THE BIG DATA GAP: IS THERE CONGRUENCE BETWEEN DATA SKILLS DEMANDED BY THE INDUSTRY AND CANADIAN ACADEMIC PREPARATION?

By

Michal Wieczorek
Bachelor of Business Administration Brock University 2015

A thesis
presented to Ryerson University
in partial fulfillment of the
requirements for the

Degree of Master of Science in Management in the program of

Master of Science in Management

Toronto, Ontario, Canada, 2019

AUTHOR'S DECLARATION

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I authorize Ryerson University to lend this thesis to other institutions or individuals for the purpose of scholarly research I further authorize Ryerson University to reproduce this thesis by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

I understand that my thesis may be made electronically available to the public.

Abstract

The Big Data Gap: Is there Congruence between Data Skills Demanded by the Industry and Canadian Academic Preparation?

Degree of Master of Science in Management in the program of

Master of Science in Management

Ryerson University, Toronto, Ontario, Canada, 2019

Michal Wieczorek

This thesis examines occupations in the Canadian Big Data industry. Through analysing Canadian job advertisements, the aim is to understand what skills a professional in a data occupation requires. Supplementing with a content analysis of graduate master's programs focusing in Data Science, Analytics or Big Data the study explores Canadian educational institution offerings which prepare students for jobs in the field. Using topic modeling methods and typologies results show a fit between universities' presented content and what skills are demanded by the industry. An updated framework of Todd et al. (1995) is presented for easier comparison and recommendations of creating undergraduate data programs are discussed. Contributions made by this study can aid universities in a structuring their curricula for "Big Data" programs. Furthermore, this study contributes to the literature by explaining multiple job qualifications which allows for more standardized job descriptions, benefiting the employers, job seekers and universities.

Acknowledgements

First and foremost, I would like to thank Dr. Murtaza Haider, my main supervisor. My committee Dr. Morteza Zihayat Kermani and Dr. Farid Shirazi who carefully and generously supported my research. I have learned a great deal from all and am grateful to them.

I would like to thank all library specialists for their aid in the gathering of data who supported me, they helped me achieve results of better quality. I would also like to thank the members of my committee for their constructive criticism and suggestions in improving my work numerous times.

I would like to thank the all the professors who made my goal much more attainable by supporting me. I would like to thank my fellow master’s peers for their feedback, cooperation and of course friendship.

Lastly I would like to thank my family: my parents and sisters for supporting me throughout my research and writing this thesis.

Table of Contents

List of Tables

List of Figures

Chapter 1 – Introduction

1.1 Background

Data Science and Big Data are increasing in importance for corporations, research institutions and health care organisations. Growing interest from these entities translates into a larger demand for workers with the necessary data skill sets. The interest in data is driven by many factors but the most common one among all organizations is value. The outcome of the term "value" may differ from place to place but the definition stays the same, value is derived from data which is important or useful. The value of Big Data for organizations has been referred to as, ‘‘the mother lode of disruptive change in a networked business environment" (Baesens et al., 2014 p.629). Through the use of Big Data technologies, organizations try to extract value from stockpiles of data to turn domains such as e-commerce, security, health, etc… into profit (Chen et al., 2012). Through well trained human resource talent, organizations can even use Big Data as a competitive advantage (Davenport, 2006). Big Data is a recent phenomenon which does not hold a static definition however; it can be narrowed down with the help of its "V's". The "V's" of Big Data are characteristics which describe the state of current data environment from multiple angles. Since their coinage by Laney (2001) with three V's (volume, variety, and velocity) they have grown to include value, veracity, viscosity, virality and visualisation with the latter descriptions being used seldom than the former. "Volume" is most often associated with Big Data since "voluminous" and "big" are similar terms in their grammatical use. Big Data is BIG, there are a lot of data, in firms' data centers, in research facilities, on your hard drive, in your cell phone, the sheer amount of data stored is enormous, Google's search data is estimated to be 10-15 Exabytes roughly translated into 1 billion Gigabytes (Tharan, 2017; Munroe, R. 2013) which excludes Google's phone data, GoogleDrive services, and everything else other than search history metadata. In order to understand Big Data, it is beneficial to understand "Velocity" which puts the amount of data created in a period of time into perspective. Big Data is fast, it is created quickly and on a large scale, most notable examples of the speed at which data is created are the social media platforms. In every single minute 473,400 tweets are sent, 49,380 Instagram pictures are posted, Amazon ships 1,111 packages, Netflix streams 97,222 hours of videos, Uber users take 1,389 rides and many more. "Variety" is defined as the different types and formats of data. The structure of current data does not match the structure of data from the past (name, phone number, address, financials, etc) earlier data was easily fit into a data table while data in 2018 is unstructured. This category includes photos, videos, sensor data, etc. Big Data allows structured and unstructured data to be harvested, stored, and used simultaneously. The lesser known V's are complimentary to explaining Big Data but do not provide as much explaining power as the initial three.

Taking advantage of Big Data has been revealed to be a key factor in the success of corporations (Davenport, 2006) however in order to take advantage of Big Data there must first be data professionals to make the data valuable (Chen et al., 2012; Tambe, 2014). McKinsey Global Institute (MGI), in their 2011 report, predicted a shortage of employees with skills in Big Data (Manyika et al., 2011). Related reports confirmed the trend that there is a need for data savvy employees, especially for organizations which are generating abnormally large amount of information. MGI estimated a shortage of data knowledgeable employees to be between 140,000 and 190,000 alongside 1.5 million managers and analysts (Manyika et al., 2011). The supply of data professionals is not meeting the demand of the market while human capital has been proven necessary to the success of technology intensive organizations (Colombo & Grilli, 2010; Delgado-Verde, Martín-De Castro, & Amores-Salvadó, 2016; Morales-Alonso, Pablo-Lerchundi,

& Núez-Del-Río, 2016; Siepel, Cowling, & Coad, 2017). A supply shortage causes organizations to try to quickly obtain employees with the best skills however, not always succeeding. Entire industries are vying for data professionals in numbers greater than ever before to make sense of the information that are lying idle in their digital storage. Organizations are progressively struggling with the challenges of handling Big Data, finding its proper fit and human resource challenges. Thus the knowledge of Big Data concepts is a highly attractive quality to have (McAfee and Brynjolfsson, 2012), but the required proficiencies are often overshadowed by the glaring need to fill the seats.

Data Science is a domain similar to Big Data, sometimes overlapping in concepts but both complement each other. It is a domain where many quantitative disciplines intersect to extract insights from data in structured or unstructured formats (Dhar, 2013; Leek, 2013). Methods, algorithms, processes and systems from areas like mathematics, statistics, information science (IS), and computer science (CS) converge to analyze and understand events with the help of data. Sometimes call the "Fourth Paradigm" of science beyond theoretical, empirical and computational, it aims to solve problems using data-driven means (Tansley, 2009; Bell, 2009). The amalgamation of the skills from specific disciplines listed above is a difficult task to manage, let alone be an expert in. Data Science shares the same demand problems as Big Data where the industry is young, the talent pool is small and demand is growing. The grip on Data Science appears to be in the hands of the industry and so, very little has come from the academia about the subject. Whether it is the academia who are lagging behind or the speed of economic change, research has not caught up to interpret many aspects of Data Science's discipline. The current state of the data job market is out of balance with many organizations demanding data scientists but being met with none or little competent applicants. There can be many reasons for the gap however, this study will focus on the problem of supply of competent workers. With the growing need for skilled employees there is a need to train them. Due to the short supply and high demand of data professionals, a number of Data Science and Big Data graduate level programs in universities have been created. The demand for Data Science and Analytics programs has increased beyond capacity, to a point where educational institutions receive up to twenty applications per open position (Turel & Kapoor, 2015). With the influx of graduate level programs there is a need to create guides and curricula as to teach the students proper knowledge, methods and concepts (Wixom, 2014). There is evidence of a skill gap between data professionals and students graduating from current university programs (Turel & Kapoor, 2015). There are nearly 250 graduate programs in Data Science, Big Data and Analytics in the U.S. (NCSU, 2018) and nearly 30 in Canada, this amount of programs will not be able to produce enough analysts for the market to be anywhere close to satisfied. There are only few undergraduate programs in Data Science or similar which means that organizations will not be able to find such talent sustainably (Dubey and Gunasekaran, 2015; Manyika et al., 2011; Wixom et al., 2014). Even though the number of programs and courses related to Data Science is growing, their curricula and topics covered are not in line with what the industry demands their employees know (Turel and Kapoor, 2016; Wixom et al., 2014). Organizations expect graduates to use and apply the knowledge to bring value to the organization (Bhimani & Willcocks, 2014; James, Maringer, Palade, & Serguieva, 2015; Loebbecke & Picot, 2015; Ulrich & Dulebohn, 2015). The lack of properly trained employees brings costs in hiring, retraining, lost time, mistakes which all take away from revenues. Since not all organizations have the resources to train or re-train employees, universities are in the position where proper education can shrink or close the skill gap. The ability to prepare, analyse and explain data is an essential skill to have (McAfee and Brynjolfsson, 2012) but specific skills of data scientists have not been explored in depth. With some exceptions (Debortoli et al., 2014), a proper and practical set of data skills has not been synthesized in North America. Handling data is a relevant skill for many jobs however this study will focus on Data Science environment with complimentary engineering, research and other positions.

In the job market, jobs have various titles originating from multiple firms, from different sectors and industries. This poses a problem of subjectivity where a specific position – example "Data Scientist" – will have different requirements for onboarding and different responsibilities when on the job, while having the same title. The description of requirements of data professions is often vague and unclear, firms rely on their subjective needs which may be understandable within house but confusing for external applicants. The market must define borders for data occupations and differentiate between positions for greater clarity among employees and employers (Provost and Fawcett, 2013). Simplifying or generalizing descriptions is also not a good strategy since it neglects the complexity of specific competencies that are required to organize and transform data into actionable insights which will lead to value. It has also been noted there exists a gap between the formal definition of data jobs and their required educational needs (Miller, 2014; Song & Zhu, 2015). This study will look at job descriptions covering positions related to Data Science, Data Analysis, Data Mining, Machine Learning, Big Data, Artificial Intelligence, Business Intelligence and Engineering.

1.2 Research Objective

The objectives of this study are to explore the diverse set of competencies required from Data Science professionals, examine the current state of Data Science education in graduate level programs and ultimately compare those requirements with the training universities provide to determine a connection between what students are taught and what they should know. For this purpose, I propose a framework based on Todd et al. (1995) typology. Written for the information systems (IS) occupations, Todd et al. (1995) competence framework stratifies jobs and skills into categories, subcategories and provides definitions for them. The new framework will aid in the research objectives by supplementing areas which Todd et al. framework has missed. In order to understand and quantify what the industry demands of data professionals job advertisements are collected and analyzed from the Canadian job search engine Indeed (www.indeed.ca).

Job advertisements are the main method of hiring and finding new employees (Walsh et al., 1975), it is a means of notifying intent to hire employees to perform tasks by organizations. Job advertisements give an overview or a description of the position in question including requirements set by the hiring party. One limitation of this way is the knowledge and certainty of the necessary requirements by the hiring party. In this study an assumption is made that the author has complete knowledge of necessary tasks and activities essential for the successful execution of the job.

To determine any connection between the training of educational institutions and the demanded skills set by the industry a content analysis of 18 Canadian university programs has been conducted. The programs are all graduate, excluding certificates, all are in-class, stemming from multiple departments including Management, Engineering, Mathematics or Computer Science and are presented as focusing on “Big Data”, “Data Science” or “Analytics”.

On the basis of those statements the following research questions are formulated:

RQ1: What are the required competencies for data professionals in Canada?

RQ2: What data skills are taught in data programs in Canada?

RQ3: Do data skills taught in data programs impart the competencies required of data scientists?

The remainder of this thesis is organized as follows. Section 2 deals with the thorough review of current literature on required skills of data professionals, graduate level university programs focused on Big Data, Data Science or Analytics. Section 3 deals with the theoretical framework of typologies and the Todd et al. (1995) competence framework. Section 4 covers the data collection, explanation and application of the research methodology, analysis of job ads and the analysis of curricula from graduate programs. Section 5 describes the problem, displays helpful statistics as well as the theoretical model which is used to graphically depict the methodology. Section 6 discusses the results along with contributions and successful implications of the research. The study concludes with a recap of assumptions versus findings, limitations of the study and suggestions for future research in section 7.

Chapter 2 – Literature Review

2.1 Big Data and Data Science

Data Science is a field of academia with a multidisciplinary nature. It includes methods, processes and algorithms from quantitative areas such as mathematics, statistics, computer science, information systems and economics (Zawadzki, 2014; Schoenherr & Speier-Pero, 2015). It overlaps topics with Big Data, Machine Learning, Data Mining and uses data focused approaches to solve complex problems.

Big Data is the concept of constantly increasing pools of information being created from multiple sources at different rates and captured by various methods. Big Data was originally described by three V's: volume, variety, and velocity (Laney, 2001), it has since added variety, value, veracity and visualisation with the latter descriptions being used seldom than the former. Big Data has become more popular in the recent years (Gandomi & Haider, 2014) and is widely used to gain competitive advantage (Davenport, 2006).

2.2 Data Occupations

This section describes the state of Big Data and Data Science occupations. In the current job market companies have a constant demand for data professionals, but the amount of professionals available is inadequate (Dubey and Gunasekaran, 2015; Manyika et al., 2011; Wixom et al., 2014; Wixom et al., 2011). The consequences of this imbalance are inflated salaries, rushed employment and lax standards (N.N., 2016). Big Data and Data Science occupations deal with the collection, extraction, storage, cleaning, manipulation, loading and presenting data. By combining multiple disciplines data occupations differ between companies and may include titles such as Data analyst, Data Mining engineer, Data Scientist, Database administrator, Machine learning engineer, Big Data architect, Consultant and others. These occupations also encompass cloud technology, artificial intelligence, Internet of things (IoT) as well as machine learning. The rise of new platforms and applications allows for a number of occupations to be created every day. Trends show an upswing in adoption of enterprise platforms, Big Data technology, data security, decentralization of computing power, open source software (Persistence Market Research, 2018). With the explosion of such concepts the data professions have emerged and one specific title stood out among the rest, the “Data Scientist”. Termed as the “Sexiest job of the 21st century” (Davenport & Patil, 2012) it brought immense attention to Big Data and Data Science. According to Davenport & Patil (2012) a data scientist is a highly positioned professional with capabilities and curiosity to discover new knowledge from the data at hand. Capability in this context is the ability to perform or achieve certain actions or outcomes successfully. The umbrella term “capabilities” used in this study is meant to encompass interchangeable language containing terms like abilities, competences, capacity, knowledge, qualifications, skills, etc… for easier understanding (Müller et al., 2014). The increasing amount of data produced daily (Ikemoto, Marsh, 2008; McKendrick, 2015) increases the demand for workers who know how to deal with such data and who can transform it into a useful result. Other capabilities mentioned are identification of data sources, teamwork, communication skills, coding and database knowledge (Davenport & Patil, 2012). However, with only promised rewards and unclear boundaries of this field the attention shifted to the whether a data scientist can provide value for a company.

Data analyst is perhaps the most common job title among the data professions. With a “catch-all” name, a data analyst is someone who cleans, transforms, models, analyzes or is involved in data. Usually the objective of a data analyst is extraction of useful information, through pattern discovery. Different from a data scientist, an analyst is often employed below a scientist however the titles are interchangeable depending on the company in question. Data analysts can be found in every domain where data is found, they are not anchored to information systems, statistics, economics or engineering, they can be successful in social sciences like psychology, hard sciences such as chemistry or biology and politics.

Data Mining is a field of occupations very alike data analysis, it deals with pattern discovery in data sets using statistical, machine learning and other quantitative methods (KDD, 2018). Most often it is under the computer science domain where it functions as a knowledge discovery in databases (KDD) engine. It covers more of machine learning and artificial intelligence than data analysis does and can be synonymous with Data Science (Bouckaert ,2010). Differences arise between Data Mining and data analysis when comparing the methods, mining has a more automatic or semi-automatic approach to dealing with data and is usually done in an unstructured environment while data analysis usually deals with more structured data and not as much automation within the methods.

Machine learning is the next field of occupations this study examines; it is classified under artificial intelligence and uses statistical methods to allow computers "learn" from data, through its pre-programmed patterns (Koza, 1996). It is the study and of algorithms which learn from data and can be used for prediction. The algorithms used in machine learning are not static formulas, they actively change based on data input, partitioning, set weights and activation functions (Nasrabadi ,2007). On the market knowledge of machine learning is a valuable skill to have since the number of jobs have increased almost ten-fold since 2012 (Bowlby, 2017) making it the fastest growing job in 2017.

Big Data analysis, development and engineering are included in data professions and play a key role on the market. Big Data professionals deal with "Big Data" problems for which are specific tools such as Hadoop, NoSQL, MongoDB and Apache packages just to name a few. Big Data is heavily emphasized in medicine, retail, financial services and transportation where the majority of Big Data professionals reside. What differentiates Big Data jobs from the rest in the data professions are the tools and the data environment. The tools are designed to handle the large amount of information, distribute the computing power and compile the results from various sources, the data environment is characterised by the “V”s previously mentioned, volume, velocity, etc…

Artificial intelligence (AI) is the study of “intelligent machines”, it also includes methods like machine learning. It is the process of “teaching” a computer to solve problems using input data. Artificial intelligence jobs are mainly in research and development (R&D), decision support systems and various automation fields. Little has been written about AI occupations and not all evidence is conclusive of their nature. Artificial intelligence can be a blanket term for the other occupations mentioned depending on the company.

Business intelligence, interchangeable with Business Analytics, is the process of data analysis with a focus on presentation of descriptive statistics and key performance indicators to make informed decisions. BI jobs center on analysts, who facilitate improved decision-making, optimize efficiency and report key business statistics.

Engineering in the context of this paper refers to data engineers, software engineers and is a catch all term for any job title, within the data professions, which contains the term “engineer” in its name. Engineering is the application of known methods to the creation, improvement, operation and maintenance of devices with the goal of creating value. Its main tenet is application rather than theory where it differs from the rest of the data professions. Where Data Science, analysis or machine learning have unstructured elements where exploration takes place, engineering is more static. Engineering jobs experienced less explosive growth over the last five years than machine learning or Data Science however, they are still in demand and still valued part of the team.

2.3 Problem of Supply

There is a shortage of qualified workers in the Big Data field (Boutlon, 2013, Baesens, Bapna, Marsden, Vanthienen, Zhao, 2014). Many organizations are demanding data scientists but are being met with none or little supply. Since the industry is young, there are no definitive experts. McKinsey stated there will be a shortage of 140,000 to 190,000 analysts and over 1.5 million managers by 2020 (Manyika, 2011). The projected need for 1.5 million managers in the United States is for people who can ask the proper questions and generate problem statements. The consumption of the results of the analysis of Big Data should be done by executives who have the power to influence company's decisions. Data generated every day is insurmountable when compared to the number of people who try to parse through it and identify key insights. The process takes time and companies have different requirements for what qualifies as a "data scientist", according to the organizations the concept of a data scientist is subjective which narrows down both parties' options for employment. This type of talent is difficult to produce, the time spent in school or practice will not aid anyone until the worker is employed and produces results (Clarke, 2016). Currently there are over 175 graduate programs in Data Science or the like (Big Data, Analytics) in the U.S. (NCSU, 2018), this amount of programs will not be able to produce enough analysts for the market to be anywhere close to satisfied. There are few undergraduate programs in Data Science or similar which means that organizations will not be able to find such talent sustainably. In order to succeed in 2018 employment market, students must have the ability to parse the vast amount of data, with emphasis on information collection

(Boyles, 2013). Entrepreneurs have stated that the low number new employees and skilled managers slows down the growth of their companies and thus the development of the market (Ewing Marion Kauffman Foundation, 2007). The increasing demand for a skilled and educated labor force has increased the importance of obtaining a university degree. Between 1970s and 2000s, the amount of workers with some postsecondary education increased by 110% and the amount of workers with bachelor's degrees increased by over 120% (Carnevale and Derochers 2002). Studies show that newly graduated students do not have the necessary information based skills at the levels that employers expect. Aguinis and Kraiger created a study that finds companies rank "Leadership", "Creativity" and "Critical Thinking" as the highest desirable skills (Aguinis, Kraiger, 2009). These skills are not being shown in the labour force. There is also a growing need to fill occupations other than Data Science still dealing with data such as Big Data specialists (Miller, 2014), artificial intelligence personnel or machine learning engineers.

2.4 Foundational Big Data Knowledge

When speaking about requirements, a distinction between capabilities or skills and foundational knowledge must be made. Foundational knowledge is the ground on which skills are built upon and must be known before any skill can be acquired. Foundational knowledge talked about in the literature focuses on programming knowledge, statistics, business calculus and data management (Jafar, Babb, Abdullat, 2016). The basic methods which these disciplines advocate are descriptive statistics, regression and ANOVA (Wixom, 2014). Big Data and Data Science courses are often found within computer science, engineering or information systems programs rather than as separate majors (Jafar, Babb, Abdullat, 2016; Lyon, Brenner 2015; Ransbotham, Kiron, Prentice, 2016). One aspect they have in common is the strong quantitative foundation required in each of the subjects. Other foundational knowledge includes database design, data modeling, and data normalization.

Dubey and Gunasekaran (2015) talked about education and training specifically in Big Data and Business Analytics. In their paper they interviewed multiple heads of Business Analytics companies and departments. According to the heads of Business Analytics the most important skills required for Business Analytics the profession are IT skills mixed with communication skills. A majority also indicated that operations research or a statistics are really important as well. They separate formal and informal education options where they make the distinction that an informal education is flexible and the learning pace can be adjusted by the learner whereas formal education is not as flexible.

Provost and Fawcet (2013) tried to understand what kinds of relationships does Data Science have with other important related concepts and identify fundamental principles underlying Data Science. One of the closest concepts relating to Data Science is Data Mining, extracting insights from data using technology. Another very important aspect is business knowledge and viewing business problems from a data point of view. All of this done with statistical and causal methods. Visualization is vital, intuition, creativity, common sense, and knowledge of software is required. Other fundamental concepts include (Provost and Fawcet, 2013):

1. "Extracting useful knowledge from data to solve business problems can be treated systematically by following a process with reasonably well-defined stages.",
2. "Evaluating data-science results requires careful consideration of the context in which they will be used.",
3. "The relationship between the business problem and the Analytics solution often can be decomposed into tractable sub problems via the framework of analyzing expected value."
4. "Information technology can be used to find informative data items from within a large body of data."
5. "Entities that are similar with respect to known features or attributes often are similar with respect to unknown features or attributes."
6. "If you look too hard at a set of data, you will find something—but it might not generalize beyond the data you're observing."

7. “To draw causal conclusions, one must pay very close attention to the presence of confounding factors, possibly unseen ones.”

2.5 Skills Required of Data Professionals

The skills needed to succeed in Big Data are not the same as the foundational knowledge one must learn in order to understand what it is. Many researchers have mentioned the skills which one needs to have as a Big Data analyst. Companies and employers value leadership, critical thinking and creativity among students (Finzer, 2013). These skills are helpful when designing a curriculum which adheres to 2018 industry requirements. The most commonly mentioned core competencies in literature include: analytical problem solving, innovation, initiative, flexibility, communication and collaboration skills. Examples of tasks performed by analysts are setting up servers for data, understanding of outliers particular data sets, generalizing and compiling results drawn from the data, visualization design suitable for easy data consumption of non-experts, data collection, clustering methods, regression analysis, principal component analysis, pattern recognition and systems set up (Finzer, 2013). Researchers such as Lyon (2016) talk about the communication and documentation skills for writing up reports as well as the ability to work collectively in teams. Big Data projects are continuously reliant on a team of analysts and moving away from a “full stack” analyst that can do everything.

A capability is defined as a work-related knowledge, skill or ability held by an individual (Nordhaug 1993). In broad terms, knowledge refers to the theoretical understanding of a concept (e.g. what the notion of a “business process” means), while skills relate to the practical application of that knowledge (e.g. how to model a business process). In contrast to knowledge and skills, which can be learned, abilities are attributes that are innate to an individual (e.g. a person’s ability to abstract). Although there is a consensus on the conceptualisation of competence, most empirical IS studies use the notions of work-related knowledge, skills and abilities interchangeably (Lee, Trauth, and Farwell 1995; Todd, McKeen, & Gallupe 1995). The same holds true for the vast majority of job ads. Big Data and Data Science jobs are rapidly filling up job posting boards, search engines, company web sites and recruiter lists. However, the criteria set for data jobs are often general in nature and are crafted with subjective views of organizational needs (De Mauro, 2018). Vague definitions of job roles such as "Data Scientist" discourage possibly qualified individuals from applying for the position. The lack of formally agreed upon requirements has been plaguing the Big Data field since its inception and was identified as one of its biggest research gaps (Miller, 2014). Research within the required Data Science capabilities has been limited to information systems (IS) jobs however, recent studies have looked at business intelligence (Debortoli et al., 2014), Big Data (Debortoli et al., 2014; Dubey and Gunasekaran, 2015), supply chain management (Schoenherr & Speier-Pero, 2015) and mobile analyst (Brauer & Wimmer, 2016) competencies to fill in the gaps. The required Data Science capabilities have been mentioned by Davenport and Patil (2012) however, they were not researched, rather they were anecdotal descriptions in specific contexts. Majority of literature is in agreement that Data Science is a multidisciplinary field drawing from other branches of knowledge like mathematics, statistics, computer science and information systems (Zawadzki, 2014; Schoenherr & Speier-Pero, 2015). Which is why the boundaries for Big Data/Data Science professionals are difficult to find. There is no rigorous definition of Big Data as well as its sub-fields in prior research (Mayer-Schönberger & Cukier, 2014). Davenport (2006) was one of the first to popularize the concept of a data analyst and described it by referring to CapitalOne's job description:

"High conceptual problem-solving and quantitative analytical aptitudes… Engineering, financial, consulting, and/or other analytical quantitative educational/work background. Ability to quickly learn how to use software applications. Experience with Excel models. Some graduate

work preferred but not required (e.g., MBA). Some experience with project management methodology, process improvement tools (Lean, Six Sigma), or Statistics preferred." (p.73)

The description encompasses a variety of educational and work backgrounds, with general skills of "problem-solving" and "quick learning". He goes on to include social skills such as communication of results with visual aids as an integral part of the job. Such a position is placed in between technical and business fields where an analyst speaks both languages. Attempts at setting or clarifying skill requirements for data professionals have been made by a limited number of studies. Lee, Trauth, and Farwell (1995) investigated the changes in the information systems (IS) professions to understand the impact they may have on future skills and knowledge of IS professionals. IS field having a lot of skill overlap with Data Science is included into the group of educational backgrounds suitable for data professionals to come from. Lee, Trauth, and Farwell tried reaching as broad a perspective as possible which lead them to include both industry and academic skill requirements into their study. Their results showed requirements becoming more demanding in the areas of business functionality and interpersonal management. As patterns in staffing choices changed over time Lee, Trauth, and Farwell found that IS organizations transformed their systems from traditional to decentralized or end-user-oriented. The organizations' goals were to align IS solutions with business goals which were the main motivations for hiring IS professionals. Most of the skills identified in literature are technical skills (Rao, 2014). In his paper Rao (2014) motioned to emphasize soft skills as important for effective performance. Although technical skills are important, soft skills like communication, leadership or passion allow a company to grow and are regarded as critical for building future skills. Debortoli et al. (2014) compared the required Big Data skillsets to business intelligence skillsets. In their analysis they referenced Big Data skills as an amalgam of computer science and statistics competencies. They break down the industry demands into topics extracted from groups of keywords found in job advertisements, separating the topics into “Business” and “IT” sub-categories and further separating each category into “Domain”, “Management”, “Concepts and Methods” and “Products”.

Schumann et al. (2016) has compiled a literature review of Data Science competences and categorized them into three sections: professional (ex: technical skills, extraction, transformation, analysis, and presentation of data), social (ex: communication) and personal (ex: team building, inquisitiveness and imagination). Zawadzki (2014) differentiates Data Science capabilities by four categories: math, computer science, domain knowledge and communication. Chen et al. (2012) described business intelligence and Analytics environment, evolution and key characteristics. Through market research, literature review and industry papers they break the discipline into three evolutionary stages: BI&A 1.0 with structured data in relational database management systems (RDBMS), BI&A 2.0 a web intelligence form of data, and BI&A 3.0 focusing on mobile and sensor sources. Each subsequent stage includes the previous one and builds on top of existing infrastructure while adding a new layer of challenges. Chen et al. (2012) also described key characteristics of the domain but did not focus on individual capabilities. Knowledge and skill topics were a call to action for higher education facilities to develop programs rather than an exhaustive list of which competencies to prioritize. The stated areas of education include accounting, finance, statistical analysis, recommender systems and machine learning among others. Murawski and Bick (2017) add enterprise business processing, decision making, data management and analytical tool skills to the list of competencies. Due to the ever expanding list of qualifications and abilities a Data Science professional needs it has become difficult to find individuals with the entire skillset. To solve this issue, researchers (Schumann et al., 2016; Waller & Fawcett, 2013; Zawadzki, 2014) proposed the formulation of Data Science teams composed of multiple data scientists; each with at least one strong background in a quantitative domain. De Mauro et al. (2018) attempted to clarify the nature of skills required in Big Data professions. They compiled a list of sources that partially described Data Science characteristics. They split the information into two groups: technology oriented and business focused. They scraped the web for job titles and grouped them based on words surrounding the job title. They came up with four Big Data job families. Using LDA they clustered most common skills for each job and created a table then merged the tables together which connected jobs with skills.

2.6 Software and Technologies

The most commonly used applications for Data Science are the language R, Python, Excel, SQL, and SAS. Using software to prepare and clean up the data to make it more readable, free of errors and allow for proper analysis is at the foundation of Data Science. Data comes from various sources and is usually very dirty, not labeled and full of missing information (Constantiou, Kallinikos, 2015; Newell, Marabelli, 2015). In order to interpret the results students must grapple with inconsistencies in the data and software limitations. Big Data software, such as NoSQL, Hadoop, Hbase, MongoDB, CouchDB, and others have been at the forefront of Big Data technologies. Big Data technologies are mainly used for data engineering, distribution of computing power or implementing data-mining techniques. Big Data software can also be used for data processing and data-analysis.

2.7 Data Science Education in Canada

In 2018 Data Science education in Canada is specialized to bachelor's programs, certificates, master's degrees, and concentrations in educational programs. Bachelor's programs are rare and only four such programs are available in Canada offered by University of Waterloo, University of British Columbia, University of Ontario Institute of Technology and University of Laurier. Certificates are offered by a larger number of schools than the bachelor; they are normally completed in less than one year, and are usually positioned as continuation of education. Master’s degrees are the focus of this study and are currently the most popular form of Data Science education in Canada. The programs’ durations range from ten months up to twenty-four months, there are comprehensive options and concentrations of existing degrees. There are programs which are offered online however they will not be discussed in this study; this study will focus on in-class courses and degrees only. As of August 2018 there are no PhD programs in Canada that deal with Data Science alone. There are Computer Sciences as well as Information Systems PhD programs with minor concentrations but no doctorate program has compiled a curriculum strictly for Data Science, Big Data or Analytics.

Graduate programs in universities are aimed at increasing the employability of their students or readying them for further education. From the industry perspective, graduate programs should educate the students to a point where they can enter the workforce. With the varying requirements and short lifespan of Data Science programs there are no conclusive results whether the teachings are useful or not. Davenport and Patil (2012) state that there are no university programs available in Data Science, this has changed in the past years and the market has seen a number of universities introduce Data Science courses. This leads into the problem of how to assess the specific curricula taught in those courses which assume the industry requirements. This issue has been looked at by Schoenherr and Speier-Pero (2015) who viewed IS curricula.

To further investigate Canadian education, the only programs chosen were the ones that emphasized “Data Science”, “Big Data” or “Analytics”, regardless of the department hosting the program. A criticism can be made here that “Business Analytics” programs are different from “Data Science” due to their focus on business acumen however both programs utilize data in the same ways, trying to accomplish similar goals and the results are used for the same reasons such as decision support, value addition or cost optimization. As of November 2018 there are 32 programs available in Canada. In most academic institutes one can find Data Science courses which are a collaboration of multiple disciplines. Through their collaboration the departments can draw on each other’s strengths with the added benefit of a larger faculty pool. In learning from different schools of thought the students would benefit by having a wider array of opportunities and experiences to develop as well as see problems from multiple perspectives.

2.8 Courses in Graduate Programs

This study investigates Canadian universities, graduate programs which are aimed at providing education about data, the program curricula as well as courses offered in the programs. Certain core courses have topics which overlap, this provides a semblance of evidence for setting specific courses as “core”. There are 95 universities in Canada (Univcan, 2018), a total of 13 universities are chosen with 218 courses between them. The information was found by viewing university websites, course description pages and program brochures. Extended information about courses, universities and their descriptions can be seen in the appendix.

2.9 Structure of data-related curricula

The academic environment has been criticised for ages for not imparting skills which can translate easily into the industry in every domain. Similarly, with information systems, IT, now Big Data and Data Science are criticised in the same fashion (Lee et al., 1995; Noll & Wilkins, 2002; Targett, 1991). These studies only considered IS jobs as a whole with no discrepancy between positions. IS is a broad field and emergence of Big Data as well as Data Science only furthers its boundaries which is why a more detailed view is necessary. Positions in new areas have been created, such as distributed computing, visualization, Big Data architecture and cyber security have widely different requirements, training and knowledge thus for a more accurate measure data collected for the analysis is divided into multiple occupations. With the popularity of Big Data and Data Science increasing many interesting parties, including academia, have begun research into the skillsets needed for data jobs. Pushed by the demand of the market, new high level educational programs came into existence. Majority of initial seminal research papers call for the development of graduate level curricula to be established (Dubey & Gunasekaran, 2015; Wixom et al., 2014; Wixom et al., 2011, Provost, 2013) which exemplifies how young the domain still is. Certain studies actively try to design or propose preliminary data-related curricula (Gupta et al., 2015; Mitri & Palocsay, 2015). Articles are written by journalists, data practitioners, industry analysts, however too many are written about the explosion of "lucrative" data jobs while overestimating demand by using simple search results with no sound underlying methodology. Through constraining of search queries and thorough data processing, the data is cleaner, clearer and provides a more accurate assessment of the data job market.

Journal articles such as Turel and Kapoor (2016) have conducted an analysis of 124 US business schools targeting "Business Analytics" courses to understand their structure and offerings. The courses were categorized into four types: data analysis, Business Analytics, data warehousing and database. The authors also controlled for school prestige and their results show that business schools are not acceptable authorities in terms of Business Analytics to certify students to be industry ready. The issues the previous papers contain are the preconceived notions and assumptions of what skills and capabilities data professionals should possess. They did not base their claims on tangible data and did not define any required capabilities. Schoenherr and Speier-

Pero (2015) were one of the first ones to include data job requirements within an industry sector with university training content for said sector. In the supply chain management industry Schoenherr and Speier-Pero (2015) gathered survey results and conducted interviews as the benchmark for industry requirements after which they analyse post-secondary graduate programs to determine the evolution of business education.

2.10 Inadequacies of Current Education

One of the biggest themes among current literature in Big Data education seems to be the inadequate preparation of students by their educational institutions (Carillo 2017, Bullen, Abraham, Galup 2007). College students seem to lack real world experience (BIC3 employer survey respondent). Employers who look for new graduates must retrain them in order to get them ready for work what should have been done mostly by the university. Common complaints from employers are: "Universities can better prepare their students by providing opportunities for them to work in simulated industry projects instead of just teaching programming and database skills" (BIC3 employer survey respondent) (Wixom, B. et al, 2014). Employers identified that internships, report development and practical experience are the most significant whilst searching for an employee (Wixom, B. et al, 2014). To be more specific, Wixom's et al. (2011) panel report notes industry trends raise concerns that education may be lagging behind the curve of progress. The delivery of effective business intelligence programs is not up to par to the employers' standards. Based on surveys conducted at BI practitioner events, Wixom et al. (2011) formulate four academic BI best practices that would close the gap between BI market needs and the content of IS education programs: (1) provide a broader range of BI skills, (2) take an interdisciplinary approach to BI programs, (3) develop reusable teaching resources, and (4) align with practice. Besides arguing for the need for technical skills, Wixom et al. (2011) argue that a deep understanding of business subjects (e.g., finance, marketing) and strong communication skills are required. Universities have not taken the time to recognise that future managers must also prepare for the new data-driven era (Chatfield, Shlemoon, Redublado, Rahman, 2014). With the growing need for data scientists there is a need for managers who know how to handle, assign and delegate data scientists. This need is not being satisfied by current university courses and programs (Carillo 2017). The low enrollment of IT-related fields has been increasing year to year due to the Big Data popularity however the universities have not adapted to the shifting tastes (Bullen, Abraham, Galup, 2007). Even with the low enrollment, graduates are finding jobs and universities are taking the credit. This allows universities to not change their practices and focus on their current successes while not adapting to the changing market needs. The need to connect academia and apply a disciplinary pedagogical model for business education is vital.

Chapter 3 – Theoretical Framework

3.1 Typological Structure

Typologies are classification methods used by academic fields to create structure for the phenomena under study (Croft, 2002). They are often linguistic in nature, most commonly used in psychology, anthropology, linguistics, statistics and urban planning. Typologies are fundamentally cross linguistic and analyze language to discover any underlying patterns or structures. They are meant to be functional approaches to the examination of variations in language for the purpose of explaining the universal structure of a given text. Typologies are different from standard classifications or taxonomies by not being generative and referring to the derived finding as “interrelated sets of ideal types” (Doty & Glick, p. 232, 1994). The purpose of using and creating a typology in this study is to identify exemplary cases of required capabilities in Data Science professions beyond simple definitions. Typologies are meant to be effective in modeling the relationships between variables to a point where they are prescriptive, which is the reason for their use in this paper. The results of this study are meant to develop a typology for data professional capabilities.

3.2 Todd et al. Competence Framework

Todd et al. (1995) have created a framework for information systems job skills which became a standard in IS. They studied the evolution of IS job competences through newspaper advertisements over the span of 20 years. Through the extracted competences they defined duties for systems analysts and IS managers and categorized them into three classes of knowledge. The classes described technical knowledge, business knowledge and systems knowledge, they were further divided into sub-categories of hardware, software for technical, business, management and social for business and problem solving, development for systems. Their descriptions are displayed in the table below.

The "ACM categories" are categorizations made by the Association for Computing Machinery in the curriculum model made by Nunamaker (1982).

Table 1. Todd et al. (1995) Typology Framework

Category	**Sub-Category**	**Description**	**ACM Categories**
Technical Knowledge	1. Hardware	Servers and personal computers. Other devices such as storage devices, controllers, printers, and networks.	Computers
	2. Software	Application systems, operating systems, packaged products, networking software, and programming languages.	
Business Knowledge	3. Business	Functional expertise (finance) and industry expertise (retail, mining).	Organizations
	4. Management	General management skills including leadership, project management, planning, controlling, training, and organisation.	
	5. Social	Interpersonal skills, communication skills, personal motivation, and ability to work independently.	Society
Systems Knowledge	6. Problem Solving	Creative solutions, quantitative skills, analytical modelling, logical capabilities, deductive/inductive reasoning, innovation.	Models
	7. Development	Knowledge of systems development methodologies, systems approach, implementation issues, operations and maintenance issues, general development phases, documentation, and analysis/design tools and techniques.	Systems

Other studies have tried to set up new classification schemes and competence dimensions for IS jobs such as Litecky et al. (2009). Previous iterations include a taxonomy of job advertisements on the website Monster.com (Prabhakar, Litecky, & Arnett, 2005), a taxonomy of IS jobs by Trauth et al. (1993) and another newspaper advertisement analysis by Athey and Plotnicki (1988). Todd et al. typology is chosen from among all the others because it provides a broad enough coverage of technical and business aspects to cover both management programs, science programs and the industry. Their classification method is a better fit for this study than all the rest. This approach has been tested by Muller et al. (2014) and Murawski (2017) with great success in varied situations such as business process management. Lastly, this typology allows for easy comparison of job skills and academic teachings.

3.3 Framework for Data Competencies

The purpose of this section is to review the existing framework while highlighting some of the pitfalls and providing recommendations of how to improve it to be more relevant to the modern conditions of data industries. The new framework keeps all the currently set criteria and expands on some categories by including and defining new ones. The additions to the existing typology are introduced due to a lack of fit between prior categories and currently existing needs. The previous framework dealt with IS jobs such as "Programmers" and "System Analysts", due to the emergence of a broader range of jobs, titles, skills required and industries the previous typology lacked certain necessary classes. Todd et al. framework is based upon the Association for Computing Machinery (ACM) 1964 Computing Classification System (CCS) categorized by Nunamaker (1982), the new framework also draws from the categories as well as builds on the previous framework. The ACM CCS is a taxonomy used for classifying computing subjects by structure, goals and scope (Coulter, 1997), it is used to categorize university courses, research journal subjects and more. The proposed framework draws sub-categories from the ACM 2012

CCS in order to be consistent with previous successful literature as well as due to its reliability and robustness.

The “Systems Knowledge” sub-category “Problem Solving” is removed and placed into a newly formed category of “Method Knowledge”. Problem solving as a skill is important in all aspects of work and management however, leaving it in “Systems Knowledge” implies that problem solving is an ordered or systematic ability. Problem solving is necessary in situations where there is a lack of structure and limited information. As a sub-category it is also very broad and a “catch-all” term for any knowledge which requires taking advantage of new information. By including “Problem Solving” into the “Methods Knowledge” class it stands to reason there are specific methods which if used may provide an optimal solution backed by previous results of similar events. A new sub-class added to the “Method Knowledge” is “Mathematics and Statistics”. Math and statistics are integral parts of Data Science, Machine Learning and Big Data knowledge which need to be included. The “Problem solving” sub-class already contains “quantitative skills” and “analytical modeling” in its definition, a case can be made that mathematics and statistics already fit into “Problem Solving” however; the sub-category covers multiple skillsets with various requirements and therefore does not describe all of them properly. By adding more narrow sub-categories the data occupations, educational programs and industry skills can be categorized with more accuracy and efficiency. The educational courses of graduate programs contain an abundance of mathematic and statistical subjects all of which would be melded together if they were to be in the sub-category of “Problem Solving”. It is thus important to distinguish mathematical and statistic skills from general problem solving skills.

The second “Method Knowledge” sub-category added, “Artificial Intelligence”, contains skills and methods needed to tackle specific industry problems which differ from, as an example, mathematical problems. Occupations studied in this paper require the knowledge of artificial intelligence and machine learning; some graduate courses teach specific artificial intelligence subjects which cannot be fit into “Problem Solving” or “Mathematics and Statistics” sub-categories. Thus the new sub-category of “Artificial Intelligence” is necessary to include due to increasing explanatory and categorization power of the typology by being more accurate when differentiating between machine learning skills and other skill types.

The new framework is used to classify job postings as well as educational courses; with the newly added sub-categories it will improve classification of both lists. Categorizing both the academia and the industry listings using the same rubric will aid in the comparison of the required skills and imparted knowledge.

Table 2. Data Skills and Capabilities Classification

Category	**Sub-Category**	**Description**
Technical Knowledge	1. Hardware	Servers and personal computers. Other devices such as storage devices, controllers, printers, and networks.
	2. Software	Application systems, operating systems, packaged products, networking software, and programming languages.
Business Knowledge	3. Domain	Functional expertise (finance) and industry expertise (retail, mining).
	4. Management	General management skills including leadership, project management, planning, controlling, training, and organisation.
	5. Social	Interpersonal skills, communication skills, personal motivation, and ability to work independently.
Systems Knowledge	6. Development	Knowledge of systems development methodologies, systems approach, implementation issues, operations and maintenance issues, general development phases, documentation, and analysis/design tools and techniques.
	7. Database management	Physical design, languages, database administration, applications, database machines, storage and retrieval
Method Knowledge	8. Mathematics and Statistics	Numerical analysis, linear algebra, optimization, non-linear equations, differential equations, integral equations, discrete mathematics and probability
	9. Problem Solving	Creative solutions, quantitative skills, analytical modelling, logical capabilities, deductive/inductive reasoning, innovation.
	10. Artificial intelligence	Automatic programming, machine learning, natural language processing, robotics, distributed A.I, control methods

Chapter 4 – Data and Methodology

4.1 Job Data Collection

In order to answer the first research question; "What are the required competencies of data scientists in Canada?" job advertisements we chosen as a representation of market needs. Job advertisements are the main method of hiring and finding new employees (Walsh et al., 1975). In order to obtain the largest amount of data, in as little time as possible, in a reliable fashion a web scraping tool was compiled and used. A computer program or a web crawler can quickly parse, save and transform webpages into clear data (Kobayashi & Takeda, 2000). The process of web scraping used for this study entails searching for specific data elements in unstructured job descriptions within quasi-structured web pages. The web scraper was custom built by the author in Python. In order to understand and quantify what the industry demands of data professionals job advertisements are collected and analyzed from the Canadian job search engine Indeed (www.indeed.ca). The entire population of ads on www.indeed.ca were scraped between August 1st and August 31st. Data Extraction, storage and transformation of said data is done for online job posts relating to Big Data. The indicators the web scraper search for are the job titles of occupations and the return results are the descriptions. Data collection for "Data Science" jobs relies on the key terms "data sci" to reside in the job title. Terms "data sci" are chosen because they include the word "data" and the beginning letters of "science". This specific combination of words is chosen because it incorporates job titles that include "junior" and "senior" levels, consultants, instructors and others while removing any niche specificity the title might call for such as location or "Intern – Analyst, Business Insights and Analytics (4 or 8-month contract)". The "data" keyword stops unwanted science positions appearing in the results such as "Geospatial Scientist" and the "sci" keyword stops all other data positions from appearing in the results such as “Engineer”, “Data Mining analyst”, etc… The same process is repeated for “Data Mining” positions. The keywords used are “data mini” where “mini” represents “mining”. Data analysis jobs were selected on the basis of “data analy” keywords. “Data” to gather data jobs and “analy” for discrimination against different positions such as “scientist” or “engineer”. Machine learning job class was set by finding jobs with keywords “machine learning”, “machine” and “ML” in the titles. “Big Data” jobs were selected based on keywords “Big Data” within the job title. Artificial intelligence jobs were chosen based on “artificial intelligence” and “AI” keywords. “Intelligence” was also considered however, once tested, jobs with “Cognitive Intelligence” keywords appeared as well as “Business intelligence”. Business intelligence jobs were chosen by their use of “BI” and “Business intelligence” keywords in the titles. Variations of “BI” were also used such as “BI ” with a space before and after the keyword for better clarity. Engineering job class was selected by applying the keyword “engineer” to job titles. Each class is not mutually exclusive, which means that within one job title there may be multiple titles (ex: Senior and Junior Applied AI/Data Science/Data Mining Developers). Job titles which include multiples are used in each sample as long as they include the name of the sampled class. A total of 24,153 job descriptions were scraped.

4.2 Education Data Collection

The educational data gathered for this study centers around graduate programs focussed on Big Data, Data Science and Analytics. Graduate studies were chosen because they are more representative of the labour market. Undergraduate programs are not common within the data domain and thus would not provide a large enough sample. Similarly, for PhD programs the scarcity of data programs would result in a non-representative experiment. The search has been conducted through the exploration of Canadian universities’ websites, course overviews and syllabi. Graduate programs containing the terms “Big Data”, “Data Science” and “Analytics” in their name were chosen for further analysis. These programs are scrutinized for their course offerings and subjects taught. The analysis is meant to obtain the results of what courses are being taught within those programs, what are the methods taught in said courses and what software.

4.3 Data Cleansing

Data, once obtained, is manipulated and cleaned with RStudio (R). Firstly, all the duplicate and non-English job postings were removed, punctuation and other special characters were also removed from the text. Given names, empty spaces and incorrect job postings (ex. “Sorry, this position has been filled”) have also been removed. Changed job titles and job posting text into lower case for easier cleaning provided greater clarity for further manipulation. Another removal of non-English job postings as well as duplicate postings via job description. Removal of special characters which do not belong to ASCII (American Standard Code for Information Interchange) aided in removing the lesser known punctuation.

Substitutions to certain keywords were made to make sure they will not be removed by a “stop word” dictionary, these include “r” to “rprogram”, “c++” to “cplusplus”, concatenating various bi-grams into single words like “communication skills” into “communicationskills” and “computer science” into “computerscience”. Removal of numbers and a stop word dictionary provided by R package “tm” (removeWords) cleaned the text of common words such as “I”, “this” and the like. A stop-word dictionary is created to remove unnecessary stop-words alongside a stop word dictionary included in the “tm” R package.

4.4 Classification

In order to achieve the most holistic view of the industry the eight occupations in question have been scraped for through the use of keywords creating job categories: Data Science, Data Mining, Data analysis, Machine learning, Big Data, Artificial Intelligence, Business Intelligence, Data Engineering.

In order to gather data specific enough to analyze and broad enough to generalise the field of Data Science, jobs with “data sci” in their title represent the “Data Science” category. Keywords “data sci” are chosen because they include the word “data” and the beginning letters of “science”. This choice was made due to the specific combination of words which incorporate job titles that include “junior” and “senior” levels, consultants, instructors and others while removing any niche specificity the title might call for such as location or “Intern- Data Scientist/Analyst, Business Insights and Analytics (4 or 8-month contract)”. The “data” keyword stops unwanted science positions appearing in the results such as “Geospatial Scientist” and the “sci” keyword stops all other data positions from appearing in the results such as “data engineer”, “Data Mining analyst”, etc…

The same is repeated for “Data Mining” positions. The keywords used are “data mini” where “mini” represents “mining”. Data analysis jobs were selected on the basis of “data analy” keywords. “Data” to gather data jobs and “analy” for discrimination against different positions such as “scientist” or “engineer”. Machine learning job class was set by finding jobs with keywords “machine learning”, “machine” and “ML” in the titles. Big Data jobs were selected based on keywords “Big Data” within the job title. Artificial intelligence jobs were chosen based on “artificial intelligence” and “AI” keywords. “Intelligence” was also considered however, once tested, jobs with “Cognitive Intelligence” keywords appeared as well as “Business intelligence”.

Business intelligence jobs were chosen by their use of “BI” and “Business intelligence” keywords in the titles. Variations of “BI” were also used such as “BI ” with a space before and after the keyword for better clarity. Data engineering job class was selected by applying the keyword “data eng” to job titles. Each class is NOT mutually exclusive, which means that within one job title there may be multiple. Each job can belong to multiple job classes (ex: data science and machine learning). Job titles which include multiples are used in each sample as long as they include the name of the sampled class. The full list of searched keywords and the corresponding categories can be seen in Table 3.

Table 3. Searched Keywords in Job Titles

Job Category	Keywords searched
Data Science	"data sci"
Data Mining	"data mini"
Data analysis	"data analy"
Machine learning	"machine learning", "ml"
Big Data	"big data"
Artificial Intelligence	"artificial intelligence", "ai"
Business Intelligence	"business intelligence", "bi"
Data Engineer	"data eng"

4.5 Term Frequency – Inverse Document Frequency (TF-IDF)

Term frequency–inverse document frequency (TF-IDF) is a natural language processing tool used for retrieval of information. Its statistical scores reflect how important a word is to a document in a collection or corpus (Ullman, 2011). The matrix of term frequencies it creates results in values which give each term a weight. The value increases based on the number of times the specific term appears in a document, it is then divided by the number of documents in the corpus that contain that term. This process allows for the adjustment of terms which appear more frequently and don’t hold much value. TF-IDF is a tool used in ranking a document's relevance. Utilizing the matrix output the words with lowest scores are added into the stop-word dictionary which is used to discard words without value, terms such as “data” are discarded because they do not hold value.

4.6 Latent Derilict Alocation (LDA)

Latent Dirichlet allocation (LDA) (Blei, 2012) is a popular topic modeling method that explains observations using unobserved groups. Handling each document as a mixture of topics, it extracts a specified amount and outputs scores relevant for each term per each topic. In the context of this study LDA is used to model job classes mentioned previously: Data Science, Data Mining, Data Analysis, Machine Learning, Big Data, Artificial Intelligence, Business Intelligence and Engineering. Due to every document being a mixture of topics each document may contain words from multiple overlapping topics. The topics will have different proportions and the ones with highest scores are set as dominant. It’s important to mention that different skills can be required for the same job by different companies, thus ordinary clustering methods like k-means are not the best choice to represent multifaceted sets of capability requirements. Mixed membership models are more suited to this situation and their assumption that a term may belong to multiple clusters is a benefit that other methods do not have (Airoldi, Blei, Fienberg, & Xing, 2008). LDA has proven to be effective in analyzing job postings as seen in De Mauro (2018), Murawski (2017) and Muller (2016). The terms LDA displays in its results are connected by a theme which can be used to describe job tasks. The results are the requirements per job class and are later matched with the content analysis of data courses.

4.7 Perplexity

The LDA inputs are the job descriptions and the number of topics needed. To select an optimal number of topics perplexity is evaluated. A simulation of multiple LDA outputs with k ranging from 2 to 30 has been run on both the job descriptions and course offerings from which the best outcome was selected based on the lowest perplexity and log likelihood score. Selection was done by sub setting the dataset into training and testing with a 75% to 25% split. The test set contains test documents W_d, the model is defined by the matrix of topics and the parameter α for topic distribution. The formula for log-likelihood evaluation is:

$$L(\boldsymbol{w})=\log p(\boldsymbol{w}|\boldsymbol{\Phi},\alpha)=\sum d \log p(\boldsymbol{w}d|\boldsymbol{\Phi},\alpha).$$

The likelihood is used to compare between models, with higher likelihood denoting a better model. Perplexity is the preferred evaluation measure for topic modelling; it measures the degree to which a model predicts a sample. Perplexity is a decreasing function of the log-likelihood $L(\boldsymbol{w})$ of the test documents W_d; the lower the perplexity, the better the model.

$$\text{perplexity(test set W)} = \exp(-L(\boldsymbol{w}) \text{ / count of tokens})$$

Chapter 5 – Results and Discussion

5.1 Descriptive Statistics

From a collection of 24,153 gathered job postings, after data cleaning, frequency analysis and descriptive statistics are calculated and there are 11,640 observations left. Table 4 summarises the number of jobs per job class for further analysis.

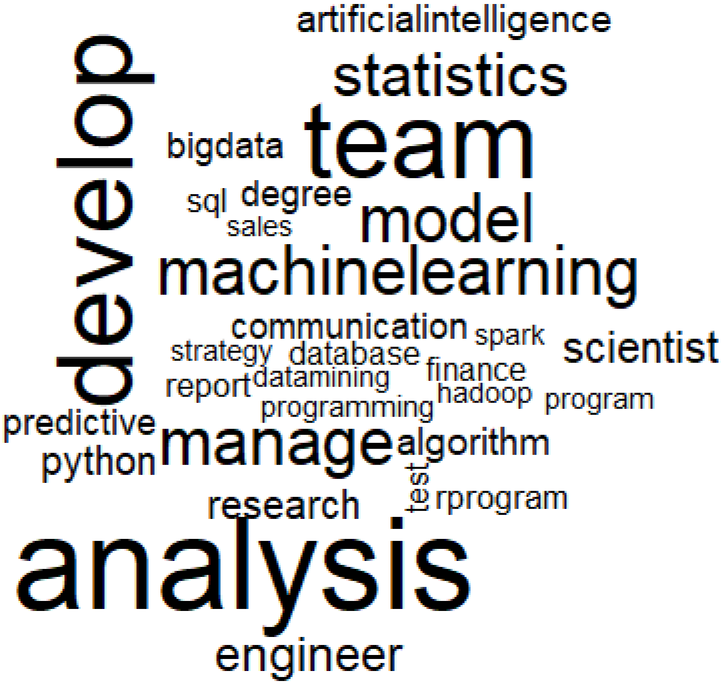
Table 4. Frequency of Job Postings based on Job Class

Jobs Class	Count of Postings
Data Science	528
Data Mining	16
Data Analysis	2104
Machine Learning	2065
Big Data	3386
Artificial Intelligence	1386
Business Intelligence	2065
Data Engineering	346

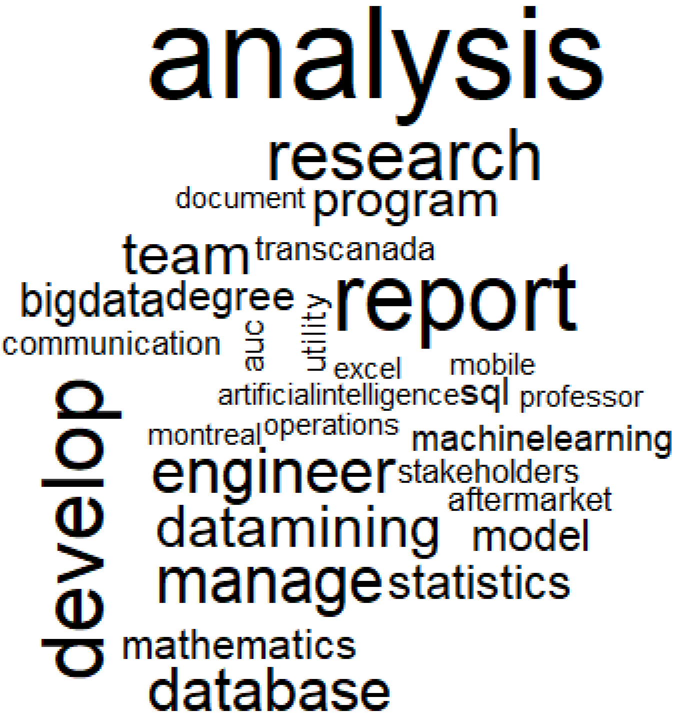
The education data shows that as of November 2018 there are 35 programs in Canada which relate to Big Data, Data Science or Analytics however, only 18 of them were selected since their program offerings included terms “Big Data”, “Data Science” or “Analytics” in their titles. The 18 programs selected are a population of the Canadian data education and represent all the available master level degrees in the topics mentioned above, within constraints. The rejected ones are graduate programs in computer science, mathematics, information studies and certificates which may have overlapping material however they do not advertise themselves as part of the Big Data program list. The programs are housed in 13 universities located in various parts of Canada. Table 5 lists the programs researched and the list of universities offering the programs, respectively.

Figures 1-8 show word clouds with the top 30 most frequent terms recurring in the job classes.

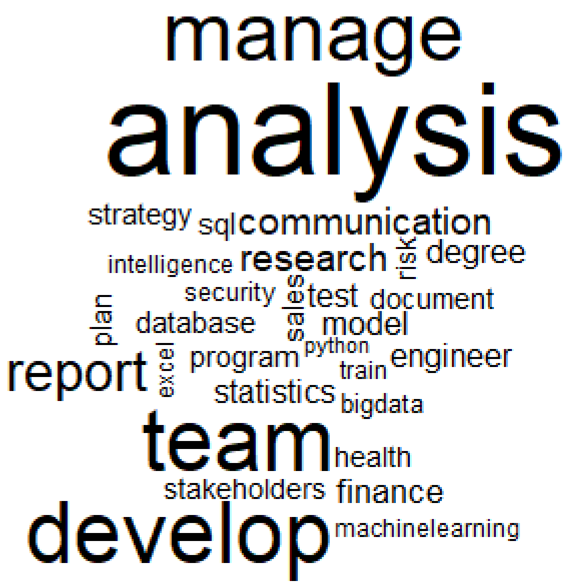
Data Science



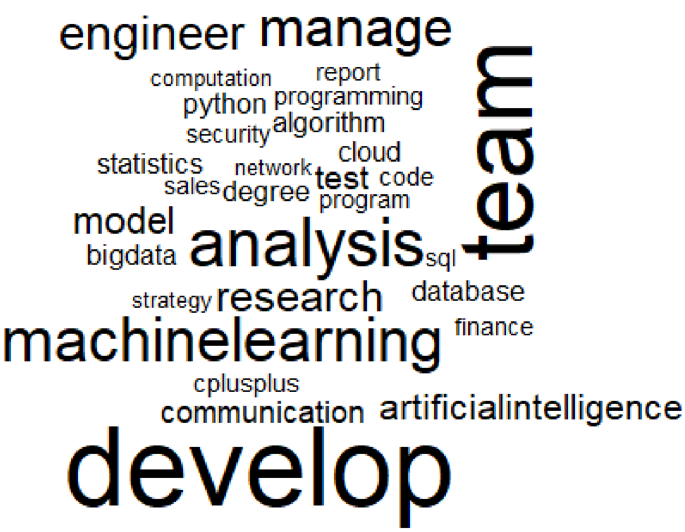
Data Mining



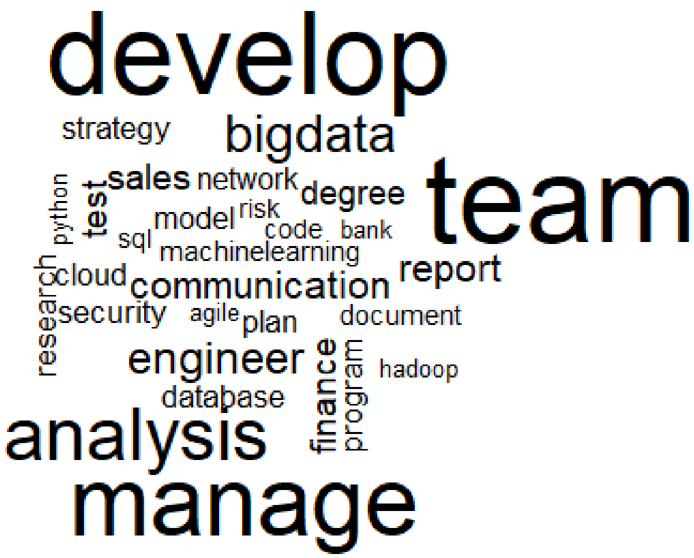
Data Analysis



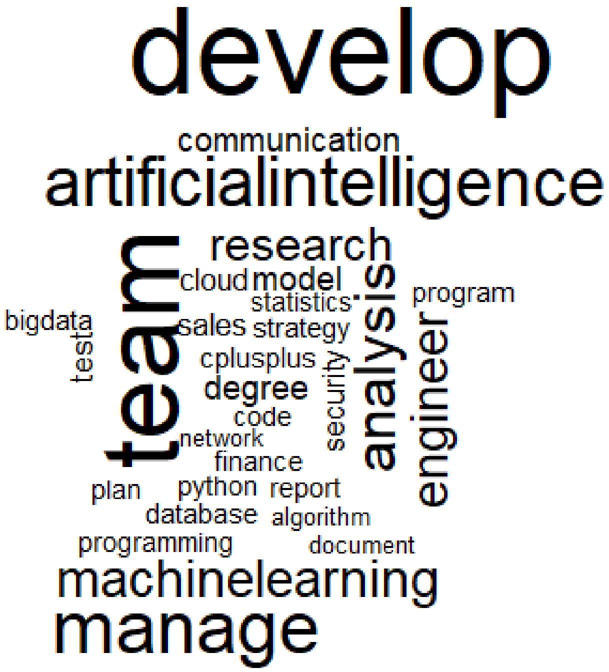
Machine Learning



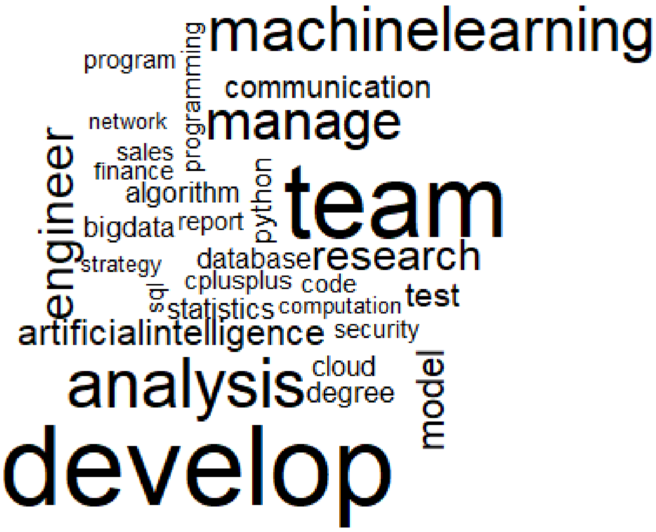
Big Data



Artificial Intelligence



Business Intelligence



Data Engineering

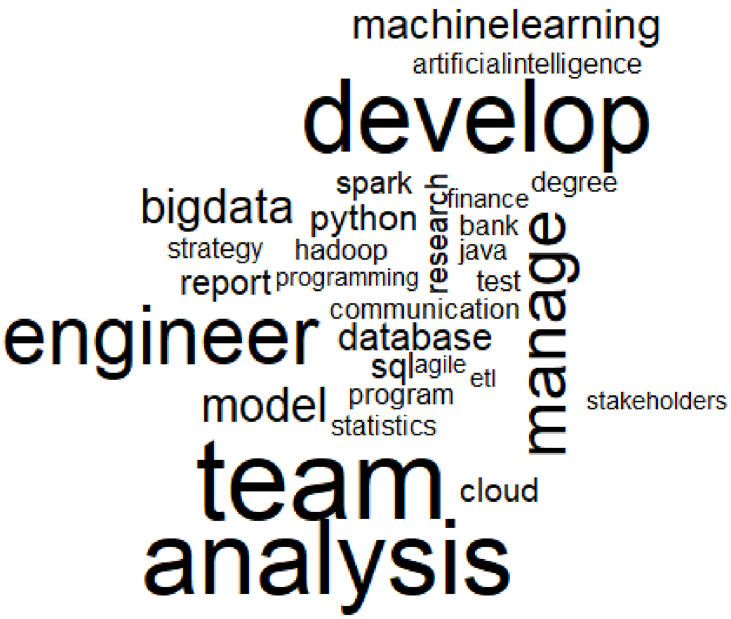


Table 5. List of Data Programs

Program	University
Master of Data Science - Vancouver	University of British Columbia Vancouver
Master of Data Science - Okanagan	University of British Columbia Okanagan
Master of Data Science - Computational Linguistics	University of British Columbia
Master of Business Analytics	University of British Columbia
M.Sc. in Computing & Data Analytics	Saint Mary's University
Master's in Big Data	Simon Fraser University
M.Sc. in Applied Computing - Data Science concentration	University of Toronto
Masters of Management in Analytics	University of Toronto
Masters of Management in Analytics	Queen's University
Masters of Business Analytics	York University
M.Sc. Data Science and Business Analytics	HEC Montreal
M.Sc. management: business analytics	University of Western Ontario
Masters of Data Analytics	University of Western Ontario
Masters of Management in Analytics	McGill University
Diploma in Data Science and Analytics	University of Calgary
Big Data Analytics M.Sc. Applied Modelling and Quantitative Methods	Trent University
Masters of Mathematics Statistics - Data Science Specialization	University of Waterloo
M.Sc. in Data Science and Analytics	Ryerson University

Table 6. Programs by Education Type

Program Type	Frequency
Big Data	2
Data Science	8
Analytics	8

Using the n-gram method (bi-grams), a list is compiled of every bi-gram which contained the key word "year". Using "year" as the denominator to gather number of years of experience required per posting results show what the market requires from applicants in terms of number of years of experience. Ex: 5_years, 8_years, 1_year, three_years, threefive_years, etc… Not all postings mentioned years of experience or education as a requirement. The bi-grams are then grouped into a table with year ranges from 1-3, 3-5, 5-7, 7-10 and 10+. The groupings chosen are standard in job descriptions and generalize the level of experience of an applicant well.

Table 7. Bi-grams of demanded experience

Bi-gram	Frequency
1_year	85
1_years	4
10_years	55
12_years	21
15_years	28
2_years	300
3_year	6
3_years	343
35_years	27
3rd_year	4
4_years	78
4year	11
4year_science	11
5_years	290
6_years	17
7_years	43
710_years	12
8_years	73
9_years	15
eleven_year	8
experience9_years	9
field4_years	14
four_year	1
four_years	8
one_year	10
qualifications3_years	4
qualifications5_years	10
roles1_year	14
six_years	14
skills1_year	19
skills23_years	6
sql2_years	10
ten_years	11
three_years	32
threefive_years	11
tools_4year	11

topics2_years	3
two_years	11

Table 8. Bi-gram Experience Summary

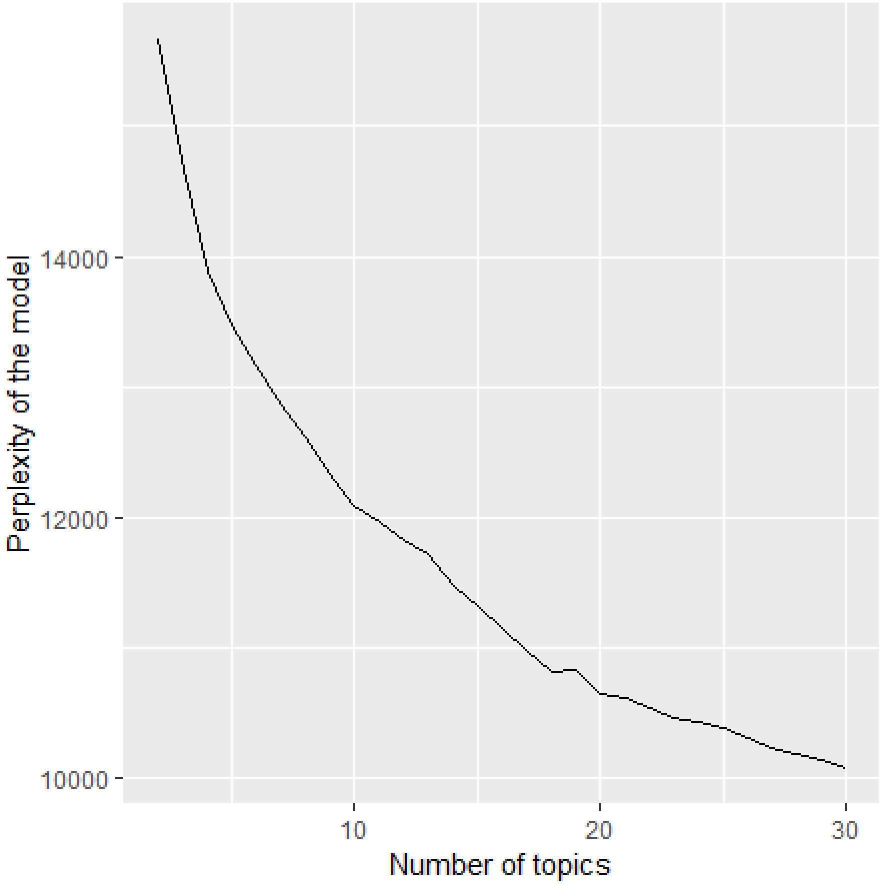
Experience Bracket	Frequency
3 years	343
2 years	300
5 years	290
1 year	85

5.2 Results

5.2.1 LDA Results and Discussion

LDA requires a set number of topics prior to running the analysis, through the application of LDA, every topic is a grouping of terms taken from the corpus with a distribution that infers a degree of membership. The groupings of top terms with highest "beta" are latent variables known as topics. The number of topics is selected prior to running the algorithm and choosing the most optimal number of topics is done through selecting the lowest perplexity score. Testing various numbers of topics from 2 to 30 the perplexity score kept decreasing as the number of topics increased. Ultimately k = 10 topics were chosen due to a large perplexity score difference from k = 9, a small difference between k = 11 and human evaluation. Figure 9 displays the perplexity score compared to the number of topics for the job analysis. With an increasing number of topics meaning is lost and the latent variables become much more difficult to label. A lower number of topics would result in broader and more general descriptions while a higher number would be too specific and would have difficulties finding valuable terms. Table 9 shows the results of the LDA with latent topics labeled manually by the researcher. It shows the top 15 terms most connected with the topic. The 10 topics best outline the skill groupings represented in job advertisements. Each grouping, labeled at the top describes a subject or field of expertise for data professionals.

Figure 9. Job LDA Topics vs. Perplexity



5.2.2 Combining LDA with new Framework

This section provides a combination of the LDA results and the new data skill classification. Each skill group as well as their requirements are assessed using the same dimensions allowing for easier comparison. Table 10 displays the LDA modeled topics along with their terms in the new framework for easier comparison. Table 11 shows the count of the terms for each topic and which category the terms belong to. Labeled topics are based on the proportion of terms within specific sub-categories. Figure 10 is a stacked column chart which visualizes how each topic is constructed based on percentage of terms in each sub-category. The resulting topics contain similar terms between each other which indicate a skill overlap between data occupations. Each topic contains 15 terms and each term is assigned to one category in the skills and capabilities framework, the collection of terms is labeled manually as the topic which meaningfully connects most of terms inside the grouping.

The first topic “Machine Learning” obtains its name from the heavy weight of “Artificial Intelligence” sub-category where 3 terms (20%) correspond with a machine learning oriented job description. This topic had the largest amount of terms corresponding with the “Artificial Intelligence” sub-category compared to the rest of the topics. The “Software” sub-category with 4 terms (27%) also fits under the topic’s title. Due to a high number of job postings calling for machine learning professionals it is unsurprising to find such a topic.

Topic two was called “Medical Research” due to its heavy weights of the “Domain” sub-category, with specific terms relating to health and medicine, 5 terms (33%) associated with the healthcare industry as well as heavy weights for “Management” and “Problem Solving” sub-categories. Research in healthcare appears to be a large portion of job postings which symbolizes the demand for such skills. Domains like medicine and finance which appear in the results signify the market heading in a specific direction which should be followed by education creating a fit.

Topic three named “Financial analysis” consists of the “finance”-like terms in the “Domain” sub-category, high number of reporting terms in the “Management” sub-category and the “Problem Solving” terms. Topic 4 “Management” was named due to the high amount of terms associated with the sub-category “Management”. The second of domain skills financial analysis require special designations and education, although it is not a data profession in itself financial knowledge as well as analysis of financial data is necessary for a large part of data professionals. This information allows individuals from accounting and finance professions to transition into data jobs more easily.

Table 9. Job LDA Results

Machine Learning	Medical Research	Financial Analysis	Management	Statistics and Math	Database Management	Security Operation	Testing	Software Engineering	Sales Forecasting
team	research	finance	manage	analysis	develop	security	test	develop	sales
develop	health	manage	analysis	develop	team	network	analysis	team	team
machinelearning	develop	team	develop	statistics	engineer	manage	develop	engineer	manage
research	analysis	report	team	model	cloud	analysis	manage	cloud	develop
artificialintelligence	manage	bank	test	manage	database	team	document	mobile	strategy
engineer	clinical	analysis	report	communication	test	develop	engineer	test	analysis
algorithm	team	develop	plan	report	sql	test	team	code	forecasting
analysis	report	excel	document	sql	analysis	report	research	javascript	communication
mobile	communication	communication	communication	train	code	document	security	manage	plan
python	medical	sales	strategy	excel	java	communication	plan	agile	retail
cplusplus	document	plan	stakeholders	strategy	manage	train	report	cplusplus	bigdata
model	healthcare	forecasting	sap	team	agile	engineer	communication	python	excel
rprogram	train	document	train	machinelearning	cplusplus	plan	train	communication	report
test	statistics	train	risk	risk	python	linux	stakeholders	hadoop	train
deeplearning	pharmaceutical	word	consulting	lift	hadoop	cloud	sap	html	cloud

The topic “Statistics and Math” obtained its name due to the “statistics” key words in the “Statistics and Math” sub-category along with the large amount of terms in “Software” and “Artificial Intelligence”. The standard mathematical knowledge requirement is not surprising in an industry which deals with finance, supply chain, artificial intelligence, prediction and reporting, this skillset is widely believed to be constant in the data science field.

“Database Management” is the name of the topic which focuses on the storage of data, based on the large percentage of terms in the “Software” sub-category which relate to databases as well as the largest amount of terms in the “Database Management” sub-category. Storage of information is necessary to the lifecycle of data thus the knowledge of operating the system which houses all the information is required of data professions. Some professions will require solely these types of skills such as Database Administrators, Systems Analysts or Data Infrastructure Engineers but majority of the professions require the individual to be able to use technology like SQL for data management to an extent.

“Security Operation” topic was named due to the “Development” sub-category being dominant as well as “Management”. The term “security” was also the highest rated term within the grouping. Topic 8 “Testing” was named after the top term in the grouping “testing” as well as the high percentage of “Problem Solving” terms. Security of data similar to privacy is an important subject and many papers have researched these precise topics and it is not unexpected to see a semblance in this analysis. The security operation skillset is a mindset which must be carried into every aspect while working with data, since the beginning of designing the infrastructure, while gathering the data and towards the end when reporting or implementing the insights. Security operation is about the safety of data and its sources, it is the creation of a system where no breaches can be exploited and the rules which govern the access of data.

"Software Engineering" is the 9th topic and is named due to its high amount of "Software" sub-category terms and the "engineering" term from "Problem Solving". Work with software was expected to appear in the results since most of data professions deal with one type of software or another and knowledge of how to use it is quite vital. Some job postings mention multiple software with varying experience requirements however each and every job posting contains at least one software the applicant needs to be proficient in which is a pattern exemplified in this topic. Software engineering need not refer to the sole creation of software but rather work with software, including but not limited to troubleshooting, debugging and verifying the work others have done. In every case the individual at question is required to know the software's inner workings, its applications and limits.

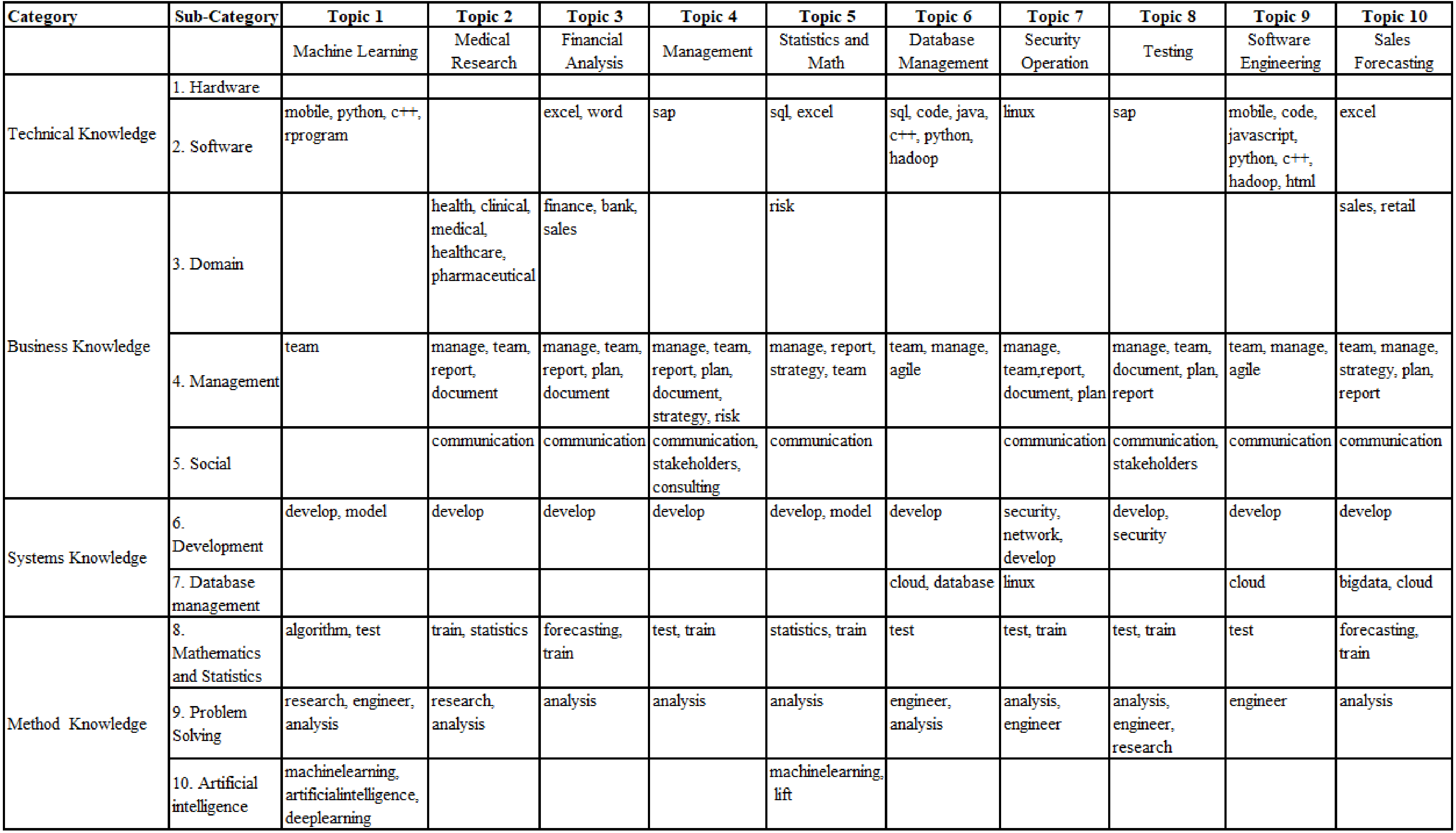
The last topic "Sales Forecasting" obtains its name from the top term "sales" and the corresponding terms such as "retail" and "forecasting". Sales Forecasting is more narrow of a skillset when compared to other topics such as "Statistics and Math" or "Management" and can be classified as a set of domain skills where sales expertise may be a significant advantage. The "forecasting" key word symbolizes the prediction of sales of products or services which is widely used in retail stores, banks and can serve a function in any industry which can be measured. Sales forecasting is associated strongly with some statistical or mathematical methods like exponential smoothing or various forms of regression, this is the reason it can be thought of as more domain rather than a broad skillset since it implies the application of those specific methods.

5.2.3 What are the required competencies for data professionals in Canada?

Results of the LDA analysis narrowly explain which skillsets are required by the industry as well as the most popular skills indicated in job postings. Knowing the granular skillsets and separate abilities, like which software, are needed categorizing them into larger groups allows for a more holistic view of the industry and all data professions. The Data Skills and Capabilities Classification sorts the skillsets into ACM categories for easier and broader interpretation. Summing the count of all terms in each sub-category results in a metric which approximates the popularity of the sub-category in the industry, out of 150 terms found by the LDA model each sub-category contains a portion and the largest portions display the importance of that sub-category. The top 5 sub-categories for job postings are “Management”, “Software”, “Mathematics and Statistics”, “Problem Solving” and “Development”.

“Management” is the most prominent sub-category containing 28% of all terms resulting from the LDA, management terms are seen in every topic with the highest numbers occurring in “Management”, “Financial Analysis”, “Security Operations”, “Testing” and “Sales Forecasting” with 7, 5,5,5 and 5 terms respectively. Half of the topics contain at least 33% of terms in the “Management” sub-category which emphasizes the importance of managerial skills among data positions. Many other researchers stressed this point and the results are in line with the current literature (Manyika et al., 2011) where managers of data are in much higher demand than the analysts of data. These results may allow certain educational programs to add managerial subject to their curricula, it also inadvertently promotes more business oriented education like Masters of Business Administration (MBA) where management is favored.

Table 10. Combination of LDA with Data Skills and Capabilities Classification



Category	**Sub-Category**	**Topic 1**	**Topic 2**	**Topic 3**	**Topic 4**	**Topic 5**	**Topic 6**	**Topic 7**	**Topic 8**	**Topic 9**	**Topic 10**
		Machine Learning	Medical Research	Financial Analysis	Management	Statistics and Math	Database Management	Security Operation	Testing	Software Engineering	Sales Forecasting
Technical Knowledge	1. Hardware										
	2. Software	mobile, python, c++, rprogram		excel, word	sap	sql, excel	sql, code, java, c++, python, hadoop	linux	sap	mobile, code, javascript, python, c++, hadoop, html	excel
Business Knowledge	3. Domain		health, clinical, medical, healthcare, pharmaceutical	finance, bank, sales		risk					sales, retail
	4. Management	team	manage, team, report, document	manage, team, report, plan, document	manage, team, report, plan, document, strategy, risk	manage, report, strategy, team	team, manage, agile	manage, team,report, document, plan	manage, team, document, plan, report	team, manage, agile	team, manage, strategy, plan, report
	5. Social		communication	communication	communication, stakeholders, consulting	communication		communication	communication, stakeholders	communication	communication
Systems Knowledge	6. Development	develop, model	develop	develop	develop	develop, model	develop	security, network, develop	develop, security	develop	develop
	7. Database management						cloud, database	linux		cloud	bigdata, cloud
Method Knowledge	8. Mathematics and Statistics	algorithm, test	train, statistics	forecasting, train	test, train	statistics, train	test	test, train	test, train	test	forecasting, train
	9. Problem Solving	research, engineer, analysis	research, analysis	analysis	analysis	analysis	engineer, analysis	analysis, engineer	analysis, engineer, research	engineer	analysis
	10. Artificial intelligence	machinelearning, artificialintelligence, deeplearning				machinelearning, lift					

Table 11. Term count in Job LDA

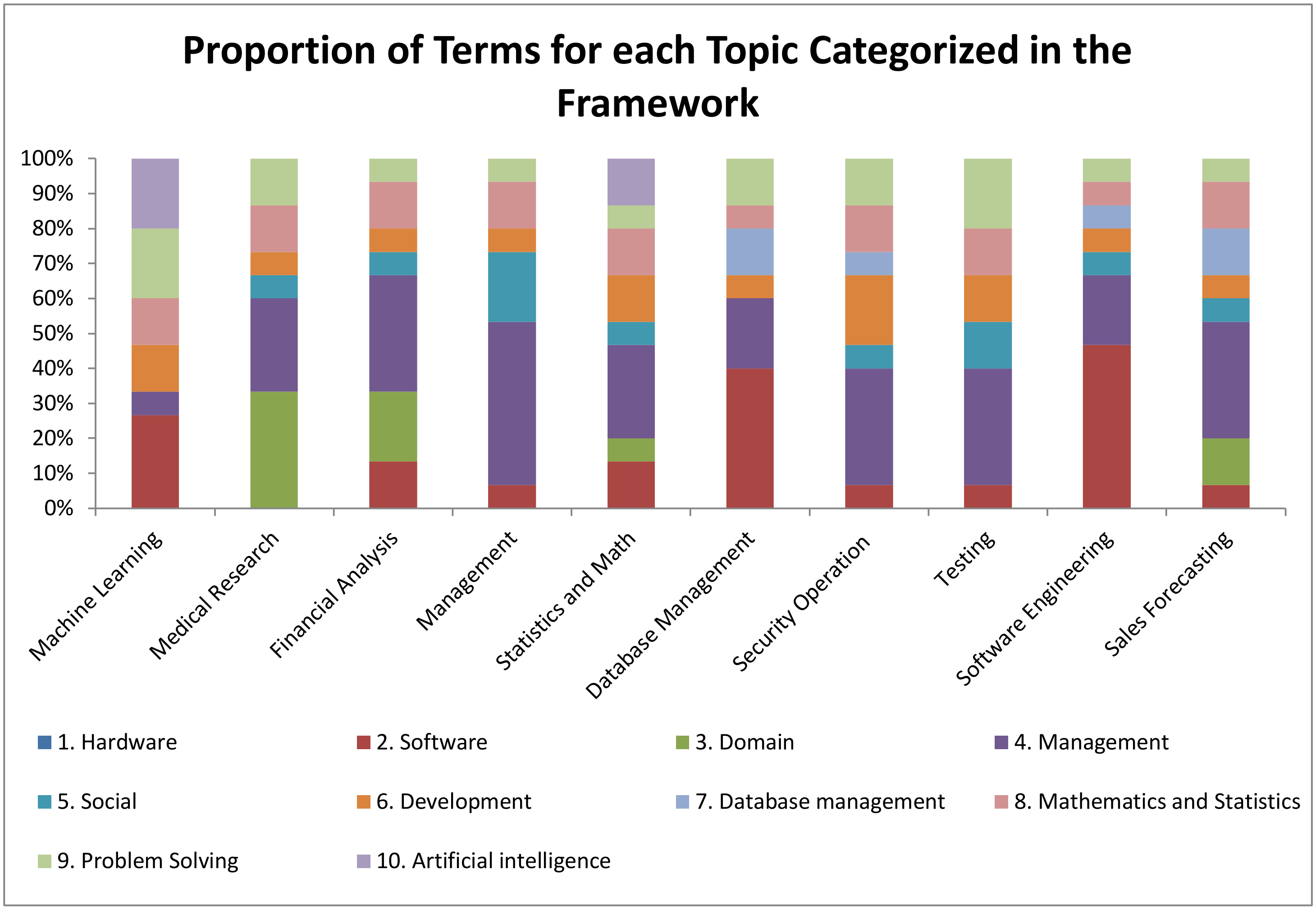
Category	Sub-Category	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
		Machine Learning	Medical Research	Financial Analysis	Management	Statistics and Math	Database Management	Security Operation	Testing	Software Engineering	Sales Forecasting
Technical Knowledge	1. Hardware	0	0	0	0	0	0	0	0	0	0
	2. Software	4	0	2	1	2	6	1	1	7	1
Business Knowledge	3. Domain	0	5	3	0	1	0	0	0	0	2
	4. Management	1	4	5	7	4	3	5	5	3	5
	5. Social	0	1	1	3	1	0	1	2	1	1
Systems Knowledge	6. Development	2	1	1	1	2	1	3	2	1	1
	7. Database management	0	0	0	0	0	2	1	0	1	2
Method Knowledge	8. Mathematics and Statistics	2	2	2	2	2	1	2	2	1	2
	9. Problem Solving	3	2	1	1	1	2	2	3	1	1
	10. Artificial intelligence	3	0	0	0	2	0	0	0	0	0

"Software" has the second most terms, 12% of the total, software terms appear in every skillset, except for "Medical Research", with the most prominent skillsets being "Software Engineering", "Database Management" and "Machine Learning". Technical knowledge of software is important since majority of data work is done using specific software. Different software is needed in different professions, for example machine learning professions are more inclined to work with C++, Python and R where database professions require SQL, Hadoop or Java.

Third most popular category is "Mathematics and Statistics" with 12% of the total terms this sub-category describes math oriented knowledge which lends itself to skillsets such as "Financial Analysis", "Sales Forecasting" and "Statistics and Math". These skillsets rely on mathematics or statistical formulas in conjunction with managerial and AI abilities. Mathematics are the base of all calculations and quantitative professions and it is natural for data jobs to require a healthy amount of quantitative knowledge.

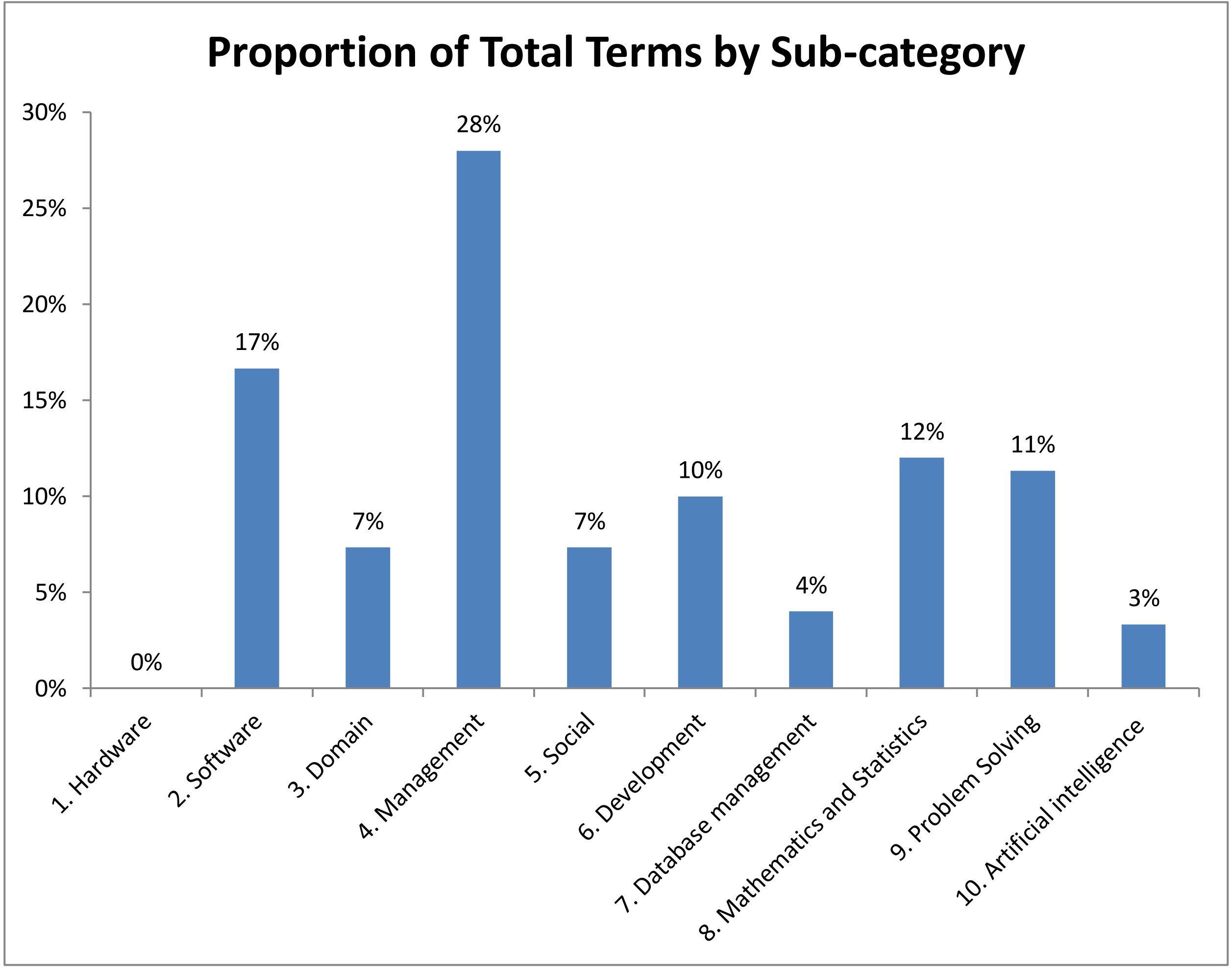
Fourth most popular category is "Problem Solving" which includes creative thinking in new situations with a certain amount of unknowns, problem solving itself is a broad category and should be used in all professions. Data shows that "analysis" or "engineering" solutions are required in every skillset stressing the importance of the skill. This subset in the original typology of Todd et al (1995) contained majority of terms due to its "catch all" nature, problem solving is an important skill to have in any managerial position and every data profession requires even slight amounts of it.

Figure 10. Proportion of Job Terms for each Topic Categorized in the Framework



The fifth most important sub-category is "Development" which is found to have 10% of the total terms and is also found among all skillsets. Development is the knowledge of systems, implementation, operations and maintenance, development lifecycle, and creation of tools necessary for success. In the skillset "Security Operation" 20% of the terms are in "Development" which is the highest percentage between all the topics indicating high involvement with systems creation and networks. Development of models and systems is therefore highly valued by the industry. The lowest sub-category is "Hardware" with 0% meaning no terms in the top 150 were associated with hardware, this may be due to the large amount of time data professionals spend on Business or Methods knowledge. Hardware itself is necessary for majority of data tasks and while these results downplay its importance, any software which is used needs to be on a hardware machine.

Figure 11. Proportion of Total Terms by Sub-category



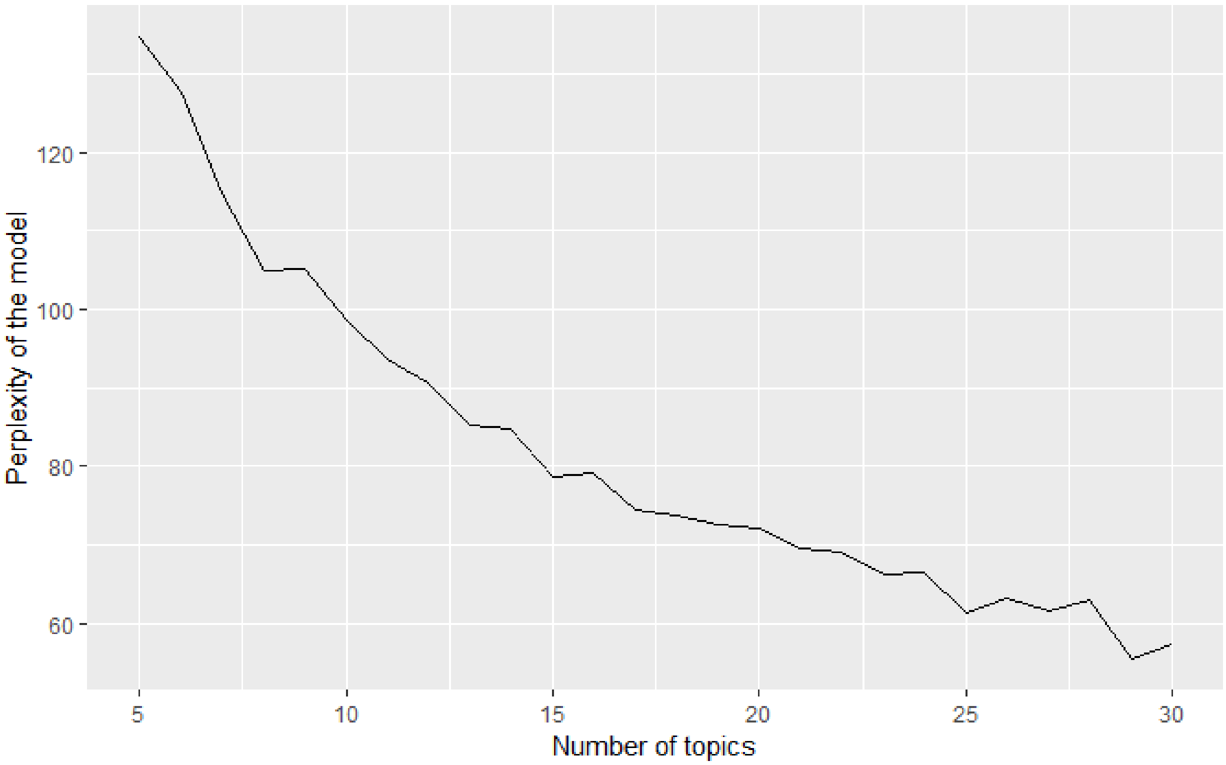
5.2.3 Educational categorization

This section identifies LDA topics for courses taught at graduate level programs.Table 12 shows the result of the LDA model for 11 topics as well as their terms. The number of courses varies between programs and not all the courses have same lengths of duration, in order to control for these variables, the courses are subject to a LDA model. The data was gathered from

Canadian universities offering programs which include keywords “Data Science”, “Big Data” and “Analytics” in their names. No specific domain was chosen for the schools as to obtain the most robust sample available. For the 218 courses offered among the 18 programs course descriptions were collected, cleaned and prepared for LDA analysis. Each program was thoroughly studied at the course level. Other research suggests qualitative content analysis using a deductive category assignment by the author (Schoenherr &Speier-Pero, 2015; Turel & Kapoor, 2016). The limitation of such a method is the results may be biased by the opinions or lack of expertise of the author. The objective of this step is to understand what topics are being taught in master level courses. Like the job advertisements, course LDA requires a pre-set number of topics prior to running. Once again the choice of the most optimal number of topics is done through selecting the lowest perplexity score which is k = 11. Figure 12 shows the perplexity scores for models run for topics 5 through 30. The analysed courses were grouped into 11 topics: Marketing Domain, Database Management, Statistics, Supply Chain Domain, Management, Machine Learning, Software Programming, Visualization, Mathematics, Forecasting and Prediction and Uncertainty Modeling, topics and terms can be seen in Table 12.

“Marketing Domain” is the first topic which resulted from the LDA model being run on all the courses available in the 18 data programs. The terms which identify this topic are “marketing” and “product” from the “Domain” sub-category as well as 20% of terms from the “Development” sub-category. This topic suggests that marketing is one of the more popular subjects taught at Canadian Universities or that the knowledge learned can be applied in the marketing industry.

Figure 12. Course LDA Topics vs. Perplexity



The second topic is labeled “Database Management” due to 40% of its terms being in the “Database Management” as well as having 4 (27%) terms in “Software” which relate to database management software such as SQL and Hadoop. Database management courses or its variations appear to be taught consistently across all programs, noting its importance by universities.

The “Statistics” topic contains terms in the “Statistics and Mathematics” sub-category which relate specifically to statistics and can be differentiated from the 9th topic “Mathematics”. Terms found in topic 3 also relate to software which is used for statistical analysis such as “rprogram” and “sas”. 10% of terms are included in the “Artificial Intelligence” sub-category, these terms do not sway the topic since the use of statistics is common in artificial intelligence and certain algorithms are used to produce AI.

Topic “Supply Chain” is a domain specific topic where the majority of terms are in the “Domain” sub-category and are associated with the supply chain industry. The top term “supplychain” is distinctive, similar to “transportation” and “distribution”, the terms “markovchain” and moderate amount of terms in “Database Management” creates this topic.

“Management” topic is named for the large amount of terms in the “Management” The topic called “Machine Learning” is named due to the terms “machine learning”, “artificial intelligence” and the large amount of terms in the “Artificial Intelligence” sub-category. The “Statistics and Mathematics” sub-category is also represented as the second most prominent category.

“Software Programming” is the name of the seventh topic and is named for its abundance of software related terms like “programming”, “coding” and “python”. The “Software” sub-category is the largest and contains 33% of the terms in the topic, the additional terms in “Artificial Intelligence” help label the topic since it includes the purpose the software is meant for. “Visualization” is the topic which centers around presenting information in a visual fashion, it gains its name from the ‘visualization” term as well as the multiple “Development” terms. The term “tableau” is a software for visualization which helps establish the label. Visualization of information is a skill mentioned in multiple job descriptions although it is not reflected in the current job LDA results. “Mathematics” is the topic which comprises of mathematical terms such as “mathematical” and “optimization”, “stochastic” and “linear”. It is a topic taught in the universities distinct from “Statistics” even though they share the “Statistics and Mathematics” sub-category, the “Mathematics” topic is focused on the theoretical aspects while “Statistics” is more akin to probability. “Forecasting and Prediction” is the tenth topic and it is labeled such due to the high amount of terms in the “Statistics and Mathematics” sub-category, different from

“Statistics” and “Mathematics” by the specific terms of “predictive”, “forecasting” and “smoothing”. These terms describe prediction as well as methods used for it, also using statistical software R and SAS. The last topic is “Uncertainty Modelling” certain courses share the name of this topic which is labeled for the terms “uncertainty” and “probability” from the “Statistics and Mathematics” sub-category and “model” and “develop” from “Development” sub-category.

Majority of the terms are concentrated in the “Method Knowledge”, summing the rows in order to know how many terms are in each sub-category results in an approximation of what is taught in Canadian Universities. Studying the results, it seems majority of the topics taught at Canadian Universities are quantitative in nature. Statistics, Mathematics, Forecasting and Prediction and Uncertainty Modelling have a strong mathematical base and share qualities and methods. With 23% of total terms “Statistics and Mathematics” is the category which is taught most often and is included in majority of the courses.

Second most popular sub-category is “Development” with 15% of total terms. Terms most often seen in that sub-category include “model”, “design” and “develop” which describe processes or results of work.

The third most popular sub-category is “Software” which includes a variety of statistical, database or programming languages. This sub-category suggests knowledge of software is an important part of training data professionals as 13% of all terms are located in this sub-category. Tied for third most popular subject is “Management”, also with 13% of all terms in the course LDA. Management is the knowledge of handling people, strategy, leadership and ethical conduct in the workplace, according to term frequency it is as important as “Software” knowledge. The lowest scoring sub-category is “Hardware” with only 1% and one term which the topic

“Uncertainty Modeling” houses, the term is “computer” which is used in nearly every data heavy profession.

Figure 13. Proportion of Course Terms for each Topic Categorized in the Framework

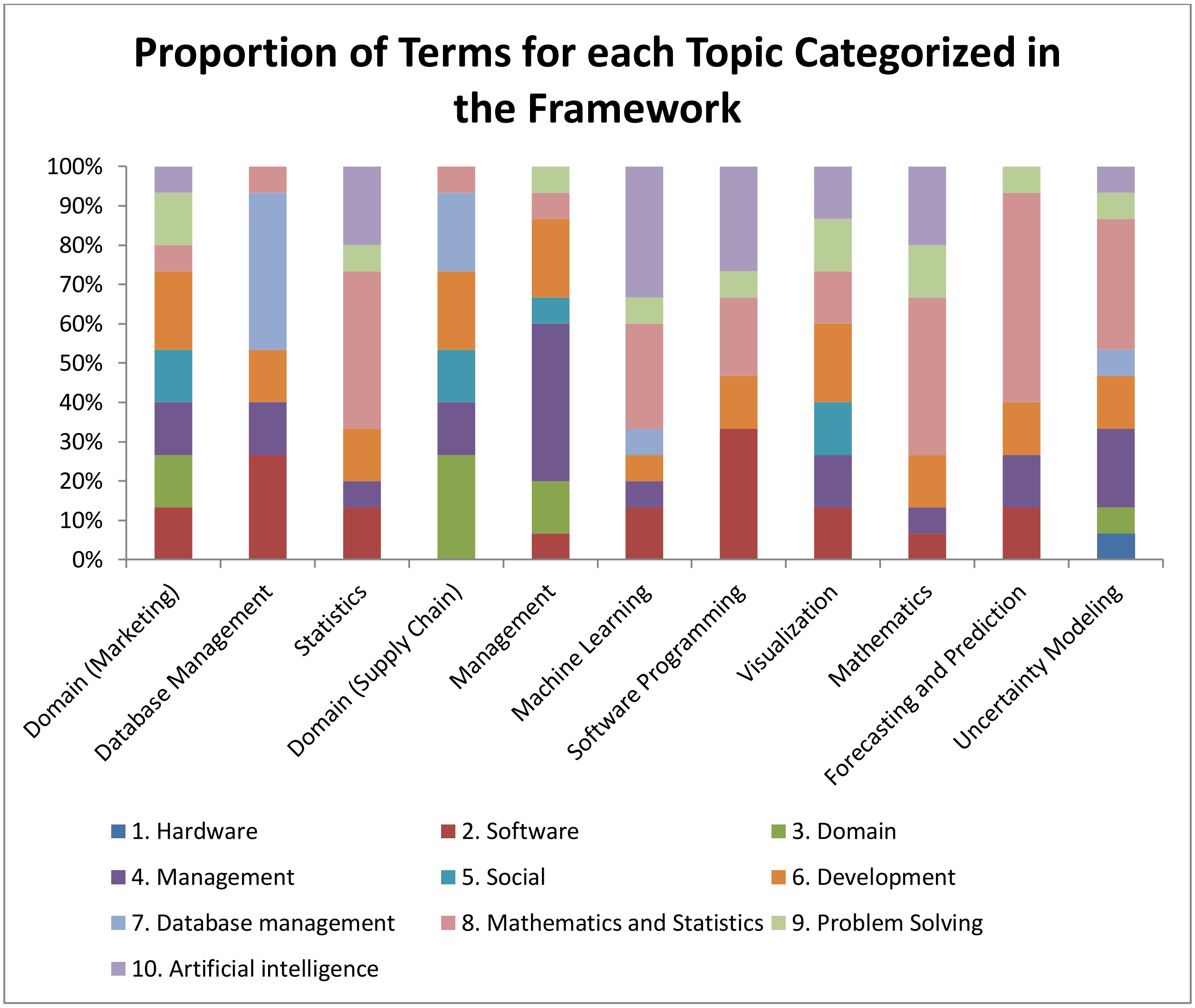


Table 12. Course LDA

Domain (Marketing)	Database Management	Statistics	Domain (Supply Chain)	Management	Machine Learning	Software Programming	Visualization	Mathematics	Forecasting and Prediction	Uncertainty Modeling
analysis	database	statistics	supplychain	lead	analysis	analysis	analysis	model	model	model
marketing	manage	machinelearning	design	manage	machinelearning	programming	model	probability	analysis	analysis
manage	relational	test	manage	develop	regression	predictive	visualization	mathematical	predictive	decision
communication	design	model	database	ethics	statistics	model	nlp	machinelearning	regression	simulation
develop	hadoop	probability	product	team	classification	rprogram	design	analysis	linear	risk
consulting	entityrelationship	regression	plan	plan	unsupervisedlearning	simulation	strategy	optimization	manage	uncertainty
strategy	mapreduce	linear	coordination	privacy	clustering	software	consulting	programming	forecasting	computer
model	markovchain	rprogram	security	strategy	bayesian	coding	neuralnetwork	manage	collaboration	decisiontrees
test	spark	clustering	system	finance	model	unsupervisedlearning	clean	svm	smoothing	stochastic
research	query	gain	model	analysis	team	python	develop	test	timeseries	system
software	schema	sas	simulation	operations	artificialintelligence	dimensionreduction	manage	stochastic	rprogram	markovchain
product	nosql	develop	present	forecasting	cloud	boosting	predictive	linear	develop	probability
design	team	analysis	transportation	excel	python	cluster	python	classification	arima	document
rprogram	system	manage	dataset	design	apis	optimization	optimization	develop	test	develop
gain	model	prediction	distribution	model	train	network	tableau	algebra	sas	manage

Table 13. Combination of Course LDA with Data Skills and Capabilities Classification

Category	**Sub-Category**	**Topic 1**	**Topic 2**	**Topic 3**	**Topic 4**	**Topic 5**	**Topic 6**	**Topic 7**	**Topic 8**	**Topic 9**	**Topic 10**	**Topic 11**
		Domain (Marketing)	Database Management	Statistics	Domain (Supply Chain)	Management	Machine Learning	Software Programming	Visualization	Mathematics	Forecasting and prediction	Uncertainty Modeling
Technical Knowledge	1. Hardware											computer
	2. Software	software, rprogram	hadoop, mapreduce, spark, noSQL	rprogram, sas		excel	python, APIs	programming, rprogram, software, coding, python	python, tableau	programming	rprogram, sas	
Business Knowledge	3. Domain	marketing, product			supplychain, product, transportation, distribution	finance, operations						risk
	4. Management	manage, strategy	manage, team	manage	manage, plan	lead, manage, ethics, team, plan, strategy	team		strategy, manage	manage	manage, collaboration	decision, document, manage
	5. Social	communication, consulting			coordination, present	privacy			visualization, consulting			
Systems Knowledge	6. Development	develop, model, design	design, model	model, develop	design, security, model	develop, model, design	model	model, network	model, design, develop	model, develop	model, develop	model, develop
	7. Database management		database, relational, entityrelationship, query, schema, system		database, system, dataset		cloud					system
Method Knowledge	8. Mathematics and Statistics	test	markovchain	statistics, test, probability, regression, linear, prediction	simulation	forecasting	regression, statistics, bayesian, train	predictive, simulation, optimization	optimization, predictive	probability, mathematical, optimization, stochastic, linear, algebra	predictive, regression, linear, forecasting, smoothing, timeseries,	simulation, uncertainty, stochastic, markovchain, probability
	9. Problem Solving	analysis, research		analysis		analysis	analysis	analysis	analysis, clean	analysis, test	analysis	analysis
	10. Artificial intelligence	gain		machinelearning, clustering, gain			machinelearning, classification, unsupervisedlearning, clustering, artificialintelligence	unsupervisedlearning, dimensionreduction, boosting, cluster	NLP, neuralnetwork	machinelearning, SVM, classification		decisiontrees

5.2.5 Do data skills taught in data programs impart the competencies required of data scientists?

Comparing the topics extracted from job postings with topics extracted from courses the outcomes show that the skills taught by data master`s programmes do not fit with the skills required by the industry. These results are exemplified in Turel and Kapoor's paper (2016). Table 14 contains the terms percentage represented in each of the sub-categories.

Table 14. Proportion of terms per Sub-Category

	Job	Education
1. Hardware	0%	1%
2. Software	17%	13%
3. Domain	7%	5%
4. Management	28%	13%
5. Social	7%	4%
6. Development	10%	15%
7. Database management	4%	7%
8. Mathematics and Statistics	12%	23%
9. Problem Solving	11%	7%
10. Artificial intelligence	3%	12%

According to job requirements "Hardware" as a set of skills is irrelevant. Considering the fact that most data professions require technology for work simple knowledge of hardware is beneficial however, extensive knowledge of inner working part of machines is not. Education understands only 1% of curriculum is required to train data professionals which is why majority of the terms are focused elsewhere. Software is the second most coveted sub-category by the industry and educational training fits with the amount of time dedicated towards teaching about software. Present programs concentrate on Method Knowledge with heavy emphasis on math and statistics which may be the base for analytical skills but is not the complete picture. The System Knowledge such as development and database management seem to fit well while personnel management and business skills seem to be understated. The results show that courses in areas such as database management and supply chain do not further students' analytical skills. This information raises question as to the differences between programs which teach these study fields and those which do not. One explanation would be that certain programs do not possess the knowledge or strategy of creating a curriculum ready for the ever-changing landscape of data science.

The analysis revealed most data programs do not consider the development of managerial skills as a priority, it seems the universities are more focused on the mathematical side of data. It is not a surprise that the classic knowledge is taught in school however the industry calls for skills which are domain specific. The domains most often seen among the job postings center on medical research and financial analysis while certain skills may be learned on the job, more competencies must be explored while in school. This method may be more employability oriented and lead to improving the market demand with the university teachings.

In order to compare what the educational institutions offer and what the market requires a direct comparison of LDA topics is beneficial. The first educational topic “Marketing” is not represented in the market needs or the skillsets it requires. The educational programs teach an abundant amount of courses in the marketing domain which is not represented in the needs the market is showing, this result does not criticize programs which contain marketing courses. Since there is no complete fit between the marketing domain and data professions some suggestions include decreasing the amount of courses taught or restructuring the marketing course curricula towards being more data oriented. Terms such as “rprogram” found in the topic are a sign the courses are already including statistical software in their teachings which can be a stepping stone towards more data driven marketing courses.

Database management from the course LDA results as a topic does fit into the market needs. Looking deeper into the topics shows almost no difference in terms between education and industry. Education’s “Database Management” sub-category consists of “database, relational, entityrelationship, query, schema, system" terms are comparable to job posting "Database Management" terms "cloud, database". Even though there are six terms in education and two in the job posting category terms such as "relational", "entityrelationship", "query", "schema" all relate to the term "database". The "Software" sub-category for education terms contains "hadoop, mapreduce, spark, noSQL" while job postings encompass "sql, code, java, c++, python, hadoop" the amount of like terms is visible such as "Hadoop" and "SQL". The high number of these comparable terms can conclude that students are indeed prepared well in the "Database Management" area.

The topics "Statistics" and "Mathematics" in the educational institutions must be compared to "Statistics and Math" topic from the industry since the market condensed the two while the courses LDA provided different results. In the industry, the "Statistics and Math" skillset contains only two terms from the "Statistics and Mathematics" sub-category while course "Statistics" and "Mathematics" both contain six terms each. The "Software" sub-category from the industry contains two terms, similarly "Statistics" and "Mathematics" contain two and one term respectively. Each of the topics contain terms in the "Artificial Intelligence" sub-category where the industry has two terms while education contains three each. Comparing the "Management" sub-category it is clear the industry possesses four terms while education contains one each showing a difference. The industry requires mathematic skills to be supplemented with managerial ones. Using this information, it can be concluded that students are over prepared and spend an overabundant amount of time on statistical and mathematic courses while not obtaining enough management talent. These results are consistent with the proportion of "Management" terms in the entire job posting corpus.

The educational topic “Supply Chain” is quite domain specific. In the industry LDA there is no comparable similar to the “Marketing” domain topic. Due to the seeming lack of demand for this topic in the industry the university programs can modify the curricula to be more data oriented or consider changing the name of the program. This is not to say the courses are without merit in the data industry, the lack of demand may indicate other fields and skills are more significant when compared.

“Management” in the educational institutions is quite comparable to the industry needs. The fact that both education and the market have the same topic is a sign of fit. Education contains six terms in the “Management” sub-category where the market contains seven, the difference is one term and a larger difference cay be seen in the “Social” sub-category where industry contains three terms while education only one. Education appears to contain more terms in the “Domain” sub-category while industry seems to contain one more in “Statistics and Mathematics”. The education and industry seem to be have a good fit in regards to “Management” however, when looking at the proportions of all terms the conclusion is not the same. As stated previously, the amount of terms in the overall industry required is much larger than the amount of terms education provides. Based on this knowledge, one conclusion may be to suggests increasing management courses.

The topic “Machine Learning” can be seen on both sides, the industry and education, which suggests a good fit. The education topic consists of five terms while the industry topic consists of three. The education’s sub-categories favored “Statistics and Mathematics” while the industry’s favored “Problem Solving”, both sides had “python” as a software however the industry had four terms to the education’s two in the “Software” sub-category. The amount of terms in each sub-category is comparable and thus it stands to reason the "Machine Learning" topics do prepare students well enough to enter the workforce.

"Software Programming" is a topic on the education's side while the industry possesses "Software Engineering". The difference in names results from the inclusion of the term "engineer" in the industry's topic while lacking in education. Software programming contains five terms in "Software" while Software Engineering contains seven terms. This discrepancy conveys that Software Engineering, the industry standard, requires a larger portion of knowledge dedicated to software. The total proportion of "Software" terms of the industry is also larger than the education's which signifies that software knowledge can be improved on the education's end.

"Visualization" is the eighth topic from the course LDA model and it too does not have a counterpart among the industry skillsets. Visualization itself has been promoted by multiple data professionals and researchers as an important tool for conveying information (Stukowski, 2009; Tansley & Tolle, 2009). The fact it is not in the top ten skillsets in the industry means it is not as essential as other skills but it is still taught.

"Forecasting and prediction" is a topic about calculating and estimating future patterns, educational courses often teach methods and techniques to predict future metrics. This topic is quite similar to the last topic seen on the industry's side: Sales forecasting. Both topics include the term "forecasting" in their titles as well as in their terms. The courses dealing with "Forecasting and prediction" focus on math and statistical approaches while the industry appears to contain more managerial and communication skills. The software on both sides are different however can be used to achieve the same outcomes, Excel may not be as sophisticated as SAS or R but can still use forecasting techniques.

The last topic is “Uncertainty Modeling” which is quite similar to “Forecasting and prediction” both in its use of statistics and the proportions of term distribution. Uncertainty modeling is also similar to “Sales Forecasting” found in the industry. The only topic containing terms in the “Hardware” sub-category and majority of its terms in the “Statistics and Mathematics” describes a large amount of courses dealing with risk, probability and uncertainty. It does not fit with “Sales Forecasting” to the highest degree as it contains no “Software” terms while “Sales Forecasting” does, no “Social” terms while “Sales Forecasting” does and it contains one “Artificial Intelligence” term while “Sales Forecasting” does not. It can thus be concluded that students are being taught skills which may be useful but not are not optimal.

Unmentioned skillset from the industry which is required is the “Testing” skillset. The characteristics of this skillset are various forms of analysis and simulations. Testing new formulas or simulating experiences is an important set of skills which is based on creative thinking and problem solving as seen from the 20% of terms which are in the “Problem Solving” sub-category. Testing requires innovative approaches to current problems or prevention of future problems. Having no comparable topics on the education’s side it can be concluded that students are not being prepared enough for this type of work

In summary the needs of the market are not met and do not fit with the education provided by the Canadian Universities in regards to subjects. The data skills taught in data programs do not impart the competencies required of data scientists in full amount. University programs and courses are called to improve their curricula in order to better train data professionals.

5.2.14 Contributions

This study focused on data occupations based on job descriptions to extract requirements and skills resembling the needs of the market. Contributions made enrich the literature by offering a structured review regarding skills, competencies and abilities from both the industry and academia for data professionals. In addition, I create a framework for further exploration of competency requirements in Data Science. Furthermore, a practical contribution is created in the form of a list compiled of Canada's graduate data programs included with significant keyword summaries, which can be used by any practitioner or novice seeking knowledge in this field. The results contribute to academia and industry by providing an understanding of what institutes the boundaries between occupations in Big Data in organizations, aiding human resources and other managers to improve recruitment and develop talent in the direction of their choice.

5.3 Implications

The literature is in agreement, for the most part, where the academia is slow to react to new trends in technology or economy especially to curriculum changes. In order to structure a proper curriculum for Big Data, Data Science or an Analytics program will require multiple attempts at uncovering the needs of the market, principal knowledge and standardizing methods. To remove the gap between what the market expects of graduate students and what knowledge the graduate students were imparted will require effort and fresh, ordered perspectives. Skills in analysis will continue to be demanded by the market since there is never a sufficient amount of intellect in the market, the skills however will continue to evolve to encompass more critical and complex business problems. Data professionals must obtain the skills to deconstruct old concepts and create new value using novel methods. Although current curricula adequately prepare students for data work they do so without a clear goal in mind. Greater emphasis needs to be placed on topics which are required demanded by the industry and not to waste time on topics which are not. Optimizing this curriculum may be difficult however a number of publications have shown evidence it will be valuable. Neither professionals nor academics can change the state of education alone. The issue is not to define the one curriculum for all data but rather how to structure various curricula so they may educate the learning who proves to be valuable to others and proud of their accomplishments. In order to do so other researchers must fine-tune the topics and methods currently imparted. This study is but a step in a direction of a more concise and specific curriculum to produce well educated, properly prepared data practitioners.

Chapter 6 – Conclusion

6.1 Conclusion

This study sought to determine whether there is congruence between the data and analytic skills demanded by the Canadian labour market and the education received by students (especially those enrolled in business and management faculties) entering the workforce. In conclusion, the results show evidence that there are discrepancies between the education of data professionals and the skills demanded by the workplace.

Graduate programs focussing on data and analytics do not adequately prepare students to enter into the industry. Results show that only 30% of graduate programs sufficiently teach the proper topics which are necessary for success in the industry, other programs mislabel their name and send unclear signals as to what is actually being taught. In addition, the number of graduates these programs produce is not sufficient to service the demand. Although Big Data education is a relatively new discipline, educational institutions must proliferate and expand their data programs to satisfy the demands of the industry.

One possible way of minimizing the job gap is to introduce Bachelor and undergraduate level programs alongside the graduate counterparts. With a newly compiled curriculum, undergraduate level programs can equip students with skills and knowledge to be successful in their work as well as produce a larger adequately skilled labour force to deal with the ever-growing demand for data-focussed jobs.

Based on the empirical findings, this study contributes to the development of a typology of data professionals and expands on the discussion whether existing educational curricula meet industry demands. This research enables educational institutions to critically evaluate their offerings based on their graduates' employability and adopt and revise curricula accordingly.

Some of the limitations of the study lie in its topic modeling approach, specifically its absence of significance measurements (Debortoli, 2016; Müller, 2016). Even though it has previously been addressed as a limitation (Murawski, 2017), topic modeling has been used in the past with success and with the increase of interest in NLP and Big Data methods, these issues can be rectified in the near future.

There are also some limitations in the curriculum data, since online syllabi only offer so much information there may be discrepancies between what is taught in a classroom and what is documented in a syllabus. In addition, while curricula are developed by educational standard boards such as Association for Computing Machinery (ACM), additional factors such as the experience of the educator must be considered for better understanding of what is being taught. The shortage of employees in the market cannot be solved by educational institutions alone and re-training of similar occupations will be necessary. Ultimately, research into educational standards is continually evolving, shifting with demands of the market. This study aims to uncover what the requirements are for data practitioners in Canada for the year 2018, what skills should be taught by educational institutions to be successful in a data industry and whether there is congruence between the two.

6.2 Limitations

Some of the limitations in this study include the methodology of LDA and topic modeling overall. There are no significance measures for topic modeling (Debortoli et al., 2016; Lau et al., 2014; Müller et al., 2016), some researchers have addressed this issue (Boyd-Graber et al., 2014) however no solutions are available at this time. The data itself also contains certain limitations such as the timeframe, amounts and geographical location. The data in question is gathered in August 2018 for a month on one online, job search engine which may not be a reliable sample. The sample of jobs used in this study is significantly larger than any which came before it but it does not mean the sample is reliable or valid to the fullest degree. The results are for the current state of data professions which may change with time as well as the graduate offerings and programs.

Another limitation is how the job posting was written in the company, who was it written by, whether it is standardized or whether the author is an expert in the field they are employing for. This study assumes the author knows exactly what skills and requirements the position should contain to be profitable to the company and the employee to be useful and valuable. The other information about the author is resource intensive to obtain, may be impossible to fully obtain a significant sample and a subject for an entirely different set of research questions and answers. The data came from an online source which means it did not originate from physical sources such as newspapers, word of mouth, job fairs and other physical or social sources. This method limits the scope of the study to only the online, even if it is the largest source of postings available on the market, it does not represent the population.

The categorization of graduate programs did not include hardware skills like the Todd et al. (1995) did, the conversion from the graduate categorization to Todd et al. may have lost some valuable information in the process. Even though some courses explicitly state "cloud" technology is taught in their description, it was not included in its fullest capacity in the calculations. This issue can be remedied by categorizing graduate programs through the updated Todd et al. framework.

Furthermore, graduate education is only one part of obtaining credentials and knowledge, this study does not take into account self-teaching which is a large area of education in information systems.

6.3 Future research

For future research opportunities the same methodology can be used to cover different topics or used in different domains. Other data related programs, not only graduate masters, can be used for future study such as doctorate and bachelor level programs. This study can be improved upon with a more granular perception through term-by-term comparisons. There are also multiple opportunities in researching re-training and cross-training of employees who have certain backgrounds and transitioned into data roles from various other domains.

References

Aguinis, H., Kraiger, K. (2009). Benefits of training and development for individuals and teams, organizations, and society. *Annual review of psychology*, *60*, 451-474.

Airoldi, E. M., Blei, D. M., Fienberg, S. E., & Xing, E. P. (2008). Mixed membership stochastic blockmodels. *Journal of Machine Learning Research, 9* , 1981–2014.

Association of College and Research Libraries. (2013). Working Group on Intersections of Scholarly Communication and Information Literacy. *Intersections of Scholarly Communication and Information Literacy: Creating Strategic Collaborations for a Changing Academic Environment*

Athey, S., & Plotnicki, J. (1988). A comparison of information system job opportunities for it professionals. Journal of Computer Information Systems, 38(3), 71-88

Baesens, B., Bapna, R., Marsden, J.R., Vanthienen, J., Zhao, J.L. (2014). Transformational issues of Big Data and Analytics in networked business. *MIS Quart*. 38(2), 629–632.

Bell, G., Hey, T., Szalay, A. (2009). COMPUTER SCIENCE: Beyond the Data Deluge. Science. 323 (5919): 1297–1298

Bhimani, A., & Willcocks, L. (2014). Digitisation, "Big Data" and the transformation of accounting information. Accounting and Business Research, 44 (4), 469-490.

Blei, D. M. (2012). Introduction to probabilistic topic models. *Communications of the ACM, 55* , 77–84. doi: 10.1145/2133806.2133826 .

Bouckaert, R. R., Frank, E., Hall, M. A., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2010). WEKA Experiences with a Java Open-Source Project. *Journal of Machine Learning Research*, *11*(Sep), 2533-2541.

Boutlon, C. (2013) "Much Hadoop About Nothing", *CIO Journal, The Wall Street Journal*,

Bowley, R. (2017) The Fastest-growing Jobs in the U.s. Based on Linkedin Data https://blog.linkedin.com/2017/december/7/the-fastest-growing-jobs-in-the-u-s-based-on-linkedin-data

Boyd-Graber, J., Mimno, D., and Newman, D. (2014). 'Care and feeding of topic models: problems,
diagnostics, and improvements.' In: *Handbook of Mixed Membership Models and Their Applications*. Ed. by E. Airoldi, D. Blei, E.A. Erosheva, and S.E. Fienberg, S.E. CRC Press, pp. 225-254.

Boyles, T. (2012). 21st century knowledge, skills, and abilities and entrepreneurial competencies: A model for undergraduate entrepreneurship education. *Journal of Entrepreneurship Education*, 15, 41- 55.

Brauer, C., & Wimmer, A. (2016). Der Mobile Analyst: Ein neues Berufsbild im Bereich von Business Analytics als Ausprägungsform von Big Data. *HMD Praxis der Wirtschaftsinformatik*, *53*(3), 357-370.

Bullen, C., Abraham, T., Galup, S. D. (2007). IT workforce trends: Implications for curriculum and hiring. *Communications of the Association for Information Systems.* 20, 545 - 554

Carillo, K. (2017). Let's stop trying to be "sexy" – preparing managers for the (big) data-driven business era. *Business Process Management Journal*, 598-622. https://doi.org/10.1108/BPMJ-09-2016-0188

Carnevale, A. P., Desrochers, D. M. (2002). The Missing Middle: Aligning Education and the Knowledge Economy.

Chatfield, A. T., Shlemoon, V. N., Redublado, W., Rahman, F. (2014). Data scientists as game changers in Big Data environments. *University of Wollongong*

Chen, H., Chiang, R.H.L., Storey, V.C. (2012). Business intelligence and Analytics: from Big Data to big impact. *MIS Quarterly*. 36 (4), 1165–1188.

Chiang, R. H. L., Goes, P., Stohr, E. A. (2012). Business intelligence and Analytics education, and program development: A unique opportunity for the information systems discipline. *ACM Transactions on Management Information Systems*, 3, 3, http://doi.acm.org/10.1145/2361256.2361257

Clarke, R. (2016). Big Data, big risks. *Information Systems Journal*. 26 (1), 77–90.

Colombo, M. G., & Grilli, L. (2010). On growth drivers of high-tech start-ups : Exploring the role of founders' human capital and venture capital. *Journal of Business Venturing, 25* , 610–626. doi: 10.1016/j.jbusvent.2009.01.005.

Constantiou, I.D., Kallinikos, J., 2015. New games, new rules: Big Data and the changing contextof strategy. *Journal of Information Technology*. 30 (1), 44–57

Coulter, N. (1997), ACM's computing classification system reflects changing times, Communications of the ACM, New York, NY, USA: ACM, 40 (12): 111–112

Croft, W. (2002). *Typology and universals*. Cambridge University Press.

Davenport, T.H. (2006). Competing on Analytics. *Harvard Business Review*. 84 (1), 98–107.

Davenport, T. H., & Patil, D. J. (2012). Data scientist. *Harvard business review, 90*(5), 70-76.

Provost, F., & Fawcett, T. (2013). Data Science and its relationship to Big Data and data-driven decision making. *Big Data*, *1*(1), 51-59.

KDD (2018) Data Mining Curriculum: A Proposal, SIGKDD https://www.kdd.org/curriculum/index.html

Debortoli, S., Müller, O., & vom Brocke, J. (2014). Comparing business intelligence and Big Data skills. *Business & Information Systems Engineering*, *6*(5), 289-300.

Delgado-Verde, M., Martín-De Castro, G., & Amores-Salvadó, J. (2016). Intellectual capital and radical innovation: Exploring the quadratic effects in technology-based manufacturing firms. Technovation, 54 , 35–47. doi: 10.1016/j.technovation.2016.02.002

De Veaux, R. D., Agarwal, M., Averett, M., Baumer, B. S., Bray, A., Bressoud, T. C., . . . Ye, P. (2016). Curriculum Guidelines for Undergraduate Programs in Data Science. *Annual Review of Statistics and Its Application,* (4), 15-30.

De Mauro, A., Greco, M., Grimaldi, M., & Ritala, P. (2018). Human resources for Big Data professions: A systematic classification of job roles and required skill sets. *Information Processing & Management*, *54*(5), 807-817.

Dhar, V. (2013). Data Science and prediction. *Communications of the ACM. 56 (12): 64*

Domo (2018) https://www.domo.com/learn/data-never-sleeps-6

Doty, D.H. and Glick, W.H. (1994). 'Typologies as a unique form of theory building: toward improved understanding and modelling.' *Academy of Management Review* (19) 2, 230-251.

Dubey, R., & Gunasekaran, A. (2015). Education and training for successful career in Big Data and Business Analytics. *Industrial and Commercial Training*, *47*(4), 174-181

Fairlie, R. W. (2010). Kauffman Index of Entrepreneurial Activity. *Ewing Marion Kauffman Foundation*

Finzer, W. (2013). The Data Science Education Dilemma. *Technology Innovations in Statistics Education*, 7(2). uclastat_cts_tise_13891. http://escholarship.org/uc/item/7gv0q9dc

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big Data concepts, methods, and Analytics. *International Journal of Information Management*, *35*(2), 137-144.

Gantz, J., Reinsel, D. (2012). The digital universe in 2020: Big Data, bigger digital shadows, and biggest growth in the far east. *IDC iView: IDC Analyze the future*, *2007*, 1-16

Gupta, B., Goul, M., Dinter, B. (2015) Business Intelligence and Big Data in Higher Education: Status of a Multi-Year Model Curriculum Development Effort for Business School

Undergraduates, MS Graduates, and MBAs. *Communications of the Association for Information Systems*, 36, 23.

Hansen, J. S., Ikemoto, G. S., Marsh, J. A., Barney, H. (2008). School finance systems and their responsiveness to performance pressures: A case study of North Carolina. RAND

Hey, T., Tansley, S., & Tolle, K. M. (2009). *The fourth paradigm: data-intensive scientific discovery* (Vol. 1). Redmond, WA: Microsoft research.

Hunt, K. (2004). The challenges of integrating data literacy into the curriculum in an undergraduate institution. *IASSIST Quarterly*, Summer/Fall, 12–16

Jafar, M. J., Babb, J., Abdullat, A. (2016). Emergence of Data Analytics in the Information Systems Curriculum. *EDSIG Conference*.

James, J., Maringer, D., Palade, V., & Serguieva, A. (2015). Special issue of quantitative finance on "financial Data Analytics" foreword. Quantitative Finance, 15 (10), 1617-1617.

Kedem, Z. et al. (2012) *The ACM Computing Classification System*. ACM

Kobayashi, M., & Takeda, K. (2000). Information retrieval on the web. *ACM Computing Surveys, 32* , 144–173. doi: 10.1145/358923.358934

Koltay, T. (2014). Big Data, big literacies? *TEMA*, 24, 3 -8.

Koza, J. R., Bennett, F. H., Andre, D., & Keane, M. A. (1996). Automated design of both the topology and sizing of analog electrical circuits using genetic programming. In *Artificial Intelligence in Design'96* (pp. 151-170). Springer, Dordrecht.

Laney, D. (2001). 3D data management: Controlling data volume, velocity and variety. *META group research note*, *6*(70), 1.

Lau, J.H., Newman, D., and Baldwin, T. (2014). 'Machine reading tea leaves: automatically evaluating topic coherence and topic model quality.' *Conference of the European Chapter of the Association for Computational Linguistics Proceedings*, pp. 530-539.

LeClair, D. (2016). Big Data's big future in business education. AACSB Blog

Lee, D.M., Trauth, E.M., and Farwell, D. (1995), 'Critical skills and knowledge requirements of IS professionals: a joint academic/industry investigation.' *MIS Quarterly* (19) 3, 313-340

Leek, J. (2013). The key word in "Data Science" is not Data, it is Science. *Simply Statistics*

Litecky, C., Aken, A., Prabhakar, B., and Arnett, K. (2009). 'Skills in the MIS job market.' *Americas*
Conference on Information Systems Proceedings, Paper 225.

Lusher, S. J., McGuire, R., van Schaik, R. C., Nicholson, C. D., & de Vlieg, J. (2014). Data-driven medicinal chemistry in the era of Big Data. *Drug discovery today*, *19*(7), 859-868.

Lyon, L., Mattern, E. (2016). Education for Real-World Data Science Roles (Part 2): A Translational Approach to Curriculum Development. *International Journal of Digital Curation, 11*(2), 13-26. doi:10.2218/ijdc.v11i2.417

Lyon, L., Brenner, A. (2015). Bridging the Data Talent Gap: Positioning the iSchool as an Agent for Change. *International Journal of Digital Curation, 10*(1). doi:10.2218/ijdc.v10i1.349

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., Byers, A. H. (2011). Big data: The next frontier for innovation, competition and productivity. *McKinsey Global Institute.* http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation

Mayer-Schönberger, V., & Cukier, K. (2014). *Learning with Big Data: The future of education*. Houghton Mifflin Harcourt.

McAfee, A. and Brynjolfsson, E. (2012). 'Big Data: the management revolution.' *Harvard Business*
Review (90) 10, 61-67

McKendrick, J. (2015). Data driven and digitally savvy: The rise of the new marketing organization. *Forbes Insights*.

McLeod, A. J., Bliemel, M., Jones, N. (2017). Examining the adoption of Big Data and Analytics curriculum, *Business Process Management Journal*, 3, 506-517. https://doi.org/10.1108/BPMJ-12-2015-0174

Miller, S. (2014). Collaborative approaches needed to close the Big Data skills gap. Journal of Organization Design, 3, 26–30. doi:10.7146/jod.3.1.9823.

Morales-Alonso, G., Pablo-Lerchundi, I., & Núez-Del-Río, M. C. (2016). Entrepreneurial intention of engineering students and associated influence of contextual factors. International Journal of Social Psychology . doi: 10.1080/02134748.2015.1101314

Müller, O., Junglas, I., Vom Brocke, J., and Debortoli, S. (2016). 'Utilizing Big Data Analytics for
information systems research: challenges, promises and guidelines.' *European Journal of Information Systems* (25) 4, 289-302.

Müller, O., Schmiedel, T., Gorbacheva, E., and Vom Brocke, J. (2014). 'Towards a typology of business process management professionals: identifying patterns of competences through latent semantic analysis.' *Enterprise Information Systems* (10)1, 50-80.

Munroe, R. (2013) Google's Datacenters on Punch Cards https://what-if.xkcd.com/63/
Nasrabadi, N. M. (2007). Pattern recognition and machine learning. *Journal of electronic imaging*, *16*(4), 049901.

Negash, S. (2004). Business intelligence. *Communications of the Association for Information Systems*, 13, 177 - 195.

Newell, S., Marabelli, M. (2015). Strategic opportunities (and challenges) of algorithmic decision-making: a call for action on the long-term societal effects of 'datafication'. *Journal of Strategic Information Systems*. 24 (1), 3–14.

N.N. (2016). *25 Best Jobs in America*. URL: https://www.glassdoor.com/List/Best-Jobs-in-America-
LST_KQ0,20.html

Noll, C.L. and Wilkins, M. (2002). 'Critical skills of IS professionals: a model for curriculum development.' *Journal of Information Technology Education* (1) 3, 143-154.

Nordhaug, O. (1993). *Human Capital in Organizations: Competence, Training, and Learning*, Oslo,
New York: Scandinavian University Press; Oxford University Press

Nunamaker Jr, J. F., Couger, J. D., & Davis, G. B. (1982). Information systems curriculum recommendations for the 80s: undergraduate and graduate programs. *Communications of the ACM*, *25*(11), 781-805.

OECD Skills Outlook (2013). http://skills.oecd.org/OECD_Skills_Outlook_2013.pdf

Persistence Market Research (2018) *Data Science Platform Market: Global Industry Analysis and Forecast 2017 - 2025*

Prabhakar, B., Litecky, C. R., & Arnett, K. (2005). IT skills in a tough job market. *Communications of the ACM*, *48*(10), 91-94.

Provost, F., & Fawcett, T. (2013). Data Science and its relationship to Big Data and data-driven decision making. *Big Data*, *1*(1), 51-59.

Qin, J., D'Ignazio, J. (2010). The central role of metadata in a science data literacy course. *Journal of Library Metadata*, 10(2-3), 188-201. DOI: 10.1080/19386389.2010.506379

Ransbotham, S., Kiron, D., Prentice, P.K. (2016). Beyond the hype: the hard work behind Analytics success. *MIT Sloan Management Review*. 57 (3), 3–16.

Rao, M. S. (2014). Enhancing employability in engineering and management students through soft skills. *Industrial and Commercial Training*, *46*(1), 42-48.

Ridsdale, C., Rothwell, J., Smit, M., Ali-Hassan, H., Bliemel, M., Irvine, D., ... Wuetherick, B. (2015). Strategies and Best Practices for Data Literacy Education: *Knowledge Synthesis Report.*

Rous, B. (2012) *Major update to ACM's Computing Classification System*. ACM

Schoenherr, T., & Speier-Pero, C. (2015). Data Science, predictive Analytics, and Big Data in supply chain management: Current state and future potential. *Journal of Business Logistics, 36*(1), 120-132

Schumann, C., Zschech, P., and Hilbert, A. (2016). 'Das aufstrebende Berufsbild des Data Scientist.' HMD Praxis der Wirtschaftsinformatik (53) 4, 453-466.

Shrimplin, A. K., Yu, J. C. (2004). Focusing in on student learning outcomes: How SDA helped us get data into the classroom. *IASSIST Quarterly, 28*, 55-57.

Siepel, J., Cowling, M., & Coad, A. (2017). Non-founder human capital and the long-run growth and survival of high-tech ventures. Technovation., 59 , 34–43. doi: 10.1016/j.technovation.2016.09.001

Sircar, S. (2009). Business intelligence in the business curriculum. *Communications of the Association for Information Systems*, 24, 289 - 302.

Song, I.-Y., & Zhu, Y. (2015). Big Data and Data Science: What should we teach? *Expert Systems* . doi: 10.1111/exsy.12130 .

Stukowski, A. (2009). Visualization and analysis of atomistic simulation data with OVITO–the Open Visualization Tool. *Modelling and Simulation in Materials Science and Engineering, 18*(1), 015012.

Tambe, P. (2014). 'Big Data investment, skills, and firm value.' Management Science (60) 6, 1452-1469.

Tansley, S., & Tolle, K. M. (2009). *The fourth paradigm: data-intensive scientific discovery* (Vol. 1). A. J. Hey (Ed.). Redmond, WA: Microsoft research.

Targett, D. (1991). 'The ITBE 2000 initiative: are business schools meeting the challenge of management and IT in the 21st century? Case study.' *Journal of Strategic Information Systems* (1) 1, 43-46.

Teal, T. K., Cranston, K. A., Lapp, H., White, E., Wilson, G., Ram, K., Pawlik, A. (2015). Data carpentry: workshops to increase data literacy for researchers. *International Journal of Digital Duration, 10*(1), 135-143.

Todd, P. A., McKeen, J. D., & Gallupe, R. B. (1995). The evolution of IS job skills: a content analysis of IS job advertisements from 1970 to 1990. *MIS quarterly*, 1-27.

Topi, H., Valacich, J. S., Wright, R.T., Kaiser, K., Nunamaker, J., Sipior, J. C., de Vreede, G. J.

(2008). Revising Undergraduate IS Model Curriculum: New Outcome Expectations, *Communications of the Association for Information Systems*, 23, 32.

Trauth, E. M., Farwell, D. W., & Lee, D. (1993). The IS expectation gap: Industry expectations versus academic preparation. *Mis Quarterly*, 293-307.

Turel, O. and Kapoor, B. (2016). 'A Business Analytics maturity perspective on the gap between business schools and presumed industry needs.' *Communications of the Association for Information Systems* (39) Article 6.

Ullman, J. D. (2011). *Mining of massive datasets*. Cambridge University Press.

Univcan (2018) https://www.univcan.ca/universities/facts-and-stats/enrolment-by-university/

Ulrich, D., & Dulebohn, J. H. (2015). Are we there yet? What's next for HR? Human Resource Management Review, 25 (2), 188-204

Walsh, J., Johnson, M., & Sugarman, M. (1975) Help Wanted: Case Studies of Classified Ads, Olympus Publishing Company Salt Lake City

Watson, H. J. (2009). Tutorial: Business intelligence past, present, and future. *Communications of the Association for Information Systems*, 25, 487 - 510

Whitmer, A., Blanchette, C., Caron, B. (2004). Bringing real-time data into the ocean science. *Carleton University*

Wixom, B., Ariyachandra, T., Goul, M., Gray, P., Kulkarni, U., and Phillips-Wren, G. (2011). 'The
current state of business intelligence in academia.' *Communications of the Association for Information Systems* (29) Article 16.

Wixom, B., Ariyachandra, T., Douglas, D., Goul, M., Gupta, B., Iyer, L., Kulkarni, U., Mooney, J., Phillips-Wren, G., Turetken, O. (2014). The Current State of Business Intelligence in Academia: The Arrival of Big Data. *Communications of the Association for Information Systems*, 34, 1.

Zawadzki, K. (2014). *Data Science Skill-Set Explained*. URL:
http://www.marketingdistillery.com
/2014/08/30/data-science-skill-set-explained/

Zikopoulos, P. C., Eaton, C., Deroos, D., Deutsch, T., & Lapis, G. (2012). Understanding Big Data: Analytics for enterprise class Hadoop and streaming data. McGraw - Hills. http://public.dhe.ibm.com/common/ssi/ecm/en/iml14296usen/IML14296USEN.PDF

https://Analytics.ncsu.edu/?page_id=4184 https://www.stoodnt.com/blog/top-data-science-and-Analytics-programs-in-canada-best-big-data-Analytics-courses-in-canada/