correlated due to the massive size of the dataset. Fan and Lv (2008) show this phenomenon through a simulation example, where the [correlation coefficient](https://www.sciencedirect.com/topics/social-sciences/correlation-coefficient) between independent random variables is shown to increase with the size of the dataset. As a result, some variables that are scientifically unrelated (due to their independence) are erroneously proven to be correlated as a result of high dimensionality.  Incidental endogeneity. A common assumption in regression analysis is the exogeneity assumption: the explanatory variables, or predictors, are independent of the residual term. The validity of most statistical methods used in regression analysis depends on this assumption. In other words, the existence of incidental endogeneity (i.e., the dependence of the residual term on some of the predictors) undermines the validity of the statistical methods used for regression analysis. Although the exogeneity assumption is usually met in small samples, incidental endogeneity is commonly present in big data. It is worthwhile to mention that, in contrast to spurious correlation, incidental endogeneity refers to a genuine relationship between variables and the error term.  The irrelevance of statistical significance, the challenges of computational efficiency, and the unique characteristics of big data discussed above highlight the need to develop new statistical techniques to gain insights from predictive models. | **3. Аналитика больших данных**  Большие данные ничего не стоят в вакууме. Их потенциальная ценность раскрывается только тогда, когда они используются для принятия решений. Чтобы обеспечить принятие решений на основе фактических данных, организациям необходимы эффективные процессы для превращения больших объемов быстро меняющихся и разнообразных данных в содержательные выводы. Общий процесс извлечения информации из больших данных можно разбить на пять этапов (Labrinidis &  Jagadish, 2012), показанные на рис. 3. Эти пять этапов образуют два основных подпроцесса: управление данными и аналитика. Управление данными включает процессы и вспомогательные технологии для получения  и хранения данных, а также подготовки и извлечения их для анализа. Аналитика, с другой стороны, относится к методам, используемым для анализа и получения из больших данных. Таким образом, аналитику больших данных можно рассматривать  как подпроцесс в общем процессе "извлечения информации" из больших данных.  больших данных.  Изображение выглядит как текст  Автоматически созданное описание  Рисунок Процессы извлечения информации из больших данных  В следующих разделах мы кратко рассмотрим методы анализа больших данных для структурированных и неструктурированных данных. Учитывая широту спектра методов, исчерпывающий список методов не укладывается в рамки одной статьи. Таким образом, следующие техники представляют собой соответствующее подмножество инструментов, доступных для больших данных и аналитики.  3.1. Текстовая аналитика  Текстовая аналитика (text mining) относится к методам, которые извлекают информацию из текстовых данных. Сообщения в социальных сетях, электронные письма, блоги, онлайн-форумы, ответы на опросы, корпоративные документы, новости, и журналы колл-центров являются примерами текстовых данных, хранящихся в организациях. Текстовая аналитика включает в себя статистический анализ, вычислительную лингвистику и машинное обучение. Текстовая аналитика позволяет компаниям преобразовывать большие объемы сгенерированного человеком текста в осмысленные резюме, которые помогают принимать решения на основе фактических данных. Например, текстовая аналитика может быть использована для прогнозирования фондового рынка на основе информации, извлеченной из финансовых новостей (Chung, 2014). Ниже мы приводим краткий обзор методов анализа текста. Методы извлечения информации (ИИ) извлекают структурированные данные  из неструктурированного текста. Например, алгоритмы ИЭ могут извлекать структурированную информацию, такую как название лекарства, дозировка и частота приема из медицинских рецептов. Двумя подзадачами в ИЭ являются распознавание сущностей (Entity Recognize (ER) и извлечение связей (RE) (Jiang, 2012). ER находит имена  в тексте и классифицирует их по заранее определенным категориям, таким как пер-.  сын, дата, местоположение и организация. RE находит и извлекает семантические  отношения между сущностями (например, людьми, организациями, лекарствами,  генами и т.д.) в тексте. Например, в предложении «Стив Джобс стал одним из основателей Apple Inc. в 1976 году», система RE может извлечь такие отношения  такие как FounderOf [Steve Jobs, Apple Inc.] или FoundedIn [Apple  Inc., 1976].  Методы сжатия текста автоматически производят четкое резюме одного или нескольких документов. Полученное резюме передает ключевую информацию, содержащуюся в исходном тексте (текстах).  Приложения включают научные и новостные статьи, рекламные объявления, электронные письма и блоги. В широком смысле сжатие основывается на двух подхода: экстрактивный подход и абстрактный подход. При экстрактивном обобщении резюме создается из исходных единиц текста (обычно предложений). Полученное резюме представляет собой подмножество исходного документа. На основе экстрактивного подхода составление резюме включает в себя определение наиболее значимых частей текста и объединение их в единое целое. Важность единиц текста оценивается путем анализа их расположения и частоты встречаемости в тексте. Методы экстрактивного обобщения не требуют “понимания” текста.  В отличие от этого, абстрактное обобщение включает извлечение семантической информации из текста. Резюме содержит текстовые единицы, которые не обязательно присутствуют в оригинальном тексте. Для того чтобы разобрать исходный текст и создать  резюме, абстрактное обобщение включает в себя продвинутые методы обработки естественного языка (NLP). В результате абстрактные системы, как правило, генерируют более связные резюме, чем экстрактивные системы (Hahn & Mani, 2000). Тем не менее, экстрактивные. Однако, экстрактивные системы легче внедрить, особенно для больших данных.  Методы ответов на вопросы (QA) предоставляют ответы на вопросы, заданные на естественном языке. Siri от Apple и Watson от IBM являются примерами коммерческих систем QA. Эти системы были внедрены в здравоохранении, финансах, маркетинге и образовании. Подобно абстрактному обобщению, системы QA полагаются на ком-. сложные методы НЛП. Методы QA далее классифицируются на три категории: подход, основанный на информационном поиске (IR), подход, основанный на знаниях, и гибридный подход. Системы ОК часто состоят из трех подкомпонентов. Первый - это вопрос обработки, используемый для определения деталей, таких как тип вопроса,  направленность вопроса и тип ответа, которые используются для создания  запрос. Второй - обработка документов, которая используется для извлечения  соответствующих предварительно написанных отрывков из набора существующих документов используя запрос, сформулированный в процессе обработки вопроса. Третье - обработка ответов, которая используется для извлечения ответов-кандидатов из выходных данных предыдущего компонента, ранжирования их и возврата кандидата с наивысшим рейтингом. Основанные на знаниях системы QA системы генерируют семантическое описание вопроса, которое затем используется для запроса структурированных ресурсов. Основанные на знаниях системы QA особенно полезны для ограниченных областей, таких как туризм, медицина и транспорт, где не существует больших объемов предварительно написанных документов не существует. В таких областях отсутствует избыточность данных, которая требуется для систем контроля качества на основе ИК. Siri от Apple является примером QA-системы, использующей подход, основанный на знаниях.  В гибридных QA системах, таких как Watson от IBM, в то время как вопрос семантически  анализируется, ответы-кандидаты генерируются с помощью ИК-методов.  Методы анализа настроений (opinion mining) анализируют текст, который содержит мнения людей в отношении объектов таких как продукты, организации, люди и события. Предприятия получают все больше данных о настроениях своих клиентов, что привело к распространению анализа настроений (Liu, 2012). Маркетинг, финансы, политические и социальные науки являются основными областями применения анализа настроений.  Методы анализа настроений подразделяются на три подгруппы: анализ на уровне документа, анализ на уровне предложения, анализ на уровне предложения, анализ на уровне документа, анализ на уровне предложения. Методы на уровне документа определяют, выражает ли весь документ негативное или позитивное настроение.  Предполагается, что документ содержит настроения об одном субъекте. Некоторые методы делят документ на два класса, негативные и позитивные, другие включают в себя больше классов настроений (например, пятизвездочная система Amazon) (Feldman, 2013).  Методы на уровне предложений пытаются определить полярность одного чувства об известном объекте, выраженного в одном предложении. Методы на уровне предложения должны сначала отличить субъективные настроения от объективных. Следовательно, методы на уровне предложений являются более сложными по сравнению с методами на уровне документов.  Методы, основанные на аспектах, распознают все настроения в документе и определяют аспекты сущности, к которой относится каждое настроение. Например, отзывы покупателей о продукции обычно содержат мнения о различных аспектах (или особенностях) продукта. Используя данный метод, продавец может получить ценную информацию о различных характеристиках продукта, которая была бы упущена, если бы мнение классифицировалось только с точки зрения полярности.  3.2. Аудио-аналитика  Аудио-аналитика анализирует и извлекает информацию из неструктурированных аудиоданных. В применении к человеческому разговорному языку аудио-аналитика также называется аналитикой речи. Поскольку эти методы в основном применяются к разговорной речи, термины "аудио-аналитика" и "речевая аналитика" часто используют как синонимы. В настоящее время основными областями применения аудио-аналитики являются центры обработки вызовов клиентов и здравоохранение.  области применения аудио-аналитики. Центры обработки вызовов используют аудио-аналитику для эффективного анализа тысяч или даже миллионов часов записанных звонков. Эти методы помогают улучшить качество обслуживания клиентов, оценить эффективность работы агентов, повысить коэффициент оборачиваемости продаж, следить за соблюдением правил (например, политики конфиденциальности и безопасности), получить представление о поведении клиентов, а также выявлять проблемы с продуктами и услугами, среди многих других задач. Системы аудио-аналитики могут быть разработаны для анализа живого разговора, формулирования рекомендаций по перекрестным/повышающим продажам на основе прошлого и настоящего взаимодействия с клиентом, и предоставлять обратную связь агентам в режиме реального времени. Кроме того, автоматизированные центры обработки вызовов  используют платформы интерактивного голосового ответа (IVR) для идентификации и  работы с разочарованными абонентами. В здравоохранении аудио-аналитика поддерживает диагностику и лечение определенных медицинских состояний, которые влияют на коммуникацию пациента (например, депрессия, шизофрения и рак)  (Hirschberg, Hjalmarsson, & Elhadad, 2010).  Также аудио-аналитика может помочь проанализировать плач младенца, который содержит информацию о здоровье и эмоциональном состоянии младенца (Patil, 2010). Огромный объем данных, записанных с помощью речевых клинических систем регистрации - еще один фактор, способствующий внедрению аудио-аналитики  в здравоохранении.  Речевая аналитика следует двум общим технологическим подходам: подход на основе транскрипта (широко известный как непрерывное распознавание речи с большим словарным запасом, LVCSR) и фонетический подход. Ниже приводится их описание.  - Системы LVCSR следуют двухфазному процессу: индексирование и поиск.  индексирование и поиск. На первом этапе они пытаются расшифровать речевое  содержание аудиозаписи. Для этого используются алгоритмы автоматического распознавания речи (ASR), которые сопоставляют звуки со словами. Слова определяются на основе заранее составленного словаря. Если система не может найти точное слово в словаре, она возвращает наиболее похожее. Результатом работы системы является индексный файл с возможностью поиска слов, произнесенных в речи. На втором этапе используются стандартные текстовые методы для нахождения поискового термина в индексном файле.  - Фонетические системы работают со звуками или фонемами. Фонемы  это перцептивно различимые единицы звука в определенном  языке, которые отличают одно слово от другого (например.  фонемы/к/и/б/ различают значения слов "кошка" и "летучая мышь").  Системы на основе фонетики также состоят из двух этапов: фонетического индексирования и поиска. На первом этапе система переводит входную речь в последовательность фонем. Это отличается от систем LVCSR, где речь преобразуется в последовательность слов. На втором этапе система на выходе первой фазы в виде фонетического представления ищет термин.  3.3. Видео-аналитика  Видео-аналитика, также известная как анализ видеоконтента (VCA), включает в себя различные методы мониторинга, анализа и извлечения значимой информации из видеопотоков. Несмотря на то, что видео-аналитика все еще находится в зачаточном состоянии по сравнению с другими видами анализа данных (Panigrahi, Abraham, & Das, 2010), различные техники уже разработаны для обработки как видео в реальном времени, так и предварительно записанного видео. Растущая распространенность телевизионных камер (CCTV) и стремительно растущая популярность видео-сайтов для обмена видео - вот два основных фактора, способствующих росту компьютерного видеоанализа. Однако одной из основных проблем является огромный объем видеоданных. Для сравнения, одна секунда видео высокой четкости по своему размеру эквивалентна более чем 2000 страницам текста (Manyika et al., 2011). Теперь подумайте о том, что 100 часов видео загружается на YouTube каждую минуту (YouTube Statistics,  n.d.). Технологии больших данных превращают этот вызов в возможность. Очевидно, что  необходимость в дорогостоящей и рискованной ручной обработке, технологии больших данных могут быть использованы для автоматического просеивания и извлечения информации из тысяч часов видео. В результате технология больших данных является третьим фактором, способствовавшим развитию видео-аналитики.  Основное применение видео-аналитика в последние годы находит в автоматизированных системах безопасности и наблюдения. В дополнение к высокой стоимости, системы видеонаблюдения, основанные на трудовых ресурсах, как правило, менее эффективны, чем автоматические системы. Например, в Хаким и др., 2012 утверждается, что сотрудники службы безопасности не могут сосредоточиться более чем на 20 минут. Видео-аналитика может эффективно выполнять такие функции наблюдения, как обнаружение нарушения запретных зон, идентификация предметов, унесенных или оставленных без присмотра, выявление бездомных в определенной зоне, распознавание подозрительных действий и обнаружение камер. При обнаружении угрозы система видеонаблюдения может уведомить сотрудников службы безопасности в режиме реального времени или запустить сигнал тревоги. Данные, полученные с помощью камер видеонаблюдения в торговых точках, можно извлечь для бизнес-аналитики. Маркетинг и управление операциями — это основные области применения. Например, интеллектуальные алгоритмы могут собирать демографическую информацию о покупателях, например, возраст, пол и этническую принадлежность. Аналогичным образом, розничные торговцы могут подсчитывать количество покупателей, измерять время их пребывания в магазине, определять характер их передвижения, измерять время пребывания в различных зонах, а также отслеживать очереди в режиме реального времени. Ценные сведения можно получить, соотнеся эту информацию с демографическими данными для принятия решений по размещению товара, оптимизации цен, оптимизации ассортимента, дизайна промоакций, перекрестных продаж, оптимизации планировки, оптимизации выкладки и подбора персонала.  Еще одним потенциальным применением видео-аналитики в розничной торговле является изучение покупательского поведения групп. Среди членов семьи, которые делают покупки вместе, только один взаимодействует с магазином на кассе, в результате чего традиционные системы упускают данные о покупательском поведении других членов семьи.  Видео-аналитика может помочь ритейлерам устранить эту упущенную возможность, предоставляя информацию о размере группы, ее демографических характеристиках и индивидуальном покупательском поведении отдельных членов группы.  Автоматическое индексирование и поиск видео представляет собой еще одну  область применения видео-аналитики. Широкое распространение онлайн и офлайн видео выявило необходимость индексирования мультимедийного контента для удобного поиска и извлечения. Индексирование видео может быть выполнено на основе различных уровней информации, доступной в видео, включая метаданные, звуковую дорожку, транскрипты и визуальное содержимое видео.  В подходе, основанном на метаданных, реляционные системы управления базами данных (РСУБД) используются для поиска и извлечения видео. Аудио-аналитика и текстовая аналитика могут быть применены для индексирования видео на основе  соответствующих звуковых дорожек и расшифровок, соответственно. Полный обзор подходов и методов для индексирования видео представлен в работе Hu, Xie, Li, Zeng и Maybank (2011).  С точки зрения архитектуры системы, существует два подхода  к видео-аналитике, а именно серверный и пограничный:  - Архитектура на основе сервера. В этой конфигурации видео, снятое  через каждую камеру направляется обратно на централизованный и выделенный  сервер, который выполняет видеоанализ. Из-за ограничения пропускной способности видео, генерируемое источником, обычно сжимается путем уменьшения частоты кадров и/или разрешения изображения. В результате потеря информации может повлиять на точность анализа. Однако подход на основе сервера обеспечивает экономию масштаба и облегчает обслуживание.  - Пограничная архитектура. При таком подходе аналитика применяется на  на "границе" системы. То есть, видеоанализ выполняется локально и на необработанных данных, захваченных камерой. В результате, для анализа доступно все содержимое видеопотока, что обеспечивает более эффективный анализ контента. Системы, основанные на краевом принципе, более дорогостоящие в обслуживании и имеют меньшую вычислительную мощность по сравнению с системами на базе сервера.  3.4. Аналитика социальных сетей  Аналитика социальных сетей относится к анализу структурированных и неструктурированных данных из каналов социальных сетей. Социальные медиа - это  широкий термин, охватывающий различные онлайн-платформы, которые позволяют пользователям создавать и обмениваться контентом. Социальные медиа можно разделить по следующим типам: Социальные сети (например, Facebook  и LinkedIn), блоги (например, Blogger и WordPress), микроблоги (например,  Twitter и Tumblr), социальные новости (например, Digg и Reddit), социальные  закладки (например, Delicious и StumbleUpon), обмен медиафайлами  (например, Instagram и YouTube), вики (например, Wikipedia и Wikihow),  сайты вопросов и ответов (например, Yahoo! Answers и Ask.com) и  сайты отзывов (например, Yelp, TripAdvisor) (Barbier & Liu, 2011; Gundecha  & Liu, 2012). Кроме того, многие мобильные приложения, такие как Find My Friend,  предоставляют платформу для социальных взаимодействий и, следовательно, служат социальными медиаканалами.  Хотя исследования в области социальных сетей появились еще в начале  1920-х годов, тем не менее, аналитика социальных сетей является зарождающейся областью, которая возникла после появления Веб 2.0 в начале 2000-х годов. Ключевой характеристикой современной аналитики социальных сетей является ее ориентированность на данные. Исследования по аналитике социальных сетей охватывают  несколько дисциплин, включая психологию, социологию, антропологию, информатику, математику, физику и экономику. В последние годы основным применением аналитики социальных сетей является маркетинг. Это можно объяснить широким распространением и ростом использования социальных сетей потребителями во всем мире (He, Zha, & Li, 2013), причем по прогнозам Forrester Research, Inc. социальные медиа станут вторым по темпам роста маркетинговым каналом в США в период с 2011 по 2016 год (VanBoskirk, Overby, & Takvorian, 2011).  Создаваемый пользователями контент (например, настроения, изображения, видео и  закладки), а также отношения и взаимодействия между сетевыми субъектами (например, людьми, организациями и продуктами) являются двумя источниками информации в социальных сетях. Основываясь на этой категоризации, аналитика социальных медиа может быть классифицирована на две группы:  - Аналитика на основе содержания. Контентная аналитика фокусируется на  данных, размещенных пользователями на платформах социальных сетей, таких как отзывы клиентов, обзоры продукции, изображения и видео. Такой контент в социальных сетях часто бывает объемным, неструктурированным, шумным и динамичным. Текстовый, аудио- и видеоанализ, как обсуждалось ранее, могут применяться для извлечения информации из таких данных. Для решения проблем обработки данных можно использовать технологии больших данных.  - Структурная аналитика. Также называемая анализом социальных сетей. Этот тип аналитики занимается синтезом структурных атрибутов социальной сети и извлечением информации из отношений между участвующими субъектами. Структура социальной сети моделируется через набор узлов и ребер, представляющих участников и отношения,  соответственно. Модель может быть визуализирована как граф, состоящий из узлов и ребер. Мы рассматриваем два типа сетевых графов, а именно социальные графы и графы активности (Heidemann, Klier, & Probst, 2012).  В социальных графах ребро между парой узлов означает только наличие связи (например, дружбы) между соответствующими сущностями. Такие графы могут быть использованы  для выявления сообществ или определения хабов (т.е. пользователей с относительно большим количеством прямых и косвенных социальных связей). В сетях активности, однако, ребра представляют собой фактическое взаимодействие между любой парой узлов. Взаимодействие включает обмен информацией (например, лайки и комментарии).  Графы активности предпочтительнее социальных графов, потому что активная взаимосвязь более значима для анализа, чем простое взаимодействие.  В последнее время появились различные методы извлечения информации из структуры социальных сетей. Мы кратко обсудим их ниже.  Обнаружение сообществ, также называемое выявлением сообществ, извлекает неявные сообщества в сети. Для социальных сетей сообщество относится к подсети пользователей, которые взаимодействуют друг с другом более активно, чем с остальной частью сети.  сети. Часто содержащие миллионы узлов и ребер, социальные сети, как правило, имеют колоссальный размер. Выделение сообществ помогает обобщить огромные сети, что впоследствии облегчает выявление существующих моделей поведения и предсказание возникающих свойства сети. В этом отношении обнаружение сообществ схоже с кластеризацией (Aggarwal, 2011), техникой интеллектуального анализа данных. используемой для разделения набора данных на разрозненные подмножества на основе  сходства точек данных. Обнаружение сообществ нашло несколько областей применения, включая маркетинг и Всемирную паутину (Parthasarathy, Ruan, & Satuluri, 2011). Например, обнаружение сообществ позволяет компаниям разрабатывать более эффективные системы рекомендации продуктов.  Анализ социального влияния относится к методам, которые занимаются  моделированием и оценкой влияния участников и связей в социальной сети. Естественно, на поведение пользователя в социальной сети влияют другие. Таким образом, желательно оценить влияние участников, количественно оценить силу связей, и выявить закономерности распространения влияния в сети. Методы анализа социального влияния могут быть использованы в вирусном маркетинге для эффективного повышения узнаваемости и принятия бренда.  Важным аспектом анализа социального влияния является количественная оценка  важности узлов сети. Для этого были разработаны различные показатели, включая центральность степени, центральность между узлами, центральность близости и центральность собственного вектора (подробнее см. Tang & Liu, 2010). Другие меры оценивают силу связей, представленных ребрами, или моделируют распространение влияния в социальных сетях. Линейная пороговая модель (LTM) и независимая каскадная модель (ICM) являются двумя известными примерами таких схем (Sun & Tang, 2011).  Прогнозирование связей конкретно рассматривает проблему прогнозирования будущих связей между существующими узлами в базовой сети. Как правило, структура социальных сетей не является статичной и постоянно растет за счет создания новых узлов и ребер. Поэтому естественной целью является понимание и прогнозирование динамики сети. Методы прогнозирования связей предсказывают возникновение взаимодействия, сотрудничества или влияния между субъектами сети в определенный промежуток времени. Методы предсказания связей превосходят чистую случайность в 40-50 раз, что говорит о том, что текущая структура сети наверняка содержит скрытую информацию о будущих связях (Liben-Nowell & Kleinberg, 2003). В биологии методы предсказания связей используются для обнаружения связей или ассоциаций в биологических сетях (например, межбелковые сети), что устраняет необходимость в поиске связей или ассоциаций, избавляя от необходимости проведения дорогостоящих экспериментов (Hasan & Zaki, 2011). В сфере безопасности прогнозирование связей помогает выявить потенциальное сотрудничество в террористических или криминальных сетях. В контексте социальных сетей основное применение прогнозирования связей предсказание ссылок применяется в разработке рекомендательных систем, таких как Facebook "Люди, которых вы можете знать", YouTube "Рекомендовано для вас", а также рекомендательные системы Netflix и Amazon.  3.5. Предиктивная аналитика  Предиктивная аналитика включает в себя целый ряд методов, которые стараются предсказывать будущие результаты на основе исторических и текущих данных. В  на практике предиктивная аналитика может применяться практически во всех дисциплинах – от предсказания отказа реактивных двигателей на основе потока данных с нескольких тысяч датчиков, до прогнозирования дальнейших действий клиентов на основе того, что они покупают, когда и что покупают и даже того, что они говорят в социальных сетях.  В своей основе предиктивная аналитика стремится выявить закономерности и улавливать взаимосвязи в данных. Методы прогнозной аналитики подразделяются на две группы. Одни методы, такие как скользящие средние, пытаются обнаружить исторические закономерности в итоговой переменной(ых) и экстраполировать их на будущее. Другие, такие как линейная регрессия, направлены на выявление взаимозависимостей между итоговой переменной (переменными) и объясняющими переменными, и используют их для составления прогноза. Исходя из основ методологии, методы можно также можно разделить на две группы: методы регрессии (например, мультиномиальные логит-модели) и методы машинного обучения (например, нейронные сети). Другая классификация основана на типе переменных результатов: такие методы, как линейная регрессия, рассматривают непрерывные итоговые переменные (например, цена продажи домов), в то время как другие, такие как Random Forests, применяются к дискретным переменным (например, кредитный статус).  Методы прогнозной аналитики по большей части основаны на статистических методах. Несколько факторов требуют разработки новых статистических методов для больших данных. Во-первых, традиционные статистические методы основаны на статистической значимости: небольшая выборка берется из населения и результат сравнивается со случайностью для изучения значимости конкретной взаимосвязи. Затем вывод обобщается на всю популяцию. В отличие от этого, выборки больших данных являются массовыми и представляют большинство, если не всю популяцию. Как следствие, понятие статистической значимости не столь актуально для больших данных. Во-вторых, с точки зрения вычислительной эффективности, многие традиционные методы для малых выборок не подходят для больших данных. Третий фактор связан с отличительными особенностями присущим большим данным: неоднородность, накопление шума, ложные  корреляции и случайная эндогенность (Fan, Han, & Liu, 2014).  Мы опишем их ниже.  - Гетерогенность. Большие данные часто получаются из разных источников  и представляют собой информацию из разных субпопуляций. В результате большие данные очень неоднородны. Субпопуляциями в небольших выборках считаются выбросами из-за их недостаточной фициентной частоты. Однако огромный размер больших массивов данных создает уникальную возможность моделировать неоднородность, возникающую из субпопуляционных данных, что потребует сложных статистических методов.  - Накопление шума. Оценка прогностических моделей для больших данных часто включает в себя одновременную оценку нескольких параметров. Накопленная ошибка оценки (или шум) для различных параметров может доминировать над величинами переменных, которые имеют истинные эффекты в модели. Другими словами, некоторые переменные со значительной объяснительной силой могут быть упущены из виду в результате накопления шума.  - Надуманная корреляция. Для больших данных ложная корреляция означает, что  некоррелированные переменные, которые ложно считаются коррелированными из-за  огромного размера набора данных. Fan и Lv (2008) показали это явление на примере моделирования, где коэффициент корреляции между независимыми случайными величинами увеличивается с ростом размера набора данных. В результате некоторые переменные, которые с научной точки зрения не связаны между собой (из-за их независимости), ошибочно считаются коррелированными в результате высокой размерности.  - Случайная эндогенность. Распространенным допущением в регрессионном анализе является предположение об экзогенности, когда объясняющие переменные или предикторы влияют на остаточный член. Валидность большинства статистических методов, используемых в регрессионном анализе, зависит от этого предположения. Другими словами, существование случайной эндогенности (т.е. зависимости остаточного члена от некоторых из предикторов) подрывает обоснованность статистических методов используемых для регрессионного анализа. Хотя предположение об экзогенности обычно соблюдается в небольших выборках, случайная эндогенность обычно присутствует в больших данных. Стоит отметить, что, в отличие от ложной корреляции, случайная эндогенность относится к истинной связи между переменными и членом ошибки. Несоответствие статистической значимости, проблемы эффективности, а также уникальные характеристики больших данных, рассмотренные выше, подчеркивают необходимость разработки новых статистических технологий для получения информации из прогностических моделей. |
| 4. Concluding remarks  The objective of this paper is to describe, review, and reflect on big data. The paper first defined what is meant by big data to consolidate the divergent discourse on big data. We presented various definitions of big data, highlighting the fact that size is only one dimension of big data. Other dimensions, such as velocity and variety are equally important. The paper's primary focus has been on analytics to gain valid and valuable insights from big data. We highlight the point that predictive analytics, which deals mostly with structured data, overshadows other forms of analytics applied to unstructured data, which constitutes 95% of big data. We reviewed analytics techniques for text, audio, video, and social media data, as well as predictive analytics. The paper makes the case for new statistical techniques for big data to address the peculiarities that differentiate big data from smaller data sets. Most statistical methods in practice have been devised for smaller data sets comprising samples.  Technological advances in storage and computations have enabled cost-effective capture of the informational value of big data in a timely manner. Consequently, one observes a proliferation in real-world adoption of analytics that were not economically feasible for large-scale applications prior to the big data era. For example, sentiment analysis (opinion mining) have been known since the early 2000s (Pang & Lee, 2008). However, big data technologies enabled businesses to adopt sentiment analysis to glean useful insights from millions of opinions shared on social media. The processing of unstructured text fueled by the massive influx of social media data is generating business value by adopting conventional (pre-big data) sentiment analysis techniques, which may not be ideally suited to leverage big data.  Although major innovations in analytical techniques for big data have not yet taken place, one anticipates the emergence of such novel analytics in the near future. For instance, real-time analytics will likely become a prolific field of research because of the growth in location-aware social media and mobile apps. Since big data are noisy, highly interrelated, and unreliable, it will likely lead to the development of statistical techniques more readily apt for mining big data while remaining sensitive to the unique characteristics. Going beyond samples, additional valuable insights could be obtained from the massive volumes of less ‘trustworthy’ data. | 4. Заключение  Цель данной работы - описать, рассмотреть и осмыслить понятие больших данных. Сначала в статье было дано определение того, что подразумевается под большими данными, чтобы урегулировать разноречивые рассуждения о больших данных. Мы представили различные определения больших данных, подчеркивая тот факт, что размер — это только один из аспектов больших данных. Другие, такие как скорость и разнообразие, являются не менее важными. Основное внимание в статье было уделено аналитике для получения достоверной и ценной информации из больших данных. Мы подчеркиваем, что предиктивная аналитика, которая в основном имеет дело со структурированными данными, затмевает другие формы аналитики, применяемые к неструктурированным данным, которые составляют 95% больших данных. Мы рассмотрели аналитические методы анализа текста, аудио, видео и данных социальных сетей, а также предиктивную аналитику. В статье приводятся аргументы в пользу новых статистических методов для больших данных, чтобы учесть особенности, которые отличают большие данные от меньших наборов данных. Большинство статистических методов были разработаны для небольших наборов данных, состоящих из выборок.  Технологические достижения в области хранения данных и вычислений позволили  позволили экономически эффективно и своевременно получать информационную ценность больших данных своевременно. Как следствие, мы наблюдаем распространение в реальном мире аналитики, которая не была экономически оправдана до наступления эры больших данных. Например, анализ настроений (поиск мнений) известен с начала 2000-х годов (Pang & Lee, 2008). Однако технологии больших данных позволили предприятиям использовать анализ настроений для извлечения полезные сведения из миллионов мнений, которыми делятся в социальных сетях. Обработка неструктурированного текста, подпитываемая огромным потоком данных из социальных сетей, приносит прибыль бизнесу, если не использовать традиционных (до появления больших данных) методов анализа настроений, которые могут не подходить для работы с большими данными.  Хотя крупные инновации в аналитических методах для больших данных еще не произошли, можно ожидать появления таких методов анализа в ближайшем будущем. Например, аналитика в реальном времени, вероятно, станет обширной областью исследований из-за роста числа социальных сетей и мобильных приложений с учетом местоположения. Поскольку большие данные являются шумными, сильно взаимосвязанными и ненадежными, это, скорее всего, приведет к разработке статистических методов, которые будут более эффективными для обработки больших данных, сохраняя при этом чувствительность к их уникальным характеристикам. Помимо образцов, дополнительные ценные сведения могут быть получены из огромных объемов менее "надежных" данных. |