Emerging Technology Series

Big Data Analytics in Health

White Paper (Full Report)

Canada Health Infoway
Inforoute Santé du Canada

April 2013

1 Disclaimer

This white paper represents solely the views of Canada Health Infoway (*Infoway)*. It is based on *Infoway's* research and analysis as well as information from various sources. *Infoway's* views are based on information and analysis which *Infoway* believes is sound and reliable, as of the publication date of this white paper.

This white paper is informative only and cannot be interpreted as providing any indication of *Infoway's* present or future strategies or investment criteria.

This white paper is provided as is. No representation or warranty of any kind whatsoever is made by *Infoway* as to the accuracy, infringement of third party intellectual property, completeness, fitness for any reader's purpose, or correctness of any information or other contents contained in the white paper, and *Infoway* assumes no responsibility or liability if there is any inaccuracy, infringement of third party intellectual property, incompleteness, failure to meet any reader's purpose or incorrectness with respect to any of the information or other contents contained in the white paper.

Infoway does not assume any responsibility or liability related directly or indirectly to the white paper, including without limitation with respect to any person who seeks to implement or implements or relies or complies with any part or all of the ideas, recommendations or suggestions set forth in the white paper.

Infoway does not implicitly or explicitly endorse any particular technology or solution of any vendor or any other person, it does not guarantee the reliability or any proposed results related to the use of such technology or solution and this notwithstanding that reference may be made directly or indirectly to any such technology or solution in the white paper.

Infoway does not make any implicit or explicit commitment of any kind or nature whatsoever to make any investment in any particular technology or solution, and this notwithstanding that reference may be made directly or indirectly to any such technology or solution in the white paper.

Anyone using the enclosed material should rely on his/her/its own judgment as appropriate and seek the advice of competent professionals and experts.

Table of Contents

1 DISCLAIMER 2

2 EXECUTIVE SUMMARY 5

2.1 Introduction 5

2.2 BDA Defined 6

2.3 Big Data in Health Care 7

2.4 The Economic Value of BDA 8

2.5 Opportunities for BDA in Health Care 8

2.6 Challenges in the Introduction of BDA 9

2.7 A Call to Action 10

3 BDA DEFINED 12

3.1 BDA Characteristics 13

3.2 Examples of BDA Use Cases in Other Industries 17

4 ECONOMIC VALUE AND OPPORTUNITIES FOR BDA IN DIGITAL HEALTH 19

4.1 International BDA Investments in Digital Health 22

4.2 Opportunities for BDA in Digital Health in Canada 23

4.3 Implementation Progress of BDA within the Health Care Sector 24

5 CONSIDERATIONS FOR BDA IN DIGITAL HEALTH 30

5.1 BDA Business Considerations 30

5.2 BDA Governance and Legislation Considerations 31

5.3 BDA Operational Considerations 32

5.3.1 Strategic Operational Considerations 32

5.3.2 Tactical Operational Considerations 34

5.4 BDA Process Considerations 36

5.5 BDA People Considerations 37

5.6 Privacy and Security, Ethics and Legal Considerations 38

5.7 BDA Technical and Deployment Considerations 39

5.7.1 Technical and Infrastructure Considerations 39

5.7.2 Technical Deployment Considerations 42

6 NEXT STEPS FOR BDA IN DIGITAL HEALTH 46

6.1 BDA: Call to Action in Health Care 46
6.2 BDA: Next Steps for the Near and Short Term............................... 49
6.3 BDA: Next Steps in the Medium Term... 50
6.4 BDA: Next Steps in the Long Term.. 51

7 CONCLUSION 52

8 BIBLIOGRAPHY.. 53

APPENDIX A – ACRONYMS AND ABBREVIATIONS 55

APPENDIX B – GLOSSARY.. 56

APPENDIX C – BDA TECHNICAL FRAMEWORKS 59

APPENDIX D – GENERAL INDUSTRY EXAMPLES OF BDA 62

APPENDIX E – ADDITIONAL HEALTH CARE EXAMPLES 65

9 CONTACT 68

2 Executive Summary

2.1 Introduction

Big data analytics (BDA) has emerged from two distinct concepts – big data and analytics. Together it represents a new information management approach that has been designed to derive previously untapped intelligence and insights from data to address many new and important questions. Within the health sector, it provides stakeholders[1] with new insights that have the potential to advance personalized care, improve patient outcomes and avoid unnecessary costs.

> *Big data analytics is a new information management approach and set of capabilities for uncovering additional value from health information.*

This white paper defines big data analytics and its characteristics, comments on its economic value, opportunities and challenges in health care, and provides recommendations to stakeholders on next steps.

Big data has become the new frontier of information management given the amount of data today's systems are generating and consuming. It has driven the need for technological infrastructure and tools that can capture, store, analyze and visualize vast amounts of disparate structured and unstructured data.[2] These data are being generated at increasing volumes from data intensive technologies including, but not limited to, the use of the Internet for activities such as accesses to information, social networking, mobile computing and commerce. Corporations and governments have begun to recognize that there are unexploited opportunities to improve their enterprises that can be discovered from these data.

Analytics when applied in the context of big data is the process of examining large amounts of data, from a variety of data sources and in different formats, to deliver insights that can enable decisions in real or near real time. Various analytical concepts such as data mining, natural language processing, artificial intelligence and predictive analytics can be employed to analyze, contextualize and visualize the data. Big data analytical approaches can be employed to recognize inherent patterns, correlations and anomalies which can be discovered as a result of integrating vast amounts of data from different data sets.

[1] Refers to a broad group of affected parties, including: governors (health ministries and departments), administrators (health service delivery organizations, local and regional health networks), clinicians (physicians, specialists, nurses, pharmacists, technicians), consumers (general public, individuals and groups), researchers (individuals and organizations), and vendors.

[2] Structured data is well organized, well defined and computer identifiable data. It is easily processable by a computer. For example, a number would be uniquely identified as to what it represents, say a body temperature or body weight. Those values may also be coded with a controlled medical vocabulary. Unstructured data has ambiguity in a computer processing sense and may even be free text.

There are many potential use cases for BDA in health care. BDA can be used to: help researchers find causes of, and treatments for diseases; actively monitor patients so clinicians are alerted to the potential for an adverse event before it occurs; and personalize care so precious resources associated with a treatment are not administered to a patient who cannot benefit from the intervention.

Until the advent of BDA solutions, scientists, government statisticians and market researchers routinely encountered problems with analyzing their large and complex data sets. The fact is that even with BDA, many will still encounter problems for some time. Like any new technology, BDA platforms and related programs require significant capital and operational investments, as well as people with specialized skills to establish and operate them. Realistically only the largest enterprises and health care corporations are likely able to afford to make the necessary investments in these technologies. For health enterprises that are fortunate enough to lead in this space, it often plays out as time limited pilot projects that are anchored by research and innovation grants.

Nevertheless, the potential return on investment – estimated to represent as much as five to six per cent of the hundreds of billions of dollars[3] Canadian governments spend on health care each year – is too promising to ignore.

2.2 BDA Defined

"Big data" is a term used to describe a collection of data sets with the following three characteristics:

- Volume – large amounts of data generated;
- Velocity – frequency and speed of which data are generated, captured and shared; and
- Variety – diversity of data types and formats from various sources.

The size and complexity of big data makes it difficult to use traditional database management and data processing tools. Data is being created in much shorter cycles, from hours to milliseconds. There is also a trend underway to create larger databases by combining smaller data sets so that data correlations can be discovered.

Big data provides new opportunities to store and index previously unusable, siloed and unstructured data for additional uses by health care stakeholders.

Analytics creates new business value by transforming this previously unusable data into new predictive insights and actionable knowledge.

[3] Analytics: The New Path to Value, MIT Sloan Management Review, IBM Global Business Services, Business Analytics and Optimization, By *Steve LaValle, Michael Hopkins, Eric Lesser, Rebecca Shockley and Nina Kruschwitz.*

Given these characteristics, big data requires the use of new frameworks, technologies and processes to manage it. Yet its arrival in the enterprise software space has created some confusion as business leaders try to understand the differences between it and traditional data warehousing (DW) and business intelligence (BI) tools.

There are important distinctions and sufficient differentiating value between BDA and DW/BI systems which make BDA unique.

Gartner defines a data warehouse as "a storage architecture designed to hold data extracted from transaction systems, operational data stores and external sources. The warehouse then combines that data in an aggregate, summary form suitable for enterprise-wide data analysis and reporting for predefined business needs."[4]

Forrester Research has defined business intelligence as "a set of methodologies, processes, architectures, and technologies that transform raw data into meaningful and useful information used to enable more effective strategic, tactical, and operational insights and decision-making."[5]

BDA solutions will not replace DW/BI, rather they will co-exist side-by-side to unlock hidden value in the massive amount of data that exists within and outside the enterprise.

BDA functions are unique because they:

- handle open ended "how and why" type questions whereas BI tools are designed to query specific "what and where"; and
- process unstructured data to find patterns, whereas DW systems process structured and mostly aggregated data.

2.3 Big Data in Health Care

The types of data anticipated to be of use in BDA include:

- Clinical data – up to 80 per cent of health data is unstructured as documents, images, clinical or transcribed notes;
- Publications – clinical research and medical reference material;
- Clinical references – text-based practice guidelines and health product (e.g., drug information) data;
- Genomic data – represents significant amounts of new gene sequencing data;
- Streamed data – home monitoring, telehealth, handheld and sensor-based wireless or smart devices are new data sources and types;
- Web and social networking data – consumer use of Internet – data from search engines and social networking sites; and
- Business, organizational and external data – administrative data such as billing and scheduling and other non-health data.

[4] http://www.gartner.com/it-glossary/data-warehouse/Accessed January 21, 2013.

[5] Evelson, Boris (November 21, 2008), "Topic Overview: Business Intelligence."

There are many sources of big data within the health sector; however, it is unrealistic to assume that all data can be put to use for BDA due to a range of operational and technical challenges and privacy considerations.

2.4 The Economic Value of BDA

A study conducted by McKinsey Global Institute (MGI) [6]in 2011 indicates that BDA can potentially transform the global economy, make significant improvements to organizational performance and work to improve national and international policies. For health care specifically, MGI predicted that if U.S. health care organizations were to use BDA creatively and effectively to drive efficiency and quality, the sector could create more than $300 billion in value annually.

McKinsey's research points out valuable insights such as patient behaviours, along with demands and efficiencies about the environment surrounding the patient, are buried in unstructured or highly varied data sources. The report cites successful pilot projects in the U.S. and internationally that have used BDA to find efficiencies in clinical operations, analyze data from remotely monitored patients, assess clinical and cost efficiencies of new treatments, and use analytics in public health surveillance and disease response.

A significant component of the $300 billion in forecasted value in the MGI report comes from clinical operations and how BDA may affect the way clinical care is provided. Examples include:

- outcomes-based research to determine which treatments will work best for specific patients ("optimal treatment pathways") by analyzing comprehensive patient and outcome data to compare the effectiveness of various interventions;
- pre-diagnosis that automatically mines medical literature to create a medical expertise database capable of suggesting treatment options to clinicians based on patients' health records; and
- remote patient monitoring for chronically ill patients and analyzing the resulting data to monitor treatment adherence, reduce patient in-hospital bed days, cut emergency department visits, improve the targeting of nursing home care and outpatient physician appointments, and reduce long-term health complications.

2.5 Opportunities for BDA in Health Care

Big data analytics represents a new approach to analytics. It does not yet have a large or significant footprint in Canada or internationally. However, the continuing digitization of health records together with the interoperable electronic health record (EHR), presents new opportunities to investigate a myriad of clinical and administrative questions.

New insights derived from big data analytics will serve to advance personalized care, improve patient outcomes and avoid unnecessary costs.

[6] Source: MGI Big Data The next frontier for innovation, competition, and productivity, June 2011.

There is potential to layer BDA-type applications, in a privacy-protective manner, on top of the foundational health IT infrastructure to derive value that might not otherwise be found. What follows are some innovative ideas and solutions.

- Clinical decision support – BDA technologies that sift through large amounts of data, understand, categorize and learn from it, and then predict outcomes or recommend alternative treatments to clinicians and patients at the point of care.
- Personalized care – Predictive data mining or analytic solutions that can leverage personalized care (e.g., genomic DNA sequence for cancer care) in real time to highlight best practice treatments to patients. These solutions may offer early detection and diagnosis before a patient develops disease symptoms.
- Public and population health – BDA solutions that can mine web-based and social media data to predict flu outbreaks based on consumers' search, social content and query activity. BDA solutions can also support clinicians and epidemiologists performing analyses across patient populations and care venues to help identify disease trends.
- Clinical operations – BDA can support initiatives such as wait-time management, where it can mine large amounts of historical and unstructured data, look for patterns and model various scenarios to predict events that may affect wait times before they actually happen.
- Policy, financial and administrative – BDA can support decision makers by integrating and analyzing data related to key performance indicators.

2.6 Challenges in the Introduction of BDA

For BDA to be successful in health care, it needs to be accompanied by a range of enablers, some of which may require a substantial rethinking of the way health care is provided and funded. The following represent enablers as well as challenges, risks and barriers that stakeholders should be aware of.

- Governance – BDA impacts existing legislation, governance and information management processes within and outside the enterprise.
- Funding models – Incentive funding can be reallocated to clinicians who sustain high quality services from clinicians who fall below accepted standards.
- Business model – How can organizations measure the return on investment when there is insufficient business case evidence in health at this time?
- Data custodianship – Who will own, operate and govern BDA concepts like personalized care, streaming sensor and device data, and social media data?
- Skilled resources – There will be a need to have the required skills and competencies to be able to execute BDA. There is already a global skills shortage in the data scientist and data analyst roles.
- Privacy and security – BDA uses challenge our traditional definitions of collection, use and disclosure of personal health information, highlighting a need to review existing policies to address privacy and security, legal and ethics considerations.

2.7 A Call to Action

Like all new technologies, BDA is being introduced to the health sector with much promise. However, its capabilities, opportunities and benefits still need to be proven through clinical and administrative applications. In light of these challenges, for those who decide to invest in BDA, *Infoway* strongly advises stakeholders to invest through scope-limited pilot projects. Iterative and step-wise project investments will limit capital requirements and permit stakeholders to study the feasibility and utility of the new capabilities before committing additional resources. Pilot projects should focus on providing insights into specific and pressing clinical or business problems which are supported by clear use cases, while allowing for some experimentation. Wherever possible, existing infrastructure assets, information sets and partnerships should be leveraged.

For those who decide to invest in BDA strongly advises stakeholders to cautiously invest through:

a) scope-limited pilot projects that focus on solving specific and pressing clinical or business problems which are supported by clear use cases, while allowing for some experimentation; and

b) iterative, step-wise project investments to limit capital requirements and permit stakeholders to study the feasibility and utility of the new capabilities before committing additional resources.

For BDA to move from the periphery to mainstream health service delivery, *Infoway* suggests several foundational steps be considered:

- Identify champions and representatives from the clinical, administrative, business, privacy, data governance and technical roles, as well as vendors and consumers;
- Invest in the learning and development of new data analytics professionals;
- Establish collaborations with scientific communities working in BDA, and form communities of interest especially with peers in their particular sector;
- Focus initial efforts on developing or adapting other stakeholder organizations' use cases, business cases, strategies and road maps, data governance, privacy and security approaches, and BDA deployment models;
- Form partnerships with trusted public and/or private sector organizations that have common interests, experience, BDA solutions and services;
- Understand how to leverage and invest in the BDA research agenda in Canada, especially as it applies to health care;
- Assess the lessons learned from initial BDA pilots to understand how to approach future investments, experiment with analytics, determine resource and skill requirements, and determine changes to data collection processes;
- Look for new greenfield opportunities where innovative technologies may come together (e.g., remote patient monitoring devices that transmit sensor data to the cloud where BDA is used to process it).

Infoway believes that BDA will continue to evolve and become a prominent component of most stakeholders' information management plans and activities. Therefore, in the near to medium term (3-5 years), stakeholders should address key challenges to BDA:

- Data policies – Policy, legislative and governance may need to be reviewed, including, but not limited to: privacy, security, intellectual property and liability;
- Legacy technology and techniques – Stakeholders may need to consider redeploying or refreshing these technologies;
- Business and operational models – Models that tackle the cost curve head-on must be prioritized over going it alone with in-house programs;
- Culture of innovation and experimentation – Enabling experimentation to discover new needs and to improve performance requires cultural alignment around innovative use of health information.

Not surprisingly, in many ways this innovation journey is not about the BDA tools or methods that will enable transformational benefits, but about the people and processes that surround it. While cautious first steps have been recommended, the application of BDA at scale in health care has the potential to yield significant benefits for future generations of Canadians.

3 BDA Defined

What is BDA?

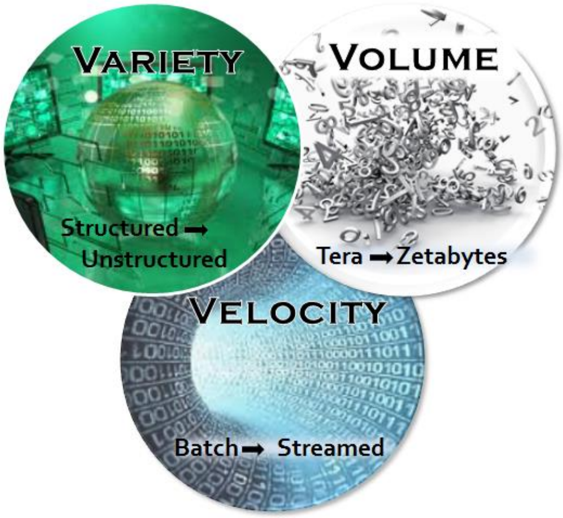
"Big data" is a term used by the IT industry to describe the voluminous amount of unstructured data an organization creates. It represents information that has not been normalized or harmonized, comes from many different sources, and in the past has been too expensive or not practical operationally to normalize for typical online transactional processing (OLTP) or data warehouse type data stores. Big Data (BD) has the characteristic of vast size that exceeds the capability of traditional data management technologies and requires the use of new capabilities and processes to source, process and manage it.

Big data provides new opportunities to store and index previously unusable, siloed and unstructured data for additional uses by health care stakeholders.

Analytics creates new business value by transforming this previously unusable data into new predictive insights and actionable knowledge.

BD is described using three terms:

- **Volume** is the amount of data generated by organizations or individuals. Enterprises in all industries are looking for ways to handle the ever-increasing data volume that's being created every day.
- **Velocity** is the frequency and speed at which data is generated, captured and shared. Consumers as well as businesses now generate more data and in much shorter cycles, from hours, minutes, seconds down to milliseconds.



- **Variety** is the proliferation of new data types including those from social, machine and mobile sources. New types include content, location or geo-spatial, hardware data points, log data, machine data, metrics, mobile, physical data points, process, radio frequency identification (RFID), search, sentiment, streaming data, social, text and web. Also, variety includes traditional unstructured clinical data (i.e., free text).

The term “Analytics” refers to the logic and algorithms, both deduction and inference, performed on BD to derive value, insights and knowledge from it. Analytical methods such as data mining, natural language processing, artificial intelligence and predictive analytics are employed to analyze, contextualize and visualize the data. These computerized analytical methods recognize inherent patterns, correlations and anomalies which are discovered as a result of integrating vast amounts of data from different datasets.

Together, the term “Big Data Analytics” represents, across all industries, new data-driven insights which are being used for competitive advantage over peer organizations to more effectively market products and services to targeted consumers. Examples include real-time purchasing patterns and recommendations back to consumers, and gaining better understandings and insights into consumer preferences and perspectives through affinity to certain social groups.

The origin of BDA comes from web-based search engines such as Google and Yahoo, the popularity of social media and social networking services such as Facebook and Twitter, and data-generating sensors, telehealth and mobile devices. All have increased and generated new data and opportunities for new insights on customer behaviours and trends. While BDA frameworks have been in operation since 2005, they have just recently moved into other industries and sectors including financial services firms and banks, online retailers and health care.

For health care, BDA represents opportunities to exploit personalized care, streamline health operations, support clinical and policy decision making, and improve patient engagement.

3.1 BDA Characteristics

Today, across all industries, the typical sources[7] of BD include:

- **Internet transactions** – By 2015, more than three billion people will be online. Billions of online purchases, stock trades, social networking exchanges, Internet searches and other transactions happen every day, including countless automated transactions. Each creates a number of data points collected by retailers, banks, credit card issuers, credit agencies, social networking and search engine service providers and others.
- **Mobile devices** – There are more than 5.6 billion mobile phones in use worldwide. Each call, text and instant message is generating data. The average teen texts 4,700 times per month. Mobile devices, particularly smart phones and tablets, also make it easier to use social networking and other data-generating applications. Mobile devices also collect and transmit location data.
- **Social networking and media** – There are currently more than 955 million active Facebook users, 500 million Twitter users and 156 million public blogs. By 2015, more than two billion videos will be watched over YouTube in one day.

[7] Estimates come from a variety of research, including Gartner, IDC, Forrester, TDWI and materials from IT vendors such as IBM, ORACLE, SAP, SAP, Cloudera, Greenplum and Vertica.

Each Facebook update, tweet, blog post and comment creates multiple new data points – structured, semi-structured and unstructured – sometimes referred to as data exhaust.

- **Networked devices and sensors** – Electronic devices of all sorts – including servers and other IT hardware, smart energy meters and temperature sensors, patient monitors and aides – all create semi-structured log data that record every action.

Specific to health care, the types of data anticipated to be available for use by BDA include:

- **Genomic data** – Represents significant amounts of new gene sequencing data being made available through new investments, BDA capabilities and business models.
- **Streamed data** – Home monitoring, telehealth, handheld and sensor-based wireless and smart devices are new data sources and types. They represent significant amounts of real time data available for use by the health system.
- **Web and social networking-based data** – Web-based data comes from Google and other search engines, consumer use of the Internet, as well as data from social networking sites.
- **Health publication and clinical reference data** – This includes text-based publications (clinical research and medical reference material) and clinical text based reference practice guidelines and health product (e.g., drug information) data.
- **Clinical data** – Eighty per cent of health data is unstructured as documents, images, clinical or transcribed notes. These semi-structured to unstructured clinical records and documents represent new data sources.
- **Business, organizational and external data** – Data which previously has not been linked, such as financial, billing, scheduling, administrative, external and other non-clinical and non-health data.

It is important to note that while there are many sources of BD within the health sector, it is unrealistic to assume that all data can be put to use for BDA due to a range of governance, privacy, operational and technical considerations.

Gartner Group's analysis of BDA shows that vendors are enabling BDA with a wide variety of new and old technologies, in different ways and at different rates. Overall, Gartner depicts an IT market that is still fairly immature, with larger traditional DW/BI entities engaged and investing millions of dollars, and smaller BD pure-players ramping up their go-to-market strategies purely focused on BDA. Gartner's research points to a marketplace in the early adopter phase, despite the large valuation[8] of $5 billion (US).

[8] Valuation represents total estimated vendor revenues for BDA Market.

Gartner points out that outside of some new frameworks for BDA, no one DW/BI component will enable BDA by itself. Most of these will complement BDA to enable benefits from mining and analytics performed on the data. In addition, BDA will be extremely broad across organizations, as it will have implications on business, IT, information management capacities, skillsets and utilization. Gartner also predicts BDA will introduce a new need to address existing information management processes inside an organization. Gartner predicts those who are able to do this will be more successful at BDA. For more information on Gartner's research and hype curve, please refer to Gartner's web site. See: http://www.gartner.com/technology/research/hype-cycles/.

Separately, International Data Corporation[9] (IDC) released a worldwide BDA technology and services forecast showing the market is expected to grow from $3.2 billion in 2010 to $16.9 billion in 2015. This represents a compound annual growth rate (CAGR) of 40 per cent, or about seven times that of the overall information and communications technology (ICT) market. Other industry experts such as Wikibon[10] and IT investment analysts such as UBS[11] take this prediction further by forecasting that the BDA market will grow at an astounding CAGR of 58 per cent between now and 2017, hitting the $50 billion mark within five years.

While BDA is fairly new, some components being leveraged by BDA have existed for several years (e.g., data integration software that moves data, approaches to process and analyze text based data, and content management and document management for managing unstructured data). The size and complexity of BD makes it difficult to use traditional database management and data processing tools. This issue is being compounded by the growth in data generated by consumer, enterprise medical devices and digitized patient records where the majority of data are in different formats. Data are being created in much shorter cycles, from hours to milliseconds. There is also a trend underway to create larger datasets by combining smaller datasets so that data correlations can be discovered.

Nonetheless, the arrival of BDA in the enterprise software space has created some confusion as business leaders try to understand the differences between it and traditional data warehousing and business intelligence (DW/BI) tools. There are important distinctions and sufficient differentiating value between BDA and DW/BI systems which make BDA unique.

BDA frameworks are not intended to replace traditional analytics capabilities for structured relational data. They are meant to be complementary, as traditional DW/BI tools do not deal with unstructured data, or the volume of big data.

[9] International Data Corporation (IDC) Worldwide BDA technology and services forecast, 2010. Refer to http://www.idc.com/getdoc.jsp?containerId=IDC_P23177.

[10] Wikibon is a professional community solving technology and business problems through an open source sharing of free advisory knowledge.

[11] UBS is a global investment firm providing highly specialized research in response to the changing demands of global client base.

Gartner defines a data warehouse as "a storage architecture designed to hold data extracted from transaction systems, operational data stores and external sources. The data warehouse then combines that data in an aggregate, summary form suitable for enterprise-wide data analysis and reporting for predefined business needs."[12] Forrester Research defines business intelligence as "a set of methodologies, processes, architectures and technologies that transform raw data into meaningful and useful information used to enable more effective strategic, tactical and operational insights and decision-making."[13]

BDA should be seen as complementary, not a replacement, for traditional DW/BI capabilities used for analyzing structured and relational data. These traditional tools cannot deal with unstructured and high volumes and sizes of BD. BDA solutions will co-exist with DW/BI to unlock hidden value in the massive amount of data that exists within and outside the enterprise.

There is also a belief that BDA is the analytics performed on health information from recent implementation and digitization of EHRs and EMRs, or the integration of clinical data with other data (e.g., financial or administrative data). This is not BDA. This is traditional analytics fueled by new structured data sources, new analytic features and existing DW/BI technologies and is in some cases now referred to as "small data".

BDA functions are unique from traditional analytic methods because they:

- support an experimental type of analytics, whereas, traditional DW/BI and statistical analyses are based on answering known questions or hypotheses;
- handle open ended "how and why" type questions, whereas BI tools are designed to query specific "what and where";
- process unstructured data to find patterns whereas DW systems process structured, related and mostly aggregated data;
- process, generate and index large sets of data, handling the complexities of network communication, parallel programming and fault tolerance;
- process large datasets, in a distributed environment, across clusters of computers designed to scale up from single servers to thousands of machines, each having its own local computation and storage;
- systemically mine and flag data that is relevant for other uses and for further analytics by traditional analytics tools;
- leverage analytic concepts such as geo-mapping, data mining and predictive modelling to help with disease outbreak monitoring and forecast modelling;
- deliver results faster than BI systems, even in real time; and
- scale to petabytes and even exabytes of data.

[12] http://www.gartner.com/it-glossary/data-warehouse/Accessed January 21, 2013.

[13] Evelson, Boris (November 21, 2008), "Topic Overview: Business Intelligence."

While BDA frameworks are experimental in nature, they do have some similar characteristics found in knowledge management and health informatics work that has been done by the university and research community. It will be necessary to understand how these knowledge management initiatives align to BDA, especially as it applies to work done around concepts such as semantic web, natural language processing, data mining, visualization and other like capabilities.

Through the use of tools such as natural language processing, text analytics, neural networks and data mining, and visualization, the results of big data analytics can be displayed as tag or word cloud (i.e., a weighted visual list). This type of visualization helps the reader quickly grasp the most salient concepts in a large body of text. From an analytics and visualization perspective, by clicking on one or two of the highlighted words, additional analytics and visualization could be performed on the data to limit a set number of items presented, to highlight commonality or relationships, or predict new patterns based on previous patterns. For more information on data visualization, please refer to section 5.7.1.

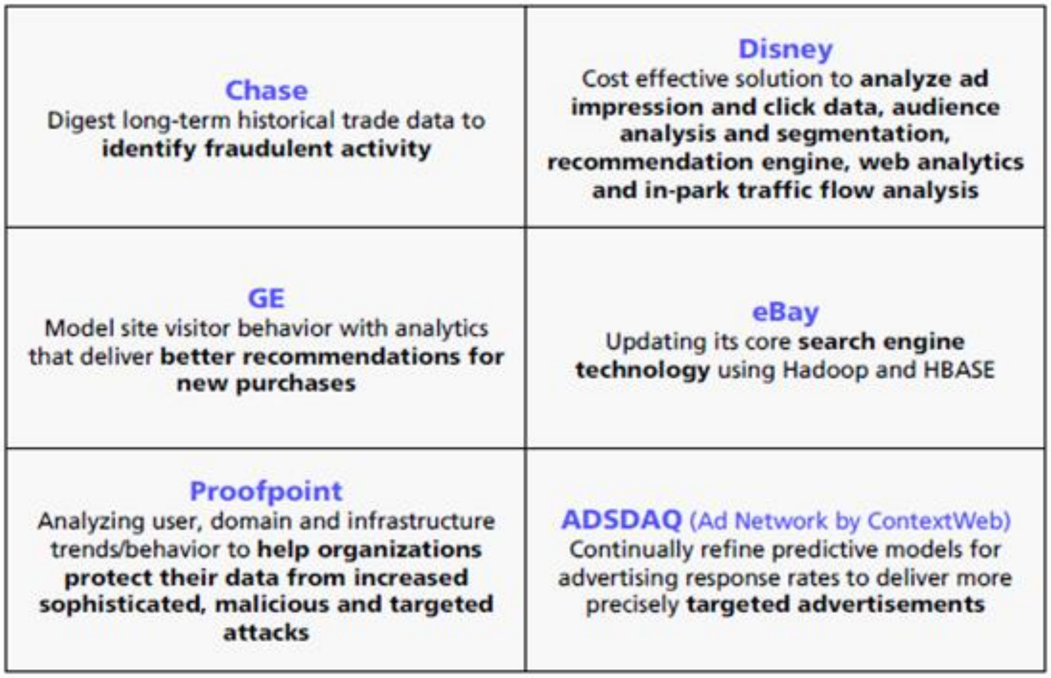
Once the analytics have been completed, the processed data is ready for further analysis by “data scientists” or data analysts. The data are manipulated and analyzed using any of a number of tools to search for hidden insights and patterns or to use as the foundation to build user-facing analytic interfaces or applications. The resulting data can also be modeled and transferred from the BDA environments into existing relational databases, data warehouses and other traditional IT systems for further analysis or to support transactional processing.

For more technical information on BDA, please refer to Appendix C.

3.2 Examples of BDA Use Cases in Other Industries

Facebook is a prime example of a use case for BDA. When a user posts a picture or writes something on his or her page, BDA brings up an ad based on mining the user’s page content or recent postings. Another example of BDA is when Netflix recommends movies to customers based on their previous uses, web navigation and traffic. It’s estimated that 70 per cent of the movies that people pick on Netflix come from a “recommendation engine” powered by BDA.

Figure 1 – Examples of BDA in Other Industries[14]



Chase Digest long-term historical trade data to **identify fraudulent activity**	Disney Cost effective solution to **analyze ad impression and click data, audience analysis and segmentation, recommendation engine, web analytics and in-park traffic flow analysis**
GE Model site visitor behavior with analytics that deliver **better recommendations for new purchases**	eBay Updating its core **search engine technology** using Hadoop and HBASE
Proofpoint Analyzing user, domain and infrastructure trends/behavior to **help organizations protect their data from increased sophisticated, malicious and targeted attacks**	ADSDAQ (Ad Network by ContextWeb) Continually refine predictive models for advertising response rates to deliver more precisely **targeted advertisements**

UBS Source: Cloudera and Hadoop World 2011.

Figure 1 provides additional examples of how leading organizations in other industries are currently leveraging BDA to analyze web-based traffic for sales and purchasing patterns, customer segmentation and risk or fraud management. For more information or examples of BDA in other industries, please refer to Appendix D.

[14] Figure used with the permission of UBS.

4 Economic Value and Opportunities for BDA in Digital Health

Research[15] on BDA's impact on the health care sector predicts a new wave of productivity, efficiencies and innovation in the areas of personalized care, health operations, clinical and health policy decision making, and improved patient engagement. It is widely believed by the research community that the use of BDA can reduce the cost of health care while improving its quality, by making health care more preventive and personalized, and basing it on more extensive continuous monitoring.

Leading industry research points out that BDA has potential to add value in all health care settings. Some specific examples include:

- Research published in the Journal of IHIMA[16] indicates that BDA solutions can help stakeholders personalize care, engage patients, reduce variability and costs, and improve quality of health delivery. The research points to BDA providing a rich context to shape many areas of health care, especially genomics where massive amounts of data are required and costs are rapidly decreasing.
- Research by Frost and Sullivan[17] depicts a new era where stakeholders are facing a flood of digital data as never before experienced. BDA is being hailed as the key to improving health outcomes and reducing health care costs, by managing and harnessing the analytical power of these large datasets.
- Research by Harvard School of Public Health (HSPH)[18] points to more scientific data being generated in the last five years than in the entire history of mankind. HSPH is sequencing and analyzing human genomes to ferret out clues to infections, cancer and non-communicable diseases. In just two weeks, Dr. Winston Hide, associate professor of bioinformatics at HSPH, joined a cancer database with a stem cell dataset – and got a big payoff. "We discovered a single gene that we think is responsible for the initiation of a whole class of leukemias," Hide said.

  Current research points to many potential public health uses of BD that extend beyond genomics. Environmental scientists are capturing huge quantities of air quality data from polluted areas and attempting to match it with equally bulky health care datasets for insights into respiratory disease. Epidemiologists are gathering information on social and sexual networks to better pinpoint the spread of disease and even create early warning systems. Comparative-effectiveness researchers are combing government and clinical databases for proof of the best, most cost-effective treatments for hundreds of conditions – information that could transform health care policy.

[15] Research by industry analysts such as IDC, Frost and Sullivan, Forrester, MIT Sloan, Harvard, and IT vendors IBM, ORACLE, SAS and SAP.

[16] Lisa Khorey, Vice-President of Enterprise Systems and Data Management, Information Technology at the University of Pittsburgh Medical Center. "Big Data, Bigger Outcomes." *Journal of AHIMA* 83, no.10 (October 2012): 38-43.

[17] Frost and Sullivan, Drowning in Big Data? Reducing Information Technology Complexities and Costs For Healthcare Organizations.

[18] Research published by Harvard School of Public Health, Spring and Summer series of 2012.

> Disease researchers have access to human genetic data and genomic databases of millions of bacteria – data they can combine to study treatment outcomes. These innovative methods for mining BD are transforming the way health care is being delivered.

In a study conducted by MGI in November 2011: *Big Data – The Next Frontier for Innovation, Competition, and Productivity*[19], high expectations are set for how BDA can potentially transform the global economy, improve organizational performance and work for better national and international policies.

MGI's research points to BDA potentially generating significant value in many areas by transforming processes, altering corporate ecosystems and facilitating innovation. Their analysis suggests BDA can be a key basis for competition, enabling new waves of productivity growth, innovation and consumer value across the world, through new insights that lead to new product ideas or help identify ways to improve operational efficiencies. The research is supported by evidence of benefits from existing BDA uses by web giants such as Google, Facebook and LinkedIn and from the more traditional financial, entertainment and retail industries. MGI predicts BDA will create new value through enabling experimentation, better segmenting and understanding populations, augmenting human decision making with automated algorithms and insights and supporting innovative new business models, products and services.

MGI studied BDA across the following domains:

- health care in the United States;
- the public sector in Europe;
- retail in the United States;
- manufacturing; and
- personal-location data globally.

MGI points out that at least $600-850 billion of the health care spending in the U.S. goes to embedded inefficiencies that increase the cost and decrease the overall quality of the public's health. MGI's research points to the health care sector as a potential gold mine where valuable insights are buried in unstructured or highly varied data sources that can now be leveraged through BDA. More specifically, their research predicts that creative and effective use of BDA in the U.S. health care system would drive efficiencies and quality, and create more than $300 billion in value annually. This represents approximately eight per cent estimated health care spending at 2010 levels, two-thirds of which would be through the reduction of national health care expenditures.

> *MGI's research states that the application of BDA can potentially remove $200-$300 billion in cost inefficiencies from the U.S. health care system.*

[19] See report at mckinsey.com/mgi.

The MGI research breaks down these long term value estimates across several broad categories:

- **Clinical operations** – This encompasses transparency of medical data, clinical decision support (CDS), remote patient monitoring, outcomes based research and comparative effectiveness research. By leveraging BDA, MGI estimates a potential reduction of U.S. national health care expenditures by up to $165 billion (US) a year. This represents a significant portion of the $300 billion. Examples from the MGI report include:
    - outcomes-based research determines which treatments will work best for specific patients ("optimal treatment pathways") by analyzing comprehensive patient and outcome data to compare the effectiveness of various interventions;
    - BDA for pre-diagnosis that automatically mines medical literature to create a medical expertise database capable of suggesting treatment options to physicians based on patients' medical records; and
    - remote patient monitoring for chronically ill patients and analysis of the resulting data to monitor treatment adherence and reduce patient in-hospital bed days, cut emergency department visits, improve the targeting of nursing home care and outpatient physician appointments, and reduce long-term health complications.
- **Remote, homecare, telehealth and smart devices** – Real time collection of data from remote patient monitoring devices, as well as numerous applications in use by telehealth organizations or home care service providers, creates a lot of new valuable and relevant data. This data can be analyzed to monitor adherence, improve drug and treatment options and prevent hospitalization or re-admission.
- **Research and development** – BDA can improve R&D productivity in the pharmaceutical industry by providing new insights for designing clinical trials and evolving personalized care by using genomics. These represent more than $100 billion (US) in value, about $25 billion (US) in the form of lower U.S. national health care expenditures.
- **Public health** – BDA could ensure the rapid, coordinated detection of infectious diseases and a comprehensive, integrated disease outbreak surveillance and response program. Benefits include a smaller number of claims and payouts due to a lower incidence of infection resulting from a timely public health response.
- **New business models** – BDA offers new business models that can aggregate and analyze data and patient records to provide data analytics services to third parties. Innovative companies could build robust datasets that would enable a number of related businesses. These might include licensing and analyzing clinical outcomes data for payors and regulators to improve clinical decision making.

Two examples of BDA specifically mentioned in MGI's report include:

- Kaiser Permanente, USA: its EHR provided the crucial dataset that led to the discovery of Vioxx's adverse drug effects and the subsequent withdrawal of the drug from the market.
- National Institute for Health and Clinical Excellence, UK: pioneered the use of large clinical datasets to investigate the clinical and cost effectiveness of new drugs and expensive treatments. The result of the analysis was that the agency issued appropriate guidelines on such costs for the National Health Service and negotiated prices and market-access conditions with pharmaceutical vendors.

4.1 International BDA Investments in Digital Health

In March 2012, the U.S. announced an initiative designed to harness massive amounts of information, including health care data, for research purposes.

The U.S. initiative aims to uncover innovative ways to use BDA for scientific discovery, biomedical research and other health care purposes. Aiming to make the most of the fast-growing volume of digital data, government agencies and departments have committed more than $200 million (US) to the effort including the Defense Advanced Research Projects Agency, Department of Defense, Department of Energy, National Institutes of Health (NIH), National Science Foundation (NSF) and U.S. Geological Survey.

The U.S. government has invested $200 million to make the most of the fast-growing volume of BDA. A significant portion of this investment is going toward health care.

The health care system in general stands to reap big rewards from this investment. As part of the initiative, NIH and NSF plan to collaborate on a project to find new technologies and methods for data analysis, data management and machine learning and natural language processing. The first wave of agency commitments to support the above initiatives includes:

- **National Science Foundation and the National Institutes of Health** – Core Techniques and Technologies for Advancing Big Data Science & Engineering. NIH is interested in imaging, molecular, cellular, electrophysiological, chemical, behavioural, epidemiological, clinical and other datasets related to health and disease.
- **National Institutes of Health Genomes Project** – Data is so massive that few researchers have the computing power to make the best use of it. Amazon Web Services (AWS) is storing the 1,000 Genomes Project as a publicly available dataset for free, and researchers will pay only for the computing services they use.

4.2 Opportunities for BDA in Digital Health in Canada

BDA is depicted as predictive and prescriptive analytics toolkits which can potentially reduce inefficiencies in the health care system and generate significant return on investment, much of which is still to be proven and documented.

Infoway believes that there is significant potential in Canada to justify use of BDA-based solutions.

Industry analysts[20] believe that if the health care system is to overcome current challenges, it must adopt new and innovative methods to customize and personalize care, improve quality of care and reduce inflationary trends by making it possible to measure and deliver true value to the health system and patients. BDA is seen by many as an essential tool in bringing these new opportunities to the health care system.

Infoway's analysis points to several types of BDA solutions emerging in the health care industry:

- **Clinical decision support (CDS) solutions** – BDA includes self-learning, question and answer, and predictive analytical solutions that can be targeted to specific business problems. These solutions can sift through large amounts of data, understand, categorize and learn from it, and then predict outcomes or recommend alternative treatments to clinicians and patients at the point of care.
- **Personalized care solutions** – Predictive data mining or analytic solutions that can leverage personalized care (also referred to as personalized medicine, for example genomic DNA sequence and biomarkers[21] for cancer care,) in real time to highlight best practice treatments for patients. These solutions can potentially offer early detection and diagnosis before a patient develops disease symptoms. More effective therapies can be employed because patients with the same diagnosis can be segmented according to molecular signature matching. Drug dosages could be adjusted to minimize side effects and maximize response.

[20] Industry sources examples include, but are not limited to – Gartner: Gartner hype cycle on Gartner's web site. http://www.gartner.com/technology/research/hype-cycles; _IDC: Global Overview Big Data technologies and services Forecast, website http://www.idc.com/getdoc.jsp?containerId=IDC_P23177; PricewaterhouseCoopers: PWC Needles in a Haystack; Seeking Knowledge with Clinical Informatics, Health Research Institute, February 2012; Acccenture: Top 10 Healthcare Game Changers: Canada's Emerging Health Innovations and Trends; UBS: Big Data Report, UBS Investment Research, Q-Series®: Next Biggest Driver in Tech? Big Data From A to Zettabyte; and IBM: Analytics: The New Path to Value, MIT Sloan Management Review, IBM Global Business Services, Business Analytics and Optimization.
For more complete list please refer to Bibliography.

[21] For medicine, a genomic biomarker is a measurable characteristic that reflects the severity or presence of some disease state. More generally a biomarker is anything that can be used as an indicator of a particular disease state or some other physiological state of an organism.

- **Public health and population solutions** – BDA solutions that can mine web-based and social media data to better understand what is happening. For example, predicting flu outbreaks based on consumers' search, social content and query activity. BDA solutions can also support clinicians and researchers or epidemiologists in performing analyses across patient populations and care venues to help identify disease trends.
- **Clinical operations solutions** – BDA solutions that could support clinical and administrative operations across programs such as wait-time management or chronic disease management, homecare and long term care, by:
  - mining large amounts of historical and unstructured data, looking for patterns and modelling various scenarios to predict events that may impact programs before they actually happen;
  - facilitating early detection, diagnosis and prediction capabilities to provide better insights on inpatient care for home and long-term care; and
  - supporting home care programs such as "aging in place" with new insights which allow patients to stay in their homes longer and reduce the use of more expensive institutionalized care.
- **Policy, financial and administrative solutions** – BDA solutions that would support decision makers by integrating and analyzing previously stove-piped networks of policy and research data to provide new insights specific to key health policy and program decisions and key performance indicators.

While the above opportunities depict BDA solutions that imply use by only administrators, decision makers, clinicians and researchers, they can also be extended to include perspectives of individuals who as consumers of health services would have access to these insights to help them make better health decisions for themselves.

4.3 Implementation Progress of BDA within the Health Care Sector

Infoway's research and analysis of current implementation and adoption trends for BDA are presented in this section. Examples of BDA in health care are also provided in this section.

Infoway's research indicates BDA as:

- growing expectations by industry analysts for applicability in health care;
- an emerging, relatively immature technology within health care, where business value is still being proven and sustainable business models are still unclear;
- a new type of analytics, experimental in nature including a predictive approach to analytics, which is quite different than traditional analytics;
- limited implementations of commercial solution-based applications in health care; and
- mostly used on the periphery of health care (i.e., in research studies, genomics research in cancer) and is not yet viewed as a management tool in mainstream practice.

Nevertheless, over the course of time, *Infoway* believes that BDA can be exploited by stakeholders across Canada to create new insights and value to support a health care system focused on better patient and health system outcomes. While little evidence of true benefits exists within health care, studies of other industries[22] have concluded that organizations effectively using analytics in their decision-making processes derived five to six per cent better "output and productivity" than if they had not used analytics.

For other industries, BDA enables businesses to better understand their customers and markets, predict buying preferences and manage corporate risk. Why should health care be any different? BDA can potentially provide better views, insights and understanding of individual health consumers and larger patient population. Through BDA's analytics and predictive capabilities, health care can potentially gain new insights about a patient's health conditions, which can serve to identify risks and alert clinicians and patients to the onset of an adverse event.

Studies of other industries have concluded that data-driven organizations who effectively use analytics in their decision-making processes derived five to six per cent better "output and productivity" than if they had not used analytics. If we use this same five-six per cent average for BDA for health care in Canada, it would translate into $10 billion in cost savings annually (five per cent of $207 billion expenditures in 2012).

The following are examples of BDA implementations in health care, within and outside of Canada.

Example1: Disease outbreak

BDA used for monitoring of disease networking. An example is Google.org's use of BDA to study the timing and location of search engine queries to predict disease outbreaks. Research shows[23] that one-third of consumers currently use social networking for health care purposes (Facebook, YouTube, blogs, Google, Twitter). As demands for access to health information from social networking sites continue to proliferate, BDA can potentially support key prevention programs such as disease surveillance and outbreak management.

Google analyzes and pulls results from related searches and online discussions. By using flu-related search request data, Google is also able to track a disease's spread thus providing another tool for surveillance in addition to the methods employed by public health agencies.

This trend to provide more public health data for various kinds of data visualizations is setting in motion new capabilities in the U.S. The U.S. Department of Health and Human Services (HHS) now publishes the research and allows the data visualization community to work with it and bring it to life and show additional value.

[22] Analytics: The New Path to Value, MIT Sloan Management Review, IBM Global Business Services, Business Analytics and Optimization, *By Steve LaValle, Michael Hopkins, Eric Lesser, Rebecca Shockley and Nina Kruschwitz.*
[23] PricewaterhouseCooper's Health Research Institute (HRI) *Social media "likes" healthcare*
From marketing to social business *April 2012*.

Example 2: Question and answer – clinical decision support

Research[24] indicates that 79 per cent of provider organizations in the U.S. are turning to clinical informatics in an effort to prevent medical errors. 61 per cent expect analytics to improve population health, and 52 per cent indicate that analytics-driven preventive care will help rein in costs. BDA can potentially help improve existing workflow and outcomes from business processes such as appointment brokering, scheduling, e-referral and e-discharge.

BDA can also assist in providing insights around gaps in the continuum of care across settings and highlight best practices in care processes and clinical outcomes.

Question and answer solutions integrated with computerized provider order entry (CPOE), e-referral, e-discharge and other process based activities can be used to analyze and predict trends in health care.

BDA can mine volumes of medical literature and other unstructured data and integrate these results with the increasing volumes of discrete data captured in EHRs, EMRs and PHRs. BDA can combine content analysis, evidence-based data and through natural language processing technology can understand, learn and then predict future events. These analytics are then fed back to clinicians as considerations in their decision making.

CDS systems are becoming substantially more intelligent by including technologies that use image analysis and recognition in databases of medical images (x-ray, CT, MRI) for pre-diagnosis. CDS systems can automatically mine medical literature to create a medical expertise database capable of suggesting treatment options to physicians based on patients' medical records. IBM's Watson and use of natural language processing tools can accurately extract medical facts and quickly understand and learn from relationships buried in large volumes of data. Several health care delivery organizations in the U.S. (Memorial Sloan-Kettering Cancer Centre and Seton) are involved in research projects where Watson is being piloted for its ability to support questions and answers for diagnosis. Natural language processing technology can help accelerate and improve clinical decisions, reduce operational waste and enhance patient outcomes.

Patients or consumers also use BDA to get answers for their own conditions. Data could be presented back in a meaningful way and encourage patient participation in their health care plans and potentially reduce re-admissions or adverse outcomes.

[24] PricewaterhouseCooper's Needles In A Haystack: Seeking Knowledge With Clinical Informatics, 2012.

Example 3: Data streaming

BDA solutions are providing early or predictive insights for clinicians and patients about treatment compliance and adverse events. The example to the left shows the use of BDA at Toronto’s Hospital for Sick Children to synthesize the deluge of information that monitors capture from neonates (more than 1,000 recordings per second of physiological measures such as body temperature, heart rate, respiratory rate and blood pressure). The BDA solution provides insights which allow researchers to create algorithms to predict when a baby is at risk of infection. As personal health devices and mobile sensor applications continue to proliferate, more information will become available as a result that can employ BDA.

Researchers at Toronto's Hospital for Sick Children are developing a new technology that would track by-the-second data for premature babies, and predict and alert physicians that a life-threatening infection could develop before a child even shows signs of illness.

The solution lets doctors monitor subtle changes in the heart rate. The monitor's computations alerts doctors, who then make the decision to proceed with treatment or monitor more closely. These alerts can potentially help prevent premature babies from getting sicker, resulting in shorter hospital stays and reduced costs. This new technology is seen as an enabler, but doesn't replace the work and decision making of the physician.

Example 4: Genomics and personalized care

Personalized care is being presented as the next wave of transformation in the delivery of medical treatment to patients.

The world’s largest set of data on human genetic variation – produced by the international 1000 Genomes Project – is now freely available on the Amazon Web Services (AWS) cloud. At 200 terabytes – the equivalent of 16 million file cabinets filled with text, or more than 30,000 standard DVDs – the current 1000 Genomes Project dataset is a prime example of BD, where datasets become so massive that few researchers have the computing power to make the best use of them. AWS is storing the 1000 Genomes Project as a publicly available dataset for free, and researchers will pay only for the computing services they use.

Genome 1000 and biomarker investments in cancer care are enabling the ability to predict individual disease risk, detect disease early, and improve diagnostic classification to better inform individualized care treatments.

In the future, if someone were to get cancer, the tumour itself could be tested. BDA could be used to provide insights into:

- *what genes are turned on;*
- *what genes are turned off; and*
- *what caused this to happen.*

Most importantly, BDA could provide insights into what needs to be done to turn off those cancer genes and destroy the disease.

In January 2012, the Government of Canada announced that it is investing $67.5 million into a "personalized medicine" health care strategy that will factor in a patient's genetics and the specific character of his or her illness before customizing a treatment plan. Personalized care has the potential to lead to improved quality of life for patients and their families, as well as cost savings across the health care system. Experts say personalized care may eliminate painful, toxic treatments that may not work on certain patients due to their genetic makeup or the type of condition they have. The federal government hopes to see results in roughly three or four years.

Genomic sequencing and biomarker data can be analyzed, combined with clinical data and presented back to clinicians and patients to provide insights that potentially would encourage behavioural changes, reduce side effects and patient safety challenges related to prescribing medicine.

BDA represents a potential "gold mine" in fostering innovation and new research for improving clinical trials and health product outcomes. The mining of genome data for research purposes is already being done. As personalized care evolves, research capacity is expected to increase and generate better outcomes.

Specifically, personalized care is being applied to create customized cancer treatments and management programs that are designed to target aberrant molecular pathways in a subset of patients with a given cancer type (see example above). Tissue-derived molecular information can be combined with an individual's medical history, family history and data from imaging and other laboratory tests to develop more effective treatments. Drug therapies could be tailored at a dosage that is most appropriate for an individual patient, with the potential benefits of increasing the efficiency and safety of the medications and optimal combination of treatments.

Example 5: Consumer based social media

WebMD is a popular website that provides health information, supportive communities and in-depth reference material on health issues. BDA powers this website. It provides the ability to enter your symptoms and ask specific questions, and through the use of BDA returns possible conditions and possible recommendations. It also provides the ability to manage weight, exercise and other general health conditions.

Another example includes social networking websites such as patientslikeme, where network based data is mined to provide insights to users on conditions and treatments for consumers (patients) with similar characteristics.

This example involves use of social media and networks to inform consumers of conditions and self-management recommendations. The example depicted to the left is the integration and provisioning of population based information within social networking environments to better inform and engage patients in managing their health and to influence or change their habits for conformance to treatment options, plans and best practices. However, the use of social network search engine toolkits does pose some interesting legal or liability considerations. Vendors of these toolkits may not provide evidence-based recommendations and treatment options to patients. In fact, patients may be at risk when they self-diagnose based on these recommendations.

Example 6: Supporting health innovation through the use of open health data

The U.S. Open Health Data Initiative is a public-private effort that aims to help Americans understand health and health care performance in their communities. This work centres on catalyzing the advent of a network of community health data suppliers (HHS) and "data appliers" who utilize that data to create BDA applications that raise awareness of community health performance, increase pressure on decision makers to improve performance, and help facilitate and inform action to improve performance.

Working with a growing array of health care stakeholders, they will be seeking to identify the uses of this data that would do the most to raise awareness of health performance, and help motivate civic leaders and citizens to improve performance.

Potential examples of uses include:

- interactive health maps on the web that allow citizens to understand health performance in their area versus other areas, with tremendous ease and clarity;
- "dashboards" that enable mayors and other civic leaders to track and publicize local health performance and issues;
- social networking applications that allow health improvement leaders to connect with each other, compare performance, share best practices, and challenge each other;
- competitions regarding how communities can innovate to improve health performance;
- viral online games that help educate people about community health;
- utilization of community health data to help improve the usefulness of results delivered by web search engines when people do health-related searches and further raise awareness of community health performance; and
- integration of community health-related data into new venues, such as real estate websites, which could be highly effective disseminators of such information.

For more information refer to: http://www.hhs.gov/open/initiatives/hdi/about.html.

For more examples of BDA in health care, see Appendix E.

5 Considerations for BDA in Digital Health

Infoway expects that the BDA hype in health care will continue to grow and stakeholders will be challenged to determine the value they will get from BDA, and how they should incorporate it into their digital health and analytic strategies.

For the past 10 years, most stakeholders in Canada have been immersed in digitizing and operationalizing health information and their electronic health record system (EHRS) infostructures. As a new technology, few stakeholders have looked into how BDA applies to them, their stakeholders, or how it fits into their digital health strategy and existing analytic investments. Even fewer have the financial scale and skills to embark on a BDA program. Also, the lack of data sharing and integration among various institutions continues to remain a major barrier to concepts such as BDA.

For the few organizations that have experimented with BDA, this activity has remained on the periphery of health care (i.e., in research studies) and is not yet viewed as a management tool in mainstream practice.

In addition to the above, *Infoway*'s research indicates little to no enterprise based, production ready BDA solutions exist, and little proven or documented benefits or return on investment (ROI) studies exist for BDA within health care.

Like any new technology, there are challenges, risks and drawbacks. For BDA to be successful in health care, BDA needs to be accompanied by a range of enablers. The following section provides business, governance, people, process and technical considerations that stakeholders should be aware of in moving forward with BDA.

5.1 BDA Business Considerations

This section highlights the business considerations that may apply to various stakeholders, namely policy makers, administrators and funders.

Infoway foresees BDA as a new "experimental" approach to analytics, which will challenge traditional approaches and investments in health data analytics.

- **Business model** – What will the business and sustainability models look like for BDA? How can organizations measure the return on investment when there is insufficient business case evidence in health care at this time? What are the sustainable options available for stakeholders for BDA?
- **BDA maturity** – BDA represents potentially a risky venture for stakeholders. Decisions about partners, key business problems, use case definitions, new data collection processes, the right resources and skillsets, the variety and types of solutions, and scope of control all require significant consideration in minimizing these risks. Early success of BDA initiatives can be optimized by ensuring key decisions are carefully considered up-front.

- **Business case driven** - Establishing an initial set of sound use cases and a resulting business case for BDA will be foundational to advancing the BDA agenda within the organization. As stakeholders move past initial pilots, proof of concepts and experimentation, *Infoway* suggests stakeholder's strategic plans answer these key questions:
    - What does our BDA road map or end state look like?
    - What types of BD and technologies are relevant to our organization?
    - How or where do we leverage our existing investments in EHRS and DW/BI?
    - Does this work support a business case and are we able to measure the benefits and ROI of these investments?
    - What are the implications of BDA to my organization? On IT governance processes or legislation? On IT and business unit skillsets? On IT and business processes? (e.g., information management and data stewardship, data collection processes, privacy and security, and infrastructure perspectives).

  Stakeholders should carefully assess decisions and trade-offs for whether to build or buy, train in-house staff versus using external or third party service providers, and how existing and new partners fit into BDA initiatives. Investment decisions need to be based on solid use cases and business case and ROI, clear investment criteria, and an end game strategy for BDA should be in place.

5.2 BDA Governance and Legislation Considerations

BDA may impact existing governance and legislation within and outside the stakeholder's realm of control. This section highlights the organizational and governance considerations that may apply to various stakeholders, namely senior decision makers, administrators and funders.

- **Data Governance** – Governance and risk frameworks may need to be enhanced to accommodate BDA. As BDA moves from an experimental type of analytics to a strategic initiative within an organization, executive support will be necessary. From data quality and common data definitions, to ensuring proper context and understanding of the data, these governance best practices represent challenges in reaching consensus. This will require financial, in-kind and collaborative resource support.

  Personalized care, streaming data and leveraging unstructured data are truly daunting data driven tasks. Stakeholders with formal governance and information management (IM) programs led by an executive manager (e.g., CIO or CMIO) may be more ready to undertake BDA initiatives. These organizations may also be more open to sharing data outside an organization's boundaries and to performing more advanced informatics functions. They may also be open to participating in new delivery models which may include their external stakeholders or partners.

- **Legislation** – Enabling BDA may require a review of existing legislation, as legislation defines and governs collection, use and disclosure of data. As new sources of data are determined for BDA, stakeholders may have to assess what changes may be required in existing legislation to enable BDA within their organization.

5.3 BDA Operational Considerations

This section highlights some of the strategic and tactical operational perspectives that stakeholders may need to consider when undertaking BDA. Given the experimental nature of BDA, we have split this section into two areas – strategic and tactical.

5.3.1 Strategic Operational Considerations

Traditional IT and DW/BI vendors as well as new market entrants are all branding themselves as BDA vendors. These BDA solutions may be costly upfront and initially pose higher risks of failure.

Stakeholders may not have the necessary upfront investment funding or skilled data scientists for BDA, and should consider options such as cloud and Software-as-a-service (SaaS) models. Stakeholders should carefully assess which model is best against a set of criteria, for example: cross organizational (regional, jurisdictional, practice), price point, level of maturity, governance, privacy and security, and ROI.

In other industries, a majority of companies have adopted the cloud for BDA in some form. More than one-third (38 per cent) of corporate-suite executives report that their company solely uses cloud technology to store and manage big data, 27 per cent say their firm uses private cloud or off-premises server farms, and 11 per cent use a public cloud. (Source: Harris Interactive, Inc., Communications C-Suite Study, April 2012).

BDA will challenge existing practices for most stakeholders. Traditionally, analytics is driven by a specific set of known questions or hypotheses the organization is trying to answer or metrics it wants to track. They are either deployed as subject specific data marts for a purpose, as statistical based models and algorithms or they leverage specific components of an existing data warehouse.

Figure 2 - BDA Experimental Model

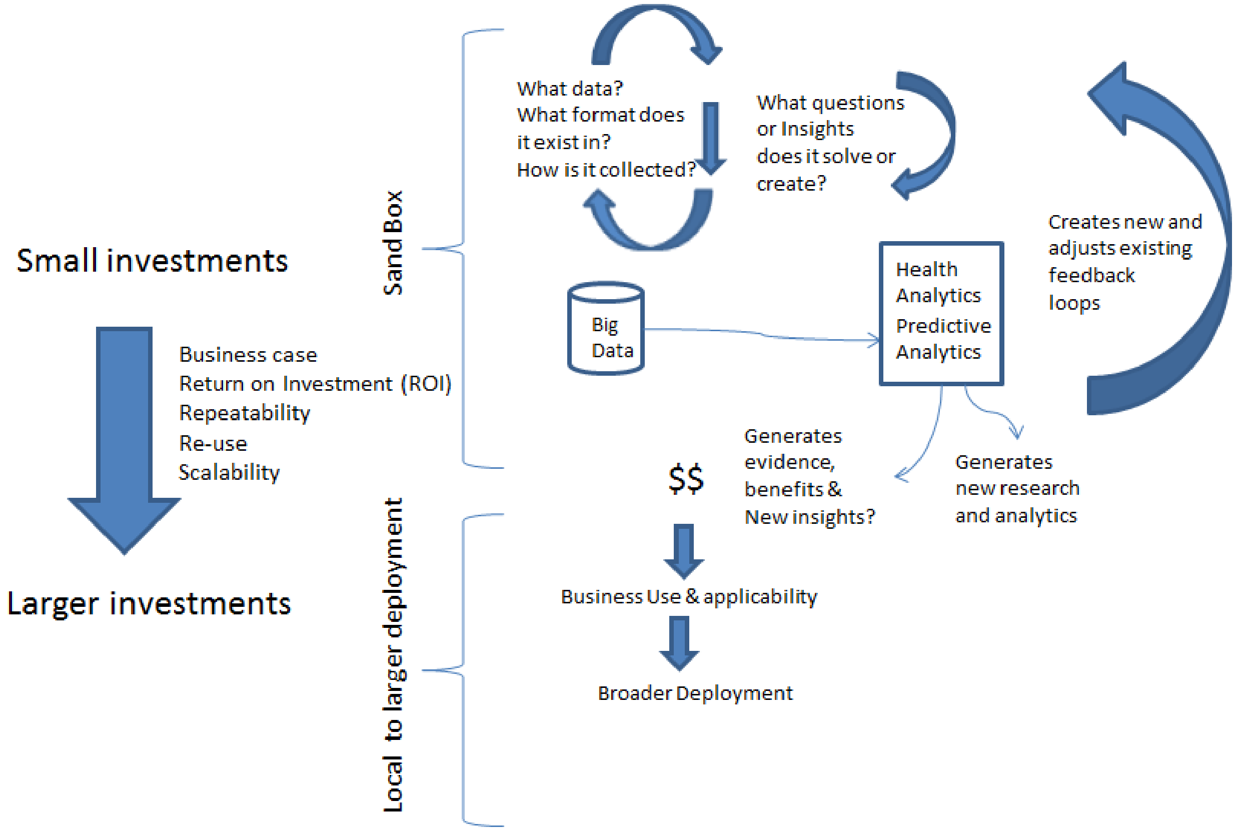

Small investments
Business case
Return on Investment (ROI)
Repeatability
Re-use
Scalability
Larger investments
Sand Box
Local to larger deployment
What data?
What format does it exist in?
How is it collected?
What questions or Insights does it solve or create?
Big Data
Health Analytics
Predictive Analytics
Creates new and adjusts existing feedback loops
$$
Generates evidence, benefits & New insights?
Generates new research and analytics
Business Use & applicability
Broader Deployment

Figure 2 depicts a collaborative data driven methodology that complements BDA. While the "sand box" depicted in the top of the diagram is not new in the DW/BI industry, it represents more than supporting a known requirement or set of individuals. The sand box for BDA represents an environment which provides a flexible, secure but open, and collaborative environment for using and sharing BDA technologies on BD, thereby minimizing the risk, investment and operational implications to the stakeholder.

This environment encourages cooperation and collaboration between researchers, clinicians, governors and the private sector through experimentation with BDA tools. The sand box is a place for business analysts, policy makers and researchers to collaboratively investigate, experiment and find insights about key program and policy challenges of the health care system. The sand box provides the raw materials for the development of new tools and resources.

The above operational model is intended to be a low investment, low risk, collaborative, but highly secure environment, based on agreements between stakeholders around the flow, privacy and security requirements of PHI. It encourages reduced risk and exposure for the organization by minimizing rollout across an organization until benefits have been proven, ensuring lessons learned have been leveraged, and the re-use and applicability of solutions (not necessarily the data) are identified and documented.

5.3.2 Tactical Operational Considerations

As mentioned earlier, BDA represents a new type of analytics that requires new and different operational processes, skillsets and considerations specific to BD. Figure 3 illustrates how BDA solutions may need to be structured, operated and supported in the future. It depicts a data-driven knowledge creation model, which takes into consideration new processes and skillsets for operating and integrating health data and information (knowledge) with other parts of the corporate and information architecture. It also attempts to depict how it would leverage traditional analytics assets.

Figure 3 – BDA – A Data Driven Knowledge Based Operational Example

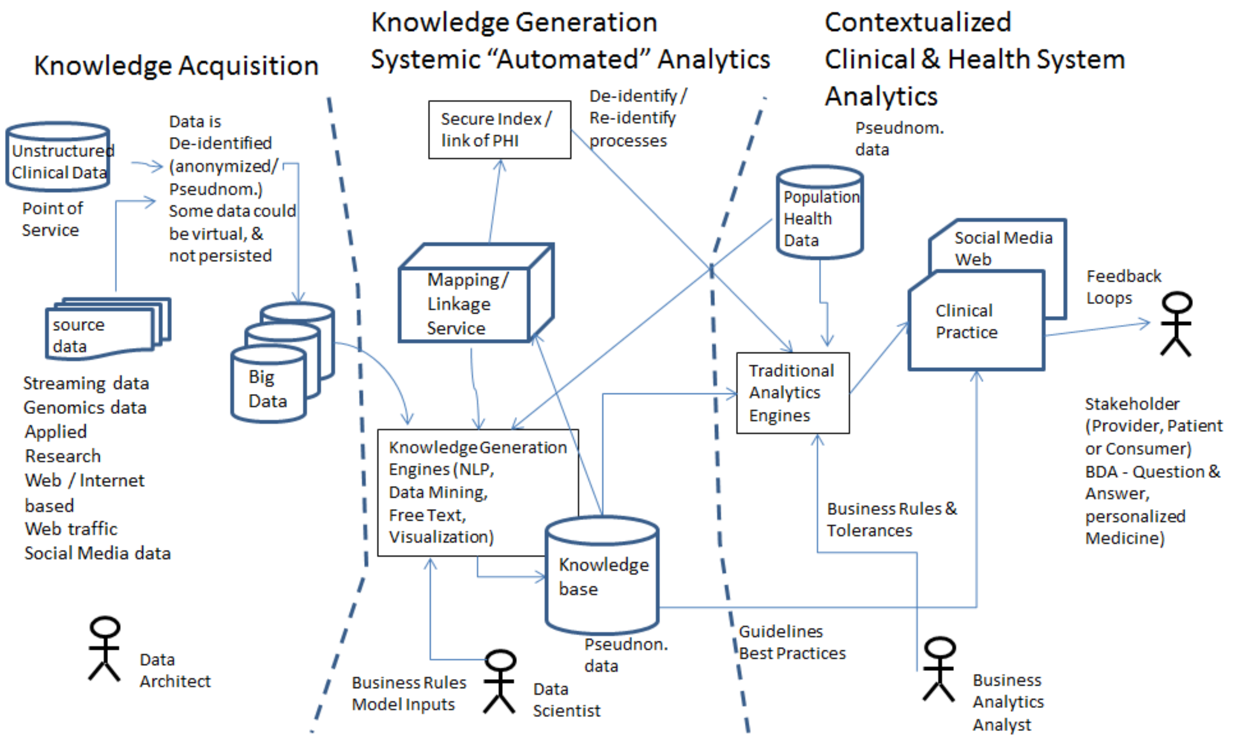


Figure 3 breaks BDA into three main areas of knowledge creation:

Knowledge Acquisition

On the left side of the diagram, new data collection flows and processes get the raw BD acquired, prepared and available for analytics. At this point, the data may only be streamed, not persisted, and processed when required by an "analytic or automated" business rule or service (e.g., the data is collected for its purpose, analyzed and results produced, but actual raw data may not be persisted).

The raw BD sources on the left come from a variety of sources:

- data streamed and sourced from a remote or smart device, monitor, sensor or telehomecare of telehealth device;

- genomics data which profiles specific gene patterns, biomarkers and their impacts and relationship to specific health products, diseases, and potentially profiles of what's worked or hasn't worked for specific gene sequences;
- web-based traffic or social networking data from web searches or network discussions for specific disease or health events and concerns;
- unstructured data from clinical, health domains or non-health sources; and
- clinical publications, applied research or clinical guidelines and reference data (e.g., drugs).

The data architect works with a variety of stakeholders to model, de-identify, integrate, categorize and catalogue metadata about the sources of BD. The data architect's role is much different than that of the data scientist or business analytics role. A data architect is knowledgeable on key health policy and program agendas, understands data across the organization, and is able to identify and prioritize these requirements against a specific business priorities and problems and integrate this data across the organization.

In addition to the above, the data architect can incorporate unstructured and traditional structured data into the BDA architecture to provide better informed insights of the particular business problem being tackled.

Knowledge Generation

At the centre are key knowledge generation processes, tools and applications that systemically go through the BD in near real-time to discover things that are important about these data. This requires development and maintenance of business rules, computer based algorithms, predictive models and applications that mine the data. As new insights are discovered and best practices are identified, these are moved into a central knowledge base.

The data scientist works with knowledge generation tools [data mining, natural language processing (NLP), etc.] and creates the business rules and models which look for patterns and relationships between data, creating new insights and knowledge. Skillsets include knowledge of data visualization, data mining and artificial intelligence toolkits. They also work with a variety of clinical and administrative experts (e.g., quality improvement leads) in maintaining a knowledge base and evolving knowledge from this base that may become guidelines, best practices, evidence and alerts or new business rules for other stakeholders and health information technology (HIT) applications to incorporate into their business processes and analytic environments.

Contextualized Clinical and Health System Analytics

On the far right side, BDA ultimately feeds into the health analytics and traditional DW/BI environments, as contextualized feedback loops back to all authorized stakeholders. In addition to the normal business of analytics, it leverages the knowledge base created by BDA, and integrates these learnings though feedback loops to various stakeholders (e.g., alerts, predictive insights and actual new data, such as best practices, from the BDA platform).

The business analytics analyst works with stakeholders to ensure that they are able to derive value and action from the analytics. The analyst is able to tweak the platform to support concepts such as self-service and adjust business rules and tolerances to create proactive and actionable insights.

5.4 BDA Process Considerations

BDA may impact existing information management and other processes within and outside the enterprise. This section highlights the process and organizational considerations that apply to various stakeholders, namely policy makers, administrators and funders.

- **Information management processes** – Processes that support the traditional DW/BI environment may need to be re-assessed as BDA begins to play a more important role within an organization. For example, it will be important to ensure that data validation, data sourcing business rules, processes and services exist to support concepts of BDA at the point of data capture.
- **Data custodianship** – Data may need to flow to other custodians from the original sources for analysis and research purposes. Who will own, operate and govern BD sources like personalized care, streaming sensor and device data, and social media data?
- **Data stewardship** – Significant challenges may emerge about ownership of the newly created data, the rights of the patient or creator and fair use of the data. New criteria for managing data from different data points in near real-time may be required. BDA may require dialogue on the following:
    - Where do the official versions of data sources need to be located?
    - What can be or needs to be distributed versus centralized?
    - Who owns or is custodian of the newly created information?
    - How are these data sources maintained?
    - Where, how and by whom are they standardized, integrated and aggregated?
    - Should this data remain inside or outside the organization?
    - How is external data integrated back into the health care system and organization?

- **Master data management**[25] – Stakeholders need to consider how existing or traditional information management best practice processes such as master data management (MDM) and identity management fit into BDA. BDA will create new views into previously complex assets that may not be supported by traditional data management best practices of the organization. BDA may also challenge traditional methods of linking data to a common health data model. Stakeholders should be aware that as they invest in BDA initiatives they may discover gaps in their existing data management practices (e.g., metadata, data model, reference data, relationships, tagging, filtering, indexing) generated through the use of BDA.
- **Metadata and data quality management** – BDA may introduce the need for data to move across an organization's business processes and outside of the enterprise. New standards and metadata may be required to ensure that data is understandable and of the highest quality possible. These new types of standards, semantics and metadata will represent necessary enablers for fully leveraging BDA across and outside of the enterprise for health care.

5.5 BDA People Considerations

This section highlights the people considerations that may apply to various stakeholders, namely clinicians and operators or IT organizations.

- **Clinical value, relevance and change management** – As BDA moves into clinical practice, capabilities such as computer-based predictive or prescriptive analytics, natural language processing (NLP) solutions and personalized care may result in changes in the way that clinicians are accustomed to working with health information. Through BDA, clinicians may need to adapt to new sources of potentially near real-time health information to support their decision making at the point of care. Concepts such as predictive analytics and personalized care will be of value to the clinicians only when business practices and processes are refined to take advantage of these capabilities. Measurable outcomes need to be determined to define the exact value of BDA to stakeholders.

  BDA is expected to come with significant change management and education with clinicians as well as the clinical informatics groups who manage the processes for health care organizations. There is also a need to better align clinical research (using these types of BDA) and its adoption and implementation within clinical practice. Predictive analytics and personalized care will introduce new concepts to clinicians that they'll need to be trained on how to interpret, act on and explain it to their patients.

[25] In IT, master data management (MDM) comprises a set of processes and tools that consistently defines and manages the master data (i.e. non-transactional data entities – client, provider, organization, etc.) of an organization (which may include reference data). MDM has the objective of providing processes for collecting, aggregating, matching, consolidating, quality-assuring, persisting and distributing such data throughout an organization to ensure consistency and control in the ongoing maintenance and application use of this information.

The lack of data science, statistics and programming skills needed by BDA will be a significant barrier to implementation.

- **Skilled resources** – There is already a global skills shortage in the data scientist and data analyst roles, especially when combined with health system and clinical knowledge. For stakeholders getting into BDA, there will be a gap in new roles and skillsets required to support these technologies. The role is referred to as a "data scientist."[26] *Infoway* expects there will be a significant gap for this skillset for most stakeholders and globally across the IT industry in general, which may become a prime barrier in moving forward with BDA. BDA requires the combined skillsets of programmers, data scientists, data miners and information analysts. Few staff will possess all of these competencies. On the IT services sector side, *Infoway* also expects that there will be a lack of highly skilled data scientists.

5.6 Privacy and Security, Ethics and Legal Considerations

As BDA progresses within the Canadian health care system, challenges, risks and concerns will arise about privacy, ethics and the legal implications of BDA. These risks and challenges represent significant barriers and potential drawbacks which may impact the success, pace and rate of adoption of BDA in Canada and the potential benefits it can bring. Stakeholders should consider the following:

- **Privacy and security** – Mining of unstructured clinical data or streamed clinical data potentially containing PHI represents a significant privacy risk and challenge. The emergence of BDA challenges our traditional definitions of collection, use and disclosure of personal health information. The emergence of BDA highlights the need for re-assessing existing policies around privacy and security. Agreement on common, clear and consistent terminology is required for appropriate uses of BD and the information (insights) created as a result. Best practice concepts such as Privacy by Design[27], should be considered to ensure privacy of the patient is maintained, while enabling the health system and patient to benefit from BDA.

  In the context of research, consideration should be given to how personal health information collected for the purpose of care and treatment can be used for research while respecting obligations to obtain an individual's consent for a different type of use. This requires jurisdictions to consider if and how BDA may require changes to current consent related legislation.

[26] A data scientist possesses a combination of analytic, machine learning, data mining and statistical skills as well as experience with algorithms and coding. Perhaps the most important skill a data scientist possesses, however, is the ability to explain the significance of data in a way that can be easily understood by others.

[27] For more information, refer to http://privacybydesign.ca/.

- **Ethical and medico-legal aspects** – Policies on the ethical use of these data types and sources may need to be established. Medico-legal implications for patient's and clinician's use or non-use of BDA should be assessed to understand if there are any new implications created by BDA.
    - **Ethical** – BDA represents a new capability that allows stakeholders to answer new questions or needs for knowledge that were not anticipated when the original data collection methods were put in place. While this represents a huge opportunity for health research, ethics policies should be reviewed to assess whether they allow and enable these experimental analytics.
    - **Medico-legal** – Legal and liability implications of leveraging BDA technologies such as genomics, natural language processing and artificial intelligence need to be considered. BDA solutions provide access to health knowledge bases for various stakeholders (e.g., patient-consumers) that can be searched instantly and displayed back as medical advice and treatment recommendations. On the one hand, BDA may engage stakeholders by providing instant access to health information regardless of location. On the other hand, it may fuel self-diagnosis, which may prevent consumers from seeking proper medical attention or result in misdiagnosis.

5.7 BDA Technical and Deployment Considerations

This section highlights the technical and deployment considerations for BDA that may apply to various stakeholders.

When architecting for BDA, stakeholders need to consider technical perspectives of performance, fault-tolerance and query integration and interface. They need to consider that the traditional analytical community may continue to communicate through structured query language (SQL) and messaging interfaces.

Considerations include:

5.7.1 Technical and Infrastructure Considerations

BDA introduces new requirements into EHRS and DW/BI environments. BDA is driven by volume, near real-time data analytics and analytic processes on different types of data. New investments or upgrades to infrastructure may be required to address requirements for availability, flexibility, portability, system integrity (security attributes of the system), performance, reliability, reusability, robustness (error and exception management), scalability and usability.

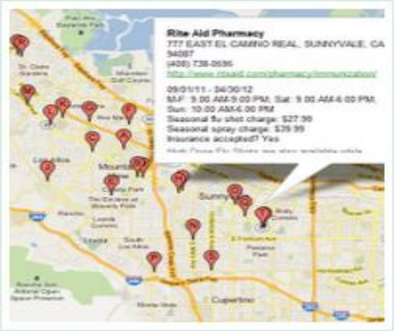
- **Batch Versus Near Real-time Latency Trade-offs** – Stakeholders should consider the trade-offs between batch and near real-time latency. Natural language processing and text analytic processes on BDA need to take into consideration latency expectations. Some BDA technologies are batch based tasks and may take anywhere from minutes to hours to process or sift through large amounts of BD. These capabilities need to be aligned with end user expectations, workflow and business requirements.
- **BDA and Traditional DW/BI Technologies** – BDA will challenge traditional data warehousing infostructure, skillsets and practices. While BDA does not replace them, and should be considered complementary, it will contribute new data or new findings to the existing DW/BI environment. For stakeholders that choose to transition into BDA, they should look to leverage and integrate, where possible, their BDA and existing DW/BI environments.
- **IT and Vendor Legacy Challenges** – BDA increases the complexity of health data collection and use of this information for analytics by legacy HIT/PoS systems. The new BDA platform has to function seamlessly and adapt to multiple user needs in the health sector and partner sectors. Traditionally, the transition has been slower in health care around the adoption of enterprise architecture, service oriented architecture and standards. These same challenges will emerge as BDA integrates with the legacy platforms to incorporate analytics back into clinical and administrative processes and workflows. Stakeholders should look to work with private industry to ensure that the HIT industry focuses on building a standardized set of BDA capabilities that are accessible by any vendor product. The less desirable alternative is for all stakeholders to build their own proprietary BDA capabilities within every HIT/PoS, mobile or DW/BI application implementation.
- **BDA and Open Source** – A significant amount of BDA is being driven by open source toolkits and frameworks. Stakeholders should understand that open source doesn't necessarily mean free. While frameworks such as Hadoop, MapReduce, R (open source data mining) and NLP are all open source, the process of developing and deploying it is far from free. Hadoop's, R's data mining and NLP's concepts and complexity makes it necessary to acquire or build, and maintain a significant infrastructure and skillsets.[28]

  Stakeholders should understand some key questions around what they are trying to accomplish with BDA, and once they understand this, they can better understand the technologies, resources and data they require to support BDA.
    - What components of the toolkits are open source versus established products such as text analytics, natural language processing and data mining?
    - How aligned should my solutions be with these open source frameworks?
    - What internal (or external) expertise must be acquired to support open source and complementary BDA toolkits?

[28] For more information on Hadoop, MapReduce and R, please refer to Appendix C.

- **Data Visualization** – Visualization will be a key enabler for obtaining value for stakeholders from BD. Visualization must systemically make sense of the data, dynamically and visually alert or notify, and present this data to stakeholders thereby generating value by making sure that the results of analytics are contextualized, relevant and actionable. Merging BDA and new modalities of graphic visualization will create a powerful tool for predicting trends at a glance for individual patients and populations.

  Some references or links to visualization work in health care include:
  - http://blog.visual.ly/health-data-visualization-contest-winners/
  - http://infosthetics.com/
  - http://visual.ly/



Rite Aid Pharmacy
777 EAST EL CAMINO REAL, SUNNYVALE, CA 94087
(408) 738-0696
http://www.riteaid.com/pharmacy/immunization/

09/01/11 - 04/30/12
M-F: 9:00 AM-9:00 PM, Sat: 9:00 AM-6:00 PM,
Sun: 10:00 AM-6:00 PM
Seasonal flu shot charge: $27.99
Seasonal spray charge: $29.99
Insurance accepted? Yes

Up-to-date vaccine data
Local providers maintain the Flu Vaccine Finder with accurate and up-to-date data for locations offering flu shots. **See site**.

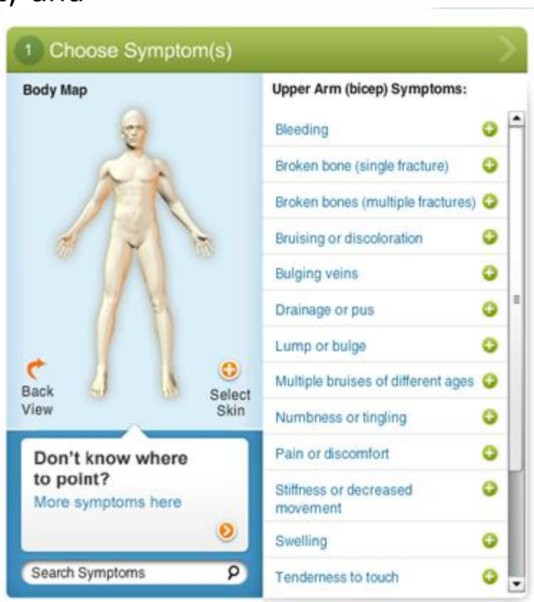
New sophisticated graphical and visual representations are necessary to enable rapid and accurate observations and actions from the data. Visualization helps one spot critical trends and patterns.

By visualization, we mean:
- display of data using techniques such as 3D bar and pie charts, histograms, data constellations, multi-scapes, and more;
- dynamic views on the fly, by selecting, zooming, pivoting or re-colouring objects, viewing information from different perspectives and uncovering new relationships in data;
- automated conditional styling and stop-lighting that leverage predictive data and text mining;
- visual dashboards that are integrated into business processes and point-of-care systems as live (near real-time) integration of PDF files, live PowerPoint presentations and other visualization objects; and
- context-preserved alerts, notifications and drill downs to more detailed data that highlight underlying causes behind problems or anomalies.

Examples of techniques leveraged include:
- Visual animation – watching data change over time gives a velocity and direction to the values it represents;
- Immersive visualization – such as a virtual-reality video game, places the user in a virtual environment containing three (or multi) dimensional representations of complex data;
- Interactive visualization – interaction with the data to provide a full information experience. Image can be brushed over with a mouse to reveal context-specific information about a data point, including the data value and various dimensional attributes such as geographic location, demographic group and organizational entity;

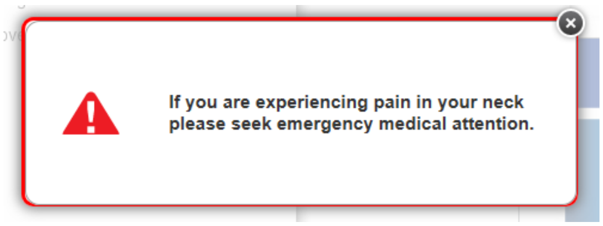
- ○ Textual animation – key perspectives and relationships of the data are depicted through the display of text; and
- ○ Geospatial visualization - increasing the strength of dashboards and maps as analytical tools through location based data.

A visualization example is depicted on the right with webMD. Visualization techniques allow a consumer to navigate through an image of a person and confirm symptoms and ask questions in an interactive manner to obtain conditions and recommendations for the symptoms.


1 Choose Symptom(s)
Body Map
Back View
Select Skin
Don't know where to point?
More symptoms here
Search Symptoms
Upper Arm (bicep) Symptoms:
Bleeding
Broken bone (single fracture)
Broken bones (multiple fractures)
Bruising or discoloration
Bulging veins
Drainage or pus
Lump or bulge
Multiple bruises of different ages
Numbness or tingling
Pain or discomfort
Stiffness or decreased movement
Swelling
Tenderness to touch

The consumer can pick a specific set of symptoms and ask a set of questions to the website. Through visualization the website will respond with a set of conditions and recommendations.

The website has specific logic that, depending on the symptom, will recommend immediate actions to seek medical attention.



If you are experiencing pain in your neck please seek emergency medical attention.

5.7.2 Technical Deployment Considerations

There are several deployment models and options available for stakeholders to consider. In this paper, we discuss two illustrative examples of deployment options:

1) Leveraging the EHRS and Health Information Access Layer (HIAL) for BDA, and

2) Integrating BDA directly into a HIT or Point of Service (PoS) system.

Each of these options has pros and cons, and each should be considered based on its merits against a common set of criteria. Some example criteria include:

- In-house versus cloud and software-as-a service model considerations;
- Centralized versus distributed (or federated) governance considerations;
- Realm of control:
  - ○ What needs to remain within the organization versus what can be outside of the organization?
  - ○ What is self-contained versus a shared resource or part of a larger community or ecosystem?

  - Is the need local versus pan-Canadian, multi-jurisdictional or across a discipline or specialty?
- Level of risk to PHI and privacy and security considerations for components of the solution within and outside of the organization;
- Role of BDA as a common analytic application that others can consume as a service across the enterprise; and
- Business and sustainability model and how BDA would operate or pay for itself (ROI).

1) Leveraging the EHRS and HIAL for BDA

Stakeholders will require an extensible and agile services framework which can deploy the appropriate new processing power and services required to support BDA. These can span more than one variety of data asset type (e.g., RDBMS, but also columnar, streams and text). BDA will expect a certain level of maturity within the organization in terms of services frameworks, governance and information management best practices. Existing investments in EHRS and their ability to scale and support BDA will need to be assessed.

Figure 4 - BDA Illustrative Example 1: Leveraging EHRS

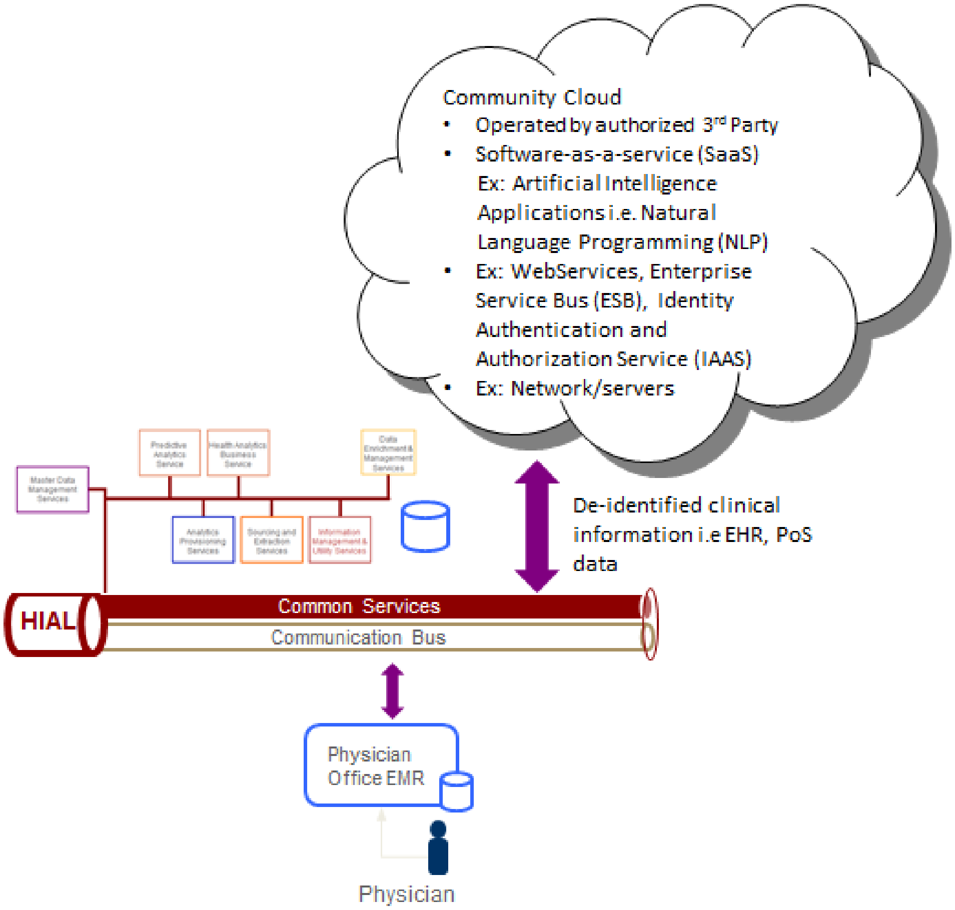

Community Cloud
• Operated by authorized 3rd Party
• Software-as-a-service (SaaS)
Ex: Artificial Intelligence Applications i.e. Natural Language Programming (NLP)
• Ex: WebServices, Enterprise Service Bus (ESB), Identity Authentication and Authorization Service (IAAS)
• Ex: Network/servers
De-identified clinical information i.e EHR, PoS data
Master Data Management Services
Predictive Analytics Service
Health Analytics Business Service
Data Enrichment & Management Services
Analytics Provisioning Services
Sourcing and Extraction Services
Information Management & Utility Services
HIAL
Common Services
Communication Bus
Physician Office EMR
Physician

Figure 4 depicts a cloud-based solution integrated through the EHRS. The BDA service is available via a private or community-based cloud and provisions all the hardware, software services and processes. This service could be provided by a third party, private sector partner or internal IT staff.

This deployment model leverages the EHRS from several points, including:

- Obtaining information from the repositories "above the HIAL" which is made available to the BDA service(s).
- Handling management of subscription and publication to services of the BDA service by stakeholders.
- Handling all software service calls in the HIAL such as authentication, authorization and consent services.
- Presenting the analytics results back to the requesting system as a visualization and presentation service.

Services description

- BD is stored on commodity hardware and supported by dedicated staff.
- Knowledge creation applications (data mining, NLP, predictive modelling) are stored on separate hardware and supported by dedicated staff.
- Knowledge generation services are accessed by stakeholders (e.g., clinicians, consumers) via their HIT (EMR, CIS, PHR and PoS applications).
- HIT/PoS can subscribe and leverage analytics based clinical and non-clinical data stores.
- HIT/PoS can subscribe and access authorized analytic data stores.

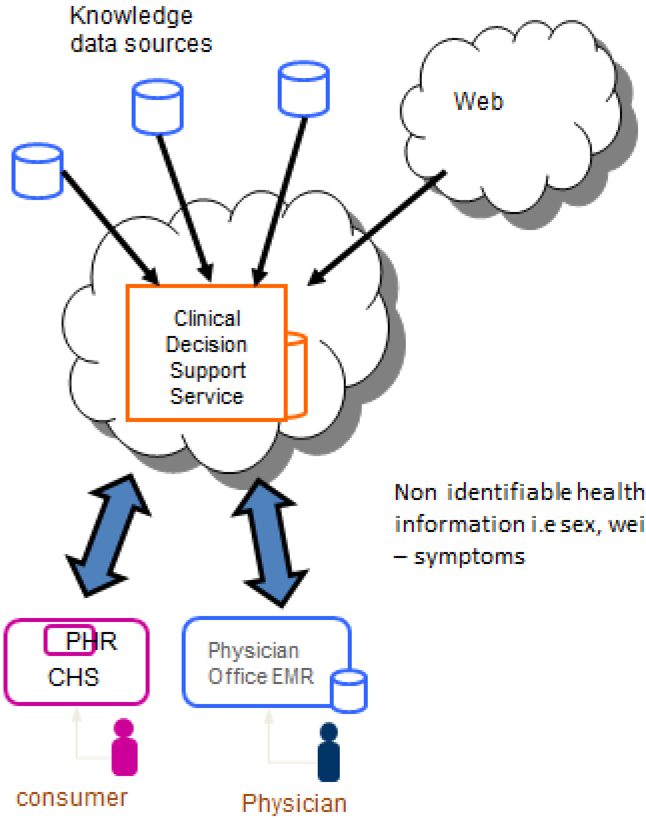
Key assumptions

- Governance, policies and best practices exist for de-identifying and sharing BD.
- Only de-identified PHI are sent to service, shared or stored on a community/private cloud.
- Privacy and security services on cloud manage access controls and identity management services for the BDA components.
- Cloud and SaaS models could be in any jurisdiction, health service delivery organization, or an authorized third party.
- Natural language processing service is called by the HIT/PoS system to process free text clinical notes and reports which have value to the analytics engine.
- In-house staff would require training on BDA architecture and knowledge creation concepts.

2) Integrating BDA Directly by a HIT or PoS System

The model in Figure 5 illustrates a BDA solution such as a "question and answer service," available to various stakeholders (e.g., consumers-patients and clinicians), but which isn't enabled or tied to an EHRS. Note that the key assumption for this service is no identifiable data is passed to and from the service.

Figure 5 – BDA Illustrative Example 2: BDA Service Accessed Directly by Point of Service



Assumption: Theses types of Solutions are standalone & don't require integration to EHRS (e.g. for authentication, or other services)

Non identifiable health information i.e sex, weight – symptoms

Figure 5 is based on a service offered in a private, community or public cloud and as SaaS. Hardware and software processes are managed by the cloud provider. Only non-identifiable information can be requested by the end user question (e.g., condition, sex, weight or other symptoms) and the response is provided to the end user.

Services description

- The business model is subscription or consumption based.
- HIT or PoS use of the service and analytics back to the requesting system are within a specific context (e.g., patient, consumer or provider).
- Extract, transform, load (ETL) and analytic functions and processes for the knowledge base are self-contained within a cloud operated by a commercial or third party.
- A central knowledge base used by the Q&A query logic for ease of maintenance.

Key assumptions

- No identifiable PHI is sent to service.
- All services are run on massively parallel hardware solution(s).
- Cloud service model can be operated centrally for a large user base resulting in some economies of scale.

6 Next Steps for BDA in Digital Health

The health care sector in Canada is undergoing significant transformation through digitization and IT investments in EHRs and EMRs. Digitization is creating many new sources of information for stakeholders across the health system that was not previously available.

6.1 BDA: Call to Action in Health Care

In the past the health care industry has been characterized as being data rich and information poor. BDA presents new opportunities to combine domain expertise with data analytics expertise to provide new insights to help decision makers transform health care delivery.

Like any new technology, BDA is being introduced to the health sector with much promise. However, its capabilities, opportunities and benefits still need to be proven through application in clinical and administrative practice. In light of the research and the challenges and risks depicted in this paper, *Infoway* strongly advises that for stakeholders who consider investing in BDA, that they do so through scope-limited pilot projects. Iterative, step-wise project investments will limit capital requirements and permit organizations to study the feasibility and utility of the new capabilities before committing additional resources. These projects should focus on solving specific and pressing clinical or business problems which are supported by clear use cases, while allowing for some experimentation. Wherever possible, existing infrastructure assets, information sets and partnerships should be leveraged.

BDA presents significant challenges for all stakeholders:

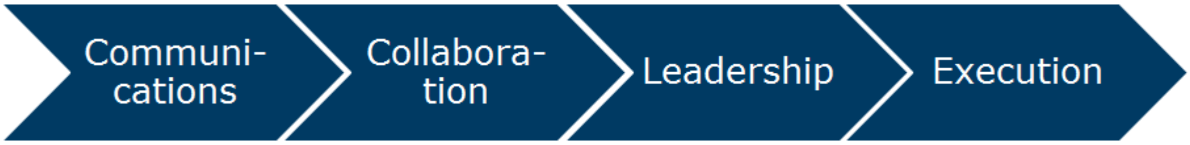
- *assessing and prioritizing where it fits into their overall digital health strategies;*
- *assessing and determining what the value and ROI of this technology is related to the hype;*
- *determining how, where and when to start; and*
- *determining what the impacts of BDA will be on the organization in terms of processes, capabilities and skills.*

Stakeholders who are ready to look seriously at BDA:

- *are the larger health care corporations who are likely able to afford the necessary investments in these technologies;*
- *have reached a certain level of maturity with their EHRS, DW/BI and information governance capabilities; or*
- *have demonstrated some success with knowledge management initiatives and are ready to take the next step to BDA.*

For others, it should be a cautious and collaborative approach.

For BDA to move from the periphery of research to more mainstream health service delivery, several inter-dependent steps are suggested be undertaken.



BDA Communications

Stakeholders should better understand BDA, the potential changes and new requirements that it will bring, by:

- performing initial education and research to understand what BDA is and how it will impact stakeholders;
- developing a strategy for how it will unfold across their organization and align with their digital health strategy;
- identifying champions and representatives from the clinical, administrative, business, privacy, data governance and technical roles, as well as vendors and patient advocates; and
- investing in staff learning and development.

BDA Collaboration

Stakeholders should not go it alone or start from scratch. They should look to leverage work done elsewhere and work collaboratively to share experiences, lessons learned and success stories for BDA. Some potential suggestions to consider include:

- form partnerships with trusted public and/or private sector organizations that have common interests, problems, experiences, BDA solutions and services;
- join or form communities of interest especially with members of their eco-system;
- collaborate with other stakeholders (e.g., jurisdictions) to create best practices, methodology, risk management, and the value proposition and business use cases for BDA;
- look to potentially leverage similar work around knowledge management by Canadian universities and the research community. The research community has accomplished some practical work in the area of knowledge management in health care and *Infoway* expects they will continue to lead and evolve the ROI for knowledge management in health care as part of larger initiative under BDA; and
- understand how to leverage and invest in the research agenda of BDA in Canada, especially as it applies to health.

BDA Leadership

BDA is a new type of analytics and requires stakeholders to assess some key issues that will in the future enable successful implementations and create value. Potential suggestions to consider include:

- look for opportunities where multiple innovative technologies may have common business problems and issues and can be combined to deliver additional value to their stakeholders (e.g., mobile or remote patient monitoring that deploys BDA, cloud computing technologies that also have BDA, or other emerging trends or technologies such as social media and visualization);
- enable experimentation within the organization to discover new insights and to improve performance. BDA represents a new culture of innovation and experimentation. Increasingly, stakeholders may need to acquire access to third-party data sources and integrate this external information with their own, to capture the full potential of BD;
- tweak or adapt existing data governance, information management processes, and privacy and security approaches to align with initial BDA deployment models; and
- work closely and invest with research and the private sector on focused initiatives where clinical and stakeholder value can be demonstrated through the use of predictive analytics and visualization techniques in such areas as genomics, personalized care and social media.

BDA Execution

In *Infoway's* opinion, BDA needs to be integrated into stakeholders' workflow and decision making processes. One key perspective of BDA, "value," is still an unknown for health care. BDA solutions need to contextualize results from BD for stakeholders to create value.

For stakeholders that decide to invest, they should do so cautiously and selectively target investments through limited scope pilot(s) focused on a key use case or business problem to demonstrate and document BDA benefits. Other suggestions include:

- assessing lessons learned from initial BDA pilots to understand how to approach future investments;
- experimenting with analytics to determine resource and skill requirements, as well as changes to data collection and other information management processes; and
- focusing initial efforts on developing (or adapting other stakeholder's) use cases, business cases, benefits evaluation framework, strategies and road maps.

6.2 BDA: Next Steps for the Near and Short Term

Infoway believes that BDA will continue to evolve and has the potential to become a prominent component of most stakeholders' information management plans and activities.

In the near to medium term, stakeholders should look to assess the following key challenges, risks or barriers:

BDA, EHRS and Analytic Investments

Stakeholders should look to understand how best they can leverage EHRS and health analytics assets for incorporating BDA. Stakeholders need to assess:

- how EHR investments can be extended to incorporate components of BDA;
- how traditional DW/BI tools can be leveraged and how the two will co-exist;
- how open source can be incorporated into their EHR infostructure to enable BDA; and
- how business and operational models that tackle the cost curve head-on can be prioritized instead of going it alone with in-house programs. Outsourced cloud computing services and SaaS for BDA should be seriously evaluated.

Governance and Legislation

Stakeholders should look to assess the more complex challenges and risks around legislation and governance. They should consider conducting a review of existing legislative obligations and what gaps may potentially exist through BDA in governance guidelines and models.

Some suggestions include:

- What is the governance required on how these solutions are deployed? What new data needs to be collected? What collaboration is required? Where are these solutions most needed? At point-of-care, within a hospital?
- As an ever larger amount of data is digitized, collected and travels across organizational boundaries, is there a need to review existing legislative obligations and gaps in governance guidelines and models?
- What are the information governance best practices that are applicable to BDA?

Information Management Processes

Stakeholders should considerer reviewing and assessing whether their existing information management processes require revision to incorporate concepts required to support BDA.

Privacy, Ethics and Legal Issues

Stakeholders should start addressing some of the larger risks and challenges based on decisions made around legislation and governance.

Considerations include:

- Need for new policies in privacy and security based on changes to legislation. Agreement on common, clear and consistent terminology for appropriate uses of BDA. Some best practice considerations are required to ensure privacy of the patient is maintained, while enabling the health system and patient to benefit from BDA. Assess changes required to policies on the ethical use of these data types and sources need. As well, the medico-legal implications for clinician's and patient's use of BDA needs should be looked at.
- Breaking down traditional silos and encouraging privacy enhanced data sharing between stakeholders such as research centres and jurisdictions.

Legacy, Technology and Techniques

Today, legacy systems and incompatible standards and formats may prevent the integration of data and the more sophisticated analytics that create value from BD. Stakeholders, as part of revising their procurement processes, should look to refresh technologies, based on new standards, and requirements specific to BDA.

6.3 BDA: Next Steps in the Medium Term

Over the next three to five years, *Infoway* anticipates progress in BDA in Canada as repeatable use cases evolve. While we expect a good portion of the investments will still go into research initially (e.g., personalized care and genomics), we do expect initial deployment of predictive analytics or CDS solutions (e.g., natural language processing, artificial intelligence), remote patient monitoring and BDA based social media or networks solutions in health care.

In the short to medium term, stakeholders should consider addressing the following challenges:

Skilled Workforce

Stakeholders should look to work with research, associations, academic and university communities to understand the wide range of skillsets required for BDA, what the gaps in health are and what actions are required to address the skills gaps for data scientist, health informatics staff, clinicians and researchers.

BDA and the Patient-consumer

Stakeholders are still challenged in engaging consumers to take a more active role in managing their own health. As digitization of health information continues over the next three-four years, *Infoway* expects that patients will increasingly demand more access to their personal health information, including mobile access, and will expand their use of social media for health-related purposes. Patients will be empowered by access to unbiased information on treatment options, benefits and drawbacks, and information to help them make informed and healthy choices about their lifestyle.

Research shows that the Internet has big potential for significantly engaging consumers in managing their health[29]. BDA can provide new ways to use health IT tools and informatics to get to know consumers better and what motivates them to change.

6.4 BDA: Next Steps in the Long Term

Infoway expects BDA will become mainstream in clinical and administrative practice in larger health care settings across Canada within five to ten years. Investments by various levels of government will drive an increased capacity for integrating BDA-based research into clinical practice through genomics and personalized care.

Not surprisingly, this innovation journey is not about the BDA tools or methods that will enable transformational benefits but about the people and processes that surround it. Yet while cautious first steps have been recommended, the application of BDA at scale in health care will hopefully yield handsome benefits to future generations of Canadians.

[29] PwC and PWC Needles in a Haystack; Seeking Knowledge with Clinical Informatics, Health Research Institute, February 2012.

7 Conclusion

As the technological and business aspects of BDA mature, there is a significant opportunity for stakeholders in Canada to consider its use as a strategic enabler for delivering new insights that will serve to advance personalized care, improve patient outcomes and avoid unnecessary costs. BDA presents new opportunities to combine health science domain expertise with data science to provide the insights that decision makers need to transform health delivery.

To take advantage of these capabilities, stakeholders across Canada need to start out with foundational initiatives to gain a better understanding of what it will take to experiment with BDA, including the development of use cases and business cases. *Infoway* believes that BDA will continue to evolve and become a prominent component of most stakeholder's information management plans and activities. Business leaders' attention will need to focus on key challenges of governance, policy, and privacy of personal health information.

Infoway hopes that this white paper has provided enough context and meaningful information to contribute to strategic discussion about the approaches to the adoption of BDA in the health care sector in Canada.

8 Bibliography

1. Analytics: The New Path to Value, MIT Sloan Management Review, IBM Global Business Services, Business Analytics and Optimization, *By Steve LaValle, Michael Hopkins, Eric Lesser, Rebecca Shockley and Nina Kruschwitz.*

2. Content and Predictive analytics in Healthcare, IBM - February 2012.

3. Data Deluge; Mastering Medicine's Tidal Wave, Special Report, BIG DATA: What it Means for our Health and the Future of Medical Research, By Krista Conger Stanford Medicine, Summer 2012.

4. Drowning in Big Data, Reducing Information Technology Complexities and costs for Health Care, Frost and Sullivan.

5. Frost and Sullivan, Drowning in Big Data? Reducing Information Technology Complexities and Costs For Healthcare Organizations, 2012.

6. Gartner Research, Magic quadrant for business intelligence, Information governance - 12 things to go in 2012, Gartner_Keynote_Big_Data_and_Master Data Management.

7. HIMSS Whitepaper "11 Disruptive Technologies That Will Change the Face of EHRs – and How Your Competition is Using Them,_Armour, Quinton, Thizy, Didier, Macadamian Oct 2010.

8. Journal of AHIMA 83, no.10 (October 2012): 38-43. "Big Data, Bigger Outcomes." Lorraine Fernandes, RHIA; Michele O'Connor, MPA, RHIA, FAHIMA; and Victoria Weaver, RHIA,Journal of AHIMA 83, no.10 (October 2012).

9. MGI_big_data_exec_summary, McKinsey Global Institute (MGI) in November 2011: *Big Data - The Next Frontier for Innovation, Competition, and Productivity.* For more, see the McKinsey Global Institute report *Big data: The next frontier for innovation, competition, and productivity*, available free of charge online at mckinsey.com/mgi.

10. McKinsey Quarterly big data. 1) Are you ready for the era of 'big data'? Brad Brown, Michael Chui, and James Manyika, October 2011. 2) Big data Competing through data: Three experts offer their game plans.

11. PWC Needles in a Haystack; Seeking Knowledge with Clinical Informatics, Health Research Institute, February 2012.

12. PWC Transforming healthcare through secondary use of health data, Health Industries, David Chin, MD, PricewaterhouseCoopers.

13. Redefining Value and Success in Health Care, IBM Healthcare and Life Sciences, February 2012.

14. Rethinking Health Information Technology on the Journey to Personalized Medicine, Brett J Davis, August 2012.

15. Smart Health and Well Being, the Role of Basic Computing Research, Computing Community Consortium, Computing Research Association, Computing Research Association (CRA), Spring 2011.

16. TDWI_Best Practices Report on Big Data Analytics_4th Quarter 2011.

17. Technology Vision 2011: The technology waves that are reshaping the business landscape, Rippert, Don; Michael, Dr. Gavin; Swaminathan, Dr. Kishore; Accenture 2011.

18. The Canadian Health Care Debate: A Survey and Assessment of Key Studies, The Conference Board of Canada, May 2012.

19. The Big Data Revolution in Healthcare Accelerating Value and Innovation, McKinsey Center for U.S. Health System Reform, June 2012.

20. *The new gold rush* Prospectors are hoping to mine opportunities from the health industry PCW Health Research Institute, May 2011.

21. The Promise of Big Data, Harvard School of Public Health, Spring and Summer series of 2012.

22. The Role of Analytics in Transforming Healthcare, ORACLE Health Sciences, By John Russell.

23. Thomson_Reuters_White_Paper_on_Healthcare_Waste, Where can $700 Billion in Waste be Cut Annually from the US Healthcare Waste, October 2009.

24. Top 10 Healthcare Game Changers: Canada's Emerging Health Innovations and Trends, Sanjay Cherian, Health Industry Lead, Canada, Accenture, 2011.

25. UBS Big Data Report, UBS Investment Research, Q-Series®: Next Biggest Driver in Tech? Big Data From A to Zettabyte, 5 January 2012.

26. Understanding Big Data, Analytics for Enterprise Class Hadoop and Streaming Data, McGraw-Hill, April 2012.

27. Watson and health care How NLP and semantics will transform healthcare, IBM - April 2011, Michael J Yuan, Ringful Health.

Appendix A – Acronyms and Abbreviations

Term	Description
BI	Business intelligence
CDS(S)	Clinical decision support (clinical decision support system)
CHS	Consumer health solution
CIHI	Canadian Institute for Health Information
DW	Data warehouse
EHR/EHRs	Electronic health record/electronic health records
EHRi	Electronic health record infostructure
EHRS	Electronic health record system
EMR/EMRs	Electronic medical record/electronic medical records
ETL	Extract, transform and load
HIAL	Health information access layer
HIS	Hospital information system(s)
HL7	Health level 7
IT/ICT/HIT	Information technology/information and communications technology/health information technology
PHI/PHR	Personal health information/personal health record
POS	Point of service
SaaS	Software as a service
RFID	Radio frequency identification

Appendix B – Glossary

Term	Description
Business intelligence	Forrester Research defines business intelligence as “a set of methodologies, processes, architectures, and technologies that transform raw data into meaningful and useful information used to enable more effective strategic, tactical, and operational insights and decision-making."
Clinical decision support (CDS)	Typically used when referring to a type of system that assists clinicians in making medical decisions. These types of systems typically require input of patient-specific clinical variables, and as a result, provide patient-specific recommendations.
Column oriented DBMS	A database management system (DBMS) that stores its content by column rather than by row. This has advantages for data warehouses and library catalogues where aggregates are computed over large numbers of similar data items. It is possible to achieve some benefits of column-oriented and row-oriented organization with any database. By denoting one as column-oriented we are referring to the ease of expression of a column-oriented structure and the focus on optimizations for column-oriented workloads. This approach is in contrast to row-oriented or row-store databases and correlation databases, which use a value-based storage structure. A column-oriented database serializes all of the values of a column together, then the values of the next column, and so on. Online transaction processing (OLTP)-focused RDBMS systems are more row-oriented, while online analytical processing (OLAP)-focused systems are a balance of row-oriented and column-oriented.
Consumer health solution	An application or suite of applications that provide health care consumers with access to personal health care information, education and health care management tools.
Data scientist	A job title for an employee or business intelligence (BI) consultant who excels at analyzing data, particularly large amounts of data, to help a business gain a competitive edge. The position is gaining acceptance with large enterprises who are interested in deriving meaning from big data. A data scientist possesses a combination of analytic, machine learning, data mining and statistical skills as well as experience with algorithms and coding. Perhaps the most important skill a data scientist possesses, however, is the ability to explain the significance of data in a way that can be easily understood by others.
Data mining	A field at the intersection of computer science and statistics, it attempts to discover patterns in large datasets.

Term	Description
Data warehouse	Gartner defines a data warehouse as "a storage architecture designed to hold data extracted from transaction systems, operational data stores and external sources. The warehouse then combines that data in an aggregate, summary form suitable for enterprise-wide data analysis and reporting for predefined business needs."
De-identified data	Health information that has been manipulated using appropriate de-identification processes. The directly identifying variables have been adequately manipulated and quasi-identifiers adequately disguised to ensure that the re-identification risk is acceptable.
Electronic health record	Provides each individual in Canada with a secure and private lifetime record of their key health history and care within the health system. The record is available electronically to authorized health care providers and to the individual anywhere, anytime in support of high-quality care.
Electronic medical record	A general term describing computer-based patient record systems. It is sometimes extended to include other functions like order entry for medications and tests, among other common functions. For the purposes of this glossary, EMR is the software application used in ambulatory, community clinic or doctors' office settings.
Extract, transform, load	System or utilities that extract data from various data sources, transform them to the required destination format and load them in the target data sources.
Feedback loop	Generic term that describes information flowing back to end users via notification, alerts, reports, dashboards and other decision support tool mechanisms.
Health information access layer	An interface specification for the EHR infostructure (OSI Layer 7) that defines service components, service roles, information model and messaging standards required for the exchange of EHR data and execution of interoperability profiles between EHR services.
Health information	A broad term including but not limited to financial information about health and health care, health information, de-identified data and aggregate data.
Health record	A health record, medical record or medical chart is a systematic documentation of a patient's individual medical history and care across time within a particular health care provider's jurisdiction. From a current state perspective, it is meant to describe the electronic information that exists and in most cases is siloed to particular organizations or care settings. In the future, this will refer to a longitudinal patient record with information from many sources.
Health system use	Refers to using health information to improve the health of Canadians by making better decisions about Canada's health system. It includes the use of health data to improve clinical programs and care protocols, planning and resource allocations across the system, public health and health research.
Personal health record	A health record controlled by the person, or representative of the person, to whom it pertains.

Term	Description
Personalized care	Also known as personalized medicine, personalized care is a rapidly advancing field of health care that promises greater precision and effectiveness than traditional medicine because it is informed by each person's unique clinical, social, genetic, genomic, and environmental information. Personalized care takes an integrated, coordinated, evidence-based approach to individualizing patient care across the continuum from health to disease.
Point of service	A system that is below the HIAL and that is used by end-users at the point of care or service. For example, an EMR, a pharmacy management system (PMS), a HIS or a laboratory information system (LIS).
Predictive analytics	A variety of techniques from statistics, modeling, machine learning and data mining that analyze current and historical facts to make predictions about future events.
Radio frequency identification	Use of a wireless non-contact system that uses radio-frequency electromagnetic fields to transfer data from a tag attached to an object, for the purposes of automatic identification and tracking.
Stakeholder	Refers to a broad group of affected parties, including: governors (health ministries and departments), administrators (health service delivery organizations, local health integration networks), health service providers (physicians, specialists, nurses, technicians), consumers (general public, individuals and groups), researchers (individuals and organizations), and vendors.

Appendix C – BDA Technical Frameworks

Two technology concepts or frameworks that are enabling BD today are MapReduce and Hadoop.[30] These frameworks have been in operation since 2005 at web-based search and online gaming companies. They are now just moving into other industries and sectors including financial services firms and banks, and online retailers.

MapReduce

MapReduce was developed by Google in 2004. It represents a programming framework for processing, generating and indexing large sets of data on the web. Google developed MapReduce as a general-purpose execution engine that handles the complexities of network communication, parallel programming and fault-tolerance for any kind of analytic application (hand-coded or analytics tool based).

MapReduce is a framework for processing highly distributable problems across huge datasets. It is an algorithm based framework for computing distributed problems using xdivide and conquer[31] approach cluster of nodes. MapReduce serves as the compute layer of another technology called Hadoop (see below). MapReduce jobs are divided into two parts. The "Map" function divides a query into multiple parts and processes data at the node level. The "Reduce" function aggregates the results of the "Map" function to determine the "answer" to the query. It consists of Master node which maps input into smaller sub problems/distributes work to clusters, where these worker nodes process smaller problems, return answers back to master node and the master node reduces set of answers back to master node.

It divides the basic problem into a set of smaller manageable tasks and assigns them to a large number of computers (nodes). An ideal MapReduce task is too large for any one node to process, but can be accomplished by multiple nodes efficiently.

MapReduce is named for the two steps at the heart of the framework.

- **Map step** – The master node takes the input, divides it into smaller sub-problems, and distributes them to worker nodes. Each worker node processes its smaller problem, and passes the result back to its master node. There can be multiple levels of workers.
- **Reduce step** – The master node collects the results from all of the sub-problems, combines the results into groups based on the key and then assigns them to worker nodes called reducers. Each reducer processes those values and sends the result back to the master node.

MapReduce can be a huge help in analyzing and processing large chunks of data: buying pattern analysis, customer usage and interest patterns in e-commerce, processing the large amount of data generated in the fields of science and medicine, and processing and analyzing security data, credit scores and other large data-sets in the financial industry.

[30] While this profile depicts the concepts of MapReduce/Hadoop, there are other vendors who are working on complementary or similar concepts to these BDA technologies (e.g., NoSQL, document based, and others). Please refer to Appendix D for more information.

[31] A divide and conquer algorithm works by recursively breaking down a problem into two or more sub-problems of the same (or related) type, until these become simple enough to be solved directly. The solutions to the sub-problems are then combined to give a solution to the original problem.

Hadoop

Hadoop was developed by Yahoo in 2005. It represents a set of software and open source distributed platform for processing large sets of data. The Apache™ Hadoop™ project develops open-source software for reliable, scalable, distributed computing. The Apache Hadoop software library is a framework that allows for the distributed processing of large datasets across clusters of computers using a simple programming model. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage. The library itself is designed to detect and handle failures at the application layer, delivering highly-available services on top of a cluster of computers, each of which may be prone to failures.

Hadoop was designed to handle petabytes and exabytes of data distributed over multiple nodes in parallel. Apache Hadoop's open-source software library is available from ASF at http://www.apache.org. In addition to Hadoop and MapReduce, the Apache Hadoop library also includes other concepts to support BD, which either extend or complement:

- Hadoop distributed file system (HDFS) can manage the storage and access of any data type as long as you can put the data in a file and copy that file into HDFS. As outrageously simplistic as that sounds, it's largely true, and it's exactly what brings many users to Apache HDFS.
- Hive is a Hadoop-based data warehouse developed by Facebook. It is a system that lets users structure and query data in a SQL-like code. This allows SQL programmers with no MapReduce experience to use the warehouse and makes it easier to integrate with business intelligence and visualization tools.
- Hbase is a non-relational database that allows for low-latency, quick lookups in Hadoop. It adds transactional capabilities to Hadoop, allowing users to conduct updates, inserts and deletes. EBay and Facebook use HBase heavily.
- Pig Latin is a Hadoop-based language developed by Yahoo. It is relatively easy to learn and is adept at very deep, very long data pipelines[32](a limitation of SQL).

BD vendors and the IT industry are contributing to the "Apache Hadoop project," and the technology is advancing rapidly, becoming more powerful and easier to implement and manage.

There are many others components of these frameworks such as Zookeeper, Flume, Sqoop, Oozie, Hue and Cassandra. As an example, Cassandra was developed by Facebook as an open source database and is used by Netflix, Twitter, Constant Contact, Cisco and others. For users desiring a more enterprise-ready package, new vendors are emerging which are offering Hadoop distributions that include additional administrative tools and technical support.

These products can be combined in various ways, but the Hadoop HDFS and MapReduce (perhaps with Hbase and Hive) constitute a useful technology stack for new BDA applications in BI, DW and analytics.

[32] Realtime Data Pipelines. Data is fed in one end, results are computed in real time as data flows down the pipeline and comes out the other end whenever relevant changes we care about occur. Data Pipelines/workflow/streams are becoming much more relevant for processing massive amounts of data with real time results. Moving relevant forms of analytics out of large repositories into the actual data flow from producer to consumer is becoming a fundamental step forward in BD management. These new applications and data analysis are being engineered to use data constantly rather than a series of push and pulls from a persisted store. Data is originating from somewhere and is heading for consumption somewhere else, either in raw state or part of newly derived data.

Other Analytic Components of the BD Framework

In addition to traditional DW concepts and advanced analytic tools like data mining, text analytics and natural language processing, organizations and companies are also taking advantage of open source tools for building analysis tools for BD. One example of these open source environments being leveraged by a large number of vendors is called R[33]. R is an open source statistical language that was launched in the 1990s. R is one of many tools being adopted by enterprises because it is a very comprehensive open source tool and has similar capabilities to traditional enterprise analytic/data mining tools. In addition to the language, there are also several open source analytic development environments (ADE) or front-end Graphical User Interface (GUI) tools for creating analytics and R-based models. When people refer to R they are referring to the huge library of thousands of open source analytics packages called CRAN (Comprehensive R Archive Network). The R CRAN library includes general interest categories such as statistics, time series, econometrics, machine learning, high-performance computing and natural language processing.

[33] See http://www.rdatamining.com/ for more information on R.

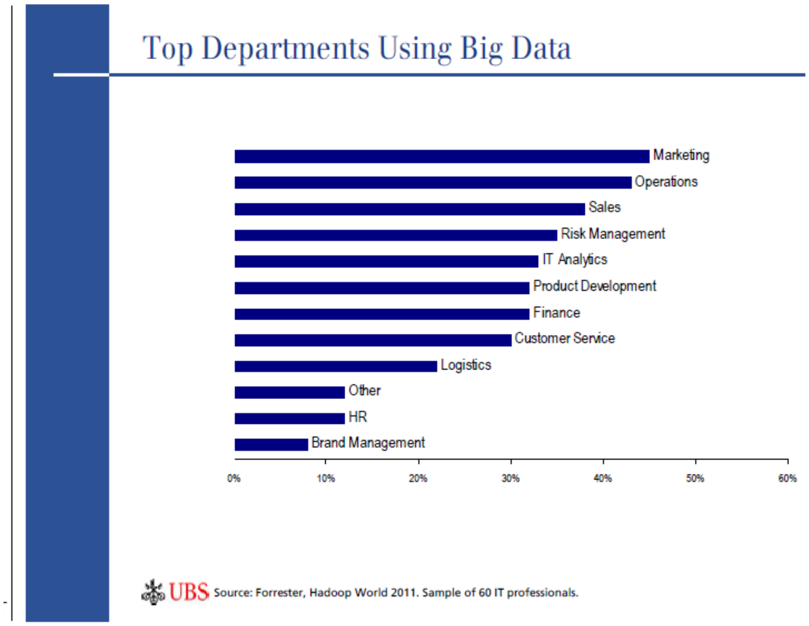
Appendix D – General Industry Examples of BDA

BDA technologies and approaches are allowing enterprises to find answers to questions they didn’t even know to ask. This can result in insights that lead to new product ideas or help identify ways to improve operational efficiencies.

There is significant evidence through identified use-cases such as web giants like Google, Facebook and LinkedIn and for the more traditional financial and retail industries.

As an example, when posting things on Facebook brings a related ad to a person’s page or posting, this is driven by BDA. Another example is the use of BDA by Netflix when recommending movies to customers. The estimate is that about 70 per cent of the movies that people pick on Netflix come from a “recommendation engine” powered by BDA.

Figure D1 - Top Uses of BDA [34]



UBS Source: Forrester, Hadoop World 2011. Sample of 60 IT professionals.

Figure D1 represents top use cases for BDA across various industries and departments. As one would expect, high on the list are marketing, customer analytics, sales analytics and risk management, based mostly on traditional retail and financial sectors.

The following are some additional examples and sampling of BDA use cases:

Recommendation Engine

Web properties and online retailers use BDA technologies to match and recommend users to one another or to products and services based on analysis of user profiles and behavioral data.

[34] Figure used with the permission of UBS.

LinkedIn uses this approach to power its “People You May Know” feature, while Amazon uses it to suggest related products for purchase to online consumers.

Sentiment Analysis

Used in conjunction with BDA technologies, advanced text analytics tools analyze the unstructured text of social media and social networking posts, including Tweets and Facebook posts, to determine the user sentiment related to particular companies, brands or products. Analysis can focus on macro-level sentiment down to individual user sentiment.

Risk Modeling

Financial firms, banks and others use BDA technologies and analytics to analyze large volumes of transactional data to determine risk and exposure of financial assets, to prepare for potential “what-if” scenarios based on simulated market behavior, and to score potential clients for risk.

Fraud Detection

BDA technologies and techniques for combining customer behavior, historical and transactional data, are used to detect fraudulent activity. Credit card companies, for example, use BDA technologies to identify transactional behaviour that indicates a high likelihood of a stolen card.

Marketing Campaign Analysis

Marketing departments across industries have long used traditional data warehousing technology to monitor and determine the effectiveness of marketing campaigns. BDA technologies allow marketing teams to incorporate higher volumes of increasingly granular data, like click-stream data and call detail records, to increase the accuracy of analysis.

Customer Churn Analysis

Enterprises use BDA technologies to analyze customer behaviour data to identify patterns that indicate which customers are most likely to leave for a competing vendor or service. Action can then be taken to save the most profitable of these customers.

Social Graph Analysis

In conjunction with BDA technologies and analytics, social networking data is being mined to determine which customers pose the most influence over others inside social networks. This helps enterprises determine which are their “most important” customers. They are not always those who buy the most products or spend the most but those who tend to influence the buying behaviour of others the most.

Customer Experience Analytics

Consumer-facing enterprises use BDA technologies to integrate data from previously siloed customer interaction channels such as call centres, online chat and Twitter, to gain a complete view of the customer experience. This enables enterprises to understand the impact one customer interaction channel has on another to optimize the entire customer lifecycle experience.

Network Monitoring

BDA technologies have been used to ingest, analyze and display data collected from servers, storage devices and other IT hardware to allow administrators to monitor network activity and diagnose bottlenecks and other issues. This type of analysis can also be applied to other forms of networks (e.g., transportation networks, to improve fuel efficiency).

Research and Development

Enterprises, such as pharmaceutical manufacturers, use BDA technologies to comb through enormous volumes of text-based research and other historical data to assist in the development of new products.

An example of this is ClickFox. ClickFox tracks every individual interaction that every customer has with a given company, at every step in the path, across all channels, and delivers insights on the trends: where breakdowns and successes occur, and where opportunities lie for improvement and growth. ClickFox takes numerous existing interactional data systems (e.g., web, phone centre, retail store) and combines them into a unified, cross-channel, experience analytics framework. ClickFox then renders easy-to-understand visualizations that represent behaviours within and between the various data systems.

ClickFox is able to look beyond the "what" of customer behaviour to get to the more important "why," the reliable and actionable insights into data to help organizations make smart business decisions.

Appendix E – Additional Health Care Examples

Example 1: Clinical Decision Support (CDS) systems and use of BDA in the U.S.

In the future using BDA, CDS systems can become more intelligent by including modules that use image analysis and recognition in databases of medical images (X-ray, CT, MRI) for pre-diagnosis. They would automatically mine medical literature to create a medical expertise database capable of suggesting treatment options to physicians based on patients' medical records.

IBM is working with several organizations in the U.S. on research projects where Watson is being piloted for its ability to support questions and answers for diagnosis. The hope is that, by providing additional diagnostic advice/insights. IBM's Watson is expected to represent a technology breakthrough that can help physicians manage the information "tsunami" of BDA to affect clinical decisions and patient outcomes. Watson also represents a new class of BDA solutions and type of decision support system that uses deep content analysis and evidence-based reasoning and natural language processing on BD. Watson can accurately extract medical facts and quickly understand and learn from relationships buried in large volumes of data, such as electronic medical records, family medical history, and the latest clinical research. Natural language processing technology can help accelerate and improve clinical decisions, reduce operational waste and enhance patient outcomes.

As an example, IBM and Memorial Sloan-Kettering Cancer Center are using Watson to treat oncology patients. Watson relies on parallel probabilistic algorithms to analyze millions of pages of unstructured text in patient records and the medical literature to locate the most relevant answers to diagnostic and treatment related questions. This represents a massive collection of information and papers that otherwise people would have to read. Watson reads it for them.

Sloan-Kettering has about 2,000 order sets it can pull from when choosing a cancer treatment. Finding the best fit for each patient is no easy task. Sloan-Kettering can tap its own massive database, called Darwin, which includes everything that has happened to all of its 1.2 million inpatients and outpatients over 20-plus years.

The two organizations are combining all of Darwin's intelligence with all of Watson's natural language processing capabilities. IBM is using all of Sloan-Kettering's structured patient data and its NLP tools to convert the medical center's free text consult notes into usable data. The team will first use this approach to tackle non-small-cell lung cancer. Watson is focusing on 14 to 20 data elements, including the size and location of a patient's tumor, the presence of any genetic mutations (Sloan-Kettering does a full genomic analysis on all of its lung and colon cancer patients), and whether the tumour has spread to other tissues.

Watson's task will be to follow the protocol outlined above and come back with a list of diagnostic and treatment options for physicians to choose from, with confidence ratings for each option. Ideally, a treatment regimen that Watson concludes has a 95 per cent confidence rating, for example, would help oncologists choose from the 28 different chemotherapy cocktails they have at their disposal. The organization hopes to launch by the end of 2013 a pilot program that will allow the supercomputer to work on real cases.

Example 2 – Genomes and cancer research

The world's largest set of data on human genetic variation – produced by the international 1000 Genomes Project – is now freely available on the Amazon Web Services (AWS) cloud. At 200 terabytes – the equivalent of 16 million file cabinets filled with text, or more than 30,000 standard DVDs – the current 1000 Genomes Project dataset is a prime example of BD, where datasets become so massive that few researchers have the computing power to make best use of them. AWS is storing the 1000 Genomes Project as a publicly available dataset for free and researchers will pay only for the computing services they use.

A further extension of this example of BDA is cancer research, which is progressing from a single human genome to the HapMap[35], and now to inexpensive whole-genome sequencing, and is part of the 1000 Genomes Project. Genome-wide association studies have identified hundreds of genotype–disease linkages, some of which have strong clinical implications. All this work will soon result in the almost overwhelming growth of the fundamental substrate for personalized medicine: data.

Genome Canada was established in February 2000 with a mandate from the Government of Canada to develop and implement a national strategy for supporting large-scale genomics and proteomics research projects, for the benefit of all Canadians. Since 2000, Genome Canada has received $915 million in funding commitments from the Government of Canada. Genome Canada has used this funding, along with co-funding of more than $900 million from other organizations, to finance innovative, large-scale research projects in genomics, proteomics and related activities. Recently, Genome Canada and the Canadian Institutes of Health Research (CIHR), through its Personalized Medicine Signature Initiative, are jointly supporting research projects in the area of genomics and personalized health. Their 2012 Large-Scale Applied Research Project Competition in Genomics and Personalized Health aims to support projects that will demonstrate how genomics-based research can contribute to a more evidence-based approach to health and improve the cost-effectiveness of the health care system.

Example 3 - Online platforms and communities

New business models are being enabled by BDA online platforms and communities, which are already generating valuable data. Examples of this business model in practice include web sites such as PatientsLikeMe.com, where individuals can share their experience as patients in the system; Sermo.com, a forum for physicians to share their medical insights; and Participatorymedicine.org, a web site that encourages patient activism.

Examples include:

- Cambridge Temperature Concepts makes a product designed to help women with fertility problems conceive. Women wear a sensor which records movement and changes in body temperature (an indication of ovulation), and gives up to 20,000 readings a day. This is valuable medical data. Women using this product report anything that might affect their body temperature, allowing the development of profiles of illnesses without ever setting foot in a hospital. More importantly it is possible to see what those illnesses look like in the general population who are otherwise normally healthy. People who have a cold don't go to the doctor.

[35] The HapMap project is an organization that aims to develop a haplotype map (HapMap) of the human genome, which will describe the common patterns of human genetic variation. HapMap is a key resource for researchers to find genetic variants affecting health, disease and responses to drugs and environmental factors. The information produced by the project is made freely available to researchers around the world.

Also, the product records this information while the women are asleep. For the first time, extensive data on what normal sleep looks like is being collected. This gives researchers a huge control group to do things like comparing sleep patterns of "normal" people with those of pain sufferers.

- San Francisco-based SeeChange offers unique ways of designing health insurance plans, with what it calls "value-based benefits." In addition to paying for care when a patient is sick or in an accident, SeeChange's plans offer financial incentives, cash rewards or lower out-of-pocket costs to patients who complete specific action plans. The company uses a substantial amount of data gleaned from personal health records, claims databases, lab feeds and pharmacy data to identify patients with chronic illnesses who would benefit from a customized compliance program. Once a patient is enrolled, the data analytics engine is used to mine and monitor compliance and to continue to customize the patient's experience to meet clinical goals.

  SeeChange Health Solutions offers these as Software as a Service (SaaS)-based turnkey solutions that enable large self-funded employers and health plans nationwide to offer highly customized engagement and incentive programs that motivate employees to improve their health. The solution includes:

  - ***Robust data analytics*** to identify population health risks by analyzing large datasets of historical claims data to identify the likely percentage of undiagnosed conditions that exist within a workforce that require intervention and ongoing care.
  - ***Early detection screening model*** to identify an individual's health status and chronic health conditions utilizing tools such as historical claims analysis, health questionnaire, biometric (lab) test and a preventive exam including age/gender appropriate cancer screenings to detect: pre-diabetes, diabetes, asthma, coronary artery disease, elevated cardio-vascular disease, metabolic syndrome, depression and cancers (breast, colon, prostate and cervical).
  - ***Personalized Health Action Plan*** that creates a "plan of action" for each member identifying the specific health actions each should complete based on age, gender and condition-specific guidelines to better manage their own health. Online coaching and disease management programs can be offered based on each individual's results. The Personalized Health Action Plan is continuously updated with medical claims, biometric (lab) results and pharmacy claims so progress in achieving preventive and condition-specific health actions, and the rewards earned as a result, are easy to track and manage.
  - ***Health action*** rewards that reward individuals for participation and/or outcome improvements. The specific rewards are determined by the employer and may include cash rewards, premium credits, enhanced benefits or merchandise.

9 Contact

Infoway established the Emerging Technology Group (ETG) in 2011 to identify and guide the use of information and communications technologies (ICTs) in health care innovation. The ETG's role is to identify and evaluate emerging technologies, and mature technologies that haven't been fully applied, that look most likely to provide significant benefits to our health system and the health of Canadians.

This white paper is part of an ETG white paper series which aims to provide information and analysis that could benefit those who make decisions about technologies for health in Canada.

For more information about the ETG and its work, contact ETG@infoway-inforoute.ca or visit our website.