

A Vision for Data Monetization in Canada

Canada's Big Data Consortium
May 2018

We would like to thank the following organizations for their leadership and in-kind contributions to Canada's Big Data Consortium, Canada's Big Data Talent Gap Study, and to this paper, "A Vision for Data Monetization in Canada."





A Vision for Data Monetization in Canada

Canada's Big Data Consortium
May 2018

Contents

05 Abstract

06 Executive Summary

09 1 Introduction

10 1.1 Strategies for Data Monetization

12 1.2 Conditions for Data Monetization

12 1.2.1 The Six V's

14 1.2.2 Cost

15 1.2.3 Ethics

16 1.3 Data Assessment

18 1.4 Open Data Sets

19 1.5 Public Data

20 1.6 Private Data

21 1.7 Things to consider

22 2 Marketing

24 2.1 A Case Study: Retail Marketing

24 2.1.1 Data sets

25 2.1.2 Data Preparation: Understanding
Data Characteristics

25 2.1.3 Methodology and Results

29 2.2 Opportunities and Challenges

30 2.3 Conclusion

32 3 Finance

34 3.1 Opportunities

34 3.1.1 Case Study: Analysis of Financial
Conference Call Transcripts

35 3.1.2 Monetizing Data in Banking

36 3.1.3 Monetizing Data in Insurance

37 3.2 Challenges

38 3.3 Conclusion

40 4 Real Estate

40 4.1 Introduction

42 4.2 Opportunities and Challenges

44 4.3 A Case Study: Toronto's Housing Market

44 4.3.1 Unemployment Rate Effects

44 4.3.2 Senior Population Effects

44 4.3.3 Housing Occupancy Effects

44 4.3.4 Pollution and Poverty Effects

44 4.3.5 Crime Rate Effects

45 4.3.6 Demographic Effects

46 4.4 Conclusion

48 5 Security

48 5.1 Introduction

50 5.2 Opportunities

50 5.2.1 Current Effectiveness and Potential

50 5.2.2 Techniques of Big Data Security Analytic

51 5.2.3 Advantages and Applications of Big
Data Analytics

52 5.2.4 Cyber Security Workforce Opportunities

53 5.3 Challenges

53 5.3.1 Threat Response
53 5.3.2 Technical Challenges
55 5.4 Conclusion

56 6 Health Care

56 6.1 Introduction
58 6.2 Opportunities
58 6.2.1 Monetization
59 6.2.2 Additional references:
60 6.3 Challenges
60 6.3.1 Legislation
60 6.3.2 Privacy and Security
62 6.4 Conclusion

64 7 Social Media

64 7.1 Introduction
65 7.2 Opportunities
66 7.3 Challenges
66 7.3.1 Privacy Issues
67 7.3.2 Noise and Potential Biases
68 7.3.3 Context Consideration
69 7.4 Conclusion

70 8 Energy and Mining Industries

70 8.1 Introduction
72 8.2 Opportunities
74 8.3 Challenges

75 8.4 Conclusion

76 9 Manufacturing

76 9.1 Introduction
78 9.2 Big Data, Transparency, and Predictive Manufacturing
78 9.2.1 Conceptual Framework of a Predictive Manufacturing System
80 9.2.2 Deployment of the Internet of Things (IoT) and Real-Time Data
81 9.2.3 Predicting Future Opportunities
82 9.3 Two approaches for Monetizing Data in Manufacturing
83 9.4 Conclusion

84 10 Government

84 10.1 Introduction
86 10.2 The Role of big data Analytics in the Government Sector
86 10.2.1 Homeland Security and Law Enforcement
86 10.2.2 Economic Growth and General Welfare
88 10.3 Opportunities and Challenges
89 10.4 Conclusion

90 11 Conclusion

92 Glossary

95 WorkshopParticipants

96 Bibliography

Principal Contributors

Shariyar Murtaza, Ph.D
TELUS Communications Company

Tamer Abdou, Ph.D
Ryerson University

Natalie kasparian
Ryerson University

Dalia Shanshal
Ryerson University

Contributors

Shiva Amiri
BioSymetrics Inc.

Daniel Gent
Canadian Electricity Association

Claire Gabillard
Government of Canada

Ayse Bener, Ph.D
Ryerson University

Tess McDonald
Ryerson University

Malcolm White
Signature.ci

Bryan Smith
ThinkData Works

Gregory Richards
University of Ottawa

This Big-data Consortium project is supported by the Office of the Provost and Vice-President, Academic and led by Data Science Laboratory at Ryerson University.

Abstract

We live in a world increasingly driven by data; 2.5 quintillion bytes of data are produced everyday. With all this data at hand, companies and organizations stand in front of an abundance of information. The challenge is to understand the data and turn it into valuable insights to be used for performance improvement. In this paper, we explore the topic of data monetization, namely, how we can convert big data into a source of wealth. We cover variety of aspects of data monetization in this paper in a meaningful capacity without taking deep dive in topics. This paper is divided into 10 chapters. In the first chapter, the introductory chapter, we discuss the different strategies and conditions for data monetization, as well as the sources of data. In the following 9 chapters, we examine big data opportunities and challenges specific to each sector. This includes real estate, marketing, security, health care, social media, energy and mining, finance, government, and manufacturing. Our research shows that there are exceptional opportunities hidden under data mines. However, strong industrial and academic collaboration is needed to address the challenges of lack of skilled human resources in big data analytics and the lack of publicly available big datasets. These challenges can be mitigated with the help of new government initiatives and policies.

Executive Summary

This study examines the means, methods and outcomes of data monetization within different sectors, and highlights the opportunities and challenges of data monetization for each sector. Monetization through data could mean better use of the data assets to improve customer acquisition, customer experience, brand loyalty, fraud detection, loss control and generating revenue by selling data analytics services.

In our research, different general strategies for data monetization are uncovered, and these are applicable in all sectors. The different strategies are: leveraging proprietary data for internal growth, trading data with business partners, adopting an “open by default” data strategy to increase access to information, making use of external data to increase revenue, selling data as service bundles, and selling data on pay per use basis.

Throughout our research, we noticed that data is an under-utilized asset; currently, only 30 percent of companies use data analytics. If used efficiently, the value derived from data is very profitable. That is true across different sectors. Below is a brief summary of the advantages of data analytics with examples drawn from various sectors.

Higher Prediction Accuracy

- In the real estate sector, higher accuracy in economic forecasting permits stakeholders to make more rational investment decisions.
- In the manufacturing sector, the enhanced prediction of machine performance and operational performance makes the production process more efficient.

Analysis of Consumer Behavior

- In the marketing sector, a thorough analysis of consumer behavior through data analysis leads to more efficient marketing strategies and better audience target. The results translates into increased promotions and sales.
- The energy sector also benefits from consumer analysis behavior. The tracking of consumers’ pattern of energy consumption allows for the optimization of asset management and the creation of better energy conservation strategies.

Sentiment Analysis

- In the social media sector, the use of the vast available data has proven to be a successful tool for business exposure and sales promotion. It has also proved to be successful when used to track socio-political sentiment.
- Sentiment analysis is particularly interesting in the finance and banking sector. Sentiment analysis of conference calls data for example, can bring immediate value to financial practitioners, in such a way that they can adjust their recommendations to clients and optimize their portfolios.

Real Time Analysis

- In the security sector, anomaly detection and misuse detection are effective methods of big data for the enhancement of cyber-security. They monitor real time activities and automatically flags any suspicious activity. This type of data analysis usage makes it easier to detect hackers and fraudulent activities.

Exploratory Analysis

- In the health care sector, biologists are moving from traditional labs to genomic data as well as data found in medical imaging or public and private health databases to predict epidemics, cure diseases and avoid preventable death. An interesting finding in the health sector is the use of new technology increasingly used for diagnosis purposes and personalized treatments. Health applications that monitor an individual's health measurement is an example of such technology.
- In the government sector, machine learning methods can be used to draw a better segmentation of the population. This in turn, permits governments to gain better insights on the population they serve and design programs that can yield better social outcomes. Overall, the application of big data analysis in the government sector contributes to the economic growth and general welfare.

In order to use data efficiently, companies must take into consideration the following factors that could influence the value of data as an asset:

Machine Capabilities

- **Storage** - A competent data storage is needed to store the massive amount of data available.
- **Performance** - Performing processing capabilities are needed to run fast and continuous flow of large and real-time data. We noted that this is particularly important in the security sector, as real-time data is consistently being monitored for threat detection.

Data Formatting

- Data normalization - With the variety of data types, such as videos, emails, spreadsheets, audio, pdf etc. It is necessary to adopt a process that can normalize, maintain and operationalize different types of data. We noticed, for example, that handling unstructured data and choosing the right tool and methodology for analysis can be quite challenging in finance sector.
- Data cleaning - Constant cleaning must be performed in order to remove biases, noise and abnormality in the data. Social media data for example, contains a considerable amount of missing values, inaccurate data or noise from advertisement. Cleaning the data is therefore a significant challenge in this sector.

Bias and Context

- Population representativeness within sample selection could lead to erroneous results when inaccurately selected. Similarly, without a deep understanding of the context, data analysis results could have negative impacts.
- In the social media sector for example, it is important to understand the context before coming to conclusions. It is not sufficient to notice a trend in social media. A trend can have positive connotations to some users and negative connotations to others, therefore it is important to understand the context.

Lack of Skilled Workers

- As the use of data analytics is expected to increase, so too is the demand for experts in this field. It is important to find professionals possessing the necessary skills for handling such complex data.
- In the cyber-security sector specifically, it is expected that a workforce shortage of more than 1.5 million cyber-security professional will take place by 2019.

Nature of Market

- Data analytics can have limitations especially when it comes to prediction, as some sectors are volatile and complex by nature. A clear example of this challenge is found in the real estate sector, as this sector is very sensible to sudden economic fluctuations.

Privacy

- There is a thin line between accessing individuals' data and engaging in privacy breach issues. Regulations and legislations are being studied to overcome this obstacle. The privacy issue is the most

recurrent challenge that we found in our study in almost all sectors. An interesting example is mentioned in the social media chapter, where insurance companies are finding posts about a customer and adjusting their premiums accordingly.

Cost Another challenge is the high cost of investment in the field of Big Data Analytics. However, the returns could compensate for the cost of that investment.

Recommendation We recommend that governments and businesses invest in data analytics. We also highly recommend collaboration between public and private sectors to leverage fully the data and analytics on data. Another important collaboration needed is an inter-disciplinary collaboration, either in different departments within a company, or within different academic fields. Finally, our research indicates that big data technology is rapidly and constantly evolving, as such, businesses and governments must keep up with this technology in order to improve their performance and efficiency.

1 Introduction

Businesses see the value in utilizing their data, which consists of vast amounts of customer, process and operational information, in order to get strategic insight, realize profits and generate revenue. Massive volumes of structured and unstructured data, decreasing storage costs, data-driven marketing campaigns, and improved business intelligence give data monetization methods the perfect environment to expand.

Data monetization could mean better use of the data assets to improve transactional performance, customer acquisition, customer experience, brand loyalty, fraud detection and loss control. But for a number of reasons, data remains an under-utilized asset in most industries. But several challenges – talent shortages, technical constraints, compliance, privacy, and security issues – are limiting the ways data is being used by enterprises. EMC corporation [1] narrows down the common challenges across all sectors to organizational resistance, conservative regulatory requirements, inflexible organizational structure and outdated market strategies.

In this paper, we will examine multiple strategies for monetization of data by looking at concrete examples and case studies across nine areas of business and government. Our objective is to build a better understanding of the various ways organizations can monetize data to support profitability and gain competitive advantages. This understanding can facilitate Canadian industries to embrace big data analytics and leverage the potential of data in their organizations to support competitiveness and maintain global leadership positions in their industries. In this chapter, we will discuss the basics and conditions of the data monetization that can be used by organizations. In following chapters, we will discuss data analytics for different sectors with a focus on nine sectors and business areas, as depicted in Figure 1.1.

Data Monetization Strategies



Figure 1.1: Overview

1.1 Strategies for Data Monetization

Today, we are just beginning to tap into the potential of data. Businesses are swimming in a pool of data that could generate valuable business insights. The volume of data is growing exponentially and the cost of data storage is decreasing, making data an underutilized asset in many organizations. A McKinsey study found that the U.S. health care sector alone could generate \$300 billion a year by using big data analytics [2]. Manirul [2] states that retailers can potentially increase revenue by more than 60% through the use of data science tools. Currently only 30 percent of companies use data analytics in their projects [2] and only 21 to 38 percent of marketing departments are currently using big data techniques [3]. In two years this range will double to 51 to 62 percent. Within companies that are analyzing big data, insights developed in one department are often not shared across the organization, limiting their potential to drive better business decisions. Governments host an enormous amount of valuable information but, for various reasons, are not able yet to use data to its full potential. There is high demand for data scientist professionals and 80 percent of positions for this skill set cannot be filled.

At the same time, the growing quantity of data is changing how data assets are measured and perceptions of its economic value to buyers, owner, sellers, and traders of data assets [4]. Higher processing velocity, greater precision and greater scale of data are altering business decisions and proving useful to businesses and markets. The demand for data monetization methods is continuously increasing. Below are the four key strategies for data monetization [4].

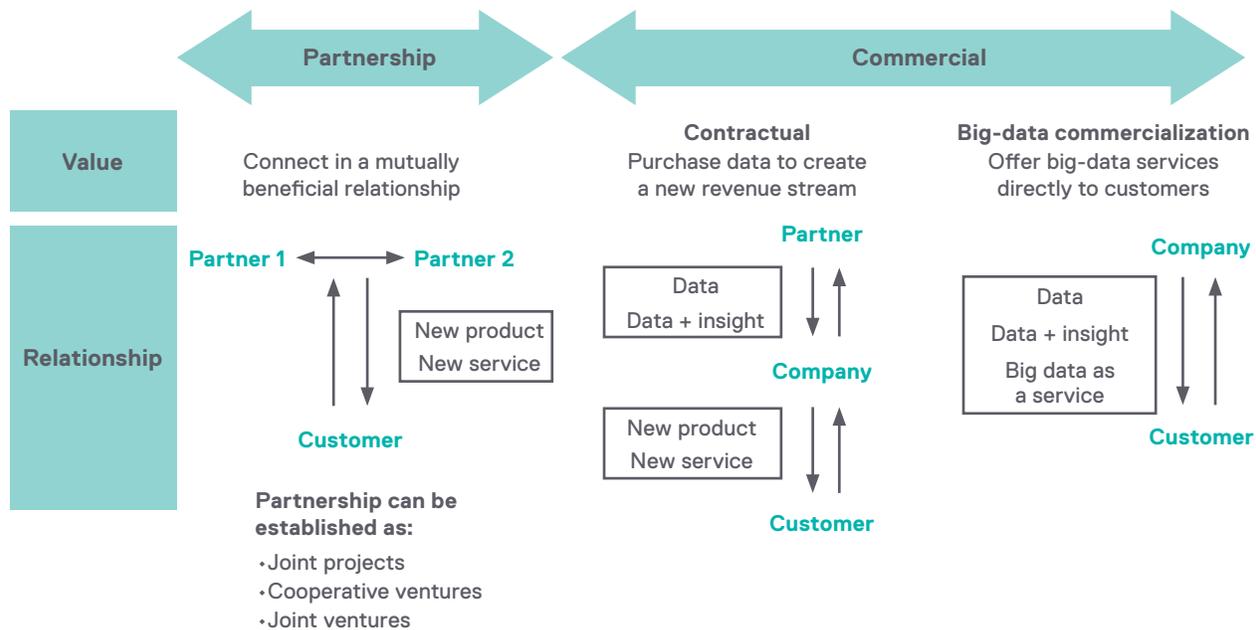
- Keep the data proprietary and leverage it for internal growth and efficiency. In this strategy, an organization uses data as an opportunity to assess operations and find opportunities to improve.
- Trade the data to business partners to improve business interactions (i.e., speed and accuracy

- during transactions). Vertical integration of data upward to suppliers and downward to business partners can improve the joint business relationships and overall market offering. This method can be useful when the operation and success of a firm requires the participation of many firms. Platt et al. [5] differentiates companies into those with large amounts of existing transactional data and those without sufficient amounts of data to create value. Often the first type sells to the second type. Platt et al. [5] also breaks down profit methods of big data into seven different types where three differ in terms of the product or service delivered as illustrated in Figure 1.2
- Sell the data to possible clients, such as a third party which collects data from a variety of places and sells segments of personalized relevant data to buyers [6]. Companies frequently contact third parties but data partnerships and alliances with other companies are rare.
- Make a portion of enterprise data publicly available, or adopt an “open by default” data strategy that would make non-proprietary data publicly available for consumption as it is created. The availability of timely and relevant data, even if it is unstandardized data, will promote an open-air marketplace for external data.

Platt et al. [5] also identifies different monetization strategies for data: build to order, service bundles, plug and play, pay per use, commission, value exchange and subscription. “Build to order” implies that data products and services are tailored to customers’ specifications. “Service bundle” represents several offerings combines into one package deal. “Plug and play” occurs when the same product is sold to every buyer. “Pay per use” is an option that gives customers easy access to many different offerings where only the offerings used are taken into account at pay. The “commission” model

is usually more long lasting than the pay-per-use model because of the continuous revenue-sharing. “Value exchange” works when there is a partner between the company and the customer where the

customer creates rebate, discounts or additional services depending on the business. “Subscription” models requires the client to periodically pay a fee for unlimited access to services during a set period of time.



Source: BCG analysis

Figure 1.2

KPMG [7] defines an effective data monetization strategy as one which combines strategy and execution in a dynamic and evolving way. KPMG [7] simplifies some methods of gaining comparative advantage through the use of unstructured and structured internal and external data, such as vertical value delivery, horizontal value delivery and cross-market value delivery. Vertical value delivery occurs when data from one segment of an industry can be used in another. For example, pharmaceutical companies will find prescription data very valuable. Horizontal value delivery occurs with various similar industries that use the same type and form of data. For instance,

a real estate firm would use retail sales data. Crossmarket value delivery occurs when a method typically used for one purpose in an industry is valuable for another purpose in other businesses. For example, insurance companies and car lease rates can use driver information captured from vehicles. Figure 1.3 depicts the level of value that big data can bring to a sector compared to the ease in which the industry players can start utilizing big data [8].

To leverage these strategies, we need to understand the conditions and capabilities that support data monetization.

1.2 Conditions for Data Monetization

1.2.1 The Six V's

Normandeau [9] describes the “six V’s” of big data: volume, variety, velocity, veracity, validity and volatility. Volume relates to the massive amounts of storage needed to store billions of terabyte of information. Volume constraints are being solved by using hyperscale computing environments, such as Hadoop, NoSQL, MapReduce, etc., which are designed to deal with enormous data sets [10].

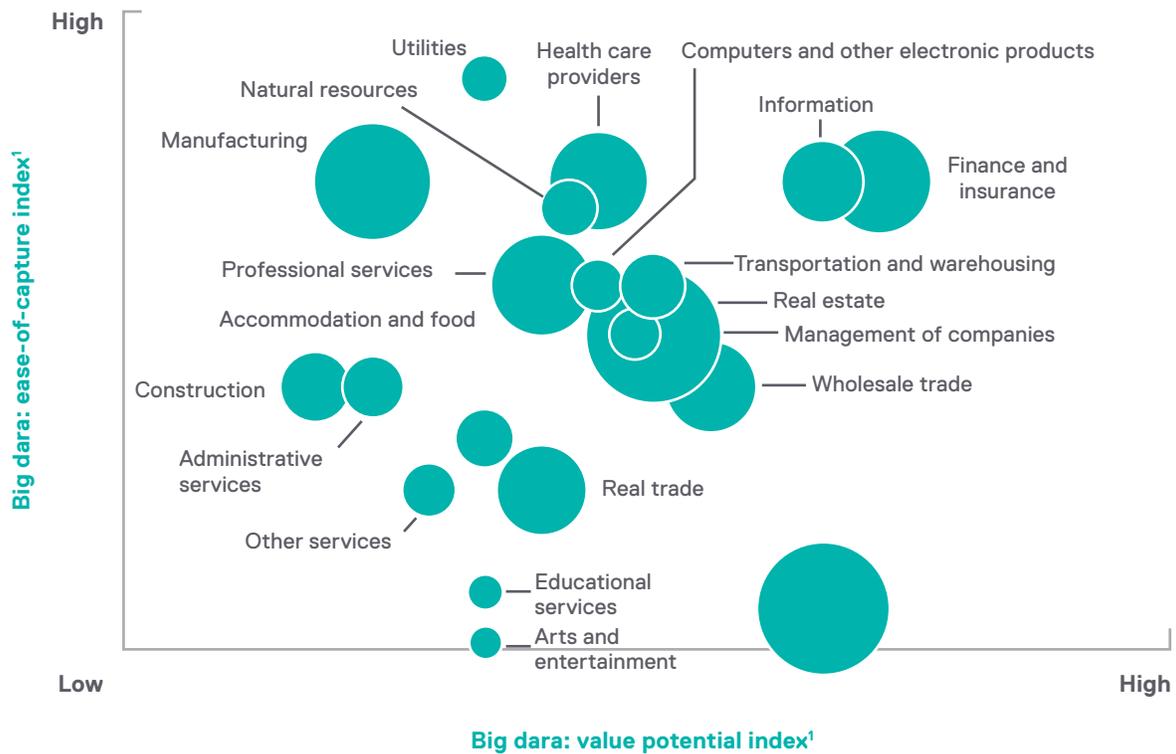


Figure 1.3

Variety refers to the diverse types of structured and unstructured data, such as videos, emails, spreadsheets, audio, PDFs, etc., and the multiple ways of representing data, such as XML, JSON, graph, etc. Connecting multiple, complex types of data from different sources requires skill and experience and a very different kind of platform than what is available through traditional data infrastructure solutions. In order to effectively leverage the variety of data that is currently available, it is necessary to adopt a solution that can normalize, maintain, and operationalize data that does not correspond to internal specifications in its raw form. Data management companies such as ThinkData Works and Import.io grew out of the need to transform available raw data into accessible structured data. But whereas Import.io aims to digitize raw information from web pages, ThinkData’s external data management platform, Namara, gives users the opportunity to connect to standardized external data from any number of sources. Mason [11] states that this challenge has given rise to many integration frameworks and APIs (Application Program Interface).

Velocity refers to fast and continuous data flow supplied by real-time data collection. Many environments have been created to deal with high-velocity data, such as Spark, Storm and Kafka. Kafka serves as a real-time data aggregator and log service provider while Storm and Spark offer processing capabilities for large volumes of real-time data. Mason [11] provides a solution for velocity and variety, which is depicted in Figure 1.4. Veracity refers to the biases, noise and abnormalities in data that can cause unreliable or wrong insights.

Data quality is relatively the biggest challenge for big data analytics. As data is generated, it needs to constantly cleaned and processes need to be run to keep “dirty data” from accumulating [9]. Validity requires data correctness and accurate use of the data. Volatility refers to the length of data validity and decisions about when data is no longer relevant to the current analysis.

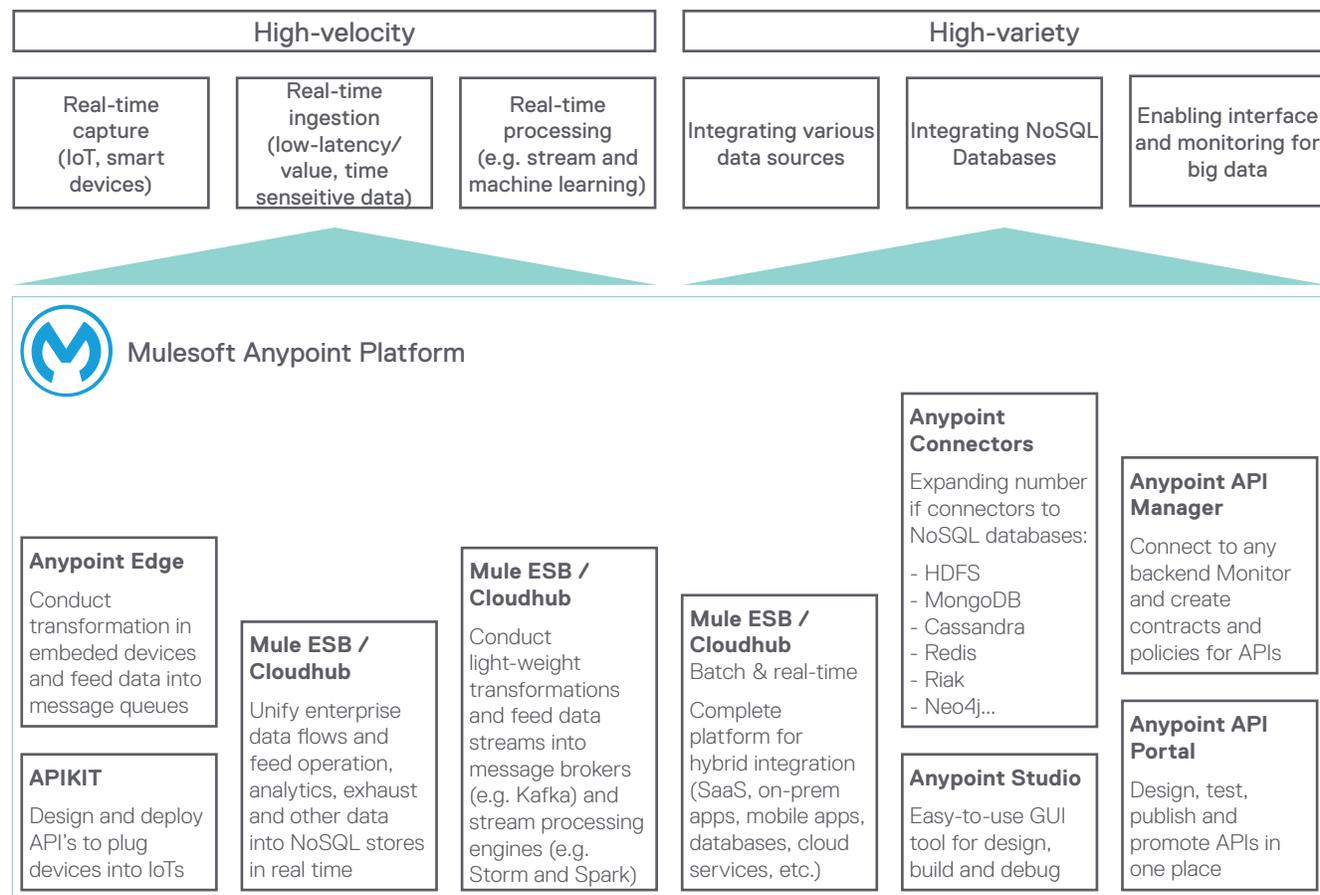


Figure 1.4

1.2.2 Cost

Big data offers potential for big business gains, but the cost of initializing data monetization strategies must be weighed against expected revenue gains and added business complexity. Some organizations may not be able to justify a large infrastructure upgrade, but there are methods of improving data monetization with low cost. Suavity [12] states that a petabyte Hadoop cluster will require between 125 and 250 nodes costing around \$1 million and the cost of supported Hadoop distribution will have similar annual costs (4,000 per node). In comparison, enterprise data warehousing costs (\$10 to \$100 million). Hadoop is a new player in the game but is making great strides toward improving reliability and ease of use of big data. There is no shortage of innovation coming from start-ups and major contributors to the Apache open source project. Two areas that will have the most serious impact in both ease-of-adoption and cost [12]:

- Leverage existing SQL query language and existing BI tools against data within Hadoop.
- Compress data at the most granular level, which will reduce storage requirements, drive down the number of nodes and simplify the infrastructure.

Suavity [12] concludes that the above listed capabilities will help keep up with growing business demands and keep data growth rates at a manageable level so that the cost of scaling petabytes of data daily is kept at a minimum.

Enterprise data warehousing costs 10 times more than the Hadoop cluster

1.2.3 Ethics

Companies are generally collecting more and more digital information, sparking privacy concerns related to big data. These concerns stem from organizations collecting information that may reflect an individual's interests, affiliations, habits or sexual orientation and so on, without there being a clear understanding of how information will be used [13]. Likewise, owning data that relates to private information, such as financial or health issues can raise important ethic issues and comes with great responsibility. Lawsuits, negative media coverage and malpractice of companies can have reverse effects on the data monetization aim of analytics.

Developing technical solutions that can identify trends and patterns while protecting individual identities is crucial. However, governments share the burden of the responsibility to protect privacy by implementing good regulations about the means of collecting data and its use. While governments must be careful to anonymize data to protect their citizens, it is also critical that they release as much of their internal data as is possible in order to ensure a lake of public data large enough to represent actual value to businesses, individuals, and the government itself. Rather than looking at the release of data as a mandated obligation, governments should be aware of the economic and social benefits that will accrue to them by virtue of this unhampered transparency.





1.3 Data Assessment

Generating valid insights from big data analytics depends largely on the quality of data used for analysis. Conducting a data assessment provides strategic and operational insight into a company's performance. It is used when a business is implementing a new system or looking to improve the quality of data in existing systems [14]. By using data experts to break down all master, operational and transaction data, executives can figure out whether their data meets industry standards, if they have the right skills on their teams, and whether they data they have is usable for different requirements.

The initial decision for an assessment is often because of confusion around a company's data quality. Michale Collins [14] lists duplication, misalignment, accounting issues and human resources errors as common data problems. The following list provides examples for these types of data inaccuracies [14].

- **Overlapping data:** Overlapping data refers to similar records that correspond to the same customer or product and is the main problem in data quality. For example, a sales person may create a duplicate record to execute a sales order, causing delivery and payment errors and miscommunication. Customer satisfaction can also be affected.
- **Misaligned data in multiple systems:** Problems occur if multiple systems that share common data are not synchronizing properly or if there are data entry errors. For example, one system may be intended vendor and procurement information and another for managing details around maintenance for physical assets.

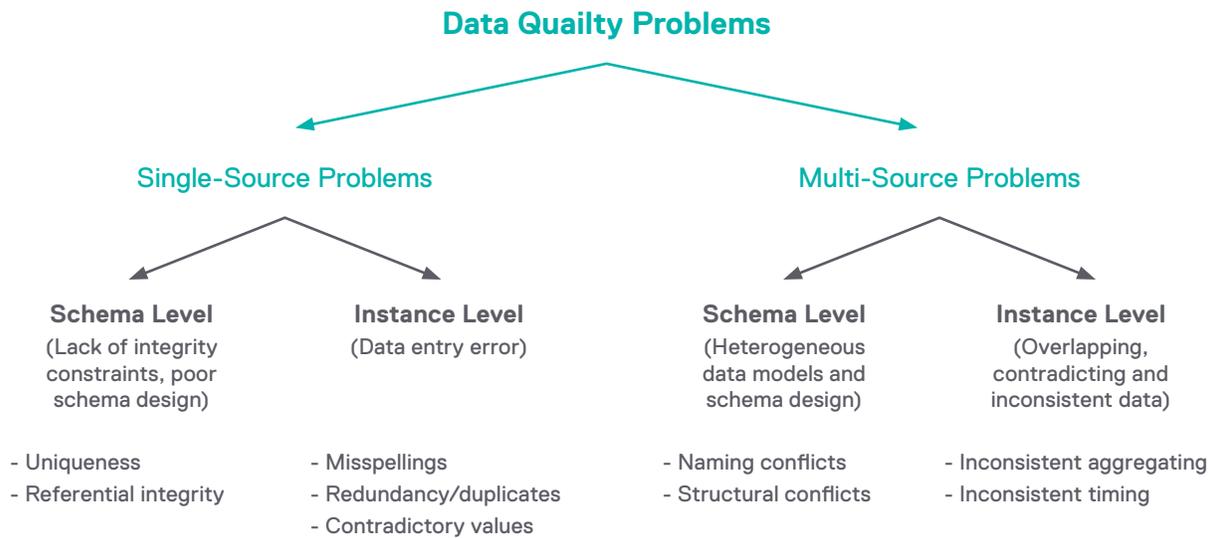


Figure 1.5: Classification of data quality problems in data sources

If someone mistakenly used the second system to add vendors or include information about purchase orders, the first system would consistently have different vendors and vendor attribute data, creating issues around missing purchase orders and misaligned budgeting.

- Transforming data: Data may need to be transformed to usable and scalable formats in order to be deemed useful for analysis. Huge volumes of a variety of non-standard formats (i.e., video, email, etc.) is a common challenge across all industries and requires modern infrastructure that can process and manipulate big data, connect to external data, and create aggregates based on an ideal schema. This infrastructure, however, must be efficient, cost-effective and a timely way to make data accessible and extract insights. Companies such as Oracle are offering services that provide this solution [15]. These services manage unstructured and structured raw data from a variety of relevant sources, organize the data and make it readily available for SQL queries. It may sometimes be more costly and timely to hire experts who can engineer big data

rather than hire a corporation such as Oracle Big Data Appliance solution.

- Payment terms: Non-standard payment terms are common. If payment data is not accounted for and shared across internal systems, misunderstood cash flow and accounting issues can arise.
- Chart of Accounts (COAs): To increase efficiency and effectiveness, assessing the COAs across all relevant systems is crucial. Accuracy of financial reporting affects the company's statistics such as tracking key performance indicators or closing the books on time.
- Human Resources mistakes: Examples are continuing to pay former employees who have not been removed from payroll, or switching birth and hiring dates, which can impact benefits eligibility and cause inadvertent regulatory violations and security vulnerabilities.

Figure 1.5 classifies data quality problems that analysts need to consider during the data assessment process [16].

1.4 Open Data Sets

Open data sets have the power to bridge the information gap between industries and public sectors. Open data support data diversification, which has potential to create new insights and drive increased productivity. Private and public collaboration and co-operation with open data is also key to strengthening the global data structure. Protecting the privacy of the personal information is a major concern with open data sets; however, a 2013 McKinsey study highlights several advantages to sharing open data sets: [17]:

- Supporting increased efficiency and profitability, and development of innovative products and services
- Supporting better decision-making, business models, products and services through experimentation and transparency of the data. An example would be allowing a company to customize its products or services for unique population clusters.
- More potential for business opportunities through increased efficiency and productivity which can enable new ideas for new and existing businesses.
- Benefits for customers as well as business through transparency of products and prices, and new methods of providing feedback to companies.
- Raising awareness of legal and regulatory concerns about open data that need to be addressed

Although open data initiatives have grown exponentially in the past decade, there is a gulf between data providers who have information and data consumers who want it. Unfortunately, neither data providers nor consumers are ideally suited to bridge this divide. Data providers, such as governments, are not able to accurately anticipate the many possible uses for their data and cannot, therefore, effectively standardize their output for all the known and unknown intended uses. Data consumers, since they cannot be sure that the product is worth the effort, are not eager to spend the time, energy, or money that will be necessary to bring it into their internal environment. As a result, the gulf persists. This has created opportunities for digital “infomediaries” to shuttle data from provider to consumer without the significant expansion of open data sets.

The following sections distinguish between public and private data.

Market size of open data is expected to increase by 37% between the years 2016 and 2020 and have a value of up to 76 billion euros by 2020 [18]. According to predictions made by Gartner, 80% of businesses will be using open data in the coming years.

1.5 Public Data

Public data is open access information that come from a variety of sources in the public or private sector. Research from McKinsey [17] suggests that \$3 trillion a year could be generated through improved efficiency, operational effectiveness and innovation in seven areas of the global economy if more data was open to the public. These areas are education, transportation, consumer products, electricity, oil and gas, health care, and consumer finance.

Below are some examples of public data sources:

- Canada public sector data: <http://open.canada.ca/en/open-data>
- Locations and mappings - City of Toronto: www1.toronto.ca
- United States public sector data: www.data.gov
- United States Food and Drug Administration data: www.fda.gov/default.htm
- Australia public sector data: www.data.gov.au

The problem with leveraging data from any of these sources is that they do not correspond to common standards, formats, or use restrictions. Using two of these sources is exactly twice as much work as using one, which makes scaling an external data

solution difficult. ThinkData Works' external data management platform, Namara, has provided the data from over 1,000 sources across North America in order to create a public-facing warehouse of public data that can be accessed using an API. In order to monetize and operationalize the value of public data, it is clear that this kind of API-based vascular system must be in place.

Market size of open data is expected to increase by 37% between the years 2016 and 2020 and have a value of up to 76 billion euros by 2020 [18]. According to predictions made by Gartner, 80% of businesses will be using open data in the coming years.

1.6 Private Data

Private data is proprietary or commercial information that is not open to the public, and includes some public sector data. Consumer preferences can be found through thorough analysis of open data provided by companies which would help uncover new marketing methods, find anomalies and innovative products.





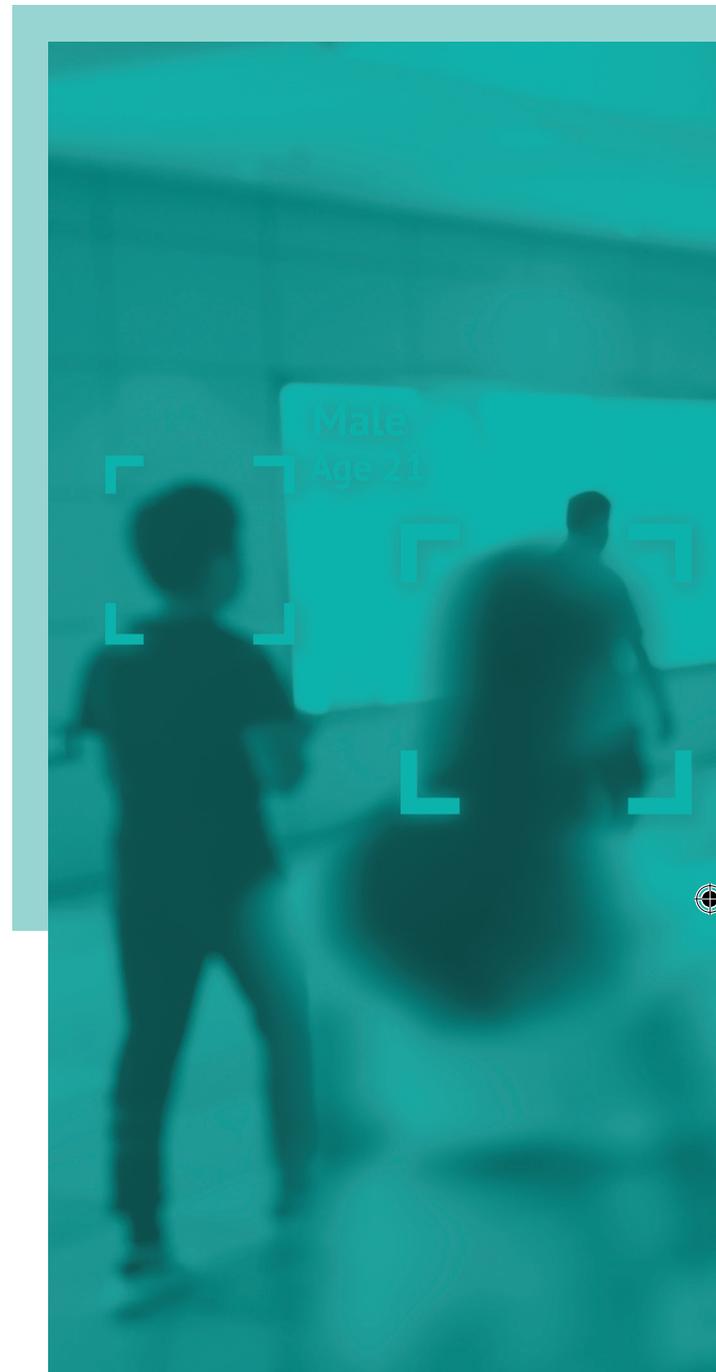
1.7 Things to consider

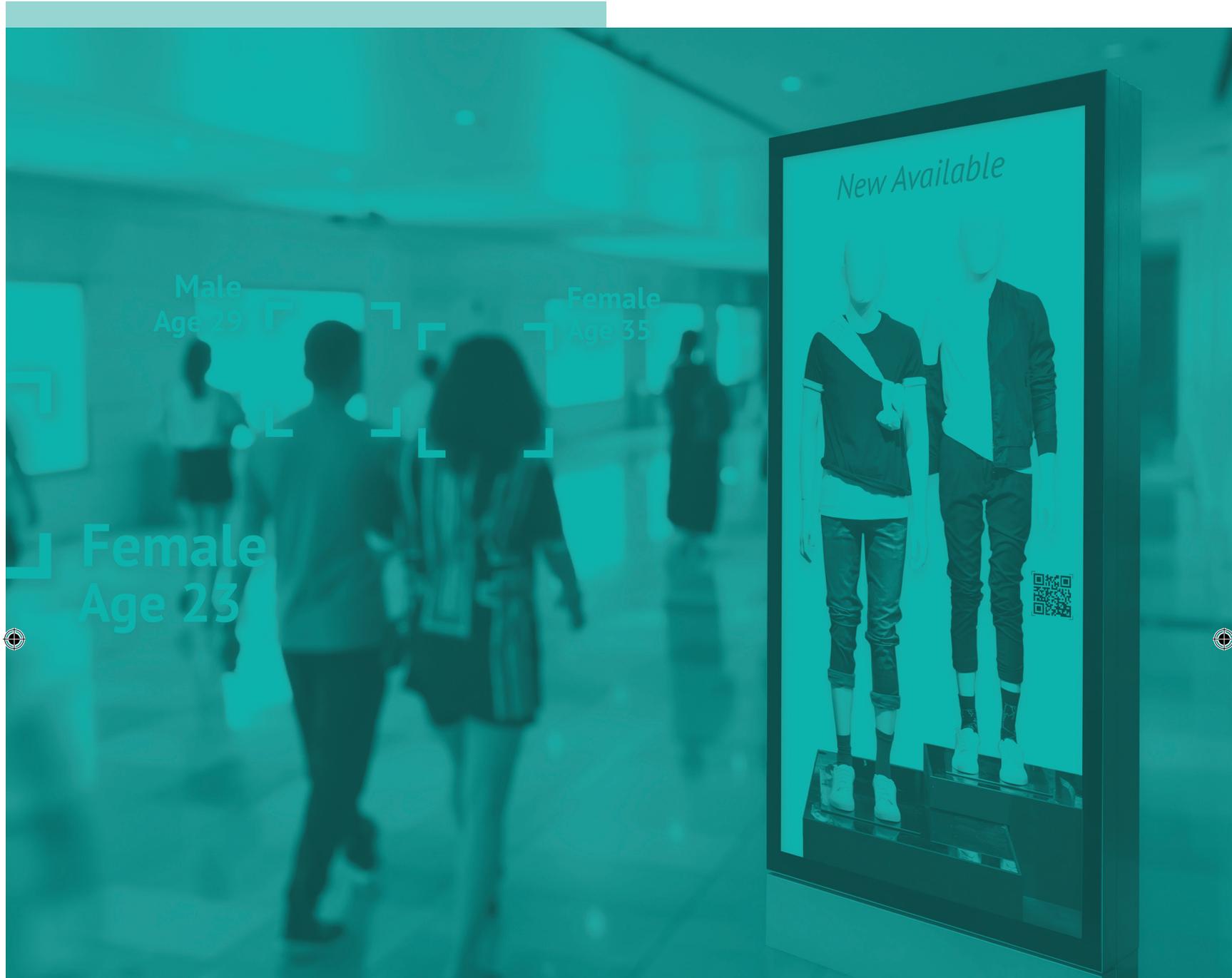
In summary, here are the things to consider for data monetization:

- Look for ways to increase the under-utilized value.
- Uncover the potential of the enterprise data and monetize it.
- Create new opportunities and revenue streams.
- Identify conditions for data monetization, such as massive volumes of structured and unstructured data.
- The role of third parties in data monetization.
- Refine and transform data in usable format.
- Develop a data monetization strategy, such as strategy to provide insights into consumer behaviour, strategy to deal with high-frequency transactions, strategy to deal with overlapped data, etc.
- Adopt or refine existing internal data infrastructure that can be paired with an external data infrastructure in order to marry data together to derive new insight.

2 Marketing

Big data can be used to understand people's behavioural patterns and buying intentions. Although there has been increasing innovation in this field, issues emerge in the areas of data ownership and privacy. Regulations are fragmented across different industries and geographic regions and often lag technology changes. Moore [19] found that negative press from privacy breaches is more likely to impact company decisions than regulations. On the other hand, while consumers value their privacy, many will give away their personal information in return for free access to a service or for financial benefit; examples include initial membership sign up discounts for online stores. Companies need to abide by privacy regulation development and determine their risk tolerance in relation to how they want to monetize customer data. In this chapter, we will first present the advantages of big data in retail marketing with specific examples and case studies, then, we will outline the general opportunities and challenges of big data within the marketing sector.





Male
Age 29

Female
Age 35

Female
Age 23

New Available



2.1 A Case Study: Retail Marketing

Opher [20] explains how retail stores are constantly experimenting with Wi-Fi, Bluetooth and camera technologies to monitor movements of customer segments. To understand these segments and improve marketing initiatives, companies are willing to pay for data. In a highly competitive market, brands are shifting away from mass promotional offers and instead leveraging personalization to nurture the loyalty of the customer segments that promote long-term profits. Large volumes of transactional data exist that enable retailers to perform Market Basket Analysis. This predictive technique identifies products that consumers frequently purchase and helps retailers predict the items they would buy next. Among the many strategic insights this technique provides, retailers can learn how to optimize promotional offers, in-store product placement and e-commerce design.

As part of this study, Tess McDonald, a student of Big Data Analytics certificate program at Ryerson University [20.c], performed a case study by applying data analytics techniques on a data set for an online retailer. The author evaluates consumers' shopping behaviour based on transactional data and identifies the unique products the retailer should recommend to their distinct customer segments.

2.1.1 Data sets

Tess McDonald's study used a UK online retail store's data set provided by the University of California's Machine Learning Repository. The data set includes 395,000 records and over 18,000 unique transactions that represent the shopping behaviour of 4,000 customers during a one year period (Dec. 2010 - Dec. 2011).

RFM analysis (which stands for Recency, Frequency, Monetary value) depends on multiple attributes such as a customer identifier key, transaction date and purchase amount. These attributes correspond to CustomerID, InvoiceDate and TotalSpend. TotalSpend is calculated by multiplying Quantity by Unit- Price and then aggregating the values that correspond to each single customer by using InvoiceNo. The TotalSpend attribute identifies the total expenditure per transaction.

Market Basket Analysis requires transaction and product identifiers in order to calculate association rules. Attributes InvoiceNo and StockCode are used where StockCode and Description can be linked to one another to identify names of each product.

Market basket analysis (association rules based-pattern mining) can help in discovering commonly occurring patterns in the data.

2.1.2 Data Preparation: Understanding Data Characteristics

Summary statistics were calculated to identify outliers in the data. A negative TotalSpend value represents a product return, and these values were included in the RFM segmentation data because customers are segmented based on their average spend per transaction. Customers who spend a lot but later return the merchandise do not offer significant value to a retailer and this behaviour needs to be considered when identifying a retailer's most valuable customers. The average TotalSpend in the data set is approximately \$430, but initial values ranged from \$-7,000 to more than \$70,000. To eliminate extreme values, the average and standard deviation of TotalSpend were calculated. Any transaction that had a TotalSpend of less than -\$3,000 or greater than 3,862 (3 standard deviations away from the mean) were eliminated to avoid skewing the data.

2.1.3 Methodology and Results

This study was performed in two phases:

- Phase 1: Customers were segmented into distinct groups by calculating RFM metrics that determine Customer Lifetime Value (CLV).
- Phase 2: Market Basket Analysis was performed on each customer segment to identify unique shopping behaviour types. These insights could be leveraged to increase marketing personalization, as well as the retailer's profits and the competitive advantage.

2.1.3.1 Phase 1: RFM Analysis

The analysis has been started by using a clustering technique, such as RFM, which segments each data record into several groups with similar types of records. The RFM Analysis identified three distinct customer segments in the data set. These customer segments were assigned to Customer Lifetime Value (CLV) stages and were based on purchasing recency, frequency and purchase value. This stage of analysis shows that there are variations in how customers shop with the retailer - how long ago, how often, and how much they spend when they shop. This phase does not suggest what customers will buy and if customer segments demonstrate loyalty to different products. A common assumption is that the retailer's most valued and expert customers are those who shop often (recency) and have been loyal to the retailer for an extended period of time (frequency). Table 2.1 depicts the distribution of recency and frequency of all customers among the three CLV. The three CLV are "best customers / Experts", "New Customers / Beginners" and "One Time Customers / Beginners".

Table 2.1 shows that the best customers make frequent purchases of about 24 times and bought something in around the last eight days. The author can infer that best customers have a high probability of making

another purchase in the future and hence the retailer should personalize a marketing strategy to fit this shopping behaviour.

New customers have been shopping with the retailer for a shorter period of time and hence, have lower recency and frequency rates compared to best customers. However, they make purchases at a similar rate with regards to frequency as the best customers segment. A targeted marketing strategy for this customer segment has potential to develop brand loyalty and enable the new customers to transition to the retailer's best customers segment over time.

One-time customers segment can also be referred to as the retailer's beginner level customers. These customers have made the least number of purchases and they do not make purchases regularly. It is possible that these customers only make a purchase when a preferred product is on sale, or these customers were not satisfied with the retailer's product following their first purchase and decided to switch to a competitor. Whatever the reason, these customers do not offer significant value to the retailer at the moment but by understanding the shopping behaviour of this customer segment, the retailer has the potential to develop this segment's brand loyalty and build a foundation for long-term profitability.

Table 2.1: The distribution of recency and frequency of customers among the three CLVs

CLV Stage	Recency	Frequency	Monetary	Records
Best/Expert	8.8 Days	24.2	\$398.96	224
New/Beginner	41.3 Days	3.9	\$374.34	2866
One-Time/Beginner	219.3 Days	1.9	\$202.80	1275

2.1.3.2 Phase 2: Market Basket Analysis

Phase two provides the retailer with insights required to develop a personalization strategy that will enhance their competitive advantage. The retailer can use Market Basket Analysis results to create personalized e-commerce experiences by featuring a customer segment's most frequently purchased products on the homepage and by suggesting segment-specific complementary products in an item's "You May Also Like..." section. This type of analysis uses association rules to calculate the probability of two or more items co-occurring in a sales transaction. Association rules use the Apriori algorithm which consists of two metrics being the support and confidence. Support identifies how often a set of items occur together in the transactional data while confidence depicts how likely a transaction will include a complete set of items given that one or more items is already present.

Market Basket Analysis of this data set also determined that the Top 10 products purchased by each customer

segment are statistically different. The Top 10 products purchased by each customer segment are listed in Table 2.2. The fact that the Top 10 products purchased by each customer segment are different suggests that each segment is loyal to a distinct set of products and that the retailer can leverage personalization to improve consumer engagement through promotions.

The top three products purchased across all transactions are the White Hanging Heart T-Light Holder, Regency Cakestand 3 Tier, and Jumbo Bag Red Retrosport. To take this analysis further, the author identified what products are complimentary to the top three products purchased across all transactions, and found that complimentary products for each of the top three brands are statistically different between the three customer segments. An example of Market Basket Analysis is shown in Table 2.3 and depicts which complimentary product a customer in each segment will likely add to their shopping card after initially adding White Hanging Heart T-Light Holder StockCode:85123A.

Table 2.2: Top 10 purchased products by customer segments

Top 10 Products Purchased by "One-time Customers"	Top 10 Products Purchased by "best customers"	Top 10 Products Purchased by "best customers"
White Hanging Heart T-Light Holder	White Hanging Heart T-light Holder	Jumbo Bag Red Retrosport
Regency Cakestand 3 Tier	Regency Cakestand 3 Tier	White Hanging Heart T-light Holder
Party Bunting	Jumbo Bag Red Retrosport	Regency Cakestand 3 Tier
Assorted Colour Bird Ornament	Assorted Colour Bird Ornament	Lunch Bag Red Retrosport
Rex Cash and Carry Jumbo Shopper	Party Bunting	Party Bunting
Set of 3 Cake Tins Pantry Design	Lunch Bag Red Retrosport	Small Popcorn Holder
Natural Slate Heart Chalkboard	Paper Chain Kit 50's Christmas	Lunch Bag Black Skull
Heart of Wicker Small	Postage	Lunch Bag Suki Design
Victorian Glass Hanging T-Light	Jumbo Bag Doiley Patterns	Set of 3 Cake Tins Pantry Design
Jam Making Set with Jars	Set of 3 Cake Tins Pantry Design	Jumbo Bag Doiley Patterns

Table 2.3: Market Basket Analysis

Complimentary Products for "best customers"	Probability	Complimentary Products for "New Customers"	Probability	Complimentary Products for "One-Time Customers"	Probability
Red Hanging Heart T-Light Holder	23%	Red Hanging Heart T-Light Holder	24.10%	Red Hanging Heart T-Light Holder	19.40%
Jumbo Bag Red Retrosport	16.20%	Wooden Picture Frame White Finish	17.50%	Candleholder Pink Hanging Heart	18.40%
Lunch Bag Red Retrosport	15.40%	Heart of Wicker Large	16.10%	Heart of Wicker Large	15%
Party Bunting	14.10%	Natural Slate Heart Chalkboard	16.10%	Heart of Wicker Small	15%
Wooden Picture Frame White Finish	13.70%	Candleholder Pink Hanging Heart	15.70%	Assorted Colour Bird Ornament	13.30%

Table 2.3 used association rules to predict unique customer behaviour depending on the customer segment and can be defined as follows: "If a best customer adds StockCode:85123A (White Hanging Heart T-Light Holder) to their shopping cart there is a:

- 23% probability that the best customer will add StockCode: 21733 (Red Hanging Heart T-Light Holder) to their shopping cart next.
- 16.2% probability that the best customer will add StockCode: 85099B (Jumbo Bag Red Retrosport) to their shopping cart next.
- 15.4% probability that the best customer will add StockCode: 20725 (Lunch Bag Red Retrosport) to their shopping cart next.
- 14.1% probability that the best customer will add StockCode: 47566 (Party Bunting) to their shopping cart next.
- 13.7% probability that the best customer will add StockCode: 82482 (Wooden Picture Frame White Finish) to their shopping cart next. "

The author also developed a list of substitute goods for each of the Top 10 products purchased by each customer segment. The process used by Ho [21] involves a two-step approach. Specification of the level of similarity between the original product and the alternate products is required and based on the product chosen by the user and the desired similarity level to the chosen product. The process employs an algorithm to retrieve alternate products that satisfy the user requirements when compared to the original product. The alternate products are ranked in order of their similarity to the original product and sorted to give them an index value. The alternate products are then ranked based on the index. Finally, the list of alternate products are presented to the user.

2.2 Opportunities and Challenges

Through the conclusions derived from big data analytics, such as the ones derived from the case study above, marketers are in a better place to market their products more efficiently. In his 2016 article, Goldfein identifies four ways that marketers can use big data to their advantages [22]:

- **Segmentation:** Classifying customers into groups such as age, location or gender in order to tailor marketing campaigns and thereby improve the marketing strategy.
- **Identification:** Identifying what actions buyers take in an attempt to improve sales strategies
- **Evaluation:** Assessing and comparing the success of different campaigns in order to identify what campaigns are most efficient given the desired marketing campaign goals
- **Understanding behaviour:** analyzing customer data to better understand how people use a particular product, and in turn, using this information to improve marketing strategies

By applying these methods, marketers can personalize marketing strategies to customers which may increase sales. In fact, in his article, Goldfein mentions that “according to MarketingSherpa Marketers, neglecting the development of personalized lead nurturing can expect as many as 79 per cent fewer of their leads to convert” [22]

Nonetheless, common big data problems exist in the retail marketing sector. The main problem relates to the quality of data. Most marketers do not assess the quality and accuracy of their data, and many use data that is invalid or incomplete which affects the understanding of the customer. A second problem is associated with the overwhelming scale of data available. The extensive amount of data available to marketers makes it difficult to derive meaningful insights. To overcome this problem, marketers need to have good data management and maintenance, as well as a unified data storage and quality control. In fact, when data is not managed and handled well, it can cause adverse effects for marketing companies, including reduced customer satisfaction and retention rates, distortion of success metrics, higher failure rate with marketing initiatives, and negative comments on social media. Privacy issues also have important concern in marketing [23]. Another problem in the marketing area is related to the lack of skilled data workers. There is a shortage for workers with the appropriate skill for handling handle data volume, velocity, and variety, as well as those who can deal with the three “perspectives” - descriptive, predictive, and prescriptive analytics. [24] categorize this problem as an education challenge, and suggests that, to overcome this challenge, educational institution must consider adding a new stream of courses into marketing specializations, such as data mining, text mining, opinion mining, social media/network analytics, web mining, and predictive analytics.

2.3 Conclusion

The case study by Tess McDonald [20.c] provided two important insights. First, customers demonstrate variable patterns in how they shop with a retailer such as how long ago they made a purchase, how many times they have made a purchase and how much they have spent on each purchase. This allows a retailer to assign customers to unique customer segments based on the Customer Lifetime Value (CLV). Second, the way a customer shops influences what they buy which proves that personalization is a strategic advantage when nurturing customer loyalty in a highly competitive market. The author also discussed when the personalized recommendations should be delivered and the most efficient way of presenting promotions to different types of customers. The findings point to marketers' ability to harness big data to improve and personalize their marketing strategies, therefore increasing their sales revenue potential. However, marketers also face challenges, mainly related to the quality of data, the lack of technical skills in the labour market, and privacy concerns.

Personalization is a strategic advantage when nurturing customer loyalty in a highly competitive market. It can be leveraged through RFM analysis and association rule mining.



3 Finance

Banking, insurance and other financial companies are rich in customer and transactional real-time data that are only beginning to be utilized as business assets. The next evolution in the world of finance and banking will come from monetizing real-time data to derive profitable insights. But the challenge lies in creating capabilities to operationalize real-time data insights fast enough to be valuable to the business.

There are many examples big data applications being developed in and for the financial sector. For instance, Malcolm White, a student of Big Data Analytics certificate program at Ryerson University [25.c], has developed a way to analyze financial conference call transcripts to better understand the economic condition of companies. This new technique is described in detail in this chapter. Other potential applications for big data could mitigate customer churn risk in banking sector and help to build customer bases for insurance companies. We will discuss these three examples in further detail in Opportunities section of this chapter to demonstrate the opportunities and gains that the financial companies can derive from big data. Then, we will look into the challenges.





3.1 Opportunities

As mentioned in previous chapters, data represents a valuable resource that companies can harness for competitive insights but that remains largely under-utilized. Financial companies can use big data insights to generate sales leads, improve products, create innovative channels and technologies (such as mobile applications), adjust pricing and increase customer loyalty, all leading to an increase in their revenues [25]. In this section, we present a case study on the automated analysis of conference call transcripts that can help to predict how well companies will perform, as well as several big data applications in the banking and insurance industry.

3.1.1 Case Study: Analysis of Financial Conference Call Transcripts

Industry professionals rely heavily on quarterly financial conference calls and transcripts to understand a company's performance and outlook. Sentiment is manually derived by looking for cues in what was said in either the prepared remarks from management and the question and answer session initiated by industry analysts. Trillions of dollars in assets and billions of dollars of trading flow use these calls as their primary source of updated information pertaining to a company or economy.

As a key part of this study, Malcolm White [25.c] developed a method to automate sentiment analysis for conference call transcripts using Natural Language Processing and machine learning techniques. Natural Language Processing was used to extract sentiments and a machine learning classifier was used to automatically predict the economic condition of the corporation using the sentiments.

The primary data sets used are financial conference call transcripts posted on the Seeking Alpha website (<http://seekingalpha.com>). A North American corporation typically hosts four conference calls a year that update the market on their economic progress and performance. The top 250 companies in technology were selected [for this study] based on inclusion in the NASDAQ, S&P 500 or the Russell 2000 index with a market capitalization over US\$150M. After some data was filtered out, this data set included approximately 2,300 conference calls. Bloomberg data was used to collect trading behaviour of the corporations around the period where they announced updated financial results. Market behaviour at the time of release was recorded using the price volatility of the stock as a proxy. The market reaction, either positive, neutral or negative, was used as a label for the data. Transcript text was subsequently preprocessed using natural language processing techniques (e.g., text mining technique) to create a financial dictionary that was then trained and tested with a Naive Bayes classifier. This approach classified the economic condition of a company with a maximum accuracy approaching 70%.

The ability to quickly quantify economic sentiment using machine learning adds new analytic capabilities for financial practitioners. Tens of thousands of financial documents can now be summarized for sentiment in minutes, something that is not possible with manual techniques. Economists can use these insights to supplement their understanding of current economic conditions with data points that reside outside of their surveys. Asset managers can see trends at an aggregated industry level and adjust their asset

allocations accordingly to optimize their portfolios. Dynamic visualizations can be created that best illustrate economic conditions and their corresponding trends. The author conservatively estimates that this approach can generate tens of millions of dollars of economic value for the Canadian financial industry.

3.1.2 Monetizing Data in Banking

Banking data is divided and owned by the commercial, consumer, retail banking and mortgage sectors. But banks are realizing that to fully utilize and exploit the value of data, data silos must be broken down. Numerous, complex factors are putting pressure on revenue gains and reducing traditional means of revenue. Risk and regulatory data management are high priorities in Canada's heavily regulated financial companies.

Marr [26] states that Citibank has integrated open source data platforms with their big data solutions. Citi works closely with open source vendors that demonstrate high integration capability with the technology provided by Citi. In another example, the commercial banking division at a large Canadian bank realized that by using external data they could create a more granular picture of the landscape outside their existing client base. Rather than tailoring banking products and sales strategies on internal data, the bank looked to new sources of information that would enable them to expand their market penetration, leading to the addition of over 60,000 prospects to their pipeline quickly.

Customers of financial services have a very low tolerance for poor customer service. Shaeffer [27] states that equipping call centers with big data tools is an efficient tactic for detecting when a customer is unhappy. Using data this way can identify at-risk customers for intervention and increase retention. Techniques such as text and speech analytics, as well as Natural Language Processing, help institutions mine and filter large volumes of customer content in order to identify those that suggest a business problem or opportunity. For example, negative sentiment calls near a renewal date could alert customer service intervention in order to avoid losing the renewal and customer. This can be accomplished by applying sentiment analysis to unstructured data types such as real-time speech, voice recordings, customer comments left on self-service applications, etc. Based on the customer's words and product portfolio, streaming sentiment analysis interprets product sale opportunities and delivers real-time suggestions to call center agents or financial adviser. Integrating existing customer profiles with their social profile and website browsing history can deliver personalized and highly relevant up-sell and cross-sell recommendations that meet customers' real needs. Technology investments are required to build this capability but this method will convert the call center to a profit center by providing revenues from cross-sell and up-sell, increased referrals from more satisfied customers and increased customer lifetime value from customer retention.

3.1.3 Monetizing Data in Insurance

Insurance companies have been exploring ways to provide personalized policies by monitoring customer's day-to-day tasks. For example, by using a sensor in the car or mobile app, a car insurance company may offer proof of the driver's insurance policy, provide available parking spots, offer roadside assistance, file claims, schedule vehicle inspections, and reserve rental cars. Data gathered through applications will help sales correctly estimate auto-insurance rates by offering discounts and rewards to improve customer retention and tailoring products and premiums based on accurate risk factors and customer car usage. By comparing a driver's daily data to other driver, claims, actuarial and policy data, insurances companies can lower or adjust premiums for individuals to account for behaviour, history and risk. Schaeffer [27] describes a U.S. insurance company that offers mobile application downloads to non-customers that provide benefits such as real-time traffic displays, cheap gas locations, parking finders and trip information. This application also provides an estimate of auto insurance rate a driver would receive if he or she were to switch auto insurance providers. This method achieved a 7 percent conversion rate from non-customers for the insurance provider. Moreover, these new customers are considered the lowest risk and most profitable. Similarly, he [27] also proposes that life insurance

providers can monitor customer's lifestyles through wearable technology or social data. Companies are wary that this level of monitoring may not be accepted or permitted by most customers as it would be viewed as an invasion of privacy; however, they may select the level and type of monitoring. The following list created by Schaeffer [27] describes how big data can be used to increase the customer base of insurance providers.

- If an insurance company learns from social data that a customer made an expensive jewelry purchase recently, it could identify that a customer recently experienced a significant life event. In this event, being the first to offer a relevant insurance policy deal (e.g., new policy or incentives on existing policy) can increase customer conversion rates.
- If social or household data reveals that an auto policy customer has a child turning 16 soon, an offer aimed at adding additional drivers to the policy, first time credit cards or college tuition savings accounts can further increase customer share.
- If an existing customer queries 401K plans on the company website, an insurance provider can end the next call center question with an offer to speak to a 401K specialist.

3.2 Challenges

A key issue that is found in all sectors is concern for privacy. Armah [28] highlights the importance for financial companies to ensure they do not breach personal privacy laws or ethics or use personal information when collecting data for analytical purposes. Armah [28] suggests the need for “a balanced regulatory framework to address these concerns.”

To make the best use of its data, financial companies must also combine the efforts of both the business and IT professionals [25]. This step towards inter-departmental collaboration could be somewhat challenging as it is a new approach to operations in financial companies. Methodological and technical constraints present still more obstacles to big data analytics. Armah [28] describes several difficulties in selecting unstructured data for analytics and choosing the best analytical tools needed for specific contexts, especially as data analytics methodologies rapidly evolve.

Moreover, population subsets sampled for analysis must be representative of the desired population to achieve accurate insights.

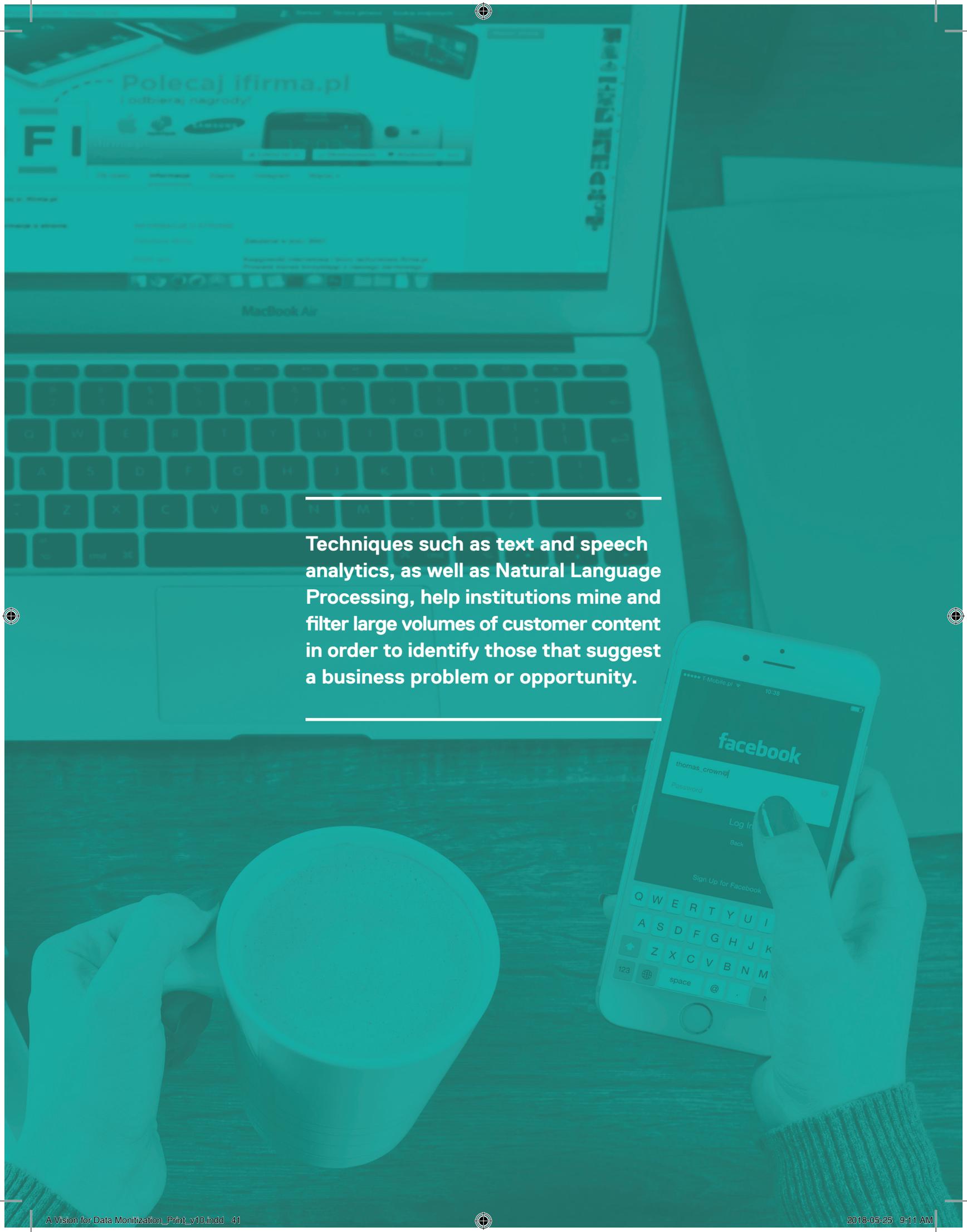
Finally, most of the existing big data is archived in a way that is difficult to access and analyze. Armah [28] states that it is important to overcome this obstacle and improve access to data in order to gain the most value of it. The Government of Canada and the

Government of Ontario have together taken a step towards facing this obstacle. Through a collaboration with IBM and a consortium of seven universities, a new Ontario-based data center, worth 210 million has been established to help researchers use high-performance cloud computing infrastructure to better exploit big data [28]

The ability to quickly quantify economic sentiment using machine learning adds new analytic capabilities for financial practitioners. Tens of thousands of financial documents can now be summarized for sentiment in minutes, something that is not possible with manual techniques. Economists can use these insights to supplement their understanding of current economic conditions.

3.3 Conclusion

Machine learning and Natural Language Processing techniques have already started leaving an impact on financial data analysis. Organizations are analyzing financial transcripts to predict the economic behaviour of the company in future. Personalized up-sell and cross-sell recommendations generated by integrated customer profiles, social profiles, and website browsing history. Insurance firms are using driving history to recommend low premium rates to good drivers and even win new customers through the navigation applications. In short financial organizations are monetizing data in a variety of ways. However, to make the most of big data, the financial industry should implement strategies of collaboration between their business and IT professionals, access to data should be improved, and responsible privacy measures must be taken into consideration and followed.



Techniques such as text and speech analytics, as well as Natural Language Processing, help institutions mine and filter large volumes of customer content in order to identify those that suggest a business problem or opportunity.

4 Real Estate

4.1 Introduction

2016 was a record breaker for the Canada's housing market. According to media reports, the average house price in Toronto increased by \$122,000 in a single year, to about \$730,000" [29]. Rapid price escalation in this market calls for more effective analyses of market conditions and with the large amount of real estate data available, big data can provide new insights into the real estate sector. In this chapter, we will first explore how the real estate sector can harness big data to its advantage, and then we will look into some challenges. Finally, we present a case study of Toronto's real estate market using big data, performed by Tess McDonald, a student of Big Data Analytics certificate program at Ryerson University [29.c], specifically for the white paper.





4.2 Opportunities and Challenges

One noticeable advancement that big data has brought to the realty market is the ability to make more rational investments, with higher accuracy in prediction. According to Lohr S. [30], the use of big data has proved to generate a higher accuracy level when it comes to economic forecasting. This has been demonstrated by Google searches of housing prices and sales predictions [31]. Through the analysis, modeling and manipulation of data available in Google search engine, researchers were able to estimate the demand and supply equilibrium and estimated price index in the real estate market with high accuracy [33]. In fact, they find that their model predicts future US home sales 23.6% more accurately than the National Association of Realtors. As per Danyange Du et al. [32] “Google is several times the efficiency of the government at a fraction of the cost”.

That being said, big data allowed for better decision making when it comes to realty investment, but it has also allowed for more innovative investments. With the emergence of Internet and the growth of big data, a broad range of data that is not limited to real estate information, has been made accessible to companies (i.e., buyers’ habits and customs, preferred travel routes, etc.). Through data mining, Fantasia company, a real estate company located in China [33] plans to use vast data of home buyers to create a community platform to identify business and investment opportunities. As such, the vast amount of data available has broadened the investment possibilities for real estate companies. Other applications of big data realty development are stated in Table 4.1 [32].

Table 4.1: Applications of Big Data Realty Development

Enterprise	The big data resources	Relata development and investment
Google	Key words	estimate the demand-supply equilibrium and predict the price index in realty market by analyzing the relationship between key words and the data of housing price, providing strong support for rational developments
Vanke	land resources	analysis big data of land resources to deal with the rising land price
Wanda, Greenland	realty development	reveal potential value of big data for diversified investment
Fantasia Group	buyers’ requirement	build community e-commerce creatively and expands its big data business to the financial sector, hotel services, culture, and tourism
Vanke	owners’ personal information	put forward the concept of building city supporting services
Shimao Group	owners’ health conditions	introduce the “health clouds” business management to its property owners for the health monitoring and advisory opinion
Godland, Greenland	owners’ personal information	open up new operations such as Intelligent City and Cloud Service
Windermere	information for from drivers’ GPS	plan for the potential buyer with there commute routes and the cost of time

In addition, providing insights for better decision-making, big data analytics has been able to improve the effectiveness of realty sales promotions. With precise data, marketing can be made more effectively targeted to specific audiences based on their locations. Figure 4.1 [32] explains how big data can be used in the realty marketing process. For example, Hoawu, a third-party marketing platform located in China, amassed a big data warehouse of home buyers' personal information and then harnessed this information to match buyers with houses according to their demands, using an algorithmic function.

Use of big data is new in the realty market and it is developing quickly and new data sets being generated almost every day. But there are challenges; most data available in the real estate sector is in an unstructured form where relevant information is scattered throughout documents, and data storage and processing [34] issues are being observed.

Privacy concerns are another major challenge that recurs in other sectors. The use of personal and private information in big data can put individuals privacy in jeopardy. Danyange Du et al. [32] finds that even without personal information being included, big data analytics can recognize a personal identity with a recognition rate of more than 99 per cent. Clearly, privacy issues are one of the main concerns for big data analytics use.

Finally, Danyange Du et al. [32] sheds a light on the challenges originating from the complexity and metastability of the real economy, which affects the real estate market and can thus put at risk the application of big data. Estate companies should therefore have a sound judgment when taking advantage of big data.

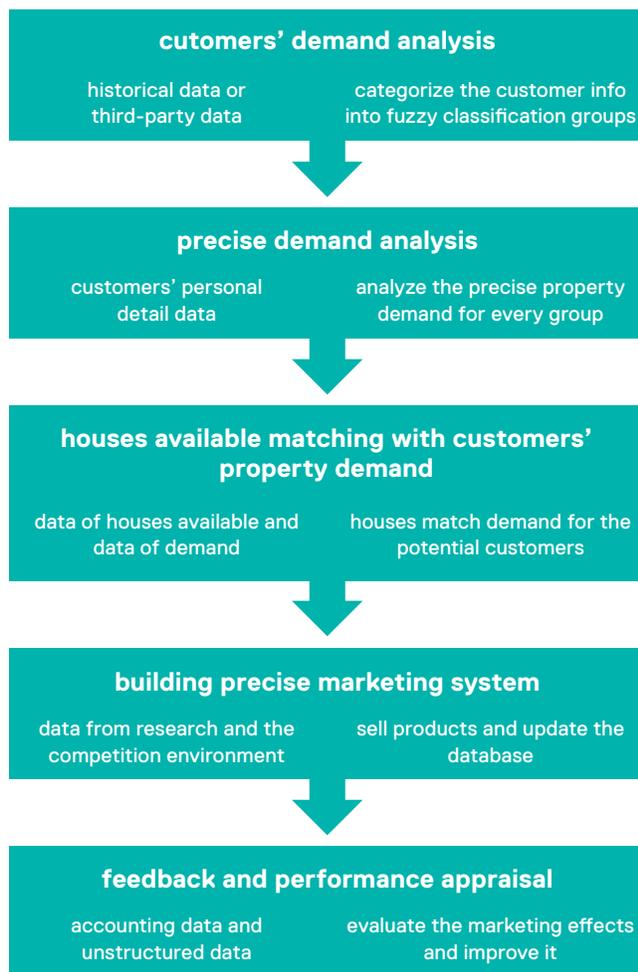


Table 4.1: Realty Precise Marketing Flow Chart

4.3 A Case Study: Toronto's Housing Market

Toronto's housing landscape has changed dramatically in recent years. The City of Toronto has published the data of many Business Improvement Areas (BIAs) that include local business owners along with the demographic information about many neighbourhoods. In this study, Tess McDonald [29.c] uses data mining methods to uncover the relationship between demographics and business types present in a neighbourhood. The author used Apriori Association Rule mining algorithm to extract these patterns. To perform this pattern mining, the author first mapped business types to neighbourhoods and then used the most useful attributes in the analysis. This analysis identified patterns that can help business owners find optimal locations.

4.3.1 Unemployment Rate Effects

Pattern: If 20-39% of a neighbourhood has low-income, then there is a 65% probability that 10-24% of the businesses in that neighbourhood are in the Retail Trade category. Pattern from [35]:

4.3.2 Senior Population Effects

Pattern: If 25-39% of a neighbourhood is aged 55 years or older, and 25-39% of the neighbourhood is youth, then there is a 94% probability that up to 19% of businesses in that neighbourhood are in the Accommodation and Food Services category.

4.3.3 Housing Occupancy Effects

Toronto is a diverse city that is expected to grow to over three million people by 2031. With population growth comes an added demand for housing. Pattern: Of the new households in the City over the past 15 years, nearly 70% were in high-rise apartments and condominiums. The shift to high-rises was most common among those under the age of 44 years of

age, while the other age groups maintained their share in ground-related housing.

4.3.4 Pollution and Poverty Effects

Pattern: Poverty maps identify seven Toronto neighbourhoods that have high releases of toxic pollutants and poverty rates above the national average (11.8%), and 17 neighbourhoods that have high releases of combined air pollutants and poverty rates above the national average.

4.3.5 Crime Rate Effects

Some parts of Toronto are reported to have higher crime rates than other parts. Data to determine the following patterns used police-reported statistics with neighbourhood characteristics for 2006. This combination allows us to better understand how the complex social geography and demographics of Toronto is related to crime rates.

- Pattern from Wallace [36]: If an area is highly populated and has high commercial activity, it also has a high probability of having a high crime rate.
- Pattern from Wallace [36]: Areas near the downtown core, east and northwest of Toronto which corresponds to where residents earn the lowest per capita income experience the highest violent and property crime rates. In those areas, the most concentrated areas of crime are in Danforth, east side of downtown, Lawrence and Morningside, Jane and Finch and Jane and Eglinton intersections.
- Pattern from Wallace [36]: In North Toronto along Yonge Street residents earn a higher income experience and lower violent crime rates than the average.

- Pattern from Wallace [36]: Neighbourhoods with a high rate of violent crime are more densely populated and have a higher percentage of residents living in multi-unit dwellings. They also have the highest percentages of children (under the age of 15), renters, single-parent families and visible minorities. The residents of these neighbourhoods are also less likely to have a university degree, more likely to earn a lower wage, and more likely to live in low-income households.
- Pattern from Wallace [36]: A neighbourhood resident's economic vulnerability and the neighbourhood's available socio-economic resources are proportional to the level of violent crimes.
- Pattern from Wallace [36]: These results are consistent with previous research studies that have found that a lack of access to socio-economic resources and the "urban" character of neighbourhoods impede social control of crime by limiting social cohesion and sense of belonging.
- Pattern from Wallace [36]: Level of income in a neighbourhood is closely associated to violent crime rates. However, violent crime rates are highest in "urban" neighbourhoods with high population density and mobility.
- Pattern from Wallace [36]: Areas characterized by greater social control with a high percentage of immigrants and elderly people see a lower rates of crime compared to neighbourhoods with equal access to socio-economic resources.
- Pattern from Wallace [36]: A neighbourhood's physical environment such as the number of buildings that require repairs is directly associated to high crime rates in that area.

The following patterns retrieved from Charron [37] and suggest that criminal rates by young people

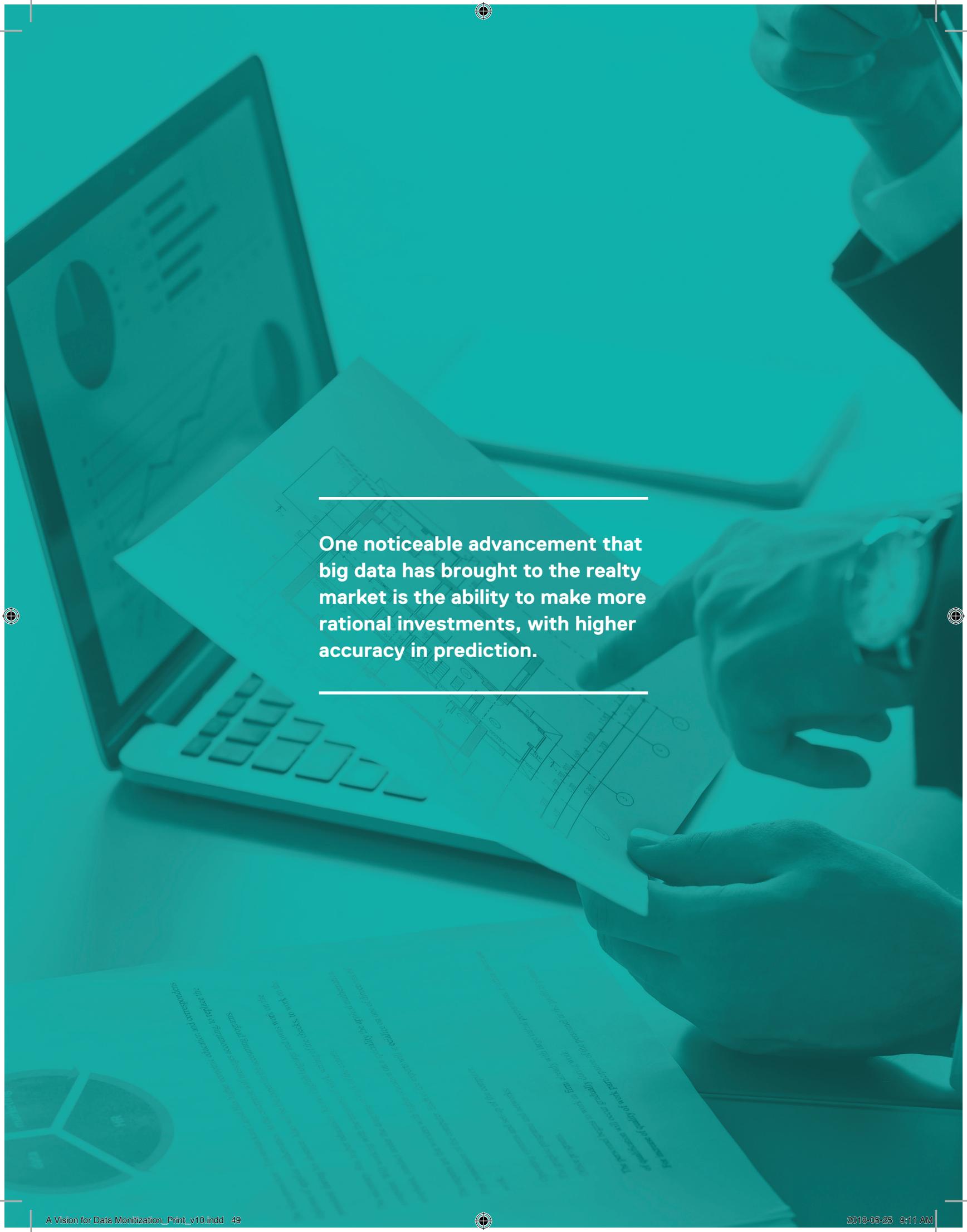
- Pattern: People between the ages of 12 to 17 are influenced by the neighbourhood they live in
- Pattern: There is an association between commercial activity and incidents of youth crime in commercial establishments such as restaurants and convenience stores that are situated in residential settings and fewer crimes in commercial areas such as industrial parks and stores.
- Pattern: Living in a neighbourhood characterized by economic insecurity increases a young person's risk of being accused of committing a crime.
- Pattern: If a neighbourhood has a high violent crime rate, youths are affected and influenced by witnessing the violence by residents in that area.

4.3.6 Demographic Effects

- Pattern: If 50-74% of a neighbourhood has a post-secondary education and 35-49% of couples are living without children, then there is a 79% probability that 20-39% of the businesses in that neighbourhood are in the Accommodation and Food Services category.
- Pattern: If a neighbourhood has 10-19 distinct business types and the dominant ethnicity is English, then there is a 90% probability that only 5% of the businesses in that neighbourhood are in the Arts, Entertainment and Recreation category.

4.4 Conclusion

In summary, big data offers opportunities to increase profitability in the realty market, which are accomplished through better investment decision-making, better prediction and accuracy of the market, new innovative streams of investments as well as a more efficient real estate marketing. Also, through data exploration, such as the one undertaken in our second section of this chapter, we derive new information and insights in the real estate sector. However, as with any other sectors, challenges exist, and these are mainly categorized as: privacy protection, data processing skills, as well as the challenge of the unstable economy and real estate market nature.



One noticeable advancement that big data has brought to the realty market is the ability to make more rational investments, with higher accuracy in prediction.

5 Security

5.1 Introduction

The growth of cloud computing has made it easier for hackers to gain unauthorized access to data. Companies are often the first target of such cyber-criminal activities. As one of many examples, in March 2011, Epsilon company, a Dallas-based marketing firm, was victim of a cyber attack that cost the company between \$255 million and \$4 billion in loss [38]. Organizations generate huge amounts of data daily and the risk of loss and reputation damage through security threats are major concerns. The use of big data analytic techniques can help in detection and prevention of criminal activities, safeguarding enterprises from potential losses. There are also significant opportunities for cyber security companies to monetize their developed analytical models by providing security analytical services to others.

In this chapter, we will present the opportunities of big data analytics in the security sector, along with some of its applications, and then we will explore key challenges.





5.2 Opportunities

5.2.1 Current Effectiveness and Potential

Despite significant potential to use big data analytics to avoid fraud and data theft, some research suggests that big data analytics are, so far, not as effective as anticipated in preventing cyber attacks and intrusions, primarily due to technical challenges. Sentinel One [39] performed a survey and found that 53 percent of companies are using analytics for their overall cyber-defense strategy yet they have been compromised at least once every month. However, less than half of the 53 percent of participants say their efforts are highly effective. Big data will not be effective for threat analysis if it is poorly mined and while meta-data is available, it can be difficult to get the maximum benefit from it. Finding the right people who know how to mine data for trends is another aspect of the problem. The study listed several challenges enterprises face:

- 49% said their systems have been compromised because of volume of data is overwhelming and the analysis cannot keep up
- 33% said they do not have the right systems to collect data
- 30% said the data is stale when it finally gets to a cyber security manager

These findings point to a need to improve the techniques and systems that use big data analysis to prevent and detect cyber threats. These needs translate into opportunities that could be monetized by security companies. And even though outright prevention of security threats has so far proven difficult to achieve with big data, many people see fraud detection (as opposed to prevention) as an ideal application for data monetization. This idea is explored further in Section 5.2.3.

5.2.2 Techniques of Big Data Security Analytics

Terabytes of data are routinely collected from security-relevant sources for compliance reasons and analysis. Hadoop and other big data tools are simplifying the process of analyzing large data sets at remarkable speeds and scales and creating new opportunities to create affordable security measures. BD Analytics [40] states that security analysts are now able to correlate, consolidate and contextualize even more diverse sources of security data for longer periods of time. Although real-time analysis of active threat indicators are easy to miss, this analysis can create new possibilities for predictive models, statistical models, and machine learning when analyzed over time. Big data also helps visualize cyber attacks by removing complexity from various data sources and

simplifying the patterns into visualizations.

Still, the data generated from organizational assets is so big that without appropriate infrastructure and analytics techniques, malicious activities are hard to detect as they happen. This is where security analytics companies can provide data visualization, forensic analysis, anomaly detection and malware detection services to other organizations. Security analytics organizations are monetizing the data generated from other organizations by providing big data SIEM (Security Information and Event Management) services. Telus and Bell are among the leading Canadian organizations providing managed security services.

The two main services that are being monetized through third-party data are anomaly detection and misuse detection [41]. Using historical data, an anomaly detection system can create statistical baselines that serve as “normal” data and help to identify threats that deviate from the norm. To illustrate this method, let’s assume that an employee usually logs in at 9 am, reads email, performs database transactions, and take lunch from 1 pm to 2 pm. If the user logs in at, say, 3 am, the system will flag this activity as suspicious and categorize it as an anomaly. This method helps to detect unknown

attacks only by flagging “unusual” activities but generates high false-positive rates. Misuse detection involves defining a set of attack descriptions, also called “signatures”, that scan against an audit data stream looking for known attacks. Unlike the anomaly detection method, this method generates low false-positive rates [but cannot identify new threats].

5.2.3 Advantages and Applications of Big Data Analytics

Marr [42] states that big data techniques such as advanced analytics and cutting-edge technology like machine learning has been helping to detect hackers, malicious actors and illegal activities such as money laundering. While it has proven difficult to prevent fraud and attacks outright using big data so far, fraud detection is seen by many as an ideal application for data monetization. When fraud prevention fails, fraud detection can be implemented using supervised or unsupervised methods. Unsupervised fraud detection learning would be used when the firm needs to profile observations and detect anomalies. Supervised fraud detection learning uses a database of known fraudulent and legitimate cases to construct a model that assigns scores and ranks new suspicious activity. Marr [26] describes how scanning transactional records to spot anomalies helps to identify incorrect

or unusual credit card charges, which can be easily corrected if spotted early. The only footprints hackers leave is the sequence of commands used to enter and compromise the system and so sequence analysis techniques are used during supervised and unsupervised machine learning. These attacks are sometimes carried out to steal data or as a show of strength to let someone know that the attacker is serious and has cyber-firepower.

BD Analytics [40] also talks about Advanced Persistent Threat detection (APT) which operates using “low-and-slow” mode. This method looks for low-profile, long-term activities that victims may not notice quickly or at all, as well as money laundering activities. To detect these types of fraud, very large amounts of diverse data from internal data sources and external shared intelligence data need to be joined and analyzed. For example, while a single large deposit may not draw attention, analysis of wire transfer data can be very useful in detecting money laundering activity by monitoring transaction patterns over a longer term to identify suspicious behavior.

5.2.4 Cyber Security Workforce Opportunities

Given these potential advantages and the rapid evolution of technology, cyber security professionals are in high demand. Not surprisingly, demand exceeds supply of skilled cyber security workers according to recent research. A few examples captured from Marie A. Wright’s article on ‘Improving cyber security Workforce Capacity and Capability’ are stated below:

- From 2007 to 2013, cyber security jobs have increased by 74 percent, more than twice the rate of all other IT jobs.
- From 2014 to 2019, global demand for security experts is expected to grow annually by 10.8 percent while supply will grow by only 5.6 percent. (Both figures are the expected compound annual growth rates). Based on these numbers, we expect a workforce shortage of more than 1.5 million cyber security professionals in this time frame.

The U.S. government has been paying particular attention to security and has been spending billions of dollars to find and hire the people with the right knowledge, skills and abilities.

5.3 Challenges

5.3.1 Threat Response

As cyber threats become increasingly harmful and prevalent, companies need to take action towards improving cyber security so that they are able to respond to threats effectively in real time. Unless they invest in required expertise and infrastructure, data collection and analysis will prove useless.

The key is being able to automatically respond to threats identified through analysis, and in being able to trust the accuracy of the data[39]. Responding to a detected security threat can take many forms. It can generate an alert about the intrusion, or it can take more intrusive forms such as paging a system administrator, sounding a siren, or escalating it to a form of counter-attack. A counter attack typically consists of router reconfiguration in order to block the attacker's address or even attacking the offender. This latter can be quite risky and challenging if the offender or hacker is initiating a spoofing attack in which he or she successfully conceal his or her address by using someone else's address. In this case, a victim person can be falsely targeted and accused for being a threat to security [41].

5.3.2 Technical Challenges

Dealing with data volume and variety, and the ever-advancing nature of security threats presents a number of technical challenges. Reducing the numbers of false positives generated in detection systems and keeping signature sets (known attack descriptors) up to date in misuse detection systems as new attacks are discovered are key issues. In fact, Marr [42] claims that the struggle of defeating hackers will always remain as people are inclined to steal data for as long as there is technological means and incentive to do so. As security becomes tighter, hackers will become more ingenious, challenging security experts to stay a step ahead. To combat these problems, researchers advocate for a new approach combining misuse detection and anomaly detection. This method is getting popularity among security service providers.

Keeping up with the high-speed networks and high-performance network nodes also represents a challenge. BD Analytics [40] states that the volume and variety of big data is becoming overwhelming and conventional methods of identifying threats, such as analyzing logs, network flows and system

events, are outdated. These methods cannot handle large quantities of data, perform analytics and complex queries on large data sets with noisy features. On top of this, companies are increasingly moving to cloud architectures that are more difficult to fortify.

Bolton [43] suggests making fraud detection more efficient relies largely on the availability of better data and system performance, such that there would be more legitimate records for each fraudulent case and faster and more accurate data processing to catch fraudulent behaviour in real time. Therefore, reducing fraud to its lowest level would require significant effort and cost that must be balanced against the potential savings to the enterprise. All models of fraud detection, whether supervised or unsupervised models, have to be adaptive as fraud patterns and the methods to fight them change constantly.

Kemmerer and Vigna [41] provide two solutions to system performance issues. The first method is to split the event stream into slimmer and more manageable streams for the detection sensors to analyze in real time. The disadvantage with this

method is that when events are randomly divided, sensors may not be able to receive enough data to detect an intrusion. The second method is to use peripheral network sensors. This involves deploying sensors close to the hosts that the system must protect, but managing a highly distributed set of sensors is more challenging.

Computer resources generate huge amount of data which many organizations do not have the capability to monitor for security threats. Security service providers are using this opportunity to monetize their security analytics models (services) by using the data generated from other organizations.

5.4 Conclusion

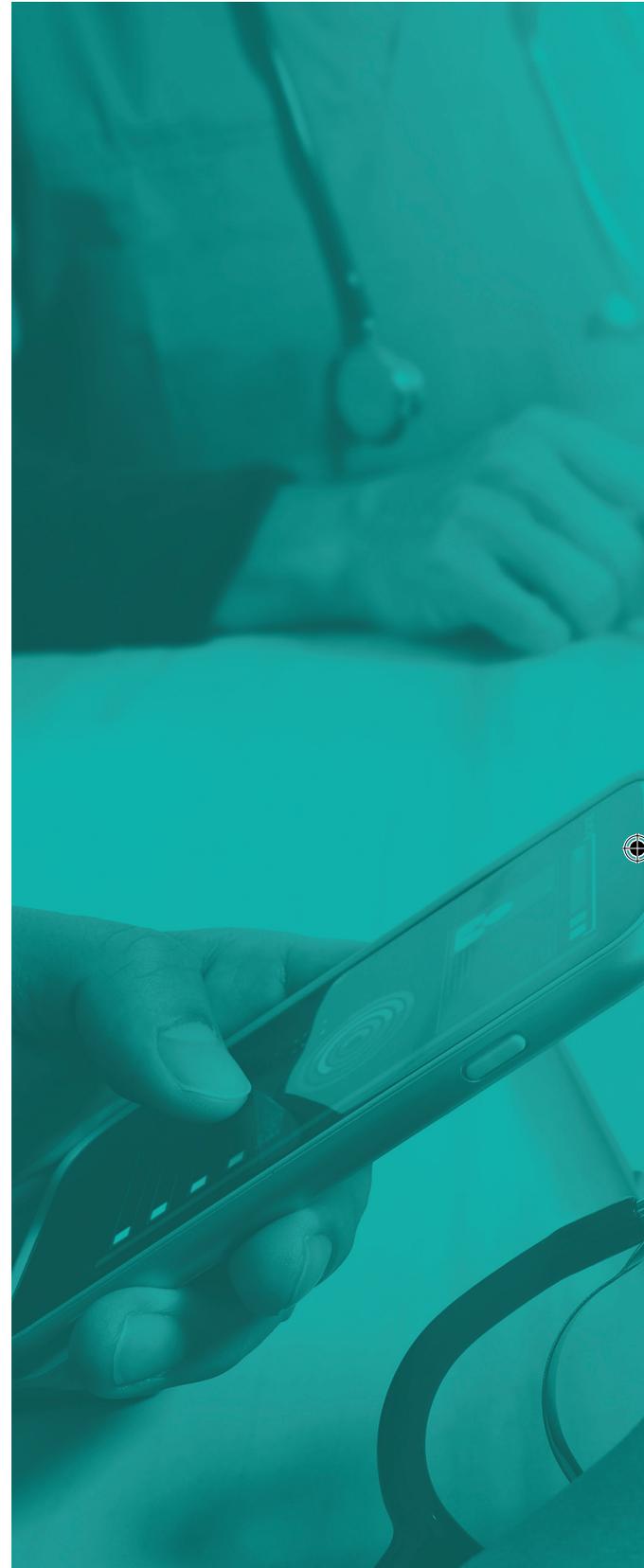
Big data analytics play a major role in preventing threats such as fraud, money laundering and cyber threat security, which, if detected and prevented, can bring implicit but immediate monetary value. However, companies have yet to take action towards improving cyber security, such as detecting real-time threat and immediately responding to the threat because they lack expertise and infrastructure needed for robust data analysis. Security analytics organizations are using this opportunity to monetize their security analytics models (services) by using the data generated from other organizations.

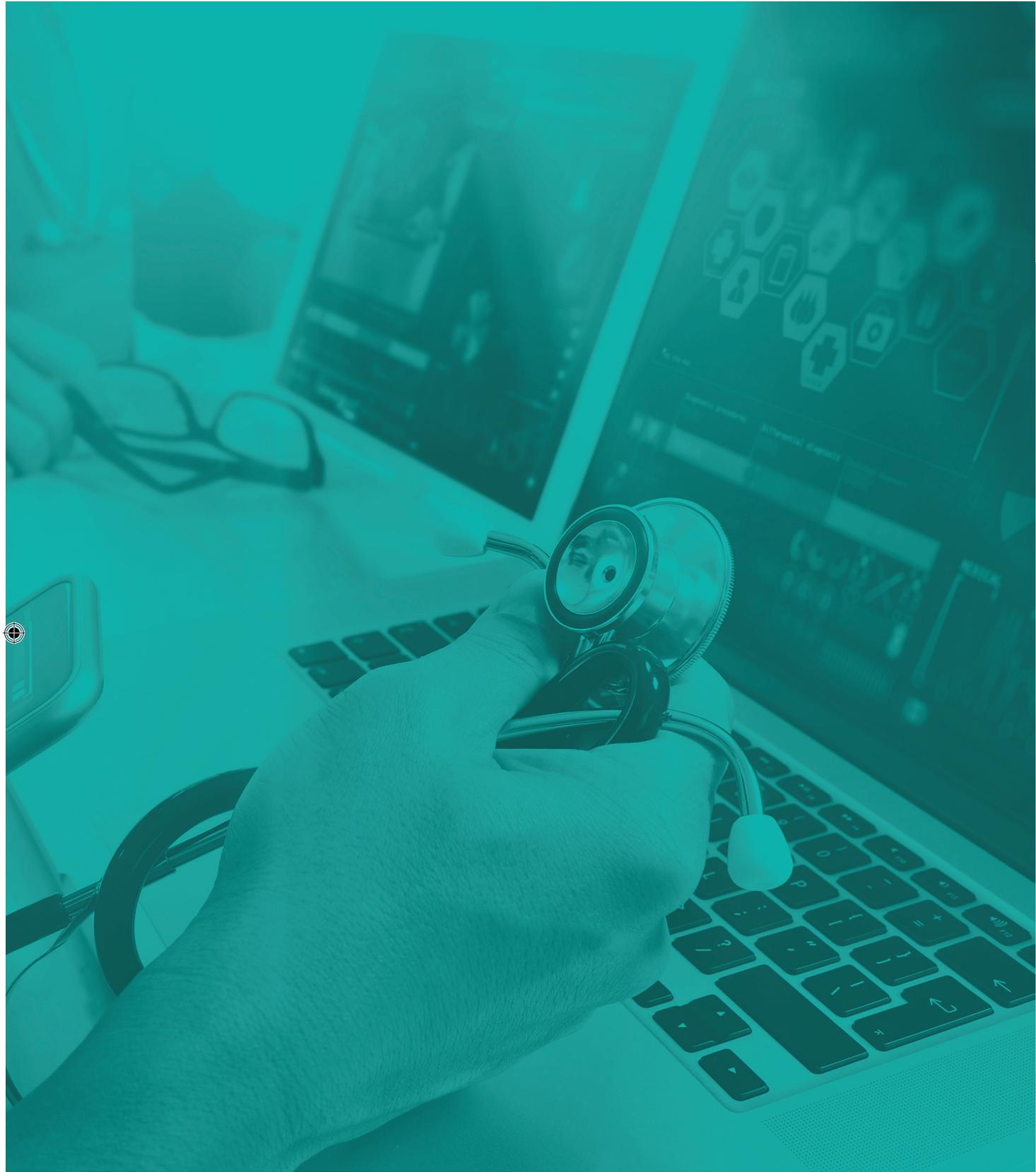


6 Health Care

6.1 Introduction

It has been reported that data from the U.S. health care system alone reached 150 exabytes in 2011 and will reach the Zettabyte 1021 gigabyte scale. The rising data volume is also supported by decreasing computing costs and technological advances that are increasing analytics throughput. Biologists no longer use traditional labs to discover a novel biomarker for a disease; instead, they depend on large and continuously growing genomic data generated by multiple research groups. In this chapter, we first explore the opportunities to use big data in the health care sector, as well as the streams for data monetization. Then, in our second section, we present the challenges.





6.2 Opportunities

U.S. and Canadian hospitals, clinics and health care providers use big data to increase profits, reduce overhead costs and, more importantly, to predict epidemics, cure diseases and avoid preventable deaths. Applications are starting to track our daily calorie counts and physical activity levels. Big data can allow for this data to be shared with doctors who may monitor a patient's treatment or get insight into the patient's lifestyle for a more efficient and accurate diagnosis and treatment [44].

Access to data from medical and data professionals, such as medical and insurance records or wearable sensors, can be used to draw a comprehensive picture of the patient and offer a personalized treatment plan. Treatments of millions of patients can be analyzed and used to find patterns and trends that enable predictive modeling. For instance, data can make a patient's reaction to a treatment easier to predict so that more accurate prescriptions can be given. In this same way, big data can also be used in clinical trials where data on applicants can be analyzed and the best candidates selected.

Data-related changes are taking place in health care. Although not as rapidly as one would expect, this

shift can be called somewhat of an "information revolution". Health care organizations are adopting information analytics systems to improve both business operations and clinical care. Health care data is being generated through various health information technology (HIT) applications, such as electronic health records (EHRs), medical imaging, public health databases and the proprietary systems of health care providers (e.g., physicians, insurance companies, Health Maintenance Organizations (HMO), hospitals, government agencies and others). In addition, data are accumulated on social media

6.2.1 Monetization

Given that we have large and valuable data sets from various sources, there are several avenues of monetization of health care data:

- Reports from analytics of large and integrated data sets to third parties.
- Operational data analytics to reduce costs of hospitals and other health care institutions.
- Lower health care costs through:

- Better and more targeted treatment plans for patients.
- Improved diagnosis - for example, hospitals and biotech companies could buy data to improve their systems and R&D.
- Enhanced discovery, which in itself is a major revenue generator for pharmaceutical companies, biotechnology companies, and others.
- Clinical trial recruitment from the appropriate gathering and analytics of patient related data sets.
- Development of improved care plans for insurance companies.
- Contextual Genomics - www.contextualgenomics.com
- Deep Genomics - www.deepgenomics.com
- BlueDot - www.bluedot.global
- PointClickCare - www.pointclickcare.ca
- ThinkResearch - www.thinkresearch.com/ca
- BioSymetrics - www.biosymetrics.com
- Dateva - www.dateva.biz
- DNASTack - www.dnastack.com
- Sequence Bio - www.sequencebio.com

Another key monetization opportunity exists for health care analytics companies and the health related Internet of Things (IoT) industry that will benefit from increased availability and standardization of health care data types and data sets.

A few notable Canadian startups in the space making headway through the effective use of health care related data are:

6.2.2 Additional references:

1. BCC Research Healthcare Analytics- Technologies and Global Markets 2015.pdf
2. BCC Research Advanced Analytics Technologies- Global Markets 2016.pdf

6.3 Challenges

The health care industry is data rich; however, many hospitals and health systems make poor use of data in improving health care. There is significant opportunity, specifically in Canada with the single payer system, to utilize the vast amounts of data available within health care organizations for enhanced discovery and as a direct result for monetization. This requires a shift in thinking and perhaps of priorities from various stakeholders, including data custodians as well as the privacy community, governments, and researchers. These stakeholders need to work together to enable data sharing and build capabilities that will have local and global benefits. The gap between data collection and processing time also needs to be addressed along with ownership, governance and standards of health care data sets. Other challenges include legislature and privacy and security, which we explore next.

6.3.1 Legislation

The importance of legislation in encouraging and even mandating health care organizations to make their data available to third parties for analysis cannot be underestimated, especially initially. In almost every area from retail to business operations, we are learning that the more data you have on certain topics, the better you can optimize your systems, assumptions and models.

A key challenge in the health care space is incentivizing health care organizations to share their data so that it can be combined with other large data sets. In many cases, there are no incentives for producers to monetize data or the perceived risks could be too high, although the data itself could very well be monetizable. Where there is interest in monetizing data, often stakeholders do not know how to navigate the space or the data's monetary value.

Legislation is a key driver for opening up health data, particularly in countries where health care is partly or fully funded by the state. By emphasizing the importance of data sharing, governments will create new channels not only for enhanced discovery but of revenue generation for data producers and third parties. Also it is important to develop provisions which set out clear rules and expectations for data producers and data users in order to minimize grey areas and confusion to the extent possible. In many cases, unless there is legislative support, many organizations will not feel compelled or comfortable in sharing, analyzing and monetizing valuable data.

6.3.2 Privacy and Security

The chief concern surrounding health care data is its sensitive and private nature. Health care institutions are wary of being exposed in any way to malpractice

¹ <https://www.privacy-analytics.com/>

risks. Legislation has not kept up with the advances in this fast moving field to offer the necessary protections so that organizations can better use data while respecting privacy requirements. On the other hand, we should recognize that there are many companies and technologies that are fully geared up to handle this type of data from storage and management to analytics and reporting.

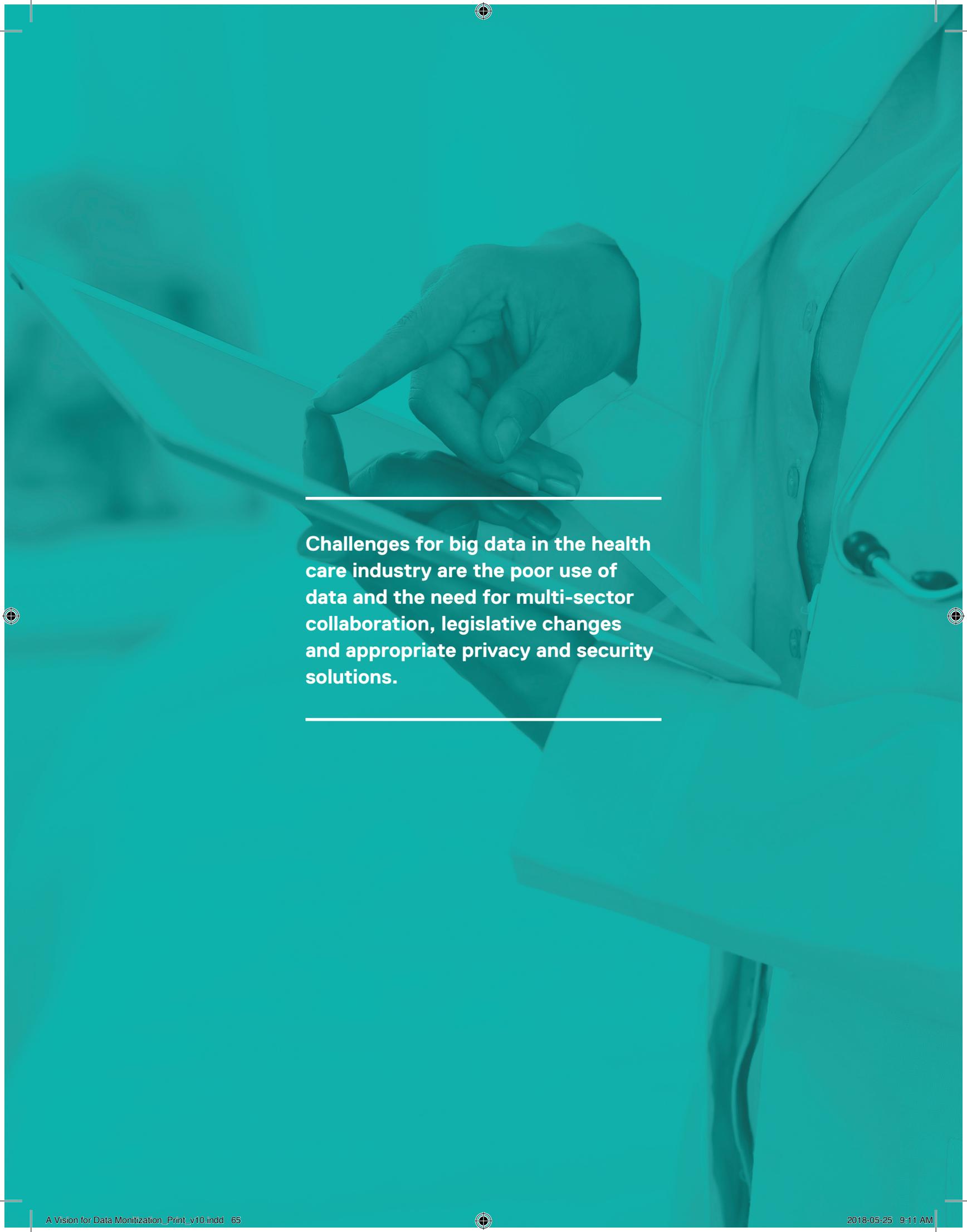
Often it is patients who are most open to sharing their data because they want a diagnosis or cure. It is vital

that health care organizations enable this type of data sharing while following the best privacy protection and security practices. New developments are making it possible to share health care data without risking personal privacy. For instance, Privacy Analytics¹, a Canadian spin-off from the Children's Hospital of Eastern Ontario in Ottawa, was recently acquired by IMS, the largest vendor of U.S. physician prescribing data. Privacy Analytics and other companies in Canada and internationally have been developing powerful tools for de-identification and anonymization of health data, to enable their effective handling, sharing, and analytics. The trade-offs are clear. We need to use the available technologies to open up data in health care so that health care and outcomes can improve.



6.4 Conclusion

In this chapter, we have seen how the health sector can take advantage of big data by using not only data found in health care platforms, but also data available from social media and wearable technology. When handled responsibly, big data can increase profits, lower hospital and health care costs, predict epidemics, cure diseases, avoid preventable death, and personalize treatments. However, challenges for big data in the health care industry are the poor use of data and the need for multi-sector collaboration, legislative changes and appropriate privacy and security solutions.



Challenges for big data in the health care industry are the poor use of data and the need for multi-sector collaboration, legislative changes and appropriate privacy and security solutions.

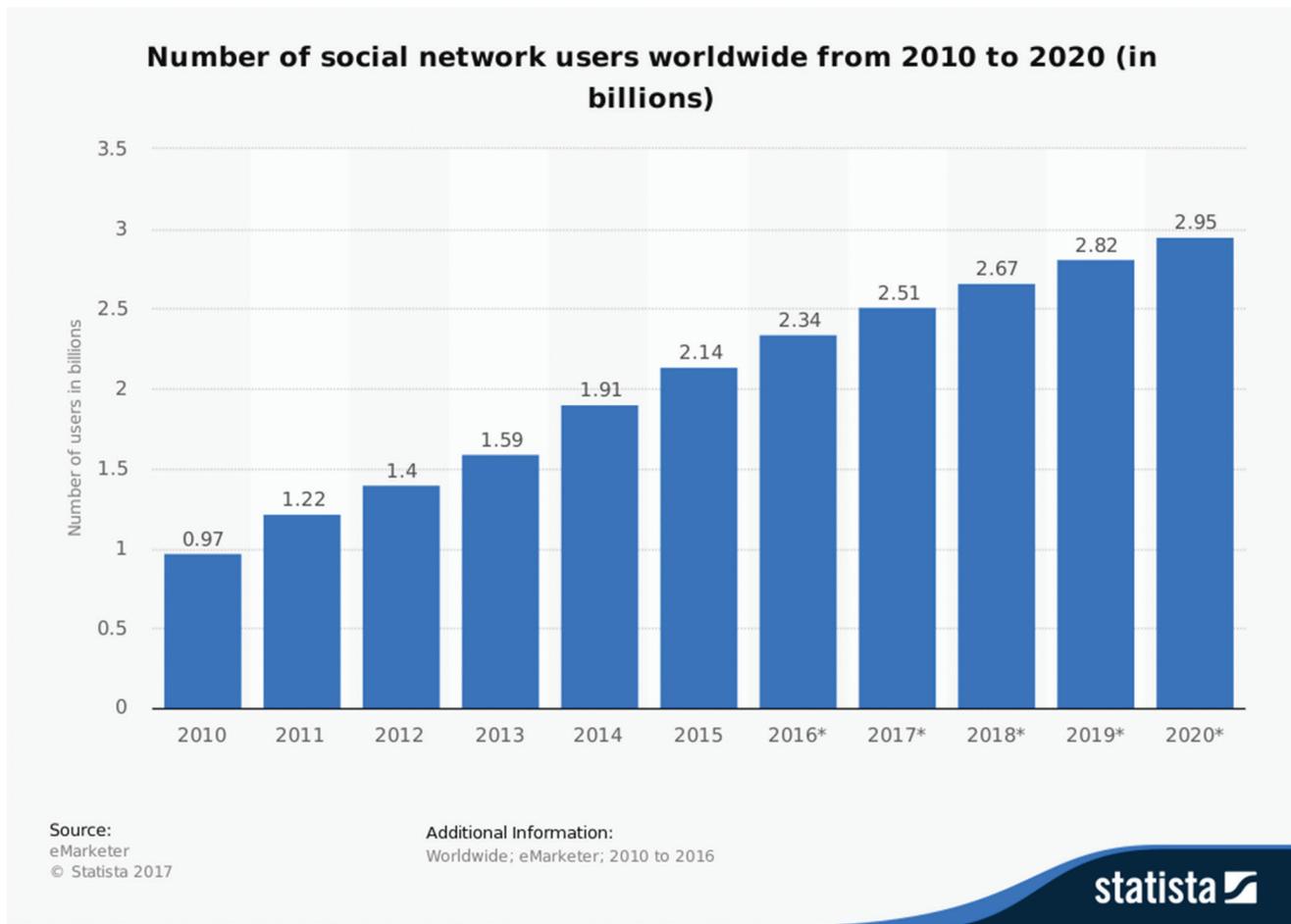
7 Social Media

7.1 Introduction

The number of worldwide social media users reached 2.34 billion and is expected to grow to almost 3 billion users by 2020, as illustrated in figure 7.1. The vast availability of data generated through social media allows researchers to gain valuable information and knowledge that can be used to tackle real world

problems. In this section, we will first look into the opportunities of big data in social media, and how it can be used to gain business insight. Then, in a second section, we will explore the different challenges facing social media big data.

Figure 7.1: Number of social network users worldwide from 2010 to 2020



7.2 Opportunities

In this section, we explore the opportunities of social media big data on business and socio-political issues. In the business realm, social media data has been used as a modern tool for business exposure and sales promotion. It has also been used as an insight into human sentiment and behaviour. Access to those insights can give businesses competitive advantages, particularly by adopting more efficient marketing strategies. To illustrate this point, we explore a case study reported by Wu Hea, Shenghua Zhab, and Ling Li [45]. Using text mining, they analyze the social media data of popular pizza chains: Pizza Hut, Domino's and Papa John's. The data in this study has been collected via social media sites, particularly Facebook and Twitter. Their findings concluded that a positive customer experience leads to increased brand loyalty and referrals, which in turn, increases the business' revenues and profits for brands that make efforts to interact with customers and sustain loyalty through social media.

Social media big data has also been widely used to monitor global social and political events and movements, disease outbreaks, natural disasters, and high-profile events. It has been used to understand human behaviours, communications and movements by a large range of institutions. Smart phones, mobile devices, and computers have made it possible for people to build a digital footprint on the Internet. One of the most prominent examples of this phenomena was the use the Arab Spring of hashtag[46] that spread information from within the Arab uprisings territories to the outside world[47].

7.3 Challenges

7.3.1 Privacy Issues

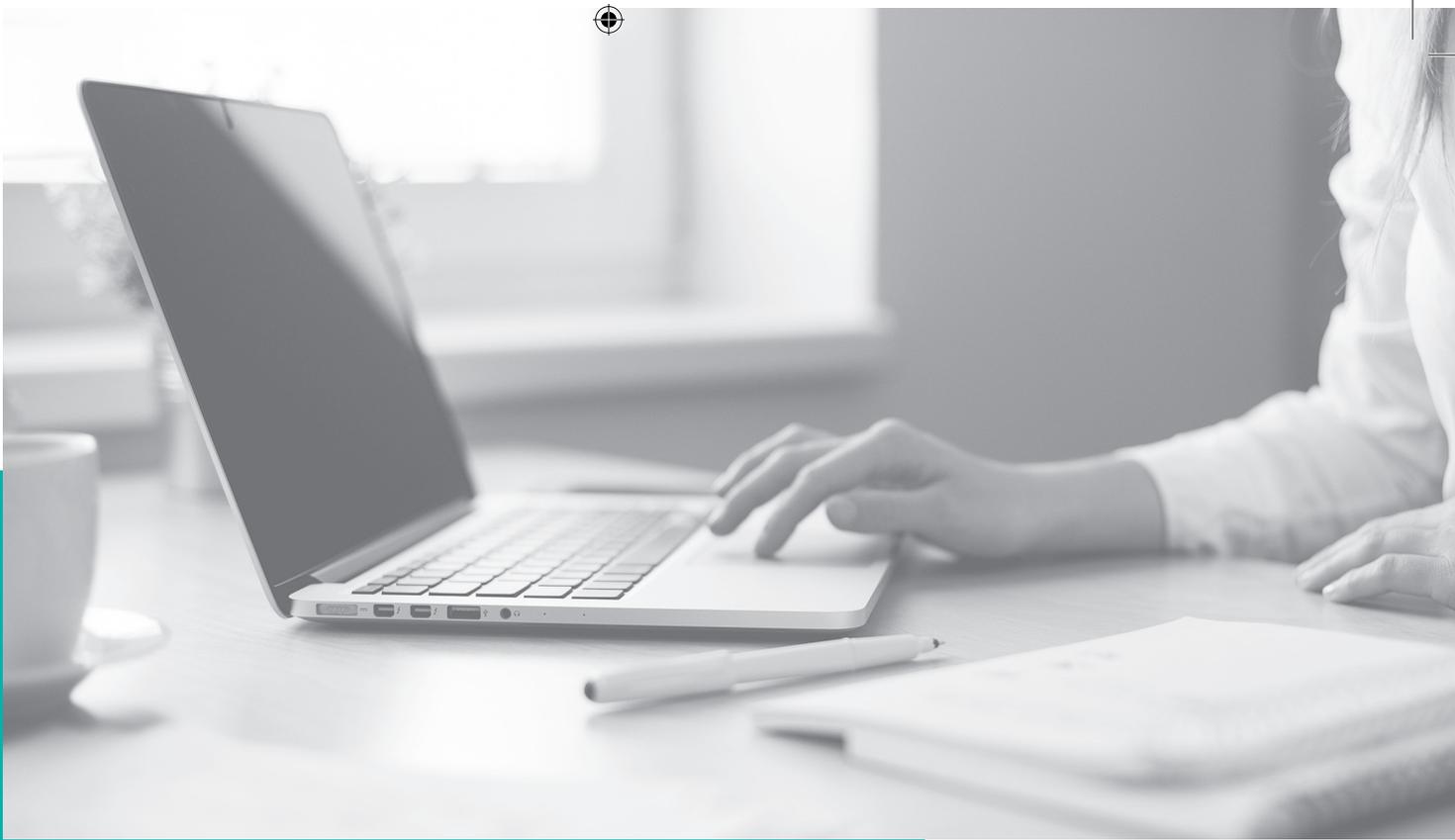
Privacy is one of the main concerns in the social media big data field – specifically, what companies do with the information that is posted on their websites or mobile applications. Smith [48] states that these concerns are being addressed by introducing regulatory rules that define what companies are allowed to do with the data they are given permission to gather.

Another privacy issue involves geo-data – location-based information embedded in modern devices that is used for applications like Google Maps and Yelp and allow people’s daily locations to be tracked. Users affect their own privacy by choosing to use such applications but further measures to protect users’ privacy should be examined.

An extension of this issue is that locations and other meta-data contained in pictures and videos can affect the privacy of people captured in the content other than the uploader. Uploaders and others captured in images have no control over who sees posts or pictures and, [unless individuals are tagged in content], there is no formal method of informing

people when their images or locations have been disclosed. Smith [48] analyzed the meta-data in a set of 5,000 random photos from the social media platform Locr and found that 10 percent of all photos could potentially harm other people’s privacy. Smith’s study [48] highlights an scenario in which insurance companies or credit rating companies might alter customers’ premiums or claims based on images or content posted by a friend of the customer. This issue has not been sufficiently addressed.

To tackle this issue, Smith [48] suggests a new type of service that would protect the peoples’ privacy when someone else posts content they do not wish to share on social media. This type of service locates the users’ surrounding (spatial and temporal) and acts like an indexing search machine. It examines publicly available social media and its meta-data and alerts the user when it detects an image of him or her. Smith [48] suggests some incentives for the development of this type of standalone service, including pay-per-use, subscription based, ad-based revenue. As mentioned in the introduction, these incentives represents some of the monetization strategies described by Platt et al. [5].



7.3.2 Noise and Potential Biases

Tsou [49] suggests that the noise found in social media data – advertisements, marketing message and other irrelevant information – make up to 70 percent of social media data. Proper data filtering and data cleaning process are essential to derive valid insights about behaviour, actions, etc., from social media data.

Researchers should also be aware that a large number of studies rely on data extracted from one social media platform – Twitter – and this reliance can introduce biases. Tufekci[47] states that almost half of the papers written for the International AAAI Conference on Web and Social Media (ICWSM), one of the prominent conferences in the field of social media, have focused primarily on Twitter data. The excessive use of Twitter for research could be explained by its clean and simple structure. However, its dominance can skew analysis and result in a biased sample selection that may not represent sentiment by behaviour in the general population.

In fact, Tufekci [47] notes that only 20 percent of the U.S. population uses Twitter. Moreover, Twitter has far fewer users than Facebook. By looking at the active users on different social network sites worldwide, see Figure 7.2, we can see that as of January 2017, there is approximately 317 million Twitter users, as opposed to 1,871 Facebook users. These statistics illustrate why it is important to integrate data from different social media networks, particularly Facebook, when conducting research. Facebook not only has a wider diffusion, but its data is also structured by race, gender, class, etc.[47], which better represents the population.

Researchers should also be conscious of peoples' tendency to change their behaviour when they know they are being observed on social media platforms[47], which affects what content they post. This change in behaviour may create some inconsistencies in the results obtained.

7.3.3 Context Consideration

The context of the data found in social media is as important as the content. Algorithms lack the ability to detect the context of extracted data. The same hashtag, for example, can have different implicit meanings. It can be used to express support, endorsement, denouncement, or even to relay sarcastic messages. Without taking context into consideration, the analysis of social media big data can be prone to wrong conclusions.

Tufecki [47] gives the example of the Aurora hashtag. In 2012, two simultaneous events related to the word “Aurora” appeared: the first one was associated with a Kim Kardashian-inspired dress designed by a celebrity boutique; the second referred to a tragic shooting at a movie theatre in Aurora, Colorado. As the hashtag Aurora was trending on social media due to the massacre, the fashion store boutique tweeted to express its enthusiasm in regards to the Aurora trend, mistakenly associating the trend to its Aurora dress. Consequently, a flood of negative sentiments emerged towards the boutique. Disregarding the context of a hashtag could potentially lead to misleading conclusions. One way to overcome this challenge is to collaborate with multi-disciplinary teams, as explained by Tsou [49].

Moreover, although algorithms are useful for extracting content from social media, they cannot access all the data on the platform. For instance, Twitter has certain characteristics, such as sub-

tweets or quoting text via screen captures, that hinders an algorithm’s ability to locate data. Privacy and context issues need to be considered when relying on findings derived from social media big data.

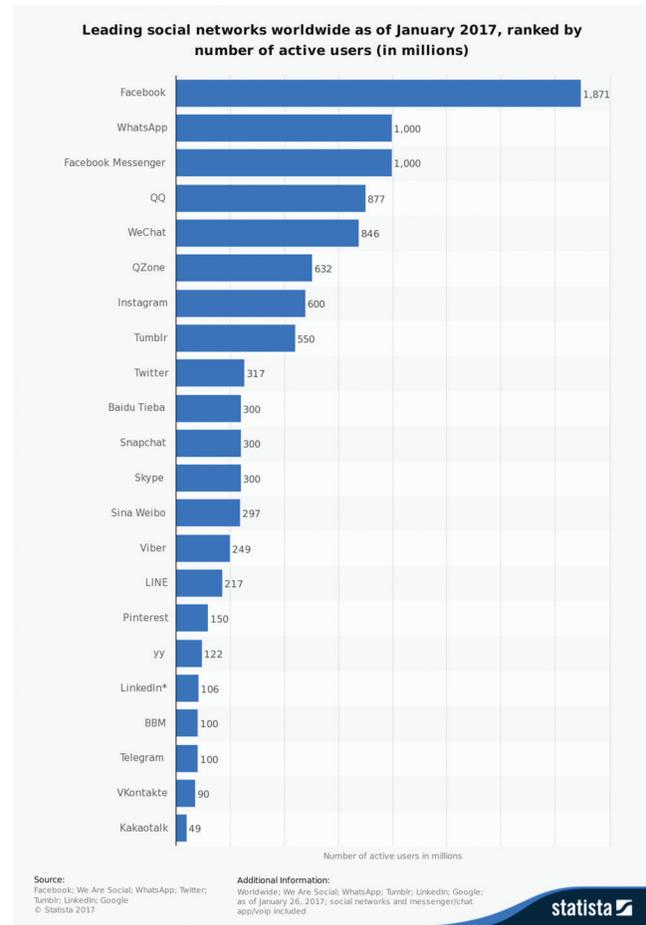


Figure 7.2: Leading social networks worldwide as of January 2017

7.4 Conclusion

In this section, we have glimpsed how the vast amount of big data available on social media can be used for business purposes. It can be used to increase businesses' profits, to gain extensive knowledge on a subject matter, or to monitor socio-political movements. However, serious privacy issues, noise found in data sets, potential biases, and disregard for context remain a challenge for the effective use of social media big data.



8 Energy and Mining Industries

8.1 Introduction

Shukla [50] defines three essential aspects of the data monetization process: 1) capturing, storing, and managing appropriate data; 2) performing analytics to identify key trends and patterns; and 3) sharing the discovered insights with internal groups who can create value from them. This process can bring an immense source of information and knowledge that can be used towards increasing gains and reducing losses in several sectors. In this chapter, we will first present the big data analytics opportunities and applications within the energy and mining sector before reviewing key challenges for the sector.





8.2 Opportunities

Energy and mining companies have significant opportunities to use big data analytics for cost and production optimization, improved services and system performance, asset management optimization, and fault and theft detection. As well, advancing technologies that connect users to real-time consumption data can benefit consumers and the community through energy cost savings and conservation. We explore those opportunities in more details in this section.

Increased Efficiency: The oil industry offers good examples of how companies can use valuable data to their competitive advantage. Oil fields constantly read data on wellhead conditions, pipelines and mechanical systems. Analysis on this data is fed to real-time operations centers that adjust oil flow to optimize production and increase efficiency. Brown [8] states that one oil company, in particular, decreased operating and staff costs by 10 to 25 percent and ramped up production by 5 percent by implementing data analysis methods.

Better System Monitoring: With the deployment of smart meters, household energy consumption data can be collected in near real-time [51]. By analyzing this data, companies can find trends and patterns in the daily or monthly energy consumption of their customers, which they can use to improve

their marketing strategies and offer personalized services [51]. Shukla [50] explains the impact data analysis has on the communication service providers by creating an intelligent network that identifies successful services and creates a personalized customer experience which will increase customer loyalty. They use big data to design competitive offers, prices and packages; recommending the most attractive offers to subscribers during the shopping and ordering process; communicating with users about their usage, spending and purchase options; configuring the network to deliver more reliable services; and monitoring Quality of Experience (QoE) to proactively correct any potential problems. In fact, as with many industries, business continuity plans are key to their survival. In the electricity industry it is not only critical for the industry, but for the livelihood of others who depend on a continuous power supply. Analyzing timely weather data and asset management condition data for example, provide utilities the ability to determine service areas that are at risk of power outages. Once identified, utilities may improve their resiliency and reliability by dispatching crews to high risk areas, prior to the arrival of incoming storms. Thereby minimizing power outages.

Moreover, a number of utilities are exploring how to use data to maintain and extend the life of

infrastructure assets. This can be done through condition monitoring, volt/VAR control, temperature monitoring, GPS data and more. It is made possible through communications systems, analytics software and system monitors that make the Internet of Things (IoT) a possibility for utility asset management. The potential results are improved maintenance programs, timely maintenance, and ultimately reducing costs and increasing the longevity of key assets. Predictive modeling enables companies to more accurately match electricity supply to demand.

Introducing smart meters has also given utilities the ability to monitor electricity usage and system performance in near real time. Identifying fault locations is key in maintaining the grid and providing service to customers. Advanced Metering Infrastructure (AMI) data along with Supervisory Control and Data Acquisition (SCADA) data from substations can identify energy losses. Areas and homes that are experiencing higher losses than acceptable can be investigated for theft or other fault problems (i.e., meter tampering, faulty equipment or line damage), which reduces energy losses.

Consumption and Conservation: Big data analytics findings also benefit consumers and the community as a whole, mainly through savings on energy

consumption cost and energy conservation. With the emergence of real-time interaction between power companies and energy consumers, consumers can monitor their real-time consumption and adjust their consuming behaviour to save on energy costs. Zhou and Yang [51] found that households can cut energy consumption by 10-30 percent just by changing the consumers' behaviour. The authors estimate that "up to 27 percent of current households' energy use can be saved through more efficient energy use."

Consumers ability to shift their consumption patterns to match the supply of power was highlighted in a 2016 pilot project [18]. Hydro Ottawa participated in the pilot project along with five other municipalities to encourage consumer participants to reduce energy consumption. Customers were encouraged to reduce consumption on key days during a specified 5-hour period, over the course of several months. Using realtime applications that fed data and patterns of energy usage, participants were able to manage and change their consumption. (Ottawa in 1st place for energy saved amongst the participating cities.) However, there remain many challenges in this field, which we explore in our next section.

8.3 Challenges

The challenges include collaboration between the government, local communities and companies, as well as inter-disciplinary collaboration, energy market competition and technical challenges.

In fact, big data analytics research has been able to provide valuable insights on the energy consumption behaviour [51] which is an asset for governments and companies. Good analysis promotes better management and policy decisions that can optimize conservation and control energy costs. As such, it is important that effective strategies be put in place to promote energy consumption reduction, which is one of the challenges we face in the energy and mining sector. As mentioned by Zhou and Yang [51], the government, local communities and power companies should collaborate together, and implement more efficient measures to promote householders’ behavioural changes towards energy consumption.

Another challenge is related to the increased need for collaboration between the following three disciplines: Data and Information, Energy and Environment, and Behaviour and Psychology,[51]. With better collaboration, more efficient and effective conclusions can be generated from big data analytics, with the attempt to understand and change energy consumption.

The two broad technical issues are data scaling and data variety. Companies must scale and manage realtime data that is increasing in volume exponentially. For example, as seen in Table 8.1 [51], 1 million smart meters that capture data every 15 minutes can generate 2920 terabytes of data within a year. Managing data this big involved storage problems, processing problems related to delivering meaningful information to decision makers in near real time, and analytics challenges of making sense of so much data in near real time for complex decision making. That being said, we recognize that that data-handling challenges that arose from the sheer weight of big data have now largely been met by tools developed in the open-source community.

Systems also need to be able to handle a variety of data[51]. Energy data comes in many forms: structured data, such as the energy consumption data; semi-structured data, such as data exchanged amongst smart energy management platforms and third-party aggregators; and unstructured data, such as email or SMS notifications about energy use. In addition, there are also some inter-industry data such as electric vehicle related data, and outside-industry data such as weather data. This mix in data types and sources adds to the complexity of energy big data analytics and requires sophisticated systems to handle it.

Table 8.1: Amount of data collected by 1 million smart meters in one year assuming 5KB per records collected

Collection frequency	1/day	1/h	1/30 min	1/15 min
Records collected	365 million	8.75 billion	17.52 billion	35.04 billion
Terabytes collected	1.82 TB	730 TB	1460 TB	2920 TB

8.4 Conclusion

With big data analytics, energy companies can optimize costs and production processes, improve services and marketing strategies, better maintain their assets, and limit faults, theft and energy losses. Consumers and the community as a whole also benefit by means of energy consumption savings and energy conservation.

The challenges include collaboration between the government, local communities and companies, as well as inter-disciplinary collaboration, energy market competition and technical challenges.



9 Manufacturing

9.1 Introduction

The manufacturing industry mainly gathers its data from historical data and sensors distributed through the manufacturing process. In the manufacturing environment, sensors encompass networks or wireless sensors, Radio Frequency Identification (RFID) tags and readers, GPS system, inventory monitoring and control sensors, etc. [52]. The emergence of these technologies has produced a vast amount of data in the manufacturing sector, translating into big data. In fact, with big data and the Internet of Things (IoT), the manufacturing operations have evolved from a “5M” framework to a more data-driven and connected “5C” framework [53]. The “5M” framework consists of: Materials, Machines, Methods, Measurements and Modeling. Whereas the “5C” framework consists of: Connection, Cloud, Content, Community, and Customization. In this chapter, we will first present the value added that big data provides to the manufacturing system, and then we will provide an example of data monetization within this sector, and discuss additional data monetization opportunities.





9.2 Big Data, Transparency, and Predictive Manufacturing

Transparency is the capability of a company to decipher uncertainties and create efficient solutions. According to Lee [53], there are two types of uncertainties: internal and external. Internal uncertainty relates chiefly to the deterioration of machines and manufacturing processes, and includes variable cycle time due to inconsistent operation or breakdowns. External uncertainty refers to unreliable downstream capacity (i.e., shipping and warehousing); unpredictable delivery of raw materials or parts; quantity and quality issues; or incomplete product design. In this section, we will first explain what a predictive manufacturing system

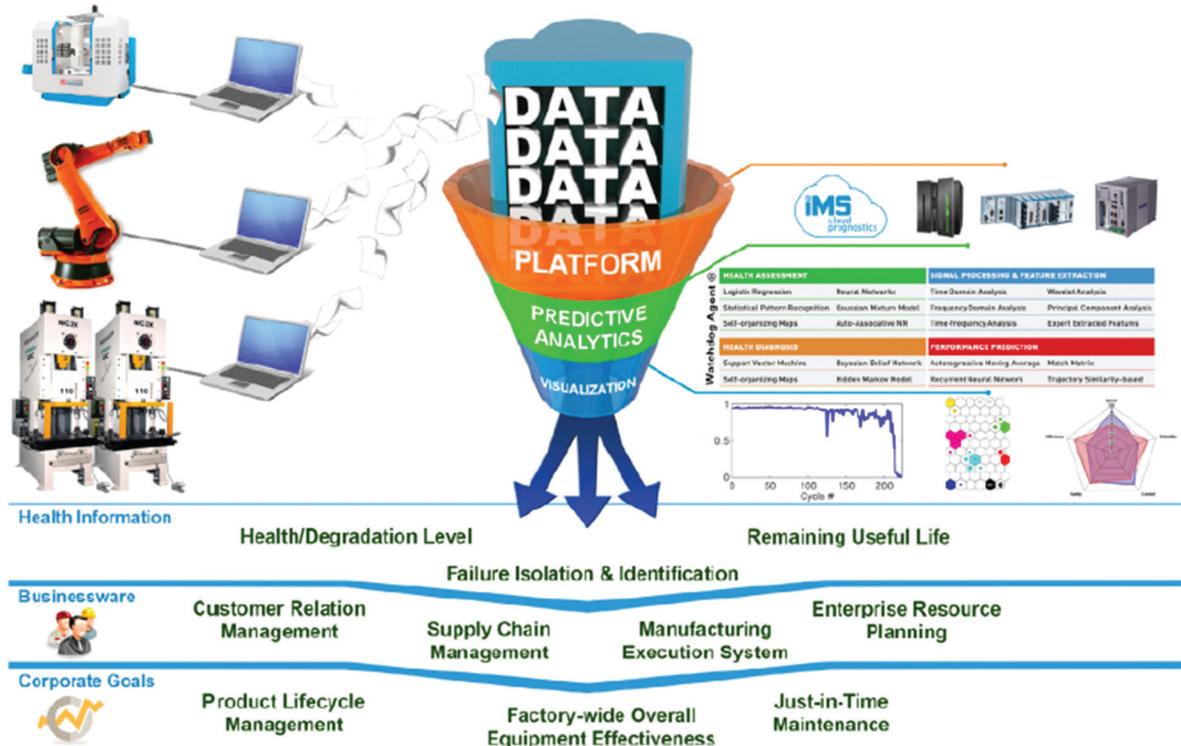
is, and then we will look at how big data is used to manage common manufacturing uncertainties.

9.2.1 Conceptual Framework of a Predictive Manufacturing System

In their paper, Lee, Lapira, Bagheri and Kao [53] illustrated the conceptual framework of a predictive manufacturing system as per Figure 9.1. Following is a summary of figure 9.1:

- Data is collected from system sensor or through data mining of historical data. The aggregation of all the data forms what we call “big data”.

Figure 9.1: Predictive manufacturing framework



- Data is transformed through various components: an integrated platform, predictive analytics and visualization tools.
- The platform is selected based on different characteristics, such as the speed of computation, investment cost, and the ease of deployment .
- Big data is then transformed into valuable manufacturing performance information with the help of predictive analytics such as the Watchdog Agent developed by the Intelligent Maintenance System (IMS) center. The algorithms in the

Watchdog can be classified into four categories: signal processing and feature extraction; health assessment such as the current condition; performance prediction; and fault diagnosis [53].

- The information generated is communicated in a comprehensible format using visualization tools.

Using a predictive framework, manufacturers can make sure the operations of the system are smooth and how likely will remain smooth.

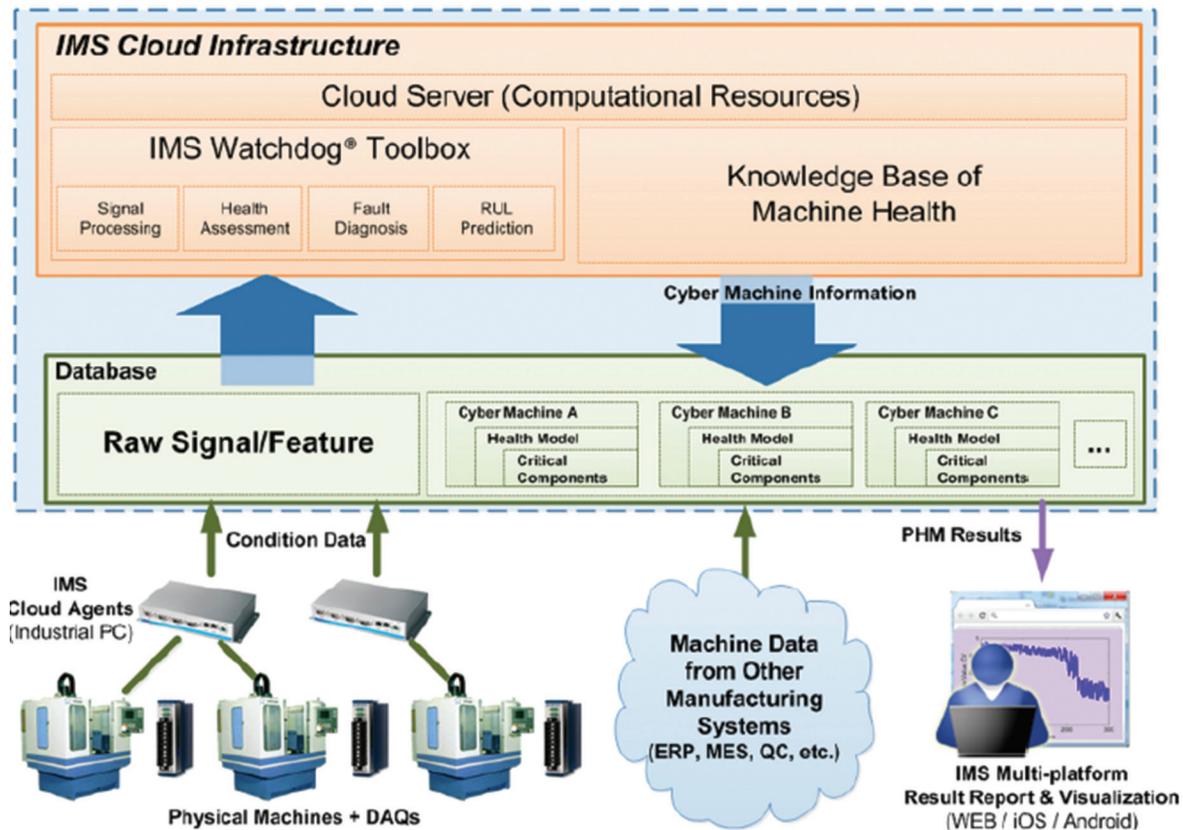
9.2.2 Deployment of the Internet of Things (IoT) and Real-Time Data

After generating information from the data, predictive manufacturing systems share and integrate information with other management systems via Internet of Things (IoT) technology. In fact, the Internet of Things (IoT) is primarily what allows big data analytics to be used in the manufacturing industry in real time. Babicneau and Seker [52] state that “the emerging Internet of Things technology enables human operators and decision-makers to talk to the physical objects, as well as enables the physical objects to talk to the human operators and decision-makers.” This is done through the communication of data generated by the sensors and sensor networks, which allows for efficient

monitoring, control, and adjustment at any stage of the manufacturing operations. This combination of technologies allows for efficient collaboration in manufacturing processes, which can add more value to the enterprise.

Cloud computing plays a particular role in the predictive manufacturing framework. Connectivity allows the company’s managers or employees to track the real-time manufacturing process and machine conditions, regardless of where they are located [53]. They can do so by accessing the «cyber-physical» model scheme [53], illustrated in Figure 9.2. It is basically a «digital factory» [54] mirroring the entire manufacturing process, and created through advanced computational methods.

Figure 9.2: Cyber-physical model for enhanced predictive manufacturing system



9.2.3 Predicting Future Opportunities

So far, we have seen that when it comes to the production process, preventing of system performance deterioration can bring about important gains to companies. But there are other ways the manufacturing industry can make use of its valuable data. For example, it can be used to improve other areas across the manufacturing value-chain from product development to after-sales. Manufacturers can also harness data to improve forecasting opportunities to better manage demand volatility. Better information on forecasted demand allows manufacturers to optimize their supply chain process[54].

The vast amount of data available from customers can be used towards product development strategies. This data can answer questions such as: What is the most important feature according to customers? Or how much is a customer willing to pay for a certain

product or service? By answering these questions, product designers can enhance current products or create new products matching the customers' needs, while minimizing cost production and development costs. Some manufacturers even use web-based platforms to include stakeholders in the product development process.

As for marketing, sales, distribution, and after sales, these can also be optimized with the use of big data. For example, Nedelcu [54] presents the concept of inserting sensors in the retail products in order to collect actual time data about product usage and performance. This would allow manufacturers to systematically adjust the production process when defects appear. But for this strategy to succeed, companies must have the customers' consent to collect their data without violating their privacy.

9.3 Two approaches for Monetizing Data in Manufacturing

A good example of data monetization through operational efficiency is found in case study [55] that examined how a precious-metals mine was able to increase profitability. The grade of ore was declining at the mine and the company wanted to maintain production levels. To recover precious metals, the mine considered 10 and 15 variables and used more than 15 pieces of machinery. The production and process data was extremely fragmented, so data mining had to be run on the mine's information using mathematical approaches to fix inconsistencies and recover missing data. The team then examined the data and recognized that variability in levels of dissolved oxygen (a key parameter in the leaching process) seemed to have the biggest impact on yield. The analysis also showed that the best demonstrated performance at the mine occurred on days in which oxygen levels were highest. By accounting for this finding, the mine increased its average yield by 3.7 % within three months, translating 10 million to 20

million annual profit with minimal additional capital investments costs.

Manufacturing companies can take another approach to monetizing their data; they can sell it to other organizations. In fact, some companies are already selling their data and are generating significant revenue from this practice. In 2012, 27% of the companies in Asia, Latin America, Europe and North America were selling their data, generating average revenue of around 21.6 million for the companies [54]. With more than two-thirds of companies not yet involved in selling data, there is still plenty of room to monetize data within the manufacturing industry.

9.4 Conclusion

This chapter sheds light on how big data is utilized in the manufacturing sector and how it adds value to this sector. The manufacturing industry can gain valuable insights from the large amount of data it possesses, and can use it to improve all aspects of its manufacturing value chain. A major application is the real-time monitoring of manufacturing machines to prevent performance deterioration. Another big data application in the manufacturing industry is the integration of sensors in the retailed products which will generate real-time data on actual usage and product performance. This approach offers manufacturing companies an opportunity to analyze real-time data, predict the most accurate demand level, and optimize production. It is important to note, however, that one major challenge facing the manufacturing companies is the high cost of investment in the field of big data analytics, as well as the shortage in talent to tackle such complex issues.



10 Government

10.1 Introduction

The opportunities big data presents for the public sector are significant and cover many aspects of government's role, such as security, promotion of general welfare, and economic growth. Government can integrate data analysis in public sector operations and services to support the public good and betterment of life for citizens. Government also has a leadership role to play supporting the use of big data analytics by creating good privacy policy and standards for the responsible use of data, and by supporting open data initiatives. In this chapter, we will first explore the role of big data in multiple areas of the government sector. Then, we will present the opportunities and challenges facing governments when it comes to harnessing big data.





10.2 The Role of big data Analytics in the Government Sector

10.2.1 Homeland Security and Law Enforcement

Government can use big data and data analytics to support homeland security and law enforcement. With big data analytics, authorities can be proactive in detecting and preventing malicious activities (e.g., terrorist attacks) and increasing public safety. Video data for surveillance and criminal investigation is a crucial component for public and private security agencies. Use of surveillance data has expanded to include police car cameras, body-worn video cameras and tazer gun cameras, to name a few. These devices will more efficiently secure lawful and accurate information compared to witness observations, but the wealth of data they generate is difficult to manage. Michael [56] describes challenges that arise when the big data becomes too large and expensive to store and process quickly.

On another note, citizens can now take on a partnership role with the government as a result of the emergence of crowd-sourcing or crowd-reporting (reporting an incident on-line) platforms. Crowd-reporting online platforms allowed for the citizen to take the role of the council inspector, therefore creating a more direct cooperation between public and government agencies and enhancing the citizens level

of engagement. This activity, for example, translates into a lower cost of civil services (less council inspector needed), and enhance the transparency of public service processes (with the option tracking of reported issues). This information also adds on the wealth of information collected by governments for increased accuracy of the analysis.

10.2.2 Economic Growth and General Welfare

Many countries, including Canada, are working on initiatives related to big data individually and together. To unravel the economic potential of public data, the European Union Commission, for instance, established the “Digital Agenda for Europe” in 2010 to investigate the means through which the digital market can provide sustainable economic and social benefits to EU citizens [57]. When harnessed in the government sector, big data analytics can produce returns through enhanced operational efficiency and public services, reduced fraud and error, and increased tax collection. Yiu estimates the returns for the UK public sector of using big data have amounted to between pound 16 to pound 33 billion [58] and provides some examples of big data applications for public services could be done by means machine learning techniques. Yiu states that “in the welfare

arena, better segmentation and personalization could help to identify the support that unemployed people need and get them into long-term work.”

Further, governments aim to create value through big data research. As a result of a shift from small-sample government surveys to large-sample ones, researchers can now examine a sub-population of their choice, resulting in a more rigorous study related, for example, to wages, health, productivity and other matters [59]. As in the private sector, using big data in the public sector can help governments gain better insights about the population they serve and design programs that can yield better social outcomes. For example, sentiment analysis about a specific government program helped identify the common hurdles from citizens in using the program and led the program to work on addressing the identified issue.

A number of municipalities in Canada (e.g., Edmonton, AB - https://www.edmonton.ca/city_government/initiatives_innovation/smart-cities.aspx) are using data to help better manage their resources (e.g., energy, water). This approach is aimed at reducing costs, as well as improving citizens life through better traffic and transit management (e.g., traffic, transit) and reducing pollution.

Another big data application example comes from the newly launched federal Big Data Analytics Centre of the Communications Research Centre. The Centre is using big data and artificial intelligence technologies to look for ways to use economically critical wireless spectrum more efficiently. Wireless data traffic is expected to increase 1,000 times by 2020, which will apply significant pressure on Canada’s infrastructure.

The Centre is experimenting with new technology to locate unused radio waves in real time and put surplus capacity to work. ([https:// www.canada.ca/en/innovation-science-economic-development/news/2017/05/official_openingofcrc_bigdata_analyticscentre.html](https://www.canada.ca/en/innovation-science-economic-development/news/2017/05/official_openingofcrc_bigdata_analyticscentre.html))

10.3 Opportunities and Challenges

Promoting Open Data and Collaboration: Canada recognizes that big data analytics is an essential component of the use of evidence-based policy development, results-based programs and the evaluation of efficiency. A number of projects are underway to integrate data analytics in government operations to create better outcomes for Canadians. In addition, the federal government is part of the international open government movement – the Open Government Partnership (OPG) – which not only aims at increasing transparency and citizens engagement in their democracy, but also in spurring innovation through public involvement.

As part of this initiative, the federal government plans to make data open (public) by default, and is also working towards making publicly funded research open by default (Commitment 14: Increase Openness of Federal Science Activities (Open Science)). Encompassed in Canada’s third biennial plan to the OGP, the federal government is committed to assess the benefit of open government data (Commitment 5: Define an Approach for Measuring Open Government Performance). (<http://open.canada.ca/en/content/third-biennial-plan-open-government-partnership#toc5-1-5>). A 2015 study from Natural Resources Canada looked at the economic impact of open geospatial data and estimated that it added \$650 million to Canada’s GDP as a result of its use. (<http://geoscan.nrcan.gc.ca/starweb/geoscan/servlet.starweb?path=geoscan/fulle.web&search1=R=296426>). Quantifying the potential economic value of open data is difficult and through a number of case studies, the GovLab is assessing the value of open data for governments around the world (<http://odimpact.org/>).

The Public Sector Executives Network of Canada is learning from the British government’s evolving methods for collecting and analyzing data, and developing

predictive techniques to acquire valuable insight. The Public Sector Executives Network [60] states that Canada is focusing on the following objectives:

- Understanding how other governments are implementing open data policies and the international legal framework used from bodies such as the United Nations, the Organization for Economic Cooperation and Development and the European Union.
- Learning the various methods that Canadian governments are using for open data and data analytics in industries such as public safety, citizen involvement, health care, public health and social programs, roads, bridges and other infrastructure, transportation and the environment.
- Exploring how open data and data analytics can promote citizen participation, create new business opportunities and build constructive partnerships.

Moreover, Tableau[61] mentions two points that governments should consider to optimize the use of data. First, governments should identify problems that could be solved by big data initiatives rather than sticking to low-cost, short-term solutions. Rather, Tableau argues government should focus on solutions that would achieve long-term goals. Second, it should support collaboration between the public and private sectors. Tableau [61] lists open source initiatives with the private sector and inter-agency collaboration across the public sector as key issues that the government should work to address. The realization that much of their data is shared among different government agencies and more useful with appropriate data from private sector firms could help bring value to big data initiatives as it would make analysis efficient.

10.4 Conclusion

Developing Good Privacy Policies: As in the private sector, governments face privacy issues in the responsible use of big data. Privacy issues affecting the use of big data in a broad range of sectors have been discussed thoroughly in preceding chapters. Government has a major role in protecting privacy by developing (and evolving) balanced policies that set transparent standards for how data can and cannot be used[60]. Ethical dilemmas also arise. Unlike surveillance cameras, the data collected from smart phones and wearable devices will not allow for the privacy of bystanders who are captured in a video by coincidence because they were in the wrong place at a certain time.

Sharing Data Internationally and Nationally: Along with privacy concerns, one of the most prominent challenges for government is data sharing. It is a particular challenge to share data between countries, but if overcome, it can greatly enhance homeland security and other aspects of the government's role. For example, Kim [57] states that the Boston Marathon tragedy [62] could have been prevented had the Russian government shared information with the U.S. government. Global collaboration is therefore necessary for each individual nation's well-being. Because governments are large and complex, Kim finds that 58 Contents sharing data within and across government departments and agencies also requires sustained effort and planning[57].

Managing Technical and Talent Constraints: Other challenges facing governments include capturing, categorizing, storing, searching, sharing, transferring, analyzing and visualizing information. Capturing data is a difficult task for governments as it not only comes from different platforms (social network, the Web etc), but also from multiple sources (such as countries, institutions, agencies etc). Also, an efficient analysis of the data requires highly skilled data professionals that are in high demand[57].

In this chapter, we presented concrete examples of the benefits big data analytics brings to the government, whether in the security domain, economic growth or general welfare . However, it was also evident that, like in any other sector, governments too have some challenges with the utilization of big data analytics, mainly privacy issues and local and global data sharing difficulties. On another note, given the shifting landscape of data requirements and use restrictions, it has become clear that overcoming these obstacles cannot be accomplished by government alone. Infomediaries and data management solutions must be leveraged in some capacity in order to unlock the value of external data.

11 Conclusion

Industry and the public sector can benefit from the increasing amount of data available. When harnessed well, data provides valuable insights which, in turn, translates into profits and sustainable development. To achieve these results, it is important for organizations to grasp of how data can be monetized across different sectors and business areas. By examining concrete examples of how big data analytics is being used currently, or could be better used, across nine areas and sectors, we conclude there are several strategies that private and public enterprises can use to leverage their valuable data to create competitive advantages.

Increasing sales promotion and revenue potential:

The analysis of data has proven to be a useful tool to make more informed decisions that can increase company revenue or value. Companies across multiple sectors can use data analyses to their advantage for sales promotion and increased revenue to improve their customer bases and bottom lines. Many businesses and retailers use customer sales and social media data to improve their marketing strategies by better targeting promotions to specific customers segments or audiences to generate more sales potential.

Banks and other financial institutions can take advantage of their internal database (transaction records, customer service call recordings, etc.) to create personalized selling strategies that meet customers' needs and to promote cross-selling. Machine learning can be used in the client service or telephone banking department where vocal data

can be analyzed to generate more sales opportunities and identify customer solutions that will improve satisfaction and retention. Based on the customer's words and product portfolio, sentiment analysis delivers real-time suggestions to call center agents or financial advisers that can benefit customers, as well as protect or increase profitability.

Trading or selling data: The main strategies for data monetization discussed in this paper include, but are not limited to: trading data with partners or selling the data through different means, such as service bundles or pay-per-use. Some manufacturing companies have begun selling data to increase revenue.

Optimizing performance and driving efficiency:

Companies in the energy and mining industry, particularly the utilities sector, can use big data advantageously to monitor system performance and energy loss, optimize operational processes, and improve efficiency. Manufacturers are using predictive analysis to capture and transform diverse data into valuable real-time information to improve manufacturing performance and all aspects of the manufacturing value chain. Similarly, we find that health care providers can benefit from big data insights and real-time feedback to improve health care services, increase efficiency and reduce costs. Even energy consumers can use real-time big data insights to change their energy consumption patterns and save on energy costs.

Better prediction insights and more informed decision-making: Combining and analyzing huge amounts of data in big data sets offers opportunities to create more reliable insights with higher prediction accuracy in the public and private sectors. The health care field has a wealth of big data that can be used to predict epidemics, cure diseases, avoid preventable deaths and predict patient outcomes in order to personalize more effective treatments. Researchers can use vast social media data to gain insights and understanding about behaviour, public sentiment, and global socio-political movements in order to tackle real-world problems. In the financial industry, banks are beginning to use big data to better understand the business landscape outside their internal customer base. Advanced machine learning and other analytical techniques offer financial organizations the ability to analyze sentiment in thousands of financial transcripts in order to better predict future economic conditions in companies and offer better recommendations to clients. In the real estate industry, by analyzing neighborhood demographics, commercial and residential buyers can make better decisions when it comes to investing in real estate

Fraud detection and loss control: Big data is also key to security. Analytics can be used to prevent cyber threats (although outright prevention is difficult as hackers become more sophisticated) and to detect anomalies that may signal fraudulent activities. Fraud detection provides high benefits to private enterprises

and governments by minimizing losses, and governments also make use of big data to improve homeland security, public safety and general welfare.

Talent, technical and privacy challenges: We find that most industries face common technical challenges in dealing with big data related to data storage and volume, data format varieties, analytics processing power and other technical difficulties. Common and recurrent challenges across industries relate to talent, data security and privacy concerns. Industry and government lack of skilled human resources to perform big data analytics. Educational institutions must consider adding courses and specializations in areas such as data mining, text mining, predictive analysis, network analytics, among other data-related skills. All sectors face the challenges of protecting data privacy, security of data, as well as understanding and sharing the different types of data existing in organizations.

Need for more open data and collaboration: Academia faces the challenge of lack of access to publicly available big data sources. A strong collaboration between industry and academia is needed to create skilled resources capable of addressing big data needs of organizations, and such strong collaboration can not be created without government's support.

Glossary

Advanced Metering Infrastructure In the energy sector context, it refers to an infrastructure of smart meters different from the traditional one, in such a way that communication from the meter to the network can be done via wired or wireless connections. 48

Apriori Association Rule Algorithm used to identify frequent individual items in a database and the most common item associated with the first one. It is often used to discover trends and patterns in the database. 28

biomarker A measurable substance in an organism indicating the presence of disease. 37

Chart of Accounts A list of a company's accounts used to organize the finances of an entity. 13 cloud computing Shared computer processing and data resources provided through Internet-based computing. 26, 33, 53

clustering A technique that groups data into categories in such a way that the data found in one group have greater similarities than those found in another group. 18

CLV Customer Lifetime Value - The estimation of the net profit of a future relationship with a customer. 18, 19, 22

data mining The process of analyzing large databases in order to derive information. 22, 27, 28, 51, 54

data storage Data storage, also called memory, is the computer component that saves and digital data. 5, 7, 21, 28

exabytes A measure of computer storage capacity. One exabyte is equal to one quintillion bytes. 37

false-positive A result erroneously indicating that a specific condition is present. 34

genomic data Data related to the study of genomes. A genome is a series of genes found in a cell. 5, 37

gigabyte A measure of computer storage capacity. One gigabyte is equal to one 1000000000 bytes. 37

Hadoop One of the most commonly used big data open source platforms. 9, 11, 34

Internet of Things It refers to all physical devices and machines able to exchange data via network connectivity. i, 38, 48, 51, 52

JSON JavaScript Object Notation - An open standard file format or data format used effectively for browser/server communication. 10

Kafka Apache Kafka is an open-source stream processing computation platform originally written in Scala and Java. 10

machine learning A set of techniques in which algorithms are used allowing computers to learn by and of themselves. 5, 23, 24, 26, 34, 56

MapReduce The main component of the Apache Hadoop software framework. It provides a performing distributed and parallel processing system for large amounts of unstructured data sets across computer clusters. 9

Market Basket Analysis A data analysis and data mining technique that highlights relationships and associations among activities performed by specific individuals or populations. 17, 19-21

meta-data A data describing and providing information about other data. 42

Natural Language Processing A computer science subject interested in the communication between computer and human language. 23, 24, 26

NoSQL Not Only SQL or Non SQL - A NoSQL database is a non relational database which allows for a non relational database processing. 9

petabyte A measure of computer storage capacity. One petabyte is equal to 1,024 terabyte. 11

predictive analytics The use of data and statistical analysis to foresee future events. 22, 51, 52

Quality of Experience A measure of customer's satisfaction from a service. 47

Radio Frequency Identification type of electromagnetic sensor used to track objects electronically. 51

RFM Recency Frequency Monetary Value - A technique used for analyzing customer value. 17, 18

sequence analysis In the bioinformatics context, sequence analysis is a process used to analyze DNA, RNA or peptide sequence using data analytics. 34

Spark An open source cluster computing framework. It is now the largest big data open-source project, providing the best processing speed compared to other open source projects. 10

speech analytics The operation of obtaining information through analytics from data contained in recorded calls. 24

spoofing attack The act of maliciously masking one's address by using another address. 35

SQL Structured Query Language - A language used for managing and stream processing data contained in a relational database management system (RDBMS). 11, 13

Storm Apache Storm is a distributed stream processing computation originally written in Clojure language. 10

structured Data that is presented in an organized manner or in a fixed field within a record or file. Data in relational databases for example are considered structured data. 7, 9, 10, 13, 15, 49

supervised The machine learning task of inferring a function from labeled training data. 34

terabyte A measure of computer storage capacity. One terabyte is equal to 1,024 gigabytes. 9

text mining Also called text analytics. It is the operation of obtaining information through analytics from data contained in a text format. 22, 42

unstandardized data The opposite of standardized data. Unstandardized data refers to data received in different types and formats. 8

unstructured The opposite of structured data. Unstructured data is data that is presented in a disorganized manner and is more difficult to read or manipulate. 7, 9, 10, 13, 15, 25, 26, 28, 49

unsupervised The machine learning task of inferring a function to describe hidden structure from unlabeled data. 34

XML Extensible Markup Language - A language that contains specific rules for encoding documents in a format that can be easily read by humans and machines. 10

Zettabyte A measure of computer storage capacity. One exabyte is equal to one sextillion bytes. 37

List of Workshop Participants

Vikas Shreedhar, Accenture

Andrew Hall, aero Info

Atul Varde, Affinity Credit Union

Michael Poyser, Aimia

Shiva Amiri, BioSymetrics Inc.

Joseph Bou-Younes, Canada's Open Data Exchange

Daniel Gent, Canadian Electricity Association

Anne-Marie Brinsmead, Chang School RU

Fred Anger, Chang School RU

Katherine Goff Inglis, Cineplex Inc.

Newton Asare, ClientDesk

Paul Macmillan, Deloitte

Steven Maynard, Ernst & Young

Gord Edall, Globe and Mail

Colin McKay, GOOGLE

Laura Morin, Government of Canada

Claire Gabillard, Government of Canada

Krista Campbell, Govt of Canada

Zeljko Cakic, GTAA

Clement Li, HydroOne

Giuliana Rossini, HydroOne

Rob Quail, HydroOne

Karen Harracksingh, Industry Canada

Sarah Hobson, Industry Canada

Andrew Matte, Lighthouse Labs

Sam Seo, Livegauge

Igor Ikonnikov, Manulife

Andrey Cavalcanti, Microsoft Canada

Dele Ibitoye, RBC

Tess McDonald, Ryerson University

Ayse Bener, Ryerson University

Dalia Shanshal, Ryerson University

Natalie kasparian, Ryerson University

Tamer Abdou, Ryerson University

Abidin Akkok, Ryerson University

Gerri Sinclair, Ryerson University

Dr. Fred Popowich, Simon Fraser University

Glenda Crisp, TD Securities

Peter Husar, TD Securities

Shariyar Murtaza, TELUS

Bryan Smith, ThinkData Works

David Orzel, TMX Group

Eric Sinclair, TMX Group

Thomas Wadden, TMX Group

Gregory Richards, University of Ottawa

Darwin Sauer, Vancity

Marco Wu, Vancity

Robin Mathews-Kanhai, Vancity

Bernie Ip, YVR

Bibliography

- [1] EMC Corporation. *Monetizing your data to create new revenue streams*. Tech. rep. 2011.
- [2] Manirul L. *Big Data: Is it under-utilized?* - 3 Roots Studios. 2016. url: <http://www.3rootsstudios.com/is-big-data-under-utilized/>.
- [3] Scott Brinker. *Data is the most underutilized asset in marketing - Chief Marketing Technologist*. 2013. url: <https://chiefmartec.com/2013/08/data-is-the-most-underutilized-asset-in-marketing/>
- [4] R Walker. *From Big Data to Big Profits: Success with Data and Analytics*. New York, USA: Oxford University Press, 2015, p. 312. isbn: 978-0199378326.
- [5] James Platt, Robert Souza, Enrique Checa, and Ravi Chabalda. *bcg.perspectives - Seven Ways to Profit from Big Data as a Business*. 2014. url: <https://www.bcg.com/en-ca/publications/2014/technology-digital-seven-ways-profit-big-data-business.aspx>
- [6] Virginia Lunt. *A Marketer's Guide to Using Third-Party Data for Programmatic Advertising - 360i Digital Agency Blog*. 2015. url: <http://blog.360i.com/media-planning-buying/a-marketers-guide-to-using-third-party-data-for-programmatic-advertising>
- [7] KPMG. "Framing a winning data monetization strategy". In: KPMG International Cooperative (2015).
- [8] Brad Brown, Michael Chui, and James Manyika. "Are you ready for the era of 'big data'?" In: *McKinsey Quarterly* 4.1 (2011), pp. 24–35.
- [9] Kevin Normandeau. *Big data volume, variety, velocity and veracity*. 2013. url: <http://insidebigdata.com/2013/09/12/beyond-volume-variety-velocity-issue-big-data-veracity/>.
- [10] Antony Adshead. *Big data storage: Defining big data and the type of storage it needs*. 2013. url: <https://www.computerweekly.com/podcast/Big-data-storage-Defining-big-data-and-the-type-of-storage-it-needs>
- [11] Ross Mason, John D'emic, Ken Yagen, Dan Diephouse, and Alex Li. "Big Data's Velocity and Variety Challenges". In: *MuleSoft* (). url: <https://www.mulesoft.com/lp/whitepaper/soa/big-data>
- [12] Eric Suavity. *The Big Cost Of Big Data*. 2012. url: <https://www.forbes.com/sites/ciocentral/2012/04/16/the-big-cost-of-big-data/#58dbe18e5a3b>
- [13] Ellen Martin. *EMCVoice: The Ethics Of Big Data*. 2014. url: <https://www.forbes.com/sites/emc/2014/03/27/the-ethics-of-big-data/#345cdab76852>
- [14] Collins. Michael. *How to Build a Data Assessment Business Case*. 2014. url: <http://www.enterpriseappstoday.com/data-management/how-to-build-a-data-assessment-business-case.html>
- [15] Oracle Corporation. "Bringing the Value of Big Data to the Enterprise". In: *Oracle Magazine* (2013).
- [16] Erhard Rahm and Hong H Do. "Data Cleaning: Problems and Current Approaches". In: *IEEE Data Eng. Bull.* 23.4 (2000), pp. 3–13. doi: 10.1.1.101.1530. url: <http://sites.computer.org/debull/A00DEC-CD.pdf>

- [17] James Manyika, Michael Chui, Peter Groves, Diana Farrell, Steve Van Kuiken, and Elizabeth Almasi Doshi. "Open Data: Unlocking Innovation and Performance with Liquid Information". In: *McKinsey* October (2013), P. 24. url: https://www.mckinsey.com/~/media/mckinsey/business%20functions/mckinsey%20digital/our%20insights/open%20data%20unlocking%20innovation%20and%20performance%20with%20liquid%20information/mgi_open_data_fullreport_oct2013.ashx
- [18] UrbanTide. *Open Data - Is the Open Private Sector the Next Frontier?* 2016. url: <https://urbantide.com/fullstory2/2016/10/24/open-data-is-the-open-private-sector-the-next-frontier>
- [19] Susan Moore. *How to Monetize Your Customer Data - Smarter With Gartner*. 2015. url: <http://www.gartner.com/smarterwithgartner/how-to-monetize-your-customer-data/>.
- [20] Albert Opher. *THINK Driving Value Through IoT Data Monetization*. 2016. url: <https://www.ibm.com/blogs/think/2016/03/iot-data-monetization/>
- [20.c] Tess McDonald, "Personalization in Retail Marketing," Capstone Project for CKME 136 Course, August 2016, Chang School, Ryerson University.
- [21] S. Iyer and P. Soral. *Product substitution search method*. 2000. url: <https://patents.google.com/patent/WO2000079453A2/en>
- [22] Joshua Goldfein. *Big Data and Marketing: Value, Problems, and Solutions*. 2016. url: <http://www.mercurycreative.net/blog/digital/big-data-marketing>.
- [23] Boyd, ;, D Crawford, K Rubinstein, I S Hartzog, W Selinger, and E Ur. 2015. "The Big Data Industry 1 2." *Communication & Society Stanford Law Review Online Stanford Law Review Online* 14153 (466): 1309-56.
- [24] Hsinchun Chen, Roger H L Chiang, Carl H Lindner, Veda C Storey, and J Mack Robinson. "BUSINESS INTELLIGENCE AND ANALYTICS: FROM BIG DATA TO BIG IMPACT". In: *MIS quarterly* 36.4 (2012), pp. 1165-1188.
- [25] IBM Global Business Services Business Analytics and Optimization Analytics. "The real-world use of big data in financial services". In: *IBM Global Services* (2013).
- [25.c] Malcom White, "Analysis of Financial Conference Call Transcripts," Capstone Project for CKME 136 Course, December 2016, Chang School, Ryerson University.
- [26] Bernard Marr. *Big Data In Banking: How Citibank Delivers Real Business Benefits With Its Data- First Approach*. 2016. url: <https://www.forbes.com/sites/bernardmarr/2016/09/09/big-data-in-banking-how-citibank-delivers-real-business-benefits-with-their-data-first-approach/#6acd166947e0>
- [27] Chuck Schaeffer. *How Banks and Insurance Companies are using Big Data to grow Sales*. 2016. url: <http://www.crmsearch.com/fiserv-sales.php>.

- [28] Nii Ayi Armah. “Big Data Analysis: The Next Frontier.” In: *Bank of Canada Review* (2013), pp. 32–39. issn: 00451460. url: <http://ra.ocls.ca/ra/login.aspx?url=https://www.bankofcanada.ca/wp-content/uploads/2013/08/boc-review-summer13-armah.pdf>
- [29] Daniel Tencer. “History Repeating Itself? Toronto’s Long Record Of Housing Busts”. In: *The Huffington Post Canada* (2017). url: https://www.huffingtonpost.ca/2017/01/18/house-prices-toronto-historic-boom-bust_n_14230574.html
- [29.c] Tess McDonald, “Case Study on Toronto’s Real Estate Market,” Project for A Vision for Data Monetization, December 2016, Chang School, Ryerson University.
- [30] Lohr, S 2012, “The Age of Big Data”, New York Times 11 February, accessed 4 May 2017, <https://www.nytimes.com/2012/02/12/sunday-review/big-datas-impact-in-the-world.html>
- [31] Reference is: Wu, L., & Brynjolfsson, E. (2009). The Future of Prediction: How Google Searches Foreshadow Housing Prices and Sales. SSRN Electronic Journal. url: <https://doi.org/10.2139/ssrn.2022293>
- [32] Danyang Du, Aihua Li, and Lingling Zhang. “Survey on the applications of big data in Chinese real estate enterprise”. In: *Procedia Computer Science*. 2014. isbn: 8618511897. doi: 10.1016/j.procs.2014.05.377.
- [33] Bloomberg Real Estate Management and Development. (2017). Company Overview of Shenzhen Fantasia Real Estate Group Co., Ltd. Retrieved from <http://www.bloomberg.com/research/stocks/private/snapshot.asp?privcapId=223402215>
- [34] Douglas Burdick, Michael Franklin, Paulo Issler, Rajasekar Krishnamurthy, Lucian Popa, Louiqa Raschid, Richard Stanton, and Nancy Wallace. “Data Science Challenges in Real Estate Asset and Capital Markets”. In: *Proceedings of the International Workshop on Data Science for Macro-Modeling* (2014), pp. 1–5.
- [35] Statistics Canada. *Toronto Employment Survey 2015*. Tech. rep. Toronto, Canada: City Planning Division, Strategic Initiatives, Policy & Analysis, 2014. url: www.toronto.ca/demographics/surveys.
- [36] Marnie Wallace, Michael Wisener, and Krista Collins. *Neighbourhood Characteristics and the Distribution of Crime in Regina*. Tech. rep. 85. 2006, pp. 1–62. url: <https://www.statcan.gc.ca/pub/85-561-m/85-561-m2006008-eng.htm>
- [37] Mathieu Charron. *Neighbourhood Characteristics and the Distribution of Crime in Toronto: Additional Analysis on Youth Crime*. Vol. 85. 22. 2011, pp. 1–31. isbn: 9781100195520.
- [38] Ritchey, Diane. “Big data, big security.” *Security* 49, no. 7 (2012): 28–30.
- [39] SentinelOne. “How big data is improving cyber security”. In: CSO (2016). url: <https://www.sentinelone.com/blog/big-data-improving-cyber-security/>
- [40] Sreeranga P. Rajan Alvaro A. Cardenas Pratyusa K. Manadhata. “Big Data Analytics For Security”. In: *IEEE Computers and Reliability Societies* November/December (2013), pp. 74–76. doi: 10.1109/MSP.2013.138

- [44] Bernard Marr. *How Big Data Is Changing Healthcare*. 2015. url: <https://www.forbes.com/sites/bernardmarr/2015/04/21/how-big-data-is-changing-healthcare/#634f19132873>
- [45] Wu He, Shenghua Zha, and Ling Li. “Social media competitive analysis and text mining: A case study in the pizza industry”. In: *International Journal of Information Management* 33 (2013), pp. 464–472. doi: 10.1016/j.ijinfomgt.2013.01.001.
- [46] The Economist Group Limited. *The Arab Spring, Five Years On*. 2016. url: <http://www.economist.com/blogs/graphicdetail/2016/01/daily-chart-8>.
- [47] Zeynep Tufekci. “Big questions for social media big data: Representativeness, validity and other methodological pitfalls”. In: *ICWSM '14: Proceedings of the 8th International AAAI Conference on Weblogs and Social Media* (2014), pp. 505–514.
- [48] Matthew Smith, Christian Szongott, Benjamin Henne, and Gabriele von Voigt. “Big data privacy issues in public social media”. In: *2012 6th IEEE International Conference on Digital Ecosystems and Technologies (DEST)*. IEEE, June 2012, pp. 1–6. isbn: 978-1-4673-1703-0. doi: 10.1109/DEST.2012.6227909.
- [49] Ming-Hsiang Tsou. “Research challenges and opportunities in mapping social media and Big Data”. In: *Cartography and Geographic Information Science* 42.sup1 (2015), pp. 70–74. doi: 10.1080/15230406.2015.1059251.
- [50] Vibha Shukla and Pawan Kumar Dubey. “Big Data: Beyond Data Handling”. In: *International Journal Of Scientific Research And Education* 2.09 (2014), pp. 1929–1935.
- [51] Kaile Zhou and Shanlin Yang. *Understanding household energy consumption behavior: The contribution of energy big data analytics*. 2016. doi: 10.1016/j.rser.2015.12.001.
- [52] Radu F Babiceanu and Remzi Seker. “Service Orientation in Holonic and Multi-Agent Manufacturing”. In: 640 (2016), pp. 157–164. issn: 1860-949X. doi: 10.1007/978-3-319-30337-6.
- [53] Jay Lee, Edzel Lapira, Behrad Bagheri, and Hung-an an Kao. “Recent advances and trend in predictive manufacturing systems in big data environment”. In: *Manufacturing Letters* 1.1 (2013), pp. 38–41. issn: 22138463. doi: 10.1016/j.mfglet.2013.09.005.
- [54] Bogdan Nedelcu. “About Big Data and its Challenges and Benefits in Manufacturing”. In: *Database Systems Journal* IV.3 (2013), pp. 10–19. issn: 20693230.
- [55] Eric Auschitzky, Markus Hammer, and Agesan Rajagopaul. “How Big Data Can Improve Manufacturing”. In: *McKinsey and Company* (2014). url: <http://www.mckinsey.com/business-functions/operations/our-insights/how-big-data-can-improve-manufacturing>.
- [56] Katina Michael and Keith W Miller. “Big data: New opportunities and new challenges”. In: *Computer* 46.6 (2014), pp. 22–24. issn: 00189162 (ISSN).

- [41] D Kemmerer and G Vigna. "Intrusion detection: a brief history and overview". In: *Computer-IEEE Computer Magazine* (2002).doi: 10.1109/MC.2002.1012428
- [42] Bernard Marr. *How Big Data Is Used To Fight Cyber Crime And Hackers: Fascinating Use Case From BT*. 2016. <https://www.forbes.com/sites/bernardmarr/2016/12/01/how-big-data-is-used-to-fight-cyber-crime-and-hackers-fascinating-use-case-from-bt/#6bb9468076b9>
- [43] Richard J. Bolton, David J. Hand, Foster Provost, Leo Breiman, Richard J. Bolton, and David J. Hand. "Statistical Fraud Detection: A Review". In: *Statistical Science* 17.3 (2002), pp. 235–255. issn: 08834237. doi: 10.1214/ss/1042727940. url: <http://www.jstor.org/stable/i359432>
- [57] Gang-Hoon Kim, Silvana Trimi, and Ji-Hyong Chung. "Big-Data Applications in the Government Sector". In: *Association for Computing Machinery. Communications of the ACM* 57.3 (2014), p. 78. issn: 00010782. doi: 10.1145/2500873. url: <http://search.proquest.com/docview/1516150205?accountid=34461>.
- [58] Chris Yiu. "The Big Data Opportunity". In: *Policy Exchange* (2012), p. 34. url: <http://www.policyexchange.org.uk/images/publications/thebigdataopportunity.pdf>.
- [59] Liran Einav and Jonathan Levin. "Economics in the age of big data". In: *Science* 346.6210 (2014). issn: 0362-4331. doi: 10.1126/science.1243089.
- [60] PUBLIC SECTOR EXECUTIVES NETWORK. "Open Data and Big Data Analytics: Governments on the Dance Floor". In: *The Conference Board of Canada* (2015).
- [61] Varun Singh, Ishan Srivastava, and Vishal Johri. "Big Data and the Opportunities and Challenges for Government Agencies". In: *International Journal of Computer Science and Information Technologies* 5.4 (2014), pp. 5821–5824.
- [62] History.com Staff. *Boston Marathon Bombings*. 2014. url: <http://www.history.com/topics/boston-marathon-bombings>.



We would like to thank the following organizations for their leadership and in-kind contributions to Canada's Big Data Consortium, Canada's Big Data Talent Gap Study, and to this paper, "A Vision for Data Monetization in Canada."

